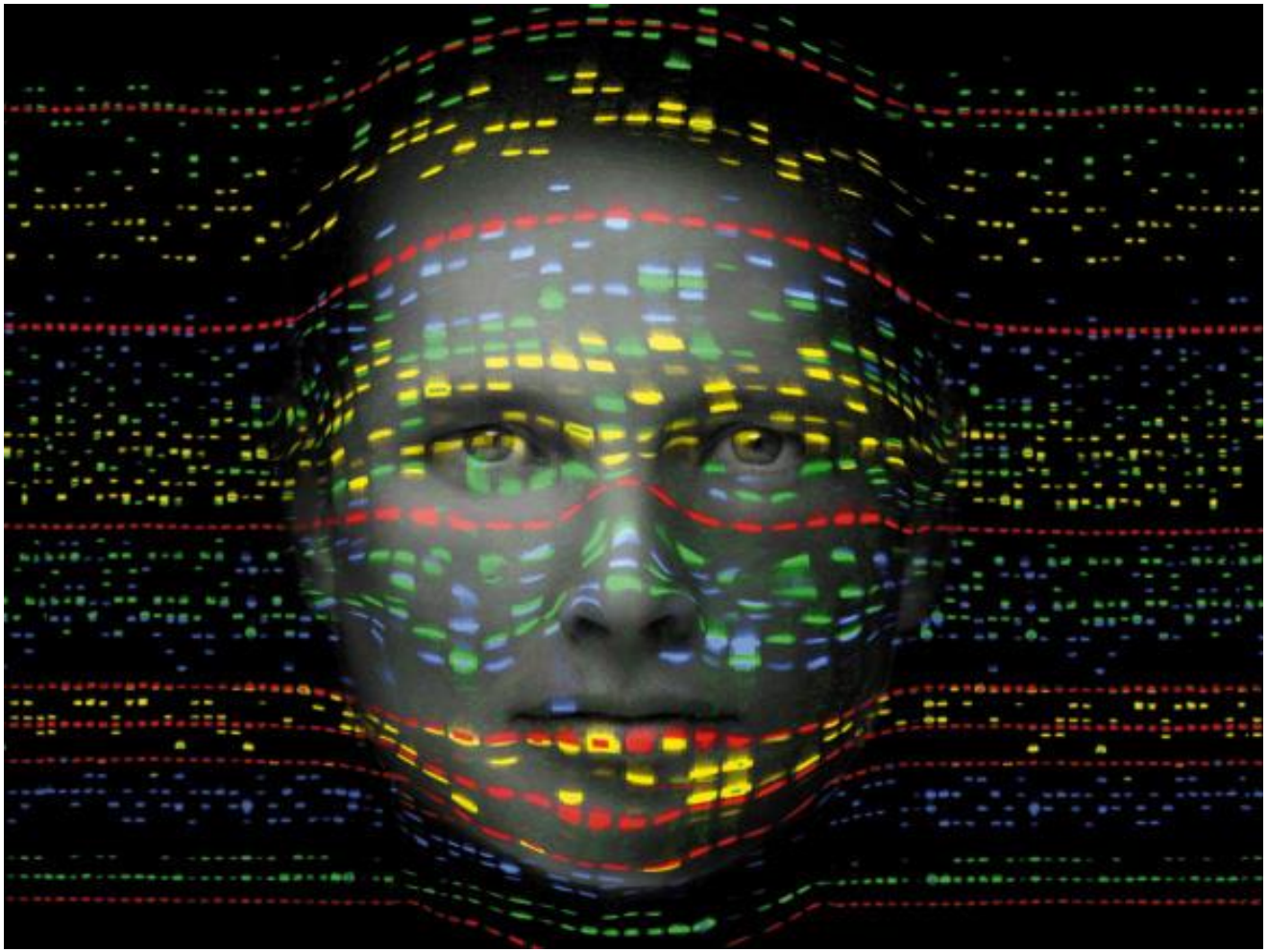
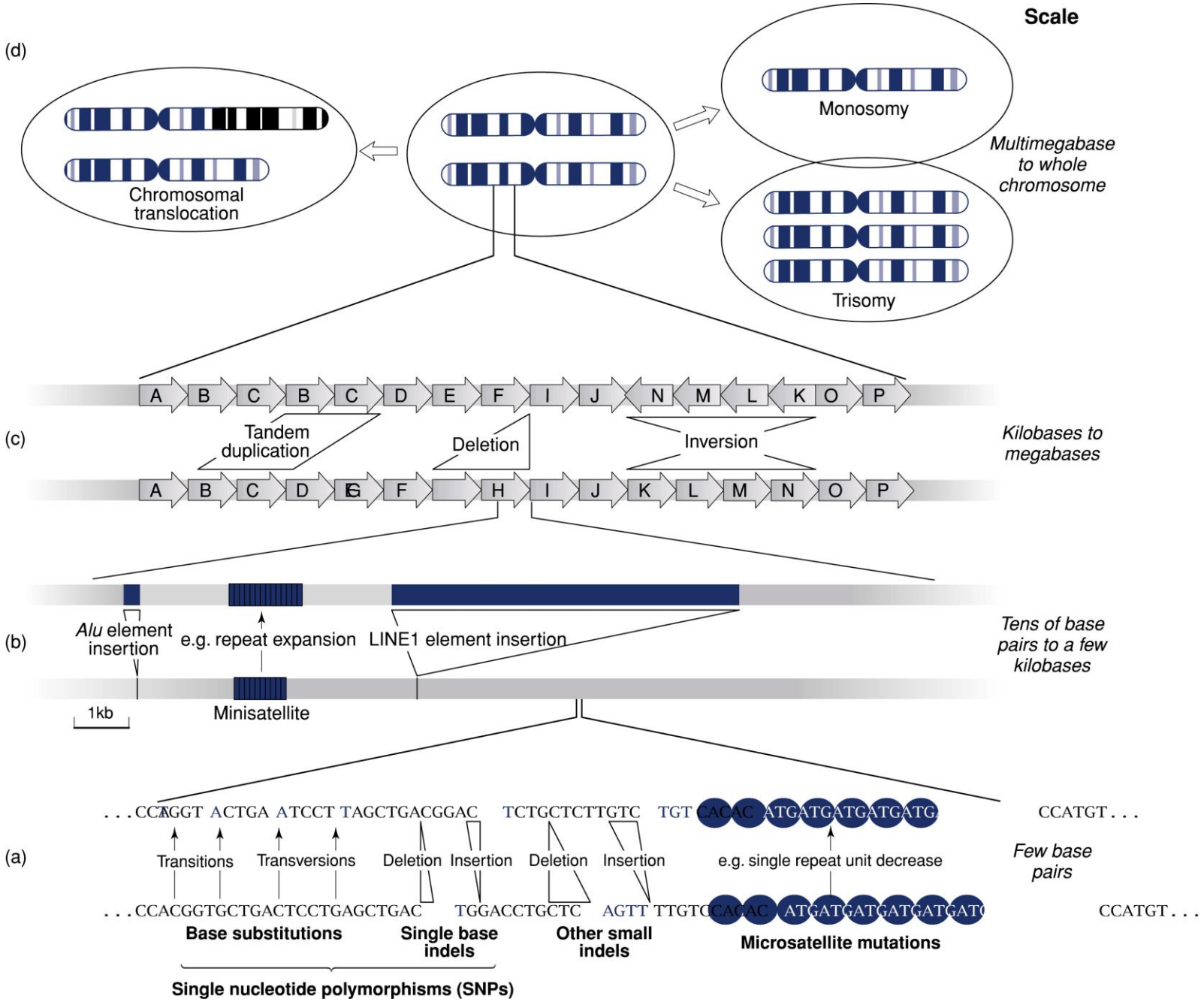
GENOMICS

Variations in the genomes: Genetic variability and phenotype



Department of Genetics, Eötvös Loránd University



Human Evolutionary Genetics, Jobling, 2004

Current Biology 19, 1453–1457, September 15, 2009 ©2009 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2009.07.032

Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree

Y chromosome resequencing:

Illumina

Forensic Science International: Genetics 4 (2010) 59–61

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Review

The hare and the tortoise: One small step for four SNPs, one giant leap for SNP-kind

Yali Xue, Chris Tyler-Smith *

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, UK

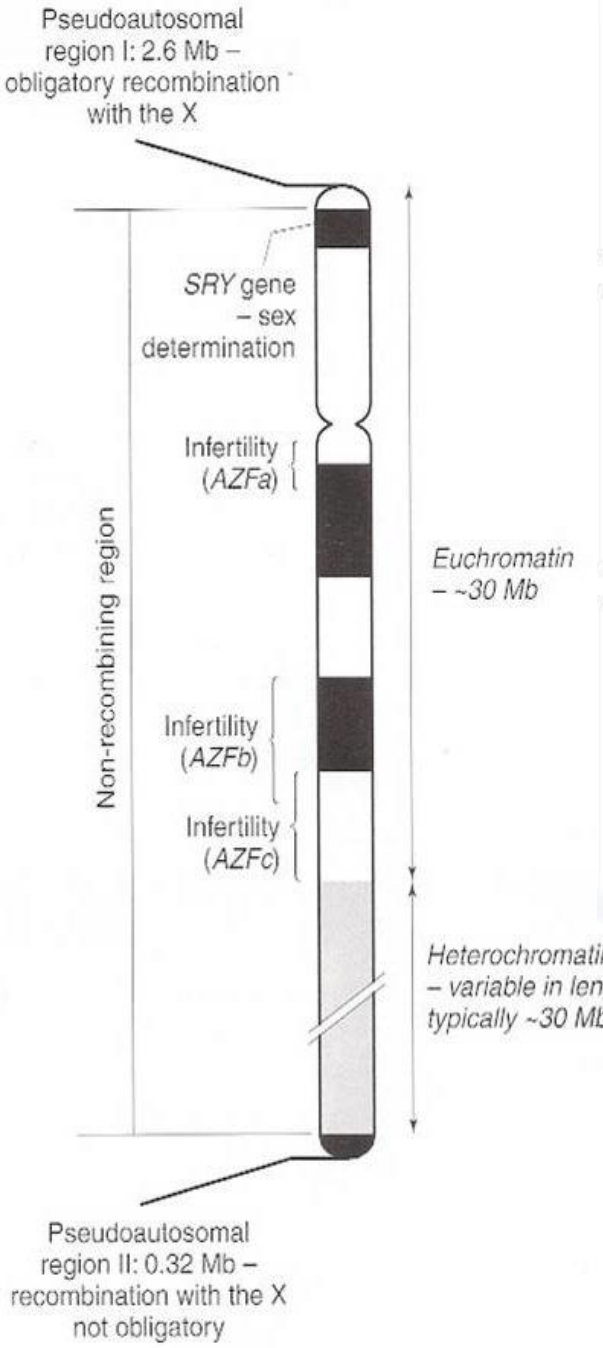
ARTICLE INFO

Article history:
Received 31 July 2009
Accepted 6 August 2009

Keywords:
Next-gen sequencing
Y-SNP
Y-STR
Haplotype resolution
Forensic applications

ABSTRACT

A recently published study has used next-gen sequencing technology to resequence two Y chromosomes separated by 13 generations and discovered four single-base differences in ~10 Mb DNA, suggesting that the Y chromosome euchromatin accumulates around one mutation per generation. Y-SNPs therefore now offer the best resolution of Y haplotypes and promise to distinguish almost every Y chromosome. This work illustrates the promise of current sequencing technology for forensically relevant applications.



Sample Number	M175	M5	M122	PN31	LLY22G	M134	M7	M113	M121	M159	M164	B_DYS388I	B_DYS389II	B_DYS390	B_DYS456	G_DYS19	G_DYS385a	G_DYS385b	G_DYS458	R_DYS437	R_DYS438	R_DYS448	R_Y_GATA_H4	Y_DYS391	Y_DYS392	Y_DYS393	Y_DYS439	Y_DYS635
66	A(1)	G(0)	G(1)	T(0)	C(0)	C(0)	G(0)	T(0)	A(0)	T(0)	A(0)	14	30	23	15	15	12	21	18	14	10	19	11	11	13	12	11	22
101	A(1)	G(0)	G(1)	T(0)	C(0)	C(0)	G(0)	T(0)	A(0)	T(0)	A(0)	14	30	23	15	15	12	21	18	14	10	19	11	11	13	12	11	22

Sample number	DYS472 (B)	DYS508 (G)	DYS487 (Y)	DYS570 (G)	DYS 583(B)	DYS579 (G)	DYS525 (Y)	DYS531 (B)	DYS488 (G)	DYS589 (Y)	DYS578(B)	DYS580(G1)	DYS636(Y)	DYS590(B)	DYS533(G1)	DYS617(Y)	DYS594(G2)	DYS505(B)	DYS641(G)	DYS638(Y)	DYS476(B)	DYS492(G)	DYS540(Y)	DYS537	DYS405S1(G)	DYS568(Y)	DYS459(B)	DYS572(G)
66	8	12	14	18	9	9	10	13	13	13	8	10	10	8	12	16	12	11	7	12	11	12	12	11	11	11	8	12
101	8	12	14	18	9	9	10	13	13	13	8	10	10	8	12	16	12	11	7	12	11	12	12	11	11	11	8	12

Sample number	DYS485(Y)	DYS490(B)	DYS495(G)	DYS567(Y)	DYS494(B)	DYS578(G)	DYS565(Y)	DYS481(G)	DYS576(B)	DYF390S1(G)	DYS569(Y)	DYS618(G)	DYS511(Y)	DYS643(B)	DYS556(G)	DYS573(Y)	DYS530(B)	DYS491(G)	DYS549(Y)	DYS640(G)	DYS554(B)	DYS497(G)
66	10	12	15	10	9	10	11	22	17	3	10	12	11	11	11	10	9	13	12	15	9	11
101	10	12	15	10	9	10	11	22	17	3	10	12	11	11	11	10	9	13	12	15	9	11

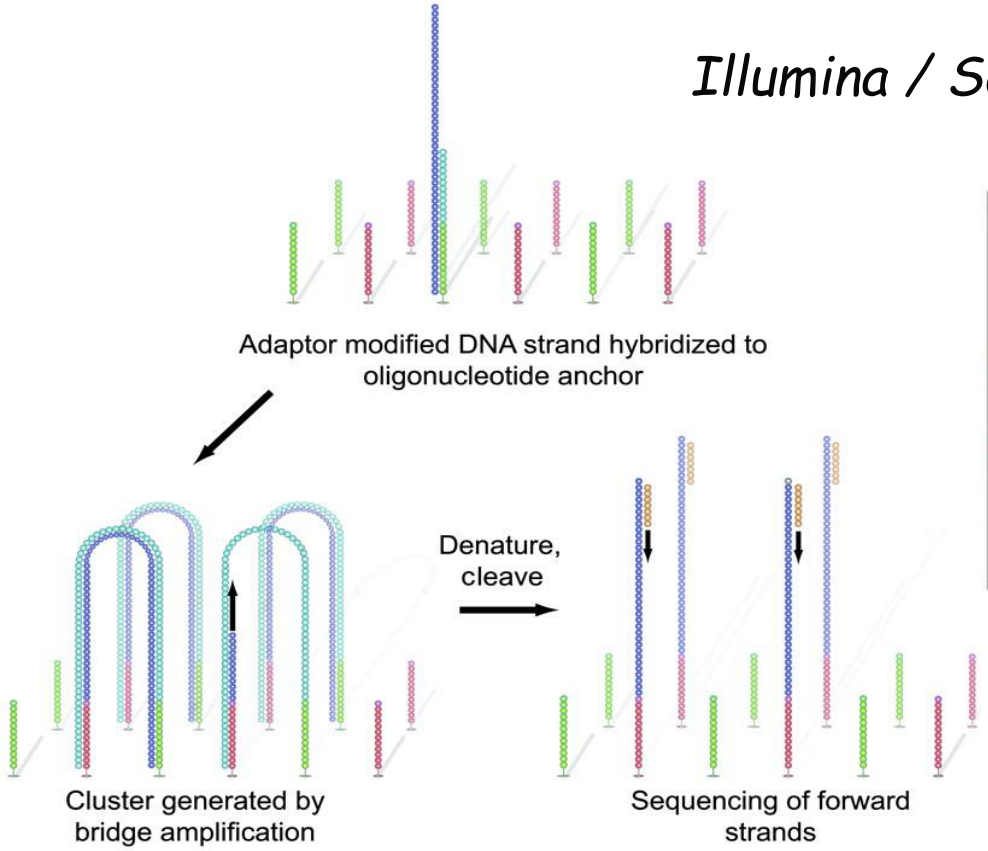
Table S1. Y-SNP and Y-STR haplotypes of the DFNY1-66 and DFNY1-101 chromosomes

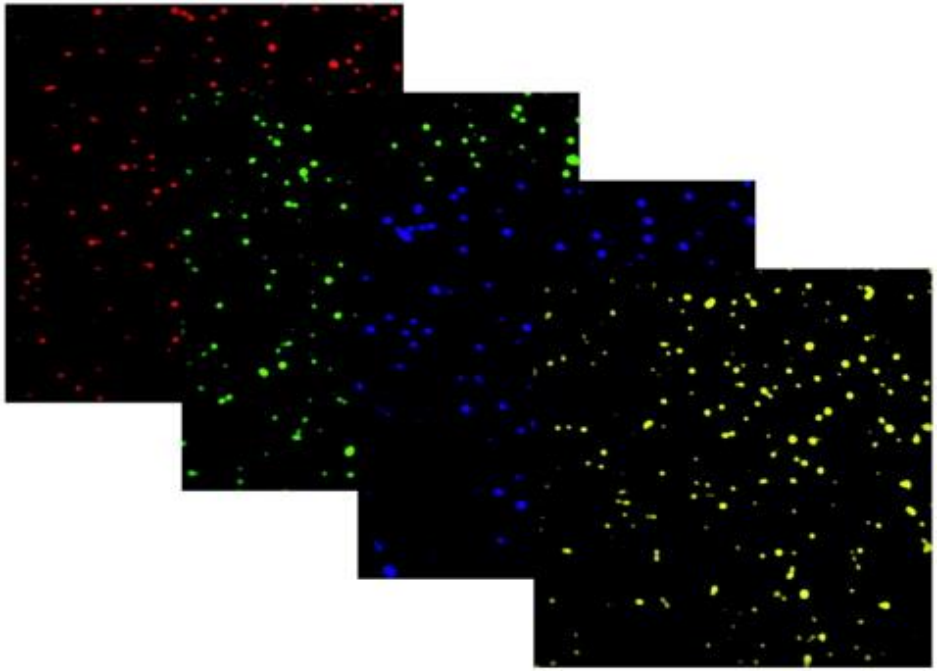
Identical Y chromosome haplotype at 67 microsatellites and 11 SNPs in the linage

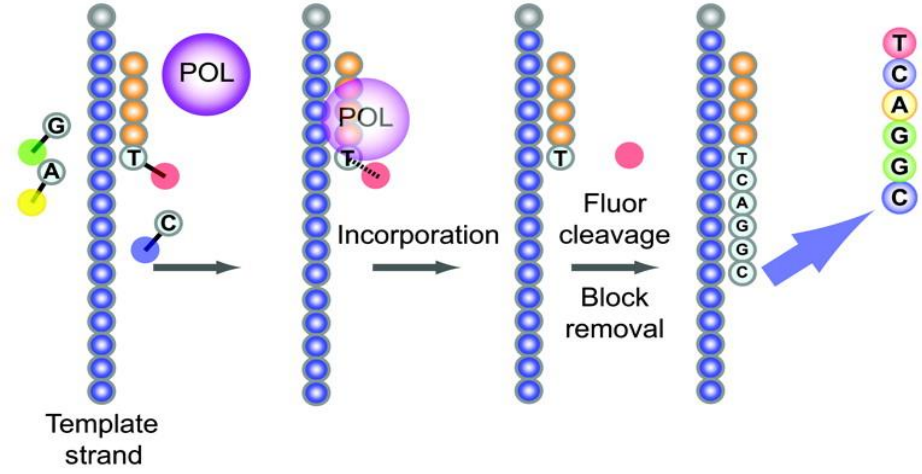
-generation distance: 13 generations apart

-markers localisation: euchromatin

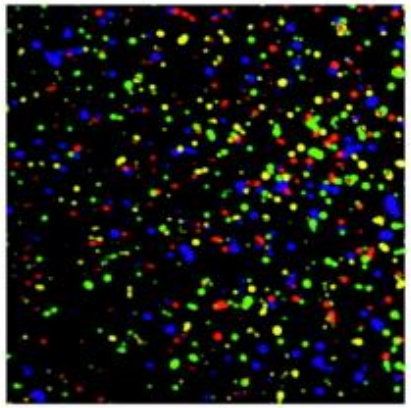
Illumina / Solexa NGS genome sequencing







Sequencing by reversible dye terminators



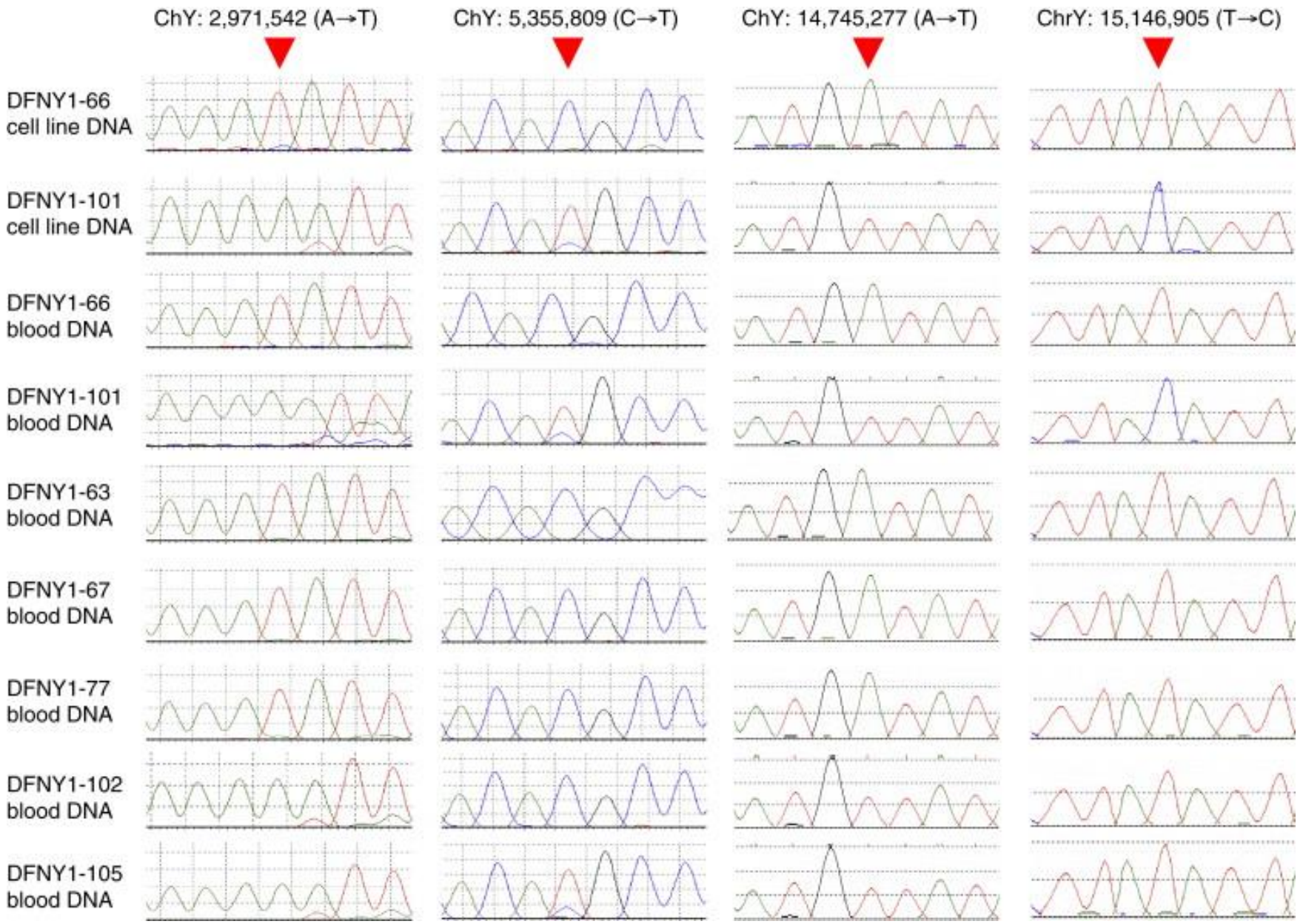
Observed mutations in Y chromosome euchromatic region

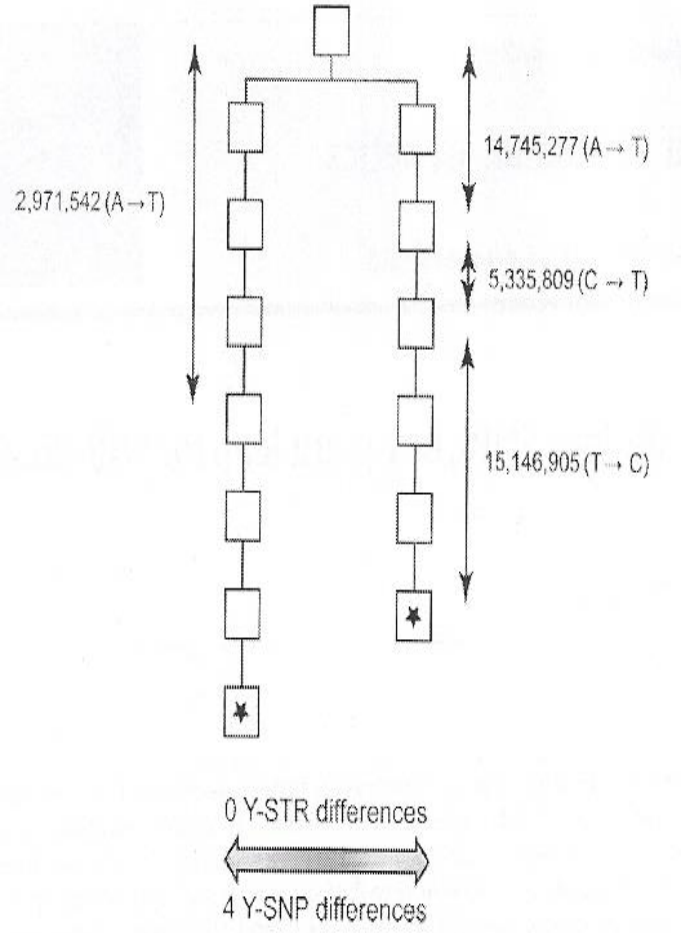
Table 2. Details of the Filtered Candidate Mutations

		DFNY1_101 Pileup		DFNY1_66 Pileup		Confirmation	
Chromosome Coordinate	Base	Coverage	Calls[1]	Coverage	Calls[1]	Cell-Line DNA	Blood DNA
First Class							
chrY:3,957,219	G	7	AAaaAAA	10	GGgGGGGgGG	Yes	No
chrY:4,633,474	C	4	tttT	6	cCCccc	Yes, het	No
chrY:4,939,256	T	13	cCccCcccCCCCC	13	TTTTTTTTTTttT	Yes	No
chrY:4,980,623	T	5	ggggg	7	TtTTTTT	Yes, het	No
chrY:5,355,809*	C	12	TtTTTTTTTtTt	9	cCccccCcC	Yes	Yes
chrY:6,555,594	G	13	TgTttTTtTTtTT	12	GGGGGgGGgGGG	No	
chrY:7,381,330	G	7	cCcCCCc	12	GGGGGgGGgGGG	No	
chrY:12,063,011	C	5	gggGG	8	ccccCCCC	Yes	No
chrY:14,745,277*	A	9	TtTTtTttt	6	aaAaAa	Yes	Yes
chrY:15,126,873	T	7	cccCccc	8	tttTttTT	Yes	No
chrY:15,146,905*	T	4	CCcC	9	tTtTTTTtT	Yes	Yes
chrY:20,627,064	C	9	gGGgGGGG.	5	Ccccc	Yes	No
chrY:27,095,961	T	7	CCcCCCc	8	TTtttTTt	Yes	No
chrY:2,971,542*	A	4	aAAA	14	tTTtTTttTtttT	Yes	Yes
chrY:4,097,585	C	7	CCcaacc	2	aa	No	
chrY:4,876,956	T	11	aatTTTTTTTT	4	AAAA	No	
chrY:11,970,133	T	10	tttTTTTTTt	6	aaAAaa	No	
chrY:19,883,785	A	5	aAaaA	4	cccc	No	
Second Class							
chrY:13,445,456	G	4	GGGg	1	t	No	
chrY:13,568,272	G	13	aAAgggggggggg	11	aaaAaAaaAAa	No	
chrY:13,833,351	C	17	cCccCCggccCcCcccc	16	CCcCcCCcCttCtttc	No	
chrY:14,573,532	A	21	GAAAAaaAaAAaAaaAAaAAg	5	AAggg	No	
chrY:15,375,202	G	4	GGGg	4	TTTT	No	

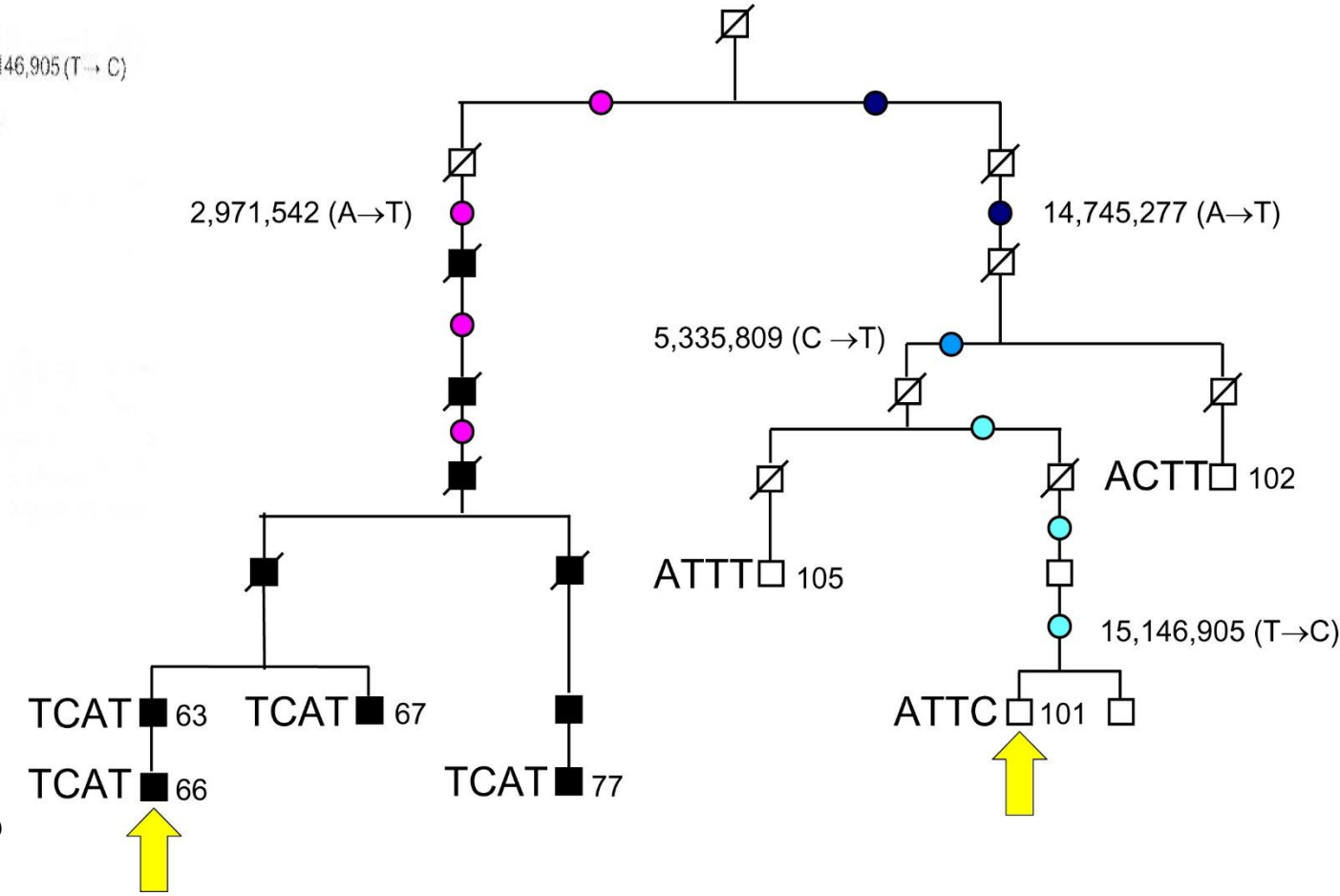
An asterisk denotes mutations that were confirmed in blood DNA.

[1] Upper case = forward strand; lower case = reverse strand.


ChY: 2,971,542 (A→T)
ChY: 5,355,809 (C→T)
ChY: 14,745,277 (A→T)
ChrY: 15,146,905 (T→C)
DFNY1-66 cell line DNA
DFNY1-101 cell line DNA
DFNY1-66 blood DNA
DFNY1-101 blood DNA
DFNY1-63 blood DNA
DFNY1-67 blood DNA
DFNY1-77 blood DNA
DFNY1-102 blood DNA
DFNY1-105 blood DNA



de novo substitutions in the Ychromosome euchromatic region based on pedigree analysis



Y chromosome mutation rate:

3.0×10^{-8}

Consistent with human-chimp Y chr. comparison

ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

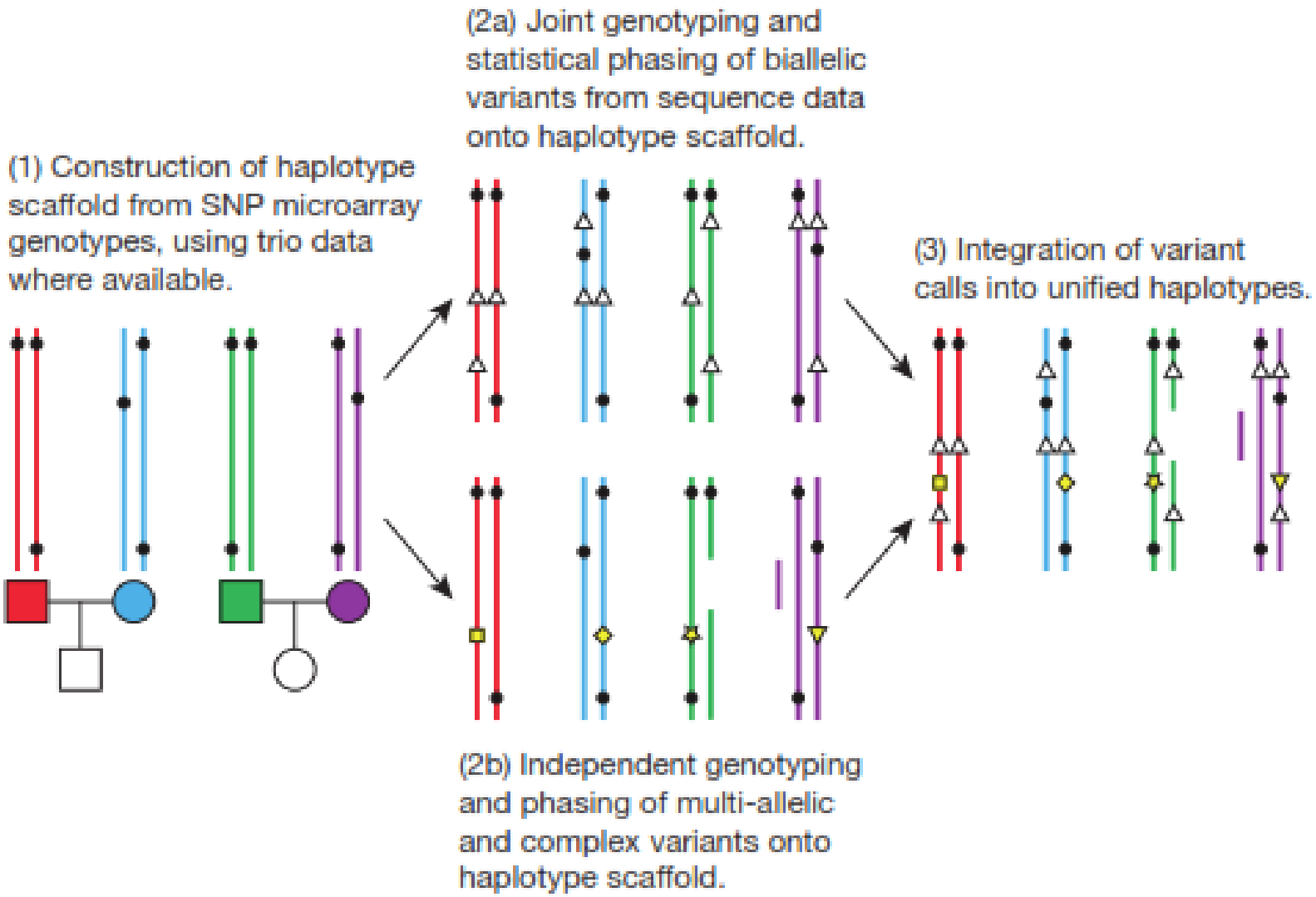
The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 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.

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

Structural variants are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight structural variant classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype blocks in 26 human populations. Analysing this set, we identify numerous gene-intersecting structural variants exhibiting population stratification and describe naturally occurring homozygous gene knockouts that suggest the dispensability of a variety of human genes. We demonstrate that structural variants are enriched on haplotypes identified by genome-wide association studies and exhibit enrichment for expression quantitative trait loci. Additionally, we uncover appreciable levels of structural variant complexity at different scales, including genic loci subject to clusters of repeated rearrangement and complex structural variants with multiple breakpoints likely to have formed through individual mutational events. Our catalogue will enhance future studies into structural variant demography, functional impact and disease association.

Building a haplotype scaffold

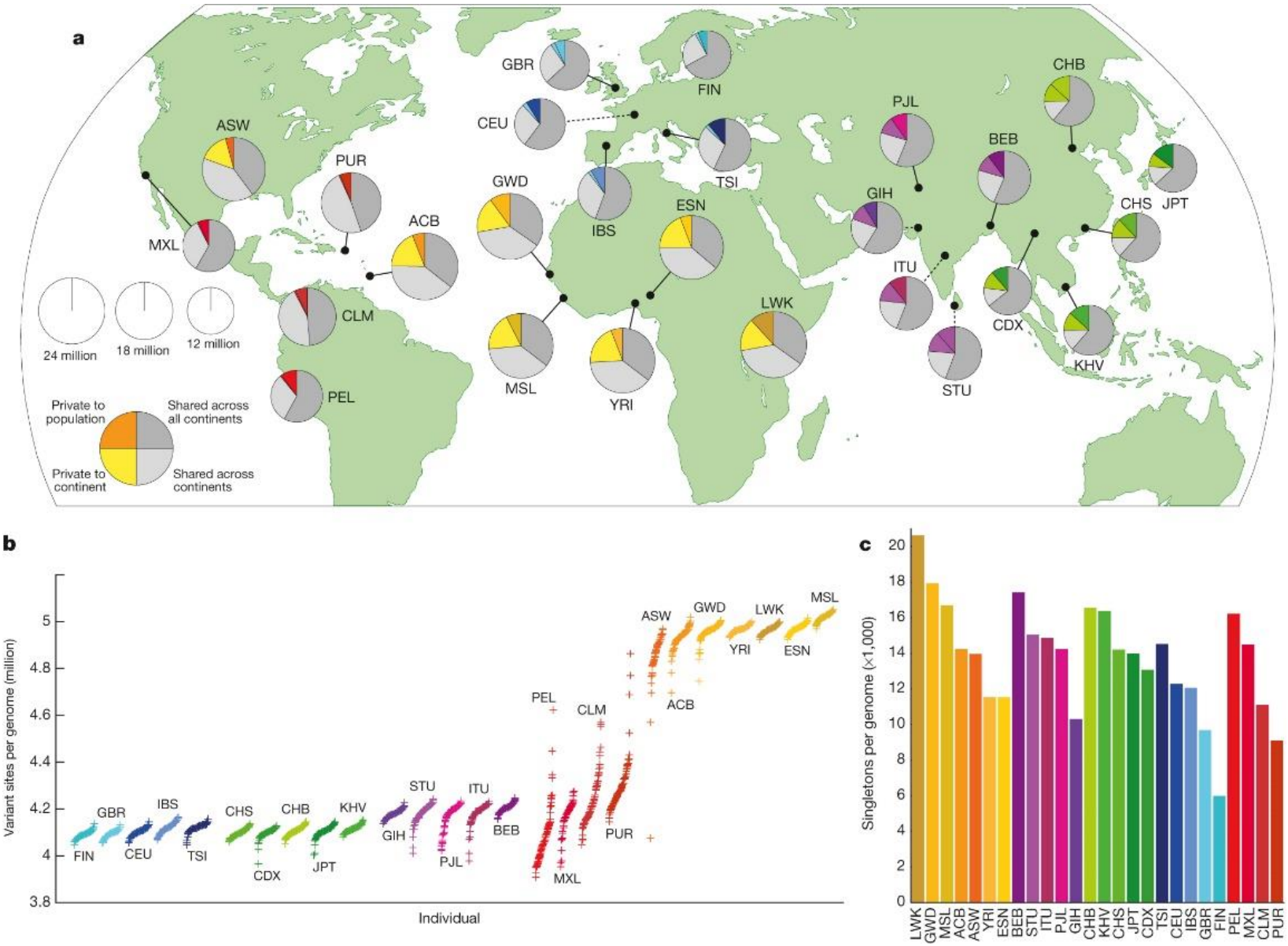


	Autosomes	Exome target regions**	chrX***	chrY***	Totals
Samples	2,504	2,504	2,504	1,233	-
Total Raw Bases (Gb)	85,426	18,273	3,213	291	-
Mean Mapped Depth (X)*	8.45	75.25	6.20	2.60	-
Total Variant Sites	84,801,880	1,416,049	3,468,093	62,042	88,332,015
Biallelic SNPs	81,102,777	1,383,927	3,223,927	60,505	84,387,209
Indels	3,196,364	19,832	212,196	1,427	3,409,987
Mean Indel Length (bp)	2.94	3.46	2.64	2.00	-
Multiallelic sites	444,026	6,153	30,996	-	475,022
Multiallelic SNPs	274,425	4,706	15,055	-	289,480
Multiallelic Indels	169,601	1,447	15,941	-	185,542
Structural Variants	58,713	6,137	974	110	59,797
ALU Insertion	12,491	52	-	-	12,491
LINE1 Insertion	2,910	10	-	-	2,910
Large Deletion	33,336	2,684	974	-	34,310
Duplication	5,896	2,513	-	-	5,896
SVA Insertion	822	5	-	-	822
Other Insertion	165	1	-	-	165
Inversion	100	8	-	-	100
CNV	2,993	864	-	110	3,103

Supplementary Information Table 3: Integrated callset summary. *Assuming 2.84Gb as the genome size. The mapping of exome sequence to targeted pull down regions was calculated by Picard function *calculateHsMetrics*. **The exome targeted regions were exome pulldown targets derived from CCDS (NimbleGen EZ Exome v1 and Agilent SureSelect v2). These variant totals are included in the other columns. ***chrX and chrY statistics are for the entire chromosomes.

- a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites.
- >99.9% of variants consist of SNPs and short indels.
- structural variants affect more bases:
- typical genome contains an estimated 2,100 to 2,500 structural variants (1,000 large deletions, 160 copy-number variants, 915 Alu insertions, 128 L1 insertions, 51 SVA insertions, 4 NUMTs and 10 inversions) affecting 20 million bases of sequence.

Population sampling



nature

ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The International Genome Sample Resource

https://www.internationalgenome.org/

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.

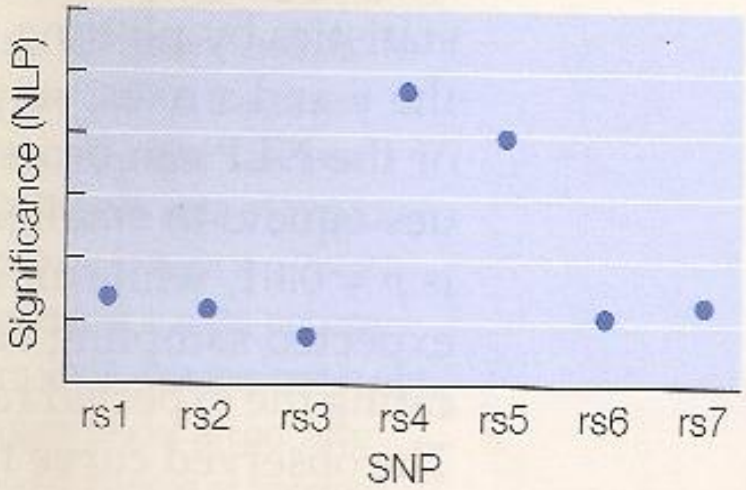
GWAS: Genome wide association studies

First study: Age related macular degeneration, 100k SNPs array (Klein et al., 2005)
- Complementer factor H gene (CFH), 4x risk in hetero- and 7x in homozygous patients

SNP array

	SNP							
Individual	**rs1**	**rs2**	**rs3**	**rs4**	**rs5**	**rs6**	**rs7**	**Status**
ID001	A	T	C	C	A	G	A	Case
ID002	A	C	T	C	A	G	A	Case
ID003	A	C	C	T	G	–	A	Control
ID004	T	T	C	T	A	G	A	Control
ID005	A	T	C	T	G	G	A	Control
ID006	A	C	C	C	A	–	A	Case
ID007	T	C	T	T	G	–	A	Control
ID008	T	T	C	C	A	G	A	Case
ID009	T	C	T	C	A	–	G	Case
ID010	T	T	T	T	G	–	A	Control

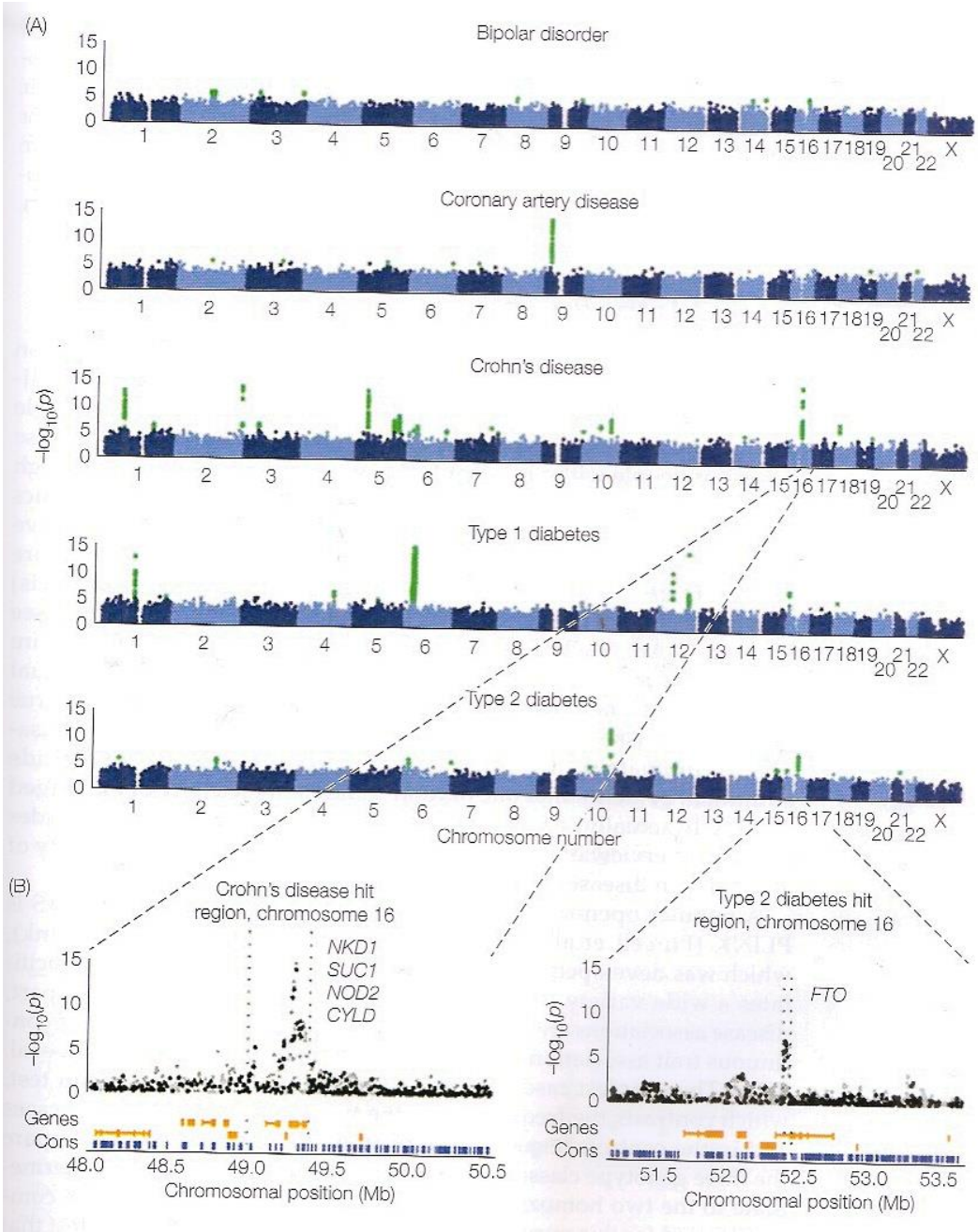
Manhattan plot



Significance threshold $p < 10^{-8}$ (Negative logarithm of p-value)

Gibson G, A primer of Human genetics, 2015

Welcome Trust Case Control Consortium 2007



360K SNP array

Case samples: 2.000

Controls: 3.000

Significance threshold:

$p < 10^{-5}$

<u>23 associations:</u>

-No in bipolar disorder

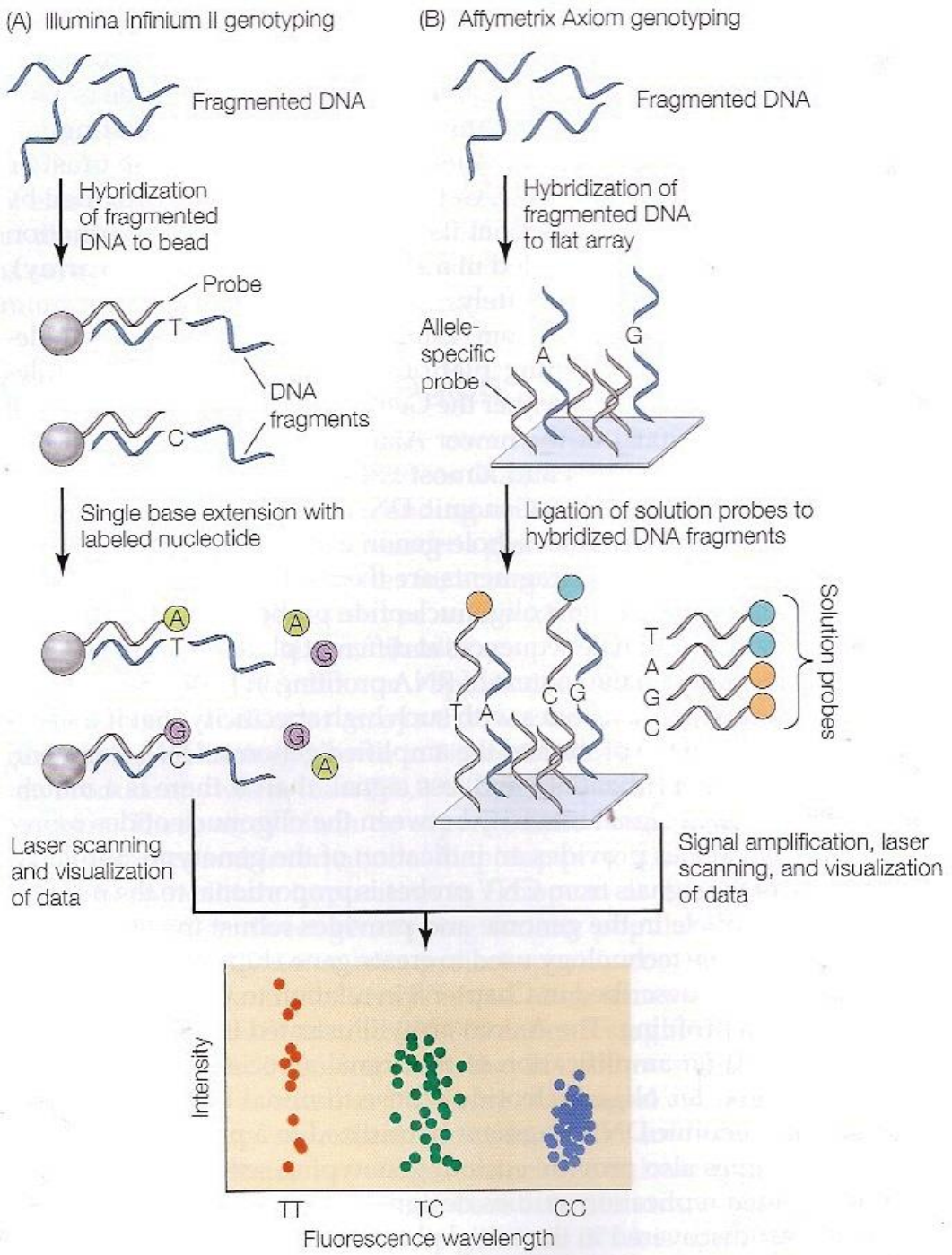
-1 coronary artery

-3 diabetes type 2

-7 diabetes type 1

-9 Crohn's disease

WTCCC, Nature 447:661-678, 2007



GWAS: genotyping

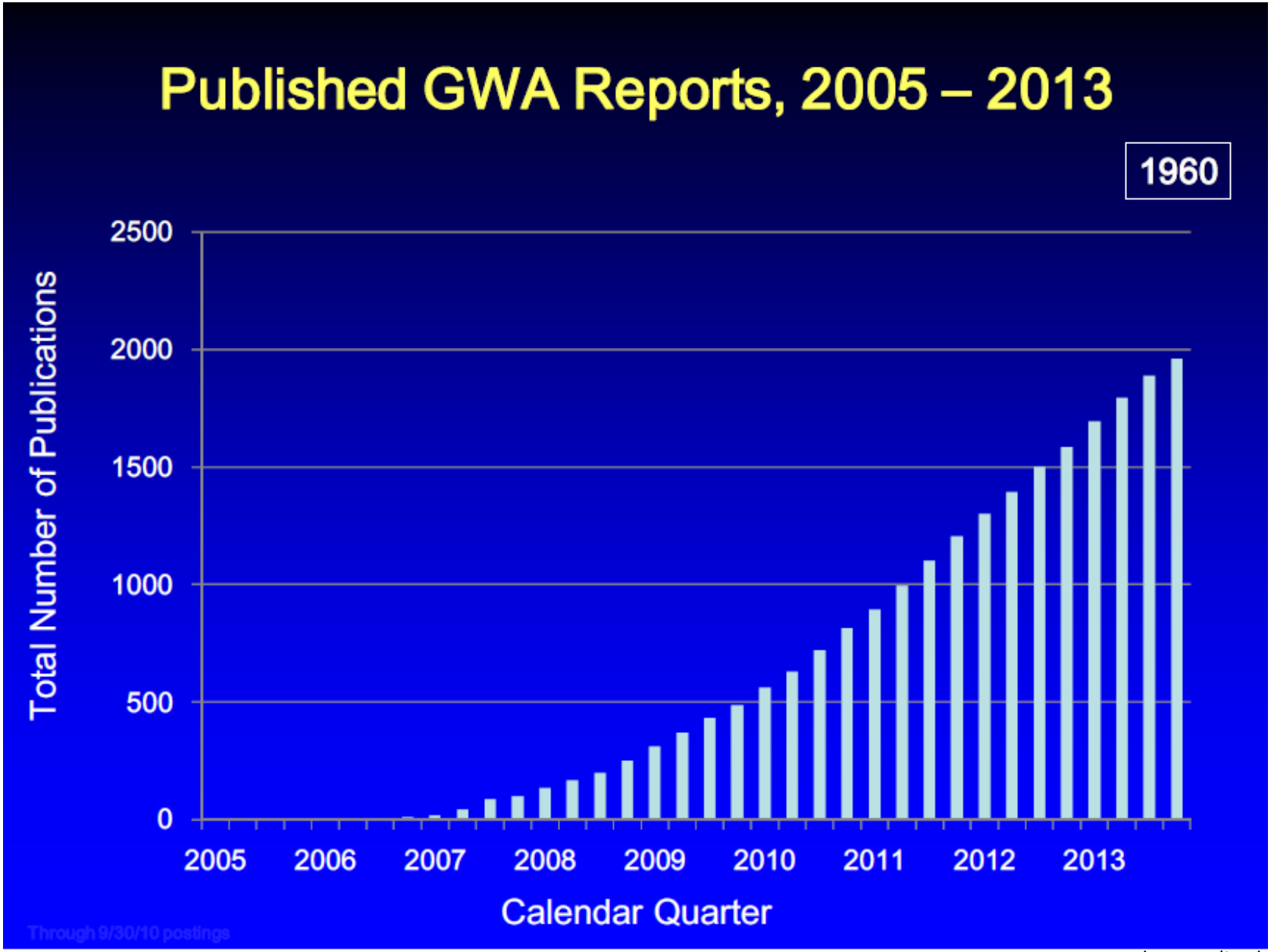
-Illumina Infinium II

-Affymetrix Axiom

(Copy Number Variation)

Gibson G, A primer of Human genetics, 2015

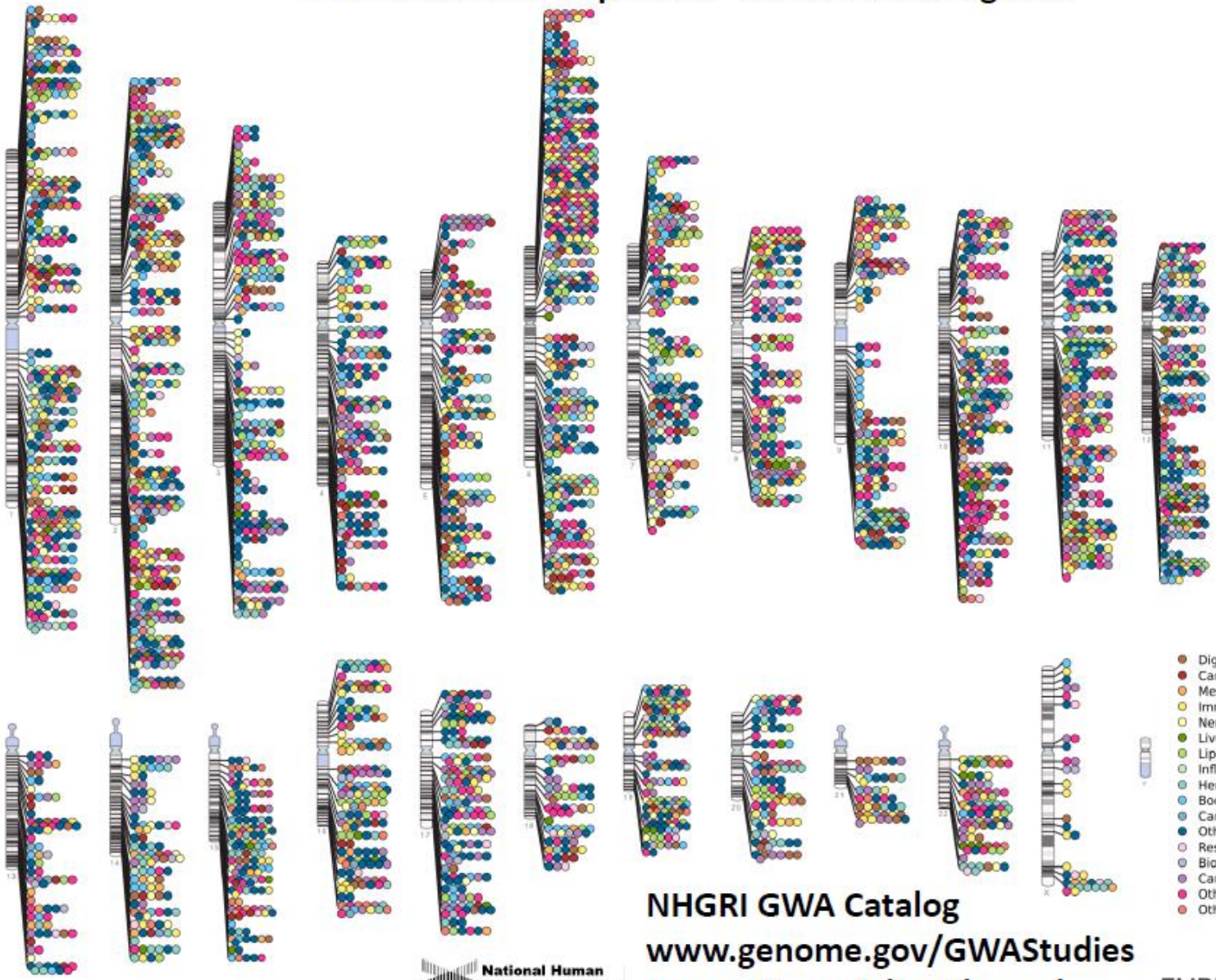
GWAS: Genome wide association studies



www.genome.gov/gwastudies/

Published Genome-Wide Associations through 12/2013

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

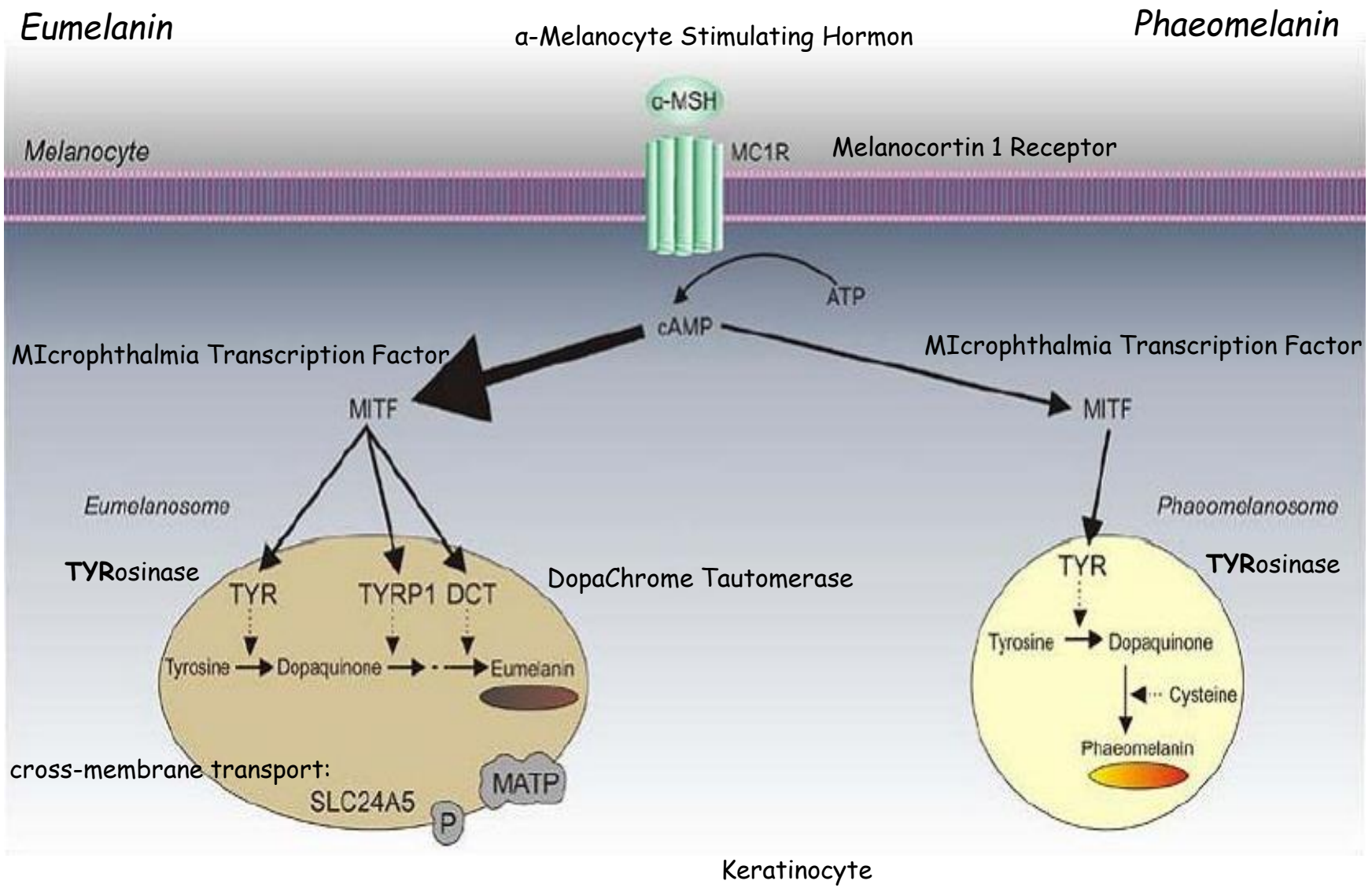
www.genome.gov/gwastudies/

Autosome SNPs in the Human Genome

TABLE 12.2 Categories of SNP Markers (See Budowle & van Daal 2008, Butler et al. 2008).

Category	Characteristics	Examples
Identity SNPs Individual Identification SNPs (IISNPs)	SNPs that collectively give very low probabilities of two individuals having the same multi-locus genotype	FSS 21plex (Dixon et al. 2005) SNPforID 52plex (Sanchez et al. 2006) Kidd group SNPs (Pakstis et al. 2010)
Lineage SNPs Lineage Informative SNPs (LISNPs)	Sets of tightly linked SNPs that function as multi-allelic markers that can serve to identify relatives with higher probabilities than simple bi-allelic SNPs	mtDNA coding region SNPs (Coble et al. 2004) Japanese Y-SNPs (Mizuno et al. 2010) Haplotype blocks (Ge et al. 2010)
Ancestry SNPs Ancestry Informative SNPs (AISNPs)	SNPs that collectively give a high probability of an individual's ancestry being from one part of the world or being derived from two or more areas of the world	SNPforID 34plex (Phillips et al. 2007b) 24 SNPs (Lao et al. 2010) FSS YSNPs (Wetton et al. 2005)
Phenotype SNPs Phenotype Informative SNPs (PISNPs)	SNPs that provide a high probability that the individual has particular phenotypes, such as a particular skin color, hair color, eye color, etc.	Red hair (Grimes et al. 2001) "Golden" gene pigmentation (Lamason et al. 2005) IrisPlex eye color (Walsh et al. 2010)

The human melanogenesis



Bouakaze et al., Int J Legal Med (2009) 123:315

Genes underlying skin pigmentation

Principal skin pigmentation candidate genes

Locus	Chromosome	Protein	Mut phenotype	Function
Melanosome proteins				
TYR	11q14-11q21	Tyrosinase	OCA1	Oxidation of tyrosine
TYRP1	9p23	Gp75, TRYP1	OCA3	DHICA-oxidase, TYR stabilisation
DCT	13q32	DCT, TRYP2		Dopachrome tautomerase
OCA2	15q11.2-15q12	P-protein	OCA2 (eye)	pH of melanosome
SLC45A2	5p14.3-5q12.3	MATP, AIM-1	OCA4 (skin)	Melansome maturation
SLC24A5	15q21.1	Cation exchanger		Melanosome precursor
Signal proteins				
ASIP	20q11.2-20q12	Agouti signal protein		MC1R antagonist
MC1R	16q24.3	MSH receptor	Red hair (skin)	G-protein coupled receptor
POMC	16q24.3	MSH receptor	Red hair	MC1R antagonist
OA1	Xp22.3	OA1 protein	OA1	G-protein coupled receptor
MITF	3p12.3-3p14.1	MITF	Waardenburg	Transcription factor
Proteins involved in melanosome transport or uptake by keratinocytes				
MYO5A	15q21	MyosinVa	Griscelli	Motor protein
RAB27A	15q15-15q21.1	Rab27a	Griscelli	RAS family protein
HPS1	10q23.1-10q23.3	HPS1	Hermansky-Pudlak	Organelle biogenesis and size
HPS6	10q24.32	HPS6	Hermansky-Pudlak	Organelle biogenesis

ACTH: adrenocorticotrophin hormone; DCT: dopachrome tautomerase; DHICA: 5,6-dihydroxyindole-2-carboxylic acid; MATP: membrane-associated transporter protein; MC1R: melanocortin-1 receptor; MITF: microphthalmia-associated transcription factor; MSH: melanocyte stimulating hormone; OCA: oculocutaneous albinism; POMC: pro-opiomelanocortin; TYRP1: tyrosinase-related protein 1.

Pulker et al., FSI Genet (2007) 1:100

MC1R gene mutation

Mutations in the MC1R gene, their penetrance and functional significance (where known)

Mutation	Type	Designation	Penetrance (odds ratio)	Functional significance	References (for functional significance and penetrance)
R151C	Mis-sense	R	63.3	Altered cellular location	[16,26]
R160W	Mis-sense	R	63.3	Altered cellular location	[16,26]
D294H	Mis-sense	R	63.3	Impaired G coupling ability	[26,27]
D84E	Mis-sense	R	63.3	Altered cellular location	[16,26]
I155T	Mis-sense	Lack of statistical data—strong familial association		Altered cellular location	[16,26]
V92M	Mis-sense	r	5.1	Reduced α-MSH binding	[26,28,29]
V60L	Mis-sense	r	5.1		[26]
R163Q	Mis-sense	r	5.1	Slightly reduced α-MSH binding	[26,29]
R142H	Mis-sense	Lack of statistical data—strong familial association			[26]

- MC1R allele variants possess different activities
- 317 amino acids, 7 transmembrane domains
- SNPs, RHC phenotype
- Neanderthal pigmentation
- Genetic tests, phenotype prediction

Tully, FSI Genetics (2007) 1:105

SNPs located in pigmentation genes

- ASIP (aguti): 3'UTR 8818A - MSH anagonist - phaeomelanin production
- MATP: melanosome pH regulation, 374Leu allele – dark colour, albinism
- SLC24A5: „gold" gene, zebrafish, Ala111Thr allele, light shade, fixed in kaukasoid race, selection?
- OCA2: albinism gene, 305 Arg/Trp, Africa / Europe

Gene	Location	Protein	Reference SNP ID (rs#)[a]	Alleles	Variation type
MC1R	16q24.3	MC1R: melanocortin 1 receptor	rs1805007	C/T	ns coding, c.451C>T, p.R151C
			rs1805008	C/T	ns coding, c.478C>T, p.R160W
HERC2	15q13	Unknown	rs12913832	A/G	Non-coding, intron 86
OCA2	15q11.2-15q12	P-protein: NA+/H+ antiporter or glutamate transporter	rs7495174	T/C	Non-coding, intron 1
			rs6497268 or rs4778241	G/T	
			rs11855019 or rs4778138	T/C	
			rs1545397	G/A	Non-coding intronic
SLC45A2	5p13.3	MATP: membrane-associated transporter protein	rs16891982	C/G	ns coding, c.1122C>G, p.F374L
SLC24A5	15q21.1	SLC24A5 (or NCKX5): solute carrier family 24, member 5; potassium-dependent sodium–| calcium ion exchanger	rs1426654	G/A	ns coding, p.A111T
DCT	13q32	DCT or TYRP2/TRP-2: dopachrome tautomerase or tyrosinase-related protein-2	rs2031526	G/A	Non-coding, intronic

ns non-synonymous

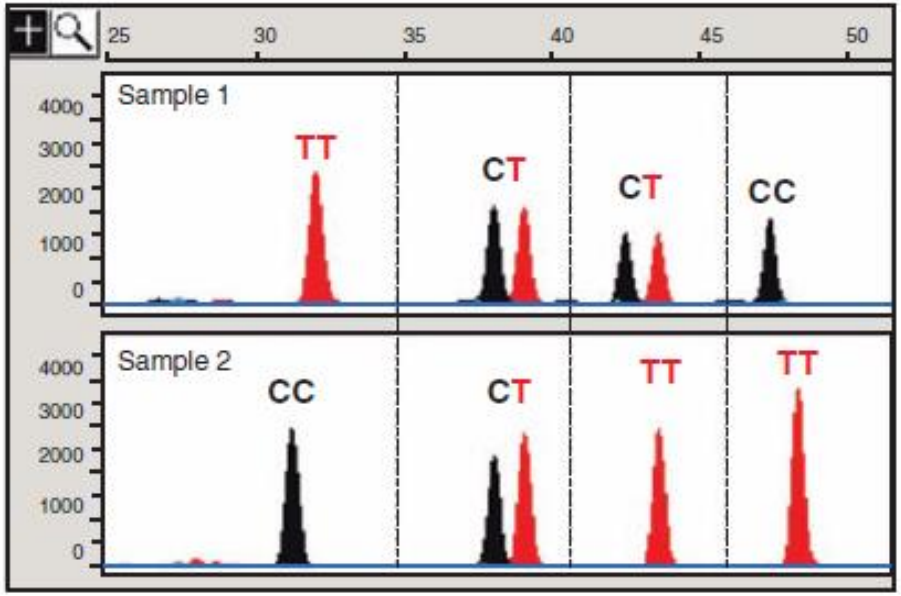
[b] Reference SNP ID refer to the reference sequence identifier given to the SNP in the dbSNP database

Bouakaze et al., Int J Legal Med (2009) 123:315

SNaPshot: A Primer Extension Assay Capable of Multiplex Analysis

Minisequencing (SNaPshot assay)	Allele-specific primer extension across the SNP site with fluorescently labeled ddNTPs; mobility modifying tails can be added to the 5′-end of each primer in order to spatially separate them during electrophoresis.

(a)



(b) (TTTTT)–**primer1** (chromosome 20)-**ddT/ddT**

(TTTTT)–(TTTTT)–**primer2** (chromosome 6)-**ddC/ddT**

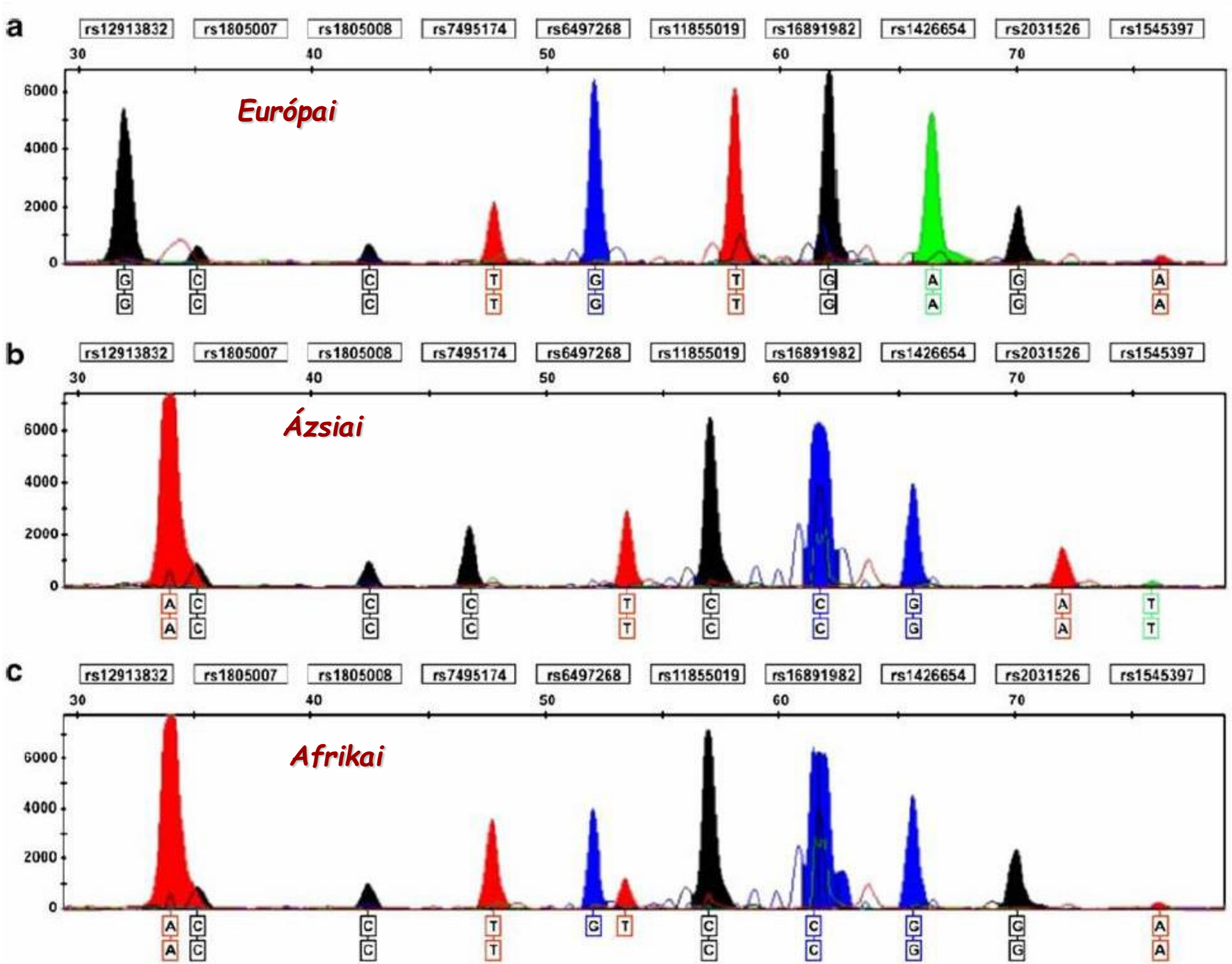
(TTTTT)–(TTTTT)–(TTTTT)–**primer3** (chromosome 14)-**ddC/ddT**

(TTTTT)–(TTTTT)–(TTTTT)–(TTTTT)–**primer4** (chromosome 1)-**ddC/ddC**

FIGURE 12.2 Allele-specific primer extension results using four autosomal SNP markers on two different samples (a). SNP loci are from separate chromosomes (1, 6, 14, and 20) and therefore unlinked. Electrophoretic resolution of the SNP primer extension products occurs due to poly(T) tails that are 5 nucleotides different from one another (b).

Butler, J.M. (2011) Advanced Topics in Forensic DNA Typing, p. 354

10 pigmentation genes SNP genotyping (SNaPshot)



Bouakaze et al., Int J Legal Med (2009) 123:315

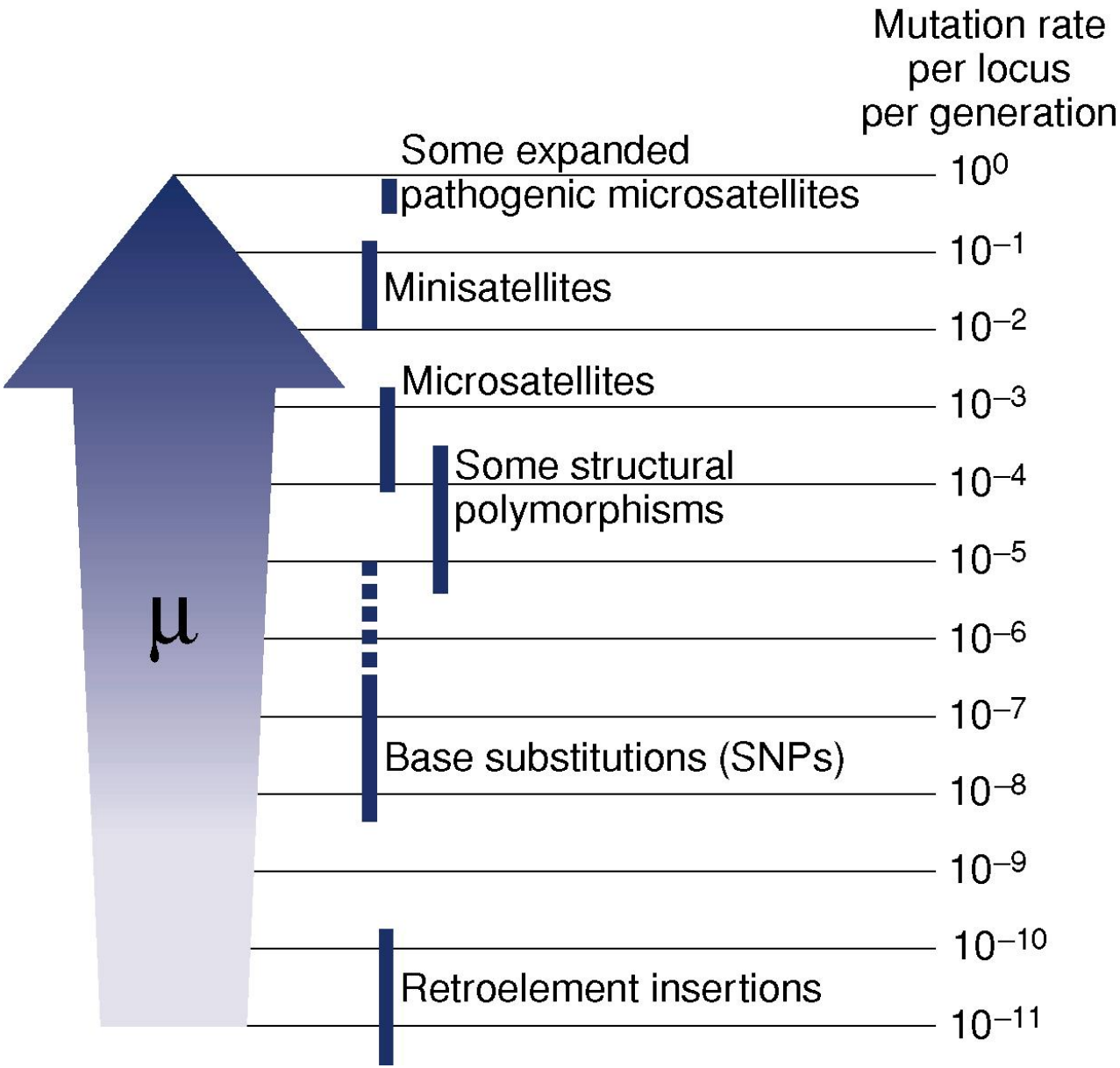
Sample	Self-reported pigmentary traits			rs12913832 HERC2	rs1805007 MC1R	rs1805008 MC1R	OCA2 diplotype[a]	rs16891982 SLC24A2	rs1426654 SLC24A5	rs2031526 DCT	rs1545397 OCA2	Inferred ancestry of individuals[b]		
	Eye color	Hair color	Skin color									European	Asian	African
E1	Blue	Red	Fair	G/G	C/C	C/T	TGT/TGT	G/G	A/A	G/G	A/A	**0.963**	0.012	0.024
E2	Green	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	G/G	A/A	A/G	A/A	**0.954**	0.021	0.025
E3	Blue	Blond	Fair	G/G	C/C	C/C	TGT/TGT	G/G	A/A	A/G	A/A	**0.954**	0.024	0.022
E4	Blue	Blond	Fair	G/G	C/C	C/C	TGT/TGT	G/G	A/A	A/G	A/A	**0.960**	0.020	0.020
E5	Blue/gray	Auburn	Fair	G/G	C/T	C/C	TGT/TGT	G/G	A/A	G/G	A/A	**0.961**	0.013	0.026
E6	Green/gray	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	C/G	A/A	G/G	A/A	**0.787**	0.038	0.175
E7	Green/hazel	Light brown	Fair	A/G	C/C	C/C	TGT/TGT	G/G	A/A	A/G	A/A	**0.955**	0.022	0.024
E8	Green/hazel	Dark brown	Fair	A/A	C/C	C/C	TGT/CTC	G/G	A/A	G/G	A/A	**0.961**	0.013	0.027
E9	Green/hazel	Dark brown	Fair	A/A	C/C	C/C	TTT/CTC	G/G	A/A	G/G	A/A	**0.963**	0.013	0.024
E10	Blue	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	C/G	A/A	G/G	A/A	**0.789**	0.049	0.163
E11	Green	Auburn	Fair	G/G	C/T	C/C	TGT/TGC	G/G	A/A	G/G	A/A	**0.958**	0.014	0.028
E12	Blue/hazel	Light brown	Fair	A/G	C/C	C/C	TGT/TTT	G/G	A/A	G/G	A/A	**0.962**	0.012	0.026
E13	Blue/hazel	Light brown	Fair	A/G	C/C	C/C	TGT/TTT	G/G	A/A	G/G	A/A	**0.965**	0.013	0.022
E14	Green	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	C/G	A/A	G/G	A/T	**0.763**	0.165	0.073
E15	Brown	Dark brown	Fair	A/G	C/C	C/C	TGT/TGT	G/G	A/A	A/G	A/A	**0.957**	0.022	0.021
E16	Brown	Dark brown	Fair	A/A	C/C	C/C	TGT/CTC	C/G	A/A	A/G	A/T	**0.669**	0.283	0.048
E17	Green/hazel	Dark brown	Medium	A/G	C/C	C/C	TGT/TTT	C/G	A/A	G/G	A/T	**0.755**	0.170	0.076
E18	Blue	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	G/G	A/A	G/G	A/T	**0.935**	0.045	0.021
E19	Brown	Red	Fair	A/G	C/T	C/C	TGT/TGT	G/G	A/A	G/G	A/A	**0.964**	0.013	0.022
E20	Green	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	C/G	A/A	G/G	A/A	**0.792**	0.047	0.161
E21	Green/gray	Blond	Fair	G/G	C/C	C/C	TGT/TGT	G/G	A/A	A/G	A/A	**0.957**	0.022	0.021
E22	Blue	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	G/G	A/A	G/G	A/A	**0.959**	0.014	0.026
E23	Green/hazel	Light brown	Fair	A/G	C/C	C/C	TGT/TTT	G/G	A/A	A/G	A/A	**0.957**	0.020	0.022
E24	Green	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	C/G	A/A	G/G	A/A	**0.786**	0.049	0.166
E25	Brown	Red	Fair	A/G	C/C	T/T	TGT/TGC	G/G	A/A	G/G	A/A	**0.963**	0.014	0.023
E26	Blue	Light brown	Fair	G/G	C/C	C/C	TGT/TGT	G/G	A/A	A/G	A/A	**0.954**	0.021	0.025
E27	Blue	Red	Fair	G/G	C/C	C/T	TGT/TGT	G/G	A/A	G/G	A/A	**0.958**	0.014	0.028
Af1	Brown	Black	Dark	A/A	C/C	C/C	TGC/TTC	C/C	G/G	A/G	A/A	0.028	0.094	**0.878**
Af2	Brown	Black	Dark	A/A	C/C	C/C	TGC/TTC	C/C	G/G	G/G	A/A	0.023	0.031	**0.946**
Af3	Brown	Black	Dark	A/A	C/C	C/C	TGC/TTC	C/C	A/G	G/G	A/A	0.164	0.041	**0.795**
As1	–	–	–	A/A	C/C	C/C	TTT/CTC	C/C	G/G	A/G	A/T	0.042	**0.649**	0.308
As2	–	–	–	A/A	C/C	C/C	CTC/CTC	C/C	G/G	A/G	T/T	0.020	**0.921**	0.060
As3	–	–	–	A/A	C/C	C/C	CTC/CTC	C/C	G/G	A/A	T/T	0.013	**0.964**	0.023
As4	–	–	–	A/G	C/C	C/C	TTT/CGC	C/C	A/G	A/A	A/T	0.212	**0.708**	0.080
As5	–	–	–	A/A	C/C	C/C	TTC/CGC	C/C	G/G	A/G	T/T	0.019	**0.922**	0.059
As6	–	–	–	A/A	C/C	C/C	CTC/CTC	C/G	G/G	A/A	T/T	0.119	**0.858**	0.023

E European modern sample, *Af* African modern sample, *As* Asian modern sample

[a] OCA2 diplotype correspond to markers rs7495174/rs6497268/rs11855019. OCA2 diplotype and rs12913832 genotype predictive of blue eye color phenotype are underlined

[b] Probability of being from European/Asian/African population determined using the STRUCTURE program. The greatest probability, most likely estimate of ancestry, is indicated in bold

Mutation rate of polymorph sequences (μ)

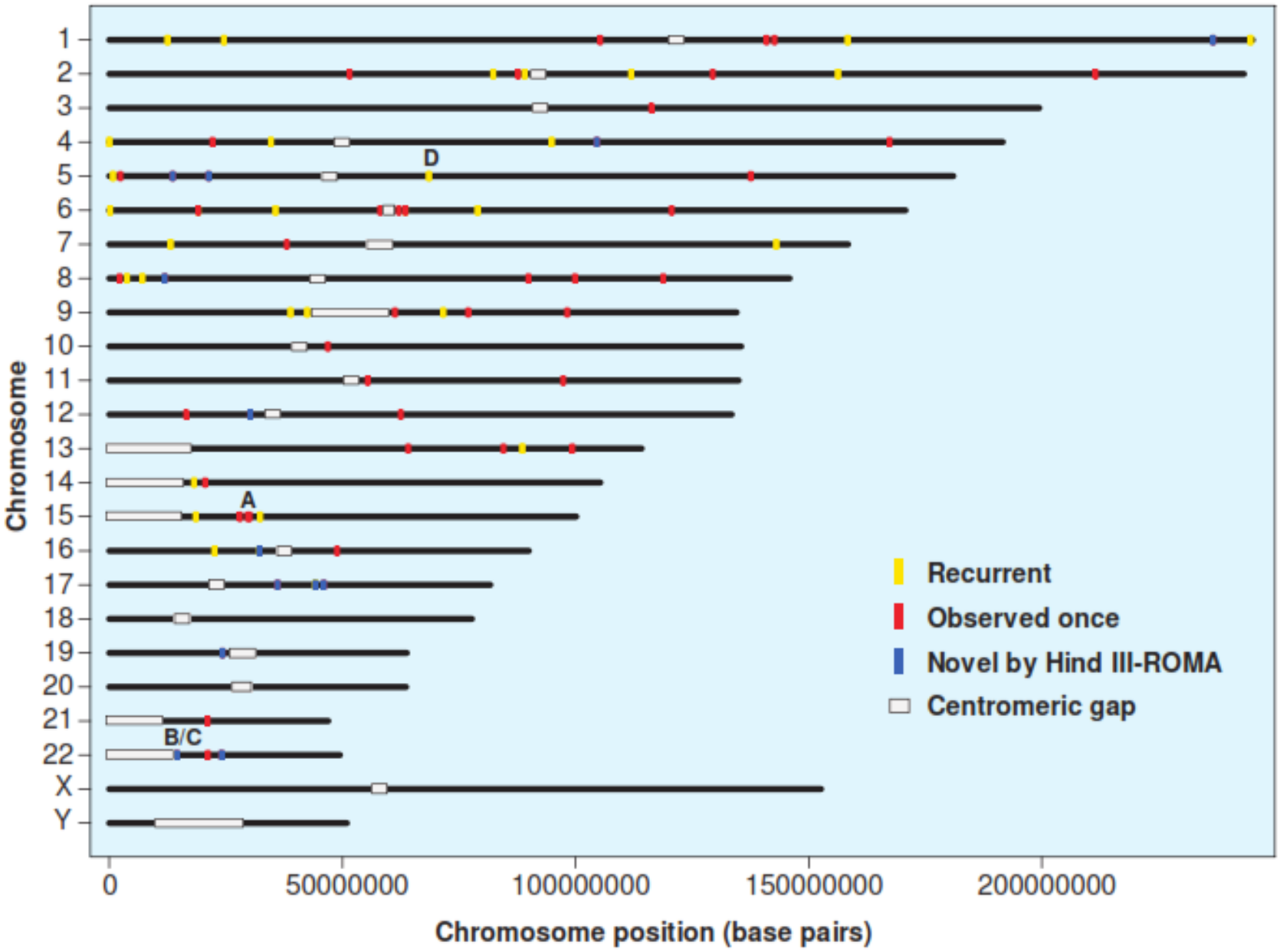


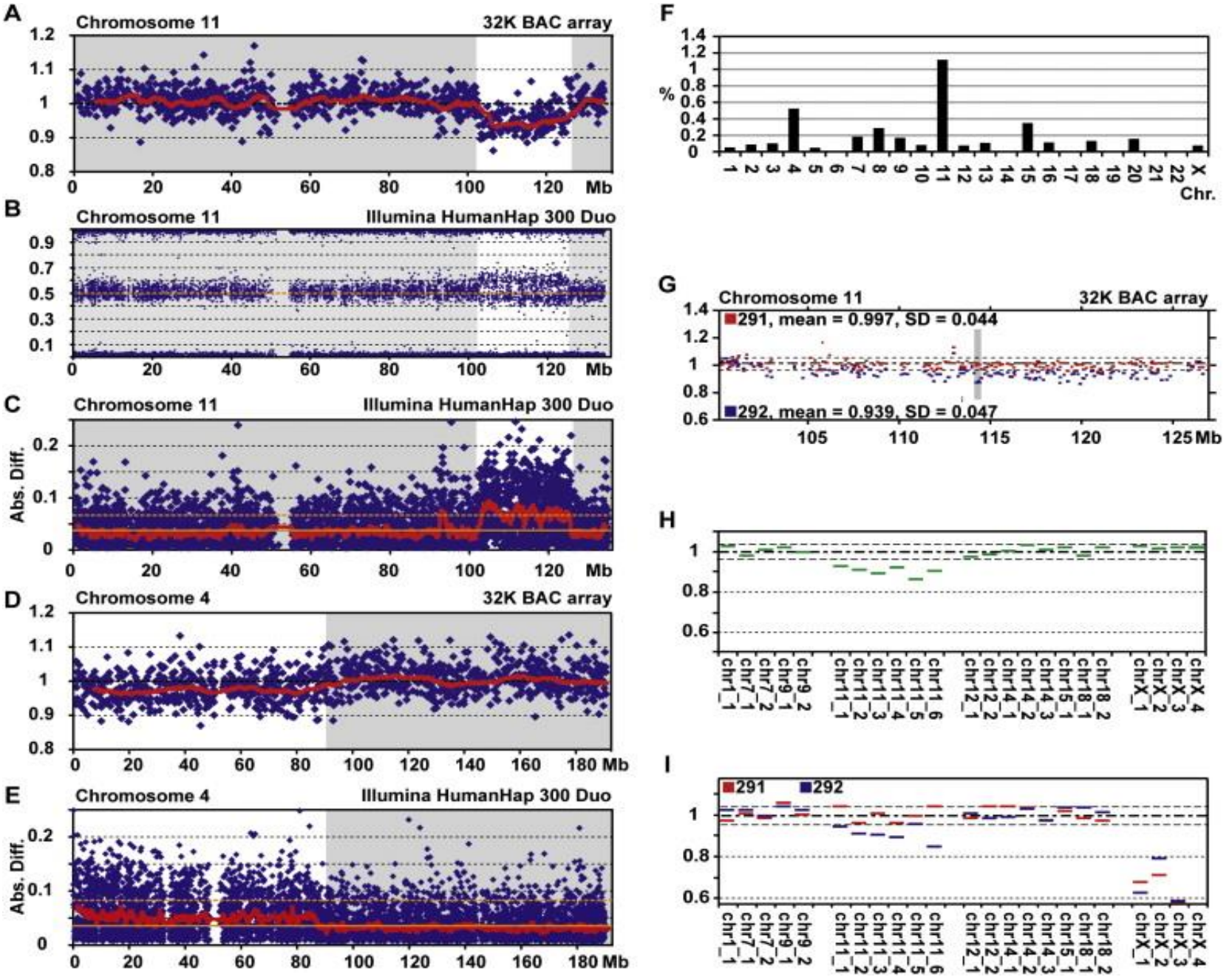
Human Evolutionary Genetics, Jobling, 2004

Large-Scale Copy Number Polymorphism in the Human Genome

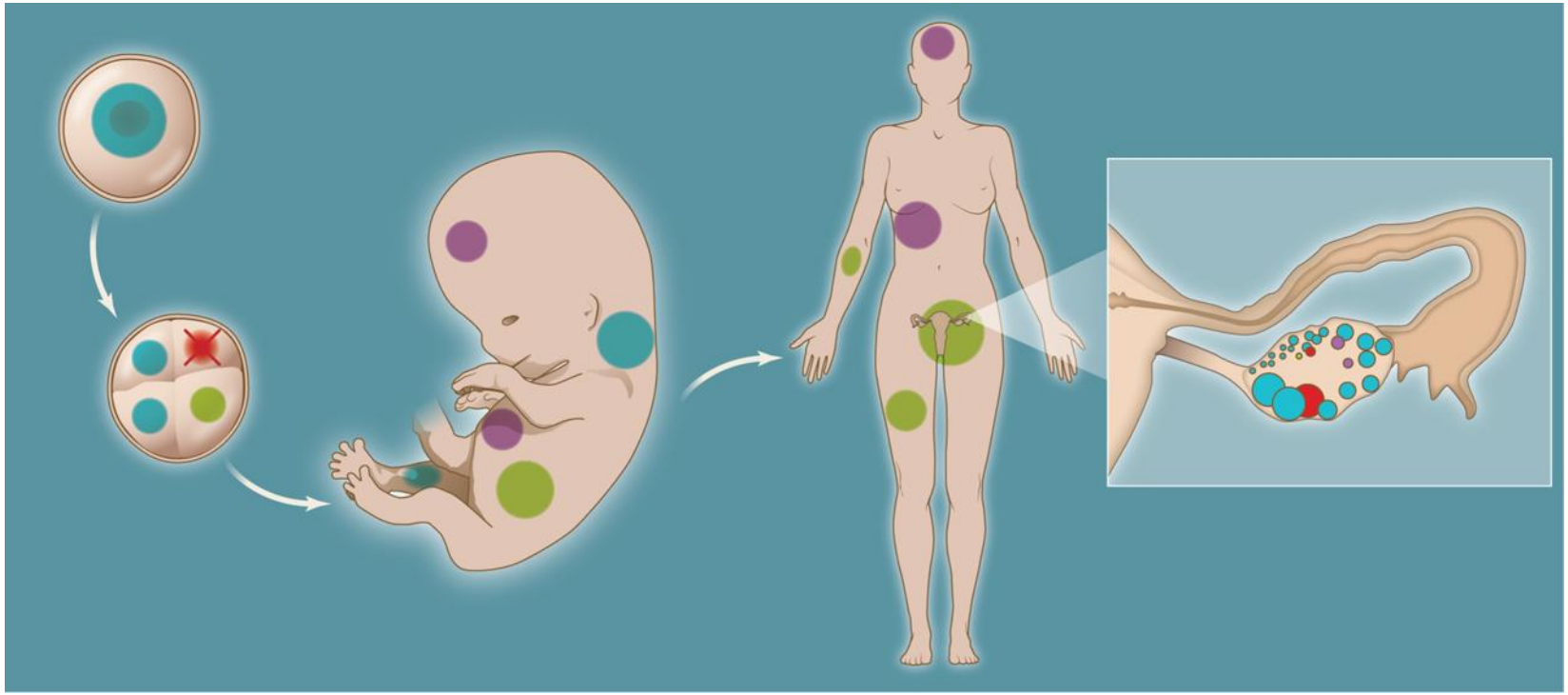
Jonathan Sebat,[1] B. Lakshmi,[1] Jennifer Troge,[1] Joan Alexander,[1] Janet Young,[2] Pär Lundin,[3] Susanne Månér,[3] Hillary Massa,[2] Megan Walker,[2] Maoyen Chi,[1] Nicholas Navin,[1] Robert Lucito,[1] John Healy,[1] James Hicks,[1] Kenny Ye,[4] Andrew Reiner,[1] T. Conrad Gilliam,[5] Barbara Trask,[2] Nick Patterson,[6] Anders Zetterberg,[3] Michael Wigler[1]*

The extent to which large duplications and deletions contribute to human genetic variation and diversity is unknown. Here, we show that large-scale copy number polymorphisms (CNPs) (about 100 kilobases and greater) contribute substantially to genomic variation between normal humans. Representational oligonucleotide microarray analysis of 20 individuals revealed a total of 221 copy number differences representing 76 unique CNPs. On average, individuals differed by 11 CNPs, and the average length of a CNP interval was 465 kilobases. We observed copy number variation of 70 different genes within CNP intervals, including genes involved in neurological function, regulation of cell growth, regulation of metabolism, and several genes known to be associated with disease.


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
D
A
B/C
0
50000000
100000000
150000000
200000000
Chromosome position (base pairs)
Recurrent
Observed once
Novel by Hind III-ROMA
Centromeric gap

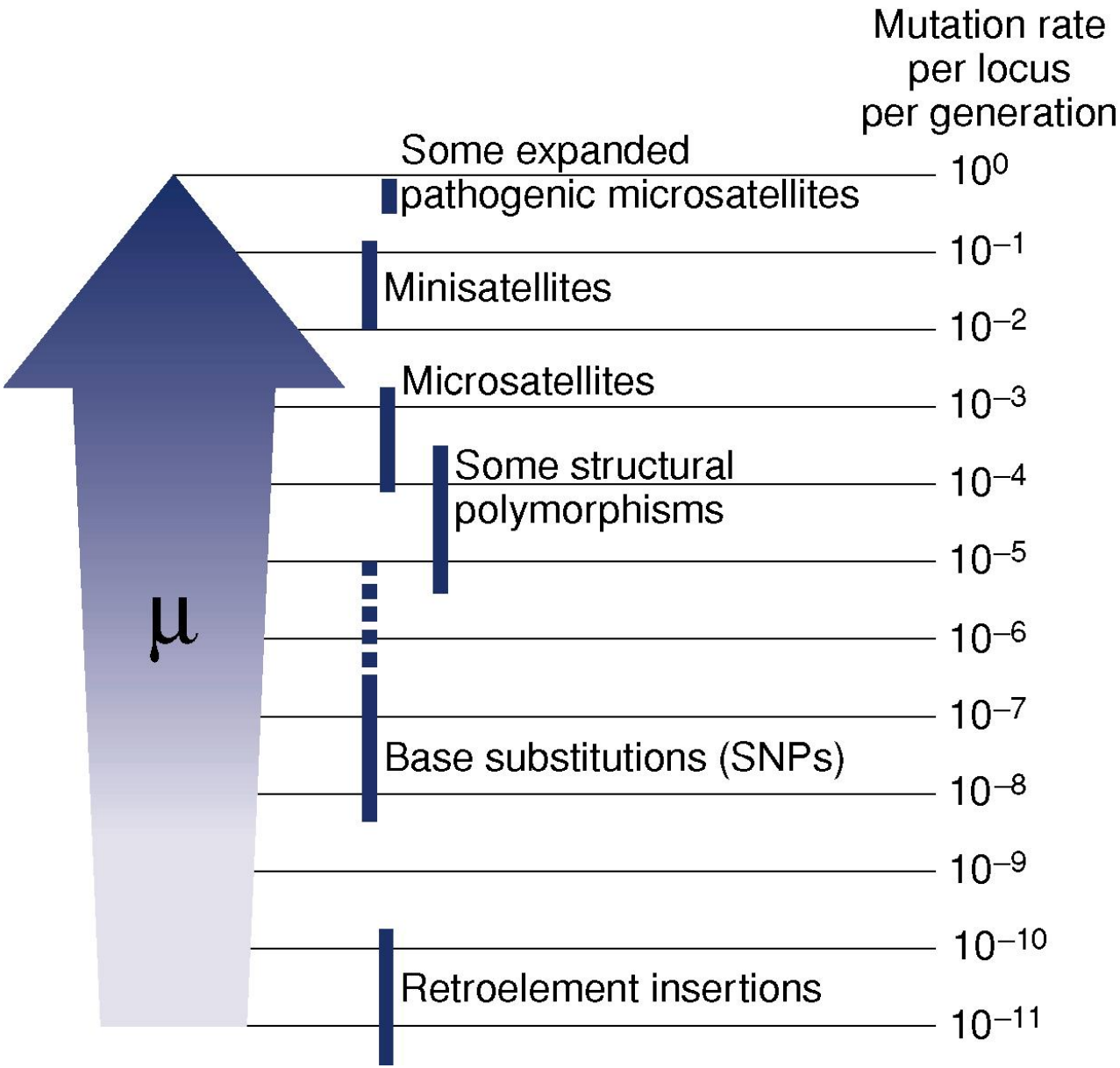

A
Chromosome 11
32K BAC array
1.2
1.1
1
0.9
0.8
0 20 40 60 80 100 120 Mb
B
Chromosome 11
Illumina HumanHap 300 Duo
0.9
0.7
0.5
0.3
0.1
0 20 40 60 80 100 120 Mb
C
Chromosome 11
Illumina HumanHap 300 Duo
Abs. Diff.
0.2
0.1
0
0 20 40 60 80 100 120 Mb
D
Chromosome 4
32K BAC array
1.2
1.1
1
0.9
0.8
0 20 40 60 80 100 120 140 160 180 Mb
E
Chromosome 4
Illumina HumanHap 300 Duo
Abs. Diff.
0.2
0.1
0
0 20 40 60 80 100 120 140 160 180 Mb
F
%
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chr.
G
Chromosome 11
32K BAC array
291, mean = 0.997, SD = 0.044
292, mean = 0.939, SD = 0.047
1.4
1.2
1
0.8
0.6
105 110 115 120 125 Mb
H
1
0.8
0.6
chr1_1 chr7_1 chr7_2 chr9_1 chr9_2 chr11_1 chr11_2 chr11_3 chr11_4 chr11_5 chr11_6 chr12_1 chr12_2 chr14_1 chr14_2 chr14_3 chr15_1 chr18_1 chr18_2 chrX_1 chrX_2 chrX_3 chrX_4
I
291
292
1
0.8
0.6
chr1_1 chr7_1 chr7_2 chr9_1 chr9_2 chr11_1 chr11_2 chr11_3 chr11_4 chr11_5 chr11_6 chr12_1 chr12_2 chr14_1 chr14_2 chr14_3 chr15_1 chr18_1 chr18_2 chrX_1 chrX_2 chrX_3 chrX_4

Acquiring mosaicism.Human development from a single fertilized cell to a multicellular organism requires many cell divisions and the genetic material to be replicated many times.



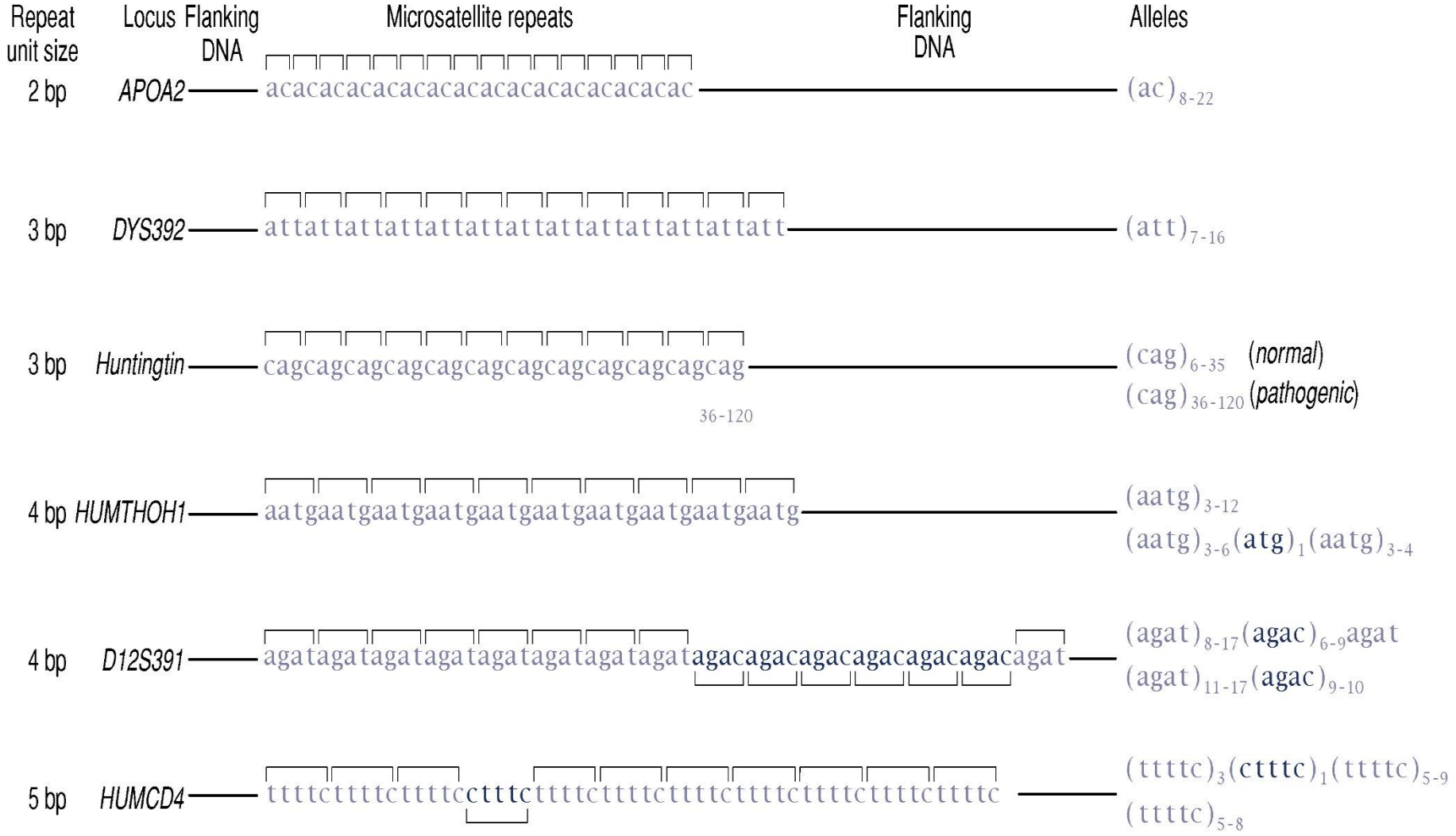
J R Lupski Science 2013;341:358-359

Mutation rate of polymorph sequences (μ)

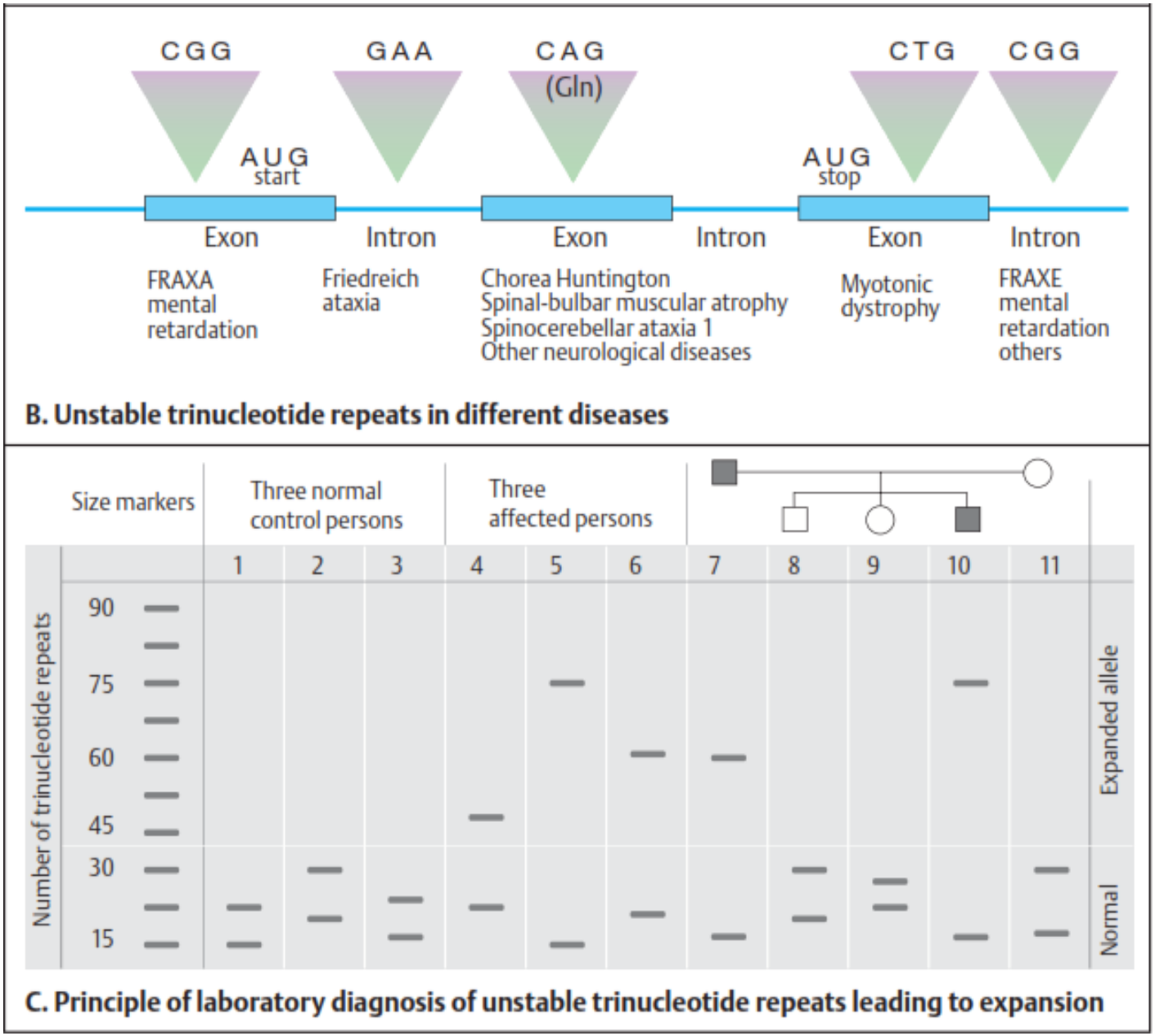


Human Evolutionary Genetics, Jobling, 2004

Microsatellite structure



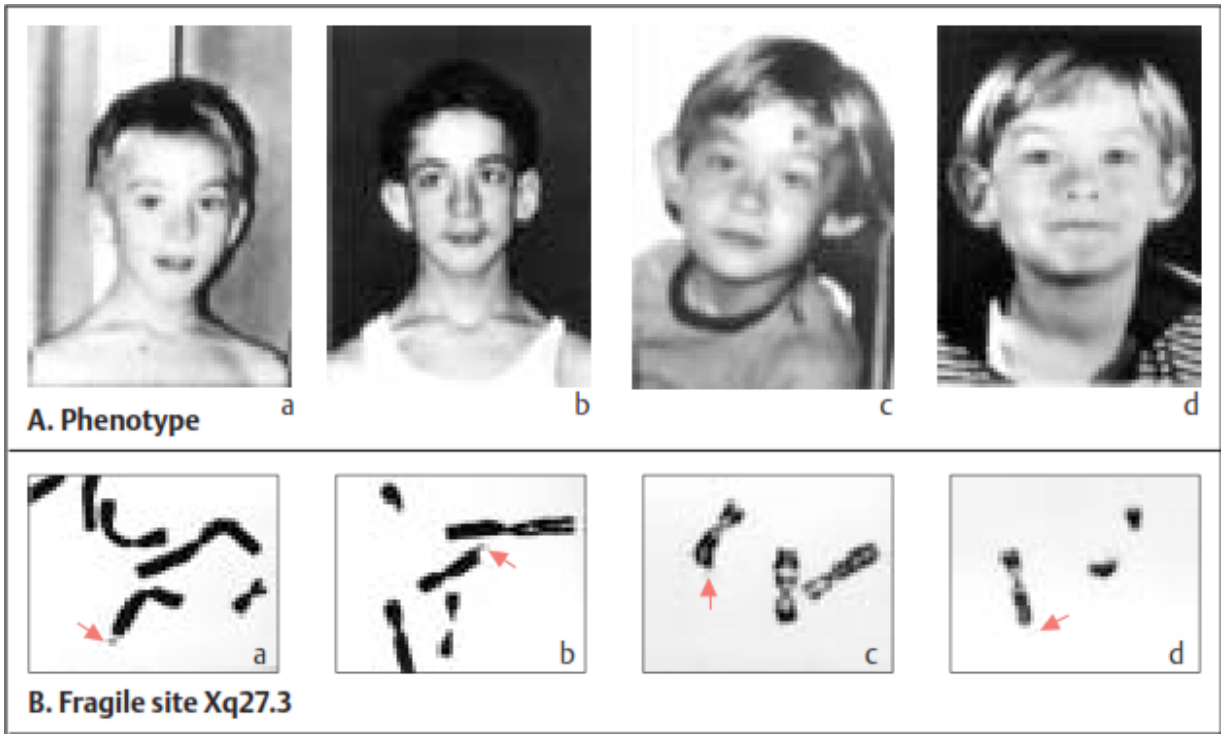
Trinucleotide repeat expansion



Passarge, 2001

Genetic diseases due to repeat expansion

Disease (Examples)	Gene	Frequency	Tri-nucleotide	Normal Number	Mutant Allele	Chromosome
Huntington disease	*HD*	1:10 000	$(CAG)_n$	0–26	36–121	4p16.3
Fragile X syndrome	*FMR1*	1:5 000	$(CGG)_n$	6–50	52–500	Xq27.3
Myotonic dystrophy	*DMPK*	1:8 000	$(CTG)_n$	5–37	50–500	19q13.2
Spinal-bulbar muscular atrophy (Kennedy)	*SBMA*	<1:50 000	$(CAG)_n$	11–31	36–65	Xq11-12



Fragile X

Huntington disease

Myotonic dystrophy

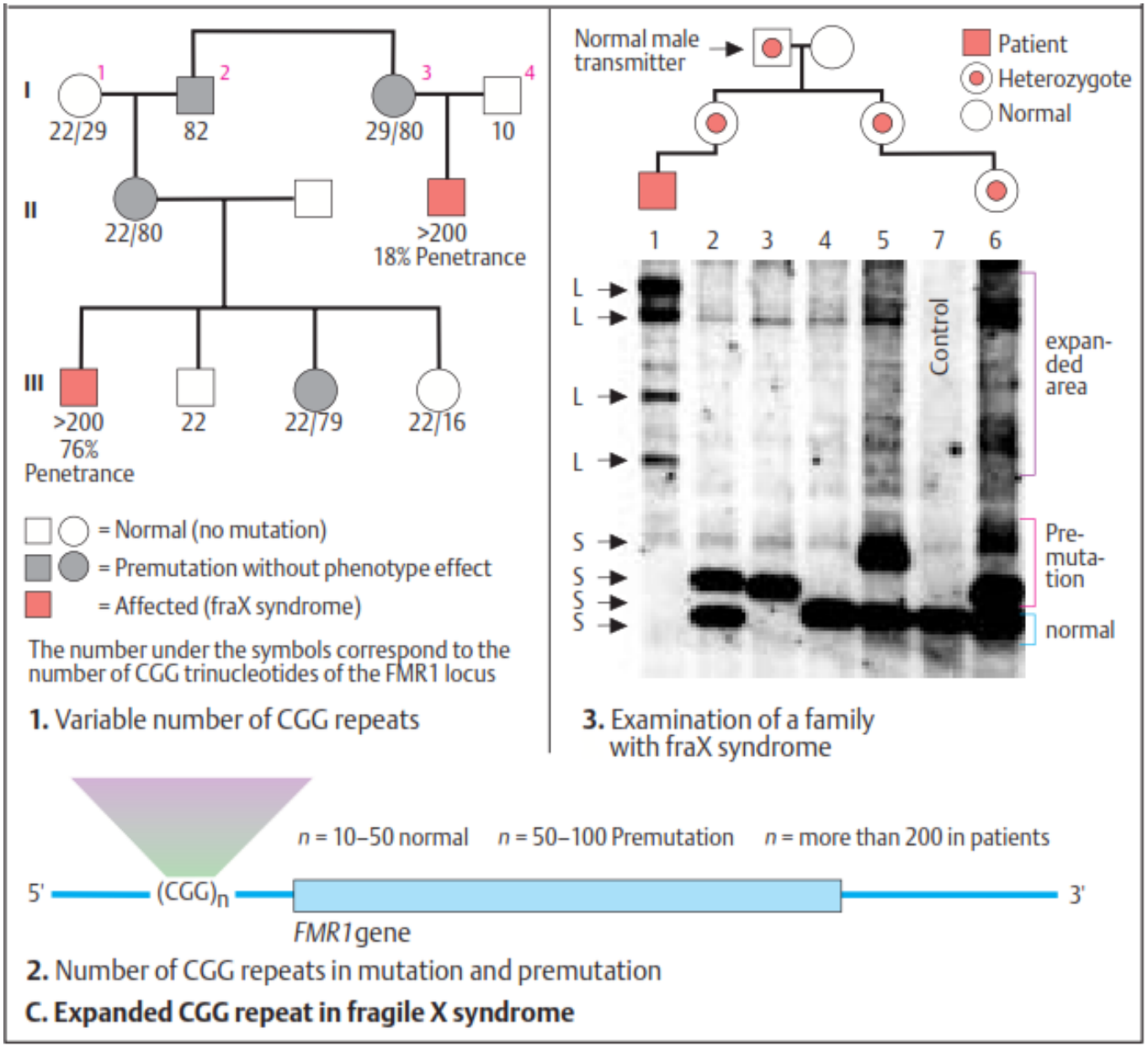
Friedrich ataxia

SMA

etc.

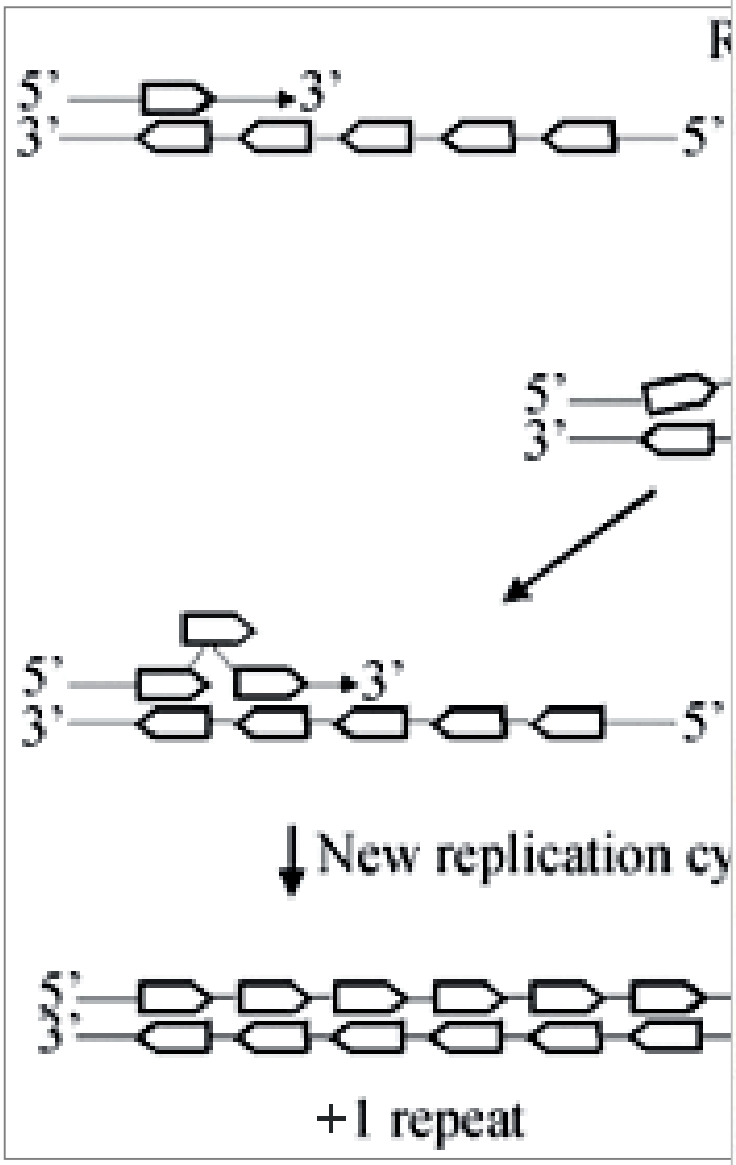
Passarge, 2001

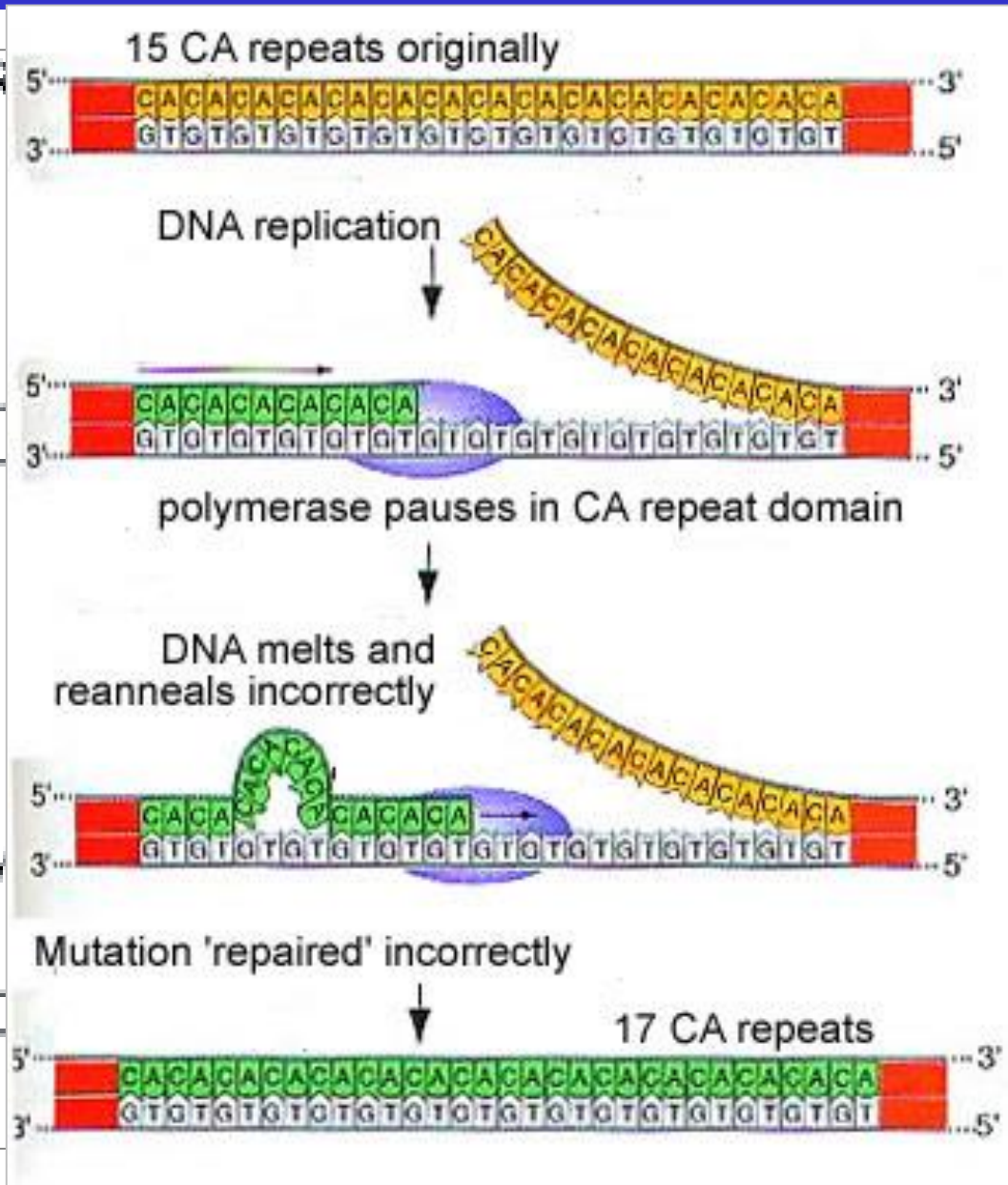
Expanded *CGG* repeats in the Fragile X syndrome



Passarge, 2001

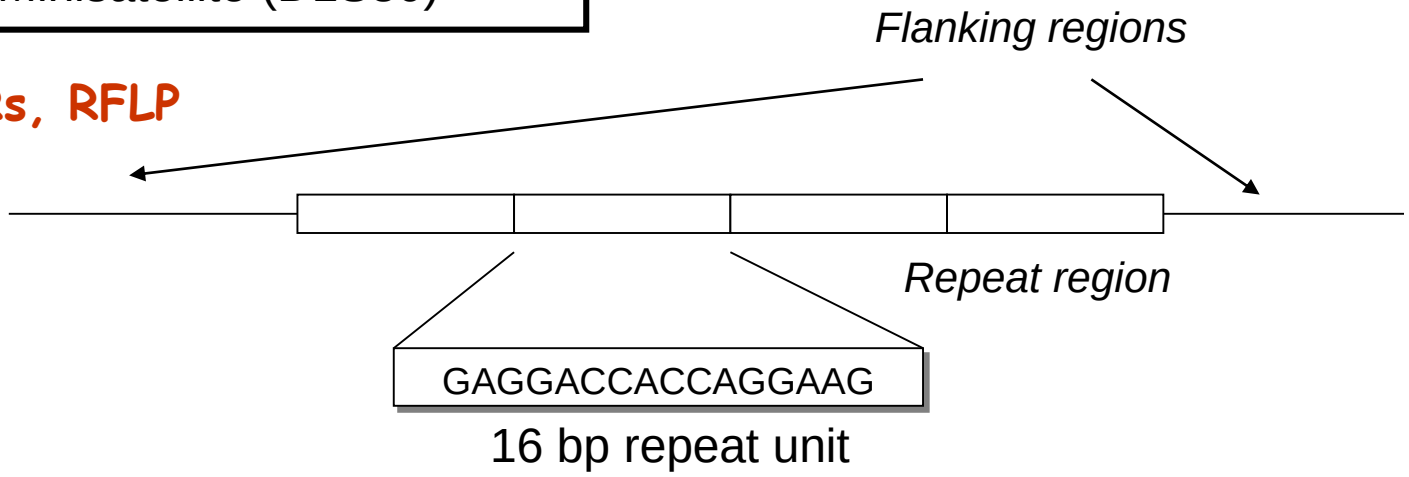
Microsatellite evolution





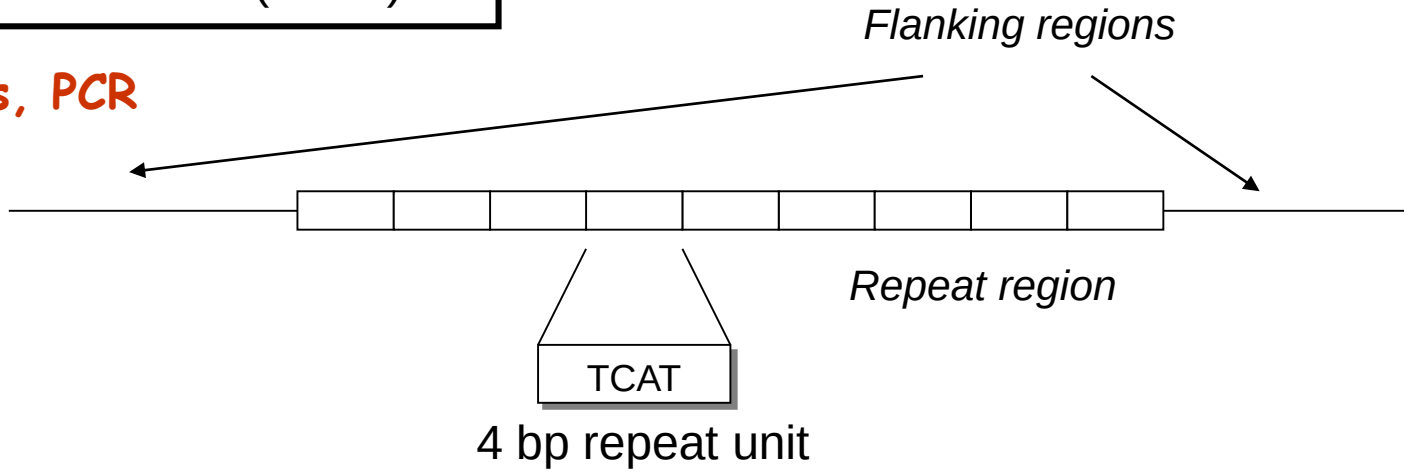
Minisatellite (D1S80)

VNTRs, RFLP



Microsatellite (TH01)

STRs, PCR



Principle of STR allelic ladder formation

Separate PCR products from various samples amplified with primers targeted to a particular STR locus

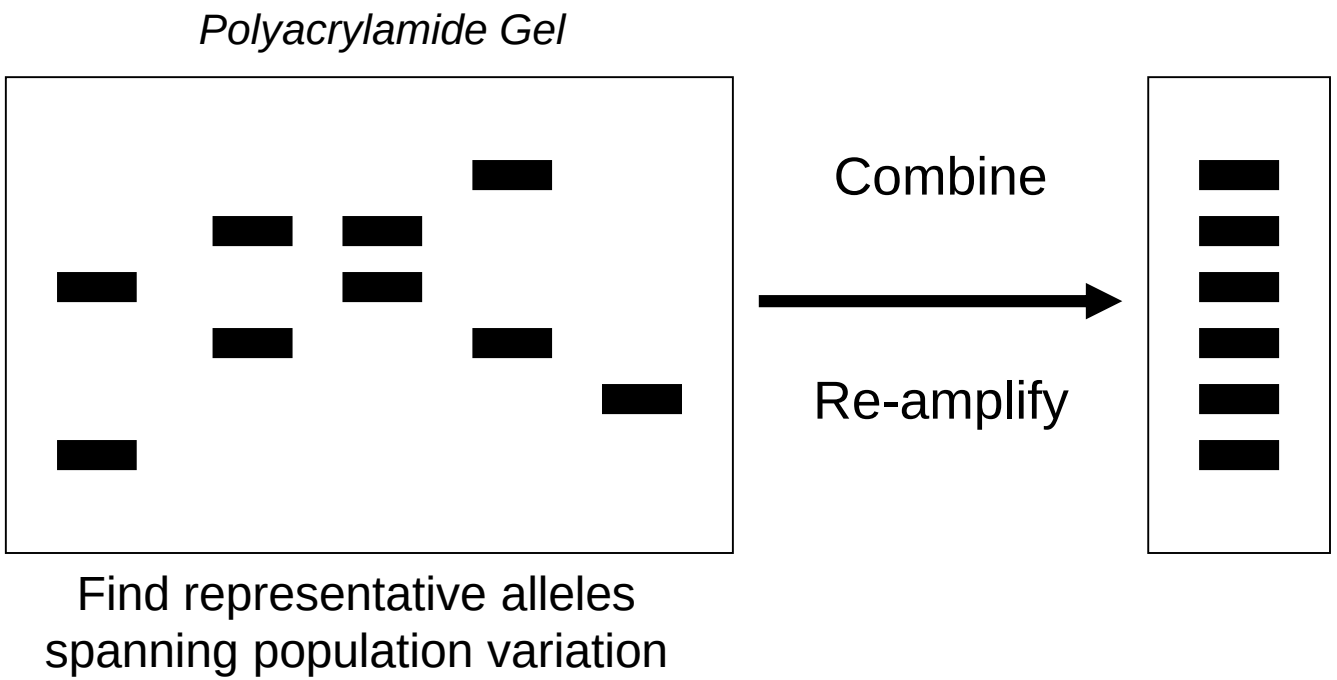
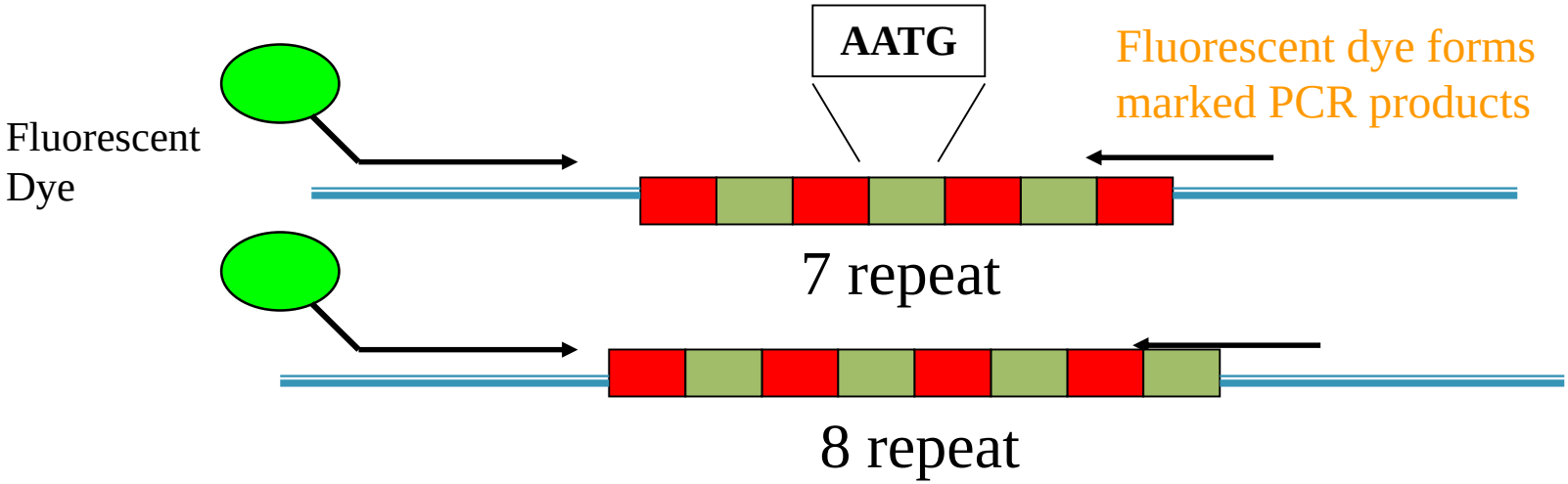


Figure 5.3, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

Microsatellite – STR – marker (Short Tandem Repeat)



Repeat region is variable between tested persons, while flanking region where PCR primers anneal is invariant

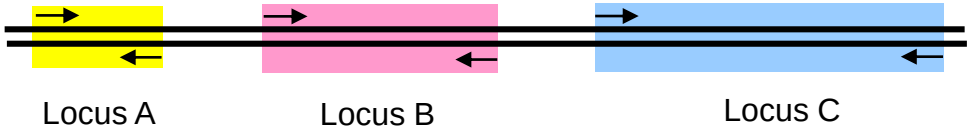
Homozygous = tested alleles are same

Heterozygous = tested alleles differ and can be separeted

Primer binding site determines PCR product size!

Schematic of Multiplex-PCR

(A) Three loci parallel amplification in one reaction



(B) PCR products separated based on fragment size

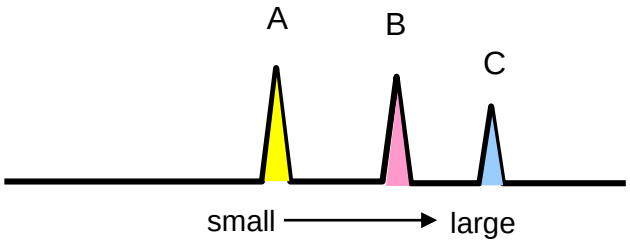
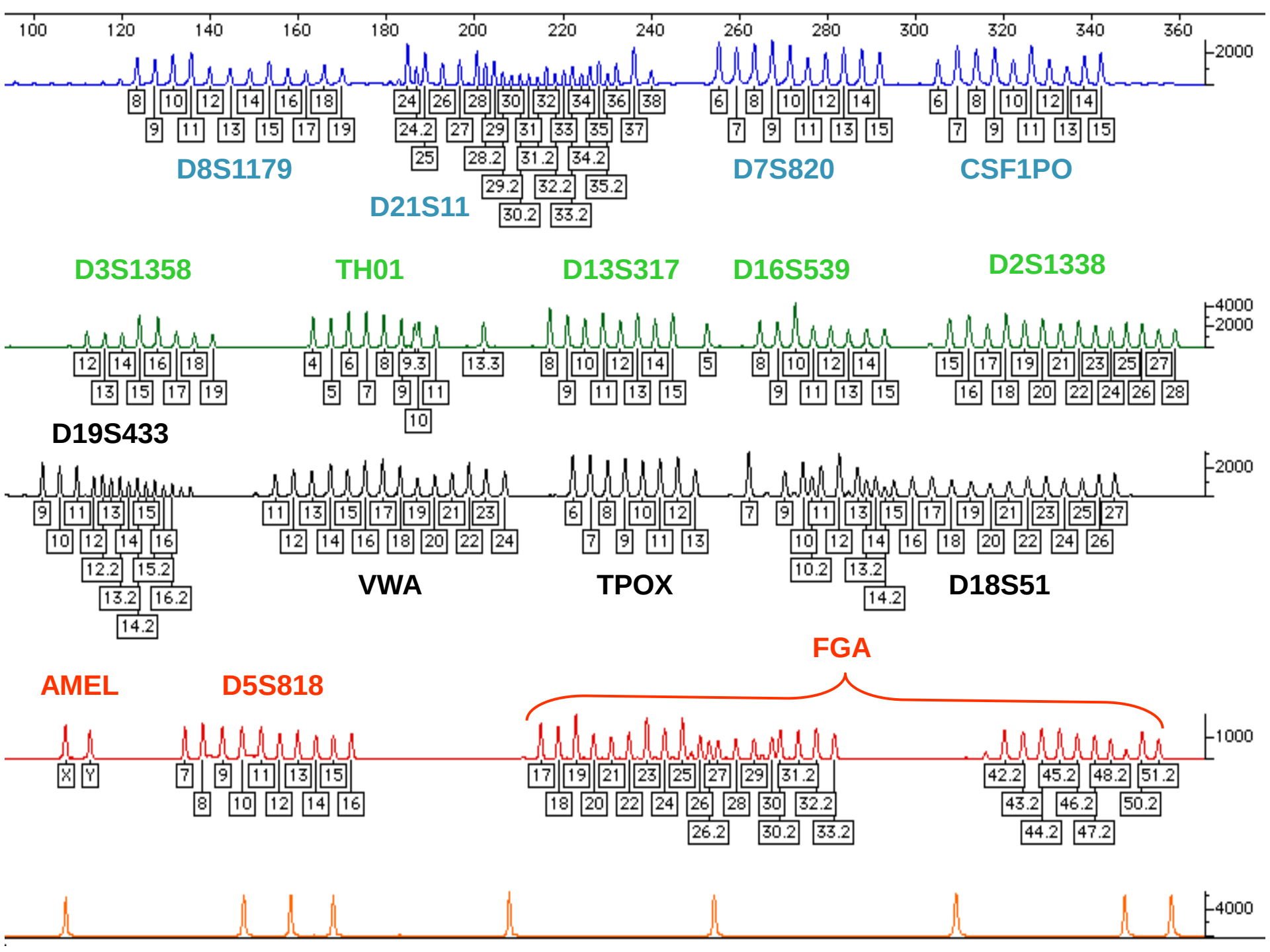
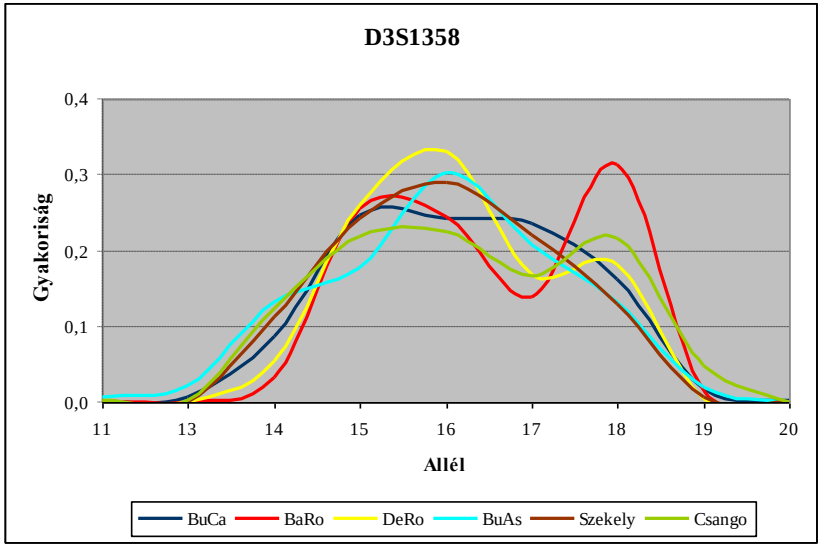
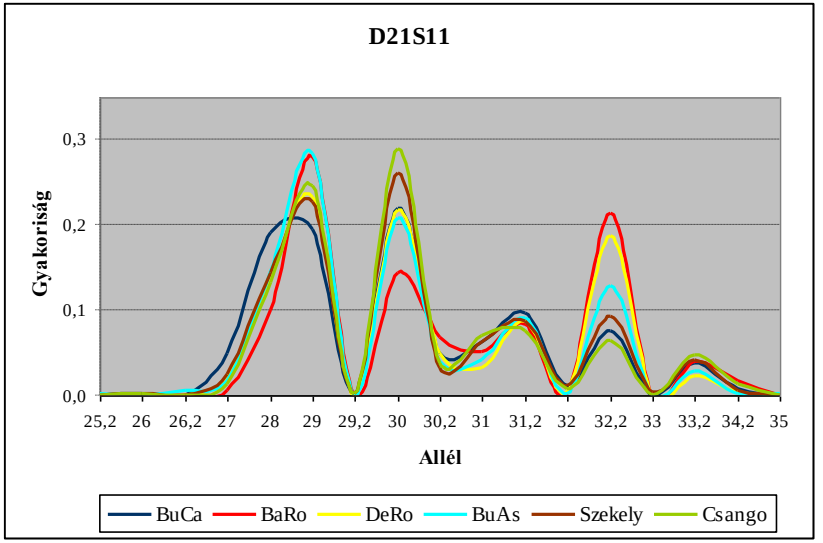


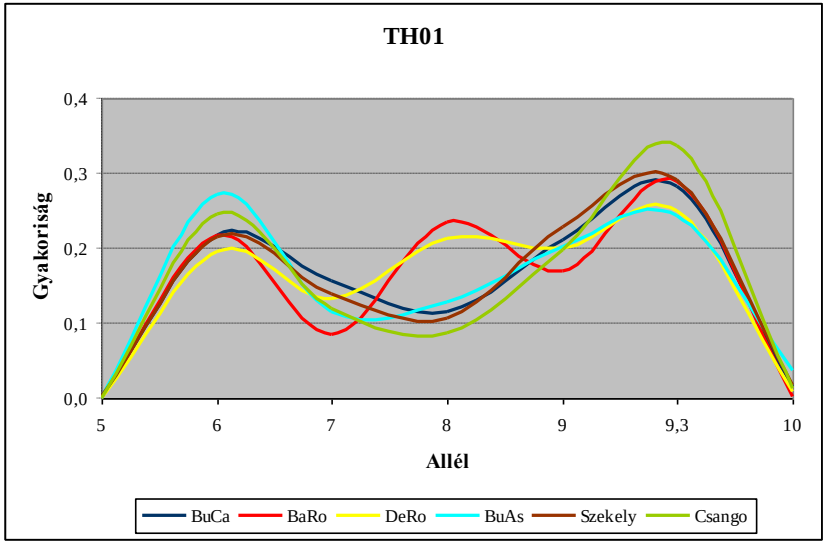
Figure 4.3, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

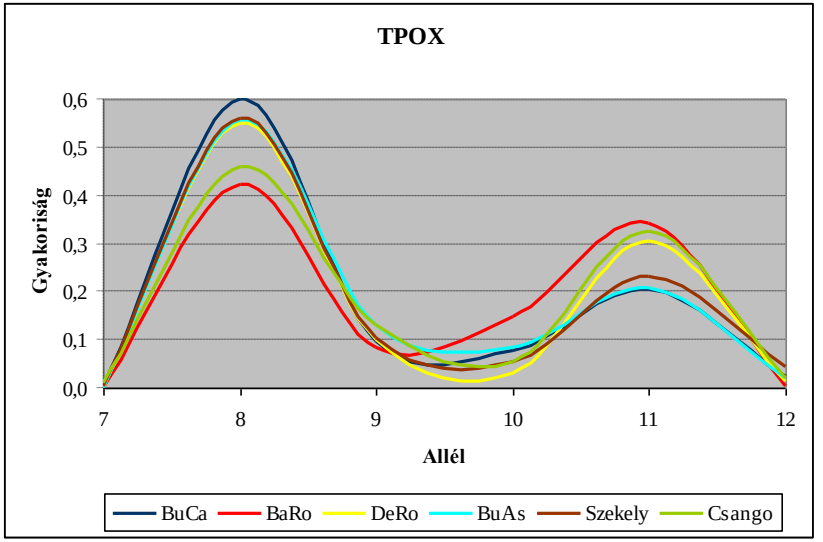

100 120 140 160 180 200 220 240 260 280 300 320 340 360
2000
D8S1179
D21S11
D7S820
CSF1PO
D3S1358
TH01
D13S317
D16S539
D2S1338
4000
2000
D19S433
2000
VWA
TPOX
D18S51
AMEL
D5S818
FGA
1000
4000

Microsatellite allele frequency distributions









How Statistical Calculations are Made

- Generate data with set(s) of samples from desired population group(s)
  - Generally only 100-150 samples are needed to obtain reliable allele frequency estimates

- Determine allele frequencies at each locus
  - Count number of each allele seen

- Allele frequency information is used to estimate the rarity of a particular DNA profile
  - Homozygotes (p^2), Heterozygotes (2pq)
  - Product rule used (multiply locus frequency estimates, PM = (P1)(P2)…(Pn))

Assumptions with Hardy-Weinberg Equilibrium

The Assumption	The Reason
Large population	Lots of possible allele combinations
No natural selection	No restriction on mating so all alleles have equal chance of becoming part of next generation
No mutation	No new alleles being introduced
No immigration/emigration	No new alleles being introduced or leaving
Random mating	Any allele combination is possible

None of these assumptions are really true…

Table 20.6, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

STR Cumulative Profile Frequency with Multiple Population Databases

STR Locus	Profile Computed	Number of Populations Used	Cumulative Profile Frequency Range (1 in ...)	Cumulative Profile Frequency against U.S. Caucasians (Appendix II)
D3S1358	16,17	166	5.24 to 62.6	9.19
VWA	17,18	166	37.6 to 1080	81.8
FGA	21,22	166	737 to 119 000	1010
D8S1179	12,14	166	8980 to 5 430 000	16 400
D21S11	28,30	166	165 000 to 248 000 000	186 000
D18S51	14,16	166	3.85×10^{6} to 2.68×10^{10}	4.88×10^{6}
D5S818	12,13	166	2.28×10^{7} to 4.22×10^{11}	4.51×10^{7}
D13S317	11,14	166	4.32×10^{8} to 1.69×10^{13}	1.38×10^{9}
D7S820	9,9	166	1.17×10^{10} to 2.98×10^{16}	4.22×10^{10}
D16S539	9,11	97	4.06×10^{11} to 1.11×10^{18}	5.82×10^{11}
TH01	6,6	97	9.30×10^{12} to 1.45×10^{19}	1.05×10^{13}
TPOX	8,8	97	3.33×10^{13} to 1.54×10^{20}	3.63×10^{13}
CSF1PO	10,10	97	3.43×10^{14} to 2.65×10^{21}	7.43×10^{14}

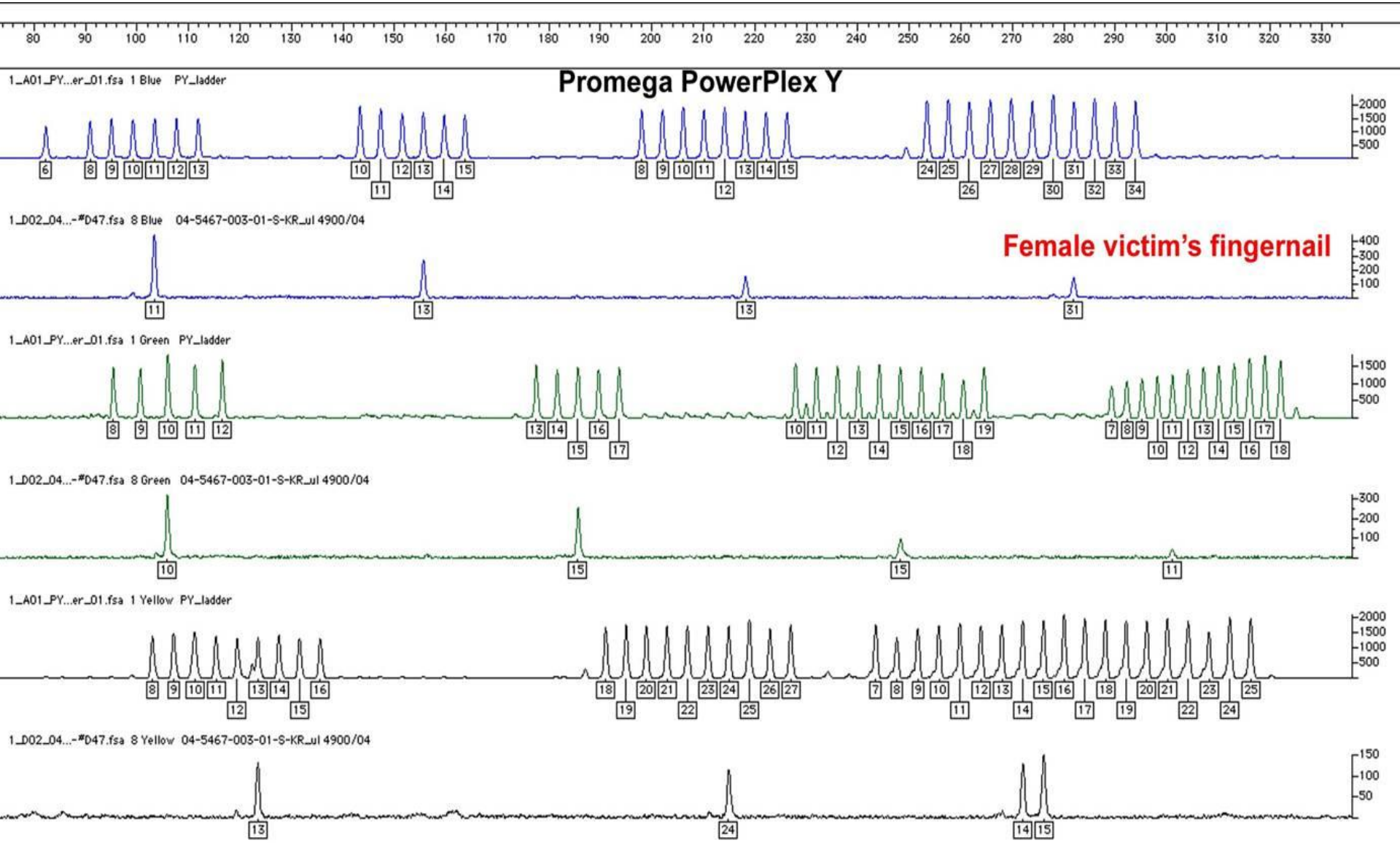
10^{14} to 10^{21}

D.N.A. Box 21.1, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

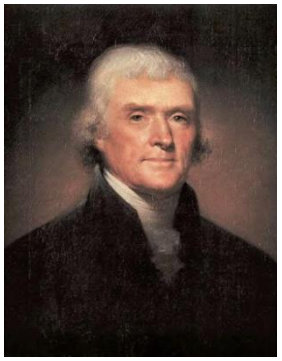
STR System	Maternal Meioses (%)	Paternal Meioses (%)	Number from either	Total Number of Mutations	Mutation Rate
CSF1PO	95/304,307 (0.03)	982/643,118 (0.15)	410	1,487/947,425	**0.16%**
FGA	205/408,230 (0.05)	2,210/692,776 (0.32)	710	3,125/1,101,006	**0.28%**
TH01	31/327,172 (0.009)	41/452,382 (0.009)	28	100/779,554	**0.01%**
TPOX	18/400,061 (0.004)	54/457,420 (0.012)	28	100/857,481	**0.01%**
VWA	184/564,398 (0.03)	1,482/873,547 (0.17)	814	2,480/1,437,945	**0.17%**
D3S1358	60/405,452 (0.015)	713/558,836 (0.13)	379	1,152/964,288	**0.12%**
D5S818	111/451,736 (0.025)	763/655,603 (0.12)	385	1,259/1,107,339	**0.11%**
D7S820	59/440,562 (0.013)	745/644,743 (0.12)	285	1,089/1,085,305	**0.10%**
D8S1179	96/409,869 (0.02)	779/489,968 (0.16)	364	1,239/899,837	**0.14%**
D13S317	192/482,136 (0.04)	881/621,146 (0.14)	485	1,558/1,103,282	**0.14%**
D16S539	129/467,774 (0.03)	540/494,465 (0.11)	372	1,041/962,239	**0.11%**
D18S51	186/296,244 (0.06)	1,094/494,098 (0.22)	466	1,746/790,342	**0.22%**
D21S11	464/435,388 (0.11)	772/526,708 (0.15)	580	1,816/962,096	**0.19%**
Penta D	12/18,701 (0.06)	21/22,501 (0.09)	24	57/41,202	**0.14%**
Penta E	29/44,311 (0.065)	75/55,719 (0.135)	59	163/100,030	**0.16%**
D2S1338	15/72,830 (0.021)	157/152,310 (0.10)	90	262/225,140	**0.12%**
D19S433	38/70,001 (0.05)	78/103,489 (0.075)	71	187/173,490	**0.11%**
SE33 (ACTBP2)	0/330 (<0.30)	330/51,610 (0.64)	None reported	330/51,940	**0.64%**

Mutation rate of generally used STR markers: 10^{-3} - 10^{-4} / meiosis

Y chromosome microsatellite profile



Genetic History



Thomas Jefferson II

Field Jefferson

Peter Jefferson

President Thomas Jefferson

?

Eston Hemings

Same Y Haplotype

John Weeks Jefferson born in 1946

Thomas Woodson

Different Y Haplotype

Jefferson Y Haplotype

Jefferson Y Haplotype

Figure 9.10, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

DNA Marker Tested	Field Jefferson Male-Line	Eston Hemings Male-Line	John Carr Male-Line	Thomas Woodson Male-Line
Number of individuals typed	5	1	3	5
Y STR Loci				
DYS19	15	15	14 ←	14 ←
DYS388	12	12	12	12
DYS389A	4	4	5 ←	5 ←
DYS389B	11	11	12 ←	11
DYS389C	3	3	3	3
DYS389D	9	9	10 ←	10 ←
DYS390	11	11	11	11
DYS391	10	10	10	13 ←
DYS392	15	15	13 ←	13 ←
DYS393	13	13	13	13
DXYS156Y	7	7	7	7
Y SNP Loci	(0 = ancestral state; 1 = derived state)			
DYS287 (YAP)	0	0	0	0
SRYm8299	0	0	0	0
DYS271 (SY81)	0	0	0	0
LLY22g	0	0	0	0
Tat	0	0	0	0
92R7	0	0	1 ←	1 ←
SRYm1532	1	1	1	1
Minisatellite Locus				
MSY1	(3)–5	(3)–5		
	(1)–14	(1)–14	(1)–17 ←	(1)–16 ←
	(3)–32	(3)–32	(3)–36 ←	(3)–27 ←
	(4)–16	(4)–16	(4)–21 ←	(4)–21 ←

Table 9.8, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385
16	13	31	25	12	11	13	14.16

National database | Metapopulations | SNP

Whole database

Search

DYS438	DYS439	DYS437	DYS448	DYS456	DYS458	DYS635	YGATAH4
10	13	15	20	15	18	23	11

Please note: The database size will vary based on the loci you have entered.

- 7 loci haplotype (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393): **101055 haplotypes**
- 9 loci haplotype (+ DYS385a/b): **99258 haplotypes**
- 11 loci haplotype (+ DYS438, DYS439): **72171 haplotypes**
- 12 loci haplotype (+ DYS437): **52628 haplotypes**
- 17 loci haplotype (+ DYS448, DYS456, DYS458, DYS635, YGATAH4): **40987 haplotypes**

Y-SNPs:

- **124 Y-SNP** branches (defined by **134 Y-SNP markers**)
- **9039 haplotypes** with Y-SNP information

Search

Haplotypes
SNPs
Populations
Contributors
Contributions

Analyse | Research | Contribute | Meet

DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385
16	13	31	25	12	11	13	14.16

National database | Metapopulations | SNP

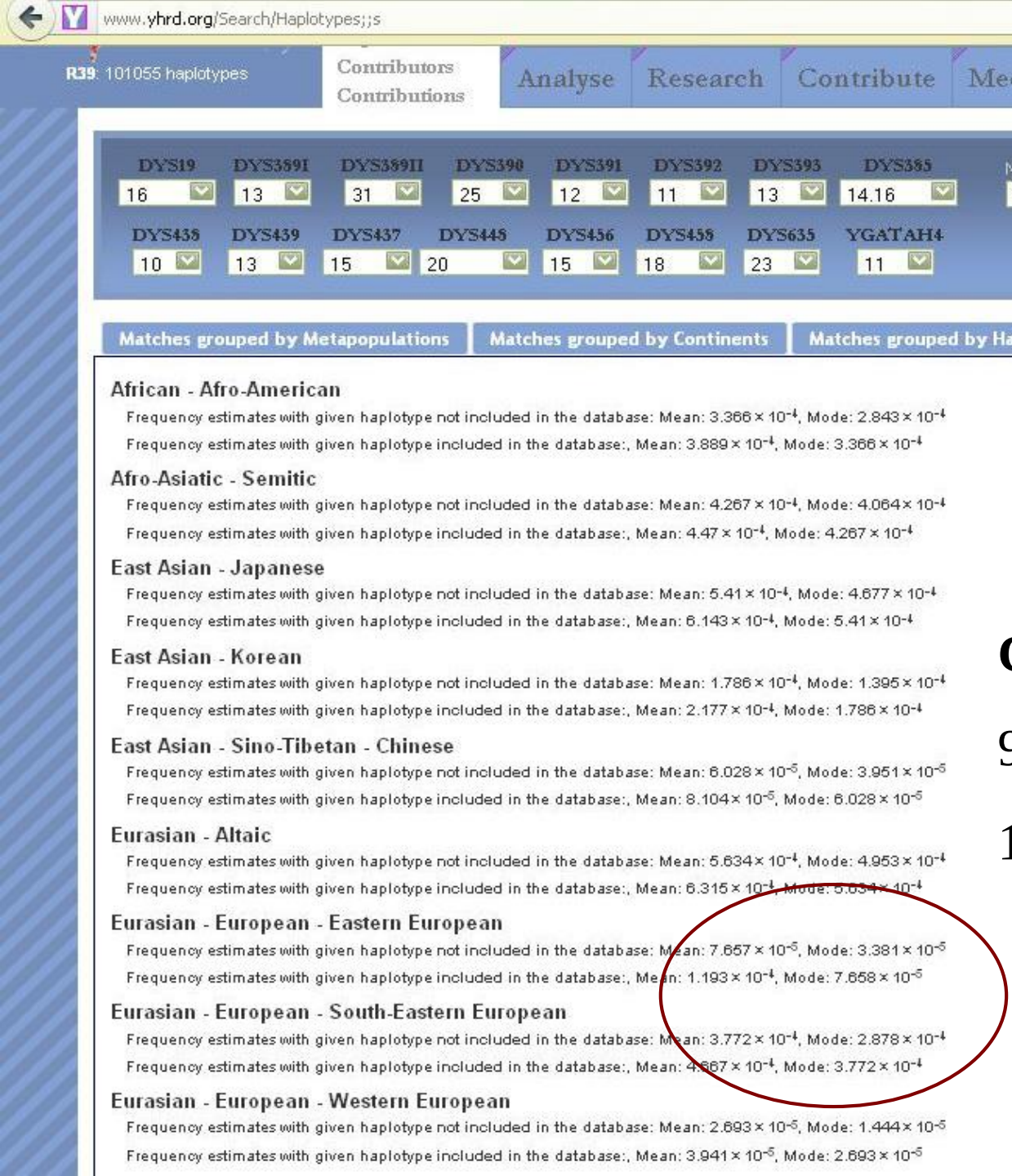
Whole database

Search

DYS438	DYS439	DYS437	DYS448	DYS456	DYS458	DYS635	YGATAH4
10	13	15	20	15	18	23	11

Matches grouped by Metapopulations | Matches grouped by Continents | Matches grouped by Haplogroups | Frequency surveying estimates

- **All Metapopulation:** Found 0 of 40987 matching haplotypes [f=0 (95% CI: 0 – 9×10^{-5})] in 0 of 282 populations.
  - **Eurasian Metapopulation:** Found 0 of 16733 matching haplotypes [f=0 (95% CI: 0 – 2.204×10^{-4})] in 0 of 130 populations.
  - **East Asian Metapopulation:** Found 0 of 12674 matching haplotypes [f=0 (95% CI: 0 – 2.91×10^{-4})] in 0 of 64 populations.
  - **Australian Aboriginal Metapopulation:** Found 0 of 766 matching haplotypes [f=0 (95% CI: 0 – 4.804×10^{-3})] in 0 of 1 populations.
  - **African Metapopulation:** Found 0 of 1533 matching haplotypes [f=0 (95% CI: 0 – 2.403×10^{-3})] in 0 of 10 populations.
  - **Native American Metapopulation:** Found 0 of 384 matching haplotypes [f=0 (95% CI: 0 – 9.56×10^{-3})] in 0 of 9 populations.
  - **Eskimo Aleut Metapopulation:** Found 0 of 301 matching haplotypes [f=0 (95% CI: 0 – 1.218×10^{-2})] in 0 of 2 populations.
  - **Afro-Asiatic Metapopulation:** Found 0 of 1854 matching haplotypes [f=0 (95% CI: 0 – 1.988×10^{-3})] in 0 of 21 populations.
  - **Admixed Metapopulation:** Found 0 of 6742 matching haplotypes [f=0 (95% CI: 0 – 5.47×10^{-4})] in 0 of 45 populations.



Counting Method

95% confidence interval

$1-(0.05)^{1/N} \approx 3/N$