UNIVERSITY OF CALIFORNIA, SAN DIEGO

Genetic regulation of RNA splicing and expression in cancer and stem cells

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Christopher Mark DeBoever

Committee in charge:

Professor Kelly Frazer, Chair
Professor Trey Ideker, Co-Chair
Professor Rafael Bejar
Professor Terry Gaasterland
Professor Catriona Jamieson

2016

The Dissertation of Christopher Mark DeBoever is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2016

DEDICATION

This thesis is dedicated to my family and friends.

EPIGRAPH

Facts are meaningless.
You could use facts to prove anything that's even remotely true!

Homer Simpson

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTARY FILES

Supplementary File 1.1: Metadata for samples used in this study. *SF3B1* mutated samples have columns for frequency of *SF3B1* mutation in RNA-seq data, mutation type, codon change and whether the mutation is in the HEAT 5-9 repeats. These columns are empty for *SF3B1* wild-type tumor samples.

Supplementary File 1.2: Summary of differential junction usage results from DEXSeq. DEXSeq was used to test for differential splice junction usage in a joint analysis of the CLL, BRCA, and UM samples as well as individually for each cancer type. “Novel” indicates that the junction is not annotated in Gencode. Proximal indicates that a novel 3’SS is 10-30 bp upstream of a canonical Gencode 3’SS.

Supplementary File 1.3: 619 cryptic 3’SSs located 10-30 bp upstream of canonical 3’SSs from joint BRCA, CLL, and UM analysis. Location of 5’ splice sites and 3’SSs are one-based coordinates that denote the start and end of the intron. The columns COSMIC, TSgene, and ncg denote whether the gene is present in COSMIC, TSGene, or the Network of Cancer Genes respectively.

Supplementary File 1.4: 417 distal cryptic 3’SSs used more often in *SF3B1* mutants from joint BRCA, CLL, and UM analysis. Location of 5’ splice sites and 3’SSs are one-based coordinates that denote the start and end of the intron. The columns COSMIC, TSgene, and ncg denote whether the gene is present in COSMIC, TSGene, or the Network of Cancer Genes respectively.

Supplementary File 1.5: GSEA results for 912 genes containing 619 proximal and 417 distal cryptic 3’ splice sites used more often in *SF3B1* mutants.

Supplementary File 1.6: 325 significant cryptic 3’SSs located 10-30 bp upstream of canonical 3’SSs and used more often in *SF3B1* mutants from CLL-only DEXSeq analysis. Location of 5’ splice sites and 3’SSs are one-based coordinates that denote the start and end of the intron. The columns COSMIC, TSgene, and ncg denote whether the gene is present in COSMIC, TSGene, or the Network of Cancer Genes respectively.

Supplementary File 1.7: Percent spliced in for 325 cryptic 3’SSs located 10-30 bp upstream of canonical 3’SSs from CLL-only DEXSeq analysis. Note that there

are only 324 values because one canonical 3'SS was filtered due to low coverage so a PSI value could not be calculated.

Supplementary File 1.8: 272 genes that are differentially expressed between *SF3B1* mutant and wild-type samples from joint analysis of CLL, BRCA, and UM using DESeq2.
Supplementary File 1.9: GSEA results for 272 genes differentially expressed genes from joint CLL, BRCA, and UM DESeq2 analysis.

Supplementary File 1.10: 192 significant cryptic 3'SSs located 10-30 bp upstream of canonical 3'SSs and used more often in *SF3B1* mutants from BRCA-only DEXSeq analysis. Location of 5' splice sites and 3'SSs are one-based coordinates that denote the start and end of the intron. The columns COSMIC, TSgene, and ncg denote whether the gene is present in COSMIC, TSGene, or the Network of Cancer Genes respectively.

Supplementary File 1.11: Percent spliced in for 192 cryptic 3'SSs located 10-30 bp upstream of canonical 3'SSs from BRCA-only DEXSeq analysis. Note that there are only 191 values because one canonical 3'SS was filtered due to low coverage so a PSI value could not be calculated.

Supplementary File 1.12: 33 genes that are differentially expressed between *SF3B1* mutant and wild-type CLL samples using DESeq2.

Supplementary File 2.1: Classification of 5,619 eGenes. Number of genes tested and significant in primary eQTL analysis using 215 subjects stratified by gene type according to Gencode v19. Percent eGenes is the percentage of the eGenes that the indicated gene type comprise.

Supplementary File 2.2: eQTL results. Lead variants and all significant variants for 5,619 eGenes as well as 668 eGenes with second eQTLs and 201 eGenes with third eQTLs. "leads01" contains the lead variants for the primary eQTL analysis and "all01" contains all significant associations for the primary eQTL analysis. "leads02" and "all02" contain the lead variants and all significant associations for the second eQTL analysis conditioning on the lead variant from the first analysis. "leads03" and "all03" contain the lead variants and all significant associations for the third eQTL analysis conditioning on the lead variant from the first and second analyses.

Supplementary File 2.3: Noncoding lead variant enrichment results. Results for enrichment of 4,491 noncoding SNVs and indels (Fisher exact test) in Roadmap and ENCODE DNase hypersensitivity sites and ENCODE transcription factor ChIP-seq peaks.

Supplementary File 2.4: Putative expression quantitative trait nucleotides. Putative expression quantitative trait nucleotides (peQTNs) that overlap transcription factor ChIP-seq peaks and disrupted an associated motif. There are 3,058 distinct peQTNs although some are associated with the expression of more than one gene so there are more than 3,058 rows in this table.

Supplementary File 2.5: GWAS enrichments. Enrichment results (Fisher exact test) for lead variants, peQTNs, and peQTNs without HLA gene associations in GWAS hits for 33 phenotypes from the GRASP database. “number_markers” is the number of GWAS hits with $p < 10^{-5}$ and “number_independent” is the number of these hits with LD < 0.8 (1000 Genomes phase 3 EUR).

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Kelly Frazer for mentoring me during my time in her laboratory. I would like to thank all of the Frazer laboratory members and my collaborators for their help and support.

Chapter 1, in full, is a reprint of material as it appears in *PLoS Computational Biology* 2015, Christopher DeBoever, Emanuela M. Ghia, Peter J. Shepard, Laura Rassenti, Christian L. Barrett, Kristen Jepsen, Catriona H. M. Jamieson, Dennis Carson, Thomas J. Kipps, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, has been submitted for publication of the material as it may appear in *Cell Stem Cell*, 2016. Christopher DeBoever, He Li, David Jakubosky, Angelo Arias, Joaquin Reya, William Biggs, Efren Sandoval, Hiroko Matsui, Paola Benaglio, Agnieszka D'Antonio-Chronowska, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

VITA

2010 Bachelor of Science, Harvey Mudd College

2016 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

DeBoever C, Arias A, Frazer KA. Genetic regulation of gene expression in a collection of 215 human induced pluripotent stem cells. Submitted.

Cui B, Ghia EM, Chen L, Rassenti L, DeBoever C, Yu J, Zhang L, Neuberg DS, Wierda WG, Rai KR, Kay NE, Brown JR, Byrd JC, Gribben JG, Greaves AW, Frazer KA, Kipps TJ. High-level ROR1 associates with accelerated disease-progression in chronic lymphocytic leukemia. Submitted.

Barrett CL, DeBoever C, Jepsen K, Saenz CC, Carson DA, Frazer KA. Systematic transcriptome analysis revals tumor-specific isoforms for ovarian cancer diagnosis and therapy. PNAS 2015; 112(23): E0350-E3057; doi: 10.1073/pnas.1508057112.

Smith EN, Ghia EM, DeBoever CM, Rassenti L, Jepsen K, Yoon KA, Matsui H, Rozenzhak S, Alakus H, Shepard P, Dai Y, Khosroheidari M, Bina M, Gunderson K, Messer K, Muthuswamy L, Hudson T, Harismendy O, Barrett C, Jamieson CHM, Carson D, Kipps TJ, Frazer KA. Genetic and epigenetic profiling of CLL disease progression reveals limited somatic evolution and suggests a relationship to memory-cell development. Blood Cancer Journal 2015; 5, e303; doi:10.1038/bcj.2015.14.

DeBoever C, Ghia EM, Shepard PJ, Rassenti L, Jepsen K, Jamieson CHM, Carson D, Kipps TJ, Frazer KA. Transcriptome sequencing reveals potential mechanism of 3' splice site selection in *SF3B1*-mutated cancers. PLoS Computational Biology 2015; 11(3): e1004105. doi:10.1371/journal.pcbi.1004105.

Cheng CP, DeBoever C, Frazer KA, Liu YC, Tseng VS. MiningABs: mining associated biomarkers across multi-connected gene expression datasets. BMC Bioinformatics 2014; 15:173. Doi:10.1186/1471-2105-15-173

DeBoever C, Reid EG, Smith EN, Wang X, Dumaop W, Harismendy O, Carson D, Richman D, Masliah E, Frazer KA. Whole transcriptome sequencing enables discovery and analysis of viruses in archived primary central nervous system lymphomas. PLOS One 2013; 8(9): e73956. Doi: 10.1371/journal.pone.0073956

Bush EC, Clark AE, DeBoever CM, Haynes LE, Hussain S, Ma S, McDermott MM, Novak AM, Wentworth JS. Modeling the Role of Negative Cooperativity in Metabolic Regulation and Homeostasis. PLOS One 2012; 7(11): e48920. Doi:10.1371/journal.pone.0048920

Nevarez PA, DeBoever CM, Freeland BJ, Quitt MA, Bush EC. Context dependent substitution biases vary within the human genome. BMC Bioinformatics 2010; 11(1): 462. Doi: 10.1186/1471-2105-11-462.

FIELDS OF STUDY

Major Field: Bioinformatics

Studies in Quantitative Genetics
Professor Kelly Frazer

Studies in Cancer Genetics
Professor Kelly Frazer

ABSTRACT OF THE DISSERTATION

Genetic regulation of RNA splicing and expression in cancer and stem cells

by

Christopher Mark DeBoever

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2016

Professor Kelly Frazer, Chair

Professor Trey Ideker, Co-Chair

A central question in genetics is how different classes of DNA variants affect RNA splicing and expression. While there has been substantial progress in

associating single nucleotide polymorphisms and small indels with these phenotypes, only recently has affordable high throughput sequencing provided the opportunity to assess the impact of somatic, rare, and copy number variants (CNVs) on RNA splicing and expression. In this thesis, I use high throughput sequencing to investigate the effect of somatic variants in *SF3B1* on RNA splicing and characterize the genetic regulation of gene expression in induced pluripotent stem cells (iPSCs). In the first part, I examine the effect of recurrent somatic mutations in the splicing factor *SF3B1* on RNA splicing in three different cancer types and find that *SF3B1* mutants use hundreds of cryptic 3' splice sites that are rarely used in samples without *SF3B1* mutations. Sequence properties of these cryptic 3' splice sites suggest altered sterics may allow usage of cryptic 3' splice sites in *SF3B1* mutants. I also identify several candidate genes with out-of-frame cryptic splice sites that are used in a majority of transcripts in the mutants and may contribute to oncogenesis. In the second part, I examine the genetic regulation of gene expression in a collection of 215 human iPSCs using transcriptome and whole genome sequencing. I identify expression quantitative trait loci (eQTLs) for nearly six thousand genes including markers of pluripotency such as *POU5F1*, *LCK*, *IDO1*, and *CXCL5*. A comparison to GTEx eQTLs reveals that iPSCs are well powered statistically for finding eQTLs and have a unique regulatory landscape. I identify biallelic and multiallelic CNVs eQTLs and find that a substantial proportion of CNV eQTLs appear to affect intergenic regulatory regions. I also find that rare promoter variants weakly disrupt gene

expression while rare CNVs that overlap genes tend to disrupt gene expression with relatively high effect sizes. Overall, this thesis helps define the roles of somatic, rare, and copy number variants in the regulation of gene expression and splicing and provide key insights into *SF3B1*-mutated cancers and iPSCs as a model system for molecular association analyses.

Chapter 1: Transcriptome Sequencing Reveals Potential Mechanism of Cryptic 3' Splice Site Selection in *SF3B1*-mutated Cancers

Chapter 1.1: Abstract

Mutations in the splicing factor *SF3B1* are found in several cancer types and have been associated with various splicing defects. Using transcriptome sequencing data from chronic lymphocytic leukemia, breast cancer and uveal melanoma tumor samples, we show that hundreds of cryptic 3' splice sites (3'SSs) are used in cancers with *SF3B1* mutations. We define the necessary sequence context for the observed cryptic 3' SSs and propose that cryptic 3'SS selection is a result of *SF3B1* mutations causing a shift in the sterically protected region downstream of the branch point. While most cryptic 3'SSs are present at low frequency (<10%) relative to nearby canonical 3'SSs, we identified ten genes that preferred out-of-frame cryptic 3'SSs. We show that cancers with mutations in the *SF3B1* HEAT 5-9 repeats use cryptic 3'SSs downstream of the branch point and provide both a mechanistic model consistent with published experimental data and affected targets that will guide further research into the oncogenic effects of *SF3B1* mutation.

Chapter 1.2: Introduction

One of the biggest surprises to emerge from the growing catalog of somatic mutations in various cancer types is the recurrent mutation of gene encoding the RNA spliceosome (Watson, Takahashi, Futreal, & Chin, 2013). Recurrent mutations in the highly conserved HEAT 5-9 repeats of splicing factor 3B subunit 1 (*SF3B1*) have been reported in myelodysplastic syndrome, chronic lymphocytic leukemia (CLL), breast cancer (BRCA), uveal melanoma (UM), and pancreatic cancer (Biankin et al., 2012; Harbour et al., 2013; M. Martin et al., 2013; Papaemmanuil et al., 2011; Wan & Wu, 2013; Yoshida et al., 2011). *SF3B1* mutation is associated with poor prognosis in CLL but improved prognosis in myelodysplasia and UM (Harbour et al., 2013; Quesada et al., 2012; Schwaederle et al., 2013; Wan & Wu, 2013). Prior studies have shown that mutated *SF3B1* CLL samples have differential exon inclusion and use some cryptic 3' splice sites (3'SSs) relative to wild-type *SF3B1* CLL samples (Ferreira et al., 2013; Papaemmanuil et al., 2011; Quesada et al., 2012; L. Wang et al., 2011; Yoshida et al., 2011). However, it is unknown whether *SF3B1* mutation is associated with the same 3'SS selection defects in different cancers. The mechanism underlying the cryptic 3'SS selection and the functional consequences thereof remain unresolved as well.

SF3B1 is a core part of the U2-small nuclear ribonucleoprotein (U2-snRNP) complex and stabilizes the binding of the U2-snRNP to the branch point (BP), a degenerate sequence motif usually located 21-34 bp upstream of the

3'SS (Gao, Masuda, Matsuura, & Ohno, 2008; Padgett, 2012). SF3B1 also interacts with other spliceosomal proteins such as U2AF2, which binds the polypyrimidine tract (PPT) downstream of the BP (Gozani, Potashkin, & Reed, 1998; Wan & Wu, 2013; C. Wang et al., 1998). The binding of the U2-snRNP and other spliceosome proteins around the BP prevents 3'SS selection in a ~12-18 bp region directly downstream of the BP due to steric hindrance (Chua & Reed, 2001; Smith, Chu, & Nadalginard, 1993). Inherited *cis*-acting splicing mutations beyond this ~12-18 bp region downstream of the BP that result in the use of cryptic 3'SSs have been shown to occur in Mendelian disease genes (Kralovicova, Christensen, & Vorechovsky, 2005). Additionally, a competitive region exists ~12 bp downstream from the first 3'SS after the protected region where AG dinucleotides can compete to be used as 3'SSs based on sequence characteristics such as the PPT length, distance from the BP, nucleotide preceding the AG dinucleotide, and other features (Chua & Reed, 2001).

The role of SF3B1 and the U2-snRNP in recognizing and binding the BP and the localization of mutations to HEAT 5-9 repeats suggest that *SF3B1* mutations are dominant drivers that may alter 3'SS selection (Papaemmanuil et al., 2011). To test this, we examined splice site usage in transcriptome data from *SF3B1* mutant and *SF3B1* wild-type CLL, UM and BRCA cases. We identified 619 cryptic 3'SSs used more frequently in *SF3B1* mutants and clustered 10-30 bp upstream of canonical 3'SSs. The majority of these cryptic 3'SSs were observed in all three tumor types despite the divergent clinical implications of

SF3B1 mutation. Our analysis of tumors with *SF3B1* mutations shows that cryptic 3'SS selection occurs only in samples with missense mutations at ~10 amino acid hotspots in the fifth to ninth HEAT repeats. We analyzed the organization of splicing motifs around the cryptic 3'SSs and found that only introns with an AG dinucleotide at the boundary of the sterically protected region downstream of the BP but >10 bp upstream of the canonical 3'SS are susceptible to cryptic 3'SS selection in *SF3B1* mutants. We assessed the functional impact of *SF3B1* mutation and found that the cryptic 3'SSs are typically used at low frequency in the *SF3B1* mutants (<10% relative to the canonical splice site) and are sometimes present in the *SF3B1* wild-types but at an even lower frequency (<0.5% relative to the canonical splice site). However, we identified 10 candidate genes, some previously implicated in tumorigenesis, for which there is a high amount of out-of-frame cryptic splice site usage that may affect the function of these genes.

Chapter 1.3: Results

Chapter 1.3.1: Cryptic 3' splice sites 10-30 bp upstream of canonical 3' splice sites are used in *SF3B1* mutants

We used RNA-sequencing data from *SF3B1* mutated and *SF3B1* wild-type chronic lymphocytic leukemia (CLL; seven mutant, nine wild-type), breast cancer (BRCA; 14 mutant, 18 wild-type), and uveal melanoma (UM; four mutant, four wild-type) samples (Supplementary Figure 1.1, Supplementary File 1.1) to test

219,476 splice junctions present in the Gencode v14 gene annotation (Harrow et al., 2012) along with 87,941 novel splice junctions (not annotated in Gencode) for differential usage by comparing junction-spanning reads using a generalized linear model as implemented in DEXSeq (Anders, Reyes, & Huber, 2012). A splice junction is considered differentially used between mutant and wild-type samples if the expression level of that junction differs significantly after accounting for overall expression differences of the corresponding gene locus. All tested junctions were covered by at least 20 reads summed over all cancer samples in a given analysis, shared a 5' splice site and/or 3'SS with a Gencode splice junction, and had a known splice site motif. We identified 1,749 junctions that were significantly differentially used between the *SF3B1* mutant and *SF3B1* wild-type samples across the three tumor types including 1,330 novel junctions, of which 1,117 are novel 3'SSs (BH-adjusted $p < 0.1$, Supplementary File 1.2). These 1,749 significant junctions were highly enriched for novel splice junctions compared to annotated junctions (Fisher exact, $p < 10^{-200}$) and the novel junctions were enriched for novel 3'SSs (Fisher exact, $p < 10^{-200}$) showing that *SF3B1* mutations result in the usage of a large number of novel 3'SSs. These 1,749 significant junctions include 61 of 79 splice sites recently reported as specific to CLL cases with *SF3B1* mutations (Ferreira et al., 2013) supporting the specificity of our approach while demonstrating an increased sensitivity that has allowed us to identify many more cryptic 3'SSs than previously reported. We plotted the distance between each significant novel 3'SS and its associated

canonical 3'SS (defined as the nearest Gencode 3'SS that shared the same 5' splice site - see Methods). Of the 1,117 significant novel 3'SSs, 619 were proximal cryptic 3'SSs clustered 10-30 bp upstream of their associated canonical 3'SSs while the remaining 498 cryptic 3'SSs were widely distributed (herein referred to as distal cryptic 3'SSs) (Figure 1.1A, Supplementary File 1.3). All of the 619 proximal cryptic 3'SSs were used more often in the *SF3B1* mutant samples compared to the wild-type samples and 58% were out-of-frame relative to the nearby canonical 3'SSs, suggesting that these are not canonical 3'SSs missing from Gencode. 417 of the 498 distal cryptic 3'SSs were also used more highly in the *SF3B1* mutants (Supplementary File 1.4). The distribution of the 1,117 significant novel 3'SSs is different from that of novel 3'SSs whose usage did not differ significantly between the *SF3B1* mutants and wild-types (Figure 1.1B,C), further demonstrating that the usage of proximal cryptic 3'SSs is a property of *SF3B1* mutants. Examining each tumor type individually, we observed the same enrichment of cryptic 3'SSs 10-30 bp upstream of canonical splice sites (Supplementary Figure 1.2). Given these observations, SF3B1's role in binding the BP, and the organization of the BP and splicing motifs in the last 30 bp of the intron (Padgett, 2012), we focused our initial analyses on the 619 proximal cryptic 3'SS.

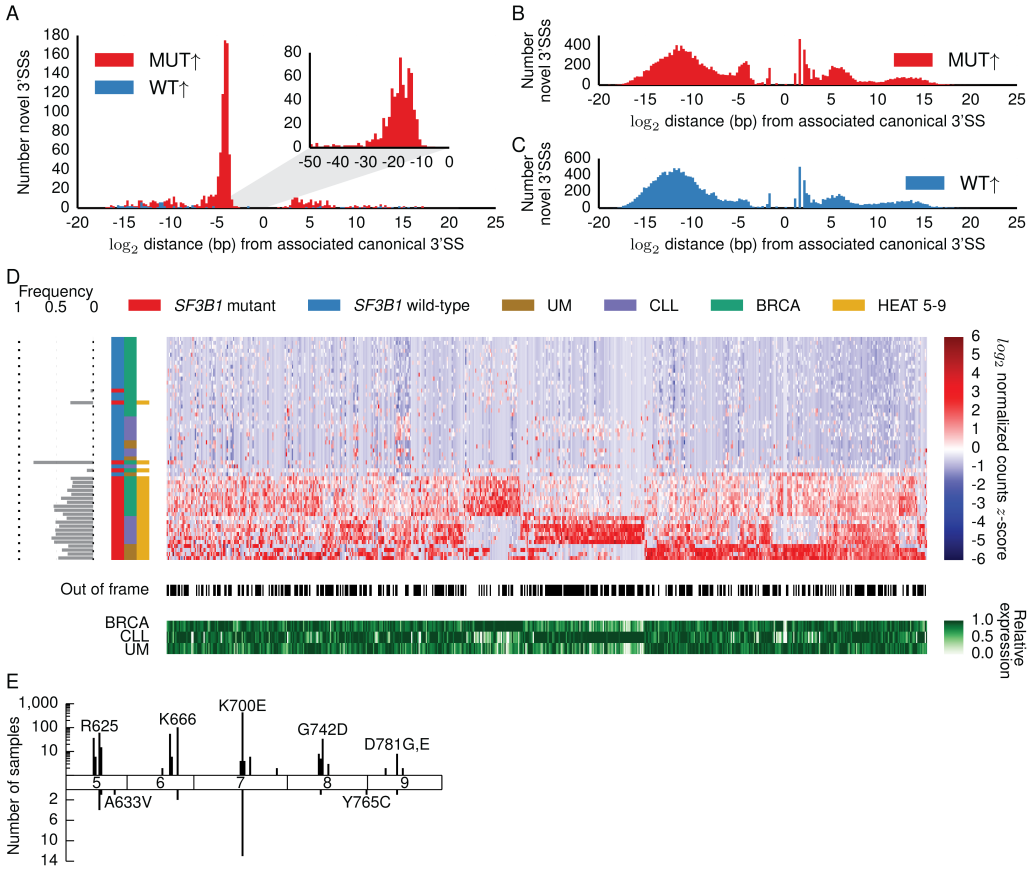


Figure 1.1: Proximal cryptic 3'SSs used significantly more often in cancers with SF3B1 hotspot mutations. $\log_2$ distance in base pairs from associated canonical 3'SSs to (A) 1,117 significantly differentially used novel 3'SSs, (B) 16,673 novel 3'SSs with canonical intron motifs (GT/AG) used more highly in the mutants but not significant, and (C) 18,660 novel 3'SSs with canonical intron motifs (GT/AG) used more highly in the wild-types but not significant. Zero represents the position of the canonical 3'SS. Negative and positive distances indicate that the cryptic 3'SS is respectively upstream or downstream from the canonical 3'SS. Inset in (A) shows base-by-base binning from zero to 50 base pairs upstream of canonical 3'SS. Red and blue histograms represent junctions with significantly higher usage in SF3B1 mutants or SF3B1 wild-type samples, respectively. (D) Upper red and blue heatmap shows for each sample the $\log_2$ library-normalized count z-score for 619 cryptic 3'SSs used significantly more often in the SF3B1 mutants and located 10-30 bp upstream of canonical 3'SSs (DEXSeq, BH-adjusted $p < 0.1$). Grey bars at left indicate frequency of SF3B1 mutant allele in RNA-seq data. Colorbars indicate SF3B1 mutation status, cancer type, and whether the SF3B1 mutation is located in the HEAT 5-9 repeats. Black and white colorbar indicates whether novel 3'SSs are out-of-frame (black) relative to canonical 3'SSs. Bottom green heatmap shows relative expression levels for the genes containing each cryptic 3'SS. We calculated the average expression of each gene in each cancer type and normalized by the maximum expression for each gene so that the maximum value in each column is one (see Methods). Cryptic 3'SSs not observed in all cancer types tend to have differing gene expression levels between cancers. (E) Locations and frequency of SF3B1 mutations in HEAT repeats 5-9. Mutations observed more than once in COSMIC (upper axis) cluster in ~10 amino acid hotspots in each HEAT repeat; most frequent mutation in each hotspot is labeled. Bottom axis shows locations and frequency of mutations in our study. BRCA samples with A663V and Y765C mutations do not show evidence for cryptic 3'SS selection.

Chapter 1.3.2: Cryptic 3'SS selection is limited to tumors with mutations in HEAT repeat hotspots

We clustered all samples based on the read coverage of the 619 proximal cryptic 3'SSs and found that four *SF3B1*-mutated BRCA samples did not cluster with the other mutants (Fig 1D). The *SF3B1* mutation for one of these BRCA samples was a nonsense mutation not located in the HEAT 5-9 repeats while another sample had a subclonal (8.4%) HEAT 5-9 mutation with attenuated cryptic 3'SS selection (Supplementary Figure 1.3). The other two samples had mutations in the HEAT 5-9 repeats but outside of the apparent ~10 amino acid mutational hotspots (Fig 1E). We observed cryptic 3'SS selection in a TCGA lung adenocarcinoma sample with a hotspot mutation but not in lung cancer samples with *SF3B1* mutations outside of the five hotspots (Supplementary Figure 1.4). These results show that cryptic 3'SS selection only occurs in tumors carrying mutations in one of the five ~10 amino acid hotspots in the HEAT 5-9 repeats and is not limited to cancers in which *SF3B1* is recurrently mutated.

Chapter 1.3.3: Cryptic 3'SSs are shared across different cancer types

The majority of the 619 proximal cryptic 3'SSs were used in *SF3B1*-mutated samples in all three cancer types suggesting that the mechanism of cryptic 3'SS selection in *SF3B1*-mutated tumors is the same between different cancers (Fig 1D). Some cryptic 3'SSs were not used in one or two of the cancer types due to lower expression of the corresponding genes in those cancers.

Differences in cryptic 3'SS usage due to varying gene expression may contribute to the divergent prognostic implications of *SF3B1* mutation in various cancers (Harbour et al., 2013; Wan & Wu, 2013).

To characterize the roles of the genes affected by cryptic 3'SS usage, we performed a gene set enrichment analysis for the 912 genes that contained the 619 proximal and 417 distal cryptic 3'SSs used significantly more often in the *SF3B1* mutant samples (Supplementary File 1.5). The gene set with the second smallest *p*-value consists of genes up-regulated in chronic myelogenous leukemia and the seventh gene set contains genes up-regulated in aggressive uveal melanoma samples (GSEA (Subramanian et al., 2005), $q < 10^{-35}$). These results may reflect the fact that we are more likely to identify cryptic 3'SSs in genes that are highly expressed which may bias such a gene set enrichment analysis. Nonetheless, several gene sets with potential importance for cancer development are enriched such as genes positively correlated with *BRCA1*, *ATM*, and *CHEK2* expression across normal tissues (GSEA, $q < 10^{-28}$).

Chapter 1.3.4: Cryptic 3'SSs are located ~13-17 bp downstream of the branch point

We characterized the sequence features of the 619 proximal cryptic 3'SSs and their associated canonical 3'SSs to gain further insights into the mechanism of cryptic 3'SS selection (Figure 1.2A). We chose 23,066 control 3'SSs (see Methods) and plotted the nucleotide frequency (Crooks, Hon, Chandonia, &

Brenner, 2004) for the last 50 bp of the introns for all control, associated canonical, and cryptic 3'SSs as well as the enrichment of adenines relative to the control introns. The control introns have a typical nucleotide composition with a 4-24 bp PPT preceding the 3'SS (Figure 1.2B) (Gao et al., 2008). The associated canonical 3'SS introns are enriched for adenines ~15-20 bp upstream of the 3'SS since the proximal cryptic 3'SSs are located in this region (Figure 1.2C). However, the introns for proximal (Figure 1.2D) and distal (Figure 1.2E) cryptic 3'SSs have a strong enrichment of adenines concentrated ~15 bp upstream of the splice sites. These results suggest that the increased usage of the 619 proximal and 417 distal cryptic 3'SSs in the *SF3B1* mutants may result from the same mechanism. The human BP motif is highly degenerate except for a largely invariant adenine (Gao et al., 2008) leading us to suspect that the adenine signal upstream of the cryptic 3'SSs is caused by the associated canonical 3'SSs' BP adenines. We used SVM_BP (Corvelo, Hallegger, Smith, & Eyras, 2010) to predict BPs for the associated canonical 3'SSs and calculated the distance from the highest scoring predicted BPs to the cryptic splice sites. We found that AG dinucleotides that serve as cryptic 3'SSs are enriched ~13-17 bp downstream from the predicted BP (Figure 1.3A) relative to random AG dinucleotides present in control 3'SS introns (Figure 1.3B, $p < 10^{-7}$, Mann Whitney U). For cryptic 3'SSs not located 13-17 bp downstream from the highest scoring BP in Figure 1.3A, we calculated the distance from the second highest scoring BP to the cryptic 3'SSs

and found that overall, the majority of the cryptic 3'SSs were located 13-17 bp from either the highest or second highest scoring BP (Figure 1.3C).

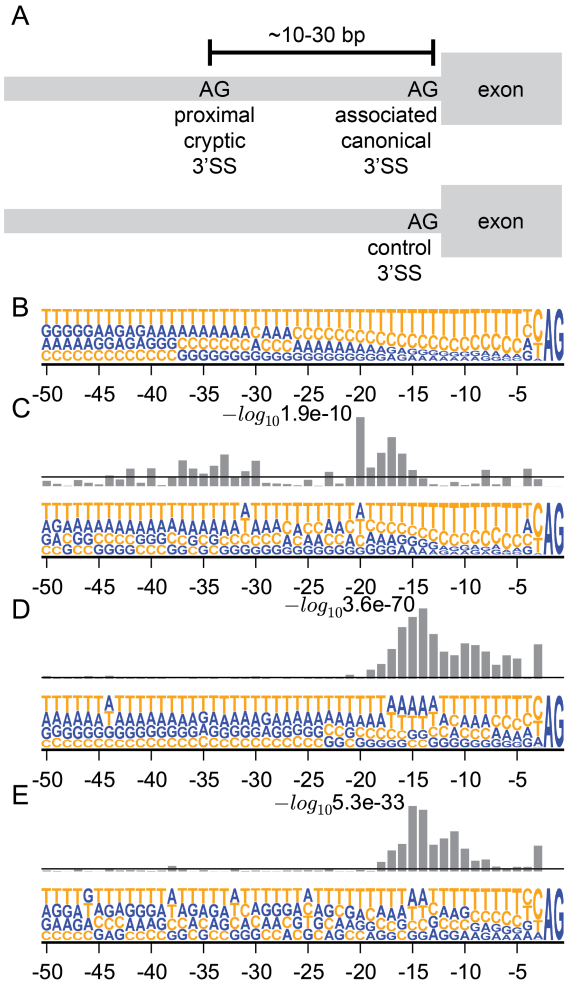


Figure 1.2*:* 3' intron nucleotide composition for control, associated canonical, and cryptic 3'SSs. (A) We identified 23,066 control 3'SSs whose junctions had a mean coverage greater than 100 reads over all CLL, BRCA, and UM samples to compare to the cryptic and associated canonical 3'SSs. Nucleotide frequency for the last 50 bp of the intron for (B) 23,066 control 3'SSs; (C) 613 associated canonical 3'SSs; (D) 619 proximal cryptic 3'SSs; and (E) 417 distal cryptic 3'SSs. Bar plots above each nucleotide composition plot are $-\log_{10}$ *p*-values from Fisher exact tests for enrichment of adenines at each position relative to control 3'SSs. Horizontal line marks significance level of *p*=0.05 ($-\log_{10} 0.05 \approx 1.3$). The *p*-value box plots have different scales in (C), (D), and (E); the smallest *p*-values for each panel are labeled.

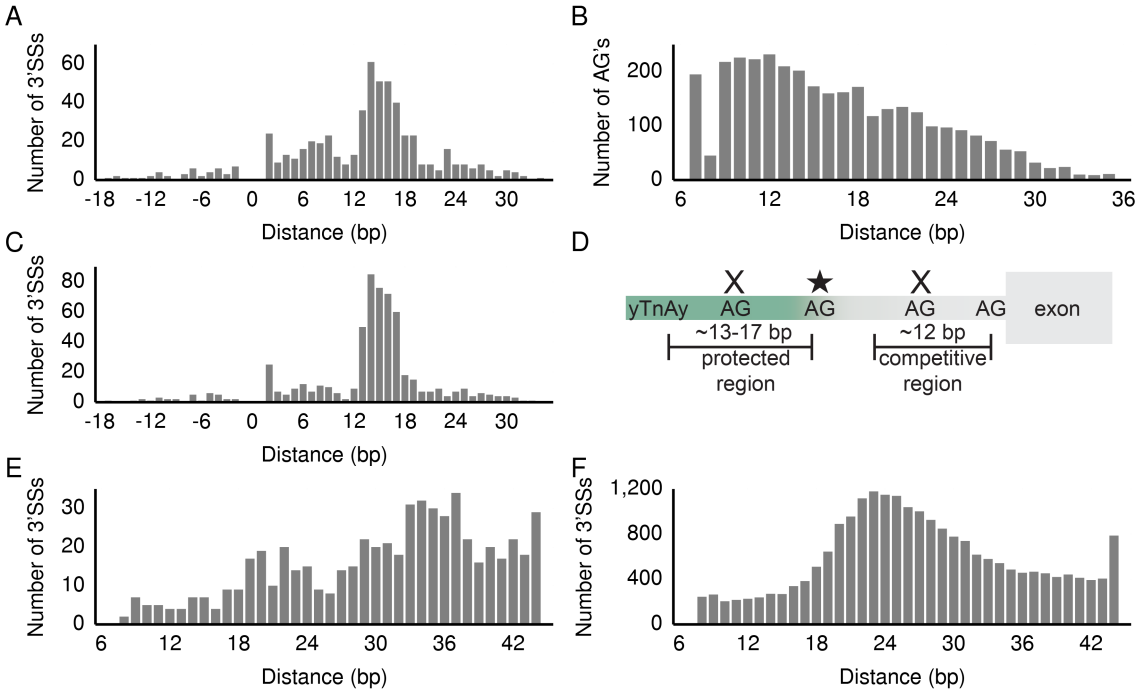


Figure 1.3: Location of predicted branch point relative to cryptic and canonical 3'SSs and model of cryptic 3'SS selection. (A) Distance from highest scoring BP predicted for associated canonical 3'SSs to the corresponding proximal cryptic 3'SSs. A negative distance indicates that the cryptic 3'SS is upstream of the BP predicted for the canonical 3'SS. The small spike at 2 bp indicates that in a few cases the adenine in the cryptic 3'SS is predicted to be the BP adenine for the canonical 3'SS. (B) Distance from highest scoring BP predicted for control 3'SSs to downstream intronic AG dinucleotides that are not annotated as 3'SSs. (C) Distance from either highest or second highest scoring BP predicted for canonical 3'SSs to their associated cryptic 3'SSs (see Methods). (D) Model for proximal cryptic 3'SS selection in SF3B1 mutants. yTnAy is the human BP motif. AG dinucleotides located at the edge of the sterically protected region can be used as 3'SSs in SF3B1 mutants (star). AG dinucleotides located in the protected or competitive regions (X's) are respectively sterically hindered from being selected as 3'SSs or out-competed by the canonical 3'SS. Distance from predicted BP to 3'SS for (E) associated canonical 3'SSs and (F) control 3'SSs (see Methods) is significantly different ($p < 10\text{-}23$, Mann Whitney U).

Chapter 1.3.5: Proposed mechanism of cryptic 3'SS selection

3'SSs are typically not located within ~12-18 bp downstream of the BP because the proteins bound to the BP sterically hinder AG dinucleotides in this region and prevent them from being used as 3'SSs (Smith et al., 1993). Our results suggest that AG dinucleotides serving as cryptic 3'SSs in *SF3B1* mutants are located at the end of this sterically protected region downstream of the BP (Figure 1.3D). Additionally, during the splicing reaction, the spliceosome

searches ~12 bp downstream from the first 3'SS after the BP for any other 3'SSs and chooses the strongest 3'SS based on sequence features (Smith et al., 1993). The lack of cryptic 3'SSs in the last 10 bp of the intron (Fig 1A) indicates that cryptic 3'SSs used in *SF3B1* mutants are located far enough upstream of the associated canonical 3'SSs to avoid competition for splicing. We observed that the distance between associated canonical 3'SSs and their predicted BPs is significantly greater than the distance between control 3'SSs and their BPs such that the cryptic 3'SSs at the edge of the protected region do not compete with the canonical 3'SS for splicing ($p < 10^{-23}$, Mann Whitney U, Figure 1.3E,F). We also predicted BP's for the 619 proximal and 417 distal cryptic 3'SSs (as opposed to above where we predicted BP's for the canonical 3'SSs associated with the 619 proximal 3'SSs) and found that the majority of these cryptic 3'SSs were 13-17 bp downstream of their predicted BP's (Supplementary Figure 1.5) providing further evidence that most cryptic 3'SSs (both proximal and distal) associated with *SF3B1* mutations are located at the edge of the sterically protected region.

Our results suggest that the mechanism of cryptic 3'SS selection in *SF3B1* mutants is not altered BP recognition because a more varied distribution of distances from the cryptic 3'SS to the canonical 3'SS BP would be expected if BP recognition was altered. Studying the role of cryptic 3'SS in inherited Mendelian disease genes, Královicová *et al.* 2005 used splicing reporters with cryptic 3'SSs located in the PPT and found that moving the cryptic 3'SS into the ~12-18 bp sterically protected region reduced or eliminated cryptic 3'SS selection. On the

other hand, moving an AG dinucleotide out of the sterically protected region allowed for its selection as a cryptic 3'SS (Kralovicova et al., 2005). These published experimental results and the rigid distance between the BP and the cryptic 3'SSs observed in our study are consistent with a model of altered 3'SS selection in *SF3B1* mutants due to a change in the size of the sterically hindered region downstream of the BP.

To test whether the sequences requirements defined here are sufficient for cryptic 3'SS usage, we identified 11,302 introns whose canonical 3'SSs passed our coverage cutoff of 20 reads summed over all samples and had potential cryptic 3'SSs (intronic AG dinucleotides that were 10-30 bp upstream of an annotated 3'SS and 13-17 bp downstream of the highest-scoring predicted BP). For 900 of these introns, the potential cryptic 3'SSs also passed the coverage cutoff, of which 310 were used significantly more often in the *SF3B1* mutants. This analysis demonstrates that not every potential cryptic 3'SS is differentially used in the mutants, so the sequence requirements described here appear to be necessary for cryptic 3'SS usage but not sufficient.

Chapter 1.3.6: Cryptic 3'SSs are used infrequently relative to canonical 3'SSs

Although the cryptic splice sites described here are used significantly more often in the *SF3B1* mutants, the biological effects are likely dependent on the proportion of transcripts that use the cryptic 3'SSs relative to the canonical 3'SSs. We therefore calculated the percent spliced in (PSI) for the proximal cryptic 3'SSs

relative to their associated canonical 3'SSs in the CLL samples since they have a higher sequencing depth than the other tumor samples (Supplementary Figure 1.1) that allows for more accurate quantification of splicing and because the distribution of well-characterized low- and high-risk CLL prognostic factors was similar between the *SF3B1* mutated and wild-type samples (Figure 1.4A). To calculate PSI for the 325 proximal cryptic 3'SSs used significantly more often in the *SF3B1* mutants from the CLL-only analysis (Supplementary File 1.6, Supplementary File 1.7), we divided the number of reads that span the cryptic 3'SS by the number of reads that span both the cryptic 3'SS and its associated canonical 3'SS. We observed that some cryptic 3'SSs are used exclusively in *SF3B1* mutants while others are also used in *SF3B1* wild-type samples but at a lower frequency relative to the mutants (Figure 1.4A). 67% of the cryptic 3'SSs were included in <10% of transcripts compared to their associated canonical 3'SS. These results suggest that the cryptic splice sites are either included rarely even in the *SF3B1* mutants or that transcripts with cryptic splice sites are subject to a higher rate of nonsense-mediated decay (NMD). To investigate the potential role of NMD, we identified differentially expressed genes between the *SF3B1* mutant and wild-type samples in a joint analysis of all three cancers and performed a gene set enrichment analysis. We found that genes in the "Reactome NMD enhanced by the exon junction complex" set were enriched (GSEA (Subramanian et al., 2005), $q < 10^{-28}$) among the 272 differentially expressed genes (DESeq2, BH-adjusted $p < 0.1$, Supplementary File 1.8,

Supplementary File 1.9) suggesting that NMD may be different between the *SF3B1* mutants and wild-types. 33 of the 582 genes that contained the 619 proximal cryptic 3'SSs were differentially expressed with the expression of 29/33 of these genes lower in the *SF3B1* mutants. Genes containing a proximal cryptic 3'SSs were more likely to be differentially expressed (Fisher exact, $p < 10^{-8}$) and more likely to have lower expression in *SF3B1* mutants (Fisher exact, $p = 0.0009$). These results suggest that cryptic 3'SS selection may affect gene expression for a subset of genes. However, the observation that in-frame cryptic 3'SSs likely not subject to NMD and out-of-frame cryptic 3'SSs potentially subject to NMD are included at similar rates relative to their associated canonical 3'SSs (Figure 1.4A) suggests that most genes' expression are not affected by cryptic 3'SS selection and most cryptic 3'SSs are observed at a low frequency because they are spliced in infrequently compared to their associated canonical 3'SSs.

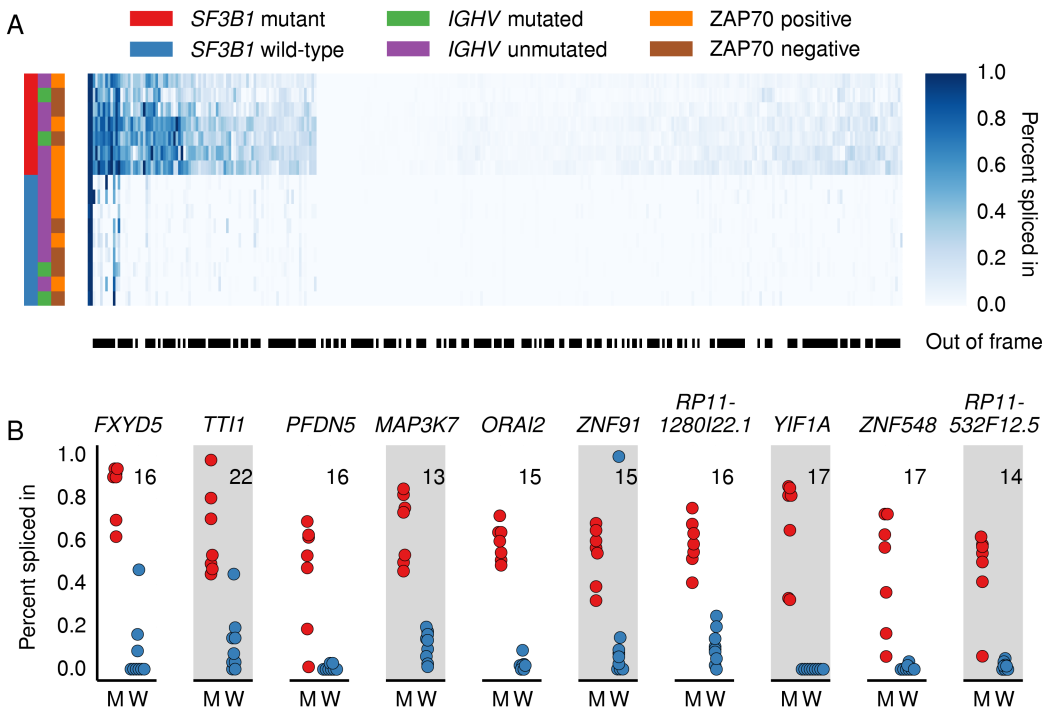


Figure 1.4: Percent spliced in for cryptic 3' splice sites in CLL analysis. (A) Heatmap shows the percent spliced in (PSI) values for cryptic 3'SS relative to the canonical 3'SS in CLL *SF3B1* mutated or wild-type samples for 325 proximal cryptic 3'SSs used significantly more often in the CLL mutants (DEXSeq, BH-adjusted $p < 0.1$). *SF3B1* mutation presence and the status of prognostic factors *IGHV* and ZAP70 are shown in left colorbars. Black and white colorbar indicates whether novel 3'SSs are out-of-frame (black) relative to canonical 3'SSs. In-frame and out-of-frame cryptic 3'SSs are used at similar rates relative to their associated canonical 3'SSs. (B) Beeswarm plots indicating the PSI values for the cryptic 3'SS relative to the associated canonical 3'SS in ten genes with high levels of cryptic 3'SS inclusion in CLL *SF3B1* mutants (M) compared to wild-type (W) samples. No reads were observed spanning the cryptic *YIF1A* junction in any wild-type CLL samples. The number in the upper corner of each plot is the distance in base pairs from the highest or second-highest scoring BP predicted for the associated canonical 3'SS to the cryptic 3'SS.

To identify cryptic 3'SSs with relatively high PSI values in the *SF3B1* mutant versus wild-type samples, we searched for cryptic 3'SSs that were 1) used more than 50% of the time in the CLL *SF3B1* mutants; 2) used less than 20% of the time in wild-type samples; and 3) had an average coverage of at least 30 junction-spanning reads in the mutant samples. Despite the generally low PSI values for the 325 cryptic 3'SSs from the CLL-only analysis, we identified four genes previously implicated in cancer (*TTI1* (Fernandez-Saiz et al., 2013; Hurov,

Cotta-Ramusino, & Elledge, 2010; Kaizuka et al., 2010), *MAP3K7* (Hofer-Warbinek et al., 2000; Kimura, Matsuo, Shibuya, Nakashima, & Taga, 2000; Yamaguchi et al., 1999), *FXYD5* (Nam, Hirohashi, & Wakefield, 2007), *PFDN5 (Fujioka et al., 2001)*) and six others (*YIF1A*, *ORAI2*, *ZNF91*, *ZNF548*, *RP11-1280I22.1*, *RP11-532F12.5*) with out-of-frame cryptic 3'SSs that were consistently preferred to the associated canonical 3'SS in the CLL *SF3B1* mutant samples (Figure 1.4B). Ferreira *et al.* identified the junctions in *ORAI2*, *ZNF91*, and *TTI1* in CLL SF3B1 mutants as well (Ferreira et al., 2013). Nine of the ten junctions were significant in our BRCA-only analysis and showed high differences in relative inclusion (Supplementary Figure 1.6, Supplementary File 1.10, Supplementary File 1.11). These genes are not differentially expressed between the CLL *SF3B1* mutant and wild-type samples (Supplementary File 1.12) but the frequent inclusion of out-of-frame cryptic 3'SSs may affect their biological function.

Chapter 1.4: Discussion

Here we have shown that a consequence of *SF3B1* mutations in different cancer types is genome-wide selection of hundreds of cryptic 3'SSs. We have shown the cryptic 3'SSs have specific sequence requirements; AG dinucleotides used as cryptic 3'SSs in *SF3B1* mutants are located at the end of the sterically protected region ~13-17 bp downstream of the BP but are >10 bp upstream of nearby canonical 3'SSs allowing them to avoid competition for splicing. These

sequence requirements limit the introns susceptible to cryptic 3'SS selection to those where the BP is located farther from the 3'SS than the typical ~24 bp. While these requirements appear necessary for cryptic 3'SS usage, they are not sufficient, as we did not detect cryptic 3'SS usage in all introns with AG dinucleotides that satisfy these requirements. Characteristics such as RNA conformation, RNA binding protein sites, BP prediction inaccuracies, cryptic or downstream canonical 3'SS strength, gene/transcript expression, sequencing depth, or other factors may also play a role in determining whether cryptic 3'SSs are used and detected by RNA sequencing.

Examining differential splice junction usage allowed us to identify many more cryptic 3'SSs than previous studies while still identifying 61 of 79 cryptic 3'SSs recently reported for CLL *SF3B1* mutants using a method based on relative inclusion (Ferreira et al., 2013; Papaemmanuil et al., 2011; Quesada et al., 2012; L. Wang et al., 2011; Yoshida et al., 2011). When examining the three cancer types in our study individually, the number of cryptic 3'SSs identified was highly dependent on the sequencing depth of the samples (Supplementary Figure 1.1 Supplementary Figure 1.2, Supplementary File 1.2). Additionally, examining cryptic 3'SS expressed higher in the *SF3B1* mutants but not significantly (Fig 1B) shows a modest enrichment of novel 3'SSs 10-30 bp upstream of canonical 3'SSs. These observations suggest that deeper sequencing will continue to reveal proximal cryptic 3'SSs in *SF3B1* mutants that are used very infrequently or are present in lowly expressed genes.

Selection of cryptic 3'SSs in the region downstream of the BP has been reported for some inherited diseases including those resulting from disrupted tumor suppressor genes such as *ATM*, *NF1*, and *TP53* (Kralovicova et al., 2005). Using a curated a list of aberrant splice sites associated with different diseases from the literature, Královicová *et al.* 2005 found that in cases where cryptic 3'SS selection was not caused by mutation of the 3'YAG consensus sequence, cryptic 3'SSs were often located ~19 bp upstream of associated canonical 3'SSs and ~11-15 bp downstream of the BP (Kralovicova et al., 2005). Most of the diseases considered in Královicová *et al.* 2005 are Mendelian diseases where a cryptic 3'SS disrupts or abolishes the function of a single disease gene. In these cases, a mutation in the PPT between the sterically protected and competitive regions has introduced a cryptic 3'SS (Figure 1.3D). For cancers with *SF3B1* mutations, we suspect that the size of the sterically protected region is slightly altered allowing for existing AG dinucleotides to be used as cryptic 3'SSs in hundreds of genes. It is also possible *SF3B1* mutations could cause destabilization of the U2 snRNP complex or alter interactions with U2AF2, affecting the ability to recognize the canonical 3'SS and leading to cryptic 3'SS selection. However, the rigid distance (~13-17 bp) from the predicted BPs to the cryptic 3'SSs for most of the cryptic 3'SSs is most consistent with a change in the size of the sterically protected region downstream of the branch point.

We found that cryptic 3'SS selection is limited to tumors with mutations in the five ~10 amino acid hotspots in the *SF3B1* HEAT 5-9 repeats and that these

mutations are associated with cryptic 3'SS selection across different cancer types and even in cancers in which *SF3B1* is not recurrently mutated. 58% of these cryptic 3'SSs are out-of-frame relative to nearby canonical 3'SSs, but the biological impact of these cryptic 3'SSs is likely a function of how frequently they are used relative to the nearby canonical 3'SSs. We found that while the cryptic 3'SSs are used more often in the *SF3B1* mutated samples compared to wild-type samples, they are used relatively infrequently (<10%) compared to nearby canonical 3'SSs. While the differentially expressed genes between the *SF3B1* mutated and wild-type samples are enriched for genes in the NMD pathway, even in-frame cryptic 3'SSs are used at a low frequency indicating that the associated canonical 3'SS is mostly preferred to the cryptic 3'SS even in *SF3B1* mutants. Nonetheless, we identified ten genes, including four with known roles in cancer, which had a high frequency of cryptic splice site usage relative to the nearby canonical splice site. Further studies are required to determine whether low-frequency cryptic 3'SS selection in hundreds of genes, high-frequency cryptic 3'SS selection in a small group of genes, and/or other splicing alterations drive the oncogenic effect of *SF3B1* mutation.

Chapter 1.5: Methods

Chapter 1.5.1: Sample selection

Ethics statement

For the chronic lymphocytic leukemia (CLL) samples, the UCSD IRB approved the study and all subjects gave informed consent (Project #080918). Refer to the informed consent for The Cancer Genome Atlas and Harbour *et al.* for consent information for other cancer samples (Harbour et al., 2013).

CLL

Seven *SF3B1*-mutated CLL cases and nine *SF3B1* wild-type CLL cases were identified from the CLL Consortium database. The mutations were originally characterized by PCR and verified in the RNA-sequencing data (Schwaederle et al., 2013). Sample dates were chosen on average 95 days prior to treatment and at least 287 days after prior treatment to select samples with high tumor cell count. Samples were chosen to have relatively similar numbers of *IGHV* mutated/unmutated and ZAP-70 positive/negative samples (Fig 4).

BRCA, LUAD, and LUSC

SF3B1 mutant samples were identified using the Broad GDAC TCGA analysis (http://gdac.broadinstitute.org/runs/analyses__2013_02_22/) in TCGA tumor types with no publication restrictions. Samples with *SF3B1* mutations outside of Gencode version 14 exons were excluded. We excluded any cancer types with less than four *SF3B1* mutants or for which paired-end RNA-

sequencing data was not available leaving breast cancer (BRCA), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC). We chose 1.25 as many *SF3B1* wild-type controls as mutated samples for each cancer type randomly from samples without mutations in *SF3B1* or other splicing factors. RNA sequencing data was downloaded from CGHub (Wilks et al., 2014).

UM

Uveal melanoma samples were downloaded from the Short Read Archive (SRA062359) (Harbour et al., 2013). As reported in Furney *et al.*, four uveal melanoma samples had *SF3B1* mutations in codon 625 and four had wild-type copies of *SF3B1* (Furney et al., 2013).

Chapter 1.5.2: Library preparation and sequencing for CLL samples

RNA was extracted from peripheral blood mononucleocytes from seven *SF3B1*-mutated CLL cases and nine *SF3B1* wild-type cases per the manufacturer's specifications using Qiagen RNeasy mini-spin columns, and RIN scores determined using an Agilent Bioanalyzer. RNA was polyA selected and processed using SMART cDNA synthesis (Clontech) to prepare sequencing libraries. Samples were sequenced on Illumina HiSeq2000 instruments generating an average of 239 million paired 75 bp reads per sample (Supplementary Figure 1.1).

Chapter 1.5.3: Adapter trimming

Sequencing adapters and poly-A/T tails were trimmed for CLL samples only using cutadapt version 1.1 (-m 20 –n 10 –b AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTT –b AAGCAGTGGTATCAACGCAGAGTACGCGGG –b AAGCAGTGGTATCAACGCAGAGT –b TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT –b AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA) (Marcel Martin, 2011). Read pairs where or one of both reads were of length less than 20 were removed.

Chapter 1.5.4: Read alignment

RNA-seq reads were aligned to the human genome (hg19) using STAR 2.3.0e (--alignSJDBoverhangMin 1 –seedSearchStartLmax 12 --alignSplicedMateMapLminOverLmate 0.08 –outFilterScoreMinOverLread 0.08 –outFilterMatchNminOverLread 0.08 –outFilterMultimapNmax 100 –outFilterIntronMotifs RemoveNoncanonicalUnannotated –outSJfilterOverhangMin 6 6 6 6) and a splice junction database consisting of junctions from Gencode, UCSC knownGene, AceView, lincRNAs, and H-Inv (Cabili et al., 2011; Dobin et al., 2013; Harrow et al., 2012; Hsu et al., 2006; Thierry-Mieg & Thierry-Mieg, 2006; Yamasaki et al., 2010). Duplicate read pairs were removed prior to

alignment by comparing the sequences of all read pairs and keeping only one read pair per set of read pairs with identical sequences.

Chapter 1.5.5: Splice junction read coverage

Splice junction read coverages were obtained from the SJ.out.tab output file from STAR.

Chapter 1.5.6: Novel splice junction identification

Novel splice junctions were defined as those junctions identified by STAR not present in Gencode version 14 that (i) were covered by at least 20 reads summed over all cancer samples in a given analysis, (ii) shared a 5' splice site and/or 3'SS with a Gencode junction, and (iii) had one of the following motifs: GU/AG, CU/AC, GC/AG, CU/GC, AU/AC, GU/AU. Novel junctions were calculated separately for each analysis.

Chapter 1.5.7: Splice junction usage

Known and novel junctions that had a coverage of at least 20 reads over all samples, used a known intron motif, and contained a known Gencode 5' splice site or 3'SS were aggregated by gene and tested for differential usage using DEXSeq's testForDEUTRT function (Anders et al., 2012). Splice junctions used in more than one Gencode gene were removed. When multiple cancer types were analyzed, we provided cancer type as a covariate to DEXSeq. Raw *p*-values were adjusted for multiple hypothesis testing using the Benjamini

Hochberg procedure. To examine the impact of the coverage cutoff of 20 reads summed over all samples on our results, we increased the cutoff to 50, 75, and 100 reads summed over all samples and found that 42%, 32%, and 24% of the significant novel 3'SSs remained at each of these cutoffs. The enrichment for proximal cryptic 3'SS remained at all cutoffs, so we used the 20 read cutoff to maximize sensitivity.

Chapter 1.5.8: Identification of associated canonical 3'SSs for cryptic 3'SSs

Associated canonical 3'SSs were identified for novel/cryptic 3'SSs as follows. First, all Gencode splice sites that shared a 5' splice site with the novel 3'SS were identified. Then, the closest Gencode 3'SS from these splice sites that was downstream of the cryptic 3'SS was chosen as the associated canonical 3'SS for that cryptic 3'SS. If there was no Gencode 3'SS downstream of the cryptic 3'SS, the closest Gencode 3'SS upstream of the cryptic 3'SS was chosen as the associated canonical 3'SS.

Chapter 1.5.9: Gene set enrichment for genes with cryptic 3'SS usage

We performed a gene set enrichment analysis using GSEA (Subramanian et al., 2005) for the genes that contained cryptic 3'SSs by combining the genes that contained the 619 proximal (File S3) and the 417 distal cryptic (File S4).

Chapter 1.5.10: Identification of control 3'SSs

We identified 23,065 control 3'SSs by choosing splice sites that are annotated in Gencode, whose average coverage over BRCA, CLL, and UM samples is greater than 100, and whose 5' splice site does not have any novel 3'SSs. We characterized intronic AG dinucleotides for these control junctions by analyzing the intronic sequence downstream of the predicted branch points minus the last 10 bp of the intron since 3'SSs can be located in the last 10 bp of the intron.

Chapter 1.5.11: Hierarchical clustering

All heatmap rows and columns were clustered using scipy.cluster.hierarchy.linkage with either the "complete" or "single" distance metric.

Chapter 1.5.12: *SF3B1* mutant allele frequency

Mutant allele frequency was determined by calculating per-base coverages using unique properly paired reads with samtools mpileup for the *SF3B1* locus and counting the number of reads supporting either the reference or alternate alleles.

Chapter 1.5.13: Gene expression

Reads that were not contained within Gencode v14 exons in the STAR genomic alignment were discarded. The remaining reads were re-aligned to the Gencode v14 transcriptome using Bowtie2 (v2.1.0, -t -k 400 -X 400 --no-mixed --no-discordant) and transcript expression was estimated using eXpress (v1.3.0, --max-indel-size 20) (Langmead & Salzberg, 2012; Roberts & Pachter, 2013). Gene expression was estimated by summing together the effective counts or FPKM values for all transcripts contained in a gene.

Chapter 1.5.14: Relative average expression of genes with cryptic 3'SSs

For the green heatmap in Fig 1D, the average expression (FPKM) of each gene containing a cryptic 3'SS was determined for each cancer type. The average expression values were then normalized for each gene by dividing by the largest average expression of the three cancers for that gene. Therefore each column in the green heatmap in Fig 1D has one value of 1.0 while the other two values are between 0.0 and 1.0 and represent the expression of the gene in that cancer relative to the maximum.

Chapter 1.5.15: Definition of HEAT repeats

HEAT repeat locations were defined according to the definition of HEAT repeats in Wang *et al.* 1998 (C. Wang et al., 1998).

Chapter 1.5.16: COSMIC SF3B1 mutations

COSMIC v66 complete export was downloaded and the number of mutations at each location in the *SF3B1* heat domains 5-9 was plotted for locations with at least two observed mutations in COSMIC (Forbes et al., 2011).

Chapter 1.5.17: Nucleotide frequency plots

Nucleotide frequency plots were constructed using WebLogo (unit_name='probability') (Crooks et al., 2004). Adenine enrichment was calculated by counting the number of adenines and non-adenines at each intron position for a given splice site class and comparing to the number of adenines and non-adenines in control 3'SSs using a Fisher exact test.

Chapter 1.5.18: Branch point identification

SVM_BP was used to predict branch points (Corvelo et al., 2010). The SVM_BP code was altered to allow for branch points eight bp from the 3'SS by setting mindist3ss=3 in svm_getfeat.py (see https://github.com/cdeboever3/svm-bpfinder). SVM_BP was run with options "Hsap 50." When multiple branch points were predicted for one 3'SS, we chose the branch point with the highest sequence score (bp_scr). In some instances, there was more than one cryptic 3'SS associated with a canonical 3'SS, so we randomly chose only one of these cryptic splice sites for further analysis. For Fig 3C, we plotted the distance from highest scoring BP predicted for canonical 3'SSs to their associated cryptic 3'SSs

as in Fig 3A. However, the distances for cryptic 3'SSs located less than 13 bp or more than 17 bp from the BP in Fig 3A were replaced with the distance from the second highest scoring BP. Supplementary Figure 1.5 C and D were created similarly.

Chapter 1.5.19: Differential gene expression

Gene expression was estimated as described above. We summed the effective counts from eXpress for all transcripts from each gene to obtain effective read counts for each gene. We provided these read counts to DESeq2 (v1.2.10, R v3.0.3) and tested for differential gene expression using nbinomWaldTest using cancer type as a covariate for the analysis with different cancers (Anders & Huber, 2010). We only tested genes where the sum of effective read counts over all samples was greater than 100. *p*-values were adjusted using the Benjamini-Hochberg procedure. Gene set enrichment analysis was performed using GSEA (Subramanian et al., 2005).

Chapter 1.5.20: Percent spliced in for cryptic 3'SSs relative to associated canonical 3'SSs

Percent spliced in (PSI) values for cryptic 3'SSs relative to canonical 3'SSs were calculated by dividing the number of reads that span the cryptic 3'SS (*c*) by the number of reads that span the cryptic 3'SS plus the number of reads that span the canonical 3'SS (*a*), $\frac{c}{c+a}$, for each sample. The ten 3'SSs with high

PSI values in CLL were identified by identifying cryptic 3'SSs whose median PSI was greater than 50% in the CLL *SF3B1* mutants but less than 20% in the wild-type samples and whose average coverage was at least 30 junction-spanning reads in the CLL mutant samples. These junctions were also chosen to be out-of-frame although the cryptic 3'SS in *ORAI2* is located in the 5' untranslated region.

Chapter 1.5.21: Code, data, and reproducibility

We have made the code and intermediate data files needed to replicate this study available on Github (https://github.com/cdeboever3/deboever-sf3b1-2015) and Figshare (http://dx.doi.org/10.6084/m9.figshare.1120663). Instructions are provided in the Github repository for reproducing our figures, tables, and statistical analyses. Sequencing data is available through dbGaP (phs000767).

Chapter 1.6: Contributions

Conceived and designed the experiments: EMG LR KJ CHMJ DC TJK KAF. Analyzed the data: CD PJS. Contributed reagents/materials/analysis tools: EMG LR CLB TJK. Wrote the paper: CD EMG LR KJ TJK KAF.

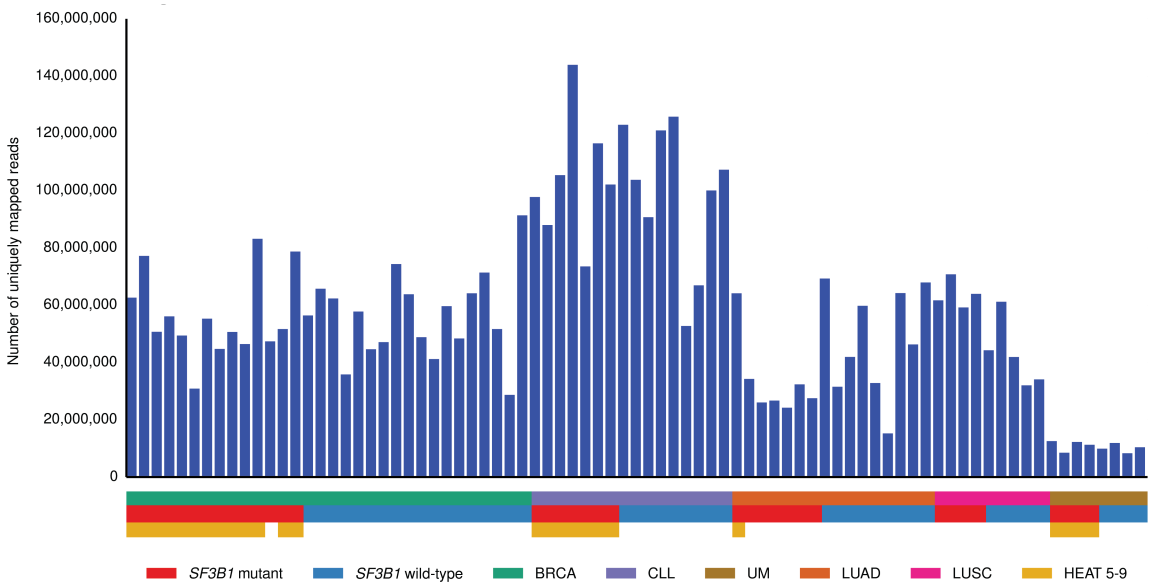
Chapter 1.7: Acknowledgements

The authors are grateful to the Chronic Lymphocytic Leukemia Research Consortium for providing the CLL samples and Marco A. Marra, Richard A.

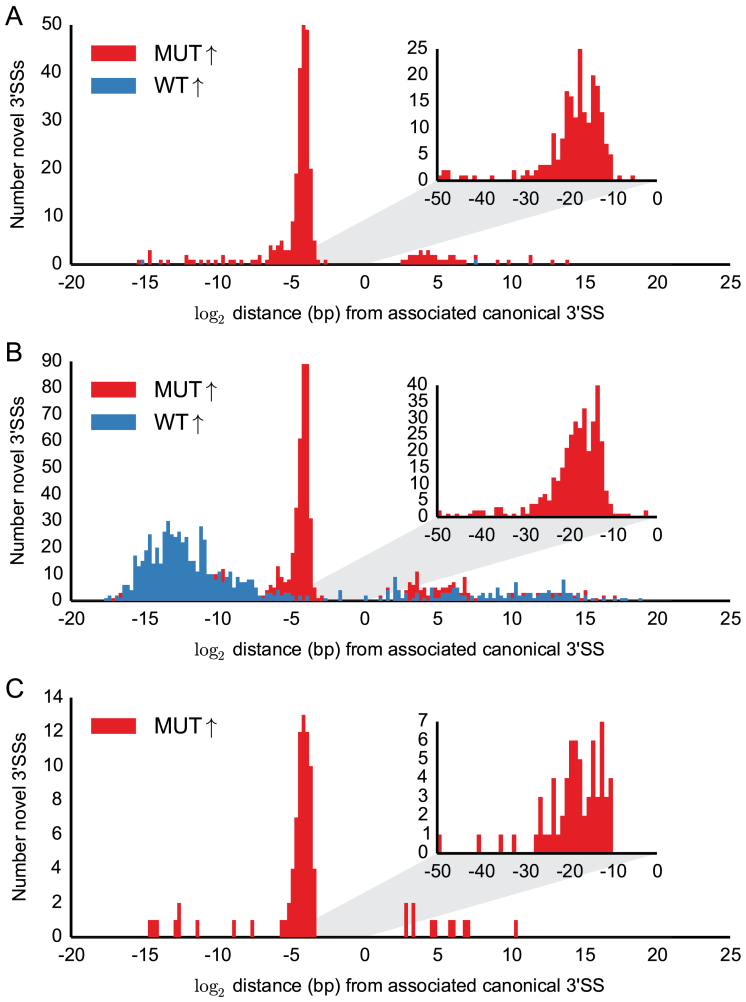
Moore, Joanne E. Johnson, Andrew J. Mungall and the Transcriptome Consortium at the Canada's Michael Smith Genome Sciences Centre for performing transcriptome sequencing for the CLL samples.

Chapter 1, in full, is a reprint of material as it appears in *PLoS Computational Biology* 2015, Christopher DeBoever, Emanuela M. Ghia, Peter J. Shepard, Laura Rassenti, Christian L. Barrett, Kristen Jepsen, Catriona H. M. Jamieson, Dennis Carson, Thomas J. Kipps, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

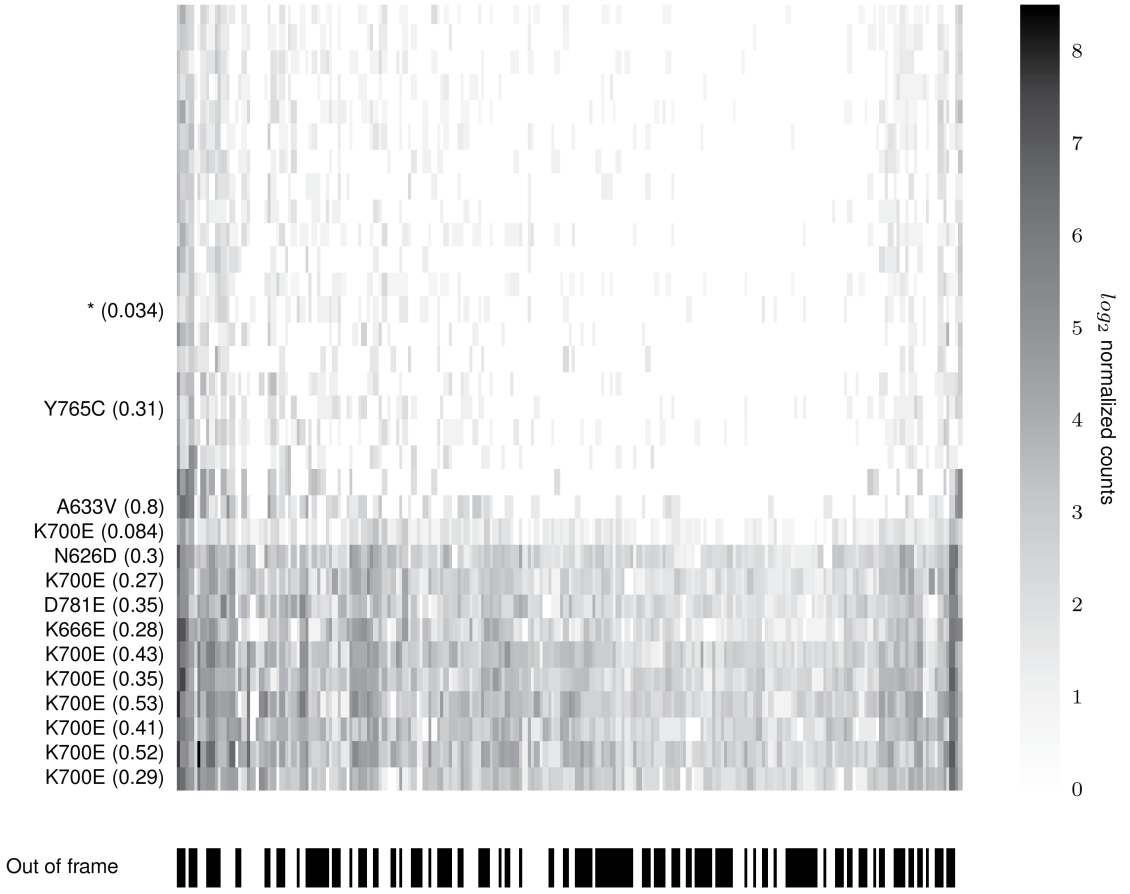
Chapter 1.8: Supplementary Figures



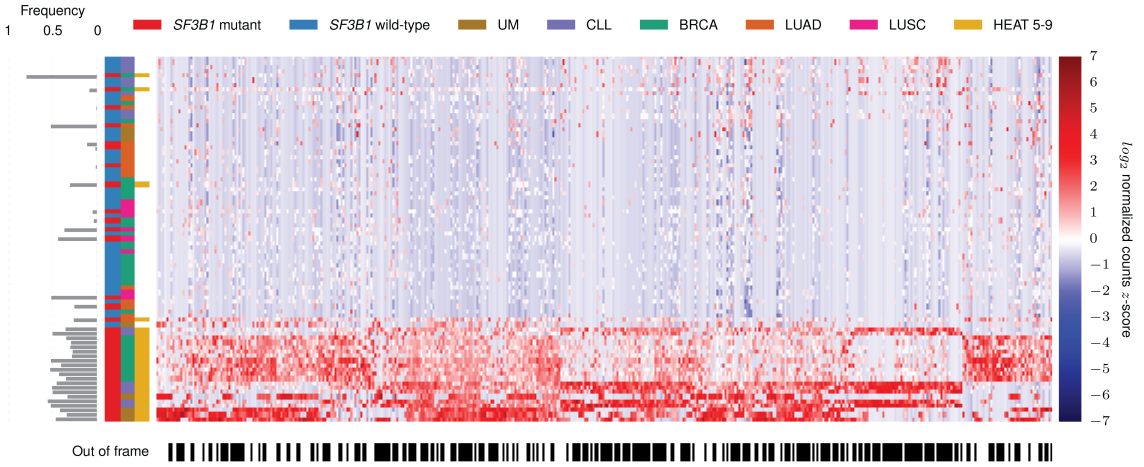
Supplementary Figure 1.1: Number of uniquely mapped RNA-seq reads from STAR alignment. We sequenced the transcriptomes of peripheral blood mononucleocytes from seven *SF3B1*-mutated chronic lymphocytic leukemia (CLL) cases and nine *SF3B1* wild-type cases. We also obtained data from breast cancer (BRCA; 14 mutant, 18 wild-type), lung squamous cell carcinoma (LUSC; four mutant, five wild-type) and lung adenocarcinoma (LUAD; seven mutant, nine wild-type) samples from the TCGA and uveal melanoma (UM; four mutant, four wild-type) samples from Harbour *et al.* 2013.



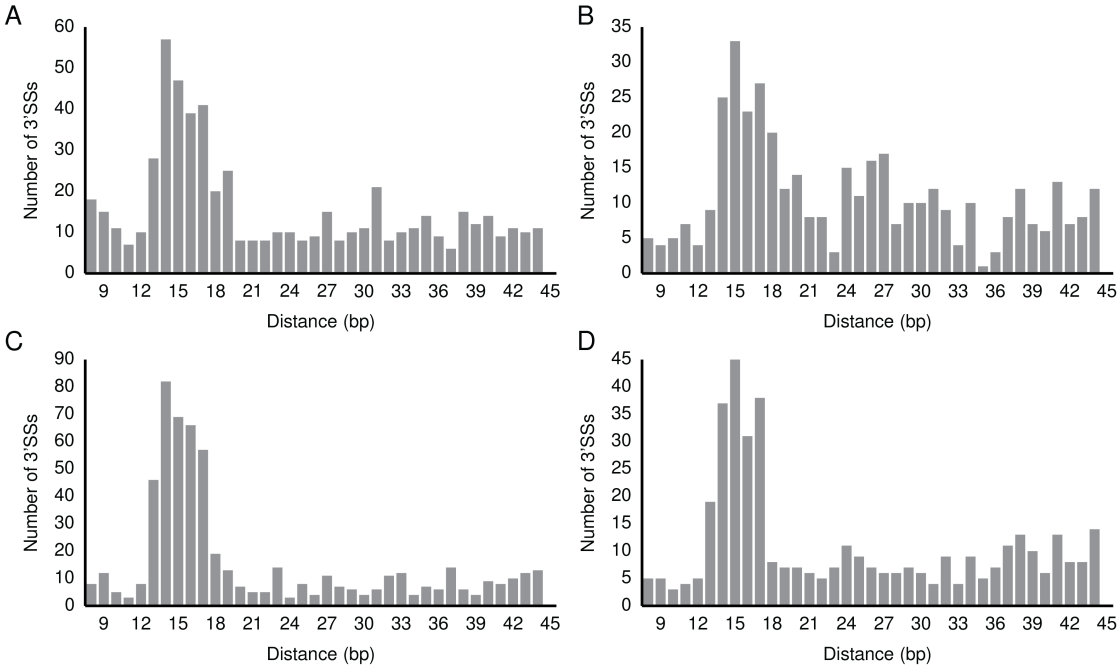
Supplementary Figure 1.2: Proximal cryptic 3'SSs in individual cancer analyses. $\log_2$ distance in base pairs from 280, 1,476, and 86 significantly differentially used novel 3'SSs (File S2) to their associated canonical 3'SSs in (A) BRCA, (B) CLL, and (C) UM analyses respectively. Novel 3'SSs were associated with canonical 3'SSs only if they shared the same 5' splice site. Zero represents the position of the canonical 3'SS. Negative and positive distances indicate that the cryptic 3'SS is respectively upstream or downstream from the canonical 3'SS. Inset shows base-by-base binning from zero to 50 base pairs upstream of canonical 3'SS. Red and blue histograms represent junctions with significantly higher usage in *SF3B1* mutants or *SF3B1* wild-type samples respectively. The number of cryptic 3'SS identified varied with the overall sequencing depth of the different data sets.



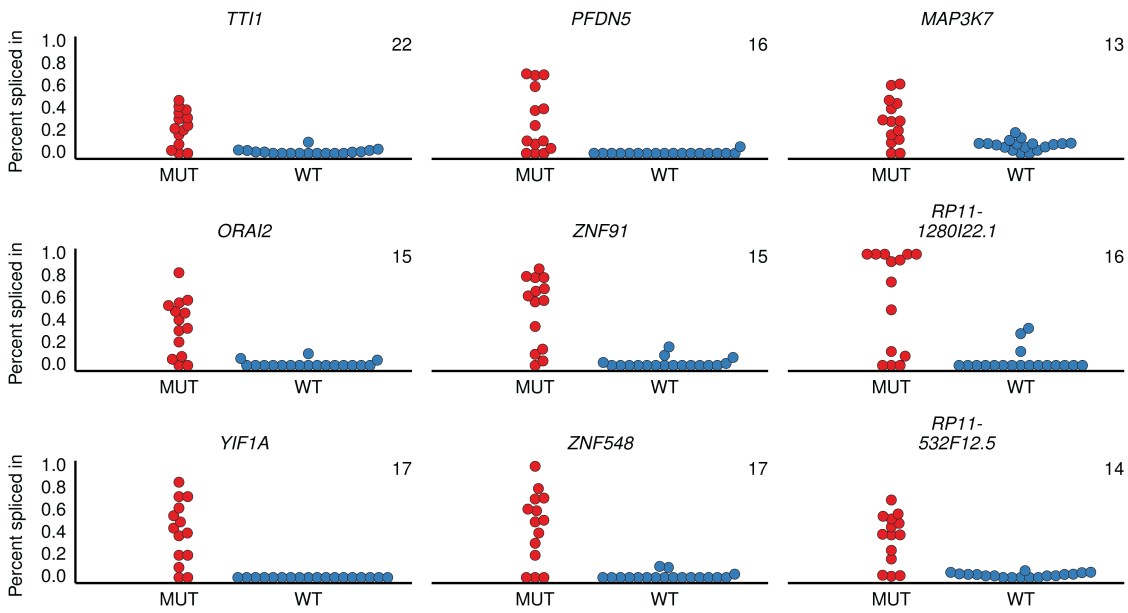
Supplementary Figure 1.3: Breast cancer proximal cryptic 3'SS coverage. Heatmap shows for each BRCA sample the $\log_2$ library-normalized count *z*-score for 192 proximal cryptic 3'SSs used significantly more often in the *SF3B1* mutants and located 10-30 bp upstream of canonical 3'SSs (File S2). *SF3B1* mutants are labeled with the observed missense or nonsense (*) mutation as well as the frequency of the mutant allele in the RNA-sequencing data. Attenuated cryptic 3'SS selection is visible for the K700E mutant with only 8.4% allele frequency. A633V and Y765C mutants do not show evidence for cryptic 3'SS selection. Black and white colorbar indicates whether novel 3'SSs are out-of-frame (black) relative to canonical 3'SSs.



Supplementary Figure 1.4: Proximal cryptic 3'SSs used significantly more often in cancers with *SF3B1* hotspot mutations including TCGA lung cancer samples. Heatmap shows for each sample the $\log_2$ library-normalized count *z*-score for the 578 proximal cryptic 3'SSs used significantly more often in the *SF3B1* mutants in the CLL, BRCA, UM, LUAD, and LUSC joint analysis (File S2). Grey bars indicate frequency of *SF3B1* mutant allele in RNA-seq data. Colorbars indicate *SF3B1* mutation status, cancer type, and whether the *SF3B1* mutation is located in the HEAT 5-9 repeats. Black and white colorbar indicates whether novel 3'SSs are out-of-frame (black) relative to canonical 3'SSs.



Supplementary Figure 1.5: Cryptic 3'SSs have branch points located ~13-17 bp upstream. Distance from 3'SS to highest scoring predicted branch point (BP). We were able to predict BPs for (A) 584 of 619 proximal cryptic 3'SSs and (B) 405 of 417 distal cryptic 3'SSs (as opposed to predicting the BPs for the associated canonical 3'SSs as in Fig 3). Distance from either highest or second highest scoring predicted BP to (C) proximal cryptic 3'SSs and (D) distal cryptic 3'SSs. Cryptic 3'SSs that are used more often in *SF3B1* mutants have BPs located ~13-17 bp upstream regardless of whether they are 10-30 bp upstream of canonical 3'SSs.



Supplementary Figure 1.6: Percent spliced in (PSI) in BRCA analysis for junctions with high PSI in CLL analysis. Beeswarm plots showing the PSI values for the cryptic 3'SS relative to the associated canonical 3'SS in nine of ten genes with high levels of cryptic 3'SS inclusion in CLL *SF3B1* mutants (M) compared to wild-type (W) samples that were also expressed in the BRCA samples. The number in the upper corner of each plot is the distance in base pairs from the highest or second-highest scoring BP predicted for the associated canonical 3'SS to the cryptic 3'SS.

Chapter 1.9: References

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology, 11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research, 22*(10), 2008-2017. doi:10.1101/gr.133744.111

Biankin, A. V., Waddell, N., Kassahn, K. S., Gingras, M. C., Muthuswamy, L. B., Johns, A. L., . . . Grimmond, S. M. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature, 491*(7424), 399-405. doi:10.1038/nature11547

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic

noncoding RNAs reveals global properties and specific subclasses. *Genes & development, 25*(18), 1915-1927. doi:Doi 10.1101/Gad.17446611

Chua, K., & Reed, R. (2001). An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. *Molecular and Cellular Biology, 21*(5), 1509-1514. doi:Doi 10.1128/Mcb.21.5.1509-1514.2001

Corvelo, A., Hallegger, M., Smith, C. W. J., & Eyras, E. (2010). Genome-Wide Association between Branch Point Properties and Alternative Splicing. *Plos Computational Biology, 6*(11). doi:Artn E1001016
Doi 10.1371/Journal.Pcbi.1001016

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research, 14*(6), 1188-1190. doi:Doi 10.1101/Gr.849004

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Fernandez-Saiz, V., Targosz, B. S., Lemeer, S., Eichner, R., Langer, C., Bullinger, L., . . . Bassermann, F. (2013). SCFFbxo9 and CK2 direct the cellular response to growth factor withdrawal via Tel2/Tti1 degradation and promote survival in multiple myeloma. *Nature Cell Biology, 15*(1), 72-U164. doi:Doi 10.1038/Ncb2651

Ferreira, P. G., Jares, P., Rico, D., Gomez-Lopez, G., Martinez-Trillos, A., Villamor, N., . . . Guigo, R. (2013). Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Research*. doi:10.1101/gr.152132.112

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., . . . Futreal, P. A. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research, 39*, D945-D950. doi:Doi 10.1093/Nar/Gkq929

Fujioka, Y., Taira, T., Maeda, Y., Tanaka, S., Nishihara, H., Iguchi-Ariga, S. M., . . . Ariga, H. (2001). MM-1, a c-Myc-binding protein, is a candidate for a tumor suppressor in leukemia/lymphoma and tongue cancer. *Journal of Biological Chemistry, 276*(48), 45137-45144. doi:10.1074/jbc.M106127200

Furney, S. J., Pedersen, M., Gentien, D., Dumont, A. G., Rapinat, A., Desjardins, L., . . . Marais, R. (2013). SF3B1 mutations are associated with alternative

splicing in uveal melanoma. *Cancer discovery*. doi:10.1158/2159-8290.CD-13-0330

Gao, K. P., Masuda, A., Matsuura, T., & Ohno, K. (2008). Human branch point consensus sequence is yUnAy. *Nucleic Acids Research, 36*(7), 2257-2267. doi:Doi 10.1093/Nar/Gkn073

Gozani, O., Potashkin, J., & Reed, R. (1998). A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Molecular and Cellular Biology, 18*(8), 4752-4760.

Harbour, J. W., Roberson, E. D. O., Anbunathan, H., Onken, M. D., Worley, L. A., & Bowcock, A. M. (2013). Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nature genetics, 45*(2), 133-135. doi:Doi 10.1038/Ng.2523

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., . . . Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research, 22*(9), 1760-1774. doi:Doi 10.1101/Gr.135350.111

Hofer-Warbinek, R., Schmid, J. A., Stehlik, C., Binder, B. R., Lipp, J., & de Martin, R. (2000). Activation of NF-kappa B by XIAP, the X chromosome-linked inhibitor of apoptosis, in endothelial cells involves TAK1. *The Journal of biological chemistry, 275*(29), 22064-22068. doi:10.1074/jbc.M910346199

Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M., & Haussler, D. (2006). The UCSC Known Genes. *Bioinformatics, 22*(9), 1036-1046. doi:Doi 10.1093/Bioinformatics/Btl048

Hurov, K. E., Cotta-Ramusino, C., & Elledge, S. J. (2010). A genetic screen identifies the Triple T complex required for DNA damage signaling and ATM and ATR stability. *Genes & development, 24*(17), 1939-1950. doi:10.1101/gad.1934210

Kaizuka, T., Hara, T., Oshiro, N., Kikkawa, U., Yonezawa, K., Takehana, K., . . . Mizushima, N. (2010). Tti1 and Tel2 are critical factors in mammalian target of rapamycin complex assembly. *The Journal of biological chemistry, 285*(26), 20109-20116. doi:10.1074/jbc.M110.121699

Kimura, N., Matsuo, R., Shibuya, H., Nakashima, K., & Taga, T. (2000). BMP2-induced apoptosis is mediated by activation of the TAK1-p38 kinase pathway that is negatively regulated by Smad6. *Journal of Biological Chemistry, 275*(23), 17647-17652. doi:Doi 10.1074/Jbc.M908622199

Kralovicova, J., Christensen, M. B., & Vorechovsky, I. (2005). Biased exon/intron distribution of cryptic and de novo 3' splice sites. *Nucleic Acids Res, 33*(15), 4882-4898. doi:10.1093/nar/gki811

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Martin, M. (2011). *Cutadapt removes adapter sequences from high-throughput sequencing reads* (Vol. 17).

Martin, M., Masshofer, L., Temming, P., Rahmann, S., Metz, C., Bornfeld, N., . . . Zeschnigk, M. (2013). Exome sequencing identifies recurrent somatic mutations in EIF1AX and SF3B1 in uveal melanoma with disomy 3. *Nature genetics, 45*(8), 933-U296. doi:Doi 10.1038/Ng.2674

Nam, J. S., Hirohashi, S., & Wakefield, L. M. (2007). Dysadherin: A new, player in cancer progression. *Cancer letters, 255*(2), 161-169. doi:Doi 10.1016/J.Canlet.2007.02.018

Padgett, R. A. (2012). New connections between splicing and human disease. *Trends in Genetics, 28*(4), 147-154. doi:Doi 10.1016/J.Tig.2012.01.001

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., . . . Campbell, P. J. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *The New England journal of medicine, 365*(15), 1384-1395. doi:10.1056/NEJMoa1103283

Quesada, V., Conde, L., Villamor, N., Ordonez, G. R., Jares, P., Bassaganyas, L., . . . Lopez-Otin, C. (2012). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature genetics, 44*(1), 47-52. doi:10.1038/ng.1032

Roberts, A., & Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods, 10*(1), 71-U99. doi:Doi 10.1038/Nmeth.2251

Schwaederle, M., Ghia, E., Rassenti, L. Z., Obara, M., Dell'Aquila, M. L., Fecteau, J. F., & Kipps, T. J. (2013). Subclonal evolution involving SF3B1 mutations in chronic lymphocytic leukemia. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 27*(5), 1214-1217. doi:10.1038/leu.2013.22

Smith, C. W. J., Chu, T. T., & Nadalginard, B. (1993). Scanning and Competition between Ags Are Involved in 3' Splice-Site Selection in Mammalian Introns. *Molecular and Cellular Biology, 13*(8), 4939-4952.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550. doi:Doi 10.1073/Pnas.0506580102

Thierry-Mieg, D., & Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome biology, 7*. doi:Artn S12
Doi 10.1186/Gb-2006-7-S1-S12

Wan, Y., & Wu, C. J. (2013). SF3B1 mutations in chronic lymphocytic leukemia. *Blood, 121*(23), 4627-4634. doi:10.1182/blood-2013-02-427641

Wang, C., Chua, K., Seghezzi, W., Lees, E., Gozani, O., & Reed, R. (1998). Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. *Genes & development, 12*(10), 1409-1414.

Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., . . . Wu, C. J. (2011). SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *The New England journal of medicine, 365*(26), 2497-2506. doi:10.1056/NEJMoa1109016

Watson, I. R., Takahashi, K., Futreal, P. A., & Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nature reviews. Genetics*. doi:10.1038/nrg3539

Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., . . . Maltbie, D. (2014). The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database (Oxford), 2014*. doi:10.1093/database/bau093

Yamaguchi, K., Nagai, S., Ninomiya-Tsuji, J., Nishita, M., Tamai, K., Irie, K., . . . Matsumoto, K. (1999). XIAP, a cellular member of the inhibitor of apoptosis protein family, links the receptors to TAB1-TAK1 in the BMP signaling pathway. *The EMBO journal, 18*(1), 179-187. doi:10.1093/emboj/18.1.179

Yamasaki, C., Murakami, K., Takeda, J., Sato, Y., Noda, A., Sakate, R., . . . Gojobori, T. (2010). H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Research, 38*, D626-D632. doi:Doi 10.1093/Nar/Gkp1020

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., . . . Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature, 478*(7367), 64-69. doi:10.1038/nature10496

Chapter 2: Genetic Regulation of Gene Expression in Human Induced Pluripotent Stem Cells

Chapter 2.1: Summary

In this study, we identified genetic variants associated with RNA expression for 5,619 genes using 215 human induced pluripotent stem cell (iPSC) lines from different donors. These expression quantitative trait loci (eQTLs) were enriched in stem cell regulatory regions and have evidence for disrupting transcription factor binding. We used whole genome sequencing to identify copy number variant (CNV) eQTLs, including some that appear to affect gene expression by altering the copy number of intergenic regulatory regions. We found that rare genic CNVs have a relatively strong effect on gene expression that is positively correlated with copy number, whereas rare regulatory single nucleotide variants have a weak negative effect. Additionally, X chromosome gene reactivation in female-derived iPSCs was dependent on gene chromosomal position. This work demonstrates the utility of iPSCs for genetic association analyses and provides a unique resource for investigating the genetic regulation of gene expression in stem cells.

Chapter 2.2: Introduction

Since their discovery 10 years ago, induced pluripotent stem cells (iPSCs) have been used to model a multitude of “diseases in a dish” by utilizing lines

derived from a relatively small number of diseased and healthy donors (Avior, Sagi, & Benvenisty, 2016; Takahashi et al., 2007; Takahashi & Yamanaka, 2006). Several recent initiatives have begun to scale the generation of iPSC lines to create large banks of hundreds or thousands of iPSCs derived from diverse donors for studying stem cells and differentiated tissues in a variety of genetic backgrounds (Martin, 2015; McKernan & Watt, 2013). Due to iPSCs' capacity for self-renewal, these banks potentially provide an unprecedented opportunity for performing genetic association analyses (Pai, Pritchard, & Gilad, 2015) and investigating developmental phenomena like X chromosome inactivation during reprogramming (Lessing, Anguera, & Lee, 2013; Pasque & Plath, 2015). While there is evidence suggesting genetic association studies will be possible in iPSCs (Rouhani et al., 2014; Thomas et al., 2015), such analyses could be confounded by non-genetic factors affecting expression such as reprogramming heterogeneity, somatic mutations (Gore et al., 2011; J. Ji et al., 2012), or epigenetic drift during passaging (Papp & Plath, 2013). Thus the suitability of these large sets of iPSCs for examining the effects of inherited genetic variants on molecular and physiological phenotypes remains largely unknown.

Expression quantitative trait loci (eQTL) mapping is a type of genetic association analysis that identifies genomic regions that harbor polymorphisms associated with the RNA expression of a gene. Over the last 15 years, eQTL mapping has been performed in a variety of cell types and model organisms and has contributed to our understanding of how genetic variants regulate gene

expression (Albert & Kruglyak, 2015). eQTL mapping has not yet been performed in iPSCs due to the lack of hundreds of systematically reprogrammed lines with corresponding genotype and gene expression data, and thus the statistical power to map eQTLs in iPSCs compared to other cell types is not known. eQTL studies conducted to date have largely utilized arrays and low-depth whole genome sequencing (WGS) for genotyping. While these methods can accurately genotype single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels), neither method is ideal for identifying and genotyping copy number variants (CNVs). Therefore, though CNVs likely play an important role in human diseases (Gamazon, Nicolae, & Cox, 2011; Sudmant et al., 2015), their effects on gene expression are not well understood. High-depth WGS combined with new CNV-calling algorithms that utilize information across samples to identify CNVs and estimate integer copy number genotypes (Handsaker et al., 2015; Layer, Chiang, Quinlan, & Hall, 2014) greatly enhances our ability to investigate the contributions of and mechanisms by which CNVs regulate gene expression.

Rare variants (minor allele < 0.5% in general population) constitute another class of variation that has been poorly assessed in previous eQTL studies despite their established importance for disease (U. K. Consortium et al., 2015; Shendure & Akey, 2015; Zanoni et al., 2016) because they are generally not included on genotyping arrays and are harder to identify using low-depth WGS. Recent studies have begun to investigate the effect of the vast number of

rare regulatory variants on gene expression by leveraging unique family structures (Montgomery, Lappalainen, Gutierrez-Arcelus, & Dermitzakis, 2011) or deep targeted sequencing (Zhao et al., 2016) but no studies using high-depth WGS to identify rare variants and quantify their effect on gene expression in a large set of subjects have been reported yet. Thus the extent to which rare variants contribute to gene expression is not known, and it remains difficult to predict which of the estimated 40k-200k rare variants per genome may affect gene expression (Genomes Project et al., 2015).

In this study, we leverage high-depth WGS to explore the genetic regulation of gene expression in a set of 215 iPSC lines. We demonstrate that iPSCs are well-powered for eQTL mapping and have a distinct regulatory landscape relative to somatic tissues. We functionally annotate the iPSC eQTLs and show they are enriched in stem cell regulatory elements and for overlapping the binding sites of transcription factors (including NANOG and POU5F1) important for establishing and maintaining pluripotency. To identify putative causative variants underlying the eQTL signals, we identify variants that are both associated with gene expression and alter transcription factor binding. We observe that a large proportion of common CNVs associated with gene expression levels are located in intergenic regulatory regions. We also find that rare genic CNVs have relatively large effects on gene expression that can be positive or negative dependent on their location relative to the gene while rare promoter SNVs overall have a small negative effect on gene expression. Finally,

we investigate X chromosome reactivation during reprogramming for iPSC lines from female donors and find that overall X reactivation is heterogeneous across lines but that the reactivation statuses of nearby genes are correlated. This work establishes iPSCs as a useful model for quantitative molecular association studies, identifies genetic regulators of gene expression in iPSCs, provides new information about the impact of CNVs and rare variants on gene expression, and reveals novel insights into X chromosome reactivation during reprogramming.

Chapter 2.3: Results

To investigate the genetic regulation of gene expression in iPSCs, we generated 30x germline WGS and RNA sequencing (RNA-seq) data for 215 human iPSC lines from a diverse set of donors (median age 48.3, 55% female) consisting of both unrelated individuals as well as families (Supplementary Figure 2.1) (Frazer, 2016). The donors represent several ancestries although the majority (66%) are European. We used the high-depth WGS data to identify 22,461,624 single nucleotide variants (SNVs) and insertions/deletions (indels) using GATK and 15,735 CNVs using LUMPY and GenomeSTRiP after filtering for 1% minor allele frequency among our 215 subjects and violations of Hardy Weinberg equilibrium (Methods) (Handsaker et al., 2015). We compared the expression levels of nine pluripotency and 25 mesoderm markers from (Tsankov et al., 2015) in our iPSCs to publicly available RNA-seq data from human embryonic stem cells (hESCs), iPSCs, and fibroblasts and found little or no

expression of mesoderm markers but high expression of pluripotency markers in our iPSCs compared to fibroblasts and other stem cell lines indicating that our stem cell lines are of high quality (Supplementary Figure 2.2).

Chapter 2.3.1: eQTL mapping in iPSCs

We used gene expression estimates from RNA-seq and germline variant calls to map eQTLs in 215 CARDiPS iPSC lines from different donors (Frazer, 2016). We estimated expression using RSEM (transcripts per million, TPM) and used PEER to remove confounding variation from the gene expression estimates to increase our power for identifying *cis* eQTLs (B. Li & Dewey, 2011; Stegle, Parts, Durbin, & Winn, 2010). For each gene, we identified biallelic variants, including indels and CNVs, within 1Mb of a transcription start site (TSS) and used EMMAX (Kang et al., 2010) to calculate association *p*-values that accounted for relatedness amongst our donors. We then calculated per-gene *p*-values using a permutation approach (G. T. Consortium, 2015) and performed multiple hypothesis testing correction (Storey & Tibshirani, 2003). Of the 17,819 autosomal genes tested we found 5,619 (32%) with eQTLs (eGenes) including 4,495 protein coding genes (33% of tested), 356 long non-coding RNAs (34% of tested), 342 pseudogenes (25% of tested), and 269 antisense genes (29% of tested) (Figure 2.1A, Supplementary File 2.1). An eQTL typically contains multiple variants associated with a gene's expression due to the linkage disequilibrium structure of the human genome. For instance, the 5,619 eGenes

identified here have in total 253,231 significant variant-expression associations involving 195,589 unique variants. In total there were 4,892, 1,346, and 111 eGenes with SNV, indel, and CNV lead variants respectively (some eGenes had multiple variants with equal significance) (Supplementary File 2.2). Consistent with previous eQTL studies (G. T. Consortium, 2015), lead variants were enriched around the transcription start sites (TSSs) of genes (Supplementary Figure 2.3) and eGenes were highly enriched for allele specific expression (ASE) (OR = 3.1, $p < 10^{-292}$, Fisher exact test) which supports the presence of *cis* regulatory effects at these loci. We also found on average 93% agreement for lead SNV direction of effect compared to 44 GTEx v6 tissues demonstrating that our eQTLs are of high quality (G. T. Consortium, 2015). We tested for additional independent eQTLs in the 5,619 eGenes by using the lead variant as a covariate and found that 668 of the 5,619 eGenes had a second independent eQTL and 201 had a third eQTL.

Since gene expression is often used to estimate stem cell pluripotency, we compared our eGenes to nine stem cell marker genes from (Tsankov et al., 2015) and found that four (*CXCL5*, *IDO1*, *LCK*, and *POU5F1*) had eQTLs (Figure 2.1B-C, Supplementary File 2.2). The lead variants for these four genes explained respectively 18%, 11%, 6%, and 19% of the variance in gene expression in a model using only batch, sex, and donor age as covariates demonstrating that the genetic background of an iPSC line could affect estimations of pluripotency based on gene expression markers. We also

identified eQTLs for 35 of 191 genes involved in stem cell population maintenance (GO:0019827) such as the oncogene *BCL9* and the developmental regulator *FGFR1* (Ashburner et al., 2000) (Figure 2.1D-E, Supplementary File 2.2) indicating that eQTLs might affect maintenance of pluripotency or differentiation.

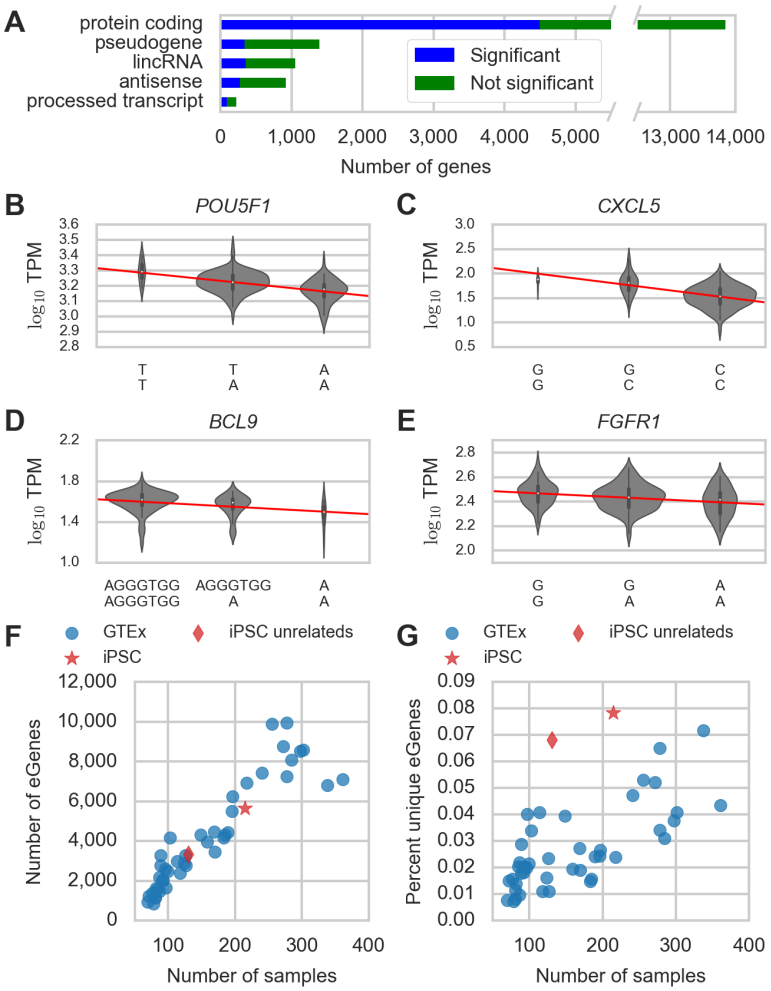


Figure 2.1 Summary of eQTL Results and Power Analysis. (A) Number of genes tested (green) and significant (blue) by Gencode gene type (see Supplementary File 2.1 for all gene types). (B-E) log10 RSEM TPM gene expression estimates stratified by lead variant genotype for (B) *POU5F1*, (C) *CXCL5*, (D) *BCL9*, and (E) *FGFR1*. The x-axis is labeled with the genotypes for the lead variant for each eQTL. We used residual expression values to identify eQTLs but plot raw TPM here to demonstrate the effect of the eQTL on the raw expression data. (F) Number of eGenes and (G) percent unique eGenes versus number of samples for 43 GTEx v6 tissues (blue

circles), 131 unrelated subjects from this study (red diamond), or all 215 subjects from this study (red star).

To investigate if the power to detect eQTLs in iPSCs is diminished by non-genetic sources of variation in gene expression, we compared the number of eGenes discovered in our study to the number identified in 43 GTEx v6 tissues, taking sample numbers in both studies into account (Figure 2.1F). Although we used EMMAX to account for relatedness among our 215 subjects, using related subjects can affect the power to detect eQTLs since some subjects may share substantial portions of their genomes. Therefore, to more accurately compare to the GTEx results, we performed our eQTL analysis again using 131 of the 215 samples that were genetically unrelated and found eQTLs for 3,310 of 17,819 genes compared to 5,619 eGenes for all 215 samples. The number of eGenes for both the 131 unrelateds and full 215 samples follow the same general trend observed in the GTEx data of an increase of about 30 eGenes per additional sample indicating that iPSCs are powered similarly to GTEx tissues for detecting eQTLs (Figure 2.1F). Since GTEx mostly focuses on somatic tissues, we hypothesized that the iPSCs might contain more unique eGenes (i.e. not found in other tissue types) than a typical GTEx tissue. To test this, we compared the percentage of eGenes that were unique to a given tissue relative to all GTEx eGenes plus the iPSC eGenes reported here (Figure 2.1G). GTEx tissues with more samples have a higher percentage of unique eGenes, with an increase of roughly 1.3% unique eGenes per 100 samples, likely reflecting the discovery of small effect size, tissue-specific eQTLs. Given this trend in the GTEx tissues, we

would expect 2.3% (95% confidence interval [0.1%, 4.4%]) of the 3,310 eGenes identified using the 131 iPSCs to be unique to iPSCs but instead observed that 6.8% of these eGenes are unique to iPSCs. These results demonstrate that iPSCs are well-powered for identifying eQTLs and that the gene regulatory landscape of iPSCs differs significantly compared to the primary tissues and transformed cell lines in GTEx.

Chapter 2.3.2: iPSC eQTLs Enriched in Stem Cell Regulatory Regions

To determine whether our eQTLs correspond to annotated stem cell regulatory regions, we calculated the enrichment of 4,491 noncoding lead eQTL SNVs and indels in DNase hypersensitivity sites (DHSs) from 53 Roadmap Epigenomics cell types by determining if lead variants overlapped DHSs more often than expected given the density of DHSs in 5kb windows centered on the lead variants (Figure 2.2A, Tables S2 and S3) (G. T. Consortium, 2015; Roadmap Epigenomics et al., 2015). Although the lead eQTL variants are enriched in DHSs from most of the Roadmap cell types (most likely due to shared regulatory architecture across cell types), they are most enriched in DHSs from hESCs and iPSCs consistent with being located in stem cell regulatory regions. Lead variants also had a relatively high enrichment in DHSs from *in vitro* differentiated lines which likely reflects incomplete/heterogeneous differentiation or retention of some stem cell regulatory features in these lines. We also calculated the enrichment of noncoding lead SNVs and indels for 209 ENCODE

DHS experiments comprising 134 different cell types and again found that noncoding lead variants were most enriched in DHSs from stem cells followed by *in vitro* differentiated cells (Figure 2.2B, Supplementary File 2.3) (E. P. Consortium, 2012). The 209 ENCODE DHS experiments included nine skin fibroblast experiments that ranked from the 16th to the 207th most enriched, so there does not appear to be a strong signal of epigenetic memory for the original cell type (Supplementary File 2.3). The fact that lead variants are enriched in DHSs from both hESCs and iPSCs agrees with previous work showing that these two cell types have highly similar gene expression and epigenetic marks (Rouhani et al., 2014) and enables us to use the substantial amount of functional genomics data publicly available for the H1 hESC line to annotate our eQTLs. We calculated the enrichment of the 4,491 noncoding lead SNVs and indels among peaks from 49 ENCODE H1 hESC transcription factor (TF) ChIP-seq experiments and found that NANOG and POU5F1 were the most enriched TFs consistent with these factors' known roles in reprogramming and pluripotency (Figure 2.2C, Supplementary File 2.3). Overall, these results demonstrate that the eQTLs identified here are strongly enriched for overlapping stem cell regulatory regions and provides further evidence that functional genomics annotations from hESCs are relevant to iPSCs.

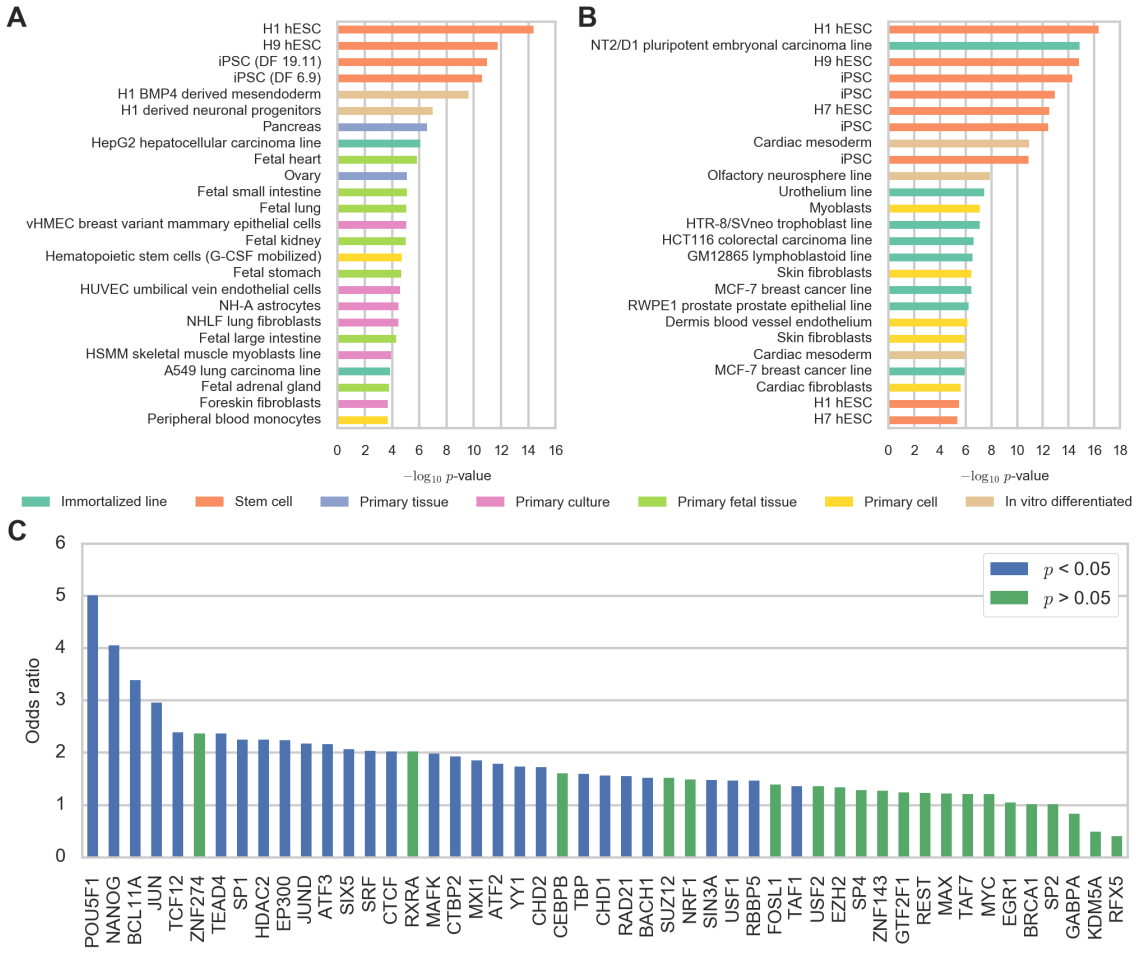


Figure 2.2: eQTL Functional Annotation Enrichments. -$\log_{10}$ Fisher exact *p*-values for 4,491 eQTL lead SNVs/indels in (A) Roadmap Epigenomics DNase hypersensitivity sites (DHSs) and (B) ENCODE DHSs. The replicate H1 and H7 hESC DHS experiments in (B) were performed in different laboratories which may account for their different levels of enrichment. (C) Fisher exact odds ratios for ENCODE H1 hESC transcription factor CHiP-seq peaks. Color indicates whether the enrichment was significant which can vary due to the number of ChIP-seq peaks for each particular mark.

Chapter 2.3.3: Disruption of Transcription Factor Binding Sites by eQTL Variants

Given that altered TF binding is thought to be one of the primary causes of eQTLs (Pai et al., 2015), we investigated how many eQTL SNVs and indels overlapped TF ChIP-seq peaks and disrupted motifs associated with those TFs. While an eQTL typically contains multiple variants due to linkage disequilibrium,

generally only one variant (not necessarily the lead variant) is the functional, or causal variant, termed the expression quantitative trait nucleotide (eQTN). To identify putative eQTNs (peQTNs) that disrupt TF binding, we focused on 5,437 of the 5,619 eGenes that did not overlap a CNV eQTL and did not have an eQTL predicted to cause NMD since these eQTLs are unlikely to be caused by altered TF binding. We overlapped the 186,656 eQTL SNVs and indels associated with the expression of these 5,437 eGenes with H1 hESC ChIP-seq peaks from 40 ENCODE experiments for 34 TFs (some TFs had multiple subunits assayed, like JUN and JUND for AP1) and identified 7,630 variants that overlapped a TF peak. We then predicted which of these 7,630 variants disrupted motifs associated with the overlapped TF ChIP-seq peak in (Kheradpour & Kellis, 2014) and found that 3,058 (40%) distinct variants disrupted motifs (Supplementary File 2.4). While the peQTNs are predicted to disrupt motifs enriched in ChIP-seq peaks for specific TFs, these motifs do not always correspond to the known motif for the particular TF. Previous studies show that peaks for most TFs are enriched for motifs of other TFs (Kheradpour & Kellis, 2014) which likely occurs for several reasons including cooperative/interfering binding or motif similarity. Accordingly, we find that 70% of our 3,058 peQTNs disrupt a motif that is associated with the particular ChIP TF but is similar to a known motif for a different TF highlighting the cooperative nature of TF binding.

In total, the 3,058 peQTNs we identified corresponded to 1,475 of the 5,437 eGenes. 50% of these 1,475 eGenes have only one peQTN and 92% have

five or less peQTNs indicating that most eGenes have few peQTNs (Figure 2.3A). Interestingly, a lead variant was a peQTN for only 19% of the 1,475 genes suggesting that lead variants may not often be the causal eQTL variant. Though we did not consider distance to the TSS when identifying peQTNs, 54% of the peQTNs were within 20kb of the nearest TSS for the associated eGene consistent with previous estimates of the distribution of eQTLs around the TSS (Wen, Luca, & Pique-Regi, 2015). We observed that 90% of the peQTNs overlap a DHS present in at least one of the four Roadmap stem cell lines (Figure 2.2A) and 61% overlap a DHS present in all four lines (Figure 2.3B). Figure 2.3C shows three example peQTNs for *POU5F1* that are within 1,700 bp of a *POU5F1* TSS, overlap TAF1 ChIP-seq peaks, and disrupt motifs associated with TATA TFs. One of these variants also overlaps an H1 hESC DHS and falls into a predicted strong enhancer. This variant disrupts motif TATA_disc7 from (Kheradpour & Kellis, 2014) which is highly similar to motifs for NFE2L2 and ETV6 (two TFs for which H1 ChIP-seq data is not available), suggesting that this variant may actually disrupt binding of one of these TFs. Overall, our observations are consistent with the peQTNs playing a role in the regulation of nearby genes and underscore the difficulty in interpreting eQTLs even for relatively well-characterized cell types with substantial amounts of functional genomics data.

We next sought evidence that the peQTNs we identified cause differential TF binding. (Maurano et al., 2015) tested ~360k heterozygous SNVs located in DHSs for allelic bias caused by differential TF binding *in vivo* and found that 18%

of tested variants affected TF binding. Of the 186,656 eQTL variants we used to identify peQTNs, (Maurano et al., 2015) assayed 13,366 including 974 peQTNs. We found that 37% of the 974 peQTNs showed evidence for altered TF binding in (Maurano et al., 2015) compared to only 19% of the 12,392 eQTL variants assayed by (Maurano et al., 2015) that we did not classify as peQTNs. Thus peQTNs are highly enriched for altering TF binding relative to eQTL variants that we did not classify as peQTNs (OR = 2.5, $p < 10^{-36}$, Fisher exact test) and relative to all ~360k variants tested by Maurano (OR = 2.7, $p < 10^{-43}$, Fisher exact test) indicating that these variants likely modulate gene expression via differential TF binding. peQTNs were also nearly five times more likely to interact with the promoter of the associated eGene according to ChIA-PET interactions from naive hESCs (OR = 4.6, $p < 10^{-20}$, Fisher exact test) (X. Ji et al., 2016). That our peQTNs are strongly enriched for variants that alter TF binding and interact with the promoter of the associated eGene suggests that these variants are good eQTN candidates.

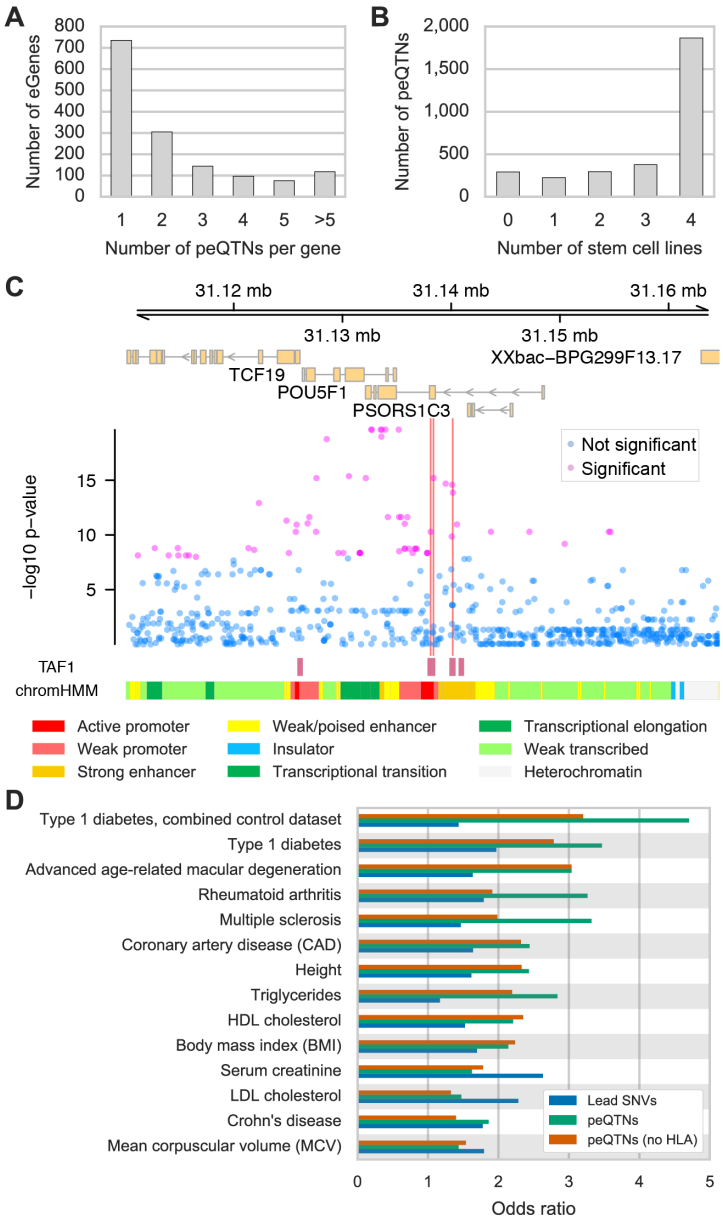


Figure 2.3: peQTN Characteristics and GWAS Enrichments. (A) Number of peQTNs per eGene for 1,475 eGenes with at least one peQTN. (B) Number of stem cell DHSs overlapped by eQTNs for four stem cell lines from Roadmap Epigenomics (H1, H9, iPS DF 6.9, iPS DF 19.11). (C) Putative eQTNs (red lines) for *POU5F1* eQTL that overlap TAF1 ChIP-seq peaks and disrupt motifs associated with TATA TFs. Scatter plot shows -$\log_{10}$ association *p*-value from EMMAX for variants that are significantly (purple points) and not significantly (blue points) associated with *POU5F1* expression. ENCODE H1 hESC TAF1 ChIP-seq peaks (red rectangles) and chromHMM chromatin state predictions (multi-color track) are displayed. (D) Enrichment odds ratios of lead eQTL SNVs and peQTNs among GWAS associations for traits and diseases from the GRASP database.

Chapter 2.3.4: iPSC eQTLs are Enriched Among GWAS Associations

To determine whether iPSC eQTLs are also associated with human diseases and phenotypes, we calculated the enrichment of eQTL lead SNVs and peQTNs amongst GWAS associations for 33 phenotypes from the GRASP GWAS catalog (Supplementary File 2.5) (Leslie, O'Donnell, & Johnson, 2014). We found that peQTNs were generally more enriched among GWAS hits than lead variants, especially for immune-related traits (Figure 2.3D). We hypothesized that this enrichment was driven by HLA genes, most of which have eQTLs, and were concerned that the extensive polymorphism in HLA genes might have resulted in misalignment of RNA-seq reads to the reference genome. We repeated the GWAS enrichment using peQTNs after removing HLA eQTLs and still found higher enrichments for most traits using peQTNs (Figure 2.3D). This demonstrates that peQTNs are enriched among variants that are associated with organismal traits and that iPSC eQTLs capture regulatory associations important for known diseases and complex phenotypes.

Chapter 2.3.5: Intergenic CNVs Affect Gene Expression

To examine the effect of CNVs on gene expression, we used our high-depth WGS to identify 15,271 autosomal biallelic CNVs that were within 1Mb of at least one TSS and included these CNVs when testing for eQTLs as described above. Overall, we found significant CNV-expression associations (CNV eQTLs) for 247 genes including 111 genes for which the CNV was the lead variant (52

deletions, 38 duplications, and 26 mixed CNVs) and 67 genes whose expression was associated with a CNV but not a SNV or indel. While eGenes with no significant CNV associations showed only a slight bias toward lower expression for the alternate allele (p=0.001, binomial, Figure 2.4A), eGenes with CNV lead variants were heavily skewed toward positive associations between copy number and expression (Figure 2.4B). Lead CNVs also had larger effect sizes than lead SNVs and indels (Figure 2.4C). We compared all eQTL CNVs to CNVs that were not associated with the expression of any genes and found that eQTL CNVs are longer (median 2,386 bp versus 528 bp), more likely to overlap genes ($p < 10^{-16}$, Fisher exact test), and closer to transcription start sites (median 3,547 bp versus 12,390 bp) even after removing CNVs that overlap TSSs (median 5,606 bp vs 13,705 bp) (Supplementary Figure 2.4). These data show that CNV eQTLs tend to be larger than a typical CNV and generally have strong effects on gene expression that are positively correlated with copy number.

To explore the mechanisms underlying the correlation between CNV copy number and eGene expression levels we stratified the eQTLs on whether or not a significant CNV overlapped the eGene. Of the 111 eGenes with lead CNVs, 60 (54%) overlapped an associated CNV while only 102 of the 247 (41%) genes with at least one significant CNV association overlapped an associated CNV. As expected based on previous studies (Handsaker et al., 2015; Schlattl, Anders, Waszak, Huber, & Korbel, 2011; Sudmant et al., 2015), the expression levels of eGenes that overlapped CNV eQTLs were positively associated with copy

number (data not shown) most likely due to altered gene dosage. Only considering the 51 intergenic CNV lead variants, we still observed a bias toward positive associations between gene expression and copy number suggesting that these CNV eQTLs may act by altering the copy number of regulatory regions (Figure 2.4D, Supplementary Figure 2.5). We calculated the enrichment of DHSs and several histone modifications from Roadmap iPSC and ESC lines in these intergenic CNV eQTLs and found that the intergenic CNV eQTLs are enriched for marks of active regulatory regions but not repressive marks or marks of active transcription (Figure 2.4E). Repeating the same analysis for CNV eQTLs that do overlap their associated eGenes revealed that they are also enriched for marks of active regulatory regions but are most enriched for H3K36me3, a mark of transcribed regions, consistent with these CNVs overlapping genic regions (Figure 2.4F). These data suggest that intergenic CNV eQTLs can affect gene expression levels by altering the dosage of intergenic regulatory regions.

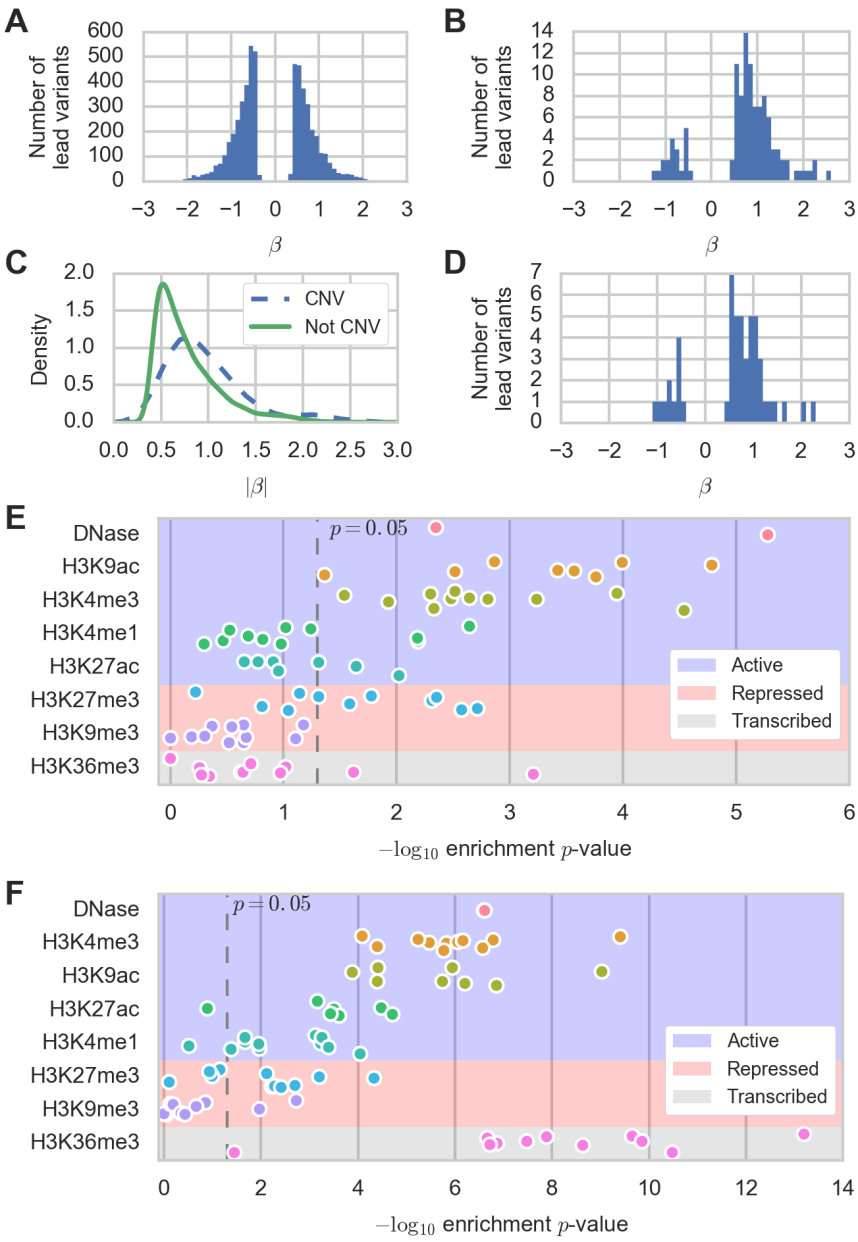


Figure 2.4: CNV eQTL Effect Sizes and Functional Annotation. (A-B) Distribution of effect sizes for (A) lead SNVs/indels for eGenes with no significant CNV associations and (B) lead CNVs. (C) Density plot of absolute effect size for effect sizes from (A) and (B). (D) Effect sizes for lead CNVs for eGenes where no significant CNV overlaps the eGene. (E-F) Enrichment *p* values (Fisher exact test) of Roadmap stem cell DHS and histone modification ChIP-seq peaks in lead CNVs for eGenes where (E) no significant CNV overlaps the eGene or (F) a significant CNV overlaps the eGene versus CNVs that were not eQTLs for any gene. Different points for each mark represent different Roadmap stem cell lines.

It was recently reported that multiallelic CNVs (mCNVs) are an important class of CNVs that can affect gene expression (Handsaker et al., 2015). Since

EMMAX is limited to testing for associations with biallelic variants and cannot test for associations with multiallelic loci, we identified mCNV eQTLs by regressing gene expression estimates against genotype using a linear model for the 131 unrelated individuals. After filtering (Methods), we identified 152 mCNVs segregating in the 131 unrelated individuals that were within 1 Mb of one or more genes and found mCNV eQTLs for 89 genes of which 33 overlapped an associated mCNV and 56 did not. The effect sizes for mCNV eQTLs were again skewed toward positive associations between gene expression and copy number for both mCNV eQTLs that overlapped genes and those that did not (Supplementary Figure 2.6) indicating that mCNVs may also affect gene expression by altering the dosage of regulatory regions. For example, we identified a 2kb mCNV on chromosome seven whose diploid copy number estimates ranged from one to eight and that was associated with the expression of seven nearby genes (Figure 2.5A). While this mCNV slightly overlaps one of the genes it is associated with, it also overlaps a DHS, CEBPB TF ChIP-seq peak, and predicted enhancer in the H1 hESC line suggesting that the CNV alters gene expression in the region by changing the copy number of this regulatory region (Figure 2.5B). Although most of the intergenic mCNV eQTLs were associated with the expression of only one or two genes, the bias toward positive associations between copy number and gene expression and the scaling of gene expression with the dosage of intergenic regions indicates that intergenic CNVs can cause eQTLs.

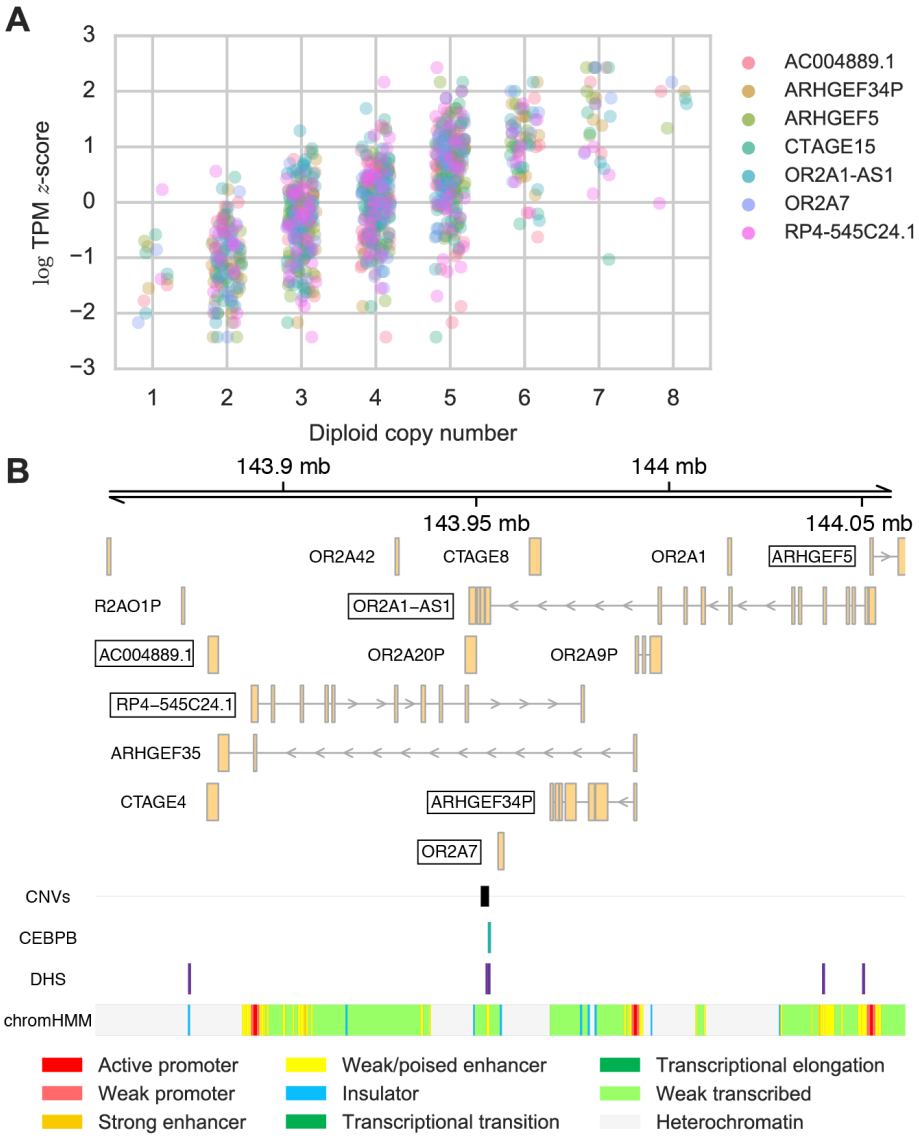


Figure 2.5: mCNV eQTL Example. (A) Gene expression estimates for seven genes associated with a single mCNV in 131 unrelated donors. Jitter was added to the diploid copy number estimates (x-axis) to aid in visualization. (B) Genomic location of mCNV on chromosome seven along with six of seven associated genes (indicated by boxes). The mCNV overlaps a CEBPB ChIP-seq peak, DHS, and predicted enhancer from the H1 hESC line.

Chapter 2.3.6: Effect of Rare Variants on Gene Expression

Rare variants are another class of variants whose effect on gene expression has been difficult to investigate because accurate identification of rare

variants within an individual requires high-depth WGS, and large sample sizes are needed to achieve sufficient statistical power to detect rare variant associations. While we expect to observe many rare variants across our 215 subjects, most of these will not fall in genes or regulatory regions and are not likely to affect gene expression. Therefore to investigate the effects of noncoding rare variants on gene expression, we decided to focus on rare variants located in the promoters of expressed genes. We identified 65,552 SNVs that (1) were located in the promoters of 18,556 robustly expressed genes (including genes on the sex chromosomes), (2) overlapped a DHS from at least one of the four Roadmap stem cell lines (Figure 2.2A), (3) had only one minor allele observed among the 131 unrelated subjects, and (4) were either not observed in 1000 Genomes or whose minor allele frequency was less than 0.5% in all 1000 Genomes populations (Genomes Project et al., 2015). We refer to these 65,552 SNVs as rare promoter DHS SNVs (rpdSNVs). In total, 14,599 of 18,556 robustly expressed genes had an rpdSNV in at least one of the 131 unrelated subjects.

To determine the effect of rpdSNVs on gene expression, we stratified the 18,556x131=2,430,836 expression estimates for the 18,556 robustly expressed genes in the 131 unrelated subjects into two groups based on the presence or absence of an rpdSNV in the gene's promoter in a given subject and obtained 69,041 expression estimates from genes with an rpdSNV and 2,361,795 estimates for genes without an rpdSNV. We restricted this analysis to the 131 unrelated subjects to avoid confounding due to relatedness and used the PEER

residual gene expression estimates transformed into *z*-scores to compare the two groups. We compared the distribution of these 69,041 and 2,361,795 expression values and found that expression estimates for samples with rpdSNVs were slightly lower than estimates for samples without rpdSNVs indicating that the presence of an rpdSNV has a small but significant effect on gene expression (Mann Whitney U, $p = 0.0026$, Figure 2.6A). Additionally, genes were more likely to have significant ASE in samples with an rpdSNV versus samples without an rpdSNV (OR = 1.09, $p = 0.015$, Fisher exact test) consistent with rare variants regulating gene expression in *cis*. It was reported previously that evolutionary constraint and functional annotations can help predict which rare variants may affect gene expression (X. Li et al., 2014) so we filtered the rpdSNVs according to phyloP conservation and CADD scores (Kircher et al., 2014; Pollard, Hubisz, Rosenbloom, & Siepel, 2010). We found that the bias toward lower expression estimates was stronger for genes with an rpdSNV with a CADD Phred score greater than 20 ($p < 10^{-4}$, Mann Whitney U) or a phyloP score greater than 3 ($p < 10^{-5}$, Mann Whitney U, Figure 2.6B). We also observed higher rates of ASE among genes with rpdSNVs with a CADD Phred score greater than 20 (OR = 1.47, $p = 0.011$, Fisher exact test) or a phyloP score greater than 3 (OR = 1.49, $p = 0.004$, Fisher exact test) compared to genes that did not have an rpdSNV. These results show that rpdSNVs that affect gene expression generally cause a decrease in expression and that rpdSNVs are more likely to affect gene expression if they are in conserved sequences or have higher CADD scores.

We next asked whether rare CNVs that overlap genes may affect gene expression by altering the dosage or structure of the overlapped gene. We defined rare genic CNVs as CNVs that overlapped introns and/or exons of genes and were observed in only one of the 131 unrelated subjects in our study. In total, we identified 431 rare genic duplications and 2,157 rare genic deletions. We stratified expression estimates into three groups based on the presence or absence of either a rare genic duplication or deletion for a given gene and subject. We found that the 431 rare genic duplications had a much stronger effect on gene expression than rpdSNVs and generally caused increased gene expression (Figure 2.6C). This effect was stronger if we restricted to the 226 rare duplications that were predicted to overlap exons as opposed to the larger set of deletions which includes some deletions that are only intronic. We did observe some instances where rare duplications appeared to decrease gene expression which seemed dependent on whether the CNV duplicated the entire gene or just part of the gene which would likely disrupt the gene's structure (data not shown). As observed for rpdSNVs, genes were much more likely to have significant ASE in subjects with rare genic duplications (OR = 5.55, $p < 10^{-9}$, Fisher exact test) with nearly 16% of such genes demonstrating significant ASE. The presence of higher rates of ASE among genes with rare duplications indicates that the altered expression of these genes is likely caused by these duplications. As opposed to duplications, we found that the 2,157 rare genic deletions generally caused lower expression (Figure 2.6D). This effect was much stronger for the 511 rare

deletions that were predicted to overlap exonic parts of the genes (Figure 2.6D). 9.1% of genes with rare exonic deletions in a given sample had significant ASE compared to 3.2% of the genes that did not have rare exonic deletions (OR = 3, $p < 10^{-3}$, Fisher exact test). These results indicate that rare genic CNVs are more likely to affect gene expression than rpdSNVs and that the effect of rare CNVs is dependent on the location of the CNV relative to coding regions of the gene.

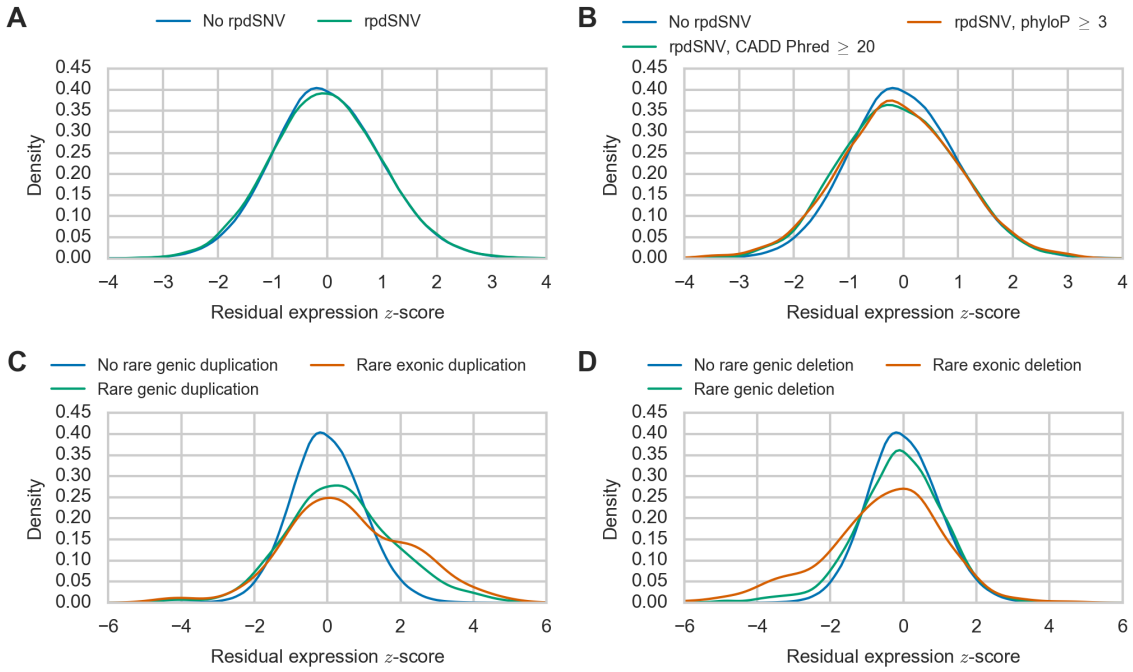


Figure 2.6: Effect of Rare Variants on Gene Expression. Distribution of gene expression estimates for genes (A) with (green) or without (blue) a rare promoter DHS SNV (rpdSNV) and (B) without an rpdSNV (blue), with an rpdSNV with CADD Phred greater than 20 (green), or with an rpdSNV with a phyloP score greater than three (orange). Distribution of gene expression estimates for genes (C) without rare genic duplications (blue), with rare genic duplications (green), or with rare exonic duplications (orange) and (D) without rare genic deletions (blue), with rare genic deletions (green), or with rare exonic deletions (orange).

Chapter 2.3.7: X Reactivation Status Varies According to Gene Chromosomal Position

X inactivation (Lyon, 1961) has been studied in iPSCs derived from female donors to determine the behavior of the inactive X chromosome during reprogramming and passaging (Lessing et al., 2013; Pasque & Plath, 2015) but the heterogeneity of X chromosome reactivation across a large set of systematically reprogrammed lines is unknown. Since our iPSCs are clonally derived from single fibroblasts, female-derived iPSCs should have one inactive and one active X unless the inactive X has been reactivated during reprogramming or passaging. iPSCs with residual X inactivation should have a higher amount of ASE for genes on the X chromosome relative to autosomal genes (i.e. ASE for genes on the X chromosome is a proxy for X inactivation). We calculated the percentage of X chromosome and autosomal genes with significant ASE per sample for 144 RNA-seq samples from the 116 iPSC lines derived from female donors (predominantly assayed at passage 12) and found that the X chromosome is highly enriched for ASE relative to autosomes with an average of 44% of X chromosome genes displaying significant ASE per sample compared to only 3% of autosomal genes per sample. We identified 120 robustly expressed X chromosome genes and stratified each gene's expression estimates into two groups based on whether or not the gene had significant ASE in a given sample. We calculated the average expression of each gene in the two groups and observed that 78% of the genes had lower average expression in the group

of samples with significant ASE consistent with allelic silencing of these genes by X inactivation (Figure 2.7A). These results indicate that X inactivation persists at some level for most iPSCs derived from female subjects and affects the gene expression of X chromosome genes.

To examine the heterogeneity of X reactivation across the iPSCs, we defined the strength of ASE for a given gene as the percentage of RNA transcripts estimated to originate from the parental haplotype with higher expression, referred to as the major haplotype frequency (MHF) (Mayba et al., 2014). The distribution of MHFs for X chromosome genes was bimodal with some genes showing relatively balanced expression (MHF near 0.5) and other genes displaying nearly mono-allelic expression (MHF near 1.0) consistent with some X chromosome genes remaining silenced following reprogramming (Figure 2.7B). In contrast, the MHFs for most autosomal genes was near 0.5 with few genes showing evidence for strong allelic bias (Figure 2.7C). Stratifying the MHFs by sample showed that there is considerable variation between samples with some iPSC displaying low levels of X reactivation and others displaying high levels of reactivation (Figure 2.7D). The percentage of X chromosome genes with significant ASE per sample (a proxy for the overall amount of reactivation per sample) is correlated with *XIST* ($r=0.72$, $p<10^{-24}$, Spearman) and *TSIX* gene expression ($r=0.51$ and $p<10^{-11}$, Spearman) showing that *XIST* and *TSIX* are down-regulated as the inactive X is reactivated. *XIST* ($r=-0.18$, $p=0.029$, Spearman) and *TSIX* ($r=-0.17$, $p=0.044$, Spearman) expression are also

negatively correlated with passage although passage is not correlated with the percentage of X genes with significant ASE (r=-0.07, p=0.43). However, most of our iPSC lines were at passage 12 so it is possible that we are not powered to find this latter association. These results suggest that *XIST* and *TSIX* are downregulated as the inactive X is reactivated during early passages.

While we observed that the overall amount of X reactivation differs between lines (Figure 2.7C), we also asked whether the X reactivation status of genes was correlated with respect to their location on the X chromosome. We plotted the major haplotype frequency estimates for each gene in each sample versus the position of the gene on the X chromosome and observed that clusters of nearby genes tended to show similar levels of reactivation even in different lines (Figure 2.7E,F). Our data suggest that while the overall amount of reactivation differs between lines, reactivation follows the same physical pattern in different lines with some clusters of nearby genes consistently becoming reactivated faster than others.

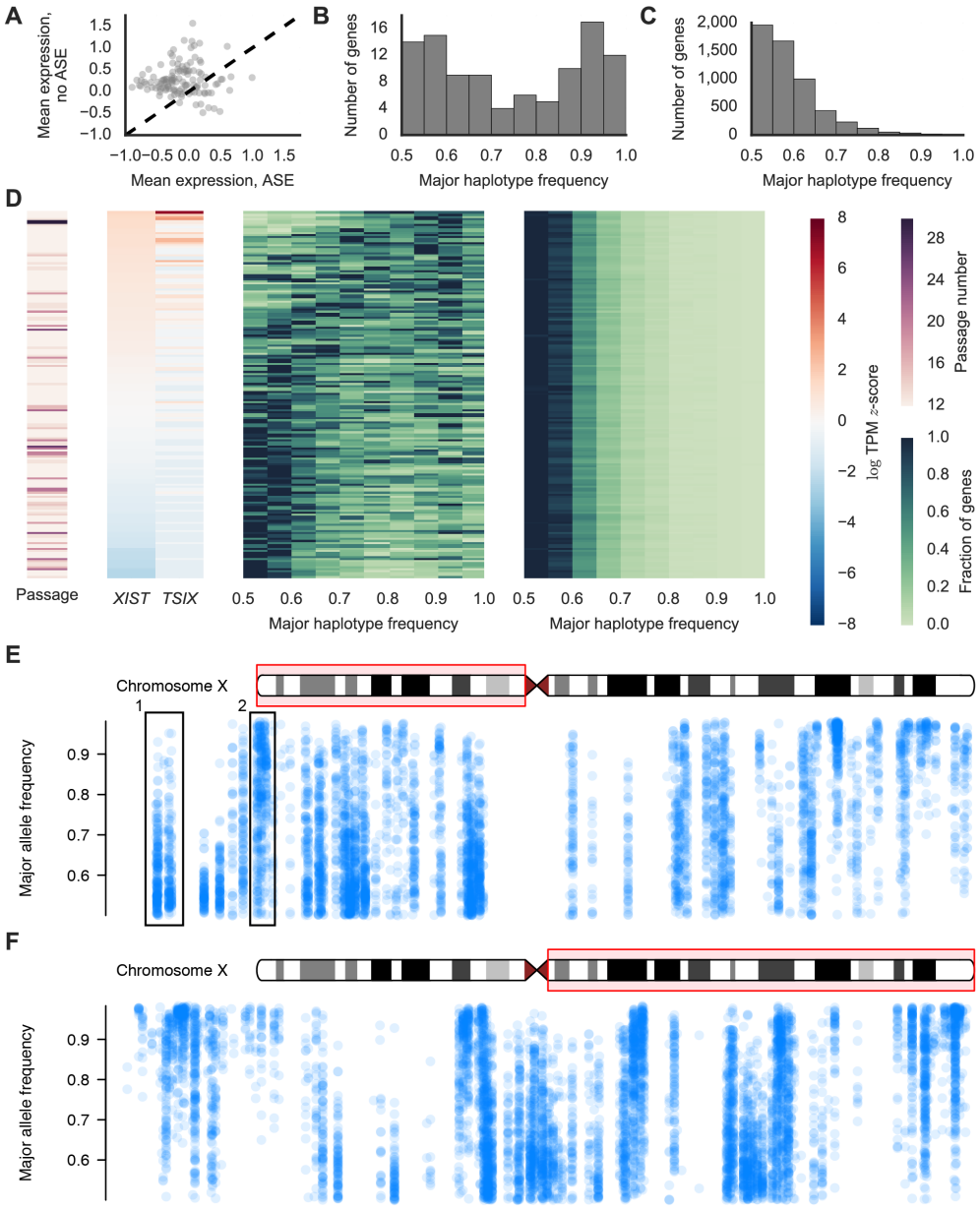


Figure 2.7: Heterogeneity of X Reactivation Following Reprogramming. (A) Average expression for 120 X chromosome genes in samples with significant ASE versus samples without significant ASE. (B and C) Distribution of estimated major haplotype frequencies (MHFs) for (B) X chromosome and (C) autosomal genes for one representative RNA-seq sample. MHF is the percentage of transcripts estimated to come from the haplotype that is more expressed. (D) From left to right, the heatmaps show passage; *XIST and TSIX* expression; X chromosome MHF distribution; and autosomal MHF distribution for each RNA-seq sample from female-derived iPSCs. Each row corresponds to one one sample. Samples were sorted by *XIST* expression before plotting. (E and F) Estimated MHFs across the (E) p and (F) q arms of the X chromosome. Each point represents an estimate of the MHF for a gene/sample pair. Box 1 shows a largely reactivated cluster of genes with most MHFs near 50% while box 2 shows a cluster of genes that have not been reactivated with MHFs closer to 100%.

Chapter 2.4: Discussion

We present here an analysis of the genetic regulation of gene expression in a set of 215 systematically reprogrammed human iPSCs. We used WGS variant calls and RNA-seq expression estimates to map eQTLs in iPSCs and found eQTLs for 5,619 genes including markers and regulators of pluripotency. We found that relative to the somatic tissues profiled in GTEx, iPSCs are well-powered for identifying eQTLs and have a distinct gene regulatory landscape. These results show that iPSCs are a suitable system for genetic association analyses and provide important insights into the effect of genetic background on gene expression in stem cells.

We demonstrated that iPSC eQTLs are enriched in regulatory regions of both iPSCs and hESCs from the ENCODE and Roadmap Epigenomics projects supporting previous findings that iPSCs and hESCs have similar gene expression and epigenetic profiles (Choi et al., 2015; Papp & Plath, 2013; Rouhani et al., 2014). Since altered TF binding has been proposed to be one of the primary causes of eQTLs (Pai et al., 2015), we intersected our eQTL variants with TF ChIP-seq peaks from the H1 hESC line for 34 different TFs and identified putative eQTNs that both overlapped a TF peak and disrupted a motif associated with the TF. We identified 3,058 peQTNs and found that they were highly enriched for evidence of disrupted TF binding *in vivo*. In total, these 3,058 variants corresponded to 1,475 of 5,619 eGenes demonstrating that it may be necessary to profile more TFs in order to dissect the majority of eQTLs. Interestingly, only

19% of the 1,475 eGenes had a lead variant that overlapped a TF peak and disrupted a motif suggesting that for eQTLs lead variants may not often be causal variants. The high frequency of peQTNs that are not lead variants may be due to the presence of multiple independent eQTLs for most genes (many of which we may not be able to detect at this sample size) that affect which variants are most significant. We also found that iPSC eQTLs are enriched for GWAS hits from several human phenotypes. Due to the shared genetic architecture of gene regulation across tissues (G. T. Consortium, 2015; Flutre, Wen, Pritchard, & Stephens, 2013), these associations do not necessarily mean that stem cells are especially important for these particular phenotypes but do show that examining the regulatory landscape of iPSCs is relevant to human traits and diseases.

We also explored the contribution of CNVs and rare variants to the regulation of gene expression. We showed that both common and rare CNVs can affect gene expression either by altering the copy number/structure of genes or intergenic regulatory regions. We presented an example of a mCNV causing gene expression changes by altering the dosage of a regulatory sequence though further experiments are needed to determine what fraction of eQTL CNVs act through this mechanism. We also observed that rare SNVs in promoters can cause decreased gene expression and that this effect is stronger for conserved variants or those that overlap functional annotations. It remains to be seen whether this signal is driven by a small fraction of functional rare variants with larger effect sizes or whether many rare variants have small effects on gene

expression. While our genotypes are derived from blood and fibroblasts, it is notable that somatic variants that arise during the reprogramming process may have similar effects as inherited rare variants. We anticipate future studies may benefit from genotyping iPSCs and germ cells to profile somatic variants genome-wide and incorporate them into association analyses.

We investigated the heterogeneity of X chromosome reactivation following reprogramming from female donors and found that most samples retain some amount of silencing on the X chromosome although the amount differs from sample to sample. We found that all lines share similar physical reactivation patterns across the X chromosome, with clusters of genes in some areas escaping silencing more quickly than clusters of genes in other areas. This may be related to the fact that XIST physically coats the X chromosome. The differing rates of reactivation between samples and across the X chromosome shown here will need to be accounted for when investigating X-linked molecular quantitative traits.

iPSCs are a promising system for mapping expression and other molecular trait QTLs for several reasons including their ability to self-renew and differentiate into other cell types (Pai et al., 2015). Genetic association analyses in iPSCs and differentiated cell types are not limited to gene expression or other molecular phenotypes like methylation levels, however, but can be extended to physiological phenotypes like electrophysiological responses or cellular phenotypes like cell survival after drug treatment (Avior et al., 2016). Merging

“disease in a dish” modeling approaches with large-scale genetic association analyses like the one presented here will useful for dissecting complex diseases and drug-genotype interactions and will likely become an important strategy for exploring the genetic and molecular causes of disease.

Chapter 2.5: Experimental Procedures

Complete information on the experimental procedures for this work can be in the Supplemental Experimental Procedures.

Chapter 2.5.1: Sample collection, reprogramming, and cell culture

Samples were collected, reprogrammed, and cultured as described in (Frazer, 2016). The UCSD IRB approved the study and all subjects gave informed consent (Project #110776ZF).

Chapter 2.5.2: Whole Genome and RNA Sequencing

Whole genome and RNA sequencing and data processing were performed using standard protocols and are described in detail in the Supplemental Experimental Procedures.

Chapter 2.5.3: Data Analysis

We mapped eQTLs using the permutation approach from (G. T. Consortium, 2015) except that we used EMMAX (Kang et al., 2010) to calculate association *p*-values that account for the relatedness between donors. We

calculated per gene *p*-values using the permutations and corrected these empirical *p*-values using the Storey method (Storey & Tibshirani, 2003). Allelic mapping bias was accounted for using WASP (van de Geijn, McVicker, Gilad, & Pritchard, 2015) and allele specific expression was identified using MBASED (Mayba et al., 2014). Complete details are described in the Supplemental Experimental Procedures.

Chapter 2.5.4: Data and Code Availability

Sequencing data are deposited in dbGaP (phs000924). Code for this project is available on Github at https://github.com/frazer-lab/cardips-ipsc-eqtl.

Chapter 2.6: Author Contributions

CD designed and performed computational analyses. CD and KAF designed experiments. HL performed WGS alignment and GATK variant calling. DJ performed GenomeSTRiP and LUMPY CNV calling. ADC, AA, PB performed iPSC cell culture and collected RNA for sequencing. WB and ES generated whole genome sequencing data. CD, HM, and JR processed data and organized data in a database. CD and KAF wrote the manuscript. All authors edited the manuscript.

Chapter 2.7: Acknowledgments

We would like to thank Naoki Nariai, Erin Smith, and Terry Solomon for useful input on this project. This work was supported in part by a CIRM grant GC1R-06673 (to KAF) and NIH grants HG008118-01 (to KAF) and HL107442-05 (to KAF). We would like to thank the UC San Diego Institute for Genomic Medicine (IGM) core for sequencing services with support from NIH grant P30CA023100. CD is supported in part by the University of California, San Diego, Genetics Training Program through an institutional training grant from the National Institute of General Medical Sciences (T32GM008666) and the California Institute for Regenerative Medicine (CIRM) Interdisciplinary Stem Cell Training Program at UCSD II (TG2-01154).

Chapter 2, in full, has been submitted for publication of the material as it may appear in *Cell Stem Cell*, 2016. Christopher DeBoever, He Li, David Jakubosky, Angelo Arias, Joaquin Reya, William Biggs, Efren Sandoval, Hiroko Matsui, Paola Benaglio, Agnieszka D'Antonio-Chronowska, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

Chapter 2.8: Supplemental Experimental Procedures

Chapter 2.8.1: Sample collection and reprogramming

Samples were collected, reprogrammed, and cultured as described in (Frazer, 2016). The UCSD IRB approved the study and all subjects gave informed consent (Project #110776ZF).

Chapter 2.8.2: RNA sequencing

Library preparation and sequencing

Total RNA was extracted from 222 iPSC lines using AllPrep RNasy Blood & Tissue Kit (Qiagen) following the manufacturer's protocol. RNA quality was assessed based on RNA integrity number (RIN) using an Agilent Bioanalyzer. Libraries were prepared using the Illumina TruSeq stranded mRNA kits and sequenced using an Illumina HiSeq2500 (~11 samples per lane). Samples were sequenced to an average of ~22 million read pairs. Biological replicates were sequenced for some lines.

Alignment and quality control

2x150 bp RNA-seq reads were aligned with STAR (2.5.0a) to the hg19 reference (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit) using Gencode v19 splice junctions with default alignment parameters except --outFilterMultimapNmax 20, --outFilterMismatchNmax 999, --alignIntronMin 20, --alignIntronMax 1000000, --alignMatesGapMax 1000000 (Dobin et al., 2013; Harrow et al., 2012). Bam files were coordinate sorted using Sambamba (0.5.9)

(Tarasov, Vilella, Cuppen, Nijman, & Prins, 2015) and duplicate reads were marked using biobambam2 (2.0.21) bammarkduplicates (Tischler & Leonard, 2014).

We repeated library preparation and sequencing for samples that were outliers for percent uniquely mapped reads as reported by STAR or percent duplicate reads or 5'/3' bias as estimated by Picard Tools. Seven of the 222 samples that were outliers after the second sequencing run were not used resulting in RNA-seq for 215 of the 222 lines. The minimum uniquely mapped read percentage was 86% and the median was 91%.

Gene expression

We estimated transcript and gene expression using the STAR transcriptome bam file and RSEM (1.2.20) rsem-calculate-expression (--seed 3272015 --estimate-rspd --forward-prob 0) (B. Li & Dewey, 2011).

Allele specific expression

Uniquely mapped reads that were not marked as duplicates were tested for mapping bias using the WASP mapping pipeline (van de Geijn et al., 2015). Reads that mapped uniquely to the same location after swapping in alternate alleles were used to calculate the coverage of heterozygous variants overlapping Gencode v19 exons for all exonic regions unique to one gene using the ASEReadCounter from GATK (3.4-46) (Van der Auwera et al., 2013). MBASED was used to estimate per-gene and per-heterozygous variant allele specific

expression (ASE) p-values (Mayba et al., 2014). Heterozygous variants that met the following criteria were used as input for MBASED: (1) coverage greater than or equal to 8, (2) reference allele frequency between 2-98%, (3) located in unique mappability regions according to wgEncodeCrgMapabilityAlign100mer track, (4) not located within 10 bp of another variant in a particular subject (heterozygous or homozygous alternative). Additionally, for heterozygous variants within 300 bp of each other, only one variant was used to avoid double counting variant coverage from the same read pair. These filters are based on the GTEx and MBASED ASE pipelines (G. T. Consortium, 2015; Lappalainen et al., 2013; Mayba et al., 2014). A gene was considered significant for ASE if the MBASED "p_val_ase" was less than or equal to 0.005 (G. T. Consortium, 2015).

Chapter 2.8.3: DNA sequencing

Library preparation and sequencing

Genomic DNA was isolated from blood (DNEasy Blood & Tissue Kit), or in 19 cases directly from the fibroblast. The samples were normalized to 1 ug and submitted for whole genome sequencing. DNA was isolated (DNeasy kit, Qiagen), quantified, normalized and sheared with a Covaris LE220 instrument. DNA libraries were prepared (TruSeq Nano DNA HT kit, Illumina), characterized in regards to size (LabChip DX Touch, Perkin Elmer) and concentration (Quant-iT, Life Technologies), normalized to 2-3.5nM, combined into 6-sample pools, clustered and sequenced on the HiSeqX (150 base paired-end). In total, germline

whole genome sequencing (WGS) was performed for 274 subjects though only 222 were reprogrammed into iPSCs.

Alignment and quality control

We esitimated the quality of fastq files using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were aligned against human genome b37 with decoy sequences (Genomes Project et al., 2015) using BWA-mem and default parameters (H. Li & Durbin, 2009). The resulting bam files were sorted using Sambamba (Tarasov et al., 2015) and duplicate reads were marked using biobambam2 (Tischler & Leonard, 2014).

SNV/indel calling

The bam files were split into individual chromosomes to maximize the efficiency of the variant calling process on our cluster. We applied the GATK (McKenna et al., 2010) best-practices pipeline for variant calling that includes indel-realignment, base-recalibration, genotyping using HaplotypeCaller, and finally joint genotyping using GenotypeGVCFs (DePristo et al., 2011; Van der Auwera et al., 2013). We performed quality control for the genotypes of single nucleotide variants and indels using GATK's Variant Quality Score Recalibration (VQSR) (Van der Auwera et al., 2013). We performed variant calling for sex chromosomes in males and females separately and resolved the pseudoautosomal regions of the sex chromosomes independently (considered as diploid in both males and females).

CNV calling

CNVs were called using two algorithmic approaches with the goal of finding variants across a wide spectrum of sizes and making use of both read-pair and read depth information. We used the population level read-depth and split-read caller Genome STRiP (svtoolkit 2.00.1611) (Handsaker et al., 2015) to discover and genotype biallelic and multiallelic CNVs using whole genome sequencing data from 274 subjects. We supplemented this CNV call set using the split and discordant read-pair caller LUMPY (Layer et al., 2014) as implemented in the SpeedSeq software (version 0.1.0) (Chiang et al., 2015). Speedseq SV calling was done individually on each of the 274 samples, excluding areas identified by the LUMPY developers with very high read-depth in family CEPH 1463 (Chiang et al., 2015). Calls for each sample were then filtered further, removing calls overlapping regions with over 200 split or discordant reads in a given sample, and calls that overlapped centromeres, telomeres, or low complexity regions (H. Li, 2014). Calls were then merged using svtools lsort and lmerge (https://github.com/hall-lab/svtools), before running the SVtyper Bayesian genotyping algorithm on these positions in each sample. Following genotyping, sites that were predicted as reference in all samples were removed as well as sites supported by less than 10 reads.

Chapter 2.8.4: eQTL analysis

We first selected one iPSC RNA-seq sample per subject for which WGS variant calls were also available. We constructed an empirical kinship matrix for all subjects with WGS variant calls by intersecting biallelic SNVs with 1000 Genomes phase 3 variants and LD pruning the resulting variants using plink 1.90b3x (--biallelic-only --indep-pairwise 50 5 0.2) for unrelated EUR 1000 Genomes subjects (Chang et al., 2015; Genomes Project et al., 2015). We used the remaining LD-pruned variants to construct the kinship matrix using EPACTS 3.2.6 (epacts make-kin --min-maf 0.01 --min-callrate 0.95) keeping variants whose frequency was above 1% in our cohort and that were called in at least 95% of our samples.

We filtered RSEM gene TPM values by removing any genes whose expression was not greater than 2 TPM in 10 or more samples. We then transformed the expression values for each of the 17,819 autosomal genes passing these filters to match a standard normal distribution and ran PEER for 15 factors (Stegle et al., 2010). We transformed the PEER residuals to match a standard normal distribution to minimize the effect of outliers on the eQTL analysis (G. T. Consortium, 2015).

We filtered WGS variant calls by removing variants whose call rate was less than 95% or with Hardy-Weinberg $p < 0.000001$ for 104 unrelated European samples from our cohort.

We filtered GenomeSTRiP CNV calls to keep those that were observed in three diploid copy number states for at least 95% of our 215 eQTL samples. If a CNV was observed in three diploid copy number states for 95% of samples but also had other copy number states, we set those genotypes to missing. All CNVs were encoded as 0/0, 0/1, and 1/1 for increasing diploid copy number for the purposes of association. We filtered LUMPY CNV calls to keep calls with minor allele frequency greater than 1% in the 215 eQTL samples.

We tested autosomal genes for eQTLs using EMMAX (assoc --maxMAF 1 --maxMAC 1000000000 --minRSQ 0 --minCallRate 0.5 --minMAC 3) using the standard normal transformed PEER residuals and the empirical kinship matrix described above (Kang et al., 2010). We provided the sex of each subject as a covariate for EMMAX. For each gene, we tested variants within 1Mb of any TSS for that gene from the Gencode v19 gene annotation and whose call rate was greater than 95%. We identified genes with significant eQTLs (eGenes) using the permutation approach from (G. T. Consortium, 2015). For each gene, we performed 1,000-10,000 permutations of the expression values and recorded the minimum p-value from EMMAX. We stopped performing permutations when we obtained 15 minimum p-values less than the minimum p-value observed for the real data or when we reached 10,000 permutations. We calculated an empirical p-value for each gene as the fraction of permutations with minimum p-values less than the observed minimum p-value. We corrected these empirical p-values using the Storey method (Storey & Tibshirani, 2003).

We identified additional independent eQTLs for eGenes by providing the lead variant as a covariate for EMMAX and performing the same permutation procedure. We corrected these permutation p-values using the Storey method.

Comparison to GTEx eQTLs

We compared our eQTLs to those reported in GTEx v6 (phs000424.v6.p1). When plotting the number of and percent unique eGenes versus the number of samples for Figure 2.1F-G, we omitted the GTEx testis results because they were highly different than all other GTEx tissues.

Chapter 2.8.5: GO comparison

Genes in the "stem cell population maintenance" (GO:0019827) category were downloaded on March 17, 2016 from the AmiGO database (Ashburner et al., 2000; Carbon et al., 2009; Gene Ontology, 2015) and intersected with the 5,619 eGenes.

Chapter 2.8.6: Functional Annotation

Roadmap Epigenomics DNase hypersensitivity site (DHS) enrichments

We downloaded DHS data for 53 Roadmap Epigenomics cell types from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/. We then defined one lead variant per eGene by randomly break ties and kept only SNVs and indels, resulting in 5,420 lead SNVs/indels. We intersected these

5,420 variants with exonic regions from Gencode v19 and removed any SNVs intersecting an exon, leaving 4,491 noncoding SNVs/indels remaining. We calculated how many SNV/indel bases did and did not overlap a DHS for each DHS experiment. We then created 5kb windows centered on each SNV/indel and calculated the number of base pairs that did and did not overlap a DHS in the window (excluding the lead variant). We used these counts to perform a Fisher exact test (fisher_exact, scipy) to determine an odds ratio and enrichment *p*-value for each DHS experiment as in (G. T. Consortium, 2015).

ENCODE DHS enrichments

We searched for all ENCODE DHS experiments with narrowPeak files for the hg19 assembly using the ENCODE web API (encodeproject.org) and pyencodetools (https://github.com/cdeboever3/pyencodetools). We used the most recent narrowPeak file for each experiment or chose randomly when the date was malformed. We used the same set of noncoding lead SNVs/indels described above and calculated odds ratios and enrichment *p*-values as described above.

ENCODE transcription factor (TF) enrichments

We identified ENCODE TF ChIP-seq experiments for the H1 hESC cell line using pyencodetools as described above. We used the same set of noncoding lead SNVs/indels described above and calculated odds ratios and enrichment *p*-values as described above.

Chapter 2.8.7: Identification of putative eQTNs

To identify putative expression quantitative trait nucleotides (peQTNS) for the 5,619 eGenes, we considered all significant associations with SNVs and indels but filtered out CNV associations because their mechanism of action is likely different than disrupting a TF binding site. We also removed any eGenes that overlapped a significant CNV or had an eQTL variant that was predicted to cause nonsense mediated decay according to SnpEff (Cingolani et al., 2012). We then overlapped the remaining 186,656 variants with ENCODE TF ChIP-seq peaks (Supplementary File 2.4). For each variant that overlapped a peak, we calculated motif scores for motifs associated with the particular TF that the variant overlapped (http://compbio.mit.edu/encode-motifs/motifs.txt) (Kheradpour & Kellis, 2014). We calculated the motif scores using MOODS (Korhonen, Martinmaki, Pizzi, Rastas, & Ukkonen, 2009) for both the reference and alternate alleles. If the MOODS scores for the reference and alternate alleles differed by more than 2.5, we said that the variant disrupted TF binding (Supplementary File 2.4). When comparing to the data from (Maurano et al., 2015), we considered $q <$ 0.05 as evidence for significant TF allelic bias.

Chatper 2.8.7: GWAS enrichments

We downloaded the GRASP v2 database (Leslie et al., 2014). For each phenotype, we identified independent GWAS hits with *p*-values less than 10^{-5}. We identified independent SNPs by creating a graph whose nodes were

significant variants that shared an edge if the two variants were in LD > 0.8 (1000 Genomes phase 3 EUR). We then chose the variant with the smallest *p*-value per connected component. We created 50 random sets of null SNPs matched on minor allele frequency, number of SNPs in LD > 0.8, and distance to the nearest protein coding gene; these statistics were obtained from SNPsnap (EUR population) (Pers, Timshel, & Hirschhorn, 2015).

We then LD pruned eQTL lead SNVs and counted the number of independent GWAS SNPs that were in LD (LD > 0.8, 1000 Genomes phase 3 EUR) with an independent eQTL lead variant for both the real and null data. We summed the results for the 50 null sets and calculated enrichments using a Fisher exact test (fisher_exact, scipy). We LD pruned the peQTNs and calculated enrichments in the same way. We also calculated enrichments after removing any peQTNs associated with HLA genes (defined as any gene with HLA in the gene name).

Chapter 2.8.8: CNV eQTL Analysis

CNVs eQTLs

We included GenomeSTRiP and LUMPY CNVs when mapping eQTLs as described above. While GenomeSTRiP calls multiallelic CNVs, we only used CNVs with at most three biallelic copy number states for 95% of the 215 subjects when identifying eQTLs. Mixed CNVs are defined by GenomeSTRiP as CNVs with diploid copy numbers consistent with both deletions and duplications relative

to the reference. We encoded the three copy number states as 0/0, 0/1, 1/1 in order of increasing copy number for use with EMMAX. For LUMPY, we used the genotypes from SVtyper.

CNVs overlapping genes

We took a conservative approach for identifying which eGenes overlapped CNVs. We observed that in some instances, GenomeSTRiP called one CNV as two different CNVs. This was apparent because the two CNVs were in perfect LD and next to each other on the genome. We therefore merged nearby CNVs with highly correlated copy number estimates. For a given eGene, we also merged all CNVs associated with that eGene for the purpose of determining whether the eGene overlapped a significant CNV. Thus if there were two CNVs on either side of a gene and they were both associated with the expression of the gene, we would merge these two CNVs and consider that eGene to overlap a significant CNV.

CNV functional annotation

We overlapped the eQTL mCNVs with functional annotations from Roadmap Epigenomics (http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/) for stem cell lines ES-I3, ES-UCSF4, ES-WA7, HUES48, HUES6, HUES64, iPS, iPS-15b, iPS-18, iPS-20b. Some annotations were not available for some lines. We calculated enrichments using scipy's fisher_exact.

Multiallelic CNV (mCNV) eQTLs

We identified mCNVs using the GenomeSTRiP CNV calls by first removing any CNVs that did not have more than three diploid copy number states among the 131 unrelated subjects or that were in the MHC region (chr6:29,600,000-33,100,000). We further filtered the mCNVs to only include mCNVs for which at least 6 subjects had diploid copy number states that differed from the three most prevalent diploid copy number states to avoid including CNVs that may have been classified as mCNVs due to erroneous copy number estimates for a small number of samples. We identified eQTLs by regressing PEER residual expression values for genes within 1Mb of an mCNV against the diploid copy number estimates for the 131 unrelateds for that mCNV. We included sex as a covariate. In total, there were 152 distinct mCNVs that we tested for eQTLs with 1,493 genes (2,952 total tests). We corrected these 2,952 test for multiple testing using the Benjamini Hochberg procedure. We determined whether an mCNV overlapped an eGene after merging the mCNVs as described above for CNVs.

Chapter 2.8.9: Rare Variant Analysis

Rare variant identification

We first intersected GATK SNVs with promoters from Gencode v19. Promoters were defined as 2kb upstream and 200 bp downstream of a TSS for all Gencode genes. We only used promoters from 18,556 genes (including sex

chromosome genes) with TPM > 2 in at least 10 of the 215 samples. We obtained DHSs for the H1, H9, iPS DF 6.9, and iPS DF 19.11 cell lines from Roadmap Epigenomics and merged them into one bed file. We the intersected the promoter variants with these merged DHSs. We next annotated each SNV with its minor allele frequency (MAF) from the 1000 Genomes phase 3. We kept variants whose MAF was less than 0.5% in all 1000 Genomes population groups and that only had one observed minor allele among the 131 unrelated individuals. We identified 65,552 rare promoter DHS SNVs (rpdSNVs) in total.

Effect of rare promoter DHS SNVs on gene expression

To determine the effect of rpdSNVs on gene expression, we focused on the expression the 18,556 genes in the 131 unrelated subjects to avoid confounding due to relatedness. We used the PEER residual gene expression estimates transformed into *z*-scores so that we could compare across genes. We stratified each of the 18,556x131=2,430,836 expression estimates into two groups based on whether a given gene had an rpdSNV in a given sample. In total, there were 69,041 estimates from genes/samples with an rpdSNV and 2,361,795 from genes/samples without an rpdSNV. We compared the distribution of these 69,041 and 2,361,795 expression values using a Mann Whitney U test to test whether the distributions differed. We also calculated whether a given gene/sample was more likely to have ASE if it contained an rpdSNV using a Fisher exact test.

We calculated CADD scores (Kircher et al., 2014) for all variants and used the CADD Phred scores to filter the 69,041 estimates from genes/samples with rpdSNVs to only include rpdSNVs with Cadd Phred greater than 20. We also filtered based on phyloP score (Pollard et al., 2010) greater than 3.

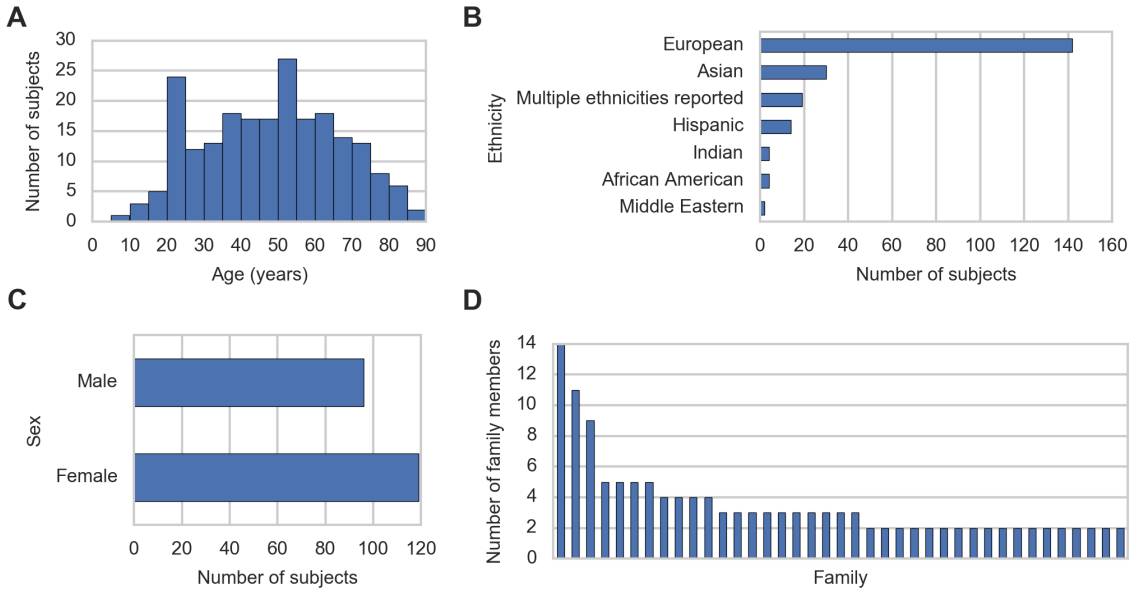
Effect of rare genic CNVs on gene expression

We identified rare CNVs that overlapped genes where a gene was defined as the entire region from its 5'-most TSS to its 3'-most UTR. A CNV was defined as rare if it was observed in only one of the 131 unrelateds. We also characterized whether the CNV overlapped any exonic part of the gene. We stratified the 2,430,836 estimates as described above based on the presence of a genic duplication or a genic deletion. We similarly compared the distributions of expression values using a Mann Whitney U and used a Fisher exact test to test for ASE enrichment.

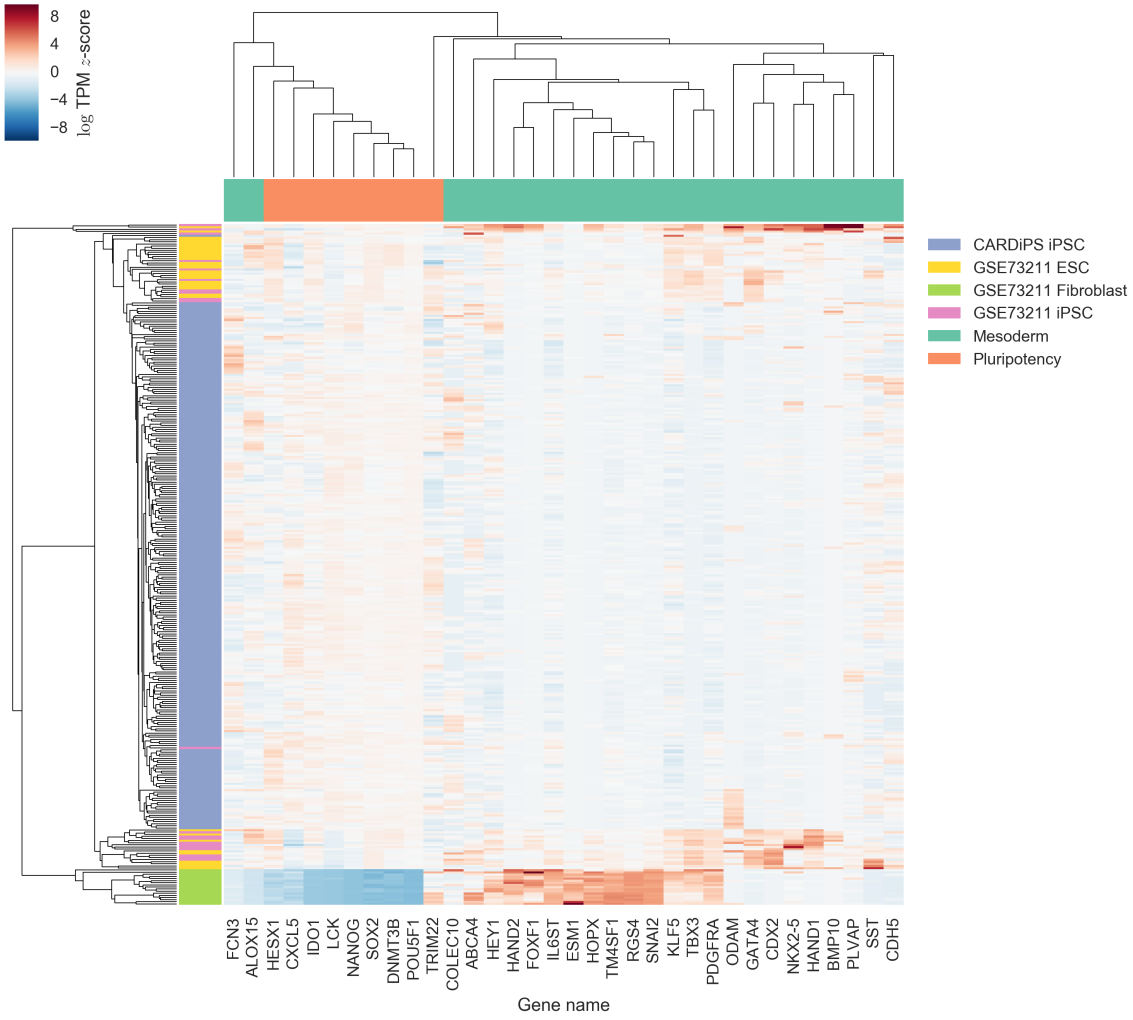
Chapter 2.8.10: X Reactivation

We used 144 separate RNA-seq experiments from 116 iPSC lines derived from female subjects (some lines had biological replicates). We restricted the analysis to lines with no evidence of reprogramming-associated CNVs on the X chromosome (Frazer, 2016). We used the ASE results from MBASED described above. The major haplotype frequency estimates were also produced by MBASED.

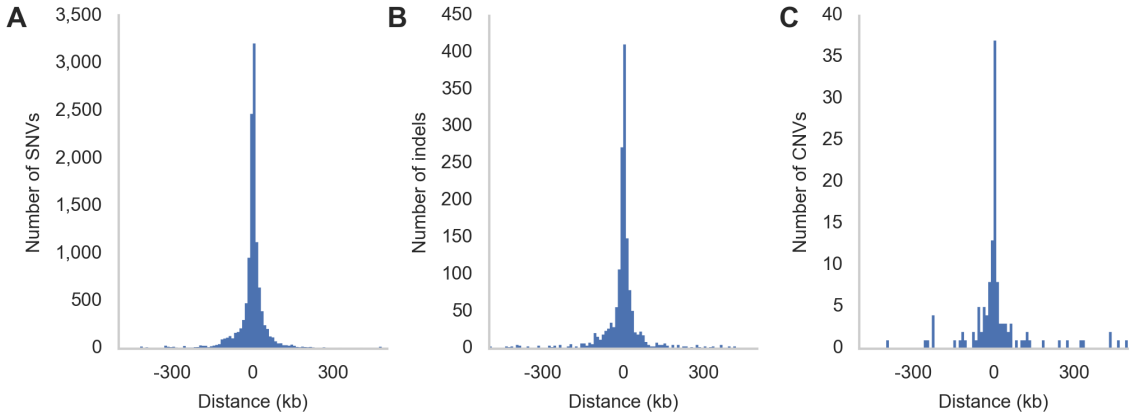
Chapter 2.9: Supplementary Figures



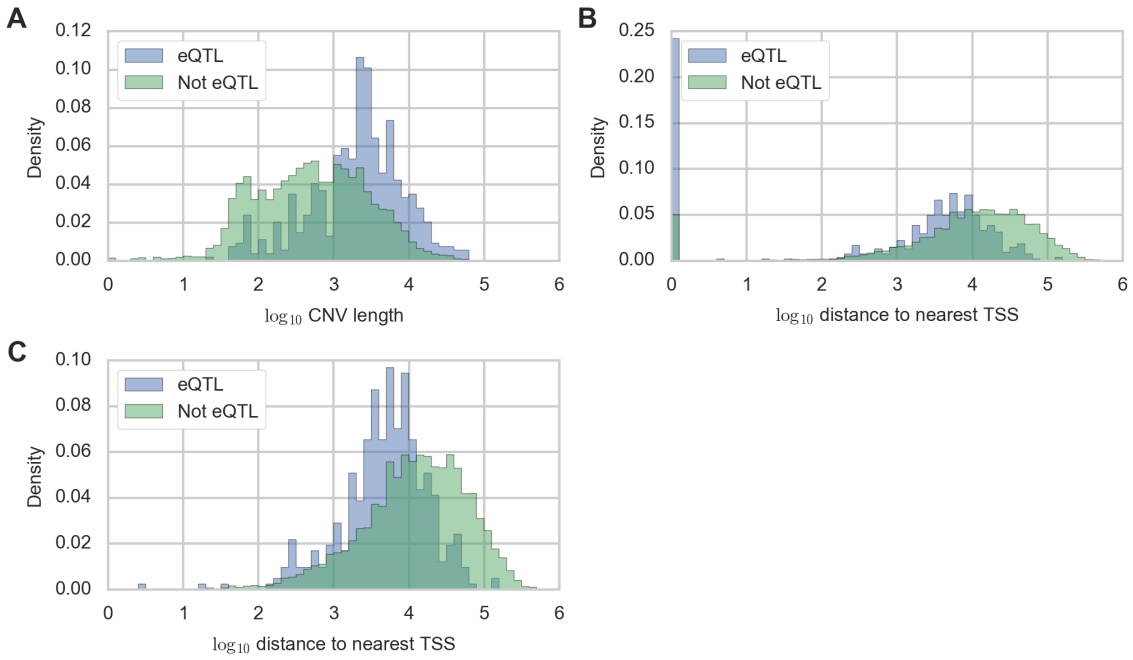
Supplementary Figure 2.1: Donor characteristics of the 215 iPSC lines used for eQTL mapping. (A) Distribution of ages, (B) self-reported ethnicity, and (C) sex. (D) Number of family members per family for families with more than one person in the 215 donor set. Note that some individuals within a family are genetically unrelated to each other due to marriage. In total, there were 131 genetically unrelated individuals out of the 215 subjects used for eQTL mapping.



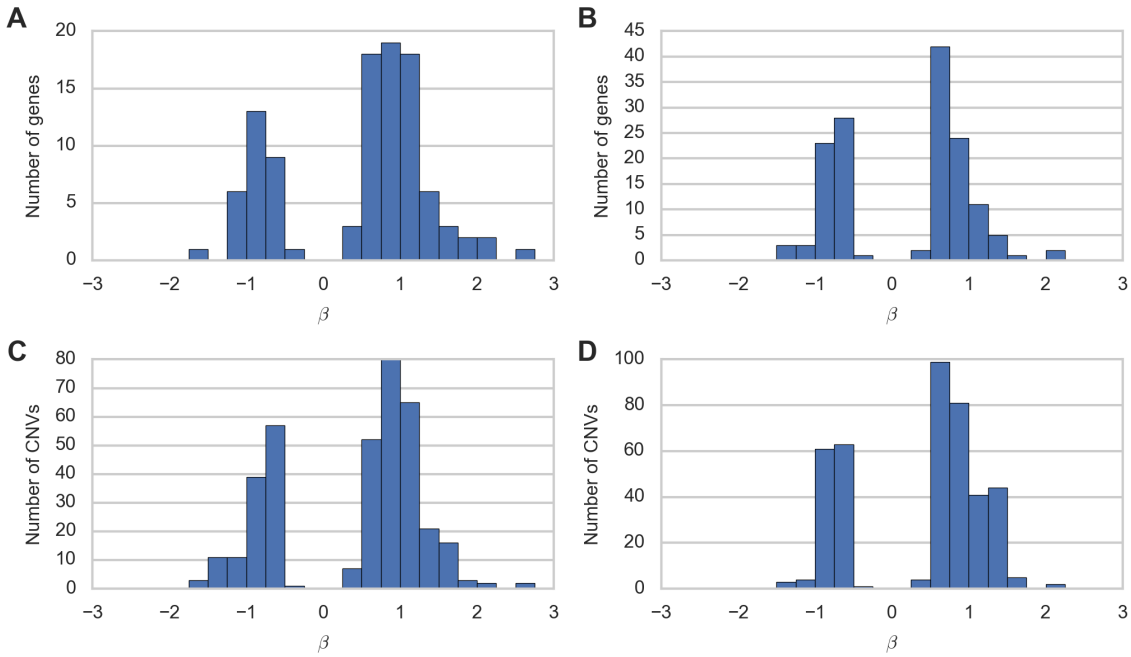
Supplementary Figure 2.2: Gene expression for markers of pluripotency and mesoderm. Expression (log TPM *z*-scores) for nine pluripotency and 25 mesoderm markers from Tsankov *et al.* 2015 in 250 RNA-seq samples (some of the 215 subjects had biological replicate RNA-seq samples) and 35 ESCs, 21 iPSCs, and 17 fibroblasts from GEO accession GSE73211. Fibroblasts (green) have low expression of pluripotency markers but higher expression of most mesoderm markers relative to stem cells.



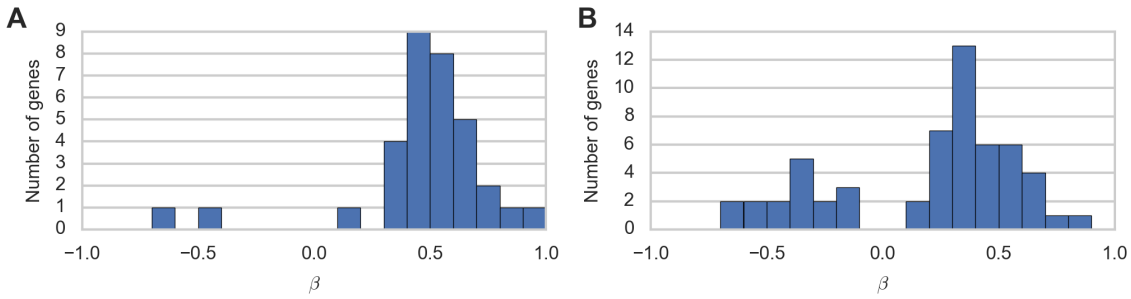
Supplementary Figure 2.3: Distance from lead variants to transcription start sites. (A-C) Distribution of distance from (A) SNVs, (B) indels, and (C) CNVs to the nearest transcription start site of the associated eGene for lead variants. If a gene had multiple lead variants due to *p*-value ties, all lead variants were included.



Supplementary Figure 2.4: CNV eQTL characteristics. Distribution of (A) CNV length and (B) distance to nearest transcription start site (TSS) in $\log_{10}$ base pairs for 545 CNVs that had at least one significant gene expression association and 14,726 CNVs that did not have any significant associations. (C) Distribution of distances to nearest TSS in log10 base pairs after removing CNVs that overlapped a TSS.



Supplementary Figure 2.5: CNV eQTL effect sizes. There were 247 eGenes with at least one significant CNV eQTL. 126 of these genes had more than one significant CNV. Distribution of effect sizes for most significant CNV for (A) 102/247 eGenes where at least one significant CNV overlapped the gene and (B) 145/247 eGenes where no significant CNVs overlapped the gene. Distribution of effect sizes for all significant CNVs for (C) 102/247 eGenes where at least one significant CNV overlapped the gene and (D) 145/247 eGenes where no significant CNVs overlapped the gene.



Supplementary Figure 2.6: mCNV eQTL effect sizes. Distribution of effect sizes for the most significant mCNV association for (A) 33 genes where at least one significant mCNV overlapped the gene and (B) 56 genes where no significant mCNV overlapped the gene.

Chapter 2.10: References

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature reviews. Genetics, 16*(4), 197-212. doi:10.1038/nrg3891

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics, 25*(1), 25-29. doi:10.1038/75556

Avior, Y., Sagi, I., & Benvenisty, N. (2016). Pluripotent stem cells in disease modelling and drug discovery. *Nat Rev Mol Cell Biol, 17*(3), 170-182. doi:10.1038/nrm.2015.27

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., . . . Web Presence Working, G. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics, 25*(2), 288-289. doi:10.1093/bioinformatics/btn615

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*, 7. doi:10.1186/s13742-015-0047-8

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., . . . Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods, 12*(10), 966-968. doi:10.1038/nmeth.3505

Choi, J., Lee, S., Mallard, W., Clement, K., Tagliazucchi, G. M., Lim, H., . . . Hochedlinger, K. (2015). A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. *Nature Biotechnology, 33*(11), 1173-1181. doi:10.1038/nbt.3388

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin), 6*(2), 80-92. doi:10.4161/fly.19695

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:10.1038/nature11247

Consortium, G. T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science, 348*(6235), 648-660. doi:10.1126/science.1262110

Consortium, U. K., Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., . . . Soranzo, N. (2015). The UK10K project identifies rare variants in health and disease. *Nature, 526*(7571), 82-90. doi:10.1038/nature14962

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics, 43*(5), 491-498. doi:10.1038/ng.806

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet, 9*(5), e1003486. doi:10.1371/journal.pgen.1003486

Frazer, K. A. (2016). The CARDiPS resoruce. *TBD*.

Gamazon, E. R., Nicolae, D. L., & Cox, N. J. (2011). A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet, 7*(2), e1001292. doi:10.1371/journal.pgen.1001292

Gene Ontology, C. (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res, 43*(Database issue), D1049-1056. doi:10.1093/nar/gku1179

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Gore, A., Li, Z., Fung, H. L., Young, J. E., Agarwal, S., Antosiewicz-Bourget, J., . . . Zhang, K. (2011). Somatic coding mutations in human induced pluripotent stem cells. *Nature, 471*(7336), 63-67. doi:10.1038/nature09805

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature genetics, 47*(3), 296-303. doi:10.1038/ng.3200

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., . . . Hubbard, T. J. (2012). GENCODE: The reference human genome

annotation for The ENCODE Project. *Genome Research, 22*(9), 1760-1774. doi:Doi 10.1101/Gr.135350.111

Ji, J., Ng, S. H., Sharma, V., Neculai, D., Hussein, S., Sam, M., . . . Batada, N. N. (2012). Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. *Stem Cells, 30*(3), 435-440. doi:10.1002/stem.1011

Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., . . . Young, R. A. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell, 18*(2), 262-275. doi:10.1016/j.stem.2015.11.007

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., . . . Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics, 42*(4), 348-354. doi:10.1038/ng.548

Kheradpour, P., & Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res, 42*(5), 2976-2987. doi:10.1093/nar/gkt1249

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics, 46*(3), 310-315. doi:10.1038/ng.2892

Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P., & Ukkonen, E. (2009). MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics, 25*(23), 3181-3182. doi:10.1093/bioinformatics/btp554

Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A., Monlong, J., Rivas, M. A., . . . Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature, 501*(7468), 506-511. doi:10.1038/nature12531

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome biology, 15*(6), R84. doi:10.1186/gb-2014-15-6-r84

Leslie, R., O'Donnell, C. J., & Johnson, A. D. (2014). GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics, 30*(12), i185-194. doi:10.1093/bioinformatics/btu273

Lessing, D., Anguera, M. C., & Lee, J. T. (2013). X chromosome inactivation and epigenetic responses to cellular reprogramming. *Annu Rev Genomics Hum Genet, 14*, 85-110. doi:10.1146/annurev-genom-091212-153530

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi:10.1186/1471-2105-12-323

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics, 30*(20), 2843-2851. doi:10.1093/bioinformatics/btu356

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, X., Battle, A., Karczewski, K. J., Zappala, Z., Knowles, D. A., Smith, K. S., . . . Montgomery, S. B. (2014). Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *Am J Hum Genet, 95*(3), 245-256. doi:10.1016/j.ajhg.2014.08.004

Lyon, M. F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature, 190*, 372-373.

Martin, J. (2015). NHLBI Next Gen Consortium partners with WiCell to distribute over 1,500 well-characterized, novel cell lines for use in disease research. Retrieved from http://www.wicell.org/home/about-wicell/news-room/62415/6-24-15.cmsx

Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., & Stamatoyannopoulos, J. A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nature genetics, 47*(12), 1393-1401. doi:10.1038/ng.3432

Mayba, O., Gilbert, H. N., Liu, J., Haverty, P. M., Jhunjhunwala, S., Jiang, Z., . . . Zhang, Z. (2014). MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome biology, 15*(8), 405. doi:10.1186/s13059-014-0405-3

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

McKernan, R., & Watt, F. M. (2013). What is the point of large-scale collections of human induced pluripotent stem cells? *Nature Biotechnology, 31*(10), 875-877. doi:10.1038/nbt.2710

Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M., & Dermitzakis, E. T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet, 7*(7), e1002144. doi:10.1371/journal.pgen.1002144

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet, 11*(1), e1004857. doi:10.1371/journal.pgen.1004857

Papp, B., & Plath, K. (2013). Epigenetics of reprogramming to induced pluripotency. *Cell, 152*(6), 1324-1343. doi:10.1016/j.cell.2013.02.043

Pasque, V., & Plath, K. (2015). X chromosome reactivation in reprogramming and in development. *Curr Opin Cell Biol, 37*, 75-83. doi:10.1016/j.ceb.2015.10.006

Pers, T. H., Timshel, P., & Hirschhorn, J. N. (2015). SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics, 31*(3), 418-420. doi:10.1093/bioinformatics/btu655

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research, 20*(1), 110-121. doi:10.1101/gr.097857.109

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi:10.1038/nature14248

Rouhani, F., Kumasaka, N., de Brito, M. C., Bradley, A., Vallier, L., & Gaffney, D. (2014). Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet, 10*(6), e1004432. doi:10.1371/journal.pgen.1004432

Schlattl, A., Anders, S., Waszak, S. M., Huber, W., & Korbel, J. O. (2011). Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Research, 21*(12), 2004-2013. doi:10.1101/gr.122614.111

Shendure, J., & Akey, J. M. (2015). The origins, determinants, and consequences of human mutations. *Science, 349*(6255), 1478-1483. doi:10.1126/science.aaa9119

Stegle, O., Parts, L., Durbin, R., & Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *Plos Computational Biology, 6*(5), e1000770. doi:10.1371/journal.pcbi.1000770

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9440-9445. doi:10.1073/pnas.1530509100

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81. doi:10.1038/nature15394

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell, 131*(5), 861-872. doi:10.1016/j.cell.2007.11.019

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell, 126*(4), 663-676. doi:10.1016/j.cell.2006.07.024

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics, 31*(12), 2032-2034. doi:10.1093/bioinformatics/btv098

Thomas, S. M., Kagan, C., Pavlovic, B. J., Burnett, J., Patterson, K., Pritchard, J. K., & Gilad, Y. (2015). Reprogramming LCLs to iPSCs Results in Recovery of Donor-Specific Gene Expression Signature. *PLoS Genet, 11*(5), e1005216. doi:10.1371/journal.pgen.1005216

Tischler, G., & Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine, 9*, 13-13. doi:10.1186/1751-0473-9-13

Tsankov, A. M., Akopian, V., Pop, R., Chetty, S., Gifford, C. A., Daheron, L., . . . Meissner, A. (2015). A qPCR ScoreCard quantifies the differentiation potential of human pluripotent stem cells. *Nature Biotechnology, 33*(11), 1182-1192. doi:10.1038/nbt.3387

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods, 12*(11), 1061-1063. doi:10.1038/nmeth.3582

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., . . . DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics, 11*(1110), 11 10 11-11 10 33. doi:10.1002/0471250953.bi1110s43

Wen, X., Luca, F., & Pique-Regi, R. (2015). Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet, 11*(4), e1005176. doi:10.1371/journal.pgen.1005176

Zanoni, P., Khetarpal, S. A., Larach, D. B., Hancock-Cerutti, W. F., Millar, J. S., Cuchel, M., . . . Global Lipids Genetics, C. (2016). Rare variant in scavenger receptor BI raises HDL cholesterol and increases risk of coronary heart disease. *Science, 351*(6278), 1166-1171. doi:10.1126/science.aad3517

Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T. J., Lee, C. M., Banskota, S., . . . Gibson, G. (2016). A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *Am J Hum Genet, 98*(2), 299-309. doi:10.1016/j.ajhg.2015.12.023