Uniparental disomy and mosaic structural variation in developmental disorders

Daniel Alexander King, M.D.

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

November 2015

DECLARATION OF ORIGINALITY

This dissertation describes my research performed between January 2012 and June 2015 under the supervision of Dr. Matthew Hurles. This work is my own and includes nothing which is the outcome of work done in collaboration except where explicitly stated in the text. It has not been previously submitted for any qualification, and it complies with the prescribed word limit set by the Degree Committee for the Faculty of Biology.

ABSTRACT

Developmental disorders (DDs) are diseases of impaired childhood development and include congenital anomalies, neurodevelopmental disorders, and abnormalities in growth and behaviour. Determining the genetic causes underlying DD is a major goal of contemporary medical research and the recent entrance of exome sequencing data into the rare-disease field has been transformative in uncovering the importance of *de novo* point mutations as a major source of DD-associated mutations. Recent efforts have successfully harnessed exome sequencing data to detect constitutive copy-number variation, a form of large-scale structural abnormality. However, at the inception of my doctoral work, no software tools had yet been developed to identify, from exome sequencing data, uniparental disomy (UPD), a form of copy-neutral variation, nor large-scale ('structural') abnormalities, which have long been implicated as important contributors to DD. The research underlying this work aimed to fill this void.

This dissertation addresses the development of new software tools, UP*D*io and MrMosaic, which have extended the diagnostic reach of sequencing data to identify UPD and structural mosaicism, and have been made freely available. Simulation analyses show that these tools can detect the large-scale abnormalities identified by karyotyping or microarray in standard clinical testing. Implementation on nearly 5,000 children with undiagnosed diseases demonstrated that UPD and structural mosaicism are enriched in children with developmental disorders compared with healthy children and suggested that most of the detected abnormalities are likely to be pathogenic. Investigation of the clinical impact of the detected events identified several disease-causing mechanisms, including UPD-associated imprinting and recessive diseases, and genomic disorders associated with large mosaic deletions and duplications.

The five chapters of this dissertation are: 1) an introduction, to describe the context of this doctoral work; 2) a description of UP*D*io, a new method for detecting uniparental disomy from exome trio data; 3) a burden analysis of mosaic structural variation and the clinical consequences of mosaic structural variants found in children with DD; 4) a description of MrMosaic, a new method for the detection of mosaic structural variation using next generation sequence data; and lastly, 5) a discussion that recapitulates the results of these analyses, describes their limitations, and considers future directions.

ACKNOWLEDGEMENTS

Great thanks are owed to many. In chronological order, my parents deserve credit, overdue and too understated, for teaching me how to walk, talk, and create (thanks mom) and inspiring my passions in science (thanks dad). Josh and Jason provided a sibling rivalry that was a great cure of laziness and I look forward to one day raising our families together.

One of my first patients in medical school was a middle-aged dad, experiencing new, severe headaches, and died from glioblastoma a few months later. Thank you for allowing me to follow your case as the experience led me to pursue medical research at the NIH. There, Dr. Les Biesecker nurtured my interests in programming, advocated for me, and encouraged me to become a physician scientist. Jamie Teer was a patient and forgiving scripting coach and Larry Singh taught me that statistics is really not so boring at all.

I am sincerely grateful to my Ph.D. supervisor, Matt Hurles, with whose intuition this dissertation research has greatly benefited, and with whose guidance the pleasure of research was conspicuous and reinforcing. One of my favourite phrases of yours is ‘it’s a testable hypothesis’, a maxim demonstrating your perpetual loyalty to evidence and methodical experimentation. I am grateful to have pursued with your mentorship and collaboration science as it evolved in front of us, yielding findings often unexpected and fascinating, and hope that such discovery can continue in partnership for many years to come.

Thank you to everyone in Team 29 for putting up with my cheeky birthday surprises and for your friendship. Saeed, Manu and Raheleh, thanks for being good coffee-break companions. Ray, Alejandro, Jeremy, and Tom, I enjoyed our brainstorming sessions. Thank you to the DDD laboratory and informatics teams for performing so many upstream analyses. The work presented here was not possible without the participation of thousands of children and their families; thank you for joining DDD. Wellcome Trust kindly funded this research. Annabel & Christina, and Carol, much appreciation for organising the logistics for all my academic activities.

Thank you Tanya for always being there for me, and for enabling my life inside and outside the lab these Ph.D. years to be adventurous and fulfilling.

CONTENTS

PUBLICATIONS

King, D.A., Sifrim, A.S., Fitzgerald, T.W. & Hurles, M.E. Detection of structural mosaicism from targeted and whole-genome sequencing data. *In review.*

King, D.A., Jones, W.D., Crow, Y.J., Dominiczak, A.F., *et al.* Mosaic structural variation in children with developmental disorders. *Human molecular genetics* **24**, 2733-2745 (2015).

Carvalho, C.M.B., Pfundt, R., King, D.A., Lindsay, S.J., *et al.* Absence of heterozygosity due to template switching during replicative rearrangements. *The American Journal of Human Genetics* **96**, 555-564 (2015).

Akawi, N., McRae, J., Ansari., M, Balasubramanian, M., et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature genetics* **47**, 1363-1369 (2015).

Wright, C.F., Fitzgerald, T.W., Jones W.D., McRae, J.F., *et al*. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet* **385**, 1305-1314 (2015).

TDDD Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **12**, 223-228 (2014).

King, D.A., Fitzgerald, T.W., Miller, R., Canham, N., *et al*. A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome research* **24**, 673-687 (2014).

LIST OF TABLES

LIST OF FIGURES

TABLE OF ABBREVIATIONS AND ACRONYMS

1000G	1000 Genomes study
aCGH	array comparative genomic hybridisation
ALSPAC	Avon Longitudinal Study of Parents and Children
AUC	area under the curve
BAC	bacterial artificial chromosome
BAF	b allele frequency
B_{dev}	BAF deviation
C_{dev}	copy-number deviation
CNV	copy number variation
DD	developmental disorders
DDD	Deciphering Developmental Disorders
DECIPHER	DatabasE of genomiC varIation & Phenotype in Humans using Ensembl Resources
FISH	fluorescent *in situ* hybridisation
GADA	genome alteration detection analysis (software)
GRC	genome reference consortium
GRCh37	genome reference consortium, human genome reference 37
GWAS	genome-wide association studies
HGP	human genome project
HMM	hidden Markov model
HPO	human phenotype ontology
indels	insertions and deletions
LOH	loss of heterozygosity
LRR	log r ratio
MAD	mosaic alteration detection
MAF	minor allele frequency

Mb	megabases
MrMosaic	Mosaic Rearrangements by Merging Of Sequenced Alleles using their Identity and Counts
OMIM	online Mendelian inheritance in man
QC	quality control
RFLP	restriction fragment length polymorphism
ROH	region of homozygosity
SFHS	Scottish Family Health Study
SNP	single nucleotide polymorphism
SV	structural variation
TEDS	Twins Early Development Study
UK10K	United Kingdom 10,000 Genomes Project
UPD	uniparental disomy
VCF	variant call format
WES	whole-exome sequencing
WGS	whole-genome sequencing