

# **Stony Brook University**



OFFICIAL COPY

**The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.**

**© All Rights Reserved by Author.**

# **Graphical and machine learning algorithms for large-scale genomics data**

A Dissertation Presented

by

**Han Fang**

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

**Doctor of Philosophy**

in

**Department of Applied Mathematics & Statistics**

Stony Brook University

**August 2017**

Copyright by  
Han Fang  
2017

**Stony Brook University**

The Graduate School

**Han Fang**

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dr. Michael C. Schatz

**Dissertation Advisor**

**Adjunct Professor, Dept. of Applied Mathematics and Statistics**

**Adjunct Associate Professor,**

**Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory**

**Associate Professor, Dept. of Computer Science and Biology, Johns Hopkins University**

Dr. Gholson J. Lyon

**Dissertation Co-advisor**

**Assistant Professor, Cold Spring Harbor Laboratory**

Dr. Song Wu

**Chairperson of Defense**

**Associate Professor, Dept. of Applied Mathematics and Statistics**

Dr. Tom MacCarthy

**Assistant Professor, Dept. of Applied Mathematics and Statistics**

Dr. Rob Patro

**Assistant Professor, Dept. of Computer Science**

This dissertation is accepted by the Graduate School

Charles Taber

Dean of the Graduate School



Abstract of the Dissertation

**Graphical and machine learning algorithms for large-scale genomics data**

by

**Han Fang**

**Doctor of Philosophy**

in

**Applied Mathematics and Statistics**

Stony Brook University

**2017**

One fundamental question in computational genomics is to understand the relationship between genotype and phenotype. In this dissertation, I developed graphical and machine learning algorithms for large-scale genomics data, allowing accurate genotyping and molecular phenotype quantification. This work has helped to shed new light on the genetic contributions to autism spectrum disorders, intellectual disability, and other psychiatric disorders, as well as enabled detailed analysis of the molecular biology of several model organisms.

The first major theme of my research has been in the study of genomic variations, in particular insertion and deletion (indel) mutations. As the second most common type of variations in the human genome, indels have been linked to many diseases, but indels of more than a few bases are still challenging to discover from short-read sequencing data. We present an open-source algorithm, Scalpel, which combines mapping and assembly for sensitive and specific discovery of indels. A detailed repeat analysis coupled with a self-tuning k-mer strategy allows Scalpel to outperform other state-of-the-art approaches for indel discovery, particularly in regions containing near-perfect repeats. We characterized various types of sequencing data to investigate the sources of indel errors. We also developed a classification scheme to rank high and low quality calls.

In a second major theme of research, I present new methods for analyzing ribosome profiling (Riboseq) data, a powerful technique for monitoring protein translation *in vivo*. This, combined with detailed genomic variation data allows researchers to study how the genome influences transcription, translation, and ultimately the overall phenotype of an organism. However, there are prevalent sampling and biological biases in Riboseq data, limiting our ability to understand translation control. To tackle these issues, I developed scikit-ribo, the first open-source software for accurate genome-wide inference of translation efficiency (TE) and A-site prediction. Scikit-ribo accurately identifies ribosome A-site locations even with different mRNA digestion protocols and nearly perfectly reproduces the codon elongation rates in several datasets ( $r=0.99$ ). Next we show the commonly used RPKM-derived TE is very sensitive to sampling errors and biological biases, skewing the TE estimates in all previous studies. To address this, I developed a codon level generalized linear model with ridge penalty to correctly estimate TE while inferring codon elongation rates and mRNA secondary structure. We performed a large-scale validation using mass spectrometry data of 1200 genes and showed very high correlation. Scikit-ribo is particularly robust to low abundance genes that are most commonly distorted by lesser approaches and successfully corrected the TE biases for more than 2000 genes in *S. cerevisiae*. These improvements allow us to discover the Kozak-like consensus sequence in *S. cerevisiae* and a previously undiscovered biological significance in the Dhh1p study. Together, these results show that scikit-ribo substantially improves Riboseq analysis and deepens the understanding of translation control.

## **Dedication Page**

To my advisors and colleagues. Thank you for all the intellectual discussions and incredible collaborations. To my parents, my brother, and all of my friends. Thank you for all of being extremely supportive over the years. To my wife, Pan. Thank you for always being there with me.

## Table of Contents

List of Figures/Tables.....	viii
<b>Chapter 1 Background and significance .....</b>	<b>1</b>
<b>Motivation .....</b>	<b>1</b>
<b>Genome sequencing, insertions and deletions .....</b>	<b>1</b>
<b>Common variant calling approaches.....</b>	<b>2</b>
<b><i>De novo</i> assembly with de Bruijn graph.....</b>	<b>2</b>
<b>Protein translation and ribosome profiling .....</b>	<b>3</b>
<b>Generalized linear model .....</b>	<b>3</b>
<b>Lasso, ridge penalized GLM .....</b>	<b>4</b>
<b>Chapter 2 Indel calling with de Bruijn graph assembly .....</b>	<b>5</b>
<b>Summary of Contribution .....</b>	<b>5</b>
<b>Abstract.....</b>	<b>5</b>
<b>Introduction .....</b>	<b>5</b>
Overview of the Scalpel micro-assembly strategy .....	5
Comparison to other methods.....	6
Overview of the computational protocol .....	7
<b>Methods.....</b>	<b>8</b>
Experimental Design .....	8
The Scalpel pipeline .....	10
Graph construction .....	10
Repeat Analysis.....	11
Graph traversal .....	11
Computation protocol.....	12
<b>Results .....</b>	<b>13</b>
High accuracy of Scalpel.....	13
Expected distribution of indels and signatures of low-quality calls .....	14
Expected number of indels during the filtering cascade.....	14
A list of frame-shift mutations in the family .....	14
Limitations of the protocol and software.....	15
<b>Tables and figure in this chapter.....</b>	<b>16</b>
Tables .....	16
Figures.....	16
<b>Chapter 3 Characterizing the sources of indel errors .....</b>	<b>25</b>
<b>Summary of Contribution .....</b>	<b>25</b>
<b>Abstract.....</b>	<b>25</b>
<b>Introduction .....</b>	<b>25</b>
<b>Methods.....</b>	<b>27</b>

Analysis of Simulated Data, WGS and WES data .....	27
Generation of MiSeq validation data.....	29
Classifications of indels with calling quality.....	30
<b>Results .....</b>	<b>31</b>
Characterizing alignment and assembly based callers at different coverage .....	31
WGS vs. WES: Low concordance on indel calling.....	32
Coverage distributions of different regions in WGS and WES data.....	32
MiSeq validation of indels in WGS and WES data on the sample K8101-49685s .....	33
Assessment of the indels calls sets from WGS and WES.....	34
Sources of multiple signatures in WGS and WES data .....	34
Standard WGS vs. PCR-free: assessment of indels calling quality .....	35
What coverage is required for accurate indel calling?.....	36
<b>Discussion .....</b>	<b>38</b>
<b>Tables and figures in this chapter .....</b>	<b>39</b>
Figures.....	39
Tables.....	44
<b>Chapter 4 Benchmarking and applications of Scalpel .....</b>	<b>47</b>
<b>Summary of Contribution .....</b>	<b>47</b>
<b>Abstract.....</b>	<b>47</b>
<b>Introduction .....</b>	<b>47</b>
<b>Results .....</b>	<b>48</b>
<b>Figures and tables in this chapter .....</b>	<b>51</b>
Figures.....	51
<b>Chapter 5 Accurate inference and robust modelling of translation dynamics at codon resolution.....</b>	<b>53</b>
<b>Summary of Contribution .....</b>	<b>53</b>
<b>Abstract.....</b>	<b>53</b>
<b>Introduction .....</b>	<b>53</b>
<b>Results .....</b>	<b>56</b>
Accurate A-site codon prediction with different organisms and nuclease digestion .....	56
Paused ribosomes and biological biases of TE .....	57
Sampling errors for low abundance genes using Riboseq .....	58
Accurate inference reveals the interplay between cognate tRNA availability and mRNA secondary structure.....	58
Simultaneously correcting sampling errors and biological biases for TEs .....	59
Scikit-ribo discovers Kozak-like consensus in <i>S. cerevisiae</i> .....	60
Large-scale validation showed Scikit-ribo’s accurate TE estimation, especially for low-abundance genes.....	61
Refined TE analysis revealed Dhh1p’s role in translation repression .....	62
Overview of Scikit-ribo.....	64
Ribosome A-site codon prediction .....	64

Calculating RPKM-derived TE .....	65
Correcting for biological biases with the Scikit-ribo GLM .....	65
Correcting for sampling errors with ridge penalty .....	66
Deriving relative protein abundance .....	67
Sequencing reads processing.....	67
Scikit-ribo input processing.....	68
Data and statistical analysis in this paper .....	68
Simulation, sequence enrichment, and gene enrichment analysis .....	68
<b>Figures and tables in this chapter .....</b>	<b>70</b>
Figures.....	70
<b>Supplemental figures.....</b>	<b>76</b>
<b>Chapter 6 Applications on genome informatics .....</b>	<b>98</b>
<b>Summary of contribution .....</b>	<b>98</b>
<b>Applications.....</b>	<b>98</b>
Application 1 - Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine .....	98
Application 2 - GenomeScope: fast reference-free genome profiling from short reads .....	99
<b>Figures and tables in this chapter .....</b>	<b>101</b>
Figures.....	101
<b>Chapter 7 Conclusions and Perspectives .....</b>	<b>103</b>
<b>Conclusions and contributions of this thesis .....</b>	<b>103</b>
<b>Applications of this research .....</b>	<b>104</b>
<b>Further directions .....</b>	<b>104</b>
<b>Bibliography.....</b>	<b>106</b>

## List of Figures/Tables

<b>Figures</b>	<b>Page</b>
Figure 2.1. Main steps in the Scalpel protocol	17
Figure 2.2. Overview of the indel variant filtering cascade	18
Figure 2.3. Higher coverage can improve Scalpel's sensitivity performance of indel detection with WGS data	18
Figure 2.4. Comparison of standard WGS and PCR-free data based on indel quality	19
Figure 2.5. Whole genome mutational concordance	20
Figure 2.6. High accuracy of indel detection using Scalpel on WGS data	20
Figure 2.7. Size distribution of inherited and <i>de novo</i> indels	21
Figure 2.8. Histograms of low quality homopolymer indels by category	21
Figure 2.9. Variant allele fractions (VAF %) of the inherited indels	22
Figure 2.10. Filtering cascade of inherited and <i>de novo</i> indel calls	22
Figure 2.11. Frame-preserving indels are more abundant within coding sequences (CDS)	23
Figure 2.12. Screenshot of the alignment of the <i>de novo</i> deletion in IGV browser	24
Figure 3.1. Performance comparison between the Scalpel and GATK-UnifiedGenotyper in terms of sensitivity	39
Figure 3.2. Mean concordance of INDELS over eight samples between WGS (blue) and WES (green) data	39
Figure 3.3. Coverage distributions of the exonic targeted regions in (A) the WGS data, (B) the WES data	40
Figure 3.4. Coverage distributions of the WGS-specific INDELS regions in (A) the WGS data, (B) the WES data	40
Figure 3.5. Percentage of high quality, moderate quality and low quality INDELS in three call set.	41
Figure 3.6. Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR in three call set	41
Figure 3.7. Numbers of genomic locations containing multiple signature INDELS	42
Figure 3.8. Percentage of reads near regions of Non-homopolymer, poly-N, poly-A, poly-C, poly-G, poly-T	42
Figure 3.9. Concordance of INDEL detection between PCR-free and standard WGS data on NA12878	43
Figure 3.10. Percentage of high quality, moderate quality and low quality INDELS.	43
Figure 3.11. Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR	44
Figure 3.12. Sensitivity performance of INDEL detection with eight WGS datasets at different mean coverages on Illumina HiSeq2000 platform	44
Figure 4.1. Overview of the Scalpel algorithm workflow	52
Figure 4.2. Concordance of indels between pipelines	53

Figure 4.3. MiSeq validation	53
Figure 5.1. Sources of biases in using ribosomes densities per mRNA (RPKM-derived TE) as a proxy for TE	70
Figure 5.2. Overview of the analysis workflow in scikit-ribo	71
Figure 5.3. Accurate inference of codon elongation rates and mRNA secondary structure	72
Figure 5.4. Pair-wise comparisons of estimates between Scikit-ribo and RPKM-derived TE	73
Figure 5.5. Large-scale validation with mass spectrometry data showed Scikit-ribo's accurate TE estimates, especially for low-abundance genes	74
Figure 5.6. Reproducing and new findings of the <i>Dhh1p</i> data	75
Supplemental figure 5.S1. RPKM-derived log <sub>2</sub> (TE) and scikit-ribo log <sub>2</sub> (TE)	76
Supplemental figure 5.S2. Multi-class ROC curves for A-site prediction	77
Supplemental figure 5.S2. Feature importance from the random forest model	78
Supplemental figure 5.S4 Analysis of mRNA abundance in TPM by region	79
Supplemental figure 5.S5. Violin plots of stAI for genes in the six regions	80
Supplemental figure 5.S6. Statistically enriched sequences based on scikit-ribo's TIE estimates	81
Supplemental figure 5.S7. Statistically enriched sequences based on RPKM-derived TE	82
Supplemental figure 5.S8. Higher correlation between scikit-ribo derived PA and SRM measurement, after considering protein degradation rate	83
Supplemental figure 5.S9. Substantial differences of TE between strains	84
Supplemental figure 5.S10. Highly reproducible TE estimates between replicates	85
Supplemental figure 5.S10. Highly reproducible TE estimates between replicates	86
Supplemental figure 5.S12. Codon dwell time (DT) comparisons between strains	87
Supplemental figure 5.S13. Reproducing Radhakrishnan et al's findings on TE changes and codon optimality	88
Supplemental figure 5.S14. Genes with extreme TE changes that were unique to scikit-ribo	89
Supplemental figure 5.S15. The complete workflow of scikit-ribo analysis	90
Figure 6.1 Screenshot of three heterozygous <i>de novo</i> deletions	101
Figure 6.2. GenomeScope Heterozygosity Categories	102

<b>Tables</b>	<b>Page</b>
Table 2.1. Comparisons and validation rates of indel detection with WGS and WES.	16
Table 2.2. Expected QC-passed read and mapping statistics	16
Table 3.1. Mean concordance and discordance rates of INDEL detection between WGS and WES data in different regions	45
Table 3.2. Mean coefficients of variation of coverage with respects to the three regions	45
Table 3.3. Validation rates of intersection, WGS-specific, and WES-specific INDELS	45
Table 3.4. Number and fraction of large INDELS in the following INDEL categories	45
Supplemental Table 5.S1. Prediction accuracy of A-site locations	91
Supplemental table 5.S2. Interpretation of the pair-wise comparison in Figure 5.4	92

Supplemental Table 5.S3. Gene set enrichment in region 4 genes	93
Supplemental Table 5.S4. Relative codon elongation rate (ER) and dwell time (DT) in the Dhh1p study	94
Supplemental Table 5.S5. The GO enrichment of gene with reduced TE in OE, relative to WT in the Dhh1p analysis	96
Supplemental Table 5.S6. The GO enrichment of gene with increased TE in KO, relative to WT in the Dhh1p analysis	97



## Acknowledgments

During my graduate career, I was lucky enough to have worked with the most amazing scientists in the field.

First I would like to thank my advisor, Dr. Michael Schatz. You have always been a scientific mentor and a friend to me. It was quite a privilege to pursue my PhD research with you for three and a half years. Throughout the process, I enjoyed tackling the most challenging problems with you, many of which were previously undefined in the field. I am extremely grateful for both the guidance and freedom that you gave me. To me, you found the perfect balance between the two sides, which makes my fruitful graduate career possible, evolving from a student to a researcher. But more than that, you are a role model to me; I am deeply impressed and also inspired by your way of leading the lab, interacting with colleagues, and solving scientific problems. This is definitely the kind of scientist I wish to become one day. And I hope to continue learning from you going forward. Thank you.

I also want to thank my advisor, Dr. Gholson Lyon. You were one of the first to recognize the potentials in me. At our first interview, I was still trying to sort out my life and my future directions. Your passion and critical thinking of science, deeply inspired me to pursue this path. You encouraged me to be independent from day one, which I appreciate since then. The ability to have independent thinking for complex problems has become one of my most unique assets. Thank you.

I would like to thank the members of my dissertation committee, Dr. Rob Patro, Dr. Tom MacCarthy, and Dr. Song Wu. Thank you for the valuable advice and discussions on the dissertation. You have all selflessly contributed to help improve my research, taking the quality of my work to a whole new level. Thank you.

I would also like to thank my Masters research advisor, Dr. Stephen Finch. You opened up the door of statistical genetics to me, helped me realize the fascinating field of computational biology. Thank you.

I would like to thank my fellow students and colleagues, Giuseppe Narzisi, Fritz Sedlazeck, Ruibang Luo, Hayan Lee, Tyler Garvin, James Gurtowski, Srividya Ramakrishnan, Rob Aboukhalil, Maria Nattestad, Yiyang Wu, Charlotte Darby, Samuel Kovaka, and many others. You all have provided really useful feedbacks on the many practice talks that I made you sit through. It was my pleasure to have worked with you.

I would also like to thank my parents, Yanfen Lin and Liqun Fang, my brother Mingtao Fang, my parents-in-law, Naiqin Li and Xuefeng Teng, my grandparents, Suliang li and Huiqing Li, and all of my families and relatives. Thank you for your love and your continuous support on my career.

Finally, I would like to thank my wife, Pan, for supporting me through the ups and downs. Thank you.

Again, thank you to everyone.

## Vita, Publications and/or Fields of Study

### First-author manuscripts:

- **Fang**, Huang, Radhakrishnan, Siepel, Lyon, Schatz, "Scikit-ribo: Accurate estimation and robust modelling of translation dynamics at codon resolution", In submission (2017)
- **Fang**, Wu, Yoon, Jiménez-Barrón, Mittelman, Robison, Wang, Lyon, "Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine", *BMC Medical Genomics* (2017)
- **Fang**, Bergmann, Arora, Vacic, Zody, Iossifov, O’Rawe, Wu, Jimenez Barron, Rosenbaum, Ronemus, Lee, Wang, Dikoglu, Jobanputra, Lyon, Wigler, Schatz, Narzisi, "Indel variant analysis of short-read sequencing data with Scalpel", *Nature Protocols* (2016)
- **Fang**, Wu, Narzisi, O’Rawe, Jimenez Barron, Rosenbaum, Ronemus, Iossifov, Schatz, Lyon, "Reducing INDEL calling errors in whole genome and exome sequencing data", *Genome Medicine* (2014)

### Co-author manuscripts:

- Yang, Chen, Lima, **Fang**, Jimenez, Li, Lyon, He, Wang, "PennCNV-Hadoop: Accurate Detection of Copy Number Variation from Whole Genome Sequencing Data", Under review (2017)
- Vurture, Sedlazeck, Nattestad, Underwood, **Fang**, Gurtowski, Schatz, "GenomeScope: Fast reference-free genome profiling from short reads", *Bioinformatics* (2017)
- Doerfel, **Fang**, Crain, Klingener, Weiser, Lyon, "Proteomic and genomic characterization of a yeast model for Ogden syndrome", *Yeast* (2016)
- O’Rawe, Wu, Doerfel, Rope, Billie Au, Parboosingh, Moon, Kousi, Kosma, Smith, Tzetis, Schuette, Hufnagel, Prada, Martinez, Orellana, Crain, Caro-Llopis, Oltra, Monfort, Jiménez-Barrón, Swensen, Ellingwood, Smith, **Fang**, Ospina, Stegmann, Den Hollander, Mittelman, Highnam, Robison, Yang, Faivre, Roubertie, Rivière, Monaghan, Wang, Davis, Katsanis,

Kalscheuer, Wang, Metcalfe, Kleefstra, Innes, Kitsiou-Tzeli, Rosello, Keegan, Lyon, "TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations", *American Journal of Human Genetics* (2015)

- Jimenez-Barron, O’Rawe, Wu, Yoon, **Fang**, Iossifov, Lyon, "Genome Wide Variant Analysis of Simplex Autism Families with an Integrative Clinical-Bioinformatics Pipeline". *Molecular Case Studies* (2015)
- Narzisi, O’Rawe, Iossifov, **Fang**, Lee, Wang, Wu, Lyon, Wigler, Schatz, "Accurate detection of de novo and transmitted INDELS within exome-capture data using micro-assembly", *Nature Methods* (2014)
- O’Rawe, **Fang**, Rynearson, Robison, Kiruluta, Higgins, Eilbeck, Reese, Lyon, "Integrating precision medicine in the study and clinical treatment of a severely mentally ill person", *PeerJ* (2014)

# Chapter 1 Background and significance

## Motivation

In the field of computational genomics, one fundamental question is to understand the relationship between genotype and phenotype. This requires accurate methods to be developed for identifying genetic mutations and quantifying molecular phenotypes. Reductions in the cost of whole genome sequencing (WGS) and whole exome sequencing (WES) are opening the door for affordable sequencing of patients and the development of precision medicine<sup>1</sup>. Historically, genomic studies have focused on single nucleotide polymorphisms (SNPs) due to their high prevalence and relative simplicity to detect<sup>2</sup>. However, recent advancements in sequencing technologies and computational methods have broadened the focus to include the role of insertion and deletion (indel) mutations, which were very challenging to detect from short-read sequencing data. With respect to molecular phenotypes, there has been a low correlation between mRNA abundance and protein abundance, suggesting a role of translation regulation<sup>3</sup>. Previously, scientists used the mass spectrometry assay to quantify the abundance of peptides, but the throughput of this technique is low<sup>4</sup>. Polysome profiling can be used to measure the mRNA constituents of different ribosome number fractions, but it does not provide information about ribosome position<sup>5</sup>. Recently, ribosome profiling (Riboseq) emerged as a high throughput method for monitoring protein translation, while providing precise ribosome locations on the mRNA. In this dissertation, I developed graphical and machine learning algorithms for large-scale genomics data, in particular new methods for the accurate genotyping of indels, and molecular phenotype quantification from the Riboseq data. I also demonstrate how these methods can be applied to help understand the genetic contributions to autism spectrum disorders and other major human disorders as well as to better measure and understand how genomic variants interplay with gene transcription and translation in several model species.

## Genome sequencing, insertions and deletions

Researchers use genome sequencing to determine the order of nucleotides within the DNA molecules of a sample, especially so that they can be compared to a reference genome to identify any genetic mutations. The most widely used sequencing assays include HiSeq and NextSeq series sequencing platforms from Illumina ([illumina.com](http://illumina.com)) that can produce many billions of short sequencing reads per run. For studies focusing on the coding regions, scientists typically choose whole exome sequence (WES) as a cost-effective approach, where molecular probes are used to enrich for those molecules spanning exonic sequences. If non-coding regions and structural variants (SV) are of interest, one will choose whole genome sequencing (WGS) instead which sequences the entire sample without enrichment for any particular regions. Other technologies include DNA microarray arrays, which have been a popular choice as they can be scaled to a population study, although with limited resolution of the genome. Long-read sequencing, including single molecule real time sequencing from Pacific Bioscience ([pacb.com](http://pacb.com)) and nanopore sequencing from Oxford Nanopore Technologies ([nanoporetech.com](http://nanoporetech.com)) have proved their power in genome assembly and SV detection. More recently, linked-read technologies from 10x Genomics ([10xgenomics.com](http://10xgenomics.com)) have enabled us to easily phase the genome, as well as performing single cell sequencing more efficiently.

However, despite this diversity of technologies, one of the most challenging problems remaining is to accurately call insertions and deletions (indels) from the genome. Indels are the second most common type of variations in the human genome. They are defined by the addition or loss of one or more nucleotides of a DNA sequence. In coding regions of the genome, if an indel occurs and its length is not a multiple of 3, it is considered as a frameshift mutation, because it disrupts the canonical open reading frame (ORF) of the gene. Otherwise, it is considered as an in-frame mutation; one or more codons will be inserted/deleted from the gene and hence one or more amino acids will be added/removed to the protein but otherwise the amino acid sequence will remain the same. Frame-shift mutations are a highly disruptive class of indel mutations, which have been strongly implicated in neurodevelopment disorders, cardiovascular diseases, cancer, and many other human diseases<sup>6-9</sup>. Studies have shown widespread occurrences of loss-of-function variants, especially indels, in protein-coding genes of human, plant and other species<sup>10-12</sup>.

### Common variant calling approaches

A common approach for variant calling (SNPs, indels, or other types of variants) is to align reads one at a time to a reference genome, and to recognize when the reads disagree from the reference<sup>13, 14</sup>. We refer to these methods as alignment based variant calling. Although this approach works well for SNPs, it is less reliable for indel detection. For example, reads containing a long insertion will contain few bases matching the reference and will fail to map correctly. While reads supporting a deletion consist of bases from the reference, it may be hard to unambiguously map both sides of the deletion. In both cases the aligner may ignore parts of the reads (“soft-clip”) in order to place them on the reference or fail to map them at all. Earlier methods for indel detection relied on paired-end and split-read information as a computational signature for the presence of an indel. Some tools such as GATK UnifiedGenotyper<sup>13</sup>, SAMtools<sup>15</sup>, Dindel<sup>14</sup>, and 16GT<sup>16</sup> use paired-end information to screen for indels where one read of a pair aligns well but the other pair does not. After identifying such regions, the algorithms use a local realignment of the reads to detect indels, although the sensitivity declines quickly for mutations longer than 5bp<sup>17</sup>. By using split-read information where the alignment for an individual read is split into two segments spanning structural variation breakpoints, methods like Pindel<sup>18</sup> and Splitread<sup>19</sup> are able to detect indels, especially deletions. Theoretically, this approach should be effective for deletions of any size, but the sensitivity is reduced due to the read length of current sequencing technologies. Recently, there has been much interest in developing specialized local assembly and micro-assembly methods for variant calling<sup>20</sup>, including Platypus<sup>21</sup>, GRIDSS<sup>22</sup>, and SvABA<sup>23</sup>. It was shown that micro-assembly based methods were more sensitive in detecting larger indels than alignment based methods<sup>17</sup>.

### *De novo* assembly with de Bruijn graph

In computer science graph theory, a de Bruijn graph is a directed graph representing overlaps between sequences of symbols<sup>24</sup>. For a given  $n$ -dimensional de Bruijn graph with  $m$  symbols, there are in total  $m^n$  vertices, consisting of all possible length- $n$  sequences and the same symbol may occur multiple times in a sequence. It has some practical usage and applications in grid network, distributed hash table, and bioinformatics. Specifically, it has demonstrated its unique power in *de novo* genome assembly from short reads<sup>25</sup>. Many have developed short read assemblers based on this technique, including Velvet<sup>26</sup>, MEGAHIT<sup>27</sup>, Allpath-LG<sup>28</sup>. Some have

further attempted to utilize whole-genome *de novo* assembly with de Bruijn graphs for variant calling and developed Cortex<sup>29</sup>. But in practice this method is less sensitive than expected and accurate indel detection instead requires a fine-grained and localized analysis. That is why many in the community started to investigate the practical usage of local *de novo* assembly for variant calling. The basic idea is to localize the reads in a sliding window, using alignment towards a given reference genome. This avoids over-exhaustive search of the entire genome. Then the reads in the same window are sheared into *k*-mers and assembled into a de Bruijn graph. Once the contig is formed, one can then align it back to the reference genome and retrieve the mutations.

### Protein translation and ribosome profiling

Once the genome sequence has been determined, the central dogma of molecular biology explains the flow of genetic information within a biological system, which has been described as "DNA makes RNA and RNA makes protein"<sup>30</sup>. Thanks to the high throughput methods for mRNA profiling, many studies have analyzed gene regulation at the transcriptional level. However, in terms of protein translation, there have been limited numbers of assays that could be used at a genome-wide level. Fortunately, Riboseq is a powerful technique for monitoring protein translation *in vivo*. The original protocol was introduced by Ingolia et al in 2009<sup>3</sup>. It allows researchers to investigate genome-wide translation regulation in a high throughput manner<sup>31</sup>, and has led to discoveries of new mechanisms involving translational defects in different forms of cancer<sup>32-35</sup> and other important human diseases<sup>36, 37</sup>. Reports of novel drug targets<sup>38, 39</sup> and new biological processes around translation<sup>40, 41</sup> have also been made using Riboseq. One key measurement from the Riboseq data is the translational efficiency (TE), which is defined as the rate of protein production per mRNA<sup>42</sup>. This measurement tells us the level of translational control for each gene, independently of mRNA transcription.

### Generalized linear model

The generalized linear model (GLM) is a generalization of ordinary linear regression, that unifies linear regression, logistic regression and Poisson regression by allowing response variables that have error distribution models other than a normal distribution<sup>43</sup>. This is achieved by establishing the relationship between the linear predictor and the mean of the distribution function using a link function. For example, in Poisson regression, a log link function is usually employed, while in a logistic regression, a logit link function is employed instead. The maximum likelihood estimation of the GLM's model parameters is typically done by iteratively reweighted least squares (IRLS)<sup>44</sup>. In a GLM, each outcome  $\mathbf{Y}$  (dependent variables) is assumed to be generated from a distribution in the exponential family. The mean ( $\boldsymbol{\mu}$ ) of the distribution depends on the independent variables ( $\mathbf{X}$ ), parameters ( $\boldsymbol{\beta}$ ), and a link function  $g(\cdot)$ :

$$E[\mathbf{Y}] = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \quad \text{Equation 1.1}$$

The linear predictor ( $\boldsymbol{\eta}$ ) is related to the expected value of the data through the link function, which can be expressed as a linear combination of parameters ( $\boldsymbol{\beta}$ ) and independent variables ( $\mathbf{X}$ ):

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \quad \text{Equation 1.2}$$

In the Chapter 5 of this dissertation, I aim to model the count data for Ribosome profiling with a Poisson regression, the un-penalized log likelihood function for the observations  $\{\mathbf{x}_k, y_k\}_1^N$  is given by

$$l(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \sum_{i=k}^N (y_k(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_k) - e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_k}) \quad \text{Equation 1.3}$$

### Lasso, ridge penalized GLM

In modern statistics, high-dimensional data ( $p > n$ ) imposes a huge challenge for the traditional GLM because it is more likely to result in multicollinearity and overfitting. When overfitting occurs, the fitted model becomes excessively complex, and the model tend to describe random errors instead of the underlying relationship<sup>45</sup>. To overcome these issues, many regularization and variable selection methods have been proposed, including ridge ( $l_2$ ) and lasso ( $l_1$ ) regularization. Ridge regression was proposed by Hoerl and Kennar<sup>46</sup>. It finds the coefficients minimizing the sum of squared error loss subject to an  $l_2$  norm constraint on the coefficients.

$$\hat{\boldsymbol{\beta}}_{ridge}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \quad \text{Equation 1.4}$$

Lasso (least absolute shrinkage and selection operator) was introduced by Tibshirani<sup>47</sup>. It penalizes the size of the  $l_1$  norm of the coefficients, and determines the coefficients with the following:

$$\hat{\boldsymbol{\beta}}_{lasso}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \quad \text{Equation 1.5}$$

In the context of the generalized linear model, one aims to fit a model with a penalized maximum likelihood with ridge ( $l_2$  norm), lasso ( $l_1$  norm), or elastic-net<sup>48</sup> (the mixture of both). Given the weights  $w_i$ , the elastic-net penalty  $\alpha$ , the overall penalty  $\lambda$ , and the likelihood function  $l(\cdot)$ , the problem can be formulated as the following:

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^N w_i l(y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) + \lambda [(1 - \alpha) \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1] \quad \text{Equation 1.6}$$

In the Chapter 5 of this dissertation, I optimize the  $l_2$  norm penalized log likelihood for the Poisson regression model:

$$\operatorname{argmin}_{\beta_0, \boldsymbol{\beta}} -\frac{1}{N} l(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) + \lambda \left( \sum_{k=1}^K \beta_k^2 / 2 \right) \quad \text{Equation 1.7}$$

# Chapter 2 Indel calling with de Bruijn graph assembly

## Summary of Contribution

This chapter describes the de Bruijn graph assembly based variant caller, Scalpel and its computational protocol. The algorithms and benchmarking results of Scalpel were published in *Nature Methods*<sup>49</sup>. The computation protocol and related new results were published in *Nature Protocols*<sup>50</sup>. Han Fang developed the computational protocol for performing indel variant analysis, the benchmarking against competing algorithms, and contributed to the development of Scalpel. Giuseppe Narzisi implemented the local de Bruijn graph assembler inside Scalpel, led the analyses in the first paper, and continues to be the lead developer. Michael Schatz contributed to the development of Scalpel and wrote the microsatellite detector module. Permission for republication of this material has been granted and is available upon request.

## Abstract

As the second most common type of variations in the human genome, insertions and deletions (indels) have been linked to many diseases, but indels of more than a few bases are still challenging to discover from short-read sequencing data. Scalpel (<http://scalpel.sourceforge.net>) is an open-source software for reliable indel detection based on the micro-assembly technique. To date, it has been successfully used to discover mutations in novel candidate genes for autism, and is extensively used in other large-scale studies of human diseases. This chapter gives an overview of the algorithm and describes how to use Scalpel to perform highly accurate indel calling from whole genome and exome sequencing data. I provide detailed instructions for an exemplary family-based de novo study and I also characterize the other two supported modes of operation for single sample and somatic analysis. Indel normalization, visualization, and annotation of the mutations are also illustrated. Using a standard server, indel discovery and characterization in the exonic regions of the example sequencing data can be finished in around 5 hours after read mapping.

## Introduction

### Overview of the Scalpel micro-assembly strategy

Scalpel is a computational tool specifically designed to detect indels in next-generation sequencing (NGS) data. **Figure 2.1** outlines the main steps for the analysis of a sequencing dataset using Scalpel. To highlight the main focus of this protocol, the left panel of **Figure 2.1** depicts the specific scenario of detecting *de novo* indels in a quartet family composed of two parents and two children. I highly recommend reviewing the original Scalpel publication for a more extensive description of the method<sup>51</sup>. Here I describe the main ideas used in the micro-assembly strategy employed by Scalpel, the strategies and filters that can be applied for optimizing the accuracy with different experimental designs or sequencing conditions, and describe the new developments since the original publication of the software (v0.1.1 beta).

Before running Scalpel, the sequencing reads (whole genome, whole exome, or custom capture) must be aligned to a reference genome using a short-read mapping algorithm such as BWA-MEM, similar to the steps used for SNP calling or other analyses. It is worth noting that



computationally expensive procedures like indel realignment and base quality recalibration are not necessary with Scalpel. Unlike in those analyses, the alignments are not directly used to find indels but instead are used to localize the analysis into computationally tractable regions. After alignment, Scalpel examines all the genomic regions provided in the input by the user in BED format (right panel, **Figure 2.1**). For each region, reads that align in the region or whose mates align in the region are extracted from the alignment and assembled independently of the reference using a de Bruijn assembly paradigm. If the size of a region is larger than the user-defined window size parameter, a sliding-window approach will be performed over this target region based on the window size and step size parameters. In order to reduce the number of errors in locally highly repetitive regions, Scalpel automatically performs a local repeat analysis coupled with a self-tuning  $k$ -mer strategy that iteratively increases the  $k$ -mer size until a “repeat-free” local assembly graph is built. A repeat-free graph is a graph without exact repeats, which would introduce cycles in the de Bruijn graph, as well as near-identical repeats (up to 3 mismatches by default). The advantage of this strategy is that every genomic window will be analyzed using an optimal  $k$ -mer specifically tuned according to its sequence composition. The graph is then exhaustively explored to identify end-to-end paths spanning the selected region. These paths, representing *de novo* assembled sequences of the short reads, are then aligned to the reference window to detect candidate mutations using a sensitive gapped sequence aligner based on the Smith–Waterman algorithm.

Scalpel supports three modes of operation: *single*, *de novo*, and *somatic*. In the single mode, Scalpel detects indels in one single dataset (e.g., one individual exome or genome). In the *de novo* mode, Scalpel detects *de novo* indels in a quad family (father, mother, affected child, unaffected sibling). In the somatic mode, Scalpel detects somatic indels from the sequencing data coming from matched tumor and normal samples. In the first version of Scalpel (v0.1.1), all possible paths in the final graph were exhaustively examined using a breadth-first-search traversal approach. This strategy worked well for the majority of the human genome with limited numbers of mutations leading to the generation of one or two paths. However, this step is computationally expensive for a small number of regions with high level of heterozygosity or higher sequencing error rate that generate exponentially many alternative paths due variants not linked by the same  $k$ -mer. Since the release of a new version (0.4.1), Scalpel instead enumerates only the minimum number of source-to-sink paths that cover every edge of the graph using a network flow approach. This strategy still detects all the mutations in the graph but significantly reduces the computational requirements by aligning to the reference a much smaller set of paths. Another important addition in the new version of Scalpel is the ability to better handle regions characterized by sudden drops in coverage. After removal of low-coverage nodes, the de Bruijn graphs associated to these regions can be disconnected into multiple connected components, which are now analyzed independently. Finally, the somatic mode of Scalpel is entirely new since the previous publications.

### Comparison to other methods

Several hundred software packages are now available for analyzing WGS and WES sequence data <sup>52</sup>, including dozens of methods each for quality assessment, read alignment, variant identification, annotation, and other applications. Most of the variant analysis packages are specialized for detecting one or a few types of mutations because each type requires a different computational and statistical framework. For example, SNPs are generally found

directly from read alignments, copy number variations (CNVs) and structural variations (SVs) from read coverage and/or split-read approaches, while the leading methods for detecting indels rely on alignment or localized sequence assemblies. A few other indel-finding software packages implement a localized sequence assembly strategy similar to the one employed by Scalpel. These include GATK HaplotypeCaller<sup>53</sup>, SOAPindel<sup>54</sup>, Platypus<sup>21</sup>, ABRA<sup>55</sup>, TIGRA<sup>56</sup>, DISCOVAR<sup>57</sup>, Bubbleparse<sup>58</sup>, Manta<sup>59</sup>, and ScanIndel<sup>60</sup>. Although they all employ a local read assembly step, these tools differ in how they explore the graphs and in their relative ability to handle repeat structures. Scalpel is unique because of on-the-fly repeat analysis that it uses to automatically optimize the parameters used for different regions of the genome, and the extensive set of filters that can be applied to correct for different sequencing conditions, among several other enhancements. The combination of these features enables Scalpel to accurately identify indel variants in diverse sequencing conditions and sequence contexts. Small-scale repeats are especially challenging for most other indel finding algorithms, although these are detected and properly analyzed by Scalpel.

Most indel-finding tools, including Scalpel, have been designed to be general variant callers for detecting mutations across every region of the reference genome. However, some class of indels, specifically the one located within short tandem repeats (STR), are known to be inherently more difficult to detect due to the high level of replication slippage events (e.g., homopolymers) of Illumina technology. Very few tools have been designed to specifically deal with the complexity of calling within STR regions. Users that specifically require to call variants within STRs are highly recommended to employ the following two tools: RepeatSeq (Highnam et al., 2013) and lobSTR (Gymrek et al., 2012). More recently more complex classes of indels have been also discovered and analyzed where a simultaneous deletion and insertion of DNA fragments of different sizes can co-occur at the same genomic location. A new tool has been specifically designed to handle these complex indels<sup>61</sup>, Pindel-C, and I encourage the user to utilize such a tool for detecting complex indels in cancer-associated genes.

## Overview of the computational protocol

In the protocol section, I present a step-by-step guideline for identifying *de novo* variants in a HapMap family from PCR-free Illumina HiSeq2000 data. Here, I provide an overview of using Scalpel to discover *de novo* and inherited indel mutations within a quad family of two parents and two children, one affected and one unaffected with a certain phenotype. It should be noted that internally within the algorithm the two children are treated identically, which can support additional use cases. The input to the algorithm can be data from WGS, WES, or targeted sequencing experiments. A two-pass search mode is employed by Scalpel when calling *de novo* or somatic mutations. In the first pass, Scalpel identifies indels in each of the samples using parameters designed to balance between sensitivity and specificity. In the second pass, Scalpel performs a more sensitive search in the parents for the indels identified in the children to reduce false positives *de novo* calls in regions of low coverage in parents. I also show how to extract indel calls that fall into target regions and filter out false-positive calls with respect to their sequence composition and variant quality (**Figure 2.2**). Finally, I present one of the available methods for annotating the mutations, especially to prioritize any potential disease-related mutations. Although I use Scalpel for generating the indel calls, the protocol provides general guidelines for standard operations required to analyze and evaluate indel calls. I also illustrate several sources of indel calling errors, which could be introduced by library

construction, sequencing or alignment. Whenever possible, visualization of data/results is performed using IGV alignments, and auxiliary scripts are provided for plotting size, allele fraction distribution, etc.

## Methods

### Experimental Design

In this protocol, I use publicly available WGS data to detect and analyze indels within a family. However, when designing a new study, researchers are typically faced with the problem of choosing suitable sequencing and bioinformatics strategies to answer the relevant scientific questions. There are many factors that play a role in study design, including depth of coverage, read length, parameter tuning, WGS versus WES protocols, the use of PCR amplification, cost per basepair, etc. In this section, my goal is to provide some guidelines on the impact of such different experimental design choices on the sensitivity and accuracy of indel detection.

Although WES is a cost-effective approach to identify genetic mutations within the coding region, it suffers from several major limitations due to a combination of coverage biases, low capture efficiency, and errors introduced by PCR amplification. For example, an indel located near the end of a target region may not be well covered by sequencing reads, which limits detection ability. Also, the exome capture kits are typically designed to pull down a region of about 400bp around an exon, which can limit detection of large indels within coding regions or near splice sites. On the other hand, albeit with higher cost, WGS comes with several significant benefits, including more uniform coverage, freedom from capture efficiency biases, and the inclusion of the non-coding genome. In the context of detecting indels, it has been shown that the accuracy of indel detection with WGS data is much greater than WES data even within the targeted regions<sup>17</sup>. **Table 2.1** shows that a much higher validation rate of WGS-specific indels, compared to WES-specific indels (84% vs. 57%). Specifically, WGS has a unique advantage over WES in identifying many more indels longer than 5 bp (25 vs. 1). When using WGS, it was estimated that 60X depth of coverage from the HiSeq platform would be needed to recover 95% of the indels detected by Scalpel. In particular, detecting heterozygous indels naturally requires deeper sequencing coverage relative to homozygous indels (**Figure 2.3**). WGS at 30X using the HiSeq platform is not sufficient for sensitive indel discovery, resulting in at least 25% false negative rates for heterozygous indels. But these requirements can rapidly change with the longer reads and lower error rates provided by newer instruments.

PCR is a widely used and useful technique to amplify DNA fragments of interest and for attaching various linkers or barcodes for sequencing. However, small amounts of contaminating material can also be amplified without discrimination. Also, PCR amplification introduces errors during the library construction step, especially in regions near STRs such as homopolymer A or T runs. These types of errors are due to replication slippage events and result in high variability in the number of repeat elements (**Figure 2.4**). It becomes then very difficult to distinguish true events at these loci from stutter errors. For indel analysis, I recommend using PCR-free protocols, which can significantly reduce the number of errors around those loci. Moreover, as reported in this protocol, filtering based on the combination of alternative allele coverage and  $k$ -mer  $\chi^2$  score is an effective strategy to filter out additional false-positives without sacrificing much sensitivity.

Large-scale sequencing studies, involving hundreds or thousands of samples, are now becoming more and more widespread. Here I aim to introduce some of the advantages of having access to a collection of sequenced individuals. Even though Scalpel does not directly provide an API for joint calling on more than four samples, we provide examples and a recommendation on how to take advantage of such information if available. The basic idea is to aggregate all the genetic variants detected in the samples into a database framework with associated genotypes and genomic annotation. There are existing flexible systems for exploring genetic variation for disease and population genetics, such as GEMINI <sup>62</sup>. Analyzing the genetic code of a large cohort of individuals has the potential to shed light on the underpinning mechanisms of complex diseases such as autism and schizophrenia. These studies are generally focused on the detection and analysis of rare variant, that can explain the phenotype of the affected individuals.

The population frequency of such rare mutations is usually so low that it is obscured by the noise in the sequencing data, making any real biological signal undetectable. In these circumstances the population can be used to devise effective filtering strategies. For example, in a large-scale autism study where Scalpel was employed <sup>9</sup>, the population database was used to identify rare variants by filtering highly polymorphic loci with many more mutations than expected in the general population as well as common variants using minor allele frequency (MAF) cutoffs. Typically, variants for which the minor allele is present in a population above 1% are considered common. By removing these locations from the analysis, the biological signal started to emerge: an enrichment of frame-shift *de novo* mutations in the affected child compared to the unaffected sibling. The highly polymorphic regions were later found to enrich for homopolymers and other STRs, which are known to be more susceptible to sequencing errors. In the case of *de novo* studies, it is extremely unlikely that the same mutation is present as *de novo* in multiple individuals; in this case, the population information can be used again to filter out these candidates as artifacts in the sequencing.

Detection of somatic variation in tumor-normal matched samples is complicated by different factors such as ploidy, clonality, and purity of the input material. Moreover, the sensitivity and specificity of any somatic mutation calling approach varies along the genome due to differences in sequencing read depths, error rates, variant allele fractions (VAF) of mutations, etc. Accounting for all these variables poses a very complex and challenging problem. However, the proper filtering parameters can eliminate the majority of Scalpel's false-positive calls. For example, **Figure 2.5** show the effects of different phred-scaled Fisher's exact score cutoffs used for filtering on a pair of highly concordant primary and metastatic samples from Branon et al. <sup>63</sup>. **Figure 2.5** demonstrates that indels with a phred-scaled Fisher's exact score below 10 tend to have low VAF and are much more likely to be sequencing errors. In fact, the allele fraction of mutations exclusive to either the primary tumor or the metastasis is significantly lower with higher (more stringent) cutoffs. Similarly, the VAF distribution of the indels found only in the primary tumor shifts towards the expected distribution for these samples (with a peak at ~20%) as more conservative Fisher's exact test cutoffs are used. Not all errors are eliminated though, especially in regions where very low support for a mutation in the normal or the tumor precludes the assembly of the reads.

## The Scalpel pipeline

Scalpel is designed to perform localized micro-assembly of specific regions of interest in a genome with the goal of detecting insertions and deletions with high accuracy. It is based on the de Bruijn graph assembly paradigm where the reads are decomposed into overlapping  $k$ -mers, and directed edges are added between  $k$ -mers that are consecutive within any read<sup>64</sup>. **Figure 2.1** shows the high-level structure of the pipeline. (1) The pipeline begins with a fast alignment of the reads to the reference genome using BWA. Importantly, these alignments are not directly used to call variations, but only to localize the analysis by identifying all the reads that have similarity to a given locus. Reads are then extracted in the region of interest (e.g., exon) including: (i) *well-mapped* reads, (ii) *soft-clipped* reads, and (iii) reads that *fail to map*, but are anchored by their mate. The latter two classes correspond to locations where the mapper encountered trouble aligning the reads, especially because of the large indels present, so it's necessary to include them in the assembly. (2) Once localized, the algorithm computes an on-the-fly assembly of the reads in the current region using the de Bruijn graph paradigm, specifically, reads are decomposed into overlapping  $k$ -mers (starting with a default  $k=25$ ) and the associated graph is constructed. (3) Using the reference sequence, one source node and one sink node are then selected according to the procedure described later in the “Graph traversal” section. (4) An on-the-fly analysis of the repeats in each region is used to automatically select the  $k$ -mer size to be used for the assembly, described in section “Repeat analysis”. (5) The graph is then exhaustively examined to find end-to-end paths that span the region. (6) After the sequences are assembled, they are aligned to the reference to detect candidate mutations using a sensitive gapped sequence aligner based on the Smith Waterman algorithm<sup>65</sup> targeted at the reference window. Finally, the above assembly process is applied using a sliding window approach over each target region. By default, a window size of 400bp is used with a sliding factor of 100bp. The sliding window strategy is fundamental to handle the highly non-uniform read distribution across the target. A window size of 400bp is large enough to assemble the majority of the exons into a single contig since ~95% of the human exon-targets are shorter than 400bp, however each assembly task is small enough for using in-depth techniques to optimize the assembly.

## Graph construction

Two critical components of the Scalpel algorithm are (i) construction of the de Bruijn graph and (ii) detection of sequence paths spanning the targeted region. Reads aligning to the region are extracted and decomposed into overlapping  $k$ -mers. In order to model the double stranded nature of the DNA, a bidirected de Bruijn graph is constructed<sup>66</sup>. The graph is then compressed by merging all non-branching chains of  $k$ -mers into a single node. Tips and low coverage nodes are removed according to input threshold parameters to remove obvious sequencing errors. Note that, differently from traditional de Bruijn graph assemblers, Scalpel does not use any threading strategy to resolve collapsed repeats. Threading allows resolution of repeats whose lengths are between  $k$  and the read length. However, we observed in both real and simulated data that, due to the localized graph construction, if a bubble was not covered end-to-end by the reads, threading would either disconnect the graph or introduce errors. Repeats are instead handled differently, as explained in the next section.

## Repeat Analysis

Due to the highly non-uniform read depth distribution across the targeted region and the presence of near-perfect repeats that can mislead the assembly, we implemented a detailed repeat composition analysis coupled with a self-tuning  $k$ -mer strategy. Specifically, when assembling a window, Scalpel inspects both the base pair composition of the corresponding reference sequence as well as the resulting de Bruijn graph for the presence of cycles in the graph or near-perfect repeats in the assembled sequences. If a repeat structure is detected, the graph is discarded and a larger  $k$ -mer is selected. This process continues until a maximum  $k$ -mer length is reached, which is a function of the read length. If no  $k$ -mer value can be chosen to avoid the presence of repeats, the region is skipped and the next available region from the sliding window scheme is analyzed. This conservative strategy reduces the number of false-positive calls in highly repetitive regions, and, according to our experiments, skips less than 2% of possible windows in the human exome. Note also that, once  $k$  is selected by the self-tuning  $k$ -mer strategy, the graph is “repeat free”, and there is no need to use threading to resolve small repeats. The proposed self-tuning  $k$ -mer strategy is similar to the dynamic approach used by SOAPindel and TIGRA to reconnect a broken path in low coverage regions. However, SOAPindel searches for unused reads with gradually shorter  $k$ -mers until a path is formed or the lower bound on  $k$ -mer length has been reached; while TIGRA allows only 2 possible values for the  $k$  (15 and 25). Scalpel instead starts from a small  $k$ -mer value (input parameter) first and then gradually increases it, such that the smallest possible  $k$ -mer value is used for each region. This strategy has the advantage of better handling of repetitive sequences, highly polymorphic regions, and sequencing errors: source and sink have higher chance to be selected (see section “Graph traversal”) and a smaller  $k$ -mer reduces the chance of fragmented assembly in low coverage regions.

## Graph traversal

Once a valid de Bruijn graph is constructed, Scalpel examines the graph to find end-to-end sequence paths that span the target window. Because the coverage from exome capture data is highly non-uniform, a special selection algorithm is used to find the edges of each window where coverage is present. First, two nodes in the graph are labeled as *source* and *sink* according to the following procedure: the reference sequence of the target region is scanned left-to-right to detect the first sequence of  $k$  bases that exactly matches one of the  $k$ -mers from the nodes in the graph, this node will be marked as the source. In a similar fashion the sink node is detected scanning the reference sequence right-to-left. Since every region is first inspected for repeats, source and sink can be safely selected at this stage. The automated strategy used by Scalpel to select the boundaries of the reference sequence improves over TIGRA’s approach, where the reference region is selected based only on input parameters. After the source and sink nodes are identified, all possible source-to-sink paths are enumerated up to a max number (default 100,000) using a depth-first search (DFS) traversal of the graph, similarly to Sutta assembly algorithm<sup>67</sup>. Note that since the regions to assemble are very small, time and space computational complexities associated with large-scale whole-genome assembly are not relevant and an exact brute-force strategy can be efficiently applied. If there are no repeat structures in the graph, all the candidate paths are enumerated and aligned to the portion of the reference sequence delimited by source and sink  $k$ -mers using the standard Smith-Waterman-Gotoh alignment



algorithm with affine gap penalties. The list of candidate mutations is then generated. Under typical conditions, the assembler reports a single path for homozygous mutations and two paths for heterozygous mutations. For example, if the sample has an insertion in only one of the two haplotypes, the assembler would discover the indels and also the unmodified reference sequence. Note that a traditional sequence assembler would have selected only one of these two paths (usually with higher coverage) and discarded the other one. Scalpel instead examines both paths to distinguish, for example, between homozygous and heterozygous mutations. However, in practice, various factors in real data complicate the detection process and, sometimes, multiple paths are reported in the case of more exotic variations. For example, the Illumina sequencing platform is particularly error prone around microsatellites (e.g., homopolymer runs) and, as a consequence, multiple candidate alleles are elucidated by the data at these loci. Highly polymorphic regions are also prone to generate multiple paths and could be computationally demanding: if the distance between multiple nearby mutations is larger than the (automatically) selected  $k$ -mer value, each of the associated bubbles in the graph will give rise to two different paths.

### Computation protocol

This protocol includes 24 steps encompassing the whole procedure from downloading the input datasets to identification of frame-shift variants. The protocol bundle, available within the Scalpel software package, contains a master script called `run_protocol_0.53.sh` with the complete list of commands (<https://github.com/hanfang/scalpel-protocol>) required to replicate the results presented in this procedure. This script can also be modified to automate the processing of user samples. To align the NGS reads to the genome:

- 1| Convert the \*.2bit genome to \*.fa format and index it with bwa (Note you can also download the fasta file directly, although this may take much longer):
- 2| Align reads to reference for each sample separately with bwa mem:
- 3| Sort the bam files by chromosome coordinates with samtools and then delete the unsorted versions:
- 4| Mark duplicated reads within the alignment with picard tools:
- 5| Perform a basic quality control of the alignment files with samtools:

In order to generate reliable indel calls, accurate alignment of the NGS short reads are of great importance. If the DNA is derived from blood sample, the mapping rate of Illumina HiSeq reads is typically higher than 90%. Lower mapping rates indicate either contaminations of DNA from other species (e.g. bacterial DNA from saliva samples) or poor quality of the sequencing experiments. In addition, excessive numbers of duplicated reads are usually due to issues with library construction and PCR amplification. **Table 2.2** lists the number of reads generated for each sample and the reads mapped to the human genome hg19. To perform indel variant calling and downstream filtering:

- 6| Run Scalpel in the “de novo” mode to perform multi-sample calling for a quad family. In this example, we use NA12882 as the affected individual. The NA12881 is the unaffected individual accordingly:
- 7| Export the inherited and denovo mutations from the Scalpel database (in target only):
- 8| Identify and mark indels within STR regions using the micro-satellite annotation software (msdetector) distributed with the protocol bundle:

- 9| Save indels within and outside STR regions into different variant calling format (vcf) files (Note: the number of fields to keep with UNIX cut command depends on the number of samples in the vcf file):

In both “denovo” and “somatic” mode, Scalpel is optimized to achieve high sensitivity, but may include some false positives. To control for this, I recommend using the --two-pass option in Scalpel, which undergoes a second round of indel verification to reduce the likely false calls. Low-quality indel calls (potential false-positives) are usually found within low coverage regions, or have an unbalanced number of reads supporting the alternative allele.

- 10| Filter out false positive calls by adjusting coverage and/or chi-square score thresholds for your data:
- 11| (Optional) Perform additional filtering of the de novo calls using the python script provided in the Scalpel resource bundle. This script supports filtering indels by alternative allele coverage, chi-square scores (chi), and parental coverage (pc):
- 12| (Optional) Extract a subset of indels based on other annotations using bedtools. Here we show how to extract the variants that overlap any of the mutations in the ClinVar main database.
- 13| Summarize indel calls with a histogram of mutations by size:
- 14| Characterize low quality homopolymer indels calls with a histogram of mutations by VAF:
- 15| Summarize inherited indels with variant allele fractions (VAF %):
- 16| Determine the number of indels remained after each step of the filtering:
- 17| Split the multi-sample VCF to an individual file for NA12882:
- 18| Filter the single VCF files based on Chi-Square score and allele coverage

There are usually much higher sequencing biases in GC-extreme regions. Indels within STRs, especially homopolymer A or T runs, are major source of false positive variant calls. The filtering cascade should not reduce the sensitivity of inherited indels by a lot. One should expect a relatively balanced number of reads support each inherited indel, indicating high confidence for these calls. To annotate and visualize of the indel calls:

- 19| Prepare and create the input format required by Annovar:
- 20| Annotate and intersect indels with gene regions using Annovar:
- 21| Summarize coding region indels by size in R:
- 22| Filter the indels based on population allele frequencies:
- 23| Annotate novel indels that were not reported by a population database before (1000G, ESP6500, ExAC, CG46):
- 24| Retrieve frame-shift mutations, which are potentially loss-of-function

## Results

### High accuracy of Scalpel

One of the most sensitive and accurate approaches for indel detection from short read data is a micro-assembly algorithm, Scalpel. It was previously demonstrated to have substantially improved accuracy over eight algorithms including GATK-HaplotypeCaller<sup>53</sup> (v3.0) and SOAP-indel<sup>54</sup> (v2.01), while other methods report a large number of false negative calls<sup>51</sup>. In fact, Scalpel achieves very high accuracy (positive predictive value=90%) of indel detection even on 30X WGS data (**Figure 2.6**). In this thesis, I describe the use of Scalpel for indel detection from



whole genome and exome capture sequencing experiments. I introduce three different modes of indel detection: *de novo*, somatic, and single-sample for different study designs. First, the *de novo* mode is useful for calling germline *de novo* variants in nuclear families up to four people. Second, the somatic mode is useful for identifying somatic changes within matched samples, especially tumor/normal pairs in cancer studies. Finally, the single mode is useful for studies of a single proband.

### Expected distribution of indels and signatures of low-quality calls

After filtering for alternative allele coverage, Chi-square scores, and STR regions, the size of the high-quality inherited indels should follow a log-normal distribution (**Figure 2.7**). Similar observations of such a size distribution were also reported in the 1000 Genomes Project<sup>68</sup> and an analysis of 179 human genomes<sup>69</sup>. I also observed a much higher abundance of homopolymer A or T indels, relative to homopolymer C or G indels in the low-quality call set (**Figure 2.8**). Homopolymer A or T indels usually have low variant allele fraction (VAF), because homopolymer A or T molecules are enriched for PCR stutter/slippage artifacts. Conversely, the VAF of high quality inherited indels approximately follow a normal distribution with a mean of around or slightly less than 50% (**Figure 2.9**). This indicates that I observed equal read evidence of both alleles in the genome.

### Expected number of indels during the filtering cascade

Since calling *de novo* indels requires a more sensitive analysis of the family members, I recommend using the `--two-pass` search option when discovering *de novo* events. Many more inherited indels will persist through the filtering cascade, relative to the number of *de novo* events. This is because *de novo* events are extremely rare in comparison to inherited indels. *De novo* mutations are also particularly vulnerable to batch effects and random errors, as a correct analysis requires both high sensitivity and specificity in the entire family. In fact, among the in target indels, about 51% of the inherited ones are of high quality while only 5% of the *de novo* ones survived the filtering cascade (**Figure 2.10**). Because frameshift mutations can cause loss of function of a gene, these mutations are expected to be less frequent than frame-preserving mutations in the coding region. As shown in **Figure 2.11**, indels whose size is a multiple of three are much more abundant than others with similar sizes (+1 or -1).

### A list of frame-shift mutations in the family

Although this family has been investigated in many studies, many frameshift indels were not discovered in any public databases, including 1000G, ExAC, and ESP. We observe a total of 6 novel frameshift mutations. Many of these indels are of a size larger than five base pair. Based on Sanger validation of these loci, all 20 genotypes in four family members were successfully validated/confirmed. With the improvement of indel calling protocol introduced in this manuscript, we are able to identify these previously undiscovered loss-of-function mutations. We also inspected the VCF file generated by the Illumina Platinum Genome project (release 8.0.1) for the presence of the six discovered frameshift. Although the VCF file was generated using five different variant callers (Freebayes, Platypus, GATKv3, Cortex, and Issac2), it only contained two out of the six indels. This indeed further demonstrates the power of Scalpel over other methods, especially on detecting large indels.

High-quality *de novo* indels usually share the following characteristics: 1) the number of reads in the region is close to the genome-wide mean coverage, 2) there are balanced number of

reads supporting both the reference and alternative allele, 3) these indels are not located within or near short tandem repeat regions, 4) in the parents' genome, there are no reads supporting the same indel presented in the child's genome. For example, I found a one-bp heterozygous frameshift deletion located in the exon 4 of the gene *HFMI*. This *HFMI de novo* deletion was also successfully validated in Sanger experiments. The genomic coordinate is chr1: 91859889, relative to the reference genome hg19. This variant has not been reported before in any of the widely used variant databases, such as 1000G, ESP6500, ExAC and CG46. **Figure 2.12** shows the screenshot of the IGV alignment of all four genomes. We can see a distinct signature of the deletion only presented in the affected child, but not in anyone else in the family.

### Limitations of the protocol and software

Scalpel provides several advantages to standard mapping approaches but, like any bioinformatics algorithm, it does not attempt to address all possible types or sizes of mutations at once. In these experiments, Scalpel was able to reliably detect deletions up to 400 bp (including deletions of *Alu* mobile elements) and insertions shorter than 200 bp, but the sensitivity is reduced for longer indels given the available read lengths (data not shown). Even within this size range, Scalpel, and all pipelines, has lower sensitivity for indels in low coverage regions that are supported by very few reads. In the worst scenario, a combination of low coverage within a complex repeat region may require a *k*-mer size too large for assembling across the mutation, leading to false negatives. Phasing of the discovered mutations is not supported and, given the locality of the assembly, it would be possible to phase only mutations within the same window (400bp by default). Thanks to the new advances in long-molecule sequencing technologies (e.g., PacBio, 10X Genomics), in the near future it will be possible to combine such technologies for phasing mutations hundreds of kilobases to megabases apart.

For variant calling purposes, it is ideal to have a high-quality reference genome available. This is also true for indel calling with Scalpel because assembly errors might falsely increase the number of variants and the read localization will not be effective unless a complete representation of the genome is available. Users working with data from a genome without a reference should first generate a high-quality assembly using one of the several whole-genome assemblers<sup>70, 71</sup>. This procedure can be easily adapted to work with a draft assembly, but no testing has been performed and the results could be unpredictable. Tumor/normal and multiple family members can be analyzed together, but joint calling across a large number of samples is not supported by Scalpel, although population frequencies can be used to identify systematic sequencing errors. This protocol also assumes that sequencing was performed using the Illumina sequencing platform, including MiSeq, HiSeq 2000, and HiSeq X sequencers. Other sequencing technologies (e.g. Ion Torrent, Sanger, SOLiD) can be also used for studies like the one reported here, but the software pipeline used in this thesis does not support them. Finally, no graphical user interface is available for the steps performed in this protocol; all the operations are performed through the UNIX shell. Some of the tools used here, such as BWA and Picard tools, are now available through cloud-based web interface systems such as Galaxy (<https://usegalaxy.org/>). We look forward to seeing Scalpel integrated into such systems in the near future.

Tables and figure in this chapter

Tables

**Table 2.1. Comparisons and validation rates of indel detection with WGS and WES**

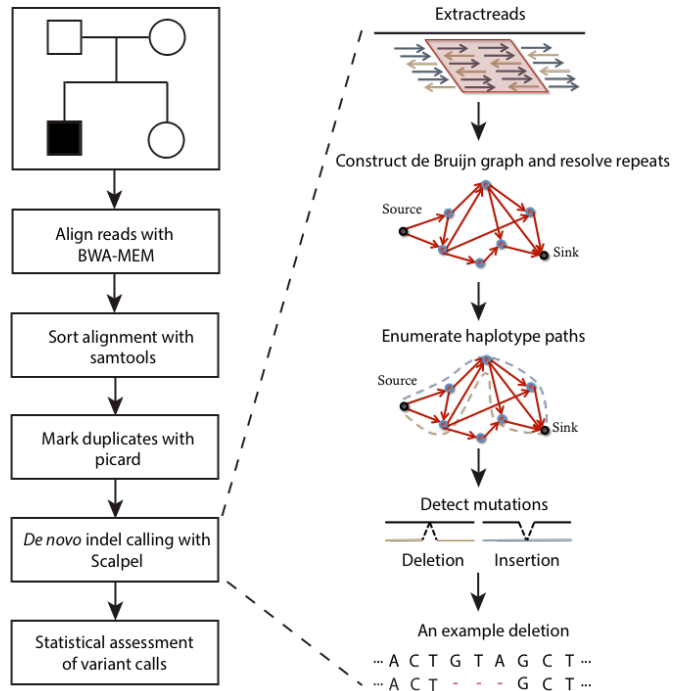
	Indels	Valid	PPV	Indels (>5 bp)	Valid (>5 bp)	PPV (>5 bp)
WGS-WES intersection	160	152	95.0%	18	18	100%
WGS-specific	145	122	84.1%	33	25	75.8%
WES-specific	161	91	56.5%	1	1	100%

Note: The validation rate, positive predictive value (PPV), is computed by the following:  $PPV = \#TP / (\#TP + \#FP)$ , where #TP is the number of true-positive calls and #FP is the number of false-positive calls. Both WGS (mean coverage = ~70X) and WES (mean coverage = ~330X) were done on Illumina HiSeq 2000 sequencers under 2x100bp mode (described in Fang et al.)<sup>17</sup>. The construction of WGS libraries here involved a procedure of PCR amplification. The exome capture kit used for WES was NimbleGen SeqCap EZ Exome v2.0, which was designed to pull down 36 Mb of the human genome.

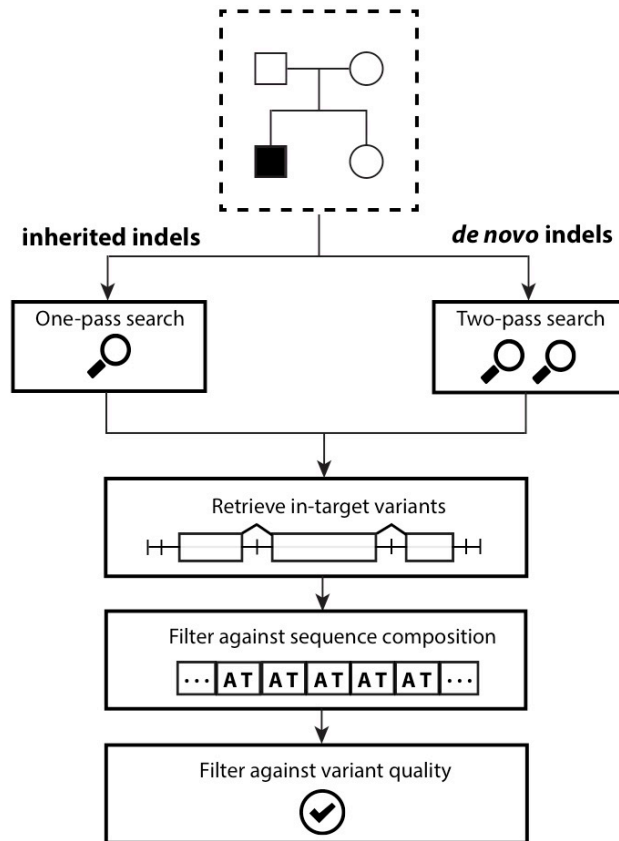
**Table 2.2. Expected QC-passed read and mapping statistics**

Sample	QC-passed	Duplicated	Duplicated rate	Mapped reads	Mapping rate
NA12877	1,629,579,046	39,439,277	<b>2.42%</b>	1,618,796,107	<b>99.34%</b>
NA12878	1,578,485,183	36,266,744	<b>2.30%</b>	1,568,334,656	<b>99.36%</b>
NA12881	1,559,137,724	39,169,529	<b>2.51%</b>	1,547,550,351	<b>99.26%</b>
NA12882	1,617,281,311	38,592,443	<b>2.39%</b>	1,607,709,559	<b>99.41%</b>
Average	1,596,120,816	38,366,998	<b>2.40%</b>	1,585,597,668	<b>99.34%</b>

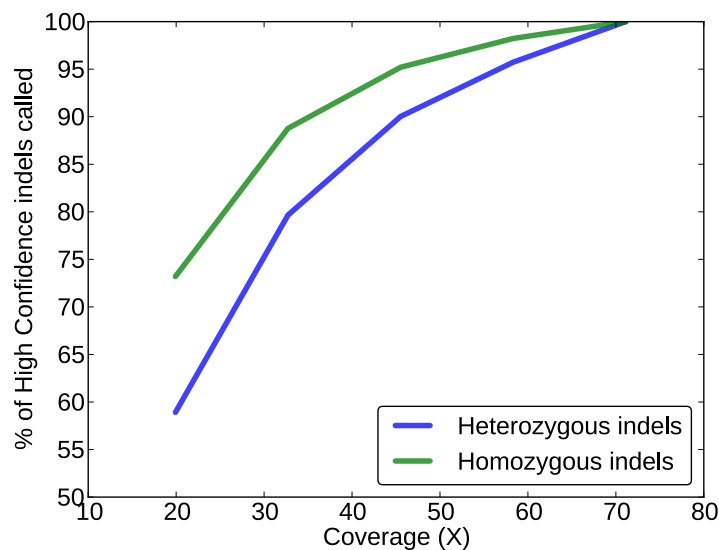
Figures



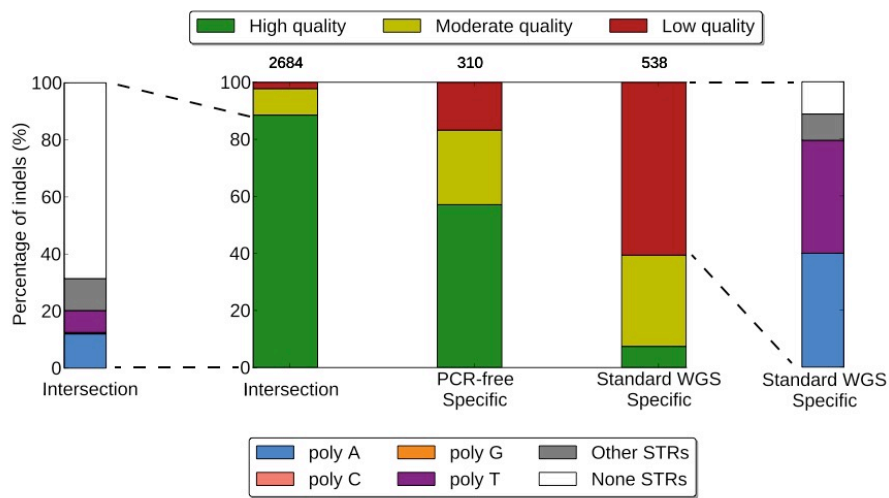
**Figure 2.1. Main steps in the Scalpel protocol.** Starting from raw sequencing data, reads are first aligned to the human genome using the BWA<sup>72</sup> software package (step 4 in Procedure). Following the standard practices in the field, the alignments are sorted (using Samtools<sup>15</sup>, step 5 in Procedure) and duplicates are marked (using Picard tools - <http://broadinstitute.github.io/picard/>, step 6 in Procedure). Finally, indels can be called with Scalpel (steps 8-9 in Procedure) and statistical assessment of the variant calls can provide diagnostics of the data (step 10-20 in Procedure). Note that, since Scalpel locally re-assembles the reads, this procedure is free of computationally expensive techniques such as indel realignment and base quality recalibrations. The BAM files obtained after the earlier steps are the input for Scalpel micro-assembly procedure. Scalpel then localizes the reads with a window, constructs a de Bruijn graph, resolves repeat structure, and enumerate haplotype paths. Figure is adapted from Narzisi et al<sup>51</sup>.



**Figure 2.2. Overview of the indel variant filtering cascade.** This figure is a conceptual representation of the filtering cascade in the materials section. It is used to report high quality *de novo* and inherited indels within the target region; coding regions in this case. (1) Inherited and *de novo* indels are analyzed separately; (2) only variants within the target regions are exported; (3) Low quality indels are identified and removed based on sequence composition (e.g., STRs); (4) Additional filters based on supporting coverage and allele balance are used to reduce the number of false-positives.



**Figure 2.3. Higher coverage can improve Scalpel’s sensitivity performance of indel detection with WGS data.** The sensitivity performance is assessed using the high confidence call set shared by WGS and WES data (both Illumina HiSeq2000 platform) from eight samples using all available coverage (70x mean coverage). We down-sampled the reads to a fraction of the original number and performed indel calling again. Compared to the original set at 70X mean coverage, we report the percentage of variants that could still be called at a reduced coverage. The Y-axis represents the percentage of the high confidence indels revealed at a down-sampled dataset. The X-axis represents the mean coverage of the eight down-sampled genomes. Among the entire call set, about 61% of the indels are heterozygous and the remaining 39% are homozygous. Performance of heterozygous (blue) and homozygous (green) indel detection are shown separately. Reduced coverage indeed affected the detection of heterozygous indels more than homozygous ones.



**Figure 2.4. Comparison of standard WGS and PCR-free data based on indel quality.** Indel quality was defined with respect to alternative allele coverage and chi-square score, which is described in details in the materials section and Fang et al.<sup>17</sup>. “Intersection” represents the shared indels from both the PCR-free and standard WGS indels. The number reported above a call set represent the total number of indels in that subset; the two data shared 2684 variants, while 310 and 538 were specific to standard WGS and PCR-free, respectively. Indel calls are further categorized (side-bars) based on their sequence composition: poly\*-A, poly-C, poly-G, poly-T, other-STR, and non-STR. To be noted, although poly-C and poly-G indels existed in the call-set, their fractions are too minimal to be visualized in the plot. In fact, poly-A, poly-T and non-homopolymer STRs dominate the STR indels. \*Homopolymer – poly.

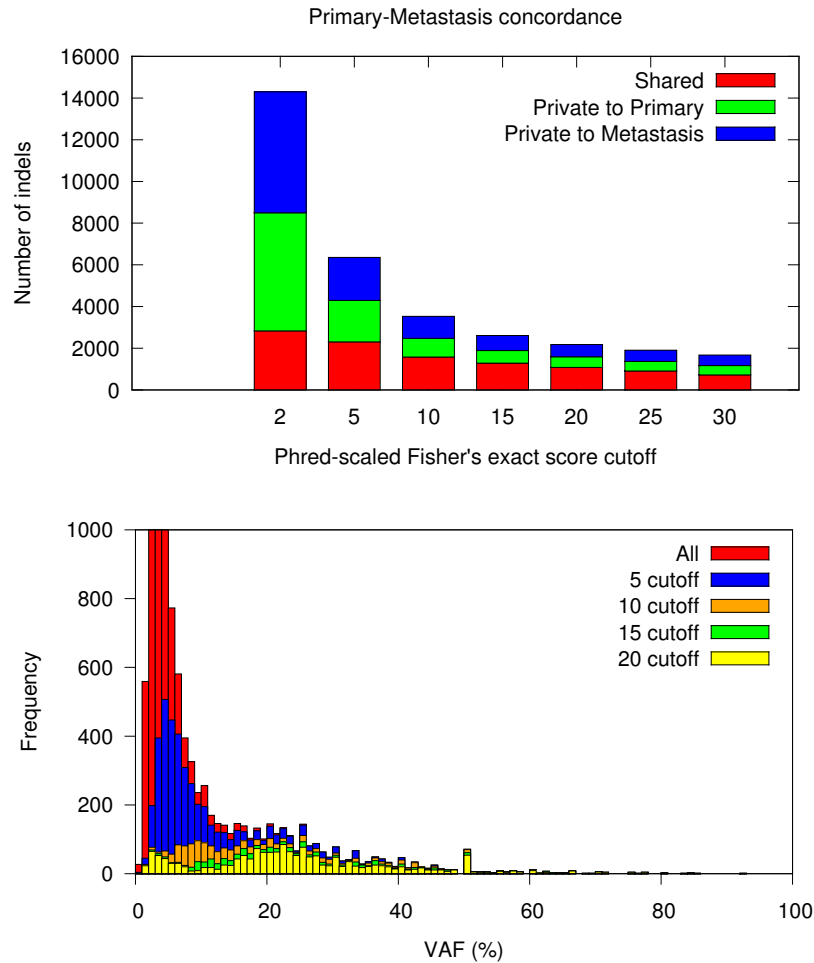


Figure 2.5. Whole genome mutational concordance. (a) **Concordance and discordant indel mutations as a function of the phred-scaled Fisher's exact score cutoff between primary and metastasis for a pair of highly concordant colorectal cancer samples from Branon et al.<sup>63</sup>** Increasing the Fisher's exact score cutoff substantially reduces the number of private indels, while maintaining a similar amount of shared ones. This demonstrates the Fisher's exact score ability to discriminate true mutations from the false positive ones. (b) **Distribution of variant allele fraction (VAF) as a function of different phred-scaled Fisher's exact score cutoffs for the somatic indels detected in the primary tumor.** Increasing the cutoff shifts the distribution to the expected 20% VAF for these samples.

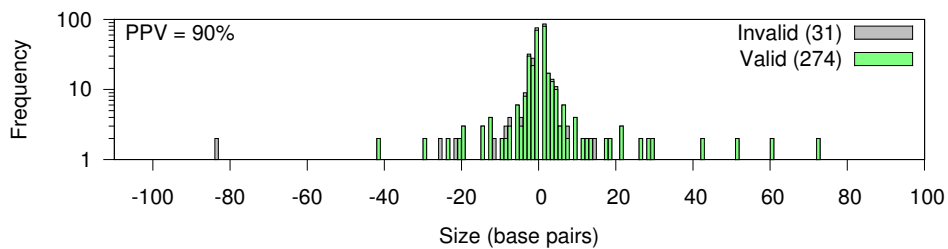
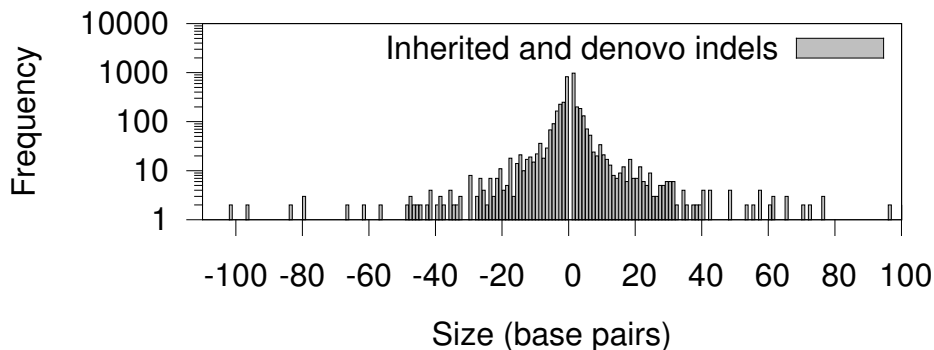
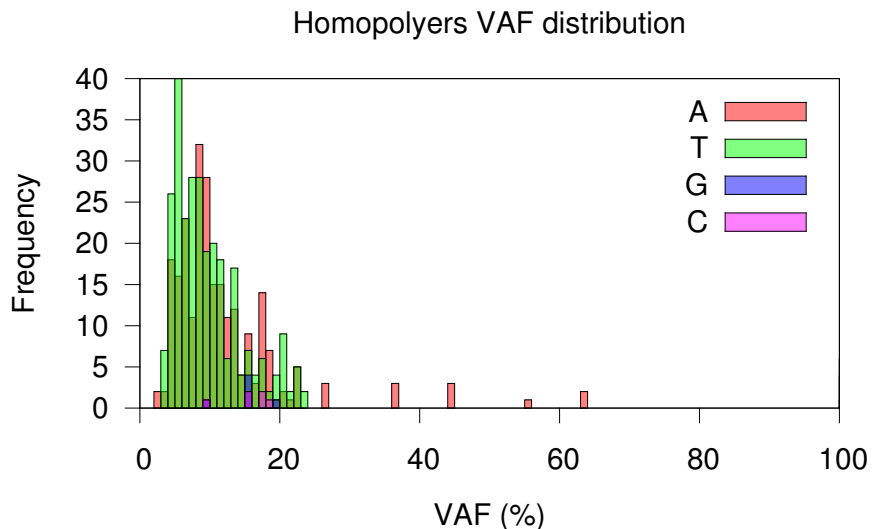


Figure 2.6. **High accuracy of indel detection using Scalpel on WGS data.** Scalpel was run in the single mode on 30X Illumina Hiseq 2000 2x100bp WGS data described in Narzisi et al.<sup>51</sup> and later analyzed in Fang et al.<sup>17</sup>. This figure shows the size distribution of valid (green) and

invalid (gray) indels that are randomly selected for validation (using targeted resequencing) in the two previous studies. This validation set includes 160 and 145 candidate variants that were WGS-WES intersected and WGS-specific, respectively. Among a total of 305 candidates, 90% of them (274) were successfully validated. Positive-predictive value (PPV) is computed by  $PPV = \frac{\#TP}{\#TP + \#FP}$ , where #TP is the number of true-positive calls and #FP is the number of false-positive calls.

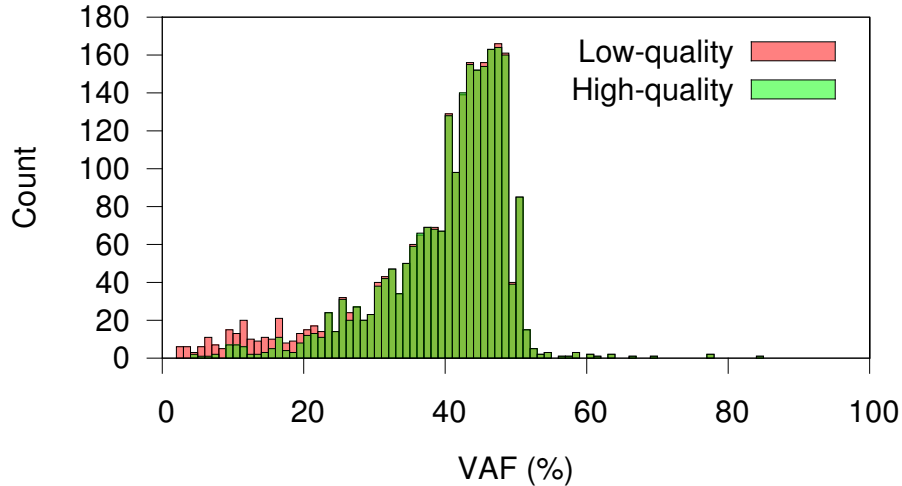


**Figure 2.7. Size distribution of inherited and denovo indels.** The Y-axis represents the number of indels, while the X-axis represent the size of indels in base pair. We should expect a log-normal distribution of indels with majority of them being short, i.e. less than 5bp in the human exonic regions<sup>51</sup>. This figure was generated using the data from step 15.

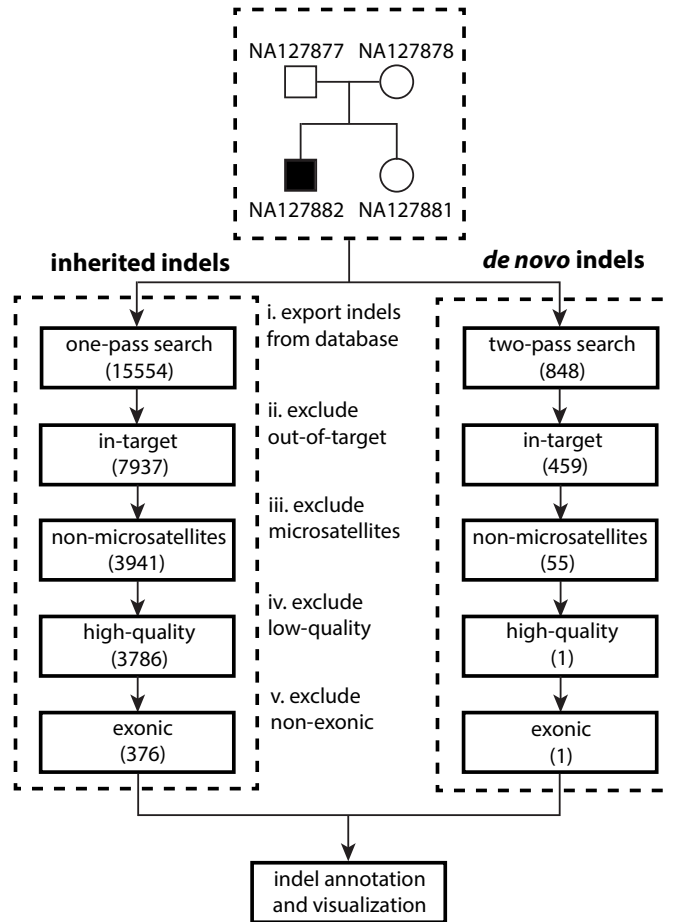


**Figure 2.8 Histograms of low quality homopolymer indels by category.** The Y-axis represent the number of indels, while the X-axis represent the variant allele fraction (VAF). Homopolymer A or T indels should be more abundant than C or G indels in the call set, especially indels with very low VAF. Due to the limitation of PCR amplification, homopolymer A or T runs are more like result in inaccurate molecules<sup>17</sup>. This figure was generated using the data from step 16.

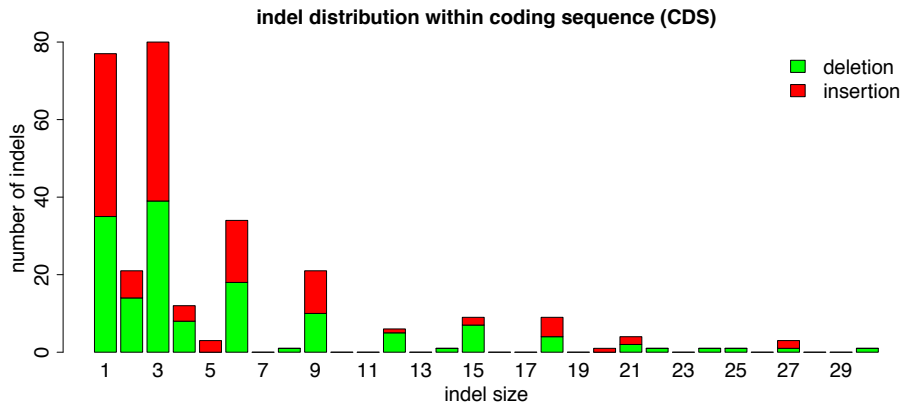




**Figure 2.9 Variant allele fractions (VAF %) of the inherited indels.** Low/High quality indels here were defined with respect to the coverage and Chi-square scores described in step 11 and 12. VAF of high quality inherited indels should approximately follow a normal distribution with a mean of about 50%. In practice, due to sequencing and alignment biases, the mean of the normal distribution is usually slightly smaller than 50%. Low-quality indels usually have low VAF, especially towards lower than 20%. This figure was generated using the data from step 17.



**Figure 2.10. Filtering cascade of inherited and *de novo* indel calls.** The numbers in each box denote the expected numbers of indel calls remained after filtering. The *de novo* indels were undergone a two-pass search to reduce the number of false positives. The numbers in this figure was obtained from step 9, 10, 11, 12, and 22. It is important to use a two-pass search in *de novo* indel calling. Because false-positive calls can be reduced with a more sensitive parameter setting used on the parents' data.



**Figure 2.11 Frame-preserving indels are more abundant within coding sequences (CDS).** This figure was generated using the data from step 23, which was the set of inherited indels from the proband, NA12882. Y-axis represents the number of indels, while the X-axis shows the indel size. Stacked bar plots of insertions (red) and deletions (green) are shown in this figure. Indels with size that is a multiple of three (frame-preserving) are more abundant than the frame-disrupting ones.



**Figure 2.12. Screenshot of the alignment of the *de novo* deletion in IGV browser.** From the top to the bottom are: NA12877 (father), NA12878 (mother), NA12881 (sibling), NA12882 (proband). The black lines in the alignment of NA12882 show the T deletion in the genome. It is clear that this deletion is only present in the proband but not in any other family members.

## Chapter 3 Characterizing the sources of indel errors

### Summary of Contribution

This chapter describes the characterization of the sources of indel calling errors. The analysis and results were published in *Genome Medicine*<sup>17</sup>. Han Fang designed the experiments, performed the simulation, and analyzed the results. Yiyang Wu generated the MiSeq validation data in this chapter. Permission for republication of this material has been granted and is available upon request.

### Abstract

Indels, especially those disrupting protein-coding regions of the genome, have been strongly associated with human diseases. However, there are still many errors with indel variant calling, driven by library preparation, sequencing biases, and algorithm artifacts. We characterized whole genome sequencing (WGS), whole exome sequencing (WES), and PCR-free sequencing data from the same samples to investigate the sources of indel errors. We also developed a classification scheme based on the coverage and composition to rank high and low quality indel calls. We performed a large-scale validation experiment on 600 loci, and find high-quality indels to have a substantially lower error rate than low quality indels (7% vs. 51%). Simulation and experimental data show that assembly based callers are significantly more sensitive and robust for detecting large indel (>5bp) than alignment based callers, consistent with published data. The concordance of indel detection between WGS and WES is low (52%), and WGS data uniquely identifies 10.8-fold more high-quality indels. The validation rate for WGS-specific indels is also much higher than that for WES-specific indels (85% vs. 54%), and WES misses many large indels. In addition, the concordance for indel detection between standard WGS and PCR-free sequencing is 71%, and standard WGS data uniquely identifies 6.3-fold more low-quality indels. Furthermore, accurate detection with Scalpel of heterozygous indels requires 1.2-fold higher coverage than that for homozygous indels. Lastly, homopolymer A/T indels are a major source of low-quality indel calls, and they are highly enriched in the WES data. Overall, we show that accuracy of indel detection with WGS is much greater than WES even in the targeted region. We calculated that 60X WGS depth of coverage from the HiSeq platform is needed to recover 95% of indels detected by Scalpel. While this is higher than current sequencing practice, the deeper coverage may save total project costs because of the greater accuracy and sensitivity. Finally, we investigate sources of indel errors (e.g. capture deficiency, PCR amplification, homopolymers) with various data that will serve as a guideline to effectively reduce indel errors in genome sequencing.

### Introduction

With the increasing use of next-generation sequencing (NGS), there is growing interest from researchers, physicians, patients and consumers to better understand the underlying genetic contributions to various conditions. For rare diseases and cancer studies, there has been increasing success with exome/genome sequencing in identifying mutations that have a large effect size for particular phenotypes<sup>73-75</sup>. Some groups have been trying to implement genomic and/or electronic health record approaches to interpret disease status and inform preventive

medicine<sup>76-80</sup>. However, we are still facing practical challenges for both analytic validity and clinical utility of genomic medicine<sup>81-85</sup>. In addition, the genetic architecture behind most human disease remains unresolved<sup>86-91</sup>. Some have argued that we should bring higher standards to human-genetics research in order to return results and/or reduce false-positive reports of “causality” without rigorous standards<sup>92,93</sup>. Others have reported that analytic validity for WES and WGS is still a major issue, pointing out that the accuracy and reliability of sequencing and bioinformatics analysis can and should be improved for a clinical setting<sup>82, 83, 94-97</sup>.

At the time of this work in 2014, there was debate whether we should primarily use whole genome sequencing (WGS) or whole exome sequencing (WES) for personal genomes. Some suggested that a first-tier cost-effective WES might be a powerful way to dissect the genetic basis of diseases and to facilitate the accurate diagnosis of individuals with Mendelian disorders<sup>98,99</sup>. Others showed that targeted sequencing misses many things<sup>100</sup> and that WGS could reveal structural variants (SVs), maintain a more uniform coverage, is free of exome capture efficiency issues, and actually includes the noncoding genome, which likely has substantial importance<sup>101-104</sup>. Some groups directly compared WGS with WES, but thorough investigation of indel errors was not the focus of these comparisons<sup>82, 95, 96, 105</sup>. Substantial genetic variation involving indels in the human genome had been previously reported but accurate indel calling was still difficult<sup>106-108</sup>. There has been a dramatic decrease of sequencing cost over the past few years, and this cost has decreased further with the release of the Illumina HiSeq X Ten sequencers which have capacity for nearly 18,000 whole human genomes per instrument per year. However, it is still unclear whether we can achieve a high-accuracy personal genome with a mean coverage of 30X from the Illumina HiSeq X Ten sequencers. In addition, there have been questions on the use of PCR amplification in the library preparations for NGS, although very few had characterized the PCR errors that might be complicating the detection of insertions and deletions.

Concordance rates among indels detected by the GATK Unified Genotyper (v1.5), SOAPindel (v1.0) and SAMtools (v0.1.18) are reportedly low, with only 26.8% agreeing across all three pipelines<sup>82</sup>. Another group also reported low concordance rates for indels between different sequencing platforms, further showing the difficulties of accurate indel calling<sup>96</sup>. Other efforts have been made to understand the sources of variant calling errors<sup>84</sup>. Common indel issues, such as realignment errors, errors near perfect repeat regions, and an incomplete reference genome have caused problems for approaches working directly from the alignments of the reads to reference<sup>109, 110</sup>. De novo assembly using de Bruijn graphs has been reported to tackle some of these limitations<sup>111</sup>. Fortunately, with the optimization of micro-assembly, these errors have been reduced with a novel algorithm, Scalpel, with substantially improved accuracy over GATK-HaplotypeCaller (v3.0), SOAP-indel (v2.01), and six other algorithms<sup>112</sup>. Based on validation data, the positive prediction rate (PPV) of algorithm specific indels was high for Scalpel (77%), but much lower for GATK HaplotypeCaller (v3.0) (45%) and SOAP-indel (v2.01) (50%)<sup>112</sup>.

Thus, we set out to investigate the complexities of indel detection on Illumina reads using this most accurate indel-calling algorithm. Firstly, we used simulation data to understand the limits of how coverage affects indel calling with Illumina-like reads using GATK-UnifiedGenotyper and Scalpel. Secondly, we analyzed a dataset including high coverage WGS and WES data from two quad families (mother, father and two children), in addition to extensive

high-depth validation data on an in-house sample, K8101-49685s. In order to further understand the effects of PCR amplification on indel calling, we also downloaded and analyzed two WGS datasets prepared with and without PCR from the well-known HapMap sample NA12878. We characterized the data in terms of read depth, coverage uniformity, base-pair composition pattern, GC contents and other sequencing features, in order to partition and quantify the indel errors. We were able to simultaneously identify both the false-positives and false-negatives of indel calling, which will be useful for population-scale experiments. We observed that homopolymer A/T indel are a major source of low quality indel and multiple signatures. As more and more groups start to use these new micro-assembly based algorithms, practical considerations for experimental design should be introduced to the community. Lastly, we explicitly addressed the question concerning the necessary depth of coverage for accurate indel calling using Scalpel for WGS on HiSeq sequencing platforms. This work provided important insights and guidelines to achieve a highly accurate indel call set and to improve the sequencing quality of personal genomes.

## Methods

### Analysis of Simulated Data, WGS and WES data

We simulated Illumina-like 2\*101 paired-end reads with randomly distributed indel, which ranged from 1 bp to 100 bp. The simulated reads were mapped to human reference genome hg19 using BWA-mem (v0.7-6a) using default parameters<sup>113</sup>. The alignment was sorted with SAMtools (v0.1.19-44428cd)<sup>15</sup> and the duplicates were marked with Picard using default parameters (v1.106), resulting in a mean coverage of 93X. We down-sampled the reads with Picard to generate 19 sub-alignments. The minimum mean coverage of the sub-alignments was 4.7X and increased by 4.7X each time, before it reached the original coverage (93X). Scalpel (v0.1.1) was used as a representative of assembly based callers to assemble the reads and call indel from each alignment separately, resulting in 20 indel call-sets from these 20 alignments, using the following parameter setting: “--single --lowcov 1 --mincov 3 --outratio 0.1 --numprocs 10 --intarget”. We also used GATK-UnifiedGenotyper (v3.2-2) as a representative of alignment based callers to call indel from each set of alignments<sup>114</sup>. We followed the best practices on the GATK website, including all the pre-processing procedures, such as indel realignment and base recalibration. Scalpel (v0.1.1) internally left-normalized all the indel so we only used GATK-LeftAlignAndTrimVariants on the indel calls from UnifiedGenotyper. We then computed both the sensitivity and false discovery rate (FDR) for both indel callers, with respects to all and large (>5 bp) indel. The same versions and the same sets of parameter settings for BWA-mem, Picard, and Scalpel, were also used in the rest of the study, including the analysis of WGS/WES data, standard WGS and PCR-free data.

Blood samples were collected from eight humans of two quartets from the Simons Simplex Collection (SSC). Both WGS and WES were performed on the same genomic DNA isolated from these eight blood samples. The exome capture kit used was NimbleGen SeqCap EZ Exome v2.0, which was designed to pull down 36Mb (approximately 300,000 exons) of the human genome hg19. The actual probe regions were much wider than these targeted regions, because probes also covered some flanking regions of genes, yielding a total size of 44.1Mb. All of the libraries were constructed with PCR amplification. We sequenced both sets of libraries on

Illumina HiSeq2000 with average read length of 100 bp at the sequencing center of Cold Spring Harbor Laboratory (CSHL). We also generated WGS (mean coverage=30X) and WES (mean coverage=110X) data from an in-house sample K8101-49685s (not from SSC), which was extensively investigated in the later validation experiment. Exome capture for this sample was performed using the Agilent 44Mb SureSelect protocol and the resulting library was sequenced on Illumina HiSeq2000 with average read length of 100 bp. All of the HiSeq data from K8101-49685s have been submitted to the Sequence Read Archive (SRA) (<http://www.ncbi.nlm.nih.gov/sra/>) under project accession number SRX265476 (WES data) and SRX701020 (WGS data). All of the HiSeq data from eight SSC samples have been submitted to the National Database for Autism Research (NDAR) (<http://ndar.nih.gov/>) under collection “Wigler SSC autism exome families” (project number: 1936).

We excluded all of the low quality raw reads, aligned the remaining high quality ones with BWA-mem, and mark-duplicated with Picard. We used Scalpel to assemble the reads and identify indels under both single mode and quad mode. The single mode outputs all of the putative indels per person, and the quad mode outputs only the putative de novo indels in the children in a family. We expanded each of the exons by 20 bp upstream and 20 bp downstream in order to cover the splicing sites and we called this set of expanded regions the “exonic targeted regions”. The exonic targeted regions are fully covered by the exome capture probe regions. We excluded indels that were outside the exonic targeted regions in the downstream analysis. We left-normalized the indels and compared the two call sets for the same person using two criteria: exact-match and position-match. Position-match means two indels have the same genomic coordinate, while exact-match additionally requires that two indels also have the same base-pair change(s). We called the indels in the intersection based on exact-match as WGS-WES intersection indels. Further, we named the indels only called from one dataset as “WGS-specific” and “WES-specific” indels, respectively. Regions of the above three categories of indels were partitioned and investigated separately. In particular, we focused on regions containing short tandem repeats (STR) and homopolymers. We used BedTools (v2.18.1) with the region file from lobSTR (v2.04) to identify homopolymeric regions and other STR (dual repeats, triplets and etc.) in the human genome <sup>115-117</sup>.

We used Qualimap (v0.8.1) to generate summary statistics of the alignment files of interest <sup>118</sup>. For a certain region, we define the proportion of a region covered with at least X reads to be the coverage fraction at X reads. In addition to the coverage histograms, we also computed the coefficient of variation ( $C_v$ ) to better understand the coverage uniformity of the sequencing reads. An unbiased estimator of  $C_v$  can be computed by  $\widehat{C}_v^* = \left(1 + \frac{1}{4n}\right) * \left(\frac{s}{\bar{x}}\right)$ , where  $s$  represents the sample standard deviation and  $\bar{x}$  represents the sample mean. In our case,  $\widehat{C}_v^*$  asymptotically approaches to  $\left(\frac{s}{\bar{x}}\right)$  as the sample size ( $n$ ) of the data is usually greater than 10,000. The reference genome used here is hg19. There were four region files that we used for this part of the analysis. The first one is the exon region bed file from NimbleGen. We generated the other three region files by expanding 25 bp upstream and downstream around loci of WGS-WES intersection indels, WGS-specific indels, and WES-specific indels, respectively. We followed all of the default settings in Qualimap except for requiring the homopolymer size to be at least five (-hm 5). Finally, we used Matplotlib to generate the figures with the raw data from Qualimap under the Python environment 2.7.2 <sup>119</sup>.

We downloaded PCR-free WGS data of NA12878 (access Code: ERR194147), which is publicly available in the Illumina Platinum Genomes project. We also download another WGS dataset of NA12878 with PCR amplification during library preparation, and we called it standard WGS data (SRA access Code: SRR533281, SRR533965, SRR539965, SRR539956, SRR539947, SRR539374, SRR539357). Both data were generated on the Illumina HiSeq 2000 platform. Although the PCR-free data was not supposed to have any PCR duplicates, we observed a duplication rate of 2% as reported by Picard, and we excluded these reads, yielding 50X mean coverage for both datasets after removing PCR duplicates. We used the same methods for alignment, indels calling, and downstream analysis as described above. indels outside the exonic targeted regions were not considered in the downstream analysis.

We were interested to know how depth of coverage affects the sensitivity of indel detection in WGS data. To accurately measure this sensitivity, one needs a robust call set as a truth set. Fortunately, we had exact-match indels concordant between high coverage WGS and high coverage WES data. We therefore measured sensitivity based on these WGS-WES intersection indels, rather than on the whole set of indels, which might contain more false positives. We down-sampled each WGS dataset to mean coverages of 20X, 32X, 45X and 57X. We then used Scalpel to call indels from the resulting 4 sub-alignment files for each sample and computed the sensitivity at a certain mean coverage (X) for each sample by the equation:

$$\text{Sensitivity at X coverage} = \frac{\# \text{ intersection indels at X coverage}}{\# \text{ intersection indels at the full coverage}} \quad \text{Equation 3.1}$$

This metric measures how many of the WGS-WES intersection indels can be discovered as a function of read depth. We also analyzed the WGS-WES intersection indel call set in terms of zygosity: WGS-WES intersection heterozygous and homozygous indel, subsequently measuring the sensitivity with respect to different zygosityes.

### Generation of MiSeq validation data

We randomly selected 200 indels for validation on an in-house sample K8101-49685s from each of the following categories: 1) indels called from both WGS and WES data (WGS-WES intersection), 2) WGS-specific indels, 3) WES-specific indels. Out of these 600 indels, 97 were covered with more than 1,000 reads in the previous MiSeq data set reported by Narzisi *et al.* Hence, we only performed additional MiSeq validation on the remaining 503 loci<sup>112</sup>. PCR primers were designed using Primer 3 to produce amplicons ranging in size of 200 - 350 bp, with indels of interest located approximately in the center. Primers were obtained from Sigma-Aldrich in 96-well mixed-plate format, 10 μmol/L dilution in Tris per oligonucleotide. 25 μL PCR reactions were set up to amplify each indel of interest using K8101-49685s' genomic DNA as template and LongAmp Taq DNA polymerase (New England Biolabs). PCR products were visually inspected for amplification efficiency using 1.5% agarose gel electrophoresis, and then pooled for ExoSAP-IT (Affymetrix) cleanup. The cleanup product was purified using QIAquick PCR Purification Kit (Qiagen) and quantified by Qubit dsDNA BR Assay Kit (Invitrogen). Subsequently, a library construction was performed following the TruSeq Nano DNA Sample Preparation Guide for the MiSeq Personal Sequencer platform (Illumina). Before loading onto the MiSeq machine, the quality and quantity of the sample was reevaluated using the Agilent DNA 1000 Kit on the Agilent Bioanalyzer and with quantitative PCR (Kapa Biosystems).



We generated high quality 250 bp paired-end reads with an average coverage of 55,000X over the selected indels. We aligned the reads with BWA-MEM (v0.7.5a) to hg19, sorted the alignment with SAMtools (v0.1.18) and marked PCR duplicates with Picard (v1.91). The alignment quality control showed that 371 out of the 503 loci were covered with at least 1,000 reads in the data and we only considered these loci in the downstream analysis. Therefore, we have validation data on 160, 145 and 161 loci from the WGS-WES intersection, WGS-specific, and WES-specific indels, respectively. As reported by Narzisi *et al*, mapping the reads containing a large indel (near or greater than half the size of the read length) is problematic. This was particularly difficult when the indel is located toward either end of a read<sup>112</sup>. To avoid this, we used very sensitive settings with Bowtie2 (--end-to-end --very-sensitive --score-min L,-0.6,-0.6 --rdg 8,1 --rfg 8,1 --mp 20,20) to align the reads because it can perform end-to-end alignment and search for alignments with all of the read characters<sup>120</sup>. We generated the true indel call set by two steps: 1) used GATK UnifiedGenotyper to call indels from the BWA-MEM alignment, 2) performed manual inspection on the large indels from the Bowtie2 alignment (require at least 25% of the reads supporting an indel)<sup>114</sup>. The alignments were realigned with the GATK (v2.6-4) IndelRealigner and base quality scores were recalibrated before variants were called with UnifiedGenotyper. Left-normalization was performed to avoid different representations of a variant. An indel was considered valid if a mutation with the same genomic coordinate and the same type of variation exists in the validation data. For example, an insertion call would not be considered valid if the variant with the same coordinate in the validation data was instead a deletion. All of the MiSeq data can be downloaded from the Sequence Read Archive under project accession number SRX386284.

### Classifications of indels with calling quality

We previously benchmarked Scalpel with respect to the coverage of the alternative allele ( $C_o^{Alt}$ ) and the k-mer Chi-Square scores ( $\chi^2$ ). Scalpel applied the standard formula for the Chi-Square statistics and applied to the K-mer coverage of both alleles of an indel.

$$\chi^2 = \frac{(C_o^{Ref} - C_e^{Ref})^2}{C_e^{Ref}} + \frac{(C_o^{Alt} - C_e^{Alt})^2}{C_e^{Alt}} \quad \text{Equation 3.2}$$

where  $C_o^{Ref}$  and  $C_o^{Alt}$  are the observed k-mer coverage for the reference and alternative alleles,  $C_e^{Ref}$  and  $C_e^{Alt}$  are the expected k-mer coverage, i.e.  $C_e^{Ref} = C_e^{Alt} = \frac{C_o^{Ref} + C_o^{Alt}}{2}$ .

Here we used 466 indels from the validation data to understand the relationship between the FDR and these two metrics (Supplemental Figure S4). Our validation data showed that with the same  $\chi^2$ , indels with a lower  $C_o^{Alt}$  tend to have a higher FDR, especially for indels with  $C_o^{Alt}$  not greater than 10 (Supplemental Figure S4). For indels with relatively the same  $C_o^{Alt}$ , a higher  $\chi^2$  also made them less likely to be valid. We noticed that the calling quality could be determined by the error rate inferred by these two metrics. To achieve a consistent accuracy for indels with different  $C_o^{Alt}$ , we classified indel calls and determined the calling quality with the below criteria:

High quality indels: low error-rate (7%) indels meeting any of the three cutoffs:  $C_o^{Alt} > 10$  and  $\chi^2 < 10.8$ , or  $5 < C_o^{Alt} < 10$  and  $\chi^2 < 4.5$ , or  $C_o^{Alt} < 5$  and  $\chi^2 < 2$ ;

Low quality indels: high error-rate (51%) indels meeting the following cutoff:  $C_o^{Alt} < 10$  and  $\chi^2 > 10.8$ ;

Moderate quality: The remaining indels that do not fall into the above two categories.

## Results

### Characterizing alignment and assembly based callers at different coverage

We started our study with asking whether depth of sequencing coverage affect different kinds of indels calling algorithms. (e.g. assembly based callers and alignment based callers). Thus, we began with simulated reads with known error rates across the genome to answer this question. We used GATK-UnifiedGenotyper (v3.2-2) and Scalpel (v0.1.1) as a representative of alignment based callers and assembly based callers, respectively. Figure 3.1A shows that for both algorithms, higher coverage improves sensitivity of detecting both general indels (i.e. any size starting from 1 bp) and large indels (i.e. size greater than 5 bp). For general indels detection with both algorithms, this improvement did not saturate until a mean coverage of 28X. Furthermore, detecting large indels was more difficult than general indels because the increase of sensitivity did not saturate until reaching a mean coverage of 42X. However, there were substantial differences of sensitivity performance between these two algorithms for large indel detection. We noticed that even at a very high coverage (mean coverage = 90X), GATK-UnifiedGenotyper could call only about 52% of the large indels while Scalpel could reveal more than 90% of them. This is because GATK-UnifiedGenotyper tries to infer genotypes from alignment and large indels could complicate or distort the correct mapping. To achieve a sensitivity of 90% with Scalpel, a mean coverage of 30X was required for general indel detection while 90X was needed to detect large indels at a similar sensitivity. This showed that much higher coverage is needed for large indel detection, especially to maintain coverage across the indel and to have enough partially mapping or soft-clipped reads to use for the micro-assembly.

The FDRs of Scalpel were robust to the changes in coverage while GATK-UnifiedGenotyper's FDRs were affected by coverage. For the detection of large indels with Scalpel, the FDRs marginally decreased as the mean coverage increased from 5X to 28X, and remained basically the same again from 33X to 93X (Figure 3.1B). This indicates that for large indels, insufficient coverage results in more assembly errors, which results in a higher error rate for micro-assembly variant calling. Based on the simulation data, a mean coverage of at least 30X is needed to maintain a reasonable FDR for Scalpel. In contrast, FDRs of GATK-UnifiedGenotyper are much higher and more unstable at different coverages, especially for large indels. Nonetheless, since these results were based on simulation data, which does not include the effects of any sequencing artifacts on indel calling, these values establish the upper bound of accuracy and performance compared to genuine sequence data. Previous studies reported that local assembly allows to call indels much larger than those that can be identified by the alignment<sup>85, 112, 121</sup>. Consistent with previous reports, our simulated data suggested that assembly based callers can reveal a much larger spectrum of indels than alignment based callers, in terms of their size. Furthermore, Narzisi *et al.* recently reported that Scalpel is more accurate than GATK-HaplotypeCaller and SOAPindel, especially within regions containing near-perfect repeats<sup>112</sup>. Thus, in order to control for artifacts from callers, we chose to use Scalpel as the only indel caller in our downstream analysis on the experimental data, which could help to better clarify differences between data types.

## WGS vs. WES: Low concordance on indel calling

We analyzed a dataset including high coverage WGS and WES data from eight samples in the SSC. To make a fair comparison, the indel calls were only made from the exonic targeted regions as explained in the Methods. The mean concordance between WGS and WES data was low, 53% using exact-match and 55% using position-match (Figure 2, Table 3.1). Position-match means the two indels have the same genomic coordinate, while exact-match additionally requires that the two indels also have the same base-pair change(s) (see Methods). When we excluded regions with less than one read in either dataset, the mean concordance rates based on exact match and position-match increased to 62% and 66%, respectively (Table 3.1). If we excluded regions with base coverage in either dataset with less than 20, 40, 60, or 80 reads, the mean concordance rate based on exact-match and position-match both continued to increase until reaching a base coverage of 80 reads (Table 3.1). This showed that some indels were missing in either dataset because of low sequencing efficiency in those regions. Although WES data had higher mean coverage than WGS data, we were surprised to see that in regions requiring at least 80 reads, there were more indels that were specific to WGS data than WES data (21% vs. 4%). Regions with excessive coverage might indicate problems of sequencing or library preparation, and this highlights the importance of coverage uniformity in WGS (Figure 3.3AB, Table 3.2). It should be noted that mapping artifacts could also be a possible reason. For example, the reads may originate in regions which are absent from the reference genome, such as copy number variants<sup>122</sup>. Based on exact-match, the proportion of the WGS-specific indels was 2.5-fold higher than that of WES-specific indels (34% vs. 14%). This difference was even larger based on position-match (3-fold). In principle, the reasons for this could be either high sensitivity of indel detection with WGS data or high specificity of indel detection with WES data, and we will examine these options in more detail.

## Coverage distributions of different regions in WGS and WES data

An ideal sequencing experiment should result in a high number of reads covering a region of interest uniformly. Using the eight SSC samples, we investigated the coverage behaviours of the WGS and WES data by the following: distribution of the read depth, mean coverage, coverage fraction at X reads, coefficient of variation ( $C_v$ ) (See methods). Hence, ideally one should expect to see a normal distribution of read depth with a high mean coverage and a small  $C_v$ . Comparisons of the coverage distributions are shown in the following order: 1) Exonic targeted regions, i.e. the exons that the exome capture kit was designed to pull down and enrich; 2) WGS-WES intersection indel regions, i.e. the regions where WGS and WES revealed the identical indels based on exact-match; 3) WGS-specific indel regions, i.e. the regions where only WGS revealed indel based on position-match; 4) WES-specific indel regions, i.e. the regions where only WES revealed indels based on position-match.

First, in the exonic targeted regions, the mean coverages across eight samples were 71X and 337X for WGS and WES data, respectively (Figure 3.3AB, Supplemental Table S1). We noticed that there was a recovery issue with WES in some regions, as the coverage fraction at 1X was 99.9% in WGS data but only 84% in WES data, meaning that 16% of the exonic targeted regions were not recovered, which could be due to capture inefficiency or other issues involving DNA handling during the exome library preparation and sequencing protocols (Figure 3.3CD, Supplemental Table S2). The coverage was much more uniform in the WGS data than that in the WES data because  $C_v$  of the WGS data was much lower (39 % vs. 109%, Figure 3.3AB, Table

3.2). Second, in the WGS-WES intersection indel regions, the mean coverage across eight samples were 58X and 252X for WGS and WES data, respectively (Supplemental Figure S1A & B, Supplemental Table S1). We noticed that there was an increase of coverage uniformity for WES in the WGS-WES intersection indel regions, relative to the exonic targeted regions, because  $C_v$  was lower (109% vs. 97%) (Table 3.2, Figure 3.3B, Supplemental Figure S1b). We noticed WGS was able to reveal WGS-WES intersection indels at a much lower coverage relative to WES, which we attribute to a better uniformity of reads across the genome ( $C_v$ : 47% vs. 97%, Table 3.2). The coverage distributions were skewed in the WES data, with some regions poorly covered and other regions over saturated with redundant reads.

Third, in WGS-specific indel regions, the mean coverages across eight samples were 61X and 137X for WGS and WES data, respectively (Figure 3.4, Supplemental Table S1). Compared to the entire exonic targeted regions, the mean coverage for WES data was significantly reduced in these regions (137X vs. 337X), and 44% of the regions were not covered with a single read (Figure 3.4). We noticed that compared to the WGS data, the WES data poorly covered these regions with 20 reads or more (94% vs. 31%, Figure 3.4CD). In these regions, the coverage uniformity of the WES data was much lower than that of the WGS data ( $C_v$ : 282% vs. 75%, Figure 3.4AB, Table 3.2). The reason why WES data missed these indels could be insufficient coverage around the indels in these regions. Finally, in WES-specific indels regions, the mean coverages across eight samples were 41X and 172X for WGS and WES data, respectively (Supplemental Figure S2A & B, Supplemental Table S1). In these regions, both data had a relatively high coverage and the WES data covered most these regions with at least one read (Supplemental Figure S2C & D). However, we noticed that the WES data still had a much lower coverage uniformity ( $C_v$ : 117% vs. 56%, Table 2). In order to better understand these issues, we used the WGS-WES intersection indel set as a positive control and proceeded to assess each call set with newly developed quality criteria.

#### MiSeq validation of indels in WGS and WES data on the sample K8101-49685s

In order to understand error rates and behaviours of the indel call from the WGS and WES data, we randomly selected 200 indels for MiSeq validation on the sample K8101-49685s from each of the following categories: 1) indels called from both WGS and WES data (WGS-WES intersection indels), 2) WGS-specific indels, 3) WES specific indels. First, the validation rate of WGS-WES intersection indels was in fact very high (95%), indicating indels called from both WGS and WES data were mostly true-positives (Table 3.3). Second, the validation rate of WGS-specific indels was much higher than that of WES-specific indel (84% vs. 57%). Third, among the validation set, large indels (> 5 bp) that were called from both the WGS and WES data were 100% valid, while the validation rate of large indels that were specific to the WGS data was only 76%. However, we noticed that there was only one large indel specific to the WES data that we selected for validation. Since the sampling was performed randomly, we examined the original call set to understand this phenomenon. Only 9% of the WGS-WES intersection indels (176) and 21% of the WGS-specific indels (106) were greater than 5 bp (Table 3.4). But only 1.5% of the WES-specific INDELS were greater than 5 bp, meaning only 10 indels were large according to our definition. This showed that the WES data missed most large indels, which I speculate might be due to capture deficiency or some other procedure related to the process of exome capture and sequencing. In particular, large indels could disrupt

the base pairing that occurs during the exome capture procedure, which would then result in insufficient coverage in those regions (Figure 3.4).

### Assessment of the indels calls sets from WGS and WES

To understand the error profile of the WGS and WES data with a larger sample size, I developed a classification scheme based on the validation data and applied them to the eight samples in the Simons Simplex Collection (SSC). Three combinations of thresholds were used to define the calling quality of an indel call as either high, moderate or low quality based on the following two metrics: the coverage of the alternative allele and the k-mer Chi-Square score of an indel (see Methods). Based on those cutoffs, there was 7.3-fold difference between high-quality and low-quality INDELs in terms of their error rates (7% vs. 51%). This suggests that my classification scheme is able to effectively distinguish behaviours of problematic indel calls from likely true-positives. Our classification scheme is also useful for eliminating false *de novo* indel calls in family-based studies. Furthermore, WGS-WES intersection and WGS-specific indel seem to be reliable calls, and the majority of the indels in these two call sets were of high-quality, 89% and 78% respectively. Only a very small fraction of them were of low-quality, 2% and 7% respectively. (Figure 3.5, Supplemental Table S3). In contrast, for WES-specific indels, there was a striking enrichment of low-quality events (41%), and a 4.1-fold decrease of the high-quality events (22%). Notably, among these 8 samples, there were 991 WGS-specific indels and 326 WES-specific indels, and from these, 769 of WGS-specific indels and 71 of the WES-specific indels were of high quality. This comparison determined that WGS yielded 10.8-fold more high quality indels than WES according to our classification scheme. Furthermore, WES produced 133 low quality indels per sample, while WGS only produced 71 low quality indels per sample. That being said, WES yielded 1.9-fold more low quality indels. This indicates WES tends to produce a larger fraction of error-prone indels, while WGS reveals a more sensitive and specific set of indels.

In order to understand what was driving the error rates in different data sets, we partitioned the indels according to their sequence composition: homopolymer A (poly-A), homopolymer C (poly-C), homopolymer G (poly-G), homopolymer T (poly-T), short tandem repeats (STR) except homopolymers (other STR), and non-STR. We noticed that for the high quality events, the majority of the WGS-WES intersection indels (70%) and WGS-specific indels (67%) were within non-STR regions (Figure 3.6, Supplemental Table S4 & S5). On the contrary, the majority of the high quality indels specific to WES were within poly-A (24%) and poly-T regions (30%). When we compared the low quality indels to the high quality indels, there were consistent enrichment of homopolymer A or T (poly-A/T) indels in all three call sets, 2.3-fold for WGS-WES intersection events, 2.1-fold for WGS-specific events, and 1.5-fold for WES-specific events. The WES-specific call set contained a much higher proportion (83%) of Poly-A/T indels from the low-quality indels, relative to the WGS-WES intersection call set (44%), and the WGS-specific call set (45%). This suggested that poly-A/T is a major contributor to the low quality indels, which gives rise to much more indel errors. I explored this further in the comparison of PCR-free and standard WGS data below.

### Sources of multiple signatures in WGS and WES data

Another way of understanding indel errors is to look at multiple signatures at the same genomic location. Multiple signatures means that for the same genomic location, there are more



than one indels called. If we assume only one signature can be the true indel in the genome, any additional signatures would represent false-positive calls. So if we have a higher number of multiple signatures, it means that these reads contained more indel errors or the algorithm tends to make more mistakes in these regions. We combined the call sets from both datasets and identified multiple signatures in the union set for each sample. In order to understand the error behaviors in the above assessment, we also partitioned the signatures by the same regional criteria. We noticed that the poly-A/T indels are the major source of multiple signatures, which are enriched in WES data (72% for WES vs. 54% for WGS). In particular, there is a higher number of poly-A (35 vs. 25) and poly-T (36 vs. 16) indel errors in the WES data than in the WGS data (Figure 3.7). We investigated the source of multiple signatures by the numbers of reads containing homopolymer indels inferred by the CIGAR code (Figure 3.8). Figure 3.8 showed that there is a much higher proportion of poly-A/T indels in the WES-specific regions from both WGS (56%) and WES data (64%), relative to other regions. In addition, WES data has also 6.3-fold more reads than WGS data in the regions with indels specific to WES data (11251 vs. 1775, Supplemental Table S7). According to Qualimap, a large number of homopolymer indels might indicate a problem in sequencing for that region. Here I particularly identified the effects of these problematic sequencing reads on indel calling, which revealed more multiple signatures of poly-A/T indels.

#### Standard WGS vs. PCR-free: assessment of indels calling quality

The concordance rate within the exonic targeted regions between standard WGS (defined as WGS involving PCR during library construction) and PCR-free data on NA12878 using exact-match and position-match were 71% and 76%, respectively (Figure 3.9). Note that both data used here are WGS data, so it is not surprising that these concordance rates were higher than those between WGS and WES, even for regions having at least one read in both datasets. Based on exact-match, the proportion of indels specific to standard WGS data was 18%, which is 1.6-fold higher than the proportion of indels specific to PCR-free data (11%). This ratio was similar based on position-match (1.7-fold). Like previous assessments, we classified the three call sets with respect to calling quality. We again used the indels called from both standard WGS and PCR-free data as a positive control. Figure 3.10 shows that 89% of the standard WGS & PCR-free intersection indels are considered as high quality, 9% as moderate quality, and only 2% as low quality. However, for indels specific to standard WGS data, there is a large proportion of low quality events (61%), and a very limited proportion are of high quality (7%). There were on average 310 indels specific to PCR-free data and 538 indels specific to standard WGS data. Notably, 177 of the PCR-free-specific indels and 40 of the standard-WGS-specific indels were of high quality, suggesting that in these specific regions, PCR-free data yielded 4.4-fold more high quality indels than standard WGS data. Furthermore, 326 of the standard-WGS-specific indels were of low quality, while in the PCR-free-specific call set, 52 indels were of low quality. That being said, in regions specific to data types, standard WGS data yielded 6.3-fold more low quality indels. Consistent with the comparisons between WGS and WES data, this suggested PCR amplification induced a large number of error-prone indels to the library, and we could effectively increase indel calling quality by reducing the rate of PCR amplification.

To understand the behaviors of errors in the poly-A/T regions, we partitioned the indel call set by the same six regions again. We noticed that for the high quality events, a majority of the standard WGS & PCR-free intersection indels (68%) were within non-STR regions (Figure

3.11). The proportion of poly-A/T indels was small for the standard WGS & PCR-free intersection call set (20%), larger for PCR-free-specific call set (35%), and even larger for standard-WGS-specific call set (51%). This was similar to the WGS and WES comparisons because there would be more poly-A/T INDELS when a higher rate of PCR amplification was performed. A majority of the high quality INDELS specific to standard WGS data were within poly-A (24%) and poly-T regions (38%). When we compared the low quality indels to the high quality ones, there was consistent enrichment of poly-A/T indels in all three call sets, 2.3-fold for standard WGS & PCR-free intersection events, 2.3-fold for PCR-free-specific events, and 1.3-fold for standard-WGS-specific events. For indels specific to standard WGS data and PCR-free data, poly-A/T indels represented a large proportion of the low quality indels: 80% and 62%, respectively. Ross *et al.* previously reported that for human samples, PCR-free library construction could increase the relative coverage for high AT regions from 0.52 to 0.82, resulting in a more uniform coverage<sup>94</sup>. This again suggested that PCR amplification could be a major source of low quality poly-A/T indels, and a PCR-free library construction protocol might be one possible solution to improve the accuracy of calls.

#### What coverage is required for accurate indel calling?

Ajay *et al.* 2011 reported that the number of SNVs detected exponentially increased until saturation at 40-45X average coverage<sup>123</sup>. However, it was not clear what the coverage requirement should be for indel detection. To answer this question, we down-sampled the reads, called indels again, and measured corresponding sensitivity for each sample using the WGS-WES intersection calls as our truth set (Methods). Figure 3.12A shows that we are missing 25% of the WGS-WES intersection indels at a mean coverage of 30X. Even at 40X coverage recommended by Ajay *et al.* 2011<sup>123</sup>, we could only discover 85% of the WGS-WES intersection indels. We calculated that WGS at 60X mean coverage (after removing PCR duplicates) from the HiSeq 2000 platform is needed to recover 95% of indels with Scalpel, which is much higher than current sequencing practice (Figure 3.12A). If economically possible, WGS at 60X mean coverage with PCR-free library preparation would generate even more ideal sequencing data for indel detection. Some groups previously reported that determining heterozygous SNPs requires higher coverage than homozygous ones<sup>124</sup>. The sensitivity of heterozygous SNP detection was limited by depth of coverage, which requires at least one read from each allele at any one site and in practice much more than one read to account for sequencing errors<sup>125</sup>. However, the read depth requirement of indel detection in terms of zygosity has not been well understood. To answer this question, we took the WGS-WES intersection indels and partitioned them by zygosity. We first plotted the pair-wise coverage relationship between WGS and WES for each WGS-WES intersection indel. Supplemental Figure S3 shows that the detection of homozygous indels starts with a lower coverage, which is consistent in both WGS and WES datasets, although the rest of the homozygotes and heterozygotes were highly overlapping. To further understand this phenomenon, we measured the sensitivity again for heterozygous indels and homozygous indels separately. At a mean coverage of 20X, the false negative rates of WGS-WES intersection indels was 45% for heterozygous indels and 30% for homozygous indels, which is consistent with the fact that homozygous indels are more likely to be detected at a lower coverage shown above (Figure 12B). This shows that one should be cautious about the issue of false-negative heterozygous indels in any sequencing experiment with a low coverage (less than 30X). Figure 12B also shows that detection of heterozygous indels indeed requires higher coverage than homozygous ones

(sensitivity of 95% at 60X vs. 50X). Notably, the number of heterozygous indels was 1.6-fold higher than homozygous ones (1600 vs. 635 per sample). This re-affirms the need for 60X mean coverage to achieve a very high accuracy call set.



## Discussion

Despite the fact that both WES and WGS have been widely used in biological studies and rare disease diagnosis, limitations of these techniques on indel calling are still not well characterized. One reason is that accurate indel calling is in general much more difficult than SNP calling. Another reason is that many groups tend to use WES, which we have determined is not ideal for indel calling for several reasons. We report here our characterization of calling errors for indel detection using Scalpel. As expected, higher coverage improves sensitivity of indel calling, and large indel detection is uniformly more difficult than detecting smaller indels. We also showed that assembly based callers are more capable of revealing a larger spectrum of indels, relative to alignment based callers. There are several reasons for the low concordance for WGS and WES on indel detection. First, due to the low capture efficiency, WES failed to capture 16% of candidate exons, but even at sites that were successfully captured, there were more coverage biases in the WES data, relative to the WGS data. Second, PCR amplification introduces reads with higher indel error rate, especially in regions near homopolymer A/T's. Lastly, STR regions, especially homopolymer A/T regions were more likely to result in multiple candidates at the same locus. We recommend controlling for homopolymer false calls with a more stringent filtering criteria. This is essential for population-scale sequencing projects, because the expense of experimental validation scales with the sample size.

Our validation data showed that indels called with both WGS and WES data were indeed of high quality and with a low error rate. Even though the WGS data has much lower depth coverage in general, the accuracy of indel detection with WGS data is much higher than that with WES data. We also showed that the WES data is missing many large indels, which we speculated might be related to the technical challenges of pulling down the molecules containing large indel during the exon capture process. Homopolymer A/T indels are a major source of low quality indels and multiple signature events, and these are highly enriched in the WES data. This was confirmed by the comparison of PCR-free and standard WGS data. In terms of sensitivity, we calculated that WGS at 60X mean coverage from the HiSeq platform is needed to recover 95% of indels with Scalpel. As more and more groups are moving to use new micro-assembly based algorithms such as Scalpel, practical considerations for experimental design should be introduced to the community. Here I presented a novel classification scheme utilizing the validation data, and I encouraged researchers to use this guideline for evaluating their call sets. The combination of alternative allele coverage and the k-mer Chi-Square score is an effective filter criterion for reducing indel calling errors without sacrificing much sensitivity. This classification scheme can be easily applied to screen indels calls from all variant callers. For consumer genome sequencing purposes, we recommend sequencing human genomes at a higher coverage with a PCR-free protocol, which can substantially improve the quality of personal genomes. Although this recommendation might initially cost more than the current standard protocol of genome sequencing used by some facilities, we argue that the significantly higher accuracy and decreased costs for validation would ultimately be cost-effective as the sequencing costs continue to decrease, relative to either WES or WGS at a lower coverage. However, it is important to point out that with the release of Illumina HiSeq X-Ten and other newer sequencers, the coverage requirement to accurately detect indels may decrease because reads with longer read length can span repetitive regions more easily. Besides, bioinformatics algorithms are another important consideration, and we expect the further enhancements of Scalpel and other algorithms will help reduce the coverage requirement while maintaining a high accuracy.

Figures

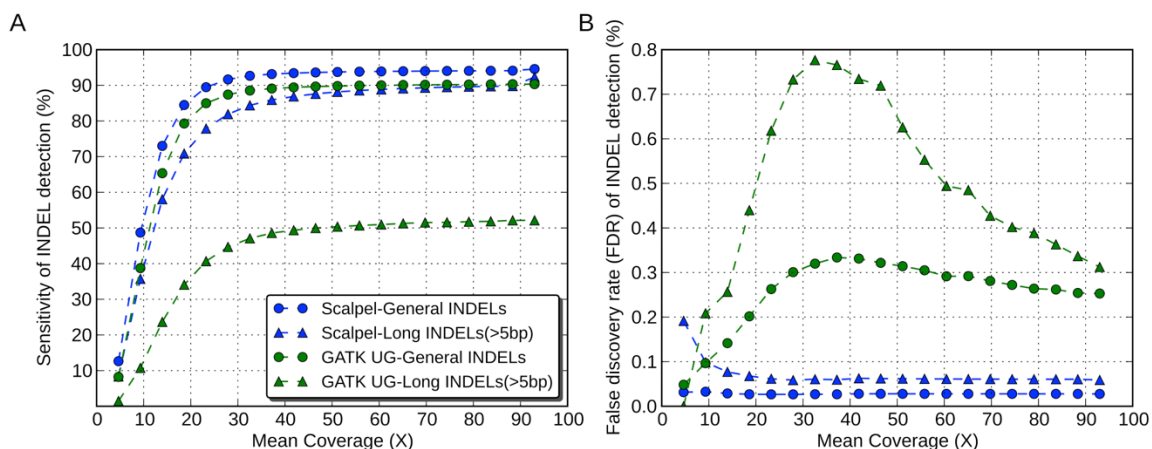


Figure 3.1. Performance comparison between the Scalpel and GATK-UnifiedGenotyper in terms of sensitivity (A) and false discovery rate (B) at different coverage based on simulation data. Each dot represent one down-sampled experiment. Round dots represent performance of general INDELs (i.e. INDELs of size starting at 1 bp) and triangles represent performance of large INDELs (i.e. INDELs of size greater than 5 bp). The data of Scalpel was shown in blue while GATK-UnifiedGenotyper was shown in green.

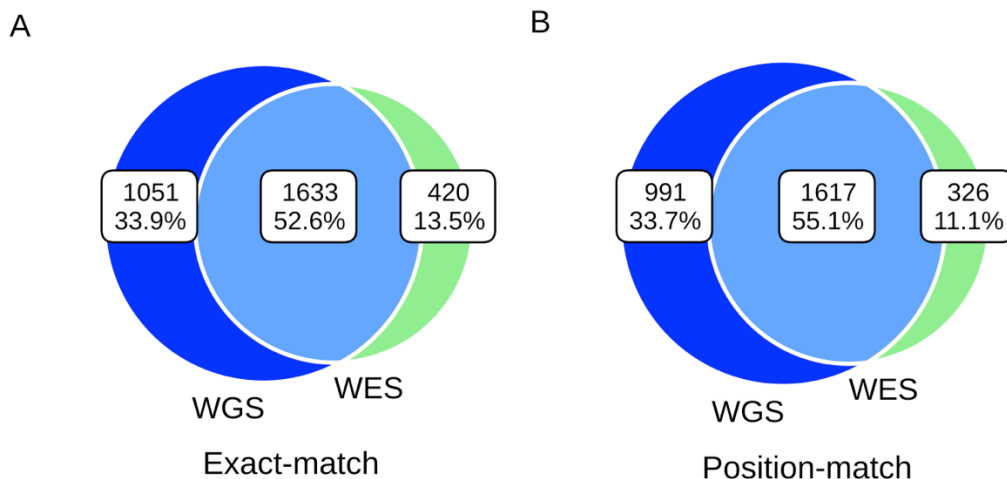


Figure 3.2. Mean concordance of INDELs over eight samples between WGS (blue) and WES (green) data. Venn diagram showing the numbers and percentage of shared between data types based on (A) Exact-match (B) Position-match. The mean concordance rate increased when we required at least a certain number of reads in both data (Table 1).

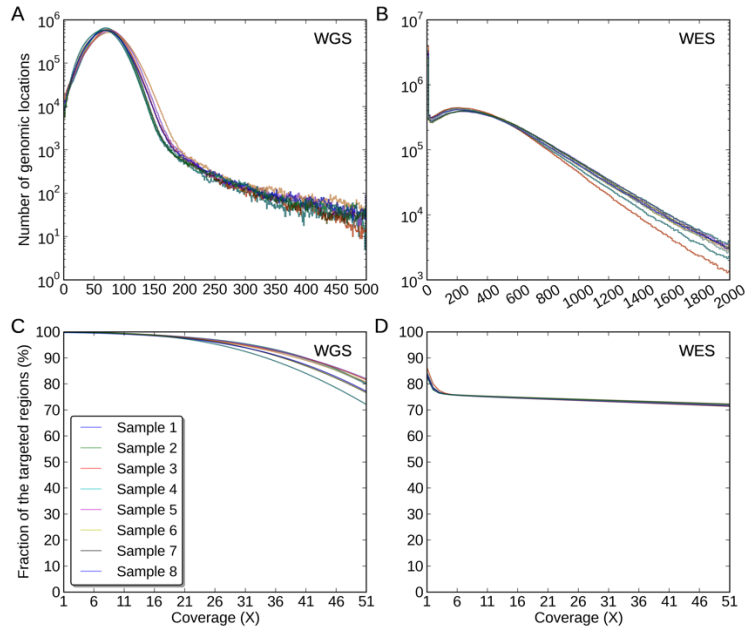


Figure 3.3. Coverage distributions of the exonic targeted regions in (A) the WGS data, (B) the WES data. The Y-axis for A) and B) is of log<sub>10</sub>-scale. The coverage fractions of the exonic targeted regions from 1X to 51X in (C) the WGS data, (D) the WES data.

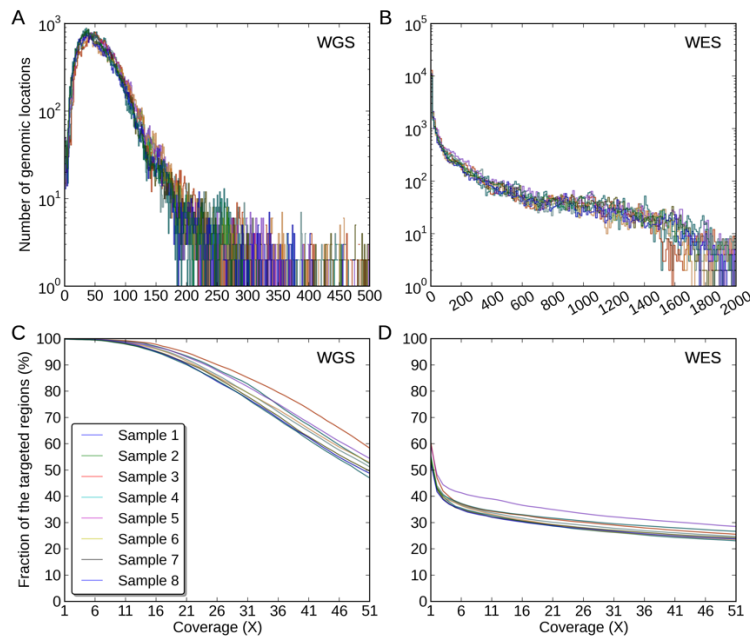


Figure 3.4. Coverage distributions of the WGS-specific INDELs regions in (A) the WGS data, (B) the WES data. The Y-axis for A) and B) is of log<sub>10</sub>-scale. The coverage fractions of the WGS-specific INDELs regions from 1X to 51X in (C) the WGS data, (D) the WES data.

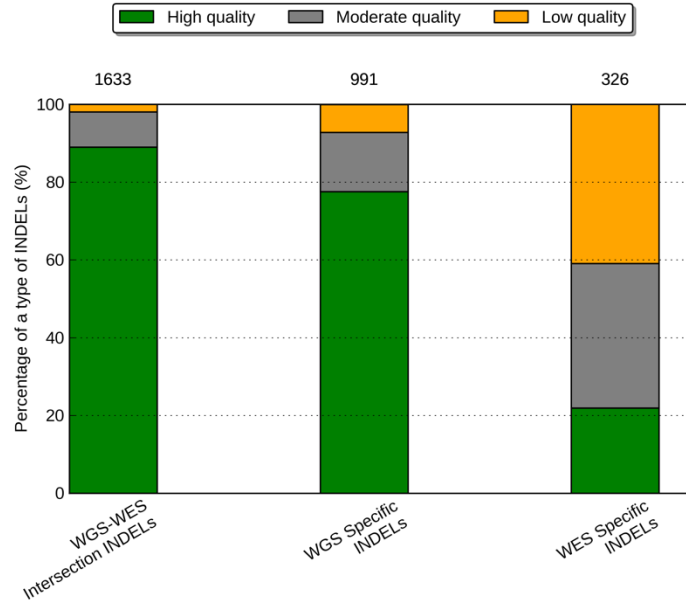


Figure 3.5. **Percentage of high quality, moderate quality and low quality INDELs in three call set.** (A) the WGS-WES intersection INDELs, (B) the WGS-specific INDELs, (C) the WES-specific INDELs. The numbers on top of a call set represent the mean number of INDELs in that call set over eight samples.

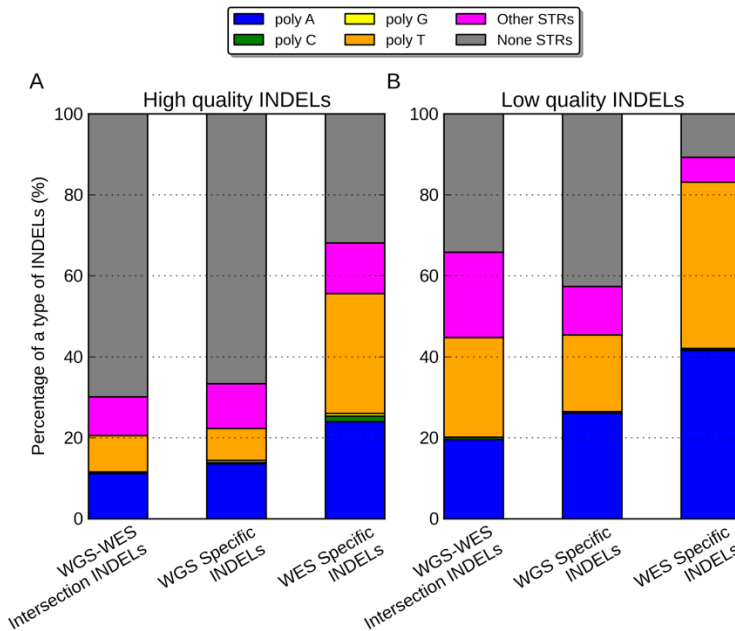


Figure 3.6. **Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR in three call set.** (A) high quality INDELs, (B) low quality INDELs. In both figures, from left to the right are WGS-WES intersection INDELs, WGS-specific INDELs, and WES-specific INDELs.

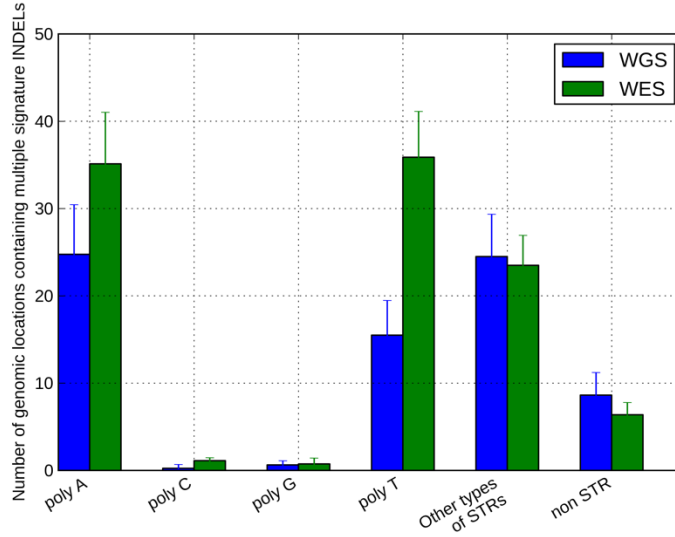


Figure 3.7. Numbers of genomic locations containing multiple signature INDELs in WGS (blue) and WES data (green). The height of the bar represents the mean across eight samples and the error bar represent the standard deviation across eight samples.

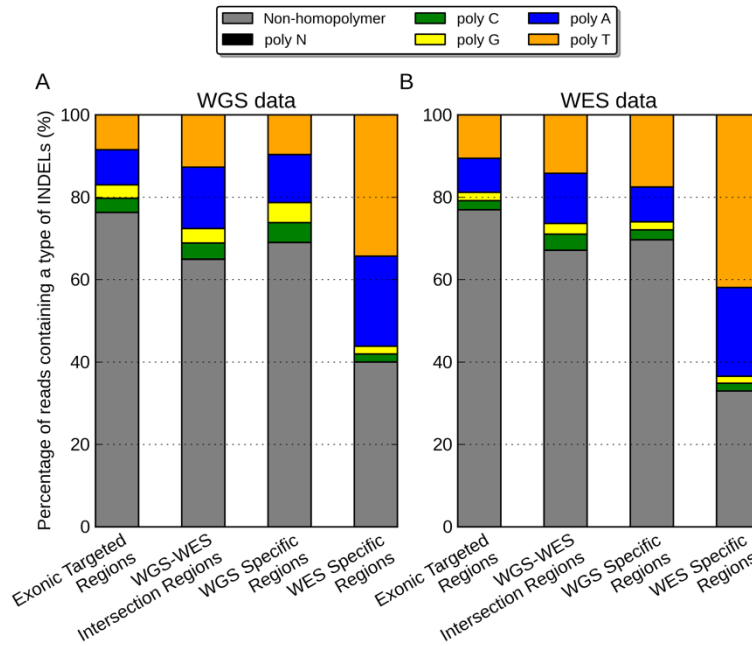


Figure 3.8. Percentage of reads near regions of Non-homopolymer, poly-N, poly-A, poly-C, poly-G, poly-T in (A) WGS data, (B) WES data. In both figures, from left to the right are exonic targeted regions, WGS-WES intersection INDELs, WGS-specific INDELs, and WES-specific INDELs.

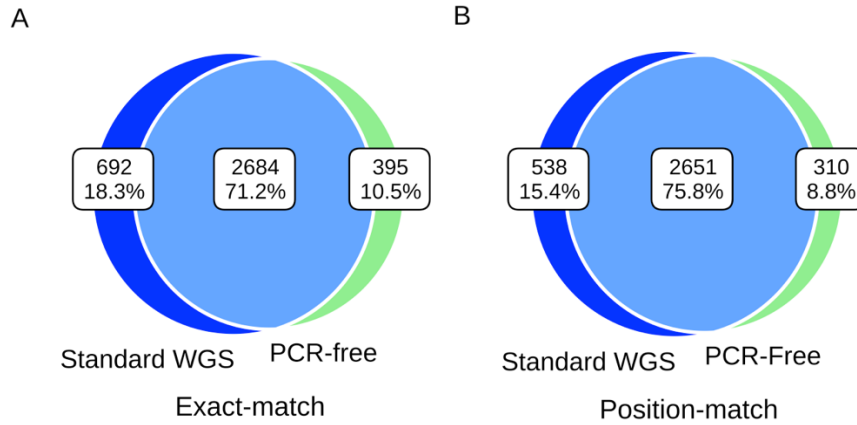


Figure 3.9. **Concordance of INDEL detection between PCR-free and standard WGS data on NA12878.** Venn diagram showing the numbers and percentage of shared between data types based on (A) Exact-match (B) Position-match.

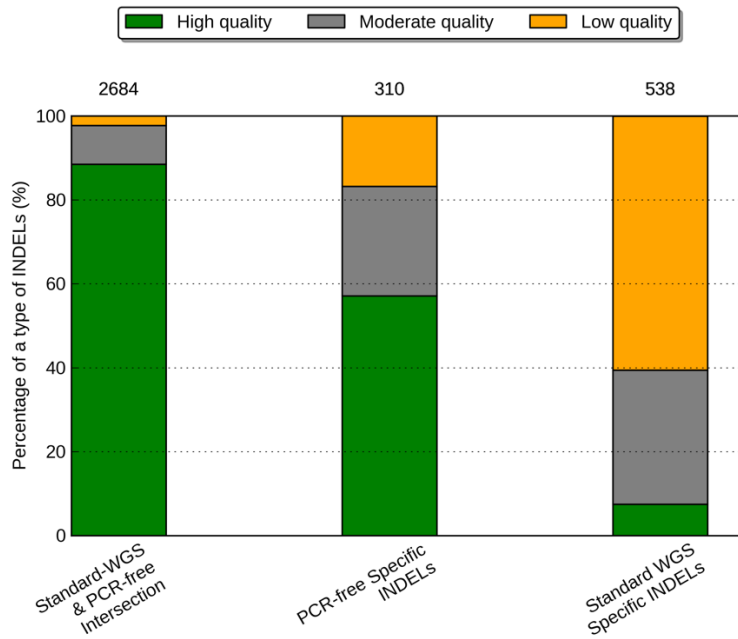


Figure 3.10. **Percentage of high quality, moderate quality and low quality INDELs in two datasets.** (A) the PCR-free & standard WGS INDELs, (B) the PCR-free-specific INDELs, (C) the standard-WGS-specific INDELs. The numbers on top of a call set represent the number of INDELs in that call set.

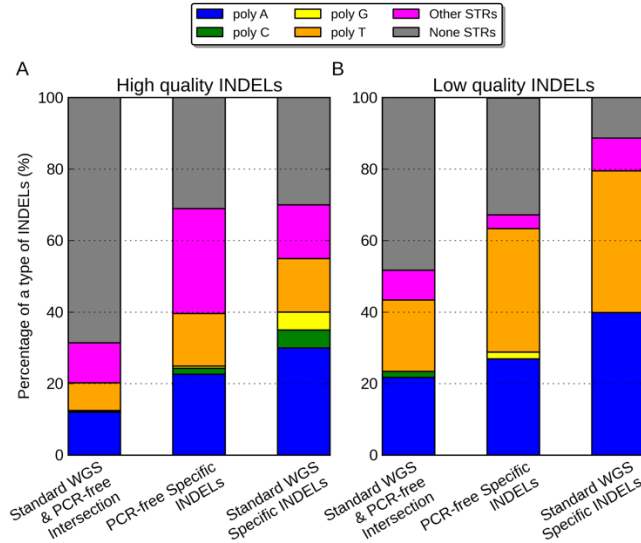


Figure 3.11. **Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR in (A) high quality INDELS, (B) low quality INDELS.** In both figures, from left to the right are PCR-free & standard WGS INDELS, INDELS specific to PCR-free data, and INDELS specific to standard WGS data.

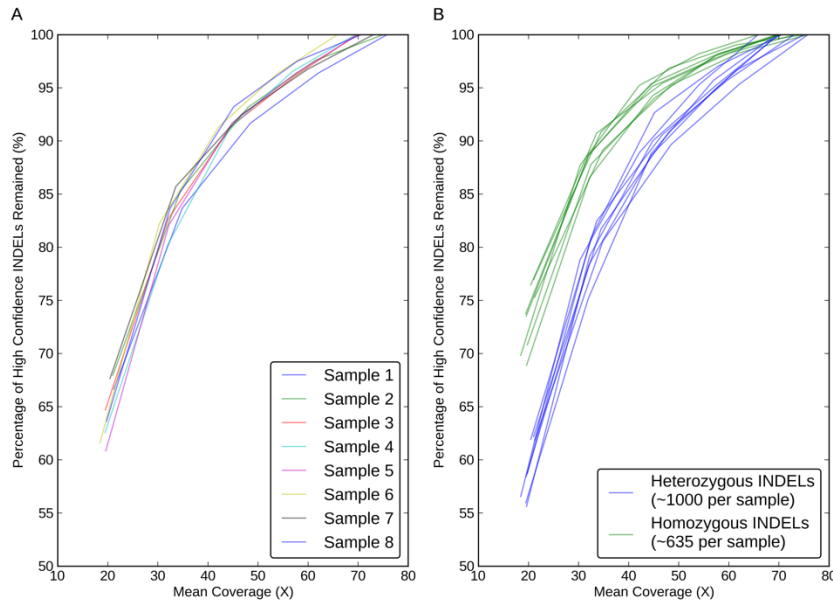


Figure 3.12. **Sensitivity performance of INDEL detection with eight WGS datasets at different mean coverages on Illumina HiSeq2000 platform.** The Y-axis represents the percentage of the WGS-WES intersection INDELS revealed at a certain lower mean coverage. (A) Sensitivity performance of INDEL detection with respects with each sample, (B) Sensitivity performance of heterozygous (blue) and homozygous (green) INDEL detection were shown seperately.

Tables

Table 3.1. **Mean concordance and discordance rates of INDEL detection between WGS and WES data in different regions.** The data is shown in the following order: 1) regions without filtering, and regions filtered by requiring base coverage to be at least 2) one read, 3) 20 reads, 4) 40 reads, 5) 60 reads, or 6) 80 reads in both data. The mean discordance rate is calculated based on position-match, which is the percentage of INDELS specific to either dataset. The standard deviation is shown in parenthesis.

Concordance Rate	Without filtering	$\geq 1$ read	$\geq 20$ reads	$\geq 40$ reads	$\geq 60$ reads	$\geq 80$ reads
<b>Exact-match</b>	53% (0.8%)	62% (1.1%)	69% (1.5%)	73% (2.3%)	76% (1.6%)	74% (1.3%)
<b>Position-match</b>	55% (0.8%)	66% (1.0%)	73% (1.1%)	77% (1.8%)	79% (1.1%)	76% (1.3%)
Discordance Rate	Without filtering	$\geq 1$ read	$\geq 20$ reads	$\geq 40$ reads	$\geq 60$ reads	$\geq 80$ reads
<b>WGS-Specific</b>	34% (1.4%)	20% (1.5%)	14% (1.6%)	14% (2.2%)	15% (2.5%)	20% (3.2%)
<b>WES-Specific</b>	11% (1.2%)	14% (1.4%)	13% (1.3%)	9% (2.6%)	6% (2.2%)	4% (1.5%)

Table 3.2. **Mean coefficients of variation of coverage with respects to the following regions: WGS-WES intersection INDELS, WGS-specific INDELS, and WES-specific INDELS.** WGS-WES intersection INDELS means the INDELS called from both WGS and WES data. WGS-specific INDELS means the INDELS only called from the WGS data. The standard deviation is shown in parenthesis.

	Exonic targeted regions	WGS-WES intersection INDEL regions	WGS-specific INDEL regions	WES-specific INDEL regions
<b>WGS</b>	39.4% (1.9%)	47.2% (3.0%)	75.3% (5.7%)	56.1% (9.6%)
<b>WES</b>	109.3% (1.5%)	96.8% (3.2%)	281.5% (13.3%)	117.4% (22.8%)

Table 3.3. **Validation rates of WGS-WES intersection INDELS, WGS-specific, and WES-specific INDELS.** We also calculated the validation rates of large INDELS (>5 bp) in each category. The validation rate, positive predictive value (PPV), is computed by the following:  $PPV = \#TP / (\#TP + \#FP)$ , where #TP is the number of true-positive calls and #FP is the number of false-positive calls.

	INDELS	Valid	PPV	INDELS (>5 bp)	Valid (>5 bp)	PPV (>5 bp)
<b>WGS-WES intersection</b>	160	152	95.0%	18	18	100%
<b>WGS-specific</b>	145	122	84.1%	33	25	75.8%
<b>WES-specific</b>	161	91	56.5%	1	1	100%

Table 3.4. **Number and fraction of large INDELS in the following INDEL categories: 1) WGS-WES intersection INDELS, 2) WGS-specific, and WES-specific.**



	All INDELS	Large INDELS (>5 bp)	Fraction of large INDELS (>5 bp)
<b>WGS-WES intersection</b>	2009	176	8.8%
<b>WGS-specific</b>	494	104	21.1%
<b>WES-specific</b>	674	10	1.5%

## Chapter 4 Benchmarking and applications of Scalpel

### Summary of Contribution

This chapter describes the benchmarking results and applications on population data of Scalpel. The analysis and results were published in *Nature Methods*<sup>49</sup>. Han Fang assisted on the benchmarking against competing algorithms. Giuseppe Narzisi performed the benchmarking based on simulation data, and conducted the population data analysis. Permission for republication of this material has been granted and is available upon request.

### Abstract

We present an open-source algorithm, Scalpel, which combines mapping and assembly for sensitive and specific discovery of indels in exome-capture data. A detailed repeat analysis coupled with a self-tuning k-mer strategy allows Scalpel to outperform other state-of-the-art approaches for indel discovery, particularly in regions containing near-perfect repeats. We analyze 593 families from the Simons Simplex Collection and demonstrate Scalpel's power to detect long ( $\geq 20$ bp) transmitted events, and enrichment for *de novo* likely gene-disrupting indels in autistic children.

### Introduction

While the analysis of Single Nucleotide Variations (SNVs) has become a standard technique to study human genetics<sup>13</sup>, insertions and deletions in DNA sequences (indels) cannot be detected as reliably<sup>126</sup>. Indels are the second most common sources of variation in human genomes and the most common structural variant<sup>106</sup>. Within microsatellites (simple sequence repeats, SSRs, of 1 to 6bp motifs), indels alter the length of the repeat motif and have been linked to more than 40 neurological diseases<sup>127</sup>. Indels also play an important genetic component in autism: *de novo* indels that are likely to disrupt the encoded protein are nearly twice as abundant in affected children than in their unaffected siblings<sup>8</sup>.

Detecting indels is challenging for several reasons: (1) reads overlapping the indel sequence are more difficult to map and may be aligned with multiple mismatches rather than with a gap; (2) irregularity in capture efficiency and non-uniform read distribution increase the number of false positives; (3) increased error rates makes their detection very difficult within microsatellites; and, as shown in this study, (4) localization, near identical repetitive sequences can create high rates of false positives. For these reasons, the size of indels detectable by available software tools has been relatively small, rarely more than a few dozen base pairs<sup>69</sup>.

Two major paradigms are currently used for detecting indels. The first and most common approach is to map all of the input reads to the reference genome using a read mapper (e.g., BWA, Bowtie, Novoalign), although the available algorithms are not as effective for mapping across indels of more than a few bases. Advanced approaches exploit paired-end information to perform local realignments to detect longer mutations (e.g., GATK UnifiedGenotyper<sup>13</sup> and Dindel<sup>14</sup>), although, in practice, their sensitivity is greatly reduced for longer variants ( $\geq 20$ bp).

Split-read methods (e.g., Pindel<sup>18</sup> and Splitread<sup>128</sup>) can theoretically find deletions of any size, but they have limited power to detect insertions due to the short-read length of current sequencing technologies. The second paradigm consists of performing *de novo* whole-genome assembly of the reads, and detecting variations between the assembled contigs and the reference genome<sup>129</sup>. While having the potential to detect larger mutations, in practice this paradigm is less sensitive since detecting indels requires a fine-grained and localized analysis to correctly report homozygous and heterozygous mutations. Recently, three approaches have been developed that use *de novo* assembly for variation discovery: GATK HaplotypeCaller, SOAPindel<sup>54</sup>, and Cortex<sup>29</sup>. Another recent approach, TIGRA<sup>56</sup>, also uses localized assembly, but it has been tailored for breakpoints detection, without reporting the indel sequence.

## Results

We present a DNA sequence micro-assembly pipeline, Scalpel, for detecting indels within exome-capture data (**Figure 4.1**). By combining the power of mapping and assembly, Scalpel carefully searches the de Bruijn graph for sequence paths (contigs) that span each exon. The algorithm includes an on-the-fly repeat composition analysis of each exon, coupled with a self-tuning *k*-mer strategy. We confirm previous findings that nine standard algorithms have reduced power to detect large ( $\geq 20$ bp) indels using simulated reads: Scalpel, SOAPindel, GATK-HaplotypeCaller, GATK-UnifiedGenotyper, SAMtools<sup>15</sup>, FreeBayes, Platypus ([www.well.ox.ac.uk/platypus](http://www.well.ox.ac.uk/platypus)), and lobSTR<sup>130</sup>. We also performed a large-scale validation experiment involving  $\sim 1000$  indels from one single exome. The individual was sequenced to  $\geq 20\times$  coverage over 80% of the exome target using the Agilent SureSelect capture protocol and Illumina HiSeq2000 paired-end reads, averaging 90bp in length, after trimming. Indels were called using the three pipelines that had performed best with our simulated reads: Scalpel v0.1.1 beta, SOAPindel v2.0.1 and GATK HaplotypeCaller v2.4.3. Interestingly, there is only  $\sim 37\%$  concordance among calls made by all of the pipelines, and each method reports hundreds of indels unique to that pipeline (**Figure 4.2a**), which is in close agreement with a recent analysis. An update to GATK to version 3.0 was released after our initial validation experiments, but we also assessed its accuracy with a second blinded re-sequencing experiment (**Figure 4.2**).

From the concordance rate alone, it is hard to judge the quality of indels unique to each pipeline, as these could either represent superior sensitivity or poor specificity. The size distribution of indels called by the HaplotypeCaller (v2.4.3) has a bias towards deletions whereas SOAPindel has a bias towards insertions (**Figure 4.2b**). Scalpel and HaplotypeCaller (v3.0) instead show a well-balanced distribution, in agreement with other studies of human indel mutations. We further investigated the performance of the algorithms by a focused re-sequencing of a representative sample of indels using the more recent 250bp Illumina MiSeq sequencing protocol. Based on the data depicted in **Figure 4.2a**, we selected a total of 1,000 indels according to the following categories: (1) 200 random indels from the intersection of all pipelines; (2) 200 random indels only found by HaplotypeCaller (v2.3.4); (3) 200 random indels only found by SOAPindel; (4) 200 random indels only found by Scalpel; (5) 200 random indels of size  $\geq 30$ bp from the union of all three algorithms. Due to possibly ambiguous representation, indels positions are “left-normalized”. However, some ambiguity can still remain, especially within microsatellites, so we computed validation rates using two different approaches. (1) *Position-based*: an indel is considered valid if a mutation with the same coordinate exists in the

validation data (**Figure 4.3a**). (2) *Exact-match*: an indel is considered valid if there is a mutation with the same coordinate and sequence in the validation data (**Figure 4.3b**).

As expected, indels detected by all pipelines have a high validation rate and their sizes follow a lognormal distribution. However, the validation rate varies dramatically for each tool. Respectively, only 22% and 55% of the HaplotypeCaller (2.4.3) and SOAPindel specific indels could be validated even when the less strict position-based approach was used, whereas 77% of Scalpel's specific indels are true positive. For the long indels: less than 10% called by SOAPindel and HaplotypeCaller passed validation. (**Figure 4.3c**). The new version of GATK (v3.0) has largely removed the bias towards deletions (**Figure 4.2b**), but find that Scalpel still outperforms HaplotypeCaller. Scalpel shows substantially higher validation rate (76%) for longer indels (>5bp) compared to HaplotypeCaller v3.0 (27%).

We further divide the results to separately report the validation rate for indels within microsatellites. SOAPindel shows an appreciably higher rate of false-positives within microsatellites ("SSRs-only" in **Figure 4.3a-b**). When microsatellites are excluded ("no-SSRs" in **Figure 4.3a-b**), the performance of SOAPindel and HaplotypeCaller decline, while Scalpel's validation rate is only slightly reduced. **Figure 4.3a** and **Figure 4.3b** also illustrate the relative abundance of indels within microsatellites called by each tool, although HaplotypeCaller seems to filter against these. Finally, when switching from position-based to exact-match, indels within microsatellites show notable reduction in validation rate. This phenomenon is due to their high instability and higher error rates, and in fact it is not unusual to have more than one candidate mutation at a microsatellite locus.

We further inspected the sequence composition of all false-positive long indels. Specifically, we reanalyzed the 129 SOAPindel invalid long mutations using Scalpel. The majority of these mutations (115) overlap repeat structures where the reference contains a perfect or near-perfect repeat. In contrast, of the 62 false-positive long indels from HaplotypeCaller, only 16 overlap a repeat. The remaining false positive deletions appear to be due to an aggressive approach used by the algorithm when processing soft-clipped sequences. The soft-clipped reads in false positive indels for HaplotypeCaller are highly variable, and are conjectured to be mapping artifacts of reads from different genomic locations. Finally we investigated the relationship between the false-discovery rate (FDR) and characteristic features (e.g., chi-square score and coverage) for 614 indels detected by Scalpel and validated by re-sequencing. In addition to highlighting the common trends, this analysis provides recommendation on how to select a chi-square score cutoff to achieve a given FDR.

Using Scalpel we detected a total of 3.3 million indels in exomes from 593 families from the Simons Simplex Collection, corresponding to an average of ~1,400 ( $=3,388,139/(4*593)$ ) mutations per individual. Accounting for population frequencies, there were 27,795 distinct transmitted indels across the exomes. We find close agreement to the size distributions reported by Montgomery *et al*<sup>69</sup> using low coverage whole-genome data from 179 individuals. Direct comparison to those detected by the GATK-UnifiedGenotyper based mapping pipeline used by Iossifov *et al*<sup>8</sup> shows that Scalpel has superior power to detect longer insertions. To estimate Scalpel's ability to discover transmitted mutations, we performed targeted re-sequencing of 31 long ( $\geq 29$ bp) transmitted indels. Excluding indels that failed to sequence (4), 21

passed validation (out of 27), which gives a 78% true positive rate. Three of the indels that did not pass validation were indeterminate with ambiguous alignments because they were either too long ( $\geq 70$ bp) or embedded in a repetitive region.

Within the coding sequence (CDS), frame-preserving indels are more abundant than frame-shifts. In agreement with MacArthur *et al*<sup>131</sup>, we detect a large number of transmitted loss-of-function (LOF) variants in protein-coding genes. Frame-shift mutations are found at lower frequency in the population when located in protein-coding sequences compared to intronic regions. Finally, we observe an enrichment of deletions over insertions, with an overall 2:1 ratio across all annotation categories. Similar trends were reported in previous studies<sup>14</sup>. Here we reanalyzed the data on autistic children and unaffected siblings with Scalpel with the goals of examining *de novo* likely gene disrupting (LGD) mutations. We confirm an overabundance of frame-shift mutations in autistic patients<sup>6</sup> predict additional candidates, and extend the analysis to a larger number of families. Our re-analysis of a previous study with 200 SSC families<sup>132</sup> reports an enrichment of 11 LGD indels in autistic children compared to 4 in their healthy siblings. In targeted re-sequencing of 102 candidate indels we confirmed 84 as *de novo* mutations, invalidated 11 and failed to sequence 7, giving an 88% *de novo* positive predictive rate. In order to focus the list of candidate genes, we excluded mutations that are common in the population, and used stringent coverage filters to select a total of 97 high quality *de novo* indels. Even after extending the population size from 343 to 593, the same 2:1 enrichment for LGD mutations is confirmed: 35 frame shifts in autistic children vs. 16 in siblings ( $p$ -value 0.01097), other smaller studies came to similar conclusions<sup>133, 134</sup> This result also holds for a larger collection of 1303 SSC families (not presented in this study). All together, in agreement with the previously reported results<sup>8</sup>, we find a significant overlap between the LGD target genes and the 842 FMRP-associated genes<sup>135</sup>. Specifically, 8 out of 35 LGDs in autistic children overlap with the 842 FMRP-associated genes.

## Figures and tables in this chapter

### Figures

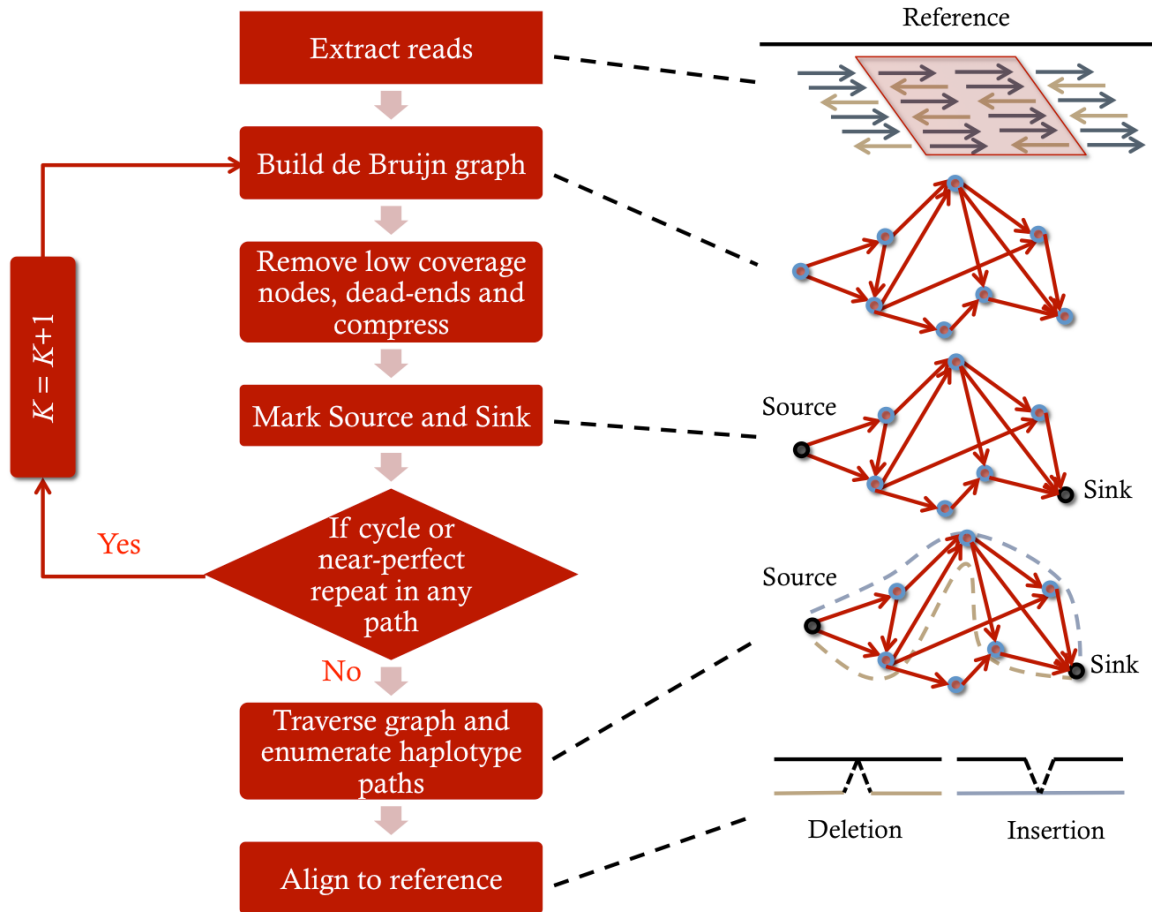
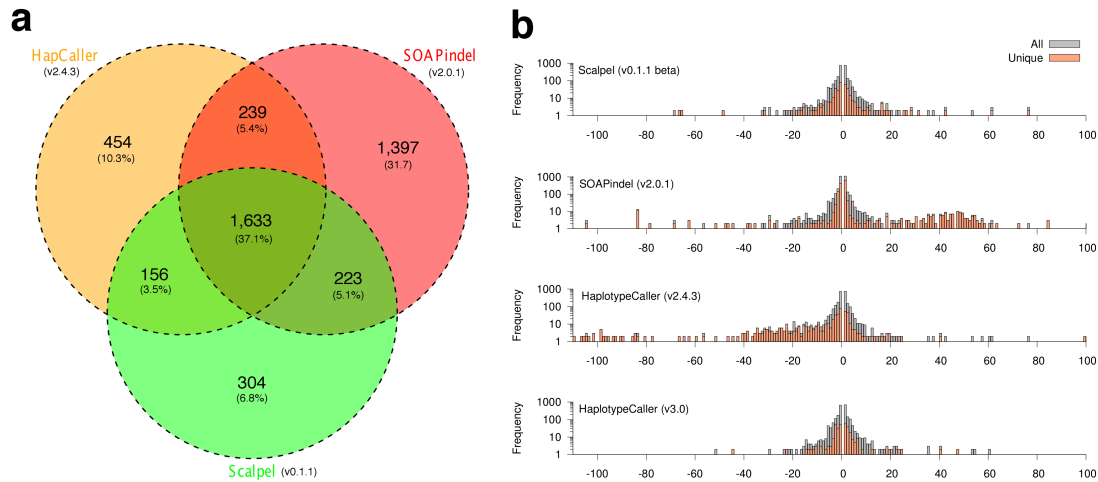
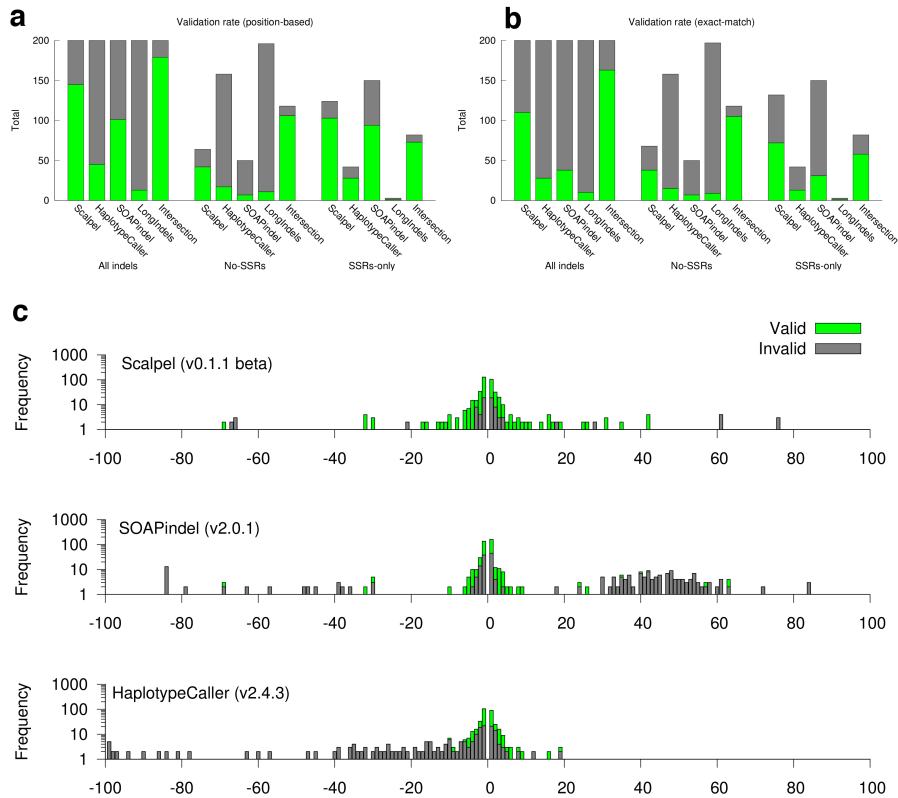


Figure 4.1. Overview of the Scalpel algorithm workflow. **Extracted reads include: well-mapped reads, soft-clipped reads, and reads that fail to map, but are anchored by their mate.** The assembled sequences are aligned to a reference using the standard Smith-Waterman-Gotoh alignment algorithm with affine gap penalties.



**Figure 4.2. Concordance of indels between pipelines.** (a) Venn Diagram showing the percentage of indels shared between the three pipelines. (b) Size distribution for indels called by each pipeline. 1,000 indels from five categories were analyzed by focused resequencing.



**Figure 4.3. MiSeq validation.** Ratio of valid (green) and false (grey) indel calls based on position-based matches (a) or exact matches (b) for the indicated tools, for indels of size  $\geq 30$ bp from the union of the mutations detected by all three pipelines (“LongIndels”), and for indels in the intersection (“Intersection”). Validation for all indels (“All indels”), validation only for indels within microsatellites (“SSRs-only”), and validation for indels that are not within microsatellites (“No-SSRs”). (c) Stacked histogram of validation rate by indel size for each variant caller.

# Chapter 5 Accurate inference and robust modelling of translation dynamics at codon resolution

## Summary of Contribution

This chapter describes the methods for analyzing Riboseq data. The model and results is in submission. Han Fang conceived the project, built the model, developed the software, performed the analysis. This chapter has been written in preparation for submission, but is unpublished at this time. Permission for republication of this material has been granted and is available upon request.

## Abstract

Ribosome profiling (Riboseq) is a powerful technique for measuring protein translation, however, sampling errors and biological biases are prevalent and poorly understand. Addressing these issues, we present Scikit-ribo (<https://github.com/hanfang/scikit-ribo>), the first open-source software for accurate genome-wide A-site prediction and translation efficiency (TE) estimation from Riboseq and RNAseq data. Scikit-ribo accurately identifies A-site locations and reproduces codon elongation rates using several digestion protocols ( $r = 0.99$ ). Next we show commonly used RPKM-derived TE estimation is prone to biases, especially for low-abundance genes. Scikit-ribo introduces a codon-level generalized linear model with ridge penalty that correctly estimates TE while accommodating variable codon elongation rates and mRNA secondary structure. This corrects the TE errors for over 2000 genes in *S. cerevisiae*, which we validate using mass spectrometry of protein abundances ( $r = 0.81$ ). From this, we determine the Kozak-like sequence directly from Riboseq and discover novel roles of the DEAD-box protein Dhh1p, deepening our understanding of translation control.

## Introduction

First introduced by Ingolia et al in 2009<sup>3</sup>, ribosome profiling (Riboseq) allows researchers to investigate genome-wide *in vivo* protein synthesis through deep sequencing of ribosome-protected mRNA footprints<sup>31</sup>. Since the original introduction, several improved versions have been developed to mitigate biases in the data<sup>136-138</sup> and address new biological questions<sup>139-141</sup>. After the protocol became standardized in 2012, there was a rapid increase in adoption<sup>5</sup>, leading to discoveries of new mechanisms involving translational defects in different forms of cancer<sup>32-35</sup>, other important human diseases<sup>36, 37</sup>, and the identification of novel drug targets<sup>38, 39</sup>. Riboseq has also revealed new insights into many steps in the translation process itself<sup>40, 41</sup>.

Riboseq provides genome-wide insights into the regulation of gene expression at the level of translation. A key metric of measuring translational control is translational efficiency (TE), defined as the level of protein production per mRNA<sup>3, 42</sup>. Assuming minimal ribosome fall-off, Li showed that TE is the same as translation initiation efficiency (TIE) in the steady state<sup>42</sup>. Shah *et al* showed that TIE is the rate limiting factor for translation<sup>142</sup>. In practice, this metric is



calculated for a given gene by taking the ratio of the ribosome density from Riboseq to the mRNA abundance measured by RNAseq. We refer to this ratio as RPKM-derived TE (ribosome density per mRNA, **Equation 5.1**), because both values have RPKM units, reads per kilobase of transcript per million mapped reads (**Equation 5.2**). Although this metric is commonly used in the Riboseq and RNAseq literature, it is not a direct measure of protein output but ribosome density, and the two are only correlated assuming the same elongation rate across genes<sup>42</sup>. However, this assumption does not hold in many cases, especially genes with extensive ribosome pausing<sup>143-147</sup>.

Technical shortcomings in the Riboseq workflow can introduce bias and systematic error into the analysis, masking the true ribosome density on an mRNA. Ribosome footprints come in many sizes depending on the organism, nuclease, and cell lysis conditions, making it difficult to identify the ribosome position on the fragment. Sampling only part of the footprint distribution can yield misleading results<sup>144</sup>. Another source of the noise in the data can be attributed to ligation bias in cloning ribosome footprints and amplification by PCR<sup>148</sup>. Finally, early protocols used antibiotics such as cycloheximide (CHX) to arrest translation prior to cell lysis; CHX treatment distorts ribosome profiles because initiation continues even though elongation is blocked<sup>138</sup>. This artifact leads to high levels of ribosome density at alternative initiation sites and the 5'-end of ORFs. CHX also masks the local translational landscape at the single-codon level<sup>149</sup>. Weinberg *et al* produced excellent quality reference datasets and showed that RNAseq libraries are subject to their own problems; isolation of mRNA through interaction with the poly-A tail leads to error in measuring mRNA abundance<sup>136</sup>. All of these problems confound the accurate determination of TE. Below, we summarize the major experimental and analytical challenges and proposed solutions to overcome them.

Analytically, it is first essential to correctly determine the location of the ribosome within the Riboseq reads, and particular, the location of the codon bound in the ribosomal A-site. Decoding of the A-site codon by incoming aminoacyl-tRNAs is rate limiting during elongation<sup>41</sup>; low levels of specific aminoacyl-tRNA species lead to pausing as indicated by changes in the codon-specific elongation rate (ER). Precise determination of the A-site codon of a Riboseq read is needed to determine whether a given read belongs to the canonical open reading frame (ORF) of a gene, especially when genes are overlapping. RiboDeblur<sup>150</sup> models ribosome profiles as blurred position signals, but it is not suitable for downstream analysis beyond finding the A-site. Most other studies followed the 15-nucleotide (nt) rule from Ingolia *et al*<sup>3</sup>, based on the work of Wolin and Walter<sup>151</sup>; the A-site codon starts at 15 nt in 28mer reads produced by RNase I. Reads of other lengths are commonly excluded from consideration, significantly reducing the data for downstream analysis, and perhaps missing important signals that affect footprint size. Correct identification of the ribosome position is particularly problematic in bacteria<sup>144, 152</sup> and *Arabidopsis*<sup>153</sup> where MNase generates a broad distribution of footprints<sup>152</sup>. Here, we introduce a novel method of finding the A-site codon that substantially improves the resolution of the downstream analysis.

Next, in almost every published Riboseq study, the distributions of RPKM-derived  $\log TE$  are severely skewed with a long tail on the negative side<sup>3, 154, 155</sup> (**Supplemental 5.S1A**). This observation is also reported by Weinberg *et al* in their analysis of wild-type *S. cerevisiae* data from ten different labs<sup>136</sup>. One of the main reasons for the skewed distribution is sampling

error from low-abundance genes: the range of gene expression level spans 8 to 11 orders of magnitude, but a limited amount of sequencing coverage is available. As a result, the sampling of low-abundance transcripts is more error-prone (**Figure 5.1A**), yielding higher dispersion of RPKM among low-abundance genes, and subsequently even higher dispersion of RPKM-derived TE (**Figure 5.1A**). To address this same problem in analyses of RNAseq data, fold change shrinkage methods (e.g. empirical Bayesian shrinkage) have been widely adapted in differential expression (DE) methods such as DEseq2<sup>156</sup>, edgeR<sup>157</sup>, and Slueth<sup>158</sup>. In order to perform shrinkage with between-sample normalization, however, these methods rely on at least three replicates, which are not typically available in Riboseq studies. Even where multiple replicates are available, it is not appropriate to use RNAseq DE methods to compute TE, because those methods were developed to estimate changes of gene expression under perturbation, while TE reflects the level of translation control under a single condition<sup>159, 160</sup>. To overcome this limitation, we developed a robust model for estimating TE using a shrinkage method that is compatible with a single library of Riboseq data.

Finally, traditional techniques for mRNA quantification and DE testing rely on a strong assumption: random fragmentation and uniform sequencing of mRNA molecules. However, this assumption does not apply to Riboseq data, given that the abundance of ribosome-protected fragments is strongly influenced by local translational elongation rates. In fact, peaks due to paused ribosomes (**Figure 5.1B**) have been observed in the literature<sup>143, 161, 162</sup>. Two major determinants of ribosome pausing are slow codons<sup>163</sup> and downstream mRNA secondary structure<sup>164</sup> (**Figure 5.1B**), although their importance and relative contributions have been controversial in Riboseq studies<sup>144, 165-167</sup>. The presence of paused ribosomes problematizes the use of ribosome density for calculating TE<sup>146</sup> (**Figure 5.1C**). Genes with paused ribosomes have more reads than expected, depleting coverage on other genes. Traditional read counting methods do not control for these biases (when using RPKM to derive TE). In contrast, our proposed method correctly estimates TE while accounting for biological biases simultaneously, enabling us to separate out the effects of translation initiation and elongation.

There were earlier attempts to model TE that are relevant for this work, although the published methods have significant restrictions and have seen limited application so far. Pop *et al* developed a queuing model for translation, but it failed to recover significant correlation between codon dwell time and cognate tRNA availability, and the source code is not publicly available<sup>168</sup>. Weinberg *et al* proposed a comprehensive model to estimate TE<sup>136</sup> in *S. cerevisiae* (budding yeast) using the analytical approximations of initiation probability, but this required parameterizations from a whole-cell simulation from Shah *et al*<sup>142</sup>, making it difficult to apply to other organisms. Duc and Song developed a simulation-based inference algorithm to estimate translation initiation and local elongation rates, but it could only be applied to ~900 (13%) genes in *S. cerevisiae*, because it requires filtering genes by length and coverage<sup>169</sup>. None of these methods addressed the prevalent sampling errors and biological biases in Riboseq data described above.

Here, we present Scikit-ribo, the first statistical model and open-source software package for accurate genome-wide TE inference from Riboseq data (**Figure 5.2**). The software is written in python and is freely available at <https://github.com/hanfang/scikit-ribo>. Scikit-ribo is very fast; it can analyze more than 6000 genes from a high-coverage *S. cerevisiae* Riboseq data (over 75

million reads) in less than one hour with single-codon resolution. It can accurately infer A-site codons with a variety of different mRNA digestion methods. We applied it to 10 Riboseq data sets and demonstrated its robustness to low-abundance genes while automatically correcting biases across different genes. We next show that the commonly used RPKM-derived TE is very sensitive to sampling errors and biological biases, creating substantial discrepancies and skewing the values of this key metric in previous studies. To address this, we developed a codon-level generalized linear model (GLM) with a ridge penalty to shrink the TE estimates. The GLM also serves as a mechanistic model for translation elongation and initiation, incorporating codon-specific elongation rates, local mRNA secondary structure, and gene-specific translational initiation efficiencies. We validate the model using *in silico* analysis as well as large-scale experimental mass spectrometry data and show a very high correlation in predicted protein abundance ( $r=0.81$ ). This successfully corrects the biases for ~2000 genes, and resolves the negative skew in TE observed in previous studies of Riboseq data. Finally, we show the importance of accurate TE estimation for interpreting Riboseq data. Our refined TE analysis using Scikit-ribo helped recover the Kozak-like consensus sequence in *S. cerevisiae* and reveal novel roles of the DEAD-box protein Dhh1p<sup>145</sup>. Together, these results showed that Scikit-ribo substantially improves Riboseq analysis and deepen the understanding of translation control.

## Results

### Accurate A-site codon prediction with different organisms and nuclease digestion

Using a supervised learning approach, Scikit-ribo trains a model for identifying the A-site codon within Riboseq data using reads that contain start codons (**Figure 5.2A**). Briefly, the algorithm uses a random forest model to evaluate eight features of how the Riboseq reads align to the genome: the length of the read, the distance from the 5' or 3' end of the read to the start codon, and the nucleotides flanking the ends of the Riboseq reads (**Methods**). Unlike other methods, Scikit-ribo can easily accommodate different types of Riboseq data because of its recursive feature selection technique. For a given dataset, Scikit-ribo uses cross validation (CV) to find the optimal features with the lowest prediction error. This is an effective way to remove irrelevant features for the given data and avoids overfitting an unnecessarily complex model.

Using this approach on the *S. cerevisiae* data prepared with RNase I by Weinberg *et al*, the accuracy of the prediction of the A-site codon was extremely high (mean accuracy=0.98, SD=0.003, 10-fold CV)<sup>136</sup>. Unlike the basic 15-nt rule, our model's predictions are consistent across reads with different lengths or A-site locations, as demonstrated using the multi-class ROC curves (**Supplemental Figure 5.S2A**). This means that we can utilize the full complement of reads for downstream analysis; this is especially helpful for low-abundance genes. Our model also achieved very high accuracies in seven other *S. cerevisiae* datasets (**Supplemental Table 5.S1**). Interestingly, for all eight *S. cerevisiae* datasets the most important features learned were the phase of the 5'-end of a read (whether it falls in the first, second, or third frame) and the read length (**Supplemental Figure 5.S3A**). This is consistent with the previous findings that RNase I was not always precise in generating ribosome footprints<sup>137</sup>. When we look at elongating ribosomes within the canonical ORF (not overlapping the start codon), 94.3% of the predicted A-sites are in the correct frame, confirming Scikit-ribo's very high accuracy.

To test whether Scikit-ribo can maintain high accuracy in different model organisms or with different nuclease digestions protocols, we next applied it to the Riboseq data from *E. coli*. Bacterial ribosome profiling protocols use MNase instead of RNase I because as an *E. coli* protein, RNase I is inhibited by bacterial ribosomes. The resulting read distributions are broad and have posed challenges in assigning ribosome position<sup>162, 170</sup>. One promising approach is to employ MNase together with the endonuclease RelE, taking advantage of RelE's ability to cleave the A-site codon within the ribosome with high precision. In the resulting ribosome footprints, the A-site codon is found at the 3'-end of reads, rather than 12 to 18 nt away from the 5'-end of a read as in *S. cerevisiae*. In spite of these differences, the accuracy of Scikit-ribo on the *E. coli* data generated with RelE was still very high (mean accuracy=0.91, SD=0.041, 10-fold CV, **Supplemental Figure 5.S3B**) and showed 99.8% assignment of the A-site codon to canonical ORFs for reads not overlapping the start codons. Interestingly, for the RelE data, the optimal feature was the phase of 3'-end of a read, while the 5'-end did not have a strong effect (**Supplemental Figure 5.S3B**). This is consistent with the report in Hwang *et al* that RelE preferentially cleaves at the ribosome A-site codon, generating precise 3'-ends<sup>152</sup>. Using Scikit-ribo, we also analyzed *E. coli* Riboseq libraries prepared with MNase alone, but the accuracy was much lower (0.70) than observed in libraries prepared with RelE. This indicates that RelE improves the precision of the ribosome sub-codon position and thus is a better nuclease for analyses requiring codon resolution.

### Paused ribosomes and biological biases of TE

Ribosome pausing (RP) events are prevalent in several different model organisms<sup>143</sup>. Pausing can occur for a number of reasons, including slow recruitment of tRNAs and mRNA secondary structure<sup>167</sup>. These biological effects can introduce biases in ribosome profiles on different genes, leading to overestimation of TE in genes with high levels of pausing. In Weinberg *et al*<sup>136</sup>, the distribution of RPKM-derived  $\log_2 TE$  is negatively skewed with a mean of -0.5 (**Supplemental Figure 5.S1B**), although this is likely an artifact of RPKM-derived TE. We hypothesized that the distribution of RPKM-derived TE was largely skewed due to RP events. To illustrate this, we simulated both Riboseq and RNAseq data, with and without paused ribosomes in *S. cerevisiae* (**Methods**). Upon comparing  $\log_2 TE_{RP}$  (i.e. the  $\log_2 TE$  in the data with RP) with  $\log_2 TE_{Baseline}$  (i.e. the  $\log_2 TE$  in the data without RP), we observed that several genes had inflated TEs, while the remaining majority had decreased estimates. We also observed that the  $\log_2 TE_{RP}$  distribution for paused data became broader and negatively skewed, similar to what has been observed in previous reports. These results suggest the possibility that this skew arises from the fact that genes with significant pausing will have more Riboseq reads and higher RPKM-derived TE, although their protein abundance remains the same. Pausing on these genes also reduces the available Riboseq reads available on other non-paused genes, so that their TE estimates of those genes are deflated.

Since pauses can be induced by non-optimal codons and downstream mRNA secondary structure<sup>167</sup>, we developed a statistical model to jointly correct for these effects that we refer to as biological biases. Since the observed ribosome profiles are affected by changes in elongation rates, and not simply initiation rates, Scikit-ribo uses a codon-level generalized linear model (GLM) to separate out these two processes, considering three categorical covariates and one continuous covariate (**Methods, Equation 5.5-6**). The general model to explain the data is that at a codon position, the ribosome coverage is proportional to mRNA abundance and gene specific

TE, reflecting initiation levels, as well as downstream mRNA secondary structure and codon specific dwell time, reflecting limiting steps in elongation rates (**Figure 5.2B**).

### Sampling errors for low abundance genes using Riboseq

Another difficulty in estimating TE is caused by sampling error for low-abundance genes due to lack of depth in the sequencing data. Similar trends have been reported in DE analysis of RNAseq data, where low abundance genes can have extreme fold changes if not corrected for dispersion<sup>156</sup>. This is a side-effect of modeling high-dispersion count data; measurements are inherently noisier when counts are low<sup>156</sup>. Riboseq data shares the same issue. Since most of the Riboseq experiments are done in two or fewer replicates, estimation of between-sample variability and subsequent shrinkage of dispersion has not been feasible<sup>159</sup>. Thus, most published Riboseq studies used the RPKM-derived TE:  $RPKM^{Ribo} / RPKM^{mRNA}$  (**Equation 5.1**)<sup>3</sup>.

However, low abundance genes, especially those with a “transcripts per million” (TPM, **Equation 5.3**) value less than one, tend to show much more dispersed TE values, compared with other genes (**Figure 5.1A**). This is true even if the TPM cutoff is increased to 10 (**Supplemental Figure 5.S1D**). Consequently, the standard deviation (SD) of  $\log_2 TE$  in low abundance genes from the Weinberg *et al*<sup>136</sup> data was 3-fold higher than for other genes (Levene test, p-value= $3 \times 10^{-89}$ ), the overall range in TE was 5-fold larger (99 vs 20), and the median absolute deviation (MAD) was also larger (1.9 vs 1.0). In fact, the high dispersion of TEs was driven by the high variance of the ratio between the numbers of reads per gene (**Equation 5.4**).

One ad-hoc solution is to remove low abundance genes from downstream analysis, although this is not very effective as the chosen threshold is arbitrary and cannot be determined rigorously. Furthermore, this filtering approach reduces the sensitivity of finding genuinely extreme TE genes and reduces the power of finding significance. Instead of imposing arbitrary thresholds, Scikit-ribo uses a shrinkage method based on ridge penalty to account for the sampling uncertainty for low abundance genes (**Methods, Equation 5.7-8**). This method helps address the sampling errors issues even without having replicates. As a result, Scikit-ribo reports balanced  $\log_2 TE$  distributions while the distributions of RPKM-derived  $\log_2 TE$  are negatively skewed (**Supplemental Figure 5.S1**).

### Accurate inference reveals the interplay between cognate tRNA availability and mRNA secondary structure

Having described how Scikit-ribo addressed the errors and biases, we asked whether it can reveal new aspects of biology that were not detectable using previous methods. To investigate whether the biological covariates from Scikit-ribo were meaningful, we analyzed the CHX-free *S. cerevisiae* Riboseq data from Weinberg *et al*<sup>136</sup>. The codon dwell time (DT) estimates from the GLM are the inverse of the codon elongation rates (ER). Scikit-ribo almost perfectly reproduced the codon DT (Pearson  $r = 0.99$ ) from Weinberg *et al*<sup>136</sup>, in which the three slowest codons are CGG, CGA, and CCG (**Figure 5.3A**). The tRNA adaptation index (tAI) measures the efficiency of a coding sequence recognized by the intra-cellular tRNA pool, taking into account each gene’s codon compositions, mRNA expression levels, and the availability of the conjugate tRNA<sup>171</sup>. Reis *et al*<sup>171</sup> estimates tAI by taking the geometric mean of its codons’ relative adaptiveness value (RAV). A codon with lower RAV means that it is sub-optimal for translation elongation, i.e. slower codon. We found CGG, CGA, and CCG have very low RAV values<sup>171</sup> and are among the rarest codons in the *S. cerevisiae* transcriptome. Following

Weinberg *et al* and others<sup>136, 143, 167, 169</sup>, we compared the relative codon ERs with RAV and their cognate tRNA abundance (measured by microarray<sup>136</sup>), and reproduced a positive correlation against both (Spearman  $\rho_{tAI} = 0.54$ ,  $\rho_{tRNA} = 0.47$ , **Figure 5.3B-C**).

Although our findings confirm that ribosomes have lower DT on codons with higher cognate tRNA levels, it still cannot solely explain the variation in ER given the imperfect correlation. Consequently, we tested whether part of the missing contribution was from downstream mRNA secondary structure. We adjusted the within-gene ribosome densities by the inferred codon ERs, which controlled for the codon-specific effects on local translational elongation. We used RNAfold<sup>172</sup> to predict the optimal mRNA secondary structure and test if large downstream stem-loops would increase ribosome density (**Methods**). We found that the ribosomes move slower with the presence of a downstream mRNA stem-loops (t-test, p-value=  $5 \times 10^{-3}$ ). We computed the average adjusted ribosome density in a five-codon sliding window and notice a peak right at the junction (**Figure 5.3D**). This finding is consistent with previous reports that downstream stem-loops decrease the ribosome ER, i.e. increase the DT as ribosomes wait for the downstream stem-loops to be unfolded<sup>165, 173, 174</sup>. Taken together, our analyses show that ribosome elongation rates are affected by a complex interplay of cognate tRNA availability and downstream mRNA secondary structure. These results also confirm that Scikit-ribo accurately estimates codon-specific DT and the effect of mRNA secondary structure, after it correctly predicted the A-site codon and fit the GLM.

### Simultaneously correcting sampling errors and biological biases for TEs

To understand how Scikit-ribo corrects the biases in the Riboseq analysis, we compared the Scikit-ribo  $\log_2 TE$  with the RPKM-derived  $\log_2 TE$  from the Weinberg *et al* data (**Figure 5.4A**). The correlation between the estimates was high ( $r=0.82$ ), but the RPKM-derived TE estimates showed clear trends of systematic biases (negative skew) that were successfully corrected by Scikit-ribo (**Figure 5.4B**). We calculated the differences between the two estimates,  $\Delta \log_2 TE = \log_2 TE_{scikit-ribo} - \log_2 TE_{RPKM}$ , and colored them with respect to the values: 1)  $\Delta \log_2 TE > 0.5$ , previously underestimated (green), 2)  $\Delta \log_2 TE < -0.5$ , previously overestimated (orange), and 3) other genes in between (gray) (**Supplemental Table 5.S2**). The green points in the left half of the plot shifted upward from the diagonal line, while the points in the right half were more consistent (**Figure 5.4A**). There were 1957 genes with large differences ( $|\Delta \log_2 TE| > 0.5$ ); 897 being under-estimated and 1060 being over-estimated. Compared with RPKM-derived TE, we found the  $\log_2 TE$  of some genes were previously underestimated by as much as 11 (2048 fold), while other genes were overestimated by almost 3 (8 fold) (**Supplemental Figure 5.S4B**).

We further defined six regions based on  $\Delta \log_2 TE$  and the sign of Scikit-ribo  $\log_2 TE$ . For example, region 1 corresponds to genes with  $\Delta \log_2 TE$  greater than 0.5 with negative Scikit-ribo  $\log_2 TE$  (n=629); most of these genes were of low abundance with a TPM less than 10 (**Figure 5.4C**, **Supplemental Figure 5.S4**). This means given 75 million reads, these genes had fewer than 750 reads on average, i.e.  $\sim 2$  reads per codon. The sampling of such genes is highly unstable, causing the ratio of the read counts to have even higher variance. As a result, the RPKM-derived TE reports a very high dispersion and incorrect TE estimates in region 1, while Scikit-ribo successfully corrected the sampling errors by leveraging the power of shrinkage estimates.



While improvements in TE estimates in region 1 arise from a better treatment of sampling error on low abundance genes, how can we address differences in regions with more highly expressed genes? For this part of the analysis, we excluded low abundance genes with TPM less than 10 to focus on the effects on biological covariates, codon specific ER and mRNA structure. There were 268 and 981 genes in the highly-translated regions 4 and region 6, respectively. If downstream mRNA secondary structure had an effect, one would expect the RPKM-derived  $\log_2 TE$  of genes with high levels of structure would be inflated as additional ribosomes are paused at the loop; the  $\Delta \log_2 TE$  becomes smaller with a higher stem loop density (normalized by ORF length). We found this was indeed the case: there is a negative correlation between  $\Delta \log_2 TE$  and stem loop density (**Figure 5.4D**, Spearman  $\rho = -0.33$ ). This bias was automatically adjusted by the mRNA secondary structure covariate of the Scikit-ribo GLM as we found enrichment of 15% more ribosome density when there was a downstream secondary structure.

Second, we investigated the influences of variation in codon-specific ER values. The gene level tRNA-adaptation index (tAI) indicates whether a gene is enriched for optimal or non-optimal codons: higher tAI means the gene is enriched for faster codons, while a lower tAI means the gene is enriched for slower codons. The middle regions (gray), 2 and 5, served as baseline for genes with negative and positive  $\log_2 TE$ , respectively (**Figure 5.4E**). For negative  $\log_2 TE$  genes, there were no significant difference of tAI between genes in the region 1 and 2, but the region 3 genes had significantly lower tAI than those in region 2 (**Supplemental Table 5.S2**, t-test, p-value= $2 \times 10^{-6}$ ). We conclude that the differences in TE for region 1 between RPKM-derived TE and our TE estimates is not due to tAI but is instead due to the shrinkage estimates via the ridge penalty of the Scikit-ribo model. In contrast, the TE values of region 3 genes were previously overestimated because they contained more non-optimal/slow codons. When  $\log_2 TE$  is positive, tAI values have a stronger effect: region 4 genes had much higher tAI values than region 5 genes (t-test, p-value= $1 \times 10^{-17}$ ) while genes in region 6 had lower tAI (t-test, p-value= $5 \times 10^{-55}$ ). This means the genes in the region 4 and 6 were previously underestimated and overestimated, respectively, because their genes tend to enrich for fast and slow codons.

We further found the region 4 genes are enriched for the biological process of cytoplasmic translation [GO:0002181] (**Supplemental Table 5.S3**, p-value= $3 \times 10^{-25}$ ). Genes encoding ribosomal proteins are enriched for optimal codons and genes with more optimal codons are preferentially translated<sup>175</sup>. Since ribosomes move faster on mRNAs encoding ribosome proteins, RPKM-derived TE values are underestimated for these genes and corrected by Scikit-ribo. These observations do not depend on the use of the tAI metric that is based on gene expression data (including ribosome proteins: the same conclusion holds true using the species-specific tAI (stAI)<sup>176</sup> metric developed to provide a similar measurement of codon efficiency without using gene expression data (**Supplemental Figure 5.S5**).

#### Scikit-ribo discovers Kozak-like consensus in *S. cerevisiae*

The Kozak consensus sequence, GCCRCCATGG, promotes translation initiation in vertebrates<sup>177</sup>. In *S. cerevisiae*, the Kozak-like sequence was shown to be AAAAAAATGTCT<sup>178</sup>, and it has been widely used as a positive control to train translation initiation start (TIS) site prediction methods<sup>140, 179, 180</sup>. The Kozak sequence has been re-discovered in Riboseq studies in

humans (*homo sapiens*), mice (*Mus musculus*) and maize (*Zea mays*)<sup>181-183</sup>. However, no clear signal of Kozak-like sequences in *S. cerevisiae* has been found using Riboseq data, only a very weak resemblance of the Kozak-like sequence (4 out of 12 bases) was reported by Pop *et al*<sup>168</sup>. Thus, we were interested in whether the improved TE estimates from Scikit-ribo can help re-discover this mRNA element associated with high TE.

We collected the 5'UTR sequences from genes with  $\log_2 TE > 2$ , and scanned for enriched sequences using HOMER<sup>184</sup>. Based on HOMER's suggested p-value threshold, there were two statistically significant sequences. Strikingly, the top hit exactly matched the Kozak-like sequence from Hamilton *et al*<sup>178</sup>, AAAATGTCT (p-value= $1 \times 10^{-21}$ , **Figure 5.4F**). This is the first report of the identical Kozak-like sequence in the *S. cerevisiae* Riboseq analyses. The other enriched sequence was AAATAAGCTCCC, which has never been reported *in vivo* (p-value= $1 \times 10^{-11}$ , **Supplemental Figure 5.S6**). Interestingly, this sequence contains the motif ATAAG, one of the top five sequences that leads to higher TE in a large-scale *HIS3* reporter assay from Cuperus *et al*<sup>185</sup>. In contrast, using the same threshold, RPKM-derived TE failed to discover either of these Kozak-like sequences. Instead, it only found a weak signal of CAACATGGCT with a much less significant p-value ( $1 \times 10^{-11}$ ) and weak resemblance to the Kozak-like sequence (**Supplemental Figure 5.S6**). This failure of RPKM-derived TE to yield the Kozak-like motif was likely because that approach provided skewed estimates where some lower TE genes had artificially high RPKM-derived TE. This therefore contaminated the gene set for enrichment analysis, and reduced the ability to find motifs with high statistical significance.

Large-scale validation showed Scikit-ribo's accurate TE estimation, especially for low-abundance genes

To further understand the discrepancies between Scikit-ribo and RPKM-derived TE, we performed a large-scale validation using the selected reaction monitoring (SRM) mass spectrometry data from a recent reference proteome dataset containing high quality measurements of about 1,800 gene in *S. cerevisiae*<sup>186</sup>. Based on the master equations relating mRNA transcription and protein translation (**Equation 5.9**)<sup>42</sup>, the relative protein abundance (PA) is proportional to the product of mRNA abundance and TE, assuming a consistent protein degradation rate across genes (**Equation 5.10**). There were 1,180 genes in the validation set, with a mean of 55,012 copies per cell, ranging from 6 to 4,366,751. The correlation between the protein abundance derived by Scikit-ribo and derived by mass spectrometry was indeed very high (Pearson  $r = 0.81$ , **Figure 5.5A**) and the fitted line was close to the diagonal (linear regression,  $\beta = 0.83$ ). When we further considered protein degradation rates from Christiano *et al*<sup>187</sup>, the correlation became even higher (Pearson  $r = 0.83$ , **Supplemental Figure 5.S8**). In comparison, RPKM-derived  $\log PA$  reported a lower correlation (Pearson  $r = 0.77$ ) and the fitted line is more distant from the diagonal ( $\beta = 0.75$ , **Figure 5.5C**). In addition, many of the outliers in the RPKM-derived PA were low abundance genes, suggesting a systematic bias (**Figure 5.5C**). Focusing on a set of 933 low abundance genes with a TPM less than 100, the Scikit-ribo derived  $\log PA$  maintained a high correlation with mass spectrometry derived  $\log PA$  (Pearson  $r = 0.6$ ,  $\beta = 0.48$ , **Figure 5.5B**). In contrast, RPKM-derived PA became more inaccurate with a much lower correlation (Pearson  $r = 0.35$ ,  $\beta = 0.29$ , **Figure 5.5D**). This analysis demonstrates that Scikit-ribo more accurately estimates genome-wide TE regardless of



mRNA abundance, while the RPKM-derived TE performed poorly among low abundance mRNAs.

### Refined TE analysis revealed Dhh1p's role in translation repression

The DEAD-box protein Dhh1p is a sensor for codon optimality and ribosome speed, targeting an mRNA for repression and subsequent decay<sup>145</sup>. Radhakrishnan et al performed ribosome profiling in three *S. cerevisiae* strains: wild-type (WT), *dhh1Δ* (KO), and overexpressed (OE) Dhh1p<sup>145</sup> with substantial differences in TE between the strains (**Supplemental Figure 5.S9**). Here, we re-analyze their data to make use of Scikit-ribo's more refined analysis to reproduce major findings and to yield new biological insights into Dhh1p's activity.

First, regarding reproducibility, the mean correlation of  $\log TE$  and the codon DT were all very high between the biological replicates for a given strain ( $\bar{r}_{te} = 0.95$ ,  $\bar{r}_{dt} = 0.99$ , **Supplemental Figure 5.S10-S11**), indicating that the data are of high quality and that the inference procedures in Scikit-ribo are stable. When comparing codon DTs between different strains, we observe OE and KO have the largest and smallest standard deviation, respectively (**Supplemental Figure 5.S11, 5.S12A-C**). This is consistent with Radhakrishnan *et al*<sup>145</sup>. They also showed a pattern of increased ribosome density per mRNA on non-optimal genes in the OE strain<sup>145</sup>, which was successfully reproduced by Scikit-ribo as well (**Supplemental Figure 5.S13**). Compared with WT, codon-optimal genes (higher tAI) had enhanced TE in KO, while non-optimal genes had much lower TE (**Supplemental Figure 5.S13A**). Overall, whenever Dhh1p was overexpressed, codon-optimal genes exhibited reduced TE (**Supplemental Figure 5.S13B**), which became even more distinct when comparing OE with KO (**Supplemental Figure 5.S13C**).

We next refined the estimation of the codon DT differences between strain using the log ratios of DTs in OE and KO relative to those in WT. A lower log ratio indicates the codon becomes faster, and a higher ratio indicates the codon becomes slower. In Radhakrishnan *et al*<sup>145</sup>, the AGG codon was an outlier and had large differences, although it is an optimal codon. In our analysis, it only had minimal differences (log ratio=-0.03, **Supplemental Figure 5.S12B, Table 5.S4**). The two slowest codons in WT (CCC, CCG) had the most changes in DT, which became much faster in OE (log ratio=-0.68, -0.50, **Supplemental Table 5.S4, Figure 5.6F**). Radhakrishnan *et al*<sup>145</sup> showed that Dhh1p stimulated the degradation of low codon optimality mRNAs and increased of their ribosome densities per mRNA, meaning the number of non-optimal codons decreased in OE while the amount of tRNA availability stayed unchanged. Thus, the pairing of non-optimal codons became more efficient and these codons elongated faster. We also sorted the 61-sense codons by their DT in WT, and discovered a strong negative correlation against the log ratios (*Spearman*  $\rho = -0.63$ , **Supplemental Figure 5.S12D**). This means slower codons in WT reported larger changes of DT in OE. The findings of codon DT differences are particularly interesting for Scikit-ribo because its GLM infers codon DTs directly from the data, without the need of pre-defined parameters.

Finally, genes with large changes in TEs ( $\Delta \log_2 TE$ ), might provide insights about Dhh1p's role in translation regulation. We examined this with a conservative approach, focused on the two extreme tails, and compare results from Scikit-ribo with RPKM-derived TE

**(Methods).** Using both methods, we found genes with reduced TE in OE enriched for optimal codons (t-test,  $p\text{-value}=1\times 10^{-51}$ , **Figure 5.6B**). This set of genes is significantly enriched for the GO categories “cytosolic ribosome” (GO:0022626,  $p\text{-value}=3\times 10^{-16}$ ) and “cytosolic small ribosomal subunit” (GO:0022627,  $p\text{-value}=6\times 10^{-11}$ ) (**Supplemental Table 5.S5**). Dhh1p is a known mRNA translation repressor<sup>188-190</sup> and associates with the eukaryotic ribosome<sup>145</sup>. Here, we further speculate that Dhh1p might reduce translation of the 40S ribosomal subunit mRNA, in addition to inhibiting the production of a stable 48S preinitiation complex to form on mRNA<sup>191</sup>. In contrast, genes with increased TE in KO tend to be codon sub-optimal (t-test,  $p\text{-value}=3\times 10^{-52}$ , **Figure 5.6A**). Found with Scikit-ribo but not with RPKM-derived TE, these genes are enriched for “inner mitochondrial membrane protein complex” (GO:0098800,  $p\text{-value}=3\times 10^{-4}$ , **Supplemental Table 5.S6**). To investigate the GO enrichments that are specific to Scikit-ribo, we selected the tail genes that correspond to the significant GO categories (**Methods**). Among these genes, we only kept the ones specific to Scikit-ribo (not found with RPKM-derived TE), and we again observed the same patterns with respect to codon optimality and Dhh1p expression (**Figure 5.6C-E**). This means the Scikit-ribo-specific genes are consistent with the global patterns, thus strengthening our understanding of Dhh1p’s role in translation.

## Discussion

For nearly 60 years, the central dogma of molecular biology has been the guiding model for explaining how genetic information flows from DNA to RNA and then to proteins. Through widespread genome and transcriptome sequencing, the first half of this process has been extensively explored, revealing many important relationships between genomic sequences, gene expression, and gene regulation in evolution, development, and disease. In contrast, relatively little is known about the final phases of this process, largely because of the difficulties in acquiring high throughput and high quality data about translation and translational control. Riboseq is a powerful approach poised to fill this void. Several methods have been developed for selected aspects of Riboseq analysis, including differential TE testing<sup>192-195</sup>, identifying ORFs and alternative translation initiation sites<sup>196, 197</sup>, and predicting the shape of ribosome profiles<sup>198</sup>. But few practical statistical methods have been developed for robust TE estimation and most previous analyses were not performed in a systematic fashion. This had led to conflicted findings about the roles of codons and mRNA secondary structure on translation, and has prevented biological discoveries from being made in some cases. Here, through a systematical characterization and validation using mass spectrometry data, we exposed some of the more troubling issues of RPKM-derived TEs, including sampling errors and biological biases, especially for the low abundance genes.

We argue that Scikit-ribo is the first statistically robust model and open-source software package for accurate genome-wide TE inference from Riboseq data. The core of Scikit-ribo is a codon-level generalized linear model that unifies our study of translation elongation and initiation including the effects of codon specific elongation rates, mRNA secondary structure, and gene specific translation initiation efficiency. When paired with a powerful ridge regression regularization method, Scikit-ribo corrects the negative skew in TE observed in most previous papers, especially for low expressed genes. Using three case studies involving ten different datasets, we showed how these statistical advancements allow universal improvement to Riboseq data analysis. This particularly improves the estimation of genome-wide TE, allowing us to discover the Kozak-like consensus sequence in *S. cerevisiae*, and yield novel insights into

Dhh1p's role on translation repression. Our findings showcase the interplay between biology and statistics; biological knowledge informs statistical methods development, and statistical improvement yields novel biological insights. Together, we demonstrate that Scikit-ribo substantially improves Riboseq analysis and our understandings of translation control. In the future, we foresee more researchers applying Riboseq to address their biological questions related to protein translation and Scikit-ribo can unlock the full potential of this technique.

## Methods

### Overview of Scikit-ribo

Scikit-ribo has two major modules (**Figure 5.2**): (1) Ribosome A-site codon location prediction, and (2) TE inference using a codon-level generalized linear model (GLM) with ridge penalty. A complete analysis with Scikit-ribo involves two steps: 1) data pre-processing to prepare the ORFs and codons for a genome of interest, 2) the actual model training and fitting. The few inputs to Scikit-ribo includes the alignments of Riboseq reads (i.e. BAM file), gene-level quantification of RNAseq reads (i.e. from Salmon<sup>199</sup> and Kallisto<sup>200</sup>), a gene annotation file (i.e. gtf file) and a reference genome (i.e. fasta file) for the model organism of interest. The main outputs include  $\log_2 TE$  estimates for the genes, and the translation elongation rates for the 61-sense codons. Scikit-ribo also has modules to automatically produce diagnostic plots of the random forest model and the GLM. The ribosome profile plots for each gene can also be plotted using Scikit-ribo. For details of preparing the inputs, see data processing steps in Methods. For a complete workflow from raw sequencing reads to results, see **Supplemental Figure 5.S15**. Scikit-ribo can be easily installed with a single command: "`pip install scikit-ribo`". The documentation of Scikit-ribo is available at <http://scikit-ribo.readthedocs.io/>.

### Ribosome A-site codon prediction

Scikit-ribo uses a random forest<sup>201</sup> classifier from Scikit-learn<sup>202</sup> to predict the ribosome A-site locations over the 61-sense codons in the ORFs after excluding the start and stop codons. (**Figure 5.2A**). Low mapping quality (MAPQ<20) and clipped alignments are removed from downstream analysis. After filtering out overlapping genes, it collects all reads that intersect the start codons as training data. In the Weinberg *et al* data, the sample size of the training data is ~700,000, with ~85,00 in each class. The feature set of the classifier include 1) read length, 2) reading frame phase of the 5'-end and 3'-end nucleotides (1st, 2nd, or 3rd), 3) the edge and the flanking nucleotides of the Riboseq reads. In the RNase I data, the label of the training data is the distance between the 3'-end of the start codon and the 5'-end of the read. In the ReIE data, the label of the training data is the distance between the 3'-end of the start codon and the 3'-end of the read, which is enabled by the flag `-r` of the Scikit-ribo program.

The training of the random forest classifier involved two steps: recursive feature selection with CV, and training the classifier with reduced feature set. The first step of the training uses CV to find the optimal features that gives the lowest prediction error. During each step of the CV, the features are re-ranked and the lowest ranked feature is dropped. This is similar to finding the "elbow" point in the feature importance plot (**Supplemental Figure 5.S3**), which indicates the last sharp decrease of feature importance. Once the optimal feature set is selected, Scikit-ribo performs another ten-fold CV to measure the accuracy (1 - error rate) of the model and learns the

weights for each feature. After this, the learned classifier is applied to all the reads in the ORF and the A-site location on each read is predicted. Finally, Scikit-ribo compares the A-site locations to the canonical ORF, and reads that do not match it will be dropped from downstream analysis.

### Calculating RPKM-derived TE

We refer to ribosome density per mRNA as RPKM-derived TE. It is a commonly used proxy for TE, which can be calculated by the ratio of RPKM for a given gene  $i$ <sup>3, 42</sup>:

$$\text{Ribosome density per mRNA}_i = \frac{\text{RPKM}_i^{\text{Ribo}}}{\text{RPKM}_i^{\text{mRNA}}} \quad \text{Equation 5.1}$$

where  $\text{RPKM}_i^{\text{Ribo}}$  and  $\text{RPKM}_i^{\text{mRNA}}$  are the relative abundance of gene  $i$  in the Riboseq data and RNAseq data, respectively.

RPKM and TPM are defined by:

$$\text{RPKM}_i = \frac{R_i}{\left(\frac{l_i}{10^3}\right) \left(\frac{\sum_i R_i}{10^6}\right)} = \frac{R_i}{l_i \cdot \sum_i R_i} \cdot 10^9 \quad \text{Equation 5.2}$$

$$\text{TPM}_i = \left(\frac{\text{RPKM}_i}{\sum_i \text{RPKM}_i}\right) \cdot 10^6 \quad \text{Equation 5.3}$$

where  $R_i$ ,  $l_i$  are the sequencing coverage and coding sequence length of a gene, respectively.

In Riboseq studies, rather than using fragments per kilobase of gene per million reads mapped (FPKM), RPKM is employed (**Equation 5.1**). This is because the Riboseq reads are single stranded, and the companion RNAseq libraries were also made using a single stranded protocol to mimic the Riboseq data. Since  $l_i$  is a shared term between the two data, RPKM-derived TE can be further derived as:

$$\text{RPKM - derived TE}_i = \frac{R_i^{\text{Ribo}} / \sum_i R_i^{\text{Ribo}}}{R_i^{\text{mRNA}} / \sum_i R_i^{\text{mRNA}}} = \frac{R_i^{\text{Ribo}} / R_i^{\text{mRNA}}}{\sum_i R_i^{\text{Ribo}} / \sum_i R_i^{\text{mRNA}}} \quad \text{Equation 5.4}$$

The total number of reads  $\sum_i R_i^{\text{Ribo}}$  and  $\sum_i R_i^{\text{mRNA}}$  are fixed normalization factors shared between genes. Thus, the variance of the nominator, the ratio of the number of reads, determines the dispersion of RPKM-derived TE. That is why low abundance genes, either in the Riboseq or RNAseq data, report highly dispersed TE derived with RPKM.

### Correcting for biological biases with the Scikit-ribo GLM

The joint inference of TE and codon DT is achieved via a codon-level GLM with a penalized likelihood function<sup>203</sup> (**Equation 5.5**). The model can be fit using a python implementation of glmnet ([https://github.com/hanfang/glmnet\\_python](https://github.com/hanfang/glmnet_python)<sup>204</sup>). In Scikit-ribo, the

design matrix is loaded as a `scipy`<sup>205</sup> compressed sparse column matrix. This can effectively reduce memory usage, as the size of the design matrix grows exponentially with respect to the number of categorical variables. As a quality control, low MAPQ regions and genes with TPM less than one are excluded from the analysis. If a gene has fewer than 10 effective codons remaining, it is also excluded. The model assumes that the number of ribosomes  $Y_{ij}$  for each codon at position  $j$  of gene  $i$  follows a Poisson distribution with the mean equal to  $\mu_{ij}$  (**Equation 5.5**). A log link function is employed.

$$\begin{aligned}
 Y_{ij} &\sim \text{Poisson} (\text{mean} = \mu_{ij}) \text{ for position } j \text{ of gene } i \\
 \log \mu_{ij} &= \beta_0 + \beta^T x_{ij} \\
 \text{where } i &\in [0, I], j \in [0, J]
 \end{aligned}
 \tag{Equation 5.5}$$

To correct for the biological biases, Scikit-ribo considers the below three categorical covariates and a continuous covariate (**Figure 5.2B, Equation 5.6**). The first continuous covariate  $X_i^m$  represents mRNA abundance in TPM and its coefficient is fixed to be one, indicating the ribosomes are proportional to mRNA abundance. Before putting into the model, the  $\log \text{TPM}_i$  values are normalized by their mean and SD. The coefficients  $\beta_i^t$  (in  $\log_e$  scale) of the first categorical covariate  $X_i^t$  represent TE/TIE for each gene. The  $\log_2 TE_i$  can further be computed by using median normalization:  $\log_2 TE_i = (\beta_i^t - \text{median}(\beta^t)) / \log_e 2$ . The second categorical covariate  $X_{ij}^c$  represent the 61-sense codons. Their coefficients,  $\beta^c$  (in  $\log_e$  scale) are proportional to the relative codon DT, which are the inverse of codon ERs. The start and stop codons in each ORF are excluded, because of their relevance to translation initiation and termination, rather than elongation. Finally, the third categorical covariate  $X_{ij}^s$  indicates whether a likely double-stranded stem loop exists within 18 nt downstream of the current ribosome, as predicted from the optimal minimum free energy structure from RNAfold<sup>172</sup>. The current ribosome is likely to reside at a single strand part of the mRNA molecule.

$$g(\mu_{ij}) = \beta_0 + \underbrace{x_i^m}_{\text{mRNA}} + \underbrace{x_i^t \beta_i^t}_{\text{TE}} + \underbrace{x_{ij}^c \beta^c}_{\text{codon}} + \underbrace{x_{ij}^s \beta_{ij}^s}_{\text{secondary structure}}
 \tag{Equation 5.6}$$

where  $g(\cdot)$  is a log link function,  $\mu_{ij} = E[Y_{ij}]$ ,

$x_i^m$  is the mRNA abundance for gene  $i$  with its coefficient fixed to 1,

$\beta_i^t$  is the translational efficiency coefficient for gene  $i$ ,

$\beta^c$  is the codon dwell time (inverse of elongation rate) for codon  $c$ ,

$x_{ij}^s$  denotes whether secondary structure exists downstream of position  $j$  in gene  $i$ ,

$\beta_0$  is the intercept.

### Correcting for sampling errors with ridge penalty

To correct for the sampling errors, i.e. the high dispersion of TE among low-abundance genes, Scikit-ribo employs a GLM with a ridge penalty<sup>203</sup> ( $l_2$  norm) to provide shrinkage estimates of TEs (**Equation 5.7 and 5.8**). This is computed by setting the  $\alpha$  parameter in `glmnet` to zero. The lasso penalty is not considered here because we wish to infer all the coefficients (e.g. TEs of all genes), rather than performing variable selection. To optimize the log-likelihood,

Scikit-ribo calls `glmnet`<sup>203</sup>, which uses a Newton quadratic approximation (outer loop) and then coordinate descent on the resulting penalized weighted least-squares problem (inner loop). A ten-fold CV is performed to find the optimal  $\lambda$ , which controls the strength of  $l_2$  norm regularization. If one wishes to utilize or inspect the coefficients from an un-penalized GLM, this could be done by setting  $\lambda = 0$  when printing the coefficients.

The log likelihood for the observations  $\{x_{ij}, y_{ij}\}$  is given by

$$l(\beta|X, Y) = \sum_{i=0}^I \sum_{j=0}^J (y_{ij}(\beta_0 + \beta^T x_{ij}) - e^{\beta_0 + \beta^T x_{ij}}) \quad \text{Equation 5.7}$$

We optimize the  $l_2$  norm penalized log likelihood w. r. t. a total of N observations and K parameters:

$$\operatorname{argmin}_{\beta_0, \beta} -\frac{1}{N} l(\beta|X, Y) + \lambda \left( \sum_{k=1}^K \beta_k^2 / 2 \right) \quad \text{Equation 5.8}$$

where the optimal  $\lambda$  with the smallest Poisson deviance is decided via CV.

### Deriving relative protein abundance

As per the master equations for mRNA transcription and protein translation from Li<sup>42</sup>, for a gene  $i$ ,

$$\frac{d}{dt} P_i = k_i^2 M_i - \lambda_i^2 P_i \quad \text{Equation 5.9}$$

where  $M_i$  and  $P_i$  are the concentration of mRNA and protein, respectively.  $k_i^1$  and  $k_i^2$  are the transcription and translation efficiency, while  $\lambda_i^1$  and  $\lambda_i^2$  are the degradation rates of mRNA and protein. Under steady state,  $\frac{d}{dt} P_i = 0$ , thus, the relative protein abundance (PA) can be derived from Riboseq and RNAseq data using:

$$P_i = \frac{k_i^2}{\lambda_i^2} M_i = \frac{TE_i}{DR_i} M_i \propto TE_i M_i \quad \text{Equation 5.10}$$

where  $TE_i$  is the translation efficiency,  $M_i$  is the relative mRNA abundance in TPM, and  $DR_i$  is the relative protein degradation rates, which can be assumed identical across genes. For the Riboseq data alone,  $P_i$  approximates to the relative ribosome density/abundance in TPM.

### Sequencing reads processing

The complete sequencing reads processing workflow is shown in **Supplemental Figure 5.S15**. Each time a new fastq file is generated, it is recommended to run `fastqc` to ensure the expected outcome and replace runs with excessive quality errors. For both Riboseq and RNA-seq data, the first step is to identify and trim the 3'-end adapters from each read using `cutadapt`<sup>206</sup> (v1.13). The first base of the reads' 5'-end is also clipped to avoid contamination on the 5'-end. To filter out ribosomal RNA (rRNA) sequences, the resulting reads are aligned to the known rRNA using `Bowtie`<sup>207</sup> (v1.2.0). As a quality control, the reads that are too short or too long are removed using `Prinseq`<sup>208</sup>, keeping reads in a range from 15nt to 35nt (v0.20.4). In *E. coli*, the size range of the Riboseq reads is larger, so this filtering step on read size should be adjusted accordingly. The remaining reads are then aligned with `STAR`<sup>209</sup> (v2.4.0j) in a single pass mode



with parameters tuned for short reads (--sjdbOverhang 35). The quality control report file of the resulting bam is generated using Qualimap<sup>210</sup> (v2.0.2). From there, the RNAseq data is used to quantify the gene-level mRNA abundance in TPM using a quantifier. Salmon<sup>199</sup> and Kallisto<sup>200</sup> are recommended here because they are extremely fast and their file formats are automatically supported by Scikit-ribo.

### Scikit-ribo input processing

Scikit-ribo uses the pandas<sup>211</sup> data frame as the main data structure: a codon-level data frame for the GLM, and a read-level data frame for A-site prediction. The codon-level data frame consists of the following variables: chromosome, start, end, codon, secondary structure pairing probability, mRNA abundance in TPM, number of ribosomes at this codon. Scikit-ribo filters and converts the provided Riboseq bam file into a bed file using pysam(v0.10.0)<sup>15</sup> and pybedtools(v0.7.9)<sup>212, 213</sup>, which is subsequently converted into a read-level data frame. To prepare the codon-level data frame, it retrieves the cDNA sequence (includes ORF, 5'/3'-UTR) given a reference genome and a gene annotation file. The 24 nucleotides in both the 5'UTR and 3'-UTR are included for calculating mRNA secondary structure. The cDNA sequence is then used to predict the optimal secondary structure under minimal free energy using RNAfold(v2.3.4)<sup>172</sup>. By parsing the postscript files, Scikit-ribo finds the lbox entries, which represent the pairing of nucleotides in the optimal structure. With that, it identifies the positions on the ORF with a likely stem loop downstream (i.e. nine nucleotides downstream of the A-site), while the ribosome is residing at a likely single-strand region (i.e. from six nucleotides upstream to nine nucleotides downstream). Due to the uncertainty of RNAfold prediction, a likely stem loop requires at least 17 out of the 18 nucleotides to be paired, while a single-strand region requires no more than three nucleotides paired. Given the canonical ORF of a gene, Scikit-ribo splits the sequences into tri-nucleotides as codons.

### Data and statistical analysis in this paper

For the wild-type *S. cerevisiae* analysis, the Riboseq (flash-freeze protocol) and RNA-seq (Ribo-zero protocol) data were from Weinberg *et al*<sup>136</sup>. The accession numbers are GSM1289257, GSM1289256. For the analysis involving *Dhh1p*, the Riboseq and RNA-seq data were from Radhakrishnan *et al*<sup>145</sup> under the accession number GEO: GSE81269. The reference genome of *S. cerevisiae* used is S288C R64-2-1. The gene annotation file was the SGD annotation downloaded from UCSC. For the *E. coli* analysis, the Riboseq (ReLE protocol) and RNA-seq data were from Hwang *et al*<sup>152</sup>. The accession number is GSE85540. The reference genome of *E. coli* used is the MG1655 genome. For more details of how these data were generated, please refer to the original papers. All the figures in the paper were plotted using matplotlib<sup>214</sup> (v2.0.0) and seaborn<sup>215</sup> (v0.7.1). The Pearson correlation and Spearman correlation are denoted as  $r$  and  $\rho$ , respectively.

To ensure reproducibility, all source codes for data processing, statistical analyses and figure plotting are available in the iPython notebooks under the GitHub repository: [https://github.com/hanfang/scikit-ribo\\_manuscript](https://github.com/hanfang/scikit-ribo_manuscript)

### Simulation, sequence enrichment, and gene enrichment analysis

The simulation of the *S. cerevisiae* Riboseq and RNAseq data were done with polyester<sup>216</sup> and the  $\log TE_{baseline}$ , followed a balanced normal distribution. To mimic paused ribosomes, we randomly sampled 2500 sites (occurring within ~20% of the genes) and added

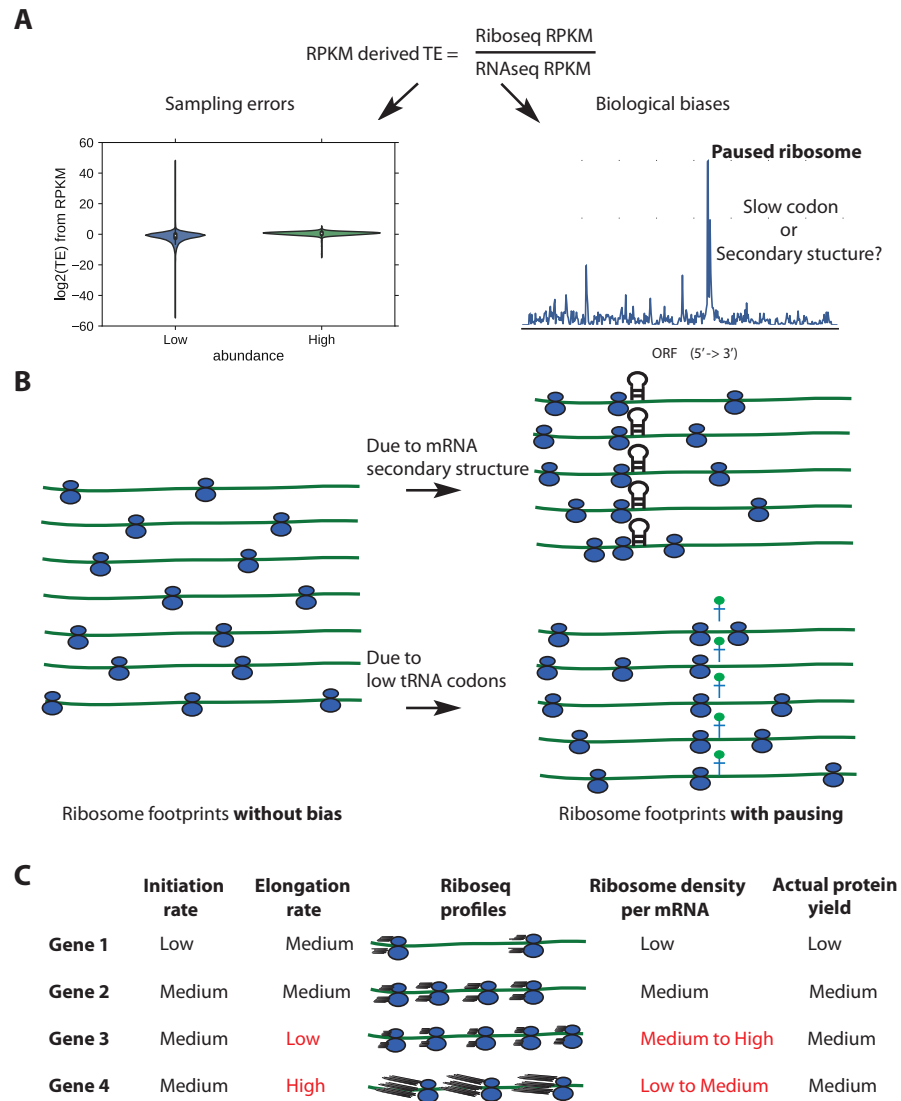
1000 additional reads into these locations of the Riboseq data. We then sampled back to the same number of reads as the original data and computed the new RPKM-derived  $\log TE_{RP}$ . For the sequence enrichment analysis, we collected 5'UTR sequences from genes with  $\log_2 TE$  greater than two. The 5'UTR region is from 50 nt upstream to 6nt downstream of the translation start site. Then we used HOMER (v4.9) to scan for enriched sequences from the 56nt windows<sup>184</sup>, using the HOMER recommended p-value cutoff of  $1 \times 10^{-10}$ .

Gene set enrichment analysis required three steps. First, we excluded low abundance gene (TPM < 10) to focus on effects of the biological covariates (e.g. codon ER). Second, we selected 50 genes from the left and right tails, i.e. genes with the most changes of TE. This cutoff gave bounds of  $\Delta \log_2 TE$  about -0.9 and +1.7 in the three comparisons. Finally, we uploaded the gene sets to <http://www.yeastgenome.org/> and performed enrichment analysis<sup>217</sup>.

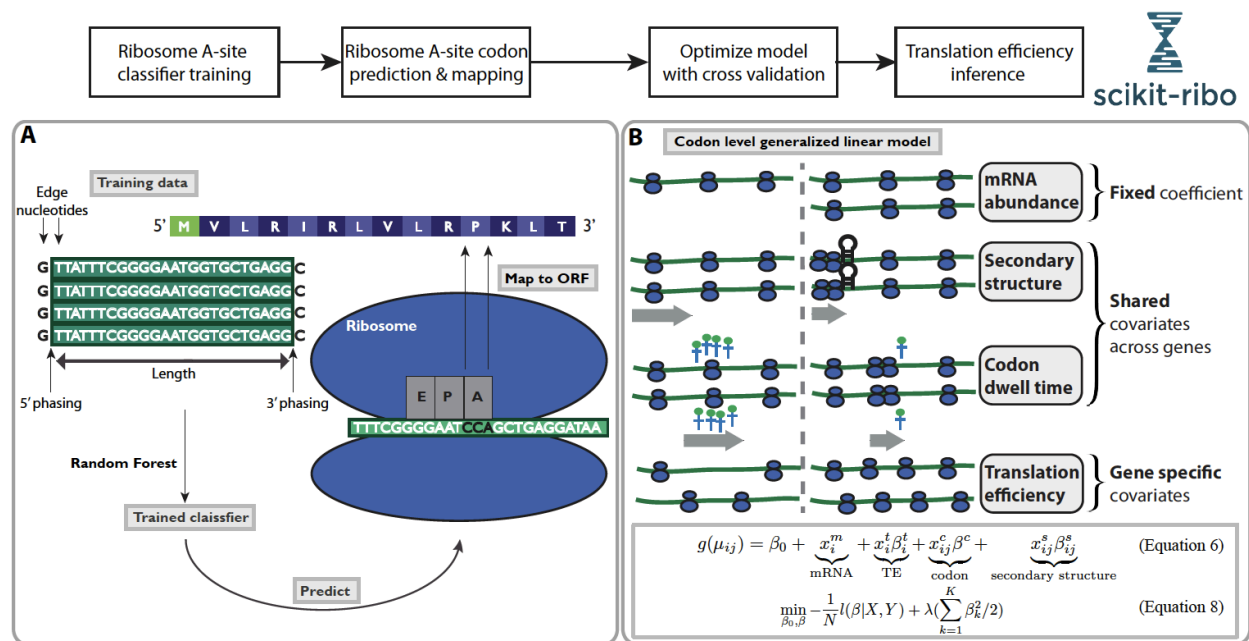


## Figures and tables in this chapter

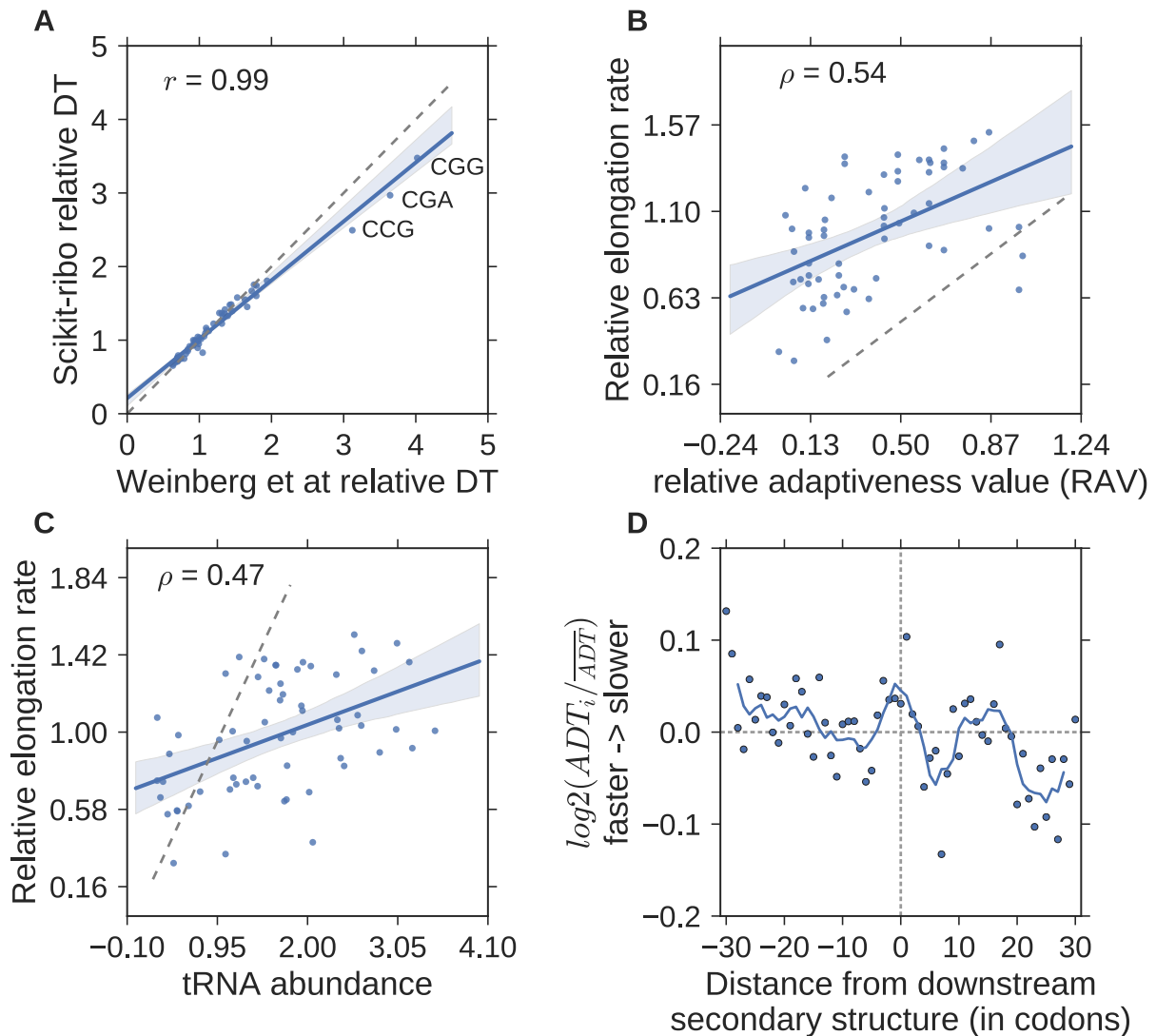
### Figures



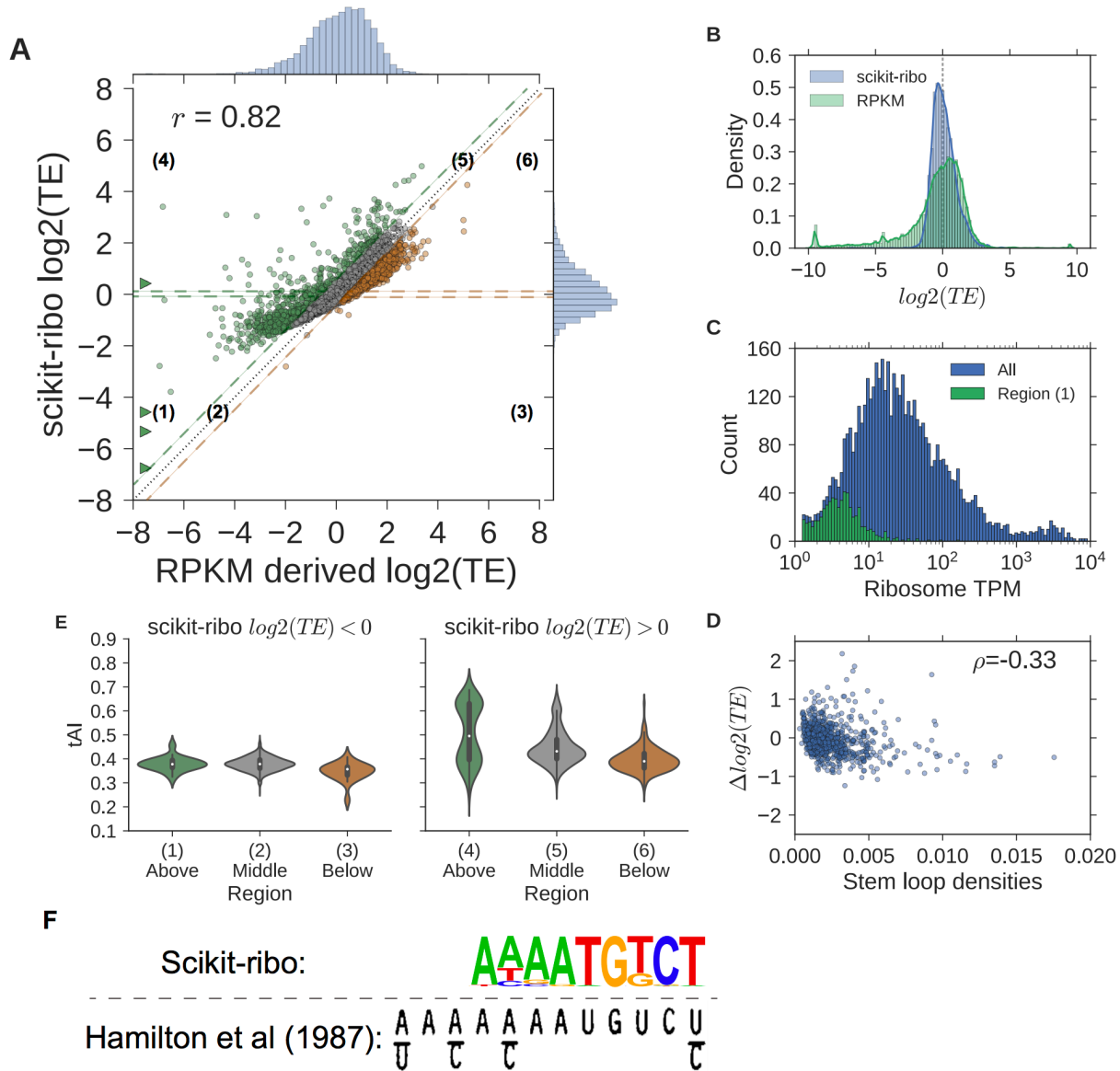
**Figure 5.1. Sources of biases using ribosomes densities per mRNA (RPKM-derived TE) as a proxy for TE.** (A) Sampling biases towards low abundance genes (left), and biological biases due to paused ribosomes (right). (B) Idealized ribosome footprints distribution without biases (left), or with downstream mRNA secondary structure and low conjugate tRNA availability for the A-site codon (right). (C) Confounding effects of translation initiation and elongation on Riboseq profiles, figure adapted from Quax *et al* 2013. Initiation rate should be proportional to actual protein yield.



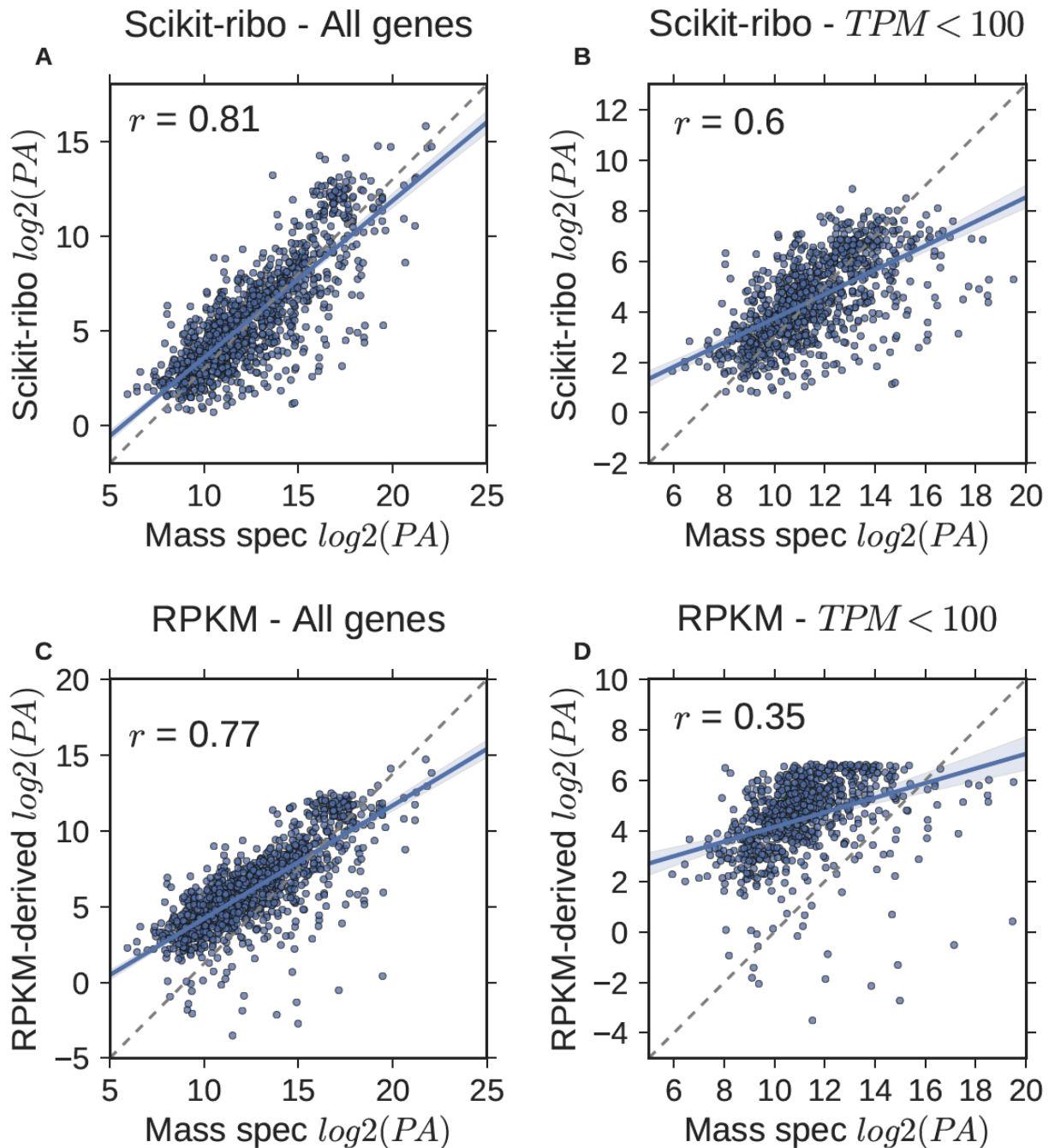
**Figure 5.2. Overview of the analysis workflow in Scikit-ribo.** The complete workflow consists of Ribosome A-site classifier training, A-site codon prediction and mapping, and translation efficiency inference. **(A)** Ribosome A-site training and prediction, gray text boxes denote the major steps. **(B)** Illustration of the covariates in the codon level generalized linear model. In the model, the mRNA abundance (in TPM) are considered as offset with fixed coefficient equal to one. Codon dwell time and mRNA secondary structure are shared covariates across genes. Translation efficiencies are gene specific covariates.



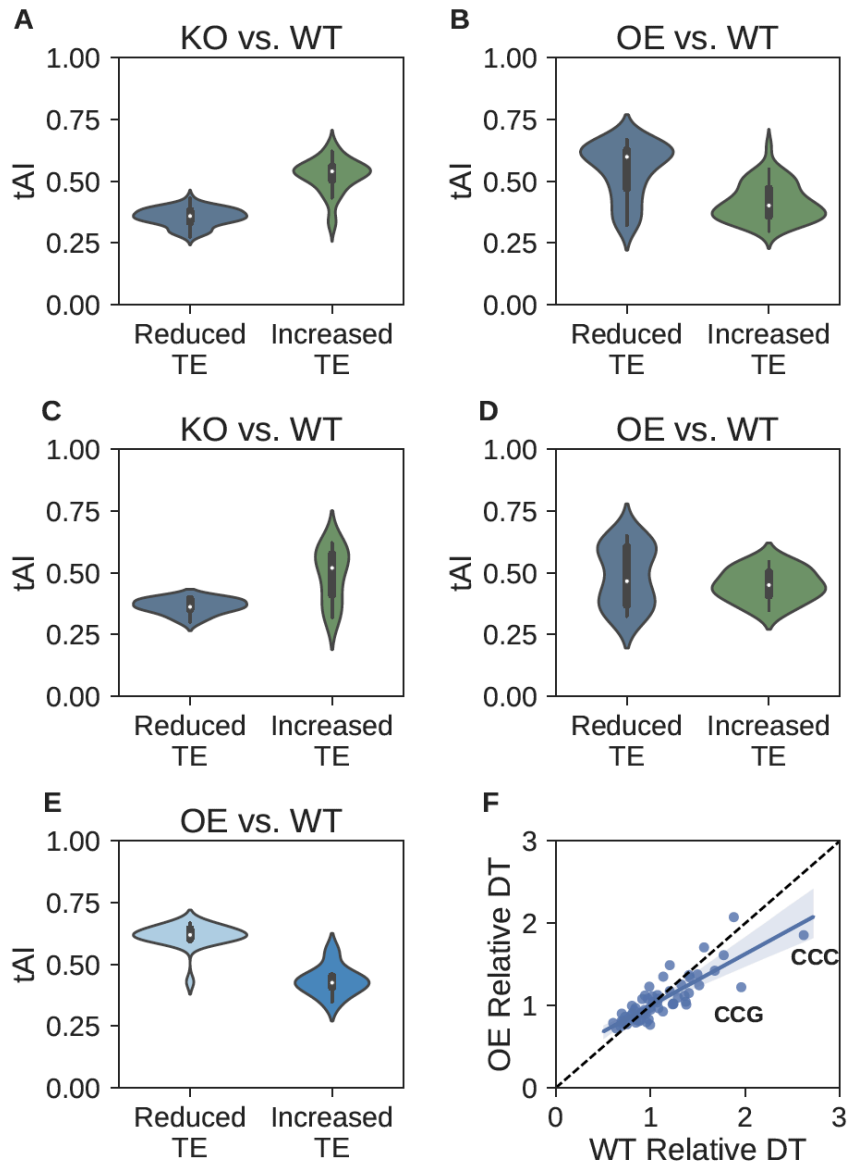
**Figure 5.3. Accurate inference of codon elongation rates and mRNA secondary structure.** (A) Almost perfectly reproduced codon dwell time (DT), inverse of elongation rate) from Weinberg *et al* ( $r=0.99$ ). (B) Correlation with the codon's adaptiveness value (RAV,  $r=0.5$ ), (C) Correlation with tRNA abundance ( $r=0.47$ ). In A-C, the gray dashed line denotes the diagonal line;  $y=x$ . The RAV scales from 0 to 1. A codon with lower RAV means that it is less optimal for translation elongation, i.e. slower codons. (D) Meta gene analysis of the log ratio of adjusted DT (ADT), divided by the mean adjusted DT. The solid line denotes the average ADT in a five-codon sliding window. A log ratio greater than zero means ribosomes at this position are faster than average. The log ratios on the left were significantly higher than the ones on the right (T-test,  $p\text{-value}=5 \times 10^{-3}$ ). The unit of the distance is codon.



**Figure 5.4. Pair-wise comparisons of estimates between Scikit-ribo and RPKM-derived TE.** (A) Scatter plot of Scikit-ribo and RPKM derived  $\log_2(TE)$ . Difference in  $\log_2(TE)$ :  $\Delta \log_2(TE)$ .  $\Delta \log_2(TE) > 0.5$ , previously underestimated (green),  $\Delta \log_2(TE) < -0.5$ , previously overestimated (orange), and other genes in between (gray). The genes with  $\Delta \log_2(TE)$  less than -8 are indicated by triangles. (B) Histograms of scikit-ribo and RPKM-derived  $\log_2(TE)$ ,  $\log_2(TE)$  values less than -10 are adjusted to -10 (C) Histograms of ribosome TPM in all genes (blue), and region 1 (green). (D) Violin plots of  $\Delta \log_2(TE)$  by the number stem loops. (E) Violin plots of tAI for genes in the six regions, left:  $\log_2(TE) < 0$ , right:  $\log_2(TE) > 0$ . (F) The Kozak consensus sequence, AAAATGTCT, found with the TE estimates from Scikit-ribo (p-value= $1 \times 10^{-21}$ ). The lower panel is adapted from the original paper, Hamilton *et al* (1987).

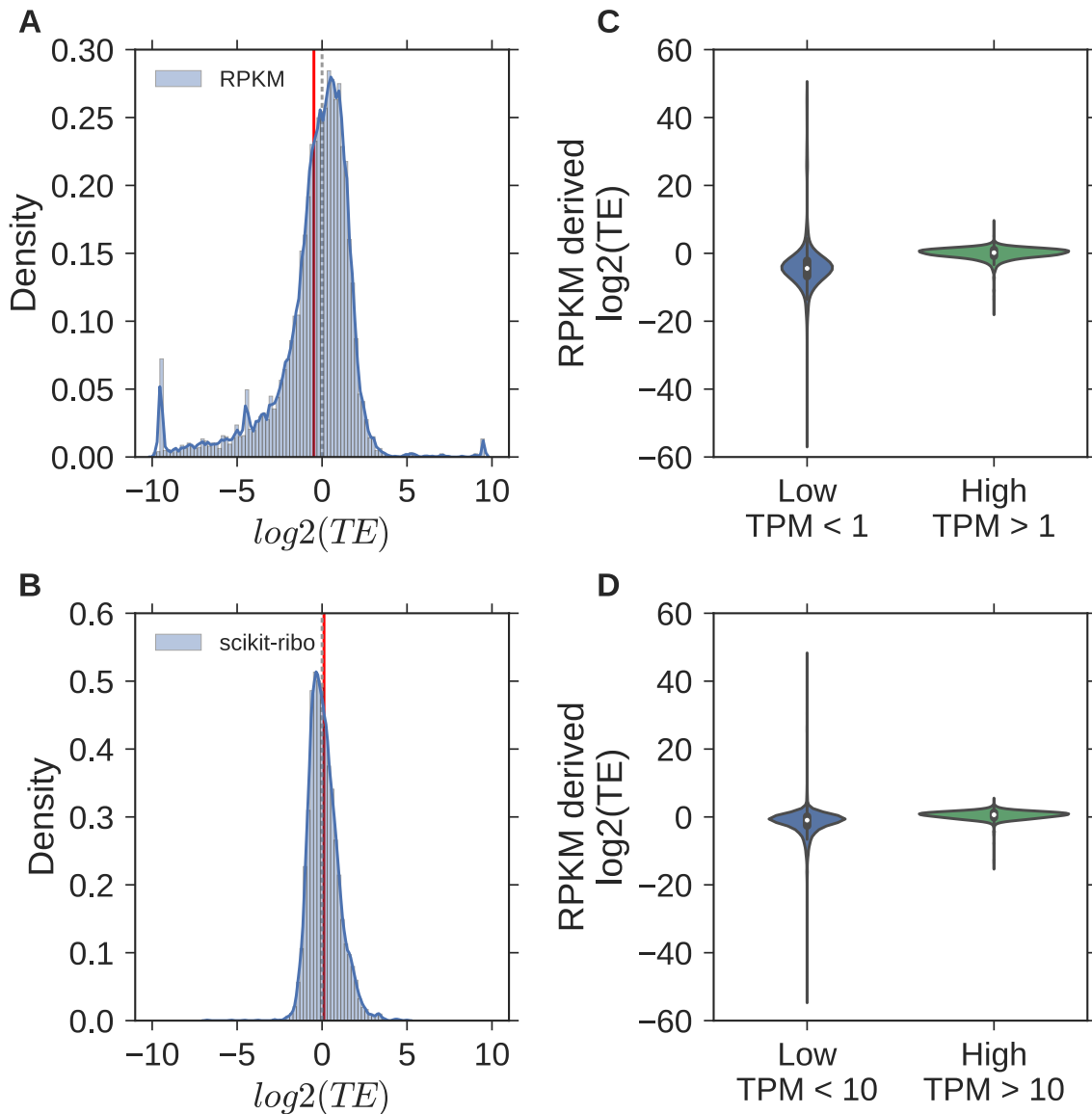


**Figure 5.5. Large-scale validation with mass spectrometry data confirmed Scikit-ribo's accurate TE estimates, especially for low-abundance genes.** (A) Scikit-ribo derived protein abundance (PA) for all genes in the validation set ( $r = 0.81, \beta = 0.83$ ). (B) Scikit-ribo derived PA for genes with TPM less than 100 ( $r = 0.6, \beta = 0.48$ ). (C) RPKM-derived PA for all genes in the validation set ( $r = 0.77, \beta = 0.75$ ). (D) RPKM-derived PA for genes with TPM less than 100 ( $r = 0.35, \beta = 0.29$ ). The black dashed line denotes the identity line;  $y=x$ .

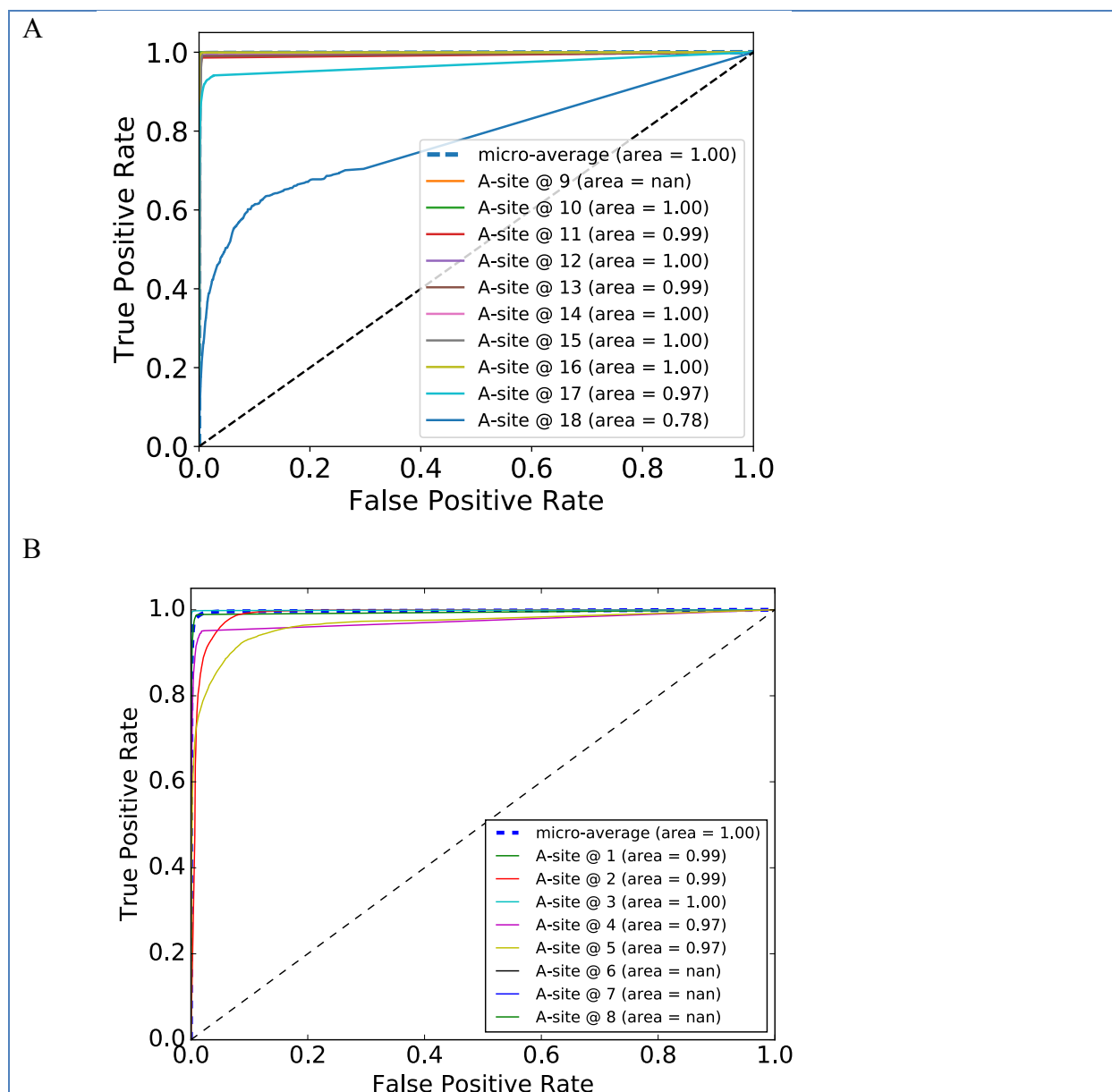


**Figure 5.6. Analysis of the *Dhh1p* data using Scikit-ribo.** Violin plots of tAI for genes with decreased/increased TE in (A) Knock Out (KO), (B) Over Expressed (OE), relative to Wild Type (WT). Violin plots of tAI for tail genes unique to Scikit-ribo in (C) KO, (D) OE. (E) Violin plot of tAI for genes, left: reduced TE in OE, and right: increased TE in OE. (F) Scatter plot of DT comparing OE and WT. WT: wild type, KO: knock out *Dhh1p*, OE: Overexpression of *Dhh1p*. The black dashed line denotes the identity line;  $y=x$ .

Supplemental figures

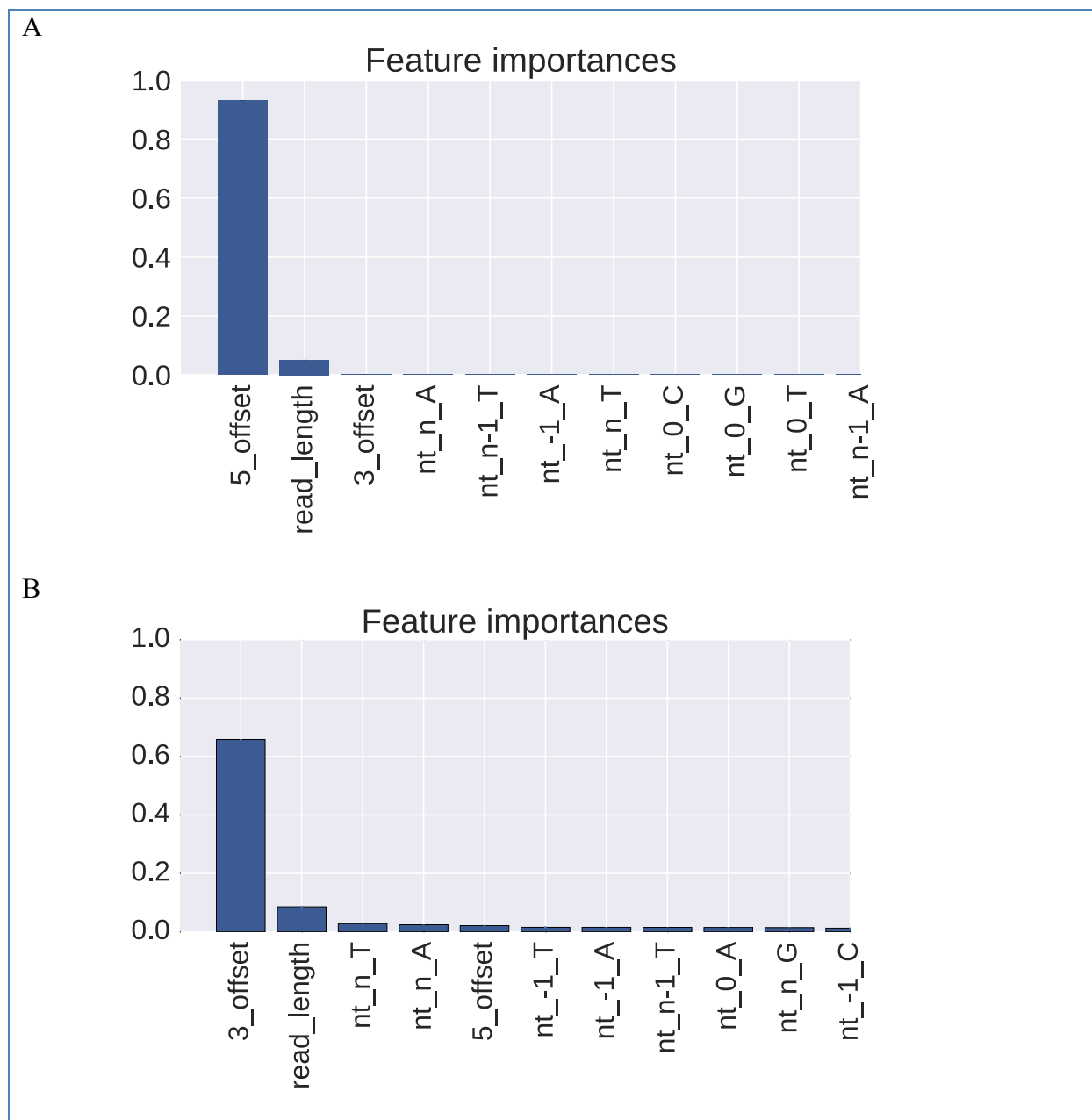


**Supplemental figure 5.S1. RPKM-derived  $\log_2(TE)$  and scikit-ribo  $\log_2(TE)$ .** (A) Scikit-ribo reported a balanced  $\log_2(TE)$  distribution (mean=0.1). The red solid line denotes the mean. (B) The RPKM-derived  $\log_2(TE)$  reported high dispersion among low abundance genes (TPM<1), while the genes with TPM > 1 still reported a long tail on the negative side. (C) The RPKM-derived  $\log_2(TE)$  reported a skewed distribution (mean=-0.5). (D) Even increasing the TPM cutoff to 10, the RPKM-derived  $\log_2(TE)$  still reported a long tail on the negative side.

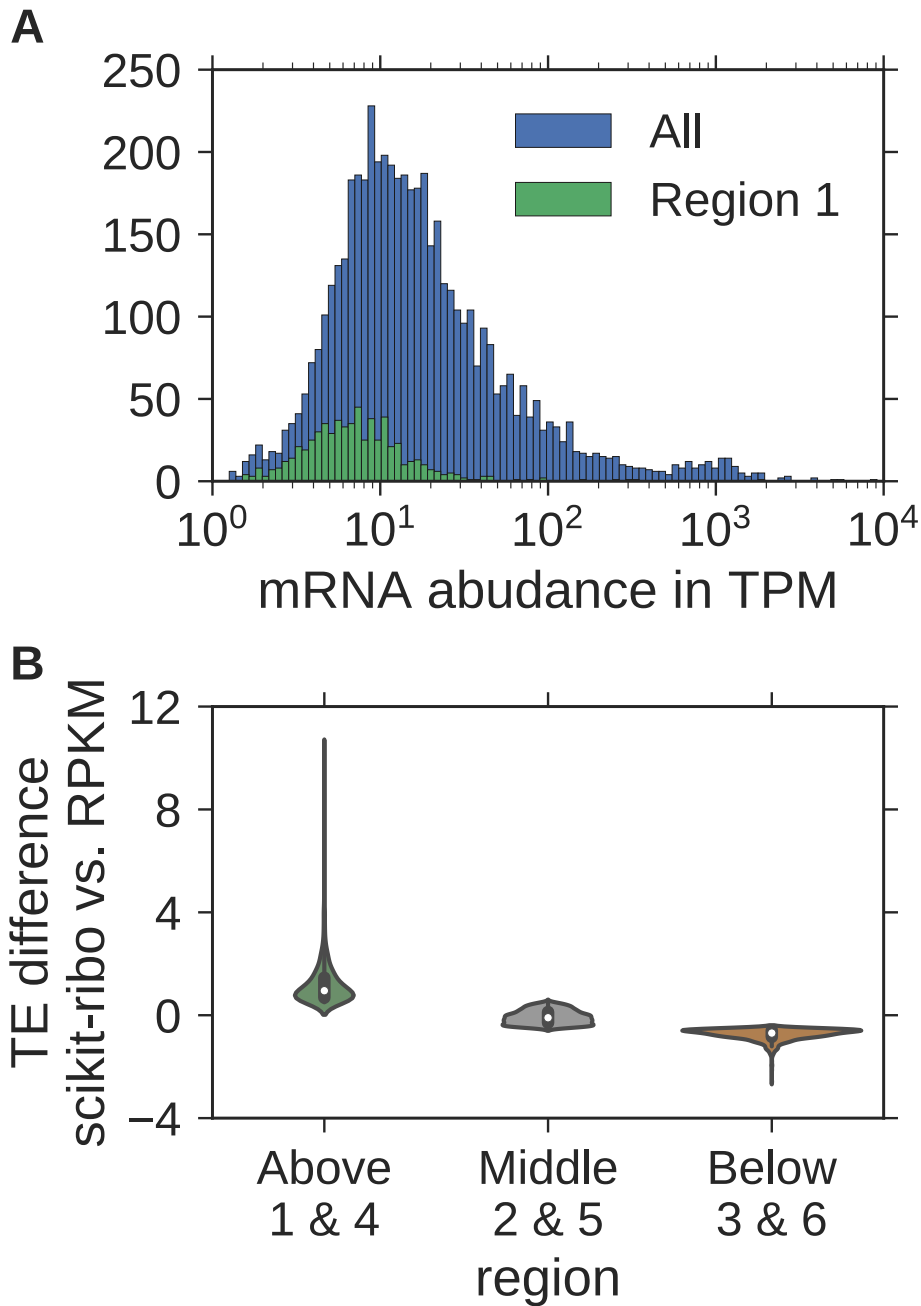


**Supplemental figure 5.S2. Multi-class ROC curves for A-site prediction.** (A) *S. cerevisiae* RNase I data. (B) *E. coli* RelE data. Each curve represents the data with different A-site locations (12 to 18 in RNase I, 1 to 8 in RelE). The dash line represents the micro-average across classes.

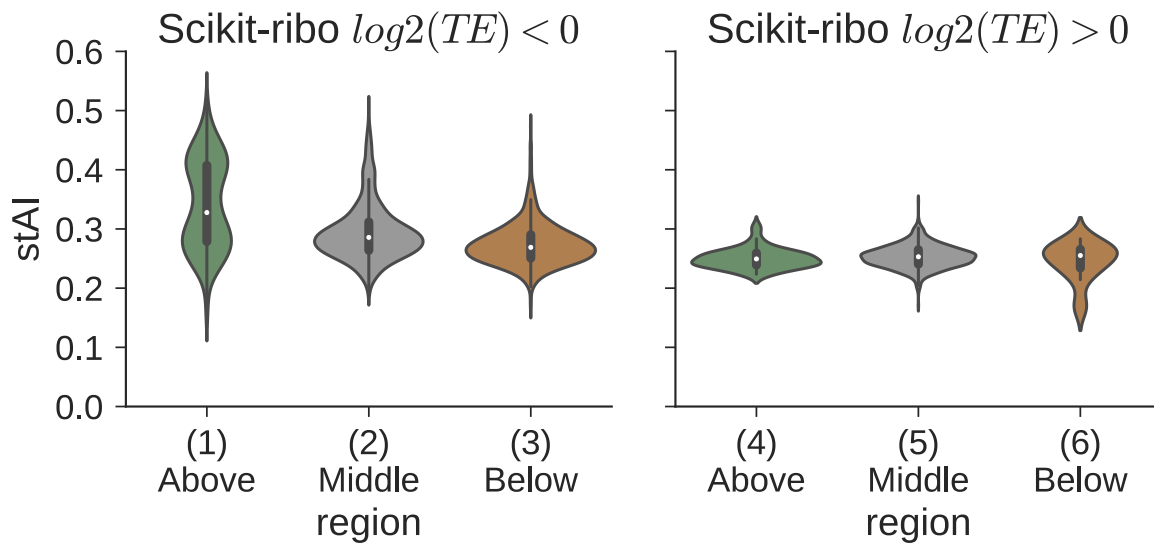




**Supplemental figure 5.S3. Feature importance from the random forest model.** (A) *S. cerevisiae* RNase I data. (B) *E. coli* RelE data. 5/3\_offset represents whether the 5'/3' end of the read is in the first/second/third reading frame. Nt\_-1/0/n-1/n represents the nucleotide at that position.

















**Supplemental figure 5.S4. Analysis of mRNA abundance in TPM by region.** Related to Figure 5.4; (A) Histograms of mRNA TPM in all genes (blue), and region 1 (green). (B) Violin plots of TE difference in the three regions, similar to Figure 4.













**Supplemental figure 5.S5. Violin plots of stAI for genes in the six regions.** Related to Figure 5.4; left:  $\log_2(TE) < 0$ , right:  $\log_2(TE) > 0$ .

\* - possible false positive

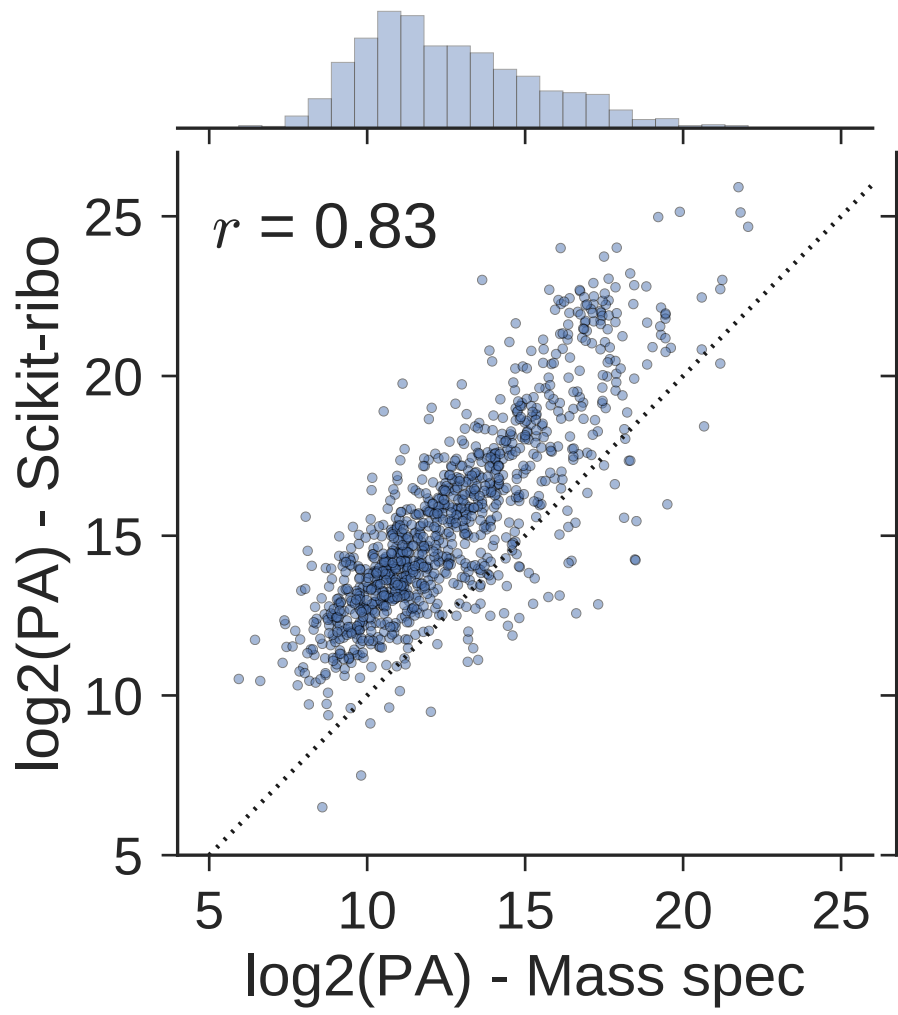
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-21	-4.854e+01	21.09%	2.13%	29.2bp (31.9bp)
2 *		1e-11	-2.569e+01	13.61%	1.88%	14.6bp (16.2bp)
3 *		1e-10	-2.440e+01	4.08%	0.04%	8.7bp (14.6bp)
4 *		1e-9	-2.266e+01	14.29%	2.49%	28.6bp (36.0bp)
5 *		1e-8	-2.062e+01	5.44%	0.21%	11.6bp (15.4bp)
6 *		1e-8	-1.945e+01	5.44%	0.24%	9.2bp (16.0bp)
7 *		1e-8	-1.897e+01	5.44%	0.26%	20.4bp (31.5bp)
8 *		1e-8	-1.878e+01	16.33%	4.06%	27.9bp (38.8bp)
9 *		1e-7	-1.777e+01	4.08%	0.11%	6.1bp (16.6bp)
10 *		1e-6	-1.479e+01	7.48%	1.01%	17.3bp (17.7bp)
11 *		1e-6	-1.436e+01	2.72%	0.04%	9.5bp (11.8bp)
12 *		1e-4	-1.031e+01	25.17%	12.76%	28.9bp (35.5bp)
13 *		1e-1	-3.783e+00	0.68%	0.02%	0.5bp (15.8bp)
14 *		1e0	-2.205e+00	33.33%	28.41%	23.6bp (33.8bp)

**Supplemental figure 5.S6. Statistically enriched sequences based on Scikit-ribo's TIE estimates using HOMER.** Related to Figure 5.4; The Homer's suggested p-value threshold is  $1 \times 10^{-10}$  to  $1 \times 10^{-12}$ .

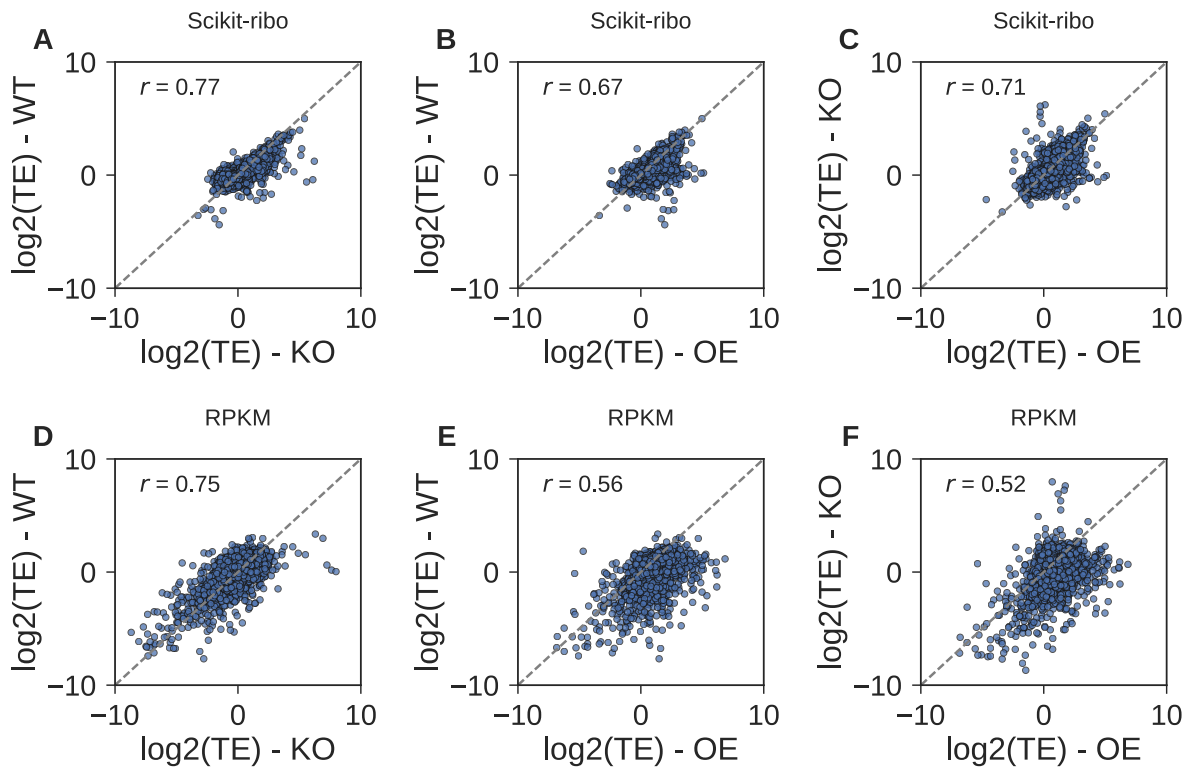
\* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)
1 *		1e-11	-2.587e+01	9.95%	1.34%	12.2bp (14.3bp)
2 *		1e-9	-2.123e+01	2.49%	0.02%	10.1bp (17.9bp)
3 *		1e-7	-1.829e+01	4.98%	0.40%	12.6bp (16.0bp)
4 *		1e-7	-1.705e+01	4.98%	0.45%	11.2bp (14.7bp)
5 *		1e-6	-1.586e+01	4.98%	0.52%	15.6bp (18.8bp)
6 *		1e-6	-1.442e+01	6.47%	1.12%	14.5bp (17.0bp)
7 *		1e-6	-1.393e+01	5.47%	0.80%	14.2bp (14.0bp)
8 *		1e-6	-1.385e+01	8.46%	2.03%	9.1bp (17.7bp)
9 *		1e-5	-1.199e+01	5.47%	0.98%	9.3bp (21.4bp)
10 *		1e-4	-1.111e+01	3.98%	0.53%	11.5bp (13.9bp)

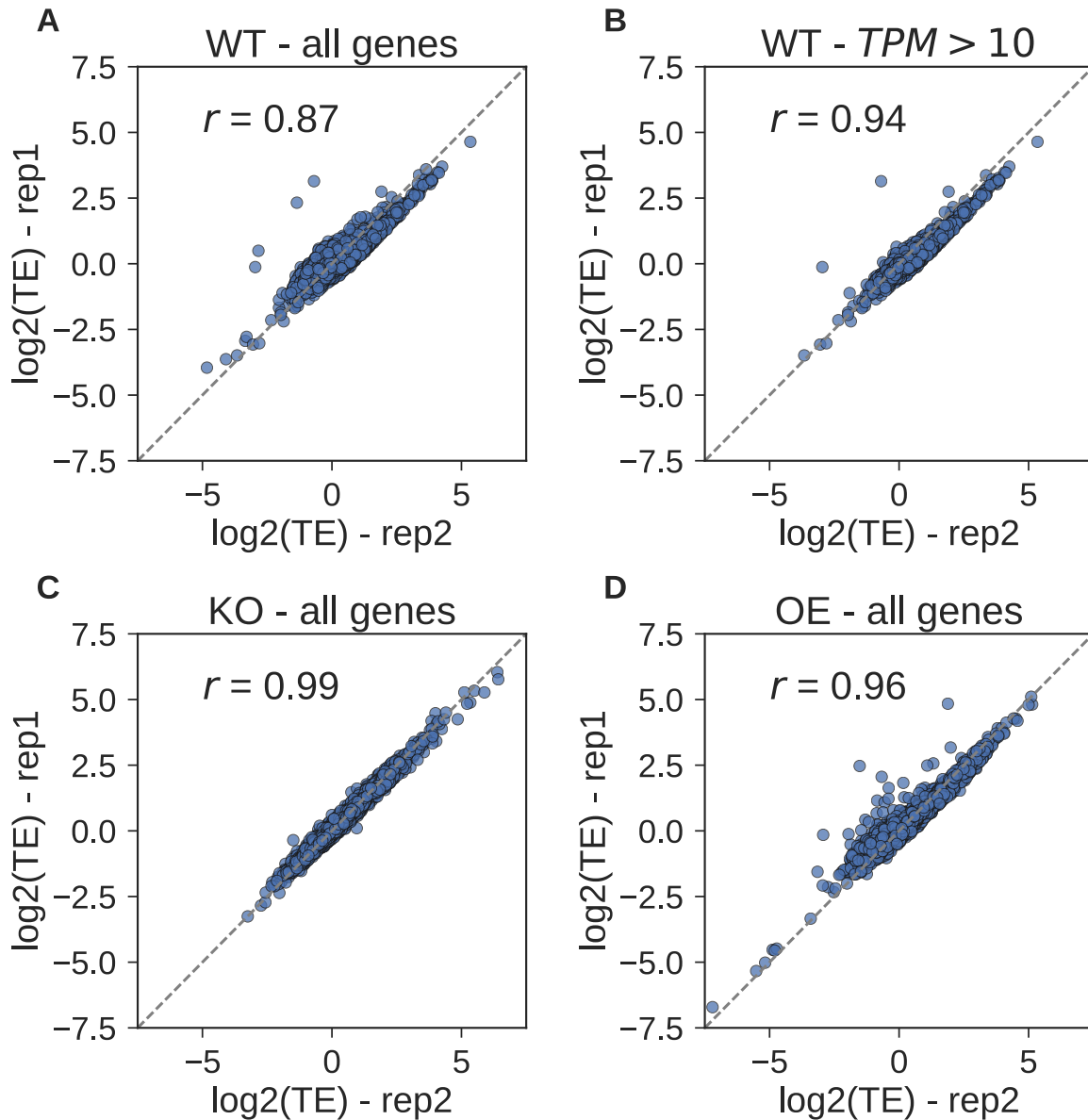
**Supplemental figure 5.S7. Statistically enriched sequences based on RPKM-derived TE estimates using HOMER.** Related to Figure 5.4; The Homer's suggested p-value threshold is  $1 \times 10^{-10}$  to  $1 \times 10^{-12}$ .



**Supplemental figure 5.S8. Higher correlation between scikit-ribo derived PA and SRM measurement, after considering protein degradation rate.** Related to Figure 5.5; The protein degradation rate was obtained from Christiano et al ( $r = 0.83$ ).

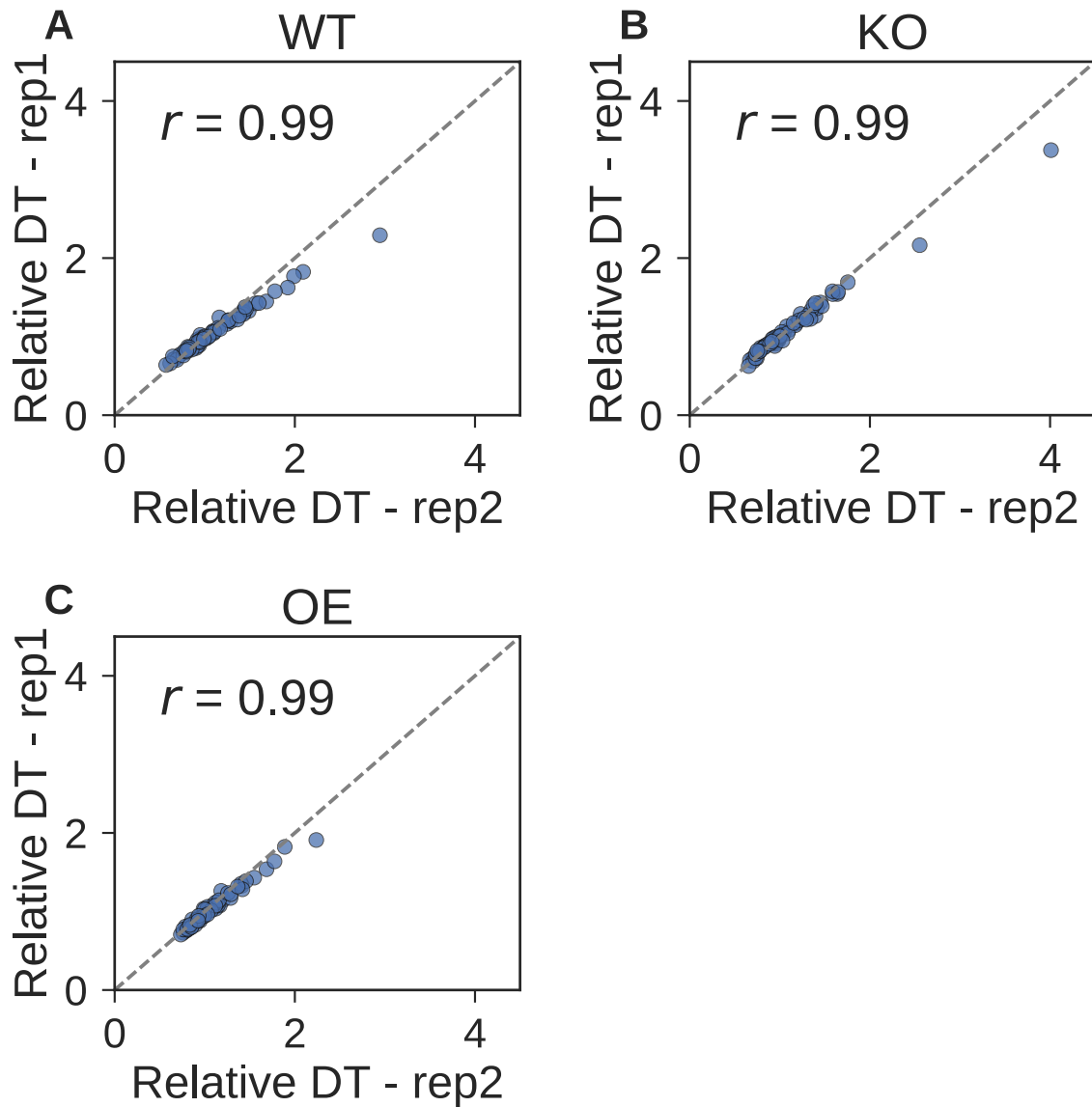


**Supplemental figure 5.S9. Substantial differences of TE between strains.** Related to Figure 5.6; (A-C) based on TE estimates from scikit-ribo, (D-F) based on RPKM-derived TE. WT: wild type, KO: knock out *Dhh1p*, OE: Overexpression of *Dhh1p*.

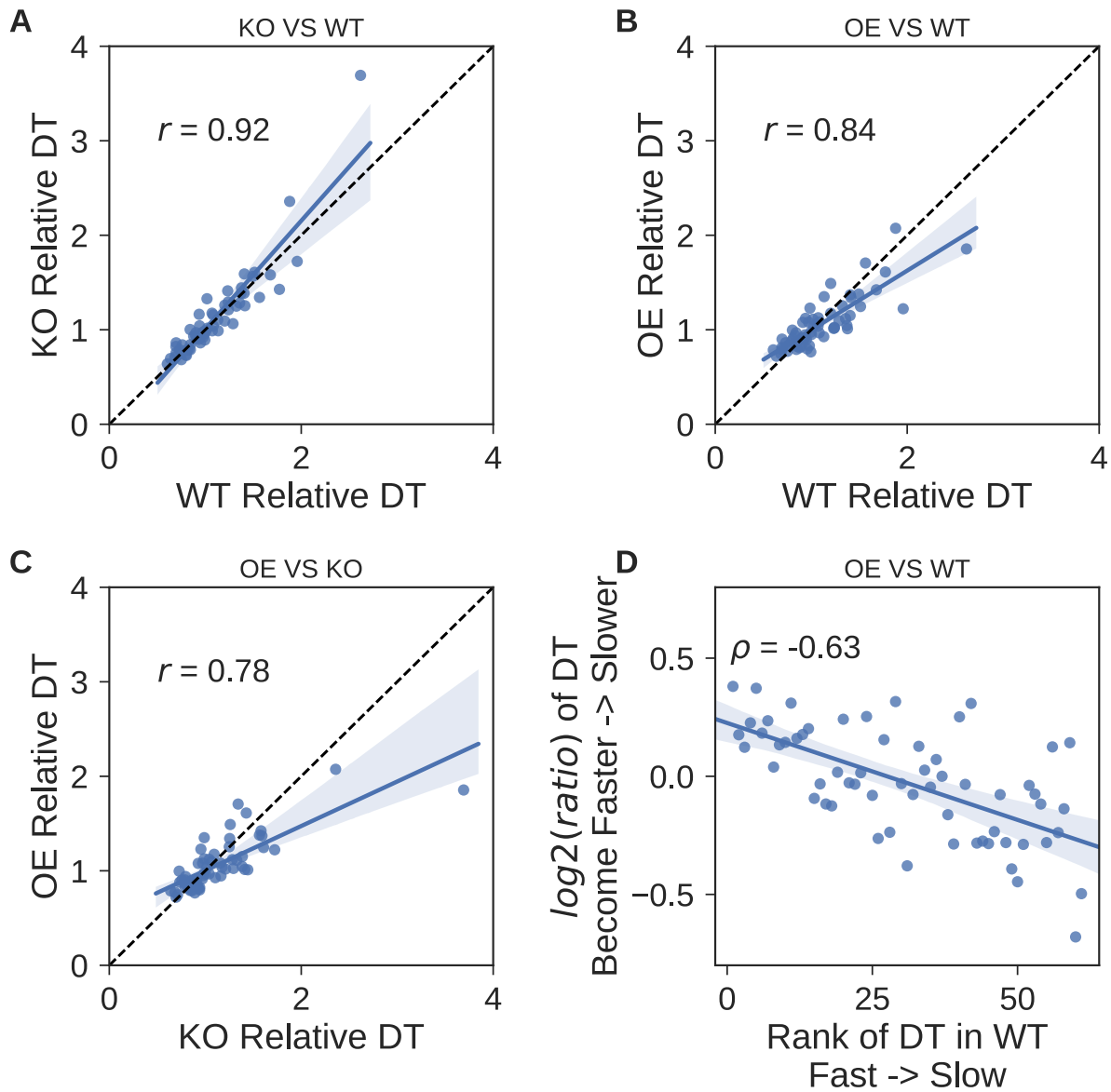


**Supplemental figure 5.S10. Highly reproducible TE estimates between replicates.** Related to Figure 5.6; (A) WT: wild type, 55 million and 16.7 million in replicate 1 and 2 ( $r=0.87$ ). (B) WT with TPM greater than ( $r=0.94$ ). (C) KO: knock out *Dhh1p* ( $r=0.99$ ), 74 million and 56 million in replicate 1 and 2. (D) OE: Overexpression of *Dhh1p*, 80 million and 39 million in replicate 1 and 2 ( $r=0.96$ ). The correlation was a function of the number of reads in each replicate.

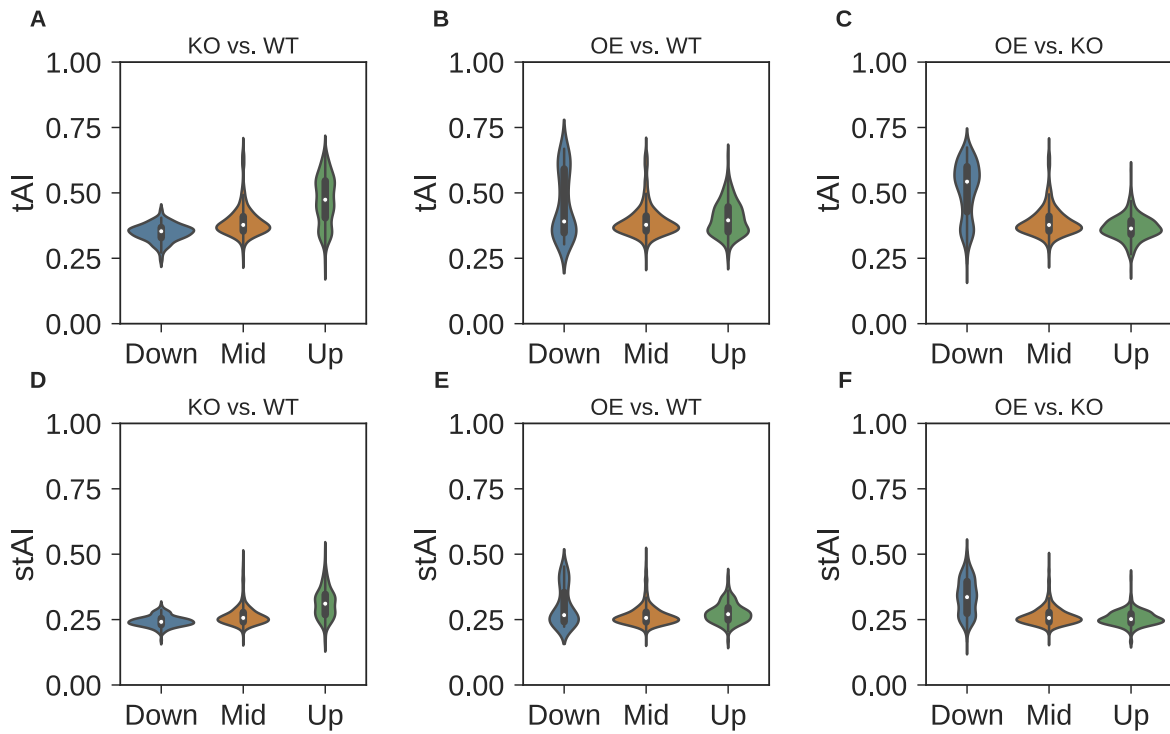




**Supplemental figure 5.S11. High correlation of codon dwell time (DT) between biological replicates.** Related to Figure 5.6; (A) wild-type, range of DT: 2.01, SD: 0.36, (B) KO, range: 3.05, SD: 0.45, (C) OE, range: 1.35, SD: 0.27. WT: wild type, KO: knock out *Dhh1p*, OE: Overexpression of *Dhh1p*.

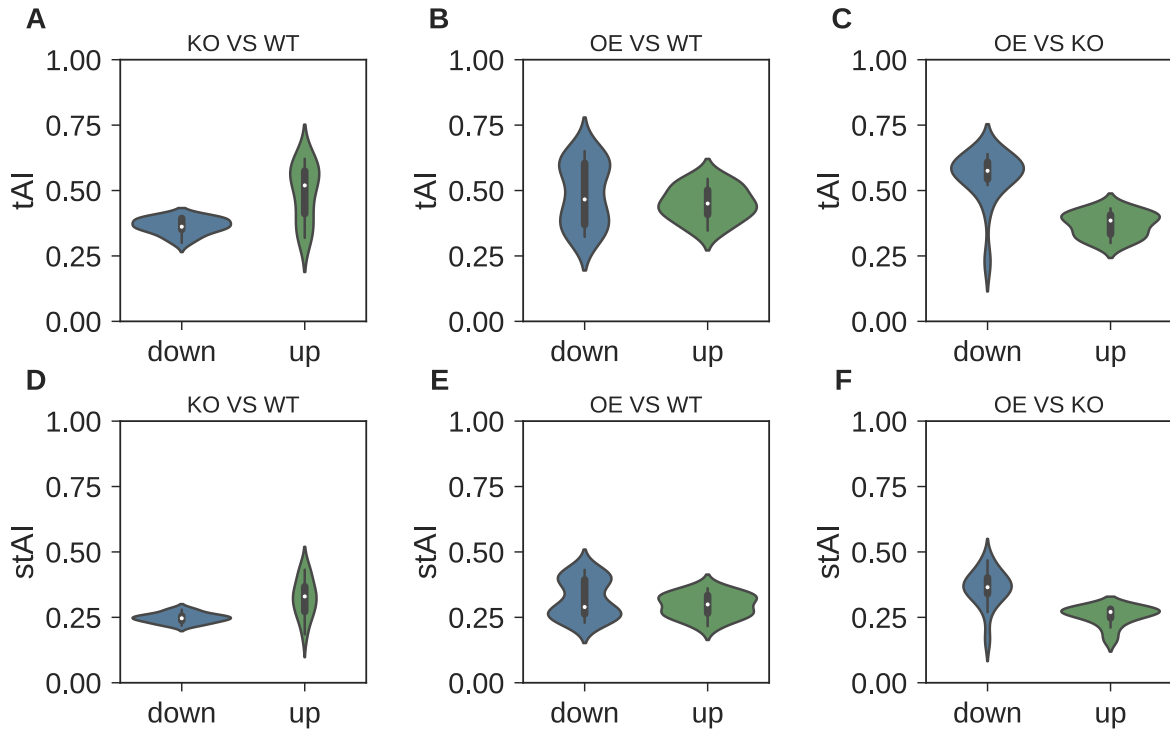


**Supplemental figure 5.S12. Codon dwell time (DT) comparisons between strains.** Related to Figure 5.6; (A) KO versus WT, (B) OE versus WT, (C) OE versus KO. (D) Compare the log ratio of DT with the original rank of DT in WT. WT: wild type, KO: knock out *Dhh1p*, OE: Overexpression of *Dhh1p*.

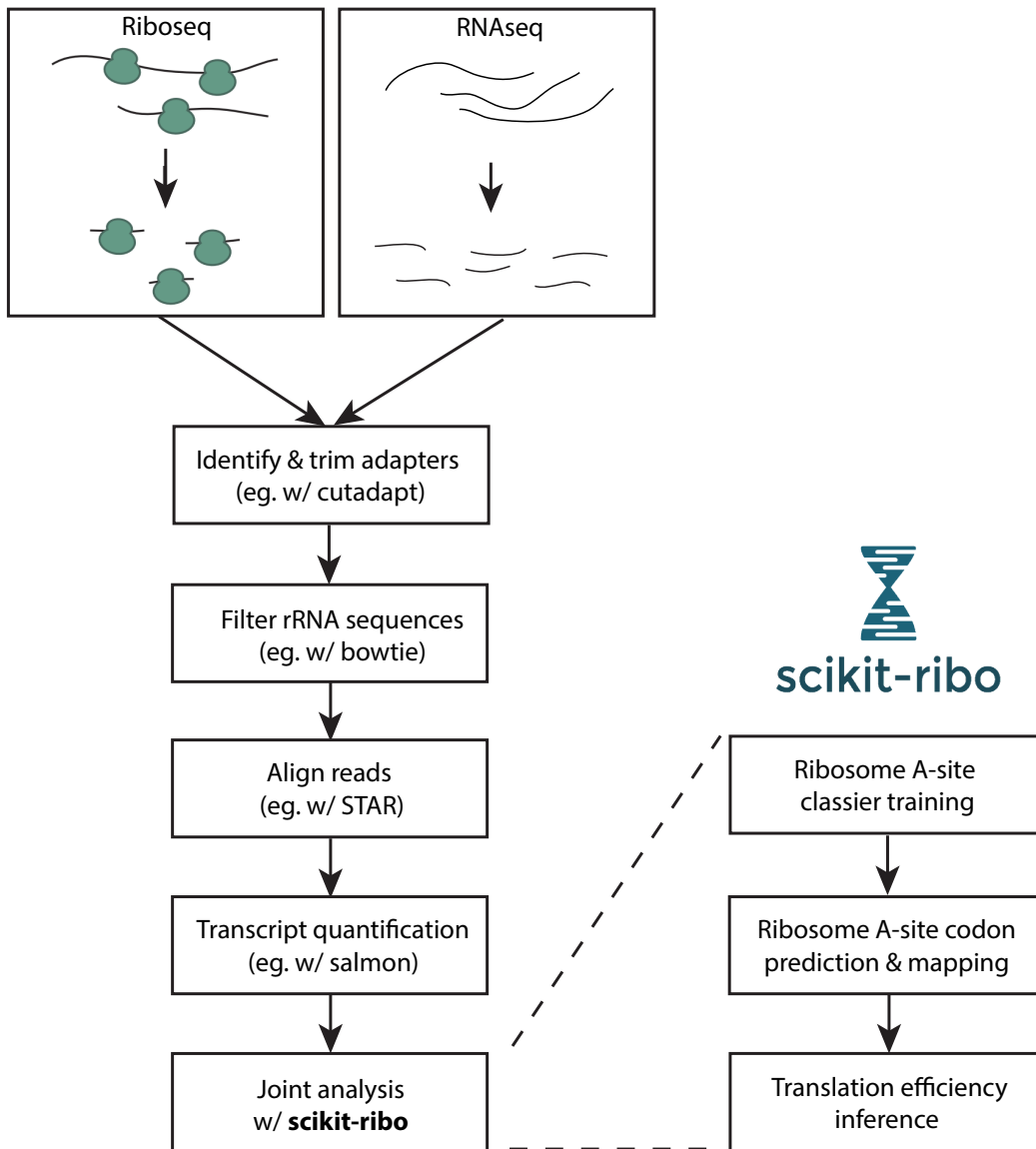


**Supplemental figure 5.S13. Reproducing Radhakrishnan *et al*'s findings on TE changes and codon optimality.** Related to Figure 6; (A-C) are based on tAI, (C-E) are based on stAI. The category down:  $\Delta\log_2(TE) < 1$ , the category up:  $\Delta\log_2(TE) > 1$ , the category mid: genes in between. WT: wild type, KO: knock out *Dhh1p*, OE: Overexpression of *Dhh1p*.

Genes unique to scikit-ribo



**Supplemental figure 5.S14. Genes with extreme TE changes that were unique to scikit-ribo.** Related to Figure 5.6; (A-C) are based on tAI, (C-E) are based on stAI. The category down:  $\Delta\log_2(TE) < 1$ , the category up:  $\Delta\log_2(TE) > 1$ . WT: wild type, KO: knock out *Dhh1p*, OE: Overexpression of *Dhh1p*.



**Supplemental figure 5.S15. The complete workflow of scikit-ribo analysis.**

Supplemental Tables

Study	SRR #	Mean accuracy	SD	# Optimal features
<i>S. cerevisiae</i> RNase I				
Weinberg et al (2016)	SRR1049521	0.987	0.004	3
Radhakrishnan et al (2016)	SRR3493886	0.981	0.008	2
Radhakrishnan et al (2016)	SRR3493887	0.929	0.036	2
Radhakrishnan et al (2016)	SRR3493890	0.982	0.008	4
Radhakrishnan et al (2016)	SRR3493891	0.963	0.022	2
Radhakrishnan et al (2016)	SRR3493894	0.941	0.019	7
Radhakrishnan et al (2016)	SRR3493895	0.936	0.025	2
Radhakrishnan et al (2016)	SRR3493898	0.938	0.03	2
<i>E. coli</i> RelE				
Hwang et al (2016)	SRR4023280	0.910	0.041	1
Hwang et al (2016)	SRR4023281	0.810	0.043	1

**Supplemental Table 5.S1. Prediction accuracy of A-site locations.** Mean and SD were computed via 10-fold cross validation. SD: standard deviation.

Region	Comparison	Sign of log <sub>2</sub> (TE)	# genes	Color
1	Under-estimated by RPKM	Negative	629	Green
2	Similar	Negative	1846	Gray
3	Over-estimated by RPKM	Negative	79	Orange
4	Under-estimated by RPKM	Positive	268	Green
5	Similar	Positive	1305	Gray
6	Over-estimated by RPKM	Positive	981	Orange

**Supplemental table 5.S2. Interpretation of the pair-wise comparison in Figure 4A.** Related to Figure 5.4; The sign of log(TE) are based on TE of Scikit-ribo.  $\Delta \log_2(TE) = \log_2(TE_{scikit-ribo}) - \log_2(TE_{RPKM})$ . For gene with  $\Delta \log_2(TE) < -0.5$ , they were previously underestimated by RPKM-derived TE, and genes with  $\Delta \log_2(TE) > 0.5$  were previously overestimated, and other genes have similar TE.

GO Term	Accession #	p-value	# genes
cytoplasmic translation	GO:0002181	$3 \times 10^{-25}$	49
translational elongation	GO:0006414	$1 \times 10^{-8}$	59
ribosome assembly	GO:0042255	$2 \times 10^{-6}$	19
translation	GO:0006412	$3 \times 10^{-6}$	63
peptide biosynthetic process	GO:0043043	$4 \times 10^{-6}$	63

**Supplemental Table 5.S3. Gene set enrichment in region 4 genes.** Related to Figure 5.4; There were 268 genes in region 4: 1) positive Scikit-ribo  $\log_2(\text{TE})$ , 2) previously underestimated by RPKM derived TE. The p-values shown were adjusted with Bonferroni correction.



codon	DT_W T	DT_K O	DT_O E	RA V	DT_WT rank	log2_ratio_OE VS_WT	log2_ratio_KO VS_WT	log2_ratio_OE VS_KO
CAT	0.60	0.64	0.79	0.19	1.00	0.38	0.08	0.30
CAA	0.64	0.70	0.72	0.55	2.00	0.18	0.12	0.05
ACC	0.68	0.71	0.75	0.49	3.00	0.12	0.06	0.07
AAC	0.69	0.82	0.81	0.62	4.00	0.23	0.25	-0.02
TTA	0.70	0.76	0.90	0.43	5.00	0.37	0.12	0.25
ATC	0.70	0.86	0.79	0.58	6.00	0.18	0.31	-0.13
AAT	0.73	0.79	0.86	0.27	7.00	0.24	0.11	0.13
ACT	0.75	0.68	0.77	0.68	8.00	0.04	-0.14	0.17
ATT	0.76	0.84	0.83	0.80	9.00	0.13	0.14	-0.01
CAC	0.80	0.73	0.88	0.43	10.00	0.14	-0.12	0.26
TGT	0.80	0.73	1.00	0.11	11.00	0.31	-0.14	0.45
TTT	0.81	0.79	0.90	0.27	12.00	0.16	-0.04	0.20
TTG	0.83	0.79	0.94	0.75	13.00	0.18	-0.06	0.24
TCA	0.84	1.00	0.97	0.19	14.00	0.20	0.25	-0.05
GTT	0.85	0.84	0.79	0.86	15.00	-0.09	0.00	-0.09
TCT	0.85	0.79	0.83	0.68	16.00	-0.03	-0.10	0.07
AAG	0.88	0.93	0.81	1.00	17.00	-0.12	0.09	-0.20
GTC	0.90	0.94	0.82	0.62	18.00	-0.13	0.07	-0.19
AAA	0.90	0.97	0.91	0.43	19.00	0.02	0.10	-0.08
AGT	0.91	0.93	1.08	0.05	20.00	0.24	0.02	0.22
ATG	0.92	0.91	0.90	0.62	21.00	-0.03	-0.01	-0.02
TCC	0.93	0.90	0.91	0.49	22.00	-0.03	-0.05	0.02
TAT	0.94	1.16	0.95	0.22	23.00	0.01	0.31	-0.30
AGC	0.94	1.04	1.12	0.12	24.00	0.25	0.15	0.11
GAT	0.95	0.86	0.90	0.43	25.00	-0.08	-0.14	0.06
GCC	0.96	0.94	0.80	0.49	26.00	-0.26	-0.03	-0.24
ACA	0.97	0.97	1.08	0.25	27.00	0.15	0.00	0.15
TTC	0.98	0.92	0.83	0.62	28.00	-0.24	-0.08	-0.15
CTT	0.99	0.95	1.23	0.03	29.00	0.32	-0.05	0.36
CGT	1.00	0.99	0.97	0.37	30.00	-0.03	-0.01	-0.02
GCT	1.00	0.89	0.77	0.68	31.00	-0.38	-0.16	-0.22
GAC	1.00	0.98	0.95	0.99	32.00	-0.08	-0.04	-0.04
ATA	1.02	1.33	1.11	0.12	33.00	0.13	0.38	-0.26
GTA	1.02	1.02	1.04	0.12	34.00	0.03	-0.01	0.04
AGA	1.07	1.18	1.04	0.68	35.00	-0.05	0.13	-0.18
CTA	1.07	0.99	1.13	0.19	36.00	0.07	-0.12	0.19

<b>CAG</b>	1.08	1.16	1.08	0.24	37.00	0.00	0.10	-0.10
<b>GGT</b>	1.08	1.04	0.97	0.43	38.00	-0.16	-0.06	-0.10
<b>GAA</b>	1.13	1.10	0.93	0.86	39.00	-0.29	-0.03	-0.25
<b>TGC</b>	1.13	0.99	1.35	0.25	40.00	0.25	-0.20	0.45
<b>ACG</b>	1.20	1.09	1.17	0.14	41.00	-0.03	-0.14	0.10
<b>CTC</b>	1.20	1.26	1.49	0.06	42.00	0.31	0.07	0.24
<b>TAC</b>	1.23	1.41	1.01	0.49	43.00	-0.28	0.19	-0.48
<b>GCA</b>	1.24	1.29	1.02	0.31	44.00	-0.27	0.06	-0.34
<b>GAG</b>	1.24	1.21	1.02	0.40	45.00	-0.28	-0.03	-0.25
<b>GTG</b>	1.29	1.06	1.10	0.16	46.00	-0.23	-0.28	0.04
<b>TCG</b>	1.32	1.25	1.25	0.12	47.00	-0.08	-0.08	0.01
<b>GCG</b>	1.36	1.28	1.12	0.10	48.00	-0.28	-0.08	-0.20
<b>CCT</b>	1.37	1.38	1.05	0.12	49.00	-0.39	0.01	-0.41
<b>CCA</b>	1.38	1.44	1.01	0.62	50.00	-0.45	0.07	-0.51
<b>GGC</b>	1.40	1.39	1.15	0.99	51.00	-0.29	-0.02	-0.27
<b>AGG</b>	1.41	1.59	1.37	0.28	52.00	-0.04	0.18	-0.22
<b>TGG</b>	1.41	1.25	1.34	0.37	53.00	-0.08	-0.17	0.10
<b>CGC</b>	1.50	1.56	1.38	0.27	54.00	-0.12	0.07	-0.18
<b>GGG</b>	1.51	1.61	1.25	0.18	55.00	-0.28	0.09	-0.37
<b>CTG</b>	1.57	1.34	1.71	0.06	56.00	0.12	-0.22	0.35
<b>GGA</b>	1.68	1.58	1.42	0.19	57.00	-0.24	-0.09	-0.15
<b>CGG</b>	1.77	1.43	1.61	0.06	58.00	-0.14	-0.31	0.18
<b>CGA</b>	1.88	2.36	2.07	0.00	59.00	0.14	0.33	-0.19
<b>CCC</b>	1.96	1.72	1.22	0.09	60.00	-0.68	-0.18	-0.50
<b>CCG</b>	2.62	3.69	1.86	0.20	61.00	-0.50	0.50	-0.99

**Supplemental Table 5.S4. Relative codon elongation rate (ER) and dwell time (DT) in the Dhh1p study.** Related to Figure 5.6; RAV: relative adaptation value.

GO Term	Accession #	p-value	# genes
<b>cytosolic ribosome</b>	[GO:0022626]	$3 \times 10^{-16}$	20
<b>cytosolic part</b>	[GO:0044445]	$1 \times 10^{-14}$	21
<b>ribosome</b>	[GO:0005840]	$2 \times 10^{-12}$	20
<b>ribosomal subunit</b>	[GO:0044391]	$5 \times 10^{-12}$	19
<b>cytosolic small ribosomal subunit</b>	[GO:0022627]	$6 \times 10^{-11}$	12
<b>small ribosomal subunit</b>	[GO:0015935]	$9 \times 10^{-9}$	12
Pathway		p-value	# genes
<b>glucose fermentation</b>		$4 \times 10^{-3}$	5

**Supplemental Table 5.S5. The GO enrichment of gene with reduced TE in OE, relative to WT in the Dhh1p analysis.** Related to Figure 5.6; 50 genes the most changes in TE were used as inputs for the GO enrichment analysis. The p-values shown were adjusted with Bonferroni correction.

GO Term	Accession #	p-value	# genes
<b>inner mitochondrial membrane protein complex</b>	[GO:0098800]	$3 \times 10^{-4}$	8
<b>cytochrome complex</b>	[GO:0070069]	$1 \times 10^{-3}$	5
<b>mitochondrial protein complex</b>	[GO:0098798]	$2 \times 10^{-3}$	8
<b>mitochondrial part</b>	[GO:0044429]	$3 \times 10^{-3}$	17
<b>mitochondrial respiratory chain</b>	[GO:0005746]	$4 \times 10^{-3}$	5
<b>respiratory chain complex</b>	[GO:0098803]	$5 \times 10^{-3}$	5
<b>ATP metabolic process</b>	[GO:0046034]	$5 \times 10^{-3}$	8
<b>purine ribonucleoside triphosphate metabolic process</b>	[GO:0009205]	$7 \times 10^{-3}$	8
<b>hydrogen ion transmembrane transport</b>	[GO:1902600]	$7 \times 10^{-3}$	7
<b>purine nucleoside triphosphate metabolic process</b>	[GO:0009144]	$8 \times 10^{-3}$	8
<b>Pathway</b>		<b>p-value</b>	<b># genes</b>
<b>aerobic respiration, electron transport chain</b>		$1 \times 10^{-2}$	5

**Supplemental Table 5.S6. The GO enrichment of gene with increased TE in KO, relative to WT in the Dhh1p analysis.** Related to Figure 5.6; 50 genes the most changes in TE were used as inputs for the GO enrichment analysis. The p-values shown were adjusted with Bonferroni correction.

## Chapter 6 Applications on genome informatics

### Summary of contribution

This chapter describes the applications on genome informatics. The first application and analyses were published in *BMC Medical Genomics*<sup>218</sup>. Han Fang developed the computational pipeline for analyzing the whole genome sequencing data. The second application and methods were published in *Bioinformatics*<sup>219</sup>. Michael Schatz led the development of the mixture model to estimate the mutation rate, duplication rate, and genome size. Han Fang further refined the prior probabilities and the mixture model. Permission for republication of this material has been granted and is available upon request.

### Applications

#### Application 1 - Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine

Human Phenotype Ontology (HPO) has risen as a useful tool for precision medicine by providing a standardized vocabulary of phenotypic abnormalities to describe presentations of human pathologies; however, there have been relatively few reports combining whole genome sequencing (WGS) and HPO, especially in the context of structural variants. We illustrate an integrative analysis of WGS and HPO using an extended pedigree, which involves Prader-Willi Syndrome (PWS), hereditary hemochromatosis (HH), and dysautonomia-like symptoms. A comprehensive WGS pipeline was used to ensure reliable detection of genomic variants. Beyond variant filtering, we pursued phenotypic prioritization of candidate genes using Phenolyzer. Regarding PWS, WGS confirmed a 5.5Mb *de novo* deletion of the parental allele at 15q11.2 to 15q13.1 (**Figure 6.1**). Phenolyzer successfully returned the diagnosis of PWS, and pinpointed clinically relevant genes in the deletion. Further, Phenolyzer revealed how each of the genes is linked with the phenotypes represented by HPO terms. For HH, WGS identified a known disease variant (p.C282Y) in HFE of an affected female. Analysis of HPO terms alone fails to provide a correct diagnosis, but Phenolyzer successfully revealed the phenotype-genotype relationship using a disease-centric approach. Finally, Phenolyzer also revealed the complexity behind dysautonomia-like symptoms, and seven variants that might be associated with the phenotypes were identified by manual filtering based on a dominant inheritance model. The integration of WGS and HPO can inform comprehensive molecular diagnosis for patients, eliminate false positives and reveal novel insights into undiagnosed diseases. Due to extreme heterogeneity and insufficient knowledge of human diseases, it is also important that phenotypic and genomic data are standardized and shared simultaneously.

ERDS and CNVnator both detected three *de novo* heterozygous deletions with a total size of about 5.5 Mb, in the chromosome regions from 15q11.2 to 15q13.1 of the proband with PWS (K10031-10232). The hg19 genomic coordinates of the breakpoints are chr15:22,749,401-23,198,800 (~449 Kb), chr15:23,608,601-28,566,000 (~4.96 Mb), and chr15:28,897,601-28,992,600 (~95 Kb). Notably, these deletions are relatively close to one another; the distances between each deletion are ~410 Kb and ~332 Kb, respectively. Within the regions containing the *de novo* deletions, the depth of coverage in the proband's genome is 20X, about half of the

genome-wide mean coverage (40X). Due to the lack of the proband's mother's sequencing data (as she refused to participate), analysis was performed to determine which allele (paternal or maternal) is deleted. This can be inferred through SNVs where the mendelian inheritance law is violated; meaning those instances in which the proband (K10031-10232) does not carry certain paternal or maternal SNVs that his brother (K10031-10233) does carry. In total, there are 2,987 SNVs where the proband's father (K10031-10231) is a homozygote and the proband's brother is a heterozygote. Out of the 2112 SNVs where the father is homozygous to the reference allele, the proband is homozygous to the alternative allele at 1944 loci (92%, **Figure 6.1**). Among 875 SNVs where the father does not carry any reference allele, the proband carries only the reference allele at 861 SNVs (94%, **Figure 6.1**). This indicates that the proband only carries the maternal alleles in those regions. These deletions were not detected in either the proband's father or his brother using the WGS data. The Illumina microarray data further confirmed this discovery; the proband carries these deletions while his father and his brothers (K10031-10233 and K10031-10234) do not carry any of these deletions in their genome. Probe distributions of Log-R ratios and B allele frequencies are not uniform in the microarray because the density of SNV varies between genomic regions. This highlights the higher resolution and completeness of WGS over microarray for precise molecular diagnosis of such diseases. Thus, we confirm that the proband carries the *de novo* PWS Type I deletion (spanning breakpoints BP1 and BP3) defined by previous publications<sup>220, 221</sup>.

## Application 2 - GenomeScope: fast reference-free genome profiling from short reads

GenomeScope is an open-source web tool to rapidly estimate the overall characteristics of a genome, including genome size, heterozygosity rate and repeat content from unprocessed short reads. These features are essential for studying genome evolution, and help to choose parameters for downstream analysis. We demonstrate its accuracy on 324 simulated and 16 real datasets with a wide range in genome sizes, heterozygosity levels and error rates. Availability and Implementation: <http://genomescope.org>, <https://github.com/schatzlab/genomescope.git>.

The full GenomeScope model builds on this analysis to consider the interplay between heterozygosity, repeats, sequencing depth, and sequencing biases. Central to our method is the mixture model to describe the impact of heterozygosity and repeats on the *k-mer* distribution:

$$f(X; \alpha, \beta, \gamma, \delta, \lambda, \rho, G) = G \cdot \left[ \alpha NB \left( X; \lambda, \frac{\lambda}{\rho} \right) + \beta NB \left( X; 2\lambda, \frac{2\lambda}{\rho} \right) + \gamma NB \left( X; 3\lambda, \frac{3\lambda}{\rho} \right) + \delta NB \left( X; 4\lambda, \frac{4\lambda}{\rho} \right) \right] \quad \text{Equation 6.1}$$

where

*G* is a scaling parameter w. r. t the genome size,

$\alpha, \beta, \gamma, \delta$  are the prior probability (weights) for each distribution,

$\lambda$  is the mean of the first Negative binomial distribution,

$\rho$  is the variance parameter of the Negative binomial distribution

We further determined the coefficients  $\alpha, \beta, \gamma, \delta$  were related to the underlying genomic properties through the following system of equations (**Equation 6.2-6.6**, See also **Figure 6.2**). The parameter  $r$  is the mutation date, and  $k$  is the duplication rate. The sum of  $\alpha, \beta, \gamma, \delta$  will be greater than 1 if there is a non-zero rate of heterozygosity. This is because introducing heterozygosity will create new *k-mers* relative to the haploid genome length  $G$  similar to what is described above for repeat-free genomes. The maximum value of the sum may be as large as

3\*G at extreme rates of duplication and heterozygosity, consisting of G k-mers from the (duplicated) maternal haplotype, and 2\*G heterozygous *k-mers* from the paternal haplotype. The four coefficients can be scaled by  $\frac{1}{(\alpha+\beta+\gamma+\delta)}$  so that they will sum to 1 and form a proper probability distribution for the mixture model. Equivalently, GenomeScope infers a value for G which has been scaled by  $\frac{1}{(\alpha+\beta+\gamma+\delta)}$ .

$$\alpha = 2(1 - d)[1 - (1 - r)^k] + 2d[1 - (1 - r)^k]^2 + 2d[(1 - r)^k][1 - (1 - r)^k] \quad \text{Equation6.2}$$

$$\beta = (1 - d)[(1 - r)^k] + d[1 - (1 - r)^k]^2 \quad \text{Equation6.3}$$

$$\gamma = 2d[(1 - r)^k][1 - (1 - r)^k] \quad \text{Equation6.4}$$

$$\delta = d[(1 - r)^{2k}] \quad \text{Equation6.5}$$

## Figures and tables in this chapter

### Figures

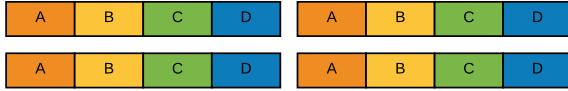


**Figure 6.1** Screenshot of three heterozygous *de novo* deletions between the region 15q11.2 to 15q13 in proband K10031-10232. The deleted regions are denoted by the red boxes. The yellow tagging SNVs represent the SNVs that violate the Mendelian inheritance law. The non-deleted regions are denoted by the green tagging SNVs. Genome-wide average coverage (40X) is denoted by the grey dashed line. The breakpoints of these deletions (PWS Type I deletion) are chr15:22,749,401-23,198,800 (~449 Kb), chr15:23,608,601-28,566,000 (~4.96 Mb), and chr15:28,897,601-28,992,600 (95 Kb) (hg19). These deletions are not detected either in the proband's father or the unaffected brother.



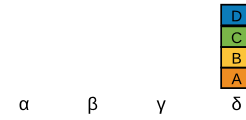
## Heterozygosity categories

Duplicated homozygous case:

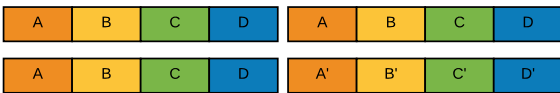


total contribution to  $\delta$  peak:  $d(1-r)^{(2k)}$

## Contribution to k-mer profile



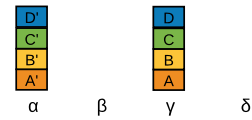
Duplicated homozygous and one heterozygous case:



OR



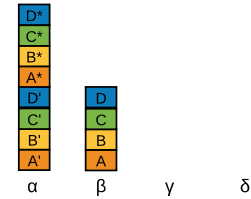
total contribution to  $\alpha$  peak  $2d(1-r)^k(1-(1-r)^k)$  and  $\gamma$  peak  $2d((1-r)^k)(1-(1-r)^k)$



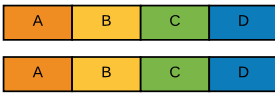
Duplicated heterozygous case:



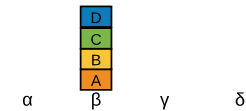
total contribution to  $\alpha$  peak  $2d(1-(1-r)^k)^2$  and  $\beta$  peak  $d(1-(1-r)^k)^2$



Unique homozygous case:



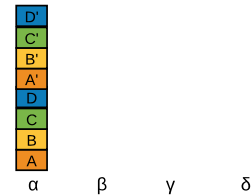
total contribution to  $\beta$  peak:  $(1-d)((1-r)^k)$



Unique heterozygous case:



total contribution to  $\alpha$  peak:  $2(1-d)(1-(1-r)^k)$



Legend:

kmer: X mutated kmer: X' X\*

**Figure 6.2. GenomeScope Heterozygosity Categories.** The figure shows how duplications and heterozygosity impacts the  $k$ -mer profile by contributing  $k$ -mers to the four possible peaks.

## Chapter 7 Conclusions and Perspectives

As described in the central dogma of molecular biology, biological information flows from DNA to RNA, and from RNA to protein. To understand the relationship between genotype and phenotype, it is essential to develop methods for decrypting the information hidden in the sequences at each step. Thanks to the advancement in the sequencing technologies, high throughput readout of the nucleic acid sequence information has become routine in research laboratories and soon for clinical laboratories. In particular, DNA sequencing has revolutionized the way to study a person's genome. More and more research groups are now able to investigate not only SNPs, but also small indels and large structural variants. Although many groups had demonstrated successful approaches for SNP calling, indel calling remained challenging until recently, because the error model of indels is quite different and requires new analysis methods. From Chapter 2 to Chapter 4, I presented a set of methods to investigate its error model and improve indel calling.

Knowing the genotype is only one part of the process to decrypt the genome. We also need to quantify the effects of gene regulation in order to understand the sequence function. The development of RNAseq enables researchers to study gene expression levels genome-wide. This led to many novel discoveries on biological processes of transcription regulation. However, another significant part of the central dogma, translation, has received substantially less attention and has ample opportunities for development and study. Riboseq data provide new opportunities for studying translation regulation genome-wide, but there are prevalent sampling errors and biological biases that are unknown. Compared with RNAseq, there is a limited number of statistical methods developed for Riboseq. In Chapter 5, I develop a software package for Riboseq analysis, as well as providing detailed characterizations of the common errors in Riboseq data.

### Conclusions and contributions of this thesis

In Chapter 2, the development of the Scalpel algorithm improves our ability to identify indels from short-read sequencing data. This chapter describes the algorithm involving the de Bruijn graph assembly. The protocol in this chapter provides a standardized workflow for analyzing indel variants, making the analysis more reproducible. In Chapter 3, I investigated the sources of indel errors from whole genome sequencing and exome sequencing data. It illustrates the existing challenges that are intrinsic to different properties of data. This chapter highlights the key factors of indel calling, such as algorithmic artifacts, coverage requirement, library preparation, and etc. For the first time, a detailed comparison and validation on “whole genome sequencing vs exome capture sequencing” on indel calling is described. An important finding about how PCR amplification and homopolymer runs create many false positives is also discussed in this chapter.

In Chapter 4, I demonstrated the accuracy of Scalpel via a three-way benchmarking against other start-of-the-art methods. Further, this comparison also exposes limitations of those competing algorithms, prompting their developers to address the issues in later development. In terms of applications, we analyze 593 families from the Simons Simplex Collection and demonstrate Scalpel's power to detect long transmitted events, and enrichment for de novo likely

gene-disrupting indels in autistic children. The accuracy of indel calling is essential for finding the statistical significance in large-cohort and population-scale studies, which can be warranted by Scalpel.

In Chapter 6, I develop Scikit-ribo, the first statistically robust model and open-source software package for accurate genome-wide TE inference from Riboseq data. The core of Scikit-ribo is a codon-level generalized linear model that unifies our study of translation elongation and initiation including the effects of codon specific elongation rates, mRNA secondary structure, and gene specific translation initiation efficiency. When paired with a powerful ridge regression regularization method, Scikit-ribo corrects the negative skew in TE observed in most previous papers, especially for low expressed genes. Using three case studies involving ten different datasets, we showed how these statistical advancements allow universal improvement to Riboseq data analysis. This particularly improves the estimation of genome-wide TE, allowing us to discover the Kozak-like consensus sequence in *S. cerevisiae*, and yield novel insights into Dhh1p's role on translation repression.

### Applications of this research

From an algorithm point of view, Scalpel is a successful attempt and advancement on applying localized de Bruijn graph assembly for indel calling. I describe many scenarios where the algorithm might be in error if one is not careful with the distinct error models, which are natures of the high-throughput genomics data. In addition to the Simons Simplex Collection, Scalpel has been successfully used to discover mutations in novel candidate genes for large cohort studies of cancers, autism, and other important human diseases. This shed light on the complex mechanisms and biological pathways for these devastating diseases, which might help identify new drug targets and treatment methods for patients.

Our findings based on Scikit-ribo showcase the interplay between biology and statistics; biological knowledge informs statistical methods development, and statistical improvement yields novel biological insights. Together, we demonstrate that Scikit-ribo substantially improves Riboseq analysis and our understandings of translation control. In the future, we foresee more researchers applying Riboseq to address their biological questions related to protein translation and Scikit-ribo can unlock the full potential of this technique.

### Further directions

In terms of genotyping, a major challenge remains for somatic variant calling. Cancer cells even from the same tissue usually have multiple clones, in addition to multi-ploidy. This leads to the low allele frequencies of somatic variants, making it very difficult to distinguish between true signals and noises in cancer genomes. From a computational perspective, the low allele frequencies also pose new issues for constructing and enumerating de Bruijn variant graphs, as the previous assumption on diploid genomes no longer holds. Thus, the next big question is to develop an accurate and efficient algorithm based on localized graph assembly for somatic variant calling. Several groups have been working along these lines, including a new method called Lancet.

Regarding ribosome profiling, another potential improvement is to address the issues of isoform-specific expression. Because in mammalian model organisms, alternative splicing is prevalent and various isoforms might be present for the same gene. Within a gene, Riboseq

coverage might be higher for regions with higher exon usage; this will confound the analysis of local translation elongation rates. Therefore, a more comprehensive model should incorporate the mRNA expression levels of different transcripts and subsequently determine the relative expression level of different exons. This improvement will be highly valuable for the community and further deepens our understanding of translation regulation in higher order model organisms.

## Bibliography

1. Collins, F.S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* 372, 793-795 (2015).
2. Highnam, G. et al. An analytical framework for optimizing variant discovery from personal genomes. *Nat Commun* 6, 6275 (2015).
3. Ingolia, N.T., Ghaemmighami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218-223 (2009).
4. Miyagi, M. & Rao, K.C. Proteolytic 18O-labeling strategies for quantitative proteomics. *Mass Spectrom Rev* 26, 121-136 (2007).
5. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. & Weissman, J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* 7, 1534-1550 (2012).
6. Nik-Zainal, S. et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* 149, 979-993 (2012).
7. Zaidi, S. et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 498, 220-223 (2013).
8. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285-299 (2012).
9. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216-221 (2014).
10. MacArthur, D.G. et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* 335, 823-828 (2012).
11. Fukuoka, S. et al. Loss of function of a proline-containing protein confers durable disease resistance in rice. *Science* 325, 998-1001 (2009).
12. Denver, D.R., Morris, K., Lynch, M. & Thomas, W.K. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* 430, 679-682 (2004).
13. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491-498 (2011).
14. Albers, C.A. et al. Dindel: accurate indel calls from short-read data. *Genome Res* 21, 961-973 (2011).
15. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079 (2009).
16. Luo, R., Schatz, M.C. & Salzberg, S. 16GT: a fast and sensitive variant caller using a 16-genotype probabilistic model. *bioRxiv* (2017).
17. Fang, H. et al. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* 6, 89 (2014).
18. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865-2871 (2009).
19. Karakoc, E. et al. Detection of structural variants and indels within exome data. *Nat Methods* 9, 176-178 (2012).

20. Narzisi, G. & Schatz, M.C. The challenge of small-scale repeats for indel discovery. *Front Bioeng Biotechnol* 3, 8 (2015).
21. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 46, 912-918 (2014).
22. Cameron, D.L. et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *bioRxiv* (2017).
23. Wala, J. et al. Genome-wide detection of structural variants and indels by local assembly. *bioRxiv* (2017).
24. de Bruijn, N.G. A Combinatorial Problem. *Koninklijke Nederlandsche Akademie Van Wetenschappen* 49, 758-764 (1946).
25. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* 98, 9748-9753 (2001).
26. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18, 821-829 (2008).
27. Li, D., Liu, C.M., Luo, R., Sadakane, K. & Lam, T.W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674-1676 (2015).
28. Ribeiro, F.J. et al. Finished bacterial genomes from shotgun sequence data. *Genome Res* 22, 2270-2277 (2012).
29. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 44, 226-232 (2012).
30. Crick, F. Central dogma of molecular biology. *Nature* 227, 561-563 (1970).
31. Ingolia, N.T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* 15, 205-213 (2014).
32. Sendoel, A. et al. Translation from unconventional 5' start sites drives tumour initiation. *Nature* 541, 494-499 (2017).
33. Hsieh, A.C. et al. The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature* 485, 55-61 (2012).
34. Wurth, L. et al. UNR/CSDE1 Drives a Post-transcriptional Program to Promote Melanoma Invasion and Metastasis. *Cancer Cell* 30, 694-707 (2016).
35. Goodarzi, H. et al. Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. *Cell* 165, 1416-1427 (2016).
36. Schafer, S. et al. Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat Commun* 6, 7200 (2015).
37. Wein, N. et al. Translation from a DMD exon 5 IRES results in a functional dystrophin isoform that attenuates dystrophinopathy in humans and mice. *Nat Med* 20, 992-1000 (2014).
38. Su, X. et al. Interferon-gamma regulates cellular metabolism and mRNA translation to potentiate macrophage activation. *Nat Immunol* 16, 838-849 (2015).
39. Thoreen, C.C. et al. A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature* 485, 109-113 (2012).
40. Brar, G.A. & Weissman, J.S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 16, 651-664 (2015).
41. Michel, A.M. & Baranov, P.V. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip Rev RNA* 4, 473-490 (2013).
42. Li, G.W. How do bacteria tune translation efficiency? *Curr Opin Microbiol* 24, 66-71 (2015).

43. Nelder, J.A. & Wedderburn, R.W.M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* 135, 370-384 (1972).
44. Burrus, C.S. Iterative Reweighted Least Squares.
45. Hawkins, D.M. The problem of overfitting. *J Chem Inf Comp Sci* 44, 1-12 (2004).
46. Hoerl, A.E. & Kennard, R.W. Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 55-& (1970).
47. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* 58, 267-288 (1996).
48. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). *J R Stat Soc B* 67, 768-768 (2005).
49. Narzisi, G. et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat Meth* 11, 1033-1036 (2014).
50. Fang, H. et al. Indel variant analysis of short-read sequencing data with Scalpel. *Nat Protoc* 11, 2529-2548 (2016).
51. Narzisi, G. et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat Methods* 11, 1033-1036 (2014).
52. Pabinger, S. et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 15, 256-278 (2014).
53. Van der Auwera, G.A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 11, 11 10 11-11 10 33 (2013).
54. Li, S. et al. SOAPindel: efficient identification of indels from short paired reads. *Genome Res* 23, 195-200 (2013).
55. Mose, L.E., Wilkerson, M.D., Hayes, D.N., Perou, C.M. & Parker, J.S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* 30, 2813-2815 (2014).
56. Chen, K. et al. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res* 24, 310-317 (2014).
57. Weisenfeld, N.I. et al. Comprehensive variation discovery in single human genomes. *Nat Genet* 46, 1350-1355 (2014).
58. Leggett, R.M. et al. Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs. *PLoS One* 8, e60058 (2013).
59. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* (2015).
60. Yang, R., Nelson, A.C., Henzler, C., Thyagarajan, B. & Silverstein, K.A. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. *Genome Med* 7, 127 (2015).
61. Ye, K. et al. Systematic discovery of complex insertions and deletions in human cancers. *Nat Med* 22, 97-104 (2016).
62. Paila, U., Chapman, B.A., Kirchner, R. & Quinlan, A.R. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput Biol* 9, e1003153 (2013).
63. Brannon, A.R. et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol* 15, 454 (2014).
64. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat Rev Genet* 14, 157-167 (2013).

65. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* 147, 195-197 (1981).
66. Medvedev, P., Georgiou, K., Myers, G. & Brudno, M. Computability of models for sequence assembly. *Lect N Bioinformat* 4645, 289-301 (2007).
67. Narzisi, G. & Mishra, B. Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons. *Bioinformatics* 27, 153-160 (2011).
68. Genomes Project, C. et al. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073 (2010).
69. Montgomery, S.B. et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* 23, 749-761 (2013).
70. Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108, 1513-1518 (2011).
71. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* 33, 623-630 (2015).
72. Li, H. in *ArXiv e-prints*, Vol. 1303 3997 (2013).
73. Gudmundsson, J. et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature genetics* 44, 1326-1329 (2012).
74. Rope, A.F. et al. Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *American journal of human genetics* 89, 28-43 (2011).
75. Biesecker, L.G. & Green, R.C. Diagnostic Clinical Genome and Exome Sequencing. *New England Journal of Medicine* 370, 2418-2425 (2014).
76. Patel, C.J. et al. Whole genome sequencing in support of wellness and health maintenance. *Genome Med* 5, 58 (2013).
77. O'Rawe, J.A. et al. Integrating precision medicine in the study and clinical treatment of a severely mentally ill person. *PeerJ* 1, e177 (2013).
78. Chen, R. et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293-1307 (2012).
79. Hood, L. & Rowen, L. The human genome project: big science transforms biology and medicine. *Genome Med* 5, 79 (2013).
80. Tarczy-Hornoch, P. et al. A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record. *Genet Med* 15, 824-832 (2013).
81. Lyon, G.J. & Wang, K. Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome medicine* 4, 58-58 (2012).
82. O'Rawe, J. et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 5, 28 (2013).
83. Dewey, F.E. et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA* 311, 1035-1045 (2014).
84. Zook, J.M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotech* 32, 246-251 (2014).
85. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 46, 912-918 (2014).
86. Lupski, J.R., Belmont, J.W., Boerwinkle, E. & Gibbs, R.A. Clan genomics and the complex architecture of human disease. *Cell* 147, 32-43 (2011).
87. Lyon, G.J. & O'Rawe, J. in *The Genetics of Neurodevelopmental Disorders* (ed. K. Mitchell) (Cold Spring Harbor Labs Journals, 2014).



88. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* 141, 210-217 (2010).
89. Ober, C. & Vercelli, D. Gene-environment interactions in human disease: nuisance or opportunity? *Trends in genetics* : TIG 27, 107-115 (2011).
90. Clerget-Darpoux, F. & Elston, R.C. Will formal genetics become dispensable? *Hum Hered* 76, 47-52 (2013).
91. Weiss, K.M. & Terwilliger, J.D. How many diseases does it take to map a gene with SNPs? *Nat Genet* 26, 151-157 (2000).
92. Lyon, G.J. Personalized medicine: Bring clinical standards to human-genetics research. *Nature* 482, 300-301 (2012).
93. MacArthur, D.G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469-476 (2014).
94. Ross, M.G. et al. Characterizing and measuring bias in sequence data. *Genome Biol* 14, R51 (2013).
95. Clark, M.J. et al. Performance comparison of exome DNA sequencing technologies. *Nature biotechnology* 29, 908-914 (2011).
96. Lam, H.Y. et al. Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* 30, 78-82 (2012).
97. Linderman, M. et al. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Medical Genomics* 7, 20 (2014).
98. Bamshad, M.J. et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics* 12, 745-755 (2011).
99. Bamshad, M.J. et al. The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. *Am J Med Genet A* 158A, 1523-1525 (2012).
100. Lam, H. et al. Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* 30, 78 - 82 (2012).
101. Cech, Thomas R. & Steitz, Joan A. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell* 157, 77-94.
102. Li, S. & Mason, C.E. The Pivotal Regulatory Landscape of RNA Modifications. *Annual Review of Genomics and Human Genetics* 15, 127-150 (2014).
103. Metzker, M.L. Sequencing technologies - the next generation. *Nature reviews. Genetics* 11, 31-46 (2010).
104. Zhu, M. et al. Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet* 91, 408-421 (2012).
105. Meynert, A., Ansari, M., FitzPatrick, D. & Taylor, M. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* 15, 247 (2014).
106. Mullaney, J.M., Mills, R.E., Pittard, W.S. & Devine, S.E. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* 19, R131-136 (2010).
107. Mills, R.E. et al. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* 21, 830-839 (2011).
108. Mills, R.E. et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research* 16, 1182-1190 (2006).
109. Li, H. in *ArXiv e-prints*, Vol. 1404 929 (2014).
110. Highnam, G. et al. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research* 41, e32 (2013).

111. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 44, 226-232 (2012).
112. Narzisi, G. et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat Meth* advance online publication (2014).
113. Toyama, B.H. et al. Identification of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell* 154, 971-982 (2013).
114. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491-498 (2011).
115. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842 (2010).
116. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research* (2012).
117. Willems, T.F. et al. The landscape of human STR variation. *Genome Research* (2014).
118. García-Alcalde, F. et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28, 2678-2679 (2012).
119. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9, 90-95 (2007).
120. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* 9, 357-359 (2012).
121. Van der Auwera, G.A. et al. in *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2002).
122. Kidd, J.M. et al. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat Meth* 7, 365-371 (2010).
123. Ajay, S.S., Parker, S.C., Abaan, H.O., Fajardo, K.V. & Margulies, E.H. Accurate and comprehensive sequencing of personal genomes. *Genome Res* 21, 1498-1505 (2011).
124. McKernan, K.J. et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research* 19, 1527-1541 (2009).
125. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872-876 (2008).
126. Zook, J.M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 32, 246-251 (2014).
127. Pearson, C.E., Nichol Edamura, K. & Cleary, J.D. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* 6, 729-742 (2005).
128. Karakoc, E. et al. Detection of structural variants and indels within exome data. *Nat Methods* 9, 176-178 (2011).
129. Li, H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* 28, 1838-1844 (2012).
130. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* 22, 1154-1162 (2012).
131. MacArthur, D.G. & Tyler-Smith, C. Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet* 19, R125-130 (2010).
132. Sanders, S.J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237-241 (2012).
133. O'Roak, B.J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246-250 (2012).

134. Neale, B.M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485, 242-245 (2012).
135. Darnell, J.C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 146, 247-261 (2011).
136. Weinberg, D.E. et al. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep* 14, 1787-1799 (2016).
137. Gerashchenko, M.V. & Gladyshev, V.N. Ribonuclease selection for ribosome profiling. *Nucleic Acids Res* 45, e6 (2017).
138. Gerashchenko, M.V. & Gladyshev, V.N. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res* 42, e134 (2014).
139. Oh, E. et al. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* 147, 1295-1308 (2011).
140. Lee, S. et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* 109, E2424-2432 (2012).
141. Archer, S.K., Shirokikh, N.E., Beilharz, T.H. & Preiss, T. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. *Nature* 535, 570-574 (2016).
142. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J.B. Rate-limiting steps in yeast protein translation. *Cell* 153, 1589-1601 (2013).
143. Zhang, S. et al. ROSE: a deep learning based framework for predicting ribosome stalling. *bioRxiv* (2016).
144. Mohammad, F., Woolstenhulme, C.J., Green, R. & Buskirk, A.R. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep* 14, 686-694 (2016).
145. Radhakrishnan, A. et al. The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell* 167, 122-132 e129 (2016).
146. Quax, T.E., Claassens, N.J., Soll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell* 59, 149-161 (2015).
147. Quax, T.E. et al. Differential translation tunes uneven production of operon-encoded proteins. *Cell Rep* 4, 938-944 (2013).
148. Lecanda, A. et al. Dual randomization of oligonucleotides to reduce the bias in ribosome-profiling libraries. *Methods* 107, 89-97 (2016).
149. Hussmann, J.A., Patchett, S., Johnson, A., Sawyer, S. & Press, W.H. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet* 11, e1005732 (2015).
150. Wang, H., McManus, J. & Kingsford, C. Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast. *J Comput Biol* (2016).
151. Wolin, S.L. & Walter, P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J* 7, 3559-3569 (1988).
152. Hwang, J.Y. & Buskirk, A.R. A ribosome profiling study of mRNA cleavage by the endonuclease RelE. *Nucleic Acids Res* 45, 327-336 (2017).
153. Hsu, P.Y. et al. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc Natl Acad Sci U S A* (2016).
154. Gonzalez, C. et al. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J Neurosci* 34, 10924-10936 (2014).

155. Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R. & Weissman, J.S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *Elife* 2, e01179 (2013).
156. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).
157. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139-140 (2010).
158. Pimentel, H.J., Bray, N., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv* (2016).
159. Albert, F.W., Muzzey, D., Weissman, J.S. & Kruglyak, L. Genetic influences on translation in yeast. *PLoS Genet* 10, e1004692 (2014).
160. Csardi, G., Franks, A., Choi, D.S., Airoidi, E.M. & Drummond, D.A. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet* 11, e1005206 (2015).
161. Schuller, A.P., Wu, C.C.-C., Dever, T.E., Buskirk, A.R. & Green, R. eIF5A Functions Globally in Translation Elongation and Termination. *Molecular Cell* (2017).
162. Woolstenhulme, C.J., Guydosh, N.R., Green, R. & Buskirk, A.R. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep* 11, 13-21 (2015).
163. Thanaraj, T.A. & Argos, P. Ribosome-mediated translational pause and protein domain organization. *Protein Sci* 5, 1594-1612 (1996).
164. Doma, M.K. & Parker, R. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature* 440, 561-564 (2006).
165. Chen, C. et al. Dynamics of translation by single ribosomes through mRNA secondary structures. *Nat Struct Mol Biol* 20, 582-588 (2013).
166. Mortimer, S.A., Kidwell, M.A. & Doudna, J.A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* 15, 469-479 (2014).
167. Gorochoowski, T.E., Ignatova, Z., Bovenberg, R.A. & Roubos, J.A. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res* 43, 3022-3032 (2015).
168. Pop, C. et al. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol* 10, 770 (2014).
169. Dao Duc, K. & Song, Y.S. Identification and quantitative analysis of the major determinants of translation elongation rate variation. *bioRxiv* (2017).
170. Li, G.W., Oh, E. & Weissman, J.S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484, 538-541 (2012).
171. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* 32, 5036-5044 (2004).
172. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* 6, 26 (2011).
173. Mao, Y., Liu, H., Liu, Y. & Tao, S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in *Saccharomyces cerevisiae*. *Nucleic Acids Res* 42, 4813-4822 (2014).
174. Zur, H. & Tuller, T. Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution. *Nucleic Acids Res* 44, 9031-9049 (2016).
175. Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. *Mol Syst Biol* 7, 481 (2011).

176. Sabi, R. & Tuller, T. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Res* 21, 511-526 (2014).
177. Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* 15, 8125-8148 (1987).
178. Hamilton, R., Watanabe, C.K. & de Boer, H.A. Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs. *Nucleic Acids Res* 15, 3581-3593 (1987).
179. Raj, A. et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* 5 (2016).
180. Michel, A.M., Andreev, D.E. & Baranov, P.V. Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. *BMC Bioinformatics* 15, 380 (2014).
181. Chew, G.L., Pauli, A. & Schier, A.F. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat Commun* 7, 11663 (2016).
182. Cenik, C. et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res* 25, 1610-1621 (2015).
183. Lei, L. et al. Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J* 84, 1206-1218 (2015).
184. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38, 576-589 (2010).
185. Cuperus, J.T. et al. Deep Learning Of The Regulatory Grammar Of Yeast 5' Untranslated Regions From 500,000 Random Sequences. *bioRxiv* (2017).
186. Lawless, C. et al. Direct and Absolute Quantification of over 1800 Yeast Proteins via Selected Reaction Monitoring. *Mol Cell Proteomics* 15, 1309-1322 (2016).
187. Christiano, R., Nagaraj, N., Frohlich, F. & Walther, T.C. Global proteome turnover analyses of the Yeasts *S. cerevisiae* and *S. pombe*. *Cell Rep* 9, 1959-1965 (2014).
188. Collier, J.M., Tucker, M., Sheth, U., Valencia-Sanchez, M.A. & Parker, R. The DEAD box helicase, Dhh1p, functions in mRNA decapping and interacts with both the decapping and deadenylase complexes. *RNA* 7, 1717-1727 (2001).
189. Fischer, N. & Weis, K. The DEAD box protein Dhh1 stimulates the decapping enzyme Dcp1. *EMBO J* 21, 2788-2797 (2002).
190. Sweet, T., Kovalak, C. & Collier, J. The DEAD-box protein Dhh1 promotes decapping by slowing ribosome movement. *PLoS Biol* 10, e1001342 (2012).
191. Collier, J. & Parker, R. General translational repression by activators of mRNA decapping. *Cell* 122, 875-886 (2005).
192. Zhong, Y. et al. RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* 33, 139-141 (2017).
193. Olshen, A.B. et al. Assessing gene-level translational control from ribosome profiling. *Bioinformatics* 29, 2995-3002 (2013).
194. Xiao, Z., Zou, Q., Liu, Y. & Yang, X. Genome-wide assessment of differential translations with ribosome profiling data. *Nat Commun* 7, 11194 (2016).
195. Larsson, O., Sonenberg, N. & Nadon, R. anota: Analysis of differential translation in genome-wide studies. *Bioinformatics* 27, 1440-1441 (2011).
196. Zhang, S., Hu, H., Jiang, T., Zhang, L. & Zeng, J. TIDE: predicting translation initiation sites by deep learning. *bioRxiv* (2017).

197. Malone, B. et al. Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res* 45, 2960-2972 (2017).
198. Liu, T.Y. & Song, Y.S. Prediction of ribosome footprint profile shapes from transcript sequences. *Bioinformatics* 32, i183-i191 (2016).
199. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14, 417-419 (2017).
200. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34, 525-527 (2016).
201. Breiman, L. Random forests. *Mach Learn* 45, 5-32 (2001).
202. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825-2830 (2011).
203. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33, 1-22 (2010).
204. Balakumar, B.J., Fang, Han, Hastie, Trevor, Friedman, Jerome H., Tibshirani, Rob, & Simon, Noah. (Zenodo; 2017).
205. Jones, E., Oliphant, T., Peterson, P. & others (2001).
206. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 17 (2011).
207. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* Chapter 11, Unit 11 17 (2010).
208. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863-864 (2011).
209. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21 (2013).
210. Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32, 292-294 (2016).
211. McKinney, W. in *Proceedings of the 9th Python in Science Conference*. (eds. S.e. van der Walt & J. Millman) 51 - 56 (2010).
212. Dale, R.K., Pedersen, B.S. & Quinlan, A.R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* 27, 3423-3424 (2011).
213. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842 (2010).
214. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9, 90-95 (2007).
215. Waskom, M.L. & Wagner, A.D. Distributed representation of context by intrinsic subnetworks in prefrontal cortex. *Proc Natl Acad Sci U S A* 114, 2030-2035 (2017).
216. Frazee, A.C., Jaffe, A.E., Langmead, B. & Leek, J.T. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* 31, 2778-2784 (2015).
217. Cherry, J.M. et al. *Saccharomyces Genome Database: the genomics resource of budding yeast*. *Nucleic Acids Res* 40, D700-705 (2012).
218. Fang, H. et al. Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine. *BMC Med Genomics* 10, 10 (2017).
219. Vurture, G.W. et al. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* (2017).
220. Cassidy, S.B., Schwartz, S., Miller, J.L. & Driscoll, D.J. Prader-Willi syndrome. *Genet Med* 14, 10-26 (2012).

221. Christian, S.L. et al. Molecular characterization of two proximal deletion breakpoint regions in both Prader-Willi and Angelman syndrome patients. *American Journal of Human Genetics* 57, 40-48 (1995).