Variants in the genome and possible consequences

Jan Traeger-Synodinos
Laboratory of Medical Genetics
National & Kapodistrian University of Athens
Categories of variants
- at the level of DNA

Change in nucleotide sequence

Change in amount of sequence (Copy Number Variants)

Change in position of sequence (Translocations/Inversions)
DNA changes can be caused by:
• An error during DNA replication during cell cycle
• Environmental factors (e.g. UV radiation)

**Hereditary mutations (or germline mutations):**
• Present throughout a person’s life
• Present in every cell in the body
• Can be inherited/transmitted

**Acquired (or somatic) mutations:**
• Arise at some time during a person’s life
• Present only in certain cells, not in every cell in the body
• Cannot be passed on to the next generation (unless in gametes)
Variation in a human genome

Based on 1000 Genomes Project, a typical genome differs from the reference human genome at approx. 4.1 - 5.0 million sites:

- 99.9% variants are single nucleotide variants & short indels

- 2,100 to 2,500 approx are structural variations, including:
  - 1,000 large deletions
  - 160 copy-number variants
  - 915 Alu insertions
  - 128 L1 insertions
  - 51 composite SINE/VNTR/Alu insertions
  - 4 nuclear mitochondrial DNA variants
  - 10 inversions

Medical Genetics, Athens University
Variation in a human genome

Additional complexity arises from the fact that the genome sequence of each individual has a few hundred thousand “private” variants (not found in the general population)!

Medical Genetics, Athens University

1000 Genomes Project Consortium, Nature 2015
Variant databases

Numerous variants, reported in numerous databases:

- Locus Specific Mutation Databases
- Disease Centered Mutation Databases
- Central Mutation & SNP Databases
  - HGMD
  - dbSNP-polymorphism repository
  - European Bioinformatics Institute EBI
  - ClinVar
  - OMIM
- National & Ethnic Mutation Databases
- Mitochondrial Mutation Databases
- Chromosomal Variation Databases
  - DECIPHER
  - Database of Genomic Variants
- Clinical & Patient Aspects Databases
- Etc. etc. etc.

http://www.hgvs.org/content/databases-tools
## Variant databases

E.g. Human Gene Mutation Database (HGMD®): an attempt to collate published gene lesions responsible for human inherited disease.

<table>
<thead>
<tr>
<th>Table:</th>
<th>Description:</th>
<th>Mutation totals (as of 2016-08-08)</th>
<th>Public entries: This site. Academic/non-profit users only</th>
<th>Total entries: HGMD Professional 2016.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene symbol</td>
<td>The gene description, gene symbol (as recommended by the HUGO Nomenclature Committee) and chromosomal location is recorded for each gene. In cases where a gene symbol has not yet been made official, a provisional symbol has been adopted which is denoted by lower-case letters.</td>
<td>134732</td>
<td>5158</td>
<td>7473</td>
</tr>
<tr>
<td>cDNA sequence</td>
<td>cDNA reference sequences are provided, numbered by codon.</td>
<td>5076</td>
<td>7709</td>
<td></td>
</tr>
<tr>
<td>Genomic coordinates</td>
<td>Genomic (chromosomal) coordinates have been calculated for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.</td>
<td>0</td>
<td>165408</td>
<td></td>
</tr>
<tr>
<td>HGVS nomenclature</td>
<td>Standard HGVS nomenclature has been obtained for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.</td>
<td>0</td>
<td>166488</td>
<td></td>
</tr>
<tr>
<td>Missense/nonsense</td>
<td>Single base-pair substitutions in coding regions are presented in terms of a triplet change with an additional flanking base included if the mutated base lies in either the first or third position in the triplet.</td>
<td>74707</td>
<td>105236</td>
<td></td>
</tr>
<tr>
<td>Splicing</td>
<td>Mutations with consequences for mRNA splicing are presented in brief with information specifying the relative position of the lesion with respect to a numbered intron donor or acceptor splice site. Positions given as positive integers refer to a 3’ (downstream) location, negative integers refer to a 5’ (upstream) location.</td>
<td>12498</td>
<td>17120</td>
<td></td>
</tr>
<tr>
<td>Regulatory</td>
<td>Substitutions causing regulatory abnormalities are logged in with thirty nucleotides flanking the site of the mutation on both sides. The location of the mutation relative to the transcriptional initiation site, initiation codon, polyadenylation site or termination codon is given.</td>
<td>2644</td>
<td>3522</td>
<td></td>
</tr>
<tr>
<td>Small deletions</td>
<td>Micro-deletions (20 bp or less) are presented in terms of the deleted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).</td>
<td>20803</td>
<td>27975</td>
<td></td>
</tr>
<tr>
<td>Small insertions</td>
<td>Micro-insertions (20 bp or less) are presented in terms of the inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).</td>
<td>8565</td>
<td>11668</td>
<td></td>
</tr>
<tr>
<td>Small indels</td>
<td>Micro-deletions (20 bp or less) are presented in terms of the deleted/inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).</td>
<td>2001</td>
<td>2662</td>
<td></td>
</tr>
<tr>
<td>Cross deletions</td>
<td>Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.</td>
<td>9555</td>
<td>14164</td>
<td></td>
</tr>
<tr>
<td>Cross insertions</td>
<td>Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.</td>
<td>2204</td>
<td>3410</td>
<td></td>
</tr>
<tr>
<td>Complex rearrangements</td>
<td>Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.</td>
<td>1344</td>
<td>1743</td>
<td></td>
</tr>
<tr>
<td>Repeat variations</td>
<td>Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.</td>
<td>411</td>
<td>495</td>
<td></td>
</tr>
</tbody>
</table>

Curators D.N. Cooper, E.V. Ball, P.D. Stenson, A.D. Phillips, K. Evans, S. Heywood, M.J. Hayden, M.E. Mort and M. Hussain

http://www.hgmd.cf.ac.uk/ac/index.php
Variants in the genome - the challenge!

To precisely interpret the effect of genome variants!

Important in:

- A diagnostic setting to support prognostic, therapeutic and reproductive advice with respect to human phenotypes
- The research context to support understanding of pathophysiological mechanisms supporting more effective patient management (therapy and even cure)
## Number of genes

Version 25 (March 2016 freeze, GRCh38)  
- Ensembl 86 Excerpt from General stats

<table>
<thead>
<tr>
<th>Element</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total No of Genes</td>
<td>58037</td>
</tr>
<tr>
<td>Protein-coding genes</td>
<td>19950</td>
</tr>
<tr>
<td>Long non-coding RNA genes</td>
<td>15767</td>
</tr>
<tr>
<td>Small non-coding RNA genes</td>
<td>7258</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>14650</td>
</tr>
<tr>
<td>Immunoglobulin/T-cell receptor gene segments</td>
<td>650</td>
</tr>
</tbody>
</table>
Possible consequences of variation on normal gene expression

- Aberrant quantity of gene product
- Aberrant quality of gene product
- Aberrant timing of expression (during development)
- Aberrant location of expression (abnormal cell or tissue)

Each variant will likely have its time and place of effect
Categories of variants – at the level of DNA

Change in nucleotide sequence

<table>
<thead>
<tr>
<th>C T G A C T C C T G A G G A G A A G</th>
<th>Substitution</th>
</tr>
</thead>
<tbody>
<tr>
<td>C T G A C T C T T G A G G A G</td>
<td>Deletion</td>
</tr>
<tr>
<td>C T G A C T C T G A G G A A A G</td>
<td>Insertion</td>
</tr>
</tbody>
</table>

Change in amount of sequence (Copy Number Variants)

Change in position of sequence (Translocations/Inversions)
Change in position of sequence

These include Translocations and Inversions

Translocations

Balanced

Unbalanced

Inversions

Usually balanced

Paracentric

Pericentric

Rare complex rearrangements

Balanced or unbalanced
Unbalanced Chromosomal rearrangements

Change of position in genome:
• Simple or Complex rearrangements

Possible consequences:
• Loss or gain of genetic material
  • Many possible consequences of variation on normal gene expression!!

![Diagram of chromosome rearrangements](image-url)
Balanced Chromosomal rearrangements

Change of position in genome
But NO loss or gain of genetic material

Possible consequences:
• Disruption of gene structure
  • No expression
  • Aberrant product (Fusion gene)
• Disruption of regulatory elements
  • No expression
  • Aberrant expression (quantity? timing?)
## Chromosomal rearrangements - methods for detecting

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Detection</th>
<th>Maximum resolution</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deletions and duplications</td>
<td>Insertions</td>
<td>Unbalanced translocations</td>
</tr>
<tr>
<td>Early 1970s Karyotyping/G-banding</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>FISH-based</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early 1990s CGH</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Mid 1990s M-FISH/SKY/COBRA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Late 1990s RxFISH</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Array-based</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early 2000s 1-Mb BAC array-CGH</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Tiling-path BAC array-CGH</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Oligonucleotide array-CGH</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Late 2000s SNP arrays</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NGS-based</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Categories of variants
- at the level of DNA

Change in nucleotide sequence

<table>
<thead>
<tr>
<th>Original</th>
<th>Substitution</th>
<th>Deletion</th>
<th>Insertion</th>
</tr>
</thead>
</table>

Change in amount of sequence (Copy Number Variants)

Change in position of sequence (Translocations/Inversions)
Copy Number Variants

Copy number variants are characterized by additional or missing copies of some segments of DNA.

CNVs are a widespread feature of the human genome in health and disease.

An estimated 4.8–9.5% of the genome contributes to copy number variation.

Sebat et al, Science 2004
Iafrate et al, Nat Genet, 2004
Copy Number Variants

CNVs can include genes and also intragenic regulatory regions

They can be:
• Simple or Complex deletions
• Simple or Complex duplications
• May also include inversions

Medical Genetics, Athens University

Zarrei M et al, Nat Rev Genet 2015
Copy Number Variants

The mechanisms through which CNVs may have deleterious effects include:

• Aberrant quantity of gene product (dosage sensitive genes)
• Aberrant gene product (eg if fusion gene arises from CNV event)
• Exposure of deleterious (recessive) variants on the trans allele (LOH) - indirect effect of CNV

Zarrei M et al, Nat Rev Genet 2015
Copy Number Variants

The phenotypic effect of CNVs can include:
• No effect!
  • Approximately 100 genes have been identified which can be completely deleted without apparent phenotypic consequences!
• Genetic predisposition to certain conditions
• Contribution to adaptive traits
• Severe disease expression (microdeletion/microduplication syndromes)

Zarrei M et al, Nat Rev Genet 2015
Copy Number Variants - methods for detecting

Arrays (aCGH or SNP)

MLPA

qPCR

NGS

Dear PH, Trends in Biotechnol, 2009
Categories of variants
- at the level of DNA

Change in nucleotide sequence

<table>
<thead>
<tr>
<th>Original Sequence</th>
<th>Mutation Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>C T G A C T C C T G A G G A G A A G</td>
<td>Substitution</td>
</tr>
<tr>
<td>C T G A C T C T T G A G G A G A A G</td>
<td>Deletion</td>
</tr>
<tr>
<td>C T G A C T C T - T G A G G A G A A G</td>
<td>Insertion</td>
</tr>
</tbody>
</table>

Change in amount of sequence (Copy Number Variants)

Change in position of sequence (Translocations/Inversions)

Medical Genetics, Athens University
# Change in nucleotide sequence

Usually within genes and their flanking regions  
Most common cause of monogenic diseases

<table>
<thead>
<tr>
<th>Involving single nucleotides:</th>
<th>Involving up to a few nucleotides:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Substitutions</td>
<td>• Micro-deletions</td>
</tr>
<tr>
<td>• Deletions</td>
<td>• Micro-insertions</td>
</tr>
<tr>
<td>• Duplications</td>
<td>• Indels (Combined micro-insertions/micro-deletions)</td>
</tr>
<tr>
<td>• Insertions</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Repeat variations</th>
<th>Complex variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Triplet repeats</td>
<td></td>
</tr>
</tbody>
</table>
Change in nucleotide sequence

Nucleotide variants within a gene can disrupt the correct process of protein synthesis at one (or more) of the key stages: transcription, RNA processing and translation.

Finally the protein (if produced) can have aberrant structure and/or function and/or stability
Variants disrupting transcription

Variants can be located in the basal promoter, distal promoter or in the 5’UTR

The example of beta thalassemia

Reduced gene transcription and thus protein synthesis (β-globin chain)

(see HbVar for correct HGVS nomenclature)
Variants disrupting RNA splicing

Most mutations that alter splicing patterns are cis-acting. They can be located in:

- core consensus sequences (5′ss, 3′ss and branch point (BP))
- regulatory elements that modulate spliceosome recruitment
  - exonic splicing enhancer (ESE)
  - exonic splicing silencer (ESS)
  - intronic splicing enhancer (ISE)
  - intronic splicing silencer (ISS)

---

Medical Genetics, Athens University

Scotti & Swanson, Nature Reviews Genetics, 2016
Variants disrupting RNA splicing

Mutations that change the canonical dinucleotides (5’ss or 3’ss) at the splice site – no protein product

The example of beta thalassemia

(see HbVar for correct HGVS nomenclature)
Variants disrupting RNA splicing

Intronic splicing enhancer (ISE)
Compete with consensus splice site sequence and reduce efficiency of splicing, resulting in a lower level of correctly spliced mRNA (and protein)

The mis-spliced mRNA is usually unstable
The example of beta thalassemia

Splicing consensus

(see HbVar for correct HGVS nomenclature)

Treisman et al, Nature. 1983
Variants disrupting RNA splicing

Activation of exonic splicing enhancers (ESE’s)

- Silent amino acid changes (e.g. Ala→Ala)
- Missense mutations (e.g. Hb E)

Causes activation of usually silent splice site sequences which competes with the correct splice site

The example of beta thalassemia

Traeger et al, Nature 1980
Pawar et al, Blood 1997
Variants disrupting RNA splicing

>90% of human protein-coding genes producing multiple mRNA isoforms through alternative splicing

![Diagram showing different types of alternative splicing events](image)
## Variants disrupting RNA splicing

Many mechanisms of aberrant splicing and resulting phenotypes

The example of the LMNA gene

<table>
<thead>
<tr>
<th>Disease</th>
<th>LMNA mutation</th>
<th>Mis-splicing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limb girdle muscular dystrophy 1B (LGMD1B)</td>
<td>Mutant 5’ss (c. 1608+5G&gt;C)</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td><img src="#" alt="LGMD1B Diagram" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intron 9 retention → LMNA RNA turnover</td>
</tr>
<tr>
<td></td>
<td></td>
<td>truncated lamin A/C</td>
</tr>
<tr>
<td>Familial partial lipodystrophy type 2 (FPLD2)</td>
<td>Mutant 5’ss (c. 1488+5G&gt;C)</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td><img src="#" alt="FPLD2 Diagram" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intron 8 retention → LMNA RNA turnover</td>
</tr>
<tr>
<td></td>
<td></td>
<td>truncated lamin A/C</td>
</tr>
<tr>
<td>Hutchinson-Gilford progeria syndrome (HGPS)</td>
<td>Alternative 5’ss (c. 1824C&gt;T)</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td><img src="#" alt="HGPS Diagram" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exon 11 150-nucleotide deletion → progerin</td>
</tr>
<tr>
<td>Dilated cardiomyopathy (DCM)</td>
<td>Alternative 3’ss (c. 640-10A&gt;G)</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td><img src="#" alt="DCM Diagram" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exon 4 5’ extension → lamin A/C+3 amino acids</td>
</tr>
</tbody>
</table>

---

**Diagram Legends:**
- **Control:** Normal splicing is indicated by thin black lines, and disease-associated splicing is indicated in dotted lines or purple.
- **PTC:** Premature termination codon.
- **RNA turnover:** RNA turnover leads to RNA degradation and cell death.
- **exon:** Exons are shown in blue, introns are shown as thick black lines, alternative splicing sites are shown in dotted lines or purple.

---

*Medical Genetics, Athens University*  
*Scotti & Swanson, Nature Reviews Genetics, 2016*
Variants affecting RNA processing

Mutations in the 3’UTR or polyA addition signal (AATAAA), causing RNA instability and resulting in a reduced level mRNA (and protein)

The example of beta thalassemia

- AATAAA > AACAAAA
- AATAAA > AATGAA
- AATAAA > AATATA
- AATAAA > AATAGA
- AATAAA > AATAAG
- AATAAA > del AATAA
- AATAAA > del AT or TA
- AATAAA > del AA

PROMOTER

5’UTR | exon1 | IVS 1 | exon2 | IVS2 | exon3 | 3’UTR

+1480C>G

(see HbVar for correct HGVS nomenclature)

Orkin et al, EMBO, 1985
Variants affecting translation

The example of beta thalassemia

Frameshift (leading to premature stop)

Initiation codon

Premature termination codon (PTC) = nonsense mutation

(see HbVar for correct HGVS nomenclature)

Orkin and Goff, J Biol Chem, 1981
Pirastu et al, J Biol Chem, 1984
Variants affecting translation

Nonsense mutations

- PTC > Premature termination of translation
- Usually leads to nonsense-mediated mRNA decay (NMD) and no protein product
Variants affecting translation

Frameshift mutations

Caused by insertion or deletion of nucleotides not divisible by three, leading to an altered reading frame from the position of the mutation to the next in-frame termination codon

NMD and sometimes an unstable protein product
Variants affecting translation

The example of alpha thalassemia

Termination codon changed to amino-acid codon
RNA translation continues until next in-frame termination codon producing aberrant protein product (unstable)

Example Hb Constant Spring Stop>Gln, leading to modified C-terminal sequence:
(142)Gln-Ala-Gly-Ala-Ser-Val-Ala-Val-Pro-Pro-Ala- Arg-Trp-Ala-Ser-Gln-Arg-Ala-Leu-Leu-Pro- Ser-Leu-His-Arg-Pro-Phe-Leu-Val-Phe-(172)Glu-COOH

TAA → CAA (Hb Constant Spring)
TAA → AAA (Hb Icaria)
TAA → GAA (Hb Seal Rock)
TAA → TAT (Hb Pakse)
TAA → TCA (Hb Koya Dora)
TAA → TTA (Hb Kinshasa)
Variants leading to altered protein - missense mutations

Missense mutation

Original DNA code for an amino acid sequence.

DNA bases

Amino acid

Replacement of a single nucleotide.

Incorrect amino acid, which may produce a malfunctioning protein.

U.S. National Library of Medicine
Nucleotide variants – predicting effect

• Missense changes are usually deleterious; synonymous changes are usually benign
• However, some missense changes are benign, whereas some synonymous variants can be deleterious

• Frameshift and truncating mutations are generally considered deleterious

But
• Truncating variants may be deleterious in a recessive disease with loss-of-function model, but irrelevant in a dominant disease caused by a gain of abnormal function
Variants in DNA sequence - methods for detecting

Targeted methods
• Reverse dot blot hybridization
• ARMS-PCR
• Restriction enzyme (RE)-PCR
• Real-time PCR (with probes)

Gene scanning methods
• Denaturing Gradient Gel Electrophoresis (DGGE)
• High resolution melting analysis (HRMA)

Generic methods
• Sanger sequencing (Automated)
• **Next Generation Sequencing**!!

Medical Genetics, Athens University
Repeat expansion variants

Instability of repetitive DNA sequences within the genome is associated with a number of human diseases. The expansion of trinucleotide repeats is recognized as a major cause of neurological and neuromuscular diseases.

Examples

<table>
<thead>
<tr>
<th>Disease</th>
<th>Repeated sequence</th>
<th>Normal range</th>
<th>Disease range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinal and bulbar muscular atrophy</td>
<td>CAG</td>
<td>11–33</td>
<td>40–62</td>
</tr>
<tr>
<td>Fragile-X syndrome</td>
<td>CGG</td>
<td>6–54</td>
<td>50–1500</td>
</tr>
<tr>
<td>Jacobsen syndrome</td>
<td>CGG</td>
<td>11</td>
<td>100–1000</td>
</tr>
<tr>
<td>Spinocerebellar ataxia (several types)</td>
<td>CAG</td>
<td>4–44</td>
<td>21–130</td>
</tr>
<tr>
<td>Autosomal dominant cerebellar ataxia</td>
<td>CAG</td>
<td>7–19</td>
<td>37–220</td>
</tr>
<tr>
<td>Myotonic dystrophy</td>
<td>CTG</td>
<td>5–37</td>
<td>44–3000</td>
</tr>
<tr>
<td>Huntington disease</td>
<td>CAG</td>
<td>9–37</td>
<td>37–121</td>
</tr>
<tr>
<td>Friedreich ataxia</td>
<td>GAA</td>
<td>6–29</td>
<td>200–900</td>
</tr>
<tr>
<td>Dentatorubral-pallidoluysian atrophy</td>
<td>CAG</td>
<td>7–25</td>
<td>49–75</td>
</tr>
<tr>
<td>Myoclonus epilepsy of the Unverricht-Lundborg type*</td>
<td>CCCCGCCCGCGCG</td>
<td>2–3</td>
<td>12–13</td>
</tr>
</tbody>
</table>
Repeat expansion variants

The repeats that are expanded can be found within the genes (introns and exons) or the immediate flanking regions.

![Diagram showing repeat expansion variants](image-url)

- (CGG) FMR-1
- (GAA) Friedreich Ataxia
- (CAG) PolyQ ataxias (HD, SCAs 1-3, 6, 7, 17)
- (ATTCT) SCA10
- (CTG) SCA8, DM1
Repeat expansion variants - methods for detecting

Southern blotting

Fluorescent PCR and fragment sizing
- Simple PCR
- Triplet primer PCR

**NGS not suitable** – current challenges
- expanded repeat does not map to reference sequence (too many mismatches)
- GC repeats have low coverage
Unusual variants and/or unusual effect of variants – caveats in variant effect prediction
Deletions causing silencing of nearby genes

Functional genes are present but loss of intergenic regulatory elements kilobases distant abolishes their expression

Example of the *HBA* gene cluster

About 50 described deletions – most *de novo*
Deletions causing silencing of nearby genes

A rare deletion causing α-thalassemia via an antisense RNA

Due to deletion, the RNA transcript from the truncated LUC7L gene expresses across the HBA2 gene and its CpG island. This mediates methylation of the HBA2 gene CpG island and silencing of α2 gene expression.

Medical Genetics, Athens University

Intragenic single nucleotide variant and gain-of-function

A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter

Marco De Gobbi,1* Vip Viprakasit,2* Jim R. Hughes,1 Chris Fisher,1 Veronica J. Buckle,1 Helena Ayyub,1 Richard J. Gibbons,1 Douglas Vernimmen,1 Yuko Yoshinaga,3 Pieter de Jong,3 Jan-Fang Cheng,4 Edward M. Rubin,4 William G. Wood,1 Don Bowden,5 Douglas R. Higgs1‡

Science, 2006
Intragenic single nucleotide variant and gain-of-function

α-Thalassemia resulting from competition for upstream regulatory elements

• Five patients from Melanesia with hemoglobinopathy H, (the severest post-natal form of alpha thalassemia) who had none of known causes of alpha thalassemia

• They all had a common α-gene haplotype, indicating a likely cis acting mutation

• Fine sequencing spanning 213kb around the alpha gene cluster found 283 SNPs

• RNA from the patient’s erythroid cells was hybridized to a tiled array of the region and the SNPs were aligned with the expression profile (and compared to a normal (αα/αα) expression profile)

Intragenic single nucleotide variant and gain-of-function

Region expressing RNA contains 17 SNPs, 10 present in nonthalassemics

Intragenic single nucleotide variant and gain-of-function

The new peak of mRNA expression (located between the ζ and ψζ genes) coincides with a single nucleotide variants that creates a binding site for an erythroid transcription factor also known as GATA-binding factor 1 (GATA-1)

\[ 5'\text{TAATAA}3' \text{ (T allele)} \text{ to } 5'\text{TGATAA}3' \text{ (C allele)} \]

It is hypothesized that the new promoter–like element between the upstream regulatory element (HS-40) and the HBA gene promoters significantly down-regulates expression of the alpha genes by successfully competing for preferential interaction with HS-40

Trans genomic-contextual variants

 Mutation-dependent recessive inheritance of NPHS2-associated steroid-resistant nephrotic syndrome

Kálmán Tory1–3, Dóra K Menyhárd4, Stéphanie Woerner2,3, Fabien Nevo2,3, Olivier Gribouval2,3, Andrea Kerti1, Pál Stráner4, Christelle Arrondel2,3, Evelyne Huynh Cong2,3, Tivadar Tulassay1,5, Géraldine Mollet2,3, András Perczel4,6 & Corinne Antignac2,3,7

Nature Genetics, 2014
Trans genomic-contextual variants

Nephrotic syndrome type 2 (MIM 600995) is an autosomal-recessive disorder caused by mutations in NPHS2 (MIM 604766 1q25-q31), which encodes podocin.

Podocin is located, in dimeric or oligomeric forms, in lipid raft microdomains at the podocyte slit diaphragm, which is the key component of the glomerular filtration barrier in the kidney.

Tory et al, Nature Genetics, 2014
Trans genomic-contextual variants

Podocin (a stomatin protein family) is a transmembrane protein with a hairpin-like intramembrane loop and intracellular N and C termini.

The C-terminal portions of podocin are responsible for dimerization.
Trans genomic-contextual variants

>150 disease-associated *NPHS2* variants have been described

Patients with nephrotic syndrome type 2 typically inherit 2 variant *NPHS2* alleles

One missense variant p.R229Q (c.686G>A) reaches polymorphic frequencies in most populations (0.03-0.13) and has an ambivalent pathogenicity:

Homozygotes for p.R229Q are asymptomatic (unaffected) whereas compound heterozygotes are sometimes but not always symptomatic

Bouchireb et al, Hum Mutat, 2013
Tory et al, Nature Genetics, 2014
Trans genomic-contextual variants

Recent studies have demonstrated that the R229Q (p.Arg229Gln) allele in fact leads to a disease phenotype only when associated with certain *NPHS2* mutations located in the C-terminal end of the protein, because of an altered heterodimerization leading to mislocalization of the Gln229 podocin and its retention within cytoplasmic compartments.

Expression of Nephrotic syndrome type 2

- p.Arg229 (wild-type)
- p.Gln229 (variant)
Cis genomic-contextual variants

LETTER

Identification of cis-suppression of human disease mutations by comparative genomics

Daniel M. Jordan¹*, Stephan G. Frangakis²*, Christelle Golzio², Christopher A. Cassa¹, Joanne Kurtzberg³, Task Force for Neonatal Genomics†, Erica E. Davis², Shamil R. Sunyaev¹§ & Nicholas Katsanis²§

Nature 2015
Cis genomic-contextual variants

- The number of pathogenic missense variants found in orthologues is surprisingly high: 5.6% ±0.5% of ClinVar variants and 6.7% ±0.4% of HumVar variants were found in the alignment of mammals.
Cis genomic-contextual variants

- Thus an allele can appear damaging in one sequence yet be neutral in an orthologous sequence of another species - a phenomenon referred to as compensated pathogenic deviation (CPD).
- A minimum estimate of CPDs in humans is 3%

Cis genomic-contextual variants

• Through a combination of bioinformatics, computational models and functional experiments Jordan et al investigated and demonstrated the structure of genetic interactions underlying CPD.
• Most CPDs involve simple pairwise interactions such that discrete *cis* amino acid residues can rescue the pathogenicity of human mutations *in vivo*.
• The findings:
  – provide an insight into the complexity of allele effect on phenotype
  – are likely to assist methods for predicting allele pathogenicity
Retrotransposition and human disease

Over 95 retrotransposition events have been reported causing single-gene disease in humans.
Large variation in characteristics of insertions of elements.
Consequences on gene expression include complete disruption or reduced synthesis of normal protein.

Both germline and somatic transposition events have been observed.
# Variants in non-coding RNAs

Genecode Version 25 (March 2016 freeze, GRCh38)  
Ensembl 86 General stats

<table>
<thead>
<tr>
<th>Genomic element</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total No of Genes</td>
<td>58037</td>
</tr>
<tr>
<td>Protein-coding genes</td>
<td>19950</td>
</tr>
<tr>
<td>Long non-coding RNA genes</td>
<td>15767</td>
</tr>
<tr>
<td>Small non-coding RNA genes</td>
<td>7258</td>
</tr>
</tbody>
</table>

Medical Genetics, Athens University
Variants in non-coding RNAs

Examples of non-coding RNA gene associated genetic disorders

<table>
<thead>
<tr>
<th>Genomic Element</th>
<th>Name</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIR</td>
<td>MIR96</td>
<td>DFNA50 (Autosomal Dominant deafness 50)</td>
</tr>
<tr>
<td></td>
<td>MIR184</td>
<td>EDICT syndrome</td>
</tr>
<tr>
<td></td>
<td>MIR17HG</td>
<td>Feingold syndrome 2</td>
</tr>
<tr>
<td>Long ncRNA</td>
<td>TERC</td>
<td>AD dyskeratosis congenita; susceptibility to aplastic anemia</td>
</tr>
<tr>
<td></td>
<td>RMRP</td>
<td>CHH (cartilage hair hypoplasia) syndrome; anauxetic dysplasia; metaphyseal dysplasia without hypotrichosis</td>
</tr>
<tr>
<td></td>
<td>CISTR-ACT IncRNA</td>
<td>Type E polydactyly</td>
</tr>
<tr>
<td></td>
<td>HELLP lincRNA</td>
<td>HELLP (Hemolysis, Elevated Liver enzymes, Low Platelets) syndrome</td>
</tr>
<tr>
<td></td>
<td>ATXN8/ATXN8OS</td>
<td>Spinocerebellar ataxia 8 (SCA8)</td>
</tr>
<tr>
<td>Small ncRNA</td>
<td>snRNA RNU4ATAC</td>
<td>Microcephalic Osteodysplastic Primordial Dwarfism, type I (MOPD I)</td>
</tr>
</tbody>
</table>

The variants observed to date include either single nucleotide variants within, or deletions of, non-coding RNA genes. The mechanism underlying disease-expression seems to be haploinsufficiency.
Variants in the genome - the challenge!

So, how can we rise to the challenge and ensure precise interpretation of the effect of genome variants?
Variant effect prediction

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

Sue Richards, PhD1, Nazneen Aziz, PhD2,16, Sherri Bale, PhD3, David Bick, MD4, Soma Das, PhD5, Julie Gastier-Foster, PhD6,7,8, Wayne W. Grody, MD, PhD9,10,11, Madhuri Hegde, PhD12, Elaine Lyon, PhD13, Elaine Spector, PhD14, Karl Voelkerding, MD15 and Heidi L. Rehm, PhD15, on behalf of the ACMG Laboratory Quality Assurance Committee

Guidelines for investigating causality of sequence variants in human disease

Thank-you