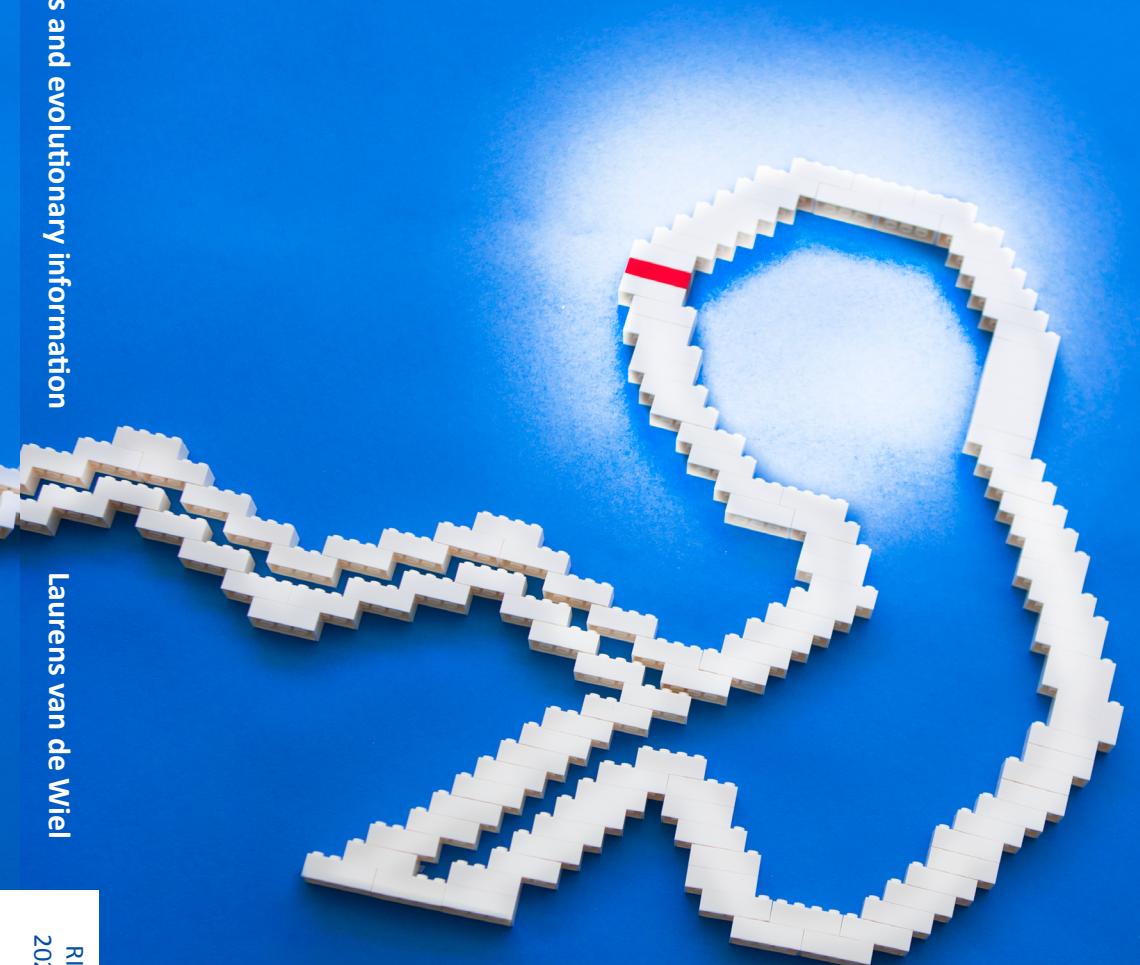


# Interpreting genomic variation using protein structures and evolutionary information

*Laurens van de Wiel*



## Propositions

1. Genetic tolerance indicates parts of a protein that are not important for function. (*This thesis*)
2. Most disease-causing missense mutations are found in protein domains. (*This thesis*)
3. Damaging missense mutations cluster in the 3D protein structure and can provide insights into disease-mechanisms. (*This thesis*)
4. Missense mutations of unknown clinical significance that cluster are damaging. (*This thesis*)
5. Damaging missense mutations predict damaging effects at equivalent locations in other proteins. (*This thesis*)
6. Genomic data growth shall uncover increasingly complex concepts that will require easily-accessible and user-friendly interfaces. (*This thesis*)
7. Identification of increasingly rare genetic disorders will require increasingly large, international, and interdisciplinary collaborations. (*This thesis*)
8. Genetic tolerance will take decades to reach saturation, if ever. Meta-domains can help reach this saturation sooner. (*This thesis*)
9. "As ge niks makt, makte ok niks kapot" // "If you never attempt anything, you will never break anything". (Oma Lies van de Wiel-van Moorsel)
10. "Life is too short to drink bad beer". (Derivative of a quote by Johann Wolfgang von Goethe)

# **INTERPRETING GENOMIC VARIATION USING PROTEIN STRUCTURES AND EVOLUTIONARY INFORMATION**

LAURENS VAN DE WIEL

The work presented in this thesis was carried out within the Radboud Institute for Molecular Life Science, at the Department of Human Genetics and at the Center for Molecular and Biomolecular Informatics of the Radboud university medical center in Nijmegen, The Netherlands.

ISBN: 978-94-6416-647-7

Cover Design: RuudRocks photography | [www.ruudrocks.com](http://www.ruudrocks.com)

Lay-out: Publiss | [www.publiss.nl](http://www.publiss.nl)

Print: Ridderprint | [www.ridderprint.nl](http://www.ridderprint.nl)

© 2021, Laurens van de Wiel | [www.wiel.science](http://www.wiel.science)

All rights reserved. No part of this publication may be reproduced, stored in retrieval systems or transmitted in any form or by any means, electronic mechanical, photocopying, recording, or otherwise, without prior permission.

# **Interpreting genomic variation using protein structures and evolutionary information**

## **Proefschrift**

ter verkrijging van de graad van doctor  
aan de Radboud Universiteit Nijmegen  
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,  
volgens besluit van het college van decanen  
in het openbaar te verdedigen op  
woensdag 25 augustus 2021 om 14:30 uur precies

door

Laurentius Johannes Maria van de Wiel

geboren op 3 mei 1988  
te Oss

**Promotoren:**

Prof. dr. ir. J.A. Veltman

Prof. dr. G. Vriend

**Copromotor:**

dr. C.F.H.A. Gilissen

**Manuscriptcommissie:**

Prof. dr. D.J. Lefeber (Voorzitter)

Prof. dr. J. Heringa (Vrije Universiteit Amsterdam)

Prof. dr. L.H. Franke (Universitair Medisch Centrum Groningen)

# **Interpreting genomic variation using protein structures and evolutionary information**

## **Doctoral thesis**

to obtain the degree of doctor  
from Radboud Universiteit Nijmegen  
on authority of the rector magnificus prof. dr. J.H.J.M. van Krieken,  
according to the decision of the council of deans  
to be defended in public on  
Wednesday, August 25, 2021 at 14:30 hours

by

Laurentius Johannes Maria van de Wiel

**Supervisors:**

Prof. dr. ir. J.A. Veltman

Prof. dr. G. Vriend

**Co-supervisor:**

dr. C.F.H.A. Gilissen

**Doctoral Thesis Committee:**

Prof. dr. D.J. Lefeber (Chair)

Prof. dr. J. Heringa (VU University Amsterdam)

Prof. dr. L.H. Franke (University Medical Center Groningen)

*"We sift over our fingers the first grains of this great outpouring of information and say to ourselves that the world be helped by it. The Atlas is one small link in the chain from biochemistry and mathematics to sociology and medicine."*

-

Margaret Oakley Dayhoff (1968)  
on the first Atlas of Protein Sequence and Structure



## **Contents**

<b>1. General Introduction</b>	<b>13</b>
What makes a protein?	15
How does the genome relate to the protein structure?	21
How can changes in the genome affect proteins?	22
What can we learn from evolution?	26
Scope of this thesis	34
<b>2. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics</b>	<b>37</b>
<b>3. MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains</b>	<b>59</b>
<b>4. Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorder-associated genes</b>	<b>77</b>
<b>5. Evidence for 28 genetic disorders discovered by combining healthcare and research data</b>	<b>93</b>
<b>6. De novo mutation hotspots in homologous protein domains point to new candidate developmental disorder genes</b>	<b>111</b>
<b>7. Discussion</b>	<b>129</b>
Data integration is important for DNA variant interpretation	130
The completeness of genetic variation	130
The limitations of meta-domains	132
The future of meta-domains	136
<b>Appendix</b>	<b>143</b>
Bibliography	144
Statement on FAIR research data management	158
<b>Addendum</b>	<b>163</b>
Summary	164
Samenvatting	166
Curriculum vitae	169
Acknowledgements	176



*"Bilbo: Can you promise that I will come back?  
Gandalf: No. And if you do... you will not be the same."*

-

The Hobbit: An Unexpected Journey (2012)  
conversation between Bildo and Gandalf  
before Bilbo embarks on this adventure



# 1

## **Chapter 1**

---

General introduction

Proteins are fascinating, large, complex molecular machines that have developed over millions of years of evolution. Without proteins, life as we know it would not exist. Proteins are the work horses of the body. Antibodies recognize and bind to viruses or bacteria to protect the host. Enzymes trigger chemical reactions and assist in chemical processes. Messengers signal between cells. Structural components provide structural integrity and support for cells. Transport proteins assist in carrying chemical elements and molecules in and to other cells. Proteins are responsible for the structure, function, and regulation of all critical processes in every form of life. Life, however, is faced with constant selective pressures. These selective pressures are the drivers of natural selection. Given enough time and iterations, they lead to diversification of species in a process that is called evolution.<sup>1</sup> Evolution on a molecular level occurs in the form of mutations that could have a structurally altering effect on proteins. Protein structural changes can directly affect the protein function. These changes are damaging when they drastically disrupt the protein function and can result in reduced fitness, disease, or, death of the host. Selective pressures favour changes that lead to higher fitness. Most variations are neutral to fitness,<sup>2</sup> which resulted in the evolution of many 'optimally enough' proteins suited for a certain task. Identifying which changes are neutral and which are damaging is one of the key challenges in modern-day genetics and also the main motivation for this thesis.

The completion of the Human Genome Project in 2003 gave a boost to the now approximately 22,300 protein-coding genes that have been identified in humans.<sup>3-5</sup> In the almost two decades that followed, a massive accumulation of human genetic data have become publicly available.<sup>6</sup> These genetic data have allowed scientists to look at a fine scale of possible variations to protein-coding genes within a single species. Of all disease-causing genetic variation discovered to date, 58% alters or impairs the protein structure.<sup>7</sup> The accumulation of genetic information from a multitude of human individuals have led to notions of 'tolerated genetic variation': variation that occurs in high-frequency in the general population and are therefore likely harmless.<sup>8-11</sup>

Despite these vast resources, it remains a challenge to predict if genetic variation is damaging. Small changes in the genome can have a major effect on a protein's structure and thus function. To begin to understand why, it is crucial to first learn how proteins are constructed.

## What makes a protein?

Proteins consist of hundreds to thousands of smaller units called amino acids. All amino acids contain an amino ( $\text{NH}_2$ ) group and a carboxyl ( $\text{COOH}$ ) group. The amino group can bind via a peptide bond to the carboxyl group of another amino acid to form a dipeptide. To form a protein, multiple amino acids are chained together in a polypeptide. The first residue in a polypeptide is called the N-Terminus, and the last residue is called the C-Terminus. When represented in the form of letters a polypeptide is called a protein sequence, or the primary protein structure (**Figure 1**).<sup>12</sup>

There are 22 different proteinogenic amino acids, each commonly denoted by a unique 1-, or, 3-letter combination (A/Ala, C/Cys, D/Asp, E/Glu, F/Phe, G/Gly, H/His, I/Ile, K/Lys, L/Leu, M/Met, N/Asn, O/Pyl, P/Pro, Q/Glu, R/Arg, S/Ser, T/Thr, U/Sec, V/Val, W/Trp, Y/Tyr). Every amino acid has the same neutral backbone and a characteristic side-chain (or R-group). The side-chain determines the amino acid type and has a unique set of different structural and chemical properties.

## The importance of side-chains

The side-chains determine the amino acid type. The properties of side-chains shape the protein, and these properties can be of structural or chemical nature. These properties play an especially important role in the folding of the primary protein structure into a tertiary structure. Side-chain features that are particularly important for the structural formation or function of the protein are the size, electrical charge, presence of a reactive sulphur atom, ability to form salt bridges, overall atomic rigidity, and, hydrophobicity.

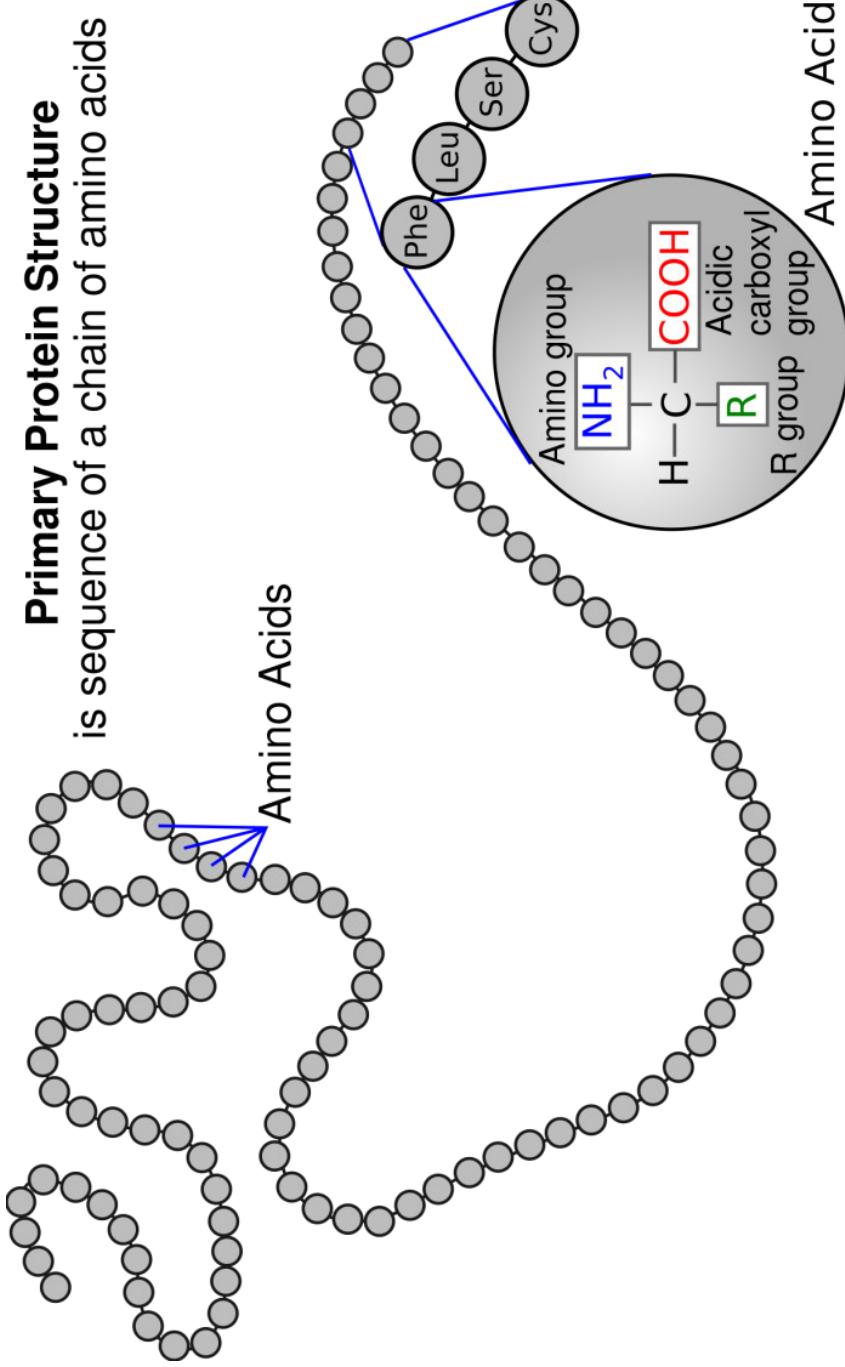
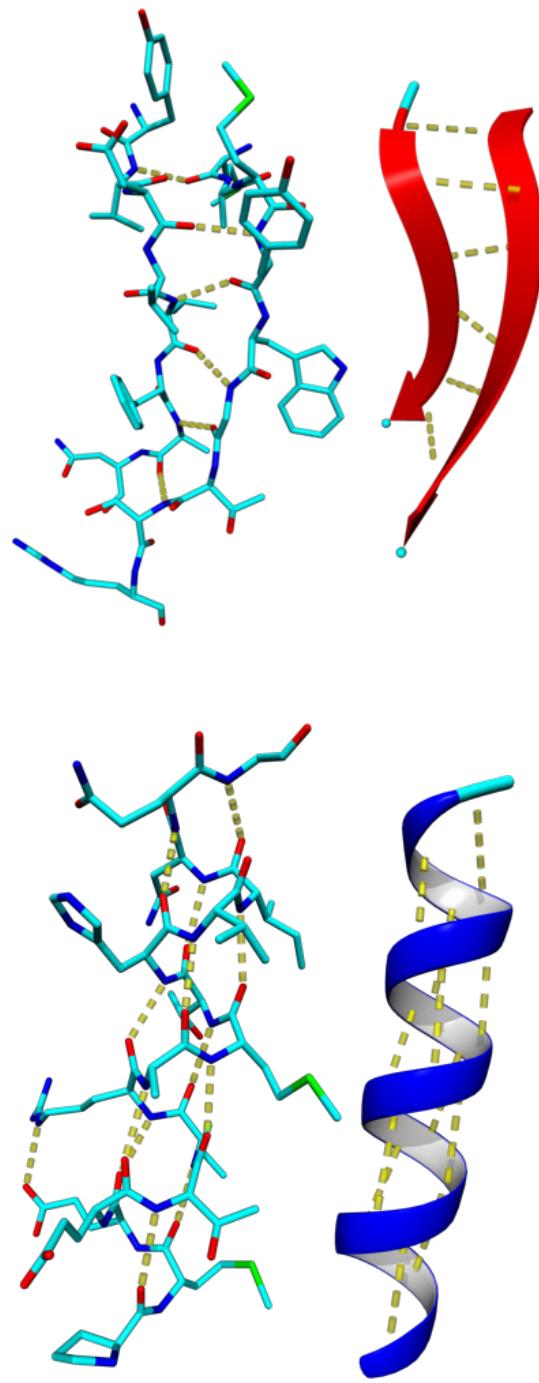


Figure 1. (Image adapted from the National Human Genome Research Institute – Genome.gov).

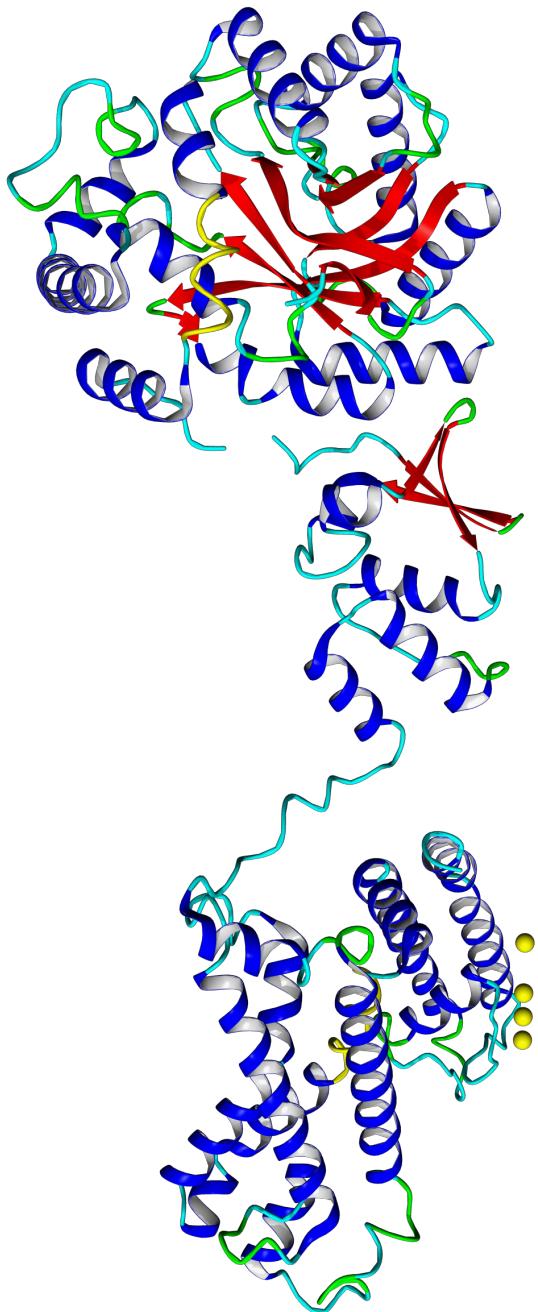
Protein folding is the process in which a polypeptide chain conforms into a 3-dimensional molecule: the tertiary structure. The tertiary structure shape is determined by the environment and the chemical and structural properties of amino acids in the polypeptide. The tertiary protein structure consists of three generic patterns,  $\alpha$ -helices,  $\beta$ -sheets, and, loops. These generic patterns are called secondary protein structures (**Figure 2A**). The type of secondary protein structure is influenced by the forming of hydrogen bonds between amino acids.  $\alpha$ -helices are right-hand-coiled structural conformations that consist of a multitude of repetitive patterns: four amino acids, wherein each first and last residue forms a hydrogen bond using their backbone.  $\beta$ -sheets consists repeated stretched of 3 to 10 amino acids, called  $\beta$ -strands, that are interconnected via hydrogen bonds and assisted by loops and turns.

The tertiary protein structure (**Figure 2B**) is the native conformation of a single polypeptide chain. If multiple polypeptide chains are involved to form a shared conformation, it is called a quaternary protein structure (**Figure 2C**). In the components of the quaternary protein structure are not held together by covalent bonds. Instead they are bound by hydrophobicity, salt bridges, or, disulphide bridges, to name a few. The forming of a quaternary protein structure is also directly influenced by the side-chain properties of the amino acids. Quaternary protein structures that are formed by multiple proteins are commonly referred to as polymers, with 1 = monomer, 2 = dimer, 3 = trimer, etc. And, in the case of dimers or larger polymers, homo- or hetero- prefixes indicates if the quaternary structure is made from identical (homo) or different (hetero) polypeptides. In **Figure 2C** an example of a homo-tetrameric protein structure is provided. In this tetrameric conformation, four identical protein structures join together to form the pore-like structure necessary for channelling  $K^+$  ions. All of the structural examples in **Figure 2** are taken from a mammalian voltage-gated  $K^+$  channel in an inactivated state (PDB: 5WIE<sup>13</sup>). This particular protein structure was used to model and analyse mutation hotspots in **Chapter 6**.

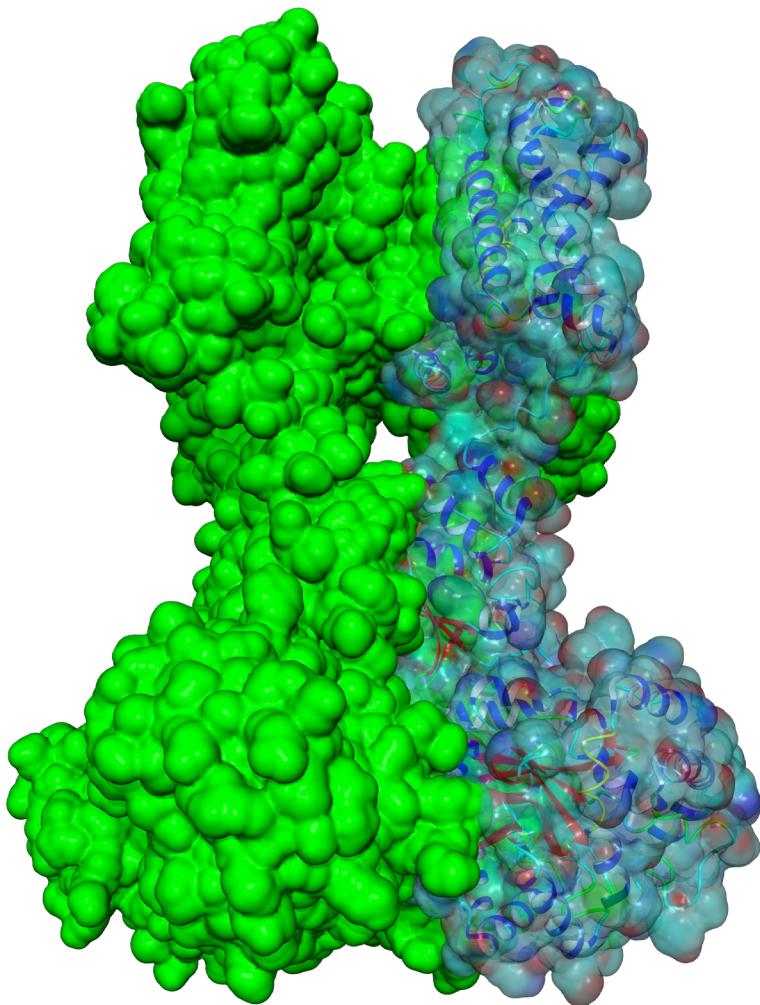
**Figure 2.**  
Parts from the same crystal structure (PDB: 5WIE).<sup>13</sup> (Images were created using YASARA<sup>14</sup> modeling software)



**A.** The structures on the left are representations of an  $\alpha$ -helix (*p*.*Pro*165-*p*.*Gly*180 in PDB: 5WIE-A). The structures on the right are representations of a  $\beta$ -sheets (*p*.*Tyr*151-*p*.*Asn*15 and *p*.*Met*181-*p*.*Arg*189 in PDB: 5WIE-A). The top are represented as atomic stick (carbon: light blue, nitrogen: dark blue, sulphur: green, oxygen: red, hydrogen bonds: yellow dotted lines). The bottom representations are ribbon cartoons, with blue ribbons as  $\alpha$ -helices and red ribbons the  $\beta$ -strands that form the  $\beta$ -sheet.



**B.**  
The complete solved crystal structure of a mammalian voltage-gated  $K^+$  channel in an inactivated state (chain A in PDB:5WE). The yellow balls on the top right are potassium ions moving through the channel.



C

The homo-tetrameric protein structure with the solvent-accessible surface indicated blobs. The visible structure within the blobs is the structure from 2B. This tetramer is a quaternary protein structure and consists of a four-time duplication of the same protein. This quaternary structure is a pore through which ions travel. This particular structure is natively located in the transmembrane of a cell.

## How does the genome relate to the protein structure?

The genome is the collection of all genetic information necessary for the building, maintaining, and, reproduction of organisms. It is passed from parents to offspring. In cellular organisms, like humans, every cell has a copy of the genome. The genome is contained in multiple large molecules that are called chromosomes. The chromosomes are composed of Deoxyribonucleic acid (DNA) molecules. DNA consists of even smaller molecules called nucleotides which are chained together in the shape of a double helix. There are four different nucleotides (A, C, T, G) and each nucleotide is paired with another nucleotide to form base pairs that constitute the double helix shape.<sup>15</sup> Similar to the primary protein structure, where the sequence consists of amino acids, the DNA can be represented as a sequence of letters corresponding to the nucleotides. The human genome consists of 23 chromosome pairs, totalling to 46 chromosomes. Half of these are inherited from the father and the other half from the mother. Combined, the chromosomes contain approximately 6 billion base pairs. Potentially, a change to any one of these 6 billion base pairs can influence the entire organism. In the human genome most of the essential information is located in regions that are called genes. A recent assessment of the human genome identified 60,669 different genes, of which 32.9% are protein-coding, 42.1% non-coding RNA genes, and, 24.3% pseudogenes.<sup>16</sup> The protein-coding genes make up roughly 1-2% of the entire genome.<sup>3</sup> They encode the amino acid arrangement of every protein in human cells.

### Protein-coding genes are blueprints

Protein-coding genes describe how to construct a primary protein structure via sets of instructions. These genes ensure the consistency of how proteins are composed throughout all cells of an organism. The genomic structure of protein-coding genes in eukaryotes consists of regulatory sequences and the open reading frame. The regulatory sequences consist of enhancers, silencers, promoters and the 5' and 3' untranslated regions (UTR). These parts of the protein-coding genes primarily regulate the expression level of proteins. Additionally, they contain instructions for isoforms in the form of transcripts. These isoforms are alternative protein sequence conformations. According to GENCODE there are 84,068 possible transcripts for the 19,959 curated human protein-coding genes (GENCODE Release Version 34).<sup>16</sup> In theory, these transcripts could each result

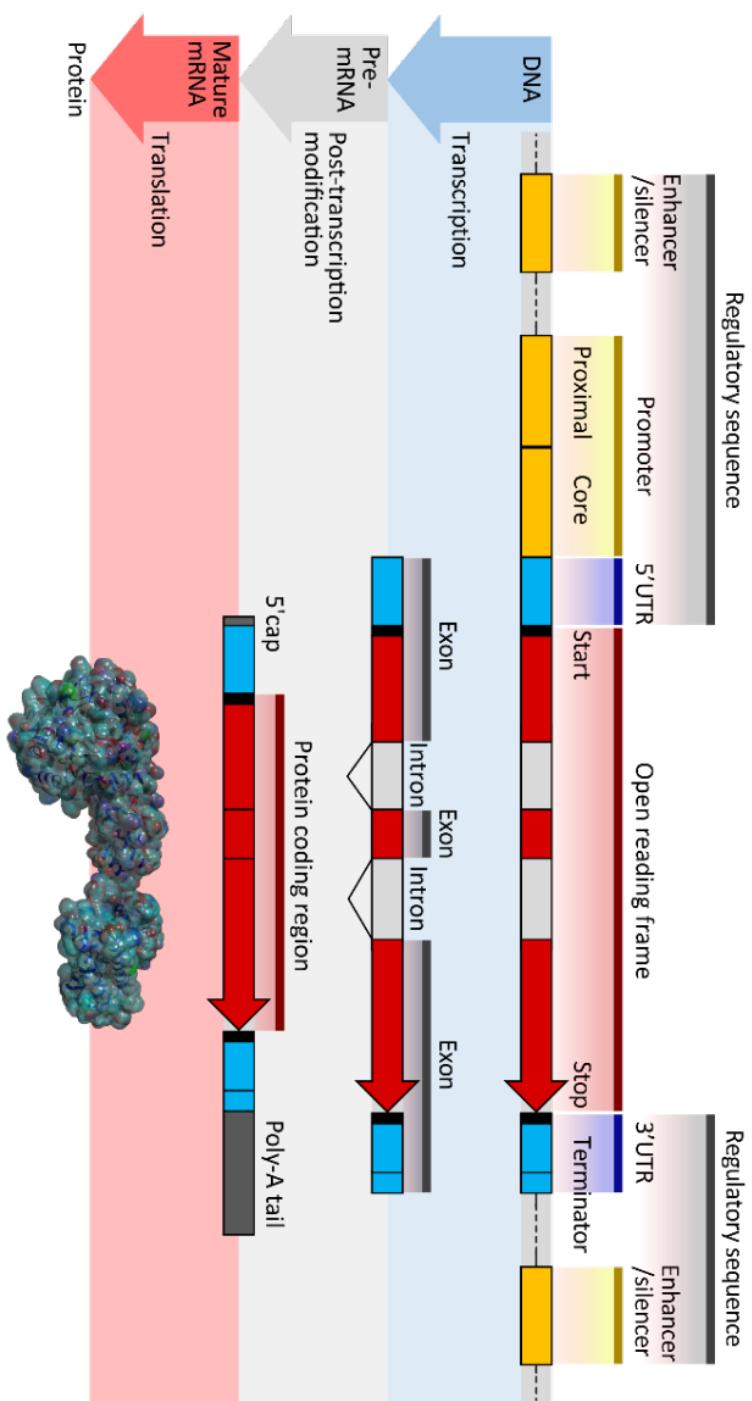
in a different protein sequence. However, most differences between transcripts are in the non-coding UTR regions and will therefore not affect the final protein sequence.<sup>17</sup>

The open reading frame is composed of regions called introns and exons. Introns are non-coding and important for isoform formation and the protein expression level. Exons code for parts of the amino acid sequence via triplets of nucleotides called codons. Each codon directly corresponds to one of 20 amino acids or indicate the termination of the coding region via a 'stop-codon'.<sup>18</sup> The amino acid sequence is constructed from a protein-coding gene with three steps called "central dogma of molecular biology". Protein folding could be seen as the final step (**Figure 3**):<sup>19</sup>

1. Transcription: The 5'UTR, the introns and exons and 3'UTR are transcribed into precursor messenger RNA (pre-mRNA). In this step DNA, with the help of ribosomes, is copied into an RNA representation.
2. Post-transcriptional modification: The intronic regions are removed from the pre-mRNA, this way the exons form the complete, untranslated, protein sequence in RNA, which is called mature messenger RNA (mRNA).
3. Translation: the mRNA is translated into a chain of amino acids (a polypeptide).
4. Protein folding: The polypeptide chain conforms into the tertiary protein structure.

## How can changes in the genome affect proteins?

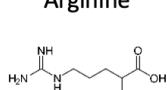
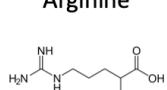
Genetic variations are alterations to the nucleotide mark-up of the genome. These variations can affect only one nucleotide (e.g. transitions), one or a stretch of nucleotides (e.g. insertions and deletions also called indels, or substitutions), or affects a region of nucleotides (e.g. structural variations). Structural variations can be deletions, insertions, inversions, duplications, or, copy number variations. If any of these variations occur within the region of a protein-coding gene, they may have a direct effect on the protein.



**Figure 3.**  
The genomic structure of a protein-coding gene and the different steps to form the primary protein structure. (Image courtesy of Thomas Shafee, adapted from Wikimedia and licensed under Creative Commons CC BY-SA 4.0; Image of the protein structure was created using YASARA<sup>14</sup> modeling software for PDB: 5WIE<sup>15</sup>).

Most of the work in this thesis is focused on single nucleotide variants (SNVs) that occur within protein-coding regions. There again is specific jargon for different SNVs. If an indel SNV in the coding region affects the reading frame of codons it is called a frameshift variation, and, can result in an entirely different protein sequence. Substitution SNVs can have multiple effects on the protein. If the substitution does not change the amino acid translation it is called synonymous, otherwise it is called missense. When the translation is changed to a stop codon, it is called nonsense or stop-gained (**Table 1**). Nonsense and missense variants are also referred to as non-synonymous variations.

**Table 1.** Example of single nucleotide variants in codons and the effect on encoding. (Structural formula representations courtesy of NEUROtiker, adapted from Wikimedia and are licensed under public domain).

	No variation	Synonymous	Nonsense	Missense
Codon	CGA	CG <u>G</u>	<u>T</u> GA	<u>G</u> GA
Translation effect	Arginine 	Arginine 	Stop-introduced No amino acid	Glycine 

## How genetic variations in protein-coding genes can result in disease

The amino acid composition of proteins is encoded in protein-coding genes, and, therefore, the genetic code plays an important role in dictating the composition of a protein. Genetic variations may affect proteins in a positive, neutral, or, negative way. Positive and negative changes can alter the protein in a loss-of-function (LoF) or a gain-of-function (GoF) effect. A variant with a negative effect is called damaging or deleterious. If the damaging variant leads to disease, it is called a pathogenic or disease-causing mutation.

Nonsense variants generally have the largest effect on the protein structure. These variants induce the termination of the open reading frame. The result may be a partial structure, that is often ‘cleaned up’ by a process called nonsense mediated decay (NMD). If the partial protein is cleaned up by NMD there is no protein expressed at all.<sup>20</sup> This can affect the protein expression level also called

the dosage. Disease may occur due to this lack of dosage, and, if this is the case, the mechanism of disease is called haploinsufficiency (HI). The effect of a missense variant greatly depends on the location of that variant in the protein structure and the difference between the original amino acid residue and the one it changes into. If the residue introduced disrupts the folding of the protein structure, the structure could also be cleaned up in the NMD process. Therefore, missense variants may trigger a HI disease-mechanism. Alternatively, damaging missense variants that do not disrupt folding may still disrupt the function, or functional sites, of the protein. If this leads to disease, the disease-mechanism is called non-haploinsufficiency (NHI). Determining if variants are damaging, and how, can require the need for functional testing and replications studies, and, therefore is often a laborious task. In **Chapter 4 and 6** we show that clustering of missense variants found in patient with neurodevelopmental disorders indicate a likely disease-mechanisms and help identify candidate disease-genes.

### Identifying genetic variations in a diagnostic setting

In the two decades following the completion of the Human Genome Project, the technology involved in analysing the human genome advanced immensely. The Human Genome Project provided the first version of the human reference genome.<sup>3,4</sup> The reference genome can be used to identify genetic variation. Genetic variations are differences in nucleotide composition of a patient compared to the reference genome. To find these differences, whole exome sequencing (WES)<sup>21</sup> or whole genome sequencing (WGS)<sup>22</sup> can be used. A patient undergoing WES or WGS will result in many small genome sequence pieces that are called reads. These reads are then mapped to the reference genome. The total number of mapped reads at the same location indicates the quality and certainty of any genetic variants that are identified at that location. Nowadays, whole exome sequencing and whole genome sequencing are part of routine diagnostic protocols.<sup>23,24</sup> Since the first version of the human reference genome, disease-gene associations have increased by a four-fold.<sup>25</sup>

The first step in a present-day genetic diagnostic procedure is to identify all genetic variation in a patient. The second step is variant effect prediction. In a diagnostic setting the goal for variant effect prediction is to find the variant, or variants, that explain the phenotype of the patient. Typically each sequenced

individual has between 20,000 and 26,000 genetic variants in the coding regions, which can be reduced to 150-500 candidate variants by various filtering strategies.<sup>26</sup> This commonly includes considering variants that alter the protein-coding region, are rarely encountered in the general population, are located in a previously disease-associated genomic regions, or, are present in genes that have a specific biological role. Computer-aided variant effect predictors have evolved over the last two decades as well. Deleteriousness predictors, such as SIFT<sup>27</sup>, Polyphen-2<sup>28</sup> and CADD<sup>29</sup>, make use of an aggregate of information resources and proven metrics to determine the likelihood of a variant to have a deleterious effect. HOPE<sup>30</sup> attempts to explain the functional effect of a missense variant in the protein structure. Despite these predictors, it remains challenging to accurately diagnose patients. Another way to gather evidence for diagnosis is to combine genetic data from patients. In **Chapter 5** we combined genetic data from 31,058 patients with developmental disorders. By combining this data, we found 285 genes significantly enriched with rare mutations. Of these, 28 genes were not yet associated to developmental disorders.

## What can we learn from evolution?

The selective pressures that drive evolution induce changes in the genome. These changes may have an effect on the protein structure and function. Given enough iterations these changes enable diversification into different species.<sup>31</sup> The effects of evolutionary-driven genetic variations on genomes are an active topic for scientific studies. Changes that occurred only a short while ago, or hundreds, or millions of years ago can be traced back by sequence analysis. There are many ways to approach this resource of information. For example, these data help estimate how, and when exactly, species diversified by constructing genome-based phylogenetic trees.<sup>32,33</sup> From a shorter time-perspective these data help in uncovering history of human geological migration patterns.<sup>34</sup> Or these data help explain why certain African populations carry a disease-enabling copy of the gene that causes sickle cell anaemia, as it offers protection against malaria.<sup>35</sup>

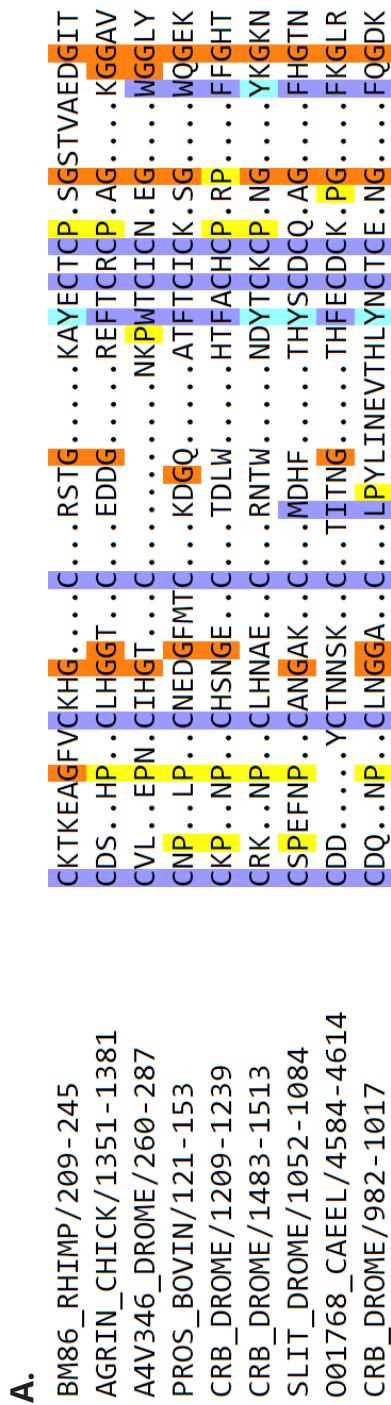
Genetic changes can be damaging, neutral, or, beneficial. There are many possible exceptions and it is difficult to identify which is which. The genome is so complex that not every change will have an everlasting negative or beneficial effect for the following generations. Instead most changes are expected to be neutral.<sup>2</sup> The most

common way to predict likely damaging changes is by evolutionary conservation.<sup>36</sup> Evolutionary conservation can be computed by comparing lack of changes between highly similar proteins from different species.<sup>37</sup> Popular pathogenicity predictors make use of evolutionary conservation (e.g. SIFT<sup>27</sup>, Polyphen-2<sup>28</sup> and CADD<sup>29</sup>). The underlying assumption of evolutionary conservation is that there are a great number of iterations needed to diversify into different species. If the residues at equivalent proteins rarely change during this diversification, then they are probably important. On the other hand, if these residues change often, they are likely neutral.

Highly similar sequences are necessary to compute evolutionary conservation. The *de facto* standard to find analogy in sequences is the basic local alignment search tool (BLAST).<sup>38</sup> BLAST requires an input sequence and then scores sequences based on the similarity to that input sequence. Analogy is often an indication of homology. Similar sequences (>25% sequence identity) can indicate a shared evolutionary ancestor and are called homologous.<sup>37</sup> Homologous relationships can accommodate the transfer of information, and help elucidate important residues and regions within sequences. Transfer of information can be achieved via sequence alignment or multiple sequence alignment (MSA). Sequence alignments are generally made on similar sequences via Clustal<sup>39</sup>. MSA allows nucleotides or amino acids to be aligned to corresponding positions (**Figure 4A**). In homologous proteins, mutations at corresponding locations across an MSA are known to result in similar effects.<sup>40</sup>

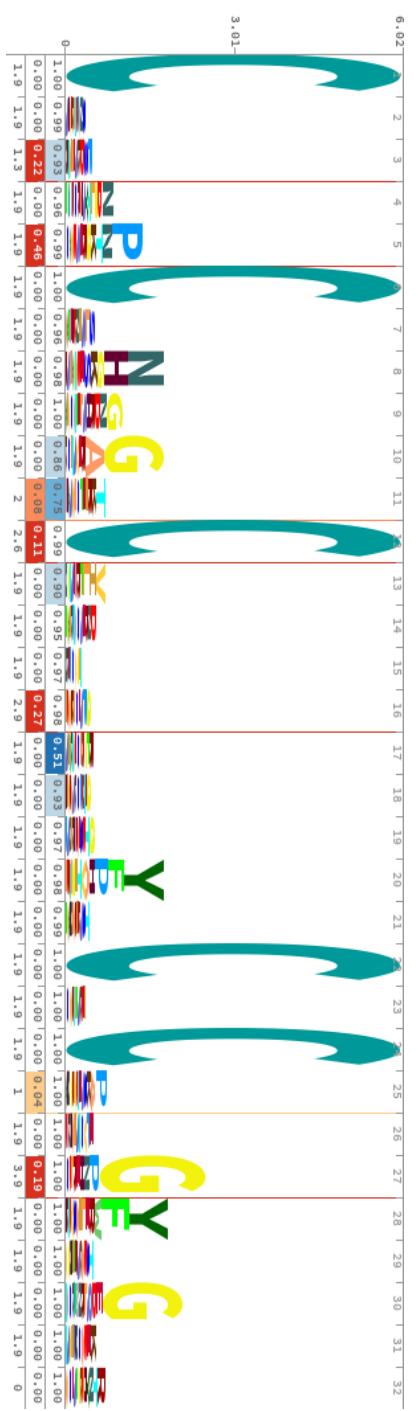
Evolutionary conservation can be calculated by considering the amount of different amino acids encountered. This is computed per column in an MSA, and preferably calculated over homologous protein sequences from evolutionary distant species. The result per position can be expressed as relative entropy<sup>37</sup>. Using relative entropy, in **figure 4B**, the letter-size indicates how conserved residues are based on the MSA from **Figure 4A**.

Figure 4.



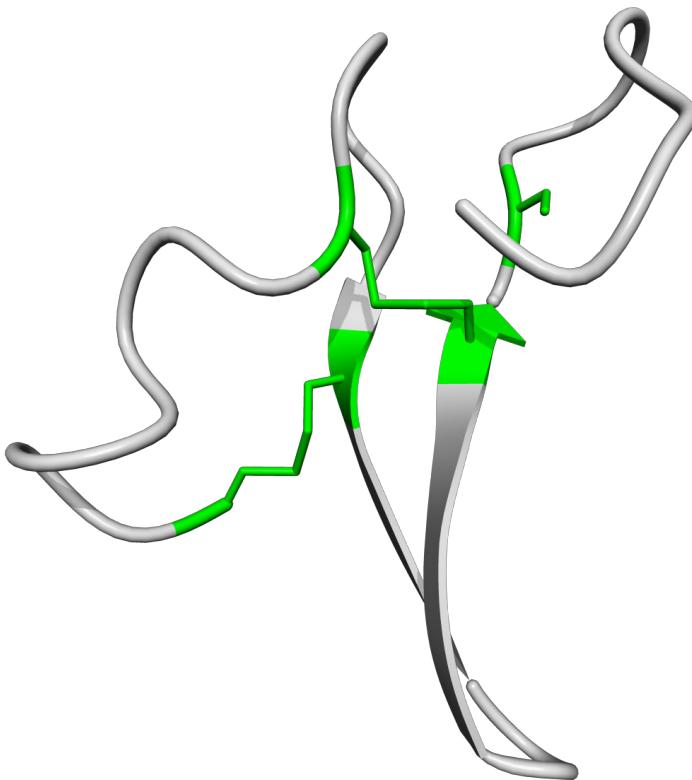
An example of a multiple sequence alignment. This is a small excerpt of the seed alignment for the EGF-like domain (PF00008), a Pfam<sup>41</sup> protein domain family. For this domain family, a total of 67 sequences were used to identify the sequence characteristics for this family. Here you can see that the Cysteines represented as 'C's with a blue background rarely change between species and are therefore conserved.

B.



The Pfam HMM sequence logo generated via the Skylign tool<sup>42</sup> for the EGF-like domain (PF00008). The height of each residue is based on the inverted relative entropy for that position. The height indicates how conserved each residue is in multiple sequence alignment from **44**. In this example the big Cs correspond to highly conserved cysteines. The thin red vertical lines in the sequence logo denote regions prone to contain deletions and the orange lines are regions prone to insertions.

C.



A part of PDB structure 1VO<sup>43</sup> (Chain C, p.6-41). This region is a structural representation of an EGF-like domain (PF00008). The structure is coloured grey and the five conserved cysteines are coloured green. The sulphur ions form disulphide bridges between two pairs of cysteines in this structure adding to the overall structural rigidity. The fifth cysteine, not connected, typically connects to another cysteine that is part of the connecting structure.

## Protein domains and homology

Protein structure is more evolutionary conserved than sequence.<sup>44</sup> The protein structure determines the function. Protein functions rely on elementary functional elements. These elements are for example the binding of an ion, voltage-gating, a specific structural shape, etc. These elementary functions have been optimized over the course of evolution. When these elements have a similar protein structure and/or sequence they are called protein domains. Protein domains can be detected from sequences by locating evolutionary conserved regions. When these evolutionary conserved regions have a similar sequence composition and/or structure, then these often have the same function. When these regions are homologous, and can be located in multiple proteins, they can be part of a protein domain family.

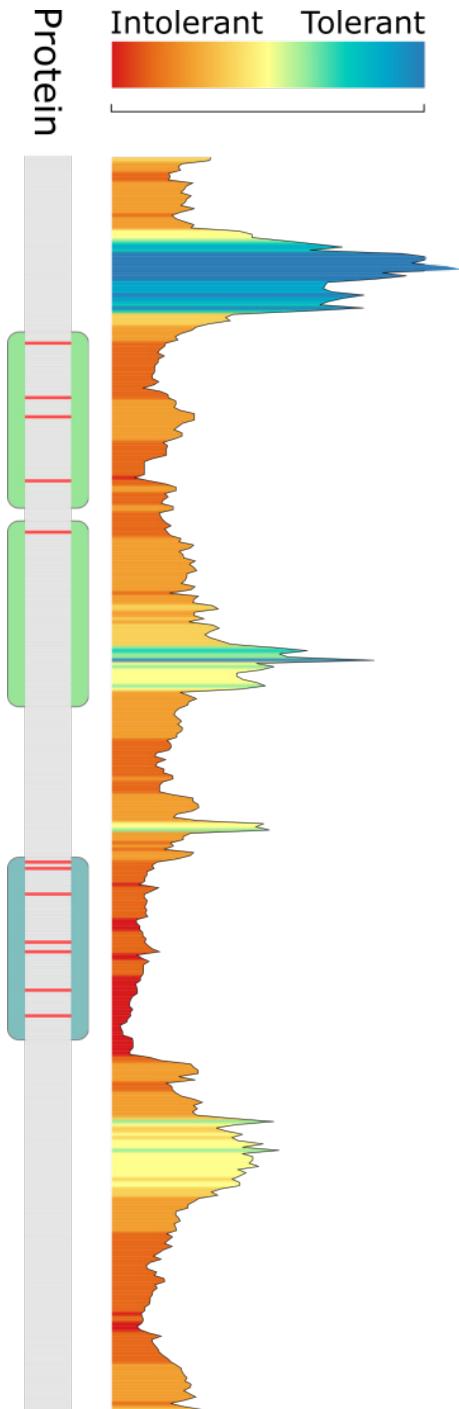
The example in **Figure 4** is an EGF-like domain (PF00008) that we analysed in-depth in **Chapter 2** of this thesis. This is a structural domain and most parts in this protein domain, from a sequence perspective, are variable (**Figure 4B**). The large C's, however, indicate conserved cysteines. The structural importance of the conserved cysteines can be seen in **Figure 4C** as they form rigid disulphide bridges. In EGF-like domains, any changes to the conserved cysteines will cause loss of a stabilizing disulphide bond necessary for the structure of the domain.<sup>45</sup>

## Understanding the human genome from an evolutionary perspective

The UniProt Knowledgebase (UniProtKB) currently contains 37,670 proteomes of which 1,832 are part of the Swiss-Prot collection that have been reviewed by experts (release 2020\_03).<sup>46</sup> Evolutionary conservation between-species can be computed from these data. This helps to discover homologous genes, proteins and protein domains. Most proteomes contained in the UniProtKB result from a single to a few sequencing samples. It will require considerably more sequencing efforts to analyse the within-species variability for each of these proteomes. For humans, however, sequence data is becoming more readily available. This is gradually leading to a more accurate estimation of within-human variation. Patients and controls involved in genetic studies can consent to their genetic data be used for scientific purposes. Contributing to the formation of large population-size catalogues of genetic variation.<sup>47-51</sup> The largest dataset to date is gnomAD, representing 141,456 individuals.<sup>51</sup> From these datasets, the frequency

of rare and commonly encountered genetic variations can be determined. These measurements have led to the notion of genetic tolerance. Genetic tolerance is a measurement from a within-species perspective, and, has a likeness to evolutionary conservation. However, it is different in that evolutionary conservation is based mostly on single sequence comparisons between related species. In genetic tolerance there are hundreds of thousands of sequences that we can compare from a single species. This abundance of data can uncover much finer details than ‘conserved’ versus ‘variable’. Genetic tolerance can indicate positions and regions that are highly variable or not variable at all. Genetic tolerance can help to determine the likely pathogenicity of genetic variants.<sup>26,52</sup>

In recent years, metrics such as RVIS<sup>8</sup>, subRVIS<sup>9</sup> and pLI<sup>50</sup> have been developed that provide an indication of potential deleteriousness of variants. Perhaps inspired by evolutionary conservation, these methods use the absence of population-based variation to determine variant deleteriousness. Genes vary in their tolerance to variation and this can be used to determine their essentiality.<sup>8</sup> Regions within genes vary in tolerance to variation as well. Regions that are intolerant to variation correspond to important parts of the gene and disease variants are more likely found within these regions.<sup>9,10</sup> For example, **Figure 5** depicts a ‘tolerance landscape’ for the gene *LMX1B* created by our webserver MetaDome (**Chapter 3**). The regions that are intolerant to missense variants correspond to the protein domain regions and where disease-causing variants are encountered.



**Figure 5.** Visualization of Tolerance Landscape of *LMX1B* (transcript: ENST00000355497.5, protein: 060663-3) created by MetaDome<sup>53</sup> (accessed July 23 2020). Tolerant regions are coloured blue and intolerant regions red. The light-green blocks are LIM domains (PF00412, p.56-p.110 and p.115-p.172), the dark-green a Homeobox domain (PF00046, p.220-p.276). The 12 red bars indicate locations where pathogenic variants are recorded in ClinVar54, corresponding to 6 missense (p.Q559Phe, p.Cys118Phe, p.Arg223Gln, p.Arg246Gln, p.Arg261Cys, p.Asn269Lys) and 7 nonsense (p.Tyr102\*, p.Arg221\*, p.Arg231\*, p.Arg231\*, p.Arg246\*, p.Arg249\*). In this visualization, tolerance is based on a missense over synonymous ratio, using the genetic variation from gnomAD.51

## Scope of this thesis

In this thesis I have combined structural biology and human genetics. I integrate protein information with publicly available human genetic variation. This combination allowed validation of the following hypotheses:

**Hypothesis I:** The parts of a protein that are tolerant to population-based genetic variation are not important for protein function.

**Hypothesis II:** Genetic variants that are damaging to a part of a protein can be used to predict damaging effects in highly similar parts in other proteins.

Investigating these hypotheses led to integrate human genetic data with protein domain and protein structure information. This combination resulted in the following chapters.

### Meta-domains and the MetaDome web server

Integrating human genome data with homologous protein domains resulted in meta-domains (**Chapter 2**). Meta-domains allow transfer of information between equivalent residues in different protein domains. This transfer of information helps interpret genetic variation. The meta-domain concept has been implemented in the MetaDome web server (**Chapter 3**).

### Clustering of *de novo* missense mutations suggest disease mechanisms

*De novo* mutations (DNMs) are rare genetic variants. In patients with developmental disorders (DD), DNMs are the likely cause. We identified that missense DNMs clustered in 15 genes in publicly available DD patient data (**Chapter 4**). Of these, 3 genes were novel DD-associations. Analysis of these clusters in the protein 3D structure suggest an N-HI disease-mechanism.

### Deleterious *de novo* missense mutations locate to protein domains

We formed the largest cohort to date of DNMs identified in 31,058 DD-patients (**Chapter 5**). We found 285 genes significantly enriched with DNMs. Of these, 28 genes were novel DD-associations. Specifically, I showed that missense DNMs are more likely located in protein domains. This is not the case for stop-gained and synonymous DNMs. Furthermore, specific protein domain families are enriched with missense DNMs identified in DD-associated genes.

## Gene DD-association based on a single *de novo* mutation

I combined meta-domains (**Chapter 2**) with the insights that missense DNM clusters indicate disease-mechanisms (**Chapter 4**), and, that protein domains are enriched with missense DNMs (**Chapter 5**). This led to the identification of missense DNM hotspots in meta-domains (**Chapter 6**). The hotspot DNMs were located in 25 genes. Analysis of these hotspots in the protein 3D structure confirmed deleteriousness. Six of these genes are novel candidate DD-associations based on a single DNM in a hotspot.

In **Chapter 7** I discuss the limitations and implications of this thesis.



# Chapter 2

---

Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics

*Laurens van de Wiel, Hanka Venselaar,  
Joris A. Veltman, Gert Vriend, and Christian Gilissen*

Published in Human Mutation  
November 2017 Nov; 38(11):1454-63

## Abstract

Whole exomes of patients with a genetic disorder are nowadays routinely sequenced but interpretation of the identified genetic variants remains a major challenge. The increased availability of population-based human genetic variation has given rise to measures of genetic tolerance that have been used, for example, to predict disease-causing genes in neurodevelopmental disorders. Here, we investigated whether combining variant information from homologous protein domains can improve variant interpretation. For this purpose, we developed a framework that maps population variation and known pathogenic mutations onto 2,750 “meta-domains.” These meta-domains consist of 30,853 homologous Pfam protein domain instances that cover 36% of all human protein coding sequences.

We find that genetic tolerance is consistent across protein domain homologues, and that patterns of genetic tolerance faithfully mimic patterns of evolutionary conservation. Furthermore, for a significant fraction (68%) of the meta-domains high-frequency population variation re-occurs at the same positions across domain homologues more often than expected. In addition, we observe that the presence of pathogenic missense variants at an aligned homologous domain position is often paired with the absence of population variation and vice versa. The use of these meta-domains can improve the interpretation of genetic variation.

## Acknowledgements

This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (916-14-043 to C.G. and 918-15-667 to J.A.V.), and from the Radboud Institute for Molecular Life Sciences, Radboud university medical center (R0002793 to G.V.). We thank Susanne Roosing for her suggestions on data presentation, and Tom Heskes for his advice on the statistical analyses.

## Introduction

Next generation sequencing technologies now allow for the comprehensive identification of all genetic variation in an individual, and exome and genome sequencing are increasingly being used in clinical care to provide a diagnosis for patients with a genetic disorder.<sup>23,24</sup> The interpretation of the large number of genetic variants present in the exome or genome of a patient is now the major remaining challenge.<sup>26</sup> Filtering strategies that reduce the number of candidate disease-causing variants make use of information such as the occurrence of variants in the normal and in the diseased population, knowledge about the role of genes in disease, and the predicted effect of specific mutations.<sup>30</sup> Algorithms such as Polyphen-2<sup>28</sup> and CADD<sup>29</sup> are able to predict the pathogenicity of individual variants, but leave room for improvement, especially within a clinical context.<sup>55-57</sup> Other methods have used population-wide genetic variation from healthy individuals that is available in large public databases such as the NHLBI Exome Sequencing Project (ESP),<sup>58</sup> and the Exome Aggregation Consortium (ExAC)<sup>50</sup> to construct metrics that estimate the genetic tolerance of a gene. Various studies have shown that genetic intolerance of a gene is a strong indicator for a role in severe human diseases such as intellectual disability and other neurodevelopmental disorders.<sup>8,59</sup> Metrics such as RVIS<sup>8</sup> and pLI<sup>50</sup> are now being used in conjunction with variant pathogenicity prediction algorithms to improve the interpretation of variants of unknown significance in patients suffering from these disorders.

The continuous growth of catalogues of human genetic variation has made it feasible to investigate genetic tolerance at a finer scale, such as for individual exons of a gene or even domains of a protein. This was done, for example, by Gussow *et al.*<sup>9</sup> who developed subRVIS and found that tolerance within a gene varies, and that specific protein domain coding parts of a gene are sometimes much more intolerant than the whole gene. Moreover, the authors found that intolerance to genetic variation within genic sub-regions significantly correlates with reported pathogenic mutations. These patterns of region-specific variation in genetic tolerance were also used by Ge *et al.*<sup>10</sup> to detect missense-depleted regions to confirm the pathogenicity of individual variants of unknown significance.

Since its introduction, one of the applications of BLAST<sup>38</sup> was to identify homologous proteins. Mutations at corresponding locations in these homologues were found to result in similar effects on protein stability.<sup>40</sup> Protein domains are

especially interesting as they have homologous relationships spanning many proteins. Because of this, protein domains can also have many homologues that occur within the same species. An example of a framework that annotates protein domains to proteins is Pfam.<sup>41</sup> The Pfam database is a large collection of protein domain families represented by curated multiple sequence alignments (MSAs) and a hidden Markov model (HMM). In recent work Miller *et al.* combined mutation information from different protein domain homologues to identify mutation hotspots in cancer, and Melloni *et al.* used a similar approach to identify cancer driver mutations.<sup>60,61</sup> We hypothesized that genetic tolerance found in the regions coding for protein domains, may be consistent across other within-human homologues of that domain and that therefore interpretation of variants in a protein domain can be improved by aggregating population variation over homologous protein domains.

## Materials and Methods

### Mapping of human genomic variation to Pfam domains

We performed a Protein-Protein BLAST 2.2.31+<sup>62</sup> for each of the longest translations for all 18,651 human protein-coding genes in the GENCODE Basic set release 19 GRCh37.p13<sup>63</sup> to canonical and isoform human protein sequences in UniProtKB/Swiss-Prot Release 2016\_09 (Swiss-Prot).<sup>64</sup> We then selected the top BLAST result with 100% identity to the query sequence and a BLAST E-value of 0.01 or less. Pfam-A 30.0<sup>41</sup> protein domains in the matched Swiss-Prot sequences were annotated using InterProScan 5.20-59.0.<sup>65</sup> ClustalW2 v2.1<sup>39</sup> was used to create pair-wise alignments between the gene translations and Swiss-Prot sequences. The resulting alignment was then used to map genomic variation onto residues in Swiss-Prot protein sequences.

### Datasets of population genetic variation and disease-causing missense variants

Population variation was obtained from the Exome Aggregation Consortium (ExAC) v0.3.1<sup>50</sup> by selecting all synonymous and missense variants with the PASS filter criteria. For the creation of meta-domains we considered missense variants from ExAC with an allele frequency > 0.1%. For validation purposes we also used two additional sets of ExAC missense variants having >0.5% and >0.05% allele frequency.

We selected a set of disease-causing missense variants from the Human Gene Mutation Database (HGMD) 2016.2<sup>66</sup> that have disease-causing (DM) status, which were subsequently filtered by removing all variants that are identical to PASS variants in ExAC with >0.1% allele frequency. This filtering reduced the original set of HGMD DM missense variants by 0.17%. In addition, we used missense variants from ClinVar (downloaded for GRCh37 on 2017-06-15), with disease-causing (Pathogenic) status, as an additional validation to HGMD DM variants. The filtering of identical PASS variants in ExAC with >0.1% allele frequency, that was used for the HGMD DM set, was applied to this set as well.

### **Aggregation of genetic variation into meta-domains**

In order to aggregate genetic information over protein domain homologues we considered each Pfam identifier found in more than one gene as a within-human homologue. In this study, when we mention homologous protein domains, or domain homologues, we refer to Pfam protein domains that are homologous in the protein-coding regions of the human genome. For each domain found this way, we retrieved the Pfam HMM and the domain protein sequence. We used all the domain sequences that had the same Pfam identifier, together with the Pfam HMM, to generate a MSA using the HMMER 3.1b2 tool.<sup>67</sup> We used our mapping to combine genetic variants on positions that were aligned to the same Pfam domain positions. Variations on Swiss-Prot residues in insertions with respect to the Pfam domain were ignored. The percentage of homologous domains aligned to a position (MSA coverage) was determined based on the number of gaps with respect to the Pfam domain.

### **Gene Ontology Biological Process enrichment analysis in protein domains**

Gene Ontology Biological Process (GOBP) enrichment analysis was performed using the R package dCGOR 1.0.6.<sup>68</sup>

### **Computing genetic tolerance via the missense over synonymous ratio**

We use the non-synonymous over synonymous ratio, or  $d_N/d_S$  score, to quantify genetic tolerance in genes and domains. In our setting this score is based on the single nucleotide missense and synonymous variants (SNVs) from ExAC in a

protein-coding region ( $missense_{obs}$  and  $synonymous_{obs}$ ). This score was corrected for the sequence composition of the protein coding region based on the total possible missense and synonymous SNVs ( $missense_{bg}$  and  $synonymous_{bg}$ ):

$$d_N/d_S = \frac{missense_{obs}/missense_{bg}}{synonymous_{obs}/synonymous_{bg}}$$

## Consistency of genetic tolerance across protein domain homologues

We calculated the Median absolute deviation:

$MAD(x) = \text{median}(|d_N/d_S(x_i) - \text{median}(d_N/d_S(x))|)$  to measure whether genetic tolerance scores are consistent across homologous domains. For each domain occurrence ' $x_i$ ' of a homologous domain group ' $x$ ' we calculate the difference of  $d_N/d_S$  score to the median. The median of all these differences is then computed as the MAD. The minimal and optimal value of the MAD score is zero, meaning that no score deviates from the median. To test whether the MAD score per homologous domain group is significantly different from another randomly selected group of homologues, we permuted the MAD scores for each homologous domain group using the  $d_N/d_S$  score of each member in that group and comparing it to the median  $d_N/d_S$  of another homologous domain group that we selected via the numpy function random.permutation in Python. This permutation test was repeated 10,000 times.

## Evolutionary conservation and population variability

We measured sequence conservation via the relative entropy per position<sup>37</sup> in a multiple sequence alignment (MSA) to compute the evolutionary conservation and population variability:  $\text{relative\_entropy}(j) = \frac{-\sum_{R=1}^{20} f_{Rj} \ln f_{Rj}}{\ln 20}$ . Here ' $j$ ' is an aligned position, ' $R$ ' is the amino acid residue type, ' $f_{Rj}$ ' is the frequency of how often a residue of type ' $R$ ' occurs at position ' $j$ '. The relative entropy ranges from 0.0 to 1.0 for conserved to variable. We used the Pfam-A full alignment for each Pfam domain to compute evolutionary conservation. We used our mappings to assess population variability by extracting missense and synonymous variants and their respective allele frequencies from ExAC to compute the ' $f_{Rj}$ ' variable. To achieve a sufficiently high MSA resolution and certainty of correct entropy we only considered positions for computing the relative entropy that had at least 25 sequences with 80% MSA coverage.

## Quantifying patterns of missense variants in meta-domains

We created a metric to quantify how often a consensus position in a meta-domain contains identical missense variants (i.e. two or more homologous domains wherein the aligned residues both are identical in reference and alternative amino acid residues). We call this metric the characteristic missense variant score:  $CMVS = \sum_j^{L_x} \frac{C_x[j]}{M_x[j]}$ . Here ' $L_x$ ' is the size of meta-domain ' $x$ ', ' $j$ ' is an aligned domain position, ' $M_x[j]$ ' are the number of missense variants found in all domain homologues aligned to position ' $j$ ' and ' $C_x[j]$ ' are the number of missense variants in ' $M_x[j]$ ' that are of identical change in amino acid (i.e. that have identical reference residues and change to the same alternate residue). The  $NCMVS = \frac{CMVS}{L_x}$  normalizes the CMVS with respect to the domain size.

We assigned values of significance to patterns of missense variants observed in meta-domains by comparing these to permuted meta-domains resulting from Monte Carlo experiments. In these experiments we shuffled missense variants in each domain occurrence ' $x_i$ '. To perform this shuffling, we first estimated the probability of a missense variant to occur in ' $x_i$ ' via  $\frac{M_{x_i}}{L_{x_i}}$  if  $M_{x_i} > 0$ , else  $\frac{1}{L_{x_i}}$ , where ' $L_{x_i}$ ' are the number of aligned residues and ' $M_{x_i}$ ' are the number of missense variants found in domain ' $x_i$ '. Then we estimated the probability for any missense variant to occur on an aligned position ' $j$ ' by considering the codon of that position with respect to the codon table:  $\frac{\# \text{possible missense}(x_i[j])}{9}$ . Finally, we distributed missense variants on the domain occurrence by combining these two probabilities and assessing each possible missense variant. The distribution of missense variants was subsequently used to reconstruct a permuted meta-domain over 1,000 experiments for each meta-domain.

The patterns of missense variants across homologues were then tested for significance in two different ways. First we computed per aligned position the ratio of missense variants observed in contrast to the number of domain occurrences aligned. We checked if a position is significantly enriched for either the reference allele or the missense variant allele as compared to the same position in the permuted meta-domains. We report the meta-domains for which more than 75% of the positions are significantly different from the permuted meta-domains. Secondly, we tested whether the entire meta-domain is significantly enriched for identical variants via NCMVS as compared to the permuted meta-domain. In both cases we made our comparisons with the Welch's t-test and used Bonferroni correction for multiple testing.

## Results

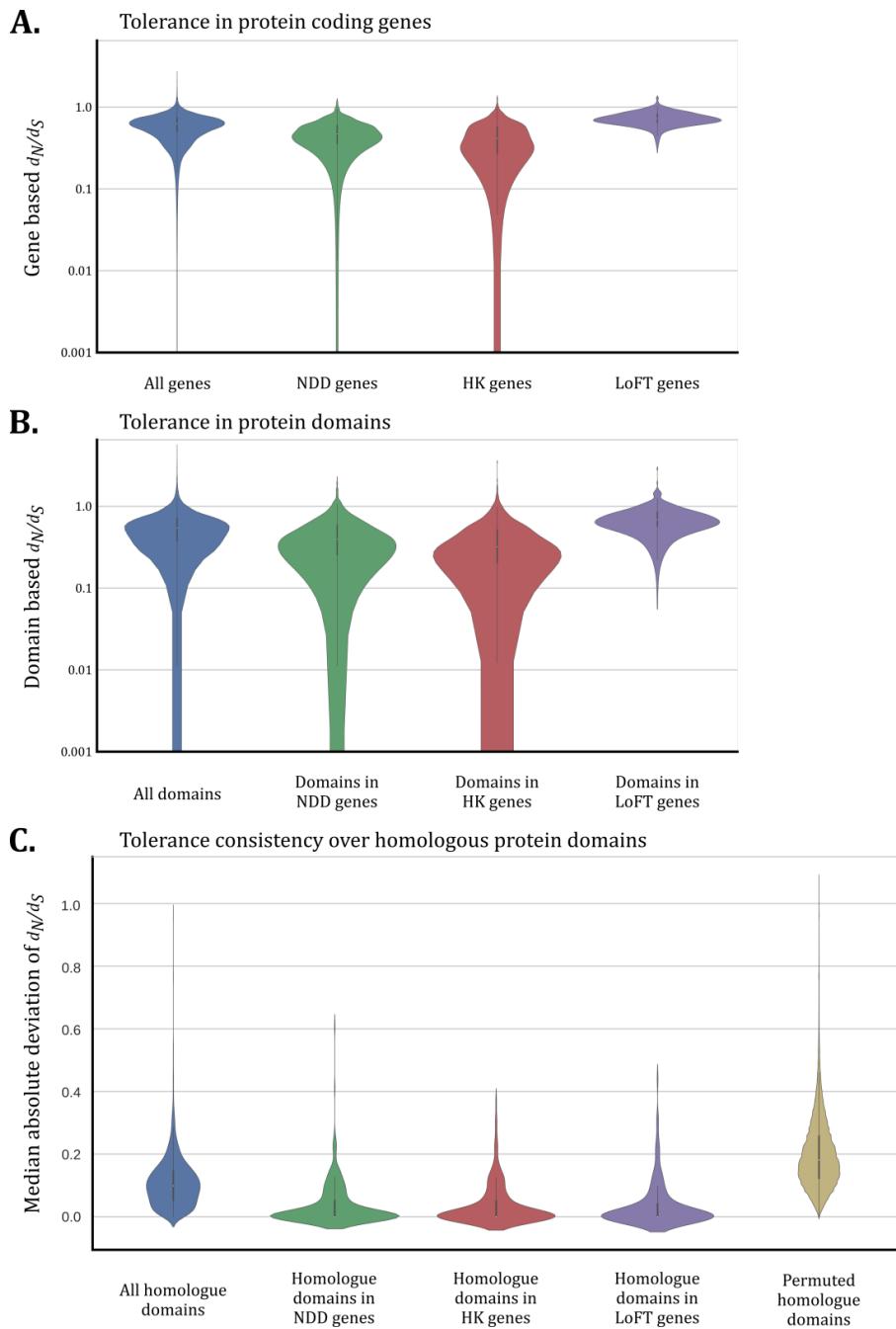
In total 16,684 GENCODE genes were mapped to Swiss-Prot protein sequences and annotated with protein domains from Pfam (**Methods**). We found 5,250 Pfam domains spanning 33,638 domain occurrences in these genes, of which 30,853 made up 2,750 within-human Pfam domain homologues (**Supp. Table S1**). We found 961 Pfam domain homologues to occur in exactly two different genes and, on average, a within-human homologous protein domain occurs in at least six different human genes. The most prevalent domains were the “KRAB domain” (PF01352), “Zinc finger, C2H2 type” (PF00096) and “Protein kinase domain” (PF00069), each being present in more than 300 different human genes. Pfam protein domains covered approximately 41% of coding sequences of the 16,684 genes. In total 1,493,414 synonymous, 2,892,092 missense variants from ExAC, 58,968 DM missense variants from HGMD, and 14,016 Pathogenic missense variants from ClinVar are present in the coding regions of our set of genes. 71% of disease-causing missense variants from HGMD and 72% pathogenic missense variants from ClinVar occur in Pfam domain regions (**Supp. Table S2**).

### Tolerance to genetic variation of protein domains

Regions that code for protein domains are sometimes much less tolerant than the whole coding region of a gene.<sup>9</sup> Therefore, we first wanted to test how similar tolerance patterns in protein domains are to their respective genes. We used the population-based variation from ExAC to compute the ratio of missense over synonymous variants ( $d_N/d_S$ ). This, we used as a measure of genetic tolerance scores for all genes and Pfam domains (**Supp. Data S1 and S2; Methods**). We compared the tolerance measured in genes of different gene sets that are known to have a particular pattern of genetic tolerance<sup>59</sup> to the tolerance of the regions with protein domains in these genes. We found that protein domains in genes known as intolerant, such as housekeeping genes<sup>69</sup> and genes involved in neurodevelopmental disorders,<sup>70</sup> are indeed intolerant too (Welch’s t-test p=4.33e-61 and p=5.24e-57 respectively; **Supp. Table S3, S4**). Conversely, we found that domains in genes that are known to be tolerant to protein truncating variation and variation in general<sup>71</sup> are also tolerant to missense variation (Welch’s t-test p=7.42e-23; **Supp. Table S3 and S4; Figure 1a and 1b**). Thus we find that protein domains have a similar trend of tolerance as their genes.

After establishing that genetic tolerance of a domain mimics that of its respective gene we wondered whether  $d_N/d_S$  scores are consistent across domain homologues. We used the Median Absolute Deviation (MAD) computed over the homologues of a domain to test for the consistency of genetic tolerance (**Supp. Data S3; Methods**). We find that 2,741 out of 2,750 (99%) aggregated homologues show a consistent pattern of  $d_N/d_S$  scores as compared to what may be expected by chance (Welch's t-test  $p<0.05$ , Bonferroni corrected; **Methods; Supp. Table S5; Figure 1c**). The most consistently intolerant domain was the "SRF-type transcription factor (DNA-binding and dimerisation domain)" (PF00319) whereas the "Keratin, high-sulphur matrix protein" (PF04579) is the most consistently tolerant domain (**Supp. Table S6, S7**). These results show that domains have tolerance patterns that are consistent over homologues, and thus that genetic variation in one protein domain is therefore not fully independent from the variation measured in the homologues of that domain. This potentially allows us to aggregate variant information across protein domain homologues.

Interestingly, enrichment analysis for Gene Ontology Biological Process (GOBP) on the top 5% of most intolerant domains ( $n = 134$ ) found that these are strongly enriched for biological processes such as chromatin condensation, chromosome organization and DNA packaging ( $p=5.90e-08$ ,  $p=7.10e-05$ ,  $p=1.10e-05$  respectively, **Supp. Data S4**). This connection to chromatin remodelling has also been observed among dominant genes for neurodevelopmental disorders.<sup>72-74</sup>

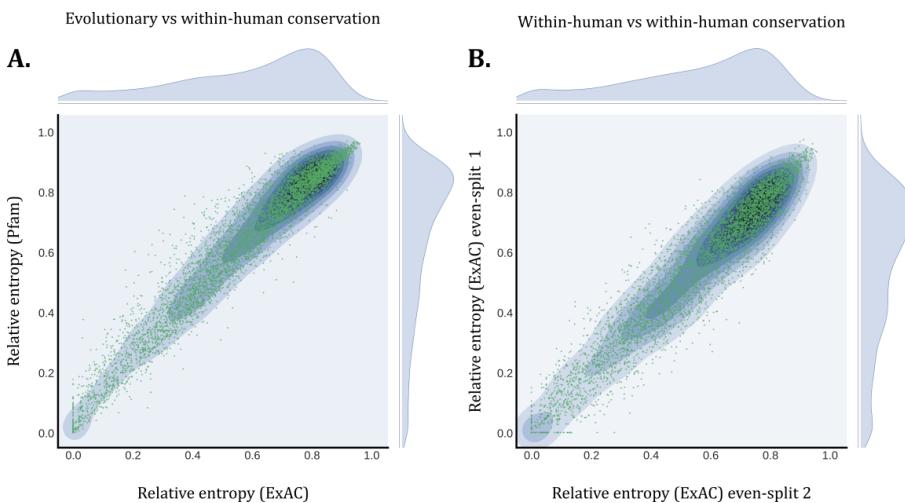


◀ **Figure 1.** Tolerance in genes, domains and domain homologues

**A.)** Tolerance to normal genetic variation as measured via the  $d_N/d_S$  ratio (**Methods**). A higher  $d_N/d_S$  ratio means that the gene is more tolerant to genetic variation and vice versa. From left to right data is presented for all 16,684 genes (blue), 398 genes involved in neurodevelopmental disorders (green),<sup>70</sup> 361 housekeeping genes (red),<sup>69</sup> 157 loss-of-function tolerant genes (purple).<sup>71</sup> All groups are significantly different (*Supp. Table S3*). **B.)** As A. with the exception that the  $d_N/d_S$  ratio is now computed only for domain regions. All 33,638 domains (blue), 1,302 domains in genes involved in neurodevelopmental disorders (green), 811 domains in housekeeping genes (red), 358 domains present in loss-of-function tolerant genes (purple). All groups are significantly different (*Supp. Table S4*). **C.)** The consistency of  $d_N/d_S$  scores across homologous domains computed via the MAD of the  $d_N/d_S$  (**Methods**). The lower the MAD score the more consistent is the  $d_N/d_S$  ratio. There are 2,750 Pfam domains that have homologues in our set of genes with a total of 30,853 occurrences (blue). Of the Pfam domains, 383 have a homologue occurring in a gene involved in neurodevelopmental disorders (green), 223 have a homologue occurring in a housekeeping gene (red), and 178 have a homologue occurring in a loss-of-function tolerant gene (purple). The permuted domains (yellow) consists of 27,500,000 permuted MAD scores that resulted by computing the MAD score using the median  $d_N/d_S$  of another Pfam domain (**Methods**). All groups have been found significantly different from the permuted domain group (*Supp. Table S5*). The impact of different domain sizes on the MAD score is minimal (*Supp. Figure S5 and S6*).

## Population variability across domain homologues mimics evolutionary conservation

Although many methods have made use of population-based genetic variation to assess genetic tolerance, it has remained unclear to what extent population variability complements information from evolutionary conservation. Within-human protein domain homologues offer the unique opportunity to answer this question. We compared the consistency of population-based genetic variation with evolutionary conservation across homologous domain positions by investigating 81 Pfam domains that have at least 50 homologous instances in our set of human protein-coding genes, twice of what we need to ensure high-quality alignments (**Methods**). In total, for 6,536 positions of these 81 domains we measured relative entropies based on population and evolutionary variation in 14,059 human domain instances. We observe a high degree of correlation between these two groups (Pearson = 0.97, p-value < 1e-308; **Methods; Figure 2a**). We validated this result further by splitting the population-based entropies evenly into two separate groups, each consisting of 25 or more homologous instances. This way we can test for any noise in the computation of within-human conservation. Again, the relative entropies results in an almost perfect correlation (Pearson = 0.96, p-value < 1e-308; **Figure 2b**). These results show that variation in the human population measured across homologous protein domains faithfully mimics evolutionary conservation, thereby providing support for our proposed approach to aggregate genetic variation across domain homologues.

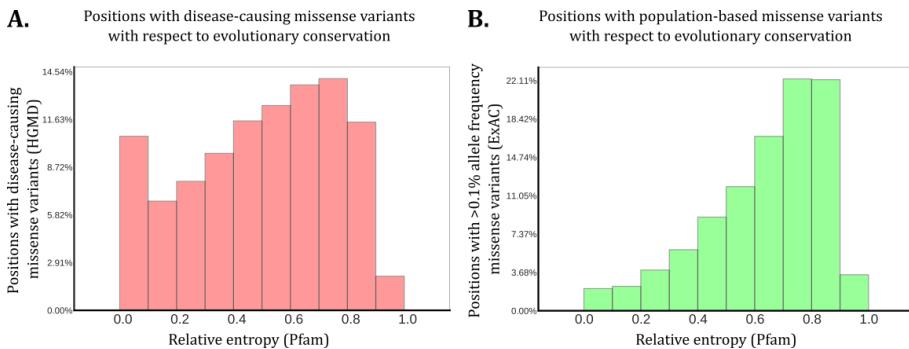


**Figure 2.** Evolutionary conservation and within-human conservation in Pfam domains

For 81 domains that have 50 or more homologues within the human genome we computed the relative entropy to measure the conservation of amino acid residues per position in these domains for both evolutionary conservation based on Pfam and within-human conservation based on ExAC (**Methods**). In both plots the x and y-axis represent the relative entropy for a single position in a domain that ranges from 0.0 to 1.0; conserved to variable. **A.** On the y-axis evolutionary conservation is represented by the relative entropy per position based on Pfam. The x-axis shows variability measured solely in the human genome, based on relative entropy computed from ExAC. These two measurements show almost perfect correlation. (Pearson correlation coefficient = Pearson = 0.97, p-value < 1e-308). **B.** A validation of the results presented in A where we split the relative entropy measured solely in the human genome in two, hereby comparing the conservation solely between human protein domains. Again we observe an almost perfect correlation (Pearson correlation coefficient = 0.96, p-value < 1e-308).

To establish whether population variation adds additional information for variant interpretation compared to evolutionary conservation we assessed how disease-causing and population-based missense variants are distributed with respect to evolutionary conservation. We expected to find that positions containing disease-causing variants are conserved in general, whereas positions with genetic missense variants common in the human population are expected to be variable. Therefore we investigated 17,195 positions in 1,079 Pfam domains with 31,732 disease-causing missense variants from HGMD. Contrary to what we expected, more than 54% of the positions with a disease-causing missense variant were found to be evolutionary variable with a relative entropy of 0.5 or higher (**Figure 3a**). The local maxima, observed between 0.0 and 0.1 relative entropy in **Figure 3a**, was expected to degrade gradually for higher levels of entropy. As this is a measurement on protein domains, we hypothesize that this local maxima is

caused by mutations that affect active site residues. In line with our expectations, when we performed the same analysis for positions with missense variants that have >0.1% allele frequency in ExAC, we found that 77% of these positions was highly variable (**Figure 3b**). These results highlight that evolutionary conservation is not the perfect indicator for pathogenic mutations, and that population-based genetic tolerance scores may function as a complementary approach in variant interpretation.



**Figure 3.** Number of missense variants per position in a meta-domain in perspective of conservation

Plotted here is the binned distribution of positions that contain one or more missense variant of interest with respect to the evolutionary conservation of the position where these variants occur. The x-axes are denoted by "Relative entropy (Pfam)" and the y-axes are marked as the overall percentage of these positions. The figure shows that disease-causing missense variants also affect very variable sites. **A.** 17,195 different positions spanning 1,079 Pfam domains. On these positions 31,732 disease-causing missense variants from HGMD were found in 22,651 domain occurrences in the human genome. Of these positions, 54% have relative entropy 0.5 or higher. **B.** 13,571 different positions spanning 1,965 Pfam domains. On these positions 17,258 missense variants with an allele frequency above 0.1% in ExAC were found in 27,767 domain occurrences. 77% of these positions have relative entropy 0.5 or higher.

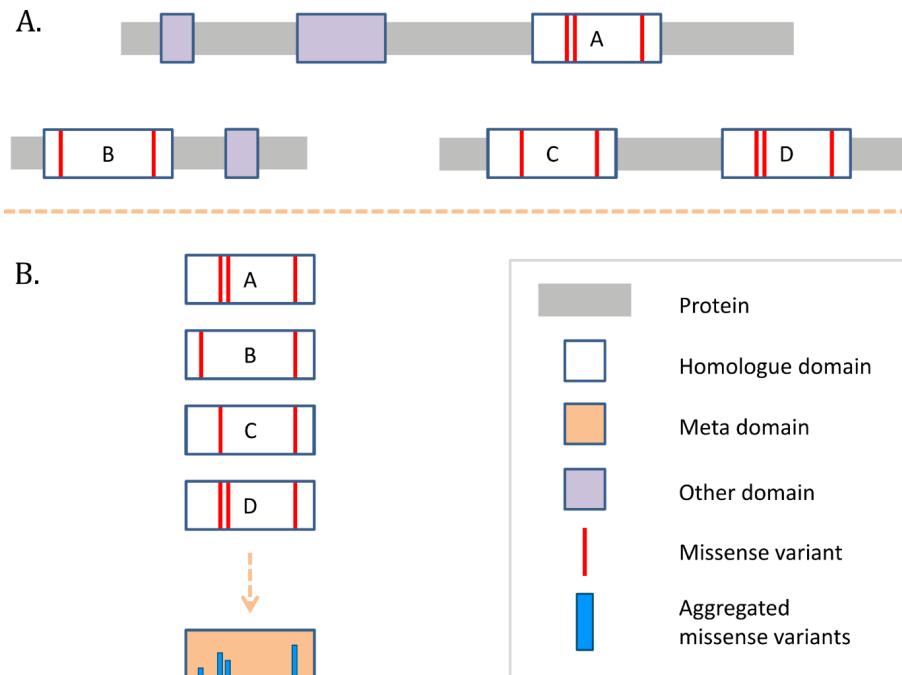
## Creation of meta-domains by aggregating genetic variation over domain homologues

Based on our results that genetic variation is consistent across human protein domain homologues, and that population-based genetic variation correlates faithfully with evolutionary conservation, we hypothesized that genetic variation can be aggregated across homologous domains to provide a more detailed map of genetic variation. Hence, we projected disease-causing and population-based missense variation found in human protein domains onto Pfam domain consensus positions giving rise to a "meta-domain" (**Methods; Figure 4**). In total

we successfully projected 20,404 population-based missense variants with >0.1 % allele frequency from ExAC, 35,069 disease-causing missense mutations from HGMD and 8,569 pathogenic missense mutations from ClinVar (**Supp. Data S5; Methods**). We tested whether there was any overlap between the pathogenic and population-based missense variants on aligned positions by comparing HGMD DM with ExAC and found a negative correlation (Pearson = -0.51, p-value < 1e-308; **Supp. Figure S1**) indicating that disease-causing missense variants at aggregated domain positions often are paired with the absence high-frequency population missense variants and *vice versa*. This suggests that the information annotated to the meta-domains may be used to enhance variant interpretation.

To further confirm that aggregation of variants to Pfam domain consensus positions is meaningful, we perform two separate analyses. We first performed Monte Carlo experiments to test whether missense variants re-occur at the same position in domain homologues more often than could be expected by chance. We find that high-frequency population missense variants in 68% of the meta-domains re-occur at the majority of the aligned positions, and that this is significantly different from what may be expected by chance (Bonferroni corrected p<0.05 Welch's t-test; **Supp. Data S6 and S7; Methods**). Similarly we find that HGMD DM and ClinVar Pathogenic missense variants, in 65% and 62% of the meta-domains respectively, re-occur at the majority of the aligned positions (Bonferroni corrected p<0.05 Welch's t-test; **Supp. Data S6 and S7**). This analysis shows that the re-occurrence of missense variants found at aligned positions over all domain homologues follows a non-random pattern.

In our second analysis, again we perform Monte Carlo experiments and compute for each meta-domain our NCMVS metric to quantify how many missense variants, which re-occur at the same position, are also of identical change in amino acid (**Methods**). This way we find that high-frequency population missense variants in 21% of the meta-domains have significantly more variants of identical change at aligned positions across homologues as compared to what may be expected by chance. The pathogenic missense variants from HGMD DM and ClinVar Pathogenic datasets show a similar signal, with 23% and 18% respectively, of the meta-domains having an enriched NCMVS (Bonferroni correction p<0.05 Welch's t-test; **Supp. Data S7; Methods**). This second analysis shows that the change in amino acid of missense variants found over all domain homologues is for a large set of domains more often identical than what may be expected by chance.



**Figure 4.** Meta-domain construction in a schematic representation.

Genetic information is aggregated into a meta-domain based on domain homology. **A.** In this specific example there are three human proteins (indicated by the grey bars) with four domains that are found to have the same Pfam domain identifier and therefore belong to the same homologous domain group (indicated by A, B, C, and D). Red vertical lines in these domains indicate missense variants. There are other domains found in these proteins, but these are not further used in this specific example. **B.** The homologous domains together with their respective missense variants are extracted from the proteins and are aligned according to the Pfam domain. Based on the alignment the missense variants are then aggregated into a meta-domain. Some of these missense variants were aligned to the same position, in the meta-domain this is expressed with a higher blue column.

The results of these two analyses find that missense variation in domains follow a non-random pattern. Such a non-random pattern in pathogenic variants suggests that specific positions in domains are more likely to have a pathogenic effect via missense variants as compared to other positions. Conversely, finding a non-random pattern for re-occurring high-frequency population missense variants provides insight into positions that are genetically tolerant. These findings support our hypothesis that variant information can be aggregated across homologous domains, and that aggregation may help to interpret variants of unknown significance.

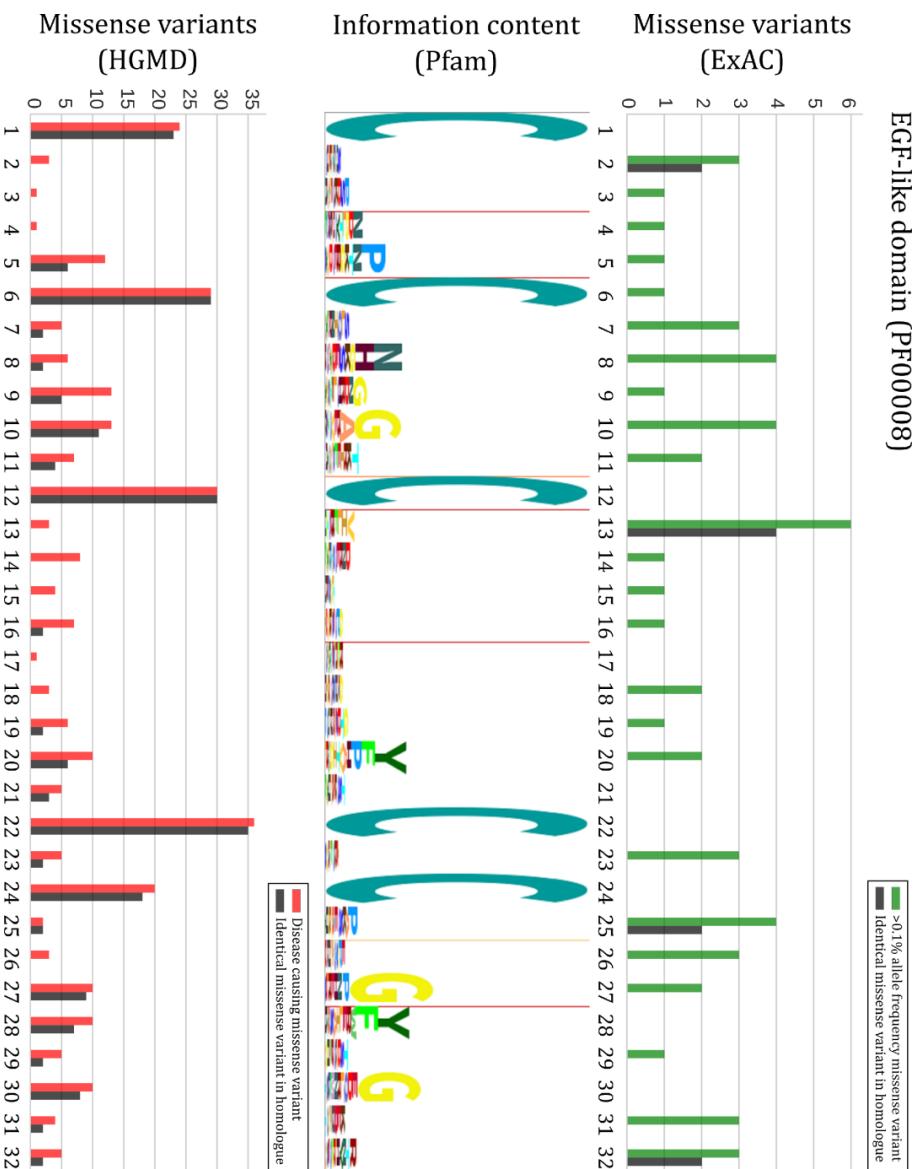
## Investigating a meta-domain in detail

To illustrate how these meta-domains can straightforwardly be used to improve variant interpretation we investigated one meta-domain in detail; the “EGF-like domain” (PF00008). This domain has 244 homologous occurrences in 60 different human genes (**Figure 5**). The “EGF-like domain” has the second highest NCMVS in the context of HGMD DM missense variants, and the 13<sup>th</sup> highest based on high-frequency population variants (**Supp. Data S7**). This suggests that the majority of variants often re-occur at aligned positions across the 244 homologues as identical changes in amino acids. Based on what is known from EGF-like domains, any changes to the conserved cysteines will cause loss of a stabilizing disulphide bond that are necessary for the structure of the domain.<sup>45</sup> As expected, we find that the highly conserved cysteines are indeed enriched for disease-causing variants across the 244 homologues. Furthermore, all of the conserved cysteines are depleted for population-based missense variants, with the exception of consensus position six, confirming the importance of these residues. For consensus position six we observe that population variation is present in only one homologue. This specific variant in *NOTCH4* (p.Cys815Gly, rs150079294) has an allele frequency 0.1632% in ExAC. dbSNP suggests that this variant is benign based on a single study<sup>75,76</sup> whereas our results further support the notion that this variant is problematic for this domain because of almost complete absence of common variation across the homologues. Even more interesting are the positions that are not evolutionary conserved (>0.6 relative entropy), but nevertheless depleted of population-based missense variation. In this “EGF-like domain” example, we find one such position at 21. In support of our hypothesis, we find multiple disease-causing missense mutations in different homologous domains at this position. We find that these

►**Figure 5.** An example of the EGF-like domain, represented as a meta-domain.

The “EGF-like domain” (PF00008) occurring in 60 different human genes found to be significantly enriched for identical disease-causing missense variants across 244 homologues. X-axis shows the amino acid positions of this domain. The green bars in the top panel indicate how many missense variants with >0.1% allele frequency from ExAC are found over the 244 homologous domains and. The black bars indicate the number of missense variants that are of identical chance in amino acid (i.e. having identical reference and alternate residues). The middle panel denotes the Pfam HMM sequence logo generated via the Skylign tool<sup>42</sup> where the height of each stack of residues indicates the relative entropy for that position. The thin red vertical lines in the sequence logo denote regions prone to contain deletions and the orange lines are regions prone to insertions based on the Pfam HMM. In the bottom panel red bars indicate the number of a disease-causing variant found across the 244 homologous domains. Black bars again indicate identical mutations. A comparison with ClinVar was made as well, albeit the dataset is much sparser as compared to HGMD (**Supp Figure S7**).

## EGF-like domain (PF00008)



disease-causing mutations have been previously linked to CADASIL (OMIM #125310, p.Tyr337Cys, p.Tyr1021Cys, p.Tyr1069Cys in *NOTCH3* (Q9UM47). CADASIL is an adult-onset autosomal dominant hereditary stroke disorder.<sup>77</sup> Other mutations aligned to this consensus position are p.Tyr690Asp in *JAG1* (P78604) associated with Biliary atresia extrahepatic (OMIM %210500), a disorder in infants that is fatal within the first two years of life when untreated,<sup>78,79</sup> and p.Arg628Cys in *CRB2* (Q5IJ48) associated with Nephrotic syndrome steroid resistant (OMIM #616220), a childhood onset renal disorder.<sup>80</sup>

These results illustrate how meta-domains can be straightforwardly used to improve the interpretation of genetic variants of unknown significance. We have made our mapping of genomic positions to meta-domain identifiers and consensus positions available for the wider genetic community to make use of in **Supp. Data S8**.

## Discussion

Here we combined two distinct concepts into a novel method for variant interpretation. Firstly, we used the observation that mutations at aligned positions in homologous proteins commonly lead to the same or similar effects on those proteins' structure and function. Secondly, large datasets of population scale exome data have made it possible to determine the degree of intolerance to genetic variation for individual genes in order to identify potential disease genes. We combined these two concepts by aggregating population variation across homologous protein domain positions and thereby achieving single base resolution for genetic intolerance. As genetic data accumulates in the coming years, our method will become more and more accurate in predictions of intolerance at the single base pair level (**Supp. Figure S2 and S3**).

To quantify genetic tolerance in genes, protein domains and domain homologues (**Figure 1**) we made use of the  $d_N/d_S$  score rather than other well-established tolerance scores such as pLI,<sup>50</sup> RVIS,<sup>8</sup> and subRVIS.<sup>9</sup> The  $d_N/d_S$  metric was originally intended for detecting selective evolutionary pressure in protein-coding regions and genomes,<sup>81–83</sup> and has previously been used by us and others to measure genetic tolerance and predict disease genes.<sup>10,59,84</sup> Our choice for this score was motivated by the fact that the mentioned tolerance scores typically capture a

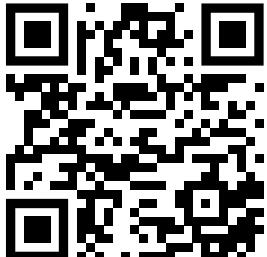
more general notion of tolerance to genetic variation and are not designed to measure tolerance for any specific genic region of interest.

Contrary to our expectations we found that 54% of disease-causing missense variants are evolutionary variable. There are some explanations why we find this result: Firstly, we did not take into account whether disease-causing variants asserted their effect in a dominant or a recessive fashion. We know that mutations in dominant disease genes are in general more conserved than mutations in recessive genes. Secondly, we know that not all disease-causing variants have the same severity in terms of fitness. For example, mutations causing infertility will be much more selected against than mutations causing genetic deafness. Thirdly, a large percentage of HGMD DM variants used to be present in recent population databases and may therefore be incorrect.<sup>85</sup> Although in the version we used, this number was significantly reduced, some may still be present.<sup>86,87</sup> Finally, our comparison does not account for unobserved (potentially lethal) variants, as many of these variants are likely to have never been observed, nor ever will be.

In our meta-domains, we tested whether high-frequency missense variants with an allele frequency  $> 0.1\%$  in ExAC are repeatedly enriched or depleted on Pfam domain consensus positions. This strict cut-off of 0.1% may cause us to miss variants with allele frequencies smaller than 0.1% at corresponding positions in homologues. We choose this cut-off in order to exclude the possibility of artefacts in the ExAC database, and for increasing the likelihood that variation is truly benign. Setting a stricter threshold such as 0.5% decreases the number of ExAC missense variants in meta-domains by 56%. Allowing for a less stringent cut-off will add a substantial amount of genetic variation to our model that would improve our sensitivity, but likely at the cost of specificity (**Supp. Figure S4, Supp. Data S9**). We expect there is still much to be gained from these ‘rare’ variants found in population cohorts. Furthermore we note that by aggregating genetic variation, the specific context such as haplotype information or interactions with other proteins, may be lost. An aggregation may only encapsulate general biological or molecular functions attributed to the domain. Nonetheless, we believe these meta-domains can be used to better interpret variants of unknown significance simply based on our pre-calculated meta-domains (**Supp. Data S5 and S8**), but also by incorporating these results in existing methods for variant effect prediction.

## Supporting Information

All supplementary information can be found online with the published article at



<https://doi.org/10.1002/humu.23313>





# 3

## Chapter 3

---

MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains

**Laurens van de Wiel**, Coos Baakman, Daan Gilissen,  
Joris A. Veltman, Gerrit Vriend, and Christian Gilissen

Published in Human Mutation  
August 2019; 40(8):1030-8

## Abstract

The growing availability of human genetic variation has given rise to novel methods of measuring genetic tolerance that better interpret variants of unknown significance. We recently developed a concept based on protein domain homology in the human genome to improve variant interpretation. For this purpose, we mapped population variation from the Exome Aggregation Consortium (ExAC) and pathogenic mutations from the Human Gene Mutation Database (HGMD) onto Pfam protein domains. The aggregation of these variation data across homologous domains into meta-domains allowed us to generate amino acid resolution of genetic intolerance profiles for human protein domains.

Here, we developed MetaDome, a fast and easy-to-use web server that visualizes meta-domain information and gene-wide profiles of genetic tolerance. We updated the underlying data of MetaDome to contain information from 56,319 human transcripts, 71,419 protein domains, 12,164,292 genetic variants from gnomAD, and 34,076 pathogenic mutations from ClinVar. MetaDome allows researchers to easily investigate their variants of interest for the presence or absence of variation at corresponding positions within homologous domains. We illustrate the added value of MetaDome by an example that highlights how it may help in the interpretation of variants of unknown significance. The MetaDome web server is freely accessible at <https://stuart.radboudumc.nl/metadome>.

## Acknowledgements

This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (916-14-043 to C.G. and 918-15-667 to J.A.V.), and from the Radboud Institute for Molecular Life Sciences, Radboud university medical center (R0002793 to G.V.). We thank Hanka Venselaar for her critical reading of the manuscript.

## Introduction

The continuous accumulation of human genomic data has spurred the development of new methods to interpret genetic variants. There are many freely available web servers and services that facilitate the use of these data by non-bioinformaticians. For example, the ESP Exome Variant Server<sup>58,88</sup> and the Genome Aggregation Database (gnomAD) browser<sup>50,89</sup> help locate variants that occur frequently in the general population. These services are used for the interpretation of unknown variants based on the assumption that variants occurring frequently in the general population are unlikely to be relevant for patients with Mendelian disorders.<sup>90</sup> There are also methods that derive information from these large human genetic databases. For example genetic intolerance, which is commonly used to interpret variants of unknown significance by assessing whether variants stand out because they occur in regions that are genetically invariable in the general population.<sup>9,10</sup> Examples of such methods are RVIS<sup>8</sup> and subRVIS.<sup>9</sup> The strongest evidence for the pathogenicity of a genomic variant comes from the presence of that variant in any of the clinically relevant genetic variant databases such as the Human Gene Mutation Database (HGMD)<sup>91</sup> or the public archive of clinically relevant variants (ClinVar).<sup>54</sup> These databases are gradually growing in the amount of validated pathogenic information.

Another way to provide evidence for the pathogenicity of a genomic variant is to observe the effect of that variant in homologous proteins across different species. Mutations at corresponding locations in homologous proteins are found to result in similar effects on protein stability<sup>40</sup> and can facilitate variant interpretation between disease genes and their paralogues.<sup>92</sup> Finding homologous proteins is one of the key applications of BLAST.<sup>38</sup> Transferring information between homologous proteins is one of the oldest concepts in bioinformatics, and can be achieved by performing a multiple sequence alignment (MSA) and locating equivalent positions between the protein sequences. We have previously used this concept and showed that it also holds for homologous Pfam protein domain relationships within the human genome. We found that ~71-72% of all disease-causing missense variants from HGMD and ClinVar occur in regions translating to a Pfam protein domain and observed that pathogenic missense variants at equivalent domain positions are often paired with the absence of population-based variation and *vice versa*.<sup>93</sup> By aggregating variant information over homologous protein domains, the resolution of genetic tolerance per position is increased to the number of aligned positions.

Similarly, the annotation of pathogenic variants found at equivalent domain positions also assists the interpretation of variants of unknown significance. This use of variant information from homologous protein domains was dubbed ‘meta-domains’. We realized that this type of information could be of great benefit to the genetics community and therefore developed ‘MetaDome’.

MetaDome is a freely available web server that uses our concept of meta-domains to optimally use the information from population-based and pathogenic variation datasets without the need of a bioinformatics intermediate. MetaDome is easy to use and utilizes the latest population datasets by incorporating the gnomAD and ClinVar datasets.

## Methods

### Software architecture of MetaDome

MetaDome is developed in Python v3.5.1<sup>94</sup> and makes use of the Flask framework v0.12.4<sup>95</sup> for the web server part which communicates between the front-end, the back-end, and the database. The software architecture (**Supp. Figure S5**) follows the Domain-driven design paradigm.<sup>96</sup> The entities in the domain part of this software architecture are rich data representations that are based on the internal database (**Creating the mapping database**) and annotations from external resources. These entities are stored after their first creation and afterward directly used for data retrieval to make the lookup in MetaDome as efficient as possible. The code is open source and can be found at our GitHub repository: <https://github.com/cmbi/metadome>. Detailed instructions on how to deploy the MetaDome web server can be found there too.

To ensure MetaDome can be deployed to any environment and provide a high degree of modularity, we have containerized the application via Docker v17.12.1.<sup>97</sup> We use docker-compose v1.17.1 to ensure that different containerized aspects of the MetaDome server can work together. The following aspects are containerized to this purpose: 1.) The Flask application, 2.) a PostgreSQL v10 database wherein the mapping database is stored, 3.) a Celery v4.2.0 task queue management system to facilitate the larger tasks of the MetaDome web-based user requests, 4.) a Redis v4.0.11 for task result storage, and 5.) RabbitMQ v3.7 to mediate as a task broker between client and workers. For a full overview of the docker-compose architecture we refer to **Supp. Figure S6**.

The visualization medium of the MetaDome web server is a fully interactive and responsive HTML web page. This page is generated by the Flask framework and the navigation aesthetics are made using the CSS framework Bulma v0.7.1.<sup>98</sup> The visualizations of the various landscapes and the schematic protein are created with JavaScript, JQuery v3.3.1, and the D3 Framework v4.13.0.<sup>99</sup> As the visualization by the D3 Framework is highly dependent on the user's cpu power, so are the visualizations of MetaDome.

3

## Datasets of population and disease-causing genetic variation

MetaDome makes use of single nucleotide variants (SNVs) from population and clinically relevant genetic variation databases. Population variation was obtained from the gnomAD r2.0.2 VCF file by selecting all synonymous, nonsense, and missense variants that meet the PASS filter criteria. Variants meeting the PASS criteria are considered to be true variants.<sup>50</sup> The variants in the VCF file from ClinVar release 2018\_05\_03 with disease-causing (Pathogenic) status are used as the disease-causing SNVs in MetaDome.

## Creating the mapping database

MetaDome stores a complete mapping between genomic, protein positions, and all domain annotations (**Supp. Figure S7**) in a PostgreSQL relational database.<sup>100</sup> This mapping is auto-generated and stored in the PostgreSQL database by the MetaDome web server upon the first run. The genomic positions consist of each chromosomal position in the protein-coding transcripts of the GENCODE release 19 GRCh37.p13 Basic set.<sup>63</sup> The protein positions correspond to protein sequence positions in the UniProtKB/Swiss-Prot Release 2016\_09 databank entries for the human species.<sup>64</sup> These mappings are created with Protein-Protein BLAST v2.2.31+<sup>62</sup> for each protein-coding translation in the GENCODE Basic set to human canonical and isoform Swiss-Prot protein sequences. We exclude sequences that do not start with a start codon (i.e. ATG encoding for methionine), or end with a stop codon. We checked if the cDNA sequence of the transcripts match the GENCODE translation via Biopython's translate function,<sup>101</sup> if they are not identical then these are excluded too. The global information on the transcript (e.g. identifiers, sequence length) is registered in the database in the table 'genes' and, for each Swiss-Prot entry with an identical sequence match, the global information is stored in the table 'proteins'. All tables are indexed by the fields that are used in the lookups.

Next, for each identical match between translation and Swiss-Prot sequence a ClustalW2 v2.1<sup>39</sup> alignment is made between these two sequences. Each nucleotide's genomic position is mapped to the protein position and stored in the 'mappings' table. Each entry in mapping represents a single nucleotide of a codon and is linked to the corresponding entry in the 'genes' and 'proteins' table (i.e. the corresponding GENCODE translation, transcription and Swiss-Prot sequence).

Each Swiss-Prot sequence in the database is annotated via InterProScan v5.20-59.0<sup>65</sup> for Pfam-A v30.0 protein domains<sup>41</sup> and the results are stored in the 'interpro\_domains' table. After the construction of the database is finished, all meta-domain alignments can be constructed.

### **Composing a meta-domain**

Meta-domains consist of homologous Pfam protein domain instances that are annotated using InterproScan. Meta-domains consist of domains that have at least two homologues within the human genome. MSAs are made using a three step process. 1.) Retrieve all sequences for the domain instances, 2.) Retrieve the Pfam HMM corresponding to the Pfam identifier annotated by InterproScan, and 3.) Use HMMER 3.1b2<sup>67</sup> to align the sequences from the first step. The resulting Stockholm format MSA files can be inspected with alignment visualization software like Jalview.<sup>102</sup> In this Stockholm formatted file, all columns that correspond to the domain consensus represent the same homologous positions.

These Stockholm files are retrieved by the MetaDome web server when a user request meta-domain information for a position of their interest. Upon retrieval of this Stockholm file, the mapping database is used to obtain the corresponding genomic positions for each residue. These genomic positions are subsequently used to retrieve corresponding gnomAD or ClinVar variation.

### **Computing genetic tolerance and generating a tolerance landscape**

The non-synonymous over synonymous ratio, or  $d_N/d_S$  score, is used to quantify genetic tolerance. This score is based on the observed (obs) missense and synonymous variation in gnomAD ( $missense_{obs}$  and  $synonymous_{obs}$ ). This score is corrected for the sequence composition by taking into account the background (bg) of possible missense and synonymous variants based on the codon table  $missense_{bg}$  and  $synonymous_{bg}$ ):

$$d_N/d_S = \frac{\text{missense}_{obs}/\text{missense}_{bg}}{\text{synonymous}_{obs}/\text{synonymous}_{bg}}$$

The tolerance landscape computes this ratio as a sliding window of size 21 (i.e. ten residues before and ten after the residue of interest) over the entirety of the gene's protein, similar to the Missense Tolerance Ratio (MTR) presented by.<sup>103</sup> The edges (e.g. start and end) are therefore a bit noisy as they are not the result of averaging over a full length window.

3

## Results

### Accessibility

The MetaDome web server is freely accessible at <https://stuart.radboudumc.nl/metadome>. MetaDome features a user-friendly web interface and features a fully interactive tour to get familiar with all parts of the analysis and visualizations.

All source code and detailed configuration instructions are available in our GitHub repository: <https://github.com/cmbi/metadome>.

### The underlying database: a mapping between genes and proteins

The MetaDome web server queries genomic datasets in order to annotate positions in a protein or a protein domain. Therefore, the server needs access to genomic positional information as well as protein sequence and protein domain information. The database maps GENCODE gene translations to entries in the UniProtKB/Swiss-Prot databank in a per-position manner and corresponding protein domains or genomic variation. With respect to our criteria to map gene translations to proteins (**Methods; creating the mapping database**), 42,116 of the 56,319 full-length protein-coding GENCODE Basic transcripts for 19,728 human genes are linked to 33,492 of the 42,130 Swiss-Prot human canonical or isoform sequences. Of the total 591,556 canonical and isoform sequences present in Swiss-Prot, 42,130 result from the Human species. The resulting mappings contain 32,595,355 unique genomic positions that are linked to 19,226,961 residues in Swiss-Prot protein sequences.

71,419 Pfam domains are linked to 30,406 of the Swiss-Prot sequences in our database. Of these Pfam domain instances, 5,948 are from a unique Pfam domain

family and 3,334 of these families have two or more homologues and are therefore suitable for meta-domain construction. Thus, by incorporating every protein-coding transcript, instead of only the longest ones, we increase the previously 2,750<sup>93</sup> meta-domains to 3,334. These meta-domains, on average, consist of 16 human protein domain homologues with a protein sequence length of 158 residues. **Table 1** summarizes the counting statistics for sequences, domains, etc.

Database	What	# of entries
GENCODE	Protein-coding genes	20,345
MetaDome	Protein-coding genes	19,728
GENCODE	Protein-coding transcripts	57,005
MetaDome	Protein-coding transcripts	56,319
Swiss-Prot	Canonical and isoform protein sequences	591,556
Swiss-Prot	Human canonical and isoform protein sequences	42,130
MetaDome	Gene translations identically mapped to a canonical or isoform protein sequence	42,116
MetaDome	Canonical and isoform protein sequences	33,492
MetaDome	Pfam protein domain regions	71,419
MetaDome	Unique Pfam protein domain families	5,948
MetaDome	Unique Pfam protein domain families with two or more within-human occurrences	3,334
MetaDome	Chromosome to protein position mappings	70,261,143
MetaDome	Unique chromosome positions	32,595,355
MetaDome	Unique residues (as part of a protein)	19,226,961
MetaDome	Unique protein sequences with at least one Pfam domain annotated	30,406

**Table 1.** Statistics on the number of entries present in GENCODE, Swiss-Prot, and our mapping database.

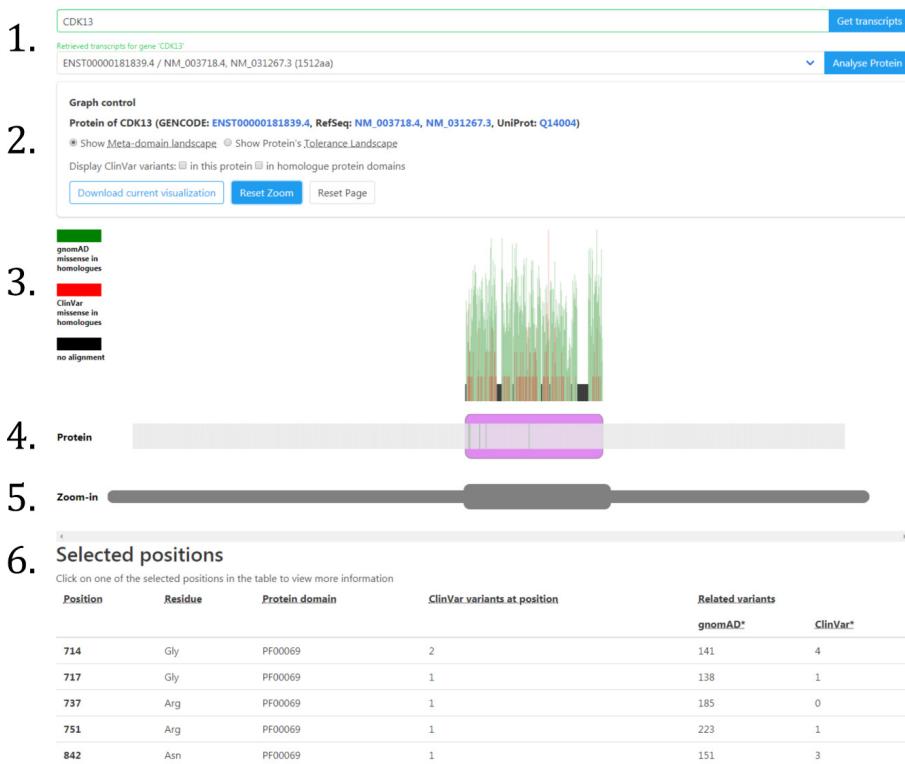
## How to use the MetaDome web server

At the welcome page users are offered the option to start an interactive tour or start with the analysis. The navigation bar at the top is available throughout all web pages in MetaDome and allow for further navigation to the 'About', 'Method', 'Contact' page (**Supp. Figure S1**). The user can fill in a gene symbol in the 'gene of interest' field and is aided by an auto-completion to help you find your gene of interest more easily (**Supp. Figure S2**). Clicking the 'Get transcripts' fills all GENCODE transcripts for that gene in the dropdown box. Only the transcripts that

are mapped to a Swiss-Prot protein can be used in the analysis, the others are displayed in grey (**Supp. Figure S3**).

Clicking the 'Start Analysis' button starts an extensive query to the back-end of the web server for the selected transcript. Firstly, all the mappings are retrieved for the transcript of interest. Secondly, the entire transcript is annotated with ClinVar and gnomAD single nucleotide variants (SNVs) and Pfam domains. Thirdly, if there are any Pfam domains suitable for meta-domain relations then all mappings for those regions are gathered and annotated with ClinVar and gnomAD variation (**methods; Composing a meta-domain**).

The web-page provided to the user as a result of the 'Analyse Protein' can best be explained using an example. Therefore, we have generated this result for gene *CDK13* for transcript 'ENST00000181839.4' (**Figure 1**). The result page features four main components that we will describe from top to bottom. Located at the top is the graph control field. Directly below the graph control is the landscape view of the protein. Below the landscape view, a schematic and interactive representation of the protein and an additional representation of the protein which controls the zooming option. Lastly, at the bottom of the page there is the list of selected positions. All of these components are interactive and the various functionalities are described in **Table 2**.



**Figure 1.** MetaDome web server result for the gene CDK13

The result provided by the MetaDome web server for the analysis of gene CDK13 with transcript ENST00000181839.4, as provided in 1.). In 2.), there is additional information that the translation of this transcript corresponds to Swiss-Prot protein Q14004. Here also various alternative visualizations can be selected. The visualization starts by default in the 'meta-domain landscape', a mode selectable in the graph control in 2.). The landscapes are visualized in 3.), and in the meta-domain landscape the domain regions are annotated with missense variation counts found in homologous domains as bar plots. The schematic protein representation, located at 4.), is per-position selectable, and the domains are presented as purple blocks. Selected positions are highlighted in green. The 'Zoom-in' section at 5.) features a selectable greyed-out copy of schematic protein representation that can zoom-in on any part of the protein. Any selected positions are in the list of selected positions in 6.). Here more information can be obtained by clicking on one of these positions. A detailed description of the functionality of each component is described in Table 2.

Component	Functionality
Gene and transcript input field (Figure 1.1)	<ul style="list-style-type: none"> <li>• Input of gene of interest</li> <li>• Retrieving transcripts for gene of interest</li> <li>• Selecting a transcript</li> <li>• Starting the analysis for selected transcript</li> </ul>
Graph control field (Figure 1.2)	<ul style="list-style-type: none"> <li>• Toggling between different landscape representations</li> <li>• Reset the zoom on the landscape</li> <li>• Reset the web page</li> <li>• Toggle ClinVar variants to be displayed in the schematic protein</li> <li>• Download the visual representation</li> </ul>
Landscape view (Figure 1.3)	<ul style="list-style-type: none"> <li>• Displays the meta-domain landscape</li> <li>• Displays the tolerance landscape</li> </ul>
Schematic protein (Figure 1.4)	<ul style="list-style-type: none"> <li>• Displays a schematic representation of the gene's protein with Pfam protein domains annotated</li> <li>• Hovering over a position displays positional information</li> <li>• Clicking on a position highlights the position and adds the position to the list of 'Selected Positions'</li> <li>• Controls the zooming of particular parts of the protein (Figure 1.5)</li> </ul>
Selected Positions (Figure 1.6)	<ul style="list-style-type: none"> <li>• Displays any positions selected in the schematic protein</li> <li>• Displays per selected position: if that position is part of a Pfam protein domain, any known gnomAD or ClinVar variants present at this position, and any variants that are homologously related to this position</li> <li>• Provides more detailed information as a pop-up when clicking on one of the positions in this list.</li> </ul>

**Table 2.** Descriptions of the various functionalities on the MetaDome result page.

Another way to use population-based variation in the context of the entire protein is via the tolerance landscape representation in MetaDome that can be selected in the graph control component (**Figure 1.2**). The tolerance landscape depicts a missense over synonymous ratio (also known as  $K_d/K_s$  or  $d_N/d_S$ ) over a sliding window of 21 residues over the entirety of the protein of interest (e.g. calculated for ten residues left and right of each residue) based on the gnomAD dataset (**methods; Computing genetic tolerance and generating a tolerance landscape; Figure 2A**). Previously, the  $d_N/d_S$  metric has been used by others and us to measure genetic tolerance and predict disease genes,<sup>59,84,104</sup> and it is suitable for measuring tolerance in regions within genes.<sup>10</sup>

## An example of using the MetaDome web server for variant interpretation

The MetaDome analysis result for *CDK13* (**Figure 1**) is the longest protein coding transcript for *CDK13* with a protein sequence length of 1,512 amino acids. In the

resulting schematic protein representation we can observe the Pkinase Pfam protein domain (PF00069) between positions 707 and 998 as the only protein domain in this gene (**Figure 2B**). The Pkinase domain is highly prevalent throughout the human genome with as many as 779 homologous occurrences in human proteins, of which 353 are unique genomic regions. It is the 8th most occurring domain in our mapping database. The meta-domain landscape is the default view mode and shows any missense variation found in homologous domain occurrences throughout the human genome. Population-based (gnomAD) missense variation is displayed in green and pathogenic (ClinVar) missense variation is annotated in red bars, with the height of the bars depicting the number of variants found at each position (**Figure 2B**).

At the 'Display ClinVar variants' the user is provided two options; to highlight all known pathogenic information known for the current protein and/or highlight any ClinVar variants that are present at homologous positions (**Figure 2A**). All ClinVar variants highlighted are displayed in red. In total six known disease-causing SNVs are present in the *CDK13* gene itself according to ClinVar, and these all fall within the Pkinase protein domain. All of these are missense variants. If we add variants

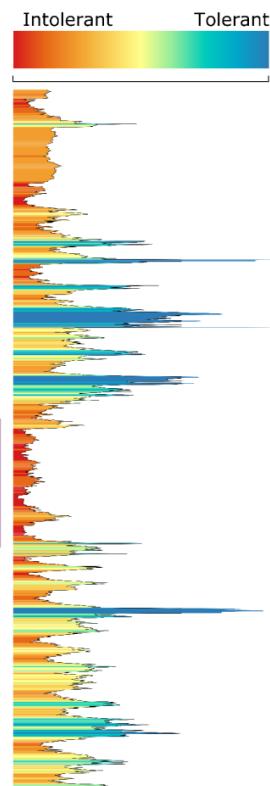
► **Figure 2.** Examples of a MetaDome analysis for the gene *CDK13*

**A.)** The tolerance landscape depicts a missense over synonymous ratio calculated as a sliding window over the entirety of the protein (**methods; Computing genetic tolerance and generating a tolerance landscape**). The missense and synonymous variation are annotated from the gnomAD dataset and the landscape provides some indication of regions that are intolerant to missense variation. In this *CDK13* tolerance landscape the Pkinase Pfam protein domain (PF00069) in purple can be clearly seen as intolerant if compared to other parts in this protein. The red bars in the schematic protein representation correspond to pathogenic ClinVar variants found in this gene and in homologous protein domains. All of these variants are contained in the intolerant region of the landscape.

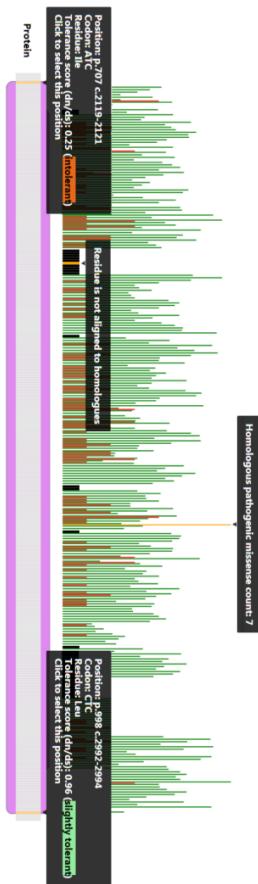
**B.)** A zoom-in on the meta-domain landscape for *CDK13*. The Pkinase Pfam protein domain (PF00069) is located between protein positions 707 and 998 and annotated as a purple box in the schematic protein representation. The meta-domain landscape displays a deep annotation of the protein domain: the green (gnomAD) and red (ClinVar) bars correspond to the number of missense variants found at aligned homologous positions. Unaligned positions are annotated as black bars. All of this information is displayed upon hovering over these various elements.

**C.)** The positional information provides a detailed overview of a position from the 'Selected Positions' list, especially if that position is aligned to domain homologues. Here, for position p.Gly714 we can observe in 1.) the positional details for this specific protein position. In 2.) is any known pathogenic information for this position. We can observe here that for this position there are two known pathogenic missense variants. In 3.) meta-domain information is displayed and we can observe that p.Gly714 is aligned to consensus position 10 in the Pkinase Pfam protein domain and related to 329 other codons. This consensus position has an alignment coverage of 93.5% for the meta-domain MSA. There are also four pathogenic variants found in ClinVar on corresponding homologous positions as can be seen in 4.) and in 5.) there is an overview of all corresponding variants found in gnomAD.

A.



B.



C.

### Positional information (p.714)

**Protein details**  
Protein of CDK13 (GENCODE ENST000001818394, RefSeq NM\_003718.4, NM\_003718.3, Uniprot Q14004)

**Location details**  
Chr: chr7 strand: +  
Gene: 9 (00370857-400393959  
Protein: p.714 (9)  
CDKA: c.216\_216\_GGT  
Tolerance score (in 0.05): 0.25 (intolerant)  
Position is part of protein domain(s): PRO0069

1.

Gene	Position	Variant	Residue change	Type	ClinVar.ID
CDK13	c.7400_7400GAT>CTT	G>A	Gly>Arg	missense	375738
CDK13	c.7400_7400GAT>CTT	G>A	Gly>Arg	missense	449234

2.

**Known pathogenic ClinVar SNVs at position**

Gene	Position	Variant	Residue change	Type	ClinVar.ID
CDK13	c.7400_7400GAT>CTT	G>A	Gly>Arg	missense	375738
CDK13	c.7400_7400GAT>CTT	G>A	Gly>Arg	missense	449234

3.

**Meta-domain information for domain PRO0069:**  
Aligned to consensus position 0, reduced to 32 other colors throughout the genome (with a 95.5% alignment coverage).

**Potentially ClinVar SNVs at homologous positions:**

Gene	Position	Variant	Residue change	Type	ClinVar.ID
MAK	chr6:10830945	G>A	Gly>Ser	missense	29783
PRKG1	chr19:54401351	G>A	Gly>Ser	missense	42129
CTT	chr12:21202854	G>T	Gly>Val	missense	254134
PRKG1	chr14:30095714	G>A	Gly>Arg	missense	375740

4.

**Variants in gnomAD SNVs at homologous positions:**

Gene	Position	Variant	Residue change	Type	gnomAD Frequency	Allele
SPP1	chrX:130470703	C>A	Gly>Gly	synonymous	0.000123	-
MONK1	chr13:141046263	C>A	Gly>Val	missense	0.000008	-
PNCK	chrX:12989151	C>G	Gly>Arg	missense	0.000006	-
MELK	chr9:36583624	G>T	Gly>Val	missense	0.000029	-

5.

Gene	Position	Variant	Residue change	Type	gnomAD Frequency	Allele
SPP1	chrX:130470703	C>A	Gly>Gly	synonymous	0.000123	-
MONK1	chr13:141046263	C>A	Gly>Val	missense	0.000008	-
PNCK	chrX:12989151	C>G	Gly>Arg	missense	0.000006	-
MELK	chr9:36583624	G>T	Gly>Val	missense	0.000029	-

found in homologous domains there are 64 positions with one or more reported pathogenic variants (**Supp. Data S1**). Four of these positions overlap with the positions on which ClinVar variants were found in the gene itself and on position p.883 (**Supp. Figure S4**) we can observe a peak of eight missense variants annotated from other protein domains.

MetaDome helps to look in more detail to a position of interest. If we do this for protein position 714 (**Figure 2C**) in *CDK13* we find that it corresponds to consensus position 10 in the Pkinase domain (PF00069). At this position in *CDK13* there are two variants reported in ClinVar: p.Gly714Arg (ClinVar ID: 375738) submitted by,<sup>105</sup> and p.Gly714Asp (ClinVar ID: 449224) submitted by GeneDX. The first is reported as a *de novo* variant and is associated to Congenital Heart Defects, Dysmorphic Facial Features, and Intellectual Developmental Disorder. For the second there is no associated phenotype provided. As MetaDome annotates variants reported at homologous positions, we can find even more information for this particular position. At the homologues aligned to this position we find a variant of identical change in *PRKD1*: p.Gly600Arg (ClinVar ID: 375740) reported as pathogenic and *de novo* in the same study.<sup>105</sup> It is also associated to Congenital Heart Defects as well as associated to Ectodermal Dysplasia. There are three more reported pathogenic variants aligned to this position: *MAK*:p.Gly13Ser (ClinVar ID: 29783) associated to Retinitis Pigmentosa 62,<sup>106</sup> *PRKCG*:p.Gly360Ser (ClinVar ID: 42129) associated to Spinocerebellar Ataxia Type14,<sup>107</sup> and *CIT*:p.Gly106Val (ClinVar ID: 254134) associated to Microcephaly 17, primary, autosomal recessive.<sup>106</sup> These homologously related pathogenic variants and the severity of the associated phenotypes contributes to the evidence that this particular residue may be important at this position. Further evidence can be found from the fact that in human homologue domains this residue is extremely conserved. There are 330 unique genomic regions encoding for a codon aligned to this position (**Supp. Data S2**). Only in the gene *PIK3R4* (ENST00000356763.3) does this codon encode for another residue than Glycine, namely a Threonine at position p.Thr35.

In the same way that we explored pathogenic ClinVar variation we can also explore the variation reported in gnomAD. In *CDK13* at protein position 714 there is no reported variant in gnomAD, but there are homologously related variations. There are 65 missense variants with average allele frequency of 1.24E-05 and 76 synonymous with average allele frequency 8.71E-03 and there is no reported nonsense variation (**Supp. Data S1**).

When we inspect the tolerance landscape for *CDK13* (**Figure 2A**) we can see that all of the ClinVar variants (either annotated in *CDK13* or related via homologues) fall within the Pkinase Pfam protein domain (PF00069). In addition, the protein domain can clearly be seen as more intolerant to missense variation as compared to other parts of this protein, thereby supporting the ClinVar variants likely pathogenic role.

3

## Conclusion

The MetaDome web server combines resources and information from different fields of expertise (e.g. genomics and proteomics) in order to increase the power in analysing population and pathogenic variation by transposing this variation to homologous protein domains. Such a transfer of information is achieved by a per-position mapping between the GENCODE and Swiss-Prot databases. 79.4% of the Human Swiss-Prot protein sequences are of identical match to one or more of 42,116 GENCODE transcripts. This means that 25.7% of the GENCODE transcriptions differ in mRNA but translate to the same Swiss-Prot protein sequence. GENCODE previously reported that this is due to alternative splicing, of which a substantial proportion only affect untranslated regions (UTRs) and thus have no impact on the protein-coding part of the gene.<sup>108</sup>

MetaDome is especially informative if a variant of interest falls within a protein domain that has homologues. This is highly likely as 43.6% of the positions in the MetaDome mapping database are part of a homologous protein domain. Pathogenic missense variation is also highly likely to fall within a protein domain as we previously observed for 71% of HGMD and 72% of ClinVar pathogenic missense variants.<sup>93</sup> By aggregating variation over protein domain homologues via MetaDome, the resolution of genetic tolerance at a single amino-acid is increased. Furthermore, we can obtain variation that could disrupt the functionality of a protein domain, as annotated throughout the entire human genome, which may potentially be disease-causing. It should be noted, that by aggregating genetic variation in this way the specific context such as haplotype information or interactions with other proteins may be lost. Aggregation via meta-domains only encapsulates general biological or molecular functions attributed to the domain. Nonetheless, we believe MetaDome can be used to better interpret variants of unknown significance through the use of meta-domains and tolerance landscapes as we have shown in our example.

As more genetic data accumulates in the years to come, MetaDome will become more and more accurate in predictions of intolerance at the base-pair level and the meta-domain landscapes will become even more populated with variation found in homologue protein domains. We can imagine many other ways of integrating this type of information to be helpful for variant interpretation. Future directions for the MetaDome web server could lead to machine learning empowered variant effect prediction, or visualization of the meta-domain information in a protein 3D structure.

## Supporting Information

All supplementary information can be found online with the published article at



3

<https://doi.org/10.1002/humu.23798>



# Chapter 4

---

Spatial clustering of *de novo* missense mutations identifies candidate neurodevelopmental disorder-associated genes

Stefan H. Lelieveld<sup>1</sup>, **Laurens van de Wiel**<sup>1</sup>, Hanka Venselaar, Rolph Pfundt,  
Gerrit Vriend, Joris A. Veltman, Han G. Brunner, Lisenka E.L.M. Vissers<sup>2</sup>,  
and Christian Gilissen<sup>2</sup>

*1, 2: These authors contributed equally*

Published in The American Journal of Human Genetics  
7 September 2017; 101(3):478-84

## Abstract

Haploinsufficiency (HI) is the best characterized mechanism through which dominant mutations exert their effect and cause disease. Non-haploinsufficiency (NHI) mechanisms, such as gain-of-function and dominant-negative mechanisms, are often characterized by the spatial clustering of mutations, thereby affecting only particular regions or base pairs of a gene. Variants leading to haploinsufficiency might occasionally cluster as well, for example in critical domains, but such clustering is on the whole less pronounced with mutations often spread throughout the gene. Here we exploit this property and develop a method to specifically identify genes with significant spatial clustering patterns of de novo mutations in large cohorts. We apply our method to a dataset of 4,061 de novo missense mutations from published exome studies of trios with intellectual disability and developmental disorders (ID/DD) and successfully identify 15 genes with clustering mutations, including 12 genes for which mutations are known to cause neurodevelopmental disorders. For 11 out of these 12, NHI mutation mechanisms have been reported. Additionally, we identify three candidate ID/DD-associated genes of which two have an established role in neuronal processes. We further observe a higher intolerance to normal genetic variation of the identified genes compared to known genes for which mutations lead to HI. Finally, 3D modeling of these mutations on their protein structures shows that 81% of the observed mutations are unlikely to affect the overall structural integrity and that they therefore most likely act through a mechanism other than HI.

## Acknowledgements

This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (916-14-043 to C.G. and 918-15-667 to J.A.V.), the European Research Council (ERC Starting grant DENOVO 281964 to J.A.V.) and from the Radboud Institute for Molecular Life Sciences, Radboud university medical center (R0002793 to G.V.). We thank Stephan Boersma for help with writing the software for analysis of cluster mutations.

De novo mutations affecting protein-coding genes are a major cause of intellectual disability (ID) and other developmental disorders (DDs).<sup>59,109</sup> Several whole exome sequencing (WES) studies have identified ID syndromes molecularly characterized by very specific spatial clustering of de novo missense mutations.<sup>110-113</sup> Similarly, large-scale WES studies of individuals affected by ID/DD have recently leveraged this phenomenon as supporting evidence of the involvement of a gene in disease.<sup>70,114</sup> This spatial clustering of de novo mutations (DNMs) is typical for missense mutations in genes without clear, or limited numbers of, truncating mutations subsequently degraded by nonsense mediated mRNA decay, suggesting that these clustered mutations act through a different mechanism than haploinsufficiency (HI).<sup>115</sup> Alternative pathophysiological mechanisms that might underlie (de novo) mutation clustering are gain-of-function or dominant-negative effects, resulting in the alteration or impairment of specific protein function.<sup>116,117</sup> We note that while spatial clustering is commonly taken to indicate a mechanism different from loss-of-function,<sup>118</sup> this is not an absolute rule, and a loss-of-function mechanism cannot be excluded without functional evidence.<sup>119</sup> Here, we developed a method to identify genes with spatially clustered DNMs and applied this to DNMs identified in a large cohort of individuals with ID/DD.<sup>120</sup>

We downloaded all DNMs occurring in individuals with ID/DD from denovo-db version 1.3<sup>120</sup> identified through WES and whole genome sequencing which were then re-annotated with our in-house variant annotation pipeline. The de novo mutations included in the analysis were previously validated by a second independent method or showed a high validation rate for a subset of de novo mutations. In addition, we added 1,183 de novo variants identified in the exomes of an in-house ID cohort that was previously published.<sup>70</sup> To further reduce the risk of including sequencing artifacts and/or genotyping errors, we excluded all de novo variants that were present more than once in the ExAC dataset (**Table S1**).<sup>50</sup> These efforts resulted in 6,495 protein coding DNMs, including 4,061 missense mutations, in 5,302 individuals with ID/DD (**Table S2**).

We set out to determine for any gene whether the observed de novo missense mutations cluster more than expected compared to random permutations. Hereto, we selected for each the longest representative transcript (i.e. part of the GENCODE basic set)<sup>63</sup> and calculated the geometric mean distance  $\delta_g$  over all missense DNMs on cDNA.  $\delta_g$  was calculated by taking the mean distance

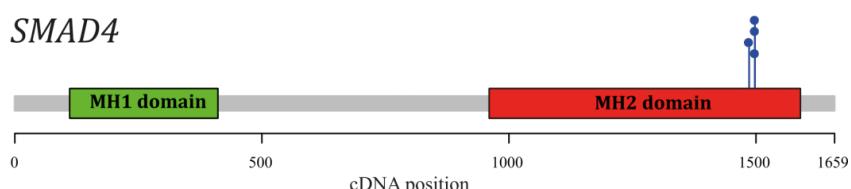
normalized for transcript length  $l$  over all ( $M$ ) combinations of  $x_i$  and  $x_j$  of the missense DNMs (Equation 1.), where  $x$  represents the position for mutation  $i$  and  $j$  respectively. Statistical significance was determined by performing 1.00E+08 (or  $N$ ) permutations and calculating for each permuted geometric mean distance ( $\delta_g$ ) how many times this resulted in the same or smaller geometric mean distance as observed (Equation 2.) Permutation p-values were corrected for multiple testing via Bonferroni procedure based on the 19,280 genes of the Agilent SureSelect v5 exome enrichment kit.

$$\delta_g = \left( \prod_{\substack{i,j=1 \\ i < j}}^M \frac{|x_i - x_j| + 1}{l + 1} \right)^{\frac{1}{M}} \quad 1.$$
$$p = \frac{\sum_{i=1}^N [\delta_g' \leq \delta_g] + 1}{N + 1} \quad 2.$$

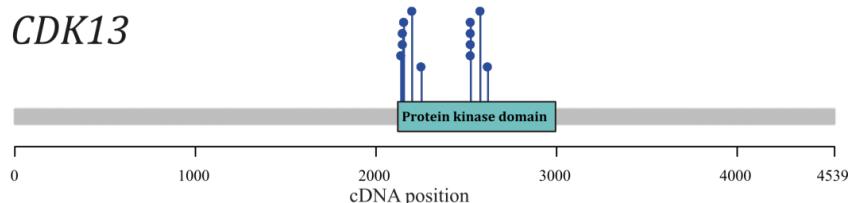
We first validated our method on a dataset of DNMs identified in 2,448 unaffected siblings and healthy control studies<sup>120-126</sup> (**Table S3**). In this cohort, we failed to identify genes for which clustering of de novo missense mutations reached statistical significance (**Table S4**). However, application of our method to the dataset of 4,061 DNMs, containing 583 genes with more than one de novo missense mutation, revealed 15 genes with significant clustering<sup>70,114,127-129</sup> (**Table 1, Figure 1, Figures S1-S15**). In these genes, a total of 107 de novo missense mutations contributed to mutation clustering, ranging from three to 20 mutations per gene with an average distance ranging from 0 to 354 bp. To exclude a correlation between the extent of clustering and the total number of de novo missense mutations analyzed, we applied our method to a cohort of 6,154 de novo missense variants present in Denovo-db excluding the five studies incorporated in the ID/DD cohort, and found no such correlation (**Figure S16**). To examine whether this set of 15 genes is relevant in the context of ID/DD, we compared these genes to a list of 1,541 genes for which mutations are known to cause ID/DD (**Table S5**). This list of genes was a compilation of two manually curated lists of disease related genes including “confirmed” unique genes from DDG2P (n=1,098; **see Web Resources**) and 1,034 genes offered for diagnostic testing in individuals with ID/DD by our

in-house diagnostic facility (**see Web Resources**). Among the 15 identified genes with mutation clustering, we find 12 genes for which mutations have previously been implicated in ID/DD, constituting a significant enrichment ( $p=3.09e-03$ ; Fisher's exact test; **Tables S6 and S7**), and confirming that our method is valid for its purpose.. The inclusion of exome data of two large DDD-studies in both the DDG2P gene list and the ID/DD cohort of this study could introduce a potential bias<sup>109,114</sup>. To exclude such bias we repeated this analysis while excluding the DDD specific genes identified in the two exome studies yielding a significant enrichment ( $p=3.68E-02$ ; **Table S7A-C**).

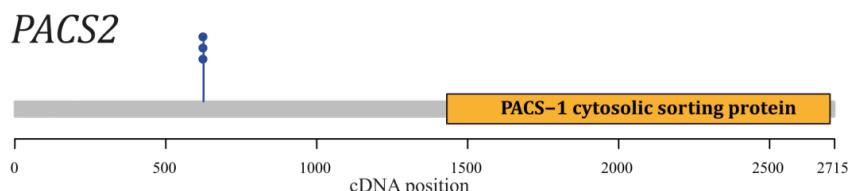
### A SMAD4



### B CDK13



### C PACS2



**Figure 1.** Examples of Identified Genes with Clustering Mutations

Protein domains are annotated based on Pfam HMM search.<sup>41</sup> cDNA locations of de novo missense mutations are depicted by blue pins. Genes shown here are as follows: SMAD4 (A), CDK13 (B), PACS2 (C). Figures visualizing the clustering of de novo missense mutations in the other 12 genes are provided in **Figures S1-S15**.

**Table 1. List of identified genes with clustering de novo missense mutations.** Genes previously known to be involved in neurodevelopmental disorders are indicated in italics. P-values are based on a permutation test ( $N=1.00E+08$ ). Adj. p-values are corrected by Bonferroni correction. The three identified genes with that have not yet been implicated in ID/DD are indicated by an <sup>a</sup>.

Gene name	Transcript ID	# de novo missense	Median distance (bp)	P-value	Adj. p-value
ACTL6B <sup>a</sup>	ENST00000160382	3	0	5.70E-07	1.10E-02
ALG13	ENST00000394780	3	0	1.50E-07	2.89E-03
CDK13	ENST00000181839	12	273	<1.00E-08	<1.93E-04
COL4A3BP	ENST00000380494	6	18	2.60E-07	5.01E-03
GABBR2 <sup>a</sup>	ENST00000259455	3	0	9.00E-08	1.74E-03
GRIN2B	ENST00000609686	11	354	1.57E-06	3.03E-02
KCNH1	ENST00000271751	7	65	1.00E-07	1.93E-03
KCNQ2	ENST00000354587	20	301	5.00E-08	9.64E-04
KIF5C	ENST00000435030	3	0	1.40E-07	2.70E-03
PACS1	ENST00000320580	9	0	<1.00E-08	<1.93E-04
PACS2 <sup>a</sup>	ENST00000458164	3	0	1.50E-07	2.89E-03
PCGF2	ENST00000360797	3	0	1.11E-06	2.14E-02
PPP2R1A	ENST00000322088	4	5	4.60E-07	8.87E-03
PPP2R5D	ENST00000485511	16	10	<1.00E-08	<1.93E-04
SMAD4	ENST00000398417	4	6	1.60E-07	3.08E-03

We also identified three genes with clustered de novo missense mutations that have not yet been implicated in ID/DD: *ACTL6B* (MIM:612458), *GABBR2* (MIM:607340) and *PACS2* (MIM:610423). None of these genes would have been identified based on enrichment for de novo mutations in this cohort (**Table S8**). Further systematic evaluation of gene function supports a role in (neuro)development for two of these genes (**Table 2 and Table S9**). *ACTL6B*, encoding Actin-like 6B (also known as *BAF53B*), is a pivotal co-factor for the SWI/SNF neuron-specific chromatin remodeling complex nBAF, which is required for neural development and dendritic outgrowth.<sup>130,131</sup> Also, *GABBR2*, which is a component of the G-protein-coupled GABA receptor, plays a critical role in the fine-tuning of inhibitory synaptic transmission,<sup>132–134</sup> and other members of the GABA receptor family have already been conclusively linked to neurodevelopmental disorders.<sup>135,136</sup> *GABBR2* was very recently also reported by others to show significant de novo mutation clustering in a neurodevelopmental cohort.<sup>113</sup>

**Table 2. Gene function for candidate genes with clustered mutations.** First column indicates the gene name, second column a summary of the known gene functions; third column indicates whether the gene has physical interactions with other proteins. (See **Table S9** for extended information).

Summary of gene function	Interactions
ACTL6B Belongs to the neuron-specific chromatin remodeling complex (nBAF complex) and is required for postmitotic neural development and dendritic outgrowth.	Complex formation with ACTB, ARID1A, SMARCA2, SMARCA4, SMARCE1, SMARCC1, SMARCC2, SMARCD2, SMARCB1
GABBR2 Postsynaptic GABAB Receptor Activity Regulates Excitatory Neuronal Architecture and Spatial Memory.	Heterodimerization is required for the formation of a functional GABA-B receptor.
PACS2 Multifunctional sorting protein, controlling endoplasmic reticulum-mitochondria communication and Bid-mediated apoptosis.	N/A

Our method might potentially identify clustering based on identical mutations in multiple individuals only as a result of issues in the underlying cohort. It could for instance be that the same individual was included in multiple studies and therefore occurs twice in the cohort. For 99 out of 107 de novo missense mutations (92.5%) occurring in the 15 genes with clustering mutations we could decisively conclude that they occurred as unique events in separate individuals based on a combination of the gender of the affected individual and the presence of additional de novo mutations (**Table S10**). Nevertheless, it might be possible that siblings of affected individuals were included who share a DNM due to parental gonadal mosaicism.<sup>137</sup> Alternatively, DNMs might occur multiple times in disease cohorts as a consequence of a locally increased mutation rate. Examples of the latter might for instance incur a selective growth advantage (i.e. selfish mutations<sup>138</sup>) and thereby result in a pattern of mutational clustering such as known for *FGFR2* (MIM: 176943) mutations in Apert syndrome (MIM: 101200).<sup>138</sup> However, biological relevance for the mutations in the identified genes in the context of ID/DD is suggested by the fact that in our control cohort genes with significant clusters were absent, and that for the majority of our identified genes experimental evidence in literature supports a NHI mutational mechanism (**Table S11**).

We hypothesized that the clustering de novo missense mutations of the 15 genes might exert their effects through mechanisms other than haploinsufficiency. To validate this hypothesis, we compiled a set of 116 genes known for mutations

that exert disease through non-haploinsufficient (NHI) mechanisms. Hereto, we selected for genes that have a “confirmed” status in the DDG2P list, or are present on both the Radboudumc ID/DD diagnostic testing and DDG2P lists (irrespective of the DDG2P status). Furthermore, genes were selected to be (i) dominant (mono-allelic), with the pathophysiological mechanism being either “activating”, “all missense/in frame” and/or “dominant negative” (**Table S12**). In addition, we generated a set of 183 haploinsufficient genes for which mutations are associated with ID/DD from the DDG2P gene list by selecting “loss-of-function” as the “mutation consequence” and “mono-allelic” for the “allelic requirement” in the DDG2P gene list (**Table S13**).

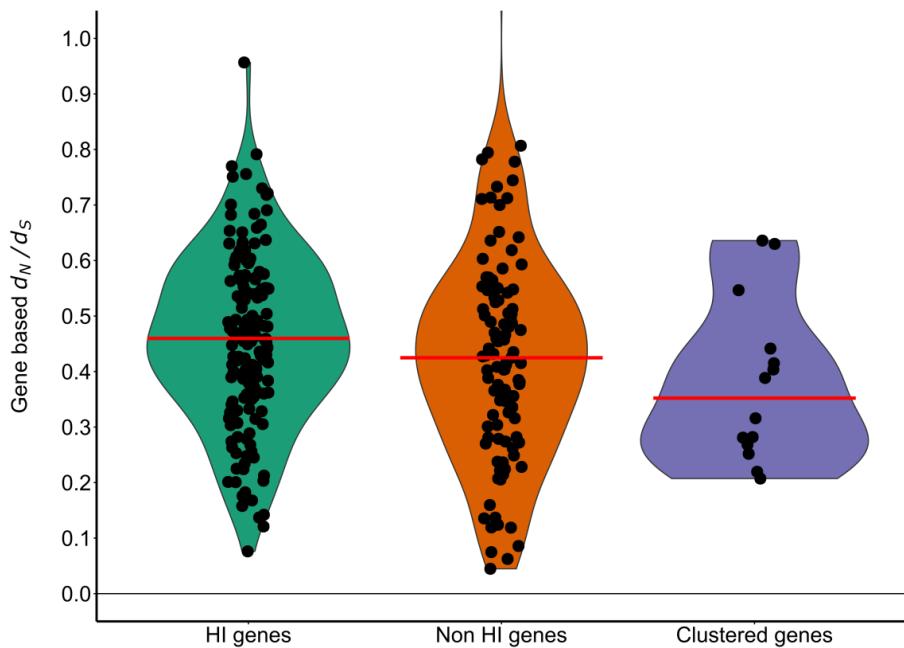
Interestingly, for eight of the 12 genes for which mutations are known to cause ID/DD and for which we identified mutation clustering, the disease mechanism on the constructed gene list was reported to be NHI. For these eight genes, it is either gain-of-function or dominant negative, thereby showing statistical enrichment for NHI mechanisms ( $p=2.66E-03$ , Fisher’s exact test; **Table S14 and S15**). For two of the three remaining genes (*GRIN2B* [MIM:138252] and *SMAD4* [MIM: 600993]) both HI and NHI consequences have been reported,<sup>139-142</sup> suggesting that for mutations in these genes more complex genotype-phenotype relations might exist, where HI and NHI mechanisms cause clinically distinct ID/DD-related disorders. For *KCNQ2* (MIM: 602235), the reported mutational mechanism is HI although a literature search also revealed cases with dominant-negative effects.<sup>143</sup> We also investigated the extent of the evidence for NHI mechanisms and found that extensive functional work of mutations supporting NHI mechanisms has been previously published for eight of the 12 known genes (**Table S11**).

Further we hypothesized that NHI genes should be depleted for truncating mutations in individuals with ID/DD, i.e. mutations resulting in premature translation termination, whereby the mRNA is targeted for nonsense mediated decay. In our initial analyses focusing on de novo missense mutations only, we excluded truncating mutations from our dataset. Retrospectively, we searched for truncating DNM s in the 15 identified genes with clustering de novo missense mutations. We found only three predicted truncating mutations in two of 15 genes, which is significantly less than expected based on the total number of DNM s found in the total cohort for all HI genes ( $p<1.00e-05$ ; Permutation test).

We have previously hypothesized that genes with mutations acting through NHI mechanisms might be more intolerant to normal variation than genes with mutations acting through a HI mechanism for ID/DD.<sup>70</sup> To test for tolerance to variation, existing scores like pLI<sup>50</sup> are not useful as these capture tolerance to mRNA truncating variation rather than tolerance to variation in general. Therefore, we measured tolerance to variation as the ratio of missense over synonymous variation ' $d_N/d_S'$ , which has been used by us and others previously for predicting disease genes.<sup>10,59,84</sup> We downloaded all PASS-filtered single nucleotide variants (SNVs) from ExAC (n=9,035,134) and constructed a ' $d_N/d_S'$ ' measure by counting the unique missense SNVs  $missense_{obs}$ , and the unique synonymous SNVs  $synonymous_{obs}$ , while correcting for sequence composition using the total possible unique missense and synonymous SNVs ( $missense_{bg}$  and  $synonymous_{bg}$  respectively)(**Table S16**):

$$d_N/d_S = \frac{missense_{obs}/missense_{bg}}{synonymous_{obs}/synonymous_{bg}}$$

Based on calculations of these scores for the sets of 116 NHI, and 183 HI genes, we indeed find that genes with mutations acting through a NHI mechanism are significantly more intolerant to missense variation than genes with mutations acting through a HI mechanism (p=2.24e-03; permutation test, **Figure 2**). In line with our hypothesis, also our set of 15 genes with clustered DNMs was significantly less tolerant to missense variation compared to the set of 183 genes with mutations acting through a HI mechanism (p=8.45e-03; permutation test, **Figure 2**).



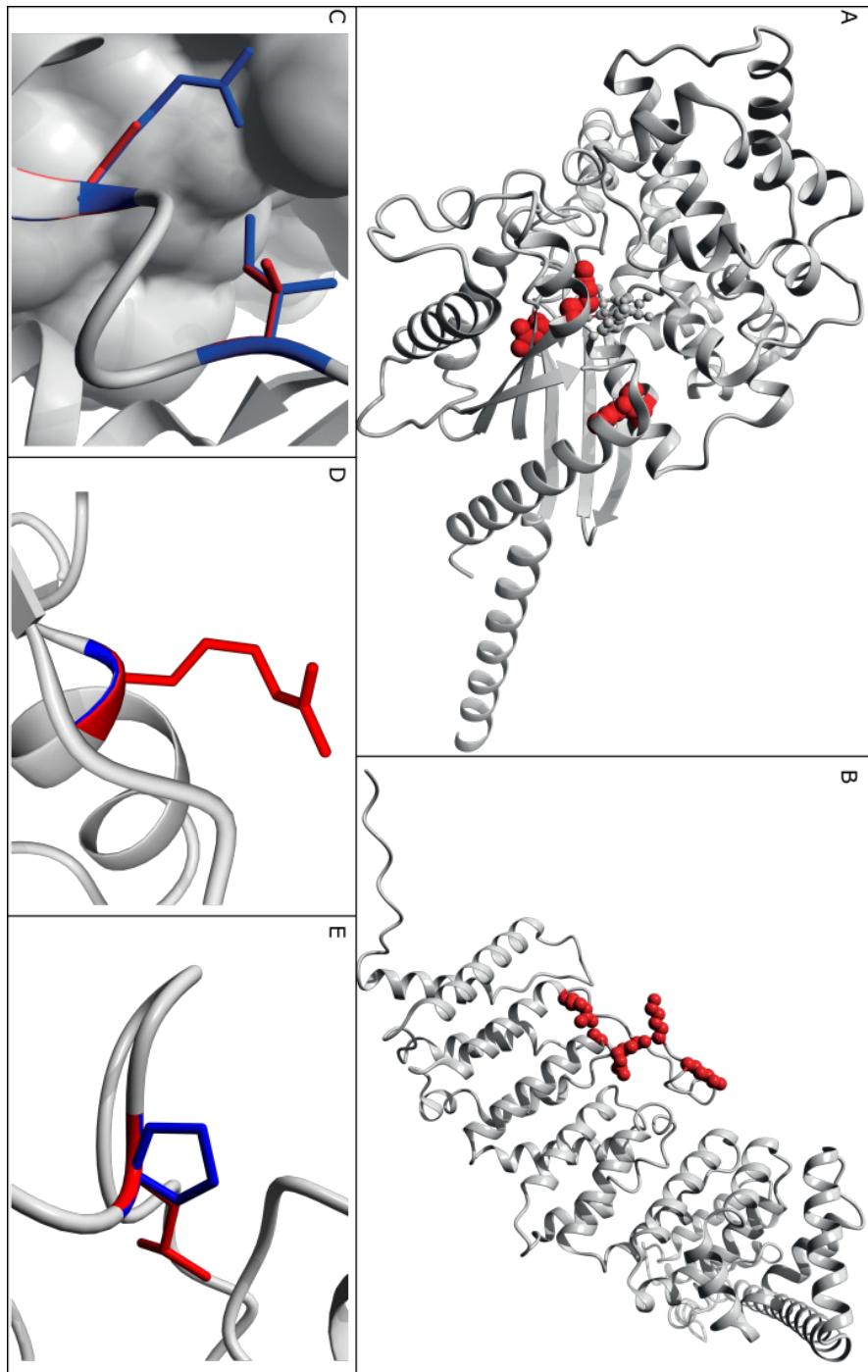
**Figure 2. Intolerance to Missense Variation**

Violin plots show the distribution of the gene-based  $dN/dS$  (y axis) per gene set (x axis). The median  $dN/dS$  is indicated by a red horizontal line. The NHI genes are more intolerant to missense variation than HI genes (HI genes median: 0.460; NHI genes median: 0.428;  $p = 2.24e-03$ ). In addition, the identified genes with clustering mutations are more intolerant to missense variation than HI genes (genes with clustering mutations median: 0.352;  $p = 8.45e-03$ ).

► **Figure 3. Examples of Modeling of Missense Mutations on 3D Protein Structures**

Wild-type residues are marked in blue; de novo mutations are indicated as red globes or lines (Tables S17).

- A.** 3D structure of GNA1, acting through HI, showing that the modeled missense mutations are buried and likely to disrupt protein folding.
- B.** Structure of PPP2R5D, acting through NHI, where the modeled missense mutations affect mostly surface residues and are expected to have no or only local structural effects.
- C.** Zoom-in of known missense variants p.Arg496Cys and p.Ile500Val in SMAD4 known to act through a gain-of-function mechanism. These variants are located on the surface of the monomer and in contact with another SMAD4 monomer.<sup>141</sup>
- D.** Zoom-in of the missense variant p.Gly343Arg in ACTL6B which is located at the surface. The side-chain points toward the solvent, therefore the larger Arginine will fit.
- E.** Zoom-in of the missense variant p.Pro65Leu in PCGF2 close to the interaction site with other molecules.



Modeling of missense mutations in a 3D protein structure is helpful to gain more insight into the possible structural and functional effects.<sup>144</sup> Conceptually, mutations in the core of the protein structure are more likely to prevent proper folding than mutations on the protein surface.<sup>145</sup> The impact of a surface change however depends entirely on the spatial context and is therefore less likely to result in misfolding and subsequent protein degradation.<sup>146</sup> Consequently, de novo disease-causing missense mutations preventing proper folding cause protein degradation, and thus indirectly lead to HI, similar to protein truncating mutations in such genes. To test the hypothesis that our clustered de novo missense mutations do not generally result in HI due to protein misfolding we modeled mutations onto the 3D protein structure using YASARA & WHAT IF Twinset.<sup>14,147</sup> A (partial) protein 3D structure was available or could be created via homology modeling for 10 of the 15 identified genes. We assessed 48 missense mutations on the 3D structure (i.e. buried, at the surface, or semi-buried) and whether the mutation was likely to affect protein folding (no effect, local effect, or large effect; **Figure 3, Table S17**). To compare the results of 3D modeling of clustered mutations, we also modeled 75 de novo disease-causing missense mutations in 25 genes with mutations acting through HI (**Table S13**) for which a structure was available (**Table S17**). For the HI genes, 42% of missense mutations were buried and 34% of mutations were located at the protein surface. In the 10 genes for which a mutational NHI effect is proposed, only 11% of mutations was buried whereas 61% was located at the protein surface ( $p=1.26E-03$ , chi-square test; **Table S17**). Even more strikingly, only 19% of the clustering de novo missense mutations were likely to result in a large structural change that would affect protein function whereas this was observed for 63% of de novo missense mutations in HI genes ( $p = 8.43E-06$ , chi-square test). These results support the notion that the majority of clustered de novo disease-causing missense mutations do not result in haploinsufficiency at the protein structure level, but enact their effect through other mechanisms. Possibly this could be through the functional impairment of protein-protein interactions, as we noted that two of the three candidate ID/DD genes require complex formation or joining of protein subunits (e.g. multimerisation) to be functional (**Table 2**).

In conclusion, we developed a method for the identification of disease genes based on the significance of spatial mutation clustering within a gene. We show that our method successfully identifies genes previously implicated in ID/DD. Moreover,

we identified three genes with similar clustering patterns that we propose as candidate ID/DD genes. Our findings support the concept that these mutations mostly exert their pathogenic effect through disease mechanisms other than haploinsufficiency. Thus, our findings might indicate a larger contribution of non-haploinsufficient mechanisms to ID/DD than previously thought.

## Supporting Information

All supplementary information can be found online with the published article at



<https://doi.org/10.1016/j.ajhg.2017.08.004>





# Chapter 5

---

Evidence for 28 genetic disorders  
discovered by combining healthcare  
and research data

*Joanna Kaplanis<sup>1</sup>, Kaitlin E. Samocha<sup>1</sup>, Laurens van de Wiel<sup>1</sup>, Zhancheng Zhang<sup>1</sup>, Kevin J. Arvai, Ruth Y. Eberhardt, Giuseppe Gallone, Stefan H. Lelieveld, Hilary C. Martin, Jeremy F. McRae, Patrick J. Short, Rebecca I. Torene, Elke de Boer, Petr Danecek, Eugene J. Gardner, Ni Huang, Jenny Lord, Iñigo Martincorena, Rolph Pfundt, Margot R. F. Reijnders, Alison Yeung, Helger G. Yntema, DDD Study, Lisenka E. L. M. Vissers, Jane Juusola, Caroline F. Wright, Han G. Brunner, Helen V. Firth, David R. FitzPatrick, Jeffrey C. Barrett, Matthew E. Hurles<sup>2</sup>, Christian Gilissen<sup>2</sup>, Kyle Retterer<sup>2</sup>*

*1, 2: These authors contributed equally*

Published in Nature  
14 October 2020; 586:757-762

## Abstract

*De novo* mutations in protein-coding genes are a well-established cause of developmental disorders.<sup>114</sup> However, genes known to be associated with developmental disorders account for only a minority of the observed excess of such *de novo* mutations.<sup>114,148</sup> Here, to identify previously undescribed genes associated with developmental disorders, we integrate healthcare and research exome-sequence data from 31,058 parent-offspring trios of individuals with developmental disorders, and develop a simulation-based statistical test to identify gene-specific enrichment of *de novo* mutations. We identified 285 genes that were significantly associated with developmental disorders, including 28 that had not previously been robustly associated with developmental disorders. Although we detected more genes associated with developmental disorders, much of the excess of *de novo* mutations in protein-coding genes remains unaccounted for. Modelling suggests that more than 1,000 genes associated with developmental disorders have not yet been described, many of which are likely to be less penetrant than the currently known genes. Research access to clinical diagnostic datasets will be critical for completing the map of genes associated with developmental disorders.

## Acknowledgements

We thank the families and their clinicians for their participation and engagement, and our colleagues who assisted in the generation and processing of data. Inclusion of RadboudUMC data was in part supported by the Solve-RD project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 779257. This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (917-17-353 to C.G.). The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003). This study makes use of DECIPHER, which is funded by the Wellcome Trust. The full acknowledgements can be found at [www.ddduk.org/access.html](http://www.ddduk.org/access.html). The DDD study authors acknowledge the work of R. Kelsell. Finally, we acknowledge the contribution of an esteemed DDD clinical collaborator, M. Bitner-Glindicz, who died during the course of the study.

It has previously been estimated that around 42–48% of patients with a severe developmental disorder (DD) have a pathogenic *de novo* mutation (DNM) in a protein-coding gene.<sup>114,148</sup> However, most of these patients remain undiagnosed despite the identification of hundreds of DD-associated genes. This indicates that there are more DD-relevant genes to find. Existing methods to detect the gene-specific enrichment of damaging DNMs do not incorporate all of the available information about which variants are more likely to be disease-associated; missense variants and protein-truncating variants (PTVs) vary in their impact on protein function.<sup>11,29,149,150</sup> Known dominant DD-associated genes are strongly enriched in the minority of genes that exhibit strong selective constraint on heterozygous PTVs.<sup>50</sup> To identify additional DD-associated genes, we need to increase our power to detect gene-specific enrichments of damaging DNMs by both increasing sample sizes and improving our statistical methods. In previous studies of pathogenic copy number variations, the use of healthcare data has been key to achieve larger sample sizes than would be possible in a research setting alone.<sup>151,152</sup>

## Identification of 285 DD-associated genes

Following clear consent practices and only using aggregate, deidentified data, we pooled DNMs from patients with a DD from three centres: GeneDx (a US-based diagnostic testing company), the Deciphering Developmental Disorders study and Radboud University Medical Center. We performed stringent quality control on variants and samples to obtain 45,221 coding and splicing DNMs in 31,058 individuals (**Supplementary Fig. 1; Supplementary Table 1**), including data on 24,348 trios that have not previously been published. These DNMs included 40,992 single-nucleotide variants (SNVs) and 4,229 insertions or deletions (indels). The three cohorts have similar clinical characteristics, male-to-female ratios, enrichments of DNMs by mutational class and prevalences of known disorders (**Supplementary Fig. 2**).

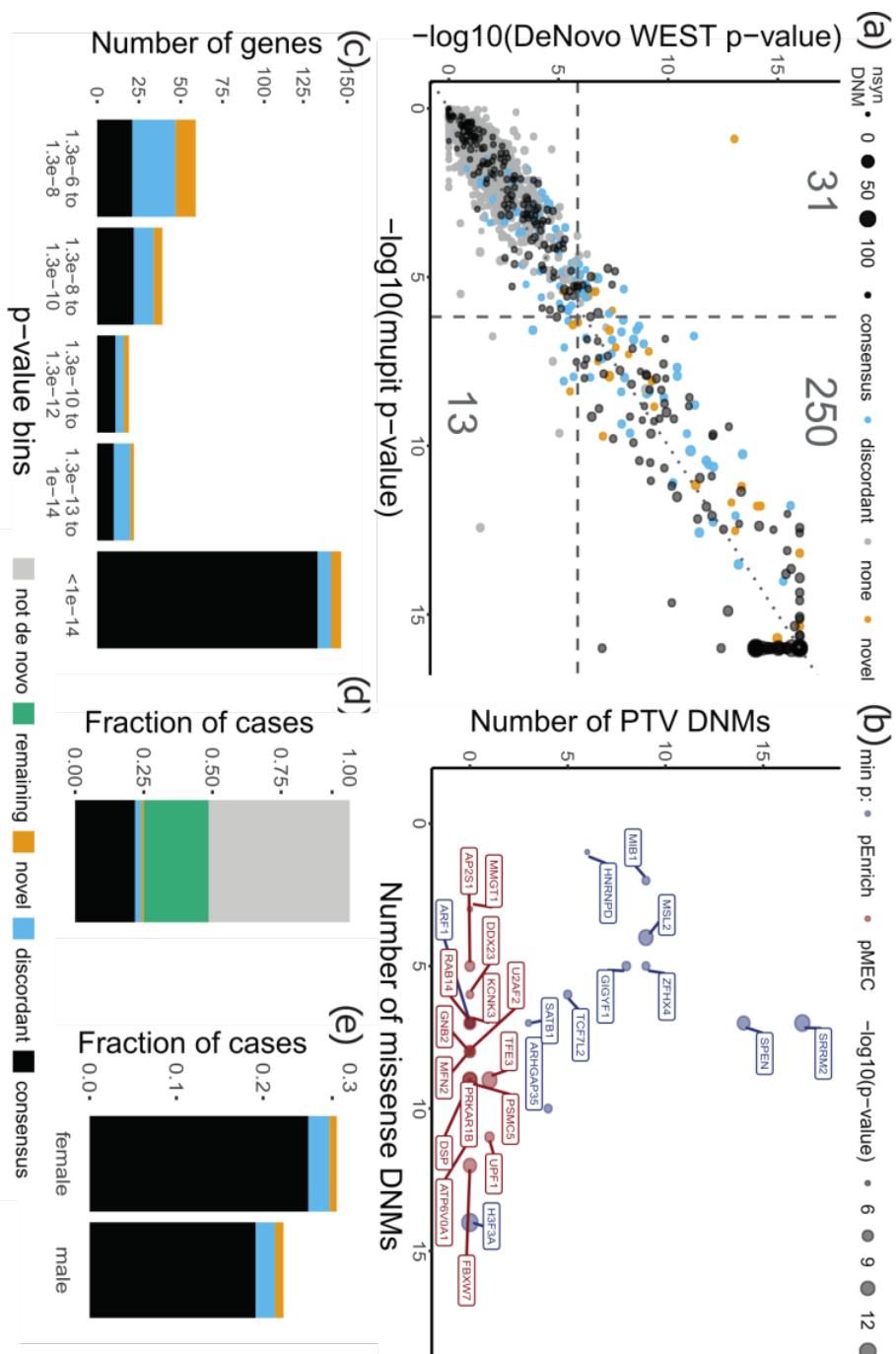
To detect gene-specific enrichments of damaging DNMs, we developed a method named DeNovoWEST (*De Novo* Weighted Enrichment Simulation Test, <https://github.com/queenjobo/DeNovoWEST>). DeNovoWEST scores all classes of sequence variants on a unified severity scale based on empirically estimated positive predictive values of being pathogenic (**Supplementary Fig. 3, 4**). We perform two tests per gene: an enrichment test on all nonsynonymous DNMs

and a test designed to detect genes that probably act through an altered-function mechanism, which combines a missense enrichment test with a missense clustering test. We then applied a Bonferroni multiple-testing correction accounting for the number of genes ( $n = 18,762$ ) and two tests per gene.

We first applied DeNovoWEST to all individuals in our cohort and identified 281 significantly enriched genes, 18 more than when using our previously published method<sup>114</sup> (**Figure 1a; Supplementary Fig. 5**). The majority (196 out of 281; 70%) of the significantly enriched genes already had sufficient evidence of an association with DDs to be considered of diagnostic utility (as of late 2019) by all three centres, and we refer to these genes as ‘consensus’ genes. A further 54 out of 281 of the significantly enriched significant genes were previously considered diagnostic by one or two centres (‘discordant’ genes). Applying DeNovoWEST to synonymous DNMs, as a negative control analysis, identified no significantly enriched genes (**Supplementary Fig. 6**).

To discover novel DD-associated genes with greater power, we applied DeNovoWEST to DNMs in patients without damaging DNMs in consensus genes (we refer to this subset as ‘undiagnosed’ patients) and identified 94 significant genes (**Supplementary Fig. 7; Supplementary Table 2**), of which 33 were putative ‘novel’ DD-associated genes. To ensure robustness to potential mutation rate variation between genes, we determined whether any of the putative novel DD-associated genes had significantly more synonymous variants in the Genome

**►Figure 1: Results of DeNovoWEST analysis.** **A.** Comparison of  $P$  values using DeNovoWEST versus the previous published method (mupit),<sup>114</sup> run on the full cohort. Dashed lines indicate the threshold for genome-wide significance (one-sided, Bonferroni correction). Point size is proportional to the number of nonsynonymous DNMs in our cohort. The number of genes that fall into each quadrant are annotated. **B.** The number of missense and PTV DNMs in the novel genes. Point size is proportional to the  $-\log_{10}(P)$  value of the analysis of the undiagnosed subset. Point colour corresponds to which test  $P$  value was more significant: blue, the nonsynonymous enrichment test ( $p\text{Enrich}$ ); red, the missense enrichment and clustering test ( $p\text{MEC}$ ). H3-3A is also known as H3F3A. **C.** The distribution of significant  $P$  values from analysis of the undiagnosed subset for discordant and novel genes;  $P$  values for consensus genes come from the full cohort analysis. The number of genes in each  $P$ -value bin is coloured by diagnostic gene group ( $n = 285$  significant genes; one-sided Bonferroni-corrected  $P$  values). **D.** The fraction of patients ( $n = 31,058$ ) with a nonsynonymous mutation in each diagnostic gene group. Green, the remaining fraction of patients (the offspring of the parent-offspring trios) expected to have a pathogenic de novo coding mutation; grey, the fraction of patients that are likely to be explained by other factors. **E.** The fraction of patients with a nonsynonymous mutation in each diagnostic gene group split by sex ( $n = 13,636$  female patients;  $n = 17,422$  male patients). In all panels, black, blue and orange represents consensus, discordant and novel genes, respectively.



Aggregation Database (gnomAD)<sup>51</sup> of population variation than expected under our null mutation model. We identified 11 out of 33 genes with a significant excess of synonymous variants. For these 11 genes, we repeated the DeNovoWEST test, increasing the null mutation rate by the ratio of observed to expected synonymous variants in gnomAD. Five of these genes fell below our exome-wide significance threshold and were removed, leaving 28 novel genes, with a median of 10 nonsynonymous DNMs (**Fig. 1c; Supplementary Table 3**). There were 314 patients with nonsynonymous DNMs in these 28 genes (1.0% of our cohort); all of these DNMs were inspected in the Integrative Genomics Viewer (IGV)<sup>153</sup> and, of the 198 patients for which experimental validation was attempted, all variants were confirmed to be DNMs. The DNMs in these novel genes were distributed randomly across the three datasets (no genes with  $P < 0.001$ , heterogeneity test). In addition, 6 of the 28 novel DD-associated genes were corroborated by OMIM entries or publications, including *TFE3*, which was described in two recent publications.<sup>154,155</sup>

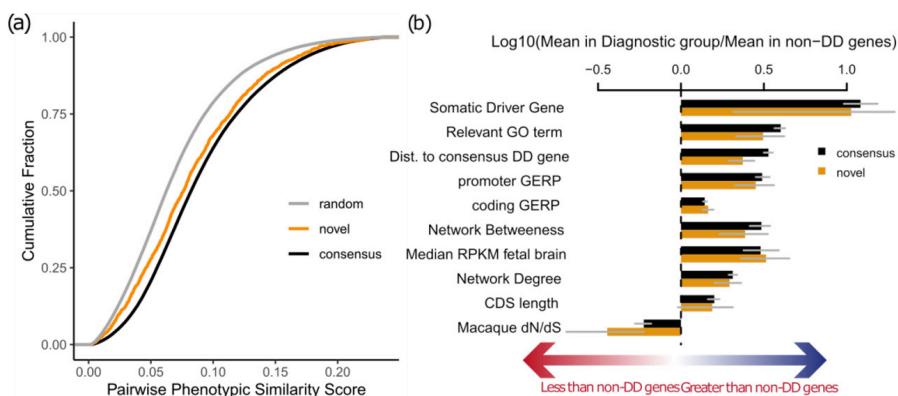
We also investigated whether some of the synonymous DNMs might be pathogenic by disrupting splicing. We identified a small but significant enrichment of synonymous DNMs with high values of the splicing pathogenicity score SpliceAI<sup>156</sup> ( $\geq 0.8$ , 1.56-fold enriched,  $P = 0.0037$ , Poisson test; **Supplementary Table 4**). This enrichment corresponds to an excess of around 15 splice-disrupting synonymous DNMs in our cohort, of which 6 are accounted for by a recurrent synonymous DNM in *KAT6B* that is known to disrupt splicing.<sup>157</sup>

Taken together, 25.0% of our cohort has a nonsynonymous DNM in one of the consensus or significant DD-associated genes (**Fig. 1d**). We noted significant sex differences in the autosomal burden of nonsynonymous DNMs (**Supplementary Fig. 8**). The rate of nonsynonymous DNMs in consensus autosomal genes was significantly higher in female individuals than male individuals ( $OR = 1.16$ ,  $P = 4.4 \times 10^{-7}$ , Fisher's exact test; **Fig. 1e**), as noted previously.<sup>114</sup> However, the exome-wide burden of autosomal nonsynonymous DNMs in all genes was not significantly different between undiagnosed male and female participants ( $OR = 1.03$ ,  $P = 0.29$ , Fisher's exact test). This indicates that there are subtle sex differences in the genetic architecture of DDs, especially with regard to known and undescribed disorders. This could include sex-biased contributions of polygenic, oligogenic and/or environmental modifiers of phenotypic variation and thus clinical ascertainment.

## Characteristics of the novel DD-associated genes

Based on semantic similarity<sup>158</sup> between human phenotype ontology terms, patients with DNM in the same novel DD-associated gene were less phenotypically similar to each other, on average, than patients with DNM in a consensus gene ( $P = 2.3 \times 10^{-11}$ , Wilcoxon rank-sum test; **Fig. 2a and Supplementary Fig. 9**). This suggests that these novel disorders less often result in distinctive and consistent clinical presentations, which may have made these disorders more difficult to discover using a phenotype-driven approach. Each of these novel disorders requires genotype–phenotype characterization, which is beyond the scope of this study.

Overall, novel DD-associated genes encode proteins that have very similar functional and evolutionary properties to consensus genes (**Fig. 2b; Supplementary Table 5**). Despite the high-level functional similarity between known and novel DD-associated genes, nonsynonymous DNM in the more recently described DD-associated genes are much more likely to be missense DNM, and less likely to be PTVs (discordant and novel;  $P = 1.2 \times 10^{-25}$ , chi-squared test). Of the 28 novel genes, 15 (54%) had only missense DNM. As a consequence, we expect that the effects of a greater proportion of the novel genes act through altered-function mechanisms



**Figure 2: Properties of the novel genes.** **A.** The phenotypic similarity of patients with DNM in novel and consensus genes. Random phenotypic similarity was calculated from random pairs of patients. Patients with DNM in the same novel gene were less phenotypically similar than patients with DNM in the same consensus gene  $P = 2.3 \times 10^{-11}$ , Wilcoxon rank-sum test. **B.** Comparison of properties of consensus ( $n = 380$ ) and novel ( $n = 28$ ) DD-associated genes known to be differential between consensus and non-DD-associated genes (95% bootstrapped confidence intervals are shown). GO, Gene Ontology; GERP, genomic evolutionary rate profiling; RPKM, reads per kilobase of transcript per million mapped reads; CDS, coding sequence; dN/dS, the ratio of substitution rate at nonsynonymous and synonymous sites.

(for example, as dominant-negative or gain-of-function disorders). For example, the novel gene PSMC5 (DeNovoWEST  $P = 2.6 \times 10^{-15}$ ) had one in-frame deletion and nine missense DNMs, eight of which altered two structurally important amino acids in the AAA+ ATPase domain; the effect of PSMC5 alterations are therefore probably generated through an altered-function mechanism (**Supplementary Fig. 10 a, b**). None of the novel genes exhibited significant clustering of de novo PTVs.

We observed that missense DNMs were more likely to affect functional protein domains than other coding regions. We observed a 2.63-fold enrichment ( $P = 2.2 \times 10^{-68}$ , G-test) in missense DNMs that reside in protein domains among consensus genes and a 1.80-fold enrichment ( $P = 8.0 \times 10^{-5}$ , G-test) in novel DD-associated genes, but no enrichment in synonymous DNMs (**Supplementary Table 6**). Four protein domain families in consensus genes were enriched in missense DNMs (**Supplementary Table 7**): ion transport protein (PF00520,  $P = 6.9 \times 10^{-4}$ , Bonferroni-corrected G-test), ligand-gated ion channel (PF00060,  $P = 4.0 \times 10^{-6}$ ), and protein kinase domain (PF00069,  $P = 0.043$ ) and kinesin motor domain (PF00225,  $P = 0.027$ ). Missense DNMs in all four enriched domain families have previously been associated with DDs (**Supplementary Table 8**).<sup>159–161</sup>

We observed a significant overlap between the 285 DNM-enriched DD-associated genes and a set of 369 previously described cancer-driving genes<sup>162</sup> (overlap of 70 genes;  $p = 1.7 \times 10^{-49}$ , logistic regression correcting for selection on heterozygous PTVs ( $s_{het}$ )), as observed previously,<sup>163,164</sup> as well as a significant enrichment in nonsynonymous DNMs in both overlapping and non-overlapping cancer genes (**Supplementary Table 9**). We observe 117 DNMs in 76 recurrent somatic mutations that were observed in at least three patients in The Cancer Genome Atlas (TCGA).<sup>165</sup> By modelling the germline mutation rate of these somatic driver mutations, we found that recurrent nonsynonymous mutations in the TCGA are enriched 21-fold in our cohort ( $p < 10^{-50}$ , Poisson test, **Supplementary Fig. 11**), whereas recurrent synonymous mutations in the TCGA are not significantly enriched (2.4-fold,  $p = 0.13$ , Poisson test). These results suggest that this observation is driven by the pleiotropic effects of these mutations in development and tumorigenesis, rather than because of hypermutability of these variants.

## Recurrent mutations

We identified 773 recurrent DNMs (736 SNVs and 37 indels), observed in 2–36 individuals, which enabled us to systematically interrogate the factors that drive recurrent germline mutations. We considered three potential contributory factors: (1) clinical ascertainment that enriches for pathogenic mutations; (2) greater mutability at specific sites; and (3) positive selection that confers a proliferative advantage in the male germline.<sup>166</sup> We observed evidence that all three factors contributed to the occurrence of recurrent germline mutations; however, these factors are not mutually exclusive. Clinical ascertainment drives the observation that 65% of recurrent DNMs were in consensus genes, a 5.4-fold enrichment compared with DNMs that were observed only once ( $p < 10^{-50}$ , proportion test). Hypermutability underpins the observation that 64% of recurrent de novo SNVs occurred at hypermutable CpG dinucleotides,<sup>167</sup> a 2.0-fold enrichment over DNMs that were observed only once ( $p = 3.3 \times 10^{-68}$ , chi-squared test).

Positive germline selection can increase the apparent mutation rate more strongly<sup>166</sup> than either clinical ascertainment (10–100X in our dataset) or hypermutability (around 10× for CpGs). However, only a minority of the most highly recurrent mutations in our dataset are in genes that have been previously associated with germline selection. Nonetheless, several lines of evidence suggested that the majority of these most highly recurrent mutations are likely to confer a germline selective advantage. On the basis of the observations above, DNMs under germline selection should be more likely to be activating missense mutations, and should be less enriched for CpG dinucleotides. **Extended Data**

**Table 1** shows the 16 de novo SNVs that were observed 9 or more times in our cohort, only 2 of which are in known germline selection genes. All but 2 of these 16 de novo SNVs cause missense changes, all but 2 of these genes cause disease by an altered-function mechanism, and these DNMs were depleted for CpGs relative to all recurrent mutations. Two of these genes with highly recurrent de novo SNVs, in *SHOC2* and *PPP1CB*, which encode interacting proteins that regulate the RAS–MAPK pathway; pathogenic variants in these genes are associated with a Noonan-like syndrome.<sup>168</sup> Moreover, two of these recurrent DNMs are in the same gene (*SMAD4*), which encodes a key component of the TGFβ signalling pathway, potentially expanding the pathophysiology of germline selection beyond the RAS–MAPK pathway. Confirming germline selection of these mutations will require deep sequencing analyses of the testes and/or sperm.<sup>169</sup>

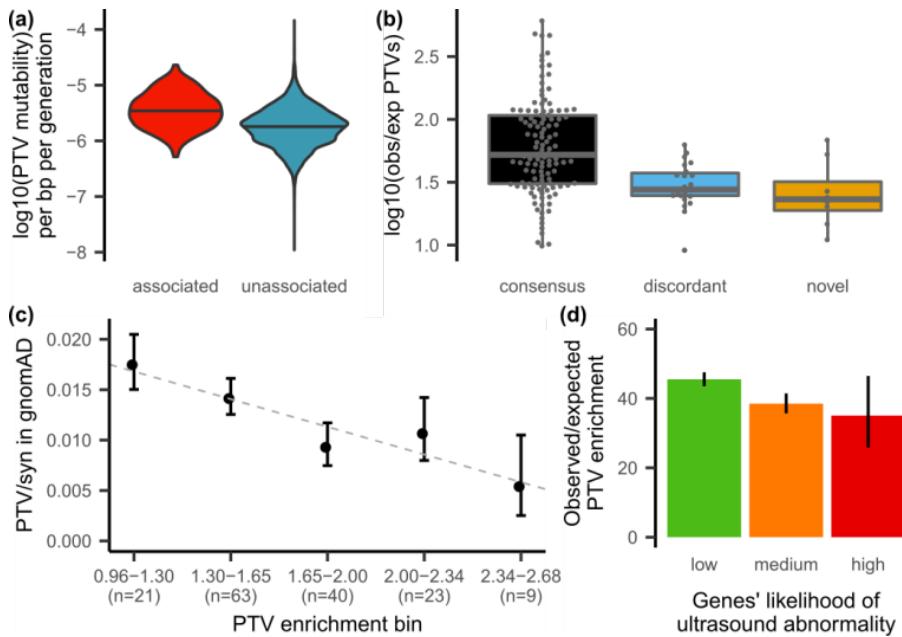
**Extended Data Table 1.** *De novo SNVs with more than nine recurrences in our cohort annotated with relevant information, such as CpG status, whether the affected gene is a known somatic driver or germline-selection gene, and diagnostic gene group (for example, consensus). 'Recur' refers to the number of recurrences. 'Likely mechanism' refers to the mechanisms attributed to this gene in the published literature.*

Symbol	Chr	Position	Ref	Alt	Consequence	Recur	Likely mechanism	CpG	Somatic Driver Gene	Germline Selection Gene	DD status
PACS1	11	65978677	C	T	missense	36	activating	Yes	-	-	consensus
PPP2R5D	6	42975003	G	A	missense	22	dominant negative	-	-	-	consensus
SMAD4	18	48604676	A	G	missense	21	activating	-	Yes	-	consensus
PACS2	14	105834449	G	A	missense	13	dominant negative	Yes	-	-	discordant
MAP2K1	15	66729181	A	G	missense	11	activating	-	Yes	Yes	consensus
PPP1CB	2	28999810	C	G	missense	11	all missense/in frame	-	-	-	consensus
NAA10	X	153197863	G	A	missense	11	all missense/in frame	Yes	-	-	consensus
MECP2	X	153296777	G	A	stop gain	11	loss of function	Yes	-	-	consensus
CSNK2A1	20	472928	T	C	missense	10	activating	-	-	-	consensus
CDK13	7	40085606	A	G	missense	10	all missense/in frame	-	-	-	consensus
SHOC2	10	112724120	A	G	missense	9	activating	-	-	-	consensus
PTPN11	12	112915523	A	G	missense	9	activating	-	Yes	Yes	consensus
SMAD4	18	48604664	C	T	missense	9	activating	Yes	Yes	-	consensus
SRCAP	16	30748664	C	T	stop gain	9	dominant negative	Yes	-	-	consensus
FOXP1	3	71021817	C	T	missense	9	loss of function	Yes	-	-	consensus
CTBP1	4	1206816	G	A	missense	9	dominant negative	Yes	-	-	discordant

## Incomplete penetrance and pre- or perinatal death

Nonsynonymous DNMs in consensus or significant DD-associated genes accounted for half of the exome-wide nonsynonymous DNM burden associated with DD (**Fig. 1b**). Despite our identification of 285 significantly DD-associated genes, there remains a substantial burden of both missense and protein-truncating DNMs in unassociated genes (those that are neither significant in our analysis nor on the consensus gene list). This residual burden of protein-truncating DNMs is greatest in genes that are intolerant to PTVs in the general population (**Supplementary Fig. 12**), which suggests that many haploinsufficient disorders have not yet been described. We observed that PTV mutability (estimated from a null germline mutation model) was significantly lower in unassociated genes compared with DD-associated genes ( $p = 4.5 \times 10^{-68}$  Wilcox rank-sum test; **Fig. 3a**), which leads to reduced statistical power to detect DNM enrichment in unassociated genes, consistent with our hypothesis that numerous haploinsufficient disorders have not yet been identified.

A key parameter in estimating statistical power to detect novel haploinsufficient disorders is the fold enrichment of de novo PTVs expected in undescribed haploinsufficient disorders. We observed that novel DD-associated haploin-



**Figure 3: Factors that influence power to detect DD-associated genes.**

**A.** PTV mutability is significantly lower ( $p = 4.6 \times 10^{-68}$ , two-sided Wilcoxon rank-sum test) in genes that are not significantly DD-associated (blue) than in DD-associated genes (red). Median is shown as a black horizontal line. bp, base pairs. **B.** Distribution of PTV enrichment in significant, likely haploinsufficient genes by category (118 consensus, 23 discordant and 8 novel genes). Lower and upper hinges correspond to first and third quartiles. Median is shown by a horizontal grey line. The upper and lower whiskers extend 1.5 $\times$  the interquartile range. **C.** Comparison of PTV enrichment in our cohort compared with the PTV to synonymous (syn) ratio in gnomAD, for genes that are significantly PTV-enriched in our cohort (without variant weighting;  $n = 156$  genes). PTV enrichment bins are calculated as  $\log_{10}(\text{enrichment})$ . The dashed line shows the regression line. Confidence intervals are the 95% intervals of the rate ratio. **d.** Overall PTV enrichment across genes grouped by the likelihood of individuals showing a structural malformation on a prenatal ultrasound (145 low, 65 medium, 6 high genes). PTV enrichment is significantly higher for genes with a low likelihood compared to other genes ( $p = 4.6 \times 10^{-5}$ , two-sided Poisson test). Poisson 95% confidence intervals are shown.

sufficient genes had significantly lower PTV enrichment compared with the consensus haploinsufficient genes ( $p = 0.005$ , Wilcoxon rank-sum test; **Fig. 3b**). Two additional factors that could lower DNM enrichment, and thus the power to detect a novel DD association, are reduced penetrance and increased pre- or perinatal death (due to spontaneous fetal loss, termination of pregnancy because of a fetal anomaly, stillbirth or early neonatal death). To evaluate incomplete penetrance, we investigated whether haploinsufficient genes with a lower enrichment of de novo PTVs in our cohort are associated with a greater prevalence of PTVs in the general population. We observed a significant negative correlation ( $p = 0.031$ ,

weighted linear regression) between PTV enrichment in our cohort and the ratio of PTV to synonymous variants in gnomAD<sup>149</sup>, which suggests that incomplete penetrance does lower de novo PTV enrichment in our cohort (**Fig. 3c**).

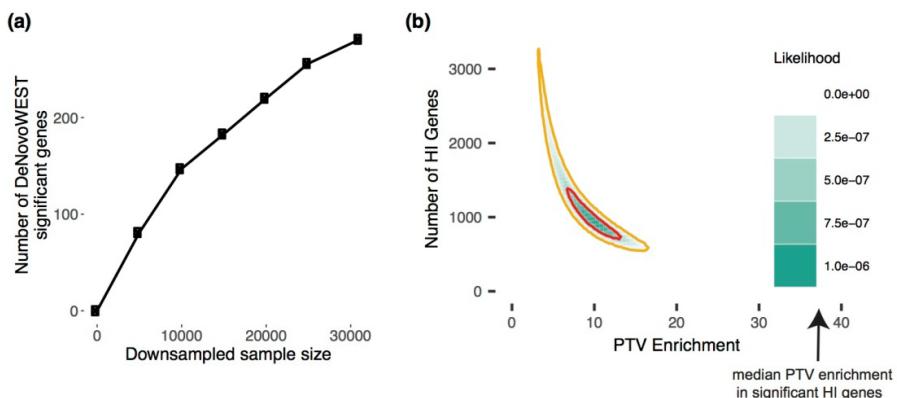
Additionally, we observed that the fold enrichment of de novo PTVs in consensus haploinsufficient DD-associated genes in our cohort was significantly higher for genes with a low likelihood of presenting with a structural malformation of the fetus during prenatal screening ( $p = 4.6 \times 10^{-5}$ , Poisson test, **Fig. 3d**), which indicates that pre- or perinatal death decreases our power to detect some of the novel disorders (see Supplementary Information for details).

## Hundreds of DD genes have not yet been discovered

Downsampling of our cohort and repeating enrichment analyses showed that the discovery of DD-associated genes has not plateaued (**Extended Data Fig. 1a**). Increasing the sample size should result in the discovery of many novel DD-associated genes. To estimate how many haploinsufficient genes have not yet been described, we modelled the likelihood of the observed distribution of de novo PTVs among genes as a function of varying numbers of undiscovered haploinsufficient DD-associated genes and fold enrichments of de novo PTVs in those genes. We found that the remaining PTV burden is most likely spread across around 1,000 genes with an approximately 10-fold PTV enrichment (**Extended Data Fig. 1b**). This fold enrichment is three times lower than in known haploinsufficient DD-associated genes, which suggests that incomplete penetrance and/or pre- or perinatal death is more prevalent among undiscovered haploinsufficient genes. We modelled the missense DNM burden separately and also observed that the most likely architecture of undiscovered DD-associated genes is one that comprises more than 1,000 genes with a substantially lower fold enrichment than in currently known DD-associated genes (**Supplemental Fig. 13**).

We calculated that a sample size of around 350,000 parent-offspring trios would be needed to have 80% power to detect a tenfold enrichment of de novo PTVs for an average gene. Using this inferred tenfold enrichment among undiscovered haploinsufficient genes, from our current data we can evaluate the likelihood that any gene  $i$  is an undiscovered haploinsufficient gene, by comparing the likelihood of the number of de novo PTVs observed in each gene to have arisen from the null mutation rate or from a tenfold increased PTV rate. Among the approximately

19,000 non-DD-associated genes, around 1,200 were more than three times more likely to have arisen from a tenfold increased PTV rate, whereas approximately 7,000 were three times more likely to have no de novo PTV enrichment.



**Extended Data Fig. 1 Exploring the remaining number of DD genes.**

*a*, Number of significant genes after downsampling the full cohort and running the enrichment test of DeNovoWEST. *b*, The likelihood of the observed distribution of de novo PTV mutations was modelled. This model varies the numbers of remaining haploinsufficient (HI) DD genes and PTV enrichment in those remaining genes. The 50% credible interval is shown in red and the 90% credible interval is shown in orange. Note that the median PTV enrichment in genes that are significant and known to operate through a loss-of-function mechanism (as indicated by an arrow) is 39.7.

## Discussion

Here we describe 28 novel developmental disorders by developing an improved statistical test for mutation enrichment and applying it to a dataset of exome sequences from 31,058 parent–offspring trios. Most of the increased power to detect novel disorders comes from the increase in sample size, rather than the improved statistical test. These 28 novel genes account for 1.0% of our cohort, and their inclusion in diagnostic workflows will help to improve diagnosis of similar patients globally. The value of this study for improving diagnostic yield extends beyond these 28 novel genes; the total number of genes added to diagnostic workflows of the three participating centres (including newly validated discordant genes) ranged from 48 to 65 genes. We show that both incomplete penetrance and pre- or perinatal death reduced our power to detect novel DDs postnatally, and hypothesize that one or both of these factors are operating more strongly among undiscovered DD-associated genes. In addition, we identify a set of highly

recurrent mutations that are strong candidates for novel germline selection mutations, which should result in a higher than expected disease incidence that increases markedly with increased paternal age.

Our study is approximately three times larger than a recent meta-analysis of DNMs from a collection of individuals with autism spectrum disorder, intellectual disability and/or a developmental disorder.<sup>170</sup> We identified around 2.3 times as many significantly DD-associated genes as this previous study when using Bonferroni-corrected exome-wide significance (285 compared with 124). In contrast to meta-analyses of published DNMs, the harmonized filtering of candidate DNMs across cohorts in this study should be more robust to cohort-specific differences in the sensitivity and specificity of detecting DNMs.

We inferred indirectly that developmental disorders with higher rates of detectable prenatal structural abnormalities had a greater likelihood of pre- or perinatal death. The potential size of this effect can be quantified from the recently published PAGE study of genetic diagnoses in a cohort of fetal structural abnormalities.<sup>171</sup> In the PAGE study, genetic diagnoses were not returned to participants during the pregnancy, and so genetic diagnostic information could not influence the incidence of pre- or perinatal death. In the PAGE study data, 69% of fetal abnormalities with a genetically diagnosable cause died perinatally or neonatally. This emphasizes the substantial effect that pre- or perinatal death can have on reducing the ability to discover novel DDs from postnatal recruitment alone, and motivates the integration of genetic data from prenatal, neonatal and postnatal studies in future studies.

To empower our mutation enrichment testing, we estimated positive predictive values that each DNM was pathogenic on the basis of their predicted protein consequence, CADD score,<sup>29</sup> selective constraint against heterozygous PTVs across the gene ( $s_{het}$ <sup>172</sup>), and, for missense variants, presence in a region under selective missense constraint.<sup>11</sup> These positive predictive values should also be informative for variant prioritization in the diagnosis of dominant developmental disorders. Further work is needed to investigate whether these positive predictive values could be informative for recessive developmental disorders, and in other types of dominant disorders. More generally, we hypothesize that empirically estimated positive predictive values based on variant enrichment in large datasets will be similarly informative in many other disease areas.

We adopted a conservative statistical approach to identifying DD-associated genes. In two previous studies using the same significance threshold, we identified 26 novel DD-associated genes.<sup>109,114</sup> All 26 are now regarded as being diagnostic, and have entered routine clinical diagnostic practice. Had we used a significance threshold with a false-discovery rate of <10% as used previously,<sup>173</sup> we would have identified 770 DD-associated genes. The false-discovery rate of individual genes depends on the significance of other genes being tested, which means that it is not appropriate for assessing the significance of individual genes, but can be useful for defining gene sets. There are 184 consensus genes that did not cross our significance threshold in this study. It is likely that many of these genes cause disorders that were underrepresented in our study due to the ease of clinical diagnosis on the basis of distinctive clinical features or targeted diagnostic testing. These ascertainment biases will not affect the representation of novel DDs in our cohort.

Our modelling suggests that there are probably more than 1,000 DD-associated genes that remain to be discovered, and that reduced penetrance and pre- or perinatal death will reduce our power to identify these genes using DNM enrichment. Identifying these genes will require both improved analytical methods and greater sample sizes. As sample sizes increase, accurate modelling of gene-specific mutation rates becomes more important. In our analyses of 31,058 trios, we observed evidence that mutation rate heterogeneity among genes can lead to overestimation of the statistical significance of mutation enrichment based on an exome-wide mutation model. We advocate the development of more granular mutation rate models, based on large-scale population variation resources, that correct for all technical and biological complexities, to ensure that larger studies are robust to mutation rate heterogeneity.

We anticipate that the variant-level weights used by DeNovoWEST will improve over time. As reference population samples, such as gnomAD,<sup>149</sup> increase in size, weights based on selective constraint metrics (for example,  $s_{het}$  or regional missense constraint) will improve. Weights could also incorporate more functional information, such as expression in disease-relevant tissues. For example, we observe that DD-associated genes are significantly more likely to be expressed in the fetal brain (**Supplementary Fig. 14**). Furthermore, new metrics based on gene co-regulation networks can predict whether genes function within a disease-relevant pathway.<sup>174</sup> As a cautionary note, including more functional information

may increase power to detect some new disorders while decreasing power for disorders with a pathophysiology that is different from known disorders. Our analyses also suggest that variant-level weights could be further improved by incorporating other variant prioritization metrics, such as upweighting variants predicted to affect splicing, variants in particular protein domains or variants that are somatic driver mutations during tumorigenesis. In developing DeNovoWEST, we explored the application of both variant-level weights and gene-level weights in separate stages of the analysis; however, subtle but pervasive correlations between gene-level metrics (for example,  $s_{het}$ ) and variant-level metrics (for example, regional missense constraint or CADD) present statistical challenges to implementation. Finally, the discovery of less penetrant disorders can be empowered by analytical methodologies that integrate both DNMs and rare inherited variants, such as TADA.<sup>175</sup> Nonetheless, using current methods focused on DNMs alone, we estimated that around 350,000 parent-child trios would need to be analysed to have around 80% power to detect haploinsufficient genes with a tenfold PTV enrichment. Discovering non-haploinsufficient disorders will need even larger sample sizes. Reaching this number of sequenced families will not be possible for an individual research study or clinical centre; it is therefore essential that genetic data generated as part of routine diagnostic practice are shared with the research community such that it can be aggregated to drive discovery of previously undescribed disorders and improve diagnostic practice.

## Supporting Information

All supplementary information can be found online with the published article at



<https://doi.org/10.1038/s41586-020-2832-5>



# Chapter 6

---

*De novo mutation hotspots in homologous protein domains point to new candidate developmental disorder genes*

**Laurens van de Wiel, Hanka Venselaar, Juliet E. Hampstead, Lisenka E. L. M. Vissers,  
Rolph Pfundt, Gerrit Vriend, Joris A. Veltman, and Christian Gilissen**

In preparation

## Abstract

Variant interpretation remains one of the major challenges in medical genetics. Previously we showed how genetic variation, when aggregated over homologous protein domains, help interpret variants of unknown significance. Here, we created the Meta-Domain HotSpot (MDHS) p-value to identify mutation hotspots in homologous domains. The MDHS p-value was used to identify hotspots of *de novo* mutations (DNMs) in a dataset of 45,221 DNMs from 31,058 patients with developmental disorders (DDs). Of these, 15,392 DNMs locate to evolutionary equivalent positions in protein domain regions across 6,910 genes. The MDHS p-value identified three missense DNM hotspots, and no hotspots for synonymous or nonsense DNMs. All missense DNM hotspots are in the ion transport protein domain family (PF00520). The 57 missense DNMs driving enrichment result from 25 genes, of which 19 were previously associated to DDs. Function altering disease-mechanisms have been described for some of the DNMs at these hotspots in literature. 3D Protein structure modelling of the 25 genes consistently confirmed the same function of the native residues at each of these hotspots. One hotspot is located at the ion channel gate and the other two at voltage-sensing positions critical for the in/activation of the ion channel. Therefore all DNMs at these hotspots are function-altering and likely pathogenic. Six genes (*CACNA1B*, *TPCN1*, *TPCN2*, *KCNH5*, *KCNG1*, and *TRPM5*) are now suggested as new candidate genes for DD based on DNMs at these hotspots. In conclusion, we show a novel approach to identify candidate disease genes based on homologous protein domain mutation hotspots.

## Acknowledgements

This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (916-14-043 to C.G. and 918-15-667 to J.A.V.), and from the Radboud Institute for Molecular Life Sciences, Radboud university medical center (R0002793 to G.V.). We thank Erin Torti, Heather Mefford, and Kyle Retterer for confirmation of patient phenotypes in personal communication.

## Introduction

*De novo* mutations (DNMs) in protein-coding genes are an established cause for developmental disorders (DDs).<sup>176</sup> An estimated 2-5% of all children are born with severe DDs in the form of congenital malformations or neurodevelopmental disorders.<sup>177,178</sup> Of these, ~42-48% are caused by a DNM in a protein-coding gene.<sup>114,148</sup> On average, any individual has about 1-2 DNMs in protein-coding regions.<sup>59</sup> Statistical models use this to identify DNM enrichment in patient cohorts that point to candidate disease-causing genes. These efforts have resulted in a growing number of genes that are now associated with DD, and has led to the publication of a growing collection of DNMs from patient cohorts with DDs.<sup>109,114,120,128,179</sup> Nevertheless, DD-association of genes has far from saturated and over 1,000 DD-associated genes are expected to await discovery.<sup>179</sup> To continue DD-association this way, larger and larger cohorts are required.

The largest cohort of 31,058 patients with DDs was recently published in a study by Kaplanis *et al.* This enabled novel DD-association for 28 genes. Remarkably, 15 of these genes were enriched by missense mutations only, suggesting that these genes may not act through a classical mechanism of haploinsufficiency. This could partly explain the difficulties in identifying novel DD genes, since the decreased mutational target would give rise to fewer patients with mutations in these genes, than would be expected if these genes were to act through haploinsufficiency. Non-haploinsufficient DD genes can be identified by mutation clustering patterns in particular gene regions.<sup>104,113</sup> However, DNMs are rare and therefore these methods require large sample sizes to be successful.

Protein domain regions are of particular interest, because ~71% of curated disease-causing missense variants in Human Gene Mutation Database (HGMD)<sup>91</sup> and ClinVar<sup>54</sup> occur in protein domains.<sup>93</sup> DD-associated missense DNMs are up to a 2.63 fold more likely to be found in these regions.<sup>179</sup> It has been shown that the evolutionary conserved architecture underlying homologous protein domains can be used to aggregate genetic variation across the human genome.<sup>60,93,180-183</sup> Disease-causing missense variants aggregated to equivalent protein domain positions are depleted of population-based variation and *vice versa*.<sup>93</sup> In addition, disease-causing missense variants on identical homologous protein domain positions, modelled in yeast, result in similar disease-phenotypic changes.<sup>181</sup>

We developed a novel methodology to perform mutation clustering of DNMs across homologous protein domains. By aggregating across homologs we increase the statistical power to identify mutation clusters. Using this method on DNMs from 31,058 patients with DDs and suggest novel disease gene candidates.

## **Materials and Methods**

### **Dataset of de novo mutations and developmental disorder diagnostic gene lists**

We obtained all 45,221 DNMs from the Kaplanis *et al* study.<sup>179</sup> These DNMs were identified in 31,058 patients with DDs from three centres. The genetic testing approach of these patients were described previously per centre: DDD,<sup>114</sup> GeneDX,<sup>184</sup> and, Radboudumc.<sup>128</sup> All individuals that underwent genetic testing provided informed consent.<sup>179</sup> Subset of these patients have been analysed and reported in previous publications.<sup>70,114,184,185</sup> We also make use of the diagnostic lists of DD-associated genes from the Kaplanis *et al*. study, namely the novel (n=28), consensus (n=380) and discordant (n=607) diagnostic gene lists.<sup>179</sup>

### **Annotation of transcript details, protein and meta-domain position annotation**

The DNMs were annotated with corresponding GENCODE<sup>63</sup> transcripts from release 19 GRCh37.p13 Basic set, protein information from UniProtKB/Swiss-Prot<sup>64</sup> Release 2016\_09, Pfam-A<sup>41</sup> v30.0 protein domains information, and meta-domain<sup>93</sup> positions using a local version of the MetaDome<sup>53</sup> web server (code available at <https://github.com/cmbi/metadome>). Meta-domains are based on multiple sequence alignments of parts of human protein-coding genes that correspond to Pfam protein domain families. The genetic variants which correspond to homologous protein domain positions receive additional annotation of the corresponding Pfam domain ID and consensus position.

### **Filtering the annotated DNMs**

The annotation process can result in multiple GENCODE gene transcripts per DNM. To ensure a single GENCODE transcript per gene we performed a filtering step by the following order of criteria:

1. Only keep variants that have the following transcript consequence: missense, synonymous, or, stop-gained
2. The transcript corresponds to a human canonical or isoform entry in Swiss-Prot
3. This transcript contains all (or most) of the *de novo* mutations for the corresponding gene
4. The transcript translates to the longest protein sequence length
5. If multiple transcripts remain for a gene, one of these is selected
6. Filter variants only to those that are in a Pfam protein domain

### Detection of variant hotspots in homologous protein domains

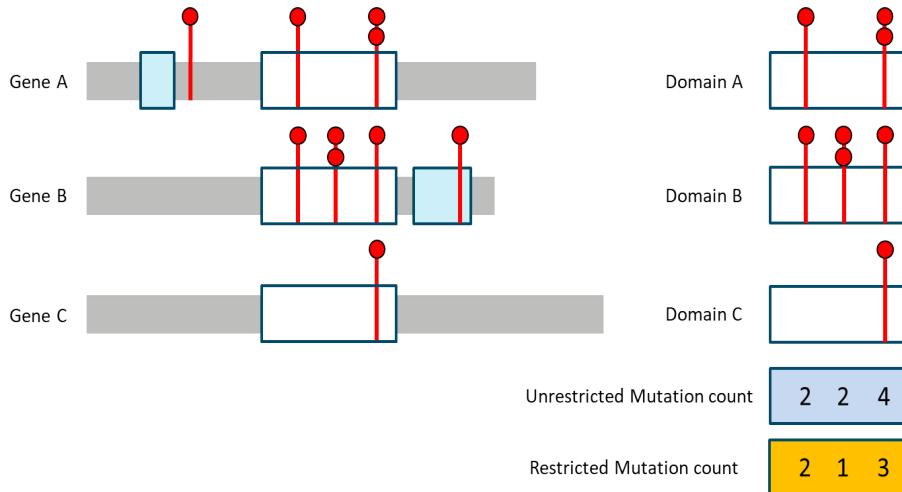
The Pfam domain ID in combination with the consensus position allows for aggregation of variants. Using these aggregated variants, we can identify which of the meta-domain positions are significantly enriched with variants. For this purpose we created the MDHS (Meta-Domain HotSpot) p-value to identify mutational hotspots in homologous protein domains defined as follows:

$$\text{MDHS p-value} = \Pr\left(x < k; \text{Bin}\left(n, \frac{1}{L}\right)\right) \quad (1.)$$

In the context of meta-domains,  $n$  corresponds to the total number of aggregated genetic variants for the Pfam domain ID,  $L$  is the total number of possible consensus positions for a Pfam domain ID,  $k$  is the total number of genetic variants aggregated at a single consensus position, and,  $x = k - 1$ , which depicts the chance of finding less than observed genetic variants at the consensus position. The MDHS p-value is adapted from the mCluster<sup>183</sup> and DS-Score<sup>186</sup>. In line with these methods, variants are assumed to follow a Binomial distribution. We correct the MDHS p-value via the Bonferroni method for the total number of Pfam protein domain IDs considered. If a Bonferroni corrected MDHS p-value  $< 0.05$  we consider it as a significant mutational hotspot.

We consider two ways of counting genetic variants to represent variable  $k$  in the MDHS p-value (**Equation 1**): an ‘unrestricted mutation count’ and a ‘restricted mutation count’ (**Figure 1**). The unrestricted mutation count would include every DNM, even when multiple DNMs occur at exactly the same chromosomal position

(i.e. recurrent DNM). The restricted count considers mutated chromosomal positions only once, thereby reducing the impact of recurrent mutations at a single position in a gene.



**Figure 1.** Graphical example of the two ways we count mutations that are aggregated over homologous protein domain regions. On the left there are three protein representations of hypothetical genes A, B and C with the mutations displayed as red lollipops, the domains as blue and white boxes. The white boxes represent domains that are homologous and are extracted including their mutations and displayed on the right part of this image as domains A, B, and C. The mutations encountered in the domains are aggregated over corresponding homologous domain positions. The aggregated mutations are displayed as ‘unrestricted mutation count’, which includes all observed mutations. The ‘restricted mutation count’ counts uniquely occurring mutation per position.

## Protein 3D structural modelling

We have created structural homology models using YASARA & WHAT IF Twinset<sup>14,147</sup> of the Ion Transport protein domain regions for each of the 25 genes in which a DNM missense was located at the identified DNM missense hotspot. The locations of each missense DNM present at one of the hotspots have been coloured purple in the YASARA scenes and the remainder of the structures are grey (**Supp. Data S1**).

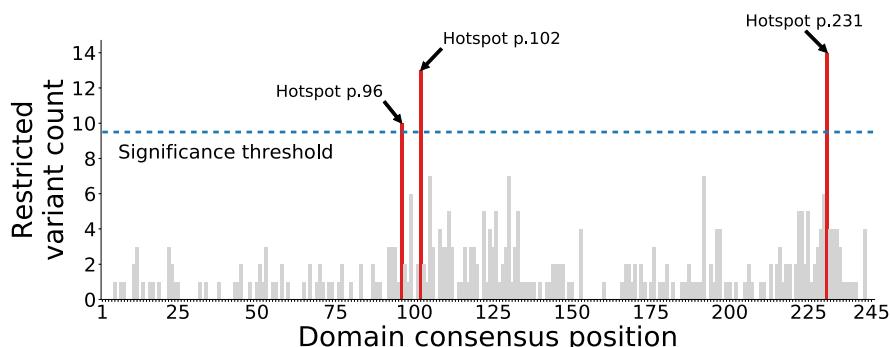
## Results

To identify hotspots of *de novo* mutations in protein domains, we count DNM s in a manner that reduces any mutational gene-bias (**Figure 1**), which then can be used to compute protein domain based positional enrichment (**Equation 1**) for each *de novo* variant type separately. We first mapped the original 45,221 DNMs resulting from 31,058 patients with developmental disorders from the Kaplanis *et al.*<sup>179</sup> study onto gene transcripts (**Methods**). After this mapping, of the original DNMs 37,089 single nucleotide variants remained of which 15,322 are located on a total of 12,389 meta-domain positions. These 15,322 DNMs resulted from protein domain regions of 6,910 protein-coding genes, and these protein domain regions consist of 2,311 protein domain families. The distribution of variant types of these 15,322 DNMs are ~73.7% missense, ~21.1% synonymous, and, ~5.3% stop-gained (**Supp. Data S2; Supp. Table 1**).

Using all 15,322 DNMs in protein domains the MDHS p-value identified 32 significant hotspots. These hotspots were enriched by 326 missense DNMs from 16 protein domain families (**Supp. Data S3**). There were no synonymous or nonsense DNMs driving significant enrichment (**Supp. Data S4 & S5**). Upon close examination, we found 9 of these hotspots to be enriched due to a large numbers of DNMs located in a single gene codon. Meaning that gene-specific DNM burdens are picked up via the MDHS method. To reduce the gene-specific DNM burden, we further filtered the 32 hotspots with the criteria that the DNMs driving their enrichment should span at least two different gene-codons. After this filtering, there remain 23 missense DNM hotspots in 12 protein domain families based on 245 DNMs from 67 genes. Nineteen of these 67 genes were not associated to DDs in the Kaplanis *et al.* study, representing a 2.53-fold enrichment of known DD-associated genes ( $p = 1.26^{-31}$  chi-square test; **Supp Table 2**). This suggests that our approach could potentially point to new candidate DD genes. However, as this analysis picked up gene-specific DNM burdens, we cannot attribute the DNMs that drive hotspot enrichment as purely domain-specific.

We repeated the hotspot discover analysis with a more restricted way of counting the DNMs to reduce gene-specific enrichment patterns being picked up (**Figure 1**). In this restricted counting analysis, the MDHS p-value identifies three significant hotspots comprised of 57 missense DNMs from 25 genes (**Supp Data S6**). Strikingly, all three hotspots are located in the Ion Transport protein domain

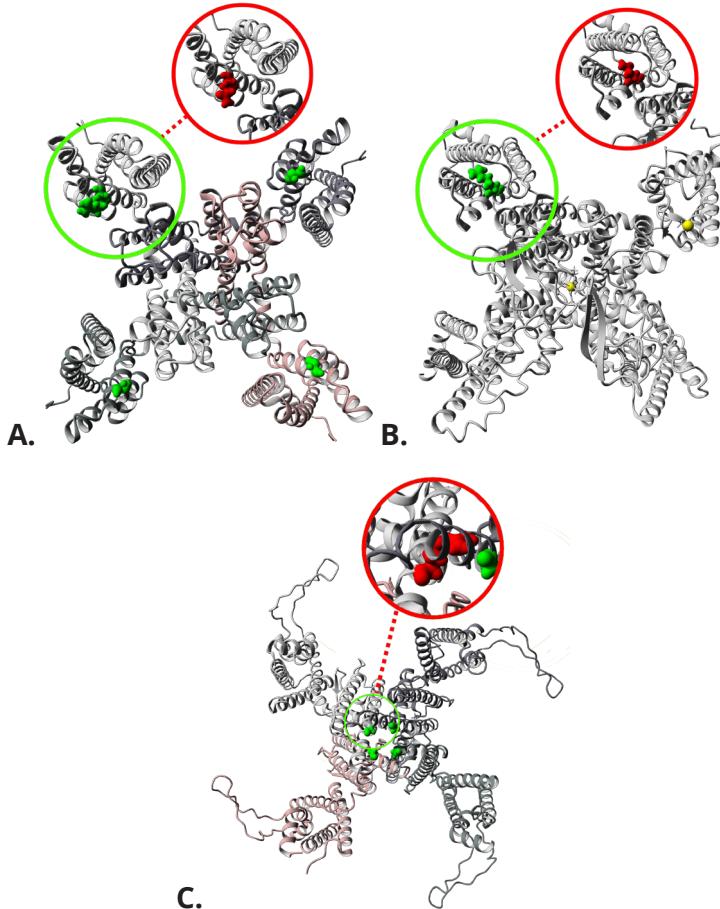
family (PF00520) (**Figure 2**). Again there are no hotspots revealed for synonymous and nonsense DNMs. The three significant hotspots are located on the domain consensus positions p.96 (10 DNMs, restricted MDHS  $p = 3.6 \times 10^{-2}$ , 16 DNMs unrestricted MDHS  $p = 1.7 \times 10^{-6}$ ), p.102 (13 DNMs, restricted MDHS  $p = 7.1 \times 10^{-5}$ , 20 DNMs, unrestricted MDHS  $p = 1.6 \times 10^{-10}$ ), and p.231 (14 DNMs, restricted MDHS  $p = 8.0 \times 10^{-6}$ , 21 DNMs, unrestricted MDHS  $p = 1.4 \times 10^{-11}$ ). The fact that all hotspots occur within the same domain family strengthens the hypothesis that these positions are likely of functional importance. The Ion Transport protein domain family is one of four protein domain families that we previously found to be significantly enriched with missense DNMs in genes that are associated to DDs.<sup>179</sup> Of the 25 genes identified with a missense DNM at a hotspot, 19 were listed as diagnostic DD-associated gene in Kaplanis *et al.* representing a 3.17-fold enrichment of known DD-associated genes ( $p = 1.78 \times 10^{-13}$  chi-square test; **Supp Table S3**).



**Figure 2.** The restricted count distribution of missense DNMs aggregated over the Ion Transport protein domain family (PF00520). The total consensus length of this domain is 245 and the sum of the restricted count distribution is 350. The significance threshold is displayed as a dotted blue line, computed via the MDHS  $p$ -value (**Equation 1**). The bars that exceeded the significance threshold are colored in red and represent the mutational hotspots p.96, p.102, and, p.231.

We created 3D protein structure homology models for each of the 25 genes (**Supp. Data S1**). Then we analysed if the missense changes were at functionally important positions in the Ion Transport protein domain in each of these 3D structures (**Supp. Data S7**). There is a large 3D protein structural overlap between for the Ion transport protein domains, they are a 3-fold less diverse in structural conformation then their observed sequences (CATH-Gene3D ID: 1.20.120.350).<sup>187</sup> Due to the structural overlap, we validated if molecular effects of missense variants

at these hotspots are likely to have similar impact on domain function across the 25 genes. Using the 25 homology models we found that hotspot p.96 (**Figure 3A**) and p.102 (**Figure 3B**) are part of the voltage-sensing helix that is important for the channel in/activation.<sup>188</sup> Hotspot p.231 (**Figure 3C**) is part of the channel gate at the end of the transmembrane helix (**Supp. Data S7**). In addition, we found that missense mutations follow a specific pattern for each of these hotspots. Of the 13/16 missense DNMs located at hotspot p.96 and 20/20 at p.102 change the positively charged wild-type residue to lose the positive charge. Losing positive charges at these locations has previously been described to trigger a function altering disease-mechanism (**Figure 3A&B**).<sup>189,190</sup> At hotspot p.231 20/21 of the missense DNMs changes the wild-type residue from a small into a larger residue. This change in residue size likely impacts the pore closure. This estimation is shared by Kortüm *et al.* as they suggest this likely causes a steric hindrance and result into a function-altering mechanism of disease (**Figure 3C**).<sup>191</sup> Furthermore, previous studies have concluded that missense DNMs in ion-channel genes are likely to result in DDs: The location and type of missense mutations in these channel genes may result in different phenotypes depending on whether they alter the function or disrupt the entire structure of the proteins.<sup>192,193</sup> Additionally, we previously found this protein domain family to be significantly enriched with missense DNMs in genes that are associated to DDs.<sup>179</sup> Considering these analyses, we argue that missense mutations at the identified hotspots are likely deleterious to the domain function.



**Figure 3.** Structural changes due to missense DNM in DD-associated genes for each hotspot.

**A.** Homology model of the KCNQ3 complex (Source PDB ID: 5VMS) with missense DNM p.R227Q marked as a green to red change. The KCNQ3 complex is a tetramer constructed from four copies of the KCNQ3 monomer. All monomers are marked in different color shades. This DNM is located at identified hotspot p.96. The wild-type Arginine residue is part of the voltage-sensing helix and changed into a Glutamine. This change causes it to lose the positive charged that was previously found to cause a function-altering mechanism of disease.<sup>189</sup>

**B.** Homology model of CACNA1A (Source PDB ID: 6JPB) with missense DNM p.R1663Q marked as a green to red change. This DNM is located at identified hotspot p.102. The wild-type Arginine residue is part of the voltage-sensing helix and changed into a Glutamine. This change causes it to lose the positive charged that was previously found to cause a function-altering mechanism of disease.<sup>190</sup>

**C.** Homology model of the KCNH1 complex (Source PDB ID: 5K7L) with missense DNM p.G496R marked as a green to red change. The KCNH1 complex is a tetramer constructed from four copies of the KCNH1 monomer. All monomers are marked in different color shades. This DNM is located at identified hotspot p.231. The wild-type Glycine residue is near the pore-closing region and changed into a much larger Arginine. This may impact pore closure and is previously reported to result into a function altering mechanism of disease.<sup>191</sup>

We focused on 6 mutations in the 6 genes that had no developmental disorder association: *TRPM5*, *TPCN2*, *TPCN1*, *KCNH5*, *KCNG1*, and *CACNA1B*. We conducted a literature review for each of the 6 genes. The *CACNA1B* gene was recently established as a DD-associated gene on the basis of nonsense DNM enrichment.<sup>194</sup> For *KCNH5* a DNM, identical to the one in our analysis, was described as a variant of unknown significance (VUS) in a patient with epileptic encephalopathy.<sup>195</sup> Protein structural analysis revealed that this variant weakens ionic interactions between other neighbouring negatively charged residues that destabilizes channel resting and activation states of the ion channel.<sup>196</sup> The variant was not from the patient included in the Kaplanis *et al.* study, meaning there are two patients with similar phenotypes and the same potentially causative variant. The patient from the Kaplanis *et al.* study is part of a cohort in an upcoming study which proposes *KCNH5* as a novel candidate gene for DD-association based on more likely pathogenic variants identified in this gene (personal communication with Heather Mefford and Erin Torti 25<sup>th</sup> of September & 6<sup>th</sup> of October 2020). Both *TPCN1* and *TPCN2* have no DD-association at the moment, however, they are both part of the mTOR complex. Genes that are part of this complex have previously been associated to DDs.<sup>197</sup> Specifically, variations in genes related to mTOR are associated to intracranial volume and intellectual disability.<sup>198</sup> In-house phenotypic data for the patient with the missense DNM in *TPCN1* (p.265R>Q) at hotspot p.96 fits this hypothesis as this patient has macrocephaly and severe ASD. To the best of our knowledge, no current literature points to a DD-association for *KCNG1* or *TRPM5*. Finally, we classified each variant according to the ACMG guidelines (**Table 1**, **Supp. Table 4**). The DNMs in *KCNH5* and *CACNA1B* are class 5 (Pathogenic) and the other DNMs as class 4 (Likely Pathogenic). A detailed description of the missense DNMs in these six candidate genes are described in **Table 1**.

Variant	Gene	Literature evidence for likely DD-association	gnomAD AF / SIFT / Polyphen-2 / MPC / CADD / MetaDome score	ACGM classification
<b>ENST00000452833.1</b> <i>c.2558G&gt;C; p.850R&gt;Q; PF00520:p.102</i>	<i>TRPM5</i> *604600	Unknown	1,20E-02 // Deleterious (0) // Probably damaging (1) // - // 28.7 // intolerant (0.49)	Likely Pathogenic (Class 4)
<b>ENST00000294309.3</b> <i>c.1734C&gt;A; p.545R&gt;S; PF00520:p.96</i>	<i>TPCN2</i> *612163	Part of the mTOR complex <sup>197</sup>	- // Deleterious (0.02) // Probably damaging (0.965) // 0.80 // 23.5 // slightly intolerant (0.67)	Likely Pathogenic (Class 4)
<b>ENST00000550785.1</b> <i>c.963G&gt;A; p.265R&gt;Q; PF00520:p.96</i>	<i>TPCN1</i> *609666	Part of the mTOR complex <sup>197</sup>	7.97e-06 // Tolerated (0.1) // Possibly damaging (0.903) // 2.35 // 26.1 // tolerant (1.03)	Likely Pathogenic (Class 4)
<b>ENST00000322893.7</b> <i>c.1249G&gt;A; p.327R&gt;H; PF00520:p.102</i>	<i>KCNH5</i> *605716	Identical VUS (p.327R>H) in unrelated patient with epileptic encephalopathy <sup>195</sup>	- // Deleterious (0) // Probably damaging (0.999) // 1.93 // 32 // intolerant (0.19)	Pathogenic (Class 5)
<b>ENST00000371571.4</b> <i>c.1332G&gt;A; p.349R&gt;H; PF00520:p.102</i>	<i>KCNG1</i> *603788	Unknown	- // Deleterious (0) // Probably damaging (1) // 2.74 // 32 // highly intolerant (0.13)	Likely Pathogenic (Class 4)
<b>ENST00000371372.1</b> <i>c.1887G&gt;A; p.581R&gt;H; PF00520:p.102</i>	<i>CACNA1B</i> *601012	Nonsense DNM <sup>s</sup> in CACNA1B lead to a neurodevelopmental disorder with seizures and nonepileptic hyperkinetic movements #618497 <sup>194</sup>	4,58E-03 // Deleterious (0) // Probably damaging (0.999) // 1.32 // 26.1 // highly intolerant (0.13)	Pathogenic (Class 5)

**Table 1.** Overview of the variants found at the hotspots that are located in genes that are not in the consensus and discordant gene lists of Kaplanis et al.<sup>179</sup> We used the Ensembl Variant Effect Predictor (VEP)<sup>199</sup> to annotate gnomAD allele frequency (AF)<sup>51</sup>, SIFT<sup>200</sup>, Polyphen-2<sup>28</sup>, MPC<sup>11</sup>, and the CADD\_Phred<sup>29</sup>. MetaDome<sup>53</sup> tolerance indication based on regional  $d_N/d_S$  was obtained manually. ACMG<sup>201</sup> classification was obtained through variant curation by a laboratory specialist.

## Discussion

Robustly predicting the pathogenicity of mutations is fundamental to improve patient diagnostics and to the advancement of our understanding of the biology underlying disease. Previously, the re-occurrence of missense mutations at identical homologous domain positions has been used to successfully implicate function and separate driver from passenger mutations in cancer.<sup>61,182,186,202-204</sup> The mCluster<sup>183</sup> scoring and the DS-Score<sup>186</sup> approaches were both developed specifically to this purpose, and, we based the MDHS p-value (**Equation 1**) on these previous methods. The MDHS p-value identified three hotspots enriched with missense DNMs in patients with DDs, and, all hotspots are located in the Ion Transport protein domain family (PF00520). In contrast to computing DNM enrichment per gene, we computed enrichment of DNMs at equivalent protein domain positions. This way we identified functionally important mutational hotspots. We have shown that the missense mutations at these hotspots disrupt domain function in the 3D protein structure. Six of the 25 genes that had missense DNMs at these hotspots had no previous diagnostic DD-association. Although this does not without a doubt confirm that these DNMs are the cause of disease for these six novel candidates, it does show that these genes are worth extra consideration for further functional DD-association studies. To that extent, of the six novel candidate genes we found, *KCNH5* and *CACNA1B* have been recently associated to DDs. *TPCN1* and *TPCN2* are likely candidates as they are part of the mTOR complex. For *KCNG1* we could not find anything in particular pointing to DD-association. We evaluated any other DNMs identified in these patients in order to exclude an existing diagnosis (**Supp. Data S8**). The patient with a missense DNMs at the hotspot in *TRPM5* also has DNMs in established DD-associated genes *SLC9A1* and *ADNP*, making *TRPM5* a less likely candidate gene for DD-association. None of the other 5 patients have DNMs in a gene with a currently known DD-association.

In line with previous finding that missense clusters indicate functional importance,<sup>104,119,205,206</sup> here we identified DNM hotspots in protein domains of likely functional importance. Using our MDHS p-value we found 32 missense DNM hotspots based on 15,322 DNMs. After filtering these 32 hotspots with the criteria that the DNMs driving enrichment should span at least two different gene-codons, 23 hotspots remained. We cannot exclude that some of these 23 hotspots may have been identified due to gene-based DNM hotspots. However, the three hotspots

that we analysed extensively were also part of these 23 hotspots, indicating that maybe more of the 23 hotspots are due to domain-based enrichment. The 245 missense DNMs that led to the identification of the 23 hotspots resulted from 67 genes. Nineteen of these 67 genes were not associated to DDs in the Kaplanis *et al.* study. In our extensive analysis we discussed six of these nineteen genes. Of these six we proposed five as likely DD-associations.

Methods that make use of protein domain architecture to aggregate variants will increase in precision with the influx of larger datasets.<sup>93</sup> Therefore, in the future more discoveries of protein domain missense DNM hotspots in DD patients is possible if cohort size increases, and, this will further drive candidate association of genes to DDs and understanding of molecular mechanisms of DNMs on protein structure and function.

## Supporting Information

All supplementary information can be found online



[https://wiel.science/publications/domain\\_dnm\\_hotspots](https://wiel.science/publications/domain_dnm_hotspots)



"Pass on what you have learned. Strength. Mastery. But weakness, folly, failure also. Yes, failure most of all. The greatest teacher, failure is."

-

Yoda in Star Wars Episode VIII – The Last Jedi (2017)  
conversation between Yoda and Luke



# **Chapter 7**

---

Discussion

## Data integration is important for DNA variant interpretation

Human DNA is complex and contains much more information than just the nucleotide sequence. For example, the DNA has a particular 3-dimensional structure which is folded around histones at specific locations and encodes for proteins, furthermore parts of these proteins can contain protein domain regions with a particular structure and function. Representing DNA as just a sequence of letters allows for both human and computer readability at the cost of losing information, this can be recovered by adding it as metadata and annotations. Integration of many layers of information is challenging, but crucial for the interpretation of genetic variation.

I have integrated genome data with protein domain sequences in so-called meta-domains, which have been mapped on 3D structures. The concept of meta-domains, that allows for transfer of information between equivalent residues in different proteins (**Chapter 2**),<sup>93</sup> has been implemented in MetaDome (**Chapter 3**).<sup>53</sup> MetaDome was instrumental in a series of developmental disorder studies that contributed to the identification of 36 novel candidate gene associations (**Chapter 4, 5, and 6**).<sup>104,179</sup> Mapping the mutations on 3-dimensional protein structures revealed the likely disease-mechanism for eight of these candidate genes (**Chapter 4 and 6**). A single variant in a meta-domain mutation hotspot helped identify six of these 36 candidate disease-genes (**Chapter 6**).

I will discuss how meta-domains have helped increase our understanding of genetic variation, and I will discuss the limitations of meta-domains and their potential future use.

## The completeness of genetic variation

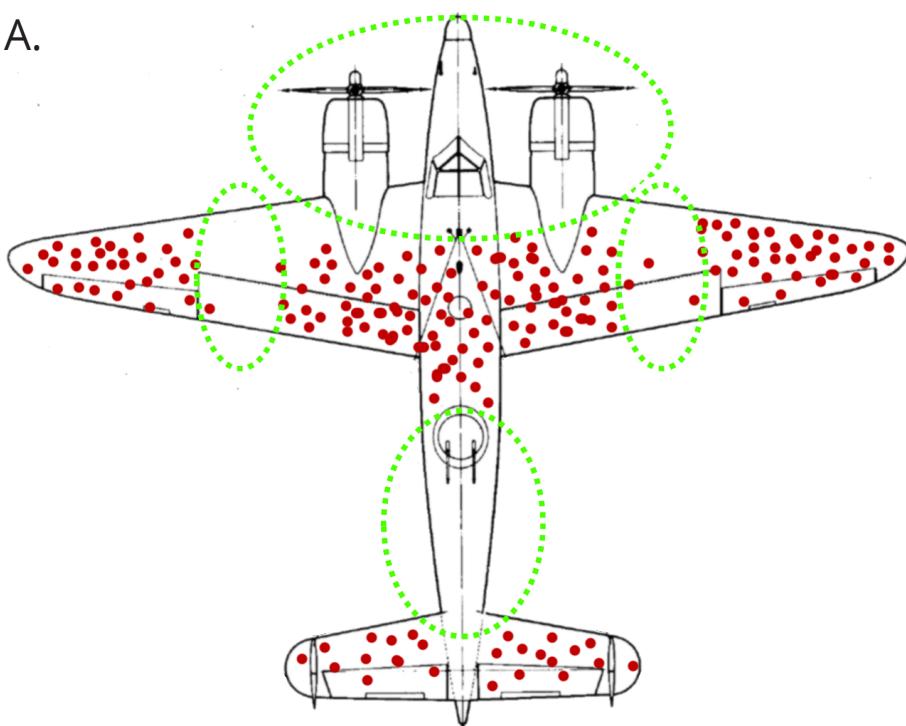
### Reaching saturation of tolerated genetic variation

In 1943, mathematician Abraham Wald calculated which parts of B-17 bomber planes needed extra armour in order to increase their survivability. He aggregated bullet hole location data from B-17s that returned from missions. He visualised this on a schematic representation of a B-17 (**Figure 1A**). One might intuitively suggest adding armour to the areas with the most damage, but Wald suggested

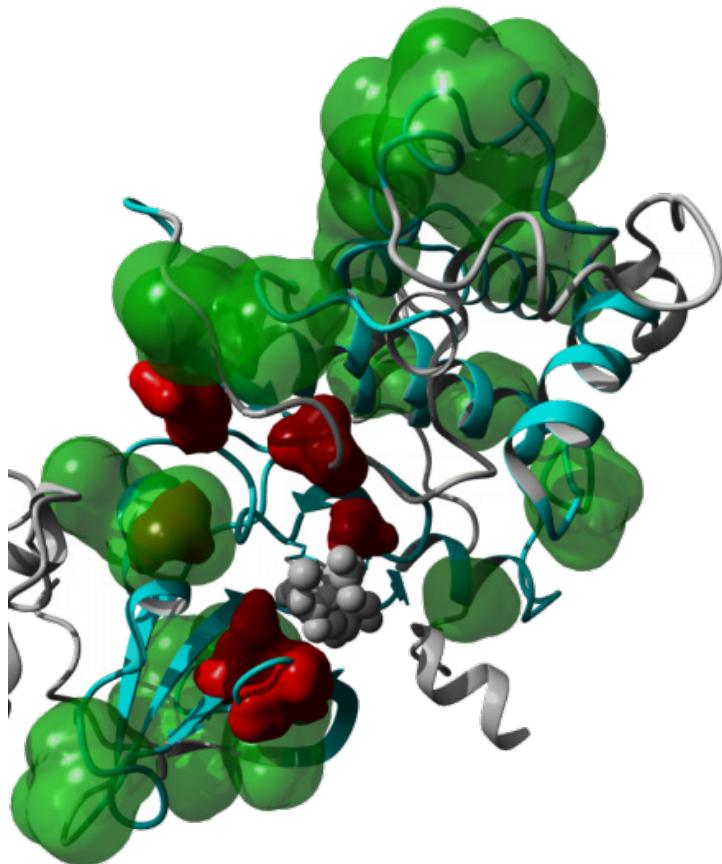
adding extra armour to the areas that are rarely damaged. This indeed increased survivability because airplanes with damage in these areas never returned.<sup>207,208</sup>

This example of ‘survivorship bias’<sup>208</sup> resembles the genetic tolerance that was discussed in the Introduction. In **Chapter 2** we showed that genetic tolerance of regions is preserved across domain homologues. We illustrated that this principle, similar to Wald’s case, can be used to predict regions that are essential, and thus predict likely deleteriousness of novel genetic variants. It has been shown that novel missense mutations observed in intolerant regions tend to be disease-causing too.<sup>11,209–211</sup>

A.



**Figure 1. A.** Example of ‘survivorship bias’ based on a schematic representation of a B-17 airplane. The red dots indicate bullet holes found on B-17s that returned after missions. The green dotted ellipses indicate areas where almost no bullet hole was found on surviving B-17s. (Airplane image courtesy of McGeddon, adapted from Wikimedia and licensed under Creative Commons CC BY-SA 4.0)

**B.**

**Figure 1. B.** Example of the protein kinase domain (PF00069) in the CDK13 protein structure. The green blobs are positions that are consistently enriched in meta-domain based on ExAC<sup>50</sup> missense variants from 353 homologues. De novo mutations known to cause a developmental disorder are displayed as dark red blobs. These red blobs are located in the region depleted of aggregated population-based variation. (Protein image was created using YASARA<sup>14</sup> modeling software based on the PDB structure 5EFQ<sup>212</sup>).

It is estimated that saturation of tolerated variation will require population studies in the hundreds of millions of individuals.<sup>51</sup> The publicly available genetic data resulting from population-based sequencing studies increased from 1,000 to 141,456 individuals in the past decade.<sup>47,50,51,58</sup> The size of population-based studies will likely continue to grow in the coming years, but it may take several decades before saturation is reached. Additionally, not all variation will be found through population sequencing studies alone. For example, only half of all protein truncating variants are expected to be found in such studies. The other half is

expected to be heterozygous lethal.<sup>51</sup> Sequencing studies of rare variation with low propagation chance may help compensate here, for example, large scale genetic studies on stillbirth or infertility.<sup>213-215</sup>

Meta-domains can help to reach saturation of all tolerated genetic variation much sooner. Meta-domains can aggregate variation found in regions that cover 41% of the human genome. For example, we can aggregate missense variants encountered in gnomAD over all 353 instances of the Protein Kinase domain (PF00069) in the human genome. Then, a pattern of tolerated missense variation emerges in the 3D protein structure (**Figure 1B**). Located inside the intolerant region in **Figure 1B** are confirmed pathogenic missense mutations which indicate the intolerant region's deleteriousness. This example is only an indication of the usefulness of data integration to reduce the need for sequencing studies.

### The importance of missense variant location and genetic data of unknown clinical significance

In contrast to Wald's situation, in genetics we sometimes do have "bullet hole information from the planes that did not survive": pathogenic mutations in patients with a genetic disorder. Distinguishing benign variation from pathogenic mutations, even in intolerant regions, remains challenging. Combining novel variants from patients with a suspected genetic disorder can help identify commonly affected regions in a gene, protein, or, over protein domain homologue. Data such as the growing collection of DNM<sup>s</sup> from patient cohorts with undiagnosed DDs<sup>59,109,114,120,123,127,128,179,216-223</sup> have proven especially helpful: sequence-based clusters of missense variants of unknown clinical significance can predict dominant vs recessive inheritance patterns,<sup>116</sup> distinguish haploinsufficiency and non-haploinsufficiency disease mechanisms (**Chapter 4**),<sup>104</sup> and identify positions where variation may trigger a function-altering mechanism of disease (**Chapter 6**). All in all, clustering of variants of uncertain significance is likely to indicate functional importance.<sup>104,119,205,206,224</sup> Discovery of the location of disease-causing missense variants uncovers a great deal of understanding behind a possible disease-mechanism, and perhaps in the future, treatments.

These methods are very powerful but require large datasets. It will therefore be crucial to keep forming large collaborations to combine large datasets of missense variants with pathogenic, benign, or unknown clinical significance.

## The limitations of meta-domains

### The need for high abundance and high quality data for meta-domains

Biological data is inherently noisy and varies in data quality due to a myriad of reasons.<sup>225</sup> A well understood computer science principal is “*Garbage in, Garbage out*” meaning that the quality of results of any method is dependent on the quality of the input data. Meta-domains are no exception to this. In fact, Meta-domains are highly dependent on the quality and amount of genetic information that is integrated. For example, whereas individuals with a childhood-onset disease were excluded from gnomAD, pathogenic variation part of recessive, polygenic, and/or late-onset genetic disorders is still present in gnomAD.<sup>51</sup> Computing tolerance based on gnomAD could therefore result in assigning regions as tolerant, whereas they are not. For this reason we removed rare variants and variants present in HGMD<sup>91</sup> or ClinVar<sup>54</sup> from our analyses in **Chapter 2**. On the other hand, a large percentage of variants are incorrectly marked as pathogenic in clinically relevant databases.<sup>85–87</sup> One of the reasons for this is that most small scale studies cannot be successfully replicated.<sup>226</sup> Luckily, these databases are gradually improved in both quantity and quality and thereby meta-domains will increase in efficacy too.

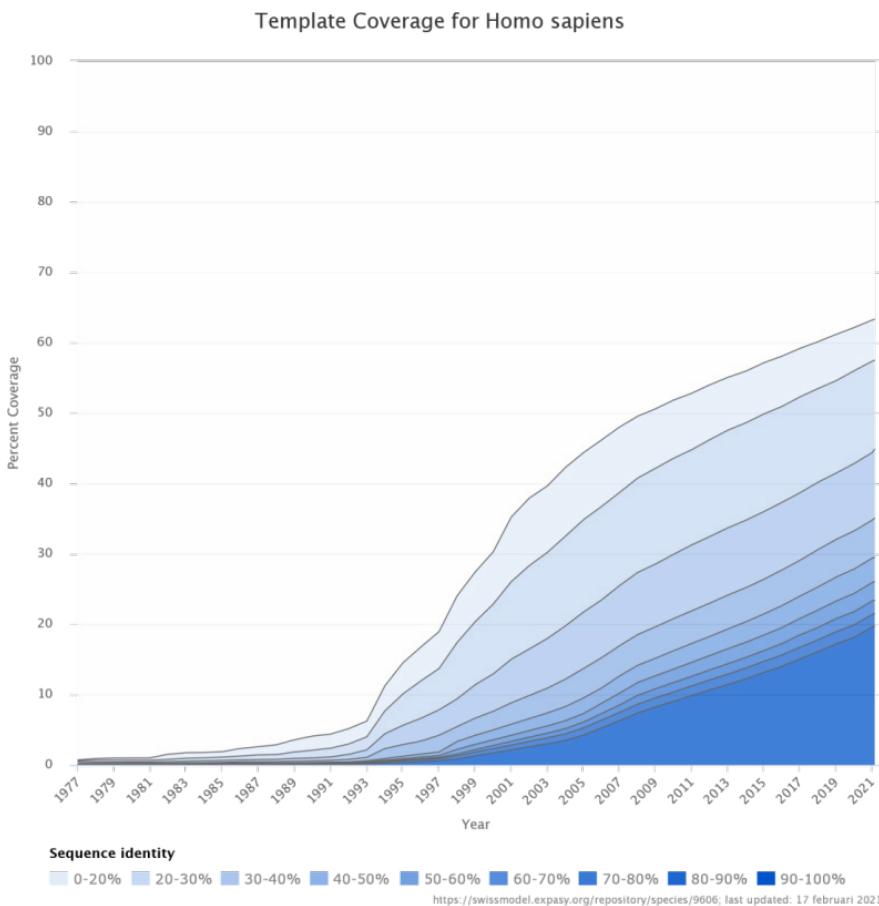
### Reasoning on variant pathogenicity with meta-domains

Protein domains cover 41% of the human genome sequence.<sup>93</sup> Meta-domains can utilise the within-human domain homologues to aggregate variant information. There are currently 3,334 Pfam-based meta-domains. These meta-domains have two, up to hundreds, of domain occurrences throughout the human genome. For example, the Protein kinase domain (PF00069) meta-domain in **Figure 1B** has 353 occurrences throughout the human genome.<sup>53</sup> The usefulness of variant aggregation in meta-domains scales with the number of homologue occurrences. Meaning that, for a low number of occurrences, the chance of encountering multiple variants from different protein domains diminishes. Additionally, the chance to observe biological signals from aggregated variation in meta-domains becomes smaller. Therefore, the efficacy of meta-domains is dependent on the number of homologue occurrences. Still, I would argue that even a single hit of a pathogenic variant at an equivalent protein position may be informative for evaluating a patient’s candidate missense variant. First of all, homology is powerful: Mutations at equivalent locations in homologous proteins result in

similar effects on protein stability.<sup>40</sup> Because protein domains are evolutionary conserved regions, pathogenic missense mutations in human protein domains have similar pathogenic effects in yeast.<sup>181</sup> Therefore, finding a novel missense variant in a meta-domain, which at an equivalent protein domain position leads to disease, provides more support for pathogenicity than when encountering identical novel variants in two patients.

### Meta-domains are constructed on sequence-based protein domain identification methods

Protein domains are at the core of the meta-domain concept and therefore the quality and completeness of protein domains determine to a large degree the meta-domain efficacy. We have built meta-domains upon Pfam<sup>41</sup> domain families, which are based on multiple sequence alignments of conserved sequences that overlap between evolutionary related species. Pfam is one of many protein domain identification methods. Typically, protein domain identification methods can be categorised into sequence-based and protein structure-based approaches.<sup>227</sup> Meta-domains can aggregate variant information over homologous protein domains, and, using a sequence-based approach has major benefits for aggregation. Firstly, sequence alignments force residues that are evolutionary conserved to be 'aligned' at identical positions. This allows for ease of aggregation from a computational perspective while retaining high certainty that these positions are identical from an evolutionary perspective. Secondly, sequence information is much more prevalent than protein structure information (**Figure 2**). Thus, Pfam-based meta-domains cover a large part of the human genome but at the cost of losing molecular structure information, however, highly related sequence homologues in general share the same protein structure.<sup>227</sup> In addition, structural information can be annotated to sequence-based protein domains. Recently, Pfam expanded annotation of PDB structures and structural models for 88% of the Pfam domain families.<sup>228</sup> This suggests that the loss of molecular structure information may soon be less of an issue for sequence-based meta-domains.

**Figure 2.**

Growth of protein structural coverage of the human genome based on sequence. This coverage does not take into account the 44% of protein sequences without a rigid structure, also called the dark proteome<sup>229</sup> (Image adapted from the SWISS-MODEL Repository and is licensed under Creative Commons CC BY-SA 4.0 – swissmodel.expasy.org).<sup>230</sup>

## The future of meta-domains

### Sequence-based protein structure prediction

So, why focus on structure-based methods for meta-domains at all? Structure is more conserved than sequence,<sup>37,44</sup> which is why structure-based domain detection is more sensitive than sequence-based detection. It can identify homologues with less than 15% sequence identity that are undetected by sequence-based

methods.<sup>227</sup> Meta-domains that are built on structure-based domain identification methods will therefore add to the overall coverage of the genome. A meta-domain-like approach using CATH<sup>187,231</sup> structure-based protein domain families was recently proposed by Ashford *et al.* (2019). This approach was suggested to be complementary to meta-domains.<sup>182</sup> However, the major limiting factor with a pure structure-based approach remains the structural coverage of sequences. As of 2008 the amount of unique structural folds have stagnated in the Protein Data Bank.<sup>227</sup> The growth of structural protein coverage of sequences so far has mostly happened due to the betterment of structure determination and 3D homology modelling methods (**Figure 2**).<sup>230</sup>

A full structural coverage of the human genome sequence will likely never be reached. In fact, 44% of the total human genome protein-coding sequence consists of (partial) natively unfolded proteins called the dark proteome<sup>229</sup> (e.g. proteins that do not conform to a rigid structure). Nevertheless, there is still much to improve (**Figure 2**). Accurate prediction of protein structure based on sequence alone has been an unsolved challenge since the 1980s. Interestingly, however, it may have been solved only very recently with AlphaFold2's participation in the annual Critical Assessment of Structure Prediction (CASP) challenge.<sup>232</sup> The preliminary results of this novel computational approach show a tremendous leap forward compared to previous years. Two-thirds of structure predictions resulting from AlphaFold2 are indistinguishable from experimentally determined protein structures. A full description of AlphaFold2 has yet to be released but it builds further on the previous AlphaFold approach.<sup>233</sup> The impact of AlphaFold2 will likely lead to a large part of protein-coding genes to have a predicted protein structure. This will have a major effect on protein domain identification methods from at least two perspectives. Firstly, sequence-based domain methods are more likely to have a structure assigned. This increases the potential to analyse the molecular effects of mutations in these protein domains. Secondly, with more (predicted) structure, structure-based domain methods might uncover protein domains for which previously the structural coverage was too limited. These sequence-based and structure-based perspectives will both be beneficial for meta-domains; the catalogue of homologous protein domains will grow and so will the potential to analyse molecular effects of mutations in protein domains.

## **Population-based variant allele frequency can complement evolutionary conservation**

Evolutionary conservation is a strong predictor of pathogenicity and heavily used in pathogenicity predictors such as Polyphen-2<sup>28</sup> and CADD<sup>29</sup>. These predictors perform well, but leave room for improvement especially within a clinical context.<sup>55-57</sup> In **Chapter 2** we show that 54% of the aggregated protein domain positions with one or more disease-causing missense variants were found to be evolutionary variable. If we assume that most disease-causing variants are correctly marked as pathogenic in the clinically relevant databases, then this could be an indication that evolutionary conservation could be complemented by population-based data. One way meta-domains can complement evolutionary conservation is by the inclusion of the population frequency of variants. For example, in most of our analyses including the display in the MetaDome web server, we only use the aggregated missense counts. Instead, the frequency of missense variants encountered in general population could be used. This way, MetaDome could represent amino acid frequency across homologous domain positions as a complement to evolutionary conservation scores.

## **The validation of the full mutational spectra in meta-domains**

To quote George Box (1978) - “*All models are wrong but some are useful*”, which also applies to meta-domains. Therefore, validation of models remains essential. In a study by Peterson *et al.* (2013) yeast was used as a model organism to validate deleterious effects of recurring pathogenic missense mutations at homologous human protein domain positions.<sup>181</sup> This study suggests that pathogenic missense mutations in protein domains have similar deleterious effects across species. Furthermore, this opens the door to use the vast amount of clinical data from the human genome to predict deleteriousness in other organisms. To achieve validation on a much larger scale in the future, we may be able to combine data from deep mutational scanning with meta-domains. Deep mutational scanning is a high-throughput method that allows for editing and analysing the effect of every single nucleotide variant change over a stretch of nucleotides in a massively parallel manner.<sup>234-237</sup> Already, data resulting from deep mutational scanning projects are empowering protein structure determination, co-variation and variant effect prediction.<sup>238-240</sup> However; deep mutational scanning has two major limitations. Firstly, it is only applicable on small regions of nucleotides. Secondly, to determine

mutational effects there needs to be a clear functional readout. To date, the protein coding region of *BRCA1* is the largest stretch of nucleotides that has been tested with deep mutational scanning.<sup>241</sup> In this study by Findlay *et al.* (2018), each possible variant in *BRCA1* was tested for its functional effect on homology-directed DNA repair, a mechanism that is necessary for tumour suppression. Meta-domains are a perfect candidate for deep mutational scanning projects. First of all, protein domains cover small parts of a protein. Secondly, a large proportion of protein domains have a specific function that may be very suitable for a functional readout. Thirdly, performing deep mutational scanning on a single protein domain has implications for all homologous occurrences. If we reason from the findings of the yeast study of Peterson *et. al.* (2013) these implications may be cross-species.<sup>181</sup> I therefore believe that deep mutational scanning is the next step to filter out domain-specific from protein-specific mutational effects.

### The future of meta-domains without reinventing “de Wiel”

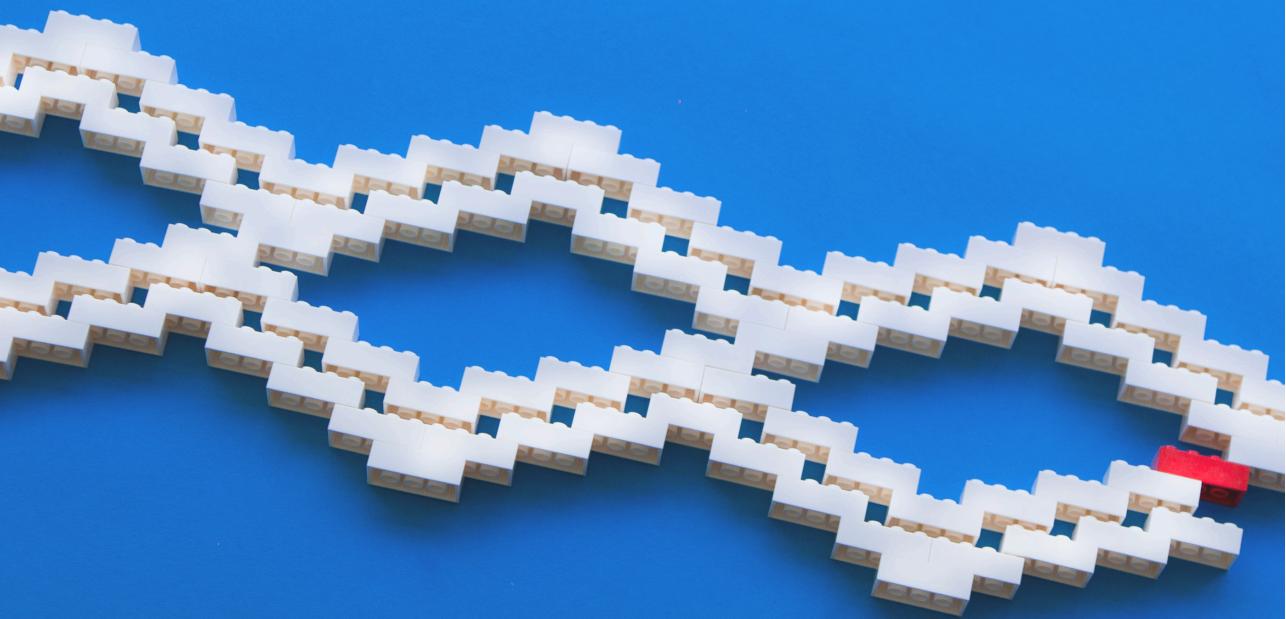
In this thesis we have discussed how to evaluate variant pathogenicity using the meta-domain concept. We have shown how meta-domains can lead to a deeper understanding of disease-mechanisms by capturing signals from ‘noisy’ clinical data of unknown significance. Previous studies in cancer genetics have utilised reasoning and concepts that bear some resemblance to meta-domains.<sup>61,180–183,203,204,242–244</sup> These studies, to me, are an indication that with the growth of genomic information, methods that were originally intended for different purposes may become more and more relevant.

Reuse of previously proved concepts also underlies the design of MetaDome. The goal for constructing the MetaDome web server was to make the abstract concept of meta-domains available in a user-friendly manner. My computer science background allowed me to set up MetaDome in accordance to up-to-date standard practice. Firstly, the code is completely documented and publicly available as open source on a code repository. Secondly, MetaDome is ‘containerized’, meaning that anyone willing can run MetaDome on their own personal computer set-up identically to the actual server. The containerization will ensure the identical environment, regardless of the host machine’s operating system or software versions. Containerization is especially helpful in scientific projects. For example, it ensures that referees are able to run the identical software during peer review.

Furthermore, containerization ensures indefinitely operational software. This last part often proves to be a problem in scientific labs where it is common that developers of the software migrate to different labs together with their expertise. This expertise is especially missed when hardware or software dependencies are upgraded that may result in non-functional software. This is not a problem for containerized software. Lastly, incorporating usage tracking in a web server helps in understanding how much it is used more than citation count would. I therefore believe that usage statistics should be part of grant applications that are specific to the continuation of scientific software projects. MetaDome has now (March 2021) been used by ~5,100 individuals from 80 countries since the initial release (November 2018). MetaDome is still steadily growing in monthly users with 460 users in the last month. The growth of users for MetaDome is a testament to its success, and, for the need of providing easily-accessible and user-friendly ways to handle the increasingly complex concepts that arise from the growth of genomic data. I believe these are lessons long learned in computer science at that computational biologists do not need reinvent the wheel.

In this discussion I have explored the potential applications of meta-domains that are outside of the scope of this thesis. I have examined the limitations to the concept of meta-domains, and, how they may be resolved. Lastly, I have provided a glimpse into the potential future of meta-domains. These future perspectives may lead to more accurate prediction and better understanding of mutational effects in protein domains and a better understanding of genetic variation.





# **Appendix**

---

## Bibliography

1. Darwin, C. *On the origin of species by means of natural selection: Or the preservation of the favoured races in the struggle for life. On the origin of species by means of natural selection: Or the preservation of the favoured races in the struggle for life.* (John Murray, 1860). doi:10.1037/14088-000
2. Kimura, M. *The neutral theory of molecular evolution.* (Cambridge University Press, 1983).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).
4. Venter, J. C. *et al.* The sequence of the human genome. *Science* 291, 1304–51 (2001).
5. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* 422, 835–847 (2003).
6. Prawira, A., Pugh, T. J., Stockley, T. L. & Siu, L. L. Data resources for the identification and interpretation of actionable mutations by clinicians. *Ann. Oncol.* 28, 946–957 (2017).
7. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* (2020). doi:10.1007/s00439-020-02199-3
8. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet.* 9, e1003709 (2013).
9. Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* 17, 9 (2016).
10. Ge, X. *et al.* Missense-depleted regions in population exomes implicate ras superfamily nucleotide-binding protein alteration in patients with brain malformation. *npj Genomic Med.* 1, 16036 (2016).
11. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* 148353 (2017). doi:10.1101/148353
12. Alberts, B. *et al.* *Molecular Biology of the Cell.* (W.W. Norton & Company, 2014).
13. Pau, V., Zhou, Y., Ramu, Y., Xu, Y. & Lu, Z. Crystal structure of an inactivated mutant mammalian voltage-gated K<sup>+</sup> channel. *Nat. Struct. Mol. Biol.* 24, 857–865 (2017).
14. Krieger, E., Koraimann, G. & Vriend, G. Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. *PROTEINS Struct. Funct. Bioinforma.* 47, 393–402 (2002).
15. WATSON, J. D. & CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171, 737–738 (1953).
16. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773 (2019).

17. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017).
18. Nirenberg, M. W. The Genetic Code: II. *Sci. Am.* 208, 80–95 (1963).
19. F. H. Crick. On protein synthesis. *Symp. Soc. Exp. Biol.* 12, 138–163 (1958).
20. Baker, K. E. & Parker, R. Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr. Opin. Cell Biol.* 16, 293–299 (2004).
21. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272–276 (2009).
22. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876 (2008).
23. Stavropoulos, D. J. *et al.* Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *npj Genomic Med.* 1, 15012 (2016).
24. Stark, Z. *et al.* A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet. Med.* 18, 1090–1096 (2016).
25. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* 577, 179–189 (2020).
26. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.* 20, 490–7 (2012).
27. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–4 (2003).
28. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–9 (2010).
29. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315 (2014).
30. Venselaar, H., te Beek, T. A., Kuipers, R. K., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* 11, 548 (2010).
31. Huynen, M. A. & Bork, P. Measuring genome evolution. *Proc. Natl. Acad. Sci.* 95, 5849–5856 (1998).
32. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425 (1987).
33. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* 21, 108–110 (1999).
34. Llamas, B., Willerslev, E. & Orlando, L. Human evolution: a tale from ancient genomes. *Philos. Trans. R. Soc. B Biol. Sci.* 372, 20150484 (2017).
35. Luzzatto, L. SICKLE CELL ANAEMIA AND MALARIA. *Mediterr. J. Hematol. Infect. Dis.* 4, e2012065 (2012).
36. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* 11, 863–874 (2001).

Ap

37. Sander, C. & Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56–68 (1991).
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410 (1990).
39. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948 (2007).
40. Ashenberg, O., Gong, L. I. & Bloom, J. D. Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* 110, 21071–6 (2013).
41. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285 (2016).
42. Wheeler, T. J., Clements, J. & Finn, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* 15, 7 (2014).
43. Ogiso, H. *et al.* Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell* 110, 775–87 (2002).
44. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence-A study of structural response in protein cores. *Proteins Struct. Funct. Bioinforma.* 77, 499–508 (2009).
45. Wouters, M. A. *et al.* Evolution of distinct EGF domains with specific functions. *Protein Sci.* 14, 1091–1103 (2005).
46. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515 (2019).
47. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–73 (2010).
48. Yang, R.-Q. *et al.* New population-based exome data question the pathogenicity of some genetic variants previously associated with Marfan syndrome. *BMC Genet.* 15, 74 (2014).
49. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
50. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291 (2016).
51. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020).
52. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469–476 (2014).
53. Wiel, L. *et al.* MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. *Hum. Mutat.* 40, humu.23798 (2019).
54. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D862–8 (2016).
55. Walters-Sen, L. C. *et al.* Variability in pathogenicity prediction programs: impact on clinical diagnostics. *Mol. Genet. Genomic Med.* 3, 99–110 (2015).

56. Miosge, L. a. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proc. Natl. Acad. Sci.* 112, E5189–E5198 (2015).
57. Masica, D. L. & Karchin, R. Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. *PLOS Comput. Biol.* 12, e1004725 (2016).
58. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493, 216–220 (2012).
59. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511, 344–347 (2014).
60. Miller, M. L. *et al.* Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Syst.* 1, 197–209 (2015).
61. Melloni, G. E. M. *et al.* LowMACA: exploiting protein family analysis for the identification of rare driver mutations in cancer. *BMC Bioinformatics* 17, 80 (2016).
62. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421 (2009).
63. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774 (2012).
64. Boutet, E. *et al.* UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol. Biol.* 1374, 23–54 (2016).
65. Finn, R. D. *et al.* InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190–D199 (2017).
66. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* 133, 1–9 (2014).
67. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* 43, W30–W38 (2015).
68. Fang, H. dcGOR: An R Package for Analysing Ontologies and Protein Domain Annotations. *PLoS Comput. Biol.* 10, e1003929 (2014).
69. Zhu, J., He, F., Song, S., Wang, J. & Yu, J. How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 9, 172 (2008).
70. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* 19, 1194–1196 (2016).
71. MacArthur, D. G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science (80-. ).* 335, 823–828 (2012).
72. Hendrich, B. & Bickmore, W. Human diseases with underlying defects in chromatin structure and modification. *Hum. Mol. Genet.* 10, 2233–42 (2001).
73. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.* 25, R157–R165 (2016).
74. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol.* 12, 228 (2011).

75. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–11 (2001).
76. Chassaing, N. *et al.* Targeted resequencing identifies PTCH1 as a major contributor to ocular developmental anomalies and extends the SOX2 regulatory network. *Genome Res.* 26, 474–85 (2016).
77. Joutel, A. *et al.* Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature* 383, 707–710 (1996).
78. Bates, M. D., Bucuvalas, J. C., Alonso, M. H. & Ryckman, F. C. Biliary atresia: pathogenesis and treatment. *Semin. Liver Dis.* 18, 281–93 (1998).
79. Leyva-Vega, M. *et al.* Genomic alterations in biliary atresia suggest region of potential disease susceptibility in 2q37.3. *Am. J. Med. Genet. Part A* 152A, 886–895 (2010).
80. Ebarasi, L. *et al.* Defects of CRB2 Cause Steroid-Resistant Nephrotic Syndrome. *Am. J. Hum. Genet.* 96, 153–161 (2015).
81. Li, W. H., Wu, C. I. & Luo, C. C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150–74 (1985).
82. Yang & Bielawski. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503 (2000).
83. Yang, Z., Swanson, W. J. & Vacquier, V. D. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* 17, 1446–55 (2000).
84. Ge, X., Kwok, P.-Y. & Shieh, J. T. C. Prioritizing genes for X-linked diseases using population exome data. *Hum. Mol. Genet.* 24, 599–608 (2015).
85. Cassa, C. A., Tong, M. Y. & Jordan, D. M. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum. Mutat.* 34, 1216–20 (2013).
86. Abouelhoda, M., Faquih, T., El-Kalioby, M. & Alkuraya, F. S. Revisiting the morbid genome of Mendelian disorders. *Genome Biol.* 17, 235 (2016).
87. Pinard, A. *et al.* Actionable Genes, Core Databases, and Locus-Specific Databases. *Hum. Mutat.* 37, 1299–1307 (2016).
88. NHLBI GO Exome Sequencing Project (ESP). Exome Variant Server. (2011). Available at: <http://evs.gs.washington.edu/EVS/>. (Accessed: 14th May 2015)
89. Karczewski, K. J. *et al.* The ExAC browser: Displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* 45, D840–D845 (2017).
90. Amr, S. S. *et al.* Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation. *Genet. Med.* 1–9 (2016). doi:10.1038/gim.2016.134
91. Stenson, P. D. *et al.* 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, 1–13 (2017).

92. Lal, D. *et al.* Gene family information facilitates variant interpretation and identification of disease-associated genes. *bioRxiv* 159780 (2017). doi:10.1101/159780
93. Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G. & Gilissen, C. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Hum. Mutat.* 1–10 (2017). doi:10.1002/humu.23313
94. Rossum, G. Van & Drake, F. L. Python Tutorial. *History* 42, 1–122 (2010).
95. Ronacher, A. Flask. (2010).
96. Evans, E. *Domain-driven design: tackling complexity in the heart of software*. (Addison-Wesley Professional, 2004).
97. Hykes, S. Docker. (2013).
98. Thomas, J., Potekhin, O., Lauhakari, M., Shah, A. & Berning, D. *Creating Interfaces with Bulma*. (Bleeding Edge Press, 2018).
99. Bostock, M., Ogievetsky, V. & Heer, J. D<sup>3</sup>: Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* 17, 2301–9 (2011).
100. PostgreSQL Global Development Group. PostgreSQL. (1996).
101. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–3 (2009).
102. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191 (2009).
103. Traynelis, J. *et al.* Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.* 27, 1715–1729 (2017).
104. Lelieveld, S. H. *et al.* Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. *Am. J. Hum. Genet.* 8, 52–56 (2017).
105. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* 48, 1060–5 (2016).
106. Özgül, R. K. *et al.* Exome sequencing and cis-regulatory mapping identify mutations in MAK, a gene encoding a regulator of ciliary length, as a cause of retinitis pigmentosa. *Am. J. Hum. Genet.* 89, 253–264 (2011).
107. Klebe, S. *et al.* New mutations in protein kinase Cy associated with spinocerebellar ataxia type 14. *Ann. Neurol.* 58, 720–729 (2005).
108. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7 Suppl 1, S4.1–9 (2006).
109. Deciphering Developmental Disorders Study *et al.* Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 519, 223–228 (2015).
110. Srour, M. *et al.* Gain-of-Function Mutations in RARB Cause Intellectual Disability with Progressive Motor Impairment. *Hum. Mutat.* 37, 786–93 (2016).

Ap

111. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* 42, 483–5 (2010).
112. Schuurs-Hoeijmakers, J. H. M. *et al.* Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *Am. J. Hum. Genet.* 91, 1122–7 (2012).
113. Geisheker, M. R. *et al.* Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* 20, 1043–1051 (2017).
114. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 542, 433–438 (2017).
115. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* 6, e1001154 (2010).
116. Turner, T. N. *et al.* Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. *Hum. Mol. Genet.* 24, 5995–6002 (2015).
117. Wilkie, A. O. M. The molecular basis of genetic dominance. *J. Med. Genet.* 31, 89–98 (1994).
118. Stehr, H. *et al.* The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol. Cancer* 10, 54 (2011).
119. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci.* 112, E5486–E5495 (2015).
120. Turner, T. N. *et al.* denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* 45, D804–D811 (2017).
121. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLoS Genet.* 12, e1006315 (2016).
122. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* 46, 818–825 (2014).
123. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221 (2014).
124. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* 47, 582–8 (2015).
125. Turner, T. N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* 98, 58–74 (2016).
126. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* 43, 712–714 (2011).
127. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674–1682 (2012).
128. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* 367, 1921–9 (2012).

129. Halvardson, J. *et al.* Mutations in HECW2 are associated with intellectual disability and epilepsy. *J. Med. Genet.* 53, 697–704 (2016).
130. Choi, K.-Y., Yoo, M. & Han, J.-H. Toward understanding the role of the neuron-specific BAF chromatin remodeling complex in memory formation. *Exp. Mol. Med.* 47, e155–e155 (2015).
131. Kuroda, Y., Oma, Y., Nishimori, K., Ohta, T. & Harata, M. Brain-specific expression of the nuclear actin-related protein ArpNalpha and its involvement in mammalian SWI/SNF chromatin remodeling complex. *Biochem. Biophys. Res. Commun.* 299, 328–34 (2002).
132. Ramírez, O. A. *et al.* Dendritic assembly of heteromeric gamma-aminobutyric acid type B receptor subunits in hippocampal neurons. *J. Biol. Chem.* 284, 13077–85 (2009).
133. Robbins, M. J. *et al.* GABA(B2) is essential for g-protein coupling of the GABA(B) receptor heterodimer. *J. Neurosci.* 21, 8043–52 (2001).
134. Jones, K. A. *et al.* GABA(B) receptors function as a heteromeric assembly of the subunits GABA(B)R1 and GABA(B)R2. *Nature* 396, 674–9 (1998).
135. Møller, R. S. *et al.* Mutations in GABRB3: From febrile seizures to epileptic encephalopathies. *Neurology* 88, 483–492 (2017).
136. Johansen, K. *et al.* Phenotypic spectrum of GABRA1: From generalized epilepsies to severe epileptic encephalopathies. *Neurology* 87, 1140–51 (2016).
137. Acuna-Hidalgo, R. *et al.* Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *Am. J. Hum. Genet.* 97, 67–74 (2015).
138. Goriely, A. & Wilkie, A. O. M. Missing heritability: paternal age effect mutations and selfish spermatogonia. *Nat. Rev. Genet.* 11, 589–589 (2010).
139. Endelev, S. *et al.* Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* 42, 1021–6 (2010).
140. Lemke, J. R. *et al.* GRIN2B mutations in West syndrome and intellectual disability with focal epilepsy. *Ann. Neurol.* 75, 147–54 (2014).
141. Le Goff, C. *et al.* Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre syndrome. *Nat. Genet.* 44, 85–8 (2011).
142. Gallione, C. *et al.* Overlapping spectra of SMAD4 mutations in juvenile polyposis (JP) and JP-HHT syndrome. *Am. J. Med. Genet. A* 152A, 333–9 (2010).
143. Wuttke, T. V *et al.* Peripheral nerve hyperexcitability due to dominant-negative KCNQ2 mutations. *Neurology* 69, 2045–53 (2007).
144. Lee, C. & Levitt, M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* 352, 448–51 (1991).
145. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353, 459–73 (2005).
146. Venselaar, H. *et al.* Status quo of annotation of human disease variants. *BMC Bioinformatics* 14, 352 (2013).
147. Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* 8, 52–56 (1990).

Ap

148. Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* (80-. ). 362, 1161–1164 (2018).
149. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210
150. Kosmicki, J. A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* 49, 504–510 (2017).
151. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* 43, 838–846 (2011).
152. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* 46, 1063–1071 (2014).
153. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26 (2011).
154. Villegas, F. *et al.* Lysosomal Signaling Licenses Embryonic Stem Cell Differentiation via Inactivation of Tfe3. *Cell Stem Cell* 24, 257–270.e8 (2019).
155. Diaz, J., Berger, S. & Leon, E. TFE3-associated neurodevelopmental disorder: A distinct recognizable syndrome. *Am. J. Med. Genet. Part A* 182, 584–590 (2020).
156. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176, 535–548.e24 (2019).
157. Yilmaz, R. *et al.* A recurrent synonymous KAT6B mutation causes Say-Barber-Biesecker/Young-Simpson syndrome by inducing aberrant splicing. *Am. J. Med. Genet. Part A* (2015). doi:10.1002/ajmg.a.37343
158. Wu, X., Pang, E., Lin, K. & Pei, Z. M. Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method. *PLoS One* (2013). doi:10.1371/journal.pone.0066745
159. Catterall, W. A., Dib-Hajj, S., Meisler, M. H. & Pietrobon, D. Inherited neuronal ion channelopathies: New windows on complex neurological diseases. *J. Neurosci.* 28, 11768–11777 (2008).
160. Lasser, M., Tiber, J. & Lowery, L. A. The Role of the Microtubule Cytoskeleton in Neurodevelopmental Disorders. *Front. Cell. Neurosci.* 12, (2018).
161. Hamilton, M. J. *et al.* Heterozygous mutations affecting the protein kinase domain of CDK13 cause a syndromic form of developmental delay and intellectual disability. *J. Med. Genet.* 55, 28–38 (2018).
162. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* (2017). doi:10.1016/j.cell.2017.09.042
163. Qi, H., Dong, C., Chung, W. K., Wang, K. & Shen, Y. Deep Genetic Connection Between Cancer and Developmental Disorders. *Hum. Mutat.* (2016). doi:10.1002/humu.23040
164. Ronan, J. L., Wu, W. & Crabtree, G. R. From neural development to cognition: unexpected roles for chromatin. *Nat. Rev. Genet.* 14, 347–359 (2013).

165. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics* (2013). doi:10.1038/ng.2764
166. Goriely, A. & Wilkie, A. O. M. Paternal age effect mutations and selfish spermatogonial selection: Causes and consequences for human disease. *American Journal of Human Genetics* (2012). doi:10.1016/j.ajhg.2011.12.017
167. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* (1980). doi:10.1038/287560a0
168. Young, L. C. *et al.* SHOC2-MRAS-PP1 complex positively regulates RAF activity and contributes to Noonan syndrome pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* (2018). doi:10.1073/pnas.1720352115
169. Maher, G. J. *et al.* Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Res.* (2018). doi:10.1101/gr.239186.118
170. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* (2019). doi:10.1038/s41588-018-0288-4
171. Lord, J. *et al.* Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet* (2019). doi:10.1016/S0140-6736(18)31940-8
172. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* 49, 806–810 (2017).
173. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* 180, 568–584.e23 (2020).
174. Deelen, P. *et al.* Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. *Nat. Commun.* (2019). doi:10.1038/s41467-019-10649-4
175. He, X. *et al.* Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. *PLoS Genet.* 9, e1003671 (2013).
176. Veltman, J. a & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* 13, 565–75 (2012).
177. Ropers, H. H. Genetics of Early Onset Cognitive Impairment. *Annu. Rev. Genomics Hum. Genet.* 11, 161–187 (2010).
178. Sheridan, E. *et al.* Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the Born in Bradford study. *Lancet* 382, 1350–1359 (2013).
179. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* (2020). doi:10.1038/s41586-020-2832-5
180. MacGowan, S. A. *et al.* Human Missense Variation is Constrained by Domain Structure and Highlights Functional and Pathogenic Residues. *bioRxiv* 127050 (2017). doi:10.1101/127050

Ap

181. Peterson, T. A., Park, D. H. & Kann, M. G. A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC Genomics* 14 Suppl 3, (2013).
182. Ashford, P., Pang, C. S. M., Moya-García, A. A., Adeyelu, T. & Orengo, C. A. A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations. *Sci. Rep.* 9, 263 (2019).
183. Yue, P. *et al.* Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Hum. Mutat.* 31, 264–271 (2010).
184. Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genet. Med.* 18, 696–704 (2016).
185. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 385, 1305–1314 (2015).
186. Peterson, T. A., Nehrt, N. L., Park, D. H. & Kann, M. G. Incorporating molecular and functional context into the analysis and prioritization of human variants associated with cancer. *J. Am. Med. Informatics Assoc.* 19, 275–283 (2012).
187. Sillitoe, I., Lewis, T. & Orengo, C. Using CATH-Gene3D to Analyze the Sequence, Structure, and Function of Proteins. *Curr. Protoc. Bioinforma.* 50, 1.28.1-1.28.21 (2015).
188. Bezanilla, F. How membrane proteins sense voltage. *Nat. Rev. Mol. Cell Biol.* 9, 323–332 (2008).
189. Sands, T. T. *et al.* Autism and developmental disability caused by KCNQ3 gain-of-function variants. *Ann. Neurol.* 86, 181–192 (2019).
190. Luo, X. *et al.* Clinically severe CACNA1A alleles affect synaptic function and neurodegeneration differentially. *PLoS Genet.* 13, e1006905 (2017).
191. Kortüm, F. *et al.* Mutations in KCNH1 and ATP6V1B2 cause Zimmermann-Laband syndrome. *Nat. Genet.* 47, 661–667 (2015).
192. Heyne, H. O. *et al.* De novo variants in neurodevelopmental disorders with epilepsy. *Nat. Genet.* 50, 1048–1053 (2018).
193. Heyne, H. O. *et al.* Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci. Transl. Med.* 12, eaay6848 (2020).
194. Gorman, K. M. *et al.* Bi-allelic Loss-of-Function CACNA1B Mutations in Progressive Epilepsy-Dyskinesia. *Am. J. Hum. Genet.* 104, 948–956 (2019).
195. Veeramah, K. R. *et al.* Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia* 54, 1270–1281 (2013).
196. Yang, Y. *et al.* Multistate structural modeling and voltage-clamp analysis of epilepsy/autism mutation Kv10.2-R327H demonstrate the role of this residue in stabilizing the channel closed state. *J. Neurosci.* 33, 16586–93 (2013).
197. Cang, C. *et al.* mTOR regulates lysosomal ATP-sensitive two-pore Na(+) channels to adapt to metabolic state. *Cell* 152, 778–790 (2013).

198. Reijnders, M. R. F. *et al.* Variation in a range of mTOR-related genes associates with intracranial volume and intellectual disability. *Nat. Commun.* 8, 1052 (2017).
199. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 1–14 (2016).
200. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081 (2009).
201. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–24 (2015).
202. Peterson, T. A., Gauran, I. I. M., Park, J., Park, D. H. & Kann, M. G. Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. *PLoS Comput. Biol.* 13, 1–24 (2017).
203. Gauthier, N. P. *et al.* MutationAligner: A resource of recurrent mutation hotspots in protein domains in cancer. *Nucleic Acids Res.* 44, D986–D991 (2016).
204. Yang, F. *et al.* Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations. *PLOS Comput. Biol.* 11, e1004147 (2015).
205. Fujimoto, A. *et al.* Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* 48, 500–509 (2016).
206. Gao, J. *et al.* 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* 9, 4 (2017).
207. Wald, A. A Method of Estimating Plane Vulnerability Based on Damage of Survivors. (Columbia University, 1943).
208. Mangel, M. & Samaniego, F. J. Abraham Wald's Work on Aircraft Survivability. *J. Am. Stat. Assoc.* 79, 259 (1984).
209. Hicks, M., Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Functional characterization of 3D protein structures informed by human genetic diversity. *Proc. Natl. Acad. Sci.* 116, 8960–8965 (2019).
210. Iqbal, S. *et al.* Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl. Acad. Sci.* 117, 28201–28211 (2020).
211. Sivley, R. M., Dou, X., Meiler, J., Bush, W. S. & Capra, J. A. Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am. J. Hum. Genet.* 102, 415–426 (2018).
212. Greifenberg, A. K. *et al.* Structural and Functional Analysis of the Cdk13/Cyclin K Complex. *Cell Rep.* 14, 320–331 (2016).
213. Wojcik, M. H. Genomic Insights into Stillbirth. *N. Engl. J. Med.* 383, 1182–1183 (2020).
214. Stanley, K. E. *et al.* Causal Genetic Variants in Stillbirth. *N. Engl. J. Med.* 383, 1107–1116 (2020).

Ap

215. Oud, M. S. *et al.* A *de novo* paradigm for male infertility. *bioRxiv* 2021.02.27.433155 (2021). doi:10.1101/2021.02.27.433155
216. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 498, 220–223 (2013).
217. Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* 501, 217–221 (2013).
218. Appenzeller, S. *et al.* De Novo Mutations in Synaptic Transmission Genes Including DNM1 Cause Epileptic Encephalopathies. *Am. J. Hum. Genet.* 95, 360–370 (2014).
219. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506, 179–184 (2014).
220. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* 74, 285–299 (2012).
221. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246–250 (2012).
222. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209–215 (2014).
223. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241 (2012).
224. Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* 14 Suppl 3, (2013).
225. Tsimring, L. S. Noise in biology. *Rep. Prog. Phys.* 77, 026601 (2014).
226. Ioannidis, J. P. a. Why most published research findings are false. *PLoS Med.* 2, 0696–0701 (2005).
227. Dawson, N., Sillitoe, I., Marsden, R. L. & Orengo, C. A. The Classification of Protein Domains. *Methods Mol. Biol.* 1525, 137–164 (2017).
228. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* 117, 1496–1503 (2020).
229. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.* 112, 15898–15903 (2015).
230. Bienert, S. *et al.* The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res.* 45, D313–D319 (2017).
231. Orengo, C. *et al.* CATH – a hierachic classification of protein domain structures. *Structure* 5, 1093–1109 (1997).
232. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* 588, 203–204 (2020).
233. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710 (2020).
234. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* 11, 801–7 (2014).
235. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* 12, 203–6, 4 p following 206 (2015).

236. Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* 200, 413–422 (2015).
237. Findlay, G. M., Boyle, E. a., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* 513, 120–3 (2014).
238. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35, 128–135 (2017).
239. Munro, D. & Singh, M. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btaa1030
240. Stiffler, M. A. *et al.* Protein Structure from Experimental Evolution. *Cell Syst.* 10, 15-24.e5 (2020).
241. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 562, 217–222 (2018).
242. Kuipers, R. K. *et al.* 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins Struct. Funct. Bioinforma.* 78, 2101–2113 (2010).
243. Peterson, T. A. *et al.* DMDM: Domain mapping of disease mutations. *Bioinformatics* 26, 2458–2459 (2010).
244. Yates, C. M. & Sternberg, M. J. E. Proteins and Domains Vary in Their Tolerance of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs). *J. Mol. Biol.* 425, 1274–1286 (2013).

Ap

## **Statement on FAIR research data management**

### **Ethical compliance, consent, and FAIR patient data**

Work in this thesis is based on the results from human studies. These studies were conducted in accordance with the principles of the Declaration of Helsinki. All families involved in these human studies gave informed consent. The human studies in this thesis were approved and conducted in accordance to the guidelines set forth by:

- The medical and ethical review board Committee on Research Involving Human Subjects Region Arnhem Nijmegen (2011/188).
- The UK Research Ethics Committee (10/H0305/83 granted by the Cambridge South Research Ethics Committee, and GEN/284/12 granted by the Republic of Ireland Research Ethics Committee).
- The Western Institutional Review Board, Puyallup, WA (WIRB 20162523).

All analysed data, resulting from these human studies, has been included in the published articles. Any additional files are available from the associated corresponding authors on request. Raw sample material and identifiable clinical information were not part of the publications. Sequence and variant-level data and phenotypic data for the DDD study data are available from the European Genome-phenome Archive (EGA) with study ID EGAS00001000775. The Radboudumc sequence and variant-level data are stored on the Radboudumc Human Genetics department server and cannot be made available through the EGA owing to the nature of consent for clinical testing. In accordance to 'Wet op de geneeskundige behandelingsovereenkomst' (WGBO), Radboudumc patient data will be kept for fifteen years after publication. To access this data, please contact Christian Gilissen with a request. Data sharing will be dependent on patient consent, diagnostic status of the patient, the type of request and the potential benefit to the patient. GeneDx data cannot be made available through the EGA owing to the nature of consent for clinical testing. GeneDx-referred patients are consented for aggregate, deidentified research and subject to US HIPAA privacy protection. As such, GeneDx is not able to share patient-level BAM or VCF data, which are potentially identifiable without a HIPAA Business Associate Agreement. Access to the deidentified aggregate data used in this analysis is available upon request to

GeneDx. Clinically interpreted variants and associated phenotypes from the DDD study are available through DECIPHER. Clinically interpreted variants from RUMC are available from the Dutch national initiative for sharing variant classifications (VKGL) as well as LOVD, where they are listed with 'VKGL-NL\_Nijmegen' as the owner. Clinically interpreted variants from GeneDx are deposited in ClinVar under accession number 26957.

## Usage of public resources

The analyses in **Chapter 2, 3, 4, and, 6** were performed on scientifically published public datasets and resources. Citations, version numbers, and, identifiers were used to link back to these resources. See the methods and material section of each article for step-by-step ways to reproduce the results. See the **Web links & resources** and **Availability and identifiability of supporting data** for further details.

## Availability and identifiability of supporting data

All supporting data of work in this thesis have been (or will be) made accessible upon publication of the corresponding articles. Each element in the supporting data has been assigned an identifier:

- Data resulting from patient material samples are provided with a 'sample identifier'. This identifier, publicly, cannot directly be linked back to identifiable information of a patient. Internally, at Radboudumc, DDD, and GeneDX, these links may be made by the appropriate clinicians. If such information is required, it may be obtained via the associated corresponding authors of the studies.
- All genes references in were marked by GENCODE and/or RefSeq IDs.
- All disease related phenotypes or genotypes had associated OMIM IDs.
- Any references to variants had a corresponding ClinVar, HGMD, ExAC, or gnomAD IDs.
- Every protein structure has an associated PDB ID. The template structure's PDB ID was noted if it was a homology modelled structure.
- Protein domains have a Pfam and/or Interpro ID.
- Meta-domains make use of Pfam IDs and Pfam consensus positions.

## Source code availability

All Source code used for **Chapter 2** was made available with release of MetaDome in **Chapter 3**. The code for SpatialClustering in **Chapter 4**, and DeNovoWest and Phenopy from **Chapter 5** are available on Github. Source code for the analyses in unpublished **Chapter 6** will be made available via GitHub upon publication. See **Code repositories** for links to the repositories and a listing of **Tools, Frameworks and Programming languages** used.

## Code repositories

DeNovoWest:	<a href="https://github.com/queenjobo/DeNovoWEST">https://github.com/queenjobo/DeNovoWEST</a>
MetaDome:	<a href="https://github.com/cmbi/metadome">https://github.com/cmbi/metadome</a>
Phenopy	<a href="https://github.com/GeneDx/phenopy">https://github.com/GeneDx/phenopy</a>
SpatialClustering:	<a href="https://github.com/laurensvdwiel/SpatialClustering">https://github.com/laurensvdwiel/SpatialClustering</a>

## Tools, Frameworks and Programming languages

Biopython:	<a href="https://biopython.org/">https://biopython.org/</a>
BULMA:	<a href="https://bulma.io/">https://bulma.io/</a>
Celery:	<a href="http://www.celeryproject.org/">http://www.celeryproject.org/</a>
Docker:	<a href="https://www.docker.com/">https://www.docker.com/</a>
D3.js:	<a href="https://d3js.org/">https://d3js.org/</a>
Flask:	<a href="https://palletsprojects.com/p/flask/">https://palletsprojects.com/p/flask/</a>
Jupyter:	<a href="https://jupyter.org/">https://jupyter.org/</a>
PostgreSQL:	<a href="https://www.postgresql.org/">https://www.postgresql.org/</a>
Python:	<a href="https://www.python.org/">https://www.python.org/</a>
RabbitMQ:	<a href="https://www.rabbitmq.com/">https://www.rabbitmq.com/</a>
Redis:	<a href="https://redis.io/">https://redis.io/</a>

## Web links & resources

CADD:	<a href="https://cadd.gs.washington.edu/">https://cadd.gs.washington.edu/</a>
ClinVar:	<a href="https://www.ncbi.nlm.nih.gov/clinvar/">https://www.ncbi.nlm.nih.gov/clinvar/</a>
CMBI PDB facilities:	<a href="http://swift.cmbi.ru.nl/gv/facilities/">http://swift.cmbi.ru.nl/gv/facilities/</a>
DECIPHER	<a href="https://decipher.sanger.ac.uk">https://decipher.sanger.ac.uk</a>
Denovo-db:	<a href="https://denovo-db.gs.washington.edu/">https://denovo-db.gs.washington.edu/</a>
DDG2P:	<a href="https://www.ebi.ac.uk/gene2phenotype/downloads">https://www.ebi.ac.uk/gene2phenotype/downloads</a>
DDD:	<a href="https://decipher.sanger.ac.uk/ddd">https://decipher.sanger.ac.uk/ddd</a>
EGA	<a href="https://www.ebi.ac.uk/ega/">https://www.ebi.ac.uk/ega/</a>
ExAC:	<a href="https://exac.broadinstitute.org/">https://exac.broadinstitute.org/</a>
GENCODE:	<a href="https://www.gencodegenes.org/">https://www.gencodegenes.org/</a>
gnomAD:	<a href="https://gnomad.broadinstitute.org/">https://gnomad.broadinstitute.org/</a>
HGMD:	<a href="http://www.hgmd.cf.ac.uk/">http://www.hgmd.cf.ac.uk/</a>
HMMER:	<a href="http://hmmer.org/">http://hmmer.org/</a>
HOPE:	<a href="http://www.cmbi.ru.nl/hope/">http://www.cmbi.ru.nl/hope/</a>
InterPro:	<a href="https://www.ebi.ac.uk/interpro/">https://www.ebi.ac.uk/interpro/</a>
Lift Over tool:	<a href="https://genome.ucsc.edu/cgi-bin/hgLiftOver">https://genome.ucsc.edu/cgi-bin/hgLiftOver</a>
LOVD	<a href="https://databases.lovd.nl/shared/variants">https://databases.lovd.nl/shared/variants</a>
MetaDome:	<a href="https://stuart.radboudumc.nl/metadome/">https://stuart.radboudumc.nl/metadome/</a>
MRS:	<a href="https://mrs.cmbi.umcn.nl/">https://mrs.cmbi.umcn.nl/</a>
NHLBI ESP:	<a href="https://evs.gs.washington.edu/">https://evs.gs.washington.edu/</a>

OMIM:	<a href="https://www.omim.org/">https://www.omim.org/</a>
Pfam:	<a href="https://pfam.xfam.org/">https://pfam.xfam.org/</a>
RCSB PDB:	<a href="http://www.rcsb.org">http://www.rcsb.org</a>
RefSeq	<a href="https://www.ncbi.nlm.nih.gov/refseq/">https://www.ncbi.nlm.nih.gov/refseq/</a>
RVIS:	<a href="http://genic-intolerance.org/">http://genic-intolerance.org/</a>
subRVIS:	<a href="http://www.subrvis.org/">http://www.subrvis.org/</a>
UniProtKB/Swiss-Prot:	<a href="https://www.uniprot.org/">https://www.uniprot.org/</a>
VKG	<a href="https://www.vkg.nl/nl/diagnostiek/vkg-dashshare-database">https://www.vkg.nl/nl/diagnostiek/vkg-dashshare-database</a>
wwPDB:	<a href="http://www.wwpdb.org">www.wwpdb.org</a>
YASARA:	<a href="http://www.yasara.org/">http://www.yasara.org/</a>

Ap



# Addendum

---

## Summary

About 2-5% of all children are born with severe developmental disorders (DDs) and about half of these cases have a genetic cause. Despite that hundreds of DD-associated genes have been identified, the genetic cause for two-third of these patients remains undiagnosed. A genetic diagnosis helps families in many ways. They can join support networks, get information about possible treatments, and they learn about the risks for having further children. DDs are often the result of *de novo* mutations (DNMs) that are thus not inherited from the parents. Every individual has about 1-2 *de novo* mutations in the protein-coding regions of the genome. This low number of DNMs makes it hard to gather enough data for proper statistical treatment. Large-scale projects have been performed over the past decade to gather data on a world-wide scale. This has allowed for the association of a series of DDs to genes; often by finding a larger number of DNMs in one gene than expected in the patient cohort.

This thesis integrates human genome data with 3D protein structures with a focus on structure domains. This integration allows for the detection of disease-causing effects of mutations and has contributed to the identification of 36 candidate disease-gene associations for DDs. These newly associated genes directly enabled diagnosis for 500 families included in the studies and many more to follow world-wide.

## The meta-domain framework and the MetaDome web server

1-2% of the human DNA codes for proteins, of these proteins 42% are formed by recurring protein domains. Domains are small parts of a protein with specific structure and function. Often very different proteins share structurally highly similar domains that are homologs.

In **Chapter 2** we introduce meta-domains as the alignment of the proteome on human Pfam domains. These meta-domains are annotated with data extracted from pathogenic and population-based variation databases, genomic locations, evolutionary conservation, etc. Meta-domains allow for transfer of information between equivalent residues in different proteins. In **Chapter 3** we describe the MetaDome web server that uses this concept to support the analysis of genetic variants of unknown clinical significance. MetaDome makes the abstract concept of meta-domains widely available, including to scientists with limited bioinformatics

expertise.

## Identification of candidate developmental disorder genes and disease mechanisms

In **Chapter 4** we used spatial clustering on publicly available DNM DD-patient data to identify missense DNM clusters in fifteen genes, three of which were not previously associated with DD. Analysis of these clusters in the protein 3D structure suggested a Non-Haploinsufficiency disease-mechanism.

In **Chapter 5** we describe how a unique international collaboration between the Radboudumc, GeneDX, and the Wellcome Sanger Institute shared healthcare data of 31,058 parent-offspring trios of patients with DDs. In the resulting article in Nature we describe a series of innovations that were only possible thanks to the large size of this dataset. 285 DD-associations to genes could be made, of which 28 were previously unknown. We estimate that for at least ~1,000 genes a DD-association is still to be discovered, indicating how much work remains. This article will serve as a reference point to understand the genomic architecture of DNMs and DDs for years to come.

In **Chapter 5** we also observed that more than two-third of all missense DNMs in DD-associated genes are found in domains, of which ion transport domains, ligand-gated ion channels, protein kinase domains, and kinesin motor domains are most enriched. In **Chapter 4** we concluded that disease associated DNMs tend to cluster in the 3D protein structure, and in **Chapter 6** we invert this reasoning and use the clustering of DNMs in 3D protein structures as an indication of their disease association. After removing the genetic redundancy from the 45,221 DNMs from **Chapter 5**, MetaDome found three missense DNM hotspots in the ion transport domain. These were found in 25 genes, 19 of which were already DD-associated. Human analyses of the 3D protein structures suggested a similar functional role for the native residue at each hotspot, suggesting that the DNMs in the six novel genes are deleterious too. **Chapter 6** thus shows that a novel way of data integration leads to an enhanced interpretation of the pathogenicity of genetic variants.

Ad

## Samenvatting

Ongeveer 2-5% van alle kinderen wordt geboren met een ernstige ontwikkelingsstoornis (OS) en ongeveer de helft van deze stoornissen heeft een genetische oorzaak. Ondanks dat er honderden OS-geassocieerde genen bekend zijn, blijft de genetische oorzaak onbekend voor tweederde van de patiënten. Een genetische diagnose helpt gezinnen op veel manieren. Ze kunnen lid worden van ondersteunende netwerken, informatie krijgen over mogelijke behandelingen en ze leren over de risico's van het krijgen van meer kinderen. OS zijn vaak het resultaat van *de novo* mutaties (DNMs) die dus niet van de ouders worden geërfd. Elk individu heeft ongeveer 1-2 *de novo* mutaties in de eiwit coderende regio's van het genoom. Dit lage aantal DNMs maakt het moeilijk om voldoende data te verzamelen voor een statistische verband legging. Er zijn het afgelopen decennium grootschalige projecten uitgevoerd om gegevens van patiënten met een OS op wereldwijde schaal te verzamelen. Door in een patiënten cohort een groter aantal DNMs in één gen te vinden dan verwacht maakt het mogelijk om dat gen met OS te associëren.

Dit proefschrift integreert menselijke genoom gegevens met 3D-eiwitstructuren met een focus op structuur domeinen. Deze integratie maakt de detectie van ziekteverwekkende effecten van mutaties mogelijk en heeft bijgedragen tot het ontdekken van 36 kandidaat-ziekte tegen associaties voor OS. Deze nieuw geassocieerde genen maakten de diagnose direct mogelijk voor 500 families die deel uit maakte van de studies en er zullen er wereldwijd nog veel meer volgen.

## Het meta-domein framework en de MetaDome webserver

1-2% van het menselijk DNA codeert voor eiwitten. Van deze eiwitten wordt 42% gevormd door eiwitdomeinen. Domeinen zijn kleine onderdelen van een eiwit met een specifieke structuur en functie. Vaak kun je over verschillende eiwitten een vergelijkbaar structureel domein terugvinden die homoloog zijn.

In **Hoofdstuk 2** introduceren we meta-domeinen als een framework die samenvoeging van vergelijkbare menselijke Pfam-domeinen mogelijk maakt. Deze meta-domeinen worden geannoteerd met: gegevens uit pathogene en algemene populatie variatie databases, locaties op het genoom, evolutionaire conservering, enz. Meta-domeinen maken overdracht van informatie tussen equivalenten

residuen in verschillende eiwitten mogelijk. In **Hoofdstuk 3** presenteren we de MetaDome webserver die dit concept gebruikt om de analyse te ondersteunen van genetische varianten met onbekende klinische significantie. MetaDome maakt het abstracte concept van meta-domeinen algemeen beschikbaar voor onder andere wetenschappers met een beperkte bioinformatica-expertise.

### **Ontdekking van nieuwe kandidaat ziektegenen voor ontwikkelingsstoornissen en de ziektemechanismen daarvan**

In **Hoofdstuk 4** hebben we ‘spatial clustering’ gebruikt op de DNMs uit publiekelijk beschikbare OS-patiëntgegevens om clusters van missense DNMs in vijftien genen te identificeren. Drie van deze genen waren niet eerder geassocieerd met OS. Analyse van deze clusters in de 3D-eiwitstructuur suggereerde een ziektemechanisme van niet-haploinsufficiëntie.

In **Hoofdstuk 5** beschrijven we hoe een unieke internationale samenwerking tussen het Radboudumc, GeneDX en het Wellcome Sanger Institute leidde tot het gezamenlijk combineren van DNMs van 31.058 patiënten met een OS. In het resulterende artikel in Nature beschrijven we een reeks innovaties die alleen mogelijk waren dankzij de grote omvang van deze dataset. Er konden 285 OS-associaties met genen gemaakt worden, waarvan er 28 voorheen nog niet waren ontdekt. We schatten dat voor minstens ~1.000 genen er nog een OS-associatie ontbreekt en dat geeft aan hoeveel werk er nog ligt. Dit artikel zal dienen als een referentiepunt om de genomische architectuur van DNMs en OS in de komende jaren beter te begrijpen.

In **Hoofdstuk 5** hebben we ook waargenomen dat meer dan tweederde van alle missense DNMs in OS-geassocieerde genen worden gevonden in eiwitdomeinen, waarvan ionentransport domeinen, ligand-geactiveerde ionkanalen, proteïne kinase domeinen en kinesine-motor domeinen het meest verrijkt zijn. In **Hoofdstuk 4** concludeerden we dat met ziekte geassocieerde DNMs de neiging hebben om te clusteren in de 3D-eiwitstructuur, en in **Hoofdstuk 6** keren we deze redenering om en gebruiken we de clustering van DNMs in 3D-eiwitstructuren als een indicatie van hun ziekteassociatie. Na het verwijderen van de genetische redundantie van de 45.221 DNMs uit **Hoofdstuk 5**, vond MetaDome drie missense DNM-hotspots in het ionentransport domein. Deze werden gevonden in 25 genen, waarvan er al 19 een OS-associatie hadden. Menselijke analyses van de 3D-eiwitstructuren

Ad

suggereerden een vergelijkbare functionele rol voor het originele residu op elke hotspot, wat vervolgens suggereert dat de DNMs in de zes nieuwe genen ook schadelijk zijn. Hiermee laten we dus in **Hoofdstuk 6** zien dat een nieuwe manier van data-integratie leidt tot een verbeterde interpretatie van de pathogeniteit van genetische varianten.

# Curriculum vitae

## About the author

Laurens van de Wiel was born on the 3<sup>rd</sup> of May 1988 in Oss, the Netherlands. He grew up in Berghem, a typical Brabant village with an agricultural and catholic history, as the 7<sup>th</sup> generation of the 'van de Wiel' family.



Laurens graduated in 2006 from the Maasland College in Oss. He developed an aptitude for computers and fell in love with biology. Laurens survived cancer at the age of 12 and again at 15. This experience shaped his future ambition to fight diseases by inventing drugs. He obtained a Bachelor degree in software engineering at the Avans Hogeschool in Den Bosch in 2010 and a Master degree in Computing Science at the Radboud University in Nijmegen in 2014. His Master thesis "*Differentiating Shigella from E. coli using hierarchical feature selection on MALDI-ToF MS data*" was conducted at TNO and at the Radboud University. His thesis was awarded "Best master thesis in computing science of 2014" and supervised by dr. Evgeni Levin, dr. Armand Paauw, and, Prof. dr. Tom Heskes.

After his Master's, Laurens became a data scientist at the start-up FLXone in Eindhoven. Here, he optimized online marketing campaigns by engineering machine learning solutions for real-time data flows. However, Laurens missed conducting research on biologically and medically relevant questions. Therefore, in 2015, he accepted a bioinformatics PhD candidacy at the Radboudumc in Nijmegen. His PhD was a joint collaboration between the department of Human Genetics and the Centre for Molecular and Biomolecular Informatics. He was supervised by Prof. dr. ir. Joris Veltman, Prof. dr. Gert Vriend, and, dr. Christian Gilissen. Laurens' PhD focused on studying the effects of genomic variation on protein structures using large scale genomic datasets. By integrating structural and evolutionary biology with genomics he identified novel gene candidate-associations to developmental disorders. These discoveries enabled diagnosis for over 500 families worldwide. His work led to multiple scientific publications.

Ad

## **List of Honors, Grants and Awards**

2015 – AIA Software Master Award 2014 – Best master thesis in computing science of 2014, Radboud University Nijmegen, the Netherlands

2015 – SNUF Individual Travel Grant – To visit Discovery Science 2015

2016 – Best Poster Prize – 29th Course of Medical Genetics, Bertinoro, Italy

2017 – RIMLS Travel Grant – To visit ESHG2017

2017 – Simonsfonds Travel Grant – To visit ASHG2017

2019 – Radboud Travel grant for outgoing PhD candidates – To visit ESHG2019

## List of Publications

**KeCo: Kernel-Based Online Co-agreement Algorithm.**

*Discovery Science, Oct 2015, 10.1007/978-3-319-24282-8\_26*

L. Wiel; T. Heskes; E. Levin

**Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts**

*Nucleic acids research, Aug 2017, 10.1093/nar/gkx704*

R. van der Lee; L. Wiel; T.J.P. van Dam; M.A. Huynen

**Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorder-associated genes**

*The American Journal of Human Genetics, Sep 2017, 10.1016/j.ajhg.2017.08.004*

S.H. Lelieveld\*; L. Wiel\*; H. Venselaar; R. Pfundt; G. Vriend; J.A. Veltman; H.G. Brunner; L.E.L.M. Vissers#; C. Gilissen#

**Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics**

*Human Mutation, Nov 2017, 10.1002/humu.23313*

L. Wiel; H. Venselaar; J.A. Veltman; G. Vriend; C. Gilissen

**Heterozygous missense variants of *LMX1A* lead to nonsyndromic hearing impairment and vestibular dysfunction**

*Human Genetics, May 2018, 10.1007/s00439-018-1880-5*

M. Wesdorp; P.A.M. de Koning Gans; M. Schraders; J. Oostrik; M.A. Huynen; H. Venselaar; A.J. Beynon; J. van Gaalen; V. Piai; N. Voermans; M.M. van Rossum; B.P. Hartel; S.H. Lelieveld; L. Wiel; B. Verbist; L.J. Rotteveel; M.F. van Dooren; P. Lichtner; H.P.M. Kunst; I. Feenstra; R.J.C. Admiraal; H.G. Yntema; L.H. Hoefsloot; R.J.E. Pennings; Hannie Kremer

***De Novo* and Inherited Pathogenic Variants in *KDM3B* Cause Intellectual Disability, Short Stature, and Facial Dysmorphism**

*The American Journal of Human Genetics, Apr 2019, 10.1016/j.ajhg.2019.02.023*

I.J. Diets; R. van der Donk; K. Baltrunaite; E. Waanders; M.R.F. Reijnders; A.J.M. Dingemans; R. Pfundt; A.T. Vulto-van Silfhout; L. Wiel; C. Gilissen; J. Thevenon; L. Perrin; A. Afenjar; C. Nava; B. Keren; S. Bartz; B. Peri; G. Beunders; N. Verbeek; K. van Gassen; I. Thiffault; M. Cadieux-Dion; L. Huerta-Saenz; M. Wagner; V. Konstantopoulou; J. Vodopiutz; M. Griese; A. Boel; B. Callewaert; H.G. Brunner; T. Kleefstra; N. Hoogerbrugge; B.B.A. de Vries; V. Hwa; A. Dauber; J.Y. Hehir-Kwa; R.P. Kuiper; M.C.J. Jongmans

Ad

**MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains**

*Human Mutation, May 2019, 10.1002/humu.23798*

**L. Wiel**; C. Baakman; D. Gilissen; J.A. Veltman; G. Vriend; C. Gilissen

**Featured on the Front Cover 10.1002/humu.23892**

**Selected as Editor's Choice Article**

**Top Accessed Article in Human Mutation (2019)**

**Top Cited Article in Human Mutation (2020)**

***De novo* variants disturbing the transactivation capacity of POU3F3 cause a characteristic neurodevelopmental disorder**

*The American Journal of Human Genetics, Aug 2019, 10.1016/j.ajhg.2019.06.007*

L. Snijders Blok; T. Kleefstra; H. Venselaar; S. Maas; H.Y. Kroes; A.M.A. Lachmeijer; K.L.I. van Gassen; H.V. Firth; S. Tomkins; S. Bodek; the DDD study; K. Önnap; M.H. Wojcik; C. Cunniff; K. Bergstrom; Z. Powis; S. Tang; D.N. Shinde; C. Au; A.D. Iglesias; K. Izumi; J. Leonard; A. Abou Tayoun; S.W. Baker; M. Tartaglia; M. Niceta; M.L. Dentici; N. Okamoto; N. Miyake; N. Matsumoto; A. Vitobello; L. Faivre; C. Philippe; C. Gilissen; **L. Wiel**; R. Pfundt; P. Deriziotis; H.G. Brunner; S.E. Fisher

***De Novo* Variants in SPOP Cause Two Clinically Distinct Neurodevelopmental Disorders**

*The American Journal of Human Genetics, Mar 2020, 10.1016/j.ajhg.2020.02.001*

M.J. Nabais Sá; G. El Tekle; A.P.M. de Brouwer; S.L. Sawyer; D. del Gaudio; M.J. Parker; K. Farah; M.H. van den Boogaard; K. van Gassen; M. Van Allen; K. Wierenga; G. Purcarin; E.R. Elias; E. Torti; T. Bernasocchi; **L. Wiel**; C. Gilissen; H. Venselaar; R. Pfundt; L.E.L.M. Vissers; J. Theurillat#; B.B.A. de Vries#

***De novo* CLTC variants are associated with a variable phenotype from mild to severe intellectual disability, microcephaly, hypoplasia of the corpus callosum, and epilepsy**

*Genetics in Medicine, Apr 2020, 10.1038/s41436-019-0703-y*

M.J. Nabais Sá; H. Venselaar; **L. Wiel**; A. Trimouille; E. Lasseaux Pharma; S. Naudion; D. Lacombe; A. Piton; C. Vincent-Delorme; C. Zweier; A. Reis; R. Trollmann; A. Ruiz; E. Gabau; A. Vetro; R. Guerrini; S. Bakhtiari1; M. Kruer; D.J. Amor; M. Cooper; E.K. Bijlsma; T.S. Barakat; M.F. van Dooren; M. van Slegtenhorst; R. Pfundt; C. Gilissen; B.B.A. de Vries; A.P.M. de Brouwer; D.A. Kooken

## **Evidence for 28 genetic disorders discovered by combining healthcare and research data**

*Nature*, Oct 2020. 10.1038/s41586-020-2832-5

J. Kaplanis\*; K.E. Samocha\*; **L. Wiel\***; Z. Zhang\*; K.J. Arva; R.Y. Eberhardt; G. Gallone; S.H. Lelieveld; H.C. Martin; J.F. McRae; P.J. Short; R.I. Torene; E. de Boer; P. Danecek; E.J. Gardner; N. Huang; J. Lord; I. Martincorena; R. Pfundt; M.R.F. Reijnders; A. Yeung; H.G. Yntema; DDD Study; L.E.L.M. Vissers; J. Juusola; C.F. Wright; H.G. Brunner; H.V. Firth; D.R. FitzPatrick; J.C. Barrett; M.E. Hurles\*; C. Gilissen\*, K. Retterer\*

## **Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction**

*The American Journal of Human Genetics*, Feb 2021, 10.1016/j.ajhg.2021.01.007

J. den Hoed\*; E. de Boer\*; N. Voisin; A.J.M. Dingemans; N. Guex; **L. Wiel**; C. Nellaker; S.M. Amudhavalli; S. Banka; F.S. Bena; B. Ben-Zeev; V. R. Bonagura; A. Bruel; T. Brunet; H.G. Brunner; H.B. Chew; J. Chrast; L. Cimbalistiené; H. Coon; E.C. Délot; F. Démurger; A. Denommé-Pichon; C. Depienne; D. Donnai; D.A. Dyment; O. Elpeleg; L. Faivre; C. Gilissen; L. Granger; B. Haber; Y. Hachiya; Y.H. Abedi; J. Hanebeck; J.Y. Hehir-Kwa; B. Horist; T. Itai; A. Jackson; R. Jewell; K.L. Jones; S. Joss; H. Kashii; M. Kato; A.A. Kattentidt-Mouravieva; F. Kok; U. Kotzaeridou; V. Krishnamurthy; V. Kučinskas; A. Kuechler; A. Lavillaureix; P. Liu; L. Manwaring; N. Matsumoto; B. Mazel; K. McWalter; V. Meiner; M.A. Mikati; S. Miyatake; T. Mizuguchi; L.H. Moey; S. Mohammed; H. Mor-Shaked; H. Mountford; R. Newbury-Ecob; S. Odent; L. Oreec; M. Osmond; T.B. Palculict; M. Parker; A.K. Petersen; R. Pfundt; E. Preikšaitiené; K. Radtke; E. Ranza; J.A. Rosenfeld; T. Santiago-Sim; C. Schwager; M. Sinnema; L. Snijders Blok; R.C. Spillmann; A.P.A. Stegmann; I. Thiffault; L. Tran; A. Vaknin-Dembinsky; J.H. Vedovato-dos-Santos; S.A. Schrier Vergano; E. Vilain; A. Vitobello; M. Wagner; A. Waheed; M. Willing; B. Zuccarelli; U. Kini; D.F. Newbury; T. Kleefstra; A. Reymond\*; S.E. Fisher\*; L.E.L.M. Vissers\*

## **De novo mutation hotspots in homologous protein domains point to new candidate developmental disorder genes**

*In preparation*

**L. Wiel**; H. Venselaar; L.E.L.M. Vissers; R. Pfundt; G. Vriend; J.A. Veltman; C. Gilissen

\*: These authors contributed equally, #: These authors jointly supervised

Ad

## RIMLS Portfolio

<b>PhD student:</b>	L.J.M. van de Wiel	<b>PhD period:</b> 20-07-2015 – 30-07-2021
<b>Department:</b>	Human genetics & Centre for Molecular and Biomolecular Informatics	<b>Promotor:</b> Prof. dr. ir. J.A. Veltman, Prof. dr. G. Vriend
<b>Graduate school:</b>	Radboud Institute for Molecular Life Sciences	<b>Co-promotor:</b> dr. C.F.H.A. Gilissen

TRAINING ACTIVITIES	Year(s)	ECTS
<b>a) Courses &amp; Workshops</b>		
- Radboudumc Introductory course	2015	0.25
- RIMLS Introductory course “in the lead”	2015	0.75
- FNWI Radboud University BSc: Bioinformatics A	2015	3.0
- FNWI Radboud University MSc: Human Genetics	2015	3.0
- FNWI Radboud University BSc: Structure, Function & Bioinformatics	2016	6.0
- ESHG 29th Course in Medical Genetics (Bertinoro, Italy)	2016	1.75
- RIMLS Scientific Integrity Course	2016	1.0
- Radboud in’to Languages – Cambridge English: C2 Proficiency	2017-2018	4.5
- 2nd International Summer School on Deep Learning (Genova, Italy)	2018	1.75
- Radboudumc workshop on Personal Grants	2018	0.1
- Radboud PhD course on Diversity and Inclusion	2020	1.0
- Human Genetics workshop on work pressure	2020	0.2
<b>b) Seminars &amp; Lectures</b>		
- RIMLS Technical Forum – Bioinformatics for PhD students	2015	0.1
- Radboud Grand Round – Personalized Medicine	2016	0.1
- RIMLS Seminars (Laura Zahn, Peter-Bram ‘t Hoen)	2016-2017	0.2
- Max Planck Seminar (Matthew Hurles)	2017	0.1
- 1 <sup>st</sup> & 2 <sup>nd</sup> Deep Learn Meetup Nijmegen	2018	0.2
- AIVD Security seminar	2018	0.1
- CMBI Seminar John van Dam	2019	0.1
<b>c) National Symposia &amp; Congresses</b>		
- CMBI Autumn conference 2015 (Ravenstein)	2015	0.25
- RIMLS PhD retreat 2016 (Veldhoven) [＊]	2016	0.75
- BioSB 2016 (Lunteren) [#]	2016	0.75
- Genetics Retreat 2016 (Rolduc) [#]	2016	0.75
- CMBI Spring conference 2016 (Berg en Dal) [＊]	2016	0.5
- CMBI Autumn conference 2016 (Berg en Dal)	2016	0.25
- BioSB 2017 (Lunteren) [#]	2017	0.75
- CMBI Spring Conference 2018 (Berg en Dal) [#]	2018	0.5
- NVHG 2018 (Arnhem) [#]	2018	0.75
- Human Genetics scientific infrastructure day (Nijmegen) [#]	2019	0.5

TRAINING ACTIVITIES (continued)	Year(s)	ECTS
<b>c) International Symposia &amp; Congresses</b>		
- Discovery Science 2015 (Banff, Canada) [#]	2015	1.0
- ESHG 29th Course in Medical Genetics (Bertinoro, Italy) [*, #]	2016	0.5
- ESHG 2017 (Copenhagen, Denmark) [*]	2017	1.25
- ASHG 2017 (Orlando, FL, USA) [#]	2017	1.5
- ESHG 2019 (Gothenburg, Sweden) [#]	2019	1.25
- ESHG 2020 2.0 (Online in my Livingroom) [#]	2020	1.25

**d) Other**

- Attended and presented in Machine Learning journal club 2016-2017 2.0
- Human Genetics Theme discussion [#], [#] 2017-2018 2.0
- Human Genetics Literature discussion [#] 2016 1.0
- Human Genetics GDG group meeting [\*,\*,\*,#] 2016-2019 2.5
- CMBI Comics meeting presentations [#,#,#,#] 2018-2020 4.0
- Attended PhD Thesis defense of 2016-2021 2.0  
Rocio Acuna-Hidalgo, Peer Arts, Galuh Astuti, Margo Dona, Tom Ederveen, Daniel Garza, Jakob Goldmann, Mubeen Khan, Ideke Lamers, Robin van der Lee, Stefan Lelieveld, Lisette Meerstein-Kessel, Machteld Oud, Helmi Pett, Margot Reijnders, Carolien Ruesen, Roos Schellevis, Sanne Schoenmakers, Ralph Slijkerman, Wouter Touw

TEACHING ACTIVITIES	Year(s)	ECTS
<b>a) Supervision of internship students</b>		
- Supervising UT Honours student (Lucca Derkx)	2015-2017	1.2
- Supervising RU BMW student (Wouter-Michiel Vierdag)	2016-2017	1.0
- Supervising RU MLW student (Daniel Rademaker)	2016-2017	0.5
- Supervising HAN Bioinformatics student (Daan Gilissen)	2017-2018	1.0
- Supervising HAN Bioinformatics student (Brecht van de Berg)	2018-2019	1.3

**b) Teaching & organization**

- Teacher on the UCSC Genome Browser 2016 0.7
- Organiser & Chair of Machine Learning Journal Club 2016-2017 2.0
- Co-organiser of the Human Genetics department day-out 2017-2018 2.0
- Teacher for Genomics for Health and Environment (NWI-BB086) 2018 2.4
- Co-organiser of the RTC Bioinformatics meeting 2018-2019 0.5
- Co-organiser of the Human Genetics scientific infrastructure day 2018-2019 2.0
- Co-organiser of Human Genetics Scuba diving event 2019 0.5
- Referee of scientific articles 2017-2021 0.5

<b>TOTAL</b>	<b>65.8</b>
--------------	-------------

Oral and poster presentations are indicated with a \* and # after the name of the activity, respectively.

Ad

## Acknowledgements

This thesis would not have existed without my supervisors: **Christian Gilissen**, **Joris Veltman**, and, **Gert Vriend**. **Christian**, your similar early-career track inspired me from the start. I liked how didactic you were in your supervision. You taught me how and when to properly use statistics, the result being that I will attempt to quantify every single result for the rest of my career. You also knew so perfectly well how to separate informality from being a supervisor. This led me to discover quite a bit later on what an amazing, kind and fun person you are in addition to being an excellent supervisor. A memory I will hold dear is when you joined me and the scuba group for a pizza in Flores Indonesia while you were just passing through on your holiday. **Joris**, you always gave me a lot of freedom to conduct my research, but steered me back into the right direction when I seemed to steer too far off course. When you left to Newcastle I had my fears that it would have a negative impact on my PhD. It turned out that instead you had more time for your PhD students, especially when you brought a monthly visit to the Radboudumc. I like that these visits always seemed to align with the department's end-of-the-month 'borrel'. I grew to know you as a very social person besides your supervisor role. **Gert**, you live by the Dutch proverb "*ijzer moet je smeden als het heet is*". Besides the actual meaning of that proverb, I like the analogy to my PhD that I am the 'iron' and that many times you indeed got me heated up, but you have also formed me. You are morally headstrong and have a deep dislike for unrighteousness that I idolize. I think you have the most uncompromising scientific integrity compass of anyone I have ever met in my life. You never seemed to run out of anecdotes and I have caught myself quoting you every now and then to colleagues and friends.

I would like to thank the Doctoral Thesis committee, consisting of **prof. dr. D.J. Lefever**, **prof. dr. J. Heringa**, and **prof. dr. L.H. Franke**, for their careful reading and evaluation of this thesis.

Studies in this thesis were possible by voluntarily donated materials from **patients and their families**. This thesis would not be possible without this generosity. I thank **everyone who participated in these studies** and I hope the work in this thesis will have a positive effect to your lives. Also I want to thank all **clinicians, researchers and technicians** that contributed to combining and curating these materials and resulting data.

I have had the honour and pleasure of learning and experiencing science as a team sport with colleagues all over the globe. Thank you **Coos Baakman, Elke de Boer, Arjan de Brouwer, Han Brunner, John van Dam, Illja Diets, Christian Gilissen, Daan Gilissen, Juliet Hampstead, Tom Heskes, Joery den Hoed, Matthew Hurles, Martijn Huynen, Joanna Kaplanis, Robin van der Lee, Stefan Lelieveld, Evgeni Levin, Rolph Pfundt, Kyle Retterer, Kaitlin Samocha, Maria Nabais Sá, Lot Snijders-Blok, Joris Veltman, Hanka Venselaar, Lisenka Vissers, Gert Vriend, Mieke Wesdorp, and Zhancheng Zhang**, and other collaborators for contributing to the work in this thesis and allowing me to learn from you.

It is funny how life turns out: When I was a patient at the Radboudumc I sought distraction by watching the construction of a building. I ended up conducting my PhD in that very building. Without modern medicine and the excellent healthcare I received back then, I would not be alive today. Anyone that provided medical care or support for me at that time especially at the departments **B80, B70**, and, **Ortopedie**: THANK YOU! And a very special thanks to **Bart Schreuder, Paul Brons, Jos Bokkerink, Marietje van Mellekom, Annet Bongaerts, and Irma Nuij**.

My path towards an academic career knew many twists and turns, guided by inspirational supervisors. **Bert Hoeks**, during my Bachelor you taught me to sit down, slow down, and take the time to understand complex algorithms. I learned how exciting and empowering it can be to master complex concepts. This encouraged me to pursue a Master's degree, during which I was unsure if I had the capacities needed to pursue an academic career. **Evgeni Levin and Tom Heskes**, the patience, encouragement, and, trust that you put in me during supervision of my Master's thesis helped remove these insecurities. Later on being awarded with the AIA best Master thesis in computing science of 2014 was the final push I needed to find a PhD project to pursue.

The work in this thesis was conducted at the departments **Centre for Molecular Bioinformatics** and **Human genetics** of the Radboudumc. I would like to thank both departments for allowing me the freedom to conduct my work that resulted in this thesis and making me feel very much included. **Peter-Bram 't Hoen**, you were able to strengthen the 'part-of-the-team' feeling at the CMBI and you helped lay the groundwork for many collaborations that I am sure will flourish in the coming years. I appreciated how you were always available for advice or a chat. **Han**

Ad

**Brunner**, you have a talent in making people feel included by being approachable and I was impressed how you actually knew the 400+ human genetics colleagues by name. The way you are able combine leading the department while contributing to both the clinic and science will remain an inspiration to me. **Martijn Huynen**, your experience in how the scientific field has evolved and changed over the years could always count on active listeners. I liked how much you made me experience the absence of hierarchy and how excited you always would get whenever I walked into your office with a scientific problem or an anecdote. You often also came to me with similar problems and one of those discussions resulted in contribution to **Robin van de Lee's** publication together with **John van Dam**. I will soon try to find time to watch *Lawrence of Arabia*.

My honoured paronyms, **Jakob Goldmann**, **Brooke Latour**, and **Renee Salz**, thank you for all of you support. **Jakob**, we were the first generation of Christian's PhD students together with **Stefan**. We clicked on a personal level, and found each other in our love to ferment things. Our yearly cider making day of the "Shaky Gold Wheel" always was absolute joy and I hope we will be able to keep doing that for years to come. You were always up for a scientific discussion and I admired how headstrong you could be at times. Keep up your awesome dancing moves! **Brooke**, I recall meeting you for the first time at an end-of-the-month research borrel and we immediately clicked. I kept meeting you after that everywhere. If the hospital was *a fishbowl*, I think we might be the *two lost souls swimming in it*. Our shared love of the journey to finding exquisite foods and drinks and the stories surrounding them bonded us. We ended up organising lots of fine food and drink events for colleagues together with **Alex**. The memory of when we met up with your childhood friend and ended up spending the entire afternoon back stage at the Best Kept Secret festival will always remain dear to me. **Renee**, we became offices mates at the final part of my PhD. You had a particular way of looking at life that inspired me to do some personal introspection. You have this keen ability of finding the best value deals. I dearly appreciated your brutal honesty. Many times we shared our hardships and happiness together after work by cooking, followed by watching lots of TV shows. I particularly enjoyed watching the women finals of the football world championship with you featuring America vs The Netherlands.

I was blessed with the roomies I have had over the years. I want to thank my **CMBI roomies: Jon Black, Joanna Lange, Wouter Touw, Hanka Venselaar**,

and the roomies that are better known as “**The Bondage Champions**”: **Dario Marzella, Lisette Meerstein-Kessel, Renee Salz, Joeri van Strien, and Anouk Verboven.** **Jon** thank you for inspiring me to further perfect my coding and for the many useful tech discussions we had. **Joanna**, we bonded by both being the last generation of Gert’s PhD students. I hope we will have more karaoke nights together in the future. **Wouter** you were the senior PhD student that showed me the ‘PhD-ropes’. I have since attempted to similarly be there for the next generations of PhD students. **Hanka**, somehow every time either of us had a question, it turned into an hour-long talk that switched topics between science, teaching, life, relationships, and everything else. We found each other in similar challenges faced. I will always remember your wedding party as the one that had the most and best dancing! Let’s soon plan the rollercoaster theme park we always talked about. **Dario**, somehow you have the habit of always being busy accommodating other people. You are a great host. I hope we will continue playing many more board games and walking through abandoned monasteries in the future. **Lisette**, you always resonated with calm, which greatly helped putting things I was stressing about in perspective. **Joeri**, you are the fellow founder of the soup group. I always enjoyed our walks and talks to the restaurant to get some food and conducting our joint review of the soup of the day. Thank you for being part of the CMBI borrel committee with me. **Anouk**, I was super happy to have a fellow geneticist join the CMBI and even more so that we shared the same room. Thanks for all the formal and informal talks we had.

At genetics I started out with **Jakob Goldmann, Stefan Lelieveld, Maartje van de Vorst, and Dimitra Zafiropoulou.** After a year **Jakob, Stefan** and I moved into a new office that would be known as **GINOMICS**, with the additional members over the years: **Stéphanie Cornelis, Juliet Hampstead, Brechtje Hoegen, Simone Kersten, Erdi Kucuk, Susanne Roosing, Karolis Šablauskas, Wouter Steyaert, Burcu Yaldiz, Kevin Yuay, and Jard de Vries.** **Stefan**, your social way of connecting with people was infectious. You somehow knew everyone in wet-lab, clinics, and bioinformatics. Whenever you randomly said “Hey Lou!” from behind your screen an interesting discussion always followed. Quoting Dumbert movies without you will never be the same. **Maartje**, you set the office rule: ‘always go together with roomies for a coffee’. I always tried to follow this rule well after we switched office. As bioinformaticians, the joint ‘getting coffee’ often was the social interaction we didn’t realize we needed. **Dimitra**, you always brightened

Ad

up my day with clever remarks and Greek philosophy. **Stéphanie**, our very open discussions changed my perspective and made me so much more ‘woke’. Thank you for acknowledging my blunt questions with your well thought-out answers. I have come to know you as an intellectual kind friend of many depths and one of the most inclusive of colleagues. **Juliet**, when you started we were talking a lot about our potential postdoc trajectories. I stuck around a bit longer before actually taking off. Luckily this made me experience how you coach people. I am sure you will make a great PI one day. **Brechtje**, you always had a captivating story ready and were always available for a chat. I liked how you always aimed and succeeded in keeping everyone’s spirits up. **Simone**, you joined the room and thereby increased our Aloe vera population by 400%. I am sure the added oxygen had a positive effect on all of us. We had lots of coffees together that was often joined by a talk on both serious and mundane struggles. I love the way you are artistically creative and I hope you will continue to practice your arts. **Erdi**, your enthusiasm for many obscure board games that I never heard about before sparked over to me. I can’t wait till we have another board game night together. **Suus!** Our in-office senior scientist! You were often very busy, but somehow you always found time for funny strict Dutch-English translations. Thank you for ant fucking my work so often, this thesis wouldn’t be in the same level of quality without it. It was unfortunately peanut butter for us when you transitioned to another office. I hope you will keep taking good care of Vera. And remember, you will always be welkom! **Karolis**, somehow we found each other in our love for old school hip hop after watching Office Space. I particularly enjoyed our cooking sessions. I hope you will still have some time for *Developing! Developing! Developing!* in the future. **Wouter**, you always seemed to have a listening ear ready for a colleague in need, followed by a calmly given piece of advice. **Burcu**, thank you for the fun chats we had and for broadening my knowledge on Turkish culture. **Kevin**, okay you were not really a roommate, but we included you as an honorary member of the GINOMICS. Even though you were only in Nijmegen for a short time, you fit in so well and joined many nice events. I hope to see you again at a next ASHG or ESHG. **Jard**, you were the silent disc jockey in our room. Even though you were with us for only a year, you will forever be highly valued as the co-founding member of GINOMICS.

**Judith Grolleman** and **Manon Oud**, our PhDs started around the same time and you both brought constant support throughout my PhD academically and personally. You were there to celebrate moments of happiness, or provide

comfort when work brought me down. And no, I still don't know the difference between a hamster and a guinea pig. **Juut!** We've celebrated many great moments together throughout our PhDs. You always seemed to have a bottle of champagne stashed away for the next celebration. You have unique social skills to bind and excite people towards a common goal. You were the champion of undeserved manuscript rejections, and yet you never gave up hope. You radiate kindness, trust and intelligence. **Manon**, you were the most organised PhD student I have ever known. Besides that you are kind, intelligent, and always seemed to have time for someone else even when you were very busy. It doesn't surprise me at all that you often walked away with well-deserved awards at conferences. You never wavered from tough questions. You taught me so many basics of genetics and were never too shy to dive in the more advanced topics. Speaking of diving, I enjoyed our scuba diving trips and hope you will find more time for exploring the watery depths of calmness and serenity in the future.

Working at two departments resulted in double the perks, but also double the bureaucracy. **Barbara van Kampen**, **Ineke Zaalmink**, **Baukje Oosterhof-Konings**, **Doménique Nijsten**, and **Dennis Vissers**, I appreciate how you were there for me to navigate many of these challenges. **Arthur Pistorius**, **Jon Black**, and **Coos Baakman**, I could always count on you for assistance with hardware, software or Linux issues that ended up saving me many hours of work. **Steven Castelein**, **Marc Pieterse**, **Rick de Reuver**, and **Maartje van de Vorst** thank you for assisting me figure out and navigate the infrastructure of the genetics server. This helped me make my first baby-steps in genomics. My productivity was very much boosted by having a clean desk. **Monique Jacobs** and the **cleaning team at Human Genetics**, thank you for making sure this was never a worry on my mind.

I have had the honour of being a supervisor during my PhD. **Brecht van den Berg**, **Lucca Derks**, **Daan Gilissen**, **Daniel Rademaker**, and **Wouter-Michiel Vierdag** thank you for teaching me by coaching you. **Daan**, your internship laid the groundwork design of the characteristic MetaDome tolerance landscape. Your place on the publication was well deserved. I have come to know you as a hard worker. I am glad to see you have become a colleague now at CMBI. **Daniel**, your internship introduced you to many datasets you ended up working with later on in your PhD. I admired how you always knew what everyone was busy with or working on. You had a hard time saying no, but this also resulted in many

Ad

upcoming publications. I am glad you chose to pursue a PhD in the department. You have become an essential component of the CMBI by being a true connector of people. I am sure you will excel in your PhD.

Perhaps one of the biggest perks of working in academia is the scientific conferences that are often located in fantastically vibrant or exotic locations. To all of my colleagues that joined me in creating memories that I will forever remember with a smile: **Chantal Deden, Illja Diets, Sinje Geuer, Christian Gilissen, Jakob Goldmann, Judith Grolleman, Alex Hoischen, Simone Kersten, Robin van der Lee, Stefan Lelieveld, Britt Mossink, Gaël Nicolas, Manon Oud, Margot Reijnders, Roos Schellevis, Lot Snijder-Blok, Marloes Steehouwer, and Richarda de Voer**. Thank you **Jakob** for being room buddies at the RIMLS PhD retreat 2016 with infinite patience as I lost the room key three times. Room buddy **Robin** at the BioSB 2016. I remember that **Jakob** and **Stefan** joined me in the monastery of Rolduc dungeon for a beer, where Stefan lost his way but was guided by my text message. The ESHG 2017 in Copenhagen was my first experience with the 'Nijmegen house' tradition. Thank you **Christian, Jakob, Judith, Alex, Stefan, Gaël, Manon, Margot** and **Richarda** for the spontaneous fun after-conference-borrels with amazing food in this weird and humongous house that contained revealing pictures of the owners and their 'interesting' book collection. I will forever remember the conference party. The tradition continued in the ASHG 2017 in Orlando. With the massive apartment, interesting taxi conversations together with **Han**, the need for cooking our own lunches if we didn't want the fatty foods from Denny's, stubbornly walking to the conference in a country designed for cars, and lots and lots of bacon in the morning. **Illja, Sinje, Christian, Stefan, Margot, Roos, and Lot** thank you for the fantastic memories. **Sinje** and **Roos** thank you for joining me in visiting the **Give Kids The World Village**. **Stefan, Margot, Roos, and Lot** I can't wait to go to the theme parks again with you and face the music of a mariachi band after being bombarded with impressions all day. **Chantal** and **Britt**, I enjoyed our pre-conference ESHG 2019 exploration of Gothenburg and surroundings. After which we group housed together with **Alex, Simone, Manon, Lot, Marloes, and Richarda** in another 'Nijmegen house' villa with a sauna that we didn't dare to use. I remember everyone panicked that I was gone since I had my presentation the next day, but after a more thorough search you found I was vast asleep in my bed all along.

Perhaps one of the best pieces of advice I ever got from senior PhD students was to join VrijMiBo in the Aesculaaf. I tried my best to live up to this advice and passed it on to the next generation of PhD students. We have an exceptionally social group of colleagues that always made you feel included. Every Friday there would be a core group of colleagues that never seemed to miss a single Aesculaaf Friday and every time I met new people too. We would mostly end up having grouped dinner afterwards and sometimes even visited another bar. These VrijMiBo's been a de-stressor in the weeks I needed them the most and energised me in good weeks. They helped me get to know my colleagues on a more personal level. Many collaborations and friendships started here. The going into the city afterwards helped me discover food places and parts of Nijmegen that I wouldn't have otherwise. **Thank you to everyone that was part of one of these Fridays.**

Besides the spontaneously weekly VrijMiBo's I found much joy and gratitude in organizing and hosting social events. At CMBI this started out with our 'own version of the BioSB conference named BYOSB ('Bring Your Own Speciaal Bier') that had multiple recurrences and later the Dutch Cultural movie night. At genetics the honour was bestowed upon **Chantal Deden, Suzanne de Bruijn, Jettie van Engelen, Stephan Maas, Marc Pauper**, and me to organise the yearly day-out or DUC2018. Thank you for the pre-, during-, and post- fun of our well-oiled organisation team! **Hans van Bokhoven**, thank you for the excellent idea of going deeper than ever before with human genetics. The first Human Genetics Nijmegen scuba-diving event was a success and hopefully more will follow. **Brooke Latour, Alex Hoischen**, and me clicked as lovers and explorers of food and drinks and sought to share this enthusiasm with others by organising many food and drink related events. We have grown to be like-minded friends in the process. Thank you **Alex** for teaching me the meaning of hospitality by exemplifying it in its purest form. **Thank you to everyone that joined or aided in organising these events.**

I was surrounded by many fantastic colleagues during my PhD. Some of you I haven't mentioned by name, so thank you **Rocío Acuña-Hidalgo, Peer Arts, Galuh Astuti, Sander Bervoets, Jordy Coolen, Rosanne van Deuren, Margo Dona, Tom Ederveen, Dei Elurbe, Siebren Faber, Daniel Garza, Josh Gillard, Evelien Hurkmans, Charlotte Kaffa, Mubeen Kahn, Gelana Khaveeza, Marije Klumpers, Michael Kwint, Ideke Lamers, Anne Niehues, Machteld Oud, Iris Te Paske, Gayatri Ramakrishnan, Simon van Reijmersdal, Rick de Reuver,**

Ad

**Tabea Riepe, Dimitrijs Rots, Carolien Ruesen, Bart van der Sanden, Ralph Slijkerman, Balaji Venkatasubramanian, and Petra de Vries**, for joining social events, being available for a chat, going out for lunch/dinner together, helping out with a tough question, or getting a coffee with an appelflap together. Thank you and **all other colleagues from Human Genetics and CMBI** that helped create an atmosphere at work that made lasting memories.

Thank you **Ruud van de Wiel**, for creating and shaping the cover and thematic chapter pictures in this thesis.

To the **members of De Waterman scuba diving club**, for being a weird bunch of awesome people that so selflessly and happily support the club in any way. To **Mari and Marja van der Lee**, for all the effort you put in De Waterman scuba diving club. The annual scuba dive trips you organised were always rewarded with lasting memories, and, of course, cooling down in the snow on the annual ski trips together with the **rest of the van der Lee family** to Austria. You have forever added meaning to my definition of ‘gezelligheid’.

To my dear friends **Ralph Coolen, Erik van de Lee, Tiemco Nelissen, Andy Ouwerkerk, David Strijbos, and Rob Wissenburg**. More than once I had to cancel planned social events with you to get to where I am now, somehow you forgave me every time. I could always rely on you to bring me down-to-earth again with your sobering remarks. Thank you for sticking with me all these years.

Dhanyabaad **Basanta, Meela, and Prayash Shakya** for making me feel included in a family a world apart. I wish we will soon be able to meet in-person and outside the confinements of video calling.

To my family **Willebrord, Diny, Geert van de Wiel, Wilma, Rob, Matt, and Luke van Dijk. Papa en mama**, bedankt voor jullie onuitputtelijke steun en vertrouwen in mij door de jaren heen. Bedankt dat jullie letterlijk altijd voor iedereen belangeloos en onbaatzuchtig klaar stonden. Jullie hebben mij vele levenslessen geleerd, ook al wilde ik er niet altijd naar luisteren. De vele boeren wijsheden die jullie vaak probeerden over te brengen aan ons hebben zeker bijgedragen aan de kwaliteit van dit proefschrift. **Papa**, bedankt dat je na mijn ziekzijn vaak vertelde “*leer maar goed door, want jij kan geen timmerman worden*”, die woorden hebben mij altijd gemotiveerd net weer een trede hoger proberen te stappen in

mijn academische carrière. **Mama**, bedankt voor alle kerent dat je een luisterend oor bood wanneer ik het nodig had. **The Gurt** en **Wimla**, ik kan mij geen betere brussen bedenken dan jullie. Jullie staan ook echt altijd voor mij klaar en kennen geen lengtes van opgeven om elkander te helpen. Wij zijn hechter geworden door alles wat we samen hebben meegemaakt. Zonder jullie steun en misschien ook wel opvoeding, ik ben tenslotte '*de klène*', zou ik niet zijn wie ik nu ben. **Rob**, ik ben blij dat jij goed voor mijn grote zus zorgt. Samen met jullie lieve **Matt** en **Luke** vormen jullie een prachtig gezin. Ik hoop dat jullie een goede toekomst tegemoet gaan.

Dhanyabaad my maya, dhanyabaad **Pallavi**. We were together when the world seemed to be falling apart around us, but together we were and that made all the difference. Thank you for being a shining guiding light in the darkest of days, and thank you for making bright days so much brighter. Thank you for teaching me a different perspective on western society from your own cultural identity. Thank you for your ever-present kindness and generosity. Thank you for your radiating written words in poetry and stories, and your continuous and unwavering support that helped me cover the final stretch of my PhD journey. Whatever the future may bring us, knowing we may face it together makes it all the brighter. I can't wait to dive into new adventures with you. Ma timilai maya garchu.

Ad