Human genetics in the 21st century:

Using bioinformatics to link genotype and phenotype

High School Student Program 2015

Bioinformatics and Research Computing
Whitehead Institute

http://barc.wi.mit.edu

Whitehead Institute

Genetic engineering

- Before genetic engineering is really useful, we might want to know more about
  - Gene function and regulation
  - Associations between genetic markers and physical traits
- Bioinformatics tries to address both topics

Jughead, May 1997

Bioinformatics & Research Computing

Consultation and collaboration, training and education, and software in the areas of Bioinformatics and Graphics.

at Whitehead Institute

BaRC Bioinformatics and Research Computing

George Bell, Bingbing Yuan, Prat Thiru, Yanmei Huang, Inma Barrasa — *Bioinformatics*

Tom DiCesare — *Graphics*

150 years ago (1865)

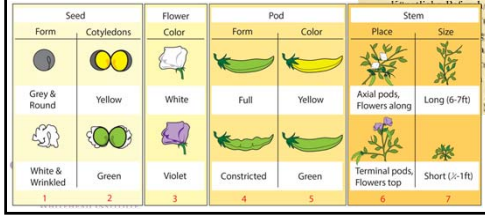
- Mendel presented his "Experiments on Plant Hybridization"

Led to the **particulate theory of inheritance**

Versuche über Pflanzen-Hybriden.

Gregor Mendel

Einleitende Bemerkungen.



Selected discoveries since Mendel

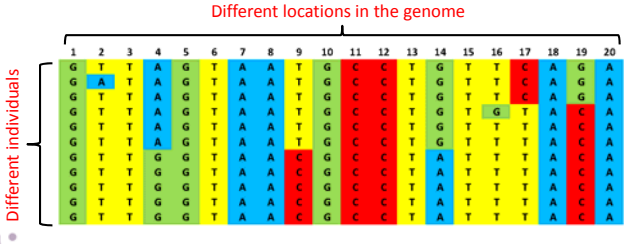
- 1950s
  - DNA is the genetic material
  - The structure of DNA
- 1960s
  - The genetic code (DNA => protein)
- 1970s
  - DNA sequencing
- 1990s-2000s
  - Genome sequencing

Linking genotype and phenotype using genomics

- To do this, we need only
  - Genome sequence (or a sample of it) for many individuals
  - Selected phenotype(s) for the same individuals
  - [Some complex statistics]

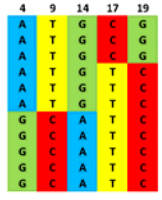
Sampling genome sequence

- Most of the genome is identical between individuals
- Let's concentrate only on the places that are the most different



Single nucleotide polymorphisms

- SNPs (pronounced "snips") because
  - Single: were looking at just one genome position
  - Nucleotide: DNA letter differs
  - Polymorphism: variation occurring commonly in a population (in at least 1% of individuals)
- SNPs can be within a gene or between genes



But humans (like peas) are diploid

- We have 2 genomes, with 2 copies of each chromosome
- Each SNP can be
  - Homozygous (ex: CC), or
  - Heterozygous (ex: TG)

Different individuals

SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC

Nearby SNPs are associated

- Nearby SNPs tend to stay together during meiosis
- As a result, they tend to be genetically linked
- One "tag SNP" can be used to represent a set of linked SNPs

Genotype block "A" (linked SNPs) | Genotype block "B"

Different individuals

SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC

Combining phenotype with genotype

- Genotype: use all SNPs or a subset of tag SNPs
- Phenotype: whatever traits we want to study (as long as we think they at least partly genetic)

Genotype block "A" | Genotype block "B"

Different individuals

SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	Height	Lactose
AA	TT	GG	CC	GG	tall	can digest
AA	TT	GG	CC	GG	short	can digest
AA	TT	GG	CC	GG	short	intolerant
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	short	can digest
AG	TC	GT	CT	GC	tall	intolerant
AG	TC	GT	CT	GC	short	intolerant
AG	TC	GT	CT	GC	tall	intolerant
AG	TC	GT	CT	GC	short	intolerant
AG	TC	GT	CT	GC	tall	intolerant

Genome-wide association study (GWAS)

- Compare every marker (SNP, tag SNP, etc.) to every trait
- Is the trait associated (correlated) with any marker?

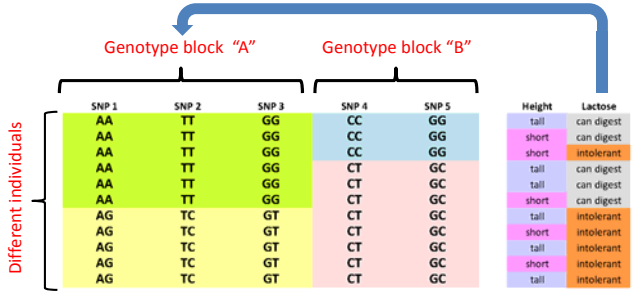
Genotype block "A" | Genotype block "B"

Different individuals

SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	Height	Lactose
AA	TT	GG	CC	GG	tall	can digest
AA	TT	GG	CC	GG	short	can digest
AA	TT	GG	CC	GG	short	intolerant
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	short	can digest
AG	TC	GT	CT	GC	tall	intolerant
AG	TC	GT	CT	GC	short	intolerant
AG	TC	GT	CT	GC	tall	intolerant
AG	TC	GT	CT	GC	short	intolerant
AG	TC	GT	CT	GC	tall	intolerant

Interpreting GWAS

- What does it mean that a trait is associated with a genetic marker?
- Does the marker cause the trait? More research needed….
- Multiple markers can be associated with the same trait

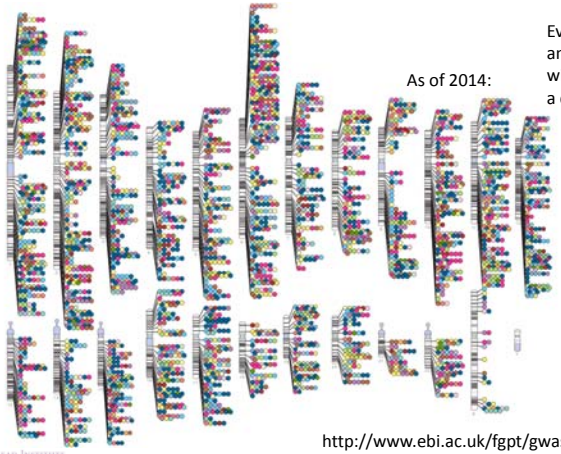


SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	Height	Lactose
AA	TT	GG	CC	GG	tall	can digest
AA	TT	GG	CC	GG	short	can digest
AA	TT	GG	CC	GG	short	intolerant
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	short	can digest
AG	TC	GT	CT	GC	tall	intolerant
AG	TC	GT	CT	GC	short	intolerant
AG	TC	GT	CT	GC	tall	intolerant
AG	TC	GT	CT	GC	short	intolerant
AG	TC	GT	CT	GC	tall	intolerant

Exercise 1

- Perform a small-scale GWAS analysis

The GWAS catalog keeps growing

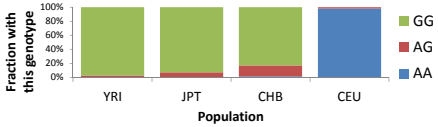


Resources for genotype-phenotype interactions

- Online Mendelian Inheritance in Man http://omim.org
- GWAS catalog http://www.genome.gov/gwastudies/
- PheGenI http://www.ncbi.nlm.nih.gov/gap/phegeni/
- ClinVar http://www.ncbi.nlm.nih.gov/clinvar/
- SNPedia http://www.snpedia.com

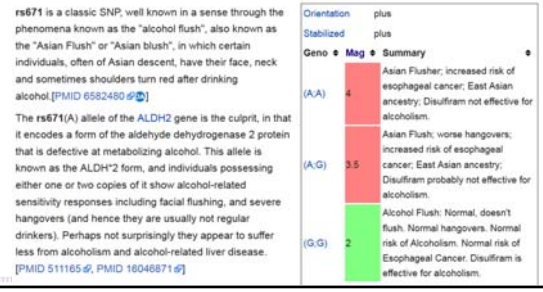
Taking ethnicity into account

- Genotypes have been collected from large-scale projects like
  - HapMap http://hapmap.ncbi.nlm.nih.gov
  - 1000 Genomes http://www.1000genomes.org
- These populations ("ethnic groups") include
  - Yoruba in Ibadan, Nigeria ("YRI")
  - Japanese in Tokyo, Japan ("JPT")
  - Han Chinese in Beijing, China ("CHB")
  - Utah residents with ancestry from northern and western Europe ("CEU")
- Sample HapMap data for SNP rs1834640



Exploring genotype-phenotype associations

- Detecting an association is only the first step.
- One typically also wants to associate a trait with the choices of genotypes at that SNP.
- SNPedia often reports these, such as for rs671



rs671 is a classic SNP, well known in a sense through the phenomena known as the "alcohol flush", also known as the "Asian Flush" or "Asian blush", in which certain individuals, often of Asian descent, have their face, neck and sometimes shoulders turn red after drinking alcohol.[PMID 6582480]

The **rs671**(A) allele of the ALDH2 gene is the culprit, in that it encodes a form of the aldehyde dehydrogenase 2 protein that is defective at metabolizing alcohol. This allele is known as the ALDH*2 form, and individuals possessing either one or two copies of it show alcohol-related sensitivity responses including facial flushing, and severe hangovers (and hence they are usually not regular drinkers). Perhaps not surprisingly they appear to suffer less from alcoholism and alcohol-related liver disease. [PMID 511165, PMID 16046871]

Orientation plus
Stabilized plus

Geno	Mag	Summary
(A;A)	4	Asian Flusher; increased risk of esophageal cancer; East Asian ancestry; Disulfiram not effective for alcoholism.
(A;G)	3.5	Asian Flush; worse hangovers; increased risk of esophageal cancer; East Asian ancestry; Disulfiram probably not effective for alcoholism.
(G;G)	2	Alcohol Flush: Normal, doesn't flush. Normal hangovers. Normal risk of Alcoholism. Normal risk of Esophageal Cancer. Disulfiram is effective for alcoholism.

Interpreting associations

- An association means that the genotype of a marker (like a SNP) can help predict the presence of a trait, BUT
  - The SNP itself might have no effect – it may just be linked to another DNA element that has the effect
  - The effect may be very small (but still > 0)
  - The association may be present only in certain individuals
  - The association may be one of many for this trait
  - it may only appear to be present (until other independent studies can verify it)
- What experiment(s) could you design to verify that a SNP causes some effect?

Exercise 2

- Link the genotype of your "alter ego" to potential traits

Summary

- One's genotype can reveal hints about one's
  - physical (and even mental) characteristics
  - medical and disease risks
- The genetic contribution to a phenotype can be linked to many genes (unlike Mendel's pea traits)
- Most genotype-phenotype associations require a lot more research to be useful
- In the future, (how) will doctors (and us) use this information?