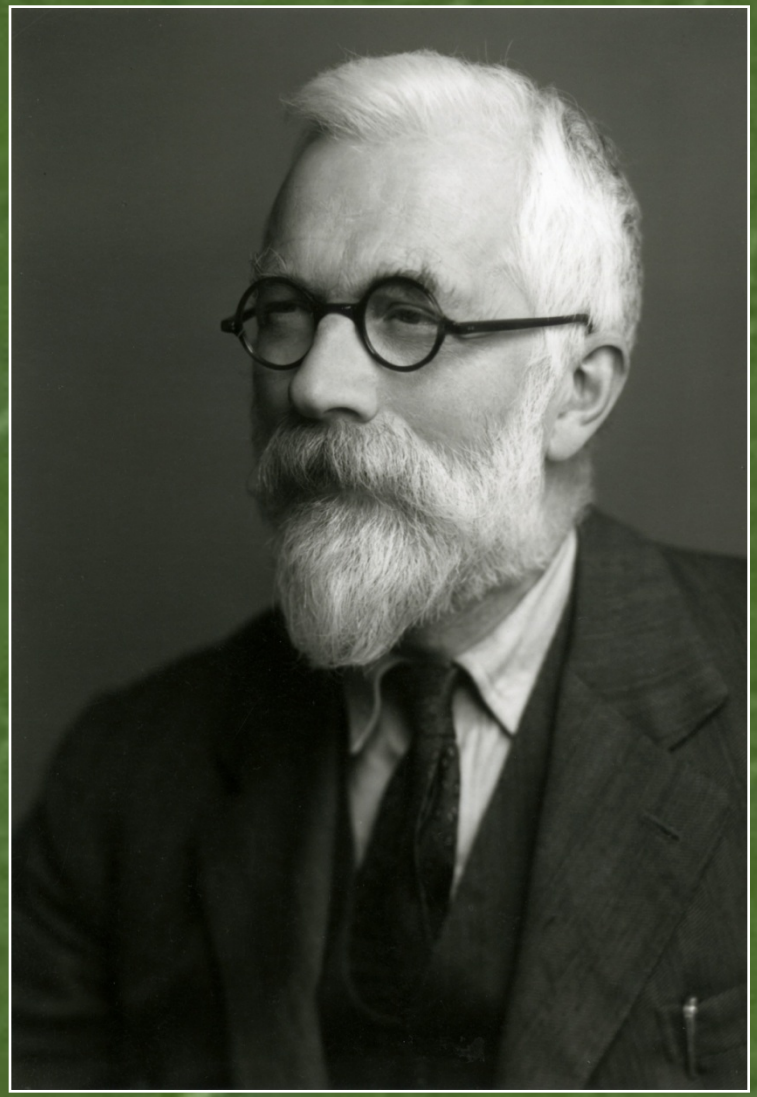
Describing DNA sequence variation in ways that help us understand its causes

Take-home:

The best measures depend on the kinds of questions we're asking, and on the kind and SCALE of the data.

New measures and methods are still being invented, to capture patterns in huge genome-scale data sets.



The most-studied molecular polymorphism: alcohol dehydrogenase (*Adh*) in *Drosophila melanogaster*

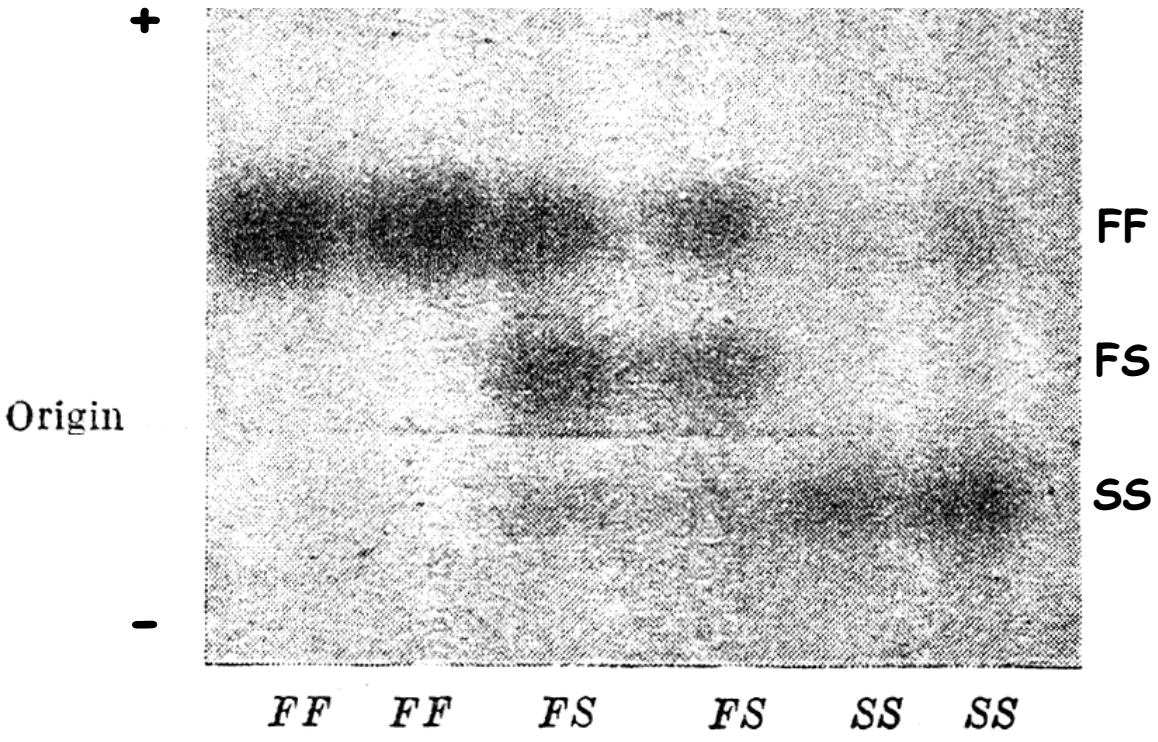
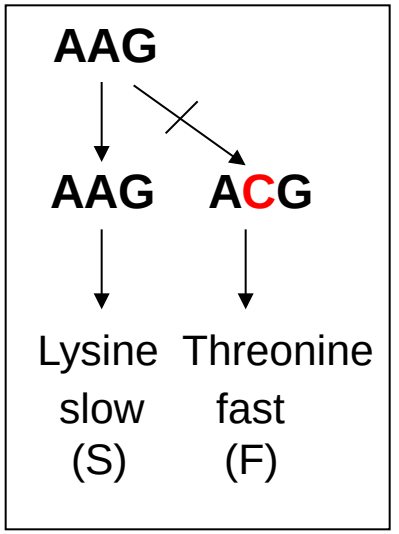


Fig. 2. Photograph of starch gel showing fast (*FF*), heterozygous (*FS*), and slow (*SS*) alcohol dehydrogenase

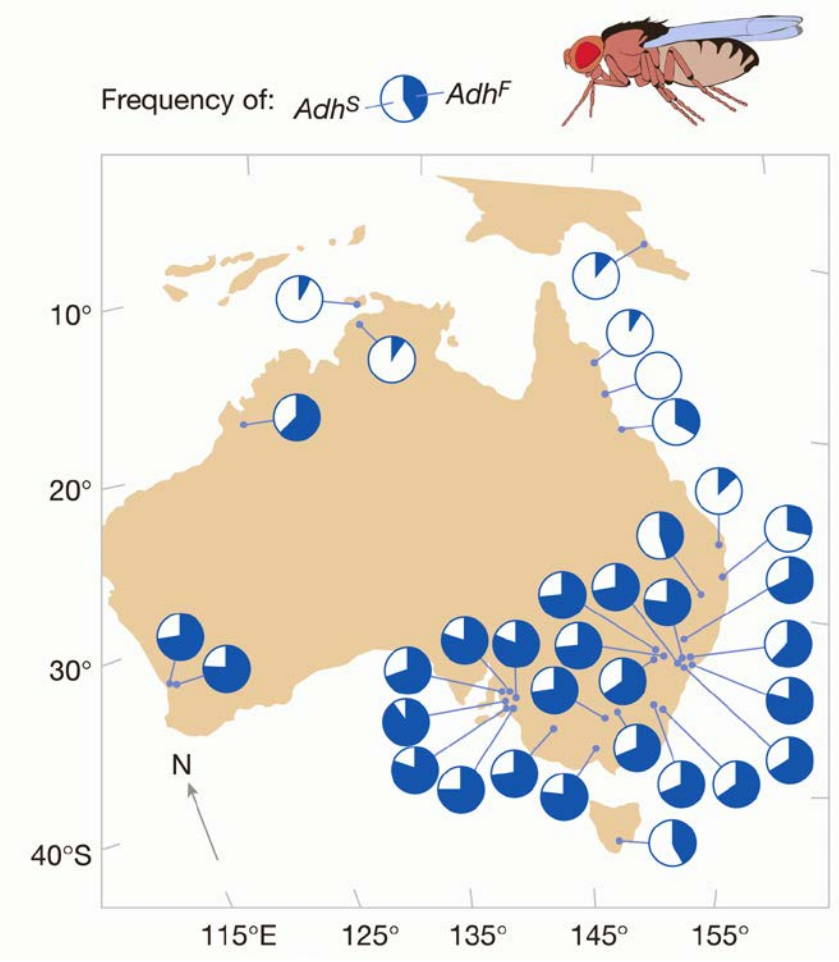


From FM Johnson & C Denniston (1964) *Nature* **204**, 906-907.

```
1                                                   g
atg.tcg.ttt.act.ttg.acc.aac.aag.aac.gtg.att.ttc.gtt.gcc.ggt.ctg.gga.ggc.att.ggt
Met.Ser.Phe.Thr.Leu.Thr.Asn.Lys.Asn.Val.Ile.Phe.Val.Ala.Gly.Leu.Gly.Gly.Ile.Gly
61
ctg.gac.acc.agc.aag.gag.ctg.ctc.aag.cgc.gat.ctg.aag.aac.ctg.gtg.atc.ctc.gac.cgc
Leu.Asp.Thr.Ser.Lys.Glu.Leu.Leu.Lys.Arg.Asp.Leu.Lys.Asn.Leu.Val.Ile.Leu.Asp.Arg
121
att.gag.aac.ccg.gct.gcc.att.gcc.gag.ctg.aag.gca.atc.aat.cca.aag.gtg.acc.gtc.acc
Ile.Glu.Asn.Pro.Ala.Ala.Ile.Ala.Glu.Leu.Lys.Ala.Ile.Asn.Pro.Lys.Val.Thr.Val.Thr
181                                                       t
ttc.tac.ccc.tat.gat.gtg.acc.gtg.ccc.att.gcc.gag.acc.acc.aag.ctg.ctg.aag.acc.atc
Phe.Tyr.Pro.Tyr.Asp.Val.Thr.Val.Pro.Ile.Ala.Glu.Thr.Thr.Lys.Leu.Leu.Lys.Thr.Ile
241
ttc.gcc.cag.ctg.aag.acc.gtc.gat.gtc.ctg.atc.aac.gga.gct.ggt.atc.ctg.gac.gat.cac
Phe.Ala.Gln.Leu.Lys.Thr.Val.Asp.Val.Leu.Ile.Asn.Gly.Ala.Gly.Ile.Leu.Asp.Asp.His
301
cag.atc.gag.cgc.acc.att.gcc.gtc.aac.tac.act.ggc.ctg.gtc.aac.acc.acg.acg.gcc.att
Gln.Ile.Glu.Arg.Thr.Ile.Ala.Val.Asn.Tyr.Thr.Gly.Leu.Val.Asn.Thr.Thr.Thr.Ala.Ile
361                                 t       a
ctg.gac.ttc.tgg.gac.aag.cgc.aag.ggc.ggt.ccc.ggt.ggt.atc.atc.tgc.aac.att.gga.tcc
Leu.Asp.Phe.Trp.Asp.Lys.Arg.Lys.Gly.Gly.Pro.Gly.Gly.Ile.Ile.Cys.Asn.Ile.Gly.Ser
421                         a
gtc.act.gga.ttc.aat.gcc.atc.tac.cag.gtg.ccc.gtc.tac.tcc.ggc.acc.aag.gcc.gcc.gtg
Val.Thr.Gly.Phe.Asn.Ala.Ile.Tyr.Gln.Val.Pro.Val.Tyr.Ser.Gly.Thr.Lys.Ala.Ala.Val
481                                         a       c               g           t
gtc.aac.ttc.acc.agc.tcc.ctg.gcg.aaa.ctg.gcc.ccc.att.acc.ggc.gtg.acc.gct.tac.acc
Val.Asn.Phe.Thr.Ser.Ser.Leu.Ala.Lys.Leu.Ala.Pro.Ile.Thr.Gly.Val.Thr.Ala.Tyr.Thr
541                                                   c
gtg.aac.ccc.ggc.atc.acc.cgc.acc.acc.ctg.gtg.cac.aag.ttc.aac.tcc.tgg.ttg.gat.gtt
Val.Asn.Pro.Gly.Ile.Thr.Arg.Thr.Thr.Leu.Val.His.Lys.Phe.Asn.Ser.Trp.Leu.Asp.Val
601   t           c                                           c
gag.ccc.cag.gtt.gct.gag.aag.ctc.ctg.gct.cat.ccc.acc.cag.cca.tcg.ttg.gcc.tgc.gcc
Glu.Pro.Gln.Val.Ala.Glu.Lys.Leu.Leu.Ala.His.Pro.Thr.Gln.Pro.Ser.Leu.Ala.Cys.Ala
661                             a
gag.aac.ttc.gtc.aag.gct.atc.gag.ctg.aac.cag.aac.gga.gcc.atc.tgg.aaa.ctg.gac.ctg
Glu.Asn.Phe.Val.Lys.Ala.Ile.Glu.Leu.Asn.Gln.Asn.Gly.Ala.Ile.Trp.Lys.Leu.Asp.Leu
721
ggc.acc.ctg.gag.gcc.atc.cag.tgg.acc.aag.cac.tgg.gac.tcc.ggc.atc.
Gly.Thr.Leu.Glu.Ala.Ile.Gln.Trp.Thr.Lys.His.Trp.Asp.Ser.Gly.Ile.
```

Figure 1.1: The DNA sequence for the coding region of the reference allele from the alcohol dehydrogenase locus of *Drosophila melanogaster*. The translation, given below the DNA sequence, uses the three-letter codes for amino acids. The letters over certain bases indicate the variants for those nucleotides found in a sample from nature. The variant at position 578 changes the amino acid of its codon from lysine to threonine.

Adh is polymorphic in *D. melanogaster* populations all over the world.



What is the *allele frequency* of adh^F in Miami?

$p(F) = (1)(1/110) + (\frac{1}{2})(24/110) + (0)(85/110)$

$= 1/110 + 12/110 = 13/110 =$ **0.118**

or

$p(F) = 2/220 + 24/220 = 26/220 = 13/110 =$ **0.118**

What is the *heterozygosity*?

$H = 24/110 =$ **0.218**

What is the *expected* heterozygosity?

$E(H) = 2pq = 2(0.118)(0.882) =$ **0.208**

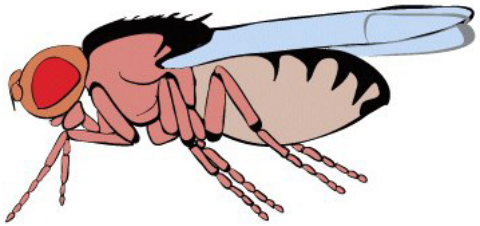
site	N	FF	FS	SS	P(F)	s.e.
Miami FL	110	1	24	85	0.118	0.022
Orlando FL	120	3	35	82	0.171	0.024
Raleigh NC	566	67	225	274	0.317	0.014
Winchester VA	252	29	123	100	0.359	0.021
Boston MA	378	31	164	183	0.299	0.017
Portland ME	307	45	174	88	0.430	0.020
Erie PA	122	37	53	32	0.520	0.032

E(heterozygosity) is a natural way to measure genetic variation.

Why? Because it's the *expected variance* of allele numbers per fly!

How many *Fast* alleles does a typical Miami fly have?

site	N	FF	FS	SS	P(F)	s.e.
Miami FL	110	1	24	85	0.118	0.022



mean(#*F*) = (2)(1/110) + (1)(24/110) + (0)(85/110) = **0.236** = **2*p***

var(#*F*) = $(2-0.236)^2(1/110) + (1-0.236)^2(24/110) + (0-0.236)^2(85/110)$ = **0.199**

And what did we *expect* (given the observed allele frequencies)?

E[var(#F)] = $(2-2p)^2(p^2) + (1-2p)^2(2pq) + (0-2p)^2(q^2)$ = **2*pq*** = 2(0.118)(0.882) = **0.208**

This is exactly the same as the expected heterozygosity (2*pq*)!

And var(#*F*) = var(#*S*), as long as we are considering just two alleles.

The variance and the heterozygosity are both *highest* when $p = q = \frac{1}{2}$.

Exercise: Calculate the ***frequencies*** of the three ***alleles*** in this sample of alkaline phosphatase genotypes from humans. Then calculate the ***expected genotype frequencies***, which are shown in the last column.

Genotype	Number	Frequency	Expected
SS	141	0.4247	0.4096
SF	111	0.3343	0.3507
FF	28	0.0843	0.0751
SI	32	0.0964	0.1101
FI	15	0.0452	0.0471
II	5	0.0151	0.0074
Total	332	1.0000	1.0000

Table 1.2: The frequencies of alkaline phosphatase genotypes in a sample from the English people. The expected Hardy-Weinberg frequencies are given in the fourth column. The data are from Harris (1966).

What are the ***expected numbers*** of the six genotypes ?

What is the ***expected heterozygosity*** ?

At what ***allele frequencies*** would heterozygosity be ***highest*** ?

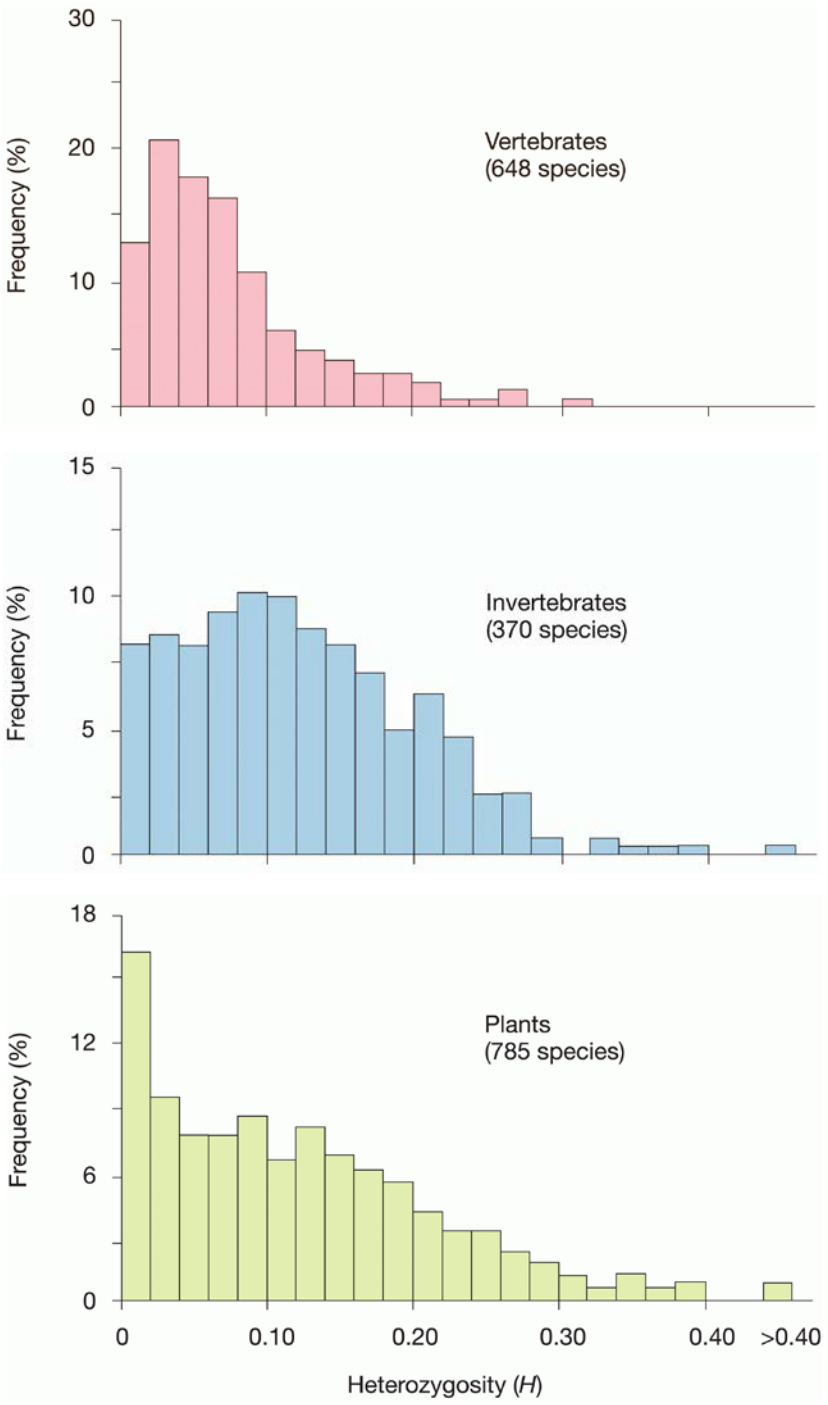
Enzyme polymorphisms are abundant at many loci and in *all* species.

These histograms show the average enzyme heterozygosities of individuals in hundreds of species that were surveyed at several to many loci.

A typical *species* is polymorphic at somewhere between one third and one half of its loci.

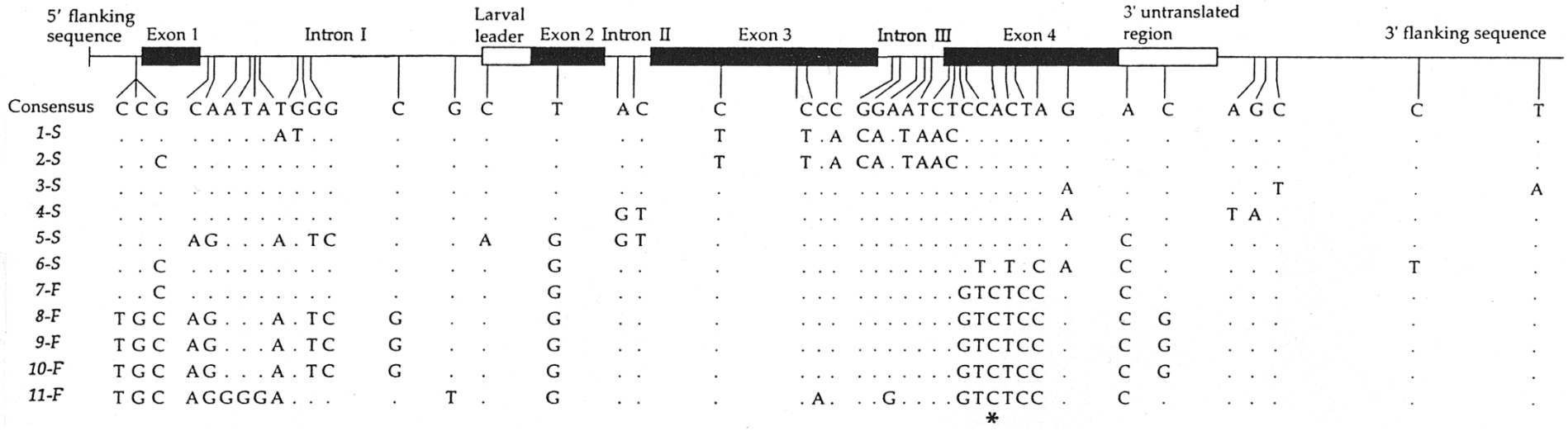
A typical *individual* is heterozygous at 4-15% of its loci.

That's a lot of genetic variation!



But there's *much more* at the DNA level!

Marty Kreitman sequenced 11 alleles of *Adh* for his Ph.D. thesis project in the early 1980s.



The *Fast* alleles are slightly more similar to each other, on average, than the *Slow* alleles.

But all 11 alleles differ from each other, at least by an insertion/deletion difference.

On *average* they differ by about 15 nucleotide differences (0.6%).

The average per-base-pair nucleotide difference or *heterozygosity* is called π.

For these *Adh* sequences, **π = 0.006**.

TABLE 9.1 Pairwise percent nucleotide differences (per 100 sites) among 11 sequences of the alcohol dehydrogenase locus in *Drosophila melanogaster*[a]

	Sequence									
Sequence	***1-S***	***2-S***	***3-S***	***4-S***	***5-S***	***6-S***	***7-F***	***8-F***	***9-F***	***10-F***
1-S										
2-S	0.13									
3-S	0.59	0.55								
4-S	0.67	0.63	0.25							
5-S	0.80	0.84	0.55	0.46						
6-S	0.80	0.67	0.38	0.46	0.59					
7-F	0.84	0.71	0.50	0.59	0.63	0.21				
8-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38			
9-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38	0.00		
10-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38	0.00	0.00	
11-F	1.22	1.18	0.97	1.05	0.84	0.67	0.46	0.42	0.42	0.42

From Nei (1987). Data from Kreitman (1983).

[a]Total number of compared sites is 2379. *S* and *F* denote the slow and fast migrating electrophoretic alleles, respectively.

How should we represent genetic variation?

There's no "best" way. It depends on the question!

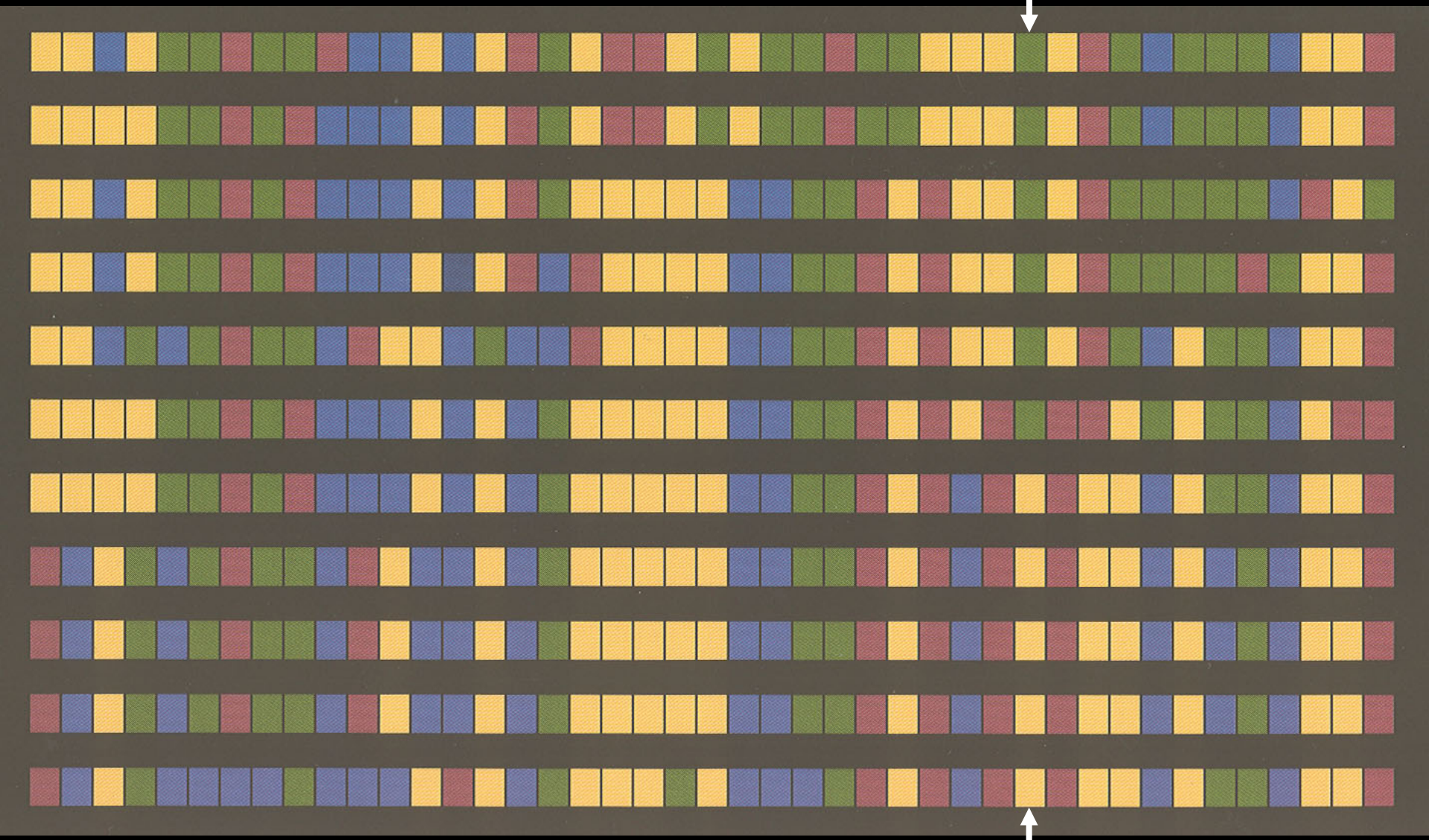
Allele	39	226	387	393	441	513	519	531	540	578	606	615	645	684
Reference	T	C	C	C	C	C	T	C	C	A	C	T	A	G
Wa-S	.	T	T	.	A	A	C	.	.	.	.	.	.	.
Fl-1S	.	T	T	.	A	A	C	.	.	.	.	.	.	.
Af-S	.	.	.	.	.	.	.	.	.	.	.	.	.	A
Fr-S	.	.	.	.	.	.	.	.	.	.	.	.	.	A
Fl-2S	G	.	.	.	.	.	.	.	.	.	.	.	.	.
Ja-S	G	.	.	.	.	.	.	.	T	.	T	.	C	A
Fl-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Fr-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Wa-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Af-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Ja-F	G	.	.	A	.	.	.	G	T	C	T	C	C	.

K slow

T fast

Table 1.1: The 11 *ADH* alleles. A dot is placed when a nucleotide is the same as the nucleotide in the reference sequence. The numbers refer to the position in the coding sequence where the 14 variant nucleotides are found (see Figure 1.1). The first two letters of the allele name identify the place of origin. The S alleles have a lysine at position 192 of the protein; the F alleles have a threonine.

All of the polymorphic sites in color



Implication: The genetic differences *between* species are *evolutionary substitutions* caused by the fixation of polymorphisms *within* species.

In other words, **the differences between species are just like the differences within them, only more so.**

If this is true, then to understand how evolution works, we need to understand the processes that generate and shape genetic polymorphism.

The big three:
mutation, drift and selection.
(Also recombination, migration, etc.)

Figure 1.2: The DNA sequence for *D. melanogaster* ADH with those bases and amino acids that differ in *D. erecta* shown below. The *erecta* sequence is from Jeffs et al. (1994).

And the differences *within* species are just like those *between* them, only *less* so!

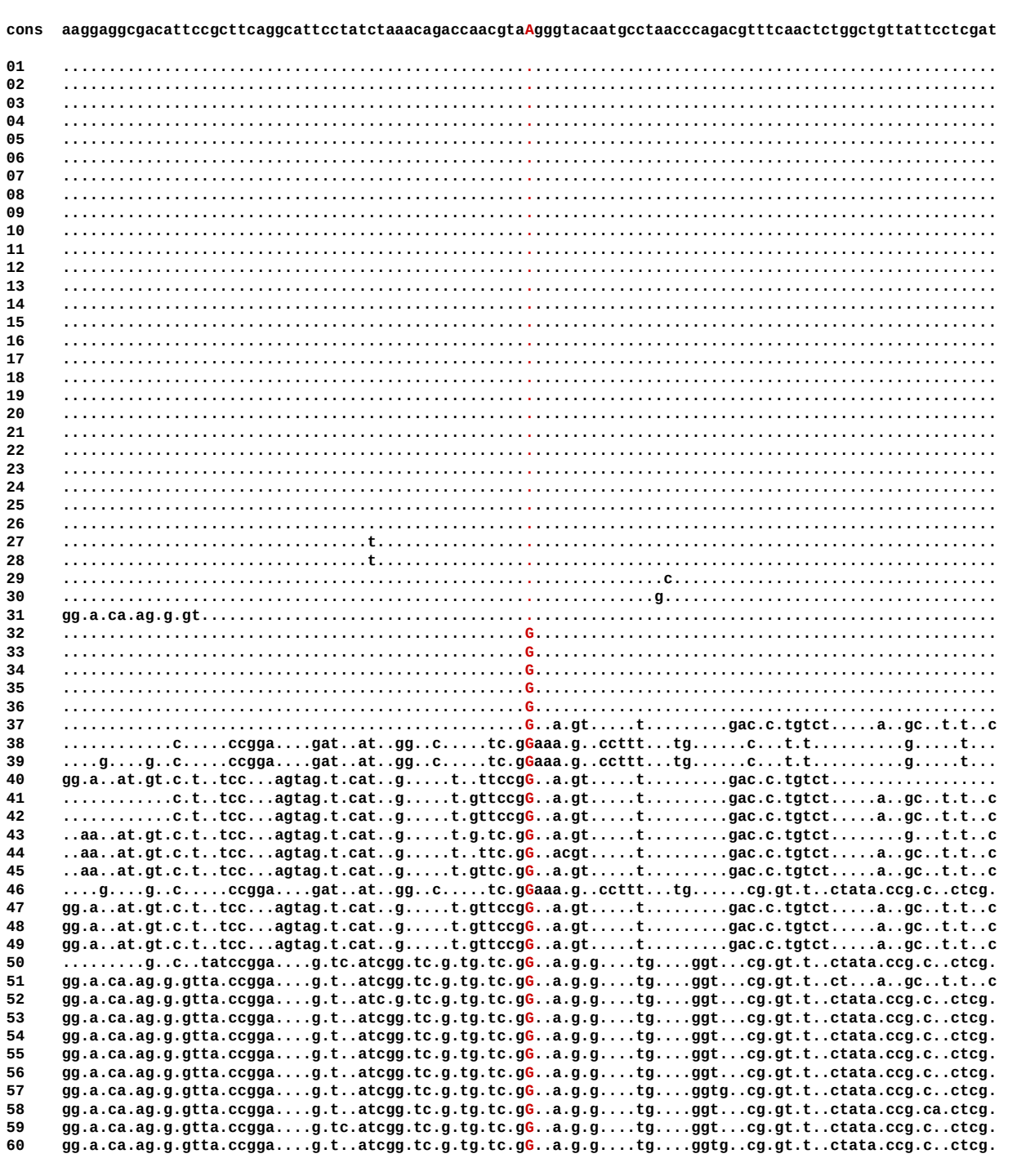
SNPs in the lactase gene region on chromosome 2, sampled from Utahns of European ancestry.

The consensus (majority) nucleotides are shown at the top ("cons").

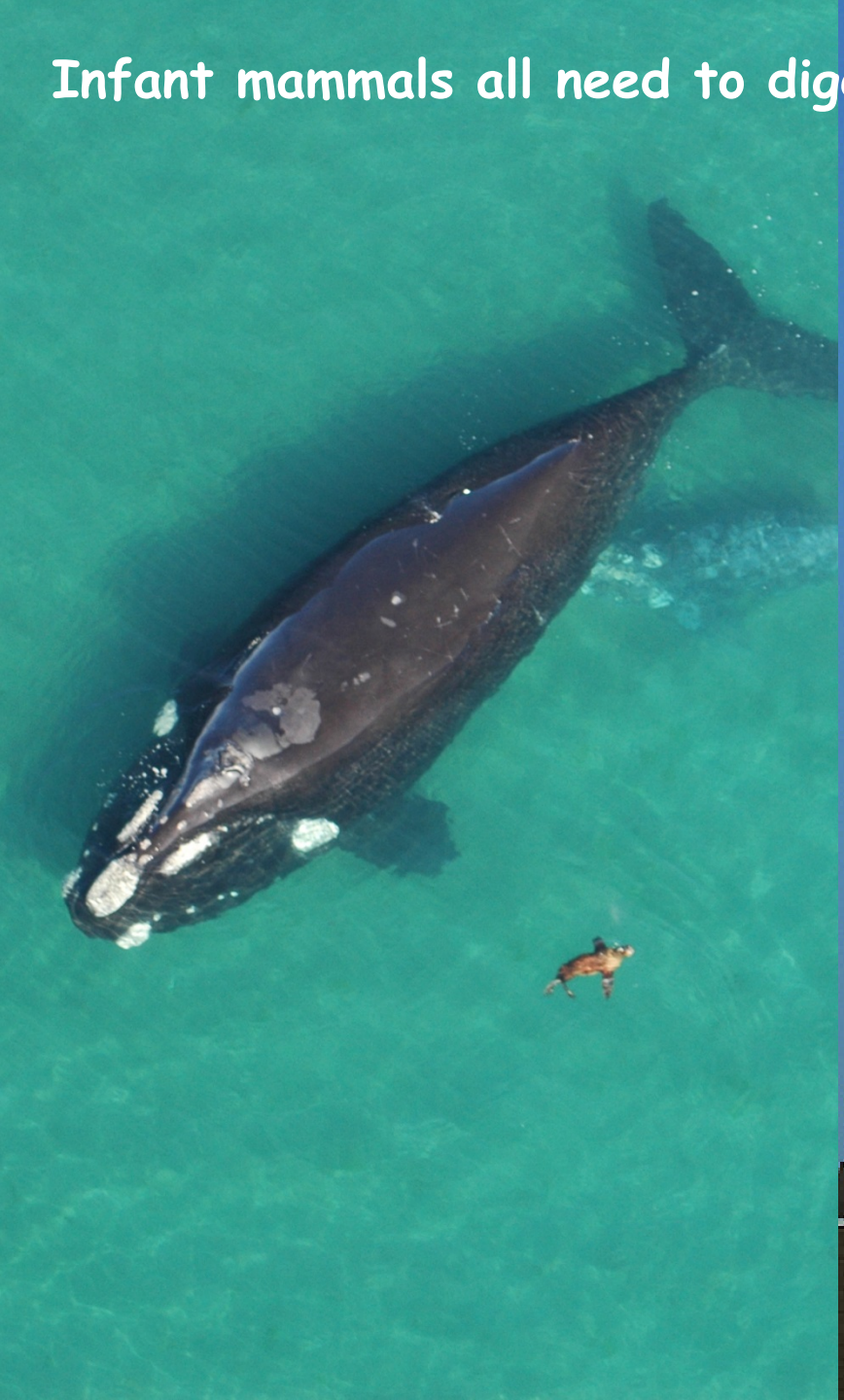
The first 26 chromosomes have exactly this sequence (as do 60 others which were omitted to save space).

The others are shown as their *differences* from the consensus.

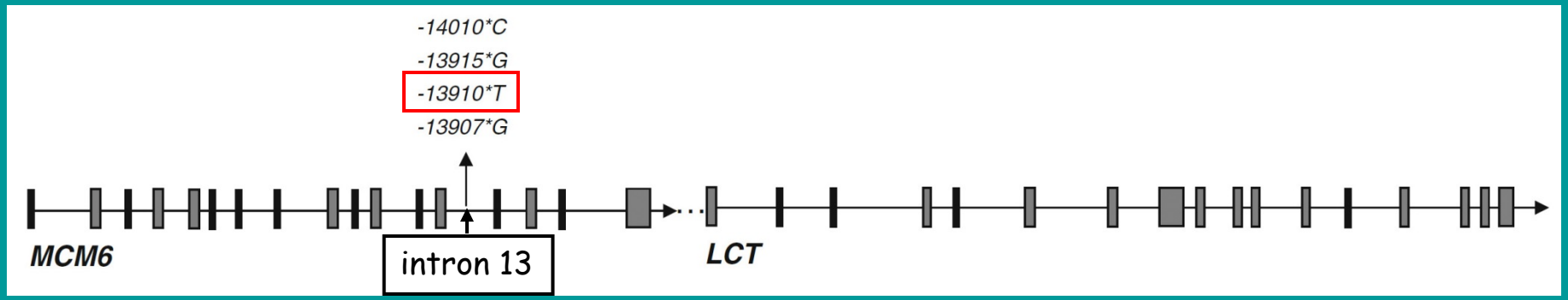
These 101 variable sites are embedded in a region of roughly 140,000 base pairs.



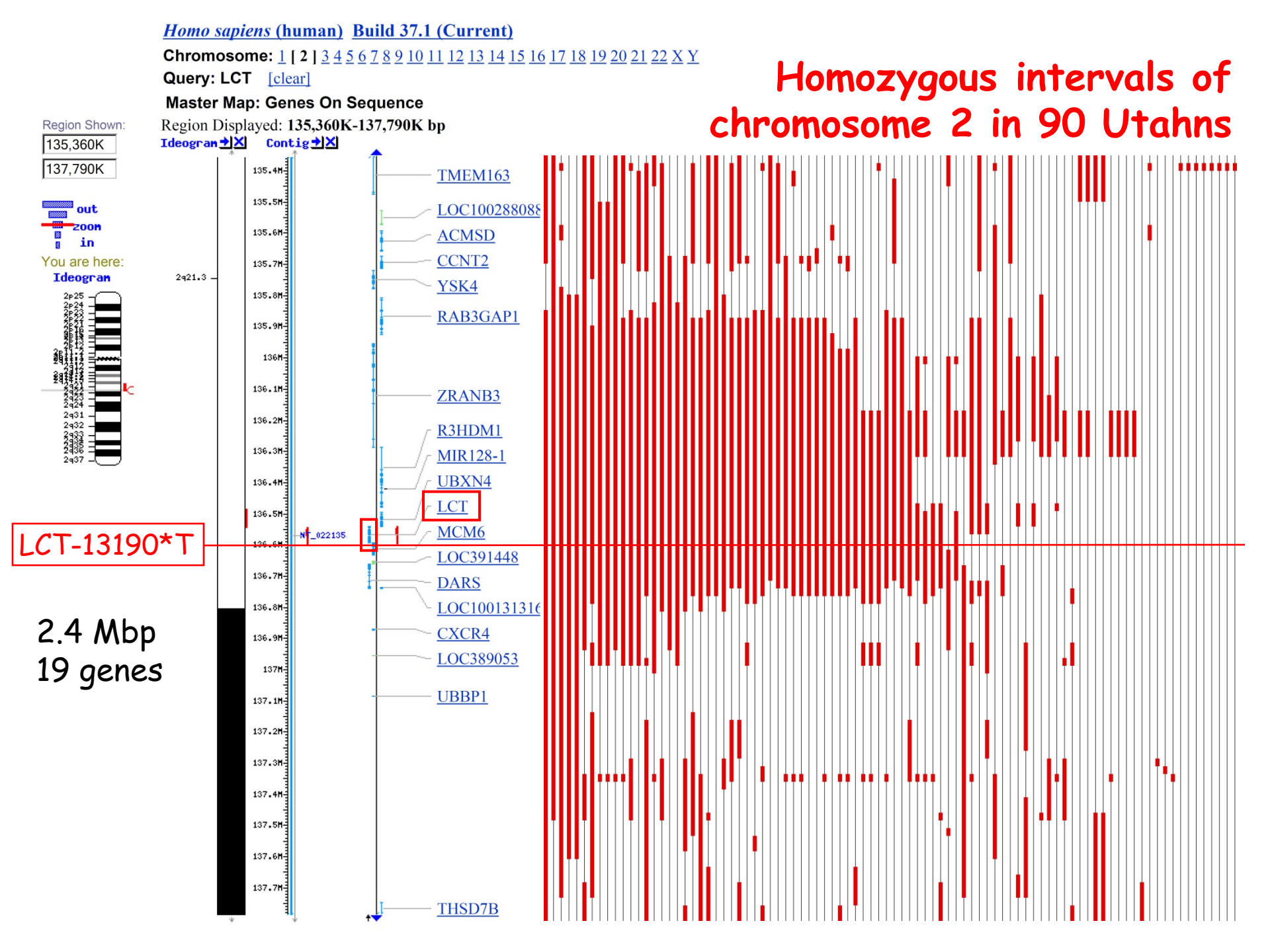
Infant mammals all need to digest lactose, but adults normally don't.



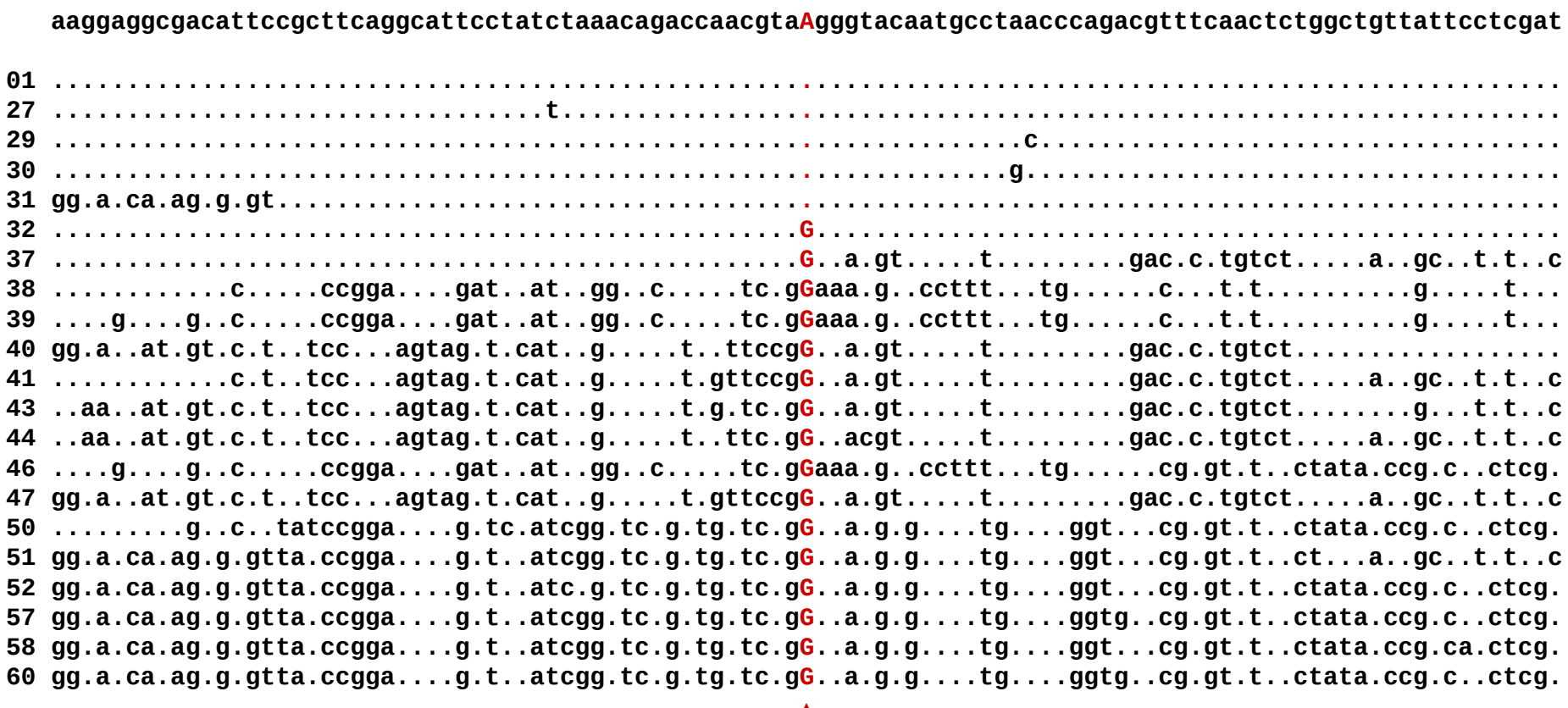
LCT region and putative lactase persistence mutations



Ingram *et al.* (2009) *Human Genetics* **124**:579-591


Homozygous intervals of chromosome 2 in 90 Utahns
Homo sapiens (human) Build 37.1 (Current)
Chromosome: 1 | 2 | 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Query: LCT [clear]
Master Map: Genes On Sequence
Region Shown:
135,360K
137,790K
Region Displayed: 135,360K-137,790K bp
Ideogram
Contig
out
zoom
in
You are here:
Ideogram
2q21.3
TMEM163
LOC100288088
ACMSD
CCNT2
YSK4
RAB3GAP1
ZRANB3
R3HDM1
MIR128-1
UBXN4
LCT
MCM6
LOC391448
DARS
LOC100131316
CXCR4
LOC389053
UBBP1
THSD7B
NT_022135
LCT-13190*T
2.4 Mbp
19 genes

How should we talk about (describe) this variation?



The "*A*" in the central position (T on the other strand) is the mutation that causes the lactase gene to remain "on" in adulthood, conferring tolerance to the consumption of lactose (milk sugar).

Measures of variation among DNA sequences

Gene diversity per sequence (a.k.a. heterozygosity) The probability that two random sequences differ.

Number, S, of segregating sites A "segregating site" is one that is polymorphic in the data.

Mean pairwise difference, Π, per sequence The average number nucleotide site differences between pairs of sequences.

Mean pairwise difference, π, per nucleotide Equals $\pi = \Pi/L$, where L is sequence length.

Mismatch distribution A histogram whose ith entry is the number of pairs of sequences that differ by i sites.

Site frequency spectrum A histogram whose ith entry is the number of polymorphic sites at which the mutant allele is present in i copies within the sample.

```
                0000000001 1111111112 2222222223 3333333334
                1234567890 1234567890 1234567890 1234567890

Sequence01      AATATGGCAC CTCCCAACCC TCTAGCATAT ACCACTTACA
Sequence02      .......T.. .C......TG C......C.. ..........
Sequence03      ..C....... .......... .......... ..........
Sequence04      .......T.. .C......TG C......... G.........
Sequence05      .......... .......... .......... ..........
Sequence06      .....A.... ........T. C......... G....C....
Sequence07      ..C....T.. .C......TG C......... G.........
Sequence08      .....A.T.. TC......TG C......... G.........
Sequence09      .......... .......... C......... ..........
Sequence10      .G...A.... ........T. C......C.. .T....C..G
Segregating:     ^^   ^ ^   ^^       ^^ ^       ^   ^^   ^^  ^
```

15 segregating sites

The number, $\binom{k}{2}$, of ways to choose 2 items from k

There are k ways to choose the first item. Having chosen the first, there are $k-1$ ways to choose the second, so there are $k(k-1)$ pairs. But this counts pair AB separately from BA. We are interested in unordered pairs, so

$$\binom{k}{2} = k(k-1)/2$$

A a set of made-up DNA sequences

```
          00000 00001
          12345 67890
S1        AAACT GTCAT
S2        ..... A....
S3        ..... A...C
S4        ..G.. A....
S5        ..G.. A....
```

Calculate the mean pairwise difference, the number of segregating sites, the mismatch distribution and the site frequency spectrum.

Mean pairwise difference (MPD)

```
    00000 00001        Pair Diff   Pair Diff
    12345 67890       (1,2)   1   (2,5)   1
S1  AAACT GTCAT       (1,3)   2   (3,4)   2
S2  ..... A....       (1,4)   2   (3,5)   2
S3  ..... A...C       (1,5)   2   (4,5)   0
S4  ..G.. A....       (2,3)   1   Sum diffs: 14
S5  ..G.. A....       (2,4)   1   MPD/seq  : 14/10
```

Column	Differences
03	2×3
06	1×4
10	1×4
Sum	14

Number of pairs: $(5 \times 4)/2 = 10$

MPD per sequence: $\Pi = 14/10$

MPD per site: $\pi = 14/(10 \times 10)$

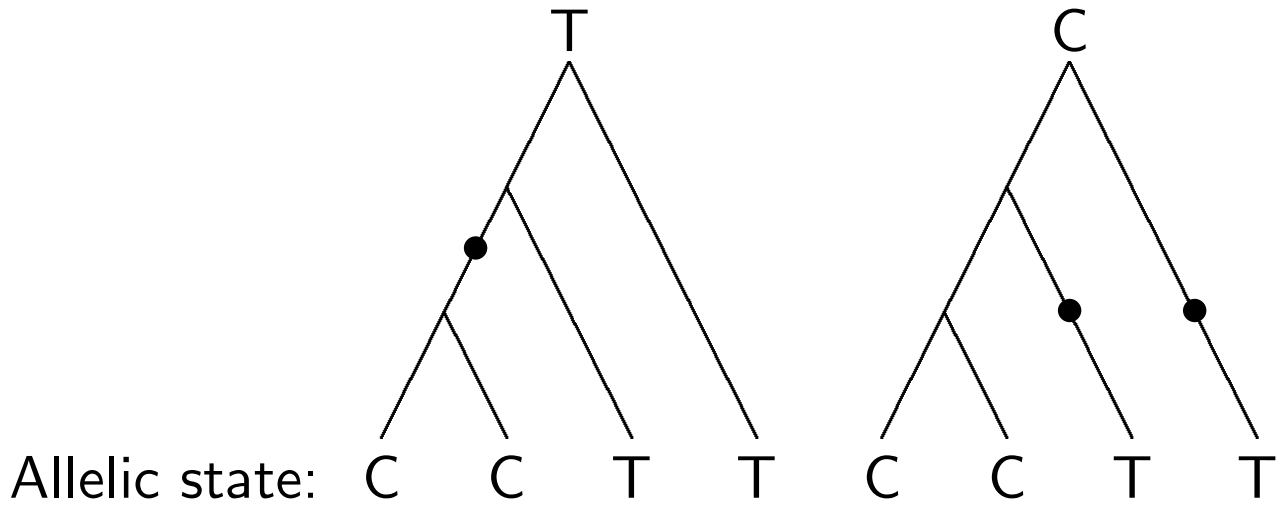
Mismatch distribution

```
    00000 00001        Pair Diff    Pair Diff
    12345 67890       (1,2)   1    (2,5)   1
S1  AAACT GTCAT       (1,3)   2    (3,4)   2
S2  ..... A....       (1,4)   2    (3,5)   2
S3  ..... A...C       (1,5)   2    (4,5)   0
S4  ..G.. A....       (2,3)   1
S5  ..G.. A....       (2,4)   1
```

Mismatch distribution

Differences	0	1	2
Count	1	4	5

Calling ancestral and derived alleles



- Two hypotheses about which allele is ancestral.
- "C" requires 2 mutations; "T" requires 1.
- Because mutations are rare, "T" is more likely.
- When the in-group is polymorphic, the ancestral allele is usually the one present in the out-group.

Unfolded site frequency spectrum

```
        123456
Human1  AATAGC
Human2  ..AC..
Human3  .TACT.
Human4  ..ACT.
------------
Chimp   AAAATC
```

1: fixed.
2: T derived; singleton.
3: T derived; singleton.
4: C derived; tripleton.
5: G derived; doubleton.
6: fixed.

Singletons	2
Doubletons	1
Tripletons	1

Big picture from genome sequencing (*e.g.*, 1000 Genomes Project)

Individual nucleotide heterozygosities (π) are 0.0005 – 0.001 (1.5 – 3 M/genome)

Higher in Africa, lower in Eurasia.

Some cause amino-acid polymorphisms in proteins (10-20 K/genome, ~0.5/protein!)

Many insertion-deletion polymorphisms ("indels" of 1-50 bp, ~550 K/genome)

Many loss-of-function (LOF) mutations (~150/person, ~20 in known disease genes)

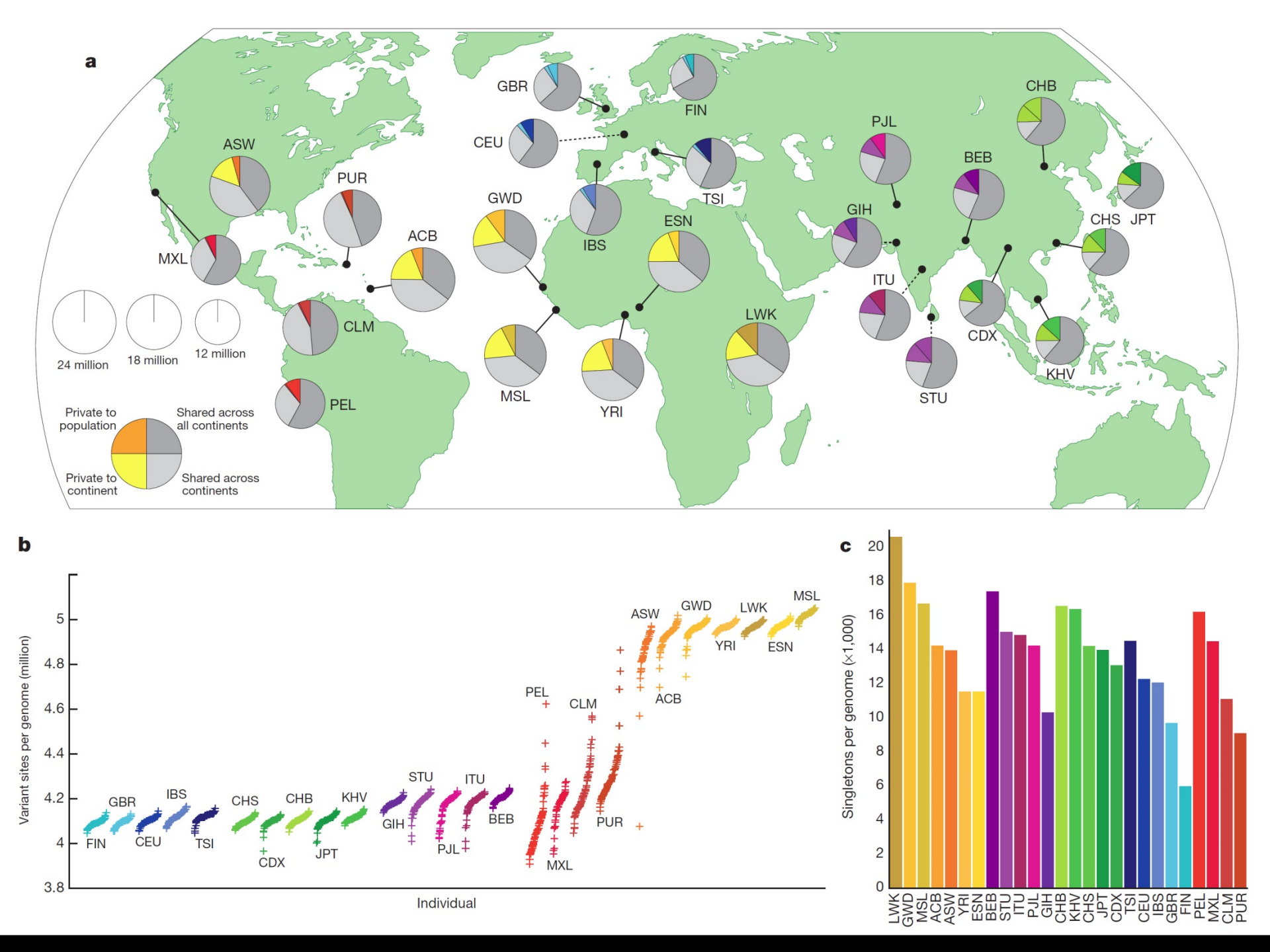
Many copy-number variants ("CNVs" of 2 kb – 2 Mb, ~200/genome)

But allele-frequency differences *between* populations are modest:

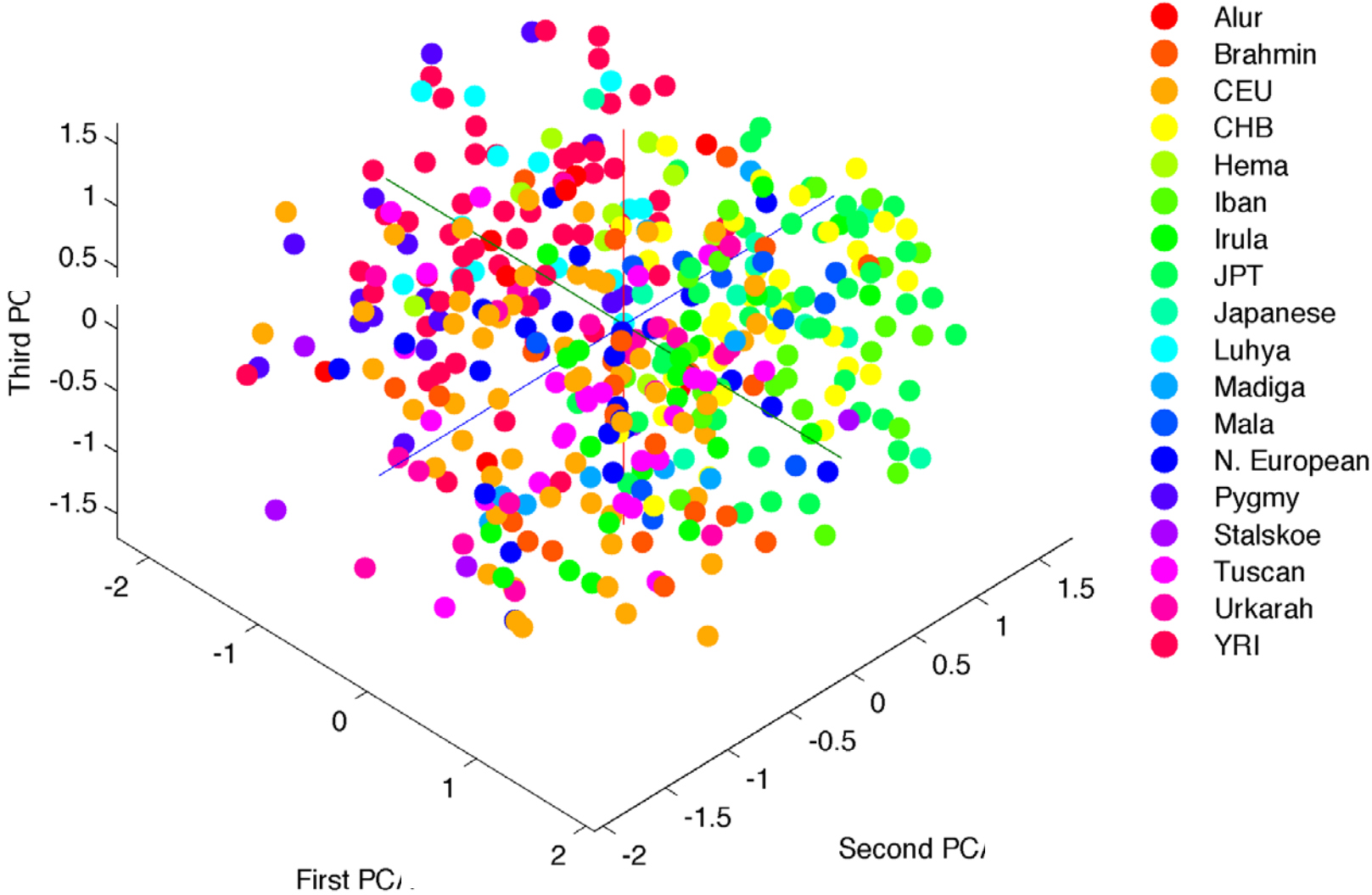
For typical loci, ~85% of the world-wide variation is *within* local populations.

Only 10-15% is accounted for (added) by differences *between* major groups.

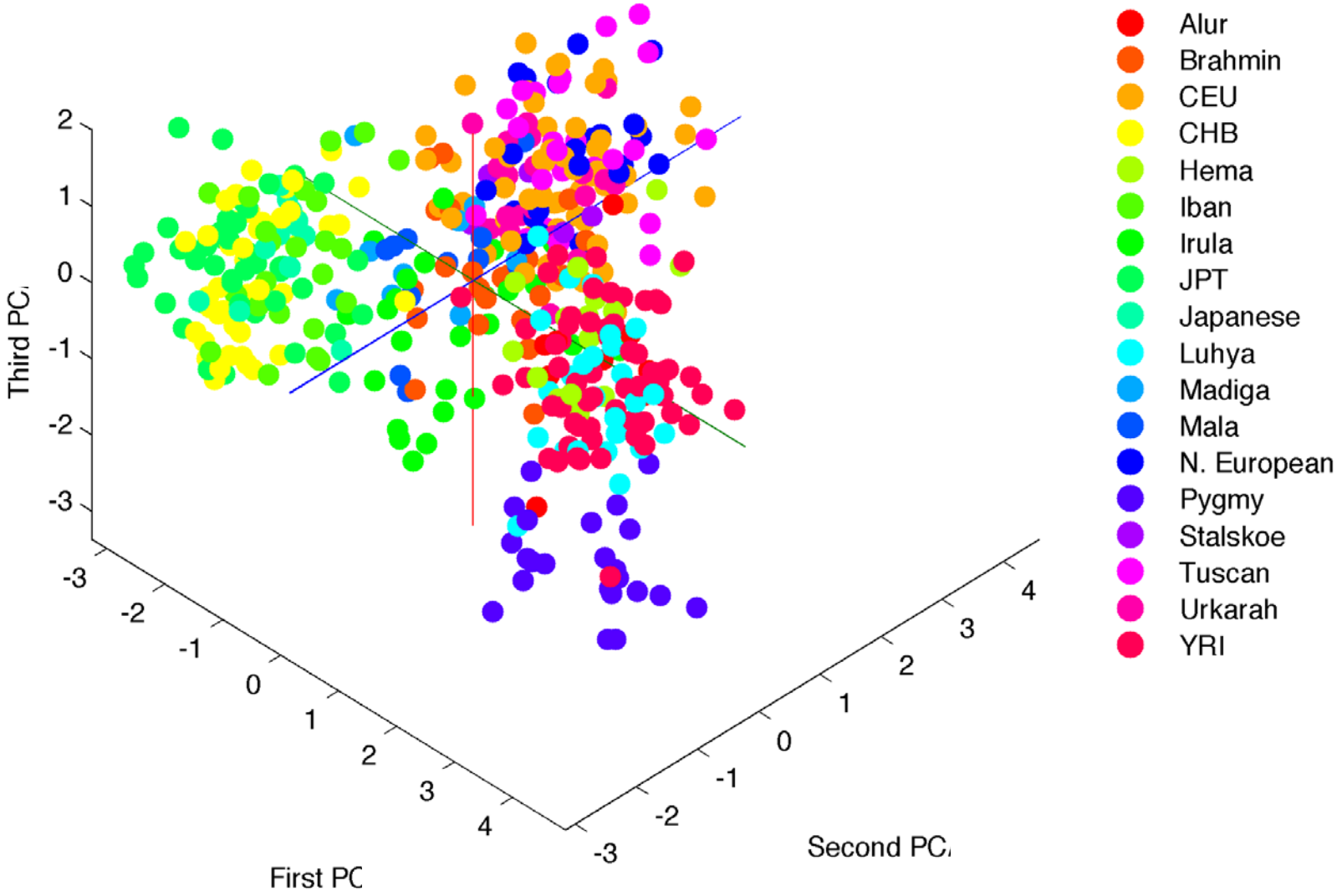
Long-standing question: Does this mean the "races" are *imaginary*?


a
GBR
FIN
CEU
TSI
IBS
GWD
ESN
MSL
YRI
LWK
ASW
ACB
PUR
MXL
CLM
PEL
PJL
GIH
ITU
STU
CHB
BEB
CHS
JPT
CDX
KHV
24 million
18 million
12 million
Private to population
Shared across all continents
Private to continent
Shared across continents
b
Variant sites per genome (million)
5
4.8
4.6
4.4
4.2
4
3.8
Individual
FIN
GBR
CEU
IBS
TSI
CDX
CHS
JPT
CHB
KHV
GIH
STU
PJL
ITU
BEB
MXL
PEL
CLM
PUR
ACB
ASW
YRI
GWD
LWK
ESN
MSL
c
Singletons per genome (×1,000)
20
18
16
14
12
10
8
6
4
2
0
LWK
GWD
MSL
ACB
ASW
YRI
ESN
BEB
STU
ITU
PJL
GIH
CHB
KHV
CHS
JPT
CDX
TSI
CEU
IBS
GBR
FIN
PEL
MXL
CLM
PUR

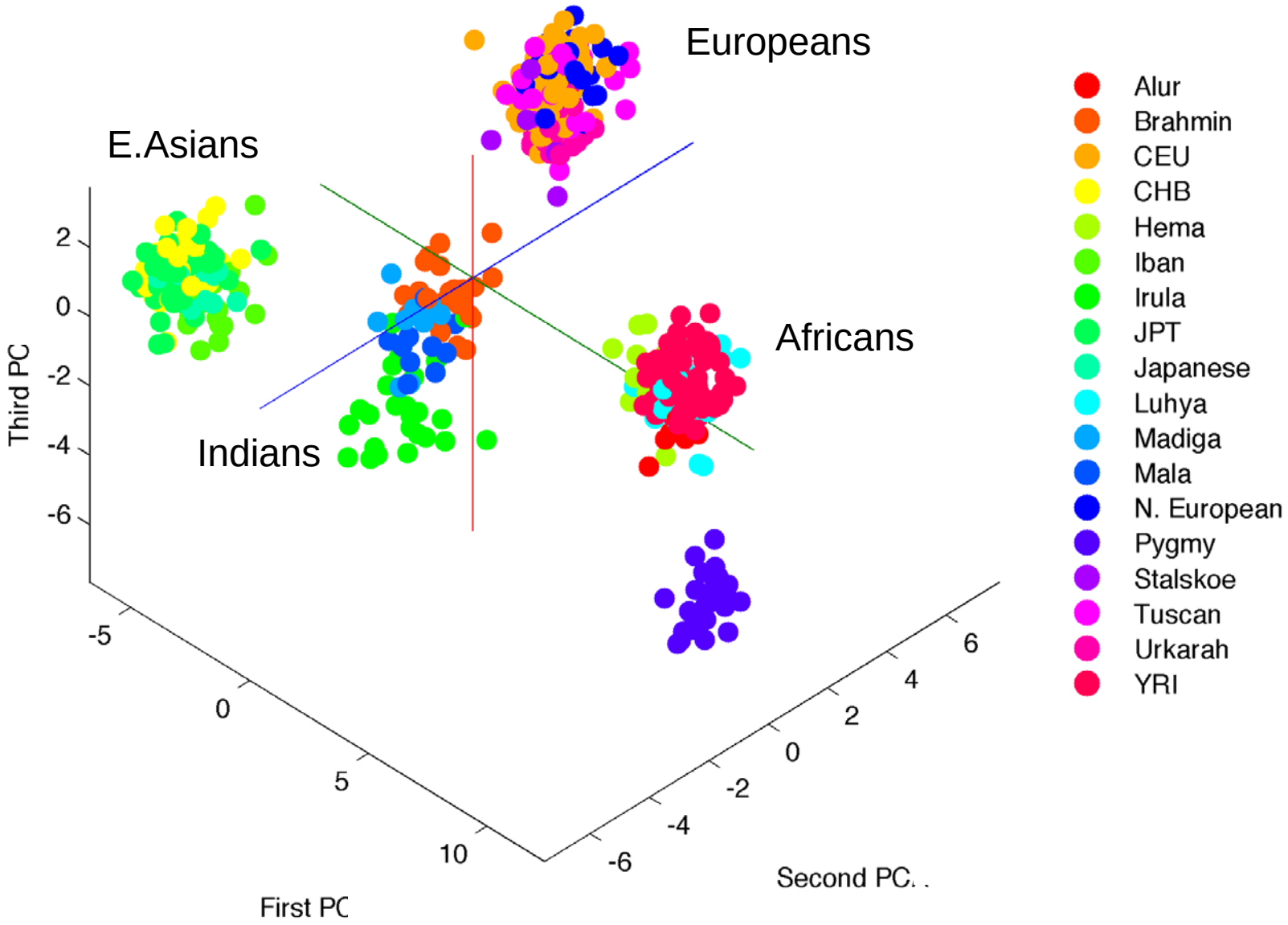
Genetic distances among 467 individuals: 10 SNPs



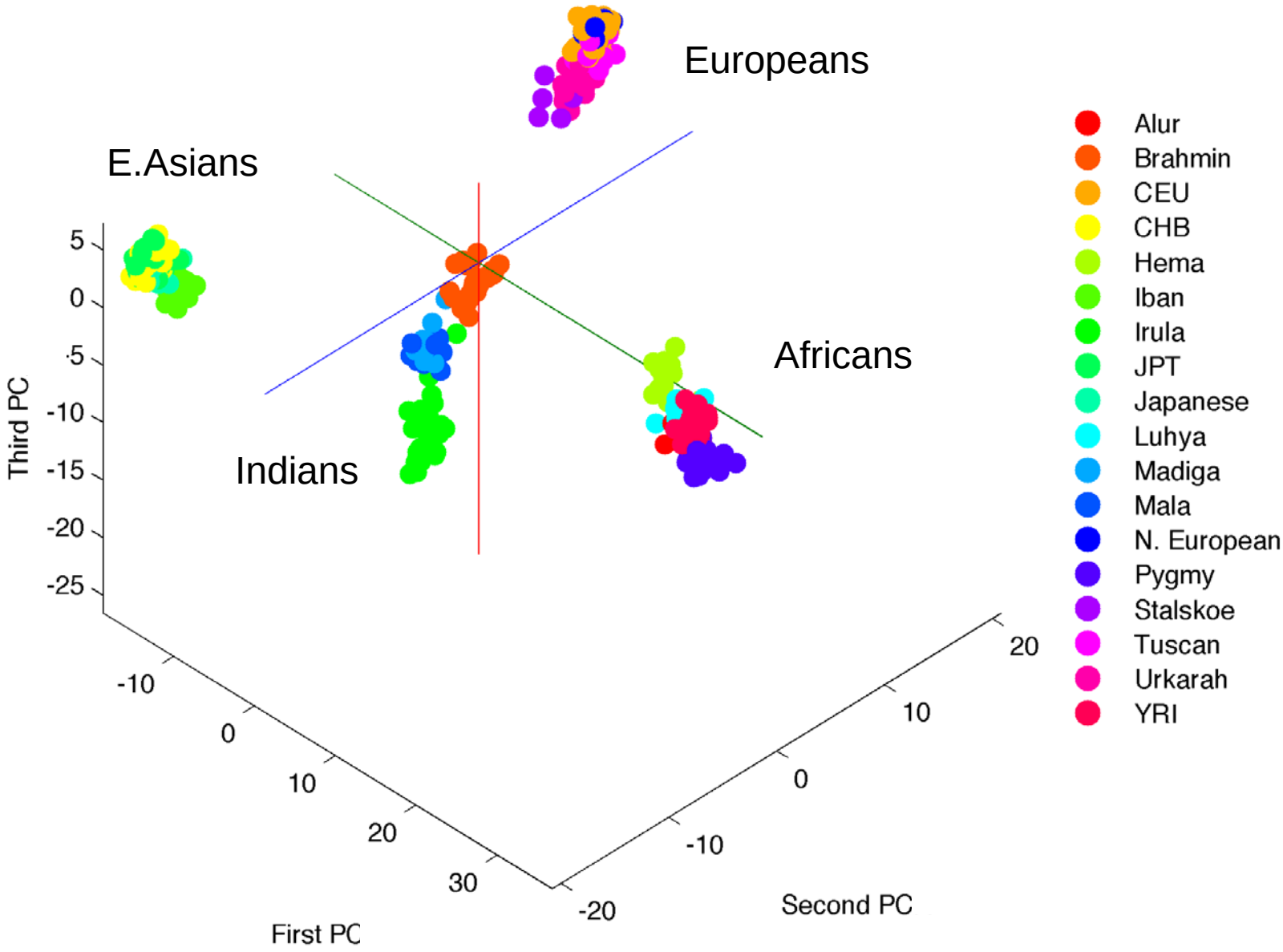
Genetic distances among 467 individuals: 100 SNPs



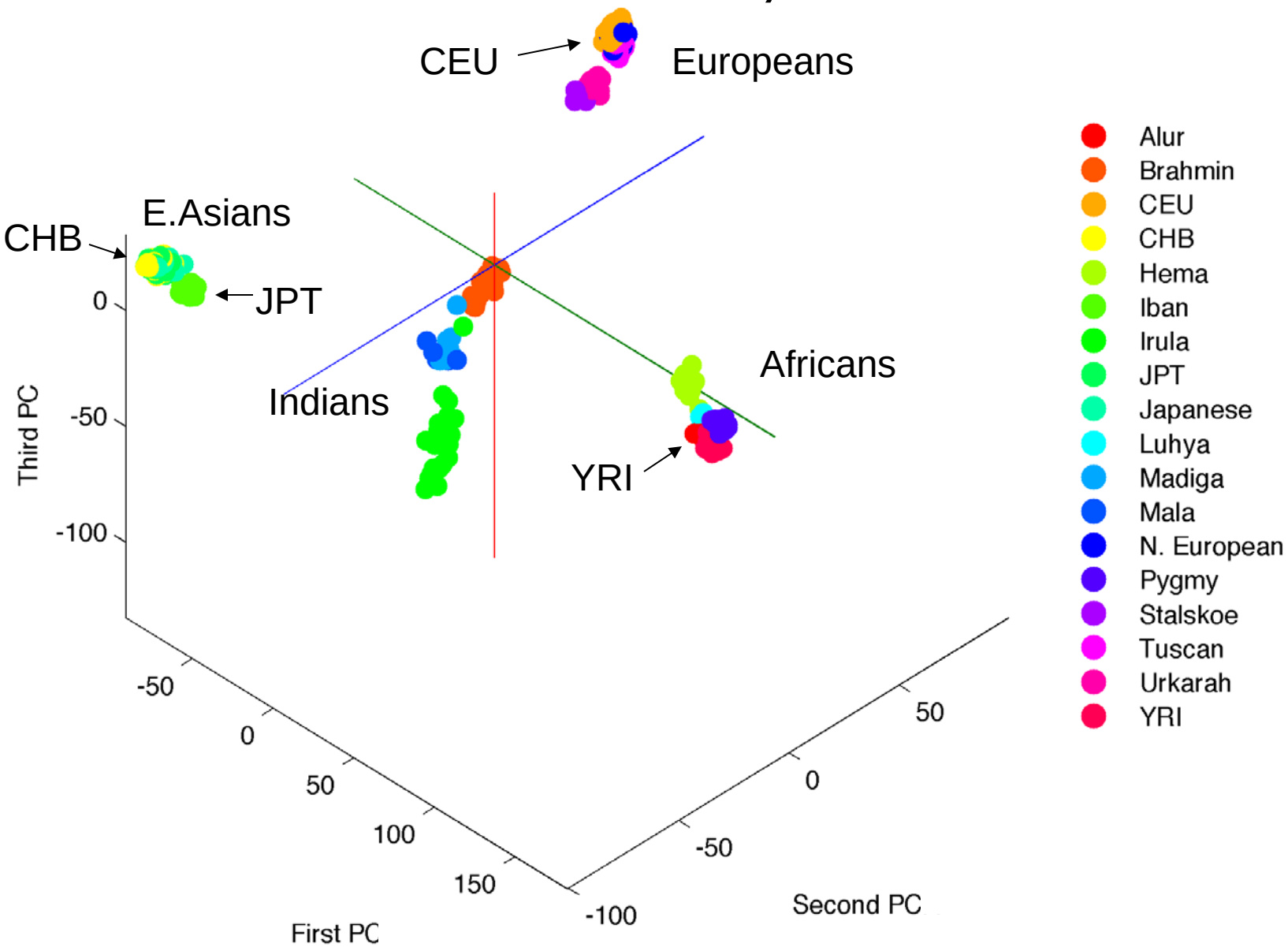
Genetic distances among 467 individuals: 1000 SNPs



Genetic distances among 467 individuals: 10,000 SNPs



Genetic distances among 467 individuals: 261,000 SNPs



Summary about genomic variation

Heterozygosity (expected) is a natural measure of genetic variation.

It can be described at many levels, for example:
- individual nucleotide or amino-acid sites
- whole genes or chromosome segments
- (sequence, or presence/absence)

Per-site heterozygosities are low: ($\pi \approx 0.0008$ for humans).

But staggering per genome: ($0.0008 \times 3 \times 10^9 = 2.4 \times 10^6$ per person).

Most genomic variation is not in genes, so probably meaningless.

Even so, the part with phenotypic effects is absolutely enormous.

Per-site, most human variation (>85%) occurs *within* local populations.

Less than 15% is explained by differences among continental "races".

And there are *no* diagnostic (fixed) differences among "races".

However, allelic states are *not independent* among loci.

So *per-genome*, long-separated populations may "cohere" *statistically*.

The vocabulary (-ies) of homologs: a muddle, beware!

vocabulary system:	*1*	*2*	*3*
Position on chromosome	*locus*	*locus*	*locus*
Protein- or RNA-coding locus	*gene*	*gene*	*gene*
1 of 2 or more sequence variants at locus	*allele*	*allele*	*allele*
Physical instance of the DNA at a locus	*gene*	*allele*	*gene copy*

Gillespie favors system 2, but we think system 3 is clearer because it distinguishes all four aspects of "gene-iness"

Orthologous genes occupy the *same* locus in *different* species (*i.e.*, they are separated by a speciation event).

Paralogous alleles occupy *different* loci in the *same* or *different* species (*i.e.*, they are separated by a gene-duplication event).

Alleles, orthologs and paralogs are all *homologs,* because they descend from a common ancestral sequence.