

# Analysis of short tandem repeat variation in large scale resequencing data



Weldon W. Whitener

Wellcome Trust Sanger Institute

St. Edmunds College

University of Cambridge

A thesis submitted for the degree of

*Doctor of Philosophy*

September 2011

I would like to dedicate this thesis to my loving parents...

## Acknowledgements

None of this would of been possible without the love, support and guidance from the best parents a son could ever hope to have. Thank you Mom and Dad for always believing in me.

Now, where to begin?

There have been so many people who have contributed in making this document possible. Along the way, I've had more support than anyone could ask for, as well as, plenty of lucky breaks! I want to begin by thanking all my professors back at my *alma mater*, North Carolina State University. Specifically, I would like to thank Dr. Frank Abrams for always allowing me to use the 'big words', as well as, Dr. Elizabeth Lobo for giving me my first shot at scientific exploration.

I want to thank Dr. Mark van Dyke for allowing a scrappy biomedical engineer into one of the most amazing labs I've ever had the good fortune of being a part of. You are one of the most inspiring and friendly persons I've ever met and I am forever grateful for your support and guidance along the way. None of my subsequent success would of been possible had I not met you by chance at a career event my third year at North Carolina State University.

The next person I'd like to thank is someone whose guidance and support cannot be fully conveyed in writing. Thank you Jennie LaMonte for always going above and beyond. More than anything, you have been a great friend and mentor to me for the past six years and

I hope that we remain friends for many years to come.

The past four years have been some of the most challenging, yet thought provoking, years of my life. Richard Durbin accepted me into his group and helped me cultivate a probing and scientific outlook that has helped me answer some of the most difficult questions I've ever been posed. Richard's continued support, guidance and occasional knock on the head has given me a true confidence in myself that I will be forever grateful for. Thank you for teaching me to stand on my own two feet and giving me the tools to help me be successful no matter where life takes me.

This thesis would not of been possible without the help of countless people at the Sanger Institute and University of Cambridge. Thank you Leopold Parts and Aylwyn Scally for helping me come to grips with statistical modeling. You were the best labmates I could of asked for and the help you offered me will never be forgotten. Thank you Jim Stalker, Thomas Keane and the entire vertebrate sequencing group for making my life so much easier by doing a great job at curating all the sequencing data – an almost insurmountable task! Thank you Avril Coghlan for your continued collaboration throughout the duration of my thesis. Thank you Stijn van Dongen, Sergei Manakov, David Adams, Theo Whipp and Steve Russen for your support and friendship throughout my time at the Sanger Institute. I also wish to send a huge thank you to my favorite collaborator, David Knowles. Without David's masterful grasp of machine learning and diligence, most of the higher level analysis conducted in the latter chapters of this thesis would not of been possible. David even managed to make it a fun process – through his energy and good nature – which by no means is no small feat!

Lastly, I want to thank a person who, as much as my parents, has helped me become the the person I am today. Without my sister

Witnee's guidance, I would not of been able to achieve all that I have. She taught me how to conduct myself in a manner that has opened doors for me both socially and professionally. She also offered the best advice while keeping my spirits high. I could not of done this without her, and for that, I am forever grateful.

---

## Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This thesis does not exceed the length limit set by the Biology Degree Committee.

Weldon W. Whitener  
30 September 2011

## Abstract

The average eukaryotic genome contains many types of variation; from single nucleotide polymorphisms, small, medium and large insertions and deletions to copy number variation, translocations and inversions to name a few. The genome is also highly non-uniform, with some regions more variable than others. Tandem repeats are stretches of DNA comprised of a short motif repeated end-to-end multiple times. They are of interests to geneticists because they exhibit a high rate of length variation and are relatively frequent in the genome. However, until now they have been hard to assay using new sequencing technologies, which have revolutionized the study of other types of genetic variation. In this thesis, we address this deficit by developing methods to genotype short tandem repeats from shotgun short sequencing reads and applying them to human genome data.

To begin, I present a statistical model based on a Bayesian framework which uses Illumina paired end sequencing reads to determine the genotype of a diploid individual at a given short tandem repeat locus. This method is applied to all triplet tandem repeats (repeat motifs three bases in length) in the human genome for an individual sequenced deeply from multiple libraries as part of the 1000 Genomes project. We show that our method has good sensitivity and specificity for both homozygous and heterozygous indel genotypes measuring over three bp in length.

Next, we build upon the previous chapter by utilizing our model for genotyping across nine deeply sequenced individuals. We use the putative indel calls made in this data set to gain an understanding of

what factors of a tandem repeat have the largest effect on observing an indel at a given locus. We look at the effect that various measures of repeat length, repeat purity, GC content and tandem repeat motif have on triplet repeat variation. This analysis furthers our understanding of tandem repeat variation.

Lastly, we reformulate our individual genotyping model to take sequencing data from multiple, low sequence depth individuals in a population to understand the population distributions of variants at tandem repeat loci. This uses machine learning approaches including the expectation-maximization algorithm and Gibbs sampler, that help elucidate which loci show evidence of variation in the sample population, and allow us to explore the distribution of alternate alleles at a locus. As well as cataloguing variation efficiently, this allows us to examine a broader picture of the contribution the previously described factors have in influencing variation at a tandem repeat locus.



# Contents

|   |             |
|---|-------------|
| <b>Contents</b>   | <b>viii</b> |
| <b>List of Figures</b>  | <b>xiii</b> |
| <b>1 Introduction</b>   | <b>1</b>    |
| 1.1 New age of technology . . . . .   | 1           |
| 1.2 Sequencing technology and bioinformaticians . . . . .   | 2           |
| 1.3 Genomic variation . . . . .   | 3           |
| 1.4 Detecting small scale insertions and deletions . . . . .  | 4           |
| 1.5 Tandem Repeats . . . . .  | 6           |
| 1.6 Small scale insertions and deletions in tandem repeats . . . . .  | 10          |
| 1.6.1 Background significance of tandem repeat indels . . . . .   | 10          |
| 1.6.2 Detection of indels using paired end mapping information . . . . .                                    | 11          |
| 1.6.3 Relevance of an indel caller for short tandem repeats . . . . .                                       | 20          |
| 1.7 Ascertaining tandem repeat allele frequencies in large populations . . . . .                            | 23          |
| 1.8 Proposal . . . . .  | 26          |
| <b>2 Genotyping short tandem repeats using short paired end reads from two deeply sequenced individuals</b> | <b>28</b>   |
| 2.1 Locating tandem repeats in the human reference genome . . . . .   | 29          |
| 2.1.1 Translating NCBI build 36 coordinate to GRCh build 37 coordinates . . . . .                           | 29          |
| 2.2 Sources of sequence . . . . .   | 30          |
| 2.2.1 Individual NA12878 sequence . . . . .   | 30          |
| 2.2.1.1 Illumina read sequence data . . . . .   | 31          |

|         |   |    |
|---------|---|----|
| 2.2.1.2 | 454 sequence data . . . . .   | 31 |
| 2.2.1.3 | Capillary sequence data . . . . .   | 31 |
| 2.2.2   | Individual NA18507 sequence . . . . .   | 31 |
| 2.2.2.1 | Illumina read sequence data . . . . .   | 31 |
| 2.2.2.2 | Capillary sequence data . . . . .   | 31 |
| 2.3     | Mapping of paired end reads to the human reference genome . . .   | 32 |
| 2.4     | Determining the empirical distribution of a given library's mapped<br>paired end read separations (MPERS), $P(M)$ . . . . . | 32 |
| 2.4.1   | The empirical distribution of mapped paired separations<br>(MPERS) . . . . .  | 35 |
| 2.4.1.1 | Individual NA12878 . . . . .  | 35 |
| 2.4.1.2 | Individual NA18507 . . . . .  | 35 |
| 2.5     | Detecting indels in tandem repeat loci using long capillary reads<br>from the Trace Archive . . . . .                       | 36 |
| 2.6     | Detecting indels in tandem repeat loci using short read sequence<br>data . . . . .  | 38 |
| 2.6.1   | Background on indel detection using paired end sequence<br>data . . . . .   | 38 |
| 2.6.2   | The empirical distribution of MPERS for read pairs that<br>span a STR locus of a given length, $P_l(M)$ . . . . .           | 39 |
| 2.6.3   | Estimating the genotype of a tandem repeat locus . . . . .  | 45 |
| 2.6.3.1 | Rationale behind analysing MPERS distributions<br>to detect indels in STR loci . . . . .                                    | 45 |
| 2.6.4   | Prior Probabilities . . . . .   | 51 |
| 2.6.5   | Odds ratio and normalized posterior . . . . .   | 56 |
| 2.7     | Software . . . . .  | 57 |
| 2.8     | Simulations . . . . .   | 58 |
| 2.8.1   | Reference . . . . .   | 59 |
| 2.8.2   | Homozygous indel . . . . .  | 60 |
| 2.8.3   | Heterozygous with one reference allele . . . . .  | 66 |
| 2.8.4   | Heterozygous with no reference allele . . . . .   | 69 |
| 2.9     | Results on real data . . . . .  | 70 |
| 2.9.1   | Inferring genotypes at repeat loci in individual NA12878 .  | 70 |

|          |  |           |
|----------|--|-----------|
| 2.9.2    | Accuracy in inferring genotypes at repeat loci . . . . .                     | 72        |
| 2.9.2.1  | Validation data from capillary and 454 alignments . . . . .                  | 72        |
| 2.9.2.2  | Accuracy at homozygous reference loci . . . . .                              | 77        |
| 2.9.2.3  | Accuracy at homozygous indel loci . . . . .                                  | 78        |
| 2.9.2.4  | Accuracy at heterozygous loci . . . . .                                      | 78        |
| 2.9.3    | Comparison with MoDIL . . . . .  | 82        |
| 2.10     | Discussion . . . . .   | 84        |
| 2.10.1   | Specific adaptations for detecting indels in STR loci . . . . .              | 84        |
| 2.11     | Conclusion . . . . .   | 86        |
| <b>3</b> | <b>Factors influencing polymorphism in short tandem repeats</b>              | <b>87</b> |
| 3.1      | Sources of sequence . . . . .  | 88        |
| 3.1.1    | 1000 Genomes pilot trios . . . . .   | 88        |
| 3.1.1.1  | Sequencing statistics . . . . .  | 88        |
| 3.1.2    | Illumina Trio . . . . .  | 90        |
| 3.2      | MPERS distributions . . . . .  | 90        |
| 3.3      | Detecting indels in short tandem repeats . . . . .                           | 95        |
| 3.4      | Short tandem repeat criteria . . . . .                                       | 96        |
| 3.4.1    | STR metrics . . . . .  | 97        |
| 3.4.2    | Tandem repeat length in reference (reflen) . . . . .                         | 98        |
| 3.4.3    | Tandem repeat motif family (motif) . . . . .                                 | 98        |
| 3.4.4    | Purity of tandem repeat in reference . . . . .                               | 99        |
| 3.4.4.1  | Longest pure stretch (purls) . . . . .                                       | 100       |
| 3.4.4.2  | Percent match (purnew) . . . . .   | 100       |
| 3.4.5    | GC content in and around tandem repeat (GCref, GC100<br>and GOnly) . . . . . | 101       |
| 3.4.6    | Whether a tandem repeat is in a transcript (trans) . . . . .                 | 102       |
| 3.5      | Results . . . . .  | 102       |
| 3.5.1    | Modeling of factors . . . . .  | 104       |
| 3.5.1.1  | Bias in modeling of purity . . . . .   | 110       |
| 3.6      | Discussion . . . . .   | 111       |
| 3.6.1    | Sample family correlations . . . . .   | 112       |
| 3.6.2    | GC composition correlations . . . . .  | 112       |

|          |  |            |
|----------|--|------------|
| 3.6.3    | Motif correlations . . . . .   | 114        |
| 3.6.4    | Purity correlations . . . . .  | 115        |
| 3.6.5    | Further correlations: number of spanning read pairs, repeat<br>length in reference and located within a transcript . . . . . | 115        |
| 3.6.6    | Independent analysis and comparison of each factors' effect<br>on the magnitude of a variant at non-reference loci . . . . . | 116        |
| 3.6.6.1  | All variants . . . . .   | 116        |
| 3.6.6.2  | Independent analysis of insertions and deletions<br>compared to the reference . . . . .                                      | 117        |
| 3.7      | Conclusion . . . . .   | 118        |
| <b>4</b> | <b>Population based analysis of short tandem repeats</b>   | <b>119</b> |
| 4.1      | Low coverage individuals in the 1000 Genomes Project . . . . .   | 120        |
| 4.1.1    | Sources of sequence . . . . .  | 120        |
| 4.1.2    | Sequencing statistics . . . . .  | 120        |
| 4.1.3    | Population MPERS distributions . . . . .   | 122        |
| 4.2      | Modeling . . . . .   | 123        |
| 4.2.1    | Priors . . . . .   | 128        |
| 4.2.2    | EM algorithm . . . . .   | 128        |
| 4.2.3    | Gibbs sampling . . . . .   | 130        |
| 4.3      | Simulation . . . . .   | 131        |
| 4.3.1    | Simulation of MPERS for spanning read pairs . . . . .  | 132        |
| 4.3.2    | Simulation results . . . . .   | 134        |
| 4.3.2.1  | Reference allele frequency . . . . .   | 134        |
| 4.3.2.2  | Two and three allele population frequency alleles  | 137        |
| 4.3.3    | Simulation results comparisons . . . . .   | 141        |
| 4.3.4    | Test statistics . . . . .  | 147        |
| 4.3.4.1  | Entropy . . . . .  | 147        |
| 4.3.4.2  | Off reference/ $\pm 3$ bp . . . . .  | 148        |
| 4.3.5    | False discovery rate . . . . .   | 149        |
| 4.4      | Results . . . . .  | 154        |
| 4.5      | Discussion . . . . .   | 155        |
| 4.5.1    | Factors . . . . .  | 156        |

|          |   |            |
|----------|---|------------|
| 4.6      | Conclusion . . . . .                                      | 162        |
| <b>5</b> | <b>Conclusions</b>  | <b>164</b> |
| 5.1      | Conclusions, discussion and future work . . . . .         | 164        |
| 5.1.1    | Modeling variation in STRs . . . . .                      | 164        |
| 5.1.1.1  | Future work . . . . .                                     | 165        |
| 5.1.2    | Characterizing STR variation . . . . .                    | 166        |
| 5.1.2.1  | Future work . . . . .                                     | 166        |
| 5.1.3    | Modeling STR loci in large population data sets . . . . . | 167        |
| 5.1.3.1  | Future work . . . . .                                     | 167        |
|          | <b>References</b>   | <b>170</b> |

# List of Figures

|     |   |    |
|-----|---|----|
| 1.1 | Histogram of number of loci of each length across the human genome with loci longer than 150 bp binned in the last bin. . . . .   | 9  |
| 1.2 | The underlying paired end sequencing methodology used to detect structural variation by fosmid pairing. . . . .   | 15 |
| 1.3 | Example of a homozygous (1.3a) and heterozygous (1.3b) deletion with the observed distribution of mapped distances shown in gray. . . . .   | 19 |
| 1.4 | Graph of expected number of spanning reads (physical coverage) and reads that extend across various genomic lengths at base pair coverages of 10, 15 and 20x. . . . .                               | 23 |
| 1.5 | Map of populations in 1000 Genomes Project Phase 1 build. . . . .   | 25 |
| 2.1 | Four of the various mapping scenarios related to paired end reads. . . . .  | 33 |
| 2.2 | Graphic of mapped paired end read alignments of an individual whose locus matches the reference (top) and whose locus contains a deletion in respect to the reference (bottom). . . . .             | 40 |
| 2.3 | Mapped paired end reads sequenced from an individual whose reads align to both the reference repeat length (top) or to a deletion in the repeat tract in respect to the reference (bottom). . . . . | 41 |
| 2.4 | Empirical distributions of MPERS for individual NA12878 library g1k-sc-NA12878-CEU-1. . . . .   | 42 |
| 2.5 | Simulation results of the number of spanning (a) and hanging (b) reads across different coverages and repeat lengths from a constant fragment length library. . . . .                               | 43 |
| 2.6 | Cartoon representation of actual mapping positions of two paired end reads across a poly-A repeat of length 20 bp. . . . .  | 44 |

|      |  |    |
|------|--|----|
| 2.7  | Heatmap of likelihoods at a selected repeat locus of length 60 bp from a simulated homozygous reference genotype with average base pair coverage 15x. . . . .  | 49 |
| 2.8  | Prior probability distribution of haploid indel calls made in individual NA18507 from capillary reads. . . . .   | 52 |
| 2.9  | Symmetric prior probability distribution of haploid indel calls made in individual NA18507 from capillary reads. . . . .   | 54 |
| 2.10 | Heat map of the estimated prior probabilities for the varying genotypes of a triplet repeat locus in an individual (shown in log space). . . . .   | 55 |
| 2.11 | Histogram of spanning paired end read separations and MPERS distribution for a reference genotype simulation. . . . .  | 60 |
| 2.12 | Histogram of spanning paired end read separations across an individual whose repeat length is 21 bp shorter than that in the reference graphed against the MPERS distribution for a reference length genotype. . . . . | 62 |
| 2.13 | Histogram of spanning paired end reads across an individual whose two copies differ in length from the reference graphed against the MPERS distribution for a reference length genotype. . . . .                       | 69 |
| 2.14 | Histogram for loci containing a given number of spanning paired end reads for every triplet repeat loci in individual NA12878. . . . .   | 72 |
| 2.15 | Samtools tview of a 454 alignment for an unambiguously genotyped locus. . . . .  | 75 |
| 2.16 | Samtools tview of a 454 alignment for an inconclusive genotyped locus. . . . .   | 76 |
| 2.17 | Plot of the 454 indel genotypes when our method called a reference genotype, {0, 0}. . . . .   | 77 |
| 2.18 | Comparison of true homozygous indel genotypes as called from 454 sequence to that of our method's calls at these loci. . . . .   | 79 |
| 2.19 | Join plot comparison of actual genotype (red dot) compared to the genotype called by our method (blue dot). . . . .  | 81 |
| 2.20 | Histogram of differences in proximal allele lengths between genotype calls made by 454 and our method. . . . .   | 82 |

|      |   |     |
|------|---|-----|
| 2.21 | Distribution of the MPERS for two separate libraries for sequenced individual NA12878. . . . .  | 86  |
| 3.1  | Distributions of each library in the nine individuals from the three trios data set. . . . .  | 91  |
| 3.2  | Graph of coefficients determined by full logistic regression of factors giving contradictory results because of confounding between correlated factors. . . . .   | 106 |
| 3.3  | Bar graph of absolute values of coefficients from a logistic linear model for a STR being non-reference. . . . .  | 107 |
| 3.4  | Bar graph of absolute values of coefficients from a linear model for the magnitude of an indel at variant STR loci. . . . .   | 108 |
| 3.5  | Bar graph of absolute values of coefficients from a linear model for the magnitude of an insertion at variant STR loci. . . . .   | 109 |
| 3.6  | Bar graph of absolute values of coefficients from a linear model for the magnitude of a deletion at variant STR loci. . . . .   | 109 |
| 3.7  | Boxplot of repeat purity across varying repeat lengths. . . . .   | 111 |
| 3.8  | Boxplot of differences in GOnly and GCref at a locus binned by the number of observed spanning read pairs at a locus. . . . .   | 114 |
| 4.1  | Plot of MPERS distributions for every library in the 1000 Genomes Project data set. . . . .   | 123 |
| 4.2  | Plot of MPERS distributions whose mean of each library is arbitrarily set at zero. . . . .  | 124 |
| 4.3  | Plots of the raw MPERS for each of the fourteen populations in the 1000 Genomes Project data set. . . . .   | 126 |
| 4.4  | Allele frequency distribution predictions for the EM algorithm (blue bars) in 14 simulated loci in chromosome 20 from an underlying allele frequency distribution comprised solely reference alleles based on a CHS population (red bars). . . . .                          | 135 |
| 4.5  | Allele frequency distribution predictions of alleles for the Gibbs sampler algorithm (blue bars) in 14 simulated loci in chromosome 20 from an underlying allele frequency distribution comprised solely of reference alleles based on a CHS population (red bars). . . . . | 136 |



|      |   |     |
|------|---|-----|
| 4.6  | Allele frequency distribution predictions of alleles for the EM algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency distribution of $\pm 9$ bp each at a 0.5 frequency (red bars) based on a CLM population. . . . .  | 138 |
| 4.7  | Allele frequency distribution predictions of alleles for the Gibbs sample algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of $\pm 9$ bp each at a 0.5 frequency (red bars) in a CLM population. . . . .   | 139 |
| 4.8  | Allele frequency distribution predictions of alleles for the EM algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of 0.45 at both -12 bp deletion and 9 bp insertion alleles and a 0.1 frequency at the reference allele (red bars) based on a CLM population. . . . .            | 140 |
| 4.9  | Allele frequency distribution predictions of alleles for the Gibbs sampler algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of 0.45 at both -12 bp deletion and 9 bp insertion alleles and a 0.1 frequency at the reference allele (red bars) based on a CLM population. . . . . | 141 |
| 4.10 | Averages of allele frequency distributions (blue bars) across chromosome 20 for three simulation scenarios (red bars) for the EM algorithm. . . . .   | 142 |
| 4.11 | Averages of allele frequency distributions (blue bars) across chromosome 20 for three simulation scenarios (red bars) for Gibbs sampling algorithm. . . . .   | 142 |
| 4.12 | Comparison of the EM and Gibbs sampler algorithms for a reference allele frequency distribution. . . . .  | 144 |
| 4.13 | Comparison of the EM and Gibbs sampler algorithms for a two allele frequency simulation. . . . .  | 145 |
| 4.14 | Comparison of the EM and Gibbs sampler algorithms for a three allele frequency simulation. . . . .  | 146 |
| 4.15 | Plot of FDR versus true calls for the ASW population for triplet repeat loci on chromosome 20. . . . .  | 151 |

|  |     |
|--|-----|
| 4.16 Plot of FDR versus true calls for the MXL population for triplet repeat loci on chromosome 20. . . . .  | 152 |
| 4.17 Plot of FDR versus true calls for the PUR population for triplet repeat loci on chromosome 20. . . . .  | 153 |
| 4.18 Venn diagram of intersection of significant loci called by entropy and off $\pm 3$ bp. . . . .  | 155 |
| 4.19 Bar graph of absolute values of coefficients from logistic linear model on whether a locus's entropy value is significant against various factors. . . . .        | 157 |
| 4.20 Bar graph of absolute values of coefficients from logistic linear model on whether a locus's off $\pm 3$ bp value is significant against various factors. . . . . | 158 |
| 4.21 Bar graph of absolute values of coefficients from linear model of significant entropy loci values and the various explanatory factors.                            | 160 |
| 4.22 Bar graph of absolute values of coefficients from linear model of significant off $\pm 3$ bp loci values and the various explanatory factors.                     | 161 |