Variations in biological sequences

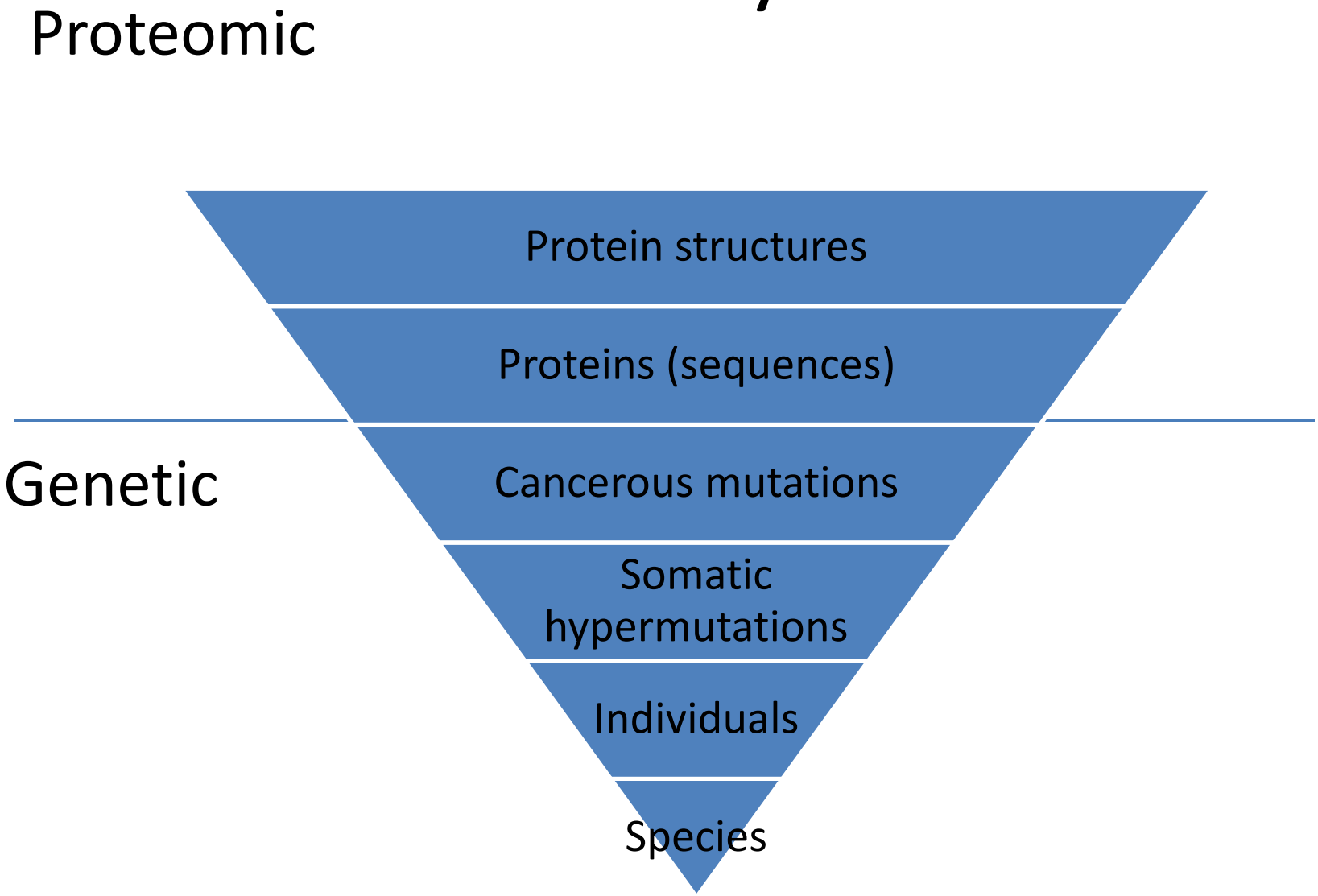
Peter Antal
antal@mit.bme.hu

2012.12.30.

Overview

- Variations in biological sequences
  - Species, individuals, cancerous cells, immune cells, (protein seqs/structs)
  - Substitutions, indels, inversions, duplications, repeats, rearrangements, fractals
  - Dependencies/independencies/patterns of variations
    - SNPs, TFs, domains
  - **PRACTICAL STUDY DESIGN ;-)**
- Models
  - A stochastic single state model:
    - Optimality, matching
    - Phylogeny, PAM, BLOSUM
  - Indels, gaps
    - Alignment, dynamic programming, BLAST
  - Position-specific scoring
    - Stochastic finite-state automaton
    - Hidden Markov Models, HMMer, multiple alignment
  - Chomsky hierarchy…

Variations in biological sequences and beyond



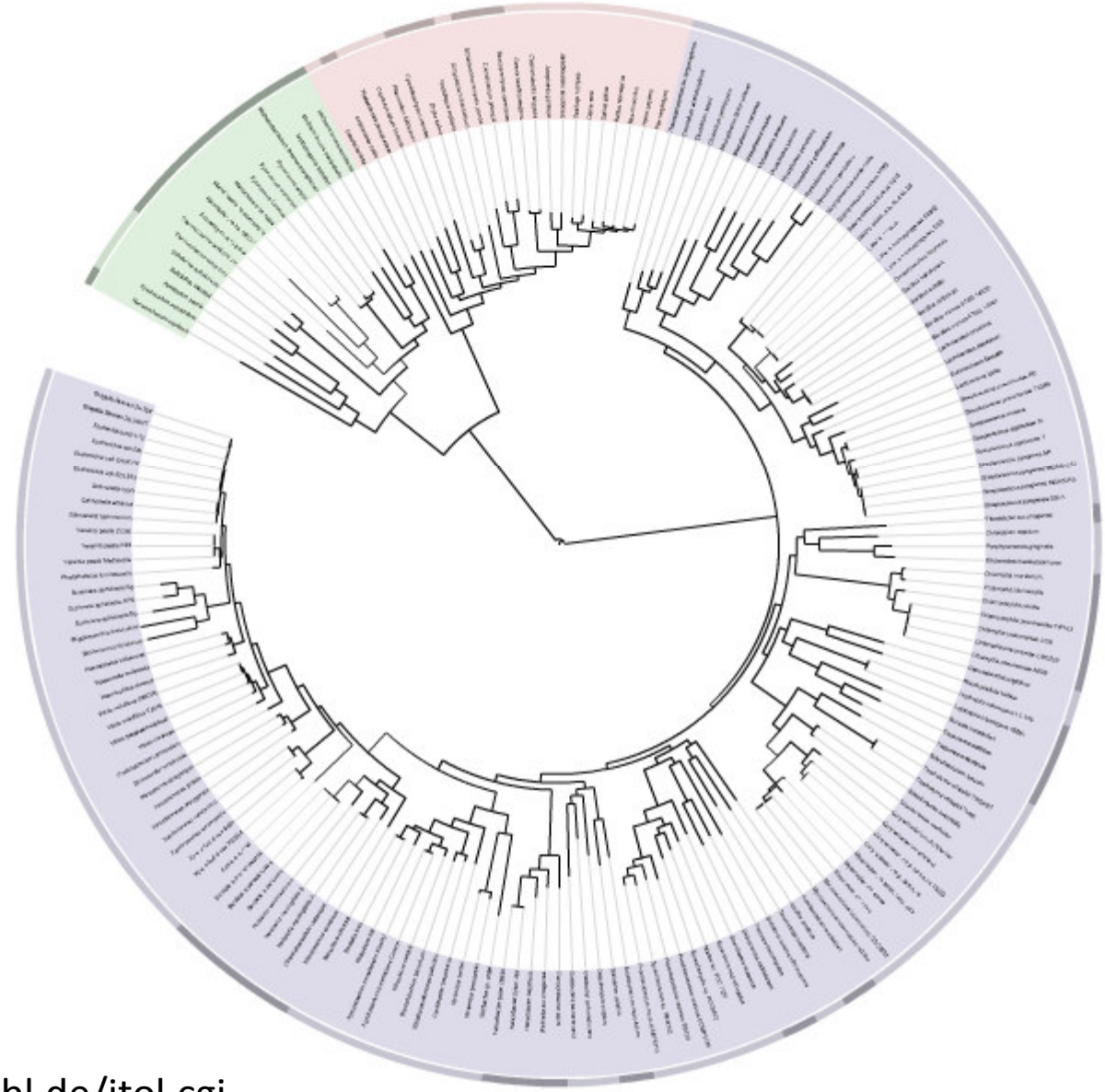
Variations

- Phylogenetic – Species (multi-cellular: ~10^6)
- Inherited – Individuals (humans: ~10^{10})
  - Genetic (SNPs, CNVs,..)
  - Epigenetic (e.g. methylome, histone)
- Somatic – Cells (in humans: ~10^{14}, ~10^{16} , <bacteria!)
  - T/B-cell repertoire (in humans: ~10^{12})
  - Cancerous cells (in humans: ~10^{10})
- Proteomic – Proteins
  - Alternative splicing
  - Post-translational modifications
- Metabolome - (in humans: ~10^4)
- Glycome (carbohydrates/saccharides) (in humans: ~∞)

The monophyletic view of life

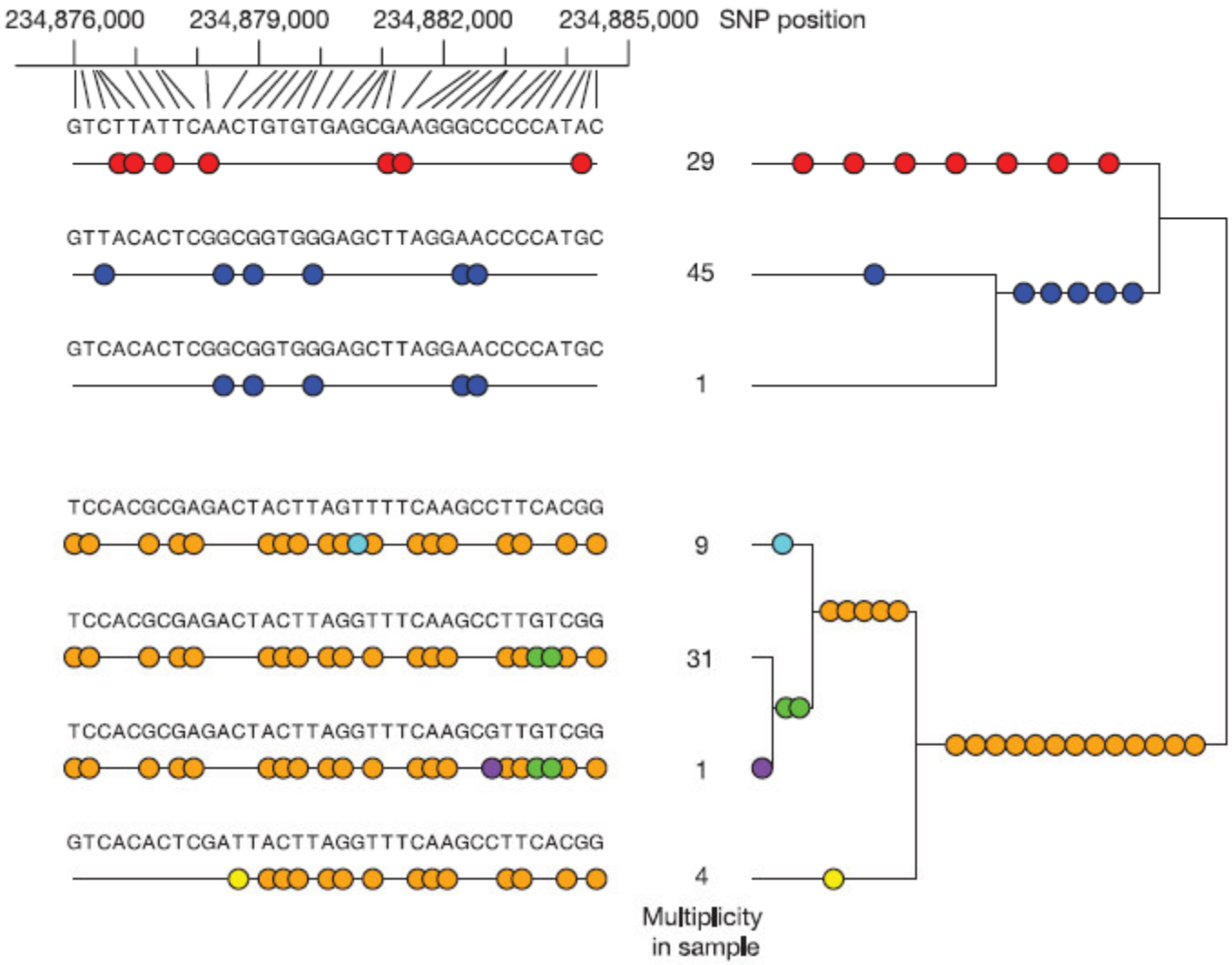
- C.Darwin: „A history of the world, imperfectly kept and written in a changing dialect. Of this history we posses the last volume alone. Of this volume, only here and there a short chapter has been preserved; and of each page only here and there a few lines.”

Tree Of Life



http://itol.embl.de/itol.cgi

Genealogy and variations



Mutation rates

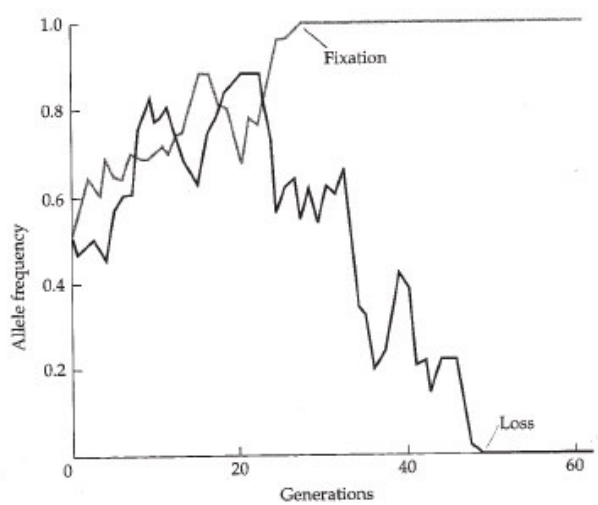
- Error rate in DNA replication is $10^{-9}/10^{-7}$ with/without mismatch repair (consensus belief is 10^{-12} site/cell generation, but in certain genes can be as high 10^{-7}.
- Note that the human body contains 10^{13} cells with limited lifetime...).
- Average rate in mammalian DNA is 10^{-9} substitution/site/year, but mitochondrial is 10x,
- Functional constraints on mutations.

Population genetics

Locus is the genomic location of a gene, alternatives (1%<) are the alleles.
Natural selection is the differential reproduction of different genotypes (through different phenotypes). Fitness of $A_1\ A_1$,$A_1\ A_1$,$A_2\ A_2$ diploids is denoted with w_{ij} or $1 + s_{ij}$. If the frequency for A_1 (old) is p and A_2 $q = 1 - p$, then the Hardy-Weinberg equilibrium is $p^2, 2pq, q^2$.

$$q_{t+1} = \frac{pqw_{12} + q^2 w_{22}}{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}} \Rightarrow q(t) = \frac{1}{1 + (\frac{1-q_0}{q_0})e^{-st}} \qquad (1)$$

Random genetic drift is the consequence of sampling a small population (fixation or loss).



Fixation probability for 1,1+s,1+2s

$$\begin{aligned} p &= \frac{1 - e^{-4Nsq}}{1 - e^{-4Ns}} \\ &\approx_{s=0} \frac{1}{2N} \\ &\approx_{s\neq 0} 2s \end{aligned}$$

Fixation time, conditional fixation time ($\approx 4N$)

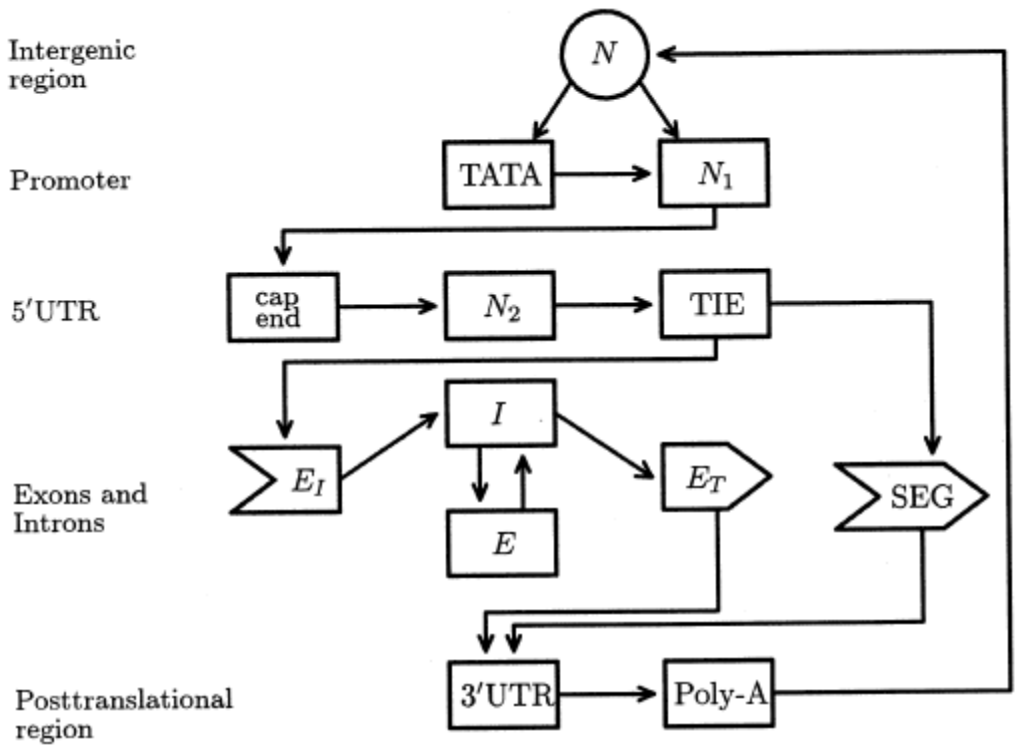
Some terminology

- Point or segmental mutations (deletion, insertion, inversion, duplication, repetations).
- Homology(common ancestor)
  - Orthology (speciation)
  - Paralogy (gene duplication)
- Point mutations can be
  - transition/transversion within/between purins (A,G) and pyrimidins (T,C).
  - synonymous(silent)/nonsynonymous:missense/nonsense(preter mination codons).
  - non-degenerate site or 2/3/4fold degenerate site
  - Frameshift mutation (reading frame, open reading frame)
  - Alternative splicing, miRNA, transcription factors,...

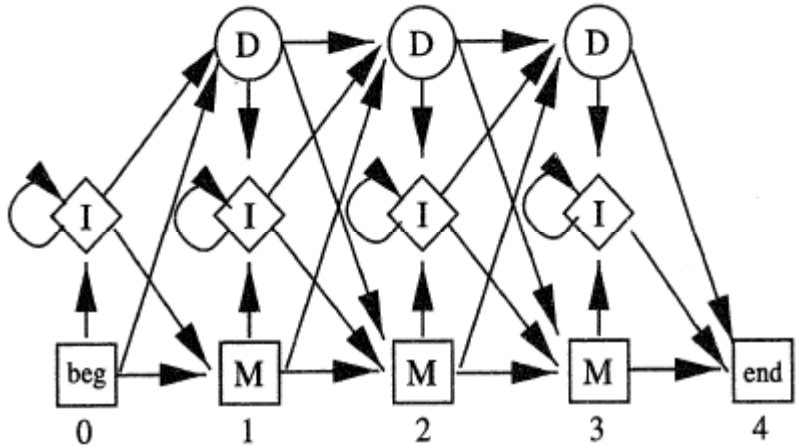
Quantifying sequence similarity

- Distances between strings: edit distance
  - Hamming/Manhattan distance
  - Levenshtein distance: minimum number of insertion, deletion, or substitution of a single character
- Cost matrices: PAM, BLOSUM
  - Distance in simple case: $d(x,y)=\sum_i d(x_i,y_i)$
  - Alignment
    - Global, semi-local, local, multiple (*BLAST*)
    - Hidden Markov Models (profileHMM,...)

(Semi-)Hidden Markov Models



Genescan
(for gene detection)



ProfileHMM
(for protein families)

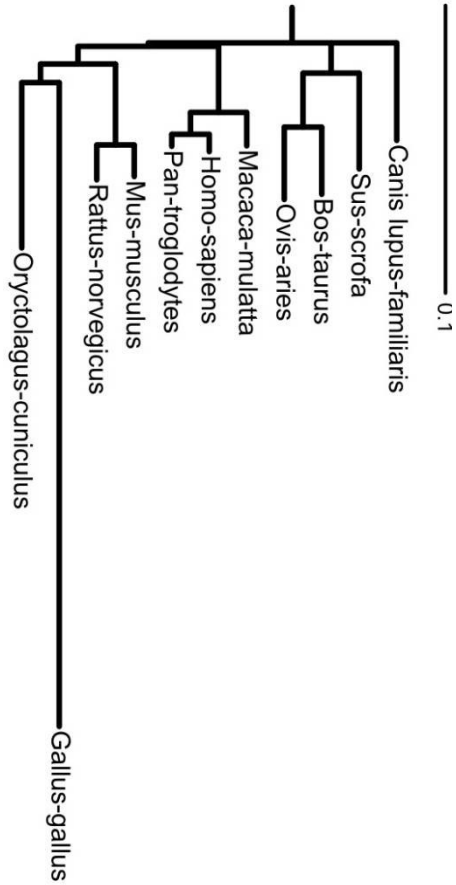
Molecular phylogeny

Goal: **pheneticist** vs **cladist**
Goal': topology, labeling, extinct characters/sequences

Sources for phylogeny: phenotype based, paleontology, geologic,

Molecular phylogeny include/uses molecular relations in the reconstruction:
- Immune response (1902),
- cross-hybridization of denaturated genomes (1950<),
- protein sequencing (1960) ➔ **Molecular clock hypothesis!**
    - rate=substitution/2xTime
    - substitution =-3/4ln(1-4/3 Difference)
    - Zuckerkandl,Pauling: homologous proteins

Recall: standard pointwise approach, pairwise alignment, multiple alignment, PAM, BLOSUM, BLAST….. From multiple alignment to phylogeny: are the implied distances from a multiple alignment consistent?



Construction of phylogenic tree

Data: characters and distances. Characters can represent not just genomic, but phenotypic information as well. Distances can represent (dis)similarity-score?, |difference|, surrogate distances such as |substitutions|, unit-time, time. Note the easy combination/fusion of distances.
Any ultrametric distance is tree-derived (see proof of **Ultrametric Tree Learning**).
Tree can be reconstructed correctly with the **UPGMA**, but it can fail for not ultrametric cases.
For additive, but not ultrametric distances the **neighbour-joining** method can reconstruct the correct labelled tree.
Two further opposite families are the **maximum parsimony** (without evolutionary theory and time labelling) and **maximum likelihood** methods (with evolutionary theory and time labelling).
Methods:

1. Ultrametric reconstruction
2. UPGMA
3. neighbour-joining
4. maximum parsimony
5. maximum likelihood

UPGMA

UPGMA: unweighted pair group method using arithmetic averages.

Define distance d_{ij} between clusters C_i, C_j as the average of between-pairwise distances:

$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$ (variants use min,max instead of arithmetic mean.).

Require: pairwise distances

Ensure: topology and labelling

Ini: Define a cluster C_i and leaf at height 0 in tree T for each sequence i

for i=1 to L-1 **do**

Select pair of clusters i,j with minimal d_{ij}

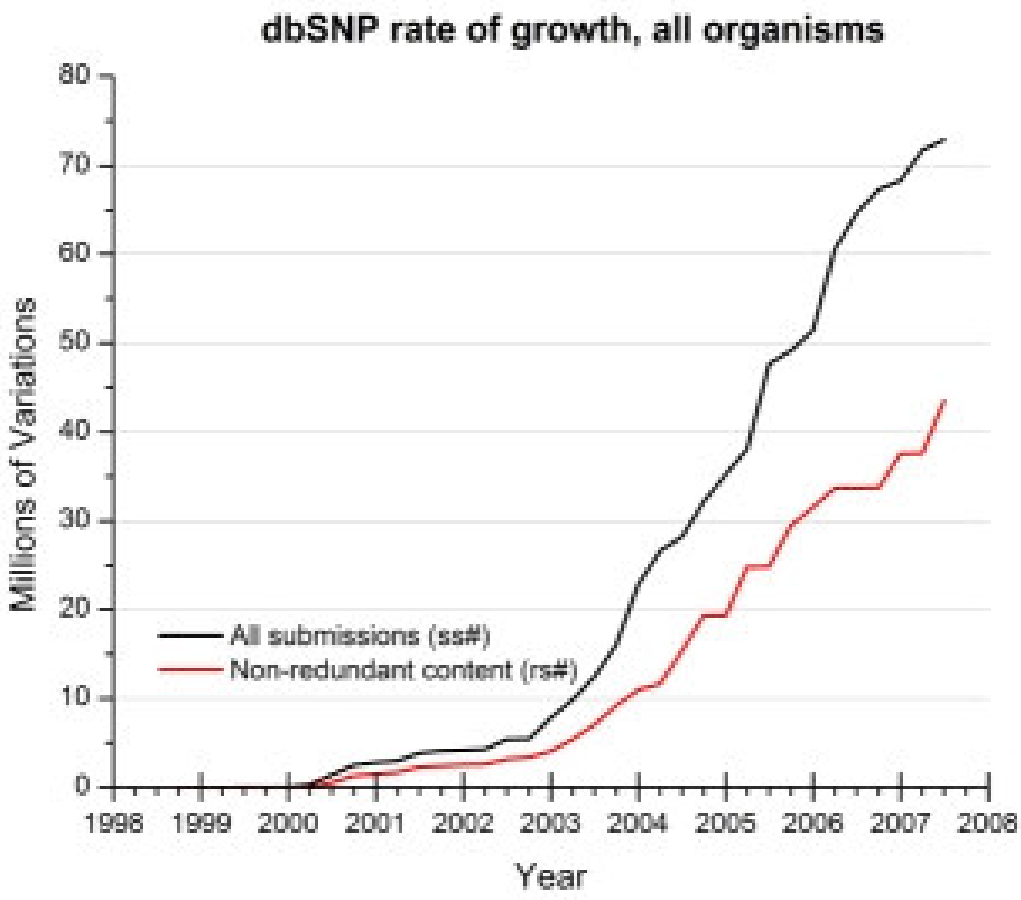
Define new cluster $C_k = C_i \cup C_j$

Define a new node in T to be the parent of i and j at height $d_{ij}/2$

Remove C_i, C_j from set of clusters and insert C_k

End: Insert root for the final two clusters i,j at height $d_{ij}/2$

Discovery rate of single nucleotid polymorphisms (SNPs)



HapMap Project

	Phase 1	**Phase 2**	**Phase 3**
Samples & POP panels	269 samples (4 panels)	270 samples (4 panels)	1,115 samples (11 panels)
Genotyping centers	HapMap International Consortium	Perlegen	Broad & Sanger
Unique QC+ SNPs	1.1 M	3.8 M (phase I+II)	1.6 M (Affy 6.0 & Illumina 1M)
Reference	Nature (2005) 437:p1299	Nature (2007) 449:p851	Draft Rel. 1 (May 2008)

HAPMAP.ORG

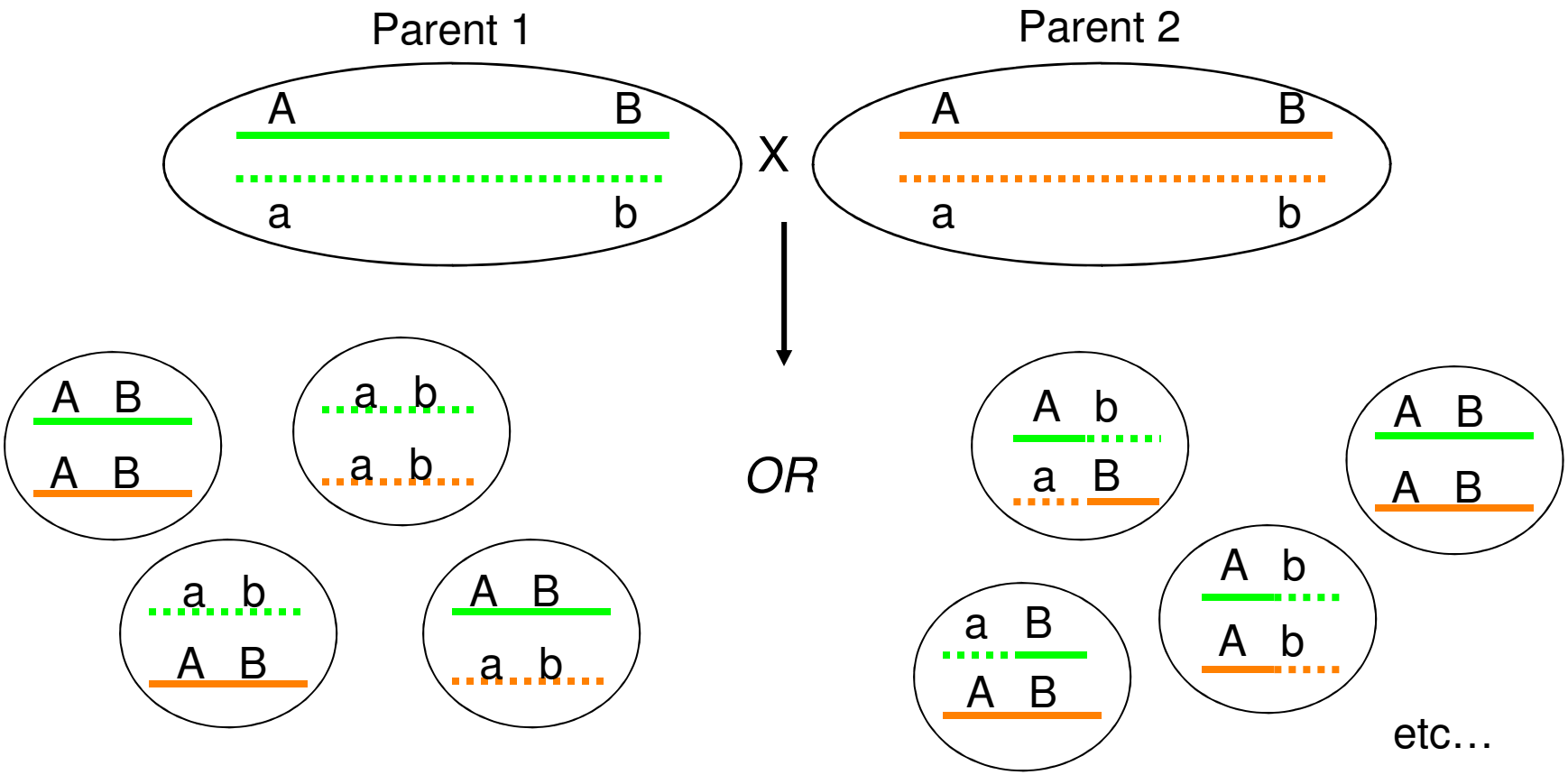
Phase 3 Samples

label	population sample	# samples	QC+ Draft 1
ASW*	African ancestry in Southwest USA	90	71
CEU*	Utah residents with Northern and Western European ancestry from the CEPH collection	180	162
CHB	Han Chinese in Beijing, China	90	82
CHD	Chinese in Metropolitan Denver, Colorado	100	70
GIH	Gujarati Indians in Houston, Texas	100	83
JPT	Japanese in Tokyo, Japan	91	82
LWK	Luhya in Webuye, Kenya	100	83
MEX*	Mexican ancestry in Los Angeles, California	90	71
MKK*	Maasai in Kinyawa, Kenya	180	171
TSI	Toscans in Italy	100	77
YRI*	Yoruba in Ibadan, Nigeria	180	163
		1,301	**1,115**

* Population is made of family trios

HAPMAP.ORG

Basic Concepts



High LD -> No Recombination
(r^2 = 1) SNP1 “tags” SNP2

HAPMAP.ORG

Low LD -> Recombination
Many possibilities

Basic Concepts

	SNP1	SNP2
alleles:	A/a	B/b

C1

C2

POP allele freqs:	A (80%)	B (60%)
	a (20%)	b (40%)

genotypes:

Person 1	Person 2	Person 3
AA	AA	Aa
BB	Bb	Bb

phased haplotypes (C1/C2):

A — B	A — B	A — B
A — B	A — b	a — b
		OR
		A — b
		a — B

HAPMAP.ORG

Haplotypes vs. genotypes

- Problem: loss of information!
  - For a biallelic SNP the (unphased) genotype is
    - 0: wild homozygous
    - 1: heterozygous
    - 2: mutant homozygous

Haplotypes	AA	AT	TT
GG	AG AG	AG AT	TG TG
GC	AG AC	AG TC or AC TG	TG TC
CC	AC AC	AC TC	TC TC

Haplotype	00,00	01,10 or 11,00	11,11
Genotype1	0	1	2
Genotype2	0	1	2

HapMap Glossary

- LD (linkage disequilibrium): For a pair of SNP alleles, it's a measure of deviation from random association. Measured by D', r^2
- Phased haplotypes: Estimated distribution of SNP alleles. Alleles transmitted from Mom are in same chromosome haplotype, while Dad's form the paternal haplotype.
- Tag SNPs: Minimum SNP set to identify a haplotype. r^2= 1 indicates two SNPs are redundant, so each one perfectly "tags" the other.

HAPMAP.ORG

Linkage disequilibrium and recombination

For a given locus A and alleles A_1, A_2:

Hardy-Weinberg equilibrium: independence ($p_{A1A2}=p_{A1}p_{A2}$)

For two loci A, B

they are in linkage diseqilibrium, „LD”, if not independent.

$D=p_{AB}-p_Ap_B$

$D'=D/\min(p_A(1-p_B), p_B(1-p_A))$ if $D>0$,

otherwise $D'=D/\max(-p_A(1-p_A), -p_B(1-p_B))$

$r=D/\mathrm{sqrt}(p_A(1-p_A)p_B(1-p_B))$

$D'=1$: **complete LD**,

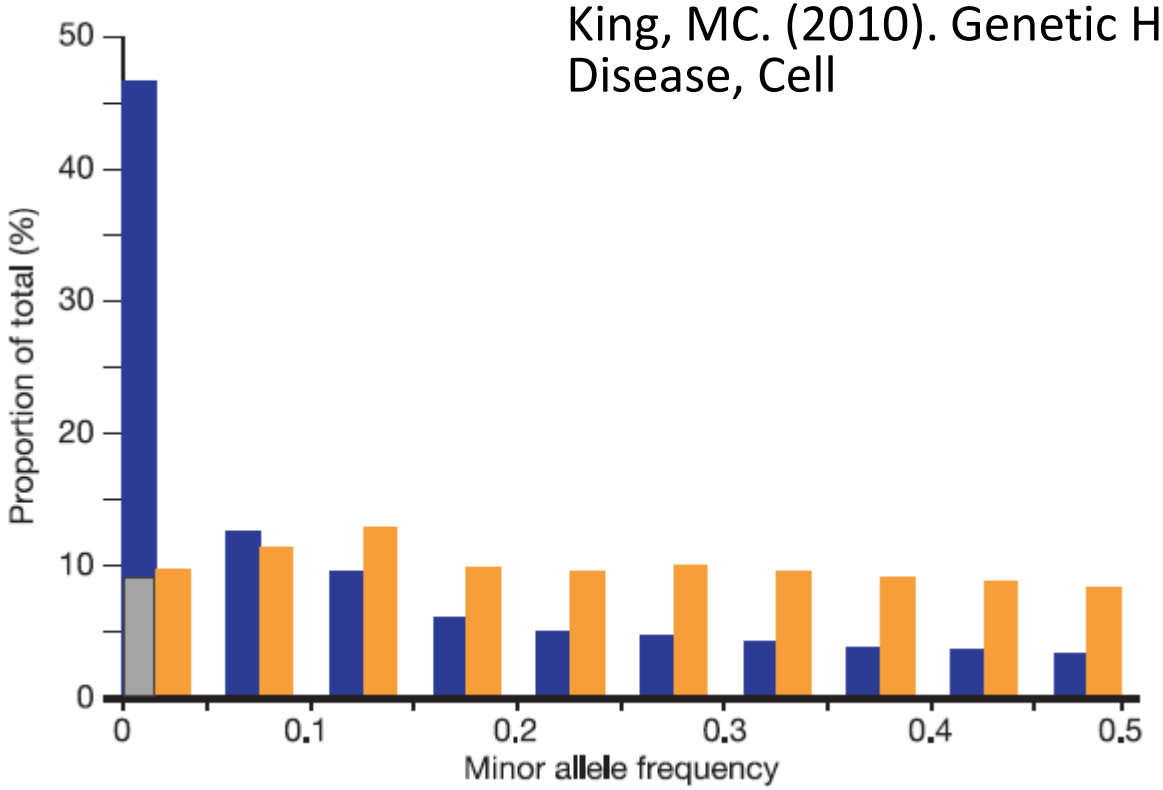
$r=1$: **total LD** (allele frequencies are also the same).

Centimorgan (cM): 1% chance of crossing over in one generation

(~1Mb in human).

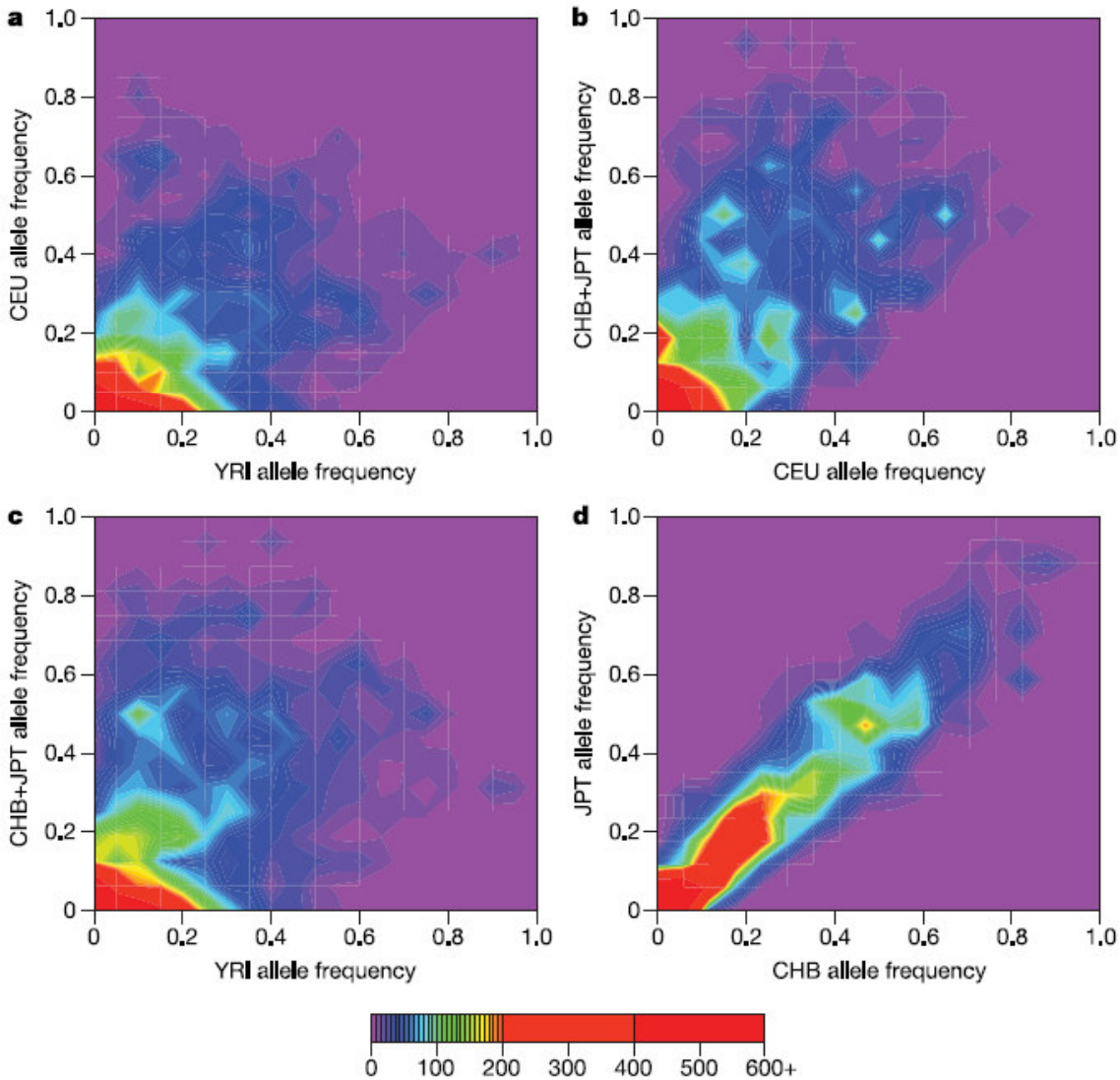
Distribution of minor allele frequences (MAFs)

- „every point mutation compatible with life is likely present in someone, somewhere", McClellan, J., and King, MC. (2010). Genetic Heterogeneity in Human Disease, Cell



Blue: Proportion of MAFs , singletons (gray). Orange: Proportion of SNPs with a given MAF.

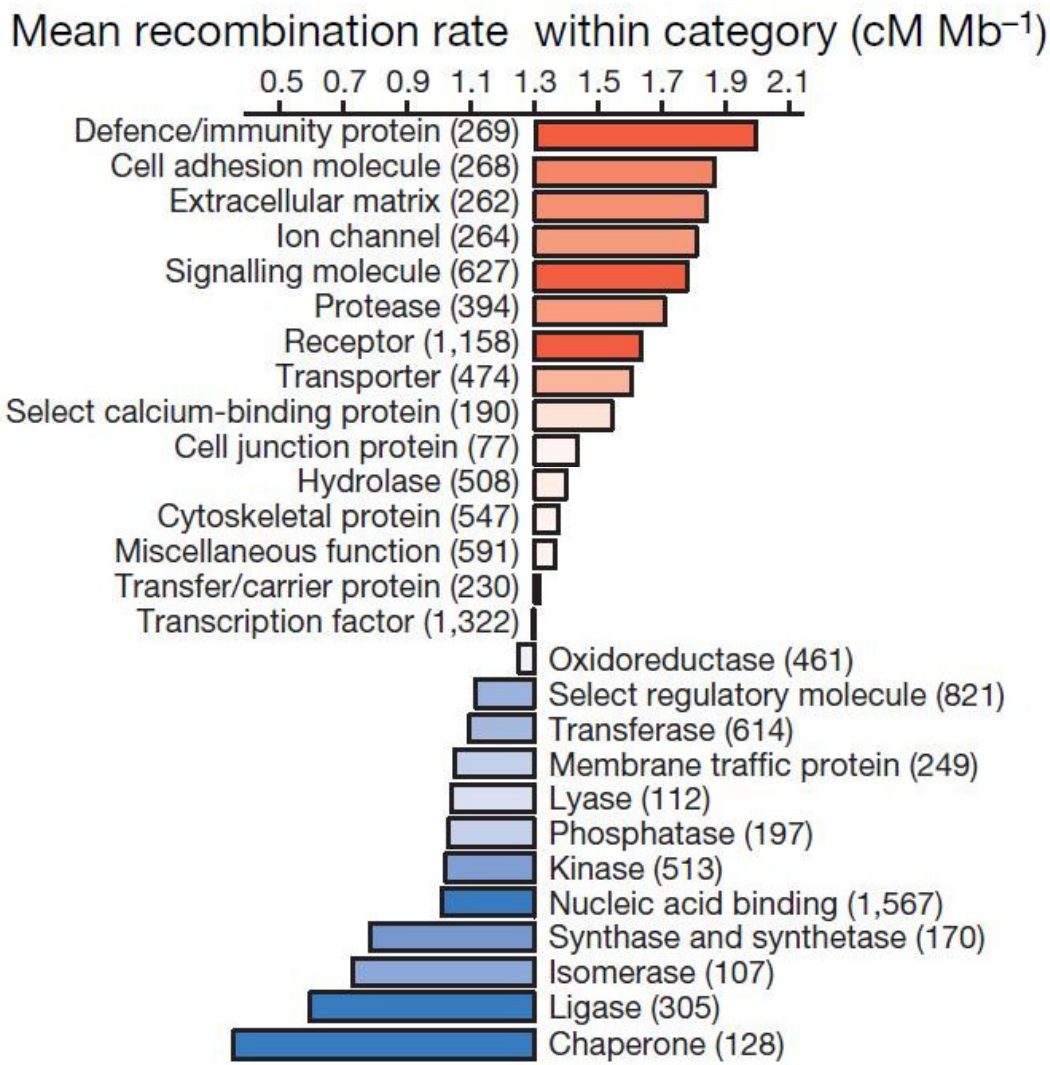
Allele frequencies in subpopulations



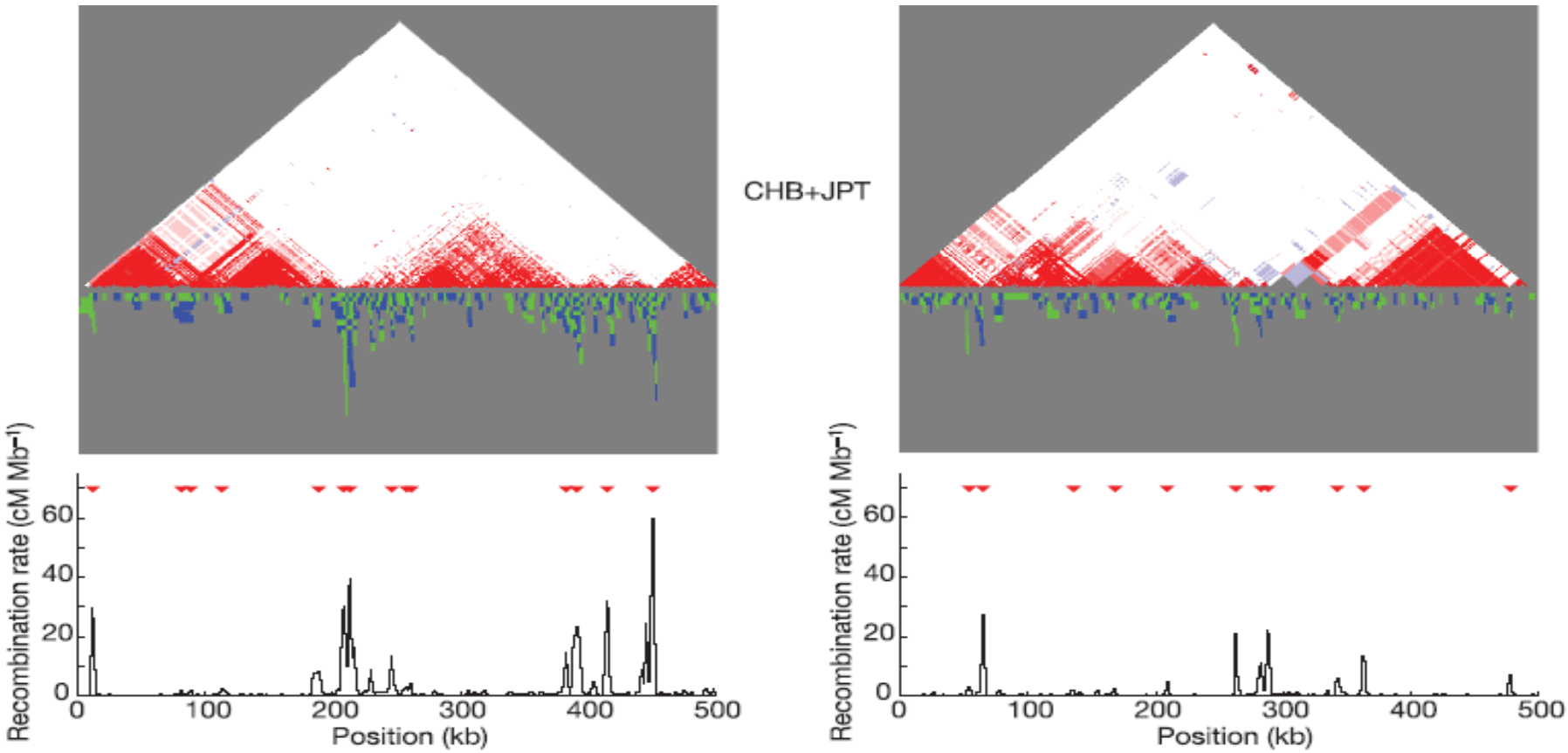
dbSNP

- Google: dbSNP
- http://www.ncbi.nlm.nih.gov/projects/SNP/
- Select MAFs for SNPs in your selected genes.
- E.g. OXTR: **rs11706648**

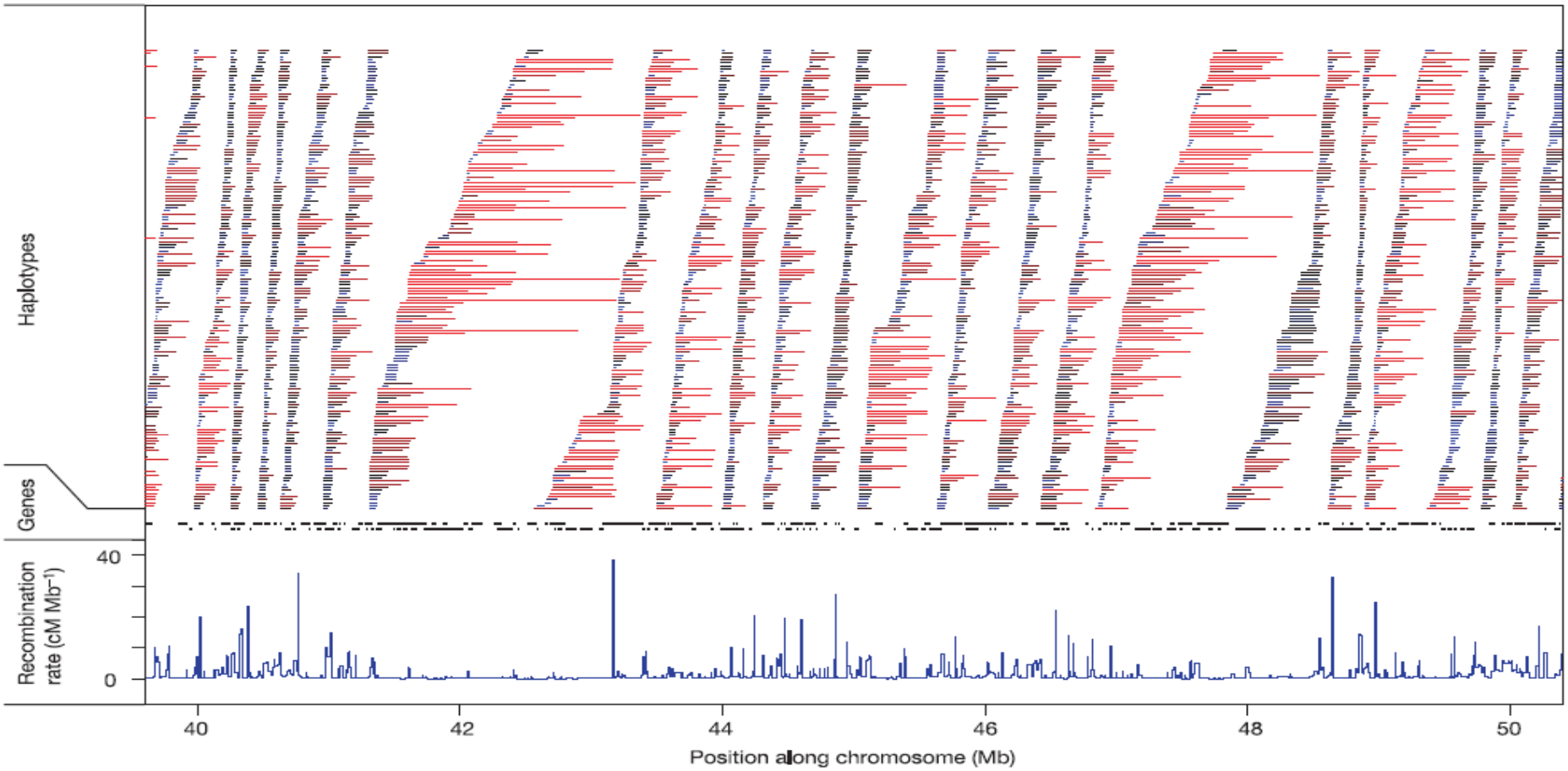
Recombination rate



Linkage disequilibrium



Recombination, haplotypes, and genes



Haplotype: Combination of alleles transmitted together (e.g., SNPs in LD).

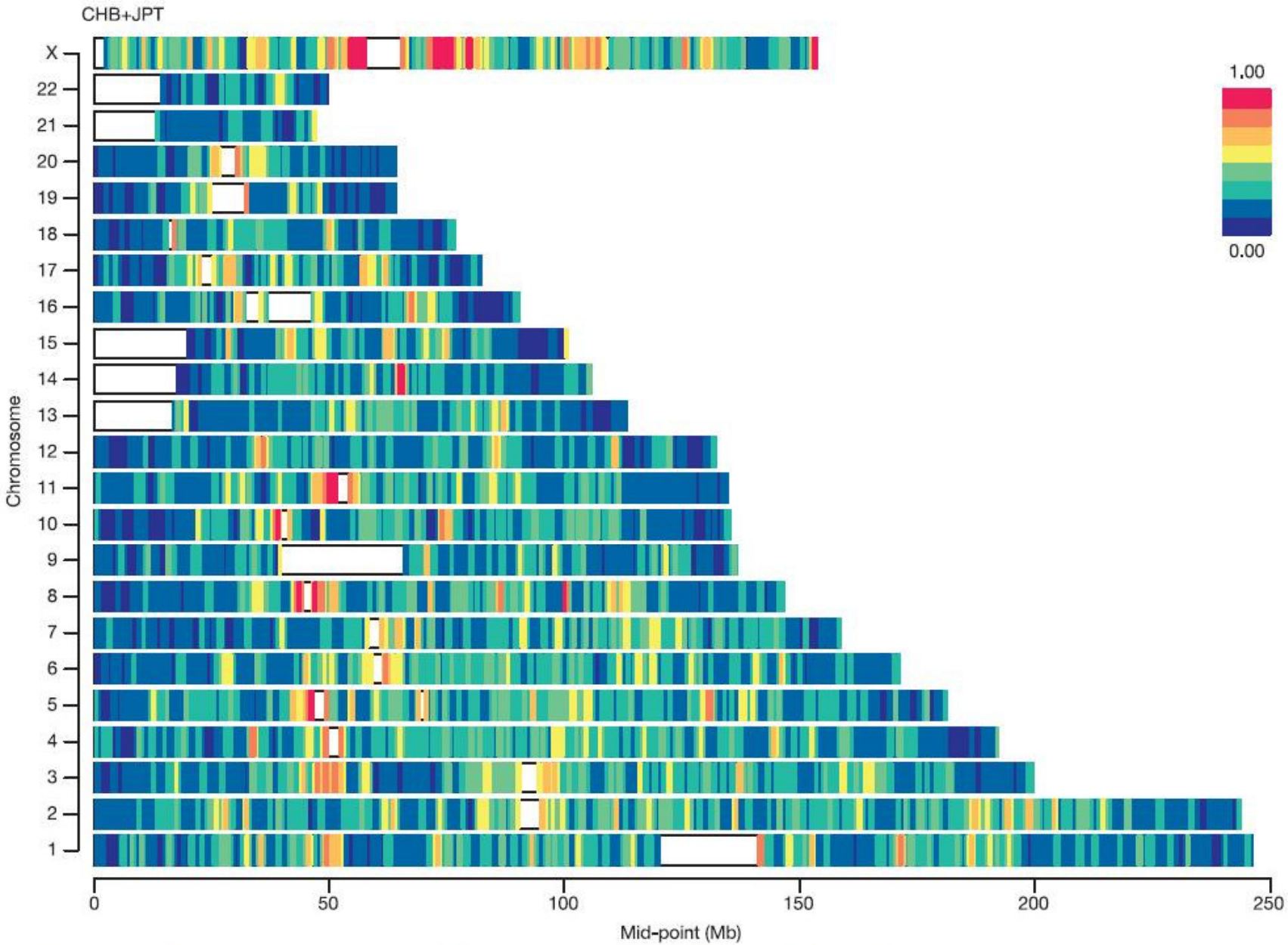
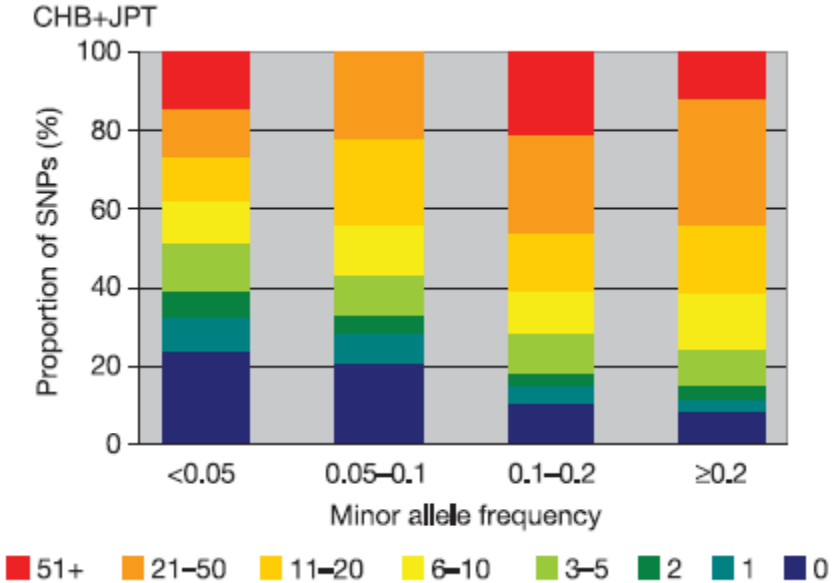


Figure 15 | Length of LD spans. We fitted a simple model for the decay of linkage disequilibrium[103] to windows of 1 million bases distributed throughout the genome. The results of model fitting are summarized for the CHB+JPT analysis panel, by plotting the fitted r^2 value for SNPs separated by 30 kb. The overall pattern of variation was very similar in the other analysis panels[84] (see Supplementary Information).

Proxy and tag SNPs



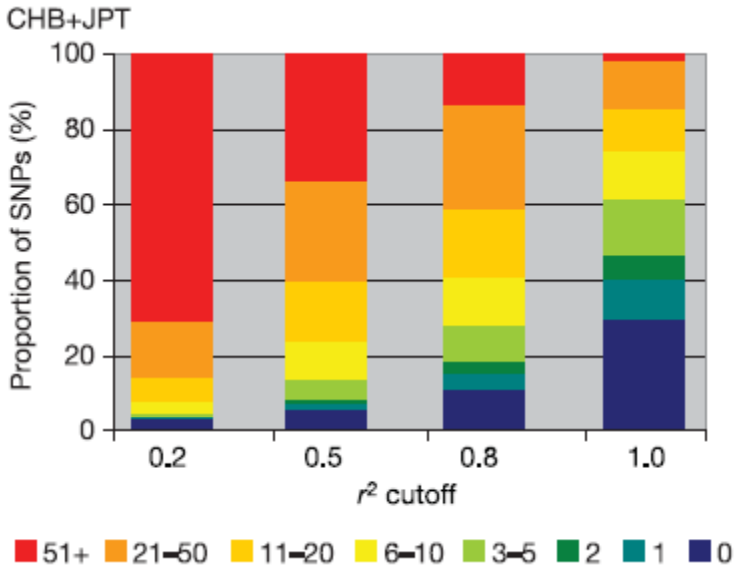
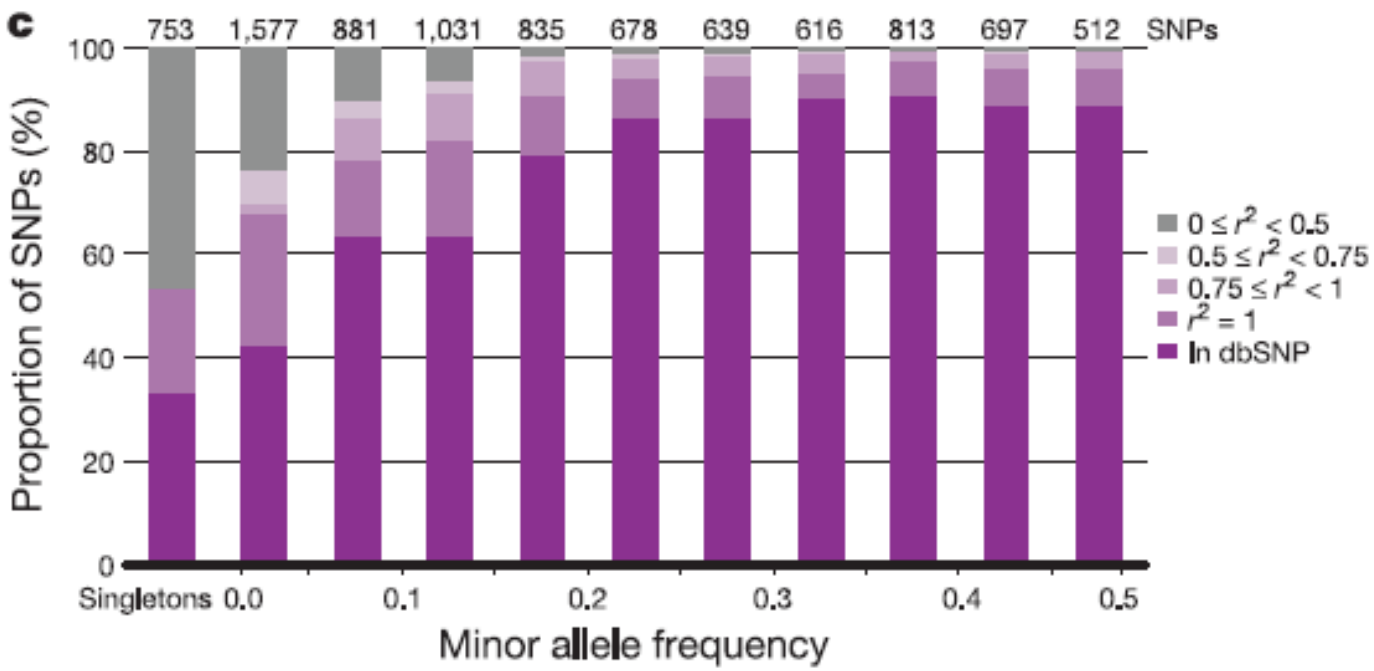


Table 3 | Number of tag SNPs required to capture common (MAF ≥ 0.05) Phase II SNPs

Threshold	YRI	CEU	CHB+JPT
$r^2 \geq 0.5$	627,458	290,969	277,831
$r^2 \geq 0.8$	1,093,422	552,853	520,111
$r^2 = 1.0$	1,616,739	1,024,665	1,078,959

SNPs vs. sequencing (HAPMAP-Phase I.)



Copy number variations (CNVs)

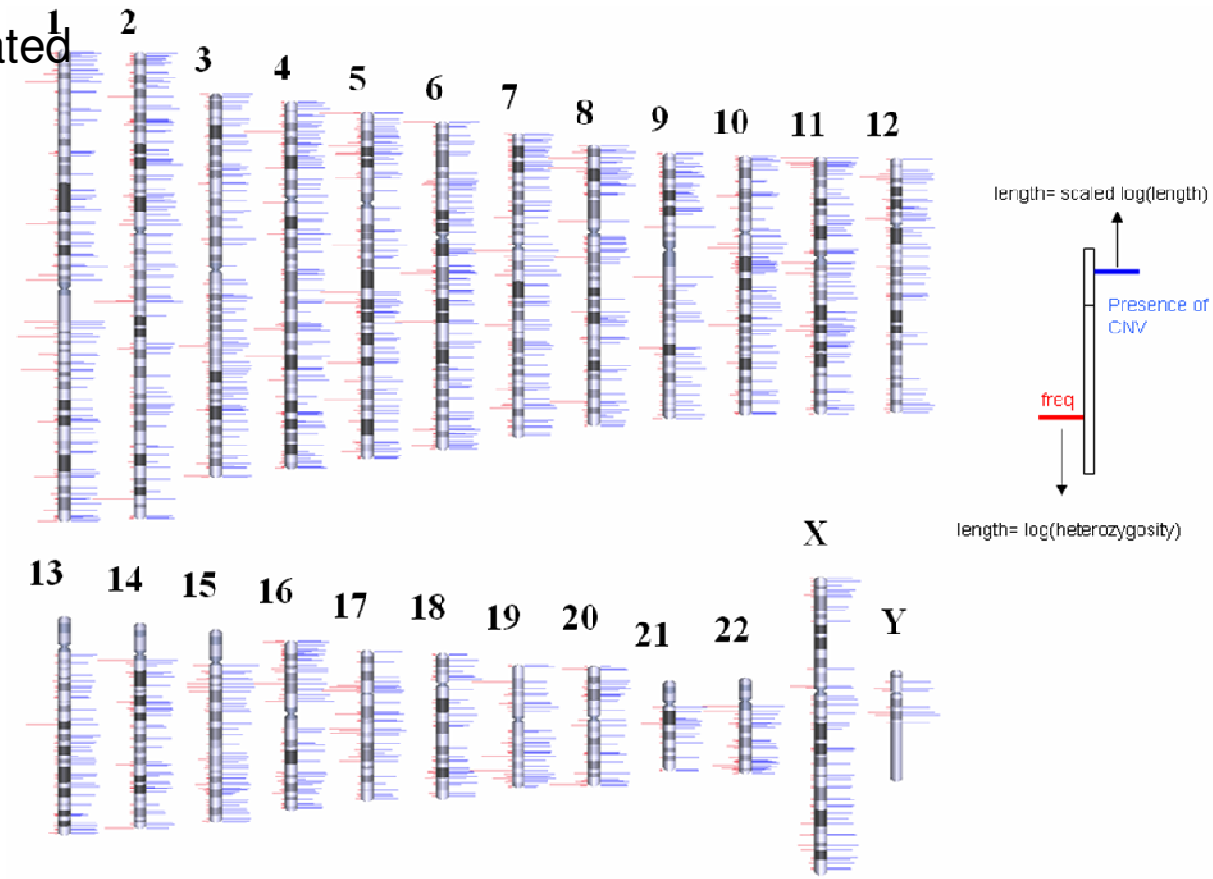
0.4% difference between unrelated individuals

[Database of Genomic Variants]

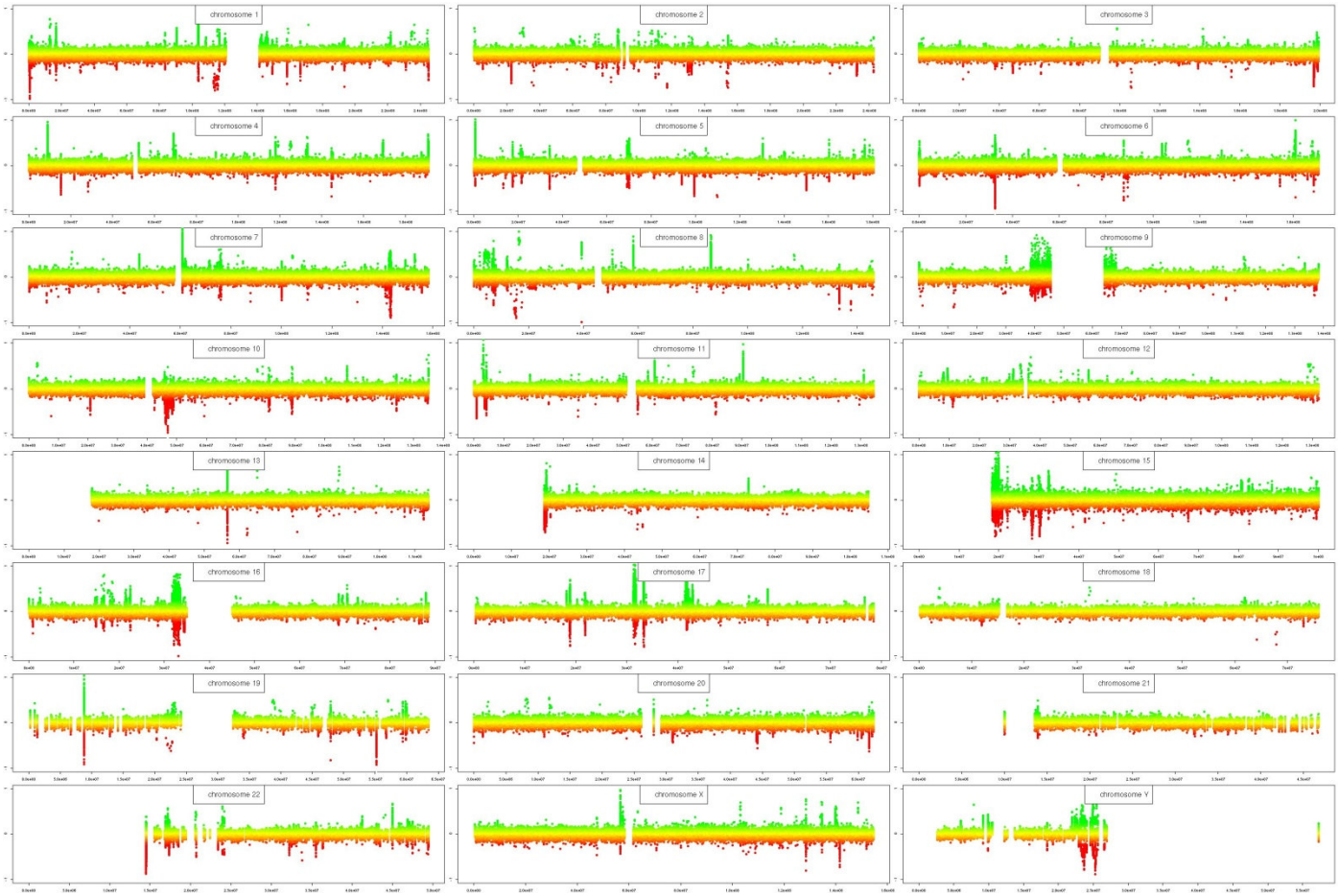
CNVs: 66741

Inversions: 953

InDels (100bp-1Kb): 34229



CNV map II.


chromosome 1
chromosome 2
chromosome 3
chromosome 4
chromosome 5
chromosome 6
chromosome 7
chromosome 8
chromosome 9
chromosome 10
chromosome 11
chromosome 12
chromosome 13
chromosome 14
chromosome 15
chromosome 16
chromosome 17
chromosome 18
chromosome 19
chromosome 20
chromosome 21
chromosome 22
chromosome X
chromosome Y

Rare variants

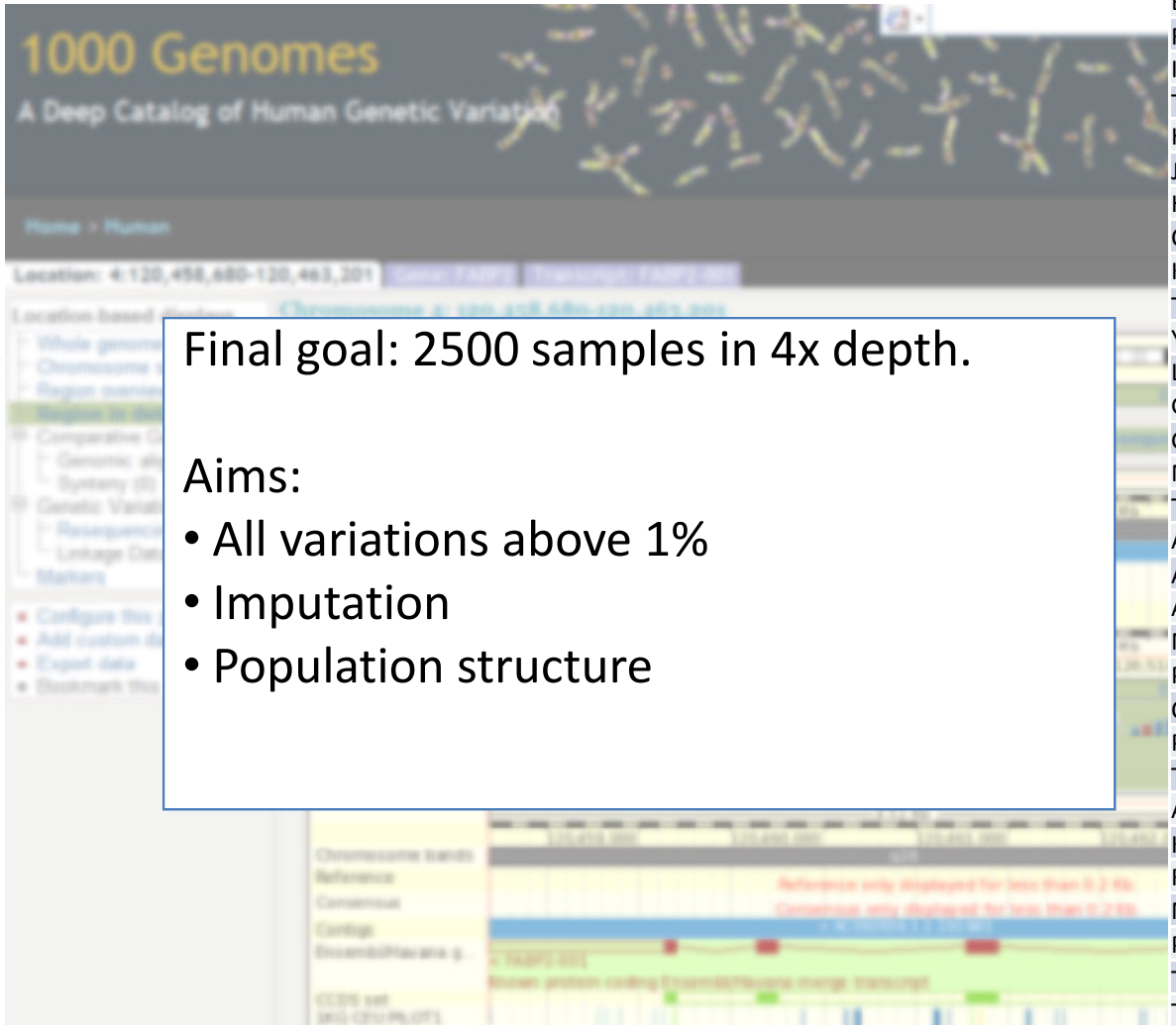
Phenotype	Method	Sample size (cases versus controls)	Genes or genomic regions sequenced	Variants found	Variants associated with phenotype	Comments	Ref.
HTG levels	CAST	438/327	4	187	154	Associated variants across four genes	133
Type 1 diabetes	CS and FET	480/480	10	212	4	Four rare variants in one gene	22
Plasma HDL and TG levels	FET	3,551	4	93	NP	Rare NS cSNPs more frequent in low TG subjects	134
Plasma HDL levels	Observe	154/102	1	NP	3	Five carriers so far are variants with low HDL	135
Folate response	FET	564	1	14	5	Functional evaluation of NS mutations	136
Blood pressure	FET	3,125	3	138	30	Rare mutations affect blood pressure	137
Plasma HDL levels	FET	95/95	1	51	3	Variants in *ABCA1* influence HDL-C	138
Colorectal cancer	FET	691/969	1	61	NP	Rare NS variants in patients	139
Pancreatitis	CS	216/350	1	20	18	Rare variants common in patients	140
Tuberculosis	FET	1,312	5	179	NP	Rare NS variants in tuberculosis cases	141
BMI	CS	379/378	58	1,074	NP	Rare NS variants in obese versus lean	142
HTG levels	CS	110/472	3	NP	10	Single common variant combined with rare variants	143
Heart disease	CS	3,363	1	2	2	Rare variants associated with lower plasma LDL	144
Plasma LDL levels	FET	3,543	4	17	1	*PCSK9* variants associated with low LDL	145
Plasma LDL levels	NP	512	1	26	NP	Variants in *NPC1L1* associated with low cholesterol	146
Plasma LDL levels	NP	128	1	2	2	Two missense mutations associated with low LDL	147
Plasma AGT levels	FET	29/28	1	93	11	Rare haplotypes associated with high AGT levels	45
Plasma HDL levels	FET	519	3	NP	NP	Used collapsing of rare variants	148
Colorectal adenoma	NP	124/483	4	NP	NP	25% rare variants are in cases versus 12% in controls	149
Complex I	Observe	Pooled	103	898	151	More likely deleterious variants in complex I deficiency	150

Statistical analysis strategies for association studies involving rare variants, NRG, 2010

1000Genomes - pilots

Pilot	Purpose	Coverage	Strategy	Status
1 - low coverage	Assess strategy of sharing data across samples	2-4X	Whole-genome sequencing of 180 samples	Sequencing completed October 2008
2 - trios	Assess coverage and platforms and centers	20-60X	Whole-genome sequencing of 2 mother-father-adult child trios	Sequencing completed October 2008
3 - gene regions	Assess methods for gene-region-capture	50X	1000 gene regions in 900 samples	Sequencing completed June 2009

1000Genomes – final goal



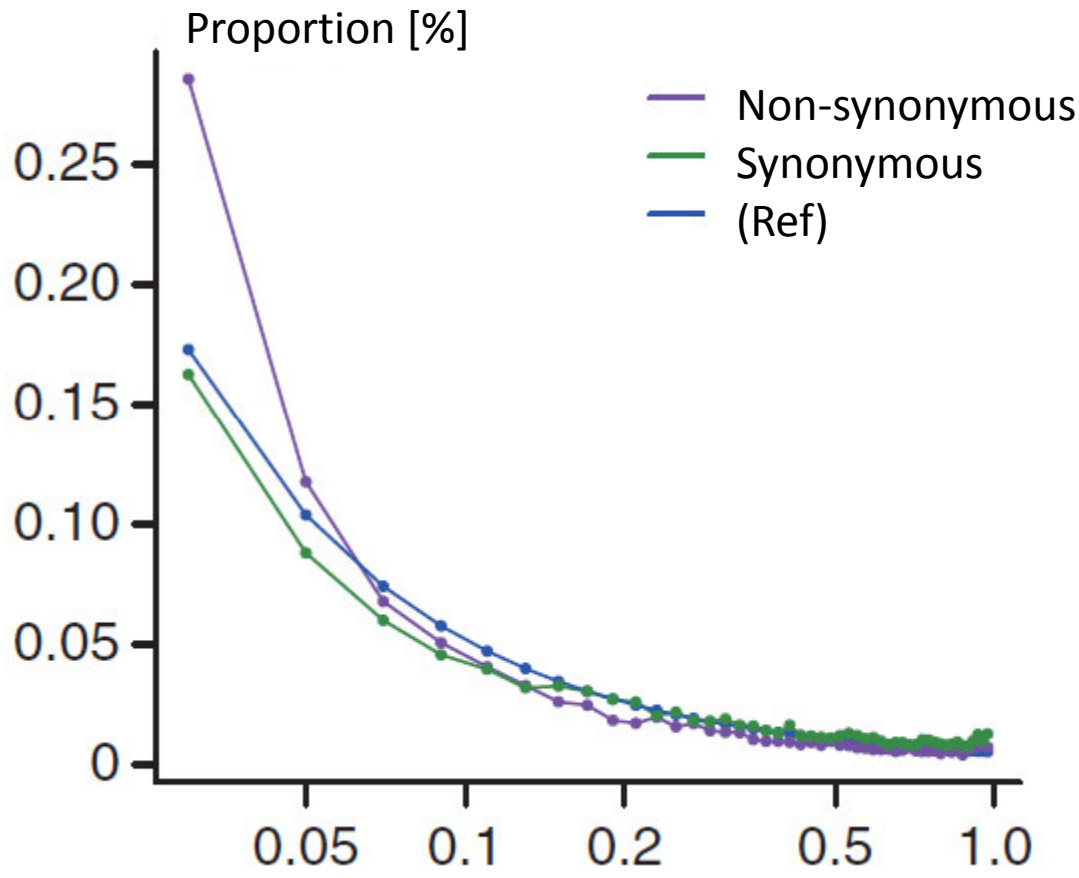
Final goal: 2500 samples in 4x depth.

Aims:

- All variations above 1%
- Imputation
- Population structure

1000 Genomes Samples	
Population	Total
Utah residents (CEPH) with Northern and Western European ancestry (CEU)	100
Toscani in Italia (TSI)	100
British from England and Scotland (GBR)	100
Finnish from Finland (FIN)	100
Iberian populations in Spain (IBS)	100
TOTAL European ancestry	500
Han Chinese in Beijing, China (CHB)	100
Japanese in Tokyo, Japan (JPT)	100
Han Chinese South (CHS)	100
Chinese Dai in Xishuangbanna (CDX)	100
Kinh in Ho Chi Minh City, Vietnam (KHV)	100
TOTAL East Asian ancestry	500
Yoruba in Ibadan, Nigeria (YRI)	100
Luhya in Webuye, Kenya (LWK)	100
Gambian in Western Division, The Gambia (GWD)	100
Ghanaian in Navrongo, Ghana (GHN)	100
Malawian in Blantyre, Malawi (MAB)	100
TOTAL West African ancestry	500
African Ancestry in Southwest US (ASW)	61
African American in Jackson, MS (AJM)	80
African Caribbean in Barbados (ACB)	79
Mexican Ancestry in Los Angeles, CA (MXL)	70
Puerto Rican in Puerto Rico (PUR)	70
Colombian in Medellin, Colombia (CLM)	70
Peruvian in Lima, Peru (PEL)	70
TOTAL Americas	500
Ahom in the State of Assam, India	100
Kayadtha in Calcutta, India	100
Reddy in Hyderabad, India	100
Maratha in Bombay, India	100
Punjabi in Lahore, Pakistan	100
TOTAL South Asian ancestry	500
TOTAL	2500

Functional constraints?



Li et al.:Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants, 2010, Oct., Nature Genetics

Summary

- Basics: LD....

Tumorgenetics

- Virus/Viral origin: HPV, hepatitis-B,..
- Carcinogenes: smoking
- Number of „trials”: life span, breast/ovarian cancer
- Heredity vs somatic mutation
- Dominant vs recessive model?
- Evolutionary modeling / population genetics?
- Multicellular level?

The multistage cancer model

Fisher: ~1930

Multiple steps with different

p $$E_0 \xrightarrow{p_1} E_1 \xrightarrow{p_2} E_2 \cdots \xrightarrow{p_n} E_n.$$

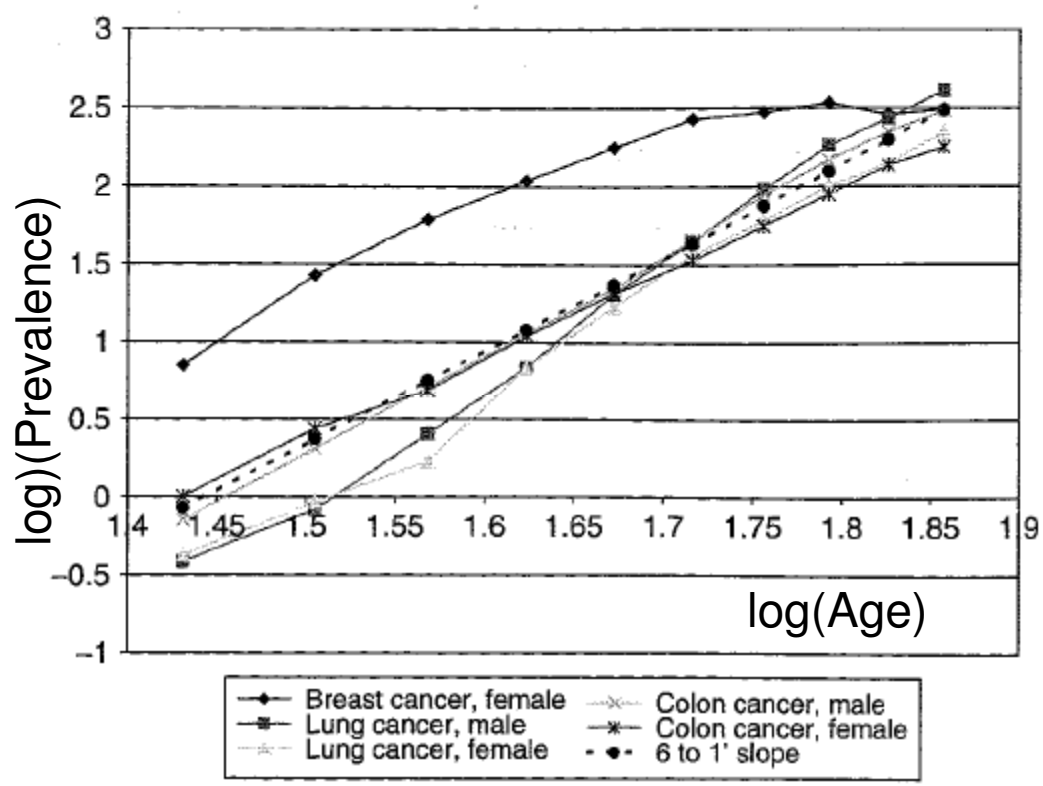
$$p_1 p_2 p_3 p_4 p_5 p_6 p_7 t^7.$$

prevalence:

$$\frac{p_1 p_2 \cdots p_n t^{n-1}}{(n-1)!}.$$

incidence:

Two-stage model: Loss-of-heterozygosity, LOH research



Barriers

- Proto-oncogenes, oncogenes
- Tumor suppressors
- DNA repair pathways
- Growth-related/signal transduction
- Telomerase related
- Cell death related
- Cell cycle related
- Immune system related
- Proliferation
  - Leaving the host organ + metastates

DNA damage and repair

- Endogenous (replication/mitosis)
- Exogenous: radiation/UV-A/B, mutagens, viruses
- Repairs:
  - Direct
  - Single-strand damage
    - Base excision repair (BER)
    - Nucleotide excision repair (NER)
    - Mismatch repair (MMR)
  - Double-strand breaks
    - Non-homologous end joining (NHEJ),
    - Microhomology-mediated end joining (MMEJ),
    - Homologous recombination (HR)
- Global responses:
  - Senescence
  - Apoptosis: ATM/ATR... BRCA1...P53

Driver vs passenger mutation

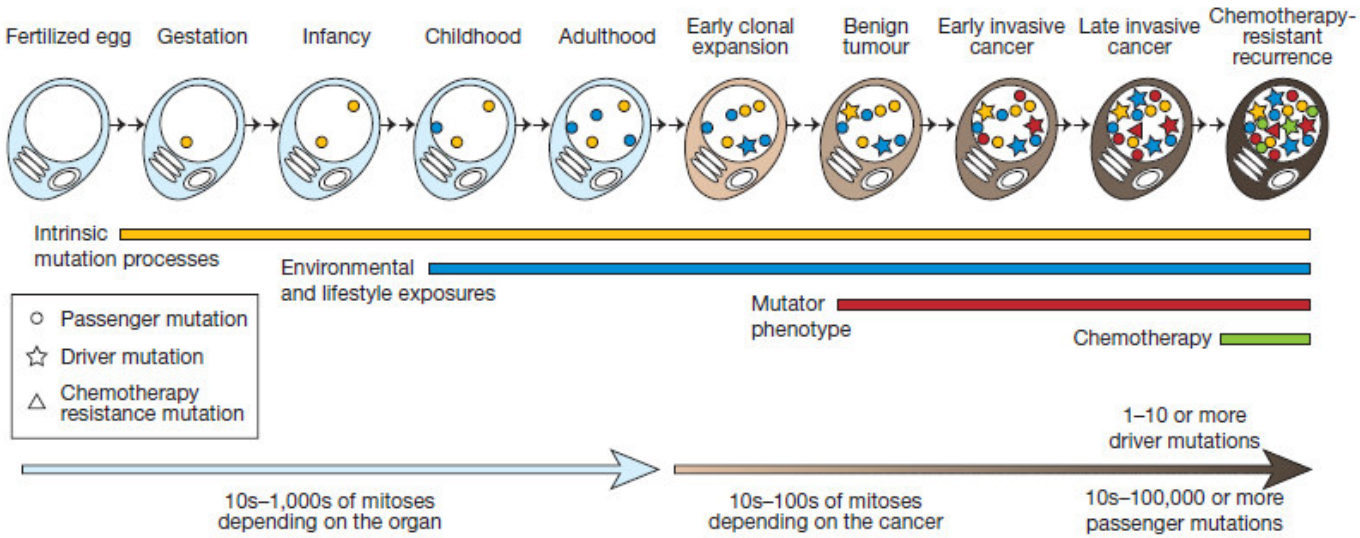
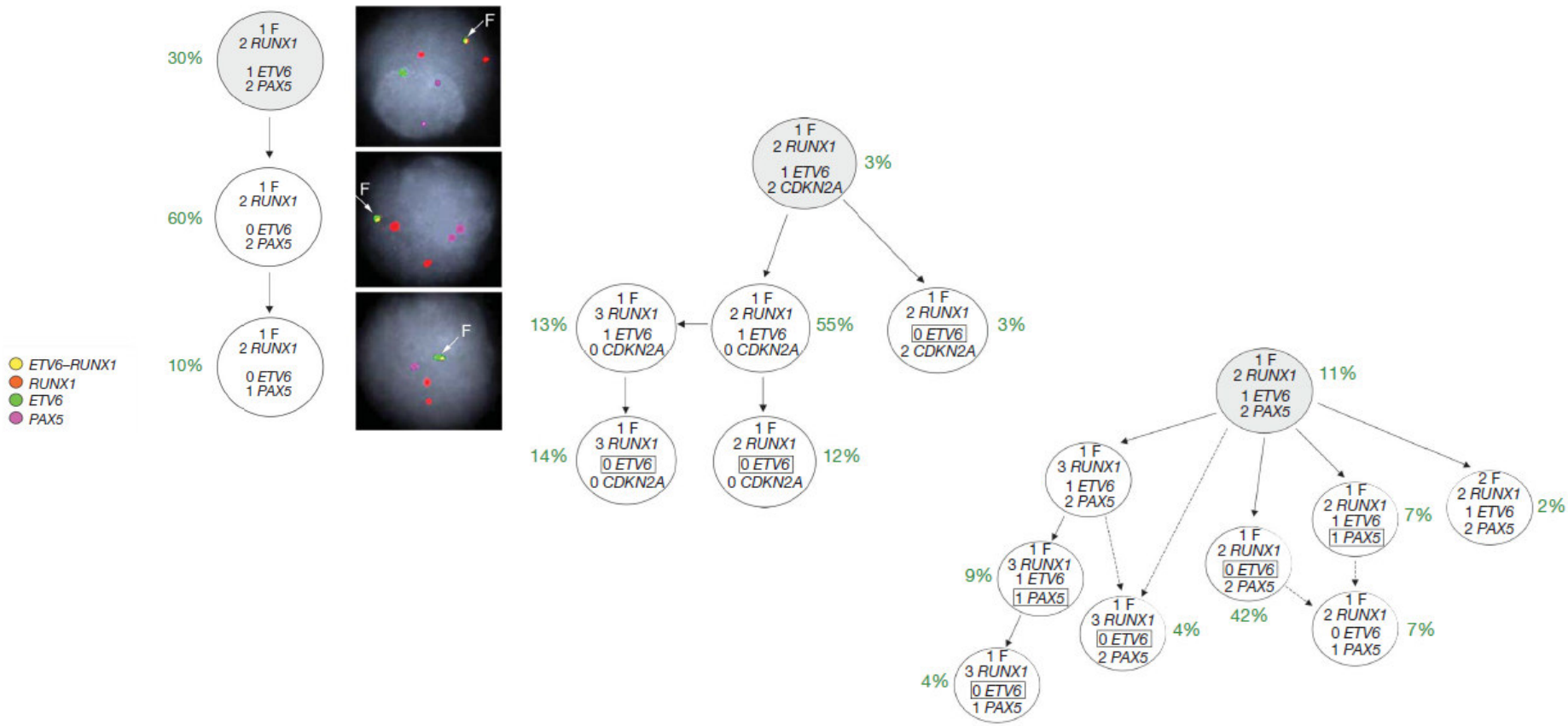


Figure 1 | The lineage of mitotic cell divisions from the fertilized egg to a single cell within a cancer showing the timing of the somatic mutations acquired by the cancer cell and the processes that contribute to them. Mutations may be acquired while the cell lineage is phenotypically normal, reflecting both the intrinsic mutations acquired during normal cell division and the effects of exogenous mutagens. During the development of the cancer other processes, for example DNA repair defects, may contribute to the mutational burden. Passenger mutations do not have any effect on the cancer cell, but driver mutations will cause a clonal expansion. Relapse after chemotherapy can be associated with resistance mutations that often predate the initiation of treatment.

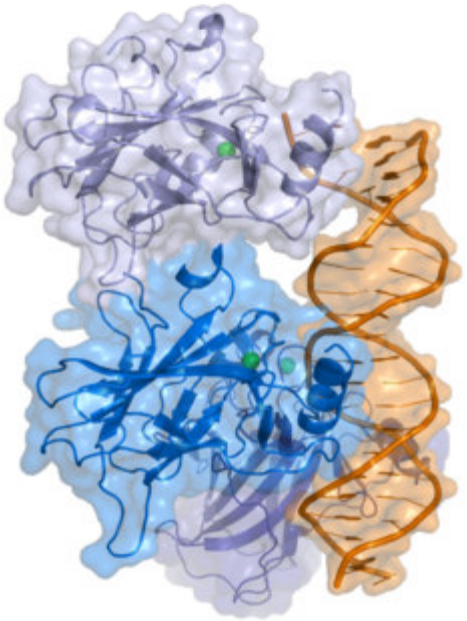
The clonal architecture in leukaemia

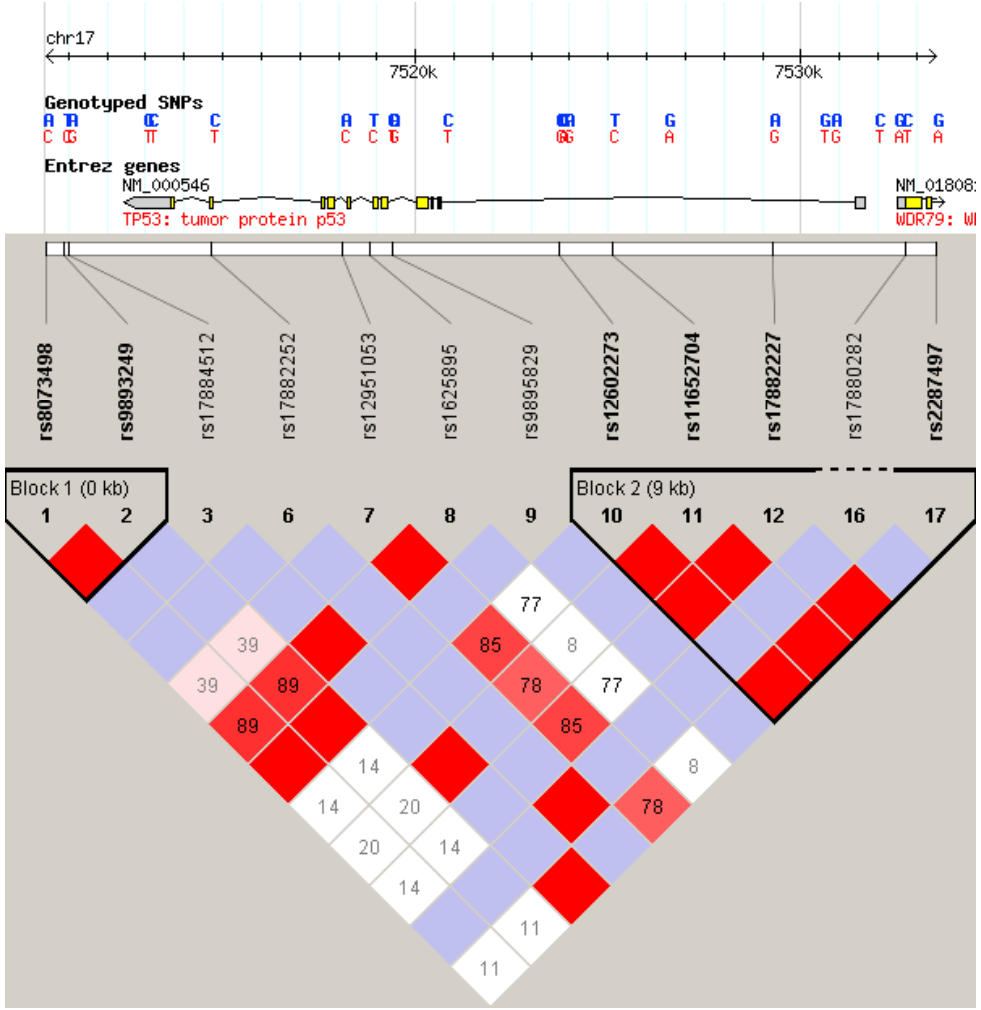


Genetic variegation of clonal architecture and propagating cells in leukaemia, 2011

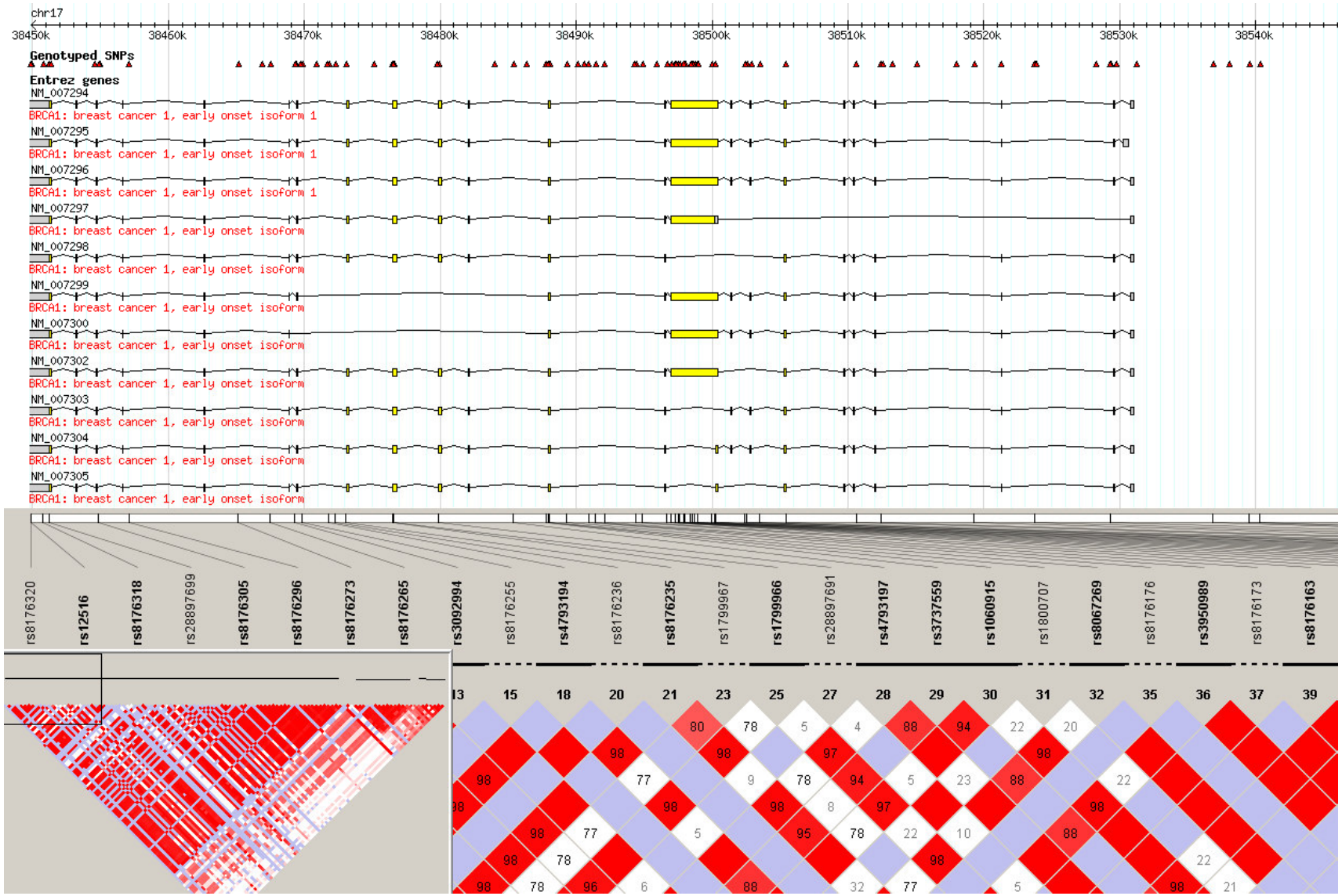
P53 – inherited SNPs and LD

- 18 phosphorylations...





BRCA1-inherited SNPs and LD



International projects

- The Cancer Genome Atlas (TCGA)
- The Cancer Genome Project (CGP)
- The International Cancer Genome Consortium (ICGC)
- Catalogue Of Somatic Mutations In Cancer (COSMIC)
- Cancer Genome Anatomy Project (CGAP)
- Cancer Genome Characterization Initiative (CGCI)

Cancer gene census

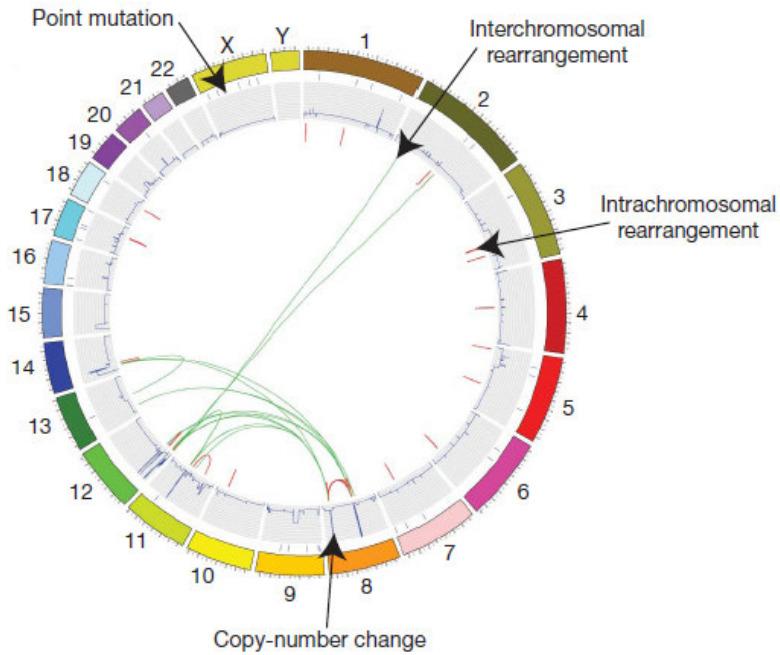


Figure 2 | Figurative depiction of the landscape of somatic mutations present in a single cancer genome. Part of catalogue of somatic mutations in the small-cell lung cancer cell line NCI-H2171. Individual chromosomes are depicted on the outer circle followed by concentric tracks for point mutation, copy number and rearrangement data relative to mapping position in the genome. Arrows indicate examples of the various types of somatic mutation present in this cancer genome.

Cancer Gene Census

Complete working list.xls

Cancer Gene Census

Sorted By	Number
Amplification	15
Chromosome	457
Frameshift mutation	88
Germline mutation	75
Large deletion	34
Missense mutation	126
Nonsense mutation	85
Other mutation	20
Somatic mutation	415
Splicing mutation	53
Symbol	457
Translocation	315

An optimistic(?)/sceptic view

- Bert Vogelstein
  - Overviewing solid cancers (for 353 cancer subtypes): potential 130,072 mutations in 3142 genes.
  - Conclusion: largely known
    - 12 pathways
    - 320 driver genes
    - 10-100 mutations in each type of tumor

AML sequencing: comparison

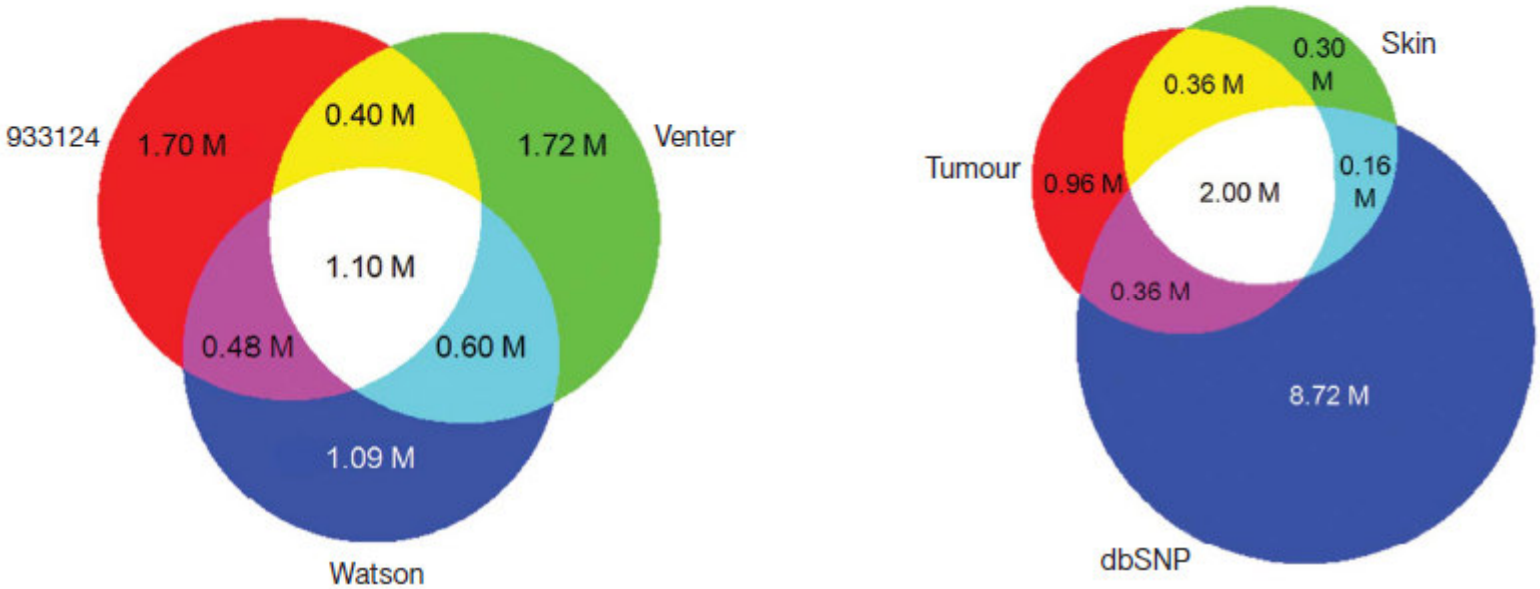


Figure 1 | Overlap of SNPs detected in 933124 and other genomes. a, Venn diagram of the overlap between SNPs detected in the 933124 tumour genome and the genomes of J. D. Watson and J. C. Venter. **b,** Venn Diagram of the overlap among the 933124 tumour genome, the skin genome and dbSNP (ver. 127). SNVs were defined with a Maq SNP quality ≥15.

DNA sequencing of a cytogenetically normal AML, 2008

AML sequencing: filtering

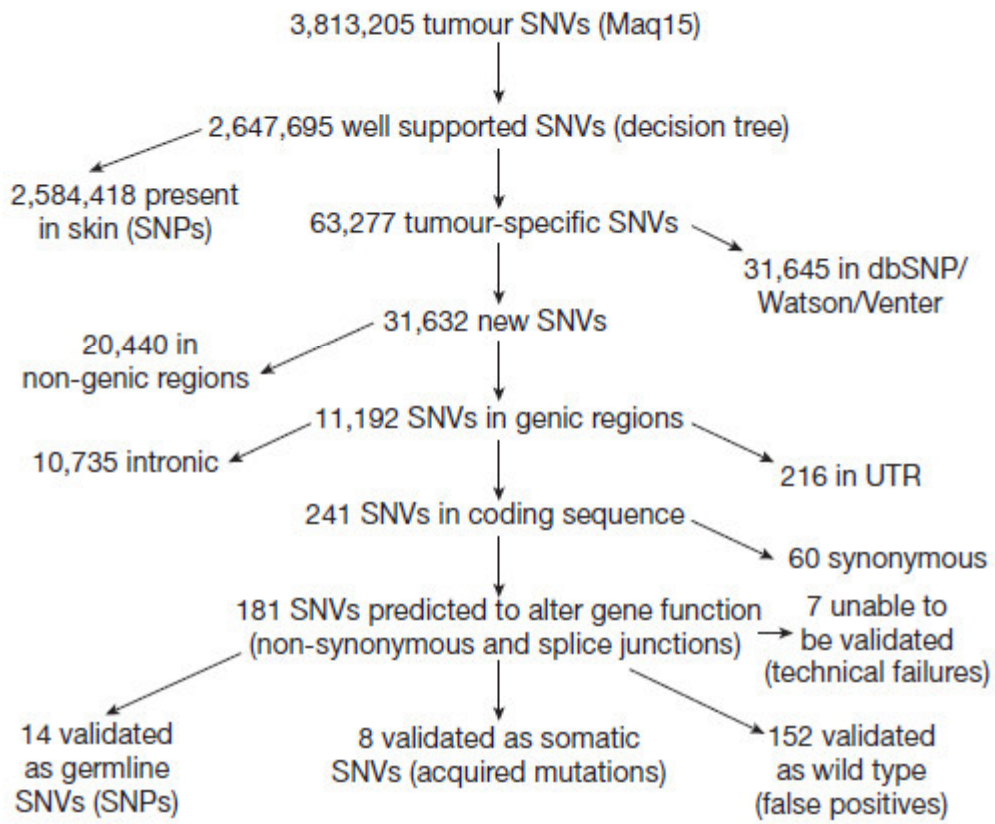
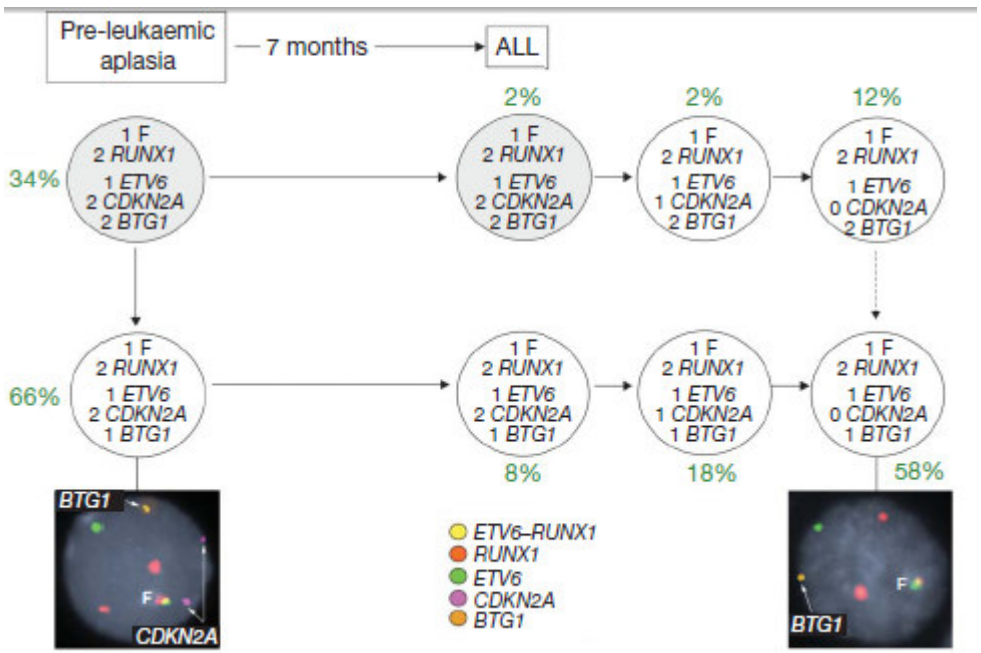


Figure 2 | Filters used to identify somatic point mutations in the tumour genome. See text for details. UTR, untranslated regions.

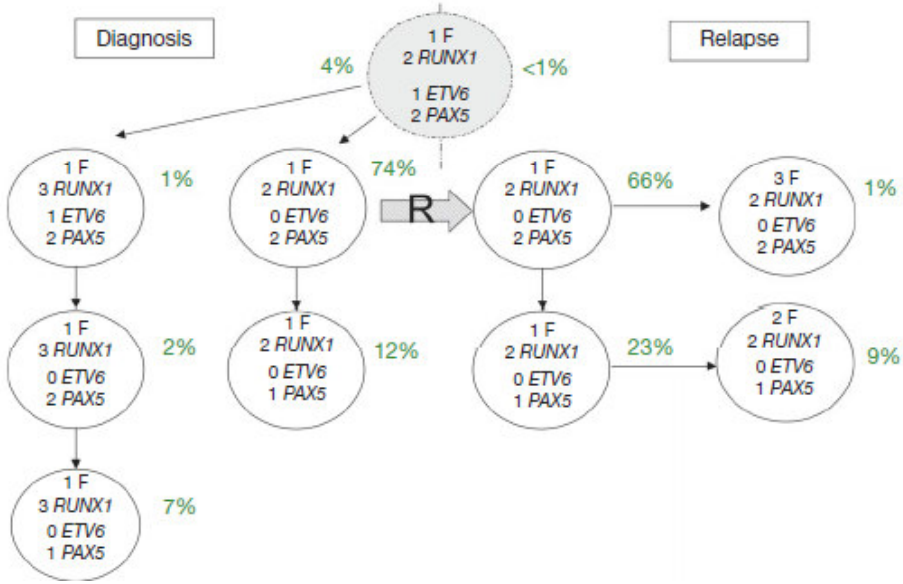
DNA sequencing of a cytogenetically normal AML, 2008

Dynamics of clonal architecture

Effect of time



Effect of treatment



Genetic variegation of clonal architecture and propagating cells in leukaemia, 2011

Breast cancer (BC), Ovarian cancer (OC)

- The first/second most commonly diagnosed gynecologic malignancy.
- The leading cause of death from gynecological malignancy.
- The fifth leading cause of cancer deaths in women.
- Poor prognosis if diagnosed at an advanced stage.
- Early diagnosis is important (screening, prevention).
- Multifactorial disease.

Meta-analysis of BC related variants

- 24500 publications
- 1059 included
- 128 genes
- 279 genetic variants
- 51/40 strongly sign.
- 37 weak sign.
- 47 none

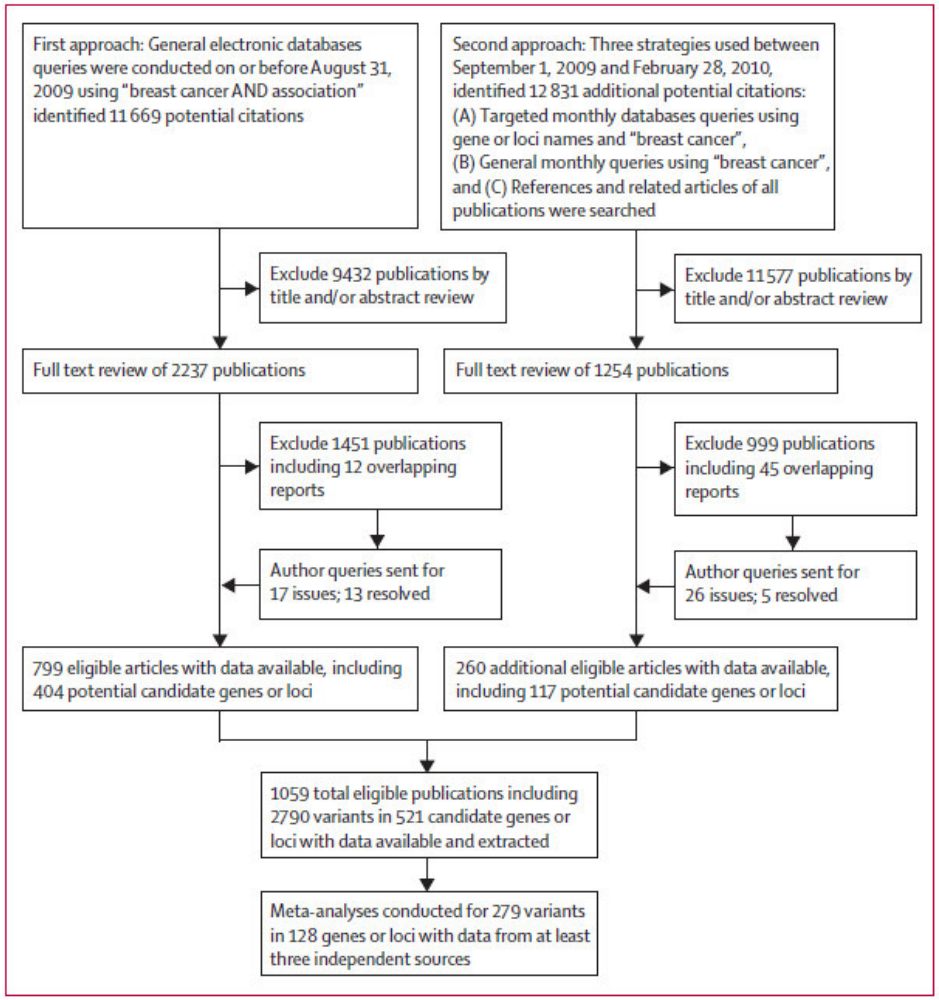


Figure: Two-stage search strategy to identify and select breast-cancer candidate-gene association studies

OC/BC variants

	Variants	Relative risk	Population frequency (%)
BRCA1	Multiple mutations	>10	0·1
BRCA2	Multiple mutations	>10	0·1
TP53	Multiple mutations	>10	<0·1
PTEN	Multiple mutations	>10	<0·1
ATM	Truncating and missense mutations	2–4	<0·5
CHEK2	1100delC	2–5	0·7
BRIP1	Truncating mutations	2–3	0·1
PALB2	Truncating mutations	2–5	<0·1

Table 1: **High-penetrance and moderate-penetrance breast-cancer susceptibility genes**

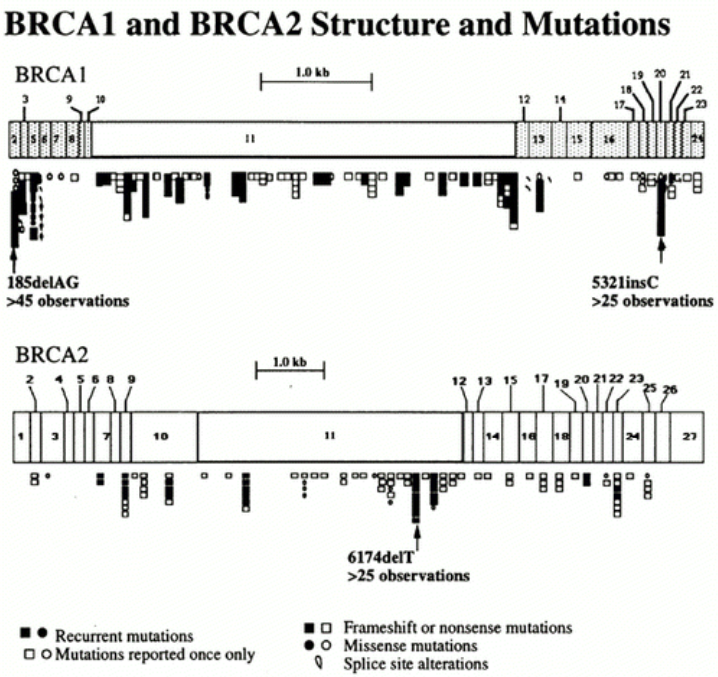
	Variant	OR (95% CI)*	MAF (%)
1p11 *NOTCH2, FCGR1B*	rs11249433	1·14 (1·10–1·19)	39
2q35	rs13387042	1·20 (1·14–1·26)	50
3p24 *SLC4A7, NEK10*	rs4973768	1·11 (1·08–1·13)	46
5p12 *MRPS30*	rs4415084 rs10941679	1·16 (1·10–1·21) 1·19 (1·13–1·26)	40 24
5q11 *MAP3K1*	rs889312	1·13 (1·10–1·16)	28
6q22 *ECHDC1, RNF146*	rs2180341	1·41 (1·25–1·59)	21
6q25 *ESR1, C6orf97*	rs2046210	1·29 (1·21–1·37)	35
8q24	rs13281615	1·08 (1·05–1·11)	40
9p21 *CDKN2A, CDKN2B*	rs1011970	1·09 (1·04–1·14)	17
10p15 *ANKRD16, FBXO18*	rs2380205	0·94 (0·91–0·98)	43
10q21 *ZNF365*	rs10995190	0·86 (0·82–0·91)	15
10q22 *ZMIZ1*	rs704010	1·07 (1·03–1·11)	39
10q26 *FGFR2*	rs2981582 rs1219648	1·26 (1·23–1·30) 1·27 (1·18–1·36)	38 40
11p15 *LSP1*	rs3817198	1·07 (1·04–1·11)	30
11q13	rs614367	1·15 (1·10–1·20)	15
14q24 *RAD51L1*	rs999737	0·89 (0·85–0·93)	24
16q12 *TOX3, LOC643714*	rs3803662 rs4784227	1·28 (1·21–1·35) 1·25 (1·20–1·31)	27 24
17q22 *COX11*	rs6504950	0·95 (0·92–0·97)	27
19p13 *ABHD8, ANKLE1, C19orf62*	rs2363956	0·80 (0·74–0·87)	47

OR=odds ratio. MAF=minor-allele frequency. *From original report of association with breast-cancer risk.

Table 2: **Low-penetrance loci associated with breast-cancer risk, identified by genome-wide association studies**

BRCA genes

- Mutations of BRCA1 and BRCA2 responsible for the majority of hereditary breast cancer cases (5-10%)
- Lifetime risks of breast cancer are as high as 80% among women with mutations in these genes
- BRCA1 is also associated with ovarian cancer (LR: 40 and 20%).
- The BRCA genomic regions are each about 80-90 kb (including introns that span 81,155 bp and 84,193 bp, respectively)
- OMIM database: 68 known mutations
- Human Gene Mutation Database: lists 1,215 and 966 mutations for BRCA1 and BRCA2
- Breast Cancer Information Core (BIC): 3500< variant



Ellisen, L. W. and D. A. Haber (1998). "Hereditary breast cancer." Annu Rev Med **49**: 425-436.
http://www.matud.iif.hu/05aug/08.html

- outcomes of BRCA genetic testing:
  - confirmation of hereditary nature of their disease
  - unaffected relatives to undertake presymptomatic testing and, if positive, to receive early screening and appropriate surgery/treatment

BRCA1, BRCA2

Table 2. Point mutations and small insertions and deletions identified by the assay

					Mutant sites identified			No. of reads		
Gene	Nucleotide	Effect	Type	Size (bp)	Chromosome	Start	End	Wild type	Variant	% Variant
BRCA1	4510 del3ins2	1465 stop	Deletion-insertion	1	17	41,228,596	41,228,597	525	596	0.53
BRCA1	5083 del19	1657 stop	Deletion	19	17	41,222,949	41,222,968	700	644	0.48
BRCA1	5382 insC	1829 stop	Insertion	1	17	41,209,080	41,209,081	606	596	0.50
BRCA2	999 del5	273 stop	Deletion	5	13	32,905,141	32,905,146	363	229	0.39
BRCA2	1983 del5	585 stop	Deletion	5	13	32,907,366	32,907,371	304	258	0.46
BRCA2	6174 delT	2003 stop	Deletion	1	13	32,914,438	32,914,439	565	661	0.54
BRCA2	9179 C > G	2984 stop	Nonsense	1	13	32,953,650		391	361	0.48
BRIP1	3401 delC	1149 stop	Deletion	1	17	59,761,006	59,761,007	651	486	0.43
CDH1	591 G > A	157 stop	Nonsense	1	16	68,842,406		421	359	0.46
CHEK2	1100 delC	381 stop	Deletion	1	22	29,091,857	29,091,858	3,293	586	0.15
MLH1	ivs14(-1) G > A	568 stop	Splice	1	3	37,083,758		1,024	683	0.40
MSH2	1677 T > A	537 stop	Nonsense	1	2	47,693,895		575	552	0.49
p53	721 G > A	R175H	Missense	1	17	7,578,406		449	306	0.41
PALB2	509 delGA	183 stop	Deletion	2	16	23,647,357	23,647,359	1,283	1,233	0.49
STK11	ivs6(-1) G > A	316 stop	Splice	1	19	1,221,947		722	572	0.44

Table 3. Genomic deletions and duplication identified by the assay

		Mutant sites identified by assay				
Gene	Genomic event	Chromosome	Start*	End*	Size (bp)	Ratio†
BRCA1	Deletion exons 1–15	17	41,226,145	41,327,157	101,013	0.509
BRCA1	Duplication exon 13	17	41,230,562	41,235,836	5,275	1.578
BRCA1	Deletion exons 14–20	17	41,203,975	41,229,297	25,323	0.519
BRCA1	Deletion exon 17	17	41,219,596	41,219,755	160	0.495
BRCA2	Deletion exons 1–2	13	32,889,020	32,890,900	1,881	0.489
BRCA2	Deletion exon 21	13	32,950,734	32,952,070	1,337	0.544

*Breakpoints are flanked by *Alu* and other repeats, which are not captured.
†Reads per base pair for deletion or duplication/reads per base pair for wild-type genotype.