Genotype to phenotype

09.01.14

Valborg Gudmundsdottir
Outline for the afternoon

• The human genome
• Genomic variation
• Genotypes, homozygotes and heterozygotes
  – Short exercise
• Genotyping
• Phenotypes
• Associating genotype with phenotype, genome-wide associated studies (GWAS)
• Genotype to phenotype predictions
  – Main exercise
THE HUMAN GENOME
The Human Genome Project
99.9% of nucleotide bases are identical in all people, proving that we really are all the same on the inside.

HUMAN DNA IS 98% IDENTICAL TO CHIMPANZEE DNA.
The Human Genome Project

The HGP identified the approximately 25,000 genes in human DNA.

The vast majority of DNA in the human genome, (97%) consists of non-genetic sequence with unknown function, often called “junk DNA.”
Main components of the eukaryotic genome

- **Protein-coding genes**: 25.9%
- **Introns**: 11.6%
- **Miscellaneous unique sequences**: 11.6%
- **Miscellaneous heterochromatin**: 8%
- **Simple sequence repeats**: 5%
- **Segmental duplications**: 3%
- **DNA transposons**: 2.9%
- **SINEs**: 8.3%
- **LINEs**: 20.4%
- **LTR retrotransposons**: 13.1%
- **Non-coding sequences such as introns (almost 26%)**
- **Small portion of mammalian genomes. The human genome is estimated to contain about 19,000 pseudogenes**
- **Non-functional copies of coding genes, the original meaning of the term 'junk DNA', were once thought to explain variation in genome size**
- **Non-functional copies of coding genes, the original meaning of the term 'junk DNA', were once thought to explain variation in genome size**
- **Non-functional copies of coding genes, the original meaning of the term 'junk DNA', were once thought to explain variation in genome size**
- **Non-functional copies of coding genes, the original meaning of the term 'junk DNA', were once thought to explain variation in genome size**
GENOMIC VARIATION
Genomic variation

- Single nucleotide polymorphisms (SNPs)
- Insertions/deletions
- Copy number variations (large)
- Variable (short) number tandem repeats
Single nucleotide polymorphisms (SNPs)

A single nucleotide (A,T,C,G) DNA sequence alteration

... ACGGCTAA ...
... ATGGCTAA ...

C and T are the alleles for this position

DNA is double stranded
- “C” or “T” on red strand
- “G” or “A” on green strand
Single nucleotide polymorphisms (SNPs)

- Occur in at least 1% of the population
- Most common kind of human genetic variation
- 10-30 million SNPs in the human genome
- Occur every 100-300 bases along the 3-billion-base human genome
- Evolutionary stable
Single nucleotide polymorphisms (SNPs)

- Non-coding region
- Coding region
  - Synonymous
  - Nonsynonymous
SNPs in Coding Regions – Synonymous: No Changes in Protein

DNA SNP C to G

RNA Codon CUG to CUC

Leucine to Leucine

No change in shape
SNPs in Coding Regions – Nonsynonymous: Changes in Protein

- DNA SNP: A to C
- RNA Codon: GAU to GAG
- Protein: Aspartic acid to Glutamic acid
- Slight change in shape
dbSNP database

- rs numbers
- chromosome and positions
- Strand orientation
Ensembl database

rs1801282 SNP

Variation displays

- Explore this variation
- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Individual genotypes
- Linkage disequilibrium
- Phenotype data
- Phylogenetic context
- Citations
- External Data
- LOVD

Original source

Variants (including SNPs and indels) imported from dbSNP (release 138) | View in dbSNP

Alleles

C/G | Ancestral: C | Ambiguity code: S | MAF: 0.07 (G)

Location

Chromosome 3:12393125 (forward strand) | View in location tab

Co-located

with HGMD-PUBLIC CM981614

Most severe consequence

Missense variant | See all predicted consequences (Genes and regulation)

Evidence status

Clinical significance

Synonyms

This variation has 7 synonyms - click the plus to show

HGVS names

This variation has 17 HGVS names - click the plus to show

Genotyping names

This variation has assays on 7 chips - click the plus to show

Explore this variation

- Genomic context
- Genes and regulation
- Population genetics
- Individual genotypes
- Linkage disequilibrium
- Phenotype data
- Citations
- Phylogenetic context
GENOTYPES
One copy of rs17822931 from the father and one copy from the mother

Copy 1:  
5’...GGCCTGAGT...3’ (+) 
3’...CCGGGACTCA...5’ (-) 

Copy 2:  
5’...GGCCCGGAGT...3’ (+) 
3’...CCGGGCTCA...5’ (-) 

Genotype for rs17822931 on plus strand: 

\[
\begin{align*}
&\text{T;C (T;C)} \\
&\text{(C;T) } CT \quad C;T \quad T,C \quad (T,C) \\
&\text{rs17822931(C;T)} \quad \text{rs17822931(T;C)} \quad \text{TC}
\end{align*}
\]

Genotype for rs17822931 on minus strand: 

\[
\begin{align*}
&A;G (A;G) \\
&(G;A) \quad GA \quad G;A \quad C,T \quad A,G \quad (A,G) \\
&\text{rs17822931(G;A)} \quad \text{rs17822931(A;G)} \quad \text{AG}
\end{align*}
\]
**exercise**

<table>
<thead>
<tr>
<th></th>
<th>rs4788084</th>
<th>rs17822931</th>
<th>rs73546424</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy 1:</td>
<td>5′...TCCCCCTGGG...GGCCCTGAGT...TGCA<strong>T</strong>GTGA... 3′ (+)</td>
<td>3′...AGGGGACCC...CCGGACTCA...ACGT<strong>T</strong>ACACT... 5′ (−)</td>
<td></td>
</tr>
<tr>
<td>Copy 2:</td>
<td>5′...TCCCCCTGGG...GGCCCGAGT...TGCA<strong>T</strong>GTGA... 3′ (+)</td>
<td>3′...AGGGGACCC...CCGGGCTCA...ACGT<strong>T</strong>CACT... 5′ (−)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>rs4788084</th>
<th>rs17822931</th>
<th>rs73546424</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype on “plus strand”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>genotype on “minus strand”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>genotype on “dbSNP strand”</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HOMOZYGOTES AND HETEROZYGOTES
**Homozygous:**
Genotype consisting of two identical alleles at a given locus
(For a SNP: the same base at both copies, eg. C;C or A;A)

**Heterozygous:**
Genotype consisting of two different alleles at a locus
(For a SNP: the same base at both copies, eg. A;C)
exerc**ise**

rs4788084  rs17822931  rs73546424

Copy 1:

5′...TCCCTCTGGG...GGCCTGAGT...TGCA**T**GTGA... 3′ (+)

3′...AGGGGACCC...CCGGACTCA...ACGTACACT... 5′ (-)

Copy 2:

5′...TCCCTCTGGG...GGCCCGAGT...TGCA**A**GTGA... 3′ (+)

3′...AGGGGACCC...CCGGGCTCA...ACGT**T**CACT... 5′ (-)

**heterozygous** or homozygous?

**heterozygous** or homozygous?

**heterozygous** or homozygous?
GENOTYPING
SNP arrays

- Microarray technique
- 0.3 – 4.3 million SNPs
Genomic enlightenment
Medicinsk Museion
Next Generation Sequencing

• Different parts of the genome can be sequenced:
  ✷ Whole genome
  ✷ Exome
  ✷ Targeted

• Different methods for different platforms

• Is increasing in popularity due to increasingly lower costs
PHENOTYPES
An observable characteristics or trait

obesity

eye color

height
Connecting genotypes and phenotypes:

GENOME-WIDE ASSOCIATION STUDIES (GWAS)
Monogenic and polygenic traits

- Some traits are determined by a single gene, where a one mutation can cause a disease
  - Often called Mendelian diseases
  - Examples are Huntington’s disease and sickle cell anemia
- Most common traits and diseases are caused by a large number of genes
  - Often called complex traits/diseases
  - Examples are human height, obesity, type 2 diabetes and cardiovascular disease
- GWAS studies usually focus on complex polygenic traits
GWAS

Association of common variants (SNPs) across the whole genome with a particular phenotype

Science 2007:

BREAKTHROUGH OF THE YEAR

Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another.
Cases vs controls

- Obtain DNA from a disease group (e.g. asthma) and a control group
- Obtain genotypes
- Identify variants that are significantly more common among cases than controls
- Those SNPs are associated with the disease (in this study)
- Not necessarily causal
Example of GWAS results (BMI)
Manhattan plot displays all SNPs on x-axis (ordered by chromosome) and $-\log_{10}$ of p-values on y-axis

Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index

Speliotes et al, Nature Genetics, 2010
Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).

NHGRI catalog
As of 07/01/14, the catalog includes 1782 publications and 12151 SNPs.
SNPedia.com

SNPedia

SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by Promethease to analyze and help explain your DNA.

Help! [edit]

- look at the example rs1234
- learn more about SNPs
- browse
  - genes
  - genomes
  - genosets
  - genotypes
  - medicines
  - medical conditions
Challenges of GWAS

- Missing heritability
- Small effect sizes (OR < 1.5)
- Not much translation into clinical practice
- Biological role of variants unclear, majority (93%) outside of coding regions

Five Years of GWAS Discovery

Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang
For most common diseases, the sum of individual effects found so far is much less than the total estimated heritability.
GENOTYPE TO PHENOTYPE PREDICTIONS
It is difficult, perhaps impossible, to accurately predict complex traits from the information we have today!

- Clinical setting
  - High penetrance, often rare mutations, such as \textit{BRCA1}
- Commercial genotyping
  - 23andMe
- Ancient genomes
Discover your ancestral origins and lineage with a personalized analysis of your DNA.

- Learn what percent of your DNA is from populations around the world.
- Contact relatives across continents or across the street.
- Build your family tree and enhance your experience with relatives.

$99

order now
Genotype and phenotype of an ancient genome
First Ancient Human Genome
The Saqqaq Genome Project

4,000 years
Hair sample from permafrost
DNA extraction <10% contamination
20 x coverage
Started 2009

Eske Willerslev
DNA from hair
Rasmussen et al. have sequenced the genome of a man from the Saqqiq culture, using DNA from hair preserved in permafrost in Greenland. They analysed the genome to find single nucleotide polymorphisms (SNPs) — differences in single DNA base pairs that exist between individual genomes, and that may act as markers of an individual’s physical traits. 

a. Here, a short stretch of human DNA is shown that is a marker for normal earwax.

b. In the analogous DNA from the Saqqiq individual, there is a SNP in which a C in the lower strand has been replaced by a T (C, G, T and A denote the four kinds of DNA base). This SNP shows that the Saqqiq man had dry earwax. Rasmussen and colleagues identified other SNPs indicating that the ancient human had, among other things, brown eyes, non-white skin, thick dark hair and an increased susceptibility to baldness.

V. WARD, UNIV. AUCKLAND
What can we say about his phenotypes?
From genotype to phenotype: how good are we at putting a face to an anonymous individual? While some traits manifest themselves in a tissue specific manner (highlighted in green), others are more systemic (highlighted in blue). Going from the genetic blueprint to visual appearance, physiological behaviour and medical predispositions is still an open challenge.
The Saqqaq Genome Database (NCBI36)

Enter sequence range, identifier or cheat code

Examples
- Range: 17:398382..399882 (chromosome:start..end)
- SNP ID: rs17822931 or ENSSNP22423 - Ambiguously mapped SNPs and in-dels may return several records.
- List phenotypic associations on chromosome: 1:phenotype

Note: Query is currently limited to 100000 records/nucleotides

[ home | download flat files | ancientgenome.dk ]
## The Saqqaq Genome Database (NCBI36)

### Result

<table>
<thead>
<tr>
<th>chr</th>
<th>pos</th>
<th>ref</th>
<th>is_ref</th>
<th>genotype</th>
<th>pp</th>
<th>depth</th>
<th>repeat</th>
<th>rs</th>
<th>type</th>
<th>strand</th>
<th>snp_alleles</th>
<th>trait</th>
<th>risk_allele</th>
<th>pmid</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>12368125</td>
<td>C</td>
<td>n</td>
<td>GG</td>
<td>0.0006439</td>
<td>11</td>
<td></td>
<td>rs1801282</td>
<td>single</td>
<td>+</td>
<td>CG</td>
<td>obesity, Obesity, association with Disease-associated and putatively functional polymorphism, Type 2 diabetes</td>
<td>C</td>
<td>9732554</td>
<td>NHGRI,other</td>
</tr>
</tbody>
</table>

### Columns explained

- **chr**: The chromosome
- **pos**: Position on chromosome
- **ref**: The reference nucleotide on forward strand in hg18
- **is_ref**: Indicates whether the genotype is the same as the reference nucleotide (y) or not (n)
- **genotype**: The genotype called (on forward strand)
- **pp**: For numerical reasons, we report (1-PP), where PP is the posterior probability of the genotype
- **depth**: The number of reads covering the position
- **repeat**: If the position lies in an annotated repeat, the ID is given here
- **rs**: dbSNP rs-number
- **type**: Type of dbSNP entry ("single", "indel" etc. - see the UCSC genome browser for details).
- **strand**: Strand for dbSNP entry + (or 1) or - (or -1)
- **snp_alleles**: Known SNP alleles, e.g. "AC" (or "A/C") for a SNP of type "single"
- **trait**: Associated trait or phenotype (general information – not dependant on this individual’s genotype)
- **risk_allele**: Allele associated with trait
- **pmid**: Putum ID of GWAS paper reporting the association
- **source**: Source database for association
EXERCISE
Genotype to phenotype exercise

http://wiki.cbs.dtu.dk/teachingmaterials/index.php/ExGenotype2PhenotypeLite