GENETIC DATA

Sources of Population Genetic Data

Phenotype	Mendel's peas
	Blood groups
Protein	Allozymes
	Amino acid sequences
DNA	Restriction sites, RFLPs
	Length variants: VNTRs, STRs
	Single nucleotide polymorphisms, SNPs
	Single nucleotide variants, SNVs

Mendel's Data

Dominant Form		Recessive Form	
	Seed characters		
5474	Round	1850	Wrinkled
6022	Yellow	2001	Green
	Plant characters		
705	Grey-brown	224	White
882	Simply inflated	299	Constricted
428	Green	152	Yellow
651	Axial	207	Terminal
787	Long	277	Short

Genetic Data

Human ABO blood groups discovered in 1900.

Elaborate mathematical theories constructed by Sewall Wright, R.A. Fisher, J.B.S. Haldane and others. This theory was challenged by data from new data from electrophoretic methods in the 1960's:

"For many years population genetics was an immensely rich and powerful theory with virtually no suitable facts on which to operate. ... Quite suddenly the situation has changed. The motherlode has been tapped and facts in profusion have been pored into the hoppers of this theory machine. ... The entire relationship between the theory and the facts needs to be reconsidered."

Lewontin RC. 1974. The Genetic Basis of Evolutionary Change. Columbia University Press.

STR markers: CTT set

http://www.cstl.nist.gov/biotech/strbase/seq_info.htm

Locus	Structure	Chromosome	Usual No. of repeats
CSF1PO	$[AGAT]_n$	5q	6–16
TPOX	$[AATG]_n$	2p	5–14
TH01*	$[AATG]_n$	11p	3–14

* "9.3" is $[AATG]_6ATG[AATG]_3$

Length variants detected by capillary electrophoresis.

"CTT" Data - Forensic Frequency Database

CSF1P0		TPOX		TH01	
11	12	8	11	7	8
11	13	8	8	6	7
11	12	8	11	6	7
10	12	8	8	6	9
11	12	8	12	9	9.3
10	12	9	11	6	7
10	13	8	11	6	6
11	12	8	8	6	9.3
9	10	8	9	7	9.3
11	12	8	8	6	8
11	13	8	11	7	9
11	12	8	11	6	9.3
10	11	8	8	7	9.3
10	10	8	11	7	9.3
9	10	8	8	6	9.3
11	12	9	11	9	9.3
9	11	9	11	9	9.3
11	12	8	8	6	7
10	10	9	11	6	9.3
10	13	8	8	8	9.3

Sequencing of STR Alleles

"STR typing in forensic genetics has been performed traditionally using capillary electrophoresis (CE). Massively parallel sequencing (MPS) has been considered a viable technology in recent years allowing high-throughput coverage at a relatively affordable price. Some of the CE-based limitations may be overcome with the application of MPS ... generate reliable STR profiles at a sensitivity level that competes with current widely used CE-based method."

Zeng XP, et al. 2015. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Science International: Genetics 16:38-47.

Single Nucleotide Polymorphisms (SNPs)

"Single nucleotide polymorphisms (SNPs) are the most frequently occurring genetic variation in the human genome, with the total number of SNPs reported in public SNP databases currently exceeding 9 million. SNPs are important markers in many studies that link sequence variations to phenotypic changes; such studies are expected to advance the understanding of human physiology and elucidate the molecular bases of diseases. For this reason, over the past several years a great deal of effort has been devoted to developing accurate, rapid, and cost-effective technologies for SNP analysis, yielding a large number of distinct approaches. "

Kim S. Misra A. 2007. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng. 2007;9:289-320.

AMD SNP Data

SNP	Individual														
rs6424140	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
rs1496555	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2
rs1338382	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
rs10492936	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs10489589	3	1	1	1	2	2	1	2	1	1	1	3	1	1	1
rs10489588	3	1	1	1	2	2	1	2	1	1	1	3	1	1	1
rs4472706	1	3	3	3	2	2	3	2	3	3	3	1	3	3	3
rs4587514	3	3	3	3	3	2	2	3	2	2	2	3	3	1	3
rs10492941	3	3	3	3	3	3	3	3	2	3	3	2	3	3	1
rs1112213	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs4648462	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs2455122	2	1	1	0	1	2	1	1	1	1	1	1	1	1	2
rs2455124	2	1	1	2	1	2	1	1	1	1	1	1	1	1	2
rs10492940	2	1	1	1	1	2	1	2	1	1	1	2	1	1	2
rs10492939	1	2	1	1	1	1	3	2	1	2	3	2	2	1	1
rs10492938	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs10492937	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3
rs7546189	1	2	3	3	1	3	2	2	3	3	2	2	2	2	2
rs1128474	3	2	3	2	3	3	2	3	3	3	3	3	2	1	3

Genotype key: 0 −; 1 AA; 2 AB; 3 BB.

Phase 3 1000Genomes Data

- 84.4 million variants
- 2504 individuals
- 26 populations

https://www.internationalgenome.org/data

Whole-genome Sequence Studies

Largest amount of sequence data currently is from the NHLBI Trans-Omics for Precision Medicine (TOPMed) project. www.nhlbiwgs.

For data freeze 9 of this study:

158,470 genomes.
843 million genetic variants; 781m SNVs and 62m indels.
46.4% of SNVs are singletons; 49.7% of indels are singletons.
3.4-4.5 million variants per genome.
1,000-15,000 singletons per genome.