Identifying and characterizing transcriptional regulatory elements from chromosome conformation capture data

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Michelle N. Yancoskie

aus Wiesbaden

Tübingen

2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	05.11.19
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Dr. Frank Chan
2. Berichterstatter:	Prof. Dr. Alfred Nordheim

“Wisdom and might are His. He changes the times and the seasons;
He removes kings and raises up kings;
He gives wisdom to the wise
and knowledge to those who have understanding.
He reveals deep and secret things;
He knows what is in the darkness, and light dwells with Him.”

–Daniel 2:20-22 (The Bible, New King James Version)

Acknowledgements

I would like to thank:

My advisor, Frank, for much support, kindness, and encouragement and for teaching many invaluable lessons – and not just about science. It has been a privilege to work with and learn from you.

Bruce Blumberg, Professor of Developmental and Cell Biology at UCI, for sharing wisdom, humor, and sound advice through countless office hours and emails. I am glad to call you a friend, pen pal, and mentor.

Dr. Felicity Jones for consistently helpful suggestions and guidance.

My current and past thesis committee members and evaluators: Prof. Alfred Nordheim for preparing the written review and providing early comments on this thesis, and Profs. Detlef Weigel, Klaus Harter, Bernard Moussian, and Remco Sprangers for their scientific input and advice.

Our Longshanks collaborators: Profs. Campbell Rolian, Nick Barton and John Cobb, and their lab members.

Our collaborators, Prof. Jiří Forejt and Dr. Petr Jansa, for their hard work on the hybrid sterility manuscript.

Ludmila, who made a huge impression on me from the start. You are an example worth following, with many commendable qualities – especially your kindness, quick wit and on-point assessments of all sorts of situations.

Stan for often applying his experience and considerable knowledge of the field to suggest useful project directions and improvements, and for helping me get through rough days.

Marek: You have a knack for many things. One of them is making me laugh.

Elena for commiserating, for calling me out when I'm being unreasonable, and for many intriguing discussions about God, morality, and mortality.

Muhua for all sorts of help and career advice.

Andreea for much laughter. I enjoy your brains, nerdy humor, and enthusiasm.

Enni for manuscript help, organizing events, and staying positive and supportive.

Stefano for playful banter. You excel at "layered speech", as Elena puts it.

João and Kavita for project insights and for help troubleshooting experiments.

Li and Bill for kind gestures and for benchwork and coding help.

Vrinda, Melanie, Layla, and Insa for tasty desserts and tasteful conversation.

JP and Saad for sharing code, advice and helpful papers.

Sophia Flad for her efficient and skillful benchwork.

Johanna Geiger, Aylin Yigitliler, and Leonie Keller for benchwork support.

Caroline Schmidt for taking care of the mice, Ronald Naumann (Transgenic Core Facility, Dresden) for transgenic mouse work, and Christa Lanz, Rebecca Schwab, and Ilja Bezrukov for high-throughput sequencing support.

Dagmar Sigurdardottir for thesis feedback and for advice how to get through the last stretch of the marathon.

Herta Soffel for administrative support.

Claire Siebenmorgen and Volker Soltys for help translating my summary into German.

Thomas Pavelka for translation help, and for always encouraging me to press toward the goal. “Schneller voran”!

Dr. Tobias Widmer and his wife Angi for kindly hosting me at their house and for imparting much wisdom.

Mr. Brunner and his wife Jo for their continued friendship and advice. Your seventh-grade life science class and Jo’s guest lecture is where it all started!

My dear family for their continued love and support.

Abstract

This study aims to characterize the chromatin features associated with long-range transcriptional regulation. Transcriptional regulation is a highly complex process mediated by the action of many features in the nuclear environment that affect promoter activity, gene expression and, through gene interaction networks, phenotypic output. The contacts formed by promoters are a useful readout of transcriptional regulation and can be revealed by chromosome conformation capture (3C) assays. These contacts likely work through interaction with transcriptional machinery, but it is still unclear how they find the target promoter, or which chromatin features confer transcriptional regulatory function. Potentially relevant features of promoter contacts include the presence or absence of histone tail modifications, and local chromatin structure and accessibility. In this work, I attempt to connect promoter contacts with transcriptional output by integrating 3C and other datasets at several loci in the mouse genome. In **Chapter 2**, I explored the genomic features of promoter contacts, including chromatin accessibility, GC content class, distance from the promoter, and enrichment for nearby histone modification marks. I found that promoter contacts were often located in the same topologically associating domain and that the correlation between promoter contact frequency and each chromatin feature varied across promoter gene expression level, with poised promoters less constrained than active or silent promoters when forming contacts. I explored these features and applied the principles in a mouse selective breeding experiment for longer tibia called “Longshanks” (**Chapter 3**). At loci that show evidence of selection, we identified putative enhancers by chromatin accessibility and histone modification marks, linked them to their target promoters by 4C-seq, a variation of 3C, and functionally validated them with transgenic reporter assays. This allowed us to identify the molecular changes at enhancers that we hypothesize to encode for gain-of-function of the limb growth activator *Gli3* and loss-of-function of the limb growth repressor *Nkx3-2*, contributing to longer leg length in mice under selection. These studies reveal differences in contact features across promoter gene expression levels and underline the key role of transcription factors in conferring transcriptional regulatory function. Characterizing the chromatin features that enable transcriptional regulation should facilitate our understanding of how gene expression is precisely regulated during development and how it is altered during evolution or disease.

Zusammenfassung

Das Ziel der Studie war es, die Chromatineigenschaften zu charakterisieren, welche mit der weitreichenden Transkriptionsregulation assoziiert werden. Die Transkriptionsregulation ist ein hochkomplexer Prozess, der durch das Zusammenwirken vieler Eigenschaften in der Zellkernumgebung beeinflusst wird, welche die Promotoraktivität und Genexpression beeinflussen und schließlich durch Gen-Interaktionsnetzwerke den phänotypischen Output. Die Kontakte, welche durch die Promotoren gebildet werden, können mithilfe des “chromosome conformation capture (3C)” Tests aufgezeigt werden und zur Analyse der Transkriptionsregulation genutzt werden. Diese Kontakte können den Output der Promotoren beeinflussen durch Interaktion mit der transkriptionellen Mechanismen. Dabei ist jedoch unklar, wie sie die Zielpromotoren finden und welche Chromatineigenschaften die Transkriptsionsregulatoren übertragen. Potenziell relevante Eigenschaften der Erbgutsequenz von Promotorkontakten beinhalten die An- oder Abwesenheit von Histonmodifikationen sowie die lokale Chromatinstruktur und dessen Zugänglichkeit. Das Ziel dieser Studie war es, Promotorkontakten mit dem transkriptionellen Output zu verknüpfen, indem ich 3C und andere Datensätze an mehreren Loci im Mausgenom integriere. In **Kapitel 2** habe ich die Eigenschaften der Ergbutsequenz von Promotorverbindungen untersucht, einschließlich der Chromatinzugänglichkeit, den GC-Verhältnissen, die Entfernung zum Promotor und die Anreicherung von nahegelegene Histonmodifikationen. Ich konnte zeigen, dass die Promotorkontakten sich häufig in denselben topologisch assoziierten Domänen befanden und dass die Korrelation zwischen der Häufigkeit der Promotorverbindungen und den verschiedenen Chromatineigenschaften je nach Genexpressionsniveau variierte, wobei einsatzbereite Promotoren weniger eingeschränkt waren als aktive oder inaktive Promotoren, wenn sie Kontakte aufbauten. In **Kapital 3** habe ich diese Merkmale untersucht und gefundene Prinzipien auf eine künstliche Ausleseexperiment in Mäusen namens „Longshanks“ (selektierend auf längere Oberschenkelknochen) angewandt, um die molekularen Grundlagen an zwei Loci, welche Hinweise auf Selektion zeigten, aufzuklären. An den zwei Loci wurden mutmaßliche Enhancer durch Chromatinzugänglichkeit und durch Histonmodifikationen identifizierte, und durch 4C-seq (eine 3C-Variante) miteinander verknüpft und schließlich durch transgene Reportertests funktionell validiert. Dies ermöglichte es uns, die molekularen Veränderungen an den Enhancern zu

identifizieren, von denen wir vermuten, dass sie einen Funktionsgewinn des Extremitätenwachstumsaktivators *Gli3* und einem Funktionsverlust des Extremitätenwachstumsrepressors *Nkx3-2* codieren, was zu einer längeren Beinlänge bei Mäusen in der künstlichen Auslese führte. Diese Studien zeigen Unterschiede in den Kontakteigenschaften zwischen den Genexpressionsniveaus und schildern die Schlüsselrolle der Transkriptionsfaktoren bei der Übertragung der Transkriptionsregulationsfunktionen. Die Charakterisierung der Chromatineigenschaften, die Transkriptionsregulation ermöglichen, sollte unser Verständnis vertiefen, wie die Genexpression während der Entwicklung genau reguliert wird und wie sie sich während der Evolution oder Krankheit verändert.

Translated and proofread by Thomas Pavelka, Claire Siebenmorgen, and Volker Soltys.

Table of Contents

Abbreviations

3C	Chromosome conformation capture
4C-seq	Circularized chromosome conformation capture followed by (high-throughput) sequencing
5C	Chromosome conformation capture carbon copy
ANCOVA	Analysis of Covariance
ATAC-seq	Assay for Transposase-Inducible Chromatin followed by sequencing
BAC	Bacterial artificial chromosome
BL6	C57BL/6NJ mouse strain
bp	base pair
Capture-C	A high-throughput 3C derivation which captures the chromatin contacts of up to several hundred viewpoints at once
CBP	CREB-binding protein
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CRE	*cis*-regulatory element
Cre	*Cre* recombinase (used to drive *Cre-lox* recombination)
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
DHS	DNase hypersensitive site
eRNA	Enhancer RNA
E13.5	Embryonic day 13.5 (13.5 days after fertilization)
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements followed by sequencing
FLASH	Fast Length Adjustment of SHort reads
GRO-seq	Global nuclear run-on sequencing
FISH	Fluoresence *in situ* hybridization
H3K4me1	Histone H3, lysine 4 monomethylation
H3K4me3	Histone H3, lysine 4 trimethylation
H3K27ac	Histone H3, lysine 27 acetylation
H3K27me3	Histone H3, lysine 27 trimethylation
Hi-C	A high-throughput 3C-seq derivative which captures genome-wide interacting loci
ISH	*In situ* hybridization
kbp	kilobase pair
LAD	Laminar associating domain
Mbp	Megabase pair
mESC	Mouse embryonic stem cells
mm10	*Mus musculus* genome assembly GRCm38 from Genome Reference Consortium
MNase-seq	Micrococcal nuclease-sequencing
MPRA	Massively parallel reporter assay
PCR	Polymerase Chain Reaction
PRC	Polycomb Repressive Complex
STARR-seq	Self-Transcribing Active Regulatory Region sequencing
TAD	Topologically associating domain
TF	Transcription factor

Gene names from the Capture-C analysis are listed in **Table 2.A**.

List of Figures

List of Tables

Chapter 1: General Introduction

1.1 Objectives of this study

This study attempts to characterize the genomic features of DNA regions that contact gene promoters in order to determine which of these features are important for regulating transcription. In mammals, these long-range elements often skip over closer promoters to act on targets that are up to tens of hundreds of kilobase pairs away. In its native conformation, chromatin forms looping structures that bring distal loci into contact with one another. These chromatin loops serve as a proxy for transcriptional activity and can be captured by chromosome conformation capture (3C) assays. However, only a subset of the regions contacting a given promoter actually regulates its transcriptional output. In this work, I investigate the following questions:

1. Where do the regions that regulate transcriptional output lie with respect to their target promoters?
2. How well do various genomic features predict how often a given region contacts a promoter? Are there features that should be prioritized when determining which promoter contacts to functionally validate?
3. How do these genomic features vary with promoter activity level?

In **Chapter 2**, I compared the contacts from 25 promoters on mouse Chromosome 7 with other datasets to determine which features are most highly correlated with promoter contact frequency. In **Chapter 3**, we discovered that two genes are relevant to the rapid increase in leg length in a mouse selective breeding experiment called Longshanks. These two genes are among up to a hundred genes known to regulate limb development. Despite our knowledge of limb developmental genetics, only a handful of these genes have been linked to natural variation in leg length in human or mice, and none with molecular detail. Here, my work contributed to the identification of naturally occurring variants within their enhancers and our characterization of their associated chromatin features provided new evidence on how this trait may have evolved in response to artificial or potentially natural selection.

Answering these questions should aid efforts to efficiently and accurately link promoters to the regions that transcriptionally regulate them. This, in turn, should facilitate identification of causal mutations that drive changes in gene expression during disease or evolution, and help elucidate the transcriptional regulatory mechanisms that enable precise control of gene expression during development.

1.2 Properties of *cis*-regulatory elements

There are only an estimated 21,000 protein-coding genes in the human and mouse genomes (Mouse Genome Sequencing Consortium, 2002; ENCODE Project Consortium, 2012), but potentially an infinite number of possible distinct gene expression profiles conferring cellular identity. This complexity is possible through the combined action of *cis*-regulatory elements and *trans*-acting factors (Fakhouri et al., 2010) which determines gene expression in a given cell type or developmental stage. *Cis*-regulatory elements (CREs) are DNA sequences to which *trans*-acting factors like transcription factors (TFs) and their cofactors, RNA polymerases, chromatin remodelers, histone-modifying enzymes, and chromatin looping factors like cohesin localize to affect transcriptional activity and subsequently gene expression (Meiklejohn et al., 2014; Long et al. 2016) (**Figure 1.1**). CREs include promoters, enhancers, silencers, and insulators. With the exception of promoters, we know little about what determines their activity and function, yet as argued above, they are thought to be key contributors to complexity during vertebrate evolution. This is the core motivation of our study.

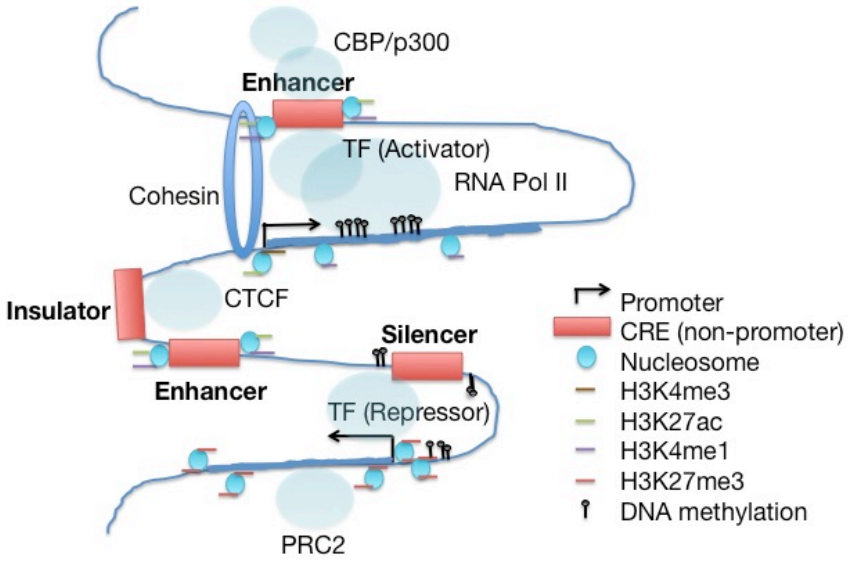


Figure 1.1. Features of the chromatin and nuclear environment that enable transcriptional regulation. Typical *cis*-regulatory elements – including (black arrows) promoters and (red boxes) enhancers, silencers, and insulators – interact with *trans*-acting factors (green spheres except for cohesin) to repress or stimulate transcription of protein-coding genes (thick blue lines). Starting from the top of the chromatin segment (blue line), an active enhancer – marked by flanking nucleosomes containing H3K27ac and H3K4me1 – comes

into contact with its target promoter by means of chromatin looping factors such as cohesin (blue ring). Transcriptional activators bind to the enhancer and interact with cofactors like CBP and p300, facilitating gene transcription by RNA Polymerase II. The active promoter is marked by H3K27ac and H3K4me3 and its gene body contains DNA methylation (black vertical lines with circles) and H3K4me1. Another enhancer is blocked by CTCF, which binds to an intervening insulator sequence, from acting on the promoter. A silencer sequence, marked by DNA methylation, provides a binding site for a transcriptional repressor that silences transcription of an adjacent gene whose promoter is marked by DNA methylation and H3K27me3. PRC2 helps add and maintain the repressive mark H3K27me3 at the silent promoter and its gene body. Abbreviations: CRE, *cis*-regulatory element; TF, transcription factor; CTCF, CCCTC-binding factor; RNA Pol II, RNA Polymerase II; PRC, Polycomb Repressive Complex; CBP, CREB-binding protein; H3K4me3, histone H3, lysine 4 trimethylation; H3K27ac, histone H3, lysine 27 acetylation; H3K4me1, histone H3, lysine 4 monomethylation; H3K27me3, histone H3, lysine 27 trimethylation.

Promoters

The core promoter is the minimal sequence around the TSS necessary to activate basal levels of transcription (Burke et al., 1998). The core promoter recruits general transcription factors, RNA Polymerase II, and the Mediator complex to form the pre-initation complex (Poss et al., 2013). It contains the TATA box, initiator, TFIIB and DNA recognition elements, downstream core element, and/or motif ten elements (Butler and Kadonga, 2002). The level of transcriptional output at a promoter is affected by the binding of TFs to promoter-proximal elements as well as to more distal CREs. Promoter-proximal elements are located up to a few hundred base pairs away from the TSS (Dikstein, 2011). Distal CREs may act on multiple promoters, particularly if they are alternate promoters of the same gene (Sanyal et al., 2012; Andrey et al., 2017).

Promoters are well annotated in mice and humans (FANTOM Consortium, 2014) because they are position- and orientation-dependent with respect to protein-coding genes and are characterized by certain motifs and chromatin modifications. They extend for up to one or two kilobase pairs around the transcription start site (TSS) (Florquin et al., 2005; Kiran et al., 2006).

In terms of their genomic features, promoters are enriched for CpG islands. About 70% of known vertebrate promoters, especially those of developmental regulator genes, contain CpG islands (Saxonov et al., 2006). CpG islands are 500-1500 base pair stretches of high C-G dinucleotide content whose methylation is associated with transcriptional repression (Andersson et al., 2014).

The nucleosomes flanking promoters and also enhancers are associated with certain types of histone modifications (**Table 1.1**). Through their interactions with

trans-acting factors, these modified histone tails affect chromatin stability and transcriptional output (Vignali et al., 2000). Histone H3, lysine 4 methylation follows enrichment gradients across the gene body starting from the promoter, with H3K4 trimethylation most highly enriched close to the promoter, and di- and monomethylation peaking farther away (Wang et al., 2011). The ratio of H3K4me1 to H3K4me3 is generally higher at enhancers than promoters (Robertson et al., 2008), but this ratio may be more linked to the activity level of a given CRE than to the class of CRE (Andersson et al., 2015). Active promoters, like active enhancers, are marked by H3K27ac (Zhang et al., 2015). Active and poised promoters, like strong enhancers, are associated with H3K4me3, likely due to its positive correlation with RNA Polymerase II (Andersson et al., 2015). In embryonic stem cells, the promoters of developmental genes tend to be enriched for both H3K4me3 and H3K27me3 (Voigt et al., 2013). Upon cellular differentiation, genes resolve these bivalent marks according to their activity level: active genes retain H3K4me3 and silent genes H3K27me3 (Bernstein et al., 2006). Silent promoters and their gene bodies are marked by trimethylation of histone H3, lysine 27 (H3K27me3), whose addition is catalyzed by the Polycomb Repressive Complex 2 (PRC2) and whose presence is associated with transcriptional repression (Cao and Zhang, 2004).

Table 1.1. Conventional associations of histone marks with CREs.

Modification	Principal associated chromatin features
H3K4me1	Active and poised enhancers
H3K4me3	Active, poised, and bivalent promoters
H3K27ac	Active promoters and enhancers
H3K27me3	Bivalent and silent promoters and gene bodies; repressive chromatin

Enhancers

Much work on CREs has focused on enhancers. Enhancers, like promoter-proximal elements, activate transcription, but often in a distance- and orientation-independent manner (Pennacchio et al., 2013). They are short sequences tens to hundreds of base pairs in length (Kulaeva et al., 2012) that recruit TFs and their co-activators, bringing RNA Polymerase II and transcriptional machinery to promoters to stimulate transcription. The pre-initiation complex which typically assembles at the core promoter may alternatively assemble at an enhancer, enabling more precise control over when transcription is initiated (Maston et al., 2006). At enhancers, TFs

outcompete nucleosomes for DNA occupancy (Long et al., 2016). Nucleosomes at enhancers acquire histone variants which cause them to become hypermobile and thus easier to displace by TFs than regular nucleosomes (Calo and Wysocka, 2013).

Nucleosomes flanking enhancers, like those flanking promoters, are often marked by combinations of histone tail modifications (**Table 1.1**) (Karlic et al., 2010). Two of the most common enhancer marks are H3K27ac and H3K4me1. H3K27ac neutralizes the positive charge of the lysine residual, making it easier for the negatively-charged DNA to dissociate from the nucleosome and be accessible to transcriptional machinery (Bao and Bedford, 2016). It is added – usually to regions already marked by H3K4me1 – by the histone acetyltransferase p300 and its paralog CBP (CREB-binding protein), which also function as TF co-activators. While H3K27ac is an accurate predictor of active enhancers (Holmqvist and Mannervik, 2013), it does not comprehensively mark all active enhancers, and it is also associated with active promoters (Wang et al., 2008; Creyghton et al., 2010). H3K4me1 aids enhancer activity by mediating interactions between histone tails, by protecting DNA from methylation and subsequent gene silencing, and by recruiting cofactors to aid TF recruitment and chromatin remodeling (Calo and Wysocka, 2013). H3K4me1 is nonspecific to CREs, often extending across the gene body. It is associated with poised and active enhancers. In mouse embryonic stem cells (mESCs), poised enhancers are marked by trimethylation of histone H3, lysine 27 (H3K27me3), or acquire it upon differentiation (Zentner et al., 2011). Enhancers can also be found within coding regions and therefore may also be characterized by histone modifications conventionally associated with genes (Pennacchio et al., 2013). Finally, strong enhancers may be marked by H3K4me3, which is strongly positively correlated with RNA Polymerase II and therefore classically considered to be a promoter mark (Andersson et al., 2015). Looking for combinations of histone marks known to be associated with enhancers can help determine their location and activity level (Pennacchio et al., 2013).

It has long been known that enhancers not only stimulate transcription of genes, but can also themselves be transcribed. Recently, this was shown to be a universal characteristic of active enhancers (Pennacchio et al., 2013), which stimulate their own transcription into short, bidirectional enhancer RNAs (“eRNAs”) (De Santa et al., 2010; Kim et al., 2010). These eRNAs may help enhance the chromatin loops between promoters and enhancers, and have also been observed to

interact with histone-modifying enzymes (Ding et al., 2018). eRNA transcript abundance can be used to predict active enhancers (Hah et al., 2013; Andersson et al., 2014). In an endogenous context, eRNA transcription can be captured in real-time through GRO-seq (Garcia-Martinez et al., 2004; Core et al., 2008). However, because this method requires the incorporation of bromouridine, which is cytotoxic (Paulsen et al., 2014), it may be more suited for experiments with cultured cells. Therefore, methods like ChIP-seq and ATAC-seq may be more appropriate for enhancer identification in cells isolated from tissues.

Enhancers, along with promoters and nucleosome binding sites, are enriched in mammals for GC-rich sequences (Fenouil et al., 2012; Wang et al., 2012; Colbran et al., 2017). TFs exhibit cell type-specific preferences for the GC content of the regions flanking their binding sites (Dror et al., 2015). The GC content of the flanking regions helps determine the thermostability and flexibility of the DNA, thus affecting the formation of chromatin looping structures that enable distal enhancers to contact their target promoters (Vinogradov, 2003). The GC content of enhancers appears to be correlated with their activity, although studies have found both positive (White et al., 2013; Kwasnieski et al., 2014) and negative (Shen et al., 2016) correlation.

Further sequence features of enhancers include higher levels of DNA methylation than those found at promoters (Sharifi-Zarchi et al., 2017), which often inhibits but sometimes enhances their activity (Chamberlain et al., 2014; Lea et al., 2018), and the prominence of dinucleotide repeats (Yanez-Cuna et al., 2014). Enhancers tend to evolve rapidly in mammals (Villar et al., 2015). While TFs and their binding motifs are well-conserved, enhancer activity varies considerably across species (Fish et al., 2017) because it depends on many factors, like TF binding and accessibility, histone modification marks, and other features of the chromatin and nuclear environment. TF binding is a better predictor of *in vivo* enhancer activity than chromatin accessibility or histone modification marks (Dogan et al., 2015; Kreimer et al., 2017), but the hundreds of different TFs in mammals tend to be highly cell type-specific, making comprehensive TF binding data impractical to collect (Khamis et al., 2018). TF binding at enhancers can be sub-optimal, which could in itself constitute a mechanism whereby enhancer activity domains become restricted during development (Farley et al., 2015). It is governed by a complex motif grammar, including spacing and orientation of the motifs on the enhancer (Spitz and Furlong, 2012; Long et al., 2016).

Histone modification marks, chromatin accessibility, DNA methylation, eRNA transcription, and other datasets are now available in a number of cell types in the mouse and human genomes (ENCODE Project Consortium, 2012), but which of these features actually confer *cis*-regulatory function and thus impact gene expression (Graur et al., 2013)? The determinants of enhancer activity are explored in greater detail in the General Discussion (**Chapter 4**).

Silencers

Silencers repress transcription by recruiting transcriptional repressors or by competing with enhancers for TF binding (Ogbourne and Antalis, 1998). Silencers have higher rates of DNA methylation than enhancers and, like enhancers, act in an orientation- and position-independent manner with respect to their target promoters (Jayavelu et al., 2018). Current genomic methods, especially 3C-derived approachs, likely confound enhancers and silencers. However, due to the difficulty in assaying the *absence* of a reporter gene compared to its activity, experimental demonstrations for silencers are scarce compared to enhancers.

Insulators

Insulators recruit the architectural and chromatin structure protein CTCF (CCCTC-binding factor) to prevent CREs from interacting with one another (Kim et al., 2015). Insulators range from 500 base pairs to around 3 kilobase pairs and are position-dependent. They shield promoters from enhancers by interfering with chromatin loop formation, and from repressive heterochromatin by blocking chromatin remodelers (Maston et al., 2006; Tokuda et al., 2011). Insulators in vertebrates are often found at imprinted loci such as the well-characterized *H19/Igf2* imprinting control region (Maston et al., 2006). There have been many comprehensive studies and reviews detailing how insulators organize chromosome domains and modulate transcription through CTCF binding (Phillips and Corces, 2009; Lupianez et al., 2015). For the purpose of this thesis, since CTCF ChIP binding data is not widely available for different cell types and much of its effect is captured within defined TADs, the role and effects of insulators have been factored into the datasets I use and I will therefore not discuss insulators separately.

Dynamic switching between classes of CREs

There is dynamic switching between distinct classes of CREs. Promoters can take on enhancer-like roles and regulate transcription of other promoters (Kowalczyk et al., 2012; Leung et al., 2015). Conversely, enhancers can take on promoter-like roles (Andersson et al., 2015). Silencers, insulators, and enhancers, like promoters and enhancers, can act interchangeably across cell types (Kolovos et al., 2012; Andersson et al., 2015). These dynamics enable precise control of gene expression but lend complexity to the genome, compounding efforts to map CREs to their target genes.

Methods to identify CREs

Since CREs tend to have highly variable sequences, often act distally on target genes, and can lie anywhere in the genome, even evolving from "junk" DNA like transposable elements (Makalowski, 2003), how can they be accurately linked to their target genes? A plethora of methods exists to aid identification and mapping of CREs to their target genes, of which a subset is listed in **Table 1.2**. These methods capture known chromatin features of CREs that differ depending on CRE function. Methods like ATAC-seq, DNase-seq, MNase-seq, or FAIRE-seq assays reveal regions of open and closed chromatin (Tsompana and Buck, 2014) and can be used to infer DNase hypersensitive sites (DHSs). DHSs have historically been used to identify all types of CREs because they often coincide with TF binding (Thurman et al., 2012; Chen et al., 2018). Additional methods that distinguish between the distinct classes of CREs – promoters, enhancers, silencers, and insulators – facilitate their identification.

Table 1.2. Methods to identify and characterize *cis*-regulatory elements. The first five methods in the table capture genome-wide data such as transcript abundance, chromatin accessibility, and histone modification marks, but do not reveal spatiotemporal domains of CREs. FISH and 3C (grey shaded boxes) explore the three-dimensionality of the genome. The last four methods enable functional validation of CREs, revealing their expression domains.

Method and Aim	Advantages	Limitations	References
ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin): Map genome-wide chromatin accessibility, inferring	Quick, simple protocol. High signal-to-noise ratio. As few as 5,000 cells required; single-cell also possible. Improvement over DNase-seq, MNase (micrococcal nuclease)-seq, and FAIRE (Formaldehyde-assisted isolation of regulatory	It is possible for closed chromatin to open during sample preparation and integrate the transposome, leading to spurious signal. Biased towards shorter fragments.	ATAC-seq: Buenrostro et al., 2015 DNase-seq: Galas and Schmitz, 1978 MNase-seq: Valouev et al., 2011 FAIRE-seq: Giresi et al., 2007

nucleosome and transcription factor occupancy.	elements)-seq.		
ChIP-seq (**Ch**romatin **i**mmuno-**p**recipitation): Capture genome-wide TF binding and nucleosome occupancy.	Binding of the p300 TF co-activator is the most accurate *in-vivo* predictor of active enhancers. Histone modification enrichment is strongly associated with active and poised enhancers and promoters. Up to single base pair resolution. Quick, high efficiency and accuracy.	Can be highly cell type-specific. Sensitive to antibody quality; biased toward high GC-content sequences. Requires high sequencing coverage. Histone modifications are not comprehensive and may miss CREs, and may also be nonspecific.	Initial development of ChIP-seq: Robertson et al., 2007 Histone profiling: Mikkelsen et al., 2007
eRNA-seq (Enhancer RNA sequencing): Determine *in-vivo* enhancer activity from the abundance of short, bi-directional enhancer RNAs, or eRNAs.	Can be gleaned from RNA-seq data if coverage is of sufficient depth; thus optimal for experiments with limited starting material (no separate experiment needed).	Requires high sequence coverage and/or enrichment techniques due to the low transcription rate and instability of eRNAs compared to mRNAs.	Kim et al., 2010; Andersson et al., 2014
GRO-seq (**G**lobal nuclear **r**un-**o**n **seq**uencing): Obtain a real-time readout of transcription.	Unbiased and highly sensitive; can detect transcripts of low abundance, including noncoding RNAs like eRNAs.	Transcription is "frozen" at the desired stage. High amount of material needed (10 million cells).	Garcia-Martinez et al., 2004; Core et al., 2008
Isochore assignment: Distinguish active from inactive chromatin by looking at GC content and comparing with chromatin compartments "A" and "B".	Already characterized in mouse and humans: is an inherent sequence feature and not cell type-specific.	Low resolution (300 kbp average length in the mouse).	Thiery et al., 1976
3C (**C**hromosome **c**onformation **c**apture) and derivatives: Capture the contacts occurring between loci in the genome.	Definitively links regulatory elements to their targets.	Prone to noise. Long, complex protocol.	Dekker et al., 2002
FISH (**F**luorescence *in* ***s****itu* **h**ybridization): Confirm the interaction between loci by hybridizing fluorescently labeled probes.	Definitively links regulatory elements to their targets.	Low resolution. Technique is confirmational rather than experimental (probe sequence must be known ahead of experiment)	Langer-Safer et al., 1982
RNA-seq: Obtain quantitative expression levels of coding and noncoding RNAs.	High throughput. Capable of detecting low-abundance transcripts like eRNAs.	Transcript size selection can lead to bias or inconsistency across experiments.	Morin et al., 2008
MPRA (**M**assively **P**arallel **R**eporter **A**ssay):	High-throughput; avoids positional (random integration) effects because tested enhancers are not integrated	Episomal features that affect enhancer activity, may differ from genomic (endogenous) features.	Melnikov et al., 2012 STARR-seq: Arnold et al., 2013

Obtain a readout of enhancer activity from co-transfection of a library of enhancer reporter constructs.	into the genome.		
Transgenic reporter assay: Visualize CRE expression domains by site-specific or random integration of enhancer reporter constructs or of minimal promoter-reporter genes.	Reveals expression domains of all CREs at the site of integration (enhancer traps), of individual enhancers (enhancer-reporter constructs), or of entire regions (through BAC reporters).	May be subject to site-specific integration effects unless integration is targeted. Construct injection is tedious and expensive.	Banerji et al., 1981
ISH (***In s**itu* **h**ybridization): Obtain expression pattern of a DNA or RNA molecule by visualizing hybridization to a labeled probe.	Material (embryo or tissue section) can be re-used repeatedly for multiple ISH; method does not use up the sample.	Time-consuming and tedious. Not sensitive at detecting low-copy molecules. Probe permeation and subsequent staining visualization of internal tissues is limited. Hybridization, post-hybridization, and proteinase K digestion conditions must be carefully titrated.	Gall and Pardue, 1969

1.3 Linking CREs to target promoters

The main methodology to directly link CREs to promoters is through the chromosome conformation capture (3C) assay and its derivatives. These methods use cross-linking, restriction enzyme cutting and re-ligation to preserve three-dimensional information such as looping contacts that would otherwise be lost during DNA extraction.

1.3.1 3C methodology

The three-dimensionality of the genome necessitates the integration of several types of data when attempting to characterize CREs. The methods in **Table 1.2** can be divided into those that rely on genome-wide data without distinguishing which promoter is acted upon, and those that do establish the physical link between distal-acting CREs and their target promoters (FISH, 3C). Fluorescence *in situ* hybridization (FISH) can be used to visualize and confirm the interaction between two or more loci, but only has single-gene (100-200 kbp) resolution at best, and fluorescent probes must be designed ahead of time to hybridize to complementary regions in the genome (Cui et al., 2016). In contrast, chromosome conformation capture (3C) methods, which are molecular as opposed to microscopic techniques,

afford sub-kilobase pair resolution and are high throughput, with no foreknowledge of interacting regions required (Barutcu et al., 2016).

To prepare a 3C library, nuclei are cross-linked to preserve their native conformation (**Figure 1.2**). Cross-linked nuclei undergo shearing or restriction digestion, followed by re-ligation, de-cross-linking, and sequencing. Captured sequences are mapped to the reference genome to identify the sequences interacting with one another (de Wit and de Laat, 2012). Selective sequencing of 3C library molecules that contain a “viewpoint”, or target sequence, yields a chromosome conformation capture profile for that viewpoint.

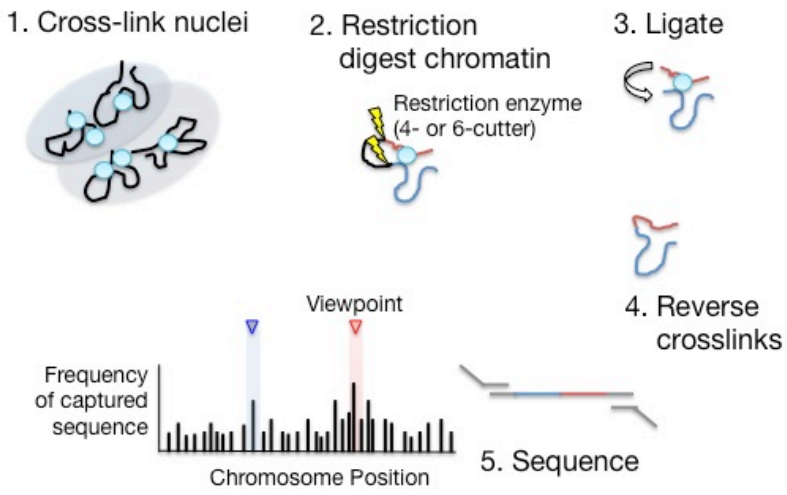


Figure 1.2. Chromosome conformation capture (3C) captures interacting genomic sequences. During 3C template preparation, the chromatin is cross-linked to its associated proteins (light blue ovals), digested into shorter fragments to increase resolution, and re-ligated to reconstitute the chromatin loops that occurred in its native conformation. Ligation of physically proximate (but linearly distant) sequences to one another results in chimeric molecules. Sequencing of the molecules which contain the “viewpoint”, or target sequence, followed by alignment of the interacting regions to the reference genome, yields a chromosome conformation capture-seq profile for that viewpoint (red arrowhead, viewpoint; blue arrowhead, genomic location of the blue interacting fragment depicted in Steps 2-5), where the height of the capture signal (y-axis) indicates how often that fragment was captured, or ligated to the viewpoint-containing fragment and sequenced.

Each step of the 3C library preparation must be carefully titrated for different cell types in order to minimize introducing or capturing spurious interactions (van de Werken et al., 2012). When properly optimized, a chromosome conformation capture signal profile typically shows exponetntial signal decay out from the viewpoint (**Figure 1.2**; **Figure 1.3B**) (van de Werken et al., 2012). This is due to the proximity ligation effect: molecular kinetics dictates that during ligation, DNA fragments that were closest to the viewpoint are more likely to become re-ligated to it due to their physical proximity on the linear (pre-digested) DNA. Ligation of chimeric molecules is further affected by cohesiveness of digested fragments (Gavrilov et al., 2013). Ultimately, capture data is semi-quantitative because biological interactions may be

captured or amplified less preferentially due to proximity ligation, PCR amplification bias, and prevalence of spurious interactions. In addition to preparation of multiple replicates, complementary chromatin feature data can help rule out noise (Denker and de Laat, 2016).

A

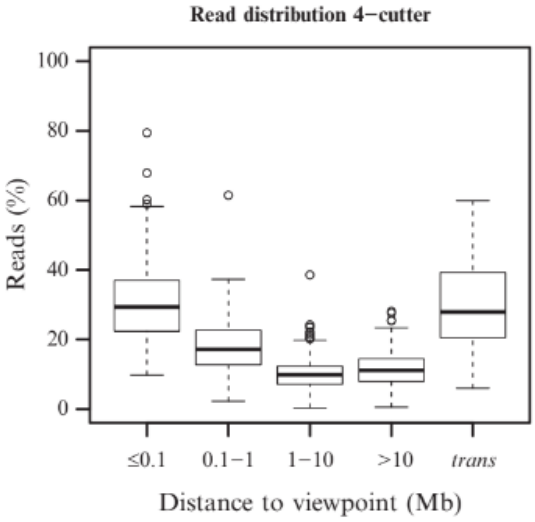
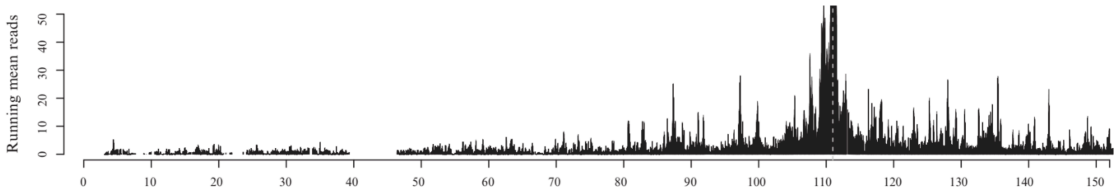


Figure 1.3. Chromosome conformation capture signal decays over distance. (**A**) The percentage of viewpoint contacts mapping in *cis*, or to the same chromosome as the viewpoint (leftmost four boxplots), increases with decreasing linear distance from the viewpoint. The median percentage of reads mapping to fragment within 100 kbp of the viewpoint (leftmost boxplot) is comparable to the median of *trans* – mapping to all other chromosomes – reads (rightmost boxplot). (**B**) A typical conformation capture profile shows that the signal emanates from the viewpoint (dashed grey line), with the highest frequency of contact closest to the viewpoint. Peak height represents the running mean reads (y-axis), and relative position (kbp) along the viewpoint chromosome is depicted on the x-axis. **Figures are Figure 4.6 and Figure 4.7B from van de Werken et al., 2012.**

B



The many derivatives of 3C-seq differ in their hybridization, digestion, and ligation conditions and subsequently, their throughput. 4C-seq (circularized chromosome conformation capture) captures all contacts for a single viewpoint, whereas Hi-C captures all genome-wide interactions. ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) combines ChIP-seq with 3C to find interacting loci bound by a given TF. 5C (chromosome conformation capture carbon copy) captures interactions for many viewpoints at once, as does Capture-C (Denker and de Laat, 2016).

1.3.2 TADs and the 3D structure of chromatin

Hi-C, which captures genome-wide interactions, revealed the presence in the genome of topologically associating domains (TADs), which are megabase-scale

regions of interacting loci (Dixon et al., 2012). Interactions tend to occur far more frequently within TADs than across TAD boundaries. This is due to the insulating action of boundary elements like CTCF, which are found at the edges of TADs. CTCF stalls or halts cohesin, preventing chromatin loops from forming across TAD boundaries (Fudenberg et al., 2016). TADs are important structural units that have been found in mESCs to contain a majority of promoter-enhancer interactions (Shen et al., 2012; Schoenfelder et al., 2015). At the *HoxD* cluster, which may be a rare example of a developmentally critical locus straddling two different TADs, different *HoxD* genes become active at specific times through their interaction with either the centromeric or the telomeric TAD flanking the cluster, allowing precise coordination of gene expression in the proximal – during early stages of limb development – and distal – during later stages – limb bud (Andrey et al., 2013). At the *Wnt6* locus, deletion of CTCF-associated boundary elements allowed CREs to act on promoters outside of their TADs, resulting in ectopic gene expression (Lupianez et al., 2015). CTCF binding site orientation is nonrandom at TAD edges and can determine enhancer directionality: over 90% of TAD boundary elements in mouse brain cells were found to have oppositely orientated (reverse-forward) CTCF binding sites relative to those of their neighboring TADs, and CRISPR/Cas9-mediated inversion of a CTCF-containing enhancer between two TADs caused the enhancer to switch the TAD it contacted (Guo et al., 2015).

TADs aid identification and mapping of CREs to their target promoters. Even though enhancers tend to act from tens of hundreds of kilobase pairs away, they tend to stay within TAD boundaries. TAD boundaries were reported to be relatively conserved across cell types and even species (Dixon et al., 2012), although exact boundaries likely vary from cell to cell (Giorgetti et al., 2014; Liu and Tijan, 2018). If TAD boundaries are available for a cell type, they can be used as a proxy for the cell type of interest and used to narrow candidate genes. Conversely, if the experimental goal is to find the CREs regulating a promoter of interest, one can look within the TAD for known CRE marks using other datatsets like chromatin accessibility or histone profiling (Andrey and Mundlos, 2017). To help confirm or functionally test CRE-promoter interactions, expression profiles of candidate genes in the TAD – for instance, from *in situ* hybridizations – can be compared with expression domains from transgenic or massively parallel reporter assays (Smith et al., 2019; Frazer et al., 2004).

On a larger scale, TADs can be organized into different chromatin compartments, designated “A” and “B”, that are characterized by their respective gene expression levels (Rao et al., 2014). Compartment A is characterized by active genes and open chromatin that localizes close to the center of the nucleus. Compartment B chromatin is closed, typically has lower GC content than Compartment A chromatin, includes silent genes and repressive marks, and localizes at the nuclear periphery (Lieberman-Aiden et al., 2009). Not only local sequence GC content but also broad-scale GC content can help determine which regions of the genome likely harbor CREs. One such type of feature has been termed isochores. Isochores are long (on average, 300 kilobase pairs in mammals) stretches of homogenous GC content (Bernardi, 2000); isochore families with moderate levels of GC content are most likely to contain protein-coding genes and hence CREs (Arhondakis et al., 2011).

In summary, the three-dimensionality of the genome allows CREs to act distally, even skipping over closer promoters or being blocked by insulators lying between them and their targets on the linear DNA. Chromosome conformation capture reveals or confirms distal interactions, but must be followed by functional validation tests or compared with other datasets.

1.4 Mouse limb development as a model system to study transcriptional regulation

The developing mouse limb bud is an ideal organ to study transcriptional regulation. Limb bud development is regulated by many transcription factor families. Recently, Andrey and coworkers obtained Capure-C data (which I analyzed in **Chapter 2**) from over four hundred limb development genes, including members of the important *Hox*, *Sox*, *Tbx*, and *Runx* transcription factor families and the IGF, FGF, WNT, and RA signaling pathways (Andrey et al., 2017). The mouse limb bud is a highly heterogeneous structure (Zeller et al., 2009). The anterior-posterior, proximal-distal, and dorsal-ventral axes form by E12.5 (Zuniga, 2015), and transcriptional dynamics differ across micro-sections (Rodriguez-Carballo et al., 2017). For instance, at the *HoxD* cluster, various *HoxD* genes are differentially transcriptionally regulated through interactions with one of two adjacent TADs depending on the stage and location of the cells (Fabre et al., 2017). The mammalian limb bud therefore represents an intriguing system in which variations in

long-range DNA interactions and transcription regulation may give rise to visible, morphological differences.

It is therefore our motive to study a biological system in which limb phenotypes show striking yet specific changes. One such system is provided in the mouse from an artificial selection experiment for long tibia length known as the "Longshanks selection experiment". This is a mouse population created by Dr. Campbell Rolian at the University of Calgary, in which outbred mice with variable tibia length were subjected to 20 generations of selective breeding. I was involved in a genomic study aimed at determining which loci may be responsible for the rapid increase in tibia length.

In the Longshanks selection experiment (**Chapter 3**), analysis of the genomic sequencing data revealed many loci genome-wide that were found to have shifted allele frequencies between the lines under selection and the control lines, providing ample candidate loci at which to investigate transcriptional dynamics. Furthermore, regions with significant allele frequency shifts were enriched for genes known to have limb knockout phenotypes. Understanding how gene expression is regulated in the limb bud has important implications for researching evolution, disease, and development.

1.5 References to Chapter 1

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jorgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Muller, A. R. R. Forrest, P. Carninci, M. Rehli and A. Sandelin (2014). "An atlas of active enhancers across human cell types and tissues." *Nature* **507**(7493): 455-461.

Andersson, R., A. Sandelin and C. G. Danko (2015). "A unified architecture of transcriptional regulatory elements." *Trends Genet* **31**(8): 426-433.

Andrey, G., T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz and D. Duboule (2013). "A switch between topological domains underlies HoxD genes collinearity in mouse limbs." *Science* **340**(6137): 1234167.

Andrey, G. and S. Mundlos (2017). "The three-dimensional genome: regulating gene expression during pluripotency and development." *Development* **144**(20): 3646-3658.

Andrey, G., R. Schopflin, I. Jerkovic, V. Heinrich, D. M. Ibrahim, C. Paliou, M. Hochradel, B. Timmermann, S. Haas, M. Vingron and S. Mundlos (2017). "Characterization of hundreds of

regulatory landscapes in developing limbs reveals two regimes of chromatin folding." *Genome Res* **27**(2): 223-233.

Arhondakis, S., K. Frousios, C. S. Iliopoulos, S. P. Pissis, G. Tischler and S. Kossida (2011). "Transcriptome map of mouse isochores." *BMC Genomics* **12**: 511.

Arnold, C. D., D. Gerlach, C. Stelzer, L. M. Boryn, M. Rath and A. Stark (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq." *Science* **339**(6123): 1074-1077.

Banerji, J., S. Rusconi and W. Schaffner (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." *Cell* **27**(2 Pt 1): 299-308.

Bao, J. and M. T. Bedford (2016). "Epigenetic regulation of the histone-to-protamine transition during spermiogenesis." *Reproduction* **151**(5): R55-70.

Barutcu, A. R., A. J. Fritz, S. K. Zaidi, A. J. van Wijnen, J. B. Lian, J. L. Stein, J. A. Nickerson, A. N. Imbalzano and G. S. Stein (2016). "C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization." *J Cell Physiol* **231**(1): 31-35.

Bernardi, G. (2000). "Isochores and the evolutionary genomics of vertebrates." *Gene* **241**(1): 3-17.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber and E. S. Lander (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." *Cell* **125**(2): 315-326.

Buenrostro, J. D., B. Wu, H. Y. Chang and W. J. Greenleaf (2015). "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Curr Protoc Mol Biol* **109**: 21 29 21-29.

Burke, T. W., P. J. Willy, A. K. Kutach, J. E. Butler and J. T. Kadonaga (1998). "The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box." *Cold Spring Harb Symp Quant Biol* **63**: 75-82.

Butler, J. E. and J. T. Kadonaga (2002). "The RNA polymerase II core promoter: a key component in the regulation of gene expression." *Genes Dev* **16**(20): 2583-2592.

Calo, E. and J. Wysocka (2013). "Modification of enhancer chromatin: what, how, and why?" *Mol Cell* **49**(5): 825-837.

Cao, R. and Y. Zhang (2004). "The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3." *Curr Opin Genet Dev* **14**(2): 155-164.

Carroll, S. B. (2008). "Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution." *Cell* **134**(1): 25-36.

Chamberlain, A. A., M. Lin, R. L. Lister, A. A. Maslov, Y. Wang, M. Suzuki, B. Wu, J. M. Greally, D. Zheng and B. Zhou (2014). "DNA methylation is developmentally regulated for genes essential for cardiogenesis." *J Am Heart Assoc* **3**(3): e000976.

Chen, A., D. Chen and Y. Chen (2018). "Advances of DNase-seq for mapping active gene regulatory elements across the genome in animals." *Gene* **667**: 83-94.

Colbran, L. L., L. Chen and J. A. Capra (2017). "Short DNA sequence patterns accurately identify broadly active human enhancers." *BMC Genomics* **18**(1): 536.

Core, L. J., J. J. Waterfall and J. T. Lis (2008). "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." *Science* **322**(5909): 1845-1848.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young and R. Jaenisch (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state." *Proc Natl Acad Sci U S A* **107**(50): 21931-21936.

Cui, C., W. Shu and P. Li (2016). "Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications." *Front Cell Dev Biol* **4**: 89.

De Santa, F., I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B. K. Tusi, H. Muller, J. Ragoussis, C. L. Wei and G. Natoli (2010). "A large fraction of extragenic RNA pol II transcription sites overlap enhancers." *PLoS Biol* **8**(5): e1000384.

de Wit, E. and W. de Laat (2012). "A decade of 3C technologies: insights into nuclear organization." *Genes Dev* **26**(1): 11-24.

Dekker, J., K. Rippe, M. Dekker and N. Kleckner (2002). "Capturing chromosome conformation." *Science* **295**(5558): 1306-1311.

Denker, A. and W. de Laat (2016). "The second decade of 3C technologies: detailed insights into nuclear organization." *Genes Dev* 30(12): 1357-1382.

Dikstein, R. (2011). "The unexpected traits associated with core promoter elements." *Transcription* **2**(5): 201-206.

Ding, M., Y. Liu, X. Liao, H. Zhan, Y. Liu and W. Huang (2018). "Enhancer RNAs (eRNAs): New Insights into Gene Transcription and Disease Treatment." *J Cancer* **9**(13): 2334-2340.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* **485**(7398): 376-380.

Dogan, N., W. Wu, C. S. Morrissey, K. B. Chen, A. Stonestrom, M. Long, C. A. Keller, Y. Cheng, D. Jain, A. Visel, L. A. Pennacchio, M. J. Weiss, G. A. Blobel and R. C. Hardison (2015). "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility." *Epigenetics Chromatin* **8**: 16.

Dror, I., T. Golan, C. Levy, R. Rohs and Y. Mandel-Gutfreund (2015). "A widespread role of the motif environment in transcription factor binding across diverse protein families." *Genome Res* **25**(9): 1268-1280.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome." *Nature* **489**(7414): 57-74.

FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). "A promoter-level mammalian expression atlas." *Nature* **507**(7493): 462-470.

Fabre, P. J., M. Leleu, B. H. Mormann, L. Lopez-Delisle, D. Noordermeer, L. Beccari and D. Duboule (2017). "Large scale genomic reorganization of topological domains at the HoxD locus." *Genome* Biol **18**(1): 149.

Fakhouri, W. D., A. Ay, R. Sayal, J. Dresch, E. Dayringer and D. N. Arnosti (2010). "Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo." *Mol Syst Biol* **6**: 341.

Farley, E. K., K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar and M. S. Levine (2015). "Suboptimization of developmental enhancers." *Science* **350**(6258): 325-328.

Fenouil, R., P. Cauchy, F. Koch, N. Descostes, J. Z. Cabeza, C. Innocenti, P. Ferrier, S. Spicuglia, M. Gut, I. Gut and J. C. Andrau (2012). "CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters." *Genome Res* **22**(12): 2399-2408.

Fish, A., L. Chen and J. A. Capra (2017). "Gene Regulatory Enhancers with Evolutionarily Conserved Activity Are More Pleiotropic than Those with Species-Specific Activity." *Genome Biol Evol* **9**(10): 2615-2625.

Florquin, K., Y. Saeys, S. Degroeve, P. Rouze and Y. Van de Peer (2005). "Large-scale structural analysis of the core promoter in mammalian and plant genomes." *Nucleic Acids Res* **33**(13): 4255-4264.

Frazer, K. A., L. Pachter, A. Poliakov, E. M. Rubin and I. Dubchak (2004). "VISTA: computational tools for comparative genomics." *Nucleic Acids Res* **32**(Web Server issue): W273-279.

Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur and L. A. Mirny (2016). "Formation of Chromosomal Domains by Loop Extrusion." *Cell Rep* **15**(9): 2038-2049.

Galas, D. J. and A. Schmitz (1978). "DNAse footprinting: a simple method for the detection of protein-DNA binding specificity." *Nucleic Acids Res* **5**(9): 3157-3170.

Gall, J. G. and M. L. Pardue (1969). "Formation and detection of RNA-DNA hybrid molecules in cytological preparations." *Proc Natl Acad Sci U S A* **63**(2): 378-383.

Garcia-Martinez, J., A. Aranda and J. E. Perez-Ortin (2004). "Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms." *Mol Cell* **15**(2): 303-313.

Gavrilov, A. A., A. K. Golov and S. V. Razin (2013). "Actual ligation frequencies in the chromosome conformation capture procedure." *PLoS One* **8**(3): e60403.

Giorgetti, L., R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana and E. Heard (2014). "Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription." *Cell* **157**(4): 950-963.

Giresi, P. G., J. Kim, R. M. McDaniell, V. R. Iyer and J. D. Lieb (2007). "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin." *Genome Res* **17**(6): 877-885.

Graur, D., Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall and E. Elhaik (2013). "On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE." *Genome Biol Evol* **5**(3): 578-590.

Guo, Y., Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis and Q. Wu (2015). "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function." *Cell* **162**(4): 900-910.

Hah, N., S. Murakami, A. Nagari, C. G. Danko and W. L. Kraus (2013). "Enhancer transcripts mark active estrogen receptor binding sites." *Genome Res* **23**(8): 1210-1223.

Holmqvist, P. H. and M. Mannervik (2013). "Genomic occupancy of the transcriptional co-activators p300 and CBP." *Transcription* **4**(1): 18-23.

Jayavelu, N. D., Jajodia, A., Mishra, A., and R. D. Hawkins. An atlas of silencer elements for the human and mouse genomes. BioRxiv 252304 [**Preprint**]. January 23, 2018 [cited 2019 Apr 25]. Available from: http://**doi**.org/10.1101/252304.

Karlic, R., H. R. Chung, J. Lasserre, K. Vlahovicek and M. Vingron (2010). "Histone modification levels are predictive for gene expression." *Proc Natl Acad Sci U S A* **107**(7): 2926-2931.

Khamis, A. M., O. Motwalli, R. Oliva, B. R. Jankovic, Y. A. Medvedeva, H. Ashoor, M. Essack, X. Gao and V. B. Bajic (2018). "A novel method for improved accuracy of transcription factor binding site prediction." *Nucleic Acids Res* **46**(12): e72.

Kim, S., N. K. Yu and B. K. Kaang (2015). "CTCF as a multifunctional protein in genome regulation and gene expression." *Exp Mol Med* **47**: e166.

Kim, T. K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman and M. E. Greenberg (2010). "Widespread transcription at neuronal activity-regulated enhancers." *Nature* **465**(7295): 182-187.

Kiran, K., S. A. Ansari, R. Srivastava, N. Lodhi, C. P. Chaturvedi, S. V. Sawant and R. Tuli (2006). "The TATA-box sequence in the basal promoter contributes to determining light-dependent gene expression in plants." *Plant Physiol* **142**(1): 364-376.

Kolovos, P., T. A. Knoch, F. G. Grosveld, P. R. Cook and A. Papantonis (2012). "Enhancers and silencers: an integrated and simple model for their function." *Epigenetics Chromatin* **5**(1): 1.

Kowalczyk, M. S., J. R. Hughes, D. Garrick, M. D. Lynch, J. A. Sharpe, J. A. Sloane-Stanley, S. J. McGowan, M. De Gobbi, M. Hosseini, D. Vernimmen, J. M. Brown, N. E. Gray, L. Collavin, R. J. Gibbons, J. Flint, S. Taylor, V. J. Buckle, T. A. Milne, W. G. Wood and D. R. Higgs (2012). "Intragenic enhancers act as alternative promoters." *Mol Cell* **45**(4): 447-458.

Kreimer, A., H. Zeng, M. D. Edwards, Y. Guo, K. Tian, S. Shin, R. Welch, M. Wainberg, R. Mohan, N. A. Sinnott-Armstrong, Y. Li, G. Eraslan, T. B. Amin, R. Tewhey, P. C. Sabeti, J. Goke, N. S. Mueller, M. Kellis, A. Kundaje, M. A. Beer, S. Keles, D. K. Gifford and N. Yosef (2017). "Predicting gene expression in massively parallel reporter assays: A comparative study." *Hum Mutat* **38**(9): 1240-1250.

Kulaeva, O. I., E. V. Nizovtseva, Y. S. Polikanov, S. V. Ulianov and V. M. Studitsky (2012). "Distant activation of transcription: mechanisms of enhancer action." *Mol Cell Biol* **32**(24): 4892-4897.

Kwasnieski, J. C., C. Fiore, H. G. Chaudhari and B. A. Cohen (2014). "High-throughput functional testing of ENCODE segmentation predictions." *Genome Res* **24**(10): 1595-1602.

Langer-Safer, P. R., M. Levine and D. C. Ward (1982). "Immunological method for mapping genes on Drosophila polytene chromosomes." *Proc Natl Acad Sci U S A* **79**(14): 4381-4385.

Lea, A. J., C. M. Vockley, R. A. Johnston, C. A. Del Carpio, L. B. Barreiro, T. E. Reddy and J. Tung (2018). "Genome-wide quantification of the effects of DNA methylation on human gene regulation." *Elife* **7**: e38555.

Leung, D., I. Jung, N. Rajagopal, A. Schmitt, S. Selvaraj, A. Y. Lee, C. A. Yen, S. Lin, Y. Lin, Y. Qiu, W. Xie, F. Yue, M. Hariharan, P. Ray, S. Kuan, L. Edsall, H. Yang, N. C. Chi, M. Q. Zhang, J. R. Ecker and B. Ren (2015). "Integrative analysis of haplotype-resolved epigenomes across human tissues." *Nature* **518**(7539): 350-354.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* **326**(5950): 289-293.

Liu, Z. and R. Tjian (2018). "Visualizing transcription factor dynamics in living cells." *J Cell Biol* **217**(4): 1181-1191.

Long, H. K., S. L. Prescott and J. Wysocka (2016). "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution." *Cell* **167**(5): 1170-1187.

Lupianez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel and S. Mundlos (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions." *Cell* **161**(5): 1012-1025.

Makalowski, W. (2003). "Genomics. Not junk after all." *Science* **300**(5623): 1246-1247.

Maston, G. A., S. K. Evans and M. R. Green (2006). "Transcriptional regulatory elements in the human genome." *Annu Rev Genomics Hum Genet* **7**: 29-59.

Meiklejohn, C. D., J. D. Coolon, D. L. Hartl and P. J. Wittkopp (2014). "The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression." *Genome Res* **24**(1): 84-95.

Melnikov, A., A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan, Jr., J. B. Kinney, M. Kellis, E. S. Lander and T. S. Mikkelsen (2012). "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay." *Nat Biotechnol* **30**(3): 271-277.

Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander and B. E. Bernstein (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." *Nature* **448**(7153): 553-560.

Morin, R., M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones and M. Marra (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." *Biotechniques* **45**(1): 81-94.

Mouse Genome Sequencing Consortium (2002). "Initial sequencing and comparative analysis of the mouse genome." *Nature* **420**(6915): 520-562.

Ogbourne, S. and T. M. Antalis (1998). "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes." *Biochem J* **331** (Pt 1): 1-14.

Paulsen, M. T., A. Veloso, J. Prasad, K. Bedi, E. A. Ljungman, B. Magnuson, T. E. Wilson and M. Ljungman (2014). "Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA." *Methods* **67**(1): 45-54.

Pennacchio, L. A., W. Bickmore, A. Dean, M. A. Nobrega and G. Bejerano (2013). "Enhancers: five essential questions." *Nat Rev Genet* **14**(4): 288-295.

Phillips, J. E. and V. G. Corces (2009). "CTCF: master weaver of the genome." *Cell* **137**(7): 1194-1211.

Poss, Z. C., C. C. Ebmeier and D. J. Taatjes (2013). "The Mediator complex and transcription regulation." *Crit Rev Biochem Mol Biol* **48**(6): 575-608.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander and E. L. Aiden (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell* **159**(7): 1665-1680.

Robertson, A. G., M. Bilenky, A. Tam, Y. Zhao, T. Zeng, N. Thiessen, T. Cezard, A. P. Fejes, E. D. Wederell, R. Cullum, G. Euskirchen, M. Krzywinski, I. Birol, M. Snyder, P. A. Hoodless, M. Hirst, M. A. Marra and S. J. Jones (2008). "Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding." *Genome Res* **18**(12): 1906-1917.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder and S. Jones (2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." *Nat Methods* **4**(8): 651-657.

Rodriguez-Carballo, E., L. Lopez-Delisle, Y. Zhan, P. J. Fabre, L. Beccari, I. El-Idrissi, T. H. N. Huynh, H. Ozadam, J. Dekker and D. Duboule (2017). "The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes." *Genes Dev* **31**(22): 2264-2281.

Sanyal, A., B. R. Lajoie, G. Jain and J. Dekker (2012). "The long-range interaction landscape of gene promoters." *Nature* **489**(7414): 109-113.

Saxonov, S., P. Berg and D. L. Brutlag (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." *Proc Natl Acad Sci U S A* **103**(5): 1412-1417.

Schoenfelder, S., M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B. M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe and P. Fraser (2015). "The pluripotent

regulatory circuitry connecting promoters to their long-range interacting elements." *Genome Res* **25**(4): 582-597.

Sharifi-Zarchi, A., D. Gerovska, K. Adachi, M. Totonchi, H. Pezeshk, R. J. Taft, H. R. Scholer, H. Chitsaz, M. Sadeghi, H. Baharvand and M. J. Arauzo-Bravo (2017). "DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism." *BMC Genomics* **18**(1): 964.

Shen, S. Q., C. A. Myers, A. E. Hughes, L. C. Byrne, J. G. Flannery and J. C. Corbo (2016). "Massively parallel cis-regulatory analysis in the mammalian central nervous system." *Genome Res* **26**(2): 238-255.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov and B. Ren (2012). "A map of the cis-regulatory sequences in the mouse genome." *Nature* **488**(7409): 116-120.

Smith, C. M., T. F. Hayamizu, J. H. Finger, S. M. Bello, I. J. McCright, J. Xu, R. M. Baldarelli, J. S. Beal, J. Campbell, L. E. Corbani, P. J. Frost, J. R. Lewis, S. C. Giannatto, D. Miers, D. R. Shaw, J. A. Kadin, J. E. Richardson, C. L. Smith and M. Ringwald (2019). "The mouse Gene Expression Database (GXD): 2019 update." *Nucleic Acids Res* **47**(D1): D774-D779.

Spitz, F. and E. E. Furlong (2012). "Transcription factors: from enhancer binding to developmental control." *Nat Rev Genet* **13**(9): 613-626.

Thiery, J. P., G. Macaya and G. Bernardi (1976). "An analysis of eukaryotic genomes by density gradient centrifugation." *J Mol Biol* **108**(1): 219-235.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford and J. A. Stamatoyannopoulos (2012). "The accessible chromatin landscape of the human genome." *Nature* **489**(7414): 75-82.

Tokuda, N., M. Sasai and G. Chikenji (2011). "Roles of DNA looping in enhancer-blocking activity." *Biophys J* **100**(1): 126-134.

Tsompana, M. and M. J. Buck (2014). "Chromatin accessibility: a window into the genome." *Epigenetics Chromatin* **7**(1): 33.

Valouev, A., S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire and A. Sidow (2011). "Determinants of nucleosome organization in primary human cells." *Nature* **474**(7352): 516-520.

van de Werken, H. J., P. J. de Vree, E. Splinter, S. J. Holwerda, P. Klous, E. de Wit and W. de Laat (2012). "4C technology: protocols and data analysis." *Methods Enzymol* **513**: 89-112.

Vignali, M., A. H. Hassan, K. E. Neely and J. L. Workman (2000). "ATP-dependent chromatin-remodeling complexes." *Mol Cell Biol* **20**(6): 1899-1910.

Villar, D., C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek and D. T. Odom (2015). "Enhancer evolution across 20 mammalian species." *Cell* **160**(3): 554-566.

Vinogradov, A. E. (2003). "DNA helix: the importance of being GC-rich." *Nucleic Acids Res* **31**(7): 1838-1844.

Voigt, P., W. W. Tee and D. Reinberg (2013). "A double take on bivalent promoters." *Genes Dev* **27**(12): 1318-1338.

Wang, J., J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder and Z. Weng (2012). "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors." *Genome Res* **22**(9): 1798-1812.

Wang, X., G. O. Bryant, M. Floer, D. Spagna and M. Ptashne (2011). "An effect of DNA sequence on nucleosome occupancy and removal." *Nat Struct Mol Biol* **18**(4): 507-509.

Wang, Z., C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang and K. Zhao (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." *Nat Genet* **40**(7): 897-903.

White, M. A., C. A. Myers, J. C. Corbo and B. A. Cohen (2013). "Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks." *Proc Natl Acad Sci U S A* **110**(29): 11952-11957.

Yanez-Cuna, J. O., C. D. Arnold, G. Stampfel, L. M. Boryn, D. Gerlach, M. Rath and A. Stark (2014). "Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features." *Genome Res* **24**(7): 1147-1156.

Zeller, R., J. Lopez-Rios and A. Zuniga (2009). "Vertebrate limb bud development: moving towards integrative analysis of organogenesis." *Nat Rev Genet* **10**(12): 845-858.

Zentner, G. E., P. J. Tesar and P. C. Scacheri (2011). "Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions." *Genome Res* **21**(8): 1273-1283.

Zhang, T., S. Cooper and N. Brockdorff (2015). "The interplay of histone modifications - writers that read." *EMBO Rep* **16**(11): 1467-1481.

Zuniga, A. (2015). "Next generation limb development and evolution: old questions, new perspectives." *Development* **142**(22): 3810-3820.

Chapter 2: Investigating chromatin features that determine promoter contacts and transcriptional regulation

2.1 Declaration of Contributions

Yancoskie, M.N., Chan, Y.F. Investigating chromatin features that determine promoter contacts and transcriptional regulation.

(Manuscript in preparation)

***Author contributions*: M.N. Yancoskie** retrieved and processed sequence data, performed and interpreted all analyses, generated 4C-seq data at the *Nkx3-2* enhancer (see also **Chapter 3**), designed the figure flow, and wrote the manuscript.

Co-author contributions: Y.F. Chan provided guidance and feedback on all stages of analyses and writing, including helping to design figures and suggesting external works to integrate into the discussion.

2.2 Abstract

Linking *cis*-regulatory elements to the genes they regulate is challenging because they tend to act at a distance from their target promoters. Features such as chromatin accessibility and certain combinations of histone modification marks can be used to identify putative regulatory regions in the genome, but chromatin looping data is needed to confirm the interaction between a regulatory element and its target promoter. However, chromosome conformation capture assays like Capture-C which reveal such chromatin loops are noisy, and thus must be combined with other types of data to separate signal from noise. I modeled promoter contact frequency, or Capture-C signal, at 25 limb development genes along mouse Chromosome 7 as a function of chromatin accessibility, GC content, and histone modification. I categorized the viewpoint genes by their expression level in E13.5 forelimb and hindlimb and compared the predictive values of various chromatin features across the different expression states. Most promoter contacts were confined to the same topologically associating domain as the viewpoint promoter. Furthermore, distance from the viewpoint was the best predictor of promoter contact frequency. Promoters in transition formed more contacts and were less constrained by distance than active or silent promoters. The features included in the model explained only a small proportion of contacts, suggesting that the best way to identify biologically meaningful contacts is to first confine the search to the viewpoint TAD and to consider the expression level of the promoter of interest.

2.3 Introduction

Identifying genomic regulatory elements and the genes they regulate is an important challenge. Finding promoter contacts is confounded by the fact that mammalian *cis*-regulatory elements (CREs) can lie tens to hundreds of kilobase pairs away from their promoters and do not always act on the nearest promoters. Rather, distal CREs contact promoters by means of chromatin loops which mediate recruitment and localization of transcriptional machinery (Kadauke and Blobel, 2009). While enhancers and promoters are often flanked by nucleosomes with certain types of histone modifications, they may be highly tissue- and stage-specific, necessitating *in vivo* and *in vitro* assays to verify their activity in a given cell type (Pennacchio et al., 2006). Active promoters can be regulated by multiple enhancers in a combinatorial manner, with enhancers acting additively (Shin et al., 2016) or

synergistically (Marinic et al., 2013; Hay et al., 2016; Thormann et al., 2018) to refine one another's activity. Conversely, one enhancer can act on multiple promoters (Mohrs et al., 2001). The sequence code of enhancers has proven difficult to crack (Yanez-Cuna et al., 2014; Bu et al., 2017), in part because enhancer activity depends on cooperation between multiple transcription factors (TFs), which can be highly cell type-specific (Long et al., 2016). Silencers, similar to enhancers, act in a distance- and orientation-independent manner, recruiting transcriptional repressors to inhibit gene expression (Ogbourne and Antalis, 1998). Insulators, which contain binding sites for the architectural and chromatin structure protein CTCF, shield loci from interacting with one another (Kim et al., 2015). Silencers, insulators, and enhancers in one cell type may even act as another class in other cell types (Kolovos et al., 2012; Andersson et al., 2015). These dynamics enable precise control of gene expression but lend complexity to the genome, compounding efforts to map CREs to their target genes.

The chromatin looping structures that enable CREs to physically contact promoters so as to regulate their transcription can be captured by chromosome conformation capture (3C). 3C uses cross-linking to preserve chromatin in its native conformation followed by digestion, re-ligation, and sequencing to obtain a readout of interacting loci (Dekker et al., 2002). A 3C derivative known as Hi-C, which captures genome-wide interactions, has revealed the presence of topologically associating domains, or TADs: megabase-scale regions of high-density inter-chromosomal contacts in Hi-C matrices (Dixon et al., 2012). At the borders of TADs are boundary elements like the architectural protein CTCF which stop chromatin loops from proceeding along the DNA, shielding loci in neighboring TADs from interacting with one another (Fudenberg et al., 2016).

3C derivatives like Capture-C (Davies et al., 2016) and 4C-seq (Simonis et al., 2006) which target specific bait regions, or "viewpoints", provide high resolution of interacting loci. 3C assays are nevertheless prone to noise, and captured contacts may not necessarily represent biologically relevant interactions. Several experimental factors can affect ligation frequency and lead to spurious signal (Denker and de Laat, 2016), including cross-linking time and temperature (Dekker et al., 2002), ligation conditions, PCR amplification bias, and the cohesiveness of digested fragments (Gavrilov et al., 2013). One way to filter contacts is by comparing other datasets which mark putative CREs but do not link them to their targets. These

include enhancer transcription (eRNA) readout (Andersson et al., 2014), chromatin immunoprecipitation (ChIP-seq) of general TFs like p300 (Visel et al., 2009) and of histone modification marks commonly associated with enhancers and promoters (Calo and Wysocka, 2013), and chromatin accessibility maps (Thurman et al., 2012; Chen et al., 2018).

Understanding transcriptional regulatory dynamics has broad implications for studying disease and development. Identification and characterization of CREs is key to understanding how gene expression is carefully controlled during development (Marinic et al., 2013; Andrey and Mundlos, 2017), and provides targets for genome-editing therapies to counteract disease (Weischenfeldt et al., 2013; Lupianez et al., 2015). By combining Capture-C, a high-throughput 3C derivative that detects contacts for up to several hundreds of viewpoints simultaneously (Hughes et al., 2014; Davies et al., 2016), with datasets that mark putative CREs, I investigated the chromatin features associated with relevant promoter contacts.

2.4 Results

I analyzed Capture-C signal from E13.5 forelimb and hindlimb (Andrey et al., 2017) because this assay directly captures putative promoter-CRE contacts. To identify which genomic features may affect such contacts, I integrated additional types of stage- and tissue-matched data (**Table 2.1**) and analyzed what relationship, if any, these features may have with Capture signal. Assay for Transposase-Accessible Chromatin, or ATAC-seq (Buenrostro et al., 2015), provides a readout of open chromatin, which likely contains active and poised CREs since it is accessible to transcription machinery. Histone profiling uses ChIP-seq to characterize histone marks associated with gene regulatory functions. For example, acetylation of lysine 27 at histone H3 (H3K27ac) is conventionally associated with active enhancers, monomethylation of histone H3, lysine 4 (H3K4me1) with poised and active enhancers, and trimethylation of histone H3, lysine 4 (H3K4me3) with poised and active promoters. Trimethylation of histone H3, lysine 27 (H3K27me3) is associated with repression of transcription. Finally, GC content is correlated with transcription rate and gene expression, with GC-rich sequences and CpG islands enriched at enhancer and promoter elements (White et al., 2013; Colbran et al., 2017; Carelli et al., 2018; Lecellier et al., 2018).

Table 2.1. Genomic and physical properties to predict Capture-C promoter contacts.

Method or genomic feature	Aim	Samples Used in Analysis	Data Source
Capture-C	Capture chromatin contacts of many viewpoints (up to hundreds) at once.	E13.5 Forelimb and Hindlimb; 2 technical replicates each	Andrey et al., 2017
Hi-C	Derive topologically associating domains (TADs).	Mouse embryonic stem cells (mESCs) and mouse cortex cells; 2 technical replicates each	mESCs: Dixon et al., 2012 Cortex: Shen et al., 2012
ATAC-seq	Find regions of open chromatin	E13.5 limb bud; 2 technical replicates	ENCODE Project Consortium, 2012
GC content	Linked to gene expression (gene-rich regions are higher GC than gene-poor).	NA	*Mus musculus* reference mm10 (GRCm38) assembly
Histone profiling	Derive K4me3, K4me1, K27ac, K27me3 enrichment.	E13.5 Forelimb and Hindlimb; 2 technical replicates each	Andrey et al., 2017

I chose to analyze these genomic features because they were the most extensive functional genomics datasets available for mouse embryonic limb bud. Since CREs are often highly specific to stage and cell type, it was important to use data that matched the Capture-C samples. Other types of data associated with transcriptional regulation such as the genome-wide binding sites of key TFs, particularly those of the histone deacetyltransferase p300 – the most accurate *in-silico* predictor of enhancers (Visel et al., 2009) – were not available for this stage of embryonic limb bud.

The Capture-C dataset from Andrey et al., 2017 is a high-quality, high-throughput resource with which to study transcriptional regulation during mouse embryonic development. It captures the contacts of 446 limb developmental regulators across seven different biological samples: E10.5, E11.5, and E13.5 forelimb and hindlimb bud, and – as a negative control – E10.5 midbrain tissue. I selected E13.5 forelimb and hindlimb since I aimed to compare Capture-C signal with other types of data, rather than comparing across tissues and stages. Although hindlimb lags half a day behind forelimb development from E9.5 to E16.5 (Zuniga, 2015), the number of genes differentially expressed between forelimb and hindlimb at the same stage is very low (Andrey et al., 2017; Cotney et al., 2013). The

Capture-C signal profiles were also similar between the two tissues. Therefore, I combined forelimb and hindlimb replicates for most analyses.

Andrey et al., 2017 categorized Capture-C interactions into several histone profile-derived functional chromatin states, including repressed, heterogeneous, and active chromatin states using a Hidden Markov Model. A minority of the Capture-C interactions in each biological sample mapped to repressed chromatin (characterized by enrichment of H3K27me3 and depletion of H3K27ac, H3K4me3, and H3K4me1). The forelimb and hindlimb samples from E13.5 had the highest proportion of contacts mapping to repressed chromatin. To maximize the power of the statistical analysis, I focused on the stage with the greatest variation in viewpoint gene expression.

I perform a statistical analysis using ANCOVA modeling in order to explore the relationship between specific genomic features and promoter contacts (**Figure 2.1**).

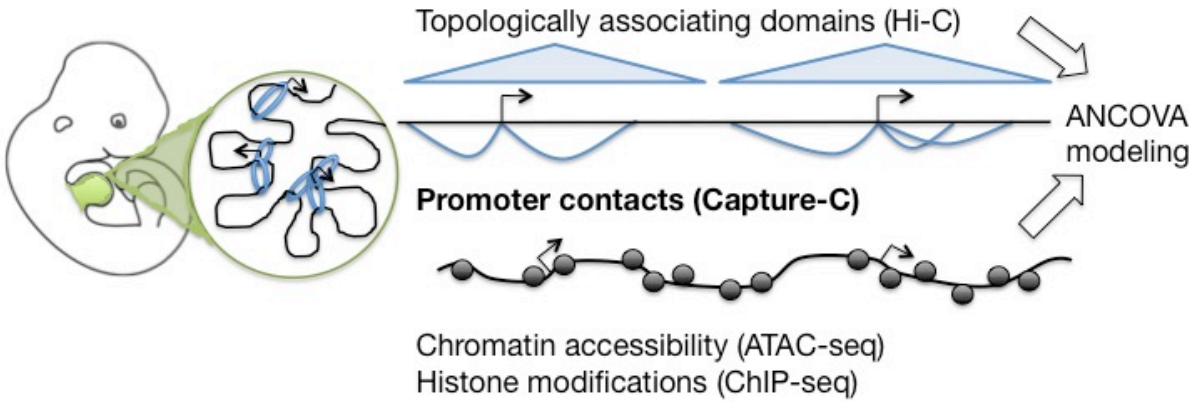


Figure 2.1. Molecular factors contributing to transcriptional regulation. Integrating multiple datasets can help identify and regulatory regions to their target promoters. Depicted on the left is an example of a chromatin segment in the limb bud cell of a developing mouse embryo. Cohesin (blue rings) translocates along the DNA (black curved line) to form looping structures, which bring promoters (black arrows) into contact with their regulatory regions. Averaging across a population of cells, promoter contacts (captured by Capture-C and Hi-C; chromatin loops depicted as blue curved lines) are generally confined to neighborhoods known as topologically associating domains (TADs; blue triangles). Through ANCOVA modeling, I explore how well chromatin and sequence features of the DNA, such as location with respect to the promoter TAD, nucleosome (black spheres) positioning (captured by ATC-seq), and histone modification marks (ChIP-seq) predict promoter contact frequency.

Distribution of Capture-C viewpoint contacts with respect to TADs

Each Capture-C viewpoint promoter had a high proportion of contacts within the same topologically associating domain, or TAD (**Figure 2.3**). Dixon et al., 2012

reported TADs to be highly stable across cell types and species. Therefore, I used mouse embryonic stem cell (mESC) TADs (Dixon et al., 2012) as a proxy for E13.5 limb bud TADs. Chromsome 7 has an exceptionally high number of TADs per unit length (**Figure 2.2**). Out of 157 mESC TADs, 20 contained Capture targets corresponding to 25 viewpoints (**Table 2.A**). The short TAD lengths mean that some of the viewpoint promoters are located close to the edges of TADs. For this reason, I chose to focus on Chromosome 7 TADs.

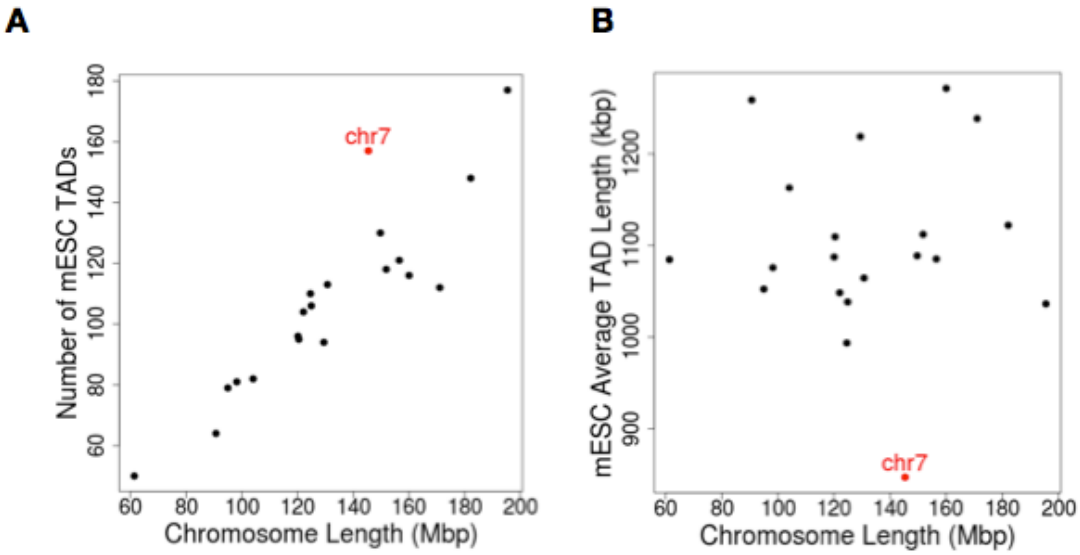


Figure 2.2. Chromosome 7 has an atypical distribution of mESC TADs. (**A**) Chromosome length and the number of mESC TADs per chromosome are positively correlated (R^2 = 0.81). Chromosome 7 has a high number of mESC TADs relative to its length. (**B**) There is no significant correlation between average TAD length and chromosome length (R^2 < 0.01). Chromosome 7 has a far shorter average mESC TAD length than the rest of the chromosomes.

Across the Chromosome 7 viewpoints, on average 45% (median; range 17-71%) of reads mapped within the same TAD (**Figure 2.3A**) despite the TADs occupying on average 0.58% of available Chromosome 7 mappable segments (**Figure 2.3C**), where segments were *DpnII* restriction fragments since *DpnII* was the four-cutter used to digest the chromatin during the Capture-C assay. By contrast, only 7% (1-27%) of reads map to the closest adjacent TADs (**Figure 2.3A**). These results reflect the signal decay that occurs over distance, and show that much of the chromosome contacts – and by extension, the transcriptional regulatory mechanisms – occur within the same TAD.

If mESC TAD boundaries were fully conserved with E13.5 limb bud TAD boundaries, and if they were fully effective at insulating neighboring domains from

interacting with one another, then viewpoints should have the same proportion of contacts in their assigned TAD irrespective of their position from the TAD boundary. Instead, the proportion of contacts in the TAD is lower for viewpoints at the edges of TADs. Viewpoints that are located closest to TAD boundaries (rightmost bars, **Figure 2.3B**) tend have more contacts in the closest neighboring TAD and fewer in the assigned TAD. This suggests that the mESC TAD boundaries may be lenient, may not be accurate for E13.5 limb bud, or both.

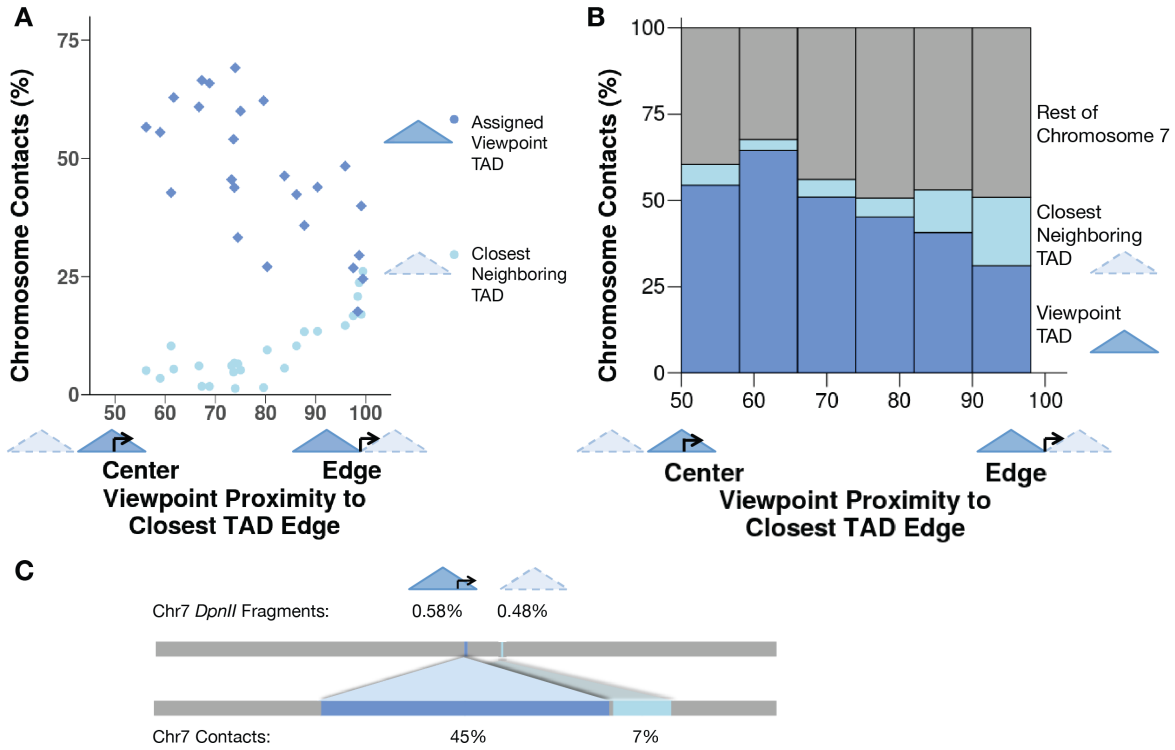
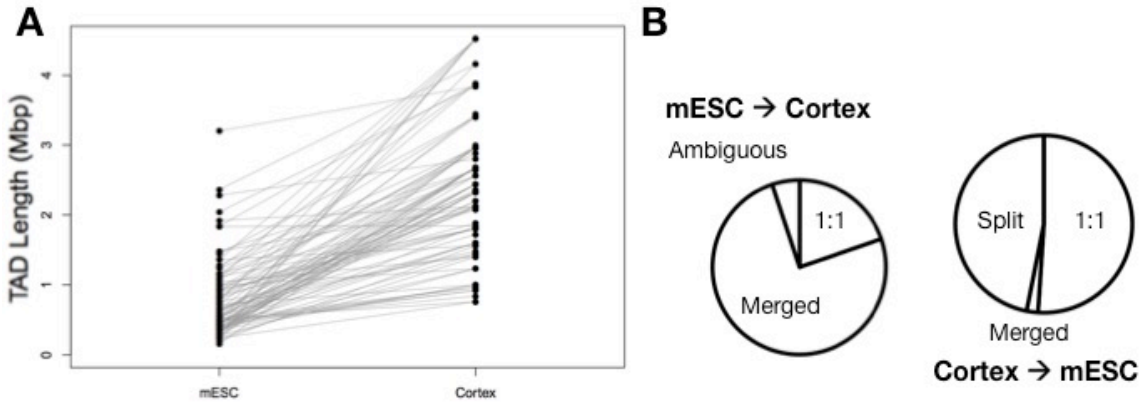


Figure 2.3. Promoters contact loci within their topologically associating domain, or TAD. (**A**) 45% (median; range 17-71%) of Chromosome 7 promoter contacts (blue diamonds) map to the same mouse embryonic stem cell TAD as the viewpoint (blue triangles with solid borders). By contrast, 7% (median; range 1-27%) of chromosome-wide contacts (light blue ovals) per viewpoint are in the closest neighboring TAD (light blue triangles with dashed borders). Each viewpoint is represented by a pair of data points with the same horizontal position. An x-axis score of 50 means that the viewpoint is at the center of its assigned TAD; an x-axis score of 100 means it is at the edge of the TAD; that is, the viewpoint is closer to the closest TAD boundary than 100% of *DpnII* fragments in the TAD. Read counts were averaged over 4 Capture samples per viewpoint (2 forelimb and 2 hindlimb replicates). To minimize the number of non-informative reads, alignments within the probe span ±1 kilobase pair were excluded from the analysis. (**B**) Viewpoints are binned into intervals based on their proximity to the closest TAD edge. Averaging across forelimb and hindlimb replicates of all viewpoints in an interval, 48% (median; range 31-54%) of Chromosome 7 contacts are in the viewpoint TAD (blue bars), 6% (3-20%) are in the closest neighboring TAD (light blue bars), and 45% (32-50%) are on the rest of Chromosome 7 (grey bars). (**C**) The viewpoint and closest neighboring TADs comprise on average 0.58% and 0.48% of the Chromosome 7 *DpnII* fragments, but the viewpoint TADs contain a high proportion of chromosome-wide contacts.

TAD boundary inaccuracy may stem from shifts in TAD boundaries over development; that is, mESC and E13.5 limb bud cells may have different TAD boundaries. To determine whether this could be the case, I compared TAD boundary positions in mESC and adult cortex cells (**Figure 2.4**). I overlapped combined Hi-C replicates from adult cortex cells (Shen et al., 2012) and from mESCs (Dixon et al., 2012) and compared their lengths. Frequently, adjacent mESC TADs overlapped at least 30% by cortex TADs were observed to merge into longer cortex TADs (**Figure 2.4A; B**). Reciprocally, cortex TADs overlapped at least 30% by mESC TADs were observed to split into multiple adjacent mESC TADs (**Figure 2.4C**). Overlapping TADs increase in length and Chromosome 7 TADs decrease in number from mESCs to cortex cells, suggesting that TAD boundaries have relaxed over developmental time. TAD boundaries tend to merge upon cellular differentiation in additional cell types (Meshorer and Misteli, 2006; Melcer and Meshorer, 2010; Gaspar-Maia et al., 2011; Boya et al., 2017; Battulin et al., 2015). The Capture-C viewpoint-containing mESC TADs on Chromosome 7 are usually shorter than the cortex TADs they overlap (**Figure 2.4D**). Even if TAD boundaries are conserved between mESCs and E13.5 limb bud cells, however, contacts may still form outside the viewpoint TAD due to the dynamic nature of TAD boundaries across different developmental contexts (Rodriguez-Carballo et al., 2017), or due to experimental noise. TAD coordinates represent an average across a population of cells whose exact boundaries may vary (Giorgetti et al., 2014; Liu and Tijan, 2018).



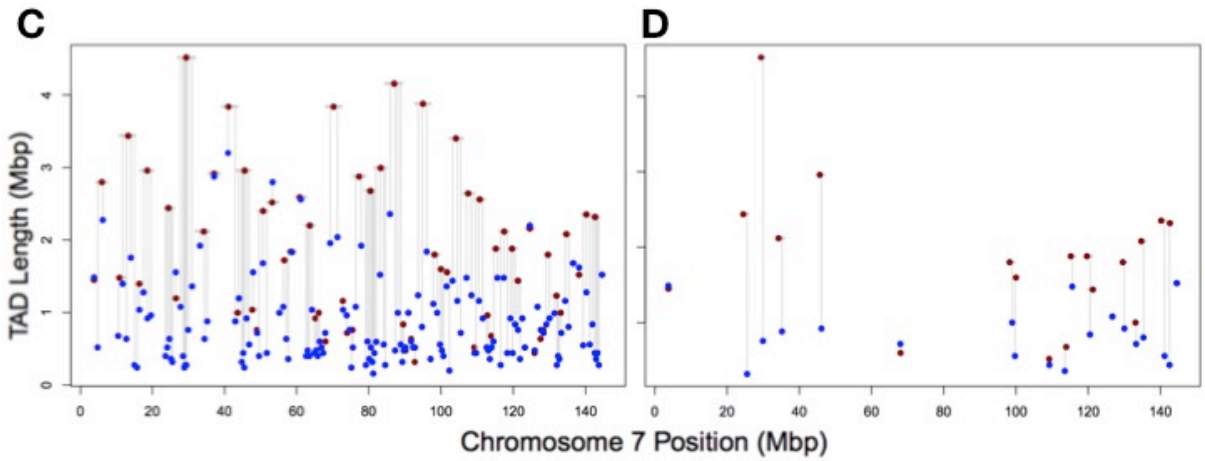


Figure 2.4. From mESC to adult cortex cells, TADs on Chromosome 7 tend to merge over developmental time. (**A**) Lengths of cortex TADs and the mESC TADs they overlap by at least 30% are plotted. (**B**) 75% of the mESC TADs overlapped at least 30% by cortex TADs can be merged into cortex TADs. Reciprocally, 47% of cortex TADs overlapped at least 30% by mESC TADs can be split into multiple adjacent mESC TADs. Cortex and mESC TAD boundaries were derived from the combined Hi-C replicates from Dixon et al., 2012. (**C**) Multiple short, adjacent TADs in mESCs (blue points) overlapped at least 30% by cortex TADs (red points) can often be merged into longer cortex TADs. (**D**) The Capture viewpoint-containing TADs follow the same overall trend as all Chromosome 7 TADs, with multiple adjacent mESC TADs tending to merge into larger cortex TADs.

Two studies showed that the majority of putative mouse promoter-enhancer contacts occur within the same TAD (Shen et al., 2012; Schoenfelder et al., 2015), and this analysis shows that the assigned viewpoint mESC TADs contain a high percentage of the promoter contacts. By focusing on the contacts within the viewpoint TAD, I increased the stringency of the analysis, limiting the amount of false positives – contacts outside the TAD that likely do not represent meaningful biological interactions – that could distort the outcome.

Proximity ligation and the confounding effect of distance

The high proportion of viewpoints within the TAD is confounded by the proximity ligation effect (Lajoie et al., 2015). During a Capture assay, the cross-linked genome is digested into restriction fragments that are subsequently ligated together. The fragments closest to the viewpoint on the linear DNA sequence have a higher likelihood of ligating to the viewpoint than do more distal fragments. This results in a typical profile where the signal is disproportionally strong at the viewpoint and decays rapidly over distance. It is visible on 4C-seq contact profiles (**Figure 1.3B**) as well as on Hi-C interaction matrices and can be modeled exponentially

(Lieberman-Aiden et al., 2009; Wijchers and de Laat, 2011). Peaks that break with the signal decay pattern often represent biologically relevant contacts.

To verify that the Capture-C signal follows the typical signal decay profile and that it is likely the result of proximity ligation, I plotted contact frequency over distance at the *Mylpf* viewpoint promoter and, as a control, at a viewpoint located in a gene desert (**Figure 2.5**). If signal decay is linked to the chromatin features found around genes, then it may not manifest at gene deserts. If, however, it stems from proximity ligation, then it should also be observed at viewpoints located far away from gene bodies.

Both at the *Mylpf* promoter (**Figure 2.5A**) and at a distal enhancer (located over 200 kbp upstream and 1.5 Mbp downstream of the closest genes) of the *Nkx3-2* gene on Chromosome 5 (**Figure 2.5B**), there is signal decay within the TAD (R^2 = 0.59 and 0.07, respectively). Since *Mylpf* is closer to the TAD edge than 96% of fragments in the TAD, many bins outside the TAD are closer to the viewpoint than within-TAD bins. These outside-TAD bins tend to have high Capture-C signal (leftmost grey data points from about 100 kbp to 1 Mbp away from the viewpoint). The line of best fit, however, is nearly horizontal (R^2 = 0.02) due to low signal in the rest of the outside-TAD bins. Likewise, signal decay rate is non-significant (R^2 < 0.01) for outside-TAD bins at the *Nkx3-2* enhancer viewpoint. The presence of signal decay at a viewpoint in a gene desert means it cannot be attributed to features found at genes. Instead, it is likely the result of proximity ligation noise.

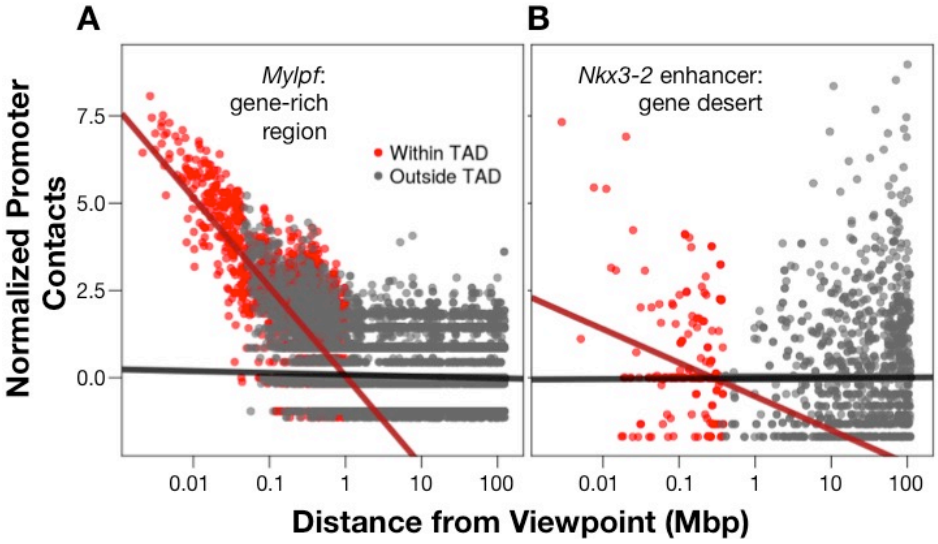


Figure 2.5. Capture-C signal decays rapidly within the viewpoint TAD. (**A**) At *Mylpf*, contact frequency ("Normalized Promoter Contacts", y-axis) is strongly negatively correlated (R^2 = 0.59) with fragment distance from the viewpoint (log of distance, x-axis) within the viewpoint TAD (red points; line of best fit in dark red). For fragments outside of the TAD (grey points; line of best fit in black), there is high Capture-C signal around 100 kbp to 1 Mbp from the viewpoint, but the line of best fit

is nearly horizontal (R^2 = 0.02) due to low signal in the rest of the outside-TAD bins. (**B**) 4C data from Longshanks (**Chapter 3**) from E14.5 forelimb bud at the distal *Nkx3-2* enhancer, located in a gene desert on Chromosome 5, displays significant signal decay within the viewpoint TAD (R^2 = 0.07) but not outside the TAD (R^2 < 0.01). Signal decay in (**A**) and (**B**) was plotted on the same scale to facilitate visual comparison between the two viewpoints. Capture-C signal was normalized across two and 4C across three technical replicates using a regularized log transformation (Love et al., 2014). Only one forelimb replicate of each viewpoint is plotted.

Modeling Promoter Contact Frequency with ANCOVA

To understand how the chromatin features explain the variation in Capture-C signal, I combined all the chromatin features into a single model. This approach also allowed me to account for potential interaction effects that may occur between the chromatin features, thus more accurately describing biological conditions. Because the genomic features used to predict promoter contact frequency (normalized Capture-C signal) include at least one categorical variable – namely, whether a given Chromosome 7 bin lies within or outside of the assigned viewpoint TAD, I used ANCOVA, or Analysis of Covariance (Eden and Fisher, 1927), to model the Capture-C signal. The ANCOVA model has the equation:

> Normalized Capture-C signal ~ TAD status (Outside | Within) + distance from viewpoint + ATAC-seq enrichment + GC content class + H3K27me3 signal + H3K27ac signal + H3K4me1 signal + H3K4me3 signal (+ interaction effects).

At the *Igf1r* viewpoint in Forelimb Replicate 1, I observed that TAD status was most highly predictive of promoter contact variance when all Chromosome 7 *DpnII* bins were considered, explaining over 40% of the variance regardless of whether pairwise interaction effects were included. When ANCOVA models were generated using only the *DpnII* bins within the TAD, distance explained over 20% of the variation. The other variables in the model each explained less than 5% of the variation (**Figure 2.6**).

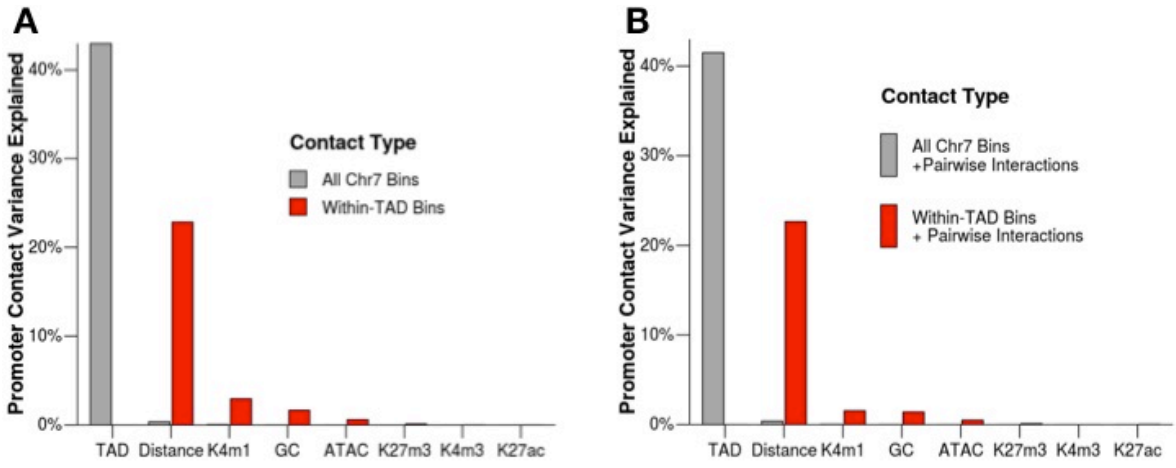


Figure 2.6. Example ANCOVA results at the *Igf1r* promoter in Forelimb replicate 1 (**A**) with no interaction effects included; (**B**) with all possible pairwise interaction effects. Grey bars depict Type II Analysis of Nested Covariance (ANCOVA) results from all Chromosome 7 *DpnII* fragments or bins. Red bars depict ANCOVA results across only the *DpnII* bins located within the same topologically associating domain (TAD) as the Capture-C viewpoint. The height of each bar plot reflects how much variation in promoter contact frequency (normalized Capture-C signal) that explanatory variable predicts. The explanatory variables include TAD location (within or outside the viewpoint TAD), log base 10 of the distance in base pairs from the viewpoint, ATAC-seq enrichment, histone profile signatures (H3K4me1, H3K4me3, H3K27ac, and H3K27me3), and GC content class as explanatory variables to predict the contact frequency, or normalized Capture-C signal.

ANCOVA model outcomes from all the Chromosome 7 viewpoints reveal similar trends to those observed at the *Igf1r* viewpoint. Within the viewpoint TAD, promoter contact frequency is strongly distance-dependent. Distance is by far the most predictive of all genomic features in the model. Across all viewpoints, the median within-TAD percent variation explained by distance is 33.7% (95.9% of the variation that can be explained), or 33.2% (96.4% of variation that can be explained) if all possible pairwise interaction terms are included in the model. Including or excluding pairwise effects does not appear to have a big influence on model outcome (**Figure 2.7**).

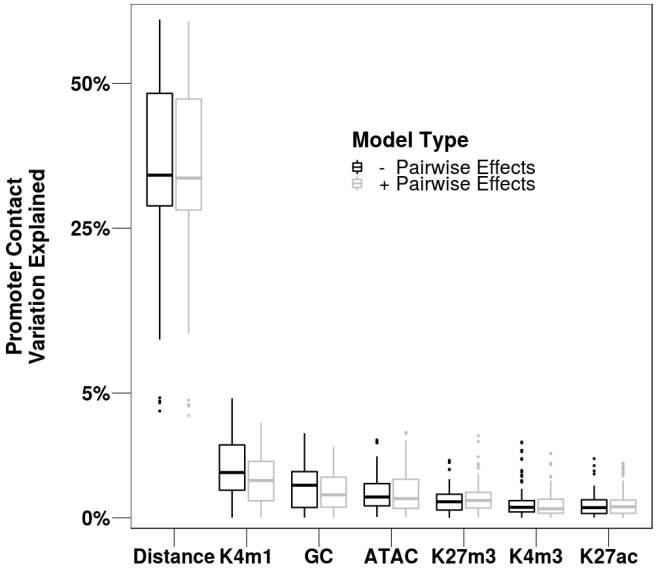


Figure 2.7. Analysis of Nested Covariance (ANCOVA) results for all Chromosome 7 Capture-C viewpoints. The black boxplots depict the amount of variation in promoter contact frequency (normalized Capture-C signal) explained or predicted by each explanatory variable when all *DpnII* bins within the viewpoint TAD (topologically associating domain) are included in the model. The grey boxplots depict the model outcome when all possible pairwise interactions between explanatory variables are included in the model.

Physical and biochemical properties of contacts with respect to viewpoint promoter activity level

The chromatin features of promoter contacts likely vary with promoter activity level, as chromatin has been observed to sequester into active and repressed hubs with distinct features (Harmston and Lenhard, 2013). Therefore, I next explored whether there were differences in ANCOVA model outcomes across different levels of viewpoint promoter activity. To determine whether the types of contacts promoters form vary with their expression levels, I assigned the viewpoint promoters to low, intermediate, or high expression categories based on their E13.5 forelimb and hindlimb RNA-seq levels (Andrey et al., 2017; **Figure 2.8**). I derived these expression categories from the sum of the normalized RNA-seq counts from all exons belonging to a viewpoint gene (low expression category genes had a median log of the normalized RNA-seq counts of -3.66; intermediate, -1.74; and high, -1.14). They predict 30% of the variation in the RNA-seq signal. Viewpoint promoters with intermediate expression levels may be in transition, switching from on to off or vice versa, or they may be a result of cellular heterogeneity where a high expression level in some cells in the limb bud is counteracted by a low level in others. To verify that the intermediate expression viewpoint genes are in transition, RNA-seq levels would need to be compared across expression stages before and after E13.5.

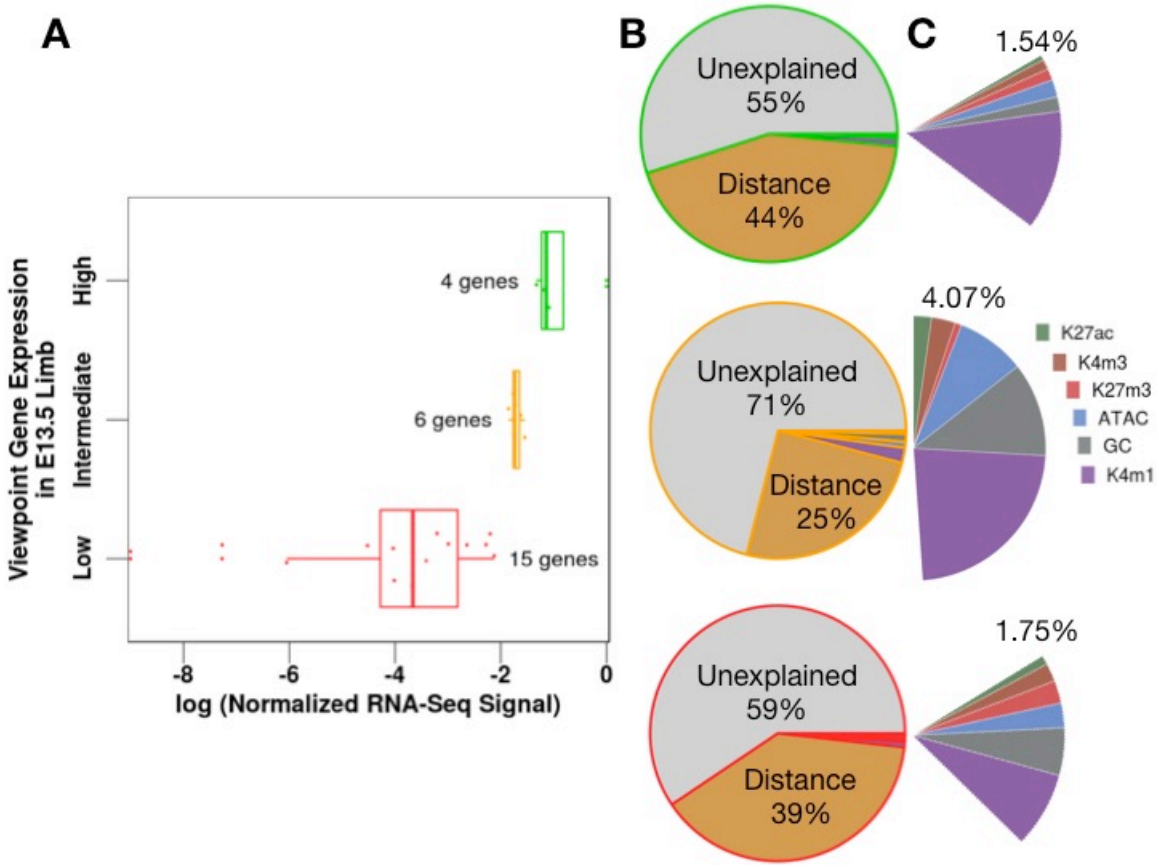


Figure 2.8. Promoter contact features vary with gene expression. (**A**) Viewpoint genes can be categorized by expression level in E13.5 forelimb and hindlimb based on cumulative RNA-seq signal. (**B** and **C**) Genomic features of promoter contacts differ across expression categories. Contacts of promoters with intermediate expression have greater unexplained variation, lower distance decay, and higher H3K4me1, GC content class, and ATAC-seq correlation. To create the pie charts, percent variation explained by each genomic feature was averaged across all viewpoints per expression category.

The directionality and magnitude of the relationship between promoter contact frequency and each genomic feature differs across the viewpoints, adding heterogeneity to each expression category (**Figure 2.9**). In accordance with its association with active and poised enhancers, H3K4me1 was positively correlated at all viewpoint promoters with intermediate to high expression levels, and at most promoters with low expression levels. A few viewpoint promoters with low or intermediate activity were positively correlated with the repressive mark H3K27me3. H3K27ac, which marks active enhancers and promoters, was positively correlated at most active and poised promoters. This matches previous findings (Simonis et al., 2006; Noordermeer et al., 2011, 2014; Vieux-Rochas et al., 2015) that active genes interact with open, active chromatin, whereas silent genes interact with inactive regions (Andrey et al., 2017).

ATAC-seq signal was nearly always positively correlated regardless of promoter activity level. In accordance, DNAse hypersensitive sites (DHSs), which ATAC-seq reveals, mark and historically have been used to identify all classes of CREs (Thurman et al., 2012; Chen et al., 2018) – including silencers, which contact repressed promoters (Ogbourne and Antalis, 1998). GC content class was positively correlated with signal at some promoters and negatively at others, seemingly independently of expression level. H3K4me3, which marks active, poised, and bivalent promoters (Heintzman et al., 2007), was negatively correlated with signal at nearly all viewpoint promoters. Although promoter-promoter interactions do play a role in transcriptional regulation by taking on enhancer-like roles (Kowalczyk et al., 2012; Leung et al., 2015) or by sharing *cis*-regulatory regions with alternate promoters of the same gene (Sanyal et al., 2012) or with promoters of genes with similar expression domains (Andrey et al., 2017), they likely represent a minority of meaningful chromatin interactions (Andersson et al., 2015).

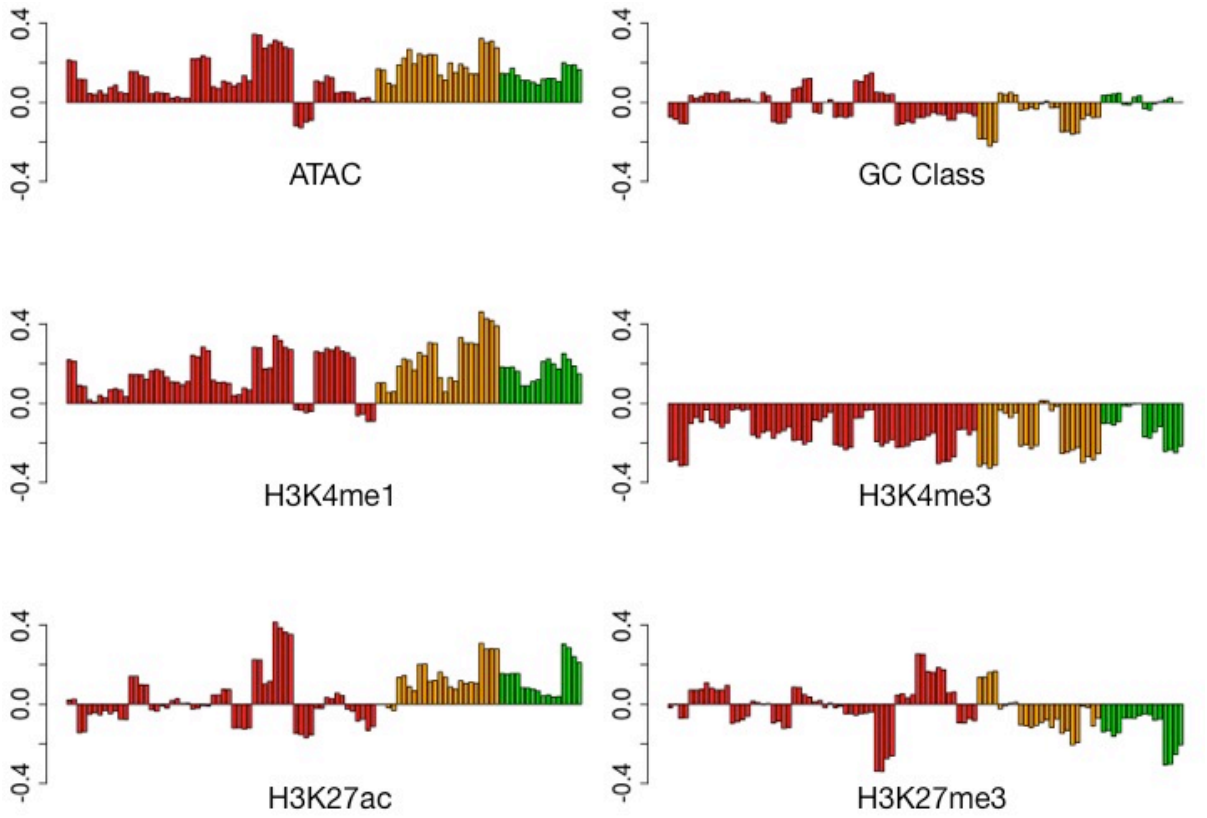


Figure 2.9. The relationship between promoter contact frequency and each explanatory variable differs across viewpoints. Correlation between promoter contact frequency (normalized Capture-C signal) and each genomic feature is plotted on the y-axis as Spearman's rho. The Chromosome 7 viewpoints are sorted by expression rank, with

viewpoints in red expressed at low levels in E13.5 limb bud; in orange at intermediate, and in green at high levels. All four replicates per viewpoint are plotted.

Low, intermediate, and high expression viewpoint genes have different averaged ANCOVA outcomes, with distance predicting less promoter contact variation at viewpoints with intermediate expression (**Figure 2.8B**). Besides distance, H3K4me1 signal, GC content class, and ATAC-seq signal explain more variation in contact frequency at promoters with intermediate expression than at active or silent promoters (H3K4me1: 0.7% at silent promoters, 1.9% at intermediate, 1.0% at active; GC content class: 0.4% at silent, 1.0% at intermediate, 0.1% at active; ATAC-seq: 0.2% at silent, 0.7% at intermediate, 0.2% at active) (**Figure 2.8C**). Intermediate promoters, however, have on average more unexplained variation (71%) than do low (59%) or high (55%) expression promoters. The contacts they form are less constrained by distance from the viewpoint, which explains 25% of the variation on average, compared to 39% at silent promoters and 44% at active promoters (**Figure 2.8B**).

2.5 Discussion

Chromosome conformation capture reveals interacting loci, which mediate transcriptional regulation. Chromatin features that confer CRE functionality can be used to find biologically relevant captured interactions. Through ANCOVA modeling of Capture-C signal with seven different genomic features, I characterized promoter contacts and how they vary with respect to gene expression. I found a high proportion of promoter contacts within the viewpoint TAD when mESC TAD boundaries were used as a proxy for the TAD boundaries in E13.5 limb bud, although this is potentially confounded by the proximity ligation effect. In line with the typically rapid signal decay produced by proximity ligation, distance was most predictive of the variation in Capture signal. Intermediate promoters, however, had on average less variation explained by distance, and more unexplained variation. Intermediate promoters may be transitioning from an active to an inactive state, or the reverse. Particularly during development, this switch may need to occur quickly in order to establish precise control of gene expression at the right stages and sections of the developing limb bud. Therefore, there may be rapid turnover or dissociation of the bulky complexes associated with transcriptional activation or repression (Swift and Coruzzi, 2017), and promoters in transition may experience

less steric hindrance in forming contacts. They may exist in a poised state in which they are more promiscuous than active or silent promoters, contacting loci enriched for genomic features other than those in the ANCOVA model (**Figure 2.10**). Interactions between poised promoters and enhancers have been observed at the Sonic hedgehog (*Shh*) promoter in anterior mouse limb bud cells, where *Shh* is not detectably transcribed despite contacting a known *Shh* limb enhancer (Amano et al., 2009).

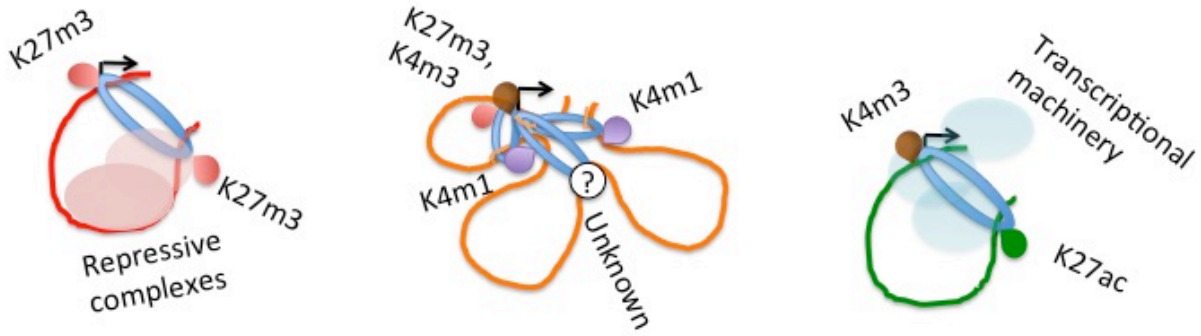


Figure 2.10. Transition promoters are less constrained than silent or active promoters in the contacts they form. The ANCOVA results suggest that active and silent promoters form fewer contacts than do transition promoters, possibly due to steric hindrance by transcriptional or repressive complexes. Curved lines, chromatin loops; blue rings, cohesin; colored teardrops, histone modification marks; black arrows, promoters; transparent spheres, large chromatin-associated complexes.

This model is supported by the dynamics of chromatin loop formation. According to loop extrusion theory, which has now been visualized in real-time in yeast (Ganji et al., 2018), loops form when loop extrusion factors such as the cohesin ring (in mammals) bind the DNA, bringing non-adjacent loci into contact with one another as they translocate along the DNA (Fudenberg et al., 2016). Depleting cohesin in human cells increased the average distance between interacting loci in the mouse, confirming its critical role in chromatin interactions (Wutz et al., 2017). When the extrusion factors encounter a boundary element like CTCF, often at the end of a TAD, translocation stalls and the loop cannot proceed beyond the TAD boundary. TAD boundaries are enriched not only for CTCF but also for active promoters and their associated transcriptional machinery. The latter is hypothesized to be capable of itself acting as a boundary element by physically disrupting the translocation of cohesin along the DNA (Fudenberg et al., 2016). Active promoters themselves, therefore, are subject to steric hindrance which prevents them from forming as many connections as transition promoters.

At inactive promoters, repressive complexes like the Polycomb Repressive Complex (PRC) may likewise contribute to steric hindrance if their dissociation rate from the DNA is sufficiently low. On a broader scale, long stretches of repressed chromatin interact with one another through the actions of PRC1 and PRC2 (Andrey et al., 2017). PRC2 both catalyzes and – in a positive feedback loop which enables it to maintain or foster the spread of repressive chromatin – recognizes the trimethylation of histone H3, lysine 27 (Berry et al., 2017). Through the spread of repressive marks, the chromatin is compartmentalized into laminar-associated domains (LADs) and inter-LADs. LADs comprise gene-poor, low GC content, closed chromatin localized to the nuclear periphery, whereas inter-LADs comprise gene-rich, higher GC content, open and active chromatin closer to the center of the nucleus (van Steensel and Belmont, 2017). Any Capture-C viewpoints located in LADs may be restricted from forming as many contacts due to the higher level of chromatin compaction in these territories. In support, 4C-seq in mouse immune cells showed that inactive viewpoints contacted fewer loci per chromatin loop than active viewpoints (Jiang et al., 2016).

The ANCOVA models in this work suggest that when attempting to identify CREs regulating a gene, one should first consider activity level of the gene – not only because chromatin interactions are known to take place between regions with similar levels of transcriptional activity (Andrey et al., 2017), but because if a gene of interest is expressed at an intermediate level or is poised, then it may be subject to fewer constraints than active or silent loci in the contacts it forms.

If relevant datasets other than those included here are available for the cell type of interest, they should be considered when attempting to predict regulatory function among promoter contacts. This is because the chromatin features I included predict less than half of the variation in Capture signal. If the lack of predictive power is because Capture signal is simply too noisy – due to exerimental conditions or to cellular heterogeneity within the limb bud (Andrey et al., 2017), including allele-specific differences (Davies et al., 2016) such as at the imprinted *Igf2* locus on Chromosome 7, then adding additional chromatin features might only result in incremental increases in model fit. However, a study predicting enhancer activity in mouse erythroid progenitors found TF occupancy to be a better predictor than either chromatin accessibility or histone modifications (Dogan et al., 2015). Two studies using chromatin features to predict contacts of active promoters in human cell lines

found DNase hypersensitivity and histone modification marks to have some predictive power (Roy et al., 2015), in accordance with this work, but also found CTCF, cohesin subunit Rad21 (Yang et al., 2017), and TF occupancy to be predictive. The abundance of predicted TF binding sites in the genome and the tendency of TFs to be highly cell type-specific and follow a complex, sometimes sub-optimal motif grammar makes accurate computational prediction of their binding sites challenging (Spitz and Furlong, 2012; Farley et al., 2015; Khamis et al., 2018; Keilwagen et al., 2019), but obtaining ChIP-seq data for each new cell type and TF would not be practical. In additon, determination of enhancer output is compounded by the cooperation between multiple TFs to out-compete nucleosomes for occupancy of the DNA (Long et al., 2016). In the absence of relevant TF binding data, consideration of the presence or absence of transcriptional or repressive complexes expected to localize at the viewpoint based on viewpoint gene expression level should guide the search for CREs.

2.6 Conclusion

Because the genomic features included in the ANCOVA models in this work predict only a small proportion of the variation in promoter contact frequency, future work characterizing promoter contacts should focus on other chromatin features that may confer *cis*-regulatory functionality. In particular, chromatin loop formation as it relate to the structural and biochemical properties of the DNA is not well understood. If two cohesin rings come into contact before either is forced to dissociate from the DNA by a boundary element at the end of the TAD, are there properties of the DNA – possibly bestowing greater TF binding affinity – that cause one chromatin loop to out-compete the other, or do the loops merge into one? As conformation capture techniques become even higher resolution, differences across individual cells need not be averaged. Indeed, such single-cell studies may reveal subtler trends in promoter contacts. In addition, to increase statistical power within each gene expression category, further explorations should incorporate many more viewpoints so as to determine whether poised promoters consistently display higher promiscuity as observed in this work.

2.7 Materials and Methods

Capture-C de-multiplexing and read count normalization

Publicly available Capture-C Sequence Read Archives from E13.5 forelimb and hindlimb (2 replicates each) were downloaded from the NCBI Gene Expression Omnibus under accession number GSE84795. Read 1 and Read 2 fastq files were derived with fastq-dump (v2.3.1; SRA-Toolkit; NCBI). Forward and reverse reads were interleaved with FLASH (v1.2.11). Interleaved reads were *in-silico* digested with *DpnII* using a custom perl script from the CC Analyser 3 pipeline, https://github.com/sudlab/Capture_C_Perl_Scripts. With bwa mem (Li, 2013), digested reads were aligned to mouse reference genome mm10 (GRCm38), sorted, and indexed. If one *DpnII* bin overlapped any of a viewpoint's Capture probe coordinates, all bins from that read pair were assigned to that viewpoint. Reads were filtered for assembly gaps from the UCSC Table Browser and for blacklisted regions (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/). Reads containing less than a mapping quality score of 20 were removed with samtools view (samtools-1.7). With bedtools coverage (v2.22.1), read counts were summed per *DpnII* bin. To minimize proximity ligation noise, reads encompassing the probe span (or, for the *Ctbp2* viewpoint, each of the two non-adjacent probe spans) and the kilobase immediately upstream or downstream of it were removed with bedtools intersect (v2.22.1). Read counts were normalized within each pair of replicates using a regularized log transformation from the DESeq2 package (v1.18.1) in R (v3.4.1).

TAD assignment

Each viewpoint was assigned to the topologically associating domain, or TAD, that overlapped its Capture probe span. mESC Hi-C domains (Dixon et al., 2012) were lifted over from mm9 to mm10 without allowing multi-mapping and were used as a proxy for E13.5 limb bud boundaries.

GC content class

GC content was calculated for each Chromosome 7 *DpnII* bin using bedtools nuc (v2.22.1), and GC content class was calculated by taking the absolute value of the standard deviation of all GC content scores across the chromosome.

ATAC-seq fold enrichment

ATAC-seq from E13.5 limb bud (2 technical replicates) were downloaded from the ENCODE Consortium as 50 nucleotide paired-end fastq files (https://www.encodeproject.org/experiments/ENCSR896XIN/). Reads were aligned to the genome with bowtie2 (v2.3.2) and blacklisted against the same regions as the Capture alignments. PCR duplicates were removed with picard-tools (v1.138) and alignments with a mapping quality score of less than 20 were excluded with samtools (v1.8). Fold enrichment of signal over noise was computed with macs2 (v2.1.1.20160309), with in-house ATAC-seq reads from genomic DNA, prepared according to Buenrostro et al., 2015, as the control. The intervals assigned by macs2 were merged where overlapping (bedtools merge v2.22.1) and were mapped onto the *DpnII* bins that overlapped at least 50% of their length. The 100^{th} quantile of ATAC-seq signal was selected per bin and was residualized for bin length. The fold enrichment residuals were averaged across the two ATAC-seq replicates.

ChIP-seq signal

Publicly available E13.5 forelimb and hindlimb (1 alignment each) ChIP-seq mm9 alignments for the histone modification marks H3K27ac, H3K27me3, H3K4me1, and H3K4me3 were downloaded from the NCBI Gene Expression Omnibus under accession number GSE84795. Overlapping intervals were merged with bedtools merge; merged intervals were lifted over to mm10 and parsed onto *DpnII* bins that overlapped them by at least 50%. The 100th quantile of ChIP-seq signal was selected for each *DpnII* bin and residualized for DpnII bin length. Forelimb and hindlimb ChIP-seq residuals were used in forelimb and hindlimb Capture-C ANCOVA models, respectively.

ANCOVA modeling

Variation in Capture-C normalized signal was modeled with ANCOVA Type II from the R package "cars" (Fox and Weisberg, 2011; v3.0.2; R version 3.4.1). The explanatory variables were the TAD status (whether each Chromosome 7 *DpnII* bin lies outside or within the assigned viewpoint TAD); log-transformed distance from each DpnII bin midpoint to the midpoint of each viewpoint Capture probe span; GC content class; ATAC-seq fold enrichment over genomic DNA; and each of the four ChIP-seq histone modification marks for E13.5 forelimb or hindlimb. Models were run with all Chromosome 7 *DpnII* bins and with just those lying within each viewpoint TAD. To account for potential interaction effects between explanatory variables, each ANCOVA model was run with either no pairwise interactions or all possible pairwise interactions.

Viewpoint Proximity to Closest TAD Edge

The distance in base pairs from the midpoint of each viewpoint's probe span to the closest TAD boundary was calculated with bedtools intersect. Distances were normalized by multiplying by 2, then dividing by TAD length (base pairs). The resulting value was subtracted from 1 to express the percentile of *DpnII* bins that were farther from the closest TAD edge than the viewpoint-containing *DpnII* bin.

Capture-C Viewpoint gene expression level

Publicly available mouse E13.5 forelimb and hindlimb RNA-seq (2 replicates each) mm9 (GRC37) bigWig alignments were downloaded from the NCBI Gene Expression Omnibus under accession number GSE84795. Coordinates were lifted over to mm10; overlapping intervals were merged with bedtools merge. Transcript ids were compiled from the UCSC Table Browser for each of the 25 viewpoints, and the unique exons for all transcripts of each viewpoint gene were retrieved. RNA-seq signal was averaged over each unique merged exon bed interval using bedtools map (v2.22.1). The average RNA-seq signal was summed across all unique exons for each viewpoint, then rescaled from 0-1 across all viewpoints. This rescaled cumulative RNA-seq signal was averaged across all limb bud RNA-seq replicates (2 replicates per tissue, hindlimb and forelimb). Each viewpoint gene was assigned to an expression level category of either low, intermediate, or high expression in E13.5 limb bud based on the distribution of the $\log_{10}$ RNA-seq values. These expression categories explained 30% of variation in the RNA-seq values (ANCOVA Type II model).

4C Signal at the Nkx3-2 Enhancer Viewpoint

For the generation of 4C data, the reader is referred to the Materials & Methods section of **Chapter 3**.

2.8 References to Chapter 2

Amano, T., T. Sagai, H. Tanabe, Y. Mizushina, H. Nakazawa and T. Shiroishi (2009). "Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription." *Dev Cell* **16**(1): 47-57.

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jorgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Muller, A. R. R. Forrest, P. Carninci, M. Rehli and A. Sandelin (2014). "An atlas of active enhancers across human cell types and tissues." *Nature* **507**(7493): 455-461.

Andersson, R., A. Sandelin and C. G. Danko (2015). "A unified architecture of transcriptional regulatory elements." *Trends Genet* **31**(8): 426-433.

Andrey, G. and S. Mundlos (2017). "The three-dimensional genome: regulating gene expression during pluripotency and development." *Development* **144**(20): 3646-3658.

Andrey, G., R. Schopflin, I. Jerkovic, V. Heinrich, D. M. Ibrahim, C. Paliou, M. Hochradel, B. Timmermann, S. Haas, M. Vingron and S. Mundlos (2017). "Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding." *Genome Res* **27**(2): 223-233.

Battulin, N., V. S. Fishman, A. M. Mazur, M. Pomaznoy, A. A. Khabarova, D. A. Afonnikov, E. B. Prokhortchouk and O. L. Serov (2015). "Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach." *Genome Biol* **16**: 77.

Berry, S., C. Dean and M. Howard (2017). "Slow Chromatin Dynamics Allow Polycomb Target Genes to Filter Fluctuations in Transcription Factor Activity." *Cell Syst* **4**(4): 445-457 e448.

Boya, R., A. D. Yadavalli, S. Nikhat, S. Kurukuti, D. Palakodeti and J. M. R. Pongubala (2017). "Developmentally regulated higher-order chromatin interactions orchestrate B cell fate commitment." *Nucleic Acids Res* **45**(19): 11070-11087.

Bu, H., Y. Gan, Y. Wang, S. Zhou and J. Guan (2017). "A new method for enhancer prediction based on deep belief network." *BMC Bioinformatics* **18**(Suppl 12): 418.

Buenrostro, J. D., B. Wu, H. Y. Chang and W. J. Greenleaf (2015). "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Curr Protoc Mol Biol* **109**: 21 29 21-29.

Calo, E. and J. Wysocka (2013). "Modification of enhancer chromatin: what, how, and why?" *Mol Cell* **49**(5): 825-837.

Carelli, F. N., A. Liechti, J. Halbert, M. Warnefors and H. Kaessmann (2018). "Repurposing of promoters and enhancers during mammalian evolution." *Nat Commun* **9**(1): 4066.

Chen, A., D. Chen and Y. Chen (2018). "Advances of DNase-seq for mapping active gene regulatory elements across the genome in animals." *Gene* **667**: 83-94.

Colbran, L. L., L. Chen and J. A. Capra (2017). "Short DNA sequence patterns accurately identify broadly active human enhancers." *BMC Genomics* **18**(1): 536.

Cotney, J., J. Leng, J. Yin, S. K. Reilly, L. E. DeMare, D. Emera, A. E. Ayoub, P. Rakic and J. P. Noonan (2013). "The evolution of lineage-specific regulatory activities in the human embryonic limb." *Cell* **154**(1): 185-196.

Davies, J. O., J. M. Telenius, S. J. McGowan, N. A. Roberts, S. Taylor, D. R. Higgs and J. R. Hughes (2016). "Multiplexed analysis of chromosome conformation at vastly improved sensitivity." *Nat Methods* **13**(1): 74-80.

Dekker, J., K. Rippe, M. Dekker and N. Kleckner (2002). "Capturing chromosome conformation." *Science* **295**(5558): 1306-1311.

Denker, A. and W. de Laat (2016). "The second decade of 3C technologies: detailed insights into nuclear organization." *Genes Dev* **30**(12): 1357-1382.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* **485**(7398): 376-380.

Dogan, N., W. Wu, C. S. Morrissey, K. B. Chen, A. Stonestrom, M. Long, C. A. Keller, Y. Cheng, D. Jain, A. Visel, L. A. Pennacchio, M. J. Weiss, G. A. Blobel and R. C. Hardison (2015). "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility." *Epigenetics Chromatin* **8**: 16.

Eden, T. and R. A. Fisher (1927). "Studies in crop variation IV The experimental determination of the value of top dressings with cereals." *Journal of Agricultural Science* **17**: 548-562.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome." *Nature* **489**(7414): 57-74.

Farley, E. K., K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar and M. S. Levine (2015). "Suboptimization of developmental enhancers." *Science* **350**(6258): 325-328.

Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur and L. A. Mirny (2016). "Formation of Chromosomal Domains by Loop Extrusion." *Cell Rep* **15**(9): 2038-2049.

Ganji, M., I. A. Shaltiel, S. Bisht, E. Kim, A. Kalichava, C. H. Haering and C. Dekker (2018). "Real-time imaging of DNA loop extrusion by condensin." *Science* **360**(6384): 102-105.

Gaspar-Maia, A., A. Alajem, E. Meshorer and M. Ramalho-Santos (2011). "Open chromatin in pluripotency and reprogramming." *Nat Rev Mol Cell Biol* **12**(1): 36-47.

Gavrilov, A. A., A. K. Golov and S. V. Razin (2013). "Actual ligation frequencies in the chromosome conformation capture procedure." *PLoS One* **8**(3): e60403.

Giorgetti, L., R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana and E. Heard (2014). "Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription." *Cell* **157**(4): 950-963.

Harmston, N. and B. Lenhard (2013). "Chromatin and epigenetic features of long-range gene regulation." *Nucleic Acids Res* **41**(15): 7185-7199.

Hay, D., J. R. Hughes, C. Babbs, J. O. J. Davies, B. J. Graham, L. Hanssen, M. T. Kassouf, A. M. Marieke Oudelaar, J. A. Sharpe, M. C. Suciu, J. Telenius, R. Williams, C. Rode, P. S. Li, L. A. Pennacchio, J. A. Sloane-Stanley, H. Ayyub, S. Butler, T. Sauka-Spengler, R. J. Gibbons, A. J. H. Smith, W. G. Wood and D. R. Higgs (2016). "Genetic dissection of the alpha-globin super-enhancer in vivo." *Nat Genet* **48**(8): 895-903.

Heintzman, N. D., R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford and B. Ren (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." *Nat Genet* **39**(3): 311-318.

Hughes, J. R., N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons and D. R. Higgs (2014). "Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment." *Nat Genet* **46**(2): 205-212.

Jiang, T., R. Raviram, V. Snetkova, P. P. Rocha, C. Proudhon, S. Badri, R. Bonneau, J. A. Skok and Y. Kluger (2016). "Identification of multi-loci hubs from 4C-seq demonstrates the functional importance of simultaneous interactions." *Nucleic Acids Res* **44**(18): 8714-8725.

Kadauke, S. and G. A. Blobel (2009). "Chromatin loops in gene regulation." *Biochim Biophys Acta* **1789**(1): 17-25.

Keilwagen, J., S. Posch and J. Grau (2019). "Accurate prediction of cell type-specific transcription factor binding." *Genome Biol* **20**(1): 9.

Khamis, A. M., O. Motwalli, R. Oliva, B. R. Jankovic, Y. A. Medvedeva, H. Ashoor, M. Essack, X. Gao and V. B. Bajic (2018). "A novel method for improved accuracy of transcription factor binding site prediction." *Nucleic Acids Res* **46**(12): e72.

Kim, S., N. K. Yu and B. K. Kaang (2015). "CTCF as a multifunctional protein in genome regulation and gene expression." *Exp Mol Med* **47**: e166.

Kolovos, P., T. A. Knoch, F. G. Grosveld, P. R. Cook and A. Papantonis (2012). "Enhancers and silencers: an integrated and simple model for their function." *Epigenetics Chromatin* **5**(1): 1.

Kowalczyk, M. S., J. R. Hughes, D. Garrick, M. D. Lynch, J. A. Sharpe, J. A. Sloane-Stanley, S. J. McGowan, M. De Gobbi, M. Hosseini, D. Vernimmen, J. M. Brown, N. E. Gray, L. Collavin, R. J. Gibbons, J. Flint, S. Taylor, V. J. Buckle, T. A. Milne, W. G. Wood and D. R. Higgs (2012). "Intragenic enhancers act as alternative promoters." *Mol Cell* **45**(4): 447-458.

Lajoie, B. R., J. Dekker and N. Kaplan (2015). "The Hitchhiker's guide to Hi-C analysis: practical guidelines." *Methods* **72**: 65-75.

Lecellier, C. H., W. W. Wasserman and A. Mathelier (2018). "Human Enhancers Harboring Specific Sequence Composition, Activity, and Genome Organization Are Linked to the Immune Response." *Genetics* **209**(4): 1055-1071.

Leung, D., I. Jung, N. Rajagopal, A. Schmitt, S. Selvaraj, A. Y. Lee, C. A. Yen, S. Lin, Y. Lin, Y. Qiu, W. Xie, F. Yue, M. Hariharan, P. Ray, S. Kuan, L. Edsall, H. Yang, N. C. Chi, M. Q. Zhang, J. R. Ecker and B. Ren (2015). "Integrative analysis of haplotype-resolved epigenomes across human tissues." *Nature* **518**(7539): 350-354.

Li, H. (2013) . “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.” arXiv:1303.3997v1 [q-bio.GN].

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* **326**(5950): 289-293.

Liu, Z. and R. Tjian (2018). "Visualizing transcription factor dynamics in living cells." *J Cell Biol* **217**(4): 1181-1191.

Long, H. K., S. L. Prescott and J. Wysocka (2016). "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution." *Cell* **167**(5): 1170-1187.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol* **15**(12): 550.

Lupianez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel and S. Mundlos (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions." *Cell* **161**(5): 1012-1025.

Marinic, M., T. Aktas, S. Ruf and F. Spitz (2013). "An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape." *Dev Cell* **24**(5): 530-542.

Melcer, S. and E. Meshorer (2010). "Chromatin plasticity in pluripotent cells." *Essays Biochem* **48**(1): 245-262.

Meshorer, E. and T. Misteli (2006). "Chromatin in pluripotent embryonic stem cells and differentiation." *Nat Rev Mol Cell Biol* **7**(7): 540-546.

Mohrs, M., C. M. Blankespoor, Z. E. Wang, G. G. Loots, V. Afzal, H. Hadeiba, K. Shinkai, E. M. Rubin and R. M. Locksley (2001). "Deletion of a coordinate regulator of type 2 cytokine expression in mice." *Nat Immunol* **2**(9): 842-847.

Noordermeer, D., M. Leleu, P. Schorderet, E. Joye, F. Chabaud and D. Duboule (2014). "Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci." *Elife* **3**: e02557.

Noordermeer, D., M. Leleu, E. Splinter, J. Rougemont, W. De Laat and D. Duboule (2011). "The dynamic architecture of Hox gene clusters." *Science* **334**(6053): 222-225.

Ogbourne, S. and T. M. Antalis (1998). "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes." *Biochem* J **331** (Pt 1): 1-14.

Pennacchio, L. A., N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel and E. M. Rubin (2006). "In vivo enhancer analysis of human conserved non-coding sequences." *Nature* **444**(7118): 499-502.

Rodriguez-Carballo, E., L. Lopez-Delisle, Y. Zhan, P. J. Fabre, L. Beccari, I. El-Idrissi, T. H. N. Huynh, H. Ozadam, J. Dekker and D. Duboule (2017). "The HoxD cluster is a dynamic

and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes." *Genes Dev* **31**(22): 2264-2281.

Roy, S., A. F. Siahpirani, D. Chasman, S. Knaack, F. Ay, R. Stewart, M. Wilson and R. Sridharan (2015). "A predictive modeling approach for cell line-specific long-range regulatory interactions." *Nucleic Acids Res* **43**(18): 8694-8712.

Sanyal, A., B. R. Lajoie, G. Jain and J. Dekker (2012). "The long-range interaction landscape of gene promoters." *Nature* **489**(7414): 109-113.

Schoenfelder, S., M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B. M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe and P. Fraser (2015). "The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements." *Genome Res* **25**(4): 582-597.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov and B. Ren (2012). "A map of the cis-regulatory sequences in the mouse genome." *Nature* **488**(7409): 116-120.

Shin, H. Y., M. Willi, K. HyunYoo, X. Zeng, C. Wang, G. Metser and L. Hennighausen (2016). "Hierarchy within the mammary STAT5-driven Wap super-enhancer." *Nat Genet* **48**(8): 904-911.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel and W. de Laat (2006). "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)." *Nat Genet* **38**(11): 1348-1354.

Spitz, F. and E. E. Furlong (2012). "Transcription factors: from enhancer binding to developmental control." *Nat Rev Genet* **13**(9): 613-626.

Swift, J. and G. M. Coruzzi (2017). "A matter of time - How transient transcription factor interactions create dynamic gene regulatory networks." *Biochim Biophys Acta Gene Regul Mech* **1860**(1): 75-83.

Thormann, V., M. C. Rothkegel, R. Schopflin, L. V. Glaser, P. Djuric, N. Li, H. R. Chung, K. Schwahn, M. Vingron and S. H. Meijsing (2018). "Genomic dissection of enhancers uncovers principles of combinatorial regulation and cell type-specific wiring of enhancer-promoter contacts." *Nucleic Acids Res* **46**(6): 3258.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford and J. A. Stamatoyannopoulos (2012). "The accessible chromatin landscape of the human genome." *Nature* **489**(7414): 75-82.

van Steensel, B. and A. S. Belmont (2017). "Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression." *Cell* **169**(5): 780-791.

Vieux-Rochas, M., P. J. Fabre, M. Leleu, D. Duboule and D. Noordermeer (2015). "Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain." *Proc Natl Acad Sci U S A* **112**(15): 4672-4677.

Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin and L. A. Pennacchio (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." *Nature* **457**(7231): 854-858.

Weischenfeldt, J., O. Symmons, F. Spitz and J. O. Korbel (2013). "Phenotypic impact of genomic structural variation: insights from and for human disease." *Nat Rev Genet* **14**(2): 125-138.

White, M. A., C. A. Myers, J. C. Corbo and B. A. Cohen (2013). "Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks." *Proc Natl Acad Sci U S A* **110**(29): 11952-11957.

Wijchers, P. J. and W. de Laat (2011). "Genome organization influences partner selection for chromosomal rearrangements." *Trends Genet* **27**(2): 63-71.

Wutz, G., C. Varnai, K. Nagasaka, D. A. Cisneros, R. R. Stocsits, W. Tang, S. Schoenfelder, G. Jessberger, M. Muhar, M. J. Hossain, N. Walther, B. Koch, M. Kueblbeck, J. Ellenberg, J. Zuber, P. Fraser and J. M. Peters (2017). "Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins." *EMBO J* **36**(24): 3573-3599.

Yanez-Cuna, J. O., C. D. Arnold, G. Stampfel, L. M. Boryn, D. Gerlach, M. Rath and A. Stark (2014). "Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features." *Genome Res* **24**(7): 1147-1156.

Yang, Y., R. Zhang, S. Singh and J. Ma (2017). "Exploiting sequence-based features for predicting enhancer-promoter interactions." *Bioinformatics* **33**(14): i252-i260.

Zuniga, A. (2015). "Next generation limb development and evolution: old questions, new perspectives." *Development* **142**(22): 3810-3820.

2.9 Appendix to Chapter 2

Table 2.A. Chromosome 7 Capture-C Viewpoint Genes.

Viewpoint Id	Gene name	Chr7 coordinates and gene length
Acan	Aggrecan	79.053-79.115 Mbp; 62 kbp
Atp2a1	ATPase, Ca2+ transporting, cardiac muscle, fast twitch 1	126.44-126.46 Mbp (-); 17 kbp
Ccnd1	Cyclin D1	144.930-144.940 Mbp (-);10 kbp
Cebpa	CCAAT/enhancer binding protein, alpha	35.119-35.122 Mbp (+); 2.6 kbp
Chrdl2	Chordin-like 2	100.006-100.034 Mbp (+); 28 kbp
Crym	Crystallin, mu	120.186-120.202 Mbp (-); 15.6 kbp
Ctbp2	C-terminal binding protein 2	132.987-133.124 Mbp (-); 137 kbp
Fgf3	Fibroblast growth factor 3	144.838-144.844 Mbp (+); 6.3 kbp
Fgf4	Fibroblast growth factor 4	144.861-144.865 Mbp (+); 3.8 kbp
Fgfr2	Fibroblast growth factor receptor 2	120.162-120.267 Mbp (-); 104 kbp
Ifitm5	Interferon induced transmembrane protein 5	140.949-140.950 Mbp (-); 1.3 kbp
Igf1r	Insulin-like growth factor 1 receptor	67.953-68.234 Mbp (+); 281 kbp
Igf2	Insulin-like growth factor 2	142.651-142.667 Mbp (-); 16 kbp
Lmo1	LIM domain only 1	109.139-109.175 Mbp (-); 36.6 kbp
Mki67	Antigen identified by monoclonal antibody Ki 67	135.690-135.716 Mbp (-); 26.6 kbp
Mylpf	Myosin light chain, phosphorylatable, fast skeletal muscle	127.209-127.214 Mbp (+); 5.4 kbp
Myod1	Myogenic differentiation 1	46.376-46.379 Mbp (+); 2.6 kbp
Pth	Parathyroid hormone	113.386-113.389 Mbp (-); 3.0 kbp
Sox6	Sex determining region Y-box 6	115.471-116.039 Mbp (-); 568 kbp
Tbx6	T-box 6	126.781-126.786 Mbp (+); 4.1 kbp
Tgfb1	Transforming growth factor, beta 1	25.687-25.705 Mbp (+); 17.1 kbp
Tnni2	Troponin 1, skeletal, fast 2	142.442-142.444 Mbp (+); 2.6 kbp
Tnnt1	Troponin T1, skeletal, slow	4.505-4.516 Mbp (-); 11.4 kbp
Wnt11	Wingless-type MMTV integration site family, member 11	98.835-98.855 Mbp (+); 19.8 kbp
Zfp260	Zinc finger protein 260	30.095-30.108 Mbp (+); 12.8 kbp

Chapter 3: An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice

3.1 Declaration of Contributions

Castro, J.L.P.*, **Yancoskie, M.N.***, Marchini, M., Belohlavy, S., Hiramatsu, L., Kučka, M., Beluch, W.H., Naumann, R., Skuplik, I., Cobb, J., Barton, N.H., Rolian, C.R.†, Chan, Y.F.† (2019) An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. *eLife* (**8**):e42014.

(* co-first authorship; † co-last authorship)

Author contributions*: M.N. Yancoskie** designed, performed, and optimized multiplexed 4C-seq experiments, processed, analyzed, and interpreted sequencing results, generated and analyzed ATAC-seq data, wrote figure legends and methods for the experiments analyzing the chromatin features and *cis*-regulatory function at the *Nkx3-2* and *Gli3* loci (molecular dissection experiments), helped design figure panels, and provided feedback at all stages of manuscript writing. ***Relevance to the collective work: The 4C data provided critical confirmation that the N1, N2, and N3 enhancers act on the *Nkx3-2* promoter, formally providing the functional links between the *Nkx3-2* gene body and the putative enhancers. The ATAC-Seq data provided evidence independent from ENCODE results to support the identification of enhancers, which were essential to our molecular dissection and validation of the top loci.

Co-author contributions: J.L.P. Castro determined *Gli3* to be a candidate gene for molecular dissection; initiated stickleback transgenic experiments; identified candidate enhancers based on histone modification enrichment; designed, performed, and analyzed results of mouse and stickleback transgenic reporter assays; and performed a subset of *in situ* hybridization experiments. M. Marchini helped set up and perform artificial selection, phenotyped Longshanks and Control mice, analyzed pedigree data, and collected tissue samples for genome sequencing. S. Belohlavy analyzed pedigree and genomic data, performed and analyzed simulations, and helped draft the revised text. L. Hiramatsu analyzed pedigree and genomic data, estimated changes to additive variance during the experiment, and helped draft the revised text and revision response. M. Kučka prepared samples for sequencing, helped clone and prepare transgenic reporter plasmids and helped design the multiplexed 4C-seq experiment. W.H. Beluch prepared samples for genome sequencing. R. Naumann performed transient transgenic reporter injections. I. Skuplik and J. Cobb performed and analyzed results from *in situ* hybridizations from the Longshanks lines. N.H. Barton performed and analyzed simulations and analyzed pedigree and genomic data. C. Rolian designed and initiated the Longshanks selection experiment, phenotyped and collected tissue samples for sequencing, designed sequencing strategy, analyzed pedigree and genomic data, designed and analyzed functional experiments, and helped write the manuscript. Y.F. Chan initiated the genomic analysis of the selection experiment, designed sequencing strategy, analyzed pedigree and genomic data, and designed, performed, and analyzed functional experiments, and wrote the manuscript and subsequent editing and revision.

3.2 Full Article

An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice

João PL Castro[1‡], Michelle N Yancoskie[1‡], Marta Marchini[2], Stefanie Belohlavy[3], Layla Hiramatsu[1], Marek Kučka[1], William H Beluch[1], Ronald Naumann[4],Isabella Skuplik[2], John Cobb[2], Nicholas H Barton[3], Campbell Rolian[2†*],Yingguang Frank Chan[1†*]

[1]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany; [2]University of Calgary, Calgary, Canada; [3]Institute of Science and Technology (IST) Austria, Klosterneuburg, Austria; [4]Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany

*For correspondence: cprolian@ucalgary.ca (CR); frank.chan@tue.mpg.de (YFC)
[†]These authors also contributed equally to this work
[‡]These authors also contributed equally to this work
Competing interests: The authors declare that no competing interests exist.

Received: 14 September 2018
Accepted: 19 May 2019
Published: 06 June 2019
Reviewing editor: Magnus Nordborg, Austrian Academy of Sciences, Austria

Abstract

Evolutionary studies are often limited by missing data that are critical to understanding the history of selection. Selection experiments, which reproduce rapid evolution under controlled conditions, are excellent tools to study how genomes evolve under selection. Here we present a genomic dissection of the Longshanks selection experiment, in which mice were selectively bred over 20 generations for longer tibiae relative to body mass, resulting in 13% longer tibiae in two replicates. We synthesized evolutionary theory, genome sequences and molecular genetics to understand the selection response and found that it involved both polygenic adaptation and discrete loci of major effect, with the strongest loci tending to be selected in parallel between replicates. We show that selection may favor de-repression of bone growth through inactivating two limb enhancers of an inhibitor, *Nkx3-2*. Our integrative genomic analyses thus show that it is possible to connect individual base-pair changes to the overall selection response.

Introduction

Understanding how populations adapt to a changing environment is an urgent challenge of global significance. The problem is especially acute for mammal populations, which are often small and fragmented due to widespread habitat loss. Such populations often show increased inbreeding, leading to the loss of genetic diversity (***Hoffmann and Sgrò, 2011***). Because beneficial alleles in mammals typically come from standing genetic variation rather than new mutations like in microbes, this loss of diversity would ultimately impose a limit on the ability of small populations to adapt. Nonetheless, mammals respond readily to selection in many traits, both in nature and in the laboratory (***Darwin, 1859***; ***Gingerich, 2001***; ***Garland and Rose, 2009***; ***Keightley et al., 2001***). In quantitative genetics, such traits are interpreted as the overall effect from a large set of loci, each with an infinitesimally small (and undetectable) effect ('infinitesimal model'). Broadly speaking, the infinitesimal model has performed remarkably well across a wide range of selection experiments, and the model is the basis for commercial breeding (***Walsh and Lynch, 2018***). However, it remains unclear what type of genomic change is associated with rapid response to selection, especially in small populations where allele frequency changes can be dominated by random genetic drift.

While a large body of theory exists to describe the birth, rise and eventual fixation of adaptive variants under diverse selection scenarios (***Maynard Smith and Haigh, 1974***; ***Barton, 1995***; ***Otto and Barton, 2001***; ***Weissman and Barton, 2012***; ***Crow and Kimura, 1965***; ***Hill and Robertson, 1966***), few empirical datasets capture sufficient detail on the founding conditions and selection regime to allow full reconstruction of the selection response. This is particularly problematic in nature, where historical samples, environmental measurements and replicates are often missing. Selection experiments, which reproduce rapid evolution under controlled conditions, are therefore excellent tools to understand response to selection—and by extension—adaptive evolution in nature (***Garland and Rose, 2009***).

Here we describe an integrative, multi-faceted investigation into an artificial selection experiment, called Longshanks, in which mice were selected for increased tibia length relative to body mass (***Marchini et al., 2014***). The mammalian limb is an ideal model to study the dynamics of complex traits under selection: it is both

morphologically complex and functionally diverse, reflecting its adaptive value; and limb development has been studied extensively in mammals, birds and fishes as a genetic and evolutionary paradigm (***Petit et al., 2017***). The Longshanks selection experiment thus offers the opportunity to study selection response not only from a quantitative and population genetics perspective, but also from a developmental (***Marchini and Rolian, 2018***) and genomic perspective.

By design, the Longshanks experiment preserves a nearly complete archive of the phenotype (trait measurements) and genotype (via tissue samples) in the pedigree. Previously, Marchini et al. investigated how selection was able to overcome correlation between tibia length and body mass and produced independent changes in tibia length during the first 14 generations of the Longshanks experiment (***Marchini et al., 2014***). Importantly, that study focused on the phenotypes and inferred genetic correlations indirectly using the pedigree. The current genomic analysis was initiated when the on-going experiment reached generation 17 and extends the previous study by integrating both phenotypic and genetic aspects of the Longshanks experiment. By sequencing the initial and final genomes, the current analysis benefits from direct and highly resolved genetic information. Here, with essentially complete information, we wish to answer a number of important questions regarding the factors that determine and constrain rapid adaptation: Are the observed changes in gene frequency due to selection or random drift? Does rapid selection response of a complex trait proceed through innumerable loci of infinitesimally small effect, or through a few loci of large effect? What type of signature of selection may be associated with this process? Finally, when the same trait changes occur independently, do these depend on changes in the same gene(s) or the same pathways (parallelism)?

Results

Longshanks selection for longer tibiae

At the start of the Longshanks experiment, we established three base populations with 14 pairs each by sampling from a genetically diverse, commercial mouse stock (Hsd:ICR, also known as CD-1; derived from mixed breeding of classical laboratory mice [***Yalcin et al., 2010***]). In two replicate ‘Longshanks’ lines (LS1 and LS2), we

bred mice by pairing 16 males and females (and excluding sibling pairs) with the longest tibia relative to the cube root of body mass for each sex. This corresponds to 15–20% of all offspring (only details essential to understanding our analysis are summarized here. See ***Marchini et al., 2014*** for a detailed description of the breeding scheme). We kept a third Control line (Ctrl) using an identical breeding scheme, except that breeders were selected at random. In LS1 and LS2, we observed a strong and significant response to selection in tibia length (0.29 and 0.26 Haldane or standard deviations (s.d.) per generation, from a selection differential of 0.73 s.d. in LS1 and 0.62 s.d. in LS2). Over 20 generations, selection for longer relative tibia length produced increases of 5.27 and 4.81 s.d. in LS1 and LS2, respectively (or 12.7% and 13.1% in tibia length), with a modest decrease in body mass (−1.5% in LS1 and −3.7% in LS2; Student's *t*-test, $p<2 \times 10^{-4}$ and $p<1 \times 10^{-8}$, respectively; ***Figure 3.1B and C***; ***Figure 3.2***; n.b. this relationship was in part biased by the F1 generation, which were fed a different diet and phenotyped three weeks later than later generations, see ***Marchini et al., 2014*** for details). By contrast, Ctrl showed no directional change in tibia length or body mass (***Figure 3.1C***; Student's *t*-test, $p>0.05$). This approximately 5 s.d. change in 20 generations is rapid compared to typical rates observed in nature (***Hendry and Kinnison, 1999***, but see ***Grant and Grant, 2002***) but is in line with responses seen in selection experiments (***Gingerich, 2001***; ***Keightley et al., 2001***; ***Falconer and Mackay, 1996***; ***Pitchers et al., 2014***).

Simulating selection response: infinitesimal model with linkage

The rapid but generally smooth increase in tibia length in Longshanks is typically interpreted as evidence for a highly dispersed genetic architecture with no individually important loci contributing to the selection response. This is classically described under quantitative genetics as the infinitesimal model. Crucially, the appropriate null hypothesis for the genomic response here should capture "polygenic adaptation" rather than a neutral model. We therefore developed a simulation that faithfully recapitulates the artificial selection experiment by integrating the trait measurements, selection regime, pedigree and genetic diversity of the Longshanks selection experiment, in order to generate an accurate expectation for the genomic response. Using the actual pedigree and trait measurements, we mapped fitness onto tibia length $\boldsymbol{T}$ and cube-root body mass $\boldsymbol{B}$ as a single composite trait $ln(\boldsymbol{TB}^{\phi})$. We estimated ϕ from actual data as −0.57, such that the ranking of breeders closely

matched the actual composite ranking used to select breeders in the selection experiment, based on ***T*** and ***B*** separately (***Marchini et al., 2014***) (***Figure 3.3A***). We assumed a maximally polygenic genetic architecture using an "infinitesimal model with linkage" (abbreviated here as H_{INF}), under which the trait is controlled by very many loci, each of infinitesimally small effect (see Appendix for details). Results from simulations seeded with actual genotypes or haplotypes showed that overall, the predicted increase in inbreeding closely matched the observed data (***Figure 3.3B***). We tested models with varying selection intensity and initial linkage disequilibrium (LD), and for each, ran 100 simulated replicates to determine the significance of changes in allele frequency (***Figure 3.3C-E***). This flexible quantitative genetics framework allowed us to explore possible changes in genetic diversity over 17 generations of breeding under strong selection.

In simulations, we followed blocks of genomes as they were passed down the pedigree. In order to compare with observations, we seeded the initial genomes with single nucleotide polymorphisms (SNPs) in the same number and initial frequencies as the data. We observed much more variation between chromosomes in overall inbreeding (***Figure 3.3B***) and in the distribution of allele frequencies (***Figure 3.5B***) than expected from simulations in which the ancestral SNPs were initially in linkage equilibrium. This can be explained by linkage disequilibrium (LD) between the ancestral SNPs, which greatly increases random variation. Therefore, we based our significance threshold tests on simulations that were seeded with SNPs drawn with LD consistent with the initial haplotypes (***Figure 3.3C and E***; see Appendix).

Because our simulations assume infinitesimal effects of loci, allele frequency shifts exceeding this stringent threshold would suggest that discrete loci contribute significantly to the selection response. An excess of such loci in either a single LS replicate or in parallel would thus imply a mixed genetic architecture of a few large-effect loci amid an infinitesimal background.

Sequencing the Longshanks mice reveals genomic signatures of selection

To detect the genomic changes in the actual Longshanks experiment, we sequenced all individuals of the founder (F0) and 17th generation (F17) to an average of 2.91-fold coverage (range: 0.73–20.6×; n = 169 with <10% missing F0 individuals;

Supplementary file 1). Across the three lines, we found similar levels of diversity, with an average of 6.7 million (M) segregating SNPs (approximately 0.025%, or 1 SNP per four kbp; ***Supplementary file 2***; ***Figure 3.5A and Figure 3.6***). We checked the founder populations to confirm negligible divergence between the three founder populations (across-line F_{ST} on the order of 1×10^{-4}), which increased to 0.18 at F17 (***Supplementary file 2***). This is consistent with random sampling from an outbred breeding stock. By F17, the number of segregating SNPs dropped to around 5.8 M (***Supplementary file 2***). This 13% drop in diversity (0.9M SNPs genome-wide) is predicted by drift. Our simulations confirmed this and moreover, showed that selection contributed negligibly to the drop in diversity (Appendix, ***Figure 3.3B, D***).

We conclude that despite the strong selection on the LS lines, there was little perturbation to genome-wide diversity. Indeed, the changes in diversity in 17 generations were remarkably similar in all three lines, despite Ctrl not having experienced selection on relative tibia length (***Figure 3.5A***). Hence, and consistent with our simulation results (***Figure 3.3B,D***), changes in global genome diversity had little power to distinguish selection from neutral drift despite the strong *phenotypic* selection response.

We next asked whether specific loci reveal more definitive differences between the LS replicates and Ctrl (and from infinitesimal predictions). We calculated Δz^2, the square of an arcsine transformed allele frequency difference between F0 and F17; this has an expected variance of $1/2N_e$ per generation, independent of starting frequency, and ranges from 0 to π^2. We averaged Δz^2 within 10 kbp windows (see Methods for details), and found 169 windows belonging to eight clusters (i.e., loci) that had significant shifts in allele frequency in LS1 and/or LS2 (corresponding to 9 and 164 clustered windows respectively at $p \leq 0.05$ under $H_{INF,\ max\ LD}$; $\Delta z^2 \geq 0.33\ \pi^2$; genome-wide $\Delta z^2 = 0.02 \pm 0.03\ \pi^2$; ***Figure 3.4***; ***Figure 3.3D***, ***Figure 3.6***, ***Figure 3.7***; see Methods for details) and 8 windows in three clusters in Ctrl (genome-wide $\Delta z^2 = 0.01 \pm 0.02\ \pi^2$). The eight loci in Longshanks each overlapped between 2 to 179 genes and together contained 11 candidate genes with known roles in bone, cartilage and/or limb development (e.g., *Nkx3-2* and *Sox9*; ***Table 3.1***; ***Figure 3.7***, ***Figure 3.8***). Four out of the eight loci contain genes with a ‘short tibia’ or ‘short limb’ knockout phenotype (***Table 3.1***; $p \leq 0.032$ from 1000 permutations, see Methods for

details). Of the broader set of genes at these loci with any limb knockout phenotypes, only fibrillin 2 (*Fbn2*) is polymorphic for SNPs coding for different amino acids, suggesting that for the majority of loci with large shifts in allele frequency, gene regulation was likely important in the selection response (***Figure 3.8***; ***Supplementary file 3***; see Appendix for further analyses on enrichment in gene functions, protein-coding vs. *cis*-acting changes and clustering with loci affecting human height).

Taken together, two major observations stand out from our genomic survey. One, a polygenic, infinitesimal selection model with strong LD among marker SNPs performed better than moderate LD or no LD (***Figure 3.3E***); and two, we nevertheless find more discrete loci in LS1 and LS2 than in Ctrl, beyond the significance threshold set by the infinitesimal model (***Figure 3.4; Figure 3.6***). Thus, we conclude that although the genetic basis of the selection response in the Longshanks experiment may be largely polygenic, evidence strongly suggests discrete loci with major effect, even when each line is considered separately.

We next tested the repeatability of the selection response at the gene/locus level using the two LS replicates. If the founding populations shared the same selectively favored variants, we may observe parallelism or co-incident selective sweeps, as long as selection could overcome random drift. Indeed, the Δz^2 profiles of LS1 and LS2 were more similar to each other than to Ctrl (***Figure 3.4 and 3.9A***; ***Figure 3.10***; Pearson's correlation in Δz^2 from 10 kbp windows: LS1–LS2: 0.21 vs. LS1–Ctrl: 0.06 and LS2–Ctrl: 0.05). Whereas previous genomic studies with multiple natural or artificial selection replicates focused mainly on detecting parallel loci (***Burke et al., 2010***; ***Jones et al., 2012***; ***Chan et al., 2012***; ***Kelly and Hughes, 2018***), here we have the possibility to quantify parallelism and determine the selection value of a given locus. Six out of eight significant loci at the $H_{INF,\ max\ LD}$ threshold were line-specific, even though all eight selected alleles were present in the F0 generation in both lines. This prevalence of line-specific loci was consistent under different significance thresholds. However, the two remaining loci that ranked first and second by selection coefficient were parallel, both with $s > 0.3$ (***Figure 3B***; note that as outliers, the selection coefficient may be substantially overestimated, but their rank

order should remain the same), supporting the idea that the probability of parallelism can be high among those loci with the greatest selection advantage (***Orr, 2005***).

Finding just two parallel loci out of 8 discrete loci may appear to be low, given the genetic similarity in the founding generation and the identical selection applied to both Longshanks replicates. However, one should bear in mind the very many genetic paths to increasing tibia length under an infinitesimal model, and that the effect of drift is expected to be very strong in these small populations. In larger populations, the shift in the balance from drift to selection should result in selection being able to favor increasingly subtle variants and thus produce a greater proportion of parallel loci. However, we expect the trend of parallelism being enriched among the top loci to hold.

In contrast to the subtle differences within each line in changes in global diversity over 17 generations (***Figure 3.4*** and ***Figure 3.6***), we found the signature of parallelism to be significantly enriched in the comparison between the selected replicates (χ^2 test, LS1–LS2: $p \leq 1 \times 10^{-10}$), as opposed to comparisons between each selected line and Ctrl (LS1–Ctrl: $p > 0.01$ and LS2–Ctrl: $p > 0.2$, both non-significant after correcting for multiple testing), or between simulated replicates (***Figure 3.10***; see Appendix for details). Because the parallel selected loci between LS1 and LS2 have the highest selection coefficients and parallelism is not generally expected in our populations, these loci provide the strongest evidence for the role of discrete major loci. As such, the top-ranked parallel locus is the prime candidate for molecular dissection (see ***Figure 3.12*** and Appendix 'Molecular dissection of *Gli3*' for an additional *a priori* candidate locus with known limb function).

Molecular dissection of the *Nkx3-2* locus highlights *cis*-acting changes

Between the two major parallel loci, we chose the locus on chromosome 5 (Chr5) at 41–42 Mbp for functional validation because it showed the strongest estimated selection coefficient, its signature of selection was clear, and crucially for functional characterization, it contains only three genes, including *Nkx3-2* (also known as *Bapx1*), a known regulator of bone maturation (***Figure 3.4 and 3.11A***) (***Provot et al., 2006***). At this locus, the pattern of variation resembles a selective sweep spanning 1 Mbp (***Figure 3.11A***). Comparison between F0 and F17 individuals revealed no

recombinant in this entire region (***Figure 3.15A***, top panel), precluding fine-mapping using recombinants. We then analyzed the genes in this region to identify the likely target(s) of selection. First, we determined that no coding changes existed for either *Rab28* or *Nkx3-2*, the two genes located within the topologically associating domain (TAD, which mark chromosome segments with shared gene regulatory logic) (***Dixon et al., 2012***). We then performed *in situ* hybridization and detected robust expression of *Nkx3-2* and *Rab28* in the developing fore- and hind limb buds of Ctrl, LS1 and LS2 E12.5, in a domain broadly overlapping the presumptive zeugopod, the region including the tibia (***Figure 3.13B***). A third gene, *Bod1l*, straddled the TAD boundary with its promoter located in the neighboring TAD, making its regulation by sequences in the selected locus unlikely. Consistent with this, *Bod1l* showed only weak or undetectable expression in the developing limb bud (***Figure 3.13A***). We next combined ENCODE chromatin profiles and our own ATAC-Seq data to identify limb enhancers in the focal TAD. Here we found three novel enhancer candidates (N1, N2 and N3) carrying three, one and three SNPs respectively, all of which showed significant allele frequency shifts in LS1 and LS2 (***Figure 3.11B and C***; ***Figure 3.15A***). Chromosome conformation capture assays showed that the N1 and N3 sequences formed long-range looping contacts with the *Nkx3-2* promoter—a hallmark of enhancers—despite as much as 600 kbp of intervening sequence (***Figure 3.11B***). We next used transgenic reporter assays to determine whether these sequences could drive expression in the limbs. Here, we were not only interested in whether the sequence encoded enhancer activity, but specifically whether the SNPs would affect the activity (***Figure 3.11C and D***). An examination of the predicted transcription factor binding sites showed that both the N1 and N3 enhancers contain multiple SNPs with consistent directional impact on the putative enhancer activity (***Figure 3.11C***). In contrast, the N2 enhancer contains only a single SNP and is predicted to have inconsistent effect on its activity. We therefore excluded the N2 enhancer from further testing. We found that the F0 alleles of the N1 and N3 enhancers (three SNPs each in about one kbp) drove robust and consistent *lacZ* expression in the developing limb buds (N1 and N3) as well as in expanded trunk domains (N3) at E12.5 (***Figure 3.11E***). In contrast, transgenic reporters carrying the selected F17 alleles of the N1 and N3 enhancers drove consistently weak, nearly undetectable *lacZ* expression (***Figure 3.11E***). Thus, switching from the F0 to the F17 enhancer alleles led to a nearly complete loss in

activity ('loss-of-function') at developmental stage E12.5. This is consistent with the role of *Nkx3-2* as a repressor in long bone maturation (***Provot et al., 2006***). It should be noted that even though our selective regime favored an increase in the target phenotype (tibia length), at the molecular level we expect advantageous loss- and gain-of-function variants to be equally likely favored by selection. In fact, in an additional functional validation example at the *Gli3* locus, we found a gain-of-function enhancer variant that may have been favored at that locus (see ***Figure 3.12*** and Appendix 'Molecular dissection of *Gli3*').

At the *Nkx3-2* locus, we hypothesize that the F17 allele causes de-repression of bone and/or cartilage formation by reducing enhancer activity and *Nkx3-2* expression. Crucially, the F0 N1 enhancer showed activity that presages future long bone cartilage condensation in the limb (***Figure 3.11E***). That is, the observed expression pattern recalls previous results that suggest that undetected early expression of *Nkx3-2* may mark the boundaries and size of limb bone precursors, including the tibia (***Sivakamasundari et al., 2012***). Conversely, over-expression of *Nkx3-2* has been shown to cause shortened tibia (even loss) in mice (***Bren-Mattison et al., 2011***; ***Tribioli and Lufkin, 2006***). In humans, homozygous frameshift mutations in *NKX3-2* cause the rare disorder spondylo-megaepiphyseal-metaphyseal dysplasia (SMMD; OMIM: 613330), which is characterized by short-trunk, long-limbed dwarfism and bow-leggedness (***Hellemans et al., 2009***). The affected bones in SMMD patients broadly correspond to the expression domains of the two novel N1 (limbs) and N3 (limbs and trunk) enhancers. Instead of wholesale loss of *Nkx3-2* expression, which would have been lethal in mice (***Akazawa et al., 2000***) or likely cause major defects similar to SMMD patients (***Hellemans et al., 2009***), our *in situ* hybridization data did not reveal qualitative differences in *Nkx3-2* expression domains between Ctrl or LS embryos (***Figure 3.13B***). Taken together, our results recapitulate the key features of a *cis*-acting mode of adaptation: *Nkx3-2* is a broadly expressed pleiotropic transcription factor that is lethal when knocked out (***Akazawa et al., 2000***). We found no amino acid changes between the F0 and F17 alleles that could impact protein function. Rather, selection favored changes in tissue-specific expression by modular enhancers. By combining population genetics, functional genomics and developmental genetic techniques, we were able to dissect a megabase-long locus and present data supporting the identification of up to six

candidate quantitative trait nucleotides (QTNs). In mice, this represents a rare example of genetic dissection of a trait to the base-pair level.

Linking molecular mechanisms to evolutionary consequence

We next aimed to determine the evolutionary relevance of the *Nkx3-2* enhancer variants at the molecular and the population levels. At the strongly expressed N3/F0 'trunk and limb' enhancer, we note that the SNPs in the F17 selected allele lead to disrupted *Nkx3-1* and *Nkx3-2* binding sites (***Figure 3.11C and 3.14A***; UNIPROBE database [***Berger et al., 2008***]). This suggests that the selected SNPs may disrupt an auto-feedback loop to decrease *Nkx3-2* activity in the limb bud and trunk domains (***Figure 3.14A***). Using a *GFP* transgenic reporter assay in stickleback fish embryos, we found that the mouse N1/F0 enhancer allele was capable of driving expression in the distal cells but not in the fin rays of the developing fins (***Figure 3.14A***). This pattern recapitulates fin expression of *nkx3.2* in fish, which gives rise to endochondral radials (homologous to ulna/tibia in mice) (***Crotwell and Mabee, 2007***). Our results suggest that strong selection may have favored the weaker N1/F17 and N3/F17 enhancer alleles in the context of the Longshanks selection regime despite the deep functional conservation of the F0 variants.

Using theory and simulations, we went beyond qualitative molecular dissection to quantitatively estimate the selection coefficient at the *Nkx3-2* locus and its contribution to the total selection response in the Longshanks mice. We retraced the selective sweep of the *Nkx3-2* N1 and N3 alleles through targeted genotyping in 1569 mice across all 20 generations. The selected allele steadily increased from around 0.17 to 0.85 in LS1 and 0.98 in LS2 but fluctuated around 0.25 in Ctrl (***Figure 3.14B***). We estimated that such a change of around 0.7 in allele frequency would correspond to a selection coefficient *s* of ~0.24 ± 0.12 at this locus (***Figure 3.15B***; see Appendix section on '*Estimating selection coefficient of the top-ranking locus, Nkx3-2, from changes in allele frequency*'). By extending our simulation framework to allow for a major locus against an infinitesimal background, we find that the *Nkx3-2* locus would contribute 9.4% of the total selection response (limits 3.6–15.5%; see Appendix section '*Estimating selection coefficient*' for details) in order to produce a shift of 0.7 in allele frequency over 17 generations. To avoid inflation stemming from estimating from outliers, we also independently estimated the contribution of the

Nkx3-2 locus using a linear mixed animal model based on the full genotyped series mentioned above (see Appendix section '*Estimating selection coefficient, animal model*' for details). Using this alternative approach, we estimated that each selected allele increases tibia length by 0.36% (*N* = 1569, 95% conf. int.: 0.07–0.64%, p=0.0171). Multiplying the effect with the increase in the allele frequency suggests that the *Nkx3-2* locus alone would account for approximately 4% of the overall 12.9% increase in tibia length. This lower estimate of around 4% is nonetheless within the bounds of the estimate from simulations. Together, both approaches indicate that the *Nkx3-2* locus contributes substantially to the selection response.

Discussion

A defining task of our time is to understand the factors that determine and constrain how small populations respond to sudden environmental changes. Here, we analyze the replicated and controlled Longshanks experiment to characterize the genomic changes that occur as small experimental populations respond to selection.

An important conclusion from the Longshanks experiment is that selection response can be steady and robust even in extremely bottlenecked populations. That is, we found that tibia length increased readily and repeatedly in response to selection even with as few as 14–16 breeding pairs per generation. The sustained response was possible because the lines were founded with enough standing variation, and generation 17 was still only a fraction of the way to the expected limit for the selection response at ~$2N_e$ generations (***Robertson, 1960***), estimated here to be around 90 (see legend for ***Figure 3.3B***; Appendix on 'Estimating the selection coefficient'). Although other selective breeding studies using a similar base population of mice encountered selection limits at around generation 20–25 (possibly due to countervailing selection rather than loss of genetic variance) for high voluntary wheel running behavior (***Careau et al., 2013***) and for nest-building behavior (***Bult and Lynch, 2000***), here all evidence suggests that the Longshanks mice should continue to show increases in tibia length for many more generations.

The estimated N_e of 46 in the Longshanks experiment, while small, is comparable to those in natural populations like the Soay sheep (***McRae et al., 2005***), Darwin's

finches (***Grant and Grant, 1992***) or Tasmanian devils (***Epstein et al., 2016***) (this last study documents a rapid and parallel evolutionary response to transmissible tumors). These populations span a wide range of time in sustained bottlenecking, from the most recent in Tasmanian devils, to likely many millions of years in Darwin's finches. Accordingly, we also expect very different dynamics during short- vs. long-term selection response: for a short bout of selection, such as the 20 generations analyzed in this study, selection response depends overwhelmingly on standing genetic variation, with little to no contribution from *de novo* mutations (***Hill, 1982***; ***Weber and Diggins, 1990***). Over the long term, however, *de novo* mutations would contribute increasingly to selection response. In the Longshanks experiment, we observe a robust early response to selection (***Figure 3.1B and Figure 3.2***), and a gradual decrease in sequence diversity, consistent with the effect of drift (***Figure 3.3B*** and ***Figure 3.5A***, ***Supplementary file 2***). There has long been broad empirical support for adaptation from standing genetic variation in nature (***Jones et al., 2012***; ***Epstein et al., 2016***; ***Hancock et al., 2011***) and breeding (***Sheng et al., 2015***). At least in the short-run, our result demonstrating robust selection response in the Longshanks experiment provides grounds for some optimism regarding the ability of populations to respond rapidly to changes in their environment.

By combining pedigree records with sequencing of founder individuals, our data had sufficient detail to allow precise modeling of trait response, with predicted shifts in allele frequency distribution that closely matched our results (e.g. ***Figure 3.3D***). Furthermore, we functionally validated loci that showed allele frequency shifts outside the model's predictions and found key enhancers of major effect. Connecting trait changes to allele frequency changes at specific loci has been a longstanding objective in selection experiments, with a number of notable early attempts (e.g., ***Keightley et al., 1996***). To date, we know of only a few studies that attempt to explicitly link traits with changes in allele frequencies (***Kessner and Novembre, 2015***; ***Rice and Townsend, 2012***; ***Chen et al., 2019***; ***Nuzhdin et al., 1999***) and none have systematically tested the underlying architecture against an infinitesimal background. Here, our results imply a mixed genetic architecture with a few discrete loci of large effect amid an infinitesimal background. It remains to be seen whether other evolve-and-resequence (E&R) studies, with different selection pressures and population parameters, may reveal similar results.

To put our finding of a mixed genetic architecture into perspective, it is worth noting that the infinitesimal model is still the most predictive model by far in practical quantitative genetics, for diverse domesticated species from cattle to crops, despite its intrinsically unrealistic assumptions (***Hill et al., 2008***; ***Lynch and Walsh, 1998***; ***Hill and Zhang, 2012***). In general, current genomic data for many traits is consistent with a very large number of loci, each with a small effect. From a practical point of view, however, the use of an infinitesimal model does not preclude the presence or indeed the importance of a few major effect loci. Rather, it simply assumes that they are rare enough to allow reasonable model fit (***Walsh and Lynch, 2018***, page 878). Here, we note that it is actually not clear how one might parameterize a generally applicable predictive oligogenic model with more than a single major effect locus. In this study, while we consider the most likely genetic architecture underlying selection response for tibia length to be a small number of major effect loci together with a polygenic background, we cannot reject other alternative models that could also account for the observed response, such as an effectively infinitesimal model with linkage, as well as models with a few major trait loci.

Among other classical examples of complex traits, such as height or body weight, that may have been subjected to selection, we observe a range of genetic architectures in ways often tightly connected to their population size and/or selection history. Height in humans is often cited as the classical complex trait under possible selection of unknown (and much debated) intensity (see ***Turchin et al., 2012***; ***Berg and Coop, 2014***; ***Barton et al., 2019***). It shows high heritability and a highly dispersed genetic architecture (with the top-ranked locus accounting for only 0.8% of the variation explained in cosmopolitan European populations) (***Weedon et al., 2007***; ***Wood et al., 2014***). In contrast, as few as 4 to 6 loci account for 83% and 50% of the variation in height in horses and dogs, respectively (***Makvandi-Nejad et al., 2012***; ***Rimbault et al., 2013***). In both horses and dogs, selection has been strong and sustained, and breed-specific populations tend to be small. Interestingly, and in line with our experiment, the major allele at the *IGF1* locus stems from a standing genetic variant, despite many factors that may theoretically favor large-effect *de novo* mutations (***Sutter et al., 2007***). In chickens, modern breeding practice and selection from large populations yielded a highly polygenic genetic architecture for body weight, with some of the best empirical evidence for epistasis (***Carlborg et al.,***

2006; ***Wahlberg et al., 2009***; ***Rubin et al., 2010***; ***Pettersson et al., 2013***). Similarly, results from many selection experiments in *Drosophila* suggest that the genetic architecture underlying selection response may involve many genes (***Jha et al., 2015***; ***Reeves and Tautz, 2017***; ***Orozco-terWengel et al., 2012***; ***Turner et al., 2011***). By contrast, the extreme tail of the effect size distribution (as inferred from Δz^2) from the Longshanks experiment appears to account for a substantial part of the selection response, presumably due to the combined effects of relatively low diversity in commercial mouse stocks and the small founding populations. But unlike these previous QTL studies or selection experiments, in which either the genetic architecture of a trait or the selection value were estimated separately, sometimes from only few parental individuals or lines, E&R studies sample a much broader pool of alleles and continually compete them against each other. Thus, our approach allowed simultaneous inference of genetic architecture and distribution of effect sizes, is more likely to be representative of the population at large, and is more akin to genome-wide association studies (GWAS), except that here we can also directly connect a trait to its selective value and capture the trajectory of any given allele.

Parallel evolution is often seen as a hallmark for detecting selection (***Chan et al., 2012***; ***Schluter et al., 2004***; ***Chan et al., 2010***; ***Martin and Orgogozo, 2013***). We investigated the factors that contribute to parallelism in allele frequency shifts over 17 generations by contrasting the two Longshanks replicates against the Control line. However, we observed little parallelism between selected lines and Ctrl, or between simulated replicates under selection, even though the simulated haplotypes were sampled directly from actual founders. This underscores that parallelism depends on both shared selection pressure (absent in Ctrl) and the availability of large-effect loci that confer a substantial selection advantage (absent under the infinitesimal model; ***Figure 3.9; Figure 3.10***). With increasing population size, selection would be better able to detect variants with more subtle effects. This would in turn lower the threshold beyond which the selective advantage of an allele would become deterministic, that is, exhibit parallelism.

Through in-depth dissection of the *Nkx3-2* locus, our data show in fine detail how the selective value of standing variants depends strongly on the selection regime: the originally common F0 variant of the N1 enhancer shows deep functional

conservation and can evidently recapitulate fin *nkx3.2* expression in fishes (***Figure 3.14A***). Yet, in the Longshanks experiment selection strongly favored the weaker allele (***Figure 3.14B***). In fact, our molecular dissection of two loci show that both gain-of-function (*Gli3*) and loss-of-function (*Nkx3-2*) variants could be favored by selection (***Figure 3.11E and 3.14A***; ***Figure 3.12D***). Through synthesis of multiple lines of evidence, our work uncovered the key role of *Nkx3-2*, which was not an obvious candidate gene like *Gli3* due to the lack of abnormal limb phenotype in *Nkx3-2* knockout mice. To our surprise, the same loss of *NKX3-2* function in human SMMD patients manifests in opposite ways in different bone types as short trunk and long limbs (***Hellemans et al., 2009***). This matches the expression domains of our N1 (limb) and N3 (limb and trunk) enhancers (***Figure 3.14A***). Evidently, in the absence of lethal coding mutations, the F17 haplotype was doubly beneficial at both enhancers for the limb and potentially also trunk target tissues under the novel selection regime in the Longshanks selection experiment. We estimate that these enhancer variants, along with any other tightly linked beneficial SNPs, segregate as a single locus, which in turn contributes ~10% of the overall selection response.

Despite our efforts to uncover the mechanism underlying the selective advantage of the *Nkx3-2* locus, much remains unknown. For example, it remains unclear how such a major allele could segregate in the general mouse stock (and as the reference C57BL/6J allele, no less). It could be that this allele has the same effect in the general mouse population but is conditionally neutral under non-selective breeding and simply escaped notice. However, our preliminary exploration in a panel of C57BL/6-by-DBA/2 ('BXD') mice suggested otherwise: mapping of tibia length or mineral density did not reveal this locus as a major QTL determining tibia length (unpublished data kindly provided by Weikuan Gu), suggesting that this allele's effect on tibia length may depend on the genetic background. Alternatively, the broader C57BL/6 allele could be linked to a compensatory mutation that became uncoupled among the founders of the Longshanks lines. Finally, although we do observe the specific N1 and N3 SNP positions as variable across the rodent and indeed the broader mammalian lineages, further work is needed to determine their effect, if any, on limb development.

Conclusion

Using the Longshanks selection experiment and synthesizing theory, empirical data and molecular genetics, we show that it is possible to identify some of the individual SNPs that have contributed to the response to selection on morphology. In particular, discrete, large-effect loci are revealed by their parallel response. Further work should focus on dissecting the mechanisms behind the dynamics of selective sweeps and/or polygenic adaptation by re-sequencing the entire selection pedigree, testing how the selection response depends on the genetic architecture, and the extent to which linkage places a fundamental limit on our inference of selection. Improved understanding in these areas may have broad implications for conservation, rapid adaptation to climate change and quantitative genetics in medicine, agriculture and in nature.

Materials and methods

Animal care and use

All experimental procedures described in this study have been approved by the applicable University institutional ethics committee for animal welfare at the University of Calgary (HSACC Protocols M08146 and AC13-0077); or local competent authority: Landesdirektion Sachsen, Germany, permit number 24–9168.11-9/2012-5.

Reference genome assembly

All co-ordinates in the mouse genome refer to *Mus musculus* reference mm10, which is derived from GRCm38.

Code and data availability

Sequence data have been deposited in the SRA database under accession number SRP165718 and GEO under GSE121564, GSE121565 and GSE121566. Non-sequence data have been deposited at Dryad, doi:10.5061/dryad.0q2h6tk. Analytical code and additional notes have been deposited in the following repository: https://github.com/evolgenomics/Longshanks (***Evolgenomics, 2019***; copy archived at https://github.com/elifesciences-publications/Longshanks). Additional raw data and code are hosted via our institute's FTP servers at http://ftp.tuebingen.mpg.de/fml/ag-chan/Longshanks/.

Pedigree data

Tibia length and body weight phenotypes were measured as previously described (***Marchini et al., 2014***). A total of 1332 Control, 3054 LS1, and 3101 LS2 individuals were recorded. Five outlier individuals with a skeletal dysplasia of unknown etiology were removed from LS2 and excluded from further analysis. Missing data in LS2 were filled in with random individuals that best matched the pedigree. Trait data were analyzed to determine response to selection based on the measured traits and their rank orders based on the selection index.

Simulations
Simulations were based on the actual pedigree and selection scheme, following one chromosome at a time. Each chromosome was represented by a set of junctions, which recorded the boundaries between genomes originating from different founder genomes; at the end, the SNP genotype was reconstructed by seeding each block of genome with the appropriate ancestral haplotype. This procedure is much more efficient than following each of the very large number of SNP markers. Crossovers were uniformly distributed, at a rate equal to the map length (***Cox et al., 2009***). Trait value was determined by a component due to an infinitesimal background (V_g); a component determined by the sum of effects of 104 evenly spaced discrete loci (V_s); and a Gaussian non-genetic component (V_e). The two genetic components had variance proportional to the corresponding map length, and the heritability was estimated from the observed trait values (see Appendix section 'Major considerations'). In each generation, the actual number of male and female offspring were generated from each breeding pair, and the male and female with the largest trait value were chosen to breed.

SNP genotypes were assigned to the founder genomes with their observed frequencies. However, to reproduce the correct variability requires that we assign founder haplotypes. This is not straightforward, because low-coverage individual genotypes cannot be phased reliably, and heterozygotes are frequently mis-called as homozygotes. We compared three procedures, which were applied within intervals that share the same ancestry: assigning haplotypes in linkage equilibrium (LE, or 'no LD'); assigning the two alleles at heterozygous sites in each individual to its two haplotypes at random, which minimizes linkage disequilibrium but is consistent with observed diploid genotypes ('min LD'); and assigning alleles at heterozygous sites in each individual to the 'reference' and 'alternate' haplotype consistently within an interval, which maximizes linkage disequilibrium ('max LD') (***Figure 3.3C***). For details, see legend in ***Figure 3.3***.

Significance thresholds
To obtain significance thresholds, we summarized the genome-wide maximum Δz^2 shift for each replicate of the simulated LS1 and LS2 lines, averaged within 10 kb windows, and grouped by the selection intensity and extent of linkage disequilibrium (LD). From this distribution of genome-wide maximum Δz^2 we obtained the critical value for the corresponding significance threshold (typically the 95^{th} quantile or p=0.05) under each selection and LD model (***Figure 3.9A***; ***Figure 3.3E***). This procedure controls for the effect of linkage and hitchhiking, line-specific pedigree structure, and selection strength.

Sequencing, genotyping and phasing pipeline
Sequencing libraries for high-throughput sequencing were generated using TruSeq or Nextera DNA Library Prep Kits (Illumina, Inc, San Diego, USA) according to manufacturer's recommendations or using equivalent *Tn5* transposase expressed in-house as previously described (***Picelli et al., 2014***). Briefly, genomic DNA was extracted from ear clips by standard Protease K digestion (New England Biolabs GmbH, Frankfurt am Main, Germany) followed by AmpureXP bead (Beckman Coulter GmbH, Krefeld, Germany) purification. Extracted high-molecular weight DNA was sheared with a Covaris S2 (Woburn, MA, USA) or 'tagmented' by commercial or purified *Tn5*-transposase according to manufacturer's recommendations. Each

sample was individually barcoded (single-indexed as N501 with N7XX variable barcodes; all oligonucleotides used in this study were synthesized by Integrated DNA Technologies, Coralville, Iowa, USA) and pooled for high-throughput sequencing by a HiSeq 3000 (Illumina) at the Genome Core Facility at the MPI Tübingen Campus. Sequenced data were pre-processed using a pipeline consisting of data clean-up, mapping, base-calling and analysis from software fastQC v0.10.1 (***Andrews, 2016***); trimmomatic v0.33 (***Bolger et al., 2014***); bwa v0.7.10-r789 (***Li and Durbin, 2010***); GATK v3.4–0-gf196186 modules BQSR, MarkDuplicates, IndelRealignment (***McKenna et al., 2010***; ***DePristo et al., 2011***). Genotype calls were performed using the GATK HaplotypeCaller under the GENOTYPE_GIVEN_ALLELES mode using a set of high-quality SNP calls made available by the Wellcome Trust Sanger Centre (Mouse Genomes Project version three dbSNP v137 release [***Keane et al., 2011***]), after filtering for sites segregating among inbred lines that may have contributed to the original seven female and two male CD-1 founders, namely 129S1/SvImJ, AKR/J, BALB/cJ, BTBR T^{+}*Itpr3* ^{tf}/J, C3H/HeJ, C57BL/6NJ, CAST/EiJ, DBA/2J, FVB/NJ, KK/HiJ, MOLF/EiJ, NOD/ShiLtJ, NZO/HlLtJ, NZW/LacJ, PWK/PhJ and WSB/EiJ based on (***Yalcin et al., 2010***). We consider a combined ~100x coverage sufficient to recover any of the 18 CD-1 founding haplotypes still segregating at a given locus. The raw genotypes were phased with Beagle v4.1 (***Browning and Browning, 2016***) based on genotype posterior likelihoods using a genetic map interpolated from the mouse reference map (***Cox et al., 2009***) and imputed from the same putative CD-1 source lines as the reference panel. The site frequency spectra (SFS) were evaluated to ensure genotype quality (***Figure 3.5A***).

Population genetics summary statistics

Summary statistics of the F0 and F17 samples were calculated genome-wide (Weir–Cockerham F_{ST}, π, heterozygosity, allele frequencies *p* and *q*) in adjacent 10 kbp windows or on a per site basis using VCFtools v0.1.14 (***Danecek et al., 2011***). The summary statistic Δz^2 was the squared within-line difference in arcsine square-root transformed MAF *q*; it ranges from 0 to π^2. The resulting data were further processed by custom bash, Perl and R v3.2.0 (***R Development Core Team, 2015***) scripts.

Peak loci and filtering for hitchhiking windows

Peak loci were defined by a descending rank ordering of all 10 kbp windows, and from each peak signal the windows were extended by 100 SNPs to each side, until no single SNP rising above a Δz^2 shift of 0.2 π^2 was detected. A total of 810 peaks were found with a Δz^2 shift ≥0.2 for LS1 and LS2. Following the same procedure, we found 766 peaks in Ctrl.

Candidate genes

To determine whether genes with related developmental roles were associated with the selected variants, the topologically associating domains (TADs) derived from mouse embryonic stem cells as defined elsewhere (***Dixon et al., 2012***) were re-mapped onto mm10 co-ordinates. Genes within the TAD overlapping within 500 kbp of the peak window ('core span') were then cross-referenced against annotated knockout phenotypes (Mouse Genome Informatics, http://www.informatics.jax.org). This broader overlap was chosen to include genes whose regulatory sequences (e.g., enhancers), but not necessarily their gene bodies, fall close to the peak window. We highlight candidate genes showing limb- and bone-related phenotypes,

e.g., with altered limb bone lengths or epiphyseal growth plate morphology, as observed in Longshanks mice (***Marchini and Rolian, 2018***), of the following categories (along with their Mammalian Phenotype Ontology term and the number of genes): 'abnormal tibia morphology/MP:0000558' (212 genes), 'short limbs/MP:0000547' and 'short tibia/MP:0002764' (223 genes), 'abnormal cartilage morphology/MP:0000163' (321 genes), 'abnormal osteoblast morphology/MP:0004986' (122 genes). Note that we excluded compound mutants or those conditional mutant phenotypes involving transgenes. To determine if the overlap with these genes was significant, we performed 1000 permutations of the core span using bedtools v2.22.1 shuffle with the -noOverlapping option (***Quinlan and Hall, 2010***) and excluding ChrY, ChrM and unassembled scaffolds. We then followed the exact procedure as above to determine the number of genes in the overlapping TAD belonging to each category. We reported the quantile rank as the *P*-value, ignoring ties. To determine other genes in the region, we list all genes falling within the entire hitchhiking window (***Supplementary file 3***).

Identification of putative limb enhancers

We downloaded publicly available chromatin profiles, derived from E14.5 limbs, for the histone H3 lysine-4 (K4) or lysine-27 (K27) mono-/tri-methylation or acetylation marks (H3K4me1, H3K4me3 and H3K27ac) generated by the ENCODE Consortium (***Shen et al., 2012***). We intersected the peak calls for the enhancer-associated marks H3K4me1 and H3K27ac and filtered out peaks overlapping promoters (H3K4me3 and promoter annotation according to the FANTOM5 Consortium [***Forrest et al., 2014***]).

Enrichment analysis

To calculate enrichment through the whole range of Δz^2, a similar procedure was taken as in Candidate genes above. For knockout gene functions, genes contained in TADs within 500 kbp of peak windows were included in the analysis. We used the complete database of annotated knockout phenotypes for genes or spontaneous mutations, after removing phenotypes reported under conditional or polygenic mutants. For gene expression data, we retained all genes which have been reported as being expressed in any of the limb structures, by tracing each anatomy ontological term through its parent terms, up to the top-level groupings, e.g., 'limb', in the Mouse Genomic Informatics Gene Expression Database (***Finger et al., 2017***). For E14.5 enhancers, we used a raw 500 kbp overlap with the peak windows because enhancers, unlike genes, may not have intermediaries and may instead represent direct selection targets.

For coding mutations, we first annotated all SNPs for their putative effects using snpEff v4.0e (***Cingolani et al., 2012***). To accurately capture the per-site impact of coding mutations, we used per-site Δz^2 instead of the averaged 10 kbp window. For each population, we divided all segregating SNPs into up to 0.02 bands based on per-site Δz^2. We then tracked the impact of coding mutations in genes known to be expressed in limbs, as above. We reported the sum of all missense ('moderate' impact), frame-shift, stop codon gain or loss sites ('high impact'). A linear regression was used to evaluate the relationship between Δz^2 and the average impact of coding SNPs (SNPs with high or moderate impact to all coding SNPs).

For regulatory mutations, we used the same bins spanning the range of Δz^2, but focused on the subset of SNPs falling within the ENCODE E14.5 limb enhancers. We then obtained a weighted average conservation score based on an averaged phastCons (***Pollard et al., 2010***) or phyloP (***Siepel et al., 2005***) score in ±250 bp flanking the SNP, calculated from a 60-way alignment between placental mammal genomes (downloaded from the UCSC Genome Browser [***Kent et al., 2002***]). We reported the average conservation score of all SNPs within the bin and fitted a linear regression on log-scale. In particular, phastCons scores range from 0 (un-conserved) to 1 (fully conserved), whereas phyloP is the $||/log_{10}||$ of the *P*-value of the phylogenetic tree, expressed as a positive score for conservation and a negative score for lineage-specific accelerated change. We favored using phastCons for its simpler interpretation.

Impact of coding variants

Using the same SNP effect annotations described in the section above, we checked whether any specific SNP with significant site-wise Δz^2 in either LS1 or LS2 cause amino acid changes or protein disruptions and are known to cause limb defects when knocked out. For each position we examined outgroup sequences using the 60-way placental mammal alignment to determine the ancestral amino acid state and whether the selected variant was consistent with purifying vs. diversifying selection. The resulting 12 genes that matched these criteria are listed in ***Supplementary file 4***.

Association with human height loci

We downloaded the set of 697 SNPs associated with loci for human height (***Wood et al., 2014***) to test if these loci cluster with the selected loci in the Longshanks lines. In order to facilitate mapping to mouse co-ordinates, each human SNP was expanded to 100 kbp centering on the SNP and converted to mm10 positions using the liftOver tool with the multiple mapping option disabled (***Kent et al., 2002***). We were able to assign positions in 655 out of the 697 total SNPs. Then for each of the 810 loci above the $H_{INF,\ no\ LD}$ threshold in the selected Longshanks lines, the minimal distance to any of the mapped human loci was determined using bedtools closest with the -d option (***Quinlan and Hall, 2010***). When a region actually overlapped, a distance of 0 bp was assigned. To generate a permuted set, the 810 loci were randomly shuffled across the mouse autosomes using the bedtools shuffle program with the -noOverlapping option. Then the exact same procedure as the actual data was followed to determine the closest interval. The resulting permuted intervals followed an approximately normal distribution, with observed results falling completely below the range of permuted results, that is, closer to height-associated human SNPs.

In situ **hybridization**

Detection of specific gene transcripts were performed as previously described in ***Brown et al., 2005***. Probes against *Nkx3-2*, *Rab28*, *Bod1l* and *Gli3* were amplified from cDNA from wildtype C57BL/6NJ mouse embryos (***Supplementary file 5***). Amplified fragments were cloned into pJET1.2/blunt plasmid backbones in both sense and anti-sense orientations using the CloneJET PCR Kit (Thermo Fisher Scientific, Schwerte, Germany) and confirmed by Sanger sequencing using the included forward and reverse primers. Probe plasmids have also been deposited with Addgene. *In vitro* transcription from the T7 promoter was performed using the MAXIscript T7 *in vitro* Transcription Kit (Thermo Fisher Scientific) supplemented with

Digoxigenin-11-UTP (Sigma-Aldrich) (MPI Tübingen), or with T7 RNA polymerase (Promega) in the presence of DIG RNA labeling mix (Roche) (University of Calgary). Following TURBO DNase (Thermo Fisher Scientific) digestion, probes were cleaned using SigmaSpin Sequencing Reaction Clean-Up columns (Sigma-Aldrich) (MPI Tübingen), or using Illustra MicroSpin G-50 columns (GE Healthcare) (University of Calgary). During testing of probe designs, sense controls were used in parallel reactions to establish background non-specific binding.

ATAC-seq library preparation and sequencing pipeline

ATAC-seq was performed on dissected C57BL/6NJ E14.5 forelimb and hindlimb. Nuclei preparation and tagmentation were performed as previously described in ***Buenrostro et al. (2013)***, with the following modifications. To minimize endogenous protease activity, cells were strictly limited to 5 + 5 min of collagenase A treatment at 37°C, with frequent pipetting to aid dissociation into single-cell suspensions. Following wash steps and cell lysis, 50,000 nuclei were tagmented with expressed *Tn5* transposase. Each tagmented sample was then purified by MinElute columns (Qiagen) and amplified with Q5 High-Fidelity DNA Polymerase (New England Biolabs) using a uniquely barcoded i7-index primer (N701-N7XX) and the N501 i5-index primer. PCR thermocycler programs were 72°C for 4 min, 98°C for 30 s, 6 cycles of 98°C for 10 s, 65°C for 30 s, 72°C for 1 min, and final extension at 72°C for 4 min. PCR-enriched samples were taken through a double size selection with PEG-based SPRI beads (Beckman Coulter) first with 0.5X ratio of PEG/beads to remove DNA fragments longer than 600 bp, followed by 1.8X PEG/beads ratio in order to select for Fraction A as described in ***Milani et al. (2016***). Pooled libraries were run on the HiSeq 3000 (Illumina) at the Genome Core Facility at the MPI Tübingen Campus to obtain 150 bp paired end reads, which were aligned to mouse mm10 genome using bowtie2 v.2.1.0 (***Langmead and Salzberg, 2012***). Peaks were called using MACS14 v.2.1 (***Zhang et al., 2008***).

Multiplexed chromosome conformation capture (4C-Seq)

Chromosome conformation capture (3C) template was prepared from pooled E14.5 liver, forelimb and hindlimb buds (n = 5–6 C57BL/6NJ embryos per replicate), with improvements to the primer extension and library amplification steps following (***Sexton et al., 2012***). The template was amplified with Q5 High-Fidelity Polymerase (New England Biolabs GmbH, Frankfurt am Main, Germany) using a 4C adapter-specific primer and a pool of 6 *Nkx3-2* enhancer viewpoint primers (and, in a separate experiment, a pool of 8 *Gli3* enhancer-specific viewpoint primers; ***Supplementary file 6***). Amplified fragments were prepared for Illumina sequencing by ligation of TruSeq adapters, followed by PCR enrichment. Pooled libraries were sequenced by a HiSeq 3000 (Illumina) at the Genome Core Facility at the MPI Tübingen Campus with single-end, 150 bp reads. Sequence data were processed using a pipeline consisting of data clean-up, mapping, and analysis based upon cutadapt v1.10 (***Martin, 2011***); bwa v0.7.10-r789 (***Li and Durbin, 2010***); samtools v1.2 (***Li et al., 2009***); bedtools (***Quinlan and Hall, 2010***) and R v3.2.0 (***R Development Core Team, 2015***). Alignments were filtered for ENCODE blacklisted regions (***ENCODE Project Consortium, 2012***) and those with MAPQ scores below 30 were excluded from analysis. Filtered alignments were binned into genome-wide *BglII* fragments, normalized to Reads Per Kilobase of transcript per Million mapped reads (RPKM), and plotted and visualized in R.

Plasmid construction
Putative limb enhancers corresponding to the F0 and F17 alleles of the *Gli3* G2 and *Nkx3-2* N1 and N3 enhancers were amplified from genomic DNA of Longshanks mice from the LS1 F0 (nine mice) and F17 (10 mice) generations and sub-cloned into pJET1.2/blunt plasmid backbone using the CloneJET PCR Kit (Thermo Fisher Scientific) and alleles were confirmed by Sanger sequencing using the included forward and reverse primers (***Supplementary file 7***). Each allele of each enhancer was then cloned as tandem duplicates with junction *SalI* and *XhoI* sites upstream of a β-globin minimal promoter in our reporter vector (see below). Constructs were screened for the enhancer variant using Sanger sequencing. All SNPs were further confirmed against the rest of the population through direct amplicon sequencing.

The base reporter construct pBeta-lacZ-attBx2 consists of a β-globin minimal promoter followed by a l*acZ* reporter gene derived from pRS16, with the entire reporter cassette flanked by double *attB* sites. The pBeta-lacZ-attBx2 plasmid and its full sequence have been deposited and is available at Addgene.

Pronuclear injection of F0 and F17 enhancer-reporter constructs in mice
The reporter constructs containing the appropriate allele of each of the three enhancers were linearized with *ScaI* (or *BsaI* in the case of the N3 F0 allele due to the gain of a *ScaI* site) and purified. Microinjection into mouse zygotes was performed essentially as described (***DiLeone et al., 2000***).

At 12 d after the embryo transfer, the gestation was terminated and embryos were individually dissected, fixed in 4% paraformaldehyde for 45 min and stored in PBS. All manipulations were performed by R.N. or under R.N.'s supervision at the Transgenic Core Facility at the Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. Yolk sacs from embryos were separately collected for genotyping and all embryos were stained for *lacZ* expression as previously described (***Mortlock et al., 2003***). Embryos were scored for *lacZ* staining, with positive expression assigned if the pattern was consistently observed in at least two embryos.

Genotyping of time series at the *Nkx3-2* N3 locus
Allele-specific primers terminating on SNPs that discriminate between the F0 from the F17 N3 enhancer alleles were designed (rs33219710 and rs33600994; ***Supplementary file 8***). The amplicons were optimized as a qPCR reaction to give allele-specific, present/absent amplifications (typically no amplification for the absent allele, otherwise average ΔCt >10). Genotyping on the entire breeding pedigree of LS1 (n = 602), LS2 (n = 579) and Ctrl (n = 389) was performed in duplicates for each allele on a Bio-Rad CFX384 Touch instrument (Bio-Rad Laboratories GmbH, Munich, Germany) with SYBR Select Master Mix for CFX (Thermo Fisher Scientific) and the following qPCR program: 50°C for 2 min, 95°C for 2 min, 40 cycles of 95°C for 15 s, 58°C for 10 s, 72°C for 10 s. In each qPCR run we included individuals of each genotype (LS F17 selected homozygotes, heterozygotes and F0 major allele homozygotes). For the few samples with discordant results between replicates, DNA was re-extracted and re-genotyped or otherwise excluded.

Transgenic reporter assays in stickleback fish

In sticklebacks, transgenic reporter assays were carried out using the reporter construct pBHR (***Chan et al., 2010***). The reporter consists of a zebrafish *heat shock protein 70* (*Hsp70*) promoter followed by an *eGFP* reporter gene, with the entire reporter cassette flanked by *tol2* transposon sequences for transposase-directed genomic integration. The *Nkx3-2* N1/F0 enhancer allele was cloned as tandem duplicates using the *NheI* and *EcoRV* restriction sites upstream of the *Hsp70* promoter. Enhancer orientation and sequence was confirmed by Sanger sequencing. Transient transgenic stickleback embryos were generated by co-microinjecting the plasmid (final concentration: 10 ng/µl) and *tol2* transposase mRNA (40 ng/µl) into freshly fertilized eggs at the one-cell stage as described in ***Chan et al. (2010)***.

Acknowledgements

We thank Felicity Jones for input into experimental design, helpful discussion and improving the manuscript. We thank the Rolian, Barton, Chan and Jones Labs members for support, insightful scientific discussion and improving the manuscript. We thank the Rolian lab members, the Animal Resource Centre staff at the University of Calgary, and MPI Dresden Animal Facility staff for animal husbandry. We thank Derek Lundberg for help with library preparation automation. We thank Christa Lanz, Rebecca Schwab and Ilja Bezrukov for assistance with high-throughput sequencing and associated data processing; Andre Noll for high-performance computing support; the MPI Tübingen IT team for computational support. We thank Felicity Jones and the Jones Lab for help with stickleback microinjections. pRS16 was a gift from François Spitz. We thank Mirna Marinič for creating an earlier version of the transgenic reporter plasmid. We are indebted to Gemma Puixeu Sala, William G Hill, Peter Keightley for input and discussion on data analysis and simulation. We are also indebted to Stefan Mundlos, Przemko Tylzanowki, Weikuan Gu for suggested experiments and sharing unpublished data. We thank Sean B Carroll, Andrew Clark, John Kelly (reviewer), David Kingsley, Jonathan Pritchard, Matthew Rockman (reviewer), Gregory Wray, and Magnus Nordborg (reviewer) for thoughtful input that has greatly improved our manuscript. JPLC is supported by the International Max Planck Research School 'From Molecules to Organisms'. SB and NB are supported by IST Austria. CR is supported by Discovery Grant #4181932 from the Natural Sciences and Engineering Research Council of Canada and by the Faculty of Veterinary Medicine at the University of Calgary. YFC is supported by the Max Planck Society.

Additional information

Funding

Funder	Grant reference number	Author
Natural Sciences and Engineering Research Council of Canada	4181932	Campbell Rolian

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

João PL Castro, Data curation, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Michelle N Yancoskie, Data curation, Formal analysis, Validation, Investigation, Methodology, Writing—original draft, Writing—review and editing; Marta Marchini, Resources, Data curation, Methodology, Writing—review and editing; Stefanie Belohlavy, Software, Formal analysis, Methodology, Writing—review and editing; Layla Hiramatsu, Formal analysis, Validation, Visualization, Writing—review and editing; Marek Kučka, Resources, Data curation, Formal analysis, Validation, Investigation, Methodology, Writing—review and editing; William H Beluch, Resources, Data curation, Validation, Writing—review and editing; Ronald Naumann, Resources, Investigation, Methodology, Writing—review and editing; Isabella Skuplik, Formal analysis, Investigation, Visualization, Methodology, Writing—review and editing; John Cobb, Supervision, Investigation, Methodology, Writing—review and editing; Nicholas H Barton, Conceptualization, Data curation, Software, Formal analysis, Supervision, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Campbell Rolian, Conceptualization, Resources, Data curation, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing; Yingguang Frank Chan, Conceptualization, Resources, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Layla Hiramatsu https://orcid.org/0000-0001-6298-6109
John Cobb https://orcid.org/0000-0002-1053-2604
Nicholas H Barton https://orcid.org/0000-0002-8548-5240
Campbell Rolian https://orcid.org/0000-0002-7242-342X
Yingguang Frank Chan https://orcid.org/0000-0001-6292-9681

Ethics

Animal experimentation: All experimental procedures described in this study have been approved by the applicable University institutional ethics committee for animal welfare at the University of Calgary (HSACC Protocols M08146 and AC13-0077); or local competent authority: Landesdirektion Sachsen, Germany, permit number 24-9168.11-9/2012-5.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.42014.048
Author response https://doi.org/10.7554/eLife.42014.049

Additional files

Supplementary files

- Supplementary file 1. Sequencing Summary. For each line and generation, we individually barcoded all available individuals and pooled for sequencing. We aimed for a sequencing depth of around 100x coverage for 50–64

haplotypes per sample. Since the CD-1 mice were founded by an original import of 7 inbred females and two inbred males, we expect a maximum of 18 segregating haplotypes at any given locus. This sequencing design should give sufficient coverage to recover allele frequencies and possibly genotypes genome-wide. Our successful genome-wide imputation results validated this strategy.

DOI: https://doi.org/10.7554/eLife.42014.019

- Supplementary file 2. Pairwise F_{ST} and segregating sites (S) between populations. As expected, there is a general trend of decrease in diversity after 17 generations of breeding. Globally, there was a 13% decrease in diversity, but F17 populations still retained on average ~5.8M segregating SNPs (diagonal). There was very little population differentiation, as indicated by low F_{ST} among the three founder populations, however F_{ST} increases by at least 100-fold among lines by generation F17 (above diagonal, orange boxes). Within-line F_{ST} is intermediate in this respect, reaching about half of the differentiation observed between lines.

DOI: https://doi.org/10.7554/eLife.42014.020

- Supplementary file 3. Full details on the eight discrete loci. Listed here are the eight loci shown in ***Table 3.1***, with additional details on the core span and the TAD span used to identify candidate genes, and a full list of genes within the full span.

DOI: https://doi.org/10.7554/eLife.42014.021

- Supplementary file 4. Detected protein-coding changes with large allele frequency shift in amino acids. Listed are genes carrying large frequency changing SNPs affecting amino acid residues. Highlighted cells indicate the line with greater frequency changes ≥0.34 (red text with shading). Other suggestive changes are also shown with red numbers in unshaded cells. The changed amino acids are marked using standard notations, with the directionality indicated as 'purifying' or 'diversifying' with respect to a 60-way protein sequence alignment with other placental mammals. The conservation score based on phastCons was calculated at the SNP position itself, ranging from 0 (no conservation) to 1 (complete conservation) among the 60 placental mammals. For each gene, reported knockout phenotypes of the 'limbs/digits/tail' category is reported, along with whether lethality was reported in any of the alleles, excluding compound genotypes. A summary of the mutant phenotype as reported by the Mouse Genome Informatics database is also included to highlight any systemic defects beyond the 'limbs/digits/tail' category (lethal phenotypes reported in bold).

DOI: https://doi.org/10.7554/eLife.42014.022

- Supplementary file 5. Oligonucleotides for *in situ* hybridization probes.

DOI: https://doi.org/10.7554/eLife.42014.023

- Supplementary file 6. Oligonucleotide primers for multiplexed 4C-seq of enhancer viewpoints at the *Nkx3-2* locus. The 4C-seq adapter and adapter-specific primer sequences are given in ***Sexton et al. (2012)***. N2-DS denotes its location as 18 kbp downstream of the actual N2 enhancer. All viewpoints are pointed towards *Nkx3-2* gene body ('+' strand).

DOI: https://doi.org/10.7554/eLife.42014.024

- Supplementary file 7. Oligonucleotide primers for amplifying the enhancers at the *Nkx3-2* locus. Each of the amplicons are tagged with *SalI* (forward) or *XhoI* (reverse) sites (underlined) for concatenation and flanked by *EcoRV*

sites (underlined and bold) for insertion into the pBeta-lacZ-attBx2 reporter vector upstream of the b-globin minimal promoter.

DOI: https://doi.org/10.7554/eLife.42014.025

- Supplementary file 8. Oligonucleotide primers for allele-specific genotyping of the N3 enhancer. The primers were designed to target two SNPs (bold) in the N3 enhancer, rs33219710 and rs33600994.

DOI: https://doi.org/10.7554/eLife.42014.026

- Transparent reporting form

DOI: https://doi.org/10.7554/eLife.42014.027

Data availability

Sequencing data have been deposited in SRA (accession number SRP165718), GEO (accession numbers GSE121564, GSE121565 and GSE121566). Non-sequence data have been deposited at Dryad (doi:10.5061/dryad.0q2h6tk). Analytical code and additional notes have been deposited in the following repository:https://github.com/evolgenomics/Longshanks (copy archived at https://github.com/elifesciences-publications/Longshanks). Additional raw data and code are hosted via our institute's FTP servers at http://ftp.tuebingen.mpg.de/fml/ag-chan/Longshanks/.

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Castro JPL, Yancoskie MN, Marchini M, Belohlavy S, Hiramatsu L, Kuc˘ka M, Beluch WH, Naumann R, Skuplik IO, Cobb J, Barton NH, Rolian CP, Chan YF	2019	An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice	http://www.ncbi.nlm.nih. gov/sra?term=SRP165718	NCBI Sequence Read Archive, SRP165718
Castro JPL, Yancoskie MN, Marchini M, Belohlavy S, Hiramatsu L, Kuc˘ka M, Beluch WH, Naumann R, Skuplik IO, Cobb J, Barton NH, Rolian CP, Chan YF	2019	An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice	http://www.ncbi.nlm.nih. gov/geo/query/acc.cgi?acc=GSE121564	NCBI Gene Expression Omnibus, GSE121564
Castro JPL, Yancoskie MN, Marchini M, Belohlavy S, Hiramatsu L, Kuc˘ka M, Beluch WH, Naumann R, Skuplik IO, Cobb J, Barton NH, Rolian CP, Chan YF	2019	An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice	http://www.ncbi.nlm.nih. gov/geo/query/acc.cgi?acc=GSE121565	NCBI Gene Expression Omnibus, GSE121565
Castro JPL, Yancoskie MN, Marchini M, Belohlavy S, Hiramatsu L, Kuc˘ka M, Beluch WH, Naumann R, Skuplik IO, Cobb J, Barton NH, Rolian CP, Chan YF	2019	An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice	http://www.ncbi.nlm.nih. gov/geo/query/acc.cgi?acc=GSE121566	NCBI Gene Expression Omnibus, GSE121566
Castro JPL, Yancoskie MN, Marchini M, Belohlavy S, Hiramatsu L, Kuc˘ka M, Beluch WH, Naumann R, Skuplik IO, Cobb J, Barton	2019	An integrative genomic analysis of the Longshanks selection experiment for	http://dx.doi.org/10.5061/dryad.0q2h6tk	Dryad Digital Repository, 10.5061/dryad.0q2h6tk

NH, Rolian CP, Chan YF		longer limbs in mice

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Keane TM, Goodstadt L, Danecek P, White MA, Wong K	2011	Mouse Genomes Project version 3 dbSNP v137 release	https://www.sanger.ac.uk/science/data/mousegenomes-project	Wellcome Sanger Institute, dbSNP v137 release
Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B	2012	A map of the cis-regulatory sequences in the mouse genome	https://genome.ucsc.edu/encode/dataMatrix/encodeDataMatrixHuman.html	ENCODE Experiment Matrix, Mouse E14.5 Limb
Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, the Mouse Genome Database Group	2018	Mouse knockout phenotypes	http://www.informatics.jax.org/downloads/reports/MGI_PhenotypicAllele.rpt	Mouse Genome Informatics, MGI_PhenotypicAllele
Wood AR, Esko T, Yang J, Vedantam S	2014	Defining the role of common variation in the genomic and biological architecture of adult human height	https://portals.broadinstitute.org/collaboration/giant/index.php?title=GIANT_consortium&oldid=251	GIANT consortium, GWAS Anthropometric 2014 Height

References

Akazawa H, Komuro I, Sugitani Y, Yazaki Y, Nagai R, Noda T. 2000. Targeted disruption of the homeobox transcription factor *Bapx1* results in lethal skeletal dysplasia with asplenia and gastroduodenal malformation. *Genes to Cells* **5**:499–513. DOI: https://doi.org/10.1046/j.1365-2443.2000.00339.x, PMID: 10886375

Akiyama R, Kawakami H, Wong J, Oishi I, Nishinakamura R, Kawakami Y. 2015. *Sall4-Gli3* system in early limb progenitors is essential for the development of limb skeletal elements. *PNAS* **112**:5075–5080. DOI: https://doi. org/10.1073/pnas.1421949112, PMID: 25848055

Andrews S. 2016. FastQC A quality control tool for high throughput sequence data. http://www.bioinformatics. babraham.ac.uk/projects/fastqc/ [Accessed August 5, 2016].

Barton NH. 1995. Linkage and the limits to natural selection. *Genetics* 140:821–841. PMID: 7498757

Barton NH, Hermisson J, Nordborg M. 2019. Why structure matters. *eLife* 8:e45380. DOI: https://doi.org/10. 7554/eLife.45380, PMID: 30895925

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. *PLOS Genetics* **10**:e1004412. DOI: https://doi.org/10.1371/journal.pgen.1004412, PMID: 25102153

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pen˜ a-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**:1266– 1276. DOI: https://doi.org/10.1016/j.cell.2008.05.024, PMID: 18585359

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**:2114–2120. DOI: https://doi.org/10.1093/bioinformatics/btu170, PMID: 24695404

Bren-Mattison Y, Hausburg M, Olwin BB. 2011. Growth of limb muscle is dependent on skeletal-derived *Indian hedgehog*. *Developmental Biology* **356**:486–495. DOI: https://doi.org/10.1016/j.ydbio.2011.06.002, PMID: 21683695

Brown SD, Chambon P, de Angelis MH, Eumorphia Consortium. 2005. EMPReSS: standardized phenotype screens for functional annotation of the mouse genome. *Nature Genetics* **37**:1155. DOI: https://doi.org/10. 1038/ng1105-1155, PMID: 16254554

Browning BL, Browning SR. 2016. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics* **98**:116–126. DOI: https://doi.org/10.1016/j.ajhg.2015.11.020, PMID: 26748515

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**:1213–1218. DOI: https://doi.org/10.1038/nmeth.2688, PMID: 24097267

Bult A, Lynch CB. 2000. Breaking through artificial selection limits of an adaptive behavior in mice and the consequences for correlated responses. *Behavior Genetics* **30**:193–206. PMID: 11105393

Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD. 2010. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467**:587–590. DOI: https://doi.org/10.1038/nature09352, PMID: 20844486

Büscher D, Bosse B, Heymer J, Rüther U. 1997. Evidence for genetic control of *Sonic hedgehog* by *Gli3* in mouse limb development. *Mechanisms of Development* **62**:175–182. DOI: https://doi.org/10.1016/S0925-4773(97) 00656-4, PMID: 9152009

Careau V, Wolak ME, Carter PA, Garland T. 2013. Limits to behavioral evolution: the quantitative genetics of a complex trait under directional selection. *Evolution* **67**:3102–3119. DOI: https://doi.org/10.1111/evo.12200

Carlborg O, Jacobsson L, Ahgren P, Siegel P, Andersson L. 2006. Epistasis and the release of genetic variation during long-term selection. *Nature Genetics* **38**:418–420. DOI: https://doi.org/10.1038/ng1761, PMID: 16532011

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**:25–36. DOI: https://doi.org/10.1016/j.cell.2008.06.030, PMID: 18614008

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jo´ nsson B, Schluter D, Bell MA, Kingsley DM. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent

deletion of a *Pitx1* enhancer. *Science* **327**:302–305. DOI: https:// doi.org/10.1126/science.1182213, PMID: 20007865

Chan YF, Jones FC, McConnell E, Bryk J, Bünger L, Tautz D. 2012. Parallel selection mapping using artificially selected mice reveals body weight control loci. *Current Biology* **22**:794–800. DOI: https://doi.org/10.1016/j. cub.2012.03.011, PMID: 22445301

Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, Clark AG, Coop G. 2019. Allele frequency dynamics in a pedigreed natural population. *PNAS* **116**:2158–2164. DOI: https://doi.org/10.1073/pnas. 1813852116, PMID: 30598449

Cingolani P, Platts A, Wang leL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single Nucleotide Polymorphisms, SnpEff: snps in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**:80–92. DOI: https://doi.org/10.4161/fly.19695, PMID: 22728672

Cox A, Ackert-Bicknell CL, Dumont BL, Ding Y, Bell JT, Brockmann GA, Wergedal JE, Bult C, Paigen B, Flint J, Tsaih SW, Churchill GA, Broman KW. 2009. A new standard genetic map for the laboratory mouse. *Genetics* **182**:1335–1344. DOI: https://doi.org/10.1534/genetics.109.105486, PMID: 19535546

Crotwell PL, Mabee PM. 2007. Gene expression patterns underlying proximal-distal skeletal segmentation in late-stage zebrafish, *danio rerio*. *Developmental Dynamics* **236**:3111–3128. DOI: https://doi.org/10.1002/dvdy. 21352, PMID: 17948314

Crow JF, Kimura M. 1965. Evolution in sexual and asexual populations. *The American Naturalist* **99**:439–450. DOI: https://doi.org/10.1086/282389

Danecek P , Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158. DOI: https://doi.org/10.1093/bioinformatics/btr330, PMID: 21653522

Darwin C. 1859. *On the Origin of Species by Means of Natural Selection*. London: John Murray.

de Villemereuil P. 2019. Estimation of a biological trait heritability using the animal model: how to use the R package MCMCglmm. http://devillemereuil.legtux.org/wp-content/uploads/2012/12/tuto_en.pdf [Accessed March 21, 2019].

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**:491–498. DOI: https://doi.org/10.1038/ng.806, PMID: 21478889

DiLeone RJ, Marcus GA, Johnson MD, Kingsley DM. 2000. Efficient studies of long-distance *Bmp5* gene regulation using bacterial artificial chromosomes. *PNAS* **97**:1612–1617. DOI: https://doi.org/10.1073/pnas.97.4. 1612, PMID: 10677507

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**:376–380. DOI: https://doi.org/10.1038/ nature11082, PMID: 22495300

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74. DOI: https://doi.org/10.1038/nature11247, PMID: 22955616

Epstein B, Jones M, Hamede R, Hendricks S, McCallum H, Murchison EP, Schönfeld B, Wiench C, Hohenlohe P, Storfer A. 2016. Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nature Communications* **7**:12684. DOI: https://doi.org/10.1038/ncomms12684, PMID: 27575253

Evolgenomics. 2019. Longshanks. abd47cd. Github. https://github.com/evolgenomics/Longshanks

Falconer DS, Mackay TF. 1996. *Introduction to Quantitative Genetics*. Fourth Edition. London: Pearson.

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Xu J, Law M, Shaw DR, Baldarelli RM, Beal JS, Blodgett O, Campbell JW, Corbani LE, Lewis JR, Forthofer KL, Frost PJ, Giannatto SC, Hutchins LN, Miers DB, Motenko H, Stone KR, et al. 2017. The mouse gene expression database (GXD): 2017 update. *Nucleic Acids Research* **45**: D730–D736. DOI: https://doi.org/10.1093/nar/gkw1073, PMID: 27899677

Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**:462–470. DOI: https://doi. org/10.1038/nature13182, PMID: 24670764

Garland T, Rose MR. 2009. *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments*. Berkeley: University of California Press.

Gingerich PD. 2001. Rates of evolution on the time scale of the evolutionary process. *Genetica* **112**-113:127– 144. PMID: 11838762

Grant PR, Grant BR. 1992. Demography and the genetically effective sizes of two populations of Darwin's Finches. *Ecology* **73**:766–784. DOI: https://doi.org/10.2307/1940156

Grant PR, Grant BR. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. *Science* **296**:707–711. DOI: https://doi.org/10.1126/science.1070315, PMID: 11976447

Hadfield J. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software* **33**. DOI: https://doi.org/10.18637/jss.v033.i02

Haldane JBS. 1932. *The Causes of Evolution*. London: Green and Co.

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. 2011. Adaptations to climate-mediated selective pressures in humans. *PLOS Genetics* **7**:e1001375. DOI: https://doi.org/10.1371/journal.pgen.1001375, PMID: 21533023

Hellemans J, Simon M, Dheedene A, Alanay Y, Mihci E, Rifai L, Sefiani A, van Bever Y, Meradji M, Superti-Furga A, Mortier G. 2009. Homozygous inactivating mutations in the *NKX3-2* gene result in spondylomegaepiphyseal-metaphyseal dysplasia. *The American Journal of Human Genetics* **85**:916–922. DOI: https:// doi.org/10.1016/j.ajhg.2009.11.005, PMID: 20004766

Hendry AP, Kinnison MT. 1999. Perspective: the pace of modern life: measuring rates of contemporary microevolution. *Evolution* **53**:1637–1653. DOI: https://doi.org/10.1111/j.1558-5646.1999.tb04550.x

Hill WG. 1982. Predictions of response to artificial selection from new mutations. *Genetical Research* **40**:255– 278. DOI: https://doi.org/10.1017/S0016672300019145, PMID: 6819185

Hill WG, Goddard ME, Visscher PM. 2008. Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics* **4**:e1000008. DOI: https://doi.org/10.1371/journal.pgen.1000008, PMID: 18454194

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetical Research* **8**:269–294. DOI: https://doi.org/10.1017/S0016672300010156, PMID: 5980116

Hill WG, Zhang XS. 2012. On the pleiotropic structure of the genotype-phenotype map and the evolvability of complex organisms. *Genetics* **190**:1131–1137. DOI: https://doi.org/10.1534/genetics.111.135681, PMID: 22214609

Hoffmann AA, Sgro` CM. 2011. Climate change and evolutionary adaptation. *Nature* **470**:479–485. DOI: https:// doi.org/10.1038/nature09670, PMID: 21350480

Jha AR, Miles CM, Lippert NR, Brown CD, White KP, Kreitman M. 2015. Whole-genome resequencing of experimental populations reveals polygenic basis of egg-size variation in *Drosophila melanogaster*. *Molecular Biology and Evolution* **32**:2616–2632. DOI: https://doi.org/10.1093/molbev/msv136, PMID: 26044351

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht AK, Brady SD, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**:55–61. DOI: https://doi.org/10.1038/nature10944, PMID: 22481358

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nella˚ ker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**: 289–294. DOI: https://doi.org/10.1038/nature10413, PMID: 21921910

Keightley PD, Hardge T, May L, Bulfield G. 1996. A genetic map of quantitative trait loci for body weight in the mouse. *Genetics* **142**:227–235. PMID: 8770600

Keightley PD, Bünger L, Renne U, Buis RC. 2001.Reeve E C R (Ed). *Encyclopedia of Genetics*. Chicago: Fitzroy Dearborn Publishers.

Kelly JK, Hughes K. 2018. An examination of the evolve-and-resequence method using *Drosophila simulans*. *bioRxiv*. DOI: https://doi.org/10.1101/337188

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Research* **12**:996–1006. DOI: https://doi.org/10.1101/gr.229102, PMID: 12045153

Kessner D, Novembre J. 2015. Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. *Genetics* **199**:991–1005. DOI: https://doi.org/10.1534/genetics.115.175075, PMID: 25672748

Koziel L, Wuelling M, Schneider S, Vortkamp A. 2005. *Gli3* acts as a repressor downstream of *Ihh* in regulating two distinct steps of chondrocyte differentiation. *Development* **132**:5249–5260. DOI: https://doi.org/10.1242/ dev.02097, PMID: 16284117

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nature Methods* **9**:357–359. DOI: https://doi.org/10.1038/nmeth.1923, PMID: 22388286

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. DOI: https://doi.org/10.1093/bioinformatics/btp352, PMID: 19505943

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595. DOI: https://doi.org/10.1093/bioinformatics/btp698, PMID: 20080505

Lynch M, Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer.

Makvandi-Nejad S, Hoffman GE, Allen JJ, Chu E, Gu E, Chandler AM, Loredo AI, Bellone RR, Mezey JG, Brooks SA, Sutter NB. 2012. Four loci explain 83% of size variation in the horse. *PLOS ONE* **7**:e39929. DOI: https://doi. org/10.1371/journal.pone.0039929, PMID: 22808074

Marchini M, Sparrow LM, Cosman MN, Dowhanik A, Krueger CB, Hallgrimsson B, Rolian C. 2014. Impacts of genetic correlation on the independent evolution of body mass and skeletal size in mammals. *BMC Evolutionary Biology* **14**:258. DOI: https://doi.org/10.1186/s12862-014-0258-0, PMID: 25496561

Marchini M, Rolian C. 2018. Artificial selection sheds light on developmental mechanisms of limb elongation. *Evolution* **72**:825–837. DOI: https://doi.org/10.1111/evo.13447

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **10**. DOI: https://doi.org/10.14806/ej.17.1.200

Martin A, Orgogozo V. 2013. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **22**:1235–1250. DOI: https://doi.org/10.1111/evo.12081

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* **23**:23–35. DOI: https://doi.org/10.1017/S0016672300014634, PMID: 4407212

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303. DOI: https://doi.org/10.1101/gr.107524.110, PMID: 20644199

McRae AF, Pemberton JM, Visscher PM. 2005. Modeling linkage disequilibrium in natural populations: the example of the Soay sheep population of St. Kilda, Scotland. *Genetics* **171**:251–258. DOI: https://doi.org/10. 1534/genetics.105.040972, PMID: 15965254

Milani P, Escalante-Chong R, Shelley BC, Patel-Murray NL, Xin X, Adam M, Mandefro B, Sareen D, Svendsen CN, Fraenkel E. 2016. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Scientific Reports* **6**:25474. DOI: https://doi.org/10.1038/srep25474, PMID: 27146274

Mo R, Freer AM, Zinyk DL, Crackower MA, Michaud J, Heng HH, Chik KW, Shi XM, Tsui LC, Cheng SH, Joyner AL, Hui C. 1997. Specific and redundant functions of *Gli2* and *Gli3* zinc finger genes in skeletal patterning and development. *Development* **124**:113–123. PMID: 9006072

Mortlock DP, Guenther C, Kingsley DM. 2003. A general approach for identifying distant regulatory elements applied to the *Gdf6* gene. *Genome Research* **13**:2069–2081. DOI: https://doi.org/10.1101/gr.1306003, PMID: 12915490

Nakamura T, Klomp J, Pieretti J, Schneider I, Gehrke AR, Shubin NH. 2015. Molecular mechanisms underlying the exceptional adaptations of batoid fins. *PNAS* **112**:15940–15945. DOI: https://doi.org/10.1073/pnas. 1521818112, PMID: 26644578

Nuzhdin SV, Dilda CL, Mackay TF. 1999. The genetic architecture of selection response. Inferences from finescale mapping of bristle number quantitative trait loci in *Drosophila melanogaster*. *Genetics* **153**:1317–1331. PMID: 10545462

Orozco-terWengel P, Kapun M, Nolte V, Kofler R, Flatt T, Schlötterer C. 2012. Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular Ecology* **21**: 4931–4941. DOI: https://doi.org/10.1111/j.1365-294X.2012.05673.x, PMID: 22726122

Orr HA. 2005. The probability of parallel evolution. *Evolution* 59:216–220. DOI: https://doi.org/10.1111/j.0014- 3820.2005.tb00907.x

Otto SP, Barton NH. 2001. Selection for recombination in small populations. *Evolution* **55**:1921–1931. DOI: https://doi.org/10.1111/j.0014-3820.2001.tb01310.x

Petit F, Sears KE, Ahituv N. 2017. Limb development: a paradigm of gene regulation. *Nature Reviews Genetics* **18**:245–258. DOI: https://doi.org/10.1038/nrg.2016.167, PMID: 28163321

Pettersson ME, Johansson AM, Siegel PB, Carlborg Örjan. 2013. Dynamics of adaptive alleles in divergently selected body weight lines of chickens. *G3: Genes|Genomes|Genetics* **3**:2305–2312. DOI: https://doi.org/10. 1534/g3.113.008375

Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. *Tn5* transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research* **24**:2033–2040. DOI: https://doi.org/ 10.1101/gr.177881.114, PMID: 25079858

Pitchers W, Wolf JB, Tregenza T, Hunt J, Dworkin I. 2014. Evolutionary rates for multivariate traits: the role of selection and genetic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**: 20130252. DOI: https://doi.org/10.1098/rstb.2013.0252

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**:110–121. DOI: https://doi.org/10.1101/gr.097857.109, PMID: 19858363

Provot S, Kempf H, Murtaugh LC, Chung UI, Kim DW, Chyung J, Kronenberg HM, Lassar AB. 2006. *Nkx3.2/ Bapx1* acts as a negative regulator of chondrocyte maturation. *Development* **133**:651–662. DOI: https://doi. org/10.1242/dev.02258, PMID: 16421188

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. DOI: https://doi.org/10.1093/bioinformatics/btq033, PMID: 20110278

R Development Core Team. 2015. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statasticial Computing.

Reeves RG, Tautz D. 2017. Automated phenotyping indicates pupal size in *Drosophila* is a highly heritable trait with an apparent polygenic basis. *G3: Genes|Genomes|Genetics* **7**:1277–1286. DOI: https://doi.org/10.1534/ g3.117.039883

Rice DP, Townsend JP. 2012. A test for selection employing quantitative trait locus and mutation accumulation data. *Genetics* **190**:1533–1545. DOI: https://doi.org/10.1534/genetics.111.137075, PMID: 22298701

Rimbault M, Beale HC, Schoenebeck JJ, Hoopes BC, Allen JJ, Kilroy-Glynn P, Wayne RK, Sutter NB, Ostrander EA. 2013. Derived variants at six genes explain nearly half of size reduction in dog breeds. *Genome Research* **23**:1985–1995. DOI: https://doi.org/10.1101/gr.157339.113, PMID: 24026177

Robertson A. 1960. A theory of limits in artificial selection. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **153**:234–249. DOI: https://doi.org/10.1098/rspb.1960.0099

Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, Hallböök F, Besnier F, Carlborg O, Bed'hom B, Tixier-Boichard M, Jensen P, Siegel P, Lindblad-Toh K, Andersson L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**:587–591. DOI: https://doi.org/10.1038/nature08832, PMID: 20220755

Schluter D, Clifford EA, Nemethy M, McKinnon JS. 2004. Parallel evolution and inheritance of quantitative traits. *The American Naturalist* **163**:809–822. DOI: https://doi.org/10.1086/383621, PMID: 15266380

Sexton T, Kurukuti S, Mitchell JA, Umlauf D, Nagano T, Fraser P. 2012. Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. *Nature Protocols* **7**:1335–1350. DOI: https://doi.org/10.1038/nprot.2012.071

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B. 2012. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**:116–120. DOI: https://doi.org/10.1038/ nature11243, PMID: 22763441

Sheng Z, Pettersson ME, Honaker CF, Siegel PB, Carlborg Örjan. 2015. Standing genetic variation as a major contributor to adaptation in the Virginia chicken lines selection experiment. *Genome Biology* **16**:219. DOI: https://doi.org/10.1186/s13059-015-0785-z

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**:1034–1050. DOI: https://doi. org/10.1101/gr.3715005, PMID: 16024819

Sivakamasundari V, Chan HY, Yap SP, Xing X, Kraus P, Lufkin T. 2012. New *Bapx1*(*Cre-EGFP*) mouse lines for lineage tracing and conditional knockout studies. *Genesis* **50**:375–383. DOI: https://doi.org/10.1002/dvg. 20802, PMID: 21913311

Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, Padhukasahasram B, Karlins E, Davis S, Jones PG, Quignon P, Johnson GS, Parker HG, Fretwell N, Mosher DS, Lawler DF, Satyaraj E, Nordborg M, Lark KG, Wayne RK, et al. 2007. A single *IGF1* allele is

a major determinant of small size in dogs. *Science* **316**:112–115. DOI: https://doi.org/10.1126/science.1137045, PMID: 17412960

Tribioli C, Lufkin T. 2006. *Bapx1* homeobox gene gain-of-function mice show preaxial polydactyly and activated *shh* signaling in the developing limb. *Developmental Dynamics* **235**:2483–2492. DOI: https://doi.org/10.1002/ dvdy.20867, PMID: 16791844

Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN, Genetic Investigation of ANthropometric Traits (GIANT) Consortium. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* **44**:1015–1019. DOI: https://doi.org/10.1038/ng.2368, PMID: 22902787

Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM. 2011. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLOS Genetics* **7**:e1001336. DOI: https://doi.org/10.1371/journal.pgen.1001336, PMID: 21437274

Wahlberg P, Carlborg O, Foglio M, Tordoir X, Syvänen AC, Lathrop M, Gut IG, Siegel PB, Andersson L. 2009. Genetic analysis of an F(2) intercross between two chicken lines divergently selected for body-weight. *BMC Genomics* **10**:248. DOI: https://doi.org/10.1186/1471-2164-10-248, PMID: 19473501

Walsh B, Lynch M. 2018. The infinitesimal model and its extensions. In: *Evolution and Selection of Quantitative Traits*. Oxford University Press.

Weber KE, Diggins LT. 1990. Increased selection response in larger populations. II. selection for ethanol vapor resistance in *Drosophila melanogaster* at two population sizes. *Genetics* **125**:585–597. PMID: 2116359

Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JR, Elliott KS, Hackett R, Guiducci C, Shields B, Zeggini E, Lango H, Lyssenko V, Timpson NJ, Burtt NP, Rayner NW, Saxena R, Ardlie K, Tobias JH, Ness AR, et al. 2007. A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nature Genetics* **39**:1245–1250. DOI: https://doi.org/10.1038/ng2121, PMID: 17767157

Weissman DB, Barton NH. 2012. Limits to the rate of adaptive substitution in sexual populations. *PLOS Genetics* **8**:e1002740. DOI: https://doi.org/10.1371/journal.pgen.1002740, PMID: 22685419

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen J, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**:1173–1186. DOI: https://doi.org/10.1038/ng.3097, PMID: 25282103

Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, Østera˚s M, Whitley A, Yuan W, Gan X, Goodson M, Klenerman P, Satpathy A, Mathis D, Benoist C, Adams DJ, Mott R, Flint J. 2010. Commercially available outbred mice for genome-wide association studies. *PLOS Genetics* **6**:e1001085. DOI: https://doi.org/10.1371/ journal.pgen.1001085, PMID: 20838427

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**:R137. DOI: https://doi.org/10.1186/ gb-2008-9-9-r137, PMID: 18798982

Appendix 1

Major considerations in constructing the simulations

In the Longshanks experiment, the highest-ranking male and the highest-ranking female from each family were chosen to breed with the highest-ranking mice from other families within a line (i.e., disallowing sibling matings). Thus, if we disregard non-Mendelian segregation, and the fraction of failed litters (15%), selection acts solely within families, on the measured traits. Such selection does not distort the pedigree and allows us to follow the evolution of each chromosome separately.

Our simulations track the inheritance of continuous genomes by following the junctions between regions with different ancestry. In principle, we should simulate selection under the infinitesimal model by following the contributions to the trait of continuous blocks of chromosomes across the whole genome. However, this is computationally challenging, since the contributions of all the blocks defined by every recombination event have to be tracked. Instead, we follow a large number of discrete biallelic loci checking that the number is sufficiently large to approach the infinitesimal limit (***Figure 3.3D***). We made a further slight approximation by only explicitly modeling discrete loci on one chromosome at a time. We divided the breeding value of an individual into two components. The first, V_g, is a contribution from a large number of unlinked loci, due to genes on all but the focal chromosome, as represented by the infinitesimal model. The values of this component amongst offspring are normally distributed around the mean of the parents, with its variance being:

$$V_M = (V_A/2)\,(1 - \beta)\,(1 - F_{ii} - F_{jj})$$

where: V_A is the initial genetic variance, and
F_{ii}, F_{jj} are the probabilities of identity between distinct genes in each parent, i, j; F_{ii}, F_{jj} are calculated from the pedigree;
β is the fraction of genome on the focal chromosome.

The second component, V_s, is the sum of contributions from a large number, n, of discrete loci, evenly spaced along the focal chromosome (here we used 10,000), and contributing a fraction β of the initial additive variance. We choose these to have equal effects and random signs, $\pm\alpha$, such that initial allele frequencies $p_0 = q_0 = \frac{1}{2}$, and equal effects α, such that $\beta V_{A,0} = 2\sum_{i=1}^{n} \alpha^2 p_{i,0} q_{i,0}$. The initial population consists of 28 diploid individuals, matching the experiment, and loci have initial frequencies of 1, 4, 12 and 28 out of the diploid total of 56 alleles, in equal proportions. Inheritance is assumed to be autosomal, with no sex-linkage. This choice of equal effects approaches most closely to the infinitesimal model, for a given number of loci.

The decrease in genetic variance due to random drift is measured by the inbreeding coefficient, defined as the probability of identity by descent, relative to the initial population. We distinguish the identity between two distinct genes within a diploid individual, F_w,from the probability of identity between two genes in different individuals, F_b. The overall mean identity between two genes chosen independently and at random from all $2N$ genes is $\bar{F} = \frac{2(N-1)F_b + F_w + 1}{2N}$. The proportion of heterozygotes in the population decreases by a factor of $1 - F_w$, the variance in allele frequency increases with $\bar{F}$, and the genetic diversity, $\mathbb{E} = [2pq]$, decreases as $1 - \bar{F}$.

Figure 3.3B shows that in the absence of selection, the identity F_b increases slower than expected under the Wright–Fisher model with the actual population sizes (compare light shaded lines with black lines). These differences are a consequence of the circular mating

scheme, which was designed to slow the loss of variation. The dotted line show the average F, estimated from the loss of heterozygosity in 50 replicate neutral simulations, each with 10^4 loci on a chromosome of length *R*=1 Morgan. These are close to the prediction from the pedigree (light shaded lines), validating the simulations.

The thick colored line in ***Figure 3.3B*** shows F, estimated in the same way from simulations that include truncation selection on a trait with within-family variance V_s/V_e = 0.584 (a value we abbreviate as $\theta = 1$), which matches the observed selection response and parent-offspring regression. The rate of drift, as measured by the gradient in F over time, is significantly faster in simulations with selection, by 6.7% in LS1 and 9.8% in LS2 (Student's *t*-test $P \leq 0.008$ in LS1 and $P \leq 0.0005$ in LS2). However, this effect of selection would not be detectable from any one replicate, since the standard deviation of the rate of drift, relative to the mean rate, is ~13% between replicates. On average, the observed loss of heterozygosity fits closely to that expected from the pedigree (large dot with error bars), though there is wide variation among chromosomes (filled dots), which is substantially higher than seen in simulations seeded with SNP at linkage equilibrium (compare filled and open dots).

We then performed 100 simulations, seeding each founding generation with actual genotypes and using actual pedigrees, selection pressure or heritability parameters (within-family heritability h^2 of the fitness dimension: 0.51). A main conclusion from our modelling is that the overall allele frequencies were hardly perturbed by varying selection from random drift to even doubling the selection intensity. Upon closer examination, it became clear that under the standard "infinitesimal" model, selection could generate a weak but detectable excess of allele frequency sweeps compared to strict neutrality with no selection (***Figure 3.3D***, SNP classes 1/56 and 4/56). However, it would take many replicates (assuming no parallelism) for this excess to become statistically significant. Taken at face value, this result echoes many "evolve-and-resequence" (E&R) experiments based on diverse base populations that show only weak evidence of selective sweeps at specific loci (***Burke et al., 2010***; ***Orozco-terWengel et al., 2012***).

Broader patterns and analyses of parallelism

On a broader scale, we also observed greater extent of parallelism globally than in the simulated results or with the empirical Ctrl line. For example, out of the 2405 and 2991 loci found above the $H_{INF,\ no\ LD}$ cut-off in LS1 and LS2, 398 were found in both lines (13%; χ^2 test, N~150,000 windows; χ^2=2901.4, d.f.=1, $P \leq 1 \times 10^{-10}$); whereas we found only 10 or 7 overlaps in Ctrl–LS1 or Ctrl–LS2 comparisons, respectively. This difference is statistically significant (940 significant Ctrl loci at the $H_{INF,\ no\ LD}$ threshold; N~150,000 windows; Ctrl–LS1: χ^2=0.7; Ctrl–LS2: χ^2=6.0; both *P* = n.s.; see also ***Figure 3.10***). In fact, there was not a single window out of a total of 8.4 million windows from the 100 replicates where both simulated LS1 and LS2 replicates simultaneously cleared the $H_{INF,\ no\ LD}$ threshold. In contrast to our earlier analysis in single LS replicates, the parallel selected loci in both LS replicates loci may provide the strongest evidence yet to reject the infinitesimal model.

Heritability estimate by an animal model

We estimated heritability using linear mixed effect "animal models" with maximum likelihood (***Figure 3.2D***) in the R package MCMCglmm v2.5 (***Hadfield, 2010***; following guide by ***de Villemereuil, 2019***). Because the animal model makes inference of the parameter estimates to the base population, to compare heritability as it changed over time we estimated heritability in blocks of 5 generations F0-4, 5-9, 10-14, and 15-19, separately for each selected line. In testing each block, we used the full pedigree to build the relationship matrix but only phenotypes from the individuals in those generations. As an alternative, we

tested each block with a truncated pedigree, in which the first generation of each block is treated as unrelated (i.e., the base population). The two methods produced similar results. In all analyses, we standardized the composite trait $ln(\boldsymbol{T}\boldsymbol{B}^{-0.57})$ ($\boldsymbol{T}$ = tibia length in mm; $\boldsymbol{B}$ = cube-root body mass in $\sqrt[3]{}$g; see *Simulating selection response: infinitesimal model with linkage* in main text) within each generation and line to account for fluctuations in mean and variance (***Careau et al., 2013***). The phenotypic variance was partitioned as V_P = fixed effects + V_A + V_R, where fixed effects were sex, age, and litter size, V_A was additive genetic variance, and V_R was residual variance. Heritability was estimated as $h^2 = V_A / (V_A + V_R)$.

Enrichment for genes with functional impact on limb development

To determine what types of molecular changes may have mediated the selection response, we performed a gene set enrichment analysis. We asked if the outlier loci found in the Longshanks lines were enriched for genes affecting limb development (as indicated by their knockout phenotypes) and found increasingly significant enrichment as the allele frequency shift Δz^2 cut-off became increasingly stringent (***Figure 3.8A***). The "limb/digital/tail" category of affected anatomical systems in the Mouse Genomic Informatics Gene Expression Database (***Finger et al., 2017***) showed the greatest excess of observed-to-expected ratio out of all 28 phenotype categories (the excluded "normal" category also showed no enrichment). In contrast, genes showing knock-out phenotypes in most other categories did not show similar enrichment as Δz^2 became more stringent (***Figure 3.8A***). For genes expressed in limb tissue, there was a similar, but weaker increase, with the enrichment only appearing at higher Δz^2 cut-off. We did not observe similar enrichment using data and thresholds derived from Ctrl (***Figure 3.8A***, lower panels). To investigate the impact on regulatory sequences, we obtained 21,211 limb enhancers predicted by ENCODE chromatin profile at a stage immediately preceding bone formation (Theiler Stage 23, at approximately embryonic day E14.5) (***Shen et al., 2012***). We found likewise an enrichment throughout the range of significance cut-offs (***Figure 3.8A***). Again, there was no similar enrichment in Ctrl.

Clustering with loci associated with human height

Since tibia lengths directly affect human height, we tested if an association exists between loci controlling human height (***Wood et al., 2014***) and a set of 810 loci at the $P \leq 0.05$ significance level under $H_{INF,\ no\ LD}$ described here. After remapping the human loci to their orthologous mouse positions (n = 655 out of 697 total height loci; data from the GIANT Consortium), we detected significant clustering with the 810 peak loci (mean pairwise distance to remapped height loci: 1.41 Mbp vs. mean 1.69 Mbp from 1000 permutations of shuffled peak loci, range: 1.45–1.93 Mbp; n = 655 height loci and 810 peak loci; $P < 0.001$, permutations). We interpret this clustering to suggest that a shared and conserved genetic program exist between human height and tibia length and/or body mass.

Genome-wide analysis of the role of coding vs. cis-*acting changes in response to selection*

We examined the potential functional impact of coding or regulatory changes as a function of Δz^2 in all three lines. For coding changes, we tracked the functional consequences of coding SNPs of moderate to high impact (missense mutations, gain or loss of stop codons, or frame-shifts). Whereas we found only mixed evidence of increased coding changes as Δz^2 increased in the LS lines, there was a depletion of coding changes in Ctrl line as Δz^2 increased, possibly due to purifying or background selection (***Figure 3.8B***; linear regression, LS1: $P \leq 0.015$, slope > 0; LS2: $P = 0.62$, n.s., slope ≈ 0; Ctrl: $P \leq 5.72\times10^{-9}$, slope < 0).

For regulatory changes, we used sequence conservation in limb enhancers overlapping a SNP as a proxy for functional impact. In contrast to the situation for coding

changes, where the correlations differed between LS1 and LS2, the potential impact of regulatory changes increased significantly as a function of Δz^2 in both LS lines (***Figure 3.8B***): within limb enhancers, SNP-flanking sequences became increasingly conserved at highly differentiated SNPs (phastCons conservation score, ranging from 0 to 1 for unconserved to completely conserved positions; linear regression, log-scale, $P < 1.05\times10^{-9}$ for both, slopes > 0). This relationship also exists for the Ctrl line, albeit principally from lower Δz^2 and conservation values ($P < 0.8\times10^{-3}$, slope > 0; ***Figure 3.8B***). Taken together, our enrichment analysis suggests that while both coding and regulatory changes were selected in the Longshanks experiment, the overall selection response may depend more consistently on *cis*-regulatory changes, especially for developmental regulators involved in limb, bone and/or cartilage development (***Table 3.1***; ***Supplementary File 3***; c.f. ***Supplementary File 4*** for coding changes). This is a key prediction of the "*cis*-regulatory hypothesis", especially in its original scope on morphological traits (***Carroll, 2008***).

Genes with amino acid changes of potentially major impact

We have further identified 12 candidate genes with likely functional impact on limb development due to specific amino acid changes showing large frequency shifts (albeit only one, *Fbn2*, cleared the stringent $P \leq 0.05$ $H_{INF,\ max\ LD}$ threshold; 6 in LS1, 9 in LS2, of which 3 were shared; ***Supplementary File 4***). Consistent with strong selection for tibia development, all 12 genes show limb or tail phenotypes when knocked out, e.g., "short limbs" for the collagen gene *Col27a1* knockout. Most of these genes encode for structural cellular components, e.g., myosin, fibrillin and collagen (*Myo10; Fbn2*; and *Col27a1* respectively), with *Fuz* (fuzzy planar cell polarity protein) being the only classical developmental regulator gene. All but one of these genes have also been shown to have widespread pleiotropic effects with broad expression domains, and their knockouts were often lethal (eight out of 12) and/or exhibit defects in additional organ systems (11 out of 12). Based on this observation, we anticipate that the phenotypic impact of these selected coding missense SNPs (n.b. not knockout) would not be restricted to tibia or bone development.

Molecular dissection of Gli3, *a candidate limb regulator, reveal gain-of-function* cis*-acting changes*

We anticipated that genes related to major limb patterning, like *Gli3*, may contribute to the selection response (***Mo et al., 1997; Nakamura et al., 2015***). We thus performed an in-depth molecular dissection of *Gli3*, an important early limb developmental regulator on chromosome 13 (Chr13; ***Figure 3.12A***). This locus showed a substantial shift in minor allele frequency of up to 0.42 in LS1 (Δq, 98th quantile genome-wide, but below the $H_{INF,\ max\ LD}$ threshold to qualify as a discrete major locus). We performed functional validation of *Gli3*, given its limb function (***Büscher et al., 1997***) and considering that *Gli3* could be among the many minor loci in the polygenic background contributing to the selection response in LS1.

At the *Gli3* locus we could only find conservative amino acid changes (D1090E and I1326V) that are unlikely to impact protein function. Because the signal in LS1 was stronger in the 5' flanking intergenic region, we examined the *Gli3 cis*-regulatory topologically associating domain (TADs, which mark chromosome segments with shared gene regulatory logic) (***Dixon et al., 2012***) and identified putative enhancers using chromatin modification marks from the ENCODE project and our own ATAC-Seq data (***Figure 3.12B***) (***Buenrostro et al., 2013; Shen et al., 2012***). Four putative enhancers carried SNPs with large allele frequency changes. Among them, an upstream putative enhancer G2 (956 bp) carried 6 SNPs along with two 1- and 3-bp insertion/deletion ("indel") with putative functional impact due to predicted gain or loss in transcription factor binding sites (***Figure 3.12C***). We tested the G2 putative enhancer in a transgenic reporter assay by placing its sequence as a tandem duplicate upstream of a *lacZ* reporter gene (see Methods for details). We found that only the F17 LS1 allele was able to drive consistent *lacZ* expression in the developing limb

buds (***Figure 3.12D***). Importantly, this enhancer was active not only in the shaft of the limb bud but also in the anterior hand/foot plate, a major domain of *Gli3* expression and function (***Figure 3.12A***). Furthermore, substitution of the enhancer sequence with the F0 allele (10 differences out of 956 or 960 bp) abolished *lacZ* expression (***Figure 3.12D***). This showed that 10 or fewer changes within this novel enhancer sequence were sufficient to convert the inactive F0 allele into an active limb enhancer corresponding to the selected F17 allele ("gain-of-function"), suggesting that a standing genetic variant of the F17 allele may have been selectively favored because it drove stronger expression of *Gli3*, a gene essential for tibia development (***Akiyama et al., 2015***, but see ***Koziel et al., 2005***).

Estimating the selection coefficient of the top-ranking locus, Nkx3-2*, from changes in allele frequency*

The significant locus on Chr5 containing *Nkx3-2* shows strong changes in SNP frequency in both LS1 and LS2. Here, we estimate the strength of selection on this locus, and the corresponding effect on the selected trait. We approximate by assuming two alternative alleles, and find the selection coefficient implied the observed parallel changes in allele frequency; we then set bounds on this estimate that take account of random drift. Finally, we use simulations that condition on the known pedigree to estimate the effect on the trait required to cause the observed strong frequency changes; these show that linked selection has little effect on the single-locus estimates.

We see strong and parallel changes in allele frequency at multiple steps. There are 14 non-overlapping 10kb windows that have a mean square change in arc-sin transformed allele frequency of $\overline{\Delta z^2} > 2$ in both LS1 and LS2, spanning a 260 kbp region and including 807 SNP. SNP frequencies are tightly clustered, corresponding to two alternative haplotypes (***Figure 3.14*** and ***Figure 3.15A***). The initial (untransformed) allele frequencies average q_0 = 0.18, 0.17 in LS1, LS2, respectively, and the final frequencies average q_{17} = 0.84, 0.98, respectively (also see ***Figure 3.15A***, lower panel). These frequencies depend on the arbitrary threshold for which windows to include. However, this makes little difference, relative to the wide bounds on our estimates.

Under constant selection, $log\frac{q}{p}$ changes linearly with time, at a rate equal to the selection coefficient, s. Therefore, a naive estimate of selection is given by $\hat{s} = \frac{1}{T}\log\left[\frac{q_{17}}{p_{17}}\frac{p_0}{q_0}\right]$ (***Haldane, 1932***) thus, $\hat{s}$ =0.19, 0.32 for *q* in LS1, LS2, and averages 0.26. Here, males and females with longest tibia are chosen to breed; the strength of selection on an additive allele depends on the fraction selected and the within-family trait variance. The former is kept constant, and there is little loss of variance due to drift (F~0.17). Thus, assuming constant selection is reasonable (***Figure 3.14B***), unless there is strong dominance.

To set bounds on this estimate, we must account for random drift. The predicted loss of diversity over 17 generations, based on the pedigree, is F=0.173, 0.175 for LS1, LS2, which corresponds to an effective size $N_e = 44.9, 44.4$, respectively (note that due to differences in estimation methodology, this N_e differs slightly from that mentioned in ***Figure 3.3*** but is largely consistent). Therefore, we calculate the matrix of transition probabilities for a Wright–Fisher population with $2N$ rounded to 90, 89 copies for LS1, LS2, over a range of selection, s. This yields the probability that the number of copies would change from the rounded values of 16/90 to 75/90 in LS1, and from 14/89 to 87/89 in LS2—that is, the likelihood of s, given the observed changes in allele frequency, and the known N_e. There is no significant loss of likelihood by assuming the same selection in both lines; overall, $\hat{s}$ = 0.24 (limits 0.13–0.36; ***Figure 3.15B***).

Estimating the selection coefficient, accounting for linked loci

The estimates above using the simple approach do not account for selection on linked loci, and do not give the effect on the composite trait. We therefore simulated conditional on the pedigree and on the actual selection regime, as described above, but including an additive allele with effect *A* at the candidate locus on Chr5. The genetic variance associated with the unlinked infinitesimal background, and across Chr5, were reduced in proportion, to keep the overall heritability the same as before $V_a/(V_a + V_e) = 0.539$. The selection coefficient inferred from the simulated changes in allele frequency was approximately proportional to the effect on the trait, with best fit $s = 0.41A/\sqrt{V_e}$ (***Figure 3.15C***, left). Assuming this relationship, we can compare the mean and standard deviation of allele frequency from simulations with linked selection, with that predicted by the single locus Wright–Fisher model (points vs. line in ***Figure 3.15C***, middle & right). These agree well, showing that linked selection does not appreciably change the distribution of allele frequencies at a single locus. This is consistent with ***Figure 3.3D***, which shows that linked selection only inflates the tail of the allele frequency distribution, an effect that would not be detectable at a single locus.

Combining our estimates of the selection coefficient with the relation $s = 0.41A/\sqrt{V_e}$, we estimate that the locus on Chr5 has effect $\hat{A} = 0.59\sqrt{V_e}$, with 2-unit support limits $0.32\sqrt{V_e}$ to $0.87\sqrt{V_e}$. This single locus is responsible for ~9.4% of the total selection response (limits 3.6–15.5%).

This analysis does not allow for the inflation of effect that might arise from multiple testing. This is hard to estimate, because it depends on the distribution of effects across the genome, and also on the excess variation in estimates due to LD in the founder population. However, we note that if the effect of this locus is large enough that it would certainly be detected in this study, then there is no estimation bias from this source.

We also assume that there are two haplotypes, each with a definite effect. There might in fact be heterogeneity in the effects of each haplotype, for two reasons. First, this region might have had heterogeneous effects in the founder population, with multiple alleles at multiple causal loci. Second, as recombination breaks up the founder genomes, blocks of genome would become associated with different backgrounds. To the extent that genetic variation is spread evenly over an infinitesimal background, this latter effect is accounted for by our simulations, and has little consequence. However, we have not tested whether the data might be explained by more than two alleles, possibly at more than one discrete locus. Testing such complex models would be challenging, and we do not believe that such test would have much power. However, the estimates of selection made here should be regarded as effective values that may reflect a more complex reality.

Estimating the contribution of the Nkx3-2 *locus using an animal model*

We used a linear mixed “animal model” to estimate the effect of the enhancer N3 (of the major locus in Nkx3-2) on the composite selected trait $ln(\boldsymbol{TB}^{-0.57})$, see Section ‘Simulating selection response: infinitesimal model with linkage’ and ***Figure 3.3A***. The model was:

V_P = fixed effects + V_A + V_R

where:

fixed effects = sex, generation, litter size (i.e., number of siblings in family), genotype at N3 (0, 1, or 2 copies of F17 allele), and replicate line
V_A = additive genetic variance
V_R = residual variance

We found a small but significant effect of the genotype at enhancer N3 on the composite trait (mean effect = 0.0036; 95% CI: 0.00069–0.0064; *P*=0.017). Given the same body mass $\boldsymbol{B}$, the mean effect corresponds to 0.36% increase in tibia length per copy of the F17 allele, or ~1% of the variance in tibia length at generation F01. The observed increase of this allele from ~0.18 to 0.91, averaged over the two lines, implies that it accounts for ~4% of the total selection response. This is within the confidence limits in the main text, based on the change in SNP frequency (3.6–15.5%) and note that the latter may be biased upwards by ascertainment. However, the exact effect of the allele is difficult to pinpoint in any given generation or population due the nature of the composite trait and change in variance in the composite trait over generations.

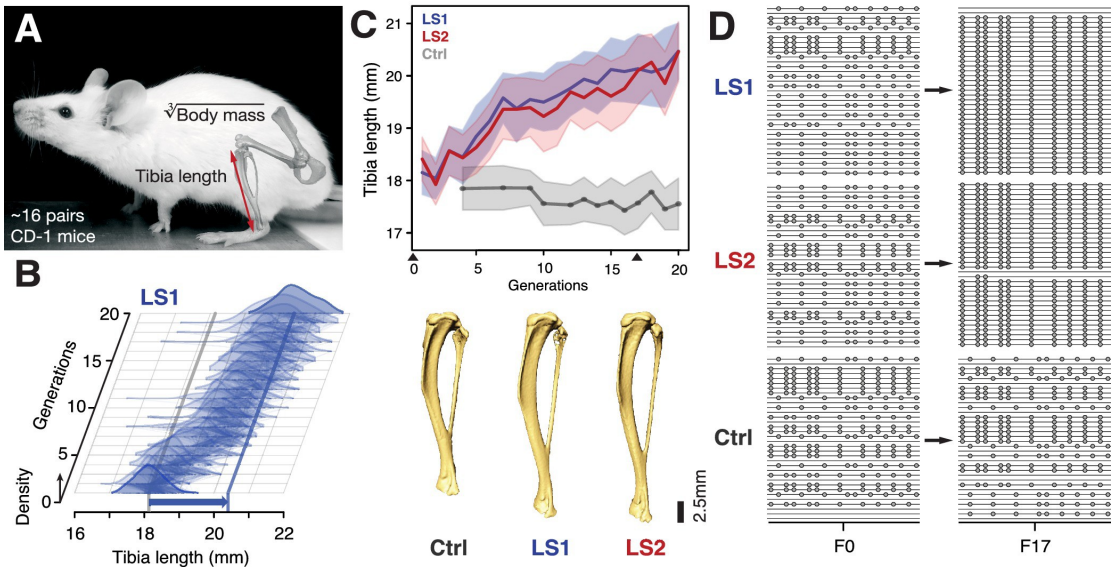


Figure 3.1. Selection for Longshanks mice produced rapid increase in tibia length. (**A** and **B**) Tibia length varies as a quantitative trait among outbred mice derived from the Hsd:ICR (also known as CD-1) commercial stock. Selective breeding for mice with the longest tibiae relative to body mass within families has produced a strong selection response in tibia length over 20 generations in Longshanks mice (13%, blue arrow, LS1). (**C**) Both LS1 and LS2 produced replicated rapid increase in tibia length (blue and red; line and shading show mean ±s.d.) compared to random-bred Controls (gray).Arrowheads along the x-axis mark sequenced generations F0 and F17. See ***Figure 3.2*** for body mass data. Lower panel: Representative tibiae from the Ctrl, LS1 and LS2 after 20 generations of selection. (**D**) Analysis of sequence diversity data (linked variants or haplotypes: lines; variants: dots) may detect signatures of selection, such as selective sweeps (F17 in LS1 and LS2) that result from selection favoring a particular variant (dots), compared to neutral or background patterns (Ctrl). Alternatively, selection may elicit a polygenic response, which may involve minor shifts in allele frequency at many loci and therefore may leave a very different selection signature from the one shown here. DOI: https://doi.org/10.7554/eLife.42014.003

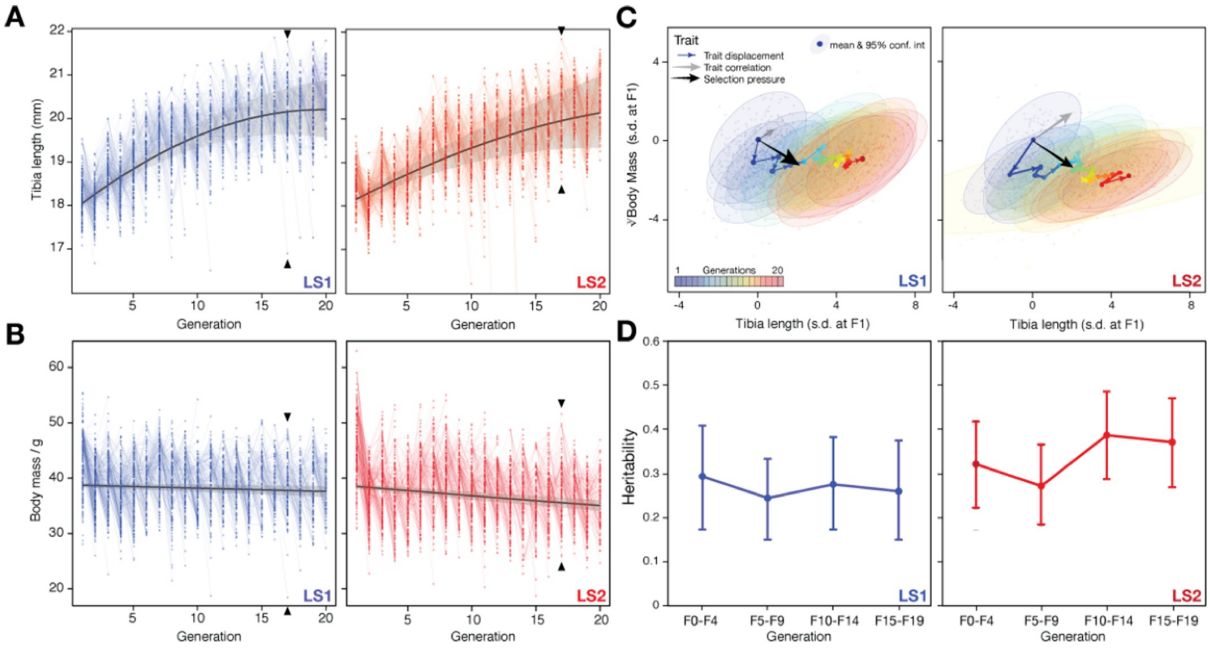


Figure 3.2. Artificial selection allowed detailed reconstruction of selection parameters. Rapid response to selection produced mice with progressively longer tibiae (**A**) and slightly lower body weight (**B**) within 20 generations. Having complete records throughout the selection experiment makes it possible to reconstruct the selection response for both phenotypes and genotypes in detail. Individuals varied in tibia length in both Longshanks lines (LS1, left; LS2, right). Lines connect parents to their offspring. The actual selection depended on the within-family and within-sex rank order of the tibia length-to-body mass (cube root) ratio (see ***Marchini et al., 2014*** for details). The overall selection response was immediate and rapid for tibia length (**A**), suggesting a selection response that depended on standing variation among the founders (black lines show the best fitting quadratic function, with shading indicating 95% confidence interval; adjusted R^2 = 0.61 for LS1; 0.43 for LS2). Strong selection response led to rapid increase in tibia length. In contrast, there was only minor decrease in body weight over the course of the experiment. (**C**) Trajectory in selection response shows decoupling of correlation between tibia length and body mass. Despite overall correlation between tibia length and body mass (gray arrow and major axes in confidence envelopes), cumulative trait displacement over the 20 generations (expressed in s. d. units at F1; arrows, dots and 95% confidence envelopes, color-coded according to generation) showed persistent increase in tibia length with only minor change in body mass along the general direction of selection pressure (black arrows from F1; vector length and directions based on logistic regression). This shows that the Longshanks selection experiment was successful in specifically selecting for increased tibia length while keeping relatively unchanged body mass. (**D**) Despite persistent and strong selection, heritability for the composite trait $ln(\boldsymbol{TB}^{-0.57})$ ($\boldsymbol{T}$ = tibia length in mm; $\boldsymbol{B}$ = cube-root body mass in $\sqrt[3]{}$g (see *Simulating selection response: infinitesimal model with linkage* in main text) was maintained over 20 generations. Heritability was estimated by a linear mixed “animal model” in which the phenotypic variance was partitioned as V_P = fixed effects + V_A + V_R, where fixed effects were sex, age, and litter size, V_A was additive genetic variance, and V_R was residual variance. Heritability was estimated as $h^2 = V_A/(V_A + V_R)$. Each tested block used the full pedigree but only phenotypic information from individuals within the block. We tested an alternate model for each block using truncated pedigrees wherein the first generation of each block was assumed to be unrelated, but found similar results.
https://doi.org/10.7554/eLife.42014.004

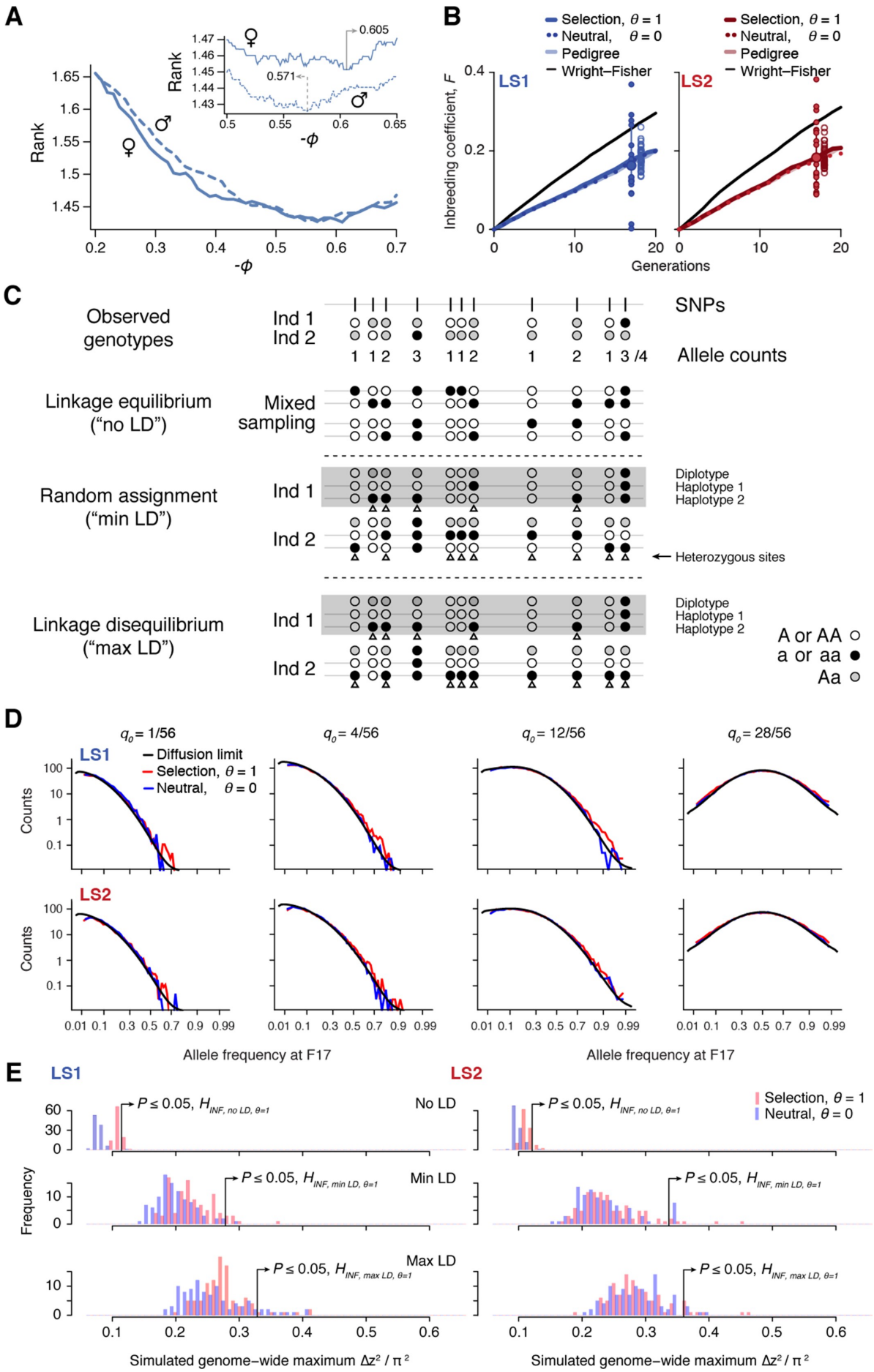

A
Rank
-ϕ
♀
♂
0.605
0.571
B
Selection, θ = 1
Neutral, θ = 0
Pedigree
Wright–Fisher
LS1
LS2
Inbreeding coefficient, F
Generations
C
Observed genotypes
Ind 1
Ind 2
SNPs
Allele counts
Linkage equilibrium ("no LD")
Mixed sampling
Random assignment ("min LD")
Diplotype
Haplotype 1
Haplotype 2
Heterozygous sites
Linkage disequilibrium ("max LD")
A or AA
a or aa
Aa
D
q_0 = 1/56
q_0 = 4/56
q_0 = 12/56
q_0 = 28/56
Diffusion limit
Counts
Allele frequency at F17
E
No LD
Min LD
Max LD
Frequency
$P \le 0.05$, $H_{INF,\ no\ LD,\ \theta=1}$
$P \le 0.05$, $H_{INF,\ min\ LD,\ \theta=1}$
$P \le 0.05$, $H_{INF,\ max\ LD,\ \theta=1}$
Simulated genome–wide maximum $\Delta z^2 / \pi^2$

Figure 3.3. Simulating selection on pedigrees. This figure summarizes the results from our analyses to determine parameters used in the simulations. For full detail, see Appendix, section '*Major considerations in constructing the simulations*'. (**A**) Finding the correct ϕ value for the composite trait $ln(\boldsymbol{TB}^{\phi})$. In each simulated family, offspring are split by sex and ranked by their composite trait. Due to occasional use of back-up crosses, the average rank of actual breeders is greater than 1. We vary ϕ to find the value where actual breeders in the LS lines have the best (lowest) rank. We find ϕ = -0.571 to show the best match for males and -0.605 for females. For subsequent analyses we set ϕ to be -0.57. (**B**) Increase in inbreeding over the course of the Longshanks experiment. The lines show the change in identity between two alleles between diploid individuals, F_b, over 20 generations, as calculated from the pedigree (light shade); the average of 50 neutral simulations without selection (dotted line); or the average of 50 simulation replicates with selection intensity at V_s/V_e = 0.584 (θ=1; thick, dark line). While the F_b trajectories based on pedigree or neutral simulations are indistinguishable, inbreeding increases slightly faster under selection (thick line). The black line shows the increase in identity expected under a Wright–Fisher model with the actual population sizes; under this model, F_w and F_b are close to each other, and to $1-(1-\frac{1}{2N_e})^t$, with N_e equal to the harmonic mean, 24.8. The large dot (with error bar showing the interquartile range among chromosomes) at right show the actual F_b, estimated from the decline in average $2p(1-p)$ over 17 generations. Small filled dots show the estimates from each of the 20 chromosomes. Open dots show 40 replicate simulations, made with the same pedigree and the same selection response θ=1 and sub-sampling from the simulated chromosome according to the actual map length of each of the mouse chromosomes (***Cox et al., 2009***). The simulation agrees well with the observed genome-wide average. Most of the observed data from chromosomes fall within a range comparable to simulated replicates (compare large dot with open dots), with LD being the likely source of this excess variance. (**C**) Three different schemes to seed founder haplotypes. We simulate founder haplotypes that are consistent with observed genotypes (shown here as black, white and gray dots as the two homozygous and the heterozygous states) by directly sampling from founder individuals in each LS line. Under the linkage equilibrium scheme, we sample from the list of allele counts at all SNPs. This produces founder haplotypes that carry no linkage disequilibrium ('no LD'). Under the random assignment scheme, we sample according to each individual (shown as 'diplotypes' within the box for easy comparison). At heterozygote sites in each individual (arrowheads), we randomly assign the alleles to the two haplotypes. This produces founder haplotypes that show minimal LD that is consistent with the observed genotypes ('min LD'). Under the 'max LD' assignment scheme, we also sample according to each individual, except that we consistently assign its haplotypes 1 and 2 with reference (white) and alternate (black) alleles, respectively. This maximizes LD in the founder haplotypes ('max LD'). (**D**) Simulated vs. expected allele frequency shifts. The distribution of minor allele frequencies q_0 at generation 17 is compared with the distribution expected with no selection (blue) or with selection (red), given a frequency of 1, 4, 12 or 28 minor alleles out of 56 founding alleles. The black line shows the diffusion limit, calculated for scaled time $\frac{17}{N_e}$, with N_e estimated to be 51.7 and 48 in LS1 and LS2 respectively, from the rate of increase in F, calculated from the pedigree in panel **A** above. (**E**) Significance threshold values under varying LD from 100 simulated replicates (blue: no selection; red: observed selection response in the actual experiment, θ=1; see panel **C** on LD assignment methods). In order to account for non-independence of adjacent windows due to linkage, a distribution of genome-wide *maximum* Δz^2 was used to determine the significance threshold at each LD level. Δz^2 is the square of arcsine transformed allele frequency difference between F0 and F17; this has an expected variance of $1/2N_e$ per generation, independent of starting frequency, and ranges from 0 to π^2. As seen in previous panels, increasing selection pressure does produce greater shifts in Δz^2 despite using the same pedigree due to a relatively greater proportion of additive genetic variance V_s. However, a far greater impact on Δz^2 is due to changes in LD. This is because weak associations between large numbers of SNP can greatly inflate the variance of Δz^2. Of the three LD levels, 'max LD' likely produced

overly conservative thresholds, whereas ‘min LD' may lead to higher false positives. We have opted conservatively to use maximal LD in our analysis.
https://doi.org/10.7554/eLife.42014.005

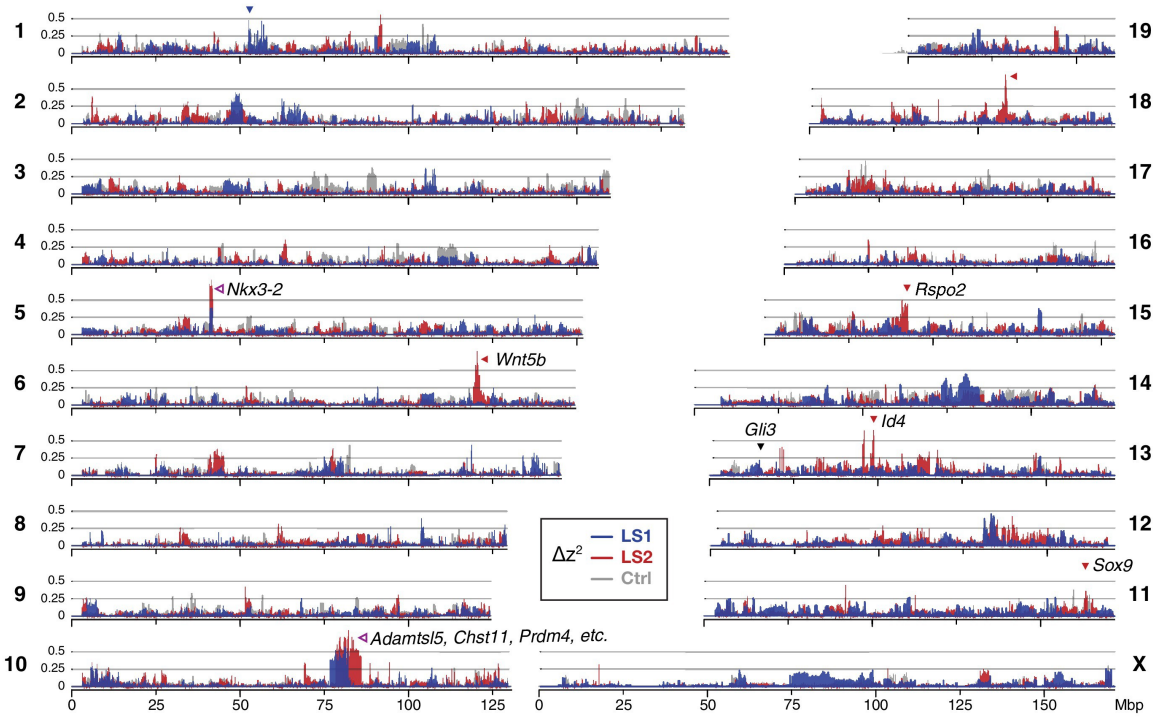
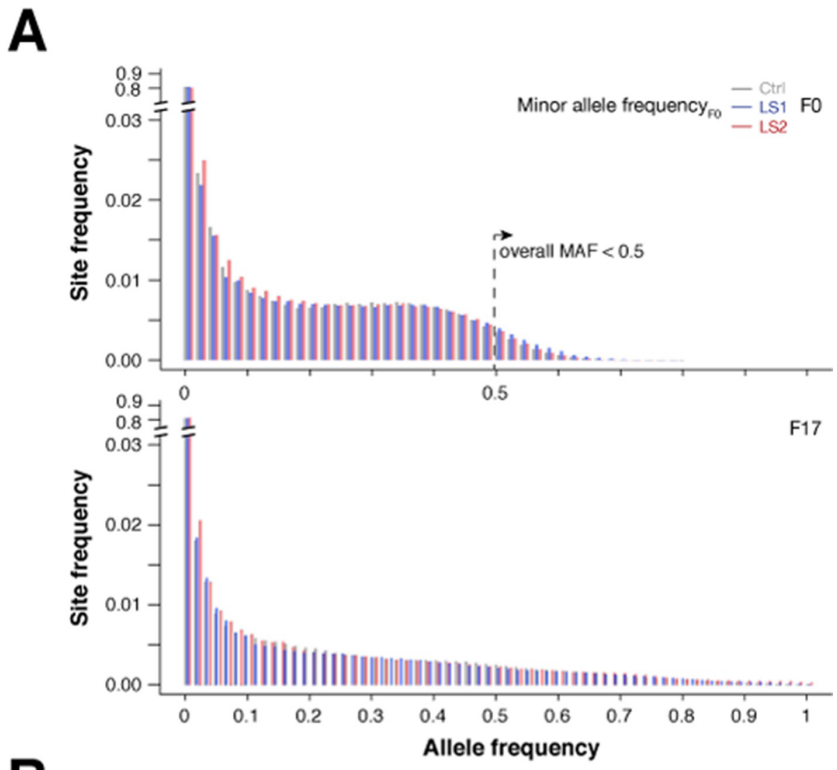
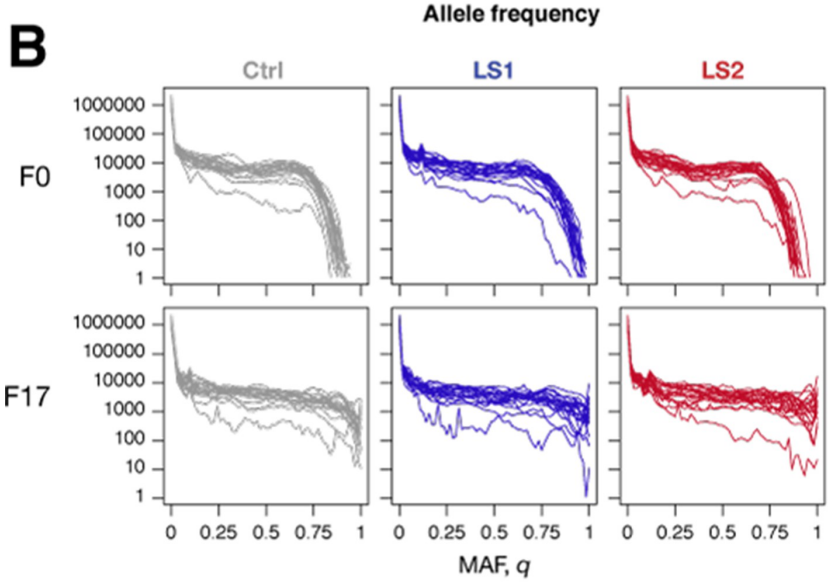


Figure 3.4. Widespread genomic response to selection for increased tibia length. Allele frequency shifts between generations F0 and F17 in LS1, LS2 and Ctrl lines are shown as Δz^2 profiles across the genome (plotted here as fraction of its range from 0 to π^2). The Ctrl Δz^2 profile (gray) confirmed our expectation from theory and simulation that drift, inbreeding and genetic linkage could combine to generate large Δz^2 shifts even without selection. Nonetheless the LS1 (blue) and LS2 (red) profiles show a greater number of strong and parallel shifts than Ctrl. These selective sweeps provide support for the contribution of discrete loci to selection response (arrowheads, blue: LS1; red: LS2; purple: parallel; see also ***Figure 3.3E***, ***Figure 3.6***, ***Figure 3.7***) beyond a polygenic background, which may explain a majority of the selection response and yet leave little discernible selection signature. Candidate genes are highlighted (***Table 3.1***). An additional *a priori* candidate limb regulator *Gli3* is indicated with a black arrowhead.
https://doi.org/10.7554/eLife.42014.006


A
Minor allele frequency F0
Ctrl
LS1
LS2
overall MAF < 0.5
Site frequency
F17
Allele frequency


B
Ctrl
LS1
LS2
F0
F17
MAF, q

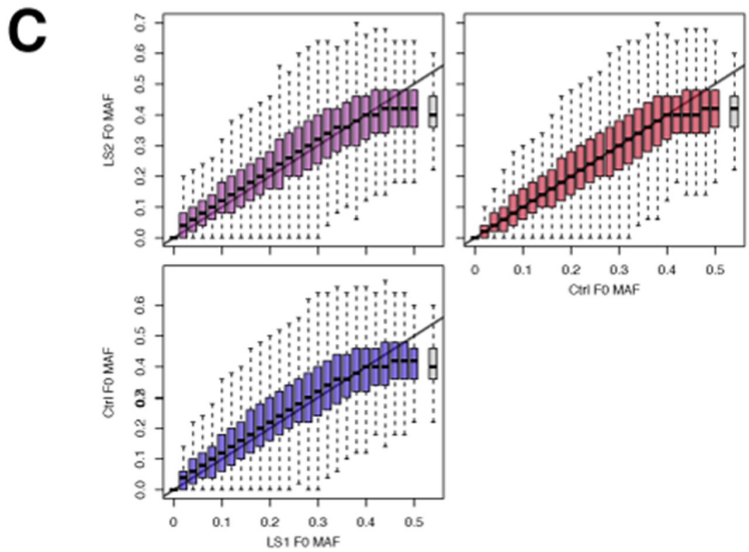

C
LS2 F0 MAF
Ctrl F0 MAF
LS1 F0 MAF

Figure 3.5. Broad similarity in molecular diversity in the founder populations for the Longshanks lines and the Control line. (**A**) Shown are the site frequency spectra from LS1, LS2 and Control lines at F0 (top; folded based on a global minor allele frequency or MAF ≤0.5) and F17 (bottom; unfolded, but tracking the same minor alleles as in F0). Overall the spectra were very similar to each other within each generation. The Control population was mostly intermediate in the decay in the rarer alleles. After 17 generations, the same alleles were generally more spread out, leading to more broadly distributed spectra. There was again little overall difference between the Longshanks and Control lines. (**B**) Variations between chromosomes (separate same-colored lines) shown in each population and generation. The unfolded site frequency spectrum is shown based on the MAF assigned as in A. There is substantial variation between chromosomes, which shows increased distortions in F17. (**C**) Allele frequencies between the founder populations were very similar. Joint minor allele frequencies shown as box plots in 2% bands between the Control and LS1 (blue), LS2 (red); or the two LS lines (purple). Outliers were omitted for clarity. The overall trends follow closely the parity line (gray line along the diagonal), except at frequencies very close to 0.5. Similar to the site frequency spectra in panel A, a small number of sites have a MAF above 0.5 (gray box), because of the use of an overall MAF ≤0.5 to determine minor allele status to enable comparisons across lines. Correlations between all pairwise combinations were around 0.93. https://doi.org/10.7554/eLife.42014.007

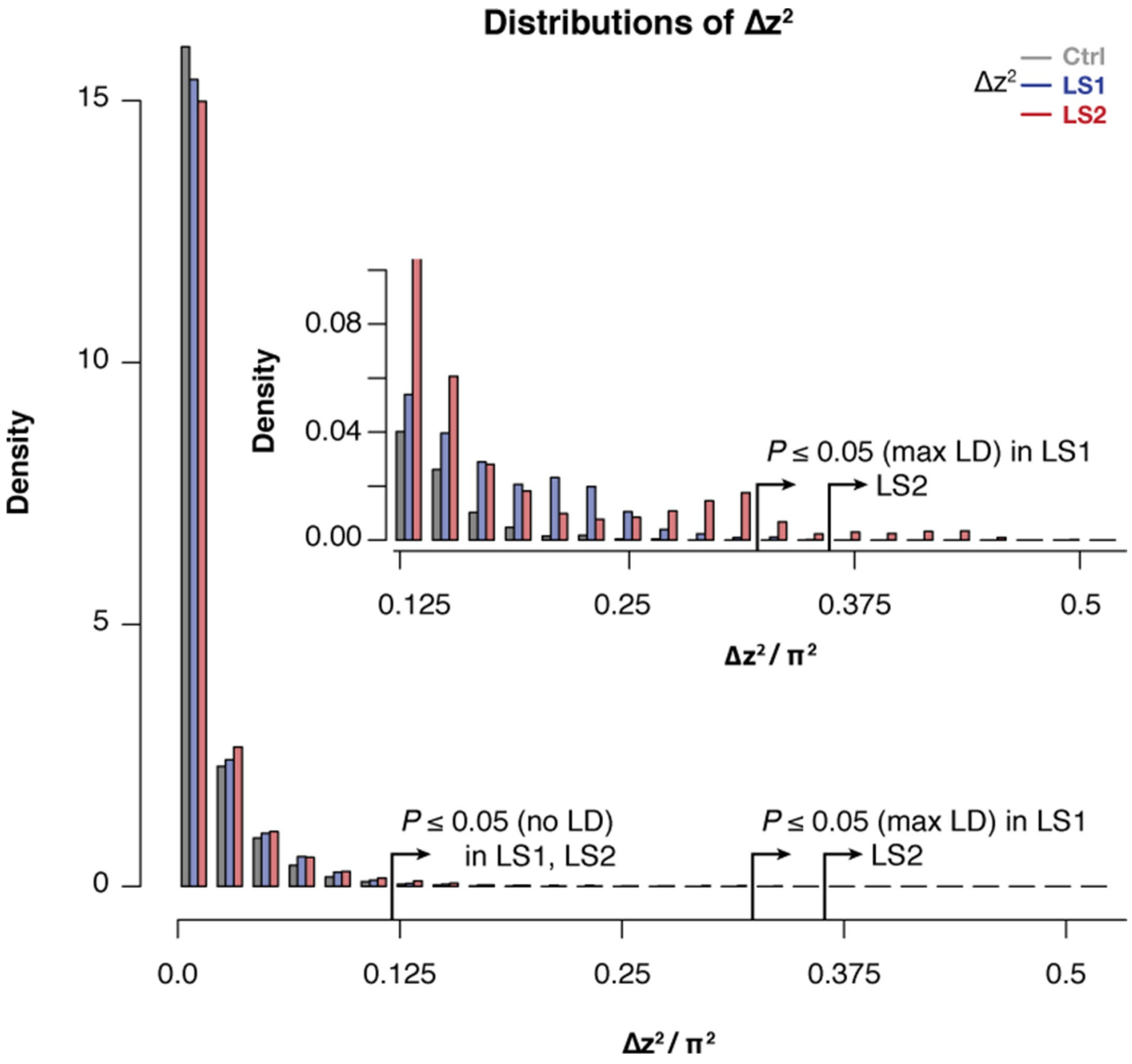


Figure 3.6. Selected lines showed more extreme values of Δz2 than the Control line. Histogram of within-line Δz^2 values in 10 kbp windows across the genome in LS1, LS2, and Control. Overall similarity is high across all three lines, but there was an excess of large Δz^2 value starting from as low as <0.1 π^2. This pattern becomes clearly distinct above the threshold value of 0.125, which corresponds to the lenient significance threshold p≤0.05 under $H_{INF,\ no\ LD}$ (inset). There was clearly an excess of windows in LS2 above the more stringent p≤0.05 threshold under $H_{INF,\ max\ LD}$. This excess supports discrete loci contributing to selection response in LS2 that give rise to greater distortion of Δz^2 spectra.
https://doi.org/10.7554/eLife.42014.008

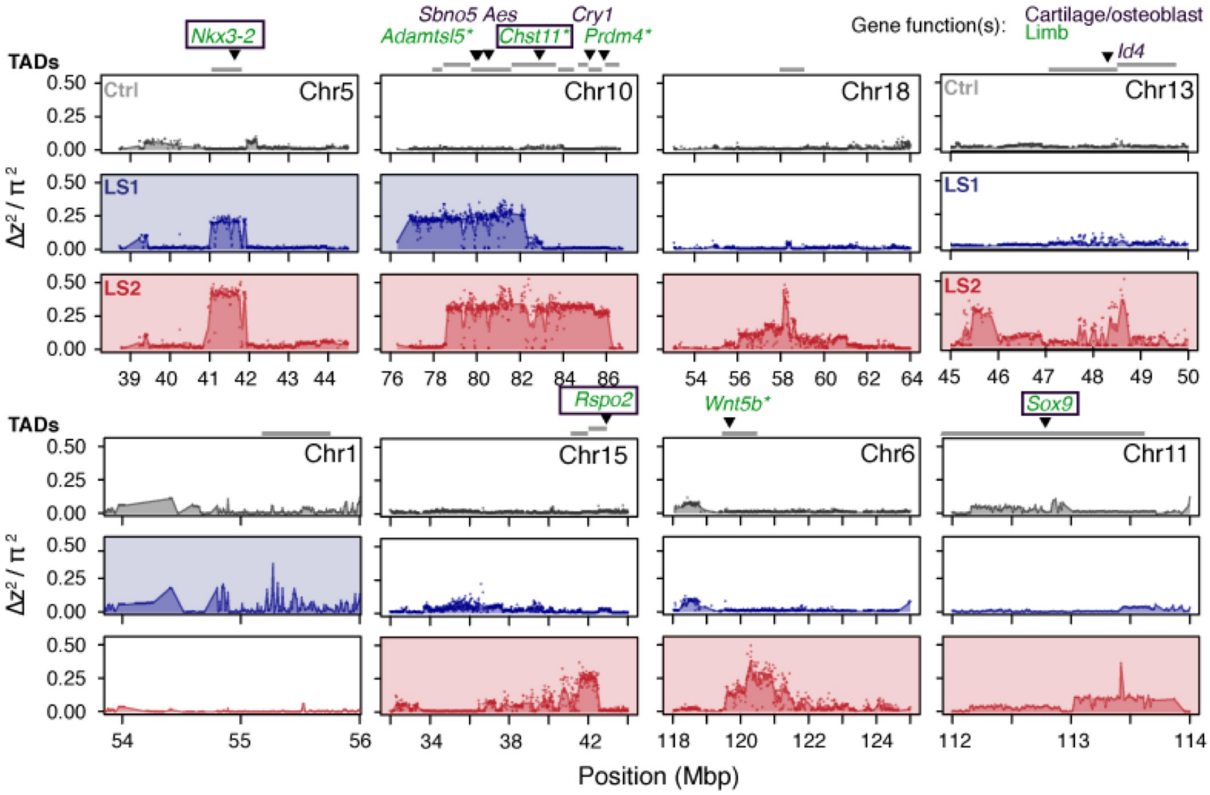


Figure 3.7. Detailed Δz^2 profiles at the 8 Longshanks significant loci. For each significant locus, Δz^2 profiles are shown for Ctrl (gray), LS1 (blue) and LS2 (red). Plots are shaded if the locus is significant in a given line. TADs within 250 kbp of the significant signals are shown as gray bars above each locus. Above the TADs are highlighted genes whose knockout phenotypes belong to the following categories: 'abnormal tibia morphology', 'short limb', 'short tibia', 'abnormal cartilage morphology', 'abnormal osteoblast morphology'. The gene symbols are colored according to the gene function(s) in limb development (green), bone development (purple) or both (boxed). Gene symbols marked by asterisks (*) have specifically reported 'short tibia' or 'short limb' knockout phenotypes. All of the above categories show significant enrichment at the eight loci (number of genes per category: 4–7, nominal p≤0.03, see Appendix, section '*Enrichment for genes with functional impact on limb development*' for details on the permutation), except 'abnormal cartilage morphology', with four genes and a nominal *P*-value of 0.083. No overlap was found with any gene in these categories from the three significant loci from the Ctrl line.
https://doi.org/10.7554/eLife.42014.009

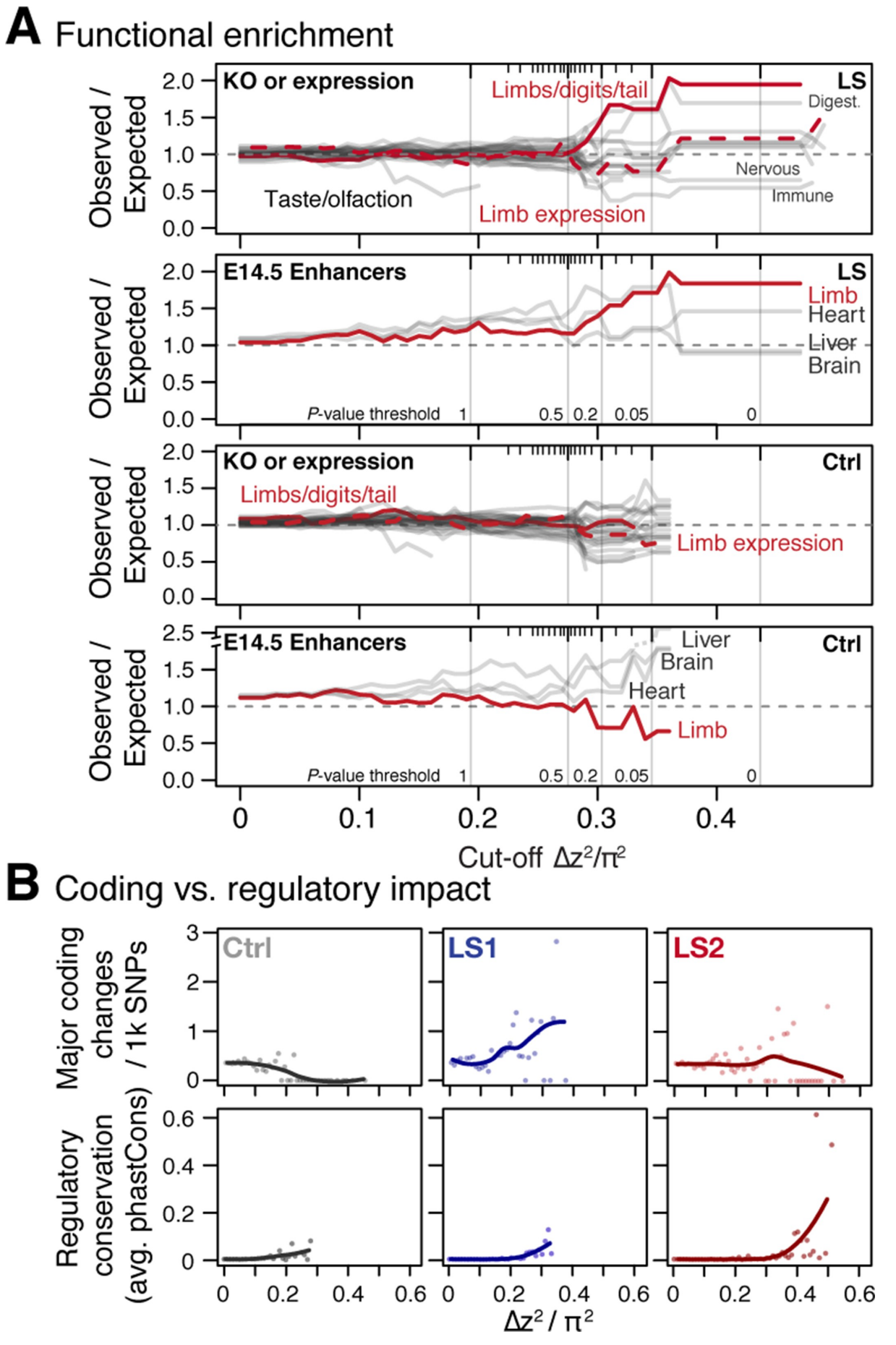

A Functional enrichment
KO or expression
Limbs/digits/tail
LS
Digest.
Nervous
Immune
Taste/olfaction
Limb expression
E14.5 Enhancers
LS
Limb
Heart
Liver
Brain
P-value threshold 1 0.5 0.2 0.05 0
KO or expression
Ctrl
Limbs/digits/tail
Limb expression
E14.5 Enhancers
Ctrl
Liver
Brain
Heart
Limb
P-value threshold 1 0.5 0.2 0.05 0
Observed / Expected
Cut-off Δz²/π²
B Coding vs. regulatory impact
Ctrl
LS1
LS2
Major coding changes / 1k SNPs
Regulatory conservation (avg. phastCons)
Δz² / π²

Figure 3.8. Loci associated with selection response in Longshanks lines show enrichment for limb function likely associated with *cis*-acting mechanisms. (**A**) Gene set enrichment analysis of knock-out phenotypes (KO) showed that selection response (here shown as Δz^2 cut-off values, see Supplementary Methods for details on cut-off values and inclusion criteria) were found among topologically associating domains (TADs) containing limb and tail developmental genes (red solid lines) or genes with limb expression (red dotted lines) in LS lines (top) but not in Ctrl (bottom). Among KO phenotypes, limb defects show the greatest excess out of 28 phenotypic categories (other gray lines, with other extreme categories labeled, the 'normal' category is excluded here). Among developmental enhancers for limb, heart, liver and brain tissue, we also observed an association with Δz^2 peaks in LS lines (top) for limb but not in Ctrl lines (bottom). The simulated significance thresholds based on $H_{INF,\ max\ LD}$ are also shown for reference (vertical gray lines). The data from the LS lines suggest that enrichment started to increase around the p≤0.5 threshold and remained largely stable at p≤0.05, corresponding to a cut-off of around 0.33 π^2. (**B**) Coding vs. regulatory impact. Frequency of moderate to major coding changes (top panels, amino acid changes, frame-shifts or stop codons), or average conservation score of regulatory sequences immediately flanking SNPs (based on conservation among 60 eutherian mammals; bottom panels) were used as proxies to estimate the functional impact of coding and regulatory mutations, respectively. In LS1, major coding changes became more common at high Δz^2 ranges; however the number of SNPs with potentially major phenotypic consequences did not increase in LS2 and in fact seemed to decrease in Ctrl. In contrast, regulatory changes showed increased conservation associated with greater allele frequency shifts or Δz^2 in all three lines, except that SNPs with large shifts and strong conservation were more abundant in LS1 and LS2. Trend lines are shown with LOESS regression but statistical comparisons were performed using linear regressions.
https://doi.org/10.7554/eLife.42014.010

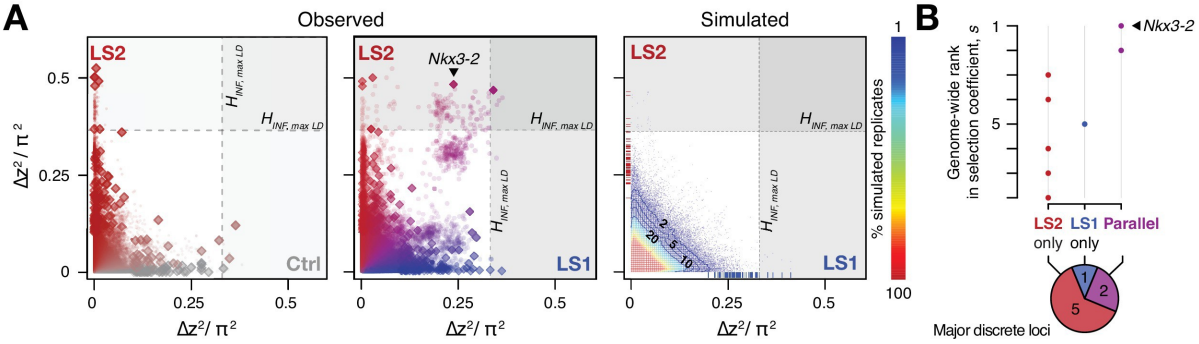


Figure 3.9. Selection response in the Longshanks lines was largely line-specific, but the strongest signals occurred in parallel. (**A**) Allele frequencies showed greater shifts in LS2 (red) than in Ctrl (gray; left panel; diamonds: peak windows; dots: other 10 kbp windows; see ***Figure 3.10*** for Ctrl vs. LS1 and Appendix for details). Changes in the two lines were not correlated with each other. In contrast, there were many more parallel changes in a comparison between LS1 (blue) vs. LS2 (red; middle panel; adjacent windows appear as clusters due to hitchhiking). The overall distribution closely matches simulated results under the infinitesimal model with maximal linkage disequilibrium ($H_{INF,\ max\ LD}$; right heatmap summarizes the percentage seen in 100 simulated replicates), with most of the windows showing little to no shift (red hues near 0; see also ***Figure 3,10*** for an example replicate). Tick marks along the axes show genome-wide maximum Δz^2 shifts in each of 100 replicate simulations in LS1 (x-axis, blue) and LS2 (y-axis, red), from which we derived line-specific thresholds at the p≤0.05 significance level. While the frequency shifts from simulations matched the bulk of the observed data well, no simulation recovered the strong parallel shifts observed between LS1 and LS2 (compare middle to right panel, points along

the diagonal). (**B**) Genome-wide ranking based on estimated selection coefficients s among the candidate discrete loci at p≤0.05 under $H_{INF,\ max\ LD}$. While six out of eight total loci showed significant shifts in only LS1 or LS2, the two loci with the highest selection coefficients were likely selected in parallel in both LS1 and LS2 (also see middle panel in A).
https://doi.org/10.7554/eLife.42014.012

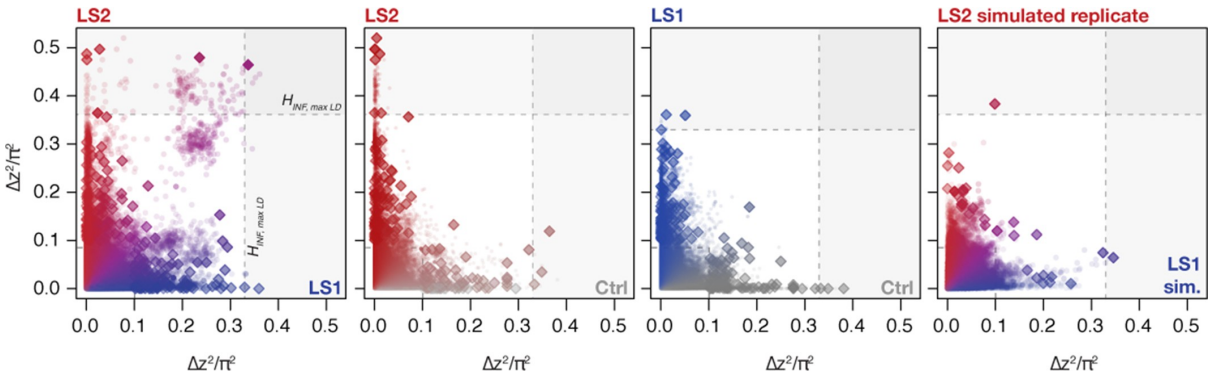
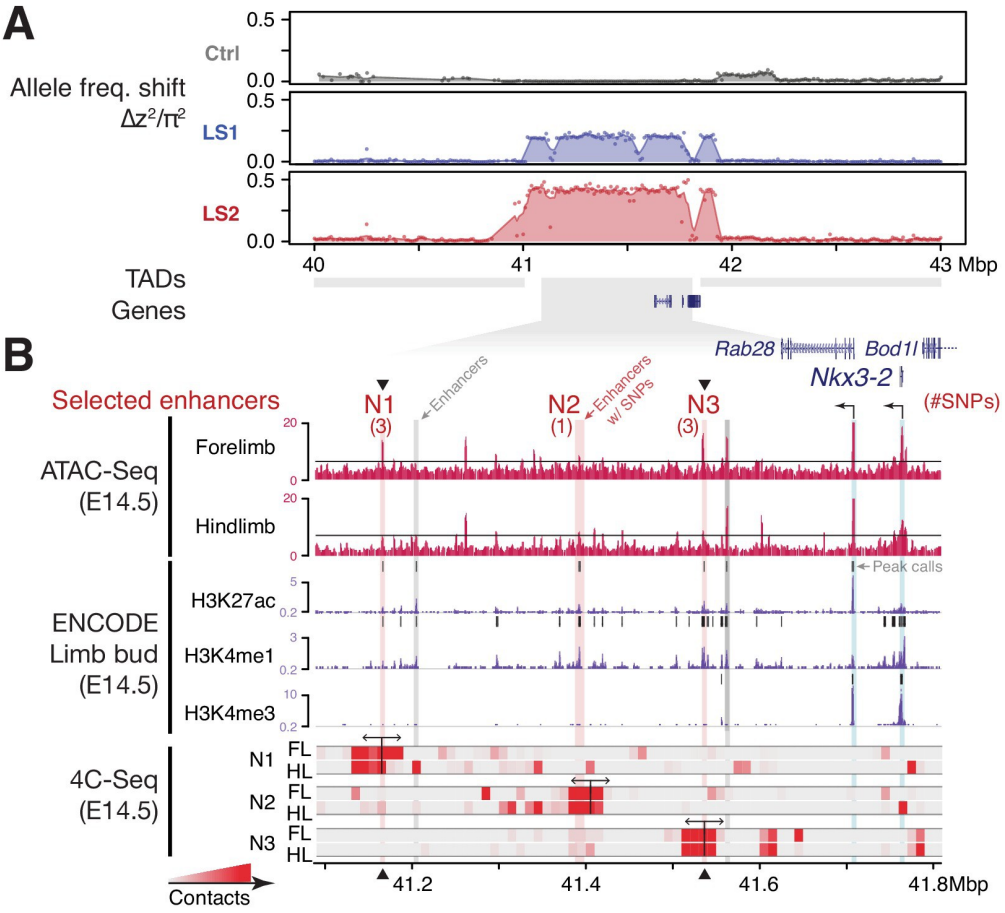


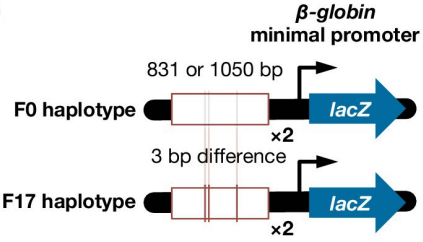
Figure 3.10. Changes in Δz^2 across lines. Shown are changes in Δz^2 in individual 10 kbp windows (all windows: circles; peak windows: diamonds). Generally there were no clear differences in Δz^2 along the axes except a slight skew toward higher values in LS2. When taken as a joint LS1–LS2 comparison, however, we observed that many windows show shifts in both LS1 and LS2 (left panel; in purple). In contrast, very few windows show parallelism in Ctrl–LS2 and Ctrl–LS1 comparisons (middle two panels). The right panel shows a single selected simulated replicate (selection pressure V_s/V_e =0.58; maximum LD) found to have among the greatest extent of parallel Δz^2 among the replicates. The excess in parallel loci in observed results is clear both among the significant loci at p≤ 0.05 under $H_{INF,\ max\ LD}$ and highly significant at the more relaxed H_{INF}, no LD threshold.
https://doi.org/10.7554/eLife.42014.013



C

Enh	Position	Alleles Sel/Alt		Allele frequency LS1 Gen F0	LS1 Gen F17	LS2 Gen F0	LS2 Gen F17	Predicted transcription factor (TF) binding sites	TF role	Impact
	41,166,187	G	A	0.19	0.84	0.14	0.97			
N1	41,166,264	C	G	0.15	0.81	0.12	0.97	Hox⊘12	+	-
	41,166,505	A	G	0.13	0.88	0.12	0.98			
N2	41,392,871	A	G	0.17	0.88	0.18	0.82	Zic1/2/3 & Gli3	-/(+)	-/(+)
	41,536,250	G	T	0.17	0.83	0.10	0.98	**2x Nkx⊘3-2**	++	----
N3	41,536,431	G	A	0.23	0.83	0.14	0.98			
	41,536,498	C	A	0.23	0.86	0.16	0.98	Nkx⊘-1,Tcf/Lef1	+/~?	-

D



E

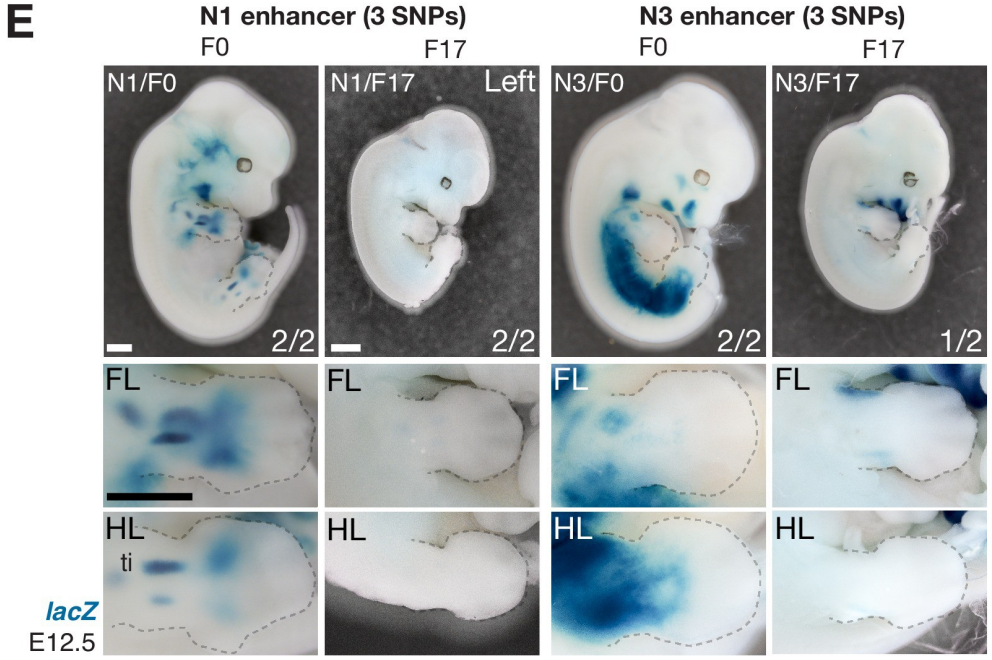


Figure 3.11. Strong parallel selection response at the bone maturation repressor *Nkx3-2* locus was associated with decreased activity of two enhancers. (**A**) Δz^2 in this region of chromosome five showed strong parallel differentiation spanning 1 Mbp in both Longshanks but not in the Control line. This 1 Mbp region contains three genes: *Nkx3-2*, *Rab28* and *Bod1l* (whose promoter lies outside the TAD boundary, shown as gray boxes). Although an originally rare allele in all lines, this region swept almost to fixation by generation 17 in LS2 (***Figure 3.15A***). (**B**) Chromatin profiles [ATAC-Seq, red, (***Buenrostro et al., 2013***); ENCODE histone modifications, purple] from E14.5 developing limb buds revealed five putative limb enhancers (gray and red shading) in the TAD, three of which contained SNPs showing significant frequency shifts. Chromosome conformation capture assays (4C-Seq) from E14.5 limb buds from the N1, N2 and N3 enhancer viewpoints (bi-directional arrows) showed significant long-range looping between the enhancers and sequences around the *Nkx3-2* promoter (heat-map from gray to red showing increasing contacts; Promoters are shown with black arrows and blue vertical shading). (**C**) Selected alleles at 7 SNPs found within the N1, N2, and N3 enhancers increased ~0.75 in frequency in both LS1 and LS2. Selected alleles at three of these sites are predicted to lead to loss (red inhibition circles) of transcription factor binding sites in the *Nkx3-2* pathway (including a SNP in N3 causing loss of two adjoining *Nkx3-2* binding sites) and thus reduce enhancer activity in N1 and N3. (**D, E**) Transient transgenic reporter assays of the N1 and N3 enhancers showed that the F0 alleles drove robust and consistent expression at centers of future cartilage condensation (N1) and broader domains of *Nkx3-2* expression (N3) in E12.5 fore- and hind limb buds (FL, HL; ti: tibia). Fractions indicate the number of embryos showing similar *lacZ* staining out of all transgenic embryos. Substituting the F17 enhancer allele (i.e., replacing three positions each in N1 and N3) led to little observable limb bud expression in both the N1/F17 and N3/F17 embryos, suggesting that selection response for longer tibia involved de-repression of bone maturation through a loss-of-function regulatory allele of *Nkx3-2* at this locus. Scale bar: 1 mm for both magnifications.
https://doi.org/10.7554/eLife.42014.014

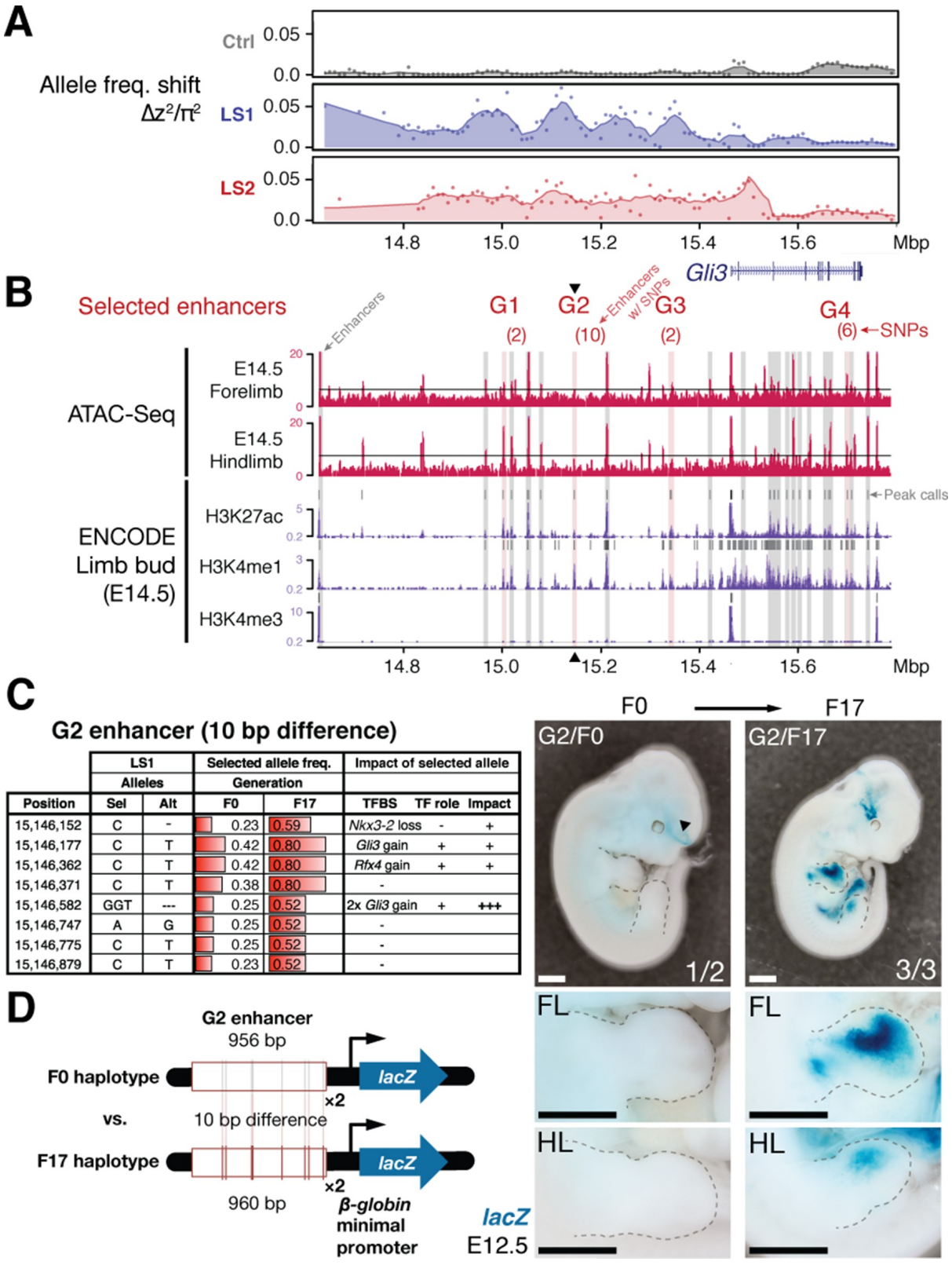


Figure 3.12. An enhancer in chromosome 13 boosts *Gli3* expression during limb bud development. (**A**) LS1 showed elevated Δz^2 in the intergenic region containing *Gli3*. (**B**) Putative limb enhancers (gray bars) were identified through peaks from ATAC-Seq (top) and histone modifications (bottom tracks, data from ENCODE project). Four of the enhancers contain mutations (in parentheses) with significant allele frequency shifts between F0 and F17 in LS1 (red shading). One of the enhancers located close to the peak Δz^2 signal (G2,

arrowhead) containing 10 bp differences was chosen for transgenic reporter assay. (**C**) Analysis of individual mutations showed an average increase of 0.33 in allele frequency, with six mutated positions affecting predicted binding of transcription factors in the *Gli3* pathway (including three additional copies of *Gli3* binding sites), all of which are predicted to boost the G2 enhancer activity. (**D**) The F17 G2 enhancer variants together drove robust and consistent *lacZ* reporter gene expression at E12.5, recapitulating *Gli3* expression in the developing fore- and hind limb buds (right; see also ***Figure 3.13***). Substitution of 10 positions (F0 haplotype) led to little observable expression in the limb buds (left). These G2 enhancer gain-of-function mutations (contrasting the major allele between F0 and F17) may confer an advantage under selection for increased tibia length. Scale bars: 1 mm for both magnifications. https://doi.org/10.7554/eLife.42014.015

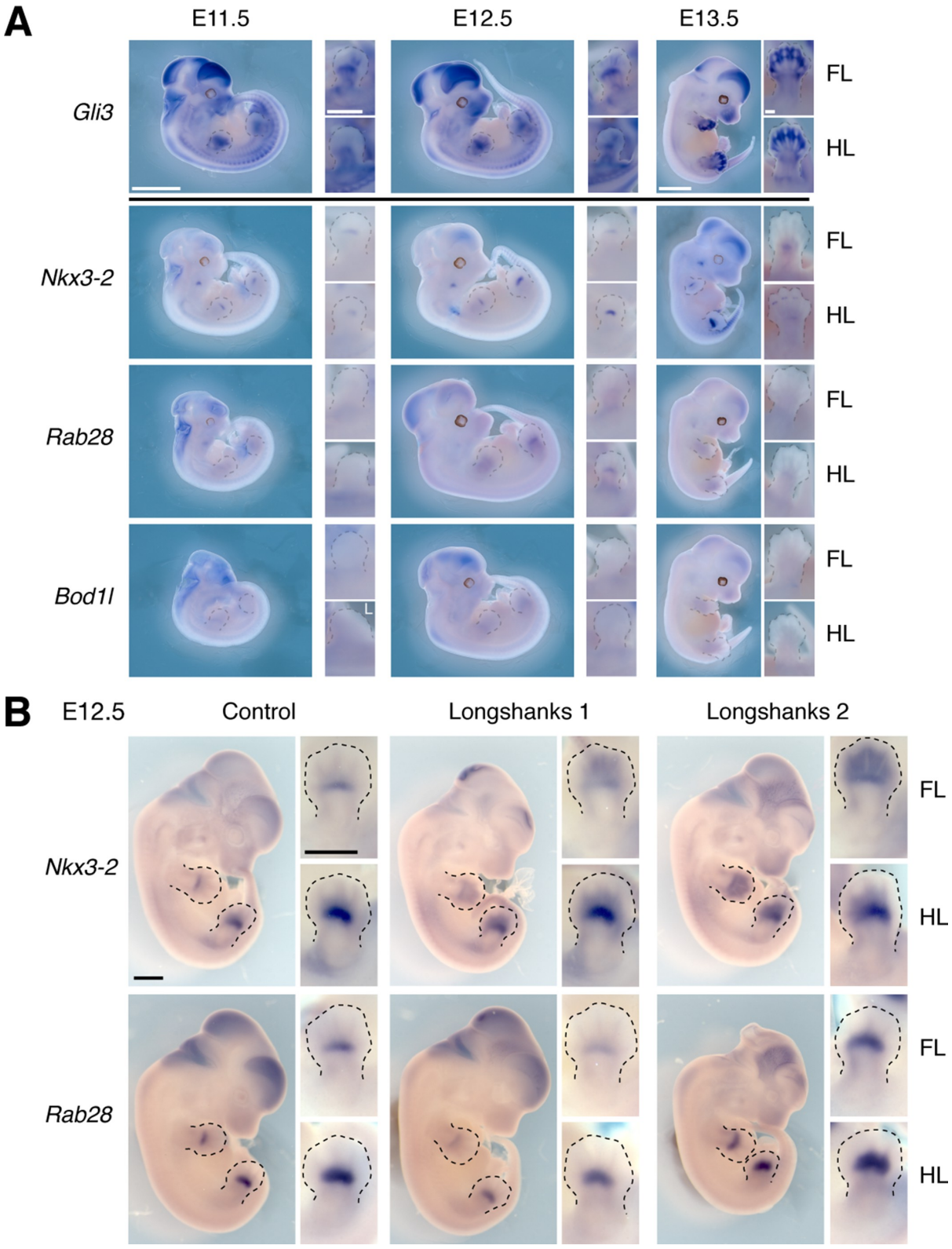


Figure 3.13. Gene expression patterns at the *Gli3* and *Nkx3-2* candidate intervals. (**A**) *Gli3* expression was determined using *in situ* hybridization. *Gli3* was robustly expressed during limb development in both developing fore- and hindlimb buds, especially in the autopod (hand/foot plate). Lower panels show expression of *Nkx3-2* and its neighboring genes *Rab28* and *Bod1l*. The stronger expression of *Nkx3-2* in the developing limb buds as well as the known role of *Nkx3-2* in bone maturation (***Sivakamasundari et al., 2012***) strongly argues for *Nkx3-2* being the gene underlying the selection response at the

chromosome 5 locus. Scale bars: 1 mm for whole-mounts; 0.5 mm for limb buds. FL, forelimb; HL, hind limb; unless otherwise indicated by 'L', all images were taken from the right side. (**B**) We collected E12.5 embryos from each line and performed in situ hybridization to determine the sites and level of expression of *Nkx3-2* and *Rab28* in the Longshanks (right columns) and Control (left column) lines. Both genes are expressed in similar sites overall and specifically in the developing fore- and hind limb buds in the region of the presumptive zeugopods. These data indicate common sites of expression and rule out qualitative presence/absence differences in expression. Note that although the N1 enhancer pattern appear to differ from endogenous *Nkx3-2* expression (***Figure 3.11E***, details in limb buds), it matches the pattern of early *Nkx3-2* expression, detectable through the use of lineage-tracing via the combined effect of a *Nkx3-2 Cre*-driver line and the *lacZ* reporter line *Rosa26R* (see Figure 2B in ***Sivakamasundari et al., 2012***).
https://doi.org/10.7554/eLife.42014.016

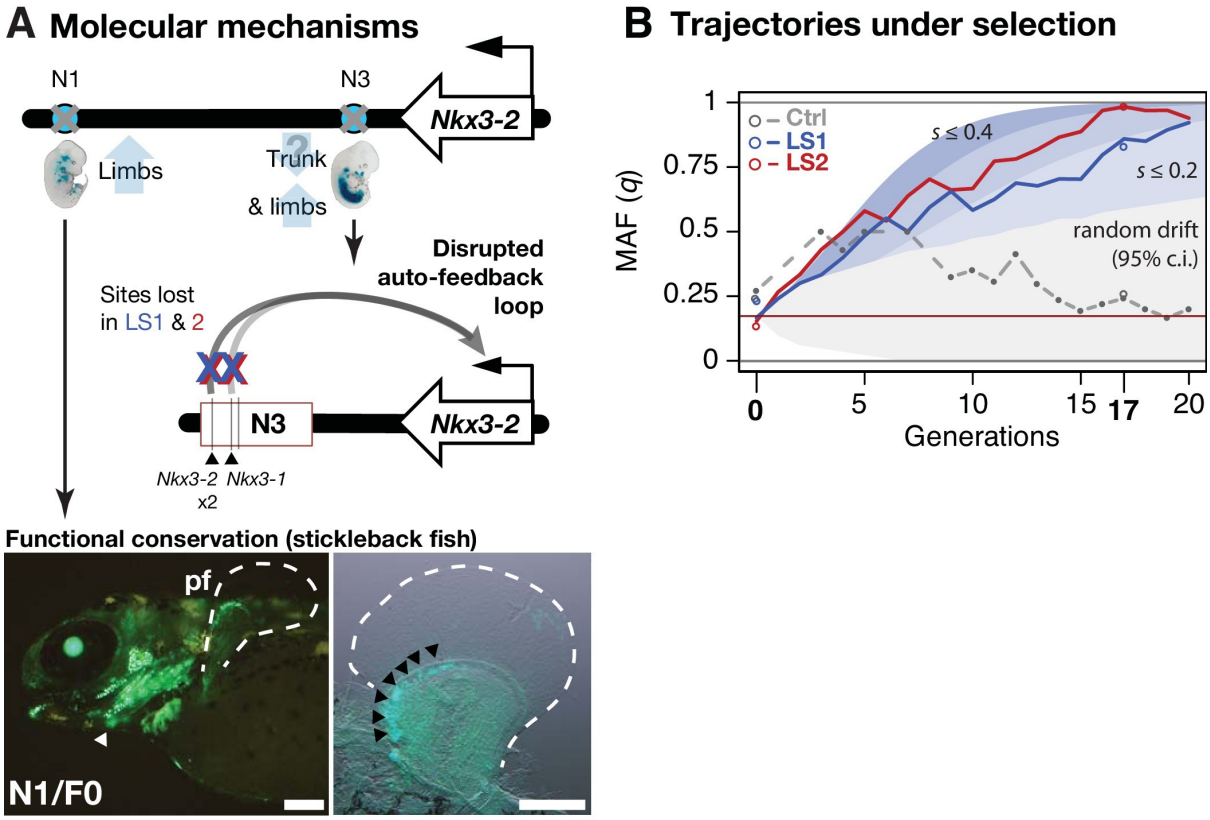


Figure 3.14. Linking base-pair changes to rapid morphological evolution. (**A**) At the *Nkx3-2* locus, we identified two long-range enhancers, N1 and N3 (circles), located 600 and 230 kbp away, respectively. During development, they drive partially overlapping expression domains in limbs (N1 and N3) and trunk (N3), which are body regions that may correlate positively (tibia length) and possibly negatively (trunk with body mass) with the Longshanks selection regime. For both enhancers, the selected F17 alleles carry loss-of-function variants (gray crosses). Two out of three SNPs in the N3 F17 enhancer are predicted to disrupt an auto-feedback loop, likely reducing *Nkx3-2* expression in the trunk and limb regions. Conversely, the enhancer function of the strong N1 F0 allele is evolutionarily conserved in fishes, demonstrated by its ability to drive consistent GFP expression (green) in the pectoral fins (pf, outlined) and branchial arches (white arrowhead, left) in transgenic stickleback embryos at 11 days post-fertilization. The N1 enhancer can recapitulate *nkx3.2* expression in distal cells specifically in the endochondral radial domain in developing fins (black arrowheads, right). Scale bar: 250 μm for both magnifications. (**B**) Allele frequency of the

selected allele (minor allele at F0, q) at N3 over 20 generations (blue: LS1; red: LS2; gray broken line: Ctrl; results from N1 were nearly identical due to tight linkage). Observed frequencies from genotyped generations in the Ctrl line are marked with filled circles. Dashed lines indicate missing Ctrl generations. Open circles at generations 0 and 17 indicate allele frequencies from whole genome sequencing. The allele frequency fluctuated in Ctrl due to random drift but followed a generally linear increase in the selected lines from around 0.17 to 0.85 (LS1) and 0.98 (LS2) by generation 17. Shaded contours mark expected allelic trajectories under varying selection coefficients starting from 0.17 (red horizontal line; the average starting allele frequency between LS1 and LS2 founders). The gray shaded region marks the 95% confidence interval under random drift.
https://doi.org/10.7554/eLife.42014.017

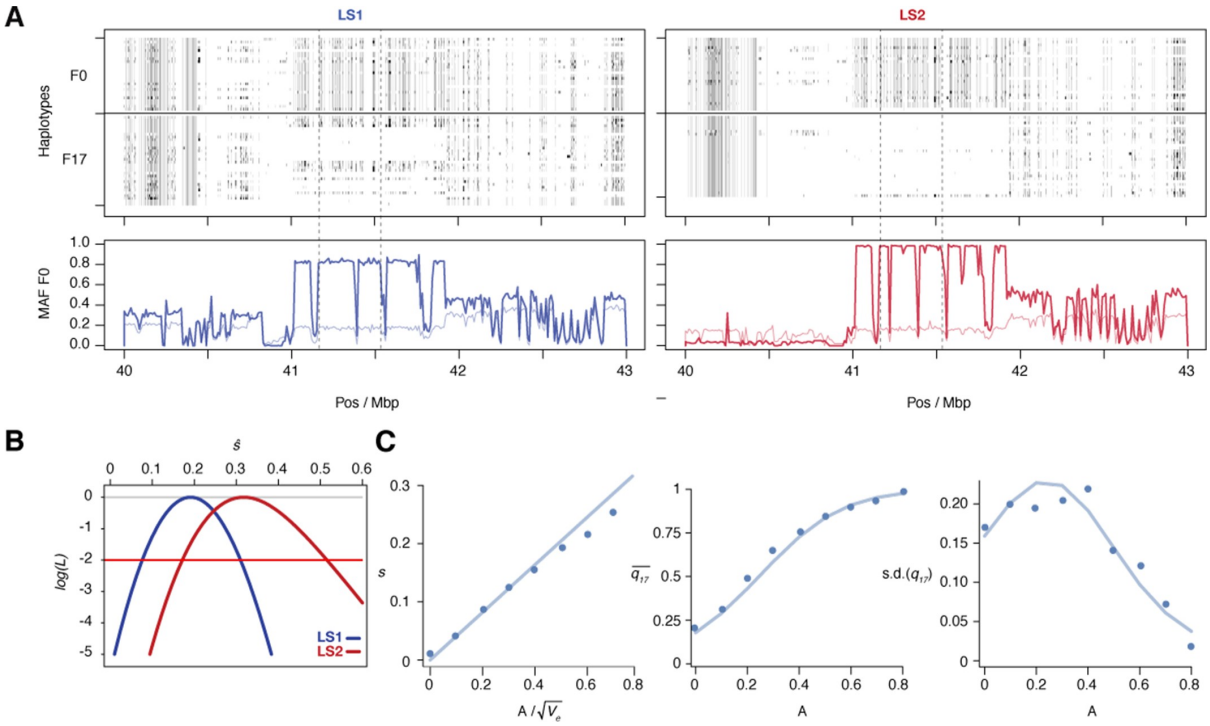


Figure 3.15. Selection at the *Nkx3-2* locus. (**A**) Raw genotypes from the F0 and F17 generations from LS1 (left) and LS2 (right) are shown, clearly indicating the area under the selective sweep. The genotype classes are shown as C57BL/6J homozygous (BL6, white), heterozygous (black) and alternate homozygous (dark gray). Lower Panel: Tracking MAF from both lines show that the originally rare F0 allele (thin line) rose to high frequency at F17 (thick lines). The plateau profile from both lines suggested that the same originally rare allele was segregating at in both founder populations and became very common by F17 in both lines (see raw genotypes). Note that in LS2 F17 the region is nearly fixed for the BL6 allele except in the bottom-most individual. (**B**) The log likelihood of the selection coefficient, *s*, for LS1 and LS2 (blue and red, respectively), based on transition probabilities for a Wright–Fisher population with the appropriate N_e. The horizontal red line shows a loss of log likelihood of 2 units, which sets conventional 2-unit support limits. (**C**) Simulations of an additive allele with effect *A* on the trait; 40 replicates for each value of *A*. Left: The selection coefficient, estimated from the change in mean allele frequency, as a function of $A/\sqrt{V_e}$; the line shows the least-squares fit $s=0.41A/\sqrt{V_e}$. Middle: dots show the mean allele frequency at generation 17; the line shows the prediction from the single-locus Wright–Fisher model, given $s=0.41A/\sqrt{V_e}$. Right: the same, but for the standard deviation of allele frequency.
https://doi.org/10.7554/eLife.42014.018

Table 3.1. Major loci likely contributing to the selection response. These eight loci show significant allele frequency shifts in Δz^2 and are ordered according to their estimated selection coefficients according to ***Haldane (1932)***. Shown for each locus are the full hitchhiking spans, peak location and their size covering the core windows, the overlapping TAD and the number of genes found in it. The two top-ranked loci show shifts in parallel in both LS1 and LS2, with the remaining six showing line-specific response (LS1: 1; LS2: 5). Candidate genes found within the TAD with limb, cartilage, or bone developmental knockout phenotype functions are shown, with asterisks (*) marking those with a 'short tibia' knockout phenotype (see also ***Figure 3.7***and ***Supplementary file 3*** for full table).
https://doi.org/10.7554/eLife.42014.011

							Δq				
Rk	Chr	Span (Mbp)	Peak	Core (kbp)	TAD (kbp)	Genes	LS1	LS2	Ctrl	Type	Candidate genes
1	5	38.95–45.13	41.77	900	720	*3*	0.69	0.86	−0.14	Parallel	*Nkx3-2*
2	10	77.47–87.69	81.07	5360	6520	*175*	0.79	0.88	−0.04	Parallel	*Sbno5, Aes, Adamtsl5*, Chst11*, Cry1, Prdm4**
3	18	53.63–63.50	58.18	220	520	*4*	0.05	0.78	−0.06	LS2-specific	-
4	13	35.59–55.21	48.65	70	2600	*22*	0.24	0.80	−0.03	LS2-specific	*Id4*
5	1	53.16–57.13	55.27	10	720	*4*	0.65	0.01	−0.23	LS1-specific	-
6	15	31.92–44.43	41.54	10	680	*3*	−0.23	0.66	0.02	LS2-specific	*Rspo2**
7	6	118.65–125.25	120.30	130	1360	*12*	−0.03	0.79	−0.15	LS2-specific	*Wnt5b**
8	11	111.10–115.06	113.42	10	2120	*2*	−0.14	0.66	−0.15	LS2-specific	*Sox9**

Rk, Rank.
Chr, Chromosome.
Core, Span of 10 kbp windows above $H_{INF,\ max\ LD}$ p≤0.05 significance threshold.
TAD, Merged span of topologically associating domains (TAD) overlapping the core span. TADs mark segments along a chromosome that share a common regulatory mechanism. Data from ***Dixon et al. (2012)***.
Candidate genes, Genes within the TAD span showing 'short tibia', 'short limbs', 'abnormal osteoblast morphology' or 'abnormal cartilage morphology'
knockout phenotypes are listed, with * marking those with 'short tibia'.
DOI: https://doi.org/10.7554/eLife

Chapter 4: General Discussion

Which factors in the nuclear environment enable transcriptional regulation by enhancers? Attempts to characterize the biochemical and physical properties that determine enhancer output have pointed to the interplay between transcription factor (TF) binding, histone modification marks, and chromatin stability and looping (Calo and Wysocka, 2013; Yanez-Cuna et al., 2014; Long et al., 2016; Bu et al., 2017). These factors are influenced both by one another and by intrinsic features of enhancers, including GC content, DNA methylation (Bu et al., 2017), and the presence of dinucleotide repeats (Spitz and Furlong, 2012; Yanez-Cuna et al., 2014; Bagshaw, 2017; Colbran et al., 2017; Carelli et al., 2018), making the determination of enhancer activity highly complex. In this discussion, I consider the evidence from the literature as well as from my work on the Capture-C and Longshanks experiments (**Chapters 2-3**) to reconstruct the chain of events that could lead to expression differences at the *Nkx3-2* and *Gli3* loci in the Longshanks mice.

4.1 Transcription factor binding

TFs bind enhancers to displace nucleosomes and recruit cofactors, which interact with RNA Polymerase II and transcriptional machinery to elevate transcription at promoters (Long et al., 2016). TF binding is difficult to predict *in silico* and impractical to comprehensively assay *in vivo*, as there are hundreds of different TFs in mammals and they tend to be highly cell type-specific (Khamis et al., 2018) and occupy only a small proportion of predicted binding sites (Grossman et al., 2017). Multiple TFs bind to each enhancer in a manner that is dependent on the spacing and order of binding motifs as well as on the interactions between the bound TFs (Spitz and Furlong, 2012; Stefflova et al., 2013; Farley et al., 2015; Long et al., 2016).

4.2 Histone modification marks

TF binding both affects and is regulated by histone modifications, although the latter are not always required for enhancer activity (Henikoff and Shilatifard, 2011; Calo and Wysocka, 2013). TF binding recruits histone-modifying enzymes, many of which double as co-activators of TFs (Xin and Rohs, 2018). The active enhancer mark H3K27ac, by neutralizing the positive charge of the lysine residue, helps lower the affinity of the nucleosome for the DNA, making it easier for TFs to bind (Bao and

Bedford, 2016). H3K4 methylation, which often precedes H3K27 acetylation (Bonn et al., 2012) and is associated with poised enhancers, enables enhancer activity by protecting the DNA from *de novo* methylation and subsequent gene silencing, by recruiting cofactors that in turn recruit TFs and chromatin remodeling complexes, and by interacting with other modified histone tails to regulate chromatin stability (Cao and Wysocka, 2013). A host of other histone modification marks, often in combination, have been found to be associated with enhancer activity (Karlic et al., 2010). In some contexts, however, TF binding occurs despite a lack of histone modification marks near the enhancer (Ghisletti et al., 2010; Heinz et al., 2010; McManus et al., 2011; Spitz and Furlong, 2012). In transgenic or massively parallel reporter assays, enhancer reporter activity is subject to positional effects such as the presence or absence of histone modifications at the site of genomic integration (Akhtar et al., 2013). In assays that rely on episomal instead of genomic integration, including STARR-seq (Arnold et al., 2013), enhancer activity is likewise affected by the histone modifications that form in episomal chromatin (Riu et al., 2007; Barde et al., 2009; Muerdter et al., 2015; Inoue et al., 2017). In short, modification of nearby histones can both facilitate and result from TF binding, though the exact cause and consequence is unclear and may depend on the context; a comprehensive assay of histone modification function has not been possible thus far due to the challenges associated with knockdown or mutagenesis of histones or the enzymes which catalyze their modification (Calo and Wysocka, 2013).

4.3 Long-distance regulatory mechanisms

Enhancers tend to act distally to their promoters, often from up to hundreds of kilobase pairs away (Kleinjan and van Heyningen, 2005). How do they recruit transcriptional machinery over long distances? Several mechanisms have been proposed. The chromatin looping model is best supported by experimental evidence, particularly from chromosome conformation capture methods (Li et al., 2012). TF binding to enhancers and subsequent binding of chromatin remodelers to TFs and their cofactors help initiate chromatin loops (Matharu and Ahituv, 2015). Subsequently, loops are maintained by loop extrusion factors like cohesin, which bring distal loci into physical proximity with one another as they translocate along the DNA (Fudenberg et al., 2016). Although cohesin can move freely over nucleosomes (Davidson et al., 2016), initial formation of the loop is influenced by histone

modification marks that regulate chromatin flexibility (Li et al., 2006). Enhancers which contact their promoters through chromatin loops may not require chromatin loops to have precise lengths: knockout of the cohesin release factor WAPL in a human cell line led to longer median chromatin loops but only affected expression of around 1,000 genes genome-wide (Haarhuis et al., 2017). Depleting cohesin from mouse immune cells, however, resulted in down-regulation of many otherwise highly expressed genes located near enhancers, underscoring the role of chromatin looping in enabling enhancer activity (Ing-Simmons et al., 2015). The prevalence of chromatin loops is further underscored by the finding that the majority of characterized promoter-enhancer contacts in mouse embryonic stem cells lie within topologically associating domains (TADs) (Shen et al., 2012; Schoenfelder et al., 2015), which were discovered by Hi-C, a technique that captures genome-wide chromatin loops.

In some cases, enhancer activity can proceed without chromatin looping: if the chromatin between an enhancer and its target promoter is sufficiently compact, the TFs recruited by enhancers may already be localized to the promoters (Pennacchio et al., 2013). Likewise, if an enhancer is sufficiently close to its target promoter, then chromatin looping is also not required (and may not even be possible due to constraints on chromatin flexibility). This is the case in transgenic reporter assays and in STARR-seq (Arnold et al., 2013), where the enhancer is directly adjacent to a minimal promoter. As an alternative to chromatin looping, enhancers may be brought into proximity with their promoters through the action of RNA Polymerase II, which binds to TFs at the enhancer and scans along the chromatin until it encounters the target promoter (Blackwood and Kadonaga, 1998). This “tracking” or “scanning” model is thought to describe only a minority of observed interactions (Calo and Wysocka, 2013; Meng and Bartholomew, 2018). In summary, while enhancers may rely on other methods to target and act on promoters (Pennacchio et al., 2013), chromatin looping is a widespread mechanism among endogenous enhancers.

4.4 Determinants of enhancer output at *Nkx3-2* and *Gli3* in the Longshanks mice

I will now turn to applying the findings from my exploration of transcription regulation in the Capture-C study to the Longshanks selection experiment,

specifically to interpret the interplay between various factors at the candidate regions we molecularly dissected. We identified candidate regions as those that had significant allele frequency shifts from F0 to F17 and looked at the genes in these regions, taking into account their function. At *Nkx3-2* and *Gli3*, because we did not see coding changes or observed only mild conservative changes, it stands to reason that gene regulatory mechanisms like enhancers would be the most likely explanation. Therefore, we focused on identifying enhancers and determining whether any functional differences could be under selection. We required putative enhancers to be located in open chromatin regions (as identified by ATAC-seq) near the candidate genes and to be enriched for H3K27ac and H3K4me1 but not for the classic promoter mark, H3K4me3. With 4C-seq, we confirmed that the putative enhancers interacted with the candidate gene promoters (**Figure 3.11B**). We then selected enhancers to test in *lacZ* transgenic reporter assays (**Figure 3.11D-E**), prioritizing those that contained SNPs with major allele frequency differences. Through differential TF binding site predictions from SNP intersection, we found that the SNPS that were swept to higher allele frequency by F17 were predicted to have resulted in loss (at *Nkx3-2*) or gain (at *Gli3*) of TF binding sites. We hypothesized based on the transgenic assays, differential TF binding site predictions, and known roles of *Nkx3-2* as a limb repressor and of *Gli3* as a limb growth activator that the SNPs in the tested enhancers led to lower expression of *Nkx3-2* (loss-of-function) and higher expression of *Gli3* (gain-of-function) in F17 mice.

The ChIP-seq (ENCODE Project Consortium, 2012), ATAC-seq, and 4C-seq experiments were performed in C57BL/6NJ (BL6) mice, whose haplotype at each of the SNP-containing enhancers matches the haplotype of the allele under selection.

Nkx3-2

At the weaker (BL6) F17 allele in the *Nkx3-2* enhancers, the chromatin is accessible and the active and poised enhancer marks (H3K27ac and H3K4me1) are present. Our 4C-seq showed that each enhancer engages in a chromatin loop with the *Nkx3-2* promoter. However, the F17 enhancers were not observed to stimulate expression of the reporter gene in E12.5 limb bud, despite the fact that each allele was cloned and injected as a tandem duplicate, which we would expect to strengthen enhancer output without changing the expression domain based on results from Farley et al., 2015 (**Figure 3.11D-E**). The lack of F17 reporter gene

expression may be due to technical aspects of the transgenic reporter assays. For example, we may not have screened enough transgenic embryos to establish a consistent *lacZ* expression pattern above and beyond the integration site effect. This is because the reporter transgene may be silenced or suppressed by histones at the integration site or may not be strong enough to recruit histone-modifying enzymes to add the appropriate marks (Akhtar et al., 2013). Another possibility is that the genetic background of the transgenic embryos may cause the enhancer to behave differently than it would in the Longshanks genetic background. Additionally, the transgenic reporter assays were screened at E12.5, whereas the other data is from E14.5 limb bud. The tested F17 enhancers might not yet be active in E12.5 limb bud, although we detected *Nkx3-2* expression as early E12.5.

If the transgenic reporter assays do reflect the activity of the endogenous enhancers in the Longshanks mice – namely, that the enhancers are weakened in F17 – then the striking difference in enhancer activity from F0 to F17 could be attributed to the predicted differential TF binding sites within the N1 and N3 enhancers. Without ChIP-seq data, it is difficult to determine actual TF occupancy at the enhancers, as combinatorial binding by multiple TFs, motif grammar, and other factors affect binding (Spitz and Furlong, 2012; Stefflova et al., 2013) such that only a small proportion of predicted binding sites are occupied *in vivo* (Grossman et al., 2017). However, from differential TF binding site predictions we found that at the N1 enhancer, the SNPs swept to higher frequency by F17 were predicted to have ablated a predicted binding site of the *hoxd12* TF (**Figure 3.11C**). SNPs in the N3 enhancer were predicted to have resulted in loss of two binding sites of *nkx3-2*, which is itself a TF. Since *nkx3-2*, like *hoxd12*, likely activates *Nkx3-2* transcription, the potential impact on enhancer activity may be amplified through a negative feedback loop whereby *nkx3-2* binds less frequently to its own enhancer. Loss or reduction of occupancy by these transcriptional activators may reduce enhancer activity either directly (if they are the primary or sole TFs to confer enhancer activity) or by impacting the combinatorial binding of additional TFs (if the additional TFs require co-binding by the transcriptional activators in order to themselves bind, or if the full set of TFs must be present in order to stimulate enhancer activity). A study of combinatorial TF binding across different species revealed that cobound TFs were sensitive to one another's binding and tended to be lost in tandem (Stefflova et al., 2013); loss of binding by TFs at N1 and N3 may decrease co-binding by other TFs.

Once TF binding is lost or reduced, the enhancers may not be able to stimulate transcription (or may still do so, but at such a low level as to be undetectable in a transgenic reporter assay) despite the presence of histone marks like H3K27ac and H3K4me1. These marks, while strongly associated with enhancers, are not sufficient for enhancer activity; not all active enhancers are marked by H3K27ac (Taylor et al., 2013), and H3K4me1 decorates gene bodies in addition to poised enhancers (Calo and Wysocka, 2013). The enzymes that add the histone modification marks are recruited through TF binding. Assuming the Longshanks F17 mice have the same H3K4me1 and H3K27ac ChIP-seq peaks present in BL6, how could these marks have been added? The transcriptional activators whose binding sites were altered by SNPs in F17 may still be able to bind the enhancers and recruit histone-modifying enzymes, albeit possibly with lower efficiency due to the more transient nature of binding. Many enhancers have been observed to function through sub-optimal binding affinity, and it is even a mechanism whereby enhancer activity can be restricted (Farley et al., 2015, 2016). Furthermore, if there are additional TFs with binding sites not impacted by the SNPs and they continue to bind to the enhancer sequence, then they could also have recruited the histone-modifying enzymes. A wide range of TFs can recruit the p300/CBP histone acetyltransferase that adds H3K27ac, for example (Goodman and Smolik, 2000).

According to our 4C-seq data, the F17 enhancers contact the *Nkx3-2* promoter. Even if the enhancers are engaged in a chromatin loop with the promoter, without TFs recruiting transcriptional machinery to the promoter, they should not be able to stimulate gene expression. Chromatin looping of an enhancer to a promoter without an accompanying increase in gene expression has been observed in the developing mouse limb bud at the *Sonic hedgehog* promoter: in cells of the anterior limb bud, no *Shh* expression was detected despite contact between a distal limb enhancer and the promoter (Amano et al., 2009). In this work, the authors hypothesized that the chromatin conformation in the anterior limb bud cells existed in a poised state whereby activation of the gene was primed by enhancer proximity through chromatin looping. This is compatible with the model proposed in the Capture-C analysis (**Chapter 2**) that suggests that promoters with intermediate expression exist in a poised state where they form contacts less constrained by distance than active or silent promoters. In F17, *Nkx3-2* behaves more like the poised than the active class of promoters. It continues to form interactions with the

now hypomorphic N1 and N3 enhancers but is not regulated by them because although they are marked by hallmark enhancer features like chromatin accessibility and H3K27ac and H3K4me1, they may lack the critical TF binding that confers functionality in the F0 mice.

In situ hybridization in Longshanks embryos from the 20th generation onwards still showed *Nkx3-2* expression (**Figure 3.13B**). Due to the semi-quantitative nature of *in situ* hybridization (Wunderlich et al., 2014), it is not possible to quantify whether *Nkx3-2* expression was lower in the Longshanks embryos than in the Control (not exposed to selection) embryos from the same generation. However, given the known role of *Nkx3-2*as a limb growth repressor, the presumed loss or weakening of transcriptional activator binding in the F17 enhancers, and the failure of the F17 enhancers to drive readily detectable *lacZ* reporter expression, we hypothesized that *Nkx3-2* expression decreased in the Longshanks mice upon selection, contributing to longer tibia length by reducing *Nkx3-2* limb growth repression. This hypothesis is consistent with the longer limb length seen in human patients who have frameshift mutations in *Nkx3-2* (Hellemans et al., 2009), and also with the shorter limb length seen in chicken embryos that overexpress *Nkx3-2* (Bren-Mattison et al., 2011). These known phenotypic impacts of *Nkx3-2* mis-expression suggest that the expression decrease we hypothesized to have happened by F17 is not negligible. While there may be other causal mutations underlying longer tibia length in Longshanks mice, the allele frequency shift at *Nkx3-2* had the strongest selection coefficient, and the changes in the enhancers arose in parallel across the two independently selected lines.

How is *Nkx3-2* expression maintained in the context of hypomorphic N1 and N3 enhancers? From the *in situ* hybridization, the expression domains of *Nkx3-2* do not visibly differ between Control and Longshanks embryos. The most parsimonious explanation is that in the F17 context, the N1 and N3 enhancers are simply weaker but still able to stimulate transcription, and that the presumed reduction of *Nkx3-2* expression is solely quantitative as opposed to spatial and thus not detectable by *in situ* hybridization. However, endogenous *Nkx3-2* showed expression in the distal (fingertip) portion of the limb bud, whereas the F0 N1 and N3 enhancers were not observed to stimulate *lacZ* expression in these cells. The absence of observable reporter gene expression in the fingertips could be a result of intrinsic features of the reporter gene assay, such as differences in how the endogenous *Nkx3-2* promoter

and the β-globin minimal promoter respond to stimulation by enhancers. Additionally, N1 and N3 may be acting in tandem in the endogenous context, possibly by forming, within the same chromatin loop, a “pocket” of *cis*- and *trans*-regulatory elements (Kragesteen et al., 2018), to stimulate more cell types to express *Nkx3-2* than each one can stimulate on their own.

There could also be other enhancers stimulating expression which we did not functionally test due to their lack of SNPs with a significant allele frequency shift, or, as at the N2 putative enhancer, a lack of SNPs with a predictable impact on TF binding sites. Our 4C-seq showed that the F17 version of the N2 enhancer also contacted the *Nkx3-2* promoter. In mammals, promoters tend to be regulated by multiple enhancers with similar expression domains (de Laat and Duboule, 2013; Osterwalder et al., 2018). These enhancers may include functionally redundant “shadow” enhancers (Hong et al., 2008) that stimulate transcription in the same cell types as do the N1 and N3 enhancers, but possibly less efficiently such that overall *Nkx3-2* expression is still lower in F17. Osterwalder et al., 2018 observed that deleting individual enhancers had no visible effect on target gene expression, and only observed a phenotypic impact when enhancers were deleted in pairs. Likewise, deletion of a hindlimb-specific distal enhancer of *Pitx1* had no observable effect on hindlimb morphology or *Pitx1* expression domains and resulted in only a small reduction of *Pitx1* expression level in the hindlimb, suggesting it is functionally redundant with other enhancers (Sarro et al., 2018). Enhancers may also function in an additive manner whereby each one contributes partially or non-overlapping expression domains, and knockout of one enhancer, such as the *PelB* enhancer at the *Pitx1* promoter in mouse hindlimb (Thompson et al., 2018), decreases but does not abolish target gene expression. Arnold et al., 2013 observed a correlation between gene expression level and the sum of enhancer strength as derived from a STARR-seq assay in *Drosophila*, suggesting enhancers acted on the same gene in an additive or redundant manner. If F17 N1 and N3 are hypomorphic, then other enhancers with additional or overlapping expression domains may at least partially compensate for their lack of activity. A CRISPR/Cas9-mediated knockout of N1 and N3 followed by *Nkx3-2 in situ* hybridization could reveal whether other enhancers can maintain *Nkx3-2* limb bud expression.

We noted that the F0 N1 and N3 enhancers stimulated *lacZ* expression in the proximal part of the limb bud, whereas endogenous *Nkx3-2* was not visibly

expressed here according to the *in situ* hybridizations. However, *Nkx3-2*-driven *Cre* in a cell lineage-tracing experiment from Sivakamasundari et al., 2012 was expressed in proximal as well as distal cells of the limb bud. This suggests that the transgenic reporter assay may have revealed subtler expression patterns not detectable by *in situ* hybridization due to its limited sensitivity (Speel et al., 1999); that is, the endogenous *Nkx3-2* promoter most likely does stimulate proximal limb bud expression.

There is disparity, however, between the F0 N3 *lacZ* expression domain and the *Nkx3-2* expression pattern captured by the in-situ and *Cre* cell-lineage tracing experiments. The F0 N3 enhancer stimulates reporter gene expression strongly in most cells of the proximal hindlimb bud, whereas proximal hindlimb expression is more scant in the cell-lineage and *in situ* hybridization E12.5 embryos. One potential reason the N3 enhancer can stimulate expression outside of the endogenous *Nkx3-2* domain is the presence of multiple enhancers acting in a combinatorial or synergistic manner to restrict the expression domain of N3. Synergistic regulation involving multiple enhancers can explain why enhancer and gene expression domains do not always match, a phenomenon also observed in large-scale transgenic reporter assays (Kragesteen et al., 2018). In the developing mouse limb bud and brain, enhancers of the *Fgf8* promoter stimulated reporter gene expression in additional cell types beside those observed to express *Fgf8*. Barring integration site effects, the expression domain of one enhancer may be restricted by the activity of other enhancers (Marinic et al., 2013). At the *Pitx1* locus, the *Pen* enhancer is expressed both in the developing forelimb and the hindlimb, yet *Pitx1* is only expressed in the hindlimb (Kragesteen et al., 2018). In cells of the proximal limb bud in F0 mice, other enhancers at the endogenous locus may hinder N3 from stimulating *Nkx3-2* transcription, which they would otherwise be able to do if acting in isolation, as in a transgenics reporter assay.

What mechanism might underlie such synergy? Chromatin conformation may sequester enhancers from their target promoters in one cell type but not in another (Kragesteen et al., 2018). Possibly, there are additional enhancers or *cis*-regulatory elements like insulators at the endogenous locus, which loop out or otherwise shield the N3 enhancer from the *Nkx3-2* promoter in many of the proximal hindlimb cells. Performing 4C-seq with the *Nkx3-2* promoter as the viewpoint could reveal other contacts that help to regulate *Nkx3-2* expression. These contacts and the

transcriptional activity hubs they may form likely take place entirely within the same TAD. In the Capture-C analysis (**Chapter 2**), I found that the majority of promoter contacts at most of the viewpoints were located within the same mESC TAD (Dixon et al., 2012) as the viewpoint, confirming previous findings in mESCs that promoter-enhancer interactions tend to take place within the same TAD (Shen et al., 2012; Schoenfelder et al., 2015). I also observed, in accordance with previous studies, that TADs merge over developmental time and during cellular differentiation (Meshorer and Misteli, 2006; Melcer and Meshorer, 2010; Gaspar-Maia et al., 2011; Battulin et al., 2015; Boya et al., 2017). However, the question of whether to use TAD boundaries from mESCS (Dixon et al., 2012) or from cortex cells (Shen et al., 2012) does not appear to be an issue in the Longshanks 4C-seq assay even though the limb bud cells (E14.5) are one day more advanced than the E13.5 limb bud cells from the Capture-C experiment, because the *Gli3* and *Nkx3-2* TADs share the same coordinates in mESCs as in cortex cells.

In the limb bud, the N1 through N3 enhancers appear to act solely on the *Nkx3-2* promoter and are not shared between multiple promoters as has been reported at certain other loci (Andrey et al., 2017). From our 4C-seq data, these enhancers were not observed to frequently contact the *Bod1l* promoter, which is located 80 kbp telomeric to the *Nkx3-2* promoter and is in the adjacent mESC TAD. They were additionally not observed to frequently contact the *Rab28* promoter, which *is* within the same mESC TAD and 60 kbp centromeric to the target promoter – thus closer to the enhancers and more likely to be affected by proximity ligation in the 4C-seq assay.

Gli3

Which molecular mechanisms at *Gli3* enabled the presumed increase in *Gli3* expression from F0 to F17? At the *Gli3* locus, the ChIP-seq and ATAC-seq profiles in BL6 mice again match the haplotype of the selected allele (**Figure 3.12B**). In this case, they match the stronger allele according to our hypothesis that a gain of function of the limb growth activator *Gli3* occurred in Longshanks. The SNPs at the G2 enhancer altered TF binding sites, with three of the ten SNPs in the enhancer predicted to have resulted in the gain of transcriptional activator – including from GLI3 itself – binding sites and one SNP resulting in loss of a binding site of NKX3-2, which likely acts to inhibit *Gli3* expression (**Figure 3.12C**). If the transcriptional

activators whose binding sites were predicted to be impacted by SNPs are key for enhancer activation, their absence in F0 and presence in F17 can explain why only the F17 version of the G2 enhancer was observed to stimulate *lacZ* expression in the limb bud.

Neither 4C-seq nor F0 histone profiles or chromatin accessibility data are available at the *Gli3* locus. Based on the chromatin features observed at the *Nkx3-2* locus in the weaker allele, the inactive or hypomorphic G2 enhancer in F0 mice may already be marked by accessible chromatin, facilitating TF binding in the F17 mice once the binding sites have been gained (and the NKX3-2 binding site has been lost). This pre-existing landscape is likely to be enriched for H3K4me1, as this mark spreads easily and nonspecifically over large swathes of DNA (Calo and Wysocka, 2013) and as the *Gli3* locus in BL6 mice includes at least twenty other putative enhancers as marked by the presence of H3K4me1 and H3K27ac and absence of H3K4me3 signal. H3K27ac may also exist at the F0 G2 enhancer, although it is less likely to be present since it is added by enzymes which double as cofactors of TFs – of which there may be few binding to the enhancer – and as it tends to disappear quickly when transcription is not stimulated (Calo and Wysocka, 2013). However, it may have been added by cofactors binding to other enhancers in the vicinity. Even in a hypomorphic state, the G2 enhancer may be part of a chromatin loop connecting multiple other enhancers to the *Gli3* promoter. *Gli3* is known to be regulated by multiple enhancers (Coy et al., 2011; Anwar et al., 2015). They likely add expression domains that together comprise the entire *Gli3* expression pattern; the expression domain of *Gli3* from our *in situ* hybridization in BL6 (**Figure 3.13A**) extends well beyond the expression domain of the *lacZ* reporter stimulated by the F17 G2 enhancer (**Figure 3.12D**), both within the forelimb and hindlimb and in the whole embryo.

4.5 References to Chapter 4

Akhtar, W., J. de Jong, A. V. Pindyurin, L. Pagie, W. Meuleman, J. de Ridder, A. Berns, L. F. Wessels, M. van Lohuizen and B. van Steensel (2013). "Chromatin position effects assayed by thousands of reporters integrated in parallel." *Cell* **154**(4): 914-927.

Amano, T., T. Sagai, H. Tanabe, Y. Mizushina, H. Nakazawa and T. Shiroishi (2009). "Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription." *Dev Cell* **16**(1): 47-57.

Andrey, G., R. Schopflin, I. Jerkovic, V. Heinrich, D. M. Ibrahim, C. Paliou, M. Hochradel, B. Timmermann, S. Haas, M. Vingron and S. Mundlos (2017). "Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding." *Genome Res* **27**(2): 223-233.

Anwar, S., R. Minhas, S. Ali, N. Lambert, Y. Kawakami, G. Elgar, S. S. Azam and A. A. Abbasi (2015). "Identification and functional characterization of novel transcriptional enhancers involved in regulating human GLI3 expression during early development." *Dev Growth Differ* **57**(8): 570-580.

Arnold, C. D., D. Gerlach, C. Stelzer, L. M. Boryn, M. Rath and A. Stark (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq." *Science* **339**(6123): 1074-1077.

Bagshaw, A. T. M. (2017). "Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes." *Genome Biol Evol* **9**(9): 2428-2443.

Bao, J. and M. T. Bedford (2016). "Epigenetic regulation of the histone-to-protamine transition during spermiogenesis." *Reproduction* **151**(5): R55-70.

Barde, I., E. Laurenti, S. Verp, A. C. Groner, C. Towne, V. Padrun, P. Aebischer, A. Trumpp and D. Trono (2009). "Regulation of episomal gene expression by KRAB/KAP1-mediated histone modifications." *J Virol* **83**(11): 5574-5580.

Battulin, N., V. S. Fishman, A. M. Mazur, M. Pomaznoy, A. A. Khabarova, D. A. Afonnikov, E. B. Prokhortchouk and O. L. Serov (2015). "Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach." *Genome Biol* **16**: 77.

Blackwood, E. M. and J. T. Kadonaga (1998). "Going the distance: a current view of enhancer action." *Science* **281**(5373): 60-63.

Bonn, S., R. P. Zinzen, C. Girardot, E. H. Gustafson, A. Perez-Gonzalez, N. Delhomme, Y. Ghavi-Helm, B. Wilczynski, A. Riddell and E. E. Furlong (2012). "Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development." *Nat Genet* **44**(2): 148-156.

Boya, R., A. D. Yadavalli, S. Nikhat, S. Kurukuti, D. Palakodeti and J. M. R. Pongubala (2017). "Developmentally regulated higher-order chromatin interactions orchestrate B cell fate commitment." *Nucleic Acids Res* **45**(19): 11070-11087.

Bren-Mattison, Y., M. Hausburg and B. B. Olwin (2011). "Growth of limb muscle is dependent on skeletal-derived Indian hedgehog." *Dev Biol* **356**(2): 486-495.

Bu, H., Y. Gan, Y. Wang, S. Zhou and J. Guan (2017). "A new method for enhancer prediction based on deep belief network." *BMC Bioinformatics* **18**(Suppl 12): 418.

Calo, E. and J. Wysocka (2013). "Modification of enhancer chromatin: what, how, and why?" *Mol Cell* **49**(5): 825-837.

Carelli, F. N., A. Liechti, J. Halbert, M. Warnefors and H. Kaessmann (2018). "Repurposing of promoters and enhancers during mammalian evolution." *Nat Commun* **9**(1): 4066.

Colbran, L. L., L. Chen and J. A. Capra (2017). "Short DNA sequence patterns accurately identify broadly active human enhancers." *BMC Genomics* **18**(1): 536.

Coy, S., J. H. Caamano, J. Carvajal, M. L. Cleary and A. G. Borycki (2011). "A novel Gli3 enhancer controls the Gli3 spatiotemporal expression pattern through a TALE homeodomain protein binding site." *Mol Cell Biol* **31**(7): 1432-1443.

Davidson, I. F., D. Goetz, M. P. Zaczek, M. I. Molodtsov, P. J. Huis In 't Veld, F. Weissmann, G. Litos, D. A. Cisneros, M. Ocampo-Hafalla, R. Ladurner, F. Uhlmann, A. Vaziri and J. M. Peters (2016). "Rapid movement and transcriptional re-localization of human cohesin on DNA." *EMBO J* **35**(24): 2671-2685.

de Laat, W. and D. Duboule (2013). "Topology of mammalian developmental enhancers and their regulatory landscapes." *Nature* **502**(7472): 499-506.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* **485**(7398): 376-380.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome." *Nature* **489**(7414): 57-74.

Farley, E. K., K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar and M. S. Levine (2015). "Suboptimization of developmental enhancers." *Science* **350**(6258): 325-328.

Farley, E. K., K. M. Olson, W. Zhang, D. S. Rokhsar and M. S. Levine (2016). "Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers." *Proc Natl Acad Sci U S A* **113**(23): 6508-6513.

Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur and L. A. Mirny (2016). "Formation of Chromosomal Domains by Loop Extrusion." *Cell Rep* **15**(9): 2038-2049.

Gaspar-Maia, A., A. Alajem, E. Meshorer and M. Ramalho-Santos (2011). "Open chromatin in pluripotency and reprogramming." *Nat Rev Mol Cell Biol* **12**(1): 36-47.

Ghisletti, S., I. Barozzi, F. Mietton, S. Polletti, F. De Santa, E. Venturini, L. Gregory, L. Lonie, A. Chew, C. L. Wei, J. Ragoussis and G. Natoli (2010). "Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages." *Immunity* **32**(3): 317-328.

Goodman, R. H. and S. Smolik (2000). "CBP/p300 in cell growth, transformation, and development." *Genes Dev* **14**(13): 1553-1577.

Grossman, S. R., X. Zhang, L. Wang, J. Engreitz, A. Melnikov, P. Rogov, R. Tewhey, A. Isakova, B. Deplancke, B. E. Bernstein, T. S. Mikkelsen and E. S. Lander (2017). "Systematic dissection of genomic features determining transcription factor binding and enhancer function." *Proc Natl Acad Sci U S A* **114**(7): E1291-E1300.

Haarhuis, J. H. I., R. H. van der Weide, V. A. Blomen, J. O. Yanez-Cuna, M. Amendola, M. S. van Ruiten, P. H. L. Krijger, H. Teunissen, R. H. Medema, B. van Steensel, T. R. Brummelkamp, E. de Wit and B. D. Rowland (2017). "The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension." *Cell* **169**(4): 693-707 e614.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. K. Glass (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Mol *Cell* **38**(4): 576-589.

Hellemans, J., M. Simon, A. Dheedene, Y. Alanay, E. Mihci, L. Rifai, A. Sefiani, Y. van Bever, M. Meradji, A. Superti-Furga and G. Mortier (2009). "Homozygous inactivating mutations in the NKX3-2 gene result in spondylo-megaepiphyseal-metaphyseal dysplasia." *Am J Hum Genet* **85**(6): 916-922.

Henikoff, S. and A. Shilatifard (2011). "Histone modification: cause or cog?" *Trends Genet* **27**(10): 389-396.

Hong, J. W., D. A. Hendrix and M. S. Levine (2008). "Shadow enhancers as a source of evolutionary novelty." *Science* **321**(5894): 1314.

Ing-Simmons, E., V. C. Seitan, A. J. Faure, P. Flicek, T. Carroll, J. Dekker, A. G. Fisher, B. Lenhard and M. Merkenschlager (2015). "Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin." *Genome Res* **25**(4): 504-513.

Inoue, F., M. Kircher, B. Martin, G. M. Cooper, D. M. Witten, M. T. McManus, N. Ahituv and J. Shendure (2017). "A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity." *Genome Res* **27**(1): 38-52.

Karlic, R., H. R. Chung, J. Lasserre, K. Vlahovicek and M. Vingron (2010). "Histone modification levels are predictive for gene expression." *Proc Natl Acad Sci U S A* **107**(7): 2926-2931.

Khamis, A. M., O. Motwalli, R. Oliva, B. R. Jankovic, Y. A. Medvedeva, H. Ashoor, M. Essack, X. Gao and V. B. Bajic (2018). "A novel method for improved accuracy of transcription factor binding site prediction." *Nucleic Acids Res* **46**(12): e72.

Kleinjan, D. A. and V. van Heyningen (2005). "Long-range control of gene expression: emerging mechanisms and disruption in disease." *Am J Hum Genet* **76**(1): 8-32.

Kragesteen, B. K., M. Spielmann, C. Paliou, V. Heinrich, R. Schopflin, A. Esposito, C. Annunziatella, S. Bianco, A. M. Chiariello, I. Jerkovic, I. Harabula, P. Guckelberger, M. Pechstein, L. Wittler, W. L. Chan, M. Franke, D. G. Lupianez, K. Kraft, B. Timmermann, M. Vingron, A. Visel, M. Nicodemi, S. Mundlos and G. Andrey (2018). "Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis." *Nat Genet* **50**(10): 1463-1473.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder and Y. Ruan (2012). "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation." *Cell* **148**(1-2): 84-98.

Li, Q., G. Barkess and H. Qian (2006). "Chromatin looping and the probability of transcription." *Trends Genet* **22**(4): 197-202.

Long, H. K., S. L. Prescott and J. Wysocka (2016). "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution." *Cell* **167**(5): 1170-1187.

Marinic, M., T. Aktas, S. Ruf and F. Spitz (2013). "An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape." *Dev Cell* **24**(5): 530-542.

Matharu, N. and N. Ahituv (2015). "Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease." *PLoS Genet* **11**(12): e1005640.

McManus, S., A. Ebert, G. Salvagiotto, J. Medvedovic, Q. Sun, I. Tamir, M. Jaritz, H. Tagoh and M. Busslinger (2011). "The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells." *EMBO J* **30**(12): 2388-2404.

Melcer, S. and E. Meshorer (2010). "Chromatin plasticity in pluripotent cells." *Essays Biochem* **48**(1): 245-262.

Meng, H. and B. Bartholomew (2018). "Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II." *J Biol Chem* **293**(36): 13786-13794.

Meshorer, E. and T. Misteli (2006). "Chromatin in pluripotent embryonic stem cells and differentiation." *Nat Rev Mol Cell Biol* **7**(7): 540-546.

Muerdter, F., L. M. Boryn and C. D. Arnold (2015). "STARR-seq - principles and applications." *Genomics* **106**(3): 145-150.

Osterwalder, M., I. Barozzi, V. Tissieres, Y. Fukuda-Yuzawa, B. J. Mannion, S. Y. Afzal, E. A. Lee, Y. Zhu, I. Plajzer-Frick, C. S. Pickle, M. Kato, T. H. Garvin, Q. T. Pham, A. N. Harrington, J. A. Akiyama, V. Afzal, J. Lopez-Rios, D. E. Dickel, A. Visel and L. A. Pennacchio (2018). "Enhancer redundancy provides phenotypic robustness in mammalian development." *Nature* **554**(7691): 239-243.

Pennacchio, L. A., W. Bickmore, A. Dean, M. A. Nobrega and G. Bejerano (2013). "Enhancers: five essential questions." *Nat Rev Genet* **14**(4): 288-295.

Riu, E., Z. Y. Chen, H. Xu, C. Y. He and M. A. Kay (2007). "Histone modifications are associated with the persistence or silencing of vector-mediated transgene expression in vivo." *Mol Ther* **15**(7): 1348-1355.

Sarro, R., A. A. Kocher, D. Emera, S. Uebbing, E. V. Dutrow, S. D. Weatherbee, T. Nottoli and J. P. Noonan (2018). "Disrupting the three-dimensional regulatory topology of the Pitx1 locus results in overtly normal development." *Development* **145**(7).

Schoenfelder, S., M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B. M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe and P. Fraser (2015). "The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements." *Genome Res* **25**(4): 582-597.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov and B. Ren (2012). "A map of the cis-regulatory sequences in the mouse genome." *Nature* **488**(7409): 116-120.

Sivakamasundari, V., H. Y. Chan, S. P. Yap, X. Xing, P. Kraus and T. Lufkin (2012). "New Bapx1(Cre-EGFP) mouse lines for lineage tracing and conditional knockout studies." *Genesis* **50**(4): 375-383.

Speel, E. J., A. H. Hopman and P. Komminoth (1999). "Amplification methods to increase the sensitivity of in situ hybridization: play card(s)." *J Histochem Cytochem* **47**(3): 281-288.

Spitz, F. and E. E. Furlong (2012). "Transcription factors: from enhancer binding to developmental control." *Nat Rev Genet* **13**(9): 613-626.

Stefflova, K., D. Thybert, M. D. Wilson, I. Streeter, J. Aleksic, P. Karagianni, A. Brazma, D. J. Adams, I. Talianidis, J. C. Marioni, P. Flicek and D. T. Odom (2013). "Cooperativity and rapid evolution of cobound transcription factors in closely related mammals." *Cell* 154(3): 530-540.

Taylor, G. C., R. Eskeland, B. Hekimoglu-Balkan, M. M. Pradeepa and W. A. Bickmore (2013). "H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction." *Genome Res* **23**(12): 2053-2065.

Thompson, A. C., T. D. Capellini, C. A. Guenther, Y. F. Chan, C. R. Infante, D. B. Menke and D. M. Kingsley (2018). "A novel enhancer near the Pitx1 gene influences development and evolution of pelvic appendages in vertebrates." *Elife* **7**: e38555.

Wunderlich, Z., M. D. Bragdon and A. H. DePace (2014). "Comparing mRNA levels using in situ hybridization of a target gene and co-stain." *Methods* **68**(1): 233-241.

Xin, B. and R. Rohs (2018). "Relationship between histone modifications and transcription factor binding is protein family specific." *Genome Res* **28**(3):321-333.

Yanez-Cuna, J. O., C. D. Arnold, G. Stampfel, L. M. Boryn, D. Gerlach, M. Rath and A. Stark (2014). "Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features." *Genome Res* **24**(7): 1147-1156.

Chapter 5: Conclusion and Future Outlook

The Capture-C and Longshanks analyses show how chromosome conformation capture data can be integrated with epigenetic, chromatin accessibility, and other datasets to help explain gene expression differences brought about by transcriptional regulation. A few main principles concerning chromosome conformation capture data emerged or were confirmed through these analyses. Firstly, Capture data can be used to establish the link between CREs and their target promoters. In the Longshanks experiment, 4C-seq data showed that the distal N1, N2, and N3 enhancers interacted with the *Nkx3-2* promoter (**Figure 3.11B**). Secondly, interacting loci tend to stay within the confines of TADs, which are indicative of physical barriers like CTCF binding sites that block chromatin loops by halting or stalling loop extrusion factors like cohesin and its subunits from translocating along the DNA. In the Capture-C analysis, most of the promoter contacts at each viewpoint mapped to the same TAD as the promoter (**Figure 2.3**). In the context of the *Nkx3-2* locus in the Longshanks mice, the N1, N2, and N3 enhancers were all found within the same TAD as the *Nkx3-2* promoter they contacted. Finally, both analyses suggested that even if CREs come into contact with their target promoters, they do not necessarily actively regulate them. Rather, the gene expression level of the target promoter should be considered, as poised promoters or those with intermediate expression levels are less constrained than active or silent promoters in the contacts they form. These contacts may not be capable of stimulating transcription, either through sequence changes in CREs that occurred during evolution (as in the Longshanks experiment) or as a result of differential *trans* factors in the nuclear environment that vary across cell types and developmental time points. These principles highlight the usefulness of Capture data in elucidating the mechanisms that underlie transcriptional regulation.

In order to efficiently both identify CREs and link them to their targets, experimenters should aim to integrate at least three types of data: TAD boundaries (derived from Hi-C), nucleosome occupancy (chromatin accessibility and/or histone modifications), and Capture data. The molecular features at the *Nkx3-2* enhancers in the Longshanks mice and the low predictive power of chromatin features in the Capture-C analysis suggest that TF binding is more directly determinative of CRE function than histone modifications or chromatin accessibility, in line with previous studies (Dogan et al., 2015; Kreimer et al., 2017). However, TF binding data is

impractical to collect. By contrast, the first two aforementioned types of data are readily available for many cell types or are easily obtainable. Each indirectly assays TF binding. TADs, derived from Hi-C data, indicate the physical barriers (CTCF binding and subsequent insulator activity) that hinder TFs from bringing transcriptional machinery to promoters. Considering TAD boundaries can quickly confine the search for putative CREs to just a few megabase pairs or hundreds of kilobase pairs, and is particularly useful for excluding contacts from Capture data of promoters that are located on the edges of TADs since these interactions are presumably due to proximity ligation, rather than biological function. Conservation of TADs between species means that TAD coordinates from one species may be used as a proxy for another; however, care should be taken to ensure that the developmental stage is as closely matched as possible, as our investigation of mESC and cortex TAD boundaries in **Chapter 2** confirmed previous studies that suggested TADs merge over development as chromatin compacts. TF binding can also be inferred from chromatin accessibility or nucleosome occupancy data, including identification of DHSs and profiling of histone marks. This is because nucleosomes and TFs compete for binding to CREs, and the enzymes that add histone marks are often cofactors of TFs. DHSs identify all types of CREs, whereas histone marks give a more detailed view linked to the gene expression level of the target gene. DHSs are easily obtainable from short, simple assays like ATAC-seq, and, like key histone marks associated with enhancers and promoters, are already widely available for a variety of cell types (ENCODE Project Consortium, 2012). Finally, chromosome conformation capture data is needed to establish the link between a promoter and a CRE, as at the Longshanks *Nkx3-2* locus, where we used 4C-seq to confirm that the N1 and N3 enhancers acted on *Nkx3-2* and not on the *Rab28* promoter, which is in the same TAD and is also expressed in the limb bud.

Although these three types of data can accurately identify and link CREs to their target promoters, they do not reveal whether a CRE is functional in a given context (the third main principle from these analyses). To uncover CRE expression domains and activity levels, functional validation assays like transgenic reporter assays are needed. These have revealed that a given CRE may not actively regulate a promoter despite forming a chromatin loop with it, as at the *Shh* locus in anterior limb bud cells (Amano et al., 2009) and presumably at the *Nkx3-2* locus in F17 Longshanks mice. However, transgenic reporter assays may not accurately reflect

the transcriptional regulation that occurs at the endogenous locus, since they do not involve chromatin looping and may suffer from integration site effects. To avoid integration effects, enhancer traps (O'Kane and Gehring, 1987) – in which a minimal promoter is cloned upstream of a reporter gene and is integrated into the genome to report on local enhancer activity – may be used instead; however, these still do not involve chromatin loops and may be confounded by the activity of multiple nearby regulatory elements at the integration site (Kvon, 2015). Both enhancer traps and transgenic enhancer reporter assays are limited to looking at individual CREs in isolation, and thus give an incomplete picture of transcriptional regulation because CREs may act synergistically on the same promoter, often refining one another's expression domains (Kragesteen et al., 2018). To capture the activity of all CREs at once, a BAC reporter construct, in which the coding sequence of the endogenous gene is replaced with a reporter gene, can instead be introduced in the genome, facilitating manipulation of putative CREs so as to molecularly dissect their activity and function (Kvon, 2015; Thompson et al., 2018). However, BAC transgenics are still prone to disparities between episomal and endogenous chromatin features (Matthaei, 2007). All in all, transcriptional regulation is a hugely complex puzzle to decipher, not only due to the interplay between *cis*- and *trans*-factors in the nuclear environment but also to the presence of multiple distal regulatory elements that act additively, synergistically, or in a functionally redundant manner on the same target promoter. This complexity means that for many developmental regulators such as *Pitx1*, there are still new distal enhancers being explored (Sarro et al., 2018; Thompson et al., 2018) even after many years and concerted efforts by multiple labs to explore transcriptional regulation of these well-known genes. Chromosome conformation capture is an invaluable tool to find transcriptional regulators of a gene so that follow-up functional validation experiments can be performed. Capture signal profiles can reveal where ectopic chromatin interactions have occurred or existing ones have been lost, as during disease (Lupianez et al., 2015) or development (Andrey et al., 2013). The advent of Capture techniques with ever higher resolution (Stevens et al., 2017) enables the detection of chromatin dynamics and subsequently possible transcriptional regulatory mechanisms with increasing rates of sensitivity.

5.1 References to Chapter 5

Amano, T., T. Sagai, H. Tanabe, Y. Mizushina, H. Nakazawa and T. Shiroishi (2009). "Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription." *Dev Cell* **16**(1): 47-57.

Andrey, G., T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz and D. Duboule (2013). "A switch between topological domains underlies HoxD genes collinearity in mouse limbs." *Science* **340**(6137): 1234167.

Dogan, N., W. Wu, C. S. Morrissey, K. B. Chen, A. Stonestrom, M. Long, C. A. Keller, Y. Cheng, D. Jain, A. Visel, L. A. Pennacchio, M. J. Weiss, G. A. Blobel and R. C. Hardison (2015). "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility." *Epigenetics Chromatin* **8**: 16.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome." *Nature* **489**(7414): 57-74.

Kragesteen, B. K., M. Spielmann, C. Paliou, V. Heinrich, R. Schopflin, A. Esposito, C. Annunziatella, S. Bianco, A. M. Chiariello, I. Jerkovic, I. Harabula, P. Guckelberger, M. Pechstein, L. Wittler, W. L. Chan, M. Franke, D. G. Lupianez, K. Kraft, B. Timmermann, M. Vingron, A. Visel, M. Nicodemi, S. Mundlos and G. Andrey (2018). "Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis." *Nat Genet* **50**(10): 1463-1473.

Kreimer, A., H. Zeng, M. D. Edwards, Y. Guo, K. Tian, S. Shin, R. Welch, M. Wainberg, R. Mohan, N. A. Sinnott-Armstrong, Y. Li, G. Eraslan, T. B. Amin, R. Tewhey, P. C. Sabeti, J. Goke, N. S. Mueller, M. Kellis, A. Kundaje, M. A. Beer, S. Keles, D. K. Gifford and N. Yosef (2017). "Predicting gene expression in massively parallel reporter assays: A comparative study." *Hum Mutat* **38**(9): 1240-1250.

Kvon, E. Z. (2015). "Using transgenic reporter assays to functionally characterize enhancers in animals." *Genomics* **106**(3): 185-192.

Lupianez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel and S. Mundlos (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions." *Cell* **161**(5): 1012-1025.

Matthaei, K. I. (2007). "Genetically manipulated mice: a powerful tool with unsuspected caveats." *J Physiol* **582**(Pt 2): 481-488.

O'Kane, C. J. and W. J. Gehring (1987). "Detection in situ of genomic regulatory elements in Drosophila." *Proc Natl Acad Sci U S A* **84**(24): 9123-9127.

Sarro, R., A. A. Kocher, D. Emera, S. Uebbing, E. V. Dutrow, S. D. Weatherbee, T. Nottoli and J. P. Noonan (2018). "Disrupting the three-dimensional regulatory topology of the Pitx1 locus results in overtly normal development." *Development* **145**(7).

Stevens, T. J., D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sanso, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klenerman and E. D. Laue (2017). "3D structures of individual mammalian genomes studied by single-cell Hi-C." *Nature* **544**(7648): 59-64.

Thompson, A. C., T. D. Capellini, C. A. Guenther, Y. F. Chan, C. R. Infante, D. B. Menke and D. M. Kingsley (2018). "A novel enhancer near the Pitx1 gene influences development and evolution of pelvic appendages in vertebrates." *Elife* **7**: e38555.

Chapter 6: Genomic structure of *Hstx2* modifier of *Prdm9*-dependent hybrid male sterility in mice

6.1 Declaration of Contributions

Lustyk, D., Kinský, S., Ullrich, K.K., **Yancoskie, M.**, Kašíková, L., Gergelits, V., Sedláček, R., Chan, Y.F., Odenthal-Hesse, L., Forejt, J., Jansa, P. (2019) Genomic structure of *Hstx2* modifier of *Prdm9*-dependent hybrid male sterility in mice. *Genetics.* 213(**3**): 1047-1063.

Author contributions: **M. Yancoskie** designed and performed BAC modification on a SPO11 BAC to replace the endogenous Spo11 coding region with a Cas9 endonuclease, cloned the CRISPR plasmids containing three different guide oligos targeting the *Hstx2* locus, wrote Methods and Materials for this part of the project, and provided manuscript feedback. ***Relevance to the collective work***: The SPO11-controlled Cas9 endonuclease induced double-stranded breaks during meiosis, increasing the rate of recombination by 15-fold around the *Hstx2* locus compared to previous recombination attempts. This strategy provided the critical recombination event to refine the critical interval of the *Hstx2* locus from 4.3 Mb to 2.7 Mb.

Co-author contributions: D. Lustyk conducted genetic and meiotic experiments, genotyped and bred recombinant transgenic mice, analyzed the data, wrote Methods and Materials for this part of the project, and provided manuscript feedback. S. Kinský designed the *Fmr1nb* TALEN knockout construct and provided the founder mice. K.K. Ullrich performed assembly of optical maps and computational analysis. L. Kašíková participated in genetic mapping, meiotic experiments and breeding of knockout transgenic mice at the beginning of the project. V. Gergelits provided statistical analysis of the data, wrote Methods and Materials for this part of the project, and provided manuscript feedback. R. Sedláček provided access to transgenic facility and manuscript feedback. Y.F. Chan designed the Spo11:Cas9 strategy to direct recombination to *Hstx2* and provided manuscript feedback. L. Odenthal-Hesse provided design of optical mapping, performed analysis of optical mapping and data interpretation, and provided manuscript feedback. J. Forejt designed the experiments, coordinated collaborative parts, analyzed the data and wrote the manuscript. P. Jansa designed the experiments, prepared the Cas9 BAC and CRISPR constructs for pronuclear injection, genotyped and bred recombinant transgenic mice, analyzed the data, wrote the manuscript, and supervised students.

6.2 Full Article

Genomic structure of *Hstx2* modifier of *Prdm9*-dependent hybrid male sterility in mice

Diana Lustyk,[*,†] Slavomír Kinský,[‡] Kristian Karsten Ullrich,[$] Michelle Yancoskie,[**] Lenka Kašíková,[*,1,2] Václav Gergelits,[*] Radislav Sedláček,[‡] Yingguang Frank Chan,[**] Linda Odenthal-Hesse,[$] Jiří Forejt,[*,3] and Petr Jansa[*,3]

*Laboratory of Mouse Molecular Genetics and [‡]The Czech Centre for Phenogenomics, Division BIOCEV, Institute of Molecular Genetics, Czech Academy of Sciences, Vestec, Czech Republic CZ-25250, Czech Republic, [†]Faculty of Science, Charles University, Prague, CZ-12000, Czech Republic, [$]Department Evolutionary Genetics, Research Group Meiotic Recombination and Genomic Instability, Max Planck Institute for Evolutionary Biology, Plön D-24306, Germany, and **Friedrich Miescher Laboratory of the Max Planck Society, Tübingen 72076, Germany

ORCID IDs: 0000-0003-0879-4012 (D.L.); 0000-0003-4308-9626 (K.K.U.); 0000-0002-5178-8833 (V.G.); 0000-0002-3352-392X (R.S.); 0000-0001-6292-9681 (Y.F.C); 0000-0002-5519-2375 (L.O-H.); 0000-0002-2793-3623 (J.F.); 0000-0002-1406-1707 (P.J.)

doi: https://doi.org/10.1534/genetics.119.302554
Manuscript received August 16, 2019; accepted for publication September 23, 2019; published Early Online September 26, 2019.
Supplemental material available at FigShare: https://doi.org/10.25386/genetics.9874460.
[1]Present address: Department of Immunology, Charles University, Second Faculty of Medicine and University Hospital Motol, Prague CZ-15000, Czech Republic.
[2]Present address: SOTIO, Prague CZ-17000, Czech Republic.
[3]Corresponding authors: Laboratory of Mouse Molecular Genetics, Division BIOCEV, Institute of Molecular Genetics, Czech Academy of Sciences, Prumyslova 595, Vestec CZ-25250, Czech Republic. E-mail: pjansa@img.cas.cz; and jforejt@img.cas.cz

ABSTRACT F_1 hybrids between mouse inbred strains PWD and C57BL/6 represent the most thoroughly genetically defined model of hybrid sterility in vertebrates. Hybrid male sterility can be fully reconstituted from three components of this model, the *Prdm9* gene, intersubspecific homeology of *Mus musculus musculus* and *Mus musculus domesticus* autosomes, and the X-linked *Hstx2* locus. *Hstx2* modulates the extent of *Prdm9*-dependent meiotic arrest and harbors two additional factors responsible for intersubspecific introgression-induced oligospermia (*Hstx1*) and meiotic recombination rate (*Meir1*). To facilitate positional cloning and to overcome the recombination suppression within the 4.3 Mb encompassing the *Hstx2* locus, we designed *Hstx2*-CRISPR and SPO11/Cas9 transgenes aimed to induce DNA double-strand breaks specifically within the *Hstx2* locus. The resulting recombinant

reduced the *Hstx2* locus to 2.70 Mb (Chr X: 66.51–69.21 Mb). The newly defined *Hstx2* locus still operates as the major X-linked factor of the F_1 hybrid sterility, and controls meiotic chromosome synapsis and meiotic recombination rate. Despite extensive further crosses, the 2.70 Mb *Hstx2* interval behaved as a recombination cold spot with reduced PRDM9-mediated H3K4me3 hotspots and absence of DMC1-defined DNA double-strand-break hotspots. To search for structural anomalies as a possible cause of recombination suppression, we used optical mapping and observed high incidence of subspecies-specific structural variants along the X chromosome, with a striking copy number polymorphism of the microRNA *Mir465* cluster. This observation together with the absence of a strong sterility phenotype in Fmr1 neighbor (*Fmr1nb*) null mutants support the role of miRNA as a likely candidate for *Hstx2*.

KEYWORDS Speciation; Hybrid sterility X2; *Prdm9*; Bionano optical mapping; SPO11Cas9 transgene; *Fmr1nb*

REPRODUCTIVE isolation is a basic prerequisite of speciation implemented by a range of prezygotic and postzygotic mechanisms under complex genetic control (Dobzhansky1951; Dion-Côté and Barbash 2017). Hybrid sterility, one of the reproductive isolation mechanisms, appears in the early stages of speciation and shares common features in many animal and plant species hybrids. They include preferential involvement of the heterogametic sex (XY or ZW), known as Haldane's rule (Haldane 1922), or the large X effect (Coyne's rule), referring to disproportionate engagement of X chromosome compared to autosomes (Dobzhansky 1951; Forejt 1996; Coyne and Orr 2004; Good *et al.* 2008; Presgraves 2018). The first hypothesis on genetic control of hybrid sterility, known as Dobzhansky–Muller epistatic incompatibility, refers to a dysfunction caused by the independent divergence ofmutually interacting genes (Dobzhansky 1951). More recently, an interaction between meiotic drive and its suppressors has been implicated in some instances of reproductive isolation (Orr 2005; Zhang *et al.* 2015; Patten 2018). However, despite extensive genetic studies in organisms of various species such as yeast, fruit fly, or house mouse, the underlying genetic architecture and molecular mechanisms of hybrid sterility remain elusive [reviewed in Maheshwari and Barbash (2011); Phifer-

Rixey and Nachman (2015); Dion-Côté and Barbash (2017); Mack and Nachman (2017); Payseur *et al.* (2018)].

The first hybrid sterility genetic factor to be identified in vertebrate, the hybrid sterility 1 (*Hst1*), was described in hybrids between laboratory and wild mice (Forejt and Ivanyi 1974; Gregorová *et al.* 1996; Trachtulec *et al.* 1997) and identified as the *Prdm9* gene encoding PR/SET domain-containing nine protein (Mihola *et al.* 2009). The PRDM9 binds genomic DNA by a zinc finger domain at allele-specific sites and trimethylates lysine 4 and lysine 36 of histone 3. In mice, humans, and other mammalian species, *Prdm9* mediates meiotic recombination by determining the genomic localization of the recombination hotspots (Baudat *et al.* 2010; Myers *et al.* 2010; Parvanov *et al.* 2010). In a mouse model of intersubspecific hybrids where *Mus musculus domesticus* subspecies is represented by inbred strain C57BL/6J (hereafter B6) and *Mus musculus musculus* by PWD/Ph (hereafter PWD) (Gregorova and Forejt 2000) *Prdm9* causes early meiotic arrest and complete male sterility by interaction with the X-linked *Hstx2* locus. Hybrids between laboratory strains PWD and B6 serve as a robust, reproducible and genetically well-defined model of hybrid sterility [reviewed in Forejt (1996); Forejt *et al.* (2012)]. Specific allelic combinations of the *Prdm9* gene ($Prdm9^{PWD/B6}$) and *Hstx2* locus ($Hstx2^{PWD}$) were shown necessary but not sufficient to fully explain the meiotic arrest in hybrids. Initially, three or more additional hybrid sterility genes of small effect complementing the *Prdm9* and *Hstx2* major hybrid sterility genes had been considered (Dzur-Gejdosova *et al.* 2012). Later, we identified chromosome-autonomous meiotic asynapsis of homeologous chromosomes [homologous chromosomes from related(sub)species] as the third requirement for meiotic arrest (Bhattacharyya *et al.* 2013, 2014). The chromosomal, non-genic effects of homeologous chromosomes in (PWD x B6) F_1 hybrids, manifested as a failure of meiotic chromosome synapsis, is most likely a consequence of evolutionary erosion of PRDM9 binding sites in each subspecies, resulting in asymmetry of DNA double-strand-break (DSB) hotspots (Davies *et al.* 2016). The explanation of hybrid sterility by expected shortage of symmetric DNA DSBs was supported by improvement of chromosome pairing and fertility after experimentally increasing the number of symmetric DNA DSBs by random stretches of a homozygous PWD sequence (Gregorova *et al.* 2018). Moreover, partial improvement of meiotic chromosome synapsis in hybrid males was

achieved by addition of exogenous DSBs generated by a single cisplatin injection (Wang *et al.* 2018).

The PWD allele of the *Hstx2* locus (*Hstx2*^{PWD}) is indispensable for full sterility of (PWD x B6) F_1 hybrids, while the *Hstx2*^{B6} allele attenuates the phenotype to partial spermatogenesis arrest in reciprocal (B6 x PWD) F_1 males (Dzur-Gejdosova *et al.* 2012; Flachs *et al.* 2012; Forejt *et al.* 2012). Admittedly, the mechanism of action of the *Hstx2* locus in meiotic arrest of F_1 hybrids remains elusive. Previously, the *Hstx2* locus was mapped to a 4.7 Mb region on X chromosome [chromosome X (Chr X): 64.9–69.6 Mb] (Bhattacharyya *et al.* 2014). The interval that encompasses 10 protein-coding genes and a cluster of microRNA (miRNA) genes is still too large to identify the true *Hstx2* candidate. The *Hstx2* locus (Chr X: 64.9–69.6 Mb) harbors two additional meiosis-related genetic factors,the hybrid sterility X1 (*Hstx1*) locus, manifested by spermhead malformations after *Hstx2*^{PWD} sequence introgression into the B6 genome (Storchová *et al.* 2004), and meiotic recombination 1 (*Meir1*), which controls meiotic recombination rate (Balcova *et al.* 2016). Since these factors have not yet been genetically separated, their phenotypes may represent a pleiotropic effect of the same gene.

In an attempt to reduce the size of *Hstx2*, we constructed an SPO11-driven CRISPR-Cas9 system to target meiotic recombination to a particular genomic locus within the *Hstx2* recombination cold spot. Although the method did not work as predicted, we recovered a single recombinant, thus reducing the *Hstx2* locus to 2.70 Mb. We show that the shortened version of *Hstx2* still carries the genetic factors or genes responsible for hybrid sterility, meiotic chromosome asynapsis, and genome-wide control of meiotic recombination rate. Using Bionano Optical mapping technology, we show high incidence of subspecies-specific insertion/deletion variants inside and outside the *Hstx2* locus. Furthermore, we interrogate the *Fmr1nb* gene as a possible *Hstx2* candidate gene.

Materials and Methods

Animals and ethics statement

The mice were maintained in the Pathogen-Free Facility of the Institute of Molecular Genetics (Czech Academy of Sciences in Prague). The project was approved by the Animal Care and Use Committee of the Institute of Molecular Genetics AS CR and by the Czech Central Committee for Animal Welfare, and ethically reviewed and performed in accordance with European Directive 86/609/EEC. Subconsomic mouse strains C57BL/6J-ChrX.1^{PWD}/Ph (abbreviated B6.DX.1) and C57BL/6J-

ChrX.1s^{PWD}/Ph (B6.DX.1s) were described earlier (Storchová *et al.* 2004). The C57BL/6J-ChrX.64-69$^{PWD/Ph}$/ForeJ (B6.DX.64-69) congenic strain was established by backcrossing B6.DX.1s to the B6 strain (**Figure 6.1**). The congenic strain C57BL/6J-ChrX.66-69^{PWD}/Ph (B6.DX.66-69) was prepared by the new CRISPR/ Cas9 *Hstx2*-targeting method. The PWD/B6 composition of the Chr X is depicted schematically in **Figure 6.1** for each consomic strains.

Genotyping, fertility parameters, and histology

Genomic DNA was prepared from tails by NaOH method (Truett *et al.* 2000). The X chromosome recombinants in the backcross 1 (BC1) populations were genotyped by PWD/B6 allele-specific microsatellite markers (**Table 6.1**). Recombination breakpoints were determined by Sanger-DNA sequencing of the PCR amplicons carrying informative PWD/B6 SNP polymorphism(s). Genotyping of the new B6.DX.66-69 strain by microsatellite markers, Sanger DNA sequencing, and next generation sequencing showed the maximum and minimum extent of the PWD sequence on Chr X. The *Fmr1nb* deletion was confirmed using primers: forward 5'CAGGAGGTTCTGGACTGCTC 3' and reverse 5'TGAAGTCCAGAAGCCAAACC 3'. All experiments were performed with at least three animals per group. Cytological and histological experiments were performed on males between 8 and 10 weeks of age, with the exception of the males after fertility test.

Quantitative Reverse Transcription-PCR (RT-qPCR) analysis

Total RNA was extracted from testes by TRI Reagent #T9424 (Sigma, St. Louis, MO) according to manufacturer's instructions. The RNA was reverse transcribed using MuMLV-RT (28025-013; Invitrogen, Carlsbad, CA). Quantitative real-time PCR was performed with the Light Cycler DNA Fast Start Master SYBR Green I kit (Roche) in a Light Cycler 480 Instrument II at Tm = 60°. The sequences of primers for *Fmr1nb* were: Fmr1nb-F – 5'-TCCTGGGATTTCTGCCTATG-3', Fmr1nb-R – 5'-CCTTCAACATCCTGTTCATCC-3'; and the primers for *Actin-b* were: Actb-F – 5'-CTAAGGCCAACCGTGAAAAG-3', Actb-R – 5'-ACCAGAGGCATACAGGGACA-3'. The *Fmr1nb* expression values were normalized to *Actin-b* expression.

Western blotting

Whole testes were snap-frozen in liquid nitrogen before extraction buffer with protease inhibitors (1836153; Roche) and bensonase (1.01654.0001; Merck) was used to homogenize the tissue (see **Table 6.2**). After a 30 min incubation, 2% SDS was added and the mixture was heated at 95° for 20 min. Total protein concentration was measured using the Pierce BCA Protein Assay Kit (#23225; Thermo Scientific). The protein samples were then size-separated by electrophoresis on a gradient Bolt 4-12% Bis-Tris plus gel (NW04120BOX; Invitrogen), transferred onto a polyvinylidene difluoride membrane, and blocked with TBST with 5% BSA overnight. Primary antibodies against FMR1NB (sc-246953, goat polyclonal; Santa Cruz Biotechnology) and alpha-tubulin (66031-1-Ig, mouse monoclonal; Proteintech) were used at the 1:1000 and 1:2000 dilutions, respectively. Secondary antibodies (a donkey anti-goat IgG-HRP antibody, sc-2020; Santa Cruz Biotechnology, and a horse anti-mouse IgG-HRP antibody, 7076; Cell Signaling Technology), conjugated to HRP were used at 1:10000 dilution. Western Blotting Substrate (#32106; Pierce ECL Plus) was used for detection of HRP enzyme activity. Images were captured using the BioRad ChemiDoc MP Imaging System and processed with ImageLab software (Bio-Rad, Hercules, CA).

Immunofluorescence microscopy
Meiotic chromosome spreads were performed as previously described (Anderson *et al.* 1999) with minor modifications. Briefly, the testes were dissected and transferred to 1ml of RPMI (Sigma). Sucrose (0.1M) was used as a hypotonic solution and cells were dropped onto a slide with 1% paraformaldehyde containing protease inhibitors (1836153; Roche). After 3 hr at 4° slides were washed and blocked with 0.5 X blocking buffer (1.5% BSA, 5% goat serum, 0.05% Triton X-100) containing protease inhibitors (1836153; Roche) for 1 hr at 4°. Primary antibodies (listed in **Table 6.2**) were added and the slides were incubated overnight in a humid chamber at 4°. The slides were then incubated with secondary antibodies conjugated to fluorophores (**Table 6.2**) for 1 hr at 4°. The slides were mounted with Vectashield mounting medium containing DAPI (H1200). The immunofluorescence images were observed by Nikon Eclipse X 400 epifluorescence microscope with single band-pass filters for excitation and emission of infrared, red, blue and green fluorescence (Chroma Technologies) and X 60 Plan Fluor objective (MRH00601; Nikon, Garden City, NY). The images were captured using a DS-QiMc monochrome CCD camera (Nikon) and NIS Elements processing program (NIS-Elements Microscope Imaging Software). The images were adjusted using Adobe Photoshop (Adobe Systems).

Construction of Fmr1nb-specific TALEN and generation of transgenic mice
TALEN nucleases were designed using TAL Effector Nucleotide Targeter 2.0 (https://tale-nt.cac.cornell.edu/), assembled using the Golden Gate Cloning system (https://international.neb.com/applications/cloning-and-synthetic-biology/dna-assembly-and-cloning/golden-gate-assembly) and cloned into the ELD-KKR backbone plasmid. TALEN containing repeats NN-NN-HD-NG-NN-NN-NG-NG-NI-NN-NI-NN-NI-HD-HD-NG-HD-HD (for 5´ site) and NG-HD-NG-HD-NG-NN-NI-HD-NG-NG-NN-NN-HD-HD-NG-NG (for 3´ site) recognized a locus close to the ATG start codon of *Fmr1nb*. Each TALEN plasmid was linearized with *Not*I and transcribed using the mMESSAGE mMACHINE T7 Kit (Ambion). Polyadenylation of resulting mRNAs was performed using the Poly(A) Tailing Kit (Ambion); the mRNA was purified with RNeasy Mini columns (Qiagen, Valencia, CA). TALEN mRNAs were diluted in nuclease free water and kept at −80°. Transgenic mice were generated in the transgenic facility of the Institute of Molecular Genetics by injecting purified mRNA of *Fmr1nb*-specific TALEN into male pronuclei of one-cell embryos of C57BL/6N or B6.DX.1s origin. Mice positive for mutations were identified by PCR reaction with Fmr1nb2outF and Fmr1RightBsrI primers followed by *Nsp*I digestion. Specific genome mutations were identified by PCR fragment sequencing. Twenty-three mouse founders (F_0), each carrying a mutated allele of the *Fmr1nb* gene were generated. After outcrossing the F_0 mice to C57BL/6N or to B6.DX.1s we obtained five B6.*Fmr1nb*$^{-}$ mouse strains and three B6.DX.1s.*Fmr1nb*$^{-}$ strains with stable deletion mutations. Here we used two lines, the B6.*Fmr1nb*^{em1ForeJ} line carrying 236 bp long deletion over the ATG start codon of the *Fmr1nb*^{B6} allele, and the B6.DX.1s.*Fmr1nb*^{em1ForeJ} line carrying 19 bp long deletion over the ATG start codon of the of *Fmr1nb*^{PWD} allele.

Preparation of CRISPR-Hstx2 and SPO11-Cas9 constructs, and generation of transgenic mice
To place the Cas9 nuclease under the control of the SPO11 promoter, the SPO11 coding region was replaced by a mouse codon-optimized Cas9 open reading frame

in a SPO11-carrying bacterial artificial chromosome (BAC) clone (RP23-20N4, distributed by BACPAC Resources, Oakland, CA) by a marker-less GalK double-selection system via liquid culture recombineering as described (Sharan *et al.* 2009). Homology arms for the SPO11 BAC were introduced by PCR with Phusion polymerase (New England Biolabs, Frankfurt am Main, Germany). The 1.3 kbp PCR product was purified with a Gel Extraction Kit (QIAGEN) and confirmed by Sanger sequencing. The Cas9 cassette was produced by excision from plasmid MLM3613 (#42251; Addgene, Watertown, MA) by enzymes *Sac*II and *Mss*I (Thermo Fisher Scientific, Schwerte, Germany) and purified by gel extraction. The homology arms were added by PCR amplification and Phusion polymerase. The CRISPR plasmid pX260 was obtained (#42229, Addgene plasmid, a gift from Feng Zhang; Cong *et al.* 2013) and the CRISPR protospacers corresponding to the *Hstx2* loci were cloned according to instructions from the Zhang Laboratory (https://media.addgene.org/cms/filer_public/e6/5a/e65a9ef8-c8ac-4f88-98da-3b7d7960394c/zhang-lab-general-cloning-protocol.pdf). Briefly, long oligonucleotides were ordered as Ultramers (oligos 20-21; Integrated DNA Technologies, Coralville, IA) for the following three target regions flanking the *Hstx2* locus: a sequence 2.2 Mb upstream of the *Ctag2* gene (Chr X: 65,069,229–65,069,258); an intergenic sequence between the *Mir465* cluster and *Gm1140* predicted protein coding gene (Chr X: 67,052,342–67,052,371); and a sequence 4 kbp upstream of the *Aff2* gene (Chr X: 69,356,143–69,356,172). After phosphorylation (T4 Polynucleotide Kinase, New England Biolabs) and annealing by temperature ramping from 95º to 30 sec by –0.1º/min increments, the duplexes were ligated into the *Bbs*I site of the cut pX260 plasmid (New England Biolabs) and transformed into DH5-Alpha *Escherichia coli* cells. The protospacer-containing plasmids were further modified by excising the Cas9 open reading frame with *Pst*I (New England Biolabs). Each final plasmid contains the U6 promoter, protospacer, the ^{1}H promoter and the *trans*-activating CRISPR RNA. These were sequence-verified before transgenic injection. The CRISPR constructs and SPO11-Cas9-BAC construct were generated in Tubingen by the laboratory of Y.F.C. The BAC transgene was injected to the pronuclei of 1-day-old mouse embryos and the founders were generated in the laboratory of R.S. in Vestec.

Bionano optical mapping

We generated optical maps for two markers (BspQ1 and DLE-1) across the whole genome of five different mice, from two mouse subspecies: C57BL/6J (B6) and C57Bl6Crl (B6N) of *M. m. domesticus* and PWD/Ph (PWD) and PWK/Ph (PWK) of *M. m. musculus* origin. Two females were from the congenic C57BL/6J-ChrX.64-69PWD/Ph strain (B6.DX64-69), carrying a small portion of Chr X including the hybrid sterility *Hstx2* locus from PWD/Ph on C57BL/6 background. First, megabase-scale high molecular weight (HMW) DNA was extracted according to the Saphyr Bionano Prep Animal Tissue DNA Isolation Soft Tissue Protocol (Document Number: 30077; Revision B). Briefly, cell nuclei were isolated from splenic tissue and embedded in agarose plugs. DNA in plugs was purified with Proteinase K and RNAse, then HMW genomic DNA was extracted from the agarose plugs using agarase, and purified by drop dialysis. HMW DNA was resuspended overnight before quantification with the Qubit BR dsDNA assay, then kept at 4° until labeling. Each sample was labeled at the recognition sites NtBspQ1 (GCTCTTC) and DLE-1 (CTTAAG), respectively, using two different methylation insensitive assays. The Bionano nicking, labelling, repairing, and staining protocol was used to label

NtBspQ1 (Document Number: 30206, Revision C), and was performed on 900 ng of purified HMW DNA for each mouse. The Bionano direct labelling and staining protocol (Document Number: 30024, Revision I) was performed on 750 ng of DNA to label all DLE-1 recognition sites. After an initial clean-up step, the labeled HMW DNA was pre-stained, homogenized, and quantified with the Qubit HS dsDNA assay, before using an appropriate amount of backbone stain YOYO-1. The molecules were then imaged using the Bionano Saphyr System (Bionano Genomics, San Diego, CA). We obtained high-quality optical reads for both labeling techniques. For example, for the nicking, labeling, repairing, and staining labeling produced an average of 437 Gbps of reads, which were longer than 150 kbps and have a minimum of nine label. It achieved an average N50 length of 0.3137 Mbp with an average label density of 14.82 labels per 100 kbp. Similarly, the direct labelling and staining labeling achieved an average output of 389 Gbps (≥150 kbp and minSites ≥9), an average N50 length of 0.2663 Mbp and an average label density of 13.72/100 kbp. (Individual outputs were collected for each animal and labeling technique in **Table 6.3**). The presence of *in-silico* recognition sites for each enzyme recognition site in the genome was used to compute separate *in-silico* optical maps for each labeling enzyme, for the mm10 genome (**Table 6.4**).

Detection and quantification of apoptotic cells: terminal deoxynucleotidyl transferase-mediated dUTP nick-end labeling assay

The males were killed and the testes dissected from, and fixed in 4% paraformaldehyde overnight at 4°. Testes were dehydrated and embedded in paraffin. Paraffin sections at 3 µm thick were deparaffinized. To perform antigen retrieval for immunohistochemistry, the slides were incubated in Citrate Antigen Retrieval solution for 15 min at pH 6.0. The slides were processed as for immunofluorescence. The apoptotic cells in the tissue sections were determined by terminal deoxynucleotidyl transferase-mediated dUTP nick-end labeling (TUNEL), using *in situ* DeadEnd Fluorometric detection kit (G3250-PROMEGA, Madison, WI) according to technical protocol (#TB235). TUNEL-treated testicular sections were mounted in Vectashield with DAPI to watch the nuclei. Images were captured from a Nikon E-400 Eclipse fluorescence microscope and captured with a Ds-Qi_Mc1 CCD camera (Nikon). The images were processed and TUNEL-positive cells counted by the NIS Elements picture analyzer, and processed using Photoshop (Adobe).

Fertility test

Each male was mated with one 8-week-old C57BL/6J virgin female for 3 months, during which the numbers of neonatal pups sired by B6.DX.1s.Fmr1nb^{-} and B6.DX.1s males were recorded.

Data availability and statistics

Strains and plasmids are available upon request. The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, tables, and in the supplemental material. The optical mapping datasets are available from Linda Odenthal-Hesse or Kristian Karsten Ullrich upon reasonable request. Statistical analyses were performed by unpaired two-tailed *t*-test, if not indicated otherwise. Statistical significance was set at *P* values of * 0.05,** 0.01, and *** 0.005. Data were processed and plotted by GraphPad Prism version 6.00 (GraphPad Software, San Diego, CA; www.graphpad.com). Other types of statistical analyses are described within the text and in the corresponding figure legends.

Supplemental material available at FigShare:
https://doi.org/10.25386/genetics.9874460.

Results

Hstx2 locus is a recombination cold spot

The *Hstx2* locus was initially defined as a 4.7 Mb PWD interval present in B6.PWD-Chr X.1s (B6.DX.1s) but absent in the partially overlapping B6.PWD-Chr X.1 (abbreviated B6.DX.1) congenic strain. (Storchová *et al.* 2004; Bhattacharyya *et al.* 2014) (**Figure 6.1A**).

Here, we specified the PWD/B6 distal border of B6.PWD-Chr X.1s by next-generation sequencing to Chr X: 69.21 Mb narrowing the *Hstx2* locus to 4.3 Mb of the PWD sequence (**Figure 6.1A**). Admittedly, such subtraction mapping could not exclude the possibility that some additional genetic information in the proximal 64.9 Mb of the PWD sequence may contribute to the genetic factors situated within *Hstx2* locus. To reduce the size of *Hstx2* locus and to check the possible role of the proximal region of the X^{PWD} sequence, 52 new recombinant X chromosomes were generated in three BC1 populations (**Table 6.5**). Genotyping of 168 (B6.DX.1s x B6) x B6 BC1 mice yielded 51 recombinants with crossovers spanning the proximal region of Chr X. A new C57BL/6J-ChrX.64-69$^{PWD/Ph}$ congenic strain (abbreviated B6.DX.64-69) derived from this backcross carried only 4.34 Mb of the PWD sequence (Chr X: 64.87–69.21 Mb; mouse genome assembly GRCm38.p6), (**Figure 6.1A**). However, not a single recombination occurred in the *Hstx2* locus tracked by markers at Chr X: 65.10 and 69.08 Mb (**Table 6.5**). In the second backcross experiment, the B6.DX.51-69 subconsomic, which carries PWD sequence in the interval 51–69 Mb was used, but again no recombinant among 111 BC1 animals was found within the *Hstx2* locus. Finally, in an attempt to change the pattern of the recombination hotspots, the B6.*Prdm9*^{Hu} strain carrying the “humanized” PRDM9 with ZnF array from the human PRDM9^{A} allele (Davies *et al.* 2016) was used in (B6.*Prdm9*^{Hu} x B6.DX.64-69) x B6 backcross. No recombinant was found within the *Hstx2* locus among 369 BC1 animals. The absence of crossovers could occur due to the lack or inaccessibility of PRDM9 binding sites, the failure of SPO11 protein to target these sites and induce DNA DSBs, or because the repair of such DSBs is implemented exclusively by noncrossovers. The available data on female B6 meiosis (Brick *et al.* 2018) showed reduced occurrence of PRDM9-dependent H3K4me3 hot

spots and absence of DMC1 hotspots within the *Hstx2* locus (**Figure 6.2**), suggesting the virtual disappearance of SPO11-generated DNA DSBs as a mechanism of recombination suppression. Remarkably, in male meiosis the strong suppression of DMC1 hotspots [data from Davies *et al.* (2016)] over the *Hstx2* locus observed in (PWD x B6) and (B6 x PWD) reciprocal F_1 hybrids was attenuated in PWD and B6 parental strains (**Figure 6.3**). To conclude, no recombinant in the *Hstx2* region was found among 648 BC1 mice, although 15 recombinants would be expected ($P = 2.495 \times 10^{-7}$, binomial test) based on the 0.526 cM/Mb mean recombination rate in the adjacent Chr X: 7.36–65.10 Mb proximal region. The recombination cold spot overlaps with the interval of low PRDM9 histone methyltransferase activity and strong suppression of DNA DSB hotspots.

Targeting homologous recombination to Hstx2 by CRISPR/Cas9

Because the *Hstx2* locus behaved as a cold spot of recombination, we attempted to bring the recombination machinery to this region by means of Cas9 endonuclease-induced DSBs. Two transgenic lines were prepared, the first carrying Cas9 endonuclease under the control of SPO11 genomic region to ensure exclusive expression of Cas9 at early prophase I of meiosis. The second transgenic strain was generated with the U6-promoter driven CRISPR cassette targeted to three sites within the *Hstx2* locus (see *Materials and Methods*). Next, the double transgenic F_1 females (B6.DX.1s.TgSPO11-Cas9 x B6.TgCRISPR-*Hstx2*) were mated to B6 males to generate the BC1 population. This approach allows the generation of targeted DSB by means of a transgene that can be removed through selective breeding in a B6 backcross design. We found that double transgenic F_1 females yielded a 15-fold higher frequency of recombination in the interval spanning 64.8-65.1 Mb immediately adjacent to the *Hstx2* locus (10 recombinants in 181 BC1 offspring, 18.42 cM/Mb) compared to previous classical backcrosses (one recombination event in 279 BC1 offspring, 1.19 cM/Mb). However, only one homologous recombination event inside the *Hstx2* locus was detected, giving rise to congenic strain B6.PWD-Chr X.66-69 (abbreviated B6.DX.66-69). The new congenic restricts the PWD sequence on Chr X to 2.70 Mb in the 66.51-69.21 Mb interval. Admittedly, all these recombinants occurred within the range bracketed by the guide RNAs but at some distance away from the sites targeted. At this point, we have not determined what may have caused the increase in recombination rate close to but not involving the targeted sites.

Phenotypes of newly defined Hstx1, Hstx2, and Meir1 loci

Hstx1 fertility phenotype: To check the *Hstx1* phenotype the fertility parameters of B6.DX.64-69 and B6.DX66-69 congenic males carrying the shortened 4.34 Mb (Chr X: 64.87–69.21) and 2.70 Mb (Chr X: 66.51–69.21 Mb) of PWD sequence were compared to B6.DX.1 and B6.DX.1s males carrying 64.9 and 69.2 Mb of proximal PWD sequence (**Figure 6.1, A and B**). Both shortened intervals of the PWD sequence reduced testes weight ($P < 0.05$, *t*-test) and caused higher frequency of morphologically malformed sperm heads compared to B6.DX.1 ($P < 0.01$, *t*-test, **Figure 6.1B**). However, compared to B6.DX.1s, the level of teratozoospermia controlled by the 4.34 Mb and 2.70 Mb stretches of PWD sequence was significantly lower (40.8% *vs.* 69%, $P < 0.01$, *t*-test, **Figure 6.1B**). Thus, some additional genetic information proximal to the Chr X: 64.87–69.21 Mb interval is necessary to fully reconstruct the *Hstx1* phenotype.

Hstx2 fertility and meiotic chromosome asynapsis phenotypes: To verify the presence of *Hstx2* in the newly derived congenic strains, testes weight and sperm count were compared in F_1 hybrid males from crosses of PWD males and B6.DX.1, B6.DX.1s, B6.DX.64-69, and B6.DX66-69 females. The quasi-fertile phenotype of (B6.DX.1 x PWD) F_1 hybrids contrasted with full sterility of the remaining three types of hybrids as shown by low testes weight ($P < 0.0001$, *t*-test) and sperm count ($P < 0.0001$, *t*-test, **Figure 6.1C**). Thus in contrast to the *Hstx1* locus, the shortest version of *Hstx2* (Chr X: 66.51–69.21 Mb) was necessary as well as sufficient to fully reconstruct the (PWD x B6) F_1 male hybrid sterility phenotype.

Recently, we have found out that meiotic asynapsis of homeologous chromosomes (homologs from different subspecies) in (PWD x B6) F_1 hybrids depends on their subspecific origin and can be abolished by introduction a short stretches (27 Mb or more) of consubspecific homology into a given chromosome pair (Gregorova *et al.* 2018). Contrary to this chromosome-autonomous *cis*-control, the substitution of the $Hstx2^{PWD}$ allele for $Hstx2^{B6}$ in (B6 x PWD) F_1 hybrids significantly reduces meiotic asynapsis *in trans*, while the $Prdm9^{PWD}/Prdm9^{B6}$ genotype remains the same as in sterile hybrids (Bhattacharyya *et al.* 2014). To evaluate meiotic chromosome synapsis we visualized the axial elements of partially or fully asynapsed chromosomes by co-immunostaining of HORMA domain-containing

protein-2, HORMAD2 (Wojtasz *et al.* 2012) and synaptonemal complex protein 3, SYCP3, in pachynemas of F_1 hybrids carrying different intervals of X^{PWD} (**Figure 6.4**). The highest proportion, 85.3± 1.3%, of pachynemas affected by asynapsis was observed in the (PWD x B6) F_1 hybrid males with intact X^{PWD} chromosome. The frequencies of pachynemas with asynapsis rates 78.9 ±1.4%, 70.5± 8.6% and 70.49% in three subconsomic F_1 hybrids (B6.DX.1s x PWD) F_1, (B6.DX.64-69 x PWD) F_1, and (B6.DX66-69 x PWD) F_1 did not differ from each other, but were significantly lower than in (PWD x B6) F_1s (**Figure 6.4A**). Importantly, the X^{B6} chromosome in (B6 x PWD) F_1 did not completely eliminate the *Prdm9* controlled asynapsis, which reached 38.9± 5.2% in (B6 x PWD) F_1 hybrid males (**Figure 6.4A**). It appears that in (B6 x PWD) F_1 hybrid genomic background this level of asynapsis rate could indicate a threshold of azoospermia because (B6 x PWD) F_1 hybrid males with <40% asynapsis rate showed 7.2 ± 4.2x10^6 epididymal sperm count, while males of the same genotype with >40% asynapsis were virtually azoospermic (0.12 ± 0.1x10^6 sperm count).

To conclude, ~three quarters of the *Hstx2* effect on *Prdm9*-controlled asynapsis rate is preserved in the newly reduced 2.70 Mb PWD sequence version (Chr X: 66.51–69.21 Mb); the remaining effect either maps elsewhere on the X chromosome or is the consequence of a hypothetical position effect of the *M. m. domesticus* genome on the introgressed *M. m. musculus* sequence.

Meir1 control of global meiotic recombination rate: The *Meir1* was localized in the *Hstx2* interval as the strongest transgressive modifier of the meiotic recombination rate in B6.DX.1s males. The $Meir1^{PWD}$ coming from the high recombination rate PWD strain lowered crossover frequency in a transgressive manner when introgressed into the B6 genome (Balcova *et al.* 2016). The crossover frequency determined by counting the MLH1 foci per pachytene spermatocyte revealed that both the 4.34 Mb and 2.70 Mb PWD interval reduced recombination compared to B6 and B6.DX.1, thus behaving as *Meir1*, but the reduction did not reach the level seen in B6.DX.1s (**Figure 6.5**). We conclude that similarly as in the case of the newly defined *Hstx1* locus some additional genetic information in the proximal PWD sequence besides the 2.70 Mb interval is necessary to fully reconstruct the *Meir1* phenotype (**Figure 6.5, A and B**).

Optical mapping of intersubspecific structural variation within and outside the Hstx2 locus

One possible cause of the recombination cold spot overlapping the *Hstx2* locus could be a structural rearrangement, typically an inversion that prevents recovery of viable recombinants. Such structural variants acting as recombination suppressors often enforce reproductive isolation between species, especially when situated on sex chromosomes (Kirkpatrick 2010; Hooper *et al.* 2018). To elucidate the physical structure of *Hstx2* locus we analyzed the region by optical mapping using the Bionano Saphyr platform, a further development of the technique described by ((Chan *et al.* 2018)). As a proof of concept, we examined the *Hstx2*^{PWD} *M. m. musculus* introgression in the *M. m. domesticus* Chr X^{B6}. Indeed, the 64–69 Mb interval of Chr X was easily recognizable in two optical maps from biological replicas of B6.DX.64-69 mice when matched with the reference B6/J *in-silico* map and with the map of a female from the C57BL/6Crl substrain. The structure of the 64–69 Mb interval of Chr X matched most closely the PWD and PWK optical maps, while the flanking intervals matched the B6 optical map. (**Figure 6.6, A-C**). To inquire into the overall divergence of the *Hstx2* locus as a possible cause of recombination suppression, optical maps of the region of the same size outside the recombination cold spot (Chr X: 59.6–64.0 Mb) was compared to the *Hstx2* region (Chr X: 64.8–69.2 Mb) from four mouse strains (B6/N, B6.DX.64-69 , PWD and PWK) by alignment to the mm10 *in-silico* reference (**Table 6.6**). Although only 0.08% of the control locus sequence was involved in deletions or insertions in B6/N and B6.DX.64-69, the same 4.3 Mb control interval included 6.92% of deleted or inserted sequence in PWD and PWK. In comparison, the *Hstx2* locus (Chr X: 64.8–69.2 Mb) displayed 2 insertions of 8.7 kb and no deletion in the B6/N, representing 0.02% of the sequence, while 4.71% of sequence was either inserted or deleted in B6.DX.64-69 , 4.58% in PWD, and 5.90% in PWK. Intraspecific comparison of the same *Hstx2* interval yielded 1.11% and 2.40% of sequence involved in PWK and PWD specific inversions and deletions. To conclude, the overall structural dissimilarity is surprisingly high between *M. m. musculus* and *M. m. domesticus* subspecies, but unlikely to explain the *Hstx2* recombination cold spot.

Fine-scale screen for the Hstx2-specific structural variants

A structural variant within the *Hstx2* locus could be a marker of the *Hstx2* candidate gene. Thus we screened for PWD-specific structural variations within the *Hstx2* locus because the *Hstx2* alleles differ between *M. m. musculus* PWK and PWD and *M. m. domesticus* B6 mice (Flachs *et al.* 2014). We first aligned *de-novo* maps of B6.DX.64-69, PWD, PWK, and C57BL/6Crl to the C57BL6/J *in-silico* reference, generating a quadruple assembly (**Figure 6.6**). We then screened for structural variants that occur in B6.DX.64-69 and PWD but not in C57BL/6Crl or PWK. This had to be done semimanually, as due to the large genetic divergence in this interval, relying only on Bionano's automated algorithms was insufficient. A fine-scale characterization of the refined *Hstx2* interval by manual label matching revealed three high-confidence structural variants. The first locus, found between Chr X positions 66.756–66.797 Mb, contains two long terminal repeats (LTRs) in the B6 reference. While PWD and PWK both possess a 4.7 kb deletion of the first LTR, the second LTR locus downstream harbors a 3.1 kb deletion in PWD, also deleting miRNA *Mir883b*. In contrast, PWK shows a large overlapping 45.0 kb insertion (**Figure 6.7A**). The second significant structural variation is located between chromosomal positions 66.819–66.840 Mb, and includes the *Mir465* cluster, which appears differentially duplicated in PWK and PWD (**Figure 6.7B**). In PWD we observed an insertion of 22.9 ± 4 kb, while the PWK map revealed a shorter insertion of only 16.3 kb. Previously, we found overexpression of the *Hstx2* miRNA cluster, particularly of *Mir465* in sterile hybrids (Bhattacharyya *et al.* 2014). A differential duplication could therefore harbor subspecies-specific differences in *Mir465* expression, which may confer dosage effects on the regulation of downstream target genes.

The third is a homozygous deletion of 4574 ± 9 bp situated at Chr X: 67,787,047–67,795,903. However, this deletion neither interrupts nor deletes any known gene, mRNA/miRNA sites, or transcripts in the available testis transcriptomics data sets (Margolin *et al.* 2014; Harr *et al.* 2016; Jung *et al.* 2018) (**Figure 6.7C**). This structural variant is thus an unlikely candidate for harboring *Hstx2*.

Probing *Fmr1nb as an Hstx2 candidate gene*

The newly reduced *Hstx2* genomic interval incorporates eight protein-coding genes, of which the Fmr1 neighbor (*Fmr1nb*) appeared as the best potential candidate for the *Hstx2* gene.

The priority was based on *Fmr1nb* expression at early meiotic prophase I (Margolin *et al.* 2014; Ball *et al.* 2016; Jung *et al.* 2019; Ernst *et al.* 2019) and two nonsynonymous single nucleotide polymorphisms between PWD and B6 parental strains (**Table 6.7**). We confirmed almost exclusive expression of *Fmr1nb* in the testis, with only traces in the spleen and heart (**Figure 6.8A**) and found 2.5-fold higher expression in sterile (PWD x B6) F_1 adult testis compared to the PWD and B6 parental strains ($P < 0.001$, $P < 0.001$; *t*-test) (**Figure 6.8B**). A continuous increase of the mRNA level of *Fmr1nb* was found in juvenile males at 10, 12, 14, and 20 days of postnatal development; however, all three genotypes showed a similar expression pattern (**Figure 6.8C**). The predicted structure of the FMR1NB protein (**Figure 6.9**) consists of two cytosolic N- and C-terminal domains, two transmembrane domains, and an extracellular part containing a P-type trefoil domain. The mouse *Fmr1nb* transcripts occur in three splice variants (ENSMUSG00000062170.12, ENSEMBL) corresponding to three isoforms of FMR1NB protein (Q80ZA7, UniProt) comprising 238, 192, and 166 amino acids, respectively. We found that in the testis, the most abundant is isoform 3 (**Figure 6.9**) made up of 166 amino acids. It lacks the complete P-type trefoil domain and most of the extracellular domain. Two FMR1NB nonsynonymous substitutions create exchanges of 31 arginine^{PWD} for threonine^{B6} and 162 leucine^{PWD} for isoleucine^{B6}.

Using fluorescent immunolabelling, we detected the FMR1NB protein on histological sections of the testis of adult B6 males in the cytoplasm and spermatocyte cell membranes. The strongest FMR1NB expression was found at the leptotene and zygotene stages of the first meiotic prophase. The signal decreased in pachynemas, and disappeared in the round and elongated spermatids (**Figure 6.8D**).

Fertility phenotypes of Fmr1nb null mutants: To test the effect of the *Fmr1nb* null allele on the *Hstx1/2* phenotypes, two mouse lines carrying *Fmr1nb* deletion mutants were generated by TALEN nuclease method (see *Material and Methods* and **Figure 6.10A**). The coisogenic mouse line B6.*Fmr1nb*^{em1ForeJ} carried 236 bp deletion within the first exon and B6.DX.1s.*Fmr1nb*^{em2ForeJ} displayed a 19 bp deletion over the ATG start codon (these lines are henceforth called B6.*Fmr1nb*$^{-}$ and B6.DX.1s.*Fmr1nb*$^{–}$). The *Fmr1nb* mRNA was detectable by quantitative Reverse Transcription-PCR in both transgenic lines as expected because the transcription start site was not

affected (not shown). Three FMR1NB isoforms were identified by Western blotting in males carrying *Fmr1nb*^{B6} and *Fmr1nb*^{PWD} alleles, while the FMR1NB protein was missing in the mutant testes (**Figure 6.10B**). Intriguingly, the most truncated isoform 3 of FMR1NB was expressed most strongly in the testes of all four genotypes, whereas the longer isoforms iso-1 and iso-2 showed low expression in B6 and (B6 x PWD) F_1, and even lower in (PWD x B6) F_1 sterile hybrids and no expression in PWD and B6.DX.1s (**Figure 6.10B**). Immunohistochemistry of testes of adult wild type males showed high expression of FMR1NB in spermatogenic cells in early stages of meiotic prophase I. The protein was missing in histological sections from the B6.DX.1s.*Fmr1nb*$^{-}$ knockout males (**Figure 6.10C**) but the overall composition of testicular tubules did not show any apparent changes (**Figure 6.10D**).

The *Fmr1nb*$^{-}$ males bred successfully, but their mean litter size was significantly lower than litter size of males carrying the wild type alleles (**Figure 6.11**). The B6.*Fmr1nb*$^{-}$ and B6 males did not differ significantly in the testes weight (165.8 ± 22.2 *vs.* 180 ± 16.8 mg; $P = 0.133$, *t*-test) or in the sperm count (54.5 ± 18.2 x 10^6 *vs.* 73.3 ± 17.5 x 10^6; $P = 0.073$, *t*-test) (**Figure 6.12, A and B**), but B6.*Fmr1nb*$^{-}$ displayed a significantly higher proportion of malformed sperm heads (32.9 ± 8.6 *vs.* 19.8 ± 4.2%; $P < 0.05$, *t*-test) (**Figure 6.12C**).

The effect of the *Fmr1nb*^{PWD} null allele was stronger on the B6.DX.1s genetic background. Testes weight of the B6.DX.1s.*Fmr1nb*$^{-}$ males was significantly lower than in B6.DX.1s (148.1± 16.1 *vs.* 171.9 ± 8.8; $P < 0.001$, *t*-test) (**Figure 6.12D**) and the sperm count was lower in B6.DX.1s.*Fmr1nb*$^{-}$ than in B6.DX.1s males (44.2 ± 12.8 *vs.* 53.5 ± 10 x 10^6; $P < 0.05$, *t*-test) (**Figure 6.12E**). Furthermore, the B6.DX.1s.*Fmr1nb*$^{-}$ males showed significantly higher proportion of malformed sperm heads than B6.DX.1s control males (76.9 ± 8 *vs.* 69 ± 7.3%; $P < 0.05$, *t*-test) (**Figure 6.12F**). The frequency of apoptotic cells in seminiferous tubules assessed by fluorescence TUNEL labeling of histological sections was higher in the B6.DX.1s.*Fmr1nb*$^{-}$ males (3.36 ± 0.23) compared to B6.DX.1s males (1.46 ± 0.39, $P < 0.005$; **Figure 6.13, A and B**).

To inquire whether *Fmr1nb* interacts with the *Hstx2* phenotype, the hybrid males were analyzed for the testes weight and sperm count. Neither the *Fmr1nb*^{B6} nor *Fmr1nb*^{PWD} null allele rescued hybrid sterility; on the contrary, the *Fmr1nb*^{PWD} null allele in (B6.DX.1s.*Fmr1nb*$^{-}$ x PWD) F_1 hybrid males significantly reduced the

testes weight when compared to (B6.DX.1s x PWD) F_1 control males (59.3 ± 4.1, and 67.7 ± 3.5 mg; $P < 0.001$, *t*-test, see **Table 6.8**).

To conclude, the *Fmr1nb* on the B6 genetic background is necessary for the normal course of spermiogenesis, with stronger effects in the PWD context of B6.DX.1s *Fmr1nb* congenic males. In intersubspecific F_1 hybrids, however, the absence of FMR1NB modifies neither the intrameiotic arrest nor hybrid sterility.

Discussion

Two-gene architecture of hybrid sterility

Our model of hybrid sterility based on (PWD x B6) F_1 hybrids is composed of three main components: the *Prdm9* gene, subspecific divergence of homeologous autosomes, and the *Hstx2* locus. It differs in its simplicity from the complex genetic control reported by other studies using the same combination of house mouse subspecies (Tucker *et al.* 1992; Payseur *et al.* 2004; Macholán *et al.* 2007, 2011; Duvaux *et al.* 2011; Janoušek *et al.* 2012; Turner *et al.* 2012).

PRDM9 protein activates high number of asymmetric DNA DSBs in prophase I of (PWD x B6) F_1 primary spermatocytes, so that $PRDM9^{B6}$-determined hotspots occur mostly on the PWD chromosome and *vice versa* (Davies *et al.* 2016; Smagulova *et al.* 2016; Hinch *et al.* 2019). The main reason of hotspot asymmetry is the evolutionary erosion of the PRDM9 DNA binding sites (Baker *et al.* 2015). The predominant role of PRDM9-induced DSB asymmetry in this model of hybrid sterility was emphasized by complete recovery of spermatogenesis and fertility of the (PWD x B6) F_1 hybrids when the zinc-finger array of $PRDM9^{B6}$ was replaced with the human orthologous sequence (Davies *et al.* 2016). The hotspot erosion and meiotic failure disappeared because $PRDM9^{Hum}$, in contrast to $PRDM9^{B6}$, has never before been in contact with mouse genome. Full recovery can be also achieved by homozygosity for the $Prdm9^{PWD}$ allele (Dzur-Gejdosova *et al.* 2012).

The importance of *cis*-interaction between homeologous chromosomes was shown in intersubspecific backcross males where asymmetry disappeared in conspecific autosomal intervals (PWD/PWD or B6/B6) (Gregorova *et al.* 2018), which initially had been misinterpreted as multiple hybrid sterility QTL (Dzur-Gejdosova *et al.* 2012). The major meiotic consequences of DSB hotspot asymmetry include persistent DNA DSBs and meiotic asynapsis, both leading to apoptosis (Davies *et al.* 2016; Gregorova *et al.* 2018; Wang *et al.* 2018).

The role of *Hstx2* is apparent from attenuated manifestation of the *Prdm9*-driven asynapsis phenotype and subsequent meiotic arrest in the reciprocal (B6 x PWD) F_1 hybrids. Previously we excluded mitochondrial inheritance, the Y chromosome, and genomic imprinting as a cause and identified the *Hstx2* locus on Chr X to be the culprit (Dzur-Gejdosova *et al.* 2012; Bhattacharyya *et al.* 2014). We have not yet identified the genetic factor behind the *Hstx2* locus, so it is difficult to guess why the same pair of homeologous autosomes with the same ratio of asymmetric/novel DMC1 hotspots (Davies *et al.* 2016; Smagulova *et al.* 2016) differs so strongly in DSB repair and meiotic synapsis in the reciprocal hybrids. Three main options can be considered: *Hstx2* could extend the time window necessary to accomplish the repair of mutated PRDM9 binding sites, it could reduce the sensitivity of putative mismatch repair anticrossover activity to sequence heterology (Spies and Fishel 2015), or it may facilitate the switch of repair partner bias by sister chromatid homologous recombination (Garcia-Muse *et al.* 2019).

A recombination cold spot overlaps the Hstx2 locus

Empirical results from rabbits and mice strongly indicate that genomic regions with suppressed recombination are more differentiated and tend to accumulate reproductive isolation genes (Nachma and Payseur 2012). Ortiz-Barrientos *et al.* (2016) predicted that "…regions of low recombination will tend to harbor genes for various forms of reproductive isolation, as well as modifiers of recombination during the early stages of speciation…" Indeed, the hybrid sterility genetic locus *Hstx2* meets both of these predictions since it is situated in a recombination cold spot and carries *Meir1*, an underdominant modifier of meiotic recombination rate. Moreover, *Hstx2* operates at early stage of speciation when reproductive isolation of *Mus musculus* subspecies is still incomplete. In an attempt to reduce the size of the *Hstx2* locus by genetic recombination, we used three genetic backcrosses, one of them employing the 'humanized' *Prdm9*^{Hu} allele known to determine a DSB hotspots landscape entirely different from the *Prdm9*^{dom2} allele. However, none of these crosses was able to break the 4.3 Mb cold spot. The only recombinant which reduced *Hstx2* to 2.7 Mb was obtained in a backcross where SPO11-driven Cas9 nuclease was targeted by CRISPR to *Hstx2* interval in female meiotic prophase. Because the recombination breakpoint lies outside the targeted sites and outside

SPO11-oligo hotspots (LANGE *et al.* 2016), the possibility that this unorthodox crossover arose by repairing a Cas9- generated DSB seems unlikely.

The cold spots of recombination are often caused by heterozygosity for large structural variations, often inversions, and these 'frozen' blocks can harbor genetic factors important for reproductive isolation (Coyne and Orr 2004; Fuller *et al.* 2018). In contrast to inversions, large copy number variants can be associated with closed chromatin and reduced gene expression in germ cells, suggesting a constitutive effect on recombination by altering chromatin structure (Morgan *et al.* 2017). A constitutive cold spot model seems to better fit to the *Hstx2* locus based on the low histone methytransferase activity of PRDM9 and strong depression DNA DSB hotspots in the *Hstx2* region in female meiosis (Brick *et al.* 2018). The conclusion is also supported by recombination data from 73 sequenced inbred strains of the Collaborative Cross project (Collaborative Cross Consortium 2012; Srivastava *et al.* 2017). We found that none of the sequenced strains carries a single recombination event within the 8 Mb (Chr X: 61.8–70.3 Mb) interval spanning *Hstx2*, while 9 and 10 recombinants occurred in the adjacent 8 Mb and 6 Mb regions (http://csbio.unc.edu/CCstatus/index.py?run=CCV). In the Diversity Outbred project that used the same eight parental strains strong association between copy number variants regions and recombination cold spots was found (Morgan *et al.* 2017).

The present results based on optical mapping of a single genomic region indicate that genome-wide optical mapping can greatly contribute to elucidating the 'fluidity' of noncoding sequences between related species as well as to clarify the greater differentiation of X chromosome compared to the autosomes (Hammer *et al.* 2008; Presgraves 2018). The optical mapping enabled unprecedentedly high resolution of the *Hstx2* locus physical map in the *M. m. musculus* (PWD) and *M. m. domesticus* (B6) genome, but did not provide evidence of an inversion that could explain the recombination cold spot. Provided that the *Hstx2* phenotype is associated with a structural variant, then it should be visible in the PWD sequence, but not in PWK or B6. Three such PWD-specific variants have been found, but only one of them, including a cluster of miRNA genes, can directly implicate functional consequences related to *Hstx2*. To conclude, these results together with the recombination data from the Collaborative Cross project show that the *Hstx2* locus is located within a constitutive recombination cold spot with the chromatin structure poorly accessible to the recombination machinery.

Hstx1 and Meir1 genetic factors located in the newly defined Hstx2 locus

The *Hstx1* was mapped on Chr X as a QTL common for several male fertility phenotypes following the transgression of Chr X^{PWD} into the B6 genome. In the same experiment the suppression of recombination in the Chr X: 59.65–72.41 Mb interval (*DXMit140–DXMit199*) was noticed for the first time and the QTL for number of offspring, testes weight and sperm morphology was mapped to the interval near the *DXMit199* marker (Storchová *et al.* 2004). Later the X-linked *Hstx2* locus controlling the early meiotic arrest in (PWD x B6) F_1 hybrids was localized in the same area (Bhattacharyya *et al.* 2014).

The effect of *Meir1* genetic factor on meiotic recombination is paralleled by the male-limited transgressive/underdominant effect of *Hstx2* on hybrid sterility, since the $Meir1^{PWD}$ allele of the high recombination rate PWD parent causes downregulation of crossover rate after introgression in the low recombination rate B6 strain. Thus the localization of *Meir1* within the *Hstx2* locus indicates a link between meiotic recombination and hybrid sterility (Balcova *et al.* 2016).

In the course of positional cloning of QTL in mice and other organisms, the QTL effect sometime weakens or even disappears with narrowing down the critical region. In most instances the weakening of QTL's effect was explained by several physically linked small effects (Flint *et al.* 2005). We have seen some weakening of all three genetic factors mapping to the 2.70 Mb interval, which can be explained in the same manner. Alternatively, an epigenetic positional *cis*-effect could be involved.

The role of the Fmr1 neighbor (Fmr1nb) gene in male fertility

In the present study, we selected the *Fmr1nb* gene as the most promising candidate of *Hstx2* based on its expression pattern during meiotic prophase I and two missense polymorphisms between PWD and B6 alleles. Although the role of *Fmr1nb* in male fertility was challenged in a study of 54 testis-expressed genes (Miyata *et al.* 2016), we showed that the *Fmr1nb* null allele induced apoptosis of spermatogenic cells, elevated the frequency of sperm head malformations and decreased sperm counts. A similar general function in cellular proliferation and apoptosis was described for human FMR1NB in glioma cells (Wu *et al.* 2018). The phenotype of *Fmr1nb* null mutants, in particular the occurrence of abnormal sperm heads mimics the *Hstx1* effect. However, since teratozoospermia is a common pathological phenotype with many possible causes, and given that the null allele of *Hstx1* does

not eliminate fertility phenotype differences between B6.DX.1 and B6.DX.1s, we consider *Fmr1nb* an unlikely candidate for *Hstx1*. Moreover, since the lack of FMR1NB protein did not modulate the pachytene arrest in (PWD x B6) F_1 hybrids, we also do not consider *Fmr1nb* as candidate of *Hstx2*.

miRNA cluster variation within the Hstx2 locus

The *Hstx2* locus harbors an evolutionary conserved group of 12 testis specific miRNAs residing in two clusters of 19 and 3 miRNAs situated between *Slitrk2* and *Fmr1* protein coding genes. The conserved location of these miRNA clusters anchored between the two X-linked genes was reported in 12 mammalian species (Zhang *et al.* 2019). In spite of the interspecific variability in number of individual miRNA genes, the levels of testicular miRNAs are under regulatory constrains because depletion as well as overexpression of specific miRNA molecules or miRNA clusters can be deleterious for male fertility (Royo *et al.* 2015; Ota *et al.* 2019). The X-linked miRNAs are actively transcribed in spermatogonia and suppressed by meiotic sex chromosome inactivation in pachytene spermatocytes (Royo *et al.* 2010). Since mouse hybrid sterility is accompanied by PRDM9-controlled meiotic silencing of unsynapsed chromatin and consequent disturbance of meiotic sex chromosome inactivation (Bhattacharyya *et al.* 2013; Campbell *et al.* 2013; Larson *et al.* 2016), the uninhibited miRNA clusters could suppress genes necessary for meiosis, thus acting as “lethal mutants” contributing to meiotic arrest. Previously we have found overexpression in pachynemas of the miR-465 miRNA cluster in sterile (PWD x B6) F_1 compared to reciprocal, quasi fertile (B6 x PWD) F_1 males (Bhattacharyya *et al.* 2013). Remarkably, this cluster is subjected to copy number variation between PWD, PWK, and B6 strains. Admittedly, until we identify the gene/sequence responsible for the *Hstx2* phenotype, such speculations have to be taken with a grain of salt. Indeed, in reciprocal crosses between the *M. m. musculus* STUS strain and B6, both reciprocal hybrid males were fully sterile, showing that in this particular cross the $Prdm9^{msc}/Prdm9^{dom2}$ hybrid sterility phenotype was not dependent on *Hstx2* allele (Bhattacharyya *et al.* 2013).

Summary

Early meiotic arrest of mouse intersubspecific hybrids depends on the interaction between the *Prdm9* gene and Hybrid sterility X2 (*Hstx2*) locus on chromosome X. Lustyk *et al.* conducted high-resolution genetic and physical mapping of the *Hstx2* locus, reduced it to 2.7 Mb interval within a constitutive recombination cold spot and found that the newly defined *Hstx2* still operates as the X-linked hybrid sterility factor, controls meiotic chromosome synapsis, and modifies recombination rate. Optical mapping of the *Hstx2* genomic region excluded inversion as a cause of recombination suppression and revealed a striking copy number polymorphism of the microRNA *Mir465* cluster.

Acknowledgements

We are grateful to Vladana Fotopulosova for technical support; Inken Beck for generation of knockout mice (https://www.phenogenomics.cz/); and Lukas Cermak, Nikol Balogova, and Tomas Lidak for help with Western blotting. We thank Simon Myers for the B6.*Prdm9*^{Hu} mice, Attila Toth for HORMAD2 antibody, Cornelia Burkhardt and Sven Künzel for sample preparation and Bionano optical mapping, and Emil Parvanov and Sarka Takacova for comments. This work was supported by LQ1604 project of the National Sustainability Program II from the Ministry of Education, Youth and Sports of the Czech Republic, and by Czech Science Foundation grant GA CR No. 16-01969S to J.F., and the Charles University Grant Agency, GA UK No. 22218 to D.L., L.O.-H. and Y.F.C were supported by the Max Planck Society.

Literature Cited

Anderson, L. K., A. Reeves, L. M. Webb, and T. Ashley, 1999 Distribution of crossing over on mouse synaptonemal complexes using immunofluorescent localization of MLH1 protein. Genetics 151: 1569–1579.

Baker, C. L., S. Kajita, M. Walker, R. L. Saxl, N. Raghupathy *et al.*, 2015 PRDM9 drives evolutionary erosion of hotspots in Mus musculus through haplotype-specific initiation of meiotic recombination. PLoS Genet. 11: e1004916. https://doi.org/10.1371/journal.pgen.1004916

Balcova, M., B. Faltusova, V. Gergelits, T. Bhattacharyya, O. Mihola *et al.*, 2016 Hybrid sterility locus on chromosome X controls meiotic recombination rate in mouse. PLoS Genet. 12: e1005906. https://doi.org/10.1371/journal.pgen.1005906

Ball, R. L., Y. Fujiwara, F. Sun, J. Hu, M. A. Hibbs *et al.*, 2016 Regulatory complexity revealed by integrated cytological and RNA-seq analyses of meiotic substages in mouse spermatocytes. BMC Genomics 17: 628. https://doi.org/10.1186/s12864-016-2865-1

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober *et al.*, 2010 PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840. https://doi.org/10.1126/science.1183439

Bhattacharyya, T., S. Gregorova, O. Mihola, M. Anger, J. Sebestova *et al.*, 2013 Mechanistic basis of infertility of mouse intersubspecific hybrids. Proc. Natl. Acad. Sci. USA 110: E468–E477. https://doi.org/10.1073/pnas.1219126110

Bhattacharyya, T., R. Reifova, S. Gregorova, P. Simecek, V. Gergelits *et al.*, 2014 X chromosome control of meiotic chromosome synapsis in mouse inter-subspecific hybrids. PLoS Genet. 10: e1004088. https://doi.org/10.1371/journal.pgen.1004088

Brick, K., S. Thibault-Sennett, F. Smagulova, K. G. Lam, Y. Pu *et al.*, 2018 Extensive sex differences at the initiation of genetic recombination. Nature 561: 338–342. https://doi.org/10.1038/s41586-018-0492-5

Campbell, P., J. M. Good, and M. W. Nachman, 2013 Meiotic sex chromosome inactivation is disrupted in sterile hybrid male house mice. Genetics 193: 819–828. https://doi.org/10.1534/genetics.112.148635

Chan, S., E. Lam, M. Saghbini, S. Bocklandt, A. Hastie *et al.*, 2018 Structural variation detection and analysis using Bionano optical mapping. Methods Mol. Biol. 1833: 193–203. https://doi.org/10.1007/978-1-4939-8666-8_16

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190: 389–401. https://doi.org/10.1534/genetics.111.132639

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto *et al.*, 2013 Multiplex genome engineering using CRISPR/Cas systems. Science 339:819–823. https://doi.org/10.1126/science.1231143

Coyne, J. A., and H. A. Orr, 2004 Speciation, Sinauer Associates, Sunderland, Massachusetts.

Davies, B., E. Hatton, N. Altemose, J. G. Hussin, F. Pratto *et al.*, 2016 Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature 530: 171–176. https://doi.org/10.1038/nature16931

Dion-Côté, A. M., and D. A. Barbash, 2017 Beyond speciation genes: an overview of genome stability in evolution and speciation. Curr. Opin. Genet. Dev. 47: 17–23. https://doi.org/10.1016/j.gde.2017.07.014

Dobzhansky, T., 1951 Genetics and the origin of Species, Columbia University, New York.

Duvaux, L., K. Belkhir, M. Boulesteix, and P. Boursot, 2011 Isolation and gene flow: inferring the speciation history of European house mice. Mol. Ecol. 20: 5248–5264. https://doi.org/10.1111/j.1365-294X.2011.05343.x

Dzur-Gejdosova, M., P. Simecek, S. Gregorova, T. Bhattacharyya, and J. Forejt, 2012 Dissecting the genetic architecture of f(1)hybrid sterility in house mice. Evolution 66: 3321–3335. https://doi.org/10.1111/j.1558-5646.2012.01684.x

Ernst, C., N. Eling, C. P. Martinez-Jimenez, J. C. Marioni, and D. T. Odom, 2019 Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. Nat. Commun. 10: 1251. https://doi.org/10.1038/s41467-019-09182-1

Flachs, P., O. Mihola, P. Simecek, S. Gregorova, J. Schimenti *et al.*, 2012 Interallelic and intergenic incompatibilities of the Prdm9 (Hst1) gene in mouse hybrid sterility. PLoS Genet. 8: e1003044. https://doi.org/10.1371/journal.pgen.1003044

Flachs, P., T. Bhattacharyya, O. Mihola, J. Pialek, J. Forejt *et al*., 2014 Prdm9 incompatibility controls oligospermia and delayed fertility but no selfish transmission in mouse intersubspecific hybrids. PLoS One 9: e95806. https://doi.org/10.1371/journal.pone.0095806

Flint, J., W. Valdar, S. Shifman, and R. Mott, 2005 Strategies for mapping and cloning quantitative trait genes in rodents. Nat. Rev. Genet. 6: 271–286. https://doi.org/10.1038/nrg1576

Forejt, J., 1996 Hybrid sterility in the mouse. Trends Genet. 12:412–417. https://doi.org/10.1016/0168-9525(96)10040-8

Forejt, J., and P. Ivanyi, 1974 Genetic studies on male sterility of hybrids between laboratory and wildmice (Musmusculus L.). Genet. Res. 24: 189–206. https://doi.org/10.1017/S0016672300015214

Forejt, J., J. Pialek, and Z. Trachtulec, 2012 Hybrid male sterility genes in the mouse subspecific crosses, pp. 482–503 in Evolution of the House Mouse, edited by Macholan, M., S. J. E. Baird, P. Muclinger, and J. Pialek. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139044547.021

Fuller, Z. L., C. J. Leonard, R. E. Young, S. W. Schaeffer, and N. Phadnis, 2018 Ancestral polymorphisms explain the role of chromosomal inversions in speciation. PLoS Genet. 14: e1007526. https://doi.org/10.1371/journal.pgen.1007526

Garcia-Muse, T., U. Galindo-Diaz, M. Garcia-Rubio, J. S. Martin, J. Polanowska *et al.*, 2019 A meiotic checkpoint Alters repair partner bias to permit inter-sister repair of persistent DSBs. Cell Rep. 26: 775–787.e5. https://doi.org/10.1016/j.celrep.2018.12.074

Good, J. M., M. D. Dean, and M. W. Nachman, 2008 A complex genetic basis to X-linked hybrid male sterility between two species of house mice. Genetics 179: 2213–2228. https://doi.org/10.1534/genetics.107.085340

Gregorova, S., and J. Forejt, 2000 PWD/Ph and PWK/Ph inbred mouse strains of Mus m. musculus subspecies–a valuable resource of phenotypic variations and genomic polymorphisms. Folia Biol. (Praha) 46: 31–41.

Gregorová, S., M. Mnuková-Fajdelová, Z. Trachtulec, J. Capková, M. Loudová et al., 1996 Sub-milliMorgan map of the proximal part of mouse Chromosome 17 including the hybrid sterility 1 gene. Mamm. Genome 7: 107–113. https://doi.org/10.1007/s003359900029

Gregorova, S., V. Gergelits, I. Chvatalova, T. Bhattacharyya, B. Valiskova et al., 2018 Modulation of Prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. eLife 7: e34282. https://doi.org/10.7554/eLife.34282

Haldane, J., 1922 Sex ration and unisexual sterility in hybrid animals. J. Genet. 12: 101–109. https://doi.org/10.1007/BF02983075

Hammer, M. F., F. L. Mendez, M. P. Cox, A. E. Woerner, and J. D. Wall, 2008 Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet. 4: e1000202. https://doi.org/10.1371/journal.pgen.1000202

Harr, B., E. Karakoc, R. Neme, M. Teschke, C. Pfeifle *et al.*, 2016 Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Sci. Data 3: 160075. https://doi.org/10.1038/sdata.2016.75

Hinch, A. G., G. Zhang, P. W. Becker, D. Moralli, R. Hinch *et al.*, 2019 Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. Science 363: eaau8861. https://doi.org/10.1126/science.aau8861

Hooper, D. M., S. C. Griffith, and T. D. Price, 2018 Sex chromosome inversions enforce reproductive isolation across an avian hybrid zone. Mol. Ecol. 28: 1246–1262.

Janoušek, V., L. Wang, K. Luzynski, P. Dufkova, M. M. Vyskocilova *et al.*, 2012 Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between Mus musculus musculus and M. m. domesticus. Mol. Ecol. 21: 3032–3047. https://doi.org/10.1111/j.1365-294X.2012.05583.x

Jung, M., D. Wells, J. Rusch, S. Ahmad, J. Marchini *et al.*, 2019 Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. Elife. 25: 8. https://doi.org/10.7554/eLife.43966

Kirkpatrick, M., 2010 How and why chromosome inversions evolve. PLoS Biol. 8: e1000501. https://doi.org/10.1371/journal.pbio.1000501

Lange, J., S. Yamada, S. E. Tischfield, J. Pan, S. Kim *et al.*, 2016 The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. Cell 167: 695–708.e16. https://doi.org/10.1016/j.cell.2016.09.035

Larson, E. L., S. Keeble, D. Vanderpool, M. D. Dean, and J. M. Good, 2016 The composite regulatory basis of the large X-effect in mouse speciation. Mol. Biol. Evol. 34: 282–295.

Mack, K. L., and M. W. Nachman, 2017 Gene regulation and speciation. Trends Genet. 33: 68–80. https://doi.org/10.1016/j.tig.2016.11.003

Macholán, M., S. J. Baird, P. Dufkova, P. Munclinger, B. V. Bimova *et al.*, 2011 Assessing multilocus introgression patterns: a case study on the mouse X chromosome in central Europe. Evolution 65:1428–1446. https://doi.org/10.1111/j.1558-5646.2011.01228.x

Macholán, M., P. Munclinger, M. Sugerkova, P. Dufkova, B. Bimova *et al.*, 2007 Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. Evolution 61: 746–771. https://doi.org/10.1111/j.1558-5646.2007.00065.x

Maheshwari, S., and D. A. Barbash, 2011 The genetics of hybrid incompatibilities. Annu. Rev. Genet. 45: 331–355. https://doi.org/10.1146/annurev-genet-110410-132514

Margolin, G., P. P. Khil, J. Kim, M. A. Bellani, and R. D. Camerini-Otero, 2014 Integrated transcriptome analysis of mouse spermatogenesis. BMC Genomics 15: 39. https://doi.org/10.1186/1471-2164-15-39

Mihola, O., Z. Trachtulec, C. Vlcek, J. C. Schimenti, and J. Forejt, 2009 A mouse speciation gene encodes a meiotic histone H3 methyltransferase. Science 323: 373–375. https://doi.org/10.1126/science.1163601

Miyata, H., J. M. Castaneda, Y. Fujihara, Z. Yu, D. R. Archambeault *et al.*, 2016 Genome engineering uncovers 54 evolutionarily conserved and testis-enriched genes that are not required for male fertility in mice. Proc. Natl. Acad. Sci. USA 113: 7704–7710. https://doi.org/10.1073/pnas.1608458113

Morgan, A. P., D. M. Gatti, M. L. Najarian, T. M. Keane, R. J. Galante *et al.*, 2017 Structural variation shapes the landscape of recombination in mouse. Genetics 206: 603–619. https://doi.org/10.1534/genetics.116.197988

Myers, S., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman *et al.*, 2010 Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876–879. https://doi.org/10.1126/science.1182363

Nachman, M. W., and B. A. Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367: 409–421. https://doi.org/10.1098/rstb.2011.0249

Orr, H. A., 2005 The genetic basis of reproductive isolation: insights from Drosophila. Proc. Natl. Acad. Sci. USA 102: 6522–6526. https://doi.org/10.1073/pnas.0501893102

Ortiz-Barrientos, D., J. Engelstadter, and L. H. Rieseberg, 2016 Recombination rate evolution and the origin of species. Trends Ecol. Evol. 31: 226–236. https://doi.org/10.1016/j.tree.2015.12.016

Ota, H., Y. Ito-Matsuoka, and Y. Matsui, 2019 Identification of the X-linked germ cell specific miRNAs (XmiRs) and their functions. PLoS One 14: e0211739. https://doi.org/10.1371/journal.pone.0211739

Parvanov, E. D., P. M. Petkov, and K. Paigen, 2010 Prdm9 controls activation of mammalian recombination hotspots. Science 327: 835. https://doi.org/10.1126/science.1181495

Patten, M. M., 2018 Selfish X chromosomes and speciation. Mol. Ecol. 27: 3772–3782. https://doi.org/10.1111/mec.14471

Payseur, B. A., J. G. Krenz, and M. W. Nachman, 2004 Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution 58: 2064–2078. https://doi.org/10.1111/j.0014-3820.2004.tb00490.x

Payseur, B. A., D. C. Presgraves, and D. A. Filatov, 2018 Introduction: sex chromosomes and speciation. Mol. Ecol. 27: 3745–3748. https://doi.org/10.1111/mec.14828

Phifer-Rixey, M., and M. W. Nachman, 2015 Insights into mammalian biology from the wild house mouse Mus musculus. eLife 4: e05959. https://doi.org/10.7554/eLife.05959

Presgraves, D. C., 2018 Evaluating genomic signatures of “the large X-effect” during complex speciation. Mol. Ecol. 27: 3822–3830. https://doi.org/10.1111/mec.14777

Royo, H., G. Polikiewicz, S. K. Mahadevaiah, H. Prosser, M. Mitchell *et al.*, 2010 Evidence that meiotic sex chromosome inactivation is essential for male fertility. Curr. Biol. 20: 2117–2123. https://doi.org/10.1016/j.cub.2010.11.010

Royo, H., H. Seitz, E. ElInati, A. H. Peters, M. B. Stadler et al., 2015 Silencing of X–linked MicroRNAs by meiotic sex chromosome inactivation. PLoS Genet. 11: e1005461. https://doi.org/10.1371/journal.pgen.1005461

Sharan, S. K., L. C. Thomason, S. G. Kuznetsov, and D. L. Court, 2009 Recombineering: a homologous recombination-based method of genetic engineering. Nat. Protoc. 4: 206–223. https://doi.org/10.1038/nprot.2008.227

Smagulova, F., K. Brick, Y. M. Pu, R. D. Camerini-Otero, and G. V. Petukhova, 2016 The evolutionary turnover of recombination hot spots contributes to speciation in mice. Genes Dev. 30: 266–280. https://doi.org/10.1101/gad.270009.115

Spies, M., and R. Fishel, 2015 Mismatch repair during homologous and homeologous recombination. Cold Spring Harb. Perspect. Biol. 7: a022657. https://doi.org/10.1101/cshperspect.a022657

Srivastava, A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon *et al.*, 2017 Genomes of the mouse collaborative cross. Genetics 206: 537–556. https://doi.org/10.1534/genetics.116.198838

Storchová, R., S. Gregorova, D. Buckiova, V. Kyselova, P. Divina *et al.*, 2004 Genetic analysis of X-linked hybrid sterility in the house mouse. Mamm. Genome 15: 515–524. https://doi.org/10.1007/s00335-004-2386-0

Trachtulec, Z., M. Mnukova-Fajdelova, R. M. Hamvas, S. Gregorova, W. E. Mayer *et al.*, 1997 Isolation of candidate hybrid sterility 1 genes by cDNA selection in a 1.1 megabase pair region on mouse chromosome 17. Mamm. Genome 8: 312–316. https://doi.org/10.1007/s003359900430

Truett, G. E., P. Heeger, R. L. Mynatt, A. A. Truett, J. A. Walker *et al.*, 2000 Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). Biotechniques 29: 52, 54. https://doi.org/10.2144/00291bm09

Tucker, P., R. Sage, J. Warner, A. Wilson, and E. Eicher, 1992 Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. Evolution 46: 1146–1163. https://doi.org/10.1111/j.1558-5646.1992.tb00625.x

Turner, L. M., D. J. Schwahn, and B. Harr, 2012 Reduced male fertility is common but highly variable in form and severity in a natural house mouse hybrid zone. Evolution 66: 443–458. https://doi.org/10.1111/j.1558-5646.2011.01445.x

Wang, L., B. Valiskova, and J. Forejt, 2018 Cisplatin-induced DNA double-strand breaks promote meiotic chromosome synapsis in PRDM9-controlled mouse hybrid sterility. eLife 7: e42511. https://doi.org/10.7554/eLife.42511

Wojtasz, L., J. M. Cloutier, M. Baumann, K. Daniel, J. Varga *et al.*, 2012 Meiotic DNA double-strand breaks and chromosome asynapsis in mice are monitored by distinct HORMAD2-independent and -dependent mechanisms. Genes Dev. 26: 958–973. https://doi.org/10.1101/gad.187559.112

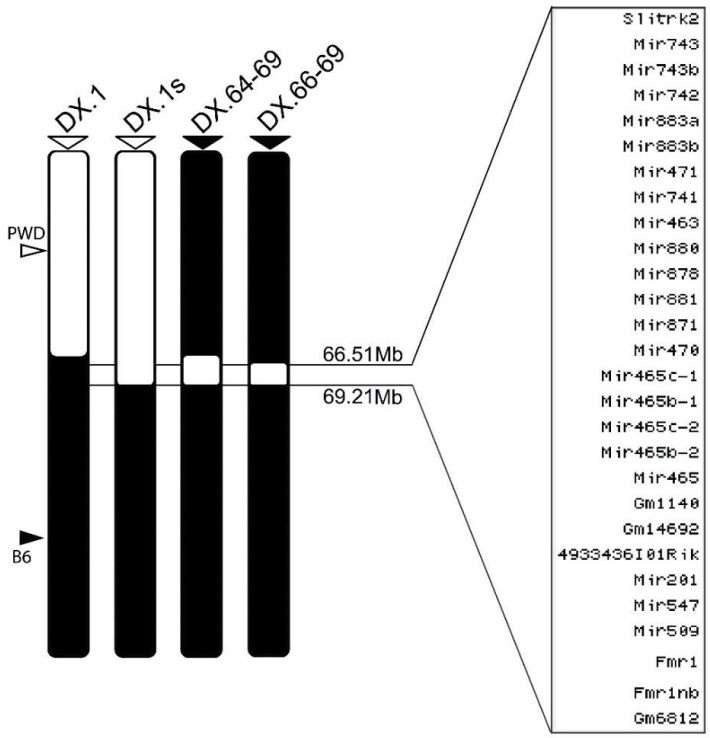
Wu, G.,W.Wang, Y. Liu, K. Zhuang, T. Cai *et al.*, 2018 RETRACTED: NY-SAR-35 is involved in apoptosis, cell migration, invasion and epithelial to mesenchymal transition in glioma. Biomed. Pharmacother. 97: 1632–1638. https://doi.org/10.1016/j.biopha.2017.11.076

Zhang, F., Y. Zhang, X. Lv, B. Xu, H. Zhang *et al.*, 2019 Evolution of an X–linked miRNA family predominantly expressed in mammalian male germ cells. Mol. Biol. Evol. 36: 663–678. https://doi.org/10.1093/molbev/msz001

Zhang, L., T. Sun, F. Woldesellassie, H. Xiao, and Y. Tao, 2015 Sex ratio meiotic drive as a plausible evolutionary mechanism for hybrid male sterility. PLoS Genet. 11: e1005073. https://doi.org/10.1371/journal.pgen.1005073

Communicating editor: F. Cole

A



B

Strain	n	Testes weight (mg)	Sperm Count ($x10^6$)	Sperm head malformation rate (%)
DX.1	9	187.4 ± 12.2	50.4 ± 12.5	24.8 ± 2.2
DX.1s	14	171.9 ± 8.8	53.5 ± 10.0	69.0 ± 7.2
DX.64-69	9	169.4 ± 26.3	50.5 ± 14.5	40.8 ± 2.8
DX.66-69	5	154.8 ± 5.0	31.3 ± 8.4	40.3 ± 4.9

C

Cross	n	Testes weight (mg)	Sperm Count ($x10^6$)
(B6 x PWD)F1	42	102.3 ± 9.4	6.2 ± 3.9
(DX.1 x PWD)F1	3	106.0 ± 8.0	7.4 ± 4.6
(PWD x B6)F1	5	65.2 ± 1.9	0.01 ± 0.02
(DX.1s x PWD)F1	20	62.3 ± 2.9	0.01 ± 0.02
(DX.64-69 x PWD)F1	25	54.7 ± 6.4	0.004 ± 0.01
(DX.66-69 x PWD)F1	20	55.5 ± 5.1	0.01 ± 0.04

Figure 6.1. Mapping of hybrid male sterility *Hstx1* and *Hstx2* loci in subconsomic and congenic strains. (**A**) Schematic view of the chromosome X architecture in subconsomic and congenic strains B6.DX.1, B6.DX.1s, B6.DX.64-69, and B6.DX.66-69. The PWD and B6 origin of chromosomal intervals is depicted in white and black. The list of protein coding genes, noncoding RNAs, and miRNAs spanning the interval of the newly defined *Hstx2* locus (66.51–69.21 Mb) is shown. (**B**) *Hstx1* locus mapping. Fertility parameters of subconsomic and congenic males; the testes weight (weight of wet testes pair in milligrams), the sperm count (number of sperms in millions per pair of epididymes) and frequency of malformed sperm heads (in percent). (**C**) *Hstx2* locus mapping. Fertility parameters of the (B6 x PWD) F_1 and the reciprocal (PWD x B6) F_1 hybrid males, and F_1 male progeny of crosses of B6.DX.1, B6.DX.1s, B6.DX.64-69, and B6.DX.66-69 congenic females with PWD males are presented as mean ±SD; *n*, number of analyzed males.

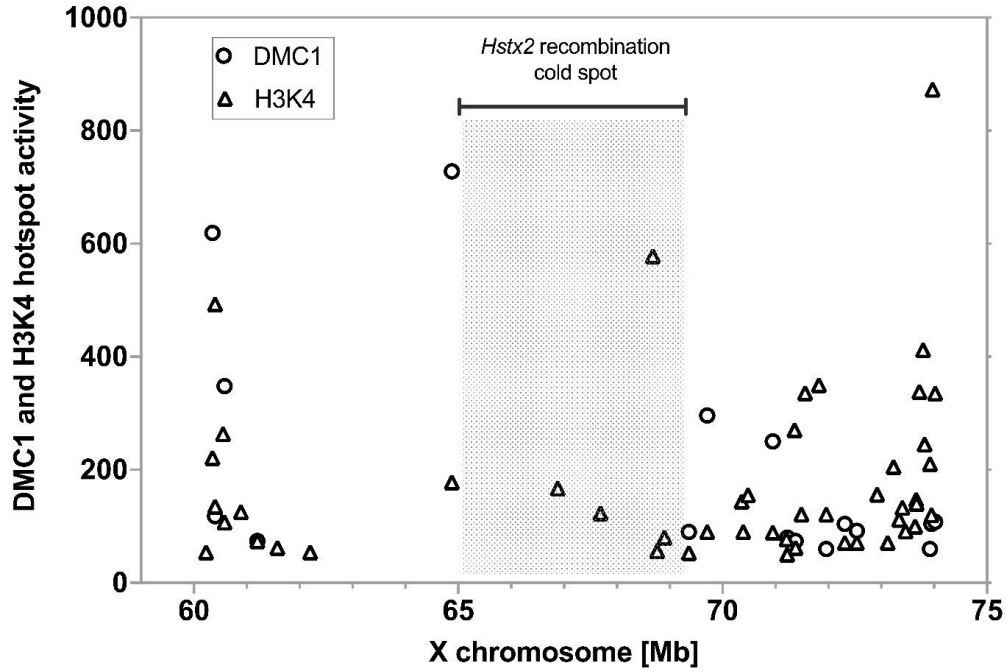


Figure 6.2. Activity of PRDM9-dependent H3K4 methylation and DMC1-marked DNA DSBs in female meiosis. The DMC1 and H3K4me3 hotspots plotted within the *Hstx2* locus and the adjacent regions of chromosome X (mm10 genome). The strong DMC1 hotspots coupled with H3K4 methylation lie outside the *Hstx2* region (shaded), which contains only H3K4 methylation marks. Data extracted from Brick *et al.* (2018); visualized are hotspots with activity >50.

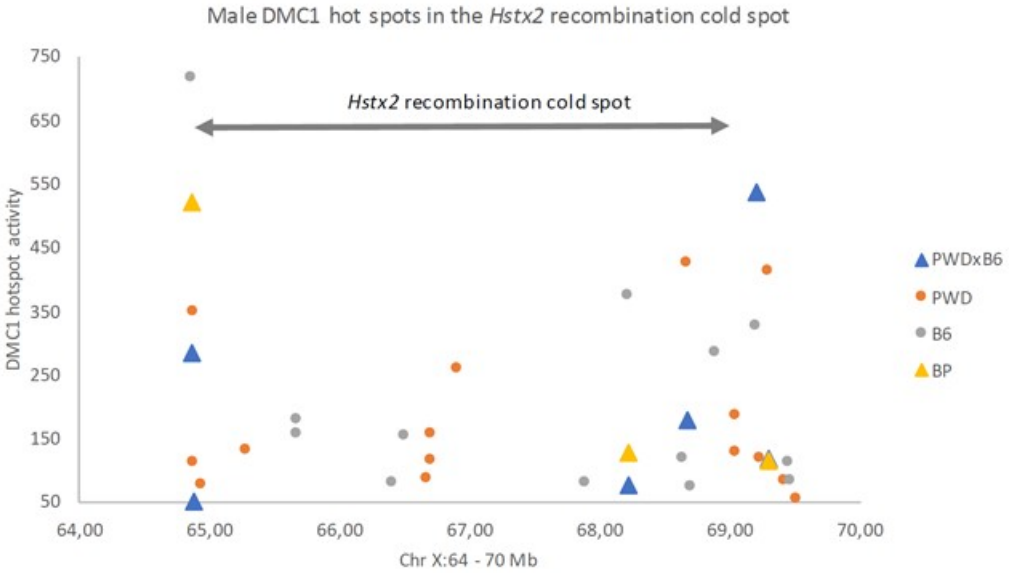
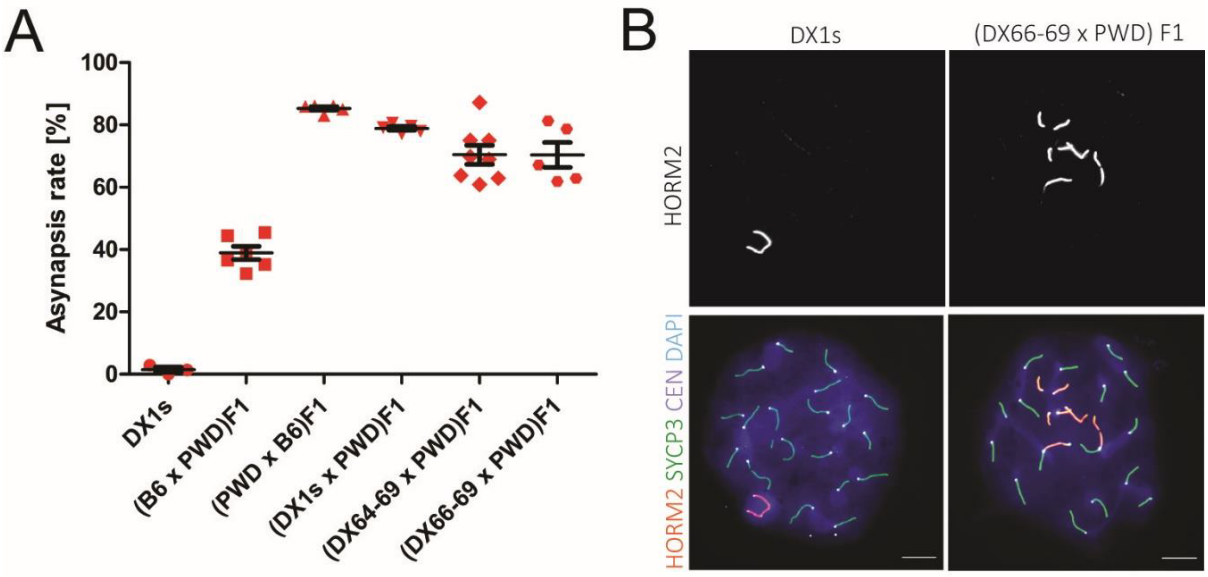


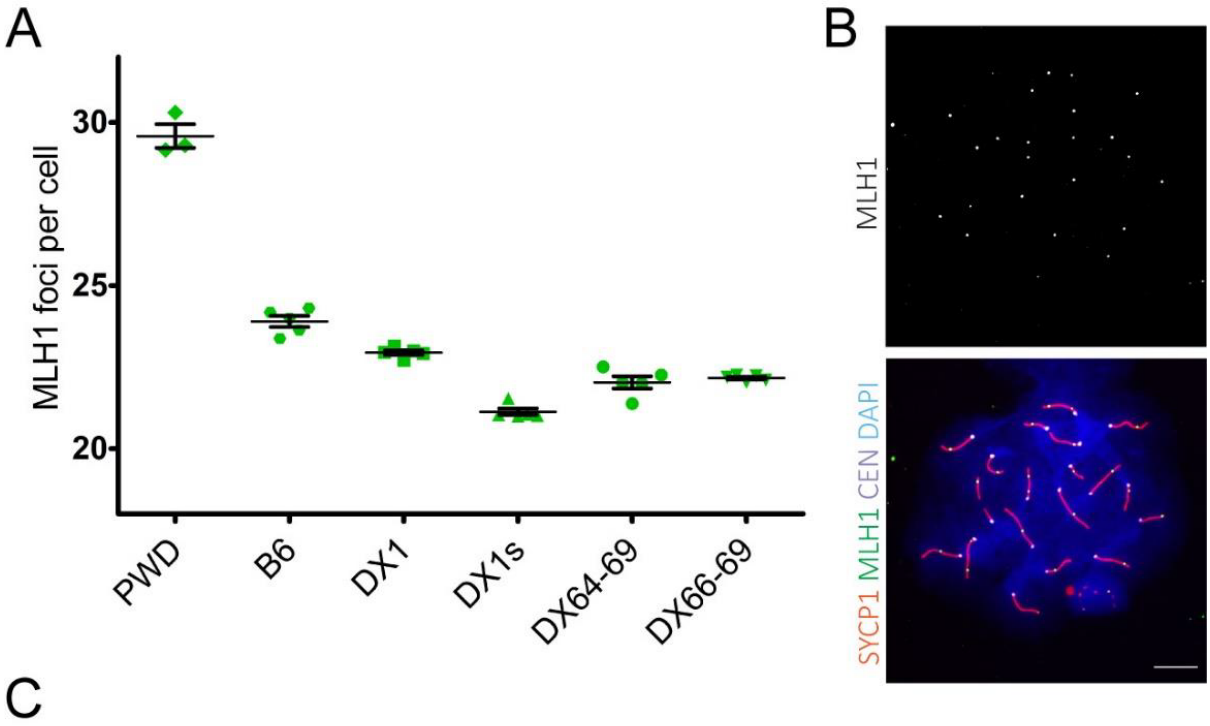
Figure 6.3. Activity of male DMC1 hotspots in the *Hstx2* recombination cold spot. The activity of DMC1 hotspots in the *Hstx2* region of the (PWD x B6) F1 and (B6 x PWD)F1 hybrid males was suppressed compared to PWD and B6 strains. Data extracted from (DAVIES et al. 2016); visualized are hotspots with activity > 50.



C

strain	DX1s	(B6 x PWD)F1	(PWD x B6)F1	(DX1s x PWD)F1	(DX64-69 x PWD)F1	(DX66-69 x PWD)F1
DX1s	X	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
(B6 x PWD)F1		X	< 0.0001	< 0.0001	< 0.0001	0.0004
(PWD x B6)F1			X	0.0001	0.0016	0.0201
(DX1s x PWD)F1				X	0.0274	0.1015
(DX64-69 x PWD)F1					X	0.9880
(DX66-69 x PWD)F1						X

Figure 6.4. Pivotal role of the *Hstx2* locus in the pachytene asynapsis rate of male F_1 hybrids. (**A**) The mean values of asynapsis rate (± SD) in F_1 and B6.DX.1s hybrid males carrying different portions of X^{PWD}. Autosomal asynapsis (frequency of pachynemas with one or more asynapsed autosomes) was examined in 5-8 animals of a given genotype, scoring at least 50 pachytene nuclei per one male. (**B**) Representative immunofluorescence micrographs show the HORMAD2-positive XY pair in a pachytene spermatocyte of B6.DX.1s congenic male and asynapsed autosomes in (DX.66-69 x PWD) F_1 hybrids. Asynapsed chromosome axes are immunostained by HORMAD2 antibody. SYCP3 visualizes lateral elements of synaptonemal complexes. CEN labels centromeric heterochromatin, and DAPI labels nuclear DNA. Bar, 10 μm. (**C**) Comparisons of the asynapsis rates between individual animal groups were performed by two-tailed *t*-test, and the *P*-values are displayed in the table.



C

	PWD	B6	DX1	DX1s	DX64-69	DX66-69
PWD	X	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
B6		X	0.0009	< 0.0001	< 0.0001	< 0.0001
DX1			X	< 0.0001	0.0019	< 0.0001
DX1s				X	0.0029	< 0.0001
DX64-69					X	0.5117
DX66-69						X

Figure 6.5. Transgressive effect of the *Hstx2*^{PWD} allele on crossover rate. (**A**) The mean crossover rate values (± SD) are shown for the subconsomic and congenic males carrying different portions of the chromosome X^{PWD} on the B6 genetic background. (**B**) Representative immunofluorescence micrograph visualizing MLH1 foci (green), synaptonemal complex protein 1, SYCP1 (red), centromeric proteins, CEN (white), and nuclear DNA (blue) in the B6.DX.1s late pachytene spermatocyte. Bar, 10 μm. (**C**) Summary of comparisons of the recombination rates between individual animal groups are shown in the table as *P*-values analyzed by unpaired two-tailed *t*-test.

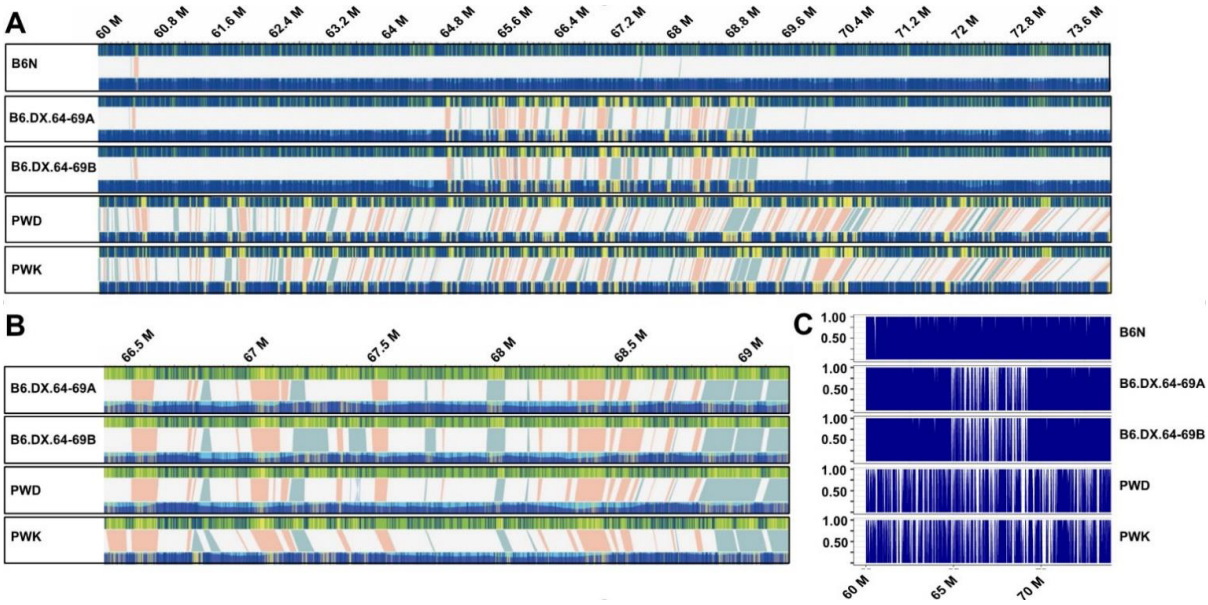


Figure 6.6. Structural variants (SVs) in the *Hstx2* locus and in flanking regions. Each box contains a comparative analysis of a *de-novo* optical map (bottom), and the mm10 *in-silico* reference B6 map (top) of a given individual. (**A**) Five maps of B6N, B6.DX.64-69A, B6.DX.64-69B, PWD, and PWK spanning Chr X 60–74 Mb (images extracted from Bionano Solve version 3.3_10252018 at maximum resolution). At this overview, individually-labeled restriction sites are not visible. However, matching intervals appear blue on both the reference and *de-novo* map, as labeled restriction sites matching their predicted position in the reference are depicted as blue lines. In contrast, labels found in either the reference or *de-novo* map, but not both, are marked by yellow lines. Therefore, clusters of mismatched labels become visible as yellow blocks. Label patterns are used to predict SVs by the Bionano Solve software. Putative SVs are depicted as shaded areas, connecting the upper reference and lower *de-novo* map. Light red areas represent putative deletions, where labels present in the *in-silico* reference, are absent in the *de-novo* map. In contrast, light blue shaded areas depict putative insertions, where additional labels were found in the *de-novo* map, but not the *in-silico* reference. (**B**) The same optical maps for B6.DX.64-69A, B6.DX.64-69B, PWD, and PWK, zoomed in to *Hstx2* position X: 66.51–69.21 Mb, which is an apparent recombination cold spot. All putative SVs are shown at higher resolution, with deletions in red and insertions in blue. Neither large inversions nor translocations have been predicted for this interval. (**C**) To quantify the number of labels matching between *in-silico* map and each of the five *de-novo* maps, we counted all labels across Chr X 60-74 M (see **Table 6.6**). Proportions of matching labels are plotted per 10 kb nonoverlapping window.

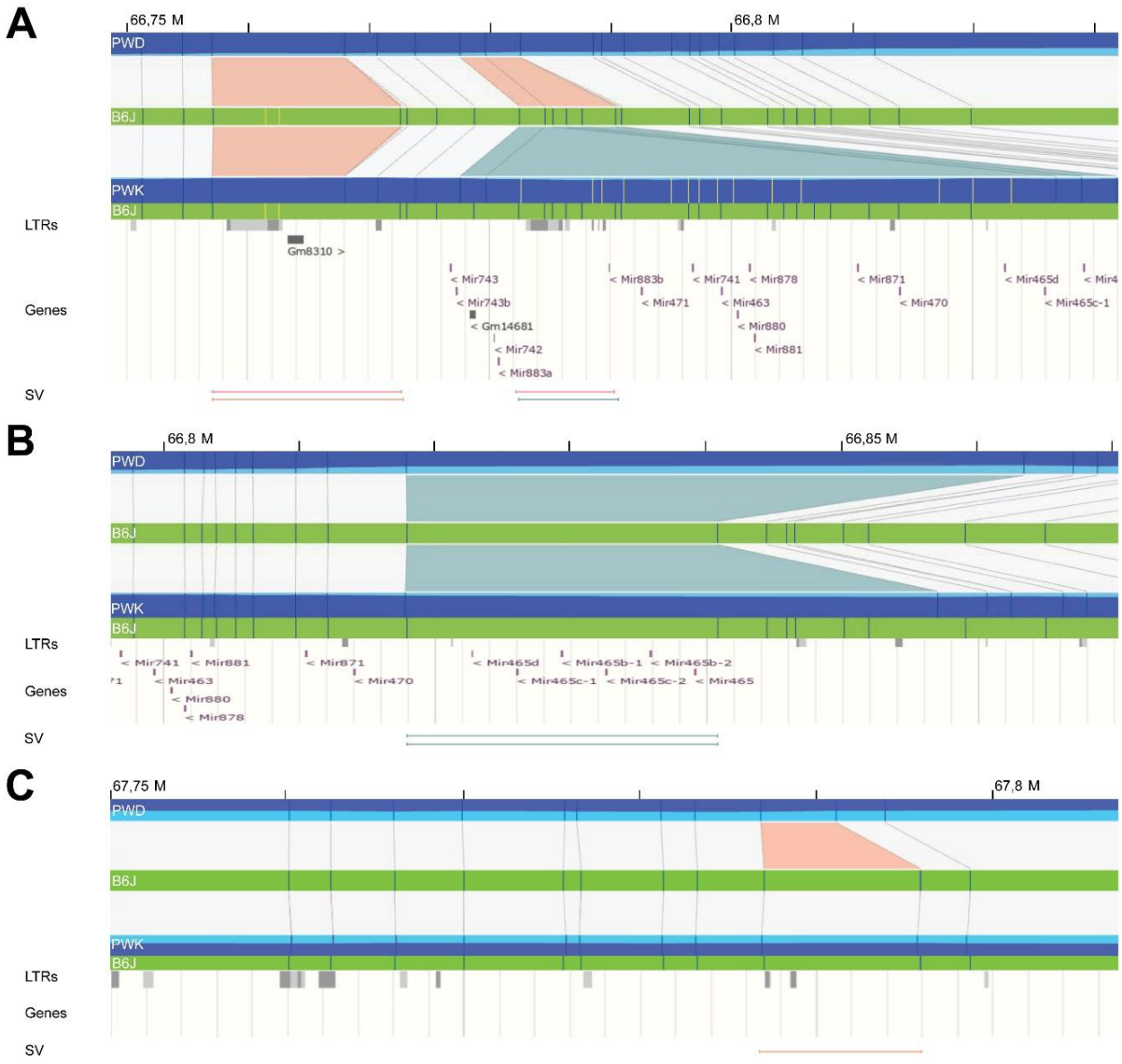


Figure 6.7. Detailed examination of polymorphic structural variation in the *Hstx2* locus. Blue vertical lines represent perfect matches to the predicted B6 *in silico* optical map (mm10), while yellow vertical lines are additional detected labels that do not match the reference. Structural variants (SVs) between the B6 reference and respective *de-novo* optical map are depicted as colored triangles, deletions in orange, and insertions in blue. At the bottom of the panel, the ENSEMBL tracks for LTRs and genes are shown, with vertical lines representing the interval affected by the SVs depicted in the top panel. (**A**) The optical maps zoomed to interval at Chr X: 66.75–66.80 Mb, revealing a polymorphic LTR region. Here, PWD possesses two deletions while PWK displays only one deletion, plus an insertion. (**B**) The optical map zoomed in at interval Chr X: 66.76–66.84 Mb. PWD and PWK both bear insertions, which duplicate the locus containing the Mir465 miRNA cluster, compared to the orthologous region in B6J. These insertions are polymorphic between the two *M. m. musculus* chromosomes, spanning only 16.2 kb in PWK but 23.3 kb in PWD (**C**) Optical map zoomed in at interval Chr X: 67.75–67.81 Mb, which possesses a deletion in PWD only. However, the deletion does not appear to disrupt any known gene.

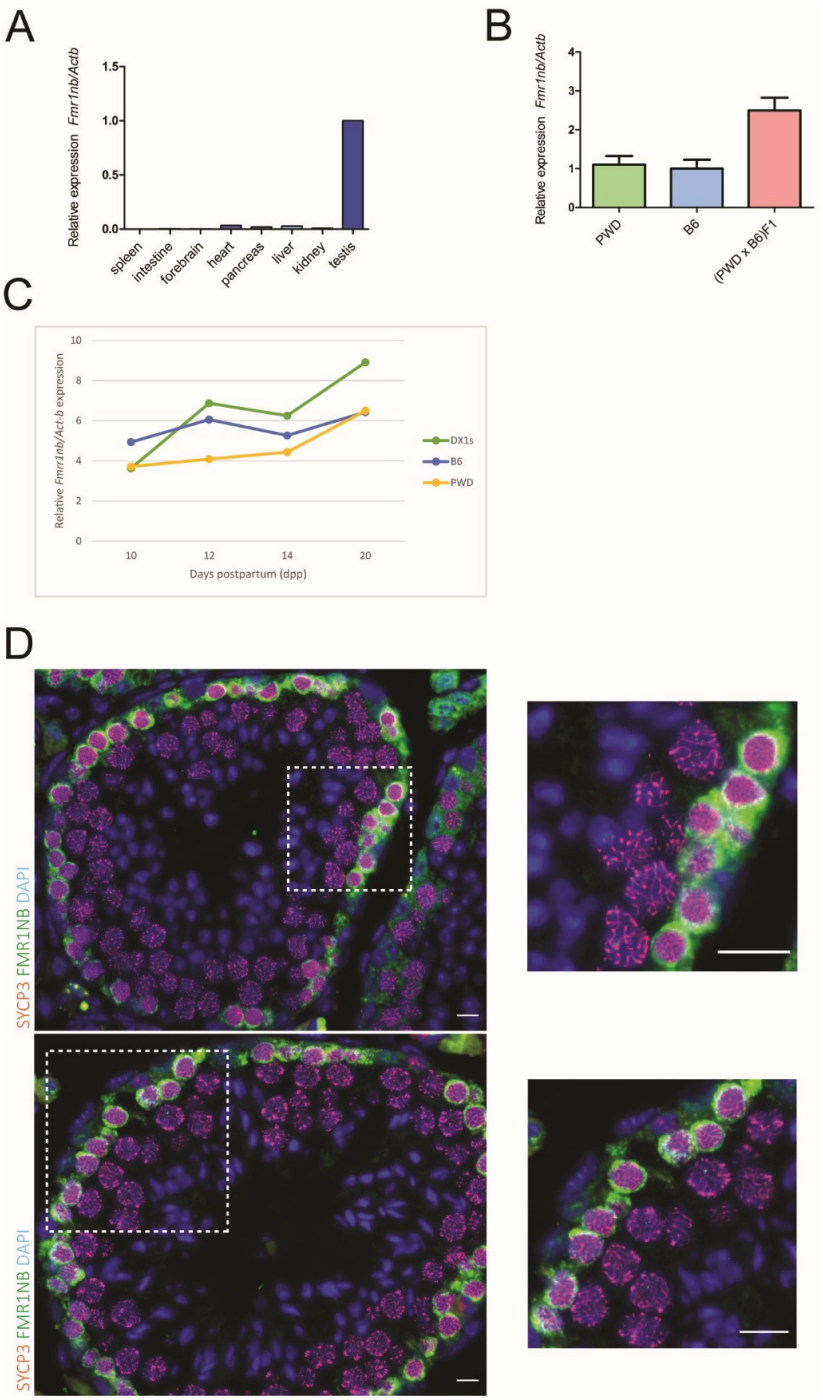


Figure 6.8. Expression of the *Fmr1nb* gene. (**A**) Tissue-specific expression of *Fmr1nb* mRNA. Relative expression of *Fmr1nb* to *Actin-b* measured by RT-qPCR and plotted for the spleen, intestine, brain, pancreas, liver, kidney and testis. (**B**) Expression of *Fmr1nb* determined by RT-qPCR in adult testis of PWD, B6 and (PWD x B6)F1 sterile hybrids. (**C**) Profiles of *Fmr1nb* mRNA in the first wave of spermatogenesis in the testis of juvenile males B6.DX.1s, PWD and B6 determined by RT-qPCR. (**D**) Immunohistochemical detection of FMR1NB and SYCP3 proteins in the histological section of testis of the B6.DX.1s mouse. FMR1NB expression appears in early stages of meiotic prophase I but not in the pachytene spermatocytes. The pachytene spermatocytes, determined by typical SYCP3 staining pattern, are shown at higher magnification. Data in B and C are presented as a mean of three independent biological replicates (±SD). FMR1NB, green; SYCP3, violet; DAPI, blue. Scale bar, 10 μm (D).

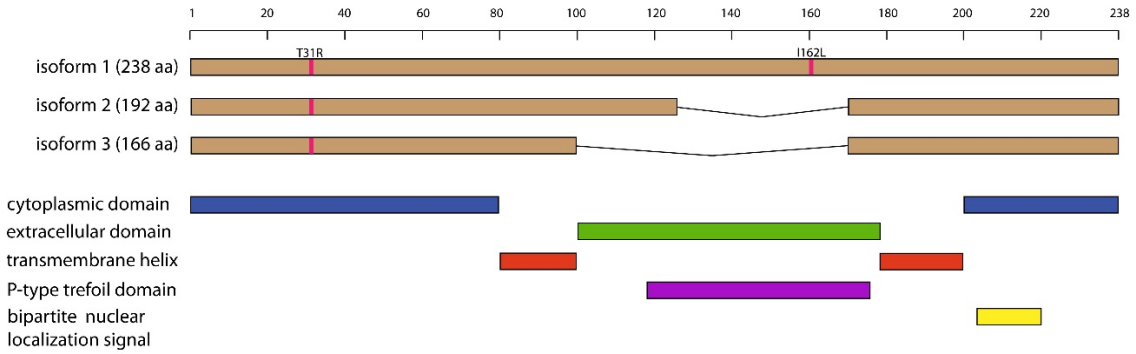


Figure 6.9. FMR1NB protein domains and isoforms. The predicted structure of the FMR1NB protein consists of two cytosolic N- and C-terminal domains, two transmembrane domains, and an extracellular part containing a P-type trefoil domain. Three isoforms of FMR1NB protein (Q80ZA7, UniProt) hold 238, 192 and 166 amino acids, respectively. The PWD and B6 allelic variants FMR1NB differ in two nonsynonymous substitutions: 31 Arginine^{PWD} / Threonine^{B6} and 162 Leucine^{PWD}/Isoleucine^{B6}. The polymorphic amino acids are highlighted in pink.

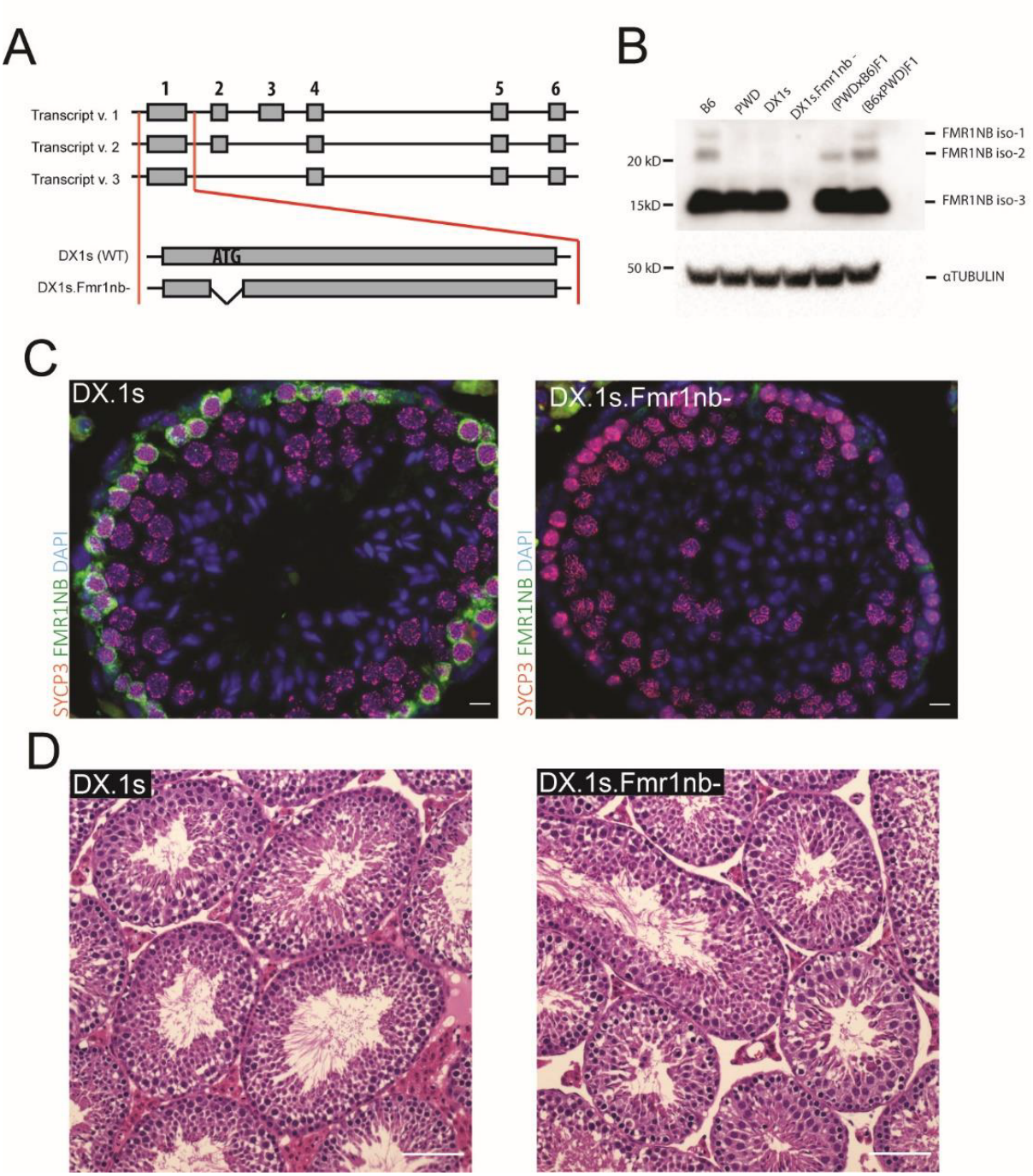


Figure 6.10. Generation of *Fmr1nb* null allele. (**A**) Transcript variants of *Fmr1nb* are shown, comprising six, five and four exons. Deletion mutants of B6 and PWD alleles of *Fmr1nb* were generated by TALEN nuclease pair constructs targeted to the ATG start codon of *Fmr1nb* in C57BL/6N (B6N) laboratory strain and C57BL/6J-ChrX.1s$^{PWD/Ph}$ (B6.DX.1s) subconsomic strain, respectively. (**B**) FMR1NB protein levels in the testes of males of indicated genotypes were assessed by Western blot. None of the three isoforms of FMR1NB was detectable in the *Fmr1nb*-deficient strain. Loading control was alpha-tubulin. (**C**) Immunolabeling of FMR1NB and SYCP3 in histological sections of testis of B6.DX.1s and B6.DX.1s.*Fmr1nb*. *FMR1NB* is shown in green, *SYCP3* is shown in violet, and DAPI is shown in blue. Bar, 10 μm. (**D**) The histological sections of testes of the B6.DX.1s and B6.DX.1s.*Fmr1nb*$^{-}$ genotype stained with hematoxylin and eosin displayed no changes in morphology and occurrence of the meiotic cells. Bar, 100 μm.

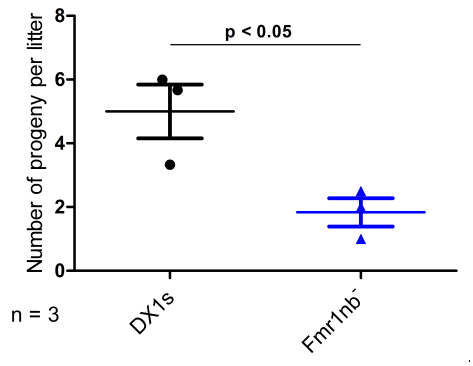


Figure 6.11. Reproductive performance of B6.DX1s and B6.DX1s.*Fmr1nb*$^{-}$ males. Mean values (± SD) of litter size sired by B6.DX.1s or B6.DX.1s.*Fmr1nb*$^{-}$ males ($p < 0.05$, t-test). The number of sired offspring was counted per each male caged individually with B6 female for three months of mating period.

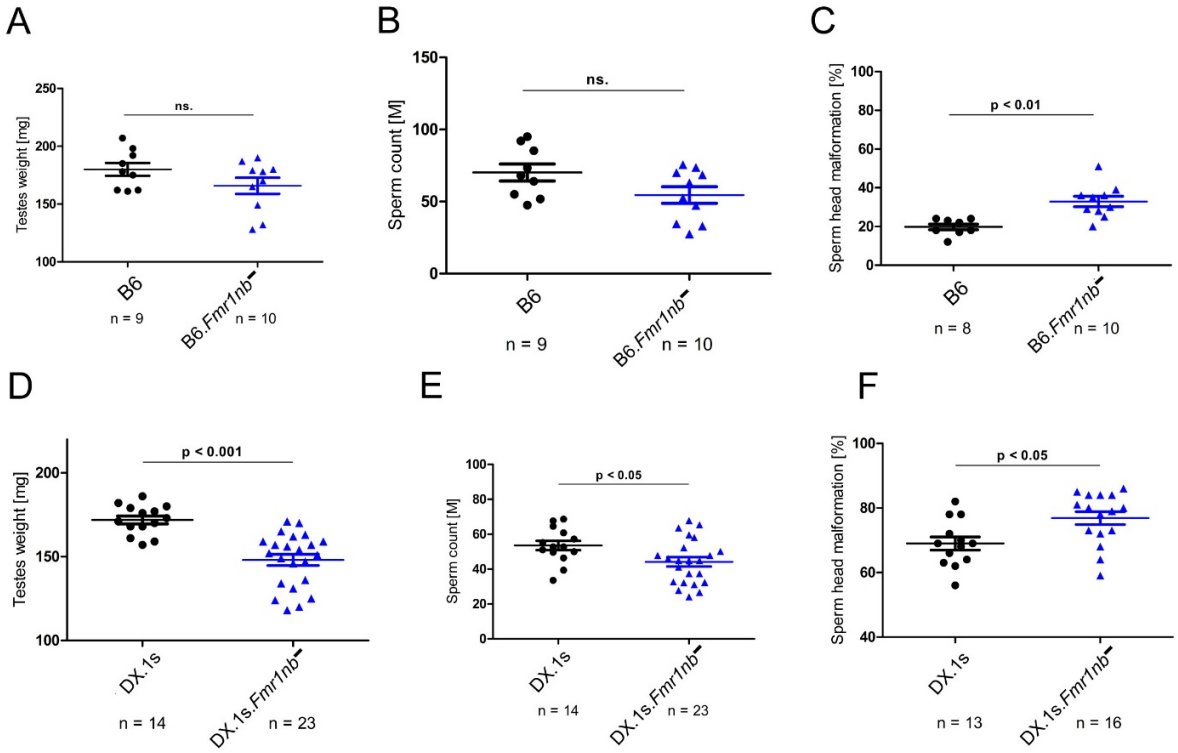


Figure 6.12. Fertility parameters of B6.*Fmr1nb*$^{-}$ and B6.DX.1s.*Fmr1nb*$^{-}$ males compared to the B6 and B6.DX.1s control counterparts. (A, B, C) *Fmr1nb*^{B6} null allele; testes weight (weight of pair of wet testes in mg), sperm count (number of sperms in millions per pair of epididymis) and frequency of malformed sperm heads (in per cent) are shown as mean (± SD) for the B6.*Fmr1nb*$^{-}$ and B6 males. **(D, E, F)** *Fmr1nb*^{PWD} null allele; Fertility parameters are plotted for the B6.DX.1s.*Fmr1nb*$^{-}$ and B6.DX.1s males. Data are presented as mean (±SD); n, number of males analyzed for the specific genotype.

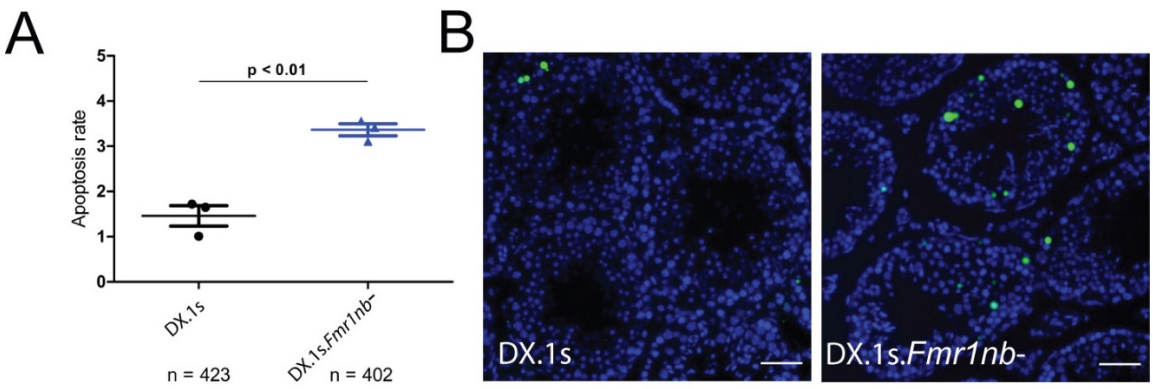


Figure 6.13. Apoptosis of spermatogenic cells in B6.DX1s and B6.DX.1s.*Fmr1nb*$^{-}$ males. (A) Average numbers of apoptotic cells per one tubule were plotted for individual males (N=3) of both genotypes (n, total number of tubules analyzed; $p<0.01$). **(B)** Apoptotic cells were visualized by FITC fluorescence using TUNEL assay in histological sections of testis of B6.DX.1s wild type and B6.DX.1s.*Fmr1nb*$^{-}$ null mutant genotypes. Scale bar, 50 µm.

Table 6.1. Microsatellite markers used for genotyping the X chromosome.

Microsatelite marker	forward primer (5' to 3')	reverse primer (5' to 3')	position (GRCm38/mm10)	PWD length (bp)	B6 length (bp)
MIT55	CTGCTTCCAGAATATTATCACTACTCC	AAAACATCCATTTATGTTAACACACA	ChrX:7360056-7360192	120	137
MIT81	GAGGAGCATCAACCTTCTCG	GAGGTGGGGAGAAACAGAGG	ChrX:36201307-36201506	190	197
MIT73	GTGCACATTTGTGTGTGTATGC	ACATGAAAGTTAGAAAGAGACCCG	ChrX:59656746-59656858	130	113
SX65100	AAAAAGGCTGCTGGAAGTCA	ATGAGGCTGGGATTCTTCCT	ChrX:65100392-65100563	162	172
SX68233	TGTGAAGTGAGGGCAGTTTG	GCTCTCCCTTTCATCGTCAA	ChrX:68233480-68233728	160	249
SX69084	AGGCCTTCTGGGCTTATCTC	AAAGCTCATGGATGAGAAAACA	ChrX:69084174-69084417	232	244

Table 6.2, Supplementary Reagent Table.

Data type	Experimental species	Symbol/name used in publication	Source – public	Source -- published	Source – unpublished	Identifiers	New reagent	Comments
Data type (mandatory) Duplicate rows as needed. Order is flexible, but row titles must be preserved.	**Experimental species** (mandatory, "NA" okay)	**Symbol/name used in publication** (mandatory)	**Source – public** [stock center; company, data repository] (one of D,E,F mandatory)	**Source -- published** [PMID or 'this paper'] (one of D,E,F mandatory)	**Source – unpublished** [description, incl. lab of origin] (one of D,E,F mandatory)	**Identifiers** [format as ID_source:identifier] Separate multiple entries with semi-colon, space	**New reagent** (mandatory for new entities) Description, progenitor(s)	**Comments** (optional) Genotypes, purpose of reagent, additional information
gene (source not applicable)	E. coli - C57BL/6J mouse BAC clone	RP23-20N4	BACPAC Resources, Oakland, CA, USA					*C57BL/6J genomic interva-Chr2:172,790,428-173,050,865*
gene (source not applicable)	NA	plasmid MLM3613	Addgene			#42251		
gene (source not applicable)	NA	plasmid pX260	Addgene			#42229		
antibody	NA	anti SYCP3 (mouse)	Santa Cruz Biotechnology			sc-74569		mouse monoclonal
antibody	NA	anti HORMAD2 (rabbit)		PMID: 22549958				rabbit polyclonal, gift from Dr. Attila Toth
antibody	NA	anti-SYCP1(rabbit)	Santa Cruz Biotechnology			ab-15087		rabbit polyclonal
antibody	NA	anti-MLH1(rabbit)	Abcam			ab14206		mouse monoclonal
antibody	NA	anti-rabbit IgG - AlexaFluor568 (goat)	Molecular Probes			A-11036		goat polyclonal
antibody	NA	anti-rabbit IgG - AlexaFluor488 (goat)	Molecular Probes			A-11029		goat polyclonal
antibody	NA	anti-mouse IgG - AlexaFluor647 (goat)	Molecular Probes			A-21235		goat polyclonal
antibody	NA	anti-Fmr1nb(goat)	Santa Cruz Biotechnology			sc-246953		goat polyclonal
antibody	NA	anti-alpha tubulin (mouse)	Proteintech			66031-1-Ig		mouse monoclonal
antibody	NA	anti-beta tubulin (mouse)	SIGMA			a5441		mouse monoclonal
antibody	NA	anti-mouse IgG - HRP (horse)	Cell Signaling Technology			#7076		horse polyclonal
antibody	NA	anti-goat IgG - HRP (donkey)	Santa Cruz Biotechnology			sc-2020		donkey polyclonal
other	NA	protease inhibitors	Roche			1836153		
other	NA	32% PARAFORMALDEHYDE AQ SOLUTION	Electron Microscopy Sciences			# 50-980-495		
other	NA	DAPI stain - mounting medium	Vectashield			H1200		
other	NA	normal goat serum	Chemicon			S26-100ML		
other	NA	TRI Reagent	SIGMA-ALDRICH			T9424		
other	NA	MuMLV-RT	Invitrogen			28025-013		
other	NA	dUTP nick end labelling (TUNEL)	Promega			G3250		
other	NA	bensonase	Merck			1.01654.0001		
other	NA	Pierce BCA Protein Assay kit	Thermo Fisher Scientific			# 23225		
other	NA	gradient Bolt 4-12% Bis-Tris plus precast gels	Invitrogen			NW04120BOX		
other	NA	Pierce™ ECL Western Blotting Substrate	Thermo Fisher Scientific			# 32106		
other	NA	mMESSAGE mMACHINE™ T7 Transcription Kit	Ambion			AM1344		
other	NA	poly (A) Tailing Kit	Ambion			AM1350		
other	NA	RNeasy Mini Kit	*Qiagen*			#74104		
other	NA	Phusion High-Fidelity DNA Polymerase	New England Biolabs			M0530		
other	NA	N*ot*I	New England Biolabs			R0189		
other	NA	N*sp*I	New England Biolabs			R0602		
other	NA	T4 Polynucleotide Kinase	New England Biolabs			M0201		
other	NA	PstI	New England Biolabs			R0140		
other	NA	SacII	Thermo Fisher Scientific			ER0201		
other	NA	MssI	Thermo Fisher Scientific			ER1341		
other	NA	Gel Extraction Kit	QIAGEN			#4993		
other	E. coli	DH5-Alpha *E. coli* competent cells	Invitrogen			#18265017		

Table 6.3. Optical mapping - Individual molecules report.

SAMPLE	Sample	Method	Enzyme	N50	Label Density (per 100 Kbp)	Total DNA >20kbp (in Mbp)	Total DNA >150kbp (in Mbp)
50058331	B6N	DLS	DLE-1	0,2647	16,36	0,452664	0,3127929
50058331	B6N	NLRS	NTBSPQ1	0,3664	12,59	0,612533	0,4533204
50065026	B6.DX64-69_A	DLS	DLE-1	0,2479	11,35	0,853258	0,3724081
0	B6.DX64-69_A	NLRS	NTBSPQ1	0,2933	14,09	0,727927	0,4395364
50065027	B6.DX64-69_B	DLS	DLE-1	0,2441	12,6	0,882681	0,4685591
50065027	B6.DX64-69_B	NLRS	NTBSPQ1	0,3139	13,39	0,656291	0,4588529
G95888	PWD	DLS	DLE-1	0,2347	15,28	0,960242	0,4761316
G95888	PWD	NLRS	NTBSPQ1	0,3056	15,3	0,675135	0,4719136
G97190	PWK	DLS	DLE-1	0,3401	12,99	0,411025	0,3144309
G97190	PWK	NLRS	NTBSPQ1	0,2891	18,75	0,483307	0,3590695

For each sample, optical mapping method and labelling enzyme, the average N50 in Megabasepairs (Mb) as well as the label density, defined as the average number of labels per 100 kilobasepair (kb) interval, was listed. As additional proxy for DNA molecule length, DNA quality and achieved optical map lengths, the table also shows the cumulative number of Mb of DNA molecules longer than 20 kb, as well as from DNA molecules longer than 150 kb.

Table 6.4. Optical mapping - Reference assemblies.

SAMPLE	Sample	Method	Enzyme	Reference	number of Contigs	Genome N50	total Genome length (Mb)
50058331	B6N	DLS	DLE-1	C57Bl/6 (mm10)	87	101,325	2615,561
50058331	B6N	NLRS	NTBSPQ1	C57Bl/6 (mm10)	1039	3,884	2615,567
50058331	B6N	DLS	DLE-1	PWK/PhJ	84	102,951	2652,997
50058331	B6N	NLRS	NTBSPQ1	PWK/PhJ	1005	4,032	2663,43
50065026	B6.DX64-69_A	DLS	DLE-1	C57Bl/6 (mm10)	109	101,496	2625,722
50065026	B6.DX64-69_A	NLRS	NTBSPQ1	C57Bl/6 (mm10)	2411	1,489	2547,894
50065026	B6.DX64-69_A	DLS	DLE-1	PWK/PhJ	127	103,381	2679,237
50065026	B6.DX64-69_A	NLRS	NTBSPQ1	PWK/PhJ	2416	1,526	2598,985
50065027	B6.DX64-69_B	DLS	DLE-1	C57Bl/6 (mm10)	73	90,738	2604,705
50065027	B6.DX64-69_B	NLRS	NTBSPQ1	C57Bl/6 (mm10)	1703	2,246	2590,975
50065027	B6.DX64-69_B	DLS	DLE-1	PWK/PhJ	75	91,015	2655,77
50065027	B6.DX64-69_B	NLRS	NTBSPQ1	PWK/PhJ	1672	2,35	2644,356
G95888	PWD	DLS	DLE-1	C57Bl/6 (mm10)	84	104,142	2609,912
G95888	PWD	NLRS	NTBSPQ1	C57Bl/6 (mm10)	1991	1,788	2614,102
G95888	PWD	DLS	DLE-1	PWK/PhJ	67	106,268	2657,694
G95888	PWD	NLRS	NTBSPQ1	PWK/PhJ	1977	1,851	2669,026
G97190	PWK	DLS	DLE-1	C57Bl/6 (mm10)	85	121,219	2641,753
G97190	PWK	NLRS	NTBSPQ1	C57Bl/6 (mm10)	1399	1,018	1220,692
G97190	PWK	DLS	DLE-1	PWK/PhJ	75	121,867	2679,769
G97190	PWK	NLRS	NTBSPQ1	PWK/PhJ	1511	1,018	1338,968

For each Sample, optical maps were obtained for two labelling enzymes. These optical maps were then aligned to optical map references of both the mm10 and the PWK/PHJ genomes. Optical map references are computed, based on the in-silico presence of enzyme recognition site in the reference genome. For each assembly, the total number of contigs, genome N50 and total assembled genome length, in Mb is summarized.

Table 6.5. Localization of PWD/B6 recombination events on the X chromosome.

	Number of recombination events (N) in the specific X chromosome intervals / recombination ratea,b (cM/Mb)					
Backcross (BC1)	**Number of BC1 (*n*)**	**X:7.36–65.10 Mb**	**X:7.36–36.20 mb**	**X:36.20–59.66 Mb**	**X:59.66–65.10 Mb**	**X:65.10–69.08 Mb**
DX.1s x B6) x B6	168	51 / 0.526[a]	17 / 0.351[a]	30 / 0.761[a]	4 / 0.438[a]	0
(DX.51-69 x B6) x B6	111	N.D.	N.D.	N.D.	1 / 0.166[a]	0
(DX.64-69 x B6.P9$^{Hu/Hu}$) x B6[c]	369	N.D.	N.D.	N.D.	N.D.	0

[a] The recombination rate (cM/Mb) was calculated from the number of recombination events (*N*) and the number of BC1 animals tested (*n*) using the length (L) of a specific region on the X chromosome.
[b] Microsatellite PCR primer sequences used for genotyping are listed in **Table 6.1**.
[c] The B6.*Prdm9*$^{Hu/Hu}$ mouse strain carries *Prdm9*$^{Hu/Hu}$ on a B6 background, which was engineered by replacing the PRDM9^{B6} zinc-finger array with the human "B-allele" zinc finger array (Davies *et al.* 2016). B6.*Prdm9*$^{Hu/Hu}$ was crossed with B6.DX.64-69, and the female progeny was backcrossed with B6 males.

Table 6.6. Insertions and deletions in the *Hstx2* locus compared to control intervals on chromosome X.

				Hstx2 locus		
Mouse strain	**Control Chr X coordinates (Mb)**	**Insertions, *n* /(kb)**	**Deletions, *n* /(kb)**	**coordinates (Mb)**	**Insertions, *n* /(kb)**	**Deletions, *n* /(kb)**
B6.N	Chr X: 59.6–64.0	1 / 0.9	1 / 2.8	Chr X: 64.8–69.2	2 / 8.8	0
B6.DX64-69_A	Chr X: 59.6–64.0	1 / 0.9	1 / 2.6	Chr X: 64.8–69.2	14 / 94.1	26 / 116.3
B6.DX64-69_B	Chr X: 59.6–64.0	1 / 0.9	1 / 2.8	Chr X: 64.8–69.2	15 / 92.6	26 / 111.6
PWD	Chr X: 59.6–64.0	22 / 105.7	21 / 174.4	Chr X: 64.8–69.2	12 / 85.8	29 / 116.4
PWK	Chr X: 59.6–64.0	21 / 113.4	24 / 192.1	Chr X: 64.8–69.2	14 / 140.3	26 / 119.4

Optical maps over the *Hstx2* region (Chr X: 64.8–69.2 Mb) and the control *Hstx2*-adjacent interval of the same size (Chr X: 59.6–64.0Mb) from five mouse genome DNA samples, representing four mouse strains, were generated and aligned to the mm10 *in silico* reference map. Coordinates are given with respect to the position in the mouse genome reference mm10 (Mb), *n* / (kb) numbers and cumulative sizes of structural variants within the intervals of the same extent in the X chromosome.

Table 6.7. *Hstx2* candidate genes.

Gene Symbol	X chromosome position [Mb][a]	Meiotic expression	SNPs (PWD/B6)
Slitrk2	66.649-66.661	POST-meiotic*	2
Gm1140	67.682-67.693	LEP, ZYG[$]	7
Gm14692	67.695-67.706	LEP, ZYG[$]	7
4933436I01Rik	67.919-67.921	RS*,[#]	7
Fmr1	68.678-68.717	LEP, ZYG*,[#]	0
Fmr1nb	68.761-68.804	LEP, ZYG*,[#]	2
Gm14698	68.821-68.825	ZG, PA*,[#]	0
Gm6812	68.892-68.893	ES*,[#]	1

The *Hstx2* locus comprises 4 protein coding and 4 predicted protein coding genes expressed in testes.
[a]Physical positions given in coordinates of mouse reference C57BL/6J genome NCBI Assembly (GRCm38/GCA_000001635.2)
Expression data were taken from: *Margolin et al. (2014), [#]Jung et al. (2018), [$]Ball et al. (2016).
Abbreviations: leptotene, LEP; zygotene, ZYG; pachytene, PA; round spermatids, RS; elongated spermatids, ES. Single nucleotide polymorphisms (SNPs) between B6 and PWD within the protein coding regions.

Table 6.8. Fertility phenotypes of (B6.*Fmr1nb*$^{-}$ x PWD) F1 and (B6.DX.1s.*Fmr1nb*$^{-}$ x PWD) F1 male hybrids.

F1 Hybrid	Number	Testes weight [mg]	Sperm Count [x106]	Sperm Head Malformation [%]
(B6.Fmr1nb- x PWD)F1	8	82.5 ± 7.62	2.73 ± 2,78	45 ± 9
(B6.Fmr1nbB6 x PWD)F1	9	82.2 ± 8.22	2.93 ± 3,78	47 ± 8
(B6.DX.1s.Fmr1nb- x PWD)F1	10	59.3 ± 4.1*	0.01 ± 0,04	N.D.
(B6.DX.1s.Fmr1nbPWD x PWD)F1	6	67.7 ± 3.5*	0.01 ± 0,01	N.D.

*Significantly different from wild type, P < 0.001, t -test; N.D. not determined