On the Origin of Phenotypic Variation: Novel Technologies to Dissect Molecular Determinants of Phenotype

Francesco Vallania

Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/etd

Part of the Computational Biology Commons

Recommended Citation
https://openscholarship.wustl.edu/etd/1201
WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biological and Biomedical Sciences

Computational & Systems Biology

Dissertation Examination Committee:
Robi D. Mitra, Chair
Barak A. Cohen
Todd E. Druley
James J. Havranek
Gary D. Stormo
Ting Wang

On the Origin of Phenotypic Variation:

Novel Technologies to Dissect

Molecular Determinants of Phenotype

by

Francesco LM Vallania

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2013
St. Louis, Missouri
Table of Contents

List of Figures ....................................................... ix
Abbreviations ......................................................... xv
Acknowledgments ....................................................... xvi
Abstract ................................................................. xxiii

1 Introduction: The molecular basis of phenotypic diversity ............... 1

1.1 The genetic basis of phenotypic diversity .................................. 1

1.1.1 The mapping of phenotype to genotype ................................ 1

1.1.2 The genetic basis of common traits and disease: the common and rare variant hypotheses ................................................. 3

1.1.3 The sequencing revolution ............................................. 6

1.2 When genetics fall short: epigenetic drivers of phenotype ............. 10

1.2.1 Genetic factors are not sufficient to explain phenotypic variation .... 10

1.2.2 Epigenetic factors control phenotypes and cell-fate decisions ...... 11

1.2.3 Role of epigenetic factors in development and cancer ............... 16
1.2.4 Chromatin modifications as novel and controversial epigenetic factors

1.2.5 DNA methylation is a stable chromatin modification and a classic epigenetic factor

1.2.6 DNA methylation stably represses gene expression in development and cancer

1.2.7 How is de novo DNA methylation targeted?

1.3 Beyond the boundaries of determinism: the stochastic nature of life

1.3.1 Stochasticity in cellular processes overcomes genetics and epigenetics

1.3.2 Stochastic variation improves fitness, guides development, and modulates differentiation dynamics

1.3.3 Gene expression is a stochastic process: noise in the system

1.3.4 How is gene expression variance regulated?

1.3.5 Noise in gene expression is modulated by multiple intrinsic and extrinsic factors

1.3.6 The relationship between mean and variance in gene expression

2 Detection of rare genetic variants in pooled DNA samples

2.1 Foreword

2.2 Quantification of rare allelic variants from pooled genomic DNA

2.2.1 Abstract

2.2.2 Intro/Results
2.2.3 Supplementary Results ........................................ 60
2.2.4 Discussion ...................................................... 67
2.2.5 Materials and methods ......................................... 69
2.2.6 Acknowledgments ............................................... 85
2.2.7 Supplementary tables .......................................... 86

2.3 High-throughput discovery of rare insertions and deletions in large cohorts .................................................. 90
   2.3.1 Abstract .......................................................... 90
   2.3.2 Introduction .................................................. 91
   2.3.3 Results .......................................................... 95
   2.3.4 Discussion .................................................... 107
   2.3.5 Materials and methods ...................................... 112
   2.3.6 Acknowledgments ............................................ 126
   2.3.7 Supplementary Tables ....................................... 127

3 High-throughput dissection of genetic factors underlying DNA methylation ................................................... 133
   3.1 Foreword ......................................................... 133
   3.2 Introduction/Results ............................................ 134
      3.2.1 Experimental design and construction of a reporter system for DNA methylation ........................................ 134
3.2.2 Dissection of the Oct4 regulatory elements as a proof of concept experiment ........................................ 142
3.2.3 Future directions ................................................................. 154
3.3 Materials and methods ............................................................. 154
  3.3.1 Construction of piggyBAC methylation reporter .................. 154
  3.3.2 Construction of the Oct4 minilibrary .............................. 158
  3.3.3 Cell culture protocols and differentiation ......................... 159
  3.3.4 Transfection and selection .............................................. 160
  3.3.5 DNA extraction and bisulfite conversion .......................... 160
  3.3.6 Next generation sequencing library preparation ................ 161
  3.3.7 Computational analysis and methylation calls .................. 162

4 Origin and consequences of the relationship between protein mean and variance ..................................................... 163
  4.1 Foreword ........................................................................ 163
  4.2 Abstract ........................................................................ 164
  4.3 Introduction .................................................................... 165
  4.4 Results ........................................................................ 167
    4.4.1 Protein mean and variance are connected by a power-law-like relationship ..................................... 167
    4.4.2 A stochastic model of gene expression recapitulates the power-law-like relationship between mean and variance ................................................................. 168
4.4.3 The power-law-like relationship between protein variance and mean depends on promoter kinetics ........................................ 170

4.4.4 The relationship between protein mean and variance separates mean-
dependent from mean-independent sources of variance ............ 173

4.4.5 Experimental confirmation of the relationship between promoter-positioned nucleosomes and mean-independent variance .............. 176

4.4.6 Promoter-positioned nucleosomes increase variance by slowing pro-
moter activation kinetics ............................................ 178

4.5 Discussion .............................................................................................................. 178

4.6 Methods ................................................................................................................. 181

4.6.1 Data Sources .................................................................................................. 181

4.6.2 Analysis of the relationship between protein mean and variance .... 182

4.6.3 Stochastic modeling of protein and mRNA variance ......................... 182

4.6.4 Correlation analysis between mean-independent variance and molecu-
lar properties ........................................................................................................ 183

4.6.5 Regression model between mean-independent variance and nucleosome occupancy ......................................................... 183

4.6.6 Experimental measurement of mean-independent variance as a func-
tion of nucleosome occupancy ......................................................... 184

4.7 Supplementary methods ......................................................................................... 185
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7.1 Stochastic Modeling of mRNA and Protein Variance in <em>S. cerevisiae</em></td>
<td>185</td>
</tr>
<tr>
<td>4.7.2 Analysis of residual protein variance</td>
<td>196</td>
</tr>
<tr>
<td>4.8 Author Contributions</td>
<td>197</td>
</tr>
<tr>
<td>4.9 Acknowledgments</td>
<td>197</td>
</tr>
<tr>
<td>5 Concluding remarks and future implications</td>
<td>199</td>
</tr>
<tr>
<td>6 Bibliography</td>
<td>205</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Examples of phenotypic differences ........................................ 2

1.2 Common Disease Common Variant Vs Common Disease Rare Variant ........ 5

1.3 Rare variants carry a large effect size .................................... 6

1.4 Cost of DNA sequencing is dropping at a faster rate than predicted by Moore's Law .......................................................... 8

1.5 Sensitivity of whole-genome sequencing in predicting disease ................. 12

1.6 Relative risk of disease in individuals testing negative by whole-genome sequencing ......................................................... 13

1.7 Phage λ fate decision: a model for epigenetics ................................. 15

1.8 Waddington's epigenetic landscape ............................................ 17

1.9 An epigenetic switch links inflammation to cellular transformation and cancer . 18

1.10 Establishment, maintenance, and removal of DNA methylation .................. 21

1.11 Dynamics of DNA methylation during development .......................... 23
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.12</td>
<td>Abnormal hyper-methylation of tumor suppressor genes in different types of tumor</td>
<td>25</td>
</tr>
<tr>
<td>1.13</td>
<td>Establishment of bimodal DNA methylation at embryonic implantation</td>
<td>27</td>
</tr>
<tr>
<td>1.14</td>
<td>Protein binding to DNA removes DNA methylation</td>
<td>30</td>
</tr>
<tr>
<td>1.15</td>
<td>Protein binding to DNA establishes DNA methylation</td>
<td>31</td>
</tr>
<tr>
<td>1.16</td>
<td>Models of β-galactosidase induction in cell populations</td>
<td>34</td>
</tr>
<tr>
<td>1.17</td>
<td>Relationship between Bernoulli $\sigma^2$ and $\mu$</td>
<td>37</td>
</tr>
<tr>
<td>1.18</td>
<td>Stochastic variation induces bacterial persistence to antibiotics</td>
<td>39</td>
</tr>
<tr>
<td>1.19</td>
<td>Responsive vs stochastic switching of phenotype following environment changes</td>
<td>40</td>
</tr>
<tr>
<td>1.20</td>
<td>Stochastic variation in gene expression</td>
<td>42</td>
</tr>
<tr>
<td>1.21</td>
<td>Noise in protein levels scales with protein abundance</td>
<td>49</td>
</tr>
<tr>
<td>2.1</td>
<td>Error model reveals hidden trends in error rates</td>
<td>58</td>
</tr>
<tr>
<td>2.2</td>
<td>SNPseeker accurately quantifies SNPs found by pooled-DNA sequencing</td>
<td>61</td>
</tr>
<tr>
<td>2.3</td>
<td>Pooled sample vs. dbSNP average heterozygosity</td>
<td>62</td>
</tr>
<tr>
<td>2.4</td>
<td>Illumina Genome Analyzer I quality score analysis</td>
<td>64</td>
</tr>
</tbody>
</table>
SNPSeeker improves SNP calling by removing sequencing errors

Supplementary Table 1. PCR primers and conditions

Supplementary Table 2. Taqman assay primers and probes.

Supplementary Table 3. Taqman validation

Supplementary Table 4. Known and putative SNP positions identified by pooled-sample sequencing

Supplementary Table 5. Comparative genomics analysis of non-synonymous SNPs

Experimental and computational pipeline for detection of indels and substitutions in large pooled-DNA samples

Visual representation of the SPLINTER algorithm

Relationship between variant detection accuracy and average sequencing coverage per base

Relationship between variant detection accuracy and average sequencing coverage per base for substitutions and indels

Comparison between SPLINTER and other variant calling algorithms

Comparison between SPLINTER and CRISP
2.17 Effect of different aligner algorithms to SPLINTER performance .......................... 103

2.18 Precise quantification of rare genetic variants in synthetic samples .......................... 105

2.19 Precise quantification of 4bp insertions and deletions ........................................ 106

2.20 Precise quantification of rare genetic variants in real samples ............................... 107

2.21 Correlation between variant quantification by SPLINTER Vs Sequenom ...................... 108

2.22 Supplementary Table 1 .................................................................................. 127

2.23 Supplementary Table 2 .................................................................................. 127

2.24 Supplementary Table 3 .................................................................................. 127

2.25 Supplementary Table 4 .................................................................................. 127

2.26 Supplementary Table 5 .................................................................................. 127

2.27 Supplementary Table 6 part1 ............................................................................. 128

2.28 Supplementary Table 6 part2 ............................................................................. 129

2.29 Supplementary Table 6 part3 ............................................................................. 130

2.30 Supplementary Table 7 .................................................................................. 130
2.31 Supplementary Table 8 .............................................. 131

2.32 Supplementary Table 9 .............................................. 131

3.1 Overview of the piggyBAC DNA transposon system ........... 136

3.2 Schematic representation of the piggyBAC methylation reporter assay ............ 137

3.3 piggyBAC recapitulates the endogenous methylation pattern of the Trf promoter 138

3.4 piggyBAC methylation reporter recapitulates endogenous patterns at a single CpG level ............. 140

3.5 DNA methylation profile across different regions of the piggyBAC−Trf construct ........ 141

3.6 piggyBAC methylation reporter recapitulates endogenous patterns .................... 143

3.7 piggyBAC methylation reporter is highly reproducible ......................... 144

3.8 Design of the Oct4 mini-library experiment ............................ 145

3.9 Methylation levels of the Oct4 mini-library in different cell types ................ 147

3.10 Single CpG methylation reproducibility in the Oct4 library .................... 148

3.11 Dinucleotide frequencies predict DNA methylation of the Oct4 inserts .......... 150

3.12 in silico occupancy of transcription factors predicts methylation of the Oct4 inserts 151
3.13 Methylation of the insert affects the methylation of the reporter ..................................... 155

3.14 Methylation of the insert affects the methylation of the reporter at a single CpG level ..................................................... 156

4.1 Relationship between mean and variance in protein expression ........................................... 168

4.2 Residual mean-independent variance in protein expression ................................................. 169

4.3 Stochastic model of gene expression ...................................................................................... 171

4.4 Model prediction of genome-wide protein variance in *S. cerevisiae* ........................................ 172

4.5 Promoter kinetics but not mRNA and protein synthesis and degradation rates modulate the relationship between mean and variance .................................................. 174

4.6 Promoter-associated factors modulate protein variance independently of the mean 175

4.7 Changes in nucleosome occupancy but not in translation rate induce changes in mean-independent variance .................................................................................................. 177

4.8 Single-cell mRNA variance connects underlying promoter kinetics to nucleosome occupancy ................................................................................................................. 179

4.9 Changes in gene expressions are driven by changes in $K_{off}$ or $K_m$ whereas $K_{on}$ remains largely constant .......................................................... 191
## Abbreviations

In order of appearance:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcriptional Start Site</td>
<td>ICM</td>
<td>Inner Cell Mass</td>
</tr>
<tr>
<td>CH$_3$</td>
<td>Methyl Group</td>
<td>COCH$_3$</td>
<td>Acetyl Group</td>
</tr>
<tr>
<td>DNMT</td>
<td>DNA Methyl Transferase</td>
<td>meC</td>
<td>5-Methyl C</td>
</tr>
<tr>
<td>A</td>
<td>Adenosyne</td>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>CH$_2$OH</td>
<td>Hydroxymethylation</td>
<td>MBP</td>
<td>Methyl Binding Proteins</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein-Barr Virus</td>
<td>GFP</td>
<td>Green Fluorescent Protein</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean</td>
<td>$\sigma^2$</td>
<td>Variance</td>
</tr>
<tr>
<td>F</td>
<td>Fano Factor</td>
<td>CV</td>
<td>Coefficient of Variation</td>
</tr>
<tr>
<td>PIC</td>
<td>Pre Initiation Complex</td>
<td>Pol II</td>
<td>RNA Polymerase II holoenzyme</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger Ribonucleic Acid</td>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>indel or IN/DEL</td>
<td>Insertion/Deletion</td>
<td>DP</td>
<td>Dynamic Programming</td>
</tr>
<tr>
<td>SPLINTER</td>
<td>Single Nucleotide Polymorphism</td>
<td>GWAS</td>
<td>Genome-Wide Association Study</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
<td>SFTPB</td>
<td>Surfactant Protein B</td>
</tr>
<tr>
<td>ACTB</td>
<td>$\beta$-Actin</td>
<td>MAF</td>
<td>Minor Allele Frequency</td>
</tr>
<tr>
<td>CDCV</td>
<td>Common Disease Common Variant</td>
<td>RDCV</td>
<td>Rare Disease Common Variant</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operator Curve</td>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
</tbody>
</table>
I would like to thank first and foremost my doctoral advisor, Robi D Mitra, or Rob, as we all know him. Rob is one of the kindest, most compassionate people I have ever had the pleasure to meet. Yet despite his gentle predisposition, an unstoppable creative energy dominates his scientific endeavors. His energy is contagious and has changed me in ways I would have never predicted. Furthermore, Rob gave me something that not many advisors give: the gift of intellectual freedom by pursuing projects that I deeply cared about. This process was not all roses as I initially thought and many unforeseen difficulties popped out along the way. Despite being one of the smartest people I have ever met, Rob never made me feel wrong or inferior when my approaches and strategies were misguided and not well thought out. Through a true collaborative and mutually respectful relationship, those obstacles were overcome over the course of long well-articulated discussions, intense brainstorming sessions, and compulsive board sketching. During my time in Saint Louis, a terrible family emergency forced me to fly back to Italy for several months. Given how expensive and bureaucratically cumbersome maintaining an international student was, it would have been no surprise if Rob decided that he could have not kept me anymore. Instead, I found in him the most personable and supportive character I could have ever hoped for. Not only I continued my work with him, but he made every possible effort to make my life as easy as possible in a time when I felt my personal world collapsing. I will never forget that.
An important character crucial in my scientific and personal development is Barak A Cohen. Although we never officiated our little arrangement, Barak was always the "other mentor", the guy in the lab downstairs whom I would often visit in times of scientific and personal confusion. Barak taught me many important lessons early on in my graduate career. One of my earliest interactions with Barak occurred when I was rotating in his lab. At that time, I accidentally referred to him as my boss, a title that seemed perfectly suitable given the scenario. One and half hours later, I was finally released from his office after being repeatedly and forcibly told that in no way whatsoever, under absolutely no circumstances, he shall ever be referred to by me or any other student or postdoc as their boss, a title not representative of the independence in judgment and thoughts that he expects from anyone. God forbid, it never happened again. In fact, that episode was incredibly important in shaping my scientific independence and often rebellious attitude toward authority, a side-effect he perhaps did not foresee at the time. That process, coupled with the aforementioned infusion with creativity and freedom, opened my mind and allowed me to pick problems that I found interesting based on curiosity and potential. Barak also made me realize that science is really not about specific projects but rather about getting at the core of important questions, no matter what it takes or what you have to do.

Todd Druley was an instrumental character in my success as a graduate student. In fact, it is fair to say that without his help, I would have not accomplished much in my career. I met Todd during my rotation in Rob’s lab and my work started as a collaboration with him. We were able to ride the wave together and he made me an integral part of the big splash
that he was working on making even though he did not have to. I owe him that much. I
like to think that we scientifically grew together and that he learnt from me as much as I
learnt from him. The most valuable lesson that I learnt from Todd is to keep things simple
and real. It is easy to lose track of the goal and become detached from reality, especially in
modeling and computation. Todd always made sure that it did not happen to me.

I would like to thank the remaining members of my committee, Gary Stormo, Jim
Havranek, and Ting Wang, for their support, constructive criticism, and help. They trusted
that everything would be all right even when the proposed projects and steps were difficult
and at high risk of failing, giving me a lot of freedom and room for independence.

One important realization in science is that everyone does not just stand on the shoul-
der of the giants that preceded them, but also has to rely on those of their peers. I find
myself hard pressed to imagine a more collaborative and fun group of people than the
one I was part of. Starting from members of the Mitra lab, I would like to thank Xuhua
Chen, Michael Brooks, Yue Yun, Kate Varley, Lee Tessler, David Mayhew, Maxim Schille-
beekx, Jessica Hoisington-Lopez, Charles "Scott" Higdon, Gabe Bien-Willner, Tom Cohen,
Sumithra Sankararaman, Zongtai "Z" Qi, Arjun Bahl, and Justin Mendelez. My interactions
with them were incredibly important and made me grow as an individual. I not only learnt
much about science but also lessons of personal nature and human interaction which I will
forever treasure. Being a foreigner in a foreign land, I never felt un-welcomed nor ignored
by my lab-mates, even if I missed some jokes along the way.
In the greater community of the Center for Genome Sciences (and Systems Biology) I found many great friends even outside of my own lab. I cannot name all the people that I have met here and have been great friends to me because it would fill up my whole dissertation. However, I would like to specially mention Adam, Brett, Carlo, Claire, Chris, Cynthia, GiNell, Hemangi, Igor, Jamie, Kate, Meng, Sung, Vasavi, and Yue for their kindness, support, and genuine interest in spending time with me discussing science and fun stuff. Their presence has proven to be a crucial component in my life.

I would also like to thank Melanie Puhar, the program manager at DBBS. More than one time I managed to cause trouble and mayhem and she ended up fixing it for me. I will miss you but you will miss me more.

Despite all my time spent in Saint Louis, none of this work would have been possible without the support of my friends from Italy. During personally difficult times and afterwards I often found myself to have to rely on their help, which they provided gladly and unquestionably. I cannot begin to thank all the people who helped me along the way on work, legal, and personal matters. However, I specifically cannot fail to mention the Maero family. Over the course of many years they have gone above and beyond the call of duty, helped me out in any possible way, and last but not least, made me feel like I was part of their family in the moments when I had none of my own.

I would like to thank my wife Stephanie for all her love and support over the 9 years I had the pleasure of being part of her life. I can recall many occasions when I neglected spending time with her due to my work and schedule and not once I have received any complaint nor
lack of support. The story of our relationship is magical and unique, something that I would never trade for anything else in the world. She is the most precious person alive to me, the love of my life, and an anchor to my sanity in difficult times. I cannot imagine spending my life without having her by my side. I would like to also thank her beautiful and kind parents, Phil and Teresa, for accepting me in their family and making me feel at home right from day one.

I would like to dedicate the best parts of this work to my parents. Unfortunately, due to unforeseen circumstances, they both passed away at a relatively young age. My father, Massimo, passed away when I was only 12. I admired him immensely and to this date he is by far the most hardworking and dedicated person I have ever met. Despite not having spent much time with him, I have wonderful memories of our moments together.

My mom, Natalina, had the difficult and challenging task of raising me by herself and making sure we would be financially stable. Not once in my entire life I felt I had less options than my peers as she gave me the freedom of choosing my own path with full un-questioned support from her side. Besides being a very sweet person, she was a beacon of inspiration, optimism, and guidance. Her drive allowed her to achieve tremendous goals during her life, no matter how difficult the obstacles were. Without her help and support I would have not been able to reach this point. Sadly, she passed away during the course of my PhD due to a terrible disease despite having been healthy throughout her life. Her passing has been the most difficult experience I have ever dealt with in my life. I deeply miss her and I still deal
with the consequences of her loss on a daily basis. My deepest hope is that my work and efforts would be a reason of pride for her.

Finally I would like to thank you, dear reader, for taking the time to go through my dissertation, if you choose to do so. The work presented here is a result of years of thinking, coding, benchwork, and interminable discussion. I hope you will enjoy it.
Dedicato ai miei amati genitori, in particolare alla mia cara mamma. Mamma ti penso tutti i giorni e mi manchi tantissimo. Non sai cosa darei per averla ancora qui. Spero che quello che ho fatto ti abbia reso orgogliosa e felice. Senza il tuo supporto e amore, non ci sarei mai riuscito. Mi impegnerò al massimo per renderti sempre fiera di me.
Abstract

On the Origin of Phenotypic Variation:
Novel Technologies to Dissect
Molecular Determinants of Phenotype

by

Vallania, Francesco LM

Doctor of Philosophy in Computational & Systems Biology,
Washington University in St. Louis, 2013.
Professor Robi D Mitra, Chairperson

This thesis describes the conception, design, and development of novel computational tools, theoretical models, and experimental techniques applied to the dissection of molecular factors underlying phenotypic variation. The first part of my work is focused on finding rare genetic variants in pooled DNA samples, leading to the development of a novel set of algorithms, SNPseeker and SPLINTER, applied to next-generation sequencing data. The second part of my work describes the creation of a reporter system for DNA methylation for the purpose of dissecting the genetic contribution of tissue-specific patterns of DNA methylation across the genome. Finally the last part of my work is focused on understanding the basis of stochastic variation in gene expression with a focus on modeling and dissecting the relationship between single-cell protein variance and mean at a genome-wide scale.
1. Introduction: The molecular basis of phenotypic diversity

1.1 The genetic basis of phenotypic diversity

1.1.1 The mapping of phenotype to genotype

The collection of a living being’s observable characteristics such as morphology, biochemical and physiological properties, life cycles and behavioral patterns is defined as ”phenotype” \[1\]. Example of observable phenotypes range from morphological difference between species \[2\], variation between individuals of the same species \[3\], and variation between different cell types within the same tissue \[4, 5\] (Figure 1.1).

An important goal in biology is to understand the molecular origin of phenotypic variation. This question has deep implication in a variety of different contexts, ranging from evolution and speciation to early diagnosis and treatment of human disease. A fundamental component underlying phenotypic variation lies in the presence of inheritance factors, collectively summarized as the ”genotype” \[1\]. The association between genotype to a phenotype is explored by the field of genetics. This field originated from the work of Gregor Mendel, a 19\textsuperscript{th} century Austrian monk whose experiments on plant hybridization provided the basis for classical genetics \[6\]. These eventually led to the discovery of ”Mendel’s law” which describe how hereditary factors would segregate in crosses, leading to an inherited predictable phe-
Figure 1.1. Examples of phenotypic differences

This figure shows interspecies differences in cranial morphology of Darwin’s finches (left, adapted from [2]), intraspecies differences as shell patterning variation in the mollusk *Donax varabilis* (center, adapted from [3], and morphological differences between cell in different layers of the epidermic tissue (right, adapted from [4]). All three cases are examples of phenotypic variation.

Genetic factors, or "genes" became an intense object of study in the early 20th century, leading to major discoveries on their shape and organization [7]. The revelation of the physical nature of genetic factors came in 1944, where the genetic material responsible for the inheritance of the phenotype was found to be deoxyribonucleic acid, or DNA [8], a polymeric molecule formed by single monomeric units called nucleotides [4]. This discovery marks the beginning of the field of "molecular genetics" and associates the concept of genetic inheritance and allelic variation to the molecular nature of DNA. Different alleles were revealed to be changes in the sequence of nucleotides. These changes are called "mutations" or "genetic
variants”, and their effects has been partially interpreted with the discovery of the ”genetic code”, which converts the information inside the coding portion of a gene into a protein sequence [4]. Despite the advancements of the understanding of the nature of genetic factors, a major difficulty lies in the fact that, in most cases, phenotypes are not easily explained by Mendel’s laws, but rather involve the interaction and contribution of different genetic factors and interaction with environment. This can be described as :

\[
\text{Genotype}(G) + \text{Environment}(E) \rightarrow \text{Phenotype}(P) \tag{1.1}
\]

Such phenotypes are termed ”complex traits” and are in many cases common across the general population [9]. Understanding the genetic basis underlying complex traits remains up to this day an important and difficult challenge.

1.1.2 The genetic basis of common traits and disease: the common and rare variant hypotheses

A fundamental goal in biomedical research is to understand the contribution of genetic variants toward common diseased phenotypes. The ability to predict the susceptibility of a single individual to develop any particular disease can indeed have large implications toward the goal of personalized medicine [10]. Two major hypotheses describing the relationship between genetic variants and common disease emerged over time [11, 12]. The first one has been described as the ”common disease common variant” hypothesis and it postulates that common diseases and phenotypes are explained by one or few causative genetic variants that
are common in the population \(13\) (Figure 1.2). A variant is defined to be common if present at a frequency of 5% or greater in the general population \(14\). The biggest corollary of this model is that causative variants are few and frequent, thereby reducing the difficulty of the problem and the number of patients required for any particular study. This hypothesis has been shown to be true in the context of rare diseases \(15\).

In contrast, the second hypothesis, which is also known as the "common disease rare variant" hypothesis, predicts a scenario where common diseases and phenotypes are instead explained by the presence of multiple causative variants that occur at a rare frequency in the population \(16, 17\) (Figure 1.2). A variant is defined to be rare if present at a frequency of 1% or lower in the general population \(14\). In this case, the variants are multiple and different, and they will co-occur very un-frequently in different patients displaying the same phenotype. As a result, large groups of patients are necessary to reach adequate statistical power.

The caveats of each model have important implications toward the design of the association study necessary to find the underlying causative variants \(18\). It is therefore important to understand which model explains most of the phenotype.

Over the recent 5 years, by reviewing the results and conclusion of multiple association studies performed on patients affected by common diseases, a clear trend emerged. Each genetic variant in our genomes carries an "effect size", namely the ability of such variant to produce a phenotypic effect. Common variants carried a very low effect size, if any at all, whereas in contrast rare variants were associated to a high effect size, pointing them
Adapted from [19], this figure presents 2 examples depicting common (left) and rare (right) genetic variants associated with a disease phenotype. Common causative variants will be one or few at most and they will segregate mostly within the diseased cohort. Rare variants will be numerous and will be mostly unique to one or a few patients.

as the likely causative variants associated to phenotype [12] (Figure 1.3). Furthermore, a computational re-analysis of previous association studies hints at the possibility that previous associations of common variants to common phenotypes were in fact artificially linked due to their covariance with underlying causative rare variants [20]. This analysis suggests that rare variants are a likely factor to explain common diseases. Therefore, in order to dissect the genetic basis underlying common phenotypes, it is fundamentally important to have the ability to detect rare variants.
Adapted from [12], this figure shows the value of sequencing cost per genome (left) and per base pair (right) over time compared to the expected trend predicted by Moore’s Law.

1.1.3 The sequencing revolution

A fundamental problem in rare variants genetics simply lies in the ability to detect them in the first place. As rare variants occur at a frequency lower than 1%, large cohorts of individuals have to be genotyped to detect such variants with adequate statistical power simply due to sampling limitations. For instance, to detect a given mutation occurring at a frequency of 1 in a 1000 in the population with a probability of 96% (i.e. to be able to find 96% of all mutations at such frequency), it is necessary to profile 5000 patients
(computed directly as a result of binomial sampling). Because of this reason, the costs and time-commitment associated with traditional genotyping strategies such a large cohort of patients would be astronomical, making this process unfeasible. Therefore, in order to find rare variants associated to disease, a new methodology had to be devised.

One of the largest scientific endeavors over the last 15 years was undoubtely the sequenc- ing of the first human genome [21]. This effort was completed in 2001 and involved multiple institutions for a cumulative costs of $100,000,000 and 10 years of time commitment. The technology adopted by the National Human Genome Research Institute [21] and its corporate competitor Celera Genomics [22] relied in both cases on ”SANGER sequencing”, a traditional method of sequencing involving fluorescently labeled dye terminators [23]. This method presents a major limitation, that is it can sequence only one molecular species at a time, limiting the applicability of sequencing to small scale projects. Most of the cost and time spent on the human genome sequencing project are a direct consequence of this limitation.

From the late 90s to the early 2000s a series of technological advancements led to the development of new sequencing platforms that would allow the parallel sequencing of millions of DNA molecules at the same time [24, 25, 26, 27]. The whole range of new sequencing technologies has been conventionally grouped under the name of ”next-generation” sequenc- ing. These new platforms can resolve individual sequences from heterogeneous mixtures of DNA molecules and as a result, bypass the limitations imposed by SANGER, paving the way for a subsequent sequencing revolution. In the field of computer hardware and electri-
In electrical engineering, the Moore’s Law is often used as a reference for the rate of technological advancement, representing the doubling rate of transistors on an integrated board \[28\]. As a direct result of next-generation sequencing platforms, the cost of sequencing dropped over time at a much faster rate predicted by Moore’s Law (Figure 1.4). At the time of writing of this thesis, the cost of sequencing a whole human genome is already 4 orders of magnitude lower than 2001.

Figure 1.4. Cost of DNA sequencing is dropping at a faster rate than predicted by Moore’s Law

![Cost per Genome](http://www.genome.gov/sequencingcosts/) ![Cost per Raw Megabase of DNA Sequence](http://www.genome.gov/sequencingcosts/)

Adapted from [http://www.genome.gov/sequencingcosts/](http://www.genome.gov/sequencingcosts/) this figure shows the value of sequencing cost per genome (left) and per base pair (right) over time compared to the expected trend predicted by Moore’s Law. Note the drop in 2007 with the introduction of next-generation sequencing.

The unprecedented power of next generation sequencing offers the promise of revolutionizing not just genetics but also general molecular biology research with numerous and unexpected applications \[29\]. As the first part of my graduate research, I developed an ex-
perimentally integrated computational framework for accurately detecting rare-variants in a cost-effective manner by leveraging the power of next-generation sequencing. This work started as a collaboration with Todd Druley, MD-PhD who lead the initial project and it is described in the second chapter of my dissertation.
1.2 When genetics fall short: epigenetic drivers of phenotype

1.2.1 Genetic factors are not sufficient to explain phenotypic variation

The promise of genetics fulfilled by the power of next-generation sequencing has indeed the potential of revolutionizing biology and medicine by revealing the past, present, and future of any individual from the information encoded in their genome. Or does it?

At this point in time, as we transition from the dawn of the sequencing revolution into the world of genomic medicine, we have yet to acquire both the necessary power and knowledge to interpret the impact of genomic variants to any relevant phenotype. A pioneering study [30] answers this question by adopting a clever strategy to bypass our technical and theoretical shortcomings. In this study, the authors analyzed a large cohort of monozygotic twin pairs and looked at the incidence of a variety of different diseases within each pair.

Even by accounting for the presence of somatic mutations in the twins (genetic alterations acquired in single cells within a single individual after conception [31]), it is very safe to assume that any twin within each pair is genetically identical to their sibling. This assumption serves as the fundamental premise for this study: if a particular disease could be completely explained through genetics, we would expect every pair to be concordant for that disease, i.e. if a twin developed a particular disease, their sibling would also have the same disease, and the other way around. Following this reasoning, we can use the number of discordant pairs to estimate what fraction of the incidence of a particular disease can be explained solely through genetics and, most importantly, how much predictive power whole-
genome sequencing provides in the ideal scenario where every causative variant has been identified and assigned to its biological function.

Despite the promise of this study, its conclusions suggested a scenario far from ideal. Using a permissive statistical threshold that ensured at least 10% of the individuals diagnosed as positive would effectively have the disease, the authors found that most of the patients affected by the disease (> 50%) would have been completely disregarded as false negatives (see Figure 1.5).

Furthermore, individuals predicted to be healthy by whole-genome sequencing (i.e. not to have a particular disease) only carried a small reduction in relative risk of disease incidence (see Figure 1.6) compared to the general population (with a few notable exceptions). This suggests that even a negative result from genomic sequencing would not substantially improve the odds of remaining healthy for most of the diseases profiled in the study.

These results indicate that even in an idealized scenario where practical and theoretical limitations are absent, whole-genome sequencing would provide only limited predictive power restricted to a selected set of diseases. From this we can conclude that, while genetic information is indeed important and necessary to understand and predict phenotype, it is clearly not sufficient to do so, leaving room for other factors to come into play.

1.2.2 Epigenetic factors control phenotypes and cell-fate decisions

A second important group of factors crucial for phenotypic specification is given by "epigenetic" factors. The adjective epigenetic has been used in different ways over the years
Adapted from [30], this figure shows the sensitivity of whole-genome sequencing in correctly identifying individuals affected by disease. For each disease, the expected maximum and minimum fraction of positive cases are presented on the y-axis as function of the statistical threshold chosen in the study. A sensitivity value of 100% indicates correct identification of every individual affected by the disease, whereas a sensitivity value of 0% indicates complete failure in identifying any individual affected by the disease.

to indicate different entities and mechanisms linked by a common theme, which lies in their non-genetic nature. The original definition of an epigenetic change states that such change "...does not involve a mutation, but that is nevertheless inherited in the absence of the signal (or event) that initiated the change" [32]. Epigenetic factors are important because they integrate information from the external environment, which act as an initiating trigger, and then maintain it over time. This process eventually translates into a phenotype.
We assumed that the risk alleles for these 24 diseases were independent. It was of interest to determine how the results described above varied overestimates. On the other hand, these frequencies may represent the same risk of developing the disease as the general population, whereas a relative risk of 0% indicates that the individuals in the group will not develop the disease.

Adapted from [30], this figure shows the relative risk of developing a particular disease given a negative result by genetic testing. For each disease, the expected maximum and minimum relative risks are presented on the y-axis as a function of the statistical threshold chosen in this study. A relative risk value of 100% represents the same risk of developing the disease as the general population, whereas a relative risk of 0% indicates that the individuals in the group will not develop the disease.

One of the first observed examples of an epigenetic switch consists in the regulatory network controlling the fate of the phage λ [33]. A bacteriophage or phage is a virus that infects bacterial cells. λ is a phage that infects the bacterial species E.coli by binding through the lamB or mannose receptor present on the surface of E.coli [34] and subsequently injecting its own DNA inside the bacterial cytosol (see Figure 1.7). Once inside E.coli, λ can then

Figure 1.6. Relative risk of disease in individuals testing negative by whole-genome sequencing

Adapted from [30], this figure shows the relative risk of developing a particular disease given a negative result by genetic testing. For each disease, the expected maximum and minimum relative risks are presented on the y-axis as a function of the statistical threshold chosen in this study. A relative risk value of 100% represents the same risk of developing the disease as the general population, whereas a relative risk of 0% indicates that the individuals in the group will not develop the disease.

One of the first observed examples of an epigenetic switch consists in the regulatory network controlling the fate of the phage λ [33]. A bacteriophage or phage is a virus that infects bacterial cells. λ is a phage that infects the bacterial species E.coli by binding through the lamB or mannose receptor present on the surface of E.coli [34] and subsequently injecting its own DNA inside the bacterial cytosol (see Figure 1.7). Once inside E.coli, λ can then
undergo two distinct fate decisions. In one case \( \lambda \) replicates itself and produces all the viral proteins necessary for the assembly of new viral particles causing the infected \textit{E.coli} cell to undergo lysis and eventually release the new particles outside. This particular fate decision is called "lytic cycle". In a second alternative scenario, the \( \lambda \) genome previously existing as a separate molecule in the cytosol integrates itself into the bacterial host genome and lies dormant. This second fate decision is called "lysogenic cycle" (see Figure 1.7).

The choice of \( \lambda \) to initiate one decision over the other is controlled by a regulatory network involving the expression and interaction of different viral proteins with the \( \lambda \) genome [33, 35].

The viral protein \( CII \) is the focal point of the epigenetic switch. \( CII \) is a "transcription factor" (TF), a DNA-binding protein that can activate or repress the expression of a target gene by binding to its regulatory signals located upstream of the "transcriptional start site" (TSS) of said gene [4, 36]. The levels of \( CII \) control the switch: \( CII \) is degraded by the activity of host proteases expressed by \textit{E.coli}. When \textit{E.coli} is in ideal growth conditions it will express high levels of proteases leading to the elimination of \( CII \) from the cell. In this scenario a TF named \( Cro \) will be expressed inducing the transcriptional program leading to the lytic cycle.

In contrast, when \textit{E.coli} is for instance "starved" and its growth conditions are not optimal, its proteases will be down-regulated, leading to increased levels of \( CII \). In this scenario \( CII \) will promote the expression of \( CI \), another TF which will induce lysogenesis. As depicted in Figure 1.7, the two programs are reinforced by the fact that they are mutually
exclusive (as one program represses the other) indicating that once the switch is activated, a memory of the current state is then established.

Figure 1.7. Phage λ fate decision: a model for epigenetics

Adapted from [35], the upper part of this figure shows the life cycle of phage λ and its two fate decisions (lytic or lysogenic) regulated by the levels of the viral protein CII. The lower part of this figure shows a model for the gene regulatory network of λ underlying its fate decision.

This example demonstrates how a simple regulatory circuit can control the fate of a virus and maximize its fitness. In this case phenotypic variation (namely the pathway chosen by λ) is driven solely by an epigenetic switch. Despite the fact that the switch itself is encoded in the DNA, the states of the switch are not as the lytic and lysogenic viral particles are genetically identical.
1.2.3 Role of epigenetic factors in development and cancer

A typical and important scenario in which the establishment of radically distinct phenotypes is observed within a clonal population can be found during the process of development [5], where a complete organism is generated from a single cell. During the first phases of development, a ”zygote”, the first cell resulting from fertilization of the oocyte by a single sperm cell, gives rise to a cellular structure named the ”blastocyst”, whose inner cell mass (ICM) consists of ”embryonic stem cells”, cells that have the potential to generate every cell type in the adult organism. During the process of ”gastrulation”, the cells in the embryo rearrange themselves forming three ”germ layers”, each layer carrying the capacity of generating different cell types [5]. This process becomes intrinsically more specialized, as with each step, each intermediate cell type acquires specific functions and loses ”potency”, the ability to differentiate into different cell types. This phenomenon has been termed ”epigenetic landscape” by Conrad Hal Waddington [37] (Figure 1.8). This process is epigenetic because every cell in the embryo is genetically identical and every step is associated to a memory state. Because of this memory state, cells do move backward in development, i.e. de-differentiate into more potent cell types.

An additional prominent example of the role of epigenetic switches can be found in pathology, specifically in the context of cancer [38]. Using isogenic cancer cell-lines as a model of cancer (therefore maintaining a uniform genetic background), the authors behind this study found that the process of transformation, that is the set of phenotypic changes that distinguish a physiologically normal cell from a tumorigenic one, was triggered by an
Historically, the word “epigenetics” was used to describe events that could not be explained by genetic principles. In the years following the enigma of Mendelian genetics, numerous biological phenomena, largely governed by changes in interactions between genes and their products, which bring the phenotype into being” (Waddington, 1942). Over the years, numerous biological phenomena have been identified, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include a variety of diverse phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics. These include phenomena, some considered bizarre and inexplicable, have been lumped into the category of epign...
induces downstream activation of STAT3, a TF involved in tumorigenesis when constitutively activated \[40\]. Finally, the authors show that the key factors mediating the epigenetic switch are also expressed in clinical samples, suggesting that these findings may extend outside of the model system that they adopted \[38\].

Adapted from \[38\], this scheme illustrates the epigenetic switch that connects inflammation to cancer. Src is the initiating factor that triggers an inflammatory response mediated by NFκB.

These results shed light on the importance of epigenetics in disease and potentially offer an explanation as to why the mono-zygotic twins in \[30\] share substantial discordance in disease incidence.
1.2.4 Chromatin modifications as novel and controversial epigenetic factors

The examples presented here so far depict the classical interpretation of an epigenetic switch, namely a self-contained regulatory network where different *trans*-acting factors interact together forming positive and negative feedback loops resulting in the initiation and maintenance of a non-genetic memory state. In recent years, the definition of epigenetic factors has been extended to the context of chromatin biology because of its role in modulating gene expression in a cell-type specific manner without underlying genetic changes.

Chromatin is defined as the combination of DNA and packaging proteins found in the nucleus \[41\]. The minimal functional unit of chromatin is dictated by a small protein complex termed "nucleosome", which consists of a 8 monomeric protein units named "histones" \[42\]. Nucleosomes have a primary important role in packaging DNA \[43\] but they also serve as important regulatory functions \[44\].

DNA-binding proteins such as transcription factors can modify chromatin by recruiting catalytic domains that can covalently add or remove modifying groups to the N-terminal tails of the histones of any given nucleosomes \[45\]. These groups range from small molecules such as acetyl (COCH$_3$) and methyl (CH$_3$) groups to small proteins such as the Small Ubiquitin-like MOfifier or "SUMO" proteins. These modifications provide the basis for a "histone code" \[46\] where single or multiple marks combined together can influence the level of expression of a genes by facilitating the recruitment of protein complexes necessary for gene expression.
The decision of including chromatin marks as epigenetic factors has stirred controversy among purists in the field [32, 47]. This discontent originated because chromatin marks are not maintained across cell division due to their intrinsically fast turn-over rate [48, 49]. As a result, they are believed to reflect transcription factor recruitment, carrying therefore a transient and dynamic role. This behavior undermines a fundamental point in the classical definition of epigenetics, namely the presence of a memory state that persists after the initiating stimulus has disappeared.

1.2.5 DNA methylation is a stable chromatin modification and a classic epigenetic factor

An important chromatin-associated factor involved in cell differentiation and development in vertebrates is DNA methylation [50]. DNA methylation consists of the covalent addition of a methyl group (CH$_3$) to a cytosine nucleotide (Figure 1.10). This reaction is catalyzed by a class of enzymes called "DNA methyl transferases" (DNMTs) [51] (specifically $DNMT3a$ and $b$) (Figure 1.10) and occurs primarily (but not exclusively) in cytosines (C) followed by a guanine (G), forming a CG di-nucleotide. Such di-nucleotide is commonly referred to as a "CpG" and a methylated cytosine is referred to as a "5-MethylC" or "meC" [52].

Unlike chromatin marks, DNA methylation abides to the classic definition of epigenetic factor as it is maintained without the presence of its initiating stimulus through a clear molecular mechanism. DNA methylation is bound to be diluted over the course of cell division. As new DNA is synthesized at the replication fork [53], DNA methylation is
lost on the nascent strand because of lack of a mechanism by the DNA polymerase to recognize and add meCs during replication. This phenomenon is biologically important and is named ”passive demethylation” [53]. DNMT1, a member of the DNMT family of enzymes, counteracts this problem by binding to hemi-methylated DNA at the replication fork and subsequently restoring methylation in the newly synthesized strand [53] (Figure 1.10). This ensures that, once DNA methylation is established in the first place, it is maintained over the course of cell division, thereby producing an epigenetic memory state.

Figure 1.10. Establishment, maintenance, and removal of DNA methylation

Adapted from [53], this figure displays establishment of DNA methylation by DNMT3, maintenance of DNA methylation upon DNA replication mediated by DNMT1, and finally removal by passive (replication dependent) and active DNA (replication independent) de-methylation.
As with any other classic epigenetic switch, if an initiating stimulus arises, the positive feedback can be broken and the switch can be turned. Equivalently DNA methylation can be actively removed through chemical conversion into hydroxymethylation (CH2OH) followed by replacement of the meC with an unmethylated C through the “base-excision repair” complex [54]. This process is defined as “active demethylation”.

The function of DNA methylation consists in silencing the expression of nearby genes. The region controlling the basal expression of a gene is called “promoter” and it is located upstream the TSS of the gene. When the promoter of a particular gene becomes methylated, its meCs can be recognized and bound by a family of proteins called “Methyl Binding Proteins” (MBP) [55]. Upon binding on the meCs, these proteins act as repressors of transcription, inducing silencing of the expression of the downstream gene. Because DNA methylation is chemically stable and it is maintained through DNA-replication, its silencing effect is stable as well, making it a suitable epigenetic switch.

1.2.6 DNA methylation stably represses gene expression in development and cancer

DNA methylation plays a prominent role in the context of development. Following fertilization, the zygote undergoes a process of complete de-methylation termed “erasure”, where the pronuclei originated from the egg and sperm cells lose their methylation patterns, effectively making tabula rasa of their previous statuses [53] (Figure 1.11). This process is followed by re-establishment of specific patterns of DNA methylation following implantation.
of the embryo, when the stem cells located in the ICM acquire a "bimodal" methylation state \[56\]. Following this process, DNA methylation mirrors the process of differentiation, where every cell acquires specific and unique patterns across its genome.

Figure 1.11. Dynamics of DNA methylation during development

![Graph showing DNA methylation dynamics](image)

Adapted from \[53\], this figure depicts the dynamics of DNA methylation during development. After fertilization, DNA methylation "erasure" of the male and female pronuclei occurs followed by denovo establishment after implantation.

The silencing effect of DNA methylation plays a fundamental role in establishing the epigenetic memory necessary for development that was previously observed by Waddington \[37\]. A notable example comes from the methylation of the gene Oct4 \[57\]. Oct4 is a TF acting at the top of the transcriptional program that defines potency and stem-cell state of a cell. Because of this reason, Oct4 is often called a "master regulator" of pluripotency \[58\]. Exogenous expression of Oct4 together with 3 other master regulators has been shown
to dedifferentiate somatic cells back to embryonic stem cells [59]. This result indicates that stable repression of $Oct_4$ is a necessary epigenetic switch that in physiological conditions prevents de-differentiation. The expression of $Oct_4$ is regulated by cis-acting elements, DNA sequences that are specifically recognized and bound by TFs [60, 61]. Methylation of these elements is observed during differentiation and is associated to repression of $Oct_4$ [62]. The role of DNA methylation in the context of $Oct_4$ is of critical importance in maintaining its epigenetic memory. $Oct_4$ is methylated by $G9a$, a histone lysine methyl-transferase that can bind and recruit $DNMT3a$ and $b$ [63]. When $G9a$ was mutated so that $Oct_4$ would be repressed but not methylated, it was observed that 15% of the stem-cells that underwent differentiation could revert back to pluripotency when chemically induced to do so, whereas cells with an intact copy of $G9a$ fully maintained their differentiated state. Interestingly, when $G9a$ was mutated so that it would methylate $Oct_4$ but not repress it through chromatin marks, only 1% of the differentiated cells reverted. This finding suggests that DNA methylation is not only necessary but also sufficient to establish epigenetic memory, confirming its role as a true epigenetic factor.

The role of DNA methylation as a repressive epigenetic switch has also been widely appreciated in the context of cancer and tumorigenesis. In fact two major phenomena have been observed: on a global scale, tumorigenesis is accompanied by a general decrease in levels of DNA methylation, defined as a state of ”global hypomethylation” [39]. The origin and consequences of this phenomenon are still poorly understood. In contrast, on a finer resolution, several genomic regions that are normally un-methylated acquire high levels of
DNA methylation, or become ”hypermethylated”. These loci tend to be promoter regions driving the expression of a class of genes called ”tumor suppressors” \[39\].

Figure 1.12. Abnormal hyper-methylation of tumor suppressor genes in different types of tumor

Adapted from \[39\], this figure depicts the distribution of cases where a particular tumor suppressor gene is found to be hyper-methylated in a particular type of tumor.

Tumor suppressor genes are normally expressed in order to prevent tumorigenesis by inhibiting cell growth. Their inactivation often leads to development of cancer. In several clinical cases, inactivation of such genes is not accompanied by the presence of mutations in their coding or non-coding regions \[39\], but instead their promoter is found to be hyper-methylated (Figure 1.12), with methylation leading to their silencing. Consequently, demethylating agents such as 5-Azacytidine have been used as chemotherapies in different types of cancer. *In vitro* analysis of its mechanism of action have revealed its ability to specifically de-methylate tumor suppressor genes that were previously hyper-methylated.
and restore their expression \[64\]. These results indicate that DNA methylation is a major epigenetic driver of development and differentiation and plays a prominent role in cancer.

1.2.7 How is \textit{de novo} DNA methylation targeted?

Because DNA methylation plays an important role as an epigenetic factor, a fundamental question is to understand how its patterns are established in the first place. As mentioned above, at the implantation phase of embryonic development the cells of the ICM will re-establish patterns of DNA methylation following erasure \[56\]. This patterning will be bimodal and its specification is determined by occupancy of an activating chromatin mark: the methylation of the 4th Lysine of Histone 3, also referred to as ”H3K4me”. When H3K4me is present, DNA methylation will be absent and vice-versa. This is explained by the requirement of \textit{DNMT3L}, a member of the DNMT family lacking catalytic activity. \textit{DNMT3L} binds to nucleosomes and recruits \textit{DNMT3a} and \textit{DNMT3b}, which in turn will methylate the underlying CpGs \[56\] (Figure 1.13). However, its interaction with the nucleosomes is inhibited by the presence of H3K4me, leading to the establishment of the bimodal pattern \[65\] (Figure 1.13).

This however can only explain bimodal patterning at implantation. In the following phases of development, the cells in the embryo will become more and more differentiated acquiring specific and unique patterns of DNA methylation \[66\]. The underlying mechanism specifying these pattern is at this stage not completely understood.
Figure 1.13. Establishment of bimodal DNA methylation at embryonic implantation

Adapted from [56], this figure shows the molecular mechanism explaining the bimodal pattern of DNA methylation at embryonic implantation. When H3K4me is absent, DNMT3L binds the nucleosome and recruits DNMT3a and/or DNMT3b leading to establishment of DNA methylation. When H3K4me is present, DNMT3L will not bind the nucleosome and consequently DNA methylation will not be established.

However, over the course of the last 20 years, independent studies have highlighted examples that may shed some light on this problem [67, 68, 69, 70, 71]. The first example comes from a class of regions present in the genome called "CpG islands" [72]. CpG islands are characterized by having an unusually high density of CpGs (>= 65% of observed to expected ratio) and GC-rich (>= 55%) sequence content [73] compared to the rest of the genome. CpG islands are found in > 40% of promoter regions in mammals and for the most
part are largely un-methylated \[72, 74\]. Furthermore, when a piece of DNA from a CpG island is isolated and methylated \textit{in vitro}, it will become de-methylated once reintroduced in the stem cells \[75\]. How do CpG islands establish and maintain their un-methylated state? The answer comes from the binding a single TF called \textit{Sp1} \[67\]. \textit{Sp1} binds to a DNA motif characterized by the canonical recognition site ”GGGCGGGG” \[76\] which is enriched in CpG islands. Its binding has been shown to be sufficient to remove DNA methylation when exogenously introduced inside a genomic region that was already methylated \[67\]. Furthermore when naturally occurring sites recognized by \textit{Sp1} are mutated, the nearby regions acquire DNA methylation, both in CpG islands as well as in dynamically methylated regions such as the \textit{Oct4} promoter \[67, 77\].

These results suggest a model where DNA methylation is specified by \textit{cis}-acting elements located in the promoter sequence and the binding of a \textit{trans}-acting factor, such as a TF, to those elements can specify patterns of DNA methylation. An additional set of experiments was performed later on to dissect the methylation patterns occurring on the genome of a human pathogen, the ”Epstein-Barr” virus (EBV) \[68\]. EBV is one of the most common human viruses, infecting 90\% of the population world-wide \[78\] and is the causative agent of infectious mononucleosis (source: [http://www.cdc.gov/ncidod/diseases/ebv.htm](http://www.cdc.gov/ncidod/diseases/ebv.htm)). EBV can infect different human cell types and can exists in a dormant phase called ”latency” inside the cell nucleus. In this state, EBV is present in the form of a circular DNA genome, called ”episome” \[79\]. EBV episomes are maintained and replicated by the host cell machinery and for all intended purposes, behave as mini-chromosomes. EBV’s replication is controlled by
a viral protein called EBNA-1 which binds to the latent replication origin oriP located on
the viral episome \cite{80}. Methylation analysis of human cells revealed the EBV episome to be
completely methylated except for the regions of oriP bound by EBNA-1 \cite{68} (Figure 1.14a).
Through genetic manipulation of EBV, it was shown that binding of EBNA-1 was sufficient
and necessary to de-methylate oriP and that its effect was replication dependent \cite{68}. This
mechanism is general and holds true for other protein/DNA interactions independently of the
nature of the protein, pointing to binding alone as a sufficient factor \cite{69, 70, 81}. Furthermore,
the amount of de-methylation induced by protein binding was proportional to its affinity to
its cis-acting sequence and its concentration \cite{69, 81} (Figure 1.14b,c).

The data presented so far points to a model where 1) deposition of DNA methylation is
a default non-specific event that occurs everywhere and 2) patterns of DNA methylation are
specified only by either blocking or removing DNA methylation. This model is however con-
tradicted by experimental evidence where DNA methylation is on the other hand established
by protein/DNA interactions \cite{63, 71, 82}. An early example comes again from Oct4, which
acquires methylation during differentiation \cite{62, 63, 82}. As mentioned above, the case of Oct4
DNA methylation is not the default state but instead it results from binding of G9a and its
direct recruitment of DNMT3a and b \cite{63, 82}. A second example comes again from cancer,
specifically from the hyper-methylation of p15, a tumor suppressor gene \cite{71}. It was shown
that the aberrant methylation of p15 was caused by ZNF217, a TF that can bind to the p15
promoter. By inhibiting ZNF217 the methylation of p15 is lost \cite{71} (Figure 1.15a), pointing
ZNF217 as necessary driver of DNA methylation. Similarly to the Oct4 promoter, ZNF217
Evidence of protein binding specifying de-methylated DNA sites: (a) Adapted from [68], this figure shows methylation status of the oriP sites bound by EBNA-1. HpaII bands at 1.2Kb and 320bp are indicative of de-methylation and are found only when cells are replicating and EBNA-1 is expressed. (b) Adapted from [69], this figure shows % demethylation of lacO DNA locus as a function of the inhibition of the protein lacI by IPTG (underlying biology of lacO and lacI is elucidated in [35]). (c) Adapted from [81], this figure shows the methylation level of *in vivo* bound binding sites by their corresponding TFs as a function of intracellular TF concentration.
can specifically recruit $DNTM3a$ through the repressive COREST complex (Figure 1.15b). This result suggests an alternative model where a TF or trans-acting factor is recruited to a genomic region and by interacting with co-factors and $DNMT$s can drive establishment of $de$ novo methylation.

Figure 1.15. Protein binding to DNA establishes DNA methylation

Adapted from [71], this figure shows the mechanism underlying aberrant methylation of the promoter of $p15$.

(a) Bisulfite sequencing reads indicating methylation status of the $p15$ promoter: $p15$ is hyper-methylated in a cancer cell line (left, dots in a line are single CpGs in a SANGER read, a black dot represents a methylated CpG, a white dot represents an unmethylated CpG). When $ZNF217$ is inhibited (right), methylation is lost.

(b) Model of direct recruitment of $DNMT3a$ by $ZNF217$ to the $p15$ promoter.
At the time of writing of this thesis, the notion of DNA methylation being encoded at the level of the DNA sequence has been generally accepted and proven \[70\]. However, the empirical evidence supporting the above mentioned models is still quite limited and the number of known *cis*-acting elements driving DNA methylation is very sparse. A major limiting factor in this endeavor is the lack of a strategy for the high-throughput dissection of the sequence basis underlying DNA methylation. In the third chapter of this dissertation, I will describe my effort to bring a solution to this problem by developing a novel ”methylation reporter system” for the high-throughput dissection of the sequence basis of DNA methylation and its application to an *in vitro* model system for neuronal development.
1.3 Beyond the boundaries of determinism: the stochastic nature of life

1.3.1 Stochasticity in cellular processes overcomes genetics and epigenetics

The addition of epigenetic factors that integrate information from the environment to the underlying genetic information should in theory be sufficient to explain any phenotypical variation observed in nature. However that is not entirely the case.

One of the first empirical observations comes from measuring levels of β-galactosidase, an enzyme important for the metabolism of the sugar galactose, inside the bacterium *E. coli* cells [83]. The authors of this paper induced the production of this enzyme and measured the amount of enzyme in single bacterial cells. The expectation of this experiment would be that every cell would display similar amounts of enzyme and that this quantity would increase proportionally to the level of induction. In contrast, what the authors found was that while the mean level of enzyme across the population was indeed proportional to the induction, the response at a single-cell level was instead binary: a single cell would either express the enzyme at fully induced levels or show almost no expression at all (Figure 1.16). Several implications can be extrapolated from this result:

1. The mean level of induction in a population of cells is dictated by a combination of induced and non-induced cells.

2. The level of enzyme expressed in either induced or non-induced cells is similar across the respective population, following a all-or-none response.
3. Different cells exhibit different phenotypes despite being genetically identical and in the same environment. Phenotypic variation is therefore very high.

As described in [83], this figure illustrates opposing models for the quantitative induction of the β-galactosidase enzyme in *E. coli*. Under the "average cell" model, each cell expresses the same amount of enzyme, leading to an identical population. In contrast, under a "all-or-none" model, cells are either fully induced (blue cells) or completely un-induced (white cells) at different frequency corresponding to different levels of induction.

The observation that phenotypic variation can occur within a given environmental context with a fixed genetic background has also been later appreciated in genetics. More specifically, a particular mutation can manifests its phenotype only in a subset of the individuals that
carry it, given fixed genetic background and environmental stimuli. This phenomenon has been described as "incomplete or partial allele penetrance" [84]. Follow-up studies have revealed the generality of this behavior in different biological phenomena [85, 86, 87, 88, 89, 90, 91, 92, 93, 94], suggesting that its origin may lie on a common general principle.

How can the same genetic instructions in the same environment be interpreted in radically different ways within the same cell population? The answer to this question lies in the context of probability. Physical-chemical, and therefore biological, processes are based on the interaction of a discrete number of molecules within a system. These interactions are probabilistic in nature and the average fraction of molecules in a particular state can be accurately captured through mathematical modeling. In the easier case when the system reaches "thermodynamic equilibrium", this can be achieved by applying the formalisms of statistical thermodynamics [95]. For instance, let’s consider a protein-DNA interaction, such as the binding of a TF to its cognate DNA recognition site. Assuming that the concentrations of TF and DNA do not change, their binding is irreversible (as it usually is), and pressure and temperature remain invariant, we can described their interaction as:

\[
[TF] + [DNA] \rightleftharpoons [TF \cdot DNA] \tag{1.2}
\]

where \([TF]\) and \([DNA]\) represent the unbound species and \([TF \cdot DNA]\) represent the bound complex. The affinity of the TF to DNA can be described as

\[
K_{eq} = \frac{[TF \cdot DNA]}{[TF][DNA]} \tag{1.3}
\]
where $K_{eq}$ is the equilibrium constant of the system. The average fraction of TF-DNA complex at equilibrium is then described by the Boltzmann distribution as

$$<TF \cdot DNA> = \frac{[TF]K_{eq}}{1+[TF]K_{eq}}$$ \hspace{1cm} (1.4)

where the average is the probability of being the bound state. This model directly relates to the Bernoulli distribution in probability space [96], in which the mean $\mu$ is equal to the probability of the Bernoulli event to occur:

$$\mu = p$$ \hspace{1cm} (1.5)

the variation in a series of Bernoulli trials is measured by its variance $\sigma^2$, which is mathematically defined as

$$\sigma^2 = p(1-p)$$ \hspace{1cm} (1.6)

This definition can be easily shown to be equivalent in thermodynamic space as well, leading to

$$\sigma^2_{[TF\cdot DNA]} = <TF \cdot DNA> (1-<TF \cdot DNA>)$$ \hspace{1cm} (1.7)

Now given that $p$ is a probability and therefore can only range between 0 and 1, the relationship between $\mu$ and $\sigma^2$ shows that $\sigma^2$ is at its peak at intermediate levels of $\mu$ (0.5) and at its lowest at the extreme values of $\mu$ (0 and 1) (Figure 1.17).
This figure shows the relationship of $\sigma^2$, which is defined as $p(1-p)$, and $\mu$, defined as $p$, for a Bernoulli random variable.

This reflects the fact that a single Bernoulli variable can exist only in two states and can explain the original all-or-none expression of $\beta$-galactosidase in *E. coli*. At any given time point a single *E. coli* cell can either express it or not and therefore can be modeled as a Bernoulli variable. Because of this reason, at intermediate levels of induction, the population of *E. coli* cells will be at its highest variance as $\sim$50% will be in the induced state and the remaining $\sim$50% will be un-induced. This behavior results from the fact that *E. coli* possess only one copy of the gene for $\beta$-galactosidase, which leads any single cell to
exist in a single state of a binary set. In contrast, hypothetically speaking, if \textit{E. coli} carried a high number of copies of $\beta$-galactosidase, each copy would behave as a Bernoulli variable and, as a result, each cell would express a continuous amount amount of $\beta$-galactosidase. This is purely a consequence of the limits of statistical sampling and the law or large numbers \cite{97}. These results indicate that phenotypic changes can occur despite uniform environmental and genetic background purely as a result of sampling from a low number of molecules, resulting in ”stochastic” variation.

1.3.2 Stochastic variation improves fitness, guides development, and modulates differentiation dynamics

An important underlying question at the base of stochastic variation is whether it provides a phenotypic advantage compared to a case of uniform phenotypic response. An illuminating example comes from experiments conducted on bacterial survival after antibiotic exposure \cite{88,98}. In one of these studies \cite{88}, single cells of \textit{E. coli} were exposed in culture to the antibiotic "Ampicillin". Because these strains of \textit{E. coli} did not carry any ”resistance factor”, i.e. genes that conferred immunity to the antibiotic used, the expected result was to observe no surviving cells after the treatment. In contrast, several cells from the original colony were shown to survive the treatment and eventually be able to repopulate the entire colony once the drug was removed. This phenomenon was called "bacterial persistence”. Its manifestation comes from the fact that inside the original population there is a subpopulation of cells replicating at a slower rate in dynamic ”equilibrium” with the remaining cells. Because
"Ampicillin" kills replicating cells [99], this population survives the exposure of antibiotic and eventually reconstitutes the entire original population (Figure 1.18).

Figure 1.18. Stochastic variation induces bacterial persistence to antibiotics

Adapted from [88], this figure shows "persistence" of *E.coli* after exposure to Ampicillin in absence of a resistance factor. Time points from 0:00 to 1:45 show *E.coli* cells growing inside a micro-fluidic cell (green lines). Time point 6:50 shows the presence of cells that persist treatment with Ampicillin and eventually repopulate the micro-fluidic chamber in subsequent time points after Ampicillin is removed.

In the absence of stochastic variation, the population would have not survived the treatment, suggesting the presence of an associated fitness advantage. Mathematical modeling of this observation [100] suggests that stochastic phenotype switching provides higher fitness compared to direct response to environmental stimuli (Figure 1.19).
Figure 1.19. Responsive vs stochastic switching of phenotype following environmental changes

Adapted from [100], this figure compares stochastic phenotypic switching to response to environmental changes.

This phenomenon is not however limited only to environmental response but it is physiologically relevant in the context of development and differentiation. One particular example comes from the development of the eye in the fruitfly *Drosophila melanogaster* [91]. The eye structure in *Drosophila* is formed by individual optical units called "ommatidia", which exist in two different subtypes (pale and yellow). These subtypes are randomly distributed in the fruitfly retina and their specification depends on the expression of the protein *spineless*. When *spineless* is inactivated, almost all ommatidia assume the pale fate whereas its over-expression induces the yellow fate. These phenotypes are randomly specified in normal
development as a result of stochasticity in the expression of \textit{spineless}, allowing the formation of a complicated biological structure simply by executing a random sampling.

Finally, stochastic switching between different cell types occurs in the context of cancer \cite{101}. In this work, the authors analyzed two breast cancer cell lines and identified 3 different subpopulations characterized by different surface markers: luminal, basal, and stem-like cells. By isolating each type using flow-cytometry the authors showed that each individual purified fraction could reconstitute the entire population. These results have important implications in the design of therapies to treat cancer as the effects of specifically targeting of only one population subtype could be nullified by the ability of every other subtype to reconstitute the original population. Taken together, these observations and others \cite{92} indicate that stochastic phenotype transitions are a fundamental component of the process of cell differentiation and overall are associated to increased fitness.

1.3.3 Gene expression is a stochastic process: noise in the system

Regulation of gene expression has been implicated as a driving factor underlying cellular differentiation, development, and disease \cite{5 36 39}. In many of the cases that we have observed so far, stochastic switching of a phenotype occurs as a downstream consequence of variation in the expression of a particular protein regulating the activation of downstream genes \cite{91 93}. Because of these reasons, in order to understand stochastic phenotypic variation, it is important to measure and dissect variation in gene expression.
One of the first and most widespread empirical appreciation of stochasticity in gene expression comes from the measurements of cell-to-cell variation in protein levels. By coupling the coding region of the green fluorescent protein (GFP) naturally found in the jellyfish Aequeorea victoria to a promoter region of interest, the intensity of the light emitted by GFP upon UV excitation can be used as a readout for protein expression [102, 103]. By adopting techniques such as fluorescence microscopy and flow-cytometry [86], GFP levels can be measured at a single-cell resolution (Figure 1.20a).

Figure 1.20. Stochastic variation in gene expression

Adapted from [104] (a) and [86] (b), this figure shows cell-to-cell variation in the expression of GFP in E.coli (a). This can be represented as a distribution of measured fluorescence units per cell (b), which is characterized by a mean $\langle p \rangle$ and a variance $\sigma_p$.

One of the first application of this technology to the context of stochastic expression comes from measuring the levels of GFP expressed by the bacterium B. subtilis [86]. In this
work, the authors show that single *B. subtilis* cells display varying levels of GFP expression, which follows a distribution across the bacterial population (Figure 1.20b). This distribution can be parameterized by two numbers. The first is the mean protein value, indicated as $<p>$ or $\mu_p$, which corresponds to the average amount of GFP produced in each cell. The second is the protein variance, indicated as $\sigma^2_p$, which quantifies the amount of variation across the distribution of single-cell protein levels. Measures of single-cell protein variance have been obtained for eukaryotes as well such as the budding yeast *S. cerevisiae* [87], the fruitfly *D. melanogaster* [91], and mammalian cells [90, 94], indicating its generality across multiple organisms.

1.3.4 How is gene expression variance regulated?

Having established its existence and importance, a fundamental goal in the field of stochastic gene expression is to understand how cell-to-cell variation is controlled and modulated. The benefits of answering this question are multiple. First, a clear understanding of how cell-to-cell variation is modulated can predict which genes are going to be susceptible to high variation and ultimately what downstream phenotypic consequences we can expect from that. Secondly, the ability to predict expression variance can lead to the ability to control and ultimately engineer genes that are expressed with a desired level of variation.

A first step into the analysis of stochastic variation in protein levels is to determine the relationship between $\mu_p$ and $\sigma^2_p$. This is important because in many probability distributions, $\mu$ and $\sigma^2$ are not independent. Through the use of an inducible system for gene expression,
several studies observed a strong positive correlation between $\sigma_p^2$ and $\mu_p$ \cite{86, 87}. To better describe the changes of noise for different values of $\mu_p$, two different normalizing metrics were introduced \cite{105}, the ”Fano factor” \cite{106} and the ”coefficient of variation” \cite{107}. These metrics are used in statistics to normalize inherent mean effects of a stochastic process and provide a measure of the dispersion of a probability distribution.

The Fano factor was first created by the Italian-American physicist Ugo Fano for the description of the fluctuations of an electric charge obtained in a detector by comparing it to a Poisson process \cite{106}. This metric is defined as

\[
F = \frac{\sigma^2}{\mu} \tag{1.8}
\]

A Poisson process is a stochastic process counting the number of events in a given time-interval such that in any given non-zero time point, the distribution of events is Poisson \cite{96}. The Poisson distribution \cite{108} is mathematically described as

\[
P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{1.9}
\]

where both $\mu$ and $\sigma^2$ are parameterized by $\lambda$. This implies that for this particular case $F = \frac{\lambda}{\lambda} = 1$. In the context of stochastic gene expression, the Fano factor has been adopted as a normalization metric with the name of ”noise strength” and its implicit assumption is that the basal level of noise follows Poissonian scaling \cite{86, 89, 109}.

The coefficient of variation or CV is normalized measure of dispersion of a probability distribution. It is mathematically described as
Analogous to the Fano factor, the CV can assume value of 1 under particular circumstances, namely whether the underlying distribution of a stochastic process is exponential \[107\]. The exponential distribution is mathematically described as

\[
P(X = x) = \lambda e^{-\lambda x}
\]  

for \( x \geq 0 \). In this context, \( \sigma \) and \( \mu \) both assume the value of \( \frac{1}{\lambda} \) and therefore \( CV = \frac{\lambda}{\lambda} = 1 \).

This can also happen in the case of a Bernoulli variable when \( p = 0.5 \) as \( \sigma^2 \) becomes \( p^2 \) following \([1.6]\), indicating that a CV of 1 can be achieved in the case of highest possible Bernoulli variance. In the context of stochastic gene expression, the CV has been adopted as a normalization metric with the name of ”noise” and has been used widely throughout the field \[110, 111, 112, 113\]. Because of this reason, the study of stochastic gene expression, or gene expression variance, has been commonly defined as the study of ”noise in gene expression”.

1.3.5 Noise in gene expression is modulated by multiple intrinsic and extrinsic factors

Using these metrics, several studies aimed at dissecting determinants of stochastic variation focusing on single promoters or genes by performing titration and time course analysis \[87, 89, 110\]. The process of gene expression is inherently highly regulated and involves multiple consecutive steps \[60\] starting by the specific binding of TFs at the promoter and
enhancer sequences regulating the expression of the downstream gene. This is followed by the active recruitment of the "pre-initiation complex" (PIC) by the bound TFs, which ultimately is bound by the active "RNA Polymerase II holoenzyme" (Pol II). This step is followed by the initiation of transcription, where the DNA sequence of the gene is copied into a messenger RNA (mRNA). The process by which the mRNA is transcribed by Pol II after initiation is called "elongation". The transcribed and mature mRNA (for in depth reference see [4]) undergoes "translation" resulting in the production of protein molecules encoded in its sequence. Finally both mRNA and protein molecules are eliminated by a process of "active degradation" combined with their dilution as a result of cell division.

Different research efforts focused on the dissection of different steps involved in the process of gene expression. An initial analysis in E. coli focused on the effects of changing transcription and translation rates to cell-to-cell variation. The results indicated that both process lead to an increase in variance but at different rates. By normalizing cell-to-cell variance using the noise strength ($\frac{\sigma^2}{\mu}$), it was shown that changes translation efficiency produced a linear increase in noise strength, whereas changes in transcription lead to constant levels of noise strength. Similar results were observed in the model eukaryotic organism S. cerevisiae, but in this case changes in transcription lead to wide changes in noise strength as a result of different promoter kinetics [8]. Specifically, the binding of chromatin regulators and the presence of the "TATA-box" sequence in the promoter has been shown to produce an increase in noise strength, connecting specific molecular factors to gene expression vari-
Furthermore, a strong link between nucleosomes and noise has been established through large scale correlation analysis \[114\] and individual perturbation experiments \[115\].

Quantitative stochastic modeling of gene expression provides a theoretical basis for understanding and interpreting single-cell experiments. A stochastic system can be accurately described through the use of a ”master equation”, a set of first order differential equations describing the evolution of a probability distribution over time \[116\]. A simple example is (source: [http://en.wikipedia.org/wiki/Master_equation](http://en.wikipedia.org/wiki/Master_equation)):

\[
\frac{d \vec{P}}{dt} = A \vec{P}
\]  

(1.12)

where $\vec{P}$ is the probability distribution expressed as a vector and $A$ is the matrix of parameters (note that $A$ can also change as a function of time, thereby becoming $A(t)$). Commonly used stochastic models consist in a direct analytical solution of a simple system \[117\] or, in more complex cases, monte-carlo simulations recapitulating the solution, the most prominent and famous algorithm being the one created by Dan Gillespie in 1977 \[118\]. The application of these models to single-cell experiments allowed the authors of the studies to recapitulate the observed data and generate hypothesis on the underlying molecular sources of noise in gene expression.

Finally, an important result in the study of stochastic expression comes from the distinction between ”intrinsic” and ”extrinsic” sources of noise \[110\]. Using a system with two fluorescent reporters driven by the same promoter, the authors were able to appreciate changes in noise that were concordant within the same cells. These changes were due to fac-
tors "extrinsic" of the process of gene expression such as cell-to-cell fluctuations in number of polymerase molecules, cell-size, and phase of the cell cycle. In contrast, factors that were "intrinsic" of the process of gene expression, such as promoter kinetics and transcription and translation rates, would instead produce discordant fluctuations of the two reporters within the same cell. These results suggest that multiple processes can modulate noise in gene expression at the same time. An important challenge is then to estimate the magnitude and generality of each effect with respect to cell-to-cell variance.

1.3.6 The relationship between mean and variance in gene expression

As described above, over the course of the recent years, multiple studies have identified different factors associated with noise in gene expression. However, different analyses have been conducted by using different metrics based on the accepted theoretical justifications at the time of the study. The purpose of using a metric is two-fold: 1) normalizing variance in gene expression is necessary when genes expressed at different levels are compared with one another 2) a metric establishes a base-line amount of variation and allows the discovery of underlying factors that can increase or decrease variance beyond the expected value. The choice of which normalization metric to use is not trivial as different metrics make different assumptions on the underlying biophysical processes. The underlying question however remains: which metric is preferred for the analysis of noise in gene expression? Two seminal studies in the field approached this issue by performing a large scale survey of values of protein mean and variance for a representative fraction of the S. cerevisiae proteome [111, 112]. The
authors of these studies observed a similar trend in their respective datasets: protein noise levels (expressed as $\text{CV}^2$) showed an inverse relationship with protein mean levels (Figure 1.21). This result has profound implications in the analysis of noise: a strong relationship between mean and noise is indicative of the inadequacy of the metric of choice to normalize the variance from all the mean effects.

![Figure 1.21. Noise in protein levels scales with protein abundance](image)

Adapted from [111, 112], this figure shows the inverse relationship between protein noise (defined as $\sigma^2/\mu_p$) and protein mean levels ($\mu_p$).

Previous analyses [111, 112] have shown that $\sigma^2$ and $\mu$ are connected to one another but the exact shape of the relationship has not been revealed. In order to address this,
several recent studies took a new unbiased approach to the problem \cite{94,119}. Because this relationship is not known, one can assume that $\sigma^2$ and $\mu$ would scale following a general exponential relationship in the form of

$$\sigma^2 = k \mu^j$$  \hspace{1cm} (1.13)

which in log-space becomes

$$\log(\sigma^2) = j \log(\mu) + \log(k)$$  \hspace{1cm} (1.14)

by fitting (4.15) to single-cell fluorescence data from a dataset of Jurkat T-cells expressing GFP \cite{94}, the authors of the study found that protein variance and mean were related together by a power-law like relationship with an exponent $j = 1.7$. Similar results were obtained in *E. coli* \cite{120}, suggesting that this relationship is indeed general. Most importantly, if we assumed an underlying Poissonian scaling or exponential scaling, the baselines implied by the Fano factor and the CV respectively, we would have expected exponents equal to 1 and 2 respectively. This result indicates that the current metrics of noise do not completely remove underlying mean residues in their normalization. As a consequence of this, it is currently not clear what processes affect protein variance solely through changes on the mean level and what processes instead regulate protein variance independently of the mean. Furthermore the biophysical origin of this relationship has not yet been revealed and modeled. In chapter 4, I will explore and dissect the biophysical origin of this relationship by applying a stochastic model of gene expression to genome-wide protein variance data collected in the
model organism *S. cerevisiae*. Using this model, I will then reveal which molecular factors increase protein variance solely through mean effects and which factors increase variance by modulating the power-law relationship. This work was done in collaboration with members of the laboratory of Barak Cohen, PhD.
2. Detection of rare genetic variants in pooled DNA samples

2.1 Foreword

This chapter describes the development of a fast and cost-effective strategy for the discovery of rare genetic variants in large cohorts of patients. This strategy combines a new experimental protocol involving next-generation sequencing with a novel computational strategy designed to analyze its data. This work was pioneered by Todd E Druley, MD PhD, who motivated the study and developed the pooled DNA sequencing protocol, which forms the basis of this work. I started this project first as a collaboration with Todd by developing the computational strategy to analyze the data generated from pooled DNA sequencing and then by extending this strategy to more types of genetic variants. The first section of this chapter has been adapted from


in which I created the computational framework to detect rare single nucleotide polymorphisms (SNPs) in pooled DNA samples that Todd generated for this work. This framework is embodied in the software package SNPSeeker. This work was driven by Todd E Druley who designed the pooled-DNA sequencing strategy with Robi D Mitra and Katherine E Var-
ley and executed all the sequencing experiments. In this section I designed and wrote the SNPSeeker algorithm and then performed data analysis with Robi D Mitra. Daniel J Wegner and Sarah W Robinson performed all the Taqman validation assays. Olivia L Knowles and Jacqueline A Bonds punched the bloodspots from filter paper. Justin C Fay and Scott W Doniger designed and executed the comparative genomics analysis. Daniel J Wegner, Aaron Hamvas and F. Sessions Cole provided reagents and advice. The second section of this chapter has been adapted from 

Francesco LM Vallania, Todd E Druley, Enrique Ramos, Jue Wang, Ingrid Borecki, Michael Province, and Robi D Mitra. 

High-throughput discovery of rare insertions and deletions in large cohorts. *Genome research*, 20(12):1711–1718, 2010

in which I extended my algorithmic framework into a new software package named SPLINTER to detect rare insertions and deletions (indels) in pooled DNA samples. In this section, I designed and developed the SPLINTER algorithm and then tested its performance by generating synthetically engineered pooled-DNA samples. The experiments performed in this study were designed by Robi D. Mitra, Todd E Druley and myself. I then performed the sequencing experiments on the synthetic and real samples and the data analysis and variant validation. Jue Wang and Enrique Ramos performed sequencing on the GWA sample and analyzed its frequency correlation. Ingrid Borecki and Mike Province provided the real pools used for sequencing in this work. The methods developed in this chapter have been successfully applied to several association studies through collaborations [122, 123, 124] and have been expanded and featured in subsequent protocol papers [125, 126].
2.2 Quantification of rare allelic variants from pooled genomic DNA

2.2.1 Abstract

We report a targeted, cost-effective method to quantify rare single-nucleotide polymorphisms from pooled human genomic DNA using second-generation sequencing. We pooled DNA from 1,111 individuals and targeted four genes to identify rare germline variants. Our base-calling algorithm, SNPSeeker, derived from large deviation theory, detected single-nucleotide polymorphisms present at frequencies below the raw error rate of the sequencing platform.

2.2.2 Intro/Results

The cumulative impact of rare variants on common disease is currently unknown, but recent studies have implicated rare genetic variants in many complex traits and diseases. Consequently, it has been suggested that the combined effects of rare deleterious mutations, could explain a substantial fraction of the genetic susceptibility to many common diseases [16, 17]. Identifying rare variants necessitates genotyping large populations of individuals, either sequentially (e.g. the 1000 Genomes Project [127]) or, to minimize cost and time, as a pooled sample. However, it has proven difficult to quantify the prevalence of deleterious alleles in pooled samples. Sanger and array-based resequencing are expensive for the amount of sequencing coverage obtained, thus incompatible with large pools. Second-generation sequencing has lowered sequencing costs by over 100-fold, but high error rates have hindered
the analysis of large pooled samples, since it is difficult to distinguish rare variants from sequencing errors.

Due to cost and time savings, pooled-sample sequencing should be useful for studying rare, human-specific genetic variation in large populations; characterizing deleterious alleles at multiple loci that may impact disease susceptibility and treatment; quantifying the abundance of rare somatic mutations; and identifying germline variants associated with disease state.

We have implemented a novel combination of molecular biology techniques and computational analysis to achieve targeted resequencing and rare variant detection in 13 kb per individual from 1,111 individuals using the Illumina Genome Analyzer I. We pooled a normalized amount of DNA isolated from dried blood on 1,111 de-identified Guthrie cards collected for newborn screening. The Missouri Department of Health and Senior Services Institutional Review Board and the Washington University Human Research Protection Office reviewed and approved use of de-identified DNA samples (Supplementary Methods). Using specific primers (Supplementary Table 1), we PCR-amplified 15 loci covering 14.5 kilobases of the surfactant protein B (SFTPB), TP53, APC and β-actin (ACTB) genes. Amplicons were ligated into long concatemers, randomly fragmented, and prepared for Illumina sequencing according to the manufacturers protocol. This generated 4.4 \times 10^7 sequences, 83.4\% (3.7 \times 10^7 reads = 1.3 gigabases) of which aligned to the reference, allowing for up to 2 mismatches. To quantify sequencing errors, we included a 1,276 bp region of the pUC19 plasmid as an internal control. We used the first 800 bases to train our algorithm and the remaining 476 bases as
a test set. We found that the first twelve bases of each Illumina read contained significantly fewer errors than later bases (Figure 2.1a), so we used only these to identify sequence variation. However, sequencing errors were still present at a higher frequency than that of a single allele in the pool, making an accurate error model essential to distinguishing bona fide variants from sequencing errors. Since existing second-generation base-calling programs cannot detect and quantify rare variants in large pooled samples, we developed SNPSeeker, an algorithm based on Large Deviation Theory (www.genetics.wustl.edu/rmlab). SNPSeeker uses a 2nd order dependency error model for SNP identification and takes into account the position in the sequencing read (i.e. cycle number) and the identity of the two upstream bases. Consequently, only mismatches at bases 312 of each sequencing read were used to identify SNPs (this reduces the effective coverage per allele to 13.8-fold in these experiments). Unexpectedly, we found that incorporating quality scores did not improve results beyond these parameters (Figure 2.4 and Results). For each machine run, we trained a new error model using the internal pUC19 control, because we found that error rates varied significantly between machine runs (Figure 2.1b).

The SNPSeeker algorithm significantly improved the specificity of SNP calling (Figure 2.5 and Results). Using SNPSeeker, no SNPs were called in the training sequence (bases 1800), in the negative control (bases 801,276) or in 656 of the 658 bp without a known SNP in the SFTPB amplicon. This yields a specificity of 99.8% and demonstrates that our base-calling algorithm is specific and able to accurately model sequencing and PCR errors. We validated called SNPs by comparing them to prior individual Sanger sequencing at the SFTPB locus.
Figure 2.1. Error model reveals hidden trends in error rates

(a) The cumulative likelihood of every possible misincorporation event for sequencing cycles 1–32 is depicted for both the sense (+) and antisense (−) strands. The Illumina data filtering process truncated the data from two dates at 32 bases instead of 36, which is why only 32 cycles are represented here. INSET. Higher resolution of the error probability across cycles 1–12. (b) The intra- and inter-day variability for the A→C misincorporation event from four different sequencing dates. The error bars represent the standard deviation between different flowcell channels from the same date. INSET. Higher resolution of cycles 1–12.
in this cohort and performing additional Taqman assays (Supplementary Tables 2 and 3). To estimate our methods sensitivity, we sequenced a 665 bp region of the SFTPB gene with 7 known SNPs. Our method identified all 7 SNPs at very similar frequencies to those found by Sanger. Three of these SNPs were present at $< 1.5\%$ in this population demonstrating that this method has the sensitivity to detect rare SNPs in this pooled sample (Supplementary Table 3a). We also identified two SNP positions within SFTPB that were not identified by Sanger. We performed individual Taqman assays on each of these positions and neither mutation was identified.

In addition to the 9 called SNPs in SFTPB, 55 additional SNPs were called in ACTB, TP53, and APC (Supplementary Table 4). Of these, 37 (67\%) were previously described in dbSNP (build 128). By chance, one would expect less than 1 SNP, on average, to be shared between these two sets. Using Fishers Exact Test, we found that the observed degree of overlap is highly significant ($P < 1.3 \times 10^{-56}$). Therefore, it is highly likely that the 37 SNPs identified by SNPSeeker and found in dbSNP are bona fide variants. Many of these SNPs were rare: 26 of the 37 dbSNPs that were identified had estimated allele frequencies of less than 1.5\% in our population (Supplementary Table 4). We also performed a comparative genomic analysis of the deleterious nature of the non-synonymous SNPs identified (Supplementary Results and Table 5).

To estimate the positive predictive value of our method, we chose seven called SNPs in these genes for independent validation by individual Taqman genotyping. All seven SNPs were predicted to be rare, with estimated minor allele frequencies (MAF) ranging from 0.5
1.2%. Three of the selected SNPs were previously reported in dbSNP, though not in our population, and the remaining four have not been previously described. Taqman genotyping validated all seven called rare variants (Supplementary Table 3b). When combined with SFTPBP results, we validated 14 of 16 predicted SNPs, giving a positive predictive value of 87%. To determine if the pooled sample sequencing method could accurately quantify allele frequencies, we plotted the predicted versus true MAFs for each of the 14 validated SNPs (Figure 2.2). The observed and predicted frequencies were highly correlated ($r^2 = 0.96$) across a wide range of frequencies: from a single allele (0.05%) up to several hundred alleles (21.2%). When we plotted predicted SNP average heterozygosities against reported average heterozygosity values for all SNPs in common with dbSNP, the correlation remained strong at $r^2 = 0.82$ (Figure 2.3). These findings indicate that pooled sample sequencing is able to accurately determine the population frequency of common and rare alleles.

### 2.2.3 Supplementary Results

**Error Model Generation using pUC19**

For all sequencing performed, an internal control consisting of a 1,276 bp PCR amplified sequence from *E.coli*-cloned pUC19 vector was included in order to model the likelihood of observing errors in a SNP free context. Bases 1-800 of this amplicon were used to parameterize the algorithm with each machine run. Bases 801-1,276 were then used as a negative control test sequence. This model specifies the probability of observing an error in a sequencing read as a function of 1) the true identity of the base being sequenced, 2) the identity of
Allele frequency by sequencing vs. genotyping. The allele frequency as determined by sequencing is plotted against the actual frequencies as determined by individual Taqman assay for the 14 validated SNPs in our dataset (correlation coefficient $r^2 = 0.96$).

The observed base, 3) the identities of the two reference bases immediately upstream of the base being sequenced, and 4) the current cycle number of the sequencing read (i.e. cycle 1 to 36). For each sequencing cycle $j$ in the read, we calculated the probability of observing a base $x$, where $x \in B, B = \{A, C, G, T, N\}$, given a base $n$ in the reference sequence such that $n \in \mathbb{R}, \mathbb{R} = \{A, C, G, T\}$. Due to observed variability in sequencing errors between machine runs (Figure 2.1B), we created a new error model for each machine run. SNPSeeker uses a 2$^{nd}$ Order Model which assumes a 2$^{nd}$ order dependency between reference sequence...
Pooled-sample versus dbSNP average heterozygosity. Of the 44 SNPs that were described in dbSNP, 37 had average heterozygosity values listed. These values (with standard error bars) are plotted on the Y-axis against the corresponding average heterozygosity value as determined from the SNP frequency in the pooled-sample (correlation coefficient $r^2 = 0.82$). The solid line is a plot of the idealized 1:1 correlation between data sets and the dashed line is the actual linear regression.

nucleotides. Thus, we assume that the likelihood of $i$ being sequenced correctly depends on $i - 1$ and $i - 2$ (this model is computed starting from read cycle 3). We found that the first twelve bases of each Illumina read contained significantly fewer errors than later bases (Figure 2.1A) so we only used these bases to identify sequence variation. Since we use a 2nd order dependency model for SNP identification, only mismatches at bases 3-12 of each read were used to identify SNPs. On average, each position in the reference sequence was
observed 30,593 times in bases 3-12 of the Illumina sequencing reads (i.e. 13.8-fold coverage of 1,111 diploid genomes).

Illumina Quality Scores

Our error models do not take into account the ILLUMINA quality scores since we did not find any improvement by including them in our error models. As shown in Figure 2.4 we have plotted the quality scores (QS) generated by the Illumina Genome Analyzer against the true quality scores, defined as the $-\log_{10}$ of the probability of incorporating a nucleotide different from the reference sequence base (calculated from the negative control vector as described). As shown in Figure 2.4 quality scores less than 10 are informative of poor bases, but higher quality scores do not accurately reflect true error rates, as is evident by the plateau in the trend line for quality scores above 10. Since most bases with quality scores less than 10 occured at the ends sequencing reads, and this information is already accounted for by our read position parameter, incorporating QS into the algorithm did not improve SNP calling. However, for other applications needing to utilize QS data, these measurements are easily integrated into the Large Deviation Theory framework.

Non-specific PCR amplification analysis.

To assure that non-specific PCR amplification across the human genome was not a source of error in our SNP calling, we performed the following analyses. First, all primer combinations were designed to avoid repetitive regions of the human genome. Second, we performed
Illumina Genome Analyzer I quality score analysis. The quality scores generated by the Illumina Genome Analyzer I are plotted along the X-axis while the true quality scores, defined as log10 of the probability of incorporating a nucleotide different from the reference sequence bases (as calculated from the negative control vector as described).

in silico PCR for each primer combination (via the UCSC Genome Browser) against the entire human genome to demonstrate that a single, unique PCR product, spanning the region of interest, was the only product expected. Finally, we used our alignment algorithm to map back all sequencing reads against the entire human genome. We found that 544,195 reads (1.45% of the total) mapped back to more than one location in the human genome. We then excluded these reads from further SNP calling analysis and found that none of our 64 SNP
calls were significantly altered in terms of identity or frequency. These results indicate that non-specific PCR amplification is not adversely affecting SNP identification.

How SNPSeeker improves SNP calling

A single allele, occurring in a population of 2222 alleles, has a frequency of 0.00045. Despite limiting the data for SNP calling at bases 1-12, the average likelihood of an error across these bases was 0.00065 and rose dramatically as more bases were included in the analysis. In order to identify, with a high degree of certainty, true polymorphisms that occurred in the pool at a frequency less than the incipient error rate of the sequencing platform, we designed SNPSeeker, an algorithm based on Large Deviation Theory, and implemented SNPSeeker into the analysis. Figure 2.5 demonstrates how SNPSeeker further refines SNP identification above and beyond simply using the first 12 bases of each read. When considering pUC19 and trying to identify a single allele in the pool, using 32 bases per read identified 785 bases (out of 800 in the training set) as potential SNPs. When only considering the first 12 bases per read without using the algorithm, 705 bases out of 800 were identified as SNPs. By implementing SNPSeeker, zero positions were identified as SNPs. Figure 2.5 demonstrates how each of these conditions would affect SNP calling in the SFTP B amplicon. Using 12 bases and a frequency cutoff calculated on pUC19, we identified 19 potential SNP sites. By applying SNPSeeker, over 50% of the sites are eliminated and only 9 SNP positions remain.
Figure 2.5. SNPSeeker improves SNP calling by removing sequencing errors

How SNPSeeker improves SNP calling. In each panel, the X-axis depicts the sequential 665 bp of the SFTPB amplicon and the Y-axis is the percent likelihood of a mismatched base when compared to the reference sequence. (a) When using 32 bases per read to perform SNP calling, there are 218 positions in the SFTPB amplicon that are considered likely to contain a SNP. (b) When using only the first 12 bases of each read and a frequency cutoff calculated by the error model generated from pUC19 data, there are 19 base positions considered to have potential SNPs. INSET: Ten-fold higher resolution plot showing a maximum 1% likelihood of mismatch. (c) When implementing SNPSeeker on the data from cycles 1-12, only 9 SNP positions remain. INSET: Ten-fold higher resolution plot showing a maximum 1% likelihood of mismatch.
Comparative Genomics Analysis Demonstrates that Rare Non-synonymous SNPs are Deleterious

Twelve nonsynonymous SNPs from all four genes tested were identified in this analysis. Of these 12 sites, SIFT identified 7 as deleterious, PolyPhen 3, and the LRT predicted that 5 would disrupt highly significant positions (dN/dS <1 and LRT, $P <0.001$) (Supplementary Table 5). Five of the 7 sites identified by SIFT are not found in dbSNP. One of these five, position 112207052 in APC (marked with †), was validated by Taqman assay. Four of the five non-dbSNP sites (2 in TP53 and 2 in APC, marked with asterisks), including APC 112207052, were previously published in the germline of individuals with cancer [128] [129] [130] [131]. Four of the five evolutionarily conserved amino acid positions identified by the LRT are perfectly conserved across all species. If recessive, the phenotypic effects of the deleterious SNPs should rarely be observed and may be quite severe.

2.2.4 Discussion

A deeper understanding of genetic variation in the human population will allow us to dissect the causative factors that contribute to a wide array of human disease, understand the genetic characteristics that make us uniquely human, and quantify the impact of selection across our genome throughout history. We have successfully resequenced 13,237 bases per 1,111 individuals at approximately 2% of the cost of the original analysis by Sanger sequencing [132]. Importantly, this cost savings did not come at the price of sensitivity or accuracy.
The positive predictive value of 87% obtained in this study is consistent with previously published values of 85% \cite{133} and 92% \cite{134} in assays identifying SNPs from multiple organisms via second-generation sequencing, but we analyzed over an order of magnitude more individuals (1,111 vs. 2 \cite{133} or 66 \cite{134}) than these studies. The ability to pool larger numbers of individuals enables the discovery of rare SNPs, which is important, since most deleterious SNPs are unlikely to be present at frequencies greater than 1%. Furthermore, our method more accurately estimates MAFs ($r^2 = 0.96$ vs. $0.67$ \cite{134}), which is important to accurately identify disease-associated alleles when comparing disease and normal cohorts.

At the SFTBP locus, we successfully detected a single mutant allele in a background of 2,221 wild-type alleles; however, there were not enough private mutations in our validation set to determine the sensitivity of our method for the detection of private SNPs (MAF $< 0.05\%$) in this population. For applications where it is important to detect singleton SNPs with a high sensitivity, we recommend choosing a pool size such that private mutations are present at frequencies similar to those of the rare SNPs validated here (MAF = 0.5 $\sim$ 1.2%).

There are various applications for this method. Sequencing large, random populations at various genetically significant loci would enable the study of human-specific variation and selection. Quantification of rare somatic mutations in tumors and precancerous lesions would facilitate improved understanding of tumorigenesis. Finally, sequencing case-control or matched sample cohorts will enable identification of rare mutations associated with complex diseases \cite{16,17}. Candidate genes can be selected based on prior knowledge \cite{16,17} or they can be informed by genome wide association studies. Combining pooled-sample se-
quencing with genomic selection strategies [135, 136, 137] makes it possible to move beyond
the candidate gene approach and perform a more systematic survey of protein-coding DNA.
Such knowledge would be a valuable tool for disease screening, assigning risk stratification,
providing longitudinal preventative care, and tailoring risk-appropriate therapy.

2.2.5 Materials and methods

Genomic DNA Samples

We extracted genomic DNA from 1,111 random, anonymous Guthrie cards collected
for newborn screening between 1993 and 2000 by the Missouri Department of Health and
Senior Services (DHSS) [132, 138]. Both the Missouri DHSS Institutional Review Board
and the Washington University Human Research Protection Office reviewed the project and
approved waiver of individual consent for use of de-identified DNA samples under regula-
tion 45CFR467.116d (for the regulation, see http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm#46.116). Each individual DNA sample was anonymously linked to
clinical characteristics in a vital statistics (birth-death certificate) database maintained by
the Missouri DHSS to determine ethnicity. Ethnicities within the population were as follows:
European-American = 871 (78.4%), African-American = 196 (17.6%), Hispanic = 34 (3.1%),
Asian = 5 (0.5%) and unknown = 5 (0.5%).
Genomic DNA Extraction

Genomic DNA extraction was adapted from previously described methods by Hamvas et al. [132]. Modifications to this process included sequential incubation in 200 µL distilled water at room temperature for 60 min and 45 min on an orbital shaker at 400 rpm. DNA was extracted in a solution of 200 µL of 10 mM Tris with 1 mM EDTA, pH 6.8 and 2% (wt/vol) Chelex 100 Molecular Biology Grade chelating resin (Bio-Rad, Hercules, CA, U.S.A.).

DNA Quantification

High-throughput DNA quantification was performed using a fluorescent nucleic acid stain in a 384-well format. To mimic the fragmented and denatured quality of the sample DNA, commercial human genomic DNA at 163 ng/µL (Promega, Madison, WI, U.S.A.) was sonicated for 15 seconds at maximal power using a Misonix XL2020 Ultrasonic Processor sonicator (Misonix, Farmingdale, NY, U.S.A.) and then heated at 100°C for 25 minutes. This DNA was then serially diluted by 50% eight times. DNA concentrations of 20.38, 10.19, 5.09, 2.55, 1.27 and zero ng/µL were added, as described below, to the 384 plate in duplicate to be used as a DNA standard. SYBR Gold (Molecular Probes, Eugene, OR, U.S.A.) was the fluorescent nucleic acid stain used due to its ability to bind both single and double-stranded DNA. Immediately prior to each quantification experiment, the stain was diluted 1:1000 in 10 mM Tris with 1 mM EDTA, pH 7.5 and protected from light. For quantification determination, each well on a black 384-well polystyrene Fluotrac 200 microtiter plate (Greiner Bio-One, Monroe, NC, U.S.A.) contained a final volume of 20 µL consisting of 2 µL DNA
suspension, 8 μL of 10 mM Tris with 1 mM EDTA, pH 7.5 and 10 μL SYBR Gold 1:1000 suspension. The plate was protected from light until processing. Fluorescent detection was done on a Synergy HT (Biotek, Winooski, VT, U.S.A.) plate reader. With excitation at 485/20 and emission 528/20, the machine automatically determined the optimum sensitivity level for each experiment by scaling fluorescence against the negative control wells. The fluorescent plate reader then determined the fluorescent level in each well on the plate. A mean fluorescent value was determined for each concentration of the DNA standard and a linear regression was then generated. The equation of the linear regression was then used to calculate the concentration of DNA in each sample well.

**Pooling of DNA**

We elected to pool 80 ng of DNA per individual. This amount was chosen simply to have a pool of DNA large enough to perform multiple PCR and sequencing reactions. Pooling was done manually and the final volume was over 29 mL. Concentration of pooled DNA was done using the Qiagen QIAvac 96-well vacuum manifold and QIAquick 96 PCR Purification Kit (Qiagen, Valencia, CA, U.S.A.). This kit is intended to purify DNA from 100 bp to 10 kb. Gel electrophoresis of the highly fragmented genomic DNA isolated from blood spots shows a mean size of approximately 3 kb with no visible smear above 10 kb (not shown). The resulting DNA suspension was approximately 2 mL with a concentration, determined by SYBR Gold staining, of 2.6 ng/μL.
PCR primers were designed using Primer 3 (http://frodo.wi.mit.edu/primer3/input.htm) with parameters previously described [139]. All primers were purchased through Integrated DNA Technologies (Coralville, IA, U.S.A.). Reference sequence for each human gene was obtained from the UCSC Genome Browser (http://genome.ucsc.edu/index.html) and, for the pUC19 vector, from New England Biolabs NEBcutter website (http://tools.neb.com/NEBcutter2/index.php). Refseq accession numbers for human reference sequences were: TP53, NM000546; SFTPB, NM000542; APC, NM000038; ACTB, NM001101. Each individual PCR reaction was designed to include an average of 30 genome equivalents per individual. Assuming a Poisson distribution for the number of molecules input into the pool, this amount was determined to minimize the likelihood that one individual DNA sample would be omitted in a PCR reaction, while keeping the volume of PCR reactions within common standards as well as optimizing the allocation of the pooled DNA. The thermostable polymerase utilized was PfuUltra High-Fidelity DNA Polymerase (Stratagene, La Jolla, CA, U.S.A.) due to its reported extreme accuracy. PCR reaction contents were as follows: 1X final concentration of 10X PfuUltra Buffer, 200 µM dNTPs (Invitrogen, Carlsbad, CA, U.S.A.), 400 nM forward primer, 400 nM reverse primer, 1M betaine (Sigma-Aldrich/Fluka, St. Louis, MO, U.S.A.), 2.5 units PfuUltra DNA polymerase and 150.8 ng of pooled genomic DNA. The final reaction volume was 100 µL. Annealing temperatures and extension times varied slightly between reactions and are listed, along with primer sequences, in Supplementary Table 1. In general, all PCR reactions were as follows: 1) 93°C x 2 minutes; 2) 93°C x 30
seconds; 3) 56-62°C x 30 seconds; 4) 65°C x 1-5 minutes; 5) steps 2-4 for 28 cycles; 6) 65°C x 10 minutes; 7) hold at 4°C. In total, there were 14 human PCR amplicons covering 13,237 bp and 19 exons in the four genes samples as well as a single amplicon from the pUC19 vector. All PCR products then went through the QIAquick PCR purification protocol (Qiagen, Valencia, CA, U.S.A.) and quantified using the Nanodrop ND-1000 spectrophotometer.

Random Amplicon Ligation.

Next-generation sequencing was designed to sequence major portions of or even whole genomes in a single machine run. Equal coverage across the genome is dependent upon random fragmentation followed by aligning and tiling the millions of small DNA sequences against the reference genome. Our computational analysis suggested that random fragmentation by sonication cannot be achieved for small DNA sequences such as PCR amplicons less than about 1500 bp (not shown). A majority of such fragments would be fragmented in the middle of the sequence resulting in overrepresentation of the unfragmented end pieces of the amplicon in the final sequencing output. To overcome this, we randomly ligated a normalized number of pooled PCR-amplified products and sonicated the resulting concatamers. From each PCR product, 40x10ⁱ⁰ molecules of each amplicon were pooled. A blunt ended ligation (PfuUltra generates blunt ended products) was performed as follows: 1X of 10X T4 DNA Ligase Buffer (New England Biolabs, Ipswitch, MA, U.S.A.), 2400 units of 400 units/µL T4 DNA Ligase (New England Biolabs, Ipswitch, MA, U.S.A.), 120 units of 10 units/µL T4 Polynucleotide Kinase (New England Biolabs, Ipswitch, MA, U.S.A.), 15%
polyethylene glycol 8000 MW (Sigma-Aldrich, St. Louis, U.S.A.) and the PCR amplicon pool brought the total volume to 600 µL. This volume was aliquoted into four equal parts and incubated for 17 hours at 22°C, followed by 20 minutes at 65°C and held at 4°C thereafter. Agarose gel electrophoresis was performed with a small amount of the resulting product to confirm concatenation. Electrophoresis confirmed concatenated products >10kb (not shown).

Sonication

Random fragmentation of the ligated concatemers was performed using the Covaris S2 sonicator (Covaris, Woburn, MA, U.S.A.). Each 150 µL ligation aliquot was diluted in 350 µL of sterile, distilled water and transferred to a 13mm x 65mm borosilicate glass tube with polypropylene screw-top (Covaris, Woburn, MA, U.S.A.). The samples were sonicated individually with the sonicator programmed as follows: 15 cycles, duty cycle 20%, intensity 10, cycles/burst 1000, time 60 seconds, bath temperature limit 20°C. Agarose gel electrophoresis confirmed fragmentation of concatenated DNA as a smear from approximately 75-3000 bp with maximal ethidium bromide staining intensity between 150-200 bp (not shown).

DNA Library Preparation for Sequencing

Following fragmentation, a DNA library for sequencing was prepared according to the protocol described in the Preparing Samples for Sequencing Genomic DNA document provided by Illumina for the Genome Analyzer and starting with the end repair step. A difference
from the described protocol included gel excision and purification of fragments from 125-400 bp. The concentration of PCR-enriched, adapter-ligated DNA fragments following clean up was determined by Nanodrop ND-1000 spectrophotometer and found to be 29.3 ng/µL.

Sequencing

While we could have generated all data from a single flowcell, we were sharing flowcell lanes with other investigators. Therefore, the same DNA library was sequenced in a total of 12 flowcell lanes from four different dates. The protocol for preparing DNA samples for sequencing is described on page 12 of the Illumina document entitled Preparing Sample DNA for Cluster Generation. Sample DNA was diluted to 10 nM in Qiagen EB buffer as recommended. In the initial sequencing run, a titration of DNA concentrations of 0.5, 1 and 2 pM was performed in three separate lanes to determine optimum cluster generation. Two pM was determined to be optimal and all subsequent sequencing was performed at this concentration. From the same library, we performed an additional 9 lanes of sequencing on three separate dates.

TAQMAN Assays

For independent, individual validation of putative SNPs, we performed Taqman assays on all individuals in our pool. The primers / probes design, manufacture, and testing is done at Applied Biosystems (ABI) manufacturing facilities. Each designed probe set contains 2 probes, with VIC (allele 1) and FAM (allele 2) reporter dyes linked to the 5’ end of each
respective probe, a MGB, and a nonfluorescent quencher (NFQ) at the 3’ end of both probes.

The primers and probes for the nine sites tested in this study are listed in **Supplementary Table 2**. Sites found in dbSNP had pre-existing primers and probes that are commercially available and proprietary to ABI. For each SNP site of interest, we submit a sequence file identifying the polymorphic base, along with 300 flanking bases both 5’ and 3’ of each SNP site to ABI. ABI then designs an assay using proprietary algorithms that minimize adverse assay effects, such as base runs or secondary structure formation. Primers and probe sequences are matched by melting temperature, permitting universal PCR assay conditions. They also use mass spectrometry to verify the oligonucleotide sequence and perform further testing to ensure proper formulation of the primer and probe mix. The assays are also functionally tested using an allelic detection test prior to delivery, then delivered in a single tube format. We assayed 12 96-well plates with no-template (blank) controls in wells H10, H11, H12 of each plate for the nine SNP positions listed in **Supplementary Table 2**. For each plate, a master mix was made of 1,250 µl 2X Taqman genotyping master mix (ABI, part # 4371357) and 125 µl 20X SNP genotyping assay mix (ABI, part # 4332027). Next, 13.75 µl master mix is aliquoted into each well of a 96-well optical reaction plate (ABI, part # 4346906). Then, 1 µl of genomic DNA from each test plate, along with 10.25 µl DNase-free water, is added to each well of the optical reaction plate. Plates are then covered with an optical adhesive film (ABI, part # 4311971) and PCR is performed on an MJ PCT-225 thermocycler. An initial enzyme activation is required at 95C for 10 minutes, followed by 40 cycles of denaturing at 92C for 15 seconds and anneal/extension at 60C for 1 minute.
Upon completion of the PCR amplification, endpoint plate reading and genotype calls are performed on the ABI 7500 FAST Real Time PCR system.

**Large Deviation Theory applied to Illumina Genome Analyzer output.**

We found that existing second-generation base calling programs were unable to detect and quantify rare variants in a large pool of multiple individuals (Vallania and Mitra, unpublished results), so we developed a new base calling method based on Large Deviation Theory and named SNPSeeker. All the software and sequence data used for the analysis is available at [http://www.genetics.wustl.edu/rmlab/](http://www.genetics.wustl.edu/rmlab/). Sequences were mapped back to reference by using an ungapped alignment algorithm and allowing 2 mismatches (this allowed unambiguous mapping) and then positions were considered separately. For each position $i$ in the reference, all the sequence bases aligned to that position are analyzed. This analysis relies on the assumption that the sequencing process generates sequences independent from one another. A second assumption is that sequences are formed by nucleotides (corresponding to sequencing cycles) that are independent from one another as well.

Given these assumptions, it is possible to define subsets of nucleotides for each cycle $j$, sequencing run $d$ and strand $s$. These nucleotides are drawn from the same background probability distribution and, as part of our assumptions, are independent from one another. Each set can be therefore defined as a series of $n$ i.i.d. random variables (where $n$ is the number of total bases for each considered subset)

$$X_{j,d,s,1}, X_{j,d,s,2}, X_{j,d,s,3}, \ldots, X_{j,d,s,n}$$
drawn according to a distribution $Q_{j,d,s}(x)$ with symbols $\chi = \{A,C,G,T,N\}$. For each subset, it is possible to define an empirical probability distribution, also known as type, or $P_{j,d,s}$, as

$$P_{j,d,s} = \left( \frac{A_s}{n}, \frac{C_s}{n}, \frac{G_s}{n}, \frac{T_s}{n}, \frac{N_s}{n} \right)$$

If we assume the null hypothesis of observing no polymorphism at position $i$, $i-1$ and $i-2$ in the reference sequence, then $Q_{j,d,s}(x)$ will exactly correspond to the distribution of errors derived from pUC19 as described in the previous paragraph. However, this can be generalized into

$$Q_{j,d,s}(x) = \sum_{n \in \mathbb{N}} Pr(x|M_i = n, j, d) \ast Pr(M_i = n|s, \tau)$$  \hspace{1cm} (2.1)

where:

- $Pr(x|M_i = n, j, d)$ corresponds to the distribution of errors computed from pUC19 and indicates the probability of seeing a base $x$ in the sequence at cycle position $j$ on run $d$ given that the original base at position $i$ in the reference, $M_i$, is equal to $n$, where $x \in B$ and $n \in \mathbb{N}$ (see Error Model Generation using pUC19 section).

- $Pr(M_i = n|S,s,\tau)$ corresponds to the probability of observing nucleotide $n$ in the reference sequence at position $i$, $M_i = n$, given the strand $s$ and the true allele frequency vector $\tau$. 

78
To determine the likelihood of observing a SNP at position $i$, we compute the amplitude of the deviation between $P_{j,d,s}$ and $Q_{j,d,s}$, formulated as the probability of observing a number of nucleotides different from the major allele (the nucleotide in the reference at position $i$) equal or greater than the number observed in the analyzed subset. According to Sanov’s theorem \[140\], this is equivalent to

$$Q_{j,d,s}^n(E) = 2^{-n_{j,d,s}} D(P_{j,d,s} \parallel Q_{j,d,s})$$

(2.2)

id est the probability of generating a set of types $E$ satisfying the above described conditions. $D(P_{j,d,s} \parallel Q_{j,d,s})$ corresponds to the relative entropy or Kullback–Leibler distance \[4\] between $P_{j,d,s}$ and $Q_{j,d,s}$. Since we expect true SNPs to be detected on both strands and errors to behave independently with respect to the strand, we calculated a cumulative p-value for each strand as

$$Q_s(E) = \prod_d \prod_j Q_{j,d,s}^n(E)$$

(2.3)

given the previous independence assumptions. Each probability value $Q_{j,d,s}^n(E)$ is Bonferroni-corrected for the total number of tests performed at position $i$ and $Q_s(E)$ is additionally corrected for the total number of tests performed at each position in the reference sequence (corresponding to its length). In order for position $i$ to be called as a SNP, $Q_s(E)$ must be below a significancy cutoff ($\alpha = 0.05$) given the appropriate corrections. If this is true for both strands then position $i$ contains at least one allele variant.
Estimation of the true SNP frequencies by Non-Linear Least Square Fit.

For every position in which a SNP was found, we estimated its true allele frequencies by performing a non-linear least square fit. Given the relative entropy term in (2.2), we can decompose it as

\[ n_{j,d,s} D(P_{j,d,s} \| Q_{j,d,s}) = -n_{j,d,s} \left[ H(P_{j,d,s}) + \sum_{x \in B} P_{j,d,s}(x) \log_2 Q_{j,d,s}(x) \right] \] (2.4)

Given \( P_{j,d,s} \) and \( Q_{j,d,s} \), their relative entropy will converge to 0 the closer they will get to each other. That will correspond to a decrease in the difference between \( n_{j,d,s} H(P_{j,d,s}) \) and \(-n_{j,d,s} \sum_{x \in B} P_{j,d,s}(x) \log_2 Q_{j,d,s}(x)\) so that eventually they will converge to

\[ n_{j,d,s} H(P_{j,d,s}) = -n_{j,d,s} \sum_{x \in B} P_{j,d,s}(x) \log_2 Q_{j,d,s}(x) \] (2.5)

If we repeat this consideration for each cycle \( j \), day \( d \) and strand \( s \), we can define an observation vector \( y \) and a least-square vector \( \hat{y} \) such that

\[
y = \begin{bmatrix} n_{1,1,+} H(P_{1,1,+}) \\ \vdots \\ \vdots \\ \vdots \\ n_{j,d,s} H(P_{j,d,s}) \end{bmatrix} \quad \hat{y} = \begin{bmatrix} -n_{1,1,+} \sum_{x \in B} P_{1,1,+}(x) \log_2 Q_{1,1,+}(x) \\ \vdots \\ \vdots \\ \vdots \\ -n_{j,d,s} \sum_{x \in B} P_{j,d,s}(x) \log_2 Q_{j,d,s}(x) \end{bmatrix}
\] (2.6)
As defined in (2.1), $Q_{j,d,s}$ will depend on $Pr(M_i = n|s, \tau)$ which is conditioned on $\tau$. $\tau$ is defined as a stochastic vector with 4 entries, each representing the probability of seeing a particular nucleotide in a given position in the reference sequence (with respect to the forward strand as the adopted convention). The $\tau$ vector that minimizes $|y - \hat{y}|$ is defined $\hat{\tau}$ and corresponds to

$$\hat{\tau} = \arg \min_{\tau} \sum_{j} \sum_{d} \sum_{s} \left\{ n_{j,d,s} \left[ H(P_{j,d,s}) + \sum_{x \in B} P_{j,d,s}(x) \log_2 \sum_{n \in \mathbb{N}} Pr(x|M_i = n, j, d) * Pr(M_i = n|S, s, \tau) \right] \right\}^2$$

(2.7)

This is computed iteratively, initially defining all possible $\tau$ probability vectors starting with a resolution $\rho$ of 1 significant digit (0.1) and then selecting the vector that minimizes $|y - \hat{y}|$. After this step, $\rho$ is decreased 10-fold (0.01 or 2 significant digits) and only vectors located within a range defined as $\hat{\tau} \pm 20 * \rho$ for each vector entry are then analyzed. This allows the values of $\hat{\tau}$ to be refined without requiring massive computational power. These steps are iterated progressively until the estimate at each step is refined to resolution of 3 significant digits (the final resolution is a user defined parameter).

**Comparative Genomics Analysis using SIFT, PolyPhen, and the Likelihood Ratio Test**

While most common SNPs are likely neutral, 50% of rare (< 5% in the general population) nonsynonymous SNPs have been estimated to be deleterious [141], many of which may
produce significant phenotypic effects, even if heterozygous. We thus sought to determine if any of the identified nonsynonymous SNPs are deleterious. The genome sequences of multiple vertebrate species make it possible to identify functional sequences by their conservation across species. Given enough evolutionary time, even a single amino acid position has a very small probability of being conserved by chance.

To test whether any of the nonsynonymous SNPs are deleterious, we used three different prediction algorithms. Two of the algorithms, SIFT [142] and PolyPhen [143], make predictions based on conservation and structural motifs, respectively. However, without a formal probabilistic framework the rate of false positive predictions is difficult to know. The third algorithm is based on a likelihood ratio test (LRT) that compares the probability of conservation across species at a single amino acid position under a neutral model and a model of selective constraint [144]. Under the neutral model, the probability of amino acid conservation is calculated using the synonymous substitution rate for each gene. Under the model of selective constraint, the nonsynonymous substitution rate is allowed to be a fraction of the synonymous rate. This codon based test is similar to a test developed for noncoding sequences.

For each gene, coding sequences were downloaded from ENSEMBL (www.ensembl.org), the translated protein sequences were aligned using ClustalW, and then translated back into their corresponding DNA sequences. The number of species ranged from 15 (ACTB) to 21 (APC) with the most distant species ranging from Platypus (SFTPB) to Zebrafish (TP53). The maximum likelihood phylogenetic relationship of each gene was obtained using Phylip
and the maximum likelihood estimate of the synonymous substitution rate was obtained using HyPhy and the MG94xHKY85 3x4 nucleotide substitution model. The total synonymous substitution rate ranged from 4.37 for SFTPB to 14.98 for TP53. The likelihood ratio test compares two models in order to determine whether a nonsynonymous SNP is deleterious and disrupts a conserved amino acid position. Under the null model, the likelihood of the data was calculated assuming no constraints such that both the synonymous and nonsynonymous substitution rate of the SNP containing codon were equal to the synonymous rate for the entire gene. Under the alternative model, the likelihood of the data was calculated assuming the SNP containing codon was constrained such that the nonsynonymous rate was a fraction of the synonymous rate. Deleterious SNPs were defined as those for which nonsynonymous rate was significantly less than the synonymous rate.

\textit{ACTB}

- Species: 16
- Tree order: human, chimp, pongo, macaca, treeshrew, mouselemur, dog, elephant, squirrel, rat, mouse, opposum, platypus, chicken, xenopus
- dS: 5.76853
- dN: 0.24075

\textit{APC}

- Species: 21
• Tree order: human, chimp, macaque, mouselemur, elephant, tenrec, cat, dog, cow, microbat, hedgehog, shrew, guineapig, mouse, rat, treeshrew, pika, rabbit, oppossum, platypus, xenopus

• dS: 8.48990

• dN: 0.68431

\textbf{SFTBP}

• Species: 15

• Tree order: human, chimp, orangutan, macaque, elephant, tenrec, cow, bat, dog, horse, rat, mouse, bushbaby, mouselemur, hedgehog, pika, opposum, platypus

• dS: 6.49362

• dN: 1.36192

\textbf{TP53}

• Species: 18

• Tree order: human, chimp, macaque, mouselemur, cat, cow, microbat, shrew, armadillo, mouse, rat, rabbit, pika, xenopus, Fugu, stickleback, medaka, danio rerio

• dS: 28.94836

• dN: 1.54825
2.2.6 Acknowledgments

This work was supported in part by the US National Institutes of Health under the Ruth L. Kirschstein National Research Service Award T32 HD 007499 from the National Institute of Child Health and Human Development (T.E.D.), the National Heart, Lung and Blood Institute (RO1HL065174, RO1HL082747, F.S.C.), the Children’s Discovery Institute Fellowship Award MC-F-2006-1 (T.E.D.), the Children’s Discovery Institute grant MC-II-2006-1 (R.D.M.) and the Saigh Foundation (F.S.C., R.D.M. and T.E.D).
2.2.7 Supplementary tables

Figure 2.6. Supplementary Table 1. PCR primers and conditions

PCR primers and conditions. For each amplified region of genomic DNA, the specific primer combination, size of the resulting amplicon, PCR annealing temperature, and extension time are listed.

Figure 2.7. Supplementary Table 2. Taqman assay primers and probes.

Taqman assay primers and probes. The specific primers (a) and probes (b) for the nine sites tested in this study are listed. Commercially available primers and probes for specific dbSNP sites are proprietary to Applied Biosystems (ABI).
**Supplementary Table 3a.** SFTPb SNP comparison between Sanger and pooled sequencing

<table>
<thead>
<tr>
<th>Genomic Position</th>
<th>dbSNP Accession Number</th>
<th>Substitution/ Gene Location</th>
<th>Function</th>
<th>Percent frequency from pooled sample data</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTB Chr7: 85744339</td>
<td>n/a</td>
<td>CGC(R)→CAC(H) Exon 7</td>
<td>Coding, nonsynonymous</td>
<td>0.5</td>
</tr>
<tr>
<td>TP53 Chr17: 7514622</td>
<td>rs17880847</td>
<td>GAC(D)→GAT(D) Exon 7</td>
<td>Coding, synonymous</td>
<td>4.2</td>
</tr>
<tr>
<td>TP53 Chr17: 7513782</td>
<td>rs17881850</td>
<td>C→A Intron 7</td>
<td>Non-coding</td>
<td>5.8</td>
</tr>
<tr>
<td>TP53 Chr5: 112202922</td>
<td>n/a</td>
<td>A→G Intron 7</td>
<td>Non-coding</td>
<td>21.2</td>
</tr>
<tr>
<td>APC Chr5: 112206391</td>
<td>n/a</td>
<td>TCT(S)→TGG(C) Exon 8</td>
<td>Coding, nonsynonymous</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**Supplementary Table 3b.** Positions validated by Taqman assay.

<table>
<thead>
<tr>
<th>Gene / Genomic Position</th>
<th>dbSNP Accession Number</th>
<th>Substitution/ Gene Location</th>
<th>Function</th>
<th>Percent frequency from pooled sample data</th>
<th>Percent frequency from Taqman assay</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTB Chr7: 85744339</td>
<td>n/a</td>
<td>CGC(R)→CAC(H) Exon 7</td>
<td>Coding, nonsynonymous</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>TP53 Chr17: 7514622</td>
<td>rs17880847</td>
<td>GAC(D)→GAT(D) Exon 7</td>
<td>Coding, synonymous</td>
<td>4.2</td>
<td>3.5</td>
</tr>
<tr>
<td>TP53 Chr17: 7513782</td>
<td>rs17881850</td>
<td>C→A Intron 7</td>
<td>Non-coding</td>
<td>5.8</td>
<td>8.0</td>
</tr>
<tr>
<td>TP53 Chr5: 112202922</td>
<td>n/a</td>
<td>A→G Intron 7</td>
<td>Non-coding</td>
<td>21.2</td>
<td>21.9</td>
</tr>
<tr>
<td>TP53 Chr5: 112206391</td>
<td>n/a</td>
<td>TCT(S)→TGG(C) Exon 8</td>
<td>Coding, nonsynonymous</td>
<td>1.1</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Validated SNPs. (a) SNP comparison between Sanger and pooled-sample sequencing at the SFTPb locus.

(b) The 7 rare SNP positions validated by Taqman assay at the TP53, APC and ACTB loci. The correlation between called and actual frequencies for these 14 SNP positions shown in the main text in Figure 2.8.
## Supplementary Table 4. Known and putative SNP positions identified by pooled-sample sequencing

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position in reference sequence</th>
<th>Reference – Observed</th>
<th>Frequency in data set (%)</th>
<th>Bonferroni-corrected P-values, log10</th>
<th>dbSNP ref # (build 128)</th>
<th>dbSNP average heterozygosity</th>
<th>Amino Acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTB</td>
<td>112202922 (exon 15) A</td>
<td>G -&gt; A</td>
<td>1.2</td>
<td>-111.0 -172.2</td>
<td>rs1800372</td>
<td>0.024 +/- 0.106</td>
<td>synonymous</td>
</tr>
<tr>
<td>TP53</td>
<td>112206694 (exon 15) G</td>
<td>C -&gt; T</td>
<td>0.1</td>
<td>-1.6</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
<tr>
<td></td>
<td>112206391 (exon 15) C</td>
<td>G -&gt; A</td>
<td>0.2</td>
<td>-3.7</td>
<td>rs7803493</td>
<td>0.012 +/- 0.076</td>
<td>Intronic</td>
</tr>
<tr>
<td></td>
<td>7520197 (exon 4) G</td>
<td>A -&gt; G</td>
<td>0.3</td>
<td>-6.3</td>
<td>rs8584899</td>
<td>0.050 +/- 0.071</td>
<td>UTR</td>
</tr>
<tr>
<td></td>
<td>7515072 (intron 8) G</td>
<td>A -&gt; G</td>
<td>0.3</td>
<td>-9.3</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
<tr>
<td></td>
<td>85744264 (intron 7) T</td>
<td>C -&gt; G</td>
<td>0.1</td>
<td>-55.8</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
<tr>
<td></td>
<td>85744311 (exon 7) C</td>
<td>C -&gt; A</td>
<td>0.9</td>
<td>-5.8</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
<tr>
<td></td>
<td>75203344 (exon 3) G</td>
<td>A -&gt; U</td>
<td>0.4</td>
<td>-216.5 -4988.7</td>
<td>rs13447934</td>
<td>0.011 +/- 0.174</td>
<td>Intronic</td>
</tr>
<tr>
<td></td>
<td>75181053 (exon 1) C</td>
<td>C -&gt; T</td>
<td>0.5</td>
<td>-29.0</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
<tr>
<td></td>
<td>5535256 (intron 2) C</td>
<td>T -&gt; G</td>
<td>0.5</td>
<td>-39.9</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
<tr>
<td></td>
<td>75201732 (exon 10) G</td>
<td>C -&gt; T</td>
<td>0.3</td>
<td>-1.6</td>
<td>rs7803493</td>
<td>0.012 +/- 0.076</td>
<td>Intronic</td>
</tr>
<tr>
<td></td>
<td>7515072 (intron 8) G</td>
<td>A -&gt; G</td>
<td>0.3</td>
<td>-9.4</td>
<td>rs7803493</td>
<td>0.012 +/- 0.076</td>
<td>UTR</td>
</tr>
<tr>
<td></td>
<td>7513344 (exon 11) G</td>
<td>C -&gt; T</td>
<td>0.3</td>
<td>-45.5</td>
<td>rs7803493</td>
<td>0.012 +/- 0.076</td>
<td>UTR</td>
</tr>
<tr>
<td></td>
<td>85744339 (exon 7) G</td>
<td>A -&gt; C</td>
<td>0.3</td>
<td>-17.7 -29.7</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
<tr>
<td></td>
<td>7515072 (intron 8) G</td>
<td>C -&gt; U</td>
<td>0.5</td>
<td>-29.8</td>
<td>rs7803493</td>
<td>0.012 +/- 0.076</td>
<td>Intronic</td>
</tr>
<tr>
<td></td>
<td>85744264 (intron 7) T</td>
<td>C -&gt; G</td>
<td>0.5</td>
<td>-24.8 -47.7</td>
<td>rs13447934</td>
<td>0.011 +/- 0.174</td>
<td>Intronic</td>
</tr>
<tr>
<td></td>
<td>75203344 (exon 3) G</td>
<td>G -&gt; C</td>
<td>0.5</td>
<td>-39.9</td>
<td>n/a</td>
<td>wrinkles</td>
<td>synonymous</td>
</tr>
</tbody>
</table>

### Additional Data

- **dbSNP dbSNP average heterozygosity**
- **Amino Acid change**
- **Reference – Observed**
- **Frequency in data set (%)**
- **Bonferroni-corrected P-values, log10**
- **dbSNP ref # (build 128)**

**Figure 2.9.** Supplementary Table 4. Known and putative SNP positions identified by pooled-sample sequencing
Figure 2.10. Supplementary Table 5. Comparative genomics analysis of nonsynonymous
SNPs Table 5. Comparative genomics analysis of non-synonymous SNPs
Supplementary
Gene
TP53

Position in reference Frequency in
sequence
data set (%)

dbSNP ref #
(build 128)

Amino Acid
change

dN/dS
ratio

rs1800371
rs1042522
**
**

CCG(P) TCG(S)
CCC(P) CGC(R)
CCG(P) CTG(L)
GGC(G) GAC(D)

1.83
0.34
0.000
0.000

GCC(A)

GTC(V)

Likelihood
ratio
P-value

SIFT

Polyphen

tolerated
tolerated

benign
benign

not tolerated

probably damaging

not tolerated

probably damaging

0.000

not tolerated

benign

6 x 10

7520273 (exon 4)
7520197 (exon 4)
7519200 (exon 5)
7518272 (exon 7)

0.4
69.2
0.1
0.4

ACTB

5534731 (exon 4)

0.1

APC

112201798 (exon 15)
112202576 (exon 15)
112203139 (exon 15)
112206694 (exon 15)
112207052 (exon 15)

0.5
0.3
1.3
1.6
1.1

rs33974176
**
rs1801166
rs2229995
**†

CCA(P)
TTG(L)
GAA(E)
GGT(G)
TCT(S)

TCA(S)
TCG(S)
CAA(Q)
AGT(S)
TGT[C]

0.33
0.14
0.40
0.00
0.44

tolerated
not tolerated
tolerated

benign
benign
benign

not tolerated
not tolerated

benign
possibly damaging

0.0314
0.0040
0.0479
-6
1 x 10
0.0760

85744339 (exon 7)
85744307 (exon 7)

0.4
0.5

rs3024809
rs36210375

CGC(R) CAC(H)
GCT(A) ACT(T)

0.09
2.24

not tolerated
tolerated

benign
benign

3 x 10
0.0290

SFTPB

0.1717
0.0256
-6
1 x 10
-8
9 x 10
-5

-4

Amino acid conservation
across species
PPPAPPSASYPSSGXXAX
PPPPAAPPPPPPPLEEEE
PPPPPPPPPPPPPPPPPP
GGGGGGGGGGGGGGGGGG
AAAAAAAAAAAAAXA
PPPPAPPTPPPPPPPPPPPPS
LLLLLLLLLLLLLLLLLLLLH
EEEEEQDEEEEEEEEEEEEGD
GGGGGGGGGGGGGGGGGGGGG
SSSSSSSSSSXSSSSSSSATL
RRRGGGGGGGGGGGGGGG
AAATTFLSGTAATIMALT

Supplementary Table 5. Comparative genomics analysis of non-synonymous SNPs demonstrating
amino acid
conservation
among
vertebrates. The 12
non-synonymous
SNPs
identifed
the data
Comparative
genomics
analysis
of non-synonymous
SNPs
demonstrating
amino
acid in
conservation
among
analysis were analyzed for their likelihood of being deleterious. The phylogenetic order of species
at each locus is listed in the Supplementary Methods. SNP positions marked with asterisks (**) have
vertebrates.
The 12 non-synonymous
SNPs identifed
in thewith
data
analysis
were analyzed
their
been published
in the germline of individuals
or families
a history
of various
cancers.for
The
SNPlikelihood
position marked with a dagger (†) was independently validated by Taqman assay.

of being deleterious. The phylogenetic order of species at each locus is listed in the Supplementary Methods.
SNP positions marked with asterisks (**) have been published in the germline of individuals or families with
a history of various cancers. The SNP position marked with a dagger (†) was independently validated by
Taqman assay.

89


2.3 High-throughput discovery of rare insertions and deletions in large cohorts

2.3.1 Abstract

Pooled-DNA sequencing strategies enable fast, accurate, and cost-effective detection of rare variants, but current approaches are not able to accurately identify short insertions and deletions (indels), despite their pivotal role in genetic disease. Furthermore, the sensitivity and specificity of these methods depend on arbitrary, user-selected significance thresholds, whose optimal values change from experiment to experiment. Here, we present a combined experimental and computational strategy that combines a synthetically engineered DNA library inserted in each run and a new computational approach named SPLINTER that detects and quantifies short indels and substitutions in large pools. SPLINTER integrates information from the synthetic library to select the optimal significance thresholds for every experiment. We show that SPLINTER detects indels (up to 4 bp) and substitutions in large pools with high sensitivity and specificity, accurately quantifies variant frequency ($r = 0.999$), and compares favorably with existing algorithms for the analysis of pooled sequencing data. We applied our approach to analyze a cohort of 1152 individuals, identifying 48 variants and validating 14 of 14 (100%) predictions by individual genotyping. Thus, our strategy provides a novel and sensitive method that will speed the discovery of novel disease-causing rare variants.
2.3.2 Introduction

Understanding the genetic basis of common diseases is an important step towards the goal of personalized medicine [10]. At present, two distinct hypotheses are under debate [11, 12]. The Common Variant, Common Disease (CVCD) hypothesis states that disease-causing alleles are common in the human population (frequency > 5%) [13]. In contrast, the Rare Variant, Common Disease (RVCD) hypothesis posits that multiple disease-causing alleles, which individually occur at low frequencies (< 1%), cumulatively explain a large portion of disease susceptibility [15, 17]. Recent evidence favors the RVCD hypothesis as common variants have failed to explain many complex traits [12] while rare genetic variants have been successfully associated with HDL levels [10], blood pressure [17], obesity [145], and colorectal cancer [146, 147]. Due to their low frequencies, identifying rare, disease-associated variants requires genotyping large cohorts in order to reach the appropriate statistical power (for example 5,000 individuals are required to detect mutations present at 0.1% in the population with a probability of 96%). Collapsing methods in which rare variants are grouped together before association with disease have been shown to improve statistical power [148] but analysis of large cohorts is still required. One recent strategy for genotyping large cohorts consists of pooled-sample sequencing, where individual samples are pooled prior to analysis on a next-generation sequencing platform [121, 134, 149, 150, 151]. By leveraging the massively parallel output of second-generation DNA sequencing, pooled-sample sequencing allows fast and accurate detection of rare variants in thousands of samples at a fraction of time and cost of traditional methods. Individual sample identities can be recovered using a combi-
natorial pooling strategy (such as DNA sudoku [151]). Despite the promise of this method for studying rare genetic variants, current computational approaches pose a bottleneck, because they are focused either on single individual genotyping [152] or on the detection of common variants in small sized pools [149]. Our previously developed SNPseeker algorithm allows the detection of single nucleotide substitutions in large pooled samples [121] but still fails to address two important key challenges in rare variant detection. First, to date, no algorithm has been able to detect indels in pools larger than 42 individuals and without the presence of many false positives (∼40%) [149], despite the fact that they account for a quarter of the known mutations implicated in mendelian diseases [10, 153]. In particular, short indels represent the most common type of this class of variation [10] and have been reported to occur as rare germ-line variants associated to genetic disease, such as breast and ovarian cancer [154]. Efforts to detect disease-associated genetic variants will therefore greatly benefit from the ability to accurately detect rare short indels. Second, in order to accurately detect rare variants in a large pooled sample, an optimal significance cutoff for the accurate discrimination of true variants from false positives must be chosen. This parameter is in practice affected by sequencing error rates and average coverage, which have been shown to change for every run [121]. Failure to define an optimal cutoff results in lower sensitivity and increased false positive rate. Since the rare variant hypothesis posits that individual disease-associated mutations will be extremely rare (but cumulatively common), it is absolutely critical to be able to specifically discriminate, in every experiment, a single heterozygous individual in a large cohort from the background noise. Until now, this has
not been reliably demonstrated. To address these important challenges, we have developed
a novel experimental and computational strategy that combines a synthetically engineered
DNA library inserted in each run and a new computational approach named SPLINTER
(Short indel Prediction by Large deviation Inference and Non-linear True frequency Estima-
tion by Recursion). This approach allows accurately detection and quantification of short
insertions, deletions and substitutions by integrating information from the synthetic DNA
library to tune SPLINTER and quantify specificity and sensitivity for every experiment
in order to accurately detect and quantify indels and substitutions (Figure 2.11 and 2.12).
SPLINTER requires the presence of two components: a negative control (1-2 KB cloned
plasmid DNA) used to generate a run-specific error model, and a positive control consisting
of a synthetic DNA library simulating an artificial pool with mutations engineered at known
position and frequency. We tested SPLINTER on synthetically engineered pooled samples
containing different mutations at different frequencies in a variety of sequence context back-
grounds, obtaining 100% sensitivity with no false positives in pools up to 500 individuals.
SPLINTER was also able to accurately quantify allele frequencies predicted and observed
allele frequencies were correlated with a correlation of 0.999. We find that SPLINTER sig-
nificantly outperforms all the other algorithms for the analysis of pooled sequencing data by
being the most sensitive approach despite also returning almost no false positives. We then
applied our strategy to multiple pooled-samples, identifying novel and already described
sequence variants, all of which were independently validated.
Pipeline for the detection of rare substitutions and indels in pooled DNA samples. DNA samples from a selected group of patients are individually pooled in a complex mixture to be used as a template for PCR amplification of selected genomic loci. The pooled PCR products are then combined in an equimolar mix containing a DNA fragment without variants (negative control) and a synthetic pool with engineered variants (positive control). The mix is then sequenced on an Illumina Genome Analyzer IIx. The negative control reads are used to generate an error model to be used in the variant calling phase. The positive control allows determination of the optimal cutoff. SPLINTER will then be used to detect and quantify indels and substitutions present in the pool.
SPLINTER detects true segregated variants by comparing the frequency vector of observed read bases to an expected frequency vector defined by the error model. If the observed vector is significantly different than the expected vector, then SPLINTER will call that position a sequence variant. For each identified variant, SPLINTER will then perform maximum likelihood fit in order to estimate its frequency in the pooled sample.

### 2.3.3 Results

**Detection of rare insertions and deletions in synthetic libraries**

For each experiment, we first pooled equimolar amounts of sample DNA together with the controls and generated a DNA library to be sequenced on the Illumina Genome Analyzer.
IIx sequencing platform. We then mapped back the sequencing reads to their reference and built a run-specific error model from the negative control reads. Next, we optimized our cutoff parameters on the positive control and then called SNPs and indels on our sample (see Supplementary Materials and Methods). We first sought to determine the upper limit of the number of samples that SPLINTER can analyze in a pool. To do so, we generated 3 synthetic DNA libraries each containing 15 different indels and substitutions (Supplementary Table 1,2 and Material and Methods) introduced at frequencies of 0.005, 0.002 and 0.001 respectively (corresponding to cohorts of 100, 250 and 500 diploid individuals). We sequenced these libraries using the workflow depicted in Figure 1. In each instance, SPLINTER was able to correctly identify every variant (15/15 variants) without making false positive calls (2254/2254 true negatives) (Figure 2.15). We concluded that SPLINTER can accurately and reliably detect single heterozygous mutations in pools of up to 500 individuals.

Estimation of required sequencing coverage for optimal indel and substitution detection

We next investigated how SPLINTERs accuracy changed as a function of average sequencing coverage. To do so, we sampled the sequencing data obtained for each of the three previous libraries at different fractions (Supplementary Materials and Methods) and then computed the accuracy of our predictions in form of an area under an receiver-operator curve (AUC), a commonly used metric of accuracy ranging from 0.5 (random guessing) to 1 (100% sensitivity and specificity). By plotting AUC as a function of average sequencing
coverage we found that accuracy increased with coverage, with high-frequency variants requiring less coverage than lower-frequency variants (Figure 2.13 A). By analyzing AUC as a function of coverage per allele, we observed a clear overlap of the curves for each pool, reaching AUC equal to 1 at 30-fold average coverage per haploid genome (Figure 2.13 B), indicating that accurate detection can be achieved given enough coverage independently of pool size. Recent re-sequencing efforts show that indel detection remains challenging as their false-positive rate is 15-fold higher than substitutions [155]. Our initial data suggested that indels can be detected as sensitively and accurately as substitutions. To test this hypothesis, we generated 5 additional DNA libraries with synthetic insertions, deletions and substitutions included at a wide range of frequencies (from 1 to 50 variants in 1000 total alleles) (Supplementary Table 2,4). We achieved 100% sensitivity for all the pools (9/9 indel variants and 10/10 substitution variants), with specificities between 99.91 and 100% (between 2263/2265 to 2259/2259 true negatives). We then plotted the relationship between AUC and coverage for each set. Indels converged to AUC equal to 1 at a rate comparable to substitutions, independently of the frequency of the mutation (Figure 2.14 A,B,C). Thus we conclude that SPLINTER detects indels as accurately and as sensitively as it does substitutions. Since many deleterious indels are 4bp or shorter [10, 154], we wanted to determine whether SPLINTER could accurately detect indels as large as 4bp. We generated and sequenced two synthetic pools containing 8 and 10 4bp indels with frequencies ranging from 0.001 to 0.020 and from 0.025 to 0.045 respectively. SPLINTER achieved 100% sensitivity (10/10 variants) and 100% specificity (2253/2253 true negatives) for allele frequencies between...
tween 0.025 and 0.045 and 100% sensitivity (8/8 variants) and 99.5% specificity (2243/2253 true negatives) between 0.001 and 0.020 (Supplementary Table 3,4). These results suggest that SPLINTER is sensitive and specific in detecting 4bp indels.

Figure 2.13. Relationship between variant detection accuracy and average sequencing coverage per base

(a) Accuracy expressed as AUC (area under the curve) (y-axis) plotted as a function of average sequencing coverage per base (x-axis) for synthetic pools with variants present at frequencies 1/200, 1/500 and 1/1000.

(b) Same as in a with average sequencing coverage per base per allele on the x-axis.

Comparison of SPLINTER with other variant discovery approaches

We next compared SPLINTER to existing tools for variant calling. We used the synthetic DNA libraries previously described to benchmark the sensitivity and positive predic-
Figure 2.14. Relationship between variant detection accuracy and average sequencing coverage per base for substitutions and indels

AUC (y-axis) as a function of average sequencing coverage per base (x-axis) for insertions (a), deletions (b) and substitutions (c). Variants are present at frequencies 1/1000, 5/1000, 10/1000, 15/1000 and 50/1000.

We compared SPLINTER with SNPseeker [121], MAQ [152], SAMtools [156] and VarScan [149] for the detection of substitutions (Figure 2.15 A,B) and with SAMtools and VarScan for the detection of indels (Figure 2.15 C,D). For each data set analyzed, SPLINTER significantly outperformed every other approach. In all of the synthetic libraries containing substitutions, SPLINTER detected all of the synthetic variants with no false positives, thus achieving a 100% sensitivity and specificity. SNPseeker also achieved perfect accuracy in the pool simulating 100 individuals, but had a 20% positive predictive value in the libraries simulating 250 and 500 individuals, and had only an 80% sensitivity in the 500 individual library. The other approaches detected variants with substantially lower sensitivity and positive predictive values in all libraries. For each indel set,
SPLINTER returned all of the true variants with no false positives, except for the indel 1 set and the 4bp 1 set (\(\sim 30\%\) and \(\sim 50\%\) positive predictive values respectively). In comparison, every other approach resulted in false positive rates greater than 80\% while achieving low sensitivity, with the exception of the second 4bp set. We also compared SPLINTER with a recently published new algorithm for pooled DNA variant detection called CRISP [19] for both substitution and indel detection (Figure 2.16). SPLINTER outperformed CRISP in both sensitivity (at most 40\% increment) and positive predictive value (at most 80\% increment). In order to distinguish whether the improved accuracy in variant finding originated from improved alignments or improved variant calling, we also compared the performance of SPLINTER using our alignment algorithm (RAPGAP) versus using reads aligned with Novoalign (www.novocraft.com). Both aligners resulted in a comparable performance in finding true variants (Figure 2.17), although our aligner showed small increases sensitivity and positive predictive value in several of the analyzed pools. This result suggests that improved variant calling accuracy mostly depended on the variant calling algorithm and not the underlying aligner. Taken together, these results demonstrate that SPLINTER outperforms other approaches at detecting single nucleotide substitutions and indels in large pools.

**Estimation of the frequency of rare insertions and deletions in synthetic libraries**

Having established that SPLINTER could detect rare variants in pooled samples, we next examined whether SPLINTER could also accurately determine the frequencies of the identified variants. We compared estimated and expected indel frequencies from all our
Figure 2.15. Comparison between SPLINTER and other variant calling algorithms

Substitutions (a,b) and indels (c,d) were analyzed independently. For each approach, performance was evaluated by assessing sensitivity (fraction of true positive hits divided by total positives in the set) and positive predictive value (fraction of true positive hits divided by total hits).

libraries (frequency range 0.001~0.050) and found a very high correlation (r = 0.969, P < 2.2 e-16; Figure 2.18A) indicating that SPLINTER was able to accurately estimate allele frequencies. We next sought to better understand the causes of the observed errors in our
Performance comparison between SPLINTER and CRISP. For each pool, sensitivity and positive predictive value for SPLINTER (blue columns) and CRISP (green columns) were generated. Both SNPs and indels were considered for this analysis.

Allele quantification can be affected by pipetting errors during DNA pooling and by preferential amplification of specific alleles in the pooled PCR. To distinguish between these two sources of error, we constructed all our plasmids so that each contained 2 mutations spaced far enough apart to be analyzed independently (i.e. with no
Comparison of the effect of different aligner algorithms to SPLINTER performance. For each pool, sensitivity and positive predictive value for SPLINTER run on reads mapped by RAPGAP aligner (blue columns) compared to SPLINTER run on reads mapped by NOVOALIGN (red columns) were generated. Both SNPs and indels were considered for this analysis.
within the same molecule were very highly correlated (r = 0.995, P < 2.2e-16; Figure 2.18 C), indicating that most of the noise in variant quantification was due to experimental error. We similarly observed very high correlations with substitutions (frequency correlation r = 0.956, P < 2.2e-16; pair correlation r = 0.993, P < 2.2e-16; Figure 2.18 B and D) and 4bp indels (frequency correlation r = 0.962, P = 1.501e-11; pair correlation r = 0.939, P = 5.599e-05) (Figure 2.19). Based on these results, we reasoned that robotic pooling of samples might improve allelic quantification. Therefore, we robotically pooled and sequenced a large cohort of 974 people previously analyzed in a GWA study (Materials and Methods). As expected, we observed an almost perfect correlation (r = 0.999, P < 2.2e-16; Figure 2.20) between the GWA frequencies and the frequencies estimated by SPLINTER, indicating that inaccurate pipetting was indeed a primary source of error.

**High-throughput discovery of rare indels in large patient cohorts**

Finally, we applied SPLINTER to a large human cohort as a real-world test of the algorithm. We sequenced 14 loci (2596bp total) in 1152 individuals, which were divided into 9 pools (94 to 178 individuals per pool) (see Materials and Methods). For every sequenced pool, we included a negative and positive control to tune SPLINTER. We identified on average 19 variants per pool (for a total of 151 variants, see Supplementary Table 6). To confirm SPLINTERs accuracy, we examined the overlap of our hits with variants listed in dbSNP. We observed large overlapping fractions – between 68.5% and 100% of the identified variants in each pool could be found in dbSNP (Supplementary Table 5,6). In all
Figure 2.18. Precise quantification of rare genetic variants in synthetic samples

(a,b) Correlation between variant frequency measured by SPLINTER (y-axis) and expected variant frequency (x-axis) from 8 synthetic pools for indels (a) and substitutions (b). (c,d) Pair correlation between mutation pairs present in the same DNA molecule for indels (c) and substitutions (d).
Figure 2.19. Precise quantification of 4bp insertions and deletions

(a) Correlation between variant frequency measured by SPLINTER (y-axis) and expected variant frequency (x-axis) from 2 synthetic pools containing 4bp indels.  (b) Pair correlation between mutation pairs present in the same DNA molecule.

cases, statistical significance was reached (Fishers Exact Test, Supplementary Table 7). We selected 14 variants (3 novel variants and 11 from dbSNP) from the largest analyzed pool for independent validation by individual genotyping using the Sequenom iPLEX platform. All 14 variants were confirmed, resulting in 100% positive predictive value. Furthermore, allele frequencies were highly correlated with those estimated by SPLINTER (r = 0.985, P = 5.958e-09, Supplementary Table 8 and Figure 2.21). Together, these results demonstrate the utility of the SPLINTER methodology for the rapid analysis of large populations of individuals. All the computational tools, source code and the experimental datasets presented
Figure 2.20. Precise quantification of rare genetic variants in real samples

Correlation between variant frequency measured from GWA study and SPLINTER estimated frequency.

in this paper can be accessed at http://cgs.wustl.edu/~fvallania/4_splinter_2010/5_splinter_webpage/SPLINTER_supporting_material.html

2.3.4 Discussion

Rare genetic variation is likely to describe a substantial portion of heterogeneity in common and complex diseases. Identifying disease-associated rare variants requires the analysis of multiple loci in large cohorts. We have shown that a novel experimental design combined with SPLINTER can accurately identify genetic variants in large pools, leading to several
Figure 2.21. Correlation between variant quantification by SPLINTER Vs Sequenom

Correlation between variant frequency measured from single individual genotyping by Sequenom and SPLINTER estimated frequency.

advantages over other computational strategies. First, we found that SPLINTER identified genetic variants with high sensitivity and precision, whereas the other methods were unable to detect a large fraction of the variation present in the samples. We found that a sequencing coverage of \(\sim 30x\) per haploid genome was required to detect mutations with high sensitivity and specificity. In earlier work, we successfully analyzed pooled-samples using SNPseeker at lower sequencing coverage (13.8-fold per haploid genome) \[121\] However, in that study most of the variants were present in many individuals in the pool, suggesting that, in order
to detect singleton alleles with $\sim 100\%$ confidence in a variety of different sequence contexts, a higher sequence coverage is required. This finding is confirmed by recent re-sequencing studies of single cancer genomes where near optimal accuracy of somatic SNP detection (3% false-discovery rate) was achieved at $\sim 40$-fold average haploid genome coverage \[155\], and by the lower performance of SNPseeker when compared to SPLINTER in detecting substitutions present at 1 in a 1000 in both sensitivity and precision. Second, our strategy incorporates a synthetic positive control and a negative control, which allow estimation of sensitivity and specificity for each experiment. This is important because run-to-run variations in sequencing error rates can influence accuracy and perturb the optimal p-value cutoffs. The inclusion of the control DNA has a negligible impact on experiment cost. One single-end sequencing lane ($\sim 30$ million 36bp long reads per lane) can provide enough coverage to analyze 25 Kb of genomic DNA in 500 patients, with the control sequences accounting for $\sim 4\%$ of the total sequencing data. Third, SPLINTER can accurately and sensitively detect indels with a high sensitivity and accuracy. Detection of indels even in single genome resequencing studies is indeed a challenging problem due to the difficulties in reducing the false positive rate while retaining good sensitivity \[155\]. In addition, previously published approaches cannot detect indels \[121, 152\], or can only be applied to small-sized cohorts (42 people)\[149\]. Together, these issues have limited the application of pooled-DNA sequencing. We have shown here that SPLINTER can accurately discriminate single indels in pools as large as 500 individuals with high sensitivity and specificity. In comparison, the best performing algorithm achieved at best 80% false positive rate. Fourth, SPLINTER can accurately quantify the frequency
of the alleles present in the pool. Although high correlations between real and estimated frequencies were observed, small discrepancies may result in errors in variant association to a phenotype if the variant is rare and the effect of the variant is high. Our pair correlation analysis shows that the major source of errors in quantification does not come from SPLINTER but rather from pipetting errors in pool construction, as indicated by the improved correlations after robotic pipetting of the pools. This issue can in fact be resolved by performing orthogonal validation of the samples, which will be highly facilitated by the overall performance of SPLINTER in detecting rare variants as opposed to other methods. In contrast, the major source of error in array-based pooled-DNA analysis is array variation, being 7 times higher than pool construction variation [157]. This observation argues that our approach shows higher accuracy even compared to other experimental platforms. Finally, our approach can be applied to any pooled cohort or any heterogeneous sample of any size and can be easily scaled up to whole-exome and whole-genome analysis. Given the presence of a positive control to infer the optimal parameters, pooled-samples can accurately be analysed without limitations on experimental design or achieved coverage. In this study, we used PCR to amplify the various genomic regions, but our strategy is also compatible with solid and liquid-phase genomic capture approaches [158]. We found that alignment errors decreased our ability to detect large indels. This explains why SPLINTER performed slightly worse in the analysis of the 4bp indel libraries relative to the 1-2 bp indel libraries. To detect the longer indels, it was necessary to allow larger gaps in our read alignments, which increased the overall alignment noise. We believe this was due to potential sequencing artifacts or
sample contaminants aligning back to the reference sequence, thereby reducing the signal coming from true variants. This limitation can be overcome with longer sequencing read lengths, which should reduce the ambiguity in aligning reads while allowing larger gaps (in this work, all sequencing reads were 36bp in length). Similarly, while whole genome analysis may present additional challenges due to increased sequence complexity compared to the analyzed synthetic controls, we expect it to mostly impact the read alignment step in the analysis pipeline, which can be overcome by generating paired-end and/or longer sequencing reads. In addition, with reduced error rate, fewer observations at a given variant position will be needed to provide confidence in the variant call. Nevertheless, our approach is the first one to accurately call short indels in large pooled-samples. One departure of our algorithm from other variant calling programs is that SPLINTER does not incorporate quality scores in any step of the analysis. We have found that our error model captures essentially the same information that is contained in quality scores (see Supplementary Materials and Methods, [121]), and so including quality score information does not improve SPLINTER’s performance. The high performance of our method compared to others that use quality scores [149, 152] suggests that this viewpoint is likely correct. Additionally, analysing reads aligned with quality scores resulted in equal or lower performance when compared to reads aligned using our aligner (Figure 2.19). To obtain a complete understanding of the molecular causes of common diseases, it is critical to be able to detect and analyze rare variants [121, 134, 149, 150, 151]. Pooled-DNA sequencing is an important method for rare variant analysis, since it enables the rapid and cost-effective analysis of thousand or tens of thou-
sands of individuals. SPLINTER will also be useful for analyzing samples that are naturally heterogeneous e.g. for the detection and quantification of rare somatic mutations in a tumor samples [159]. A second promising application is detection of induced mutations in in vitro evolution experiments [92, 160]. Thus, we expect SPLINTER will become useful tool for the analysis of data generated by next-generation sequencing methods.

2.3.5 Materials and methods

Preparation of the Synthetic Pools

Every synthetic pool library consists of a mixture of different oligonucleotides where one is referred to as the wild-type allele and the other are mutants with respect to the wild type. We used the consensus sequence of the 72 bp exon 9 from TP53 (RefSeq accession number: NM_000546) as the wild type insert into a pGEM-T Easy vector (Promega, Madison, WI, U.S.A.). We then designed a panel of different variations of this consensus sequence (see Supplementary Table 1,2,3) containing single, double and 4bp IN/DELs as well as single nucleotide substitutions. These vectors could then be pooled such that each mutation was present at different frequencies. Once pooled, a single PCR reaction was performed using primers that flanked the insertion site and generated a 335 bp amplicon. To facilitate ligation into the vector, each oligonucleotide was ordered with 5’ phosphorylation and an overhanging 3’ A from Integrated DNA Technologies (Coralville, IA, U.S.A.). Complimentary oligonucleotide pairs were annealed as follows: 1µl of sense and antisense oligonucleotide at 100µM were mixed with 5µl of 10X PCR Buffer (Sigma-Aldrich, Saint Louis, MO, U.S.A.) and
brought to a final volume of 50µl. The annealing mix was then warmed up to 95°C for 5 minutes followed by 20 minutes at 25°C. Each annealed sequence was then ligated into the pGEM-T Easy Vector (Promega, Madison, WI, U.S.A.) according to the manufacturer’s protocol and reagents. The final ligation product was then transformed into GC-10 competent cells (GeneChoice, Frederick, MA, U.S.A.) using standard cloning protocol. Colonies were screened using ‘Blue/White’ selection induced by Xgal and IPTG, White colonies were picked and grown on Luria broth agar with ampicillin for 12-16 hours. Plasmid was then recovered from the transformed bacteria suspension using Qiaprep Spin Miniprep Kit according to the manufacturer’s protocol (Qiagen, Hilden, Germany). Following insert validation by Sanger sequencing, plasmid pools were prepared by pooling each plasmid at the appropriate number of molecules in order to introduce the desired mutations at the desired frequency with respect to the wild-type background. Each pool was generated with a total number of $10^{11}$ plasmid molecules. This was chosen in order to have mimic at the best the conditions described in the original pooled-DNA sequencing protocol [121] to maximize the number of molecules available for analysis while keeping fluid volumes tractable. Each pool was then PCR amplified using primer sequences flanking the plasmid insertion site (see Supplementary Table 4). Each PCR reaction was performed as follows: 1) 93°C for 2 minutes 2) 93°C for 30 seconds 3) 56°C for 30 seconds 4) 65°C for 2 minutes 5) repeat steps 2-4 for 18 cycles 6) 65°C for 10 minutes. Each PCR mix contained 2.5µl of 10X PfuUltra Buffer, 10µM of forward and reverse primers, 1M Betaine (Sigma-Aldrich/Fluka, St. Louis, MO, U.S.A.), 1.25 units PfuUltra DNA polymerase and between 30 and 50 ng of template DNA in a final
volume of 25 µL. Each pool was then sequenced using a single lane of the Illumina Genome Analyzer II platform.

**Pooled-DNA sample sequencing from single individuals**

We performed pooled-DNA sequencing on 974 individuals enrolled in the Family Heart Study (FHS) [161] and 178 individuals enrolled in the Silent Cerebral Infarction Transfusion Trial (SIT) [162]. Each pool of human DNA was designed to contain 450 ng of DNA per individual. Pooling was performed robotically using the Eppendorf epMotion 5075 pipetting robot (Eppendorf, Hauppauge, NY, U.S.A) in order to minimize dilution errors. Patients from the FHS study were divided into 8 pools ranging from 94 to 150 individuals, whereas the remaining 178 patients were pooled into a single ninth pool. We computationally generated a list of 14 genomic loci selected on the basis that they contained at least one single or double base pair insertion or deletion reported in dbSNP129 at various frequency ranges. We defined a genomic region of 400 bp centered on the targeted IN/DEL and we designed primers in order to target the known variant in a final amplicon with length ranging from 150 to 200 bp. Primers were designed as described previously [121]. Each PCR reaction for each pool was repeated multiple times in order to minimize the likelihood of PCR errors appearing as rare variants in the sequencing output. Each PCR reaction was performed as described above with the exception of undergoing 28 total cycles. Each PCR mix contained 2.5 µl of 10X PfuUltra Buffer, 10 µM of forward and reverse primers, 1M Betaine (Sigma-Aldrich/Fluka, St. Louis, MO, U.S.A.), 1.25 units PfuUltra DNA polymerase and between 30 and 50 ng of
template DNA, representing 50 genome copies per individual (see Supplementary Table 4). For every analyzed pool a positive and a negative control were generated. The positive control consisted of a synthetic pool containing substitutions and IN/DELs at the lowest possible frequency for each of the analyzed pools (i.e. 1 divided by the total number of alleles in the pool). The positive control was prepared as described, with PCR reactions performed for 28 cycles in order to match the sample preparation. The negative control consisted of a DNA fragment for which the sequence is known in order to generate a run-specific second order error model to be used in the data analysis. In order to generate a negative control, we performed PCR amplification on the M13 plasmid (New England Biolabs, Ipswich, MA, U.S.A.) generating a 1934 bp product (see Supplementary Table 4). PCR reaction was performed as described above and repeated multiple times. We then sequenced our samples using a single lane per pool and mapped back all the sequencing reads.

Four candidate loci identified by a GWA study performed on the 974 FHS patients were targeted in a total of 36 PCR reactions spanning 20,729 base pairs per individual (data not shown). Sample preparation, sequencing and analysis was performed as described. We identified a total of genomic variants that were also represented by the Illumina 6.0 genotyping array performed on each individual. Pearson correlation coefficient between GWAS and SPLINTER variant frequencies was calculated by using the frequency of the minor allele for each variant (Figure 2.20).
Independent Validation of Putative Variants

Independent validation of putative variants identified by SPLINTER was performed by Sequenom (Sequenom, San Diego, CA, U.S.A.) according to the manufacturers protocol (for probes see Supplementary Table 5). Sanger sequencing validation was performed using the same primer pairs used for initial PCR amplification prior to pooled-DNA sequencing.

Semi-local gapped sequence alignment

In order to efficiently map sequencing reads with gaps without compromising mapping accuracy (i.e. deviating from the optimal mathematical solution of the alignment), we developed a new Smith-Waterman-like alignment strategy [163]. This allowed us to have a pure and controlled implementation of dynamic programming while being feasible in terms of speed. Quality scores are ignored for the alignment. We first build a hash-map of the reference sequence with hash key size equal to $k$, which is defined as

$$k = \left\lfloor \frac{l-c}{c+1} \right\rfloor$$

where $l$ is the length of the sequencing read and $c$ is the maximum edit distance cutoff between the read sequence $r$ and the reference sequence $s$. $k$ is guaranteed to be the largest possible stretch of perfect match nucleotides achieved in the case of maximum entropy, i.e. when the edits are distributed uniformly in the reads, minimizing the length of the shortest read fragment. While $l$ is run-dependent, $c$ is defined by the users at the time of the alignment, leading to a consequent user-defined value of $k$. 
When \( r \) is aligned to \( s \), the first step is to hash-map all the substrings of \( r \) of length \( k \) to \( s \). Every mapped substring allows to define the boundaries of a dynamic programming (DP) matrix for sequence alignment. The value of \( c \) determines the dimensions of the DP matrix, which are equal to \( l \) and \( l + 2 \times c \), assuming that the read will contain all the allowed edits.

Once the boundaries of the DP matrix are defined, we perform DP programming in the following way: first the matrix is initialized by setting the values of the first column to 0 (\( s \) dimension) as in Smith-Waterman [163] whereas the first row is set to 0 at position 0 and at progressively adding a gap penalty for every increasing position (\( r \) dimension) as in Needleman-Wunsch [164]. This strategy allows every alignment to start at any position in the reference sequence but always at the first position of the read, therefore being semi-local. Gaps are inserted according to an affine-gap penalty model [165], adopting a gap insertion penalty of 2 and gap extension penalty of 1.

Traceback is performed starting from the highest scoring position in the last column (corresponding to the last position in \( r \)) until the first position of \( r \) is reached. The final result is the optimal mathematical solution for the gapped semi-local alignment of \( r \) with respect to \( s \). \( r \) and its reverse complement are both mapped to the positive strand of \( s \).

If \( r \) aligns to multiple loci of \( s \) with the same minimum edit distance, its alignment to \( s \) is discarded in order to minimize noise due to spurious mapping. Insertions and deletions present inside a nucleotide homopolymer (i.e. AAA,CCC,GGG \ldots) are aligned at the beginning of the homopolymer on the positive strand by default, as their true position in the sequence is arbitrarily established.
We previously reported that error rates change significantly for every sequencing run \cite{121} and therefore, for every run we calculated an independent error model. Quality scores have been discarded as they have been previously shown to be less informative than an empirically derived run-specific error model \cite{121}. This finding is further supported by the lower performance of every other approach that we compared to SPLINTER that also integrated quality scores in the analysis (see Figures 2.15 and 2.17).

Since this approach does not take into account quality score information, in order to save computing time while preserving the same amount of information, we compressed the original SCARF file by keeping only unique reads sequence in it and adding a weight to each read counting the number of times a read with the same sequence appeared in the original file. This strategy generated files at the best 10% of the original size of the SCARF output, linearly reducing the alignment run-time of the same factor.

**Error model calculation**

A 2\textsuperscript{nd}-order error model was parameterized from a negative control sequence included in every sample, i.e. a DNA fragment consisting of a PCR product from the M13 vector. The negative control allows to estimate the likelihood of a sequencing error defined as the rate of observed mutations in the sequencing reads without variants being present in the analyzed DNA fragment. Briefly, for every base \( n \) and its context defined as the two preceding bases \( m \) and \( l \), we calculate the likelihood of observing a substitution \( s \), an insertion \( i \) or a deletion \( d \) where \( l, m, n, s \in \{A, C, G, T, N\}, i \in \{\text{Insertion}_{A,C,G,T}\}, \text{and} \ d \in \{D\} \). For substitutions,
we calculate \( Pr(s|n, m, l, j, r) \) for each read base \( j \) and run \( r \) as the ratio between the number of observed read bases with base equal to \( s \) and the total number of observed read bases. Deletion error rates are calculated the same way as substitutions, where \( j \) in this case is assumed to be the read base number preceding the deletion. Insertions are analyzed by selecting only reads that overlap consecutive loci \( n \) and \( o \). \( Pr(i|o \sim n, m, l, j, r) \) is therefore computed as the ratio between the reads that contain one or more inserted bases between \( n \) and \( o \) and the total number of reads overlapping \( n \) and \( o \).

**SPLINTER: IN/DEL and substitution detection using Large Deviation Theory**

Since previously designed algorithms are unable to precisely call and quantify short IN/DELs in large pools, we designed and implemented SPLINTER (Short IN/DEL Prediction by Large deviation Inference and Non-linear True frequency Estimation by Recursion), a new algorithm based on Sanov’s theorem, which is part of the information theoretic treatment of Large Deviation Theory [140]. SPLINTER takes in input aligned sequencing reads. For every position \( i \) of the reference sequence, SPLINTER stores the counts of each observed base character \( b = \{A, C, G, T, N, D\} \) as well as the counts for each inserted base stretch \( g \) between \( i \) and its consecutive position \( i+1 \) of length \( c \) (maximum number of accepted edits). SPLINTER assumes that sequencing reads are generated independently from one another and that read bases within the same read are incorporated independently from one another.

Substitution variants can be detected at a particular position \( i \) by estimating the distance of the empirical distribution \( P \) of observed nucleotides A,C,G,T,N from the expected distri-
tion \( Q \) representing the expected frequency of nucleotides assuming that \( i \) does not harbor any variant present in the pool. \( Q \) is computed as a linear product between the error model matrix \( A \) for each read base \( j \) and sequencing run \( r \) and the true frequency vector \( \tau_{i,\text{null}} \) under the null hypothesis that only the reference base is present as

\[
Q_{j,r,s} = A_{i,j,r} \ast \tau_{i,\text{null}}
\]  

(2.8)

The distance between \( P \) and \( Q \) is computed independently for each read base \( j \), sequencing run \( r \) and strand \( s \) as equation 2.2:

\[
Q_{j,r,s}^{n_j,r,s}(E) = 2^{-n_{j,r,s}D(P_{j,r,s} || Q_{j,r,s})}
\]

where \( D(P_{j,d,s} || Q_{j,d,s}) \) is the Kullback–Leibler distance between \( P \) and \( Q \). \( Q_{j,r,s}^{n_j,r,s}(E) \) is a p-value calculated by testing the hypothesis that \( P \) was sampled from \( Q \). Since \( j \) and \( r \) are independent, according to the initial assumptions, a cumulative p-value for each strand \( s \) is computed as

\[
Q_{s}^{n_s}(E) = 2^{-\sum_{r} \sum_{j} n_{j,r,s}D(P_{j,r,s} || Q_{j,r,s})}
\]

(2.9)

equivalently to equation 2.3. Deletions are detected by estimating the distance of the fraction of observed deletions \( P_D \) from the fraction of expected deletions \( Q_D \). \( P_D \) and \( Q_D \) are defined for each read base \( j \), run \( r \) and strand \( s \) as

\[
P_{D_{j,r,s}} = \left( \frac{d_{j,r,s}}{C_{j,r,s}}, 1 - \frac{d_{j,r,s}}{C_{j,r,s}} \right)
\]

(2.10)

\[
Q_{D_{j,r,s}} = \left( A_{D_{i,j,r,s}}, 1 - A_{D_{i,j,r,s}} \right)
\]

(2.11)
where \( d_{j,r,s} \) and \( C_{j,r,s} \) represent the number of observed deletions and the total observed coverage and \( A_{D_{j,r,s}} \) corresponds to the expected likelihood of observing a deletion at \( i \) for \( j, r \) and \( s \) given the error model matrix \( A \).

The distance between \( P_{D_{j,r,s}} \) and \( Q_{D_{j,r,s}} \) is again computed independently as

\[
Q_{D_{j,r,s}}^{n_{D_{j,r,s}}}(E) = 2^{-n_{D_{j,r,s}}} D(P_{D_{j,r,s}} || Q_{D_{j,r,s}})
\]  

(2.12)

and the cumulative p-value for \( s \) is computed as

\[
Q_{D_{s}}^{n_{D_{s}}}(E) = 2^{-\sum_r \sum_j n_{D_{j,r,s}}} D(P_{D_{j,r,s}} || Q_{D_{j,r,s}})
\]  

(2.13)

Insertions are analyzed as stretches of nucleotides (\( g \)) of maximum length \( c \) located between the adjacent and consecutive positions \( i \) and \( i+1 \) and are detected by comparing the observed insertion distribution \( P_{I_{i\sim i+1}} \) and the expected insertion distribution \( Q_{I_{i\sim i+1}} \), defined as

\[
P_{I_{i\sim i+1,j,r,s}} = \left( \frac{g_{i\sim i+1,j,r,s}}{C_{i\sim i+1,j,r,s}}, 1 - \frac{g_{i\sim i+1,j,r,s}}{C_{i\sim i+1,j,r,s}} \right)
\]  

(2.14)

\[
Q_{I_{i\sim i+1,j,r,s}} = \left( A_{I_{i\sim i+1,j,r,s}}, 1 - A_{I_{i\sim i+1,j,r,s}} \right)
\]  

(2.15)

where \( A_{I_{i\sim i+1,j,r,s}} \) corresponds to the expected likelihood of observing an insertion at \( i \sim i+1 \) for \( j, r \) and \( s \) given the error model matrix \( A \). Calculation of the p-value for each strand \( s \) is performed as described for deletions.

Presence or absence of any given variant at position \( i \) is assessed by asking the p-values for both strands to be equal or less than a user-defined empirical cutoff \( \alpha \) (where \( \alpha <= 0.05 \)).
We find that requiring both strands to pass $\alpha$ greatly increases the accuracy of our method, as previously observed [121].

**Frequency estimation of identified pool variants**

For every identified pool variant, SPLINTER performs estimation of the true variant frequency vector $\tau_i$ at position $i$ and/or the true insertion frequency vector $\tau_{i,i+1}$ between $i$ and $i + 1$. $\tau$ is fit by maximum likelihood

$$\tau = \arg \max_{\tau} 2^{-\sum_s \sum_r \sum_j n_{j,r,s} D(P_{j,r,s} \parallel Q_{j,r,s,\tau})} \ast P_T(\tau)$$  \hspace{1cm} (2.16)

where we implicitly assume that $P_T(\tau)$, the prior distribution for $\tau$, is uniform, leading to

$$\tau \approx \arg \min_{\tau} \sum_s \sum_r \sum_j n_{j,r,s} D(P_{j,r,s} \parallel Q_{j,r,s,\tau})$$  \hspace{1cm} (2.17)

SPLINTER is significantly different from our previous pooled DNA SNP caller algorithm, SNPseeker [121] at various levels. First, it is able to detect indels in large pools by using new models and new integrated data structures, whereas SNPseeker can only detect substitutions. Secondly, SPLINTER is more sensitive and specific than SNPseeker as it integrates information of a positive control to define the optimal cutoff for the values of $Q_{n_s}^{\text{ns}}(E)$, $Q_{D_s}^{\text{nDs}}(E)$, $Q_{I_s}^{\text{nIs}}(E)$ (p-value cutoff) at every position $i$. Thirdly, SNPseeker implemented a non-linear least-square fit for estimating the true frequency vector $\tau$ [121], whereas SPLINTER uses a maximum likelihood method. We found that this leads to more accurate frequency estimates (data not shown) but it also allows incorporation of prior information.
Evaluation of sensitivity and specificity of variant calling and accuracy of frequency estimation

In order to determine the discriminatory power of our method, we calculated sensitivity and specificity in a p-value cutoff-independent way by iterating over a range of p-value cutoff values from 0 to -3000 with increments of -0.001 at each round. The optimal cutoff was determined as the value that minimized the Euclidean distance between the corresponding specificity and sensitivity (ranging from 0 to 1) to perfect specificity and sensitivity (1,1). This strategy was repeated by analyzing the data incorporating 12, 18, 21, 24 bases per read (cycles) and comparing sensitivity and specificity of the analysis, resulting in the definition of the optimal cutoff and incorporated read bases. This was done because we have previously demonstrated that the likelihood of sequencing errors increases for later cycles [121], and different error rates will affect the accuracy of discrimination between signal and noise. The optimal combination between cycles and cutoff was then used for data analysis. Accuracy of the frequency estimation was measured by calculating Pearson’s correlation coefficient between the observed and estimated frequencies.

Monte Carlo sampling and calculation of Receiver Operating Characteristics Curves

In order to determine the relationship between p-value and coverage per base per strand for any given variant, we performed Monte Carlo sampling on aligned reads for a selected synthetic pool. We randomly sampled fractions equivalent to 0.005, 0.010, 0.015, 0.020,
0.025, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.75, 0.90 of the total number of aligned reads 100 times each and performed an analysis with SPLINTER on every sample. This allowed us to generate a distribution of p-values for each coverage point. For each set of 100 samples we calculated a Receiver Operating Characteristics (R.O.C.) curve. ROC curves plot a methods sensitivity (here, the fraction of mutant positions correctly identified) versus the false positive rate (the fraction of the bases without mutation that were incorrectly reported) for different p-value cutoffs. For each ROC curve we computed the corresponding Area Under the Curve (A.U.C.), and we used it as a metric for assessing the lowest value of coverage per base at which 100% specificity and 100% sensitivity (AUC equal to 1) are reached.

**Comparison between different variant callers**

In order to compare the performance of SPLINTER with that of other approaches, we applied SNPseeker [121], MAQ v0.7.1 [152], SAMtools [156] and VarScan [149] to the synthetic pool datasets. We separately compared the performance of each approach for detection of substitutions and indels given the fact that indels are not supported by SNPseeker and MAQ. Performance was evaluated by determining sensitivity (fraction of true positives identified by the method over total true positives in the set) and positive predicted value (fraction of true positives identified by the method over total positions identified by the method) and values were plotted and compared for each approach. For substitutions, Pools 4 and 5 were used in their entirety (renamed sub 1 and sub 2 in Figure 2.15) whereas only substitutions were considered for the pools simulating 100, 250 and 500 samples. For indels, Pools 1,
2 and 3 were used in their entirety (renamed indel 1, indel 2 and indel 3) whereas only indels were considered for the pools simulating 100, 250, and 500 samples. SNPseeker was run as previously described and performance was computed after determining the optimal p-value cutoff and that maximized its performance. MAQ was run as described in [152] with snp filtering after its execution in order to reduce the number of false positives. For SAMtools and VarScan, files were previously aligned using Novoalign at its default settings www.novocraft.com, and SAM files were then converted into BAM and then pileup files. For SAMtools, variants were called from the pileup file, variants are unfiltered because when filtering was applied no hits were returned in output for any of the tested libraries. VarScan was run on the SAMtools pileup files and results were filtered by finding the optimal p-value cutoff that maximized its performance. We compared also the performance of CRISP [19] by running the approach applied to all the pools using the default settings. We additionally compared the performance of the Genome Analysis Toolkit (GATK) framework on our set [166] using the suggested default parameters but we could not detect any of the true positives in any of the synthetic sets, so we decided not to include this analysis in the comparison. We believe that this result was due to the Unified Genotyper being optimized for single individual genotyping rather than large pools sequencing. Additionally, GATK is not able to detect indels in its current iteration (v1.0.3864).
2.3.6 Acknowledgments

This work was partly supported by the Children’s Discovery Institute grant MC-II-2006-1 (R.D.M., T.E.D.), the NIH Epigenetics Roadmap grant (1R01DA025744-01 and 3R01DA025744-02S1; R.D.M., F.L.M.V.), the Saigh Foundation (F.L.M.V., T.E.D.), and Hope Street Kids and Alex’s Lemonade Stand A Award support (T.E.D.). We thank Michael DeBaun at Washington University in Saint Louis, School of Medicine, for providing the samples from the SIT cohort; Lee Tessler, David Mayhew, and the other members of the Mitra lab for helpful discussion on the SPLINTER algorithm; the Sequenom Core facility at the Human Genetics Division at Washington University in Saint Louis, School of Medicine for the Sequenom validation; and Jessica Hoisington-Lopez at the Center for Genome Science, Washington University in Saint Louis, School of Medicine for running the samples on the Illumina GAIIx platform. This work is dedicated to Natalina Vallania.
### 2.3.7 Supplementary Tables

#### Figure 2.22. Supplementary Table 1

<table>
<thead>
<tr>
<th>Synthetic Strain</th>
<th>Mutation 1</th>
<th>Mutation 2</th>
<th>Engineered Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>356</td>
<td>300</td>
<td>230</td>
<td>226</td>
</tr>
<tr>
<td>188</td>
<td>248</td>
<td>258</td>
<td>260</td>
</tr>
<tr>
<td>152</td>
<td>180</td>
<td>210</td>
<td>240</td>
</tr>
<tr>
<td>134</td>
<td>137</td>
<td>149</td>
<td>158</td>
</tr>
<tr>
<td>152</td>
<td>180</td>
<td>210</td>
<td>240</td>
</tr>
<tr>
<td>134</td>
<td>137</td>
<td>149</td>
<td>158</td>
</tr>
<tr>
<td>152</td>
<td>180</td>
<td>210</td>
<td>240</td>
</tr>
<tr>
<td>134</td>
<td>137</td>
<td>149</td>
<td>158</td>
</tr>
<tr>
<td>152</td>
<td>180</td>
<td>210</td>
<td>240</td>
</tr>
<tr>
<td>134</td>
<td>137</td>
<td>149</td>
<td>158</td>
</tr>
</tbody>
</table>

#### Figure 2.23. Supplementary Table 2

<table>
<thead>
<tr>
<th>Synthetic Pool</th>
<th>IN/DEL 1</th>
<th>IN/DEL 2</th>
<th>IN/DEL 3</th>
<th>IN/DEL 4</th>
<th>IN/DEL 5</th>
<th>Substitution 1</th>
<th>Substitution 2</th>
<th>Substitution 3</th>
<th>Substitution 4</th>
<th>Substitution 5</th>
<th>Wild-Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic Pool 1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.968</td>
</tr>
<tr>
<td>Synthetic Pool 2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.968</td>
</tr>
<tr>
<td>Synthetic Pool 3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.968</td>
</tr>
<tr>
<td>Synthetic Pool 4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.968</td>
</tr>
<tr>
<td>Synthetic Pool 5</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.968</td>
</tr>
</tbody>
</table>

#### Figure 2.24. Supplementary Table 3

<table>
<thead>
<tr>
<th>Synthetic Pool</th>
<th>4bp IN/DEL 1</th>
<th>4bp IN/DEL 2</th>
<th>4bp IN/DEL 3</th>
<th>4bp IN/DEL 4</th>
<th>4bp IN/DEL 5</th>
<th>IN/DEL 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>4bp Pool 1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.005</td>
<td>0.010</td>
<td>0.015</td>
<td>0.001</td>
</tr>
<tr>
<td>4bp Pool 2</td>
<td>0.025</td>
<td>0.030</td>
<td>0.035</td>
<td>0.040</td>
<td>0.045</td>
<td>0.000</td>
</tr>
</tbody>
</table>

#### Figure 2.25. Supplementary Table 4

<table>
<thead>
<tr>
<th>Synthetic Pool</th>
<th>Engineered Variants</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Incorporate Read Bases (Cycles)</th>
<th>Optimal Cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1e-29.157</td>
</tr>
</tbody>
</table>

#### Figure 2.26. Supplementary Table 5

<table>
<thead>
<tr>
<th>Pool</th>
<th>Alleles</th>
<th>Engineered Variants</th>
<th>Incorporated Read Bases (Cycles)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>log10p cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pool 1</td>
<td>260</td>
<td>250</td>
<td>2C</td>
<td>1.000</td>
<td>1.000</td>
<td>-19.273</td>
</tr>
<tr>
<td>Pool 2</td>
<td>258</td>
<td>15</td>
<td>21C</td>
<td>1.000</td>
<td>0.9996</td>
<td>-29.767</td>
</tr>
<tr>
<td>Pool 3</td>
<td>248</td>
<td>15</td>
<td>21C</td>
<td>1.000</td>
<td>1.000</td>
<td>-25.115</td>
</tr>
<tr>
<td>Pool 4</td>
<td>188</td>
<td>15</td>
<td>24C</td>
<td>1.000</td>
<td>1.000</td>
<td>-36.583</td>
</tr>
<tr>
<td>Pool 5</td>
<td>246</td>
<td>15</td>
<td>24C</td>
<td>1.000</td>
<td>1.000</td>
<td>-49.156</td>
</tr>
<tr>
<td>Pool 6</td>
<td>230</td>
<td>15</td>
<td>24C</td>
<td>1.000</td>
<td>1.000</td>
<td>-29.887</td>
</tr>
<tr>
<td>Pool 7</td>
<td>238</td>
<td>15</td>
<td>24C</td>
<td>1.000</td>
<td>0.9996</td>
<td>-15.226</td>
</tr>
<tr>
<td>Pool 8</td>
<td>300</td>
<td>15</td>
<td>24C</td>
<td>1.000</td>
<td>1.000</td>
<td>-44.108</td>
</tr>
<tr>
<td>Pool 9</td>
<td>356</td>
<td>15</td>
<td>24C</td>
<td>1.000</td>
<td>0.9991</td>
<td>-17.781</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>-------------</td>
<td>------------------------</td>
<td>----------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Pool 1</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 2</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 3</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 4</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 5</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 6</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 7</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 8</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 9</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 10</td>
<td>INDEL 6</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 11</td>
<td>INDEL 7</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
<tr>
<td>Pool 12</td>
<td>INDEL 11</td>
<td>X</td>
<td>12437306</td>
<td>12437556</td>
<td>172</td>
<td>12347548</td>
</tr>
</tbody>
</table>

**Figure 2.27. Supplementary Table 6 part1**
| Pool | INDEL | X | Chr | Start | End | Size | Position | Sequence | Distance | Identity | Evalue | Score | No | Reference |
|------|-------|---|-----|-------|-----|------|----------|----------|----------|---------|--------|-----|-----------|
| 4    | INDEL | 14 | 18961224 | 18961410 | 190 | 18961303 | 18961303 | 97 | -760,041271 | -92,007839 | 0,990 | 0,010 | 64447 | rs3218085 |
| 5    | INDEL | 8 | 18124762 | 18124959 | 197 | 18124852 | 18124852 | 94 | -7517,203596 | -730,570315 | 0,755 | 0,245 | 3107  | rs8190861 |
| 5    | INDEL | 14 | 18961224 | 18961410 | 190 | 18961303 | 18961303 | 97 | -347,942,2873 | -1537,939998 | 0,991 | 0,009 | 4307  | rs3213397 |
| 7    | INDEL | 1 | 19881224 | 19881410 | 187 | 19881282 | 19881282 | 59 | -589,775478 | -434,837051 | 0,950 | 0,050 | 8305  | rs3093873 |
| 7    | INDEL | 4 | 18961224 | 18961410 | 190 | 18961303 | 18961303 | 97 | -178,443,2873 | -1537,939998 | 0,991 | 0,009 | 4307  | rs3213397 |
| 7    | INDEL | 11 | 18961224 | 18961410 | 190 | 18961303 | 18961303 | 97 | -178,443,2873 | -1537,939998 | 0,991 | 0,009 | 4307  | rs3213397 |
### Table 6 part 3

| INDEL 1 | Pool 9 | 16964636 | 90432612 | 186 | 90432758 | 81 | D | -117.060013 | -53.938475 | 0.92 | 0.049 | 9745 | 4971 | x590110 | x590110 |
| INDEL 12 | 14 | 9965736 | 90432612 | 186 | 90432758 | 82 | D | -117.154305 | -54.315186 | 0.92 | 0.046 | 9967 | 4856 | x590110 | x590110 |
| INDEL 19 | 16 | 2208156 | 2208156 | 171 | 2208156 | 76 | D | -123.737584 | -59.352464 | 0.93 | 0.010 | 7261 | 7501 | x590110 | x590110 |
| INDEL 8 | 10 | 79883850 | 74883851 | 150 | 79883932 | 117 | D | -245.151791 | -28.255203 | 0.93 | 0.026 | 9464 | 10161 | x590110 | x590110 |
| INDEL 4 | 19 | 46735707 | 46735706 | 181 | 46735766 | 96 | D | -172.583893 | -52.317592 | 0.93 | 0.010 | 8233 | 8149 | x590110 | x590110 |
| INDEL 4 | 3 | 46735707 | 46735706 | 181 | 46735766 | 97 | D | -205.858767 | -72.675544 | 0.94 | 0.002 | 8408 | 8320 | x590110 | x590110 |
| INDEL 4 | 1 | 2208156 | 2208156 | 171 | 2208156 | 76 | D | -126.810213 | -59.577470 | 0.94 | 0.002 | 7914 | 10034 | x590110 | x590110 |
| INDEL 4 | 4 | 46735707 | 46735706 | 181 | 46735766 | 97 | D | -159.594444 | -55.281919 | 0.95 | 0.009 | 2077 | 1917 | x590110 | x590110 |
| INDEL 5 | 6 | 18127660 | 18127659 | 197 | 18127659 | 162 | D | -307.965521 | -31.025490 | 0.94 | 0.016 | 1526 | 109 | x590110 | x590110 |
| INDEL 5 | 6 | 18127660 | 18127659 | 197 | 18127659 | 162 | D | -449.227356 | -281.386336 | 0.74 | 0.322 | 1601 | 117 | x590110 | x590110 |
| INDEL 16 | 12 | 51793337 | 51793334 | 160 | 51793419 | 75 | D | -132.664222 | -79.304191 | 0.96 | 0.004 | 13879 | 10276 | x590110 | x590110 |
| INDEL 8 | 14 | 51793337 | 51793334 | 160 | 51793419 | 75 | D | -135.430748 | -63.737304 | 0.97 | 0.004 | 1542 | 684 | x590110 | x590110 |
| INDEL 11 | 6 | 42016310 | 42016308 | 174 | 42016325 | 100 | D | -694.013907 | -343.977792 | 0.96 | 0.016 | 12961 | 4932 | x590110 | x590110 |
| INDEL 6 | 10 | 42016305 | 42016308 | 174 | 42016325 | 101 | D | -386.769698 | -165.559841 | 0.94 | 0.016 | 13444 | 5051 | x590110 | x590110 |

**Figure 2.29. Supplementary Table 6**

**Figure 2.30. Supplementary Table 7**

<table>
<thead>
<tr>
<th>Pool</th>
<th>dbSNP129 Variants</th>
<th>Total Variants</th>
<th>Fraction</th>
<th>Fisher's Exact Test P-value</th>
<th>Bonferroni Corrected P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pool1</td>
<td>13</td>
<td>17</td>
<td>78.87%</td>
<td>4.91E-22</td>
<td>4.42E-21</td>
</tr>
<tr>
<td>Pool2</td>
<td>16</td>
<td>22</td>
<td>72.22%</td>
<td>1.38E-26</td>
<td>1.24E-25</td>
</tr>
<tr>
<td>Pool3</td>
<td>15</td>
<td>22</td>
<td>68.18%</td>
<td>3.35E-24</td>
<td>3.02E-23</td>
</tr>
<tr>
<td>Pool4</td>
<td>14</td>
<td>19</td>
<td>73.68%</td>
<td>2.40E-29</td>
<td>2.16E-28</td>
</tr>
<tr>
<td>Pool5</td>
<td>16</td>
<td>19</td>
<td>84.21%</td>
<td>1.84E-28</td>
<td>1.65E-27</td>
</tr>
<tr>
<td>Pool6</td>
<td>17</td>
<td>20</td>
<td>80.00%</td>
<td>9.10E-28</td>
<td>8.19E-27</td>
</tr>
<tr>
<td>Pool7</td>
<td>11</td>
<td>20</td>
<td>100.00%</td>
<td>1.98E-29</td>
<td>1.80E-28</td>
</tr>
<tr>
<td>Pool8</td>
<td>22</td>
<td>32</td>
<td>68.75%</td>
<td>2.92E-36</td>
<td>2.62E-35</td>
</tr>
</tbody>
</table>
### Figure 2.31. Supplementary Table 8

<table>
<thead>
<tr>
<th>Variant</th>
<th>Presence</th>
<th>Method</th>
<th>Measured Frequency</th>
<th>SPLINTER Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3050130</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,002</td>
<td>0,002</td>
</tr>
<tr>
<td>rs3918110</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,003</td>
<td>0,002</td>
</tr>
<tr>
<td>rs2070888</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,006</td>
<td>0,005</td>
</tr>
<tr>
<td>rs17884136</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,006</td>
<td>0,005</td>
</tr>
<tr>
<td>rs3093876</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,023</td>
<td>0,022</td>
</tr>
<tr>
<td>rs11573206</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,043</td>
<td>0,026</td>
</tr>
<tr>
<td>rs9658624</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,060</td>
<td>0,056</td>
</tr>
<tr>
<td>rs17885121</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,074</td>
<td>0,093</td>
</tr>
<tr>
<td>rs8177888</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,090</td>
<td>0,116</td>
</tr>
<tr>
<td>rs10305888</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,041</td>
<td>0,116</td>
</tr>
<tr>
<td>rs15561</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,359</td>
<td>0,502</td>
</tr>
<tr>
<td>INDEL18 32</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,020</td>
<td>0,034</td>
</tr>
<tr>
<td>INDEL4 96</td>
<td>Y</td>
<td>Sequenom</td>
<td>0,048</td>
<td>0,039</td>
</tr>
<tr>
<td>INDEL1 59</td>
<td>Y</td>
<td>Sanger</td>
<td>/</td>
<td>0,413</td>
</tr>
</tbody>
</table>

### Figure 2.32. Supplementary Table 9

<table>
<thead>
<tr>
<th>Amplicon</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
<th>Product size</th>
<th>Annealing Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>M13/ Negative Control</td>
<td>TCCGAAATAGGCAGGGGGCATTAA</td>
<td>GCTGATGCAAATCCTGCAACAGA</td>
<td>1534</td>
<td>56</td>
</tr>
<tr>
<td>Synthetic library/Positive Control</td>
<td>GCCACGTGTTTCTCAGGAGGCG</td>
<td>CAGGAGATGCCTACGCACTCG</td>
<td>335</td>
<td>56</td>
</tr>
<tr>
<td>INDEL1</td>
<td>GCCCTGGGCGCCAGCTTGCG</td>
<td>ATAGGGCGGAGGGGCTGCTCAGT</td>
<td>187</td>
<td>59</td>
</tr>
<tr>
<td>INDEL2</td>
<td>AAAGAGAGGGAGGAGGCAATCCG</td>
<td>GGGGGCGAATAGTCAGTCTG</td>
<td>153</td>
<td>59</td>
</tr>
<tr>
<td>INDEL4</td>
<td>GGGGCTGCTGCTCTTCTTTTT</td>
<td>GGCTGATGCAAATGCGAGGAG</td>
<td>151</td>
<td>59</td>
</tr>
<tr>
<td>INDEL5</td>
<td>TGCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL6</td>
<td>TGCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL8</td>
<td>AGCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL10</td>
<td>ACCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL11</td>
<td>GCCGACGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL12</td>
<td>AGCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL13</td>
<td>AGCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL16</td>
<td>AGCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
<tr>
<td>INDEL18</td>
<td>AGCTGCTGCTGCTGCTGCTGCTG</td>
<td>TCCGGAGCTGCTGCTGCTGCTG</td>
<td>157</td>
<td>59</td>
</tr>
</tbody>
</table>
3. High-throughput dissection of genetic factors underlying DNA methylation.

3.1 Foreword

This chapter describes the development of a reporter system to dissect the genetic \textit{cis} architecture underlying DNA methylation patterns across the genome, with a specific focus on neuronal differentiation. This project involves experimental work as well as computational modeling aimed at inferring such elements from large scale DNA methylation datasets. This project was conceived by me and supervised by Robi D Mitra, PhD. I designed, prototyped, and tested the reporter system as well as the probabilistic models for the analysis of differentially methylated regions. Rob and I discussed and designed the experiments. Sumithra Sankararan, PhD and Zongtai Qi performed transposon-based experiments critical for the correct implementation of the experimental assay. This work is in collaboration with Barak A Cohen, PhD, who provided guidance and experimental reagents. This work is currently in progress and has not yet been published.
3.2 Introduction/Results

3.2.1 Experimental design and construction of a reporter system for DNA methylation

Understanding how DNA methylation patterns across the genome are established remains a major question in the fields of epigenetics and chromatin biology. A significant factor hindering our efforts toward answering that question is the lack of a high-throughput system to dissect the DNA sequence contribution to DNA methylation. The purpose of such system is to provide a DNA backbone where a DNA fragment of interest could be inserted in isolation from its endogenous locus of origin. Doing so would allow us to determine whether such fragment contains all the sufficient sequence elements to determine its own methylation pattern. Here we present the design and construction of a methylation reporter system based on the piggyBAC DNA transposon \[167\] and its application to the dissection of the establishment of DNA methylation across the regulatory regions of the gene Oct4 \[62\].

Design and testing of a high-throughput reporter system for DNA methylation

We first started by designing and testing our reporter system for DNA methylation. The system that we adopted for our experiments is a derivation of piggyBAC, a type II DNA transposon \[167\]. PiggyBAC is a two-component system consisting of a "donor" vector, i.e. the transposon, characterized by inverted repeats sequences, and a "helper" vector, encoding for the "transposase" (Figure 3.1). Once co-transfected inside a particular cell
type of interest, the piggyBAC transposase will integrate multiple copies of the donor vector inside the host genome. Integration events will occur at the presence of a TTAA recognition site in the host genome, but will otherwise be specified randomly. By inserting a DNA fragment of interest inside the piggyBAC donor, it is therefore possible to integrate multiple copies across the genome. Finally, after transient transfection, individual clones containing one or multiple insertions can be selected by drug screening using an integrated resistance cassette present in the inserted donor constructs. This setup can allow us to read out the methylation of the inserted fragment by averaging out the methylation across all the insertion events that occurred across the genome of all the transfected cells. This feature provides several advantages. Firstly, by integrating our signal across many insertion events in different genomic locations, site-specific position effects \[168\] can be averaged out, resulting in a clear picture of the endogenous methylation pattern specified by the inserted DNA fragment. Secondly, integration of the DNA fragment will ensure its persistence during the process of differentiation without loss of signal due to dilution of the reporter during cell replication.

In order to determine the efficiency of the reporter, we started by inserting a DNA fragment that had been previously identified as being sufficient to establish its own methylation pattern outside of its endogenous locus. The control region that we chose consists of the proximal promoter of the \(Trf\) gene, which is unmethylated in ES cells and acquires methylation during differentiation into neural progenitors \[70\]. Using the RW4 mouse embryonic stem cell line as a model system for neural differentiation (see Materials and methods), we proceeded with the experiment as described in Figure 3.2. Briefly, the \(Trf\) promoter was first
This figure is adapted from [76] and describes the piggyBAC transposon system.

cloned into the donor plasmid (piggyBAC~Trf) and then co-transfected with the transposase inside RW4 cell lines. After transfection, cells containing copies of piggyBAC~Trf will be selected using a puromycin [169] resistance cassette inside the donor. After a one-week selection, the ES cells (ESCs) will be differentiated into Neural Progenitor cells (NPCs). DNA collected before and after differentiation is then harvested, processed and then sequenced in order to read out methylation of the Trf fragment inside the piggyBAC donor (for details see Materials and methods).
This figure depicts a schematic representation of the piggyBAC methylation reporter assay: the DNA fragment of interest is inserted inside the piggyBAC transposon plasmid, which is subsequently co-transfected in the mouse RW4 embryonic stem cell line. After selection, the cells containing piggyBAC insertions in their genome will be differentiated into neural progenitor cells. DNA is collected before and after differentiation and processed for methylation readout.

The piggyBAC methylation reporter faithfully recapitulates the endogenous patterns of DNA methylation

After performing the experiment and sequencing all samples, we then compared the level of methylation of piggyBAC~Trf between ESC and NPC across 2 biological replicates. As
expected, we observed almost no methylation in ESC but, on the other hand, piggyBAC~Trf displayed intermediate levels of methylation in NPCs (Figure 3.3), in concordance to the endogenous pattern of Trf. This result suggests that the piggyBAC reporter system can recapitulate the endogenous patterns of DNA methylation. Interestingly, individual CpGs displayed a unique and distinct pattern of DNA methylation, which appeared strikingly similar between biological replicates, suggesting that specific methylation of individual CpGs is a highly controlled process.

Figure 3.3. piggyBAC recapitulates the endogenous methylation pattern of the Trf promoter

This figure shows the levels of methylation of the Trf promoter in ES cells and NP cells. Each column represents a cell type and each row is an individual biological replicate. In each barplot, the x-axis represents the position in the Trf sequence, with each blue bar representing the position of every single CpG. The red bar overlaying the blue bar is the fraction that CpG that appears methylated following the methylation scale present on the left y-axis. The green line represents sequencing coverage in log_{10} scale, following the scale present on the right y-axis.
We then asked whether the patterns of single CpG methylation in piggyBAC~Trf would also recapitulate those of the endogenous copy of Trf. A comparison between the level of single CpG methylation of endogenous Trf to the inserted copy in NP cells revealed a strong correlation ($r = 0.885$, p-value = $1.137 \times 10^{-5}$, Figure 3.4). This result indicates not only that the piggyBAC reporter system recapitulates methylation patterns down to a single CpG level, but also that methylation is a highly controlled process down to a single CpG resolution. A surprise result from our analysis was that the level of methylation from our insert was dramatically lower than the one from the endogenous locus (compare 22% vs 70% in Figure 3.4). While this may produce potential shortcomings in the fine dissection of smaller changes in DNA methylation, the data suggests that the system is still quantitatively accurate on a relative scale even down to a single CpG level.

Since we compared the level of methylation of the same construct with different cell types, a potential explanation underlying our result could derive from undirected global methylation of the piggyBAC reporter in one cell type versus the other rather than sequence specific effects. To rule out this scenario and assess whether the Trf insert specifically acquired methylation independently of the context, we profiled two different regions of the piggyBAC transposon (named CR1 and CR2). As these regions belong to the backbone of the piggyBAC transpose, we would expect them to display a methylation pattern different from the Trf. In contrast, if methylation levels were largely specified by the location of insertion in the host genome and cell type, we would expect to observe the pattern of DNA methylation displayed by Trf across the whole transposon. As shown in Figure 3.5, unlike
Figure 3.4. piggyBAC methylation reporter recapitulates endogenous patterns at a single CpG level

This figure compares levels of methylation at single CpG level between the Trf insert (x-axis) and the endogenous copy (y-axis).

Trf, CR1 and CR2 display monotonic patterns of DNA methylation, suggesting that the pattern observed in Trf is largely dependent on its sequence.

To assess the robustness of the piggyBAC reporter system in recapitulating endogenous patterns of methylation, we generated 4 additional constructs with inserts deriving from the promoters of OrmI, Mrap, Zic3, and Syt1 [70]. Unlike Trf, these promoters displayed...
Figure 3.5. DNA methylation profile across different regions of the piggyBAC~Trf construct

This figure shows level of methylation of the Trf construct and 2 control regions of the piggyBAC reporter (CR1 and CR2) respectively (structure of the reporter system is represented in the bottom panel of the figure). Each barplot is interpreted analogously to the ones depicted in Figure 3.3. Arbitrary methylation calls are shown underneath each panel.

monotonic patterns of DNA methylation across ES cells and NP cells (the first two being methylated and the latter two being unmethylated respectively). We then asked whether each construct displayed the same pattern of methylation of the endogenous locus, i.e. increase in methylation or stationary methylation. To compare the changes in methylation between different constructs, we computed the relative gain of methylation, defined as
\[ F = \frac{\text{meth}T_2 - \text{meth}T_1}{\min(\text{meth}T_1, \text{meth}T_2)} \]  

(3.1)

where \( T_1 \) and \( T_2 \) are indicative of ES cells and NP cells respectively. When we compared all the constructs to one another (Figure 3.6), we observed a large gain of methylation in \( Trf \) (~5-fold), whereas, in contrast, we observed a small increase (up to 1-fold) of methylation in the other constructs. This result is concordant with the endogenous pattern of these regions (Figure 3.6, bottom panel) and suggests that the piggyBAC reporter system can faithfully recapitulate the endogenous dynamics in DNA methylation.

Finally we asked whether the results obtained with this system held true across biological replicates. In Figure 3.7 levels of methylation at single CpGs across \( OrmI, Mrap, Zic3 \) and \( Syt1 \) were correlated between two biological replicates. We observed high concordance (\( r = 0.97, \text{p-value} < 2.2 \times 10^{-16} \)) indicating that the assay is highly reproducible. These results indicate that the piggyBAC methylation reporter system can faithfully recapitulate endogenous methylation patterns down to single CpGs and is highly reproducible across biological replicates.

### 3.2.2 Dissection of the Oct4 regulatory elements as a proof of concept experiment

\( Oct4 \) is a master regulator of stem cell fate [58]. As \( Oct4 \) is completely un-methylated in ESC and acquires methylation upon differentiation [62], we asked which parts of the regulatory sequence of \( Oct4 \) determined the establishment of its methylation pattern. The
I looked at the correlation between CpG methylation of 2 independent biological replicates: The correlation is very strong and suggests that the system is actually very robust. Also since these are independent insertion events it means that I have saturated the minimum number of insertions to ensure robustness of the assay.

I then computed the fold-change of methylation to see if I can quantitatively discriminate the region that changes with constant regions. My formula is

\[ F = \frac{\text{methNeu}}{\text{methESC}} \]  

which possibly can be generalized into

\[ F = \frac{\text{methT}_2}{\min(\text{methT}_1, \text{methT}_2)} \]  

in a time series of two points (\( T_1 \) and \( T_2 \)). This will assess the difference in methylation between the 2 time points with a normalization scaling for the effect size. That way I should really read out a strong signal coming from sequences that go from U to M or M to U (+ and - respectively).

This figure shows the gain in DNA methylation in the analyzed control regions by comparing methylation in ES cells versus NP cells. On the lower panel, the expected methylation pattern of the endogenous locus is shown.

expression of \( Oct4 \) is controlled by a distal enhancer, a proximal enhancer, and its promoter sequence spanning \( \sim 2 \text{Kb} \) [61].

Despite extensive molecular dissection and differentiation experiments, the \( cis \)-acting elements regulating the methylation patterns observed in \( Oct4 \) are still largely unknown. To
This figure shows a correlation plot of methylation levels at single CpGs across the across *OrmI,Mrap,Zic3* and *Syt1* constructs.

dissect their role to the DNA methylation of *Oct4*, I generated 14 PCR-amplicons spanning across its regulatory region covering at least one CpG (Figure 3.8) and tested their potential in establishing methylation by inserting them into the piggyBAC methylation reporter. To facilitate the construction of this library, the technique adopted for this experiment was
based on assembly PCR. This protocol generated linear transposons compared to circular molecules generated in the previous experiment (see Materials and methods for details).

After construction, I then transfected the Oct4 linear mini-library inside RW4 ESCs, put them under selection for one week, and then differentiated them to NPCs. I then harvested and processed the DNA before and after differentiation, sequenced it and measured the level of methylation (details described in Materials and methods). Additionally, I added a set of samples consisting of RW4 stem cells frozen and thawed after being transfected with the library to test whether their storage could have an effect on methylation. For each sample, two biological replicates were generated.

**Figure 3.8. Design of the Oct4 mini-library experiment**

This figure represents a snapshot of the Oct4 regulatory regions seen from the UCSC genome browser [170]. The amplicons generated for the mini-library are labeled MMOct4dpp and are represented as black bars in the lower side of the figure. Also included are regions representing the distal enhancer, the proximal enhancer, and the promoter. Mammalian conservation and DNAseI hypersensitivity tracks are also displayed.

**Different inserts reproducibly display different levels of DNA methylation**

I first asked whether the constructs showed different patterns of DNA methylation and whether those patterns changed comparing the different samples. As seen in Figure 3.9...
each construct showed distinct and specific levels of DNA methylation. Level of methylation across the constructs were highly reproducible across biological replicates ($r = 0.94\sim0.97$, Figure 3.9). Reproducibility was also high at the level of single CpGs ($r = 0.87\sim0.98$, Figure 3.10). Because of the linear nature of this library, I expected a 10-fold decrease in integration efficiency (see Materials and methods) and in the resulting number of insertions. This translates to an average of 179 insertions per construct, compared to 25000 expected insertions per construct in the previous experiments. These results further confirm the robustness in reproducibility of this assay and its amenability to a multiplexed library.

By comparing fresh RW4 versus frozen and thawed (Figure 3.9), we did not observe any substantial difference between the two conditions, indicative of the robustness of the process of methylation establishment. Contrary to our expectations, the level of methylation of these constructs was also similar when ESCs and NPSc were compared (Figure 3.9), whereas in contrast the full $Oct4$ region acquired methylation during differentiation. There are multiple explanations for this result. One possibility is that the regulatory elements driving DNA methylation act cooperatively over a larger sequence region. As a result, breaking the promoter in smaller fragments could potentially destroy their effect. Alternatively, the regulatory region controlling methylation in $Oct4$ could have also potentially been missed. As shown in Figure 3.8, by selecting only regions containing at least a single CpG, a conserved (and therefore potentially functional) region between positions -1412 and -1290 from the TSS was excluded. Currently I am working to test these hypotheses.
Figure 3.9. Methylation levels of the Oct4 mini-library in different cell types

This figure indicates methylation levels across the Oct4 methylation library inside the piggyBAC methylation reporter system. Barplots indicate methylation levels across the 14 Oct4 constructs. Each plot consists of two biological replicates in a given condition/cell type. On the right, correlation plots of the methylation between the two biological replicates for each set are displayed. The corresponding Pearson’s correlation coefficient of each correlation is shown at the bottom right of each plot.
Figure 3.10. Single CpG methylation reproducibility in the Oct4 library

This figure shows the correlation of single CpG methylation levels across biological replicates in the piggyBAC methylation reporter system for the Oct4 mini-library. Each column indicates a particular condition/cell type. The first row, characterized by blue dots, consists of correlation plots of methylation levels for single CpGs across the bisulfite PCR product of the piggyBAC reporter constructs. The second row, characterized by red dots, is only limited to CpGs belonging to the variable DNA insert inside the piggyBAC reporter (indicated by the light grey sequence above). The third row, characterized by green dots, is instead limited to CpGs belonging to the fixed region of the piggyBAC reporter (indicated by the dark grey). The corresponding Pearson’s correlation coefficient of each correlation is shown at the bottom right of each plot.
Sequence elements in the inserts explain its levels of DNA methylation

Given that each insert displayed reproducible levels of DNA methylation, I then asked whether the sequence of the insert could explain its methylation pattern. I performed this analysis in two different ways. First, I asked whether the sequences contained any statistical signature that was correlated with DNA methylation and that could be used to build a predictive model. The advantage of this strategy is that it does not rely on any prior knowledge and is therefore completely unbiased. Because of the limited size and number of regions that I have profiled in this library I limited this analysis to short words of fixed size, or “k-mers”. I chose to perform my analysis looking at k-mers with $k$ equal to 2, i.e. all the possible dinucleotides as every dinucleotide was indeed observed at least once across the sequences (by taking into account reverse complements as well). Each dinucleotide showed a distinct correlation strength with respect to DNA methylation (Figure 3.11) with extremes ranging from 0.4 to -0.4. To assess whether the dinucleotides contained enough information to recapitulate the observed methylation patterns, I built a linear regression model that used dinucleotide frequencies as features and predicted methylation in output (Figure 3.11 right).

Using forward regression as a feature selection strategy [171], I trained the models on one dataset and computed the explained methylation variance ($r^2$) on the remaining 5 (Figure 3.11, right panel, blue line). Because of limited sample size, to test for over-fitting, I performed permutation tests where the methylation-dinucleotide pairs were shuffled and the models were retrained 100 times (Figure 3.11, right panel, grey line). Using only 4 dinucleotides (GC, CA, AC, and TA) I was able to explain $\sim$70% of the variance in DNA
Figure 3.11. Dinucleotide frequencies predict DNA methylation of the $Oct4$ inserts

This figure shows correlations between DNA dinucleotide frequencies and DNA methylation (left panel) and the results of linear regression models aimed at explaining DNA methylation as a function of dinucleotide frequencies (right panel).

methylation, 50% more than what could expected by chance based on the permeation tests. Interestingly and unexpectedly, CpG frequency was poorly correlated with levels of methylation. This result suggests that CpG density per se may not have any role in specifying DNA methylation. Instead specific motifs or sequence elements that contain CpGs may instead specify DNA methylation. This result suggests that the insert sequence contains information that can be used to predict the insert methylation.

Previous evidence has implicated TFs in mediating establishment [63, 71, 82] and removal [67, 68, 69] of DNA methylation. Using Positional Weight Matrices (PWMs) from Jaspar
Figure 3.12. *in silico* occupancy of transcription factors predicts methylation of the *Oct4* inserts

This figure shows the methylation levels predicted by the *in silico* occupancy of SP1 and SP1 with CREBB.

[172] and TRANSFAC [173], I calculated the *in silico* occupancy of each TF modeled in the database and asked whether their occupancy correlated positively or negatively with methylation. *In silico* occupancy was calculated as

\[
O = \sum_i^n \log_2 \frac{P(K_i|Model)}{P(K_i|Background)} \tag{3.2}
\]
where \( i \) represents each base in a sequence of length \( n \), the model is the PWM and the background distribution assumes equiprobable nucleotide frequencies. Because \( SP1 \) has been previously involved in removing and blocking DNA methylation [67, 77], I decided to start by asking whether predicted occupancy of \( SP1 \) was negatively correlated with methylation levels. I found that all the constructs with the exception of one outlier showed strong negative correlation between their methylation levels and \( SP1 \) occupancy (\( r = -0.78 \), Figure 3.12 left panel, blue dots, \( r = -0.39 \) including all points). This result suggests that the levels of methylation of the constructs are indeed mediated by \( SP1 \) with the exception of construct 4, where other potential interactions might be taking place. I then asked which other PWM together with \( SP1 \) could explain most of the variation in the data. The result was the PWM for \( CREBB \), a transcription factor previously linked in the up-regulation of several neuropeptides and neuron-specific genes [174]. A simple linear regression model using the PWMs for both \( SP1 \) and \( CREBB \) indeed improves the predictions for DNA methylation (\( r = 0.69 \) Figure 3.12 right panel), suggesting that the regulation of the methylation patterns of these fragments might indeed depend on the interplay between these factors. Follow-up experiments with targeted mutagenesis are currently planned for dissecting the role of these sites. These results suggest that methylation levels across the \( Oct4 \) mini-library can be explained by the presence of regulatory elements in the DNA sequence.
The variable insert region affects the methylation of the methylation reporter system

A fundamental question in the establishment of DNA methylation is whether the signals specifying the pattern of a single CpG originate from the same region or whether methylation can be spread from one CpG to adjacent ones. The motivation underlying this question has a practical basis as well: if a region contains the information necessary to specify its own methylation pattern, can this region be used to modify/affect the methylation of nearby regions? To answer this question, I engineered two CpGs in the piggyBAC reporter backbone on the 5’ and 3’ regions flanking the insert. As these CpGs are fixed across all the construct, they can act as a ”reporter” to read out the effects of the insert sequence.

First, I asked whether the methylation of the CpGs of the reporter sequence was established randomly or whether it was indeed specified by the reporter system. To answer this question, I correlated the level of methylation of the CpGs of the reporter from two biological replicates and compared it to the correlation of the methylation levels of the CpGs belonging to the insert sequence (Figure 3.10, green and red dots respectively). The high correlation observed between biological replicates indicates that the methylation of the reporter is not established randomly but rather suggests that each CpG is highly regulated. Then, I asked whether methylation of the insert sequences was correlated with the methylation of the reporter CpGs. For all the analyzed experiments, I was able to observe good correlations between insert and reporter methylation (Figure 3.13). Furthermore this correlation persists between single CpGs of the reporter and the most proximal CpG in the insert sequence.
These results indicate that the CpGs in the reporter system can be used as a readout for the methylation patterns specified by the sequence of the insert.

3.2.3 Future directions:

The experiments that have been presented so far describe the creation and the testing of a methylation reporter system. The goal of this project is ultimately to ask very specific and detailed questions on the contribution of DNA sequence elements to DNA methylation. A key advantage of the piggyBAC methylation reporter system is that it can easily be scaled up to perform large scale genome-wide analysis. We are currently designing and testing a large scale oligo library spanning several differentially methylated regions between mouse ESC and NPCs.

3.3 Materials and methods

3.3.1 Construction of piggyBAC methylation reporter

We used the piggyBAC donor as the backbone from [76]. Circular donor plasmids were built for the Trf, OrmI, Syt1, Zic3, and Mrap constructs by cloning them using the restriction sites XhoI and ClaI. Restriction sites were engineered on the inserts by PCR. The sequence of the inserts were obtained from [70]. For the construction of libraries, we adopted a strategy based on linear assembly. Briefly, a linear donor vector was generated by assembling 4 linear DNA fragments or building blocks via PCR [175]. Each building block was generated by PCR and required to have approximately 20bp of overlapping sequence with the adjacent
Figure 3.13. Methylation of the insert affects the methylation of the reporter

This figure shows the correlation between the methylation of the variable insert region and the fixed reporter region in the piggyBAC methylation reporter system for the Oct4 mini-library (diagram on the left side). For each experiment, methylation level of the insert region of a single construct is plotted against the methylation level of the reporter region for the same construct. Each column represents a distinct biological replicate of the same condition/cell type. The corresponding Pearson’s correlation coefficient of each correlation is shown at the bottom right of each plot.
Figure 3.14. Methylation of the insert affects the methylation of the reporter at a single CpG level

Analogous to Figure 3.13, this figure shows the correlation between the methylation level of the CpGs of the fixed reporter region most proximal to the insert and the corresponding nearest CpG in the variable insert region (diagram on the left side). Each column represents a distinct biological replicate of the same condition/cell type. The corresponding Pearson’s correlation coefficient of each correlation is shown at the bottom right of each plot.
block. The first and the fourth building blocks contain the Inverted repeats necessary for
the integration of the transposon, the second building block contains a puromycin resistance
cassette and the third building block consists of a library of sequences to be tested. Each
fragment contains overlapping ends. The primers used for the construction are shown below:

<table>
<thead>
<tr>
<th>Building Block</th>
<th>F</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building Block 1</td>
<td>gttttccagtcacgacgtt</td>
<td>ggctgcaggaattcgataaa</td>
</tr>
<tr>
<td>Building Block 2</td>
<td>tttatcgaattcctgcagccagcccaattctgtggaatgt</td>
<td>ggaatgaaaatctgtagca</td>
</tr>
<tr>
<td>Building Block 3</td>
<td>tgctacgagatttcgattccaagtgtaaagcctggggtgcctaat</td>
<td>atatctttctttgtagataaagatcttcttctgagatcttttt</td>
</tr>
<tr>
<td>Building Block 4</td>
<td>atctataacaagaaaaat</td>
<td>ctcctcggcaaaacttttc</td>
</tr>
</tbody>
</table>

The colored primer sequence indicates the overlapping region between construct necessary
for the assembly protocol. Each block was generated using piggyBAC donor as template. The
reaction was performed by using 12.5µl of 2X Phusion HF master mix (NEB), 1M Betaine,
0.4uM primers, and 42ng of piggyBAC donor in a final volume of 25 µl. For BB1 and BB4
we used the following protocol: 98C for 3’, 29 cycles of 98C for 10”, 50C for 30”, 72C for
30”, followed by 72C for 10’. For BB2 we used the following protocol instead: 98C for 3’, 6
cycles of 98C for 10”, 56C for 30”, 72C for 1’, followed by 29 cycles of 98C for 10”, 65C for
30”, 72C for 1’, followed by 72C for 10’.

Libraries were assembled using the following PCR protocol: The reaction was performed
by using 12.5µl of 2X Phusion HF master mix (NEB), 1M Betaine, 0.4uM primers, and 0.1
pmoles of the required building blocks in a final volume of 25 µl. We then used the following
protocol: 98C for 3’, 6 cycles of 98C for 10”, 63C for 30”, 72C for 1’, 29 cycles of 98C for 10”, 72C for 30”, 72C for 1’, followed by 72C for 10’. The assembly protocol was executed as follows: first BB1,BB2, and BB3 were assembled in a single reaction. The BB1∼2∼3 product was then assembled to BB4 in a second PCR reaction.

3.3.2 Construction of the Oct4 minilibrary

We built the Oct4 mini-library using the linear library construction strategy. The following primer pairs were used to amplify fragments from the mouse genome:
<table>
<thead>
<tr>
<th>Primer</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>1R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>2F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>2R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>3F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>3R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>4F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>4R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>5F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>5R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>6F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>6R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>7F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>7R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>8F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>8R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>9F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>9R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>10F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>10R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>11F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>11R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>12F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>12R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>13F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>13R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>14F</td>
<td>AGTGTAAAGCCTGGGGTGCCTAATGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
<tr>
<td>14R</td>
<td>AAAGGATCTTCTTGGAGATCGATCGATCGATCGGCTGCAGGCATACTTGAAC</td>
</tr>
</tbody>
</table>

These fragments were then pooled together at a final concentration of 100nM and used as templates to generate BB3. The library was then assembled as described above.

### 3.3.3 Cell culture protocols and differentiation

Experiments were carried out using the RW4 cell line. Cells were grown on gelatin coated plates with High Glucose DMEM with L-Glutamine (DMEM 5796) (Sigma-Aldrich) supplemented with 10% FBS, 10% NCBS, 1% nucleosides, 0.1% Leukemia Inhibiting Factor,
and 0.2% β-mercaptoethanol (working medium). Differentiation from ESC to NPC was performed as described in [176]. Briefly, undifferentiated cells were grown in DFK5 medium [176] for 2 days in low-adhesion plates. At this stage, cells will turn in Embrioid Bodies (EBs). After 2 days, EBs were plated in gelatin coated plates with DFK5 supplemented with 2µM retinoic acid (Sigma) and 30 nM of Shh agonist Hh-Ag 1.4 (Curis) for 4 days.

3.3.4 Transfection and selection

Transfection experiments were performed on single wells of a 6-well plate(Cornig) using Lipofectamine 2000 (Invitrogen). Briefly, a well was seeded with 500,000 cells in 2ml of working medium 24hrs before transfection. Transfection was performed using 2µg of donor and helper vectors with 10ul of Lipofectamine in 500ul of DMEM 5796 as described by the manufacturer. Between 9 and 24hrs after transfection, the medium of each well was replaced with working medium supplemented with 2.5ug/ml Puromycin (Sigma-Aldrich). Selection is maintained for at least 3 days. After selection cells were passaged and grown in working media for at least 2 days.

3.3.5 DNA extraction and bisulfite conversion

Genomic DNA was harvested from the cells using standard phenol/clorophorm extraction followed by isopropanol purification [177]. Between 250ng to 1µg of DNA were then bisulfite-converted using the Epitect Bisulfite Conversion kit (Qiagen) following standard protocol. Bisulfite treated DNA was then amplified with bisulfite-specific primers (Forward: AAGTG-
TAAAGTTTGGGGTGTTTAAT, Reverse: AAAAAATCTTCTTAAAATCCTTTTT) using 0.5 units of Jumpstart TAQ (Sigma), 1M Betaine, 0.4uM primers, 40µM dNTPs, and 5 µl of bisulfite converted DNA in a final volume of 25ul. The protocol was 95°C for 10’, followed by 5 cycles of 95°C for 30”, 55°C for 30”, 72°C for 1’, 5 cycles of 95°C for 30”, 50°C for 30”, 72°C for 1’, 40 cycles of 95°C for 30”, 45°C for 30”, 72°C for 1’, followed finally by 10’ at 72C.

3.3.6 Next generation sequencing library preparation

DNA from long constructs (over 200bp) was prepped for next generation sequencing on an Illumina miSEQ (Illumina) as previously described [121]. DNA from the Oct4 minilibrary was prepped by adding sequencing adapters and indices by PCR (Forward adapter primer: ACACTCTTTCCCTACACGACGCTCTTCCGATCT AAGTGTAAAGTTTGGGTGTTTAAT, Reverse adapter primer: TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT AAAAAATCTTCTTAAAATCCTTTTT). The reaction was performed by using 12.5µl of 2X Phusion HF master mix (NEB), 1M Betaine, 0.4uM primers, and 1µ of bisulfite PCR product in a total volume of 25µl. The protocol was 98°C for 3’, followed by 6 cycles at 98°C for 10”, 63°C for 30”, 72°C for 15”, and 12 cycles of 98°C for 10”, 72°C for 30”, 72°C for 15”. Indexed primers were added to the adapter PCR products using the same PCR protocol (Forward index primer: ATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT, Reverse index primer: CAAGCAGAAGACGGCAT-
ACGAGAT [7bp index sequence] GTGACTGGAG TTCAGACGTGTGCTCTTCCGA). Libraries were then sequenced on an Illumina miSEQ machine (Illumina).

### 3.3.7 Computational analysis and methylation calls

Bisulfite reads were aligned to their reference using a custom bisulfite aligner that accounts for C/T transitions in the context of CpGs. Once aligned, methylation levels were called by estimating the frequency of C/T variants for each CpG dinucleotide using the SPLINTER algorithm [14]. Positional weight matrix analysis was performed as described in the main text. Briefly Positional Weight Matrices (PWMs) from Jaspar [172] and TRANSFAC [173] were scanned on the sequence of the Oct4 mini-library as previously described [178]. For each PWM, the *in silico* occupancy was calculated as

\[
O = \sum_{i}^{n} \log_2 \frac{P(K_i|Model)}{P(K_i|Background)}
\]

where \( i \) represents each base in a sequence of length \( n \), the model is the PWM and the background distribution assumes equiprobable nucleotide frequencies. Linear regression model fits, permutation tests, and correlation analysis were performed using the statistical programming language R.
4. Origin and consequences of the relationship between protein mean and variance

4.1 Foreword

This chapter describes the biophysical modeling of the relationship between single-cell protein variance and its corresponding mean levels. The focus of this modeling effort is to use this relationship to interpret the mechanism through which different biological processes increase protein variance through mean levels. Furthermore, we were able to recapitulate this relationship and consequently predict single-cell protein variance at a genome-wide scale by integrating information from empirically derived biophysical parameters. This work was conceived by Robi D Mitra, PhD, Ilaria Mogno, PhD, and myself. Robi D Mitra, Marc Sherman, Barak A Cohen, PhD, designed the analysis. I performed the analysis in collaboration with Marc Sherman, Zane Goodwin, and Robi D Mitra. Finally I performed the validation experiments. This work is adapted from a draft of a corresponding manuscript in the process of being submitted for publication.
Cell-to-cell variance in gene expression (noise) is a ubiquitous phenomenon that can increase fitness by generating phenotypic differences within clonal populations of cells. An important challenge is to identify the specific molecular events that control noise. This task is complicated by the strong dependence of a protein’s cell-to-cell variance on its mean expression level through a power-law like relationship \( \sigma^2 \propto \mu^{1.69} \). Here, we dissect the nature of this relationship using a stochastic model parameterized with experimentally measured values. This framework naturally recapitulates the power-law like relationship \( \sigma^2 \propto \mu^{1.6} \) and accurately predicts protein variance across the yeast proteome. Using this model we identified two distinct mechanisms by which protein variance can be increased. Variables that affect promoter activation, such as nucleosome positioning, increase protein variance by changing the exponent of the power-law relationship. In contrast, variables that affect processes downstream of promoter activation, such as mRNA and protein synthesis, increase protein variance in a mean-dependent manner following the power-law. We verified our findings experimentally using an inducible gene expression system in yeast. We conclude that the power-law-like relationship between noise and protein mean is due to the kinetics of promoter activation. Our results provide a framework for understanding how molecular processes shape stochastic variation across the genome.
4.3 Introduction

Stochastic fluctuations in the biochemical processes that underlie gene expression produce cell-to-cell variation in protein levels, or noise [89, 97, 105]. Noise performs several biological functions. In unicellular organisms, noise improves fitness by generating phenotypic differences within clonal populations of cells, thus enabling a rapid response to fluctuating environments [88, 92, 179]. In multi-cellular organisms, noise plays a role in development, allowing identical progenitor cells to acquire distinct fates [91, 93, 180]. Because of its functional importance, a fundamental goal is to identify and dissect the molecular mechanisms that generate and control noise. Single-cell studies have connected pathway-specific (extrinsic) and gene-specific (intrinsic) factors to changes in noise [89, 110, 181]. These factors include the rate of transcript elongation12, the presence of a TATA-box [89, 112, 119, 179], nucleosome positioning at the promoter sequence [89, 113, 114, 115, 182], fluctuating mRNA levels [111], translation rate [86, 87], pathway-dependent fluctuations [111, 181], and asymmetric partitioning at cell division [183]. Determining the relative contribution of these factors to the noisiness of any given gene remains an important challenge. Understanding the contribution of any particular molecular process to noise is complicated by the strong dependence of a protein’s cell-to-cell variance on its mean expression level [86, 94, 111, 181]. Several studies have revealed that a protein’s cell-to-cell variance is related to its mean expression level by a power-law-like relationship (\( \sigma^2 \propto \mu^j \)) [94, 111, 120, 184]. Any investigation into the mechanistic origins of noise must account for this power-law-like relationship, as any process that increases the mean expression level of a protein will necessarily increase its
cell-to-cell variance. Only by accounting for the power-law-like relationship is it possible to identify processes that increase noise beyond what is expected given the processes effects on mean protein levels. Despite empirical observations, the biophysical origin of the power-law-like relationship between the variance in protein levels and mean protein levels is not yet clear. To address this problem, we analyzed noise using a stochastic model of gene expression parameterized with experimentally measured kinetic rates. This model recapitulated the precise relationship \( \sigma^2 \propto \mu^{1.6} \) between mean and variance and accurately predicted protein variance on a proteome-wide scale \( r^2 = 0.935 \). By further analyzing this model, we identified two distinct classes of processes that influence noise. First, variables that affect the rate of promoter activation, such as the presence of promoter-positioned nucleosomes or TATA boxes, increase noise in a mean-independent fashion. In contrast, variables that influence processes downstream of promoter activation, such as the synthesis and degradation of mRNA and protein, increase noise in a mean-dependent manner. Our results suggest that the kinetics of promoter activation determine the power-law-like relationship between mean protein levels and variance in protein levels. In support of this claim, we found that changing the rate of promoter activation modulates the exponent of the power-law and consequently the scaling between variance and mean. By providing a mechanistic interpretation of the power-law-like relationship, our work provides the framework to achieve a better understanding of the molecular processes that lead to cell-to-cell variation in gene expression.
4.4 Results

4.4.1 Protein mean and variance are connected by a power-law-like relationship

We started by characterizing the relationship between mean protein levels and cell-to-cell protein variance across the yeast genome. We analyzed data consisting of \( \sim 2200 \) S. cerevisiae GFP fusion strains for which protein levels had been measured at a single-cell resolution by flow-cytometry [112]. We performed a log-log regression analysis of cell-to-cell protein variance as a function of the mean protein levels and observed a power-law-like relationship with an exponent of 1.69 (Figure 4.1), in agreement with previous findings [94, 120]. Ninety-seven percent of protein variance across the proteome can be explained solely by mean levels through this relationship, indicating that highly expressed genes naturally exhibit high cell-to-cell variation whereas genes expressed at low levels are more uniformly expressed across different cells. Although the mean-independent component of protein variance accounts for only 3% of the total variation, we found that, for certain genes, mean-independent contributions increased protein variance up to 20-fold higher than expected (Figure 4.2). In contrast, very few genes displayed smaller protein variances than expected given mean levels. In fact, we observed, at most, a 2-fold reduction (see Figure 4.2). Taken together, these results indicate that for most genes, protein variance is largely explained by the protein mean through a power-law-like relationship, except for a few notable cases in which protein variances are increased substantially beyond their expected values.
4.4.2 A stochastic model of gene expression recapitulates the power-law-like relationship between mean and variance

We next sought to understand the molecular origin of the relationship between protein mean and variance. One hypothesis is that this relationship originates purely as a consequence of stochasticity in the steps of gene expression. Alternatively, this relationship could result from mechanisms that are independent of expression, such as asymmetric partitioning of protein and RNA molecules at cell division [183] or pathway-dependent fluctuations in trans-acting factors [181]. To distinguish between these two hypotheses, we tested whether a
Figure 4.2. Residual mean-independent variance in protein expression

Distribution of residual variance values across the S. cerevisiae dataset. Red bars indicate residual variance value with Z-scores over 2 standard deviations from the mean.

stochastic model based only on the processes involved in gene expression could recapitulate the observed power-law relationship. We applied a model [117] that describes cell-to-cell protein variance at steady-state as a function of kinetic parameters for promoter activation/inactivation events and mRNA and protein production/degradation (Figure 4.3, Figure 4.9). For most parameter values, we used empirical measurements (see supplementary methods). This was not possible, however, for the rates of promoter activation and inactivation, which have only been measured in a few genes [185]. Since no high-throughput methods exist for measuring rates of promoter activation and inactivation, we assumed that the promoter
kinetics would be similar across the genome and fit their values from the data (supplementary methods). The model converged to a regime in which promoter activation is an infrequent event that is quickly followed by promoter inactivation, a result supported by published experimental data [20] (supplementary information, and Figure 4.9). We obtained a rate of promoter activation ($K_{on}$) of 0.59 min$^{-1}$, a value that agrees with empirically measured activation rate for the $GLT1$ gene in yeast (1.3 ± 0.72 min$^{-1}$) [185]. Using this value for $K_{on}$, the model naturally generates a power-law-like relationship between mean and variance that is similar to the one observed empirically (modeled relationship: $\sigma^2 \propto \mu^{1.60}$, observed relationship: $\sigma^2 \propto \mu^{1.69}$). Furthermore, our framework correctly predicts protein variance across the genome (log space $r = 0.962$, $p < 2.2 \times 10^{-16}$; linear space $r = 0.839$, $p < 2.2 \times 10^{-16}$, Figure 4.4). We tested for over-fitting by performing 2-fold cross-validation 100 times and again found good agreement ($r = 0.957 \pm 0.018$, $p < 2.2 \times 10^{-16}$). Taken together, these results support the validity of our model and suggest that the power-law relationship between mean and variance depends solely on the kinetics of the processes that underlie gene expression.

4.4.3 The power-law-like relationship between protein variance and mean depends on promoter kinetics

Our analysis indicates that the power-law-like relationship between mean and variance is a consequence of the steps that lead to gene expression. We next sought to determine which of these processes specify the exponent of the power-law-like relationship. Using our biophysical
Schematic representation of the model. Each step transition is determined by a rate constant. Promoter activation and inactivation occur at $K_{on}$ and $K_{off}$ rates respectively. When active, a promoter is transcribed at $K_m$ rate into an mRNA molecule. The mRNA molecule can then be either degraded at $D_m$ rate or translated at $K_p$ rate into a protein. The protein molecule can then be degraded at rate $D_p$. $K_{on}$, $K_{off}$, and $K_m$ determine the synthesis rate of mRNA, or $S_m$. Blue indicates that the parameter has been empirically measured or calculated across the dataset, red indicates that the parameter has been simplified or fit across the dataset.

model, we randomly sampled transcription and translation rates, as well as degradation rates of mRNA and protein, while maintaining the same promoter activation regime we identified above ($K_{on} = 0.59 \, \text{min}^{-1}, \, K_{on} << K_{off}$). Virtually all permutations resulted in a power-
Figure 4.4. Model prediction of genome-wide protein variance in *S. cerevisiae*

**S. cerevisiae**

genome-wide protein variance

Model performance in predicting protein variance in *S. cerevisiae*. Each point represents a single GFP fusion strain. Data is displayed in log-scale (linear scale $r = 0.836$)

law relationship between mean and variance that was nearly identical to the one observed experimentally (exponent $= 1.612 \pm 5.9 \times 10^{-3}$, 1000 permutations, Figure 4.5). This result indicates that, when $K_{on} \ll K_{off}$, the exact form of the power-law-like relationship between
mean and variance is independent of the rates of transcription, translation, and protein and mRNA degradation. In contrast, we found that the exponent of the power law was strongly dependent on promoter kinetics. Using the same modeling framework, we changed the parameters governing promoter transitions to enforce a slow kinetics regime ($K_{on}$ and $K_{off} << K_m, D_m, K_p, D_p$). We found that protein mean and variance followed a quadratic relationship (exponent = 1.97, Figure 4.5), which differs substantially from our previous results and the observed power-law. Taken together these results suggest that the power-law relationship between protein mean and cell-to-cell variance is dictated by the kinetics of promoter activation, and is largely insensitive to downstream steps.

4.4.4 The relationship between protein mean and variance separates mean-dependent from mean-independent sources of variance

A strong prediction of our model is that perturbations which affect the kinetics of promoter activation should increase noise in a mean-independent manner, whereas perturbations that affect processes downstream of promoter activation should increase noise in a mean-dependent fashion. Several variables have previously been correlated with increases in noise including changes in transcription\textsuperscript{13} and translation rates \textsuperscript{[86, 87, 89]}, the presence of a TATA box \textsuperscript{[87, 89, 112]}, and promoter positioned nucleosomes \textsuperscript{[89, 113, 114, 182]}. Our model suggested that only variables involved in promoter activation should increase protein variance independently of mean levels, in contrast to variables affecting downstream processes. To test this hypothesis, we correlated the power-law residuals with variables that reflect
Predicted relationship between mean and variance using original model with original parameter set (grey squares), original model with permuted sets of kinetic rates for mRNA/protein synthesis and degradation (purple), and slow promoter kinetics model with original parameter set (orange).

changes in promoter activation, and with variables that affect downstream processes. Genes with TATA boxes or promoter-positioned nucleosomes, factors which influence promoter activation, had high values of residual variance (Figure 4.6), indicating that they increase noise independently of the mean. In contrast, differences in measured rates of mRNA synthesis and degradation \[186\], rates of protein degradation \[187\], measures of ribosomal occupancy\[188\], and the Codon Adaptation Index \[189\] showed little or no correlation with residual variance
This result demonstrates that these variables, which affect processes downstream of promoter activation, influence cell-to-cell protein variance almost exclusively by changing mean levels of gene expression. Taken together, the results support our hypothesis and point to positioned nucleosomes in particular as a potent source of mean-independent noise.

Figure 4.6. Promoter-associated factors modulate protein variance independently of the mean

Fraction of residual variance explained ($r^2$) by sources of noise operating at the promoter/initiation level (orange) or at a post-initiation level (purple).
4.4.5 Experimental confirmation of the relationship between promoter-positioned nucleosomes and mean-independent variance

To obtain additional support for these findings, we experimentally tested whether changes in nucleosome occupancy could produce an increase in the mean-independent component of gene expression. Using in vivo nucleosome positioning data [190], we selected a set of \textit{S. cerevisiae} TATA-containing genes whose promoters are nucleosome free in glucose but which acquire a positioned nucleosome in ethanol. A prediction of our analysis is that such genes would display increased residual variance when switched from glucose-containing medium to ethanol-containing medium. We measured the distribution of fluorescence of GFP-tagged fusion strains [187] in both glucose and ethanol by flow-cytometry, and computed the residual variance above what is expected from the mean-variance relationship. We observed a significant increase in residual variance as cells were shifted from glucose to ethanol relative to a control set of genes in which nucleosomes do not change between the two conditions (p-value < 0.05, T-test across 3 biological replicates, Figure 4.7). Using this same gene set, we examined whether changes in protein translation rate affected the mean-independent component of the variance. Our model predicts that translation rate should not correlate with residual variance, and we did not observe any significant difference (p-value > 0.4, T-test across 3 biological replicates, Figure 4.7). These results support our hypothesis that positioned nucleosomes are the major source of mean independent noise. We conclude that nucleosome bound promoters showed higher protein variance as a result of slowed promoter
activation kinetics, which increases the exponent of the power-law-like relationship between protein mean and variance.

Figure 4.7. Changes in nucleosome occupancy but not in translation rate induce changes in mean-independent variance

Increment in residual variance from glucose to ethanol in genes that show increased occupancy in ethanol (orange set: test) and genes with unaltered occupancy (orange set: control) compared to the same genes ranked by high (purple set: test) or low (purple set: test) increase in translation rate (purple set) (* indicates p < 0.05, t-test).
4.4.6 Promoter-positioned nucleosomes increase variance by slowing promoter activation kinetics

Our model suggests that the increase in residual variance caused by positioned nucleosomes is the result of slower promoter activation in these genes. To test this hypothesis, we examined single-cell mRNA measurements performed for different genes in *S. cerevisiae* [191], since the relationship between mRNA mean and variance can be used to distinguish groups of genes with different promoter kinetics [120] (see materials and methods). Our prediction is that genes without promoter-positioned nucleosomes (Figure 4.8b) will have fast promoter activation kinetics and thus display an approximately linear relationship between mean and variance (Figure 4.8a, blue line, see supplementary methods). Indeed, this was observed in the single-cell mRNA data (see Figure 4a, red dots). In contrast, our model predicts that genes with promoter-positioned nucleosomes (Figure 4.8c) will have slow promoter activation kinetics and will therefore display a quadratic scaling between mean and variance (Figure 4.8a, red line, see supplementary methods). This was again confirmed – genes lacking a nucleosome-free region (NFR) displayed the predicted mean-variance relationship (Figure 4.8a, red dots).

4.5 Discussion

A fundamental property of protein variance is its dependence on mean protein levels through a power-law-like relationship. This power-law-like relationship holds in yeast ($\sigma^2 \propto \mu^{1.6}$), bacteria ($\sigma^2 \propto \mu^{1.5}$) [184], and human T-cells ($\sigma^2 \propto \mu^{1.7}$) [94], suggesting the
Figure 4.8. Single-cell mRNA variance connects underlying promoter kinetics to nucleosome occupancy

a) mRNA mean and variance in *S. cerevisiae* plotted against each other in log-scale. Blue dashed line indicates the expected relationship between mean and variance in a regime of slow activation and fast inactivation rate ($\sigma^2 = \mu$), red dashed line indicates expected relationship at slow promoter kinetics ($\sigma^2 = \mu + \mu^2$). Circles represent experimental values of mRNA mean and variance (color matches best fit to promoter kinetics regime) b) Average nucleosome occupancy between -600 to +1000 relative to the TSS of *S. cerevisiae* genes exhibiting linear mRNA mean-variance scaling. The position of the canonical nucleosome free region is indicated by the black arrow. c) Same as b) but with respect to *S. cerevisiae* genes exhibiting quadratic mRNA mean-variance scaling.
processes that determine the power-law are common across species. Our work supports this conclusion. We found that the power-law is a natural consequence of the kinetics of transcription and translation, fundamental mechanisms shared between the three organisms. Molecular processes that differ significantly between these species, such as chromatin structure, nuclear export, or unequal partitioning during the cell cycle, were not required to explain the power-law. Our work suggests that the power-law is a universal feature of gene expression whose particular shape is determined by the rates at which promoters transition between their active and inactive states. We therefore expect to observe this power-law in all cell types, with an exponent that reflects the particular promoter kinetics regime of the cell. One practical application of understanding the power law is that the mean-independent component of noise can be measured. Surprisingly, we found that, in yeast, there are few genes with a significant mean-independent variance. In fact, 97% of all protein variance across the genome is explained solely by mean protein levels, a result that only becomes obvious after using the power-law to remove intrinsic mean effects. The 3% of genes with excess variance (up to twenty-fold over what is expected) is consistent with the occurrence of slow promoter kinetics, which our data suggest is caused by positioned nucleosomes on their promoters. Finally, we did not observe any genes with variances significantly below that expected from the power-law. Reducing protein variance may be difficult for the cell due to physical constraints that render this process energetically dis-advantageous. A theoretical analysis on the limits of suppression of molecular fluctuations [192] supports this observation. Alternatively, it is possible that cells have evolved regulatory networks with intrinsic
robustness to molecular fluctuations \cite{193}, suggesting that even if achievable, noise reduction may not be necessary. Identifying the sources of noise and their underlying mechanisms is an important step in determining their role in increasing fitness. The work presented here provides a way to isolate mean-independent effects from protein variance and to connect them to their biophysical origins. A long-standing question regarding stochastic gene expression is its role in fitness \cite{179}. Through this framework, it will be possible to decouple the role of protein variance from the mean, allowing a better understanding of the functional and evolutionary constraints that shape gene expression variance.

4.6 Methods

4.6.1 Data Sources

We used single-cell protein mean and variance values from flow-cytometry measurements on \textit{S. cerevisae} GFP-fusion strains grown in YPD for \sim \num{2000} genes from Newman et al. \cite{112}. mRNA level measurements in YPD and YPEtOH were obtained from Gasch et al. \cite{194}. We acquired mRNA synthesis and degradation rates from Miller et al. \cite{186}. mRNA single-cell measurement data were obtained from Gandhi et al. \cite{191}. Nucleosome occupancy was assessed from mnase-seq datasets in YPD and YPEtOH from Kaplan et al. \cite{190}. We used protein mean and variance from synthetic promoter libraries from the work of Mogno et al. \cite{119}. Definition of TATA-containing and TATA-less were obtained from Basehoar et al. \cite{195}. We obtained in vivo ribosome occupancy profiles for each mRNA species measured in YPD from Ingolia et al. \cite{188}. Data and source code generated and used in
4.6.2 Analysis of the relationship between protein mean and variance

Using single-cell protein mean and variance values in *S. cerevisiae*, we assumed that the underlying relationship between mean and variance could be non-linear and exponential in nature. This formulation can be generally expressed as

\[ \sigma^2 = k \mu^J \]  

where \( k \) is a scaling factor and \( J \) is the exponential index. In log-space, this equation transforms into

\[ \log(\sigma^2) = J \log(\mu) + \log(k) \]  

where \( J \) can now be directly calculated as the slope of a linear regression. We estimated the fraction of variance explained by the mean as the \( r^2 \) of the regression. Variance residuals originated from this fit were defined as mean-independent variance. Regression analysis was performed using the R programming language.

4.6.3 Stochastic modeling of protein and mRNA variance

To model mRNA and protein variance in *S. cerevisiae*, we used analytical stochastic models derived from the solution of a system of stochastic differential equations as previously described [117]. This model describes the steady-state value of mRNA and protein variance.
as a function of the kinetic rates for protein activation and inactivation ($K_{on}$ and $K_{off}$), mRNA synthesis and degradation ($K_m$ and $D_m$), and protein translation and degradation ($K_p$ and $D_p$). The model for mRNA variance is described in equation (4.3), whereas the model for protein variance in equation (4.4). In order to predict genome-wide protein variance in $S.\ cerevisiae$, we assumed $K_{on}$ and $K_{off}$ to be uniform across the genome and fit their values. Fitting, prediction and cross-validation were computed in Perl. Analysis of the fit was performed in R. (for complete explanation see Supplementary Information).

4.6.4 Correlation analysis between mean-independent variance and molecular properties

We compared mean-independent variance to mRNA synthesis rate, mRNA degradation rate, ribosomal occupancy and CAI (Codon Adaptation Index). CAI was computed as previously described [189]. To determine the amount of variation of noise explained explained by each property, we correlated mean-independent variance with the log of the measure of each property and calculated the Pearson’s correlation coefficient. We used a linear regression in log scale to avoid any non-linear effects. Regression analysis was performed in R.

4.6.5 Regression model between mean-independent variance and nucleosome occupancy

We computed the Pearson’s correlation coefficient between mean-independent variance and nucleosome occupancy at a single base resolution for each base ranging from -1000 to
+600 relative to the transcription start site of each gene in _S. cerevisiae_ for which we had both nucleosome data and residual mean-independent variance. For each base, we obtained a correlation value, which was plotted as a function of its position relative to the TSS. We repeated this analysis focusing on TATA-containing and TATA-less genes only. In order to estimate the amount of variation explained by nucleosome occupancy on TATA-containing genes, we applied a linear model to predict residual mean-independent variance as a function of nucleosome occupancy. We performed a forward-regression strategy to determine the positions in the promoter sequence to be used as predictive features for our model followed by leave-one-out cross-validation to assess over-fitting (supplementary information for details). Regression analysis was performed in R.

### 4.6.6 Experimental measurement of mean-independent variance as a function of nucleosome occupancy

and YPEtOH to log phase and measured single-cell protein levels using a Beckmann-Coulter
Cytomics FC500 MPL flow-cytometer (Beckmann Coulter, Fullerton, CA) as previously
described previously [112]. We calculated residual variance from mean and variance as de-
scribed above and, for each gene we computed differential residual mean independent vari-
ance between YPEtOH and YPD. For each gene we computed the translation rate in both
conditions as described above and computed the differential translation rate ($\Delta K_p$) between
conditions. We tested for increase in residual variance between the test and control set and
between the top and the bottom half of all strains ranked by decreasing $\Delta K_p$ using one-sided
t-test. Statistics were performed in R.

4.7 Supplementary methods

4.7.1 Stochastic Modeling of mRNA and Protein Variance in S.cerevisiae

Description of the Model

We used a stochastic model describing the expression of a single gene though the process of
its promoter activation, inactivation, transcription at the active promoter state, degradation
of the transcribed mRNA, translation of the transcribed mRNA, and degradation of the
protein. A visualization of the model is described in Figure 4.3. Each step is associated with
a rate constant:

- $K_{on}$ = promoter activation rate
- $K_{off}$ = promoter inactivation rate
• $K_m$ = mRNA transcription rate of active promoter

• $D_m$ = mRNA degradation rate

• $K_p$ = mRNA translation rate

• $D_p$ = protein degradation rate

The model that we used is the analytical solution of a chemical master equation describing $\sigma^2$ for mRNA and protein at steady-state. This solution and its derivation have been previously described [117]. The value of $\sigma^2$ for mRNA levels at steady-state corresponds to:

$$\sigma^2_{mRNA} = \mu_{mRNA}[1 + \frac{K_{off}K_m}{(K_{on} + K_{off})(D_m + K_{on} + K_{off})}]$$  \hspace{1cm} (4.3)

whereas $\sigma^2$ for protein levels at steady-state is defined as:

$$\sigma^2_p = \mu_p\{1 + \frac{K_p}{D_m + D_p}[1 + \frac{(K_{off}K_m)(D_m + D_p + K_{on} + K_{off})}{(K_{on} + K_{off})(D_m + K_{on} + K_{off})(D_p + K_{on} + K_{off})}]\}$$  \hspace{1cm} (4.4)

$\mu_{mRNA}$ and $\mu_p$ are the mean mRNA and protein levels respectively and are defined as

$$\mu_{mRNA} = \frac{<P_{on}>K_m}{D_m}$$  \hspace{1cm} (4.5)

$$\mu_p = \frac{<P_{on}>K_mK_p}{D_mD_p}$$  \hspace{1cm} (4.6)
where $< P_{on} >$ is the probability of the promoter being active, which is specified by $K_{on}$ and $K_{off}$ as

$$< P_{on} > = \frac{K_{on}}{K_{on} + K_{off}} \quad (4.7)$$

**Description of input datasets and calculation of $K_p$**

We applied the model to a single-cell protein measurement dataset in *S. cerevisiae* [112] in order to predict $\sigma^2_p$. The dataset provides values of $\sigma^2_p$ and $\mu_p$ for $\sim$2200 genes. Our experimentally derived parameters were obtained from the following sources:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_m$</td>
<td>Miller et al MSB 2011</td>
</tr>
<tr>
<td>$D_m$</td>
<td>Miller et al MSB 2011</td>
</tr>
<tr>
<td>$D_p$</td>
<td>assumed to be constant ($1/90 \text{ min}^{-1}$)</td>
</tr>
</tbody>
</table>

$S_m$ is defined as the product of $K_m$ with the probability of the promoter being active, which is specified by $K_{on}$ and $K_{off}$ as

$$S_m =< P_{on} > K_m \quad (4.8)$$

and, unlike $K_m$, can be experimentally measured. Because all genes were tagged with GFP, we assumed the fusion proteins to be stable and degraded at a rate equivalent to the dilution rate, i.e. $1/90 \text{ min}^{-1}$ (reported in [http://www.yeastgenome.org/](http://www.yeastgenome.org/)). This value
represents the lower limit of protein degradation as active degradation events will result in higher observed degradation rates. Although $K_p$ was not readily available at a genome-wide scale, it can be easily computed from $\mu_p$, $\mu_{mRNA}$, and $D_p$ as:

$$K_p = \frac{\mu_p D_p}{\mu_{mRNA}}$$  \hspace{1cm} (4.9)

**Fitting $K_{on}$ and $K_{off}$ across the entire *S. cerevisiae* genome**

By applying (4.8) to (4.4), we removed $K_m$ from the model by replacing it with $S_m$, resulting in:

$$\sigma_p^2 = \mu_p \{1 + \frac{K_p}{D_m + D_p}[1 + \frac{S_m}{K_{on}} \frac{K_{off}(D_m + D_p + K_{on} + K_{off})}{(D_m + K_{on} + K_{off})(D_p + K_{on} + K_{off})}]\}$$  \hspace{1cm} (4.10)

which can be parameterized using the data described above. We estimated $K_{on}$ and $K_{off}$ on Newman’s dataset using a gradient descent non-linear fit strategy implemented in Perl (this strategy is described in detail in [14]. We obtained a value of $K_{on} = 0.59 \text{ min}^{-1}$ whereas $K_{off}$ always converged to the maximum allowed value, therefore becoming negligible. This result indicates a regime of slow promoter activation ($K_{on}$) followed by fast inactivation ($K_{off}$). We derived an expression for protein variance under this regimen.

Starting from (4.10), and by assuming that $K_{off}$ dominates over the other parameters, as determined by the fit, the equation simplifies to
\[
\sigma_p^2 \approx \mu_p \left( 1 + \frac{K_p}{D_m + D_p} \left[ 1 + \frac{S_m}{K_{on}} \right] \right)
\]

(4.11)

where \( \frac{S_m}{K_{on}} \) is equivalent to the burst size \([94]\) under this regimen. This can be shown as:

\[
\frac{S_m}{K_{on}} = \frac{K_m}{K_{on} + K_{off}} \approx \frac{K_m}{K_{off}}
\]

(4.12)

We applied (4.11) to estimate \( \sigma_p^2 \) in *S.cerevisiae* using the estimated value of \( K_{on} \) and empirically measured values for the remaining parameters. We obtained a real scale correlation of \( r = 0.839 \) and a log scale correlation of \( r = 0.962 \) between predicted and measured \( \sigma_p^2 \). The results of this prediction are plotted in Figure 4.4.

To check for over-fitting, we performed cross-validation by randomly sampling 50% of the dataset as training set and the second 50% as a validation set 100 times. For each round, we estimated \( K_{on} \) and \( K_{off} \) from the training set and evaluated the goodness of fit on the validation set \( (r = 0.957 \pm 0.018, p < 2.2 \times 10^{-16}) \).

**Validating the assumption of a global invariant activation rate across the *S. cerevisiae* genome**

The results of our fit indicated a fast inactivation rate and suggested a slow and general value of \( K_{on} \) across the *S.cerevisiae* genome. Previous experiments \([94]\) suggest that in mammalian cells it is the case, as mean and variance are affected by changes in burst size \( \left( \frac{K_m}{K_{off}} \right) \) rather than changes in burst frequency \( (K_{on}) \).
In order to test whether this assumption applied to our system, we analyzed a promoter
dataset described above [119]. In this dataset, a library of different promoters was cloned
upstream the same reporter gene. As a result, $K_p, D_m,$ and $D_p$ can be assumed as constants
in the system. We asked whether changes in promoter activity in the libraries were due to
changes in burst frequency $K_{on}$ or burst size ($\frac{K_m}{K_{off}}$). To do so, we analyzed the relationship
between the $VMR (\frac{\sigma_p^2}{\mu_p})$ and $CV^2 (\frac{\sigma_p^2}{\mu_p^2})$ as a function of $\mu_p^2$. Following the expression of $\sigma_p^2$ in
(4.11), we define:

\[
\frac{\sigma_p^2}{\mu_p} = 1 + \frac{K_p}{D_m + D_p} + \frac{K_p}{D_m + D_p} \frac{K_m}{K_{on} + K_{off}}
\]

(4.13)

\[
\frac{\sigma_p^2}{\mu_p^2} = \frac{1}{\mu_p} + \frac{K_p}{(D_m + D_p) \mu_p} + \frac{(D_m + D_p) D_m D_p}{K_{on}}
\]

(4.14)

where the parameters represented in blue indicate constants. Changes in $K_{on}$ will result
in a inverse relationship between both mean and VMR and mean and $CV^2$ (Figure 4.9a). In contrast, changes in burst size (by either $K_m$ or $K_{off}$) with constant burst frequency
would result in a linear relationship between VMR and $\mu_p$ and an asymptotically constant
relationship between $CV^2$ and $\mu_p$ (Figure 4.9b). The experimental data from [119] follows
the second model which assumes changes in burst size with constant $K_{on}$, thereby validating
our assumption (Figure 4.9c).
Figure 4.9. Changes in gene expressions are driven by changes in $K_{off}$ or $K_m$ whereas $K_on$ remains largely constant.

(a) Expected relationship of the VMR ($\frac{\sigma^2_P}{\mu_P}$) (upper half, blue line) and the CV ($\frac{\sigma^2_P}{\mu_P}$) (lower half, red line) with protein mean levels ($\mu_P$) assuming constant $K_{off}$ and $K_m$ and variable $K_on$. (b) Same as in (a) but assuming instead constant $K_on$ and variable $K_{off}$ or $K_m$. Equations indicate the slope of the line for the VMR-mean relationship (upper half) and the equation of the asymptotic line for the CV-mean relationship.

(c) Experimentally observed relationship of the VMR and CV with protein mean levels in a promoter library dataset S. cerevisiae [119].
Power-law relationship analysis

To estimate the exponent of the power-law relationship generated between protein mean levels ($\mu_p$) and predicted protein variance ($\sigma^2_p$) in the Newman et al dataset we performed log-log regression between $\sigma^2_p$ and $\mu_p$ which is defined as

$$\log(\sigma^2_p) = j \log(\mu_p) + \log(k)$$

(4.15)

corresponding to

$$\sigma^2_p = k \cdot \mu_p^j$$

(4.16)
in real space, where $j$ is the exponent of the power-law.

Power-law relationship under slow promoter kinetics

We performed the same power-law analysis described above for a promoter regime characterized by slow activation and inactivation kinetics. Under this regimen $K_{on}, K_{off} << K_m, D_m, K_p, D_p$ and the promoter contributes to $\sigma^2_p$ following a Bernoulli/Binomial process (Paulsson 2005). As described above, in this context the variance of the active promoter ($\sigma^2_{prom}$), can be expressed as

$$\sigma^2_{prom} = \langle P_{on} \rangle (1 - \langle P_{on} \rangle) = \frac{K_{on}K_{off}}{(K_{on} + K_{off})^2}$$

(4.17)
which will reach it’s maximum value when $<P_{on}>$ and $1-<P_{on}>$ will be equivalent. As $K_{on}$ and $K_{off}$ are small in this regimen, by assuming maximum promoter variance, we derived an expression of mRNA variance under bursty kinetics. As $K_{on}$ and $K_{off}$ are small in this regimen, we can re-write (4.4) as

$$
\sigma_p^2 \approx \mu_p \{1 + \frac{K_p}{D_m + D_p} [1 + \frac{K_mK_{off}(D_m + D_p)}{D_mD_p(K_{on} + K_{off})}] \} \approx \mu_p + \mu_p \frac{K_p}{D_m + D_p} + \mu_p^2
$$

(4.18)

We then estimated $\sigma_p^2$ in this regime by computing its values using the same set of parameters used above and again computed the exponent of the power-law relationship between predicted $\sigma_p^2$ and $\mu_p$ using (4.15).

**Power-law relationship: Randomization analysis of molecular parameters**

We developed a Monte-Carlo simulation in which each parameter in the parameter set ($K_{on}$, $K_{off}$, $K_m$, $K_p$, $D_m$, $D_p$) is randomly selected according to the known physiological distribution and range. The physiological ranges used for this simulation were as follows:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Distribution/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{on}$</td>
<td>0.5</td>
<td>0.6</td>
<td>Uniform in log-space</td>
</tr>
<tr>
<td>$K_{off}$</td>
<td>0.003</td>
<td>50</td>
<td>Measured by JR Chubb <em>et al</em> Current Biology 2006 and DM Suter <em>et al</em> Science 2011</td>
</tr>
<tr>
<td>$K_m$</td>
<td>0.05892</td>
<td>12.1</td>
<td>Computed from $S_m$ (measured by Miller <em>et al</em> MSB 2011)</td>
</tr>
<tr>
<td>$D_m$</td>
<td>0.0054301</td>
<td>3.013683</td>
<td>Measured by Miller <em>et al</em> MSB 2011</td>
</tr>
<tr>
<td>$K_p$</td>
<td>0.05892</td>
<td>52.8</td>
<td>Computed from Newman <em>et al</em> Nature 2006</td>
</tr>
<tr>
<td>$D_p$</td>
<td>0.000042</td>
<td>0.3465736</td>
<td>Measured by Ghaemmaghami S <em>et al</em> Nature 2003</td>
</tr>
</tbody>
</table>

$K_{on}$ was defined by using our fitted value of 0.59 min$^{-1}$ and determining a small range around it. $K_{off}$ was defined from [196] and [197]. $K_m$ was computed from $K_{on}, K_{off}$ and $S_m$ using (4.8). $K_p$ was computed from the dataset using (4.9). Each round of Monte-Carlo simulation randomly produced 2400 parameter sets (as large as the dataset from Newman *et al.*). For each set, a protein mean and variance were computed using the previous equations. A linear regression between the log values of mean and variance for the whole dataset was then computed. We repeated this procedure 1,000 times and finally we computed the average slope resulting from each linear regression.

**Derivation of stochastic models for $\sigma^2_{mRNA}$**

We derived expressions for mRNA variance under different promoter kinetics regimens. We started by expressing equation (4.3) as a function of $S_m$, resulting in
\[
\sigma_{mRNA}^2 = \mu_{mRNA}[1 + S_m \frac{K_{off}}{K_{on}(D_m + K_{on} + K_{off})}]
\]  

(4.19)

Under a regimen of fast \( K_{off} \) and slow \( K_{on} \), this equation converges to

\[
\sigma_{mRNA}^2 \approx \mu_{mRNA}(1 + \frac{S_m}{K_{on}})
\]  

(4.20)

We found that in our dataset, the average burst size \( \frac{S_m}{K_{on}} \) corresponds to 0.1042561.

\[
\sigma_{mRNA}^2 \approx \mu_{mRNA}
\]  

(4.21)

Under a regimen of slow promoter kinetics, we expressed equation (4.3) as described above for (4.18), resulting in:

\[
\sigma_{mRNA}^2 \approx \mu_{mRNA}[1 + \frac{K_{off}K_m}{(K_{on} + K_{off})D_m}] = \mu_{mRNA}[1 + (1− < P_{on} >)\frac{K_m}{D_m}]
\]  

(4.22)

if we assume that the promoter variance term is at its maximum, implying that \(< P_{on} >\) is equal to \((1− < P_{on} >)\), then the expression simplifies to

\[
\sigma_{mRNA}^2 \approx \mu_{mRNA} + \mu_{mRNA}^2
\]  

(4.23)
4.7.2 Analysis of residual protein variance

Linear regression model between mean-independent variance and nucleosome occupancy

To determine the amount of variation in mean-independent variance explained by nucleosome occupancy at the promoter sequence, we built a linear model defined as

\[ \hat{\text{miv}} = \sum_{i=1}^{n} N_i \cdot \alpha_i \] (4.24)

where \( \text{miv} \) represents mean-independent variance, \( N_i \) represents the value of nucleosome occupancy of the \( i^{th} \) descriptive position used in the model, where \( n \) is the number of single-base positions in the promoter sequence that are used as input features for the model. We chose a linear model because of its simplicity and because we lacked detailed information on the promoter architecture of each analyzed gene, making a physically motivated sequence based model difficult to apply. Because the number of potential features (positions in the promoter sequence) is large, we selected the number of features to be used by the model by applying a forward regression strategy (for details see [171]). For each subset of features incorporated in the model, we computed a set of Monte Carlo simulations where we randomized nucleosome positions and MIV pairing 100 times. We computed the average \( r^2 \) of the Monte Carlo simulations and compared it to the real \( r^2 \) of the model using the same set of features. We chose the set of features that maximized the increment of \( r^2 \) of the real set from the average Monte Carlo simulations. Our final set includes 149 features, namely positions in the promoter sequence with associated values of nucleosome occupancy across
the dataset. To test for over-fitting, we performed a leave-one-out cross-validation strategy, where we predicted the results of 1 data-point after training the parameters on the rest of the dataset. This procedure was repeated for each data point and overall correlation was then computed. This strategy was implemented in the R programming language.

Datasets, source code and results can be accessed and downloaded at http://cgs.wustl.edu/~fvallania/5_noise_2011/5_noise_website/NOISE_Project_supporting_materials.html

4.8 Author Contributions

FLMV, RDM, and IM conceived the study. FLMV, MS, RDM, and BAC designed the analysis. FLMV, MS, RDM, and ZG performed the computational analysis. FLMV, RDM and IM designed the experiments. FLMV performed the experiments. FLMV, RDM, MS, and BAC wrote the final manuscript.

4.9 Acknowledgments

This work was funded by the NIH grant 5R01DA025744 and the Children’s Discovery Institute. We thank all the members of the Cohen and Mitra lab for helpful and constructive criticism. This work is dedicated to Natalina Vallania. No conflict of interests are declared.
The advent of the post-genomic era started with promises and doubts. Twelve years after its incipit, we still are plagued with important biological questions that the pure genomic approach has failed to address. However, one important factor has changed in recent years. During the progression of my thesis, I have witnessed the establishment of massive parallelized high-throughput sequencing as an everyday tool in the research community. What this meant, on a practical, everyday level, is the ability to sequence someone’s genome in the course of a single day in a single reaction, something that even one year before the start of my PhD was considered to be unthinkable.

What, however, I find most important about this revolution is not just its technological nature but its deep implication in changing everyone’s perspective on science. Because the accessibility of such technology is now widespread to the whole scientific community, the genomic approach can now be carefully directed to answer specific biological questions. In a way, next-generation sequencing has become the 21st century analogous of PCR, with the best results yet to come. Because of such untapped power at our disposal, these few years have seen a tremendous explosion of scientific creativity where new tools, methods, and protocols were created at a rapid pace. Over the course of my research in Rob’s lab, I
have been lucky enough to contribute to the creation of tools and protocols to leverage its potential in answering biological questions.

My first project involving the creation of SNPseeker and SPLINTER has given me a lot of rewards on both a personal and professional level. Although pooled DNA sequencing is no longer a necessity due to the declining costs of sequencing and the increased automation of library creation, there are still useful and important application of this technology. Pooling groups is still faster and more convenient for large screenings, but there are also many scenarios where samples are naturally heterogenous and therefore behave as pools. One important application is the problem of accurate detection of somatic mutations in cancer samples. As cancer undergoes phases of evolution and selection through interaction with the host immune response and chemotherapy, early mutation events initially occurring at low frequency can inform clinicians for an adequate tailored therapeutic regimen. Although different groups are now exploring single-molecule labeling strategies, detecting rare somatic mutations can be optimally done using appropriate computational strategies. As of now, I am currently involved in an effort to test SPLINTER as a potential tool for the detection of such mutations in the clinic. Another important application of this technology is monitoring of changes in viral populations in a single individual during the course of an infection. Notable examples of viruses that exist as populations or quasispecies are HCV and HIV. Understanding how their diversity can impact the host phenotype can lead to tremendous benefits in the design of vaccines and understanding of viral biology.
My second project (chronologically the last) involved the creation of a methylation reporter system. This project was the most challenging and more than any other has taught me invaluable lessons. The first is the value of persistence, as without it no important discoveries or breakthroughs will ever be made. Secondly, but not less importantly, the ability to be flexible and creative in order to reach the desired goal. Sometimes a specific approach may look great on paper but in reality it will not work, and the converse is also true. Finally, any step in the development of something novel always feeds on the passion and drive of realizing a vision of promise and potential. The potential for this project is in my opinion the highest compared to everything else I have ever made as it provides a tool that can be readily used to crack untapped biological knowledge. Importantly, it provides a experimental mean to decode the regulatory function of the genome in the context of DNA methylation, a biological process whose study has greatly benefited in recent times from the advent of high-throughput sequencing. While basic biology questions of DNA methylation can be readily dissected using this tool, more clinically relevant applications come in the context of cancer and tumorigenesis. Despite the fact that a significant fraction of cancers shows hyper-methylation of tumor suppressor genes, little is known of how this process is generated, and consequently pharmaceutical therapies targeting this phenomenon are still very a-specific. I believe in the potential application of this method to discover drivers of cancer hyper-methylation.

My third project involved a deep modeling effort in order to connect a large scale dataset to its biophysical roots. This project has challenged me in many ways and the most valuable
lesson that I have gathered from its completion was quite rewarding: any question or problem will always look difficult and unapproachable if faced from the wrong angle. However, if we take the time and the mental effort to look at the same issues from a different and novel perspective, new discoveries can be made but also difficult problems may in fact be very easy and interpretable. This project differs significantly from the previous ones also because it embodies the final steps in high-throughput biology and the next challenge in the post-genomic world. Once a large dataset is generated, the next step is to analyze it correctly and really understand the underlying biology. This requires not only computational prowess but also a clear understanding of how to mathematically model biological processes. The results of our work have important implications in the field of noise but also in single-cell biology. With the increases in resolution of our experimental assays, many groups have now been able to profile individual cells, showing their underlying and unexpected heterogeneity. Through the analysis of the relationship between mean and variance, we can understand what processes regulate the gene and speculate on their downstream implications in the context of phenotype.

I would like to conclude my thesis by stating an important realization that occurred over the course of my PhD. In my experience, many approach biology in two different ways: one group operates by focusing on a well-define, well-characterized model system and dissect it using well known and established approaches. The second group on the other hand, much smaller than the first, operates at the technical side by breaking new ground and new frontiers with new approaches and algorithms, but lacks focus and clarity as no biology is
necessarily learnt directly from their endeavors. I came to realize that the best possible work lies in between these two poles. The ability of attacking specific unsolved questions by creating novel algorithms, models, and novel tools, tailored for the problem, can be quite powerful and lead to great progress. Only by framing the right question in the right way, one can bring the theory into practice and create ways and change the perspective of the problem. But again, only by knowing all aspects of each step, especially the most technical ones, one can take a perfectly framed question and attack it in the real world. The interface between biological theory and the development of new technologies has the potential to turn dreams and thoughts into reality because it can bring a fresh unbiased perspective to attack established hard problems. I think this is by far one of the most important lessons learnt during my journey. It is not so much a path toward specialization but rather a path toward balance in connecting polar opposites.

I hope this thesis will delight you, interest you, or instigate new curiosity in you, dear reader. After all, curiosity is what got it written in the first place.

August 12th 2013, Saint Louis, MO, U.S.A.
6. Bibliography


[27] Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M


220


visualized experiments: JoVE, 0(64), 2012.


tions: high-throughput DNA extraction and molecular amplification from Guthrie cards.


Stuart Macgregor. Most pooling variation in array-based dna pooling is attributable to
array error rather than pool construction error. European Journal of Human Genetics,

Lira Mamanova, Alison J Coffey, Carol E Scott, Iwanka Kozarewa, Emily H Turner,
Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J Turner. Target-enrichment

John Stingl and Carlos Caldas. Molecular heterogeneity of breast carcinomas and the

Jeffrey E Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh,
Dominique Schneider, Richard E Lenski, and Jihyun F Kim. Genome evolution and
1247, 2009.

Millicent Higgins, Michael Province, Gerardo Heiss, John Eckfeldt, R Curtis Ellison,
Aaron R Folsom, DC Rao, J Michael Sprafka, and Roger Williams. Nhli family heart
study: objectives and design. American journal of epidemiology, 143(12):1219–1228,
1996.

Bruce A Vendt, Robert C McKinstry, William S Ball, Michael A Kraut, Fred W Prior,
Bruce Barton, James F Casella, and Michael R DeBaun. Silent cerebral infarct trans-
fusion (sit) trial imaging core: application of novel imaging information technology for


