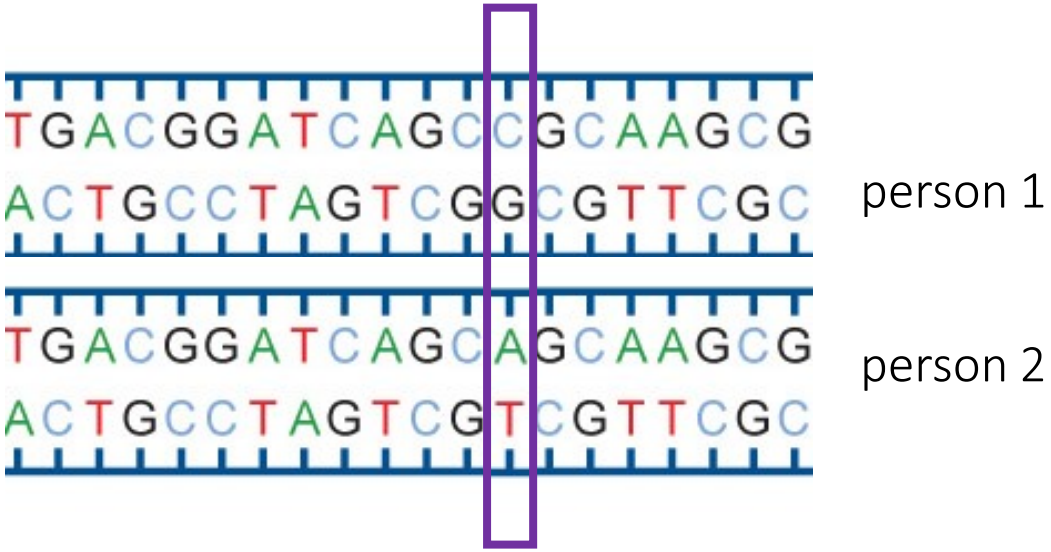
Introduction to common variation, quality control, GWAS, and PLINK (Part I)

Lucia Colodro Conde and Katrina Grasby

Introduction to common variation

adenine (A), thymine (T), cytosine (C), guanine (G)



Genetic variation: differences in the sequence of DNA among individuals.

Mutation: a newly arisen variant

Genetic variant: any specific region of the genome which differs between two genomes.

Allele: version of a variant

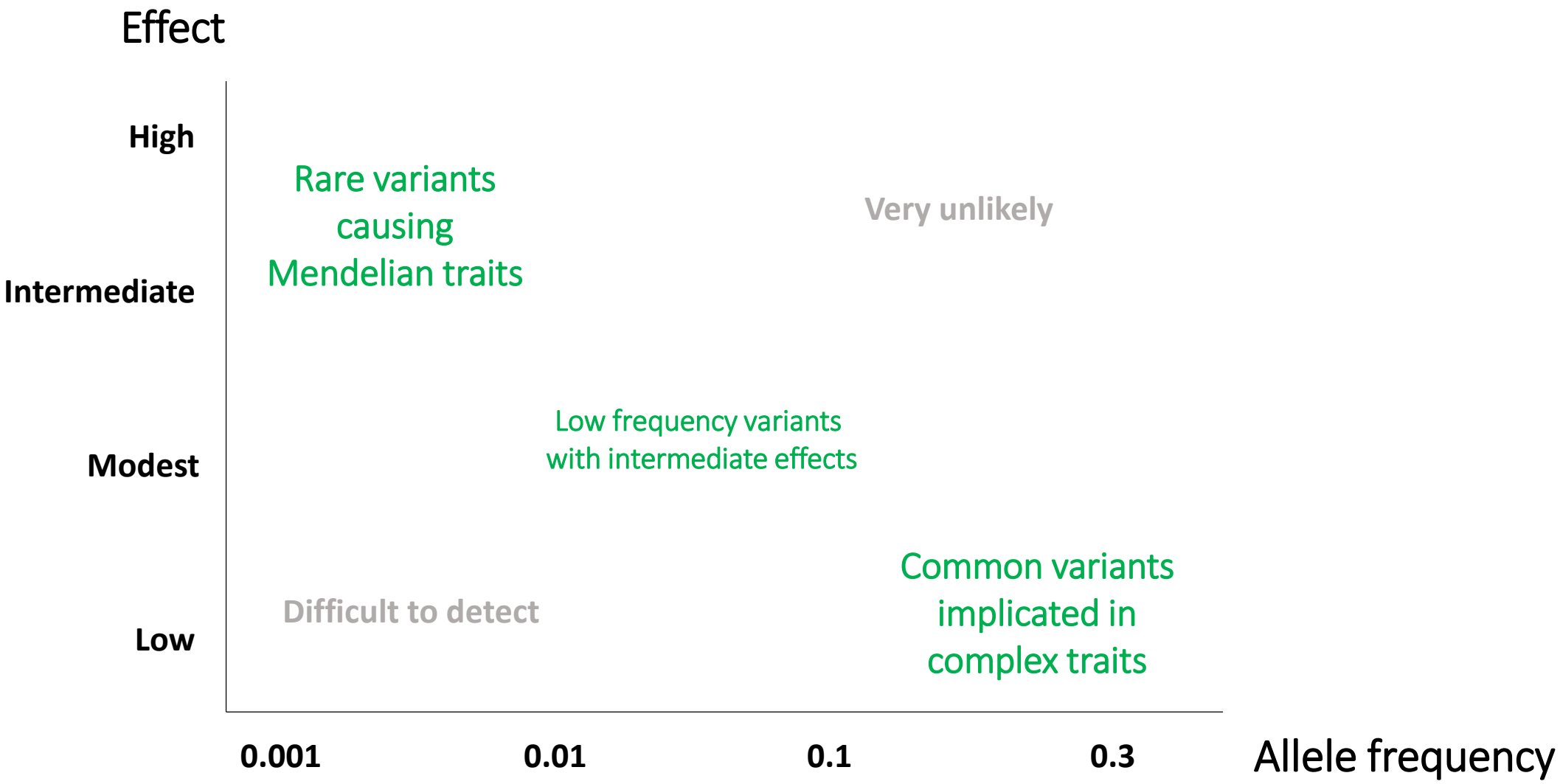
Allele frequency: incidence of an allele in a population.

Minor allele frequency (MAF): frequency at which the less common allele occurs in a given population.

Rare variant: a genetic variant present in < 1% of the alleles in the population

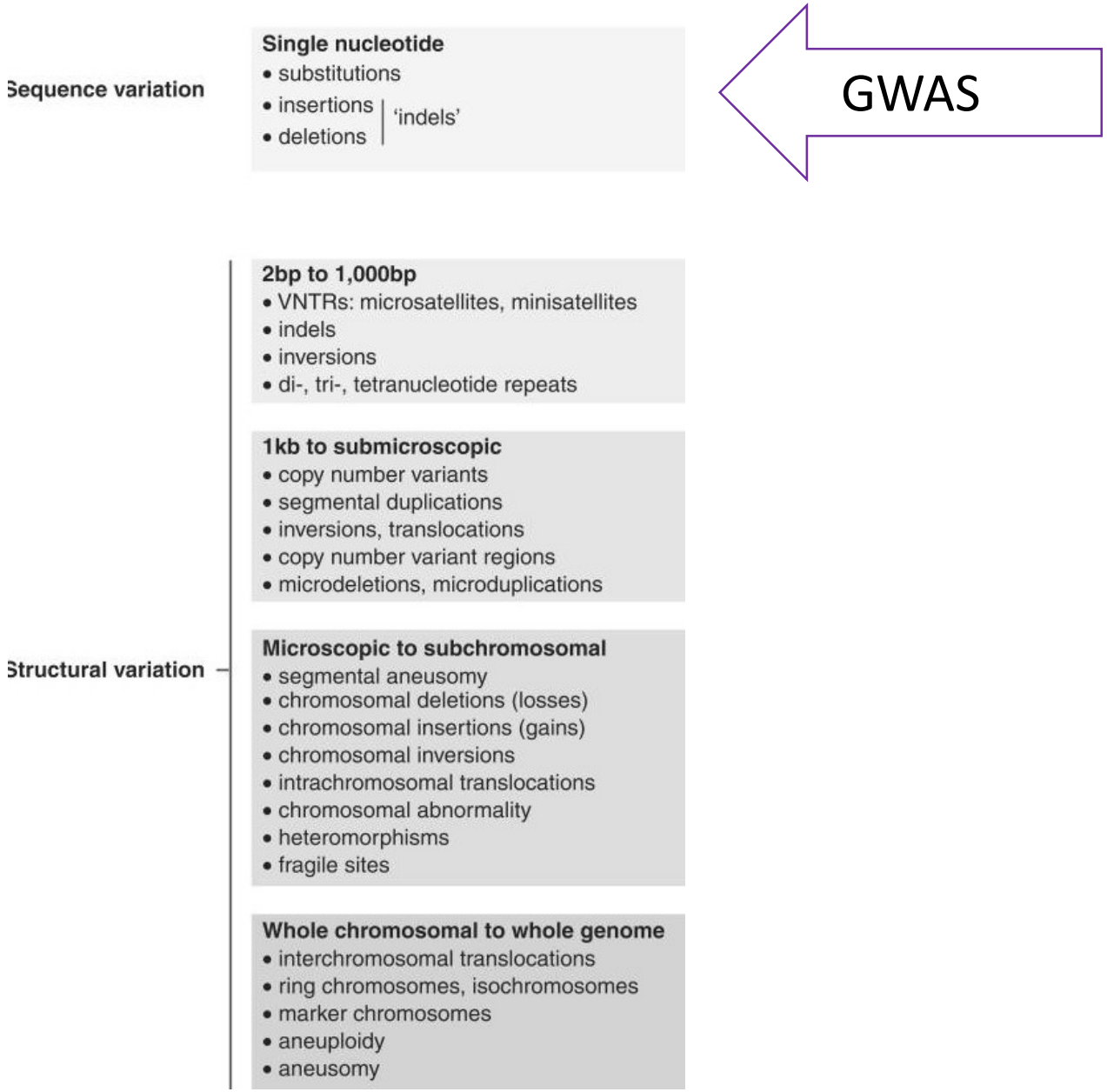
Common variant: a genetic variant present in >= 1% of the alleles in the population

Note 1% is arbitrary



Based on McCarthy et al (2008) Nature Reviews Genetics 9,

Examples of genetic variation



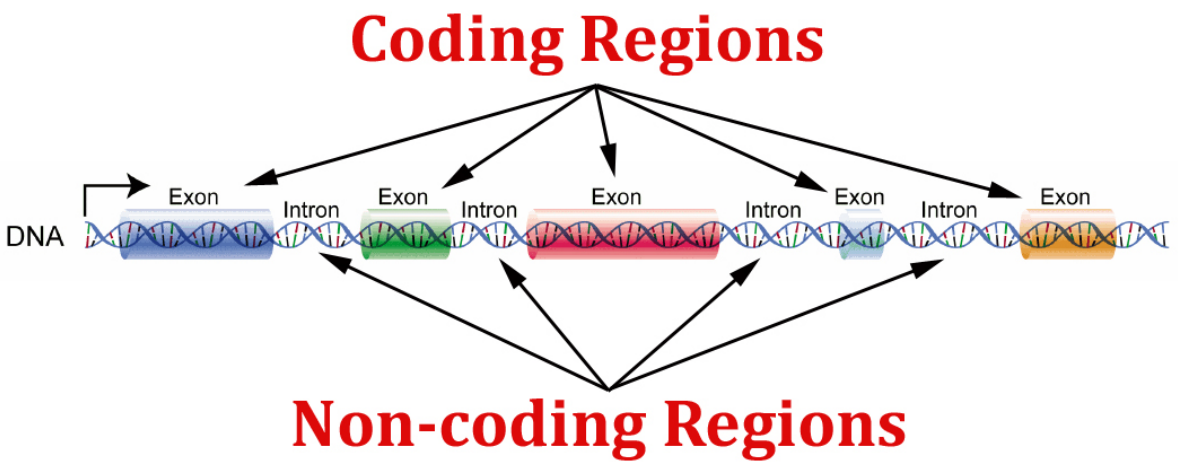
Knight JC (2009). Genetics and the general physician: insights, applications and future challenge. *QJM*..
Sherer SW et al (2007). Challenges and standards in integrating surveys of structural variation. *Nat Genet*

SNP (single nucleotide polymorphism):

variation at a single base pair in a DNA sequence among individuals.

	Chrom.	DNA sequence	Genotype
Person 1	Mat	GTAACTTGGGATCT**A**GACCAATAGAT	**A A**
	Pat	GTAACTTGGGATCT**A**GACCAATAGAT	
Person 2	Mat	GTAACTTGGGATCT**A**GACCAATAGAT	**A C**
	Pat	GTAACTTGGGATCT**C**GACCAATAGAT	
Person 3	Mat	GTAACTTGGGATCT**C**GACCAATAGAT	**C C**
	Pat	GTAACTTGGGATCT**C**GACCAATAGAT	

Slide based on Manuel Ferreira's



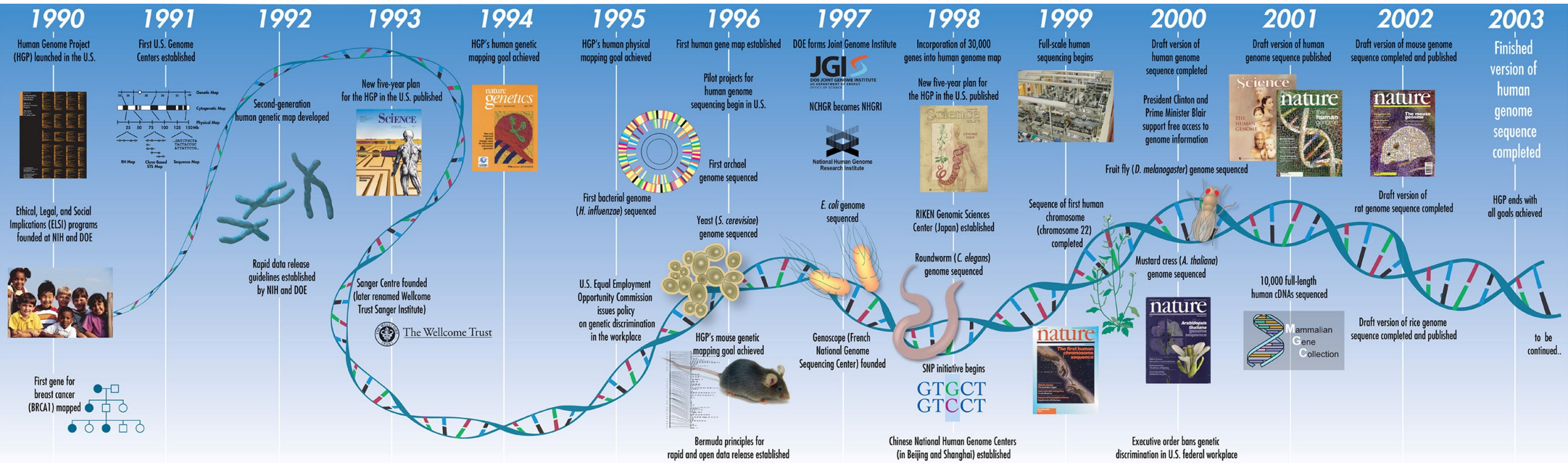
TYPES OF SNPs

- **Non-coding region**
  - They may affect regulation of genes in coding regions
- **Coding region**
  - **Synonymous**
    - They do not change the amino acid sequence
  - **Non-Synonymous**
    - They change the amino acid sequence
    - **Missense**
      - They change one amino acid in a protein (which may have an effect in the protein)
    - **Nonsense**
      - They produce a premature stop codon, that results in an incomplete and usually non-functional protein product

Insertion–deletion variants (indels):

one or more base pairs are present in some genomes but absent in others in relation to the reference

	Chrom.	DNA sequence	Genotype
Person 1	Mat	GTAACTTGGGATCT**GAT**GACCAGATAG	**R D**
	Pat	GTAACTTGGGATCT---GACCAGATAG	
Person 2	Mat	GTAACTTGGGATCT**GAT**GACCAGATAG	**R R**
	Pat	GTAACTTGGGATCT**GAT**GACCAGATAG	
Person 3	Mat	GTAACTTGGGATCT---GACCAGATAG	**D D**
	Pat	GTAACTTGGGATCT---GACCAGATAG	



Collins et al 2003, A vision for the future of genomics research, *Nature*

HapMap (haplotype map) Project

270 samples:

30 parent-offspring trios of the Yoruba from Ibadan, Nigeria (YRI)
30 trios of Utah residents with European ancestry (CEU)
45 individuals from Beijing, China (CHB)
45 individuals from Tokyo, Japan (JPT)

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*.

1000 Genomes Project

Phase 1: 1,092 individuals from 14 populations..

Phase 3: 2,504 individuals from 26 populations (~500 samples form each 5 continental ancestry groups, with ~5 populations for each group)

Population		Code	Population Color	Continental Group Color	Analysis Panel	Phase 1	Phase 3
African ancestry							
Esan in Nigeria	Esan	ESN			AFR		99
Gambian in Western Division, Mandinka	Gambian	GWD			AFR		113
Luhya in Webuye, Kenya	Luhya	LWK			AFR	97	99
Mende in Sierra Leone	Mende	MSL			AFR		85
Yoruba in Ibadan, Nigeria	Yoruba	YRI			AFR	88	108
African Caribbean in Barbados	Barbadian	ACB			AFR/AMR		96
People with African Ancestry in Southwest USA	African-American SW	ASW			AFR/AMR	61	61
Americas							
Colombians in Medellin, Colombia	Colombian	CLM			AMR	60	94
People with Mexican Ancestry in Los Angeles, CA, USA	Mexican-American	MXL			AMR	66	64
Peruvians in Lima, Peru	Peruvian	PEL			AMR		85
Puerto Ricans in Puerto Rico	Puerto Rican	PUR			AMR	55	104
East Asian ancestry							
Chinese Dai in Xishuangbanna, China	Dai Chinese	CDX			EAS		93
Han Chinese in Beijing, China	Han Chinese	CHB			EAS	97	103
Southern Han Chinese	Southern Han Chinese	CHS			EAS	100	105
Japanese in Tokyo, Japan	Japanese	JPT			EAS	89	104
Kinh in Ho Chi Minh City, Vietnam	Kinh Vietnamese	KHV			EAS		99
European ancestry							
Utah residents (CEPH) with Northern and Western European ancestry	CEPH	CEU			EUR	85	99
British in England and Scotland	British	GBR			EUR	89	91
Finnish in Finland	Finnish	FIN			EUR	93	99
Iberian Populations in Spain	Spanish	IBS			EUR	14	107
Toscani in Italia	Tuscan	TSI			EUR	98	107
South Asian ancestry							
Bengali in Bangladesh	Bengali	BEB			SAS		86
Gujarati Indians in Houston, TX, USA	Gujarati	GIH			SAS		103
Indian Telugu in the UK	Telugu	ITU			SAS		102
Punjabi in Lahore, Pakistan	Punjabi	PJL			SAS		96
Sri Lankan Tamil in the UK	Tamil	STU			SAS		102
Total						**1092**	**2504**

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*.
The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*.

The Haplotype Reference Consortium (HRC)

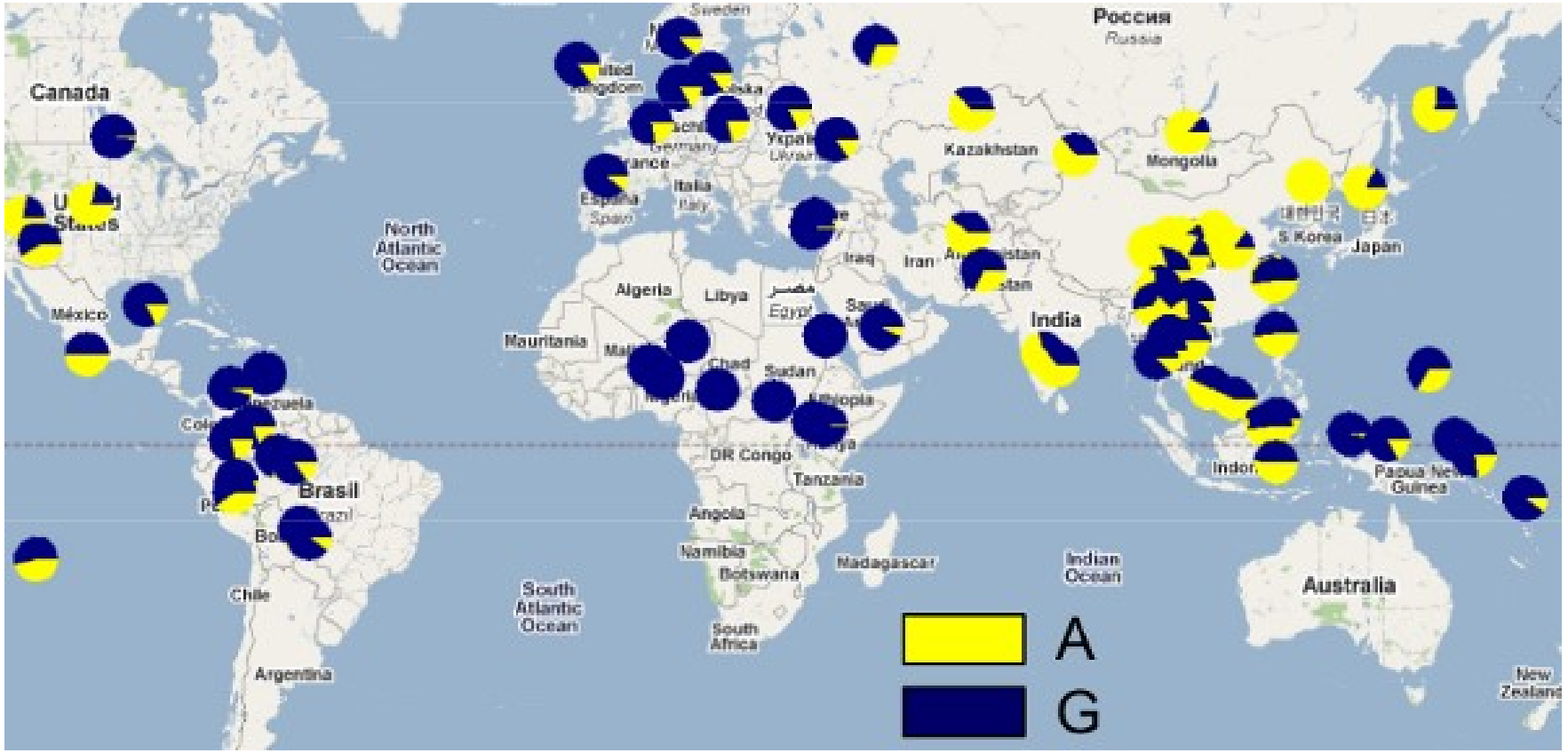
nature genetics

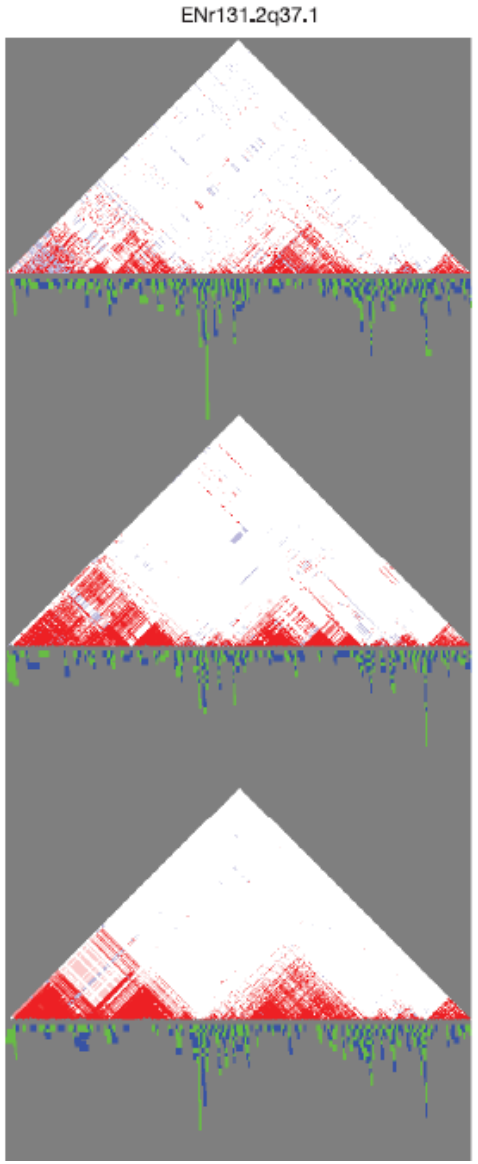
A reference panel of 64,976 haplotypes for genotype imputation

The Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*.

These projects have provided information on:

→ Patterns of human common genetic variation
→ Linkage disequilibrium (LD) and allele frequencies differences in populations
→ Tag SNPS for the design of SNP arrays to facilitate imputation and GWAS





Practical. Where to start with.

GWAS have been facilitated by the development of relatively inexpensive SNP arrays.

How do we make the information provided by SNP arrays usable?