Boston University Theses & Dissertations

http://open.bu.edu

Boston University Theses & Dissertations

2020

Transcriptional regulation landscape in health and disease

https://hdl.handle.net/2144/41919 Downloaded from OpenBU. Boston University's institutional repository.

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

TRANSCRIPTIONAL REGULATION LANDSCAPE IN HEALTH AND DISEASE

by

SEBASTIAN CARRASCO PRO

B.S., Universidad Peruana Cayetano Heredia, 2015 M.S., Boston University, 2019

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

© 2020 by

SEBASTIAN CARRASCO PRO All rights reserved except Chapter 2, which is © 2018 by Oxford University Press Approved by

First Reader

Juan Ignacio Fuxman Bass, Ph.D. Assistant Professor of Biology and Bioinformatics

Second Reader

Trevor Siggers, Ph.D. Associate Professor of Biology and Bioinformatics

DEDICATION

To my family

Sergio, Chabuca and Rodrigo

ACKNOWLEDGMENTS

My work would have not been possible without many people helping me along the way. To my family (Sergio Carrasco, Chabuca Pro, Rodrigo Carrasco) who has supported and motivated me throughout all these years, all I accomplish is thank to them and this work is no exception. To my girlfriend Agnes (Naiqiong Zhang), who has supported and motivated me, and made the PhD experience as good as it can be. To my advisor Juan, who was always available to discuss ideas and results to advance projects at a fast pace, and always pushed me with adding biological context to bioinformatics results. To my dissertation committee Trevor, Stefano, Josh and Dave for their helpful comments during meetings and special thanks to Dave for our multiple meetings for career advice and countless topics.

To the Fuxman lab members past and current, Jared, Shaleen, Alvaro, Kok, Meimei and the group of undergrads for their contributions in my projects, other lab members Anna, Jaice, Devlin for their feedback and fun interactions in lab events, and in particular to Clarissa, Xing, Luis and Sam for their support not only in the science but also in life. To Adam, who helped us in the development of project algorithms. To our collaborators in Tewhey and Siggers labs for performing the MPRA and CASCADE experiments, respectively, and providing helpful discussion of results. To the RCS members Brian and Katia for their help creating and optimizing databases, queries and other scripts. To the bioinformatics program Mary-Ellen, Gary, Dave, Johanna, Caroline and Tom for their administrative help and support throughout the PhD. To my other family members in Peru and USA, my uncles, grandparents, cousins, who were/are always supporting me in multiple ways. To my friends for all their support and fun moments, which I hope we can keep having. To my Boston friends Anita, Josh, Uros, Xingyi, Emma, David Jenkins, Jason, Ami, Tanya, Marzie, Ana, among others for all the fun times including Terrace House, squash, multiple parties, gossip and so much drama. To my fronton friends in Peru Lucan, Amanda, Ali, Rafo, Juan, among many others, who motivate me and its always a great time to play and hang out with them. To my UPCH friends Chiki, Luis, Neko, Vivi, Andrea, Josue, Jank, among others who are always there for fun times and I can't count how many hours I have spent playing LoL with most of them. To my German friends Marjan, Verena and Mary who are always there to exchange some lobster emojis. To my school friend Bryan, Parra, Pedro who I've known since we started school twenty years ago.

To everyone who has been part of my journey until today, thank you!

TRANSCRIPTIONAL REGULATION LANDSCAPE IN HEALTH AND

DISEASE

SEBASTIAN CARRASCO PRO

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2020

Major Professor: Juan Ignacio Fuxman Bass, Assistant Professor of Biology and Bioinformatics

ABSTRACT

Transcription factors (TFs) control gene expression by binding to highly specific DNA sequences in gene regulatory regions. This TF binding is central to control myriad biological processes. Indeed, transcriptional dysregulation has been associated with many diseases such as autoimmune diseases and cancer. In this thesis, I studied the transcriptional regulation of cytokines and gene transcriptional dysregulation in cancer. Cytokines are small proteins produced by immune cells that play a key role in the development of the immune system and response to pathogens and inflammation. I mined three decades of research and developed a user-friendly database, CytReg, containing 843 human and 647 mouse interactions between TFs and cytokines. I analyzed CytReg and integrated it with phenotypic and functional datasets to provide novel insights into the general principles that govern cytokine regulation. I also predicted novel cytokine promoter-TF interactions based on cytokine co-expression patterns and motif analysis, and studied the association of cytokine transcriptional dysregulation with disease. Transcriptional dysregulation can be caused by single nucleotide variants (SNVs) affecting TF binding sites (TFBS). Therefore, I created a database of altered TFBS (aTFBS-DB) by calculating the effect (gain/loss) of all possible SNVs across the human genome for 741 TFs. I showed how the probabilities to gain or disrupt TFBSs in regulatory regions differ between the major TF families, and that cis-eQTL SNVs are more likely to perturb TFBSs than common SNVs in the human population. To further study the effect of somatic SNVs in TFBS, I used the aTFBS-DB to develop TF-aware burden test (TFABT), a novel algorithm to predict cancer driver SNVs in gene promoters. I applied the TFABT to the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort and identified 2,555 candidate driver SNVs across 20 cancer types. Further, I characterized these cancer drivers using functional and biophysical assay data from three cancer cell lines, demonstrating that most SNVs alter transcriptional activity and differentially recruit cofactors. Taken together, these studies can be used as a blueprint to study transcriptional mechanisms in specific cellular processes (i.e. cytokine expression) and the effect of transcriptional dysregulation in disease (i.e. cancer).

TABLE OF CONTENTS

DEDICATION	iv
ACKNOWLEDGMENTS	v
ABSTRACT	vii
TABLE OF CONTENTS	ix
LIST OF TABLES	xii
LIST OF FIGURES	. xiii
LIST OF ABBREVIATIONS	xx
Chapter 1. Introduction	1
Transcription Factors and Gene Expression	1
Cytokine Dysregulation and Disease	2
 Transcriptional Regulation and Disease Computational approaches to identify noncoding SNVs Hotspot analysis based on mutation frequency Prediction of noncoding SNVs with high functional impact Experimental validation of differential TF binding between SNV alleles Experimental validation of altered gene expression by SNVs. SNVs affecting distal regulatory elements Noncoding SNVs affecting post-transcriptional regulation Future perspectives Funding Author contributions Dissertation aims Aim 1. Determine the transcriptional regulation landscape of cytokines in mouse and human Aim 2. Predict genome-wide effects of single nucleotide variants in gene promoters 	4
Chapter 2. Global landscape of mouse and human cytokine transcriptional regulation	26
Introduction	26
Materials and Methods	28 29 30 31 31 32 32 32
TF-drug associations	33

Prediction of novel PDIs in the human cytokine GRN	
Enhanced yeast one-hybrid (eY1H) assays	
Motif analysis	35
Transient transfections and luciferase assays	35
Code availability	
Statistical analyzes	
Software used to generate the figures	36
Results	36
Generation of CvtReg	
Association between TF connectivity and immune phenotype	
Cytokine regulation by different types of TFs	
GRN integration with TF-cofactor interactions	47
The cytokine GRN as a blueprint to study disease	52
Completeness of the cytokine GRN	57
Prediction of novel PDIs in the cytokine GRN	62
Discussion	66
	00
Funding	69
Anthon contributions	(0
Author contributions	
Chapter 3. Prediction of genome-wide effects of single nucleotide variants on	
transcription factor binding	71
Introduction	71
Materials and Methods	73
Generation of the altered TF binding site database	
Genomic region definitions	
Generation of reference parameters for altered TF binding in genomic regions	75
Analysis of parameter scores for population-wide and cis-eOTL SNVs	
Estimation of a population-wide SNV-specific reference set of TFBS parameters	77
Calculation of parameters for cancer somatic and carcinogen SNVs	77
Statistical analysis	78
Domita	70
Results	
Estimating the effects of SNVs in creating and disrupting TFBS	
aig aOTL SNVs display a high likelihood to grapts and disput TEPSs	
Cancer sometic mutations display cancer and TE family specific effects on TEPS	
Cancer somatic mutations display cancer-and TT family-specific effects on TTBS	
Discussion	89
Funding	92
Author contributions	92
Chapter 4. Discovery and characterization of cancer driver mutations in gene pro-	omoters
	93
Introduction	93
Materials and Methods	95
Altered transcription factor binding predictions.	95
ChIP-seq allelic imbalance analysis	96

Processing of PCAWG mutational data	
Generation and use of the TF-aware burden test	100
Computational validation of cancer driver candidates	101
MPRA library construction	103
MPRA library transfection into cell lines	103
RNA isolation and MPRA RNA-seq library generation	104
MPRA data analysis	105
Mutational signatures for MPRA validated drivers	106
Normalized gene expression analysis	107
Association of creation and disruption of TFBS with target gene expression	107
Cell culture for CASCADE experiments	108
CASCADE protein binding microarray experiments	109
CASCADE-based differential COF recruitment microarray design	112
Analysis of differential COF recruitment	112
Results	
Prediction of noncoding cancer driver SNVs	113
TF-aware driver candidate NCVs lead to altered transcriptional activity	118
Diver NCVs outside core promoter may affect transcriptional activity	
NCVs derived from mutational processes can affect transcriptional activity	124
Transcription factors and their effect in transcriptional activity	
Predicted driver NCVs lead to differential cofactor recruitment	127
Discussion	130
Author contributions	
Chapter 5. Conclusions	
BIBLIOGRAPHY	
CURRICULUM VITAE	166

LIST OF TABLES

Table 1.1. List of computational methods and databases to identify somatic SNVs,	
incorporate background models to predict functional noncoding SNVs, predict	
altered TF binding sites, and integrate with functional annotations. This list is not	
exhaustive, thus, the authors apologize for any method/database not referenced in	
this table10	0
Table 4.1 ChIP-seq experiments downloaded from ENCODE	9
Table 4.2 (A) Primers used in MPRA experiments and (B) Illumina Adaptor/Index	
Primers for Second PCR	5

LIST OF FIGURES

- Figure 1.1 Noncoding cancer mutations affecting transcriptional and post-transcriptional regulation. Somatic mutations (present in tumor but not in matched normal tissue samples) can affect gene regulation by affecting the binding of a transcription factor (TF) to a regulatory region, the binding of RNA binding proteins (RBPs) or miRNAs to untranslated regions (UTRs) in the mRNAs, or affect normal splicing. TF purple, RBP orange, regulatory region blue, UTR green, coding region yellow, SNV red.
- Figure 1.3 Overview of assays to measure differential TF binding between noncoding SNV alleles. (A) ChIP against a candidate TF can be performed in cells that are heterozygous for the SNV. Sequencing of the amplified regions (or allele-specific qPCR) can determine relative TF binding between wild-type (wt) and mutant (mut) alleles. Alternatively, ChIP-seq data can be analyzed to detect biases in the number of sequencing reads between alleles. The figure shows an example of loss of TF binding caused by a mutation. (B) EMSA can be performed to determine differential TF binding to oligonucleotides containing wild-type (wt) or mutant (mut) SNV alleles by using nuclear extracts (NE) followed by super-shifts using antibodies against the candidate TF (α -TF), or by incubating with extracts overexpressing the TF. (C) eY1H assays can test the binding of >1,000 TFs to wild-type and mutant allele sequences. In this assay, each DNA sequence is cloned upstream the HIS3 and LacZ reporters and integrated into the yeast genome. Interactions are tested by mating with yeast strains expressing different TFs in an arrayed format system. Differential TF interactions (highlighted in red) can be determined by comparing
- Figure 1.4 Functional assays to measure altered gene expression and phenotypic parameters induced by SNVs in regulatory regions. (A) Reporter assays can be used to determine differential expression induced by wild-type and mutant regulatory elements in transiently transfected cells. (B) In MPRAs wild-type and mutant alleles for hundreds/thousands of noncoding SNVs can be tested in parallel for changes in transcriptional activity. ~200 bp sequences containing the SNVs are cloned upstream of an inert ORF and associated with random barcodes. Cells are then

- Figure 2.1 Differentially Generation of CytReg. (A) Pipeline used for the text mining and article curation to determine literature-based PDIs between TFs and cytokine genes. (B) Search page of CytReg where PDIs can be browsed by TF, cytokine, species, assay type, and TF expression levels (mRNA and protein) in different immune cells. (C) Results page indicating the interacting cytokines and TFs, the types of assays used to determine the PDIs, whether the interaction is activating or repressing, and the Pubmed IDs of the publications referencing the PDIs. Links are provided to UniProt entries for cytokines and TFs, and to Pubmed for the references. The interactions can be downloaded as a CSV file or visualized as a network graph. (D) Network visualization of the selected PDIs. Nodes represent cytokines and TFs, edges represent the type of interaction (activating, repressing, bifunctional, or physical). Nodes can be moved to re-arrange the network. (E) Overlap of PDIs in CytReg and those annotated in InnateDB and TRRUST. (F) Overlap between mouse and human cytokine GRNs. (G) Fraction of PDIs with high evidence of direct regulatory activity (by a functional assay and an in vitro or in vivo binding assay) or

Figure 2.4 Relationship between TF connectivity and phenotype in the mouse cytokine GRN. (A) Number of cytokine targets per TF (TF degree) in the mouse cytokine GRN ordered by TF degree rank. (B) Number of interacting TFs per cytokine (cytokine degree) in the mouse cytokine GRN ordered by cytokine degree rank. (C) Fraction of TFs in the mouse cytokine GRN with annotated immune phenotypes when knocked out in mice (MGI), or associated with immune disorders in the Human Gene Mutation Database (HGMD) or in genome-wide association studies Figure 2.5 Cytokine regulation by different types of TFs. (A, B) Correlation between the percentage of PDIs involving a TF in the human cytokine GRN versus a global human GRN annotated in TRRUST, for different TF families (A) or for pathogenor stress-activated (PSA) TFs (B). (C) Average fraction of PSA and tissue-specific (TS) TFs for cytokines expressed in different cell types. (D) Fraction of PSA and TS TFs for different classes of cytokines. Correlation determined by Pearson correlation coefficient. (E) Inflammatory score (IS) for each TF based on the fraction of PDIs with pro- and anti-inflammatory cytokines. (F) Percentage of TFs with proinflammatory, anti-inflammatory, and differentiation or other functions based on Figure 2.6 TF families present in the mouse and human cytokine GRNs. (A) Correlation between the percentage of PDIs involving a TF in the mouse cytokine GRN versus a global mouse GRN annotated in TRRUST. (B) Distribution of TF families in the human and mouse cytokine GRNs compared to those annotated in the TRRUST Figure 2.7 Cooperativity and plasticity in cytokine regulation. (A) Protein-protein interaction network from Lit-BM-13 between cofactors and TFs in the human cytokine GRN. Ellipses – TFs, diamonds – cofactors. Node size indicates the number of cytokine targets (for TFs) in the cytokine GRN, and the number of protein-protein interactions with TFs (for cofactors). Only cofactors with five or more protein-protein interactions are shown. (B, C) Number of TFs (shades of grey) interacting with each human cytokine gene that interact with the different cofactors (B) or the different domains of EP300/CREBBP (C). (D, E) Fraction of cofactor (D) or EP300/CREBBP domain (E) protein-protein interactions (shades of red) involving PSA or TS TFs. Only cytokines and cofactors with five or more interactions are shown. Co-activators are shown in red font, co-repressors in blue Figure 2.8 Association of the cytokine GRN with human diseases. (A) Circos plot connecting diseases with TFs based on enrichment of the TFs in regulating cytokines upregulated in the indicated disease. Ribbon width is proportional to the percentage of cytokines upregulated in the indicated disease that are regulated by the indicated TF. (B) GRN connecting interacting TFs and human cytokine genes associated with autoimmune disorders. Edges connect interacting cytokine-TF pairs. Edge color indicates that the interacting cytokine and TF are associated with the same disease based on HGMD and GWAS. (C) The human cytokine GRN was

randomized 1,000 times by edge switching and the number of TF-cytokine-disease

Figure 2.9 Gene expression of Berry et al. (2010) 86-gene signature in TB and LTBI subjects from a South Indian population. Circos plot connecting diseases with TFs based on enrichment of the TFs in regulating cytokines upregulated in the indicated disease. Ribbon width is proportional to the percentage of cytokines upregulated in the indicated disease that are regulated by the indicated TF. The left plot is based on PDIs from the union of TRRUST and InnateDB, the right plot is based on PDIs from CytReg (as in Figure 2.8A).

Figure 2.10 Completeness of the human cytokine GRN. (A) Number of annotated PDIs, TFs, and cytokines in the human cytokine GRN over time. (B) Fraction of TFs in the human cytokine GRN with annotated immune phenotypes when knocked out in mice (MGI) or associated to immune disorders in genome-wide association studies (GWAS) and in the Human Gene Mutation Database (HGMD) over time. (C, D) Number of PDIs per TF (C) or per cytokine (D) in the human cytokine GRN over time. (E, F) Correlation between the number of PDIs in the human cytokine GRN and the number of publications per TF (E) or per cytokine (F) reported in Medline. (G, I) PDIs with the promoters of CCL27 (G) or CCL4L2 (I) were analyzed by eY1H assays. Each interaction was tested in guadruplicate. The gualitative strength of PDIs detected by eY1H compared to AD-vector control are indicated as -, +, ++, and +++ corresponding to no, weak, medium, and strong interaction, respectively. Motif location for the indicated TFs in the promoters of CCL27 and CCL4L2 are shown. (H, J) Luciferase assays to validate interactions between the promoters of CCL27 (H) or CCL4L2 (J) and the indicated TFs. HEK293T cells were cotransfected with reporter plasmids containing the cytokine promoter region (2 kb) cloned upstream of the firefly luciferase reporter gene, and expression vectors for the indicated TFs (fused to the activation domain 10xVP16). After 48 h, cells were harvested and luciferase assays were performed. Relative luciferase activity is plotted as fold change compared to cells co-transfected with the vector control (1.0). Experiments were performed 3-4 times in three replicates. Individual data points represent the average of the three replicates, the average of all experiments is indicated by the black line. *p<0.05 by one-tailed Student's t-test with Benjamini-

Figure 2.11 Completeness of the mouse cytokine GRN. (A) Number of annotated PDIs, TFs, and cytokines in the mouse cytokine GRN over time. (B) Fraction of TFs in the mouse cytokine GRN with annotated immune phenotypes when knocked out in mice (MGI) or associated to immune disorders in genome-wide association studies (GWAS) and in the Human Gene Mutation Database (HGMD) over time. (C, D)

- Figure 2.12 Prediction of novel PDIs in the human cytokine GRN. (A) Novel PDI predictions based on co-expression between cytokines and known cytokine targets of each TF (determined using the SEEK database), and motifs analysis. Prediction confidence, as defined in the methods section, is shown. (B) Correlation between the number of cytokine targets (TF degree) for known PDIs and known + predicted PDIs. Correlation determined by Spearman's rank correlation coefficient. (C) Correlation between TF degree for known (K) or known + predicted (K+P) PDIs and expression enrichment score (EES) in immune tissues, mouse immune phenotype (MGI), and human immune disorders in GWAS and HGMD. Correlation and significance determined by Spearman's rank correlation coefficient. (D, G) Top predicted cytokine targets of RORC (D) and REL (G). The co-expression rank among all genes and among cytokines is shown. CXCL8 is a known target of REL, while IL17A is a known target of RORC. (E, H) Enhanced yeast one-hybrid assays testing PDIs between the indicated human cytokine promoters and RORC (E) and REL (H). AD-vector corresponds to empty vector. The qualitative strength of PDIs compared to AD-vector control are indicated as -, +, ++, and +++ corresponding to no, weak, medium, and strong interaction, respectively. REL and RORC binding sites are indicated in red for each 2 kb promoter region. (F, I) Luciferase assavs in HEK293T cells co-transfected with reporter plasmids containing the indicated cytokine promoter region (2 kb) cloned upstream of the firefly luciferase reporter gene, and expression vectors for RORC (F) or REL (I) (fused to the activation domain 10xVP16). After 48 h, cells were harvested and luciferase assays were performed. Relative luciferase activity is plotted as fold change compared to cells co-transfected with the vector control (1.0). Experiments were performed 3-4 times in three replicates. Individual data points represent the average of the three replicates, the average of all experiments is indicated by the black line. p<0.05 by
- Figure 3.2 Differential parameter scores for population-wide and cis-eQTL SNVs. (A-D) Correlation between scores derived from SNVs from the 1000 Genomes Project (1000 genomes) and the average of 100 random sets of 1,000,000 SNVs (reference) for gainability (A), disruptability (B), hitability (C), and robustness (D). Correlation

was determined by the Pearson correlation coefficient. Significantly enriched (red) and depleted (blue) TFs are highlighted. (E-H) Δ scores (observed in set – reference) for each parameter for all TFs and specific TF families for population-wide and ciseQTL SNVs. Significant differences between the population-wide and cis-eQTL Figure 3.3 Effect of cancer somatic mutations on TFBSs. (A-D) Median Δ scores for each TF family and cancer-type combination for gainability (A), disruptability (B), hitability (C), and robustness (D). (E-F) Motifs logos for NFATC4 (E) and ELF4 (F) and impact of melanoma mutational signatures on the gain and disruption of the Figure 3.4 Effect of cancer somatic mutations in individual cancers on Againability and ∆disruptability. (A-B) For cancer samples with at least 5,000 SNVs in DHS regions, we determined for each TF the Δ gainability (A) and Δ disruptability (B) scores. Samples were clustered using hierarchical clustering, and TF were clustered by TF families. Cancer-types are indicated at the top and TF families are indicated at the right of each heatmap, respectively. (C-D) Correlation between UV-light-derived Δ gainability (C) and Δ disruptability (D) scores for each TF to those observed in skin Figure 4.1. ChIP-seq allelic imbalance F-scores versus Δ allele score threshold. Arrows Figure 4.2 Driver NCVs prediction and their association with cancer genes and pathways. (A) Number of significant NCVs with predicted gain and/or loss of TF binding per cancer type. (B) Genes with the most predicted cancer driver NCVs and the percent of patients affected per cancer type. (C) Metascape network showing the intracluster and inter-cluster similarities of enriched gene ontology terms for genes with significant NCVs. (D) Fraction of essential and fitness related genes for genes with predicted NCVs, in CGC, or all protein-coding genes. (E) Fraction of genes whose expression has favorable, unfavorable (or either) prognosis in cancer for genes with Figure 4.3. Number of predicted cancer driver NCVs and number of SNVs by cancer Figure 4.4 Predicted driver NCVs can alter transcriptional acitvity. (A) Validation rate versus q-value threshold in SK-MEL-28 for predicted driver NCVs, ChIP-seq allelic imbalance, known drivers, MPRA positive controls, germline NCVs, literature genes, no significant differential binding, no differential binding. (B) Validation rate vs q-value in SK-MEL-28 for predicted NCVs based on whether NCV caused gain, loss of TFBS or both. (C) Fraction of NCVs per frequency in patient samples. (D) Fraction of MPRA validated NCVs for genes with at least four transcriptionally active NCVs by NCV effect (up/downregulation) in each of the three cell lines...119 Figure 4.5 MPRA validation rate. Validation rates versus q-value for (A) Jurkat and (B) HT-29 cell lines for the categories referenced in figure 4.3A.....120 Figure 4.6. Three-way Venn diagram displaying the number of MPRA validated NCVs

Figure 4.7 NCV validation rate by TSS distance and mutational signature type. (A)
Validation rate of predicted driver NCVs in SK-MEL-28 by genomic distance to
TSS, and fraction of NCVs per 100 bp for predicted driver NCVs, MPRA active
NCVs and SNVs in the PCAWG cohort. (B) Validation rate for NCVs associated or
not with APOBEC mutational processes for the three cell lines. (C) Validation rate
of predicted driver NCVs associated or not with UV-light mutational signature in
SK-MEL-28
Figure 4.8 Transcription factor effect on transcriptional activity. (A) Fraction of TF
families with altered TFBS caused by predicted driver NCVs by cancer type. (B)
MPRA validation rate in SK-MEL-28 versus q-value for TF families. (C)
Normalized TPM of genes with predicted driver NCVs by TFs associated with
gain/overexpression and loss/underexpression
Figure 4.9 Predicted drivers cause differential COF recruitment. Differential COF
recruitment for predicted driver NCVs and no predicted binding NCVs (validation
rate) showed as Δz -score versus -log10(q-value) in SK-MEL-28 for (A) TBL1XR1
and (B) P300 + peptides, where dotted line on y-axis represents significance
threshold and on x-axis no differential COF recruitment (0 Δz -score). Significance
values, -log10(q-value), for 3 register versus 1 register array for predicted driver
NCVs and no predicted binding (validation rate) for (C) TBL1XR1, (D) P300 +
peptides, (E) SKP2, and (F) P300. Dotted lines represent significance thresholds. 129

LIST OF ABBREVIATIONS

aTFBS-DB	Altered transcription factor binding site database
CASCADE	Comprehensive assessment of complex assembly at DNA elements
ChIP	Chromatin immunoprecipitation
ChIP-seq.	Chromatin immunoprecipitation sequencing
DNA	
eY1H	Enhanced yeast-one hybrid
FDR	
GRN	
ICGC	International cancer genome center
MPRA	
NCVs	
PCAWG	Pan-cancer analysis of whole genomes
PWM	
RNA	
RNA-Seq	
SNV	
TCGA.	
TF	
TFABT	
TFBS	

Chapter 1. Introduction

Transcription Factors and Gene Expression

Transcription Factors (TFs) are a group of approximately 1,600 proteins that control gene expression by regulating transcription and their activity ultimately determines how cells function and respond to environmental cues (Lambert et al. 2018). The cellular processes TFs are involved with range from cell cycle progression to cellular differentiation and response to external stimuli. TFs regulate transcription by binding to highly specific short DNA sequences (6-12 bp) in regulatory regions such as gene promoters and enhancers (Wunderlich and Mirny 2009). The DNA binding specificity of TFs is determined by their DNA binding domain (DBD), which has been used to group TFs with similar DBDs into TF families, such as nuclear receptors, C2H2 zinc fingers and homeodomains (Johnson and McKnight 1989; Vaquerizas et al. 2009).

Several experimental methods have been developed to study TF binding specificities such as protein-binding microarrays (Berger et al. 2006), high-throughput systematic evolution of ligands exponential enrichment (Jolma et al. 2013), bacterial one-hybrid (Meng, Brodsky, and Wolfe 2005), and chromatin immunoprecipitation sequencing (ChIP-seq) (Valouev et al. 2008b). These methods allow us to determine the motif of a TF, the nucleotide content in each DNA binding position, which can be used to predict the binding of a TF to a new DNA sequence. However, only around 60% of TFs have motifs characterized (Lambert et al. 2018) because experimental methods have limitations such as the expression, purification and post translational modifications of the TF, availability of reagents (i.e. TF antibody for ChIP-seq), and the time and resources needed (i.e these

methods can only test one TF at a time). In addition, TFs may act in complexes and require interaction with co-factors in order to recruit (activate) or block (repress) RNA polymerase, leading to the expression or repression of its target gene (Shlyueva, Stampfel, and Stark 2014; Spitz and Furlong 2012). Further, other factors influence TF binding to DNA such as chromatin accessibility, DNA methylation status, and DNA local and global topology (Shlyueva, Stampfel, and Stark 2014; Spitz and Furlong 2012). Even though these assays characterize motifs and can be used for binding predictions, they do not provide any functional information (i.e. gene activation/repression) and the majority of them are performed *in-vitro*. Therefore, functional assays such as reporter assays and integration of sequencing and transcriptomics data are required to determine how the TF binding to a gene regulatory affects its target gene expression *in-vitro/in-vivo*.

Cytokine Dysregulation and Disease

Cytokines are small proteins predominately produced by macrophages and helper T cells, among other immune cell types (J. M. Zhang and An 2007). Known cytokines range from 132 to 261 genes, as some lists include growth factors, hormones, or cytokine receptors (Wong et al. 2016; Al-Yahya et al. 2015; Kveler et al. 2018). Indeed, 133 have been compiled to be involved primarily in the immune system (Carrasco Pro et al. 2018). Cytokines may have autocrine, paracrine, or endocrine action in cell communication (J. M. Zhang and An 2007) and they play a key role in the development of the immune system as well as response to pathogens and inflammation (J. M. Zhang and An 2007; Medzhitov and Horng 2009).

Cytokine expression dysregulation can be caused by mutations in gene regulatory regions (i.e. promoters, enhancers), changes in the TFs that regulate them, or changes in genes in related signaling pathways (Turner et al. 2014). This dysregulation has been associated with multiple diseases including autoimmune disorders, susceptibility and response to pathogens, and cancer (Carrasco Pro et al. 2018). For example, upregulation of IL-1 and IL-6 has been observed in chronic inflammatory and autoimmune disorders, including type I diabetes, rheumatoid arthritis, lupus nephritis, psoriasis and systemic sclerosis (Turner et al. 2014; Rosa et al. 2008; Kawaguchi, Hara, and Wright 1999). In addition, TNF α plays a central role in essential cellular functions, such as cell proliferation, apoptosis and necrosis (Turner et al. 2014; MacEwan 2002). However, its altered expression has been associated with rheumatoid arthritis (Arend and Dayer 1995), parkinson's diasease (Mogi et al. 1994), and alzheimer's disease (Holmes et al. 2009), among others. Further, overexpression of cytokines such as CCL2, CCL5, CCL7, IL-8, and CXC10 have been observed in bronchial biopsies of asthmatic patients and murine models (Miotto et al. 2001; Medoff et al. 2002). Finally, cytokines have been associated with cancer performing as growth factors (i.e. CXCL8 in melanoma, liver and pancreatic tumors) (Schadendorf et al. 1994; Miyamoto et al. 1998), angiogenic and angiostatic factors (i.e. CXCL8, CXCL10, CCL1, and CCL11) (Belperio et al. 2000; Bernardini et al. 2000; Salcedo et al. 2001), and playing a role in metastasis (i.e. CXCR4 and CXCR7 in breast cancer) (Müller et al. 2001). These vast implications of cytokines in disease require the study of their transcriptional regulation and the development of gene regulatory networks

describing their proper regulation as well as their dysregulation mechanisms in disease, which will ultimately lead to better disease diagnostics and therapeutics.

Transcriptional Regulation and Disease

Adapted from the following manuscript:

 Kok Ann Gan#, Sebastian Carrasco Pro#, Jared Allan Sewell, Juan Ignacio Fuxman Bass. Identification of single nucleotide non-coding driver mutations in cancer. Frontiers in genetics. 2018 Feb 2;9. doi: 10.3389/fgene.2018.00016. eCollection 2018.

co-first authors

Cancer initiation, progression, maintenance, and metastasis originate from somatic single nucleotide variants (SNVs), small insertions and deletions, structural variants, and epigenetic alterations (Helleday, Eshtad, and Nik-Zainal 2014a). In particular, recent whole-genome sequencing studies of tumor samples, through collaborative projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), have identified millions of somatic SNVs associated with different types of cancers (McLendon et al. 2008; Weinstein et al. 2013; Nik-Zainal et al. 2016). Although, these projects and follow-up studies have been successful at identifying common sets of mutated genes and pathways across many cancer types, the functional role of most mutations detected remains to be determined. Indeed, the main challenge in analyzing the genetics underlying cancer is to distinguish driver mutations (i.e., positively selected)

mutations that provide growth advantage to tumor cells) from passenger mutations (i.e., inert mutations that do not confer any growth advantages) (Khurana et al. 2016). This requires the integration of computational analyses that predict functional SNVs with experimental pipelines to validate and characterize those SNVs.

Most studies have focused on characterizing the functional impact of SNVs on coding regions given that it is relatively straightforward to computationally predict how a protein sequence and/or structure will be affected by a missense, nonsense or frameshift mutation. However, the vast majority of SNVs identified in cancer samples reside in noncoding regions of the genome (Araya et al. 2016a). These noncoding SNVs can affect the binding of transcription factors (TFs), RNA-binding proteins (RBPs), and micro RNAs (miRNAs) (Figure 1.1) (Khurana et al. 2016). This in turn affects multiple gene regulatory functions including chromatin structure or accessibility, transcription, DNA methylation, splicing, as well as 5^c and 3^c untranslated region (UTR) function, which ultimately increases or decreases the production, stability and translation efficiency of mRNA transcripts (Khurana et al. 2016).



Figure 1.1 Noncoding cancer mutations affecting transcriptional and post-transcriptional regulation. Somatic mutations (present in tumor but not in matched normal tissue samples) can affect gene regulation by affecting the binding of a transcription factor (TF) to a regulatory region, the binding of RNA binding proteins (RBPs) or miRNAs to untranslated regions (UTRs) in the mRNAs, or affect normal splicing. TF – purple, RBP – orange, regulatory region – blue, UTR – green, coding region – yellow, SNV – red.

Despite recent advances in the understanding of the downstream consequences of noncoding SNVs, it remains a challenge to identify noncoding driver mutations and the mechanisms through which they effect biological functions. First, as stated above, noncoding SNVs can affect multiple regulatory functions including transcription and post-transcriptional regulation. Second, noncoding regions present higher mutations rates than coding regions, due to weaker selective pressure (Weinhold et al. 2014). As a result, parsing through a higher number of passenger mutations to find noncoding driver SNVs becomes a difficult statistical and computational task (Vogelstein et al. 2013). Third, it is challenging to computationally predict whether a noncoding SNV affects gene expression or mRNA stability because the logic involved in regulatory element function has not yet been fully elucidated. Thus, computational predictions of altered regulatory function need to be confirmed by extensive experimental validation using reporter assays, genome editing, measurement of endogenous gene expression, and/or chromatin immunoprecipitation.

Early studies that identified noncoding driver SNVs compared the sequence of regulatory regions of candidate cancer-related genes between tumor and non-tumor samples in order to determine whether these mutations disrupt or create TF binding sites. For example, SNVs were identified in the GTAAC sequence within the first intron of MYC in samples from multiple patients with Burkitt lymphomas (Zajac-Kaye, Gelmann, and Levens 1988). These mutations, which lead to increased MYC expression, abrogated the

binding of a then unidentified TF. Since this early work, targeted studies have identified several mutations in regulatory regions, both in tumor samples and in patients with increased cancer incidence (Stenson et al. 2009).

More recently, whole-genome sequencing of matched tumor and normal samples has enabled the identification of millions of SNVs. However, the identity of the SNVs responsible for driving cancer and those that constitute passenger mutations remains to be determined. Two pioneering studies showed that mutations present in the telomerase reverse transcriptase (TERT) promoter in tumor samples of patients with melanoma lead to increased TERT mRNA expression (S. Horn et al. 2013; Huang et al. 2013). These studies identified two independent C>T transitions, at around -100 bp from the TERT transcription starting site (TSS), that create a 11 bp nucleotide stretch containing a consensus binding site for E-twenty-six (ETS) TFs. Additionally, other mutations in the TERT promoter have been found in melanoma as well as in other cancer types such as ovarian, follicular thyroid, and meningiomas (Goutagny et al. 2014; S. Horn et al. 2013; T. Liu et al. 2014; R. C. Wu et al. 2014). More recently, mutations in the regulatory regions of other cancer-related genes have been identified, including recurrent mutations in the promoters of PLEKHS1, WDR74, SDHD, and FOXA1 that alter gene expression levels, TF binding and that are associated with poor prognosis (Fredriksson et al. 2014; Nik-Zainal et al. 2016; Rheinbay et al. 2017; Weinhold et al. 2014). Here, we present an overview of state-of-the-art approaches to computationally predict and functionally validate driver somatic noncoding SNVs, as well as recent findings associated with cancer.

Computational approaches to identify noncoding SNVs

Computational approaches to predict functional SNVs within regulatory regions share a common general pipeline, including the identification of somatic SNVs, comparison with common germline variants, constraining the analysis to regulatory regions (in some cases, close to cancer-related genes), identification of mutational hotspots, and determining altered TF binding sites (Figure 1.2).



Figure 1.2 Computational pipeline to prioritize somatic SNVs in regulatory elements. Wholegenome sequencing (WGS) of matched tumor and normal samples are analyzed to identify somatic mutations. Identification of mutations within regulatory regions is performed by restricting analyses to promoter regions, generally defined around transcription start sites, and distal elements such as enhancers, predicted based on DHSs and/or histone marks. Hotspot analyses are used to identify regions with increased mutational burden compared to background models based on mutational frequency in neighboring regions and/or regions with similar functional roles. Covariates such as replication timing or gene expression levels can be included to account for mutational heterogeneity across the genome. Motif analyses are performed to predict differential TF binding between SNV alleles. Prioritized noncoding SNVs are usually validated in functional assays.

The identification of somatic SNVs requires comparing the genome sequences of tumor samples with matched normal tissue samples. This is a challenging task because somatic SNVs occur at low frequency in the genome (0.1 to 100 SNVs per megabase), which needs to be distinguished from errors derived from whole-genome sequencing and genome alignment pipelines (Alioto et al. 2015; M. S. Lawrence et al. 2013). Thus, most methods used to identify somatic SNVs require high sequencing depths (usually 30-300x) and paired-end reads, leading to elevated sequencing costs (Alioto et al. 2015). In addition, given that tumors are comprised by heterogeneous populations of cells, many functional SNVs may be present at a low frequency in patient samples (Carter et al. 2012; Nik-Zainal, Van Loo, et al. 2012). Therefore, while high-frequency SNVs can be identified provided that the sequencing depth is sufficient enough and that computational pipelines accommodate for sequence heterogeneity, low-frequency SNVs may require single-cell genome sequencing approaches (Eirew et al. 2015; Navin et al. 2011; Zong et al. 2012).

Several computational methods have been developed to identify somatic SNVs, including: (1) those that separately call SNVs in tumor and normal samples and then identify tumor-specific SNVs by comparison, such as GATK (Depristo et al. 2011), GATKcan (Hsu et al. 2017), and EBCall (Shiraishi et al. 2013); and (2) those that concurrently analyze tumor-normal samples using heuristic methods or statistical models, such as MuTect (Cibulskis et al. 2013), VarScan (Koboldt et al. 2009; 2012), and Strelka

(Saunders et al. 2012) (Table 1). While the first type of methods models sequencing errors based on statistical parameters from the sequencing reads or from non-matched normal samples, the second type of methods compare matched tumor-normal samples to distinguish true mutations from sequencing errors. Even though these algorithms have been used as stand-alone methods to call SNVs, some studies have used a combination of methods for a "wisdom of the crowd" approach with the goal of increasing the confidence in the SNVs detected (Melton et al. 2015; Weinhold et al. 2014).

Goal	Method/Database	Reference
	GATK	(DePristo et al., 2011)
	GATKcan	(Hsu et al., 2017)
Identification of somatic	EBCall	(Shiraishi et al., 2013)
SNVs	MuTect	(Cibulskis et al., 2013)
	Varscan	(Koboldt et al., 2009)
	Varscan2	(Koboldt et al., 2012)
	Strelka	(Saunders et al., 2012)
Incorporation of background	MutSigNC	(Rheinbay et al., 2017)
models for noncoding SNVs	LARVA	(Lochovsky et al., 2015)
	MOAT	(Lochovsky et al., 2017)
	FIMO	(Grant et al., 2011)
	MotifbreakR.	(Coetzee et al., 2015)
	BEEML-PBM	(Hume et al., 2015)
	TFM-pvalue	(Touzet and Varre, 2007)
Prediction of TF binding sites	MotifLocator	(Claeys et al., 2012)
	CIS-BP	(Weirauch et al., 2014)
	Jaspar	(Khan et al., 2017)
	Uniprobe	(Hume et al., 2015)
	Transfac	(Matys et al., 2003)
	RegulomeDB	(Boyle et al., 2012)
Integration with functional	Funseq2	(Fu et al., 2014)
annotation of noncoding	ENCODE Project	(Consortium, 2012)
regions	Roadmap Epigenomics	(Roadmap Epigenomics et al., 2015)
	FANTOM Consortium	(Andersson et al., 2014)
	GTEx Project	(Consortium, 2013)

Table 1.1. List of computational methods and databases to identify somatic SNVs, incorporate background models to predict functional noncoding SNVs, predict altered TF binding sites, and integrate with functional annotations. This list is not exhaustive, thus, the authors apologize for any method/database not referenced in this table.

Hotspot analysis based on mutation frequency

Among the millions of noncoding somatic SNVs identified in different cancers,

only a small number are expected to be drivers. Given that it is not currently possible to

experimentally test most of the SNVs identified, methods have been developed to prioritize which SNVs are more likely to be functional. A common approach to prioritize somatic SNVs is to determine genomic regions with high mutation frequency across different cancer samples. Given the billions of bases in the human genome, the thousands of mutations per cancer sample, and that we only have sequencing data for a few thousand tumors, the chances of detecting a significantly enriched mutation across cancers after multiple hypothesis testing correction is almost null.

Currently, there are two complementary strategies, frequently used together, to increase the power to detect noncoding driver mutations. One strategy is to focus on DNA elements that are expected to have a regulatory function. For example, promoter regions are relatively easy to determine by selecting regions up- and downstream of transcription start sites, while distal elements are usually determined based on DNase hypersensitivity sites (DHSs) or histone marks such as H3K4me and H4K27ac (Figure 1.2) (Dunham et al. 2012). Further, some studies constrain the analyses to the regulatory regions of cancerrelated genes such as those compiled in the Cancer Gene Census (Futreal et al. 2004). Overall, restricting the analysis to a set of regulatory regions reduces the search space for SNVs and, thus increases the power to detect driver mutations.

The second strategy is the identification of clusters of SNVs within short DNA windows, called hotspots, rather than single mutations (Figure 1.2). This reduces dimensionality and increases the frequency of SNVs within each DNA window leading to increased statistical power. The identification of these mutational hotspots across cancers involves comparing the SNV frequency within a DNA window to a background

distribution of SNV frequencies. These methods can be divided into local and global models, comparing the SNV frequencies to other windows in neighboring genomic regions or to functionally similar regions (e.g., other promoters or enhancers), respectively. The window size selection can vary widely between analysis, ranging from 50 bp up to 500 kb (Fujimoto et al. 2016). While short windows provide higher resolution, allowing one to identify functional promoter or enhancer regions, they lead to low statistical power and thus many functional regions may be missed (Fujimoto et al. 2016). Long windows do not have the resolution to detect functional promoters or enhancers but allow for the identification of covariates, regional features associated with genomic heterogeneity in mutation frequency, such as replication timing and gene expression levels (Fujimoto et al. 2016). Both types of methods can be integrated with one another to increase the chances of detecting driver mutations. For example, a recent study analyzing 863 human tumors has identified recurrent mutations in regulatory elements upstream of TERT, PLEKHS1, WDR74 and SDHD in different types of cancer by using 50 bp windows to find hotspots and regional recurrence approaches that take into account length and replication timing (Weinhold et al. 2014).

Although studies using low tumor sample numbers may be underpowered to identify hotspot regions, large samples sizes can also be challenging to analyze. This is because large sample sizes frequently lead to larger lists of potentially significant genes which in many cases do not have cancer-related functions, suggestive of a high false positive prediction rate (M. S. Lawrence et al. 2013). This stems from using background mutation models that do not account for mutational heterogeneity between samples and

across genomic regions (M. S. Lawrence et al. 2013). Pipelines such as MutSigNC have been developed to correct for variation in mutation frequency by considering patientspecific mutation rates, patient-specific sequencing coverage, information about regional mutation clustering, and using as background the mutation rates of promoters (Table 1) (Rheinbay et al. 2017). Other computational frameworks have also been used to also include distal elements in the analyses, including LARVA that incorporates background models for noncoding regions by integrating SNVs with a comprehensive set of noncoding functional elements based on DHSs and histone marks (Table 1) (Lochovsky et al. 2015). In addition, LARVA uses regional genomic features like replication timing allowing to better estimate local mutation rates and mutational hotspots.

Further covariates can be included while modelling mutation frequencies. For instance, recent studies have shown that some breast tumors have mutations mediated by the alipoprotein B messenger RNA-editing enzyme catalytic (APOBEC) which have been found to occur in dense hypermutated regions in the genome (kataegis) (Alexandrov et al. 2013; Nik-Zainal, Alexandrov, et al. 2012). These mutations share a sequence pattern (TCW, where W is A/T), which can be used to assign mutations a probability of being originated by APOBEC activity (Roberts et al. 2013), leading to a more conservative approach to call candidate mutations. This approach identified SNVs in breast cancer samples within the regulatory regions of FOXA1, RMRP, and NEAT1 that affect gene expression levels (Rheinbay et al. 2017). Alternatively, covariates can be avoided altogether by using a non-parametric, permutation-based approach such as MOAT, that does not make assumptions about the mutation process except for requiring that the

background-mutation rate changes smoothly with genomic features (Table 1) (Lochovsky et al. 2015). The variety of co-existing computational approaches, background models, and covariates included in those models, highlights the challenges currently faced in identifying mutational hotspots associated with cancer.

Prediction of noncoding SNVs with high functional impact

Hotspot analyses allow for the prioritization of cancer driver candidate SNVs. However, to further narrow down the set of functional SNVs and predict the functional impact of these SNVs, location and sequence context of the mutations must be integrated with functional models of noncoding regions. One of the most widely used approaches to prioritize SNVs in regulatory regions involves the identification of TF binding sites created or disrupted by the mutations (Figure 1.2). These TF binding differences between SNV alleles can be predicted based on DNA specificities determined by protein-binding microarrays, SELEX, bacterial one-hybrid assays, or chromatin immunoprecipitation (ChIP) followed by next generation sequencing (ChIP-seq) (Jolma et al. 2013; Noves et al. 2008; Weirauch et al. 2014). Currently, DNA binding specificities have been determined for nearly half of human TFs, which are available in different repositories such CIS-BP, Jaspar, Uniprobe, and Transfac (Table 1) (Hume et al. 2015; Weirauch et al. 2014; Matys et al. 2003; Khan et al. 2018). Differences in TF binding between SNV alleles can be predicted using position weight matrices (PWMs), probabilistic representations of DNA binding specificities, and motif prediction algorithms such as FIMO (Grant, Bailey, and Noble 2011), MotifbreakR (Coetzee, Coetzee, and Hazelett 2015), BEEML-PBM (Hume et al. 2015), TFM-pvalue (Touzet and Varré 2007), and MotifLocator (Aerts et al. 2005;

Claeys et al. 2012) (Table 1). For example, MotifLocator, a tool to score how mutations affect wild-type TF binding sites, led to the identification of gain of binding sites for RB1, E2F1 and ETS to multiple promoter regions in tumor samples from TCGA (Kalender Atak et al. 2017). Similarly, mutations in the promoter of FOXA1, a known gene driver in breast cancer, were found to increase E2F binding using TFM-pvalue (Rheinbay et al. 2017). Loss of TF binding sites have also been widely associated with cancer. For example, many recurrent mutated regions in cancer genomes have been found to overlap with CTCF binding sites, showing a possible selection for these mutations (Katainen et al. 2015; Lochovsky et al. 2015; Piraino and Furney 2017). In addition, disruption of FOX TF binding sites in the BCL6 promoter have been reported in follicular lymphoma using an integrative approach that identifies functional regulatory mutation blocks (Batmanov et al. 2017). Interestingly, both the creation and disruption of binding sites for the same TFs have been linked to cancer. For example, by integrating motif analyses with evolutionary conservation, creation of ETS binding sites were determined in the ANKRD53 promoter, while disruption of ETS binding sites were identified in the TAF11 and SDHD promoters (Weinhold et al. 2014).

In addition, motif analyses can integrate functional annotations of regulatory sequences (including DHSs, histone marks, and sequence conservation) and TF expression levels such as those provided by the ENCODE, Roadmap Epigenomics, FANTOM, and GTEx Projects to constrain the analyses to TFs expressed and regulatory elements active in the tissues of interest (Andersson et al. 2014; Dunham et al. 2012; Lonsdale et al. 2013; Roadmap Epigenomics Consortium et al. 2015) (Table 1). These approaches include
RegulomeDB (Boyle et al. 2012) that considers functional annotations for the regulatory regions, and Funseq2 (Fu et al. 2014) that also considers sequence conservation across species and recurrence of somatic mutations in cancer (Table 1).

Although motif analyses have been instrumental to predict altered TF binding, these methods are limited by the availability of high-quality PWMs and by the high false positive and false negative predictions rates of motif finding algorithms (Sewell and Fuxman Bass 2017; Weirauch et al. 2014; Zia and Moses 2012). Indeed, motif analyses can rarely distinguish between different members of a TF family, and often miss the TF that differentially binds to SNV alleles (Weirauch et al. 2014). Thus, SNVs in regulatory regions predicted to be functional based on hotspot and motif analyses, need to be experimentally tested to determine whether these mutations actually affect TF binding.

Experimental validation of differential TF binding between SNV alleles

Multiple complementary experimental methods can be used to determine TF binding including ChIP, electrophoretic mobility shift assays (EMSA), and enhanced yeast one-hybrid (eY1H) assays (Figure 1.3). ChIP has been successfully used to study differential TF binding between noncoding SNV alleles in vivo (Figure 1.3A). For example, several studies have identified mutations in the TERT promoter, such as G228A, that lead to the creation of de novo bind site for ETS factors (S. Horn et al. 2013; Huang et al. 2013). However, the identity of the specific ETS factor involved remained elusive until a recent study analyzing ChIP-seq data from the ENCODE Project (Dunham et al. 2012), identified GABPA as the TF that differentially binds and regulates TERT expression. In particular, GABPA was found to be bound to the TERT promoter in heterozygote cell lines harboring

the G228A mutation, specifically to the mutant allele, while other ETS factors did not show significant binding. Although ChIP is the method of choice to validate in vivo differential TF binding between alleles, this method requires a priori TF candidates as it can only test one TF at a time. Further, given that ChIP tests for in vivo TF binding, experiments need to be performed in cell lines harboring the mutations or using patient samples, which are frequently challenging to obtain.



Figure 1.3 Overview of assays to measure differential TF binding between noncoding SNV alleles. (A) ChIP against a candidate TF can be performed in cells that are heterozygous for the SNV. Sequencing of the amplified regions (or allele-specific qPCR) can determine relative TF binding between wild-type (wt) and mutant (mut) alleles. Alternatively, ChIP-seq data can be analyzed to detect biases in the number of sequencing reads between alleles. The figure shows an example of loss of TF binding caused by a mutation. (B) EMSA can be performed to determine differential TF binding to oligonucleotides containing wild-type (wt) or mutant (mut) SNV alleles by using nuclear extracts (NE) followed by super-shifts using antibodies against the candidate TF (α -TF), or by incubating with extracts overexpressing the TF. (C) eY1H assays can test the binding of >1,000 TFs to wild-type and mutant allele sequences. In this assay, each DNA sequence is cloned upstream the HIS3 and LacZ reporters and integrated into the yeast genome. Interactions are tested by mating with yeast strains expressing different TFs in an arrayed format system. Differential TF interactions (highlighted in red) can be determined by comparing screening results between alleles.

A recent study using enhanced yeast one-hybrid (eY1H) assays, a method that tests protein-DNA interactions in the milieu of the yeast nucleus, has increased the screening throughput for TF binding differences between SNV alleles by testing >1,000 TFs in parallel, without the need for antibodies or patient samples (Figure 1.3C) (Fuxman Bass et al. 2015). Although this study has focused on germline variants associated with different genetic diseases, the experimental eY1H pipeline can also be used to evaluate somatic SNVs in cancer. Given that ChIP, EMSA and eY1H assays measure physical DNA binding, rather than regulatory activity, interactions identified by these methods need to be tested in human cell lines to determine the SNV impact on gene regulation by using transient reporter assays, or endogenous gene expression measurements following TF knockdown/knockout.

Experimental validation of altered gene expression by SNVs

Driver mutations that affect regulatory regions are expected to affect the expression of a target gene. Functional validation assays such as those using luciferase reporters have been widely used to determine expression differences between noncoding SNV alleles (Figure 1.4A) (Denisova et al. 2015; Fuxman Bass et al. 2015; Huang et al. 2013; Rheinbay et al. 2017). In addition, reporter assays can be used to validate differential TF binding determined based on physical binding assays, by overexpressing or knocking down TF expression and measuring the impact on reporter activity driven by the wild-type or mutant regulatory sequences. Although useful for functional validation, reporter assays are generally low-throughput and cannot keep pace with the discovery of new mutations.



Figure 1.4 Functional assays to measure altered gene expression and phenotypic parameters induced by SNVs in regulatory regions. (A) Reporter assays can be used to determine differential expression induced by wild-type and mutant regulatory elements in transiently transfected cells. (B) In MPRAs wild-type and mutant alleles for hundreds/thousands of noncoding SNVs can be tested in parallel for changes in transcriptional activity. ~200 bp sequences containing the SNVs are cloned upstream of an inert ORF and associated with random barcodes. Cells are then transfected with the pooled library, ORF-specific mRNA is isolated, and barcode tags are counted using next-generation sequencing (NGS). By comparing the number of reads per allele in the mRNA and the plasmid populations, relative expression levels can be determined. (C) Functional validation and follow-up studies can be performed by determining differences in endogenous gene expression, proliferation, migration, and viability, among other assays, using cells engineered to carry the mutation.

Recent studies using massively parallel reporter assays (MPRAs), a highthroughput technology based on barcodes and next generation sequencing, have made progress in determining whether germline SNVs associated with genetic disorders affect transcriptional regulation (Figure 1.4B) (Melnikov et al. 2012; Tewhey et al. 2016; Ulirsch et al. 2016; Mogno, Kwasnieski, and Cohen 2013). In particular, differential transcriptional activity has been detected for hundreds of expression quantitative trait loci (eQTL) and disease-associated variants. While this method remains to be applied to cancer SNVs, it is

expected that MPRAs will constitute an essential tool for identifying functional noncoding somatic SNVs. Although powerful, MPRAs are not free of caveats. For instance, current oligonucleotide synthesis pipelines only allow for a maximum DNA fragment length of ~230 nucleotides. Thus, noncoding mutations are not usually tested within full length regulatory elements (that can be up to several kilobases), which may be hamper the ability of MPRAs to detect changes in gene expression. This limitation may be overcomed as pooled and arrayed oligonucleotide synthesis technologies are adapted to generate longer DNA sequences. Another limitation of MPRAs is that reporter activity is generally tested using episomal constructs, or randomly integrated lentiviral constructs, that do not reflect the endogenous genomic context where the noncoding mutations reside (Tewhey et al. 2016; Ulirsch et al. 2016). Thus, the functional effect of many SNVs on target gene expression may be over or underestimated. Downstream validation studies in the appropriate genomic context can be conducted by introducing the SNV in the endogenous locus using genome editing technologies such as the CRISPR/Cas9 system, zinc finger nucleases, or transcription activator-like effector nucleases (Figure 1.4C) (Claussnitzer et al. 2014; Elkon and Agami 2017). These studies, ultimately need to be followed-up using assays that demonstrate the biological significance of the SNVs in cancer by measuring different oncogenic properties such as invasion, proliferation, and viability (Figure 1.4C).

SNVs affecting distal regulatory elements

Compared to promoters, dissecting the functional effects of mutations in distal regulatory elements such as enhancers is a more complex task as it is not trivial to determine which of these elements are functional in different cells/conditions nor the identity of the target gene involved. This, and the fact that including distal elements in hotspot analyses increases the search space and reduces statistical power are the main reasons why most studies characterizing germline and somatic noncoding SNVs have focused on promoter regions (Rheinbay et al. 2017; Stenson et al. 2014).

Several technologies have been used to identify promoter-enhancer pairs interacting through chromatin loops. These methods, that involve crosslinking and ligation of spatially closed genomic regions, such as Hi-C (Lieberman-Aiden et al. 2009) and chromatin conformation capture by paired-end tag sequencing (ChIA-Pet) (G. Li et al. 2012), have been used to capture the potential regulatory effect of enhancer mutations. For example, a recent study found that a somatic SNV (C>T) four kilobases upstream of the transcriptional start site of the LMO1 oncogene generated a de novo binding site for the MYB TF in patients with T-cell acute lymphoblastic leukaemia (Yongsheng Li et al. 2017). A combination of ChIP-Seq of MYB, followed by ChIA-PET and luciferase assays revealed that this mutation induced the formation of an aberrant transcriptional enhancer complex leading to increased expression of the LMO1 oncogene. Thus, integration of chromatin interaction data can identify the gene targets of distal regulatory elements and determine how mutations in those elements affect looping interactions leading to changes in gene expression.

Noncoding SNVs affecting post-transcriptional regulation

Noncoding mutations not only affect transcriptional regulation but can also affect other biological processes such as mRNA stability, translation efficiency or splicing. Mutations in UTRs can affect mRNA stability and translation efficiency by altering interactions with RNA-binding proteins and miRNAs (Figure 1.1) (Khurana et al. 2016). For example, mutations in the 5'UTR of RB1 alter UTR conformation and mRNA stability in retinoblastoma (Kutchko et al. 2015), while mutations in the 5'UTR of BRAC1 in breast cancer patients reduce translation efficiency (Signori et al. 2001; Wang et al. 2007). In addition, mutations in the 3'UTR of BRCA1 were found to introduce a functional miRNA-103 target site in a breast cancer case leading to reduced BRAC1 levels (Brewster et al. 2012). As with SNVs in transcriptional regulatory regions, the functional impact of UTR mutations need to be tested in experimental assays. Low-throughput reporter assays have been used to quantify differences in mRNA levels by cloning the relevant UTR regions upstream or downstream of the coding region of GFP or luciferase. More recently, massively parallel functional annotation of sequences from 3' UTRs (fast-UTR) has been developed, which was used to discover 87 novel cis-regulatory elements and measure the effects of known gene variations in 3'UTRs (Zhao et al. 2014).

Mutations in the exon-intron boundaries, introns, and coding regions can affect splicing and lead to the upregulation oncogenic isoforms or the downregulation of tumor suppressor isoforms. Various cancer tumor suppressor genes such as TP53, ARID1A, PTEN, CHD1, MLL2, and PTCH1 were found to carry mutations in the exon-intron boundaries which led to intron retention (Jung et al. 2015; Supek et al. 2014). An intronic mutation in BRAF induces the expression of a splice variant that confers resistance to vemurafenib treatment in melanoma (Salton et al. 2015). These aberrant or cancer-specific isoforms are generally detected using short- and/or long-read mRNA sequencing, and are usually validated using mini-gene constructs carrying the different SNV alleles in low- or

high-throughput assay formats (Cavelier et al. 2015; Gaildrat et al. 2010; Yongsheng Li et al. 2017; Rosenberg et al. 2015).

Future perspectives

Recent studies have identified a handful of somatic SNVs in regulatory regions that affect TF binding and target gene expression. However, the number of functional noncoding SNVs associated with cancer is expected to be much higher given the low overlap between those reported in different studies, and given that noncoding SNVs seem to play an important role in disease based on the hundreds of functional noncoding SNVs identified in genome-wide association and genetic studies (Stenson et al. 2014). Advances in several areas will be needed to increase our ability to identify these driver mutations. First, larger numbers of tumor samples with available whole-genome sequence data are needed to increase statistical power in prediction algorithms. Second, more refined background models in hotspot analyses that take into account multiple covariates will help identify functional regulatory regions in cancer. Finally, improvements in motif analyses will be needed through the generation of PWMs for uncharacterized TFs and by identifying in silico parameters that can accurately predict differential TF binding between alleles.

Another source of underestimation of noncoding driver SNVs stems from the hotspot analysis itself as it assumes that driver mutations in a particular regulatory region should be present in multiple patients. Given the hundreds of thousands of regulatory elements in the human genome we may be far from having a sample size sufficiently large to detect most functional SNVs. An alternative approach would be to lower the stringency in the statistical pipelines and directly test thousands of "moderate-confidence" SNVs

using MPRAs to identify functional variants. Ultimately, a combination of computational and experimental methods along with new technical innovations will increase our ability to identify and characterize the mechanisms by which noncoding SNV drive cancer.

Funding

This work was supported by the National Institutes of Health to JFB (R00 GM114296 from the NIGMS) and to JS (5T32HL007501-34 from the NHLBI).

Author contributions

KG, SCP, JS, and JFB participated in the writing, reviewing, and critical analysis of the manuscript. JFB prepared the illustrations and coordinated the manuscript. All authors agreed and approved the final version.

Dissertation aims

The aims in this dissertation seek to develop novel algorithms and resources to aid in the analysis of transcriptional regulation in the context of cytokine expression and dysregulation of TFBS by SNVs. Together, these aims will show that we can predict and validate novel regulatory mechanisms for cytokines, determine the probabilities of creating and disrupting TFBS, and discover and validate cancer drivers in gene promoters. The three aims of the thesis are: *Aim 1. Determine the transcriptional regulation landscape of cytokines in mouse and human*

Aim 2. Predict genome-wide effects of single nucleotide variants in transcription factor binding

Aim 3. Discover and validate cancer driver single nucleotide variants in gene promoters

Chapter 2. Global landscape of mouse and human cytokine transcriptional regulation

Adapted from the following manuscript:

 Sebastian Carrasco Pro#, Alvaro Dafonte Imedio#, Clarissa Stephanie Santoso, Kok Ann Gan, Jared Allan Sewell, Melissa Martinez, Rebecca Sereda, Shivani Mehta, Juan Ignacio Fuxman Bass. Global landscape of mouse and human cytokine transcriptional regulation. Nucleic acid research 46 (18), 9321-9337.

co-first authors

Introduction

Cytokines comprise an array of polypeptides that are critical in the development of the immune system and in the regulation of immune and autoimmune responses (Griffith, Sokol, and Luster 2014). The published lists of human cytokines range from 132 to 261 genes depending on whether growth factors, hormones, or the receptors of cytokine genes are included (Wong et al. 2016; Al-Yahya et al. 2015; Kveler et al. 2018). Here, we focus on 133 cytokine genes, with a primary role in the immune system, shared by different publications.

Cytokine dysregulation is associated with myriad diseases including autoimmune disorders, susceptibility to infections, and cancer (Griffith, Sokol, and Luster 2014; Homey, Müller, and Zlotnik 2002; Netea et al. 2003; Neurath 2014; O'Shea, Ma, and Lipsky 2002). The expression of cytokine genes is primarily regulated at the transcriptional level through a combination of tissue-specific (TS) transcription factors (TFs) that control cytokine expression in different cell lineages, and pathogen- or stress-activated (PSA) TFs that respond to signaling pathways activated by pathogen-derived ligands or endogenous inflammatory mediators (Murphy and Reiner 2002; Medzhitov and Horng 2009). Although cytokine transcriptional regulation has been studied for more than three decades, including hallmark models of transcriptional regulation such as the IFNB1 enhanceosome (Thanos and Maniatis 1995), we currently lack a comprehensive view of the gene regulatory network (GRN) involved in controlling cytokine gene expression.

Several databases have been generated that annotate protein-DNA interactions (PDIs). InnateDB reports interactions between TFs and immune-related genes retrieved from different databases such as PubMed and IntAct, a subset of which have been manually curated (Breuer et al. 2013). TRRUST reports interactions involving immune and non-immune genes (Han et al. 2015), obtained by data mining and curating article abstracts from Pubmed. However, the overlap between these databases is generally low (20% overlap for cytokine genes), suggesting that they may be incomplete and/or may contain misannotated PDIs. This limits our understanding of the combinatorics involved in cytokine transcriptional regulation, especially in terms of the balance between TS and PSA TFs regulating each cytokine gene, the cooperativity and plasticity in cytokine regulation, and the relationship between TF connectivity and immune phenotype/disease.

Here, we mine through three decades of research to generate a comprehensive and userfriendly database, CytReg (http://cytreg.bu.edu), comprising 843 human and 647 mouse interactions between TF and cytokine genes. We analyze this cytokine GRN and integrate it with phenotypic and functional datasets to provide novel insights into the general principles governing cytokine regulation. In particular, we find a correlation between TF connectivity in the cytokine GRN and immune phenotype. We observe that the balance between PSA and TS TFs is shifted towards PSA TFs for interferons and pro-inflammatory cytokines and we provide a model for cooperative and plastic recruitment of cofactors to cytokine promoters. Using this cytokine GRN, we also provide a blueprint for further studies of cytokine misregulation in disease and identify novel TF-disease associations. Finally, we discuss biases and the completeness of the literature-derived cytokine GRN, and provide predictions for novel interactions which we validate using enhanced yeast onehybrid (eY1H) and reporter assays in human cells.

Materials and Methods

Generation of CytReg

To obtain a comprehensive list of physical and regulatory PDIs between TFs and cytokine genes we mined the XML files from ~26 million articles available in Medline on July 10th 2017, using NBCI's e-utilities python implementation, for studies mentioning a cytokine, a TF, and an experimental assay. Three broad categories of assays (chromatin immunoprecipitation, electrophoretic mobility shift assays, and functional assays), 1431 TFs, and 133 cytokines were considered. Alternative names for TFs and cytokines were obtained from the HUGO Gene Nomenclature Committee (www.genenames.org) and curated from the literature. Alternative spellings for names that include Greek letters or hyphens were also considered in the data mining.

The resulting 6,878 articles, together with 815 articles annotated in databases such as TRRUST (Han et al. 2015) and InnateDB (Breuer et al. 2013), were manually curated to determine whether experimental evidence for the PDIs was provided. A spreadsheet was

generated containing, for each mined interaction, the TF and cytokine HGNC names, the TF and cytokine names used in the paper, the type of assay, and the PubMed ID of the paper. Curation was performed based on the entire publication, rather than the abstract alone, because in some cases, PDIs reported in the abstract were based on indirect evidence and in other cases many PDIs identified were only reported in the body of the publication or in the figures. In addition to validating or rejecting mined PDIs, curators annotated the species, the functional activity (activating or repressing) if reported, and additional PDIs absent in the mined list but present in the body of the paper. Each PDI was curated by two independent researchers, and disagreements were resolved by a third senior curator. The resulting database contains 1,552 PDIs (843 in human, 647 in mouse, and 62 from other species) for which we annotated the assay used and the regulatory activity identified. To visualize this complex cytokine GRN we developed CytReg (https://cytreg.bu.edu), a web tool where PDIs can be browsed by species, TFs, cytokines, assay types, and TF expression patterns across different cell-types. In addition, links are provided to Uniprot entries (http://www.uniprot.org) for cytokine and TF genes, and to PubMed articles for the PDIs.

Determination of the level of evidence for PDIs

We classified PDIs as high or low evidence of being direct regulatory interactions. PDIs detected by a functional assay (e.g., reporter assays and TF knockdown) and an assay measuring direct binding (e.g., chromatin immunoprecipitation and in vitro binding assays) were classified as high evidence. PDIs detected by only one type of assay were classified as low evidence.

Determination of the relationship between TF connectivity and gene expression

The median transcript per million (TPM) expression levels in 20 immune cell-types for TFs with different connectivity in the human cytokine GRN was determined based on expression data published by the Blueprint Epigenome Consortium (Stunnenberg et al. 2016) (http://dcc.blueprint-epigenome.eu). In addition, an expression enrichment score in immune tissues compared to non-immune tissues was determined based on data from 32 tissues from the Expression Atlas (https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2836). Briefly, a pseudocount of 1 was added to all the expression data to reduce the noise from low abundant transcripts. Then, the expression of a TF in a tissue was divided by the average expression of the TF across the 32 tissues to obtain an expression enrichment score. Finally, the average enrichment score per TF was determined for the five immune tissues (lymph node, bone marrow, spleen, tonsils, and appendix) and for the remaining 27 nonimmune tissues in the dataset.

Associations between TFs and immune phenotypes and diseases

The association between TFs and immune phenotypes was determined based on phenotypes in knockout mice reported by the Mouse Genome Informatics database (www.informatics.jax.org) as of January 12th 2018. Thirty different terms including different immune cells, antibody isotypes, cytokines, inflammation, and immune tissues were used to determine whether a reported phenotype should be classified as immuneassociated.

Association between TFs and immune disorders (including autoimmune diseases and susceptibility to infections) was obtained from the Human Gene Mutation Database 2013 release and from genome-wide association studies (GWAS) downloaded on July 27th 2017 from the NHGRI-EBI Catalog (MacArthur et al. 2017; Stenson et al. 2014).

TF enrichment in PDIs with cytokines expressed in different immune cell types

For each TF, we compared the proportion of cytokine targets corresponding to cytokines expressed in a specific immune cell type, to the proportion of the remaining cytokine targets. A proportion comparison test was used to determine a p-value and a Benjamini-Hochberg adjusted p-value to account for multiple hypothesis testing.

Pathogen/stress-activated and tissue specific TFs

PSA TFs were determined from the literature based on their ability to be activated or responsive to signaling pathways triggered by pathogen-associated molecular patterns and/or stress signals (e.g., oxidative stress, heat shock, and danger-associated molecular patterns). Tissue specific TFs were determined by calculating a tissue-specificity score (TSPS) based on expression data from 34 different tissues and cells as previously described (Ravasi et al. 2010):

$$TSPS = \sum p_i . \log_2(\frac{p_i}{p})$$

where pi corresponds to the ratio between the expression level in a tissue and the sum of the expression levels across all 34 tissues; and p corresponds to the expected ratio under the assumption of equal expression across all tissues. TFs were considered tissue-specific (TS) if their TSPS ≥ 0.7 , a threshold selected based on the bimodal distribution of TSPS across all TFs. TFs for which a TSPS could not be calculated because of unavailable expression data, were excluded from the analysis.

Determination of TF inflammatory scores

For each TF, an inflammatory score (IS) was determined as the difference between the percentage of PDIs with canonical pro-inflammatory cytokines (IL1A, IL1B, IL12A, IL12B, IL18, TNF, IFNG, CSF2, CXCL8, and IL6), and the percentage of PDIs with antiinflammatory cytokines (IL10, IL11, IL13, IL19, IL1RN, IL24, IL37, IL4, IL5, CXCL17, TGFB1, TGFB2, and TGFB3). For TFs with IS \geq 0.5 or IS \leq -0.5 we determined the percentage that have a pro- or anti-inflammatory role, or a role in differentiation based of phenotypes in knockout mice (www.informatics.jax.org).

TF-disease association

For each disease (asthma, systemic lupus erythematosus, inflammatory bowel disease, type 2 diabetes, rheumatoid arthritis, tuberculosis infection, and cytomegalovirus infection) the Expression Atlas (www.ebi.ac.uk) was searched for cytokines upregulated in the disease state, using a cut-off of 2-fold induction. TFs enriched in regulating the upregulated cytokines were determined from the human cytokine GRN using the Fisher's exact test. Multiple hypothesis testing was corrected by calculating the Benjamini-Hochberg adjusted p-value and using an FDR threshold of 0.1. The resulting TF-disease associations were plotted using a Circos plot (http://mkweb.bcgsc.ca/tableviewer/).

TF and cytokine association with autoimmune diseases

TFs and cytokines associated with different autoimmune diseases were obtained from the Human Gene Mutation Database 2013 release, and from GWAS downloaded on July 27th 2017 from the NHGRI-EBI Catalog (MacArthur et al. 2017; Stenson et al. 2014). The union of gene-disease associations between both databases was considered. Crohn's disease and ulcerative colitis were grouped with inflammatory bowel disease. This list includes coding and noncoding variants, and thus variants that affect protein function or expression levels. Of note, this list of gene-disease associations is not comprehensive as it only includes associations identified in genetic studies (i.e., does not consider environmental or epistatic factors that affect cytokine expression). Significance for enrichment of shared autoimmune diseases between interacting TFs and cytokines was determined by comparing to 1,000 randomized versions of the human cytokine GRN. Network randomization was performed by edge switching as previously described (Martinez et al. 2008).

TF-drug associations

TF-drug associations and information regarding drug function were obtained from Drugbank (Wishart et al. 2018). Agonists and activators were grouped as agonists, antagonist and inhibitors were grouped as antagonists. For each cytokine, the number of TFs targetable by agonists or antagonists was determined.

Prediction of novel PDIs in the human cytokine GRN

To predict novel PDIs in the human cytokine GRN, for each TF, SEEK (Q. Zhu et al. 2015) was used to search for the top 100 genes co-expressed with the known cytokine targets of the selected TF across more than 5,000 expression profiling datasets. Then, for each cytokine within those 100 genes, the presence of binding sites for the selected TF in the cytokine promoter (2kb upstream of the transcription start site) was determined using

the Scan DNA sequence tool in CIS-BP (http://cisbp.ccbr.utoronto.ca/), the PWM-Logodds algorithm, and a stringent threshold of ten (Weirauch et al. 2014). Enrichment for human PDI predictions reported in mouse was determined by calculating an odds ratio and statistical significance was calculated using the Chi-square test. The 1,066 predicted interactions were classified according to confidence: high (two or more TF binding sites and evidence of interaction in the mouse cytokine GRN), medium (two or more TF binding sites but absent from the mouse cytokine GRN, or less than two binding sites but presence in the mouse cytokine GRN), and low (one binding site and absent from the mouse cytokine GRN).

Enhanced yeast one-hybrid (eY1H) assays

eY1H assays were used to detect interactions between TFs and cytokine gene promoters (Reece-Hoyes, Barutcu, et al. 2011; Reece-Hoyes, Diallo, et al. 2011). This method involves two components: a 'DNA-bait' such as cytokine gene promoter, and a 'TF-prey'. The DNA-bait is cloned upstream of two reporter genes (LacZ and HIS3) and both constructs are integrated into the yeast genome (Fuxman Bass, Reece-Hoyes, and Walhout 2016a; 2016b). The DNA-bait strains generated are then mated with yeast strains expressing TFs fused to the yeast Gal4 activation domain (AD), and if the TF binds the regulatory region, the AD moiety activates the reporter genes. Reporter gene activity is measured by the conversion of colorless X-gal to a blue compound, and by the ability of the yeast to grow on media lacking histidine and to overcome the addition of 3-aminotriazole (3AT), a competitive inhibitor of the His3 enzyme. Each interaction was tested in quadruplicate. Yeast DNA-baits corresponding to promoter regions (2 kb upstream of the transcription start site) of cytokine genes were generated as previously described (Fuxman Bass, Reece-Hoyes, and Walhout 2016b; Fuxman Bass et al. 2015). The promoter regions of CXCL10, CXCL8, CXCL3, CCL4, and CCL20 were screened for REL binding, while promoter regions for IL17A, IL17F, and IL26 were screened for RORC binding. To identify TFs that interact with the promoters of CCL27 and CCL4L2, the CCL27 and CCL4L2 DNA-bait strains were screened against an array of 1,086 human TFs (Fuxman Bass et al. 2015).

Motif analysis

Binding of REL, RORC, RBPJ, TFAP2A/B, PPARG, ATF3, EBF1, ZIC1/3, GCM1, and WT1 were predicted using CIS-BP via the Scan DNA sequence tool, using the PWM-LogOdds method and a stringent threshold of ten (Weirauch et al. 2014). Motif analyses were performed on the same 2 kb regions upstream of the transcription start sites used to perform the eY1H assays.

Transient transfections and luciferase assays

HEK293T cells were plated in 96-well opaque plates (~1 x 104 cells/well) 24 hours prior to transfection in 100 μ l DMEM + 10% FBS + 1% Antibiotic-Antimycotic 100X. DNA-bait luciferase reporter clones were generated by cloning the cytokine promoter regions upstream of the firefly luciferase into a Gateway compatible vector generated from pGL4.23[luc2/minP] (Fuxman Bass et al. 2015). TF-prey clones were generated by Gateway cloning the TF into a vector derived from pEZY3 (Addgene) to generate fusions with ten copies of the VP16 activation domain (TF-pEZY3-VP160). Cells were transfected with Lipofectamine 3000 (Invitrogen) according to the manufacturer's protocol using 20 ng of the DNA-bait luciferase reporter vector, 80 ng of the TF-pEZY3-VP160 vector, and 10 ng of renilla luciferase control vector. The empty pEZY3-VP160 vector co-transfected with the recombinant firefly luciferase plasmid was used as a negative control. 48 hours after transfection, firefly and renilla luciferase activities were measured using the Dual-Glo Luciferase Assay System (Promega) according to the manufacturer's protocol. Non-transfected cells were used to subtract background luciferase activities, and then firefly luciferase activity were normalized to renilla luciferase activity.

Code availability

The code used for the data mining in Medline is available at https://github.com/fuxmanlab/cytreg.

Statistical analyzes

Statistical analyzes were performed using GraphPad Prism Version 7.01, Excel 2016, or VassarStats (http://vassarstats.net). All tests performed were two-tailed tests.

Software used to generate the figures

Box, bar, histogram, and correlation plots were generated using GraphPad Prism Version 7.01. Heatmaps were generated using matrix2png (https://matrix2png.msl.ubc.ca/). Networks were generated using Cytoscape Version 3.2.1 (http://www.cytoscape.org/).

Results

Generation of CytReg

To obtain a comprehensive cytokine GRN, we systematically mined ~26 million articles in Medline for studies mentioning at least one of 133 cytokines, one of 1,431 TFs, and an experimental assay (Figure 2.1A). The resulting 6,878 articles, and 815 additional articles referenced in TRRUST (Han et al. 2015) and InnateDB (Breuer et al. 2013), were then manually curated to determine whether experimental evidence for the physical and regulatory PDIs was provided. This resulted in a list of 1,552 PDIs (843 in human, 647 in mouse, and 62 in other species), for which we annotated the assay used and the regulatory activity identified (Figure 2.1A). To visualize this GRN we developed a database, CytReg (https://cytreg.bu.edu), where users can browse PDIs by species, TF, cytokine, assay type, and TF expression patterns (Figure 2.1B). Links are provided to Uniprot entries for TFs and cytokines, and to PubMed articles reporting the PDIs (Figure 2.1C). Finally, the selected PDIs can be visualized as networks showing the TFs, cytokines, and the types of interactions (activation, repression, or bifunctional) (Figure 2.1D).



Figure 2.1 Differentially Generation of CytReg. (A) Pipeline used for the text mining and article curation to determine literature-based PDIs between TFs and cytokine genes. (B) Search page of CytReg where PDIs can be browsed by TF, cytokine, species, assay type, and TF expression levels (mRNA and protein) in different immune cells. (C) Results page indicating the interacting cytokines and TFs, the types of assays used to determine the PDIs, whether the interaction is activating or repressing, and the Pubmed IDs of the publications referencing the PDIs. Links are provided to UniProt entries for cytokines and TFs, and to Pubmed for the references. The interactions can be downloaded as a CSV file or visualized as a network graph. (D) Network visualization of the selected PDIs. Nodes represent cytokines and TFs, edges represent the type of interaction (activating, repressing, bifunctional, or physical). Nodes can be moved to re-arrange the network. (E) Overlap of PDIs in CytReg and those annotated in InnateDB and TRRUST. (F) Overlap between mouse and human cytokine GRNs. (G) Fraction of PDIs with high evidence of

direct regulatory activity (by a functional assay and an in vitro or in vivo binding assay) or low evidence (by one type of assay).

CytReg contains an additional 371 human and 264 mouse PDIs compared to TRRUST and InnateDB (Figure 2.1E). We also removed 243 PDIs annotated in TRRUST and InnateDB when: a) the article did not provide direct experimental evidence for the PDI, b) the TF interacted with the regulatory region of a cytokine receptor rather than that of a cytokine, or c) the cytokine regulated the activation pathway of a TF rather than the TF regulating a cytokine. Altogether, CytReg greatly expands the PDIs annotated in other databases and removes misannotated PDIs.

Although multiple PDIs are shared between human and mouse, 69% of human and 60% of mouse PDIs are species-specific (Figure 2.1F). This low overlap is not likely related to a lack of confidence in the interactions because a similar proportion of interactions found in one or both species were classified as high confidence based on evidence from functional (e.g., reporters assays and TF knockdowns experiments) and in vivo or in vitro binding assays (chromatin immunoprecipitation -ChIP- and electrophoretic mobility shift assays –EMSAs, respectively) (Figure 2.1G). More likely, this low overlap is related to literature bias and incompleteness of the GRN, or to different modes of regulation between mouse and human as has been previously reported (Schmidt et al. 2010). Indeed, we found that PDIs reported early on in one species were more frequently detected in the other species than PDIs reported more recently. For example, 71% of mouse PDIs reported on or before the year 2000 are also reported in human. This suggests that

literature biases may play an important role in the differences in annotated PDIs between species.

Most interactions were reported by at least two of three types of experimental assays: binding assays (e.g., EMSA and pull down assays), ChIP, and functional assays (Figure 2.2A and B). Human PDIs detected by all three types of assays were more frequently also detected in mouse (and viceversa) compared to PDIs detected by one or two types of assays (Figure S1A and B). The types of assays used to determine PDIs has changed over time, with papers in the 1990s focusing on binding and functional assays while papers in the 2010s focusing on ChIP and functional assays, reflecting the increased awareness of the importance of chromatin context in gene regulation (Figure 2.2C and D).



Figure 2.2 Distribution of experimental methods used to determine PDIs. (A, B) Number of PDIs in the human (A) and mouse (B) cytokine GRNs per assay type and the number of PDIs annotated in the mouse and human GRNs, respectively. Filled circles – PDIs involving the assay. (C, D) Number of PDIs in the human (C) and mouse (D) cytokine GRNs per assay type over time.

Association between TF connectivity and immune phenotype

As observed in other GRNs, a few TFs and cytokines are responsible for most PDIs in the cytokine GRN (Figure 2.3A and B, and Figure 2.4A and B) (Luscombe et al. 2004; Deplancke et al. 2006). For example, 12% of the TFs are responsible for more than 50% of the PDIs, including different subunits of NF-κB that when combined represent 16% of the PDIs in the human cytokine GRN (Figure 2.3A). Similarly, 8% of the cytokines, including the highly studied CXCL8, IL6, and TNF, are involved in more than 50% of the PDIs (Figure 2.3B). We obtained similar distributions for the mouse cytokine GRN (Figure 2.4A and B). These lopsided distributions in the number of PDIs can be explained by a more central role of some TFs and cytokines in the GRN, but also by research biases as discussed below.



Figure 2.3 Relationship between TF connectivity and phenotype in the human cytokine GRN. (A) Number of cytokine targets per TF (TF degree) in the human cytokine GRN ordered by TF degree rank. (B) Number of interacting TFs per cytokine (cytokine degree) in the human cytokine GRN ordered by cytokine degree rank. (C) Median expression as transcripts per million (TPM) across human immune cells obtained from the Blueprint Epigenome Consortium for TFs displaying different numbers of cytokine targets. (D) Expression enrichment in human immune tissues versus non-immune tissues for TFs with varying numbers of cytokine targets. Each box spans from the first to the third quartile, the horizontal lines inside the boxes indicate the median value and the whiskers indicate minimum and maximum values. Statistical significance determined using two-tailed Wilcoxon matched-pair ranked sign test. (E) Fraction of TFs in the human cytokine GRN with annotated immune phenotypes when knocked out in mice (MGI), or associated with immune disorders in the Human Gene Mutation Database (HGMD) or in genome-wide association studies (GWAS) based on the number of cytokine targets.

We found that TFs that interact with multiple cytokine genes show higher expression levels in immune cells (Figure 2.3C) and higher expression enrichment in immune tissues (such as the spleen, bone marrow, and lymph nodes) compared to TFs that interact with only a few or no cytokine genes (Figure 2.3D). Further, highly connected TFs are frequently PSA TFs (e.g., 71% of TFs with ten or more cytokine targets are PSA compared to 9% for TFs with one cytokine target) consistent with their function in immune responses. More importantly, highly connected TFs are more frequently associated with immune phenotypes in knockout mouse studies, and with immune disorders as reported in the human gene mutation database (HGMD) and in GWAS compared to low connected TFs (Figure 2.3E and Figure 2.4C) (MacArthur et al. 2017; Stenson et al. 2014; Eppig et al. 2017). For example, the highly connected TF IRF5 is associated with multiple autoimmune diseases, including multiple sclerosis and systemic lupus erythematosus (SLE), and leads to low type-I interferon, TNF and IL6 production in knockout mice (MacArthur et al. 2017; Stenson et al. 2014; Eppig et al. 2017). Conversely, the low connected TFs HMGA2, NDS2, and HMBOX1, to our knowledge, have not yet been associated with immune phenotypes or diseases. Overall, these observations highlight the association between TF connectivity and disease, consistent with previous findings in a developmental GRN (Fuxman Bass et al. 2015).



Figure 2.4 Relationship between TF connectivity and phenotype in the mouse cytokine GRN. (A) Number of cytokine targets per TF (TF degree) in the mouse cytokine GRN ordered by TF degree rank. (B) Number of interacting TFs per cytokine (cytokine degree) in the mouse cytokine GRN ordered by cytokine degree rank. (C) Fraction of TFs in the mouse cytokine GRN with annotated immune phenotypes when knocked out in mice (MGI), or associated with immune disorders in the Human Gene Mutation Database (HGMD) or in genome-wide association studies (GWAS) based on the number of cytokine targets.

Cytokine regulation by different types of TFs

Different cell types express different sets of cytokines in response to pathogen- or cell-mediated cues. For each immune cell type, we determined the TFs enriched in binding/regulating the cytokines expressed in the given cell type. As expected, several master regulator TFs are enriched, including TBX21 (T-bet) in Th1 cells, GATA3 and STAT6 in Th2 cells, RORC in Th17 cells, and SPI1 (PU.1) and CEBPA in monocytes. Additionally, several PSA TFs, such as RELA/NFKB1, are enriched in Th1 cells, monocytes, myeloid dendritic cells, eosinophils, and neutrophils, consistent with these cells producing pro-inflammatory cytokines upon activation; while IRF1/3/5/7 are enriched in B cells and plasmacytoid dendritic cells, producers of type-I interferons in response to viral pathogens.

Highly connected TFs in the cytokine GRN usually belong to the Ig-like plexins transcription factor (IPT/TIG/p53 - including NF-κB and NF-AT TFs), activator protein 1 (AP-1), interferon regulatory factor (IRF), and signal transducer and activator of transcription (STAT) families, which are known to play prominent roles in immune cell differentiation and immune responses (Holloway, Rao, and Shannon 2002; Taniguchi et al. 2001; Rao, Luo, and Hogan 1997; Peltz 1997). These TF families are highly enriched in the cytokine GRN compared to the GRN reported in TRRUST (Han et al. 2015), a literature-derived network not constrained to cytokine genes (Figure 2.5A and Figure 2.6A and B). Furthermore, most PSA TFs are enriched in the cytokine GRN compared to the GRN reported in TRRUST, consistent with many cytokine genes being upregulated in response to pathogens or stress conditions (Figure 2.5B).



Figure 2.5 Cytokine regulation by different types of TFs. (A, B) Correlation between the percentage of PDIs involving a TF in the human cytokine GRN versus a global human GRN

annotated in TRRUST, for different TF families (A) or for pathogen- or stress-activated (PSA) TFs (B). (C) Average fraction of PSA and tissue-specific (TS) TFs for cytokines expressed in different cell types. (D) Fraction of PSA and TS TFs for different classes of cytokines. Correlation determined by Pearson correlation coefficient. (E) Inflammatory score (IS) for each TF based on the fraction of PDIs with pro- and anti-inflammatory cytokines. (F) Percentage of TFs with pro-inflammatory, anti-inflammatory, and differentiation or other functions based on mouse knockout phenotypes. p = 0.009 by Fisher's exact test.

Cytokines are expressed in a highly tissue- and condition-specific manner. This is achieved by a specific combination of receptors and signaling pathways present in each cell type, and through the cooperation between PSA and TS TFs (Holloway, Rao, and Shannon 2002). To study the role of PSA and TS TFs in cytokine regulation, for each cytokine we determined the fraction of TFs that respond to pathogen/stress signals (e.g., NF-κB, AP-1 and IRFs) and the fraction of TS TFs determined based on each TF's gene expression variability across tissues. Our analysis revealed that cytokines expressed in plasmacytoid dendritic cells, M1 macrophages, Th1 cells, and myeloid dendritic cells are primarily regulated by PSA TFs, whereas cytokines expressed NK cells, basophils, mast cells, Th2 cells, Th17 cells, and eosinophils are also regulated by several TS TFs (Figure 2.5C). This is consistent with reports of the former cell types expressing multiple canonical pro-inflammatory cytokines and/or interferons, which are induced by pathogen-associated molecular patterns or danger signals from inflammatory microenvironments. Indeed, further analysis revealed that interferons and pro-inflammatory cytokines are regulated by broadly-expressed PSA TFs, whereas anti-inflammatory cytokines are regulated by both PSA and TS TFs (Figure 2.5D).



Figure 2.6 TF families present in the mouse and human cytokine GRNs. (A) Correlation between the percentage of PDIs involving a TF in the mouse cytokine GRN versus a global mouse GRN annotated in TRRUST. (B) Distribution of TF families in the human and mouse cytokine GRNs compared to those annotated in the TRRUST database.

Different TFs have predominantly pro- or anti-inflammatory functions. Thus, for each TF, we determined an inflammatory score (IS) based on the preference of binding to pro- versus anti-inflammatory cytokine gene targets (Figure 2.5E). TFs with an IS>0.5 more frequently had a pro-inflammatory function, while TFs with IS<-0.5 more frequently had a pro-inflammatory function, while TFs with IS<-0.5 more frequently had an anti-inflammatory function based on knockout mouse phenotypes (Figure 2.5F, p = 0.009 by Fisher's exact test). Although the dysregulation of other targets is likely involved, these analyses suggest that the cytokine targets of a TF can be important drivers of immune phenotypes.

GRN integration with TF-cofactor interactions

Different cell types express different sets of cytokines in response to pathogen- or cell-mediated cues. For each immune TFs regulate gene expression by recruiting co-activators and co-repressors that interact with the transcriptional machinery or mediator complex, or that covalently modify histones, TFs, or methylate DNA (Thomas and Chiang

47

2006; Rolland et al. 2014). Based on literature-derived protein-protein interactions reported in Lit-BM-13 (Rolland et al. 2014), we found that the TFs that bind/regulate cytokine genes interact with numerous cofactors, including multiple co-activators such as EP300, CREBBP, and nuclear co-activators 1-3 and 6 (Figure 2.7A). This is not surprising given that ~80% of the regulatory PDIs in CytReg are activating and involve potent transcriptional activators such NF- κ B and AP-1. Nevertheless, several activating TFs also interact with co-repressors which can inhibit TF function until triggered by signaling pathways (T. D. Gilmore and Herscovitch 2006).



Figure 2.7 Cooperativity and plasticity in cytokine regulation. (A) Protein-protein interaction network from Lit-BM-13 between cofactors and TFs in the human cytokine GRN. Ellipses – TFs, diamonds – cofactors. Node size indicates the number of cytokine targets (for TFs) in the cytokine GRN, and the number of protein-protein interactions with TFs (for cofactors). Only cofactors with five or more protein-protein interactions are shown. (B, C) Number of TFs (shades of grey) interacting with each human cytokine gene that interact with the different cofactors (B) or the different domains of EP300/CREBBP (C). (D, E) Fraction of cofactor (D) or EP300/CREBBP domain (E) protein-protein interactions are shown. Co-activators are shown in red font, co-repressors in blue font, and bifunctional cofactors in purple font.

In general, each cofactor interacts with multiple TFs that bind/regulate each cytokine gene (Figure 2.7B) (Rolland et al. 2014). This may be associated with TF

cooperativity to recruit cofactors to regulatory regions as has been reported for the cooperative recruitment of EP300 by RELA, IRFs, JUN, and HMGA1 to the IFNB1 enhanceosome (Thanos and Maniatis 1995). Alternatively, cofactor binding to multiple TFs may also be associated with regulatory plasticity by which cofactors can be recruited by different sets of TFs to modulate cytokine gene expression in different cell types or conditions. To evaluate these possibilities, we focused on the histone acetyltansferases EP300/CREBBP, which play key roles in immune regulation and differentiation, and whose protein-protein interactions with TFs have been mapped to their different domains (Freedman et al. 2002; Hottiger and Nabel 2000). We found that, for cytokines for which multiple PDIs have been determined, the set of TFs that bind/regulate that cytokine gene collectively interact with multiple domains of EP300/CREBBP (Figure 2.7C). This may lead to a cooperative recruitment of EP300/CREBBP to regulatory regions, as has been observed for the IFNB1, TNF, and IL6 genes (Thanos and Maniatis 1995; Berghe et al. 1999; Tsytsykova and Goldfeld 2002). This is also consistent with the observation that, even for cytokines with multiple annotated PDIs, the mutation of a single TF binding site or the inhibition of a single TF can lead to a dramatic effect on gene expression (Tsai et al. 2000; Melnikov et al. 2012). Interestingly, for each cytokine, several TFs can also interact with the same domain of EP300/CREBBP (Figure 2.7C). Although this may contribute to a cooperative recruitment of EP300/CREBBP, it may also increase regulatory plasticity in different cell types and/or under different stimuli by allowing different TF combinations to induce cytokine expression. For example, TNF induction by LPS, calcium, or viruses all

lead to EP300/CREBBP recruitment to the TNF enhanceosome, however, through different sets of TFs (Tsytsykova and Goldfeld 2002).

Some cofactors such as MAPK8, BRCA1, MDM2 and COPS5 preferentially interact with PSA TFs, consistent with their reported function in inflammation and stress responses, and associated immune phenotype in knockout mice (Figure 2.7D) (Eppig et al. 2017). Other cofactors such as NCOR1/2, NCOA1/2/3/6, RB1, NRIP1, SRC and MED1 interact primarily with TS TFs such as nuclear hormone receptors (Rolland et al. 2014) (25416956). Interestingly, different domains of EP300/CREBBP interact preferentially with PSA or TS TFs: for example, CH1, KIX and Q/I interact mostly with PSA TFs, whereas RID and CH3 interact mostly with TS TFs (Figure 2.7E). Altogether, this suggests that PSA and TS TFs cooperate in recruiting EP300/CREBBP through different domains to induce cytokine expression under the right stimuli and in the appropriate cell types. In addition, functional redundancy between different PSA TFs may allow for the activation of cytokine expression under different conditions. For example, the PSA TFs HIF1A and NF-kB, both of which interact with the CH1 domain of EP300/CREBBP, can independently induce CXCL8 expression (Kim et al. 2006). Overall, these findings are consistent with a model that contains aspects of both the enhanceosome (i.e., cooperative TF binding is required for regulatory activity) and billboard (i.e., TFs independently regulate gene expression) models of gene regulation, where only certain combinations of TFs present in particular cells or conditions can induce gene expression (Spitz and Furlong 2012). Each cytokine, depending on their regulatory flexibility, may be closer to one model or the other
The cytokine GRN as a blueprint to study disease

Cytokine expression is widely dysregulated in immune disorders and infection. This is driven by the activation of multiple signaling pathways that result in TF activation leading to the concomitant regulation of target cytokines. To explore these TF-disease relationships, we leveraged the human cytokine GRN to identify TFs enriched in regulating the cytokines overexpressed in different autoimmune diseases, Mycobacterium tuberculosis infection, and cytomegalovirus infection. We identified 46 TF-disease associations between 25 TFs and seven diseases, many of which are known (Figure 2.8A). For example, different subunits of NF- κ B were associated with all the diseases evaluated, consistent with the ubiquitous role of NF-κB in inflammation (T. D. Gilmore and Herscovitch 2006). Other TF-disease associations identified were more specific. For instance, IRFs and ATF2 (in addition to NF- κ B) were associated with cytomegalovirus infection which is consistent with these TFs being activated by viral pathogens through pattern recognition receptors (Navarro et al. 1998; Browne and Shenk 2003; Le et al. 2008). STAT1 and STAT2 were also associated with cytomegalovirus infection, in this case, likely through the activation of signaling pathways driven by the autocrine/paracrine secretion of type-I and type-II interferons induced by IRF and NF-kB activation. In addition, we identified an association between STAT6 and SLE, consistent with STAT6 deficiency being associated with a better prognosis in mouse models of SLE (Singh et al. 2003; Jacob et al. 2003), and with STAT6 polymorphisms being associated with SLE in humans (Yu et al. 2010). Further, we found known associations between KLF6, NR3C1,

XBP1, and HSF1 with inflammatory bowel disease further validating our analyses (Tanaka et al. 2007; Kaser et al. 2008; Goodman et al. 2016; Brattsand and Linden 1996).



Figure 2.8 Association of the cytokine GRN with human diseases. (A) Circos plot connecting diseases with TFs based on enrichment of the TFs in regulating cytokines upregulated in the indicated disease. Ribbon width is proportional to the percentage of cytokines upregulated in the indicated disease that are regulated by the indicated TF. (B) GRN connecting interacting TFs and human cytokine genes associated with autoimmune disorders. Edges connect interacting cytokine-TF pairs. Edge color indicates that the interacting cytokine and TF are associated with the same disease based on HGMD and GWAS. (C) The human cytokine GRN was randomized 1,000 times by edge switching and the number of TF-cytokine-disease sets in each randomized network was calculated. The number under the histogram peak indicates the average overlap in the randomized networks. The red arrow indicates the observed overlap in the real network. Statistical significance determined based on z-score calculation. (D) GRN connecting cytokines with TFs that can be targeted by approved drugs. Blue, red, and yellow ovals indicate TFs targetable by agonists,

antagonists, or both, respectively. Oval size corresponds to the number of approved drugs targeting a TF. Rectangles indicate cytokine genes. Rectangle size is proportional to the number of druggable TFs per cytokine.

More importantly, we found previously uncharacterized TF-disease associations. For example, we identified an association BCL6 and SLE (Figure 2.8A). A mouse model of SLE (Def6 and SWAP70 double knockout) showed increased BCL6 protein expression (Yi et al. 2017). However, the role of BCL6 in cytokine dysregulation in SLE has not been established. Our analyses, suggest that the increased BCL6 levels may be associated with increased levels of CCL1/2/7/8/13 observed in SLE. We also identified a previously uncharacterized association between ETS2 and cytokine upregulation in M. tuberculosis infected macrophages (Figure 2.8A). ETS2 is an activator that is upregulated 5.7 fold (p =3.6 x 10-7) in macrophages infected with M. tuberculosis for 48 hs (E-MEXP-3521). This increased ETS2 expression, together with ETS2 activation through the MAPK pathway (McCarthy et al. 1997), may contribute to cytokine upregulation in M. tuberculosis infection. Interestingly, the association between ETS2 and M. tuberculosis infection would not have been predicted only based on PDIs from InnateDB and TRRUST. Further, using PDIs from these previous databases we only predicted 21 TF-disease associations, most of them included within the 46 associations predicted based on CytReg, while missing multiple known associations such as those between NF-kB subunits and autoimmune diseases (Figure 2.9). Overall, our analyses predicted novel TF-disease associations which are consistent with known TF functions. Further studies are required to determine the mechanisms of action of BCL6 in SLE and ETS2 in M. tuberculosis infections.



Figure 2.9 Gene expression of Berry et al. (2010) 86-gene signature in TB and LTBI subjects from a South Indian population. Circos plot connecting diseases with TFs based on enrichment of the TFs in regulating cytokines upregulated in the indicated disease. Ribbon width is proportional to the percentage of cytokines upregulated in the indicated disease that are regulated by the indicated TF. The left plot is based on PDIs from the union of TRRUST and InnateDB, the right plot is based on PDIs from CytReg (as in Figure 2.8A).

Mutations in multiple TFs have been associated with immune disorders such as autoimmune diseases (MacArthur et al. 2017; Stenson et al. 2014). The role of TFs in autoimmunity is likely related to the dysregulation of immune genes, in particular cytokines, as they play a central role in immune responses and tolerance (Neurath 2014; O'Shea, Ma, and Lipsky 2002). Indeed, mutations in many cytokine genes have been associated with autoimmunity (MacArthur et al. 2017; Stenson et al. 2014). We considered the cytokines and TFs that have been associated with autoimmune diseases in GWAS and HGMD, and found that many TF-cytokine gene pairs that interact in the cytokine GRN have been associated with the same autoimmune disease (Figure 2.8B). For example, we found multiple TF-cytokine pairs associated with inflammatory bowel disease, rheumatoid

arthritis, atopic dermatitis/psoriasis, and SLE (Figure 2.8B). Overall, the number of TFcytokine pairs associated with the same autoimmune disease is higher than that determined in randomized networks derived from the human cytokine GRN (Figure 2.8C). These TFcytokine pairs identified may constitute different regulatory axes by which TFs lead to the disease. For example, AHR activation is protective in inflammatory bowel disease, partly due to increased IL10 expression (Goettel et al. 2016). Interestingly, the association between AHR, IL10, and inflammatory bowel disease, together with 19 other TF-cytokinedisease associations was absent in predictions based on PDIs from the union of TRRUST and InnateDB. Altogether, the network depicted in Figure 5B constitutes a blueprint to study other regulatory axes in autoimmunity.

Targeting cytokine activity is a widely used therapeutic approach for multiple autoimmune and inflammatory diseases (Wishart et al. 2018; Chan and Carter 2010). However, only ~15% of cytokines can currently be directly targeted with approved small molecules or specific antibodies, as reported in Drugbank (Wishart et al. 2018). An alternative strategy is to modulate cytokine production by activating or repressing TF regulatory pathways or by using TF agonists or antagonists (Wishart et al. 2018; T. D. Gilmore and Herscovitch 2006; O'Keefe et al. 1992). Although the use of antibodies is a more specific therapeutic approach to inhibit cytokine activity, antibodies cannot be used in many cases because: 1) approved antibodies blocking cytokine activity are only available for nine cytokines, 2) a therapeutic strategy may require the concomitant modulation of multiple cytokines, or 3) a strategy may require the induction of cytokine activity (e.g., the induction of anti-inflammatory cytokines such as IL10) rather than inhibition. In these cases, modulation of cytokine expression by targeting TFs may provide an effective alternative approach.

Many cytokines can potentially be targeted using drugs against their interacting TFs (or the signaling pathways that activate those TFs). Indeed, multiple TF agonists and antagonists have been approved as therapeutics, including 17 TFs with targets in the human cytokine GRN (Figure 2.8D). Combined, these TFs, which include nuclear hormone receptors, NF-κB, and AP-1, can potentially target 59 cytokine genes, most of which are dysregulated in disease. Targeting these TFs can increase or decrease cytokine expression depending on the TF regulatory function and on the drug's agonist or antagonist activity. For example, IL10 expression can be induced using AHR agonists as a protective mechanism in inflammatory bowel disease, or repressed by an endogenous VDR agonist (calcitriol) during pregnancy to enhance responses to microbial infections (Goettel et al. 2016; Barrera et al. 2012). Ultimately, multiple factors need to be considered including the off-target effect of the drugs, the number of other genes whose expression may be affected by targeting a particular TF, and how the modulation of TF activity may propagate to other immune and non-immune functions.

Completeness of the cytokine GRN

Although great progress has been made in the last three decades identifying novel PDIs, the cytokine GRN is far from complete. Indeed, we observed that the size of the cytokine GRN and the number of TFs involved have increased at a constant rate suggesting that novel PDIs remain to be identified (Figure 2.10A and Figure 2.11A). Importantly, the fraction of TFs that have been incorporated into the cytokine GRN that are associated with

immune phenotypes or diseases has remained constant suggesting that the GRN continues to grow towards immune-relevant interactions (Figure 2.10B and Figure 2.11B).



Figure 2.10 Completeness of the human cytokine GRN. (A) Number of annotated PDIs, TFs, and cytokines in the human cytokine GRN over time. (B) Fraction of TFs in the human cytokine GRN with annotated immune phenotypes when knocked out in mice (MGI) or associated to immune disorders in genome-wide association studies (GWAS) and in the Human Gene Mutation

Database (HGMD) over time. (C, D) Number of PDIs per TF (C) or per cytokine (D) in the human cytokine GRN over time. (E, F) Correlation between the number of PDIs in the human cytokine GRN and the number of publications per TF (E) or per cytokine (F) reported in Medline. (G, I) PDIs with the promoters of CCL27 (G) or CCL4L2 (I) were analyzed by eY1H assays. Each interaction was tested in quadruplicate. The qualitative strength of PDIs detected by eY1H compared to AD-vector control are indicated as -, +, ++, and +++ corresponding to no, weak, medium, and strong interaction, respectively. Motif location for the indicated TFs in the promoters of CCL27 and CCL4L2 are shown. (H, J) Luciferase assays to validate interactions between the promoters of CCL27 (H) or CCL4L2 (J) and the indicated TFs. HEK293T cells were co-transfected with reporter plasmids containing the cytokine promoter region (2 kb) cloned upstream of the firefly luciferase reporter gene, and expression vectors for the indicated TFs (fused to the activation domain 10xVP16). After 48 h, cells were harvested and luciferase assays were performed. Relative luciferase activity is plotted as fold change compared to cells co-transfected with the vector control (1.0). Experiments were performed 3-4 times in three replicates. Individual data points represent the average of the three replicates, the average of all experiments is indicated by the black line. *p<0.05 by one-tailed Student's t-test with Benjamini-Hochberg correction.

Future growth of the cytokine GRN is not expected to be uniform for all TFs and cytokines. Indeed, the number of PDIs seems to have saturated for some TFs such as RELA, NFKB1, and FOS, while other TFs such as SPI1 and MAFK do not show signs of saturation (Figure 2.10C and Figure 2.11C). The number of PDIs for some well-studied cytokines such as CCL5 have also plateaued, while new PDIs are still being identified for other cytokines such as human CXCL8 and CCL2 or mouse IL4 (Figure 2.10D and Figure 2.11D).



Figure 2.11 Completeness of the mouse cytokine GRN. (A) Number of annotated PDIs, TFs, and cytokines in the mouse cytokine GRN over time. (B) Fraction of TFs in the mouse cytokine GRN with annotated immune phenotypes when knocked out in mice (MGI) or associated to immune disorders in genome-wide association studies (GWAS) and in the Human Gene Mutation Database (HGMD) over time. (C, D) Number of PDIs per TF (C) or per cytokine (D) in the mouse cytokine GRN and the number of publications per TF (E) or per cytokine (F) reported in Medline. Correlation determined by Spearman's rank correlation coefficient. (G, H) Correlation between the number of PDIs per TF (out degree) (G) or per cytokine (in degree) (H) in the human and mouse cytokine GRNs.

We also observed a bias towards highly studied TFs and cytokines as we detected a strong correlation between the number of publications in Medline associated with a cytokine or TF and the number of PDIs in the cytokine GRN (Figure 2.10E and F; and Figure 2.11E and F). An argument can be made that highly connected TFs have more pleiotropic functions and thus, are more frequently studied. However, more than 200 TFs absent in the cytokine GRN lead to an immune phenotype when knocked out in mice, many of which are associated with alterations in cytokine expression (Eppig et al. 2017). This suggests that many TFs are absent from the cytokine GRN and that many PDIs involving infrequently studied TFs are missing.

Similarly, highly studied cytokines are involved in more PDIs (Figure 2.10F and Figure 2.11F). Although we cannot rule out the possibility that highly studied cytokines have more pleiotropic roles and are regulated by different TFs in different cells and conditions, this alone cannot explain that there are no PDIs reported for 30% of the cytokines. Further, if there is a strong selective pressure to have multiple modes of regulation for certain cytokines, we would expect the mouse and human cytokine orthologs to be regulated by a similar number of TFs, but this is frequently not the case (Supplementary Figure 2.11G and H). What is more likely is that highly studied cytokines such as TNF and CXCL8 have more PDIs because they have been studied in more cell types and conditions. To test this hypothesis, we performed eY1H assays to evaluate the binding of 1,086 human TFs to the promoters of CCL27 and CCL4L2, two under-studied cytokines absent from the GRN (Figure 2.10G and I). We detected seven interactions with the CCL27 promoter involving TFAP2A/B/E, KLF7, ZNF18, PPARG, and RBPJ (Figure 2.10G). Motif analyses for TFs with available position weight matrices (TFAP2A/B, PPARG, and RBPJ) identified multiple TF binding sites in the CCL27 promoter. We evaluated the seven eY1H interactions by luciferase assays in HEK293T cells, all of which were validated (Figure 2.10H). Of note, TF ZNF18, which is widely expressed in immune

cells, is also absent from CytReg showing that novel TFs in the cytokine GRN remain to be identified. We also detected 13 TF interactions with the promoter of CCL4L2 using eY1H assays (Figure 2.10I). Multiple TF binding sites were found in the promoter of CCL4L2 for most of the TFs for which a position weight matrix was available. We tested the 13 eY1H interactions by luciferase assays in HEK293T cells, nine of which validated (Figure 2.10J). Interestingly, ATF3 is known to regulate CCL4, a close paralog of CCL4L2 (M. Zhu et al. 2014). Further, CCL4L2 is produced by multiple cell types including monocytes, B cells, T cells, fibroblasts, endothelial, and epithelial cells, while ATF3, EBF3, REL, ZBTB10, ZNF710, WT1, TFAP2A, and TFAP2E are also expressed in one or more of these cell types (C. Wu et al. 2016). Overall, this shows that novel interactions can be detected for cytokines and TFs that have been poorly characterized.

Prediction of novel PDIs in the cytokine GRN

Different cell types express different sets of cytokines in response to pathogen- or cell-mediated cues. To predict novel PDIs in the human cytokine GRN, we leveraged the observation that co-expressed genes tend to share interactions with similar TFs (Fuxman Bass et al. 2015; Marco et al. 2009). Thus, for each TF with at least two PDIs in the human cytokine GRN, we searched for other cytokines co-expressed with the known target cytokines across more than 5,000 expression profiling datasets using SEEK (Q. Zhu et al. 2015). Potential targets were then filtered by the presence of the corresponding TF binding site in the promoter region (2 kb upstream of the transcription start site) determined using CIS-BP (Weirauch et al. 2014). The 1,066 predicted PDIs, were enriched in orthologous interactions detected in mouse but absent from the human cytokine GRN (OR = 4.43, p <

10-20 by Chi-square test). Predictions were classified as high, medium, or low confidence based on the number of TF binding sites for the corresponding TF and the presence of the interaction in mouse (Figure 2.12A). As expected, there is a strong correlation between the TF degree for known and for known plus predicted interactions, although this correlation is not perfect (Figure 2.12B). Importantly, adding the predicted interactions, maintained or even improved the correlation between TF degree and expression enrichment in immune tissues, presence of immune phenotype in mouse, and association with immune disorders in GWAS and HGMD (Figure 2.12C). Overall, this suggests that our predictions are enriched in functional PDIs.



Figure 2.12 Prediction of novel PDIs in the human cytokine GRN. (A) Novel PDI predictions based on co-expression between cytokines and known cytokine targets of each TF (determined using the SEEK database), and motifs analysis. Prediction confidence, as defined in the methods section, is shown. (B) Correlation between the number of cytokine targets (TF degree) for known PDIs and known + predicted PDIs. Correlation determined by Spearman's rank correlation coefficient. (C) Correlation between TF degree for known (K) or known + predicted (K+P) PDIs and expression enrichment score (EES) in immune tissues, mouse immune phenotype (MGI), and human immune disorders in GWAS and HGMD. Correlation and significance determined by Spearman's rank correlation coefficient. (D, G) Top predicted cytokine targets of RORC (D) and REL (G). The co-expression rank among all genes and among cytokines is shown. CXCL8 is a known target of REL, while IL17A is a known target of RORC. (E, H) Enhanced yeast one-hybrid assays testing PDIs between the indicated human cytokine promoters and RORC (E) and REL (H). AD-vector corresponds to empty vector. The qualitative strength of PDIs compared to AD-vector control are indicated as -, +, ++, and +++ corresponding to no, weak, medium, and strong interaction, respectively. REL and RORC binding sites are indicated in red for each 2 kb promoter

region. (F, I) Luciferase assays in HEK293T cells co-transfected with reporter plasmids containing the indicated cytokine promoter region (2 kb) cloned upstream of the firefly luciferase reporter gene, and expression vectors for RORC (F) or REL (I) (fused to the activation domain 10xVP16). After 48 h, cells were harvested and luciferase assays were performed. Relative luciferase activity is plotted as fold change compared to cells co-transfected with the vector control (1.0). Experiments were performed 3-4 times in three replicates. Individual data points represent the average of the three replicates, the average of all experiments is indicated by the black line. *p<0.05 by one-tailed Student's t-test with Benjamini-Hochberg correction.

Using this platform, we predicted IL26 and IL17F to be novel potential targets of RORC, whose RORγt isoform is a master regulator of Th17 cell differentiation and function (Figure 2.12D) (Ivanov et al. 2006). The interaction between RORC and IL17F, a paralog of the known RORC target IL17A, was reported in mouse (X. O. Yang et al. 2008) but, to our knowledge, not in human. IL26 is a key cytokine involved in immune cell priming, antibacterial immunity, and autoimmune diseases produced by RORγt expressing Th17 cells, but not previously shown to be directly regulated by RORγt (Manel, Unutmaz, and Littman 2008; Stephen-Victor, Fickenscher, and Bayry 2016). We validated these two novel predicted PDIs using eY1H assays, motif analyses, and luciferase assays in HEK293T cells showing even stronger activity than the well-known RORC-IL17A interaction (Figure 2.12E and F). Overall, this suggests that RORC directly regulates multiple Th17 cytokines.

Using a similar approach, we found that CCL4, CXCL3, CCL20, and CXCL10 are among the most highly correlated cytokines to the known targets of the well-studied TF REL, and that their promoters have multiple binding sites for REL (Figure 2.12G and H). Interestingly, these cytokines are known to be regulated by other subunits of NF- κ B but, to our knowledge, not by REL. We validated these predicted interactions using eY1H assays and luciferase assays in HEK293T cells (Figure 2.12H and I). Interestingly, these four novel targets of REL, a TF associated with autoimmune disorders, are also associated with and/or upregulated in autoimmune disorders (Meagher et al. 2007; Klein et al. 2004; Karin and Razon 2018; Hirota et al. 2007; Thomas D. Gilmore and Gerondakis 2011). Overall, this shows that by integrating the PDIs annotated in CytReg with co-expression data we can expand the current cytokine GRN. Additionally, our predictions provide a blueprint for further studies in cytokine regulation.

Discussion

In the present study, we mined ~26 million articles in Medline, of which we curated more than 7,000 articles, to generate comprehensive mouse and human cytokine GRNs comprising 843 and 647 PDIs, respectively. We created a user-friendly database (https://cytreg.bu.edu) where PDIs can be easily browsed by TF, cytokine, species, assay type, and TF expression patterns, and visualized as networks. Overall, CytReg is 2- to 3-fold more complete than other databases such as InnateDB and TRRUST (Breuer et al. 2013; Han et al. 2015). Using this comprehensive database, we were able to obtain novel insights into the principles involved in cytokine regulation, perform comparative analyses between mouse and human GRNs, and make functional predictions which were not previously possible with other databases.

By analyzing the cytokine GRN, we found that highly connected TFs are more highly expressed in immune cells and more frequently associated with immune phenotypes and diseases compared to low connected TFs. This is consistent with previous reports correlating network connectivity and phenotype, both in protein-protein and protein-DNA interaction networks (Fuxman Bass et al. 2015; Deplancke et al. 2006; Goh et al. 2007). Interestingly, we found that this correlation is specific to immune diseases as TFs associated with non-immune diseases do not display a high connectivity in the cytokine GRN (not shown). Overall, this suggests that the link between TF connectivity and phenotype may be a local feature of GRNs where connectivity to functionally related targets, rather than the entire GRN, dictates the type of phenotypes or diseases a TF is associated with. For example, REL which is highly connected in CytReg, but not in TRRUST, has been associated with rheumatoid arthritis, psoriasis, and Hodgkin's lymphoma but not with diseases unrelated to the immune system (MacArthur et al. 2017).

Our analysis of the combinatorics of the TFs that regulate each cytokine gene illustrates the complexity in cytokine transcriptional regulation. We observed that pro- and anti-inflammatory cytokines are regulated by a different balance between PSA and TS TFs, but ultimately a combination of both types of TFs may be required for cofactor recruitment to induce cytokine expression in the appropriate cells and conditions. This cooperativity between PSA and TS TFs, together with cell type specific expression patterns of surface receptors and signaling molecules, may ultimately be responsible for the tight control of cytokine expression in immune responses. The cooperative relationship between TFs may also explain the deleterious effects of several disease-associated single nucleotide variants (SNVs) and engineered mutations in the promoters and enhancers of cytokine genes, as affecting the binding of a single TF may result in the loss of cooperativity and lead to gene misregulation (Melnikov et al. 2012; Wei et al. 2011; Tu et al. 2013). For example, using massively parallel reporter assays it was recently shown that ~60% of all possible substitutions in the core 44 nt of the IFNB1 enhanceosome altered its activity in virus-

infected cells (Melnikov et al. 2012). Remarkably, most of the substitutions that did not affect activity were located outside of known TF binding sites or led to an alternative binding site for the same TF.

Our analyses also suggest a potential plasticity between TFs in cofactor recruitment, given that frequently multiple TFs that regulate a cytokine gene can interact with the same domain of EP300/CREBBP. Fine-mapping TF interactions with protein domains of other cofactors will indicate whether this is a unique feature of EP300/CREBBP. Further, a comprehensive functional characterization of different substitutions in cytokine promoters may determine whether the substitutions that affect the binding of potentially redundant TFs are generally more benign than those affecting the binding of cooperative TFs. However, the converse can also be true as this plasticity may be required for proper cytokine expression in different cell types and conditions.

CytReg is the most comprehensive cytokine GRN to-date, significantly increasing the number of annotated PDIs compared to previous databases, yet CytReg is not fully complete. First, articles that do not mention interactions within the information available in Medline will be missed and will not have been curated. Second, CytReg is incomplete because multiple PDIs remain to be evaluated and characterized. Indeed, by performing eY1H and luciferase reporter assays, we found interactions involving cytokines (CCL27 and CCL4L2) and TFs (e.g., ZNF18, ZBTB10, KLF17, EBF3, and ZNF710) that are absent from CytReg. Further, by leveraging CytReg, co-expression data, and motif analyses we predicted 1,066 PDIs in the human cytokine GRN, a subset of which we validated by eY1H and luciferase assays. Third, in addition to missing PDIs in the cytokine GRN, individuals may carry genomic variants in noncoding regulatory regions of cytokine genes or in TF coding sequences that lead to different TF-cytokine interactions. Indeed, several diseaseassociated SNVs have been identified in the promoters of cytokine genes that result in the gain or loss of PDIs that may be absent in CytReg (Fuxman Bass et al. 2015; Nickel et al. 2000; Sánchez et al. 2009; Knight et al. 1999). For example, a SNV in the proximal promoter of CCL5 that is associated with atopic dermatitis leads to a gain of PDI with GATA2 (Fuxman Bass et al. 2015; Nickel et al. 2000). Finally, CytReg catalogues PDIs as binary interactions between TFs and cytokine genes. However, the number of binding sites for each TF, their strength, spacing, and orientation are key for appropriate gene expression (Spitz and Furlong 2012; Smith et al. 2013). With a few exceptions (e.g., the IFNB1 and the TNF enhanceosomes), this regulatory logic is currently unknown, and thus cannot be annotated (Thanos and Maniatis 1995; Tsytsykova and Goldfeld 2002). Ultimately, the integration of different high-throughput and unbiased approaches, population-wide studies of regulatory variation, and in-depth functional characterizations of the regulatory logic will lead to a more comprehensive picture of cytokine regulation in different cell types, conditions, and individuals.

Funding

This work was supported by the National Institutes of Health [R00 GM114296 and R35 GM128625 to J.I.F.B.; and 5T32HL007501-34 to J.A.S.] and the National Science Foundation [NSF-REU BIO-1659605 to M.M.].

Author contributions

S.C.P. performed the data mining in Medline. J.I.F.B., A.D.I., K.A.G., J.A.S., M.M., R.S. and S.M. performed the literature curation. S.C.P. and A.D.I. designed the CytReg web tool. J.I.F.B. and S.C.P performed the data analysis. J.I.F.B., C.S.S. and K.A.G. performed the experiments in Figures 6 and 7. J.I.F.B. conceived the project and wrote the manuscript with contributions from S.C.P, C.S.S. and J.A.S. All authors read and approved the manuscript.

Chapter 3. Prediction of genome-wide effects of single nucleotide variants on transcription factor binding

Adapted from the following manuscript:

 Sebastian Carrasco Pro, Katia Bulekova, Brian Gregor, Adam Labadorf, Juan Ignacio Fuxman Bass. 2020. Prediction of genome-wide effects of single nucleotide variants on transcription factor binding. *In preparation*.

Introduction

Changes in gene expression caused by single nucleotide variants (SNVs) residing in transcriptional control regions have been shown to cause phenotypic changes which may be adaptive or lead to disease (Maurano et al. 2012; 2015; Hindorff et al. 2009). The mechanisms of action of these SNVs include alterations in the binding of transcription factors (TFs), in the recruitment of RNA Polymerase II, in nucleosome positioning, and in DNA modifications. Among these, the creation and disruption of TF binding sites (TFBSs) is likely the main mechanism by which SNVs affect gene expression (Maurano et al. 2015).

Experimental methods to determine changes in TFBSs driven by SNVs include electrophoretic mobility shift assays (EMSA), chromatin immunoprecipitation followed by sequencing (ChIP-seq), and enhanced-yeast one-hybrid (eY1H) assays (Gan et al. 2018). EMSA is a very low-throughput assay that tests one or few TFs and DNA sequences at a time, and requires TF purification or anti-TF-specific antibodies. ChIP can be used to study differential TF recruitment by SNVs, but can only be tested one TF at time, is limited by the availability of high-quality anti-TF antibodies, and more importantly, requires cells heterozygote for the SNV of interest. eY1H instead can determine altered TF binding to a SNV by testing the full repertoire of TFs, but can only test one SNV per experiment. Thus, current experimental methods are limited by the amount of SNVs and TFs they are able to test in a single experiment. Due to these limitations, prediction algorithms based on experimentally determined motifs have been developed for high-throughput prediction of altered TF binding by SNVs.

TFs binding preferences to DNA sequences, represented by position weight matrices (PWMs), have been used to predict the likelihood that a TF binds a DNA sequence of interest. These computational methods, that scan DNA regions to predict TFBSs, include FIMO (Grant, Bailey, and Noble 2011), RSAT (Thomas-Chollier et al. 2011), Clover (Frith et al. 2004), and ENCODE DREAM Challenge derived methods (Quang and Xie 2019; Keilwagen, Posch, and Grau 2019), among others. In addition, methods have been developed to predict the impact of SNVs in TF binding, where scores of the mutated and reference DNA sequences are compared (Coetzee, Coetzee, and Hazelett 2015; Fu et al. 2014; Weirauch et al. 2014; Boyle et al. 2012; Rentzsch et al. 2019; Movva et al. 2019). These methods have been used to predict the effect on TF binding of disease-associated SNVs such as those identified in genome-wide association and genetic studies (Xu and Taylor 2009; Tak and Farnham 2015; Schaub et al. 2012), and somatic mutations observed in tumor samples (Rheinbay et al. 2017; Yiu Chan et al. 2019; Rheinbay et al. 2020; Law et al. 2019). Furthermore, databases assessing the effect of known SNVs in the human population in gain/loss of TFBSs have been used to obtain insights into the effect of human variation on TF binding (Boyle et al. 2012; Shin et al. 2019; Kumar, Ambrosini, and Bucher 2017). However, the effect of novel or unseen SNVs, such as rare variants and somatic

mutations, on TF binding has not yet been determined. In this regard, a recent study evaluated the impact of tri-nucleotide cancer mutational signatures on TFBSs (Yiu Chan et al. 2019). This study calculated the differential probabilities of gain and loss of TFBSs corresponding to each TF for each mutational signature based on calculating the effect of SNVs across DNA k-mers found in the human genome. However, this method precludes identifying the sets of TFBSs that are poised to be gained and lost by SNVs as it assumes a uniform distribution of k-mers across the human genome.

Here, we generated a database of genome-wide altered TFBSs by in silico mutating all possible SNVs in every position in the human genome and determining gain and loss of TFBSs for 1898 PWMs corresponding to 741 human TFs. Using this resource, we show that the probability to gain (gainability) or disrupt (disruptability) a TFBS in gene regulatory regions widely differ between different TFs and TF families. We also show that functional cis-eQTL SNVs are more likely to perturb TFBSs than common SNVs in the human population. Interestingly, the difference in disruptability is driven both by a higher probability of SNVs residing within TFBSs and a lower probability of retaining existing TFBSs by cis-eQTL versus population-wide SNVs. Finally, we show that somatic mutations in different cancer-types have differential effects on TFBSs between TF families and discuss how these profiles are related to distinct cancer mechanisms. Altogether, this database provides blueprint to study the impact of SNVs associated with genetic variation and cancer on TF binding.

Materials and Methods

Generation of the altered TF binding site database

To predict the effect of all possible SNVs in the human genome on TF binding, for each possible SNV and each TF with available PWMs, we calculated the binding score for the reference and alternate SNV alleles. We downloaded 1898 PWMs corresponding to 741 human TFs from CIS-BP (Weirauch et al. 2014) on April 3 2018 and their respective TF family. Given a PWM of length n and a genomic position (hs37d5 from the 1000 Genome Project), for each of the 2n-1 DNA sequences on each strand of length n that overlap with the genomic position, we calculated a TF binding score using the function:

$$F(s,M) = \sum_{i=1}^{n} \log\left(\frac{M_{s_i,i}}{b_{s_i}}\right)$$

where s is a genomic sequence of length n, M is the PWM with n columns and each column in M contains the frequency of each nucleotide in each position i = 1,...,n, and bsi is the background frequency of nucleotide si (assuming a uniform distribution). The highest score obtained for the 4n-2 sequences was assigned as the binding score corresponding to the PWM for the reference or alternate SNV alleles. Significant scores were selected and reported based on TFM-p-value (Touzet and Varré 2007) score thresholds determined using a significance level of $\alpha = 10$ -4. This method was applied for each reference position and the three possible alternate SNVs for the complete genome (hs37d5) to create the altered TFBS database, a genome-wide catalogue of predicted SNV-PWM effects. A custom program was written in C and CUDA to generate the dataset (https://github.com/fuxmanlab/altered_TFBS). The program was executed on Nvidia GPUs that are available on the Boston University Shared Computing Cluster (SCC). The 6.1Tb dataset was stored in a compressed Parquet format on a 320-core Hadoop cluster that is also part of the SCC. In addition, a query system was developed using Python and PySpark that was run on the BU Hadoop cluster. The query system was used to search either a set of SNVs from a variant calling format (VCF) file (e.g., population-wide SNVs or somatic mutations), or all possible SNVs from genomic regions in BED files (e.g., promoter or DNase hypersitive site (DHS) regions). In both cases, the query reports the PWM scores for each reference/alternate genomic position pair where at least one of the alleles has a significant score for the given PWM. As an example, a query consisting of the human promoter coordinates from a BED file took about 60 minutes to complete on the Hadoop cluster.

Genomic region definitions

The hs37d5 human genome, downloaded from the Sanger Institute (November 2, 2018), was used as reference. Promoters were defined as regions from -2000 bp to +250 bp from all transcription start sites (TSSs) from protein coding genes available at GENCODE 19 version (June 14, 2018) (Harrow et al. 2012). We used the R package IRanges (M. Lawrence et al. 2013) and BEDTools (Quinlan and Hall 2010) to extract promoter coordinates and DNA sequences. DHS genomic coordinates were obtained by taking the union of DHS regions from all samples of the Roadmap Epigenomics Mapping Consortium (July 31, 2019) (Chadwick 2012).

Generation of reference parameters for altered TF binding in genomic regions

SNVs may affect TF binding by either creating or disrupting TFBSs. Therefore, we defined two parameters to estimate these effects for each given TF-PWM: gainability and

disruptability. Gainability was defined as the ratio between the number of SNVs that lead to gain of TFBSs and the total number of SNVs that are not located within existing TFBS for the given PWM. This corresponds to the probability of creating a TFBS for a given PWM for the set of SNVs analyzed assuming equal likelihood of nucleotide changes. Disruptability was defined as the ratio between the number of SNVs that disrupt a TFBS and the total number of possible SNVs. This corresponds to the probability of a SNV disrupting an existing TFBS for a given PWM assuming equal likelihood of nucleotide changes. Disruptability can be divided into two components: hitability, which is the probability of a random SNV residing within a TFBS corresponding to the PWM; and robustness, which is the probability of a SNV that resides within a TFBS to retain the TFBS. Thus, disruptability corresponds to the hitability multiplied by 1 – robustness of a PWM. In the case of TFs with multiple PWMs, we used the median score across PWMs as the representative one for each parameter. The four parameters (gainability, disruptability, hitability, and robustness) was calculated for each TF for the human genome, promoters, and DHS regions.

Analysis of parameter scores for population-wide and cis-eQTL SNVs

To predict the effect Population-wide SNVs were downloaded from the 1000 Genomes Project (Auton et al. 2015) in vcf format (October 1, 2019). BEDTools intersect function was used to select SNVs in promoters or DHS regions. Gainability, disruptability, hitability, and robustness scores were calculated as described above. For DHS regions, we calculated the correlation of each population-wide TF score against their population-wide specific reference set derived from a random sampling of mutations based on the mutational frequency of each of the twelve types of SNV changes in the 1000 Genomes Project set (see below). In addition, we downloaded finely mapped cis-eQTL SNVs from GTEx (Aguet et al. 2017) (October 10 2020) reported by CaVEMaN (Brown et al. 2017) and DAPG (Wen, Pique-Regi, and Luca 2017) methods. BEDTools intersect function and a custom R script were used to obtain unique cis-eQTL SNVs located in promoter and DHS regions that were identified by both cis-eQTL prediction algorithms. Then, gainability, disruptability, hitability, and robustness scores were calculated for the cis-eQTL SNVs. To determine whether the altered TF binding parameters were different than expected by chance between population-wide and cis-eQTL SNVs, we subtracted the individual scores for each TF to the reference set generated from a random sampling model (see below) to calculate Δscores for gainability, disruptability, hitability, and robustness.

Estimation of a population-wide SNV-specific reference set of TFBS parameters

A reference set of scores for gainability, disruptability, hitability, and robustness was generated for the population-wide and cis-eQTL analysis. One million randomly selected SNVs were selected matching the frequency of the twelve possible mutations from the population-wide SNVs located in DHS regions. One hundred random samples were generated and the four parameters per sample were calculated for each PWM as previously discussed. Finally, the population-wide derived reference set for each parameter correspond to the average values for each PWM across the one hundred random samples.

Calculation of parameters for cancer somatic and carcinogen SNVs

77

Somatic SNVs were obtained from 2,658 whole genome sequenced samples from the PCAWG cohort across 20 cancer types (Rheinbay et al. 2020). For each cancer type, we combined the SNVs across its associated samples and generated a unique set of SNVs per cancer type. BEDTools intersect function was used to extract SNVs in DHS regions for each cancer type. The observed gainability, disruptability, hitability, and robustness scores were calculated for each TF and were subtracted by their corresponding score from the reference set of all possible SNVs in DHS regions. This resulted in Δ scores for each PWM-cancer type combination. We also calculated the median Δ score for each TF family and generated heatmaps in Prism version 8.3.1. Furthermore, we calculated the observed ∆scores for gainability and disruptability for the 741 TFs for individual samples having more than 5,000 SNVs located in DHSs. Heatmaps comparing Ascores for individual samples and TFs were generated using the R package ComplexHeatmap (Gu, Eils, and Schlesner 2016). Finally, we downloaded SNVs caused by UV-light (Kucab et al. 2019) and these SNVs were filtered to obtain Δ scores for each parameter in DHS regions as described for the PCAWG analysis. We calculated the correlation of the UV-light derived Δ scores for gainability and disruptability to the corresponding Δ scores from skin cancer PCAWG samples.

Statistical analysis

Custom R scripts and Prism were used for statistical analysis. Correlation tests were performed using the Pearson correlation coefficient and group comparisons were performed using Kruskal-Wallis rank-sum test.

Results

Estimating the effects of SNVs in creating and disrupting TFBS

To predict the effect of each possible SNV in transcriptional control regions on TF binding, we focused DHS regions, which are generally associated with transcriptionally active or poised genomic regions. We calculated binding scores for 1,898 PWMs available in CIS-BP (Weirauch et al. 2014) corresponding to 741 human TFs, for each reference and alternative allele. For each PWM-SNV combination, we determined whether the alternative allele created or disrupted a TFBS. Then, we defined two parameters: 'gainability' as the probability of a random SNV creating a binding site for a given TF, and 'disruptability' as the probability of a random SNV disrupting an existing binding site for a given TF (Figures 3.1A-B). We also determined the gainability and disruptability scores genome-wide, and contrasted to that of DHS and gene promoter regions. We detected a wide range of distributions of gainability and disruptability scores for different TFs spanning five orders of magnitude which highly anti-correlated with the information content of the PWMs. We found a strong correlation for both scores between the different genomic regions suggesting that there is no clear a priori preference for random mutations to lead to gain or disrupt TFBSs both for regulatory regions and the whole genome. Interestingly, we found a higher disruptability for AP-1 TFs (e.g., FOS, FOSL1, FOSL2, JUN, JUNB, JUND), TAL1, and NFE2 in DHSs than in promoter regions, consistent with previous findings that these TFs are enriched in enhancer regions (Gerstein et al. 2012; Dunham et al. 2012). Conversely, SP1-9 TFs display a higher disruptability in promoter

regions, consistent with known roles of SP factors in regulating RNA Pol II recruitment to core promoters and regulating transcriptional activity.



Figure 3.1 Prediction of the effect of SNVs on TF binding in DHSs. (A-D) The distribution of gainability (A), disruptability (B), hitability (C), and robustness (D) in DHSs were calculated for all TFs with available motifs in CIS-BP and binned by TF family. Significant differences for each parameter between a TF family and all TFs were calculated using a Mann-Whitney U test. * p < 0.05. (E) The correlation between each of the four parameters was estimated using the Pearson correlation coefficient.

TFs from the same DNA binding domain (DBD) family often have similar DNA binding preferences, in particular for certain families such as homeodomains, ETS factors, bHLH factors, and nuclear receptors, and are frequently different between TFs from different families (Weirauch et al. 2014). Thus, we expected different TF families to differ in gainability and disruptability scores. Indeed, we observed that homeodomain and forkhead TFs have a higher gainability than other TFs whereas bZIP, ZF-C2H2, nuclear receptors, and T-box have a lower gainability (Figure 3.1A). A similar trend was observed

for disruptability of these TF families (Figure 3.1B), suggesting that homeodomains and forkhead TFs are more likely to be rewired by SNVs than other TF families. This is likely due to the short homeodomain and forkhead TF motifs, as we observed that gainability and disruptability are overall anti-correlated with motif length and information content.

The likelihood of SNVs disrupting TFBSs for a TF is influenced by two parameters: 1) hitability (i.e., the probability of a SNV residing within an existing TFBS), and 2) robustness (i.e., the chance that a SNV in a TFBS for such TF would not affect TF binding). In this way, disruptability is equal to the product of hitability and 1 – robustness. Of these two parameters, hitability has a larger impact on the difference in disruptability between TFs as it spans five orders of magnitude compared to robustness which spans only one order of magnitude (Figure 3.1C-D). Interestingly, although hitability, gainability, and disruptability are all highly correlated with each other (Figure 3.1E), in part driven by the information content of the PWMs, robustness is lowly correlated with these parameters (Figure 3.1E). Further, contrary to the other parameters, robustness is correlated to the information content per base in the PWM which has low variantion between TFs, rather than the total information content.

Evidence of noncoding selection in population-wide SNVs

The human population displays high variability in genome sequence with close to 100 million SNVs being reported (Auton et al. 2015). Most of these SNVs reside in noncoding regions of the genome potentially creating or disrupting TFBSs (Maurano et al. 2012; 2015; Hindorff et al. 2009). The vast majority of these SNVs are expected to be neutral and be depleted of SNVs under negative selection. Thus, we hypothesized that

SNVs present in the population would be depleted in those that alter TF binding, as changes in gene expression are expected to be evolutionarily constrained. To study the effect of population-wide genetic variation on TF binding, we analyzed SNVs from the 1000 Genomes Project (Auton et al. 2015) located in DHS regions and determined gainability, disruptability, hitability, and robustness scores for each TF. We compared these parameters to a reference set derived from a random sampling of mutations based on the mutational frequency of each of the twelve types of SNV changes in the 1000 Genomes Project set. Interestingly, 89.2% of the TFs show a significantly higher gainability score than the reference (Figure 3.2A). In contrast, 66.8% of the TFs show a significantly lower disruptability for the population-wide SNVs (Figure 3.2B). These results suggest a selection of population-wide SNVs against disrupting existing TFBSs and a positive selection towards creating TFBSs.



Figure 3.2 Differential parameter scores for population-wide and cis-eQTL SNVs. (A-D) Correlation between scores derived from SNVs from the 1000 Genomes Project (1000 genomes) and the average of 100 random sets of 1,000,000 SNVs (reference) for gainability (A), disruptability (B), hitability (C), and robustness (D). Correlation was determined by the Pearson correlation coefficient. Significantly enriched (red) and depleted (blue) TFs are highlighted. (E-H) Δ scores (observed in set – reference) for each parameter for all TFs and specific TF families for population-wide and cis-eQTL SNVs. Significant differences between the population-wide and cis-eQTL scores were determined by a Mann-Whitney U test. * p < 0.05.

We further calculated the hitability and robustness scores for population-wide

SNVs to explore the mechanisms of the negative selection observed for disruptability.

Strikingly, we found that even though hitability is similar between population-wide SNVs and the reference (Figure 3.2C), population-wide SNVs show higher values for robustness for 81.6% of TFs (Figure 3.2D). These results suggest that the negative selection towards TFBS disruption in population-wide SNVs is mainly driven by the selection for SNVs that, even though they may reside within existing TFBSs, they do not perturb TF binding.

cis-eQTL SNVs display a high likelihood to create and disrupt TFBSs

Previous studies on cis expression quantitative trait loci (cis-eQTLs) have identified functional sets of SNVs in transcriptional control regions associated with changes in target gene expression (Aguet et al. 2017). We compared the scores of cis-eQTL and populationwide SNVs for each parameter in this study to the reference score obtained from a random sampling to generate Δ scores (SNV group - reference). We found high Δ gainability and Δ disruptability scores for all TF families in the cis-eQTL SNV set compared to the Δ scores for the population-wide set (Figure 3.2E-F). This suggests that cis-eQTLs are enriched in SNVs that create or disrupt TFBSs which likely contributes to their effect in differential gene expression. We further investigated the effects on cis-eQTLs disruptability and found that cis-eQTL SNVs lead to higher Δ hitability and lower Δ robustness scores than population-wide SNVs (Figure 3.2G-H). These findings suggest that the increased disruptability by cis-eQTLs SNVs is due to both an increase in SNVs being located in existing TFBSs and by affecting bases with higher information content within those TFBSs.

Cancer somatic mutations display cancer-and TF family-specific effects on TFBS

Cancer is characterized by the presence of somatic SNVs in tumors, more than 90% of which reside in noncoding regions of the genome (Araya et al. 2016b). It has been shown that different cancer-types display different mutational signatures driven by different mutation and DNA repair mechanisms (Alexandrov et al. 2013; 2020). Given the DNA binding specificity differences between TFs, we hypothesized that mutational signatures specific to different cancer-types may affect TFBSs differentially across TF families. To investigate this hypothesis, we selected SNVs located in DHS regions from 20 cancer types from 2,658 tumor samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium (Campbell et al. 2020) and calculated, for each TF, its Δ gainability, Δ disruptability, Δ hitability and Δ robustness scores relative to the reference scores in DHSs.

We found higher Δ gainability scores for forkhead and Sox families across many cancer-types (Figure 3.3A), with the highest enrichment in colon/rectum cancer. This is consistent with studies showing that the forkhead TFs FOXO3 and FOXA1, which have a 2 and 2.4-fold increase in gainability in colon/rectum cancer respectively, promote colon cancer proliferation (Gao et al. 2019). Similarly, overexpression of FOXJ1 has been linked to progression of colorectal cancer by promoting translocation of β -catenin (K. Liu, Fan, and Wu 2017). Sox TFs are also associated with cancer, including SOX11 that shows a 1.5-fold increase in gainability in breast cancer and that has been correlated with breast cancer growth and invasion (Shepherd et al. 2016). Overall, these results support a positive selection to gaining and maintaining forkhead and sox TFBSs in multiple cancers.



Figure 3.3 Effect of cancer somatic mutations on TFBSs. (A-D) Median Ascores for each TF family and cancer-type combination for gainability (A), disruptability (B), hitability (C), and robustness (D). (E-F) Motifs logos for NFATC4 (E) and ELF4 (F) and impact of melanoma mutational signatures on the gain and disruption of the corresponding motifs.

Other associations for Δ gainability scores between TF families and cancer-types are more specific. For example, we found gain of homeodomain TFBSs to be highly enriched in colon cancer (Figure 3.3A). Indeed, HOXA3, a homeodomain TF that shows a 1.5-fold increase in gainability, has been shown to promote colon/rectum cancer (X. Zhang et al. 2018). Other TFs from the homeodomain subfamilies HOXB and HOXD have also been found to be up-regulated in cancer (S. Yang et al. 2018; de Bessa Garcia et al. 2020), displaying an average 2.8 and 2.4-fold increase in gainability across the subfamily, respectively. Furthermore, skin cancer shows an enrichment in gain of rel TFBSs, which is mainly driven by the NFAT subfamily. In particular, NFATC3 (3.8-fold increase in gainability) is highly expressed in skin cancer and is associated with cell transformation and tumor growth in this cancer type (Xiao et al. 2017). Conversely, we found a depletion to gain TFBSs from the bHLH, bZIP, and ZF-C2H2 families in skin cancer. In particular, we found that all of CREB TFs from the bZIP family show a negative Δgainability in skin cancer, where these TFs have been reported to inhibit tumor growth and metastasis (Xie et al. 1997). In addition, ZBTB7A, a ZF-C2H2 TF with a 2.3-fold decrease in gainability in skin cancer, suppresses melanoma metastasis (X. S. Liu et al. 2015).

In contrast to Δ gainability, we found negative Δ disruptability scores for forkhead, homeodomain, nuclear receptor, rel, sox and T-box families across most of the 20 cancer types analyzed (Figure 3.3B). These results suggest a negative selection towards disrupting TFBSs for these families. Contrary to what we observed for population-wide SNVs where the reduced Δ disruptability was associated to an increase in Δ robustness, the reduced disruption for cancer mutations is associated with both an increase Δ robustness and a reduced Δ hitability, suggesting negative selection (Figure 3.3C-D). The only exceptions having a higher Δ disruptability score correspond to rel and ETS TFs in skin cancer, many of which have been associated with melanoma. This is consistent with the frequency of triplets matching the mutational signatures of melanomas (TCN \rightarrow TTN and CCN \rightarrow CTN) (Alexandrov et al. 2020) within motifs of rel factors such as NFATC4 (Figure 3.3E) and ETS factors such as ELF4 (Figure 3.3F). Altogether, our results suggest that cancer
mutations lead to a net increase in TF binding sites for forkhead, homeodomain, nuclear receptor, rel, sox and T-box families.

Different tumors, even from the same cancer-type, can have different mutational signatures. Thus, we determined the Δ gainability and Δ disruptability profile for 162 highly mutated tumors (>5,000 SNVs in DHSs) across 741 TFs. We observed a similar overall clustering pattern across tumors (Figure 3.4A-B). Interestingly, all highly mutated skin cancer samples clustered together showing a similar pattern of gain and loss of TFBSs. This pattern is highly correlated to that of SNVs introduced by treating cell lines with UV light (Δ gainability, r=0.75, p-value<2x10-16 and Δ disruptability, r=0.78, p-value<2x10-16) (Figure 3.4C-D), consistent with UV light being a major mutational driver of skin cancer SNVs. Surprisingly, colon/rectum tumor show two subtypes, where one subtype shows depletion of bZIP, bHLH and C2H2 zinc finger TFs and an enrichment of homeodomain TFs and the other subtype shows the opposite profile for both Δ gainability and Δ disruptability (Figure 4A-B). The origin of these subtypes remains to be determined.



Figure 3.4 Effect of cancer somatic mutations in individual cancers on Δ gainability and Δ disruptability. (A-B) For cancer samples with at least 5,000 SNVs in DHS regions, we determined for each TF the Δ gainability (A) and Δ disruptability (B) scores. Samples were clustered using hierarchical clustering, and TF were clustered by TF families. Cancer-types are indicated at the top and TF families are indicated at the right of each heatmap, respectively. (C-D) Correlation between UV-light-derived Δ gainability (C) and Δ disruptability (D) scores for each TF to those observed in skin cancer. Correlation calculated by the Pearson correlation coefficient.

Discussion

In this study, we generated a comprehensive database of altered TFBSs by mutating

all possible SNVs across the genome. Using this resource, we determined the gainability,

disruptability, hitability, and robustness scores for 741 TFs across the genome, promoters, and DHS regions. We found differences in gainability and disruptability scores between TF families. Interestingly, we found lower gainability and disruptability values for bZIP, C2H2 ZF, nuclear receptors, and T-box, showing that binding sites for these TF families are less likely to be affected by SNVs. In contrast, forkhead and homeodomain display higher scores for both gainability and disruptability, suggesting a higher rewiring potential of the gene regulatory networks controlled by these TFs. Whether in vivo binding site occupancy for these TFs is actually rewired across evolution or between individuals in the human population, remains to be determined.

We showed that functional cis-eQTL SNVs are more likely to perturb TFBSs than common SNVs in the human population. In addition, we observed that somatic mutations in cancer have differential effects on TFBSs for multiple TF families and discuss how these profiles are related to distinct cancer mechanisms. Altogether, this database provides blueprint to study the impact of SNVs on genetic variation and cancer. In addition, our results can be implemented further in methods to identify functional SNVs in sequencing data, as our estimated probabilities can be used as background probabilities to compare germline or somatic mutations associated with disease in a given cohort.

By comparing the genome-wide gainability and disruptability to the respective gene regulatory regions, we found that score for different genomic regions are highly correlated. This suggests that SNVs are likely to affect TFBSs across the genome in a similar manner, independent of the genomic function. We hypothesize that the difference between genomewide and gene regulatory regions is determined by the complex gene regulatory logic that govern TF binding to transcriptional control region rather than the TFBSs themselves. These include factors such as the proximity, co-occurrence, and orientation of TFBSs, as well as cooperative or competitive binding/regulation between TFs (Spitz and Furlong 2012; Claussnitzer et al. 2014).

By analyzing the parameter patterns of population-wide SNVs we showed that 89% of TFs showed increased gainability. However, this increase is significantly lower to the higher gainability values found in cis-eQTLs SNVs that correspond to expression perturbing SNVs. In contrast, 67% of TFs showed a decrease in disruptability by the population-wide SNVs, whereas the cis-eQTL SNVs displayed an increase in disruptability scores. Interestingly, this difference is driven by two factors: a higher likelihood of cis-eQTL SNVs to reside within a TFBS and a higher likelihood of population-wide SNVs that land in a TFBS to retain it. These results can be explained by most population-wide SNVs being neutral, not affecting gene expression; however, there is a tendency for positive selection of gain of TFBSs and negative selection for loss of TFBSs. This suggests a higher selective pressure to maintain existing TFBSs which function together with other TFs within specific cis regulatory logics, while gain of TFBSs can provide evolutionary plasticity.

To our knowledge, this is the first database that predicts the effect of all possible SNVs on TF binding. The database of genome-wide altered TFBSs generated in this study and the gainability, disruptability, hitability and robustness parameters calculated for each TF provide a powerful resource to predict the effect of SNVs on TF binding and provide a background for further studies in specific transcriptional control regions or produced by SNVs present in specific patient cohorts. Other applications of this resource include studying the potential of repetitive elements as latent reservoirs of TFBSs and uncovering the role of other disease associated SNV sets and carcinogen signatures. Ultimately, the integration of other datasets such as i TF dimer motif specificities, TF motifs in the context of nucleosomal DNA (F. Zhu et al. 2018), and the inclusion of new TF motifs as they become available, will lead to a more comprehensive model of the effect of SNVs on TFBSs.

Funding

This work was supported by US National Institutes of Health grant R35 GM128625 to J.I.F.B..

Author contributions

J.IF.B. and S.C.P. conceived the project and wrote the manuscript. S.C.P., K.B. and B.G. generated the altered TFBS database. S.C.P. performed the data analysis which was supervised by J.F.B. and A.L.. All authors read and approved the manuscript.

Chapter 4. Discovery and characterization of cancer driver mutations in gene promoters

Adapted from the following manuscripts:

 Carrasco Pro S, Bray D, Hook HJ, Yin M, Bulekova K, Gregor B, Labadorf A, Tewhey R, Siggers T, Fuxman Bass JI, 2020. Discovery and characterization of cancer driver mutations in gene promoters. *In preparation*.

Introduction

Cancer initiation and progression often originates from environmentally induced or spontaneous mutations, and/or inherited genomic variants that increase cancer risk (Alexandrov et al. 2013; Helleday, Eshtad, and Nik-Zainal 2014b; Ding et al. 2018). Large scale projects such as the Cancer Genome Atlas (TCGA) and the International Genome Consortium (ICGC) have identified millions of somatic SNVs in tumors(Weinstein et al. 2013; Hudson et al. 2010). However, in most cases, it is not known whether these mutations affect any cellular function, confer growth advantage, and are causally implicated in cancer development (Pon and Marra 2015). This is because only a few cancer driver mutations are needed to drive tumor initiation and growth and these mutations have to be distinguished from thousands of passenger mutations (Pon and Marra 2015). The vast majority of these cancer drivers have been identified in coding regions. Even though more than 90% of somatic SNVs are located in noncoding regions, only a handful of noncoding cancer drivers have been identified (Khurana et al. 2016).

Noncoding variants (NCVs) may affect the binding of transcription factors (TFs) and cofactors (CoF) leading to changes in gene expression (Khurana et al. 2016). For

example, TERT overexpression is a major contributor to cancer and has been shown to be caused by NCVs in its promoter that create Ets factors binding sites (Susanne Horn et al. 2013; Huang et al. 2013; Shrestha et al. 2019). Other examples of characterized noncoding cancer drivers include NCVs in the promoters of FOXA1, HES1, SDHD, PLEKSH, among others (Weinhold et al. 2014; Rheinbay et al. 2017; Piraino and Furney 2017). Further, the analysis of 2,568 cancer whole genome samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) predicted driver NCVs in the promoters of 9 genes and estimated 96 potential driver NCVs gene promoters within this cohort (Rheinbay et al. 2020). Whether this is due to a limited contribution of NCVs to cancer or to limitations of current approaches to predict NCV drivers remains to be determined.

Computational methods to predict driver NCVs, collectively called mutational burden tests, are based on determining an increased mutational rate (MR) in cis-regulatory elements (CREs) compared to a background mutational rate (BMR) (H. Li 2011; Martincorena et al. 2017; Shuai et al. 2020; Lanzós et al. 2017; Lochovsky et al. 2015; M. S. Lawrence et al. 2014; Nik-Zainal et al. 2016; Juul et al. 2017; Hornshøj et al. 2018). These methods consider different parameters to estimate the BMR such as cancer-specific mutational signatures, sequence conservation, functional annotations, and mutational rates in neighboring regions or other "similar" genomic regions. In addition, other covariates may be used such as replication timing, expression levels, and motif analysis. These mutational burden tests have only identified a handful of drivers NCVs given that most NCVs are passenger and that the BMR is locus specific (Rheinbay et al. 2020). Thus, studies have focused on cancer-associated genes or proximal promoters to increase the predictive power of these methods. Given the reduced number of predicted driver NCVs, studies have used low-throughput methods for experimental validation such as report assays, EMSAs, and allelic imbalance in gene expression or TF binding.

Here, we developed a novel TF-aware burden test (TFABT) based on the hypothesis that creating (or disrupting) a TFBS at different positions within a gene promoter is likely to lead to similar effects on target gene expression. It has been reported that TF binding sites in promoters and enhancers frequently occur in homotypic clusters and regulate gene expression through cooperative and non-cooperative mechanisms. This TFABT identifies promoters containing a higher than expected number of mutations across patients that create/disrupt a specific TFBS in a CRE using a binomial test. We predicted 2,555 cancer driver NCVs in the promoters of 813 genes across 20 cancer types. These genes are enriched in cancer-related genes, essential genes, and their expression levels are associated with cancer prognosis. More importantly, we validated 765 NCVs using massively parallel reporter assays (MPRAs) and observed a similar validation rate to known drivers. Finally, we found that 604 NCVs show differential cofactor recruitment by comprehensive assessment of complex assembly at DNA elements (CASCADE).

Materials and Methods

Altered transcription factor binding predictions

To predict the effect of all possible SNVs in the human genome on TF binding, for each possible SNV and each TF with available PWMs, we determined the binding score corresponding to the reference and SNV sequences. We downloaded 1898 position weight matrices (PWMs) corresponding to human TFs from CIS-BP on April 3 2018 (Weirauch et al. 2014) and their corresponding TF family. Given a PWM of length n and a genomic position (hs37d5 from the 1000 Genome Project), for each of the 2n-1 DNA sequences on each strand of length n that overlap with the genomic position, we determined a TF binding score using the function:

$$F(s,M) = \sum_{i=1}^{n} \log\left(\frac{M_{s_i,i}}{b_{s_i}}\right)$$

where s is a genomic sequence of length n, M is the PWM with n columns and each column in M contains the frequency of each nucleotide in each position i=1,...,n, and bsi is the background frequency of nucleotide si (we assume a uniform distribution). The highest score obtained for the 4n-2 sequences was assigned as the binding score corresponding to the PWM for the reference or alternate SNV alleles. Significant scores were selected and reported based on TFM-p-value (Touzet and Varré 2007) score thresholds determined using a significance level α =10-4. This method was applied for each reference position and the three possible SNVs for the complete genome (hs37d5) to create the altered TFBS database, a genome-wide catalogue of SNV-TF effects. Custom C scripts were developed to generate this dataset using GPUs and the data was stored in the Hadoop servers at Boston University (www.github.com/fuxmanlab/altered_TFBS).

ChIP-seq allelic imbalance analysis

To estimate optimal threshold(s) of motif scores differences for a given PWM between a reference allele and SNV allele to predict allelic imbalance in TF binding, we used available ChIP-seq experimental data. ChIP-seq experiment FASTQ files were downloaded from the ENCODE Project (Davis et al. 2018) for 14 datasets (55 experiments)

performed in cell lines with normal karyotype (Table 4.1). The files were aligned using BWA (H. Li and Durbin 2009) and pre-processed using standard GATK methodology (Depristo et al. 2011). Variant calling was performed on the aligned BAM files using GATK Variant Discovery pipeline (Depristo et al. 2011) and BCF Tools (H. Li 2011). The intersection of variants from both tools was used to extract the allele read counts for each variant. Allelic imbalance analysis was performed for heterozygous positions in promoters for each experiment. A binomial test was used to identify SNV located in positions were reads were not evenly distributed (0.5 for each allele).

Motif_ID	TF_Name	Family_Name	cell_line	encode_experiment_id	
M4465_1.02	MAX	bHLH	GM12878	ENCSR000DZF	
M4596_1.02	MAX	bHLH	HUVEC	ENCSR000EEZ	
M4481_1.02	USF2	bHLH	GM12878	ENCSR000DZU	
M4479_1.02	TCF12	bHLH	GM12878	ENCSR000BGZ	
M4513_1.02	TCF12	bHLH	H1-hESC	ENCSR000BIT	
M4480_1.02	USF1	bHLH	GM12878	ENCSR000BGI	
M4514_1.02	USF1	bHLH	H1-hESC	ENCSR000BIU	
M4464 1.02	JUND	bZIP	GM12878	ENCSR000DYS	
M4452_1.02	BATF	bZIP	GM12878	ENCSR000BGT	
M4500_1.02	ATF3	bZIP	H1-hESC	ENCSR000BKC	
M4501_1.02	JUN	bZIP	H1-hESC	ENCSR000ECA	
M4591_1.02	JUN	bZIP	HUVEC	ENCSR000EFA	
M4483_1.02	ZEB1	C2H2 ZF	GM12878	ENCSR000BND	
M4469_1.02	REST	C2H2 ZF	GM12878	ENCSR000BQS	
M4508_1.02	REST	C2H2 ZF	H1-hESC	ENCSR000BHM	
M4482_1.02	YY1	C2H2 ZF	GM12878	ENCSR000BNP	
M4516 1.02	YY1	C2H2 ZF	H1-hESC	ENCSR000BKD	
M4430 1.02	CTCF	C2H2 ZF	AG04449	ENCSR000DPG	
 M4431_1.02	CTCF	C2H2 ZF	AG04450	ENCSR000DPM	
 M44331.02	CTCF	C2H2 ZF	AG09319	ENCSR000DPS	
M4436 1.02	CTCF	C2H2 ZF	BJ	ENCSR000DQI	
M4455 1.02	CTCF	C2H2 ZF	GM12878	ENCSR000AKB	
M4456 1.02	CTCF	C2H2 ZF	GM12878	ENCSR000DZN	
M4457 1.02	CTCF	C2H2 ZF	GM12878	ENCSR000DKV	
M4458 1.02	CTCF	C2H2 ZF	GM12878	ENCSR000DRZ	
M4517 1.02	CTCF	C2H2 7F	HA-sp	ENCSR000DSU	
M4518 1.02	CTCF	C2H2 7F	HBMFC	ENCSR000DTA	
M4521 1.02	CTCF	C2H2 7F	HEEpiC	ENCSR000DTR	
M4580 1.02	CTCF	C2H2 7F	HMFC	ENCSR000DUS	
M4583 1.02	CTCF	C2H2 7F	HPAF	ENCSR000DUX	
M4585 1.02	CTCF	C2H2 ZF	HRE	ENCSR000DVH	
M4586 1.02	CTCF	C2H2 ZF	HRPEpiC	ENCSR000DVI	
M4588 1.02	CTCF	C2H2 7F	HSMMtube	ENCSROOOANS	
M4593 1.02	CTCF	C2H2 ZF	HUVEC	ENCSR000DLW	
M4647 1 02	CTCF	C2H2 7F	NH-A		
M4648 1 02	CTCF	C2H2 ZF	NHDE-Ad	ENCSROOOAPM	
M4651 1 02	CTCE	C2H2 7F	NHIF	ENCSROOOANO	
M4654 1 02	CTCF	C2H2 ZF	Osteobl	ENCSROOOAPE	
M4659 1.02	CTCF	C2H2 7F	SAFC	ENCSR000DXI	
M4453 1.02	BCI 11A	C2H2 7F	GM12878	ENCSROOOBHA	
M4484 1.02	7NF143	C2H2 7F	GM12878	ENCSR000D71	
M4454 1 02	BRCA1	FIN3	GM12878	ENCSR000DZS	
M4475 1 02	SPI1	Ets	GM12878	ENCSR000BGO	
M4461 1 02	FTS1	Ets	GM12878		
M4462 1 02	GABPA	Ets	GM12878		
M4595 1 02	GATA2	GATA	HUVEC	ENCSR000EV/W/	
M4473 1 02	PRX3	Homeodomain	GM12878		
M4463 1 02	IRF4	IRF	GM12878	ENCSROOOBGY	
M4466 1 02	MFE2A	MADS hoy	GM12878	ENCSROOOBKR	
M4467 1 02	MEF2C	MADS box	GM12878	ENCSROOOBNG	
M4477 1 02	SRE	MADS box	GM12272	ENCSROOOBGE	
M4512 1 02	SRE	MADS box	H1-hESC		
M4511 1 02	RYRA	Nuclear recentor			
M4468 1 02	RELA		GM12979		
M4478 1 02	STAT2	STAT	GM12878	ENCSR000D7V	
10_1.02	51715	S 101	2141770/0	LINCONDUCLY	

Table 4.1 ChIP-seq experiments downloaded from ENCODE.

Differential binding events were calculated by comparing the motif score of each SNV to its reference allele. Thresholds of two types were generated for gain/disruption of TFBSs to determine their ability to predict ChIP-seq allelic imbalance: 1) when only the reference or alternate allele pass the binding threshold for the motif determined by TFMp-value (Touzet and Varré 2007), or 2) when at least one allele passed the motif binding threshold and the difference in score between alleles (Δ allele score) is above a certain value ranging from 0 to 7. To benchmark our predictions, for each TF, we used SNVs in allelic imbalance in ChIP-seq as true positives and those not in allelic imbalance as true negatives, and compared to predicted gain/loss of TFBSs in the same direction as the allelic imbalance. F-values and relative accuracies were calculated for all thresholds (Figure 4.1). We further selected the first threshold, and motif score differences of two and three from the second type of threshold.



Figure 4.1. ChIP-seq allelic imbalance F-scores versus Δ allele score threshold. Arrows show selected thresholds.

Processing of PCAWG mutational data

We identified coding regions by filtering "coding_regions" of the GENCODE v19 (Harrow et al. 2012) (Jun 14 2018) annotation. Promoters were defined as regions between -2 kb to +250 bp from the transcription start site (TSS) from any protein coding region. In the case of overlapping alternative promoters, we merged the regions to prevent over counting. We used the R package IRanges (M. Lawrence et al. 2013) to determine the promoter coordinates and BEDTools (Quinlan and Hall 2010) was used to remove promoter coordinates overlapping with coding regions (e.g., in cases with genes with alternative promoters). We downloaded VCF files of 2,654 samples of the PCAWG cohort (Campbell et al. 2020) from the ICGC portal (Hudson et al. 2010) (Jan 23 2019) and BEDTools intersection command (Quinlan and Hall 2010) was used to identify SNVs in promoter regions.

Generation and use of the TF-aware burden test

We designed the TF-aware burden test to determine whether the number of observed SNVs in promoter B that lead to gain (or loss) of a binding site for PWM A is more than expected by chance given the total number of mutations observed in promoter B across samples within a certain cancer type. The number of promoter SNVs that create (or disrupt) a binding site for PWM A in promoter B follows a binomial distribution P(n, p), where n is number of SNVs in promoter B across patients, and p is the probability that an SNV in B creates (or disrupts) a binding site for PWM A.

The probability (p) was estimated as equation 1, where F(Bi, Mj) is the probability of

$$p = \sum_{\substack{i=1\\j=1}}^{\substack{i=L\\j=4}} F(B_i, M_j). C(PWM A, B_i, M_j)$$

changing the reference base at position i in promoter B to the mutated base Mj, C(PWM A, Bi, Mj) is 1 if mutating Bi to Mj leads the creation (or distruption) of a binding site for PWM A and 0 otherwise, and L is the nucleotide length of promoter B. F(Bi, Mj) was calculated based on the genome-wide mutational frequencies in a cancer type, whereas C(PWM A, Bi, Mj) was determined by calculating the motif score difference between the sequence surrounding position i for the reference and alternate alleles. These motif scores were obtained by querying the altered TFBS database. We used thresholds obtained from TFMp-value algorithm (Touzet and Varré 2007) to determine whether a motif score is significant, and the three different thresholds selected from the ChIP-seq allelic imbalance analysis. For a given set of SNV samples, we calculated P(n, p) for each PWM- promoter pair and each of these three thresholds independently and corrected for multiple hypothesis testing using FDR. To increase the confidence in our predictions, only PWM-promoter associations that are significant with an FDR < 0.01 using all three Δ score thresholds were considered. Then we selected SNVs from the PCAWG samples (Campbell et al. 2020) located in the significant promoters that were associated with differential score of the corresponding PWM. For predicted driver SNVs, we used the union of significant associated PWM from any of the three thresholds. We used the TFABT for each of the 20 cancer types sample set and a pan-cancer analysis to identify predict driver SNVs.

Computational validation of cancer driver candidates

To determine the pathways associated with the 813 genes with predicted driver NCVs, we used Metascape "Express Analysis" (Zhou et al. 2019) function on this gene set to identify its significantly enriched pathways. In addition, to determine if the 813 genes are enriched in known cancer associated genes, we downloaded the Cancer Gene Census (CGC) list of genes from the COSMIC database (Sondka et al. 2018) (Aug 2 2018) and calculated the odds ratio (OR) for enrichment of the 813 genes in CGC. We also filtered the CGC gene list by the 741 TFs used in this study, to obtain a list of known cancer associated TFs. We determined the enrichment of known cancer associated TF in the 404 TF predicted to be associated with altered binding site (creation/disruption) by the TFABT predicted driver NCVs.

To determine whether our list of predicted driver genes in enriched in essential genes, we used the list essential genes from cancer cell lines from DepMap (Meyers et al. 2017) (May 5 2020) and fitness associated genes from Project Score (fitness genes for three or more cell lines) (Behan et al. 2019) (May 5 2020). We determined the proportion of the 813 predicted driver genes, and CGC genes, which are essential or are fitness related and compared to that of other protein coding genes using a proportion comparison test.

Gene expression levels have been associated with cancer prognostics (favorably/unfavorably) (The Human Protein Atlas, downloaded April 29 2019) (Uhlen et al. 2017). Genes were classified as being associated exclusively with favorable or unfavorable prognostics, or a mix (either) of the two. We determined the enrichment of prognostic associated (favorable, unfavorable, and either) genes in the 813 driver gene set and CGC gene set using a proportion comparison test.

Structural variation has been associated with changes in gene expression. We obtained genes associated with changes in gene expression caused by structural variation across 21 TCGA cohorts (A. Li et al. 2019) (May 25 2020). We filtered this gene set for genes with altered gene expression in more than five cancer types. Similarly, we calculated an enrichment of these genes in the 813 driver gene set and in the CGC genes using a proportional comparison test.

MPRA library construction

The MPRA library was constructed as previously described in Tewhey et al. (Tewhey et al. 2016). Briefly, oligos were synthesized (Agilent Technologies) as 230 bp sequences containing 200 bp of genomic sequences and 15 bp of adaptor sequence on either end. Unique 20 bp barcodes were added by PCR along with additional constant sequence for subsequent incorporation into a backbone vector by Gibson assembly. The oligo library was expanded by electroporation into NEB 10-beta E. coli, and the resulting plasmid library was sequenced by Illumina 2×150 bp chemistry to acquire oligo-barcode pairings. [DB2] The library underwent restriction digestion, and GFP with a minimal TATA promoter was inserted by Gibson assembly resulting in the 200 bp oligo.] sequence positioned directly upstream of the promoter and the 20 bp barcode falling in the 3' UTR of GFP. After expansion within E. coli the final MPRA plasmid library was sequenced by Illumina 1×31 bp chemistry to acquire a baseline representation of each oligo-barcode pair within the library.

MPRA library transfection into cell lines

Jurkat cells were grown in RPMI with 10% FBS to a density 1M cells per mL prior to transfection. HT-29 cells were cultured in Mocoy's 5a media with 10% FBS and SK-MEL-28 in EMEM supplemented with 10% FBS. Six transfection replicates were performed on separate days by collecting 90M cells and splitting across nine 100 uL transfections each containing 10 μ g of MPRA plasmid. Cells were electroporated with the Neon Transfection System (100 μ l kit) using 3 pulses at 1550v for 10ms. After transfection each replicate was split between two T-175 flasks with 150 mL of culture media for

at -80 C for later extraction.

RNA isolation and MPRA RNA-seq library generation

recovery. After 48 hours, the cells were pelleted, washed three times with PBS and stored

RNA for all cell lines was extracted from frozen cell pellets using the Qiagen RNeasy Maxi kit. Half of the isolated total RNA underwent DNase treatment and a mixture of 3 GFP-specific biotinylated primers (#120, #123 and #126) were used to capture GFP transcripts with Streptavidin C1 Dynabeads (Life Technologies). An additional DNase treatment was performed, cDNA synthesized from GFP mRNA using SuperScript III and purified with AMPure XP beads. Quantitative PCR using primers specific for the GFP transcript (#781 and #782) was used to measure GFP transcript abundance in each sample. Replicates within each cell type were diluted to approximately the same concentration based on the qPCR results. Illumina sequencing libraries were constructed using a two-step amplification process to add sequencing adapters and indices. An initial PCR amplification with NEBNext Ultra II Q5 Master Mix and primers 781 and 782 were used to extend adapters. To minimize overamplification during library construction the number of PCR

cycles used in the first amplification was selected based on where linear amplification began for each cell type (Jurkat: 10 cycles, SK-MEL-28 & HT-29: 13 cycles). A second 6 cycle PCR using NEBNext Ultra II Q5 Master Mix added P7 and P5 indices and flow cell adapters. For SK-MEL-28 samples we failed to recover enough product during the first amplification and processed the second total RNA aliquot using the same protocol, pooling the two preparations prior to sequencing. The resulting MPRA RNA-tag libraries were sequenced using Illumina single-end 31 bp chemistry (with 8 bp index read), clustered at 80-90% maximum density on a NextSeq High Output flow cell.

Α	
#120	CCTCGATGTTGTGGCGGGGTCTTGAAGTTCACCTTG/3BioTEG/
#123	CCAGGATGTTGCCGTCCTTGAAGTCGATGCCC/3BioTEG/
#136	CGCCGTAGGTGAAGGTGGTCACGAGGGTGGGCCAG/3BioTEG/
#781	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGCCCTGAGCAAAGACC
#782	ACTCTTTCCCTACACGACGCTCTTCCGATCT
P7	CAAGCAGAAGACGGCATACGAGAT(NNNNNNN)GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
P5	AATGATACGGCGACCACCGAGATCTACAC(NNNNNNN))ACACTCTTTCCCTACACGACGCTCTTCCGATC

В		
	P7 Index	P5 Index
Jurkats Rep 1	AGATGTGC	AACCTCTT
Jurkats Rep 2	TCAGCGAA	CGCATATT
Jurkats Rep 3	GAATTGCT	CTGCTCCT
Jurkats Rep 4	AGGATGTG	GCTGCACT
Jurkats Rep 5	CTAACTGG	CCTGTCAT
Jurkats Rep 6	ACATCCTT	CACTTCAT
HT-29 Rep 1	CTATTCAA	TCCATAAC
HT-29 Rep 2	GACCGAGA	CTGACATC
HT-29 Rep 3	CCTTGCTG	CCTCTAAC
HT-29 Rep 4	CTAGGTTC	CTGGTATT
HT-29 Rep 5	AGCTCTGG	CGAACTTC
Mel-28 Rep1	CCTGGTAG	CAACTGAT
Mel-28 Rep2	CAGTTGGT	GGCAATAC
Mel-28 Rep3	TACTTGCA	CCAACTAA
Mel-28 Rep4	TCGCACCT	CTTCTGGC
Mel-28 Rep5	ACATAGCG	TCCGCATA

Table 4.2 (A) Primers used in MPRA experiments and (B) Illumina Adaptor/Index Primers for Second PCR.

MPRA data analysis

Data from the MPRA was analyzed as previously described (Tewhey et al. 2016). Briefly, the sum of the barcode counts for each oligo were provided to DESeq2 (Love, Huber, and Anders 2014) and replicates were median normalized followed by an additional normalization of the RNA samples to center the average RNA/DNA activity distribution of the 506 negative control sequences over a log2 fold change of zero. This normalization was performed independently for each cell type. Dispersion-mean relationships were modeled for each cell type independently and used by DESeq2 in a negative binomial distribution to identify oligos showing differential expression relative to the plasmid input. Oligos passing a false discovery rate (FDR) threshold of 1% were considered to be active. For sequences that displayed significant MPRA activity, a paired t-test was applied on the log-transformed RNA/plasmid ratios for each experimental replicate to test whether the reference and alternate allele had similar activity. An FDR threshold of 5% was used to identify SNPs with a significant skew in MPRA activity between alleles (allelic skew).

Mutational signatures for MPRA validated drivers

SNVs can be caused by multiple mutational processes such as UV-light or APOBEC activities. We used ICGC probabilities for each SNV-donor combination to assign them a given mutational process if its probability is greater than 0.5 as described (Rheinbay et al. 2020). These processes were used to compare the MPRA validation rate difference between SNVs derived and not derived from a given mutational process. We used UV-light associated signatures (Rheinbay et al. 2020) BI COMPOSITE SNV SBS7a S, BI COMPOSITE SNV SBS7b S, BI COMPOSITE SNV SBS7c S, BI COMPOSITE SNV SBS3 P,

BI_COMPOSITE_SNV_SBS55_S,		BI_COM	POSITE_SN	V_SBS67_S,
BI_COMPOSITE_SNV_SBS75_S	and	APOBEC	related	signatures
BI_COMPOSITE_SNV_SBS2_P,		BI_COM	POSITE_SN	V_SBS13_P,
BI_COMPOSITE_SNV_SBS69_P.				

Normalized gene expression analysis

We downloaded aligned BAM files corresponding to 1,366 samples from ICGC. BAM files were converted to FASTQ files using the SAMtools fastq (H. Li et al. 2009) function. Then, we used Salmon (Patro et al. 2017) to quantify the expression of the human transcriptome (Esembl, May 30 2019) in transcripts per million (TPM). We summed the expression of each gene transcript to obtain the gene TPM expression.

We calculated a reference expression value for each gene-cancer type combination based on the median TPM expression across donors who do not have any mutation in the gene promoter. For each donor and gene with a predicted driver NCV in its promoter, we calculated the normalized TPM expression as $log10(\frac{donor-gene TPM}{reference gene-cancer type TPM})$, where values greater than 0 are associated with overexpression and values less than 0 with underexpression of genes with predicted driver SNVs. This analysis resulted in a dataset of normalized expression values for gene-donor pairs associated with predicted drivers.

Association of creation and disruption of TFBS with target gene expression

To estimate optimal threshold(s) of motif scores differences for a given PWM between a reference al Predicted driver NCVs in gene promoters may alter binding of multiple TFs. For each cell line, we determined the transcriptional effect (from MPRA) of NCVs associated with the creation and disruption of a given TF and calculated the activation ratio of TF activation/TF repression:

#SNVs create TFBS and transcript overexpression+ #SNVs disrupt TFBS and transcript underexpression+1 #SNVs create TFBS and transcript underexpression+ #SNVs disrupt TFBS and transcript overexpression+1

We selected TFs that had a log10(ratio) greater than 0.5 in at least two of the three cell lines, which will suggest these TFs may act as activators. We determined the transcriptional effect of activator TFs by comparing the normalized expression of genes with associated driver SNVs leading to a TFBS creation and compared its distribution to a μ =0 (no effect) using a Kruskal Wallis test. This determined changes in gene expression associated with the presence of SNVs affecting activator TFs. A similar approach was used for genes with associated driver NCVs leading to a given TFBS disruption. This analysis associated the effect of creation or disruption of a TFBS with the over or underexpression of its gene targets respectively.

Cell culture for CASCADE experiments

The cell lines used for CASCADE experiments were obtained from ATCC. Three cell lines were used for the CASCADE experiments: Jurkat (ATCC TIB-152), an acute T cell leukemia cell line, SK-MEL28 (ATCC HTB-72), a malignant melanoma cell line, and HT-29 (ATCC HTB-38), a colorectal adenocarcinoma cell line.

Jurkat cells were grown in suspension in RPMI 1640 Glutamax media (Thermofisher Scientific, Catalog #72400120) with 10% heat-inactivated fetal bovine serum (Thermofisher Scientific, Catalog #11360070) and 1mM sodium pyruvate (Thermofisher Scientific, Catalog #16140071). T175 (Thermofisher Scientific, Catalog #132903) non-treated flasks were used when culturing JURKAT cells for experiments. Cells were grown in 50mL of media when being cultured in T175 flasks. 3 T175 flasks, or 100 million Jurkat cells, were used for each nuclear extraction.

SK-MEL28 cells were grown in Eagle's Minimum Essential Medium (EMEM) (ATCC, Catalog #ATCC-30-2003) with 10% heat-inactivated fetal bovine serum. T225 treated flasks for adherent cells (Corning, Catalog #353138) were used when culturing SK-MEL28 cells for experiments. Cells were grown in 40mL of media when being cultured in T225 flasks. 3 T225 flasks, or 60 million SK-MEL28 cells, were used for each nuclear extraction.

HT-29 cells were grown McCoy's 5A Medium (EMEM) (ATCC, Catalog #ATCC-30-2007) with 10% heat-inactivated fetal bovine serum. T225 treated flasks for adherent cells (Corning, Catalog #353138) were used when culturing HT-29 cells for experiments. Cells were grown in 40mL of media when being cultured in T225 flasks. 3 T225 flasks, or 60 million HT-29 cells, were used for each nuclear extraction.

CASCADE protein binding microarray experiments

The nuclear extract protocols are as previously described (P. Zhang et al. 2018). Changes to the previously published protocols are detailed. To harvest nuclear extracts from Jurkat cells, the cells were collected in falcon tubes. The cells were pelleted by centrifugation at 500xg for 5 min at 4°C. The media was aspirated off, taking care to not disturb the pellet. The cell pellet was washed once with 1X PBS and 0.1mM Protease Inhibitor (Sigma-Aldrich, Catalogue #P8340) and centrifuged again at 500xg for 5 min at 4°C. The 1X PBS and 0.1mM Protease Inhibitor was aspirated off. The cell pellet was placed on ice.

To harvest nuclear extracts from SK-MEL28 and HT-29 cells, the media was aspirated off and the cells were washed once with 1X PBS. Once the 1X PBS used to wash the cells was aspirated off, enough 1X PBS was mixed with 0.1mM Protease to cover the cells was added to each flask. A cell scraper was then used to dislodge the cells from the flask. The cells were collected in a falcon tube and placed on ice. To pellet the cells, the cell volume was centrifuged at 500xg for 5 min at 4°C. The cell pellet was placed on ice.

Once the cells were pelleted, the supernatant was aspirated off. The pellet was resuspended in Buffer A and incubated for 10 min on ice (10mM HEPES, pH 7.9, 1.5mM MgCl, 10mM KCl, 0.1mM Protease Inhibitor, Phosphatase Inhibitor (Santa-Cruz Biotechnology, Catalogue #sc-45044), 0.5mM DTT (Sigma-Aldrich, Catalogue #4315)) to lyse the plasma membrane. After the 10 min incubation, a final concentration of 0.1% Igepal detergent was added to the cell and Buffer A mixture and vortexed for 10 sec. To separate the cytosolic fraction from the isolated nuclei, the sample was centrifuged at 500xg for 5 min at 4°C. The cytosolic fraction was collected into a separate microcentrifuge tube. The pelleted nuclei were then resuspended in Buffer C (20mM HEPES, pH 7.9, 1.5mM MgCl, 0.2mM EDTA, 0.1mM Protease Inhibitor, Phosphatase Inhibitor, 0.5mM DTT, and 420mM NaCl) and then vortexed for 30 sec. The nuclei were incubated in Buffer C for 1 h while mixing at 4°C. To separate the nuclear extract from the nuclear debris, the mixture was centrifuged at 21,000xg for 20 min at 4°C. The nuclear extract, they were

desalted using Zeba Spin Desalting Columns (ThermoFisher Scientific, Catalog #89882). Prior to flash freezing the nuclear extracts, glycerol was added to the nuclear extracts to reach a final concentration of 5%. Nuclear extracts were stored at -80°C until used for experiments.

Microarray DNA double stranding and PBM protocols are as previously described (Shi et al. 2016; Valouev et al. 2008a; P. Zhang et al. 2018). Any changes to the previously published protocols are detailed. Double-stranded microarrays were pre-wetted in HBS (20mM HEPES, 150mM NaCl) containing 0.01% Triton X-100 for 5 min and then dewetted in an HBS bath. Next the array was incubated with nuclear extract for 1 h in the dark in a binding reaction buffer (20mM HEPES, pH 7.9, 100mM NaCl, 1mM DTT, 0.2mg/mL BSA, 0.02% Triton X-100, 0.4mg/mL salmon testes DNA (Sigma-Aldrich, Catalogue #D7656)). The array was then rinsed in an HBS bath containing 0.1% Tween-20 and subsequently de-wetted in an HBS bath. After the protein incubation, the array was incubated for 20 min in the dark with 20µg/mL primary antibody for the TF or COF of interest (Supplementary Table 1). The primary antibody was diluted in 2% milk in HBS. After the primary antibody incubation, the array was first rinsed in an HBS bath containing 0.1% Tween-20 and then de-wetted in an HBS bath. Microarrays were then incubated with 10µg/mL of either alexa488 or alexa647 conjugated secondary antibody (see Supplementary Table 1) for 20 min in the dark. The secondary antibody was diluted in 2% milk in HBS. Excess antibody was removed by washing the array twice for 3 min in 0.05% Tween-20 in HBS and once for 2 min in HBS in coplin jars as described above. After the washes, the array was de-wetted in an HBS bath. Microarrays were scanned with a GenePix

4400A scanner and fluorescence was quantified using GenePix Pro 7.2. Exported fluorescence data were normalized with MicroArray LINEar Regression (Shi et al. 2016).

CASCADE-based differential COF recruitment microarray design

We obtained matching survival A high-throughput array-based screen was designed to profile differential COF recruitment to the 2,555 predicter driver NCVs, and 500 no predicted binding NCVs in 26-base DNA probe target regions centered at the SNP position (relative to + strand: 13 bases + SNV location + 12 bases) were obtained for each reference (REF) allele using BEDTools (Quinlan and Hall, 2010). For each REF allele probe, a probe with the corresponding SNV allele was also included in the design such that each of the comparisons above is represented by a pair of REF and SNV probes. For 1,523 of the predicted driver NCVs and 767 no predicted binding SNVs (sampled randomly from each full category above), additional REF/SNV probes were included by shifting the variant position -5 and +5 bases within the target region of the probe such that each of these comparisons were represented in three total registers. The 26-base target regions were embedded in larger 60-base PBM DNA probes as follows:

"GCCTAG" 5' flank – 26-base target region – "CTAG" 3' flank – "GTCTTGATTCGCTTGACGCTGCTG" double-stranding primer

5 replicates of each probe in both the reference (+) orientation and reverse (-) orientation were included in the final design. The microarrays were purchased from Agilent Technologies Inc. (AMAID: 085920, format: 8×60 K).

Analysis of differential COF recruitment

Each REF/SNV pair was screened for differential COF recruitment and experimental results were preprocessed as above. Z-scores were obtained for each probe as previously described (Bray et al., 2020) against the distribution of fluorescence intensities obtained at the set of variant-centered no predicted binding probes for a given experiment. Differential COF recruitment statistics were computed as previously described (Bray et al., 2020). Briefly for each REF and SNV allele pair in the design, a t-test was used to compare the fluorescence intensity distributions between the 5 REF probes and 5 SNP probes for a given COF assayed. To mitigate the influence of probe orientationspecific effects, t-tests were performed independently for each probe orientation with the p-values combined using Fisher's method. For the select sites included in three registers (see above), these t-tests were performed across each orientation and each register shift independently with the p-values combined using Fisher's method as above The Benjamini-Hochberg method was used to adjust the individual p-values for each REF/SNV pair across the total number of to account for multiple hypothesis testing. Differential COF recruitment was deemed statistically significant if the adjusted p-value (q-value) was below 0.05. The fluorescence intensity z-score difference for a given REF and SNV allele probe pair (termed Δz -score) was computed as previously described (Bray et al., 2020). Briefly, Δz scores were computed by subtracting the mean REF z-score from the mean SNV z-score such that a positive Δz -score represents a gain-of-recruitment introduced by the SNV allele and a negative Δz -score represents a loss.

Results

Prediction of noncoding cancer driver SNVs

We developed a novel TF-aware burden test (TFABT) that identifies gene promoters containing higher than expected number of SNVs across patients that create (or disrupt) a TFBS for a particular TF (741 TFs were tested). For each TF-promoter (A,B) pair, the method uses a binomial distribution P(x, n, p) to calculate the FDR for the observed number of SNVs in promoter B creating (or disrupting) interactions with TF A (x) given the total number of observed SNVs in promoter B (n) all patient samples from a cancer type and the probability that a random SNV in B creates (or disrupts) a binding site for TF A (p).

We applied the TFABT to predict cancer driver NCVs in gene promoters using 2,654 tumor samples from the PCAWG cohort corresponding to 20 cancer types (Campbell et al. 2020). Driver predictions were performed per cancer type and in a pancancer analysis. In total, we predicted 2,555 candidate driver NCVs in the promoters of 813 genes, which create/disrupt binding sites of 404 TFs. The majority of predicted driver NCVs were obtained from skin cancer (Figure 4.2A). This is not only related to skin cancer samples having the largest number of NCVs but also to a higher percentage of those NCVs being predicted drivers (Figure 4.3). The majority of predicted driver NCVs (76%) are associated with the disruption of existing TFBSs. This is likely related to a higher probability of disrupting a TFBS over its creation in cis-regulatory regions or to the disruption of a TFBS having a higher likelihood of being functional.



Figure 4.2 Driver NCVs prediction and their association with cancer genes and pathways. (A) Number of significant NCVs with predicted gain and/or loss of TF binding per cancer type. (B) Genes with the most predicted cancer driver NCVs and the percent of patients affected per cancer type. (C) Metascape network showing the intra-cluster and inter-cluster similarities of enriched gene ontology terms for genes with significant NCVs. (D) Fraction of essential and fitness related genes for genes with predicted NCVs, in CGC, or all protein-coding genes. (E) Fraction of genes with predicted NCVs, in CGC, or all protein-coding genes. (E) Fraction of genes with predicted NCVs, in CGC, or all protein-coding genes.

We identified driver NCVs in multiple genes with reported driver NCVs. For example, we identified 16 candidate driver NCVs in the promoter of TERT, which included

the two frequently mutated NCVs in chr5:1295228 C>T and chr5:1295250 C>T (Susanne

Horn et al. 2013; Huang et al. 2013). A large fraction of bladder (65%), skin (47%), and head/neck (17%) cancer samples contain at least one of TERT candidate driver NCV (Figure 4.2B). In addition, we predicted eight candidate driver SNVs in the promoter of PLEKHS1, including two previously reported mutations in chr10:15511590 C>T and chr10:115511593 C>T (Rheinbay et al. 2017). These candidate driver NCVs were found in 39% of bladder cancer samples, with no other cancer type having more than a 5% frequency (Figure 4.2B). Furthermore, the TFABT identified previously known drivers in ALDOA, DPH3, CCDC107, LEPRROTL1, and TBC1D12 (Rheinbay et al. 2017; Denisova et al. 2015). Novel driver candidate SNVs in lymphoid cancers were predicted in the BCL6 and BCL2 promoters, which were found in 23% and 21% of lymphoid cancer samples, respectively. Finally, predicted driver SNVs in RPL13A, C16orf59, CDC20, OXNAD1, PES1, and TRMT10C were found in skin cancer samples with frequencies between 21-33% (Figure 4.2B).



Figure 4.3. Number of predicted cancer driver NCVs and number of SNVs by cancer type.

We found multiple lines of evidence showing our predicted driver gene set is associated with known cancer related genes, pathways, and functions. First, our predicted driver gene set is enriched in gene ontologies associated with general and cancer related cellular processes such as cell cycle, TP53 regulation, Wnt signaling, epithelial-mesenchymal transition, and mitochondrial apoptosis (Figure 4.2C). Second, we found a significant enrichment of genes from the Cancer Gene Census (CGC) (Sondka et al. 2018) genes in the 813 promoter genes (OR=1.54, p=0.008) and their 404 associated TFs (OR=2.3, p=2x10-4). Third, we found a significant enrichment of our predicted driver genes in cellular fitness genes (Figure 4.2D) (Meyers et al. 2017), essential genes (Figure

4.2D) (Behan et al. 2019), and genes whose expression has been associated with favorable or unfavorable cancer prognosis (Uhlen et al. 2017), comparable to those of CGC genes (Figure 4.2E). Finally, we identified a significant overlap of predicted driver genes and a set of genes whose somatic copy number variation are associated with changes in their expression in multiple cancer types (OR=1.42, p=0.007) (A. Li et al. 2019). These results suggest that our predicted drivers are likely to be functional.

TF-aware driver candidate NCVs lead to altered transcriptional activity

To investigate the effect of the predicted driver NCVs on transcriptional activity, we used MPRAs (Tewhey et al. 2016) to systematically test the 2,555 predicted driver NCVs and control NCV sets in HT-29 (colorectal), Jurkat (lymphoma) and SK-MEL-28 (melanoma) cell lines. Since only a subset of DNA regions show MPRA activity for either NCV allele, we calculated the validation rate as the ratio of NCVs displaying allelic skew over the total number of active DNA regions for each NCV category. For the TF-aware predicted driver NCVs, we obtained validation rates of 33%, 53% and 27% for HT-29, Jurkat, and SK-MEL-28, respectively, higher than the percentage of NCVs with no predicted differential TF binding or no predicted TF binding that display allelic skew (Figure 4.4A, Supplementary Figure 4.5A-B). Further, 235 predicted drivers were validated across the three cell lines, and 21, 320 and 12 predicted drivers were validated exclusively in HT-29, Jurkat and SK-MEL-28 cell lines (Supplementary Figure 4.6). The high validation rates from the predicted driver NCVs are similar to experimentally characterized driver NCVs in promoters (literature), NCVs leading to allelic imbalance in ChIP-seq experiments, and disease-associated germline NCVs that lead to altered target gene expression and cause differential TF binding (germline) (Figure 4.4A). This shows that the TFABT can prioritize functional NCVs.



Figure 4.4 Predicted driver NCVs can alter transcriptional acitvity. (A) Validation rate versus q-value threshold in SK-MEL-28 for predicted driver NCVs, ChIP-seq allelic imbalance, known drivers, MPRA positive controls, germline NCVs, literature genes, no significant differential binding, no differential binding. (B) Validation rate vs q-value in SK-MEL-28 for predicted NCVs based on whether NCV caused gain, loss of TFBS or both. (C) Fraction of NCVs per frequency in patient samples. (D) Fraction of MPRA validated NCVs for genes with at least four transcriptionally active NCVs by NCV effect (up/downregulation) in each of the three cell lines.

We validated NCVs associated with both gain and loss of TFBSs. However, we observed a higher validation rate for NCVs that loose TFBSs than for NCVs that gain TFBSs or bifunctional NCVs (Figure 4.4B). This difference may be related to a higher likelihood of affecting expression by disrupting an existing TFBS in a regulatory region than by creating a TFBS that may not be in the appropriate regulatory region context or distance/orientation to other TFBSs to affect transcriptional activity. Importantly, we found that the validation rate for predicted driver NCVs is similar regardless of the NCV frequency across cancer samples (Figure 4.4C). This suggests that NCVs with low mutation frequency, such as those private to particular tumor samples, can also lead to altered transcriptional activity.



Figure 4.5 MPRA validation rate. Validation rates versus q-value for (A) Jurkat and (B) HT-29 cell lines for the categories referenced in figure 4.3A.

Multiple NCVs in a gene promoter often lead to the same transcriptional effect (over or underexpression). For example, all validated NCVs in the TERT promoter lead to increased transcriptional activity, consistent previously characterized TERT promoter drivers associated with TERT overexpression (Susanne Horn et al. 2013; Huang et al. 2013). Conversely, two MPRA validated predicted driver NCVs in the RNF20 promoter (chr9:104296044 C>T and chr9:104296134 G>A) display reduced transcriptional activity (Figure 4.4D). RNF20 underexpression due to promoter hypermethylation has been previously associated with genome instability in multiple cancer types (Guppy and McManus 2017; Nakamura et al. 2011; Shema et al. 2008). Our results suggest that reduced RNF20 promoter activity resulting from NCVs constitutes another potential cancer mechanism (Figure 4.4D). Other examples include skin cancer associated genes PARS2, GOSR2, and MBD3L1 whose promoter SNVs lead to reduced transcriptional activity (Figure 4.4D); however, they have not been previously associated with skin cancer.



Figure 4.6. Three-way Venn diagram displaying the number of MPRA validated NCVs for HT-29, Jurkat and SK-MEL-28 cell lines

Diver NCVs outside core promoter may affect transcriptional activity

Most driver NCVs have been identified and characterized in core promoter regions (-250bp to +250bp from the TSS) (Rheinbay et al. 2017; 2020). Here, we used extended promoter regions of -2kb to +250bp from the TSS, expanding the current landscape of analysis. Although the fraction of NCVs in PCAWG is mostly homogenous throughout the extended promoter region, we observed an enrichment of predicted driver NCVs in the core promoter, even though our model did not incorporate any additional information beyond TF specificities and promoter sequence (Figure 4.7A). This suggests, that considering core promoter regions likely identifies most driver NCVs in gene promoters. Nevertheless, we detected MPRA-validated NCVs beyond the core promoter (upstream of -250 from TSS) accounting for 25.8% of validated driver NCVs. For example, the

lymphoid cancer associated NCV chr18:60988772 A>G in the BCL2 promoter is located at -1441bp from the TSS and leads increase transcriptional activity in Jurkat cells. In addition, we identified the chr5:137799888 G>C NCV located at position -1291 from the EGR1 TSS that causes reduced transcriptional activity. Underexpression of the tumor suppressor gene EGR1 has been previously reported in multiple cancer types (Baron et al. 2006; Ferraro et al. 2005). Further, overexpression of USP37 has been previously associated with higher mortality rate in breast cancer (Qin et al. 2018), the chr2:219365001 located at position -1865 from its TSS was shown to cause increased transcriptional activity in Jurkat cells. These results suggest that NCVs located further from the commonly studied core promoter can also alter target gene expression.


Figure 4.7 NCV validation rate by TSS distance and mutational signature type. (A) Validation rate of predicted driver NCVs in SK-MEL-28 by genomic distance to TSS, and fraction of NCVs per 100 bp for predicted driver NCVs, MPRA active NCVs and SNVs in the PCAWG cohort. (B) Validation rate for NCVs associated or not with APOBEC mutational processes for the three cell lines. (C) Validation rate of predicted driver NCVs associated or not with UV-light mutational signature in SK-MEL-28.

NCVs derived from mutational processes can affect transcriptional activity

Somatic mutations in cancer are caused by endogenous and exogenous mutational processes, that differ between patients and cancer types leading to different mutational signatures (Alexandrov et al. 2013; 2020). We analyzed the transcriptional activity of predicted driver NCVs derived from two mutational signatures frequently excluded from

mutational burden tests: 1) defective apolipoprotein B messenger RNA-editing enzyme catalytic (APOBEC) cytidine deaminases that share a common mutational context of C>G or C>T at TCT and TCA (Alexandrov et al. 2020), and 2) UV-light associated mutational signatures consisting mainly of C>T at TCN and C>T at CCN (Alexandrov et al. 2020). We found no significant difference between the validation rate in MPRAs between APOBEC+ and APOBEC- NCVs (Figure 4.7B). Importantly, UV-light+ predicted driver NCVs validate in MPRAs in SK-MEL-28 cells at a higher rate than UV-light- NCVs (29% versus 17%, p=0.003) (Figure 4.7C). This is particularly important given that 86% of the predicted driver NCVs in MPRA-active regions in SK-MEL-28 cells are derived from the UV-light+ signature. Even though previous studies have filtered out NCVs derived from cancer associated mutational processes such as APOBEC and UV-light to increase statistical power of their analysis (Rheinbay et al. 2017; 2020), the similar or higher MPRA validation rate of predicted driver NCVs suggest that a significant fraction of these NCVs have functional activity.

Transcription factors and their effect in transcriptional activity

We further analyzed the 404 TFs involved in the predicted altered TF binding caused by the 2,555 candidate driver NCVs. We found that in the majority of cancer types four TF families are mainly involved (Figure 4.8A). Predicted driver NCVs in skin, head/neck, kidney, bone/soft tissue and pancreas cancers affect the binding sites of Ets factors, a TF family that has been largely associated with multiple cancer types (Bell et al. 2016; 2015; Yinghui Li et al. 2015). In contrast, predicted driver NCVs in cervix, uterus, lymphoid, and stomach cancer affect mostly Forkead binding sites; whereas breast, liver,

lung, and prostate cancer NCVs affect Homeodomain binding sites. These differences are likely related to the different mutational signatures between cancer types that result in altered binding of different TF families. Interestingly, we observed a higher validation rate in MPRAs for predicted driver SNVs altering binding sites of TF from the nuclear receptors (NR), Ets, and BHLH families (Figure 4.8B). Whether this reflects what occurs in the endogenous loci remains to be determined.



Figure 4.8 Transcription factor effect on transcriptional activity. (A) Fraction of TF families with altered TFBS caused by predicted driver NCVs by cancer type. (B) MPRA validation rate in SK-MEL-28 versus q-value for TF families. (C) Normalized TPM of genes with predicted driver NCVs by TFs associated with gain/overexpression and loss/underexpression.

TFs can activate or repress target gene expression, with some TFs acting mainly as activators and others mainly as repressors. To investigate the effect of predicted driver NCVs on their target gene expression, we normalized the expression levels of genes from donors with a predicted driver NCV to the median expression of those without any NCV in the corresponding gene promoter. This analysis identified 20 TFs whose gain of binding sites are associated with increased transcriptional activity and whose disruption of binding site is associated with reduced transcriptional activity (Figure 4.8C). Interestingly, we identified ten Ets, one nuclear receptor, and one STAT TFs whose creation of binding sites is associated with a significant increase of their target genes expression in patient samples. In contrast, we did not find any significant association of TFBS disruption and underexpression of target genes. This lack of significance may result from a lack of sensitivity due to gene expression for the wild type allele or due to compensatory mechanisms. In total, we identified 319 genes containing predicted driver NCVs that create or disrupt Ets binding sites. Changes in expression, alternative isoforms, or gene fusions involving multiple Ets factors have been associated with cancer. The creation or disrution of Ets TFBSs likely constitutes another widespread cancer mechanism that can also be modulated by the previously reported changes in Ets activities.

Predicted driver NCVs lead to differential cofactor recruitment

To determine the effect of our predicted driver NCVs on differential cofactor (COF) recruitment, we used the CASCADE platform (Bray et al. 2020), a protein binding

microarray-based method that allows for high-throughput profiling of COF recruitment on reference and mutant NCV pairs using nuclear extracts. We used CASCADE to study the effect of NCVs in two specifications: 1) single register, where the reference/mutant NCV alleles are located in the middle of the probe, and 2) triple register, where the reference/mutant NCV alleles are tested in the top, center and bottom of the probe. We tested the predicted drivers and controls in a single register array in SK-MEL-28 cells for differential recruitment of p300, P300 + peptides, SMARCA4 TBL1XR1, HDAC1, HDAC3, RBBP5, SKP2, and GCN5. Overall, we observed a similar or higher validation rate for predicted driver NCVs compared to other positive controls such as known driver NCVs and ChIP-seq allelic imbalance, and we found a low validation rate for negative controls such as NCVs with no predicted or no differential TF binding. These results show that the predicted driver NCVs are associated with differential cofactor recruitment.

After filtering for reference or mutant NCV alleles above background fluorescent intensity (z-score > 2), we observed a high validation rate of 49% of predicted driver NCVs for differential recruitment of TBL1XR1 in SK-MEL-28 (Fig 4.9A). These validated NCVs show a trend to disrupt TBL1XR1 recruitment. However, the handful of driver NCVs showing a gain of TBL1XR1 recruitment are associated with fifteen cancer types may serve as therapeutics candidates. Interestingly, vorinostat has been shown to inhibit TBL1XR1, it has been approved to treat cutaneous T-cell lymphoma , and clinical trial are active for breast and skin cancer types (Munster et al. 2011; Haas et al. 2014). This raises the possibility of using vorinostat as a therapeutic opportunity to treat patients carrying these mutations. In addition, we found a 7.5% validation rate of predicted drivers on

P300+peptides in SK-MEL-28 (Fig 4.9B). Similarly, P300 inhibitor, A-485, has been shown to inhibit tumor growth in multiple lineage-specific tumors including hematological malignancies and androgen receptor-positive prostate cancer (Lasko et al. 2017), and upregulates apoptosis in non-small-cell lung carcinoma cells in combination with TRAIL (B. Zhang et al. 2020).



Figure 4.9 Predicted drivers cause differential COF recruitment. Differential COF recruitment for predicted driver NCVs and no predicted binding NCVs (validation rate) showed as Δz -score versus -log10(q-value) in SK-MEL-28 for (A) TBL1XR1 and (B) P300 + peptides, where dotted line on y-axis represents significance threshold and on x-axis no differential COF recruitment (0 Δz -score). Significance values, -log10(q-value), for 3 register versus 1 register array for predicted driver NCVs and no predicted binding (validation rate) for (C) TBL1XR1, (D) P300 + peptides, (E) SKP2, and (F) P300. Dotted lines represent significance thresholds.

Furthermore, we used 3-register probes for 768 randomly selected predicted drivers and 384 NCVs with no predicted TFBS. Interestingly, we observed an increase in validation rate of 3X and 4X for TBL1XR1 and p300 + peptides recruitment (Figures 4.9C-D), respectively. Moreover, the 3-register probes were able to validate 5.2% and 2.9% of predicted drivers for SKP2 and p300, COFs that show less than 0.5% or no validation rate in the 1-register probes (Figures 4.9E-F). This is because NCVs location and orientation in the array probes may have distinct effects of TF and COF binding, and using three registers increase the likelihood of detecting altered COF recruitment. Importantly, NCVs with no predicted TFBS showed validation rates no greater than 1% in the 3-register probes. These results support the use of 3-register probes to boost the validation rate for predicted driver NCVs without leading to high false positives for negative controls.

Discussion

In this study, we developed a novel TFABT based on the hypothesis that creating (or disrupting) a TFBS at different positions within a gene promoter is likely to lead to similar effects on target gene expression. The TFABT identifies gene promoters containing higher than expected number of NCVs across patients that create (or disrupt) a TFBS for a particular TF based on a binomial test. We applied the TFABT to predict cancer driver SNVs in gene promoters using 2,654 tumor samples from the PCAWG cohort corresponding to 20 cancer types. Driver predictions were performed per cancer type and in a pan-cancer analysis. In total, we predicted 2,555 driver candidates in the promoters of 813 genes, which create/disrupt binding sites for 404 TFs. These genes included known drivers with NCVs such as TERT, ALDOA, CCDC107, LEPRROLT1, and TBC1D1.

Further, we showed multiple lines of evidence suggesting that the predicted genes and associated with known cancer related genes/TF, pathways and gene functions.

By testing the predicted driver NCVs in MPRAs in three cell lines, we found a validation rate similar or greater that known cancer driver NCVs in promoters, ChIP-seq allelic imbalance NCVs and germline NCVs associated with altered gene expression and TF binding. These results show that the TFABT can prioritize transcription perturbing NCVs. Moreover, we show that NCVs private to one sample, which constitute the majority of NCVs in the PCAWG cohort, are similarly likely to alter transcriptional activity as recurrent NCVs. These MPRA validated cancer driver NCVs greatly expand the current known drivers in literature. However, the effect of multiple NCVs located in the same regulatory region (i.e promoters, enhancers) remains to be studied. We showed that most predicted driver NCVs are located in the core promoter of a gene, which suggests that considering the core promoter regions as most other studies have done, likely identifies most drivers NCVs in promoters. Conversely, predicted driver NCVs derived from APOBEC and UV-light mutational processes show transcriptional perturbing activity, even though multiple studies filter out these types of NCVs. Further, UV-light predicted driver NCVs validate at a higher rate in MPRAs compared to non UV-light predicted driver NCVs. This suggests that excluding NCVs from burden tests based on mutational signatures may not be warranted.

Our validation using MPRAs shows that many of the potential driver NCVs identified alter transcriptional activity in an episomal construct. Whether, these NCVs alter gene expression in the endogenous locus and whether this leads to cellular change

associated with cancer phenotypes (e.g., increase proliferation, reduced apoptosis, etc.) remains to be determined. As NCV drivers have low mutational frequency, available cohorts, in most cases, lack statistical power to determine the link between NCVs and its target gene expression. Therefore, larger studies cohorts integrating WGS and RNA-seq will allow in-vivo validation of NCV drivers.

We observed a higher validation rate for NCVs associated with differential binding of nuclear receptors and Ets factors. These is consistent with the known role of Ets factors in cancer initiation and progression (Bell et al. 2016; 2015; Yinghui Li et al. 2015). Even though only a small fraction of predicted driver NCVs affect nuclear receptor binding sites, we validated driver NCVs associated with NR113 and VDR in lymphoma, and NR2C2 in skin cancer, which have known antagonists and agonists. These druggable TFs as well as the cofactors found to be differentially recruited to the NCVs using CASCADE provide a therapeutic opportunity to restore normal target gene expression in cells carrying the corresponding NCVs.

Author contributions

J.I.F.B. and S.C.P. conceived the project and wrote the manuscript. S.C.P., K.B. and B.G. generated the aTFBS-DB. J.I.F.B., A.L, and S.C.P created the TFABT. R.T. generated the MPRA data. T.S., H.J.H., and D.B. generated the CASCADE data. S.C.P performed the data analysis supervised by J.I.F.B and A.L..

Chapter 5. Conclusions

In this dissertation, I have discussed the application of literature curation and novel bioinformatics algorithms to study transcriptional regulation in health and disease. Specifically, in chapter 2 I demonstrated how mining three decades of knowledge can be used to generate a comprehensive mouse and human cytokine GRN, CytReg, with 2-3-fold more TF-cytokine gene interactions than other available databases. CytReg was implemented as a user-friendly database (https://cytreg.bu.edu) where PDIs can be easily browsed by TF, cytokine, species, assay type, and TF expression pattern, then visualized as a table or an interactive network. The integrative analysis of the cytokine GRN and other functional datasets provided insight into the general principles governing cytokine regulation, such as a correlation between TF connectivity in the cytokine GRN and immune phenotype. By characterizing the TFs and cytokines studied in the last three decades, we found biases towards specific TFs/cytokines in the literature and highlight the incompleteness of the cytokine GRN. Further, by using cytokine co-expression data and TF motif analysis, we predicted novel TF-cytokine promoter interactions that were validated with eY1H assays. This exemplifies how the integrative analysis of CytReg can be used to prioritize interaction candidates to validate experimentally. Ultimately, the integration of different high-throughput and unbiased approaches, population-wide studies of regulatory variation, and in-depth functional characterizations of the regulatory logic

will lead to a more comprehensive picture of cytokine regulation in different cell types, conditions, and individuals.

In chapter 3, I discussed the predicted genome-wide effects of SNVs on TFBS. We created a database of altered TFBS (aTFBS-DB) by calculating the effect (gain/loss) of all possible SNVs across the human genome for 741 TFs. The aTFBS-DB was used to determine "gainability" and "disruptability" scores for each TF in gene regulatory regions. We established that TFBS for bZIP, C2H2 ZF, nuclear receptors and T-box families are less likely to be altered by SNVs, whereas forkhead and homeodomain families show higher rewiring potential by their higher gainability and disruptability scores. However, whether in vivo binding site occupancy for these TFs is actually rewired across evolution or between individuals in the human population remains to be determined. By calculating gainability and disruptability scores for functional cis-eQTL SNVs and common SNVs in the human population, we determined that cis-eQTL are more likely to perturb TFBS. Altogether, this database provides a blueprint to study the impact of SNVs on genetic variation.

In chapter 4, I described how we used the aTFBS-DB to develop the TFABT, a novel algorithm to predict cancer driver NCVs in promoters. We applied the TFABT to a the PCAWG cohort of 2,654 samples across 20 cancer types and predicted 2,555 driver NCV candidates located in 813 genes that alter the binding of 404 TFs. Importantly, we identified known drivers in TERT, ALDOA, CCDC107, LEPRROLT1, and TBC1D1 and presented multiple lines of evidence suggesting the predicted genes are associated with known cancer related genes/TFs, pathways, and gene functions. By testing the predicted

drivers using MPRAs in three cell lines, we achieved a validation rate of transcriptional activity similar to or greater than known cancer driver NCVs in promoters, ChIP-seq allelic imbalance NCVs, and germline NCVs associated with altered gene expression and TF binding, showing that the TFABT can prioritize transcription perturbing NCVs. Moreover, we establish that NCVs unique to one sample, which constitute the majority of NCVs in the PCAWG cohort, are similarly likely to alter transcriptional activity as recurrent NCVs. We further identified differential COF recruitment caused by the predicted drivers using CASCADE. The study in this chapter demonstrates the functional and biophysical impact of driver NCVs and can be used as the foundation to develop novel methodologies to predict the functional impact of NCVs in distal regulatory elements.

Taken together, this thesis provides a framework to study transcriptional mechanisms in cellular processes, such as cytokine expression, and the effects of their dysregulation in diseases such as cancer.

BIBLIOGRAPHY

- Aerts, Stein, Peter Van Loo, Gert Thijs, Herbert Mayer, Rainer de Martin, Yves Moreau, and Bart De Moor. 2005. "TOUCAN 2: The All-Inclusive Open Source Workbench for Regulatory Sequence Analysis." *Nucleic Acids Research* 33 (SUPPL. 2). https://doi.org/10.1093/nar/gki354.
- Aguet, François, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, et al. 2017. "Genetic Effects on Gene Expression across Human Tissues." *Nature* 550 (7675): 204–13. https://doi.org/10.1038/nature24277.
- Al-Yahya, Suhad, Linah Mahmoud, Fahad Al-Zoghaibi, Abdullah Al-Tuhami, Haithem Amer, Fahad N. Almajhdi, Stephen J. Polyak, and Khalid S. A. Khabar. 2015.
 "Human Cytokinome Analysis for Interferon Response." *Journal of Virology* 89 (14): 7108–19. https://doi.org/10.1128/jvi.03729-14.
- Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. "The Repertoire of Mutational Signatures in Human Cancer." *Nature* 578 (7793): 94–101. https://doi.org/10.1038/s41586-020-1943-3.
- Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A.J.R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. "Signatures of Mutational Processes in Human Cancer." *Nature* 500 (7463): 415–21. https://doi.org/10.1038/nature12477.
- Alioto, Tyler S., Ivo Buchhalter, Sophia Derdak, Barbara Hutter, Matthew D. Eldridge, Eivind Hovig, Lawrence E. Heisler, et al. 2015. "A Comprehensive Assessment of Somatic Mutation Detection in Cancer Using Whole-Genome Sequencing." *Nature Communications* 6 (December). https://doi.org/10.1038/ncomms10001.
- Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. "An Atlas of Active Enhancers across Human Cell Types and Tissues." *Nature* 507 (7493): 455–61. https://doi.org/10.1038/nature12787.
- Araya, Carlos L., Can Cenik, Jason A. Reuter, Gert Kiss, Vijay S. Pande, Michael P. Snyder, and William J. Greenleaf. 2016a. "Identification of Significantly Mutated Regions across Cancer Types Highlights a Rich Landscape of Functional Molecular Alterations." *Nature Genetics* 48 (2): 117–25. https://doi.org/10.1038/ng.3471.
- Araya, Carlos L, Can Cenik, Jason A Reuter, Gert Kiss, Vijay S Pande, Michael P Snyder, and William J Greenleaf. 2016b. "Identification of Significantly Mutated Regions across Cancer Types Highlights a Rich Landscape of Functional Molecular Alterations." *Nature Genetics* 48 (2): 117–25. https://doi.org/10.1038/ng.3471.

- Arend, William P., and Jean-Michel -M Dayer. 1995. "Inhibition of the Production and Effects of Interleukins-1 and Tumor Necrosis Factor α in Rheumatoid Arthritis." *Arthritis & Rheumatism* 38 (2): 151–60. https://doi.org/10.1002/art.1780380202.
- Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature*. Nature Publishing Group. https://doi.org/10.1038/nature15393.
- Baron, V., E. D. Adamson, A. Calogero, G. Ragona, and D. Mercola. 2006. "The Transcription Factor Egr1 Is a Direct Regulator of Multiple Tumor Suppressors Including TGFβ1, PTEN, P53, and Fibronectin." *Cancer Gene Therapy*. Cancer Gene Ther. https://doi.org/10.1038/sj.cgt.7700896.
- Barrera, David, Nancy Noyola-Martínez, Euclides Avila, Ali Halhali, Fernando Larrea, and Lorenza Díaz. 2012. "Calcitriol Inhibits Interleukin-10 Expression in Cultured Human Trophoblasts under Normal and Inflammatory Conditions." *Cytokine* 57 (3): 316–21. https://doi.org/10.1016/j.cyto.2011.11.020.
- Batmanov, Kirill, Wei Wang, Magnar Bjørås, Jan Delabie, and Junbai Wang. 2017. "Integrative Whole-Genome Sequence Analysis Reveals Roles of Regulatory Mutations in BCL6 and BCL2 in Follicular Lymphoma." *Scientific Reports* 7 (1): 7040. https://doi.org/10.1038/s41598-017-07226-4.
- Behan, Fiona M., Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M. Beaver, Giorgia Migliardi, Rita Santos, et al. 2019. "Prioritization of Cancer Therapeutic Targets Using CRISPR–Cas9 Screens." *Nature* 568 (7753): 511–16. https://doi.org/10.1038/s41586-019-1103-9.
- Bell, Robert J.A., H. Tomas Rube, Alex Kreig, Andrew Mancini, Shaun D. Fouse, Raman P. Nagarajan, Serah Choi, et al. 2015. "The Transcription Factor GABP Selectively Binds and Activates the Mutant TERT Promoter in Cancer." *Science* 348 (6238): 1036–39. https://doi.org/10.1126/science.aab0015.
- Bell, Robert J.A., H. Tomas Rube, Ana Xavier-Magalhães, Bruno M. Costa, Andrew Mancini, Jun S. Song, and Joseph F. Costello. 2016. "Understanding TERT Promoter Mutations: A Common Path to Immortality." *Molecular Cancer Research*. American Association for Cancer Research Inc. https://doi.org/10.1158/1541-7786.MCR-16-0003.
- Belperio, J A, M P Keane, D A Arenberg, C L Addison, J E Ehlert, M D Burdick, and R M Strieter. 2000. "CXC Chemokines in Angiogenesis." *Journal of Leukocyte Biology* 68 (1): 1–8.

Berger, Michael F., Anthony A. Philippakis, Aaron M. Qureshi, Fangxue S. He, Preston

W. Estep, and Martha L. Bulyk. 2006. "Compact, Universal DNA Microarrays to Comprehensively Determine Transcription-Factor Binding Site Specificities." *Nature Biotechnology* 24 (11): 1429–35. https://doi.org/10.1038/nbt1246.

- Berghe, Wim Vanden, Karolien De Bosscher, Elke Boone, Stéphane Plaisance, and Guy Haegeman. 1999. "The Nuclear Factor-KB Engages CBP/P300 and Histone Acetyltransferase Activity for Transcriptional Activation of the Interleukin-6 Gene Promoter." *Journal of Biological Chemistry* 274 (45): 32091–98. https://doi.org/10.1074/jbc.274.45.32091.
- Bernardini, G, G Spinetti, D Ribatti, G Camarda, L Morbidelli, M Ziche, A Santoni, M C Capogrossi, and M Napolitano. 2000. "I-309 Binds to and Activates Endothelial Cell Functions and Acts as an Angiogenic Molecule in Vivo." *Blood* 96 (13): 4039–45.
- Bessa Garcia, Simone Aparecida de, Mafalda Araújo, Tiago Pereira, João Mouta, and Renata Freitas. 2020. "HOX Genes Function in Breast Cancer Development." *Biochimica et Biophysica Acta - Reviews on Cancer*. Elsevier B.V. https://doi.org/10.1016/j.bbcan.2020.188358.
- Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. "Annotation of Functional Variation in Personal Genomes Using RegulomeDB." *Genome Research* 22 (9): 1790–97. https://doi.org/10.1101/gr.137323.112.
- Brattsand, R., and M. Linden. 1996. "Cytokine Modulation by Glucocorticoids: Mechanisms and Actions in Cellular Studies." In *Alimentary Pharmacology and Therapeutics, Supplement*, 10:81–90. Aliment Pharmacol Ther. https://doi.org/10.1046/j.1365-2036.1996.22164025.x.
- Bray, David, Heather Hook, Rose Zhao, Jessica L Keenan, Ashley Penvose, Yemi Osayame, Nima Mohaghegh, and Trevor Siggers. 2020. "Customizable High-Throughput Platform for Profiling Cofactor Recruitment to DNA to Characterize Cis-Regulatory Elements and Screen Non-Coding Single-Nucleotide Polymorphisms." *BioRxiv*, April, 2020.04.21.053710. https://doi.org/10.1101/2020.04.21.053710.
- Breuer, Karin, Amir K Foroushani, Matthew R Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L Winsor, Robert E W Hancock, Fiona S L Brinkman, and David J Lynn. 2013. "InnateDB: Systems Biology of Innate Immunity and beyond--Recent Updates and Continuing Curation." *Nucleic Acids Research* 41 (Database issue): D1228-33. https://doi.org/10.1093/nar/gks1147.
- Brewster, Brooke L., Francesca Rossiello, Juliet D. French, Stacey L. Edwards, Ming Wong, Ania Wronski, Phillip Whiley, et al. 2012. "Identification of Fifteen Novel Germline Variants in the BRCA1 3'UTR Reveals a Variant in a Breast Cancer Case

That Introduces a Functional MiR-103 Target Site." *Human Mutation* 33 (12): 1665–75. https://doi.org/10.1002/humu.22159.

- Brown, Andrew Anand, Ana Viñuela, Olivier Delaneau, Tim D. Spector, Kerrin S. Small, and Emmanouil T. Dermitzakis. 2017. "Predicting Causal Variants Affecting Expression by Using Whole-Genome Sequencing and RNA-Seq from Multiple Human Tissues." *Nature Genetics* 49 (12): 1747–51. https://doi.org/10.1038/ng.3979.
- Browne, Edward P., and Thomas Shenk. 2003. "Human Cytomegalovirus UL83-Coded Pp65 Virion Protein Inhibits Antiviral Gene Expression in Infected Cells." *Proceedings of the National Academy of Sciences of the United States of America* 100 (20): 11439–44. https://doi.org/10.1073/pnas.1534570100.
- Campbell, Peter J., Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, et al. 2020. "Pan-Cancer Analysis of Whole Genomes." *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.
- Carrasco Pro, Sebastian, Alvaro Dafonte Imedio, Clarissa Stephanie Santoso, Kok Ann Gan, Jared Allan Sewell, Melissa Martinez, Rebecca Sereda, Shivani Mehta, and Juan Ignacio Fuxman Bass. 2018. "Global Landscape of Mouse and Human Cytokine Transcriptional Regulation." *Nucleic Acids Research* 46 (18): 9321–37. https://doi.org/10.1093/nar/gky787.
- Carter, Scott L., Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W. Laird, et al. 2012. "Absolute Quantification of Somatic DNA Alterations in Human Cancer." *Nature Biotechnology* 30 (5): 413–21. https://doi.org/10.1038/nbt.2203.
- Cavelier, Lucia, Adam Ameur, Susana Häggqvist, Ida Höijer, Nicola Cahill, Ulla Olsson-Strömberg, and Monica Hermanson. 2015. "Clonal Distribution of BCR-ABL1 Mutations and Splice Isoforms by Single-Molecule Long-Read RNA Sequencing." *BMC Cancer* 15 (1). https://doi.org/10.1186/s12885-015-1046-y.
- Chadwick, Lisa Helbling. 2012. "The NIH Roadmap Epigenomics Program Data Resource." *Epigenomics*. Epigenomics. https://doi.org/10.2217/epi.12.18.
- Chan, Andrew C., and Paul J. Carter. 2010. "Therapeutic Antibodies for Autoimmunity and Inflammation." *Nature Reviews Immunology*. Nat Rev Immunol. https://doi.org/10.1038/nri2761.
- Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." *Nature Biotechnology* 31 (3): 213–19.

https://doi.org/10.1038/nbt.2514.

- Claeys, Marleen, Valerie Storms, Hong Sun, Tom Michoel, and Kathleen Marchal. 2012. "Motifsuite: Workflow for Probabilistic Motif Detection and Assessment." *Bioinformatics* 28 (14): 1931–32. https://doi.org/10.1093/bioinformatics/bts293.
- Claussnitzer, Melina, Simon N. Dankel, Bernward Klocke, Harald Grallert, Viktoria Glunk, Tea Berulava, Heekyoung Lee, et al. 2014. "Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms." *Cell* 156 (1–2): 343–58. https://doi.org/10.1016/j.cell.2013.10.058.
- Coetzee, Simon G., Gerhard A. Coetzee, and Dennis J. Hazelett. 2015. "MotifbreakR: An R/Bioconductor Package for Predicting Variant Effects at Transcription Factor Binding Sites." *Bioinformatics* 31 (23): 3847–49. https://doi.org/10.1093/bioinformatics/btv470.
- Davis, Carrie A, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, et al. 2018. "The Encyclopedia of DNA Elements (ENCODE): Data Portal Update." *Nucleic Acids Research* 46 (D1): D794–801. https://doi.org/10.1093/nar/gkx1081.
- Denisova, Evgeniya, Barbara Heidenreich, Eduardo Nagore, P. Sivaramakrishna Rachakonda, Ismail Hosen, Ivana Akrap, Víctor Traves, et al. 2015. "Frequent DPH3 Promoter Mutations in Skin Cancers." *Oncotarget* 6 (34): 35922–30. https://doi.org/10.18632/oncotarget.5771.
- Deplancke, Bart, Arnab Mukhopadhyay, Wanyuan Ao, Ahmed M. Elewa, Christian A. Grove, Natalia J. Martinez, Reynaldo Sequerra, et al. 2006. "A Gene-Centered C. Elegans Protein-DNA Interaction Network." *Cell* 125 (6): 1193–1205. https://doi.org/10.1016/j.cell.2006.04.038.
- Depristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–501. https://doi.org/10.1038/ng.806.
- Ding, Li, Matthew H. Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L. Gibbs, et al. 2018. "Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics." *Cell* 173 (2): 305-320.e10. https://doi.org/10.1016/j.cell.2018.03.033.
- Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, et al. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

- Eirew, Peter, Adi Steif, Jaswinder Khattra, Gavin Ha, Damian Yap, Hossein Farahani, Karen Gelmon, et al. 2015. "Dynamics of Genomic Clones in Breast Cancer Patient Xenografts at Single-Cell Resolution." *Nature* 518 (7539): 422–26. https://doi.org/10.1038/nature13952.
- Elkon, Ran, and Reuven Agami. 2017. "Characterization of Noncoding Regulatory DNA in the Human Genome." *Nature Biotechnology*. Nature Publishing Group. https://doi.org/10.1038/nbt.3863.
- Eppig, Janan T., Cynthia L. Smith, Judith A. Blake, Martin Ringwald, James A. Kadin, Joel E. Richardson, and Carol J. Bult. 2017. "Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research." In *Methods in Molecular Biology*, 1488:47– 73. Humana Press Inc. https://doi.org/10.1007/978-1-4939-6427-7 3.
- Ferraro, Bernadette, Gerald Bepler, Swati Sharma, Alan Cantor, and Eric B. Haura. 2005. "EGR1 Predicts PTEN and Survival in Patients with Non-Small-Cell Lung Cancer." *Journal of Clinical Oncology* 23 (9): 1921–26. https://doi.org/10.1200/JCO.2005.08.127.
- Fredriksson, Nils J, Lars Ny, Jonas A Nilsson, and Erik Larsson. 2014. "Systematic Analysis of Noncoding Somatic Mutations and Gene Expression Alterations across 14 Tumor Types." *Nature Genetics* 46 (12): 1258–63. https://doi.org/10.1038/ng.3141.
- Freedman, Steven J., Zhen Yu J. Sun, Florence Poy, Andrew L. Kung, David M. Livingston, Gerhard Wagner, and Michael J. Eck. 2002. "Structural Basis for Recruitment of CBP/P300 by Hypoxia-Inducible Factor-1a." *Proceedings of the National Academy of Sciences of the United States of America* 99 (8): 5367–72. https://doi.org/10.1073/pnas.082117899.
- Frith, Martin C, Yutao Fu, Liqun Yu, Jiang-Fan Chen, Ulla Hansen, and Zhiping Weng. 2004. "Detection of Functional DNA Motifs via Statistical Over-Representation." *Nucleic Acids Research* 32 (4): 1372–81. https://doi.org/10.1093/nar/gkh299.
- Fu, Yao, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng J.asmine Mu, Kevin Y. Yip, Ekta Khurana, and Mark Gerstein. 2014. "FunSeq2: A Framework for Prioritizing Noncoding Regulatory Variants in Cancer." *Genome Biology* 15 (10): 480. https://doi.org/10.1186/s13059-014-0480-5.
- Fujimoto, Akihiro, Mayuko Furuta, Yasushi Totoki, Tatsuhiko Tsunoda, Mamoru Kato, Yuichi Shiraishi, Hiroko Tanaka, et al. 2016. "Whole-Genome Mutational Landscape and Characterization of Noncoding and Structural Mutations in Liver Cancer." *Nature Genetics* 48 (5): 500–509. https://doi.org/10.1038/ng.3547.

- Futreal, P. Andrew, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. 2004. "A Census of Human Cancer Genes." *Nature Reviews Cancer*. Nature Publishing Group. https://doi.org/10.1038/nrc1299.
- Fuxman Bass, Juan I., John S. Reece-Hoyes, and Albertha J.M. Walhout. 2016a. "Gene-Centered Yeast One-Hybrid Assays." *Cold Spring Harbor Protocols* 2016 (12): 1039–43. https://doi.org/10.1101/pdb.top077669.

——. 2016b. "Generating Bait Strains for Yeast One-Hybrid Assays." Cold Spring Harbor Protocols 2016 (12): 1097–1103. https://doi.org/10.1101/pdb.prot088948.

- Fuxman Bass, Juan I., Nidhi Sahni, Shaleen Shrestha, Aurian Garcia-Gonzalez, Akihiro Mori, Numana Bhat, Song Yi, David E. Hill, Marc Vidal, and Albertha J.M. Walhout. 2015. "Human Gene-Centered Transcription Factor Networks for Enhancers and Disease Variants." *Cell* 161 (3): 661–73. https://doi.org/10.1016/j.cell.2015.03.003.
- Gaildrat, Pascaline, Audrey Killian, Alexandra Martins, Isabelle Tournier, Thierry Frébourg, and Mario Tosi. 2010. "Use of Splicing Reporter Minigene Assay to Evaluate the Effect on Splicing of Unclassified Genetic Variants." *Methods in Molecular Biology (Clifton, N.J.)*. Methods Mol Biol. https://doi.org/10.1007/978-1-60761-759-4_15.
- Gan, Kok A., Sebastian Carrasco Pro, Jared A. Sewell, and Juan I. Fuxman Bass. 2018. "Identification of Single Nucleotide Non-Coding Driver Mutations in Cancer." *Frontiers in Genetics* 9 (FEB): 16. https://doi.org/10.3389/fgene.2018.00016.
- Gao, Zhuanglei, Zhaoxia Li, Yuelin Liu, and Zhonghao Liu. 2019. "Forkhead Box O3 Promotes Colon Cancer Proliferation and Drug Resistance by Activating MDR1 Expression." *Molecular Genetics and Genomic Medicine* 7 (3). https://doi.org/10.1002/mgg3.554.
- Gerstein, Mark B., Anshul Kundaje, Manoj Hariharan, Stephen G. Landt, Koon Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, et al. 2012. "Architecture of the Human Regulatory Network Derived from ENCODE Data." *Nature* 489 (7414): 91–100. https://doi.org/10.1038/nature11245.
- Gilmore, T. D., and M. Herscovitch. 2006. "Inhibitors of NF-KB Signaling: 785 and Counting." *Oncogene*. Oncogene. https://doi.org/10.1038/sj.onc.1209982.
- Gilmore, Thomas D., and Steve Gerondakis. 2011. "The C-Rel Transcription Factor in Development and Disease." *Genes and Cancer*. Genes Cancer. https://doi.org/10.1177/1947601911421925.

- Goettel, Jeremy A., Roopali Gandhi, Jessica E. Kenison, Ada Yeste, Gopal Murugaiyan, Sharmila Sambanthamoorthy, Alexandra E. Griffith, et al. 2016. "AHR Activation Is Protective against Colitis Driven by T Cells in Humanized Mice." *Cell Reports* 17 (5): 1318–29. https://doi.org/10.1016/j.celrep.2016.09.082.
- Goh, Kwang II, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert László Barabási. 2007. "The Human Disease Network." *Proceedings of the National Academy of Sciences of the United States of America* 104 (21): 8685–90. https://doi.org/10.1073/pnas.0701361104.
- Goodman, W. A., S. Omenetti, D. Date, L. Di Martino, C. De Salvo, G. D. Kim, S. Chowdhry, et al. 2016. "KLF6 Contributes to Myeloid Cell Plasticity in the Pathogenesis of Intestinal Inflammation." *Mucosal Immunology* 9 (5): 1250–62. https://doi.org/10.1038/mi.2016.1.
- Goutagny, Stéphane, Jean C. Nault, Maxime Mallet, Dominique Henin, Jessica Z. Rossi, and Michel Kalamarides. 2014. "High Incidence of Activating TERT Promoter Mutations in Meningiomas Undergoing Malignant Progression." *Brain Pathology* 24 (2): 184–89. https://doi.org/10.1111/bpa.12110.
- Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. "FIMO: Scanning for Occurrences of a given Motif." *Bioinformatics* 27 (7): 1017–18. https://doi.org/10.1093/bioinformatics/btr064.
- Griffith, Jason W., Caroline L. Sokol, and Andrew D. Luster. 2014. "Chemokines and Chemokine Receptors: Positioning Cells for Host Defense and Immunity." *Annual Review of Immunology* 32 (1): 659–702. https://doi.org/10.1146/annurev-immunol-032713-120145.
- Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data." *Bioinformatics* (Oxford, England) 32 (18): 2847–49. https://doi.org/10.1093/bioinformatics/btw313.
- Guppy, Brent J., and Kirk J. McManus. 2017. "Synthetic Lethal Targeting of RNF20 through PARP1 Silencing and Inhibition." *Cellular Oncology* 40 (3): 281–92. https://doi.org/10.1007/s13402-017-0323-y.
- Haas, N. B., I. Quirt, S. Hotte, E. McWhirter, R. Polintan, S. Litwin, P. D. Adams, et al. 2014. "Phase II Trial of Vorinostat in Advanced Melanoma." *Investigational New Drugs* 32 (3): 526–34. https://doi.org/10.1007/s10637-014-0066-9.
- Han, Heonjong, Hongseok Shim, Donghyun Shin, Jung Eun Shim, Yunhee Ko, Junha Shin, Hanhae Kim, et al. 2015. "TRRUST: A Reference Database of Human Transcriptional Regulatory Interactions." *Scientific Reports* 5 (June). https://doi.org/10.1038/srep11432.

- Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, et al. 2012. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project." *Genome Research* 22 (9): 1760–74. https://doi.org/10.1101/gr.135350.111.
- Helleday, Thomas, Saeed Eshtad, and Serena Nik-Zainal. 2014a. "Mechanisms Underlying Mutational Signatures in Human Cancers." *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3729.

—. 2014b. "Mechanisms Underlying Mutational Signatures in Human Cancers." *Nature Reviews Genetics* 15 (9): 585–98. https://doi.org/10.1038/nrg3729.

- Hindorff, Lucia A., Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. 2009. "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits." *Proceedings of the National Academy of Sciences of the United States of America* 106 (23): 9362–67. https://doi.org/10.1073/pnas.0903103106.
- Hirota, Keiji, Hiroyuki Yoshitomi, Motomu Hashimoto, Shinji Maeda, Shin Teradaira, Naoshi Sugimoto, Tomoyuki Yamaguchi, et al. 2007. "Preferential Recruitment of CCR6-Expressing Th17 Cells to Inflamed Joints via CCL20 in Rheumatoid Arthritis and Its Animal Model." *Journal of Experimental Medicine* 204 (12): 2803–12. https://doi.org/10.1084/jem.20071397.
- Holloway, A. F., S. Rao, and M. F. Shannon. 2002. "Regulation of Cytokine Gene Transcription in the Immune System." *Molecular Immunology*. Mol Immunol. https://doi.org/10.1016/S0161-5890(01)00094-3.
- Holmes, C., C. Cunningham, E. Zotova, J. Woolford, C. Dean, S. Kerr, D. Culliford, and V. H. Perry. 2009. "Systemic Inflammation and Disease Progression in Alzheimer Disease." *Neurology* 73 (10): 768–74. https://doi.org/10.1212/WNL.0b013e3181b6bb95.
- Homey, Bernhard, Anja Müller, and Albert Zlotnik. 2002. "Chemokines: Agents for the Immunotherapy of Cancer?" *Nature Reviews Immunology*. European Association for Cardio-Thoracic Surgery. https://doi.org/10.1038/nri748.
- Horn, S., A. Figl, P. S. Rachakonda, C. Fischer, A. Sucker, A. Gast, S. Kadel, et al. 2013. "TERT Promoter Mutations in Familial and Sporadic Melanoma." *Science* 339 (6122): 959–61. https://doi.org/10.1126/science.1230062.
- Horn, Susanne, Adina Figl, P. Sivaramakrishna Rachakonda, Christine Fischer, Antje Sucker, Andreas Gast, Stephanie Kadel, et al. 2013. "TERT Promoter Mutations in Familial and Sporadic Melanoma." *Science* 339 (6122): 959–61. https://doi.org/10.1126/science.1230062.

- Hornshøj, Henrik, Morten Muhlig Nielsen, Nicholas A. Sinnott-Armstrong, Michał P. Świtnicki, Malene Juul, Tobias Madsen, Richard Sallari, et al. 2018. "Pan-Cancer Screen for Mutations in Non-Coding Elements with Conservation and Cancer Specificity Reveals Correlations with Expression and Survival /631/67/69 /631/114 Article." Npj Genomic Medicine 3 (1). https://doi.org/10.1038/s41525-017-0040-5.
- Hottiger, Michael O., and Gary J. Nabel. 2000. "Viral Replication and the Coactivators P300 and CBP." *Trends in Microbiology*. Trends Microbiol. https://doi.org/10.1016/S0966-842X(00)01874-6.
- Hsu, Yu Chin, Yu Ting Hsiao, Tzu Yuan Kao, Jan Gowth Chang, and Grace S. Shieh. 2017. "Detection of Somatic Mutations in Exome Sequencing of Tumor-Only Samples." *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-14896-7.
- Huang, F. W., E. Hodis, M. J. Xu, G. V. Kryukov, L. Chin, and L. A. Garraway. 2013. "Highly Recurrent TERT Promoter Mutations in Human Melanoma." *Science* 339 (6122): 957–59. https://doi.org/10.1126/science.1229259.
- Hudson, Thomas J., Warwick Anderson, Axel Aretz, Anna D. Barker, Cindy Bell, Rosa R. Bernabé, M. K. Bhan, et al. 2010. "International Network of Cancer Genome Projects." *Nature*. Nature. https://doi.org/10.1038/nature08987.
- Hume, Maxwell A., Luis A. Barrera, Stephen S. Gisselbrecht, and Martha L. Bulyk. 2015. "UniPROBE, Update 2015: New Tools and Content for the Online Database of Protein-Binding Microarray Data on Protein-DNA Interactions." *Nucleic Acids Research* 43 (D1): D117–22. https://doi.org/10.1093/nar/gku1045.
- Ivanov, Ivaylo I., Brent S. McKenzie, Liang Zhou, Carlos E. Tadokoro, Alice Lepelley, Juan J. Lafaille, Daniel J. Cua, and Dan R. Littman. 2006. "The Orphan Nuclear Receptor RORγt Directs the Differentiation Program of Proinflammatory IL-17+ T Helper Cells." *Cell* 126 (6): 1121–33. https://doi.org/10.1016/j.cell.2006.07.035.
- Jacob, Chaim O., Song Zang, Lily Li, Voicu Ciobanu, Frank Quismorio, Akiei Mizutani, Minoru Satoh, and Michael Koss. 2003. "Pivotal Role of Stat4 and Stat6 in the Pathogenesis of the Lupus-Like Disease in the New Zealand Mixed 2328 Mice." *The Journal of Immunology* 171 (3): 1564–71. https://doi.org/10.4049/jimmunol.171.3.1564.
- Johnson, P F, and S L McKnight. 1989. "Eukaryotic Transcriptional Regulatory Proteins." *Annual Review of Biochemistry* 58 (1): 799–839. https://doi.org/10.1146/annurev.bi.58.070189.004055.
- Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. "DNA-Binding Specificities of Human Transcription Factors." *Cell* 152 (1–2): 327–39.

https://doi.org/10.1016/j.cell.2012.12.009.

- Jung, Hyunchul, Donghoon Lee, Jongkeun Lee, Donghyun Park, Yeon Jeong Kim, Woong Yang Park, Dongwan Hong, Peter J. Park, and Eunjung Lee. 2015. "Intron Retention Is a Widespread Mechanism of Tumor-Suppressor Inactivation." *Nature Genetics* 47 (11): 1242–48. https://doi.org/10.1038/ng.3414.
- Juul, Malene, Johanna Bertl, Qianyun Guo, Morten Muhlig Nielsen, Michał Świtnicki, Henrik Hornshøj, Tobias Madsen, Asger Hobolth, and Jakob Skou Pedersen. 2017. "Non-Coding Cancer Driver Candidates Identified with a Sample- and Position-Specific Model of the Somatic Mutation Rate." *ELife* 6 (March). https://doi.org/10.7554/eLife.21778.
- Kalender Atak, Zeynep, Hana Imrichova, Dmitry Svetlichnyy, Gert Hulselmans, Valerie Christiaens, Joke Reumers, Hugo Ceulemans, and Stein Aerts. 2017. "Identification of Cis-Regulatory Mutations Generating de Novo Edges in Personalized Cancer Gene Regulatory Networks." *Genome Medicine* 9 (1). https://doi.org/10.1186/s13073-017-0464-7.
- Karin, Nathan, and Hila Razon. 2018. "Chemokines beyond Chemo-Attraction: CXCL10 and Its Significant Role in Cancer and Autoimmunity." *Cytokine* 109 (September): 24–28. https://doi.org/10.1016/j.cyto.2018.02.012.
- Kaser, Arthur, Ann Hwee Lee, Andre Franke, Jonathan N. Glickman, Sebastian Zeissig, Herbert Tilg, Edward E.S. Nieuwenhuis, et al. 2008. "XBP1 Links ER Stress to Intestinal Inflammation and Confers Genetic Risk for Human Inflammatory Bowel Disease." *Cell* 134 (5): 743–56. https://doi.org/10.1016/j.cell.2008.07.021.
- Katainen, Riku, Kashyap Dave, Esa Pitkänen, Kimmo Palin, Teemu Kivioja, Niko Välimäki, Alexandra E. Gylfe, et al. 2015. "CTCF/Cohesin-Binding Sites Are Frequently Mutated in Cancer." *Nature Genetics* 47 (7): 818–21. https://doi.org/10.1038/ng.3335.
- Kawaguchi, Yasushi, Masako Hara, and Timothy M. Wright. 1999. "Endogenous IL-1α from Systemic Sclerosis Fibroblasts Induces IL-6 and PDGF-A." *Journal of Clinical Investigation* 103 (9): 1253–60. https://doi.org/10.1172/JCI4304.
- Keilwagen, Jens, Stefan Posch, and Jan Grau. 2019. "Accurate Prediction of Cell Type-Specific Transcription Factor Binding." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-018-1614-y.
- Khan, Aziz, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, et al. 2018. "JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework." *Nucleic Acids Research* 46 (D1): D1284–D1284.

https://doi.org/10.1093/nar/gkx1188.

- Khurana, Ekta, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A. Rubin, and Mark Gerstein. 2016. "Role of Non-Coding Sequence Variants in Cancer." *Nature Reviews Genetics* 17 (2): 93–108. https://doi.org/10.1038/nrg.2015.17.
- Kim, Kyoung S., Vikram Rajagopal, Caryn Gonsalves, Cage Johnson, and Vijay K. Kalra. 2006. "A Novel Role of Hypoxia-Inducible Factor in Cobalt Chloride- and Hypoxia-Mediated Expression of IL-8 Chemokine in Human Endothelial Cells." *The Journal of Immunology* 177 (10): 7211–24. https://doi.org/10.4049/jimmunol.177.10.7211.
- Klein, Robyn S., Leonid Izikson, Terry Means, Hilary D. Gibson, Eugene Lin, Raymond A. Sobel, Howard L. Weiner, and Andrew D. Luster. 2004. "IFN-Inducible Protein 10/CXC Chemokine Ligand 10-Independent Induction of Experimental Autoimmune Encephalomyelitis." *The Journal of Immunology* 172 (1): 550–59. https://doi.org/10.4049/jimmunol.172.1.550.
- Knight, Julian C., Irina Udalova, Adrian V.S. Hill, Brian M. Greenwood, Norbert Peshu, Kevin Marsh, and Dominic Kwiatkowski. 1999. "A Polymorphism That Affects OCT-1 Binding to the TNF Promoter Region Is Associated with Severe Malaria." *Nature Genetics* 22 (2): 145–50. https://doi.org/10.1038/9649.
- Koboldt, Daniel C., Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. 2009.
 "VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples." *Bioinformatics* 25 (17): 2283–85. https://doi.org/10.1093/bioinformatics/btp373.
- Koboldt, Daniel C, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." *Genome Research* 22 (3): 568–76. https://doi.org/10.1101/gr.129684.111.
- Kucab, Jill E., Xueqing Zou, Sandro Morganella, Madeleine Joel, A. Scott Nanda, Eszter Nagy, Celine Gomez, et al. 2019. "A Compendium of Mutational Signatures of Environmental Agents." *Cell* 177 (4): 821-836.e16. https://doi.org/10.1016/j.cell.2019.03.001.
- Kumar, Sunil, Giovanna Ambrosini, and Philipp Bucher. 2017. "SNP2TFBS a Database of Regulatory SNPs Affecting Predicted Transcription Factor Binding Site Affinity." *Nucleic Acids Research* 45 (D1): D139–44. https://doi.org/10.1093/nar/gkw1064.

- Kutchko, Katrina M., Wes Sanders, Ben Ziehr, Gabriela Phillips, Amanda Solem, Matthew Halvorsen, Kevin M. Weeks, Nathaniel M, Nathaniel Moorman, and Alain Laederach. 2015. "Multiple Conformations Are a Conserved and Regulatory Feature of the RB1 5' UTR." RNA 21 (7): 1274–85. https://doi.org/10.1261/rna.049221.114.
- Kveler, Ksenya, Elina Starosvetsky, Amit Ziv-Kenet, Yuval Kalugny, Yuri Gorelik, Gali Shalev-Malul, Netta Aizenbud-Reshef, et al. 2018. "Immune-Centric Network of Cytokines and Cells in Disease Context Identified by Computational Mining of PubMed." *Nature Biotechnology* 36 (7): 651–59. https://doi.org/10.1038/nbt.4152.
- Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. "The Human Transcription Factors." *Cell*. Cell Press. https://doi.org/10.1016/j.cell.2018.01.029.
- Lanzós, Andrés, Joana Carlevaro-Fita, Loris Mularoni, Ferran Reverter, Emilio Palumbo, Roderic Guigó, and Rory Johnson. 2017. "Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features." Scientific Reports 7 (January). https://doi.org/10.1038/srep41544.
- Lasko, Loren M., Clarissa G. Jakob, Rohinton P. Edalji, Wei Qiu, Debra Montgomery, Enrico L. Digiammarino, T. Matt Hansen, et al. 2017. "Discovery of a Selective Catalytic P300/CBP Inhibitor That Targets Lineage-Specific Tumours." *Nature* 550 (7674): 128–32. https://doi.org/10.1038/nature24028.
- Law, Philip J., Maria Timofeeva, Ceres Fernandez-Rozadilla, Peter Broderick, James Studd, Juan Fernandez-Tajes, Susan Farrington, et al. 2019. "Association Analyses Identify 31 New Risk Loci for Colorectal Cancer Susceptibility." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-09775-w.
- Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. "Software for Computing and Annotating Genomic Ranges." *PLoS Computational Biology* 9 (8). https://doi.org/10.1371/journal.pcbi.1003118.
- Lawrence, Michael S., Petar Stojanov, Craig H. Mermel, James T. Robinson, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, and Gad Getz. 2014. "Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types." *Nature* 505 (7484): 495–501. https://doi.org/10.1038/nature12912.
- Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, et al. 2013. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." *Nature* 499 (7457): 214–18. https://doi.org/10.1038/nature12213.

- Le, Vu Thuy Khanh, Mirko Trilling, Albert Zimmermann, and Hartmut Hengel. 2008. "Mouse Cytomegalovirus Inhibits Beta Interferon(IFN-β) Gene Expression and Controls Activation Pathways of the IFN-β Enhanceosome." *Journal of General Virology* 89 (5): 1131–41. https://doi.org/10.1099/vir.0.83538-0.
- Li, Amy, Bjoern Chapuy, Xaralabos Varelas, Paola Sebastiani, and Stefano Monti. 2019. "Identification of Candidate Cancer Drivers by Integrative Epi-DNA and Gene Expression (IEDGE) Data Analysis." *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-52886-z.
- Li, Guoliang, Xiaoan Ruan, Raymond K. Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, et al. 2012. "Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation." *Cell* 148 (1–2): 84–98. https://doi.org/10.1016/j.cell.2011.12.014.
- Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics (Oxford, England)* 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.
- Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics (Oxford, England)* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.
- Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics (Oxford, England)* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.
- Li, Yinghui, Qi Ling Zhou, Wenjie Sun, Prashant Chandrasekharan, Hui Shan Cheng, Zhe Ying, Manikandan Lakshmanan, et al. 2015. "Non-Canonical NF-KB Signalling and ETS1/2 Cooperatively Drive C250T Mutant TERT Promoter Activation." *Nature Cell Biology* 17 (10): 1327–38. https://doi.org/10.1038/ncb3240.
- Li, Yongsheng, Nidhi Sahni, Rita Pancsa, Daniel J. McGrail, Juan Xu, Xu Hua, Jasmin Coulombe-Huntington, et al. 2017. "Revealing the Determinants of Widespread Alternative Splicing Perturbation in Cancer." *Cell Reports* 21 (3): 798–812. https://doi.org/10.1016/j.celrep.2017.09.071.
- Lieberman-Aiden, Erez, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

- Liu, Kuiliang, Jianghao Fan, and Jing Wu. 2017. "Forkhead Box Protein J1 (FOXJ1) Is Overexpressed in Colorectal Cancer and Promotes Nuclear Translocation of β-Catenin in SW620 Cells." *Medical Science Monitor* 23 (February): 856–66. https://doi.org/10.12659/MSM.902906.
- Liu, T., N. Wang, J. Cao, A. Sofiadis, A. Dinets, J. Zedenius, C. Larsson, and D. Xu. 2014. "The Age-and Shorter Telomere-Dependent Tert Promoter Mutation in Follicular Thyroid Cell-Derived Carcinomas." *Oncogene* 33 (42): 4978–84. https://doi.org/10.1038/onc.2013.446.
- Liu, Xue Song, Matthew D. Genet, Jenna E. Haines, Elie K. Mehanna, Shaowei Wu, Hung I.Harry Chen, Yidong Chen, et al. 2015. "Zbtb7a Suppresses Melanoma Metastasis by Transcriptionally Repressing Mcam." *Molecular Cancer Research* 13 (8): 1206–17. https://doi.org/10.1158/1541-7786.MCR-15-0169.
- Lochovsky, Lucas, Jing Zhang, Yao Fu, Ekta Khurana, and Mark Gerstein. 2015. "LARVA: An Integrative Framework for Large-Scale Analysis of Recurrent Variants in Noncoding Annotations." *Nucleic Acids Research* 43 (17): 8123–34. https://doi.org/10.1093/nar/gkv803.
- Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics*. Nat Genet. https://doi.org/10.1038/ng.2653.
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0550-8.
- Luscombe, Nicholas M., M. Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A. Teichmann, and Mark Gerstein. 2004. "Genomic Analysis of Regulatory Network Dynamics Reveals Large Topological Changes." *Nature* 431 (7006): 308–12. https://doi.org/10.1038/nature02782.
- MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. "The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog)." *Nucleic Acids Research* 45 (D1): D896–901. https://doi.org/10.1093/nar/gkw1133.
- MacEwan, David J. 2002. "TNF Receptor Subtype Signalling: Differences and Cellular Consequences." *Cellular Signalling*. Cell Signal. https://doi.org/10.1016/S0898-6568(01)00262-5.
- Manel, Nicolas, Derya Unutmaz, and Dan R. Littman. 2008. "The Differentiation of Human TH-17 Cells Requires Transforming Growth Factor-β and Induction of the Nuclear Receptor RORγt." *Nature Immunology* 9 (6): 641–49.

https://doi.org/10.1038/ni.1610.

- Marco, Antonio, Charlotte Konikoff, Timothy L Karr, and Sudhir Kumar. 2009.
 "Relationship between Gene Co-Expression and Sharing of Transcription Factor Binding Sites in Drosophila Melanogaster." *Bioinformatics (Oxford, England)* 25 (19): 2473–77. https://doi.org/10.1093/bioinformatics/btp462.
- Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. 2017. "Universal Patterns of Selection in Cancer and Somatic Tissues." *Cell* 171 (5): 1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.
- Martinez, Natalia J., Maria C. Ow, M. Inmaculada Barrasa, Molly Hammell, Reynaldo Sequerra, Lynn Doucette-Stamm, Frederick P. Roth, Victor R. Ambros, and Albertha J.M. Walhout. 2008. "A C. Elegans Genome-Scale MicroRNA Network Contains Composite Feedback Motifs with High Flux Capacity." *Genes and Development* 22 (18): 2535–49. https://doi.org/10.1101/gad.1678608.
- Matys, V., E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, et al. 2003. "TRANSFAC®: Transcriptional Regulation, from Patterns to Profiles." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkg108.
- Maurano, Matthew T., Eric Haugen, Richard Sandstrom, Jeff Vierstra, Anthony Shafer, Rajinder Kaul, and John A. Stamatoyannopoulos. 2015. "Large-Scale Identification of Sequence Variants Influencing Human Transcription Factor Occupancy in Vivo." *Nature Genetics* 47 (12): 1393–1401. https://doi.org/10.1038/ng.3432.
- Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA." *Science* 337 (6099): 1190–95. https://doi.org/10.1126/science.1222794.
- McCarthy, S A, D Chen, B S Yang, J J Garcia Ramirez, H Cherwinski, X R Chen, M Klagsbrun, C A Hauser, M C Ostrowski, and M McMahon. 1997. "Rapid Phosphorylation of Ets-2 Accompanies Mitogen-Activated Protein Kinase Activation and the Induction of Heparin-Binding Epidermal Growth Factor Gene Expression by Oncogenic Raf-1." *Molecular and Cellular Biology* 17 (5): 2401–12. https://doi.org/10.1128/mcb.17.5.2401.
- McLendon, Roger, Allan Friedman, Darrell Bigner, Erwin G. Van Meir, Daniel J. Brat, Gena M. Mastrogianakis, Jeffrey J. Olson, et al. 2008. "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways." *Nature* 455 (7216): 1061–68. https://doi.org/10.1038/nature07385.

Meagher, Craig, Guillermo Arreaza, Andrew Peters, Craig A. Strathdee, Philippe A.

Gilbert, Qing Sheng Mi, Pere Santamaria, Gregory A. Dekaban, and Terry L. Delovitch. 2007. "CCL4 Protects from Type 1 Diabetes by Altering Islet β -Cell-Targeted Inflammatory Responses." *Diabetes* 56 (3): 809–17. https://doi.org/10.2337/db06-0619.

- Medoff, Benjamin D., Alain Sauty, Andrew M. Tager, James A. Maclean, R. Neal Smith, Anuja Mathew, Jennifer H. Dufour, and Andrew D. Luster. 2002. "IFN-γ-Inducible Protein 10 (CXCL10) Contributes to Airway Hyperreactivity and Airway Inflammation in a Mouse Model of Asthma." *The Journal of Immunology* 168 (10): 5278–86. https://doi.org/10.4049/jimmunol.168.10.5278.
- Medzhitov, Ruslan, and Tiffany Horng. 2009. "Transcriptional Control of the Inflammatory Response." *Nature Reviews Immunology*. Nat Rev Immunol. https://doi.org/10.1038/nri2634.
- Melnikov, Alexandre, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, et al. 2012. "Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay." *Nature Biotechnology* 30 (3): 271–77. https://doi.org/10.1038/nbt.2137.
- Melton, Collin, Jason A. Reuter, Damek V. Spacek, and Michael Snyder. 2015. "Recurrent Somatic Mutations in Regulatory Regions of Human Cancer Genomes." *Nature Genetics* 47 (7): 710–16. https://doi.org/10.1038/ng.3332.
- Meng, Xiangdong, Michael H. Brodsky, and Scot A. Wolfe. 2005. "A Bacterial One-Hybrid System for Determining the DNA-Binding Specificity of Transcription Factors." *Nature Biotechnology* 23 (8): 988–94. https://doi.org/10.1038/nbt1120.
- Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. "Computational Correction of Copy Number Effect Improves Specificity of CRISPR-Cas9 Essentiality Screens in Cancer Cells." *Nature Genetics* 49 (12): 1779–84. https://doi.org/10.1038/ng.3984.
- Miotto, Deborah, Pota Christodoulopoulos, Ron Olivenstein, Rame Taha, Lisa Cameron, Anne Tsicopoulos, A. B. Tonnel, et al. 2001. "Expression of IFN-γ-Inducible Protein; Monocyte Chemotactic Proteins 1 3 and 4; and Eotaxin in TH1- and TH2-Mediated Lung Diseases." *Journal of Allergy and Clinical Immunology* 107 (4): 664–70. https://doi.org/10.1067/mai.2001.113524.
- Miyamoto, Megumi, Yukihiro Shimizu, Kazuhiko Okada, Yoshiro Kashii, Kiyohiro Higuchi, and Akiharu Watanabe. 1998. "Effect of Interleukin-8 on Production of Tumor-Associated Substances and Autocrine Growth of Human Liver and Pancreatic Cancer Cells." *Cancer Immunology Immunotherapy* 47 (1): 47–57. https://doi.org/10.1007/s002620050503.

- Mogi, Makio, Minoru Harada, Peter Riederer, Hirotaro Narabayashi, Keisuke Fujita, and Toshiharu Nagatsu. 1994. "Tumor Necrosis Factor-α (TNF-α) Increases Both in the Brain and in the Cerebrospinal Fluid from Parkinsonian Patients." *Neuroscience Letters* 165 (1–2): 208–10. https://doi.org/10.1016/0304-3940(94)90746-3.
- Mogno, Ilaria, Jamie C. Kwasnieski, and Barak A. Cohen. 2013. "Massively Parallel Synthetic Promoter Assays Reveal the in Vivo Effects of Binding Site Variants." *Genome Research* 23 (11): 1908–15. https://doi.org/10.1101/gr.157891.113.
- Movva, Rajiv, Peyton Greenside, Georgi K. Marinov, Surag Nair, Avanti Shrikumar, and Anshul Kundaje. 2019. "Deciphering Regulatory DNA Sequences and Noncoding Genetic Variants Using Neural Network Models of Massively Parallel Reporter Assays." PLoS ONE 14 (6). https://doi.org/10.1371/journal.pone.0218073.
- Müller, Anja, Bernhard Homey, Hortensia Soto, Nianfeng Ge, Daniel Catron, Matthew E. Buchanan, Terri McClanahan, et al. 2001. "Involvement of Chemokine Receptors in Breast Cancer Metastasis." *Nature* 410 (6824): 50–56. https://doi.org/10.1038/35065016.
- Munster, P. N., K. T. Thurn, S. Thomas, P. Raha, M. Lacevic, A. Miller, M. Melisko, et al. 2011. "A Phase II Study of the Histone Deacetylase Inhibitor Vorinostat Combined with Tamoxifen for the Treatment of Patients with Hormone Therapy-Resistant Breast Cancer." *British Journal of Cancer* 104 (12): 1828–35. https://doi.org/10.1038/bjc.2011.156.
- Murphy, Kenneth M., and Steven L. Reiner. 2002. "The Lineage Decisions of Helper T Cells." *Nature Reviews Immunology*. Nat Rev Immunol. https://doi.org/10.1038/nri954.
- Nakamura, Kyosuke, Akihiro Kato, Junya Kobayashi, Hiromi Yanagihara, Shuichi Sakamoto, Douglas V.N.P. Oliveira, Mikio Shimada, et al. 2011. "Regulation of Homologous Recombination by RNF20-Dependent H2B Ubiquitination." *Molecular Cell* 41 (5): 515–28. https://doi.org/10.1016/j.molcel.2011.02.002.
- Navarro, Lorena, Kerri Mowen, Steven Rodems, Brian Weaver, Nancy Reich, Deborah Spector, and Michael David. 1998. "Cytomegalovirus Activates Interferon Immediate-Early Response Gene Expression and an Interferon Regulatory Factor 3-Containing Interferon-Stimulated Response Element-Binding Complex." *Molecular and Cellular Biology* 18 (7): 3796–3802. https://doi.org/10.1128/mcb.18.7.3796.
- Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, et al. 2011. "Tumour Evolution Inferred by Single-Cell Sequencing." *Nature* 472 (7341): 90–95. https://doi.org/10.1038/nature09807.

Netea, Mihai G., Jos W.M. Van Der Meer, Marcel Van Deuren, and Bart Jan Kullberg.

2003. "Proinflammatory Cytokines and Sepsis Syndrome: Not Enough, or Too Much of a Good Thing?" *Trends in Immunology*. Elsevier Ltd. https://doi.org/10.1016/S1471-4906(03)00079-6.

- Neurath, Markus F. 2014. "Cytokines in Inflammatory Bowel Disease." *Nature Reviews Immunology*. Nature Publishing Group. https://doi.org/10.1038/nri3661.
- Nickel, Renate G., Vincenzo Casolaro, Ulrich Wahn, Kirsten Beyer, Kathleen C. Barnes, Beverly S. Plunkett, Linda R. Freidhoff, et al. 2000. "Atopic Dermatitis Is Associated with a Functional Mutation in the Promoter of the C-C Chemokine RANTES." *The Journal of Immunology* 164 (3): 1612–16. https://doi.org/10.4049/jimmunol.164.3.1612.
- Nik-Zainal, Serena, Ludmil B. Alexandrov, David C. Wedge, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, et al. 2012. "Mutational Processes Molding the Genomes of 21 Breast Cancers." *Cell* 149 (5): 979–93. https://doi.org/10.1016/j.cell.2012.04.024.
- Nik-Zainal, Serena, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, et al. 2016. "Landscape of Somatic Mutations in 560 Breast Cancer Whole-Genome Sequences." *Nature* 534 (7605): 47–54. https://doi.org/10.1038/nature17676.
- Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, et al. 2012. "The Life History of 21 Breast Cancers." *Cell* 149 (5): 994–1007. https://doi.org/10.1016/j.cell.2012.04.023.
- Noyes, Marcus B., Ryan G. Christensen, Atsuya Wakabayashi, Gary D. Stormo, Michael H. Brodsky, and Scot A. Wolfe. 2008. "Analysis of Homeodomain Specificities Allows the Family-Wide Prediction of Preferred Recognition Sites." *Cell* 133 (7): 1277–89. https://doi.org/10.1016/j.cell.2008.05.023.
- O'Keefe, Stephen J., Jun'Ichi Tamura, Randall L. Kincaid, Michael J. Tocci, and Edward A. O'Neill. 1992. "FK-506- and CsA-Sensitive Activation of the Interleukin-2 Promoter by Calcineurin." *Nature* 357 (6380): 692–94. https://doi.org/10.1038/357692a0.
- O'Shea, John J., Averil Ma, and Peter Lipsky. 2002. "Cytokines and Autoimmunity." *Nature Reviews Immunology*. European Association for Cardio-Thoracic Surgery. https://doi.org/10.1038/nri702.
- Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

- Peltz, Gary. 1997. "Transcription Factors in Immune-Mediated Disease." *Current Opinion in Biotechnology* 8 (4): 467–73. https://doi.org/10.1016/S0958-1669(97)80070-5.
- Piraino, Scott W., and Simon J. Furney. 2017. "Identification of Coding and Non-Coding Mutational Hotspots in Cancer Genomes." *BMC Genomics* 18 (1): 17. https://doi.org/10.1186/s12864-016-3420-9.
- Pon, Julia R., and Marco A. Marra. 2015. "Driver and Passenger Mutations in Cancer." *Annual Review of Pathology: Mechanisms of Disease* 10 (1): 25–50. https://doi.org/10.1146/annurev-pathol-012414-040312.
- Qin, Tao, Bai Li, Xiaoyue Feng, Shujun Fan, Lei Liu, Dandan Liu, Jun Mao, et al. 2018. "Abnormally Elevated USP37 Expression in Breast Cancer Stem Cells Regulates Stemness, Epithelial-Mesenchymal Transition and Cisplatin Sensitivity." *Journal of Experimental and Clinical Cancer Research* 37 (1). https://doi.org/10.1186/s13046-018-0934-9.
- Quang, Daniel, and Xiaohui Xie. 2019. "FactorNet: A Deep Learning Framework for Predicting Cell Type Specific Transcription Factor Binding from Nucleotide-Resolution Sequential Data." *Methods* 166 (August): 40–47. https://doi.org/10.1016/j.ymeth.2019.03.020.
- Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics (Oxford, England)* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.
- Rao, Anjana, Chun Luo, and Patrick G. Hogan. 1997. "TRANSCRIPTION FACTORS OF THE NFAT FAMILY:Regulation and Function." *Annual Review of Immunology* 15 (1): 707–47. https://doi.org/10.1146/annurev.immunol.15.1.707.
- Ravasi, T., H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, et al. 2010. "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man." *Cell* 140 (5): 744–52. https://doi.org/10.1016/j.cell.2010.01.044.
- Reece-Hoyes, John S., A. Rasim Barutcu, Rachel Patton McCord, Jun Seop Jeong, Lizhi Jiang, Andrew MacWilliams, Xinping Yang, et al. 2011. "Yeast One-Hybrid Assays for Gene-Centered Human Gene Regulatory Network Mapping." *Nature Methods* 8 (12): 1050–54. https://doi.org/10.1038/nmeth.1764.
- Reece-Hoyes, John S., Alos Diallo, Bryan Lajoie, Amanda Kent, Shaleen Shrestha, Sreenath Kadreppa, Colin Pesyna, Job Dekker, Chad L. Myers, and Albertha J.M. Walhout. 2011. "Enhanced Yeast One-Hybrid Assays for High-Throughput Gene-Centered Regulatory Network Mapping." *Nature Methods* 8 (12): 1059–68. https://doi.org/10.1038/nmeth.1748.

- Rentzsch, Philipp, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. 2019. "CADD: Predicting the Deleteriousness of Variants throughout the Human Genome." *Nucleic Acids Research* 47 (D1): D886–94. https://doi.org/10.1093/nar/gky1016.
- Rheinbay, Esther, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A. Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, et al. 2020. "Analyses of Non-Coding Somatic Drivers in 2,658 Cancer Whole Genomes." *Nature* 578 (7793): 102–11. https://doi.org/10.1038/s41586-020-1965-x.
- Rheinbay, Esther, Prasanna Parasuraman, Jonna Grimsby, Grace Tiao, Jesse M. Engreitz, Jaegil Kim, Michael S. Lawrence, et al. 2017a. "Recurrent and Functional Regulatory Mutations in Breast Cancer." *Nature* 547 (7661): 55–60. https://doi.org/10.1038/nature22992.

—. 2017b. "Recurrent and Functional Regulatory Mutations in Breast Cancer." Nature 547 (7661): 55–60. https://doi.org/10.1038/nature22992.

- Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. "Integrative Analysis of 111 Reference Human Epigenomes." *Nature* 518 (7539): 317–29. https://doi.org/10.1038/nature14248.
- Roberts, Steven A., Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargo, Petar Stojanov, Adam Kiezun, et al. 2013. "An APOBEC Cytidine Deaminase Mutagenesis Pattern Is Widespread in Human Cancers." *Nature Genetics* 45 (9): 970–76. https://doi.org/10.1038/ng.2702.
- Rolland, Thomas, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, et al. 2014. "A Proteome-Scale Map of the Human Interactome Network." *Cell* 159 (5): 1212–26. https://doi.org/10.1016/j.cell.2014.10.050.
- Rosa, Jaime S., Rebecca L. Flores, Stacy R. Oliver, Andria M. Pontello, Frank P. Zaldivar, and Pietro R. Galassetti. 2008. "Sustained IL-1α, IL-4, and IL-6 Elevations Following Correction of Hyperglycemia in Children with Type 1 Diabetes Mellitus." *Pediatric Diabetes* 9 (1): 9–16. https://doi.org/10.1111/j.1399-5448.2007.00243.x.
- Rosenberg, Alexander B., Rupali P. Patwardhan, Jay Shendure, and Georg Seelig. 2015. "Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences." *Cell* 163 (3): 698–711. https://doi.org/10.1016/j.cell.2015.09.054.

Salcedo, Rosalba, Howard A. Young, M. Lourdes Ponce, Jerrold M. Ward, Hynda K.

Kleinman, William J. Murphy, and Joost J. Oppenheim. 2001. "Eotaxin (CCL11) Induces In Vivo Angiogenic Responses by Human CCR3 + Endothelial Cells." *The Journal of Immunology* 166 (12): 7571–78. https://doi.org/10.4049/jimmunol.166.12.7571.

- Salton, Maayan, Wojciech K. Kasprzak, Ty Voss, Bruce A. Shapiro, Poulikos I. Poulikakos, and Tom Misteli. 2015. "Inhibition of Vemurafenib-Resistant Melanoma by Interference with Pre-MRNA Splicing." *Nature Communications* 6 (May). https://doi.org/10.1038/ncomms8103.
- Sánchez, Elena, Rogelio J. Palomino-Morales, Norberto Ortego-Centeno, Juan Jiménez-Alonso, Miguel A. González-Gay, Miguel A. López-Nevot, Julio Sánchez-Román, et al. 2009. "Identification of a New Putative Functional IL18 Gene Variant through an Association Study in Systemic Lupus Erythematosus." *Human Molecular Genetics* 18 (19): 3739–48. https://doi.org/10.1093/hmg/ddp301.
- Saunders, Christopher T., Wendy S.W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. 2012. "Strelka: Accurate Somatic Small-Variant Calling from Sequenced Tumor-Normal Sample Pairs." *Bioinformatics* 28 (14): 1811–17. https://doi.org/10.1093/bioinformatics/bts271.
- Schadendorf, Möller, Algermissen, Worm, Sticherling, and Czarnetzki. 1994. "IL-8 Produced by Human Malignant Melanoma Cells in Vitro Is an Essential Autocrine Growth Factor." *Journal of Immunology (Baltimore, Md. : 1950)*. United States.
- Schaub, Marc A., Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. 2012. "Linking Disease Associations with Regulatory Information in the Human Genome." *Genome Research* 22 (9): 1748–59. https://doi.org/10.1101/gr.136127.111.
- Schmidt, Dominic, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, et al. 2010. "Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding." *Science* 328 (5981): 1036–40. https://doi.org/10.1126/science.1186176.
- Sewell, Jared A., and Juan I. Fuxman Bass. 2017. "Cellular Network Perturbations by Disease-Associated Variants." *Current Opinion in Systems Biology*. Elsevier Ltd. https://doi.org/10.1016/j.coisb.2017.04.009.
- Shema, Efrat, Itay Tirosh, Yael Aylon, Jing Huang, Chaoyang Ye, Neta Moskovits, Nina Raver-Shapira, et al. 2008. "The Histone H2B-Specific Ubiquitin Ligase RNF20/HBRE1 Acts as a Putative Tumor Suppressor through Selective Regulation of Gene Expression." *Genes and Development* 22 (19): 2664–76. https://doi.org/10.1101/gad.1703008.

- Shepherd, Jonathan H., Ivan P. Uray, Abhijit Mazumdar, Anna Tsimelzon, Michelle Savage, Susan G. Hilsenbeck, and Powel H. Brown. 2016. "The SOX11 Transcription Factor Is a Critical Regulator of Basal-like Breast Cancer Growth, Invasion, and Basal-like Gene Expression." *Oncotarget* 7 (11): 13106–21. https://doi.org/10.18632/oncotarget.7437.
- Shi, Wenqiang, Oriol Fornes, Anthony Mathelier, and Wyeth W Wasserman. 2016. "Evaluating the Impact of Single Nucleotide Variants on Transcription Factor Binding." *Nucleic Acids Research* 44 (21): 10106–16. https://doi.org/10.1093/nar/gkw691.
- Shin, Sunyoung, Rebecca Hudson, Christopher Harrison, Mark Craven, and Sündüz Keleş. 2019. "AtSNP Search: A Web Resource for Statistically Evaluating Influence of Human Genetic Variation on Transcription Factor Binding." *Bioinformatics* (Oxford, England) 35 (15): 2657–59. https://doi.org/10.1093/bioinformatics/bty1010.
- Shiraishi, Yuichi, Yusuke Sato, Kenichi Chiba, Yusuke Okuno, Yasunobu Nagata, Kenichi Yoshida, Norio Shiba, et al. 2013. "An Empirical Bayesian Framework for Somatic Mutation Detection from Cancer Genome Sequencing Data." Nucleic Acids Research 41 (7). https://doi.org/10.1093/nar/gkt126.
- Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. 2014. "Transcriptional Enhancers: From Properties to Genome-Wide Predictions." *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3682.
- Shrestha, Shaleen, Jared Allan Sewell, Clarissa Stephanie Santoso, Elena Forchielli, Sebastian Carrasco Pro, Melissa Martinez, and Juan Ignacio Fuxman Bass. 2019.
 "Discovering Human Transcription Factor Physical Interactions with Genetic Variants, Novel DNA Motifs, and Repetitive Elements Using Enhanced Yeast One-Hybrid Assays." *Genome Research* 29 (9): 1533–44. https://doi.org/10.1101/gr.248823.119.
- Shuai, Shimin, Federico Abascal, Samirkumar B. Amin, Gary D. Bader, Pratiti Bandopadhayay, Jonathan Barenboim, Rameen Beroukhim, et al. 2020. "Combined Burden and Functional Impact Tests for Cancer Driver Discovery Using DriverPower." *Nature Communications* 11 (1). https://doi.org/10.1038/s41467-019-13929-1.
- Signori, Emanuela, Claudia Bagni, Sara Papa, Beatrice Primerano, Monica Rinaldi, Francesco Amaldi, and Vito Michele Fazio. 2001. "A Somatic Mutation in the 5'UTR of BRCA1 Gene in Sporadic Breast Cancer Causes Down-Modulation of Translation Efficiency." Oncogene 20 (33): 4596–4600. https://doi.org/10.1038/sj.onc.1204620.

- Singh, Ram Raj, Vijay Saxena, Song Zang, Lily Li, Fred D. Finkelman, David P. Witte, and Chaim O. Jacob. 2003. "Differential Contribution of IL-4 and STAT6 vs STAT4 to the Development of Lupus Nephritis." *The Journal of Immunology* 170 (9): 4818–25. https://doi.org/10.4049/jimmunol.170.9.4818.
- Smith, Robin P., Leila Taher, Rupali P. Patwardhan, Mee J. Kim, Fumitaka Inoue, Jay Shendure, Ivan Ovcharenko, and Nadav Ahituv. 2013. "Massively Parallel Decoding of Mammalian Regulatory Sequences Supports a Flexible Organizational Model." *Nature Genetics* 45 (9): 1021–28. https://doi.org/10.1038/ng.2713.
- Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. 2018. "The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers." *Nature Reviews Cancer*. Nature Publishing Group. https://doi.org/10.1038/s41568-018-0060-1.
- Spitz, François, and Eileen E.M. Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." *Nature Reviews Genetics*. Nat Rev Genet. https://doi.org/10.1038/nrg3207.
- Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Howells, Andrew D. Phillips, David N. Cooper, and Nick S.T. Thomas. 2009. "The Human Gene Mutation Database: 2008 Update." *Genome Medicine*. Genome Med. https://doi.org/10.1186/gm13.
- Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Shaw, Andrew D. Phillips, and David N. Cooper. 2014. "The Human Gene Mutation Database: Building a Comprehensive Mutation Repository for Clinical and Molecular Genetics, Diagnostic Testing and Personalized Genomic Medicine." *Human Genetics*. Hum Genet. https://doi.org/10.1007/s00439-013-1358-4.
- Stephen-Victor, Emmanuel, Helmut Fickenscher, and Jagadeesh Bayry. 2016. "IL-26: An Emerging Proinflammatory Member of the IL-10 Cytokine Family with Multifaceted Actions in Antiviral, Antimicrobial, and Autoimmune Responses." *PLoS Pathogens*. Public Library of Science. https://doi.org/10.1371/journal.ppat.1005624.
- Stunnenberg, Hendrik G., Sergio Abrignani, David Adams, Melanie de Almeida, Lucia Altucci, Viren Amin, Ido Amit, et al. 2016. "The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery." *Cell*. Cell Press. https://doi.org/10.1016/j.cell.2016.11.007.
- Supek, Fran, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner. 2014. "Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers." *Cell* 156 (6): 1324–35. https://doi.org/10.1016/j.cell.2014.01.051.
- Tak, Yu Gyoung, and Peggy J. Farnham. 2015. "Making Sense of GWAS: Using Epigenomics and Genome Engineering to Understand the Functional Relevance of SNPs in Non-Coding Regions of the Human Genome." *Epigenetics and Chromatin*. BioMed Central Ltd. https://doi.org/10.1186/s13072-015-0050-4.
- Tanaka, Ken Ichiro, Takushi Namba, Yasuhiro Arai, Mitsuaki Fujimoto, Hiroaki Adachi, Gen Sobue, Koji Takeuchi, Akira Nakai, and Tohru Mizushima. 2007. "Genetic Evidence for a Protective Role for Heat Shock Factor 1 and Heat Shock Protein 70 against Colitis." *Journal of Biological Chemistry* 282 (32): 23240–52. https://doi.org/10.1074/jbc.M704081200.
- Taniguchi, Tadatsugu, Kouetsu Ogasawara, Akinori Takaoka, and Nobuyuki Tanaka. 2001. "IRF F AMILY OF T RANSCRIPTION F ACTORS AS R EGULATORS OF H OST D EFENSE ." Annual Review of Immunology 19 (1): 623–55. https://doi.org/10.1146/annurev.immunol.19.1.623.
- Tewhey, Ryan, Dylan Kotliar, Daniel S Park, Brandon Liu, Sarah Winnicki, Steven K Reilly, Kristian G Andersen, et al. 2016. "Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay." *Cell* 165 (6): 1519–29. https://doi.org/10.1016/j.cell.2016.04.027.
- Thanos, Dimitris, and Tom Maniatis. 1995. "Virus Induction of Human IFNβ Gene Expression Requires the Assembly of an Enhanceosome." *Cell* 83 (7): 1091–1100. https://doi.org/10.1016/0092-8674(95)90136-1.
- Thomas-Chollier, Morgane, Matthieu Defrance, Alejandra Medina-Rivera, Olivier Sand, Carl Herrmann, Denis Thieffry, and Jacques van Helden. 2011. "RSAT 2011: Regulatory Sequence Analysis Tools." *Nucleic Acids Research* 39 (Web Server issue): W86-91. https://doi.org/10.1093/nar/gkr377.
- Thomas, Mary C., and Cheng Ming Chiang. 2006. "The General Transcription Machinery and General Cofactors." *Critical Reviews in Biochemistry and Molecular Biology*. Crit Rev Biochem Mol Biol. https://doi.org/10.1080/10409230600648736.
- Touzet, Hélène, and Jean-Stéphane Varré. 2007a. "Efficient and Accurate P-Value Computation for Position Weight Matrices." *Algorithms for Molecular Biology* 2 (1): 15. https://doi.org/10.1186/1748-7188-2-15.
- Touzet, Hélène, and Jean Stéphane Varré. 2007b. "Efficient and Accurate P-Value Computation for Position Weight Matrices." *Algorithms for Molecular Biology* 2 (1). https://doi.org/10.1186/1748-7188-2-15.
- Tsai, Eunice Y., James V. Falvo, Alla V. Tsytsykova, Amy K. Barczak, Andreas M. Reimold, Laurie H. Glimcher, Matthew J. Fenton, David C. Gordon, Ian F. Dunn, and Anne E. Goldfeld. 2000. "A Lipopolysaccharide-Specific Enhancer Complex

Involving Ets, Elk-1, Sp1, and CREB Binding Protein and P300 Is Recruited to the Tumor Necrosis Factor Alpha Promoter In Vivo." *Molecular and Cellular Biology* 20 (16): 6084–94. https://doi.org/10.1128/mcb.20.16.6084-6094.2000.

- Tsytsykova, Alla V., and Anne E. Goldfeld. 2002. "Inducer-Specific Enhanceosome Formation Controls Tumor Necrosis Factor Alpha Gene Expression in T Lymphocytes." *Molecular and Cellular Biology* 22 (8): 2620–31. https://doi.org/10.1128/mcb.22.8.2620-2631.2002.
- Tu, Xin, Shaofang Nie, Yuhua Liao, Hongsong Zhang, Qian Fan, Chengqi Xu, Ying Bai, et al. 2013. "The IL-33-ST2L Pathway Is Associated with Coronary Artery Disease in a Chinese Han Population." *American Journal of Human Genetics* 93 (4): 652– 60. https://doi.org/10.1016/j.ajhg.2013.08.009.
- Turner, Mark D., Belinda Nedjai, Tara Hurst, and Daniel J. Pennington. 2014. "Cytokines and Chemokines: At the Crossroads of Cell Signalling and Inflammatory Disease." *Biochimica et Biophysica Acta - Molecular Cell Research*. Elsevier. https://doi.org/10.1016/j.bbamcr.2014.05.014.
- Uhlen, Mathias, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, et al. 2017. "A Pathology Atlas of the Human Cancer Transcriptome." *Science* 357 (6352). https://doi.org/10.1126/science.aan2507.
- Ulirsch, Jacob C., Satish K. Nandakumar, Li Wang, Felix C. Giani, Xiaolan Zhang, Peter Rogov, Alexandre Melnikov, et al. 2016. "Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits." *Cell* 165 (6): 1530– 45. https://doi.org/10.1016/j.cell.2016.04.048.
- Valouev, Anton, David S. Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M. Myers, and Arend Sidow. 2008a. "Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-Seq Data." *Nature Methods* 5 (9): 829–34. https://doi.org/10.1038/nmeth.1246.
- Valouev, Anton, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. 2008b. "Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-Seq Data." *Nature Methods* 5 (9): 829–34. https://doi.org/10.1038/nmeth.1246.
- Vaquerizas, Juan M., Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. 2009. "A Census of Human Transcription Factors: Function, Expression and Evolution." *Nature Reviews Genetics*. Nat Rev Genet. https://doi.org/10.1038/nrg2538.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A.

Diaz, and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.1235122.

- Wang, J., C. Lu, D. Min, Z. Wang, and X. Ma. 2007. "A Mutation in the 5' Untranslated Region of the BRCA1 Gene in Sporadic Breast Cancer Causes Downregulation of Translation Efficiency." *Journal of International Medical Research* 35 (4): 564–73. https://doi.org/10.1177/147323000703500417.
- Wei, Sheng, Jiangong Niu, Hui Zhao, Zhensheng Liu, Li E. Wang, Younghun Han, Wei V. Chen, et al. 2011. "Association of a Novel Functional Promoter Variant (Rs2075533 C>T) in the Apoptosis Gene TNFSF8 with Risk of Lung Cancer-a Finding from Texas Lung Cancer Genome-Wide Association Study." *Carcinogenesis* 32 (4): 507–15. https://doi.org/10.1093/carcin/bgr014.
- Weinhold, Nils, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. 2014. "Genome-Wide Analysis of Noncoding Regulatory Mutations in Cancer." *Nature Genetics* 46 (11): 1160–65. https://doi.org/10.1038/ng.3101.
- Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R.Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Chris Sander, et al. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics*. Nature Publishing Group. https://doi.org/10.1038/ng.2764.
- Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity." *Cell* 158 (6): 1431–43. https://doi.org/10.1016/j.cell.2014.08.009.
- Wen, Xiaoquan, Roger Pique-Regi, and Francesca Luca. 2017. "Integrating Molecular QTL Data into Genome-Wide Genetic Association Analysis: Probabilistic Assessment of Enrichment and Colocalization." *PLoS Genetics* 13 (3). https://doi.org/10.1371/journal.pgen.1006646.
- Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, et al. 2018. "DrugBank 5.0: A Major Update to the DrugBank Database for 2018." *Nucleic Acids Research* 46 (D1): D1074–82. https://doi.org/10.1093/nar/gkx1037.
- Wong, Henry Sung Ching, Che Mai Chang, Xiao Liu, Wan Chen Huang, and Wei Chiao Chang. 2016. "Characterization of Cytokinome Landscape for Clinical Responses in Human Cancers." *OncoImmunology* 5 (11). https://doi.org/10.1080/2162402X.2016.1214789.

Wu, Chunlei, Xuefeng Jin, Ginger Tsueng, Cyrus Afrasiabi, and Andrew I. Su. 2016.

"BioGPS: Building Your Own Mash-up of Gene Annotations and Expression Profiles." *Nucleic Acids Research* 44 (D1): D313–16. https://doi.org/10.1093/nar/gkv1104.

- Wu, Ren Chin, Ayse Ayhan, Daichi Maeda, Kyu Rae Kim, Blaise A. Clarke, Patricia Shaw, Michael Herman Chui, Barry Rosen, Ie Ming Shih, and Tian Li Wang. 2014. "Frequent Somatic Mutations of the Telomerase Reverse Transcriptase Promoter in Ovarian Clear Cell Carcinoma but Not in Other Major Types of Gynaecological Malignancy." *Journal of Pathology* 232 (4): 473–81. https://doi.org/10.1002/path.4315.
- Wunderlich, Zeba, and Leonid A. Mirny. 2009. "Different Gene Regulation Strategies Revealed by Analysis of Binding Motifs." *Trends in Genetics*. Trends Genet. https://doi.org/10.1016/j.tig.2009.08.003.
- Xiao, T., J. J. Zhu, S. Huang, C. Peng, S. He, J. Du, R. Hong, et al. 2017. "Phosphorylation of NFAT3 by CDK3 Induces Cell Transformation and Promotes Tumor Growth in Skin Cancer." *Oncogene* 36 (20): 2835–45. https://doi.org/10.1038/onc.2016.434.
- Xie, Shanhai, Janet E. Price, Mario Luca, Didier Jean, Zeèv Ronai, and Menashe Bar-Eli. 1997. "Dominant-Negative CREB Inhibits Tumor Growth and Metastasis of Human Melanoma Cells." Oncogene 15 (17): 2069–75. https://doi.org/10.1038/sj.onc.1201358.
- Xu, Zongli, and Jack A Taylor. 2009. "SNPinfo: Integrating GWAS and Candidate Gene Information into Functional SNP Selection for Genetic Association Studies." *Nucleic Acids Research* 37 (Web Server issue): W600-5. https://doi.org/10.1093/nar/gkp290.
- Yang, Seoyeon, Ji Yeon Lee, Ho Hur, Ji Hoon Oh, and Myoung Hee Kim. 2018. "Up-Regulation of HOXB Cluster Genes Are Epigenetically Regulated in Tamoxifen-Resistant MCF7 Breast Cancer Cells." *BMB Reports* 51 (9): 450–55. https://doi.org/10.5483/BMBRep.2018.51.9.020.
- Yang, Xuexian O., Bhanu P. Pappu, Roza Nurieva, Askar Akimzhanov, Hong Soon Kang, Yeonseok Chung, Li Ma, et al. 2008. "T Helper 17 Lineage Differentiation Is Programmed by Orphan Nuclear Receptors RORα and RORγ." *Immunity* 28 (1): 29–39. https://doi.org/10.1016/j.immuni.2007.11.016.
- Yi, Woelsung, Sanjay Gupta, Edd Ricker, Michela Manni, Rolf Jessberger, Yurii Chinenov, Henrik Molina, and Alessandra B. Pernis. 2017. "The MTORC1-4E-BP-EIF4E Axis Controls de Novo Bcl6 Protein Synthesis in T Cells and Systemic Autoimmunity." *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-00348-3.

- Yiu Chan, Calvin Wing, Zuguang Gu, Matthias Bieg, Roland Eils, and Carl Herrmann. 2019. "Impact of Cancer Mutational Signatures on Transcription Factor Motifs in the Human Genome." *BMC Medical Genomics* 12 (1). https://doi.org/10.1186/s12920-019-0525-4.
- Yu, H. H., P. H. Liu, Y. C. Lin, W. J. Chen, J. H. Lee, L. C. Wang, Y. H. Yang, and B. L. Chiang. 2010. "Interleukin 4 and STAT6 Gene Polymorphisms Are Associated with Systemic Lupus Erythematosus in Chinese Patients." *Lupus* 19 (10): 1219–28. https://doi.org/10.1177/0961203310371152.
- Zajac-Kaye, Maria, Edward P. Gelmann, and David Levens. 1988. "A Point Mutation in the C-Myc Locus of a Burkitt Lymphoma Abolishes Binding of a Nuclear Protein." *Science* 240 (4860): 1776–80. https://doi.org/10.1126/science.2454510.
- Zhang, Baojie, Deng Chen, Bin Liu, Frank J. Dekker, and Wim J. Quax. 2020. "A Novel Histone Acetyltransferase Inhibitor A485 Improves Sensitivity of Non-Small-Cell Lung Carcinoma Cells to TRAIL." *Biochemical Pharmacology* 175 (May). https://doi.org/10.1016/j.bcp.2020.113914.
- Zhang, Jun Ming, and Jianxiong An. 2007. "Cytokines, Inflammation, and Pain." International Anesthesiology Clinics. NIH Public Access. https://doi.org/10.1097/AIA.0b013e318034194e.
- Zhang, Peng, Ji Han Xia, Jing Zhu, Ping Gao, Yi Jun Tian, Meijun Du, Yong Chen Guo, et al. 2018. "High-Throughput Screening of Prostate Cancer Risk Loci by Single Nucleotide Polymorphisms Sequencing." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-04451-x.
- Zhang, Xianxiang, Guangwei Liu, Lei Ding, Tao Jiang, Shihong Shao, Yuan Gao, and Yun Lu. 2018. "HOXA3 Promotes Tumor Growth of Human Colon Cancer through Activating EGFR/Ras/Raf/MEK/ERK Signaling Pathway." *Journal of Cellular Biochemistry* 119 (3): 2864–74. https://doi.org/10.1002/jcb.26461.
- Zhao, Wenxue, Joshua L. Pollack, Denitza P. Blagev, Noah Zaitlen, Michael T. McManus, and David J. Erle. 2014. "Massively Parallel Functional Annotation of 3" Untranslated Regions." *Nature Biotechnology* 32 (4): 387–91. https://doi.org/10.1038/nbt.2851.
- Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. "Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-09234-6.
- Zhu, Fangjie, Lucas Farnung, Eevi Kaasinen, Biswajyoti Sahu, Yimeng Yin, Bei Wei, Svetlana O. Dodonova, et al. 2018. "The Interaction Landscape between

Transcription Factors and the Nucleosome." *Nature* 562 (7725): 76–81. https://doi.org/10.1038/s41586-018-0549-5.

- Zhu, Min, Joanne S. Allard, Yongqing Zhang, Evelyn Perez, Edward L. Spangler, Kevin G. Becker, and Peter R. Rapp. 2014. "Age-Related Brain Expression and Regulation of the Chemokine CCL4/MIP-1β in APP/PS1 Double-Transgenic Mice." *Journal of Neuropathology and Experimental Neurology* 73 (4): 362–74. https://doi.org/10.1097/NEN.00000000000000060.
- Zhu, Qian, Aaron K. Wong, Arjun Krishnan, Miriam R. Aure, Alicja Tadych, Ran Zhang, David C. Corney, et al. 2015. "Targeted Exploration and Analysis of Large Cross-Platform Human Transcriptomic Compendia." *Nature Methods* 12 (3): 211–14. https://doi.org/10.1038/nmeth.3249.
- Zia, Amin, and Alan M. Moses. 2012. "Towards a Theoretical Understanding of False Positives in DNA Motif Finding." *BMC Bioinformatics* 13 (1): 151. https://doi.org/10.1186/1471-2105-13-151.
- Zong, Chenghang, Sijia Lu, Alec R. Chapman, and X. Sunney Xie. 2012. "Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell." *Science* 338 (6114): 1622–26. https://doi.org/10.1126/science.1229164.

CURRICULUM VITAE







