ARE WE MERELY HOPEFUL MONSTERS: THE ROLE OF GENE AND GENOME DUPLICATION IN EVOLUTION AND GENETIC DISEASE

A THESIS

SUBMITTED TO THE UNIVERSITY OF MANCHESTER

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY (PHD)

IN THE FACULTY OF BIOLOGY, MEDICINE AND HEALTH.

ALEXANDRA C. MARTIN-GEARY

2022

SCHOOL OF BIOLOGICAL SCIENCES

DIVISION OF EVOLUTION AND GENOMIC SCIENCES

Detail from Hieronymus Bosch's Garden of Earthly Delights triptych

(Bosch, 1500)

CONTENTS

36,773 Words

(Excluding appendices)

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In 1933 the German geneticist Richard Goldschmidt presented his theory that major evolutionary events have been driven, not by the slow accumulation of small variants, in line with Darwinian thought, but instead large-scale mutations which he called ‘macromutations’. His theory was called “The hopeful monsters hypothesis”, as any species arising from such an event would often result in evolutionary dead end. This theory was, however, by and large rejected due to its stark contrast with that of Darwin’s theory of evolution.

86 years on, thanks to tremendous advancements in genomic technologies and evolutionary analysis, we are aware of the existence of a myriad of what could be considered ‘macromutations’ (large-scale duplications) in the human genome, and the significant role they have played in our evolution. Whilst beneficial in the process of shaping who we are as a species, these large-scale mutations are not always benevolent, having been found to have significant links with a plethora of human diseases.

Using evolutionary data, alongside important biological features of human genes, we explore the links between large-scale mutations, particularly whole genome duplications, introgressed genes, and heritable, parasitic, and viral disease.

We find that there are strong links between ancient fragility, exposed by divergence between duplicated genes, and heritable disease; that co-evolution between humans and the soil borne helminth parasite *Trichuris trichura* has been facilitated by strong selective pressures exerted on the ancient genes of the TGFβ superfamily; and that viruses target ancient and important gene families and functions.

It is clear that large-scale mutations, in particular gene and genome duplications, have provided the genetic redundancy that contributed to the evolution of the complex species we are today. However, given the fragility and propensity of genes of this kind to be associated with disease of all varieties, our future evolvability, and ability to adapt is in question. We must, therefore, readdress Goldschmidt's theory as a question – Are we merely the descendants of hopeful monsters?

LAY ABSTRACT

In 1933 the German scientist Richard Goldschmidt proposed that, major transitions in evolution, instead of being a result of small changes in the DNA as the followers of Darwin thought, are in fact the result of much larger ones. He called this theory “The hopeful monsters hypothesis” as it was very likely that most species that began in this way would die out.

We now know that these large mutations have occurred repeatedly over the course of human genetic history and have allowed us to evolve into the complex species we are today. However, the more we are able to find out, the more we find that these large mutations also have strong links to many different diseases.

We explored the evolutionary basis of some of these links, looking at inherited genetic disease, parasitic diseases with a genetic component, and the way that viruses target human genes. We found that commonly, disease of all kinds are linked to very ancient, duplicated genes with important properties, which, due to how critical they are, are resistant to change.

We can now see that Goldschmidt’s theory was right, large mutations have led to a lot of the major changes in the human genetic past, however, given that we are now unable to change without damaging these genes we must readdress Goldschmidt’s theory and ask –are we merely descendants of hopeful monsters?

DECLARATION

The University of Manchester

PhD by Alternative Format Candidate Declaration

Candidate Name: Alexandra C. Martin-Geary

Faculty: Faculty of Biology, medicine and health

Thesis Title: Are we merely hopeful monsters: The role of gene and genome duplication in evolution and genetic disease

Declaration to be completed by the candidate:

The *in vivo* experiments which provided the basis for the investigation detailed in chapter three "Keep your friends close and your enemies closer: The co-evolution of *Trichuris trichura* and the human TGFβ superfamily", was performed by Adefunke Ogunkanbi, and has been submitted in support of her PhD application. All other work was performed by Alexandra Martin-Geary.

The supplementary paper provided in appendix i "Integration of large-scale genomic data sources with evolutionary information reveals novel genetic loci for congenital heart disease" was written by and has been submitted in support of Elisavet Fotiou's PhD application. Miss

Fotiou conducted and completed the majority of the work detailed in the paper, with whole-genome data generation, assignment of paralog status, and some data manipulation performed by Alexandra Martin-Geary.

The supplementary paper provided in appendix ii “There is no place like home: demographic analysis of the undulate ray, *Raja undulata,* using non-invasive mark-recapture and DNA analysis” was written by, and has been submitted in support of Samantha Hook’s PhD application. Miss Hook conducted and completed the majority of the work detailed in the paper, with network construction and analysis, and statistical analysis of interactions within the Ray population performed by Alexandra Martin-Geary.

Signed:

Date:

COPYRIGHT STATEMENT

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=2442 0), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses

DEDICATION

The initial submission of this work was dedicated to my children,

Bramwell and Goblin, and also to Diego and Carmen.

To Bramwell, for caring about me so deeply, and striving to help me succeed.

To Goblin, for not making writing a thesis during pregnancy as difficult as he could have!

And Carmen and Diego who were both *very* good girls.

This complete and revised edition is wholly dedicated to Benjamin Stamp

Ben was a member of our research group, fellow PhD student, and friend. Ben was someone who, as both scientists and individuals we should all aspire to be. Kind, generous of heart, genuine, and full of awe and wonder at the world around us. His contribution to the work herein will form part of his lasting legacy, and the impact he had on those around him will endure forever.

ACKNOWLEDGEMENTS

I would like to acknowledge my supervisor, Professor David Robertson, for his boundless enthusiasm, constant support and unwavering kindness. Dr Joanne Pennock and Adefunke Ogunkanbi, for introducing me to the wonderful world of gut parasites and encouraging me to delve deeper! Vandana Ravindran, for providing unwavering support, proof-reading, and friendship. Professor Sam Griffiths-Jones for always having his door open, no matter the query. May Tassabehji, Marian Halfpenny, Yasmin Moore, Helena Collins, Rachel MacGregor and Shevonne Stewart, for providing five years of exemplary support, help, guidance and friendship. Past and present members of the Robertson Lab, in particular Ana Barradas, Avraam Tapinos, Bede Constantinides, Ben Stamp and Rebecca Howard for creating such a wonderful, nurturing and supportive environment.

And finally, my family and friends, in particular my long-suffering partner Dr David Wells Newman, for feeding me, taking care of me, and nodding along to my nonsense.

THE AUTHOR

Alexandra obtained a starred first-class BSc (hons) in Archaeological Sciences from The University of York in 2014. In 2015 she graduated with distinction from The University of Manchester, with an MSc in Bioinformatics and Systems biology.

During the course of her doctoral research, she has become involved with projects being conducted by fellow researchers, including the identification of causal genes in congenital heart disease (Appendix i.) and exploring the population dynamics of Undulate Rays (Appendix ii).

Outside of her research Alexandra has explored an interest in bioethics via collaboration with Cold Star Media, and The University of Salford School of Arts and Media, participating in immersive theatre, expert panel events and podcasts, with the aim of making cutting-edge scientific concepts, and bioethics accessible to the wider community.

Alexandra is very fond of cats.

ABBREVIATIONS

API: Application Program Interface

ASN: East Asian introgressed genes

BMP: Bone Morphogenetic Protein

CDS: Coding Region

EUR: European introgressed genes

GDF: Growth Differentiation Factor

GF: Growth Factor

GO: Gene Ontology

HGMD: Human Gene Mutation Database

HGNC: Hugo Gene Nomenclature Committee

HI: Haploinsufficiency/Haploinsufficient

KYA: Thousand Years Ago

LECA: Last Eukaryotic Common Ancestor

LG+G+I: Le-Gascuel model + Gamma distributed with Invariant sites

LTBP: Latent Transforming growth factor Binding Proteins

MAF: Minor Allele Frequency

MCA: Multiple Correspondence Analysis

MRC: Medical Research Council

MRCA: Most Recent Common Ancestor

MSTN: Myostatin

MYA: Million Years Ago

NCBI: National Center for Biotechnology Information

NHH: Non-Human Hominid

NTD: Neglected Tropical Disease

OMIM: Online Mendelian Inheritance In Man Database

PCA: Principle Component analysis

PPI: Protein-Protein Interaction

SSD: Small-Scale Duplication

TE: Transposable Element

TGFβ: Transforming Growth Factor beta

UTR: Un-Translated Region

VIP: Viral Interacting Protein

WGD: Whole-Genome Duplication

WHO: World Health Organisation

RATIONALE, AND INTRODUCTION TO THE THESIS

This thesis has been written in an alternative format as the work conducted has been adapted for submission to peer-reviewed journals and is presented as such.

The main subject area that has been explored within this thesis is the cause of human genetic disease from an evolutionary perspective, particularly as relates to gene and genome duplications. The thesis is presented as an introduction, three original research chapters, and a final discussion. The introduction in chapter 1 is designed to acquaint the reader with some of the major forms of large-scale mutation that have been found within the human genome, technological advancements made over the course of the last ~20 years, and some fundamental concepts in evolutionary theory. Chapter two explores the role of duplicated genes in heritable genetic disease and the basis, and underlying causes of for this association. Chapter three leverages the findings of chapter 2 to investigate how gene, particularly whole-genome, duplications have provided a platform for co-evolution of human and their eukaryotic pathogen *Trichuris trichura*. Chapter 4 leverages the findings of chapter two in a different manner to chapter 3, namely, by exploring the links between these findings and viruses. This work concludes in chapter 5, by discussing the way that gene and genome duplications have, and continue to shape the human genome over the last ~3.9 billion years, alongside some future implications.

All chapters, whilst in the style of a journal article, have been formatted to conform to standard thesis formatting regulations, and as such, references are included in a consolidated bibliography at the end of the work. Overlap between chapters has been avoided where possible, however when unavoidable, methods have been replicated between chapters, and have been signposted accordingly.

All data are available at: https://github.com/AlexMartinGeary/Hopless_Monsters

Details of author contributions for chapters 2:4 are described below:

Chapter 3: "Gene & genome duplication: Duplication is a correlate of monogenic disease, not a cause"

- **Alexandra Martin-Geary[1]** co-conceived of the project, sourced and curated the data and performed all analysis.
- Mark Reardon[1] provided vital discussion in the analysis of gene age, and haplosufficiency.
- David W. Newman[1], Benjamin Keith[1] and May Tassabehji[1] provided conducted works upon which the early concepts were based
- David L. Robertson[1,2] co-conceived of the project, advised and supervised throughout.

Chapter 4: "Keep your friends close and your enemies closer: the co-evolution of *Trichuris trichura* and the human TGFβ superfamily"

This chapter is not a stand-alone paper and will be supplemented by the *in vivo* work performed by Adefunke Ogunkanbi and Jo Pennock, and additional phylogenetic analysis performed by David Newman prior to publication.

- **Alexandra Martin-Geary[1]** co-conceived of the project, obtained and curated all data, and performed all analyses contained within this thesis.
- Adefunke Ogunkanbi[1] co-conceived of the project, performed the experimental mouse work briefly detailed within the chapter, and which provided the basis for exploration of the TGFβ superfamily.
- David W. Newman[1] performed supplementary phylogenetics not contained within this thesis, alongside vital discussion.
- David L. Roberson[1,2] provided supervision and guidance regarding the phylogenetic analyses.
- Jo I. Pennock[1] co-conceived of the project, provided supervision and guidance regarding immunology and Trichuriasis, and supervised both Adefunke and Alexandra throughout.

Chapter 5: "Viruses control the human intra-cellular systems by exploiting evolutionarily ancient molecules"

- **Alexandra Martin-Geary**[1] sourced all data, with the exception of the driver node assignments and VIP classes contained herein, she curated all data and performed all analyses presented in this thesis.
- Vandana Ravindran[2] co-conceived of the project, performed the network analysis that formed the initial basis of this project and provided the driver node assignments, and provided supervision and valuable discussion throughout.
- Benjamin Stamp[2] provided the VIP class assignments, and supplemental work that will be included in the published article.
- Haiting Chai[2]; Jose C Nacher[3] performed supplemental work which will be included in the published article.
- David L Robertson[1,2] co-conceived of the project, advised and supervised throughout.

[1]Division of Evolution and Genomic science, University of Manchester, Oxford Rd, Manchester M13 9PT

[2]MRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH

[3]Department of Information Science, Faculty of Science, Toho University, Funabashi, 274-8510, Japan

CHAPTER ONE

GENERAL INTRODUCTION

"If the eternal dance of molecules

Is too entangled for us mortal fools

To follow, on what grounds should we complain?

Who promised us that Nature's arcane rules

Would make sense to a merely human brain?"

-Peter Shor (Shor, 2019)

In 1864, inspired by Charles Darwin's "On the Origin of species" (Darwin, C., 1859), Herbert Spencer coined the phrase 'Survival of the fittest' to describe the theory of natural selection (Spencer, 1864), positing that those organisms with the most beneficial micro-evolutionary traits would outlast those without, to become new species.

In 1933 the German geneticist Richard Goldschmidt presented a theory in which he coined the term the "Hopeful Monsters" (Goldschmidt, R., 1940, 1933). This theory presented a deviation from the traditional understanding of Darwin's theory of natural selection (Darwin, C., 1859), suggesting that rather than evolution occurring through the steady accumulation of small variations, important evolutionary events, such as speciations, are brought about by larger mutations which he referred to as 'macro' mutations. The success of any species produced by such an event, he posited, would be by no means certain, with many such organisms failing to flourish, ultimately becoming evolutionary dead ends. Those that survived, he claimed, would do so as "Hopeful Monsters", alive, not because of, but despite the odds. Sadly, Goldschmidt's hypothesis was met with derision due to its perceived irreconcilable conflicts with the work of Darwin, and it seemed as though it too, would result in a dead end. Despite this chilly reception it was his sincere hope, that a future scientific community would

recognise his theory's merit (Goldschmidt, 1982), thus, ironically, rendering the hypothesis, as its subject matter - a hopeful monster.

This, however, was not the first occasion that adjustment of the traditionally accepted Darwinian mode of evolution sought. From the tail end of the 19th century a new movement in evolution had been forming, its intention- to reconcile Gregor Mendel's observations of inheritance, with Darwin's theory of natural selection. This movement, known as Neo Darwinism, gave way in 1942 to the 'Modern Synthesis' (Huxley, 1942), an attempt, in light of new population genetics approaches, and a broader understanding of mutation to resolve these developments with Darwin and Mendel and provide a more unified theory of evolution, which persists to this day.

Between Goldschmidt's pre-DNA proposal of evolution by macromutation in 1933 and the present, the tools available to biologists, and the knowledge gleaned from genomes of humans, alongside a plethora of other species, has led to a far greater understanding of the role and mechanisms of evolution, and the realization that evolution is not a simple process of survival of the fittest. Whilst it may not be wholly in line with Goldschmidt's vision of macro mutations, large-scale duplications of genetic material, that marry up Darwinian thought with Goldschmidt's theorem have been found to be integral to evolution. With the vast advancements that have been made over recent decades in the field of molecular biology, leading to the accumulation of data beyond merely animal, to include a wide diversity of our Eukaryotic and Prokaryotic brethren, the ideas underlying Darwinism, Neo Darwinism

and the Modern Synthesis are now understood to be overly simplistic, it having been proposed that we are on the verge of a new synthesis in evolution (Koonin, 2009). With every advancement and updated theory our understanding of the human condition advances, and the question emerges – Where do we go from here, or are we, as Goldschmidt theorized almost a century ago, merely hopeful monsters?

2001 heralded a major milestone in the journey towards answering this question, in the form of the publication of the first two complete human genomes. One, an amalgam of the genomes of multiple individuals, published by the Human Genome Sequencing Consortium, and the second, a composite of 5 individuals, published by Celera Genomics (Frazer et al., 2009).

In the almost 15 years since these first genomes were published, it has become possible to sequence the genomes of individuals with rapidly increasing speed and accuracy. By comparing these genomes with one another, and with other species, it has been possible to glean a great deal of information regarding inter- and intra-species variation, and the evolutionary histories of the sequences contained therein. It has now been shown that large-scale mutations can and do occur, and that they likely, led to the divergence of humans from other non-human hominids (Bailey and Eichler, 2006). The enlightenment afforded by this genomic revolution does not however end in our ability to define species, as this acceleration in the accumulation of biological data has uncovered a plethora of different types of variation, each with their own signature involvement in evolution, health, and disease.

There are a great many different types of variation that can occur within the human genome, from single point mutations in line with Darwin's theory, to the large scale structural variations (Frazer et al., 2009) of up-to whole-genome size, as proposed by Goldschmidt (1933).

Early large-scale genomic studies estimated that each individual carries a minimum of two possibly disease associated mutations (Dorschner et al., 2013; Zhu et al., 2015), and ~400 potentially deleterious variants (Xue et al., 2012). It should be noted however, that it is was estimated that only in the region of ~10% of genes were liable to have an association with disease (Barabási et al., 2011), and that the function, and phenotypic implications of large parts of the genome remain unknown, therefore these numbers are likely to be far higher than prior estimates suggest.

It is important to highlight, that as sequencing technologies advance, our ability to accurately identify variants also improves. The numbers of estimated damaging variants are likely to change therefore, with both new variants being discovered, and false positives being removed. This was highlighted by the 1000 Genome Project, which classified ~2.3 million previously identified tentative variants as being erroneous (1000 Genomes Project Consortium et al., 2015).

Progress in the fields of bioinformatics and computational biology has proven invaluable in the study of human variation, with potential for future computational analysis ever increasing. In September 2015 completion of the final phase of the thousand-genome project was announced, with 2,504 genomes being made publicly available for

analysis (Birney and Soranzo, 2015). Since then, phenotype and genotype information from a further ~500,000 individuals has been released by the UK BioBank (Bycroft et al., 2018), and the sequencing of a further ~100,000 genomes currently underway as part of an NHS project in the U.K (Caulfield et al., 2019).

Resources such as the 1000 and 100k genomes projects and Biobank have been established to permit large-scale study of human variation and disease. Further to this, an abundance of primary sequence databases are emerging, for humans and other species, containing both whole genome and exome data, such as gnomAD (Karczewski et al., 2019), COSMIC (Tate et al., 2019), and ClinVar (Landrum et al., 2018), alongside collections of collated genomic information, such as Ensembl (Hunt et al., 2018), to name but a few. Whilst these resources present unprecedented potential for the analysis of genomic data, when interpreting the output of these analyses we must exercise caution. As the quantity of published papers has increased, so too has the presentation of often conflicting information.

Genetic variation and human disease

Genetic variation of differing types are the key factors that set species apart. Be it large-scale variation leading to speciation events, or smaller-scale variation that differentiates sub populations and individuals, it is a given that variation is fundamental to diversity. Comparative genomics studies have found that there is a heightened propensity towards human disease mutations to fall within the most highly conserved regions across diverse metazoan species (Miller and

Kumar, 2001). These regions are, therefore more frequently associated with damage to protein sequence, structure, and, consequently, disease (Miller and Kumar, 2001). In order to understand genetic variability, it is imperative that we thoroughly understand the placement of these mutations in their evolutionary context (Henn et al., 2015). It has been observed that mutation patterns in disease patients often appear to have escaped the pressures of purifying selection triggered by aberrant physicochemical properties, when compared with the wild-type (Miller and Kumar, 2001).

As suggested by the UK10K consortium (The UK10K Consortium, 2015) the idea of a single human genetic profile is unrealistic, given that variation within humans has shown to be tremendously diverse. There are however population specific patterns that have emerged which, whilst unable to elucidate a 'one size fits all' profile, provide key indicators of population specific patterns of variation over time. This may be geographic populations as the UK10K discussed (2015), disease populations of phenotypic expression, or, indeed, an amalgam of the two.

Henn et al (2015) discuss various potential Demographic influences on mutation load, including bottlenecks, serial founder effects, and population growth. It has been repeatedly noted that genomes have been shaped by population dynamics and structure (Cavalli-Sforza et al., 1991; Frazer et al., 2009; Polvi et al., 2013; Rosenberg, 2002). In support of this, a 2015 study (1000 Genomes Project Consortium et al., 2015), analysed the genomes of individuals from 26 populations, and showed the

presence of variation specific to each population, with 86% of identified variants being specific to particular geographic groups (ibid).

Local adaptations in population history can play a role in the understanding of selected traits leading to disease. Take for example the disease associated sickle cell anaemia allele, which, whilst is causative of a serious blood disorder, has also conferred a degree of protection against malaria. This protective pleiotropy has allowed the otherwise deleterious allele to evade purifying selection, as it would presumably have allowed carriers to survive past reproductive age, where those without would not. This association is shown to be an instrumental factor in its persistence in regions where malaria is endemic (Allison, 1954; Henn et al., 2015). It is, therefore, imperative to consider, when assessing the effects of such mutations, the context, both biological and geographical, within which the mutation resides.

Genetic variants in human populations are generally divided into two categories: common, and rare. These categories are decided upon by the rate of occurrence of a variant in the population, or, in the case of single nucleotide polymorphisms, minor allele frequency (MAF) whereby a frequency of greater than or equal to 1% is considered common (Frazer et al., 2009). It is possible, however for a gene to have an elevated MAF and a far lower instance of phenotypic expression (Dorschner et al., 2013). Dorschner et al (2013) found that, whilst the PKP2 missense mutations under investigation in their study occurred at a frequency of 0.05%, of the 4,200 participants, the disease itself was only evident in 5 patients, rather than 21, which would be expected of an equally high

penetrance allele (Dorschner et al., 2013). It bears noting however that there are known issues when assessing such genetic information in association with disease, as many medical conditions, particularly those that are not considered 'severe', or debilitating may go unreported within the cohort. This is particularly the case if the focus of the study is on specific major diseases and syndromes. Likewise, the time of onset of disease must be considered, as many genetic diseases have persisted due to late onset, following reproductive age, therefore, it is possible for sampling to have taken place prior to disease onset.

Genetic disease is often, divided into two categories. Monogenic, otherwise known as Mendelian for Gregor Mendel (Bateson, 2010): referring to single genes that can be linked directly with a specific disease phenotype, and complex: wherein numerous genes play an interconnected role in the resultant phenotype. Barabási et al (2011) reviewed a series of methods of network analyses and concluded that it is rare for a disease to result from variants within a single gene, rather, disease is more commonly a result of mutations in multiple genes, such as has been shown to be the case in Autism (Veenstra-VanderWeele et al., 2004), and its vast spectra of potential phenotypes. This is not to imply that Mendelian disorders are insignificant, quite the contrary, in many cases these diseases require only a single mutant allele in order to produce a disease phenotype, as is the case in the majority of cancer syndromes; for example, deleterious variation in a single allele of TP53 gene, also known as "the guardian of the genome" has been found to

dramatically increase an individual's lifetime risk of developing a broad range of cancers (Hollstein et al., 1991).

As previously stated, the analysis of variants, deleterious or otherwise, has proven invaluable in the study of human health and disease, however, there are biases that bare noting (Dorschner et al., 2013), specifically, that the majority of studies to date have focused disproportionately on individuals of European descent, thus skewing the full biological picture by underrepresentation of other populations. This is of particular significance given that rare alleles tend to be population specific (The UK10K Consortium, 2015), and that it is now known that differing populations have clear patterns of admixture, both between each other and non-Human hominids, and dispersal following the initial migration out of Africa (Prohaska et al., 2019). The analysis of mutation load in populations has provided a wealth of information regarding genomic disease, whilst shedding light on the complexity and difficulties of estimating disease risk directly from sequence data (Henn et al., 2015). In particular, the assessment of penetrance in disease causing mutations requires far greater study, As reported by UK10K (The UK10K Consortium, 2015), despite their using a strict criterion, estimations of penetrance were likely to exceed the true penetrance of certain disease-causing alleles.

The majority of variants within the human genome are not thought to be deleterious (Frazer et al., 2009)with an estimated up to 9.5% of these variants, consisting of non-damaging gain or loss copy number mutations (Zarrei et al., 2015). It is proposed that in the region of 103

non-synonymous potentially deleterious mutations are present within "the average human genotype" (Sunyaev, 2001), the majority of deleterious mutations most likely falling within regions of conservation in other species (Miller and Kumar, 2001). Rare mutations, those with a low MAF, have been found, not only to be associated with increased disease risk, but also to be indicative of a propensity to have an early onset, and link to susceptibility to complex disease (Henn et al., 2015). It should be noted that an inverse correlation between frequency of variant and severity of the resultant disease is not coincidental, as the resultant loss of fitness often leads to strong selection against their persistence, whereas variants resulting in less deleterious, or later onset phenotypes are more permissible (Henn et al., 2015). An inherent, and to a degree unavoidable bias exists in the assessment of deleterious mutations, in that those individuals who participate in sequencing studies are all, to a degree 'healthy'. This by no means suggests that they are all in fine physical form, far from it in fact, as the majority of sequencing has been performed on patients with disease phenotypes. However, due to survivorship bias (Mangel and F Samaniego, 1984) embryo-lethal genetic variations cannot be identified in the majority of sequencing studies as they stand. This bias has the serious potential to skew our ability to predict genomic variation as the most severe variants are masked.

Analysis of variation is, at least to a degree, dependant on the type of sequencing conducted. As such, the reduced cost and comparative ease of exome, rather than whole genome sequencing has led to a large proportion of studies using exome data for analysis (Henn et al., 2015).

This has been, and remains, an invaluable resource for the analysis of variation, having shown, for example that an estimated 50% of exonic genes have at some juncture undergone duplication (Richard et al., 2008). It has also been found that disease-causing genes have a tendency to be more exon rich than their benign counterparts (Lieben, 2016; Wu and Hurst, 2016). Exome studies, by definition, result in analyses that specifically target DNA coding regions, without accounting for variation in the non-protein-coding portions of the genome. This is particularly biasing given that ~97% of variants are found in non-coding regulatory regions (1000 Genomes Project Consortium et al., 2015) of which relatively little is currently known of their potential impact (The UK10K Consortium, 2015), with many liable to influence phenotype (Maurano et al., 2015; The UK10K Consortium, 2015). We have, therefore, over the last half-decade entered a new phase of variant analysis, wherein, given the greater accessibility of whole genome data, non-coding regions are more readily incorporated into study (The UK10K Consortium, 2015).

Single Nucleotide Polymorphisms

The mutation of a single nucleotide, also known as a point mutation, or single nucleotide polymorphism (SNP) is the most abundant individual form of variation in the human genome (1000 Genomes Project Consortium et al., 2015). Point mutations can both occur, and influence the phenotype in a variety of ways, dependant on; their location in, or proximity to a gene; if they represent a transition, transversion, insertion or deletion; and, the degree of functional change, if any, that they induce. A very general overview of these types of variation is shown in figure 1.

Further to the potential association with disease, these polymorphisms can influence the binding affinity of targeted pharmaceuticals as discovered by Bloomfield et al (Bloomfield et al., 2016), a single polymorphism at rs6971 has a highly negative effect on the binding affinity of TSPO PET tracers.

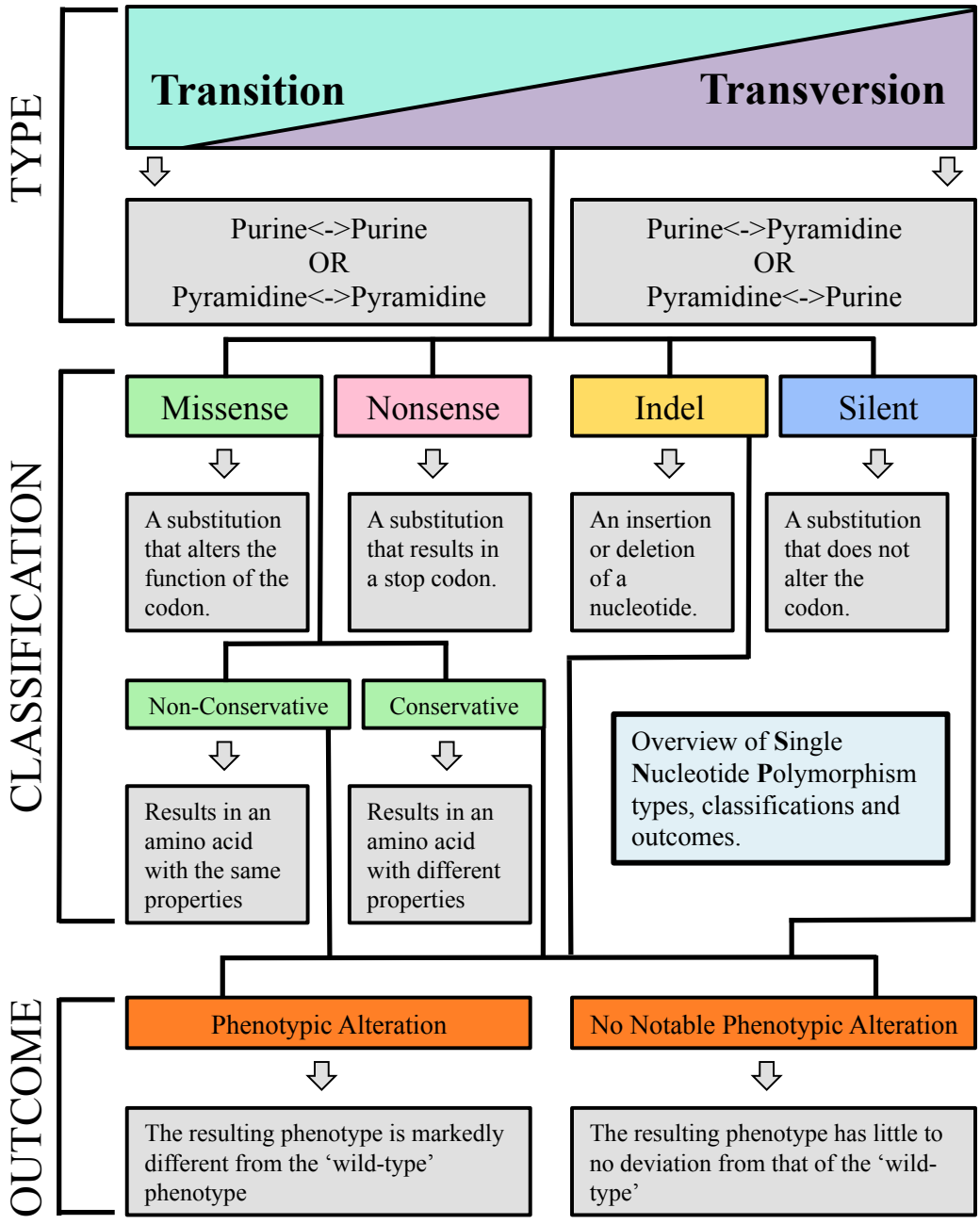


Figure 1: Overview of variation. Including type, classification and potential outcome

SNPs are often categorized, based on their properties, into two groups: Synonymous mutations; variants which do not alter the amino acid encoded for, and non-synonymous; leading to an alteration in amino acid. Studies have shown that, even in situations where a change in nucleotide does not alter the resultant amino acid from that of the wild type, noticeable phenotypic alterations have been found to occur, and should therefore be accounted for (Lieben, 2016; Wu and Hurst, 2016). Non-synonymous missense mutations, those that encode a different amino acid to the wild type, are commonly, and often erroneously, automatically classed as functional, assuming they will have a deleterious impact on any resultant product (Henn et al., 2015). However, due to this type of mutation encompassing a wide variety of polymorphisms, and any functional repercussions being reliant on both the mutation type, and the location within, or proximity to a gene, the functional repercussions of missense mutations are better viewed as a continuum, covering the span between total loss of function, no functional alteration, through to gain of function. Functional prediction is by no means easy, and therefore it may be attractive to broadly class missense mutations as being 'functional' in order to expedite analysis. This type of false dichotomy, however, is not representative of the true biological picture, and can ultimately lead to error and biases.

Insertions or deletions of nucleotides, known as 'indels' cause a frameshift, resulting in codon misalignment. These variants therefore impact on all subsequent codons in a gene and can radically alter the resultant amino acids. Whilst it is not always known what causes this

type of variation to occur, Abeysinghe et al (Abeysinghe et al., 2003) discovered a link between 'frameshift hotspots' and translocation breakpoints, highlighting an association between these areas and replication slippage.

Structural Variation

Whilst less than 0.1% of person-specific variants are classed as 'structural', due to their often very large sizes they can cover a large proportion of the genome (1000 Genomes Project Consortium et al., 2015). The distributions of structural variants (both genomic and population) are similar to that of their point and indel counterparts, with the exception of their being enriched in genomic regions that have been recently subject to large-scale duplications in recent evolutionary history (Frazer et al., 2009). Due to difficulties in sequencing such variation, and the propensity to study disease populations, the roles and frequencies of large-scale copy-number variations (CNVs) in healthy individuals are currently poorly understood (Iafrate et al., 2004).

Copy number variations (CNVs) wherein one or more genes is found in greater, or reduced abundance compared with the expectation, can occur on both an individual, and species level ranging from a single gene, multiple co-localized genes and their surrounding sequence, to the entire genome, with estimates of genetic gain per chromosome ranging from 1.1% to 16.4%, and loss from 4.3% to 19.2% (Zarrei et al., 2015). As can be seen, CNVs do not occur with uniformity across the genome, or even within each chromosome, and, with the exception of genes retained following a whole genome duplication (WGD) event, are found

to be more abundant in regions of observed pre-existing structural variation (Makino et al., 2013a; Schuster-Böckler et al., 2010). CNVs arising through WGD events, de novo duplications, singleton duplication and segmental duplications, have, in recent years taken a more prominent role in the study of human genetics, having been found to play a significant role in disease (Bailey and Eichler, 2006; Bertrand et al., 2010; Dickerson and Robertson, 2012; Diss et al., 2017; Guan et al., 2007; Rice and McLysaght, 2017), with an estimated two thirds of CNVs resulting in a functional change when compared with the wild type (Dudley et al., 2012).

The role of structural variation in disease has been pinpointed by studies into schizophrenia and autism, clarifying the involvement of specific structural variations on chromosomes 16,22,1 and 15 (Frazer et al., 2009; Stefansson et al., 2008). This finding is in support of earlier work, which identified a correlation between combined segmental duplication, copy number variations, and neurological disorders (Varki et al., 2008). It has also been suggested that, in certain cases a tolerated increase in copy number may confer disease resistance, for example, humans have one copy of the TP53 gene which, as previously noted, has an important role in protecting against cancer. Elephants however have 19 copies and subsequently a greatly reduced average lifetime cancer risk (Abegglen et al., 2015).

The Y chromosome

Unlike the autosomes, the human sex chromosomes evolved to a semblance of their current form in relatively recent evolutionary history,

with further population specific shaping occurring as a result of a bottleneck at roughly 5,000–7,000 BP (Zeng et al., 2018). The human sex chromosomes it is now known, originated as an autosomal pair, however, unlike the other autosomes which have maintained relative symmetry, one of the pair underwent chromosomal decay resulting in the current Y chromosome, with subsequent inactivation of X-linked genes in females following this event (Naqvi et al., 2018). It is not however solely the origins of allosomes that differ from their autosomal counterparts, as unusual patterns of gene CNVs are also found (Lucotte et al., 2018; Maranda et al., 2019; Naqvi et al., 2018). Allosomal CNVs exhibit an increase of 33% in processed pseudogenes (Maranda et al., 2019), and enrichment of expression of ampliconic CNV genes sharing unusual patterns of diversity between human populations, are thought to have arisen as late as 50,000 years ago (Maranda et al., 2019). CNV genes in these chromosomes, have elevated dosage sensitivity, suggested to represent an underlying ancestral dosage sensitivity. It is proposed, that this ancestral dosage sensitivity can be identified by involvement with miRNA mediated repression mechanisms, which may also explain the retention of an Y chromosomal partner, or X-inactivation. Likely influenced by differential X-inactivation, many sex-specific diseases and disorders are now known to be linked to the presence or absence of dosage sensitive genes on the sex chromosomes (Graves, 2016). RNA influenced differences in paralog expression have recently also been identified in the autosomes (El-Brolosy et al., 2019; Ma et al., 2019), elucidating the mechanistic basis by which knockouts are found to illicit

a less severe phenotypic change than expression reduction in the same gene.

Transposable elements

There are remarkable differences in genome size, across all life, with, in the region of 6,600-fold variation in size between animal species alone (Suh, 2019). A recent study investigating the role of transposable elements (TEs) in larvacean tunicates, a close relative of vertebrates not known to have undergone WGD in their evolutionary history, found that ~83% of variation in genome size between species was due to differential copy numbers of non-autonomous TEs (Naville et al., 2019). In part, the increased accumulation and fixation of many small TEs, it has been proposed, may have been tolerated due to their being less likely than larger autonomous TEs to lead to a deleterious phenotype (Suh, 2019). In contrast to this, a second study (Arkhipova and Yushenova, 2019) investigated the role of giant TEs and retrotransposons, finding that whilst there are instances of very large TEs in the human genome, their presence does not automatically infer an excess production of gene products, as their expression is heavily regulated by RNA silencing mechanisms and chromosomal relocation can allow duplicated genes within these regions to gain new, or different function which may to some degree, limit their deleteriousness. Despite their potential for disruption of stoichiometry, or other potentially deleterious consequences, large transposable elements are thought to have become fixed due to the important role they likely have played in genetic

diversification, novelty and adaptation (Arkhipova and Yushenova, 2019).

It is known that there is a strong relationship between TEs and viruses, including shared qualities between the two facilitating transposition (Arkhipova and Yushenova, 2019; Gilbert and Feschotte, 2018; Rodriguez et al., 2017). A recent study of the various epigenetic features of transposable elements found them to be dependent on the age of their component retroviral ancestors (Ohtani et al., 2018), it is suggested that a deeper understanding of the differential evolutionary histories of retrotransposons and their associations with both the genome and epigenome, including the novelty provided by their inclusion and subsequent endogenization (Johnson, 2019), may allow a greater insight into 'viral mimicry' in the human genome, and in turn allow for the development of therapeutics that exploit this machinery (Ohtani et al., 2018).

Segmental duplication

A high proportion of human structural variants have been found to be made up of segmentally duplicated regions (SDs) also known as low copy repeats, or tandem duplications (Bailey and Eichler, 2006). These arise largely as the result of non-allelic homologous recombination (NAHR), which results in recombination slippage (Bailey and Eichler, 2006; Varki et al., 2008). These regions are defined as stretches of DNA of between 1 and 10 kilobases, that differ by just 10% identity or less (Bailey et al., 2003), and occur both inter, and intra chromosomally as a result of Robertsonian Constitutional translocation (breakages in the

region of the centromeres of two chromosomes that lead to translocation of genetic material between the two) (Bailey and Eichler, 2006; Linardopoulou et al., 2005), and are influenced by chromatin organisation (Ebert et al., 2014; Emanuel and Shaikh, 2001). Flanking segmental duplications, SDs with between 50 kilobases and 10 megabases of intermediate sequence (Bailey et al., 2003; Sharp et al., 2005) have been found to be 'hotspots' of variation, with a high correlation between them and deleterious genetic traits.

It is estimated that, providing a neutral molecular clock, and a low likelihood of gene conversion, SDs which are fixed in the human genome, and share genetic similarity of 90%, emerged in the region of 35Mya (Bailey and Eichler, 2006), coinciding with the 'Grande Coupure', towards the end of the Eocene epoch, an event of rapid change in global biodiversity (Mennecart et al., 2018).

Whilst SDs can occur anywhere in the genome (Zarrei et al., 2015), they are most commonly found in proximity to the centromeres or telomeres (Bailey and Eichler, 2006; Horvath, 2000), or at sites of double strand breaks, suggested to be linked to reproductive isolation between species (Ebert et al., 2014). Analysis of SDs in humans and primates has uncovered a relationship between tandem repeats found in 3' un-translated regions, and divergence in gene expression (Bilgin Sonay et al., 2015; Waldron, 2015), which is posited to have played a key role in the evolution of the primate lineage, inclusive of rapid expansion and fixation that ultimately led to the isolation of the hominid line (Richard et al., 2008).

Gene duplication (small-scale and whole-genome)

More than 30 years after Goldschmidt presented his hypothesis, the Japanese geneticist Susumu Ohno posited his own, initially contentious theory, knowingly or otherwise supporting that of Goldschmidt's hopeful monsters. Ohno's "2R" hypothesis (Ohno, 1970), proposed that, during the course of early vertebrate evolution, the entire genome duplicated not once, but twice. These duplication events, among other large-scale mutations that followed are now known to have not only increased evolutionary capacity via the provision of repurposable genetic material, but also instigated a period of explosive sub-speciation (Richard et al., 2008), brought about by differential loss of genetic material, and resultant reproductive isolation.

Facilitated by the recent advances in genomic technologies it has now been identified that genes arising, and having been retained following, whole genome duplication events play a significant role in human health, disease, and evolution (Dickerson and Robertson, 2012; Makino and McLysaght, 2010; Richard et al., 2008; Schuster-Böckler et al., 2010; Singh et al., 2015; Van de Peer et al., 2017), with WGDs having in particular been linked to the recent divergence in the ancestor of hominid and chimpanzee (Bailey and Eichler, 2006).

There are an estimated 3,544-7,831 human genes thought to have arisen and been retained as a result WGD events alone, each displaying high degrees of mammalian conservation (Singh et al., 2015). These genes, named ohnologs for Susumu Ohno, have been linked to a propensity for deleterious mutation (Dickerson and Robertson, 2012;

Makino and McLysaght, 2010; Singh et al., 2015), exhibiting an enrichment of genes with cancer associations (21.6% to 26% of all ohnologs but only 8.3% of non ohnologs), developmental genes such as those linked to congenital heart disease (Fotiou et al., 2019), gene regulation and signalling, and, in particular autosomal dominant disease (Singh et al., 2015), with ~80% of human monogenic disease associated genes having been found to be part of a duplicated paralog pair/family (Dickerson and Robertson, 2012).

Although many genes duplicated and retained following both WGD and small-scale events (SSDs) appear non-functional (Dudley et al., 2012), this likely accounts for just a third of their total number. The remaining duplicates, if expressed, may contribute to deleterious phenotypes (Dudley et al., 2012), with an estimated 50% of exonic genes having been found to have a duplicate somewhere in the genome (Richard et al., 2008). Potentially due in part to small hominid effective population size (Bailey and Eichler, 2006), these duplicates have, even in cases where they prove fatally deleterious when disrupted, become fixed in the population (Ebert et al., 2014). This is not wholly surprising given that WGD has been found to be the mechanism by which dosage-threshold sensitive genes are duplicated. The importance of stability of copy number of these genes, and therefore their retention, having been highlighted by the fact that they are refractory to later duplication (Makino and McLysaght, 2010).

Whilst there is a proven strong association between ohnologs and disease, the story of WGDs however is not entirely an unfavourable one,

as it has been shown that not only do they confer immediate fitness benefits via reduction of expression 'noise' (Pires and Conant, 2016), but they also lead directly to the evolution of complexity, adaptive plasticity and expansion of global biodiversity, through introduction of large quantities of repurposable DNA (Innan and Kondrashov, 2010), which can lead to sub-division of the ancestral function, novel function, or differences in dosage (Conant and Wolfe, 2008).

This proposition of novel function introduction, however, is in direct contrast to that of Singh et al (Singh et al., 2012), who presented a modified compensation model of ohnolog fates (Gu et al., 2003). Singh suggested that both genes in any ohnolog pair are 'locked in' to retention, as loss would inevitably lead to, particularly dominant, disorders. A second contentious issue associated with genes arising from WGD is, to what extent asymmetrical divergence between the two genes in the pair occurs, if, in fact it occurs at all (Force et al., 1999; Ohno, 1970; Pachter, 2015). However, a number of studies investigating this issue have shown that asymmetrical evolution within gene pairs, both in sequence and function does in fact exist (Dickerson and Robertson, 2012; Kellis et al., 2004), but the degree to which it occurs, and appropriate measurement thereof is still under scrutiny (Pachter, 2015; Tasdighian et al., 2017).

Largely, the discussion around the interlocking of ohnolog fates and evolution revolves around the dosage balance hypothesis, wherein all genes are required to remain functionally constrained post duplication to maintain systemic stoichiometry (Birchler and Veitia, 2012; Veitia and

Birchler, 2010). Gene loss is at its greatest directly following a WGD event (Scannell et al., 2006), presumably due to the high energetic expense polyploidy represents. Certain genes, however, require maintenance following duplication, either to prevent failure of the complexes in which they are involved, to prevent disruption of working complexes by their incomplete interaction partners, or to act as a buffer should the partner lose function (Gu et al., 2003; Hakes et al., 2007; Hsiao and Vitkup, 2008; Lopez-Bigas, 2004) (Gu et al., 2003; Hakes et al., 2007; Hsiao and Vitkup, 2008; López-Bigas and Ouzounis, 2004), all of which have the potential to directly impact on systemic stoichiometry.

Various mechanisms to counteract issues of dosage balance are present within vertebrate species (Graves, 2016). Whilst these mechanisms are predominantly species-specific, it is thought that they have individually evolved from far older shared ancient eukaryotic mechanisms (Graves, 2016). Surprisingly this variation in compensatory mechanisms extends to the complete absence in certain vertebrate species (Graves, 2016) however, the underlying causes of this variation and absence thereof have not been fully explored. One of the most investigated within humans is nonsense mediated mRNA decay, whereby mRNA transcripts identified as containing a premature stop codon are targeted and terminated. This mechanism having been recently implicated in compensation for dosage sensitive genes (El-Brolosy et al., 2019b; Ma et al., 2019).

The disruption of stoichiometric balance posed by variations in dosage can be a significant issue to cellular health (Schuster-Böckler et

al., 2010). However, due to afore mentioned compensatory mechanisms, it can be difficult to establish the degree of imbalance posed by copy number variations from genomic data alone (Schuster-Böckler et al., 2010), as the impact of variation is a sliding scale rather than a binary event. This is also true of dosage balance. In the case of duplicated genes, variations in copy number, or altered function in existing copies may be impactful, which, whilst not necessarily considered to be significantly deleterious in a single gene, for interactions between a combination of mildly dosage sensitive genes, particularly those within an essential network this may result in major phenotypic repercussions (Makino et al., 2013b).

Adaptation, Introgression, Non-Human Hominids, and Non-Human Primates

With the advent of shotgun sequencing technologies came a new ability to obtain DNA from previously inaccessible sources – archaeological remains (Hofreiter et al., 2015). We had previously been able, based on morphological analysis of skeletal remains, to determine non-Human hominid (NHH) species and draw some conclusions regarding their origins, migration patterns, and ultimate demise, however, due to the long time periods since their extinction, any extractable DNA was often heavily fragmented and degraded, with high degrees of deamination rendering meaningful analysis difficult (Hofreiter, 2001; Hofreiter et al., 2015).

Shotgun sequencing, requiring short reads, and producing high read depths, however, was perfectly placed to address this issue, and, in

2010, 13 years after the first short mitochondrial sequences had been extracted (Hofreiter, 2011; Krings et al., 1997), Green et al were able to publish the first draft Neanderthal genome(Green et al., 2010).

Also in 2010, using just a single bone discovered in Siberia, Reich et al (Reich et al., 2010) were able to identify a new species of hominid from recovered genetic material, this species is now known as the Denisovans.

Currently, relatively little work has been conducted investigating the different duplication rates between humans and their closest known relatives- Neanderthals and Denisovans. However initial studies indicated certain contrasting characteristics between human and Neanderthal CNVs (Green et al., 2006), and with increasing quantities of archaic hominid material being made more widely accessible (reich.hms.harvard.edu Reich, 2019) this, and other work of its kind is likely to be greatly expanded in future (Marciniak and Perry, 2017).

Part of Reich's early Denisovan analysis (Reich et al., 2010) included a cursory investigation into the rates of SDs found within this species, when compared with Neanderthal and modern human. Surprisingly, it was found that Denisovans had a far higher than expected percentage of private SDs (those specific to Denisovans) (Denisovan 2.27mb; Human 1.32; Neanderthal 0.60). Beyond this, regions of similarity were identified, that were nearer chimpanzee than modern human or Neanderthal, including areas found to be associated with human disease (16p12.1. & 5q13) (Reich et al., 2010).

Comparisons between human, NHHs and non-Human primates (NHPs) may provide insight into the observed acceleration in CNVs and SDs that has occurred in modern humans by comparison with the latter and provide insights into other structural aspects of these occurrences (Varki et al., 2008). For example, between human and chimpanzee it has been found that, of nine observed human specific pericentromeric inversions, seven are associated with SDs (Varki et al., 2008).

The frequency of SDs is far from fixed across species and appears to have been increasing in humans and recently related NHPs, with human SDs accounting for approximately 5% of the genome, whilst New world monkey genomes contain just 2% (Bailey and Eichler, 2006; Ebert et al., 2014), and far greater variation than expected between human and fly or worm (Bailey and Eichler, 2006). This, it is suggested, is an artefact of the relatively low effective population size in hominids when compared to the other species (Bailey and Eichler, 2006). This variation has also been observed between chromosomes (Emanuel and Shaikh, 2001) with the estimated evolutionary age of duplications ranging from 90% similarity in chromosome 19 to 99% in Chr14 (Zarrei et al., 2015).

In just under a decade, we have been able to accrue vast quantities of ancient DNA and have discovered that early Homo sapiens sapiens walked the earth with not one, but multiple non-human hominid species, and, whilst it was once a controversial assertion, it is now known that admixture between these species occurred repeatedly (Browning et al., 2018; Enard and Petrov, 2018; Reich et al., 2010; Slon et al., 2018). This admixture is traceable not only in the genomes of our NHH brethren, but

also within the human genome, allowing a far greater understanding of the role of introgressed genes in health and disease (Dolgova and Lao, 2018; Greenbaum et al., 2018; Sanz et al., 2018). One of the major insights we have gleaned from this area of study is the intersection between introgressed Neanderthal genes in the human genome, and adaptation (Enard and Petrov, 2018). Enard et al, found that positive directional selection has guided the enrichment of viral interacting proteins within the human genome, hypothesizing that this fixation of introgressed genes occurs as a result of simultaneous admixture, and viral exposure, which they termed the poison/antidote hypothesis(Enard and Petrov, 2018).

Epigenetic variation

In 2009 it was postulated that research into the epigenome would provide "tremendous insights into the genetic architecture of complex traits" (Frazer et al., 2009). A number of early studies provided highly useful information regarding the epigenetic mechanisms of non-genetic inheritance(Szyf, 2015). These mechanisms include noncoding RNA regulation, chromatin and histone modifications leading to increased accessibility of DNA, and, DNA methylation and phosphorylation, leading to chromatin inactivation and interruption of transcription factor binding (Szyf, 2015). Szyf (2015) highlighted the difficulties in studying parental transmission of epigenetic modifications in humans. The issues presented, are largely due to the inherent ethical difficulties in studying pre-existing epigenetic triggers, such as the 1998 Quebec ice storms,

leading to heightened stress, or the epigenomes of children who have been deprived, neglected or abused.

Epigenetic silencing, particularly of regions high in copy number variation, has a notable impact on an individual's phenotype (Schuster-Böckler et al., 2010), with a correlation having been found between chromatin organization and the occurrence of segmental duplications (Ebert et al., 2014). It has also been suggested that factors such as CpG hypermutability have led to increased mutation rates and diversification (Li et al., 2010), as the rapidity of evolution of epigenetic modifications is considerably greater than that of the lengthier process of selection (Szyf 2015). Interestingly, it has also been proposed that differential loss of duplicate genes following WGD may in part be governed by epigenetic mediation (Sémon and Wolfe, 2007).

As identified by Singmann et al (Singmann et al., 2015) and Koch (Koch, 2015) epigenetic mutations reveal different patterns depending on biological sex. Singmann et al (2015) found that 1,184 of the CpG sites they studied showed differential patterns of methylation between the 1,799 cisgender men and women in their European cohort. As noted by Koch (2015) this casts light on prior findings of differences in disease risk and incidence between the cisgender sexes, and is therefore, alongside expansion to include non-binary, and trans-sexes, likely to become a valuable avenue of study.

One of the particular difficulties posed by the study of epigenetics, is understanding the transmission of somatic mutations to gametes and subsequent heritability. Of particular interest is the manner by which

epigenetic marks are retained during germ cell differentiation, as it has been suggested that such modifications are not completely erased during this process (Szyf, 2015). It is proposed that the possible mechanism by which this is likely to occur is transmission by noncoding RNAs of behavioural signals (Soubry, 2015; Szyf, 2015). Paternal gamete-mediated epigenetic inheritance in offspring may impact on the phenotype (as opposed to the gestational impacts that have mainly been studied) as it has been posited that spermatogenesis, a continuing process occurring throughout the lifespan of the male, may lead to the accumulation of epigenetic modifications in these cells, and subsequently be passed to the offspring as a result of parental exposure (Szyf, 2015). Further studies are gaining increasing evidence to support the role of exposure of toxins and addictive substances in the parent, resulting in epigenetically moderated phenotypic changes in the offspring (ibid).

Establishing gene ages

As briefly noted above, the exploration of evolutionary attributes is reliant on the adequate identification of the origin of said attribute across evolutionary time. Several methods exist that aid in identifying the age of genetic elements. The last common ancestor (LCA) method of gene assignment assigns ages based on the likely age of the common ancestor of all identified orthologues(Domazet-Lošo et al., 2007; Murahwa et al., 2019). This method has the advantage of being able to identify those genes with an ancestor rooted deep in evolutionary history, however lacks the ability to achieve the fine scale resolution needed when making between-gene comparisons between members of gene families, which have often undergone successive duplications. By contrast, the most recent common ancestor (MRCA)

method of gene assignment retrieves the shared age of the most recent divergence between orthologues in known species (Murahwa et al., 2019), hypothesizing that each gene cannot have arisen more recently than this occurrence. Whilst this method lacks the ability to detect the deep evolutionary ties of genes arising as a result of a recent duplication event, it however, it provides finer resolution when making direct comparisons between related genes.

Function- Gain, loss and co-occurrence

The analysis of functional repercussions of genetic variation is one of the most explored areas in the field of variant analysis. As previously discussed, this is best represented in a loss-gain spectrum. It has been suggested that ~20% of common, non-synonymous SNPs alter function (Sunyaev, 2001), with a 2015 study of 18,903 genes indicating that more than 30% contained a mutation likely to have functional repercussions (The UK10K Consortium, 2015).

Miller and Kumar in 2001 proposed that loss of function, rather than being an absolute, discernible by the mutation class, is instead, dependant on the mutation of particular amino acids at specific locations, which would have a discernible impact on phenotype and function. A 2013 study by Petrovski et al (Petrovski et al., 2013) highlighted within their own research the inaccuracies caused by imposing a false dichotomy. In their paper, all nonsense, splice, and missense variants were categorised as functional, whilst synonymous variants were classed as non-functional. This arbitrary dichotomy is repeatedly reported throughout the literature. Classification of ‘protein-altering variants’ inclusive of all missense variants and truncations, regardless of position,

implies that all variants of this kind have a demonstrable functional impact (Rackham et al., 2015). Zhu et al (2015) however, whilst initially classifying their 'functional mutations' in this manner, cross referenced all shortlisted genes with the OMIM database, and pre-existing patient data for the specific diseases under scrutiny, in order to further rule out any erroneously added variants. Their method, while being laudable for stringency, cannot account for the presence of de novo mutations, and therefore remains wanting in its representation of the true biological context.

The identification of loss of function as a result of genetic variation is fraught with difficulties; it has therefore been the case that researchers have often been compelled to draw a line of probability. For example, Dorschner et al (Dorschner et al., 2013) highlighted variants within the Human Gene Mutation Database (HGMD) (Stenson et al., 2017) which had not been previously identified as disease causing, they opted to further investigate these variants, implementing the criteria that if they appeared within the first 90% of the gene they were included in the study, assuming that they are more likely to result in a deleterious truncation (ibid). This is problematic for two reasons. Firstly, variants within a gene, no matter the location, do not automatically infer truncation, and if they did, this may not infer a total loss of function. Secondly, variants in the final 10%, depending on the protein product, may be equally as disruptive as a variation occurring earlier (Bell et al., 1981; Winterer et al., 2008).

Miller and Kumar (2001) within their study of cystic fibrosis, discuss degrees of loss of function, in that mutations within variable sites in conserved genes, whilst not being causative of cystic fibrosis, are involved in less severe pulmonary conditions. They conclude this to indicate that, whilst polymorphisms in invariable sites within conserved genes are most commonly highly deleterious, any change in amino acids within these genes is likely to have a degree of deleteriousness.

Petrovski et al (2013) found contrasting patterns in different disease classes relating to the likelihood of their being influenced by genes more, or less, tolerant to functional variation. For example, their developmental disease class showed a high correlation with genes intolerant to functional variation, whilst the immunological class was more directly linked to genes with a high frequency of variation (ibid). Given that genes intolerant to functional variation are more likely to result in an aberrant phenotype it is unsurprising that they are more commonly found in the developmental class, as deleterious mutations in this class are most likely to result in embryo-lethality, and cascade effects.

The co-occurrence of loss of function is of particular significance when assessing the role of variation in complex disorders. As the quantity of data available for analysis has increased, we have become more aware that genetic interactions between loci are also highly contributory to complex disease (Marchini et al., 2005), and that disease pairs with variations in functionally similar domains show a greater degree of comorbidity than more dissimilar regions in which variation

occurs (Barabási et al., 2011). Variation in genes which exhibit co-segregation has been linked to pathogenicity (Dorschner et al., 2013), combined with the analysis of variants within regions of linkage disequilibrium (Frazer et al., 2009) this presents a valuable avenue of study, and an important consideration when assessing variation in complex disease.

Barabási et al (Barabási et al., 2011) reviewed the field of genetic interactions in complex disorders. They discovered high correlations between phenotypically similar diseases, and gene and protein network interactions. Their findings support the hypothesis that disruptions to interactions within protein networks and that of gene products result in the same disease, which they found to be ten times greater than would be expected to occur at random (ibid). They postulate that predicting variants involved in complex disease may therefore be possible by identifying 'disease modules', comprised of highly associated areas of the interactome, and investigating their components.

Essentiality

Recent studies have, afforded insight into certain attributes common to the majority of genes essential to optimal cellular health. For example, essential genes have a tendency towards higher expression; are commonly found to be haploinsufficient; and are more likely to be involved in a greater degree of protein-protein interactions than their non-essential counterparts (T. Wang et al., 2015); they are less likely to have a paralog elsewhere in the genome, and have a tendency towards greater involvement in such processes as transcription and translation,

with comparably less involvement in signalling when compared to non-essential genes (Blomen et al., 2015). It has also been found that the majority of genes involved in proteasome subunits are essential, and that new essential genes exhibit higher levels of interaction with their ancient counterparts than they do with other 'new' genes (Blomen et al., 2015).

Cellular complexity and essentiality studies have indicated that, when compared with single-celled organisms, humans have a far higher degree of essentiality. It has been found that rather than added robustness through redundancy given our large genome size, increased complexity is reliant on greater degrees of essentiality (Blomen et al., 2015). It is proposed that organisms with a low degree of complexity also have a reduction in genomic fragility as they are able to tolerate more variation, due to a propensity towards more frequent functional overlap than that of humans (Gu et al., 2003). This proposal was initially contested (Makino et al., 2009), however, greater evidence has now been provided in its support (Wang et al., 2015).

Analysis of evolutionary features of genes and genomes has shown that, the most highly conserved regions are those most likely to perform an essential function (Miller and Kumar, 2001). Conservation analysis has revealed that there are large swathes of the genome that exhibit a high degree of conservation with distantly diverged species (Henn et al., 2015; Wang et al., 2015), an estimated 77% of essential genes having emerged in a pre-metazoan progenitor (Blomen et al., 2015). These regions are purportedly invariant, as any variation within them appears to have been subject to strong purifying pressures (Wang et al., 2015). This

indicates that mutation within these regions likely leads to major deleterious functional outcomes, and cannot therefore be tolerated (Henn et al., 2015; Miller and Kumar, 2001; Wang et al., 2015). Population studies, however, have shown that, disease can and does occur in genes essential to cellular health and, variations can and do persist (Barabási et al., 2011). Ohnologs, as the product of WGD, are also highly conserved. However, despite this fact, they have been found to show no more, or less essentiality than that of their singleton counterparts (Makino and McLysaght, 2010).

Protein network analysis has shown that if an essential gene contains a variant, then it is highly common (~88%) for it to be either contributory to, or the definable cause of disease (Zhu et al., 2015), however, studies to date have focused on small cohorts, therefore larger scale analysis would be required to support this. Protein network analysis has also shown that essential genes are often associated with hubs present in a variety of tissue types (Barabási et al., 2011; Goh et al., 2007). These gene hubs are subject to a slower rate of evolution and are often older than their non-hub counterparts (Barabási et al., 2011). Conversely non-essential disease associated genes have been found to rarely encode hubs, often having few interactions, and are most commonly active in only limited tissue types (Barabási et al 2011). This further lends credence to the prior proposal that; human disease variants are unlikely to represent the most severe variation capable of accumulating in the genome, as these variations would automatically infer embryo-lethality.

Evolutionary and conservation studies

Analysis of conservation and other evolutionary properties, such as substitution ratios, and phylogenetic relationships, are some of the most well-established methods of analysing genomic differences. As previously discussed, population-based methods often take into account allele frequency and investigate rates of variation both within, and between geographical populations, or as is often the case with disease populations, groupings based on phenotype. Evolutionary analyses have been used to investigate genomic variation in a broad spectra of species, assessing inter, and intra species conservation and diversity, based on amino acid characteristics or phylogenetic conservation (Braga et al., 2018; Newman et al., 2019; Olabode et al., 2016; Petrovski et al., 2013; Puigbò et al., 2019; Wood et al., 2018), and ultimately drawing conclusions about speciation, function, homology and disease based on these findings, alongside elucidating the mechanisms (such as duplications, gene transfer, transitions and transversions, selection pressures etc) by which different parts of the genome arose and evolved.

Genome Alignment

Appropriate alignment methods are a powerful tool for the identification and elimination of errors arising prior to, or during the sequencing process, however they are not infallible as, whilst appropriate alignment is able to account for many false sequencing calls, true variants, specifically de novo mutations may often be excluded (Peng et al., 2013).

Architectural differences, and the accurate alignment thereof, have been cited as being of key importance (Henn et al., 2015) with emphasis on the need for integrated experimental and bioinformatics studies to verify findings, and provide a better understanding of the role of both structural variation and chromatin organization in human variation and disease.

Beyond complications posed by sequencing method, there are numerous post-sequencing issues related to the composition of the sequence itself. The most profound of which being the alignment of regions with complex and repetitive sequence, as is often found with large-scale duplications. These regions are difficult to differentiate, and are underrepresented in sequencing output (Bailey, 2002; Bishara et al., 2015; Jain et al., 2018) this is especially the case with cross-chromosomal duplications (Horvath, 2000) and flanking segmental duplications. These biases must, therefore, be accounted for when selecting the appropriate datasets for analysis. Issues posed by both sequencing and alignment may in some way, however, be mollified by the rise of long read, and single cell sequencing, which, while currently not as widely used as shotgun techniques, are rising in popularity (Chaisson et al., 2015; Jain et al., 2018; Pendleton et al., 2015).

Genome-wide association studies

Genome wide association studies have been used extensively in the analysis of human variation. They have revealed the involvement of sub-phenotypes, suggesting that, whilst multiple pathways are associated with certain diseases, patients do not necessarily require a disruption in all

pathways, rather a combination of one or more may lead to the disease phenotype (Esposito et al., 2018; Frazer et al., 2009).

Despite their obvious power, association studies have increasingly bourn criticism, as it has been noted that they lack the ability to identify rare variants and small-scale interactions (Frazer et al., 2009; Sazonovs and Barrett, 2018; Shaw et al., 2019). A study conducted by Ladouceur et al (Ladouceur et al., 2012) compared a number of association study methods and concluded that no single method worked optimally, and all lacked power to detect variants with low causal effects. Marchini et al (Marchini et al., 2005) suggested that interaction-based searches would improve the abilities of association studies to identify variants. Whilst this method was not reviewed by Ladouceur et al (2012), a similar method combining effects and interactions, proposed by Liu and Leal (Liu and Leal, 2010), which attempts to identify rare variants by tabulating genotype data was included. Although this method was not, at the time found to have optimum performance, it was later used as a basis for improved methods (Ladouceur et al., 2012). Due to these issues, it is now accepted that GWAS studies alone cannot elucidate the true landscape of both common and rare variants, however, combined with complementary methods to incorporate rare variants their utility has continued to be proven (Sazonovs and Barrett, 2018; Shaw et al., 2019).

It is clear that genome wide association studies, whilst proving useful in certain scenarios, have profound limitations. It is therefore important when considering the use of GWAS to ensure not only

compatibility, but also, account for the limitation of each association type before drawing conclusions (Ladouceur et al., 2012).

Variant effect prediction

Assignment of degree of 'deleteriousness' of a variant depends heavily on its predicted functional outcome. As noted by Henn et al (2015) the choice of algorithm used to make such assessments can have a distinctive impact on the final analysis. In 2013 Petrovski et al (2013) conducted a study into the effects of variation on personal genomes. The initial phase of the study, as discussed earlier, erroneously classified all missense mutations as being functional, however, by supplementing the analysis with the use of PolyPhen2 (Adzhubei et al., 2010), a rule based physical and comparative selection tool to perform between-population comparisons, they found a reduction of mutations classified as functional of 33%, illustrating the utility of using such tools to supplement predictions of phenotypic outcome.

It is clear that we need to relinquish pre-existing assumptions regarding a binary understanding of mutation types, specifically the expectation that all missense mutations are functional. The Petrovski et al (2013) study presented a method of predicting disease genes based on a residual variance intolerance score (RVIS), using patterns of normal variation, cross referenced with variation in disease cohorts to produce a statistical analysis of a variant's propensity to cause disease. This method was later reviewed by Rackham et al, who presented an alternative (EvoTol) (Rackham et al., 2015), leveraging conservation and expression data to predict disease genes. It was found that when applied

to the same data, each method yielded different results. However, they clearly highlight the utility of leveraging evolutionary data in the prediction of human disease associations. Wang et al (2015) also discuss the extension of pre-existing methodologies for use in bioinformatics analysis, this time with reference to epigenetic modifications. By adapting existing techniques to target negative selection, cross referenced with essential genes, they found that combining their technique with CRISPR data, they were able to predict gene essentiality with a high degree of accuracy (Wang et al., 2015).

Maurano et al (Maurano et al., 2015) discovered upwards of 60,000 variants impacting on regulatory DNA accessibility, and transcription factor occupancy. Using DNase I hypersensitive site sequencing, combined with DNA genotyping on multiple tissues and individuals, they were able to identify 500,000 variants, common to regulatory regions, which directly impact transcription factor occupancy (ibid). This work highlights the importance of both the involvement of epigenetic modification data in genetic disease prediction, but also the need to account for both non-coding DNA, and cell specific processes.

Many of the tools and methods currently in use have been designed to assess large populations as a cohort; therefore, their clinical use for individual diagnostic purposes is often limited. This is especially true when analysing certain anonymised data, as, for example, ExAC's (Exome Aggregation Consortium et al., 2016) anonymity criteria make it impossible to extract individual participants and compare different regions of one individual's DNA. The 1000 (1000 Genomes Project

Consortium et al., 2015) and 100k genomes data (Caulfield et al., 2019) however do allow the assessment of an individual and, particularly with the latter, may prove their worth in this regard.

One key area that is still lacking in empirical support is the relationship between genotype and phenotype. Work is being conducted into this with resources such as the Gen2Phen (Webb et al., 2011) project, a standardized tool for analysis of variants which aims to create a consolidated 'biomedical knowledge environment' through which to analyse genotype to phenotype information. In addition to this, bioinformatics analyses are being conducted in an attempt to clarify the relationships between gene and protein networks and expression levels (Yu et al., 2015), which may, to some degree, provide a greater insight into this complex relationship.

Network analysis

Network analysis has become an important tool in the biologist's repertoire, having been successfully used to analyse gene and protein interactions (Boyle et al., 2018; Kuzmin et al., 2018; Monaco et al., 2018; Szklarczyk et al., 2017), ecological systems (Brodie et al., 2018; Muscente et al., 2018; Rebolledo et al., 2019), and host-pathogen interactions (Ahmed et al., 2018; Lee et al., 2018; Ravindran et al., 2019), amongst a plethora of others (Barabási et al., 2011; Emilsson et al., 2018; Gosak et al., 2018; Zhou et al., 2014).

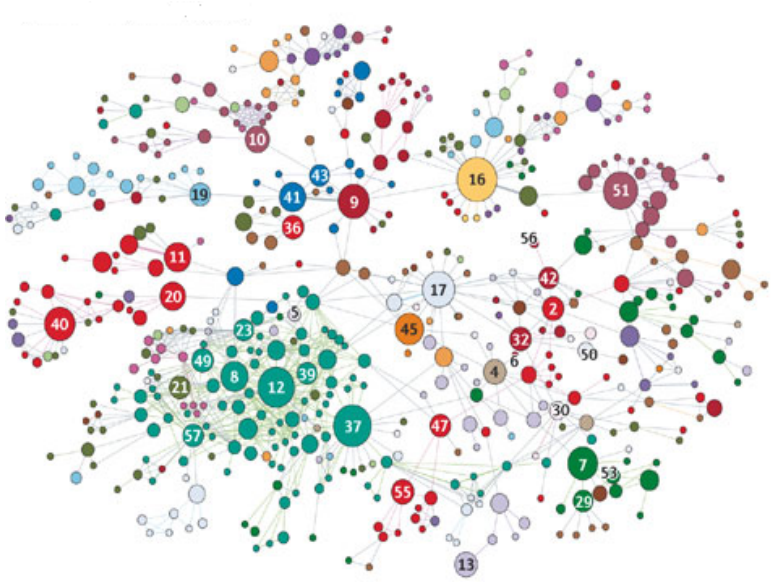


Figure 2 : Human disease network. Each node represents a disease (Barabási et al 2011)

A recent paper by Ravindran et al (2019) elucidated the manner by which viruses exploit human systems. Using network analysis, they were able to identify 'hub' genes involved in important cellular processes which, when targeted by viruses allow them to wrest control of the cell (Ravindran et al., 2019). In the investigation of heritable disease Zhong et al (Zhong et al., 2009) found that, whilst loss of function of gene product nodes are strongly linked to mendelian disorders, the disruption of interactions, termed 'edgetic perturbations' are also associated with both mendelian disorders and complex relationships between genotype and phenotype, yielding quite different consequences to that of node removal. Since this time Networks have proven their worth in the attempt to resolve the relationship between genotype and phenotype, providing a mathematical framework for the analysis of often complex systems

Barabási et al (2011) reviewed contemporary methods of network analysis in relation to the human disease network (Figure 2). They

established that methods of predicting disease genes fall into three categories: diffusion-based methods, which algorithmically calculate the likelihood of a gene being associated with disease, based on interaction with a known disease gene product; linkage methods, wherein genes are predicted based on their observed interaction with known disease gene products; and disease module-based methods, which use functional and interaction information to predict disease genes. When assessing the utility of these methods it was found that a comparison between linkage and diffusion-based methods on the same data showed the latter to be a more powerful tool (Navlakha and Kingsford, 2010).

More recently updated mathematical methods have been incorporated into the analysis of networks. Whist useful in its time, analysis no longer depends on the arbitrary counting of interaction partners to define important nodes, nor does it rely on already having knowledge of network properties in order to predict important nodes or subunits. One such new method is based on control theory, with the addition of minimum dominating sets (MDS), briefly explained in figure 3, to assign critical, intermittent and redundant nodes, control theory explores topological features of networks, and establishes network control capabilities, of all nodes in the network (Lombardi and Hörnquist, 2007; Nacher and Akutsu, 2016; Ravindran et al., 2019).

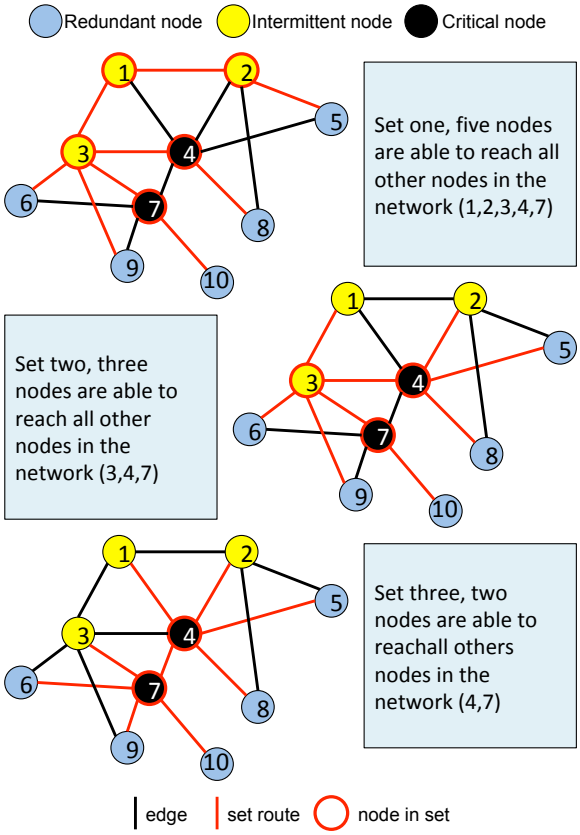


Figure 3: Example of critical, intermittent and redundant node assignment in complex biological networks using control theory and minimum dominating sets. Network traversal routes are shown as red lines, 'active' nodes, acting as conduits are outlined in red. Critical nodes, those that are integral to systemic control, are shown in black, intermittent nodes, which are inconsistently found to be critical, are shown in yellow, and redundant nodes which are not found to be essential to control, are shown in blue.

A tool more commonly associated with engineering and mathematics, control theory aims to elucidate the dynamic nature of complex systems and identify key components of these systems (Del Vecchio et al., 2018; Kremling, 2013; Tsongalis, 2018). Recently applied to biological problems, control theory as used to explore biological networks has provided strong evidence for both plasticity and essential components (de Anda-Jáuregui et al., 2018; Peyraud et al., 2018; Ravindran et al., 2019), and how this control reconciles with evolution (Badyaev, 2019).

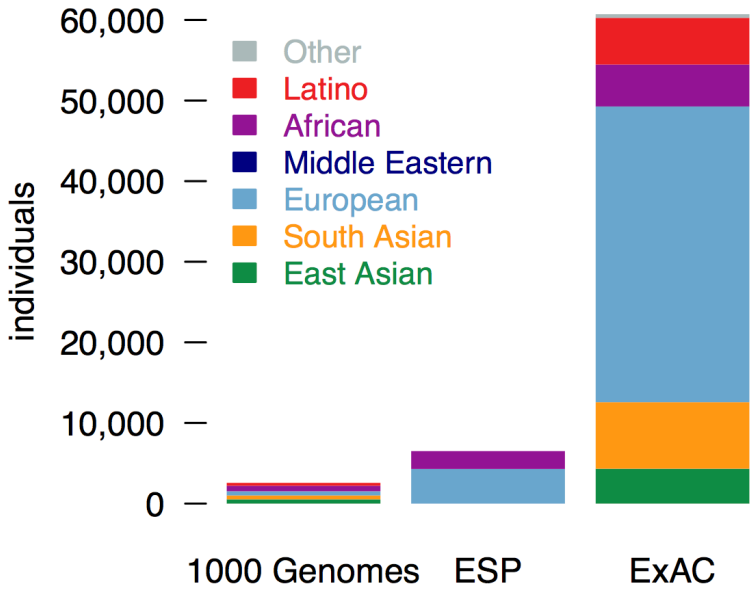


Figure 4: Number of genomes in each dataset by population (Lek et al)

Machine Learning Methods

Machine learning methods are a series of computational applications of varying complexity ranging from the simple, such as clustering techniques, to complex, such as deep learning algorithms. They can be grouped into three forms; supervised, wherein the user provides both input and output data, and the algorithm determines the route between the two (Kotsiantis, 2007; Libbrecht and Noble, 2015). Unsupervised learning, which does not require output data, rather the algorithm, determines the underlying structure of the data and draws inferences directly (Libbrecht and Noble, 2015); and semi-supervised learning, wherein only some of the data is labelled and therefore a combination of supervised and unsupervised learning is required to determine structure and output (Libbrecht and Noble, 2015).

These methods have, over the last decade been proven to assist in inference and prediction drawn from large-scale multi-dimensional biological data, either alone, or in conjunction with other methods such as network analysis (Bzdok et al., 2018; Camacho et al., 2018). It is this ability to handle large data, in an era of post-genomic biology, wherein there has been a vast increase in the quantities of data generated each year (Che et al., 2013; Jin et al., 2015; Kaisler et al., 2013), that sets machine learning methods apart from more traditional statistical methods, as these often struggle with large data (Bzdok et al., 2018).

Despite their obvious benefits to biological analysis however, machine-learning methods have some disadvantages when compared with more conventional statistical analyses. Most notably, the difficulty in equating their results with the broader biological context (Bzdok et al., 2018), which can often lead to difficulties in interpreting their results (Libbrecht and Noble, 2015). Machine learning algorithms are not, however completely distinct from established statistical methods, as a number incorporate tried and tested statistical techniques in order to produce robust predictions (Bzdok et al., 2018).

For example, random forests, one of the more commonly used machine learning methods in biology relies on bootstrapping, to inform its predictions and determine accuracy (Bzdok et al., 2018; Paul et al., 2017; Z. Wang et al., 2015), which has a proven utility in traditional statistical analysis and phylogenetics.

Machine learning has, over the last decade, revolutionised the study of biology, particularly with reference to the prediction of disease

associations in biological systems (Camacho et al., 2018; Chen et al., 2018; Gao et al., 2018; Haq et al., 2018; Kannan and Vasanthi, 2019; Nilashi et al., 2018; Poplin et al., 2018), and as a tool still in its relative infancy, this is likely to continue.

Nomenclature and data standardization

Issues with nomenclature, and a lack of standardization, are rife in pre-existing datasets and analyses. This, to a degree, is being addressed by projects such as the HGNC (Braschi et al., 2019; Stenson et al., 2017); however, standardization has by no means been adopted universally. Vihinen (Vihinen, 2015) also discuss inconsistency in reporting within pre-existing literature, for example the use of ambiguous or nuanced terms, which may lead to contrasting interpretations by different readers. Of particular significance is the misuse of terms which specifically relate to either DNA, mRNA, or proteins, being used incorrectly (i.e., a DNA term being used for proteins), this, as highlighted by Vihinen (ibid) can easily lead to misinterpretation, not only by researchers, but more so by automated analytical engines commonly used in large scale text mining projects, as these programs are less able to pare out the intended meaning based on linguistic context clues.

Datasets

Over the last two decades, as sequencing methods have become cheaper, faster, and more precise (The UK10K Consortium, 2015) there has been a deluge of data deposited in both publicly available, and proprietary datasets. The 1000 genomes project, combining open-source data and

tools to access it (The 1000 Genomes Project Consortium et al., 2012), was one of the pioneers in data redistribution, with 2,504 human genomes made publicly available. Since this time large-scale publicly available data repositories such as GnomAD (Karczewski et al., 2019), and proprietary health data such as UK biobank (Sudlow et al., 2015) and the 100k genomes project (Caulfield et al., 2019) have radically increased the data available. One of the major advantages of this increase is the ability to identify variants, which would have otherwise been missed when analysing fewer individuals. However, as has previously been noted, sequencing projects have largely preferentially focused on individuals of European ancestry (Exome Aggregation Consortium et al., 2016).

There are a wide variety of sequence types available, ranging from whole genome sequences, exomes, and epigenomes, to individual genes, each of which has its own merits and pitfalls, such as exclusion of non-coding regions, specific focus on certain disease data and sequencing shortcomings, which are too numerous to detail within the scope of this report.

As the ability to sequence living organisms has improved, so too has our ability to extract DNA from both museum specimen (Wood et al., 2018) and archaeological remains (Damgaard et al., 2015; Hofreiter et al., 2015; Leonardi et al., 2016; Slatkin and Racimo, 2016). A publicly available dataset of ancient NHH data having recently been released, to aid in answering questions, both ancient, and modern (reich.hms.harvard.edu Reich, 2019).

One of the major obstacles to the selection of data is accessibility. This is in some cases owing to certain datasets, due to the significant financial outlay that they can represent, not being open publicly, or requiring subscription. Further issue to the studies themselves, often requiring the proprietisation of subsequent data are the ethical implications of sequencing individuals. Predominantly participant's genomes have been anonymised, which, whilst an important ethical requirement, can lead to an inability to associate genomic data with patient information. Studies such as the 100k genomes project, which aim to provide genomic data alongside corresponding medical information, have been heavily scrutinised due to the potential loss of patient anonymity, despite their inherent value to the identification of disease variants within individuals whom exhibit phenotypic indicators.

Further issues are engendered by the existence of patented proprietary data, specifically pertaining to cases such as the high-profile patenting of BRCA genes by Myriad Genetics. The monopoly initially held on sequencing of the BRCA genes allowed the accumulation of large quantities of information relating to other variants in patients' genomes by a single company, which, whilst the BRCA genes are no longer covered by an active patent, will remain proprietary (Conley et al., 2014).

Of the datasets currently available, many of the largest contain consolidated and 'treated' data from other sources, which in many cases are not available individually. Some of the most high profile of these open source datasets, which have been used extensively in recent years

are ExAC (Exome Aggregation Consortium et al., 2016), containing 60K+ exomes from 17 contributory projects; The European Genome-Phenome archive (Lappalainen et al., 2015), containing 1,624 datasets comprising of 574,120 samples; UK10K (The UK10K Consortium, 2015), with 16,637 samples from 32 projects; The ClinVar database (Landrum et al., 2018), containing 172,053 submissions from 416 contributors; GnomAD (Karczewski et al., 2019) containing 125,748 exomes and 15,708 genomes; and the Leiden Open Variation Database (Fokkema et al., 2005) with more than 22,080 entries from across the globe.

The maintenance and storage of such datasets, however, pose a significant financial outlay; therefore, despite their obvious medical and scientific value, the futures of open datasets are not always certain. Over recent years the funding to support a number of the most highly accessed variant databases has been cut (Kaiser, 2016). It is not certain how this will impact on their futures in the long-term, however it would be a travesty for such valuable resources to be lost, on the basis of fiscal concerns.

Despite the large quantities of data now at our disposal, and additions to this data on the horizon, as observed by the UK10K consortium (The UK10K Consortium, 2015), without the appropriate tools to analyse, identify and clinically assess variation within the human genome our ability to draw meaningful conclusions will remain limited. It is clear that the human genome and epigenome are prone to vicissitude. Variants, both small and large are numerous and diverse, with ramifications that range from foetal inviability, to zero phenotypic

change. The analysis of variation and its role in human disease has reached a crossroads. Whilst, as has been highlighted, the rapid rise in available data has led to much advancement, it has also been the catalyst for an excess of issues. The large quantities of data now available to scientists, combined with pressures to 'publish or perish', has facilitated a flood of publication over the past decade, particularly following the advent of high throughput sequencing. Whilst much of the data presented has been beneficial, there are high quantities of publications with often flawed or conflicting results. Rather than affording insight into the field of variant analysis, this overabundance has served to cloud it further. Perhaps as a repercussion of this, the erroneous use of terminology, and proliferation of disparate nomenclature relating to the same entities has occurred, which it is imperative that we quell through standardisation. Compounding this further are a number of analytical issues. Firstly, limitations posed by the inability to segregate individual patient data within certain disease datasets. Further to this, and of particular significance to the study of variation and disease, is poor quality, or a complete absence of phenotypic data, with relation to sequence information. As has been discussed throughout, it is imperative, when analysing biological data of any kind to always consider the broader biological context. This can impact not only the conclusions drawn but is also a key governing factor in the initial data selection process, and the tools and methodologies used. The methodological issues however, it has been shown, may in certain cases be mollified by the use of multiple strategies in combination with one and other, however, inherent biases in

analytical techniques must still be accounted for. The final, and perhaps most profound issue raised within the literature presented here, is the repeated, and often arbitrary functional misclassification of variants, the most abundant of which being the classification of all missense variation as affecting function. This unnecessarily binary view of these classes of point mutation opens the door for both type one and two errors to be easily made. It would be obtuse to claim that attribution of functional consequences of variation is straightforward, far from it, however rather than the erroneous binary attribution of function, a more varied framework of 'expected functional classification' needs to be devised.

In contrast to this bleak view of variant analysis there are emergent areas, likely to facilitate profound advancement and far greater precision. Four specific points, which are likely to lead to this advancement are; single molecule sequencing, epigenetic analysis, large-scale combinatorial evolutionary studies, and the publication of comprehensive variant databases with links to health data. There is evidently already an abundance of data emerging, however, new sequences garnered from technologies such as single molecule sequencing will serve to supplement this with greater precision and clarity. A further advantage of new sequencing technologies will be the ability to cross-reference new, with existing data, and provide clarity and correction to prior analyses. It is evident that as we are able to more fully understand the role of the epigenome and the mechanisms by which these modifications occur and are transmitted, that this avenue of investigation will also prove invaluable to the study of genomic disease, particularly

within evolutionary and adaptive contexts. As, whilst still in its relative infancy, it has provided much insight into previously undetectable variants leading to aberrant phenotypes.

Large-scale and structural variations have clearly played a profound role in human evolution, having been instrumental in speciation events, and facilitation of increased complexity, however, it is also clear that they represent significant association with human disease susceptibility. The analysis of variation and genetic mutations has come a long way since Goldschmidt first presented his hypothesis almost 100 years ago. Goldschmidt's original hypothesis has been substantiated, as presented here, at the very least, we are clearly descended from hopeful monsters, however the ensuing question as-yet remains unanswered; given the extent of both small, and large-scale variation in the human genome, and the potential for negative implications that newly accrued variants pose to human health, are we still the hopeful monsters?

It is the aim of this thesis to better understand the role of gene duplications in human disease, and more specifically, the implications for human evolution of being descended from whole genome duplication events in early vertebrates. Each of the three research chapters will address this aim via covering a number of objectives. Firstly, in chapter 3, we explore the role of duplicates arising as a result of both whole-genome, and small-scale duplication events in human heritable disease. Secondly, in chapter four, we will look at the potential role shared ancestry may have in the perpetuation of the host-pathogen relationship between the parasite *Trichuris Trichura* and humans, via a deep

evolutionary relationship between genes in the TGFβ superfamily, and a *T.trichura* homolog. And lastly, in chapter five, we shed light on the interplay between paralog status, virus interactions and heritable disease in the human PPI network, and how these features relate to the control of biological systems.

CHAPTER TWO

"Among the current discussions, the impact of new and sophisticated methods in the study of the past occupies an important place"

-(Fogel, Robert William, and Elton Geoffrey Rudolph., 1983)

GENERAL METHODS

A ‘preliminary’’ dataset is used as the foundation of the three analyses presented herein. This data pertains to gene age, Disease status, paralog status, haplosufficiency, and asymmetry, on a per-gene basis. The methods for the generation of this data are presented here and may be referred to when discussing the ‘preliminary’ data (Table 1).

Table 1: Example 10 row excerpt from the preliminary data compiled using the 'General methods'. Data and generation code available at: https://github.com/AlexMartinGeary/Hopless_Monsters/blob/master/Universal_Methods/Preliminary_dataset.csv.

Preliminary_dataset

Gene_id	Gene_Name	Ortho_Age	Paralog_status	Haplosufficiency_Rank	PPI_connectivity	Disease_Association	Family_Root	Family_ID
ENSG00000000003	TSPAN6	615	Ohnolog	45.94	834	Unknown	Bilateria	1
ENSG00000000938	FGR	615	Ohnolog	5.28	3118	Unknown	LECA	5
ENSG00000001036	FUCA2	796	Ohnolog	64.33	938	Unknown	Vertebrata	7
ENSG00000001461	NIPAL3	435	Ohnolog	51.92	360	None	LECA	9
ENSG00000002746	HECW1	435	Ohnolog	28.13	1834	Unknown	LECA	20
ENSG00000002834	LASP1	796	Ohnolog	32.25	1424	Unknown	LECA	21
ENSG00000003096	KLHL13	615	Ohnolog	22.8	0	Unknown	LECA	24
ENSG00000003137	CYP26B1	615	Ohnolog	20.72	1304	Unknown	LECA	25
ENSG00000003147	ICA1	796	Ohnolog	36.19	972	Unknown	Vertebrata	26
ENSG00000003756	RBM5	615	Ohnolog	20.16	1746	Unknown	LECA	32

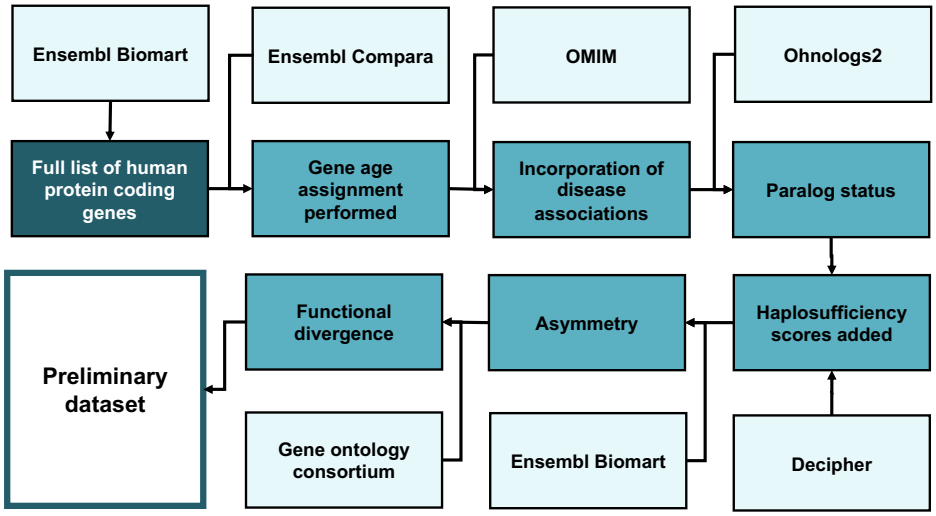


Figure 5: Flow chart showing the generation of preliminary data. Dark blue shows the template gene list, each analytical step is shown in medium blue, and light blue shows the sources of additional data incorporated at each step, culminating it the 'preliminary dataset' shown in the white box with dark blue border.

2.1 Assigning gene age

In order to assign ages to the genes, most recent common ancestors (MRCA) were calculated. A list of human protein coding genes was obtained from Ensembl BioMart release 97 (GRCh38.p7). Orthologs of these genes were obtained from Ensembl using the Ensembl Compara homology Perl API (Herrero et al., 2016) with output limited to the following:

Table 2: Example 10 row excerpt from initial homology-relationship table obtained from Ensembl compara. Data and code used to generate it available at https://github.com/AlexMartinGeary/Hopless_Monsters/blob/master/Universal_Methods/Ortho_out.csv

Ortho_out

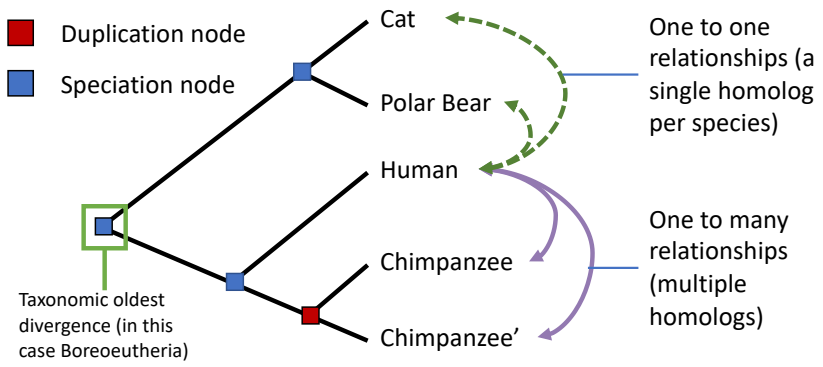
Gene_id	Taxon_between	Type	Taxonomic_level	is_high_confidence	is_tree_compliant	GOC_score	WGA_coverage
ENSG00000164404	H.sap-P.pan orthologues	ortholog_one2one	Homininae	1	1	100	100.00
ENSG00000164404	H.sap-P.tro orthologues	ortholog_one2many	Homininae	1	1	100	100.00
ENSG00000164404	H.sap-P.tro orthologues	ortholog_one2many	Homininae	1	1	100	100.00
ENSG00000164404	H.sap-G.gor orthologues	ortholog_one2one	Homininae	1	1	100	100.00
ENSG00000164404	P.abe-H.sap orthologues	ortholog_one2one	Hominidae	1	1	100	100.00
ENSG00000164404	H.sap-N.leu orthologues	ortholog_one2one	Hominoidea	1	1	100	100.00
ENSG00000164404	H.sap-C.pal orthologues	ortholog_one2one	Catarrhini	1	1	100	100.00
ENSG00000164404	H.sap-P.anu orthologues	ortholog_one2one	Catarrhini	1	1	100	100.00
ENSG00000164404	H.sap-C.sab orthologues	ortholog_one2one	Catarrhini	1	1	100	100.00
ENSG00000164404	H.sap-C.aty orthologues	ortholog_one2one	Catarrhini	1	1	100	100.00

Homology type as shown in figure 6A and Table 2 was restricted to - “ortholog_one2one” wherein there is a single identified gene in the homology relationship between the query taxa (in this case human) and target taxa; “ortholog_one2many” wherein there is a single extant gene in the homology relationship between the query taxa (in this case human), but multiple othologs are present within the target taxa, thus accounting for paralogs arising in the target taxa following the shared speciation event.

For stringency – output was restricted to recorded relationships that are defined by Ensembl as high confidence of true orthology, based on a combination of gene order conservation scores, and whole genome alignment score.

Further to this, any orthologous relationships that were identified as not being compliant with the Compara protein tree, for example those that are difficult to resolve due to gene loss, or confounded by dubious duplication nodes, were excluded. An example of the resultant output can be seen in figure 6B.

A)



B)

Homologous gene	Species between	Type	Taxonomic level
Cat_homolog_1	Human:Cat	One to one	Boreoeutheria
Polar_bear_homolog_1	Human:Polar_bear	One to one	Boreoeutheria
Chimpanzee_homolog_1	Human:Chimpanzee	One to many	Hominines
Chimpanzee_homolog_2	Human:Chimpanzee	One to many	Hominines

C)

Gene	Taxonomic level of oldest divergence	Age of divergence
Human_gene_1	Boroeutheria	96.5 MYA

Figure 6: Graphical representation of the age assignment process. A) Tree showing relationships between homologs. A one-to-many relationship (indicated here with purple solid lines) is found where duplications occurred following the divergence between species. One-to-one relationships (shown with green dashed lines) exist where no further duplications have occurred following divergence. Figure modified from Ensembl. B) Example of the relationship table obtained in step 1. C) Example of the final assignment of age resulting from data shown in A0 and B).

For each gene a presence/absence matrix was subsequently generated recording each taxonomic level in the human lineage within which an ortholog had been identified. This was output alongside stable ID and gene name, and a relative age of first identified speciation node, corresponding to the speciation point of the earliest taxonomic level with a recorded ortholog (MRCA).

2.2 Disease status, paralog status, and haplosufficiency

In addition to gene ages, data was gathered pertaining to disease association, gene age and paralog status of all 'known' genes. For paralog status, a list of ohnolog pairs within the strict and intermediate categories having q-scores for both self-comparison, and outgroup of <0.001 and 0.01 respectively was obtained from Ohnologs-2 (ohnologs.curie.fr) and compared with a list of human paralogous genes and their partners obtained from Ensembl BioMart (Zerbino et al., 2018). Genes found in the ohnolog list were defined as ohnologs, any genes not in this list, but present in the Ensembl paralog data were determined to be SSDs, and all remaining genes classed as singletons.

Data pertaining to disease association were obtained from OMIM in the form of the Genemap2 dataset (https://data.omim.org/downloads/dI1aeTBYTNet3PfqZqIS_w/genemap2.txt). Information contained therein used in the current analysis was compiled by OMIM from the following sources: Ensembl gene id – Ensembl; Phenotype data – OMIM. A text-based search was performed on this file to establish, firstly, if each gene in our 'known' set existed within the OMIM dataset, and if so, whether their association was with dominant, recessive or unknown disease inheritance, subsequent assignment of disease status was conducted according to these findings. For rare instances where genes had both dominant and recessive associations their status was defined as 'both' and genes not present in Genemap2, were defined as 'none'.

Haploinsufficiency (HI) scores (Huang et al., 2010) were obtained from Decipher (Firth et al., 2009) (Haploinsufficiency Predictions Version 3 bed file). Due to the fact that the entries in this file listed genes using HGNC identifiers, to bring them in line with the Ensembl naming convention used in the analysis it was necessary to cross reference gene names with their Ensembl counterparts. To this end Ensembl GRCh38.p7 gene names and HGNC IDs were obtained from Ensembl BioMart (Zerbino et al., 2018), and, where possible HGNC IDs substituted in the Haploinsufficiency data, for Ensembl gene IDs. Haplosufficiency scores and ranks were then obtained for each 'known' gene. A consolidated dataset was then created (Data S1), containing Ensembl gene name, disease status, paralog status and rank, which provided the basis for the initial analysis of gene age, paralog status, disease status and haplosufficiency.

2.3 Asymmetry

In order to investigate asymmetry a further dataset was compiled from data obtained from the above sources, specifically pertaining to genes in ohnolog pairs. For this, comparisons were made on a per pair basis. Each gene in the pair was named Conserved 'C' or Diverged 'D', where gene C had the lowest dN/dS ratio when compared with chimpanzee as obtained from Ensembl biomart, and was therefore considered the most conserved, and gene D the greatest ratio, and therefore considered the most diverged. Pairs with no divergence were

excluded. Disease association per pair was then calculated, and disease associations previously listed as 'both' were considered dominant.

To further detect asymmetry between the pairs a novel method of functional dissimilarity was devised to provide a dissimilarity metric to genes in paralog pairs (Table 3). The complete list of human gene ontology (GO) annotations was downloaded from the gene ontology consortium (Ashburner et al., 2000), on 06/08/19, these were then stored for each gene, and compared between genes in each pair. For each term not present in the other gene in the pair, the functional divergence score was increased by one. Any pairs with a proportional difference of less than 25% of the total recorded annotations for that pair was considered not to exhibit true asymmetry, and therefore dropped from the asymmetry data. Further to this, any pairs with a dissimilarity in dN/dS ratio of less than 0.0005 were also excluded.

Table 3 : Table showing the functional divergence scoring system method, using examples of two ohnologs pairs taken from the TGFβ superfamily. Corresponding gene ontology terms are highlighted in green, conflicting terms are highlighted in orange. The cumulative score is retrieved, and proportional score calculated (right).

GDF9	BMP15	Cumulative Score	Proportional score
GO:0005125 : enables cytokine activity	GO:0005125 : enables cytokine activity	0	0%
GO:0008083 : enables growth factor activity	GO:0008083 : enables growth factor activity	0	
GO:0070698 : enables type I activin receptor binding	GO:0070698 : enables type I activin receptor binding	0	
BMP8A	**BMP8B**	**Cumulative Score**	**Proportional score**
GO:0005576 : extracellular region	-	1	100%
GO:0005615 : is_active_in extracellular space	-	2	
-	GO:0005125 : enables cytokine activity	3	
-	GO:0008083 : enables growth factor activity	4	
-	GO:0070700 : enables BMP receptor binding	5	

Unless otherwise stated, computational analysis of data was performed using Perl 5, version 18, subversion 2 (v5.18.2) (Wall, 2000),

with all further analysis and image generation conducted using R version 3.3.2 (2016-10-31) "Sincere Pumpkin Patch" (R Core Team, 2016), and the following libraries; ggplot2 (Wickham, 2009) ggfortify (Yuan T, 2016); dplyr (Wickham et al., 2017); lattice (Sarkar, 2017); plyr (Wickham, 2016); raster (Hijmans et al., 2017) ; gridExtra (Auguie and Antonov, 2017); tidyr (Wickham et al., 2018a); cluster (Maechler et al., 2018); FactoMineR (Husson et al., 2018) ; Devtools (Wickham et al., 2018b); factoextra (Kassambara and Mundt, 2017), and, corrplot (Wei et al., 2017). The graphical representation of gene age method in Figure 6 was generated using Microsoft PowerPoint.

CHAPTER THREE

"We must, however, acknowledge,

as it seems to me, that man with all his noble qualities…

still bears in his bodily frame the indelible stamp

of his lowly origin."

– Charles Darwin (1872)

GENE & GENOME DUPLICATION: DUPLICATION IS A CORRELATE OF MONOGENIC DISEASE, NOT A CAUSE

3.1 Abstract.

The ability to predict disease association in human genes is enhanced by an evolutionary understanding. Interestingly, genes readily linked with heritable disease, particularly dominant disorders, have frequently undergone whole genome duplication (WGD) in our early vertebrate ancestors, with a strong asymmetric relationship wherein it is likely that just one gene in each pair will be disease associated. Using whole-genome comparative analysis of evolutionarily significant features, we show that contrary to the accepted compensatory model of disease evolution explaining this association, the majority of disease-associations reside with the more evolutionary constrained gene, inferred to resemble the duplicates' progenitor most closely. This indicates that the strong association between paralogs, specifically ohnologs, and dominant disorders is often a consequence of a mechanism (WGD) through which pre-existing dosage sensitive/haploinsufficient genes are successfully duplicated and retained, rather than their inherent 'dangerousness'. Heritable disease is thus as much a consequence of the fragility of evolutionarily more ancient genes as compensatory mechanisms. From these findings, we demonstrate the utility of a new model with which to predict disease-associated genes in the human genome.

Key words: Asymmetry, Ohnologs, SSDs, function

3.2 Introduction

More than 150 years since Gregor Mendel's findings were first published, the role of evolution and heritability in human genetics, and specifically disease association, is still, by necessity, being explored. We now know that whilst Mendel's foundation of dominant versus recessive alleles is broadly correct, a myriad of factors contribute to both penetrance and phenotypic severity, but also that complex genetic interactions, where the interplay between multiple genes, rather than the simple 'one gene- one trait' model is the norm (Cooper et al., 2013). In terms of mono-genic disease association, a striking association has been found to exist between genes duplicated in vertebrate evolution and disease with 80% of human heritable disease genes residing in a paralog in the human genome (Dickerson and Robertson, 2012). Why does this strong one gene to phenotype relationship exist with genes that have a duplication event in their history?

While 'duplicability' of a gene can be influenced by genomic context, such as sequence composition and chromosomal location, and accessibility, leading to much of the copy number variation observed within the human genome (Schuster-Böckler et al., 2010; Truty et al., 2018), differential retention biases have played a much more significant role in the landscape of paralogs observed in genomes particularly when small- and larger-scale duplication events are compared (Hakes et al.,

2007). Notwithstanding stochastic sampling processes in evolution associated with smaller population sizes (Lynch and Conery, 2003), this will often be a result of factors associated with gene product fitness. Whole genome duplication in particular, is hypothesised to increase the chance of retention of dosage-threshold sensitive genes, as a consequence of maintenance of stoichiometric balance in the cellular system (Papp et al., 2003), which, due to its negative/deleterious impact, would be very unlikely to be retained within the context of small-scale duplication (SSD) events (Makino and McLysaght, 2010). This is supported by the observation that copy-number associated duplicates tend to be refractory for the retention of dosage-sensitive paralogs (Rice and McLysaght, 2017a).

In order to disentangle the relationship between gene duplication and disease it is imperative to have an understanding of the processes by which genes arise and are retained (Innan and Kondrashov, 2010). Exemplifying this perspective is the strikingly strong association observed between genes which have undergone duplication in vertebrate evolutionary history, in particular WGDs, and heritable human disease (Makino and McLysaght, 2010). Following a 'compensation model', where duplicates contribute to redundancy (Gu et al., 2003), we previously hypothesised that this relationship was due to the accumulation of otherwise deleterious mutations in the context of duplication events (Dickerson and Robertson, 2012), with disease associations emerging as new genes arise and are retained, i.e., duplication introducing disease potential by 'masking' of otherwise

deleterious mutations. Singh et al (Singh et al., 2012) presented a modified version of the compensation model, in which the compensation for deleterious variation, in particular for dominant disorders, is 'locked in' by the WGD event, and subsequently neither gene in the duplicate pair can be lost without severe consequences to fitness. Furthermore, it has been proposed that whole-genome duplication (WGD) events confer an immediate fitness benefit to the organism by reducing expression 'noise' (Pires and Conant, 2016).

What is often neglected, however, is the asymmetric relationship of disease within gene families/paralogs (Dickerson and Robertson, 2012), which must be considered in any analysis. Due to random mutations (neutral or functional/adaptive), following a duplication event the two paralogous genes are unlikely to remain the same for long. Over time the accumulation of variants in either or both of the duplicates leads to divergence. This divergence can result in differing functions of the two genes, sub- or neo-functionalization, or pseudogenization amongst others. The relative proportions of these outcomes is a contentious subject, with conflicting theories surrounding not only differential retention, but also the likelihood of and degree to which evolutionary asymmetry may occur (Pachter, Lior, 2015). Fundamentally, the discussion revolves around two arguments: the first, that paralog pairs tend to show asymmetry where the less constrained copy is likely to be harmless to the organism, presented by Ohno et al (Ohno, Susumu, 1970) and the second, proposed by Force et al (Force et al., 1999) that evolution following duplication is unlikely to be asymmetric. The current

consensus agrees with Ohno's proposition that evolutionary asymmetry does exist (Kellis et al., 2004), although the statistical measurement employed remains contested (Pachter, Lior, 2015), and may be dependent on the degree of dosage sensitivity within any given paralog pair (Tasdighian et al., 2017).

It has also been hypothesised that, following duplication, redundant genes confer robustness (Hakes et al., 2007) with one paralog acting as a 'buffer' to the phenotypic manifestation of deleterious disruptions in its partner gene (Gu et al., 2003; Hakes et al., 2007; Hsiao and Vitkup, 2008; López-Bigas and Ouzounis, 2004). It is proposed that this dilution of 'deleteriousness' via the addition of a compensatory partner, permits the retention of disease-associated genes, which would otherwise be subject to purifying selection. Whilst there is some argument to be made that this phenotypic masking is a relatively infrequent event, the studies that have shown this highlighted that, within Caenorhabditis elegans there is an enrichment for older, more essential genes within the set where masking does occur, with new, more plastic duplicates being unlikely to show masking behaviours (Woods et al., 2013). Whilst the accumulation of new duplicates in C. elegans has not been directly compared to human, given the high proportion of human duplicates retained following WGD, which are both ancient, and enriched for essential genes (Makino et al., 2009) we suggest that their findings are indicative that within humans the proportion of genes which exhibit phenotypic masking is liable to relatively high, with less frequent masking behaviours associated with SSDs. In both this, and a similar

study in Saccharomyces cerevisiae (He and Zhang, 2006) mechanisms of duplication were not accounted for, which would likely strongly alter outcomes between that of C. elegans and S. cerevisiae genomes, and that of humans. Given this 'compensation' model, and the known consequences of the subsequent relaxation of purifying selection more freely permitting accumulation of slightly deleterious mutations, i.e., non-lethal disease-causing mutations (Dudley et al., 2012), we would predict that disease-associated mutations would mostly be found on the less constrained gene within the pair (Dickerson and Robertson, 2012). Thus, explaining the connection between paralogs and heritable disease.

To investigate the association between evolution, the diploid nature of human genes and human disease further, we constructed a large-scale dataset of human genes, their paralog type, and haplosufficiency status, and found that heritable disease genes tend to be evolutionary ancient and associated with gene duplication, and in the case of dominant disorders found that the underlying association is due to WGD being the mechanism for duplicating haploinsufficient genes. Furthermore, we find disease-associated mutations tended to associate with the more constrained gene in WGD paralog pairs and are thus more likely to be associated with the ancestral function. We discuss these findings and the implication that disease states are due to a pre-existing ancient fragility within the ancestral genome, rather than the product of later duplication events.

3.3 Methods

The foundation data for this analysis are the primary data generated using the general methods (see chapter 1). This data pertains to gene age, Disease status, paralog status, haplosufficiency, and asymmetry, on a per-gene basis.

Family analysis

For the stages in the analysis wherein we looked at profiles of disease association, haplosufficiency and paralog status within families we obtained a list of all human paralogous genes from ensembl biomart. We grouped these genes into sets of paralogs, which were then defined as families and added the group IDs to the previously generated data. For the asymmetry data we used a perl script to calculate and extract the oldest 25% of each family based on their calculated MRCA age, for inclusion in a separate dataset, for the youngest 75% these were obtained by extracting the oldest 25% using R.

Statistical testing

In order to ascertain enrichment of disease genes in the more ancient age categories we performed statistical testing in R. To do this all genes with an age of 435MY (roughly contemporary with the last round of whole genome duplication) or older were grouped into a “WGD and earlier” category and counted, and those with an age of less than 435MY were classified as “post WGD” and counted. We then subset these groups and counted only those genes with a heritable disease status of

"Dominant", "Recessive" or "Both". The proportion of the total genes in the "WGD and earlier", and "post WGD" groups that heritable disease genes represent was then computed, and an X-squared P-value generated using the inbuilt prop.test function in R. A prop.test in R was also conducted to compare the proportion of Older SSDs that are associated with recessive disorders, and the proportion of younger SSDs that are associated with recessive disorders.

For the statistical analysis of asymmetry in genes pairs with a single disease association the pairwise proportions of gene pairs where a that gene is the more conserved/ diverged and more/less haploinsufficient were calculated, and the proportion of the total number of pairs that each of these groups represent compared using the pairwise.prop.test function in R with Bonferroni adjustment.

3.4 Results

Gene age and disease association

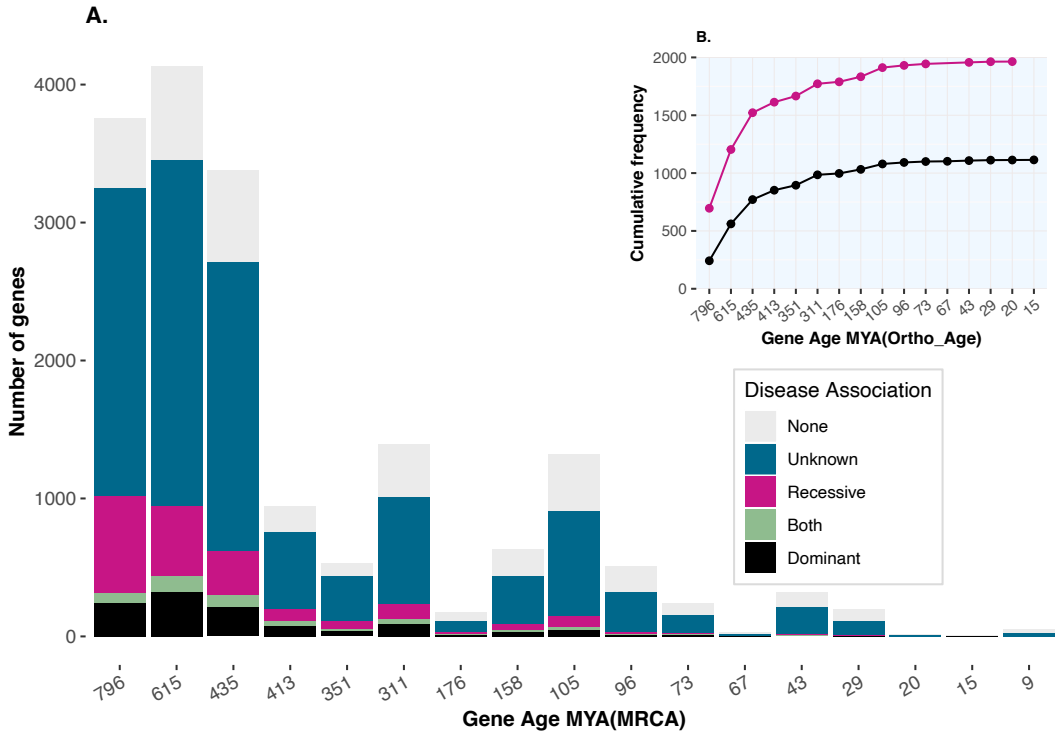


Figure 7 : Bar chart showing numbers of genes with differing disease states during our evolutionary past (inferred from taxonomic levels), showing a clear trend in older genes being disease-associated, and a spike at 311 and 105 MYA in our ancient ancestors, likely representing the diversification of the amniote line and branching of placental mammals respectively, which occurred at these time points (A). Inset: cumulative frequency of disease genes over time for dominant and recessive disease associated genes (B).

Using a method of dating genes by determining their most recent common ancestor (MRCA) (see methods), we were able to assign singleton and duplicate gene ages and compare properties of disease and non-disease genes. From this analysis, we have been able to determine that both dominant and recessive disease-associated genes, as identified within the OMIM database (McKusick-Nathans Institute of Genetic Medicine, 2018), tend to be relatively evolutionarily ancient (6A). Significantly, the

majority of these genes arose either with the two rounds of whole genome duplication (WGD, ~435 MYA in our Euteleostomi/fish ancestors before the split between cartilaginous and bony vertebrates) or predate them, with proportionally relatively few disease-associated genes arising thereafter (X-squared p= 2.2e-16). The cumulative frequency of heritable disease-associated genes within the human genome (Figure 7B) reveals a relative plateau in the introduction of obviously disease-associated genes in more recent evolutionary history, the onset of which coincides with the most recent round of WGD ~435 million years ago (MYA). This is consistent with the proposal that there has been a relatively low rate of disease associated gene introduction following the last round of WGD, with the majority of disease associated genes being evolutionarily ancient.

Haplosufficiency, disease association and gene age

Support for this finding of a high association between ancient genes and their fragility leading to disease is provided by the results of our analysis of gene age and haploinsufficiency (HI), using the Decipher haploinsufficiency scores (Firth et al., 2009) (Figure 8). Whilst haplosufficiency scores range between 1 and 100, genes falling into the *haploinsufficient* decile, as defined by Decipher are only those in the bottom 0-10 (Firth et al., 2009). These are predominantly ancient, with, in general, decreasing numbers of haploinsufficient genes arising over time. Whilst in their paper Makino and McLysaght (2010) also looked at the link between dosage and ohnologs, they did not directly measure this using haploinsufficiency rather assuming HI to be a property of the

ohnologs.

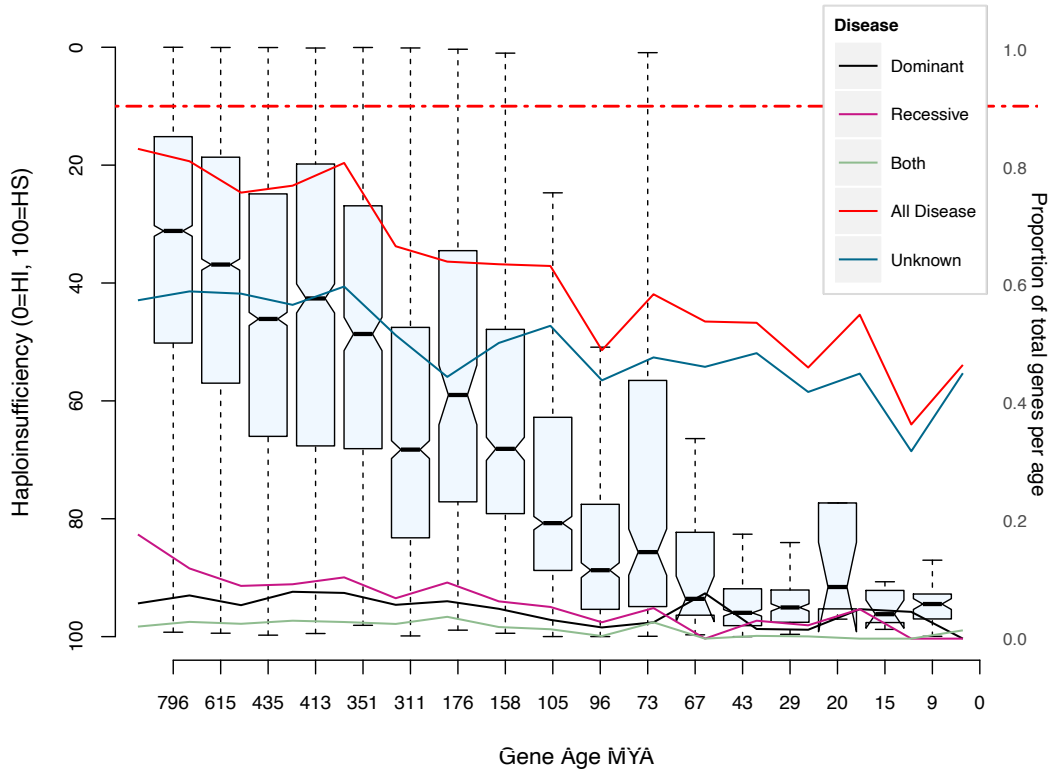


Figure 8 : Notched box and whisker plot showing haplosufficiency ranks of genes within each gene age bin, where 0 is the highest, and 100 is the lowest haploinsufficient (HI) rank. The dashed red line shows the conservative cut-off of haploinsufficiency proposed by Decipher (Firth et al., 2009), above which they predict 'true' haploinsufficient genes to reside. Overlaid is a line graph plotting the normalized frequency of disease genes in each age, between 0 and 1, arising at each time point.

Paralog status, gene age, haplosufficiency, and disease

To test the significance of this association we investigated the relationship between disease, haploinsufficiency, paralog status, and gene age (Figure 9). The multiple correspondence analysis (MCA) of these four features demonstrates a strong relationship between disease, paralog status, gene age, and haplosufficiency rank, in particular, for duplicated genes with dominant disease-associations. This indicates

haploinsufficiency is providing the underlying structure of their interactions.

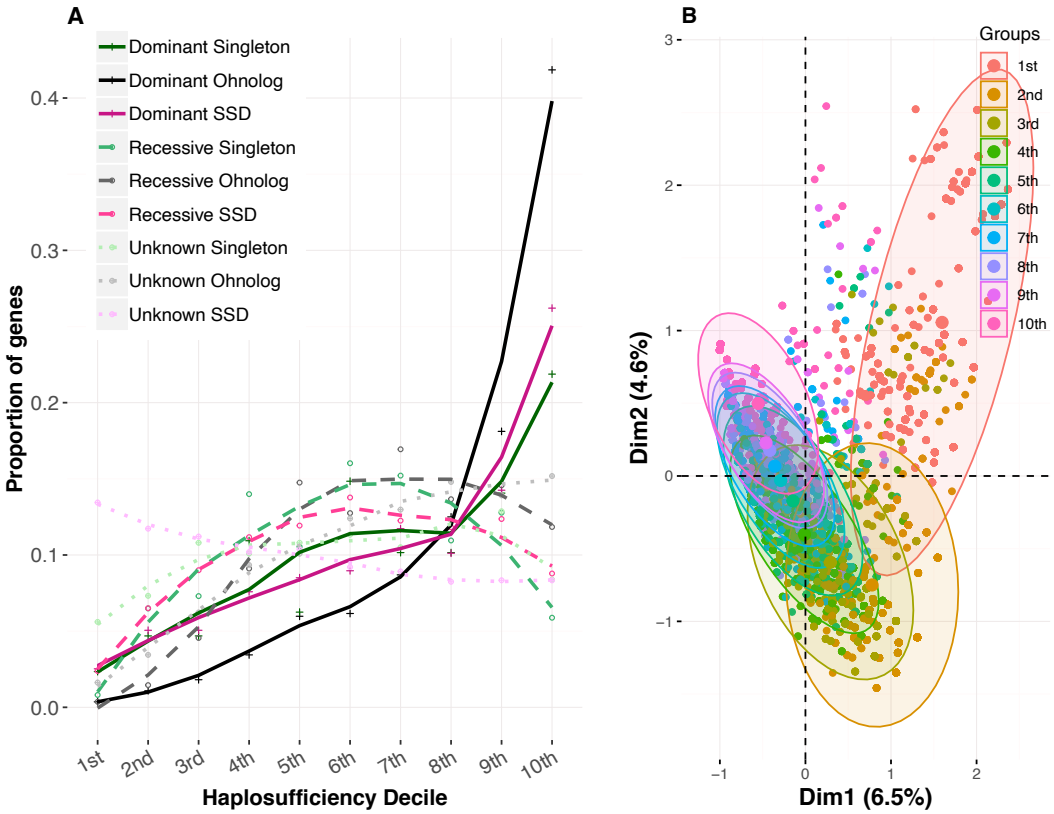


Figure 9 : Total proportion of genes from various disease and duplication states which reside in each haploinsufficiency decile (A); 1 being the least and 10 being the most haploinsufficient; solid lines show genes with dominant disease-associations, and dashed lines show those with recessive disease-associations. Multiple correspondence analysis of haploinsufficiency, gene age, paralog status and disease status (B); the 10 haploinsufficiency deciles are highlighted by the coloured ellipses.

Asymmetry in gene pairs

We tested the hypothesis that disease-association preferentially tracks to the more diverged gene within any paralog pair. In order to do this, we established relative divergence by measuring adaptive evolution: comparative synonymous to non-synonymous substitution ratios between the two genes and their closest non-human primate homologous counterparts, combined with identifying differences in functional annotation assigned to the two genes. The assumption is the functionally

diverged duplicate will have undergone more adaptive evolution, whilst subfunctionalized pairs would evolve at a similar rate to one another.

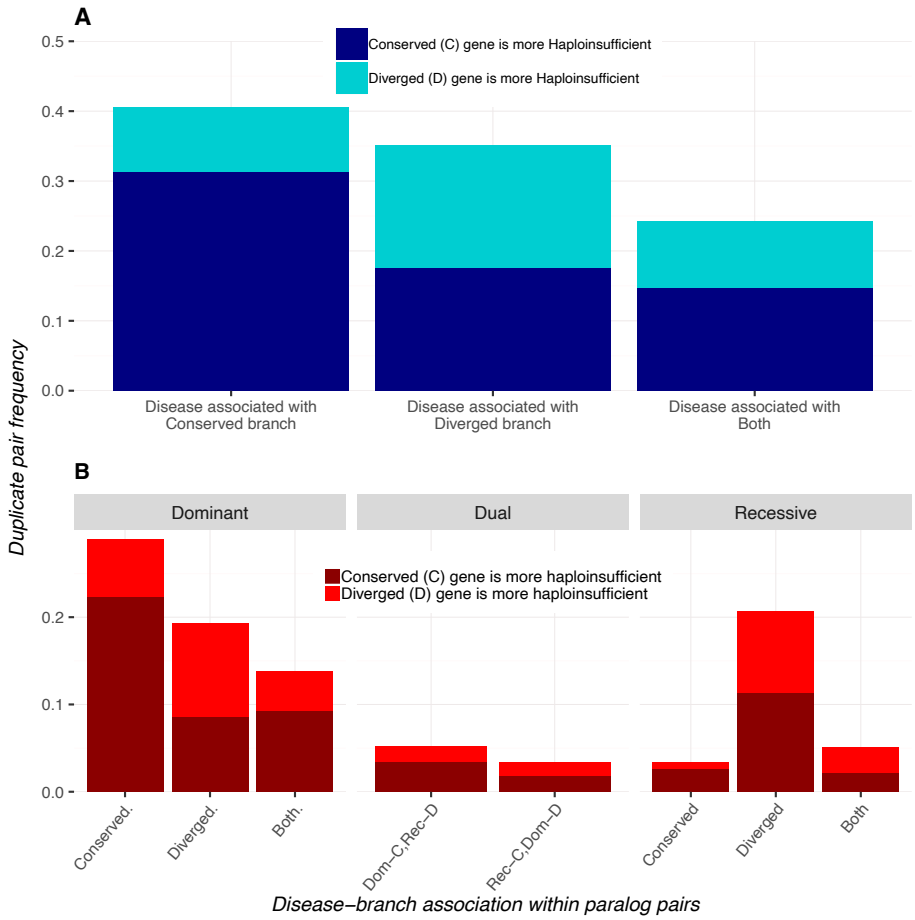


Figure 10 : Bar charts showing asymmetry between ohnolog pairs with differing disease-associations, evolutionary divergence, and haploinsufficiency for all duplicate pairs (A), ohnolog pairs by inheritance type (B). Dark bars correspond to pairs where the most conserved gene is also the most haploinsufficient; light bars are pairs where the more diverged gene is the more haploinsufficient. Disease-associations are shown on the x-axis.

Confirming our previous analysis (Dickerson and Robertson, 2012), the majority of disease associations fall on just one of the paralogs partners (76%/660 pairs), while in only 211 cases (24%) are both paralogs disease associated. Contrary to our initial expectations, however, more of the pairs containing just one gene linked with disease,

are associated with the more constrained, (54%/354 pairs), rather than the evolutionary less constrained gene (46%/306 pairs) (Figure 10A) as predicted by the disease being due to disease mutations accumulating as a consequence of relaxed selection (Dickerson and Robertson, 2012), which we have confirmed to be statistically significant (X2 P-value =0.009). This indicates that heritable disease is often a consequence of mutations disrupting the ancestral function, i.e., inferred to be retained by the more conserved paralog.

For the diseases associated with the conserved paralog these are over-represented for the more haploinsufficient genes (77%), further confirming the role of haploinsufficient genes in disease. This focus on asymmetry shows that disease propensity is most frequently associated with a combination of both haploinsufficiency and conservation, wherein the disease-associated paralog retains a greater sequence similarity to the ancestral gene than its partner.

Incorporating heritable disease type into this analysis, disease, particularly dominant disease, being associated with the conserved branch was again the case, with more tending to be haploinsufficient (Figure 10B). The major role of haploinsufficiency is supported by gene pairs that possess a dominant disease-association only in the more diverged gene (Dominant D, Figure 10), whilst these occur considerably less frequently than pairs whose disease association resides in the more conserved gene, when they do, it is predominantly the case that the disease gene also ranks more highly in the haploinsufficiency scale. Indeed, there is a highly significant trend in the increase in the number of

pairs where disease is present in only one gene, which is also both more conserved and has a greater haploinsufficiency score (pairwise proportional p-value range < 2e-16:1.5e-11).

These results indicate it is frequently the paralog that retains the ancestral function that has a tendency towards fragility and therefore disease. This asymmetric relationship confirms that disease tends in general to be associated with the more constrained/older gene function.

Haploinsufficiency, paralog status, disease association and gene families

Figure 11 shows the results of our analysis of haplosufficiency within families. It clearly shows an elevated propensity for the oldest 25% of genes arising in gene families, the most highly conserved across taxa, to be more inclined to be haploinsufficient than the younger 75%, this is especially visible within the SSD sets.

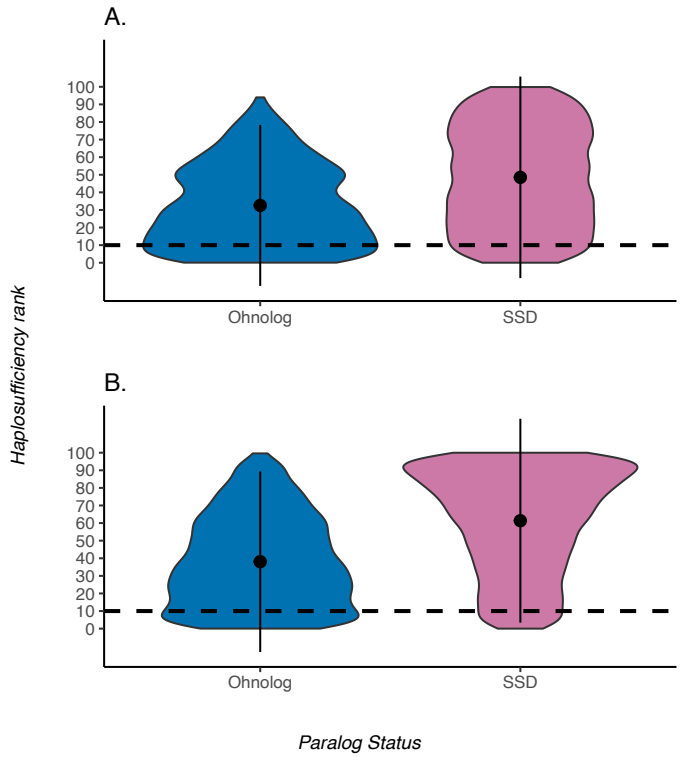


Figure 11 : Violin plots of the spread of haplosufficiency ranks in different paralog types in (A) the oldest ~25% of genes in gene families with 4 or more members, and (B) the younger ~75% of genes in families with 4 or more members. Mean and standard deviation are shown by the point and vertical lines, the horizontal lines represent the point below which decipher determine truly haploinsufficient genes to lie.

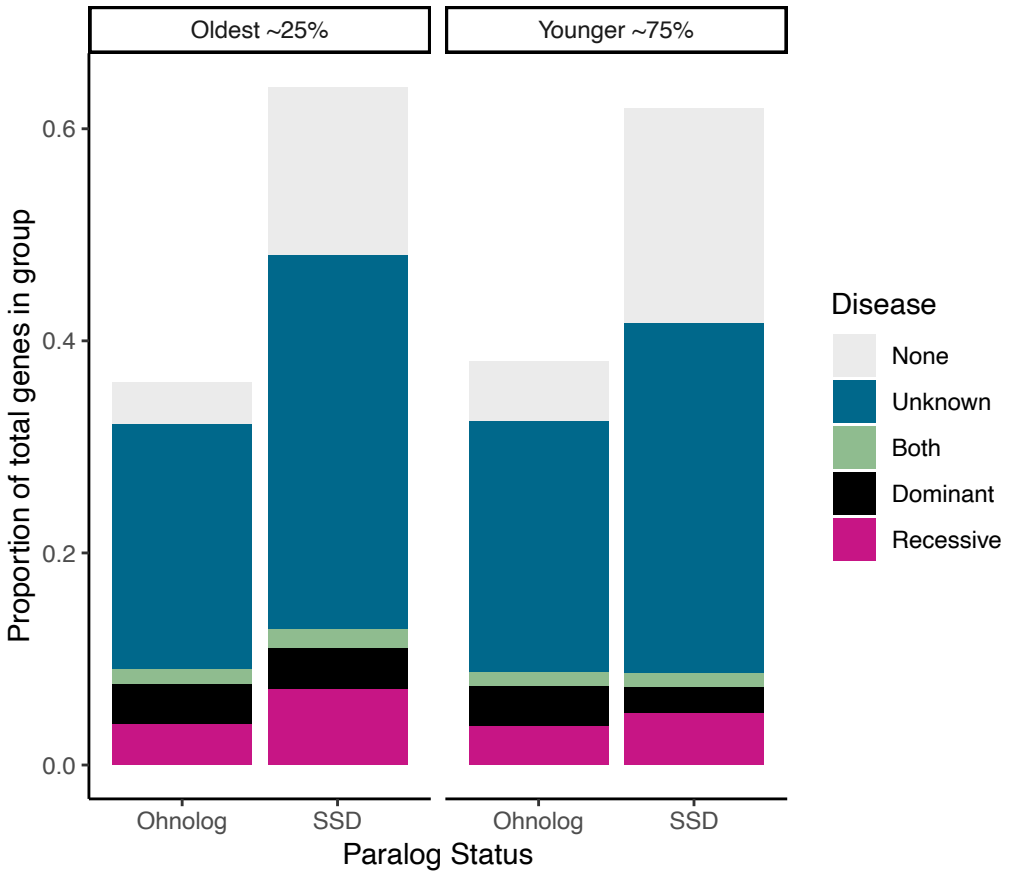


Figure 12 : Bar plot showing the proportion of genes within each age group that are disease associated, subdivided by paralog status.

As can be seen in figure 12, the proportional distribution of disease types within each age group (25% and 75%), and paralog type are broadly similar, with the exception of a marginal reduction in genes without a disease association in the older group, and interestingly, an elevation in the proportion of older SSDs, which are associated with recessive disease when compared with the younger group (P = 6.092e-05).

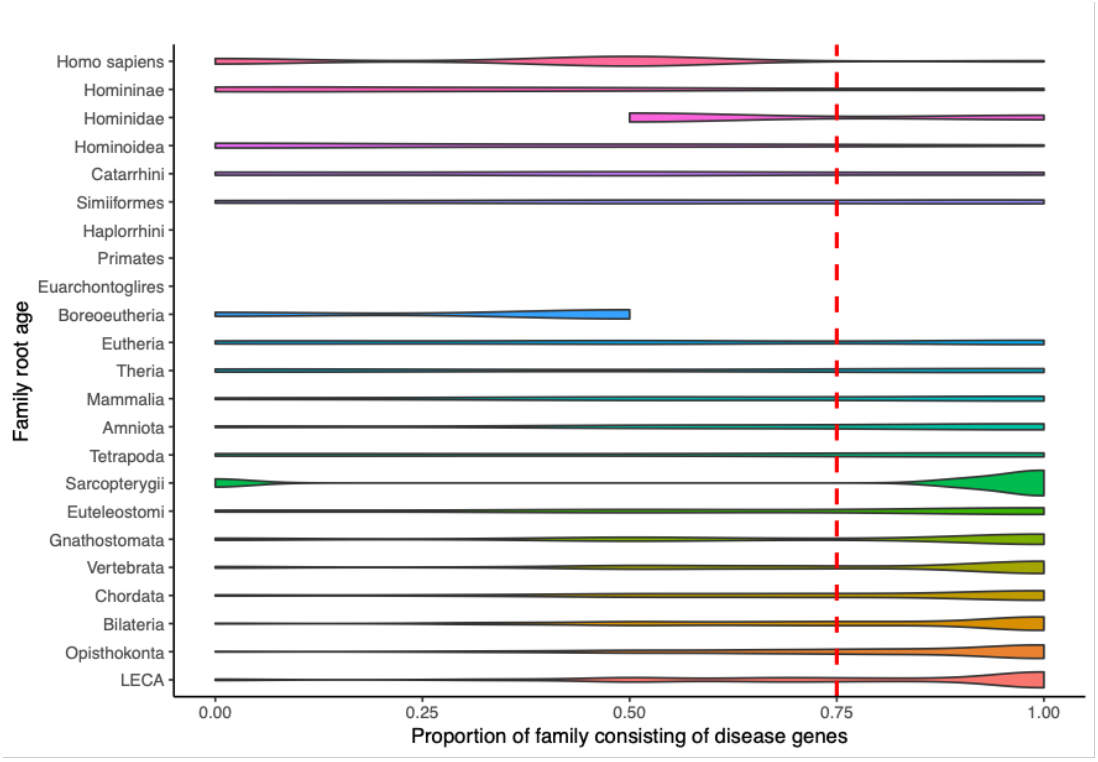


Figure 13 : Violin plots of the proportional spread of disease-associated genes in families associated with different ages of initial duplication. Mean and standard deviation are shown by the point, and horizontal line respectively. The vertical dashed line indicates 75% point, everything to the right of which represents the upper quartile of disease association.

As can be seen in figure 13, gene families with an initial duplication around the time of the last round of whole genome duplication, or earlier have a higher propensity to contain greater proportions of disease genes, with higher quantities of families being in the upper quartile of disease associations. This is particularly true of Sarcopterygii (the event directly following the proposed final round of WGD) which has both a marked enrichment for families of >75% disease associated genes, but interestingly, also a 'tail' of families wherein there are no disease genes whatsoever. One hypothesis for this may be that, given the high propensity for disease association within the other families in this group, those without disease association could be indicative of 'survivorship bias' wherein they are so intolerant to variation that they would be

embryo-lethal. These findings are supportive of our earlier results that disease arises as a result of the exposed ancient fragility of both pair, and family progenitors.

3.5 Discussion

Our findings demonstrate that the association between ohnologs and dominant disease exists for the most part due to their tendency to be haploinsufficient, a feature most commonly associated with genes arising prior to the two known WGD events in our early vertebrate ancestors. The overrepresentation of disease genes within this set then arises because WGD is the main process by which haploinsufficient genes can be both duplicated and have a high probability of being retained. This provides an understanding of why so many dominant diseases exist within human populations (Veitia and Birchler, 2010), i.e., those where a single functioning copy of the diploid gene alone are unable to 'support' the wild-type function.

Our results disentangle the 'dangerousness' of a WGD, due to locked in variation/potentially deleterious alleles (Singh et al., 2012), from the property of the gene itself, i.e., its haploinsufficiency. Haploinsufficiency is well known to be strongly associated with gene duplication (Kondrashov and Koonin, 2004; Papp et al., 2003; Pires and Conant, 2016), with haploinsufficient genes having generally been found to have a greater number of paralogs than haplosufficient genes. Our novelty here is explaining their link to WGDs. As already mentioned, this is likely due to the need to retain genes of this kind following duplication events/dosage effects as they have a greater likelihood of negatively impacting on the system, leading to stoichiometric imbalance should

aberrant copy numbers arise (Makino and McLysaght, 2010). Note, while the paper by Makino and McLysaght (2010) also looked at the link between dosage and ohnologs, they did not directly measure this using haploinsufficiency rather assumed haploinsufficiency to be a property of the ohnologs

Our results are consistent with a recent study by Diss et al (Diss et al., 2017) who investigated the impact on the system caused by the introduction of interdependent paralogs and show differential dependence in the model organism yeast, suggesting that duplication has introduced fragility and not increased robustness as expected. This supports the hypothesis of 'equivalent divergence' (and therefore absence of asymmetry) presented by Force et al (Force et al., 1999), as disruption of interdependent subfunctions of one of the duplicated paralogs in a subfunctionalised pair would produce a deleterious phenotype.

Evolution of the complexity of multi-cellular life has been facilitated by gene duplication mechanisms, particularly WGD events, which have facilitated the introduction of repurposable genetic material, despite the pre-existing fragility of more evolutionarily ancient molecular functions. Ohnologs are a source of functional divergence (Acharya and Ghosh, 2016) as they increase the probability of a dosage-dependent gene being retained, their later divergence is indicative of novel function which, in terms of disease, is likely either to be associated with complex disease or no disease. That only one of the duplicates retains dosage-sensitivity is in line with theoretical work that incorporates retention bias due to dosage constraints (Teufel et al., 2016).

Importantly, there are, however, a proportion of genes which, whilst likely to fit our criteria of greater constraint and haploinsufficiency will never be observed to be disease associated. In the case of these genes this is due to the fundamental inviability which would be introduced as a result of their disruption, and which, as highlighted by previously observed links between essentiality, developmental processes and ohnologs (Makino et al., 2009), are likely to be overrepresented in genes which have resulted from a WGD event.

Our analysis of haploinsufficiency and paralog-associated disease asymmetry within duplicated gene pairs and families, has demonstrated that the observed enrichment of dominant disease-association is an artefact of pre-existing haploinsufficiency, and any subsequent disease states are due to this, rather than the somewhat counterintuitive hypothesis that it is dominant disease-association, so-called 'dangerousness', which leads to retention (Singh et al., 2012). It should be noted that the second round of WGD (Acharya and Ghosh, 2016) will add at least one additional disease-associated gene, so ohnologs do also add disease-prone genes to the genome. These haploinsufficient duplicates, which can persist due to the initial presence of a functionally identical partner, are then able to evolve away from their haploinsufficient disease-associated state. A high abundance of haploinsufficient disease-associated genes were already extant in the ancestral genome, and what we observe within ohnologs is an increase in genes with an on-going reduction in potential for monogenic disease associations compared with the pre-WGD genome.

In conclusion, the ancestral role of the progenitor genes was likely the provision of relatively more ‘core’ functions common to most life, which leads to the observed associations with haploinsufficiency, essentiality and, as a consequence, a tendency towards fragility, all of which are traits passed on to their duplicated ‘daughters’. The lack of obvious monogenic disease-association in more recent evolutionary history will be a consequence of a reduction in the relative importance of any individual gene in the context of greater functional complexity.

CHAPTER FOUR

"Friendship sometimes rests on sharing early memories, as do brothers and schoolfellows, who often, but for that now affectionate familiarity with the same old days, would dislike and irritate one another extremely…One's friends are that part of the human race with which one can be human."

- George Santayana (1923)

KEEP YOUR FRIENDS CLOSE AND YOUR ENEMIES CLOSER: THE CO-EVOLUTION OF *TRICHURIS TRICHURA* AND THE HUMAN TGFβ SUPERFAMILY

4.1 Abstract

Trichuris trichura is a soil borne helminth, and the causal agent of Trichuriasis, a neglected tropical disease currently endemic in many low-income countries. The continued relationship between *Trichuris trichura* and human, it has been proposed, represents their 'old friendship', wherein continued exposure to the pathogen has led to interdependence between *T.trichura* and its host. Interestingly, it has been found that individuals suffering from Trichuriasis exhibit elevated levels of TGFβ1 release.

Here we use evolutionary techniques to investigate the relationship between the human TGFβ superfamily, and the *Trichuris* homolog, their shared ancestry, and relationship between the two since the two species diverged roughly 800 million years ago.

We propose that the growth factor domain of genes in the TGFβ superfamily, due to underlying important features, are liable to be under significant negative selective pressure. Whilst this is protective against heritable disease, it has provided the foundation for the old, co-evolutionary 'friendship' we observe between human and *T.trichura*

4.2 Introduction

Trichuris trichura is a soil-borne pathogenic nematode that affects human hosts, with greater than 50 species within the *Trichuris* genus that are known to be infective to a broad range of mammalian species (Hayes et al., 2010). Once ingested *Trichuris* eggs, commonly found in soil, make their way through the gastrointestinal tract to the distal small intestine, where they hatch. Following this the larvae are transported to the cecum, and proximal large intestine where they reach maturity. Once mature, these nematodes can survive for up to a year, during which time they can produce in the region 20,000 eggs a day (Hayes et al., 2010), all the while utilising strategies of immune modulation to evade expulsion (Anthony et al., 2007; Daniłowicz-Luebert et al., 2011; Lagatie et al., 2017; Maizels and McSorley, 2016).

Trichuriasis, the disease induced by active *Trichuris trichura* infection is currently classed by the World Health Organisation as a neglected tropical disease (NTD) (Centers for Disease Control and Prevention, 2011; Viswanath and Williams, 2018; World Health Organisation, 2018a). It is one of a number of diseases caused by soil transmitted helminth infections, that are currently thought to infect in the region of 1.5 billion people worldwide. Of these, between 600 and 800 million suffer with Trichuriasis (4-5%), with greater numbers of helminth infections located in low-income countries, and in the region of 4 billion people considered to be at risk worldwide. (WHO/CDC (Centers for Disease Control and Prevention, 2011, 2013; World Health Organisation, 2018a)).

A widespread programme of mass administration of various anti-helminth pharmaceuticals is currently in place to treat the adult stage of the helminth infection and to fight the spread. However, expansion of this program to provide more frequent treatment has been advised against, as, based on observed evolution of resistance in closely related animal populations, further treatment may lead to helminth resistance in humans (World Health Organisation, 2018b).

TGFβ1 is known to play an extensive role in wound healing, reproduction, cell proliferation, and differentiation amongst others, and is ubiquitously expressed across most cell types and life stages in humans (Ingman and Robertson, 2009; Moore et al., 2018).Organisms with an active *Trichuris trichura* infection exhibit increased release of TGFβ1 (Levison et al., 2010). It has been found in vivo that *Trichuris muris* TGFβ excretory/secretory component, and *Trichuris* homogenate is capable of both TGFβ1 related bioactivity and inducing the release of latent TGFβ1 in the host (Pennok and Ogunkanbi, 2019). Prior work by our group, had tentatively identified this *Trichuris trichura* TGFβ family homolog as an ortholog of a TGFβ superfamily member, suggestive that exploitation of this relationship may be one of the underlying mechanisms by which *Trichuris* is both activated, and stimulates this elevated release of TGFβ1 in the host.

The ‘hygiene’ hypothesis, first proposed in 1989 (Andreae and Nowak-Węgrzyn, 2017; Strachan, 1989) suggested that, without early exposure to certain previously endemic pathogens such as those in the helminth family, the human immune system is improperly able to mature (Briggs et al., 2016; Jackson et al., 2009; Strachan, 1989). It posited that the rise in immune and immune related disorders across Europe was, therefore, due in part to the large-scale eradication of helminth infections. This hypothesis was popular amongst researchers for a number

of years, as it seemed to provide an explanation of the correlation between the reduction in helminth infections and rise in immune disorders. However, over time the hygiene hypothesis fell out of favour, particularly with regards to the effect of helminth infections, as it does not fit with the wider biological context wherein helminths have been shown to exacerbate certain autoimmune disorders (Briggs et al., 2016). An alternative, the 'old friends' hypothesis has been proposed to replace the hygiene hypothesis (Rook et al., 2003). This postulates that due to the prolonged evolutionary history of mammalian infection by immune-modulating pathogens, the two groups have evolved a degree of interdependence (Rook et al., 2013). The hypothesis as it relates to *T. trichura* is that the human adaptive immune system has evolved dependence on certain helminths and microbial communities, selectively retaining features that are beneficial in the presence of heightened levels of regulatory, and type II T-helper cells characteristic of helminth infection. Furthermore, it posits that without their stimulus, humans have an increased susceptibility to both viral and bacterial infection (Jackson et al., 2009), autoimmune disorders (Machado et al., 2012), and cancers (Thomas et al., 2018).

It has been suggested that the once near ubiquitous prevalence of helminths in the human population led to their presence being tolerated. Such systems dependencies are commonplace for many parasites, wherein they have evolved dependence on aspects of the host systems (Decaestecker et al., 2007; Woolhouse et al., 2002).

CKGF, the proposed ancestor of TGFβ superfamily genes, is postulated to have appears to have arisen during the metazoan period (Duran *et al.*, 2018; Hinck *et al.*, 2016; Pirruccello-Straub *et al.*, 2018), prior to bilateria, which is consistent with the presence of a homologue in *Trichuris* and human. Genes within the TGFβ

superfamily are characterised by high levels of sequence similarity in a conserved Growth-Factor (GF) domain at the C-terminus, inclusive of a characteristic cysteine repeat motif, and a pro-domain which a much lower degree of conservation than that found in the GF domain (Burt, 1992; Hinck et al., 2016). The high level of sequence conservation within the GF domain, particularly the cysteine repeat motif that all genes in this family share, is indicative of purifying selection, likely due to the essential functions of the protein products that they encode.

Protein products encoded by genes in the TGFβ superfamily dimerize. For some this is a promiscuous dimerization as they are able to form a heterodimer between different TGFβ family members, while others such as TGFβ1 form homodimers (Hinck, 2012). TGFβ1, MSTN, and GDF11 dimers interact with latent transforming growth factor binding proteins (LTBPs) in the extracellular matrix, where they are stored awaiting cleavage and activation by furin, or in the case of myostatin (MSTN) bone morphogenetic protein 1 (BMP1) (Walker et al., 2016; Wolfman et al., 2003). It has been suggested that the mechanisms that allow this storage are driven by differences in sequence composition in the pro-domain, which has aided in fine-tuning tissue-specific activation of these proteins by blocking both overexpression and furin processing (Walker et al., 2016).

The TGFβ superfamily gene products undergo three stages before they become biologically active: expression, latency, and activation. As briefly mentioned, following initial expression, certain members of the TGFβ superfamily, such as TGFβ1, 2, and 3, GDF11, BMP10 and MSTN, are able to remain anchored in the extra-cellular matrix in a latent form prior to activation (Moore et al., 2018). This storage is possible due to the interaction of these proteins in their dimeric form

with extra-cellular matrix proteins such as perlecan and fibrillin, which allow them to become biologically inactive until subsequent cleavage (Harrison et al., 2011).

Receptor binding within the TGFβ superfamily can be divided into two categories; type 1 receptors, which bind with differing levels of specificity to the TGFβs (binding specific), GDF11, MSTN (promiscuous, both type 1&2 binding), and the activin genes; and Type 2 receptors which bind to the other members of the family (Hinck et al., 2016). Extracellular bioavailability of GDF11, like TGFβ is therefore, supported (Duran et al., 2018; Hinck et al., 2016; Pirruccello-Straub et al., 2018).

TGFβ superfamily genes have been well studied, and as such, are known to play pivotal roles in diverse human genetic diseases from Loeys-Dietz syndrome 4 (aortal aneurisms - TGFB2) to Osteogenesis imperfecta (BMP1) (McKusick-Nathans Institute of Genetic Medicine, 2018). The strong association between genes in the TGFβ superfamily, and often-severe human heritable disease, is indicative of the important functions that the TGFβ genes serve from early development, throughout the entirety of the human lifespan. The TGFβ superfamily genes TGFβ1, GDF11, and MSTN have been shown increase levels of phosphorylation and to be especially essential to both human development, and the aging process more generally, with particular associations with the musculo-skeletal systems (Duran et al., 2018; Walker et al., 2016). They are, as a consequence linked with disorders in these tissues, for example Camurati-Engelmann disease, characterised by limb and cranial hyperostosis (Hughes et al., 2019), muscle hypertrophy (Schafer et al., 2016), cardiovascular disease, and some cancers (Zhang et al., 2017), with GDF11 deficiency found to be embryonically lethal in mice (Walker et al., 2016).

Unlike many TGFβ superfamily members, rather than a single cleavage site being required to activate the protein, GDF11 and MSTN require two cleavages.

Firstly, a furin or PCSK5 cleavage is required, resulting in a latent complex. This is followed by a second cleavage by either a tolloid or BMP gene product, both found within the TGFβ superfamily, which finally produces the activated protein (Pirruccello-Straub et al., 2018; Walker et al., 2016)

Genes subject to differing evolutionary pressures are known to exhibit divergent rates of evolution (Yang et al., 2000), with those that have duplicate events in their evolutionary past, also tending to have a greater association with disease (Dickerson and Robertson, 2012). Of duplicate types there is a particular that have been shown to have a heightened association with haploinsufficiency (being unable to provide wild-type function without both copies functioning adequately), and disease (Makino and McLysaght, 2010; Martin-Geary et al., 2019) than their small scale duplicated (SSD) counterparts. These genes, known as ohnologs, are the result of whole genome duplication (WGD), of which two rounds are posited to have occurred at the base of the vertebrate lineage (Ohno, Susumu, 1970). The retention bias towards ohnologs has been hypothesised as being the result of dosage sensitivity, in that, interdependent components of the larger biological system, having been duplicated together, allow balance within the system to be maintained. Whilst differential loss of certain genes is tolerated, others disrupt systemic stoichiometry to such an extent that they result in a deleterious phenotype. It is proposed that these genes expose the ‘fragility’ of their ancestors, and therefore their maintenance, is essential to human health (Makino et al., 2009; Martin-Geary et al., 2019).

Duplication events, in particular WGD, have been linked to large-scale speciation (Acharya and Ghosh, 2016; Sémon and Wolfe, 2007). This is due to differential loss of non-dosage sensitive genes leading to reproductive isolation, and

subsequent diversification of function provided by the introduction of large quantities of redundant genetic material (Acharya and Ghosh, 2016) and the subsequent complexity this brings.

Here we show that the ancient host-pathogen relationship between Trichuris species and mammals is likely successful due to the close evolutionary relationship between members of the TGFβ superfamily, in particular GDF11, and, using evolutionary analysis we propose an explanation for how and why this relationship arose and has been maintained.

4.3 Methods

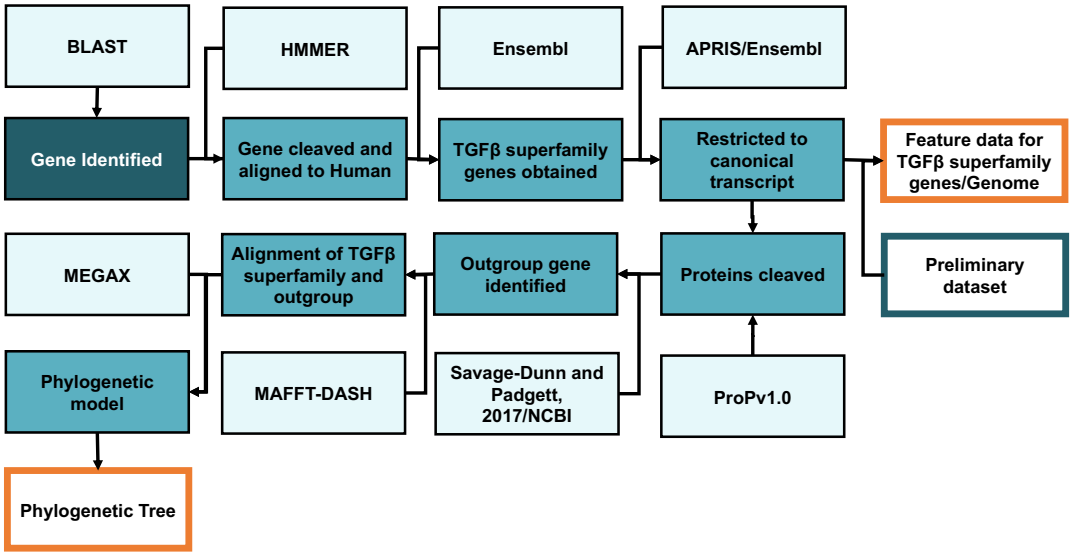


Figure 14: Flow chart showing the methods used to perform the analysis of the TGFβ superfamily. Initial identification of the gene is shown in dark blue, analysis steps are shown in medium blue, and additional tools and data are shown in light blue. The preliminary dataset generated using the general methods is shown with a dark blue border, and the resultant tree, and data used in this analysis are bordered in orange.

Identifying the gene.

Work previously conducted by our group, had tentatively identified a trichuris muris (used as a proxy for *Trichuris Trichura)*, homologous gene to human TGFβ1 using BLAST. The *Trichuris* gene was identified as a gene of interest (TMUE_2000007822 wormpep=TMP09347, henceforth referred to as T') due to its link with elevated levels of TGFB1 release *in vivo* (Pennok and Ogunkanbi, 2019).

To confirm the identity of the homolog, accounting for differential conservation of the pro-domain, and GF domain of TGFβ superfamily genes, the genomic sequence of the T' gene (obtained from WormBase:Parasite, Murrell, 2018), was cleaved, using a previously identified furin cleavage motif, into its

constituent domains. Using HMMER, a hidden markov model multiple alignment tool each domain was subsequently aligned against the human genome. This tool has the benefit of increased sensitivity of homology detection over pairwise local alignment algorithms (Finn et al., 2011).

Protein sequence data for members of the human TGFβ superfamily as defined by Ensembl, were obtained from Ensembl BioMart (Zerbino et al., 2018). A single putatively canonical amino acid sequence for each gene was chosen by cross referencing APPRIS scores (Rodriguez et al., 2013), with transcript length (including UTRs and CDS), both obtained from Ensembl. The resultant Protein IDs can be seen in table 4.

Table 4 : Ensembl gene ID, Protein stable ID, APPRIS annotation and length of selected representative proteins for each member of the TGFβ superfamily

Gene stable ID	Protein stableID	APPRIS	Transcript length
ENSG00000092969	ENSP00000355897	principal1	5868
ENSG00000101144	ENSP00000379204	principal1	4021
ENSG00000105329	ENSP00000221930	principal1	2769
ENSG00000112175	ENSP00000359866	principal1	3970
ENSG00000116985	ENSP00000361915	principal1	4826
ENSG00000119699	ENSP00000238682	principal1	2522
ENSG00000122641	ENSP00000242208	principal1	6046
ENSG00000123999	ENSP00000243786	principal1	1351
ENSG00000125378	ENSP00000245451	principal1	1931
ENSG00000125845	ENSP00000368104	principal1	3545
ENSG00000125965	ENSP00000363492	principal1	2572
ENSG00000130283	ENSP00000247005	principal1	2579
ENSG00000130385	ENSP00000252677	principal1	1179
ENSG00000130513	ENSP00000252809	principal1	1200
ENSG00000135414	ENSP00000257868	principal1	8657
ENSG00000138379	ENSP00000260950	principal1	2822
ENSG00000139269	ENSP00000266646	principal1	2460
ENSG00000143768	ENSP00000355785	principal1	2019
ENSG00000143869	ENSP00000272224	principal1	9267

ENSG00000152785	ENSP00000282701	principal1	6092
ENSG00000153162	ENSP00000283147	principal1	3784
ENSG00000156466	ENSP00000287020	principal1	3712
ENSG00000156574	ENSP00000287139	principal1	1665
ENSG00000163083	ENSP00000295228	principal1	3206
ENSG00000163217	ENSP00000295379	principal1	6036
ENSG00000164404	ENSP00000367942	principal3	2640
ENSG00000175189	ENSP00000308716	principal1	3202
ENSG00000183682	ENSP00000327440	principal1	5636
ENSG00000184344	ENSP00000331745	principal1	1236
ENSG00000243709	ENSP00000272134	principal1	1626
ENSG00000263761	ENSP00000463051	principal1	1955
ENSG00000266524	ENSP00000464145	principal1	2677

Phylogenetic analysis & sequence alignment.

An initial alignment of TGFβ superfamily proteins using clustalO (Sievers et al., 2011), implemented in SeaView Version 4 (Gouy et al., 2010), indicated that the furin cleavage site (RXXR/RXKR) previously identified in the T' gene and human TGFβ1 was not universally conserved across the TGFβ superfamily. It was therefore necessary to use a cleavage prediction tool (ProP v.1.0b Duckert et al., 2004) to establish likely cleavage locations within each of our amino acid sequences. For genes with multiple predicted possible cleavage sites, a single candidate was selected by cross-referencing the ProP score of each predicted site with similarity to the T'/TGFβ1 RXXR/RXKR motif. For genes with no predicted cleavage sites above the cut-off score threshold, sites were selected based on alignment against family members with high confidence cleavage sites.

In order to root the phylogenetic tree, it was necessary to source an appropriate out-group gene that is both a member of the TGFβ-superfamily, and within an organism whose divergence pre-dates that of *trichuris* and human. Due to

the fact that the divergence of human and *trichuris* is near-contemporary with the proposed origin of the TGFβ-superfamily very few species with identified TGFβ genes predate the *trichuris*/human split, with only Trichoplax and sponges having been tentatively identified TGFβ-superfamily homologs (Savage-Dunn and Padgett, 2017). A putative *Sycon Ciliatum* (sponge) homologous gene (TGFβD), sequenced and annotated by Leininger *et al,* and identified as having increased expression during early development was selected, the protein sequence was obtained from the NCBI (Accessions: CDO67923) and cleaved using ProP in the same manner as the human and trichuris genes.

An alignment of the resultant human, *trichuris*, and *S.ciliatum* peptides was made using the MAFFT-DASH tool (Rozewicki et al., 2019), leveraging both structural, and sequence data in order to build alignments, with the following parameters: Use DASH to add homologous structures; Adjusted direction according to first sequence; L-INS-i iterative refinement strategy providing increased accuracy for distantly related homologs; unalign level 0.0; retain gaps; BLOSUM62 scoring matrix; gap penalty = 1.53; offset value = 0.0; nzero; and Mafft-homologs selected with UniRef50, using PSI-BLAST to retrieve homologs in order to improve alignment. This was then imported into the MegaX environment (Stecher et al., 2020), and manually edited to correct any remaining misalignments.

Using the resultant alignment and the MEGAX software (Stecher et al., 2020) a phylogenetic model was selected using the MEGAX model selection tool, and the following parameters: Automatic neighbour-joining tree; maximum likelihood statistical method; Use all sites ; and a ‘very weak’ branch-swap filter, in order to explore the maximum possible space. It must be noted however, that it has recently been suggested that model selection may not have any profound impact in

accuracy of the resultant tree (Spielman, 2019). A maximum likelihood tree was built using the recommended Jones-Taylor-Thornton+G +I model, with Gamma=2, Subtree-Pruning-Regrafting (level 5), 1000 bootstraps, and eight cores. This was subsequently visualised using FigTree V1.4.2 (Rambaut, 2014), and manually adjusted to improve readability.

The TGFβ superfamily dataset.

In order to perform our analysis and infer evolutionary histories and relationships between genes in the TGFβ superfamily we used the preliminary dataset (see general methods), divided into two sets; genes within the TGFβ superfamily, and those without. Statistical analysis and figures were generated in R version 3.3.2 (R Core Team, 2016), using the following packages: plyr (Wickham, 2016), ggfortify (Yuan T, Masaaki H, Wenxuan L, 2016), gridExtra (Auguie and Antonov, 2017), grid (Murrell, 2018), and ggplot2 (Wickham, H, 2009).

In order to draw meaningful comparisons between evolutionary features as they relate to the TGFβ superfamily, statistical comparisons were made relative to the distribution of those features in the rest of the genome. Given the vast difference between quantities of genes contained within these two sets, it was necessary to perform these comparisons as proportions rather than absolute gene counts. This was done using the base R prop.test function.

4.4 Results

Identifying the *Trichuris* homolog

Having divided the *T. muris* gene (TMUE_2000007822) into its constituent domains, it was found that rather than being most closely related to TGFβ1, as had previously been indicated, both domains shared greater identity with human GDF11, with secondary matches to MSTN/Myostatin (Figure 15). This is concordant with the findings of WormBase ParaSite which also identified the T' gene as a homolog of GDF11.

	Target	Description	Species	Cross-references	E-value
>	ENSG00000135414.9	growth differentiation factor 11 [Source:HGNC Symbol;Acc:HGNC:4216]	Homo sapiens		2.1e-69
>	ENSG00000135414.9	growth differentiation factor 11 [Source:HGNC Symbol;Acc:HGNC:4216]	Homo sapiens		2.4e-69
>	ENSG00000138379.4	myostatin [Source:HGNC Symbol;Acc:HGNC:4223]	Homo sapiens		3.3e-63

Figure 15 : HMMER output for Human/T' sequence alignment

The characteristic element diagnostic of genes within the TGFβ superfamily is a highly conserved C-terminus growth factor domain containing a distinctive motif of cysteine repeats (Figure 16), with a highly divergent N-terminus pro-domain. As has previously been identified (Burt, 1992; Hinck et al., 2016) due to the high divergence of sites within the pro-domain of the TGFβ superfamily members in relation to the GF domain, reliably establishing a root for trees built using complete genes/proteins presents a serious challenge, as out-group genes tend as a result of the divergence of the pro-domain, to align to clades within, rather than external to the tree.

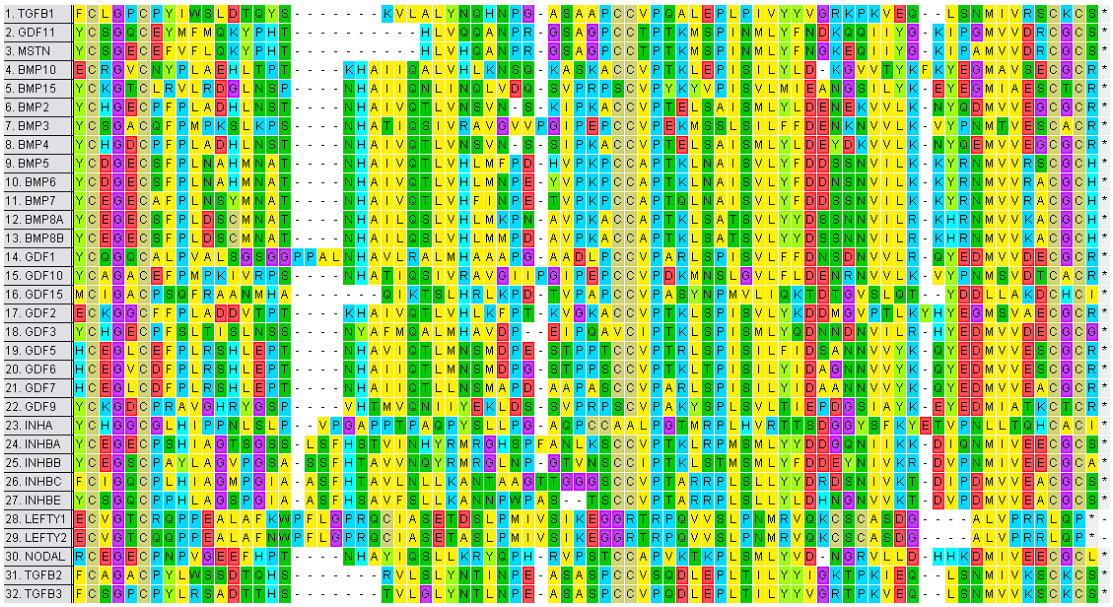


Figure 16 : Subset of the manually adjusted alignment of TGFβ superfamily canonical GF domain proteins, showing the highly conserved cysteine repeat motif in gold, alongside the T' protein (top) and Sycon Scilliatum outgroup protein (bottom).

Phylogenetic analysis

In order to infer phylogenetic relationships between members of the TGFβ superfamily, we used an amino acid sequence alignment, as these often exhibit greater conservation than their nucleotide counterparts. Similar to the method used by Hinck *et al* (2012) we only considered amino acids downstream of the furin cleavage site, thereby utilising the maximally conserved C-terminus region. This retained only homologous, highly conserved elements of the family key to our study, whilst minimising the previously discussed difficulties in root assignment, therefore allowing a potential root to be established (Figure 17).

.

Figure 17 : Phylogram showing the divergence of genes in the TGFβ superfamily, and the T' gene (shown as T.trichura. Genes in the GDF11 clade are highlighted in pink. The outgroup gene is shown at the bottom (Sycon Scilliatum). The tree was constructed using the maximum likelihood method with the JTT matrix based model. Bootstrap values are shown as percentages adjacent to their respective nodes.

As can be seen in figure 17, phylogenetic analysis indicates that the divergence of the T' gene from its human homologs occurred prior to the event that gave rise to MSTN and GDF11. The closer sequential proximity between GDF11 and the Trichuris homolog by contrast with MSTN is indicative of a higher degree of conservation of GDF11 when compared with MSTN.

Interestingly, both GDF11 and TGFβ are more proximal to the root of the tree compared with the BMP and remaining GDF genes. This result is

somewhat unexpected as previous studies have suggested that the BMP genes are some of the earliest members of the TGFβ superfamily, with the TGFβ 1/2/3 clade arising much more recently than our findings indicate (Hinck, 2012).

Paralog status in human TGFβ superfamily genes

Due to the nature of paralog families, every gene in the TGFβ superfamily arose as a result of a duplication event. What is noteworthy, is the quantity of genes within the TGFβ superfamily that are ohnologs of strict or intermediate type (~44% N=14) compared with the instance of small scale duplicates, a mild elevation of the proportions of ohnologs relative to SSDs in the genome P=0.5423 (Figure 18).

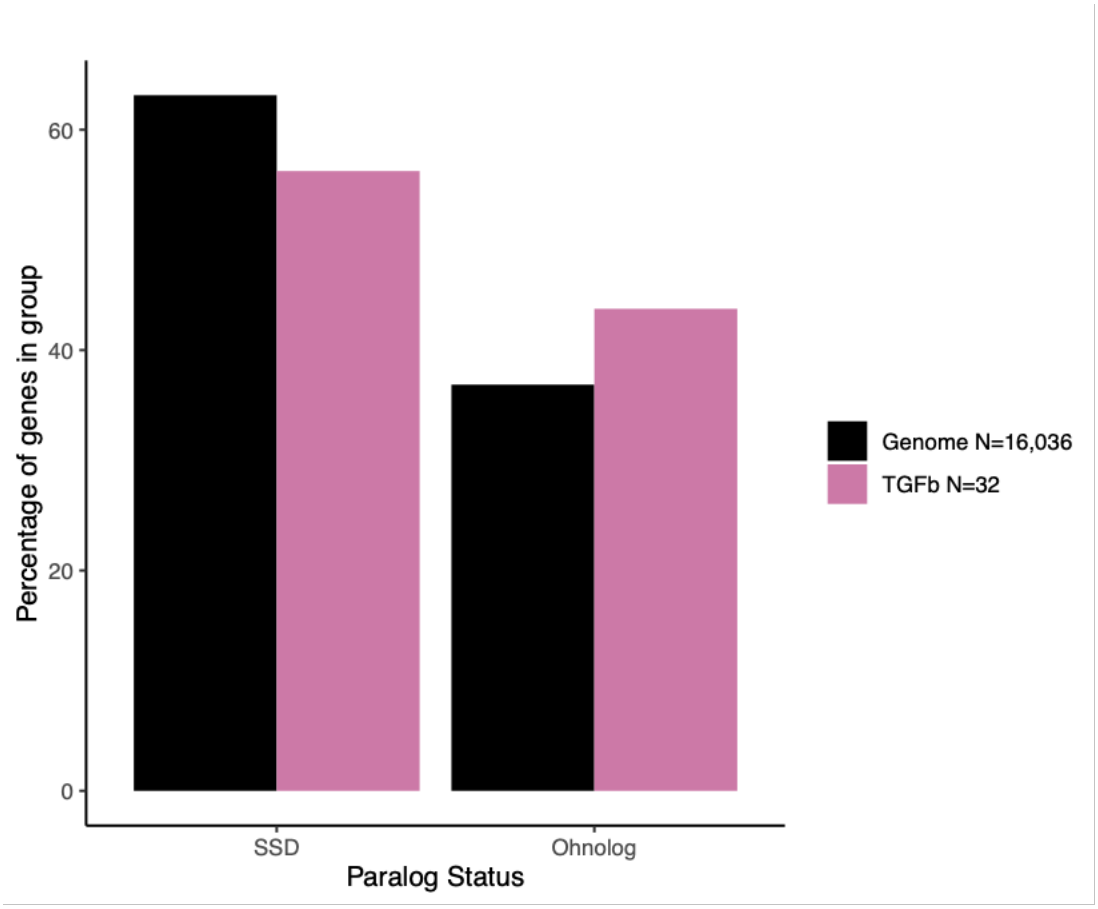


Figure 18 : Plot showing proportion of genes in the TGFβ superfamily in each duplication group (pink N=32) vs. Genome duplicate proportions (black N=16,036).

Gene ages of TGFβ superfamily genes

By dating each gene in relation to its earliest ortholog within other Ensembl Compara vertebrate species, we found that genes within the TGFβ superfamily are, on the whole, relatively evolutionarily ancient, with the mode first appearance falling at ~615mya, and no genes having arisen later than ~105mya (Figure 19).

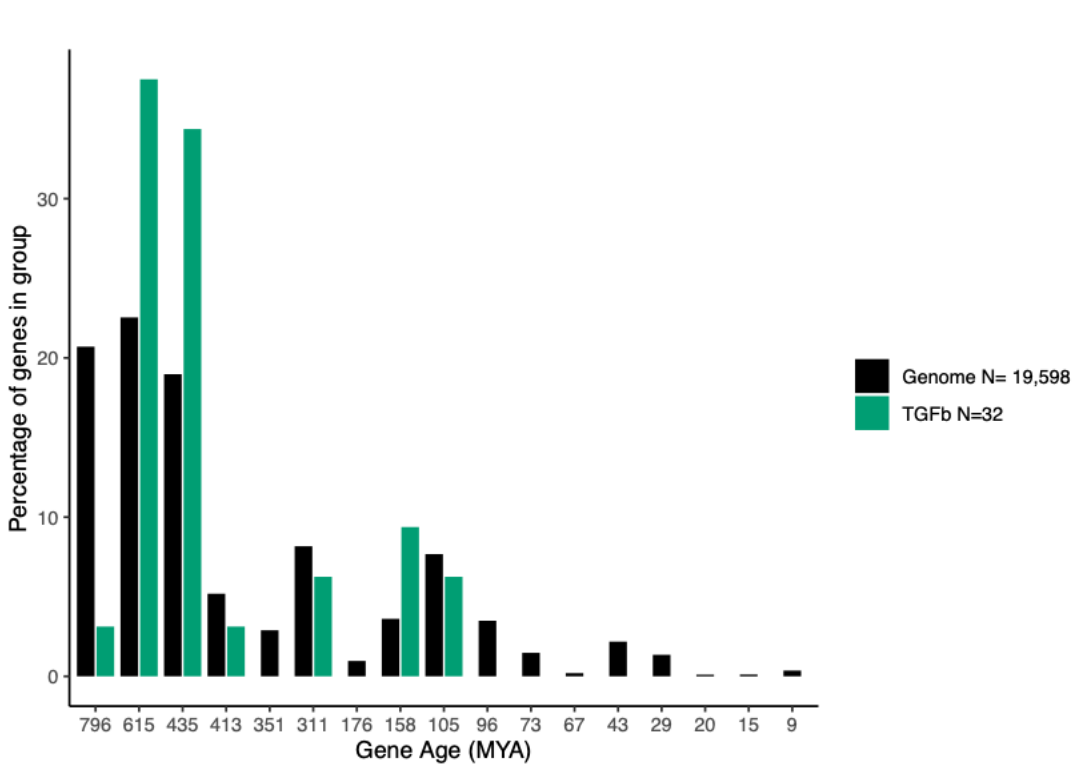


Figure 19 : The distribution of genes arising at each gene age group within the genome (black N=19,598), alongside by the distribution of genes arising at each gene age (MYA) within the TGFβ superfamily (green N=32)

Plotting the gene ages against that of the full genome (Figure 19) shows a significant difference in the distribution of TGFβ superfamily ages within two age brackets in particular, corresponding to the estimated time of the last round of whole genome duplication (~435MYA, Euteleostomi, P= 0.04625). We also found a significant depletion of genes predating the two rounds of WGD (~796mya, P= 0.0254).

Duplication relationships over evolutionary time

Figure 20 shows the distribution of gene ages within the TGFβ superfamily in relation to the major biological events occurring between the split at bilateria, and the present, capturing the earliest appearance of each individual gene as they occur within the Ensembl vertebrate tree. Surprisingly, GDF11 appears to be one of the youngest genes in the TGFβ superfamily, the oldest retained ortholog having been identified in species diverging from the *Homo sapiens sapiens* line at Eutheria. MSTN however, shares orthologs with species diverging much earlier, at Sarcopterygii.

When comparing this with our phylogenetic analysis we can see that the ages and evolutionary relationships are not concordant. This is likely due, at least in part, to the vertebrate centric data from which the ages were generated, and only a limited ability to detect both gene loss and ancestral origin. However, in reconciling the two we may infer that GDF and BMP genes are some of the oldest genes in the family, but that the TGFβ genes, MSTN and GDF11 are products of an ancient ancestor which is likely to have arisen early in the TGFβ superfamily

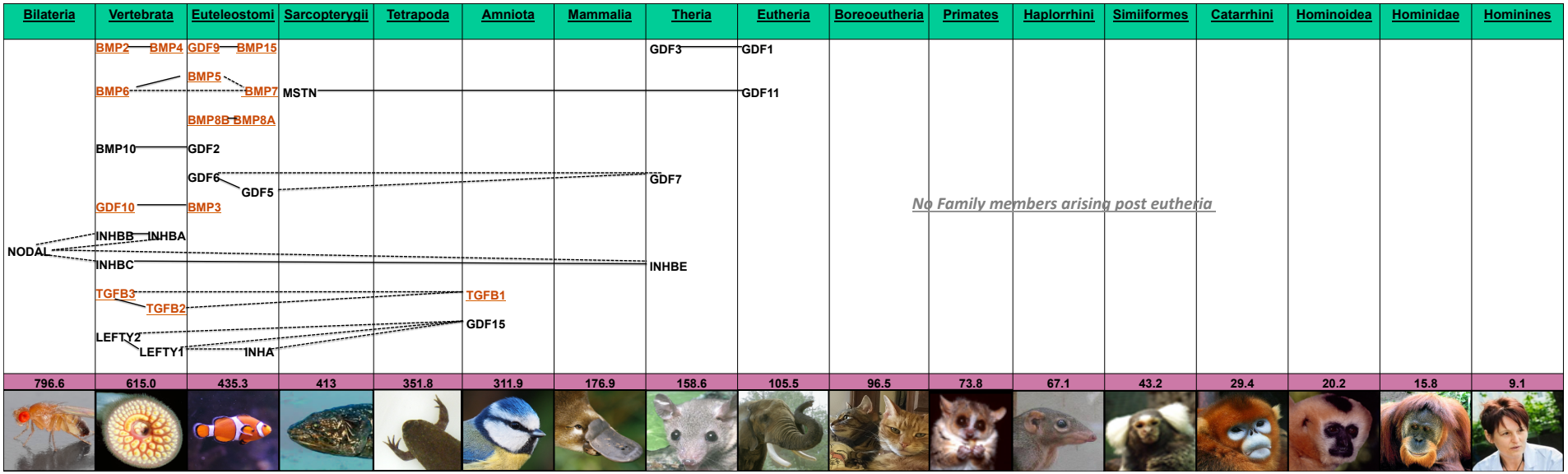


Figure 20 : The most recent common ancestor of each TGFβ superfamily gene is shown, Links between genes (solid line) represent direct ancestry observed in the phylogenetic tree (Figure 16), with deeper relationships shown as dotted lines. Genes originating as a r

Haplosufficiency in the TGFβ superfamily

Distribution of Haplosufficiency ranks within the TGFβ superfamily (Figure 21) by comparison with the rest of the genome show that a large proportion of TGFβ superfamily genes fall within the more likely to be haploinsufficient ranks (0-10, N=12, ~37.5%) as defined by decipher (Firth et al., 2009), a significant elevation compared with the genome proportions (p= 1.657e-05, genome = 10%) and likely to result in a deleterious phenotype should their diploid state be disrupted.

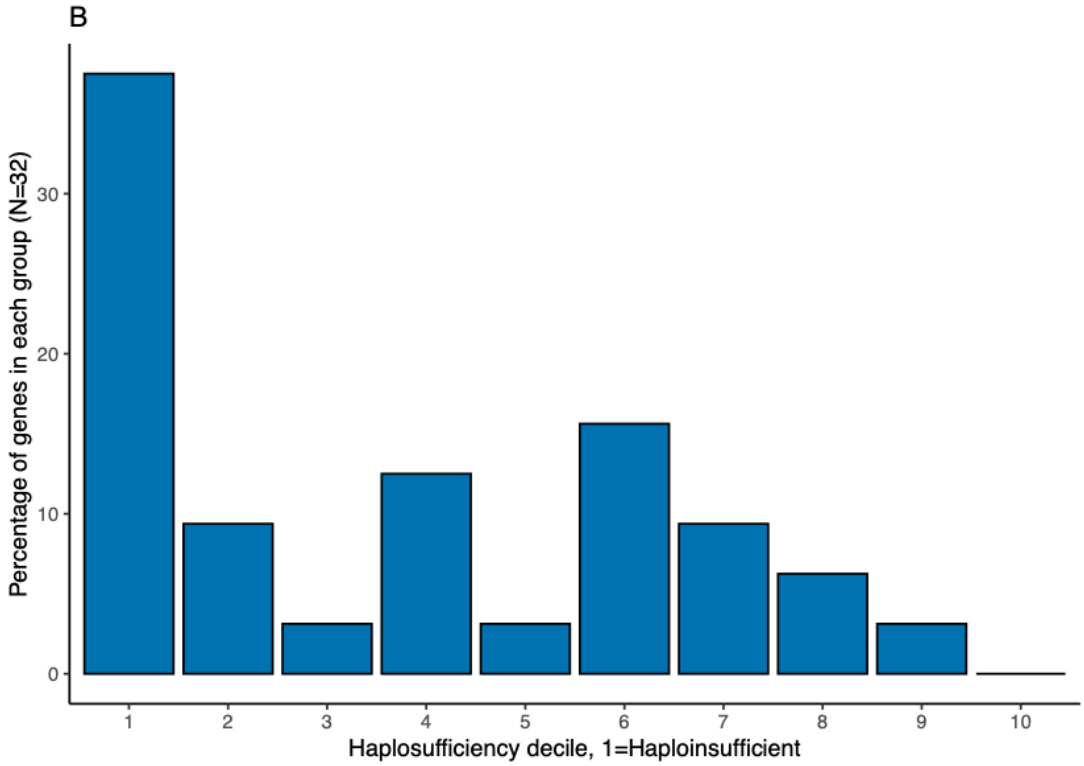


Figure 21 : Histogram showing haplosufficiency deciles for genes in the TGFβ superfamily.

For the subset of genes within the TGFβ superfamily that fall within the haploinsufficient ranks (0-10), a striking trend may be observed: these genes predominantly fall into the upper ranks, with 10 of the 12 genes within top 5% of

most haploinsufficient genes in the genome, four of which being in the top 1%. the most highly ranked of which being TGFβ 1, which is the 24th most highly ranked gene in the human genome (top 0.001%). GDF11 however, falls just outside of the bracket indicated by decipher as being haploinsufficient (0-10), with a score of 14.75, whilst MSTN is squarely within the decipher criteria with a haplosufficiency rank of 2.28. This shows that predominantly, genes in the TGF β superfamily have a propensity to be haploinsufficient, therefore resulting in deleterious phenotypes if disrupted.

Protein-protein interaction partners in TGFβ superfamily genes

The number of interaction partners a protein possesses, has long been known to be indicative of its importance within the cellular system. Proteins with large numbers of interaction partners are identified as 'hubs', exhibiting a high degree within the network (Vallabhajosyula et al., 2009), disruption or perturbation of which is liable to result in major failure of part, or the entirety of the system, and is particularly significant if it is a pleiotropic gene, which plays a part in multiple, or important pathways (Promislow, 2004).

We found (Figure 22) that only three out of the thirty-two genes in the TGFβ superfamily (~9%) fall below the mean number of protein interactions within the genome (N= 1015)**.** Of the remaining 29 genes, 22 fall within one standard deviation exceeding the mean, and 7 (~22%) exceed it. The most noteworthy of these is TGFβ1, which is within the top ~0.8% of all genes (N=22722), and the top 1.1% of genes with a connectivity score of 1 or more (N=18865) with GDF11 being found within the top 5% of all genes, and the top 7% of genes with one or more protein interaction.

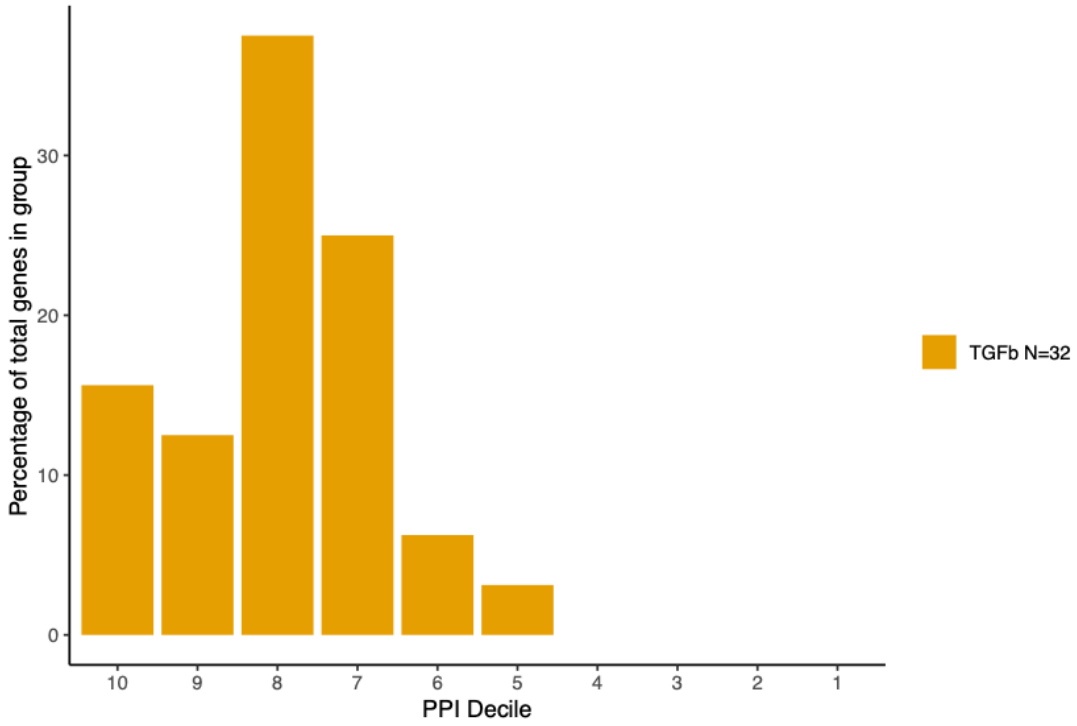


Figure 22 : Plot showing proportion of genes in the TGFβ superfamily (orange N=32) that fall within each protein-protein interaction count decile.

When looking at the distribution of TGFβ superfamily genes across the PPI deciles a significant elevation can be seen in decile 8 (top 20-30% of genes in the genome). When compared with the proportion of genes in the genome ~10%, p = 9.849e-07.

It should be noted that there are caveats associated with the use of PPI data. The STRING data contains information regarding known interactions; however, this is liable to be biased towards more highly studied genes, which may influence the TGFβ1 score for example. Of particular note, however, is that there are likely to be a proportion of genes which have no record, but for which protein interactions do exist. Whilst these false negatives can be avoided by selecting only genes with a PPI connectivity of 1 or more as the genome scale data, this introduces a new bias, by not only omitting genes which code for independently acting proteins, but also genes such as important regulators and other non protein-coding genes.

3.5 Discussion

We have found that genes in the TGFB superfamily are evolutionarily old, predominantly originating at, or prior to the last round of whole genome duplication, which is consistent with the proposal of a pre-existing progenitor in the metazoan period (Duran *et al.*, 2018; Hinck *et al.*, 2016; Pirruccello-Straub *et al.*, 2018). The finding of a significant elevation in genes contemporary with the last round of whole genome duplication is of particular interest given importance of evolutionarily ancient genes, their essentiality, and contribution to system fragility (Martin-Geary et al., 2019), which is known to lead to higher levels of conservation.

The presence of the GDF11 homolog in *Trichuris* suggests that despite the extant genes' younger appearance, GDF11, MSTN and T' genes share a common ancestor that is at least bilateria in origin. The significant depletion of genes in the oldest time point, corresponding to early in the vertebrate lineage (~796mya, P= 0.0254), suggests that diversification of the TGFβ superfamily occurred as a result of successive duplication events early in the vertebrate lineage, around the time of the two rounds of whole genome duplication. This may to some extent explain why there are so few TGFB family members in helminths, as the majority arose following divergence of helminths and humans.

Whilst we have not explored the root of GDF11's surprisingly evolutionarily youthful age compared to expectations, there are a number of explanations that may be made. Firstly, the method of dating used has the potential for bias as a result of its vertebrate focus, however, an alternative hypothesis is that GDF11 did pre-exist the date proposed by our method of dating but was lost in some of the more divergent

species. It is known that GDF11 is essential to human development, however it may not be essential to non-human lineage-specific function. The potential loss of GDF11 within other vertebrate species, therefore, would not necessarily be unexpected, given that it falls outside of Decipher's criteria for haploinsufficient genes, unlike many other members of the superfamily, therefore, its loss is less likely to result in as deleterious a phenotype as would, say, TGFβ1 or MSTN. It would be interesting to see if those species that retain GDF11 are also host species for pathogens in the *Trichuris* genus, whilst those without it are not, as this may further support the proposition of the exploitation of shared ancestry of the TGFβ in helminth infections.

The close relationship between the TGFβ and the MSTN/GDF11 clades identified in prior studies is congruent with our findings, however, the results of our phylogenetic analysis of genes within the TGFβ superfamily regarding relative clade age within the TGFβ superfamily did not confirm the findings of prior works (Hinck, 2012). This may, in part, be due to the high number of members now known to reside within the superfamily included in our analysis, by comparison with prior studies, or that, to our knowledge, ours is the first study to conduct a phylogenetic analysis using an outgroup.

We have found that many genes within the TGFβ superfamily are highly connected within the human PPI network. Loss, or significant change of these genes has the potential to cause a reduction in binding affinity, and therefore carries a higher-than-average likelihood of being disruptive to the network. Recent studies have linked highly connected nodes critical to the molecular system as being preferential targets for viruses (Ravindran *et al.*, 2019). Given that GDF11 is within the top 5% of connected genes in the human genome it may be hypothesized that this

has the potential to lead it to be a target for opportunistic pathogens, via provision of beneficial network controllability. Whilst it is beyond the scope of the current analysis to do so, it would be interesting to ascertain the degree, and control status of genes in the TGFβ superfamily, particularly as control genes are liable to be under greater constraint than their less critical counterparts. This could also in part explain the ability of *Trichuris* species to maintain a prolonged niche in their respective hosts.

It has been suggested that the pro-domains of genes in the TGFβ superfamily have been under greater pressure to diverge than the conserved growth-factor domains (Hinck *et al.*, 2016), potentially resulting in differential binding specificity, however this could also be considered to be less constraint than is found in the growth factor domain. This near fourfold greater divergence (Hinck *et al.*, 2016) in the pro-domain may be more indicative of a relaxation of constraint in this region following expansion in the superfamily, facilitating the evolution of novel functions of these genes when compared with their ancestor.

These features combined suggest that genes in the TGFβ superfamily, in particular TGFβ1, have retained a high degree of amino acid conservation over the course of the bilateria lineage, due to their being important genes within most mammals. Whilst this has been essential in order to 'protect' against human heritable genetic disease, we propose it has likely provided a platform that has proven beneficial to the evolution of the adaptive immune system in humans in the presence of once endemic opportunistic helminth parasites such as *Trichuris trichura*. This is further supported by earlier work by Machado *et al* (Machado *et al.*, 2012), who found that there is an increased susceptibility to autoimmune disease in the absence

of a *Trichuris* infection and is illustrative of susceptibility to autoimmune disorders having both a genetic and environmental component.

Our analysis shows preliminary evidence of a persistent co-evolutionary relationship between helminths and mammals throughout the course of mammalian evolution (Jackson *et al.*, 2009; Thomas *et al.*, 2018). This argument, in line with the old friends hypothesis (Rook *et al.*, 2003), hinges on an observable reduction in efficacy of the adaptive immune system in helminth-free populations by comparison with populations wherein helminths are endemic.

Our findings show that the established relationship between *Trichuris* and mammals may be seen in the close association between the *Trichuris* homolog TMUE_2000007822 (T') and GDF11/MSTN genes. This host-pathogen interaction is likely to have provided the vehicle for both their old friendship, and the continued exploitation of mammalian biology, wherein *Trichuris* is both activated, able to stimulate the release of TGFβ1 in the host.

CHAPTER FIVE

"The facts of variability, of the struggle for existence, of adaptation to conditions, were notorious enough; but none of us had suspected that the road to the heart of the species problem lay through them."

– Thomas Henry Huxley (1887)

VIRUSES CONTROL THE HUMAN INTRA-CELLULAR SYSTEMS BY EXPLOITING EVOLUTIONARILY ANCIENT MOLECULES

5.1 Abstract

Virus-host interactions leave distinctive indicators of adaptation on the genome of the host. An estimated 75% of human genetic adaptation have occurred within the most highly constrained genes in the genome, driven by a history of virus exposure (Castellano et al., 2019). This is particularly significant when you consider that heritable disease associated genes tend to also be highly conserved (Miller and Kumar, 2001; Petrovski et al., 2013). Using a combination of network analysis, and evolutionary data we explored to what extent, if any virus interacting proteins (VIPs) and control nodes, those nodes that within the human protein interaction network are integral to control, are associated with susceptibility to genetic disease and the. We confirm that there is a substantial intersection between human heritable disease, and VIPs. This observed correlation between heritable disease and VIP status of certain genes is not a result of VIPs specifically targeting disease associated genes however, but rather the underlying evolutionary importance of these genes leading to 'new' associations. Looking at the duplication status of genes that are disease associated, virus associated, or both, we also find that genes arising from duplication events, despite being heavily linked with human heritable disease (Dickerson and Robertson, 2012; Makino and McLysaght, 2010; Richard et al., 2008; Schuster-Böckler et al., 2010; Singh et al., 2015; Van de Peer et al., 2017) have potentially contributed robustness to the system rather than fragility.

5.2 Introduction

The use of network analysis to aid in the interpretation of complex interactions (Boyle et al., 2018; Kuzmin et al., 2018; Monaco et al., 2018; Szklarczyk et al., 2017), and the relationship between pathogens and their hosts (Ahmed et al., 2018; Lee et al., 2018; Ravindran et al., 2019) has revolutionised our understanding of biological data. A vast array of novel, models for network analysis have provided robust frameworks for the analysis of, often prohibitively complex systems (Barabási et al., 2011; Emilsson et al., 2018; Gosak et al., 2018; Zhou et al., 2014)

The application of such methods has proven particularly beneficial when looking at human heritable disease, having substantially aided in the elucidation of the underlying features governing human heritable diseases and their evolution. A number of methods have more recently expanded to incorporate key components from control theory, identifying key elements of biological systems integral to control (Del Vecchio et al., 2018; Kremling, 2013; Tsongalis, 2018). The application of control theory to biological data has shed light on essentiality, plasticity, and the reconciliation of control and evolution (Badyaev, 2019; de Anda-Jáuregui et al., 2018; Peyraud et al., 2018; Ravindran et al., 2019). Using components of both control and graph theory; we recently employed the method of ‘maximum matching’ and ‘minimum dominating sets’(MDS) (Ravindran et al., 2019) to uncover the way in which viruses exploit human systems, and via their interaction with human gene products, can wrest control of the cell.

Barabási et al (Goh et al., 2007) reviewed network analytical methods applied to heritable disease, and found that monogenic disease, wherein disruption of a single gene leads to a deleterious phenotype is relatively rare, instead, disease is more commonly a result of the interplay between variants in multiple interacting genes within the network. Disease causing mutations have been found to be enriched within regions of the genome that are the most highly conserved across diverse metazoan species (Miller, 2001). Unsurprisingly, this also the case when looking at related genes, with the more conserved gene in a gene pair being more commonly heritable disease associated (Martin-Geary et al., 2019).

Ohnologs, genes that arose and were retained, following, the two rounds of whole genome duplication (WGD) early in the vertebrate lineage, have been found to play an important role in human evolution and disease (Dickerson and Robertson, 2012; Makino and McLysaght, 2010; Richard et al., 2008; Schuster-Böckler et al., 2010; Singh et al., 2015; Van de Peer et al., 2017). These genes are enriched for involvement in developmental processes and subsequent disease (Fotiou et al., 2019); signalling; regulation; cancer associations; and, in particular, autosomal Dominant disease (Singh et al., 2015).Duplicated genes have a strong association with heritable disease more generally, with in the region of 80% of human monogenic disease associated genes having been identified as part of a duplicated paralog pair/family (Dickerson and Robertson, 2012). Small-scale duplications (SSDs) wherein smaller blocks of the genome have been duplicated, and their whole-genome duplication events are thought to have expanded evolutionary capacity by providing additional, redundant, genetic material (Innan and Kondrashov, 2010; Van de Peer et al., 2017). Although many duplicated genes do not appear to be functional

(Dudley et al., 2012), an estimated two thirds, if expressed, are thought to have the potential to contribute to deleterious phenotypes (*ibid*).

Potentially due in part to the fact that whole genome duplication has been found to be the mechanism by which dosage-threshold sensitive genes are able to be duplicated, alongside the small effective population size in hominids (Bailey and Eichler, 2006), ohnologs have become fixed in the population, despite their heavy association with deleterious phenotypes (Ebert et al., 2014). These negative associations are highlighted by the effects on systemic stoichiometry if they are subjected to copy-number aberrations (Birchler and Veitia, 2012; Veitia and Birchler, 2010), and underlined by the fact that they have been found to be refractory to further duplication (Makino and McLysaght, 2010). This disruption of stoichiometric balance caused by disruption or duplication of dosage threshold sensitive genes can have a profound impact on cellular health (Schuster-Böckler et al., 2010).

There are numerous fates that can befall a gene directly following a duplication event. These include, but are not limited to; neo-functionalization where new functions are gained, sub-functionalization, where the ancestral function is divided between the duplicates, and pseudogenisation wherein function is lost (Cañestro et al., 2013). It is proposed that immediately following duplication events, newly redundant genes are able to increase genomic robustness (Hakes et al., 2007) with each new gene being able to compensate for disruption to its partner gene (Gu et al., 2003; Hakes et al., 2007; Hsiao and Vitkup, 2008; Lopez-Bigas, 2004), permitting disease-associated genes, to be retained, in spite of purifying selection

pressures, acting as a form of phenotypic masking. It is not currently known however, to what extent this phenotypic masking occurs in humans.

Here we show that duplication, particularly whole-genome is likely to have played an important role in the evolution of biological systems, and that the propensity of viruses to target the nodes most closely associated with heritable disease, is not related to disease status, but the critical nature of the genes that both allow systemic control, and catastrophic results when disrupted. In contrast to our expectations, however, the role of duplicates may not have been to supplement this fragility, but rather included driving diversification of the system via the provision of alternate paths, therefore contributing to systemic robusticity.

5.3 Methods

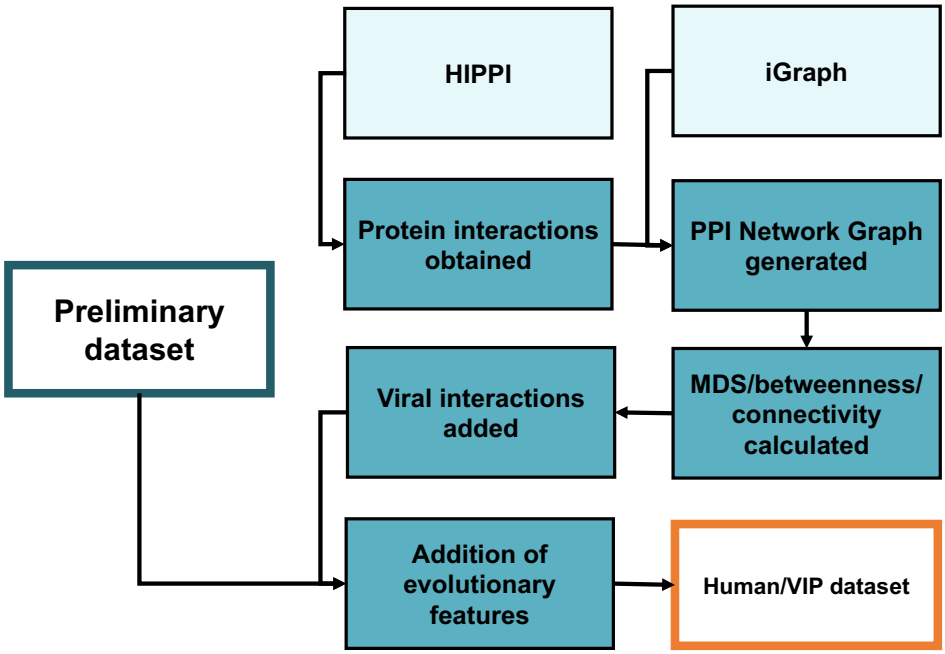


Figure 23: Flow chart showing the methods used to investigate VIPs Evolution and Disease. Primary data, as generated using the general methods is shown with a dark blue border, additional data is shown in pale blue, analysis steps mid-blue and resultant dataset in white with an orange border.

PPI and VIP Data

The initial list of gene interactions obtained from the Hippi database did not contain up-to date approved gene accessions for all of the proteins contained therein. It was therefore necessary to map the NCBI gene names to their most current approved accession. To do this a custom dataset containing approved gene names, and their corresponding NCBI counterparts was obtained from the Hugo Gene Nomenclature

Committee (HGNC 10/02/2020,), and cross referenced with the initial PPI list using perl 5 v18 to produce the list of interactions used in the network.

A list was also compiled containing high quality virus-gene associations, taxonomy , and classifications using data obtained from the International Committee on Taxonomy of Viruses (ICTV) ("International Committee on Taxonomy of Viruses ICTV,").

Processing the VIP data to map between NCBI and UniProt accessions

Each host protein contained within the compiled list of viral interactions as then mapped onto the corresponding NCBI gene using perl 5 v18 and a custom dataset of UniProt IDs and their corresponding NCBI IDs obtained from HGNC (11/02/2020). This data was subsequently used to compute the number, family and class of viral interactions for each gene.

Assignment of gene ages

Gene ages were computed using the general methods detailed in chapter 1.

Data for Ohnologs and other information

A list of approved gene symbols and their corresponding Ensembl IDs was obtained from HGNC. A preliminary file containing evolutionary information for each Ensembl gene within this set was then created using perl 5 v18. Disease associations were obtained from the Online Mendelian Inheritance in man database (OMIM https://www.omim.org) via the genemap2 file (11/02/2020, Supplementary file N). Each row within this file was cross-referenced, and only gene associations with an

observed phenotype, not previously flagged by OMIM as “non-disease” were retained.

Heritability was assigned based on a text search of the genemap2 file for the terms “Dominant” and “Recessive”, any further association achieving the initial criteria, but lacking a match to “Dominant “ or “Recessive” was assigned “Unknown” heritability. All remaining genes were classified as “None”.

Pre-computed gene and family ages (see general methods) were cross-referenced with this list and added accordingly. Haplosufficiency ranks were downloaded from Decipher (https://decipher.sanger.ac.uk) via the HI Predictions Version3 file on the 11/02/2020, and cross-referenced with the preliminary gene list.

Paralog statuses were computed using a combination of data obtained from Ohnologs2 (ohnologs.curie.fr 11/02/2020) and Ensembl. Whole genome duplication status was defined if the gene was present in either the ‘strict’ or ‘intermediate’ human pairs sets, having a q-score of less than 0.001 or 0.01 respectively for both outgroups and self-comparison. Small scale duplication (SSD) status was assigned by cross-referencing any genes not already identified as ohnologs with a list of all paralogous human genes per chromosome and their partners obtained from Ensembl 99, GRCh38.p13. Any remaining genes, not present in either the Ohnologs2 or Ensembl data were termed ‘Singleton’. These attributes were collated into a dataset of 38024 human genes, and subsequently subset using Perl 5 v18, to cover only those genes present in our network (Supplementary file N).

Statistical analysis of Disease/VIP intersect

In order to ascertain the significance of the proportion of genes that fall within the intersection between disease and VIP it was necessary to compare the observed number of genes, with that which we would expect to see given a random distribution. To do this we implemented the following in R: We first restricted the data to only genes with both a known VIP and disease status (including statuses of "none"). All genes with a disease status that was not "none" were defined as heritable disease, and all genes with a VIP type that was not "none" were defined as VIP. Using Plyr (Wickham, 2016) and VennDiagram (Chen and Boutros, 2011) packages both the total number of genes within the test population (VIPs, Ancient genes, Ohnologs and SSDs) and the number of genes from that population observed in the disease/VIP intersection were recorded. N genes (where N is the total number of genes in the population) were then sampled from the full data 10,000 times, and the mean number of genes from this sampling that were also disease and VIP associated was recorded. Using the prop.test function within R, the observed number of intersecting genes for each test population, and the expected number (the mean observed following random sampling), as proportions of the total number of genes in that population were compared, and a X-squared p-value generated.

This process was repeated to look at the specific heritable disease types (Dominant and Recessive), where, rather than retaining all "none" disease associations, defined as heritable disease, we only considered the desired disease type (Dominant or Recessive), the remaining disease genes were retained in the sample pool and treated, for the purpose of the sampling step as 'none'.

Identification and classification of driver nodes

Driver nodes for the network are identified by calculating the minimum dominating set. For a graph G(V, E), where V is set of nodes and E is set of edges, a subset S ⊆ V is called dominating set (DS) if every node in V is either an element of S or is adjacent to an element of S. That is for an undirected graph, any node v ∈ V,v ∈ S holds or there is a node u ∈ S such that there exists an edge (u,v) ∈ E then we say that v is dominated by u. S is dominating set if each node in V is either in S or dominated by some node in S. A minimum dominating set (MDS) is a dominating set with the minimum number of nodes. The MDS forms the driver node set. Since the computation of MDS is NP-hard, we used integer linear programming (ILP) to compute the MDS by assigning 0 − 1 variable to each vertex, where 1 is if v is part of MDS else 0 (Nacher and Akutsu, 2012). A graph can have multiple minimum dominating sets and hence multiple minimum driver node sets with the same size ND. So, each node is categorized based on its presence in the driver node set. If a node is always present in all MDS, it is a critical driver node, occasionally present in MDS then it is an intermittent driver node and if a node is never part of any MDS then it is a redundant node (Nacher and Akutsu, 2014). Degree centrality measures were computed using the igraph package in R (CSARDI, 2006). An example of the final compiled dataset can be seen in table 5.

Table 5 : Example of selected rows and columns from the compiled preliminary data-VIP-network dataset. Data available at: https://github.com/AlexMartinGeary/Hopless_Monsters/blob/master/VIPs/Masterfile_14Feb.csv

Gene_ID	Uniprot_ID	Aproved_gene _name	ENS_Gene_id	Ortho_Age	Paralog _status	Haplosufficiency _Rank	Disease_ Association	Family_ID	Average Shortest Path Length	Betweenness Centrality	Clustering Coefficient	Control .category	Degree	Neighborhood Connectivity	Topological Coefficient	VIP_ type
1	A1BG_HUMAN	A1BG	ENSG00000121410	96	SSD	90.38	None	1461	2.96737983	3.02E-05	0.03162055	intermittent	23	146.782609	0.06335242	NA
2	A2MG_HUMAN	A2M	ENSG00000175899	158	SSD	49.36	None	2732	2.58243957	0.000978	0.03568512	intermittent	157	115.439491	0.01496538	Multi
9	ARY1_HUMAN	NAT1	ENSG00000171428	435	SSD	59.87	None	7849	3.29974993	1.00E-08	0	intermittent	2	379.5	0.53385049	NA
10	ARY2_HUMAN	NAT2	ENSG00000156006	435	SSD	80.21	None	7849	2.84162267	6.98E-06	0.12121212	intermittent	22	245.818182	0.0622695	DNA
12	AACT_HUMAN	SERPINA3	ENSG00000196136	158	SSD	97.04	None	339	2.85468186	0.000126	0.01282051	intermittent	40	121.45	0.03226302	NA
14	AAMP_HUMAN	AAMP	ENSG00000127837	796	SSD	18.04	None	4734	2.78288413	3.39E-05	0.04836415	intermittent	38	215.894737	0.04601472	NA
15	SNAT_HUMAN	AANAT	ENSG00000129673	435	Singleton	74.21	None	0	3.05779383	1.33E-06	0.1	intermittent	5	319.4	0.22062284	NA
16	SYAC_HUMAN	AARS1	ENSG00000090861	796	SSD	NA	None	1570	2.58577383	2.89E-05	0.21812596	intermittent	63	433.587302	0.05508669	RNA
18	GABT_HUMAN	ABAT	ENSG00000183044	796	SSD	41.04	None	749	2.87096416	4.08E-05	0.09090909	intermittent	11	409.181818	0.11322085	DNA
19	ABCA1_HUMAN	ABCA1	ENSG00000165029	176	Ohnolog	9.43	None	731	2.73375938	0.000203	0.04091175	intermittent	59	170.016949	0.03157245	Multi

5.4 Results

Critical driver nodes are molecular drivers in PPI networks; Highly connected and central host proteins are critical in PPIs.

One of the primary goals of the investigation of biological networks is to expand our understanding of the behaviour of the system. As an example, this could mean elucidating the basis of changes in phenotype, such as from a healthy to a diseased state or vice versa. In order to establish the importance of a protein and its role in driving the state of the system, we computed the driver nodes for the human protein-protein interaction (PPI) network, and characterised their biological importance in terms of their interaction with viral pathogens, and involvement in heritable disease. In terms of network controllability, driver nodes are those that can steer the state of the system from an initial state to a final state in finite time. We identified the driver nodes in our PPI network using the minimum dominating set (MDS) method (Nacher and Akutsu, 2012). According to MDS, those nodes that are part of the minimum dominating set are the driver nodes. For a network, given the possibility of multiple traversal routes, multiple minimum dominating sets may exist, we therefore further classify each node in the PPI network as critical, if they are always present in all dominating sets, intermittent, if they are sometimes present and redundant if they do not appear in any MDS. In our network of 18,008 nodes and 359,379 edges, we found 203 critical driver nodes, 16,467 intermittent driver nodes and 1,338 redundant nodes.

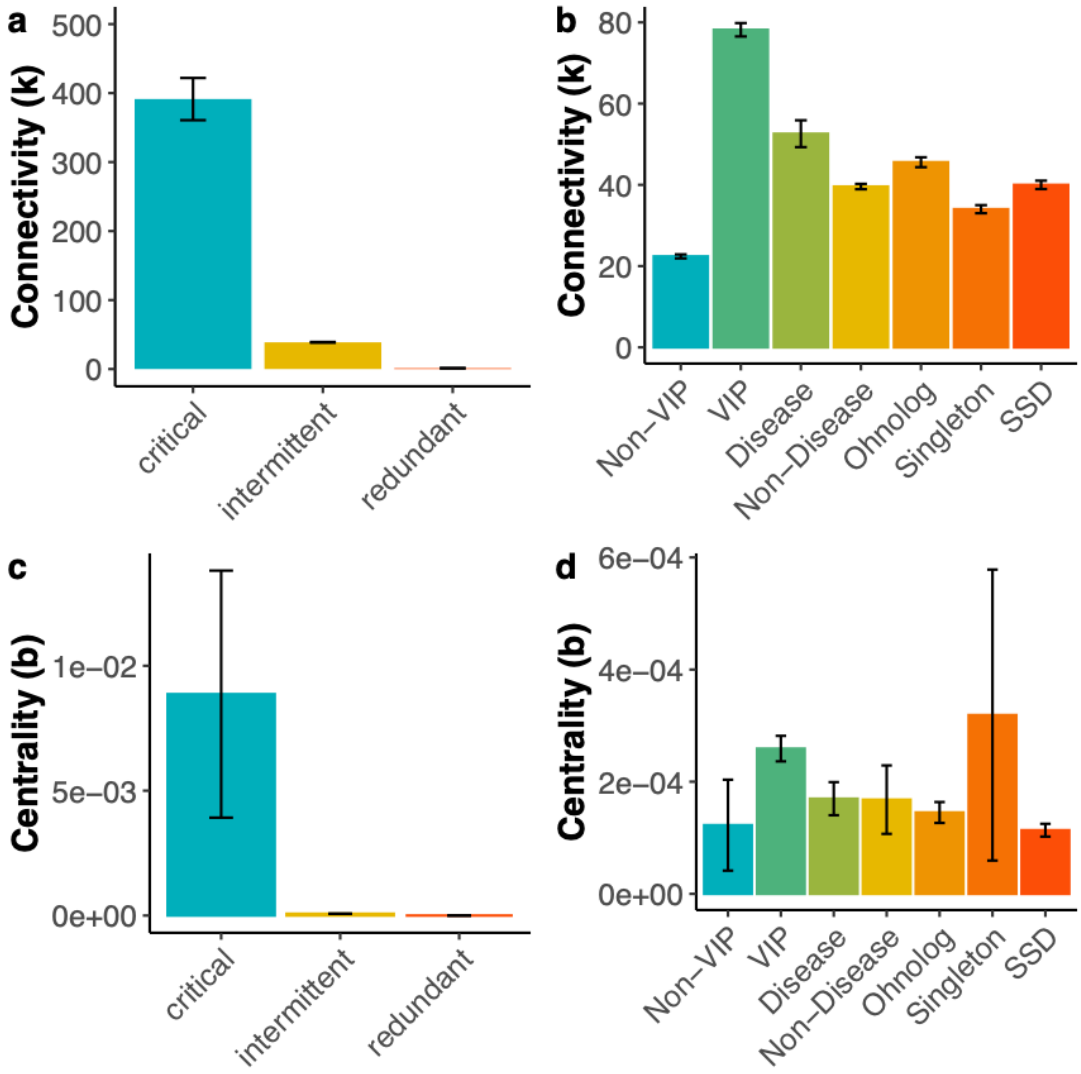


Figure 24 : Average connectivity/ degree (k) of **(a)** driver nodes and **(b)** VIP, disease, and paralog genes in the PPI network. Average betweenness centrality (b) of **(c)** driver nodes and **(d)** VIP, disease, and paralog genes

Driver Nodes

To better characterise to what extent the viral-targeted, and disease associated host proteins affect the functioning and the robustness of the PPI network, we analysed their topological properties. We calculated the connectivity/degree (k), a count of the number of direct interacting partners of each protein, and betweenness centrality (b), a measure of the number of shortest paths that pass-through a given node, for each

protein in our network. Figure 24a shows that the critical driver nodes (mean = 391.33) have a higher k compared to intermittent (mean=38.72) and redundant (mean= 1.25) nodes (P value= < 2X10−16 pairwise Wilcoxon test).

The critical driver nodes (mean=8.87X10−3) had higher betweenness centrality than intermittent (mean=7.16X10−5) and redundant nodes (mean=6.98X10−8) driver nodes (P value= < 2X10−16, Figure 24c). Further to this, the VIPs (mean= 78.18) have both a significantly higher degree k compared to non-VIP (mean=22.40) (P value= < 2X10−16 one-tailed Wilcoxon test) and had higher centrality than non-VIPs (mean=1.23X10−4) (P value= < 2X10−16). It is known that network dominating nodes are integral for signalling and cellular function (Milenković et al., 2011), and this is also what we find. We propose that this leads to highly connected, and important genes becoming ideal targets for viral interaction.

Disease Genes

Among the heritable disease genes, the disease group (mean=52.59) had a significantly higher mean k (39.58) compared to those without a known disease association (P value= 1.408X10−21). Looking at the topological features, these genes had a significantly higher centrality (mean=1.70X10−4) compared to non-disease genes (mean=1.68X10−4) (P value= 9.20X10−26 one-tailed Wilcoxon test), which we would expect given that their disruption is known to lead to deleterious phenotypes.

Paralogs

When looking at the connectivity among duplicated genes we observed that ohnologs (mean=45.57) had significantly higher connectivity compared to SSD

(mean=39.99) and singleton genes (mean=34.01) (P value= < 2X10−16 pairwise wilcoxon test) (Figure 24b). Given that ohnologs are by definition evolutionarily ancient, and have been found to be largely essential this finding makes sense, particularly given their high associations with heritable, largely dominant disease.

Interestingly there is a marked elevation in the variance of centrality (b) scores for the singleton group when compared with the other categories (Figure 24d). Looking more closely at our network it was found that this is caused by an outlier gene (IGFLR1) which, rather than being found in the full 'hairball network' graph, is central to a second fragmented graph containing only a few nodes. This fragmentation results from the fact that, at present, interaction information for biological networks is incomplete, therefore on occasion smaller 'fragmentary graphs' are found to occur alongside, and distinct from the larger network graph.

VIPs target critical driver nodes

Viral interacting proteins are those host proteins that have been identified as either being targets of, or acting against viruses. We have previously shown the importance of driver nodes as potential VIPs (Ravindran et al., 2019). For the purpose of this analysis, we not only identified a confident set of viral interacting proteins, but further classified these into three virus-type sub-categories depending on the type of virus with which the protein is known to interact; DNA, RNA and Multi (those proteins that are targeted by both DNA and RNA). We found a statistically significant enrichment (P value= < 2e−16) of VIPs among critical driver nodes (70%) compared to intermittent (33%) and redundant nodes (4%) (Figure 25). Among these critical-VIP nodes, there was no significant difference between the proportions of those associated with DNA (17%) or RNA (13%) viruses, however, in contrast

with both of these groups, we found that the proportion of genes in the 'multi' group (70%) was significantly elevated (P value= < 2e−16), demonstrating a prevalence of critical status among nodes which also interact with both DNA and RNA viruses. Among the intermittent nodes there is no particular enrichment for any specific VIP category with roughly a third belonging to each virus interacting group (Figure 25).

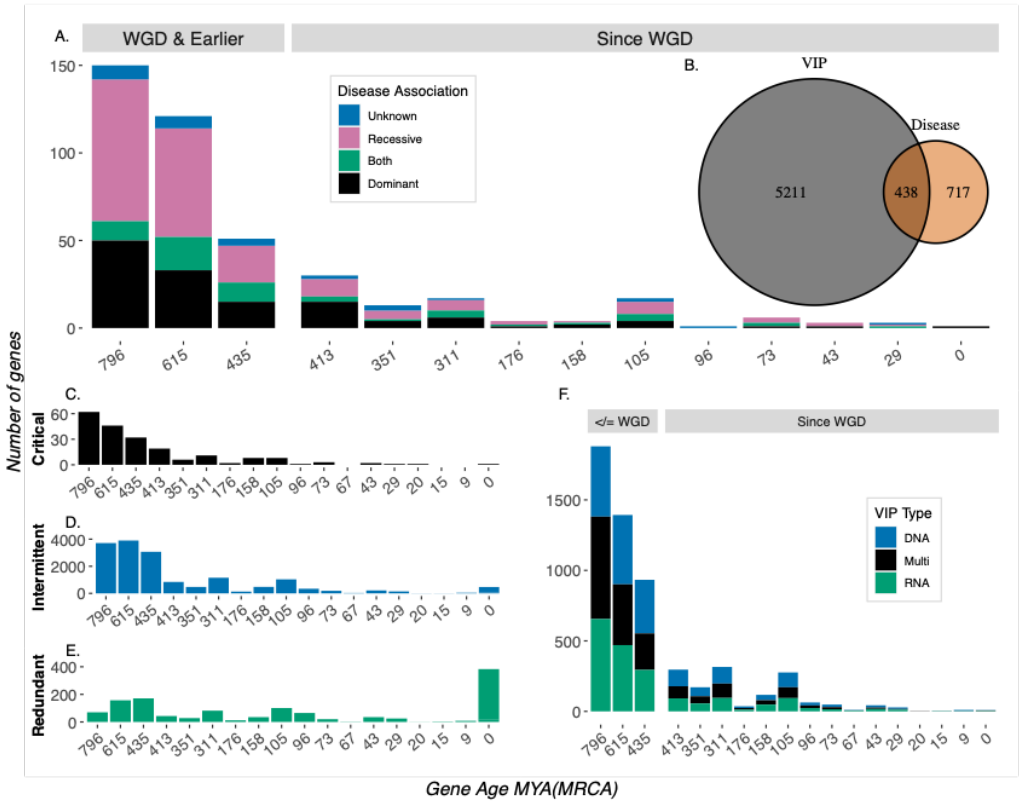


Figure 25 : Bar charts showing the ages of genes in the human genome, as calculated using the most recent common ancestor method of age assignment. A) Stacked bar chart showing the ages of disease genes. Stack colours represent the various disease associations (Unknown disease association -blue, Recessive -pink, Both -green, and Dominant -black). B) Venn diagram illustrating the previously observed intersect between VIP (grey), and Heritable disease associated genes (orange), as observed within our network. C,D & E) Bar charts showing the distribution of ages within the critical (black), Intermittent (blue), and relaxed (green) node sets respectively. F) Stacked bar chart showing the relative ages of VIPs. Stack colours represent viral associations DNA -blue, RNA -green, and Multi -black

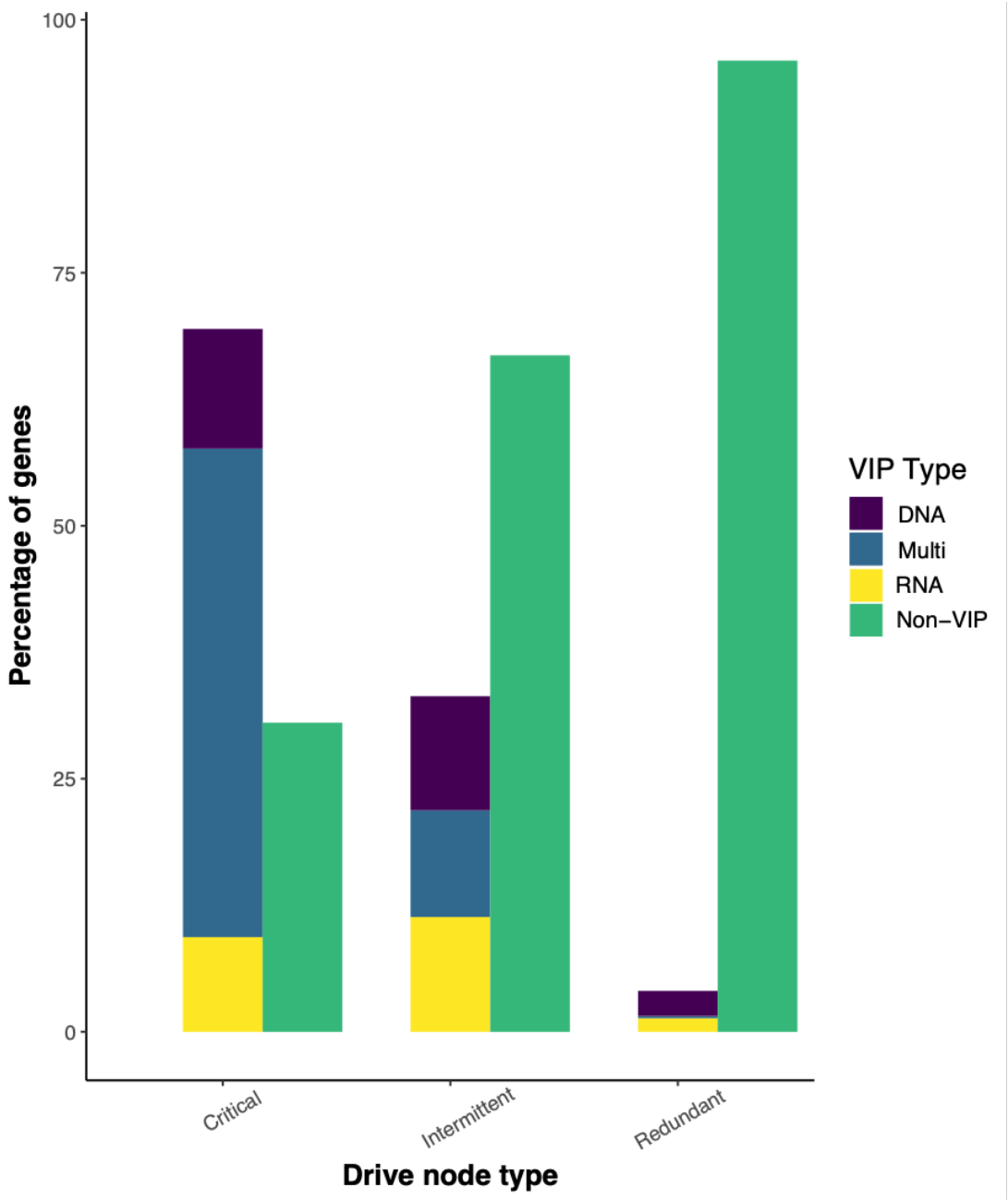


Figure 26 : Comparison of VIP types among driver nodes. Genes that interact with DNA viruses are shown in dark purple; RNA in yellow and both DNA and RNA viruses in blue. Non VIP interacting nodes are shown in the columns to the right in green.

VIPs and disease genes intersect

VIP and Heritable disease genes have been found to commonly intersect (Navratil et al., 2011), and our results confirm this (Figure 26B). Specifically, we found that a significantly greater proportion of VIP genes intersect with heritable disease than

would be expected to occur at random (P value= 0, n=438, 7.8%). This is also the case when comparing VIP status with the known disease sub-categories, as, there is a continued over-representation beyond the expectation of both Dominant (P value= 0), and Recessive (P value= 0.0246) disease groups.

Disease genes are predominantly ancient

A large proportion of human genes arose at the time of, or prior to the last round of whole genome duplication (Figure 26A). Within our network we found that, of the 16,931 genes for which an age was able to be computed, 11,220 (66%) are ancient, having occurred at the time of, or prior to the estimated two rounds of WGD (age cut-off 435 MYA), and 5,711 (34%) having arisen more recently. Given the fact that evolutionarily ancient genes are more likely to be associated with disease (Martin-Geary et al., 2019) it was our expectation that there would be an enrichment of the genetic-disease associated genes present within the ancient (615my or older) group, with very few having arisen since the last round of WGD. When comparing the proportion of disease genes arising at or prior to the last round of WGD with the total proportion of genes arising at this point this is indeed what we see (N=863/75% P value= 2.618e − 09, Figure 26A).

Duplicated genes (Ohnologs) are often disease associated

As briefly highlighted above, there is a strong link between paralog status, specifically genes that are known to have arisen following a duplication event, and heritable disease. In order to explore this relationship in genes within our network, we compared the proportions of dominant and recessive disease associated genes in the three paralog status groups (Ohnolog, SSD, Singleton), with the overall network

proportions (Table 6). We found that in the full gene set (17,696 genes for which data were available), 1062 genes (6%) had a known disease association (’dominant’, ’recessive’ or ’both’) Of which 313 (29%) were dominant, 600 (56%) recessive, and 149 (14%) had both dominant and recessive known disease associations. When looking within the three duplication types (Ohnolog, SSD, and singleton); of the 5,525 Ohnolog genes, 373 (7%) were found to be associated with heritable disease. 151 (40%) of these associations were exclusively dominant, 166 (45%) recessive, and 56 (15%) were found to be associated with both dominant and recessive diseases. Within the SSD gene group, we found that of the 8,316 genes, 476 (6%) were associated with heritable disease. Of these, 119 (25%) were associated with dominant disorders, 273 (57%) recessive, and 75 (16%) both. Lastly, within the 3,855 singleton genes, we found that 222 (6%) were associated with heritable disease. 43 (19%) of which being dominant disease associations, 161 (73%) recessive, and 18 (8%) both.

Table 6 : Profile of disease genes within the differing paralog types (Ohnolog, SSD, and Singleton) vs the whole network

	All Disease	**Dominant**	**Recessive**
Network	6.5%	1.7%	3.3%
Ohnologs	7.4%	2.7%	3%
SSDs	6.1%	1.4%	3.2%
Singletons	6.1%	1.1%	4.1%

A comparison of proportions of disease genes (Recessive/Dominant/Both/Unknown/None) within each paralog group by comparison with the proportions of each disease set observed across the full network, showed that ohnologs are enriched for Dominant disease genes (P value=

1.005e − 05) and depleted for genes without a disease association (P value= 0.02244). Singletons, however, were found to be depleted of both dominant, and dominant and multi-heritability associated genes, but enriched for recessive disease associated genes.

Ohnologs are enriched for intermittent nodes

It has previously been identified that Ohnolog genes are more likely to be associated with dominant disease (Fotiou et al., 2019; Makino and McLysaght, 2010; Martin-Geary et al., 2019). However, we wanted to further explore the role of paralogs, in particular the overlap between duplication and network control. To explore this, we calculated the proportion of genes belonging to each control category found within each of the three paralog statuses and compared these proportions with the total proportion of each control group found in the full network. We found that ohnologs are enriched for intermittent nodes (2.2e-16) and depleted for redundant ones (2.2e-16), SSDs are depleted of intermittent (0.02753) and enriched for redundant nodes (0.04631) whilst, perhaps unsurprisingly, singletons are depleted for critical (0.003143) and intermittent nodes (2.726e-09) and enriched for redundants (7.935e-14). When juxtaposing this analysis by looking at the proportions of the different paralog groups within each control category compared with the network totals, we find that critical nodes are significantly less likely to be singletons than would be expected (0.0004491). Intermittent nodes are more commonly found to be ohnologs than expected (0.0261), and Redundant nodes are significantly depleted for ohnologs (2.2e-16), whilst being enriched for SSDs (0.001645) and Singletons (2.2e-16). This means that evolutionarily important genes, when duplicated, have a propensity to become intermittent. This, we propose, is likely due to the subsequent rapid sub-

functionalisation of genes within these duplicated pairs (Su, 2005). This is particularly true of ohnolog genes, likely due to their strong links with essential function and fragility. This finding is indicative that whole genome duplication acts as a buffering mechanism in which genes with a heightened propensity for fragility (Martin-Geary et al., 2019) may become more robust to variation, whilst maintaining important functions.

How Driver/paralog status relates to VIP status

We next explored the relationship between paralog status and a gene's propensity to be viral interacting (Table 7). We found that of the 5,649 VIP genes in the full network 1,900 (34%) were Ohnologs, which, when compared with both the network proportions (5525/31%) and expected values (mean exp= 1,764) is a significant enrichment (network P value=0.0007425, expected P value=0.006591), 2,469 (44%) were SSDs (Network= 8316/47%, mean exp= 2,655), a significant depletion (Network P value= 1.717e − 05, expected P value= 0.000465), and 1,280 (23%) were singletons (Network P= 3,855/22%, mean exp P= 1,230), which is not significantly different to either the full network or the mean expected values.

Table 7 : Raw observed/expected number of critical, intermittent, and redundant nodes in each paralog status set, alongside their comparative P-values (prop.test). Statistically significant differences are shown in green.

	Ohnolog			SSD			Singleton		
Driver node	**Observed**	**Expected**	**p-value**	**Observed**	**Expected**	**p-value**	**Observed**	**Expected**	**p-value**
Critical	74	62.30	0.261	106	93.70	0.2621	23	43.48	0.00896
Intermittent	5258	5051.92	0.7024	7569	7604.56	0.01482	3425	3525.16	0.1805
Redundant	193	410.28	$2.2e-16$	641	617.89	0.392	407	286.59	$1.38e-07$

Critical and intermittent nodes tend to be ancient

When assessing the relative ages of genes in the differing network control groups, we found that, of the 202 critical nodes for which ages were calculable 140 (69%) are ancient, and 62 (31%) have arisen since the last round of WGD (Figure26 C). Of the 15,856 intermittent nodes for which age was calculable, 10,682 (67%) were found to be ancient, and 5,174 (33%) have arisen since the last round of WGD (Figure 26 D), and for the 873 redundant nodes, we found that 398 (46%) are ancient, with 475 (54%) having arisen since the last WGD (Figure 26 E). Comparing the proportions of ancient genes between each of the three control sets we found that, whilst the proportions of ancient genes in the critical and intermittent node groups are broadly similar, they are both significantly enriched for ancient genes by comparison with the redundant group (Pairwise P value= $1.8e-13$ and P value= $2e-16$ respectively). This is consistent with younger genes being less integral to network control, and having less involvement in important biological function.

Pathways analysis of VIP and disease genes

In this analysis, we show that heritable disease and VIP status can provide valuable information regarding the evolution of network control nodes. However, they alone cannot explain the complex mechanisms involved in the disease progression. To explore this issue further, we performed pathway enrichment analysis using Reactome (Jassal et al., 2019). In total we analysed 438 proteins that are both VIPs and disease associated. We found that the top pathway enriched for VIP-disease genes relates to cellular response to stress, a pathway essential to maintain tissue homeostasis when the metabolic and signalling processes are perturbed (Galluzzi et

al., 2018), with others relating to transcription, translation processes and, unsurprisingly, the immune system (Table 8).

Table 8 : Top 10 enriched pathways for VIP and Disease genes

Pathway	P-value	FDR
Cellular responses to stress	1.11E-16	4.49E-14
RNA Polymerase II Transcription	1.11E-16	4.49E-14
Generic Transcription Pathway	1.11E-16	4.49E-14
Translation	1.11E-16	4.49E-14
Cell Cycle, Mitotic	1.67E-15	5.36E-13
Regulation of expression of SLITs and ROBOs	2.33E-15	5.36E-13
Signaling by ROBO receptors	2.33E-15	5.36E-13
Eukaryotic Translation Initiation	8.06E-14	1.44E-11
Cap-dependent Translation Initiation	8.06E-14	1.44E-11
Processing of Capped Intron-Containing Pre-mRNA	1.34E-13	2.16E-11

To further understand how the properties of our data intersect with biological function more broadly, we selected a small portion of one of the top ten, and highly conserved pathways in eukaryotes (Wahle, 1999) for visualisation. Figure 27 shows the locations of the nodes within our network that fall into each combination of driver and paralog status. Predominantly, the nodes within this sub-pathway are intermittent, which is consistent with our expectations given the importance of the pathway, and the large number of intermittent nodes in the Human PPI network. What is interesting to note, is the relative frequency with which critical nodes are also Ohnologs within this pathway. Whilst It is beyond the scope of the current analysis to investigate this further, it raises some interesting questions regarding the possibility that there may be pathway specific relationships between driver and paralog status.

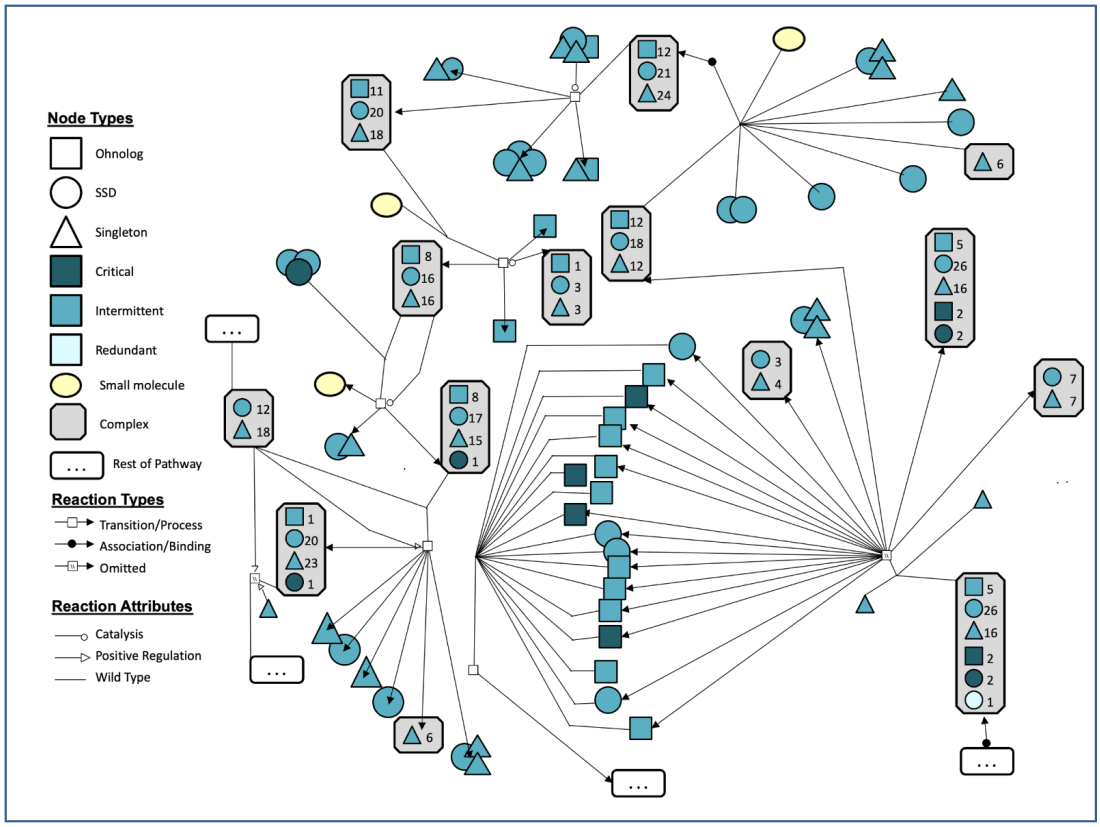


Figure 27 : An example subset of the 'Processing of Capped Intron-Containing Pre-mRNA' Pathway modified from Reactome. Ohnologs are represented as squares, SSDs as circles, and Singletons as Triangles. Critical nodes are shaded in dark blue, intermittent nodes in mid-blue and redundant nodes in pale blue. Small molecules, such as ATP are shown as yellow ovals, and protein complexes are shown as grey rectangles, with the constituent proteins totalled in each paralog-driver group.

5.5 Discussion

Our observed correlation between heritable disease and VIP status of certain genes is not a result of VIPs targeting disease associated genes, but rather the underlying evolutionary importance of these genes leading to ’new’ associations.

Critical driver nodes have a higher degree, and betweenness centrality compared to intermittent and redundant. This is indicative of their importance as ’hubs’ within the network, and the fundamental biological processes in which they are components. We have previously found that certain viruses target critical nodes that would allow them to effectively wrest control of the host cell (Ravindran et al., 2019). Our findings here, that VIPs of a much broader range of viruses are found to have significantly higher degree and betweenness centrality therefore supports both these prior findings, and the utility of the MDS method of modelling complex and dynamic biological systems.

Here we show the strong relationship between VIPs and critical nodes, the majority of critical VIPs being proteins which are targeted by both DNA and RNA viruses. This has, to an extent been alluded to previously, and explained as a propensity of viruses to target hubs for efficiency. What we would suggest however, is that it is not the high degree that renders a node more likely to be a VIP, but instead the likely fundamental biological processes that these genes encode for that makes them optimal targets for both DNA and RNA virus types, specifically the exploitation of biological machinery that is key to both replication and immune evasion.

It is not unexpected that, when incorporating heritable genetic disease associations, we find that disease associated genes also have a higher degree and betweenness centrality than their non-disease counterparts. The fact that we are seeing such high values here, underlies the capacity of these genes to lead to undesirable phenotypes if disrupted. High betweenness centrality of disease genes has previously been explored relative to pleiotropy and disease diversity (Chavali (2010) and Zhou (Barabasi 2014) and used as a feature in the prediction of disease genes (Ozgur 2008). We find that more disease genes are VIPs than would be expected to occur at random, regardless of disease type. This is likely due to the fact that disease associations expose exactly the features of important genes that viruses preferentially exploit, i.e., the high levels of connectivity and betweenness centrality that we have shown are key aspects of critical driver nodes. The underlying functions of these molecules are likely essential and their products highly expressed, which in turn leads to a greater probability of heritable disease associations if they were to be disrupted. The intersection between heritable disease and viral interaction we suggest, is therefore a correlative, rather than causal one.

D'Antonio et al (D'Antonio and Ciccarelli, 2011) found that highly conserved, ancient genes are commonly 'hub' proteins. As discussed earlier, genes arising as a result of whole genome duplication tend to be under greater levels of negative selection (Dudley et al., 2012) and, given that the two rounds of WGD in the vertebrate lineage occurred 441mya - they are by definition ancient. It is unsurprising therefore that, in line with the findings of D'Antonio et al, we have found that ohnologs have a significantly higher degree than their small-scale and singleton counterparts. It was found that the observed occurrence of ancient genes significantly exceeded the expectation, when all disease genes were considered

together, and when looking only at genes with recessive heritable disease association, but not genes associated with Dominant heritability. The most likely explanation for this is the overlap between disease and age. Ancient genes are more likely to code for important/conserved proteins that contribute fundamental functions, so an over-representation of these genes within the intersect of disease and viral interaction is to be expected. What is interesting however is that whilst there is a minor elevation, there is not a significant enrichment for Ancient genes in the VIP/Dominant disorder group. Whilst it is possible that this is due to the small number of genes in this group not allowing for detection of statistical significance, it may also be hypothesised that there will be a small number of ancient genes which are so important to the host that they do not make for good viral targets i.e., due to the likely profound negative implications of disruption of these molecules, and the subsequent negative impact this would have on viral 'success'.

Within the intersection of VIP and disease it was found that, whilst there was a minor elevation in the number of ohnologs within the 'all disease' and dominant groups, paralog status does not appear to drive a gene's propensity to be simultaneously disease associated and viral interacting. VIP status aside, we have found that contrary to our expectations given the properties of ohnologs and their high propensity to be essential, Ohnologs are not enriched for critical nodes, but rather intermittent ones. This is also true of small-scale duplicates. An explanation for this may tie-in to what we know about the main types of functional fate of duplicated genes. For example, a certain group of duplicated genes, those that retain ancestral function following duplication are able to compensate for their partner. Provision of functional redundancy in this way is a phenomenon known as phenotypic masking. This has recently been investigated (El-Brolosy et al., 2019;

Ma et al., 2019) and linked to the nonsense mediated RNA decay pathway via nonsense-induced transcriptional compensation. A second possible functional outcome arising from duplication is sub-functionalisation, wherein the ancestral function is subdivided between the 'daughter' genes. This, it is hypothesised, occurs rapidly following duplication events, and can be traced via a reduction in average number of splice forms in ancient duplicates and large gene families (Su, 2005). It may be the case that what we are seeing here is indicative of a division of ancestral function 'dampening' the impact of ancestral node removal, due to a portion of the ancestral function being provided by the sibling gene. It is possible therefore that what our results show is a combination of phenotypic masking and subfunctional dampening, turning what would otherwise have been 'critical' functions to a more 'intermediate' state. The reason that we are seeing a marked over-representation of SSDs as opposed to ohnologs may be due to ohnologs' need to retain dosage as part of a complex. This means that SSDs have a greater level of plasticity (and therefore probability of being intermittent) due to not being intrinsically linked with the fates of their interacting partners, and therefore posing a reduced risk of stoichiometric imbalance.

Whilst we found that the only instance of a greater number of genes being found in the pre-WGD 'ancient' group than the post-WGD set was within that of critical nodes, this increase was not statistically significant. However, in terms of depletion of genes within the ancient set, we found that redundant genes were significantly less likely to be ancient. This means that ancient genes are more likely to exhibit some degree of network control, albeit not necessarily in a critical capacity.

These combined findings from our analysis of evolutionary and disease related properties of the human PPI network, identify that paralog status may have played a significant role in the evolution of biological pathways, but not in the way we would expect. Rather than the underlying fragility of ohnologous genes leading to their being critical to the system, we find that ohnologs, via the provision of redundancy, may have provided new routes within the network, resulting in a greater level of robusticity than would have been the case had their fragile progenitors remained un-duplicated.

CHAPTER SIX

DISCUSSION: WE ARE THE DESCENDANTS OF HOPEFUL MONSTERS

"it must be the case that the hopeful monster

will not, in a crucial way, know what is"

(Jackson, 1997)

Understanding evolution is vital to the study of human health and disease. With a greater understanding of the mechanisms and processes that inform the evolution, not only of humans, but also their pathogens, we have provided ourselves with vital ammunition in the war on disease. Notably, at the start of 2019 the Institute of Cancer Research made headlines, recognising the importance of an evolutionary understanding of disease, by pledging £75 million to found a centre of expertise in cancer evolution (Le Page, 2019; The Institute of Cancer Research, 2019). It is not, however, solely evolution that is relevant to this thesis that has been making waves in the science community, but also, following the closure of one of the U.K.'s foremost animal facilities (Else, 2019; McKie, 2019), the increased pressure to produce robust and reliable in-silico models for scientific research.

Despite the relative youthfulness of computational analysis of biological data, computational methods have, by and large, been embraced by the scientific community. Particularly as performed from an evolutionary perspective. Phylogenetics, network analysis, and machine-learning methodologies are increasingly relevant in day-to-day biological research, and robust computational skills are essential to the modern scientist. As has been shown in chapters 3:5, the application of diverse computational and evolutionary methods has the power to shed light on a wide range of human concerns, with the understanding of heritable disease, parasitic co-evolution, and virus susceptibility, being just a few of the areas in which computational biology has proven integral.

In order to fully understand the shaping of the human genome, we must understand its journey; the root of all life is thought to have emerged on earth ~3.9 billion years ago (BYA) (Tashiro et al., 2017), with the division of animalia, fungi and plants beginning at ~1.5BYA(Wang et al., 1999), and vertebrates arriving on the scene at an estimated 530 million years ago (MYA) (Conway Morris, 2000; Newman, 2002). It is known that at least one whole genome duplication (WGD) occurred shortly before this point, however, the timing of the second is hotly contested, potentially having occurred either directly following the first round, or shortly after the emergence of vertebrates (Donoghue and Keating, 2014; Holland and Ocampo Daza, 2018; Keating *et al.*, 2018). Even without a pinpointed date for the second round, it is clear these duplications directly correspond with mass speciation events that are indelibly linked with all vertebrate life, as, coincident with the first proposed WGD was the onset of the Cambrian explosion, which ultimately can be traced as the origin of the many modern metazoan species (Deline *et al.*, 2018). Just 2-3 hundred million years ago the next major event in the human lineage occurred, with the appearance of the sex chromosomes (Bachtrog, 2013), and concluded with their morphogenesis into their current form via decay, some time after 166 MYA (Warren *et al.*, 2008). Jumping rapidly ahead to our very recent evolutionary history, humans split from chimpanzee between 4 and 7 million years ago (Gibbons, 2012) with duplications likely having also played a significant role in this event (Bailey and Eichler, 2006).

The earliest hominins are thought to have arisen in the region of 4 MYA, with the ancestor of our closest non-human hominid (NHH) relatives; Neanderthals and Denisovans, dividing from our shared ancestor in the region of 734 thousand years ago (Rogers *et al.*, 2017). This was followed just 10,000 years later by the divergence of Neanderthal and Denisovan, likely as a result of geographical isolation following their migration out of Africa (Rogers *et al.*, 2017). Their ancient Homo sapiens sapiens (Humans) cousin arose from the shared ancestor with Neanderthal and Denisovans somewhere between 1 million and just 100 thousand years ago (KYA) (Ackermann *et al.*, 2019). Human expansion out of Africa to join their NHH relatives occurred between 45 and 60 KYA, however, we now know that the genetic fidelity of these early migrants was quickly compromised, as serial founder effects rapidly led to huge loss in genetic diversity (Deshpande *et al.*, 2009; Henn *et al.*, 2019), which can still be seen today when comparing African populations with those outside of Africa (Henn *et al.*, 2019). Due to the differing proportions of introgressed genes; those found to have evolved in the NHH rather than Homo sapiens sapiens lineage in differing populations, it is now known that admixture between Neanderthal and Denisovan in the extant human lineages, occurred following the initial waves of ancient human migration out of Africa (Ackermann *et al.*, 2019; Enard and Petrov, 2018b), and at different frequencies in different populations across the globe.

It is unclear at what point on our journey across the earth we began encountering both pathogens and heritable disease, but as discussed in chapters 3:5, the duplication events that shaped our genomes, are also key to our susceptibility to both. From archaeological data and evolutionary analysis, we are able to infer that prokaryotic pathogens such as Treponema Pallidum, eukaryotic pathogens such as

Trichuris trichura, and a plethora of viruses, likely have been infecting humans for thousands of years (Anastasiou and Mitchell, 2013; Dutour, 2013; Enard and Petrov, 2018b; Johnson, 2019; Søe *et al.*, 2015, 2018), with viruses not only having co-opted the important features of certain human genes, as discussed in chapter 5, but also, in the case of RNA viruses, playing a large role in shaping whole swathes of our DNA (Enard and Petrov, 2018b). The co-option of introgressed material from our NHH relatives has proven vital to our survival as a species, having a key hand in shaping the immune genes that provide us with resistance to pathogens. Further to this however it has also, to a degree counteracted serial founder effects which occurred following our migration out of Africa, by introducing an important reservoir of tolerated variation into different human populations (Ackermann *et al.*, 2019).

Whilst relatively little genetic change has occurred in the short time since the extinction of our NHH relatives, humans as a species have come a long way. A large part of that progress has been made in the fields of science, and medicine. As discussed in the introduction to this thesis, the sequencing revolution has allowed us to gain unprecedented insight into the major mutations that have brought us to where we stand today, but it has also provided us with a good deal of insight into human disease. As previously discussed, (chapter 5) viruses have played a key role in shaping the human genome, but now, given our refined understanding of how they evolve and interact with human physiology we are able to limit their annual death toll with the use of vaccines. Modern medicine has provided an evolutionary escape, where previously, outbreaks would have resulted in a large death toll such as the case of the H1N1 pandemic in 1918, which killed an estimated fifty million people (Chandra *et al.*, 2018). It is our hope that analyses such as our study in chapter 5 may help to push these advancements further. The placement of viral targets within the

broader context of both pathways, and the larger biological system provides an insight into the key features of genes involved in viral infection, and can inform the way that new pharmaceuticals act to combat viral attack, without disruption to the broader network in which they are integral components.

Similar to the advancements in technologies for the treatment of viral infection, with the use of increasingly sophisticated methods we can now secure the survival of many sufferers of both heritable, and somatic diseases. For example, in 1995 the overall 8-year survival rate of patients undergoing the, then-new trial for treatment of Acute Myeloid Leukaemia, of which I was a member, was 38-40% whereas, by 2015 this figure had increased to 74% in the child cohort alone (Burnett *et al.*, 2010; Reedijk *et al.*, 2019). It should be noted that, to date the majority of causal variants linked to genetic disease have been uncovered in protein coding regions. Whilst it is likely that the lion's share of disease is indeed attributable to coding variation, there are vast swathes of the genome that contain important functions, for example regulatory elements and splicing, that are to date poorly characterised, but likely to be deleterious if disrupted. There is, it seems far more to health and disease than can be found in the protein coding regions alone.

Whilst it is beyond question that advancements in the treatment of human disease are outstanding, it also means that, to a degree we have managed to evade purifying selection by providing individuals who are potential carriers of deleterious genetic traits, to survive to reproductive age. It is important to make clear that, despite the often bleak or negative terminology used when discussing heritable disease, what we call disease is not universally pernicious, and in a number of cases, particularly within some of the impacted communities, this diversity is seen, rightfully, as something to be celebrated. This increasing saturation of genetic

disease in the human population, does however require us to obtain a far greater understanding of the basis and mechanisms that lead to heritable disease, in order to, for want of a better word, either 'correct' these variants, or, at the very least account for them. In chapter three we found that heritable disease, within gene pairs, most commonly tracks to both the most conserved and haploinsufficient of the two. This type of understanding is integral to both downstream prediction, and discovery of specific disease associations, which in turn, we hope, will lead to the development of novel therapeutics, or the repurposing of existing ones.

The role that large-scale mutations, particularly duplications play in the human condition is complex, both in terms of the nuanced ways in which they have shaped both disease and health, but also in their key involvement in the evolution of lineage specific complexity. Gene duplication is now known to be highly associated with eukaryotic transcriptional complexity and morphological diversity, and is linked to both chromatin remodelling, and epigenetic factors not present in prokaryotes (Hajheidari et al., 2019). Even within prokaryotes, duplicated genes have been found to be important, encoding for functions that aid in proliferation and ultimately, species survival (Bratlie *et al.*, 2010). This association with survival is not limited to prokaryotes however, as gene and genome duplications, it is proposed, have contributed to reduced extinction risk in many eukaryote species, via provision of a heightened ability to adapt, mutational robustness, increased evolutionary rates, and functional redundancy (Crow, 2006). Our interpretation of the findings we set forth in chapter 5 are in support of this increased plasticity afforded by duplication. The fact that duplicated genes provide intermittent control of human biological networks in place of their, likely-critical progenitors, is suggestive of a 'rewiring' of underlying fragility that has contributed to an increase in systemic robustness, and

may to a degree, implicate a level of defence against viruses, which favour the targeting of critical nodes.

Relatively recently, in evolutionary terms, it is evident that duplication has played a continued part in shaping the genomes, and in turn the physiology of complex organisms. The innate immune system, for example, is likely to have appeared at the same time as the emergence of vertebrates (Boehm *et al.*, 2012), with later addition, due to duplication, of class I MHC genes occurring in the human line some time following our divergence from birds and sharks (Kulski *et al.*, 2002). Lineage-specific biological differences between five species of hominoid including human and great apes, have been identified, and directly linked with copy number variations (Fortna *et al.*, 2004; Hurles, 2004), alongside human specific traits directly stemming from transcription derived gene fusion as a result of segmental duplication (Am et al., 2019).

The planet on which we live is populated by diverse species, each with a level of interdependence on each other, many having shaped each other's histories. We have discussed the co-evolution of pathogens, and the ways that viruses have helped mould the human genome, however not all interactions are negative. Early farming practices, for example, involving the cultivation of maize and secondary animal products such as milk have directly impacted human physiology, behaviour, and genetics (Prohaska *et al.*, 2019). Many species show high levels of genetic adaptation to their ecological niche. Humans, under the right circumstances have the ability to adapt to changing oxygen abundance, with modern humans living in high altitude areas having been shown to genetically adapt to the associated reduction in oxygen levels (Azad *et al.*, 2017; Gnecchi-Ruscone *et al.*, 2018; Horscroft *et al.*, 2017; Yang *et al.*, 2017). When habitats change more rapidly than a species' ability

to adapt, the results can be detrimental. Directly following the mass expansion of biodiversity due to WGD during the Cambrian explosion, were the Botomian and Toyonian periods, two major periods of mass extinction. These events it is now understood, were due, at least in part, to atmospheric oxygen depletion (He *et al.*, 2019).

As discussed in chapter 3, a good deal of biological complexity is liable to be due to the asymmetric evolution of both small-scale and whole genome duplicated gene pairs, which not only has shown to allow the provision of repurposable genetic material via the relaxation of constraint on the more divergent, less ancient gene. But has, as illustrated in chapter 5, in certain circumstances provided a degree of redundancy, wherein a second copy of an important molecule may be available to act in a compensatory manner, should its partner be lost (El-Brolosy *et al.*, 2019b; Ma *et al.*, 2019). The asymmetric relationship of duplicates in evolution, alongside added variation, and large-scale mutation in general, are therefore important to the human condition in a bi-fold manner, wherein they have provided both a degree of redundancy, and repurposable genetic material that has directly contributed to features that define our species.

It is not the provision of new material alone that has been important to human health and disease however, but the way in which this material is transcribed and expressed. A recent study into differential expression of essential duplicates found that, within gene pairs where both genes are found to be essential, i.e., resulting in a lethal phenotype should one copy be lost, expression of these genes tends to occur during different stages of development. (Kabir *et al.*, 2019). This finding is important to our understanding of the different fates of duplicated genes, as, discussed earlier and in chapter 3, compensation has provided an important

mechanism for the prevention of disease phenotypes, which has clearly been illustrated to be impossible in these instances. In line with this - rather than providing increased redundancy via the provision of compensatory copies as found in chapter 5, in some instances duplication via processes such as the subfunctionalization that lead to the division of essential function, may result in both genes in a duplicated pair becoming essential as has likely occurred in the more highly constrained ohnolog gene pairs in chapter 3. When this is the case, the result is a doubling of potential disease in the genome, and an overall increase in organismal fragility. This is particularly significant when you consider the underlying cause of disease associations in many of these genes, that of ancient fragility (chapter 3). Given the fact that whole genome duplication has occurred not once, but twice, it may be the case therefore, that whilst the first round afforded some degree of protection, the second introduced greater numbers of potential heritable disease associations into the system.

In chapters 4 and 5 we have discussed the role of duplicated genes in human susceptibility to pathogens. For both parasites and viruses, we have unveiled the underlying features by which ancient duplicates can be co-opted to the benefit of the pathogen. Largely this is due to the systemic importance of genes of this kind, in that they tend to be heavily conserved in both sequence composition and across diverse species. This is important to pathogens for two reasons, firstly the relative stability of these genes makes them less able to accumulate ‘evasive’ variants to protect against such interactions, as they are under heavy purifying selection. And secondly, their presence in diverse host species allows a greater potential for them to jump species boundaries and be infective to a greater range of hosts.

As discussed, the fixation of large variants has proven exceptionally beneficial to human evolution, however, it must be noted that beneficial mutations of this kind are fleetingly rare. It is now known that significant developmental disorders such as Edwards and Patau's syndromes are caused by chromosomal abnormalities and trisomy (Sun *et al.*, 2019; Williams and Brady, 2019). There is no cure for these conditions, however screening is in place within the UK to forewarn parents if a child is at risk of suffering from these and other genetic disorders (Sun *et al.*, 2019). This early detection however can present some major ethical concerns that underlie much research into genetic medicine. Whilst there is not the scope in this document to explore the full array of these concerns, one of the foremost, that has become increasingly prevalent in recent times is – should humans be allowed to modify their DNA? and, if there is no repair possible, at what developmental cut-off is termination acceptable? Ultimately these are issues which fall to the individual, or their representatives to make, however legislation, informed by scientists and medical professionals is required to prevent extreme cases of misuse. As technologies and techniques for the detection and modification of deleterious variants is accelerating, so too is the desire to rapidly introduce these technologies into mainstream medicine. Given the often-long timeframes that this can take however, a growing collection of individuals have decided to take matters into their own hands. These individuals, known as 'biohackers' are procuring cutting edge technology and attempting to treat themselves at home in an unregulated manner, often to detrimental effect. In August 2019, the first laws were passed in the US to regulate "biohacking" (Gent, 2019), which, while being unable to prevent people experimenting on themselves, will restrict the sale of gene therapy materials by "biohacking" companies to individuals.

In chapter 3 we discussed how important genes with ancient origins are not tolerant of variation, as, when they do vary this often leads to disease due to the underlying 'fragility', or importance of these genes' progenitors. This turns on its head the suggestion that duplication is permissive for disease by providing redundancy, and is a whole new way of thinking about the mechanics of the association between ohnologs and disease, linked to splitting first duplicates into small scale vs WGD and then identifying the "older" WGD as being disease associated.

The extent of this is highlighted in chapters 4 and 5, where we find that despite the obvious negative connotations of these genes being targets of viruses and other parasites, and are subsequently under pressure to diverge, this is likely surpassed by their need to remain constrained. This is an important finding, as, following the identification that viruses tend to target disease associated genes, it has been assumed that the link between them was through the molecules overlapping, however, we show that this is instead due to viruses targeting core/ancestral functions.

We also have discussed recent findings that within complex organisms, due to elevated levels of essentiality, reduction in robustness and redundancy occur. This poses real limitations to the future evolvability of both ancient genes arising as a result of WGD, that make up, in the region of a third of the genome, and genes which have evolved novel functions, meaning that potentially relatively little of the human genome has an ample capacity for adaptation.

With regards the future directions of human evolution, there are many pressures on a species' genome that can influence its longevity. These can include degree of diversity, which allows continual procreation without the negative effects

of inbreeding, directly tied into effective population size; the rate at which the species is able to accumulate new variants; the birth rate and 'litter' sizes of the species; and to what degree they can tolerate large-scale mutations. Living fossils, organisms found to exhibit little morphological change over time, and with few extant related species (Turner, 2019), have been found to have undergone a significant "*deceleration*" in molecular evolution (Soltis *et al.*, 2002), and require both a low extinction rate and low speciation rate in their lineage (Bennett *et al.*, 2017). By contrast humans have a higher evolutionary rate. As shown throughout this thesis however, humans struggle to adapt with any degree of immediacy without negative consequences. Dwindling within-species diversity, small effective population size, small litter sizes meaning lower likelihood of a 'temporally beneficial' large-scale mutation becoming fixed, and a rapidly changing environment, could mean that, for humans it may already be too late.

It is clear that the imposition of false dichotomies when looking at cases such as that of Goldschmidt's hopeful monsters, the dosage balance hypothesis, or the Force/Ohno argument of the existence of asymmetry, can lead to obfuscation of the true biological picture. As is the case with Goldschmidt and Darwin, it was not they who defined the argument as dichotomous, however, the result was, that for decades our understanding of biology may have been marred by Goldschmidt's theory having failed to gain traction. Since the 1950s Richard Goldschmidt's hopeful monsters hypothesis has, to a degree been revived, and did not in fact, become a dead end, but was he right? Looking at what is known about the evolutionary history of humans as presented here, a clear argument can be made for the role of large-scale mutations in vertebrate evolution being a significant one. They have provided sizable quantities of DNA for repurposing, that have allowed us to thrive and gain novelty, complexity,

and environmental plasticity. This however is not the end of their involvement, as; accompanying these benefits is a decrease in redundancy in line with the increased essentiality that accompanies complexity. This has meant that the ancestral fragility underlying many duplicated genes and structural variants, has provided a platform for the emergence of more of the same.

Evolution, at the molecular level, is a slow lottery, and we could be forgiven for feeling like we are static in the situation – however, we are not. It will be difficult to truly predict where we go from here, can we survive in a changing environment if our ability to accumulate beneficial large-scale mutations is impaired, and, if we do acquire them, due to small litter sizes could they become fixed? Are point mutations alone enough, not only to provide us with the novelty needed to cope with change, but also the diversity required to avoid species-wide deleterious homogeneity? Or, is the curtain closing on *Homo sapiens*?

Given the rate of advancement in the field of genetics it is not beyond the realm of possibility that we could soon be able to supplement natural evolutionary processes and introduce new redundancy to the system, encourage our environment to adapt to us, or better still, slow the current rate of change. Of course, given the state of global affairs as it currently stands, for humans to be able to make the discoveries required to ensure our future we are going to need to be a lot more 'hopeful' than might've been the case two decades ago. An increasing sense of mistrust in "so-called experts" instigated and fuelled largely by grandstanding charlatans for their personal, often political gain, coupled with some high-profile cases of indefensible scientific misconduct, are directly influencing the funding of vital areas of research such as these, and may have sealed our fates as hopeless monsters at a time when it is critical to understand the nature of life, the universe,

and our place within them. The research presented within this thesis is novel, however the conclusions are not. As phrased best by H.G. Wells in 1945:

> "*A series of events has forced upon the intelligent observer the realisation that the human story has already come to an end and that Homo sapiens, as he has been pleased to call himself, is in his present form played out. The stars in their courses have turned against him and he has to give place to some other animal better adapted to face the fate that closes in more and more swiftly upon mankind.*
>
> *That new animal may be an entirely alien strain, or it may arise as a new modification of the Hominidae, and even as a direct continuation of the human phylum, but it will certainly not be human. There is no way out for Man but steeply up or steeply down. Adapt or perish, now as ever, is Nature's inexorable imperative*" (Wells, 1945)

REFERENCES

1. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526:68–74. doi:10.1038/nature15393
2. Abegglen LM, Caulin AF, Chan A, Lee K, Robinson R, Campbell MS, Kiso WK, Schmitt DL, Waddell PJ, Bhaskara S, Jensen ST, Maley CC, Schiffman JD. 2015. Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans. JAMA 314:1850. doi:10.1001/jama.2015.13134
3. Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat 22:229–244. doi:10.1002/humu.10254
4. Acharya D, Ghosh TC. 2016. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics 17:71. doi:10.1186/s12864-016-2392-0
5. Ackermann RR, Arnold ML, Baiz MD, Cahill JA, Cortés-Ortiz L, Evans BJ, Grant BR, Grant PR, Hallgrimsson B, Humphreys RA, Jolly CJ, Malukiewicz J, Percival CJ, Ritzman TB, Roos C, Roseman CC, Schroeder L, Smith FH, Warren KA, Wayne RK, Zinner D. 2019. Hybridization in human evolution: Insights from other organisms. Evol Anthropol Issues News Rev 28:189–209. doi:10.1002/evan.21787
6. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. doi:10.1038/nmeth0410-248
7. Ahmed H, Howton TC, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. 2018. Network biology discovers pathogen contact points in host

protein-protein interactomes. Nat Commun 9:2312. doi:10.1038/s41467-018-04632-8

8. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel J-H, White S, Zadissa A, Flicek P, Searle SMJ. 2016. The Ensembl gene annotation system. Database 2016:baw093. doi:10.1093/database/baw093
9. Allison AC. 1954. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. BMJ 1:290–294. doi:10.1136/bmj.1.4857.290
10. Alvarez-Ponce D, McInerney JO. 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. Genome Biol Evol 3:782–790. doi:10.1093/gbe/evr073
11. Anastasiou E, Mitchell PD. 2013. Palaeopathology and genes: Investigating the genetics of infectious diseases in excavated human skeletal remains and mummies from past populations. Gene 528:33–40. doi:10.1016/j.gene.2013.06.017
12. Andreae D, Nowak-Węgrzyn A. 2017. The Effect of Infant Allergen/Immunogen Exposure on Long-Term HealthEarly Nutrition and Long-Term Health. Elsevier. pp. 131–173. doi:10.1016/B978-0-08-100168-4.00006-9
13. Andreotti S, Holtzhausen P, Rutzen M, Meÿer M, van der Walt S, Herbst B, Matthee CA. 2018. Semi-automated software for dorsal fin photographic identification of marine species: application to Carcharodon carcharias. Mar Biodivers 48:1655–1660. doi:10.1007/s12526-017-0634-2
14. Andreotti S, Rutzen M, van der Walt S, Von der Heyden S, Henriques R, Meÿer M, Oosthuizen H, Matthee C. 2016. An integrated mark-recapture and genetic approach to estimate the population size of white sharks in South Africa. Mar Ecol Prog Ser 552:241–253. doi:10.3354/meps11744

15. Anthony RM, Rutitzky LI, Urban JF, Stadecker MJ, Gause WC. 2007. Protective immune mechanisms in helminth infection. Nat Rev Immunol 7:975–987. doi:10.1038/nri2199
16. Ari C. 2015. Long-term body pigmentation changes on a manta ray (Mobulidae). Biol J Linn Soc 114:406–414. doi:10.1111/bij.12416
17. Arkhipova IR, Yushenova IA. 2019. Giant Transposons in Eukaryotes: Is Bigger Better? Genome Biol Evol 11:906–918. doi:10.1093/gbe/evz041
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29. doi:10.1038/75556
19. Auguie B, Antonov A. 2017. gridExtra: Miscellaneous Functions for "Grid" Graphics.
20. Azad P, Stobdan T, Zhou D, Hartley I, Akbari A, Bafna V, Haddad GG. 2017. High-altitude adaptation in humans: from genomics to integrative physiology. J Mol Med 95:1269–1282. doi:10.1007/s00109-017-1584-7
21. Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet 14:113–124. doi:10.1038/nrg3366
22. Badyaev AV. 2019. Evolutionary transitions in controls reconcile adaptation with continuity of evolution. Semin Cell Dev Biol 88:36–45. doi:10.1016/j.semcdb.2018.05.014
23. Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552–564. doi:10.1038/nrg1895
24. Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 73:823–834. doi:10.1086/378594
25. Bailey JA. 2002. Recent Segmental Duplications in the Human Genome. Science 297:1003–1007. doi:10.1126/science.1072047
26. Baillargeon S, Rivest L-P. 2007. Rcapture: Loglinear Models for Capture-Recapture in R.

27. Baker RE, Peña J-M, Jayamohan J, Jérusalem A. 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol Lett 14:20170660. doi:10.1098/rsbl.2017.0660
28. Barabási A-L, Gulbahce N, Loscalzo J. 2011. Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56–68. doi:10.1038/nrg2918
29. Bateson W. 2010. Mendel's principles of heredity, Dover ed. ed, Dover books on biology, psychology and medicine. Mineola, N.Y: Dover Publications.
30. Beam AL, Kohane IS. 2018. Big Data and Machine Learning in Health Care. JAMA 319:1317. doi:10.1001/jama.2017.18391
31. Beddington JR, Agnew DJ, Clark CW. 2007. Current problems in the management of marine fisheries. Science 316:1713–6. doi:10.1126/science.1137362
32. Bell GI, Karam JH, Rutter WJ. 1981. Polymorphic DNA region adjacent to the 5' end of the human insulin gene. Proc Natl Acad Sci 78:5759–5763. doi:10.1073/pnas.78.9.5759
33. Bennett DJ, Sutton MD, Turvey ST. 2017. Evolutionarily distinct "living fossils" require both lower speciation and lower extinction rates. Paleobiology 43:34–48. doi:10.1017/pab.2016.36
34. Bennett N. 2019. Random_forest.ipynb, Glasgow_ML. Glasgow.
35. Bertrand D, Gagnon Y, Blanchette M, El-Mabrouk N. 2010. Reconstruction of Ancestral Genome Subject to Whole Genome Duplication, Speciation, Rearrangement and Loss In: Moulton V, Singh M, editors. Algorithms in Bioinformatics. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 78–89. doi:10.1007/978-3-642-15294-8_7
36. Bilgin Sonay T, Carvalho T, Robinson MD, Greminger MP, Krützen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques-Bonet T, Wagner A. 2015. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome Res 25:1591–1599. doi:10.1101/gr.190868.115
37. Birchler JA, Veitia RA. 2012. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci 109:14746–14753. doi:10.1073/pnas.1207726109

38. Birney E, Soranzo N. 2015. The end of the start for population sequencing. Nature 526:52–53. doi:10.1038/526052a
39. Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Newburger DE, West R, Sidow A, Batzoglou S. 2015. Read clouds uncover variation in complex regions of the human genome. Genome Res 25:1570–1580. doi:10.1101/gr.191189.115
40. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van Diemen FR, Olk N, Stukalov A, Marceau C, Janssen H, Carette JE, Bennett KL, Colinge J, Superti-Furga G, Brummelkamp TR. 2015. Gene essentiality and synthetic lethality in haploid human cells. Science 350:1092–1096. doi:10.1126/science.aac7557
41. Bloomfield PS, Selvaraj S, Veronese M, Rizzo G, Bertoldo A, Owen DR, Bloomfield MAP, Bonoldi I, Kalk N, Turkheimer F, McGuire P, de Paola V, Howes OD. 2016. Microglial Activity in People at Ultra High Risk of Psychosis and in Schizophrenia: An [11 C]PBR28 PET Brain Imaging Study. Am J Psychiatry 173:44–52. doi:10.1176/appi.ajp.2015.14101358
42. Boehm T, Iwanami N, Hess I. 2012. Evolution of the Immune System in the Lower Vertebrates. Annu Rev Genomics Hum Genet 13:127–149. doi:10.1146/annurev-genom-090711-163747
43. Bolger DT, Morrison TA, Vance B, Lee D, Farid H. 2012. A computer-assisted system for photographic mark-recapture analysis. Methods Ecol Evol 3:813–822. doi:10.1111/j.2041-210X.2012.00212.x
44. Bosch H. 1500. The Garden of earthly delights triptych.
45. Boyle EA, Pritchard JK, Greenleaf WJ. 2018. High-resolution mapping of cancer cell networks using co-functional interactions. Mol Syst Biol 14. doi:10.15252/msb.20188594
46. Braga J, Braak CJF, Thuiller W, Dray S. 2018. Integrating spatial and phylogenetic information in the fourth-corner analysis to test trait–environment relationships. Ecology 99:2667–2674. doi:10.1002/ecy.2530
47. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, Yates B, Bruford E. 2019. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res 47:D786–D792. doi:10.1093/nar/gky930
48. Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F. 2010. Gene duplications in prokaryotes can be associated with

environmental adaptation. BMC Genomics 11:588. doi:10.1186/1471-2164-11-588

49. Briggs N, Weatherhead J, Sastry KJ, Hotez PJ. 2016. The Hygiene Hypothesis and Its Inconvenient Truths about Helminth Infections. PLoS Negl Trop Dis 10:e0004944. doi:10.1371/journal.pntd.0004944
50. Brodie JF, Redford KH, Doak DF. 2018. Ecological Function Analysis: Incorporating Species Roles into Conservation. Trends Ecol Evol 33:840–850. doi:10.1016/j.tree.2018.08.013
51. Brown KV. 2018a. What Does an Infamous Biohacker's Death Mean for the Future of DIY Science? Science.
52. Brown KV. 2018b. It Wasn't Biohacking That Killed the Biohacker: He Drowned. Bloomsberg.
53. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. 2018. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell 173:53-61.e9. doi:10.1016/j.cell.2018.02.031
54. Brownscombe JW, Chapman JM, Gutowsky LFG. 2017. Best practices for catch-and-release recreational fisheries – angling tools and tactics. Fish Res 186:693–705. doi:10.1016/J.FISHRES.2016.04.018
55. Buckley RM, Kortschak RD, Adelson DL. 2018. Divergent genome evolution caused by regional variation in DNA gain and loss between human and mouse. PLOS Comput Biol 14:e1006091. doi:10.1371/journal.pcbi.1006091
56. Burnett AK, Hills RK, Milligan DW, Goldstone AH, Prentice AG, McMullin M-F, Duncombe A, Gibson B, Wheatley K. 2010. Attempts to Optimize Induction and Consolidation Treatment in Acute Myeloid Leukemia: Results of the MRC AML12 Trial. J Clin Oncol 28:586–595. doi:10.1200/JCO.2009.22.9088
57. Burt DW. 1992. Evolutionary grouping of the transforming growth factor-β superfamily. Biochem Biophys Res Commun 184:590–595. doi:10.1016/0006-291X(92)90630-4
58. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J.

2018. The UK Biobank resource with deep phenotyping and genomic data. Nature 562:203–209. doi:10.1038/s41586-018-0579-z

59. Bzdok D, Altman N, Krzywinski M. 2018. Statistics versus machine learning. Nat Methods 15:233–234. doi:10.1038/nmeth.4642
60. Calderone A, Licata L, Cesareni G. 2015. VirusMentha: a new resource for virus-host protein interactions. Nucleic Acids Res 43:D588–D592. doi:10.1093/nar/gku830
61. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. 2018. Next-Generation Machine Learning for Biological Networks. Cell 173:1581–1592. doi:10.1016/j.cell.2018.05.015
62. Castellano D, Uricchio LH, Munch K, Enard D. 2019. Viruses rule over adaptation in conserved human proteins (preprint). Evolutionary Biology. doi:10.1101/555060
63. Caulfield M, Davies J, Dennys M, Elbahy L, Fowler T, Hill S, Hubbard T, Jostins L, Maltby N, Mahon-Pearson J, McVean G, Nevin-Ridley K, Parker M, Parry V, Rendon A, Riley L, Turnbull C, Woods K. 2019. The National Genomics Research and Healthcare Knowledgebase. doi:10.6084/m9.figshare.4530893.v5
64. Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Deegan RM, King M-C. 1991. Call for a worldwide survey of human genetic diversity: A vanishing opportunity for the Human Genome Project. Genomics 11:490–491. doi:10.1016/0888-7543(91)90169-F
65. CEC. 2010. CEC (2010). Council Regulation (EU) No 23/2010 of 14 January 2010 fixing for 2010 the fishing opportunities for certain fish stocks and groups of fish stocks, applicable in EU waters and, for EU vessels, in waters where catch limitations are required and . Off J Eur Communities L21:1–20.
66. Centers for Disease Control and Prevention. 2011. The Burden of Soil-transmitted Helminths (STH). https://www.cdc.gov/globalhealth/ntd/diseases/sth_burden.html
67. Centers for Disease Control and Prevention. 2013. Parasites-Trichutiasis (also known as Whipworm Infection). https://www.cdc.gov/parasites/whipworm/index.html

68. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE. 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611. doi:10.1038/nature13907

69. Chandra S, Christensen J, Mamelund S-E, Paneth N. 2018. Short-Term Birth Sequelae of the 1918–1920 Influenza Pandemic in the United States: State-Level Analysis. Am J Epidemiol 187:2585–2595. doi:10.1093/aje/kwy153

70. Chapman DD, Simpfendorfer CA, Wiley TR, Poulakis GR, Curtis C, Tringali M, Carlson JK, Feldheim KA. 2011. Genetic Diversity Despite Population Collapse in a Critically Endangered Marine Fish: The Smalltooth Sawfish (Pristis pectinata). J Hered 102:643–652. doi:10.1093/jhered/esr098

71. Che D, Safran M, Peng Z. 2013. From Big Data to Big Data Mining: Challenges, Issues, and Opportunities In: Hong B, Meng X, Chen L, Winiwarter W, Song W, editors. Database Systems for Advanced Applications. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 1–15. doi:10.1007/978-3-642-40270-8_1

72. Chen X, Huang L, Xie D, Zhao Q. 2018a. EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. Cell Death Dis 9:3. doi:10.1038/s41419-017-0003-x

73. Chen X, Wang C-C, Yin J, You Z-H. 2018b. Novel Human miRNA-Disease Association Inference Based on Random Forest. Mol Ther - Nucleic Acids 13:568–579. doi:10.1016/j.omtn.2018.10.005

74. Ching-Tai Lin. 1974. Structural controllability. IEEE Trans Autom Control 19:201–208. doi:10.1109/TAC.1974.1100557

75. CoC. 2015. Compliance Committee: Working group on illegal, unreported and unregulated (IUU) fishing in the GFCM area. Marrakech, Morocco.

76. Coelho R, Bertozzi M, Ungaro. N, Ellis J. 2009. Raja undulata. IUCN Red List Threat Species 2009 ET161425A5420694. http://dx.doi.org/10.2305/IUCN.UK.2009-2.RLTS.T161425A5420694.en

77. Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet 9:938–950. doi:10.1038/nrg2482

78. Conley JM, Cook-Deegan R, Lázaro-Muñoz G. 2014. MYRIAD AFTER MYRIAD: THE PROPRIETARY DATA DILEMMA. N C J Law Technol 15:597–637.

79. Conway Morris S. 2000. The Cambrian “explosion”: Slow-fuse or megatonnage? Proc Natl Acad Sci 97:4426–4429. doi:10.1073/pnas.97.9.4426

80. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132:1077–1130. doi:10.1007/s00439-013-1331-2

81. Crow KD. 2006. What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? Mol Biol Evol 23:887–892. doi:10.1093/molbev/msj083

82. CSARDI G. 2006. The igraph software package for complex network research. Int J Complex Syst 1695.

83. Damgaard PB, Margaryan A, Schroeder H, Orlando L, Willerslev E, Allentoft ME. 2015. Improving access to endogenous DNA in ancient bones and teeth. Sci Rep 5:11184. doi:10.1038/srep11184

84. Daniłowicz-Luebert E, O’Regan NL, Steinfelder S, Hartmann S. 2011. Modulation of Specific and Allergy-Related Immune Responses by Helminths. J Biomed Biotechnol 2011:1–18. doi:10.1155/2011/821578

85. Dannemann M, Andrés AM, Kelso J. 2016. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. Am J Hum Genet 98:22–33. doi:10.1016/j.ajhg.2015.11.015

86. Darwin C. 1872. The descent of man and selection in relation to sex. New York: D. Appleton and Company.

87. Darwin, C. 1859. On the origin of species by means of natural selection, or the Preservation of Favoured Races in the Struggle for Life. Leipzig: London/Die Entstehung der Arten durch natürliche Zuchtwahl.

88. de Anda-Jáuregui G, Espinal-Enríquez J, Drago-García D, Hernández-Lemus E. 2018. Nonredundant, Highly Connected MicroRNAs Control Functionality in Breast Cancer Networks. Int J Genomics 2018:1–10. doi:10.1155/2018/9585383

89. Decaestecker E, Gaba S, Raeymaekers JAM, Stoks R, Van Kerckhoven L, Ebert D, De Meester L. 2007. Host–parasite 'Red Queen' dynamics archived in pond sediment. Nature 450:870–873. doi:10.1038/nature06291

90. Del Vecchio D, Qian Y, Murray RM, Sontag ED. 2018. Future systems and control research in synthetic biology. Annu Rev Control 45:5–17. doi:10.1016/j.arcontrol.2018.04.007

91. Deline B, Greenwood JM, Clark JW, Puttick MN, Peterson KJ, Donoghue PCJ. 2018. Evolution of metazoan morphological disparity. Proc Natl Acad Sci 115:E8909–E8918. doi:10.1073/pnas.1810575115

92. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova J-L, Patin E, Quintana-Murci L. 2016. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. Am J Hum Genet 98:5–21. doi:10.1016/j.ajhg.2015.11.014

93. Deshpande O, Batzoglou S, Feldman MW, Luca Cavalli-Sforza L. 2009. A serial founder effect model for human settlement out of Africa. Proc R Soc B Biol Sci 276:291–300. doi:10.1098/rspb.2008.0750

94. Dickerson JE, Robertson DL. 2012. On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication. Mol Biol Evol 29:61–69. doi:10.1093/molbev/msr111

95. Dickerson JE, Zhu A, Robertson DL, Hentges KE. 2011. Defining the Role of Essential Genes in Human Disease. PLOS ONE 6:e27368. doi:10.1371/journal.pone.0027368

96. Diss G, Gagnon-Arsenault I, Dion-Coté A-M, Vignaud H, Ascencio DI, Berger CM, Landry CR. 2017. Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. Science 355:630–634. doi:10.1126/science.aai7685

97. Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2014. NeEstimator v2: re-implementation of software for the estimation of

contemporary effective population size (N e) from genetic data. Mol Ecol Resour 14:209–214. doi:10.1111/1755-0998.12157

98. Dolgova O, Lao O. 2018. Evolutionary and Medical Consequences of Archaic Introgression into Modern Human Genomes. Genes 9:358. doi:10.3390/genes9070358

99. Domingues RR, Garrone-Neto D, Hilsdorf AWS, Gadig OBF. 2019. Use of mucus as a non-invasive sampling method for DNA barcoding of stingrays and skates (batoid elasmobranchs). J Fish Biol 94:512–516. doi:10.1111/jfb.13919

100. Donoghue PCJ, Keating JN. 2014. Early vertebrate evolution. Palaeontology 57:879–893. doi:10.1111/pala.12125

101. Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, Gallego CJ, Bennett RL, Jones KL, Tokita MJ, Bennett JT, Kim JH, Rosenthal EA, Kim DS, Tabor HK, Bamshad MJ, Motulsky AG, Scott CR, Pritchard CC, Walsh T, Burke W, Raskind WH, Byers P, Hisama FM, Nickerson DA, Jarvik GP. 2013. Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes. Am J Hum Genet 93:631–640. doi:10.1016/j.ajhg.2013.08.006

102. Duckert P, Brunak S, Blom N. 2004. Prediction of proprotein convertase cleavage sites. Protein Eng Des Sel 17:107–112. doi:10.1093/protein/gzh013

103. Dudley JT, Kim Y, Liu L, Markov GJ, Gerold K, Chen R, Butte AJ, Kumar S. 2012. Human genomic disease variants: a neutral evolutionary explanation. Genome Res 22:1383–1394. doi:10.1101/gr.133702.111

104. Dulvy NK, Fowler SL, Musick JA, Cavanagh RD, Kyne PM, Harrison LR, Carlson JK, Davidson LN, Fordham S V, Francis MP, Pollock CM, Simpfendorfer CA, Burgess GH, Carpenter KE, Compagno LJ, Ebert DA, Gibson C, Heupel MR, Livingstone SR, Sanciangco JC, Stevens JD, Valenti S, White WT. 2014. Extinction risk and conservation of the world's sharks and rays. eLife 3:e00590. doi:10.7554/eLife.00590

105. Duran J, Troncoso MF, Lagos D, Ramos S, Marin G, Estrada M. 2018. GDF11 Modulates Ca2+-Dependent Smad2/3 Signaling to Prevent Cardiomyocyte Hypertrophy. Int J Mol Sci 19. doi:10.3390/ijms19051508

106. Dutour O. 2013. Paleoparasitology and paleopathology. Synergies for reconstructing the past of human infectious diseases and their pathocenosis. Int J Paleopathol 3:145–149. doi:10.1016/j.ijpp.2013.09.008

107. Earl DA, VonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. doi:10.1007/s12686-011-9548-7

108. Ebert G, Steininger A, Weißmann R, Boldt V, Lind-Thomsen A, Grune J, Badelt S, Heßler M, Peiser M, Hitzler M, Jensen LR, Müller I, Hu H, Arndt PF, Kuss AW, Tebel K, Ullmann R. 2014. Distribution of segmental duplications in the context of higher order chromatin organisation of human chromosome 7. BMC Genomics 15:537. doi:10.1186/1471-2164-15-537

109. El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Günther S, Fukuda N, Kikhi K, Boezio GLM, Takacs CM, Lai S-L, Fukuda R, Gerri C, Giraldez AJ, Stainier DYR. 2019a. Genetic compensation triggered by mutant mRNA degradation. Nature. doi:10.1038/s41586-019-1064-z

110. Ellis JR, McCully Phillips SR, Poisson F. 2017. A review of capture and post-release mortality of elasmobranchs. J Fish Biol 90:653–722. doi:10.1111/jfb.13197

111. Ellis JR, McCully S, Wallis RHL. 2015. Raja undulata European Regional Assessment. IUCN Red List Threat Species. doi:e.T161425A48909382.

112. Ellis JR, McCully SR, Brown MJ. 2012. An overview of the biology and status of undulate ray Raja undulata in the north-east Atlantic Ocean. J Fish Biol 80:1057–1074. doi:10.1111/j.1095-8649.2011.03211.x

113. Else H. 2019. Genomics institute to close world-leading animal facility. Nature 569:612–612. doi:10.1038/d41586-019-01685-7

114. Emanuel BS, Shaikh TH. 2001. Segmental duplications: an “expanding” role in genomic instability and disease. Nat Rev Genet 2:791–800. doi:10.1038/35093500

115. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, Hoover H, Gudmundsdottir V, Horman SR, Aspelund T, Shu L, Trifonov V, Sigurdsson S, Manolescu A, Zhu J, Olafsson Ö, Jakobsdottir J, Lesley

SA, To J, Zhang J, Harris TB, Launer LJ, Zhang B, Eiriksdottir G, Yang X, Orth AP, Jennings LL, Gudnason V. 2018. Co-regulatory networks of human serum proteins link genetics to disease. Science 361:769–773. doi:10.1126/science.aaq1327

116. Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. eLife 5. doi:10.7554/eLife.12469

117. Enard D, Petrov DA. 2017. RNA viruses drove adaptive introgressions between Neanderthals and modern humans (preprint). Evolutionary Biology. doi:10.1101/120477

118. Enard D, Petrov DA. 2018a. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. Cell 175:360-371.e13. doi:10.1016/j.cell.2018.08.034

119. Esposito MR, Binatti A, Pantile M, Coppe A, Mazzocco K, Longo L, Capasso M, Lasorsa VA, Luksch R, Bortoluzzi S, Tonini GP. 2018. Somatic mutations in specific and connected subpathways are associated with short neuroblastoma patients' survival and indicate proteins targetable at onset of disease. Int J Cancer 143:2525–2536. doi:10.1002/ijc.31748

120. Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–20. doi:10.1111/j.1365-294X.2005.02553.x

121. Exome Aggregation Consortium, Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM,

Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291. doi:10.1038/nature19057

122. Feldheim KA, Gruber SH, Marignac JRC, Ashley M V. 2002. Genetic tagging to determine passive integrated transponder tag loss in lemon sharks. J Fish Biol 61:1309–1313. doi:10.1111/j.1095-8649.2002.tb02474.x

123. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37. doi:10.1093/nar/gkr367

124. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Vooren SV, Moreau Y, Pettett RM, Carter NP. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84:524–533. doi:10.1016/j.ajhg.2009.03.010

125. Fokkema IFAC, den Dunnen JT, Taschner PEM. 2005. LOVD: Easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26:63–68. doi:10.1002/humu.20201

126. Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151:1531–1545.

127. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, Karimpour-Fard A, Glueck D, McGavran L, Berry R, Pollack J, Sikela JM. 2004. Lineage-Specific Gene Duplication and Loss in Human and Great Ape Evolution. PLoS Biol 2:e207. doi:10.1371/journal.pbio.0020207

128. Fotiou E, Williams S, Martin-Geary A, Robertson DavidL, Tenin G, Hentges K, Keavney B. 2019. INTEGRATION OF LARGE-SCALE GENOMIC DATA SOURCES WITH EVOLUTIONARY INFORMATION REVEALS NOVEL GENETIC LOCI FOR CONGENITAL HEART DISEASE. FOTIOU ET AL 2019. Circ Precis Med.

129. Fox G, Darolti J, Hibbitt J, Preziosi RF, Fitzpatrick JL, Rowntree JK. 2018. Bespoke markers for ex-situ conservation : application , analysis and challenges in the assessment of a population of endangered undulate rays. Jzar 6:50–56.

130. Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251. doi:10.1038/nrg2554

131. Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLOS Genet 12:e1005774. doi:10.1371/journal.pgen.1005774

132. Gao C, Sun H, Wang T, Tang M, Bohnen NI, Müller MLTM, Herman T, Giladi N, Kalinin A, Spino C, Dauer W, Hausdorff JM, Dinov ID. 2018. Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. Sci Rep 8:7129. doi:10.1038/s41598-018-24783-4

133. Gent E. 2019. California Passed the Country's First Law to Prevent Genetic Biohacking. Singul Hub.

134. Gibbons A. 2012. Bonobos Join Chimps as Closest Human Relatives. Science.

135. Gilbert C, Feschotte C. 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr Opin Genet Dev 49:15–24. doi:10.1016/j.gde.2018.02.007

136. Gnecchi-Ruscone GA, Abondio P, De Fanti S, Sarno S, Sherpa MG, Sherpa PT, Marinelli G, Natali L, Di Marcello M, Peluzzi D, Luiselli D, Pettener D, Sazzini M. 2018. Evidence of polygenic adaptation to high altitude from Tibetan and Sherpa genomes. Genome Biol Evol. doi:10.1093/gbe/evy233

137. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L. 2007. The human disease network. Proc Natl Acad Sci 104:8685–8690. doi:10.1073/pnas.0701361104

138. Goldschmidt R. 1982. The material basis of evolution, Silliman milestones in science. New Haven: Yale University Press.

139. Goldschmidt, R. 1933. Some aspects of evolution. Science, New Series 78:539–547.

140. Goldschmidt, R. 1940. The material basis of evolution. New Haven : Tale University Press; London : Humphrey Milford Oxford University Press.
141. González-Ramos MS, Santos-Moreno A, Rosas-Alquicira EF, Fuentes-Mascorro G. 2017. Validation of photo-identification as a mark-recapture method in the spotted eagle ray Aetobatus narinari. J Fish Biol 90:1021–1030. doi:10.1111/jfb.13215
142. Gosak M, Markovič R, Dolenšek J, Slak Rupnik M, Marhl M, Stožer A, Perc M. 2018. Network science of biological systems at different scales: A review. Phys Life Rev 24:118–135. doi:10.1016/j.plrev.2017.11.003
143. Gotelli NJ, Hart EM, Ellison AM. 2015. EcoSimR: Null model analysis for ecological data. doi:10.5281/zenodo.16522
144. Gouy M, Guindon S, Gascuel O. 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol 27:221–224. doi:10.1093/molbev/msp259
145. Graves JAM. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. Nat Rev Genet 17:33–46. doi:10.1038/nrg.2015.2
146. Green ME, Appleyard SA, White W, Tracey S, Ovenden J. 2017. Variability in multiple paternity rates for grey reef sharks (Carcharhinus amblyrhynchos) and scalloped hammerheads (Sphyrna lewini). Sci Rep 7:1528. doi:10.1038/s41598-017-01416-w
147. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S. 2010. A Draft Sequence of the Neandertal Genome. Science 328:710–722. doi:10.1126/science.1188021
148. Greenbaum G, Getz WM, Rosenberg NA, Feldman MW, Hovers E, Kolodny O. 2018. Disease and introgression explain the long-lasting

contact zone of Modern Humans and Neanderthals and its eventual destabilization (preprint). Evolutionary Biology. doi:10.1101/495515

149. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66. doi:10.1038/nature01198
150. Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional Analysis of Gene Duplications in Saccharomyces cerevisiae. Genetics 175:933–943. doi:10.1534/genetics.106.064329
151. Guirimand T, Delmotte S, Navratil V. 2015. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res 43:D583–D587. doi:10.1093/nar/gku1121
152. Guttridge TL, Gruber SH, Krause J, Sims DW. 2010. Novel Acoustic Technology for Studying Free-Ranging Shark Social Behaviour by Recording Individuals' Interactions. PLoS ONE 5:e9324. doi:10.1371/journal.pone.0009324
153. Hahn MW, Demuth JP, Han S-G. 2007. Accelerated Rate of Gene Gain and Loss in Primates. Genetics 177:1941–1949. doi:10.1534/genetics.107.080077
154. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol 8:R209. doi:10.1186/gb-2007-8-10-r209
155. Haq AU, Li JP, Memon MH, Nazir S, Sun R. 2018. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Mob Inf Syst 2018:1–21. doi:10.1155/2018/3860146
156. Harrison CA, Al-Musawi SL, Walton KL. 2011. Prodomains regulate the synthesis, extracellular localisation and activity of TGF-β superfamily ligands. Growth Factors 29:174–186. doi:10.3109/08977194.2011.608666
157. Hayes KS, Bancroft AJ, Goldrick M, Portsmouth C, Roberts IS, Grencis RK. 2010. Exploitation of the Intestinal Microflora by the Parasitic Nematode Trichuris muris. Science 328:1391–1394. doi:10.1126/science.1187703

158. He T, Zhu M, Mills BJW, Wynn PM, Zhuravlev AYu, Tostevin R, Pogge von Strandmann PAE, Yang A, Poulton SW, Shields GA. 2019. Possible links between extreme oxygen perturbations and the Cambrian radiation of animals. Nat Geosci 12:468–474. doi:10.1038/s41561-019-0357-z

159. He X, Zhang J. 2006. Higher Duplicability of Less Important Genes in Yeast Genomes. Mol Biol Evol 23:144–151. doi:10.1093/molbev/msj015

160. Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. 2015. Estimating the mutation load in human genomes. Nat Rev Genet 16:333–343. doi:10.1038/nrg3931

161. Henn BM, Cavalli-Sforza LL, Feldman MW. 2019. The great human expansion. Resonance 24:711–718. doi:10.1007/s12045-019-0830-4

162. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, Spooner W, Kulesha E, Yates A, Flicek P. 2016. Ensembl comparative genomics resources. Database 2016. doi:10.1093/database/bav096

163. Hijmans RJ, Etten J van, Cheng J, Mattiuzzi M, Sumner M, Greenberg JA, Lamigueiro OP, Bevan A, Racine EB, Shortridge A, Ghosh A. 2017. raster: Geographic Data Analysis and Modeling.

164. Hinck AP, Mueller TD, Springer TA. 2016. Structural Biology and Evolution of the TGF-β Family. Cold Spring Harb Perspect Biol 8. doi:10.1101/cshperspect.a022103

165. Hinck AP. 2012. Structural studies of the TGF-βs and their receptors - insights into evolution of the TGF-β superfamily. FEBS Lett 586:1860–1870. doi:10.1016/j.febslet.2012.05.028

166. Hinck AP. 2018. Structure-guided engineering of TGF-βs for the development of novel inhibitors and probing mechanism. Bioorg Med Chem 26:5239–5246. doi:10.1016/j.bmc.2018.07.008

167. Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, Thomas JA, Ludwig A, Collins MJ. 2015. The future of ancient DNA: Technical advances and conceptual shifts: Prospects & Overviews. BioEssays 37:284–293. doi:10.1002/bies.201400160

168. Hofreiter M. 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res 29:4793–4799. doi:10.1093/nar/29.23.4793

169. Hofreiter M. 2011. Drafting Human Ancestry: What Does the Neanderthal Genome Tell Us about Hominid Evolution? Commentary on Green et al. (2010). Hum Biol 83:1–11. doi:10.3378/027.083.0101

170. Holland LZ, Ocampo Daza D. 2018. A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution? Genome Biol 19:209. doi:10.1186/s13059-018-1592-0

171. Hollstein M, Sidransky D, Vogelstein B, Harris C. 1991. p53 mutations in human cancers. Science 253:49–53. doi:10.1126/science.1905840

172. Horscroft JA, Kotwica AO, Laner V, West JA, Hennis PJ, Levett DZH, Howard DJ, Fernandez BO, Burgess SL, Ament Z, Gilbert-Kawai ET, Vercueil A, Landis BD, Mitchell K, Mythen MG, Branco C, Johnson RS, Feelisch M, Montgomery HE, Griffin JL, Grocott MPW, Gnaiger E, Martin DS, Murray AJ. 2017. Metabolic basis to Sherpa altitude adaptation. Proc Natl Acad Sci 114:6382–6387. doi:10.1073/pnas.1700527114

173. Horvath JE. 2000. The Mosaic Structure of Human Pericentromeric DNA: A Strategy for Characterizing Complex Regions of the Human Genome. Genome Res 10:839–852. doi:10.1101/gr.10.6.839

174. Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. 2017. WormBase ParaSite − a comprehensive resource for helminth genomics. Mol Biochem Parasitol 215:2–10. doi:10.1016/j.molbiopara.2016.11.005

175. Howlett, L. Palmer, K. Flint, K. Thornhill, L. Sharky. 1994. Poison, Music for the Jilted Generation. Essex: XL/Mute.

176. Hsiao T-L, Vitkup D. 2008. Role of Duplicate Genes in Robustness against Deleterious Human Mutations. PLOS Genet 4:e1000014. doi:10.1371/journal.pgen.1000014

177. Huang N, Lee I, Marcotte EM, Hurles ME. 2010. Characterising and Predicting Haploinsufficiency in the Human Genome. PLOS Genet 6:e1001154. doi:10.1371/journal.pgen.1001154

178. Huang X, Liu H, Li X, Guan L, Li J, Tellier LCAM, Yang H, Wang J, Zhang J. 2018. Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning. BMC Neurol 18:5. doi:10.1186/s12883-017-1010-3

179. Hughes P, Hassan I, Que L, Mead P, Lee JH, Love DR, Prosser DO, Cundy T. 2019. Observations on the Natural History of Camurati-Engelmann Disease: OBSERVATIONS ON THE NATURAL HISTORY OF CAMURATI-ENGELMANN DISEASE. J Bone Miner Res e3670. doi:10.1002/jbmr.3670
180. Humphries N, Simpson S, Sims D. 2017. Diel vertical migration and central place foraging in benthic predators. Mar Ecol Prog Ser 582:163–180. doi:10.3354/meps12324
181. Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, Cunningham F. 2018. Ensembl variation resources. Database 2018. doi:10.1093/database/bay119
182. Hunter E, Buckley A a., Stewart C, Metcalfe J d. 2005. Migratory behaviour of the thornback ray, raja clavata , in the southern north sea. J Mar Biol Assoc U K 85:1095–1105. doi:10.1017/S0025315405012142
183. Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. Comput Sci Eng 9:90–95. doi:10.1109/MCSE.2007.55
184. Hunter K. 2016. The Development of Molecular Techniques for the Conservation of Captive Elasmobranchs 0–59.
185. Hurles M. 2004. Gene Duplication: The Genomic Trade in Spare Parts. PLoS Biol 2:e206. doi:10.1371/journal.pbio.0020206
186. Husson F, Josse J, Le S, Mazet J. 2018. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining.
187. Huxley J. 1942. Evolution:The Modern Synthesis. London: Allen & Unwin Ltd.
188. Huxley TH. 1887. On the reception of the origin of species.
189. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36:949–951. doi:10.1038/ng1416
190. ICES. 2016. Advice on fishing opportunities, catch, and effort Celtic Seas Ecoregion.
191. Ingman WV, Robertson SA. 2009. The essential roles of TGFB1 in reproduction. Cytokine Growth Factor Rev 20:233–239. doi:10.1016/j.cytogfr.2009.05.003

192. Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108. doi:10.1038/nrg2689

193. ISOLATE II Genomic DNA Kit Product Manual. n.d.

194. IUCN. 2014. A quarter of sharks and rays threatened with extinction | IUCN. Int Union Conserv Nat. https://www.iucn.org/content/quarter-sharks-and-rays-threatened-extinction

195. Jackson JA, Friberg IM, Little S, Bradley JE. 2009. Review series on helminths, immune modulation and the hygiene hypothesis: immunity against helminths and immunological phenomena in modern human populations: coevolutionary legacies? Immunology 126:18–27. doi:10.1111/j.1365-2567.2008.03010.x

196. Jackson T E. 1997. Charles and the hopeful monster: postmodern evolutionary theory in “The French Lieutenant’s Woman.” (protagonist in book by author John Fowles). Twent Century Lit 43:221+.

197. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O’Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. doi:10.1038/nbt.4060

198. Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806. doi:10.1093/bioinformatics/btm233

199. Jin X, Wah BW, Cheng X, Wang Y. 2015. Significance and Challenges of Big Data Research. Big Data Res 2:59–64. doi:10.1016/j.bdr.2015.01.006

200. Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. Nat Rev Microbiol. doi:10.1038/s41579-019-0189-2

201. Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol 4:22. doi:10.1186/1471-2148-4-22

202. Kabir M, Wenlock S, Doig AJ, Hentges KE. 2019. The Essentiality Status of Mouse Duplicate Gene Pairs Correlates with Developmental Co-Expression Patterns. Sci Rep 9:3224. doi:10.1038/s41598-019-39894-9

203. Kaiser J. 2016. Funding for key data resources in jeopardy. Science 351:14–14. doi:10.1126/science.351.6268.14

204. Kaisler S, Armour F, Espinosa JA, Money W. 2013. Big Data: Issues and Challenges Moving Forward2013 46th Hawaii International Conference on System Sciences. Presented at the 2013 46th Hawaii International Conference on System Sciences (HICSS). Wailea, HI, USA: IEEE. pp. 995–1004. doi:10.1109/HICSS.2013.645

205. Kannan R, Vasanthi V. 2019. Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart DiseaseSoft Computing and Medical Bioinformatics. Singapore: Springer Singapore. pp. 63–72. doi:10.1007/978-981-13-0059-2_8

206. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, The Genome Aggregation Database Consortium, Neale BM, Daly MJ, MacArthur DG. 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes (preprint). Genomics. doi:10.1101/531210

207. Kashiwagi T, Maxwell EA, Marshall AD, Christensen AB. 2015. Evaluating manta ray mucus as an alternative DNA source for population genetics study: underwater-sampling, dry-storage and PCR success. PeerJ 3:e1188. doi:10.7717/peerj.1188

208. Kassambara A, Mundt F. 2017. factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

209. Keating JN, Marquart CL, Marone F, Donoghue PCJ. 2018. The nature of aspidin and the evolutionary origin of bone. Nat Ecol Evol 2:1501–1506. doi:10.1038/s41559-018-0624-1

210. Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617–624. doi:10.1038/nature02424

211. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong Justin, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong Jason, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 172:1122-1131.e9. doi:10.1016/j.cell.2018.02.010

212. Kimura M. 1983. The neutral theory of molecular evolution. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

213. Koch L. 2015. A new foundation for non-coding variant analysis. Nat Rev Genet 16:688–688. doi:10.1038/nrg4047

214. Kohler NE, Turner PA. 2001. Shark tagging: a review of conventional methods and studies, Environmental Biology of Fishes.

215. Kondrashov FA, Koonin EV. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. Trends Genet TIG 20:287–290. doi:10.1016/j.tig.2004.05.001

216. Koonin EV. 2009. The Origin at 150: is a new evolutionary synthesis in sight? Trends Genet 25:473–475. doi:10.1016/j.tig.2009.09.007

217. Kotsiantis SB. 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica 249–268.

218. Kremling A. 2013. Systems siology: mathematical modeling and model analysis.

219. Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S. 1997. Neandertal DNA Sequences and the Origin of Modern Humans. Cell 90:19–30. doi:10.1016/S0092-8674(00)80310-4

220. Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H. 2002. Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. Immunol Rev 190:95–122. doi:10.1034/j.1600-065X.2002.19008.x

221. Kuzmin E, VanderSluis B, Wang W, Tan G, Deshpande R, Chen Y, Usaj M, Balint A, Mattiazzi Usaj M, van Leeuwen J, Koch EN, Pons C, Dagilis AJ, Pryszlak M, Wang JZY, Hanchard J, Riggi M, Xu K, Heydari H, San Luis B-J, Shuteriqi E, Zhu H, Van Dyk N, Sharifpoor S, Costanzo M, Loewith R, Caudy A, Bolnick D, Brown GW, Andrews BJ, Boone C, Myers CL. 2018. Systematic analysis of complex genetic interactions. Science 360:eaao1729. doi:10.1126/science.aao1729

222. Laake JL, Johnson DS, Conn PB. 2013. marked: an R package for maximum likelihood and Markov Chain Monte Carlo analysis of capture-recapture data. Methods Ecol Evol 4:885–890. doi:10.1111/2041-210X.12065

223. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB. 2012. The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. PLoS Genet 8:e1002496. doi:10.1371/journal.pgen.1002496

224. Lagatie O, Van Dorst B, Stuyver LJ. 2017. Identification of three immunodominant motifs with atypical isotype profile scattered over the Onchocerca volvulus proteome. PLoS Negl Trop Dis 11:e0005330. doi:10.1371/journal.pntd.0005330

225. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. 2018. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46:D1062–D1067. doi:10.1093/nar/gkx1153

226. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R,

Vaughan B, Laurent T, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres AC, de Argila JR, Llobet OM, Medina I, Puy MS, Alberich M, de la Torre S, Navarro A, Paschall J, Flicek P. 2015. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet 47:692–695. doi:10.1038/ng.3312

227. Larson SE, Daly-Engel TS, Phillips NM. 2017. Review of Current Conservation Genetic Analyses of Northeast Pacific Sharks. Adv Mar Biol 77:79–110. doi:10.1016/BS.AMB.2017.06.005

228. Le Page M. 2019. Centre for Cancer Drug Discovery to focus on anti-evolution treatments. New Sci.

229. Lee HJ, Georgiadou A, Walther M, Nwakanma D, Stewart LB, Levin M, Otto TD, Conway DJ, Coin LJ, Cunnington AJ. 2018. Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. Sci Transl Med 10:eaar3619. doi:10.1126/scitranslmed.aar3619

230. Lek M, Quinlan KGR, North KN. n.d. The evolution of skeletal muscle performance: gene duplication and divergence of human sarcomeric α-actinins. BioEssays 32:17–25. doi:10.1002/bies.200900110

231. Leonardi M, Librado P, Der Sarkissian C, Schubert M, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, Gamba C, Willerslev E, Orlando L. 2016. Evolutionary Patterns and Processes: Lessons from Ancient DNA. Syst Biol syw059. doi:10.1093/sysbio/syw059

232. Levison SE, McLaughlin JT, Zeef L, Pennock JL. 2010. OC-010 Phenotyping Trichuris colitis reveals innate and adaptive immunological commonalities to human Crohn's disease. Gut 59:A4.1-A5. doi:10.1136/gut.2009.208934j

233. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, Grarup N, Guo Y, Hellman I, Jin X, Li Q, Liu J, Liu X, Sparsø T, Tang M, Wu H, Wu R, Yu C, Zheng H, Astrup A, Bolund L, Holmkvist J, Jørgensen T, Kristiansen K, Schmitz O, Schwartz TW, Zhang X, Li R, Yang H, Wang Jian, Hansen T, Pedersen O, Nielsen R, Wang Jun. 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet 42:969–972. doi:10.1038/ng.680

234. Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. Nat Rev Genet 16:321–332. doi:10.1038/nrg3920

235. Lieben L. 2016. SNP location helps predict disease aetiology. Nat Rev Genet 17:4–5. doi:10.1038/nrg.2015.12

236. Lieber L, Berrow S, Johnston E, Hall G, Hall J, Gubili C, Sims D, Jones C, Noble L. 2013. Mucus: aiding elasmobranch conservation through non-invasive genetic sampling. Endanger Species Res 21:215–222. doi:10.3354/esr00524

237. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437:94–100. doi:10.1038/nature04029

238. Liu DJ, Leal SM. 2010. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. PLoS Genet 6:e1001156. doi:10.1371/journal.pgen.1001156

239. Lombardi A, Hörnquist M. 2007. Controllability analysis of networks. Phys Rev E 75:056110. doi:10.1103/PhysRevE.75.056110

240. Lopez-Bigas N. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32:3108–3114. doi:10.1093/nar/gkh605

241. López-Bigas N, Ouzounis CA. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32:3108–3114. doi:10.1093/nar/gkh605

242. Lucotte EA, Skov L, Jensen JM, Coll Macià M, Munch K, Schierup MH. 2018. Dynamic Copy Number Evolution of X- and Y-Linked Ampliconic Genes in Human Populations. Genetics genetics.300826.2018. doi:10.1534/genetics.118.300826

243. Lussenhop J. 2017. Why I injected myself with an untested gene therapy. BBC News.

244. Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404. doi:10.1126/science.1089370

245. Lynch M, Hill WG. 1986. PHENOTYPIC EVOLUTION BY NEUTRAL MUTATION. Evolution 40:915–935. doi:10.1111/j.1558-5646.1986.tb00561.x

246. Ma Z, Zhu P, Shi H, Guo L, Zhang Q, Chen Y, Chen S, Zhang Z, Peng J, Chen J. 2019. PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature. doi:10.1038/s41586-019-1057-y

247. Machado LR, Hardwick RJ, Bowdrey J, Bogle H, Knowles TJ, Sironi M, Hollox EJ. 2012. Evolutionary history of copy-number-variable locus for the low-affinity Fcγ receptor: mutation rate, autoimmune disease, and the legacy of helminth infection. Am J Hum Genet 90:973–985. doi:10.1016/j.ajhg.2012.04.018

248. Maechler M, original) PR (Fortran, original) AS (S, original) MH (S, maintenance(1999-2000)) KH (port to R, Studer M, Roudier P, Gonzalez J, Kozlowski K. 2018. cluster: “Finding Groups in Data”: Cluster Analysis Extended Rousseeuw et al.

249. Maizels RM, McSorley HJ. 2016. Regulation of the host immune system by helminth parasites. J Allergy Clin Immunol 138:666–675. doi:10.1016/j.jaci.2016.07.007

250. Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. Trends Genet 25:152–155. doi:10.1016/j.tig.2009.03.001

251. Makino T, McLysaght A, Kawata M. 2013. Genome-wide deserts for copy number variation in vertebrates. Nat Commun 4:2283. doi:10.1038/ncomms3283

252. Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci 107:9270–9274. doi:10.1073/pnas.0914697107

253. Mangel M, F Samaniego. 1984. Abraham Wald’s work on aircraft survivability. J Am Stat Assoc.

254. Maranda V, Sunstrum FG, Drouin G. 2019. Both male and female gamete generating cells produce processed pseudogenes in the human genome. Gene 684:70–75. doi:10.1016/j.gene.2018.10.061

255. Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417. doi:10.1038/ng1537

256. Marciniak S, Perry GH. 2017. Harnessing ancient genomes to study the history of human adaptation. Nat Rev Genet 18:659–674. doi:10.1038/nrg.2017.65

257. Marshall AD, Pierce SJ. 2012. The use and abuse of photographic identification in sharks and rays. J Fish Biol 80:1361–1379. doi:10.1111/j.1095-8649.2012.03244.x

258. Marshall TC, Slate J, Kruuk LEB, Pemberton JM. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol 7:639–655. doi:10.1046/j.1365-294x.1998.00374.x

259. Martin-Geary A, Reardon M, Keith B, Tassabehji M, Robertson DL. 2019. Human genetic disease is greatly influenced by the underlying fragility of evolutionarily ancient genes. bioRxiv. doi:10.1101/558916

260. Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, Stamatoyannopoulos JA. 2015. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat Genet 47:1393–1401. doi:10.1038/ng.3432

261. McCrea RS, Morgan BJT. 2014. Analysis of capture-recapture data.

262. McKie R. 2019. Scientists split as genetics lab scales down animal tests. The Guardian.

263. McKinney W, others. 2010. Data structures for statistical computing in pythonProceedings of the 9th Python in Science Conference. Austin, TX. pp. 51–56.

264. McKusick-Nathans Institute of Genetic Medicine M Johns Hopkins University (Baltimore. n.d. OMIM - Online Mendelian Inheritance in Man.

265. McKusick-Nathans Institute of Genetic Medicine. 2018. OMIM Genemap2. Online Mendel Inherit Man OMIM. https://omim.org/

266. Mennecart B, Geraads D, Spassov N, Zagorchev I. 2018. Discovery of the oldest European ruminant in the late Eocene of Bulgaria: Did tectonics influence the diachronic development of the Grande Coupure?

Palaeogeogr Palaeoclimatol Palaeoecol 498:1–8. doi:10.1016/j.palaeo.2018.01.011

267. Milenković T, Memišević V, Bonato A, Pržulj N. 2011. Dominating Biological Networks. PLoS ONE 6:e23016. doi:10.1371/journal.pone.0023016
268. Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328. doi:10.1093/hmg/10.21.2319
269. Miller MP. 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328. doi:10.1093/hmg/10.21.2319
270. Monaco A, Monda A, Amoroso N, Bertolino A, Blasi G, Di Carlo P, Papalino M, Pergola G, Tangaro S, Bellotti R. 2018. A complex network approach reveals a pivotal substructure of genes linked to schizophrenia. PLOS ONE 13:e0190110. doi:10.1371/journal.pone.0190110
271. Moore AL, Marshall CD, Barnes LA, Murphy MP, Ransom RC, Longaker MT. 2018. Scarless wound healing: Transitioning from fetal research to regenerative healing. Wiley Interdiscip Rev Dev Biol 7:e309. doi:10.1002/wdev.309
272. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, Taga M, Klein H-U, Patrick E, Komashko V, McCabe C, Smith R, Bradshaw EM, Root DE, Regev A, Yu L, Chibnik LB, Schneider JA, Young-Pearse TL, Bennett DA, De Jager PL. 2018. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat Neurosci 21:811–819. doi:10.1038/s41593-018-0154-9
273. Murahwa AT, Nindo F, Onywera H, Meiring TL, Martin DP, Williamson A-L. 2019. Evolutionary dynamics of ten novel Gamma-PVs: insights from phylogenetic incongruence, recombination and phylodynamic analyses. BMC Genomics 20:368. doi:10.1186/s12864-019-5735-9
274. Murrell P. 2018. R Graphics Second Edition, The R series. Boca Raton: CRC Press.
275. Muscente AD, Prabhu A, Zhong H, Eleish A, Meyer MB, Fox P, Hazen RM, Knoll AH. 2018. Quantifying ecological impacts of mass extinctions

with network analysis of fossil communities. Proc Natl Acad Sci 115:5217–5222. doi:10.1073/pnas.1719976115

276. Nacher J, Akutsu T. 2012. Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. New J Phys.

277. Nacher JC, Akutsu T. 2013. Analysis on critical nodes in controlling complex networks using dominating sets2013 International Conference on Signal-Image Technology & Internet-Based Systems. Presented at the 2013 International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). Kyoto, Japan: IEEE. pp. 649–654. doi:10.1109/SITIS.2013.106

278. Nacher JC, Akutsu T. 2014. Analysis of critical and redundant nodes in controlling directed and undirected complex networks using dominating sets. J Complex Netw 2:394–412. doi:10.1093/comnet/cnu029

279. Nacher JC, Akutsu T. 2016. Minimum dominating set-based methods for analyzing biological networks. Methods 102:57–63. doi:10.1016/j.ymeth.2015.12.017

280. Nance HA, Daly-Engel TS, Marko PB. 2009. New microsatellite loci for the endangered scalloped hammerhead shark, Sphyrna lewini. Mol Ecol Resour 9:955–957. doi:10.1111/j.1755-0998.2008.02510.x

281. Naqvi S, Bellott DW, Lin KS, Page DC. 2018. Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. Genome Res 28:474–483. doi:10.1101/gr.230433.117

282. Naville M, Henriet S, Warren I, Sumic S, Reeve M, Volff J-N, Chourrout D. 2019. Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. Curr Biol 29:1161-1168.e6. doi:10.1016/j.cub.2019.01.080

283. Navlakha S, Kingsford C. 2010. The power of protein interaction networks for associating genes with diseases. Bioinformatics 26:1057–1063. doi:10.1093/bioinformatics/btq076

284. Navratil V, de Chassey B, Combe C, Lotteau V. 2011. When the human viral infectome and diseasome networks collide: towards a systems

biology platform for the aetiology of human diseases. BMC Syst Biol 5:13. doi:10.1186/1752-0509-5-13

285. NCBI Resource Coordinators, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, Cavanaugh M, Charowhas C, Clark K, Dondoshansky I, Feolo M, Fitzpatrick L, Funk K, Geer LY, Gorelenkov V, Graeff A, Hlavina W, Holmes B, Johnson M, Kattman B, Khotomlianski V, Kimchi A, Kimelman M, Kimura M, Kitts P, Klimke W, Kotliarov A, Krasnov S, Kuznetsov A, Landrum MJ, Landsman D, Lathrop S, Lee JM, Leubsdorf C, Lu Z, Madden TL, Marchler-Bauer A, Malheiro A, Meric P, Karsch-Mizrachi I, Mnev A, Murphy T, Orris R, Ostell J, O'Sullivan C, Palanigobu V, Panchenko AR, Phan L, Pierov B, Pruitt KD, Rodarmer K, Sayers EW, Schneider V, Schoch CL, Schuler GD, Sherry ST, Siyan K, Soboleva A, Soussov V, Starchenko G, Tatusova TA, Thibaud-Nissen F, Todorov K, Trawick BW, Vakatov D, Ward M, Yaschenko E, Zasypkin A, Zbicz K. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46:D8–D13. doi:10.1093/nar/gkx1095

286. Newman D, Whelan F, Moore M, Rusilowicz M, McInerney JO. 2019. Reconstructing and Analysing The Genome of The Last Eukaryote Common Ancestor to Better Understand the Transition from FECA to LECA. bioRxiv. doi:10.1101/538264

287. Newman SA. 2002. Developmental mechanisms: putting genes in their place. J Biosci 27:97–104. doi:10.1007/BF02703765

288. Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L, Farahmand M. 2018. A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybern Biomed Eng 38:1–15. doi:10.1016/j.bbe.2017.09.002

289. Nomura T. 2008. Estimation of effective number of breeders from molecular coancestry of single cohort sample. Evol Appl 1:462–474. doi:10.1111/j.1752-4571.2008.00015.x

290. Ohno, Susumu. 1970. Evolution By Gene Duplication. Springer.

291. Ohta T, Gillespie JH. 1996. Development of Neutral and Nearly Neutral Theories. Theor Popul Biol 49:128–142. doi:10.1006/tpbi.1996.0007

292. Ohta T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. Nature 246:96–98. doi:10.1038/246096a0

293. Ohtani H, Liu M, Zhou W, Liang G, Jones PA. 2018. Switching roles for DNA and histone methylation depend on evolutionary ages of human endogenous retroviruses. Genome Res 28:1147–1157. doi:10.1101/gr.234229.118

294. Olabode AS, Gatherer D, Jiang X, Matthews D, Hiscox JA, Gunther S, Carroll MW, Lovell SC, Robertson DL. 2016. Identification of important amino acid replacements in the 2013-2016 Ebola virus outbreak (preprint). Microbiology. doi:10.1101/075168

295. Openshaw M, Openshaw S. 2018. The Undulate Ray Project.org. http://undulateray.uk/

296. Pachter, Lior. 2015. I was wrong. Bits DNA.

297. Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197. doi:10.1038/nature01771

298. Park S. 2001. Trypanotolerance in West African cattle and the population genetic effects of selection. Ph.D. Thesis. University of Dublin.

299. Paul D, Su R, Romain M, Sébastien V, Pierre V, Isabelle G. 2017. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Comput Med Imaging Graph 60:42–49. doi:10.1016/j.compmedimag.2016.12.002

300. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M. 2011. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2825–2830.

301. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MH-Y, Cao H, Cohain A, Deikus G, Durrett RE, Blanchard SC, Altman R, Chin C-S, Guo Y, Paxinos EE, Korbel JO, Darnell RB, McCombie WR, Kwok P-Y, Mason CE, Schadt EE, Bashir A. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12:780–786. doi:10.1038/nmeth.3454

302. Peng G, Fan Y, Palculict TB, Shen P, Ruteshouser EC, Chi A-K, Davis RW, Huff V, Scharfe C, Wang W. 2013. Rare variant detection using

family-based sequencing analysis. Proc Natl Acad Sci 110:3985–3990. doi:10.1073/pnas.1222158110

303. Pennok JI, Ogunkanbi A. 2019. Unpublished results.

304. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet 9:e1003709. doi:10.1371/journal.pgen.1003709

305. Peyraud R, Cottret L, Marmiesse L, Genin S. 2018. Control of primary metabolism by a virulence regulatory network promotes robustness in a plant pathogen. Nat Commun 9:418. doi:10.1038/s41467-017-02660-4

306. Pires JC, Conant GC. 2016. Robust Yet Fragile: Expression Noise, Protein Misfolding, and Gene Dosage in the Evolution of Genomes. Annu Rev Genet 50:113–131. doi:10.1146/annurev-genet-120215-035400

307. Pirruccello-Straub M, Jackson J, Wawersik S, Webster MT, Salta L, Long K, McConaughy W, Capili A, Boston C, Carven GJ, Mahanthappa NK, Turner KJ, Donovan A. 2018. Blocking extracellular activation of myostatin as a strategy for treating muscle wasting. Sci Rep 8:2292. doi:10.1038/s41598-018-20524-9

308. Polvi A, Linturi H, Varilo T, Anttonen A-K, Byrne M, Fokkema IFAC, Almusa H, Metzidis A, Avela K, Aula P, Kestilä M, Muilu J. 2013. The Finnish Disease Heritage Database (FinDis) Update-A Database for the Genes Mutated in the Finnish Disease Heritage Brought to the Next-Generation Sequencing Era. Hum Mutat 34:1458–1466. doi:10.1002/humu.22389

309. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2:158–164. doi:10.1038/s41551-018-0195-0

310. Pritchard JK, Wen X, Falush D. 2009. Documentation for structure software: Version 2.3.

311. Prohaska A, Racimo F, Schork AJ, Sikora M, Stern AJ, Ilardo M, Allentoft ME, Folkersen L, Buil A, Moreno-Mayar JV, Korneliussen T, Geschwind D, Ingason A, Werge T, Nielsen R, Willerslev E. 2019.

Human Disease Variation in the Light of Population Genomics. Cell 177:115–131. doi:10.1016/j.cell.2019.01.052

312. Promislow DEL. 2004. Protein networks, pleiotropy and the evolution of senescence. Proc R Soc Lond B Biol Sci 271:1225–1234. doi:10.1098/rspb.2004.2732
313. Puigbò P, Wolf YI, Koonin EV. 2019. Genome-Wide Comparative Analysis of Phylogenetic Trees: The Prokaryotic Forest of Life In: Anisimova M, editor. Evolutionary Genomics. New York, NY: Springer New York. pp. 241–269. doi:10.1007/978-1-4939-9074-0_8
314. R Core Team. 2016. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
315. Racimo F, Marnetto D, Huerta-Sánchez E. 2016. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Mol Biol Evol msw216. doi:10.1093/molbev/msw216
316. Rackham OJL, Shihab HA, Johnson MR, Petretto E. 2015. EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. Nucleic Acids Res 43:e33–e33. doi:10.1093/nar/gku1322
317. Rambaut A. 2014. FigTree 1.4. 2 software. Inst Evol Biol Univ Edinb.
318. Ravindran V, Nacher JC, Akutsu T, Ishitsuka M, Osadcenco A, Sunitha V, Bagler G, Schwartz J-M, Robertson DL. 2019. Network controllability analysis of intracellular signalling reveals viruses are actively controlling molecular systems. Sci Rep 9. doi:10.1038/s41598-018-38224-9
319. Raymond M, Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86:248–249.
320. Rebolledo R, Navarrete SA, Kéfi S, Rojas S, Marquet PA. 2019. An Open-System Approach to Complex Biological Networks. SIAM J Appl Math 79:619–640. doi:10.1137/17M1153431
321. Reedijk AMJ, Klein K, Coebergh JWW, Kremer LC, Dinmohamed AG, de Haas V, Versluijs AB, Ossenkoppele GJ, Beverloo HB, Pieters R, Zwaan CM, Kaspers GJL, Karim-Kos HE. 2019. Improved survival for children and young adolescents with acute myeloid leukemia: a Dutch study on incidence, survival and mortality. Leukemia 33:1349–1359. doi:10.1038/s41375-018-0314-7

322. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin J-J, Kelso J, Slatkin M, Pääbo S. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060. doi:10.1038/nature09710

323. Reich Lab. 2019. Downloadable genotypes of present-day and ancient DNA data (compiled from published papers). David Reich Lab. https://reich.hms.harvard.edu/downloadable-genotypes-worlds-published-ancient-dna-data

324. Rice AM, McLysaght A. 2017. Dosage-sensitive genes in evolution and disease. BMC Biol 15:78. doi:10.1186/s12915-017-0418-y

325. Rice AM, McLysaght A. 2017a. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. Nat Commun 8:14366. doi:10.1038/ncomms14366

326. Rice AM, McLysaght A. 2017b. Dosage-sensitive genes in evolution and disease. BMC Biol 15:78. doi:10.1186/s12915-017-0418-y

327. Richard G-F, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev MMBR 72:686–727. doi:10.1128/MMBR.00011-08

328. Roberts T. 2018. Update 3 on HIV/N6 Experiment.

329. Rodriguez F, Kenefick A, Arkhipova I. 2017. LTR-Retrotransposons from Bdelloid Rotifers Capture Additional ORFs Shared between Highly Diverse Retroelement Types. Viruses 9:78. doi:10.3390/v9040078

330. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, Valencia A, Tress ML. 2013. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res 41:D110–D117. doi:10.1093/nar/gks1058

331. Rogers AR, Bohlender RJ, Huff CD. 2017. Early history of Neanderthals and Denisovans. Proc Natl Acad Sci 114:9859–9863. doi:10.1073/pnas.1706426114

332. Rook GAW, Lowry CA, Raison CL. 2013. Microbial 'Old Friends', immunoregulation and stress resilience. Evol Med Public Health 2013:46–64. doi:10.1093/emph/eot004
333. Rosenberg NA. 2002. Genetic Structure of Human Populations. Science 298:2381–2385. doi:10.1126/science.1078311
334. RStudio Team. 2016. RStudio: Integrated Development for R. RStudio.
335. RStudio Team. 2016. RStudio: Integrated Development for R. RStudio.
336. Santayana G. 1923. Soliloquies in England and later soliloquies. New York: C. Scribner's Sons.
337. Sanz J, Randolph HE, Barreiro LB. 2018. Genetic and evolutionary determinants of human population variation in immune responses. Curr Opin Genet Dev 53:28–35. doi:10.1016/j.gde.2018.06.009
338. Sarkar D. 2017. lattice: Trellis Graphics for R.
339. Sazonovs A, Barrett JC. 2018. Rare-Variant Studies to Complement Genome-Wide Association Studies. Annu Rev Genomics Hum Genet 19:97–112. doi:10.1146/annurev-genom-083117-021641
340. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341–345. doi:10.1038/nature04562
341. Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. 2012. HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. PLoS ONE 7:e31826. doi:10.1371/journal.pone.0031826
342. Schafer MJ, Atkinson EJ, Vanderboom PM, Kotajarvi B, White TA, Moore MM, Bruce CJ, Greason KL, Suri RM, Khosla S, Miller JD, Bergen HR, LeBrasseur NK. 2016. Quantification of GDF11 and Myostatin in Human Aging and Cardiovascular Disease. Cell Metab 23:1207–1215. doi:10.1016/j.cmet.2016.05.023
343. Schuster-Böckler B, Conrad D, Bateman A. 2010. Dosage Sensitivity Shapes the Evolution of Copy-Number Varied Regions. PLOS ONE 5:e9474. doi:10.1371/journal.pone.0009474
344. Sémon M, Wolfe KH. 2007. Consequences of genome duplication. Curr Opin Genet Dev 17:505–512. doi:10.1016/j.gde.2007.09.007

345. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. 2005. Segmental Duplications and Copy-Number Variation in the Human Genome. Am J Hum Genet 77:78–88. doi:10.1086/431652

346. Neuropsychopharmacol 29:S834–S835. doi:10.1016/j.euroneuro.2017.08.096

347. Shaw A, Toma C, Allcock R, Heath A, Pierce K, Mitchell PB, Schofield P, Fullerton J. 2019. COMBINED WHOLE EXOME SEQUENCING AND LINKAGE ANALYSIS REVEALS LINKAGE TO 10Q11-10Q21 LOCUS WHICH IS NOT EXPLAINED BY GWAS-ASSOCIATED SNP OR RARE VARIANTS IN ANK3. Eur Neuropsychopharmacol 29:S834–S835. doi:10.1016/j.euroneuro.2017.08.096

348. Shor P. 2019. This Is No Clockwork Universe. Twitter.

349. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. doi:10.1038/msb.2011.75

350. Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. 2012. On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates. Cell Rep 2:1387–1398. doi:10.1016/j.celrep.2012.09.034

351. Singh PP, Arora J, Isambert H. 2015. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. PLOS Comput Biol 11:e1004394. doi:10.1371/journal.pcbi.1004394

352. Singh PP, Isambert H. 2019. OHNOLOGS v2: A comprehensive resource for the genes retained from whole genome duplication in vertebrates (preprint). Bioinformatics. doi:10.1101/717124

353. Singmann P, Shem-Tov D, Wahl S, Grallert H, Fiorito G, Shin S-Y, Schramm K, Wolf P, Kunze S, Baran Y, Guarrera S, Vineis P, Krogh V, Panico S, Tumino R, Kretschmer A, Gieger C, Peters A, Prokisch H, Relton CL, Matullo G, Illig T, Waldenberger M, Halperin E. 2015.

Characterization of whole-genome autosomal differences of DNA methylation between men and women. Epigenetics Chromatin 8:43. doi:10.1186/s13072-015-0035-3

354. Slatkin M, Racimo F. 2016. Ancient DNA and human history. Proc Natl Acad Sci 113:6380–6387. doi:10.1073/pnas.1524306113
355. Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, Hajdinjak M, Peyrégne S, Nagel S, Brown S, Douka K, Higham T, Kozlikin MB, Shunkov MV, Derevianko AP, Kelso J, Meyer M, Prüfer K, Pääbo S. 2018. The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature 561:113–116. doi:10.1038/s41586-018-0455-x
356. Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022–1024. doi:10.1038/4151022a
357. Søe MJ, Nejsum P, Fredensborg BL, Kapel CMO. 2015. DNA Typing of Ancient Parasite Eggs from Environmental Samples Identifies Human and Animal Worm Infections in Viking-Age Settlement. J Parasitol 101:57. doi:10.1645/14-650.1
358. Søe MJ, Nejsum P, Seersholm FV, Fredensborg BL, Habraken R, Haase K, Hald MM, Simonsen R, Højlund F, Blanke L, Merkyte I, Willerslev E, Kapel CMO. 2018. Ancient DNA from latrines in Northern Europe and the Middle East (500 BC–1700 AD) reveals past parasites and diet. PLOS ONE 13:e0195481. doi:10.1371/journal.pone.0195481
359. Soltis PS, Soltis DE, Savolainen V, Crane PR, Barraclough TG. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. Proc Natl Acad Sci 99:4430–4435. doi:10.1073/pnas.032087199
360. Soubry A. 2015. Epigenetic inheritance and evolution: A paternal perspective on dietary influences. Prog Biophys Mol Biol 118:79–85. doi:10.1016/j.pbiomolbio.2015.02.008
361. Speed CW, Meekan MG, Bradshaw CJ. 2007. Spot the match – wildlife photo-identification using information theory. Front Zool 4:2. doi:10.1186/1742-9994-4-2
362. Spencer H. 1864. Principles of Biology. London: Williams and Norgate.

363. Spielman SJ. 2019. Model fit does not predict accuracy in single-gene protein phylogenetics (preprint). Evolutionary Biology. doi:10.1101/698860

364. Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller H-J, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K, Kahn RS, Linszen DH, van Os J, Wiersma D, Bruggeman R, Cahn W, de Haan L, Krabbendam L, Myin-Germeys I. 2008. Large recurrent microdeletions associated with schizophrenia. Nature 455:232–236. doi:10.1038/nature07229

365. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136:665–677. doi:10.1007/s00439-017-1779-6

366. Strachan DP. 1989. Hay fever, hygiene, and household size. BMJ 299:1259–1260. doi:10.1136/bmj.299.6710.1259

367. Strona G, Nappo D, Boccacci F, Fattorini S, San-Miguel-Ayanz J. 2014. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. Nat Commun 5:4114. doi:10.1038/ncomms5114

368. Su Z. 2005. Evolution of alternative splicing after gene duplication. Genome Res 16:182–189. doi:10.1101/gr.4197006

369. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med 12:e1001779. doi:10.1371/journal.pmed.1001779

370. Suh A. 2019. Genome Size Evolution: Small Transposons with Large Consequences. Curr Biol 29:R241–R243. doi:10.1016/j.cub.2019.02.032

371. Sun X, Lu J, Ma X. 2019. An efficient method for noninvasive prenatal diagnosis of fetal trisomy 13, trisomy 18, and trisomy 21. PLOS ONE 14:e0215368. doi:10.1371/journal.pone.0215368

372. Sunyaev S. 2001. Prediction of deleterious human alleles. Hum Mol Genet 10:591–597. doi:10.1093/hmg/10.6.591

373. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering C von. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47:D607–D613. doi:10.1093/nar/gky1131

374. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. 2017. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res 45:D362–D368. doi:10.1093/nar/gkw937

375. Szyf M. 2015. Nongenetic inheritance and transgenerational epigenetics. Trends Mol Med 21:134–144. doi:10.1016/j.molmed.2014.12.004

376. Tasdighian S, Bel MV, Li Z, Peer YV de, Carretero-Paulet L, Maere S. 2017. Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity. Plant Cell 29:2766–2785. doi:10.1105/tpc.17.00313

377. Tashiro T, Ishida A, Hori M, Igisu M, Koike M, Méjean P, Takahata N, Sano Y, Komiya T. 2017. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. Nature 549:516–518. doi:10.1038/nature24019

378. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47:D941–D947. doi:10.1093/nar/gky1015

379. Teufel AI, Liu L, Liberles DA. 2016. Models for gene duplication when dosage balance works as a transition state to subsequent neo- or sub-functionalization. BMC Evol Biol 16:45. doi:10.1186/s12862-016-0616-1

380. The 1000 Genomes Project Consortium, Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M, Sherry S, Flicek P. 2012. The 1000 Genomes Project: data management and community access. Nat Methods 9:459–462. doi:10.1038/nmeth.1974

381. The British Society of Underwater Photographers. n.d. The Underwater Photographers Code of Conduct.

382. The Institute of Cancer Research. 2019. Revolutionary new Centre for Cancer Drug Discovery aims to deliver step change in cancer treatment.

383. The UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. Nature 526:82–90. doi:10.1038/nature14962

384. The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. doi:10.1093/nar/gky1049

385. Thomas F, Kareva I, Raven N, Hamede R, Pujol P, Roche B, Ujvari B. 2018. Evolved Dependence in Response to Cancer. Trends Ecol Evol 33:269–276. doi:10.1016/j.tree.2018.01.012

386. Thomas K, Benjamin R-K, Fernando P, Brian G, Matthias B, Jonathan F, Kyle K, Jessica H, Jason G, Sylvain C, Paul I, Damián A, Safia A, Carol W, Team JD. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. Stand Alone 87–90. doi:10.3233/978-1-61499-649-1-87

387. Towner A V, Wcisel MA, Reisinger RR, Edwards D, Jewell OJD. 2013. Gauging the Threat: The First Population Estimate for White Sharks in

South Africa Using Photo Identification and Automated Software. PLoS ONE 8:66035. doi:10.1371/journal.pone.0066035

388. Truty R, Paul J, Kennemer M, Lincoln SE, Olivares E, Nussbaum RL, Aradhya S. 2018. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genet Med 1. doi:10.1038/s41436-018-0033-5
389. Tsongalis GJ. 2018. Integrative Systems BiologyMolecular Pathology. Elsevier. pp. 205–215. doi:10.1016/B978-0-12-802761-5.00010-9
390. Turner DD. 2019. In defense of living fossils. Biol Philos 34:23. doi:10.1007/s10539-019-9678-y
391. Uricchio LH, Petrov DA, Enard D. 2019. Exploiting selection at linked sites to infer the rate and strength of adaptation. Nat Ecol Evol 3:977–984. doi:10.1038/s41559-019-0890-6
392. Vallabhajosyula RR, Chakravarti D, Lutfeali S, Ray A, Raval A. 2009. Identifying Hubs in Protein Interaction Networks. PLoS ONE 4:e5344. doi:10.1371/journal.pone.0005344
393. Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat Rev Genet 18:411–424. doi:10.1038/nrg.2017.26
394. van Oosterhout C, Hutchinson WilliamF, Wills DerekPM, Shipley P. 2004. MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4:535–538. doi:10.1111/j.1471-8286.2004.00684.x
395. van Rossum G. 1995. Python Software Foundation- Python Language v2.7.16.
396. Varki A, Geschwind DH, Eichler EE. 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. Nat Rev Genet 9:749–763. doi:10.1038/nrg2428
397. Veenstra-VanderWeele J, Christian SL, Cook, Jr. EH. 2004. AUTISM AS A PARADIGMATIC COMPLEX GENETIC DISORDER. Annu Rev Genomics Hum Genet 5:379–405. doi:10.1146/annurev.genom.5.061903.180050

398. Veitia RA, Birchler JA. 2010. Dominance and gene dosage balance in health and disease: why levels matter! J Pathol 220:174–185. doi:10.1002/path.2623

399. Vihinen M. 2015. Muddled genetic terms miss and mess the message. Trends Genet 31:423–425. doi:10.1016/j.tig.2015.05.008

400. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2008. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19:327–335. doi:10.1101/gr.073585.107

401. Viswanath A, Williams M. 2018. Trichuris Trichiura (Whipworm, Roundworm)StatPearls. Treasure Island (FL): StatPearls Publishing.

402. Wahle E. 1999. 3′-End processing of pre-mRNA in eukaryotes. FEMS Microbiol Rev 23:277–295. doi:10.1016/S0168-6445(99)00008-X

403. Waldron D. 2015. Tandem repeats and divergent gene expression: Human evolution. Nat Rev Genet 16:686–686. doi:10.1038/nrg4040

404. Walker RG, Poggioli T, Katsimpardi L, Buchanan SM, Oh J, Wattrus S, Heidecker B, Fong YW, Rubin LL, Ganz P, Thompson TB, Wagers AJ, Lee RT. 2016. Biochemistry and Biology of GDF11 and Myostatin: Similarities, Differences, and Questions for Future Investigation. Circ Res 118:1125–1141; discussion 1142. doi:10.1161/CIRCRESAHA.116.308391

405. Wall L. 2000. Programming Perl, 3rd ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc.

406. Wang DY-C, Kumar S, Hedges SB. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc R Soc Lond B Biol Sci 266:163–171. doi:10.1098/rspb.1999.0617

407. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. Science 350:1096–1101. doi:10.1126/science.aac7041

408. Wang Z, Lai C, Chen X, Yang B, Zhao S, Bai X. 2015. Flood hazard risk assessment model based on random forest. J Hydrol 527:1130–1141. doi:10.1016/j.jhydrol.2015.06.008

409. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang S-P, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente XS, López-Otín C, Ordóñez GR, Eichler EE, Chen L, Cheng Z, Deakin JE, Alsop A, Thompson K, Kirby P, Papenfuss AT, Wakefield MJ, Olender T, Lancet D, Huttley GA, Smit AFA, Pask A, Temple-Smith P, Batzer MA, Walker JA, Konkel MK, Harris RS, Whittington CM, Wong ESW, Gemmell NJ, Buschiazzo E, Vargas Jentzsch IM, Merkel A, Schmitz J, Zemann A, Churakov G, Kriegs JO, Brosius J, Murchison EP, Sachidanandam R, Smith C, Hannon GJ, Tsend-Ayush E, McMillan D, Attenborough R, Rens W, Ferguson-Smith M, Lefèvre CM, Sharp JA, Nicholas KR, Ray DA, Kube M, Reinhardt R, Pringle TH, Taylor J, Jones RC, Nixon B, Dacheux J-L, Niwa H, Sekita Y, Huang X, Stark A, Kheradpour P, Kellis M, Flicek P, Chen Y, Webber C, Hardison R, Nelson J, Hallsworth-Pepin K, Delehaunty K, Markovic C, Minx P, Feng Y, Kremitzki C, Mitreva M, Glasscock J, Wylie T, Wohldmann P, Thiru P, Nhan MN, Pohl CS, Smith SM, Hou S, Nefedov M, de Jong PJ, Renfree MB, Mardis ER, Wilson RK. 2008. Erratum: Genome analysis of the platypus reveals unique signatures of evolution. Nature 455:256–256. doi:10.1038/nature07253

410. Waskom M, Botvinnik O, O'Kane D, Hobson P, Ostblom J, Lukauskas S, Gemperline DC, Augspurger T, Halchenko Y, Cole JB, Warmenhoven J, Ruiter JD, Pye C, Hoyer S, Vanderplas J, Villalba S, Kunter G, Quintero E, Bachant P, Martin M, Meyer K, Miles A, Ram Y, Brunner T, Yarkoni T, Williams ML, Evans C, Fitzgerald C, Brian, Qalieh A. 2018. mwaskom/seaborn: v0.9.0 (July 2018). Zenodo. doi:10.5281/zenodo.592845

411. Webb AJ, Thorisson GA, Brookes AJ, on behalf of the GEN2PHEN Consortium. 2011. An informatics project and online "Knowledge Centre" supporting modern genotype-to-phenotype research. Hum Mutat 32:543–550. doi:10.1002/humu.21469

412. Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. 2017. corrplot: Visualization of a Correlation Matrix.

413. Wells HG. 1945. Mind At The End Of Its Tether. London: William Heinmann.
414. Wickham H, Francois R, Henry L, Müller K, RStudio. 2017. dplyr: A Grammar of Data Manipulation.
415. Wickham H, Henry L, RStudio. 2018a. tidyr: Easily Tidy Data with "spread()" and "gather()" Functions.
416. Wickham H, Hester J, Chang W, RStudio, R) RC team (Some namespace and vignette code extracted from base. 2018b. devtools: Tools to Make Developing R Packages Easier.
417. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.
418. Wickham H. 2016. plyr: Tools for Splitting, Applying and Combining Data.
419. Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.
420. Wilke CO, Sawyer SL. 2016. At the mercy of viruses. eLife 5:e16758. doi:10.7554/eLife.16758
421. Williams GM, Brady R. 2019. Patau SyndromeStatPearls. Treasure Island (FL): StatPearls Publishing.
422. Winterer G, Konrad A, Vucurevic G, Musso F, Stoeter P, Dahmen N. 2008. Association of 5′ end neuregulin-1 (NRG1) gene variation with subcortical medial frontal microstructure in humans. NeuroImage 40:712–718. doi:10.1016/j.neuroimage.2007.12.041
423. Wolfman NM, McPherron AC, Pappano WN, Davies MV, Song K, Tomkinson KN, Wright JF, Zhao L, Sebald SM, Greenspan DS, Lee S-J. 2003. Activation of latent myostatin by the BMP-1/tolloid family of metalloproteinases. Proc Natl Acad Sci U S A 100:15842–15846. doi:10.1073/pnas.2534946100
424. Wood HM, González VL, Lloyd M, Coddington J, Scharff N. 2018. Next-generation museum genomics: Phylogenetic relationships among palpimanoid spiders using sequence capture techniques (Araneae: Palpimanoidea). Mol Phylogenet Evol 127:907–918. doi:10.1016/j.ympev.2018.06.038

425. Woods S, Coghlan A, Rivers D, Warnecke T, Jeffries SJ, Kwon T, Rogers A, Hurst LD, Ahringer J. 2013. Duplication and Retention Biases of Essential and Non-Essential Genes Revealed by Systematic Knockdown Analyses. PLoS Genet 9:e1003330. doi:10.1371/journal.pgen.1003330

426. Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet 32:569–577. doi:10.1038/ng1202-569

427. World Health Organisation. 2018a. Soil-transmitted helminth infections. https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections

428. World Health Organisation. 2018b. Report of the Eleventh Meeting of the WHO Strategic and Technical Advisory Group for Neglected Tropical Diseases. Geneva.

429. Worm B, Davis B, Kettemer L, Ward-Paige CA, Chapman D, Heithaus MR, Kessel ST, Gruber SH. 2013. Global catches, exploitation rates, and rebuilding options for sharks. Mar Policy 40:194–204. doi:10.1016/J.MARPOL.2012.12.034

430. Wu X, Hurst LD. 2016. Determinants of the Usage of Splice-Associated cis -Motifs Predict the Distribution of Human Pathogenic SNPs. Mol Biol Evol 33:518–529. doi:10.1093/molbev/msv251

431. Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C. 2012. Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing. Am J Hum Genet 91:1022–1032. doi:10.1016/j.ajhg.2012.10.015

432. Yagishita N, Yamaguchi A. 2009. Isolation and characterization of eight microsatellite loci from the longheaded eagle ray, Aetobatus flagellum (Elasmobranchii, Myliobatidae). Mol Ecol Resour 9:1034–1036. doi:10.1111/j.1755-0998.2009.02568.x

433. Yang J, Jin Z-B, Chen J, Huang X-F, Li X-M, Liang Y-B, Mao J-Y, Chen X, Zheng Z, Bakshi A, Zheng D-D, Zheng M-Q, Wray NR, Visscher PM, Lu F, Qu J. 2017. Genetic signatures of high-altitude

adaptation in Tibetans. Proc Natl Acad Sci 114:4189–4194. doi:10.1073/pnas.1617042114

434. Yang Z, Hu F. 2018. Investigation of gene evolution in vertebrate genome reveals novel insights into spine study. Gene 679:360–368. doi:10.1016/j.gene.2018.09.023

435. Yu X, Zeng T, Wang X, Li G, Chen L. 2015. Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. J Transl Med 13:189. doi:10.1186/s12967-015-0546-5

436. Yuan T WL Masaaki H. 2016. ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. R J 82 478–489.

437. Yuan T, Masaaki H, Wenxuan L. 2016. ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. R J 82 478–489.

438. Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. Nat Rev Genet 16:172–183. doi:10.1038/nrg3871

439. Zeng TC, Aw AJ, Feldman MW. 2018. Cultural hitchhiking and competition between patrilineal kin groups explain the post-Neolithic Y-chromosome bottleneck. Nat Commun 9. doi:10.1038/s41467-018-04375-6

440. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. 2018. Ensembl 2018. Nucleic Acids Res 46:D754–D761. doi:10.1093/nar/gkx1098

441. Zhang Y, Wei Y, Liu D, Liu F, Li X, Pan L, Pang Y, Chen D. 2017. Role of growth differentiation factor 11 in development, physiology and disease. Oncotarget 8. doi:10.18632/oncotarget.20258

442. Zhong Q, Simonis N, Li Q-R, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D, Swearingen V, Yildirim MA, Yan H, Dricot A, Szeto D, Lin C, Hao T, Fan C, Milstein S, Dupuy D, Brasseur R, Hill DE, Cusick ME, Vidal M. 2009. Edgetic perturbation models of human inherited disorders. Mol Syst Biol 5. doi:10.1038/msb.2009.80

443. Zhou X, Menche J, Barabási A-L, Sharma A. 2014. Human symptoms–disease network. Nat Commun 5:4212. doi:10.1038/ncomms5212

444. Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu Y-F, McSweeney KM, Ben-Zeev B, Nissenkorn A, Anikster Y, Oz-Levi D, Dhindsa RS, Hitomi Y, Schoch K, Spillmann RC, Heimer G, Marek-Yagel D, Tzadok M, Han Y, Worley G, Goldstein J, Jiang Y-H, Lancet D, Pras E, Shashi V, McHale D, Need AC, Goldstein DB. 2015. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet Med 17:774–781. doi:10.1038/gim.2014.19147

APPENDIX I: FIRST YEAR PHD CONTINUATION REPORT: UNDERSTANDING DELETERIOUS VARIANTS IN THE CONTEXT OF 'NORMAL' GENETIC VARIATION

Alexandra Martin-Geary MSc[1]

[2]Division of Evolution and Genomic science, University of Manchester, Oxford Rd, Manchester M13 9PL

1st year PhD Continuation Report

Understanding deleterious variants in the context of 'normal' genetic variation

Alexandra Martin-Geary

Supervisor: David Robertson

1. ABSTRACT

It has been found that each individual carries, on average, two potentially disease causing mutations (Dorschner et al 2013; Zhu et al 2012) and between 103 (Sunyaev et al 2001) and 400 potentially damaging variants. These variants are not however, uniformally distributed throughout the genome, with certain genes being more prone to disease associations than others. Using the phase three data produced by the 1000 genomes project, which details genomic variation in 2504 individuals from five super-populations across the globe, alongside data pertaining to genes which have arisen from whole genome duplication events, known as ohnologs, and OMIM disease data, it has been found that healthy individuals are able to tolerate relatively high proportions of variation in ohnologous genes linked with both dominant and recessive disorders. This, it is hypothesized, is due to the presence of gene pairs which are able balance dosage in healthy individuals, and a pattern of compound heterozygosity leading to compensation.

2. ACKNOWLEDGEMENTS

I would like to thank a number of people for their support throughout this project. Firstly Prof. David Robertson for being tremendously supportive and infectiously enthusiastic at times when that has been greatly needed. Secondly, Ana Barradas, Nikita Abramovs, Avraam Tapinos and Bede Constantinedes for being on hand with advice and fixes when things were not working as expected. To Ralph and Ethel for keeping me company when I was tearing my hair out at 3am, and finally Kim Corbett-Ouaichai, who was tremendously encouraging towards my doing a PhD, but sadly passed away before it began.

Contents

List of Figures

List of Tables

3. INTRODUCTION

It has been found that each individual carries, on average, two potentially disease causing mutations (Dorschner et al 2013; Zhu et al 2012) and between 103 (Sunyaev et al 2001) and 400 potentially damaging variants (1000 Genomes 2012), with ~10% of genes known to have disease associations (Barabási et al 2011).

Genetic variation is by no means uniform across the genome. Conservation studies have shown that between species there is a heightened propensity towards potentially deleterious variants in the most conserved regions in the metazoan lineage (Miller and Kumar 2001), whilst regions of interspecies plasticity are far less likely to contain potentially disease associated mutations (*ibid*). In order to obtain a clearer understanding of variation, and subsequently predict the occurrence of deleterious mutations, it is important to understand the patterns of distribution and evolution of variants across the human genome (Henn *et al* 2015).

The nature of human diversity is such that it is unlikely that we will be able to establish a 'one size fits all' Human profile (UK10K 2015). However population specific patterns are beginning to be seen which may assist in the elucidation of population profiles. Of course, given increasing migration and admixture, this may in future change. It is clear that the evolution of the human genome has, in no small part been moulded by population structure and dynamics (Cavalli-Sforza et al; Polvi et al 2013, Rosenberg et al 2002, Frazer et al 2009), this having been highlighted in late 2015 when an analysis of individuals from 26 populations showed significant inter-population variation, with ~86% of variants being found to be population specific (1000 genomes 2015).

3.1 SMALL-SCALE VARIATIONS

Disease causing variation is divided, as standard, into two categories: Mendelian – wherein one gene equates to one outcome and, more commonly complex – where a composite of variants in multiple genes results in a disease phenotype (Barabási et al 2011). The majority of variants however are not thought to be deleterious (Frazer et al 2009), with a suggested 9.5% of the human genome composed of non-damaging mutations (Zarrei et al 2015). Conversely the most deleterious variants are unlikely to be seen within sequencing studies, as these variations are likely to result in foetal

inviability. Therefore it is unlikely that the 'worst case' variants will ever appear in the raw data.

Many considerations need to be made when assessing disease susceptibility and mutational load. Population specific factors have been briefly discussed, further to this however, are penetrance, frequency, hetero and homogeneity, epistasis, the presence of de-novo mutations, epigenetic modifications, and, as shown in figure 1 variant type, and location.

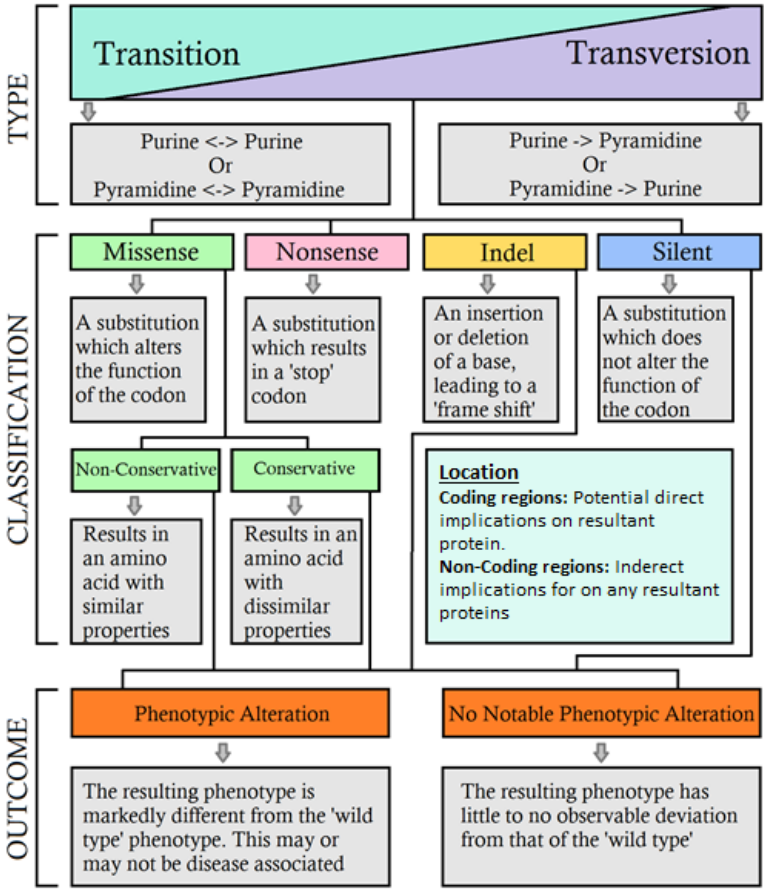


Figure 28: Overview of variation. Including type, classification and potential outcome

In September 2015 completion of the final phase of the thousand-genome project (1000g) was announced, with 2504 genomes being made publically available for analysis (Birney and Soranzo 2015). From this data it has been found that ~97% of variants are found in non-coding regulatory regions (1000 genome 2015) the impact of which relatively little is known (UK10K 2015). Many of these variants however, are likely to impact upon the resultant phenotype (UK10K 2015; Maurano et al 2015). We are therefore entering a new phase of variant analysis, where, given the greater accessibility of whole genome data, it is imperative that non-coding regions are more readily incorporated into study (UK10K 2015).

The mutation of a single base, known as point mutation, or single nucleotide polymorphism (SNP) is the most abundant individual form of mutation in the human genome (1000 Genome 2015), their influence on phenotype being dependent on; their location in, or proximity to a gene. Single nucleotide polymorphisms are often categorized, based on their properties, into either Synonymous mutations; variants which do not radically alter the genomic function, or non-synonymous; leading to altered function, which occur at an average rate of one per 1000 amino acids (Sunyaev et al 2001). It has been suggested that ~20% of common, non-synonymous SNPs alter function (Sunyaev et al 2001), with a recent study of 18,903 genes indicating that 576 of those contained a mutation likely to have functional repercussions (UK10K 2015).One highly contentious element of this latter type of mutation are missense variants (both conservative and non-conservative). These mutations are commonly, and often erroneously, classed as functional (Henn et al 2015). The functional repercussions of missense mutations are better viewed as a continuum, covering the span between total loss of function, to gain of function. In 2013 Petrovski et al conducted a study into the effects of variation on personal genomes. The initial phase of the study, mistakenly classified all missense mutations as being functional, however, by supplementing the analysis with the use of PolyPhen2 , they found a reduction of mutations classified as functional of 33%, illustrating the utility of using such tools.

The ramifications of insertions or deletions (indels) are often easier to predict, as they result in codon misalignments, however, this does not always result in an aberrant phenotype. Further to this, analysis of splice variants has shown that RNA mis-splicing and the accumulation therein of variants and polymorphisms, has led to numerous human diseases (Cáceres and Kornblihtt 2002 & Cartegni et al 2002, Scotti and Swanson 2016).

3.2 STRUCTURAL VARIATIONS

Whilst single nucleotide polymorphisms account for the largest proportion of total mutations, structural variations account for an estimated ~20 million nucleotides (1000 genome 2015) often spanning many thousands of bases. Distribution patterns of structural variants are similar to that of their point and indel counterparts, being enriched in genomic regions that have been recently subjected to large-scale

duplications (Frazer et al 2009). An estimated ~5% of the human genome is made up of such regions, which, impact on phenotypic variation from the norm (Bishara et al 2015). Targeting regions high in repeated sequence is difficult, in terms of both sequencing, and alignment. Due to these difficulties, the role of large-scale copy-number variants (CNVs) in healthy individuals are poorly understood (Iafrate et al 2004). A recent study by the 1000 genome consortium however, has attempted to quantify various types of structural variation within the human genome (Table 1).

Table 9: structural variants (1000 genome 2015).

Mutation Type	Occurrences
Large Deletions	~1000
Alu Insertions	~915
CNVs	~160
L1 insertions	~128
SVA insertions	~51
Inversions	~10
NUMTs	~4

CNVs arising through whole genome duplication events, de novo duplications, singleton duplication and segmental duplication, have, in recent years begun to be more thoroughly investigated, and found to play a significant role in disease (McLysaght et al 2014), with in the region of two thirds of CNVs resulting in functional change (Dudley et al 2012). These CNVs can represent both increase, and decrease in gene numbers, with estimates of gain per chromosome ranging from 1.1% to 16.4%, and loss from 4.3% to 19.2% (Zarrei et al 2015) alongside large scale variation caused by both chromosomal inversions, and translocations (Lupski 1998; Abeysinghe et al 2003).

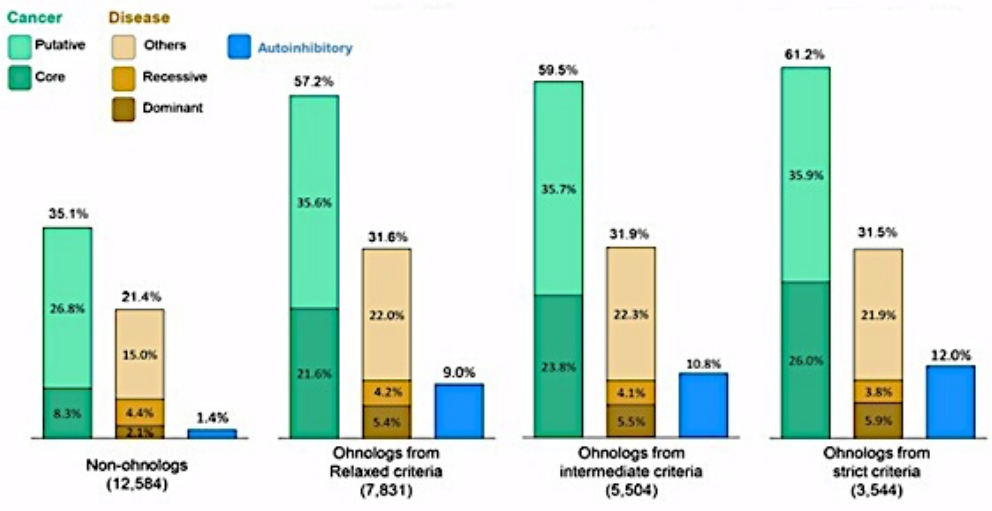


Figure 29: Involvement of Ohnologs within disease (Singh et al 2015)

Whole genome duplication is a contributory factor to structural variation, postulated to have occurred twice in human evolutionary history (Dehal and Boore 2005).

CNVs arising, and having been retained as a result of these instances of WGD, are known as Ohnologs (Wolfe 2000). There are approximately 3544-7831 of these in the human genome (Singh et al 2015), and have been linked to a propensity for deleterious mutation and association with cancer genes (Figure 2), which account for 21.6% to 26% of all Ohnologs (8.3% of non Ohnologs), developmental genes, gene regulation and signalling, and autosomal Dominant disease (Singh et al 2015).

3.3 EPIGENETICS

Whilst still being in its relative infancy, the study of epigenetic variations has provided useful information regarding mechanisms of non-genetic inheritance (Szyf 2015), including the role of noncoding RNAs, Chromatin and histone modifications leading to increased accessibility of DNA, and, DNA methylation and phosphorylation leading to chromatin inactivation and interruption of transcription factor binding (Szyf 2015). Epigenetic silencing, particularly of regions high in CNV, has a notable impact on an individual's phenotype (Schuster-Böckler et al 2009), and a high correlation between chromatin organization and the occurrence of segmental duplications (Ebert *et al* 2014). It has also been suggested that factors such as CpG hypermutability have led to increased mutation rates and diversification (Li et al 2010), as the rapidity of evolution of epigenetic modifications is considerably greater than that of the lengthier process of genetic selection (Szyf 2015).

One of the particular difficulties posed by the study of epigenetics however, is understanding heritability, as the majority of parental epigenetic modifications are erased during meiosis (Szyf 2015). It is proposed that the mechanism by which retention is most likely to occur, is through transmission by noncoding RNAs of behavioural signals (Szyf 2015; Soubry 2015). Paternal gamete-mediated epigenetic inheritance in offspring may impact on the phenotype (as opposed to the gestational impacts that have mainly been studied) as spermatogenesis, a continuing process occurring throughout the lifespan of the male, may lead to the accumulation of epigenetic modifications retained in this cell type (Szyf 2015).

3.4 GENETIC INTERACTIONS

The co-occurrence of loss of function is of particular significance when assessing the role of variation in complex disorders, it has been shown, as the quantities of data available for analysis has increased, that genetic interactions between loci are highly contributory to complex disease (Marchini et al 2005), and that disease pairs with variations in functionally similar domains show a greater degree of comorbidity than those in which variation occurs in more dissimilar regions (Barabási et al 2011).

3.5 ESSENTIALITY, TOLERANCE, AND DOSAGE BALANCE

An estimated 77% of essential genes emerged in a pre-metazoan progenitor (Blomen et al 2015). These regions are purportedly invariant, as any variation within them appears to have been subject to strong purifying pressures (Wang et al 2015). This indicates that variation within these regions likely leads to major deleterious functional outcomes, and cannot therefore be tolerated (Miller and Kumar 2001; Henn et al 2015; Wang et al 2015). Population studies however, have shown that, disease can, and does, occur in genes essential to cellular health, and, variations persist (Barabási et al 2011). Ohnologs, as the product of whole genome duplications, are also highly conserved. However, despite this fact, they have been found to show no more, or less essentiality than that of singletons (Makino & McLysaght 2010).

The disruption of stoichiometric balance posed by variations in dosage can be a significant issue to cellular health (Schuster-Böckler et al 2009). However, due to compensatory mechanisms, alongside epigenetic silencing, it can be difficult to establish the degree of imbalance posed by CNVs from genomic data alone (*ibid*). It has been found that there are a number of genes, which may be considered 'mildly dosage sensitive', in the case of these genes, variations in copy number may have an impact, which, whilst variation may not be viewed as significantly deleterious in a single gene, interactions between a combination of mildly dosage sensitive genes, particularly those within an essential network may have major phenotypic repercussions (McLysaght et al 2013).

3.6 THE RELATIONSHIP BETWEEN GENOTYPE AND PHENOTYPE

In order to better understand and predict both functional and wider disease ramifications of genetic variants with greater precision, one must consider the molecular components, global organization and interconnected environment of the cell (Wang et al 2015; Barabási et al 2011). Beyond this it is suggested that in order to understand phenotypes resulting from rare variants, allelic architecture between loci should be more thoroughly characterised (UK10K 2015).

Maurano et al's seminal study (2015) discovered upwards of 60 thousand variants impacting on regulatory DNA accessibility, and transcription factor occupancy. Using DNase I hypersensitive site sequencing, combined with DNA genotyping on multiple tissues and individuals, they were able to identify 500,000 variants, common to regulatory regions, which directly impact transcription factor occupancy (ibid). This work highlights the importance of both the involvement of epigenetic modification in genetic disease, but also the need to account for both non-coding DNA, and, cell specific processes.

3.7 METHODOLOGICAL CONSIDERATIONS

Many of the tools and methods currently used have been designed to assess large populations as a cohort; therefore, their clinical use for individual diagnostic purposes is limited. For example, ExAC's anonymity criteria make it impossible to extract a single participant, and compare regions of one individual's DNA. Over the last fifteen years, as

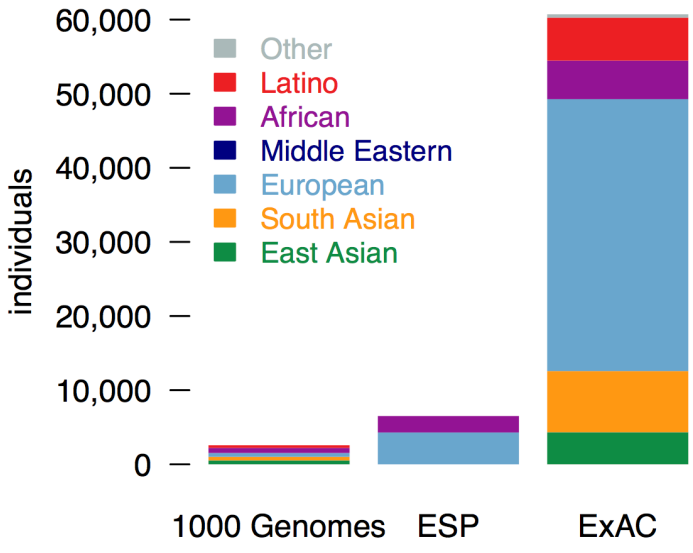


Figure 30: Number of genomes in each dataset by population (Lek et al 2015)

sequencing methods have become cheaper, faster, and more precise (UK10K 2015) there has been a deluge of data garnered in both publicly available and private datasets. The 1000 genomes project, combining open source data and tools to access it (Clarke et al 2012), is illustrative of this, with 2504 human genomes having recently been made publically available.

It has been highlighted (Dorschner et al 2013) however, that studies have focused disproportionately on individuals of European descent, thus skewing the findings by underrepresentation of other populations. This is of particular significance given that rare alleles tend to be population specific (UK10K 2015).

Despite the large quantities of data now at our disposal, and additions to this data on the horizon, as observed by the UK10K consortium (2015) (Figure 3) regarding their use of the ClinVar database; without the appropriate tools to analyse, identify and clinically assess the variation within the human genome our ability to draw meaningful conclusions will remain limited.

It is clear that the human genome and epigenome are prone to vicissitude. Variation of both small and large scales are numerous and diverse, the functional ramifications of which are wide reaching, from foetal inviability, to lethality, to zero phenotypic change, to benefit, or a combination thereof. The analysis of human variation and its role in disease has reached a crossroads. Whilst, as has been highlighted, the rapid rise in available data has led to much advancement, it has also been the catalyst for an excess of issues. Perhaps as a repercussion of this, the erroneous use of terminology, and proliferation of disparate nomenclature relating to the same entities has occurred, which it is imperative that we quell through standardisation.

3.9 HYPOTHESIS AND AIMS

It is hypothesized that there are as yet undiscovered underlying dependencies in the organisation of our system arising from being an evolved system, which lead to disparate disease propensities in differing genes, and regions of the genome. The aim of this research will be to identify these propensities and establish potential patterns of inheritance, and evolutionary histories, influencing disease in human populations.

4. METHODS

The analysis primarily focused on the 1000g phase 3 data. This data had been pre-prepared to create a database instance for use with the MongoDB client (Abramovs 2016). This was divided into a number of indexes to aid in the ease of data extraction. Given prior reporting of variant frequencies within this data (60,000 structural variants, 3.6 million indels and 84.7million single nucleotide polymorphisms (1000 Genomes Project Consortium 2015)) for the purpose of the research it was decided that initially the focus would be single nucleotide polymorphisms (SNPs) as the most abundant variation type available through the variants index. Table 2 shows an example of the data extracted from within this index.

Table 10: Data extracted from 1000g MongoDB instance.

MongoDB 1000 genomes database interface, Variants Index (N. Abramovs)		
Element	Description	Example
id	A concatenated variant identifier, containing Chromosome, position, Reference allele and alternate allele.	1_12345678_A_T
VT	Type of variant (in this instance this was refined to just the single nucleotide polymorphisms).	SNP
INDIVIDUALS	Key-Value pairs where the key represents the individual and the value is the zygosity of the variant therein.	HG00551 => 1|0

Zygosity scores within this index were taken directly from the 1000g raw data and are presented as follows: For nucleotides congruent with the reference (REF) the value is 0, therefore a diploid call, homozygous for the REF the representation is 0|0. Heterozygous variants where the alternate base (ALT) is on the first allele are 1|0, and on the second allele 0|1. Homozygous ALT calls are 1|1, and for the haploid Y chromosome, 4.Individuals within the 1000g data were also placed into five super-populations Table 3.

Table 11:Super (column 1) and Sub (column 2) populations sampled in the 1000g project

1000g Populations	
Super-Population	Sub-Populations
AFR (African)	YRI Yoruba in Ibadan, Nigeria, LWK Luhya in Webuye, Kenya, GWD Gambian in Western Divisions in the Gambia MSL Mende in Sierra Leone, ESN Esan in Nigeria, ASW Americans of African Ancestry in SW USA, ACB African Caribbeans in Barbados
AMR (Ad-mixed American)	MXL Mexican Ancestry from Los Angeles USA, PUR Puerto Ricans from Puerto Rico, CLM Colombians from Medellin, Colombia, PEL Peruvians from Lima, Peru
EAS (East Asian)	CHB Han Chinese in Bejing, China, JPT Japanese in Tokyo, Japan, CHS Southern Han Chinese, CDX Chinese Dai in Xishuangbanna, China, KHV Kinh in Ho Chi Minh City, Vietnam
EUR (European)	CEU Utah Residents (CEPH) with Northern and Western Ancestry, TSI Toscani in Italia, FIN Finnish in Finland, GBR, British in England and Scotland, IBS Iberian Population in Spain

SAS (South Asian)	GIH Gujarati Indian from Houston, Texas, PJL Punjabi from Lahore, Pakistan, BEB Bengali from Bangladesh, STU Sri Lankan Tamil from the UK, ITU Indian Telugu from the UK

Files were extracted from MongoDB using a Python script (Mongo_populations.py) written in collaboration with Nikita Abramovs. These files formed a matrix of all SNPs in each of the five super populations, subdivided by chromosome, presenting the SNP id, Individual identifier, and zygosity score. It was not possible to output data per population for the whole genome in this initial phase, as the size of data being handled, it was rapidly realized was computationally exhaustive. For the purpose of these matrices zygosities were presented as follows: 0 = homozygous for REF, 1 = Heterozygous with ALT on the first allele, 2 = Heterozygous with ALT on the second allele, 3 = Homozygous for ALT, and 4 Hemizygous for ALT (appendix).Subsequent analysis was divided into phases, with any further scripts written in the Perl language. The following flow chart (Figure 4) outlines these phases, the flow of information between them, and any supplementary data used.

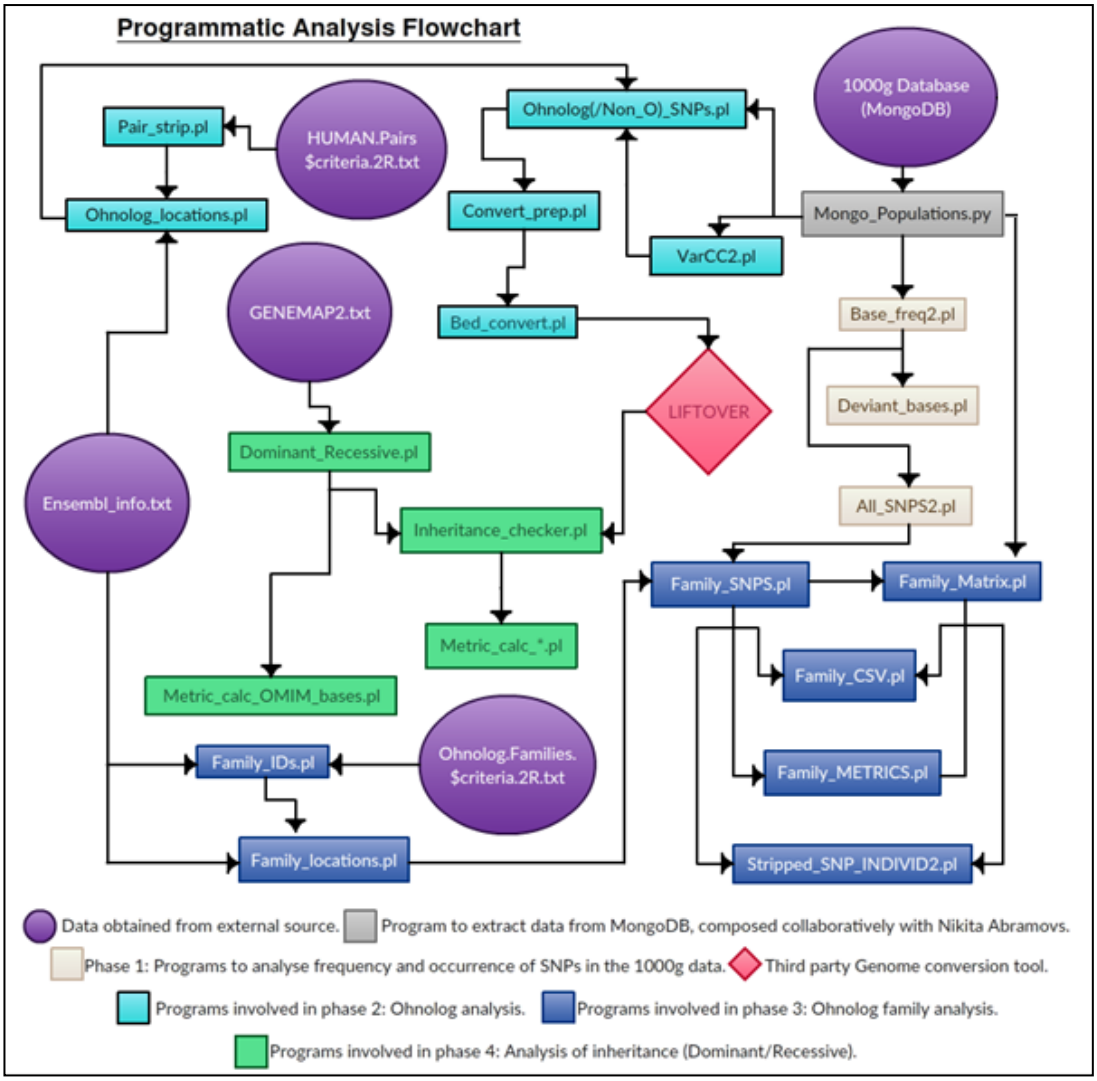


Figure 31: Programmatic analysis flowchart

4.1 ANALYSIS PHASE 1

The first phase of the analysis primarily involved the extraction and garnering a greater understanding of the 1000g data. As previously noted, a large portion of this was the extraction of suitable datasets using MongoDB and the Mongo_populations.py script.

Following extraction, this was interrogated using 3 Perl programs; Firstly the Base_freq2.pl script, designed to produce tab delimited text files per population per chromosome with SNP ids followed by the occurrence of each nucleotide within the super-population at that location (appendix). Following this the Deviant_bases.pl script was devised to establish the frequency with which the ALT occurred a greater number of times than the ref at each position (appendix). The final of the three main

Perl scripts used in this phase of analysis was All_SNPs2.pl. The function of this script was to amalgamate SNP data across the five super-populations and output a list of all SNPs, eradicating any duplicate SNPs (occurring in multiple populations) and provide a genomic representation of the SNPs for use in later phases of analysis. Beyond this, a number of smaller programs were devised to quantify various aspects of the SNP distribution within each population in the 1000g data.

4.2 ANALYSIS PHASE 2

The second phase of the analysis involved cross referencing the 1000g SNP data with a number of files from external sources. It was decided that in order to more appropriately manage such large data and address relevant questions the focus should be on a modular analysis. For this purpose two avenues were pursued; relationships with Ohnologs throughout phase 3, and Inheritance types (Dominant and recessive) as detailed in phase 4.

Firstly a list of Ohnolog pairs was obtained from http://ohnologs.curie.fr. This file (HUMAN.PAIRS.*.2R.txt) contains a list of gene pairs produced by Singh et al (2015) divided into three files conforming to the stringency criteria proposed therein (appendix). These files were prepared for use with Pair_Strip.pl by extracting the stable Id for each gene and the corresponding pair and outputting as a list (appendix). The next step was to extract from ensembl a large file (Ensembl_info.txt) containing relevant data for all genes within the database (in collaboration with Ana Barradas). This was a tab delimited text file containing: Gene name, Stable ID, Description, Gene type, Chromosome, Region start, Region end, Strand, and Synonyms (appendix).

These two files were cross referenced using Ohnolog_locations.pl to produce a list of start and end bases of each ohnolog gene using the stable id (appendix). The original matrix files were then revisited using the VarCC2.pl script. This program output a tab delimited file containing; Variant id, followed by the occurrence of each zygosity per variant for use in later analysis (appendix). These files were then cross referenced with the Ohnolog location files using Ohnolog_(non_o_)SNPs.pl to output a tab delimited file containing each variant id, if they resided within an ohnolog – the id thereof, genomic start and end points. For those SNPS which did not reside within an ohnolog, the SNP id was output as a separate list (appendix).

Due to disparities in reference chromosomes used by various data sources it was necessary to, at this point to convert the SNP coordinates from HG37 as used by Ensembl, 1000g and Singh et al to HG38, used by the Online Mendelian Inheritance in Man (OMIM) database before proceeding further. This was achieved by using Convert_prep.pl and Bed_convert.pl to produce a .bed file per chromosome. These files were then converted using the Liftover tool (available from genome.ucsc.edu/util.html) provided by UCSC on an Ubuntu platform to convert between the two reference genomes.

4.3 ANALYSIS PHASE 3

The third phase of analysis extended the ohnolog investigation to look into the relationship between SNPs and ohnolog families. The family information was obtained from http://ohnologs.curie.fr. Once again, three files were obtained representing the three stringency criteria. These files contained the ohnolog family name, number of genes in that family, followed by the names of each gene in that family. Due to the fact that the gene names were not in stable id format, it was necessary to convert these into their stable ids by cross referencing this file with the previously obtained ensembl data, using Family_IDs.pl. the output of this script was a file in the same format as the original ohnolog family files, with the gene name substituted for stable id. The locations of each gene were then established once again using the ensembl data (family_locations.pl) (appendix).

Once the Ohnolog family data had been appropriately prepared it was possible to cross reference the location files with the complete lists of SNPs present in each chromosome provided by All_SNPs2.pl using family_SNPs.pl. This provided an output file for each chromosome containing the ohnolog family name, stable id of each gene within each family, chromosome, start and end coordinates, and each SNP located therein (appendix) and a final file containing the ids of all SNPs for the whole genome.

The family_matrix.pl program was then used to cross reference the output from family_SNPs.pl with the original matrix files extracted from the 1000g data. This provided files per chromosome which were similar in format to the original matrices. The addition of two prefixing columns containing ohnlolog family name and gene (appendix) was then provided by family_CSV.pl.

The final stage in phase three was to output a file per ohnolog family containing; stable id, SNP id, followed by each individual id and the corresponding zygosity using Stripped_SNP_Individ2.pl (appendix).

4.4 ANALYSIS PHASE 4

The fourth stage of the analysis, as previously mentioned focused on establishing the relationship between SNPs and inheritance (dominant/recessive). The GENEMAP2.txt data file containing OMIM's Synopsis of the Human Gene Map, alongside genomic coordinates and inheritance was downloaded from the OMIM FTP (11/02/16). This was then parsed using the Dominant_Recessive.pl program in order to split the data into three files containing all rows listed as having associations with; dominant disorders, recessive disorders, and those with dual inheritance patterns.

The SNP files previously converted to HG38 were then cross referenced with the three OMIM files to produce files containing the original data provided, followed by a list of all SNPs contained within the start and end coordinates for each gene (appendix). Metrics for this data were then produced using the Metric_Calc.pl and Metric_calc_OMIM_bases.pl programs.

5. RESULTS

5.1 PHASE 1 RESULTS

The results from phase one were largely calculations of the composition of the 1000g data.

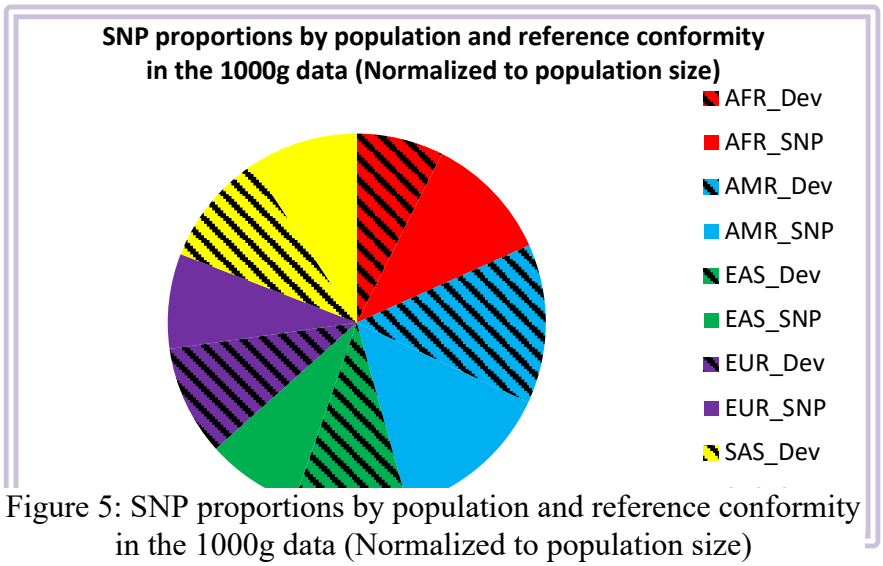


Figure 5: SNP proportions by population and reference conformity in the 1000g data (Normalized to population size)

The initial extraction, and subsequent analysis of bases which deviate from the consensus to a higher degree than the reference are shown in table 4 and figure 5.

Population	Total Individuals	Total Variants	Deviant bases	Total ALTs (-REF)
AFR	661	40910992	3364047	2953423599
AMR	347	27237493	3240789	1298710315
EAS	504	23078028	3341964	1840017336
EUR	503	23339891	3233930	1836586497
SAS	489	25974082	3277549	1820512807

Table 12: Table showing the number of individs, and SNP counts for each population, deviant base count, and Total number of SNPs occurring across all Individuals. (Total SNPs across all populations without duplicates)

As can be seen, there are a relatively high proportion of 'deviant bases' per population. When represented as a proportion of total variants the results are as follows: AFR – 8.222844; AMR – 11.89826; EAS – 14.48115; EUR – 13.85581; SAS - 12.61854. This result shows that, due to the high occurrence and the fact that these variants are largely population specific, rather than the utilization of a 'one size fits all' reference sequence, it would be beneficial to establish population specific references, particularly for super populations, if not to the sub-population level.

When the data for each population was consolidated, it provided a quantification of

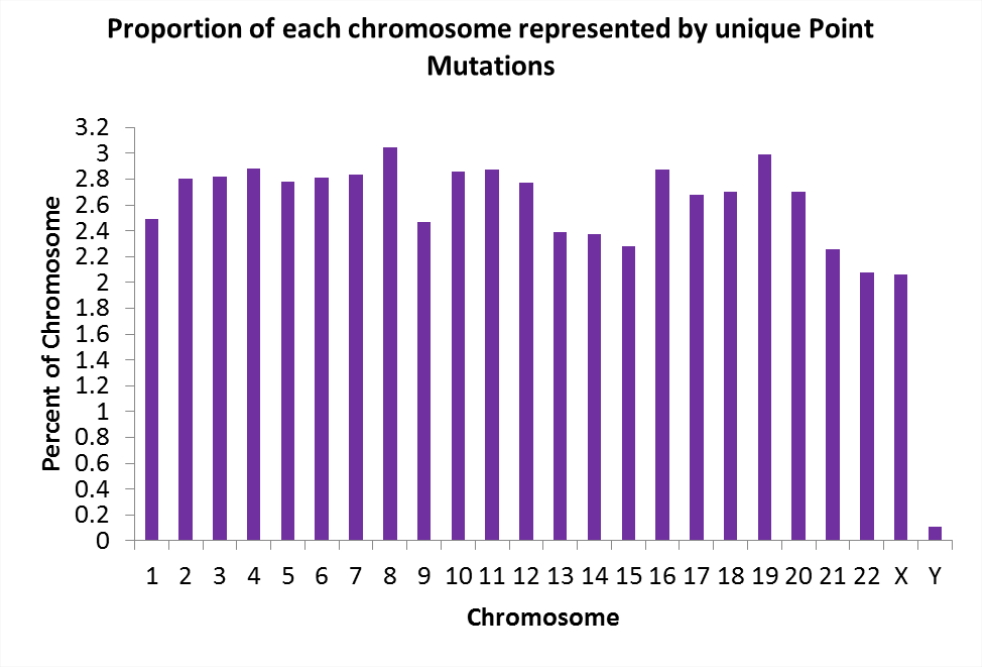


Figure 6:Proportion of each chromosome represented by unique point mutations

total unique variants per chromosome, as shown in table 5, and the relative proportion of each chromosome made up of SNPs (Fig 6).

Table 13: Number of unique point mutations per chromosome

Chromosome	Unique SNPs
1	6197318
2	6787257
3	5586665
4	5481942
5	5040053
6	4801408
7	4519682
8	4421880
9	3417598
10	3826403
11	3880403
12	3698416
13	2728216
14	2540328
15	2321569
16	2599165
17	2227931
18	2172105
19	1753777
20	1739591
21	1055458
22	1056332
X & Y	3276242

Unfortunately the autosomal data shown in figure 6 is not truly representative of the biological picture. This is due to the manner in which the variants for the allosomes were called within the 1000g data (see discussion).

5.2 PHASE 2 RESULTS

Metrics for Ohnolog pair data are shown in Table 6. Losses were incurred if, when cross referenced with gene data from Ensembl there was no matching location information. However this represents a very small quantity of the total genes, and may be rectified manually at a later date.

As can be seen in table 7, the quantities of unique genes are not, double that of the pairs. This is due to the fact that a proportion of the genes in each dataset belong to more than one pair set.

Table 14: Gene quantities within the Ohnolog pairs data files.

Criteria	Pairs	Unique Genes	Losses (no loc info)
Strict	2695	3544	8
Intermediate	4827	5504	14
Relaxed	8178	7831	19

As can be seen in table 7, the proportion of ohnologous genes in each chromosome is by no means an equal distribution, with strict criteria ohnologs per chromosome ranging from 355 in chromosome 1 to 17 in chromosome 21.

The following table (Table 8) Table 15: Ohnologous genes per chromosome

Chromosome	Strict	Intermediate	Relaxed

represents the quantification of data output by the VarCC2.pl program. It shows the quantities of each variant zygosity for each within each chromosome per population. As can be seen, on average the highest quantity of SNPs in each chromosome are Homozygous, with only a small number of instances per population which deviate from this.

1	355	569	782
2	247	373	517
3	171	292	434
4	177	240	322
5	260	344	430
6	106	223	337
7	134	229	351
8	152	231	306
9	136	198	273
10	179	256	325
11	136	242	401
12	191	319	457
13	68	103	150
14	146	177	235
15	108	195	254
16	141	194	273
17	214	317	468
18	53	85	122
19	224	330	511
20	131	176	243
21	17	27	60
22	62	119	182

Table 16: Zygosity counts per chromosome by population

Zygosity counts (VarCC2)				
AFR Chr	**Hetero strand 1**	**Hetero strand2**	**Homozygous**	**Hemizygous**
1	73575260	73363911	74246013	0
2	78994981	78892446	79875671	0
3	67264272	67285172	69335714	0
4	69723545	69675855	71795718	0
5	60115413	60113803	59468067	0
6	61452194	61580278	64266700	0
7	55931441	55929417	57827460	0
8	53201222	53358766	53833772	0
9	40897534	40946925	40622648	0
10	47418297	47506021	47240352	0
11	46913236	46834679	48318978	0
12	44692848	44501014	44570954	0
13	34096498	34137215	35016683	0
14	30548962	30517972	31829039	0
15	27927264	27910467	27953326	0
16	30176630	30163840	30664706	0
17	26102895	26101877	26369589	0
18	26963589	26954591	27032753	0
19	22576620	22613532	23452004	0
20	21112894	21221305	20937075	0
21	14141482	14109234	15408010	0
22	13175505	13184440	13773592	0
X	20090375	20156798	20602609	34466796
Y	0	0	0	362855
AMR Chr	**Hetero strand 1**	**Hetero strand2**	**Homozygous**	**Hemizygous**
1	29820054	29861685	39510757	0
2	31646936	31617395	41513055	0
3	27414298	27455761	34443142	0
4	28285220	28394733	38650205	0
5	24221412	24176232	28586496	0
6	25863540	25990161	31670980	0
7	22953237	22848551	27722978	0
8	21141754	21246771	26137751	0
9	16591567	16507402	20835549	0
10	19711088	19619533	24977358	0
11	18974813	19007088	26991685	0
12	18246649	18404678	23931575	0
13	14089166	14071717	20597952	0
14	12430483	12403472	16335684	0
15	11122736	11138454	14846824	0
16	11993672	11925600	14938512	0
17	10496178	10502469	12746760	0
18	10930731	11025631	14896615	0
19	9028460	9074262	10132362	0
20	8481763	8489826	9672395	0
21	5671713	5656597	7327607	0
22	5349895	5350878	6192052	0
X	7027240	7027241	8854258	13821006
Y	0	0	0	88015
EAS Chr	**Hetero strand 1**	**Hetero strand2**	**Homozygous**	**Hemizygous**
1	39579336	39548291	62873849	0
2	41131645	41103185	65702254	0
3	35719920	35794092	55460005	0
4	37373692	37459723	60705997	0
5	31630234	31631003	42770294	0
6	34448666	34420468	51014114	0

7	30365769	30337637	43923138	0
8	28029949	28073935	40231362	0
9	22109627	22061027	32808225	0
10	25871564	25919684	39556541	0
11	25139959	25264011	41079179	0
12	24286186	24300637	38116440	0
13	18456084	18496852	32139911	0
14	16830028	16756474	25013889	0
15	14933540	14967361	22946249	0
16	15946604	15969052	23341073	0
17	13542433	13529510	19734451	0
18	14652943	14628679	23406887	0
19	11967115	11960622	15093213	0
20	11016893	11093919	15809293	0
21	7702223	7724434	11027030	0
22	6993693	7011335	10021734	0
X	8766703	8836004	14281922	19380295
Y	0	0	0	197250
EUR Chr	**Hetero strand 1**	**Hetero strand2**	**Homozygous**	**Hemizygous**
1	41770256	41691910	56374537	0
2	44788760	44867722	58675226	0
3	38490743	38449813	48440891	0
4	39949415	39961813	55163146	0
5	34080004	34214037	41202043	0
6	36937499	36833274	43729944	0
7	32303157	32350921	38827968	0
8	29762234	29806852	37156641	0
9	23433746	23357428	29299186	0
10	27607407	27670896	36128122	0
11	27201986	27193211	38770969	0
12	25898312	25992932	34254239	0
13	19796593	19900427	30072886	0
14	17514766	17416577	23294206	0
15	15553753	15569552	21413596	0
16	16903847	16903691	21255817	0
17	14910765	14886818	18303408	0
18	15367841	15432926	21453072	0
19	12791655	12785467	14293214	0
20	11969753	11986282	13520571	0
21	8042386	8046948	10426363	0
22	7663693	7653725	8495949	0
X	9664847	9670346	12602344	18182855
Y	0	0	0	202318
SAS Chr	**Heterozygous strand 1**	**Heterozygous strand2**	**Homozygous**	**Hemizygous**
1	41552872	41634426	55660328	0
2	44207010	44010419	58381442	0
3	37874733	37843725	48702202	0
4	39614751	39598341	54298024	0
5	33572362	33641831	40601201	0
6	36185569	36017254	44774967	0
7	30918479	30999625	39824433	0
8	29818839	29882724	36660799	0
9	23349905	23379067	29175967	0
10	27473554	27439554	35521213	0
11	26606895	26531032	38026471	0
12	25569377	25580235	34151779	0
13	19587070	19585394	29463936	0
14	17572352	17641603	23099419	0
15	15722768	15680747	20654648	0
16	16734962	16842314	21262426	0
17	14511935	14538812	18014464	0

18	15179317	15194863	21088062	0
19	12798701	12823497	14381556	0
20	11772258	11771818	13813855	0
21	8010273	8012110	10448950	0
22	7585928	7639586	8754929	0
X	8831060	8883869	11617415	20594684
Y	0	0	0	178008

Table 9 shows the quantities of polymorphisms within each population per chromosome which occur within, and outside of ohnologs. This data is a quantification of the individual SNPs as separated by Ohnolog_(non_O)SNPs.pl. As would be expected, given the size of the genomic coverage of each, the majority of SNPs fall outside of ohnologous genes. Whilst it may be an artefact of differing population sizes in the sampled data, there are clear differences in the number of ohnolog SNPs per chromosome per population.

Table 17: Total Ohnolog and non ohnolog SNPs.

Pop	Chrom	Total SNPs	Ohnolog SNPs	Non ohnolog SNPs	Pop	Chrom	Total SNPs	Ohnolog SNPs	Non ohnolog SNPs
AFR	1	3103738	516341	2581113	AFR	13	1383199	192896	702725
AMR	1	2074870	341704	1729464	AMR	13	928515	127039	470080
EAS	1	1765661	294441	1469931	EAS	13	789140	108885	393494
EUR	1	1789640	297555	1490225	EUR	13	802244	108718	395542
SAS	1	1974004	328044	1643599	SAS	13	893223	123397	441507
AFR	2	3383829	430941	1659309	AFR	14	1284306	234860	516865
AMR	2	2246595	283188	1117883	AMR	14	856969	156330	344737
EAS	2	1908495	243548	939317	EAS	14	727076	133091	290544
EUR	2	1916542	241879	956315	EUR	14	729865	131842	296895
SAS	2	2149258	273645	1061514	SAS	14	819265	150385	325842
AFR	3	2791417	397408	1797293	AFR	15	1173780	190469	477487
AMR	3	1882268	266005	1203028	AMR	15	761220	121119	318592
EAS	3	1589016	224930	1023122	EAS	15	666046	108778	272901
EUR	3	1602625	227656	1030425	EUR	15	660116	107048	276282
SAS	3	1786553	253473	1146240	SAS	15	741253	119883	307272
AFR	4	2791050	431879	1605350	AFR	16	1300273	209082	2387390
AMR	4	1875246	288319	1079403	AMR	16	856287	137803	1612268
EAS	4	1566297	243859	911477	EAS	16	730794	115963	1362604
EUR	4	1602163	246483	923597	EUR	16	739391	119223	1372865
SAS	4	1781849	276268	1029407	SAS	16	823279	132108	1530593
AFR	5	2538079	402156	1186688	AFR	17	1121199	169895	2354208
AMR	5	1683594	261663	799271	AMR	17	729283	108976	1584147
EAS	5	1399884	223365	679383	EAS	17	625583	94904	1322006
EUR	5	1429827	230227	692292	EUR	17	638548	96711	1354572
SAS	5	1585255	248329	768370	SAS	17	703439	106851	1504186
AFR	6	2458725	183036	1050442	AFR	18	1101569	107828	2187695

AMR	6	1655155	119839	701415	AMR	18	744438	72543	1454335
EAS	6	1431254	103108	595525	EAS	18	627692	59419	1207785
EUR	6	1436389	103519	599114	EUR	18	625998	60052	1234459
SAS	6	1582216	115672	670306	SAS	18	702911	67231	1368865
AFR	7	2303174	239277	982108	AFR	19	894853	123114	2268523
AMR	7	1535593	156359	639512	AMR	19	605622	80432	1530828
EAS	7	1285788	133256	557552	EAS	19	511140	70928	1325876
EUR	7	1316436	134827	552964	EUR	19	518449	70722	1330027
SAS	7	1446896	149213	621303	SAS	19	576946	79388	1463210
AFR	8	2242265	351141	1087751	AFR	20	873993	169230	2057920
AMR	8	1478097	230287	716553	AMR	20	585722	114492	1375756
EAS	8	1231401	195552	614213	EAS	20	491483	97543	1151231
EUR	8	1251599	194674	619138	EUR	20	493238	97095	1179699
SAS	8	1396760	218035	689961	SAS	20	551809	109532	1295206
AFR	9	1727004	282137	951558	AFR	21	546040	27327	1885583
AMR	9	1136790	185174	620610	AMR	21	363625	17765	1244631
EAS	9	977353	158941	531680	EAS	21	305721	14705	1034678
EUR	9	982881	160044	542625	EUR	21	312524	14953	1055234
SAS	9	1097594	178193	597304	SAS	21	342569	15996	1176145
AFR	10	1955912	293322	990706	AFR	22	540844	61544	1441199
AMR	10	1314550	194717	670098	AMR	22	359704	40095	949568
EAS	10	1106260	166837	567517	EAS	22	308983	35619	817822
EUR	10	1124818	167759	564927	EUR	22	312116	35171	821785
SAS	10	1249420	186995	634478	SAS	22	348050	40052	918066
AFR	11	1953612	150619	771907	AFR	X	1557532	159210	1394385
AMR	11	1303999	97748	525481	AMR	X	994480	101284	891422
EAS	11	1108528	84051	441231	EAS	X	851930	86994	764065
EUR	11	1116458	84246	448537	EUR	X	845270	84880	759322
SAS	11	1243920	95495	498328	SAS	X	965672	97496	866671
AFR	12	1865400	255638	2945293	AFR	Y	19199	2	19200
AMR	12	1251444	169395	1959115	AMR	Y	13427	2	13428
EAS	12	1057270	144749	1663567	EAS	Y	15233	2	15234
EUR	12	1069068	144043	1672404	EUR	Y	23686	2	11844
SAS	12	1193395	162252	1873068	SAS	Y	18545	2	18546

Table 18: average losses per population per region (O- ohnolog, N- non-ohnolog)

POP	Strict	Intermediate	Relaxed
AFR O	0.014522	0.015271717	0.0137892
AMR O	0.017374	0.027119723	0.0238672
EAS O	0.02009	0.03071027	0.0271482
EUR O	0.019584	0.029788996	0.0263481
SAS O	0.017559	0.026753076	0.023494
AFR N	0.044395	0.044910012	0.0474822
AMR N	0.060726	0.061369256	0.064883
EAS N	0.067563	0.06830249	0.0722284
EUR N	0.068155	0.069073671	0.0730256
SAS N	0.060747	0.061570769	0.0652344

Due to differences in assembly between the HG37 and HG38 reference builds, when the SNPs were converted between the two losses were incurred. For each criteria the losses were as follows: Strict, ohnolog losses – Min % = 0 Max % = 0.22772766; Intermediate, ohnolog losses– Min % = 0 Max % = 0.173340991; Relaxed, ohnolog losses– Min % = 0 Max % = 0.141037442; Strict non-ohnolog losses – Min % = 0.00095992 Max % = 0.33746498; Intermediate non-ohnolog losses– Min % = 0.050363877 Max % = 0.344364976; Relaxed non-ohnolog losses– Min % = 0.051384965 Max % = 0.365478514. Table 10 shows the maximum losses per population, per stringency criteria (full raw data can be seen in appendix) .

5.3 PHASE 3 RESULTS

The original family source file, obtained from http://ohnologs.curie.fr following assignment of gene name data from ensemble, and division into the three location criteria, contained the information as shown in table 11.

Table 19: Genes contained in the ohnolog families source data

Category	Strict	Intermediate	Relaxed
Source file families	1381	2024	2642
Contained genes	3274	4994	6795
Maximum genes in families	6	11	24
Retained families	1374	1994	2623
Retained genes	3259	4928	6755
Maximum retained genes in families	6	11	24

This phase resulted in the loss of between 4 and 18 families across the stringency criteria. This was due, not to a lack of stable Ids but to complicated patterns of inheritance between duplicates of ancient and recent origin.

Table 20: SNPs per criteria per chromosome

Chrom	Strict	Intermediate	Relaxed	Chrom	Strict	Intermediate	Relaxed
1	1037947	1223071	1557974	13	311982	388027	474185
2	565327	719405	1020032	14	53331	119046	235965
3	302496	610619	954014	15	125162	228229	296043
4	461982	664159	952396	16	782540	1033909	1458576
5	360058	452418	523703	17	804535	1000245	1181723
6	444337	475683	627377	18	632404	871189	1085129
7	372063	637248	796149	19	361741	661468	848449
8	339435	404869	504782	20	477103	898815	1197664
9	327792	452575	587636	21	688088	961677	1157947
10	187254	377467	508984	22	544521	688251	849115
11	227748	308174	383769	X	275486	399769	526126
12	830415	1312965	1676837	Y	0	92	92

Table 12 shows the number of SNPs found to be contained within each criteria per chromosome.

The following tables (13-14) and figure 5 show how these SNPs were distributed across the five super populations per criteria, per chromosome.

Table 21: SNP distribution per population per criteria

Stringency criteria	Chromosome	AFR SNPs	AMR SNPs	EAS SNPs	EUR SNPs	SAS SNPs
Strict	1	510208	337654	290971	294180	324321
Strict	2	407279	267691	229902	228440	258372
Strict	3	388896	260275	219990	222804	247938
Strict	4	404557	270242	228781	231067	259447
Strict	5	311756	204212	173013	174924	194610
Strict	6	180448	118064	101569	101983	113991
Strict	7	237630	155297	132392	133972	148436
Strict	8	348530	228365	194232	193180	216872
Strict	9	273101	179410	153841	154648	172348
Strict	10	283510	188485	161083	162274	180925
Strict	11	146912	95410	82123	82159	93082
Strict	12	229690	151966	130057	129272	145821
Strict	13	179564	118453	101051	101027	114577
Strict	14	220808	147441	125266	124470	141536
Strict	15	184919	117699	105793	104042	116491
Strict	16	169622	111757	94413	96685	107297
Strict	17	161082	103396	90000	91860	101245

Strict		18	94090	63449	51984	52672	58885
Strict		19	111982	73210	64450	64245	72188
Strict		20	153051	103319	88473	87367	98764
Strict		21	27327	17765	14705	14953	15996
Strict		22	61256	39915	35449	35021	39897
Strict	X		131496	83108	71770	70246	79893
Strict	Y		0	0	0	0	0
TOTAL			5217714	3436583	2941308	2951491	3302931
Intermediate		1	599637	395782	342427	344578	382216
Intermediate		2	647464	426497	366411	365732	411601
Intermediate		3	510109	338890	289490	292174	325231
Intermediate		4	502346	333870	284082	285239	321302
Intermediate		5	429681	281393	238705	241013	268846
Intermediate		6	329981	216959	186915	186471	208139
Intermediate		7	453593	297367	250927	255633	283709
Intermediate		8	485794	318972	270625	271011	303765
Intermediate		9	343645	226086	195582	195973	218374
Intermediate		10	363010	241810	206012	206344	231249
Intermediate		11	300147	196923	167506	169582	190307
Intermediate		12	328046	216053	186641	186123	210100
Intermediate		13	226609	150125	128422	128589	144096
Intermediate		14	236354	157761	134006	133645	151666
Intermediate		15	317485	203234	181580	179275	201303
Intermediate		16	203094	133286	113009	114835	128050
Intermediate		17	223653	144398	125983	127742	141094
Intermediate		18	189644	128076	107555	108040	121674
Intermediate		19	152250	99925	86732	87509	97876
Intermediate		20	190392	128130	109827	108735	122517
Intermediate		21	60781	39246	32941	34045	36499
Intermediate		22	114055	75343	64589	65374	73086
Intermediate	X		189746	119441	102378	101155	115432
Intermediate	Y		32	28	23	50	33
TOTAL			7397548	4869595	4172368	4188867	4688164
Relaxed		1	765259	505036	437448	439569	488365
Relaxed		2	828322	546349	467346	468134	525391
Relaxed		3	720849	480627	410463	412547	461808
Relaxed		4	592277	392593	333906	336513	377799
Relaxed		5	536963	352361	297909	301339	335646
Relaxed		6	423716	278789	240905	240444	267634
Relaxed		7	603170	395918	335703	340152	377414
Relaxed		8	584908	383228	326196	327158	365993
Relaxed		9	424296	278671	241254	241276	269659
Relaxed		10	515025	344211	292356	294900	329250
Relaxed		11	470729	311056	263572	267225	299656
Relaxed		12	471167	311772	268050	267083	302117

Relaxed		13	263507	174843	150002	150774	168447
Relaxed		14	310601	207848	176437	175245	200597
Relaxed		15	397608	254732	226748	223573	251979
Relaxed		16	251378	165258	141026	142720	159373
Relaxed		17	291137	188121	163767	165949	183271
Relaxed		18	255322	171371	144478	144615	162890
Relaxed		19	190000	125445	108095	109692	122677
Relaxed		20	232404	156101	133554	132396	149464
Relaxed		21	121212	79743	67272	68970	75163
Relaxed		22	147445	96784	83421	84032	94441
Relaxed	X		248043	155499	133773	132619	151228
Relaxed	Y		32	28	23	50	33
TOTAL			9645370	6356384	5443704	5466975	6120294

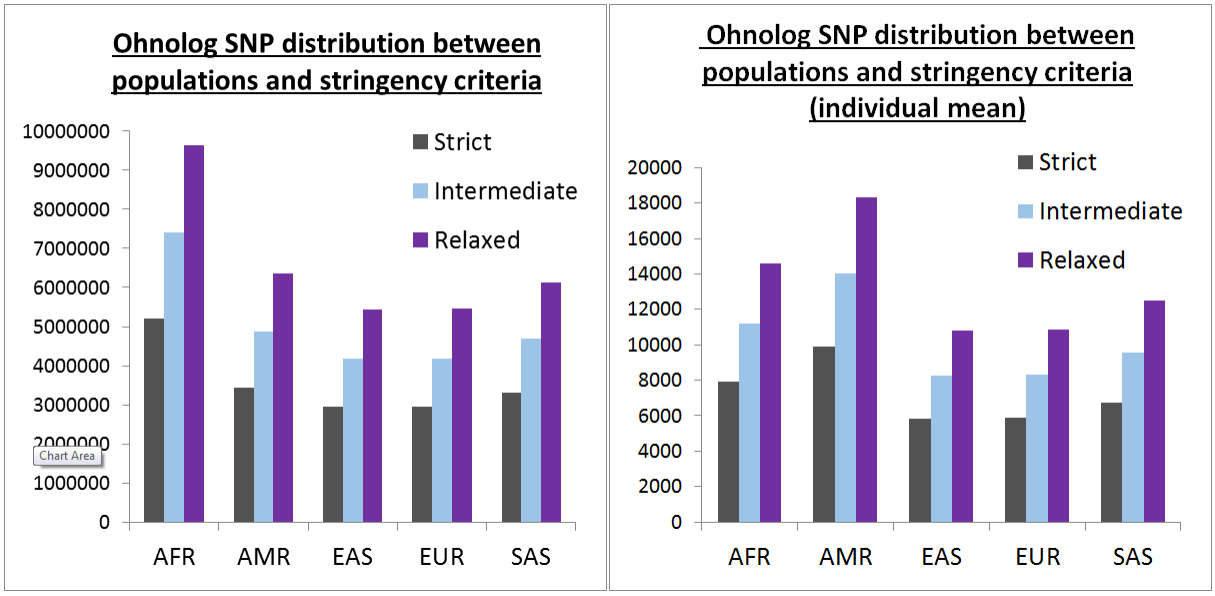


Figure 32: Ohnolog SNP distributions between population and criterion

Table 22: Ohnolog SNP distribution between populations and stringency criteria

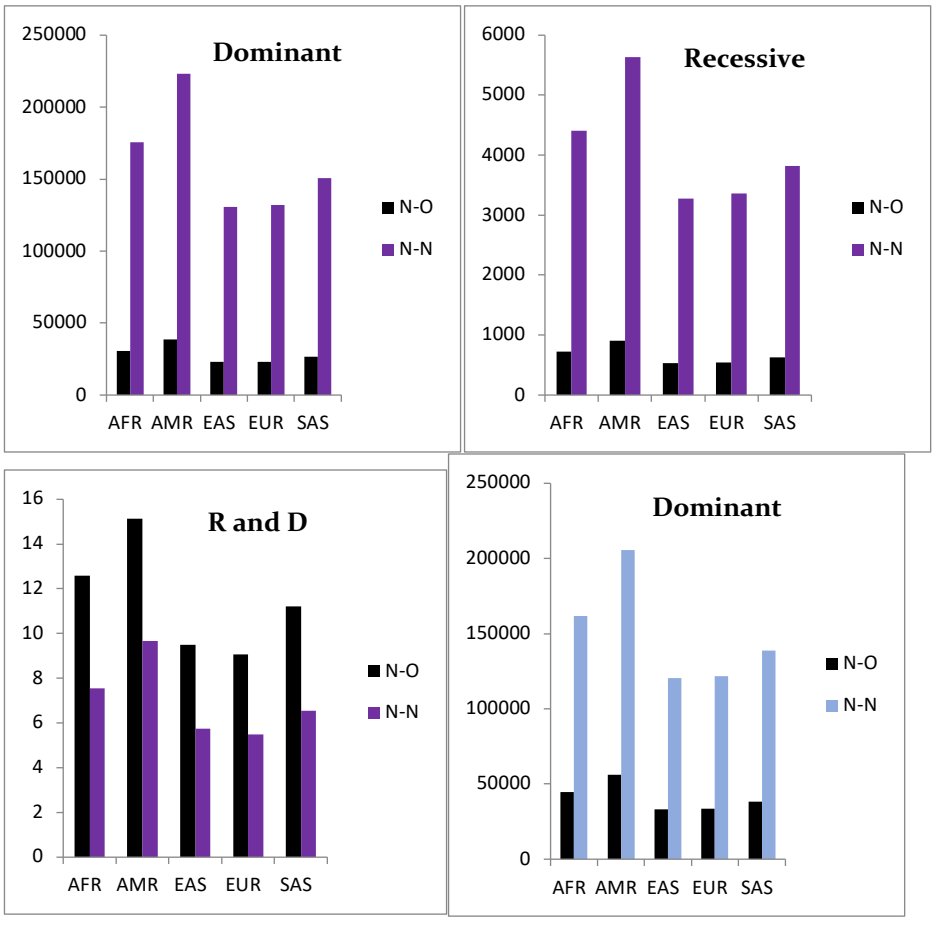
Ohnolog SNP distribution between populations and stringency criteria					
Criteria	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**
Strict	7893.667	9903.697	5835.929	5867.775	6754.46
Intermediate	11191.45	14033.41	8278.508	8327.767	9587.247
Relaxed	14592.09	18318.11	10801	10868.74	12515.94

5.4 PHASE 4 RESULTS

When divided into the differing inheritance types the Ohnolog.Families.$criteria.2R.txt file was found to contain 1348 genes specifically

associated with dominant disorders, 1747 recessive associated genes, and 400 genes which have both dominant and recessive associations.

When cross referenced with the converted HG38 SNP data the distributions of SNPs it was found that there were a quantity of ohnolog associated SNPs within disease associated genes as shown in figure 6. As can be seen, genes with associations with both dominant ad recessive diseases contain the greater variations within ohnologs than outside. This is interesting, as it suggests a greater complexity of associations with complicated inheritance types within ohnologs than occurs outside. Given that the data represented here pertains to healthy populations the findings of non-anecdotal proportions of variation within genes with associations with dominant disorders is surprising, particularly within ohnologs. These results suggest that there is an underlying mechanism by which healthy individuals are able to tolerate such variation.



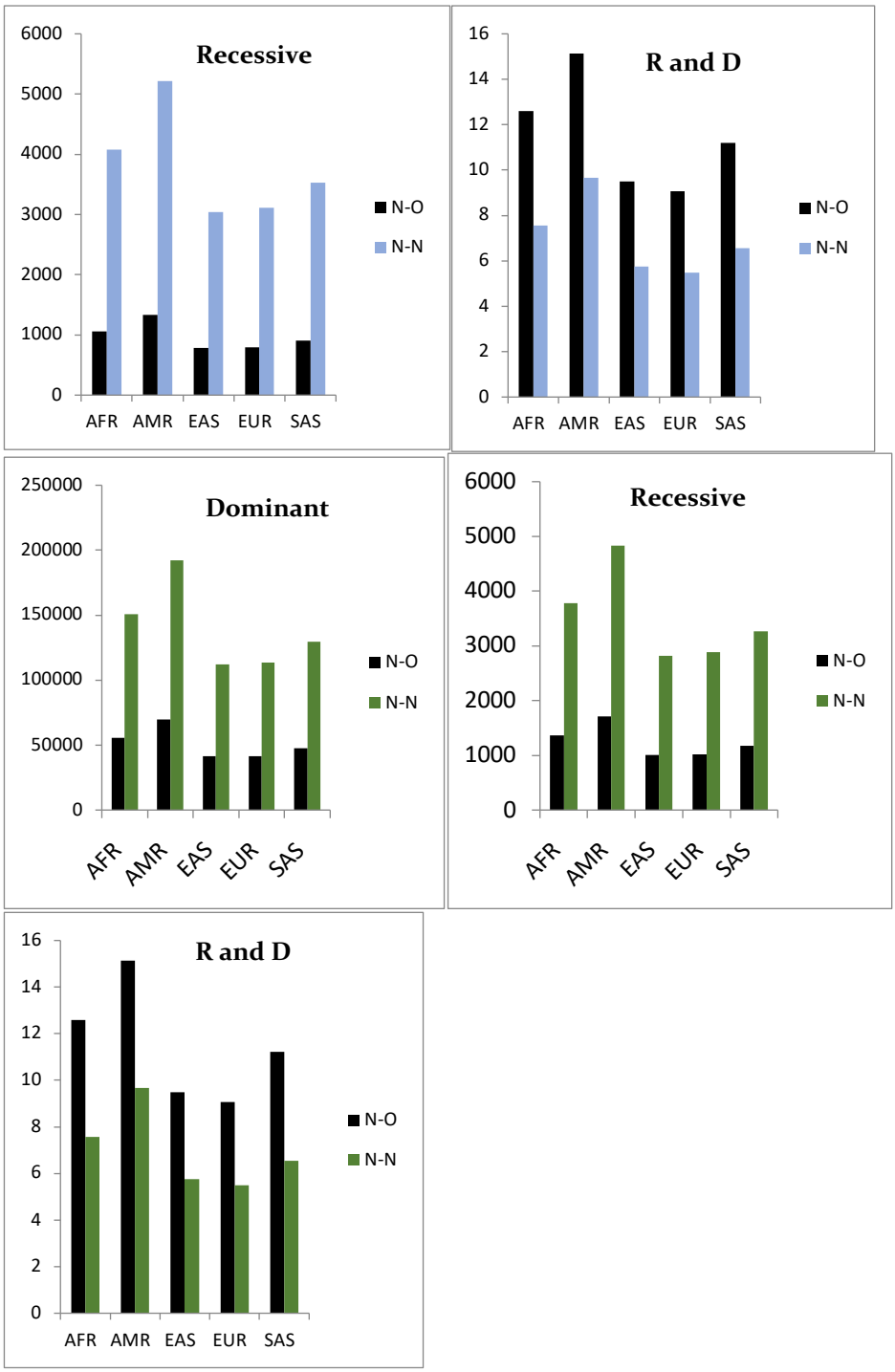


Figure 33: SNP distribution by Inheritance type and criteria (Strict in purple, Intermediate in blue and Relaxed in green)

6. DISCUSSION

In support of the findings of the UK10K the nature of human diversity is such that it is unlikely that we will be able to establish a 'one size fits all' Human profile (UK10K 2015), given the high proportions of nucleotides found in frequencies in excess of the reference in each population, it is clear that population specific variatation is highly contributory to the overall variant calls within the 1000g data, and therefore population specific reference genomes are required. Firstly this will reduce any biases introduced when quantifying variation, as these 'deviant bases' are not, in fact true variants in the classical sense, but are illustrative of the fact that the reference was built from the genomes of individuals from only one geographical location. This can be clearly seen when comparing the quantities of deviant bases in each population, as the lowest frequency (AMR – 1298710315) is in the population to which the reference individuals were native. It would be expected, however, that given the high frequencies at which these 'deviant bases' occur, it is likely that they will be synonymous. This in itself is interesting, as, by mapping the loci of commonly occurring population specific variation, we may be able to observe patterns of highly tolerated variation in healthy populations, however, the introduced bias towards European, and, in this case particularly, American populations needs to be borne in mind when analysing such data.

One of the major issues encountered when processing the 1000g data has been the manner in which variants have been called in the allosomes. For female individuals a diploid alignment for the two copies of X was aligned, in the same manner as was done for the autosomes. For the Male participants however, the portions of the Y chromosome matching X have been mapped as diploid, however, for those regions which are not shared between the two chromosomes this has been treated as haploid. It is therefore the case that the data for the X and Y chromosomes are not directly comparable to the autosomes unless pooled, as they are not a complete view of the specific chromosomes they represent. To pool the data for allosomes however, is undesirable, as this masks sex specific variation, which, given the unusual evolutionary history, and relative youthfulness of the Y chromosome, would be liable to yield interesting patterns of variation when compared with the other chromosomes. .

The most surprising result to have emerged from the research thus far, is the evidence of relatively high quantities of SNPs within ohnologous genes with known dominant disease associations in healthy individuals. Whilst this must be investigated further, particularly with regards to identifying which, if any, of these are likely to be non-synonymous, and the functional ramifications of which, it is particularly surprising given the previously identified strong relationship between ohnologs and disease (McLysaght et al 2014). It is hypothesized that, due to evolutionary constraints and underlying dependencies, the ability to tolerate such variation may be due to a composite of heterozygous variants in gene pairs compensating for dosage imbalances, which is supportive of the initial research hypothesis. Whilst it is clearly not an immediate concern, this method of compensation within the genome may become an issue as the ohnologous gene pairs continue to evolve. This is due to the fact that it has been observed that whilst ancestral ohnologous genes have high degrees of conservation, the more recent copies are more prone to the accumulation of variation. As this asymmetric evolution continues, these genes will become more functionally divergent, to the point where the function of each gene in the pair will become so disparate, that one will no longer be capable of providing dosage compensation for the other.

Much still needs to be done in order to resolve any patterns, it is clear from the initial data emerging from this research, that propensities towards the tolerance of variation in healthy individuals, may be observed. Whilst it is unlikely to be possible to plot these variations to a global reference, through the understanding of population specific variation, alongside patterns of variation in genes with differing evolutionary histories, we are gaining insight into patterns of normal variation. These findings are likely to provide a basis for comparison and prediction of disease causing variation within the human genome, and, within individual genotypes.

7. FUTURE WORK

The future analysis will be divided into a series of 'aims', for which the results will be linked. Thus far this approach has been taken in the phases detailed in this report, focusing on variants within ohnologous genes. This is by no means complete; there are a number of outstanding analyses to be conducted on this gene set. Following completion of the ohnolog analysis, the research will extend to incorporate further aims of evolutionary importance. This will include different types of CNVs, inclusive of singly duplicated genes and large structural duplications, such as segmental duplications, alongside genes without duplicates. Further to this, it is intended to complete the research by incorporating information pertaining to sources outside the genome which may act upon it, such as sites of phosphorylation and methylation.

7.1 AIM 1

The immediate steps to be taken in order to complete the ohnolog module are as follows; the primary immediate concern will be the production of phylogenetic trees for all genes. This will be achieved by creating alignments of each of the two copies per individual, to be treated as 'parental' sequences. It will be necessary at this point to account for recombination, this will be done using software, of which a number of programs are available, for example, Splitstrees. Following this, trees will be produced for the resultant 5008 genes representing the phylogenies of each gene whilst accounting for heterozygous variations. It expected that individuals within each population will cluster together, whilst differential branch lengths between trees for gene pairs will show a greater representation of asymmetrical evolution between ancestral and duplicated ohnologous genes. This will then be followed by establishing trees for these genes on a wider scale incorporating species with wider evolutionary divergence, inclusive of the ancient hominids, in order to track patterns of divergence of duplicates over larger timescales.

Following this it is intended to investigate co-occurrence of heterozygous mutations across ohnolog pairs and families. It is hypothesized that the presence of variation within disease associated genes in healthy individuals is permitted due to a process of compound heterozygosity. It will be necessary, in order to test this to define specifically disease causing individual variants in disease associated ohnologs. This

will be done by cross referencing the current data obtained from phase four, with information on known disease causing variants from ClinVar and the HGVD, alongside the 1000g project's predicted consequence data. Each pair and family set will then be tested to establish heterozygous SNPs which occur together with relative frequency between genes, and the likely outcome of the proposed consequences of these combinations of variants on dosage (hereafter referred to as compound heterozygosity). We hypothesize that the expected 'norm' would be a compensatory process wherein the older, and more highly constrained gene copies, which are known to contain fewer variants are likely to compensate for asymmetrical patterns of potentially disease causing polymorphisms in the 'daughter' copies, which are more permissive of variation and are therefore more inclined to contain deleterious polymorphisms.

The final stage of the ohnolog analysis will be to divide the 1000g data into subsets; protein coding genes, and genes coding for regulatory factors and RNA, which act upon each of the genes in the ohnolog data, to establish relationships between the abstract interactions found in ohnologous pairs and their regulation. Once these steps have been completed it is hoped that they will assist in resolving the relationships between co-occurrence of heterozygous mutations within ohnolog families, and their presence in disease associated genes. Completion of this module is expected by November 2016.

7.2 AIM 2

The second research 'aim' will incorporate a new gene set. In this instance I will be looking at structural variants known as segmental duplications, available via the Database of Genomic Variants (DGV). These are large scale duplications which occur both inter and intrachromosomally, often with close relationships with pericentromeric, and subtelomeric regions. Given that segmental duplications are not distributed evenly across each chromosome, and that there are clear differences in quantities of duplicated genes within them, alongside the fact that there is a heightened propensity towards disease genes in this type of structural variation, their inclusion in this research is imperative. The analysis of segmental duplications will follow a similar structure to that of ohnologous genes, with the core difference being

that in place of the family element, I will look at gene duplicates within paired segmentally duplicated regions. Completion of this module is expected by mid 2017.

7.3 AIM 3

To compliment the first two aims, the third will focus on singleton duplications across the genome. Again the analysis will be conducted in much the same way as the first two modules. This will then be cross referenced with the findings from module one and two in order to further refine our understanding of the role of CNVs with differing evolutionary pressures, in conjunction with small-scale variation in a normal human genome.

7.4 AIM 4

Following this, aim four will expand to incorporate the remaining types of variation identified by the 1000g project. These include insertion/deletions and structural variations. These will be overlaid on the findings from the previous three modules in order to define any potential patterns in their accumulation and regulation, the completion of which is expected by late 2017.

7.5 AIM 5

The fifth proposed aim will take the profiles of variation in normal populations as provided by modules one to four, it will then overlay this information on new genomic data pertaining to 'non-normal' populations. Currently a number of disease datasets are available, however they are predominantly anonymised in such a way that it is not possible to link variations within an individual (as is the case, for example with ExAc). There are however a collection of smaller datasets already publicly available, and proposed large scale sequencing projects which aim to provide such information due to be published, for example the 100,000 genome project, due to be released in late 2017. This key module will allow an analysis of the differences between healthy and diseased populations, providing an insight into non conformity between the two sets which will aid not only in the identification and understanding of population specific disease evolution, but also understanding disease in the context of an individual's genotype, This is liable to be far more time consuming than the previous modules, as, in order to understand the various

functional aspects of disease variants, it will be necessary to divide this data into disease subsets and treat each disease type accordingly, by geographical super population. It is therefore expected that this module will not be completed in its entirety until the end of 2018.

7.6 AIM 6

The final aim, which will be included, providing time constraints have not been encountered, will be the overlaying of epigenetic data onto the findings from modules one to five. There currently are a number of sources of epigenetic information available, such as the epigenetic roadmap, and identified sites of phosphorylation within the 1000g project data which will inform this analysis, however, it is anticipated that between 2016 and 2018 greater and more refined sources of epigenetic information will become available. This final module, it is proposed, will aid in our understanding of the external pressures acting on the genome and ascertain whether there are patterns of association between these sites and both normal and non-normal variation, and disease.

8. REFERENCES

- 1000 Genomes Project Consortium., Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E., Mort, M., Phillips, A., Shaw, K., Stenson, P., Cooper, D., Tyler-Smith, C.,(2012). 'Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing' *Am. J. Hum. Genet.* 91, 1022–1032
- Abeysinghe, S., Chuzhanova, N., Krawczak, M., Ball, E., Cooper, D. (2003). 'Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs' *Human mutation* 22(3), 229-44.
- Barabási, A.,Gulbahce, N., Loscalzo, J. (2011). 'Network medicine: a network-based approach to human disease' *Nature Reviews Genetics* 12(1), 56-68.
- Birney, E, Soranzo. N. (2015). 'Human genomics: The end of the start for population sequencing' *Nature* 526(7571), 52-53.
- Bishara, A., Liu, Y., Kashef-Haghighi, D., Weng, Z., Newburger, D., West, R., Sidow, A., Batzoglou. S. (2015). 'Read Clouds Uncover Variation in Complex Regions of the Human Genome' *Research in Computational Molecular Biology*, 9029, 30-31.
- Blomen, V., Májek, P., Jae, L., Bigenzahn, J., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F., Olk, N., Stukalov,A., Marceau, C., Janssen, H., Carette, J Bennett, K., Colinge, J., Superti-Furga, G., Brummelkamp, T. (2015). 'Gene essentiality and synthetic lethality in haploid human cells' *Science express*, 350(6264), 1092-1096
- Cáceres, J., Kornblihtt, A. (2002), 'Alternative splicing: multiple control mechanisms and involvement in human disease' *Trends in Genetics* 18(4), 186-93.
- Cartegni, L., Chew, S., Krainer, A. (2002). 'Listening to silence and understanding nonsense: exonic mutations that affect splicing' *Nature Reviews Genetics* 3(4), 285-98.

- Cavalli-Sforza, L., Wilson, A., Cantor, C., Cook-Deegan, R., King, M. (1991). 'Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project' *Genomics* 11(2), 490-491.
- Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M., Sherry, S. (2012). 'The 1000 Genomes Project: data management and community access' *Nature methods* 9(5), 459-462.
- Dehal, P., Boore, J. (2005). 'Two rounds of whole genome duplication in the ancestral vertebrate' *PLoS biology* 3(10),1700.
- Dorschner, M., Amendola, L., Turner, E., Robertson, P., Shirts, B., Gallego, C., Bennett R., Kim, J. (2015). 'Actionable, pathogenic incidental findings in 1,000 participants' exomes.' *The American Journal of Human Genetics* 93(4), 631-640.
- Dudley, J., Kim, Y., Liu, L., Markov, G., Gerold, K., Chen, R., Butte, A., Kumar, S. (2012). 'Human genomic disease variants: a neutral evolutionary explanation', *Genome research,* 22(8): 1383-1394
- Ebert, G., Steininger, A., Weißmann, R., Boldt, V., Lind-Thomsen, A., Grune, J., Badelt, S., Heßler, M., Peiser, M., Hitzler, M., Jensen, L.R., Müller, I., Hu, H., Arndt, P., Kuss, A., Tebel, K., Ullmann, R. (2014). 'Distribution of Segmental duplications in the context of higher order chromatin organisation of human chromosome 7', *BMC Genomics*, 15(1): 537-554.
- Frazer, K., Murray, S., Schork, N., Topol, E. (2009). 'Human genetic variation and its contribution to complex traits' *Nature Reviews Genetics* 10(4), 241-251.
- Henn, B., Botigué, L., Bustamante, C., Clark, A., Gravel, S. (2015). 'Estimating the mutation load in human genomes' *Nature Reviews Genetics,* 16(6), 333–343
- Iafrate, J., Feuk, L., Rivera, M., Listewnik, M., Donahoe, P., Qi, Y., Scherer, S., Lee, C. (2004). 'Detection of large-scale variation in the human genome' *Nature genetics* 36(9), 949-951.
- Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., Cummings, B., Tukiainen, T. (2015)

'Analysis of protein-coding genetic variation in 60,706 humans' *bioRxiv* 030338.

- Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen,A., Andersen, G., Cao, H., Korneliussen, T., Grarup, N., Guo, Y., Hellman, I., Jin, X., Li, Q., Liu, J., Liu, X., Sparsø, T., Tang, M., Wu, H., Wu, R., Yu, C., Zheng, H., Astrup, A., Bolund, L., Holmkvist, J., Jørgensen, T., Kristiansen, K., Schmitz, O., Schwartz, T., Zhang, X., Li, R., Yang, H., Wang, J., Hansen, T., Pedersen, O., Nielsen, R., Wang, J. (2010). 'Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants', *Nature genetics*, 42(11): 969-972
- Lupski, J. (1998). 'Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits' *Trends in genetics 1*4(10), 417-22.
- Makino, T., McLysaght, A., Kawata, M. (2013). 'Genome-wide deserts for copy number variation in vertebrates.', *Nature communications*, 4:2283
- Marchini, J., Donnelly, P., Cardon, L. (2005) 'Genome-wide strategies for detecting multiple loci that influence complex diseases' *Nature genetics* 37(4), 413-417.
- Maurano, M., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., Stamatoyannopoulos, J. (2015). 'Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo' *Nature genetics* 47(12), 1393-1401
- McLysaght, A., Makino, T., Grayton, H., Tropeano, M., Mitchell, K., Vassos, E., Collier, D. (2014). 'Ohnologs are overrepresented in pathogenic copy number mutations.', *Proceedings of the National Academy of Sciences*, *111*(1), 361-366.
- Miller, M., Kumar, S. (2001) 'Understanding human disease mutations through the use of interspecific genetic variation' *Human Molecular Genetics* 10(21), 2319-2328.
- Petrovski, S, Wang, Q, Heinzen, E., Allen, A., Goldstein, D. (2013). 'Genic intolerance to functional variation and the interpretation of personal genomes' *PLoS Genetics* 9(8), e1003709.

- Polvi, A., Linturi, H., Varilo, T., Anttonen, A., Byrne, M., Fokkema, I., Almusa, H., Metzidis, A., Avela, K., Aula, P., Kestilä, M. (2013). 'The Finnish Disease Heritage Database (FinDis) Update—A Database for the Genes Mutated in the Finnish Disease Heritage Brought to the Next-Generation Sequencing Era' *Human mutation* 34(11), 1458-1466.
- Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L., Feldman, M. (2002). 'Genetic structure of human populations' *Science* 298(5602), 2381-2385.
- Schuster-Böckler, B., Conrad, D., Bateman, A. (2009). 'Dosage sensitivity shapes the evolution of copy-number varied regions.' *PloS one*, 5(3): e9474-e9474.
- Scotti, M., Swanson, M. (2016). 'RNA mis-splicing in disease' *Nature Reviews Genetics* 17(1):19-32
- Singh, P., Arora, J., Isambert, H. (2015). 'Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes' *PLoS Comput Biol* 11(7), e1004394.
- Soubry, A. (2015). 'Epigenetic inheritance and evolution: A paternal perspective on dietary influences' *Progress in biophysics and molecular biology* 118(1-2), 79-85
- Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A., Bork, P. (2001). 'Prediction of deleterious human alleles' *Human molecular genetics* 10(6), 591-597.
- Szyf, M. (2015). 'Nongenetic inheritance and transgenerational epigenetics' *Trends in molecular medicine* 21(2), 134-144.
- UK10K Consortium. (2015). 'The UK10K project identifies rare variants in health and disease' *Nature* 526(7571), 82-90.
- Wang, T., Birsoy, K., Hughes, N., Krupczak, K., Post, Y.,Wei, J., Lander, E., Sabatini, D. (2015). 'Identification and characterization of essential genes in the human genome' *Science Express*, 350(6264),1096-101.
- Wolfe, K. (2000). 'Robustness--it's not where you think it is' *Nature genetics* 25(1), 3-4.

- Zarrei, M., MacDonald, J., Merico, D., Scherer, S. (2015). 'A copy number variation map of the human genome', *Nature Reviews Genetics*, 16:172–183
- Zhu, X, Petrovski, S., Xie, P., Ruzzo, E., Lu, Y., McSweeney, K., Ben-Zeev, B., Nissenkorn, A., Anikster, Y., Oz-Levi, D., Dhindsa, R. (2015). 'Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios' *Genetics in Medicine*, 17(10), 774-781.

APPENDIX II: INTEGRATION OF LARGE-SCALE GENOMIC DATA SOURCES WITH EVOLUTIONARY INFORMATION REVEALS NOVEL GENETIC LOCI FOR CONGENITAL HEART DISEASE. FOTIOU *ET AL* 2019

Elisavet Fotiou MRes[1], Simon Williams PhD[1], **Alexandra Martin-Geary MSc[2]**, David L. Robertson PhD[2,3], Gennadiy Tenin PhD[1], Kathryn Hentges PhD[2], Bernard Keavney MD[1,4]

(In press at Circulation: Genomic and Precision Medicine)

[1]Division of Cardiovascular Sciences, School of Medical Sciences, Faculty of Biology, Medicine, and Health, Manchester Academic Health Science Centre, University of Manchester, Manchester, M13 9PT, UK, [2]Division of Evolution and Genomic science, University of Manchester, Oxford Rd, Manchester M13 9PL, [3]MRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH, [4]Manchester Heart Centre, Manchester University NHS Foundation Trust, M13 9WL.

Short title: Ohnologs in congenital heart disease

Correspondence should be addressed to Professor Bernard D. Keavney (email: bernard.keavney@manchester.ac.uk) and Miss Elisavet Fotiou (e-mail: Elisavet.fotiou@postgrad.manchester.ac.uk), The University of Manchester, AV Hill Building, Manchester, M13 9PT.

Abstract

Background: Most cases of congenital heart disease (CHD) are sporadic and non-syndromic, with poorly understood aetiology. Rare genetic variants have been found to affect the risk of sporadic, non-syndromic CHD, but individual studies to date are of only moderate sizes, and none to date has incorporated the ohnolog status of candidate genes in the analysis. Ohnologs are genes retained from ancestral whole-genome duplications during evolution; multiple lines of evidence suggest ohnologs are over-represented among dosage-sensitive genes. We integrated large-scale data on rare variants with evolutionary information on ohnolog status to identify novel genetic loci predisposing to CHD.

Methods: We compared copy number variants (CNVs) present in 4,634 non-syndromic CHD cases derived from publicly available data resources and the literature, and >27,000 healthy individuals. We analysed deletions and duplications independently and identified CNV regions exclusive to cases. These data were integrated with whole-exome sequencing data from 829 sporadic, non-syndromic patients with Tetralogy of Fallot. We placed our findings in an evolutionary context by comparing the proportion of vertebrate ohnologs in CHD cases and controls.

Results: Novel genetic loci in CHD cases were significantly enriched for ohnologs compared to the genome (χ^2-test, $p<0.0001$, OR=1.253, 95% CI:1.199-1.309). We identified 54 novel candidate

protein-coding genes supported by both: (i) CNV and whole-exome sequencing data; and (ii) ohnolog status.

Conclusions: We have identified new CHD candidate loci, and shown for the first time that ohnologs are over-represented among CHD genes. Incorporation of evolutionary metrics may be useful in refining candidate genes emerging from large-scale genetic evaluations of CHD.

Keywords: non-syndromic, congenital heart disease, copy number variants, whole-exome sequencing, ohnologs

Abbreviations:

CHD Congenital heart disease

CNV Copy number variant

DEL Deletion

DUP Duplication

SNV Single nucleotide variant

SSD Small-scale duplication

TOF Tetralogy of Fallot

WES Whole exome sequencing

WGD Whole-genome duplication

All Tables are supplied at the end of the document.

Background

Congenital heart disease (CHD) is the most prevalent birth defect in humans, occurring in approximately 8 per 1000 live births, and consisting of malformation of the heart and/or the great vessels. Around 20% of all CHDs can be attributed to chromosomal imbalances such as Down and Turner, and 22q11 deletion syndromes; around 80% occur as sporadic, non-syndromic CHD. In such cases, CHD behaves overall as a genetically complex trait with moderate heritability. Previous genome-wide investigations into CHD have found evidence for rare causative copy number variants (CNVs) and single nucleotide variants (SNVs); and associations with common SNVs in GWAS[2-6]. It has been estimated in previous studies that several hundred genes may be involved in polygenic CHD susceptibility; therefore, many remain to be discovered[7].

CNVs are 1 kilobase (kb) to several megabase (Mb) sized regions of duplication (DUP) and deletion (DEL) in the genome. A 2014 meta-analysis of CNVs in 1694 non-syndromic CHD cases identified 79 chromosomal regions in which 5 or more CHD cases had overlapping imbalances[5]. The estimated prevalence of pathogenic CNVs in non-syndromic CHD patients is 4-14%, whereas in syndromic CHD patients it is 15-20% (the most common being 22q11 deletion syndrome)[3, 8, 9]. There are multiple mechanisms by which a CNV may lead to disease including the disruption of chromosome structure, alteration of gene expression due to disruption of regulatory elements, and changes of the relative amounts of dosage-sensitive genes[10].

The dosage-balance model postulates that, for genes that are in stoichiometric relationships (for example with other genes forming protein complexes), any perturbation in their relative ratios will be deleterious[10]. In the early course of vertebrate evolution, around 500 million years ago, two whole-genome duplications (WGD), during which gene stoichiometry throughout the genome was preserved, as all genes were duplicated, took place. Periods of gene loss followed each of these events, resulting in the retention of some WGD paralogs in the genome (termed "ohnologs") and the loss of others. The dosage-balance model would predict that ohnologs should be enriched for dosage-sensitive genes. [11] Ohnologs, of which there are around 7,000 in the human genome, have indeed been shown to exhibit characteristics consistent with dosage-sensitivity: for example, ohnologs are enriched for haploinsufficient genes[11,12]; and Makino *et al*. reported, based on CNV data in healthy individuals from the Database of Genomic Variants (DGV), that genomic regions (~2Mb in size) near ohnologs are CNV deserts, indicating that those regions are dosage-balanced[13].

The formation and fixation of gene duplications within the genome is subject to different evolutionary mechanisms –small scale duplications (SSD) involving relatively few genes, and WGD. A strong relationship between the evolutionary mechanism of duplication and phenotypic consequences, including heritable diseases, has been previously shown[14] [15, 16]. Ohnologs have a significant association with certain human genetic diseases; for example 12 out of 16 reported candidate genes within the Down syndrome critical region (21q22.12, 21q22.13 and 21q22.2) are

dosage-balanced ohnologs[11]. By contrast, genes arising from SSDs have considerably fewer associations with disease[17]. In addition, ohnologs are enriched for genes involved in signalling and gene regulation, key cardiovascular developmental processes[11]. These considerations led us to hypothesise that ohnologs may be enriched among CHD causative genes.

We tested this hypothesis in a meta-analysis of CNV data including 4,634 non-syndromic CHD cases, and integrated these data with a whole-exome sequencing (WES) study of 829 cases of Tetralogy of Fallot (TOF), the commonest cyanotic CHD phenotype, which has been previously shown to have a significant aetiological contribution from CNVs[6]. Control data were derived from large-scale genomic resources[18-21].

Methods

The appropriate institutional review bodies approved all recruitment of human participants in this study. The study corresponded with the stipulations of the Declaration of Helsinki, and all participants (or their parents, if affected probands were children too young to themselves consent) provided informed consent. Data from consortia were accessed subject to the applicable data-sharing agreements. Summary data, analytic methods, and summary study materials will be made available to other researchers for purposes of reproducing the results or replicating the analyses reported here, on request to the corresponding authors. Full Materials and Methods are available in the Data supplement of the article.

Results

Update of CHD CNV dataset and generation of control CNV dataset

We updated the previous meta-analysis of CNVs in non-syndromic CHD cases[5], in a further 2,882 non-syndromic CHD cases from DECIPHER, ECARUCA and ISCA databases and further published studies investigating the role of CNVs in CHD[4, 20, 22-32]. The updated CHD case CNV dataset consists of 4,634 unrelated individuals of different ancestries (Table 1). The outline workflow to identify candidate genes is shown in Figure 1. Filtering of the CHD case population against DECIPHER known microdeletion/microduplication syndromes resulted in 224 cases being removed; this left 4,410 CHD cases with 3,362 DEL CNVs and 2,540 DUP CNVs which were used for further analysis (Supplementary figure 1).

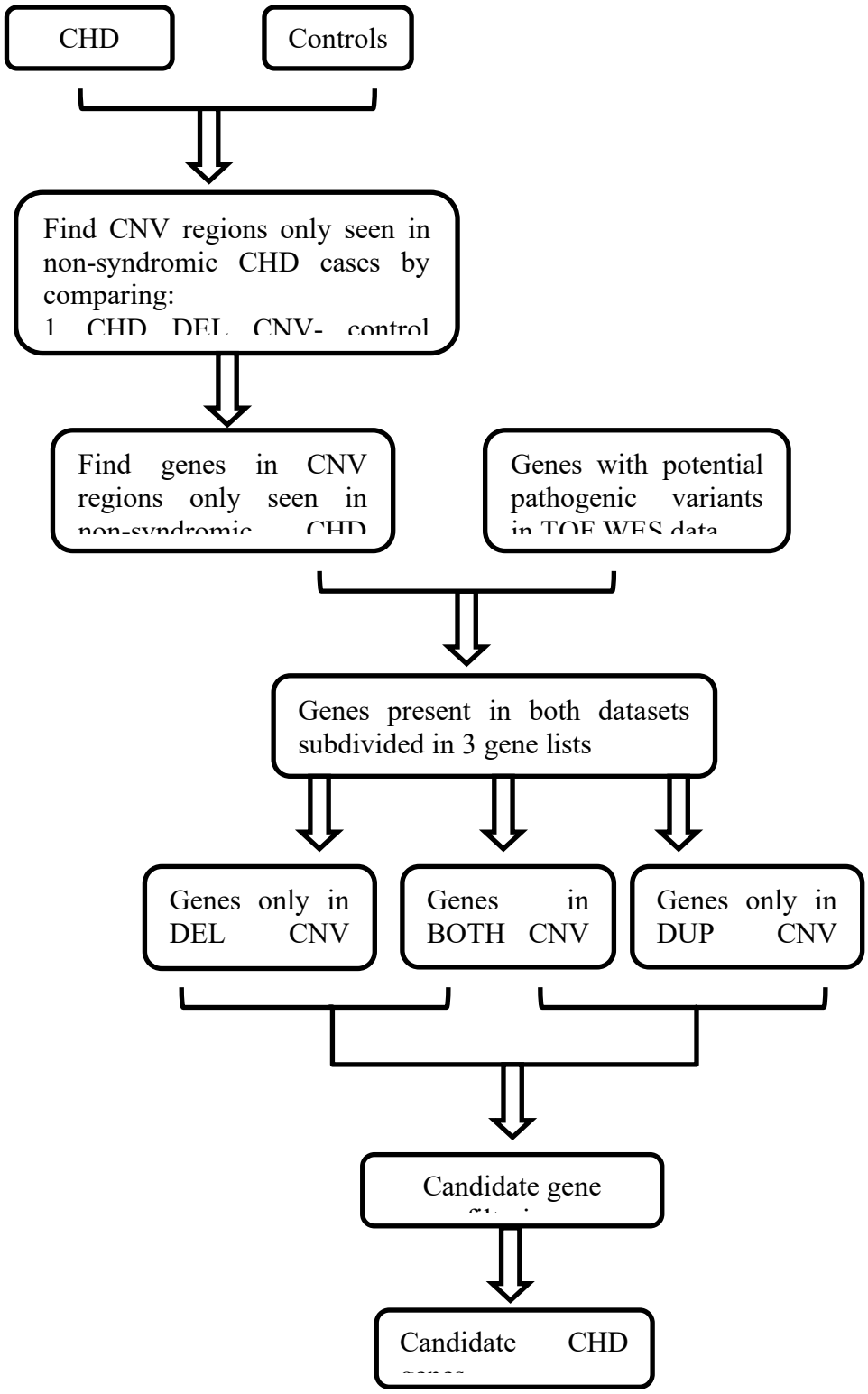


Figure 1: Overall methodology. Flowchart showing the methodology used to identify novel genetic loci for non-syndromic CHD cases. Potential pathogenic variants were novel or rare SNVs (either absent from ExAC or with frequency of <0.01). Candidate genes identified at the end of the workflow were subsequently analysed for ohnolog status.

A control CNV dataset was generated by acquiring CNV data from individuals not explicitly identified as having a developmental disorder, who were enrolled in the 1000 Genome Project Phase 3, DGV, DECIPHER, and published studies[21, 27, 28, 33, 34]. The control CNV dataset resulted in 256,511 DEL CNVs, 84,343 DUP CNVs and 6,403 BOTH CNVs, i.e. either DEL or DUP. gnomAD CNVs[35] were incorporated into the analysis as they became available, and resulted in an additional 51,420 DUP CNVs and 198,611 DEL CNVs.

Comparison of CHD CNV regions with control CNV regions

All CHD DEL and DUP CNV regions (coordinates hg19) were compared against control DEL and DUP CNV regions, respectively. Any CHD CNV regions overlapping control CNV regions were excluded. As a result, we identified DEL and DUP CNV regions only seen in non-syndromic CHD cases. The genes located in those regions were annotated using the Ensembl database. There were a number of genes that already had an assigned phenotype (OMIM)[36]; among these, 59 had been previously associated with CHD pathogenesis such as *ZIC3, NKX2-6, GATA4, JAG1, GJA1* and *TBX5*. All genes with OMIM assigned phenotypes were excluded from further analysis.

Novel genes found in the CNV regions only seen in CHD cases were then compared to an in-house list of 12,771 genes with novel or rare SNVs (either absent from ExAC or with frequency of <0.01) from WES data in 829 TOF cases[6]. Genes supported by both CNV and WES data were included for further analysis. In total, 3,082 genes in DEL CNVs, 4,297 genes in DUP CNVs and 3,068 in BOTH CNVs (i.e. genes found in

DEL and DUP CNVs) were also found in the TOF WES data with either high (nonsense variants, frameshift, splice variants) or medium (missense, splice variants) impact SNVs. This intersection of CNV and WES data led to an overall reduction of ~60% in the number of candidate genes for CHD (Figure 2).

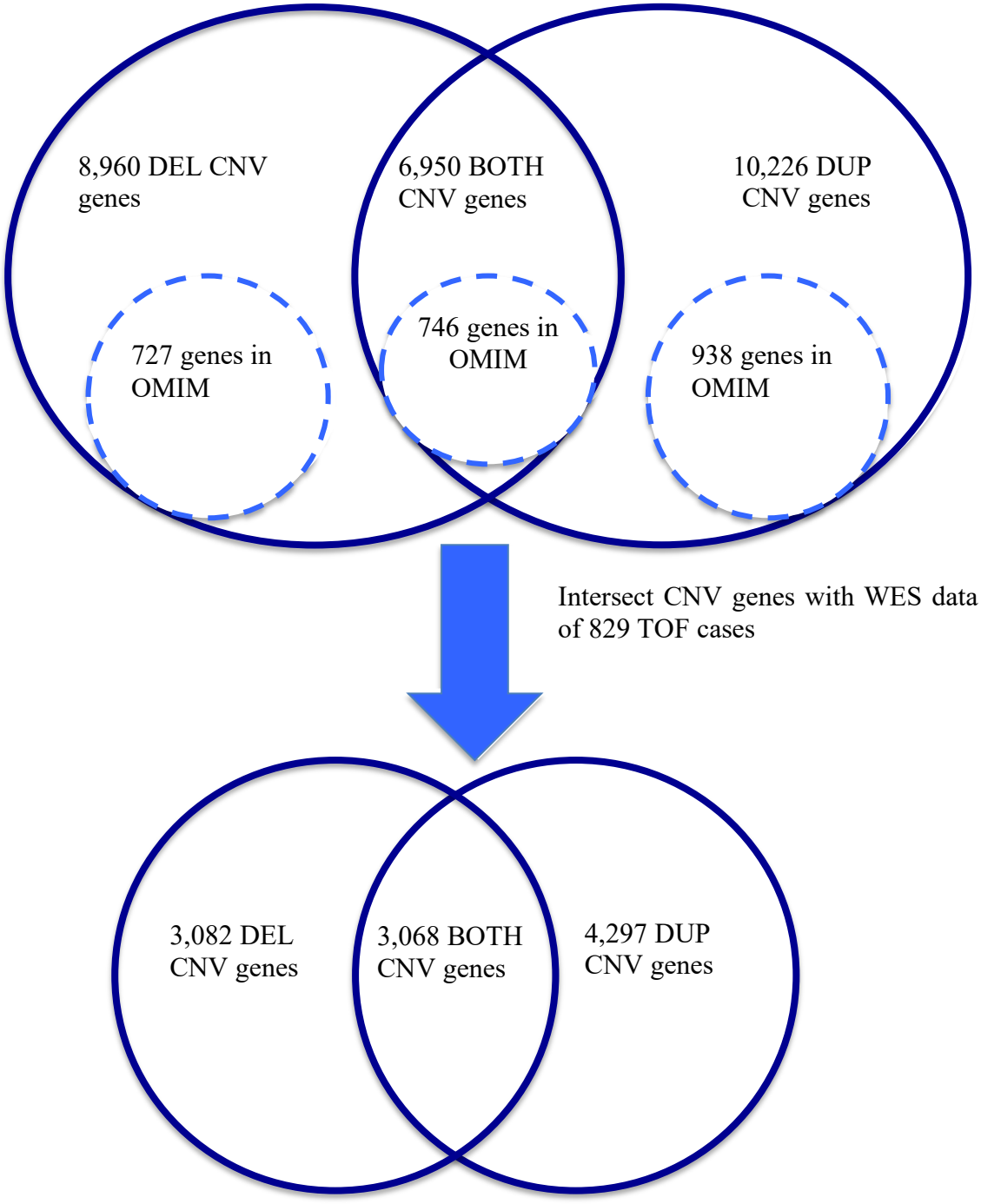


Figure 2: Intersection of CNV and WES data. Numbers of genes involved in the final stages of the workflow depicted in Figure 1 are shown. Genes with assigned phenotypes (circles with dashed line) were excluded from further analysis.

Ohnologs are highly enriched in CHD cases whereas small-scale duplications (SSD) and singleton genes are not.

Ohnologs (N=7,023) were identified using data from Singh *et al* (2015)[37], available at http://ohnologs.curie.fr/. SSDs (N=7,014) were extracted from Ensembl gene trees[12]. Any remaining genes that were neither found in the ohnolog dataset nor identified as having a direct paralog were considered for the purpose of this study to be singletons. The frequencies of ohnologs, SSDs and singletons among the candidate CHD genes were compared with their frequency in the human genome. Novel genes supported by the CNV data in CHD cases were found to be enriched for ohnologs (14.65% vs 12.05%, χ^2 test, p<0.0001, OR=1.253, 95% CI: 1.199-1.309,) (Figure 3A). There were no differences in SSDs (Figure 3B) and an under-representation for singletons (Figure 3C) compared to the human genome. There was a 2.3-fold increased enrichment of ohnologs in the genes supported by both CNV and WES data in CHD cases (χ^2 test, p< 0.0001, OR=3.751, 95% CI: 3.574-3.937). In this instance, SSDs were also enriched in CHD cases compared to the human genome (χ^2 test, p< 0.0001, OR=1.437, 95% CI: 1.356-1.905). However, ohnologs were 2-times elevated compared to SSD genes (33.94% versus 16.43%). Additionally, we assessed our methodology by applying it to a group of genes with strong *a priori* evidence for pathogenicity. The crowd-sourced Genomics England "PanelApp" gene list for CHD (available at https://panelapp.genomicsengland.co.uk/panels/212/), which represents a

consensus view of causative genes, was highly enriched for ohnologs (76.6% vs 12.05%, χ^2 test, p<0.0001, OR=23.89, 95% CI: 12.33-46.18). We therefore used ohnolog status as an additional candidate gene filter.

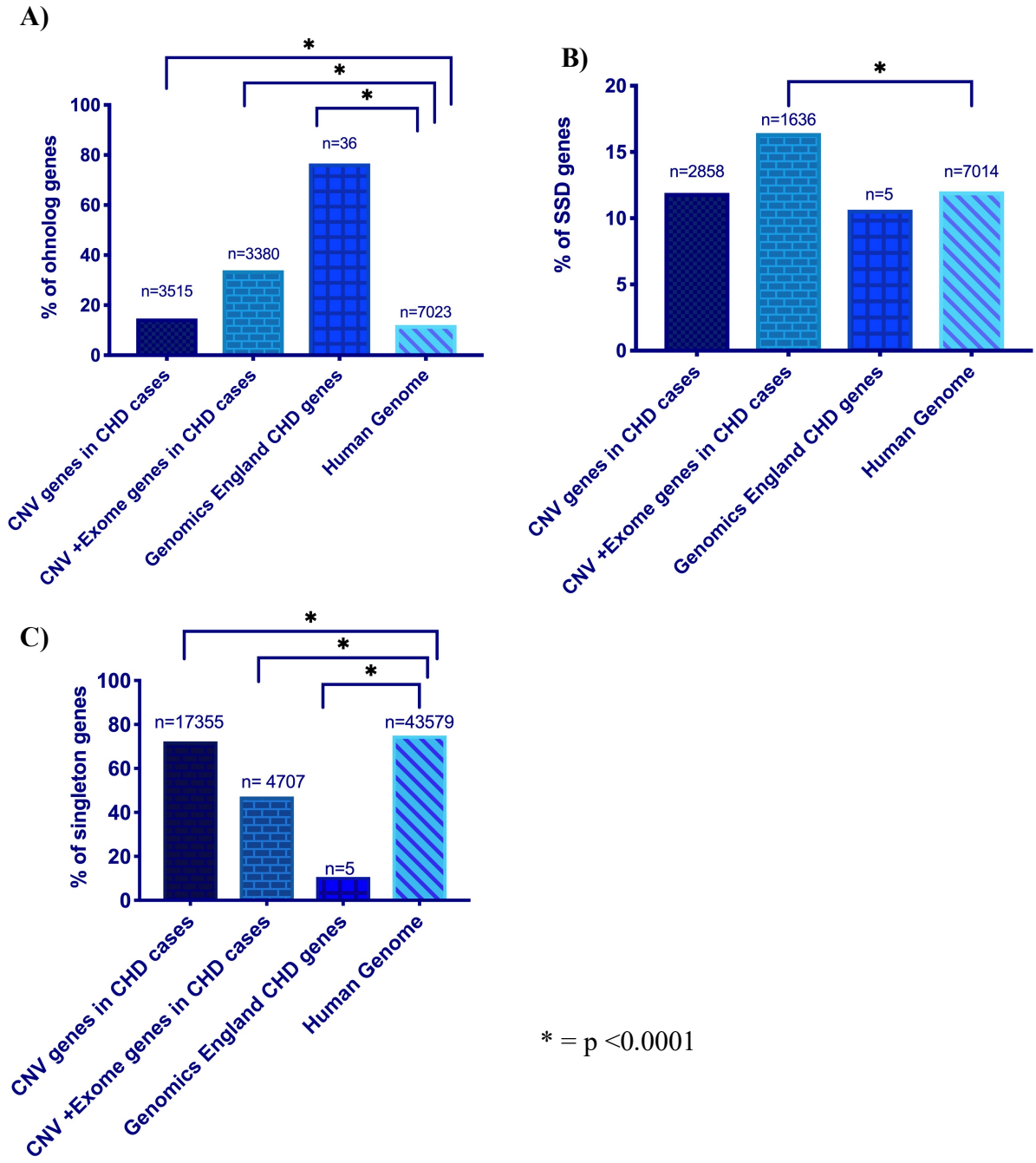


Figure 3: Ohnologs are enriched in CHD cases. Graphs show the percentage of genes that are **A)** ohnologs **B)** small scale duplications (SSD) and **C)** singletons. Statistical significance was tested using two-tailed Chi-square test with Yates's correction, p<0.05 was considered statistically significant.

Candidate genes supported by both CNV and WES data of CHD cases

In order to further refine our candidate genes, we integrated additional genomic resources including the top 5% ExAC CNV intolerance scores, probability of haploinsufficiency (pHI)[38], probability of loss-of-function intolerance (pLI)[19], and RNAseq expression data from mouse embryonic hearts[39]. Lastly, we incorporated ohnolog status. Genes from BOTH CNVs were analysed twice; once with the metrics used for genes from DEL CNVs and once with the metrics used for genes from DUP CNVs (Figure 4).

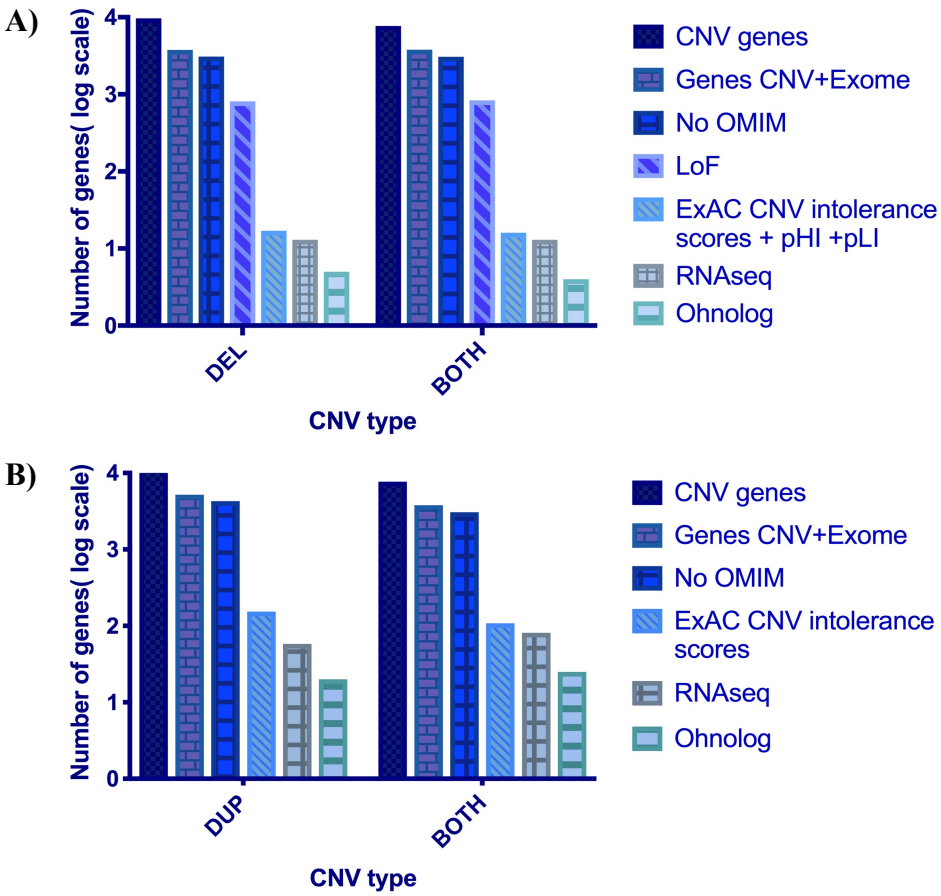


Figure 4: Filtering process using large-scale genomic data resources. Both graphs are in logarithmic scale and represent the consecutive filtering of the genes using the different metrics for **A)** deleted (DEL) and both CNV genes **B)** duplicated (DUP) and both CNV genes. There is approximately 70% reduction in the number of candidate genes when we apply the evolutionary duplication metric – ohnolog. Also, none of our candidates were present in the list of homozygous deleted genes (non-

essential) from the Sudmant study as well as not present in the list of genes curated from the DDD study.

This led to the identification of 9 candidate genes from DEL and BOTH CNVs: *BRWD1, DIP2C, EYA3, GRB10, HNRNPC, RC3H2, SLIT3, TLN1* and *UBASH3B*. All 9 have the following properties: a) loss-of-function (LoF) variants in the WES data, b) found in DEL or BOTH CNV regions only seen in non-syndromic CHD cases, c) top 5% of ExAC DEL CNV intolerance scores, d) haploinsufficient (pHI≥0.65) and/or unable to tolerate LoF variants (pLI≥0.9), e) in the top 25% of highly expressed genes in mouse heart at E9.5 and/or E14.5, f) ohnolog, g) not present in the list of genes curated from the DDD study, h) not classified as human non-essential genes from the Sudmant study[21] (Table 2).

In addition, we found 45 candidate genes from DUP and BOTH CNVs, which had the following properties: a) high or medium impact SNVs in the WES data, b) found in DUP and BOTH CNV regions only seen in non-syndromic CHD cases, c) top 5% of ExAC DUP CNV intolerance scores, d) in the top 25% of highly expressed genes in mouse heart at E9.5 and/or E14.5, e) ohnolog, f) not present in the list of genes curated from the DDD study, g) not in the list of non-essential human genes from the Sudmant study[21] (Table 2).

Pathway enrichment and gene ontology analysis

We performed pathway enrichment analysis, using the Reactome Pathways Analysis tool[40], on the final 54 candidate genes supported by both CNV and WES data in non-syndromic CHD cases. This resulted in 11 pathways, where >5 of our candidate genes were involved in those

pathways (Table 3). The top 3 pathways based on entities ratio (entities found/total entities) from Reactome were ‘’axon guidance’’, ‘’signalling by receptor tyrosine kinases’’ and ‘’cellular responses to external stimuli’’. In addition, Ingenuity pathway analysis (IPA) was also used with the only pathway including >5 genes being ‘’axon guidance signalling’’. Gene ontology analysis[41] of our candidate genes revealed 22 Gene ontology (GO) terms with particular enrichment on 4 GO terms; apoptotic process involved in luteolysis (GO0061364) (FDR corrected p-value= 0.0462), ventricular septum morphogenesis (GO0060412) (FDR corrected p-value=0.00921), ventricular septum development (GO0003281) (FDR corrected p-value=0.0343) and cardiac septum morphogenesis (GO0060411) (FDR corrected p-value=0.036). Both pathway and gene ontology analysis identified processes in which the genes *ABLIM1, ARHGEF12, SLIT2* and *SLIT3* are involved (Figure 5).

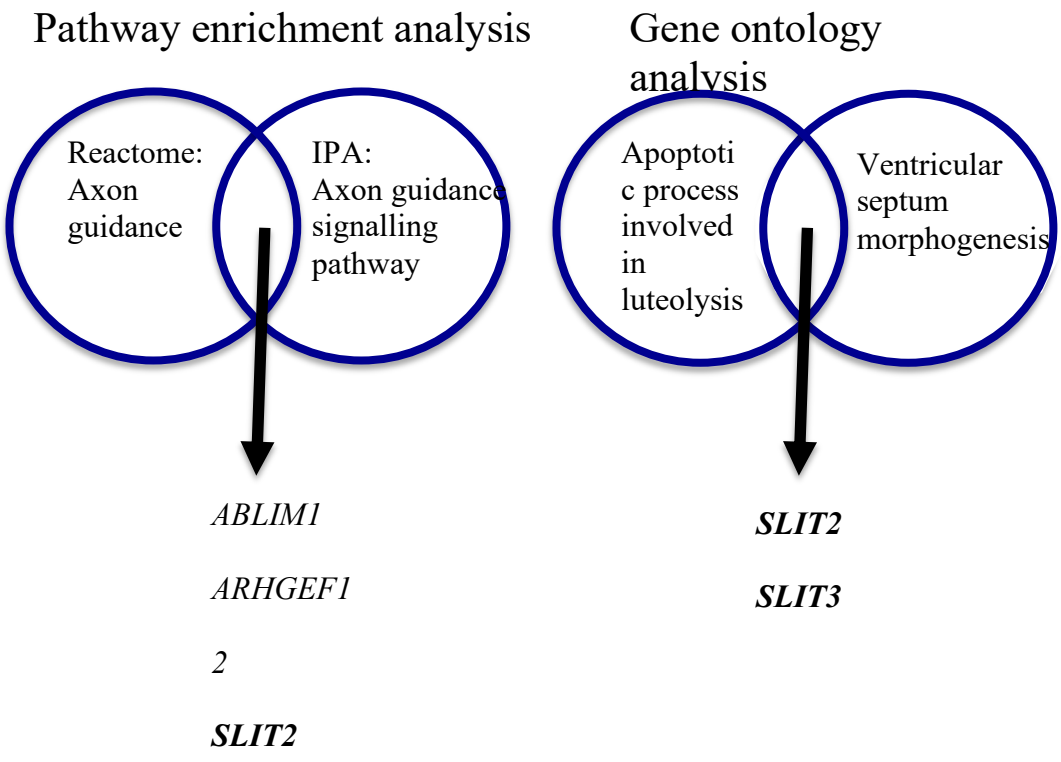


Figure 5: Genes in the top significant pathways and biological processes. *SLIT2* and *SLIT3* genes were supported by multiple lines of evidence

SLIT2 and *SLIT3* variants in CHD

SLIT2 and *SLIT3* were the most strongly supported genes found both by pathway analysis and gene ontology (Figure 5). Therefore, we further explored the phenotypic associations of these genes within our population.

In the present study, individuals with CNVs including *SLIT3* were reported with malformation of the heart and great vessels (n=1), VSD (n=1), atrial septal defect (n=3) and TOF (n=1) whereas individuals with *SLIT2* CNVs were reported with malformation of the heart and great vessels (n=1), VSD (n=2) and double outlet right ventricle (n=1). In addition, 20 missense SNVs and 3 splice-site SNVs in *SLIT3* were found in 24 out of 829 TOF cases (2.9%, 95% CI: 1.91%-4.35%) and *SLIT2* had 12 missense SNVs and 2 splice-site SNVs in 14 out of 829 TOF cases (1.7%, 95% CI: 0.9%-2.9%). Probands were available for 12 *SLIT3* variants and 5 *SLIT2* variants which were confirmed by Sanger sequencing. Remaining variants were confirmed to have good coverage using Integrative Genomics Viewer (IGV). Samples from both parents were available for 9 probands with *SLIT3* variants and were analysed for variant inheritance; 2 of the 9 *SLIT3* variants tested were identified as *de novo*. Samples from both parents were available for 5 probands with *SLIT2* variants and were all either maternally or paternally inherited.

Table 1. Number of cases in previous and current meta-analysis studies as well as controls used in the current study.

DECIPHER= Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources, ISCA= International Standards for Cytogenomic Arrays, ECARUCA= European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations, WES TOF= whole exome sequencing of Tetralogy of Fallot, DGV= Database of Genomic Variants, WTCC2= Wellcome Trust Case Control Consortium 2, DDD= Deciphering Developmental Disorders study

Databases for CHD cases	**Thorsson *et al.* study**	**Current study**	**Databases for controls**	**Current study**
DECIPHER	279	1,252	**1,000 Genome phase 3**	2,504
ISCA	331	1,107	**DGV**	>6,430
Published literature	814	1,900	**gnomAD**	10,738
CHDwiki	328	328	**Published literature**	356
ECARUCA	0	47	**WTCC2**	~6,000
WES TOF (Page *et al.* 2019)	N/A	829	**DDD**	845

Table 2. Candidate genes supported by both CNV and WES data of CHD cases. 54 protein-coding candidate genes supported by CNV and WES data in non-syndromic CHD cases. All genes in the list are strict ohnologs. Data presented includes the Ensembl ID and the nature of the chromosomal imbalance for which the gene is either deleted (DEL), duplicated (DUP) or deleted/duplicated (BOTH).

TOF= tetralog

ENS gene ID	Gene name	Chr	Start (hg19)	End (hg19)	DEL/DUP/BOTH	TOF LOF var count	TOF HIGH impact var count	TOF MED impact var count	Case CNVs overlap: FULL	Case CNVs overlap: PARTIAL
ENSG00000058668	ATP2B4	1	203595689	203713209	DUP	0	0	13	3	1
ENSG00000064042	LIMCH1	4	41361624	41702061	DUP	6	6	21	4	0
ENSG00000083223	ZCCHC6	9	88902648	88969369	DUP	1	1	8	4	0
ENSG00000092847	AGO1	1	36335409	36395211	DUP	0	0	2	3	0
ENSG00000094916	CBX5	12	54624724	54673886	DUP	0	0	1	2	0
ENSG00000101367	MAPRE1	20	31407699	31438211	DUP	0	0	1	5	0
ENSG00000108387	SEPT4	17	56597611	56618179	DUP	0	0	6	5	1

ENSG00000108389	MTMR4	17	56566898	56595266	DUP	0	0	9	10	0
ENSG00000109332	UBE2D3	4	103715540	103790053	DUP	2	2	0	3	0
ENSG00000109685	WHSC1	4	1873151	1983934	DUP	0	0	12	6	0
ENSG00000112079	STK38	6	36461669	36515247	DUP	0	0	3	5	1
ENSG00000113108	APBB3	5	139937853	139973337	DUP	1	2	8	8	0
ENSG00000122515	ZMIZ2	7	44788180	44809477	DUP	1	1	10	10	0
ENSG00000138641	HERC3	4	89442199	89629693	DUP	0	0	5	14	0
ENSG00000138835	RGS3	9	116207011	116360018	DUP	1	1	19	5	0
ENSG00000140403	DNAJA4	15	78556428	78574538	DUP	0	0	8	9	0
ENSG00000140497	SCAMP2	15	75136071	75165706	DUP	0	0	1	6	0
ENSG00000145147	SLIT2	4	20254883	20622184	DUP	0	0	14	3	0
ENSG00000146463	ZMYM4	1	35734568	35887659	DUP	0	0	12	26	1
ENSG00000179361	ARID3B	15	74833518	74890472	DUP	0	0	6	5	0
ENSG00000185658	BRWD1	21	40556102	40693485	DEL	5	5	16	10	1
ENSG00000151240	DIP2C	10	320130	735683	DEL	1	1	15	4	0
ENSG00000158161	EYA3	1	28296855	28415207	DEL	1	1	2	4	0

ENSG00000106070	GRB10	7	50657760	50861159	DEL	1	2	8	3	0
ENSG00000154127	UBASH3B	11	122526383	122685181	DEL	1	1	5	9	0
ENSG00000092199	HNRNPC	14	21677295	21737653	BOTH	1	1	2	6	0
ENSG00000056586	RC3H2	9	125606835	125667620	BOTH	1	1	8	3	1
ENSG00000184347	SLIT3	5	168088745	168728133	BOTH	2	2	21	3	0
ENSG00000137076	TLN1	9	35696945	35732392	BOTH	4	4	22	9	0
ENSG00000010017	RANBP9	6	13621730	13711796	BOTH	0	0	14	3	0
ENSG00000020577	SAMD4A	14	55033815	55260033	BOTH	0	0	11	8	0
ENSG00000033800	PIAS1	15	68346517	68483096	BOTH	0	0	4	7	1
ENSG00000064726	BTBD1	15	83685174	83736106	BOTH	1	1	7	5	0
ENSG00000083312	TNPO1	5	72112139	72212560	BOTH	0	0	4	5	0
ENSG00000091009	RBM27	5	145583113	145718814	BOTH	0	0	7	4	0
ENSG00000099204	ABLIM1	10	116190872	116444762	BOTH	3	3	14	12	0
ENSG00000100320	RBFOX2	22	36134783	36424473	BOTH	0	0	5	2	0
ENSG00000100330	MTMR3	22	30279144	30426855	BOTH	0	0	11	4	0
ENSG00000100592	DAAM1	14	59655364	59838123	BOTH	0	0	9	14	0

ENSG00000113649	TCERG1	5	145826874	145891524	BOTH	0	0	6	9	0
ENSG00000116191	RALGPS2	1	178694300	178889238	BOTH	0	0	1	2	0
ENSG00000120899	PTK2B	8	27168999	27316903	BOTH	0	0	8	4	0
ENSG00000127022	CANX	5	179105629	179157926	BOTH	0	0	10	2	0
ENSG00000135074	ADAM19	5	156822542	157002783	BOTH	2	2	10	7	1
ENSG00000137573	SULF1	8	70378859	70573150	BOTH	1	1	10	3	0
ENSG00000137962	ARHGAP29	1	94614544	94740624	BOTH	0	0	12	3	0
ENSG00000138107	ACTR1A	10	104238986	104262482	BOTH	0	0	2	2	1
ENSG00000155506	LARP1	5	154092462	154197167	BOTH	2	2	13	6	0
ENSG00000166747	AP1G1	16	71762913	71843104	BOTH	0	1	4	4	1
ENSG00000166888	STAT6	12	57489191	57525922	BOTH	0	0	7	5	0
ENSG00000180340	FZD2	17	42634925	42636907	BOTH	0	0	8	8	0
ENSG00000180776	ZDHHC20	13	21950263	22033509	BOTH	1	1	2	6	0
ENSG00000196914	ARHGEF12	11	120207787	120360645	BOTH	0	0	14	12	0
ENSG00000213079	SCAF8	6	155054459	155155192	BOTH	0	0	13	14	0

Table 3. Top pathways overrepresented in our 54 candidate genes.

= number

Pathway tool	Pathway name	# Entities found	# Entities total	Entities ratio (%)
Reactome	Axon guidance	8	583	1.372212693
Ingenuity pathway analysis	Axon guidance signalling pathway	7	501	1.397206
Reactome	Signaling by Receptor Tyrosine Kinases	5	521	0.959692898
Reactome	Cellular responses to external stimuli	5	621	0.805152979
Reactome	Signalling by Interleukins	5	641	0.780031201
Reactome	Developmental Biology	8	1177	0.679694138
Reactome	Adaptive Immune System	6	998	0.601202405
Reactome	Cytokine Signalling in Immune system	6	1056	0.568181818
Reactome	Signal Transduction	15	3202	0.468457214
Reactome	Post-translational protein modification	7	1594	0.439146801
Reactome	Immune System	11	2662	0.41322314
Reactome	Metabolism of proteins	9	2354	0.382327952

References

1. Liu Y, Chen S, Zühlke L, Black GC, Choy MK, Li N, et al. Global birth prevalence of congenital heart defects 1970-2017: updated systematic review and meta-analysis of 260 studies. *Int J Epidemiol*. 2019.
2. Cordell HJ, Bentham J, Topf A, Zelenika D, Heath S, Mamasoula C, et al. Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat Genet*. 2013;45:822-4.
3. Soemedi R, Wilson IJ, Bentham J, Darlay R, Töpf A, Zelenika D, et al. Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am J Hum Genet*. 2012;91:489-501.
4. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet Med*. 2011;13:777-84.
5. Thorsson T, Russell WW, El-Kashlan N, Soemedi R, Levine J, Geisler SB, et al. Chromosomal Imbalances in Patients with Congenital Cardiac Defects: A Meta-analysis Reveals Novel Potential Critical Regions Involved in Heart Development. *Congenit Heart Dis*. 2015;10:193-208.
6. Page DJ, Miossec MJ, Williams SG, Monaghan RM, Fotiou E, Cordell H, et al. Whole Exome Sequencing Reveals the Major Genetic Contributors to Non-Syndromic Tetralogy of Fallot. *Circ Res*. 2018.
7. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet*. 2017;49:1593-1601.
8. Greenway SC, Pereira AC, Lin JC, DePalma SR, Israel SJ, Mesquita SM, et al. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat Genet*. 2009;41:931-5.
9. Costain G, Roche SL, Scherer SW, Silversides CK and Bassett AS. Rare copy number variations in an adult with transposition of the great arteries emphasize the importance of updated genetic assessments in syndromic congenital cardiac disease. *Int J Cardiol*. 2016;203:516-8.
10. Rice AM and McLysaght A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat Commun*. 2017;8:14366.
11. Makino T and McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A*. 2010;107:9270-4.
12. Martin-Geary A, Reardon M, Keith B, Tassabehji M and Robertson DL. Human genetic disease is greatly influenced by the underlying fragility of evolutionarily ancient genes. *bioRxiv*. 2019:558916.
13. Makino T, McLysaght A and Kawata M. Genome-wide deserts for copy number variation in vertebrates. *Nat Commun*. 2013;4:2283.
14. McLysaght A, Makino T, Grayton HM, Tropeano M, Mitchell KJ, Vassos E, et al. Ohnologs are overrepresented in pathogenic copy number mutations. *Proc Natl Acad Sci U S A*. 2014;111:361-6.
15. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS).

The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol*. 2013;42:689-700.
16. Dickerson JE and Robertson DL. On the origins of Mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol*. 2012;29:61-9.
17. Singh PP, Affeldt S, Malaguti G and Isambert H. Human dominant disease genes are enriched in paralogs originating from whole genome duplication. *PLoS Comput Biol*. 2014;10:e1003754.
18. Ruderfer DM, Hamamsy T, Lek M, Karczewski KJ, Kavanagh D, Samocha KE, et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet*. 2016;48:1107-11.
19. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285-91.
20. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res*. 2014;42:D993-D1000.
21. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526:75-81.
22. Vulto-van Silfhout AT, van Ravenswaaij CM, Hehir-Kwa JY, Verwiel ET, Dirks R, van Vooren S, et al. An update on ECARUCA, the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations. *Eur J Med Genet*. 2013;56:471-4.
23. Fakhro KA, Choi M, Ware SM, Belmont JW, Towbin JA, Lifton RP, et al. Rare copy number variations in congenital heart disease patients identify unique genes in left-right patterning. *Proc Natl Acad Sci U S A*. 2011;108:2915-20.
24. Hitz MP, Lemieux-Perreault LP, Marshall C, Feroz-Zada Y, Davies R, Yang SW, et al. Rare copy number variants contribute to congenital left-sided heart disease. *PLoS Genet*. 2012;8:e1002903.
25. Glessner JT, Bick AG, Ito K, Homsy JG, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circ Res*. 2014;115:884-96.
26. Rigler SL, Kay DM, Sicko RJ, Fan R, Liu A, Caggana M, et al. Novel copy-number variants in a population-based investigation of classic heterotaxy. *Genet Med*. 2015;17:348-57.
27. Hightower HB, Robin NH, Mikhail FM and Ambalavanan N. Array comparative genomic hybridisation testing in CHD. *Cardiol Young*. 2015;25:1155-72.
28. Sanchez-Castro M, Eldjouzi H, Charpentier E, Busson PF, Hauet Q, Lindenbaum P, et al. Search for Rare Copy-Number Variants in Congenital Heart Defects Identifies Novel Candidate Genes and a Potential Role for FOXC1 in Patients With Coarctation of the Aorta. *Circ Cardiovasc Genet*. 2016;9:86-94.
29. Hanchard NA, Umana LA, D'Alessandro L, Azamian M, Poopola M, Morris SA, et al. Assessment of large copy number variants in patients with apparently isolated congenital left-sided cardiac lesions reveals clinically relevant genomic events. *Am J Med Genet A*. 2017;173:2176-2188.

30. Xie L, Chen JL, Zhang WZ, Wang SZ, Zhao TL, Huang C, et al. Rare de novo copy number variants in patients with congenital pulmonary atresia. *PLoS One*. 2014;9:e96471.
31. Xie HM, Werner P, Stambolian D, Bailey-Wilson JE, Hakonarson H, White PS, et al. Rare copy number variants in patients with congenital conotruncal heart defects. *Birth Defects Res*. 2017;109:271-295.
32. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet*. 2010;86:749-64.
33. MacDonald JR, Ziman R, Yuen RK, Feuk L and Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42:D986-92.
34. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
35. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Khera AV, et al. An open resource of structural variation for medical and population genetics. *bioRxiv*. 2019:578674.
36. McKusick-Nathans Institute of Genetic Medicine JHUB, MD). Online Mendelian Inheritance in Man, OMIM ®. 2019.
37. Singh PP, Arora J and Isambert H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput Biol*. 2015;11:e1004394.
38. Huang N, Lee I, Marcotte EM and Hurles ME. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet*. 2010;6:e1001154.
39. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature*. 2013;498:220-3.
40. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res*. 2018;46:D649-D655.
41. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*. 2017;45:D183-D189.
42. Zhao J and Mommersteeg MTM. Slit-Robo signalling in heart development. *Cardiovasc Res*. 2018;114:794-804.
43. Blockus H and Chédotal A. Slit-Robo signaling. *Development*. 2016;143:3037-44.
44. Mommersteeg MT, Yeh ML, Parnavelas JG and Andrews WD. Disrupted Slit-Robo signalling results in membranous ventricular septum defects and bicuspid aortic valves. *Cardiovasc Res*. 2015;106:55-66.
45. Medioni C, Bertrand N, Mesbah K, Hudry B, Dupays L, Wolstein O, et al. Expression of Slit and Robo genes in the developing mouse heart. *Dev Dyn*. 2010;239:3303-11.

46. Mommersteeg MT, Andrews WD, Ypsilanti AR, Zelina P, Yeh ML, Norden J, et al. Slit-roundabout signaling regulates the development of the cardiac systemic venous return and pericardium. *Circ Res*. 2013;112:465-75.
47. Liu J, Zhang L, Wang D, Shen H, Jiang M, Mei P, et al. Congenital diaphragmatic hernia, kidney agenesis and cardiac defects associated with Slit3-deficiency in mice. *Mech Dev*. 2003;120:1059-70.
48. Kruszka P, Tanpaiboon P, Neas K, Crosby K, Berger SI, Martinez AF, et al. Loss of function in ROBO1 is associated with tetralogy of Fallot and septal defects. *J Med Genet*. 2017;54:825-829.
49. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*. 2010;466:368-72.

APPENDIX III : THERE IS NO PLACE LIKE HOME: DEMOGRAPHIC ANALYSIS OF THE UNDULATE RAY, *RAJA UNDULATA*, USING NON-INVASIVE MARK-RECAPTURE AND DNA ANALYSIS. HOOK *ET AL* 2019.

Samantha A. Hook[1*], Martin Openshaw[2], Sheilah Openshaw[2], Daniel Ripley[4], **Alexandra Martin-Geary[4]**, Matthew Doggett[2, 3], Jean-Denis Hibbitt[5], Michael Buckley[1]

(Submission pending to: Conservation Biology)

[1]School of Earth and Environmental Sciences, Manchester Institute of Biotechnology, University of Manchester, Manchester, M1 7DN, UK

[2] Stardis, Kingsclere, Hampshire, RG20 4SY, UK

[3]Seven Tenths Ecology Ltd, Salisbury, SP5 2BY, UK

[4]Faculty of Biology, Medicine and Health, University of Manchester, Core Technology Facility, Manchester, M13 9NT, UK

[5] SEA LIFE Programmes and Engagement, SEA LIFE Weymouth

*Corresponding Author: Samantha A. Hook, samhook1205@gmail.com

Key words: elasmobranchs, Photo-recognition, microsatellites, Skates, *Raja undulata*, Genetic Health

Supplementary materials available upon request.

Abstract

Effective methods of individual identification and genetic sampling are necessary to properly determine elasmobranch population densities and genetic health. As populations decline, the need to develop non-invasive methods of population assessment becomes ever more urgent. The undulate ray, *Raja undulata,* is a globally endangered, but often locally-abundant species that has various protection levels throughout its range. However, there is a heavy reliance on fisheries data to conduct population analysis for the management and conservation of *R. undulata* despite being threatened with extinction. Here, we conduct the first global genetic assessment of wild *R. undulata* using microsatellites marker analysis of 143 individuals from mucus and tissue sampling. To collect underwater mucus samples, we developed a successful non-invasive method for resting *R. undulata* using scuba divers. Secondly, between 2012 and 2018, we conducted a capture mark-recapture study on an underwater sampling site on the coast of Dorset, UK, and processed images through Wild-ID to identify 263 individuals. The genetic results exhibited an overall high average genetic diversity (H_o = 0.66, H_e = 0.85, average alleles per locus = 19.8). Capture mark-recapture analysis demonstrated the highest number of *R. undulata* were present in autumn, and that the average estimated population size was 228 (maximum standard error ± 87). Despite a low probability of recapture (p = 0.035), the individual survivorship rate between visits was high (φ = 0.969). Finally, we use network analysis to investigate the social behaviour of *R. undulata*, demonstrating that distinct pairs of rays are present at a higher frequency than is expected by chance (95% confidence interval $P \leq 0.02$), indicating a high probability of same pair migration and social interaction.

Introduction

Overfishing is the main cause of the decline of shark, skate and ray (elasmobranchs) populations around the globe (Dulvy et al., 2014; Worm et al., 2013). Coincidentally, because of the difficulties surrounding the direct observation of individuals within the marine environment, fishing data also form the main source of the fish stock assessments that influence conservation management (Beddington et al., 2007). Monitoring the change in capture rates is the simplest and most common method of estimating population size, while sample collection from fisheries of either tissue or whole specimens has assisted in species identification and population genetics (Larson et al., 2017). As of 2014, a quarter of all elasmobranchs were classified as threatened with extinction under the IUCN Red List (Dulvy et al., 2014; IUCN, 2014). Thus, there is an urgent need to develop and implement better methods of population assessment.

The undulate ray, *Raja undulata,* is a globally endangered species of skate with a fragmented population distribution from the North-east Atlantic to the equator and the Mediterranean Sea (Coelho et al., 2009). In 2009, the EU enforced laws that prohibited landings in the North-east Atlantic, and placed restrictions on landings in the Mideast-Atlantic and Mediterranean (CEC, 2010). In North Africa however, high levels of illegal, unregulated and unreported (IUU) (CoC, 2015) fishing means there is little management enforced for this species. The landing restrictions within the North-east Atlantic were a controversial topic due to claims that the fisheries were catching large quantities (Ellis et al., 2012). Independent trawler surveys were conducted to provide evidence for abundance and the results contributed towards a regional delisting of the species to Near Threatened and a gradual increase in quota sizes (Ellis et al., 2015). However, despite these stock assessments relatively little is

known about the movements and the connectivity of *R. undulata* between sites (Ellis et al., 2012).

More recently, studies have employed tagging and satellite tracking to investigate elasmobranch movements and population abundance through capture mark-recapture (CMR) (Feldheim et al., 2002; Guttridge et al., 2010; Hunter et al., 2005). Capture mark-recapture uses tagging to identify individuals, allowing recording of presence and absence data, and movements between study areas (McCrea and Morgan, 2014). Analysis of CMR data can provide population estimates, and population viability and survivorship probabilities (McCrea and Morgan, 2014), without the need to destroy individuals. Furthermore, due to developments in recognition software (Speed et al., 2007), non-invasive methods such as photographs can now be used to confidently identify individuals. Large image databases of specific regions of the fish that hold unique markings, such as patterns or marks, pigmentation, or long-standing scars, can be used to identify individuals over long periods of time. Examples include images of the dorsal fin in white sharks *Carcharodon carcharias* (Andreotti et al., 2018), spot patterns on the dorsal side of the spotted eagle ray *Aetobatus narinari* (González-Ramos et al., 2017) and the ventral side's natural pigmentation and spots on manta rays *Manta alfredi* and *Manta birostris* (Ari, 2015; Marshall and Pierce, 2012). Recognition software overcomes inherent sources of bias associated with invasive tagging methods such as mortality from tag application, non-reported or non-recovered tags, and tag shedding (Kohler and Turner, 2001).

Genetic techniques have previously been used to validate the use of recognition software in elasmobranch species (Andreotti et al., 2016). Furthermore, genetic sampling is a powerful tool for understanding population dynamics, separate to or

alongside, CMR data. However, the collection of DNA no longer depends on invasive methods such as tissue or blood sampling, which often increase the risk of mortality or rely on the collection of fished individuals. The use of mucus sampling to collect DNA is a recently established non-invasive technique that has proven a viable alternative for *in situ* populations of the basking shark *Cetorhinus maximus* (Lieber et al., 2013) and manta ray *Manta birostris* (Kashiwagi et al., 2015), and *ex situ* captive populations of *R. undulata* (Fox et al., 2018).

Here we use 17 microsatellite markers, previously defined by Hunter *et al.* (2016) and Fox *et al.* (2018), to investigate the genetic relationships of *R. undulata* between 12 sample sites across their global distribution. We analyse genetic health and review differences between previously prohibited fishing regions (England, North-east Atlantic), areas with restricted quotas (mainland Europe, Mideast-Atlantic) and areas with no known quotas/high levels of IUU fishing activity (Morocco, North Africa). Secondly, with the use of CMR on a known site in England, we compare population estimates with genetic effective population size measured as effective number of breeders. Finally, we review the probability of recapture, survivorship and relationships between individuals within our single population of *R. undulata*, which, for most of this study, has been a protected species in the UK.

Methods

Global genetic analysis

DNA sampling

Tissue samples of *R. undulata* were collected as a by-product from fish markets in Portugal, Spain, and Morocco between 2015 and 2018 (Figure 1.). Samples were stored in RNA*later*® at -4˚C before being transferred to -80°C at the Manchester Institute of Biotechnology.

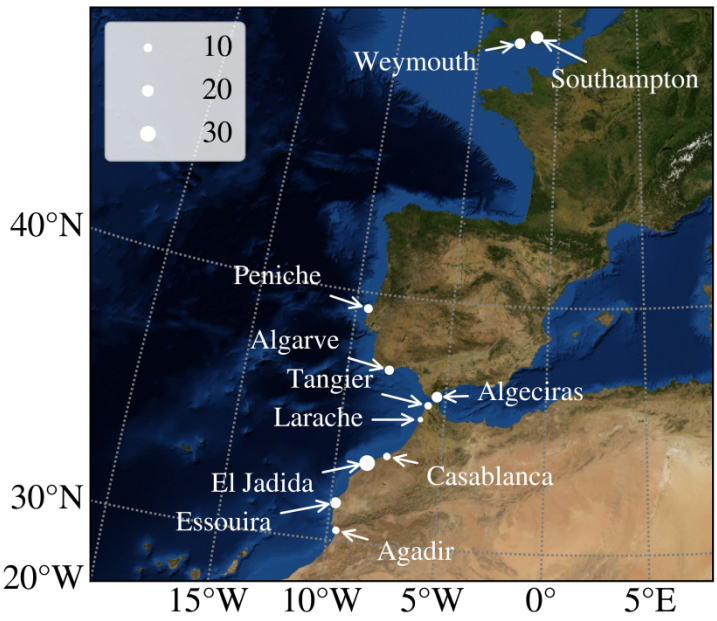


Figure 1. A map displaying the locations and total number of either tissue or mucus samples per site created using Python (v 2.7.16, (van Rossum, 1995) from the basemap library. Note, Rays' Repose was not displayed due to its close location between Weymouth and Southampton.

Non-invasive underwater mucus swabs from *R. undulata* were collected by a group of volunteer recreational British Sub Aqua Club (BSAC) accredited scuba divers between April 2017 and October 2018, on a site within the Studland to Portland Special Area of Conservation, Dorset, England. The site, hereafter referred to as Rays' Repose, is approximately 50 metres wide and over 200 metres from north to

south, forming part of the Kimmeridge Ledges (50° 35.5’ N 2° 7.5' W); a series of shallow flat ledges reaching out to sea for up to a mile. Qualified divers followed Supplementary Methods 1. to collect samples and take photographs for identification. Swab samples were taken using an autoclaved heavy-duty green scrubber (Robert Scott Ltd. Code 102450) attached to a 10 cm handle (Supplementary Methods 1.). The scrubber was gently pressed on the upper dorsal of the nearest wing and moved towards the tail end between one and three times, before placing the sample into a marked zip locked bag (Supplementary Methods 1.). A photograph was then taken of the marked zip lock bag to correspond to the identification image. After the dive, the green scrubber with the mucus was removed from the handle, placed into a 50 ml tube containing 40 ml of 98% ethanol and transferred onto ice. A further set of non-invasive mucus samples were collected from recreational sea anglers in both Southampton (27 miles due east) and Weymouth (15 miles due west) in 2017. Between June and October 2017 we used the same swabbing method on non-targeted, accidental by-catch *R. undulata*, which were collected, sampled, and returned following Brownscombe *et al.,* (2017) “best angling practices guide”. For each individual we collected three mucus swabs to increase the probability of successful DNA analysis. Once the samples were in a controlled environment, they were kept at -20°C before returning to the lab where they were stored at -80°C.

DNA extraction, amplification and genotyping

A Bioline (“ISOLATE II Genomic DNA Kit Product Manual,” n.d.) was used to extract DNA from the tissue samples, following the manufacturer’s protocol. DNA was extracted from the swabs using an adapted method with an E.Z.N.A Mollusc DNA Kit (Omega Bio-Tek, Norcross, USA). In addition to the manufacturer

protocol, we added a two stage digestion to obtain the maximum amount of mucus from each sample. Firstly, the 0.5 cm^3 of the scrubber with the most visible mucus was selected, together with the top layer of the remaining scrubber with any further visible mucus. The sections were added to a 1.5 ml tube and left for five minutes in a fume hood to remove the remaining ethanol. Secondly, the ethanol-fixed mucus that had fallen from the scrubber within the 50 ml falcon tube was centrifuged at 5000 rpm at 4˚C for 1 hour to create a mucus pellet. The ethanol was gently poured off the mucus pellet and left for five minutes in a fume hood to remove the residual ethanol. A total of 350 µl of ML1 lysis buffer was added to the falcon tube and vortexed for 15 seconds to re-suspend the mucus pellet into the buffer. This solution was then pipetted into the corresponding 1.5 ml tube containing the cut scrubber and visible mucus, and 25 µl of proteinase K was then added to the tube and digested at 60˚C for five hours, or 37°C overnight, to digest contaminating proteins. Once digested, the DNA was extracted following the original E.Z.N.A Mollusc DNA Kit protocol but with a single elution extended to 10 minutes at 70˚C to maximize yield. DNA extractions were quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Carlsbad, USA) to ensure yields were ≥10 ng/µl. Samples were stored at -20˚C.

A total of 17 *R. undulata* species-specific microsatellite primers were used from Hunter *et al*. (2016) and Fox *et al.* (2018). Two different universal tails were added to the primers and Polymerase Chain Reactions (PCRs) were conducted under a three primer approach (6-FAM or HEX; Blacket *et al*., 2012) to create five multiplexes (Supplementary Table 1.). Reaction volumes (5 µl) consisted of 0.5 µl multiplex primer mix, 1.5 µl Type-it® microsatellite master mix and 2.5 µl double distilled H_2O added to 0.5 µl of the 10-70 ng/µl genomic DNA used for

amplification. Thermal cycler conditions were as follows: initial denaturation at 5 minutes at 95˚C, 35 cycles of 30 seconds at 94˚C, 90 seconds of annealing at 60˚C and 30 seconds of extension at 72˚C, followed by 1 cycle of 30 minutes at 60˚C (Fox et al., 2018). PCR products were visualized on a 1.5% agarose gel (using a Gel Green nucleic acid stain) under a UV light source to confirm successful amplification. Following successful amplification, the products were genotyped using an ABI™ sequencer at the University of Manchester DNA Sequencing Facility with GeneScan™ 500 LIZ™ dye size standard and scored using GeneMapper v.4.0 (Applied Biosystems). Allele scores were checked for user error in Microchecker (van Oosterhout et al., 2004) and Microsatellite Toolkit (Park, 2001).

Genotype statistical analysis: genetic health and structure

We investigated observed (H_o) and expected heterozygosity (H_e) as a measure of genetic diversity for each of the loci, using GenePop on the Web v4.2 (Raymond and Rousset, 1995). To identify sample location diversity levels we collected average H_o, H_e, number of alleles per locus and the number of unique alleles (private alleles) for each sample location using the allele frequencies calculated in GenePop on the Web, and confirmed this with Cervus v3.0.7 (Marshall et al., 1998). To calculate an approximate number of genetic populations (K) we ran STRUCTURE's (Pritchard et al., 2009) systematic Bayesian clustering approach, that applies Markov Chain Monte Carlo (MCMC) estimation at 10,000 repetitions to model the possible number of clusters (K = 1 to 15) at fifteen iterations. This data was inputted into STRUCTURE HARVESTER software (Earl and VonHoldt, 2012) which uses the Evanno method (Evanno et al., 2005) to calculate ΔK, and CLUMPP (Jakobsson and Rosenberg, 2007) as a more accurate predictor of the cluster number. The Evanno method uses the rate of change in the log probability of the data provided from

STRUCTURE to account for non-homogeneous dispersal among populations (ΔK) (Evanno et al., 2005). CLUMPP defines the number of K by implementing three algorithms against the STRUCTURE analysis to align clusters via a membership coefficient (Jakobsson and Rosenberg, 2007).

Lastly, we calculated the effective number of breeders (N_{eb}) as a measure of genetic population size under the molecular co-ancestry method (Nomura, 2008) using software NeEstimator v2.1 (Do et al., 2014). This method provides unbiased estimates of N_{eb} without the need for demographic information, such as age. The molecular co-ancestry method also overcomes issues found in previous methods which may not be suitable to study natural populations of endangered species (Nomura, 2008). We analysed each sample site separately before grouping locations in the North Atlantic (Figure 1.) to gain a better understanding of N_{eb} in this region, which is likely linked with the single site capture mark-recapture site Rays' Repose.

Single Site Capture Mark-Recapture (CMR)

Dataset collection

Between April 2012 and October 2018, photographs were collected from resting *R. undulata* on Rays' Repose by the same group of scuba divers who conducted the genetic sampling. Each dive consisted of up to 3 experienced scuba divers with an average dive time of 51 minutes and maximum dive depth of 17.5 metres (tide dependent) (Supplementary Methods 1.).

Diving was conducted during daylight hours when the rays were found on the seabed in a resting position, consistent with ray behaviour described by Humphries et al. (2017) for other UK skate species (Humphries et al., 2017). The number of individuals photographed on any dive was limited by the dive conditions (such as visibility, tides and currents), individual diver constraints and may have been

influenced by individual rays becoming 'trap-shy' (avoidance of the divers). Full protocols were made to minimise ray disturbance, increasing the probability of retrieving dorsal pattern photographs (Supplementary Methods 1.). Photographs of the dorsal side of the fish were taken in .jpg and RAW format with various compact and single-lens reflex (SLR) cameras. In total, 144 dives were completed on the site, collecting CMR data for 263 individuals.

Computer assisted photo-ID

Each photograph had the colour removed, and was converted into a standard orientation and on-screen size (20-cm x 20-cm) using Adobe® Photoshop (Figure 2.). To assist the computer recognition program, areas of surrounding seabed were cropped from the image to leave only the ray's dorsal surface (Figure 2.). Each formatted photograph was entered into Wild-ID (Bolger et al., 2012). Wild-ID compares each new image and provides a numerical matching coefficient for the 20 most likely existing photos already in the dataset (Bolger et al., 2012). Where the dorsal pattern was clear, the software identified images of the same ray and clearly discriminated from other rays with a higher numerical matching coefficient. However, where the dorsal pattern was obscured, matching images were discriminated less clearly from images of the other rays, hence, the final decision for a true match was made manually by the same two users from the 20 most likely candidates identified by the software. On first capture each individual received a unique sighting number of which all future recaptures would then be associated to. The results from Wild-ID were cross-examined with a second photo-recognition software, I^3S Pattern, revealing the same exact matches between images, and thus validating the use of Wild-ID (Speed et al., 2007).

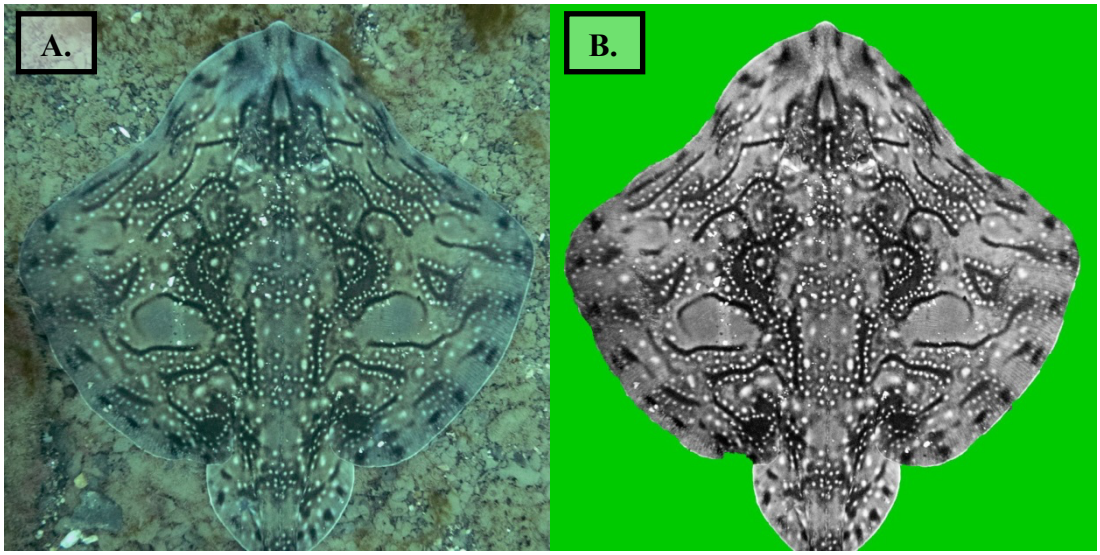


Figure 2. (A) The first cropped image of the dorsal pattern with the original substrate, orientated with the nose at the top of the image, and the tail at the bottom of the image. (B) Image A edited without colour and removal of the surrounding substrate. Both images are of the same individual ray ID 831, named 'Watson'.

Statistical analysis

In 2017, one individual from Rays' Repose was photographed by a recreational diver at Chesil Beach, Portland, Dorset (Ray ID 598), approximately 34 miles, west of Rays' Repose (Openshaw and Openshaw, 2018). This indicates that the population is open, and therefore we conducted the analysis under this assumption. We used RStudio v.1.0.143 (RStudio Team, 2016) to investigate whether the seasons had an influencing factor on the number of rays captured using one-way ANOVA test of equal variances from a Welch's *t*-test. Secondly, we created loglinear models to estimate population abundance between years in the R-package Rcapture (Baillargeon and Rivest, 2007).

To investigate an individual's probability of survival (φ) and recapture (*p*) we used a Cormack-Jolly-Seber model (CJS) in the R-package "marked" (Laake et al., 2013; RStudio Team, 2016). Duration between site visits varied in length due to weather restrictions, tide, and other time constraints. We therefore were interested in the φ

with the variation of time between site visits as a parameter. We also used individual sex as a parameter to determine whether there was a bias on φ and *p* estimates in separate models.

Network analysis

Sampling visits that occurred within a five day period were clustered together to address a potential low probability of recapture, where individuals could be present but may not be observed. The total number of clustered mark-recapture visits was 40 (original number of visits = 73). Using the clustered mark-recapture data we investigated whether the co-occurrence of observed paired individuals appearing on the site exceeded the modelled prediction which could be expected to occur at random. To model our predicted data we used the EcoSimR package (v6.0, (Gotelli et al., 2015) in RStudio v.1.0.143 ((RStudio Team, 2016), which uses the curveball algorithm (Strona et al., 2014) of matrix shuffling to generate 'random' matrices, based on the observed data, whilst maintaining row and column totals (in our case individual and time point respectively). In using the curveball algorithm rather than the more traditional sequential swap, transient effect biases are minimised and therefore the resulting matrices have demonstrably greater reliability (Strona et al., 2014). We ran the EcoSimR algorithm 5 times, with 10,000 iterations each.

Results

Global genetic results

A total of 143 individuals from the 12 sites were amplified using the 17 loci defined by Hunter *et al.* (2016) and Fox *et al.* (2018). Total average genetic diversity measures per locus were H_o = 0.66, H_e = 0.85, and mean number of alleles per locus = 19.8 (Supplementary Table 1.). Further to this, we investigated mean H_o, mean H_e, mean polymorphic information content (PIC), and mean number of private alleles (P_a) for each sample site (Table 1.). We found the largest number of private alleles to be at El Jadida (Morocco, *Pa* = 37), whilst the lowest number was found at Larache and Casablanca (Morocco, *Pa* = 0). Overall the number of private alleles for each country was: England, *Pa* = 34; Portugal, *Pa* = 19; Spain, *Pa* = 17, Morocco, *Pa* = 47.

Table 1. Genetic diversity levels for the samples taken at each site, measured as average observed and expected heterozygosity (Ho and He), Polymorphic Information Content (PIC), number of private alleles (Pa) and estimated effective number of breeders (Neb) and Neb at 95% confidence interval (CINeb)

Sample Site	**N**	H_o	H_e	**PIC**	P_a	N_{eb}	**CI**N_{eb}
Rays' Repose	16	0.141	0.296	0.242	12	-	-
Southampton	20	0.545	0.601	0.551	7	-	-
Weymouth	13	0.553	0.623	0.551	5	-	-
Peniche	8	0.539	0.688	0.597	9	-	-
Algarve	9	0.667	0.791	0.704	10	28.9	80.4
Algeciras	13	0.733	0.798	0.651	17	12.5	21.1
Tangier	5	0.515	0.606	0.489	1	2585.8	12980.8
Larache	2	0.441	0.461	0.294	0	-	-
Casablanca	5	0.549	0.617	0.504	0	133.2	668.5
El Jadida	33	0.678	0.848	0.788	37	21.8	46.1
Essouira	14	0.706	0.808	0.721	7	-	-
Agadir	5	0.500	0.564	0.284	2	44.6	224.0

Population Structure

Through the use of STRUCTURE (Pritchard et al., 2009) and STRUCTURE HARVESTER (Earl and VonHoldt, 2012; Evanno et al., 2005) we found that there were six genetically distinct populations (K) within the sample set (Figure 3.), and that there is a higher level of connectivity between sample locations that are geographically closer together (Figure 3.). From the STRUCTURE results, the Evanno method and the CLUMPP analysis confirmed a *K* of 6 ($K = 6$, $\Delta K = 4.03$, iterations = 15).

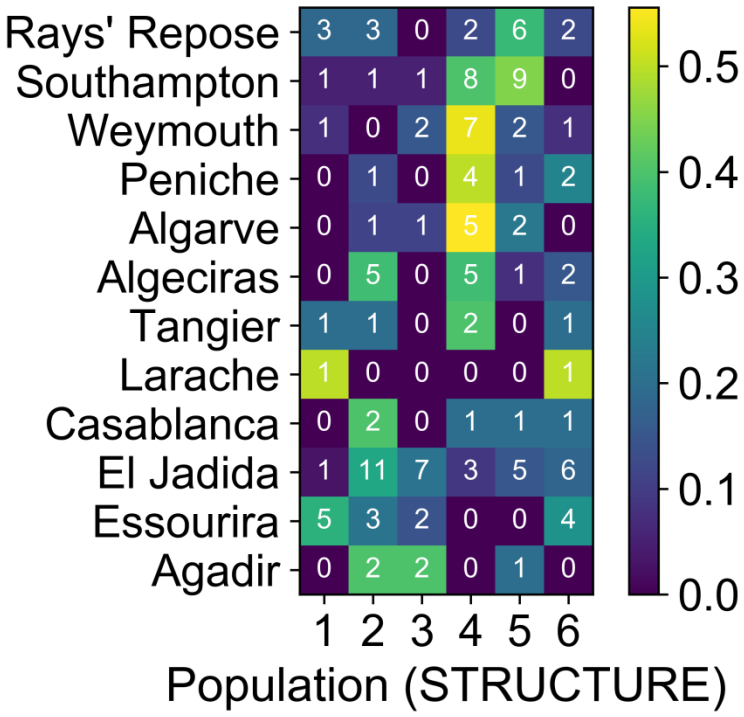


Figure 3. A colour map of the normalized data displaying the number of individuals from each sample location (y) compared with the population assignment *K*= 6 calculated in STRUCTURE and Structure HARVESTER with use of the Evanno method and CLUMPP. Colour bar represents the normalized data (fractions). Sample locations (y) are arranged north to south. The colour map was created using Python program Matplotlib.

Molecular co-ancestry effective number of breeders (N_{eb})

When combining the North Atlantic samples (N = 49) and using the molecular co-ancestry method as defined by Nomura (2008), the estimated number of breeders as

a measure of population size was unknown because the data were not sufficiently informative at a 95% confidence level. This was the case for 50% of the sample sites (Table 1.).

Single Site Capture Mark-Recapture (CMR) Results

Single site capture mark-recapture

In total, we identified 263 individuals that were present on the Ray's Repose site between spring of 2012 and autumn of 2018. Of the 263 individuals, six previously unmarked individuals were photographed on the last visit and therefore were removed from the CJS CMR model. Of the total 263 individuals, 82 were males, 173 were females, seven were juveniles and one was unknown, although these figures may be inaccurate being based only on an underwater visual assessment of maturity and gender.

Population distribution and estimates

The number of rays that were found varied across the seasons, dependent on the year (Figure 4.). Overall the highest number of rays was found in autumn and the lowest number in spring (one-way ANOVA test of equal variance $p = 0.122$, Welch *F*-test of unequal variances $p = 0.050$); months for each season where taken from the UK set dates. From the average number of individuals observed, more rays were present in summer than in any other month; average number of individuals per season was calculated from the number of rays per site visit. Population estimates for Rays' Repose fluctuated between 150 and 400 (maximum standard error +/- 125; Figure 4.) and average estimate for the site was 228 (maximum standard error +/- 87).

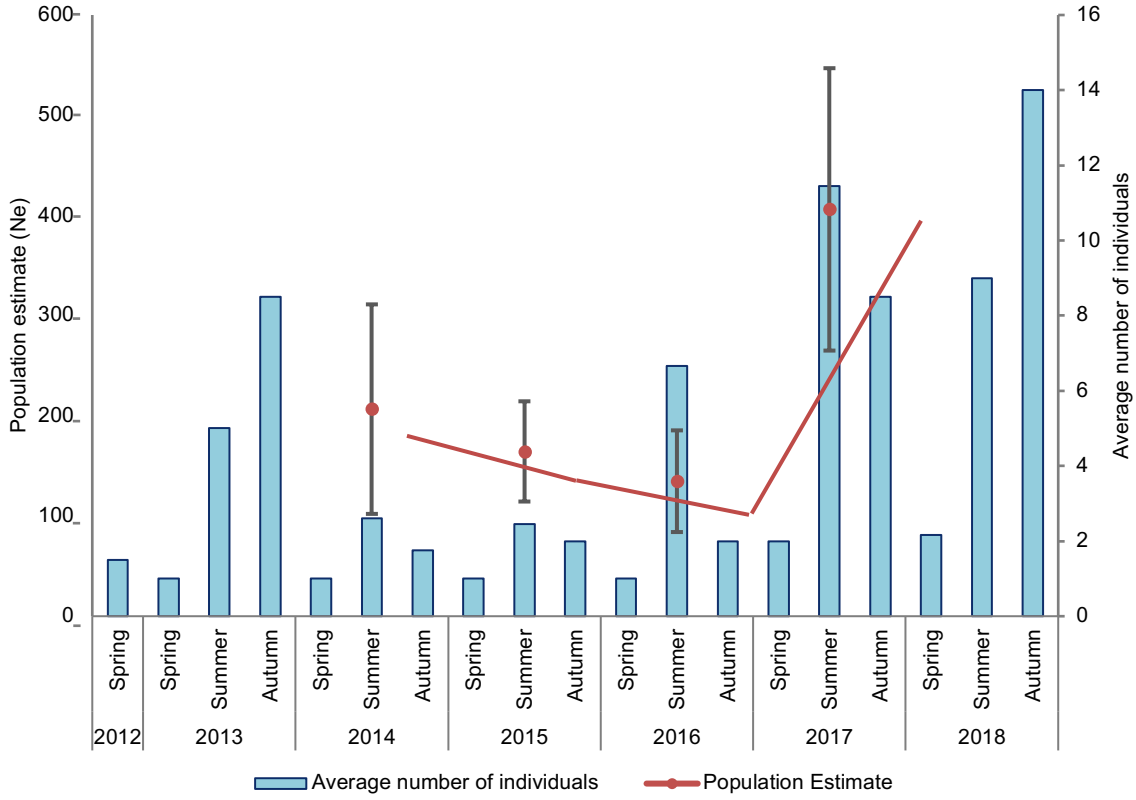


Figure 4. Average seasonal occurrence and population estimates of *R. undulata* at Rays' Repose between spring 2012 and autumn 2018.

Survival and capture probabilities

Using the CJS models under the assumption that time between visits is a variable parameter for φ and *p* we found that the overall estimate of probability of survival φ = 0.969 (standard error (SE) = 0.0054, 95% upper and lower confidence levels = 0.956 to 0.978) and capture probability *p* = 0.035, (SE = 0.004, 95% upper and lower confidence levels = 0.028 to 0.043).

Network analysis

When applying the curveball algorithm matrix reshuffling, the observed data exceeded both the one, and two tailed 95% confidence intervals ($p = \leq 0.02$), indicating that the observed co-occurrence of undulata rays on Rays' Repose is not random. Furthermore, we can see from the network analysis (Figure 5.) that pairs

types are likely to either be female to female (19 pairs) or female to male (15 pairs) rather than male to male (6 pairs).Using probability tests we found no statistical significance between the pairing type, the possible number of pairings (NP) and the actual number of pairs (AP); female to female, NP = 300, AP = 21; female to male NP = 350, AP = 18; male to male, NP = 91, AP =3.

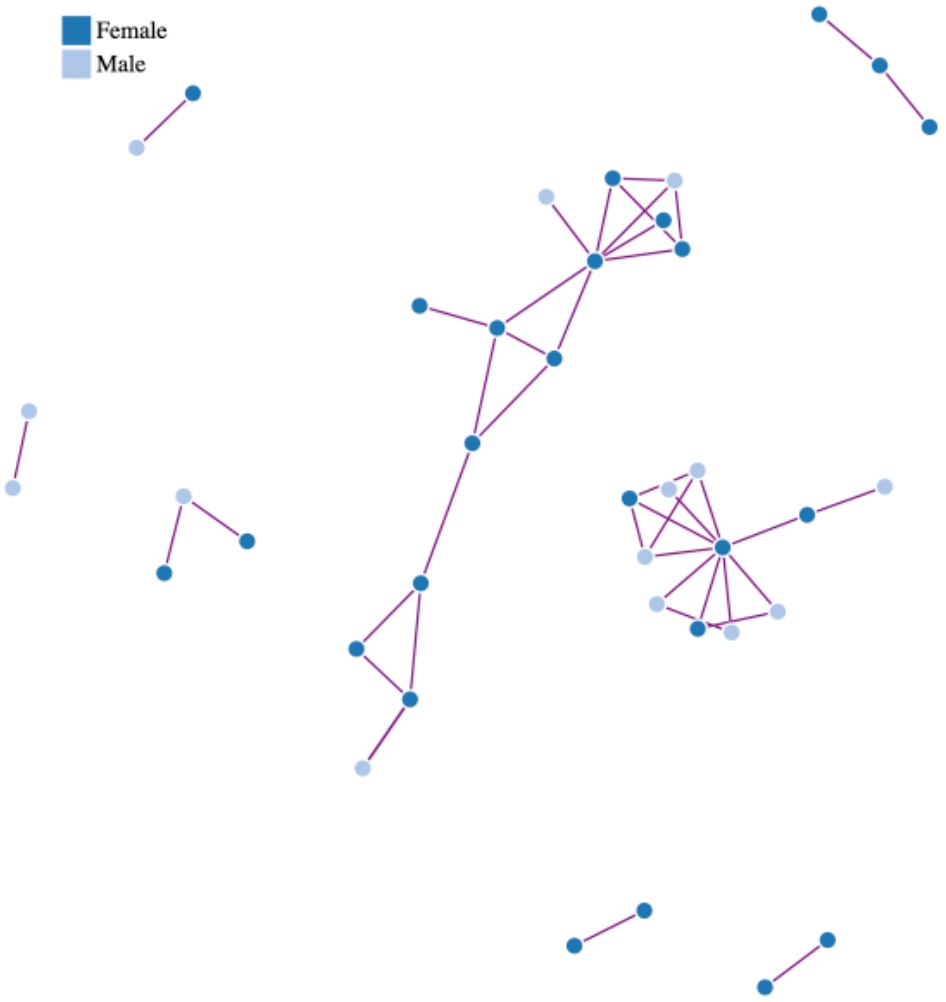


Figure 5. A network analysis showing the networks of pairs which occurred in 2 or 3 clusters together. Total number of pairs = 42, Number of retained individuals = 39.

Discussion

Here, we developed a successful method for non-invasive, underwater mucus sampling to extract DNA and examine the genetic health of *R. undulata*. This is the first study to combine mucus samples (collected by scuba divers and recreational sea anglers) and tissue samples (collected from fish markets) to review the global genetic population structure for any elasmobranch. Lastly, this is the first long term (7 year) capture mark-recapture (CMR) study conducted on any skate or ray (batoid) population, without the use of an invasive tagging method.

Prior to this research, similar mucus sampling techniques have been successfully used on other neotropical batoids (Domingues et al., 2019; Kashiwagi et al., 2015; Lieber et al., 2013) and with captive individuals (Hunter, 2016). With the exception of Kashiwagi *et al.* (2015), these studies have required invasive capture methods where the individual is removed from the water to conduct non-invasive sampling via mucus swabs (usually by fishing). Elasmobranch mortality post-capture can range depending on fishing practice such as gear type, location, fishing depth, species, and on-board conditions (Ellis et al., 2017). In our sample collection, *R. undulata* were non-targeted, accidental by-catch from recreational sea anglers using rod and line, whereby mucus swabs were non-invasively taken prior to release. Sea anglers followed the 'best angling practices guide' (Brownscombe et al., 2017) to increase probability of survivorship post release. Divers used a site-specific protocol adapted from The Underwater Photographers Code of Conduct (The British Society of Underwater Photographers, n.d.) to safely capture images and take mucus samples with little intrusion to the rays' natural behaviour and having no known effect on mortality (Supplementary Methods 1.). To our knowledge, we are the first study to employ mucus sampling method on resting batoids, collected by scuba divers.

We found that, despite *R. undulata* being globally endangered (Coelho et al., 2009), average genetic diversity was overall high (H_o = 0.66, H_e = 0.85, average alleles per locus = 19.8). This is similar to other elasmobranch species threatened with extinction such as the small sawtooth *Pristis pectinate* (Chapman et al., 2011), the scalloped hammerhead *Sphyrna lewini* (H_o = 0.67, H_e = 0.72; (Green et al., 2017; Nance et al., 2009) and the longheaded eagle ray *Aetobatus flagellum* (H_o = 0.47, H_e = 0.48; (Yagishita and Yamaguchi, 2009). No study has investigated the true effects of elasmobranch genetic diversity changes due to population decreases. As elasmobranchs have long life history traits, it may be that we do not observe a decrease in genetic diversity for many generations, as life spans of animals are often longer than the time span of data available. Furthermore, with the exclusion of Rays Repose', we observed little differences between populations that previously had full protection in the north-east Atlantic, and various fished populations in the mid-Atlantic. STRUCTURE defined six genetically unique populations. Rays' Repose individuals fell into numerous populations; however the majority of individuals were clustered within population 5, which also included the majority of Southampton individuals. However, this may be due to a reduced amplification rate from the underwater mucus swabs when compared to the sea angler mucus swabs. Structurally, there is almost a near north to south divide, with the majority of Morocco samples falling in populations 1 to 3 and the majority of northern sample sites falling into populations 4 to 6 (Figure 3.). The fact that nearly all sample sites have individuals which fall into every population identified by STRUCTURE indicates a level of shared genotypes and therefore possible geneflow between regions.

The effective number of breeders could not be determined by the molecular co-ancestry method (Nomura, 2008) for 50% of sample sites because it results in an infinite (∞) estimated N_{eb} (including confidence intervals). For all sites where samples were collected by non-intrusive methods, N_{eb} was unknown because the data is not informative enough. This was the same for only three out of the nine invasive sample sites and could be the product of a lack of population structure. It is interesting to note that N_{eb} could only be calculated in the more tropical regions around the strait of Gibraltar (south Portugal and Spain) into the mid-Atlantic (Morocco). The largest estimate, with the exception of ∞, was at Tangier, Morocco (N_{eb} = 2585.8) whilst the smallest number was in Algeciras, Spain (N_{eb} = 12.5).

Despite the success of using the unique dorsal patterns for individual recognition, certain caveats should be acknowledged. The challenge of photographing rays in their wild environment means that photographs are often lesser quality for identification purposes than what could be achieved in a controlled environment. The computer recognition process is dependent on the quality of the photographs and therefore false negatives may exist. To overcome this, we used a manual matching process to validate the identifications made within Wild.ID, which can be time consuming. We also confirmed Wild-ID with I^3S, indicating both software were able to match individuals. As photographic equipment and recognition software improves over time, we can predict that error rates will decrease.

The frequency at which individuals are caught in CMR studies can depend on the methodology, and may either influence an over- or underestimate of the population size. For example fishing for individuals may cause them to become 'trap shy', while baiting vessels to attract individuals may cause them to become 'trap happy' (attracted to the method of CMR) (Towner et al., 2013). With the exception of 2017

and 2018, individuals on Rays' Repose were only photographed, causing minimal to no disruption while rays were resting. As the *R. undulata* did not receive a known benefit from the divers, it is unlikely that they would become trap-happy; however, any minimal disruption from the divers may have encouraged individuals to move off the site (trap-shy). As divers visit on relatively few occasions (averaged ≤5.6% of the year), if individuals became trap-shy, they would likely return to the site once divers had left, causing minimal disruption to their natural behaviour. This along with sampling effort, tidal conditions, length of dive and underwater visibility could be contributing factors on why the probability of recapture is low ($p = 0.035$). It could be argued that the level of camouflage *R. undulata* have against the seabed at Rays' Repose may have also contributed to the low probability of recapture. This camouflage is likely the reason we observed them resting for such long periods of time, and a contributing factor for such a successful survivorship ($\varphi = 0.969$). A niche habitat coupled with high levels of protection and strict landing quotas since 2009 (ICES, 2016), could have contributed to the near stable population sizes we estimated. The longest period between first and last sighting of the same ray was 2,186 days, approximately six years, highlighting the importance of long term studies and specific sites to individual fish.

Lastly, from our CMR data collected at Rays' Repose, we found the first evidence that there are social interactions between individual *R. undulata*, which may influence their migratory patterns when appearing on the site. The observed patterns of co-occurrence of individuals highlight possible levels of social behaviour not previously explored in *R. undulata*. As the site appears to be used only for resting, it can be questioned whether the networks are moving off site together to conduct the same natural behaviours, such as feeding or mating. As the networks appear to be

primarily female pairings, or female to male pairings, it can be questioned whether there is an active avoidance between males. However further data and research would be required to test true significance between these pairings. Furthermore, it would be interesting to investigate movement and behaviours when individuals are on a different site as we only found evidence of resting on Rays' Repose.

The knowledge of the site, the length of study, and the presence of rays has made this unique in its field. To replicate the CMR study elsewhere, similar conditions would have to be met. With the correct equipment, diver experience and knowledge of batoid resting sites, non-invasive methods of CMR compete with traditional invasive tagging methods as a cheaper and more accessible tool for site-specific population assessments, assisting in conservation management across a range of species. Although the Ray's Repose site is already situated in a protected European Marine Site, the Studland to Portland SAC, undulate rays are not a qualifying designated feature for that site and therefore receive no specific protection. Due to the site's topography, Rays' Repose will not likely be subjected to any bottom-destructive fishing, such as trawling (offering a level of protection) as the fishing gear would be damaged. Despite this, set net fishers do operate in the area, targeting rays and flatfish amongst others; recreational anglers also fish for rays along the coastline. At the present time, continued observation could be the best management for this population before other means such as voluntary codes of conduct or local fisheries by-laws such as bag limits or closed areas need to be considered. The application of these developed sampling techniques to other areas or taxa will increase our knowledge and understanding of elasmobranch populations and behaviour more globally.

Funding

The project was funded by the Natural Environment Research Council, UK (NERC) as part of The University of Manchester Doctoral Training Program. The British Sub Aqua Club Trust Jubilee Trust partially funded some of the diving activities.

Acknowledgements

The authors would like to thank Jennifer Rowntree, Richard Preziozi and John Fitzpatrick for their early encouragement towards undulate ray genetics. We would also like to thank Simon Ruske and Merlin Entertainments Sea Life Aquariums for their advice and support. Lastly, we would like to thank the recreational anglers for their samples.

Author contributions

Samantha A. Hook[1] wrote the paper, collected the invasive DNA samples, analysed the genetic and CMR results and developed the underwater technique, Martin Openshaw[2], Sheilah Openshaw[2] voluntarily collected the capture mark-recapture data and underwater DNA samples, Matthew Doggett[2] collected underwater footage and assisted in CMR collection, Daniel Ripley[4] conducted the population estimates from CMR data, Alexandra Martin-Geary[1] completed the shark network analysis from the CMR, Jean-Denis Hibbitt[5] assisted in UK sample collection, Michael Buckley[1] supervised SAH. All authors contributed to the review prior to submission.

References Hook *et al* 2019

1. Andreotti S, Holtzhausen P, Rutzen M, Meÿer M, van der Walt S, Herbst B, Matthee CA (2018) Semi-automated software for dorsal fin photographic identification of marine species: application to Carcharodon carcharias. Mar Biodivers 48: 1655–1660.
2. Andreotti S, Rutzen M, van der Walt S, Von der Heyden S, Henriques R, Meÿer M, Oosthuizen H, Matthee C (2016) An integrated mark-recapture and genetic approach to estimate the population size of white sharks in South Africa. Mar Ecol Prog Ser 552: 241–253.
3. Ari C (2015) Long-term body pigmentation changes on a manta ray (Mobulidae). Biol J Linn Soc 114: 406–414.
4. Baillargeon S, Rivest L-P (2007) Rcapture: Loglinear Models for Capture-Recapture in R.
5. Beddington JR, Agnew DJ, Clark CW (2007) Current problems in the management of marine fisheries. Science 316: 1713–6.
6. Bolger DT, Morrison TA, Vance B, Lee D, Farid H (2012) A computer-assisted system for photographic mark-recapture analysis. Methods Ecol Evol 3: 813–822.
7. Brownscombe JW, Chapman JM, Gutowsky LFG (2017) Best practices for catch-and-release recreational fisheries – angling tools and tactics. Fish Res 186: 693–705.
8. CEC (2010) CEC (2010). Council Regulation (EU) No 23/2010 of 14 January 2010 fixing for 2010 the fishing opportunities for certain fish stocks and groups of fish stocks, applicable in EU waters and, for EU vessels, in waters where catch limitations are required and . Off J Eur Communities L21: 1–20.
9. Chapman DD, Simpfendorfer CA, Wiley TR, Poulakis GR, Curtis C, Tringali M, Carlson JK, Feldheim KA (2011) Genetic Diversity Despite Population Collapse in a Critically Endangered Marine Fish: The Smalltooth Sawfish (Pristis pectinata). J Hered 102: 643–652.
10. CoC (2015) Compliance Committee: Working Group on Illegal, Unreported and Unregulated (IUU) Fishing in the GFCM Area. Marrakech, Morocco.
11. Coelho R, Bertozzi M, Ungaro. N, Ellis J (2009) Raja undulata. IUCN Red List Threat Species 2009 eT161425A5420694. http://dx.doi.org/10.2305/IUCN.UK.2009-2.RLTS.T161425A5420694.en
12. Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N e) from genetic data. Mol Ecol Resour 14: 209–214.
13. Domingues RR, Garrone-Neto D, Hilsdorf AWS, Gadig OBF (2019) Use of mucus as a non-invasive sampling method for DNA barcoding of stingrays and skates (batoid elasmobranchs). J Fish Biol 94: 512–516.

14. Dulvy NK, Fowler SL, Musick JA, Cavanagh RD, Kyne PM, Harrison LR, Carlson JK, Davidson LN, Fordham S V, Francis MP, et al. (2014) Extinction risk and conservation of the world's sharks and rays. Elife 3: e00590.
15. Earl DA, VonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method.
16. Ellis J., McCully S, Wallis RHL (2015) Raja undulata European Regional Assessment. IUCN Red List Threat Species. https://www.iucnredlist.org/species/161425/48909382 (last accessed 12 March 2019).
17. Ellis JR, McCully Phillips SR, Poisson F (2017) A review of capture and post-release mortality of elasmobranchs. J Fish Biol 90: 653–722.
18. Ellis JR, McCully SR, Brown MJ (2012) An overview of the biology and status of undulate ray Raja undulata in the north-east Atlantic Ocean. J Fish Biol 80: 1057–1074.
19. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14: 2611–20.
20. Feldheim KA, Gruber SH, Marignac JRC, Ashley M V. (2002) Genetic tagging to determine passive integrated transponder tag loss in lemon sharks. J Fish Biol 61: 1309–1313.
21. Fox G, Darolti J, Hibbitt J, Preziosi RF, Fitzpatrick JL, Rowntree JK (2018) Bespoke markers for ex-situ conservation : application , analysis and challenges in the assessment of a population of endangered undulate rays. Jzar 6: 50–56.
22. González-Ramos MS, Santos-Moreno A, Rosas-Alquicira EF, Fuentes-Mascorro G (2017) Validation of photo-identification as a mark-recapture method in the spotted eagle ray Aetobatus narinari. J Fish Biol 90: 1021–1030.
23. Gotelli NJ, Hart EM, Ellison AM (2015) EcoSimR: Null model analysis for ecological data.
24. Green ME, Appleyard SA, White W, Tracey S, Ovenden J (2017) Variability in multiple paternity rates for grey reef sharks (Carcharhinus amblyrhynchos) and scalloped hammerheads (Sphyrna lewini). Sci Rep 7: 1528.
25. Guttridge TL, Gruber SH, Krause J, Sims DW (2010) Novel Acoustic Technology for Studying Free-Ranging Shark Social Behaviour by Recording Individuals' Interactions. PLoS One 5: e9324.
26. Humphries N, Simpson S, Sims D (2017) Diel vertical migration and central place foraging in benthic predators. Mar Ecol Prog Ser 582: 163–180.
27. Hunter E, Buckley A a., Stewart C, Metcalfe J d. (2005) Migratory behaviour of the thornback ray, raja clavata , in the southern north sea. J Mar Biol Assoc United Kingdom 85: 1095–1105.

28. Hunter K (2016) The Development of Molecular Techniques for the Conservation of Captive Elasmobranchs 0–59.
29. ICES (2016) Advice on Fishing Opportunities, Catch, and Effort Celtic Seas Ecoregion.
30. ISOLATE II Genomic DNA Kit Product Manual (n.d.).
31. IUCN (2014) A quarter of sharks and rays threatened with extinction | IUCN. Int Union Conserv Nat. https://www.iucn.org/content/quarter-sharks-and-rays-threatened-extinction (last accessed 13 February 2018).
32. Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801–1806.
33. Kashiwagi T, Maxwell EA, Marshall AD, Christensen AB (2015) Evaluating manta ray mucus as an alternative DNA source for population genetics study: underwater-sampling, dry-storage and PCR success. PeerJ 3: e1188.
34. Kohler NE, Turner PA (2001) Shark Tagging: A Review of Conventional Methods and Studies. Environmental Biology of Fishes.
35. Laake JL, Johnson DS, Conn PB (2013) marked: an R package for maximum likelihood and Markov Chain Monte Carlo analysis of capture-recapture data. Methods Ecol Evol 4: 885–890.
36. Larson SE, Daly-Engel TS, Phillips NM (2017) Review of Current Conservation Genetic Analyses of Northeast Pacific Sharks. Adv Mar Biol 77: 79–110.
37. Lieber L, Berrow S, Johnston E, Hall G, Hall J, Gubili C, Sims D, Jones C, Noble L (2013) Mucus: aiding elasmobranch conservation through non-invasive genetic sampling. Endanger Species Res 21: 215–222.
38. Marshall AD, Pierce SJ (2012) The use and abuse of photographic identification in sharks and rays. J Fish Biol 80: 1361–1379.
39. Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol 7: 639–655.
40. McCrea RS, Morgan BJT (2014) Analysis of Capture-Recapture Data.
41. Nance HA, Daly-Engel TS, Marko PB (2009) New microsatellite loci for the endangered scalloped hammerhead shark, Sphyrna lewini. Mol Ecol Resour 9: 955–957.
42. Nomura T (2008) Estimation of effective number of breeders from molecular coancestry of single cohort sample. Evol Appl 1: 462–474.
43. Openshaw M, Openshaw S (2018) The Undulate Ray Project.org. http://undulateray.uk/
44. Park S (2001) Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection. Ph.D. Thesis. University of Dublin.
45. Pritchard JK, Wen X, Falush D (2009) Documentation for Structure Software: Version 2.3.
46. Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86: 248–249.

47. RStudio Team (2016) RStudio: Integrated Development for R. RStudio.
48. Speed CW, Meekan MG, Bradshaw CJ (2007) Spot the match – wildlife photo-identification using information theory. Front Zool 4: 2.
49. Strona G, Nappo D, Boccacci F, Fattorini S, San-Miguel-Ayanz J (2014) A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. Nat Commun 5: 4114.
50. The British Society of Underwater Photographers (n.d.) The Underwater Photographers Code of Conduct.
51. Towner A V, Wcisel MA, Reisinger RR, Edwards D, Jewell OJD (2013) Gauging the Threat: The First Population Estimate for White Sharks in South Africa Using Photo Identification and Automated Software. PLoS One 8: 66035.
52. van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4: 535–538.
53. van Rossum G (1995) Python Software Foundation- Python Language v2.7.16.
54. Worm B, Davis B, Kettemer L, Ward-Paige CA, Chapman D, Heithaus MR, Kessel ST, Gruber SH (2013) Global catches, exploitation rates, and rebuilding options for sharks. Mar Policy 40: 194–204.
55. Yagishita N, Yamaguchi A (2009) Isolation and characterization of eight microsatellite loci from the longheaded eagle ray, Aetobatus flagellum (Elasmobranchii, Myliobatidae). Mol Ecol Resour 9: 1034–1036.