Methods for Variant Analysis II

Bio5488
Sheng Chih (Peter) Jin, Ph.D.
03/20/2024
Study designs linking genotype with phenotype

- Sequencing as a discovery tool
- Genotyping as a discovery tool
- Family-based studies
  - Linkage studies of multiplex families
  - Extremes of a distribution
- Case-control association studies

Effect size

Frequency of minor allele in the population
Most human genetic variants are rare; functional variants tend to be rare

**Fig. 1** Distribution of genetic variants across the genome. Common (allele frequency $\geq 0.5\%$) and rare (allele frequency $< 0.5\%$) variant counts are shown above and below the x-axis, respectively, within 1-Mb concatenated segments (see Methods). Segments are stratified by CADD functionality score, and sorted based on their number of rare variants according to the functionality category. There were 22 high CADD, 22 medium CADD and 34 low CADD coding segments, and 40 high CADD, 238 medium CADD and 2,381 low CADD noncoding segments. Noncoding regions of the genome with low CADD scores ($<10$, reflecting lower predicted function) have the largest levels of common and rare variation (noncoding plot region, dark and light blue, respectively), followed by low CADD coding regions (coding plot region, dark and light blue, respectively). Overall, the vast majority of human genomic variation comprises rare variation.

Taliun et al., Nature 2021
Variant analysis overview

1. FASTQ
2. Align (BWA)
3. BAM
4. Detect SNP/INDELs (GATK or DeepVariant)
5. VCF

IGV and validation

Statistical analysis to identify candidate genes or pathways

Annotation and filtering
From raw reads to analysis-ready data (GATK Best Practices)

1. Mapping
2. Marking duplicates
3. Base recalibration

https://gatk.broadinstitute.org/hc/en-us
Step 1: Map the sequence reads to the reference

- Raw Unmapped Reads
  - uBAM or FASTQ

- Map to Reference

- Raw Mapped Reads
  - BAM

- Mark Duplicates

- Recalibrate Base Quality Scores

- Analysis-Ready Reads
  - BAM

Mapping and alignment algorithms:
- BWA for DNA
- STAR for RNAseq

Enormous pile of short reads from sequencer

Reference genome
Reads mapped to reference

www.csc.fi/en/web/training/-/gatk2019
Output format: Binary Alignment Map

**HEADER** describing various metadata for all reads

```
@HD VN:1.6  SO:coordinate
@SQ SN:seq1 LN:394893
@SQ SN:seq2 LN:92783
@RG ID:A SM:SAMPLE_A
```

**RECORDS** containing structured read information (1 line per record)

- Mapping info summarizes position, quality, and structure for each read
- Mate information points to the read from the other end of the molecule

www.csc.fi/en/web/training/-/gatk2019
CIGAR summarizes alignment structure

CIGAR = Concise Idiosyncratic Gapped Alignment Report

RefPos: 1 2 3 4 5 6 7 8 9
Reference: C C A T A C T G A
Read: T C A T - C A G
POS: 2
CIGAR: 1S3M1D2M1I1M = 14 20 CATCTAG ...
Step 2: Mark duplicates to mitigate duplication artifacts

Duplicates = non-independent measurements of a sequence fragment

- Must be removed to assess support for alleles correctly

Raw Unmapped Reads
- uBAM or FASTQ

Map to Reference

Raw Mapped Reads
- BAM

Mark Duplicates

Recalibrate Base Quality Scores

Analysis-Ready Reads
- BAM

= error propagated in duplicates

Duplicates are not removed; they are simply tagged, unless you request their removal!

www.csc.fi/en/web/training/-/gatk2019
Where does the duplication come from?

LIBRARY DUPLICATES
Caused by PCR

OPTICAL DUPLICATES
Occur during sequencing


https://depts.washington.edu/molmicdx/mdx/tests/NGS16S.shtml

www.csc.fi/en/web/training/-/gatk2019
Step 3: Base Recalibration (BSQR) corrects for machine errors

- Sequencers make systematic errors in base quality scores
- Sequencer quality cannot include PCR-based errors
- BQSR corrects the quality scores (not the bases)
DeepVariant = Data Visualization + Deep Learning

Sequence alignments

Alignment Tensor

Each candidate +/- 7 bp

Encode all alignments to a 15 x 4 x 3 tensor

VariantNET
2 convolution layers
3 full connected layers

Genotype

[0.5, 0.0, 0.5, 0.0]

A C G T

Softmax output

[0.98, 0, 0, 0.02]

het

none

hom

complex
Comparison of GATK and DeepVariant

Total Errors (SNP and Indel) on pFDA HG002

DeepVariant v0.8.0  GATK 1.1.0 HaplotypeCaller  Strelka2.9.10  Freebayes1.3.1  Errors at GATK 4.1-HC at 30x Coverage

The hybrid model of DeepVariant for PacBio HiFi and Illumina reads represents the most accurate method for analyzing the human genome.

https://blog.research.google/2020/09/improving-accuracy-of-genomic-analysis.html?m=1
Raw VCF files are naked. Interpretation requires annotation

**Header**

```plaintext
##fileformat=VCFv4.3
##fileDate=20090805
##source=variantcallerXYZ
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=379d618ff66bebda,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

**Records**

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>MOM</th>
<th>DAD</th>
<th>KID</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>14370</td>
<td>rs6054257</td>
<td>G</td>
<td>A</td>
<td>29</td>
<td>PASS</td>
<td>NS=3;DP=14;</td>
<td>GT:GQ:DP:HQ 0</td>
<td>0:48:1:51,51</td>
<td>1</td>
<td>0:48:8:51,51</td>
</tr>
<tr>
<td>20</td>
<td>17330</td>
<td>.</td>
<td>T</td>
<td>A</td>
<td>3</td>
<td>q10</td>
<td>NS=3;DP=11;</td>
<td>GT:GQ:DP:HQ 0</td>
<td>0:49:3:58,50</td>
<td>0</td>
<td>1:3:5:65,3</td>
</tr>
<tr>
<td>20</td>
<td>1234567</td>
<td>microsat1</td>
<td>GTC</td>
<td>G,GTCT</td>
<td>50</td>
<td>PASS</td>
<td>NS=3;DP=9;</td>
<td>GT:GQ:DP 0</td>
<td>1:35:4</td>
<td>0/2:17:2</td>
<td>1/1:40:3</td>
</tr>
</tbody>
</table>

*GT*: Genotype; *GQ*: Genotype quality; *DP*: Read depth; *HQ*: Haplotype quality
Variant annotation

Types of variants
- Protein coding
  Nonsense, frameshift
  Missense
- Non-coding
  Splice site disrupters
  Promoter disrupters
  Regulatory element disrupters
- Structural variants
  May affect multiple features

Types of annotations
- Population frequencies
- Evolutionary constrain
- Biochemical consequences
- In vitro or in vivo assays
  ENCODE, PsychENCODE
  GTEx
- Phenotypes
  Mouse knockouts
  Cellular assays
Annotation provides context for interpretation

Genetic variation

...CCTCATG\textcolor{red}{G}ATGGAAAA...
...CCTCATG\textcolor{blue}{A}ATGGAAAA...
...CCTCATG\textcolor{green}{G}ATGGAAAA...
...CCTCATG\textcolor{blue}{A}ATGGAAAA...
...CCTCATG\textcolor{green}{G}ATGGAAAA...
...CCTCATG\textcolor{blue}{A}ATGGAAAA...
...CCTCATG\textcolor{green}{G}ATGGAAAA...
Types of inheritance for germline mutations

**Autosomal Dominant**
- Individuals carrying one mutated copy of a gene in each cell will be affected by the disease
- Each affected person usually has one affected parent
- Tends to occur in every generation of an affected family

**Autosomal Recessive**
- Affected individuals must carry two mutated copies of a gene
- Parents of affected individual are usually unaffected and each carry a single copy of the mutated gene (known as carriers)
- Not typically seen in every generation.

**Mitochondrial**
- Only females can pass on mitochondrial conditions to their children (maternal inheritance)
- Both males and females can be affected
- Can appear in every generation of a family

**X-linked Dominant**
- Females are more frequently affected than males
- Fathers cannot pass X-linked traits to their sons (no male-to-male transmission)

**X-linked Recessive**
- Males are more frequently affected than females
- Families with an X-linked recessive disorder often have affected males, but rarely affected females, in each generation
- Both parents of an affected daughter must be carriers
- Only mother must be carrier of affected son (fathers cannot pass X-linked traits to theirs)

https://www.ncbi.nlm.nih.gov/books/NBK132152/
Online Mendelian Inheritance in Man (OMIM): From genotype to phenotype

GROWTH/DIFFERENTIATION FACTOR 1; GDF1

HGNC Approved Gene Symbol: GDF1

Cytogenetic location: 19p13.11  Genomic coordinates (GRCh38): 19:18,868,545-18,896,158 (from NCBI)

Gene-Phenotype Relationships

<table>
<thead>
<tr>
<th>Location</th>
<th>Phenotype</th>
<th>Phenotype MIM number</th>
<th>Inheritance</th>
<th>Phenotype mapping key</th>
</tr>
</thead>
<tbody>
<tr>
<td>19p13.11</td>
<td>Congenital heart defects, multiple types, 6</td>
<td>613854</td>
<td>AD</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Right atrial isomerism (Ivemark)</td>
<td>208530</td>
<td>AR</td>
<td>3</td>
</tr>
</tbody>
</table>

MYOSIN, HEAVY CHAIN 6, CARDIAC MUSCLE, ALPHA; MYH6

HGNC Approved Gene Symbol: MYH6

Cytogenetic location: 14q11.2  Genomic coordinates (GRCh38): 14:23,381,957-23,408,273 (from NCBI)

Gene-Phenotype Relationships

<table>
<thead>
<tr>
<th>Location</th>
<th>Phenotype</th>
<th>Phenotype MIM number</th>
<th>Inheritance</th>
<th>Phenotype mapping key</th>
</tr>
</thead>
<tbody>
<tr>
<td>14q11.2</td>
<td>Atrial septal defect 3</td>
<td>614089</td>
<td>AD</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Cardiomyopathy, dilated, 1EE</td>
<td>613252</td>
<td>AD</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Cardiomyopathy, hypertrophic, 14</td>
<td>613251</td>
<td>AD</td>
<td>3</td>
</tr>
</tbody>
</table>

Alternatives:
- ClinGen (https://clinicalgenome.org/)
- DECIPHER (https://www.deciphergenomics.org/)
The objective is to identify variants that disrupt gene function.

<table>
<thead>
<tr>
<th>No mutation</th>
<th>Point mutations</th>
<th>Missense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent</td>
<td>Nonsense</td>
<td>non-conservative</td>
</tr>
</tbody>
</table>

DNA level:
- TTC
- TTT
- ATC
- TCC
- TGC

mRNA level:
- AAG
- AAA
- UAG
- AGG
- ACG

Protein level:
- Lys
- Lys
- STOP
- Arg
- Thr

Impact difficult to predict


## Predicting effects of genetic variants

### Table 1. Summary of functional impact prediction methods analyzed in our study

<table>
<thead>
<tr>
<th>Order</th>
<th>Prediction Method</th>
<th>The Employed Model</th>
<th>Feature Set</th>
<th>Variant Type</th>
<th>Update (Y/N)</th>
<th>Published Journal</th>
<th>Web Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CADD</td>
<td>Support Vector Machine</td>
<td>63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), ENCODE project and UCSC genome browser tracks</td>
<td>All types of SNPs</td>
<td>Y</td>
<td>Nature Genetics (2014)</td>
<td><a href="https://cadd.gs.washington.edu">link</a></td>
</tr>
<tr>
<td>2</td>
<td>DANN</td>
<td>Deep Learning</td>
<td>63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), ENCODE project and UCSC genome browser tracks</td>
<td>All types of SNPs</td>
<td>N</td>
<td>Bioinformatics (2015)</td>
<td><a href="https://cbcl.ics.uci.edu/public/data/DANN/">link</a></td>
</tr>
<tr>
<td>3</td>
<td>FATHMM-MKL</td>
<td>Support Vector Machine</td>
<td>10 feature groups including 46-way sequence conservation, histone modification (ChIP-Seq), transcription factor binding sites (TFBS PeakSeq), open chromatin (DNase-Seq), 100-way sequence conservation, GC content, open Chromatin (FAIRE), transcription factor binding sites (TFBS SPP), genome segmentation and footprint from ENCODE project</td>
<td>All types of SNPs</td>
<td>Y</td>
<td>Bioinformatics (2015)</td>
<td><a href="http://fathmm.biocompute.org.uk/">link</a></td>
</tr>
<tr>
<td>4</td>
<td>FunSeq2</td>
<td>Scoring System</td>
<td>4 feature groups including variants in potential regulatory elements, nucleotide-level impact of regulatory variants, variants in conserved regions and network analysis of variants associated with genes</td>
<td>All types of SNPs (It is designed for non-coding variants)</td>
<td>Y</td>
<td>Genome Biology (2014)</td>
<td><a href="http://funseq2.gersteinlab.org/">link</a></td>
</tr>
<tr>
<td>5</td>
<td>PredictSNP2</td>
<td>Ensemble Method</td>
<td>5 functional prediction scores of variants</td>
<td>All types of SNPs</td>
<td>Y</td>
<td>PLOS Computational Biology (2016)</td>
<td>[link](<a href="https://forschmidt.chenii.muni.cz/">https://forschmidt.chenii.muni.cz/</a> predictSNP2)</td>
</tr>
<tr>
<td>6</td>
<td>SIFT</td>
<td>Probability Estimation</td>
<td>Protein sequence conservation among homologs</td>
<td>Non-synonymous</td>
<td>Y</td>
<td>Nature Protocol (2009)</td>
<td><a href="http://sift.jcvi.org">link</a></td>
</tr>
<tr>
<td>7</td>
<td>PROVEAN</td>
<td>Scoring System</td>
<td>Protein sequence conservation among homologs</td>
<td>Non-synonymous</td>
<td>Y</td>
<td>PLOS ONE (2012)</td>
<td><a href="http://provean.jvri.org/index.php">link</a></td>
</tr>
<tr>
<td>8</td>
<td>MetaLR</td>
<td>Logistic Regression</td>
<td>9 functional prediction scores of variants</td>
<td>Non-synonymous</td>
<td>N</td>
<td>Human Molecular Genetics (2015)</td>
<td>No website</td>
</tr>
<tr>
<td>10</td>
<td>MutationAssessor</td>
<td>Scoring System</td>
<td>Sequence homology of protein</td>
<td>Non-synonymous</td>
<td>Y</td>
<td>Nucleic Acids Research (2011)</td>
<td><a href="http://mutationassessor.org/x/">link</a></td>
</tr>
<tr>
<td>11</td>
<td>PrimateAI</td>
<td>Deep Learning</td>
<td>The protein structure and 51-length amino acid sequence centered at the variant of interest</td>
<td>Non-synonymous</td>
<td>Y</td>
<td>Nature Genetics (2018)</td>
<td>[link](<a href="https://basespace.illumina.com/s/">https://basespace.illumina.com/s/</a> cPcCSme6vhb4)</td>
</tr>
<tr>
<td>12</td>
<td>M-CAP</td>
<td>Gradient Boosting Tree</td>
<td>Some pre-existing pathogenicity scores such as SIFT, CADD. Some pre-existing conservation scores. Four custom amino acid level features.</td>
<td>Missense</td>
<td>Y</td>
<td>Nature Genetics (2016)</td>
<td><a href="http://bejerano.stanford.edu/MCAP/">link</a></td>
</tr>
<tr>
<td>13</td>
<td>REVEL</td>
<td>Random Forest</td>
<td>Multiple functional prediction scores of variants and sequence conservation scores</td>
<td>Missense</td>
<td>N</td>
<td>The American Journal of Human Genetics (2016)</td>
<td>[link](<a href="https://sites.google.com/site/">https://sites.google.com/site/</a> revelgenomics/)</td>
</tr>
<tr>
<td>14</td>
<td>MISTIC</td>
<td>Ensemble Method</td>
<td>4 feature groups including 690 functional measures, 8 multi-ethnic MAF, 8 conservation measures and 18 functional prediction scores</td>
<td>Missense</td>
<td>N</td>
<td>PLOS ONE (2020)</td>
<td><a href="http://lbgi.fr/mistic/">link</a></td>
</tr>
</tbody>
</table>

---

Wang et al. NAR Genomics and Bioinformatics 2022

[21]
Class of loss-of-function mutations affecting protein-coding variants

MacArthur et al. Human Molecular Genetics 2010
Human genomes typically contain > 150 genuine LoF variants. So which LoF mutation is truly disease-causing?

Table 1 | Median number of protein-coding variants and effects among world super-populations*

<table>
<thead>
<tr>
<th>Super-population code</th>
<th>Synonymous (het; hom alt)</th>
<th>Missense (het; hom alt)</th>
<th>Frameshift (het; hom alt)</th>
<th>Stop gain (het; hom alt)</th>
<th>Start lost (het; hom alt)</th>
<th>Splice donor (het; hom alt)</th>
<th>Splice acceptor (het; hom alt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUR</td>
<td>6961; 4317</td>
<td>7220; 4452</td>
<td>116; 55</td>
<td>116; 38</td>
<td>151; 146</td>
<td>93; 35</td>
<td>61; 52</td>
</tr>
<tr>
<td>AFR</td>
<td>9296; 4673</td>
<td>9347; 4820</td>
<td>163; 56</td>
<td>156; 31</td>
<td>196; 150</td>
<td>123; 32</td>
<td>78; 51</td>
</tr>
<tr>
<td>AMR</td>
<td>7257; 4314</td>
<td>7449; 4479</td>
<td>121; 56</td>
<td>121; 38</td>
<td>154; 145</td>
<td>96; 34</td>
<td>62; 50</td>
</tr>
<tr>
<td>SAS</td>
<td>7180; 4397</td>
<td>7366; 4550</td>
<td>123; 56</td>
<td>121; 39</td>
<td>159; 148</td>
<td>93; 36</td>
<td>68; 49</td>
</tr>
<tr>
<td>EAS</td>
<td>6502; 4759</td>
<td>6802; 4908</td>
<td>105; 66</td>
<td>113; 45</td>
<td>143; 149</td>
<td>89; 38</td>
<td>62; 54</td>
</tr>
</tbody>
</table>

AFR, individuals of African descent; AMR, individuals of admixed descent from the Americas; EAS, individuals of East-Asian descent; EUR, individuals of European descent; PP Del, PolyPhen2 predicted the missense variant to be deleterious; SAS, individuals of South-Asian descent; SIFT Del, SIFT predicted the missense variant to be deleterious. *We measured the average number of heterozygous (het) and homozygous alternate (hom alt) genotype counts among the 2,504 individuals sequenced by the 1000 Genomes Project. All genetic variants affecting genes were annotated with the Variant Effect Predictor and categorized by their most deleterious predicted effect.

OMG! Many observed LoF variants are even homozygous!

Elbeec et al. Nature Reviews Genetics 2017
If a disease phenotype is rare, the causal variant should likewise be rare

The mutational constraint spectrum quantified from variation in 141,456 humans

Karczewski et al. Nature 2020
A causal variant responsible for a rare disease should be rare across all ancestral backgrounds.
Disease-causal genes should be under strong purifying selection (mutation intolerance)

Figure 1. Gene constraint by functional class in gnomAD. Each dot represents one gene, and its position on the plot represents the expected (x axis) and observed (y axis) number of variants in 125,748 human exomes in gnomAD. Black diagonals represent the expected relationship in the absence of natural selection; colored lines represent the actual best fit relationship. For synonymous variants, where we expect minimal natural selection, the correlation is excellent, with almost all dots lining up right on the diagonal. For missense variants, increased density below the diagonal indicates that some genes are intolerant of

Minikel et al. Nature 2020
Numerous gene constraint scores developed for prioritizing candidate variants/genes across various mendelian inheritance models
~40% of developmental disorders are caused by \textit{de novo} mutations

\textbf{Prevalence and architecture of \textit{de novo} mutations in developmental disorders}

Deciphering Developmental Disorders Study*

The genomes of individuals with severe, undiagnosed developmental disorders are enriched in damaging \textit{de novo} mutations (DNMs) in developmentally important genes. Here we have sequenced the exomes of 4,293 families containing individuals with developmental disorders, and meta-analysed these data with data from another 3,287 individuals with similar disorders. We show that the most important factors influencing the diagnostic yield of DNMs are the sex of the affected individual, the relatedness of their parents, whether close relatives are affected and the parental ages. We identified 94 genes enriched in damaging DNMs, including 14 that previously lacked compelling evidence of involvement in developmental disorders. We have also characterized the phenotypic diversity among these disorders. We estimate that 42% of our cohort carry pathogenic DNMs in coding sequences; approximately half of these DNMs disrupt gene function and the remainder result in altered protein function. We estimate that developmental disorders caused by DNMs have an average prevalence of 1 in 213 to 1 in 448 births, depending on parental age. Given current global demographics, this equates to almost 400,000 children born per year.
Recessive variants can contribute to 4% of patients with developmental disorders.

Quantifying the contribution of recessive coding variation to developmental disorders

Hilary C. Martin1, Wendy D. Jones1,2, Rebecca McIntyre1, Gabriela Sanchez-Andrade1, Mark Sanderson1, James D. Stephenson1,3, Carla P. Jones1, Juliet Handsaker4, Giuseppe Gallone1, Michaela Bruntraeger6, Jeremy F. McRae4, Elena Prigmore1, Patrick Short1, Mari Niemi1, Joanna Kaplanis1, Elizabeth J. Radford4,4, Nadia Akawi1, Meena Balasubramanian6, John Dean7, Rachel Horton9, Alice Hulbert9, Diana S. Johnson6, Katie Johnson10, Dhavendra Kumar11, Sally Ann Lynch12, Sarju G. Mehta13, Jenny Morton14, Michael J. Parker15, Miranda Splitt16, Peter D. Turnpenny17, Pradeep C. Vasudevan18, Michael Wright16, Andrew Bassett1, Sebastian S. Gerety1, Caroline F. Wright19, David R. FitzPatrick20, Helen V. Firth1,13, Matthew E. Hurles1, Jeffrey C. Barrett1, on behalf of the Deciphering Developmental Disorders Study

We estimated the genome-wide contribution of recessive coding variation in 6040 families from the Deciphering Developmental Disorders study. The proportion of cases attributable to recessive coding variants was 3.6% in patients of European ancestry, compared with 50% explained by de novo coding mutations. It was higher (31%) in patients with Pakistani ancestry, owing to elevated autozygosity. Half of this recessive burden is attributable to known genes. We identified two genes not previously associated with recessive developmental disorders, KDM5B and EIF3F, and functionally validated them with mouse and cellular models. Our results suggest that recessive coding variants account for a small fraction of currently undiagnosed nonconsanguineous individuals, and that the role of noncoding variants, incomplete penetrance, and polygenic mechanisms need further exploration.
6.0% of and 6.9% of females had a pathogenic X-linked protein-truncating or missense/inframe variant

The contribution of X-linked coding variation to severe developmental disorders


Over 130 X-linked genes have been robustly associated with developmental disorders, and X-linked causes have been hypothesised to underlie the higher developmental disorder rates in males. Here, we evaluate the burden of X-linked coding variation in 11,044 developmental disorder patients, and find a similar rate of X-linked causes in males and females (6.0% and 6.9%, respectively), indicating that such variants do not account for the 1.4-fold male bias. We develop an improved strategy to detect X-linked developmental disorders and identify 23 significant genes, all of which were previously known, consistent with our inference that the vast majority of the X-linked burden is in known developmental disorder-associated genes. Importantly, we estimate that, in male probands, only 13% of inherited rare missense variants in known developmental disorder-associated genes are likely to be pathogenic. Our results demonstrate that statistical analysis of large datasets can refine our understanding of modes of inheritance for individual X-linked disorders.
Of patients prescreened with low-resolution CMA, 2.6% had a pathogenic CNV detected by WES; the diagnostic yield of pathogenic CNVs in patients that have not previously been screened with low-resolution CMA was 5.0%
Mosaic chromosome alternations can explain ~0.5% of patients with developmental disorders

ARTICLE

Detection of mosaic chromosomal alterations in children with severe developmental disorders recruited to the DDD study

Ruth Y. Eberhardt¹, Caroline F. Wright², David R. FitzPatrick², Matthew E. Hurles¹, Helen V. Firth¹,²,³,⁴*

¹Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, Hinxton, United Kingdom; ²University of Exeter Medical School, Institute of Biomedical and Clinical Science, Royal Devon and Exeter Hospital, Exeter, United Kingdom; ³Cambridge University Hospitals NHS Foundation Trust, Clinical Genetics, Addenbrooke’s Hospital, Cambridge, United Kingdom; ⁴MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom

ABSTRACT

Purpose: Structural mosaicism has been previously implicated in developmental disorders. We aimed to identify rare mosaic chromosomal alternations (MCAs) in probands with severe undiagnosed developmental disorders.

Methods: We identified MCAs in genotyping array data from 12,550 probands in the Deciphering Developmental Disorders study using mosaic chromosome alterations caller (McCA).

Results: We found 61 MCAs in 57 probands, many of these were tissue specific. In 23 of 26 (88.5%) cases for which the MCA was detected in saliva in which blood was also available for analysis, the MCA could not be detected in blood. The MCAs included 20 polymorphisms, comprising either 1 arm of a chromosome or a whole chromosome, for which we were able to show the timing of the event (25% mitosis, 40% meiosis I, and 35% meiosis II). Only 2 of 57 (3.5%) of the probands in whom we found MCAs had another likely genetic diagnosis identified by exome sequencing, despite an overall diagnostic yield of ~40% across the cohort.

Conclusion: Our results show that identification of MCAs provides candidate diagnoses for previously undiagnosed patients with developmental disorders, potentially explaining ~4.45% of cases in the Deciphering Developmental Disorders study. Nearly 90% of these MCAs would have remained undetected by analyzing DNA from blood and no other tissue.

Insights for variant interpretation from past studies

• Careful annotation and curation of loss-of-function variants is essential
• Not all exons are expressed equally
• Not all LoF disease genes are severely depleted of putative LoF variants
• An adjacent variant can change the predicted effect
• We can now assess the frequency of structural variants
• We should explore beyond coding regions

https://rarediseasegenomics.org/blog/six-lessons-for-variant-interpretation
Why are ~50% patients undiagnosed? How can we enhance gene discovery efforts and improve diagnosis yield?
Phased assembly long-read sequencing allows identification of large, complex structural variations in the human genome

- 64 assembled human genomes that represent 26 different human populations
- Found 107,590 SVs, of which 68% are not discovered by short-read sequencing, and 278 SV hotspots
- Characterized 130 of the most active mobile element source element and found 63% of all SVs arise by homology-mediated mechanisms

Ebert et al. Science 2021
Integrating RNA-sequencing data with genome-sequencing data can improve the interpretation of genetic variation

• Detect transcript-level changes, splice-altering/splice-disrupting variants, and weird transcript isoforms

• Assess the extent of nonsense-mediated decay in LoF-containing transcripts

• Evaluate the extent of biallelic expression at recessive disease genes harboring heterozygous variants

• Feasibility
  - Tissue availability
  - Developmental versus chronic disorder
Integrating RNA-sequencing data with genome-sequencing data can enhance diagnostic yield

Fig. 1. Experimental design and quality control. (A) Overview of the number of samples that underwent RNA-seq. We performed RNA-seq on 13 previously genetically diagnosed patients, 4 patients in whom previous genetic analysis had identified an extended splice site variant of unknown significance (VUS). 12 patients in whom genetic analysis had identified a strong candidate gene, and 34 patients with no strong candidates from previous analysis. RNA-seq enabled the diagnosis of 33% of patients overall, with the rate shown above the bar plots, varying depending on previous evidence from genetic analysis. (B) PCA based on gene expression profiles of patient muscle samples passing quality control (n = 61) and GTEx samples of tissues that potentially contaminate muscle biopsies shows that patient samples cluster closely with GTEx skeletal muscle. (C) Overview of experimental setup and RNA-seq analyses performed. Our framework is based on identifying transcriptional aberrations that are present in patients and missing in GTEx controls. Upon ensuring that GTEx and patient RNA-seq data were comparable, we validated the capacity of RNA-seq to resolve transcriptional aberrations in previously diagnosed patients and performed analyses of aberrant splicing, allele imbalance, and variant calling in our remaining cohort of genetically undiagnosed muscle disease patients.

Fig. 2. Types of pathogenic splice aberrations discovered in patients. RNA-seq identified a range of aberrations caused by both coding and noncoding variants, such as (A) exon skipping caused by an essential splice site variant in patient D7, (B) exon extension caused by a donor >+3 A>C extended splice site variant in nemaline myopathy patient C9 (where disruption of splicing at the canonical splice site results in splicing from intact GT/A motifs from the intron), (C) exon splicing gain caused by a 5’ donor splice site-creating variant in patient N22 with a donor >+5 G sequence context, resulting in a stronger splice motif than the existing canonical splice site, and (D) Intrinsic splice gain in patient N33 caused by a 5’ donor splice site-creating deep intronic variant. Evidence for wild-type splicing in addition to the inclusion of the pseudogene in the patient is in line with the milder Becker’s muscular dystrophy phenotype. Splice aberrations shown in (B) to (D) result in the introduction of a premature stop codon to the transcript.
• Evaluate the possibility of using blood for RNA-seq as a diagnostic tool when tissue biopsies are not available from patients

• Generated whole-blood RNA-seq from 94 individuals with undiagnosed rare diseases spanning 16 diverse disease categories and compared with RNA-seq data from 1,594 controls

• Developed a robust approach to evaluating the impact of expression, splicing, gene and variant filtering strategies on disease gene identification

• Demonstrated a 7.5% of diagnostic rate and used an expression outlier pipeline to identify ideal gene candidates for clinical interpretation
Integrating deep learning and artificial intelligence can improve genomic analysis and variant interpretation.

**Predicting Splicing from Primary Sequence with Deep Learning**

Authors
Kishore Jaganathan, Sofia Kyrizopoulos Panagiotopoulou, Jeremy F. McRae, ..., Seraphim Batzias, Stephanie J. Sanders, Kyle Kai-Ho Won

In Brief
A deep neural network precisely recognizes mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases.

**Accurate proteome-wide missense variant prediction with AlphaMissense**

Authors
Jun Cheng, Guido Navaris, Joshua Pan, Clare Drycott, Ahlab Zongzhiyal, Taylor Appelbaum, Alexander Pitzl, Lai Heng Wong, Michael Zielke, Tobias Sargent, Rosalia G. Schneidewind, Andrew W. Senter, John Jumper, Dennis Rouskas, Pushmeet Kohli, Ziga Arava

In Brief
AlphaMissense is an artificial intelligence-based model that accurately predicts missense variants in the human proteome.

**Highly accurate protein structure prediction with AlphaFold**

Authors
John Jumper, Richard Evans, Alexander Pitzl, Tim Green, Michael Figurski, Olaf Ronneberger, Kathryn Tamara, Brian Basler, Augustine Zibl, Anna Pinto-Pereira, Alan Bridgwater, Clemens Meyer, Steven A. Kuroiwa, Andrew J. Baldi, Andrew Conley, Bernardino Romero-Paredes, Stanislav Nikolenko, Kishub Jain, John Neuhaus, Trevor Back, Greg Petersen, David Rolfsreiter, Ellen Claassen, Michael Zielke, Martin Steinhage, Michalina Pechalska, Thomas Bergmann, Sebastian Bodenmiller, David Ehe, Oriol Vizoso, Andrew W. Senter, Korya Kaszewska, Pushmeet Kohli, Dennis Rouskas

In Brief
AlphaFold is an AI-based method that predicts protein structures with high accuracy.

**Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale**

Authors
Jian Zhou

In Brief
Orca enables the prediction of specific transcription factor motifs underlying cell-type-specific genome interactions. At the compartmental scale, virtual screens of sequence activities suggest a model for the sequence basis of chromatin compartments with a prominent role of transcription start sites.
Sophisticated statistical frameworks capable of jointly analyzing *de novo* and transmitted variants, along with protein-protein interaction networks, can reveal more risk genes.
Sequencing Studies of Extreme Phenotypes to Maximize Power for Rare Disease Gene Discovery
Hypothesis: individuals with rare variants in the same gene are concentrated in one extreme of the distribution.
Sequencing of 1,795 Icelanders who had lived to over age 85 without a diagnosis of Alzheimer’s disease found a rare protective variant p.A673T in APP

A mutation in APP protects against Alzheimer’s disease and age-related cognitive decline

Thorlakur Jonsson¹, Javinder K. Atwal², Stacy Steinberg¹, Jon Snaedal³, Palmi V. Jonsson³,⁸, Sigurbjorn Bjornsson³, Hreinn Stefansson¹, Patrick Sulem¹, Daniel Gudbjartsson¹, Janice Maloney², Kwame Hoyte², Amy Gustafson², Yichin Liu², Yanmei Liu⁶, Tushar Bhangaie³, Robert R. Graham², Johanna Huttenlocher¹⁴, Gyda Bjornsdottir⁸, Ole A. Andreassen⁷, Erik G. Jonsson⁸, Arnar Palotie⁷, Timothy W. Behrens², Olafur T. Magnusson¹, Augustine Kong³, Unnur Thorsteinsdottir¹,⁸, Ryan J. Watts² & Karl Stefansson¹,⁸

APP: p.A673T
MAF in AD group: 0.13%
MAF in elderly controls: 0.62%
MAF in cognitively intact controls at age 85: 0.79%

The table shows association results, comparing patients with Alzheimer’s disease (AD) to three different control groups (top line gives numbers for patients with Alzheimer’s disease only), N_elec/number of individuals with chip-based genotype information; N_genos/number of individuals with genealogy-based genotype information.
Use of extreme phenotype approach to identify genes, pathways, and targets in common diseases

An SCN9A channelopathy causes congenital inability to experience pain


Human Hypertension Caused by Mutations in WNK Kinases

FREDERICK H. WILSON | SANDRA DISSE-NGOEME | KEITH A. CHUATS | KAZUHIKO ISHIHARA | CAROL NELSON-WILLIAMS | ISABELLE DESITT | MURAT GUNEL | DAVID V. MILFORD

Science 10.1126/science.1098844

The New England Journal of Medicine

HIGH BONE DENSITY DUE TO A MUTATION IN LDL-RECEPTOR–RELATED PROTEIN 5

LYNN M. BOYDEN | JUNHAO MAO | PH.D. | JOSEPH BELSKY | M.D. | LYLE METZGER | M.D. | ANITA FARHI | R.N. | MARY A. MITNICK | PH.D. | DIANGUO WU | PH.D. | KARL INSOOGNA | M.D. | RICHARD P. LIFTON | M.D., PH.D.
Mutations with very large effect - Nature’s gift to therapeutics

• Provide causal link between genotype and phenotype, allowing studies to determine pathophysiological mechanisms

• Rare mutations with large effect can identify genes and pathways that can be manipulated for health benefit

• Suggest direction and magnitude of beneficial effects that can be achieved through a target
  - Orexin and sleep
  - Nav1.7 and pain
  - ROMK and blood pressure
  - PCSK9 and LDL cholesterol