Computational challenges in genome wide association studies: data processing, variant annotation and epistasis

Pablo Cingolani

PhD.

School of Computer Science - Bioinformatics

McGill University

Montreal,Quebec

March 2015

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Pablo Cingolani 2015

CHAPTER 1
A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118} ; iso-2; iso-3

1.1 Preface

As this thesis is focused on extracting biological insight from sequencing data, in this chapter we examine algorithms we created for calculating "functional annotations" of genomic variants. In essence, functional variant annotations are bits of biological knowledge that allow us to make prioritize variants that are assumed to be more relevant to the phenotypic trait under study and to filter out variants assumed irrelevant. The spectrum of functional annotations for a genomic variant is wide and may involve information on which genes are affected by the variant, how the protein product is affected, how conserved is the genomic region the variant lies onto, and which clinically relevant information is associated with the loci; just to mention a few typical use cases.

When trying to find variants that affect risk of complex disease, statistical power is paramount. We need to be able to "separate wheat from chaff". In our context this means two different but closely related tasks: i) performing functional annotations, and ii) using that information for prioritizing variants (and filtering out the ones we suspect are not related to the particular trait under study). Failing to efficiently filter out irrelevant variants would reduce our statistical power as more statistical tests are calculated, thus would decrease our chances of finding the associations we are looking for. In order to efficiently annotate and filter variants, we created two software packages

called SnpEff and SnpSift that deal with the annotation and filtering aspects respectively.

The rest of the chapter is published in: Cingolani, Pablo, Rob Sladek, and Mathieu Blanchette. "BigDataScript: a scripting language for data pipelines." Bioinformatics 31.1 (2015): 10-16.

1.2 Abstract

We describe a new computer program, SnpEff, for rapidly categorizing the effects of single nucleotide polymorphisms (SNPs) and other variants such as multiple nucleotide polymorphism (MNPs) and insertion-deletions (InDels), in whole genome sequences. Once a genome is sequenced, the SnpEff program can be used to annotate and classify genetic polymorphisms based on their effects on annotated genes, such as synonymous or non-synonymous SNPs, start codon gains or losses, stop codon gains or losses; or based on their genomic locations, such as intronic, 5' untranslated region (5' UTR), 3' UTR, upstream, downstream or intergenic regions. Here the use of SnpEff is illustrated by annotating $\tilde{3}$56,660 candidate SNPs in $\tilde{1}$17 Mb unique sequences, representing a substitution rate of $\tilde{1}$/305 nucleotides, between the Drosophila melanogaster w^{1118} ; iso-2; iso-3 strain and the reference y 1 ; cn 1 bw 1 sp 1 strain. We show that $\tilde{1}$5,842 SNPs are synonymous and $\tilde{4}$,467 SNPs are non-synonymous (N/S $\tilde{0}$.28) and the remainder are in other categories, such as stop codon gains (38 SNPs), stop codon losses (8 SNPs) and start codon gains (297 SNPs) in the 5' UTR. We found, as expected, that the SNP frequency is proportional to the recombination frequency (i.e., highest in the middle of chromosome arms). We also found that start-gained and stop-lost SNPs in Drosophila melanogaster often encode N-terminal and C-terminal amino acids that are conserved in other Drosophila species. This suggests that the 5' and 3' UTRs are reservoirs of cryptic genetic variation that can be used multiple times during the

evolution of the Drosophila genus. At this time, SnpEff has been set up for annotating DNA polymorphisms of over 320 genome versions of multiple species including the human genome. It has already been used by over 50 institutions and universities in the bioinformatics community. Tools such as SnpEff are valuable because, as sequencing becomes cheaper and more available, whole genome sequencing is becoming more important in model organism genetics.

1.3 Introduction

When we re-sequenced the w^{1118} ; iso-2; iso-3 genome in 2009, 1 bioinformatics tools available then were unable to rapidly catego- rize the $\tilde{3}$56,660 SNPs as comparing to the y 1 ; cn 1 bw 1 sp 1 refer- ence strain. At the time, other available tools such as ENSEMBLs variant web application (http://ensembl.org) could only analyze a few hundred to a few thousand SNPs per batch. Therefore, over the past couple of years, we have been developing a new program, SnpEff, which is able to analyze and annotate thousands of variants per second. In addition to SnpEff, other programs to annotate genomic variants are currently now available, such as Annotate Variation (ANNOVAR) 2 and Variant Annotation, Analysis and Search Tool (VAAST). 3 However, SnpEff supports more genome versions, is open source for any user, supports variant call format (VCF) files and it is marginally faster (although the speeds of SnpEff, ANNOVAR and VAAST are comparable). Table S1 shows a feature comparison of some currently available software packages.

SnpEff, an abbreviation of "SNP effect," is a multi-platform open source variant effect predictor program. SnpEff annotates variants and predicts the coding effects of genetic variations, such as SNPs, insertions and deletions (INDELs) and multiple nucleotide polymorphisms (MNPs) (http://SnpEff.sourceforge.net/). The main features of SnpEff include: (1) speedthe ability to make thousands of predictions per second; (2) flexibilitythe ability to add custom genomes and

# SNP	Gene_name	Effect	Old_AA/new_AA	Old_codon/New_codon	Codon_Num (CDS)	CDS_size
chr2L:10006682_C/T	CG31755	UPSTREAM: 541 bases				
chr2L:10006758_G/A	CG31755	UPSTREAM: 465 bases				
chr2L:10007289_G/A	CG4747	SYNONYMOUS_CODING	L/L	TTG/TTA	489	1809
chr2L:10007319_G/C	CG4747	SYNONYMOUS_CODING	G/G	GGG/GGC	499	1809
chr2L:10007356_A/T	CG4747	INTRON				1809
chr2L:10007363_T/A	CG4747	INTRON				1809

Figure 1–1: # SNP, a description of the single nucleotide polymorphism (SNP) indicating chromosome arm (chr2L), coordinate in genome (10006682), and nucleotide change (e.g., C/T indicates that C is replaced by T in w^{1118} ; iso-2; iso-3 at this position). Gene_name, official gene symbol of gene. Effect, description of SNP (e.g., upstream of transcription start site at position 541). Old_AA/new_AA, amino acid change, if any, in one letter code. Old_codon/New_codon, if a codon contains a SNP, the old (reference) and new (w^{1118} ; iso-2; iso-3) codons are indicated. Codon_Num (CDS), the codon number of the coding sequence (CDS). CDS_size, the size of the protein in amino acids.

annotations; (3) the ability to integrate with Galaxy, an open access and web-based platform for computational bioinformatic research (http://gmod.org/wiki/Galaxy); (4) compatibility with multiple species and multiple codon usage tables (e.g., mitochondrial genomes); (5) integration with Genome Analysis Toolkit (GATK); 4 and (6) ability to perform non-coding annotations. When SnpEff was integrated into the GATK, it replaced the ANNOVAR program for variant analyses.

A simple walk-through example on how to analyze sequencing data to calculate variants and their effects is shown in Listing SL1. This example is intended for illustration purposes only since many additional steps are routinely used in re-sequencing data analysis pipelines, but design of a fully featured pipeline is beyond the scope of this paper.

Here, we report the results of SnpEff (version 1.9.6) analyses of the $\tilde{3}$56,660 candidate SNPs that we identified in w^{1118} ; iso-2; iso-3 with respect to the y 1 ; cn 1 bw 1 sp 1 reference strain as reported in our previous paper. 1 This is of great interest to the Drosophila community because thousands of transposon

insertion stocks 5 and hundreds of deficiency stocks 6,7 were generated in the w^{1118} ; iso-2; iso-3 genetic background. The large number and potential severity of many SNPs in the two laboratory strains was a surprising finding, and the possible evolutionary implications of this finding are discussed.

1.4 Results

Formats used in SnpEff. To understand the potential effects of large numbers of SNPs in genome sequence comparisons, we developed an open-source tool, SNPeff, to classify SNPs based on gene annotations. Table 1 shows the beginning portion of the output generated by SnpEff when the SNPs in w^{1118} ; iso-2; iso-3 were compared with the reference genome, y 1 ; cn 1 bw 1 sp 1 that is represented in Drosophila melanogaster release 5.3. A more complete SnpEff effect list is shown in Table 2. Before using SnpEff, an input file must be generated that lists all of the SNPs and INDELs in a genome. We published the input file for w^{1118} ; iso-2; iso-3 in our previous paper, 1 and it was derived by comparing hundreds of millions of short sequence reads ($\tilde{2}$0-fold genome coverage) and identifying SNPs based on a Sequence Alignment/Map tools (SAMtools) quality score for each nucleotide in the genome. 8

Input formats supported by SnpEff are variant call format (VCF), 9 tab separated TXT format; and and the SAMtools

Pileup format. 8 VCF was created by the 1,000 Genomes project and it is currently the de facto standard for variants in sequencing applications. The TXT and Pileup formats are currently deprecated and being phased out.

SnpEff also supports two output formats, TXT and VCF. The information provided in both of them includes four main groups: (i) variant information (genomic position, the reference and variant sequences, change type, heterozygosity, quality and coverage); (ii) genetic information (gene Id, gene

name, gene biotype, transcript ID, exon ID, exon rank); and (iii) effect information (effect type, amino acid changes, codon changes, codon number in CDS, codon degeneracy, etc.).

Whenever multiple transcripts for a gene exist, the effect and annotations on each transcript are reported, so one variant can have multiple output lines. Table 3 shows the information provided by each column in TXT format and Table 4 shows the information provided in VCF format. When using VCF format, the effect information is added to the information (INFO) fields using an effect (EFF) tag. As in the case of TXT output, if multiple alternative splicing products are annotated for a particular gene, SnpEff provides this information for each annotated version (see Sup. Data File 1 for the complete SnpEff output for w^{1118} ; iso-2; iso-3).

Predicted effects are with respect to protein coding genes. Variants affecting non-coding genes are annotated and the corresponding biotype is identified, whenever the information is available. A "biotype" is a group of organisms having the same specific genotype.

According to SnpEff (version 1.9.6), the largest number of SNPs in w^{1118} ; iso-2; iso-3 are in introns (130,126) followed by those in upstream (76,155), downstream (71,645) and intergenic (51,783) regions (Fig. 1). "Upstream" is defined as 5 kilobase (kb) upstream of the most distal transcription start site and "downstream" is defined as 5 kb downstream of the most distal polyA addition site, but these default variables can be easily adjusted. SnpEff also found thousands of SNPs within the exons. For example, there are 3,718 SNPs in the 3' untranslated regions (3' UTR) and 2,508 SNPs in the 5' untranslated regions (5' UTR). The SNPs in the upstream, downstream, 5' and 3' UTR regions might affect transcription or translation, but the actual effects have to

Effect	Note
INTERGENIC	The variant is in an intergenic region
UPSTREAM	Upstream of a gene (default length: 5K bases)
UTR_5_PRIME	Variant hits 5'UTR region
UTR_5_DELETED	The variant deletes and exon which is in the 5'UTR of the transcript
START_GAINED	A variant in 5'UTR region produces a three base sequence that can be a START codon
SPLICE_SITE_ACCEPTOR	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon)
SPLICE_SITE_DONOR	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon)
START_LOST	Variant causes start codon to be mutated into a non-start codon
SYNONYMOUS_START	Variant causes start codon to be mutated into another start codon
CDS	The variant hits a CDS
GENE	The variant hits a gene
TRANSCRIPT	The variant hits a transcript
EXON	The vairant hist an exon
EXON_DELETED	A deletion removes the whole exon
NON_SYNONYMOUS_CODING	Variant causes a codon that produces a different amino acid
SYNONYMOUS_CODING	Variant causes a codon that produces the same amino acid
FRAME_SHIFT	Insertion or deletion causes a frame shift
CODON_CHANGE	One or many codons are changed
CODON_INSERTION	One or many codons are inserted
CODON_CHANGE_PLUS_CODON_INSERTION	One codon is changed and one or many codons are inserted
CODON_DELETION	One or many codons are deleted
CODON_CHANGE_PLUS_CODON_DELETION	One codon is changed and one or more codons are deleted
STOP_GAINED	Variant causes a STOP codon
SYNONYMOUS_STOP	Variant causes stop codon to be mutated into another stop codon
STOP_LOST	Variant causes stop codon to be mutated into a non-stop codon
INTRON	Variant hist and intron. Technically, hits no exon in the transcript
UTR_3_PRIME	Variant hits 3'UTR region
UTR_3_DELETED	The variant deletes and exon which is in the 3'UTR of the transcript
DOWNSTREAM	Downstream of a gene (default length: 5K bases)
INTRON_CONSERVED	The variant is in a highly conserved intronic region
INTERGENIC_CONSERVED	The variant is in a highly conserved intergenic region

Figure 1–2: Detailed effect list from SnpEff

be confirmed case-by-case. In the next few sections, we present examples of several types of SNPs that might affect the protein function.

Heterozygosity is not considered in the w^{1118} ; iso-2; iso-3 sequence because the stock was isogenized and only high quality (i.e., homozygous SNPs) were used for this analysis. 1

The SnpEff website (http://snpeff.sourceforge.net/SnpSift. html) has a frequently asked questions (FAQ) section that addresses most issues that a user might have in operating this program.

SNPs that generate new start codons. There are 297 SNPs that potentially generate a new translation initiation codon in the 5' UTR (start-gained SNPs). The most common translation initiation codon is AUG, which is coded by ATG in the genome. To be thorough, we also included CUG and UUG codons, which code for leucine, as these codons can also be used to initiate translation in rare genes in Drosophila and mammals. 10,11 There are 60 genes with ATG start-gained SNPs (Table 5), 99 genes with CTG start-gained SNPs and 120 genes with TTG startgained SNPs in w^{1118} ; iso-2; iso-3, all by definition in 5' UTR regions, compared with the reference genome (the reading frame is indicated on the SnpEff table). Most of the ATG start-gained SNPs are within 1 kb of the annotated translation start (Table 5), but this probably reflects the fact that most 5' UTR sequences are less than 1 kb long. Less than expected by chance, only $\tilde{2}5\%$ of the ATG start-gain SNPs are in the same reading frame as the annotated translation start point (Table 5). Since 33% of in frame ATG start-gained SNPs are expected by chance, this suggests that there might be weak selection against this class of SNPs. Of the 60 genes with ATG start-gained SNPs, five genes have two ATG start-gained SNPs and one gene has three startgained SNPs; the remaining 54 genes have a single start-gained SNP. Since SnpEff does not take into account the Kozak consensus sequence flanking the AUG site, 5'-ACC AUG G-3', that is generally required for efficient translation, 12 and thus further assessment is required to determine whether a start-gained SNP is actually used.

Gene ontology (GO) pathway analysis of the genes affected by the 297 start-gain SNPs in w^{1118} ; iso-2; iso-3 was done using DAVID (Database for Annotation, Visualization and Integrated Discovery). 13,14 We found that the

Column	Notes
Chromosome	Chromosome name (usually without any leading 'chr' string)
Position	One based position
Reference	Reference
Change	Sequence change
Change type	Type of change (SNP, MNP, INS, DEL)
Homozygous	Is this homozygous or heterozygous (Hom, Het)
Quality	Quality score (from input file)
Coverage	Coverage (from input file)
Warnings	Any warnings or errors.
Gene_ID	Gene ID (usually ENSEMBL)
Gene_name	Gene name
Bio_type	BioType, as reported by ENSEMBL
Trancript_ID	Transcript ID (usually ENSEMBL)
Exon_ID	Exon ID (usually ENSEMBL)
Exon_Rank	Exon number on a transcript
Effect	Effect of this variant. See details below
old_AA/new_AA	Amino acid change
old_codon/new_codon	Codon change
Codon_Num(CDS)	Codon number in CDS
Codon_degenaracy	Codon degenaracy
CDS_size	CDS size in bases
Custom_interval_ID	If any custom interval was used, add the IDs here (may be more than one)

Figure 1–3: Information provided by SnpEff in tab separaOutput format (TXT)

GO categories "tissue morphogenesis," "immunoglobulin like," "developmental protein," and "alternative splicing" are significantly enriched after multiplecomparisons correction by false-discovery rate (FDR ¡ 0.001; Table 6). These categories are interesting because they predominantly contain proteins that show a wide degree of intra- and interspecies variability. For example, the immunoglobulin loci, which are highly divergent among humans and in other vertebrates, are used for antigen recognition. 15 Also, developmental proteins and proteins involved in tissue morphogenesis often have both conserved domains, such as the Hox domain, and highly divergent domains that maintain morphological diversity within a species, such as the trans-activation domains. 16,17

Our previous analyses suggest that most of the SNPs that we identified in w^{1118} ; iso-2; iso-3 are probably genuine and can be validated by capillary sequencing. 1 A common worry about nextgeneration sequencing data in general is that SNPs are vastly over estimated. One might thing that if a large fraction of the identified SNPs had the predicted "effects", the organism would not be viable. However, since short-read next-generation sequencing has a high error rate, such as the short-read sequences we obtained with the Illumina platform, further validation of specific SNPs is needed to be absolutely certain. Further validation of SNPs is best done with long-range DNA sequencing, such as with traditional capillary sequencing, or sequencing with the Roche, 18 and many other DNA sequencing instruments that are now available 20 (see ref.1 for validation examples with capillary sequencing).

An example of a start-gained SNP is found in the 5' UTR of Ecdysone inducible protein 63E (Eip63E) gene, which is predicted to be a cyclin J dependent kinase required for oogenesis and embryonic development (Fig. 2).

21 The potential start-gain SNP (A ¿ G) in Eip63E changes 5'-ATA-3' to 5'-ATG-3' in the same reading frame with no in-frame intervening stop codons (Fig. 2A). If translation occurs at the new start-gained SNP, it would produce a protein with 57 additional N-terminal amino acids compared with the reference gene (Fig. 2B). However, the three bases prior to the new 5'-ATG-3' sequence, 5'-AAT-3', is a poor match to the Kozak consensus sequence, 5'-ACC-3', discussed above in reference 12. Therefore, it is unclear whether the startgain SNP in Eip63E is recognized by the ribosomal machinery.

It is interesting that a BLASTp search of the protein database reveals that the N-terminal 57 amino acids in Eip63E are 63% identical (36/57) to the 58 N-terminal amino acids of the orthologous gene in Drosophila yakuba, but not to any other Drosophila species. D. yakuba is very close to D. melanogaster in the phylogeny. This suggests that the 5' UTR of Eip63E might be a source for cryptic genetic variation encoding novel N-terminal protein sequences that potentially modulates protein function (see Discussion).

SNPs that generate new stop codons. Another surprise in our SnpEff analysis was the identification of 28 stop-gained SNPs and 5 stop-lost SNPs in w^{1118} ; iso-2; iso-3 (Table 7). A stop-gained SNP, classically called a nonsense SNP, has a coding codon changed to a stop codon, UAA, UAG, UGA. 22 Three genes, oc/ otd, LRP1 and trol9, have two stop-gained SNPs. Surprisingly at least 8 of the stop-gained SNPs are in genes that encode essential proteins, and these are Dif, dp, ex, MESR4, mew, oc/otd, tai and trol. It is not known whether the other stop-gained SNPs also affect essential protein-coding genes because their functions have not yet been characterized (according to www.flybase.org). We note that what would be a stop-gained SNP in w^{1118} ; iso-2; iso-3 would be a stop-lost SNP in the reference strain, and vice versa,

because the sequence of the ancestral Drosophila melanogaster strain that gave rise to both of these strains is not known.

An important consideration with stop-gained and stop-lost SNPs is whether the C-terminal amino acids in the longest version of the protein that not present in the shortest version of the protein are conserved in other Drosophila species. If the additional C-terminal amino acids are not conserved, then these amino acids might not affect the essential function of the protein but they might exert modulatory effects. If the additional C-terminal amino acids are conserved in multiple Drosophila species, then their loss might adversely affect the function of the protein. Therefore, in Table 7, we further classify the stop-gained and stop-lost SNPs into four categories: Category 1, including 23 genes, with both the N-terminal and novel C-terminal regions conserved among Drosophila species and other organisms; Category 2, including only one gene, with the entire gene sequence not conserved even among other Drosophila species; Category 3, with two genes, with the novel C-termini not conserved among other Drosophila species. In this category, the N-termini are conserved among Drosophila species, but this conservation is not maintained beyond the Drosophila genus (this class is likely a novel gene that arose in the Drosophila genus); and Category 4, including seven genes, with the novel C-terminal regions conserved among other Drosophila species but not beyond the Drosophila genus. In this category, the N-terminus is conserved beyond the Drosophila genus (this class probably has a C-terminal domain with a modulatory role in the Drosophila genus but not beyond the genus).

An example of an essential protein-coding gene in Category 4, where the novel C-terminus is not conserved outside the Drosophila genus, is oceliless (oc), also known as orthodenticle (otd) (Fig. 3). The oc/otd gene has two in-frame stop-gained SNPs in w^{1118} ; iso-2; iso-3. The oc/otd gene is a Hox-family

Sub-field	Notes
Effect	Effect of this variant. See details below
Codon_Change	Codon change: old_codon/new_codon
Amino_Acid_change	Amino acid change: old_AA/new_AA
Warnings	Any warnings or errors
Gene_name	Gene name
Gene_BioType	BioType, as reported by ENSEMBL
Coding	[CODING | NON_CODING]. If information reported by ENSEMBL (e.g., has 'protein_id' information in GTF file)
Trancript	Transcript ID (usually ENSEMBL)
Exon	Exon ID (usually ENSEMBL)
Warnings	Any warnings or errors (not shown if empty)

Figure 1–4: Information provided by SnpEff in variant call format (VCF)

transcription factor required for photoreceptor development in the compound eye and the light-sensing ocellus, embryonic development and brain segmentation. 23,24 The Hox domain is 60 amino acids, 59 of which are identical with the human Otd protein. The Hox domains, which arose before invertebrates and vertebrates split several hundred million years ago, are among the most conserved protein domains in bilaterally-symmetric organisms in evolution. 25 The two stop-gained SNPs are in the non-conserved C-terminal region of Oc/Otd, which is thought to have a transcriptional-regulatory function. Since both strains are viable, both oc/otd genes are apparently functional although they encode a protein with 489 amino acids in w^{1118} ; iso-2; iso-3, and a protein with 543 amino acids in the reference genome (Table 6).

An example of a stop-lost gene in class c, where the C-terminus is not conserved even among the Drosophila genera, is CG13958 that encodes a protein of unknown function (Fig. 4). In w^{1118} ; iso-2; iso-3, CG13958 encodes a protein of 48 amino acids but in the reference genome it encodes a protein with 84 amino acids. When BLASTp was done with the non-redundant (nr) data set, there was not much homology beyond the 38 th amino acid within the Drosophila genus. However, there was a near perfect (37/38) identity of the first 38 amino acids in four other Drosophila species: Drosophila grimshawi, Drosophila yakuba, Drosophila erecta and Drosophila virilus (Fig. 4). This

protein likely arose in the Drosophila genus since it has no known homologs outside of this genus.

There are also five stop-lost SNPs in w^{1118} ; iso-2; iso-3 (Table 6). All of these SNPs are in predicted protein-coding genes, metabotropic GABA-B receptor subtype 1 (GABA-B-R1), CG13958, CG4975, brown (bw), and POU domain motif 3 (pdm3). It is not known whether any of the these genes are essential in Drosophila besides bw, which is not required for viability. However, the metabotropic GABA-B receptor subtype 1 (GABAB-R1) gene is required for normal behavior in mice 26 and the ortholog is therefore likely also essential in Drosophila, although no phenotypic data are available (www.flybase.org). The bw gene is classic gene first described in 1921 by Waaler, 27 which causes the eyes to be brown rather than red and encodes an ATPase binding cassette (ABC) transporter. 28 The bw 1 mutation in the reference strain is a spontaneous allele with a 412-transposon repeat insertion, 29 which would have been missed in our nextgeneration sequencing data because the input sequence we analyzed contained only short-read sequences that mapped uniquely to the reference genome.

Not much is known about the functions of several genes with in-frame stop-gained SNPs. The pdm3 gene is expressed in the larval and adult nervous system, and it encodes a highlyconserved Hox domain, but no phenotypic data are available (www.flybase.org). No phenotypic data are available for either CG13958 or CG4975. The protein encoding CG13958 has no known conserved domain, and its peak expression is observed within 0624 h of embryogenes, during early larval stages, at stages throughout the pupal period, and in the adult male (www. flybase.org). The protein encoded by CG4975 has an Armadillolike helical domain and an Ataxin-10 domain and has expression in the hind gut during the late larval and periods (www.flybase. org). 30

Some of the stop-lost SNPs have interesting consequences. For example, a stop-lost SNP in w^{1118} ; iso-2; iso-3 is in the CG13958 gene and causes an extension of eight amino acids before the next stop codon in 3' UTR sequence is reached (Fig. 5). Since the C-termini of CG13958 vary in w^{1118} ; iso-2; iso-3 and the reference strains of Drosophila melanogaster, it is conceivable that the C-terminus might also fluctuate in other Drosophila species. To test this idea, we investigated the C-terminal regions of CG13958 homologs in other Drosophila species.

We found that CG13958 homologs have variable C-terminal amino acids in different species of Drosophila. When the CG13958 protein is analyzed by protein Basic Local Alignment Search Tool (BLASTp) with the non-redundant (nr) protein database (http://www.ncbi.nlm.nih.gov/), at least two Drosophila species have extended C-terminal amino acids and at least three Drosophila species have missing amino acids at the C-termini (Fig. 5). For example, Drosophila pseudoobscura has three of the extended amino acids found in w^{1118} ; iso-2; iso-3 and Drosophila mojavenais has four of them. In contrast, Drosophila simulans is missing the last terminal amino acid, Drosophila erecta is missing the last two terminal amino acids, and Drosophila yakuba is missing the last three amino acids found in the reference strain (Fig. 5). The large number of stop-gain and stop-lost SNPs in Drosophila likely has important implications on the evolution of protein function (see Discussion).

Synonymous and non-synonymous SNPs in w^{1118} ; iso-2; iso-3. There are 15,842 synonymous SNPs and 4,467 nonsynonymous SNPs in annotated coding regions in w^{1118} ; iso-2; iso-3 (Fig. 1). A synonymous SNP (silent SNP) is defined as a SNP that does not change the amino acid in the protein, whereas a nonsynonymous SNP does. The genome-wide normalized N/S ratio (dN/dS), also called (i.e., = dN/dS), is by definition normalized to 1 in most

evolutionary studies. 31 The non-normalized N/S ratio is $\tilde{0}.28$ in w^{1118} ; iso-2; iso-3 compared with the reference genome, y 1 ; cn 1 bw 1 sp 1 (i.e., N/S = 4,467/15,842; Table 1).

We examined the distribution of synonymous and nonsynonymous SNPs genome-wide for w^{1118} ; iso-2; iso-3 and saw higher levels of both classes of SNPs in the middle of the chromosome arms and lower levels near the centromeres and telomeres (Fig. 6 and left). This was expected because the number of SNPs is proportional to the recombination frequencies in the different regions of the chromosomes. 32,33 Also, our previous analyses of the distribution of total SNPs revealed a similar pattern. 1 We observed higher N/S ratios near the telomeres and centromeres and lower N/S ratios in the middle of the chromosome arms (Fig. 6 and right).

1.5 Discussion

In this paper, we used SnpEff to categorize the $\tilde{3}$56,660 SNPs in w^{1118} ; iso-2; iso-3 and place them into 14 different classes based on their predicted effects on protein function. In order of prevalence, these 14 classes are intron, upstream, downstream, intergenic, synonymous, non-synonymous, 3' UTR, 5' UTR, start-gained, stop-gained, stop-lost, synonymous-stop, start-lost and splice-site SNPs (Fig. 1). The reason for cataloging the SNPs in w^{1118} ; iso2; iso-3 is to get a better appreciation of evolution of genome sequences and genome organization in this common laboratory strain. We appreciate the fact that both w^{1118} ; iso-2; iso-3 and y 1 ; cn 1 bw 1 sp 1 are derived and highly manipulated laboratory strains and do not represent natural populations. Therefore, we do not mean to imply that the analyses in this paper are significant but rather just observational. To be meaningful, these observations need to be followed up with natural populations. Hundreds of Drosophila natural populations have already been or are in the process of being sequenced,

so this should be feasible in the near future with a program such as SnpEff. 34

Many of the stop-gained and stoplost SNPs in w^{1118} ; iso-2; iso-3 occur in essential genes that apparently still function after amino acid truncations caused by the stop-gained SNPs (Table 6). These non-critical effects of the stop-gained SNPs are worth noting because nonsense codons in the transcribed mRNAs generally result in nonfunctional protein products. For example, some genetic disorders, such as thalassemia and Duchenne muscular dystrophy (DMD), result from nonsense SNPs. 35-37 Also, nonsense SNP-mediated RNA decay exists in yeast, Drosophila and humans, and usually ensures that mRNAs with premature stop codons are degraded. 38

The stop-gained and stop-lost SNPs in essential genes, if they are validated, could have profound evolutionary implications and suggest the involvement of prions, analogous to [PSI +], in the retention and selection of these SNPs. Brian Cox, a geneticist working with the yeast Saccharomyces cerevisiae, discovered [PSI +] in 1965 as a non-genetically transmissible trait with a cytoplasmic pattern of inheritance similar to mitochondria. 39 He isolated a yeast strain auxotrophic for adenine due to a nonsense mutation is able to survive in media lacking adenine when [PSI +] is present. 39 Reed Wickner showed in 1994 that [PSI +] resulted from a prion form of the translation termination factor, Sup35. 40 Lindquist and colleagues showed in 2008 that the [PSI +] prion provides survival advantages in sev eral stressful environments, such as high SNPs in oc/otd. (B) Protein BLAST of Oc/Otd against the non-redundant (nr) protein database shows salt conditions. 41 They have speculated that only the 60 amino Hox domain flanking amino acid 100 is conserved from Drosophila to humans. The color coding shows the alignment scores. that Sup35 is an evolutionary capacitor + that, when inactivated in

the PSI form, releases cryptic genetic variation that allow expression of novel C-terminal amino acids in hundreds of would allow a modified protein with the new C-terminal tail to proteins, some of which are beneficial in stressful environments. 41 be always expressed, even when the prion is lost. 41 Therefore, a

How might prions be involved in revealing cryptic genetic stop-lost SNP would more likely occur in a strain with benefivariation in the 5' and 3' UTRs? While most prions are thought cial codons in the 3' UTR because the cryptic C-terminal amino to not directly mutate DNA sequences, they could provide an acids encoded by these nucleotides would provide a selective environment that would make the retention and selection of advantage in stressful (i.e., [PSI +]) environments when they are beneficial SNPs more likely. For example, a stop-lost SNP translated.

It is attractive to speculate that a similar prion-mediated evolutionary mechanism might occur in Drosophila, for both stoploss and stop-gained SNPs, and that this might help explain the large number of SNPs that we see in these categories. We note that Drosophila has several Sup35 orthologs, some of which have N-terminal repeats that are known to be potentially prion-forming domains. 41 We acknowledge that this is a highly speculative explanation for the high numbers of start-gained and stop-lost SNPs, but we believe that it is worthy of further investigation.

The many potential start-gained SNPs in Drosophila might also have evolutionary implications. Similar to the cryptic genetic variation that is revealed by stop-lost mutations in the 3' UTR, start-gained SNPs reveal cryptic genetic variation in the 5' UTR. Uncovering the cryptic genetic variation in times of environmental stress, such as by inducing transcription initiation at start sites

upstream of the normally-used transcription start sites, could be one mechanism to facilitate the use of potential start-gained SNPs. Further mutations and selection of the potential start-gained SNPs, such as by introducing better Kozak consensus sequences or more commonly used 5'-AUG-3' translation initiation codons, can stabilize the cryptic genetic variation further if it leads to improved survival or reproductive fitness in a stressful environment. While amino acid extensions and deletions in known essential genes occur only 8 times in w^{1118} ; iso-2; iso-3 compared with the reference strain (Table 7), as laboratories begin to sequence hundreds or even thousands of individuals in a population, extensions and deletions are likely to be found in a large proportion of functional genes.

Finally, we recently upgraded SnpEff further by including over 320 databases for different reference genome versions that can be analyzed (http://snpeff.sourceforge.net/SnpSift.html). Sources of information for creating these databases are ENSEMBL, UCSC Genome Bioinformatics website as well as organism specific databases, such as FlyBase (Drosophila melanogaster), WormBase (C. elegans) and TAIR (Arabidopsis thaliana), to name a few. The program SnpEff is open access and additional genomes can be added and assistance in using SnpEff can be provided upon request. Rapid analyses of whole-genome sequencing data should now be feasible to perform by any laboratory

1.6 Methods

SnpEff overview. The program is divided in two main parts (i) database build and (ii) effect calculation. Part (i) Database build is usually not run by the user, because many databases containing genomic annotations are available. Databases are build using a reference genome, a FASTA file, and an annotation file, usually GTF, GFF or RefSeq table, provided by ENSEMBL,

UCSC Genome Bioinformatics website or other specific websites, such as FlyBase, WormBase and TAIR. SnpEff databases are gzip serialized objects that represent genomic annotations.

Part (ii) Effect calculations can be performed once the user has downloaded, or built, the database. The program loads the binary database and builds a data structure called "interval forest," used to perform an efficient interval search (see next section). Input files, usually in VCF format, are parsed and each variant queries the data structures to find intersecting genomic annotations. All intersecting genomic regions are reported and whenever these regions include an exon, the coding effect of the variant is calculated (hence the name of the program). A list of the reported effects and annotations is shown in Table 2, additional information produced by the program, is shown in Table 3 and Table 4, for different output formats.

SnpEff algorithms. In order to be able to process thousands of variants per second, we implemented an efficient data structure that allows for arbitrary interval overlaps. We created an interval forest, which is a hash of interval trees indexed by chromosome. Each interval tree 42 is composed of nodes. Each node has five elements (i) a center point, (ii) a pointer to a node having all intervals to the left of the center, (iii) a pointer to a node having all intervals to the right of the center, (iv) all intervals overlapping the center point sorted by start position and (v) all intervals overlapping the center point, sorted by end position.

Querying an interval tree requires O (log n + m) time, where n is the number of intervals in the tree and m is the number of intervals in the result. Having a hash of trees, optimizes the search by reducing the number of intervals per tree.

In order to create this the interval forest, genomic information can be parsed from three main annotation formats: GTF (version 2.2), GFF (versions 3 and 2), UCSC Genome Bioinformatics website RefSeqTables and tab separated text files (TXT). Once the interval forest is created, the structure is serialized and compressed (GZIP) into a binary database. There are over 250 genomic binary databases that are currently distributed with SnpEff, which include all genomes from ENSEMBL.

SnpEff accuracy. As part of our standard development cycle, we perform accuracy testing by comparing SnpEff to ENSEMBL "Variant effect predictor," which we consider it is the "gold standard." Current unity testing includes over a hundred test cases with thousands of variants each to ensure predictions are accurate.

SnpEff integration. SnpEff provides integration with third party tools, such as Galaxy, 43 which creates a web based interface for bioinformatic analysis pipelines. Integration with Genome analysis tool kit 4 (GATK) was provided by the GATK team. Detailed information on how to download, install and run, as well as usage examples of the program, can be found at http:/snpEff.sourceforge.net.

Data access. SnpEff Data can be accessed from the Supplemental data file for w^{1118} ; iso-2; iso-3 or by contacting D.M.R.

Disclosure of Potential Conflicts of Interest No potential conflicts of interest were disclosed.

1.7 Acknowledgements

This work was supported by a Michigan Core Technology grant from the State of Michigans 21 st Century Fund Program to the Wayne State University Applied Genomics Technology Center. This work was also supported by the

Environmental Health Sciences Center in Molecular and Cellular Toxicology with Human Applications Grant P30 ES06639 at Wayne

State University, NIH R01 grants (ES012933) to D.M.R. and DK071073 to X.L. We thank David Roazen, Eric Banks and Mark DePristo in the GATK team at the Broad Institute who integrated SnpEff with the Genome Analysis Toolkit (GATK).

Note Supplemental material can be found at: http://www.landesbioscience.com/journals/fly/article/19695

1.8 Epilogue

At the beginning of my Ph.D., functional annotation of genomic variants was an unsolved problem with many research labs creating in-house custom solutions that oftentimes were inefficient and lacking of rigorous testing. As a consequence, shortly after SnpEff & SnpSift were released they quickly became widely adopted by the research community as well as many private organizations. Currently SnpEff & SnpSift has over 250 downloads per week (as reported by SourceForge, where the tools are hosted). So far SnpEff & SnpSift have been cited over 400 times.

1.8.1 Data structures for annotations

A very simple approach used by ANNOVAR [106] is to create an index by dividing each chromosome into N bins of equal size. All genomic features are stored in a hash table indexed by chromosome name and bin number. This approach has running time of $O(n)$ where n is the number of features, but it can be easily tuned by creating small bins, at the cost of increased memory requirements.

Another approach [19] is to use an "interval forest", which is a hash of "interval trees" indexed by chromosome. Each interval tree is composed of nodes. Each node has five elements i) a center point, ii) a pointer to a node

having all intervals to the left of the center, iii) a pointer to a node having all intervals to the right of the center, iv) all intervals overlapping the center point sorted by start position, and v) all interval overlapping the center point sorted by end position. Querying an interval tree requires $O[log(n) + m]$ time, where n is the number of features in the tree and m is the number of features in the result. Having a hash of trees optimizes the search by reducing the number of intervals per tree.

References

[1] Marit Ackermann and Andreas Beyer. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS genetics*, 8(2):e1002463, 2012.

[2] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, and S.R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[3] Orly Agamy, Bruria Ben Zeev, Dorit Lev, Barak Marcus, Dina Fine, Dan Su, Ginat Narkis, Rivka Ofir, Chen Hoffmann, Esther Leshinsky-Silver, et al. Mutations disrupting selenocysteine formation cause progressive cerebello-cerebral atrophy. *The American Journal of Human Genetics*, 87(4):538–544, 2010.

[4] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, and AV Grimstone. Molecular biology of the cell (3rd edn). *Trends in Biochemical Sciences*, 1995.

[5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, et al. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[6] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[7] Gerd Anders, Sebastian D Mackowiak, Marvin Jens, Jonas Maaskola, Andreas Kuntzagk, Nikolaus Rajewsky, Markus Landthaler, and Christoph Dieterich. dorina: a database of rna interactions in post-transcriptional regulation. *Nucleic acids research*, page gkr1007, 2011.

[8] Manuel Ascano, Markus Hafner, Pavol Cekan, Stefanie Gerstberger, and Thomas Tuschl. Identification of rna–protein interaction networks using par-clip. *Wiley Interdisciplinary Reviews: RNA*, 3(2):159–177, 2012.

[9] D.J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[10] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.

[11] Callum J Bell, Darrell L Dinwiddie, Neil A Miller, Shannon L Hateley, Elena E Ganusova, Joann Mudge, Ray J Langley, Lu Zhang, Clarence C Lee, Faye D Schilkey, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science translational medicine*, 3(65):65ra4–65ra4, 2011.

[12] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

[13] Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, et al. Annotation of functional variation in personal genomes using regulomedb. *Genome research*, 22(9):1790–1797, 2012.

[14] Saverio Brogna and Jikai Wen. Nonsense-mediated mrna decay (nmd) mechanisms. *Nature structural & molecular biology*, 16(2):107–113, 2009.

[15] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(suppl 1):D102–D106, 2008.

[16] Susan E Celniker, Laura AL Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

[17] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. *PloS one*, 7(10):e46688, 2012.

[18] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, 2013.

[19] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):0–1, 2012.

[20] Pablo Cingolani, Viral M Patel, Melissa Coon, Tung Nguyen, Susan J Land, Douglas M Ruden, and Xiangyi Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Toxicogenomics in non-mammalian species*, page 92, 2012.

[21] G.M. Clarke, C.A. Anderson, F.H. Pettersson, L.R. Cardon, A.P. Morris, and K.T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[22] Lachlan JM Coin, Julian E Asher, Robin G Walters, Julia S El-Sayed Moustafa, Adam J de Smith, Rob Sladek, David J Balding, Philippe Froguel, and Alexandra IF Blakemore. cnvhap: an integrative population and haplotype–based multiplatform model of snps and cnvs. *Nature methods*, 7(7):541–546, 2010.

[23] FS Collins, ES Lander, J. Rogers, RH Waterston, and I. Conso. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[24] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[25] Diabetes SAT2D Consortium, Diabetes MAT2D Consortium, Anubha Mahajan, Min Jin Go, Weihua Zhang, Jennifer E Below, Kyle J Gaulton, Teresa Ferreira, Momoko Horikoshi, Andrew D Johnson, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–244, 2014.

[26] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[27] UniProt Consortium et al. Update on activities at the universal protein resource (uniprot) in 2013. *Nucleic acids research*, 41(D1):D43–D47, 2013.

[28] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

[29] Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[30] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.

[31] Jasmin Coulombe-Huntington, Kevin CL Lam, Christel Dias, and Jacek Majewski. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS genetics*, 5(12):e1000766, 2009.

[32] Robert Culverhouse, Brian K Suarez, Jennifer Lin, and Theodore Reich. A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471, 2002.

[33] Val Curwen, Eduardo Eyras, T Daniel Andrews, Laura Clarke, Emmanuel Mongin, Steven MJ Searle, and Michele Clamp. The ensembl automatic gene annotation system. *Genome research*, 14(5):942–950, 2004.

[34] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[35] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS computational biology*, 6(12):e1001025, 2010.

[36] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[37] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[38] Stanley D Dunn, Lindi M Wahl, and Gregory B Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008.

[39] R. Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[40] William G Fairbrother, Ru-Fang Yeh, Phillip A Sharp, and Christopher B Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, 2002.

[41] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.

[42] Paul Flicek, Ikhlak Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, et al. Ensembl 2013. *Nucleic acids research*, page gks1236, 2012.

[43] Hong Gao, Julie M Granka, and Marcus W Feldman. On the classification of epistatic interactions. *Genetics*, 184(3):827–837, 2010.

[44] M. Garber, M. Guttman, M. Clamp, M.C. Zody, N. Friedman, and X. Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62, 2009.

[45] G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.

[46] Jeremy Goecks, Anton Nekrutenko, James Taylor, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[47] L Guariguata, DR Whiting, I Hambleton, J Beagley, U Linnenkamp, and JE Shaw. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice*, 103(2):137–149, 2014.

[48] Lukas Habegger, Suganthi Balasubramanian, David Z Chen, Ekta Khurana, Andrea Sboner, Arif Harmanci, Joel Rozowsky, Declan Clarke, Michael Snyder, and Mark Gerstein. Vat: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*, 28(17):2267–2269, 2012.

[49] Markus Hafner, Steve Lianoglou, Thomas Tuschl, and Doron Betel. Genome-wide identification of mirna targets by par-clip. *Methods*, 58(2):94–105, 2012.

[50] D.L. Hartl and A.G. Clark. *Principles of population genetics*. Sinauer associates Sunderland, Massachusetts, 2007.

[51] David Haussler, Stephen J O'Brien, Oliver A Ryder, F Keith Barker, Michele Clamp, Andrew J Crawford, Robert Hanner, Olivier Hanotte, Warren E Johnson, Jimmy A McGuire, et al. Genome 10k: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009.

[52] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2013.

[53] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[54] Fereydoun Hormozdiari, Iman Hajirasouliha, Andrew McPherson, Evan E Eichler, and S Cenk Sahinalp. Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome research*, 21(12):2203–2212, 2011.

[55] Fan Hsu, W James Kent, Hiram Clawson, Robert M Kuhn, Mark Diekhans, and David Haussler. The ucsc known genes. *Bioinformatics*, 22(9):1036–1046, 2006.

[56] Robbie P Joosten, Tim AH Te Beek, Elmar Krieger, Maarten L Hekkelman, Rob WW Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of pdb related databases for everyday needs. *Nucleic acids research*, 39(suppl 1):D411–D419, 2011.

[57] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome browser database: 2014 update. *Nucleic acids research*, 42(D1):D764–D770, 2014.

[58] Martin Alexander Kennedy. Mendelian genetic disorders. *Encyclopedia of Life Sciences*, 2001.

[59] Ching Lee Koo, Mei Jing Liew, Mohd Saberi Mohamad, and Abdul Hakim Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed research international*, 2013, 2013.

[60] P. Kumar, S. Henikoff, and P.C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009.

[61] Lydie Lane, Ghislaine Argoud-Puy, Aurore Britan, Isabelle Cusin, Paula D Duek, Olivier Evalet, Alain Gateau, Pascale Gaudet, Anne Gleizes, Alexandre Masselot, et al. nextprot: a knowledge platform for human proteins. *Nucleic acids research*, 40(D1):D76–D83, 2012.

[62] B. Langmead and S.L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 2012.

[63] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[64] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2012.

[65] B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[66] H. Li. Improving snp discovery by base alignment quality. *Bioinformatics*, 27(8):1157–1158, 2011.

[67] H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[68] H. Li and R. Durbin. Fast and accurate short-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(5), 2009.

[69] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589, 2010.

[70] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078, 2009.

[71] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851, 2008.

[72] Chenxing Liu, Fuquan Zhang, Tingting Li, Ming Lu, Lifang Wang, Weihua Yue, and Dai Zhang. Mirsnp, a database of polymorphisms altering mirna target sites, identifies mirna-related snps in gwas snps and eqtls. *BMC genomics*, 13(1):661, 2012.

[73] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. *Human mutation*, 32(8):894–899, 2011.

[74] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia,

Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[75] D.G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J.K. Pickrell, S.B. Montgomery, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012.

[76] Baijayanta Maiti, Sandrine Arbogast, Valérie Allamand, Mark W Moyle, Christine B Anderson, Pascale Richard, Pascale Guicheney, Ana Ferreiro, Kevin M Flanigan, and Michael T Howard. A mutation in the sepn1 selenocysteine redefinition element (sre) reduces selenocysteine incorporation and leads to sepn1-related myopathy. *Human mutation*, 30(3):411–416, 2009.

[77] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[78] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

[79] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[80] Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

[81] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics*, 26(16):2069–2070, 2010.

[82] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segré, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

[83] Eszter Nagy and Lynne E Maquat. A rule for termination-codon position within intron-containing genes: when nonsense affects rna abundance. *Trends in biochemical sciences*, 23(6):198–199, 1998.

[84] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[85] Sarah B Ng, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E Eichler, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, 2009.

[86] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

[87] Jason ORawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W Evan Johnson, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med*, 5(3):28, 2013.

[88] Umadevi Paila, Brad A Chapman, Rory Kirchner, and Aaron R Quinlan. Gemini: integrative exploration of genetic variation and genome annotations. *PLoS computational biology*, 9(7):e1003153, 2013.

[89] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[90] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

[91] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1):D61–D65, 2007.

[92] Towfique Raj, Manik Kuchroo, Joseph M Replogle, Soumya Raychaudhuri, Barbara E Stranger, and Philip L De Jager. Common risk alleles for inflammatory diseases are targets of recent positive selection. *The American Journal of Human Genetics*, 92(4):517–529, 2013.

[93] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of rna-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

[94] Ho Sung Rhee and B Franklin Pugh. Chip-exo method for identifying genomic location of dna-binding proteins with near-single-nucleotide accuracy. *Current Protocols in Molecular Biology*, pages 21–24, 2012.

[95] Erin Rooney Riggs, Karen E Wain, Darlene Riethmaier, Melissa Savage, Bethanny Smith-Packard, Erin B Kaminsky, Heidi L Rehm, Christa Lese Martin, David H Ledbetter, and W Andrew Faucett. Towards a universal clinical genomics database: the 2012 international standards for cytogenomic arrays consortium meeting. *Human mutation*, 34(6):915–919, 2013.

[96] A.F. Rope, K. Wang, R. Evjenth, J. Xing, J.J. Johnston, J.J. Swensen, B. Moore, C.D. Huff, L.M. Bird, J.C. Carey, et al. Using vaast to identify an x-linked disorder resulting in lethality in male infants due to n-terminal acetyltransferase deficiency. *The American Journal of Human Genetics*, 2011.

[97] Radhakrishnan Sabarinathan, Hakim Tafer, Stefan E Seemann, Ivo L Hofacker, Peter F Stadler, and Jan Gorodkin. The rnasnp web server: predicting snp effects on local rna secondary structure. *Nucleic acids research*, 41(W1):W475–W479, 2013.

[98] Zuben E Sauna and Chava Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10):683–691, 2011.

[99] Gert C Scheper, Marjo S van der Knaap, and Christopher G Proud. Translation matters: protein synthesis defects in inherited disease. *Nature Reviews Genetics*, 8(9):711–723, 2007.

[100] Valerie Schneider and Deanna Church. Genome reference consortium. 2013.

[101] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[102] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[103] GATK team. The genome analysis toolkit. Accessed: 2015.

[104] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[105] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, et al. An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90, 2009.

[106] K. Wang, M. Li, and H. Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[107] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan FA Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome research*, 17(11):1665–1674, 2007.

[108] Lucas D Ward and Manolis Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(D1):D930–D934, 2012.

[109] Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–1106, 2012.

[110] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[111] David C Whitcomb, Michael C Gorry, Robert A Preston, William Furey, Michael J Sossenheimer, Charles D Ulrich, Stephen P Martin, Lawrence K Gates, Stephen T Amann, Phillip P Toskes, et al. Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nature genetics*, 14(2):141–145, 1996.

[112] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[113] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 2011.

[114] Yumi Yamaguchi-Kabata, Makoto K Shimada, Yosuke Hayakawa, Shinsei Minoshima, Ranajit Chakraborty, Takashi Gojobori, and Tadashi Imanishi. Distribution and effects of nonsense polymorphisms in human genes. *PloS one*, 3(10):e3393, 2008.

[115] Jian-Hua Yang, Jun-Hao Li, Peng Shao, Hui Zhou, Yue-Qin Chen, and Liang-Hu Qu. starbase: a database for exploring microrna–mrna interaction maps from argonaute clip-seq and degradome-seq data. *Nucleic acids research*, 39(suppl 1):D202–D209, 2011.

[116] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

[117] Jinying Zhao, Li Jin, and Momiao Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, 2006.

[118] Jesse D Ziebarth, Anindya Bhattacharya, Anlong Chen, and Yan Cui. Polymirts database 2.0: linking polymorphisms in microrna target sites with human diseases and complex traits. *Nucleic acids research*, page gkr1026, 2011.

[119] O. Zuk, E. Hechter, S.R. Sunyaev, and E.S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[120] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.