Human genetics in the 21st century:

Using bioinformatics to link genetics and traits

Get a clue: CSI and the science of forensics (2016)

Bioinformatics and Research Computing
Whitehead Institute

http://jura.wi.mit.edu/bio/

Whitehead Institute

Recent DNA evidence

50 Years Later, a Break in a Boston Strangler Case

By JESS BIDGOOD
Published: July 11, 2013

BOSTON — Investigators said Thursday that they had linked the man believed by many to have been the Boston Strangler to DNA found in the home of a woman thought to be the Strangler's last victim in a string of unsolved murders that petrified this city in the early 1960s and has perplexed it ever since.

But early attempts to recover usable DNA samples were inconclusive. So Mr. Hayes stored the samples and waited for technology to improve. Last fall, investigators sent the samples to two private labs for another try — and this time got a DNA profile for an unknown male.

"The evidence in this case never changed, but the scientific ability to use that evidence has surpassed every hope and expectation of investigators who were first assigned to the case," Mr. Conley said.

The New York Times

DNA analysis of other “animals”

Yeti:
- Extinct(?) polar bear

Bigfoot:
- Cow
- Horse
- Wolf/coyote/dog
- Deer
- Sheep
- Raccoon
- Porcupine
- Human

Genetic analysis of hair samples attributed to yeti, bigfoot and other anomalous primates

Bryan C. Sykes[1], Rhettman A. Mullis[2], Christophe Hagenmuller[3], Terry W. Melton[4] and Michel Sartori[5,6]

[1]Institute of Human Genetics, Wolfson College, University of Oxford, Oxford OX2 6UD, UK
[2]PO Box 40143, Bellevue, WA 98005, USA
[3]NaturAlpes, Annecy-Le-Vieux 74940, France
[4]Mitotyping Technologies, 2565 Park Center Boulevard, State College, PA 16801, USA
[5]Museum of Zoology, Palais de Rumine, Lausanne 1014, Switzerland
[6]Museum of Zoology and Grindel Biocentre, Hamburg 20146, Germany

In the first ever systematic genetic survey, we have used rigorous decontamination followed by mitochondrial 12S RNA sequencing to identify the species origin of 30 hair samples attributed to anomalous primates. Two Himalayan samples, one from Ladakh, India, the other from Bhutan, had their closest genetic affinity with a Palaeolithic polar bear, *Ursus maritimus*. Otherwise the hairs were from a range of known extant mammals.

Linking genes and traits

- To do this, we need to know something about
  - Associations between genetic markers and physical traits
  - Gene function and regulation
- Bioinformatics tries to address both topics

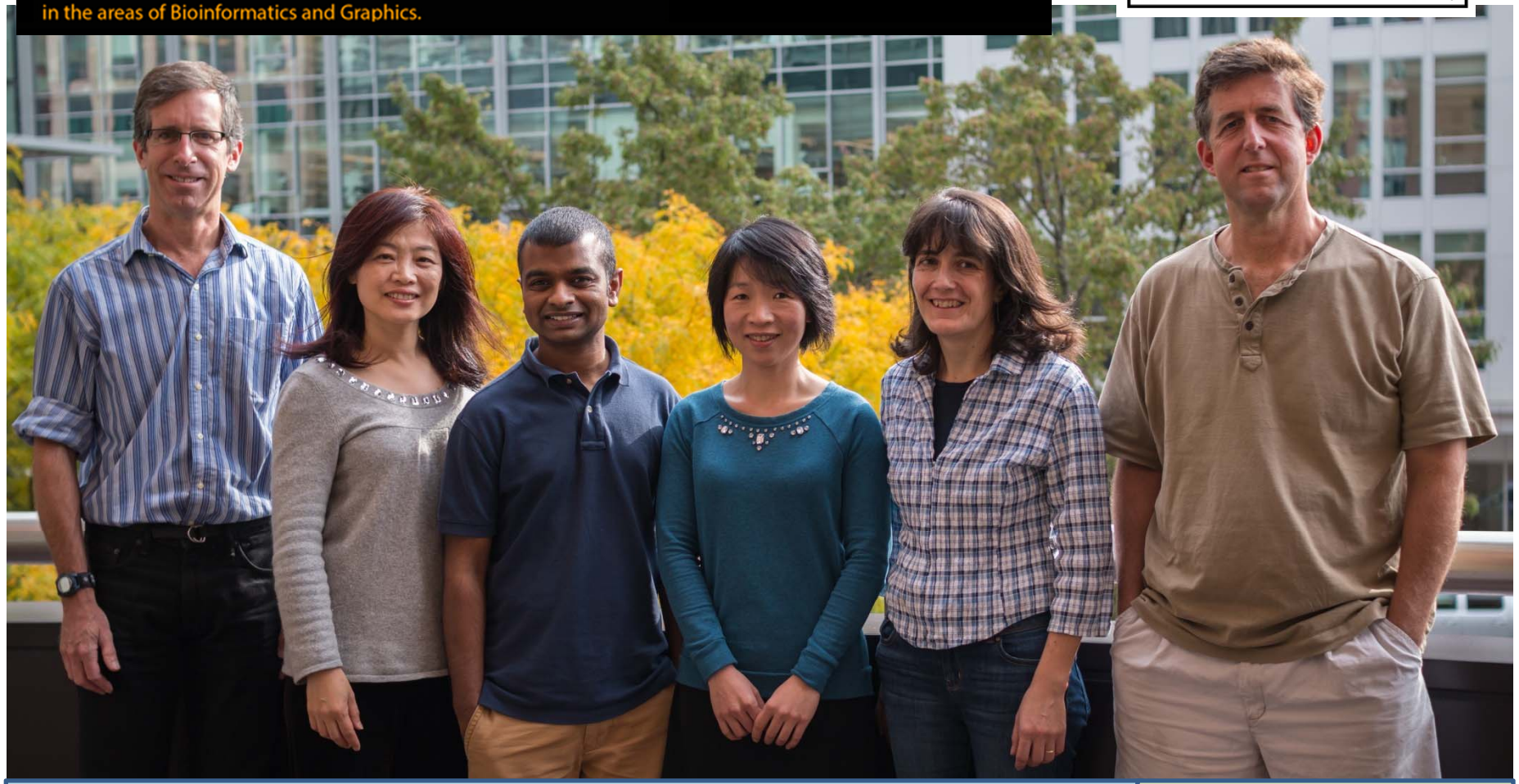
Jughead, May 1997

Bioinformatics & Research Computing

at Whitehead Institute

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

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.

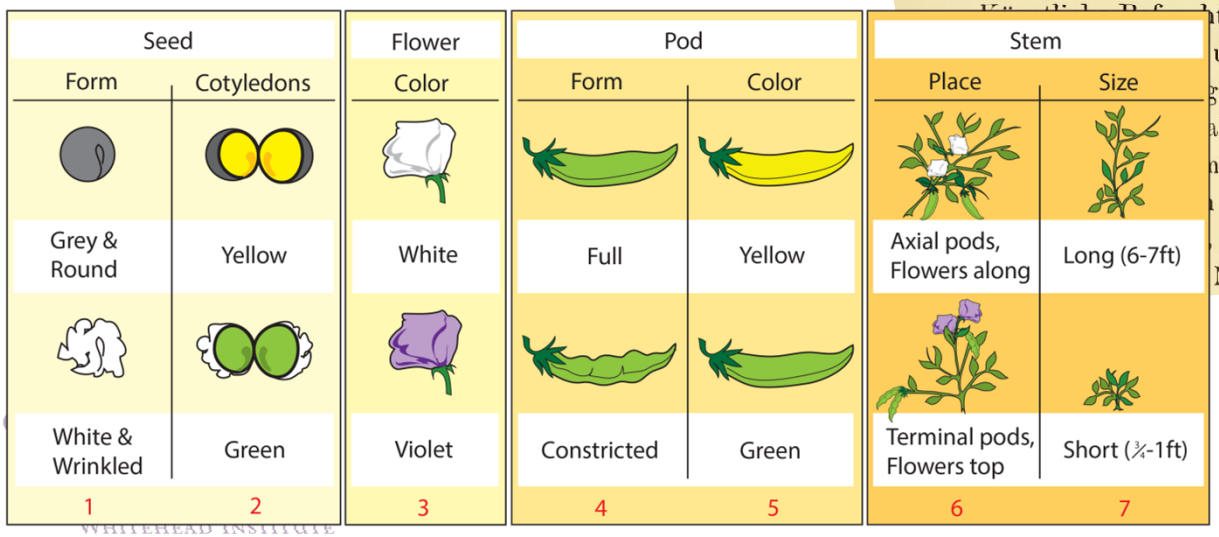
Von

Gregor Mendel.

(Vorgelegt in den Sitzungen vom 8. Februar und 8. März 1865.)

Einleitende Bemerkungen.

tungen, welche an Zierpflanzen desshalb
um neue Farben-Varianten zu erzielen,
zu den Versuchen, die her besprochen
llende Regelmässigkeit, mit welcher die-
mer wiederkehrten, so oft die Befruch-
Arten geschah, gab die Anregung zu
deren Aufgabe es war, die Entwicklung
Nachkommen zu verfolgen.



Seed		Flower	Pod		Stem	
Form	Cotyledons	Color	Form	Color	Place	Size
Grey & Round	Yellow	White	Full	Yellow	Axial pods, Flowers along	Long (6-7ft)
White & Wrinkled	Green	Violet	Constricted	Green	Terminal pods, Flowers top	Short (¾-1ft)
1	2	3	4	5	6	7

Selected discoveries since Mendel

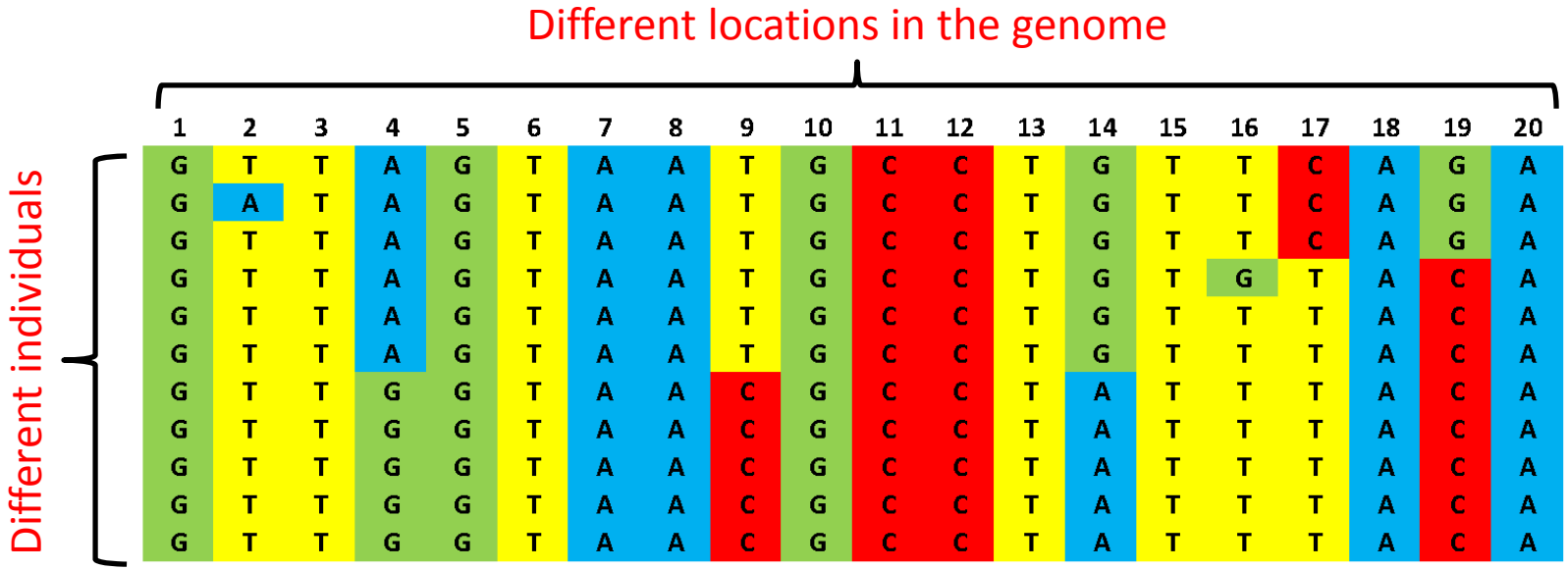
- 1950s
  - DNA is the genetic material
  - The structure of DNA
- 1960s
  - The genetic code (DNA => protein)
- 1970s
  - DNA sequencing
- 1990s-2000s
  - Genome sequencing

Linking genetics and traits using genomics

- To do this, we need only
  - Genome sequence (or a sample of it) for many individuals
  - Selected trait(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 in a gene or between genes

4	9	14	17	19
A	T	G	C	G
A	T	G	C	G
A	T	G	C	G
A	T	G	T	C
A	T	G	T	C
A	T	G	T	C
G	C	A	T	C
G	C	A	T	C
G	C	A	T	C
G	C	A	T	C
G	C	A	T	C

But humans (like peas) are diploid

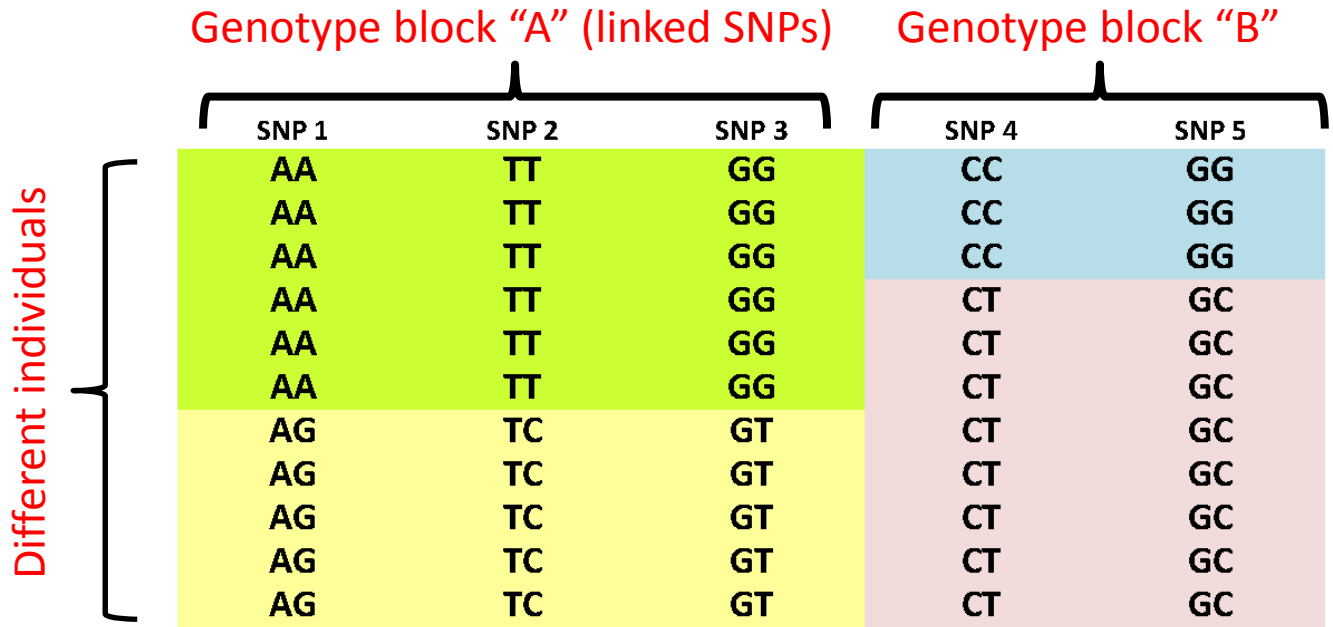
- We have 2 genomes, with 2 copies of each chromosome
- Each SNP can be
  - Homozygous (ex: CC), or
  - Heterozygous (ex: TG)

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

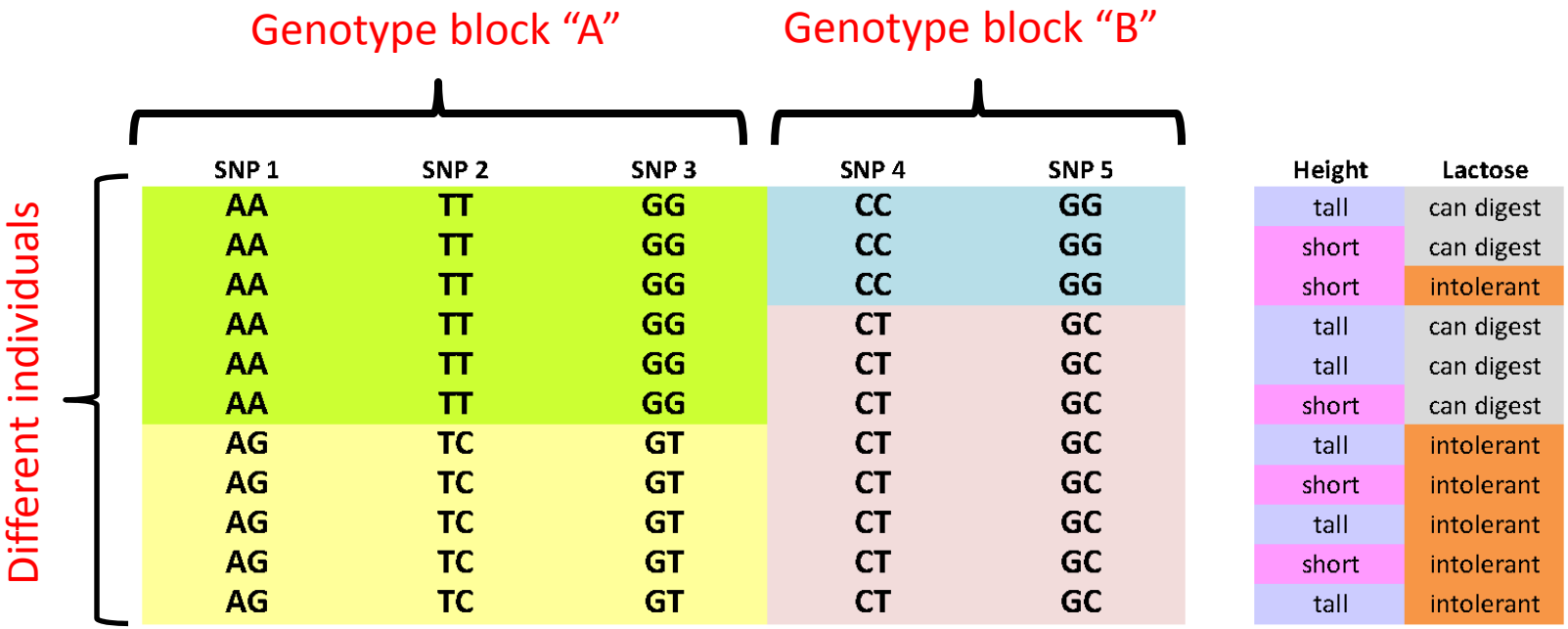
Nearby SNPs are associated

- 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



Combining phenotype with genotype

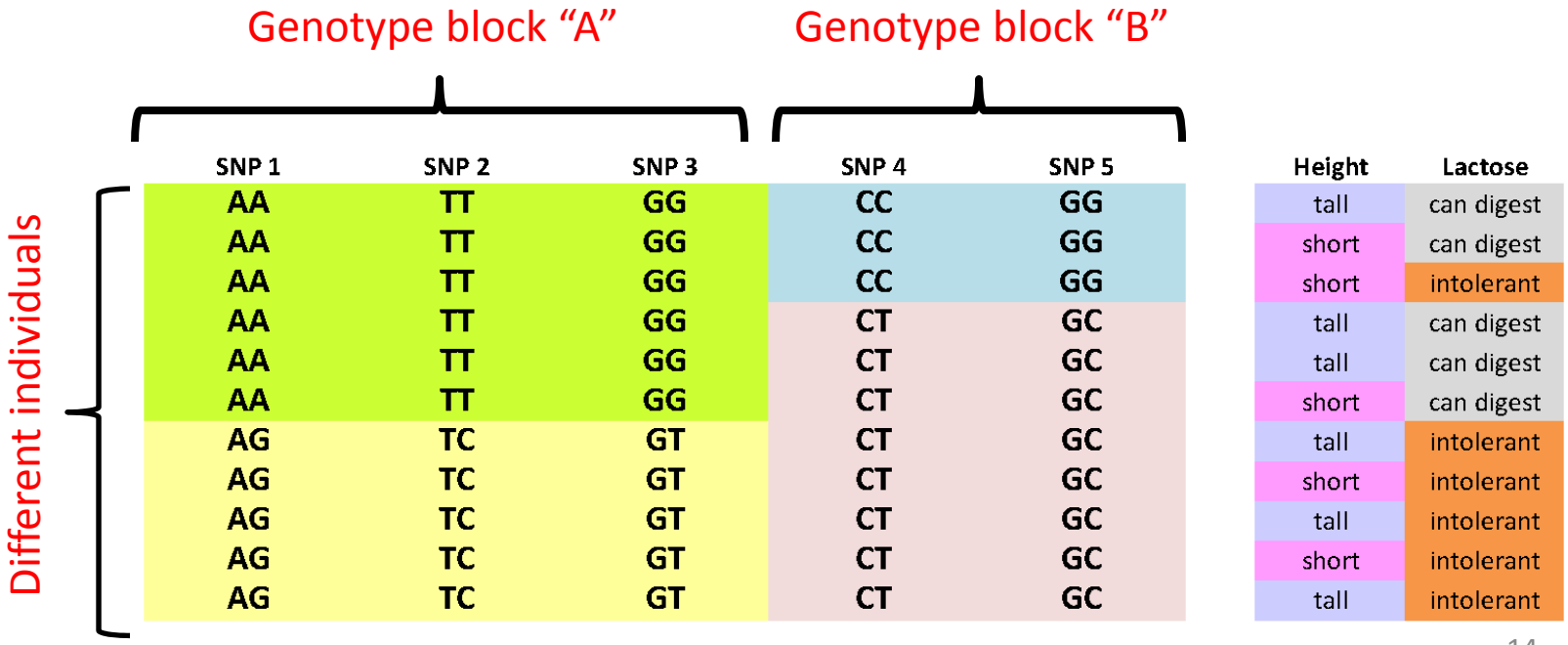
- Genotype: use all SNPs or a subset of tag SNPs
- Phenotype: whatever trait(s) we want to study (as long as we think they at least partly genetic)



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)

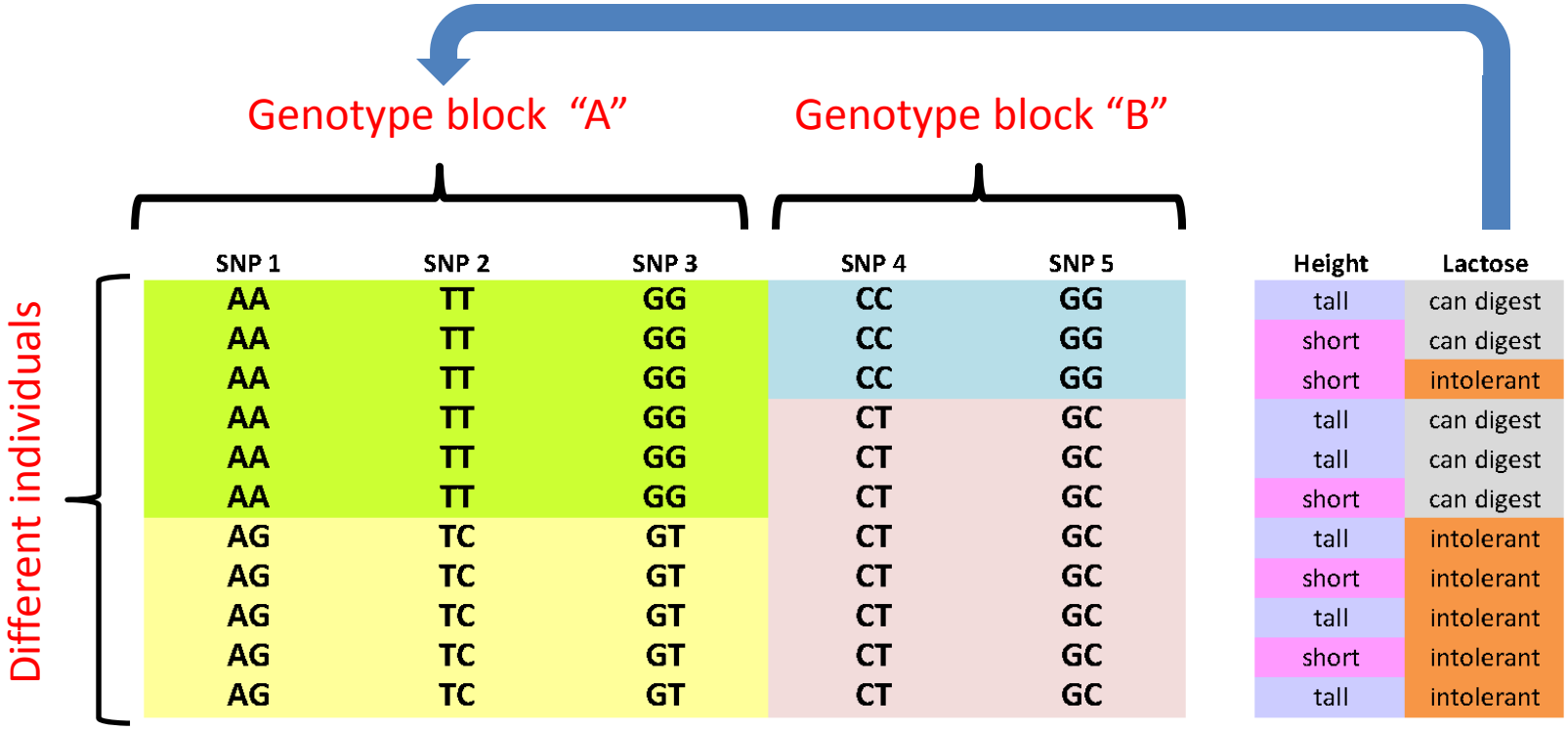
- Compare every marker (SNP, tag SNP, etc.) to every trait
- Is the trait associated (correlated) with any marker?



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

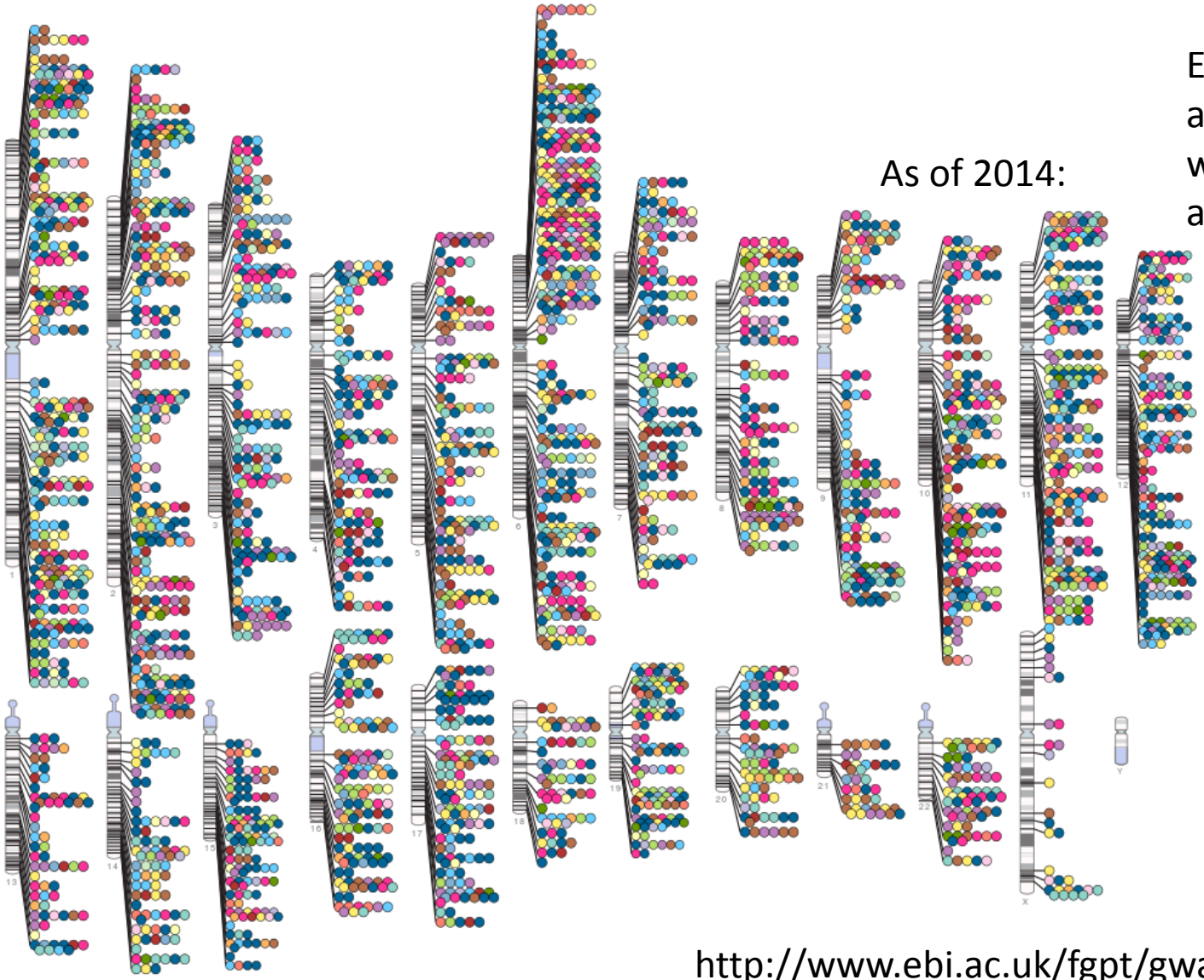


SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	Height	Lactose
AA	TT	GG	CC	GG	tall	can digest
AA	TT	GG	CC	GG	short	can digest
AA	TT	GG	CC	GG	short	intolerant
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	tall	can digest
AA	TT	GG	CT	GC	short	can digest
AG	TC	GT	CT	GC	tall	intolerant
AG	TC	GT	CT	GC	short	intolerant
AG	TC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



As of 2014:

Every point is an association with a trait or a disease

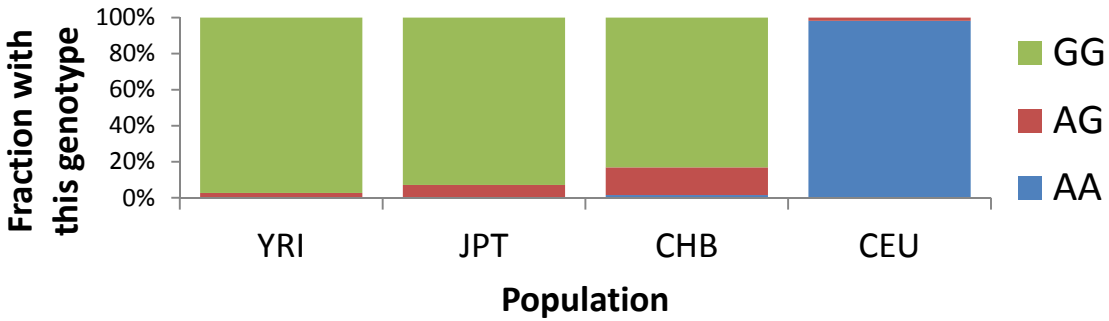
http://www.ebi.ac.uk/fgpt/gwas/

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 projects include many populations ("ethnic groups")
- Sample HapMap data for SNP rs1834640



- What's usual for one population may be very unusual for another population

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
  - 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 DNA sample to potential traits

Note that this application of genetic association is not currently applied to forensics.

Summary

- One's DNA can reveal hints about one's
  - physical (and even mental) characteristics => forensics
  - medical and disease risks => medicine
  - family history => genealogy

- A trait can be influenced by many genes (unlike Mendel's pea traits)

- In the future, how will scientists make this information even more useful?