Structure and Analysis of Genetic Variation

Sven Cichon, PhD

Division of Medical Genetics, University of Basel

Institute of Neuroscience and Medicine (INM-1), Research Center Jülich

Institute of Human Genetics, University of Bonn
Genome variation visible under the microscope already....
....but it gets enormous at the submicroscopic level

**Sequence variation**
- Single nucleotide
  - Base change – substitution – point mutation
  - Insertion-deletions (“indels”)
  - SNPs – tagSNPs
- 2 bp to 1,000 bp
  - Microsatellites, minisatellites
  - Indels
  - Inversions
  - Di-, tri-, tetranucleotide repeats
  - VNTRs
- 1 kb to submicroscopic
  - Copy number variants (CNVs)
  - Segmental duplications
  - Inversions, translocations
  - CNV regions (CNVRs)
  - Microdeletions, microduplications

**Structural variation**
- Microscopic to subchromosomal
  - Segmental aneusomy
  - Chromosomal deletions – losses
  - Chromosomal insertions – gains
  - Chromosomal inversions
  - Intrachromosomal translocations
  - Chromosomal abnormality
  - Heteromorphisms
  - Fragile sites
- Whole chromosomal to whole genome
  - Interchromosomal translocations
  - Ring chromosomes, isochromosomes
  - Marker chromosomes
  - Aneuploidy
  - Aneusomy

**Molecular genetic detection**
Sequence variation

Single nucleotide
- Base change – substitution – point mutation
- Insertion-deletions (“indels”)
- SNPs – tagSNPs

Structural variation

Molecular genetic detection

1 kb to submicroscopic
- Copy number variants (CNVs)
- Segmental duplications

Microscopic to subchromosomal
- Inversions, translocations
- CNV regions (CNVRs)
- Microdeletions, microduplications

Whole chromosomal to whole genome
- Segmental aneusomy
- Chromosomal deletions – losses
- Chromosomal insertions – gains
- Chromosomal inversions
- Intrachromosomal translocations
- Chromosomal abnormality
- Heteromorphisms
- Fragile sites

- Di-, tri-, tetranucleotide repeats
- VNTRs

Scherer et al., 2007
Single nucleotide polymorphism (SNP)

C-allele: 70% frequency
C = major allele

T-allele: 30% frequency
T = minor allele
## How many SNPs in the human genome?

<table>
<thead>
<tr>
<th>minor allele frequency</th>
<th>No. of SNPs (Mio.)</th>
<th>SNP density (1 SNP/bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.0</td>
<td>290</td>
</tr>
<tr>
<td>5</td>
<td>7.1</td>
<td>450</td>
</tr>
<tr>
<td>10</td>
<td>5.3</td>
<td>600</td>
</tr>
<tr>
<td>20</td>
<td>3.3</td>
<td>960</td>
</tr>
<tr>
<td>30</td>
<td>2.0</td>
<td>1,570</td>
</tr>
<tr>
<td>40</td>
<td>0.97</td>
<td>3,280</td>
</tr>
</tbody>
</table>

Based on mutation rate and population size it can be assumed that every base pair of the human genome exists in a mutated form in at least several individuals.
Fate of new mutations

Time (generations)

Frequency

100%

0
„connected“ SNPs - „linkage disequilibrium“ (LD)
The block structure of the human genome

- Block 1
- Hot-Spot 1
- Block 2
- Hot-Spot 2
- Block 3

SNPs

Frequent Haplotypes (htSNPs indicated)
Tagging (or “tag”) SNPs in Haplotype Blocks

Common Variants (MAF > 5%)

<table>
<thead>
<tr>
<th>Common Variants</th>
<th>SNP10</th>
<th>SNP11</th>
<th>SNP12</th>
<th>SNP20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome 1</td>
<td>A A A C G C C A .. ..</td>
<td>T T C G G G T C .. ..</td>
<td>A G T C G A C C G .. ..</td>
<td></td>
</tr>
<tr>
<td>Chromosome 2</td>
<td>A A C A C G C C A .. ..</td>
<td>T T C G A G G T C .. ..</td>
<td>A G T C A A C C G .. ..</td>
<td></td>
</tr>
<tr>
<td>Chromosome 3</td>
<td>A A C A T G C C A .. ..</td>
<td>T T C G G G T C .. ..</td>
<td>A G T C A A C C G .. ..</td>
<td></td>
</tr>
<tr>
<td>Chromosome 4</td>
<td>A A C A C G C C A .. ..</td>
<td>T T C G G G T C .. ..</td>
<td>A G T C G A C C G .. ..</td>
<td></td>
</tr>
</tbody>
</table>

Haplotypes in a block

- Haplotype 1: C T C A A A A G T A C G G T T C A G G C A
- Haplotype 2: T T G A T T T G C G C A A C A C A G T A A T A
- Haplotype 3: C C C G A T C T G T G A T A C T G G T G
- Haplotype 4: T C G A T T C C G G T T C A G A C A

Tag SNP

- only 3 tag SNPs allow to identify all 4 common haplotypes in the population!

Linkage Disequilibrium

Physical Distance

European population

African population

Block

Block 1

Block 2

bp x 10^3
How many SNPs needed to cover most of the common variation?

~300,000 tag SNPs needed to cover common variation genome-wide in Europeans at 95% level

Relative power (%) vs SNP density (SNP per kb) graph showing:
- Tag SNPs
- Random SNPs
- CEU population

Relative power levels:
- 100%
- 90%
- 80%
- 70%
- 60%
- 50%
Goals:

- Define patterns of genetic variation across human genome in different populations (CEU, CHB, JPT, YRI), is being extended in phase 3
- Guide selection of SNPs efficiently to “tag” common variants
- Public release of all data (assays, genotypes)
THE HAPMAP PROJECT
Chapter and verse on human genetic variation
Tag SNPs from HapMap enable systematic candidate gene studies
And if you don’t have good candidate genes for your brain imaging phenotype? Screen the whole genome (GWAS)
And if you don’t have good candidate genes for your brain imaging phenotype? Screen the whole genome using SNP arrays.
SNP-Arrays

Scan of an individual’s DNA with an array harbouring a genome wide set of 550,000 tag SNP markers (Illumina)
2 bp to 1,000 bp
- Microsatellites, minisatellites
  → Indels
- Inversions
- Di-, tri-, tetranucleotide repeats
- VNTRs

Microscopic to subchromosomal
- Inversions, translocations
  → CNV regions (CNVRs)
- Microdeletions, microduplications
- Segmental aneusomy
- Chromosomal deletions – losses
- Chromosomal insertions – gains
- Chromosomal inversions
- Intrachromosomal translocations
- Chromosomal abnormality
  → Heteromorphisms
- Fragile sites

Whole chromosomal to whole genome
- Interchromosomal translocations
- Ring chromosomes, isochromosomes
- Marker chromosomes
  → Aneuploidy
  → Aneusomy

Scherer et al., 2007
SNP

Microsatellite

Minisatellite

tggatc atgtctta
aatcag caca caca caca cagcagag
ccggttt tagagatccaggg tttagagatccaggg cacttt
tggatcg tgtctta
aatcag caca caca caca caca caca cagcagag
ccggttt tagagatccaggg tttagagatccaggg cacttt

DNA-Strand

Microsatellites

SNPs
5-HTTLPR (serotonin-transporter-linked polymorphic region): a degenerate repeat polymorphism
Copy number variants (CNVs) - Definition

- Variable presence or absence of a DNA sequence >1000 bp

- Different types of CNVs:
  - Deletions
  - Duplications/Triplications
  - Insertions
### Copy number variants (CNVs) - Definition

#### Deletions

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>i)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iii)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Duplications

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>i)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vi)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Wain et al., 2009**
Larger CNVs (>50-100 kb) can be detected on SNP arrays.
Comparative genome hybridisation (CGH) is more sensitive

Array with reference sequence (e.g. BAC, oligo)
CNVs reported in the **Database of Genomic Variants (DGV)**: n=15,963 (freeze Nov 02, 2010)
Molecular mechanisms leading to CNV phenotype

A) gene dosage

B) gene interruption

C) gene fusion
How CNVs occur – non-allelic homologous recombination (NAHR)

LCR – Low Copy Repeat (highly similar DNA sequences)
Most imaging phenotypes will be explained by a spectrum of common and rare functional alleles.

- **Common variants with low risk**: (SNPs detected by GWAS)
  - Low frequency variants with moderate risk
    - (Sequencing needed!)
  - Rare, higher-penetrance "mutations" (e.g. CNVs)

\[
\begin{array}{c|c|c|c|c|c}
\text{Relative Risk} & \text{Risk Allele Frequency} \\
\hline
1.1 & 0.1 & & & & \\
1.5 & 1.0 & & & & \\
2.0 & 10 & & & & \\
10 & 30 & & & & \\
\end{array}
\]
New sequencing technologies may help to identify variants relevant for imaging phenotypes not detected so far. Enable identification of phenotype-relevant rare SNPs or CNVs/Indels.
Sequencing whole exomes identifies a lot of „neutral“ background variation – how to find the phenotype-relevant variants?

- ~20,000 DNA variants in/near protein coding DNA
- ~200 rare missense variants
- ~100 loss-of-function variants (~20 rare or private)

A few things are a bit more interpretable (obvious functionality), but not absolute proof in each case...

- ~1 de novo variant per exome (only ~5% LoF)
- <5% chance that an individual has a complete knockout of a single well-preserved* gene anywhere in the genome

* well-preserved = 98-99% of genes without a common LoF mutation
Functional annotation for variants is crucial: influence on gene expression (expressed quantitative trait loci – eQTL)

(a) Cis

Cis:
DNA-sequence → regulating expression of neighbouring genes

(b) Trans

Trans:
DNA-sequence → encode for protein / mRNA → regulating target gene expression

SNP
Functional annotation for variants is crucial: influence on methylation of DNA (methylation quantitative trait loci – mQTL)

Methylated and unmethylated “C” bases can be distinguished after bisulfate sequencing.
Annotation of variants that influence gene expression or methylation in human hippocampus tissue (Bonn study)

General aspects
1) Identify eQTLs- and mQTLs in hippocampus
2) compare to other tissues/studies

Phenotype-related aspects
1) Influence of eQTLs and mQTLs on phenotype
2) Functional interpretation of GWAS and sequencing results

2.8% of genes expressed in hippocampus are regulated by SNPs
3.5% of CpG islands are differentially methylated influenced by SNPs
Functional annotation for variants is crucial: ENCODE project

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*
Summary

• The genome is highly variable.
• Especially common SNP variants as well as large structural variants (CNVs) can be tested using array-based technologies.
• The field is moving to whole-genome sequencing which allows also detection of rare SNPs and small CNVs/InDels
• Functional annotation of identified genetic variants that might play a role in brain phenotypes is of great importance:
  influence on gene regulation (incl. methylation), protein function