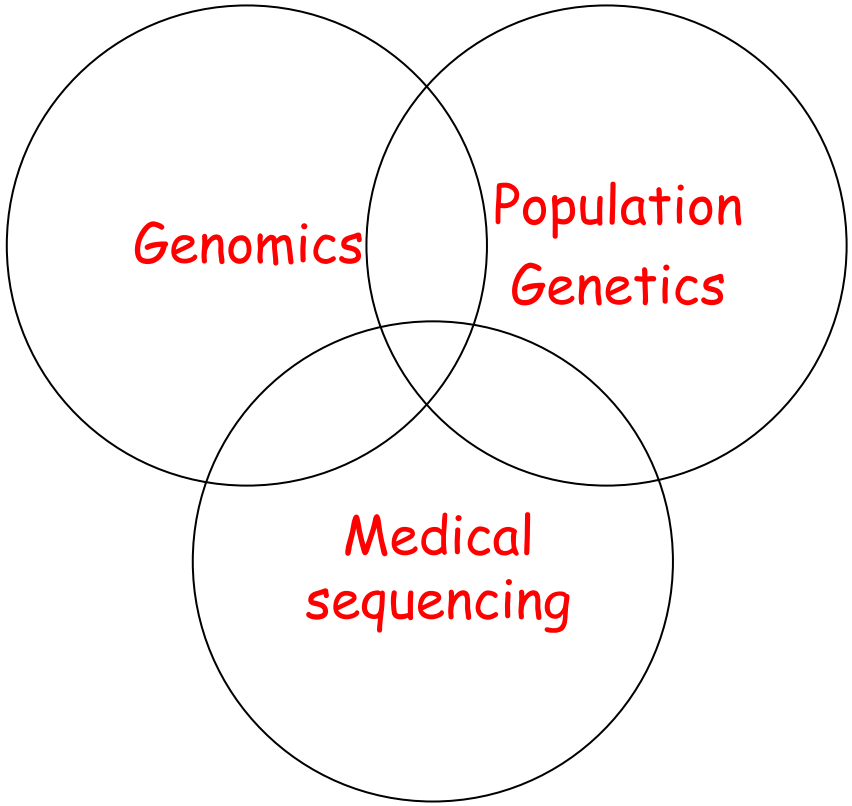
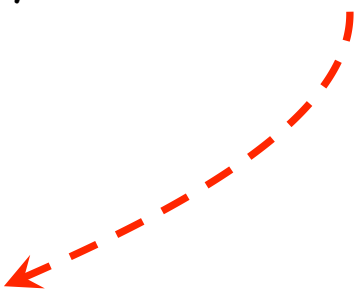

Genomics
Population
Genetics
Medical
sequencing

Why should we care (about genomics and population genetics)?

Consider a typical WGS in a clinical setting…(30x coverage of a haploid genome):

- 3 billion nt
- ~ 3,000,000 variant calls (heterozygous or homozygous; relative to "reference")
- ~ 2,997,000 are common/known (>1% population frequency)
- ~ 1,000 are rare: 950 "private" (immediate family); 50 are "de novo"
- (And ~2000 are sequencing errors)



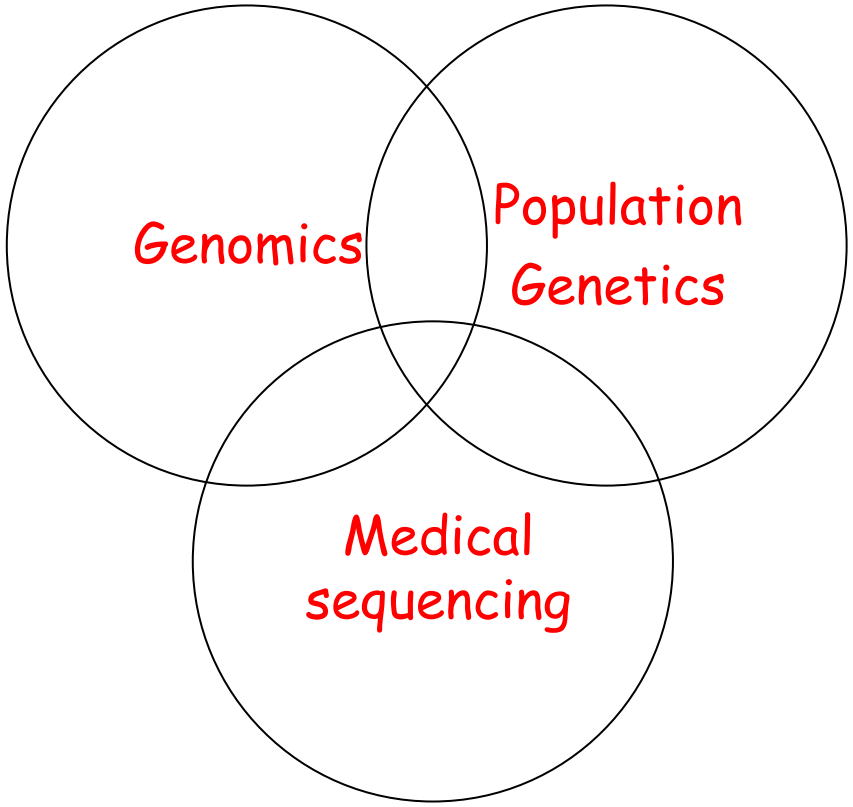
Which one(s) cause the phenotype?

Variation

Function

Evolution

Technology



Allele frequencies

Drift/selection

Population history

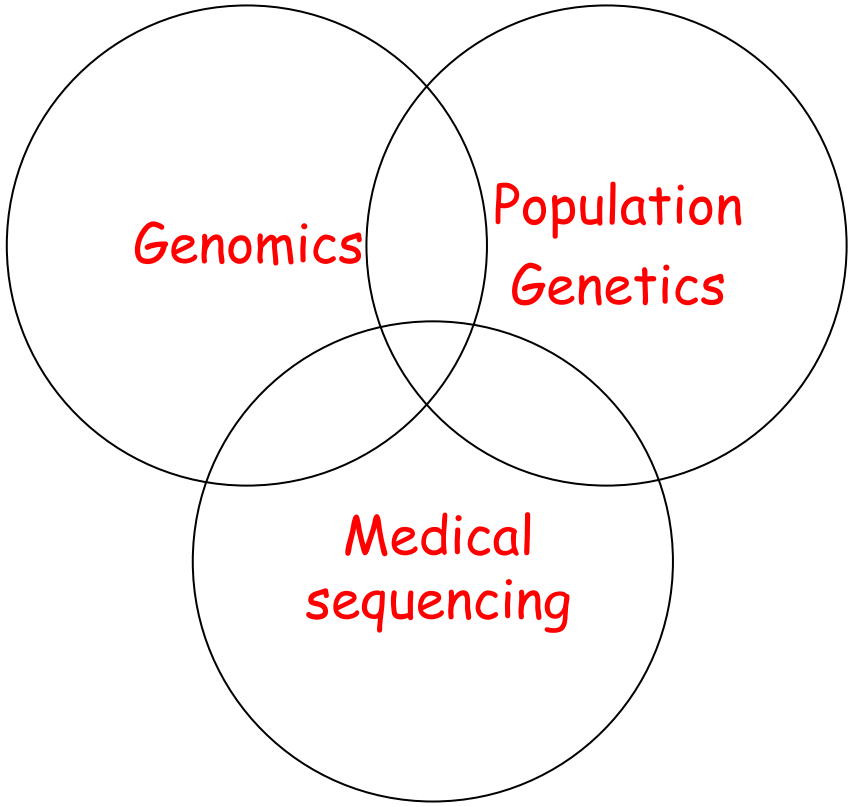
Phenotype -> WGS -> Causal variation

Variation

Function

Evolution

Technology

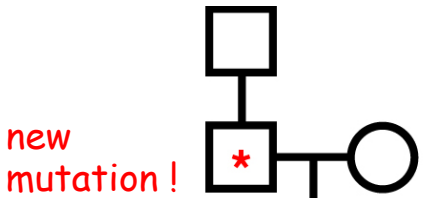


1) Sources of genetic variation
2) Measuring genetic variation
3) Population history

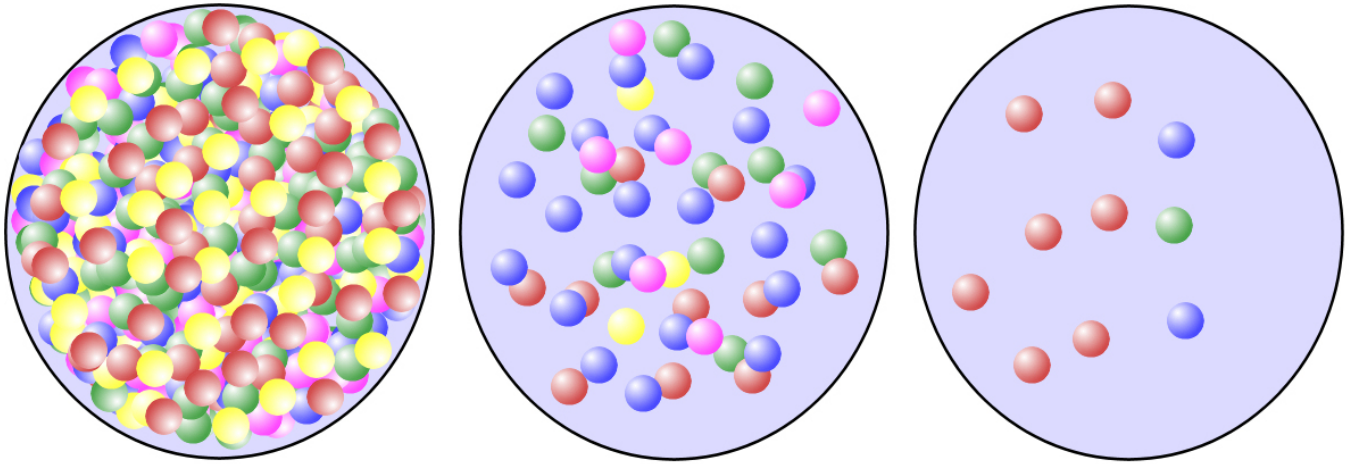
Phenotype -> WGS -> Causal variation

Where does genetic variation come from?

1) Mutation
   - ~1.5 x 10^{-8} per bp per generation
     (x 3,000,000,000 bp = ~45 new bp per genome)
   - 0-1 deleterious mutation
     (most mutations are Neutral)
2) Selection
   - Negative selection if deleterious (reduced fitness can be "balanced" by new mutations)
   - Positive selection if advantageous (rarely, but engine that drives adaptive change)
   - Most mutations are Neutral
3) Drift
   - Random fluctuation due to sampling
     ("bottlenecks", migration, unequal sex ratio)
   - Most effective in small populations…



How does genetic variation change over time?



Initial frequency distribution	
	0.20
	0.20
	0.20
	0.20
	0.20

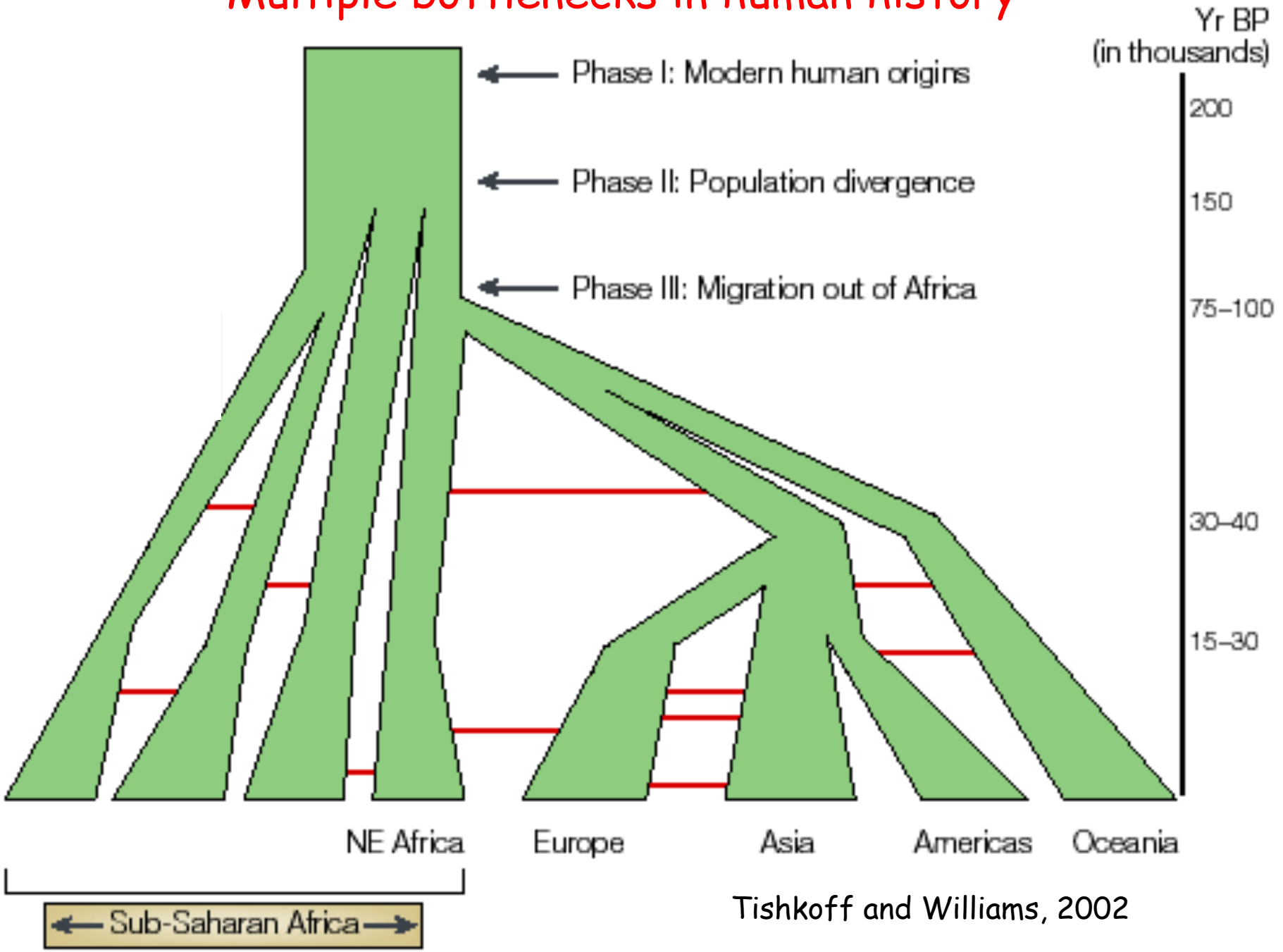
5th generation frequency distribution	
	0.40
	0.06
	0.20
	0.14
	0.20

10th generation frequency distribution	
	0.20
	0.00
	0.10
	0.00
	0.70

Q. Why do the red marbles predominate?
Q. How would a bigger jar affect things?

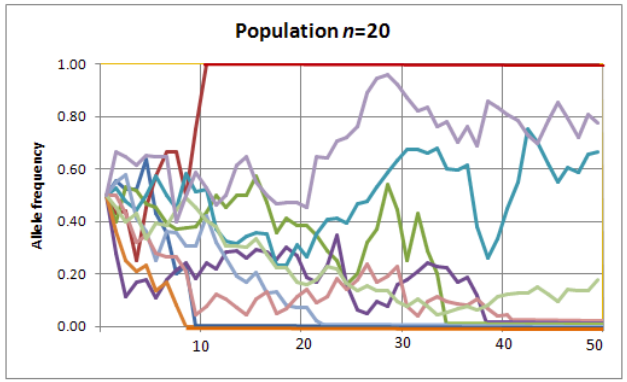
A. Drift is "more effective" in small populations; selection is more effective in large populations (fish, flies, microbes, NOT humans)

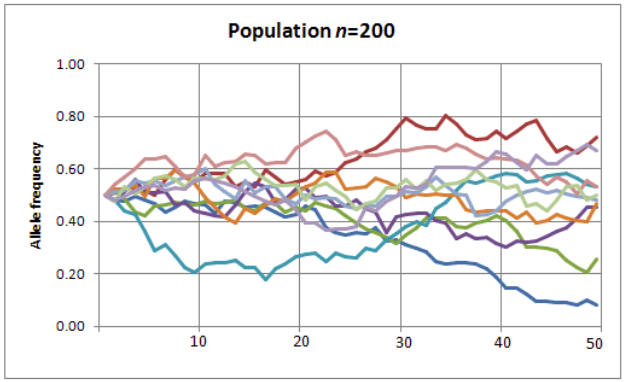
Multiple bottlenecks in human history

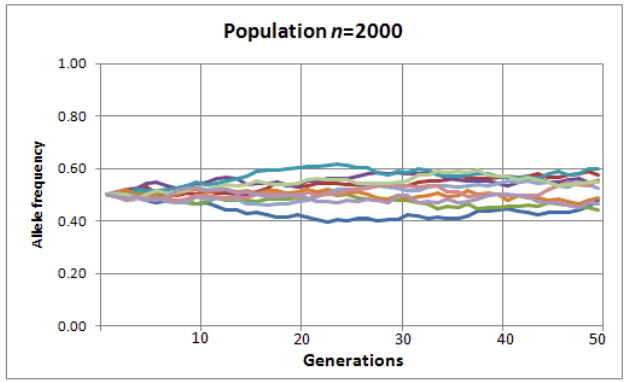


Tishkoff and Williams, 2002

Effect of population size on allele frequency drift (10 simulations each over 50 generations)







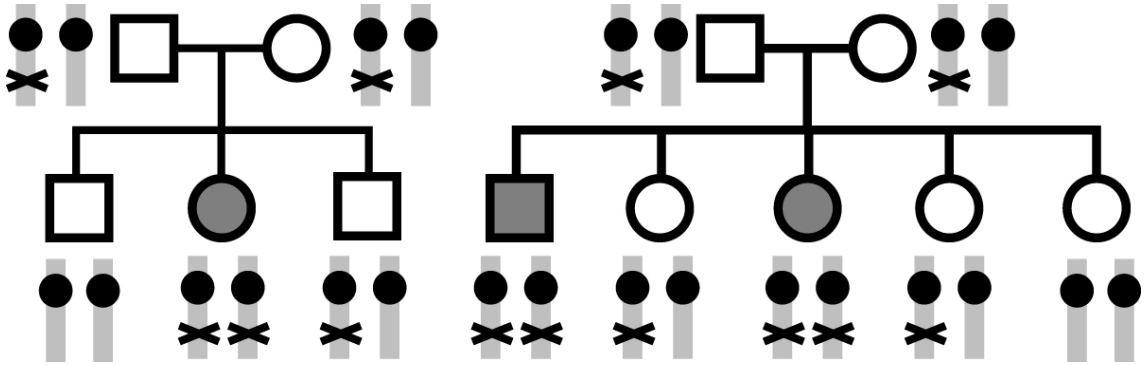
Take-home message: in a large and constant population with no selection; allele frequencies remain constant over time

N.B. this NEVER happens in reality

How do we measure genetic variation? (Frequencies of genotypes vs. frequencies of alleles)

Imagine a "population" of two families…

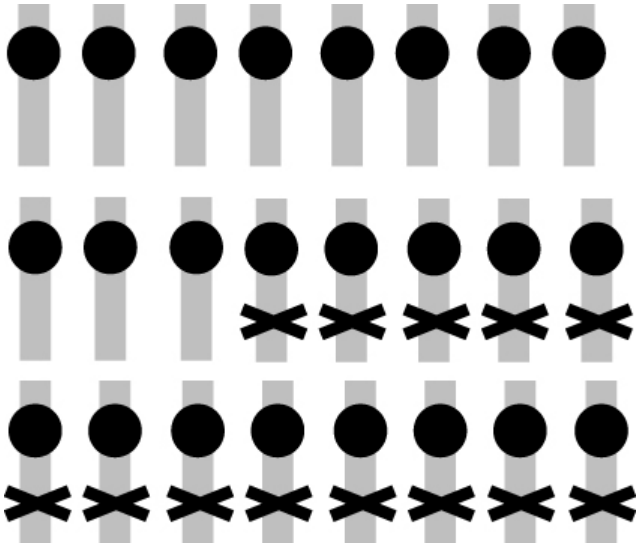
(recessive condition with mutant "-" allele marked with "x")



Genotype	No.	Proportion
+/+	2	17%
+/-	7	58%
-/-	3	25%

Count number of alleles, e.g.
(2 x 2) + (7 x 1) = 11
(7 x 1) + (3 x 2) = 13

Now consider the number of alleles…



Allele	No.	Proportion
+	11	46%
-	13	54%

Note: "p" and "q" are often used to refer to allele proportions, e.g
p = .46; q = .54

p + q = 1

Frequencies of genotypes vs. frequencies of alleles (take 2)

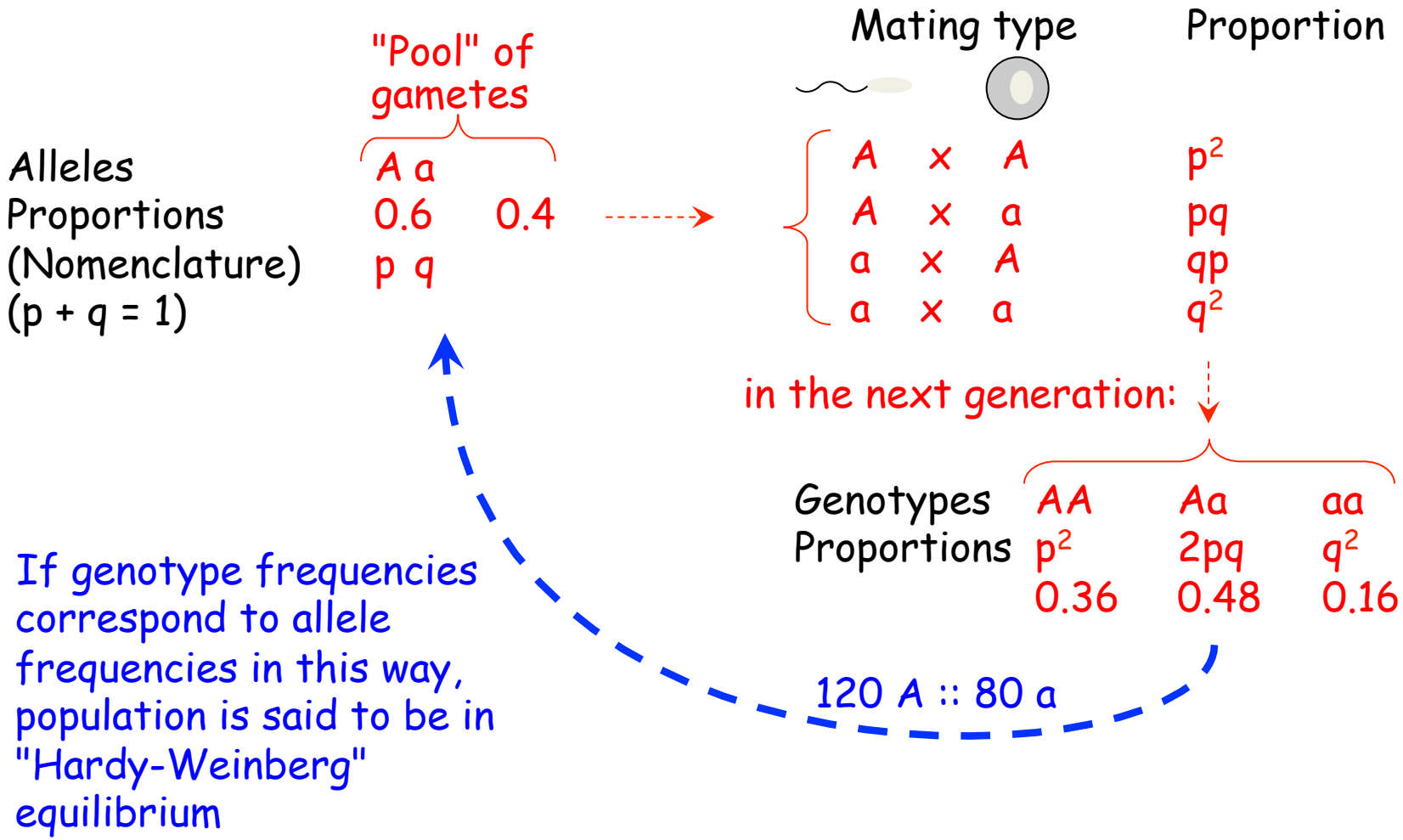
For a gene with alleles A and a, in a population with 100 individuals…

Genotypes	AA	Aa	aa
	40	40	20
Genotype proportions	0.4	0.4	0.2

Alleles					
A	80	40	0	= 120	0.6
a	0	40	40	= 80	0.4
				Allele proportions	

What proportion of genotypes will be produced by this population?

From allele frequencies to genotype frequencies…



- If genotype frequencies correspond to allele frequencies in this way, population is said to be in "Hardy-Weinberg" equilibrium

Take home message: Hardy-Weinberg can be used to estimate allele frequencies from genotype frequencies, and vice versa; lots of assumptions, but works pretty well for most situations in humans.

Population genetics and genomics

1920's: Fisher, Wright, Haldane—theoretical synthesis of Mendelian principles, changes in allele frequencies over time, and quantitative variation

1970's: Kimura—development of neutral theory (most variation in most populations is due to drift)

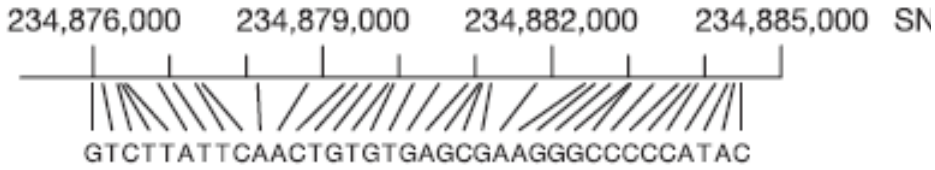
1980's: development and application of coalescent theory to population genetics and molecular evolution

2003: Human genome reference

2005: HapMap—1,000,000 SNPs genotyped on ~150 individuals: Yoruba in Ibadan, Nigeria (YRI)
Utah, USA (CEU)
Han Chinese in Beijing, China (CHB)
Japanese in Tokyo, Japan (JPT)

2012-present: 1000genomes (or more)

Typical HapMap result



9 kb region on chr 2 in which 36 SNPs discovered by resequencing

Typed on 60 CEU individuals (no recombinants)

Colored circles represent presence of minor allele

Q: How many arrangements are possible?

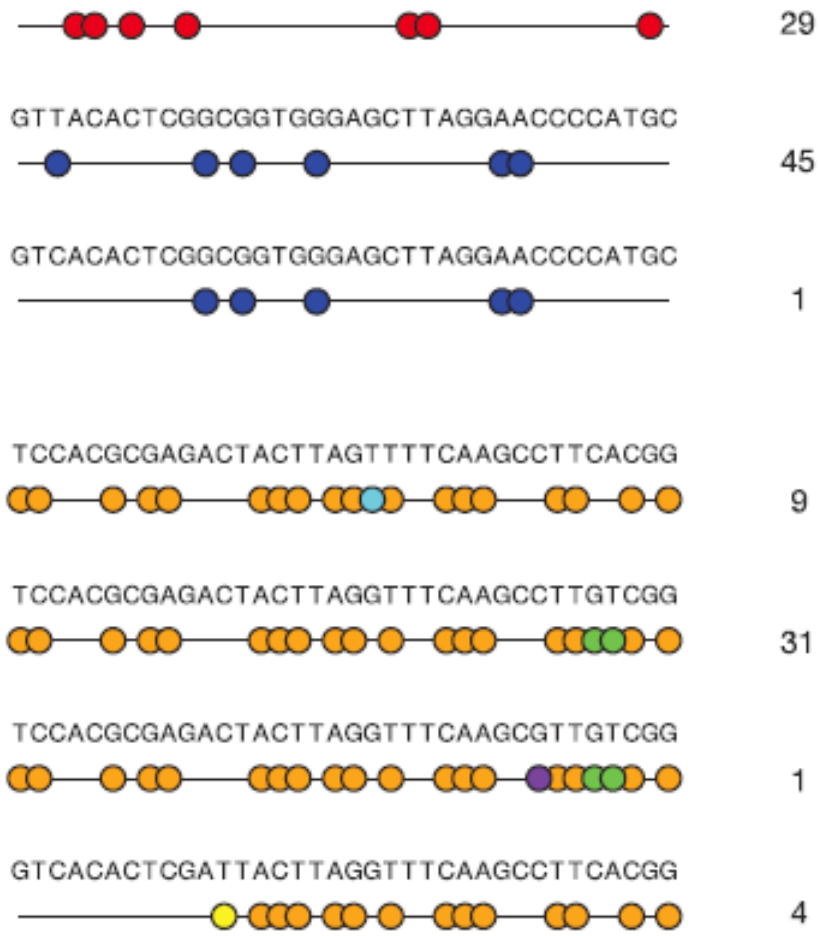
A: 2^{36} = 68719476736

Q: How many arrangements were there?

A: 7

No. of **haplotypes**

Surprisingly simple haplotype structure of the human genome



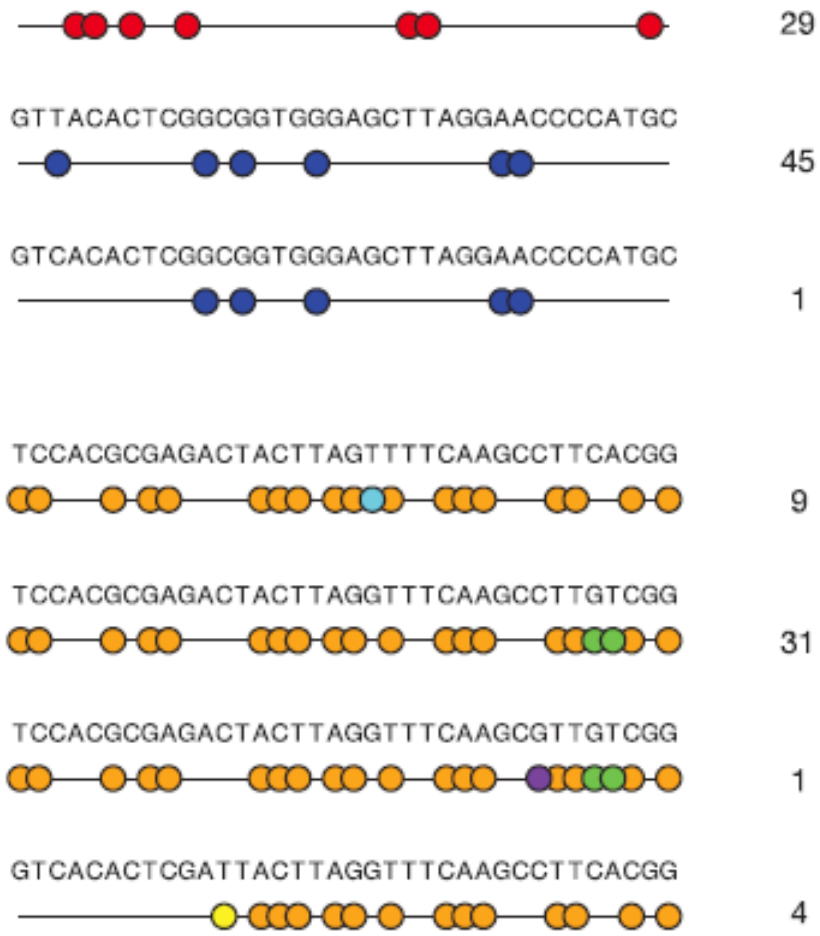
- Different haplotypes related through genealogical tree
  (Typical finding: small (5 – 15 kb regions of the human genome show diversity due to mutation rather than recombination)
- Simple structure allows haplotype identity assignment with 4 SNPs

Haplotype "blocks"

(correlated SNPs exist due to recombination, or lack thereof)

Parameter	YRI	CEU	CHB
Length (kb)	7.3	16.3	13.2
% of genome	67	87	81
No. haplotypes	5.57	4.66	4.01

Surprisingly simple haplotype structure of the human genome

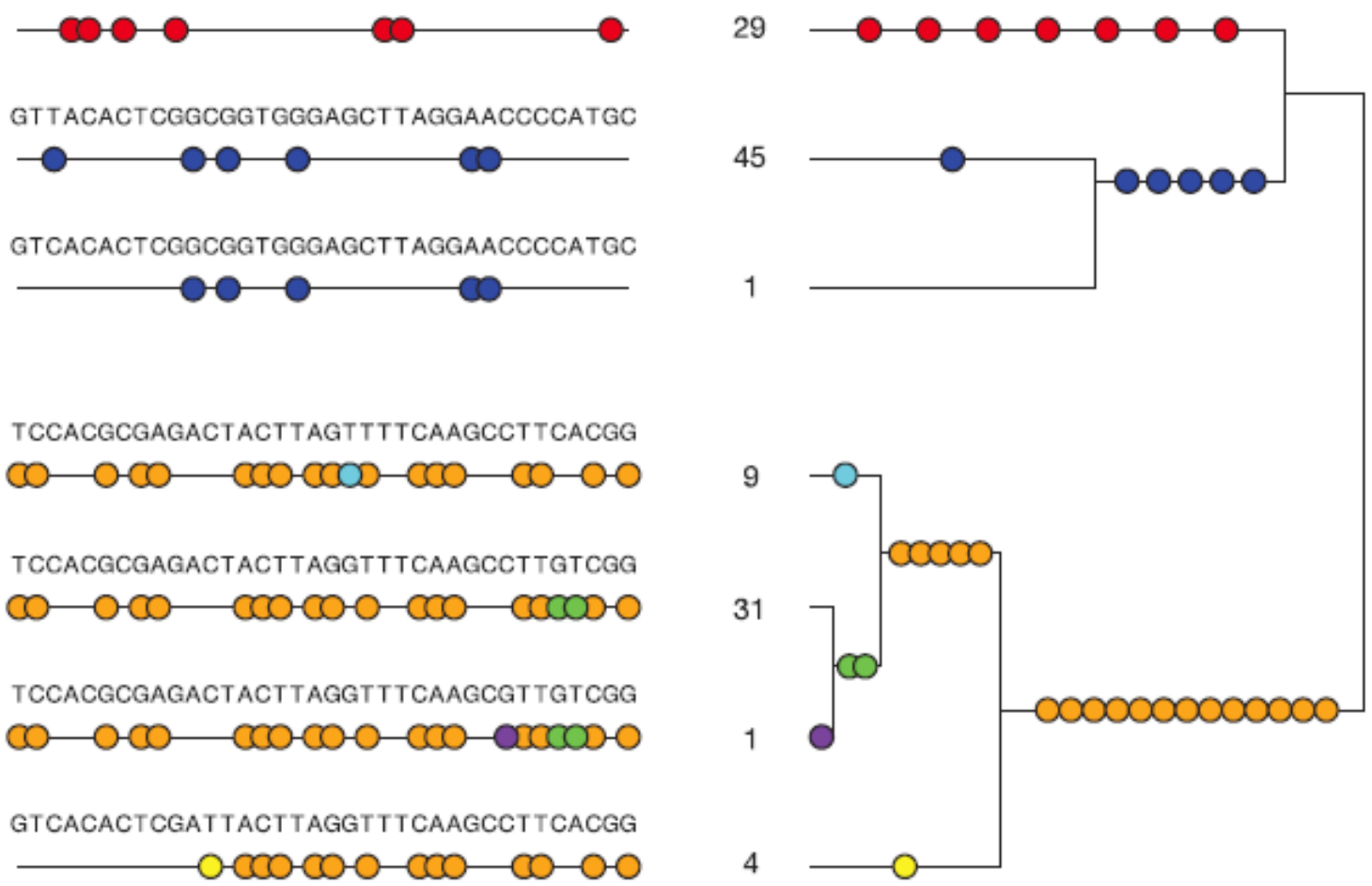


120 possibilities

but only 7 groups

- Different haplotypes related through genealogical tree
  (Typical finding: small (5 – 15 kb regions of the human genome show diversity due to mutation rather than recombination)

Surprisingly simple haplotype structure of the human genome



2012:

1092 individuals from 14 populations

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

2015:

2054 individuals from 26 populations (+ SVs)

A global reference for human genetic variation

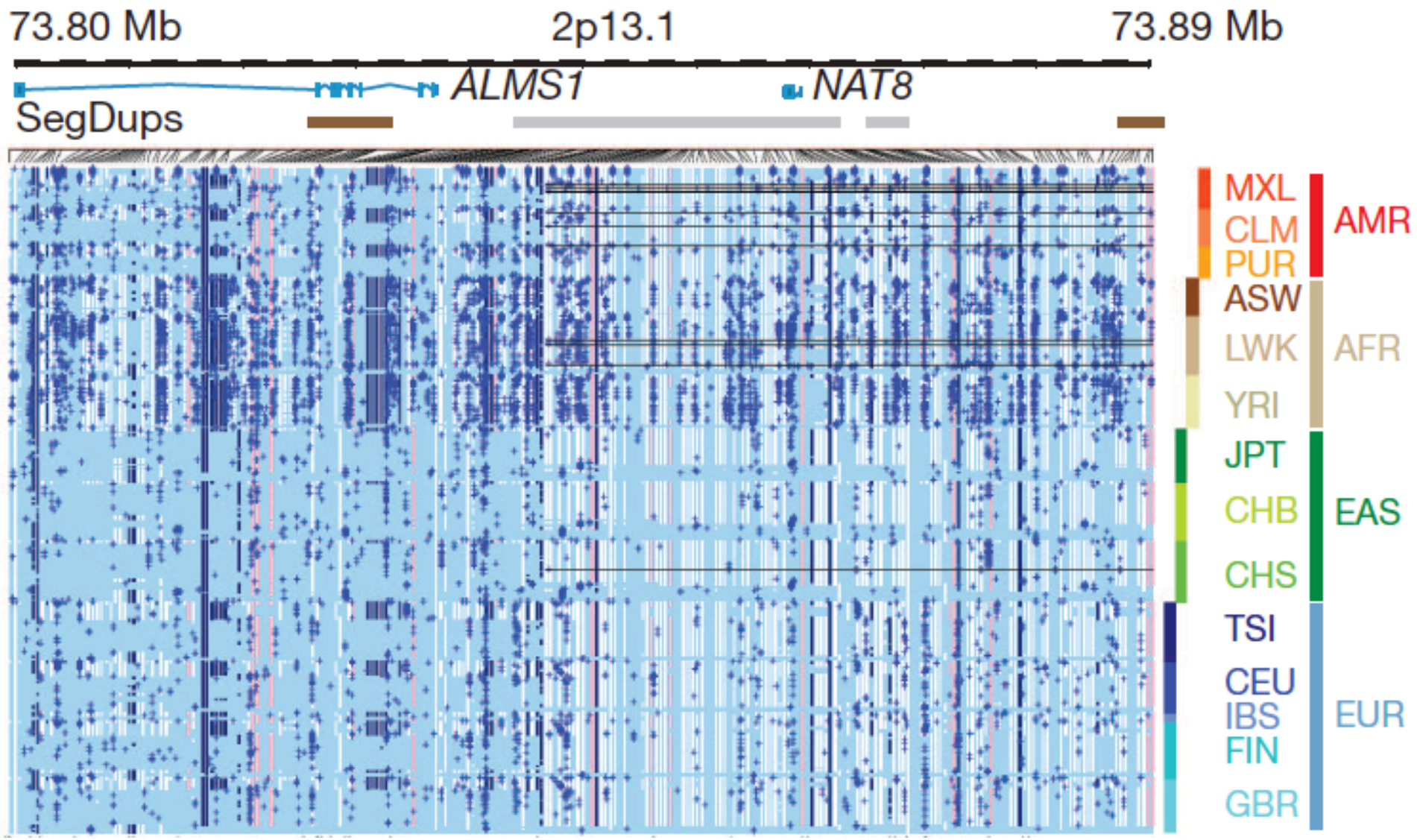
The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

Table 1 | Median autosomal variant sites per genome

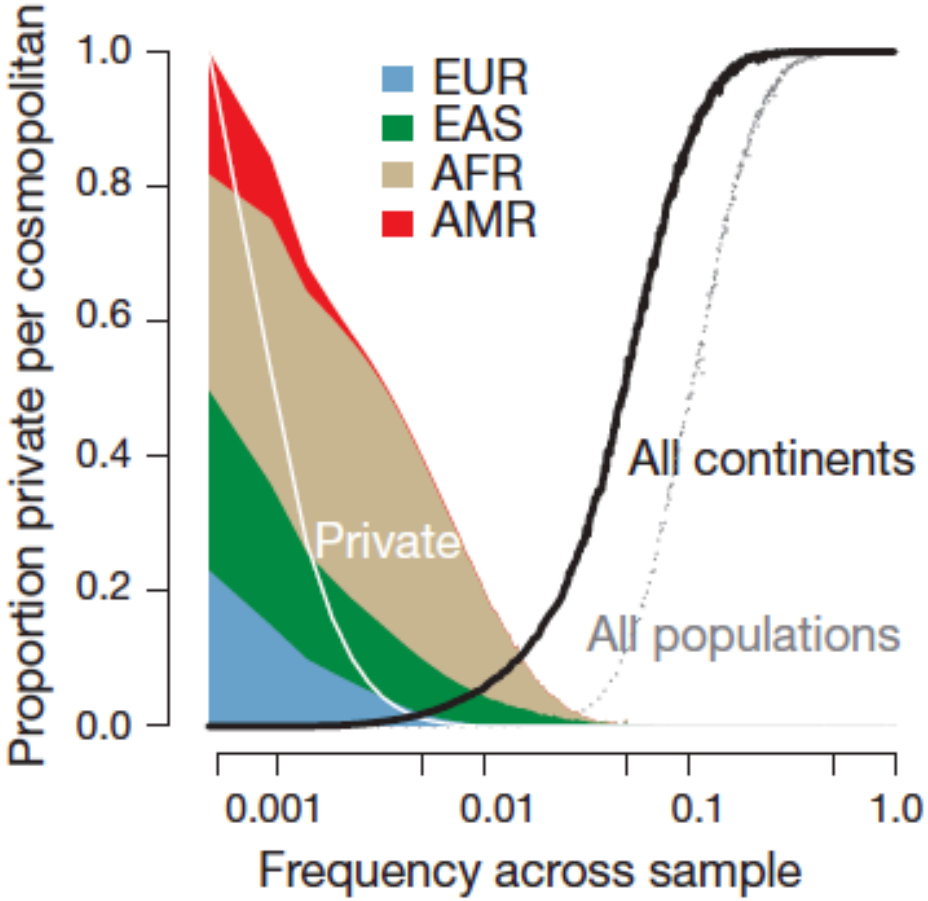
	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

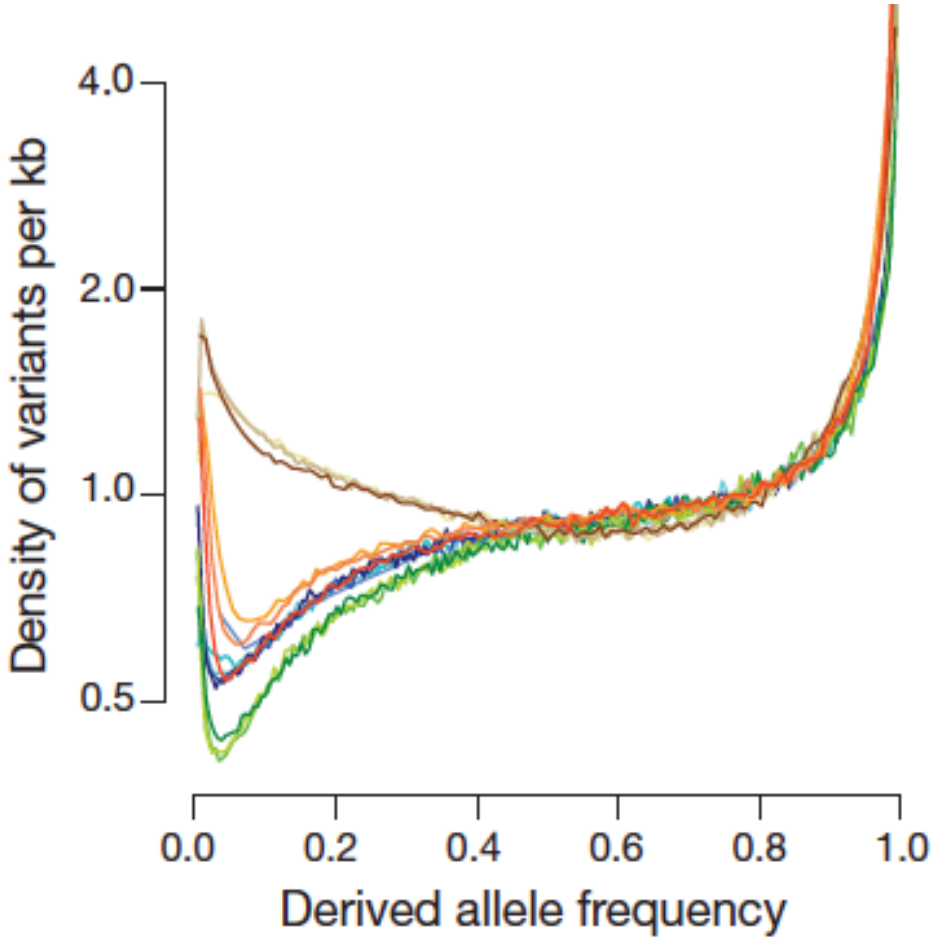
See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.



100 kb region: each row is a haplotype.

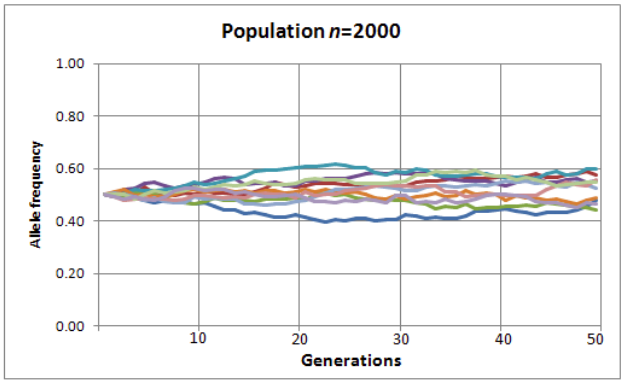
- Common (pink and white) and rare (<0.5%, dark blue) variants; deletions (black)
- (similar picture to HapMap)

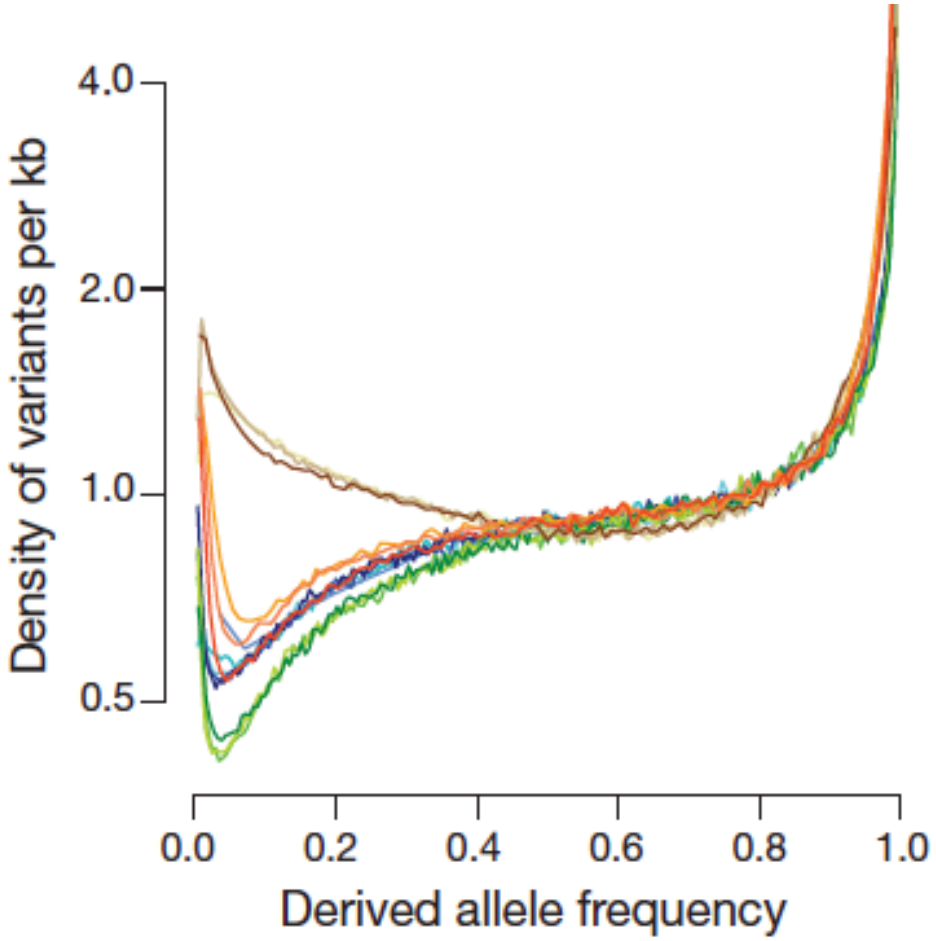




- Most variants >5% are found in all populations
- Most variants <1% are found in a single population (or individual)

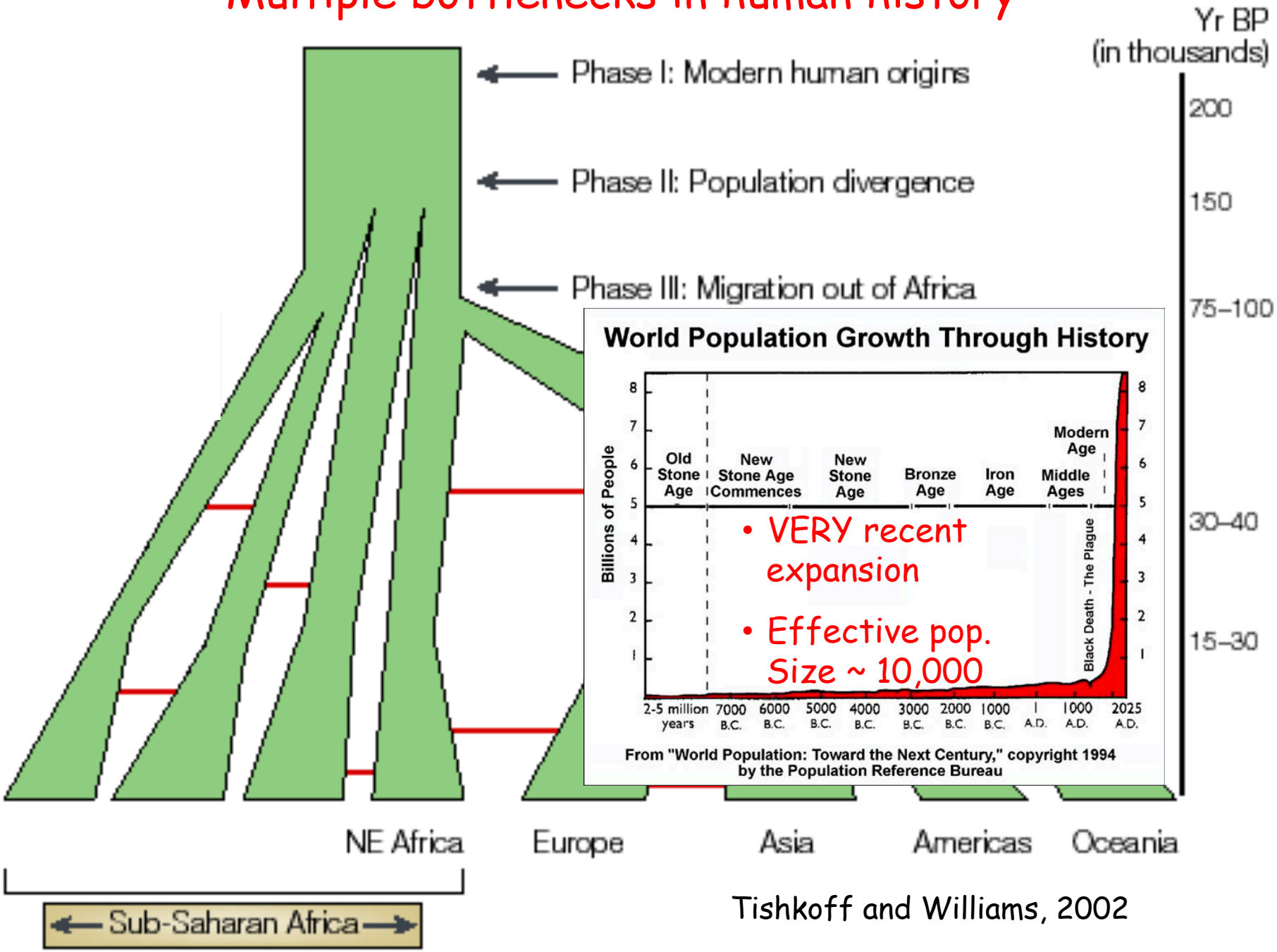
- ASW, LWK, and YRI -> excess of very rare variants
- All populations -> unusual allele frequency distribution





- ASW, LWK, and YRI -> excess of very rare variants
- All populations -> unusual allele frequency distribution

Multiple bottlenecks in human history



- VERY recent expansion
- Effective pop. Size ~ 10,000

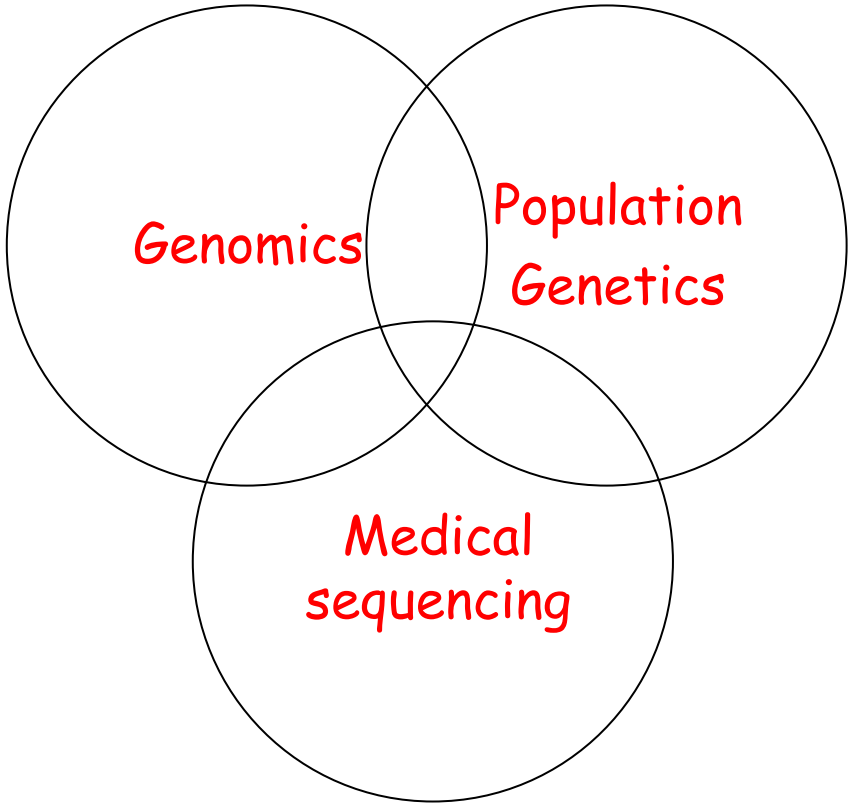
Tishkoff and Williams, 2002

Variation

Function

Evolution

Technology



Allele frequencies

Drift/selection

Population history

Phenotype -> WGS -> Causal variation