COMPUTATIONAL EXPLORATION OF THE GENETIC FACTORS BEHIND TRANSCRIPTIONAL REGULATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Shao-Pei Chou

December 2021

COMPUTATIONAL EXPLORATION OF THE GENETIC FACTORS BEHIND TRANSCRIPTIONAL REGULATION

Shao-Pei Chou, Ph. D.

Cornell University 2021

Understanding how DNA sequence affects transcription is an important first step to unravel the molecular mechanisms that cause genetic disease. Finding allele specific differences in the distribution of RNA Polymerase II (Pol II) along the genome is a powerful strategy for understanding the link between DNA sequence and the various steps in the transcription cycle. Using the natural genetic variation between the two homologous copies of the genome in diploid organisms, I can exclude most external confounding factors and identify the effect of DNA sequence differences between the copies. However, few computational methods have been developed to discover allele specific differences in functional genomic data. Existing methods either treat each SNP independently, limiting statistical power, or combine SNPs across gene annotations, preventing the discovery of allele specific differences in unexpected genomic regions. In the first part of my dissertation, I describe a new computational method, AlleleHMM, I developed which addresses this problem. AlleleHMM uses the spatial relationship among the neighboring single nucleotide polymorphisms (SNPs) to identify genomic blocks that share similar allele specific differences in mark abundance. Using both simulated and real genomic data, I found that AlleleHMM substantially outperforms

naive methods, particularly when input data has realistic levels of overdispersion. AlleleHMM is a powerful tool for discovering allele specific regions in functional genomic datasets.

In the second part of my dissertation, I describe how I used naturally occurring genetic variation in F1 hybrid mice to explore how DNA sequence differences affect the steps in the transcription cycle. To maximize allelic differences, we generated ChRO-seq data from F1 hybrids of two genetically distinct breeds of mice: C57BL/6 (B6) and Castaneus (CAST). My analysis revealed a strong genetic basis for the precise coordinates of transcription initiation and promoter proximal pause. For initiation, the data suggest that Pol II scan bidirectionally to search for an energetically favorable transcription start site within a transcription start cluster. For promoter proximal pause, the data support where paused Pol II is positioned in part through a physical interaction with pre-initiation complex. The data also show substantial allelic differences in the position of transcription termination, which frequently do not affect the composition of the mature mRNA. Finally, I identified frequent, organ-specific changes in transcription that affect mRNA and ncRNA expression across broad genomic domains. Collectively, my work reveals how DNA sequences shape core transcriptional processes at single nucleotide resolution in mammals.

BIOGRAPHICAL SKETCH

Shao-Pei was born and raised in Taichung city, Taiwan in 1983. She attended National Taiwan University and received a Bachelor of Science from Department of Life Science in 2005 and a Master of Science from Institute of Ecology and Evolutionary Biology in 2008. During her master's degree, she worked on genetic analysis of the floral shape in *Streptocarpus* (Gesneriaceae) and developed an interest in genetics. In 2008, while the next generation sequencing (NGS) was getting prevalent, Shao-Pei joined Dr. Wen-Hsiung Li's lab, one of the few labs having NSG machine in Taiwan, and was trained to prepare libraries and operate the NGS machine. With the exposure to genome studies in Li lab, she developed a strong interest in genomics. In 2012, she joined the Ph.D. program in the graduate field of Genetics, Genomics, and Development at Cornell University. She first joined Dr. Ruth Ley's lab to study the relationship between human gut microbiome and virome and had a co-first author paper published in cell host and microbe. In 2016, when Dr. Ley moved to Max Planck Institute in Germany, Shao-Pei joined Dr. Charles Danko's lab to study gene regulation. With a biology background, Shao-Pei showed her interest in computation early on during her PhD. She took courses from Department of Computer Science, from the basic 101 to machine learning. In Danko lab, she applied her machine learning skills to develop a bioinformatic tool to analyze functional genomic data from diploid organisms with allelic difference and published the tool AlleleHMM in Nucleic Acids Research. She then switched her focus and use genetics to explore the molecular functions in transcriptional regulation.

I would like to dedicate my dissertation to my family for their unwavering support.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my mentor, Dr. Charles Danko for being extremely supportive especially during the most difficult times. His enthusiasm at science, confidence in me, and the productive discussions helped me to reach this step. I also would like to thank Dr. Andrew Clark and Dr. Haiyuan Yu for serving on my Special Committee and providing insightful feedbacks during our meetings. I would like to thank each of my colleagues in Danko and Ley lab for shaping me as a scientist and for making my time at Cornell a wonderful experience. I would like to thank my officemates, Dr. Lauren Choate, Dr. Paul Munn, Dr. Zhong Wang, and Dr. Tinyi Chu for their helps and companionship, which I especially missed and appreciated during the lock down in the pandemic. I would like to thank Ed Rice and Adriana Alexander for their help with experiments. I would like to thank each person I interacted with during my time at Cornell for making the experience richer and more memorable. I would like to thank my friends at Baker Institute, especially Brynn Alford for making my time at Baker beyond enjoyable. I thank my horse Willy for being my mental support throughout the years in Ithaca. Finally, I would like to thank my family for their unconditional support.

TABLE OF CONTENTS

CHAPTER 1

AlleleHMM: a data-driven method to identify allele specific differences in distributed functional genomic marks†

Shao-Pei Chou[1, 2], Charles G. Danko[1, 3]

1. Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA.
2. Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA.
3. Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA.

† This chapter was published as an article in Nucleic Acids Resarch in 2019. The citation is doi: 10.1093/nar/gkz176 (PMID: 30918970)

1.1 Abstract

How DNA sequence variation influences gene expression remains poorly understood. Diploid organisms have two homologous copies of their DNA sequence in the same nucleus, providing a rich source of information about how genetic variation affects a wealth of biochemical processes. However, few computational methods have been developed to discover allele specific differences in functional genomic data. Existing methods either treat each SNP independently, limiting statistical power, or combine SNPs across gene annotations, preventing the discovery of allele specific differences in unexpected genomic regions. Here we introduce AlleleHMM, a new computational method to identify blocks of neighboring SNPs that share similar allele specific differences in mark abundance. AlleleHMM uses a hidden Markov model to divide the genome into three hidden states based on allele frequencies in genomic data: a symmetric state (state S) which shows no difference between alleles, and regions with a higher signal on the maternal (state M) or paternal (state P) allele. AlleleHMM substantially outperformed naive methods using both simulated and real genomic data, particularly when input data had realistic levels of overdispersion. Using global run-on sequencing (GRO-seq) data, AlleleHMM identified thousands of allele specific blocks of transcription in both coding and non-coding genomic regions. AlleleHMM is a powerful tool for discovering allele specific regions in functional genomic datasets.

1.2 Introduction

DNA encodes the blueprints for making an organism, in part by coordinating a complex cell-type and condition-specific gene expression program. Regulatory DNA affects gene expression by controlling the rates of a variety of steps during the transcription cycle, including opening chromatin, decorating core histones and DNA with chemical modifications, initiating RNA polymerase II (Pol II) at transcription start sites (TSSs), and releasing Pol II from a paused state into productive elongation (Fuda et al., 2009). In addition, mRNAs are subjected to a host of post-transcriptional regulatory processes, most of which are influenced by the sequence of the RNA (Corbett, 2018). How DNA or RNA sequences control each step during transcription, mRNA processing, and mRNA degradation remains poorly understood.

Finding allele specific differences in the distribution of marks along the genome is a powerful strategy for understanding the link between DNA sequence and the various biochemical processes that regulate gene expression (Castel et al., 2015; Rozowsky et al., 2011). Diploid organisms have two copies of their DNA sequence in the same nuclear environment, providing a rich source of information about how genetic variation affects biochemical processes. Additionally, alleles in a diploid genome share the same environmental signals, cell type-specific differences within a complex tissue, and other potential confounding factors. Therefore, allele specific signatures are a rigorous source of information about how DNA sequence affects gene expression.

Despite the general utility of allele specific expression measurements, surprisingly few computational methods have been proposed to detect allelic differences. Current methods that examine allele specific enrichment either test single-nucleotide

polymorphisms (SNPs) independently (Chen et al., 2016; Rozowsky et al., 2011) or combine the location of SNPs using gene annotations (Crowley et al., 2015). Each of these methods have important limitations. Treating SNPs independently requires a high sequencing depth, and exhibits a bias in which regions with higher abundance of the mark of interest are much more likely to be discovered. Summing up the reads within contiguous genomic regions, such as annotated genes, can improve sensitivity and reduce bias by pooling information across SNPs that are more likely to share the same allele specificity. However, combining reads requires a well-annotated reference genome, which is not available in some species, and also prevents the analysis of marks in unannotated or non-coding regions which are critical for proper genome function.

Here we introduce AlleleHMM, a novel computational tool that was designed to address these limitations. AlleleHMM identifies genomic blocks of SNPs that share the same allele specificity in mark abundance using a hidden Markov model (HMM). We show that AlleleHMM has significantly higher sensitivity and specificity when compared to tests that treat each SNP independently. AlleleHMM has similar statistical power compared to the practice of merging reads inside of gene annotations, and can also identify allele specific differences in unannotated non-coding RNAs when run genome-wide. When applied to publicly available global run-on sequencing (GRO-seq) data, a direct measurement of RNA polymerase, AlleleHMM discovers the location of the vast majority of genes discovered by merging gene annotations, and also identified over one thousand allele specific blocks that lie in unannotated genomic regions. Blocks of allele specific transcription are inversely correlated with the allele specific

differences in repressive chromatin marks. Thus, AlleleHMM is a powerful new strategy to identify allele specific differences in functional genomic data.

1.3 Materials and Methods

Overview of AlleleHMM: The primary goal of AlleleHMM is to identify allele specific blocks of signal in distributed functional genomic data. AlleleHMM relies on the key assumption that contiguous genomic regions share the same allele specificity (**Figure 1.1A**). This may happen for a variety of reasons depending on the genomic mark; for example RNA polymerase across a transcription unit shares the same allele specific differences that were derived from the rates of Pol II initiation or release from pause on the promoter which controls expression of that transcription unit (Fuda et al., 2009).

We developed a HMM (Durbin et al., 1998) that represents allele specificity in a distributed genomic mark using three hidden states: symmetric (S) distribution of the mark from both alleles (which shows no allele specificity), and an imbalance of the mark specific to either the maternal (M) or paternal (P) alleles (**Figure 1.1B**). AlleleHMM takes as input read counts corresponding to each allele, computed using AlleleDB (Chen et al., 2016; Rozowsky et al., 2011). AlleleHMM uses this information to set the parameters of the HMM using Baum–Welch expectation maximization (BAUM and L, 1972), save for a single holdout parameter used to tune the balance between sensitivity and specificity of AlleleHMM (see below). The Viterbi algorithm (Viterbi, 1967) is then used to identify the most likely hidden states through the data, resulting in a series of candidate blocks of signal that show evidence of allele specificity. We last calculated the coverage of allele specific read counts in each predicted AlleleHMM block and performed a binomial test to verify that the block predicted by the HMM is significantly allele specific. The last binomial test was performed to

eliminate any false positives that result from multiple counts of a single read that map to multiple nearby SNPs, which are difficult to handle in the context of the HMM.

AlleleHMM can be downloaded from: https://github.com/Danko-Lab/AlleleHMM.

HMM structure

There are three hidden states in AlleleHMM (**Figure 1.1B**): (S) a symmetric state which represents no allele specificity, and (M) or (P) which represent regions with evidence of allele specificity on the maternal or paternal allele. Each state can transition to the other two states or stay in the original state. We used allele specific read counts of SNPs with at least one mapped read as observed emissions for AlleleHMM. The distance between SNPs were not considered in the model.

Transition probability

Transitions are permitted between all of the hidden states (**Figure 1.1B**). To control the balance between sensitivity and specificity of AlleleHMM, we used a tuning parameter, τ (**Figure 1.1B**), to limit the transition out of the M or P states. We set the transition probability from the M or P states to any of the other states to $\tau/2$, and the transition probability to stay in the M or P state to $1-\tau$. The transition probability of the S state to either M or P were set using the Baum–Welch expectation maximization (EM) algorithm (BAUM and L, 1972; Durbin et al., 1998).

Emission probability

Each hidden state in AlleleHMM is associated with a separate probability distribution (called the "emission probability") that is used to represent the input data. The emission probabilities for all three states were calculated using the binomial distribution. AlleleHMM is given the total read count and maternal read count on each SNP in the genome. The input data is provided as:

SNP:	1,	2,	3,	…	l_c
Total read count:	n_1,	n_2,	n_3,	…	n_l
Maternal read count:	x_1,	x_2,	x_3,	…	x_l

Given these input data, the emission probability for state *j* is defined as:

Emission probability $=\prod_{autosomes}\prod_{i=1}^{l_c}\frac{n_i!}{x_i!(n_i-x_i)!}{p_j}^{x_i}(1-p_j)^{(n_i-x_i)}$

The value p_j for each state (S, M, and P) was estimated from the data using the EM algorithm, as described below.

Learning procedure for transition and emission probabilities

We use the Baum–Welch EM algorithm (BAUM and L, 1972) to learn the transition and emission parameters in AlleleHMM (with the exception of the user adjustable tuning parameter, τ). The EM algorithm uses the forward-backward algorithm to compute the probability of each state at every heterozygous SNP in the genome with one or more mapped read (E-step). These probabilities are used to maximize the likelihood of the model given the input data (M-step). After learning model parameters using EM, we identify the most likely path of hidden states at every mapped

heterozygous SNP using the Viterbi algorithm (Viterbi, 1967). These algorithms are described in detail by Durbin et al. (Durbin et al., 1998).

Tuning parameter optimization using GRO-seq data

AlleleHMM defines a user-adjustable tuning parameter, τ, that controls the balance between sensitivity and specificity by limiting the frequency of transitions between model states. Intiutively, users can think about τ as the threshold p-value that must be overcome by the input data to support a state transition.

The optimal value of τ depends on the type of data, the sequencing depth, and the heterozygosity of the genome of interest. To determine the optimal value of τ for GRO-seq data in this manuscript, we assumed that changes in allele specificity should usually arise near a transcription start site (TSSs) (**Supplementary Figure 1.1A**), because allele specificity across a gene is controlled by the rate of initiation and release from pause at the TSS (Fuda et al., 2009). We evaluated the proportion of AlleleHMM blocks that start within a fixed distance of a TSS defined using dREG (discriminative regulatory-element detection from GRO-seq) (Danko et al., 2015; Wang et al., 2018) over a range of τ in the dataset of interest. As τ increased, a larger fraction of AlleleHMM blocks occur within a predefined distance of a dREG annotated TSS (**Supplementary Figure 1.1B**). We selected the τ for each dataset at which this value approached a saturation point. For instance, in the 129/castaneus F1 hybrid mouse embryonic stem cells (mESC) dataset, as τ approached 1e-05, the fraction of AlleleHMM blocks beginning within 5 kb of a dREG site saturated at ~50% (**Supplementary Figure 1.1B black line, Supplementary Figure 1.1C**). Although the

value of saturation varied with different window sizes, the value of τ that defined the saturation point was fairly well conserved over a reasonable range of τ (**Supplementary Figure 1.1B**). In analyses that follow we fixed τ to 1e-05 (full GM12878 dataset, mESC dataset). For subsampled GM12878, the same criteria used above suggested an optimal value of τ as 1e-04 (50% of total reads), 1e-03 (25% of total reads), and 1e-2 (12.5 and 6.25% of total reads).

To provide an orthogonal validation for the value of τ selected using the TSS strategy described above, we measured the sensitivity and specificity of AlleleHMM over different values τ in each dataset, using gene annotations as a gold-standard (see below; **Supplementary Figure 1.1D-F**). We found that setting τ using the saturation point of TSSs generally produced a sensitive model with a specificity near 1.0, as desired for the problem of discovering allele specific blocks in the unbalanced setting of an entire genome.

Preparing data for input to AlleleHMM

Allele specific read counts were computed using AlleleDB (Chen et al., 2016; Rozowsky et al., 2011) with some modifications. Briefly, reads were mapped to paternal and maternal genomes separately using Bowtie (14). Reads with ambiguous mapping bias were removed, following the procedure in AlleleDB (5). The Bowtie output of each parental genome was merged and each read was assigned to either the paternal or maternal haplotype based on the amount of difference between the read and each individual genome. Reads that differed from both individual genomes by the same amount were assigned randomly to one of the individual genomes. The merged Bowtie

output was separated to plus strand and minus strand using in house scripts because AlleleDB was designed for non-strand-specific datasets. The allele specific read counts of each SNP on each strand were computed using AlleleDB (Chen et al., 2016; Rozowsky et al., 2011). The AlleleDB output was further parsed into a tab-delimited table as shown in **Supplementary Table 1.1**.

An example input file can be found here: https://github.com/Danko-Lab/AlleleHMM/blob/master/input_file_exmaples/AlleleHMM_input.txt

Identify allele specific transcribed blocks

We used the Viterbi algorithm (Viterbi, 1967) to identify the most likely set of states (M, P, or S) at each heterozygous position in the genome of interest. Nearby SNPs sharing the same hidden state were stitched into blocks. We then calculated the coverage of reads in each block as follows: Reads were mapped to diploid genome using Bowtie (Langmead et al., 2009) as implemented in AlleleDB (Chen et al., 2016). The Bowtie output, including reads and their mapping position, were separated into maternal- and paternal-specific text files. Then, the coordinates were transferred to the appropriate reference genome (mm10 or hg19) using liftOver (Casper et al., 2018). We used the older hg19 human reference genome, because existing fully phased personal references for each of the two GM12878 haplotypes were available in this coordinate system (Chen et al., 2016; Rozowsky et al., 2011), but not in the newer hg38. We calculated the number of reads from the maternal or paternal haplotype in each AlleleHMM block using bedtools (Quinlan and Hall, 2010). Binomial tests were performed for each block and the false discovery rate was estimated to correct for multiple hypothesis testing

(Benjamini and Hochberg, 1995). These steps were performed to eliminate false positives derived from multiple counts of a single read that mapped to multiple nearby SNPs, which were more complicated to consider in the context of our HMM. AlleleHMM outputs two bed files: one with all blocks and the other only reports the blocks with a FDR <= 10% as significantly allele specific.

Performance test with synthetic data

To test how the performance of AlleleHMM compared with current standards in the field, which involve testing SNPs independently, we developed a simulation strategy. We used a simulation strategy because the location and magnitude of allele specific blocks could be controlled precisely, providing a confident ground truth dataset for a rigorous performance evaluation. The synthetic data was composed of three blocks, each representing a region with allele specificity as shown on the top of **Figure 1.2**. The flanking blocks were symmetric, and were simulated using parameters that were kept consistent throughout the study.

The following parameters were changed to simulate the middle block with allele specific transcription: length, expression level, and the degree of allele specificity. Length was defined as the number of continuous SNPs sharing same allele specificity, and was set to 100 when testing other parameters, within the range of a typical gene in F1 hybrid mice (**Supplementary Figure 1.2A, top**). Expression level, or the average read count per SNP in the block, was set to 10 when testing other parameters. The degree of allele specificity was defined as the probability that a read comes from the maternal allele in a binomial or beta-binomial event, and was set to 0.9 when testing other

parameters. The total read counts of each location were simulated with a poisson distribution and the allele specific read count was simulated by either the binomial or beta-binomial distribution with overdispersion of 0.25. The overdispersion of 0.25 was chosen based on the estimates of two real data sets: GRO-seq of GM12878 and PRO-seq of 129/castaneus F1 hybrid mESCs. The estimate was performed in R using the VGAM library (Yee, 2015) using all SNPs covered by at least 5 reads. The parameters used are summarized in **Supplementary Table 1.2**.

Performance test with real biological data

To test the performance of AlleleHMM with GRO-seq data, we applied AlleleHMM and independent binomial tests implemented in AlleleDB to GRO-seq from GM12878 and 129/castaneus F1 hybrid mESCs. The allele specificity of each GENCODE gene annotation was estimated and used as a surrogate for the ground truth. Allele specific reads within each GENCODE gene annotation were counted using bedtools (Quinlan and Hall, 2010). Binomial tests were then performed for each gene annotation and the false discovery rate was used to correct for multiple hypothesis testing. We used Release 28 (mapped to hg19/ GRCh37) for GM12878 and Release M17 (mm10/ GRCm38.p6) for 129/castaneus F1 hybrid mouse. The sensitivity, specificity, and precision were calculated at the SNP-level. All SNPs inside GENCODE gene annotations with at least one read mapped were used. The allele specificity was determined by AlleleHMM, independent binomial tests, and using GENCODE gene annotations were summarized and used to calculate performance using in-house scripts.

Comparison with H3K27me3 ChIP-seq data

To test the correlation between transcription and H3K27me3 in GM12878, we mapped the H3K27me3 ChIP-seq reads to the diploid genome of GM12878 using Bowtie as implemented in AlleleDB (Chen et al., 2016). The Bowtie output, including reads and their mapping positions, was separated into maternal- and paternal-specific files. Coordinates were transferred to the reference genome (hg19) using liftOver. We used bedtools coverage to calculate the number of H3K27me3 ChIP-seq reads falling into each AlleleHMM block obtained from GRO-seq in GM12878. We then calculated the ratio of maternal-specific and paternal-specific H3K27me3 ChIP-seq reads in each GRO-seq AlleleHMM block and summarized using in-house R scripts.

Data used in this study

GRO-seq of 129/castaneus F1 hybrid mouse embryonic stem cells: SRA ID number SRR4041366.

GRO-seq of GM12878: SRA ID number SRR1552485

H3K27me3 ChIP-seq data of GM12878 were fastq files from ENCODE: ENCFF000ASV, ENCFF000ASW, ENCFF000ASZ, ENCFF001EXM, ENCFF001EXO

Software availability

The source code of AlleleHMM is available under the BSD 2-clause license https://github.com/Danko-Lab/AlleleHMM

Scripts used for each computation are available at: https://github.com/Danko-Lab/AlleleHMM/tree/master/analysis_for_AlleleHMM_manuscript

1.4 Results

1.4.1 Finding allele specific differences using a hidden Markov model

Differences in mark abundance between two heterozygous alleles are often correlated across multiple adjacent SNPs (**Figure 1.1A**). We developed AlleleHMM to identify genomic regions that share allele specific differences in functional mark abundance. AlleleHMM takes as input counts of reads mapping unambiguously to each of the two alleles in heterozygous positions of a phased reference genome. AlleleHMM models the data using a hidden Markov model (HMM) that divides the genome among three hidden states: a symmetric state (state S) which shows no allelic difference in mark abundance, and regions with a higher signal on the maternal (M) or paternal (P) allele (**Figure 1.1B; see methods**). AlleleHMM models the distribution of read counts mapping to each allele using a binomial distribution. To control the tradeoff between sensitivity and specificity, we introduced a user-adjustable tuning parameter, τ, that constrains the transition probability out of either the maternal or paternal state. Aside from τ, all other model parameters are set using expectation maximization over the provided data.

1.4.2 Performance test with binomial-distributed simulated data

To determine how AlleleHMM performed in practice, we simulated blocks of contiguous SNPs with specific levels of allele specificity. We simulated a sequence of SNPs composed of three blocks using the binomial distribution (**Figure 1.2A, top**): two blocks with equal signal in both alleles and one middle block that exhibited a known difference in signal between alleles. We evaluated the performance of AlleleHMM after

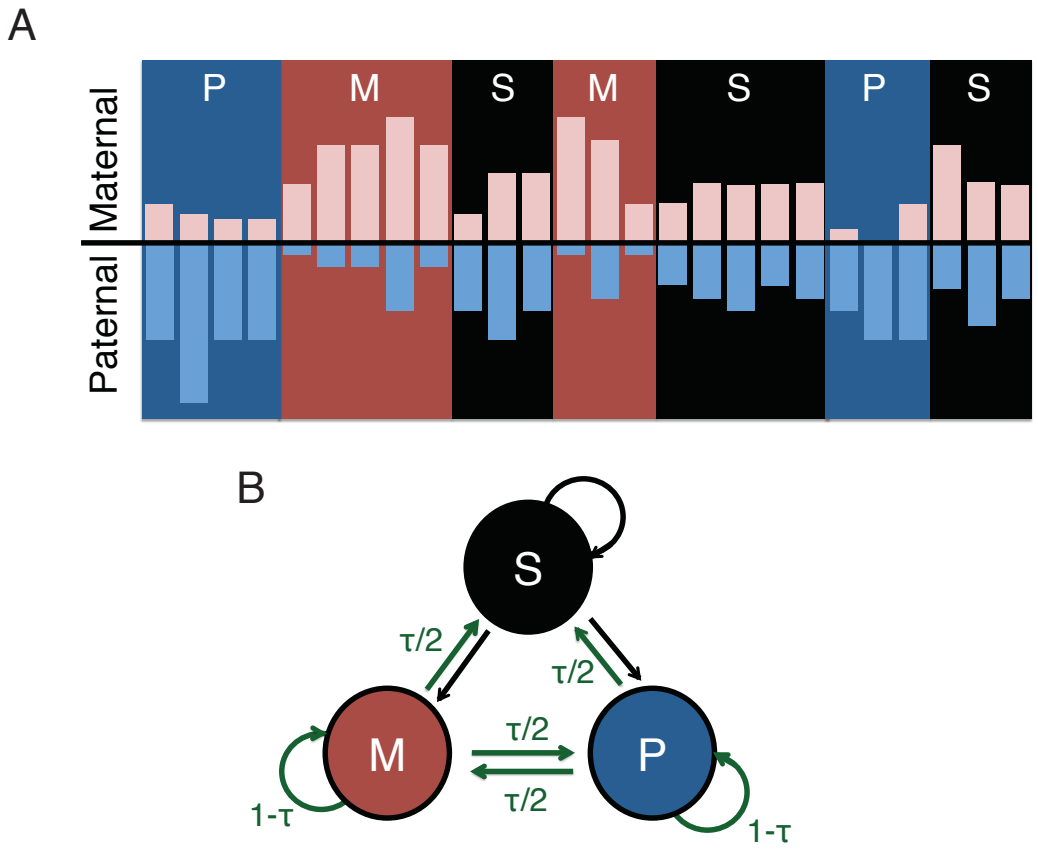


Figure 1.1: AlleleHMM uses a hidden Markov model (HMM) to infer the allele specificity of genomic markers at each SNPs.

(A) Cartoon shows the frequency of reads mapping to the patneral (light blue bars) and maternal (pink bars) allele at positions across the genome (X-axis). Nearby SNPs show similar signatures of allele specificity depicted as blue (P, paternal allele specificity), red (M, maternal allele specificity), or black (S, no evidence of allele specificity) background identified using AlleleHMM.

(B) The model structure of AlleleHMM. We model allele specificity using three hidden states: a symmetric state which shows no allele specificity (S, black), and states representing maternal- (M, red) or paternal-specific (P, blue) regions. SNPs can transition between hidden states. Green arrows represent the transition probabilities set using a user-adjustable tuning parameter, τ.

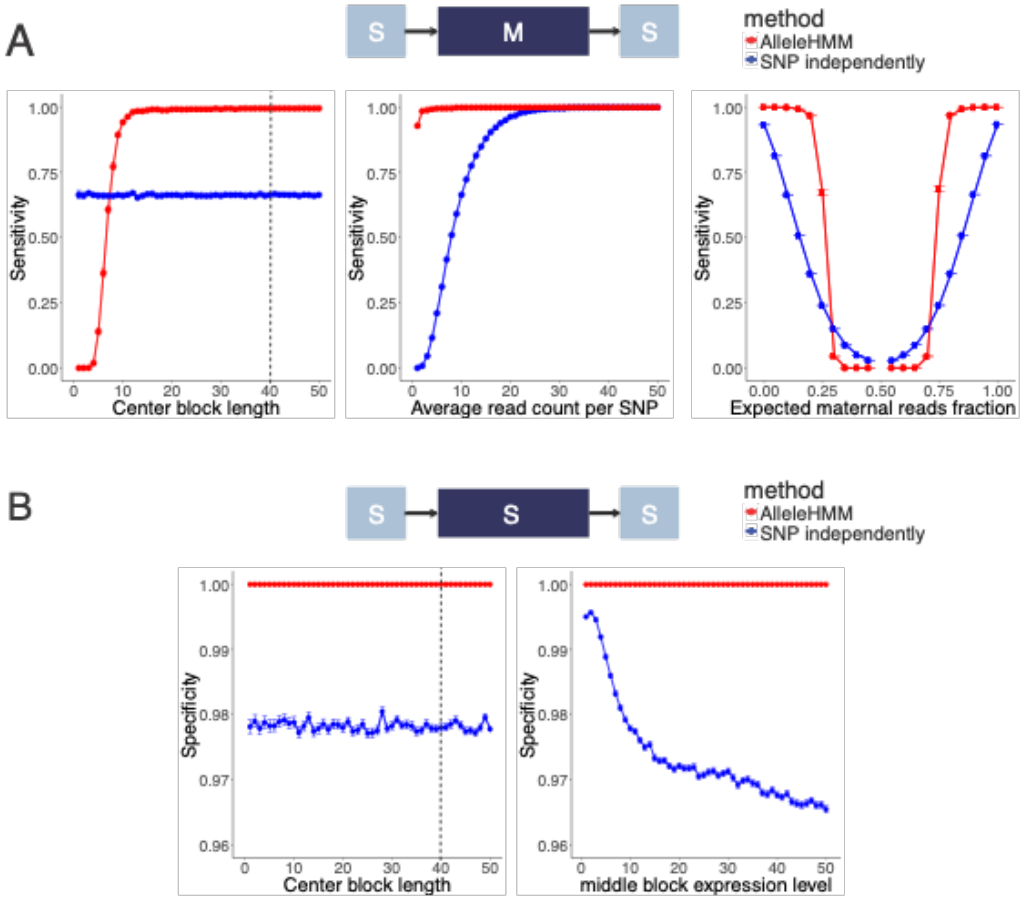


Figure 1.2: AlleleHMM had better sensitivity and specificity compared with standard methods performing independent binomial tests for each SNP.

(A) Scatterplots show the sensitivity for each SNP in the center block of AlleleHMM (red) and independent binomial tests (blue) as a function of the length of a maternal specific block (the number of continuous SNPs sharing same allele specificity, left), the average read count at each SNP (center), or the expected maternal reads fraction (right). Error bars represent the standard error of 1000 independent simulations. The dotted line indicates the average number of SNPs per human gene.

(B) Scatterplots show the specificity of AlleleHMM (red) and independent binomial tests (blue) as a function of the length of the symmetric middle block (the number of continuous SNPs sharing same allele specificity, left) or the average read count at each SNP (right). Error bars represent the standard error of 1000 independent simulations. The dotted line indicates the average number of SNPs per human gene.

we systematically changed the length, signal level, and degree of allele specificity in the middle block holding other parameters constant (see methods) (**Figure 1.2A**). We evaluated the accuracy of AlleleHMM by examining the sensitivity (the fraction of true positives recovered). We also examined the specificity (the fraction of correctly classified true negatives) and precision (the fraction of true positives over all positive calls) of AlleleHMM, incorporating simulations in which the middle block was symmetric to compute precision. We chose simulation parameters characteristic of global run-on and sequencing (GRO-seq) (Kwak et al., 2013), a direct measurement of RNA polymerase which is challenging to use in allele specific measurements because only a few reads map to each SNP in a typical dataset, resulting in poor statistical power.

AlleleHMM identified allelic differences in simulated data with higher sensitivity, specificity, and precision compared with simple methods that perform independent binomial tests at each SNP (**Figure 1.2; Supplementary Figure 1.2A**). The sensitivity of AlleleHMM for each simulated allele specific SNP in the center block increased with block length. AlleleHMM had a higher sensitivity than independent binomial tests when the center block contained as few as 8 adjacent SNPs, and a higher precision with only 10 adjacent SNPs (**Figure 1.2A, Supplementary Figure 1.2A, left**), shorter than observed in most mammalian genes (on average, 39.7 SNPs per gene for human CEPH Utah and 237.2 SNPs for 129/Castaneus F1 hybrid mouse, **Supplementary Figure 1.2B**). AlleleHMM had a higher sensitivity across the spectrum of signal levels (**Figure 1.2A, center**). Likewise, we found that AlleleHMM was more sensitive across an important range of allele specificity magnitudes (<= 0.25 or >= 0.75; **Figure 1.2A, right**). Treating SNPs independently resulted in a higher sensitivity when the allele

specificity was much lower in magnitude (0.25-0.75), which we attribute to the presence of rare individual SNPs that have a higher magnitude of allele specificity than the average for that block due to random statistical fluctuations. AlleleHMM had a higher specificity throughout the range of expression and block length parameters than treating SNPs independently (**Figure 1.2B**), and generally had superior precision as well (**Supplementary Figure 1.2A**), demonstrating that AlleleHMM does not trade a higher sensitivity for a lower specificity. Thus, we conclude that AlleleHMM had better sensitivity and specificity for allele specific transcription in synthetic data simulated using the binomial distribution.

1.4.3 Performance test with overdispersed synthetic data

Many short-read datasets exhibit overdispersion due to a variety of technical factors, which increases the rate of false positive allele specific differences (Chen et al., 2016). To test how AlleleHMM performed with overdispersed data, we applied a similar simulation strategy using a beta-binomial distribution to simulate read counts with varying degrees of overdispersion. AlleleHMM had a reasonably high sensitivity, precision, and specificity across the spectrum of distinct overdispersion values (**Figure 1.3A; Supplementary Figure 1.3A**). AlleleHMM retained a sensitivity >0.95, while maintaining both precision and specificity near 1.0 at realistic overdispersion levels estimated using two independent GRO-seq datasets: human GM12878 lymphoblastoid cells (overdispersion of 0.24) (Core et al., 2014) and 129/Castaneus F1 hybrid mouse embryonic stem cells (mESCs) (overdispersion of 0.26) (Engreitz et al., 2016).

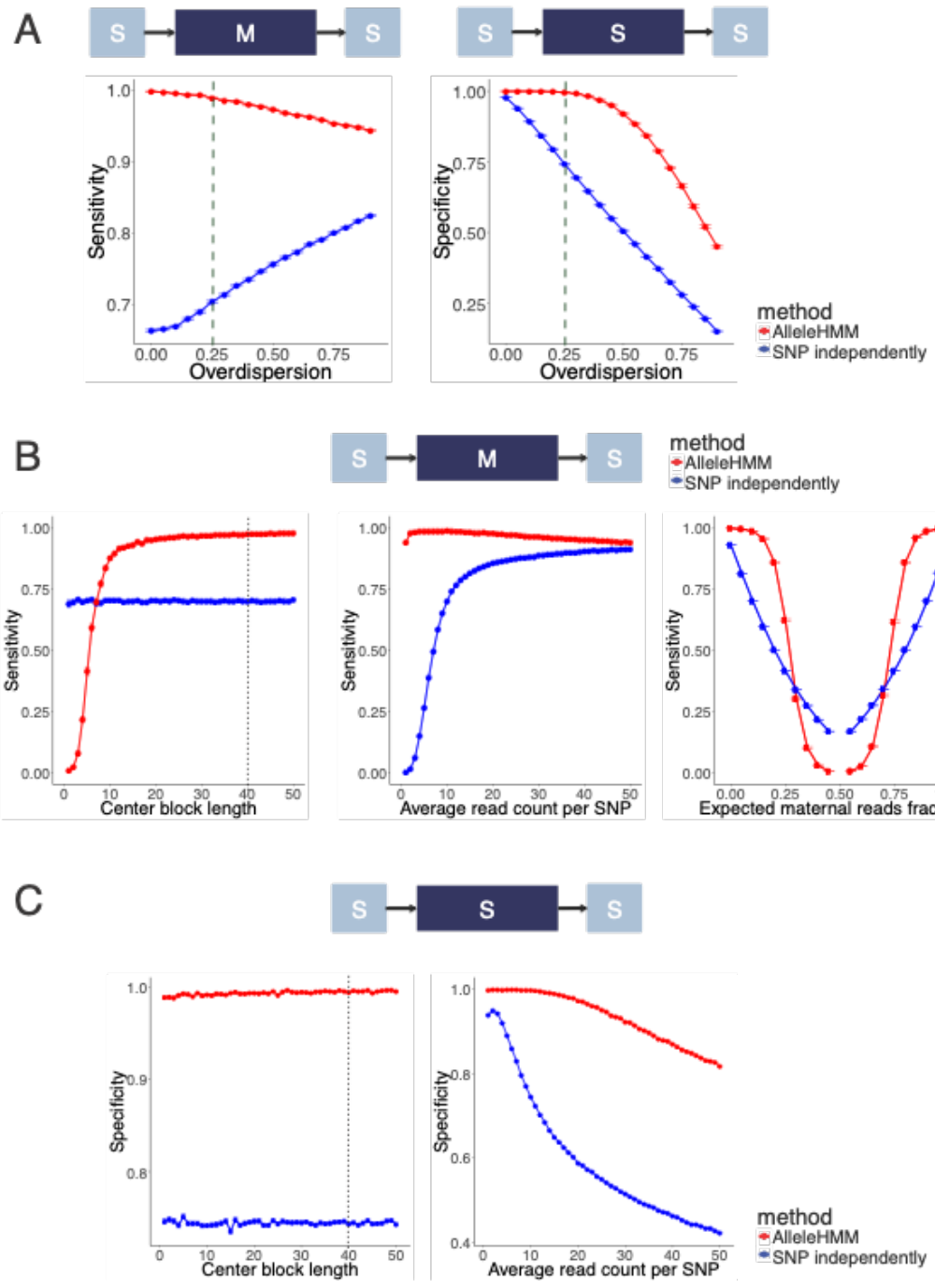


Figure 1.3: AlleleHMM had better sensitivity and specificity compared with the naive standard practice of performing a binomial test for each SNP independently in overdispersed data.

(A) Scatterplots show sensitivity (left) and specificity (right) of AlleleHMM (red) and independent binomial tests (blue) as a function of the overdispersion parameter in beta-binomial distributed simulated data. Error bars represent the

standard error of 1000 independent simulations. Dashed lines indicate the mean of overdispersion estimated from GRO-seq of GM12878 and GRO-seq of 129/Castaneus F1 hybrid mESCs.

(B) Scatterplots show the sensitivity of AlleleHMM (red) and independent binomial tests (blue) as a function of the length of a maternal specific blocks (the number of continuous SNPs sharing same allele specificity, left), the average read count at each SNP (center), or the expected maternal read fraction (right) with an overdispersion of 0.25. Error bars represent the standard error of 1000 independent simulations. The dotted line indicates the average number of SNPs per human gene.

(C) Scatterplots show the specificity of AlleleHMM (red) and independent binomial tests (blue) as a function of the length of the symmetric middle block (left) or the average read count at each SNP (right) with an overdispersion of 0.25. Error bars represent the standard error of 1000 independent simulations. The dotted line indicates the average number of SNPs per human gene.

To test how AlleleHMM performance varied with the length, signal level, and the degree of allele specificity when the input data was overdispersed, we fixed overdispersion to 0.25 (dashed lines in **Figure 1.3A and Supplementary Figure 1.3A**) and performed simulation experiments similar to those described for the binomial distribution, above. AlleleHMM sensitivity and precision increased with block length, and were higher than independent binomial tests with as few as 8 (or 5, respectively) adjacent SNPs (**Figure 1.3B, Supplementary Figure 1.3B, left**). AlleleHMM was also highly sensitive over an important range of allele specificity magnitudes (<= 0.25 or >= 0.75; **Figure 1.3B, right**). AlleleHMM had a higher specificity than independent binomial tests across the spectrum of length (the number of SNPs per gene, **Figure 1.3C, left**) and signal levels (average read counts per SNP, **Figure 1.3C, right**). The specificity of both AlleleHMM and independent binomial tests declined as read count increased (**Figure 1.3C, right**). However, AlleleHMM maintained a reasonable precision throughout this range, while independent binomial tests never achieved a precision >0.75 in this test (**Supplementary Figure 1.3, center**). Moreover, AlleleHMM exhibited a high sensitivity within the range at which it maintained a high specificity (2-20 reads supporting each SNP, **Figure 1.3B, center**), suggesting that subsampling highly expressed regions may be a viable strategy to deal with overdispersion in practice. Thus, AlleleHMM improved specificity in overdispersed data while maintaining a higher sensitivity compared with state-of-the-art tools using realistic parameters taken from GRO-seq data.

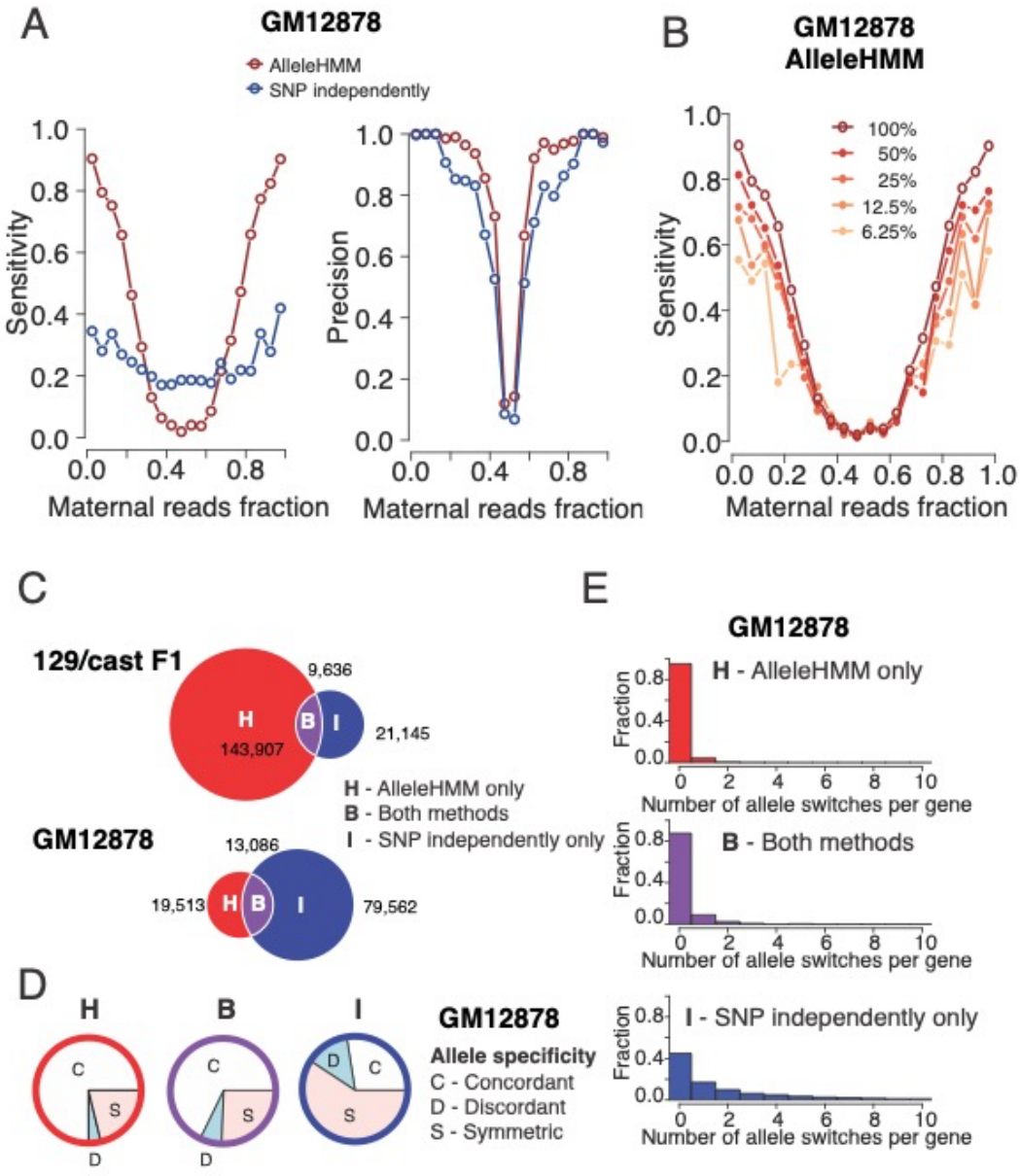


Figure 1.4: Comparison between AlleleHMM and independent binomial tests.

(A) Scatterplots show the sensitivity (left) and precision (right) of AlleleHMM (red) and independent binomial tests (blue) as a function of the maternal reads fraction in the gene annotation using GRO-seq data from GM12878.

(B) Scatterplots show the sensitivity of AlleleHMM as a function of the fraction of maternal reads in the gene annotations. Different lines indicate the read depth of the subsampled GRO-seq reads from a deeply sequenced human GM12878 dataset. The total sequencing depth is 187,896,441.

(C) Venn diagrams show the number of allele specific SNPs identified by AlleleHMM only (H, red), independent binomial tests (I, blue, implemented in AlleleDB), and intersection of both methods (B, purple) from GRO-seq data of a 129/Castaneus F1 hybrid mouse (top) and a human cell line GM12878 (bottom).

(D) Pie charts show the proportion of allele specific SNPs in GM12878 GRO-seq data that are within genes having no evidence of allele specificity over the gene (symmetric, S, pink), or genes that show higher expression on the same (concordant, C, white) or the opposite (discordant, D, light blue) haplotype.

(E) Histograms show the fraction of genes as a function of the number of allele specificity switches the gene contains. Allele specificity was determined by AlleleHMM only (H, red, top), independent binomial tests (I, blue, bottom), and intersection of both methods (B, purple, middle) using GRO-seq data from a human cell line GM12878.

1.4.4 Performance comparison with independent SNPs using GRO-seq data

We next asked whether AlleleHMM provides a higher sensitivity, specificity, or precision using real GRO-seq data as input. We used AlleleHMM to analyze two public GRO-seq datasets: one from 129/Castaneus F1 hybrid mESCs and the other from a human GM12878 lymphoblastoid cell line (Core et al., 2014; Engreitz et al., 2016).

To estimate the accuracy of AlleleHMM using real data in which there was no ground truth, we identified annotated genes with evidence for allele specific transcription by combining reads across the entire gene annotation. Because RNA polymerase in the gene body is loaded in the promoter region, SNPs residing in the same gene annotation should generally share the same level of allele specificity. Therefore, we estimated the allele specificity of each gene annotation using all reads that fall inside, and assigned this magnitude of allele specificity to every SNP that resides inside that annotation (see Methods). Using gene annotations as a surrogate for a ground truth, we found that AlleleHMM had a higher sensitivity than independent binomial tests, especially within the range of allele specificity magnitudes that was most likely to contain biologically relevant allelic differences (maternal read ratio of the gene <= 0.25 or >= 0.75, **Figure 1.4A, Supplementary Figure 1.4, left**), consistent with its performance in synthetic data. AlleleHMM had a similar or higher precision compared with independent binomial tests (**Figure 1.4A, Supplementary Figure 1.4, right**). Thus, AlleleHMM is both more sensitive and precise when using real GRO-seq data as input.

Most genomics applications are highly imbalanced, with negative examples greatly outnumbering positive examples, resulting in poor precision despite a high

specificity. To measure performance in cases where data are highly imbalanced, we generated datasets based on SNPs in genes where the magnitude of allele specificity was <0.2 or >0.8 (true positives) or between >0.45 and <0.55 (true negatives). Using these highly imbalanced datasets (ratio of positive to negative examples < 0.05), AlleleHMM retained a high precision (>0.66) at relatively high sensitivity (>0.76). AlleleHMM outperformed independent binomial tests by a wide margin in this task (sensitivity < 0.4 and precision < 0.2; see **Supplementary Figure 1.5**). We note that a precision of 1.0 is not necessarily expected or desired in these tests, because numerous differences exist between gene annotations and the actual patterns of transcription that occur in cells (as described below). These results suggest that even in a setting where less than 5% of SNPs are true positives, AlleleHMM still obtains useful information about allele specificity, whereas the use of independent tests does not.

To examine how library sequencing depth affects the performance of AlleleHMM, we subsampled GRO-seq reads from the deeply sequenced human GM12878 dataset and evaluated recovery using annotations with true positive/ negative labels generated based on the entire dataset. We found that AlleleHMM was remarkably sensitive even at sequencing depths that were <10% of the total read count (**Figure 1.4B**), at a consistent precision and specificity (**Supplementary Figure 1.6A, center and right**). In contrast, the sensitivity of independent binomial tests, which was not very high to begin with, dropped off rapidly with sequencing depth (**Supplementary Figure 1.6B, left**). Thus, even a modest sequencing depth (~20 million uniquely mapped reads) is enough to identify the majority of allele specific differences in transcription using GRO-seq data.

To more rigorously understand the differences between AlleleHMM and independent binomial tests, we divided SNPs based on whether they were identified as allele specific using AlleleHMM, independent binomial tests, or both methods. Few SNPs were identified as allele specific using both AlleleHMM and independent binomial tests on each SNP (**Figure 1.4C**). In GM12878, for example, AlleleHMM identified 32,599 heterozygous SNPs with 1 or more read in 4,026 AlleleHMM blocks. Only 13,086 of the SNPs identified using AlleleHMM were also discovered using independent binomial tests (~40% of SNPs; **Figure 1.4C, bottom**). As expected, SNPs identified only by AlleleHMM largely reflect heterozygous positions covered by too few reads to confidently assign allele specificity when treating SNPs independently, whereas those identified using both methods had a higher read depth (**Supplementary Figure 1.7**). Taken together, these observations are consistent with AlleleHMM making substantial improvements in sensitivity for allele specific differences in genes with lower expression levels.

We were more surprised to find large numbers of SNPs reported as allele specific using independent binomial tests without a corresponding discovery by AlleleHMM (n= 21,145 [mESC] or 79,562 [GM12878]). To investigate whether these SNPs were false negative calls by AlleleHMM or false positives by the binomial test, we again used SNPs within annotated genes under the assumption that RNA polymerase across a gene shares the same allele specificity. The majority of allele specific SNPs identified by AlleleHMM were found to have the same direction (maternal or paternal) of allele specificity as the gene annotation, henceforth called "concordant" (**Figure 1.4D, concordant [C] in white**). By contrast, SNPs identified as allele specific using only

independent binomial tests were most often identified within genes where the entire annotation showed no evidence of allele specificity, henceforth called "symmetric" (**Figure 1.4D, symmetric [S] in pink**). We also directly investigated whether gene annotations tend to have a single allelic state using each method, as implied by the assumption that RNA polymerase density is largely determined by events occuring at a single promoter region. We found that AlleleHMM identified a single block covering annotations in >80% of cases (**Figure 1.4E, top**). In contrast, the direction of allele specificity detected by independent binomial tests often switched across the annotation (**Figure 1.4E, bottom**), resulting in no evidence of allele specificity when SNPs within annotations were merged. Taken together, these observations provide additional support to our analysis of precision in unbalanced genomic data (**Supplementary Figure 1.5**) and suggest that many of the SNPs identified only by independent binomial tests are false positives.

1.4.5 Widespread non-coding allele specific transcription identified using AlleleHMM

Despite achieving a high concordance with gene annotations when available, AlleleHMM was also able to identify allele specific differences in unannotated transcription units. For instance, AlleleHMM identified the transcription unit upstream and antisense to *Pdpn* as sharing the same allele specificity as the *Pdpn* coding region (**Figure 1.5A**). Likewise, in cases where AlleleHMM disagreed with gene annotations, it frequently identified cell-type specific transcription units that were correctly classified upon careful examination. For instance, in the GM12878 dataset AlleleHMM found that

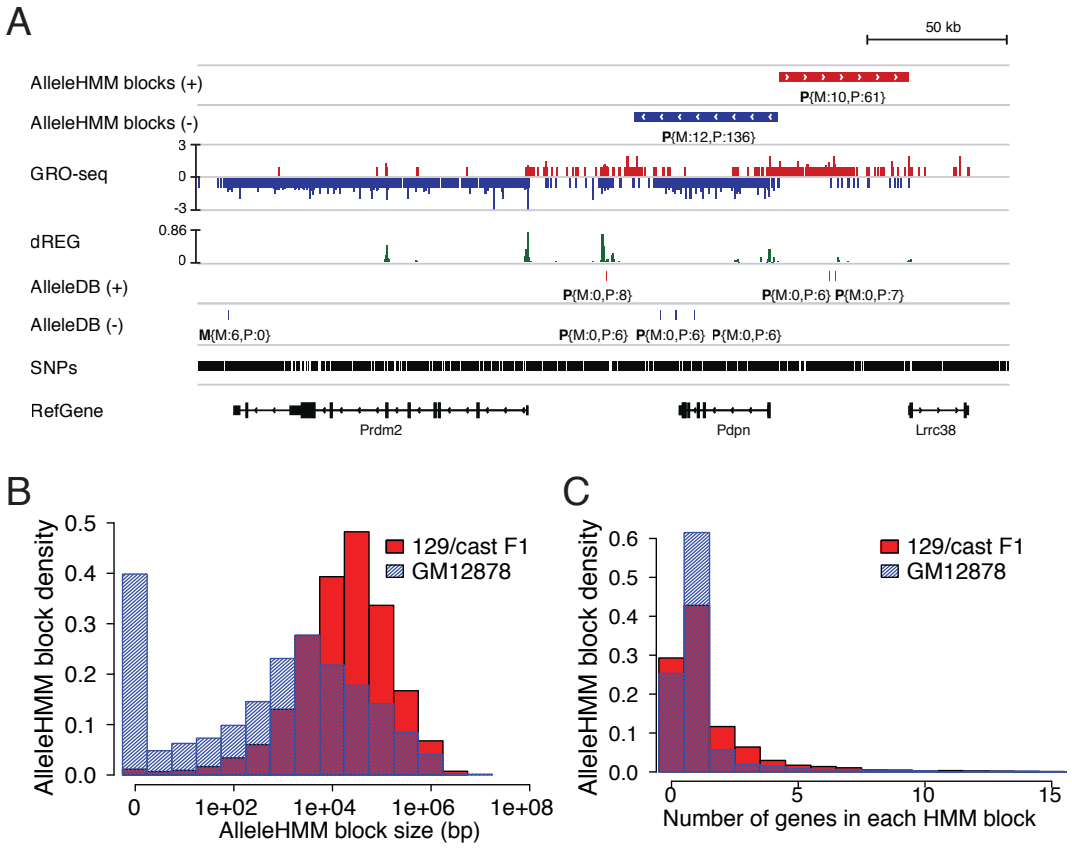


Figure 1.5: Application of AlleleHMM to GRO-seq data.

(A) Genome browser view shows the application of AlleleHMM and independent binomial tests (implemented in AlleleDB) to GRO-seq data from a 129/Castaneus F1 hybrid mouse. The allele specific read counts of the blocks and SNPs are denoted as **P**{M:12,P:136}, meaning that the block is paternal specific (**P**) with 12 maternal-specific (M) reads and 136 paternal-specific (P) reads.

(B) Histograms show the distribution of AlleleHMM block size of GRO-seq data from a 129/Castaneus F1 hybrid mouse (red) and GRO-seq data from a human cell line GM12878 (blue) in log scale (X-axis).

(C) Histograms show the fraction of AlleleHMM blocks as a function of the number of genes it contains. GRO-seq data from a 129/Castaneus F1 hybrid mouse is in red and GRO-seq data from a human cell line GM12878 is in blue.

transcription originating from enhancers within the *DTNB* gene annotation was maternal-specific, although most of the *DTNB* annotation itself was not (**Supplementary Figure 1.8**). These observations demonstrate that AlleleHMM provides substantial advantages useful for new biological discovery compared with heuristics that summarize signals within well-known gene annotations.

AlleleHMM revealed thousands of regions with maternal or paternal-specific RNA polymerase abundance genome-wide. AlleleHMM identified 3,483 and 4,026 blocks with significant allele specific differences in mESCs and GM12878, respectively. The average genome size of each block in the F1 hybrid dataset was ~166 kb (**Figure 1.5B**). Blocks were larger in the mESC dataset than in GM12878, likely owing to a combination of differences in heterozygosity and sequencing depth between datasets (**Supplementary Figure 1.9**). Approximately 25% of AlleleHMM blocks did not contain any GENCODE gene annotation (**Figure 1.5C**), for example the antisense transcription unit upstream of *Pdpn* (**Figure 1.5A**). Many AlleleHMM blocks contained more than one gene annotation (28% of F1 hybrid mouse blocks and 13% of GM12878 blocks), indicating groups of nearby genes and non-coding transcription units that share similar allele specificity in their transcription. Thus, AlleleHMM identified widespread evidence for allele specificity in non-coding transcription, as well as coordination between nearby transcription units that was not evident from strategies that merged gene annotations.

1.4.6 Allele specific transcription negatively correlates with allele specific repressive chromatin marks

To find further independent validation for blocks of allele specific transcription identified using AlleleHMM, we asked whether we could recover the negative relationship expected between transcription and histone marks associated with transcriptional repression, especially H3K27me3. We focused on GM12878, for which there is publicly available ChIP-seq data profiling the distribution of H3K27me3 (ENCODE Project Consortium, 2012; Sloan et al., 2016). Mapping H3K27me3 ChIP-seq data onto AlleleHMM blocks identified using GRO-seq revealed 113 blocks with a significant allele specificity in H3K27me3 ChIP-seq data. As expected, the degree of allele specificity in H3K27me3 ChIP-seq was inversely correlated with that of GRO-seq (Pearson's R =-0.57; **Figure 1.6**). The slope of the best fit line implies that a 2-fold change in H3K27me3 was associated with ~5.7-fold change in transcription. Assuming a similar dynamic range in both assays, this result implies that relatively modest changes in H3K27me3 may have a relatively large average impact on transcription. Thus, AlleleHMM reveals blocks of allele specificity which are largely in agreement with orthogonal genomic assays.

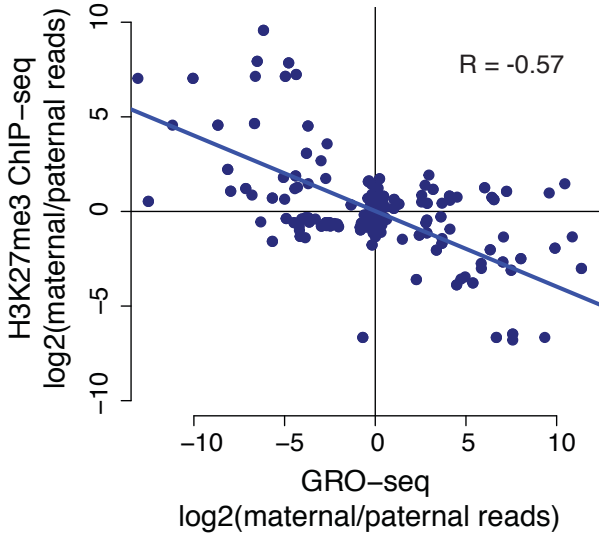


Figure 1.6: Allele specific transcription correlates with allele specific H3K27me3. Scatterplots show the allele specificity of H3K27me3 ChIP-seq data (Y-axis) as a function of allele specificity in GRO-seq data (X-axis). The trendline shows the best fit based on a total least squares regression. The Pearson correlation is shown on the plot (R = -0.57, p-value < 2.2e-16). The slope of the best fit line is -2.5.

1.5 Discussion

Here we have introduced AlleleHMM to identify allele specific differences in functional genomic marks. AlleleHMM models the spatial correlation in allele specific differences in a phased diploid genome using hidden Markov models (HMMs). We show that AlleleHMM provides substantial improvements in both sensitivity and specificity for detecting allele specific SNPs compared with existing computational tools, using both simulation studies and analyses of real GRO-seq data. AlleleHMM is applicable to any type of functional genomic data and to any diploid species with a high-quality phased reference genome. AlleleHMM can now be deployed to understand the interplay between chromatin environment, transcription, and mRNA across a wide variety of organisms, providing new insights into how DNA sequences influence biochemical processes in the nucleus.

Although there has been extensive interest in using allele specific information to understand the interplay between functional marks, surprisingly few computational methods have been developed for this task. One of the current methods to identify allelic differences requires performing independent statistical tests at each SNP (Chen et al., 2016; Rozowsky et al., 2011). This approach of treating SNPs independently requires a high sequencing depth with dozens of reads covering each SNP in order to be statistically powered to identify allele specific differences. Additionally, because signal for the majority of marks tends to be unevenly distributed along the genome, this strategy is prone to significant biases where loci with a higher signal intensity tend to be better represented due to statistical power. Finally, as has been reported elsewhere (Chen et al., 2016), and as we show here, the application of independent statistical tests

is prone to a high rate of false discoveries. Although false positives can be addressed by using more conservative statistical models (Chen et al., 2016), this more conservative strategy exacerbates issues with statistical power. AlleleHMM addresses this deficiency in an alternative way, by modelling the correlation between adjacent signal intensities, thus pooling statistical power across adjacent positions.

An alternative approach that is commonly used to identify allele specific differences in mark abundance is to use pre-established boundaries of genes or other genomic features as a way to pool SNPs within regions of the genome (Crowley et al., 2015). This alternative strategy improves upon the use of independent statistical tests by using information between nearby alleles. However, there are still a number of important limitations with this approach. Chiefly among the limitations of this strategy is that allele specific differences in functional marks cannot be identified if they fall outside of the boundaries of pre-established gene annotations. Likewise, cell-type specific differences in transcript isoforms are common, even in well annotated genomes like human and mouse, which provide a substantial source of error for annotation-based approaches. Finally, the use of annotations requires a well annotated reference genome, which is only available in well studied model organism such as Drosophila, humans, or mice. AlleleHMM addresses these limitations by providing a rigorous and statistically motivated method to identify the boundaries of allele specific blocks *de novo*.

Although AlleleHMM is a powerful tool that makes significant improvements compared with existing strategies, it does have several limitations. Chiefly among these, AlleleHMM will provide the most significant benefit for functional assays where marks are spread broadly across the genome, rather than focused within specific functional

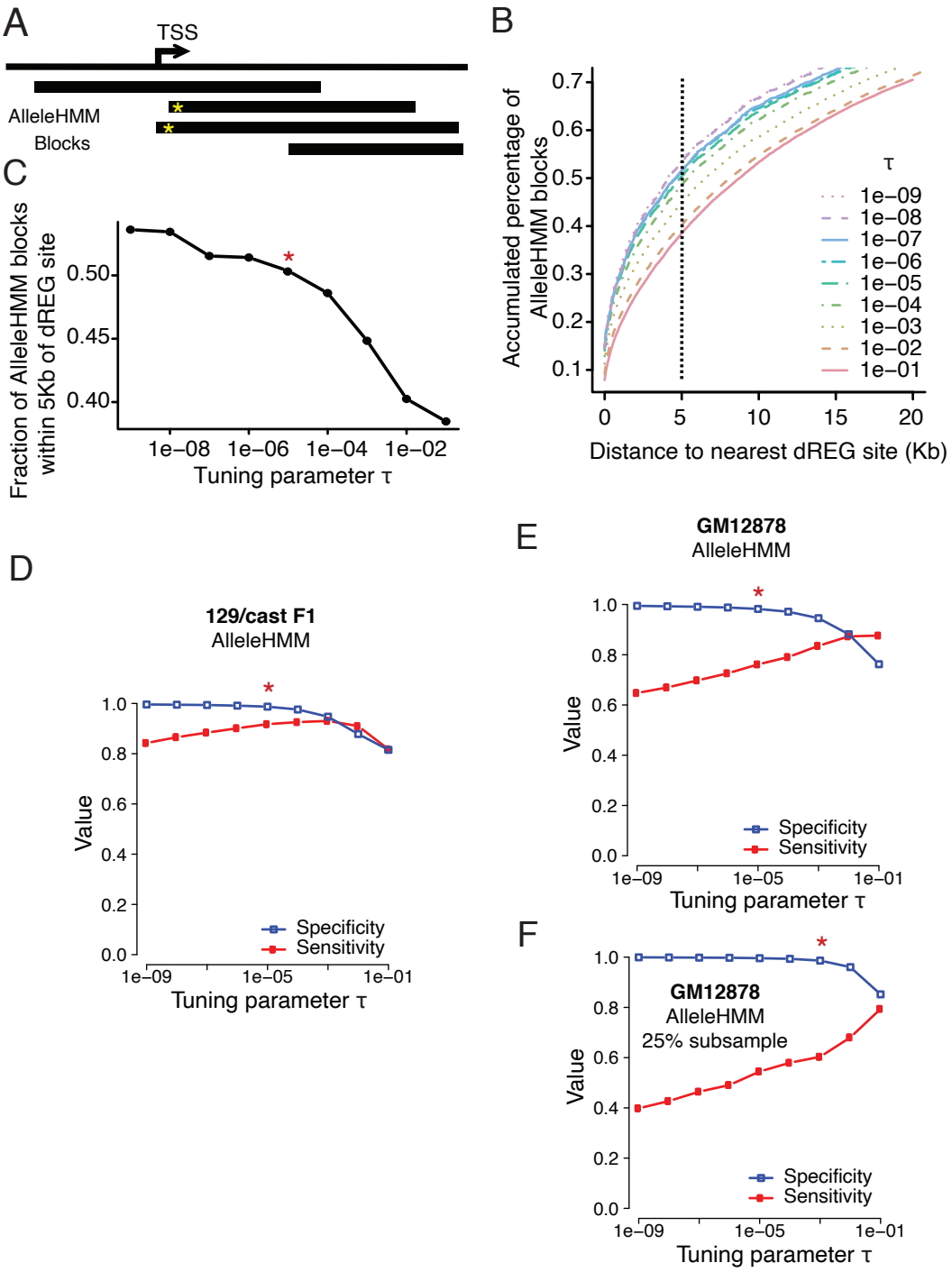
regions. This provides the broadest benefit for assays such as GRO-seq, RNA-seq, and several of the broadly distributed ChIP-seq marks. AlleleHMM may provide less benefit for assays such as 3' mRNA-seq or chromatin accessibility assays (e.g., ATAC-seq or DNase-I-seq), where signals are distributed within a specific position of the genome. Nevertheless, AlleleHMM will still work under these settings, and may still provide a substantial benefit for detecting expression differences that span multiple genes or chromatin accessible regions.

Using AlleleHMM, we have identified thousands of regions harboring allele specific differences in human GM12878 and murine ESCs. We have found that allelic differences tend to occur over large genomic regions that harbor multiple transcription units, often sharing the same gene annotation. This finding is reminiscent of the shared architecture of quantitative trait loci (QTLs) across broad genomic regions (Waszak et al., 2015). This finding may also reflect similar regulatory principles as the positionally dependent variation in gene expression across distinct biological replicates (Rennie et al., 2018). Altogether, the use of AlleleHMM provides a novel tool that will be useful to rigorously examine how homologous DNA sequences in the nucleus differ in the distribution of functional genomic marks. We are confident that future studies will use this tool to unravel multiple aspects of genome function and organization.

1.6 Acknowledgments

We thank members of the Danko laboratory for valuable discussions. Work in this publication was supported by an NHGRI (National Human Genome Research Institute) grant R01-HG009309 and a Cornell Research Grants program in Animal Health to CGD. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

1.7 Supplementary Figures and Tables



Supplementary Figure 1.1: Selection of values for the tuning parameter used to control the balance between AlleleHMM sensitivity and specificity.

(A) Cartoon illustrates how we set the tuning parameter τ. We assumed that SNPs within the same transcript have a similar allele specificity. Therefore, the optimum value of τ should approach a saturation point to maximize the fraction of state transitions near a transcription start site (TSSs) that is active in that cell type. The black bars are AlleleHMM blocks, represents a region with significant allele specificity. Those with a yellow star have state transitions near TSSs.

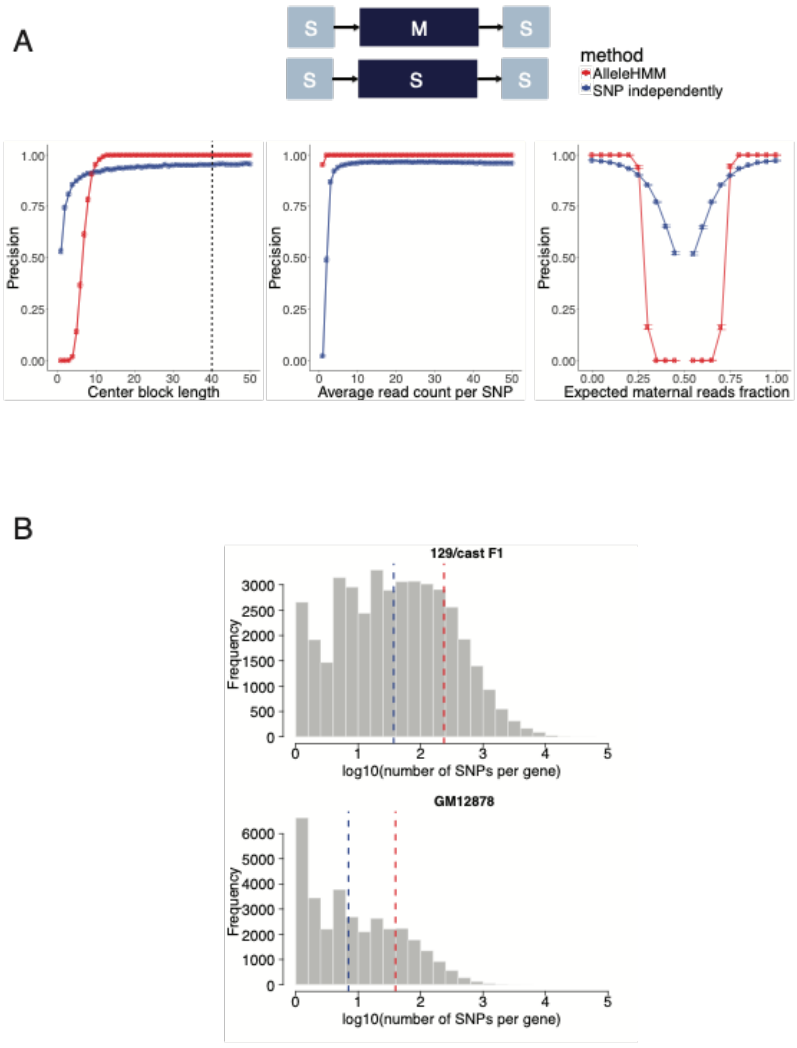
(B) Plot shows the distance between the beginning of AlleleHMM blocks and its closest TSS identified using dREG for GRO-seq data from a 129/Castaneus F1 hybrid mouse. Different lines indicate AlleleHMM blocks predicted using different values of the tuning parameter, τ.

(C) Scatterplot shows the fraction of AlleleHMM blocks within 5 kb of the nearest TSS predicted by dREG (Y-axis) as a function of the tuning parameter τ (X-axis) used for GRO-seq data from a 129/Castaneus F1 hybrid mouse. The red star indicates a value near a point of saturation (τ = 1e-5) used for the remainder of this study.

(D) Scatterplots show the sensitivity (red) and specificity (blue) of AlleleHMM as a function of the tuning parameter τ (X-axis) used for GRO-seq data from a 129/Castaneus F1 hybrid mouse. The red star indicates the value of τ (1e-5) used for the remainder of this study.

(E) Scatterplots show the sensitivity (red) and specificity (blue) of AlleleHMM as a function of the tuning parameter τ (X-axis) used for GRO-seq data from GM12878. The red star indicates the value of τ (1e-5) used for the remainder of this study.

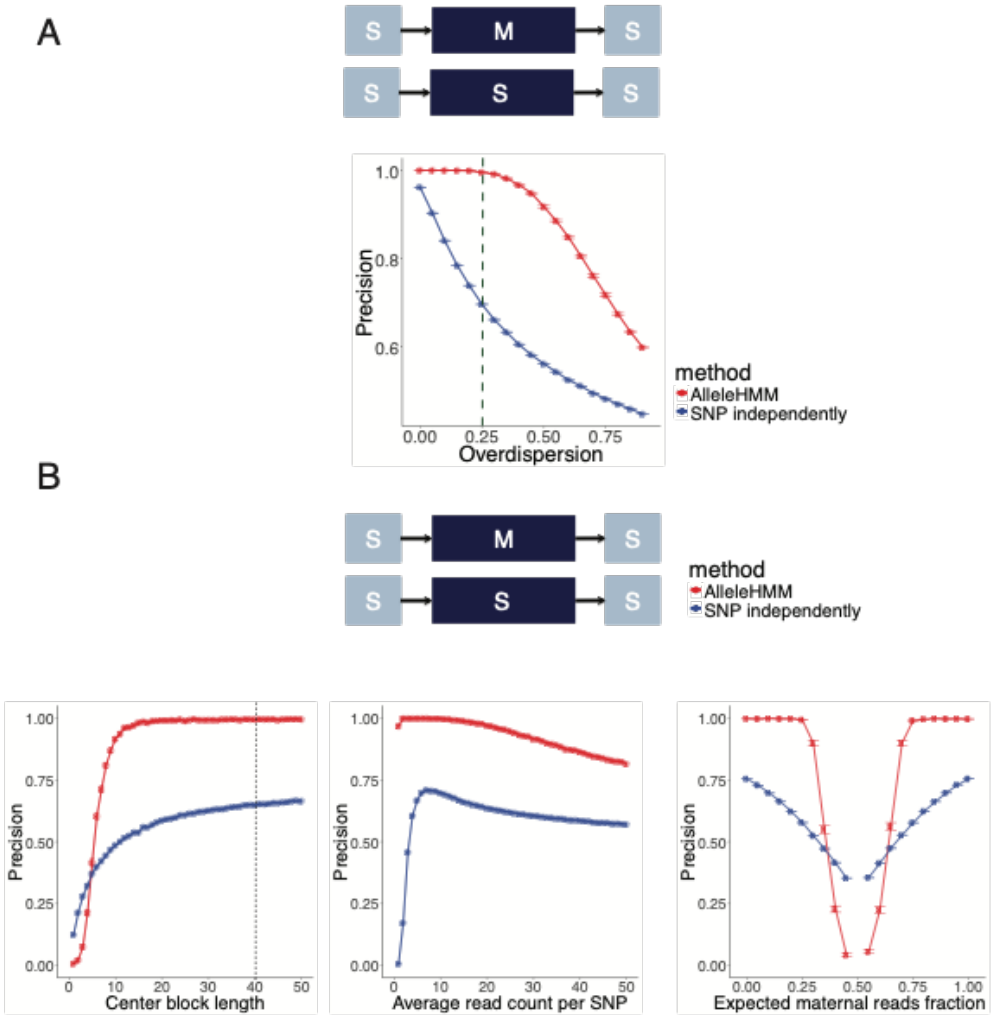
(F) Scatterplots show the sensitivity (red) and specificity (blue) of AlleleHMM as a function of the tuning parameter τ (X-axis) used for a 25% subsampled GRO-seq data from GM12878. The red star indicates the value of τ (1e-3) used for the remainder of this study.



Supplementary Figure 1.2: AlleleHMM had better precision compared with standard methods performing independent binomial tests for each SNP.

(A) Scatterplots show the precision for each SNP in the center block of AlleleHMM (red) and independent binomial tests (blue) as a function of the length of the center block (the number of continuous SNPs sharing same allele specificity, left), the average read count at each SNP (center), or the expected maternal reads fraction (right). Error bars represent the standard error of 1000 independent simulations. The dotted line indicates the average number of SNPs per human gene.

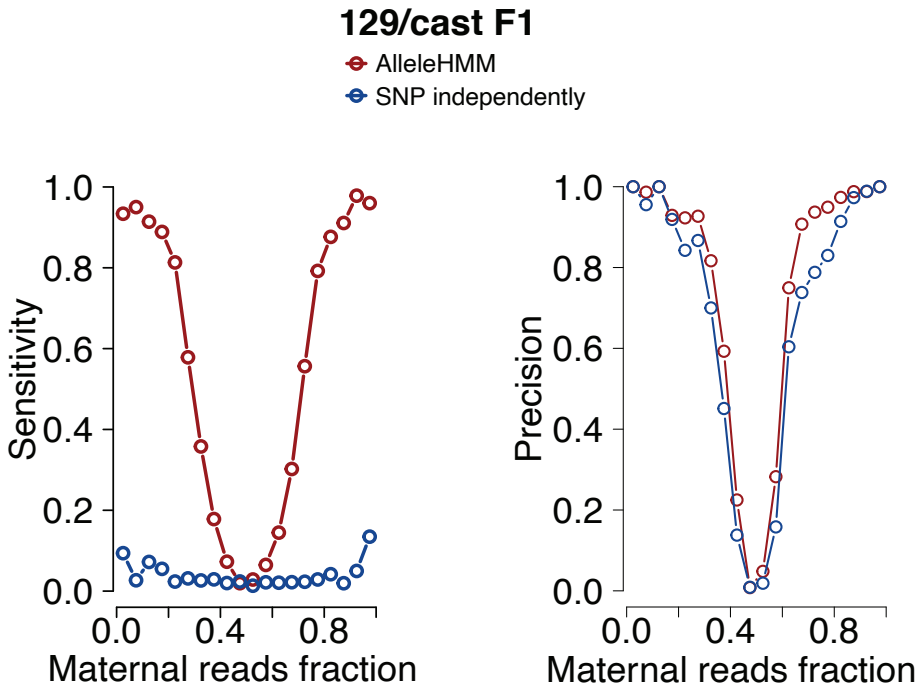
(B) Histograms show the distribution of the number of SNPs per gene in a 129/Castaneus F1 hybrid mouse (top) and a human cell line GM12878 (bottom) in log scale (X-axis). Blue dashed lines indicate the median of the number of SNPs per gene (129/*cast* is 37.0, GM12878 is 7.0) and red dashed lines indicate the mean (129/*cast* is 237.2, GM12878 is is 39.7).



Supplementary Figure 1.3: AlleleHMM had better precision compared with standard methods performing independent binomial tests for each SNP in overdispersed data.

(A) Scatterplots show precision for each SNP in the center block of AlleleHMM (red) and independent binomial tests (blue) as a function of the overdispersion parameter in beta-binomial distributed simulated data. Error bars represent the standard error of 1000 independent simulations. Dashed lines indicate the mean of overdispersion estimated from GRO-seq of GM12878 and GRO-seq of 129/Castaneus F1 hybrid mESCs.

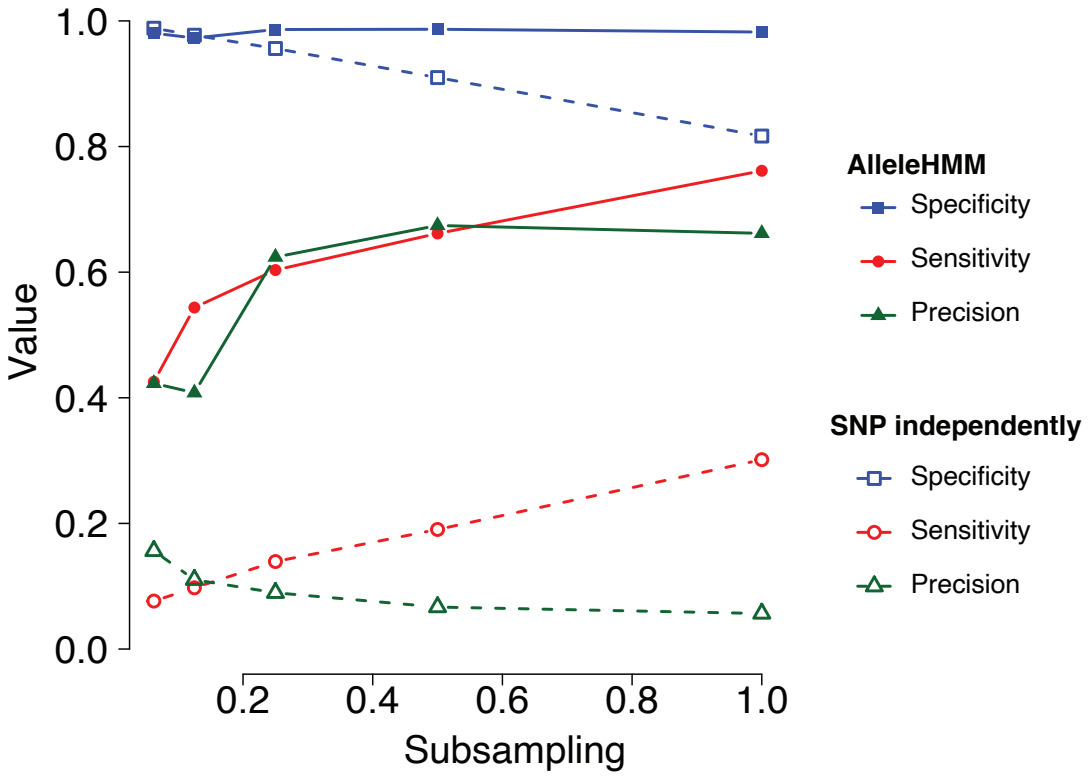
(B) Scatterplots show the precision for each SNP in the center block of AlleleHMM (red) and independent binomial tests (blue) as a function of the length of the center block (the number of continuous SNPs sharing same allele specificity, left), the average read count at each SNP (center), or the expected maternal reads fraction (right) with an overdispersion of 0.25. Error bars represent the standard error of 1000 independent simulations. The dotted line indicates an estimate number of SNPs per human gene.



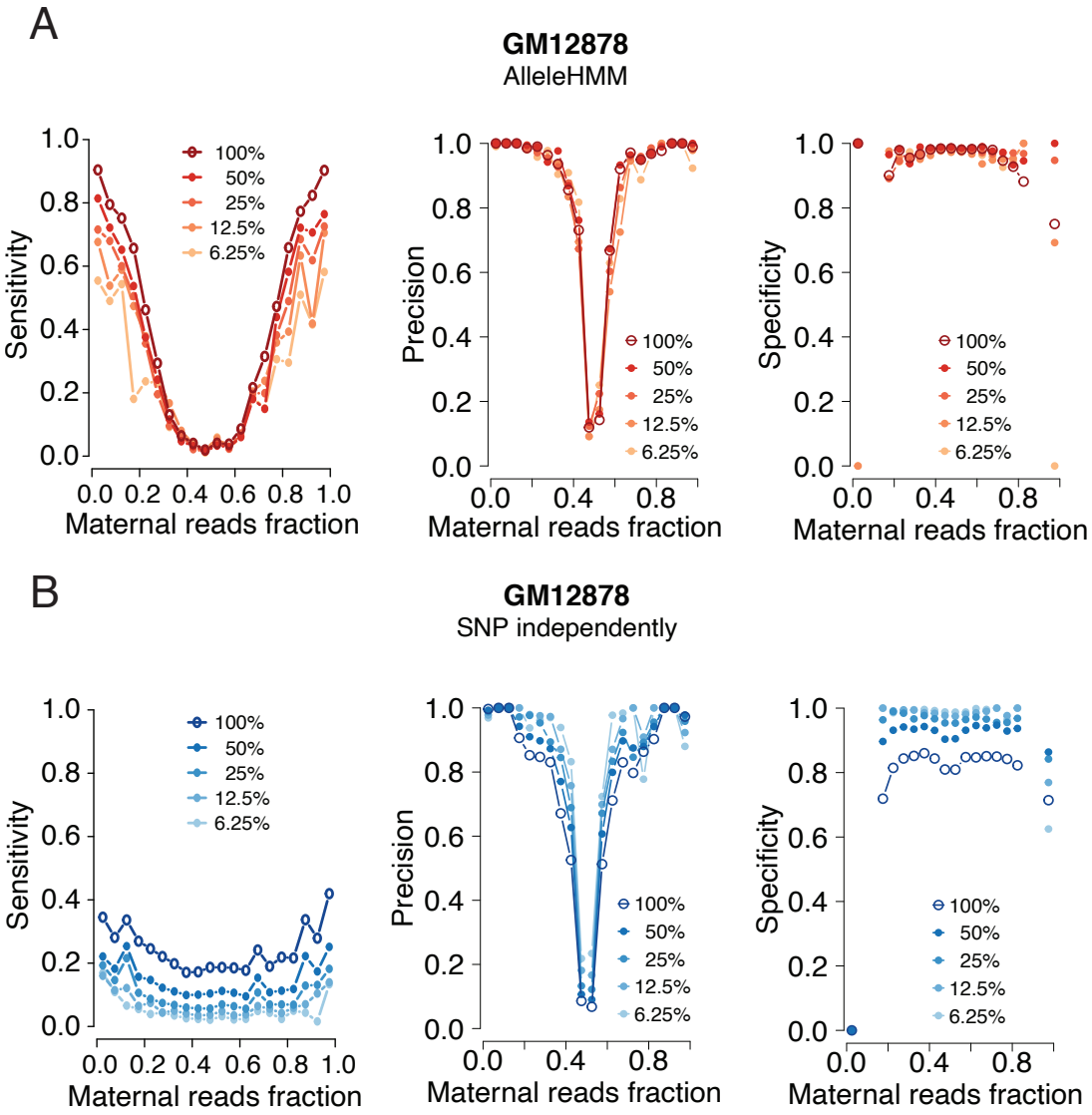
Supplementary Figure 1.4: Comparison of sensitivity and precision using GRO-seq data.

(A) Scatterplots show the sensitivity (left) and precision (right) of AlleleHMM (red) and independent binomial tests (blue) as a function of the maternal reads fraction in the gene annotation using GRO-seq of 129/Castaneus F1 hybrid mESCs.

(B) Scatterplots show the specificity of AlleleHMM (red) and independent binomial tests (blue) as a function of the maternal reads fraction in the gene annotation using GRO-seq of 129/Castaneus F1 hybrid mESCs (left) or GRO-seq of GM12878 (right).



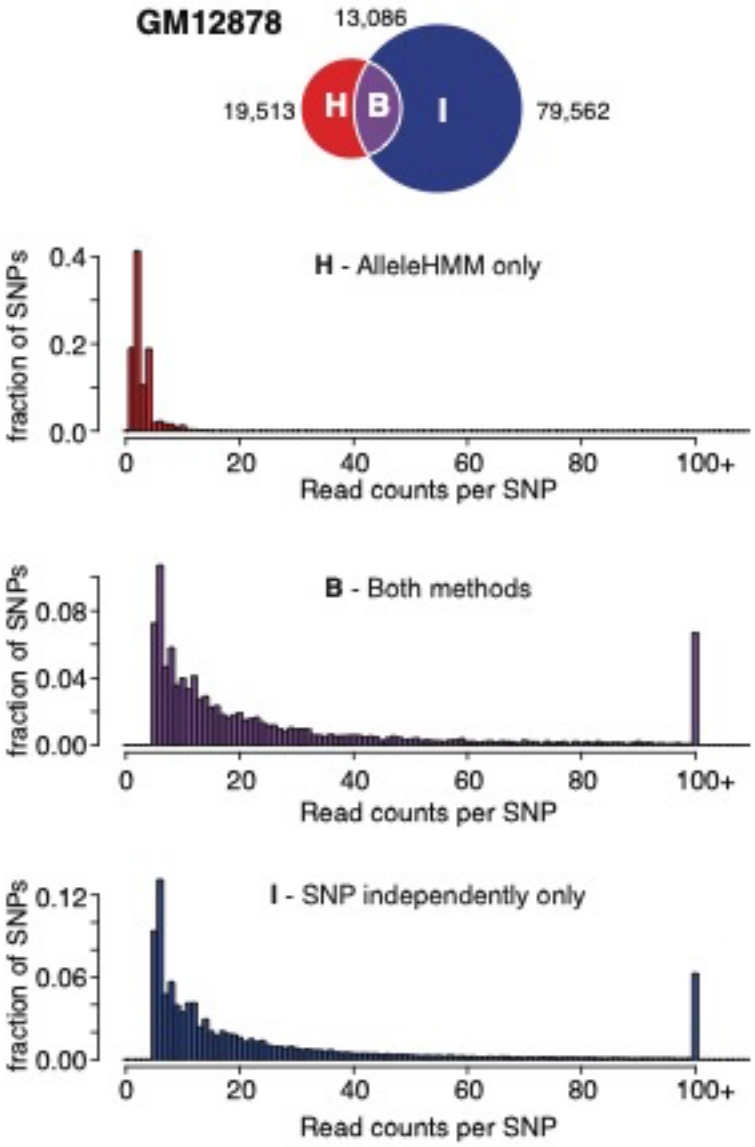
Supplementary Figure 1.5: Scatterplot shows the specificity (blue), sensitivity (red), and precision (green) of AlleleHMM (filled symbol) and independent binomial tests (unfilled symbol) as a function of read depth (subsampled from a total dataset) in human GM12878 GRO-seq data. The values were estimated using SNPs in genes defined as symmetric or allele specific. Symmetric genes have a maternal reads fraction between 0.45 and 0.55 and was classified as symmetric using all the mapped reads in the gene annotation (FDR > 0.1). Allele specific genes have maternal read fractions <0.2 or >0.8 and was classified as significantly allele specific (FDR < 0.01) by performing a binomial test using all the reads in the gene.



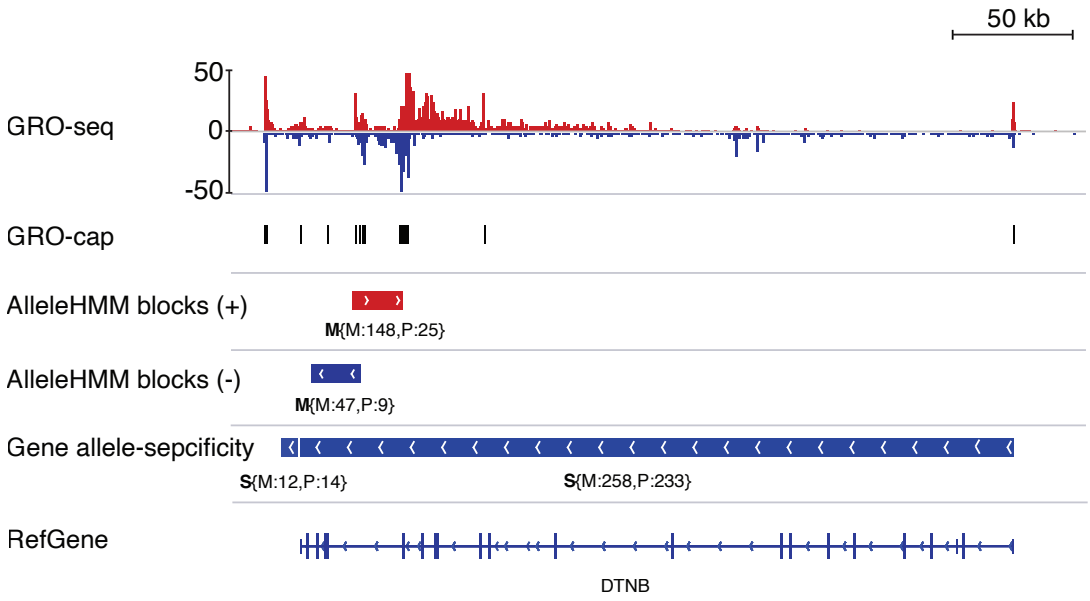
Supplementary Figure 1.6: Effects of subsampling GM12878 on sensitivity, precision, and specificity.

(A) Scatterplots show the sensitivity (left), precision (center), and specificity (right) of AlleleHMM as a function of the maternal reads fraction in the gene annotation. Different lines indicate the read depth of the subsampled GRO-seq reads from a deeply sequenced human GM12878 dataset. The total sequencing depth at 100% is 138 millions uniquely mapped reads.

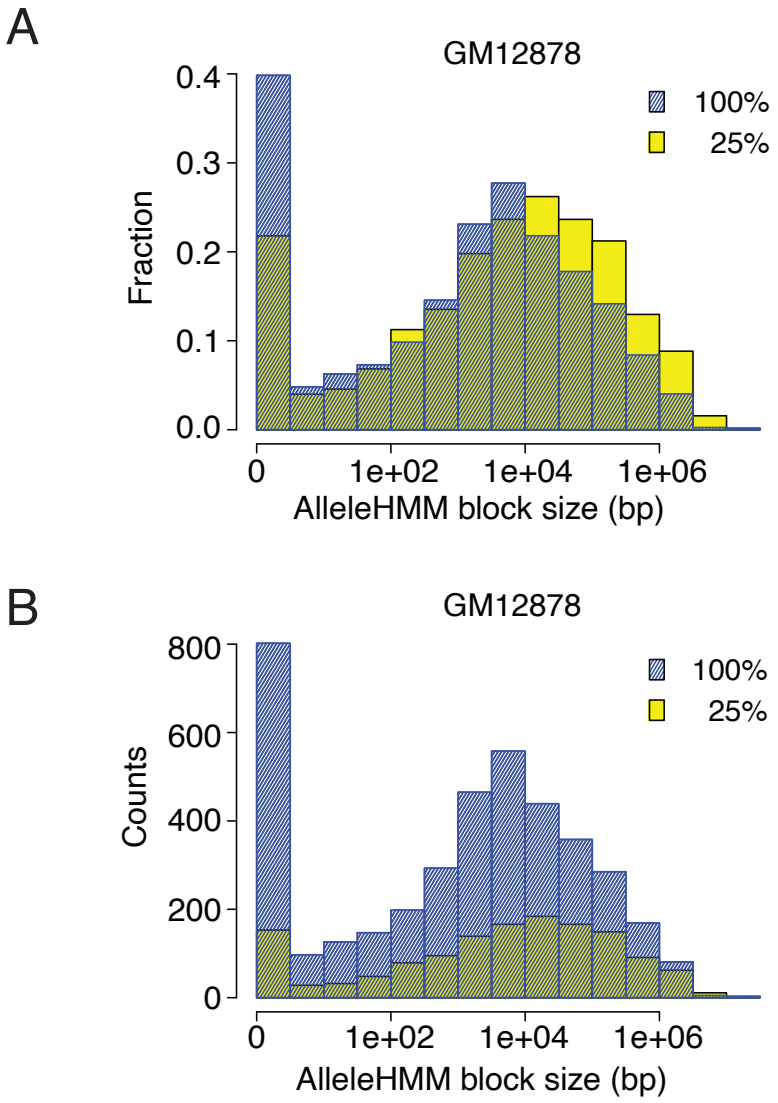
(B) Scatterplots show the sensitivity (left), precision (center), and specificity (right) of independent binomial tests as a function of the maternal reads fraction in the gene annotation. Different lines indicate the read depth of the subsampled GRO-seq reads from a deeply sequenced human GM12878 dataset.



Supplementary Figure 1.7: AlleleHMM identifies blocks with fewer reads supporting each SNP. Histograms show the fraction of SNPs as a function of the read counts per allele specific SNP identified by AlleleHMM only (H, red, top), independent binomial tests (I, blue, bottom), and the intersect of both methods (B, purple, middle) using GRO-seq data from GM12878.



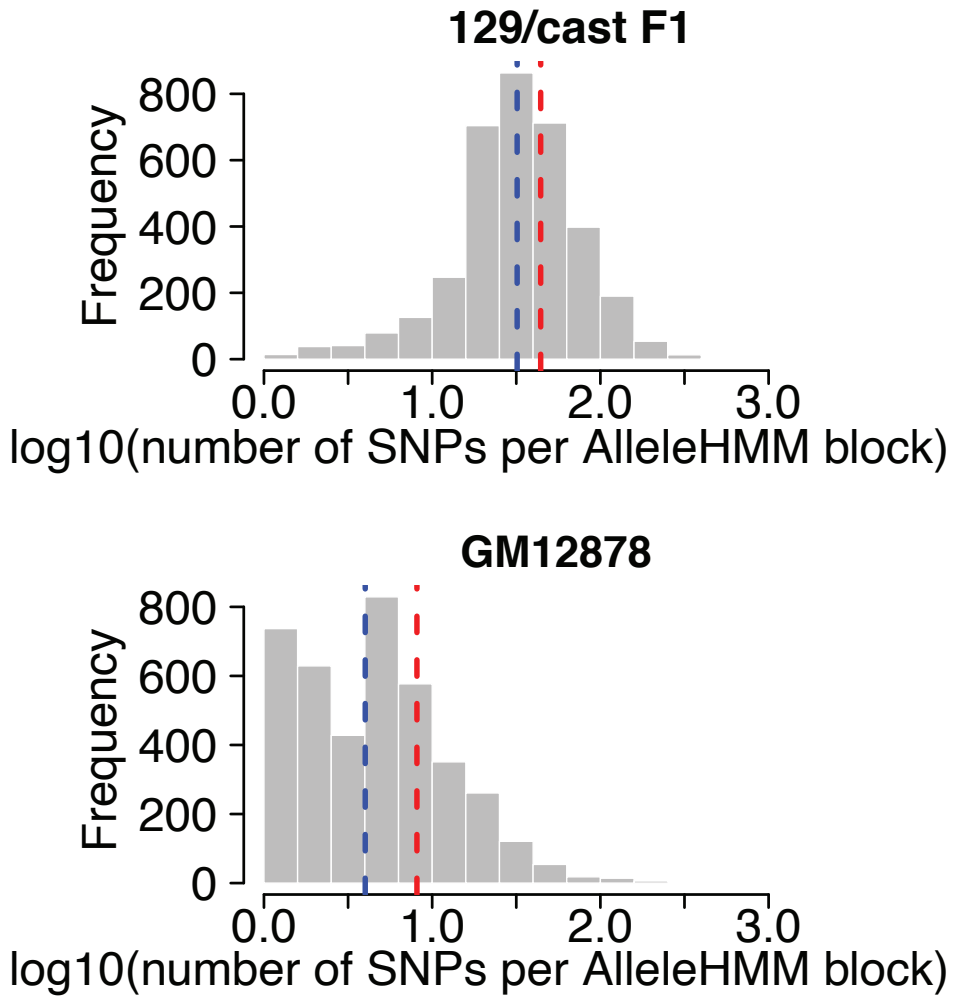
Supplementary Figure 1.8: Genome browser view shows the application of AlleleHMM to GRO-seq data from a human cell line GM12878. The allele specific read counts of the blocks are denoted as **S**{M:258,P:233}, meaning that the block is symmetric (**S**) with 258 maternal-specific (M) reads and 233 paternal-specific (P) reads. GRO-cap data denotes transcription start sites obtained from ref (Core et al., 2014).



Supplementary Figure 1.9: The size of AlleleHMM blocks increased as read depth decreased.

(A) Histograms show the fraction of AlleleHMM blocks having a block size indicated on the X axis in log scale. Data is shown for the full GM12878 GRO-seq dataset (187,896,441 input reads; blue), or a mock dataset with 25% subsample (46,974,111 input reads ;yellow).

(B) Histograms show the counts of AlleleHMM blocks as a function the block size in log scale (X-axis). The blue histogram was calculated using total GRO-seq reads (187,896,441) from GM12878, yellow was calculated using a 25% subsample dataset of 46,974,111 reads.



Supplementary Figure 1.10: Histograms show the distribution of the number of SNPs per AlleleHMM block using GRO-seq of 129/Castaneus F1 hybrid mESCs (top) and a human cell line GM12878 (bottom) in log scale (X-axis). Blue dashed lines indicate the median number of SNPs per AlleleHMM block (129/*cast* is 32.0, GM12878 is 4.0) and red dashed lines indicate the mean (129/*cast* is 44.16, GM12878 is is 8.12).

Supplementary Table 1.1: An example of input file for AlleleHMM.
The input file of AlleleHMM is a tab-delimited table with 4 columns: chrm is the chromosome number, snppos is SNP position sorted by genomic position, mat_allele_count and pat_allele_count are the maternal- or paternal- specific read counts of the SNP.

chrm	snppos	mat_allele_count	pat_allele_count
chr1	565006	0	17
chr1	565286	46	0
chr1	565406	37	0
chr1	565419	31	0
chr1	565591	27	0
chr1	566573	0	2
chr1	568214	0	6
chr1	569094	93	0
chr1	569933	0	2
chr1	724882	2	0
chr1	724883	2	0
chr1	726939	1	4
chr1	726944	5	0
chr1	940005	1	0
[…]			
chr22	50982539	0	1
chr22	50989326	0	2
chr22	51010052	2	0
chr22	51017353	2	0
chr22	51023408	0	2
chr22	51023424	0	2
chr22	51059835	1	2
chr22	51066921	0	2
chr22	51189146	0	1
chr22	51222766	0	1

Supplementary Table 1.2: Table shows the parameters used in the performance test with synthetic data. First 2 column indicates the figure panels and axises that use the specific setting of parameters.

Figures	Y-axis **Sensitivity, Precision**	**Number of SNPs per block**			**Average read count per SNP**			**Expected maternal reads fraction**		
	X-axis	S	M	S	S	M	S	S	M	S
Figure 2.2A, left Supplementary Figure 2.2A, left	Center block length	10	**1, 2, …, 50**	10	10	10	10	0.5	0.9	0.5
Figure 2.2A, center Supplementary Figure 2.2A, center	Average read count per SNP	10	100	10	10	**1, 2, …, 50**	10	0.5	0.9	0.5
Figure 2.2A, right Supplementary Figure 2.2A, right	Expected maternal reads fraction	10	100	10	10	10	10	0.5	**0, 0.05, 0.10, …, 1**	0.5
	Y-axis **Specificity, Precision**									
	X-axis	S	S	S	S	S	S	S	S	S
Figure 2.2B, left	Center block length	10	**1, 2, …, 50**	10	10	10	10	0.5	0.5	0.5
Figure 2.2B, right	Average read count per SNP	10	100	10	10	**1, 2, …, 50**	10	0.5	0.5	0.5

Figures	Y-axis **Sensitivity, Precision**	**Number of SNPs per block**			**Average read count per SNP**			**Expected maternal reads fraction**		
	X-axis	S	M	S	S	M	S	S	M	S
Figure 2.3A, left Supplementary Figure 2.3A	Overdispersion	10	100	10	10	10	10	0.5	0.9	0.5
Figure 2.3B, left Supplementary Figure 2.3B, left	Center block length	10	**1, 2, …, 50**	10	10	10	10	0.5	0.9	0.5
Figure 2.3B, center Supplementary Figure 2.3B, center	Average read count per SNP	10	100	10	10	**1, 2, …, 50**	10	0.5	0.9	0.5
Figure 2.3B, right Supplementary Figure 2.3B, right	Expected maternal reads fraction	10	100	10	10	10	10	0.5	**0, 0.05, 0.10, …, 1**	0.5
	Y-axis **Specificity, Precision**									
	X-axis	S	S	S	S	S	S	S	S	S
Figure 2.3A, right	Overdispersion	10	100	10	10	10	10	0.5	0.5	0.5
Figure 2.3C, left	Center block length	10	**1, 2, …, 50**	10	10	10	10	0.5	0.5	0.5
Figure 2.3C, right	Average read count per SNP	10	100	10	10	**1, 2, …, 50**	10	0.5	0.5	0.5

REFERENCES

BAUM, and L (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. Inequalities *3*, 1–8.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. (2018). The UCSC Genome Browser database: 2018 update. Nucleic Acids Res. *46*, D762–D769.

Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*, 195.

Chen, J., Rozowsky, J., Galeev, T.R., Harmanci, A., Kitchen, R., Bedford, J., Abyzov, A., Kong, Y., Regan, L., and Gerstein, M. (2016). A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat. Commun. *7*, 11101.

Corbett, A.H. (2018). Post-transcriptional regulation of gene expression and human disease. Curr. Opin. Cell Biol. *52*, 96–104.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320.

Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K., Kim, Y., Wang, J.R., Morgan, A.P., Calaway, J.D., Aylor, D.L., et al. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat. Genet. *47*, 353–360.

Danko, C.G., Hyland, S.L., Core, L.J., Martins, A.L., Waters, C.T., Lee, H.W., Cheung, V.G., Kraus, W.L., Lis, J.T., and Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. Nat. Methods *12*, 433–438.

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge University Press).

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.

Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186–192.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rennie, S., Dalby, M., van Duin, L., and Andersson, R. (2018). Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. Nat. Commun. *9*, 487.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. *7*, 522.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*, D726–D732.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory *13*, 260–269.

Wang, Z., Chu, T., Choate, L.A., and Danko, C.G. (2018). Identification of regulatory elements from nascent transcription using dREG. Genome Res.

Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell *162*, 1039–1050.

Yee, T.W. (2015). Vector Generalized Linear and Additive Models: With an Implementation in R (Springer).

CHAPTER 2

Genetic Dissection of the RNA Polymerase II Transcription Cycle†

Shao-Pei Chou[1,3], Adriana K. Alexander[1,2], Edward J. Rice[1], Lauren A Choate[1], Paula E. Cohen[2], and Charles G. Danko[1,2]

1. Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
2. Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
3. Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853.

† This chapter was published as a preprint in bioRxiv in 2021. The citation is doi: 10.1101/2021.05.23.445279. Shao-Pei performed all the data analysis in this manuscript.

2.1 Abstract

How DNA sequence affects the dynamics and position of RNA Polymerase II during transcription remains poorly understood. Here we used naturally occurring genetic variation in F1 hybrid mice to explore how DNA sequence differences affect the genome-wide distribution of Pol II. We measured the position and orientation of Pol II in eight organs collected from heterozygous F1 hybrid mice using ChRO-seq. Our data revealed a strong genetic basis for the precise coordinates of transcription initiation and promoter proximal pause, which was composed of both existing and novel DNA sequence motifs, and allowed us to redefine molecular models of core transcriptional processes. Our results implicate the strength of base pairing between A-T or G-C dinucleotides as key determinants to the position of Pol II initiation and pause. We reveal substantial differences in the position of transcription termination, which frequently do not affect the composition of the mature mRNA. Finally, we identified frequent, organ-specific changes in transcription that affect mRNA and ncRNA expression across broad genomic domains. Collectively, we reveal how DNA sequences shape core transcriptional processes at single nucleotide resolution in mammals.

2.2 Introduction

Transcription by RNA polymerase II (Pol II) is the core process responsible for producing mRNA for all protein-coding genes and most non-coding RNAs (ncRNAs). Transcription by Pol II is a highly stereotyped and cyclic process (Fuda et al., 2009; Jonkers and Lis, 2015). During the Pol II transcription cycle, RNA polymerase is initiated on regions of accessible chromatin by the collective actions of transcription factors and co-factors that recruit the pre-initiation complex, melt DNA, and initiate Pol II (Grünberg et al., 2012; Haberle and Stark, 2018; Murakami et al., 2013; Tsai and Sigler, 2000). After initiation, Pol II pauses near the transcription start site of all genes in most metazoan genomes (Jonkers et al., 2014; Muse et al., 2007; Rougvie and Lis, 1988). Pol II is released from pause through the collective actions of transcription factors and a key protein kinase complex (P-TEFb), a tightly regulated step that controls the rates of mRNA production (Danko et al., 2013; Dig B. Mahat et al., 2016; Rahl et al., 2010; Zeitlinger et al., 2007). After pause release, Pol II elongates through gene bodies, which in some cases cover more than 1 MB of DNA in mammals (Carninci et al., 2005). Finally the co-transcriptionally processed pre-mRNA is cleaved from the elongating Pol II complex, allowing termination and recycling of Pol II (Cho et al., 1999; O'Sullivan et al., 2004; Rosonina et al., 2006).

Each of these core stages of Pol II transcription require the coordinated efforts of dozens of macromolecules. Through outstanding research achievements during the past decades, we have a new appreciation of the molecular players involved in each stage of the transcription cycle (Gilchrist et al., 2010; Miller et al., 2001; Nechaev et al., 2010; Orphanides et al., 1998; Ranish et al., 1999), as well as the dynamics of individual

stages (Danko et al., 2013; Gressel et al., 2019, 2017; Jonkers et al., 2014; Schwalb et al., 2016).

Despite an expanding knowledge of the macromolecules involved in each stage of transcription, we still have a relatively rudimentary understanding about how DNA sequences influence each step in the Pol II transcription cycle. Previous studies have identified core promoter motifs that are correlated with transcription initiation, such as the TATA box and the initiator motif (Carninci et al., 2006; Smale and Baltimore, 1989). Likewise, weak motifs that correlate with pausing and Pol II termination are reported (Gressel et al., 2017; Schwalb et al., 2016; Tome et al., 2018). These DNA sequences that underlie initiation play an important role in gene regulation, and SNPs affecting these core transcriptional processes can alter the rates of mRNA production (Kristjánsdóttir et al., 2020). However, the DNA sequence specificity of the core promoter is extremely weak, degenerate, and spread across the promoter region. As a consequence, we are currently unable to predict how specific DNA sequence changes will affect the core transcriptional processes involved in transcription.

Here we use naturally occurring genetic variation between highly heterozygous F1 hybrid mice to understand how DNA sequence differences between alleles affect the Pol II transcription cycle. We generated an atlas of the position and orientation of RNA polymerase II in eight organs collected from three primary germ layers, using ChRO-seq to identify the position and orientation of RNA polymerase (Chu et al., 2018). Our results provide insight into how DNA sequences shape transcription initiation and pausing. We reveal substantial, heritable differences in the position of transcription termination, which frequently do not affect the composition of the mature mRNA.

Finally, we identified frequent, organ-specific changes in transcription that affect mRNA and ncRNA expression across broad genomic domains. Collectively, our results provide new insight into how genetics shape core transcriptional processes in mammals.

2.3 Results

2.3.1 Atlas of allele specific transcription in F1 hybrid murine organs

We obtained reciprocal F1 hybrids from two heterozygous mouse strains, C57BL/6 (B6) and Castaneus (CAST) (**Figure 2.1A**). Mice were harvested in the morning of postnatal day 22 to 25 from seven independent crosses (3x C57BL/6 x CAST and 4x CAST x C57BL/6; all males). We measured the position and orientation of RNA polymerase genome-wide in 8 organs using a ChRO-seq protocol that provides accurate allelic mapping by extending the length of reads using strategies similar to length extension ChRO-seq (Chu et al., 2018) (**see Methods**). We obtained 376 million uniquely mapped ChRO-seq reads across all eight organs (21-86 million reads per organ; **Supplementary Table 2.1**) after sequencing, filtering, and mapping short reads to individual B6 and CAST genomes (**see Methods**). Hierarchical clustering using Spearman's rank correlation of ChRO-seq reads in GENCODE annotated gene bodies (v.M25) grouped samples from the same organ. Additionally, organs with similarities in organ function clustered together, for instance: heart and skeletal muscle, and large intestine and stomach (**Figure 2.1B**).

Using ChRO-seq data from all eight organs, we identified 3,494 broad domains that showed consistent evidence of allelic imbalance across biological replicates (**Figure 2.1C**). Using the reciprocal hybrid cross design, we found that the majority of these domains (n = 3,466; **Figure 2.1D**) have consistent effects in each mouse strain (called strain-effect domains), the pattern expected if allelic imbalance was caused by DNA sequence differences between strains. Twenty-eight domains showed consistent evidence of genomic imprinting (imprinted domains). Both strain-effect and imprinted

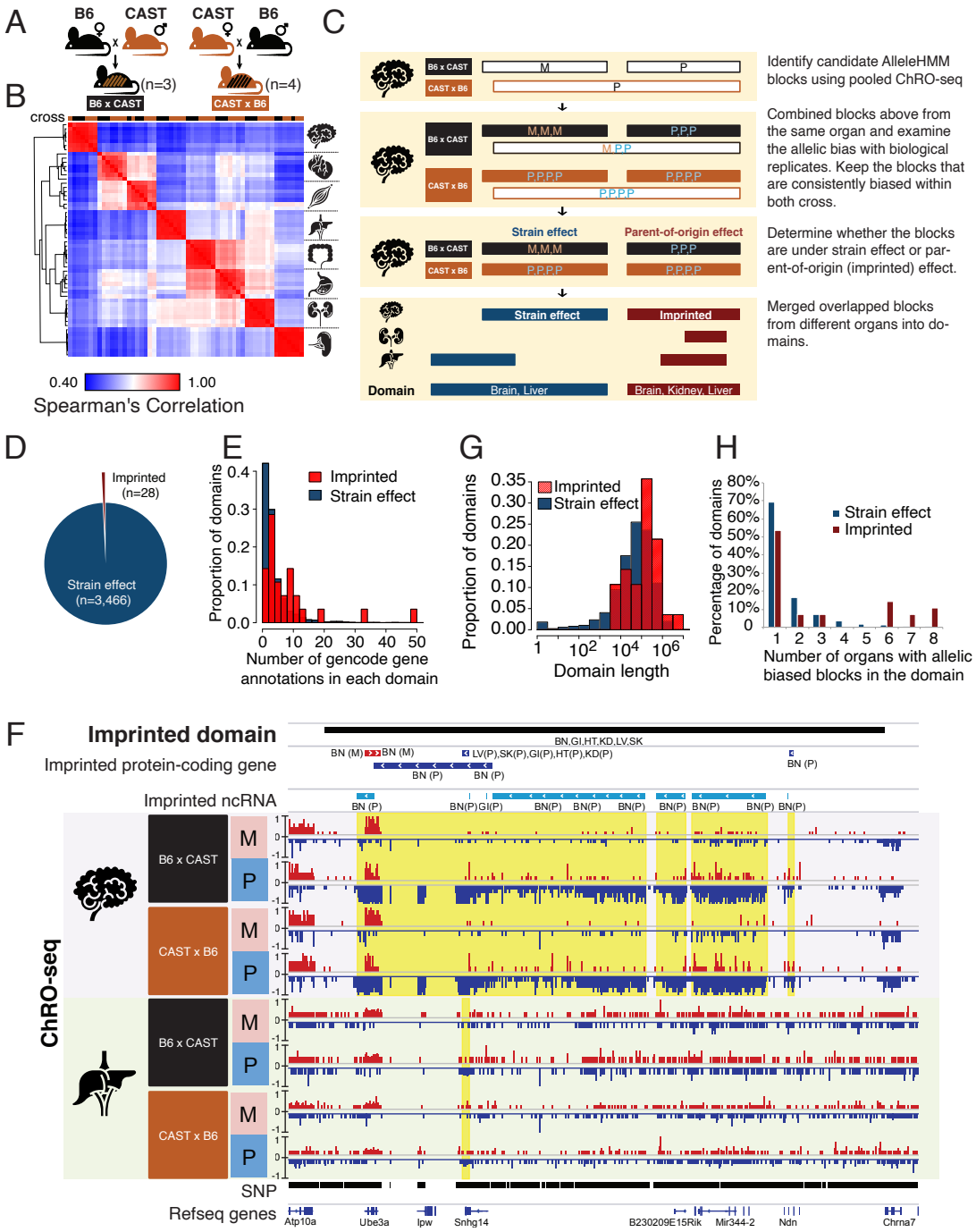


Figure 2.1: Reciprocal hybrid cross to understand the Pol II transcription cycle.

(A) Cartoon illustrates the reciprocal F1 hybrids cross design between the strains C57BL/6J (B6) and CAST/EiJ (CAST). We have seven independent crosses (3x C57BL/6 x Cast and 4x Cast x C57BL/6).

(B) Spearman's rank correlation of ChRO-seq signals in gene bodies. The color on the top indicates the direction of crosses: Black is B6 x CAST, brown is CAST x B6. The cartoon on the right indicates the organ each sample was harvested from.
(C) Cartoon depicts the methods used to identify allelic biased blocks, domains, and how they were classified as a strain-associated effect or imprinted.
(D) The pie chart shows the proportion of domains under strain effect or imprinted.
(E) The histogram shows the proportion of domains as a function of the number of gencode gene annotations in each domain.
(F) The browsershot shows an example of ChRO-seq data that has an imprinted domain (top row). The second and third rows show the imprinted protein-coding genes and imprinted non-coding RNA (ncRNA) from all organs. (BN:brain, LV:liver, SK:skeletal muscle, GI: large intestine, HT: heart, KD: kidney, P: paternal, M: maternal). The yellow shade indicates the imprinted regions in the brain and liver.
(G) The histogram shows the proportion of domains as a function of the domain length.
(H) The bar chart shows the percentage of domains as a function of the number of organs with allelic biased blocks in the domain.

domains were generally composed of multiple transcription units, including annotated genes (**Figure 2.1E**) and ncRNAs (either long intergenic ncRNAs and/ or enhancer-templated RNAs). For instance, the imprinted domain associated with Angelman syndrome consisted of twenty ncRNAs and four genes that were consistently transcribed more highly from the paternal allele and two genes transcribed from the maternal allele in brain tissue (**Figure 2.1F**). On average, both strain-effect and imprinted domains spanned broad genomic regions (~10-1,000 kb; **Figure 2.1G**) which were reminiscent of regional, coordinated effects across functional elements (Delaneau et al., 2019; Rennie et al., 2018).

Recent reports in humans suggest that organs frequently share common genetic factors that shape gene expression (GTEx Consortium et al., 2017). In contrast to these recent observations, nearly 70% of strain-effect domains and 50% of imprinted domains were organ specific (**Figure 2.1H**). Organ-specific allelic biased domains were not dominated by false-negatives in putatively unbiased organs, as the magnitude of allelic balance in putatively unbiased organs was distributed around 0 (**Supplementary Figure 2.1A,B**). Organ-specific allelic bias also could not be explained by organ-specific differences in gene expression, as organs that showed no evidence of allelic bias frequently had similar or higher transcription levels of allelic biased genes (**Supplementary Figure 2.1C**). Thus, allelic bias is organized in large multi-transcript domains that are often organ-specific.

2.3.2 Enhanced genomic imprinting in murine brain

Our reciprocal cross design allowed us to define 28 domains that show clear evidence of genomic imprinting, most of which were previously reported by others (Andergassen et al., 2017; Wang et al., 2011). We noted that genomic imprinting was frequently stronger in the brain than other adult organs analyzed here. Although imprinting was less likely to be organ-specific than strain-effect domains (**Figure 2.1H**), those which were organ specific were largely imprinted in brain (**Figure 2.2A**). Moreover, of the domains that were imprinted in 6 or more organs (n = 9 of 28; ~33%), the imprinting domain was generally larger in the brain than in other organs, consistent with previous observations (Andergassen et al., 2017; Plasschaert and Bartolomei, 2015; Wilkins, 2014). An outstanding example was the Angelman and Prader-Willi syndrome locus, which was imprinted in multiple organs including brain, liver, heart, skeletal muscle, large intestine, and kidney. However, while imprinting only affected *Snhg14* and *Snurf* in most organs, imprinting affected a much broader region in the brain, resulting in imprinting of additional protein coding genes, including *Ube3a* and *Ipw* (**Figure 2.1F**). This finding may explain why both Angelman and Prader-Willi syndrome have symptoms associated with behavior (Ho and Dimitropoulos, 2010; Pelc et al., 2008). In another example, we noted the opposite parent-of-origin effect in the brain as observed in other organs. *Grb10* was imprinted in multiple organs, but it was transcribed from the opposite allele in the brain (paternal-specific) than all other adult mouse organs (maternal-specific). Taken together, these observations show that the brain frequently has a unique pattern of genomic imprinting that is not observed in other somatic organs.

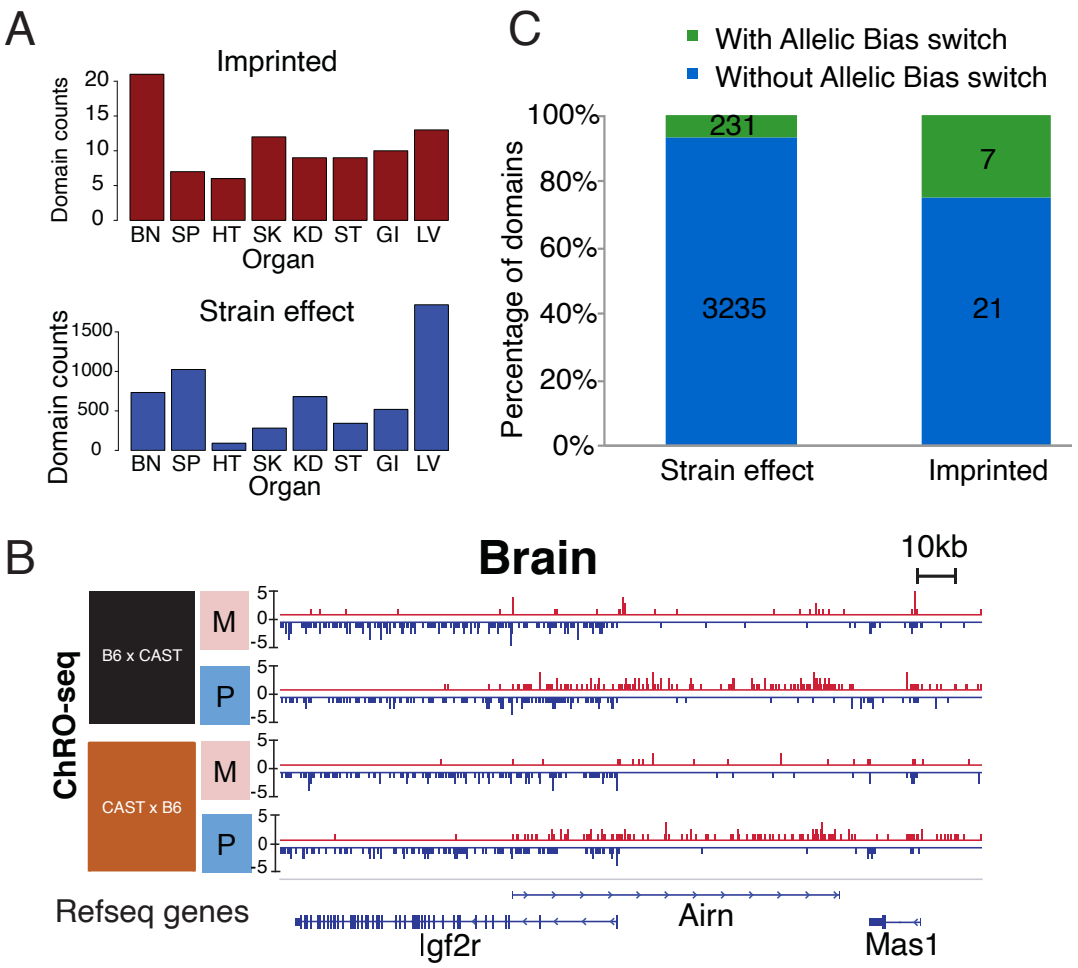


Figure 2.2: Features associated with imprinting domains.

(A) The bar charts show the number of imprinted (top) and strain effect domains (bottom) in each organ. (BN:brain, SP: spleen, HT: heart, SK:skeletal muscle, KD: kidney, ST:stomach, GI: large intestine, LV:liver)

(B) The browsershot shows allele-specific ChRO-seq signal from the *Igf2r* locus in brain, which contains the imprinting-associated ncRNA, *Airn*. M: maternal-specific reads; P: paternal-specific reads.

(C) Stacked bar charts show the percentage of strain effect and imprinted domains that contain transcription units with opposite allelic bias (called switches). Those without allelic bias switches were consistently biased to the same direction (for imprinted domain, all maternally-biased or all paternally-biased, no switch). Those domains with allelic bias switches contain blocks biased to different directions (some maternally-biased and some paternally-biased, with switches).

2.3.3 Discovery of imprinted ncRNAs

In several well-characterized cases, imprinting is initiated by a ncRNA, as elegantly illustrated at the *Igf2r* / *Airn* locus (Sleutels et al., 2002). ChRO-seq clearly defined the location of known ncRNAs, including *Airn* and *Kcnq1ot1* that mediate genomic imprinting (**Figure 2.2B**). We discovered new candidate ncRNAs at several additional imprinted loci, including in the Angelman and Prader-Willi syndrome locus (**Figure 2.1F**). Frequently, ncRNAs like *Airn* had an opposite allelic bias as the nearby imprinted protein-coding genes. Similar patterns were observed for ncRNAs in the Angelman and Prader-Willi syndrome locus (chr7), in which ncRNAs across the locus were transcribed in the opposite parental allele as several of the imprinted genes (e.g., *Ube3a*). In another interesting example, two protein coding genes, *Peg3* and *Usp29*, were transcribed with opposite allelic bias as another protein coding gene, *Zim1*, without a clear ncRNA within the locus. Overall, 7 of the 28 (25%) imprinted domains had transcripts with an opposite allelic bias occurring within the same domain, which reflects a significantly higher proportion than the 7% strain effect domains (Fisher's exact test, Odds ratio = 4.66, $P = 0.002$, **Figure 2.2C**). We note allelic differences in expression noted here are similar to well characterized examples, such as in X-chromosome inactivation, in which a ncRNA (*Xist*) is transcribed from the inactive copy of the X chromosome(Augui et al., 2011). These differences may reflect the effect of nascent RNA transcription on the chromatin architecture of the locus, in a manner reminiscent of *Airn*, and possibly *Xist*.

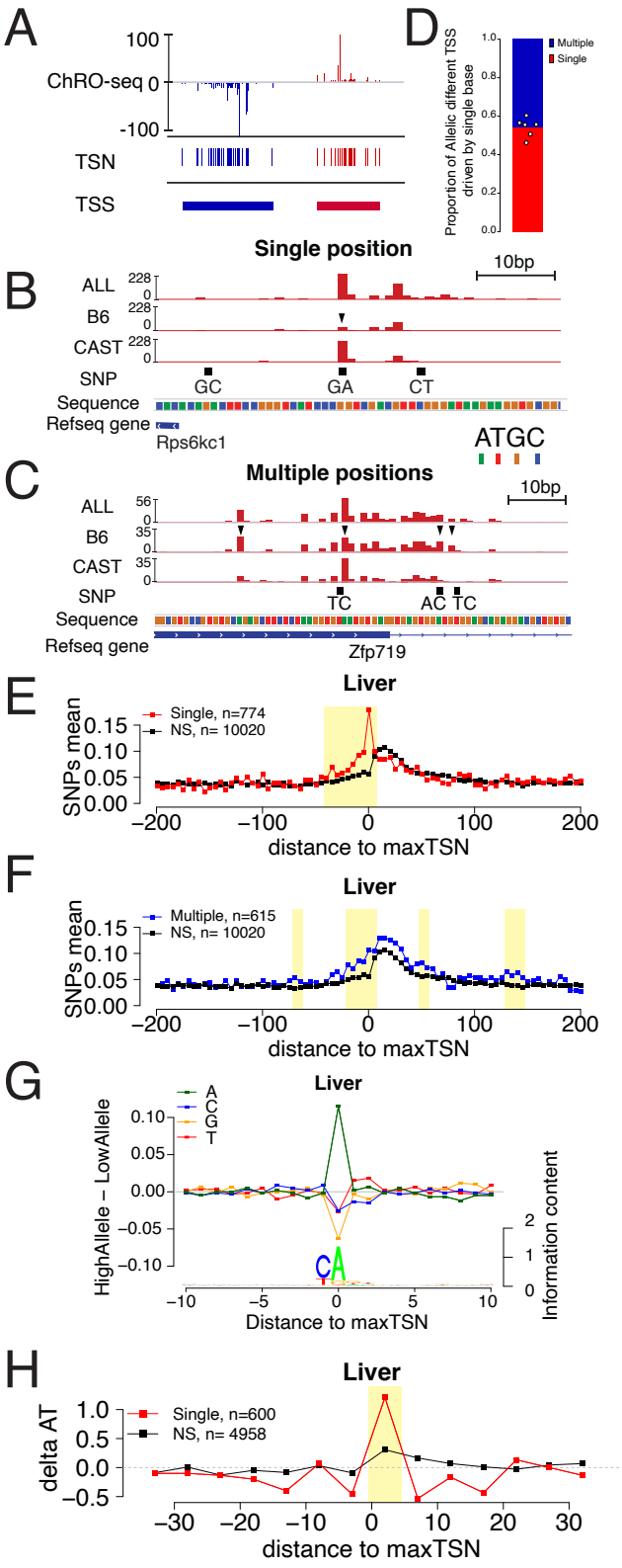


Figure 2.3: Allele specific effects on the distribution of transcription initiation.

(A) The ChRO-seq signals of the 5’ end of the nascent RNA were used to define transcription start nucleotides (TSNs), or the individual bases with evidence of transcription initiation. TSNs within 60bp were grouped into transcription start sites (TSSs).
(B) The browsershot shows an example of allelic differences in the shape of TSS that are predominantly explained by a single TSN position (arrowhead).
(C) The browsershot shows an example of allelic differences in TSS driven by multiple TSNs within the same TSS, arrowheads indicate several prominent positions with allelic differences in Poll abundance.
(D) Stacked bar chart shows the average proportion of allelic differences in TSSs driven by a single TSN (red). The specific proportion for each of the six organs are shown by dots. Six organs were selected that showed a strong signal of Inr motif at the maxTSNs (brain, heart, liver, large intestine, stomach, and spleen).
(E) Scatterplot shows the average SNP counts as a function of distance to the maxTSN at sites showing allelic differences in TSSs driven by a single base in the liver. Red denotes changes in TSS shape (Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSSs without evidence for differences in TSS shape (KS test; FDR > 0.90). Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05)
(F) Scatterplot shows the average SNP counts as a function of distance to the maxTSN at sites showing allelic differences in TSS driven by multiple bases in the liver sample. Blue denotes changes in TSS shape classified as multiple TSN driven (Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSSs without evidence for differences in TSS shape (KS test; FDR > 0.90). Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05)
(G) The scatterplot shows the average difference in base composition between the allele with high and low TSN use around the maxTSN in single-base driven allele specific TSSs. The sequence logo on the bottom represents the high allele in single-base driven allele specific TSSs. The high/low allele were determined by the read depth at maxTSN.
(H) The scatter plot shows the difference of AT contents between the high and low alleles with the maxTSN and -1 base upstream maxTSN masked. Dots represent 5 bp non-overlapping windows. Red denotes single base driven allele specific TSSs; black denotes control TSSs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT (at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher’s exact test, FDR <= 0.05).

2.3.4 Widespread genetic changes in transcription initiation

To begin dissecting the genetic basis for each stage of the transcription life cycle, we first focused on defining allele specific patterns of transcription initiation. The 5' end of nascent RNA, denoted by the 5' end of paired-end ChRO-seq reads, marks the transcription start nucleotide (TSN) of that nascent RNA (Kwak et al., 2013; Tome et al., 2018). We identified TSNs in which at least 5 unique reads share the same 5' end inside of regions enriched for transcription initiation identified using dREG after merging all samples from the same tissue (Wang et al., 2018). Using a recently described hierarchical strategy (Tome et al., 2018), we grouped candidate TSNs into transcription start sites (TSSs) and defined the maximal TSN as the position with the maximum 5' signal within each TSS (**Figure 2.3A**). To verify that our pipeline identified transcription start nucleotides accurately, even in the absence of enzymatic enrichment for capped RNAs, we examined whether candidate transcription initiation sites were enriched for the initiator DNA sequence element, a defining feature of Pol II initiation (Kaufmann and Smale, 1994; Smale and Baltimore, 1989). Most organs, especially brain and liver (which were the most deeply sequenced and are the focus of the analysis below unless otherwise specified), had high information content showing the initiator element at maxTSNs (**Supplementary Figure 2.2A**). Moreover, the relationship between read depth and TSNs was similar to those recently reported in human cells (Tome et al., 2018) (**Supplementary Figure 2.2B**).

Allelic changes in transcription initiation frequency were common, occuring in ~3-7% of different TSSs (n = 1,109 - 5,793; binomial test FDR < 0.1). Changes in the frequency of initiation from TSSs predominantly reflect changes in the rates of

transcription initiation at that gene. These changes likely reflect allelic differences in transcription factor binding and other regulatory processes that have been explored extensively elsewhere (Battle et al., 2014; Chen et al., 2016; Lappalainen et al., 2013; Montgomery et al., 2010; Pickrell et al., 2010).

To better define how DNA sequence shapes the precise genomic coordinates of transcription initiation, we focused on allele-specific changes in the position of TSNs within TSSs. We identified 1,006 (brain) and 1,389 (liver) TSSs in which the shape of the 5' end of mapped reads within that TSS changed between alleles (FDR <= 0.1; Kolmogorov-Smirnov [KS] test) (see examples in **Figure 2.3B-C**). As TSSs are typically regions spanning ~80 bp that are comprised of multiple TSNs (Carninci et al., 2006; Tome et al., 2018), which could have distinct patterns of allelic imbalance, we divided changes in TSS shape into two classes: cases that were driven predominantly by large changes in the abundance of Pol II at a single TSN position and cases in which multiple TSNs across the TSS contributed to changes in shape (see **Methods**). For example, the TSS giving rise to the transcription unit upstream and antisense to *Rps6kc1* had major differences in just one of the TSNs (**Figure 2.3B, arrow head**), whereas the promoter of *Zfp719* has multiple changes in TSNs within the same TSS (**Figure 2.3C, arrow heads**). In both examples, allelic changes result in differences in the position of the max TSN between alleles. Each of these two classes comprised approximately half of the changes in TSN shape in all organs examined here (**Figure 2.3D**).

We examined how SNPs influence the precise location of transcription initiation between alleles. We compared the distribution of SNPs at allele specific TSSs to a control set composed of non-allele-specific TSSs, which was designed to control for the

ascertainment biases associated with having a tagged SNP in each allelic read. Single position driven allele-specific TSSs were associated with a strong, focal enrichment of SNPs around maxTSNs while the multiple position driven allele specific TSSs had a weaker, broader enrichment of SNPs (**Figure 2.3E-F** [liver] and **Supplementary Figure 2.2C-D** [brain], yellow shade indicates FDR <= 0.05; Fisher's exact test). SNPs within 5 bp of the initiation site explain up to ~15-20% of single-base driven allele specific transcription start sites (**Figure 2.3E** and **Supplementary Figure 2.2C**). This was predominantly explained by SNPs at the transcription start site itself, with an A highly enriched in the allele with highest max TSN usage at that position, consistent with the sequence preference of the initiator motif (**Figure 2.3G**). By contrast, the enrichment of C in the -1 position of the initiator motif was much weaker than the A at the 0 position. Although this result may partially be explained by a bias in which the C allele does not tag nascent RNA, it was consistent between organs (**Supplementary Figure 2.2E**) and controlled based on the composition of the background set. Thus, our results suggest that the A in the initiation site may be the most important genetic determinant of transcription initiation. Changes in TSS structure driven by multiple, separate TSNs, were enriched throughout the ~30 bp upstream, and to some extent downstream, potentially implicating changes in core transcription factor binding motifs or sequence specific transcription factors that influence the precise initiation site (**Figure 2.3F** and **Supplementary Figure 2.2D**).

Next, we examined other factors near the TSS, aside from DNA within the initiator motif itself, that contributed to single base driven TSN choice. We noticed a higher frequency of A and T alleles downstream of the initiator motif on the allele with

a higher max TSN (**Figure 2.3G**). We hypothesized that the lower free energy of base pairing in A and T alleles would make them easier to melt during initiation, and could therefore increase the frequency of TSN usage at these positions. Indeed, a more direct examination of AT content in 5 bp windows near the maxTSN identified a significantly higher AT content on the allele with the highest max TSN after masking DNA at positions -1 and 0 to avoid confounding effects of the initiator element (**Figure 2.3H** and **Supplementary Figure 2.2F**). This enrichment of high AT content was consistent in both brain and liver tissue, but was unique to single base driven max TSNs (**Supplementary Figure 2.2G-H**). We conclude that multiple aspects of DNA sequence, including both sequence motif composition and the energetics of DNA melting, influence TSN choice in mammalian cells.

2.3.5 Models of stochastic search during transcription initiation

Next we examined how SNPs that affect a particular TSN impact initiation within the rest of the TSS. In the prevailing model of transcription initiation in *S. cerevisiae*, after DNA is melted, Pol II scans by forward translocation until it identifies a position that is energetically favorable for transcription initiation (Braberg et al., 2013; Kaplan et al., 2012; Qiu et al., 2020) (**Figure 2.4A**). In mammals, Pol II is not believed to scan, but rather each TSN is believed to be controlled by a separate PIC (Luse et al., 2020). We considered how mutations in a strong initiator dinucleotide (CA) would affect transcription initiation under each model. Under the yeast model, we expected CA mutations to shift initiation to the next valid initiator element downstream. Under

the mammalian model, we expected each TSN to be independent and therefore a mutation in the TSN would have no effect on the pattern of nearby initiation sites.

We analyzed 277 and 372 TSNs in brain and liver, respectively, where the high allele contained a CA dinucleotide while the other allele did not. Candidate initiator motifs within 20 bp of the CA/ non-CA initiation site had slightly more initiation signal on the non-CA allele compared with the CA allele, consistent with the shooting gallery model but inconsistent with the prevailing model of independent TSNs expected in mammals (**Figure 2.4B; purple**). By contrast, TSNs where both alleles contained a CA dinucleotide (n = 8,147 [brain] and 10,113 [liver]) did not show this same effect (**Figure 2.4B; gray**). The difference was found for both adjacent CA dinucleotides and for weaker candidate (Py)(Pu) initiator elements, and was consistent across both single-base and multiple-base TSS configurations (**Supplementary Figure 2.3**). Unexpectedly, we also observed consistently higher initiation signals on the non-CA allele both upstream and downstream of the initiator dinucleotide (**Figure 2.4C**). The signal for an increase on the non-CA allele stretched up to 20bp both upstream and downstream of the CA/ non-CA dinucleotide. For instance, a SNP in the initiator element nearly abolished the dominant max TSN of the protein coding gene *Smg9* in CAST (**Figure 2.4D, orange block**). Instead, initiation in CAST shifted to a new maxTSN located upstream (**Figure 2.4D, arrow head**), and also increased usage of several minor TSNs downstream (**Figure 2.4D**). These findings are most consistent with a model in which DNA is melted to give Pol II access to the template strand, at which point Pol II scanning for an energetically favorable initiator element occurs in both directions by a stochastic process resembling Brownian motion.

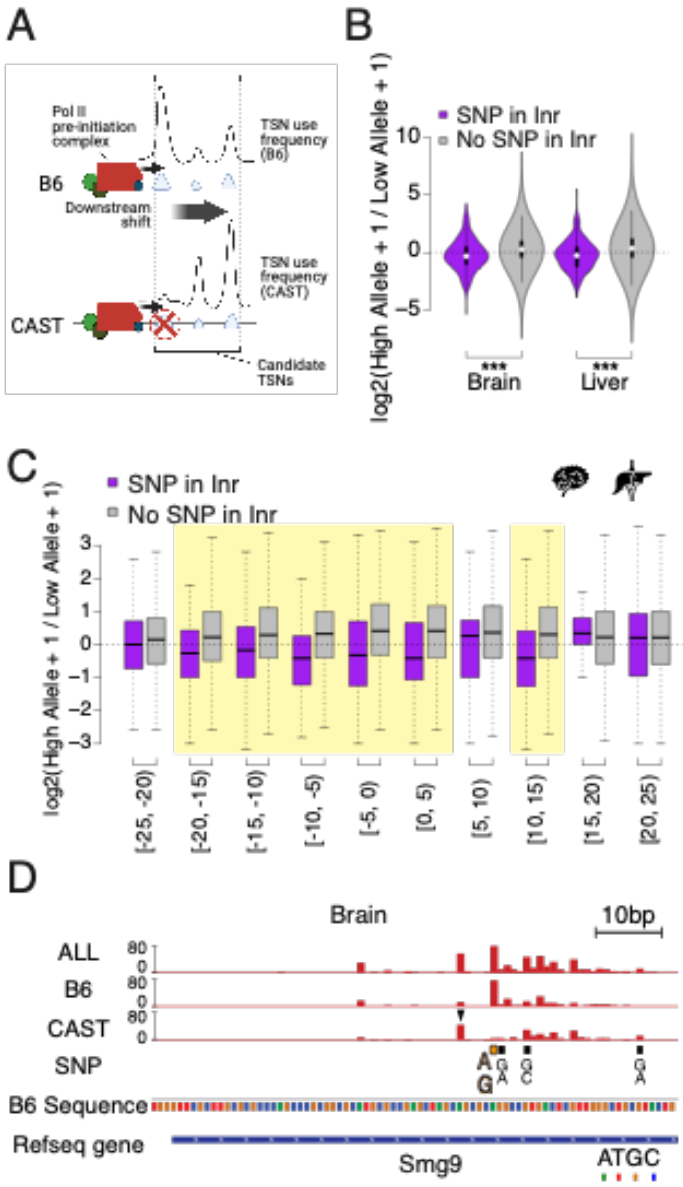


Figure 2.4: Browian motion model of transcription start nucleotide selection.

(A) Cartoon shows our expectation of the effects of allelic DNA sequence variation on transcription start nucleotide selection based on the “shooting gallery” model, in which Pol II initiates at potential TSNs (triangles) after DNA melting. We expected that mutations in a strong initiator dinucleotide (CA) on (for example) the CAST allele (bottom) would shift initiation to the initiator elements further downstream. The size of the triangle indicates the strength of the initiator.

(B) The violin plots show the distribution of ChRO-seq signal ratios on the candidate initiator motifs (including CA, CG, TA, TG) within 20bp of the maxTSNs that had a CA dinucleotide in the allele with high maxTSN (SNP in Inr, purple) or had a CA dinucleotide in both alleles (No SNP in Inr, gray). Note that the central maxTSN was not included in the analysis. Wilcoxon rank sum test with continuity correction is p-value = 5.665e-10 for Brain and p-value < 2.2e-16 for liver.

(C) The box plots show the distribution of ChRO-seq signals ratios at TSNs with any variation of the initiator motif (including CA, CG, TA, TG) as a function of the distance from the maxTSNs that had a CA dinucleotide in the allele with high maxTSN (SNP in Inr, purple) or had a CA dinucleotide in both alleles (No SNP in Inr, gray). Yellow shade indicates Wilcoxon Rank Sum and Signed Rank Tests (SNP in Inr vs no SNP in Inr) with fdr <= 0.05. The TSSs were combined from the brain and liver samples.

(D) The browser shot shows an example of a maxTSN with increased initiation upstream and downstream of an allelic change in a CA dinucleotide. The orange block AG denotes a SNP at the maxTSN, in which B6 contains the high maxTSN with CA and CAST contains CG. The ChRO-seq signals at the alternative TSN with a CA dinucleotide (arrow head) upstream of the maxTSN were higher in the low allele (CAST in this case), resulting in a different maxTSN in CAST.

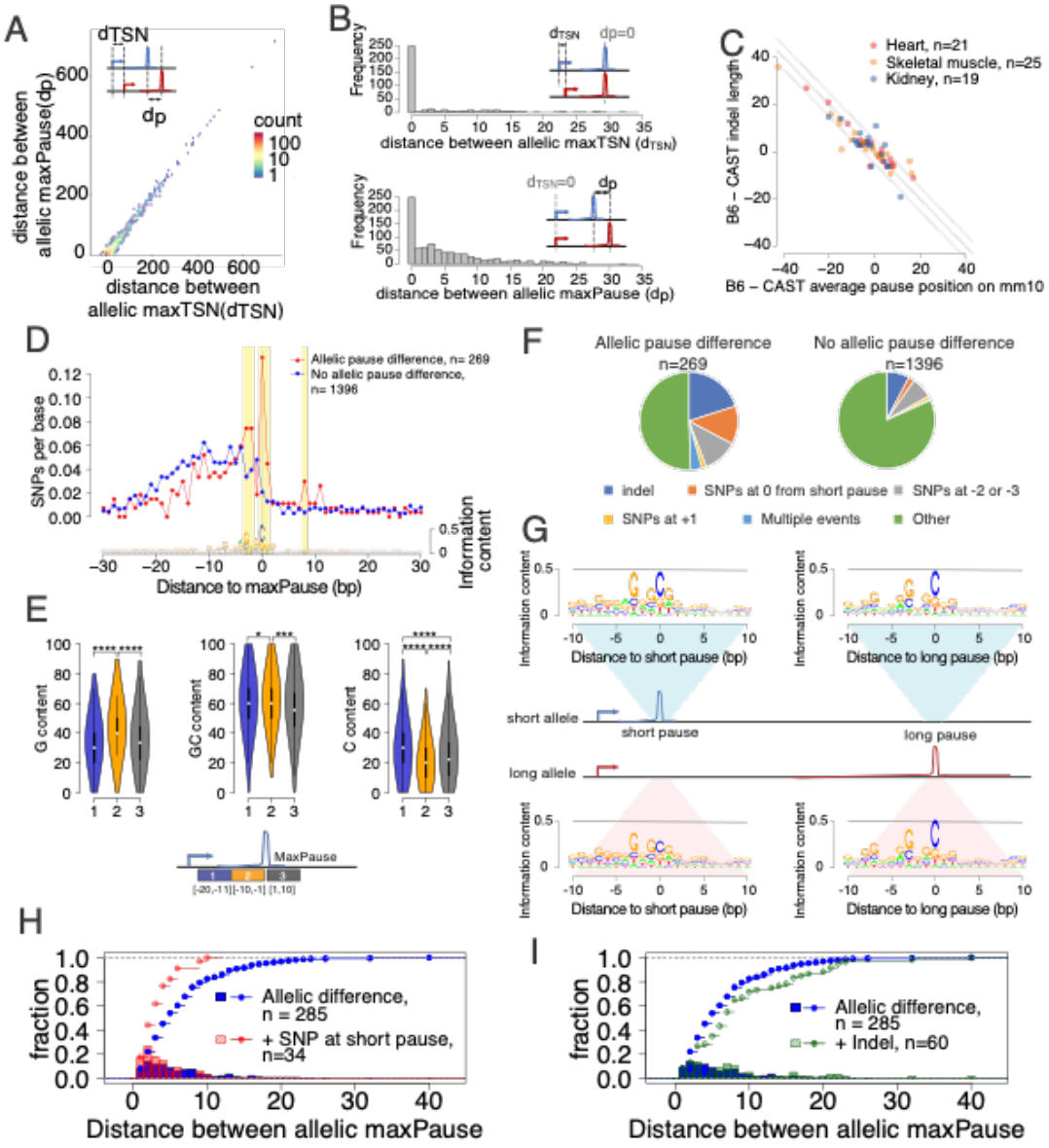


Figure 2.5: Allele specific effects on the distribution of Pol II in the promoter proximal pause.

(A) Scatterplots show the relationship between distances of allelic maxPause and allelic maxTSN within dREG sites with allelic different pause (n = 2,260).

(B) Top histogram shows the number of sites as a function of the distance between allelic maxTSN in which the allelic maxPause was identical (n = 359). Bottom histogram shows the number of sites as a function of the distance between allelic maxPause where the allelic maxTSN was identical (n = 823).

(C) Scatterplot shows the relationship between indel length and the allelic difference of the average pause position on the reference genome (mm10). The pause positions of CAST were first determined in the CAST genome and then

liftovered to mm10. Only sites initiated from the maxTSN and with allelic difference in pause shape were shown (KS test, fdr <= 0.1; also requiring a distinct allelic maximal pause). Color indicates the organs from which the TSN-pause relationship was obtained.

(D) Top: scatterplot shows the average SNPs per base around the position of the Pol II in which the distance between the maxTSN and the max pause was lowest (short pause). Red represents sites with allelic difference in pause shape (Allelic pause difference, KS test, fdr <= 0.1 with distinct allelic maxPause, n = 269), Blue is the control group (No allelic pause difference, KS test, fdr > 0.9 and the allelic maxPause were identical, n = 1,396). Bottom: The sequence logo obtained from the maxPause position based on all reads (n = 3,456 max pause sites). Sites were combined from three organs, after removing pause sites that were identical between organs.

(E) Violin plots show the G content, GC content and C content as a function of position relative to maxPause defined using all reads (combined from three organs with duplicate pause sites removed, n = 3,456), block 1 was 11 to 20 nt upstream of maxPause, block 2 was 1 to 10 nt upstream of maxPause, and block 3 was 1 to 10 nt downstream of maxPause (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$). Block 2 had a higher G content and a lower C content than the two surrounding blocks.

(F) Pie charts show the proportion of different events around the maxPause of short alleles (short pause) with or without allelic pause differences. Sites were combined from three organs with duplicated pause sites removed.

(G) Sequence logos show the sequence content of short alleles and long alleles at 270 short pause sites and 278 long pause sites.

(H) The histograms show the fraction of pause sites as a function of distance between allelic maxPause, i.e. the distance between the short and long pause. The lines show the cumulative density function. Blue represents pause sites with allelic differences (n = 285); red is a subgroup of blue sites with a C to A/T/G SNP at the maxpause (n = 34). Two-sample Kolmogorov-Smirnov test p-value = 0.002694

(I) The histograms show the fraction of pause sites as a function of distance between allelic maxPause. The lines show the cumulative density function. Blue is pause sites with allelic differences (n = 285), green is a subgroup of blue sites that contain indels between initiation and long pause sites (n=60), Two-sample Kolmogorov-Smirnov test p-value = 0.02755.

2.3.6 Correspondence and disconnect between allele specific TSN and pause position

We next examined allelic changes in the position of paused Pol II. To measure the position of the Pol II active site with single nucleotide precision, we prepared new ChRO-seq libraries in three organs (heart, skeletal muscle, and kidney from two female mice). New libraries were paired-end sequenced to identify the transcription start site and active site of the same molecule (Tome et al., 2018). New libraries clustered with those generated previously from the same organ (**Supplementary Fig 4**). Using the same pipeline we developed for transcription initiation, we identified regions enriched for transcription initiation and pausing, and validated that candidate maxTSNs were enriched for the initiator motif (**Supplementary Figure 2.5A**). Our analysis identified 2,260 dREG sites with candidate changes in the shape of paused Pol II, assessed using the position of the Pol II active site, defined as the 3' end of RNA insert (FDR <= 0.1; Kolmogorov-Smirnov [KS] test; see **Methods**).

Previous work has shown a tight correspondence between the site of transcription initiation and pausing, with pausing occurring predominantly in the window 20-60 bp downstream of the TSN (Tome et al., 2018). As expected, allelic changes in the position of paused Pol II were often coincident with changes in transcription initiation (**Figure 2.5A**), particularly when the changes were large. In addition to the main component of correlation between initiation and pausing, however, we also identified changes in both pause and initiation that were independent of the other step in the transcription cycle. In at least 111 cases (~31% of sites where paused Pol II changed shape, and the single RNA molecule was tagged by a SNP), the position

of the pause was identical between alleles, but the position of the max TSN that initiated the paused Pol II changed by 1-32 bp (**Figure 2.5B, top**). Thus, in many cases where Pol II paused at the same position in both alleles, the RNA molecules were initiated at distinct TSNs.

More commonly (n = 269; ~52% of tagged RNA molecules with putative changes in pause shape), changes in Pol II pausing occurred between alleles despite being initiated from identical TSNs (**Figure 2.5B, bottom**). Although more frequent, changes in the position of paused Pol II that shared the same TSN were slightly smaller in magnitude, typically <10 bp. These cases represent changes in the Pol II pause site without changes in the position of transcription initiation.

2.3.7 DNA sequence determinants of promoter proximal pause position

To understand the genetic determinants of pausing, we analyzed the cases in which the same TSN had different maximal pause sites in the CAST and B6 alleles. As tagged SNPs in the window between the initiation and pause site were relatively rare, we increased our statistical power by analyzing all three of the organs together after removing duplicate initiation sites when they overlapped. After filtering, we identified 269 candidate positions in which the same TSN gave rise to separate pause distributions on the B6 and CAST alleles. As a control, we used 1,396 TSN/ pause pairs that were tagged by SNPs but did not have allelic changes in the pause position.

We first examined how short insertions or deletions affect the position of paused Pol II. Paused Pol II is positioned in part through physical constraints with TFIID, a core component of the pre-initiation complex (Fant et al., 2020). As a result of such

connections with the pre-initiation complex, short insertions or deletions affecting the distance between the transcription start nucleoside and the maximal pause site altered the frequency of pausing at a model gene (D. melanogaster *HSP70*; (Kwak et al., 2013)). In our dataset, changes in pausing were highly enriched for small insertions and deletions between the maxTSN and pause site (n = 56 (21%); expected = 22 (8%); $p <$ 1e-5, Fisher's exact test). Changes in pause position were correlated with changes in the size of the insertion or deletion, such that the distance between the max TSN and the pause site in the native genome coordinates was typically less than 5 bp (**Figure 2.5C**). This finding supports a model in which paused Pol II is placed in part through physical constraints with the pre-initiation complex (Fant et al., 2020; Kwak et al., 2013).

Next, we identified single nucleotide changes that affect the position of the pause site. Previous studies have defined a C nucleotide at the paused Pol II active site (Gressel et al., 2017; Tome et al., 2018), which we recovered by generating sequence logos of max pause positions in our three murine organs (**Figure 2.5D, bottom**). Additionally, we also observed a G in the +1 position immediately after the pause, and a G/A-rich stretch in the 10 bp upstream of the pause site that lies within the transcription bubble (**Figure 2.5D, bottom**). The 10 bp upstream of the pause position had a higher G content and a lower C content than the two surrounding windows (**Figure 2.5E**). Thus, our data show that Pol II pauses on the C position immediately after a G-rich stretch.

Allelic differences at the C base had the strongest association with the pause, followed by the -2 and -3 G bases (**Figure 2.5D; Supplementary Figure 2.5B**). We also noted enrichment of SNPs downstream of the pause, especially in the +1 position, although the number of SNPs supporting these positions were small. We also noted that

multiple independent SNPs were frequently found in the same TSN/ pause pair (observed n = 10 (3.7%); expected = 1 (0.36%); $p < 2e\text{-}5$; Fisher's Exact Test; **Figure 2.5F**), suggesting that multiple changes in the weak DNA sequence motifs associated with pausing were more likely to affect the position of paused Pol II. Collectively, indels and SNPs identified as enriched in the analysis above explained 49% of allele specific differences in the pause position (**Figure 2.5F**). Thus, the DNA sequence determinants of pause position are largely found either within the pause site, the transcription bubble of paused Pol II, or insertions or deletions between the pause and initiation site.

2.3.8 Pol II pause position is driven by the first energetically favorable pause site

We extended our analysis of allelic differences in pausing to determine how multiple candidate pause positions early in a transcription unit (TU) collectively influence the position of paused Pol II. As in the analysis above, we focused on the set of allelic differences in pause shape in which the two alleles have a distinct maximal pause position (n = 269). By definition, these TUs had different maximal pause positions on the two alleles: on one allele the distance between the TSN and the pause position is shorter (which we call the 'short allele'), and on the other the distance between the TSN and the pause is longer ('long allele') (see cartoon in **Figure 2.5G, middle**). We found that the DNA sequence motif near the long pause position was similar on both alleles, recapitulating the C at the pause site and an enrichment of Gs in the transcription bubble (**Figure 2.5G, right**). By contrast, DNA sequence changes affecting pause position occurred at the short pause position on the long allele (**Figure 2.5G; bottom left**).

Our findings suggest a model in which single nucleotide changes that alter the free energy of the pause complex result in Pol II slipping to the next available position downstream. In favor of this model, when there was a SNP in the active site at the short pause, the max pause position moved downstream by <10bp, a relatively small amount compared to all changes in pause shape (**Figure 2.5H**). By contrast, indels between the initiation and short pause site tended to have a larger effect on the allelic difference between pause positions (**Figure 2.5I**). These observations support a model in which DNA sequence changes that disfavor pausing result in Pol II slipping downstream to the next pause site for which DNA sequence is energetically favorable, but maintaining physical connections that may exist between paused Pol II and the PIC.

2.3.9 Allelic changes in gene length caused by genetic differences in Pol II termination

AlleleHMM blocks (Chou and Danko, 2019), allelic biased blocks identified by AlleleHMM, were frequently found near the 3' end of genes, possibly reflecting allelic differences in Pol II termination that alter the length of primary transcription units. For example, *Fam207a* had an excess of reads on the CAST allele without a new initiation site that could explain allelic differences (**Figure 2.6A**). To systematically identify allelic differences in the termination site, we identified AlleleHMM blocks that start inside of a protein coding transcription unit and end near or after the end of the same transcription unit (see Methods). Several lines of evidence suggest that these candidate allelic differences were enriched for bona-fide termination differences: the majority did

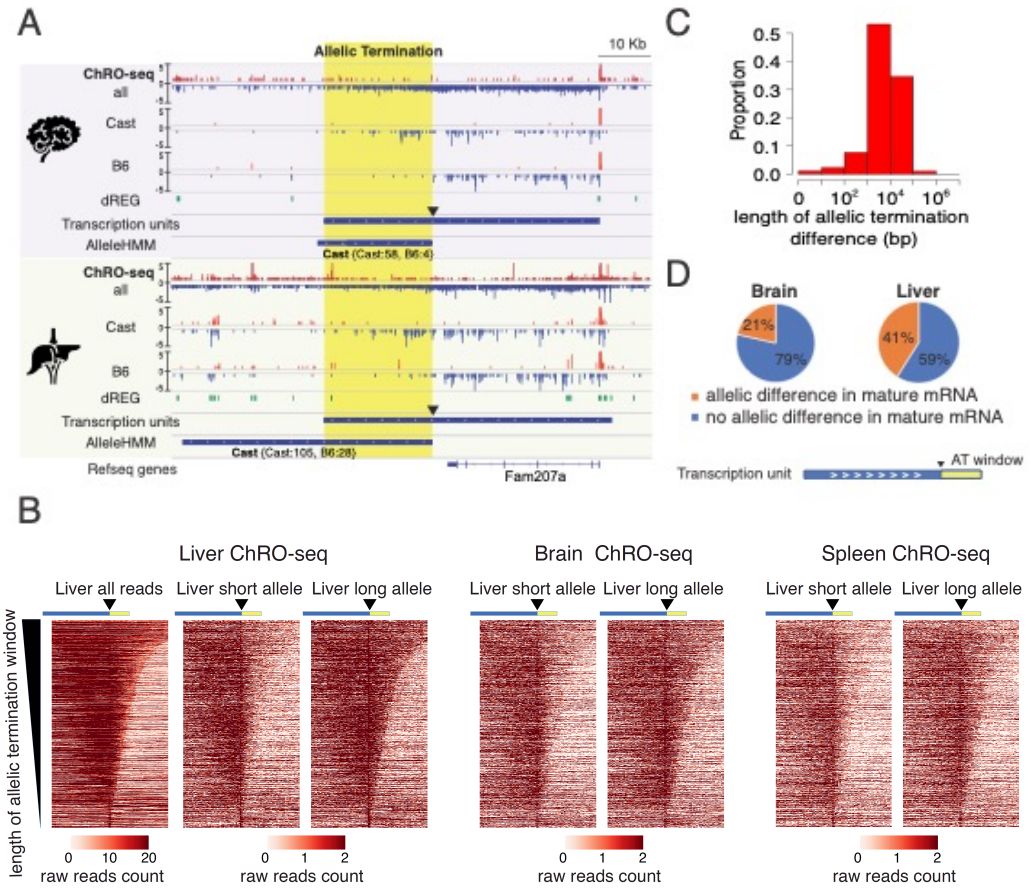


Figure 2.6: Widespread allele specific differences in the Pol II termination site.

(A) The browser shot shows an example of allelic termination differences (yellow shade) in both brain and liver. Pol II terminates earlier on the B6 allele, resulting in a longer transcription unit on the CAST allele. The difference in allelic read abundance was identified by AlleleHMM. We defined the allelic termination difference (yellow shade) using the intersection between the transcription unit and AlleleHMM blocks.

(B) Heatmaps show the raw read counts in transcription units (blue bar) with an allelic termination difference (yellow bar), centered at the beginning of allelic termination (solid triangle). The heatmap bin size is 500 bp, and 20kb is shown upstream and downstream. The rows were sorted by the length of allelic termination differences determined by ChROseq signals from Liver. The short and long alleles were determined based on analysis of the liver.

(C) The histogram shows the fraction of transcription units as a function of the length of allelic termination difference .

(D) Pie charts show the proportion of transcription units with allelic termination difference that also contains allelic difference in mature mRNA (orange).

not start near dREG sites, and overall the allele with higher expression tended to have similar ChRO-seq signal as the primary gene (**Supplementary Figure 2.6A**).

To visualize ChRO-seq signal surrounding the window of allelic termination (AT window), we generated heat maps of ChRO-seq signal centered near the start of the allelic increase in expression and sorted by the length of the allelic termination window (**Figure 2.6B**). Heatmaps showed a higher abundance of ChRO-seq reads in the allelic termination window on the allele with higher expression. Overall, we identified 317-931 candidates in each organ (total n = 3,450). These allelic termination differences varied in size between 1kb and 100kb (**Figure 2.6C**; **Supplementary Figure 2.6B**).

Allelic differences in termination were consistent between different organs. An outstanding example is *Fam207a*, which has approximately the same allelic termination window in brain and liver (**Figure 2.6A**). Furthermore, taking the subset of genes that were expressed and producing heat maps centered in the same position and the same order as in liver recovered similar patterns of allelic termination being evident in brain, spleen, or other organs (**Figure 2.6B**). This result suggests that allelic differences in termination were driven predominantly by DNA sequence, with little effect from the trans environment.

Allelic differences in termination could be caused by factors that influence the mature mRNA, such as differences in the polyadenylation cleavage site (Mittleman et al., 2021, 2020). To determine whether allelic differences in termination correlate with differences in the composition of the mature mRNA, we sequenced poly-A enriched mRNA from two liver and brain samples. Genes that had allelic differences in

termination frequently also had an AlleleHMM block identified in the mRNA-seq data (21-41% of cases), potentially consistent with allelic differences in termination driving differences in the mature mRNA composition (**Figure 2.6D**). In the majority of cases (59-79%), we could not find any evidence that changes in allelic termination of the primary transcript affected the mature mRNA composition. This indicates that many of the changes in Pol II termination do not affect the mature mRNA.

We examined whether changes in the mature mRNA composition were enriched for changes in mRNA stability. We estimated mRNA stability using the ratio of mRNA-seq to ChRO-seq signal (Blumberg et al., 2021). Our analysis did not find evidence that allelic changes in termination affecting the mature mRNA composition had a consistent effect on mRNA stability (**Supplementary Figure 2.6C**), although we are likely underpowered to detect global changes in this analysis. Nevertheless, our data allowed us to conclude that allelic differences in Pol II termination are widespread and frequently do not affect mRNA composition or stability.

2.4 Discussion

We examined how DNA sequence affects the precise, nucleotide position of Pol II during each stage of the transcription cycle. We generated and analyzed new ChRO-seq libraries which map the location and orientation of RNA polymerase in eight murine organs. Our study used an F1 hybrid cross between CAST and B6, which we contend is particularly well suited to this problem because experimental batch variation is identical between alleles. As a result, we can be confident in differences observed in our analysis, even in cases where the differences between the CAST and B6 alleles were relatively small in magnitude. Using this system, our analysis provides new insight into how DNA sequences shape the precise position and dynamics of Pol II. Here we discuss how our findings impact the prevailing models governing each stage in the Pol II transcription cycle.

DNA sequence changes within ~5-10bp of the transcription initiation site frequently determine the relative use of that initiation site. The A base at the transcription start site itself stood out as the most important determinant of transcription initiation. Changes in the cytosine nucleotide at the -1 position had surprisingly little evidence for differences in our analysis despite being as informative about initiation in the initiator sequence motif, though this may be impacted by the A at position 0 tagging the nascent RNA whereas the C does not. We also noted an enrichment of AT nucleotides surrounding the initiation site on the allele with high TSN usage. We speculate that enrichment of a relatively high AT composition may influence the initiation site because AT-rich DNA requires lower free energy to melt (Breslauer et al.,

1986). Thus, the DNA sequence composition of the initiation site itself, as well as the sequence composition of surrounding nucleotides, impact transcription initiation.

Our results have also advanced models of transcription initiation. In the prevailing model of initiation in yeast, DNA is melted and Pol II scans by forward translocation until it identifies an energetically favorable TSN (Braberg et al., 2013; Kaplan et al., 2012; Qiu et al., 2020). Mammals are not believed to scan, but rather independent PICs are believed to support initiation from a very narrow window (Luse et al., 2020). Intriguingly, DNA sequence changes in the initiator element affect the use of nearby initiation sites both upstream and downstream. We think the most straightforward interpretation of these results is that Pol II samples candidate initiation sites in both directions by a process resembling a one-dimensional brownian motion along the DNA. We note that an important assumption of our analysis is that the change in initiator sequence does not feed back and affect the DNA binding position of other components of the PIC. An alternative model that we are not able to discount completely is that feedback between transcription initiation and the binding site of other PIC components changes the site of DNA melting and leads to the PIC being assembled in nearby positions. We believe this interpretation is unlikely. Notably, the effect we report here appears confined to ~20 bp of DNA surrounding the affected initiation site. It seems likely to us that any feedback would affect all TSNs within the TID, but this possibility does not appear to be supported by our analysis. In either case, however, it is clear from our data that changes in the DNA sequence of initiator elements tend to increase the use of candidate initiators nearby.

By analyzing allelic changes in pause position from the same initiation site, we have learned much about how DNA sequence influences the precise coordinates of promoter proximal pause. Our results show that the pause position occurs at a C nucleotide downstream of a G-rich stretch, similar to motifs enriched at pause positions in prior studies (Gressel et al., 2017; Tome et al., 2018). As cytosine is the least abundant ribonucleotide (Traut, 1994), previous authors proposed it is the slowest to incorporate into nascent RNA, explaining its association with the Pol II pause (Tome et al., 2018). In addition, we also identified a G-rich stretch that coincides with the position of the transcription bubble, as well as a guanine nucleotide just downstream of the pause position. The enrichment of G nucleotides in the transcription bubble has a higher stability of RNA-DNA hybridization without using a cytosine nucleotide, and may serve to stabilize the RNA-DNA hybrid within the transcription bubble while Pol II remains paused.

We also noted widespread allelic differences in the site of transcription termination, which resulted in substantial differences in the length of primary transcription units between alleles. Allelic differences in transcription termination were largely similar between different organs, implicating DNA sequence differences as the major determinants of transcription termination. The majority (60-80%) of allelic differences in termination did not affect the sequence composition of the mature mRNA, and when they did, we found no evidence for systematic differences in RNA stability. This may be influenced by purifying selection acting to remove many of the genetic variants with a large effect on mRNA.

In summary, our study dissects how DNA sequences impact steps during the Pol II transcription cycle. The dynamics of Pol II on each position of the genome can influence the rate of mRNA production and may impact organismal phenotypes. Our work is a first step in understanding the link between the core steps in transcriptional regulation and the impact they may have on phenotypes in humans and other animals.

2.5 Materials and Methods

2.5.1 Experimental Methods:

Mouse experiments: The mice used in this study were reciprocal F1 hybrids of the strains C57BL/6J and CAST/EiJ. All mice were bred at Cornell University from founders acquired from the Jackson Laboratory. All mice were housed under strictly controlled conditions of temperature and light:day cycles, with food and water *ad libitum*. All mouse studies were conducted with prior approval by the Cornell Institutional Animal Care and Use Committee, under protocol 2004-0063.

Tissue collection: Mice were euthanized at 22 to 25 days of age by CO_2, followed by cervical dislocation. All mice were euthanized between 10 a.m. and 12 p.m., immediately after removal from their home cage. Whole brain, eye, liver, stomach, large intestine, heart, skeletal muscle, kidney, and spleen were rapidly dissected and snap frozen in dry ice.

*mRNA isolation/ RNA-seq library prep:*RNA was extracted from the brain and liver of a male and a female mouse (both 22d of age). Tissue samples were frozen using liquid nitrogen and pulverized using a mallet and a mortar. 100mg of each tissue was used for a TRIzol RNA extraction. Briefly, 1 mL of TRizol was added to each sample, chloroform was used for phase separation of the aqueous phase containing RNA, RNA was precipitated using isopropanol and washed with 75% ethanol. A total of 400 ng of RNA was input into the RNA-seq library prep. Poly-A containing mRNA was enriched

for 2 rounds using the NEBNext Poly(A) mRNA Magnetic Isolation Module. Stranded mRNA-seq libraries were prepared by the Cornell TREx facility using the NEBNext Ultra II Directional RNA Library Prep Kit. Libraries were sequenced using an Illumina NextSeq500.

Chromatin isolation: Chromatin was isolated and ChRO-seq libraries were prepared following the methods introduced in our recent publication (Chu et al., 2018). Briefly, tissue was cryo-pulverized using a cell crusher (http://cellcrusher.com). Tissue fragments were resuspended in NUN buffer (0.3 M NaCl, 1 M Urea, 1% NP-40, 20 mM HEPES, pH 7.5, 7.5 mM MgCl2, 0.2 mM EDTA, 1 mM DTT, 20 units per ml SUPERase In Rnase Inhibitor(Life Technologies, AM2694), 1× Protease Inhibitor Cocktail (Roche, 11 873 580 001)).Samples were vortexed vigorously before the samples were centrifuged at 12,500 x g for 30 min at 4C. The NUN buffer was removed and the chromatin pellet washed with 1 mL 50 mM Tris-HCl, pH 7.5 supplemented with 40 units of RNase inhibitor. Samples were centrifuged at 10,000 x g for 5 minutes at 4C and the supernatant discarded. Chromatin pellets were resuspended in storage buffer (50 mM Tris-HCl, pH 8.0, 25% glycerol, 5 mM Mg(CH3COO)2, 0.1 mM EDTA, 5 mM DTT, 40 units per ml Rnase inhibitor) using a Bioruptor sonicator. The Bioruptor was used following instructions from the manufacturer, with the power set to high, a cycle time of 10 min (30s on and 30s off). The sonication was repeated up to three times if necessary to resuspend the chromatin pellet. Samples were stored at -80C.

ChRO-seq library preparation: ChRO-seq libraries were prepared following a recent protocol (Dig Bijay Mahat et al., 2016). We prepared some libraries to achieve single nucleotide resolution for the Pol II active site. In these cases, the chromatin pellet was incubated with 2x run-on buffer (10 mM Tris-HCl, pH 8.0, 5 mM MgCl2,1 mM DTT, 300 uM KCl, 20 uM Biotin-11-ATP (Perkin Elmer, NEL544001EA), 200 uM Biotin-11-CTP (Perkin Elmer, NEL542001EA), 20 uM Biotin-11-GTP (Perkin Elmer, NEL545001EA), 200 uM Biotin-11-UTP (Perkin Elmer, NEL543001EA)) for 5 minutes at 37 C. In some libraries we modified the run-on buffer to extend the length of reads for more accurate allelic mapping at the expense of single nucleotide resolution for the Pol II active site. In these cases, the run-on reaction was performed using a different ribonucleotide composition in the nuclear run-on buffer (10 mM Tris-HCl, pH 8.0, 5 mM MgCl2,1 mM DTT, 300 mM KCl, 200 μ M ATP (New England Biolabs (NEB), N0450S), 200 μ M UTP, 0.4 μ M CTP, 20 μ M Biotin-11-CTP (Perkin Elmer, NEL542001EA), 200 μ M GTP (NEB, N0450S)).The run-on reaction was stopped by adding Trizol LS (Life Technologies, 10296-010) and RNA was pelleted with the addition of GlycoBlue (Ambion, AM9515) to visualize the RNA. RNA pellet was resuspended in diethylpyrocarbonate (DEPC)-treated water. RNA was heat denatured at 65 C for 40 s to remove secondary structure. RNA was fragmented using base hydrolysis (0.2N NaOH on ice for 4 min). RNA was purified using streptavidin beads (NEB, S1421S) and removed from beads using Trizol (Life Technologies, 15596-026). We ligated a 3' adapter ligation using T4 RNA Ligase 1 (NEB, M0204L). We performed a second bead binding followed by a 5' decapping with RppH (NEB, M0356S). RNA was phosphorylated on the 5' end using T4 polynucleotide kinase (NEB, M0201L) then

ligated onto a 5' adapter. A third bead binding was then performed. The RNA was then reverse transcribed using Superscript III Reverse Transcriptase (Life Technologies, 18080-044) and amplified using Q5 High-Fidelity DNA Polymerase (NEB, M0491L) to generate the ChRO-seq libraries. Libraries were sequenced using an Illumina HiSeq by Novogene. The adapter sequences used are depicted in **Supplementary Figure 2.7-8**.

2.5.2 Data analysis:

Read mapping, transcription start site, and transcription unit discovery:

Processing and mapping ChRO-seq reads: Paired-end reads with single nucleotide precision were processed and aligned to the reference genome (mm10) with the proseq2.0 (https://github.com/Danko-Lab/proseq2.0). Libraries in which we tailored the run-on to extend the length of reads were pre-processed, demultiplexed, and aligned to the reference genome (mm10) with the proseqHT_multiple_adapters_sequencial.bsh. AlleleDB (Chen et al., 2016; Rozowsky et al., 2011) align the R1 reads to the individual B6 and Cast genomes. In brief, the adaptor sequences were trimmed with the cutadapt, then PCR duplicates were removed using unique molecular identifiers (UMIs) in the sequencing adapters with prinseq-lite.pl (Schmieder and Edwards, 2011). The processed reads were then aligned with BWA (mm10) in analyses not using individual genome sequences (Li and Durbin, 2009), or with bowtie (Langmead et al., 2009) as input for AlleleDB. When bowtie was used, we selected either the R1 or R2 files for alignment for analyses requiring either the 5' or 3' end of the RNA insert. All scripts for mapping can be found publicly at:

https://github.com/Danko-Lab/F1_8Organs/blob/main/00_F1_Tissues_proseq_pipeline.bash

https://github.com/Danko-Lab/utils/blob/master/proseq_HT/proseqHT_multiple_adapters_sequencial.bsh

Processing and mapping RNA-seq reads: We used STAR (Dobin et al., 2013) to align the RNA-seq reads. To avoid bias toward the B6 genome, we did not use any gene annotations for mapping, but used the list of splicing junctions generated by STAR. Mapping was performed in three stages: First, reads were first mapped without annotation and STAR generated a list of splicing junctions (sj1) from the data. Second, to identify potential allele specific splicing junctions, we performed allele specific mapping using STAR which takes as input a VCF file denoting SNPs differentiating Cast and B6, using the initial splice junction list (sj1). This personalized mapping was used to generate a more complete list of splice junctions (sj2). Third, we identified allele specific alignments by using the WASP option provided by STAR (van de Geijn et al., 2015). In this final mapping, we used the splice junction list (sj2) and a VCF file. This procedure generated a tagged SAM file (vW tag) providing the coordinates of allele specific alignments and their mapping position. Scripts can be found here: https://github.com/Danko-Lab/F1_8Organs/blob/main/termination/F1_RNAseq_forManuscript.sh

dREG: For each organ, we merged all reads from each replicate and cross to increase the power of dREG. BigWig files representing mapping coordinates to the mm10 reference genome were uploaded to the dREG web server at http://dreg.dnasequence.org (Wang et al., 2018). All of the output files were downloaded and used in subsequent data analysis. Scripts used to generate the BigWig files can be found at:

https://github.com/Danko-Lab/F1_8Organs/blob/main/F1_TSN_Generate_BigWig.sh

Transcript unit prediction using tunits: We used the tunit software to predict the boundaries of transcription units *de novo* (Danko et al., 2018). We used the 5 state hidden Markov model (HMM), representing background, initiation, pause, body, and after polyadenylation cleavage site decay from tunits. To improve sensitivity for transcription unit discovery in each tissue, the input to tunits was the output of dREG and bigWig files that were merged across all replicates and crosses. Scripts can be found here:

https://github.com/Danko-Lab/F1_8Organs/blob/main/Tunit_predict_manuscript.sh

https://github.com/Danko-Lab/F1_8Organs/blob/main/run.hmm.h5_F1bedgraph.R

https://github.com/Danko-Lab/F1_8Organs/blob/main/hmm.prototypes.R

Clustering: We used all transcripts that are 10,000 bp long from GENCODE vM25. Only reads mapped to the gene body (500bp downstream of the start of the annotation to the end of the annotation) were used. We filtered the transcripts and only kept those with at least 5 mapped reads in every sample. We export rpkm normalized expression

estimates of each transcript. Morpheus was used to calculate and plot Spearman's rank correlation (https://software.broadinstitute.org/morpheus) with the following parameters: Metric = Spearman rank correlation, Linkage method = Average Linkage, distance.function.name= Spearman rank correlation. Scripts can be found here: https://github.com/Danko-Lab/F1_8Organs/blob/main/getCounts_skipfirst500.R

AlleleHMM: Maternal- and paternal- specific reads mapped using AlleleDB were used as input to AlleleHMM (Chou and Danko, 2019). We combined biological replicates from the same organ and cross, and used the allele-specific read counts as input to AlleleHMM. AlleleHMM blocks were compared with GENOCODE gene annotations to pick the free parameter, τ, which maximized sensitivity and specificity for computing entire gene annotations, as described (Chou and Danko, 2019). Most organs used a τ of either 1E-5 (brain, liver, spleen, skeletal muscle) or 1E-4 (heart, large intestine, kidney, and stomach). As reported, the primary parameter that influenced τ was the library sequencing depth. AlleleHMM scripts can be found here:
https://github.com/Danko-Lab/AlleleHMM
https://github.com/Danko-Lab/F1_8Organs/blob/main/01_F1Ts_AlleleHMM.bsh

Discovering strain effect and imprinted domains: We used the following rules to merge nearby allele specific transcription events into starin effect or imprinted domains:

1. Identify candidate AlleleHMM blocks using pooled ChRO-seq reads from samples with the same organ and same cross direction.

2. Combine blocks above from the same organ (but different crosses). Combine p-values using Fisher's method for all biological replicates within the same tissue and cross direction. Keep blocks that are biased in the same direction with a Fisher's p-value <= 0.05.
3. Determine whether the blocks are under a strain effect (allelic biased to the same strain in reciprocal crosses) or parent-of-origin imprinted effect (allelic biased to the same parent in reciprocal crosses).
4. Merge overlapping strain effect blocks from different organs into strain effect domains; merge overlapping strain effects from the same imprinted blocks into imprinted domains.

Scripts implementing these rules can be found here:
https://github.com/Danko-Lab/F1_8Organs/blob/main/Find_consistent_blocks_v3.bsh

After discovering blocks, we examined the number of gene annotations in each domain (Figure 2.1E). We used GENCODE annotated genes (vM25). We kept all gene annotations and merged those which overlapped or bookended (directly adjacent to, as defined by bedTools) on the same strand so that they were counted once. All operations were performed using bedTools (Quinlan and Hall, 2010). Scripts can be found here:
https://github.com/Danko-Lab/F1_8Organs/blob/main/Find_consistent_blocks_v3.bsh
https://github.com/Danko-Lab/F1_8Organs/blob/main/Imprinted_figures.R

Determining the allelic bias state of annotated genes: We used GENCODE gene annotations representing protein-coding genes (vM20) in which the transcription start

site overlapped a site identified using dREG (Wang et al., 2018). We used de novo annotations by the *tunits* package to identify unannotated transcription units, which do not overlap an annotated, active gene as a source of candidate transcribed non-coding RNAs. Transcription units from both sources were merged for downstream analysis. We determine if the gene/ncRNA are allelic biased by comparing mapped reads to the B6 and CAST genomes using a binomial test, retaining transcription units with a 10% false discovery rate (FDR). Allele specific transcription units were classified as being under a strain effect (allelic biased to the same strain in reciprocal crosses) or parent-of-origin imprinted effect (allelic biased to the same parent in reciprocal crosses). Scripts can be found:

https://github.com/Danko-Lab/F1_8Organs/blob/main/Genetics_or_imprinting_v2.bsh

Evaluate the contribution of false negatives to organ-specific allelic bias in organ-specific allelic biased domains (OSAB domain): In Supplementary Figure 2.1A and B, we asked whether organs in which we did not identify allelic bias were false negatives. To do this we compared distributions of the transcription level in putatively unbiased organs. For each OSAB domain identified in at least one, but not in all organs, we examined the effect size of allelic bias in the organ with the highest expression that is putatively unbiased. We defined the effect size of allelic bias as the ratio between maternal and paternal reads in the candidate OSAB domain. If the allelic-biased organ was maternally biased, the effect size was calculated as maternal reads divided by paternal reads in the blocks, otherwise the effect size was calculated as paternal reads divided by maternal reads in the blocks. Scripts implementing this can be found here:

https://github.com/Danko-Lab/F1_8Organs/blob/main/AllelicBiase_expressionLevel.bsh

https://github.com/Danko-Lab/F1_8Organs/blob/main/getNonBiasedHighest_Biased_AllelicBiaseDistribution.R

Evaluate the contribution of expression to organ-specific allelic biased domains (OSAB domain): In Supplementary Figure 2.1C, we asked whether OSAB domains were not actively transcribed in candidate unbiased organs. Using bedtools and in-house scripts, we calculated the rpkm (Reads per kilobase per million mapped reads) normalized transcription level of each strain effect block located within the OSAB domains in each organ. The full diploid genome was used for mapping. The non-allelic-biased organs with highest rpkm (nonBiasedH) were selected to compare with the rpkm of the allelic-biased organs in OSAB domains. Scripts implementing this can be found here:

https://github.com/Danko-Lab/F1_8Organs/blob/main/AllelicBiase_expressionLevel.bsh

https://github.com/Danko-Lab/F1_8Organs/blob/main/getNonBiasedHighest_Biased_TotalReadCountRatio.R

Analysis of allele-specific initiation:

Identification of candidate transcription initiation sites: We used 5 prime end of ChRO-seq reads (the R1 paired-end sequencing file, which represents the 5 prime end of the nascent RNA) to identify candidate initiation sites using methods adapted from (Tome

et al., 2018). Briefly, candidate transcription start nucleotides (TSN) from each read were merged into candidate transcription start sites, in which the max TSN was identified. We identified candidate TSNs that fall within dREG sites and were supported by at least 5 separate reads. TSNs within 60bp of each other were merged into candidate TSSs. The TSN with the maximal read depth in each TSS was defined as the maxTSN for that TSS. We allow each TSS to have more than one maxTSNs if multiple TSNs share the same number of read counts in that TSS. To test whether the candidate maxTSNs represented bona-fide transcription start sites, we generated sequence logos centered on the maxTSN using the seqLogo R package (Bembom, 2019). We retained tissues in which the maxTSN contained a clearly defined initiator dinucleotide that reflects a similar sequence composition as those previously reported (Tome et al., 2018). Additionally, we used an in-house R script to examine the relationship between TSN counts and Read counts of the TSS (Supp Fig2 B), and found a similar relationship to those reported (Tome et al., 2018).

Identify allele specific differences in TSSs abundance (ASTSS abundance): We used a binomial test to identify candidate allele specific transcription start sites, with an expected allelic ratio of 0.5. We filtered candidate allele specific differences using a false discovery rate (FDR) corrected p-value of 0.1, corresponding to an expected 10% FDR.

Identify allele specific differences in TSS shape (ASTSS shape): We used a Kolmogorov-Smirnov (K-S) test to identify TSSs where the distribution of transcription initiation

differed significantly between the B6 and CAST alleles (ASTSS shape). We used TSS sites with at least 5 mapped reads specific to the B6 genome and at least 5 mapped reads specific to CAST. Only autosomes were used. We corrected for multiple hypothesis testing using the false discovery rate (Storey and Tibshirani, 2003) and filtered ASTSS shapes using a 10% FDR. We further separated the ASTSS shape candidates into two groups: one driven by a single base (single base driven ASTSS shape), the other reflecting changes in more than one base in the TSS (multiple base driven ASTSS shape). To separate into two groups, we masked the TSN with the highest allelic difference (determined by read counts) within each TSS and performed a second K-S test. Multiple base driven ASTSS were defined as those which remained significantly different by K-S test after masking the position of highest allelic difference. Single base driven ASTSSs were defined as ASTSSs that were no longer significantly different by K-S test after masking the maximal position. In the second K-S test, we used the nominal p-value defined as the highest nominal p-value that achieved a 10% FDR during the first K-S test.

SNP analysis: We examined the distribution of single nucleotide polymorphisms (SNPs) near ASTSSs from each class. A major confounding factor in SNP distribution is the ascertainment bias of requiring at least one tagged SNP to define the allelic imbalance between the two alleles, resulting in an enrichment of SNPs within the read. To control for this bias, we compared the set of sites with a significant change in the TSS shape or abundance (FDR <= 0.1) with a background control set defined as candidate TSSs in which there was no evidence of change between alleles (FDR > 0.9) in all analyses. We

display a bin size of 5 bp. To test for differences, we merged adjacent bins by using a bin size of 10bp to increase statistical power and tested for enrichment using Fisher's exact test, FDR cutoff = 0.05. (Figure 2.3E,F). We also examined the difference in base composition between the allele with high and low initiation in each ASTSS shape difference centered on the position of the maxTSN in the allele with high initiation (in Figure 2.3G). We determined the high/low allele based on the transcription level at maxTSN. If there are more than one TSNs with the max read counts, there will be more than one maxTSNs representing each TSS.

Comparison of AT consent between alleles: As a proxy for melting temperature (in Figure 2.3H), we examined the AT content in 5 bp windows around the maxTSN on alleles with high and low maxTSN usage. As in the SNP analysis (above), we compared the set of sites with a significant change in the TSS shape or abundance (FDR <= 0.1) with a background control set defined as candidate TSSs in which there was no evidence of change between alleles (FDR > 0.9). Computations were performed using R library TmCalculator (Li, 2019). We used Fisher's exact test to examine if there was an enrichment of AT (in the high allele) to GC (in the low allele) SNPs in each 5 bp bin, and adopted an FDR corrected p-value cutoff = 0.05. In all analyses, positions at -1 and 0 relative to the maxTSN were masked to avoid confounding effects of the initiator sequence motif on computed AT content.

Shooting gallery: In our analysis of the shooting glaray model, we focused on a subset of TSSs which do not appear to change expression globally, and which have a SNP in

the initiator element. Toward this end, we identified TSSs which do not overlap AlleleHMM blocks. We set the allele with high and low expression based on allele specific reads in the maxTSN. Next, we divided data into a test and background control dataset in which the test set had a CA dinucleotide in the allele with high maxTSN use and any other combination except for CA on the other allele. The control set did not have a CA dinucleotide in the maxTSN initiator position. Next we computed the distance to the maxTSN and the allelic read count at other candidate initiator motifs (including CA, CG, TA, TG). In all analyses, we compared the set of maxTSNs with SNPs in the initiator position with the control set which did not have a SNP. Statistical tests used an unpaired Wilcoxon rank sum test. We corrected for multiple hypothesis testing using false discovery rate.

All scripts implementing analysis of allele-specific initiation can be found at: https://github.com/Danko-Lab/F1_8Organs/tree/main/initiation

Analysis of allele-specific pause:

Identification of allele specific differences in pause site shape: All pause analysis focused on ChRO-seq data in three organs (heart, skeletal muscle, and kidney) which used a single base run-on of all four biotin nucleotides. We first focused our analysis on dREG sites in each tissue to identify regions enriched for transcription start and pause sites. We retained dREG sites in which we identified at least 5 reads mapping from both B6 and Cast alleles. We performed a K-S test to identify all candidate dREG sites that

contained a candidate difference in pause, filtering for a false discovery rate of 0.1 (n= 2784). To examine the relationship between initiation and pause, we identified the maxTSN and maxPause on the B6 and Cast allele separately using reads tagged with a SNP or indel. Since maxTSN and maxPause were defined independently, the maxPause was not always correctly paired with the maxTSN (Tome et al., 2018). We therefore used 2260 dREG sites where allelic maxPause were 10 to 50 bp downstream of allelic maxTSN on both alleles. These analyses pertain to Figure 2.5A and B.

Identify genetic determinants of pausing: To focus on the genetic determinants of pausing that were independent of initiation, we identified changes in which the same maxTSN had different allelic maximal pause sites between the Cast and B6 alleles as follows. We used a K-S test to identify maxTSNs with a difference in the maxPause site between alles, filtering for maxTSNs with a different maxPause between alleles and a 10% FDR in a K-S test (n = 269). In most analyses, we also draw a background set in which there was no evidence that sites sharing the same maxTSN had different maxPause sites between alleles, by identifying maxTSNs that have the same maxPause position and a K-S test FDR >0.9 (n = 1396). In all analyses, we also filtered for maxTSNs with at least 5 allelic reads and B6/CAST read ratio between 0.5 and 2.

Comparing GC content near the maxPause position: We compared the G, C, and GC content between alleles. We computed the G, C, and GC content as a function of position relative to maxPause. All of the G, C, and GC contents were combined across unique pause sites from all three organs for which we had single base resolution data (n = 3456).

We compared three blocks: block 1 was 11 to 20 nt upstream of maxPause, block 2 was 1 to 10 nt upstream of maxPause, and block 3 was 1 to 10 nt downstream of maxPause.

All scripts implementing analysis of allele-specific pause can be found at: https://github.com/Danko-Lab/F1_8Organs/tree/main/pause

Analysis of allele-specific termination:

Definition of allelic differences in termination: We noticed frequent AlleleHMM blocks near the 3' end of annotated genes. We used the transcription units (tunits) predictions which overlapped annotated protein coding genes (vM25), as these generally retained the window between the polyadenylation cleavage site and the transcription termination site. We identified transcription units that have AlleleHMM blocks starting within the transcription unit and that end in the final 10% of the transcription unit or after the transcription unit. The overlapping region between the tunits and AlleleHMM blocks were called candidate allelic termination (AT) windows. To avoid obtaining candidate AT windows that reflected entire transcription units, we retain only AT windows whose length was less than or equal to 50% length of the host transcription unit.

RNA-seq analysis: Allele specific mapped RNA-seq reads using STAR (see above) were used as input to AlleleHMM to identify the region showing candidate allelic difference in mature mRNA. The transcription units that contain allelic termination windows, as defined above, were separated into two groups: One has an allelic

difference in mature mRNA and the other does not. Those with an allelic difference in mature mRNA were defined as having the RNA-seq AlleleHMM blocks between 10Kb upstream of the AT windows to the end of the AT windows.

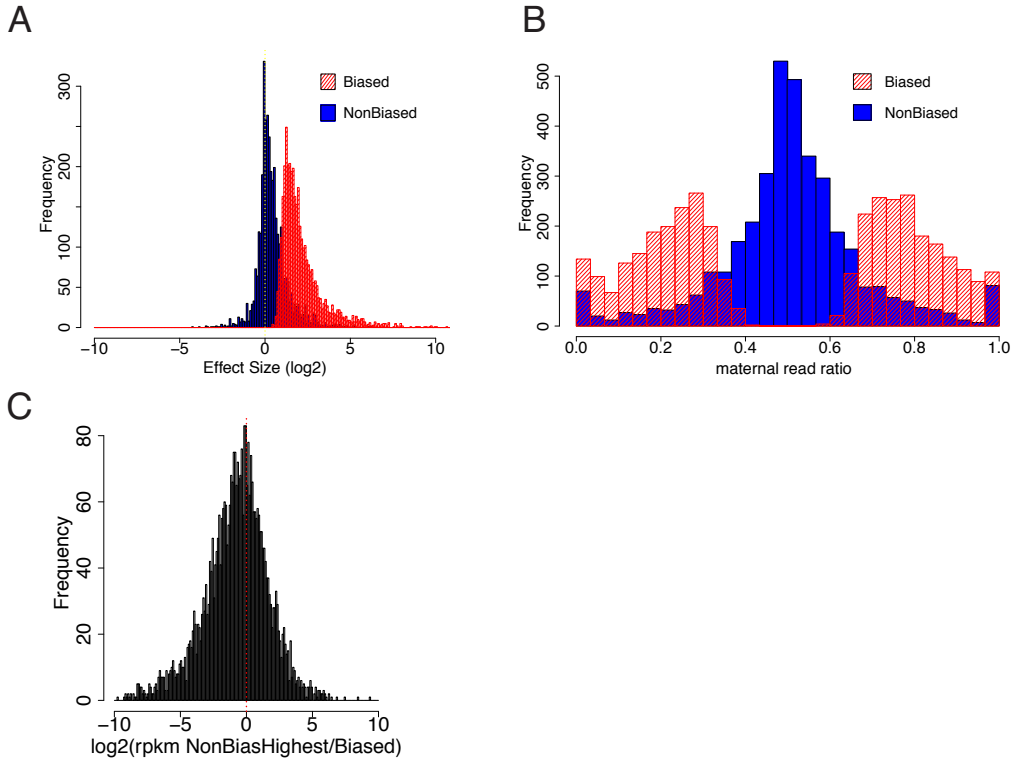
RNA stability analysis: We asked whether there was an allelic difference in mRNA stability between transcription units in which the allelic differences in termination affects the mature mRNA and those in which it does not. The RNA stability was defined as in (Blumberg et al., 2021). The stability was defined as the ratio of RNA-seq read counts in exons to ChRO-seq read counts across the gene body. We used gene annotations from GENCODE (vM25). The RNA-seq reads were counted strand specifically using htseq-count. ChRO-seq reads were counted in a strand-specific fashion using in-house R scripts. After removing the genes with less than 10 B6-specific ChRO-seq and less than 10 CAST-specific ChRO-seq reads, the cumulative distribution functions were drawn. All differences were compared using a one-sided K-S test to compare differences in allelic RNA stability between groups.

All scripts implementing analysis of allele-specific termination can be found at: https://github.com/Danko-Lab/F1_8Organs/tree/main/termination

2.6 Acknowledgments

We thank Maria Garcia-Garcia, Abdullah Ozer, John Lis, Hojoong Kwak, Gilad Barshad, Alexandra Chivu, and all members of the Danko lab for valuable discussions and suggestions throughout the life of this project. We thank Peter Borst for help preparing and working with F1 hybrid mice and Jen Grenier and the Cornell TREx facility for preparing mRNA-seq libraries. We thank C. Kaplan (U. Pittsburgh) for rapid constructive comments based on our *bioRxiv* preprint. Work in this publication was supported by R01-HG010346 and R01-HG009309 (NHGRI) to CGD. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health. Some of the figures in this manuscript were created using BioRender. All data are available at Gene Expression Omnibus under the accession number GSE174171.

2.7 Supplementary Figures and Tables

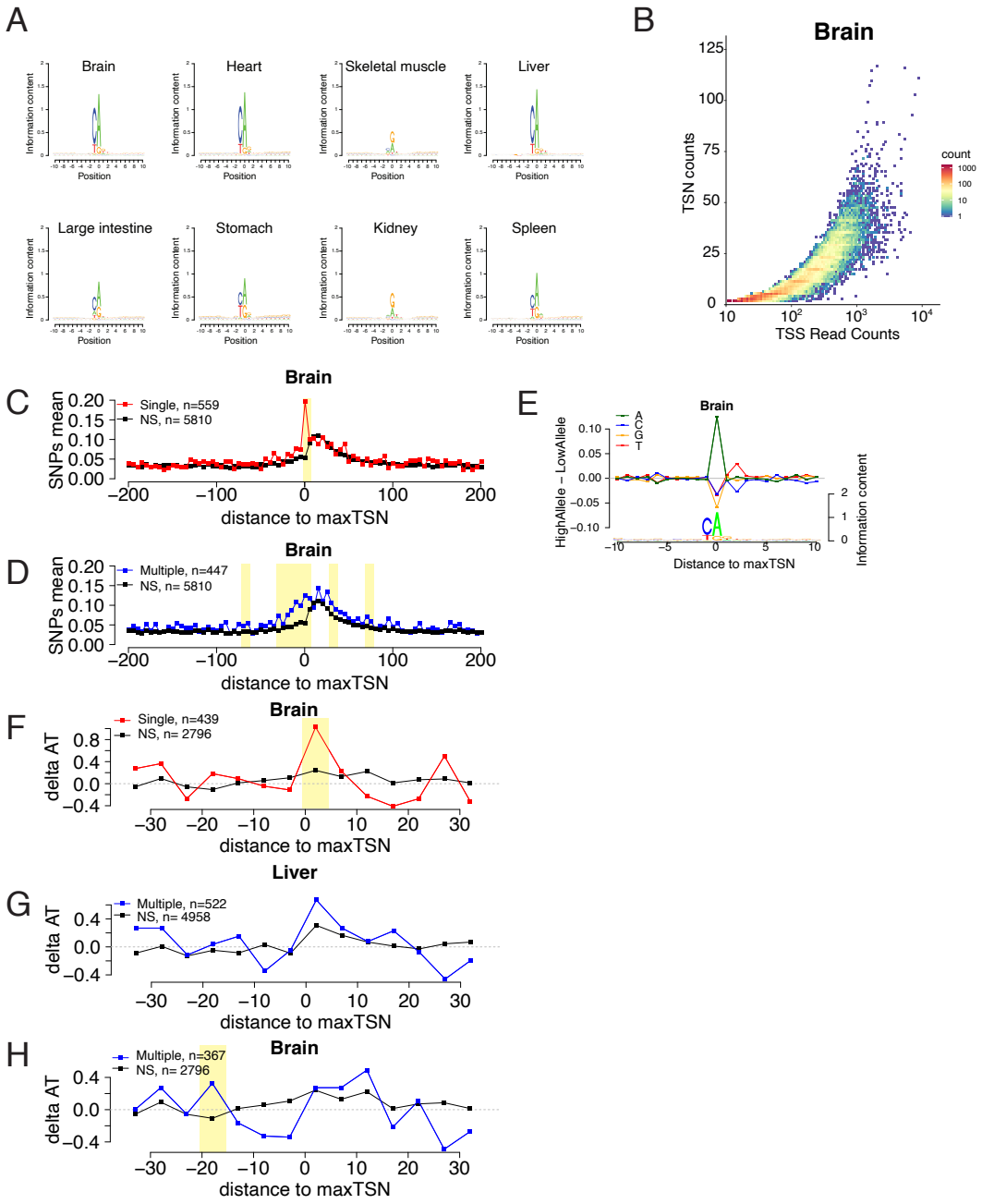


Supplementary Figure 2.1

(A) The histogram shows the frequency of blocks within organ-specific allelic biased domains (OSAB domain) as a function of effect size. Red (Biased) is from the organ with OSAB domain. Blue (NonBiased) is from the organ with the highest expression that is putatively unbiased. If the allelic-biased organ was maternally biased, the effect size was calculated as maternal reads divided by paternal reads in the blocks, otherwise the effect size was calculated as paternal reads divided by maternal reads in the blocks.

(B) The histogram shows the frequency of blocks within the OSAB domain as a function of maternal reads ratio. Red (Biased) is from the organ with OSAB domain. Blue (NonBiased) is from the organ with the highest expression that is putatively unbiased.

(C) The histogram shows the frequency of blocks within OSAB domain as a function of the log2 ratio between the rpkm of the non-allelic-biased organs with highest rpkm (nonBiasedHighest) and the allelic-biased organs in OSAB domains.



Supplementary Figure 2.2

(A) Sequence logos show the information content around the maxTSNs of each organ.

(B) Scatter plot shows the number of TSNs in a TSS as a function of the read counts in the TSS.

(C) Scatterplot shows the average SNP counts as a function of distance to the maxTSN at sites showing allelic differences in TSSs driven by a single base in the brain. Red denotes changes in TSS shape (Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSSs without evidence for differences in TSS shape (KS test; FDR > 0.90). Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05)

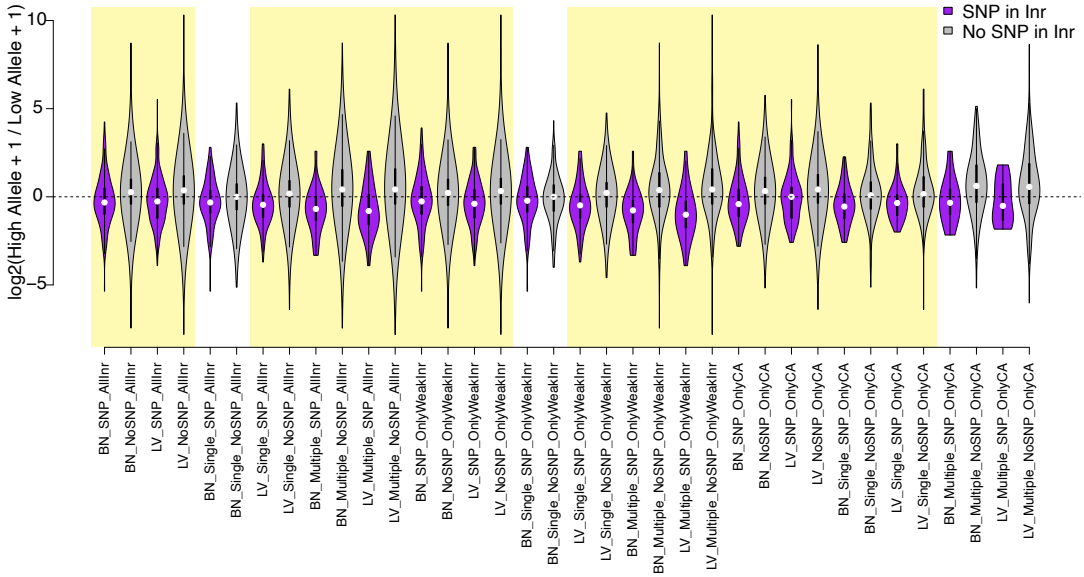
(D) Scatterplot shows the average SNP counts as a function of distance to the maxTSN at sites showing allelic differences in TSS driven by multiple bases in the brain sample. Blue denotes changes in TSS shape classified as multiple TSN driven (FDR corrected Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSSs without evidence for differences in TSS shape (KS test; FDR > 0.90). Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05)

(E) The scatterplot shows the average difference in base composition between the allele with high and low TSN use around the maxTSN in single-base driven allele specific TSSs. The sequence logo on the bottom represents the high allele in single-base driven allele specific TSSs. The high/low allele were determined by the read depth at maxTSN. This figure denotes TSSs in the brain.

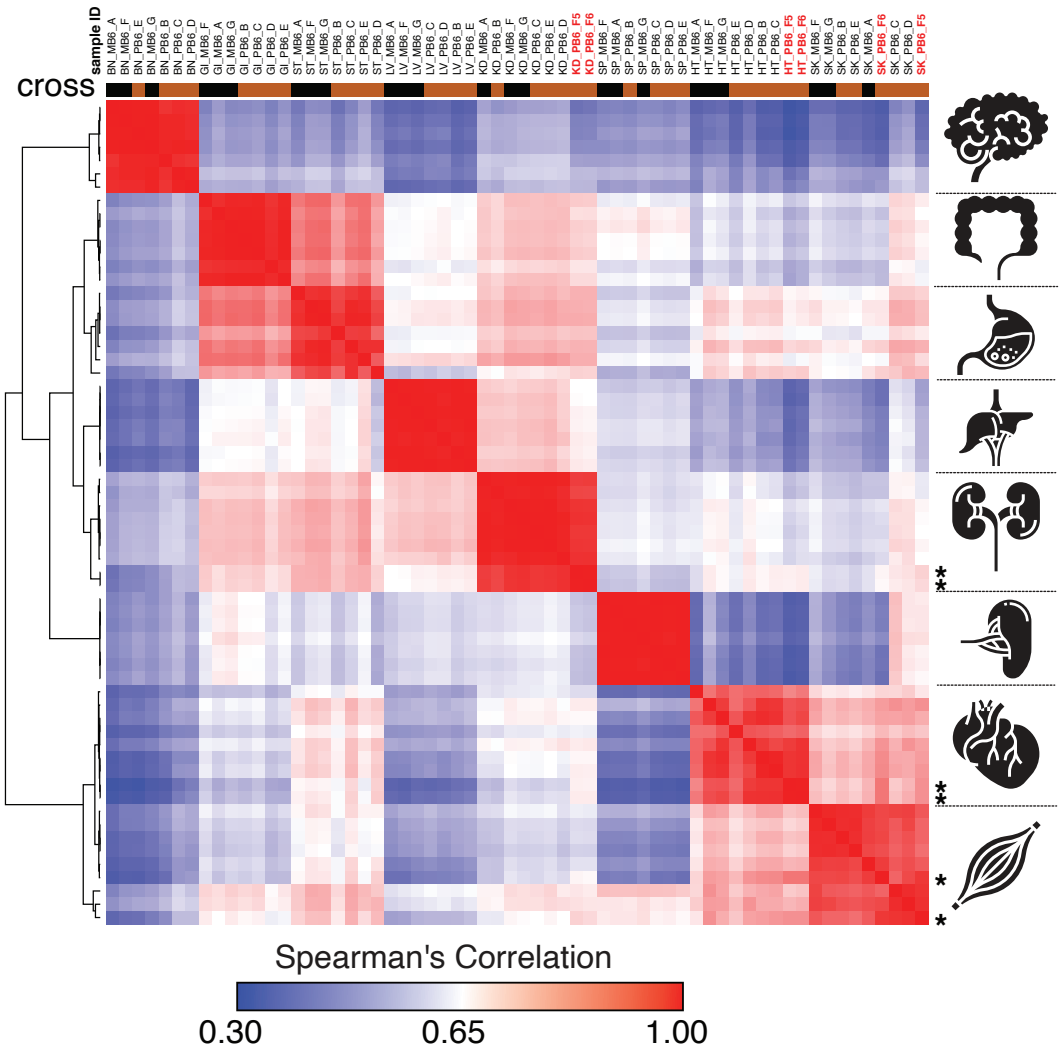
(F) The scatter plot shows the difference of AT contents between the high and low alleles in the brain with the maxTSN and -1 base upstream maxTSN masked. Dots represent 5 bp non-overlapping windows. Red denotes single base driven allele specific TSSs; black denotes control TSSs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT (at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher’s exact test, FDR <= 0.05).

(G) The scatter plots show the difference of AT contents between the high and low alleles in liver with the maxTSN and -1 base upstream maxTSN masked. This plot shows the multiple base driven allele specific TSS in the liver samples. Dots represent 5 bp non-overlapping windows. Blue denotes multiple base driven allele specific TSSs; black denotes control TSSs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT (at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher’s exact test, FDR <= 0.05).

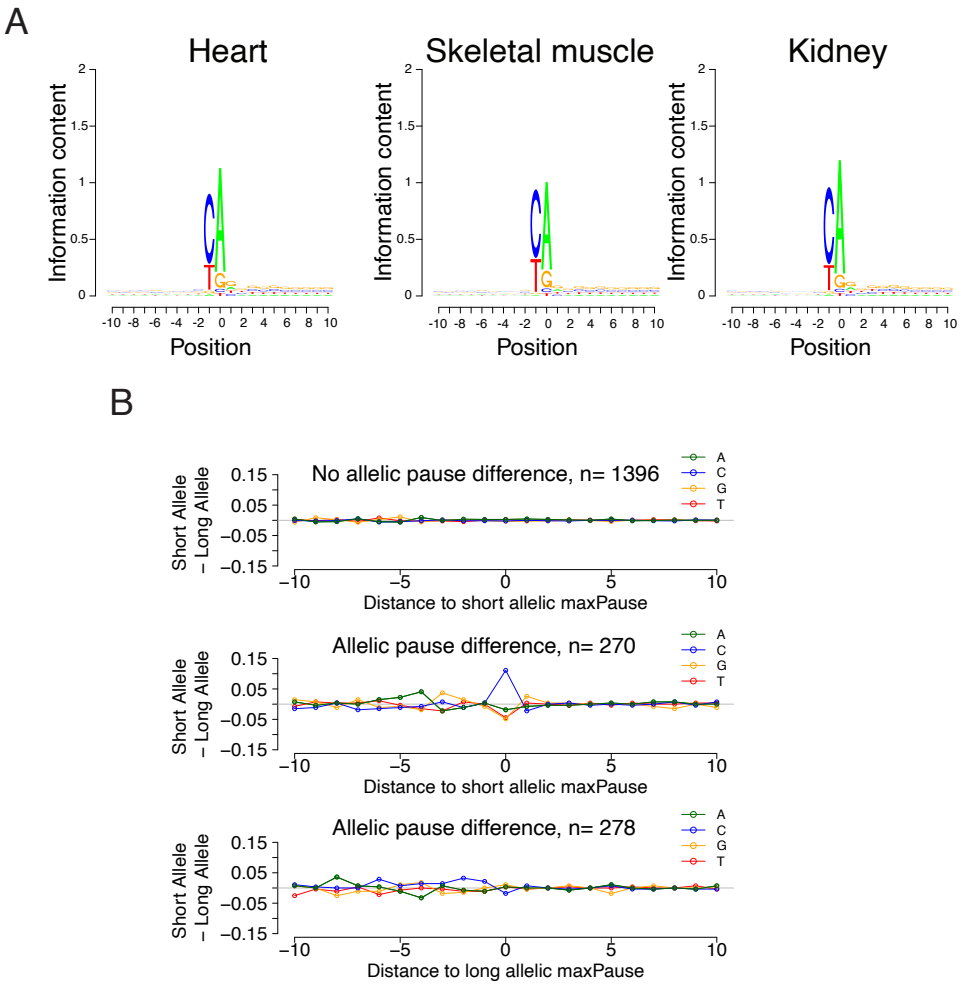
(H) The scatter plots show the difference of AT contents between the high and low alleles with the maxTSN and -1 base upstream maxTSN masked. This plot shows the multiple base driven allele specific TSS in the brain samples. Dots represent 5 bp non-overlapping windows. Blue denotes multiple base driven allele specific TSSs; black denotes control TSSs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT (at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher’s exact test, FDR <= 0.05).



Supplementary Figure 2.3
Violin plots show the distribution of ChRO-seq signals ratios between high low alleles at the candidate initiator motifs (All Inr : CA, CG, TA, TG; OnlyWeak Inr : CG, TA, TG; and OnlyCA) that are within 20bp of the maxTSNs that had a CA dinucleotide in the allele with high maxTSN (SNP in Inr, purple) or had a CA dinucleotide in both alleles (No SNP in Inr, gray) in Brain(BN) or Liver(LV). Single: indicates Single base driven allele-specific TSSs. Multiple: indicates Multiple base driven allele-specific TSSs. Yellow shade indicates Wilcoxon Rank Sum and Signed Rank Tests (SNP in Inr vs no SNP in Inr) with fdr <= 0.05.



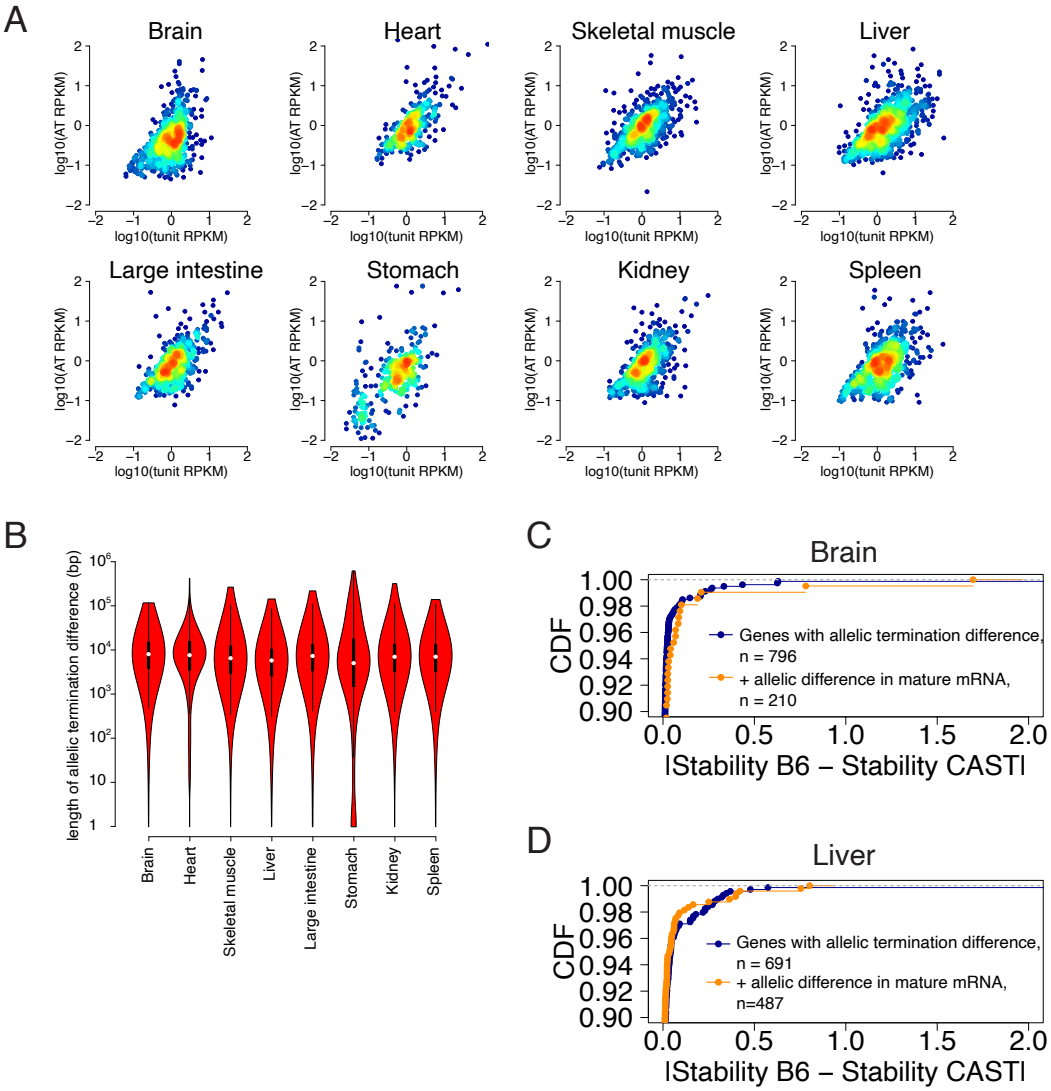
Supplementary Figure 2.4
Spearman's rank correlation of the ChRO-seq data, including samples with a single base resolution for the Pol II active site. Samples with single nucleotide precision are shown on the top in red bold font and at the right with star. (BN:brain, GI: large intestine, ST: Stomach, LV:liver, KD: kidney, SP:spleen, HT: heart, SK:skeletal muscle, MB6: B6 x CAST, PB6: CAST x B6)



Supplementary Figure 2.5

(A) Sequence logos show the information contents around the initiation site (see Methods).

(B) Scatter plots show the difference of nucleotide usage between short and long alleles as a function of distance to allelic maxPause.



Supplementary Figure 2.6

(A) Scatterplots show the relationship between the ChRO-seq signal in the transcription unit (defined using the tunits program; see Methods) and the region showing an allelic termination difference.

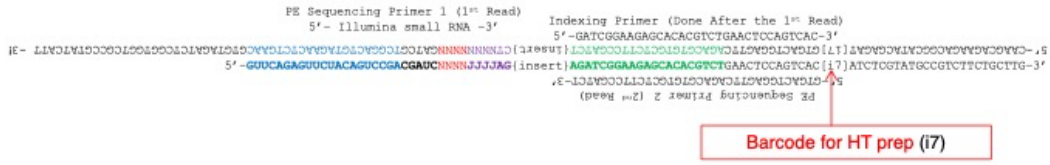
(B) Violin plots show the distribution of the length of allelic termination in eight organs.

(C) Scatterplots represent the cumulative density function of allelic RNA stability differences in brain samples. Two-sample Kolmogorov-Smirnov tests, p-value = 0.9995.

(D) The lines show the cumulative density function of the allelic RNA stability difference in the liver samples. Two-sample Kolmogorov-Smirnov tests, p-value = 0.7542.

The nucleic acid sequences used in custom barcoded high throughput adapters

Primer	Sequence	Notes
RA5-HT (DNA-RNA oligo)	5′ d{GTTCAGAGTTCTACAGTCCGACGATCNNNN}r{JJJJAG}	DNA-RNA hybrid
RA3-HT (RNA oligo)	5′ /5Phos/NNNNAGAUCGGAAGAGCACACGUCU/3InvdT/	RNA
RTP-HT(RT Primer)	5′ AGACGTGTGCTCTTCCGATCT	DNA
RP1 (PCR Primer 1)	5′ AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA Tm (during first 5 cycles) = 63.3 C	DNA
PCR3-HT (PCR Primer x)	5′CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC Tm (during first 5 cycles) = 66.0 C Tm (during remaining cycles)= 78.5 C	DNA



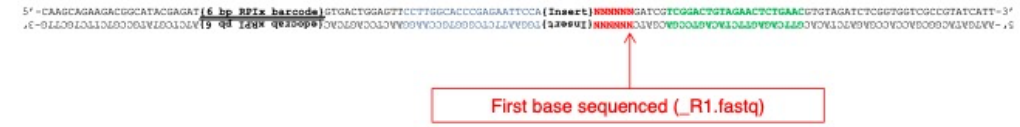
Supplementary Figure 2.7: Nucleic acid sequences used in custom barcoded high throughput adapters.

The nucleic acid sequences used in the single nucleotide resolution ChRO-seq library preparation

Primer	Sequence	Notes
Rev3 (RNA oligo)	5′ /5Phos/NNNNNNGAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/ 3′	RNA
Rev5 (RNA oligo)	5′- CCUUGGCACCCGAGAAUUCCA - 3′	RNA
RP1 (PCR Primer 1)	5′ AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA	DNA
RPIx (PCR Primer x)	5′ CAAGCAGAAGACGGCATACGAGAT{6 bp RPIx barcode}GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA 3′	DNA

RNA library before RT!"

DNA ChRO-seq library after PCR amplification:



Supplementary Figure 2.8: Figure depicts the nucleic acid sequences used in the small RNA designed adapters.

Supplementary Table 2.1: The number of uniquely mapped ChRO-seq reads to individual B6 and CAST genomes across all eight organs.

Organ	Cross	female x male	reads mapped to B6 genome	reads mapped to CAST genome
BN	MB6	C57BL/6 x Castaneus F1 hybrid	48,740,528	48,685,648
BN	PB6	Castaneus x C57BL/6 F1 hybrid	37,540,204	37,536,225
GI	MB6	C57BL/6 x Castaneus F1 hybrid	16,229,569	16,218,255
GI	PB6	Castaneus x C57BL/6 F1 hybrid	23,434,872	23,436,988
HT	MB6	C57BL/6 x Castaneus F1 hybrid	3,010,116	3,007,610
HT	PB6	Castaneus x C57BL/6 F1 hybrid	20,429,188	20,427,203
LV	MB6	C57BL/6 x Castaneus F1 hybrid	32,961,650	32,907,731
LV	PB6	Castaneus x C57BL/6 F1 hybrid	46,974,001	46,936,918
SK	MB6	C57BL/6 x Castaneus F1 hybrid	6,827,847	6,818,752
SK	PB6	Castaneus x C57BL/6 F1 hybrid	14,382,416	14,379,198
SP	MB6	C57BL/6 x Castaneus F1 hybrid	21,904,694	21,880,229
SP	PB6	Castaneus x C57BL/6 F1 hybrid	31,775,452	31,768,726
ST	MB6	C57BL/6 x Castaneus F1 hybrid	12,953,666	12,931,681
ST	PB6	Castaneus x C57BL/6 F1 hybrid	16,594,217	16,570,244
KD	MB6	C57BL/6 x Castaneus F1 hybrid	21,786,140	21,756,388
KD	PB6	Castaneus x C57BL/6 F1 hybrid	21,022,337	21,019,463

REFERENCES

Andergassen, D., Dotter, C.P., Wenzel, D., Sigl, V., Bammer, P.C., Muckenhuber, M., Mayer, D., Kulinski, T.M., Theussl, H.-C., Penninger, J.M., et al. (2017). Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. Elife *6*.

Augui, S., Nora, E.P., and Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. Nat. Rev. Genet. *12*, 429–442.

Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*, 14–24.

Bembom, O. (2019). seqLogo: Sequence logos for DNA sequence alignments.

Blumberg, A., Zhao, Y., Huang, Y.-F., Dukler, N., Rice, E.J., Chivu, A.G., Krumholz, K., Danko, C.G., and Siepel, A. (2021). Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. BMC Biol. *19*, 30.

Braberg, H., Jin, H., Moehle, E.A., Chan, Y.A., Wang, S., Shales, M., Benschop, J.J., Morris, J.H., Qiu, C., Hu, F., et al. (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell *154*, 775–788.

Breslauer, K.J., Frank, R., and Blöcker, H. (1986). Predicting DNA duplex stability from the base sequence. Proceedings of the.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626–635.

Chen, J., Rozowsky, J., Galeev, T.R., Harmanci, A., Kitchen, R., Bedford, J., Abyzov, A., Kong, Y., Regan, L., and Gerstein, M. (2016). A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat. Commun. *7*, 11101.

Cho, H., Kim, T.K., Mancebo, H., Lane, W.S., Flores, O., and Reinberg, D. (1999). A protein phosphatase functions to recycle RNA polymerase II. Genes Dev. *13*, 1540–1552.

Chou, S.-P., and Danko, C.G. (2019). AlleleHMM: a data-driven method to identify allele specific differences in distributed functional genomic marks. Nucleic Acids Res.

Chu, T., Rice, E.J., Booth, G.T., Salamanca, H.H., Wang, Z., Core, L.J., Longo, S.L., Corona, R.J., Chin, L.S., Lis, J.T., et al. (2018). Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. Nat. Genet. *50*, 1553–1564.

Danko, C.G., Hah, N., Luo, X., Martins, A.L., Core, L., Lis, J.T., Siepel, A., and Kraus, W.L. (2013). Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol. Cell *50*, 212–222.

Danko, C.G., Choate, L.A., Marks, B.A., Rice, E.J., Wang, Z., Chu, T., Martins, A.L., Dukler, N., Coonrod, S.A., Tait Wojno, E.D., et al. (2018). Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. Nature Ecology & Evolution.

Delaneau, O., Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., Kumar, S., Ongen, H., Popadin, K., Marbach, D., et al. (2019). Chromatin three-dimensional interactions mediate genetic effects on gene expression. Science *364*, eaat8266.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Fant, C.B., Levandowski, C.B., Gupta, K., Maas, Z.L., Moir, J., Rubin, J.D., Sawyer, A., Esbin, M.N., Rimel, J.K., Luyties, O., et al. (2020). TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. Mol. Cell *78*, 785–793.e8.

Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186–192.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063.

Gilchrist, D.A., Dos Santos, G., Fargo, D.C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell *143*, 540–551.

Gressel, S., Schwalb, B., Decker, T.M., Qin, W., Leonhardt, H., Eick, D., and Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. Elife *6*.

Gressel, S., Schwalb, B., and Cramer, P. (2019). The pause-initiation limit restricts transcription activation in human cells. Nat. Commun. *10*, 3603.

Grünberg, S., Warfield, L., and Hahn, S. (2012). Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. Nat. Struct. Mol. Biol. *19*, 788–796.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell Biol.

Ho, A.Y., and Dimitropoulos, A. (2010). Clinical management of behavioral characteristics of Prader-Willi syndrome. Neuropsychiatr. Dis. Treat. *6*, 107–118.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 167–177.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Kaplan, C.D., Jin, H., Zhang, I.L., and Belyanin, A. (2012). Dissection of Pol II trigger loop function and Pol II activity-dependent control of start site selection in vivo. PLoS Genet. *8*, e1002627.

Kaufmann, J., and Smale, S.T. (1994). Direct recognition of initiator elements by a component of the transcription factor IID complex. Genes Dev. *8*, 821–829.

Kristjánsdóttir, K., Dziubek, A., Kang, H.M., and Kwak, H. (2020). Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. Nat. Commun. *11*, 5963.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013).

Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Li, J. (2019). TmCalculator: Melting Temperature of Nucleic Acid Sequences.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Luse, D.S., Parida, M., Spector, B.M., Nilson, K.A., and Price, D.H. (2020). A unified view of the sequence and functional organization of the human RNA polymerase II promoter. Nucleic Acids Res.

Mahat, D.B., Salamanca, H.H., Duarte, F.M., Danko, C.G., and Lis, J.T. (2016a). Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. Mol. Cell.

Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016b). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat. Protoc. *11*, 1455–1476.

Miller, T., Krogan, N.J., Dover, J., Erdjument-Bromage, H., Tempst, P., Johnston, M., Greenblatt, J.F., and Shilatifard, A. (2001). COMPASS: A complex of proteins associated with a trithorax-related SET domain protein. Proceedings of the National Academy of Sciences *98*, 12902–12907.

Mittleman, B.E., Pott, S., Warland, S., Zeng, T., Mu, Z., Kaur, M., Gilad, Y., and Li, Y. (2020). Alternative polyadenylation mediates genetic regulation of gene expression. Elife *9*.

Mittleman, B.E., Pott, S., Warland, S., Barr, K., Cuevas, C., and Gilad, Y. (2021). Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees. Elife *10*.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D.A., Adams, C.M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B.J., et al. (2013). Architecture of an RNA Polymerase II Transcription Pre-Initiation Complex. Science *342*.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat. Genet. *39*, 1507–1511.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335–338.

Orphanides, G., LeRoy, G., Chang, C.H., Luse, D.S., and Reinberg, D. (1998). FACT, a factor that facilitates transcript elongation through nucleosomes. Cell *92*, 105–116.

O'Sullivan, J.M., Tan-Wong, S.M., Morillon, A., Lee, B., Coles, J., Mellor, J., and Proudfoot, N.J. (2004). Gene loops juxtapose promoters and terminators in yeast. Nat. Genet. *36*, 1014–1018.

Pelc, K., Cheron, G., and Dan, B. (2008). Behavior and neuropsychiatric manifestations in Angelman syndrome. Neuropsychiatr. Dis. Treat. *4*, 577.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

Plasschaert, R.N., and Bartolomei, M.S. (2015). Tissue-specific regulation and function of Grb10 during growth and neuronal commitment. Proc. Natl. Acad. Sci. U. S. A. *112*, 6841–6847.

Qiu, C., Jin, H., Vvedenskaya, I., Llenas, J.A., Zhao, T., Malik, I., Visbisky, A.M., Schwartz, S.L., Cui, P., Čabart, P., et al. (2020). Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. Genome Biol. *21*, 132.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445.

Ranish, J.A., Yudkovsky, N., and Hahn, S. (1999). Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. Genes Dev. *13*, 49–63.

Rennie, S., Dalby, M., van Duin, L., and Andersson, R. (2018). Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. Nat. Commun. *9*, 487.

Rosonina, E., Kaneko, S., and Manley, J.L. (2006). Terminating the transcript: breaking up is hard to do. Genes Dev. *20*, 1050–1056.

Rougvie, A.E., and Lis, J.T. (1988). The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell *54*, 795–804.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. *7*, 522.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J., and Cramer, P. (2016). TT-seq maps the human transient transcriptome. Science *352*, 1225–1228.

Sleutels, F., Zwart, R., and Barlow, D.P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature *415*, 810–813.

Smale, S.T., and Baltimore, D. (1989). The "initiator" as a transcription control element. Cell *57*, 103–113.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. *100*, 9440–9445.

Tome, J.M., Tippens, N.D., and Lis, J.T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. Nat. Genet.

Traut, T.W. (1994). Physiological concentrations of purines and pyrimidines. Mol. Cell. Biochem. *140*, 1–22.

Tsai, F.T.F., and Sigler, P.B. (2000). Structural basis of preinitiation complex assembly on human Pol II promoters. EMBO J. *19*, 25–36.

Wang, X., Soloway, P.D., and Clark, A.G. (2011). A survey for novel imprinted genes in the mouse placenta by mRNA-seq. Genetics *189*, 109–122.

Wang, Z., Chu, T., Choate, L.A., and Danko, C.G. (2018). Identification of regulatory elements from nascent transcription using dREG. Genome Res. *29*, 293–303.

Wilkins, J.F. (2014). Genomic imprinting of Grb10: coadaptation or conflict? PLoS Biol. *12*, e1001800.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat. Genet. *39*, 1512–1516.

CHAPTER 3

Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins[†]

J. Leonardo Moreno-Gallego[1]*, Shao-Pei Chou[2]*, Sara C. Di Rienzi[2], Julia K. Goodrich[1], Timothy Spector[3], Jordana T. Bell[3], Nicholas Youngblut[1], Ian Hewson[6], Alejandro Reyes[4,5], and Ruth E. Ley[1¥Δ]

1. Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen 72076, Germany.
2. Department of Molecular Biology and Genetics, Cornell University, Ithaca NY 14853, USA.
3. Department of Twin Research and Genetic Epidemiology, King's College London, London SE1 7EH, UK.
4. Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Universidad de los Andes, Bogotá 111711, Colombia.
5. Center for Genome Sciences and Systems Biology, Washington University School of Medicine, Saint Louis, MO 63108, USA.
6. Department of Microbiology, Cornell University, Ithaca NY 14853 USA.

* These authors contributed equally

¥ Lead contact

Δ Correspondence: rley@tuebingen.mpg.de

† This chapter was published as an article in Cell Host Microbe in 2019. The citation is doi: 10.1016/j.chom.2019.01.019 (PMID 30763537).

3.1 Abstract

The virome is one of the most variable components of the human gut microbiome. Within twin pairs, viromes have been shown to be similar for infants, but not for adults, indicating that as twins age and their environments and microbiomes diverge, so do their viromes. The degree to which the microbiome drives the vast virome diversity is unclear. Here, we examine the relationship between microbiome and virome diversity in 21 adult monozygotic twin pairs selected for high or low microbiome concordance. Viromes derived from virus-like particles are unique to each individual, are dominated by *Caudovirales* and *Microviridae*, and exhibit a small core that includes crAssphage. Microbiome-discordant twins display more dissimilar viromes compared to microbiome-concordant twins, and the richer the microbiomes, the richer the viromes. These patterns are driven by bacteriophages, not eukaryotic viruses. Collectively, these observations support a strong role of the microbiome in patterning for the virome.

3.2 Introduction

The human gut microbiome is composed of a vast diversity of bacterial cells, along with a minority of archaeal and eukaryotic cells, together forming a very dense microbial ecosystem (10^{11}–10^{12} cells per gram of feces) (Sender et al., 2016). The cells of the microbiome and the constituents of the virome (between 10^9 and 10^{12} virus-like particles [VPLs] per gram of feces) are in about equal proportion (Castro-Mejía et al., 2015, Hoyles et al., 2014, Ogilvie and Jones, 2017, Reyes et al., 2010). The virome is primarily composed of bacteriophages and prophages, and it also includes rarer eukaryotic viruses and endogenous retroviruses (Breitbart et al., 2003, Minot et al., 2011, Reyes et al., 2010). Currently, the majority of phages have no matches in databases and their hosts remain to be elucidated. Matching phages to their hosts is challenging: for instance, the host of the most common human gut phage, crAssphage, has only recently been identified as *Bacteroides* spp. (Shkoporov et al., 2018, Yutin et al., 2018). In addition to the identification of hosts, other questions remain as to the factors most important in shaping the virome, and how predictive the cellular fraction of the microbiome can be of the virome.

The temporal population dynamics of phages and their hosts might be expected to be linked. Indeed, population oscillations of viruses and their bacterial hosts are described for aquatic systems, where they indicate that viruses play a key role in regulating bacterial populations (Suttle, 2007, Thingstad, 2000, Thingstad et al., 2014, Weitz and Dushoff, 2008). But such patterns of predator-prey dynamics are not typical for the human gut virome and microbiome (Minot et al., 2011, Reyes et al., 2013, Rodriguez-Brito et al., 2010, Rodriguez-Valera et al., 2009). (For clarity, from

here on we use "microbiome" to refer to cellular fraction of the microbiome, e.g., mostly bacterial cells.) Nonetheless, the virome and microbiome do display some common patterns of diversity across hosts, such as high levels of interpersonal differences and relative stability over time (Reyes et al., 2010). The microbiome tends to be more similar for related individuals compared to unrelated individuals, possibly due to shared dietary habits, which drive similarity between microbiomes (Cotillard et al., 2013, David et al., 2014). In accord, diet has been associated with virome diversity, quite possibly through diet effects on the microbiome (Minot et al., 2011). In infants, twin comparisons have revealed viromes to be more similar between co-twins than between unrelated individuals (Lim et al., 2015, Reyes et al., 2015). This pattern was not observed in adult twins (Reyes et al., 2010), possibly due to divergence of their microbiomes. The degree to which the microbiome itself drives patterns of virome diversity across hosts has been difficult to assess due to confounding factors such as host relatedness.

Here, we focus on adult monozygotic (MZ) twin gut microbiomes to explore further the relationship between microbiome and virome diversity. By studying the viromes of MZ twin pairs, we control for host genetic relatedness. Although MZ twin pairs generally have more similar microbiomes compared to dizygotic (DZ) twin pairs or unrelated individuals, MZ twins nevertheless can display a large range of within-twin-pair microbiome diversity (Goodrich et al., 2014). We previously generated fecal microbiome data for twin pairs from the TwinsUK cohort (Goodrich et al., 2014), and based on this information we selected twin pairs either highly concordant or highly discordant for their microbiomes. We generated viromes from virus-like particles

obtained from the same samples from which the microbiomes were derived. Results indicate that microbiome diversity and virome diversity measures are positively associated.

3.3 Results

3.3.1 Selection of Microbiome-Concordant and -Discordant Monozygotic Twin Pairs

We selected twin pairs with a similar body mass index (BMI), whose microbiomes were either concordant or discordant for microbiome between-sample diversity (β-diversity) based on previously obtained 16S rRNA gene data. The adult co-twins in this study did not share a household, and we assume that other environmental variability was similar across twin pairs. We determined the degree of concordance or discordance between co-twins' microbiomes based on three β-diversity distance metrics: Bray-Curtis, weighted UniFrac, and unweighted UniFrac (Materials and Methods). As expected, the β-diversity measures were correlated (Pearson pairwise correlation coefficient > 0.4). Based on the distribution of pairwise distance measures, we selected 21 MZ twin pairs from the boundaries of all three distributions (Figure 3.1A), while maintaining a balanced distribution of age and BMI across the set (Supplementary Table 4.1). Within the 21 selected twin pairs, the microbiomes of microbiome-concordant co-twins were, as expected, more similar to each other than microbiomes of microbiome-discordant co-twins ($p = 6.31 \times 10^{-12}$). The microbiomes of the discordant co-twins differed compositionally at all taxonomic levels, particularly at the phylum level, with Firmicutes and Bacteroidetes, the two dominant phyla, contributing the most to the variation between co-twins (Figure 3.1B,C).

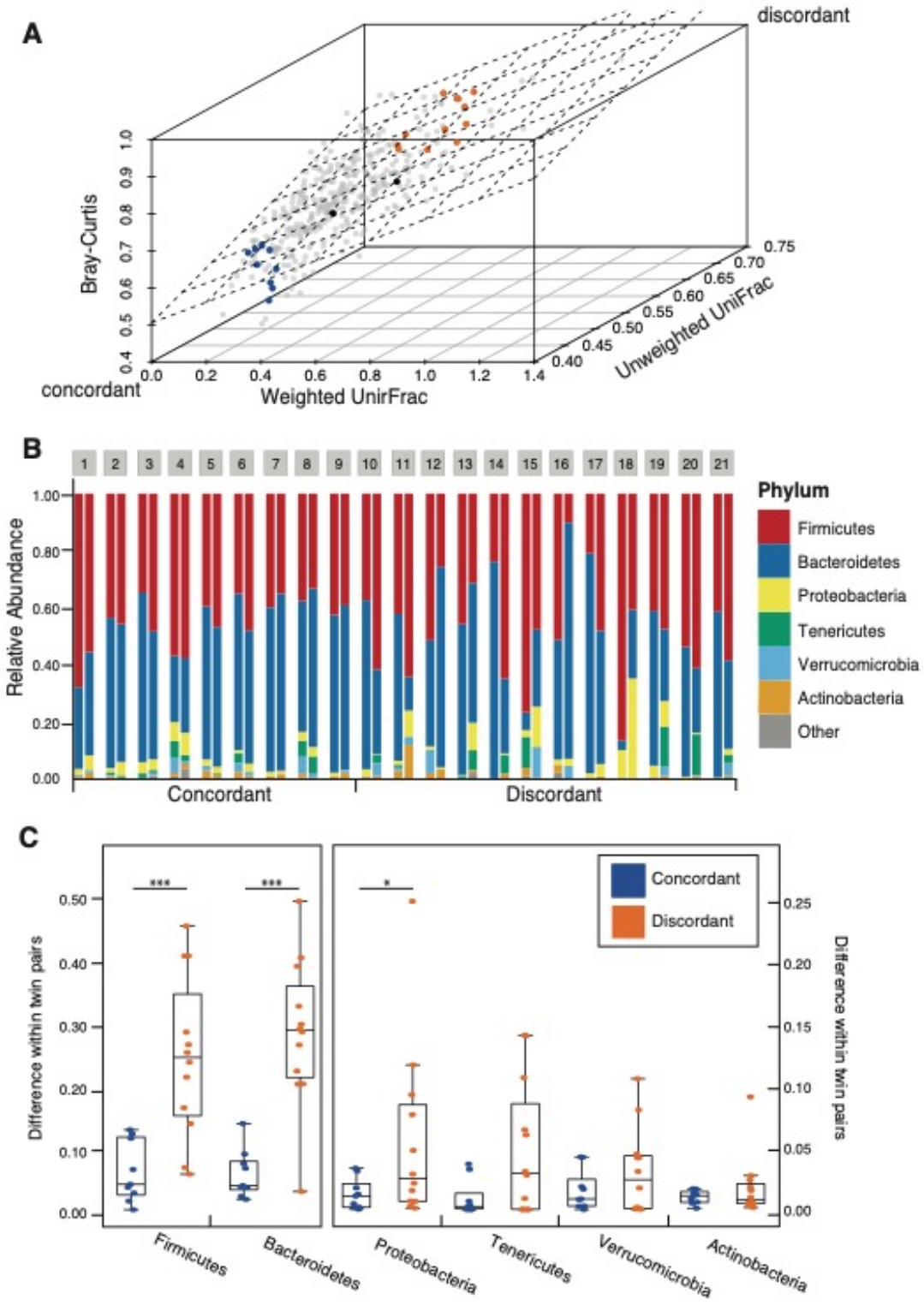

A
discordant
concordant
Bray-Curtis
Weighted UnirFrac
Unweighted UniFrac
B
Relative Abundance
Concordant
Discordant
Phylum
Firmicutes
Bacteroidetes
Proteobacteria
Tenericutes
Verrucomicrobia
Actinobacteria
Other
C
Concordant
Discordant
Difference within twin pairs
Firmicutes
Bacteroidetes
Proteobacteria
Tenericutes
Verrucomicrobia
Actinobacteria

Figure 3.1: Microbiome Discordance in Twin Pairs

(A) The β-diversity measures of the microbiotas of 354 monozygotic twin pairs from a previous study (Goodrich et al., 2014) are shown. Each dot represents the β-diversity of a pair of twins, measured by the weighted UniFrac (x axis), unweighted UniFrac (z axis), and Bray-Curtis (y axis) β-diversity metrics. The plane is the least squared fitted plane Bray-Curtis ~ Weighted UniFrac + Unweighted UniFrac. A subset of twin pairs with concordant microbiotas (blue) and discordant microbiotas (orange) was chosen from the two edges. Black dots indicate the samples used for virome and whole fecal metagenome comparison.

(B) Comparison of the taxonomic profiles (relative abundance) at the phylum level for the 21 MZ twin pairs concordant (1–9) or discordant (10–21) for their microbiotas.

(C) Differences in the relative abundances for the major phyla for concordant (blue points, n = 9) and discordant (orange points, n = 12) twin pairs. Mann-Whitney's U test. $^{***}p < 0.0005$, $^{*}p = 0.055$.

3.3.2 Shotgun Metagenomes of VLPs

We isolated VLPs from the same fecal samples that had been used for 16S rRNA gene diversity profiling (Materials and Methods). DNA extracted from VLPs was used in whole-genome amplification followed by shotgun metagenome sequencing (Materials and Methods). A first library ("large-insert-size library") was selected with an average insert size of 500 bp (34,325,116 paired reads in total; 817,265 ± 249,550 paired reads per sample after quality control) and used for *de novo* assembly of viral contigs. Smaller fragments with an average insert size of 300 bp were purified in a second library ("small-insert-size library") and sequenced. The resulting pair-end reads were merged into 25,324,163 quality filtered longer reads to increase mapping accuracy (602,956 ± 595,444 merged reads per sample).

3.3.3 Identification of Putative Bacterial Contaminants

Viromes prepared and sequenced from VLPs may be contaminated with bacterial DNA (Roux et al., 2013). However, given that phages are major agents of horizontal gene transfer and that temperate viruses often comprise up to 10% of bacterial genomes in a prophage state, removal of potential bacterial contamination risks also removing viral reads. To assess bacterial DNA contamination, we mapped virome reads against a set of 8,163 fully assembled bacterial genomes. Our strategy consisted of evaluating the coverage along the length of each genome (in bins of 100 Kb), and those genomes with a median coverage greater than 100 were considered contaminants. Reads mapping to short regions were considered to be prophages or horizontally transferred genes and

retained (Materials and Methods) (Figure 3.2A). Reads mapping to genomes determined to be potential contaminants were removed from further analyses.

We identified 65 bacterial genomes as potential contaminants, with 1% ± 1.125% (average ± SD) of reads per sample mapping to those bacterial genomes. The majority (37/68) belonged to the Firmicutes phylum; at the species level, *Bacteroides dorei*, *B. vulgatus*, *Ruminococcus bromii*, *Faecalibacterium prausnitzii*, *B. xylanisolvens*, *Odoribacter splanchnicus*, and *B. caecimuris* (in that order) were detectable in at least 50% of the samples. If the most abundant bacterial species in the microbiome are the most likely sources of contamination, then their relative abundance as contaminants should correspond to their relative abundances in the microbiome. However, we observed no significant correlation between the relative abundances of taxa represented in the contaminant DNA and in the microbiomes (Figure 3.2B).

3.3.4 Functional Profiles Support Viral Enrichment in VLP Purifications

To assess the functional content of the viromes, we annotated the "short-insert-size library" raw reads using the KEGG annotation of the Integrated Gene Catalog (IGC) (Li et al., 2014) (Materials and Methods). In line with previous reports (Breitbart et al., 2008, Minot et al., 2011, Reyes et al., 2010), the majority of reads (85.43% ± 5.74%) from our VLP metagenomes mapped to genes with unknown function (Figure 3.3A).

To further verify that sequences were derived from VLPs and not microbiomes generally, we conducted an internal check in which we generated and compared additional metagenomes from VLPs and bulk fecal DNA for an additional four

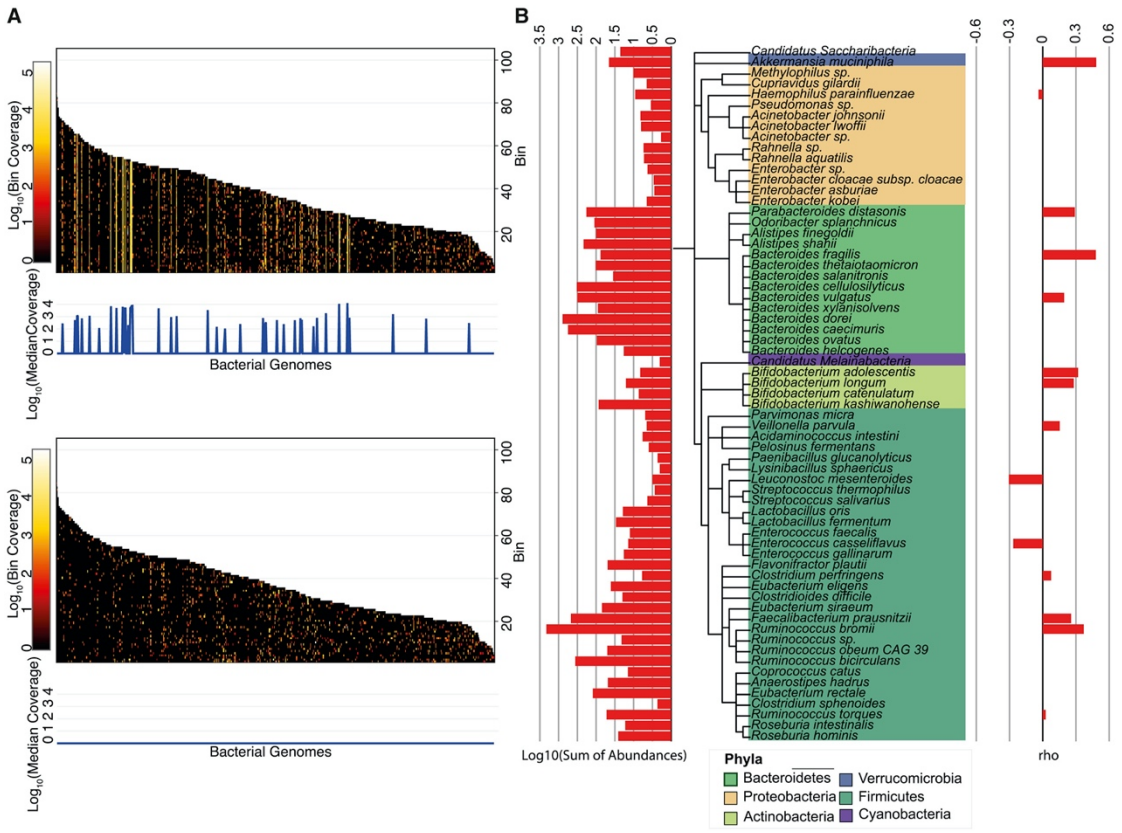


Figure 3.2: Bacterial Contamination in VLP Preparations

(A) Heatmaps of VLP reads from a single sample (4A) mapping to bacterial genomes before (upper) and after (lower) the removal of reads determined as contaminants. Bacterial genomes are represented with vertical bars, sorted by length and split into bins of 100,000 bp. Genomes with a median coverage greater than 100 were considered contaminants. The color scale to the left shows bin coverage and the scatterplots below each heatmap show the median bin coverage of each genome.

(B) Cladogram based on the NCBI taxonomy of the 65 genomes identified as contaminants across all VLP extractions. Right: Spearman rank correlation coefficient (rho) between the abundance of the bacterial genomes in the VLP extractions and 16S rRNA gene profile from the microbiome. Left: total abundance of each bacterial genome added across all individuals.

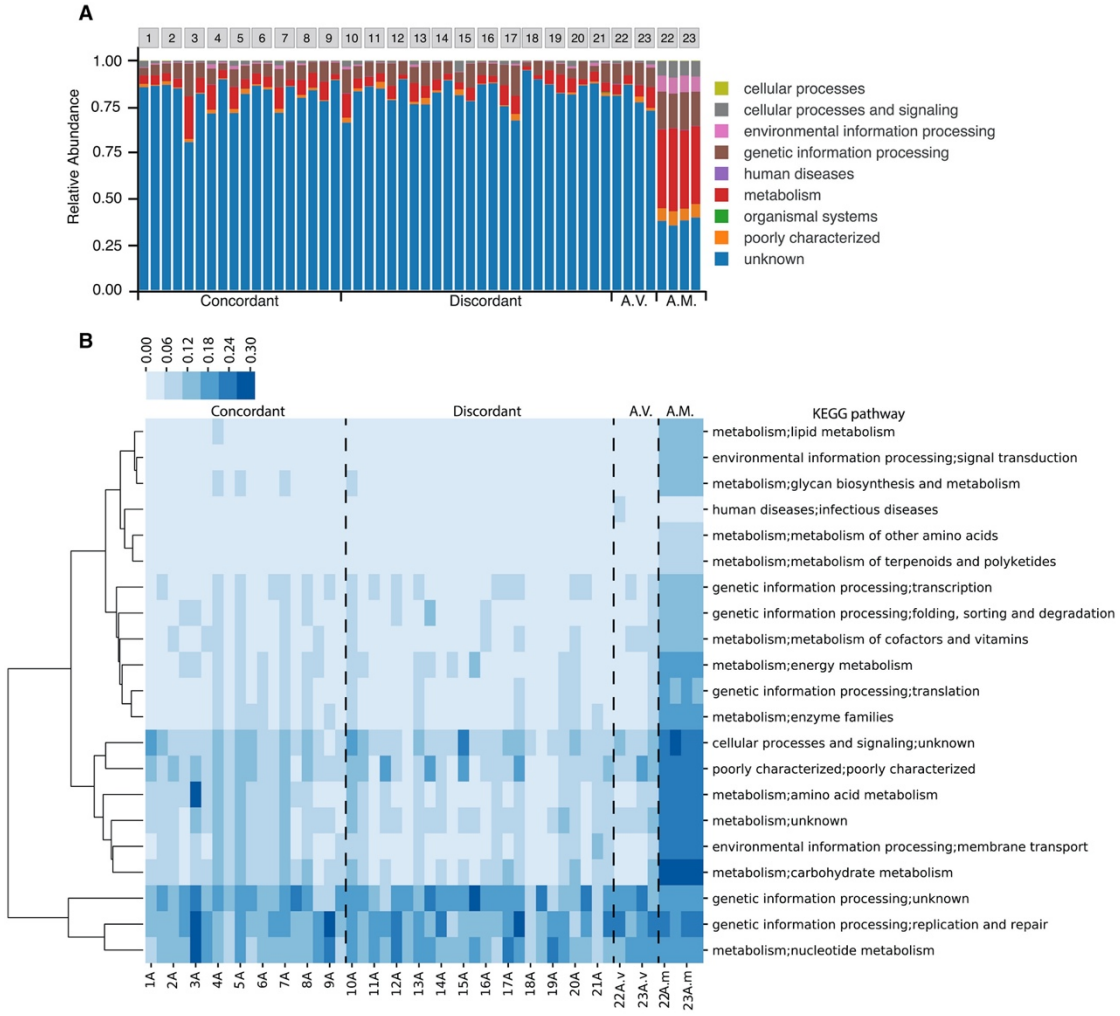


Figure 3.3: Comparison of the Gene Content of Whole Fecal Metagenomes and Viromes

(A) The relative abundance of KEGG categories in whole fecal metagenomes and viromes, including all hits to IGC genes, regardless of annotation.

(B) Heatmap of the relative abundance of the second level of KEGG categories in whole fecal metagenomes and viromes, excluding the IGC genes with unknown annotation. The color scale shows the square root transformed relative abundances. A.V., additional viromes; A.M., additional microbiomes (whole-genome extractions). Intra-class coefficient (ICC) for A.M. = 0.99; ICC for A.V. = 0.85; ICC concordant-microbiome co-twins = 0.69; ICC discordant-microbiome co-twins = 0.68.

individuals (two twin pairs; Figure 3.1A). As expected, the functional profiles of viromes and microbiome-metagenomes derived from the same samples were dissimilar. Virome reads that mapped to annotated genes were enriched in two categories: Genetic Information Process (48.87% ± 12.12%) and Nucleotide Metabolism (17.59% ± 8.81%), compared to 24.31% ± 1.28% and 5.47% ± 0.4% for the microbiome-metagenome, respectively (Figure 3.3B). Most of the other functional categories present in the bacterial metagenomes were essentially absent from the viromes. Furthermore, the functional annotations of the viromes showed greater between-sample variability than the microbiomes and a lower intraclass correlation coefficient (Figure 3.3B).

3.3.5 Viromes Are Unique to Individuals

We assembled reads from the "large-insert-size library" resulting in a total of 107,307 contigs ≥500 nt (max, 79,863 nt; mean, 1,186 nt ± 1,741; Supplementary Figure 3.1). To assess the structure and composition of the viromes, a matrix of the recruitment of reads against dereplicated contigs was built (Materials and Methods). The recruitment matrix included 14,584 contigs that were both long (>1,300 nt) and well covered (>5X); these are referred to as "virotypes" (Supplementary Figure 3.1). Analysis of the recruitment matrix showed that each individual harbored a unique set of virotypes: 3,415 virotypes (23.41% of total) were present in only one individual, 413 virotypes (2.83%) were present in at least 50% of the individuals, and only 18 virotypes (0.1%) were present in all individuals.

3.3.6 Twins with Concordant Microbiomes Share Virotypes

We checked for virotypes shared between twins and observed that co-twins did not share more virotypes than unrelated individuals ($p = 0.074$). We then assessed microbiome-concordant and -discordant twin pairs separately: twins with a discordant microbiome did not share more virotypes that unrelated individuals ($p = 0.254$), whereas twins with a concordant microbiome did share more virotypes than unrelated individuals ($p = 0.048$). Furthermore, we also found that twins with a concordant microbiome shared more virotypes than twins with a discordant microbiome ($p = 0.015$; Supplementary Figure 3.2).

3.3.7 Bacteriophage Dominance of the Gut Virome

In order to characterize the taxonomic composition of the virome, we attempted to annotate all 66,446 dereplicated and well-covered contigs (Supplementary Figure 3.1) using a voting system approach that exploited the information in both the assembled contigs and their encoding proteins (Materials and Methods). In addition, we performed a custom annotation on two highly abundant gut-associated bacteriophage families: (1) the crAssphage (Dutilh et al., 2014, Yutin et al., 2018 and (2) the *Microviridae* families (Székely and Breitbart, 2016). For this, we used profile Hidden Markov Models (HMMs) to search for crAssphage (double-stranded DNA [dsDNA] viruses) and *Microviridae* (single-stranded DNA [ssDNA] viruses) contigs (Materials and Methods).

Using HMMs allowed us to identify distant homologs, which we then incorporated into a phylogenetic tree with known reference sequences to confirm the annotation and better resolve the taxonomy. We annotated 108 contigs (19 crAssphage,

90 *Microviridae*), validated the family assignment of 68 contigs, and assigned a subfamily to 97 contigs without previous subfamily assignment. For the *Microviridae*, only 11 contigs had a previous taxonomic assignment, all belonging to the *Gokushovirinae*: we confirmed these and 23 more as *Gokushovirinae*, 54 as Alpavirinae, and 1 contig as *Pichovirinae* (Supplementary Figure 3.3A). For the crAssphage, 11 contigs were clustered with the original crAssphage, 3 contigs grouped with the reference Chlamydia phage, and 5 contigs grouped with the reference IAS virus (Supplementary Figure 3.3B).

After collating the voting system annotation and the HMM annotation, a total of 12,751 contigs (29.62%) were taxonomically assigned (Supplementary Figure 3.1). Viromes were dominated by bacteriophages with only 6.42% of contigs annotated as eukaryotic viruses. As expected, most of the contigs (96.98%) were dsDNA viruses, while only 2.43% of contigs were annotated as ssDNA viruses. Caudovirales was the most abundant order, with its three main families represented: *Myoviridae* (20.22% ± 4.83%), *Podoviridae* (10.54% ± 3.27%), and *Siphoviridae* (35.25% ± 7.19%). The crAssphage family constituted on average 13.26% (±12.24%) of the contigs, reaching a maximum contribution of 55.80% in one virome, and *Microviridae* represented 3.87% ± 2.57% of the viromes. Interestingly, we observed that *Phycodnaviridae* exceeded 1% of average abundance (1.77% ± 1.12%; Figure 3.4A) and that contigs related to any nucleocytoplasmic large DNA viruses (NCLDV) had a mean relative contribution of 3.99% ± 2.22%. The 18 contigs present in all samples included 10 annotated as crAssphage, 2 annotated as "unclassified Myoviridae," 2 "unclassified Caudovirales," 1 classified as *Microviridae*, and 3 unclassified. Within

a defined taxonomic profile for each sample, we looked for differences in composition between viromes at all taxonomic levels for concordant and discordant twin pairs. There were no significant differences between groups for any taxa at the order and family levels, including crAssphage and *Microviridae* families (Figure 3.4B).

We used CRISPR spacer mapping and the microbe-versus-phage (MVP) database (Gao et al., 2018) to predict hosts for virotypes and taxonomically characterized contigs (Materials and Methods). As host annotation was directed to bacteriophages, we did not gain any information for contigs annotated as eukaryotic viruses. These approaches allowed us to identify putative hosts for 910 contigs. Within these 910 contigs, only one was previously annotated as crAssphage, and as expected, its host was inferred to be a member of *Bacteroidetes*. In total we identified 1,280 bacterial putative host strains, including 187 species from 87 genera over several phyla: most of them from Firmicutes (92), followed by Bacteroidetes (41) and Proteobacteria (38). The median number of host for each contig was 1 (IQR = 1–2), while the median number of phages per host, at the strain level, was 2 (IQR = 1–3) (Supplementary Figure 3.4).

3.3.8 Virome Diversity Correlates with Microbiome Diversity

To assess the relationship between virome and microbiome diversity, we examined the within-sample diversity (α-diversity) and β-diversity of the viromes using three different layers of information that we recovered from the sequence data: (1) virotypes, (2) taxonomically annotated contigs, and (3) annotated genes from short reads (Supplementary Figure 3.1).

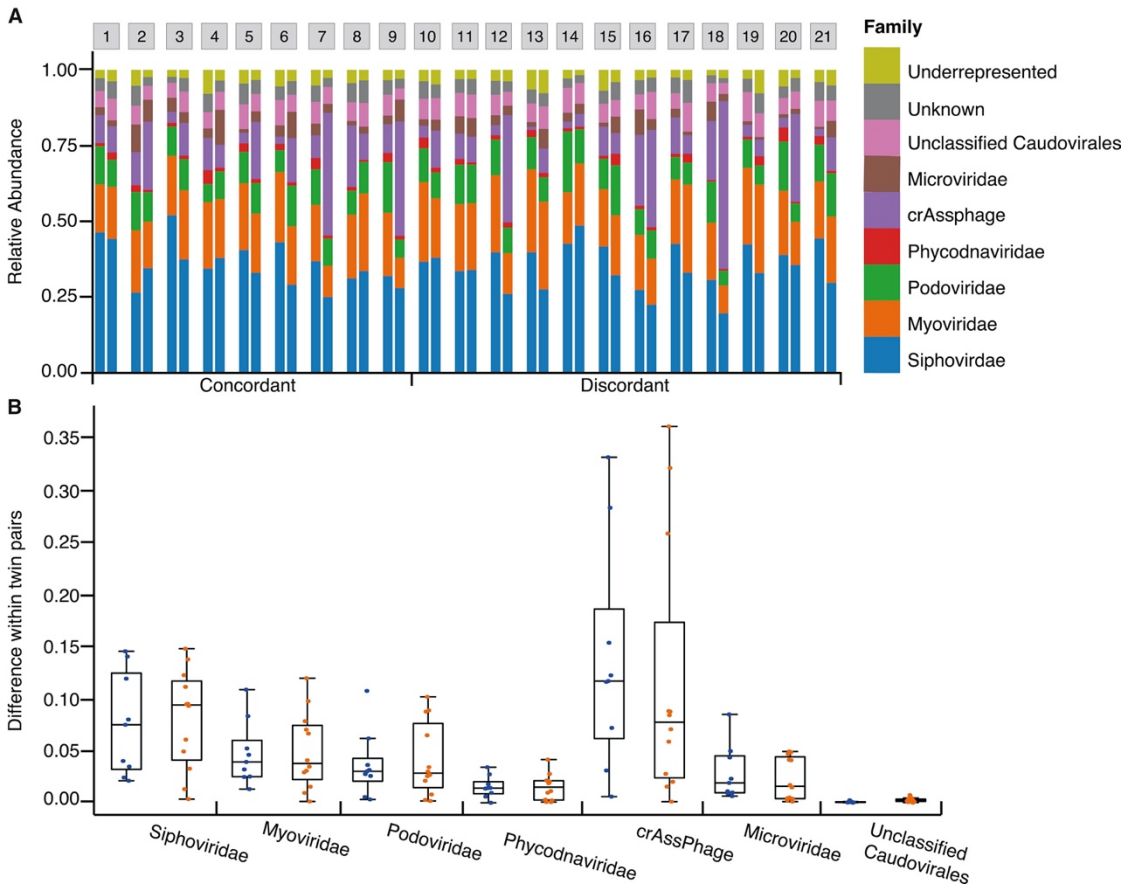


Figure 3.4: Virome Composition

Comparison of the taxonomic profiles at the family level for the 21 MZ twin pairs concordant (1–9) or discordant (10–21) for their microbiomes.

(A) The viral family composition of the MZ twins.

(B) Differences of the relative abundances of each family for concordant (blue points, n = 9) and discordant (orange points, n = 12) twin pairs.

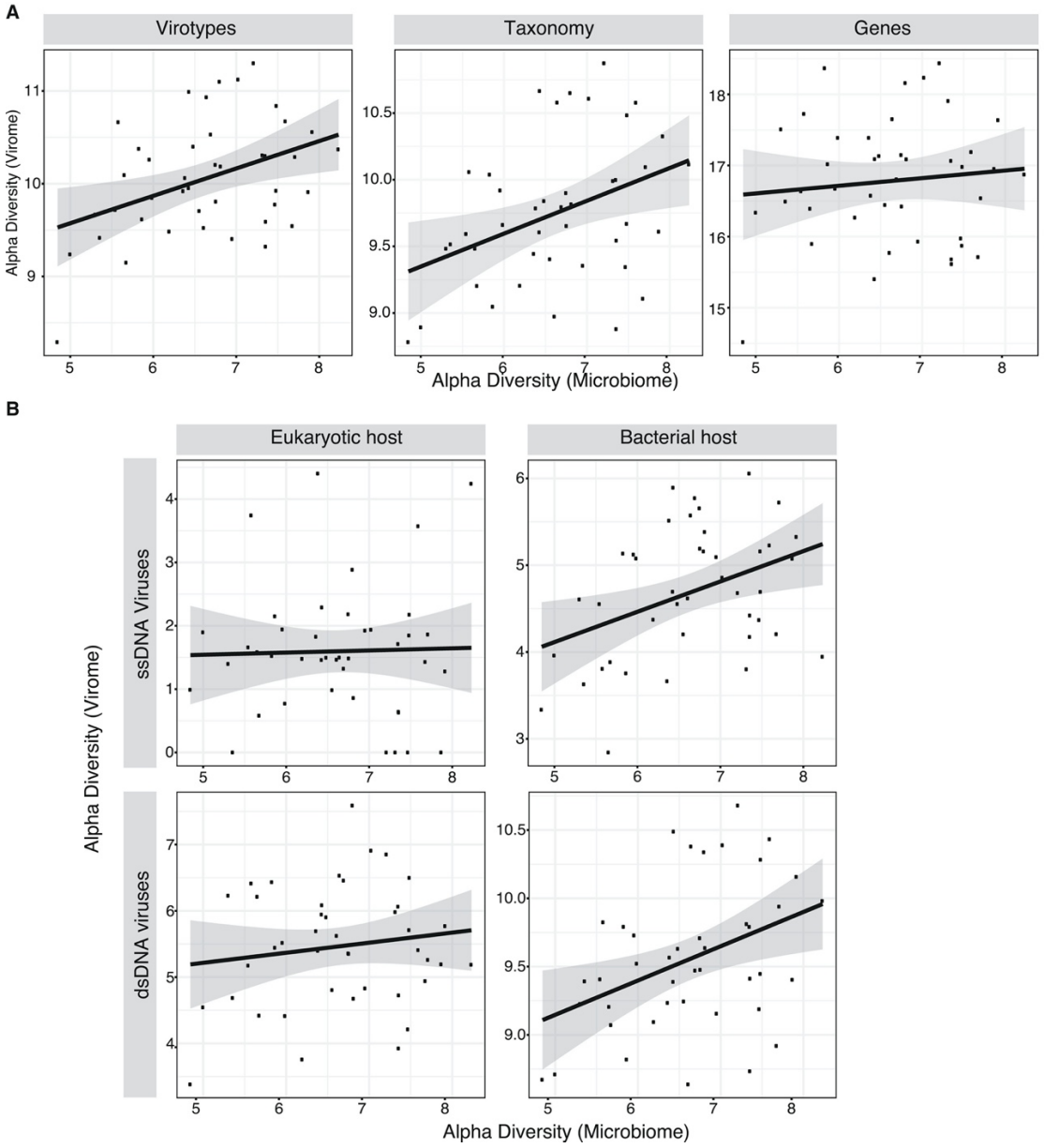


Figure 3.5: Bacteriophage Diversity Correlates with Microbiome Diversity, but Eukaryotic Virus Diversity Does Not

(A) Correlation of Shannon α-diversity of viromes to Shannon α-diversity of microbiomes (n = 42). Best-fit lines with 95% confidence intervals from linear regression are plotted. Virotypes: Pearson correlation coefficient = 0.406, m = 0.3, p = 0.007, R^2 = 0.165. Taxonomy: Pearson correlation coefficient = 0.389, m = 0.25, p = 0.010, R^2 = 0.151. Genes: Pearson correlation coefficient = 0.105, m = 0.11, p = 0.506, R^2 = 0.011.

(B) Correlation of the Shannon α-diversity of the virome, calculated from contigs annotated as ssDNA eukaryotic viruses, ssDNA phages, dsDNA eukaryotic viruses, and dsDNA phages, to Shannon α-diversity of the microbiome (n =

42). Best-fit lines with 95% confidence intervals from linear regression are plotted. ssDNA eukaryotic viruses: Pearson correlation coefficient = 0.027, m = 0.034, p = 0.863, R^2 = 0.000751. ssDNA bacteriophages: Pearson correlation coefficient = 0.394, m = 0.35, p = 0.009, R^2 = 0.155. dsDNA eukaryotic viruses: Pearson correlation coefficient = 0.143, m = 0.15, p = 0.368, R^2 = 0.020. dsDNA bacteriophages: Pearson correlation coefficient = 0.400, m = 0.25, p = 0.008, R^2 = 0.16.

Alpha-Diversity

α-diversities of the microbiome and the virome were positively correlated in two of the three layers of information used to test the correlation (virotypes and taxonomy annotated contigs, but not genes; Figure 3.5A). We used annotated contigs to ask about the α-diversity within subgroups of viruses (ssDNA eukaryotic, dsDNA eukaryotic, ssDNA bacteria, and dsDNA bacteria). Our results show that the diversity of eukaryotic viruses does not correlate with the microbiome α-diversity. In contrast, bacteriophage and microbiome α-diversity were positively correlated, for both ssDNA or dsDNA bacterial viruses (Figure 3.5B).

Beta-Diversity

We observed that concordant twins had lower virome β-diversity compared to discordant twins using Hellinger distances (Figure 3.6); the mean binary Jaccard distance and Bray-Curtis dissimilarity of viromes also showed the same trend (Supplementary Figure 3.5A,B). Similar to what

we observed with α-diversity, regardless of the layer of information used, the mean Hellinger distance of viromes within MZ twin pairs with concordant microbiomes was significantly lower than that of MZ twin pairs with discordant microbiomes ($p < 0.04$, Mann-Whitney's U test) (Figure 3.6). We did not observe significant differences in β-diversity when concordant twins or discordant twins were split by sex ($p > 0.05$, Mann-Whitney's U test). Still, any inference about sex influence is limited as the number of individuals per group is halved. Furthermore, a similar significant positive correlation was observed between microbiome and virome β-diversity when using the annotated

contigs. This relationship was driven by the bacteriophages ($p = 0.009$, Mann-Whitney's U test), but not the eukaryotic viruses ($p = 0.243$, Mann-Whitney's U test).

Finally, we compared the virome and microbiome pairwise distances among related (co-twins) and unrelated individuals. The pairwise distance matrices showed a positive correlation between virome and microbiome β-diversity measures not only within twin pairs (Pearson correlation coefficient > 0.50) but also generally across all individuals (Pearson correlation coefficient > 0.25; $p < 0.003$, Mantel test; Supplementary Figure 3.5C). These results show that regardless of genetic relatedness between hosts, individuals with more similar microbiomes harbor more similar viromes.

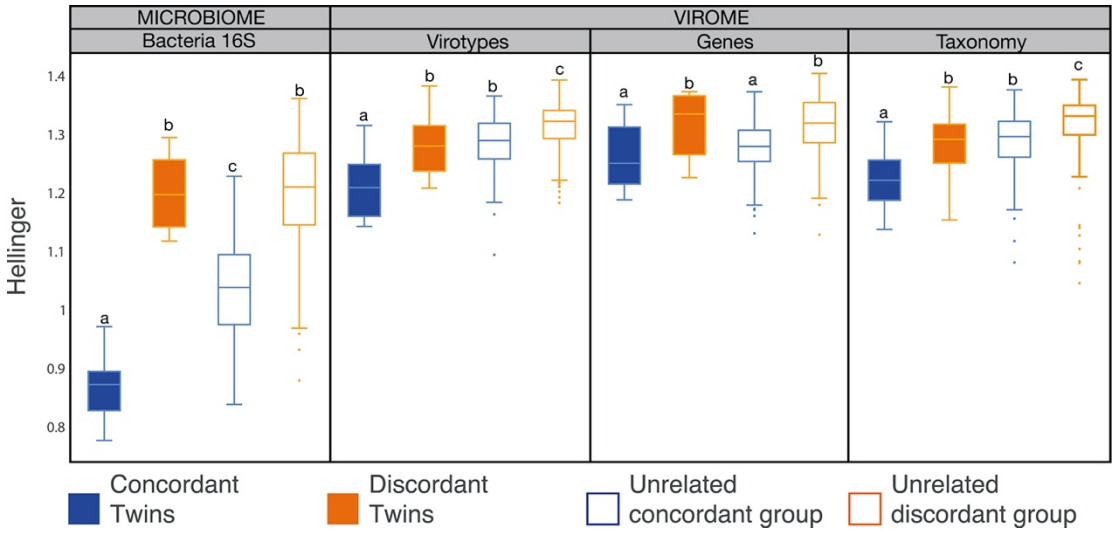


Figure 3.6: Virome Beta-Diversity Patterns Mirror Microbiome Beta-Diversity
Boxplots show the distribution of Hellinger distances for microbiomes and viromes, according to the three different layers of information recovered (virotypes, genes, and taxonomy), for concordant co-twins (solid blue, n = 9), discordant co-twins (solid orange, n = 12), unrelated samples within the concordant co-twins (blue outline, n = 144), and unrelated samples within the discordant co-twins (orange outline, n = 264). Significant differences between means (Mann-Whitney's U test, $p < 0.020$) are denoted with different letters.

3.4 Discussion

Co-twins, like other siblings, generally have more similar gut microbiomes within their twinships compared to unrelated individuals (Lee et al., 2011, Palmer et al., 2007, Tims et al., 2013, Turnbaugh et al., 2009, Yatsunenko et al., 2012). Moreover, MZ twins have overall more similar microbiomes than DZ twins, although at a whole-microbiome level this effect is small and primarily driven by a small set of heritable microbiota (Goodrich et al., 2014, Goodrich et al., 2016). Within a population of MZ twin pairs, however, the range of within-twin-pair differences in the microbiomes can be as great as for DZ twins (Goodrich et al., 2014). We took advantage of the large spread in β-diversity for MZ co-twins to select co-twins that were either highly concordant or discordant for their gut microbiomes. Our analysis of their viromes showed that despite the high variation in the gut viromes between individuals, and regardless of host relatedness, the more dissimilar their microbiomes, the more dissimilar their viromes. This pattern was driven by the bacteriophage component of the virome.

By choosing MZ twins from a distribution of β-diversities in the microbiomes, we removed host genetic relatedness as a variable possibly impacting the virome. Previous studies of the viromes and microbiomes of infant twin pairs showed that the microbiomes and viromes of co-twins were more similar than those of unrelated individuals, suggesting shared host genotype and/or environment were key (Lim et al., 2015, Reyes et al., 2015). In contrast, an early study of the virome of adult twins showed that adult co-twins did not have more similar viromes than unrelated individuals (Reyes et al., 2010); however, in light of the current study's results, this was likely a power

issue. Indeed, in our dataset we observed that regardless of whether twins were concordant or discordant for their microbiomes, co-twins had more similar viromes (virotypes and taxonomy) than unrelated individuals.

The previously reported greater virome similarity in young compared to adult twins has been related to the fact that infants have a greater shared environment compared to adult twins (Lim et al., 2015), particularly in terms of their diet. Minot et al. have also shown that individuals on the same diet have more similar gut viromes than individuals on dissimilar diets (Minot et al., 2011). It is well established that diet is a strong driver of daily microbiome fluctuation (Claesson et al., 2012, David et al., 2014, De Filippo et al., 2010, Wu et al., 2011), so the effect of diet on the virome is likely mediated by the microbiome. However, we did not control for diet, so it is possible that the microbiome discordance that we observe was caused by co-twins eating differently around the time of sampling. Regardless of what underlies the variance in microbiome concordance, it is strongly associated with virome concordance.

The relationship between virome richness and microbiome richness had not previously been directly addressed in adults. We observed that the α-diversity of the microbiome and the virome were positively correlated using two of the three layers of information describing virome diversity. Specifically, this pattern was observed for virotypes and taxonomy, but not for genes. However, since virome genes were observed to be enriched in only two categories, Genetic Information Processing and Nucleotide Metabolism, we would not expect differences in diversity of virome genes between subjects. The taxonomic annotation layer showed that the bacteriophage component of the virome, not the eukaryotic viruses, was driving this α-diversity correlation pattern.

The positive relationship between virome and microbiome α-diversity suggests that a greater availability of hosts drives a greater diversity of viruses. These observations are in accordance with the "piggyback the winner" model, which posits that in a dense environment, phages opt for a lysogenic cycle and multiply with their hosts (Knowles et al., 2016). Indeed, longitudinal studies of the human gut virome have reported genes associated with lysogeny, low mutation rate over time in temperate-like contigs, and long-term stability of the virome, suggesting preference for a lysogenic cycle (Minot et al., 2013, Reyes et al., 2010). Nevertheless, phage predation has been acknowledged as an important factor for the maintenance of highly diverse and efficient ecosystems (Rodriguez-Valera et al., 2009) and may play a role in the maintenance of diversity in a rapidly changing ecosystem as the human gut (David et al., 2014). Short-scale time series analyses of virome-microbiome interactions, along with a better understanding of the lysogenic-lytic switch in viral reproduction, would help to interpret the observed patterns in the human gut virome.

The composition of the viromes described here was similar to what has been previously reported for adult fecal viromes (Minot et al., 2011, Minot et al., 2013, Reyes et al., 2010). From the annotated fraction of the virome, the order *Caudovirales* and its families *Siphoviridae*, *Myoviridae*, and *Podoviridae*, along with crAssphage, were the dominant phages in all samples. Manrique et al. have summarized the phage colonization of the infant gut as follows: the eukaryotic viruses first dominate the newborn gut, followed by the *Caudovirales*, and by 2.5 years of age the *Microviridae* start to dominate (Manrique et al., 2017). We did observe abundant *Microviridae* in our sample set, but the *Caudovirales* were the dominant

group. Age was not related to patterns of diversity in the set of adult subjects studied here.

Despite the high diversity and uniqueness of each virome described here, we nonetheless recovered a core virome among the subjects: 18 contigs were present in all samples. More than half of these contigs were annotated as crAssphage, consistent with recent reports that this phage is widespread (Dutilh et al., 2014, Manrique et al., 2016, Yarygin et al., 2017). Other shared virotypes in our dataset were classified as *Myoviridae* and *Microviridae*. We also recovered contigs mapping to representative families of the nucleocytoplasmic large DNA viruses (NCLDVs), *Phycodnaviridae* and *Mimiviridae*. These types of viruses are increasingly reported as members of the human gut virome (Colson et al., 2013, Halary et al., 2016). A core set of bacteriophages consisting of nine representatives, including crAssphage, has previously been reported for the human gut (Manrique et al., 2016). Widely shared virotypes may indicate the wide sharing of specific hosts between individuals, or that these viruses have a broad host range within the human microbiome.

Our use of the HMMs to annotate viral contigs allowed a deep exploration into the taxonomic content of the virome. We annotated a diversity of contigs beyond what was revealed from comparisons to public databases, and also confirmed those annotations. Because each type of virus (e.g., family) requires its own HMM, we applied this method to a few key groups. When applied to the crAssphage, the HMM retrieved contigs that grouped only with sequences derived from fecal viromes and not with sequences from other environments (e.g., terrestrial or marine). This suggests that although crAssphage is a diverse group of bacteriophages, its diversity in the human gut is restricted to

sequences related to the reference crAssphage genome (Dutilh et al., 2014), the IAS virus reference (Shkoporov et al., 2018), or *Chlamydia* bacteriophage (Yutin et al., 2018). We also applied HHM to the family *Microviridae*, which are ssDNA bacteriophages. We were able to confirm the presence of diverse members of *Gokushovirinae* and Alpavirinae subfamilies. Although there is evidence that described Alpavirinae genomes constitute a third group of the *Microviridae* family (Krupovic and Forterre, 2011, Roux et al., 2012), they correspond to prophages, which makes it difficult to integrate them into the taxonomy of the International Committee on Taxonomy of Viruses (ICTV); thus, no contigs were annotated as Alpavirinae prior to application of the HMM profiles.

For each taxonomic group of viruses, there is a corresponding set of bacterial hosts. From the 16S rRNA gene diversity data we used to select the twin pairs, it is clear which bacteria phyla contribute the most to the differences in the microbiomes of concordant and discordant twins. But unlike for the bacteria, we were not able to discern such clear patterns by order or family in the virome. Indeed, most of the bacteriophage diversity is grouped in just one order, *Caudovirales*, and its three families *Myoviridae*, *Podoviridae*, and *Siphoviridae*. Representatives of these families can infect unrelated hosts (Barylski et al., 2017). Thus, we wouldn't necessarily expect specific orders or families of viruses to show the patterns observed in the bacterial phyla.

Finally, we noted an interesting pattern of complete bacterial genome coverage in the viromes for select bacterial species. As these putative contaminants were not the most abundant members of the microbiome, they are unlikely to represent random contamination by bulk DNA. Why certain bacterial genomes showed such high

coverage is unclear. One possibility is that we are observing the host species range of transposable phages. Phages such as the Mu phage randomly integrate into the host genome (Taylor, 1963), amplify by successive rounds of replicative transposition, and then can package any section of their host's genome (Hulo et al., 2011, Toussaint and Rice, 2017). Intriguingly, several of the contaminants detected here (e.g., *B. vulgatus*, *B. dorei*, *F. prausnitzii*, and *B. thetaiotaomicron*) have also been reported as contaminants in other human gut virome studies (Minot et al., 2011, Roux et al., 2013), which could indicate host specificity of transposable phages. Alternative explanations include vesicle production, gene transfer agents, and/or generalized transduction processes (Biller et al., 2014, McDaniel et al., 2010, Minot et al., 2011). Further comparisons of whole bacterial genomes recovered in diverse virome datasets may help shed light on their source, particularly if the same bacterial species are recovered across multiple studies.

Prospectus

Our results show that gut microbiome richness and diversity correlate to virome richness and diversity, and vice versa. The mechanisms underlying this association remain to be resolved for the human gut. This relationship may be useful to take into consideration when designing future studies of the virome and the factors that affect it. Baseline microbiome diversity may be important to balance between groups, for instance, prior to assessing the diversity of the virome.

3.5 Materials and Methods

3.5.1 Experimental Model and Subject Details:

Fecal Samples

Fecal samples used in this study were obtained as part of previous studies (Goodrich et al., 2014; Jackson et al., 2016). From 16S rRNA gene diversity previously measured for 354 monozygotic twin pairs whose fecal samples were received between January 28th 2013 and July 14th 2014 (Goodrich et al., 2014), we selected 9 concordant and 12 discordant MZ co-twins based on three microbiota β-diversity distances within twin pairs: unweighted UniFrac, weighted UniFrac (Lozupone et al., 2007) and Bray-Curtis (Bray and Curtis, 1957). Twins pairs in the concordant and the discordant groups were selected to be balanced between those two groups for sex, age, BMI, and BMI difference within a twin pair (Supplementary Table 4.1). Twins within the concordant group ranged in age from 23 to 77 years old and included 5 men and 4 women, while those in the discordant group ranged in age from 29 to 81 years old with 5 men and 7 women. All work involving the use of these previously collected samples was approved by the Cornell University IRB (Protocol ID 1108002388).

1.5.2 Experimental Methods:

Isolation of Virus-like Particles (VLPs) from Human Fecal Samples

VLP isolation procedures were based on the previously described protocols (Gudenkauf et al., 2014, Minot et al., 2013). For VLP isolation, ~0.5 g of fecal sample was resuspended by vortexing for 5-10 min in 15 mL PBS, previously filtered through 0.02 μm filter (Whatman). The homogenates were centrifuged for 30 min at 4,500 *x*g,

and the supernatant was filtered through 0.22 μm polyethersulfone (PES) Express Plus Millipore Stericup (150 ml) to remove cell debris and bacterial-sized particles. The filtrate was then concentrated on a Millipore Amicon Ultra-15 Centrifugal Filter Unit 100K to ∼1 ml. The concentrate was transferred to 5 Prime Phase Lock Gel and incubated with 200 μl chloroform for 10 min at room temperature. After being centrifuged for 1 min at 15,000 *x*g, the aqueous layer was transferred to a new microcentrifuge tube, and was treated with Invitrogen TURBO DNase (14 U), Promega RNase One (20 U) and 1 μl Benzonase Nuclease (E1014 Sigma Benzonase Nuclease) at 37°C for 3 hr (Gudenkauf and Hewson, 2016, Reyes et al., 2012). After incubation, 0.04 volumes 0.5 M EDTA was added to each sample. The samples were then stored at −80°C before further processing.

Viral DNA Shotgun Sequencing

The viral DNA was extracted with PureLink Viral RNA/DNA Mini Kit from Invitrogen. Each viral DNA sample was then amplified using GenomePlex Complete Whole Genome Amplification (WGA2) Kit from Sigma-Aldrich (Gudenkauf and Hewson, 2016). Two blank controls were included in this step, but very low yield precluded library construction. The amplified product was then fragmented with Covaris S2 Adaptive Focused Acoustic Disruptor with the parameters set as follows: the duty cycle set at 10%, cycle per burst 200, intensity 4 and duration 60 s. Each viral sequencing library was prepared following Illumina TruSeq DNA Preparation Protocol with one unique barcode per sample. All barcoded libraries were pooled together. Half of the pool was size selected by BluePippin (Sage Science, Beverly, MA, USA) to enrich fragments with longer inserts (425 bp to 875 bp including the adapters). Both

pools, the “large-insert-size library” and the “short-insert-size library,” were sequenced in independent lanes on an Illumina HiSeq 2500 instrument, operating in Rapid Run Mode with 250 bp paired-end chemistry at the Cornell Biotechnology Resource Center Genomics Facility.

Whole Fecal Metagenome Shotgun Sequencing

The genomic DNA was isolated from an aliquot of ~100 mg from each sample using the PowerSoil® - htp DNA isolation kit (MoBio Laboratories Ltd, Carlsbad, CA). Each sequencing library was then prepared following Illumina TruSeq DNA Preparation Protocol with 500 ng DNA using the gel-free method, 14 cycles of PCR, and with one unique barcode per sample. Sequencing was performed on an Illumina HiSeq 2500 instrument in Rapid Run mode with 2x150 bp paired-end chemistry at the Cornell Biotechnology Resource Center Genomics Facility.

1.5.3 Data analysis:

Assessment of Bacterial Contamination

A set of 8,163 finished bacterial genomes was retrieved from the NCBI FTP on 21 February 2017. Reads per sample were mapped against this bacterial genomes dataset using Bowtie2 v.2.2.8 (Langmead and Salzberg, 2012) with the following parameters:–local–maxins 800 -k = 3. Genome coverage per base was calculated considering only reads with a mapping quality above 20 using *view* and *depth* Samtools commands v.1.5 (Li et al., 2009). Next, genome coverage was averaged for 100Kbp bins. We observed that evenly covered genomes had a median bin coverage of at least 100; those genomes with a median bin coverage greater than 100 were considered as

contaminants. The reads mapping to those genomes were removed. Bacterial genomes can have one or more prophage(s) in their genomes (Munson-McGee et al., 2018); bursting events of those prophages can occur, generating several VLPs. As a conservative measure to avoid the loss of reads originating from prophages and not the bacterial genome per se, bins with a coverage over three standard deviations of the bacterial mean coverage were also identified and cataloged as prophages-like regions. Reads mapping to potential contaminant genomes were tagged as "contaminants" and removed from further analysis while reads mapping to high coverage bins were tagged as "possible prophages."

A matrix of the abundance of each potential contaminant per sample was built using an in-house Python script and normalized by RPKM. In parallel, from Goodrich et al. data (Goodrich et al., 2014), the relative abundance of each OTU was recovered and summarized at the species level using summarize_taxa.py qiime script. The Spearman rank order correlation between relative abundances of contaminants and their corresponding 16S rRNAs data was calculated for species in both sets.

Functional Profiles

The joined and trimmed reads from the "short-insert-size library" were mapped onto Integrated Gene Catalogs (IGC), an integrated catalog of reference genes in the human gut microbiome (Li et al., 2014) by BLASTX using DIAMOND v.0.7.5 (Buchfink et al., 2015) with maximum e-value cutoff 0.001, and maximum number of target sequences to report set to 25.

After the mapping onto IGC, an abundance matrix was generated using an in-house Python script. The matrix was then annotated according to the KEGG annotation

of each gene provided by IGC. The annotated abundance matrix was rarefied (subsampling without replacement) to 2,000,000 read hits per sample. The KEGG functional profile was then generated using QIIME 1.9 (Quantitative Insights Into Microbial Ecology) (Caporaso et al., 2010) using the command summarize_taxa_through_plots.py. The Intraclass Correlation Coefficient of the functional profiles for each group (additional microbiomes, additional viromes, viromes of concordant-microbiome samples and viromes of discordant-microbiome samples) was calculated using the Psych R package.

***De novo* Assembly**

Reads from the "large-insert-size library" that remain paired (forward and reverse) after the trimming step were assembled using the Integrated metagenomic assembly pipeline for short reads (InteMAP) (Lai et al., 2015) with insert size 325 bp ± 100 bp. Each sample was assembled separately. After the first run of assembly, all clean reads were mapped to the assembled contigs using Bowtie2 v.2.2.8 (Langmead and Salzberg, 2012) with the following parameters --local --maxins 800. The pairs of reads that aligned concordantly at least once were then submitted for the second run of assembly by InteMAP. Contigs larger than 500 bp from all samples were pooled together and compared all versus all, using an in-house Perl script. From this analysis, it was possible to identify potential circular genomes, and to dereplicate contigs that were contained in over 90% of their length within another contig.

The recruitment of reads to the dereplicated metagenomic assemblies was used to build an abundance matrix, applying a filter of coverage and length as recommended in Roux et al., 2017. Reads (not tagged as contaminants in the previous step) were

mapped to dereplicated contigs using Rsubread v.1.28.0 (Liao et al., 2013). Mapping outputs were parsed using an in-house Python script into an abundance matrix that was normalized by reads per kilobase of contig length per million sequenced reads per sample (RPKM) and transformed to $Log_{10}(x+1)$, x being the normalized abundance. Contigs with a normalized coverage bellow 5x were excluded. Finally, a filter on contig length was applied to obtain virotypes. A length threshold was chosen as the elbow of the decay curve generated when plotting the number of contigs as a function of length, which occurred at a length of 1,300 bp.

HMM Annotation

Independent HMM profiles were built to identify crAss-like contigs and *Microviridae* contigs. To build the HMM-crAsslike profile, sequences for the Major Capsid Protein (MCP) of the proposed crAss-like family (Yutin et al., 2018) were retrieved from https://ftp.ncbi.nih.gov/pub/yutinn/crassphage_2017/. Multiple sequence alignments (MSA) were done using MUSCLE v.3.8.31 (Edgar, 2004) and inspected using UGENE v.1.31.0 (Okonechnikov et al., 2012); positions with more than 30% of gaps were removed. Finally, the HMM-crAsslike profile was built using *hmmbuild* from the HMMER package v.3.1b2 (http://hmmer.org/) (Eddy, 1998). For the *Microviridae* case, all HMM-profiles for the viral protein 1 (VP1) developed by Alves et al., 2016 were adopted.

Predicted proteins of the assembled contigs were queried for matching the HMM-profiles using *hmmsearch* (Eddy, 1998). Matching proteins with an e-value below $1x10^{-5}$ were considered as true homologs but only proteins between the size rank of the reference proteins (crAsslike MCP: 450-510 residues; *Microviridae*: 450-800 residues),

a coverage of at least 50% and a percentage of identity of at least 40% to at least one reference sequence were used for further analysis. Coverage and identity percentages were determined with a BLASTp search of the true homologs against the reference sequences.

True homologs passing the filters mentioned above were used in phylogenetic analysis. Reference and homologous sequences were aligned using MUSCLE v.3.8.31 and sites with at least 30% of gaps were removed using UGENE v.1.31.0. A maximum-likelihood (ML) phylogenetic analysis was done using RAxML v.8.2.4 (Stamatakis, 2014), the best model of evolution was obtained with prottest v.3.4.2 (Darriba et al., 2011) and support for nodes in the ML trees were obtained by bootstrap with 100 pseudoreplicates.

Taxonomic Profiles

To infer the taxonomic affiliation of the assembled VLPs, genes were predicted from all assembled contigs larger than 500 bp using GeneMarkS v.4.32 (Besemer et al., 2001). The amino acid sequence of the predicted genes was then used in a BLASTp search against the NR NCBI viral database using DIAMOND v.0.7.5 (Buchfink et al., 2015) with maximum e-value cutoff 0.001 and maximum number of target sequences to report set to 25. Using the BLASTp results, the taxonomy of each gene was assigned by the lowest-common-ancestor algorithm in MEtaGenome ANalyzer (MEGAN5) v.5.11.3 (Huson et al., 2011) with the following parameters: Min Support: 1, Min Score: 40.0, Max Expected: 0.01, Top Percent: 10.0, Min-Complexity filter: 0.44. Independently, the taxonomy annotation of each contig was obtained using CENTRIFUGE v.1.0.4 (Kim et al., 2016) against the NT NCBI viral genomes database.

The final taxonomic annotation of each contig was then assigned using a voting system where the taxonomic annotation of each protein and the CENTRIFUGE annotation of the contig were considered as votes. With all the possible votes for a contig, an N-ary tree was built and the weight of each node was the number of votes including that node. The taxonomic annotation of a contig will be the result of traversing the tree passing through the heaviest nodes with one consideration: if all children-nodes of a node have the same weight the traversing must be stopped. The taxonomic profile was considered as a subset of the recruitment matrix containing all contigs annotated either by the voting system or annotated through the HMM profiles (see above).

Prediction of Phage-Host Interaction

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) were identified using the PilerCR program v.1.06 (Edgar, 2007) from the same set of 8,163 bacterial used to assess the bacterial contamination. Spacers within the expected size of 20 bp and 72 bp (Horvath and Barrangou, 2010) were used as queries against virotypes and taxonomically annotated contigs using BLASTn (v.2.6.0+) with short query parameters (Camacho et al., 2009). Matches covering at least 90% of the spacer and with an e-value < 0.001 were considered to be CRISPR spacer-virus associations. Additionally, virotypes and taxonomically annotated contigs were mapped against the representatives genomes of the viral clusters in the MVP database (Gao et al., 2018) using LAST-959 (Kiełbasa et al., 2011). As viral clusters in MVP comprise sequences that have at least 95% identity along at least 80% of their lengths, only matches that fulfill those constraints were kept. The host(s) of a contig was determined from its matching viral cluster.

Diversity Indexes

The Shannon diversity index within-samples (α-diversity) and the Hellinger distance within co-twins (β-diversity) were calculated using *diversity* and *vegdist* functions of Vegan R package for all three abundance matrices generated (function, taxonomy and read recruitment matrices). Correlations between virome α-diversity and microbiome α-diversity were measured using the Pearson correlation coefficient. Correlations between viromes β-diversity and the microbiomes β-diversity were computed with the Mantel test using the Pearson correlation coefficient. Additionally, the β-diversity between concordant MZ co-twins was compared to the β-diversity between discordant MZ co-twins; p values were calculated using a Mann-Whitney U test.

Quantification and Statistical Analysis

The number of twins/individuals in each group (Figures 1.1C, 1.4B, 1.6, and Supplementary Figure 3.5A,B) or the number of comparisons (Figure 3.5, Supplementary Figures 1.2 and 1.5C) is denoted using *"n"*; p values were obtained using Mann-Whitney U test or Mantel test using the Python library "scipy"; correlation coefficients were measured as the Pearson correlation coefficient using the Python library "scipy"; alpha and beta-diversity metrics were calculated with the R package "vegan"; Intra-class coefficient was calculated using the R package "psych"; maximum-likelihood phylogenetic analysis was done using RAxML.

Data and Software Availability

Jupyter notebooks and scripts describing the data analysis process are available on GitHub at https://github.com/leylabmpi/TwinsUK_virome. The sequence data have

been deposited in the European Nucleotide Archive under the study accession number ENA: PRJEB29491.

3.6 Acknowledgments

We thank Laura Avellaneda-Franco for helpful discussions. This work was supported by the Max Planck Society , by grants from the NIH (NIDDK RO1 DK093595 and DP2 OD007444), and by a Fellowship in Science and Engineering to R.E.L. from the David and Lucile Packard Foundation . The TwinsUK cohort is supported by the Wellcome Trust , European Community's Seventh Framework Programme (FP7/2007–2013), National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility, and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

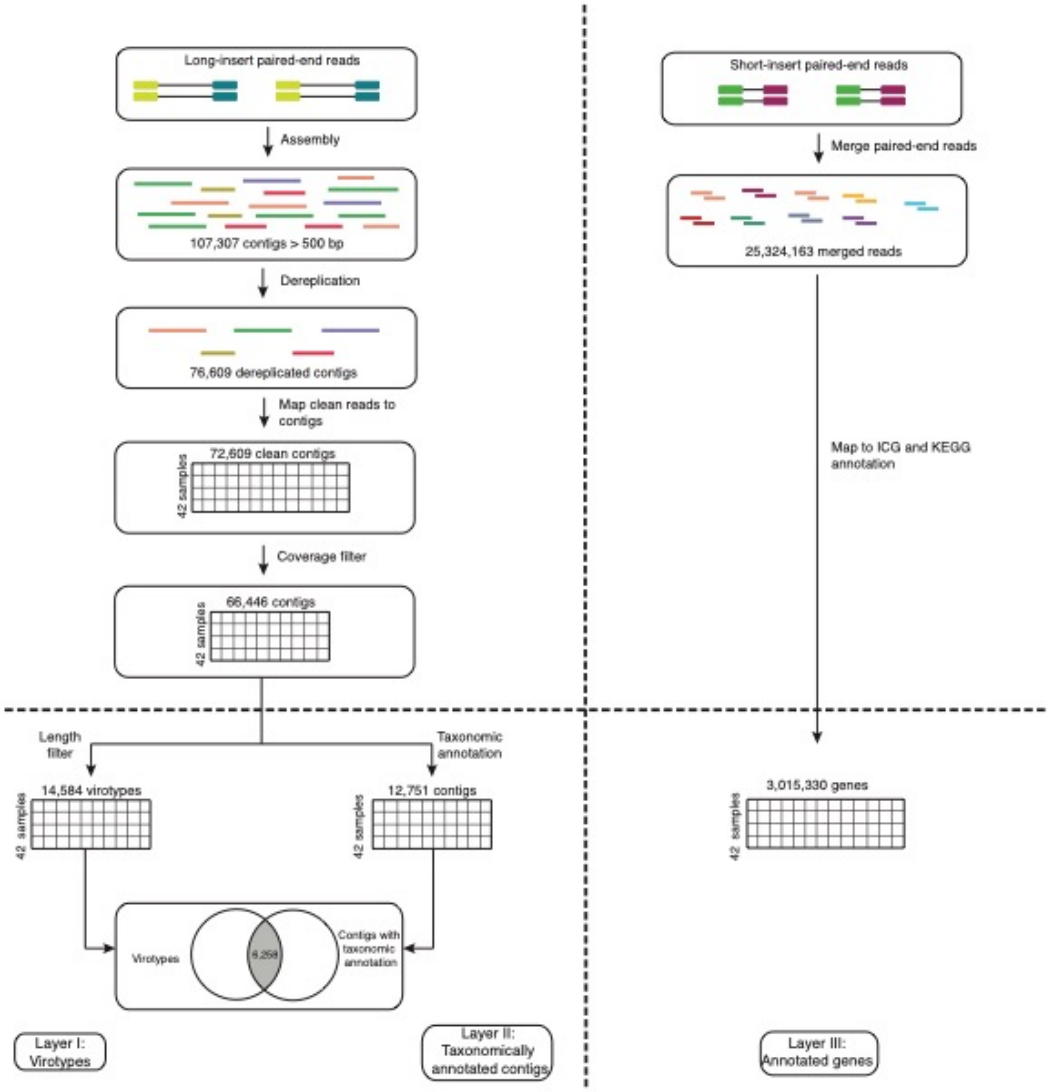
Declaration of Interests

The authors declare no competing interests.

Author Contributions

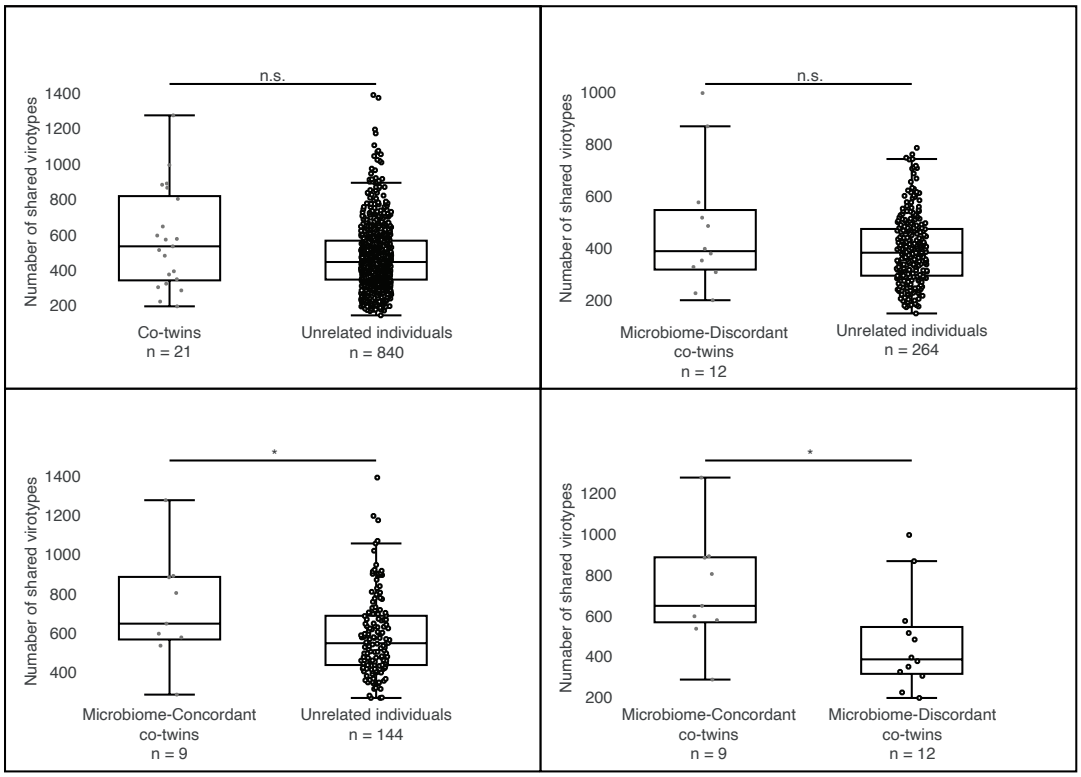
R.E.L. and S.-P.C. designed the study. T.D.S. and J.T.B. were involved in sample collection. S.-P.C. and I.H. generated the data. J.L.M.-G., S.-P.C., J.K.G., N.D.Y., A.R., and R.E.L. analyzed the data. J.L.M.-G., S.-P.C., S.C.D., A.R., and R.E.L. wrote the manuscript. All authors read and approved the final manuscript.

3.7 Supplementary Figures and Tables



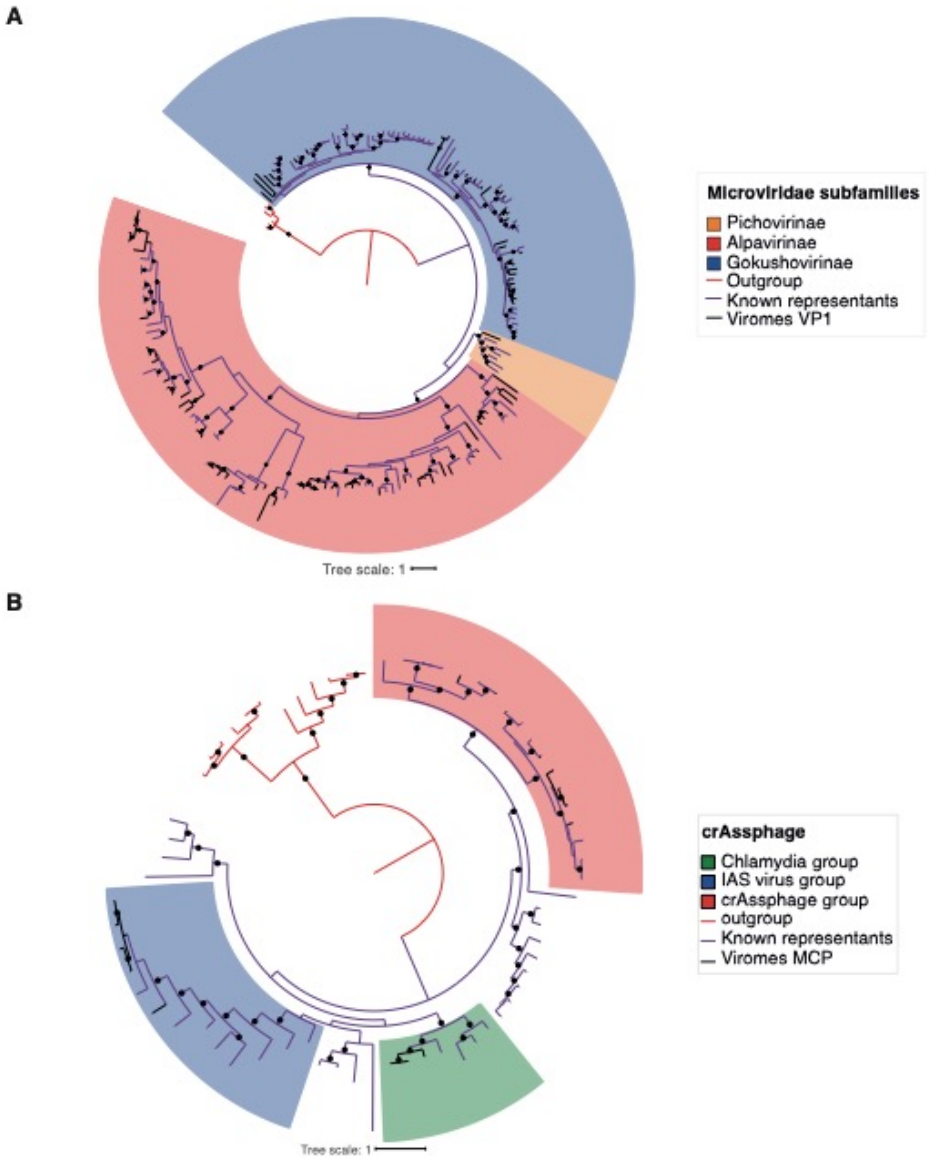
Supplementary Figure 3.1:

Related to Figure 3.1. Schematic representation summarizing the procedures applied to (left) the “large-insert-size library” and (right) the “short-insert-size library” to obtain three different layers of information used to analyze the virome diversity of the microbiome-concordant and microbiome-discordant co-twins.



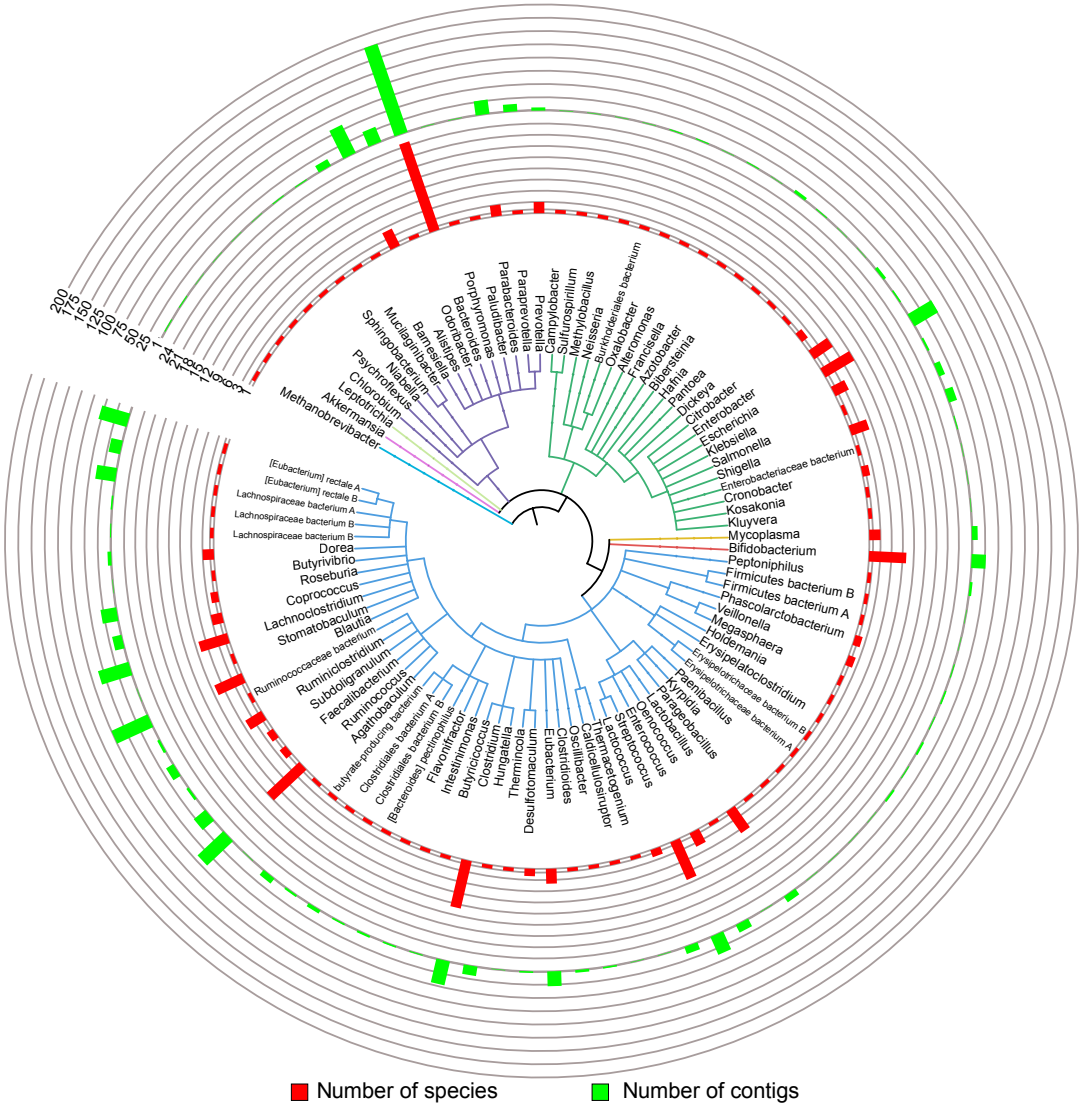
Supplementary Figure 3.2:

Related to Figure 3.4. Box plots showing the distribution of the number of shared virotypes between different groups made from the 21 MZ co-twins. (Upper left) All co-twins vs unrelated individuals. (Upper right) Microbiome-discordant co-twins vs unrelated individuals in the same group. (Lower left) Microbiome-concordant co-twins vs unrelated individuals in the same group. (Lower right) Microbiome-concordant co-twins vs microbiome-discordant co-twins. Mann-Whitney's U test. * $p < 0.05$; n.s: no significant difference.



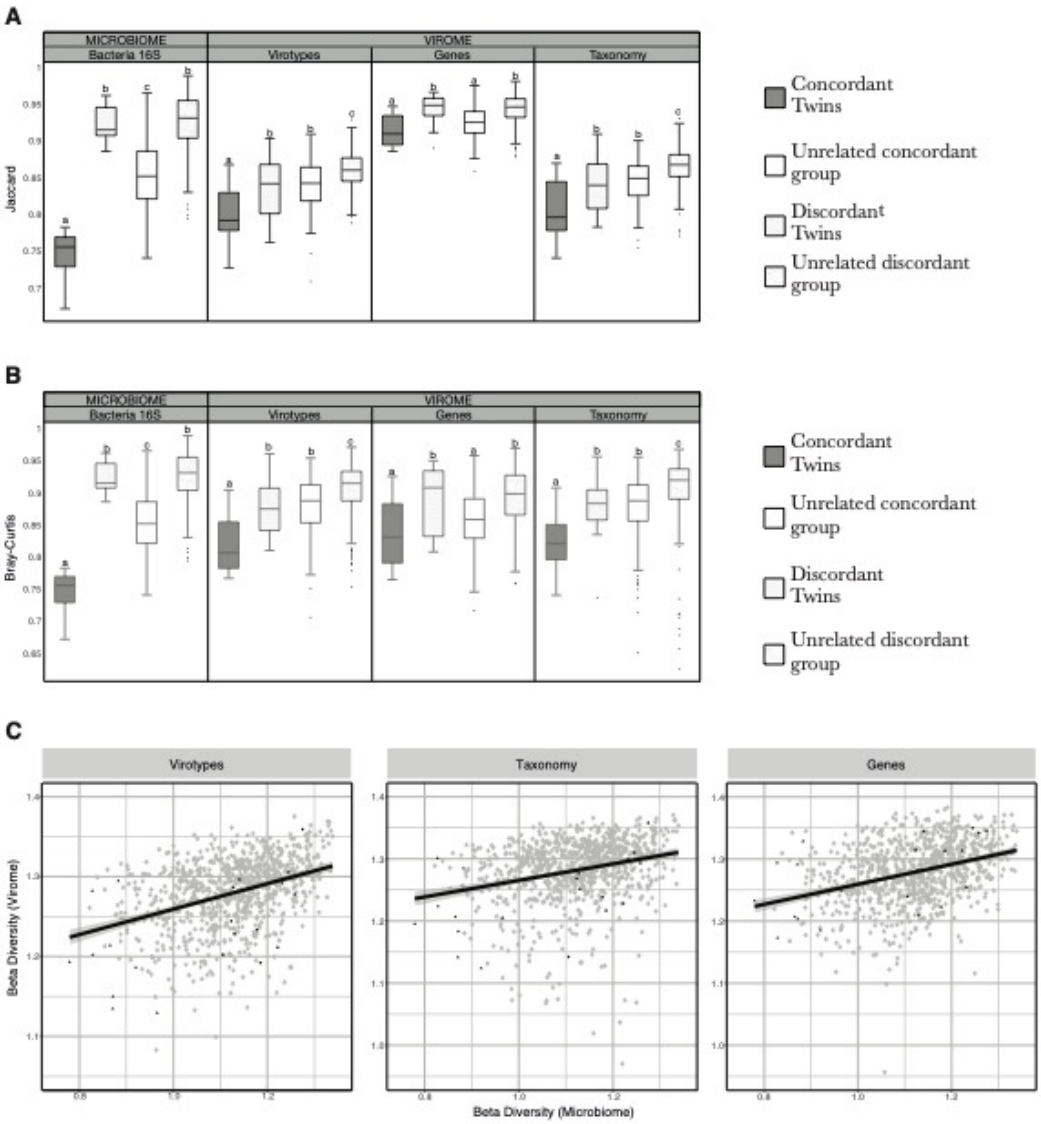
Supplementary Figure 3.3:

Related to Figure 3.4. Maximum likelihood phylogenetic analysis of (A) the VP1 protein of *Microviridae* phages and (B) the MCP protein of crAssphage found in the 42 MZ viromes. Reference sequences are in purple, outgroup sequences are in red while the different MCP or VP1 proteins found in this work are labeled in black. Circles in the nodes indicate bootstrap values above 70%. Scale: Average substitutions per site.



Supplementary Figure 3.4:

Related to *phage-host interaction prediction methods*. Cladogram based on the NCBI taxonomy showing the bacteria identified as hosts. The cladogram is summarized by genus, and clades are colored by Phylum. Blue: Firmicutes; Red: Actinobacteria; Yellow: Tenericutes; Green: Proteobacteria; Purple: Bacteroidetes; Light green: Fusobacteria; Magenta: Verrucomicrobia; Light blue: Euryarchaeota. Red bars indicate the number of species in each genus, and green bars show the dereplicated number of contigs associated to each genus (i.e., if a contig was found associated to two species in that genus, it is only shown one time).



Supplementary Figure 3.5:

Related to Figure 3.5 and Figure 3.6. Box plots showing the distribution of (A) the Jaccard distances and (B) Bray-Curtis distances for microbiomes and viromes, according to the three different layers of information recovered (virotypes, genes and taxonomy). Significant differences between means (Mann-Whitney's U test) are denoted with different letters. Groups and n values as in Figure 3.6. (C) Correlation between virome β-diversity and microbiome β-diversity (n=840). Virotypes: Pearson correlation coefficient among all individuals = 0.382 (p = 0.0005, Mantel test), m = 0.167, p = 0, R^2 = 0.157; Pearson correlation coefficient among co-twins = 0.522, m = 0.188, p = 0.015, R^2 = 0.1508; Taxonomy annotated contigs: Pearson correlation coefficient among all individuals = 0.266 (p = 0.003, Mantel test), m = 0.140, p = 0, R^2 = 0.0796; Pearson correlation coefficient among co-twins = 0.512, m = 0.186, p = 0.017, R^2 = 0.224; Genes: Pearson correlation coefficient among all

individuals = 0.344 ($p = 0.0009$, Mantel test), $m = 0.162$, $p = 0$, $R^2 = 0.123$; Pearson correlation coefficient among co-twins = 0.53, $m = 0.182$, $p = 0.012$, $R^2 = 0.248$. Lines describe linear regressions of pairwise distances among all individuals. Triangles indicate concordant-microbiome co-twins and squares indicate discordant-microbiome co-twins.

Supplementary Table 4.1:

Related to Figure 3.1. Additional information pertaining to the 21 selected MZ twin pairs.

Sample ID	Category	BMI	Gender	Age (years)
1A	Concordant	30.30	F	63
1B	Concordant	32.93	F	63
2A	Concordant	29.86	M	68
2B	Concordant	33.32	M	68
3A	Concordant	21.57	M	36
3B	Concordant	20.37	M	36
4A	Concordant	20.35	F	64
4B	Concordant	21.22	F	64
5A	Concordant	21.24	F	23
5B	Concordant	20.30	F	23
6A	Concordant	24.61	F	56
6B	Concordant	28.72	F	56
7A	Concordant	20.13	M	25
7B	Concordant	28.62	M	25
8A	Concordant	17.90	M	77
8B	Concordant	18.76	M	77
9A	Concordant	25.13	M	57
9B	Concordant	28.60	M	57
10A	Discordant	27.97	M	64
10B	Discordant	29.58	M	64
11A	Discordant	24.33	F	81
11B	Discordant	24.71	F	81
12A	Discordant	41.16	M	49
12B	Discordant	40.25	M	49
13A	Discordant	24.57	F	66
13B	Discordant	26.93	F	66

14A	Discordant	19.28	F	43
14B	Discordant	21.30	F	43
15A	Discordant	29.23	F	59
15B	Discordant	23.77	F	59
16A	Discordant	32.75	F	55
16B	Discordant	27.93	F	55
17A	Discordant	19.83	M	29
17B	Discordant	18.58	M	29
18A	Discordant	24.93	M	63
18B	Discordant	29.48	M	63
19A	Discordant	18.73	F	52
19B	Discordant	18.29	F	52
20A	Discordant	21.13	M	64
20B	Discordant	24.40	M	64
21A	Discordant	22.01	F	62
21B	Discordant	20.57	F	62

REFERENCES

Alves, J.M.P., de Oliveira, A.L., Sandberg, T.O.M., Moreno-Gallego, J.L., de Toledo, M.A.F., de Moura, E.M.M., Oliveira, L.S., Durham, A.M., Mehnert, D.U., Zanotto, P.M. de A., et al. (2016). GenSeed-HMM: A tool for progressive assembly using profile HMMs as seeds and its application in Alpavirinae viral discovery from metagenomic data. Front. Microbiol. *7*, 269.

Barylski, J., Enault, F., Dutilh, B.E., Schuller, M.B.P., Edwards, R.A., Gillis, A., Klumpp, J., Knezevic, P., Krupovic, M., Kuhn, J.H., et al. (2017). Genomic, proteomic, and phylogenetic analysis of spounaviruses indicates paraphyly of the order Caudovirales. bioRxiv. doi: https://doi.org/10.1101/220434

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. *29*, 2607–2618.

Biller, S.J., Schubotz, F., Roggensack, S.E., Thompson, A.W., Summons, R.E., and Chisholm, S.W. (2014). Bacterial vesicles in marine ecosystems. Science *343*, 183–186.

Bray, J.R., and Curtis, J.T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol. Monogr. *27*, 326–349.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. *185*, 6220–6223.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R.A., Felts, B., Mahaffy, J.M., Mueller, J., Nulton, J., Rayhawk, S., et al. (2008). Viral diversity and dynamics in an infant gut. Res. Microbiol. *159*, 367–373.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Castro-Mejía, J.L., Muhammed, M.K., Kot, W., Neve, H., Franz, C.M.A.P., Hansen, L.H., Vogensen, F.K., and Nielsen, D.S. (2015). Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. Microbiome *3*, 64.

Claesson, M.J., Jeffery, I.B., Conde, S., Power, S.E., O'Connor, E.M., Cusack, S., Harris, H.M.B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., et al. (2012). Gut microbiota composition correlates with diet and health in the elderly. Nature *488*, 178–184.

Colson, P., Fancello, L., Gimenez, G., Armougom, F., Desnues, C., Fournous, G.,

Yoosuf, N., Million, M., La Scola, B., and Raoult, D. (2013). Evidence of the megavirome in humans. J. Clin. Virol. *57*, 191–200.

Cotillard, A., Kennedy, S.P., Kong, L.C., Prifti, E., Pons, N., Le Chatelier, E., Almeida, M., Quinquis, B., Levenez, F., Galleron, N., et al. (2013). Dietary intervention impact on gut microbial gene richness. Nature *500*, 585–588.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics *27*, 1164–1165.

David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman, S.E., and Alm, E.J. (2014). Host lifestyle affects human microbiota on daily timescales. Genome Biol. *15*, R89.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. U. S. A. *107*, 14691–14696.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. *5*, 4498.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and

high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics *8*, 18.

Gao, N.L., Zhang, C., Zhang, Z., Hu, S., Lercher, M.J., Zhao, X.-M., Bork, P., Liu, Z., and Chen, W.-H. (2018). MVP: a microbe–phage interaction database. Nucleic Acids Res. *46*, D700–D707.

Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., et al. (2014). Human genetics shape the gut microbiome. Cell *159*, 789–799.

Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M.A., Knight, R., Ober, C., Spector, T.D., Bell, J.T., Clark, A.G., and Ley, R.E. (2016). Genetic determinants of the gut microbiome in UK Twins. Cell Host Microbe *19*, 731–743.

Gudenkauf, B.M., and Hewson, I. (2016). Comparative metagenomics of viral assemblages inhabiting four phyla of marine invertebrates. Frontiers in Marine Science *3*, 23.

Gudenkauf, B.M., Eaglesham, J.B., Aragundi, W.M., and Hewson, I. (2014). Discovery of urchin-associated densoviruses (family Parvoviridae) in coastal waters of the Big Island, Hawaii. J. Gen. Virol. *95*, 652–658.

Halary, S., Temmam, S., Raoult, D., and Desnues, C. (2016). Viral metagenomics: are we missing the giants? Curr. Opin. Microbiol. *31*, 34–43.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.

Hoyles, L., McCartney, A.L., Neve, H., Gibson, G.R., Sanderson, J.D., Heller, K.J., and van Sinderen, D. (2014). Characterization of virus-like particles associated with the human faecal and caecal microbiota. Res. Microbiol. *165*, 803–812.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., and Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. *39*, D576–D582.

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Res. *21*, 1552–1560.

Jackson, M.A., Goodrich, J.K., Maxan, M.-E., Freedberg, D.E., Abrams, J.A., Poole, A.C., Sutter, J.L., Welter, D., Ley, R.E., Bell, J.T., et al. (2016). Proton pump inhibitors alter the composition of the gut microbiota. Gut *65*, 749–756.

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res. *21*, 487–493.

Kim, D., Song, L., Breitwieser, F.P., and Salzberg, S.L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. *26*, 1721–1729.

Krupovic, M., and Forterre, P. (2011). Microviridae goes temperate: microvirus-related proviruses reside in the genomes of Bacteroidetes. PLoS One *6*, e19893.

Knowles, B., Silveira, C.B., Bailey, B.A., Barott, K., Cantu, V.A., Cobián-Güemes, A.G., Coutinho, F.H., Dinsdale, E.A., Felts, B., Furby, K.A., et al. (2016). Lytic to temperate switching of viral communities. Nature *531*, 466–470.

Lai, B., Wang, F., Wang, X., Duan, L., and Zhu, H. (2015). InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. BMC Bioinformatics *16*, 1–14.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lee, S., Sung, J., Lee, J., and Ko, G. (2011). Comparison of the gut microbiotas of healthy adult twins living in South Korea and the United States. Appl. Environ. Microbiol. *77*, 7433–7437.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. *32*, 834–841.

Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. *41*, e108.

Lim, E.S., Zhou, Y., Zhao, G., Bauer, I.K., Droit, L., Ndao, I.M., Warner, B.B., Tarr, P.I., Wang, D., and Holtz, L.R. (2015). Early life dynamics of the human gut virome

and bacterial microbiome in infants. Nat. Med. *21*, 1228–1234.

Lozupone, C.A., Hamady, M., Kelley, S.T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl. Environ. Microbiol. *73*, 1576–1585.

Manrique, P., Bolduc, B., Walk, S.T., van der Oost, J., de Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. Proc. Natl. Acad. Sci. U. S. A. *113*, 10400–10405.

Manrique, P., Dills, M., and Young, M.J. (2017). The human gut phage community and its implications for health and disease. Viruses *9*, 10.3390/v9060141.

McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B., and Paul, J.H. (2010). High frequency of horizontal gene transfer in the oceans. Science *330*, 50.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. *21*, 1616–1625.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. U. S. A. *110*, 12450–12455.

Munson-McGee, J.H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R.J., Weitz, J.S., and Young, M.J. (2018). A virus or more in (nearly) every cell: ubiquitous networks of virus-host interactions in extreme environments. ISME J.

Ogilvie, L.A., and Jones, B.V. (2017). The human gut virome: form and function. Emerging Topics in Life Sciences *1*, 351–362.

Okonechnikov, K., Golosova, O., Fursov, M., and UGENE team (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics *28*, 1166–1167.

Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A., and Brown, P.O. (2007). Development of the human infant intestinal microbiota. PLoS Biol. *5*, e177.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature *466*, 334–338.

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., and Gordon, J.I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. Nat. Rev. Microbiol. *10*, 607–617.

Reyes, A., Wu, M., McNulty, N.P., Rohwer, F.L., and Gordon, J.I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc. Natl. Acad. Sci. U. S. A. *110*, 20236–20241.

Reyes, A., Blanton, L.V., Cao, S., Zhao, G., Manary, M., Trehan, I., Smith, M.I., Wang, D., Virgin, H.W., Rohwer, F., et al. (2015). Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc. Natl. Acad. Sci. U. S. A. *112*, 11941–11946.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan,

J., Desnues, C., Dinsdale, E., Edwards, R., et al. (2010). Viral and microbial community dynamics in four aquatic environments. ISME J. *4*, 739–751.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasić, L., Thingstad, T.F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. *7*, 828–836.

Roux, S., Krupovic, M., Poulet, A., Debroas, D., and Enault, F. (2012). Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. PLoS One *7*, e40418.

Roux, S., Krupovic, M., Debroas, D., Forterre, P., and Enault, F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol. *3*, 130160.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ *5*, e3817.

Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. Cell *164*, 337–340.

Shkoporov, A., Khokhlova, E.V., Brian Fitzgerald, C., Stockdale, S.R., Draper, L.A., Paul Ross, R., and Hill, C. (2018). ΦCrAss001, a member of the most abundant bacteriophage family in the human gut, infects Bacteroides. bioRxiv. doi: https://doi.org/10.1101/354837

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Suttle, C.A. (2007). Marine viruses--major players in the global ecosystem. Nat. Rev. Microbiol. *5*, 801–812.

Székely, A.J., and Breitbart, M. (2016). Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. FEMS Microbiol. Lett. *363*.

Taylor, A.L. (1963). Bacteriophage-induced mutation in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *50*, 1043–1051.

Thingstad, T.F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol. Oceanogr. *45*, 1320–1328.

Thingstad, T.F., Våge, S., Storesund, J.E., Sandaa, R.-A., and Giske, J. (2014). A theoretical analysis of how strain-specific viruses can control microbial species diversity. Proc. Natl. Acad. Sci. U. S. A. *111*, 7813–7818.

Tims, S., Derom, C., Jonkers, D.M., Vlietinck, R., Saris, W.H., Kleerebezem, M., de Vos, W.M., and Zoetendal, E.G. (2013). Microbiota conservation and BMI signatures in adult monozygotic twins. ISME J. *7*, 707–717.

Toussaint, A., and Rice, P.A. (2017). Transposable phages, DNA reorganization and transfer. Curr. Opin. Microbiol. *38*, 88–94.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Weitz, J.S., and Dushoff, J. (2008). Alternative stable states in host–phage dynamics. Theor. Ecol. *1*, 13–19.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science *334*, 105–108.

Yarygin, K., Tyakht, A., Larin, A., Kostryukova, E., Kolchenko, S., Bitner, V., and Alexeev, D. (2017). Abundance profiling of specific gene groups using precomputed gut metagenomes yields novel biological hypotheses. PLoS One *12*, e0176154.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.

Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A., and Koonin, E.V. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol *3*, 38–46.

CONCLUSION

DNA is the blueprint of life. Through transcription into RNA, the information stored in the genome guides a single cell growing into an organism with a diverse array of different cells. The goal of my study is to uncover how DNA sequence differences influence the various steps in the transcription cycle. In order to accomplish this, I used the existing genetic variation between the two copies of the genome inside diploid organism. Since the two copies of the genome reside in the same environment, this eliminates external confounding factors, such as a difference in transcription factors in different cells, and allows me to draw conclusions about the direct role of DNA sequence differences in the transcription cycle.

These genetic variations, known as single nucleotide polymorphisms (SNPs), can serve as markers to identify allelic differences in transcription, and in some cases be the cause of the observed differences. Since genetic variation is naturally occurring, the majority of DNA sequence differences likely have no major influence on downstream biological functions due to purifying selection. This makes them a great tool to observe the naturally occurring variation in the steps of transcription cycles and serve as a source of information about how DNA sequence influences transcription.

Despite the general utility of allele specific analysis, few computational methods have been developed to identify allele specific expression. Existing methods either test the amount of reads mapped to each SNP independently or combine allelic reads within each gene annotation. Treating each SNPs independently has limited statistical power and is likely to result in false negatives or false positives depending on the read depth. Combining allelic reads within gene annotation can increase the statistical power, but it

requires a well-annotated reference genome and prevents studying unannotated or non-coding regions.

To address these limitations, I developed AlleleHMM to identify genomic regions showing the same allele bias using a hidden Markov model (HMM). AlleleHMM uses the spatial correlation between nearby SNPs to increase the sensitivity and specificity of detecting allelic bias blocks when compared to tests that treat each SNP independently. Since AlleleHMM uses the allelic read counts at each SNP as input to infer the allelic bias blocks, it requires no annotation and can identify allelic bias blocks in any region in the genome with SNPs and transcription activity. AlleleHMM allows the study of non-coding regions, where the majority of information about transcriptional regulation is stored.

Transcription by RNA polymerase II (Pol II) is a process that is highly regulated by many macromolecules. While there are many studies to elucidate how the macromolecules play a role in different stages of transcription, there is little understanding of what the role of DNA sequence plays in the different stages of transcription. Genome-wide association studies (GWAS) of autoimmune disease show that the majority of SNPs associated with disease fall in noncoding regions (Farh et al., 2015). Only 10-20% of those in noncoding regions alter known transcription binding motifs and only 12% score as expression quantitative trait loci (eQTLs). This suggests that other aspects of transcriptional regulation, such as the location of transcription and the act of transcription itself, can contribute to disease pathogenesis.

We devised a dataset to maximize allelic differences using genetically distant mouse strains in order to study the effect of DNA sequence changes on transcription. In

collaboration with a graduate student from Cohen and Danko lab, we generated reciprocal F1 hybrids from two mouse strains with distinct genetic backgrounds, C57BL/6 (B6) and Castaneus (CAST). We harvested eight organs and used ChRO-seq to measure the position and orientation of RNA polymerase II genome-wide. Using this rich dataset, I identified genetic factors that implicate in three steps of transcription: initiation, pause, and termination.

In my study of transcription initiation, I found transcription start sites (TSSs) with different usage of transcription start nucleotides (TSNs), resulting in distinct shapes of TSSs. The TSSs with allelic differences can be divided into two classes: cases that were driven predominantly by large changes in the abundance of Pol II at a single TSN position and cases in which multiple TSNs across the TSS contributed to changes in shape. Comparing the usage of maxTSNs between two alleles, I found that the nucleotide A in the initiation site may be the most important genetic determinant of transcription initiation. I also examined how the other TSNs were affected when the maxTSN contained a SNP that replaces the strong Initiator element (Inr) CA. We found that the potential TSNs with a Inr motif, both upstream and downstream of the maxTSN, have increased usage when the maxTSN contain a SNP. Based on this finding, we propose that RNA Pol II are more likely to scan for an energetically favorable Initiator elements in both directions by a stochastic process resembling Brownian motion.

In my study of transcription pausing, I recovered the nucleotide C at the paused Pol II active site (Gressel et al., 2017; Tome et al., 2018) by generating sequence logos of max pause positions in three murine organs. When there is a SNP at the max pause site that replaces the C in one of the alleles, I found that the alleles with SNPs paused

within 10bp downstream the original max pause sites in all cases, with 80% paused within 5bp downstream. I also found that changes in pause position were correlated with changes in the size of the insertion or deletion, such that the difference in distance between the max TSN and the pause site in the native genome coordinates was typically less than 5 bp. These findings support a model in which paused Pol II is placed in part through physical constraints with the pre-initiation complex (Fant et al., 2020; Kwak et al., 2013).

In my study of transcription termination, I identified allelic difference in termination that result in a difference in the length of the primary transcription unit. Despite the allelic difference in transcript length, the sequence composition and the length remain constant in the majority (60-80%) of the mature mRNA. Allelic differences in termination were mostly consistent between different organs indicating that DNA sequence differences are the major determinants of transcription termination.

For allelic difference in initiation and pause, there is little or no change in the transcription level of the gene bodies or mature RNA. Since the genetic variation is naturally occurring, we suspect those variations with stronger effects are likely to be eliminated due to purifying selection.

F1 hybrids with high levels of heterozygosity and allelic specific analysis are a great tool to identify naturally occurring allelic differences. These findings suggest that DNA sequence difference is an important determinant which influences steps in the transcription cycle, especially for those allelic differences that are similar across tissues. Allelic specific analysis is restricted to genomic regions containing markers. To further understand the connection between DNA sequence and transcription, machine learning

models such as convolutional neural networks can be trained to learn the latent representation from DNA sequence that impact transcription.

REFERENCES

Fant, C.B., Levandowski, C.B., Gupta, K., Maas, Z.L., Moir, J., Rubin, J.D., Sawyer, A., Esbin, M.N., Rimel, J.K., Luyties, O., et al. (2020). TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. Molecular Cell *78*, 785-793.e8.

Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

Gressel, S., Schwalb, B., Decker, T.M., Qin, W., Leonhardt, H., Eick, D., and Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. ELife *6*, e29736.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. Science *339*, 950–953.

Tome, J.M., Tippens, N.D., and Lis, J.T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. Nat Genet *50*, 1533–1541.