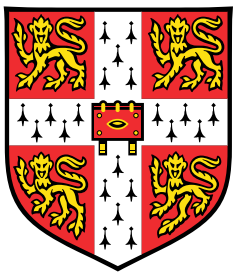
Germline mutation in rare disease



Joanna Kaplanis

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College September 2020

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text. This dissertation does not exceed the prescribed world limit set out by the Degree Committee for the Faculty of Biology.

Joanna Kaplanis
September 2020

Acknowledgements

I would like to start by thanking my supervisor, Matt Hurles. I feel very thankful to have had the opportunity to work on some truly exciting projects in the last few years. I have learnt as much about the scientific process as science itself from you, especially in your ability to balance your remarkable scientific intuition with evidence and trust in the data. Your ability to notice when the slightest detail in a figure is not quite right has set me back on the right track several times. Aside from your scientific support, I am especially grateful for your kindness and encouragement when I made the decision to have a child in the middle of my PhD.

I would like to thank everyone involved in the DDD project for your work in creating such an important resource and for all of the scientific input from meeting discussions. I am especially grateful to all of the children and their families who have participated in DDD as well as those who have contributed to other cohorts included in this work. Thank you to the Wellcome Trust for funding this work and to Annabel Smith, Christina Hedberg-Delouka and the committee of graduate studies for keeping the Sanger PhD programme running smoothly.

Thank you to the members of Team 29, past and present, for engaging scientific discussions, fun lunchtime chats and all of that delicious cake. I am especially grateful to Kaitlin who was a wonderful scientific teammate and always made time for my questions; I will miss our endless to-do lists. A special thanks also to Eugene, for always being willing to help and advise. Thank you to Carol Dunbar for handing all of the logistics of Team 29. Thank you also to everyone beyond Team 29 who has made the Sanger a fun place to be these last 5 years especially to Eva, Sophie, Alex, Fernando, the members of the now defunct breakfast club and the free biscuits at HumGen tea.

Lastly, I want to thank my family for their unwavering support in the last few years. To Paddy, for moving to Cambridge with me and for your enthusiasm and belief in me. To Theodora, my other thesis chapter, for your smiles at the end of the day and helping me keep everything in perspective. To my brother, who has always inspired me and managed to start and finish a PhD within mine. Finally to my parents, who never fail to pick up the phone, for their love, encouragement and patience. I would never have been able to do this without you, thank you for all of the opportunities you have given me.

Abstract

Germline mutation is the ultimate source of evolutionary change and disease-causing variants. Understanding the rates and patterns of human mutation can help us learn about their molecular origins, uncover our evolutionary history and improve our ability to identify the genetic causes of human disease. With the advent of exome and genome data sets of parent-offspring trios there is an unprecedented opportunity to characterise mutations at an individual level and to harness the increasing sample sizes to identify disease-causing mutations. The goal of this thesis is to understand sources of variation in germline mutation and the contribution of these mutations to rare developmental disorders. These sources of variation encompass types of mutations that have been previously underrepresented in genetic research as well as individual mutation rates and spectra across individuals and parental origin. These analyses fall into three distinct projects.

My first project in this dissertation focuses on the mutational origins and pathogenic impact of multi-nucleotide variants (MNVs). These are variants that fall within 20 base pairs of each other and are frequently misannotated in variant-calling pipelines. Using data from the Deciphering Developmental Disorders (DDD) study, I explore the pathogenicity of this type of variant and found that MNVs in protein-coding sequences can be more pathogenic than a single nucleotide variant even when the MNV falls within a single codon. I also estimate the MNV mutation rate, explore the mutational spectra of these variants and describe the contribution of *de novo* MNVs to severe developmental disorders.

The next project focuses on identifying and characterising germline hypermutators. Using sequencing data from the DDD and 100,000 Genomes Project datasets across ~20,000 parent-offspring trios, I identified fifteen children with an unusually large number of *de novo* mutations. Eight of these appear to be due to a paternal hypermutator. I describe analyses to try and identify a genetic cause for this hypermutation. For two of the individuals, I found rare homozygous paternal variants that fell into two different DNA repair genes and are the likely cause. I also explore whether variants in DNA repair genes more generally impact germline mutation rates. First by examining a well characterised cancer somatic mutator gene and second by using a broader approach across all DNA repair genes. Using the large resource of DNMs called in the 100,0000 Genomes Project dataset, I also estimate what

fraction of variance in germline mutation rate can be explained by hypermutation as well as by parental age.

In my final project, I describe analyses of *de novo* mutations in a cohort of individuals with developmental disorders (DDs). *De novo* mutations are a major cause of DDs however known genes only account for a minority of the observed excess of these mutations. Here I develop a statistical framework and apply this on *de novo* mutations from ~31,000 exome sequenced parent offspring trios from the DDD study pooled with trios from GeneDx, a US-based genetic diagnostic company, and trios from Radboud University Medical Center (RUMC). I identify 28 genes that were not previously robustly associated with DDs and explore how these genes differ from those that were previously known. I also develop a model-based approach to explore the likely properties of currently undiscovered genes which can inform future directions in the field.

Collectively, these results reveal important insights into sources of variation in germline mutation rates as well as in mutation type. This can inform how germline mutations arise and further improve our ability to assess their contribution to rare genetic disease.

Table of contents

List of figures **xiii**

List of tables **xv**

1 Introduction **1**
1.1 Motivation 1
1.2 Mutational processes 1
1.2.1 Types of mutation 1
1.2.2 Origins of mutation 2
1.2.3 DNA damage tolerance and repair 2
1.3 Estimating human germline mutation rates 4
1.3.1 Early strategies to detect mutations 4
1.3.2 Methods for recent direct mutation rate estimates 5
1.4 Variation in the human germline mutation rate 6
1.4.1 Variation within genomes 6
1.4.2 Individual level variation 8
1.4.3 Population level variation 10
1.5 *De novo* mutations in human disease 11
1.5.1 Modes of inheritance 11
1.5.2 Historical context 12
1.5.3 Developmental disorders 12
1.6 Outline of dissertation 16

2 Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations **19**
2.1 Introduction 19
2.1.1 Chapter overview 20
2.1.2 Publication and contributions 20

2.2 Methods . . . 21
2.2.1 Variant and *De Novo* calling in DDD . . . 21
2.2.2 Estimating the MNV mutation rate . . . 21
2.2.3 Estimating the enrichment of *de novo* MNVs . . . 22
2.2.4 Estimating the number of clinically reported MNVs . . . 22
2.3 Results . . . 23
2.3.1 Identifying and categorising MNVs . . . 23
2.3.2 Analysis of MNV mutational spectra . . . 26
2.3.3 Misannotation of MNVs . . . 30
2.3.4 Functional Consequences of MNVs . . . 31
2.3.5 MNVs can create a missense change with a larger physico-chemical distance compared to missense SNVs . . . 32
2.3.6 Missense MNVs are on average more damaging than missense SNVs 32
2.3.7 Estimation of the MNV mutation rate . . . 34
2.3.8 Contribution of *de novo* MNVs to developmental disorders . . . 35
2.3.9 Clinically reported MNVs in DD-associated genes . . . 38
2.3.10 MNV mutator phenotype . . . 38
2.4 Discussion . . . 39

3 Identifying and characterising germline hypermutators 41
3.1 Introduction . . . 41
3.1.1 Chapter Overview . . . 43
3.1.2 Contributions . . . 44
3.2 Methods . . . 44
3.2.1 *De novo* calling and filtering in paternal *MBD4* PTV carriers . . . 44
3.2.2 DNM filtering in 100,000 Genomes Project . . . 46
3.2.3 DNM filtering for possible DDD hypermutated individuals . . . 47
3.2.4 Parental phasing of *de novo* mutations . . . 48
3.2.5 Analysis of effect of parental age on germline mutation rate . . . 48
3.2.6 Identifying hypermutation in 100kGP . . . 48
3.2.7 Extraction of mutational signatures . . . 49
3.2.8 Defining set of genes involved in DNA repair . . . 49
3.2.9 Estimating the fraction of variance explained . . . 49
3.2.10 Analysis of contribution of rare variants in DNA repair genes . . . 51
3.3 Results . . . 52
3.3.1 Examining the effect of PTVs in *MBD4* on germline mutation rate . 52
3.3.2 Identifying germline hypermutators . . . 53

3.3.3 Characterising hypermutation in 15 individuals 60
3.3.4 Fraction of germline mutation rate variation explained 67
3.4 Discussion 70

4 Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders 77
4.1 Introduction 77
4.1.1 Chapter overview 78
4.1.2 Publication and contributions 79
4.2 Methods 79
4.2.1 Sample collection and individual quality control 79
4.2.2 Definition of diagnostic lists 81
4.2.3 Joint quality control of datasets 82
4.2.4 DeNovoWEST framework 87
4.2.5 Functional similarity between new and known genes 92
4.2.6 DNM enrichment in non-significant genes 93
4.2.7 Modelling remaining PTV DNM burden 95
4.2.8 Expression in fetal brain 97
4.3 Results 97
4.3.1 Improved statistical enrichment test identifies 300 significant DD-associated genes 97
4.3.2 Characteristics of the novel DD-associated genes and disorders . . . 102
4.3.3 Recurrent mutations and potential new germline selection genes . . 103
4.3.4 Evidence for incomplete penetrance and pre/perinatal death 105
4.3.5 Modelling reveals hundreds of DD genes remain to be discovered . 110
4.4 Discussion 111

5 Discussion 115
5.1 Summary of Findings 115
5.2 Limitations and future directions 117
5.3 Concluding remarks 120

References 123

List of figures

1.1 *De novo* mutation mechanisms and genome level variation 7
1.2 Embryogenesis and gametogenesis 9
1.3 Phenotypes in the DDD study 14

2.1 Properties of MNVs 25
2.2 Mutational spectra of *de novo* MNVs 26
2.3 Mutational Spectra of MNVs 27
2.4 Mutational spectra of adjacent trinucleotide MNVs 29
2.5 Classification of intra-codon MNV missense mutations 31
2.6 Quantifying the pathogenicity of MNVs 33
2.7 Sensitivity of MNV enrichment analysis to MNV mutation rate estimates . 36
2.8 Enrichment of *de novo* MNVs in DDD study 37

3.1 Mutational Spectra of DNMs in paternal *MBD4* paternal PTV carriers . . . 53
3.2 Mutational Spectra of all DNMS called in the 100kGP cohort 54
3.3 Distribution of number of *de novo* SNVs and InDels per person 54
3.4 Parental age and the number of DNMs 55
3.5 Proportion of paternally phased DNMs against paternal age 56
3.6 Mutational spectra and signatures for maternal vs paternal DNMs across 100kGP cohort 57
3.7 Loss of transmitted allele example leading to false positive DNMs 59
3.8 Enrichment of mutation type for hypermutated individuals 61
3.9 Mutational signature decomposition for DNMs in hypermutated individuals 62
3.10 Transcriptional strand bias for DNMs in hypermutated individuals 63
3.11 Position of paternal *MPG* missense variant in the context of the protein . . . 65
3.12 Distribution of variant allele fraction for DNMs in hypermutated individuals 67
3.13 Impact of rare variants in DNA repair genes on germline mutation rate . . . 69
3.14 Mutational spectra of DNMs from hypermutated individuals (A) 74

3.15 Mutational spectra of DNMs from hypermutated individuals (B) 75

4.1 Variant allele fraction of DNMs across cohorts pre and post filtering 86
4.2 Overview of DeNovoWEST method 88
4.3 Enrichment of consequence classes and corresponding PPV weights used for DeNovoWEST test 91
4.4 Comparison of cohorts 98
4.5 Results of DeNovoWEST analysis 100
4.6 Quality Control analyses for DeNovoWEST 101
4.7 Functional properties and mechanisms of novel genes 103
4.8 DNM enrichment in non-significant genes 106
4.9 Comparison of proportion of genes expressed in fetal brain 106
4.10 Impact of penetrance on power 107
4.11 Impact of pre/perinatal death on power 109
4.12 Exploring the remaining number of DD genes 110
4.13 Likelihood model for missense DNM enrichment 111

List of tables

1.1 Distribution of family types within the rare disease arm of the 100,000 Genomes Project 15
1.2 Distribution of disease types in the rare disease arm of the 100,000 Genomes Project 15

2.1 Numbers of MNVs in each category type 28
2.2 Numbers and proportions of consequence types for MNVs within same codon 30
2.3 *De Novo* MNVs that fall in genes associated with developmental disorders . 35

3.1 Properties of hypermutated individuals 58
3.2 Possible paternal mutator variants 63
3.3 Impact of parental rare variants in DNA repair genes on germline mutation rate 70

4.1 Table showing the GO terms selected as being relevant to consensus DD genes 94
4.2 Recurrent Mutations 104