

2415-13

Winter School on Quantitative Systems Biology

26 November - 7 December, 2011

The measure of a genome

R. Sachidanandam
*Mount Sinai Sch. of Medicine
New York
USA*

Measures of a genome

Ravi Sachidanandam
Mount Sinai School of Medicine

Questions ?

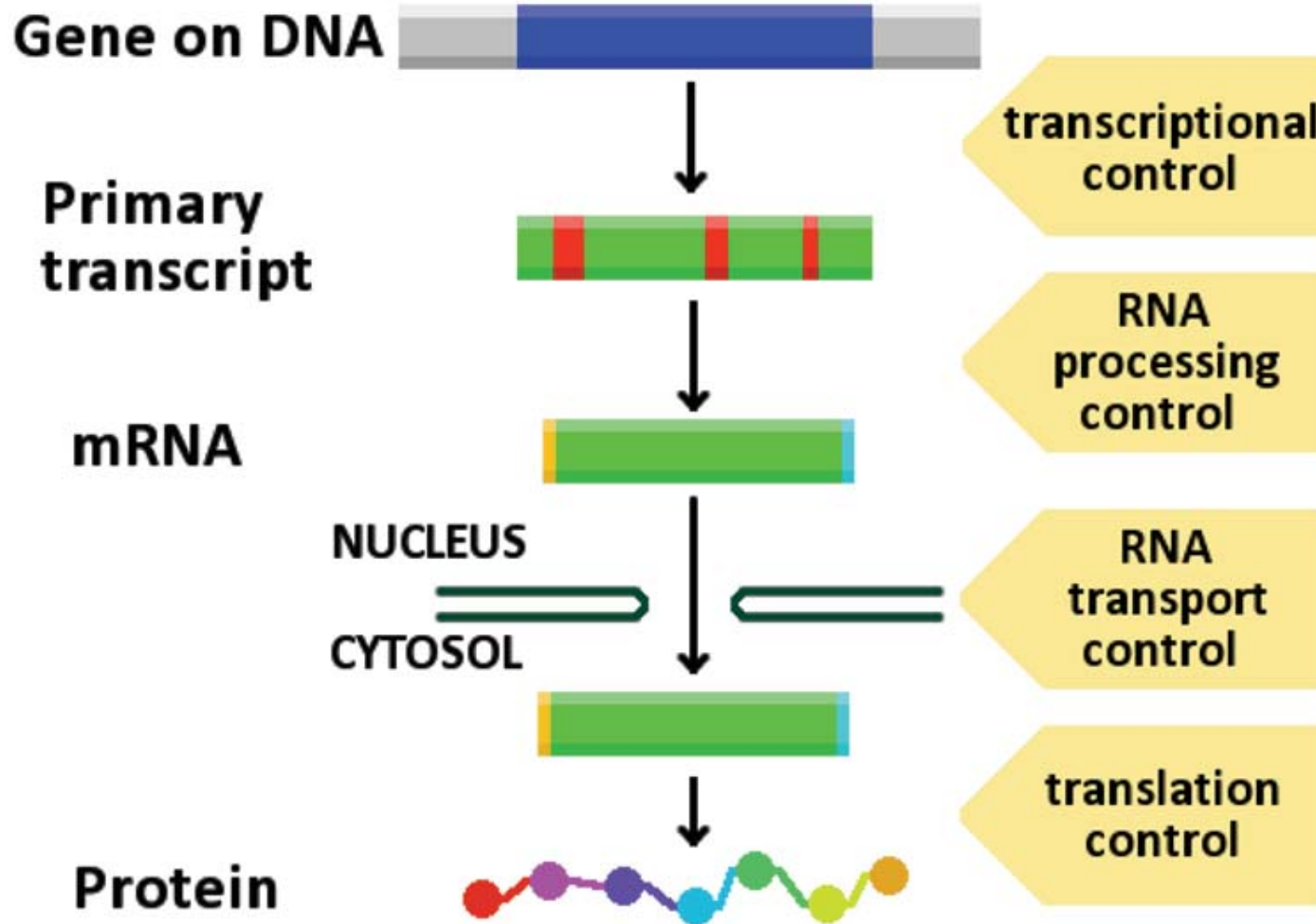
Contact me at ravi.mssm@gmail.com for criticisms, comments and questions.

I have tried to give a quick overview in lectures 1 and 2 of the broad ideas in genomics/ bioinformatics.

I have attached to this presentation, slides containing notes, for some material I worked out on the board, as well as some extra population genetics material that I did not have time to get into.



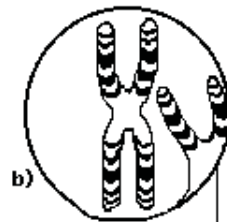
Central Dogma



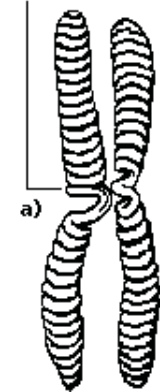
Retroviruses (HIV) go from RNA to DNA

Karyotype

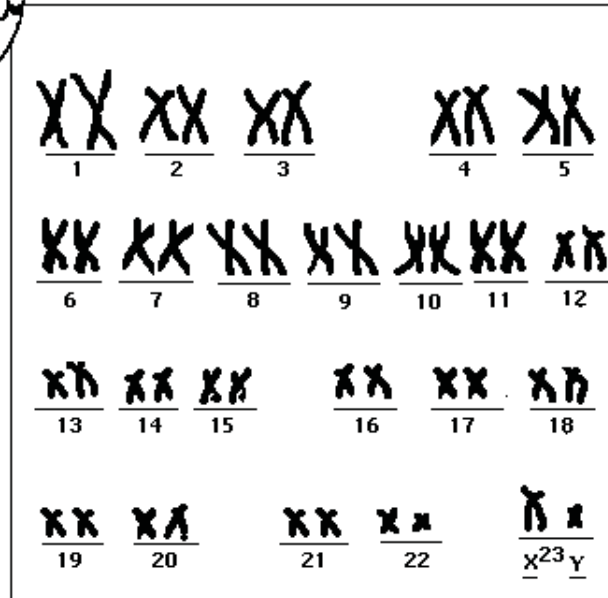
HUMAN CHROMOSOMES



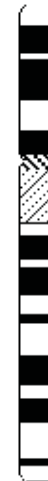
Centromere



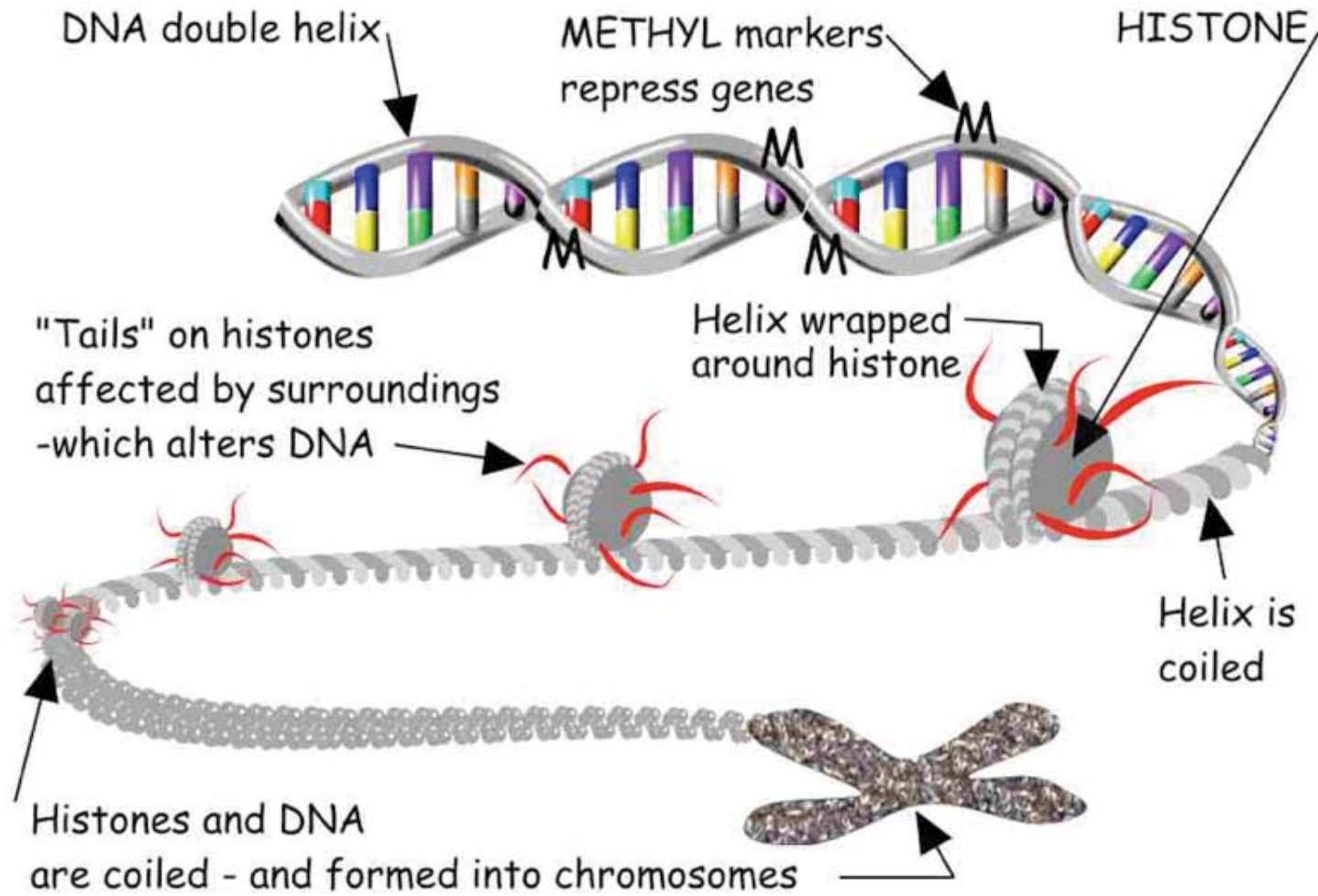
Chromatid



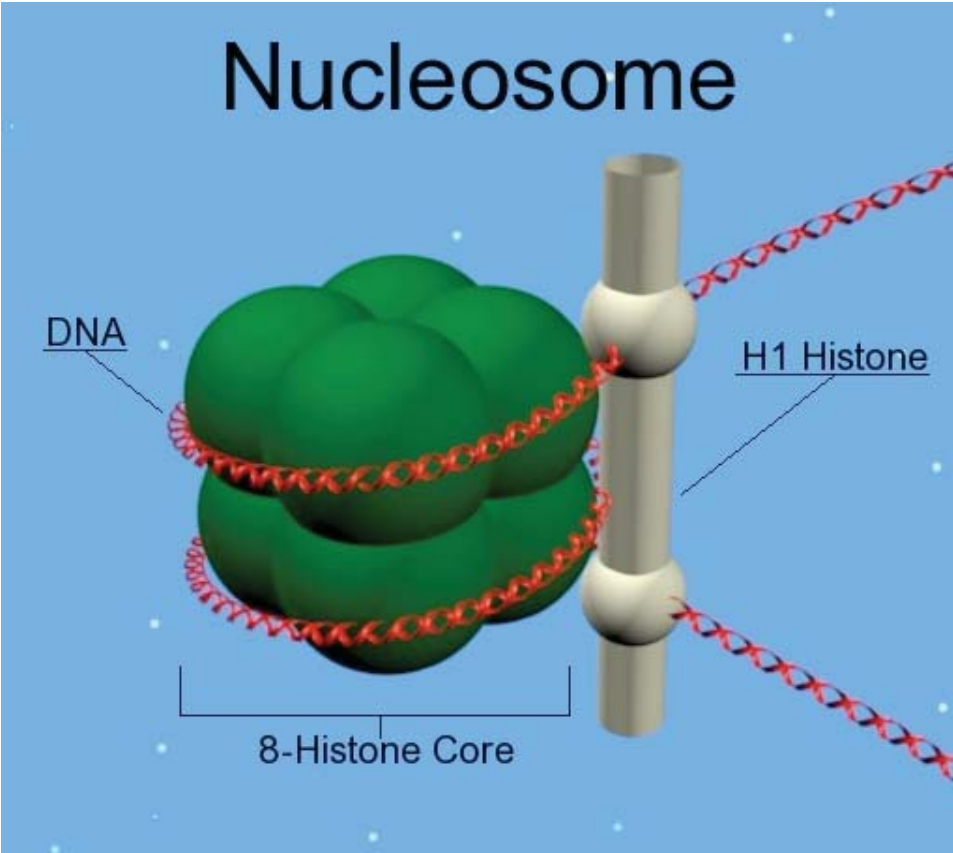
c)



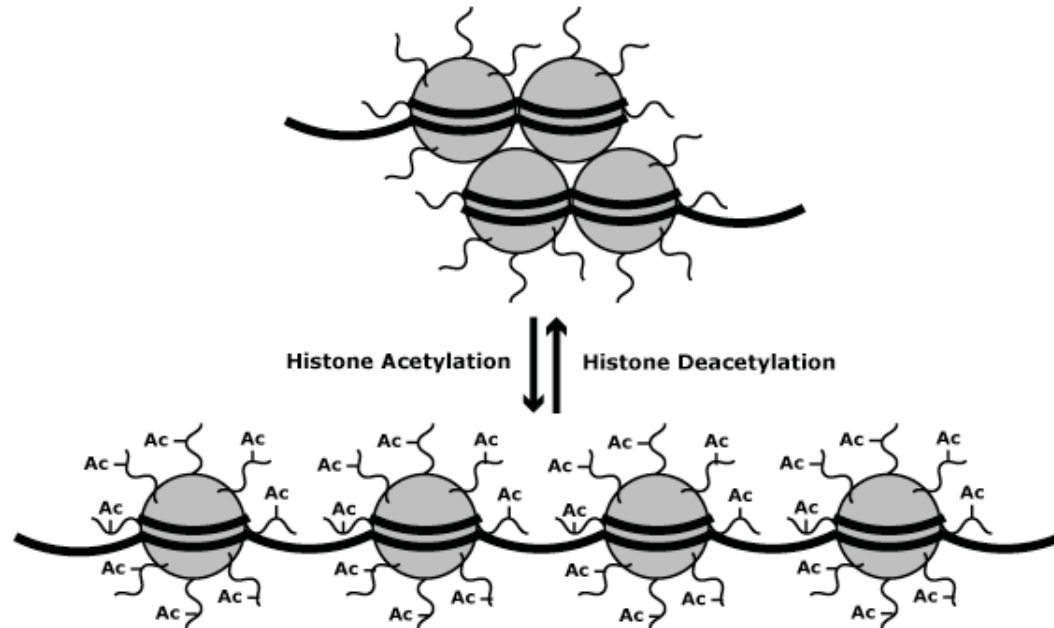
HISTONES AND METHYL MARKERS CONTROL DNA



Nucleosome



Acetylation

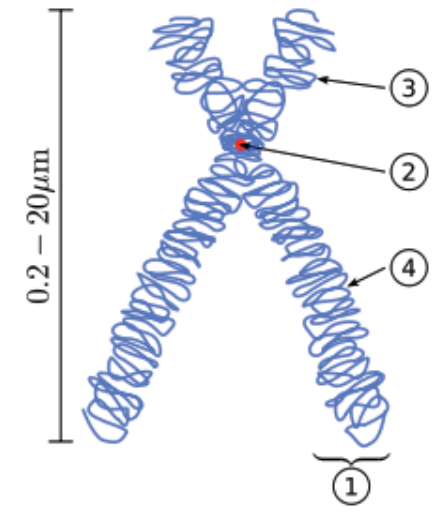
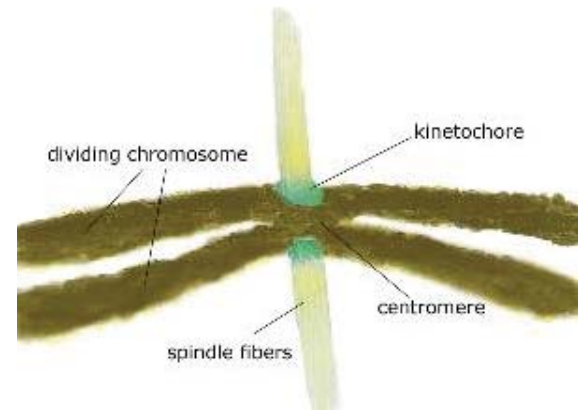
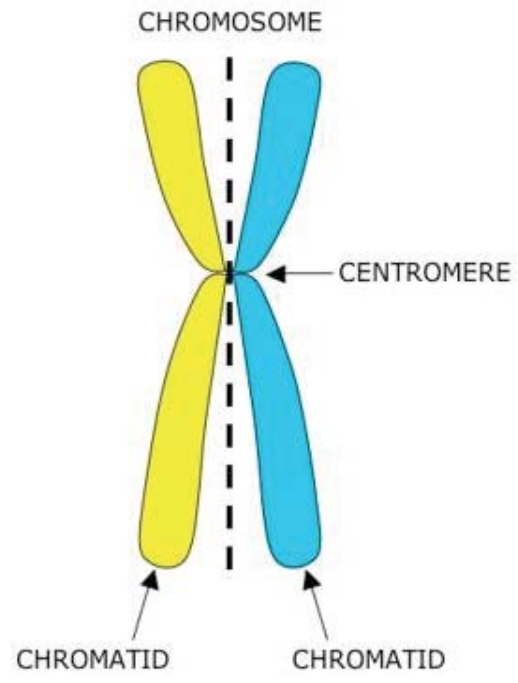


Histone Acetylation

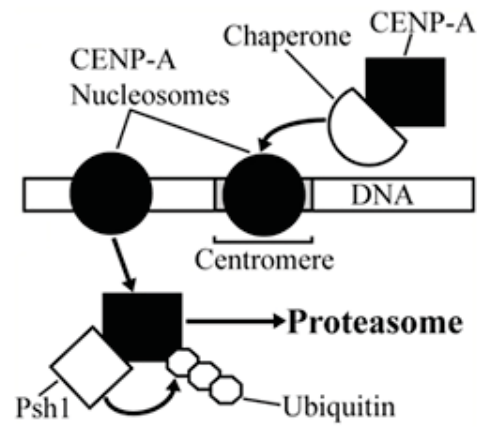
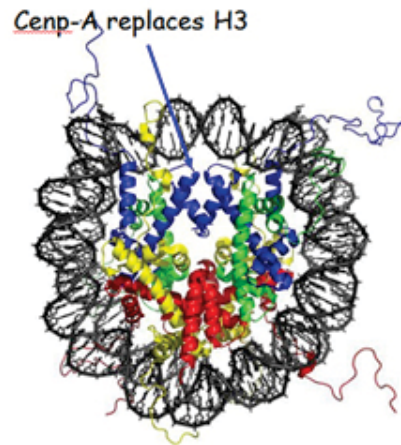
Recall that histones form octameric protein complexes, which DNA wraps around to form a nucleosome. Histones undergo various types of modifications that determine chromatin structure. The best studied modification is histone acetylation.

Histone proteins have N-terminal tails that extend out from the nucleosome core. Histone acetylation involves the attachment of acetyl groups to lysine residues in the N-terminal tails of histone proteins. It is believed that the acetylation of lysine (changing a positively charged residue to a negatively charged residue) decreases the affinity for histones for DNA (and possibly histones for other histones), thereby making DNA more accessible for transcription.

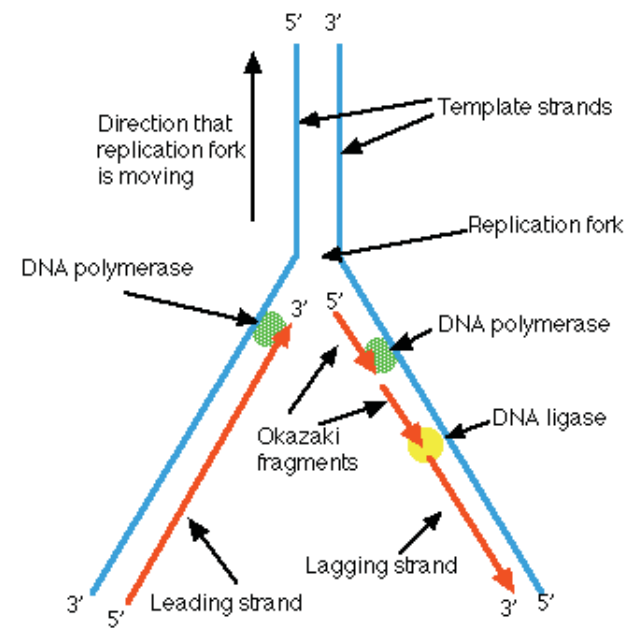
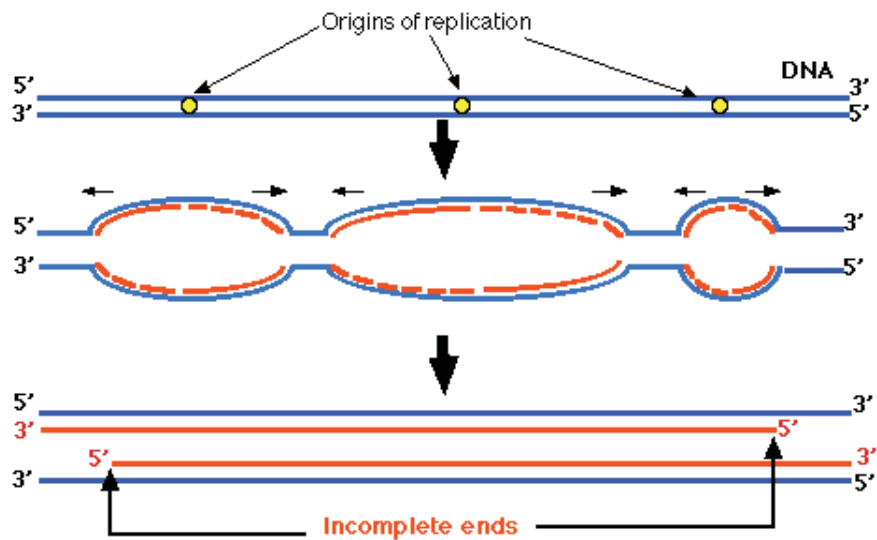
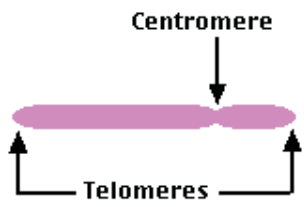
The opposite reaction (deacetylation) removes acetyl groups from lysine residues in the N-terminal tails of histone protein



Centromere



Telomeres

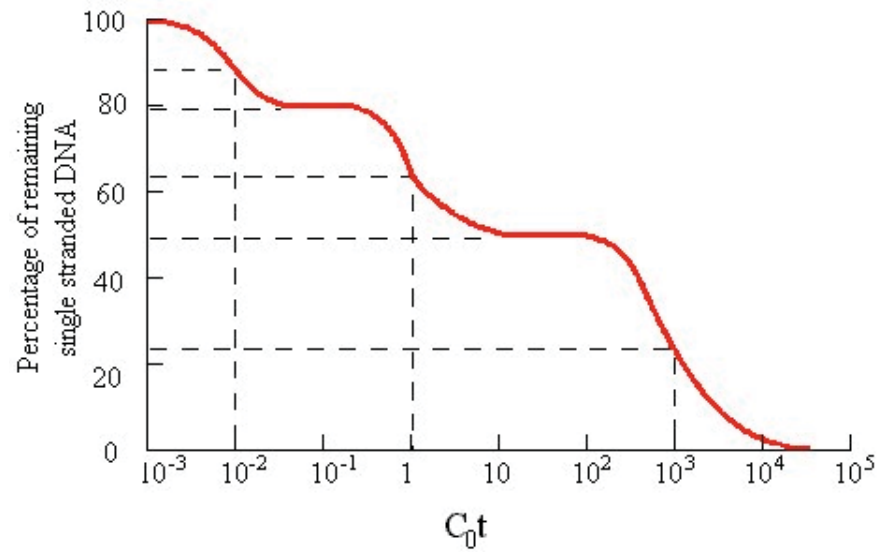


The telomeres of humans consist of as many as 2000 repeats of the sequence 5' TTAGGG 3'.

5' ...TTAGGG TTAGGG TTAGGG TTAGGG TTAGGG TTAGGG...3'
 3' ...AATCCC AATCCC AATCCC AATCCC AATCCC AATCCC...5'

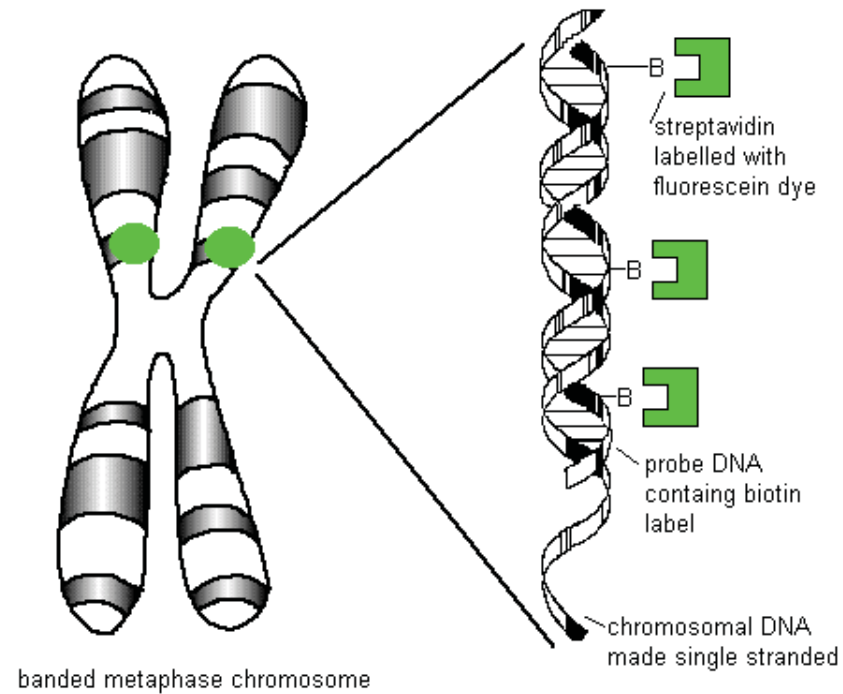
Exploring DNA

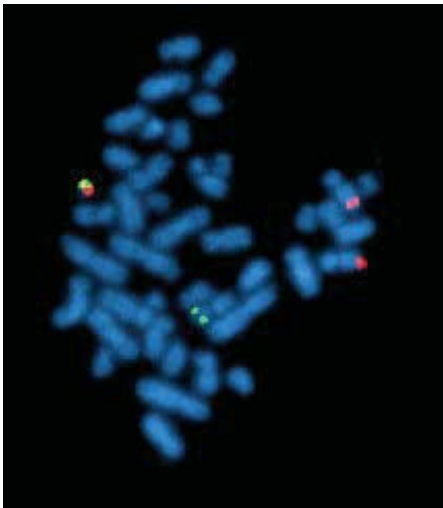
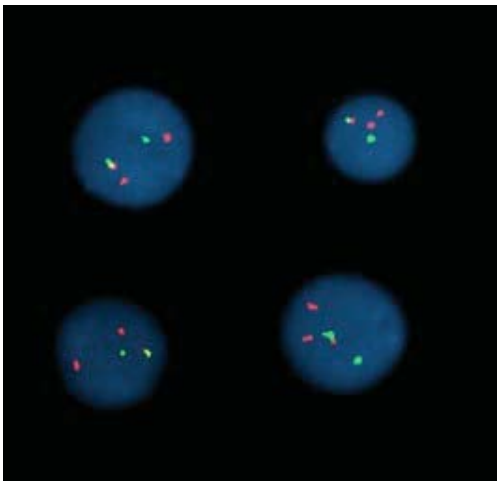
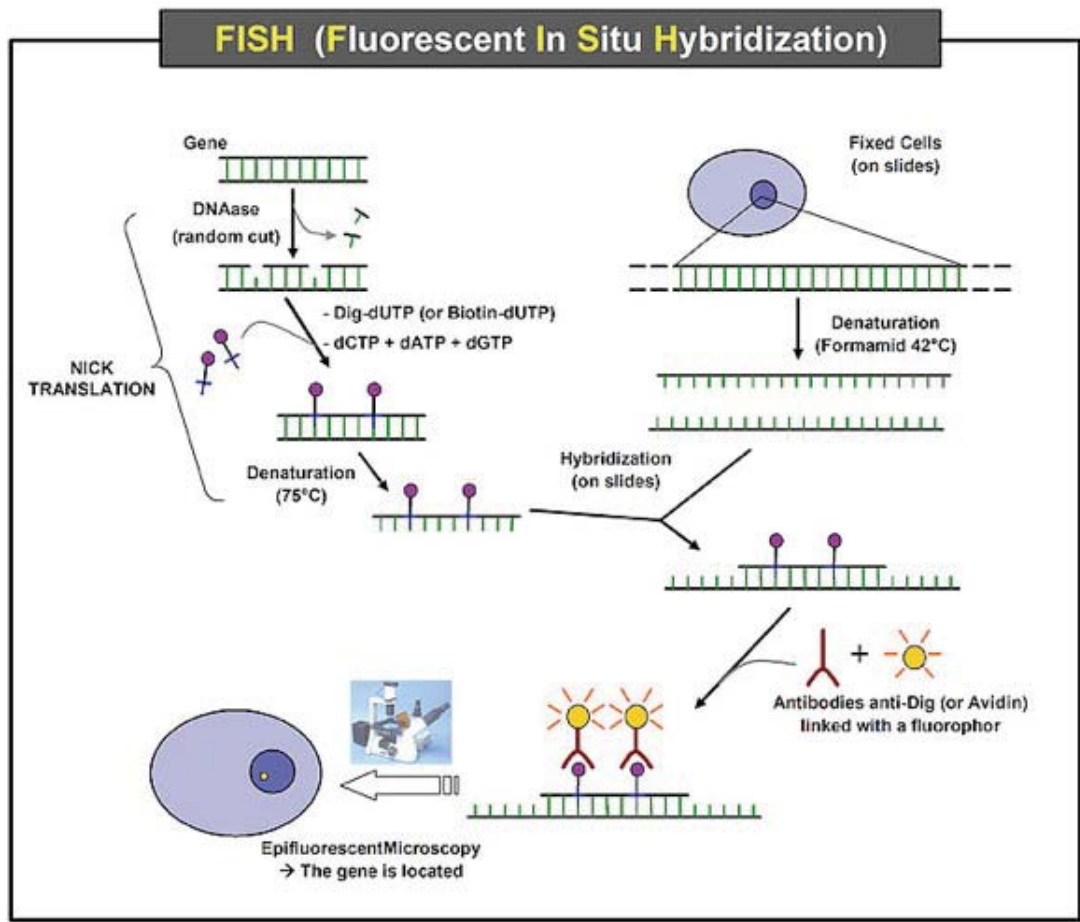
Renaturation of DNA to study complexity



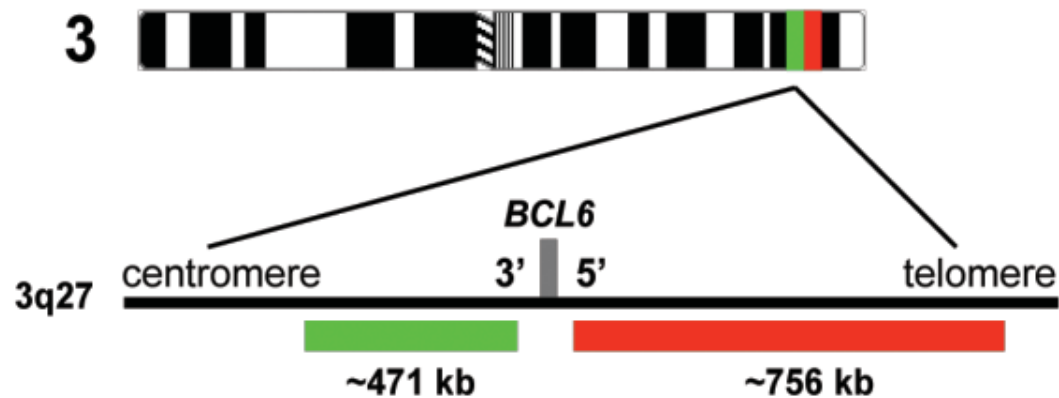
$$\frac{C}{C_0} = \frac{1}{(1 + kC_0t)}$$

FISH- fluorescent in-situ hybridisation

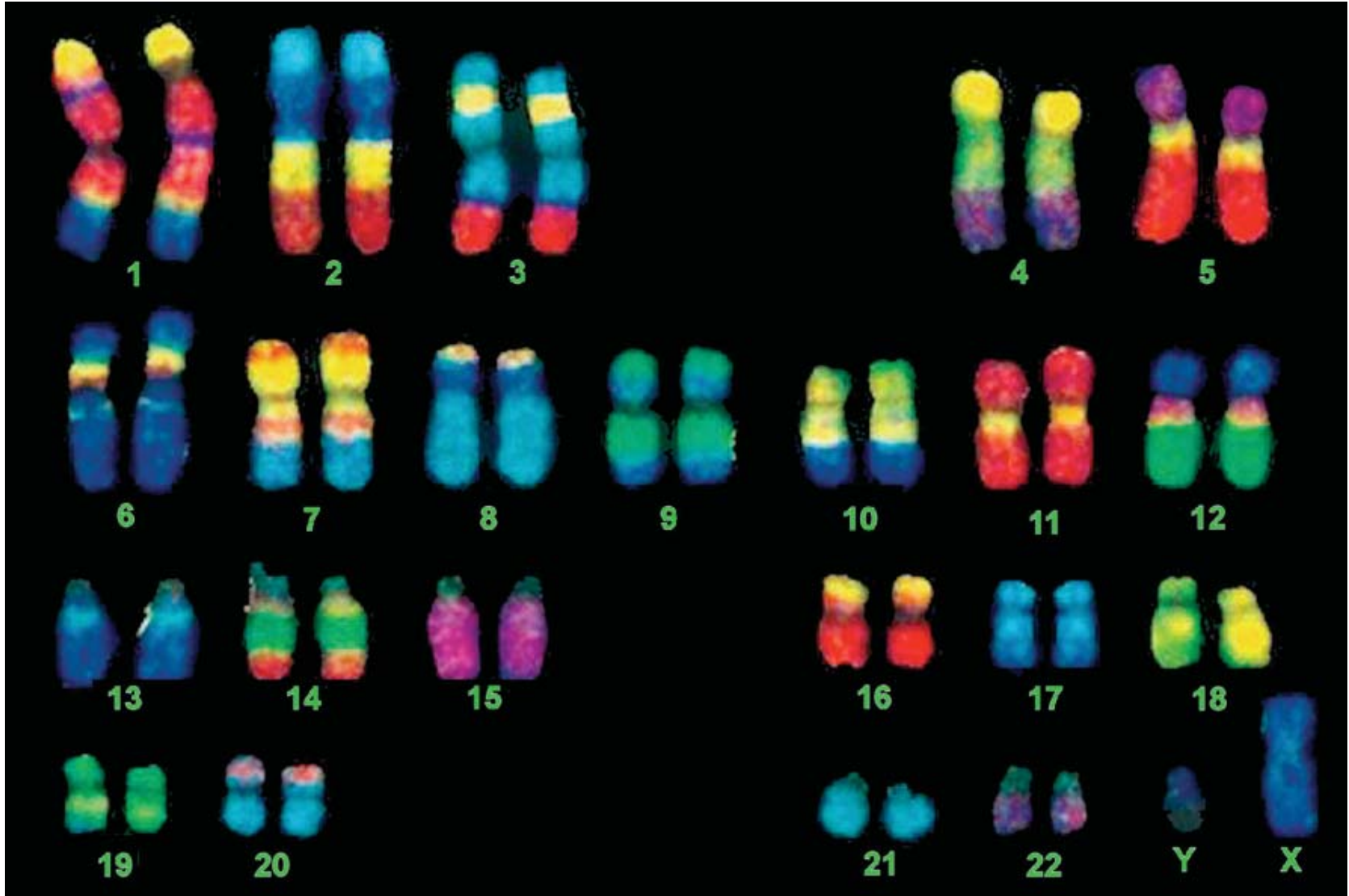




Common rearrangements in B-cell lymphoma



The *BCL6* Break Apart DNA-FISH Probe is designed to detect the translocations between the *BCL6* gene located on 3q27 and one of at least 20 known translocation partner loci as detected by fluorescence *in situ* hybridization (FISH). Translocation of the *BCL6* gene occurs in 6-26% of follicular lymphoma (FL)^[1] with higher incidence (44%) in grade 3 cases negative for t(14;18)(q32;q21).^[2] Rearrangement of the *BCL6* gene is observed at a frequency of 15~40% in diffuse large B-cell lymphomas (DLBCL).^[1-3]



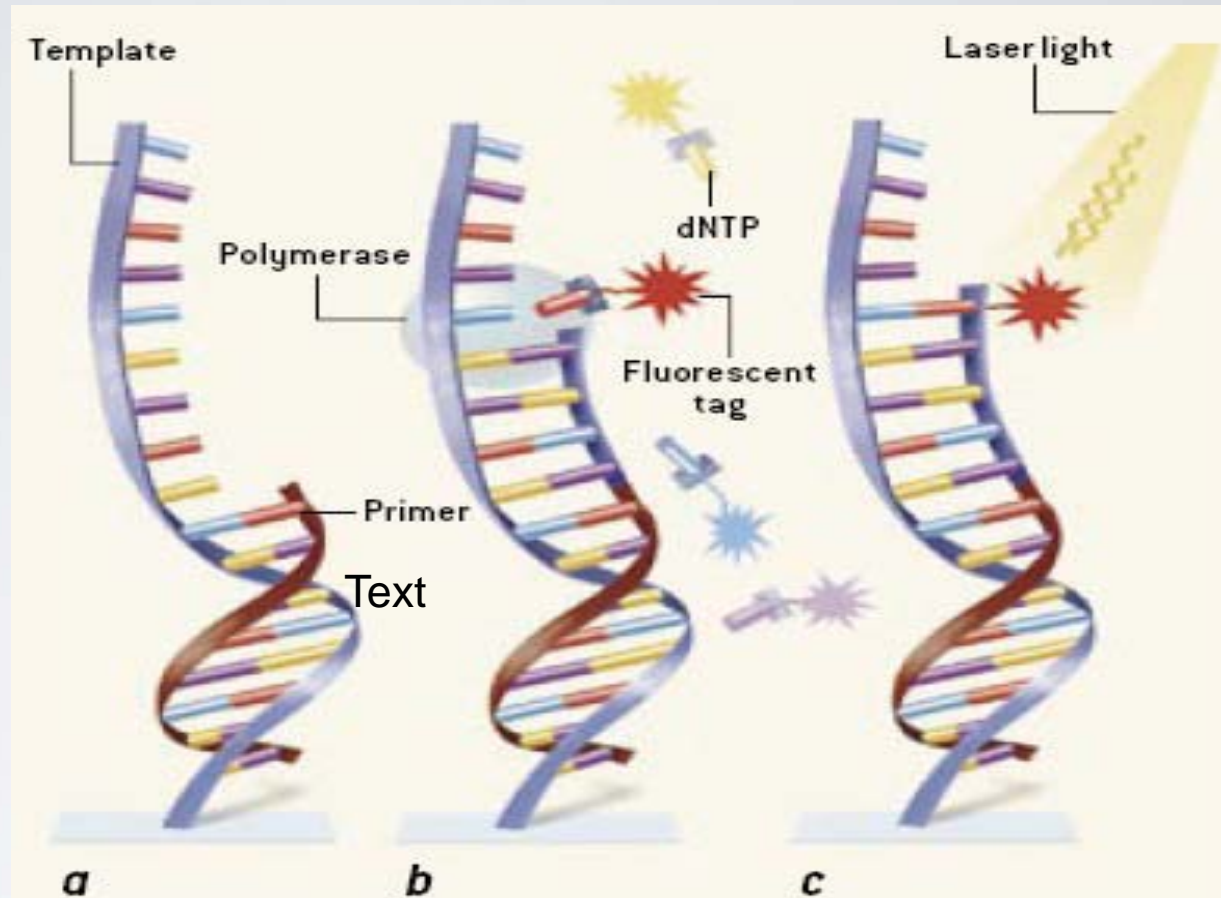
Other Techniques

- Arrays or other specialized variant detection technologies.
 - pros: Cheap, easy to perform en-masse
 - cons: Only known variants, mechanism often unknown
- Whole-exome sequencing.
 - pros: Variant discovery, potentially leading to mechanism
 - cons: limited regions (introns and intergenic regions missed), spotty coverage
- Genome sequencing.
 - pros: Capture everything
 - cons: Capture everything, spotty coverage



DNA base extension

Fluorescence



Bioluminescence



Pyrophosphate detection uses bioluminescence, instead of fluorescence, to signal base-extension events. A pyrophosphate molecule is released when a base is added to the complementary strand, causing a chemical reaction with a luminescent protein that produces a flash of light.

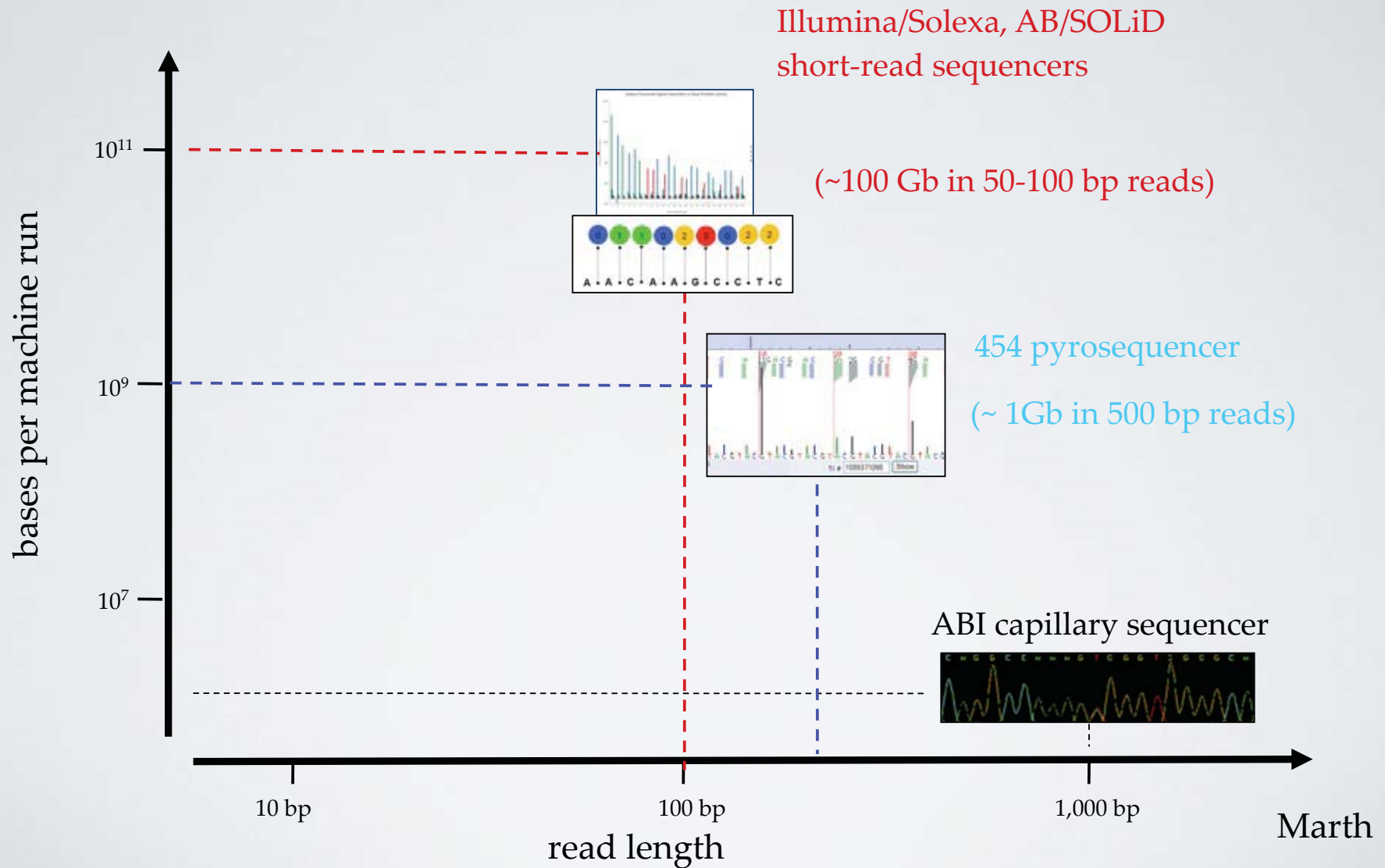
Sequencing Basics

- Traditional capillary sequencing - terminate the growth, sampling the distribution of nucleotides at different lengths. Inefficient, because lots of reactions per sequence needed.
- Massively parallel techniques observe the process in groups of molecules, or single molecules, without terminating the process.

Sequencing process

- Each base extension is converted to a pulse of light. Thus, sequencing reduces to image processing.
- Bases called from images are used to assemble linear sequences.
- linear sequences are analyzed, either assembled or mapped and annotated.

Read length and throughput

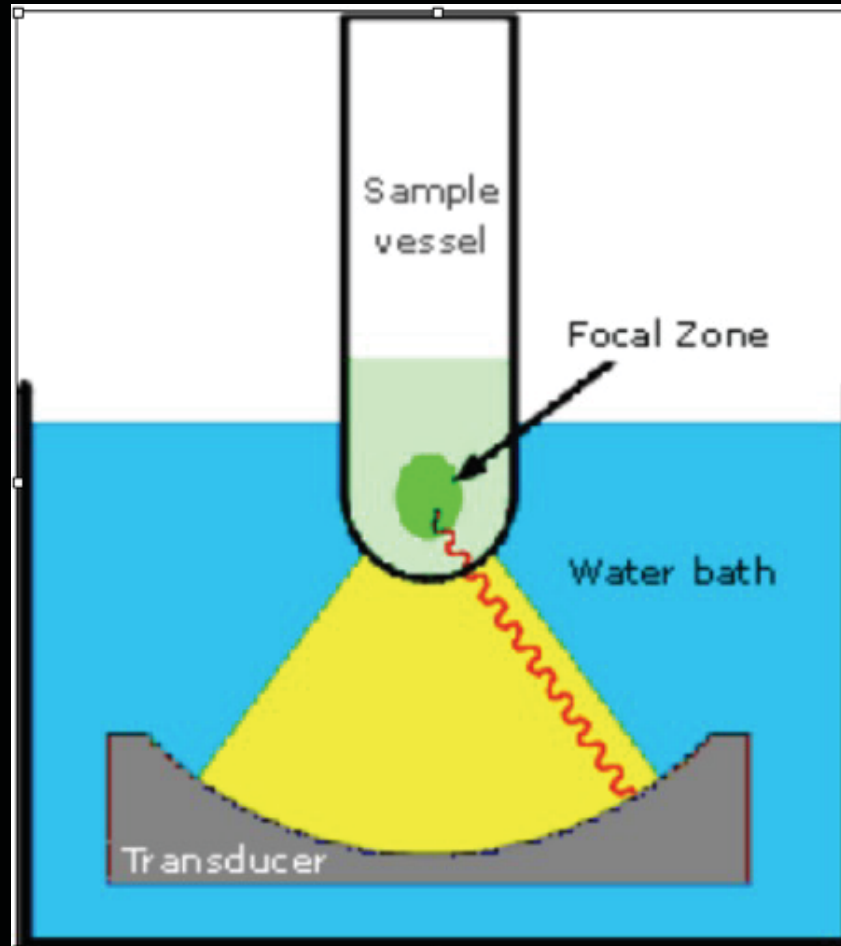


Covaris: Parameters

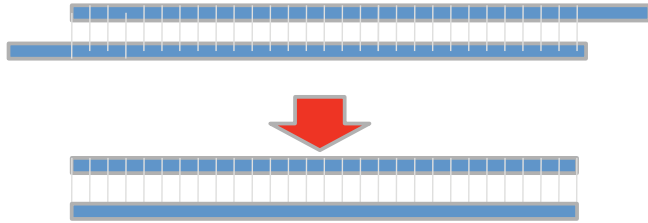
Duty cycle—or the percent of time that the transducer is creating acoustic waves

Intensity—the amplitude of the pressure created by the transducer

Cycles/burst—the number of waves generated by the transducer in a burst



5. Perform End Repair



6. Add "A" Bases to the 3' Ends



-Provides anchor for adaptor attachment which has a T overhang - Prevents concatimerization

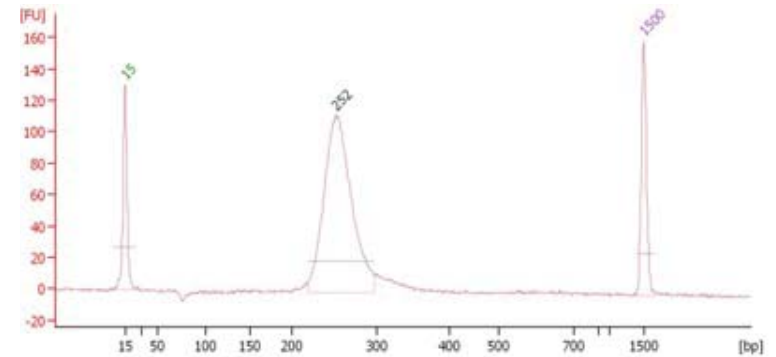
7. Ligation of Adapters to DNA Fragments



The Adapters have a 20 bp overhang, apart from the sequencing primer binding site, to attach to the Solexa flow cell

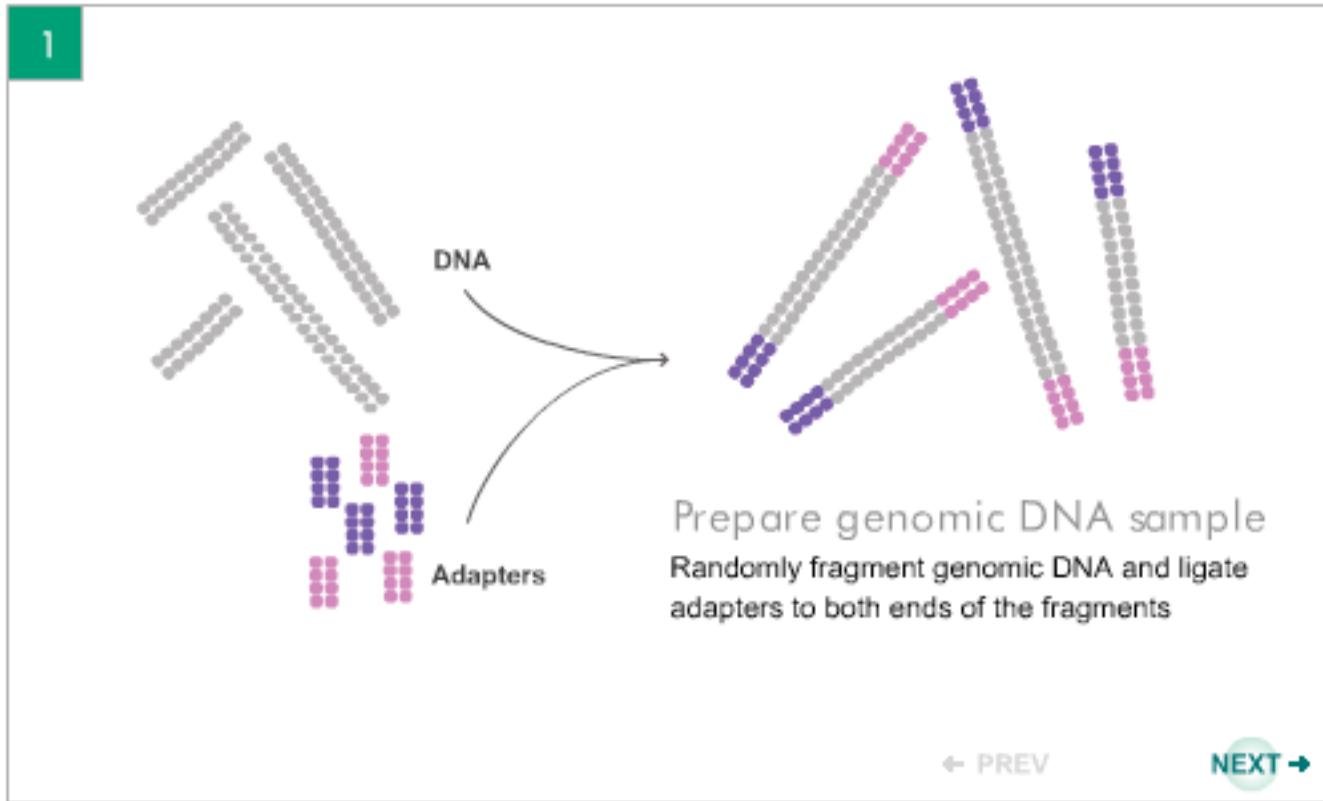
8. Size Selection for Ampl/Seq

Run samples on a 2% agarose gel, excise 250bp region from gel, Amplify

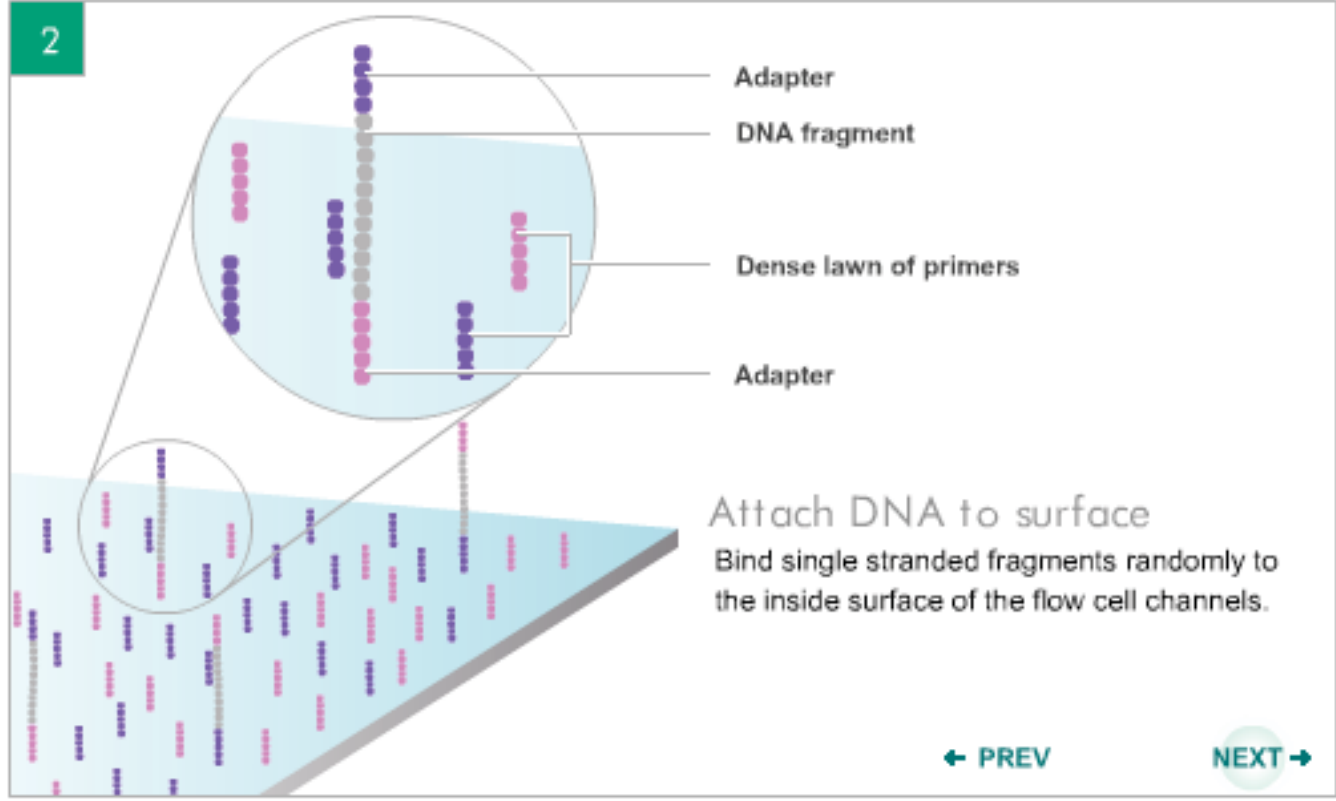


384 MEM Library, (250bp fragments)

Solexa Technology

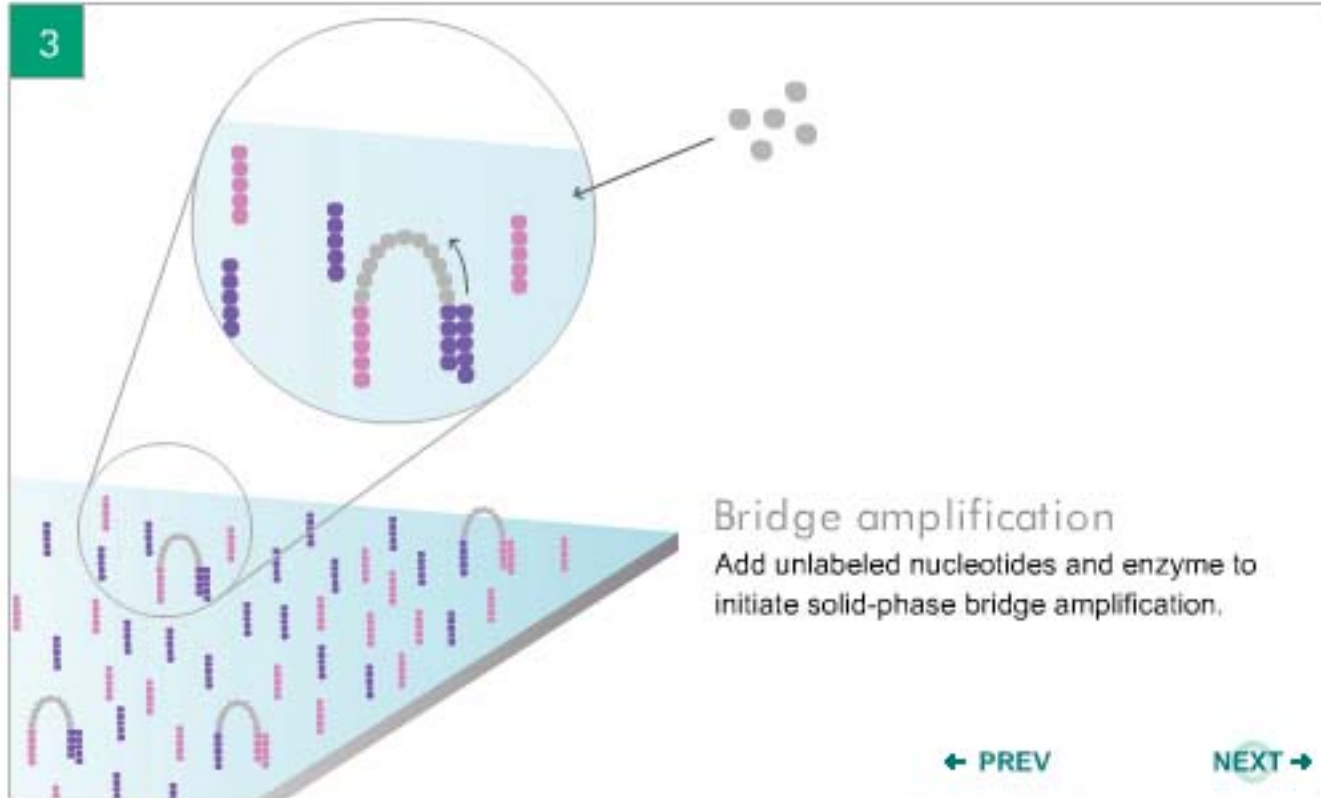


2



Illumina

3



Bridge amplification

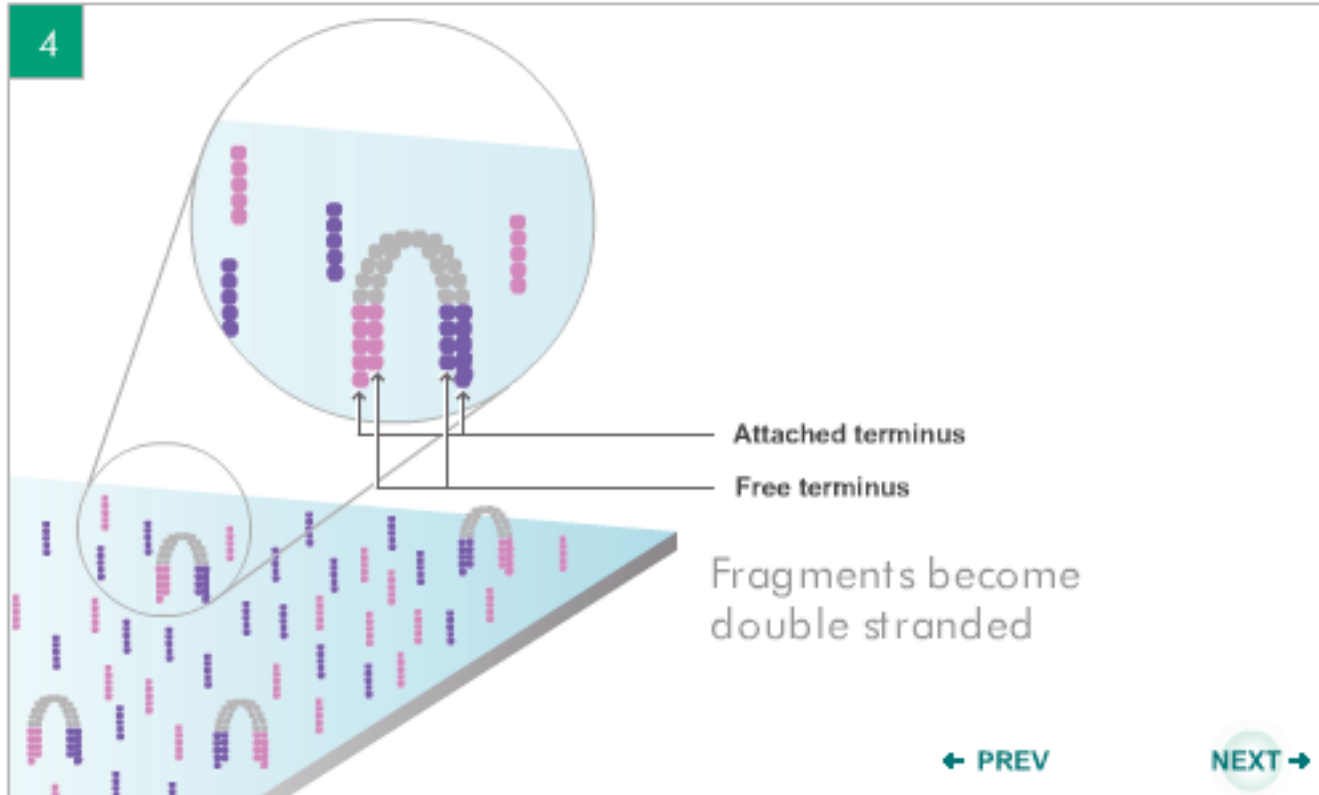
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

← PREV

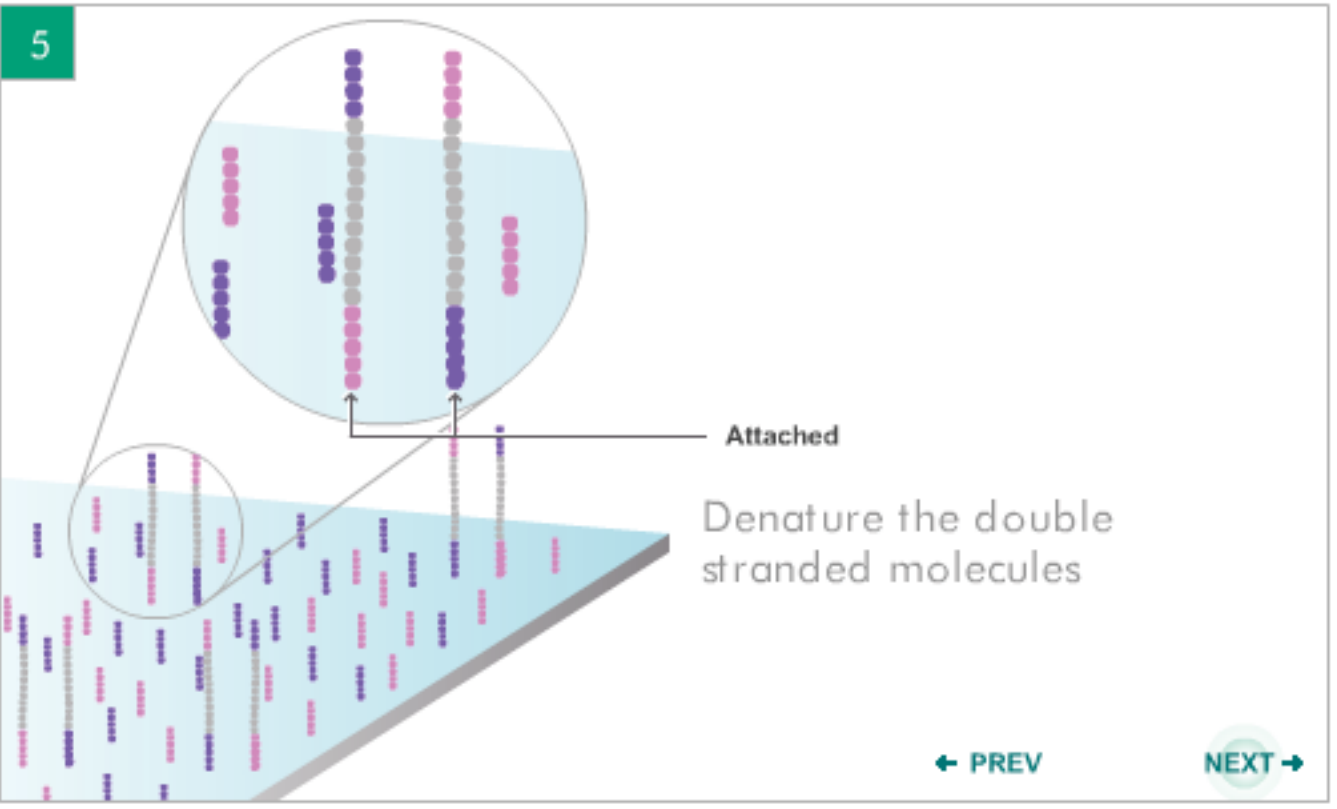
NEXT →

Illumina

4

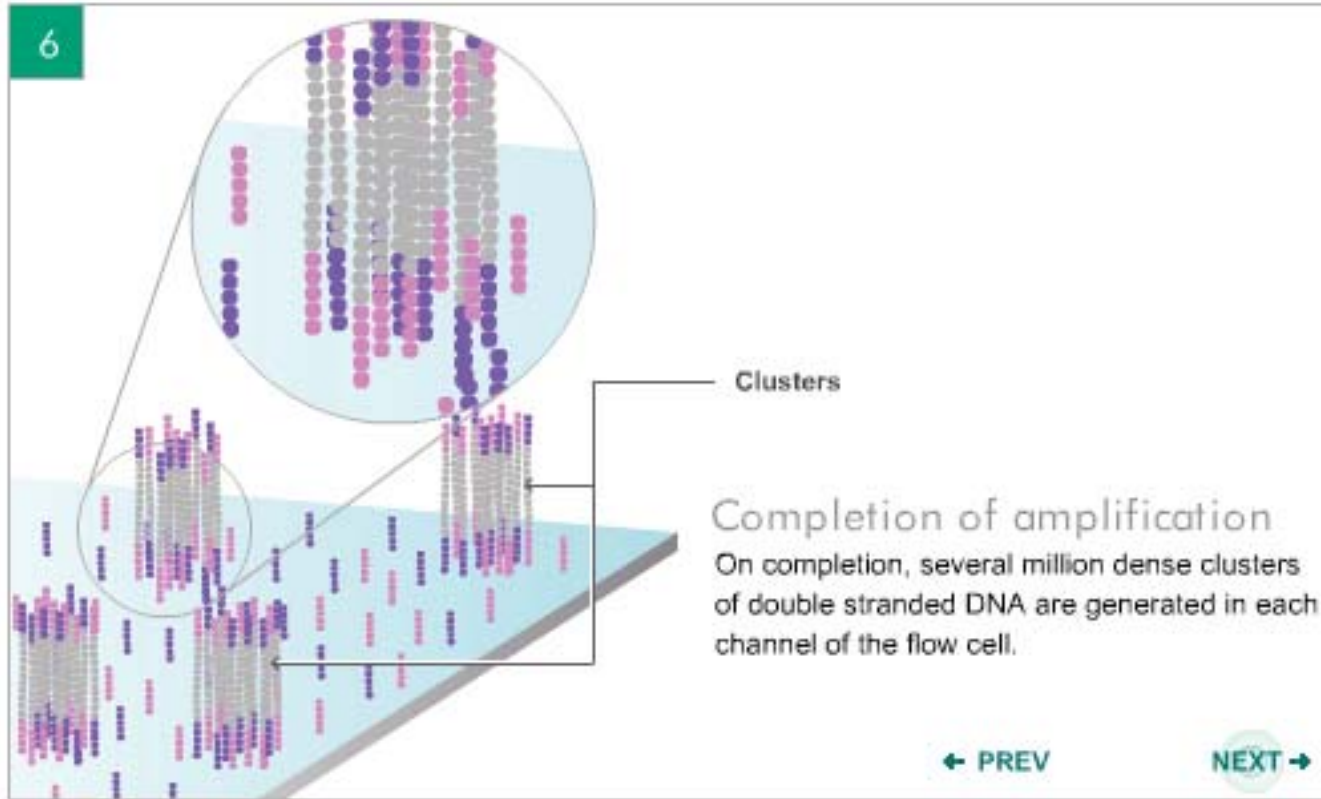


Illumina

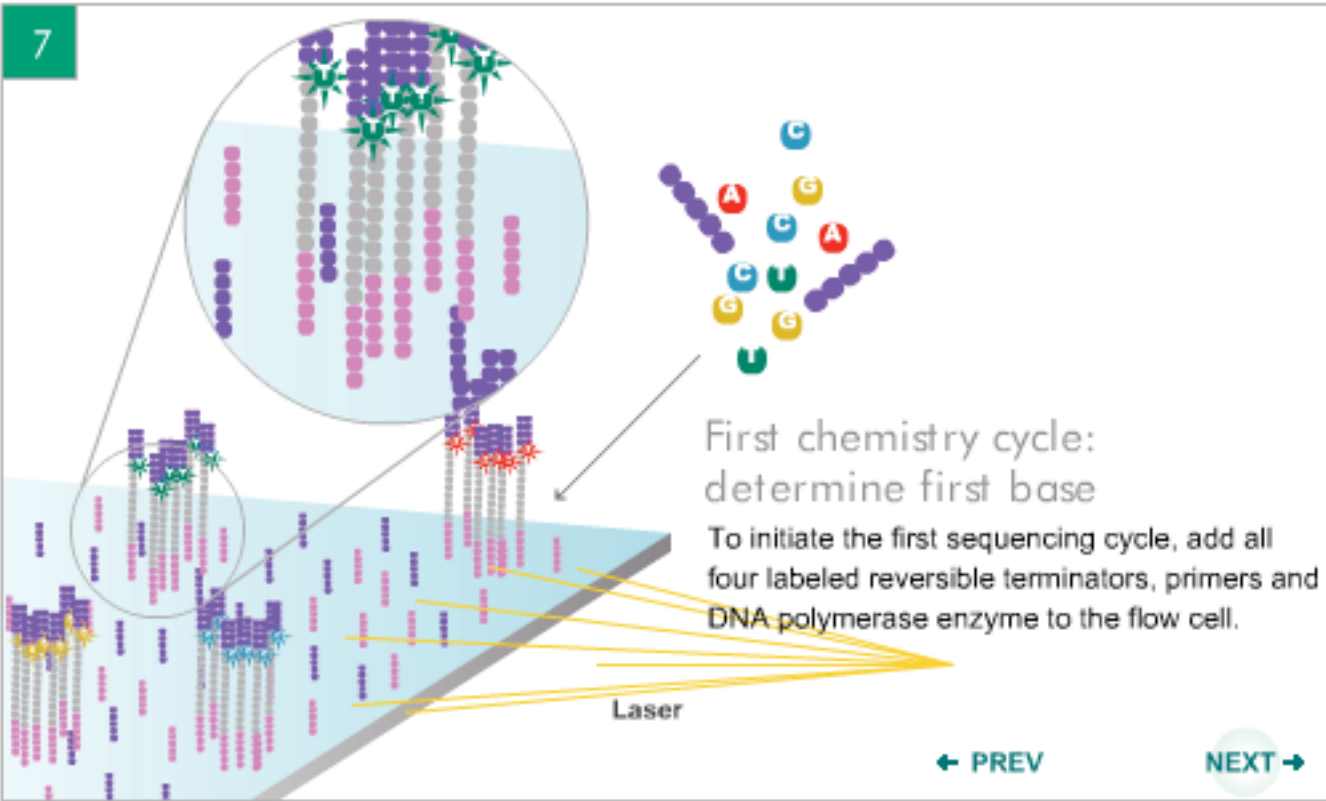


Illumina

6



Illumina



Illumina

Assembling reads

- de Bruijn graphs
 - Hamiltonian paths
 - Eulerian paths

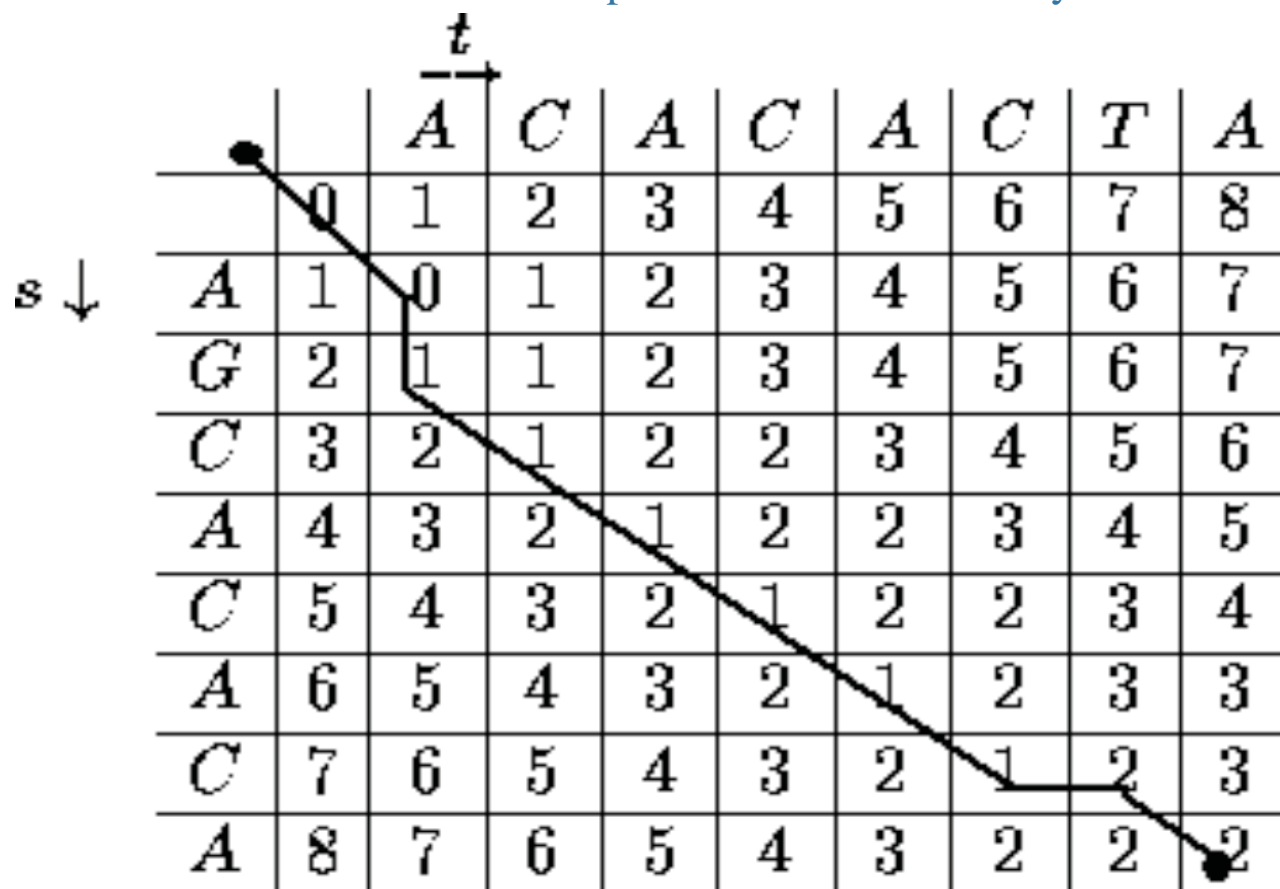
Alignments

s : $A G C A C A C - A$ or $A G - C A C A C A$
 t : $A - C A C A C T A$

All alignments are paths through the Needleman-Wunsch matrix

Global - Needleman-Wunsch- terminals of path at top-left and bottom-right corners

Local - Smith-Waterman-Gotoh - path can start and end anywhere in matrix



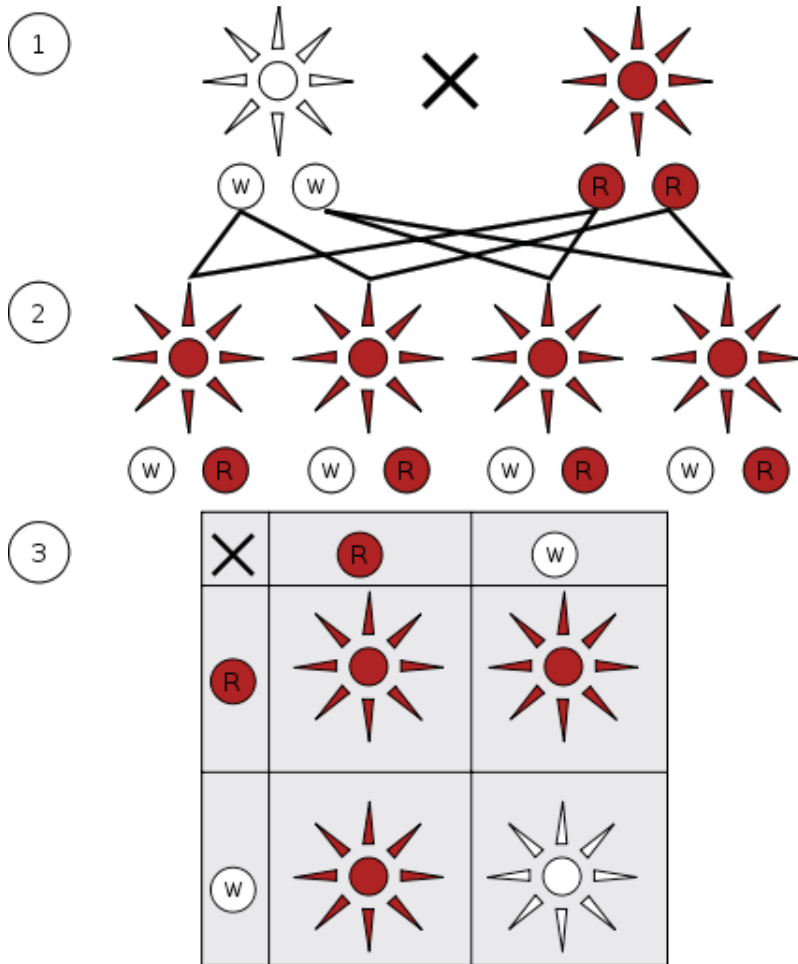
Genetics	Epigenetics
Mendelian Inheritance	Lamarckian Inheritance
DNA-sequence based	Environment-based
	DNA-methylation
	Histone modifications

Lamarckian effects:

- 1) Starvation leads to DNA-methylation changes that persist over generations, studied in children and grandchildren of dutch women who were pregnant during WWII and starved.
- 2) Imprinting is variable expression based on parental origin, There are theoretical arguments for size of fetus being controlled by father's genes.

Mendel's Laws

Law of Segregation The "First Law"

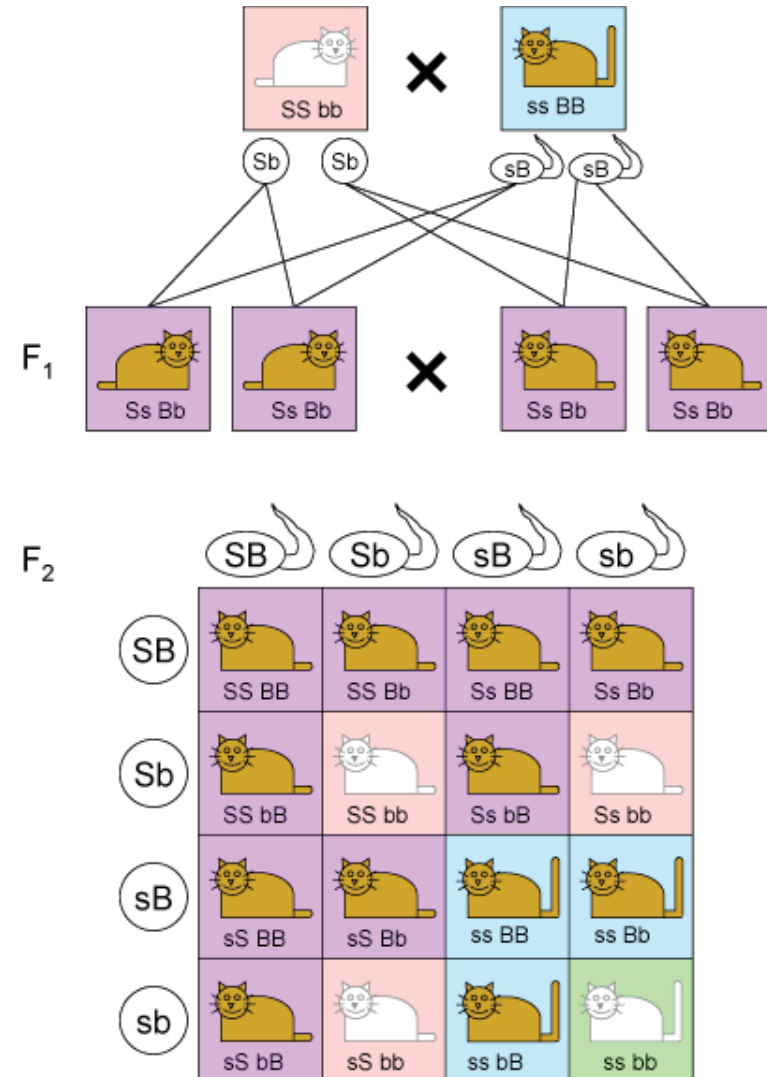


Dominant and recessive phenotypes.

(1) Parental generation.

(2) F₁ generation

Law of Independent Assortment The "Second Law"



Dihybrid cross. The phenotypes of two independent traits show a 9:3:3:1 ratio in the F₂ generation. In this example, coat color is indicated by **B** (brown, dominant) or **b** (white), while tail length is indicated by **S** (short, dominant) or **s** (long).

The Hardy-Weinberg Principle is the fundamental model of population genetics.

Today, the Hardy-Weinberg Law stands as a kind of Newton's First Law (bodies remain in their state of rest or uniform motion in a straight line, except insofar as acted upon by external forces) for evolution: Gene frequencies in a population do not alter from generation to generation in the absence of migration, selection, statistical fluctuation, mutation, etc.

Robert M. May (2004)

Not fundamental, derived from Mendel's laws

Frequencies of an allele in the population

Frequency of A	p
Frequency of a	$q = 1-p$

Frequencies of diploid genotypes in the population

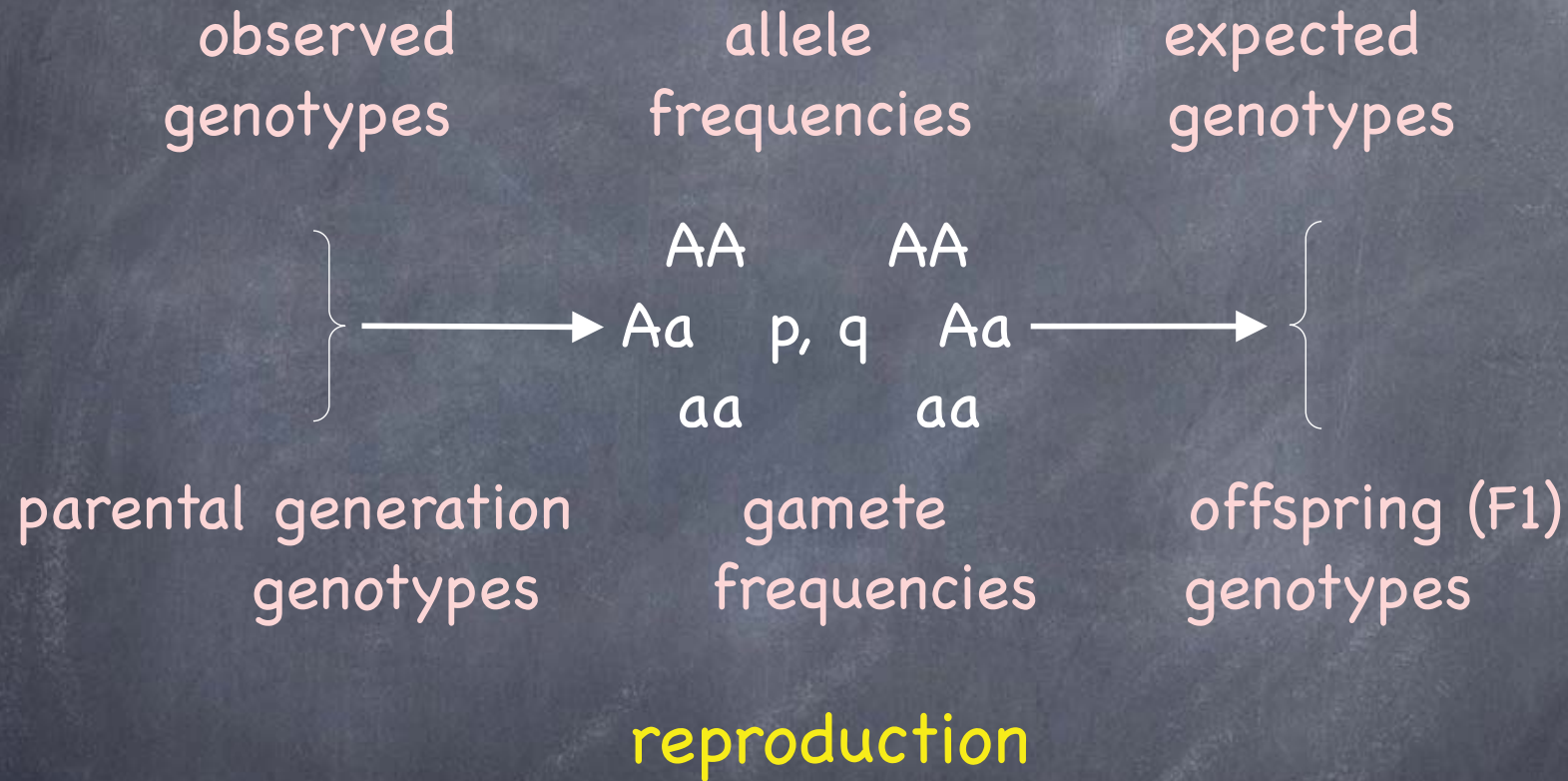
$$AA = p^2$$

$$Aa = 2pq$$

$$aa = q^2$$

$$\begin{aligned}(p + q)^2 &= p^2 + 2pq + q^2 \\ &= 1\end{aligned}$$

Hardy-Weinberg (single generation)



Assumptions of Hardy-Weinberg model

1. Random mating.
2. No mutation.
3. Large (infinite) population size.
4. No differential survival or reproduction (i.e., no natural selection).
5. No immigration

What's a G test?

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$G = 2 \sum_i O_i \cdot \ln(O_i/E_i)$$

G is closer to a true chi-square distribution and is used often in genetic studies to test if distribution is from expected models

Utility of the Hardy-Weinberg Model

Essential for understanding genetic variation in natural populations in various studies:

- Conservation
- Evolution
- Medicine
- Forensics
- Genetic counseling

Variations in DNA

- Edits:

- SNPs (Single nucleotide polymorphisms)

- Indels

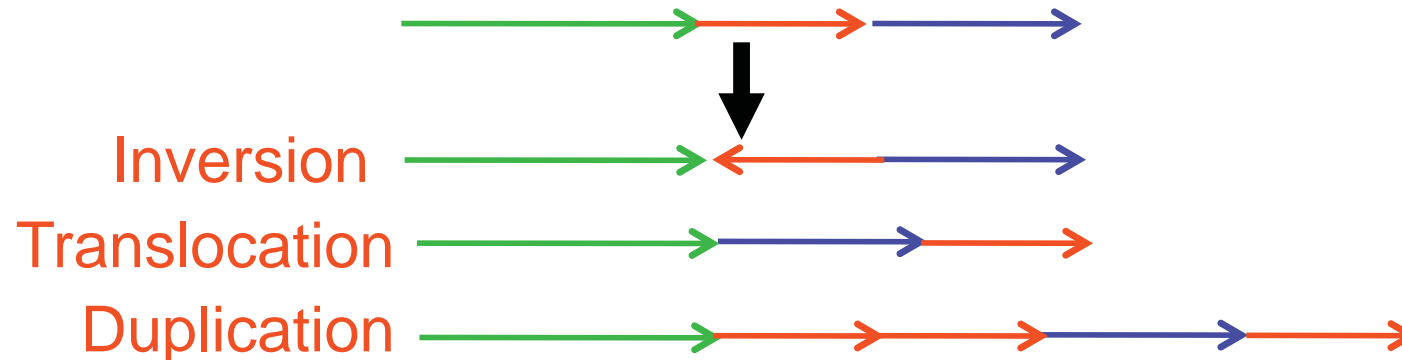


- Rearrangements

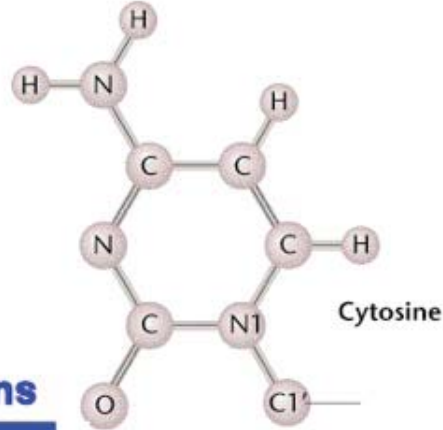
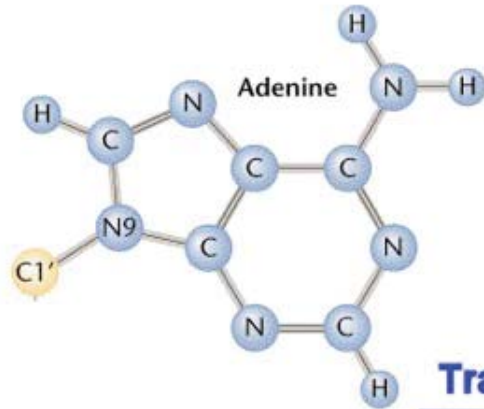
- Duplication CNV (Copy number variations)

- Inversion

- Translocation



Types of single nucleotide changes in DNA

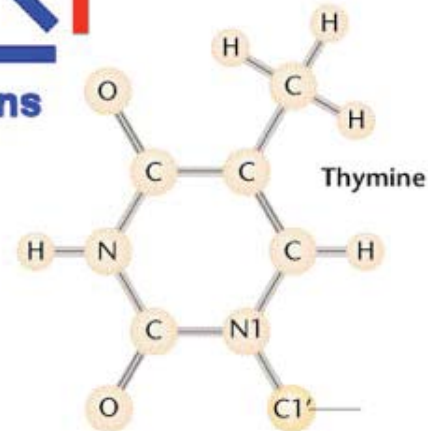
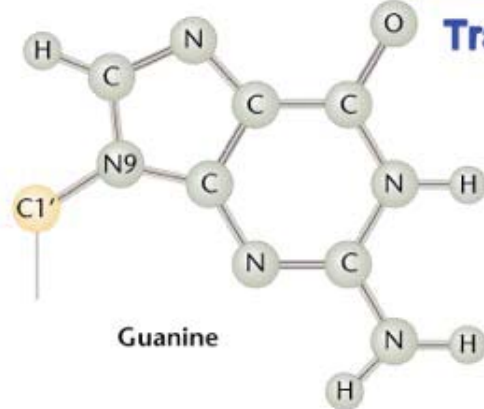


Transversions

$T_i = \text{Transitions}$

Transitions = T_v

Transversions



C → T and G → A
more frequent than
T → C and A → G
due to C-methylation followed by de-
amination

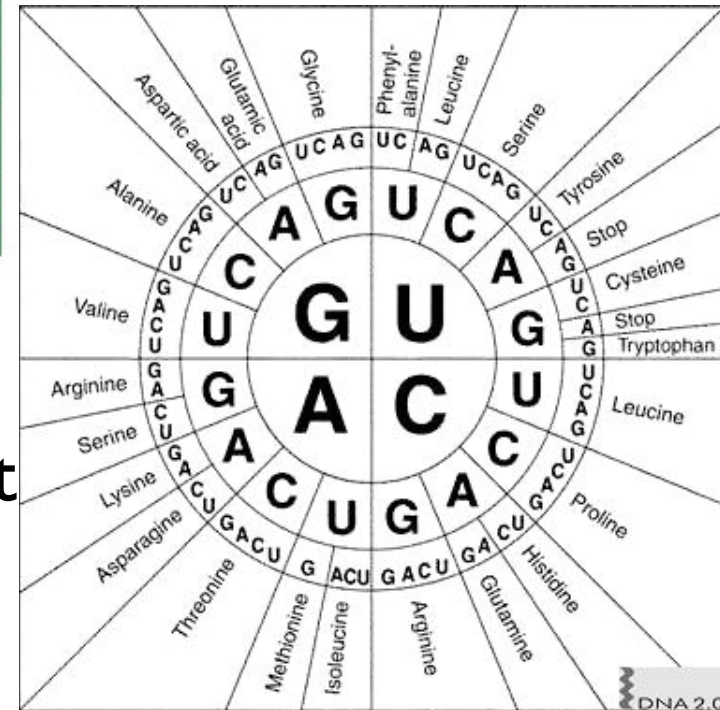
expected
 $T_i/T_v \sim 2$

in Coding
sequences
 $T_i/T_v \sim 3$

The universal code

		Second base					
		U	C	A	G		
First base	U	UUU } Phenyl-alanine F UUC } UUA } Leucine L UUG }	UCU } Serine S UCC } UCA } UCG }	UAU } Tyrosine Y UAC } UAA } Stop codon UAG } Stop codon	UGU } Cysteine C UGC } UGA } Stop codon UGG } Tryptophan W	U	C
	C	CUU } Leucine L CUC } CUA } CUG }	CCU } Proline P CCC } CCA } CCG }	CAU } Histidine H CAC } CAA } Glutamine Q CAG }	CGU } Arginine R CGC } CGA } CGG }	U	C
	A	AUU } Isoleucine I AUC } AUA } AUG } Methionine start codon M	ACU } Threonine T ACC } ACA } ACG }	AAU } Asparagine N AAC } AAA } Lysine K AAG }	AGU } Serine S AGC } AGA } Arginine R AGG }	U	C
	G	GUU } Valine V GUC } GUA } GUG }	GCU } Alanine A GCC } GCA } GCG }	GAU } Aspartic acid D GAC } GAA } Glutamic acid E GAG }	GGU } Glycine G GGC } GGA } GGG }	U	C

Actually, not totally universal, mitochondria has a slightly different code, as do several bacteria



Codon Bias

Organism	AGG arginine	AGA arginine	CUA leucine	AUA isoleucine	CCC proline
<i>Escherichia coli</i>	1.4	2.1	3.2	4.1	4.3
<i>Homo sapiens</i>	11.0	11.3	6.5	6.9	20.3
<i>Drosophila melanogaster</i>	4.7	5.7	7.2	8.3	18.6
<i>Caenorhabditis elegans</i>	3.8	15.6	7.9	9.8	4.3
<i>Saccharomyces cerevisiae</i>	9.3	21.3	13.4	17.8	6.8
<i>Plasmodium falciparum</i>	4.1	20.2	15.2	33.2	8.5
<i>Clostridium pasteurianum</i>	2.4	32.8	6.0	52.5	1.0
<i>Pyrococcus horikoshii</i>	30.3	20.4	18.0	44.9	10.1
<i>Thermus aquaticus</i>	13.7	1.4	3.2	2.0	43.0
<i>Arabidopsis thaliana</i>	10.9	18.4	9.8	12.6	5.2

Table 1

Codon Usage in Various Organisms

Codon frequencies are expressed as codons used per 1000 codons encountered. The arginine codons AGG and AGA are recognized by the same tRNA and should therefore be combined. Codon frequencies of more than 15 codons/1000 codons are shown in bold to help identify a codon bias that may cause problems for high-level expression in *E. coli*. These frequencies are updated regularly. A complete compilation of codon usage of the sequences in the gene bank database can be found at www.kazusa.or.jp/codon/.

Organisms exhibit differing preferences for synonymous codons, thought to arise from differences in tRNA efficiency

Kimura's Neutral Theory

Vast majority of base substitutions are neutral with respect to fitness

Genetic drift dominates evolution at the molecular level

Rate of molecular evolution is equal to the neutral mutation rate

What about deleterious and advantageous mutations?

Neutral theory (Motoo Kimura):

Deleterious mutations will tend to be eliminated

Advantageous mutations are very rare

Therefore, molecular evolution will be dominated by drift.

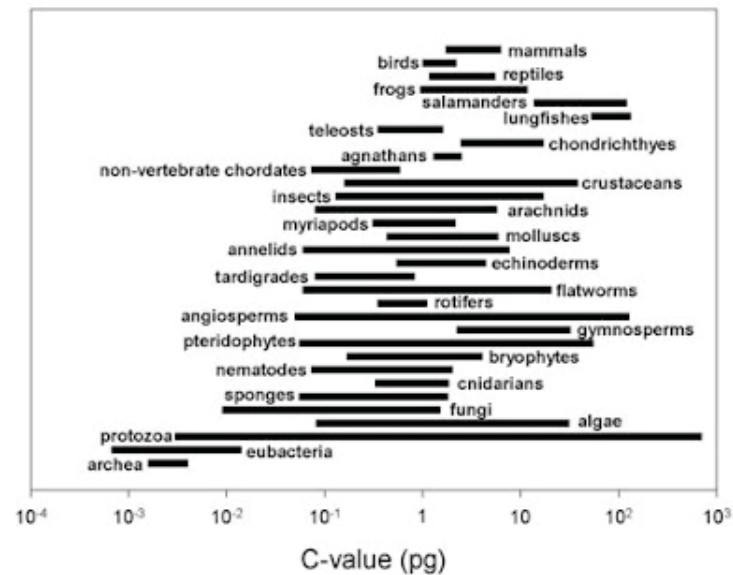
Selectionist theory (John Gillespie):

Advantageous mutations too common to be ignored

Molecular evolution will show substantial evidence of selection.

C-value paradox

Size of the genomes



$$1 \text{ pg} = 10^9 \text{ nt}$$

In amphibians where the smallest genomes are just below 10^9 bp while the largest are almost 10^{11} . It is hard to believe that this could reflect a 100-fold variation in the number of genes needed to specify different amphibians.

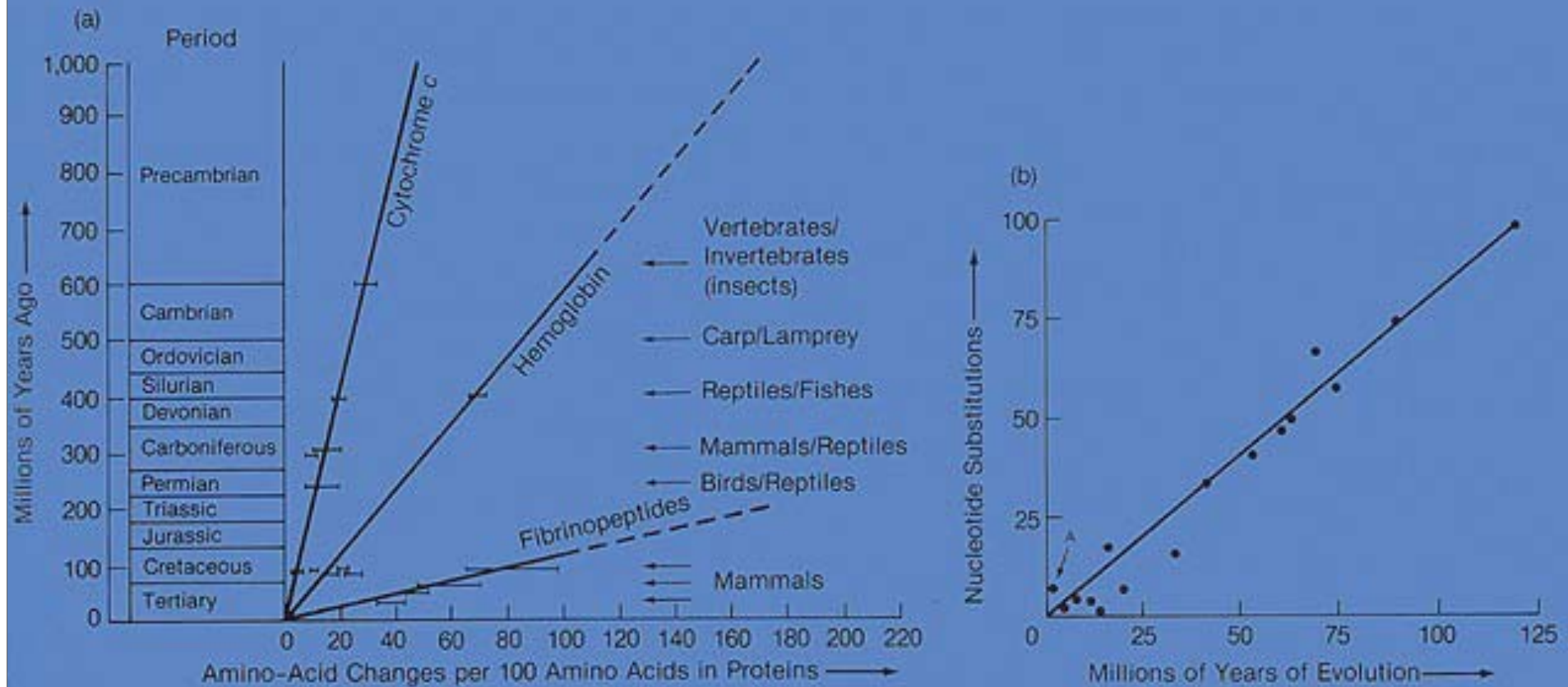
The Onion test from Ryan Gregory

The onion test is a simple reality check for anyone who thinks they have come up with a universal function for non-coding DNA. Whatever your proposed function, ask yourself this question: Can I explain why an onion needs about five times more non-coding DNA for this function than a human?

Selfish elements can increase C-value as long as they do not put undue load on replication and packaging of material.

Orgel LE, Crick FH: Selfish DNA: the ultimate parasite. Nature 1980

Observations that prompted development of Neutral Theory



Molecular clock for various proteins

Predictions of the theory

Pseudogenes are duplications of active genes (or insertions of cDNA from mRNA) that are not functional.

Pseudogenes should evolve neutrally; they should evolve faster than sequences constrained by selection, since there are more neutral mutations than beneficial ones.

Data: Pseudogene divergence among highest observed in nuclear genomes

Prediction: Different parts of genes are under different selective constraints, and should evolve at different rates

Several different types of sites need to be distinguished for protein encoding portions:

Non-degenerate sites -all possible changes at this site are nonsynonymous.

Two-fold degenerate—if one of the three possible changes is synonymous

Four-fold degenerate—if all possible changes are synonymous

Untranscribed and untranslated regions

Comparison of rates of substitution in different regions

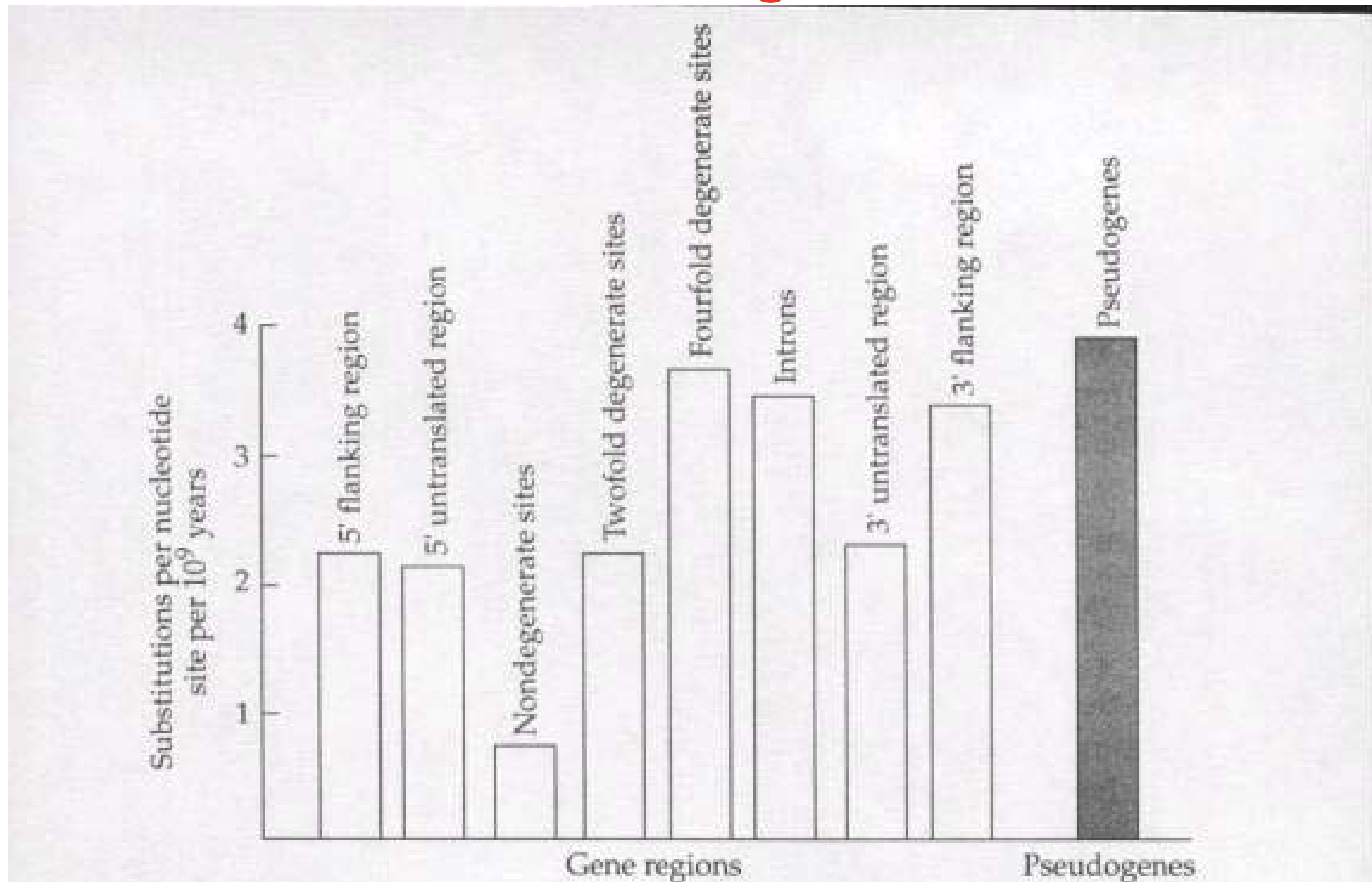
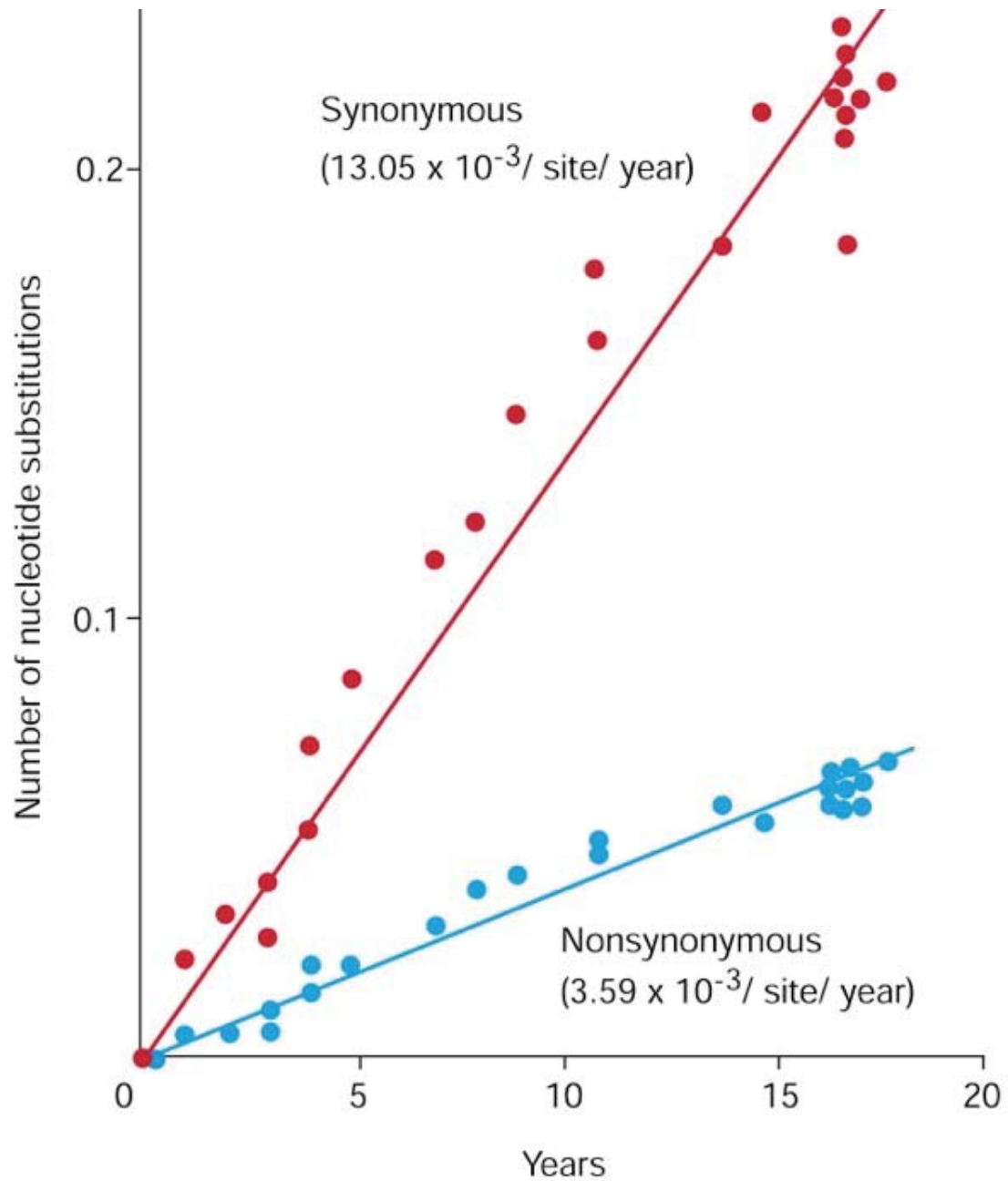


FIGURE 4.3 Average rates of substitution in different parts of genes (white) and in pseudogenes (gray). From Li (1997).

Prediction:
Synonymous
substitutions
accumulate in
genomes
faster than do
non-
synonymous
ones



Influenza virus evolution over 20 years

Other patterns

Different genes have different rates of non-synonymous evolution

Genes with most vital cellular functions change very little, even over long evolutionary time: e.g., histones, actins, insulin in mammals.

Interpretation?

Many moving parts, difficult to tolerate intermediate dis-advantageous states and coordinate changes

Prediction: highly constrained proteins should evolve slowly

Rates of nucleotide substitution (per site per billion years)

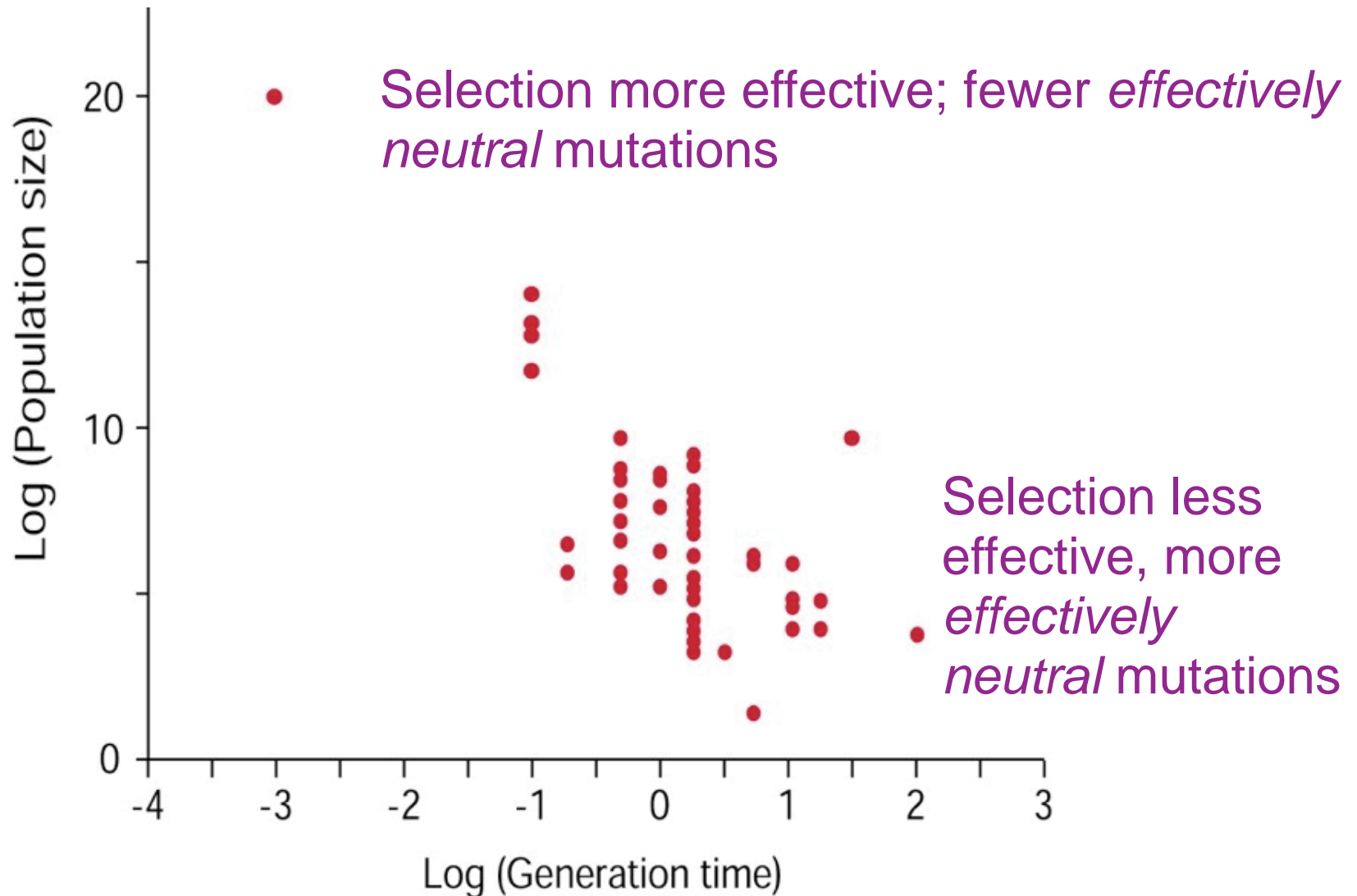
Gene	Non-synonymous rate	Synonymous rate
Histone H4	0.00	3.94
Histone H2	0.00	4.52
Actin a	0.01	3.68
Ribosomal protein S14	0.02	2.16
Insulin	0.13	4.02
α -globin	0.78	2.58
Myoglobin	0.57	4.10
β -Interferon	3.06	5.50
MHC (HLA-A)	13.30	3.5

Problems w/ Neutral Theory

The neutral mutation rate, ν should vary among species as a function of generation time; there should be *more changes per million years* for rodents than for primates.

However, some proteins evolve at constant rate among lineages without regard to generation time: same rate between mice/rats and chimp/human

Negative correlation between N_e and generation time.
Species w/ short generation time will have large N_e .



Ohta's Nearly Neutral Theory

If mutation rate to *effectively neutral* mutations varies with generation time, then generation time differences and mutation rate differences cancel each other ==>

Molecular evolution proportional to absolute time, rather than generation time.

Current status of Neutral and Nearly Neutral Theories

Probably correct for some fraction of the genome

Big Questions Now

How much of the genome evolves neutrally and how much is under selection?

How much non-neutral evolution is due to positive selection and how much is due to purifying or diversifying selection?

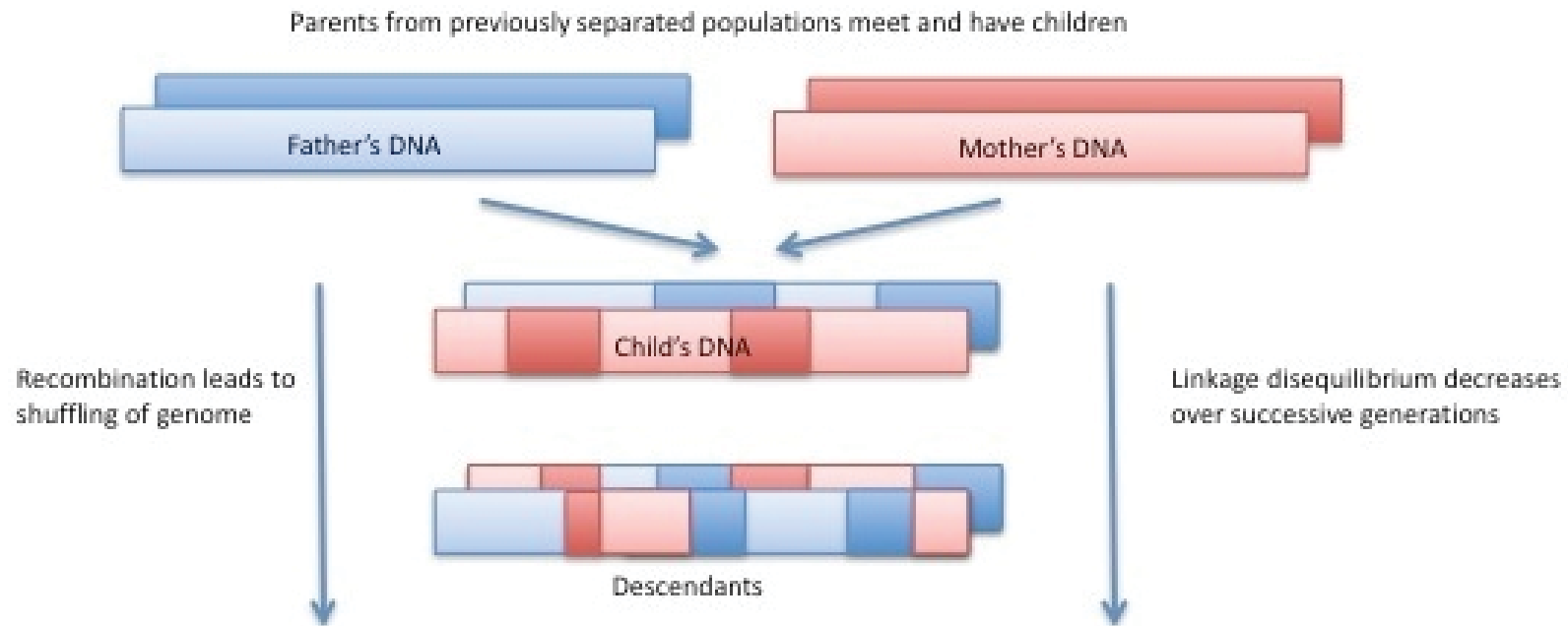
Neutral Theory Provides *Null Model* for Tests of Selection

If even replacement substitutions are effectively neutral, then the rate of replacement substitution should equal the rate of synonymous substitution.

Sequence a coding region in two different species.

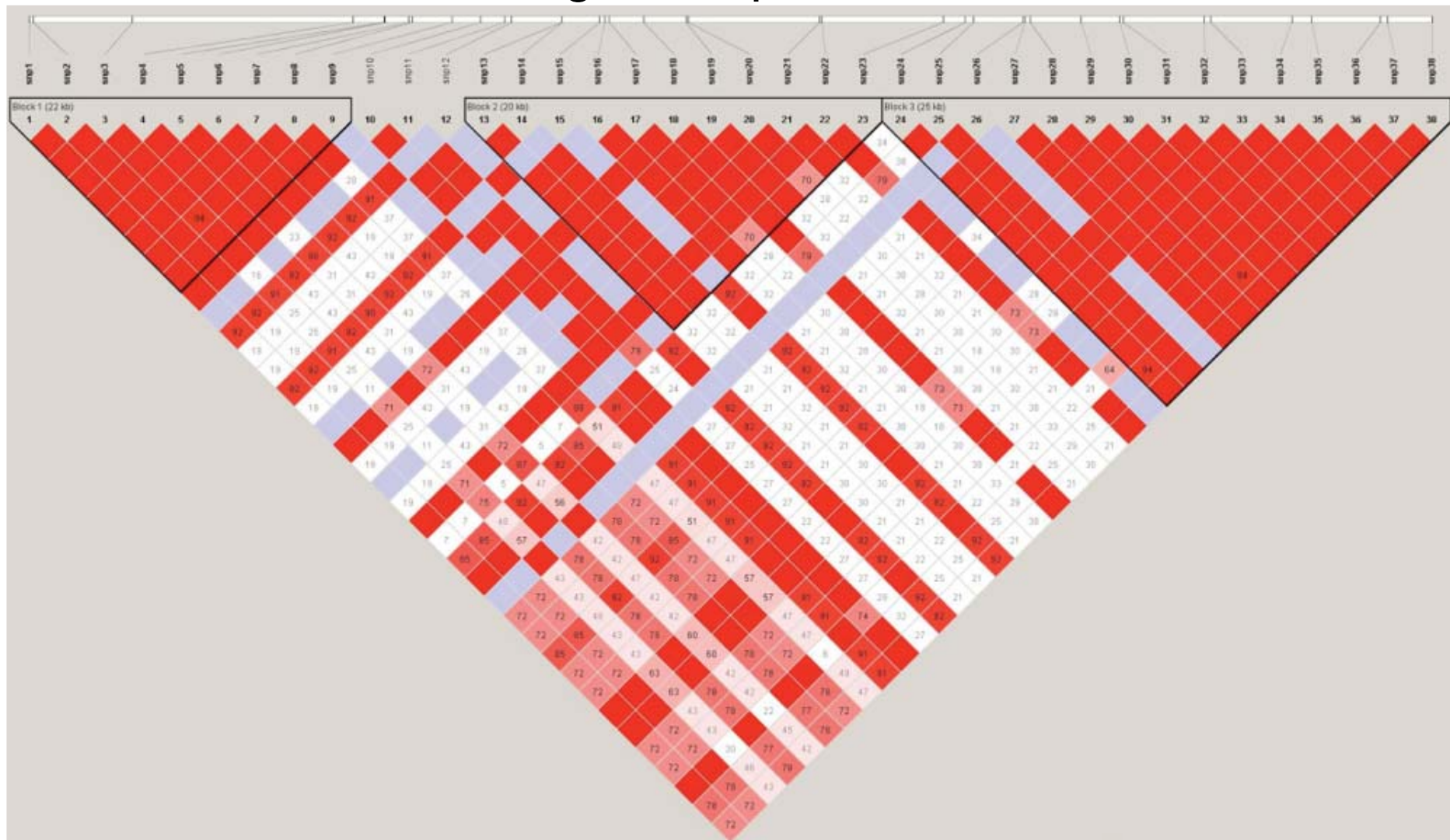
Compare the rate of synonymous substitution to the rate of replacement substitution.

Recombinations across generations



Haplotypes are blocks (regions) that have not “broken” apart and show the original order of nucleotides/markers

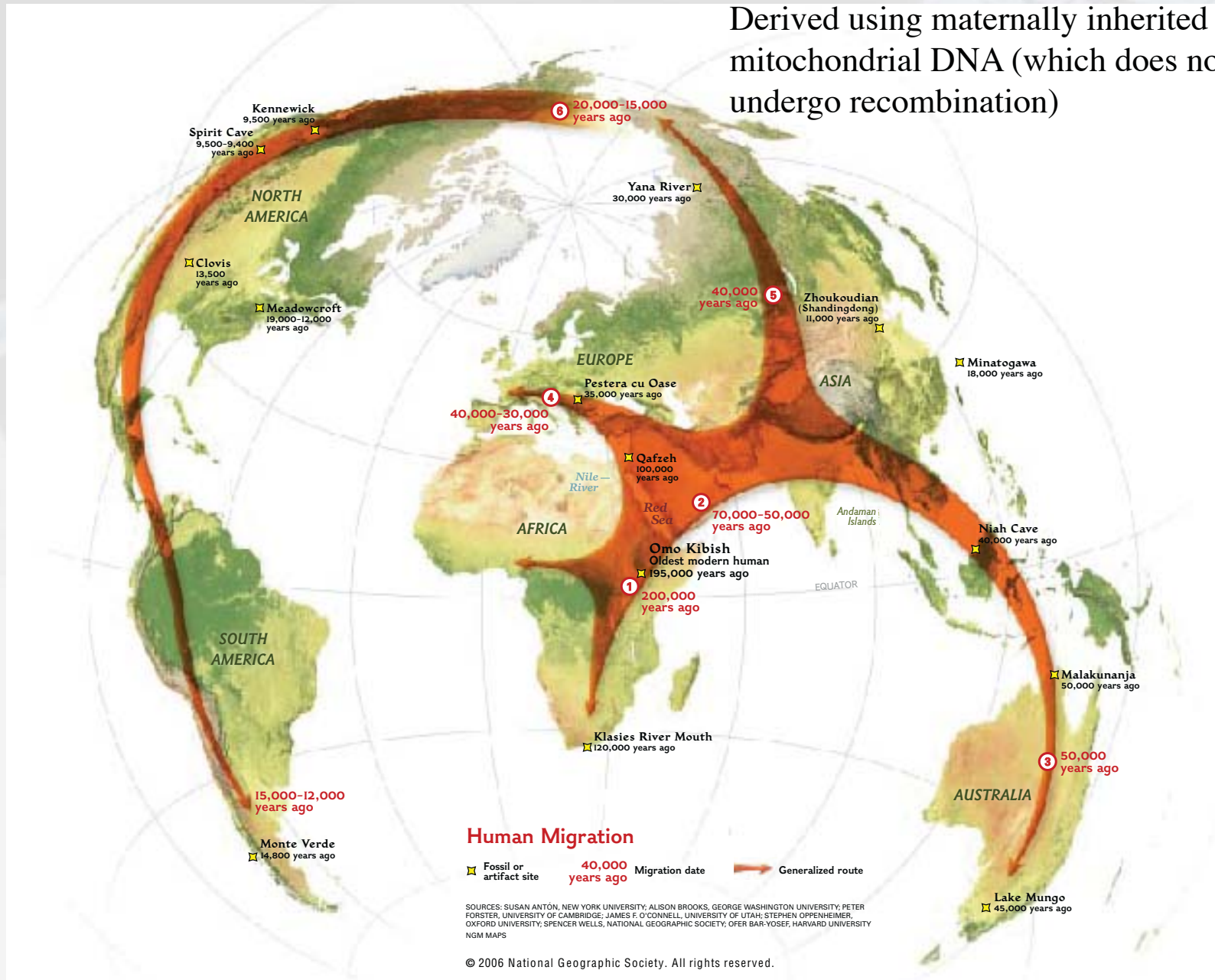
Linkage disequilibrium



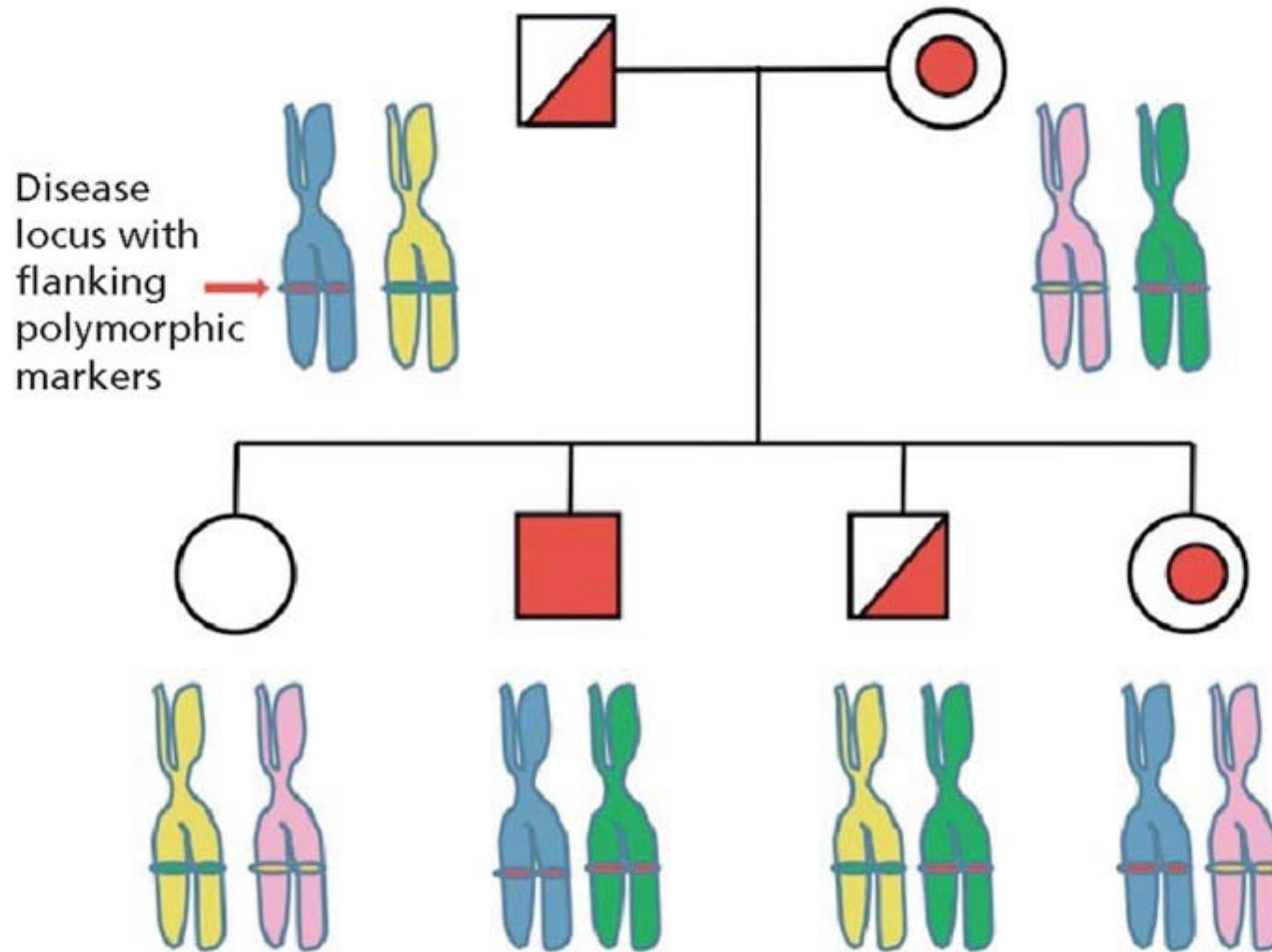
Regions that show deviations from predictions of Mendel's second law, of independent segregation. The deviations are measured using linkage disequilibrium, essentially a measure of correlation between points on the genome.

Human Migration

Derived using maternally inherited mitochondrial DNA (which does not undergo recombination)

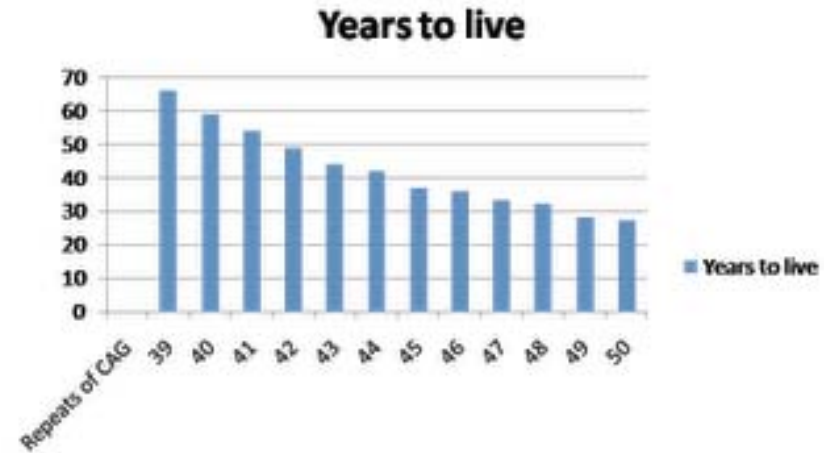
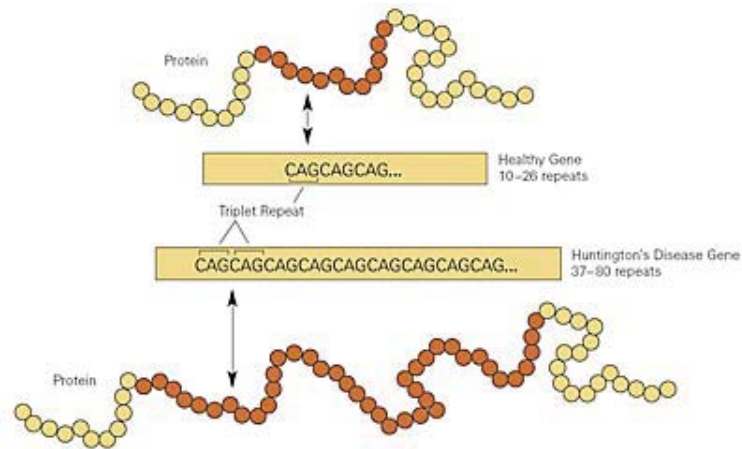


Linkage analysis



Strategy for using affected families to identify markers and regions of DNA linked to disorder

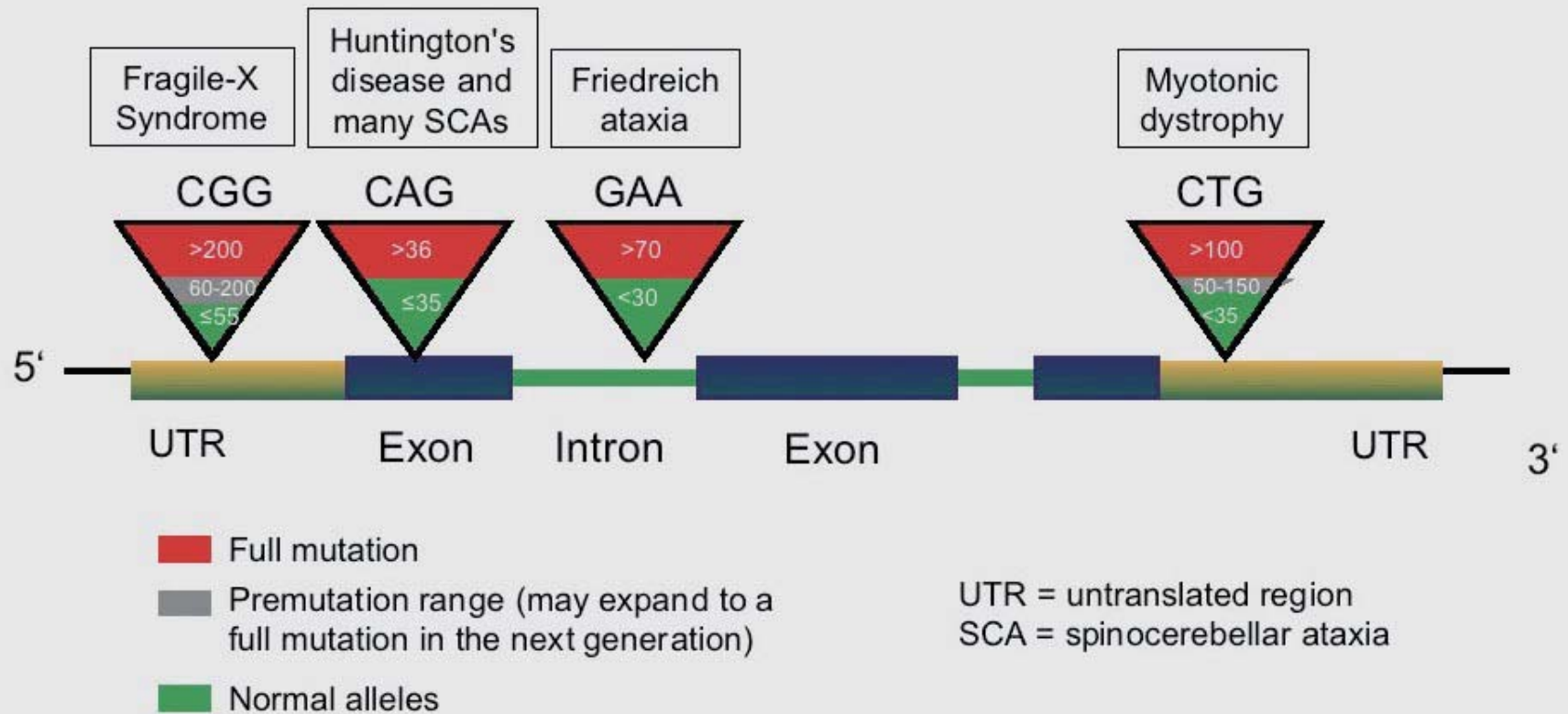
Huntington's disease



- HTT (Huntingtin gene, HD or IT15, on 4p16.3)
- Trinucleotide repeat in CDS (CAG -> Q) polyQ tract
- Genetically dominant and penetrant
- Imprinted ? severity can be influenced by sex of affected parent

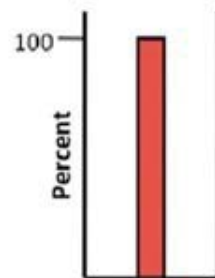
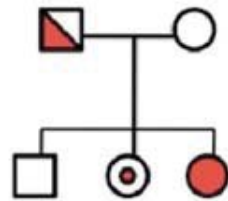
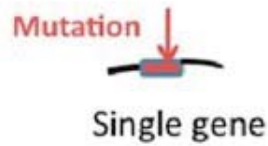
Classic disease whose cause was uncovered using linkage analysis (read book by Nancy Wexler for a fascinating personal account of this work)

Trinucleotide repeat diseases

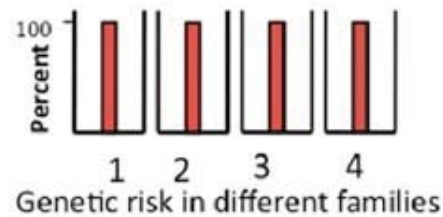


Monogenic vs complex disease

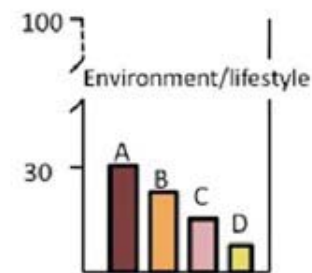
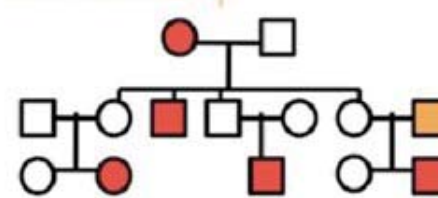
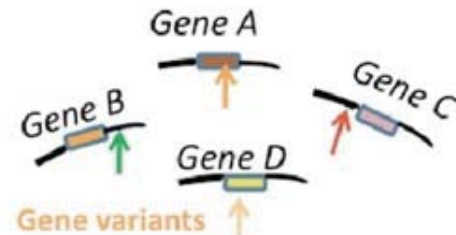
Monogenic disorder



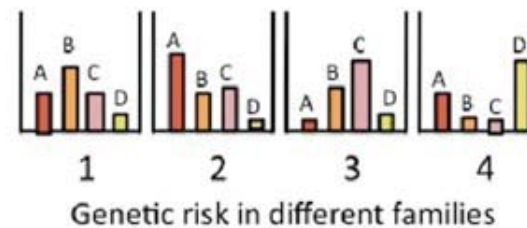
Impact of mutation on disease phenotype



Complex disorder



Influence of variations in different genes on disease phenotype



GWAS

Genome wide association studies. Study large populations of affected versus unaffecteds to pick out mutations/variations that might be associated with diseases.
Assumes common causes for common diseases

GWAS criticisms

No Mechanisms.

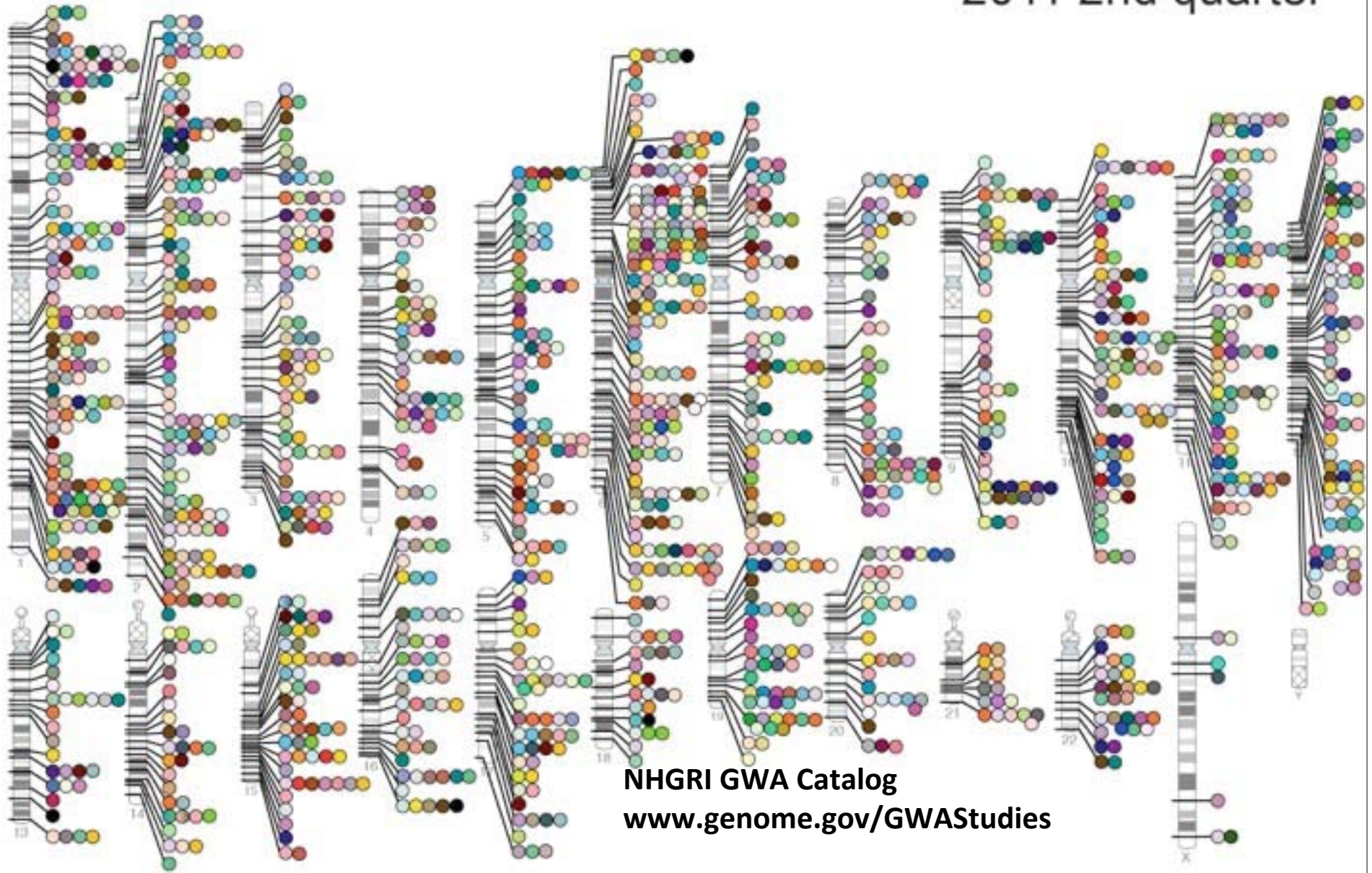
“linear models”

Complexity is not really addressed

Ascertainment bias

Published Genome-Wide Associations through 06/2011,
1,449 published GWA at $p \leq 5 \times 10^{-8}$ for 237 traits

2011 2nd quarter



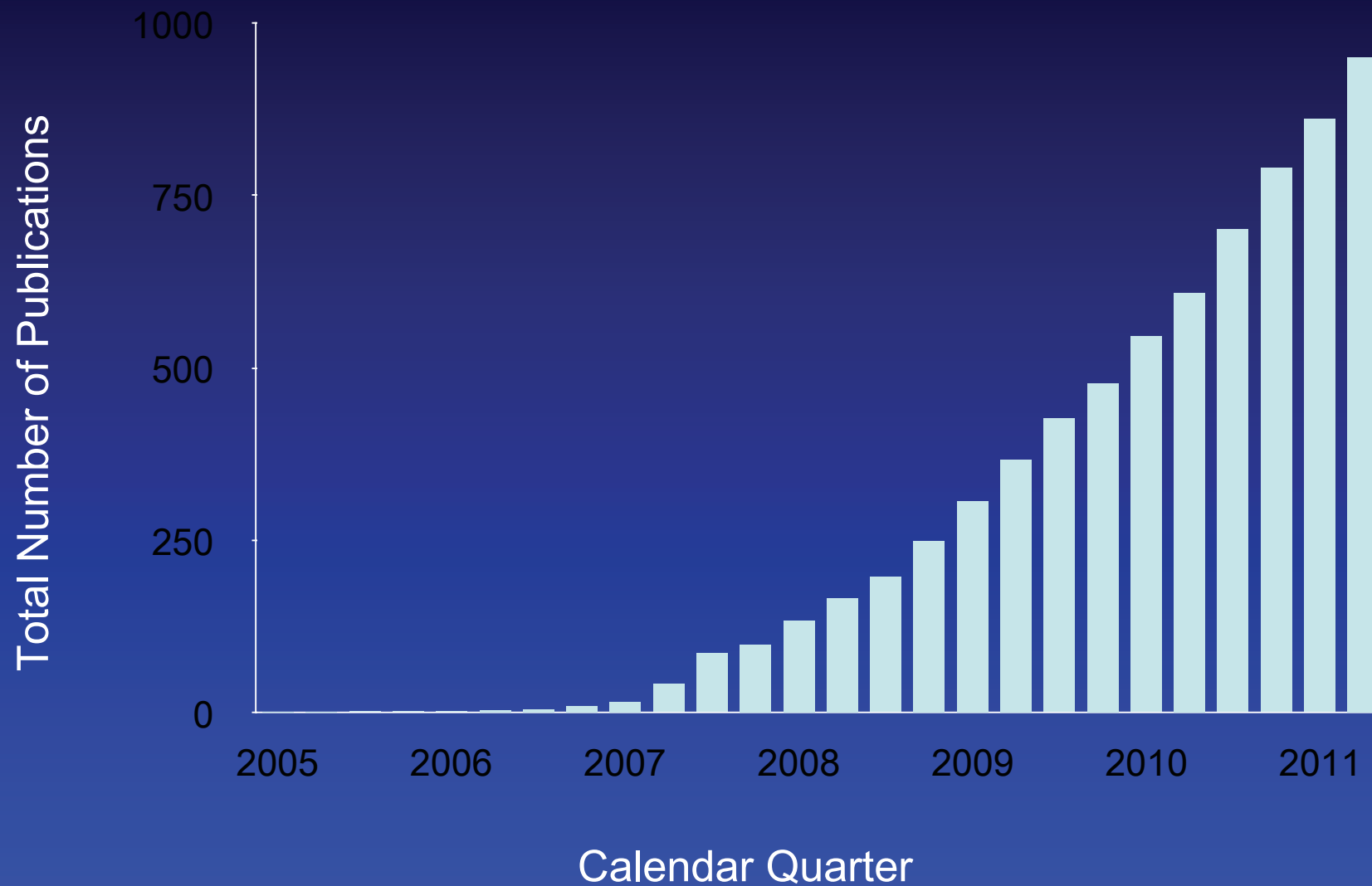
NHGRI GWA Catalog
www.genome.gov/GWASStudies

Disorders with causes uncovered by GWAS studies

- Abdominal aortic aneurysm
- Acute lymphoblastic leukemia
- Adhesion molecules
- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alcohol dependence
- Alopecia areata
- Alzheimer disease
- Amyloid A levels
- Amyotrophic lateral sclerosis
- Angiotensin-converting enzyme activity
- Ankylosing spondylitis
- Arterial stiffness
- Asparagus anosmia
- Asthma
- Atherosclerosis in HIV
- Atrial fibrillation
- Attention deficit hyperactivity disorder
- Autism
- Basal cell cancer
- Behcet's disease
- Bipolar disorder
- Biliary atresia
- Bilirubin
- Bitter taste response
- Birth weight
- Bladder cancer
- Bleomycin sensitivity
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- C-reactive protein
- Calcium levels
- Cardiac structure/function
- Cardiovascular risk factors
- Camitine levels
- Carotenoid/tocopherol levels
- Celiac disease
- Celiac disease and rheumatoid arthritis
- Cerebral atrophy measures
- Chronic lymphocytic leukemia
- Chronic myeloid leukemia
- Cleft lip/palate
- Coffee consumption
- Cognitive function
- Conduct disorder
- Colorectal cancer
- Corneal thickness
- Coronary disease
- Creutzfeldt-Jakob disease
- Crohn's disease
- Crohn's disease and celiac disease
- Cutaneous nevi
- Cystic fibrosis severity
- Dermatitis
- DHEA-s levels
- Diabetic retinopathy
- Dilated cardiomyopathy
- Drug-induced liver injury
- Drug-induced liver injury (antipsychotics)
- Endometrial cancer
- Endometriosis
- Eosinophil count
- Eosinophilic esophagitis
- Erectile dysfunction and prostate cancer treatment
- Erythrocyte parameters
- Esophageal cancer
- Essential tremor
- Exfoliation glaucoma
- Eye color traits
- F cell distribution
- Fibrinogen levels
- Folate pathway vitamins
- Follicular lymphoma
- Fuch's corneal dystrophy
- Freckles and burning
- Gallstones
- Gastric cancer
- Glioma
- Glycemic traits
- Hair color
- Hair morphology
- Handedness in dyslexia
- HDL cholesterol
- Heart failure
- Hear1 rate
- Height
- Hemostasis parameters
- Hepatic steatosis
- Hepatitis
- Hepatocellular carcinoma
- Hirschsprung's disease
- HIV-1 control
- Hodgkin's lymphoma
- Homocysteine levels
- Hypospadias
- Idiopathic pulmonary fibrosis
- IFN-related cytopeni
- IgA levels
- IgE levels
- Inflammatory bowel disease
- Insulin-like growth factors
- Intracranial aneurysm
- Iris color
- Iron status markers
- Ischemic stroke
- Juvenile idiopathic arthritis
- Keloid
- Kidney stones
- LDL cholesterol
- Leprosy
- Leptin receptor levels
- Liver enzymes
- Longevity
- LP (a) levels
- LpPLA(2) activity and mass
- Lung cancer
- Magnesium levels
- Major mood disorders
- Malaria
- Male pattern baldness
- Mammographic density
- Matrix metalloproteinase levels
- MCP-1
- Melanoma
- Menarche & menopause
- Meningococcal disease
- Metabolic syndrome
- Migraine
- Moyamoya disease
- Multiple sclerosis
- Myeloproliferative neoplasms
- Myopia (pathological)
- N-glycan levels
- Narcolepsy
- Nasopharyngeal cancer
- Natriuretic peptide levels
- Neuroblastoma
- Nicotine dependence
- Obesity
- Open angle glaucoma
- Open personality
- Optic disc parameters
- Osteoarthritis
- Osteoporosis
- Otosclerosis
- Other metabolic traits
- Ovarian cancer
- Pancreatic cancer
- Pain
- Paget's disease
- Panic disorder
- Parkinson's disease
- Periodontitis
- Peripheral arterial disease
- Personality dimensions
- Phosphatidylcholine levels
- Phosphorus levels
- Photic sneeze
- Phytosterol levels
- Platelet count
- Polycystic ovary syndrome
- Primary biliary cirrhosis
- Primary sclerosing cholangitis
- PR interval
- Progranulin levels
- Progressive supranuclear palsy
- Prostate cancer
- Protein levels
- PSA levels
- Psoriasis
- Psoriatic arthritis
- Pulmonary funct. COPD
- QRS interval
- QT interval
- Quantitative traits
- Recombination rate
- Red vs non-red hair
- Refractive error
- Renal cell carcinoma
- Renal function
- Response to antidepressants
- Response to antipsychotic therapy
- Response to carbamazepine
- Response to clopidogrel therapy
- Response to hepatitis C treat
- Response to interferon beta therapy
- Response to metformin
- Response to statin therapy
- Restless legs syndrome
- Retinal vascular caliber
- Rheumatoid arthritis
- Ribavirin-induced anemia
- Schizophrenia
- Serum metabolites
- Skin pigmentation
- Smoking behavior
- Speech perception
- Sphingolipid levels
- Statin-induced myopathy
- Stroke
- Sudden cardiac arrest
- Suicide attempts
- Systemic lupus erythematosus
- Systemic sclerosis
- T-tau levels
- Tau AB1-42 levels
- Telomere length
- Testicular germ cell tumor
- Thyroid cancer
- Thyroid volume
- Tooth development
- Total cholesterol
- Triglycerides
- Tuberculosis
- Type 1 diabetes
- Type 2 diabetes
- Ulcerative colitis
- Urate
- Urinary albumin excretion
- Urinary metabolites
- Uterine fibroids
- Venous thromboembolism
- Ventricular conduction
- Vertical cup-disc ratio
- Vitamin B12 levels
- Vitamin D insufficiency
- Vitiligo
- Warfarin dose
- Weight
- White cell count
- White matter hyperintensity
- YKL-40 levels

Published GWA Reports, 2005 – 6/2011

951

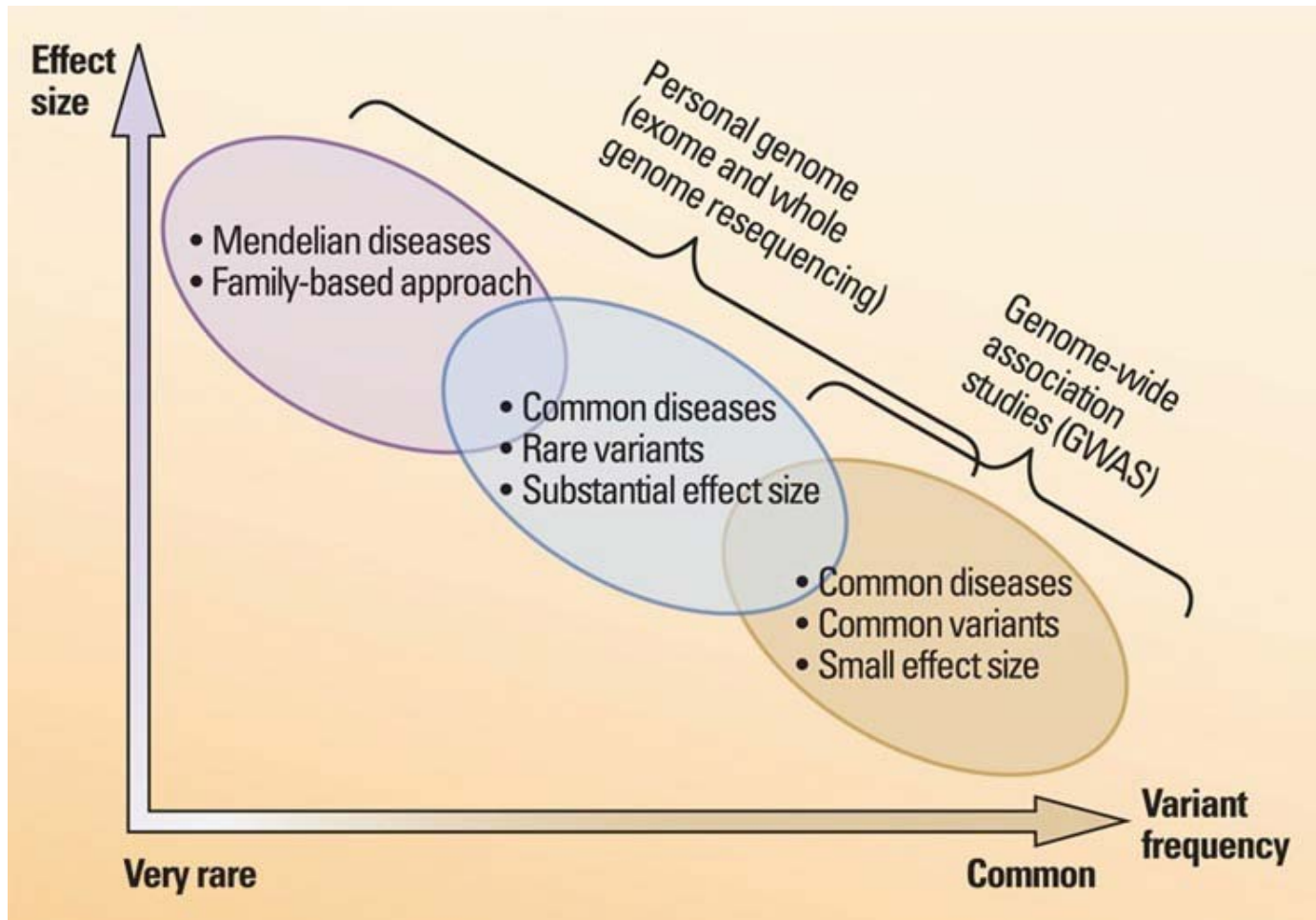


Some Examples of effects on personalized medicine

- Warfarin: oral anticoagulant, greater than 10-fold variability in therapeutic dose. Two genes, the warfarin metabolic enzyme CYP2C9 and warfarin target enzyme, vitamin K epoxide reductase complex 1 VKORC1, influence warfarin maintenance dose.
 - CYP2C9*2, CYP2C9*3 haplotypes decrease enzyme activity, decreased warfarin dose
 - VKORC1 snps are associated with dose variance.
 - VKORC1 has an approximately three-fold greater effect than CYP2C9.
- H pylori treatment: multidrug regimen that contains multiple antibiotics plus proton pump inhibitors (PPIs) or histamine-2 receptor blockers for eradication.
 - CYP2C19 snps can defined the rapid (RM), intermediate (IM) or poor (PM) metabolizers; Asian populations demonstrate a higher frequency of PMs, higher PPI efficacy, and may have better H pylori eradication at standard doses.
- MTHFR (methylene-tetra-hydrofolate) SNPs can increase toxicity of methotrexate used to treat Crohn's disease.



Hole in the middle. New studies are failing to find many rare genetic variants with large influences on common diseases (blue).



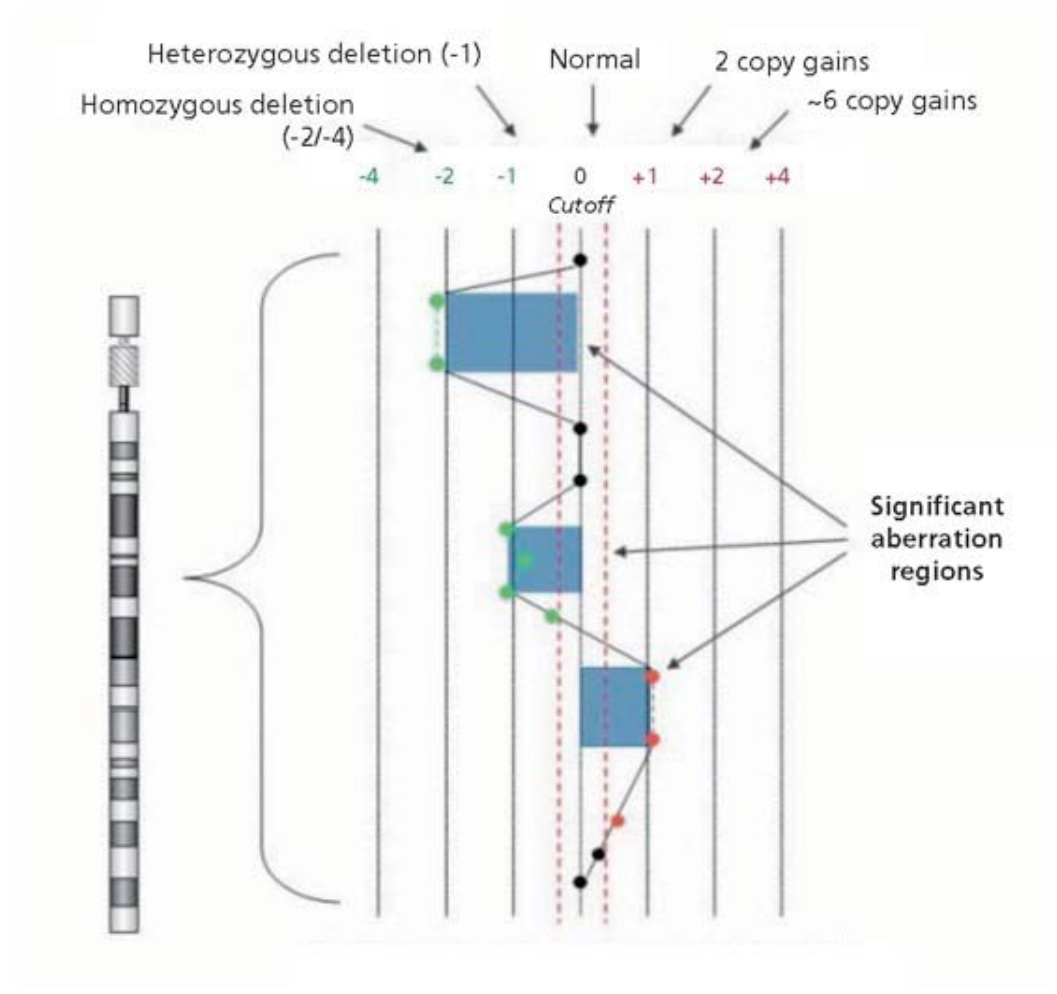
J Kaiser Science 2012;338:1016-1017



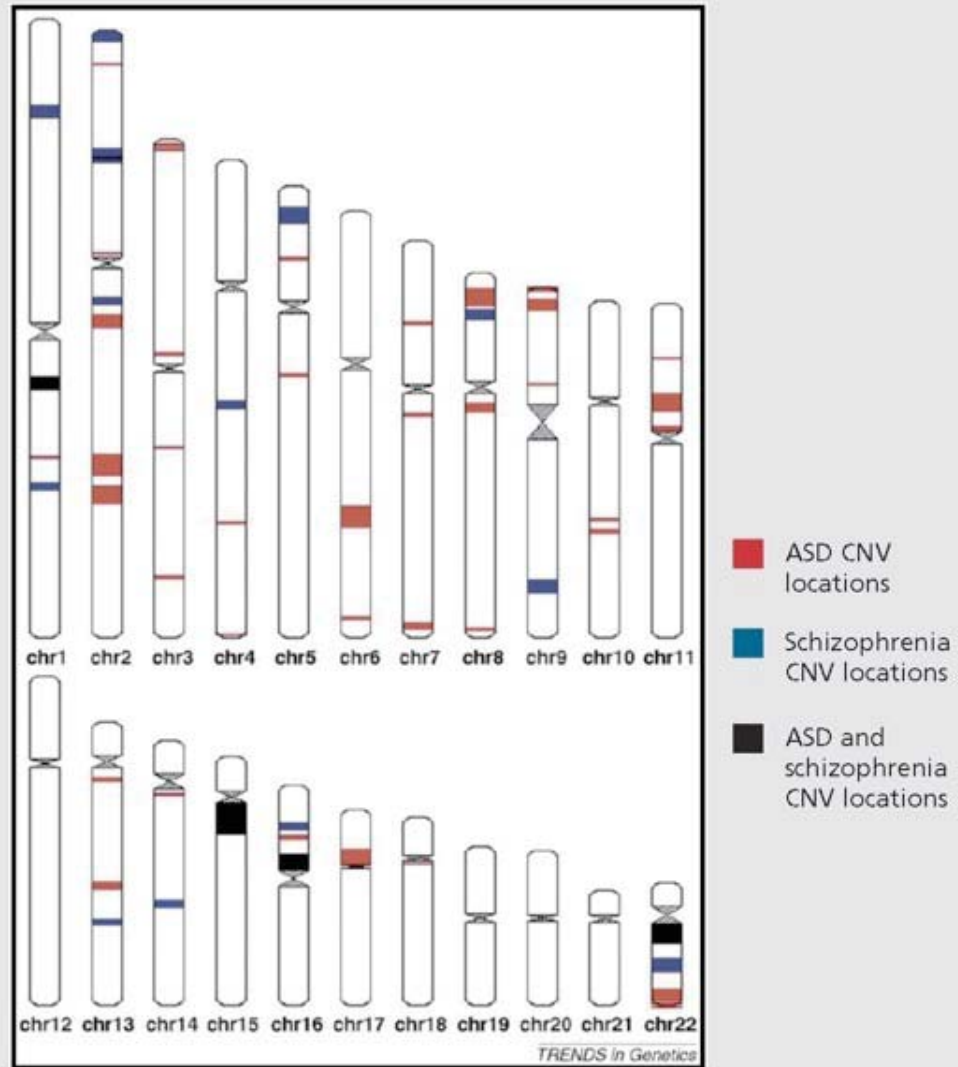
New studies are concentrating on rare variants responsible for common diseases (the category in the middle) and not finding much by way of results.

Copy number variation (CNV) is another target of studies on causes of diseases.

Copy number variation (CNV)



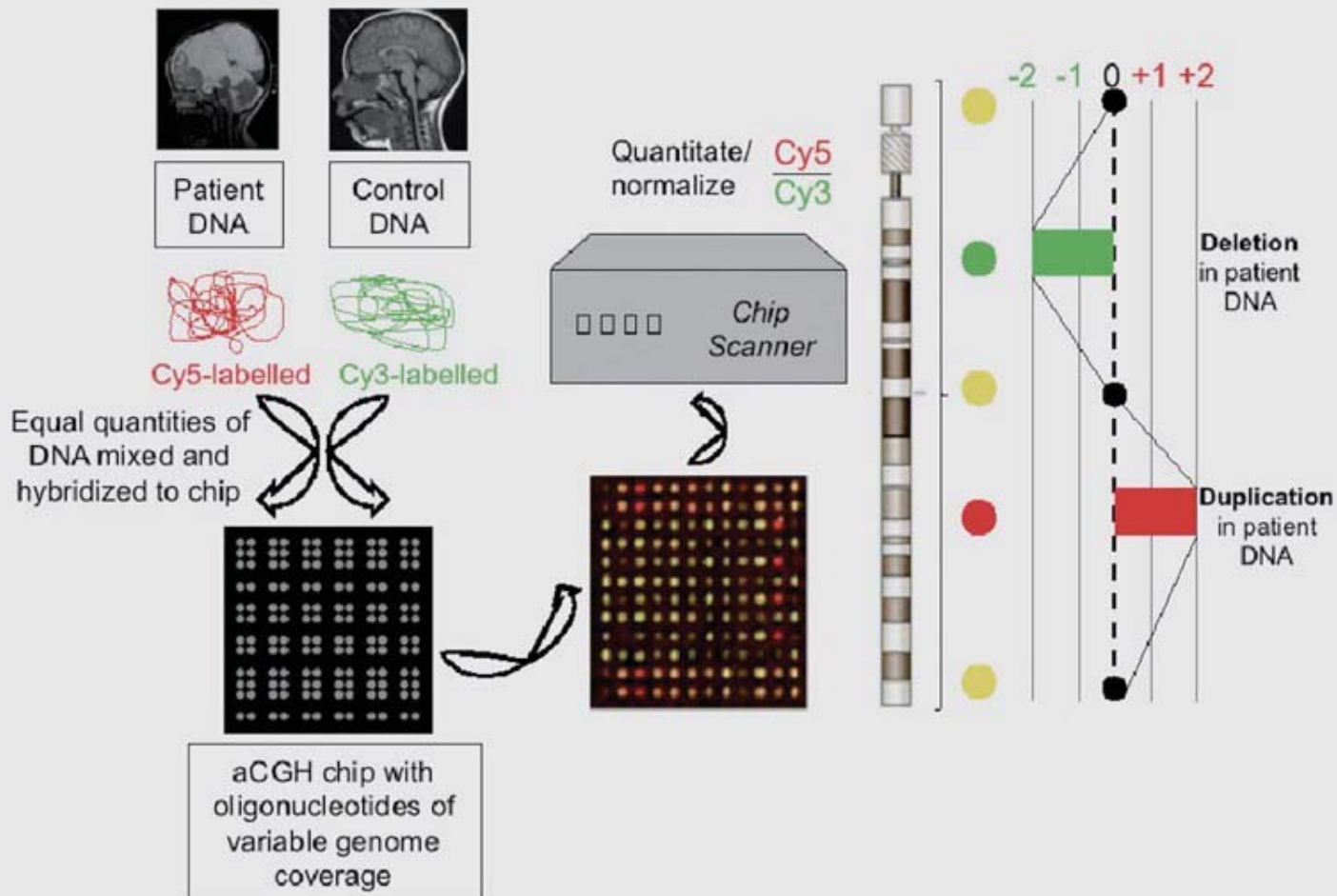
CNV association findings in schizophrenia and autism spectrum disorder (ASD)



Adapted from: Merkangas AK, Corvin AP, Gallagher L. Copy number variants in neurodevelopmental disorders: promises and challenges. *Trends Genet.* 2009;25:536-544. Copyright © Elsevier 2009

CNVs are best detected using array-CGH (comparative genome hybridization)

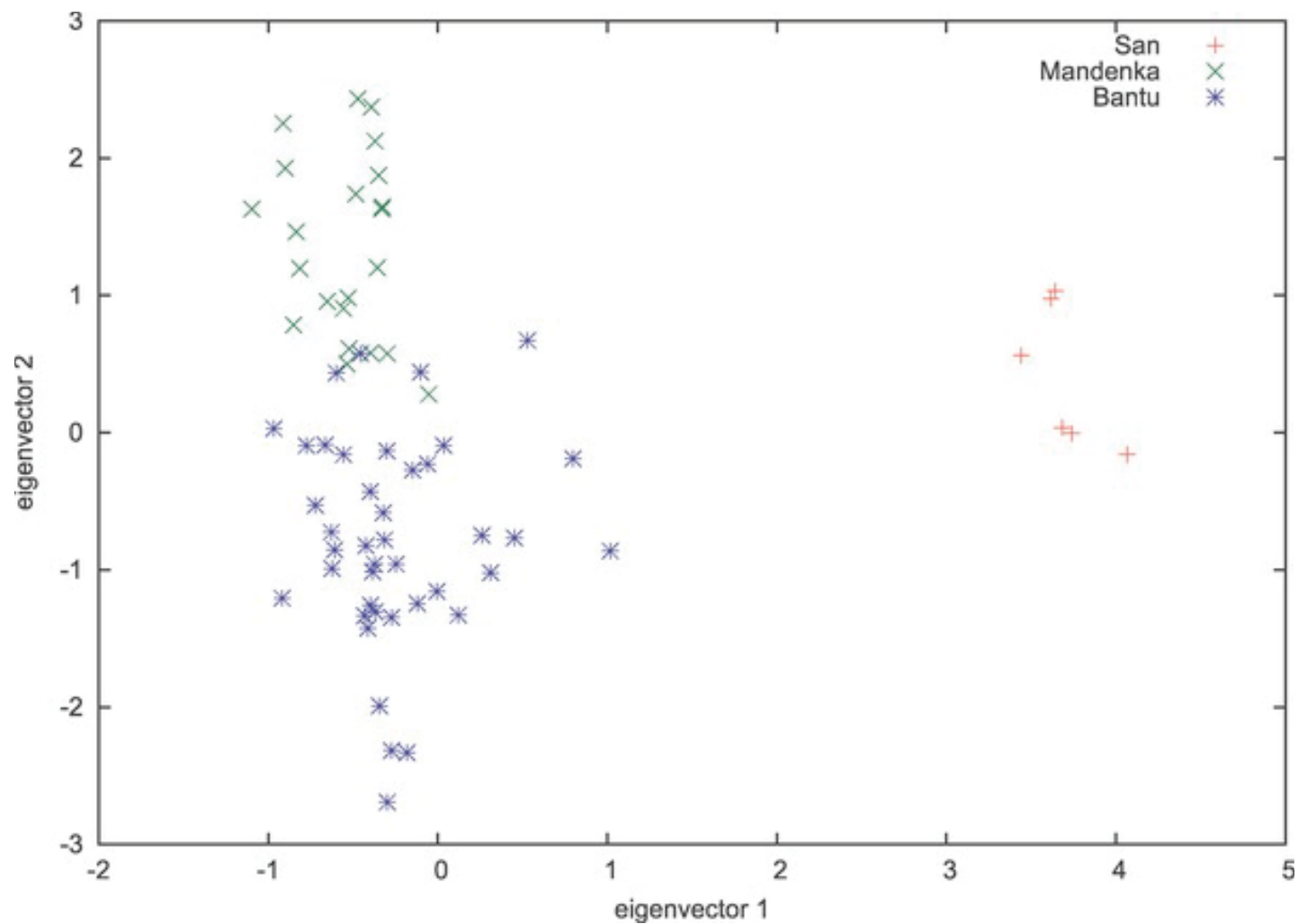
Principle of array comparative genome hybridization (aCGH)



Using SNPs for population stratification

- Bi-allelic SNPs (three values 0, 1 or 2)
- $C(i,j)$ allele for marker j , individual i
- M is obtained from $C(i,j)$ by subtracting mean over individuals and normalized by variance in allele frequency
- SVD on M , to generate eigenvectors (PCA), approach pioneered by Cavalli-Sforza

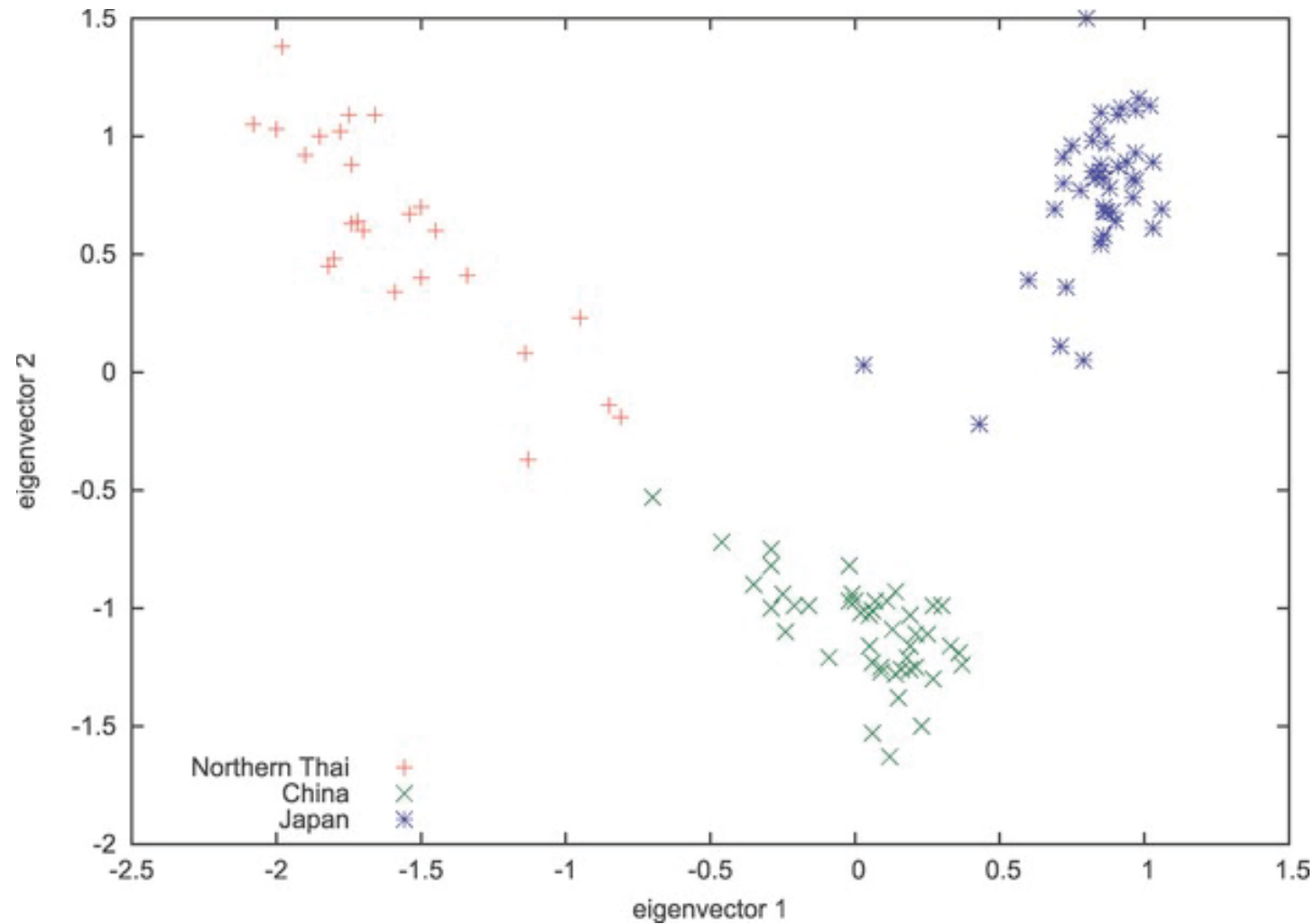
Three African Populations



Plots of the first two eigenvectors for some African populations in the CEPH–HGDP dataset [30]. Yoruba and Bantu-speaking populations are genetically quite close and were grouped together. The Mandenka are a West African group speaking a language in the Mande family [15, p. 182]. The eigenanalysis fails to find structure in the Bantu populations, but separation between the Bantu and Mandenka with the second eigenvector is apparent.

doi: 10.1371/journal.pgen.0020190.g004

Three East Asian Populations

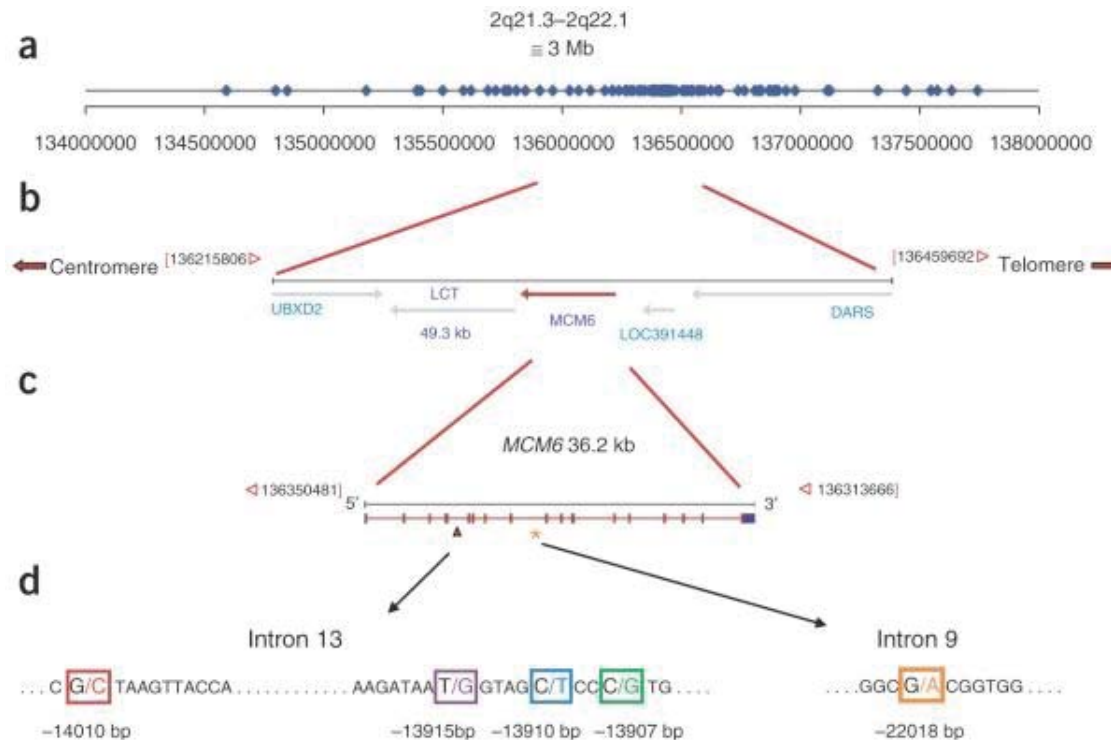


Plots of the first two eigenvectors for a population from Thailand and Chinese and Japanese populations from the International Haplotype Map [32]. The Japanese population is clearly distinguished (though not by either eigenvector separately). The large dispersal of the Thai population, along a line where the Chinese are at an extreme, suggests some gene flow of a Chinese-related population into Thailand. Note the similarity to the simulated data of [Figure 8](#).

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. *PLoS Genet* 2(12): e190. doi:10.1371/journal.pgen.0020190
<http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020190>

Is your genome controlled by cows ?

Lactose tolerance

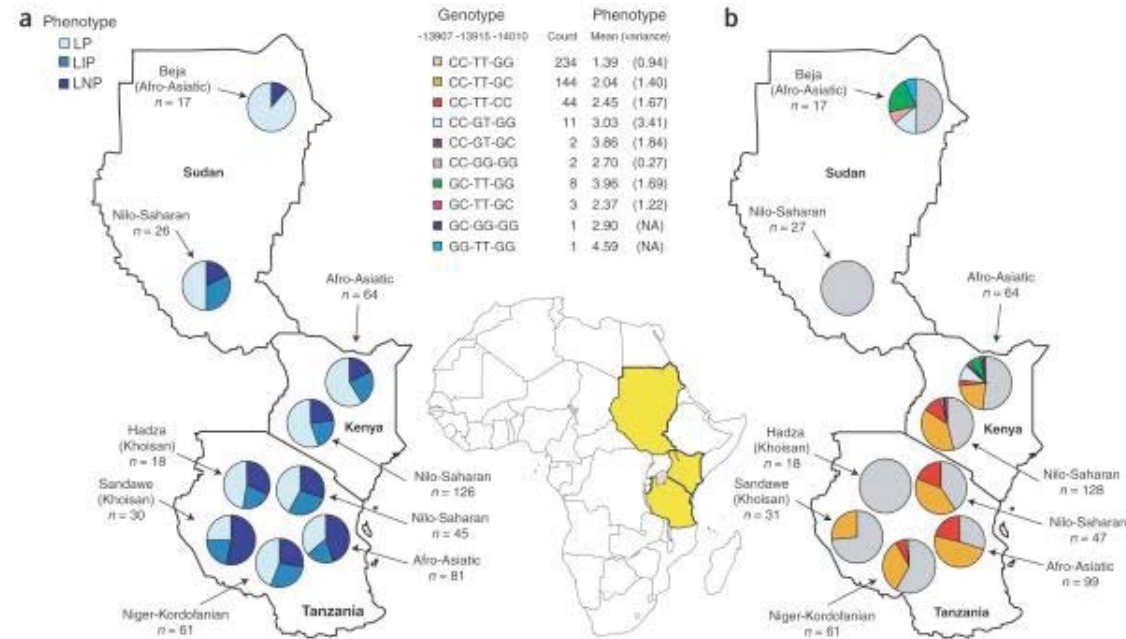


- enzyme lactase is encoded by the gene *LCT*
- two phenotypes "lactase persistent" and "lactase non-persistent (hypolactasia)"
- lactase persistence is a dominant trait (single copy is sufficient)

Map of the *LCT* and *MCM6* gene region and location of genotyped SNPs. (a) Distribution of 123 SNPs included in genotype analysis. (b) Map of the *LCT* and *MCM6* gene region. (c) Map of the *MCM6* gene. (d) Location of lactase persistence-associated SNPs within introns 9 and 13 of the *MCM6* gene in African and European populations.

Lactose tolerance in Africa

Convergent evolution. Lactose tolerance arises from different mutations in African populations, compared to the European mutation.



Map of phenotype and genotype proportions for each population group considered in this study. (a) Pie charts representing the proportion of each phenotype by geographic region. LP indicates lactase persistence, LIP indicates lactase intermediate persistence and LNP indicates lactase non-persistence. Phenotypes were binned using an LTT test according to the rise in blood glucose after digestion of 50 g lactose: lactase persistence, >1.7 mM; LIP, between 1.1 mM and 1.7 mM; LNP, <1.1 mM. (b) Proportion of compound genotypes for G/C-13907, T/G-13915 and C/G-14010 in each region. The pie charts are in the approximate geographic location of the sampled individuals.

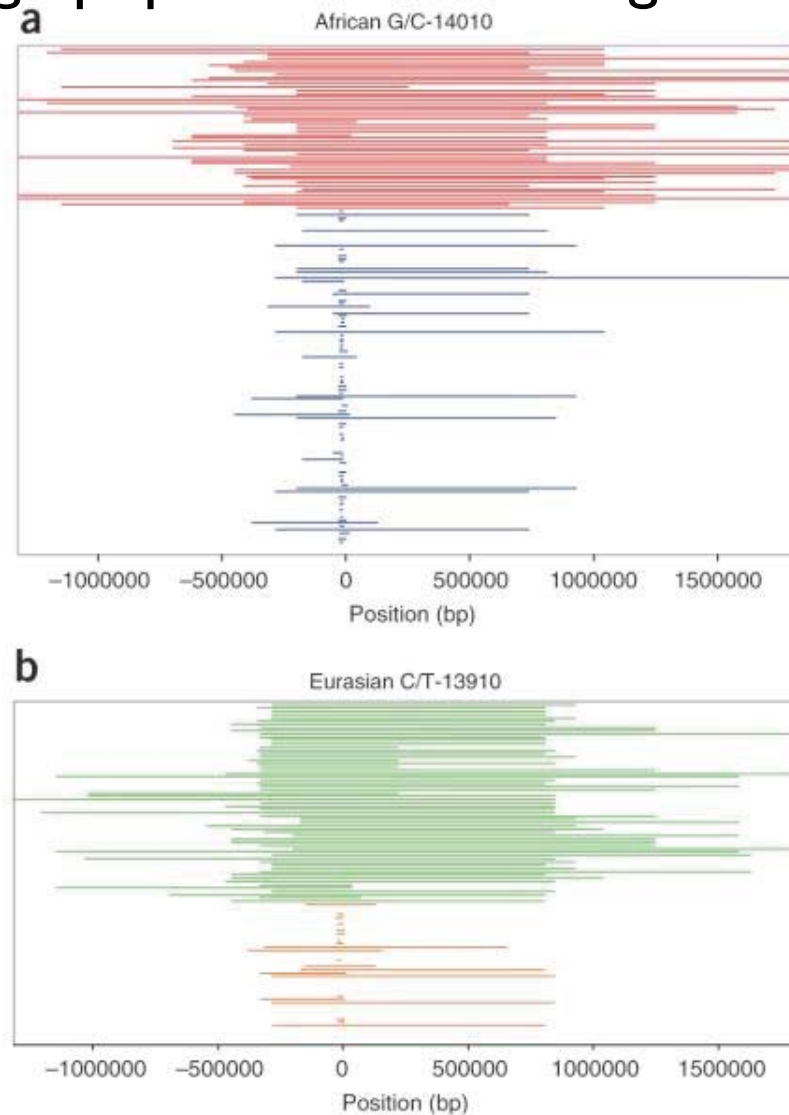
[Nat Genet. Author manuscript; available in PMC 2009 April 23.](#)

Published in final edited form as:

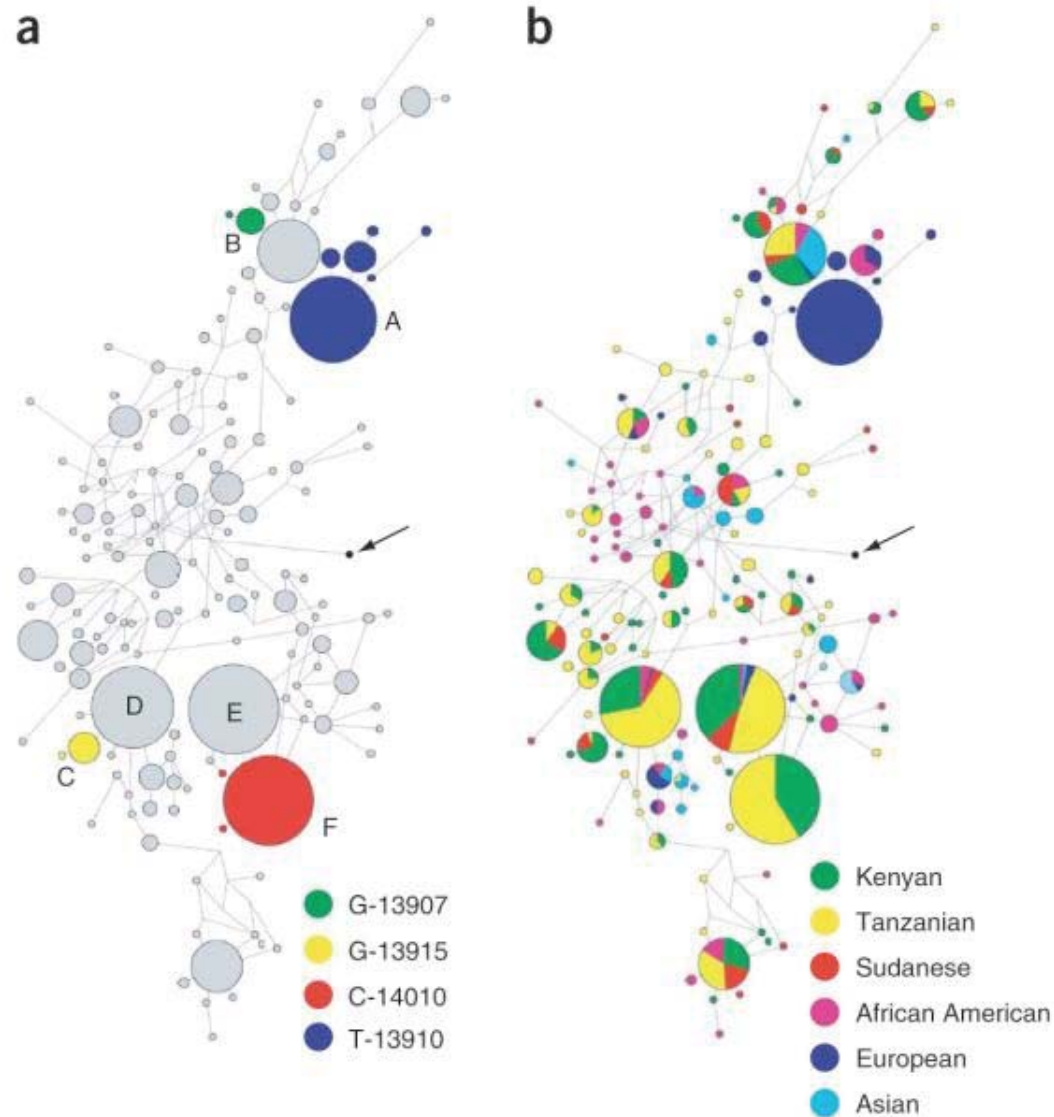
Nat Genet. 2007 January; 39(1): 31–40.

Published online 2006 December 10. doi: 10.1038/ng1946

Lactose tolerance evolved multiple times and swept through populations when it gave an advantage



Comparison of tracts of homozygous genotypes flanking the lactase persistence–associated SNPs. (a) Kenyan and Tanzanian C-14010 lactase-persistent (red) and non-persistent G-14010 (blue) homozygosity tracts. (b) European and Asian T-13910 lactase-persistent (green) and C-13910 non-persistent (orange) homozygosity tracts, based on the data from ref. 14. Positions are relative to the start codon of *LCT*. Note that some tracks are too short to be visible as plotted



Haplotype networks consisting of 55 SNPs spanning a 98-kb region encompassing *LCT* and *MCM6*. (a) Distribution of the lactase persistence–associate haplotypes. Haplotypes with a T allele at -13910 are indicated in blue, those with a G allele at -13907 in green, those with a C allele at -14010 in red and those with a G allele at -13915 in yellow. The arrow points to the inferred ancestral-state haplotype. (b) Network analysis of *LCT* and *MCM6* haplotypes indicating frequencies in the current data set and in Europeans, Asians and African Americans previously genotyped in ref. 14.

Cattle diversity versus Lactose tolerance

- Cattle genome diversity in Europe is highest where cattle farming arose.
- Diversity drops away from source, and lactose tolerance in humans wanes with diversity in cattle.

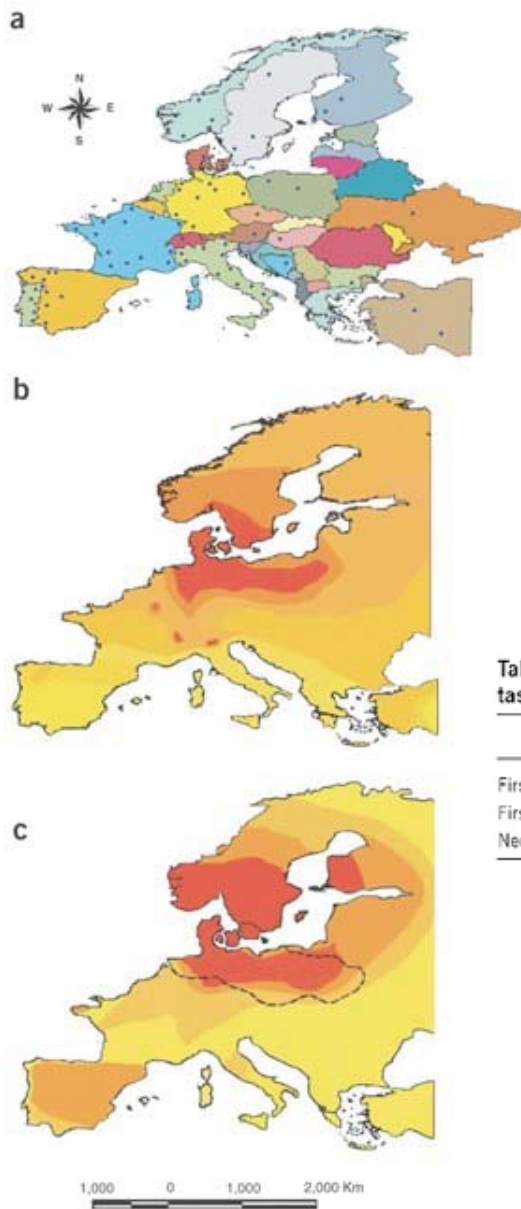


Table 1 Spearman correlation coefficient values between the first principal component from the milk protein gene frequencies, the lactase persistence allele frequency and the presence or absence of archeological evidence for Neolithic cattle pastoralists

	Spearman correlation	Degrees of freedom	P
First principal component versus neolithic cattle pastoralists	-0.750	21.2-27.3	<0.0005
First principal component versus lactase persistence allele frequency	-0.593	17.7-24.6	<0.01
Neolithic cattle pastoralists versus lactase persistence allele frequency	0.730	19.3-24.8	<0.0005

Figure 1. Geographic coincidence between milk gene diversity in cattle, lactose tolerance in humans and locations of Neolithic cattle farming sites in NCE.

(a) Geographic distribution of the 70 cattle breeds (blue dots) sampled across Europe and Turkey. (b) Synthetic map showing the first principal component resulting from the allele frequencies at the cattle genes. The dark orange color shows that the greatest milk gene uniqueness and allelic diversity occurs in cattle from NCE. (c) Geographic distribution of the lactase persistence allele in contemporary Europeans. The darker the orange color, the higher is the frequency of the lactase persistence allele. The dashed black line indicates the limits of the geographic distribution of early Neolithic cattle pastoralist (Funnel Beaker Culture) inferred from archaeological data¹⁵.

What does this have to do with India ?

- 1981 paper surveyed lactose tolerance (75% in north, 35% in south) which was taken as evidence for Aryan immigration
- Indian LT arises from same European T/C allele, hence must be of same origin.
- Indian sub-continent is more closely related to each other than to outside groups
- This allele must have been selectively swept up, by positive selection, not by mixing of outsiders.

Gallego Romero I, Basu Mallick C, Liebert A, Crivellaro F, Chaubey G, Itan Y, Metspalu M, Easwarkhanth M, Pitchappan R, VILLEMS R, Reich D, Singh L, Thangaraj K, Thomas MG, Swallow DM, Mirazón Lahr M, & Kivisild T (2011). Herders of Indian and European Cattle Share Their Predominant Allele for Lactase Persistence. *Molecular biology and evolution* PMID: [21836184](#)

Tandon et al. Lactose intolerance in North and South Indians. *Am. J. Clin. Nutr.* 34: 943-946, 1981

Razib Khan science blog on genes, culture and milk.

Fast evolving regions of the human genome

Use SNP diversity in regions between humans and variability between humans and chimps to locate regions of fast evolution. The essential idea is to identify regions that show low diversity between human groups, but show high divergence from chimp, but are identifiably conserved across several species (human, mouse, chimp, dog etc.)

Characteristic	HAR1	HAR2	HAR3	HAR4	HAR5
Location	5' region	Intron	Intron	Intergenic	Intron
Chromosome	Chromosome 20	Chromosome 2	Chromosome 7	Chromosome 16	Chromosome 12
Start ^a	61,203,966	236,556,014	1,979,228	71,686,982	844,471
Length	106 bp	119 bp	106 bp	119 bp	346 bp
Substitutions^b					
Human	13.93	11.96	5.98	4.98	8.34
Chimp	1.08	0.10	0.05	0.02	0.44
LRT statistic ^c	60.31	35.62	14.40	13.88	10.36

^aCoordinates from hg17 human genome assembly (build 35).

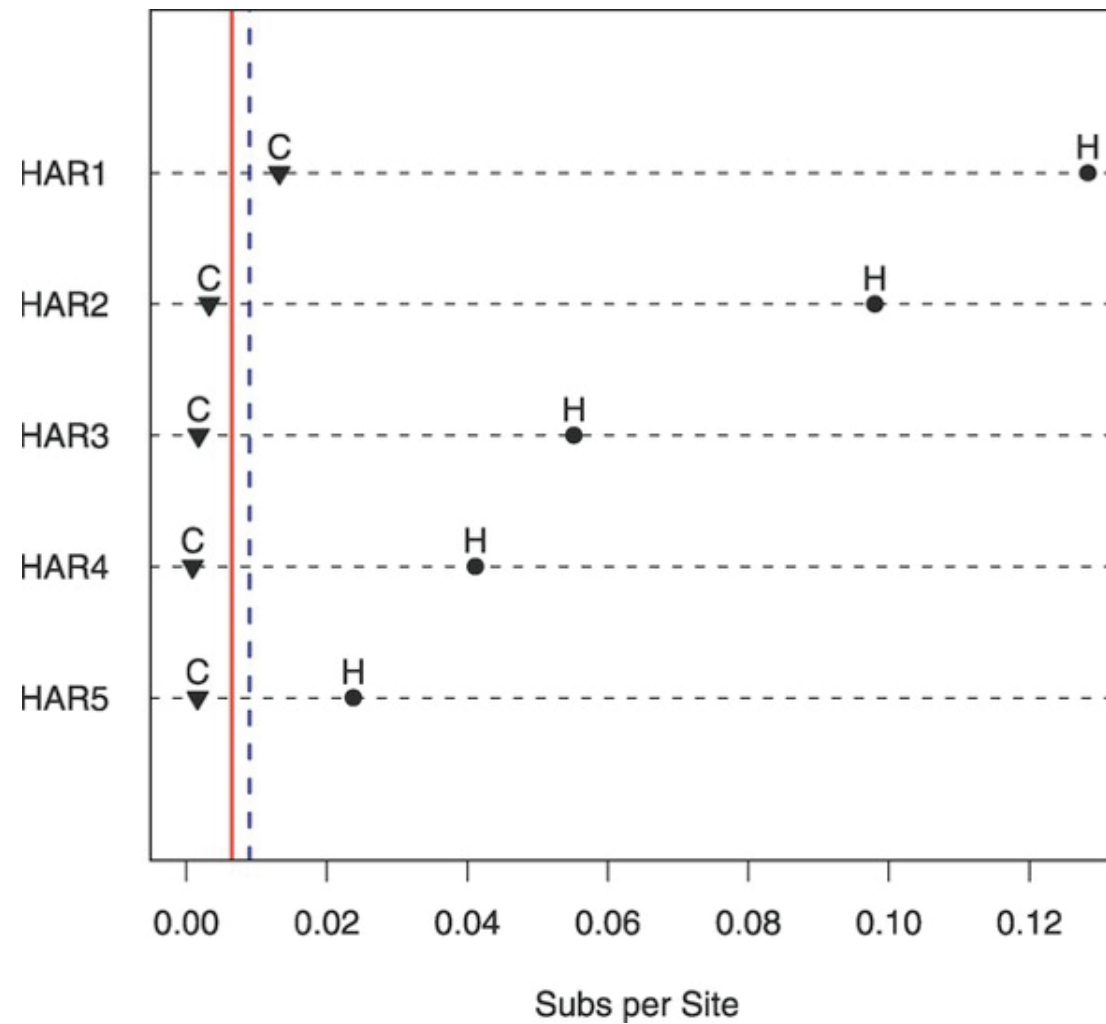
^bExpected number of substitutions reported by the phyloP program.

^cFDR adjusted $p < 4.5e-4$ for all five LRTs.

DOI: 10.1371/journal.pgen.0020168.t001

Pollard KS, Salama SR, King B, Kern AD, et al. (2006) Forces Shaping the Fastest Evolving Regions in the Human Genome. PLoS Genet 2(10): e168. doi:10.1371/journal.pgen.0020168
<http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020168>

Figure 1. Comparison of Substitution Rates in HAR1–HAR5



Pollard KS, Salama SR, King B, Kern AD, et al. (2006) Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genet* 2(10): e168. doi:10.1371/journal.pgen.0020168
<http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020168>

HAR2

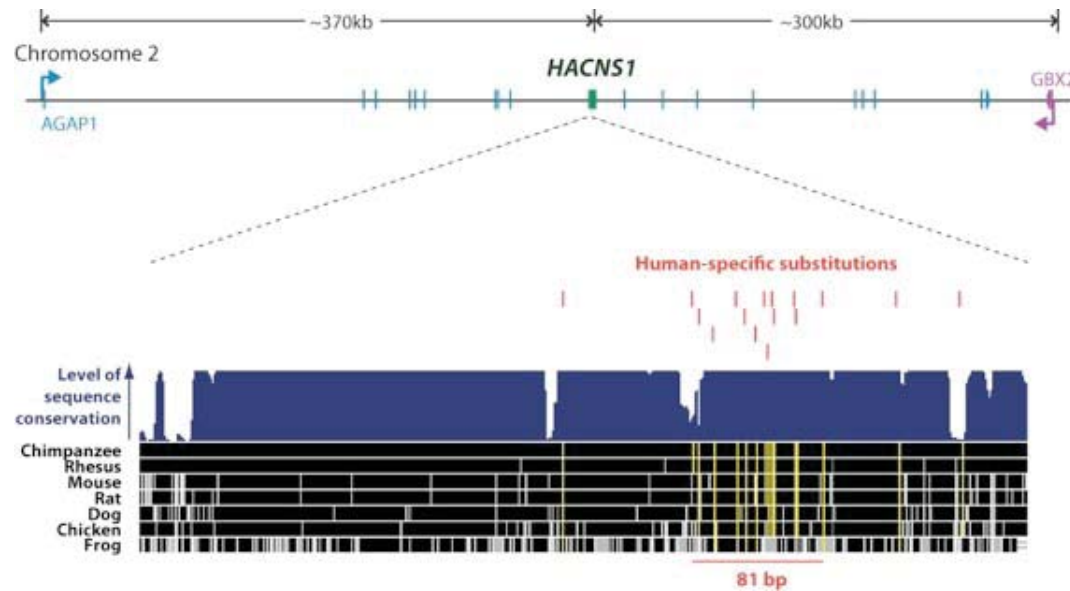


Figure 1. Genomic location of *HACNS1* and its level of conservation across terrestrial vertebrate genomes. Top: *HACNS1* is located in an intron of *AGAP1* and 300 kb downstream of *GBX2* on chromosome 2. Bottom: Sequence alignment of *HACNS1* with orthologs from other vertebrate genomes. Positions identical to human are shown in black. A plot of sequence conservation is shown in blue above the alignment. The location of each human-specific substitution is indicated by a vertical red line. The depth of nonhuman conservation at human-substituted positions is shown by a vertical yellow line that indicates whether each sequence is identical to chimpanzee and rhesus at that position. The location of a cluster of 13 substitutions in 81 bp is underlined.

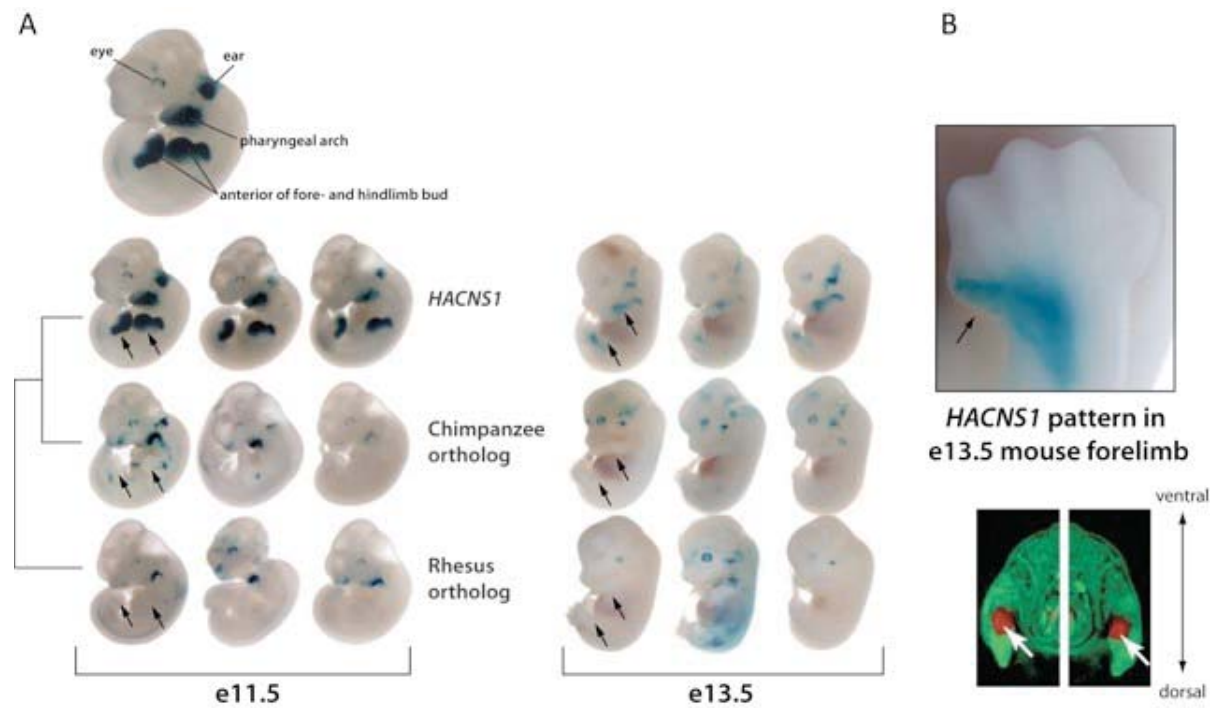


figure 2. Human-specific gain of function in *HACNS1*. **A.** *HACNS1* acts as an enhancer in transgenic mouse embryos, driving robust expression of a lacZ reporter gene in the developing anterior limb and other structures at embryonic day 11.5 (E11.5) and E13.5. However, the chimpanzee and rhesus orthologs fail to drive consistent expression in the limb at either time point. **B.** In the E13.5 limb, the human-specific expression domain extends into the handplate (and footplate, not shown) and includes digit 1, homologous to the human thumb and great toe.

Yale School of Medicine
Kavli Institute for Neuroscience

<http://www.yale.edu/noonanlab/HACNS1.html>

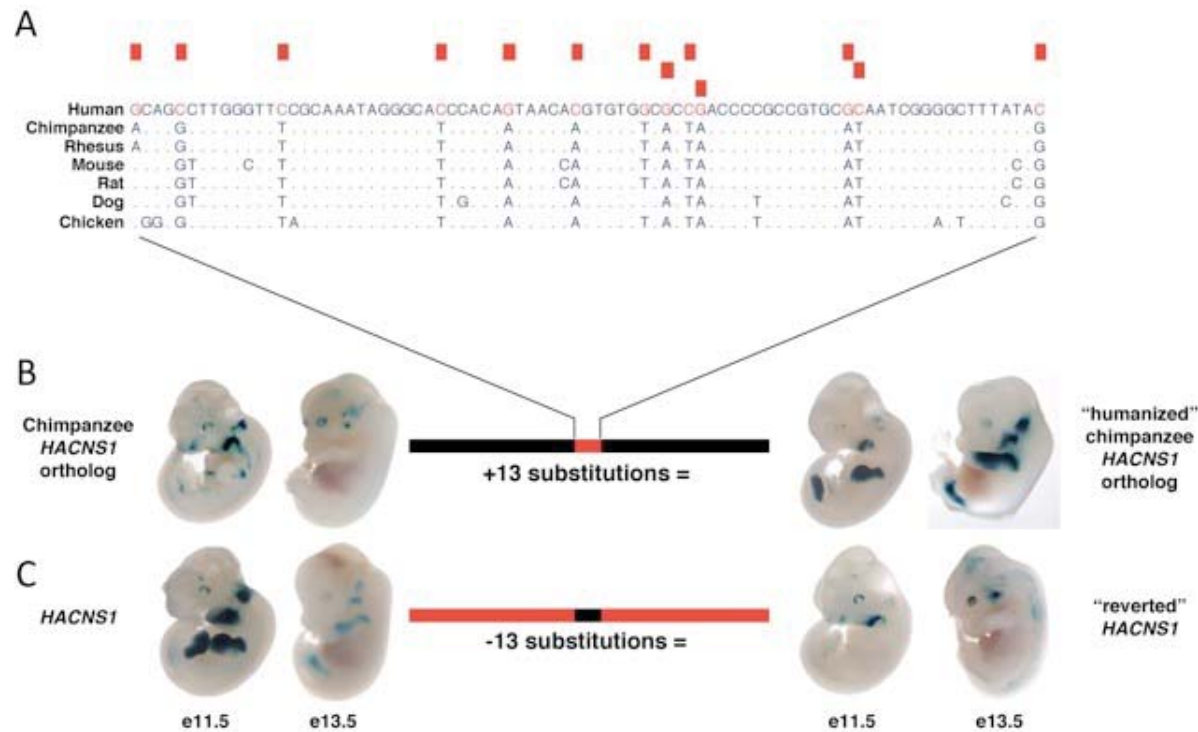


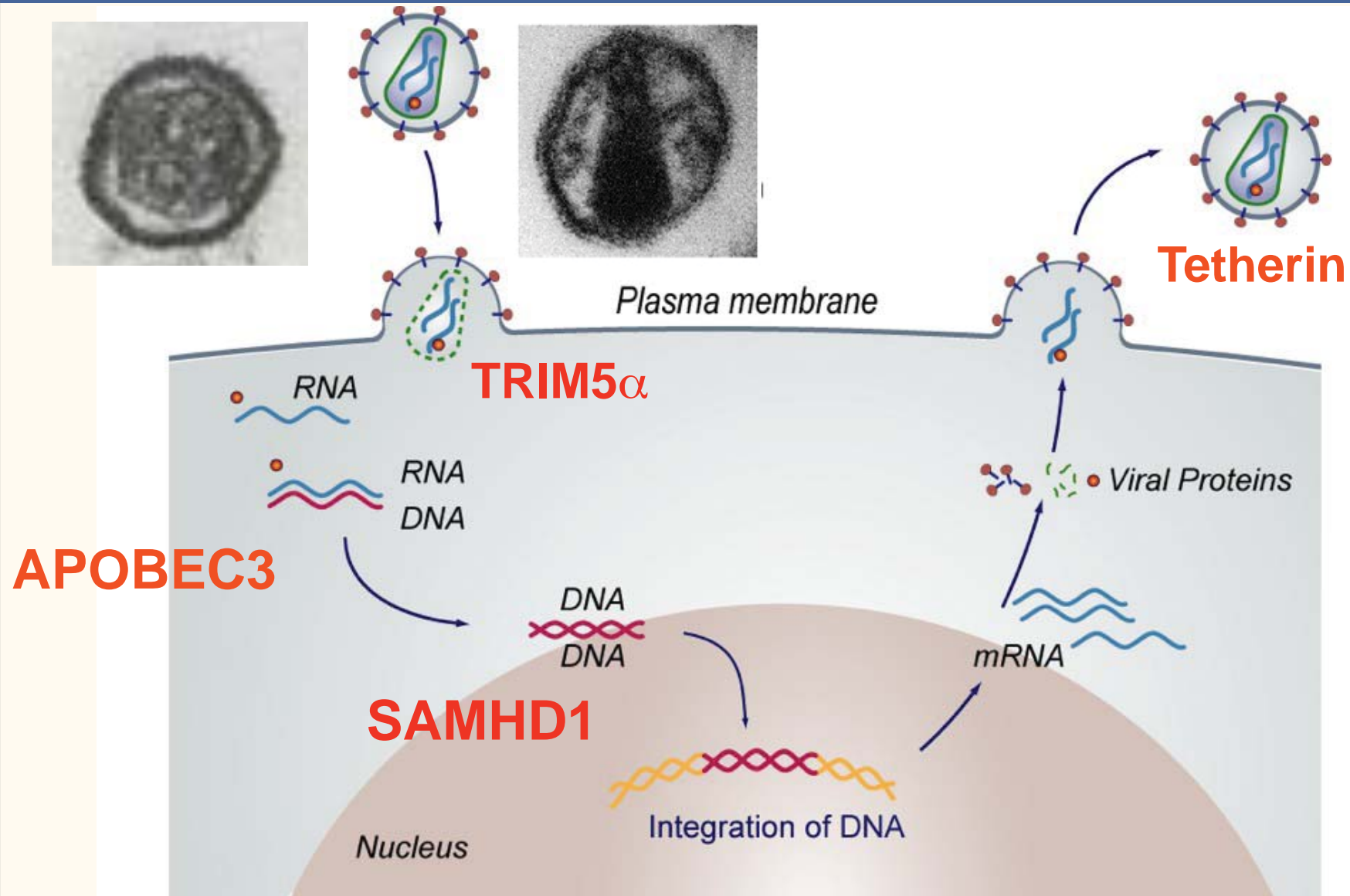
Figure 3. Molecular basis of the human-specific gain of function in *HACNS1*. **A.** Alignment of *HACNS1* with orthologous sequences from other genomes, focused on an 81-bp region in the element containing 13 human specific substitutions. Each human-specific nucleotide is highlighted in red. These 13 substitutions are sufficient to confer the gain of function. **B.** Expression pattern of a synthetic enhancer in which the 13 human-specific substitutions (red box) are introduced into the orthologous chimpanzee sequence background (black bar). **C.** Expression pattern of a synthetic enhancer obtained by reversion of these substitutions (black box) in the human sequence (red bar) to the nucleotides in chimpanzee and rhesus.

A few studies that use ideas outlines above

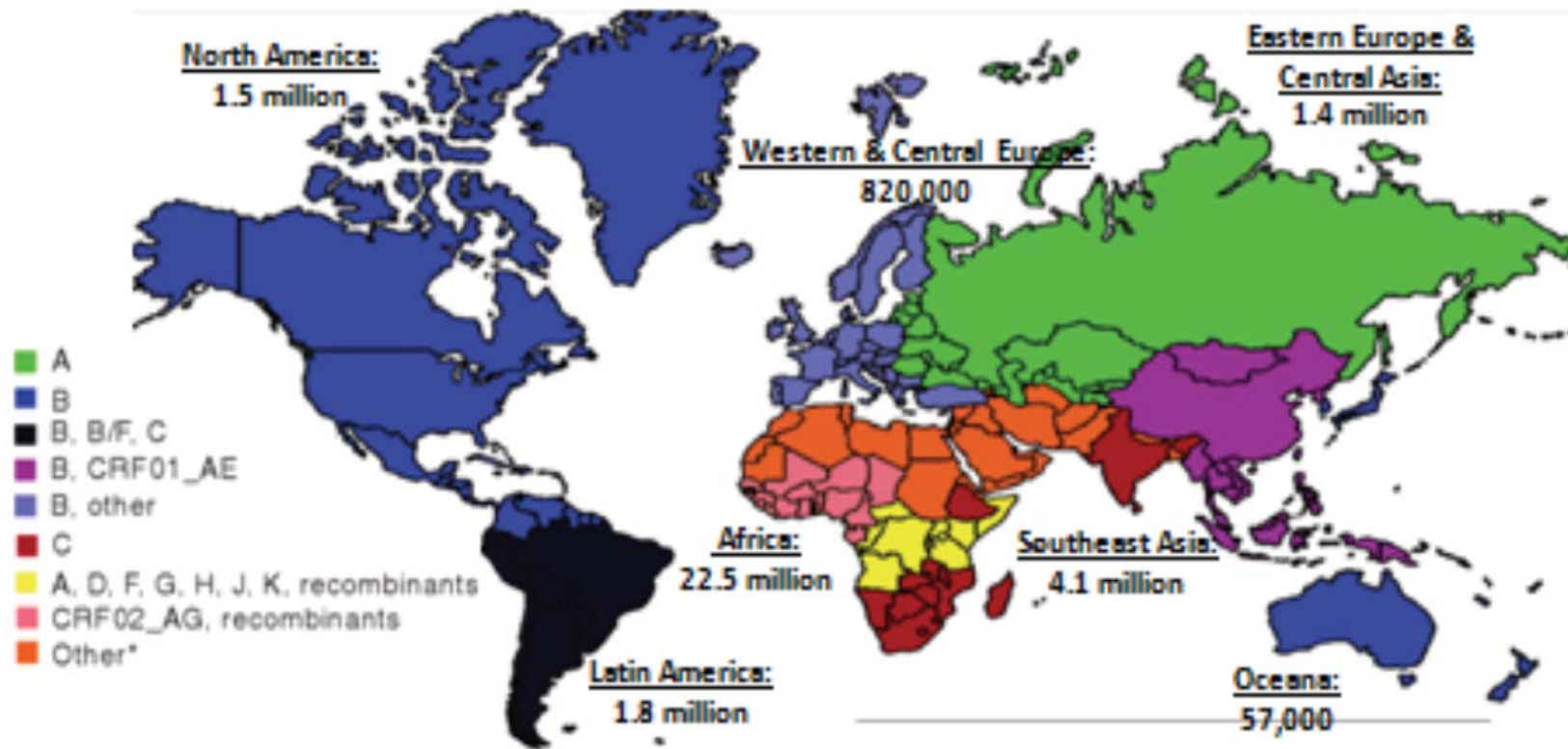
- HIV interaction with APOBEC3 which exhibits CNV across populations and is a potent anti-viral agent, which is kept in check by the VIF gene of HIV.
- TCR-repertoire. The TCR repertoire holds memory of past infections and helps the immune system fight them. Its diversity also helps recognize and fight new infections. How do we measure the diversity and what does it reflect about the state of the immune system ?

HIV-1 & APOBEC3 interactions

Host Restriction Factors

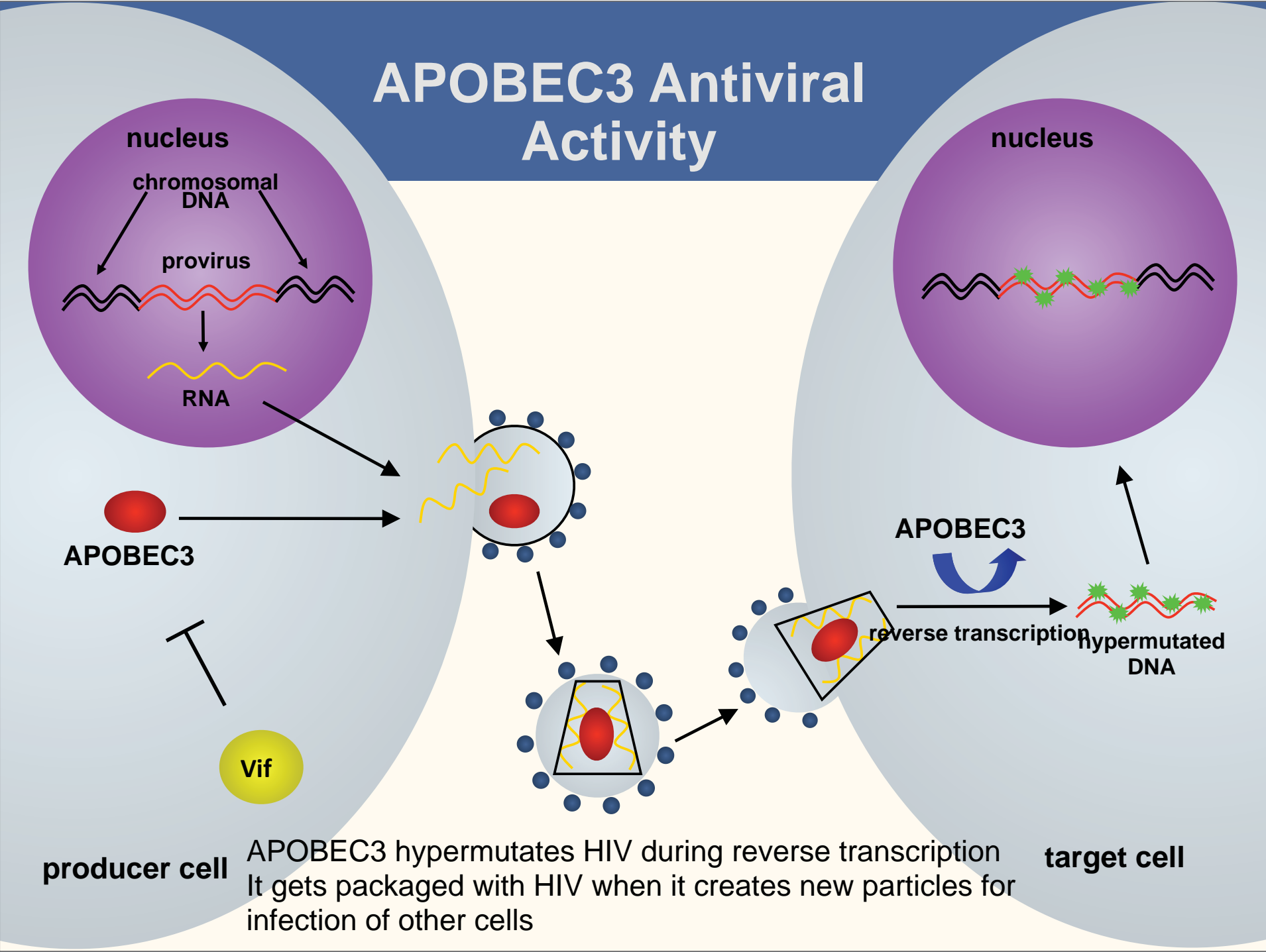


Worldwide Distribution of HIV-1

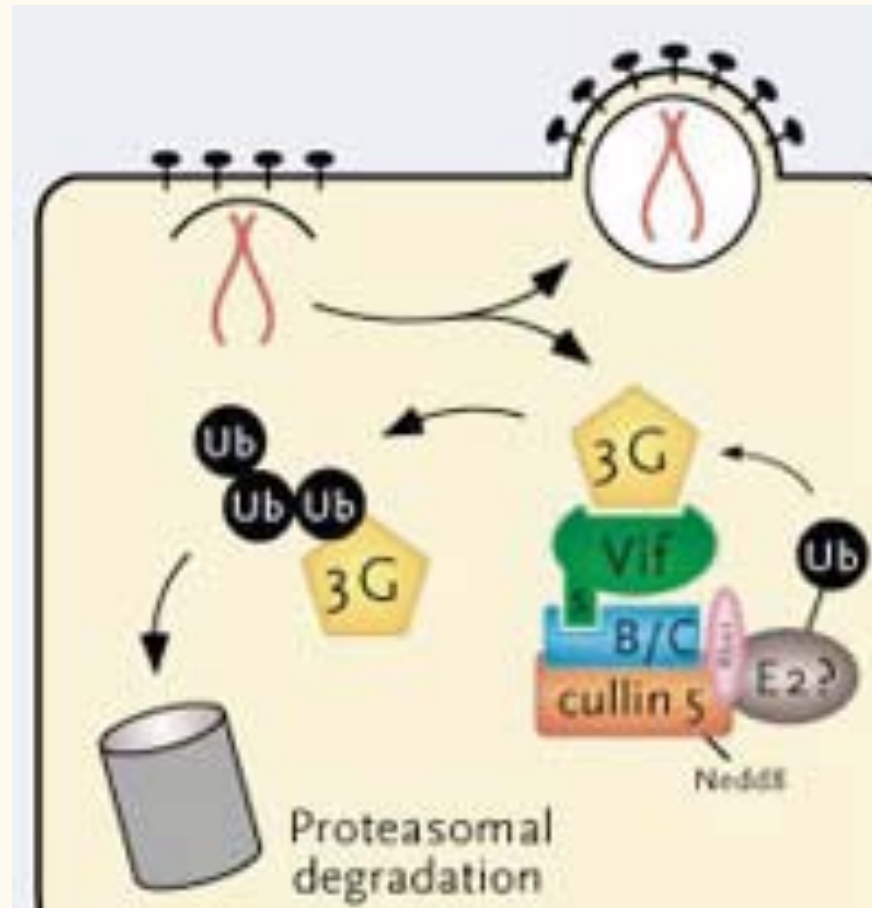


- 33 million people living with HIV

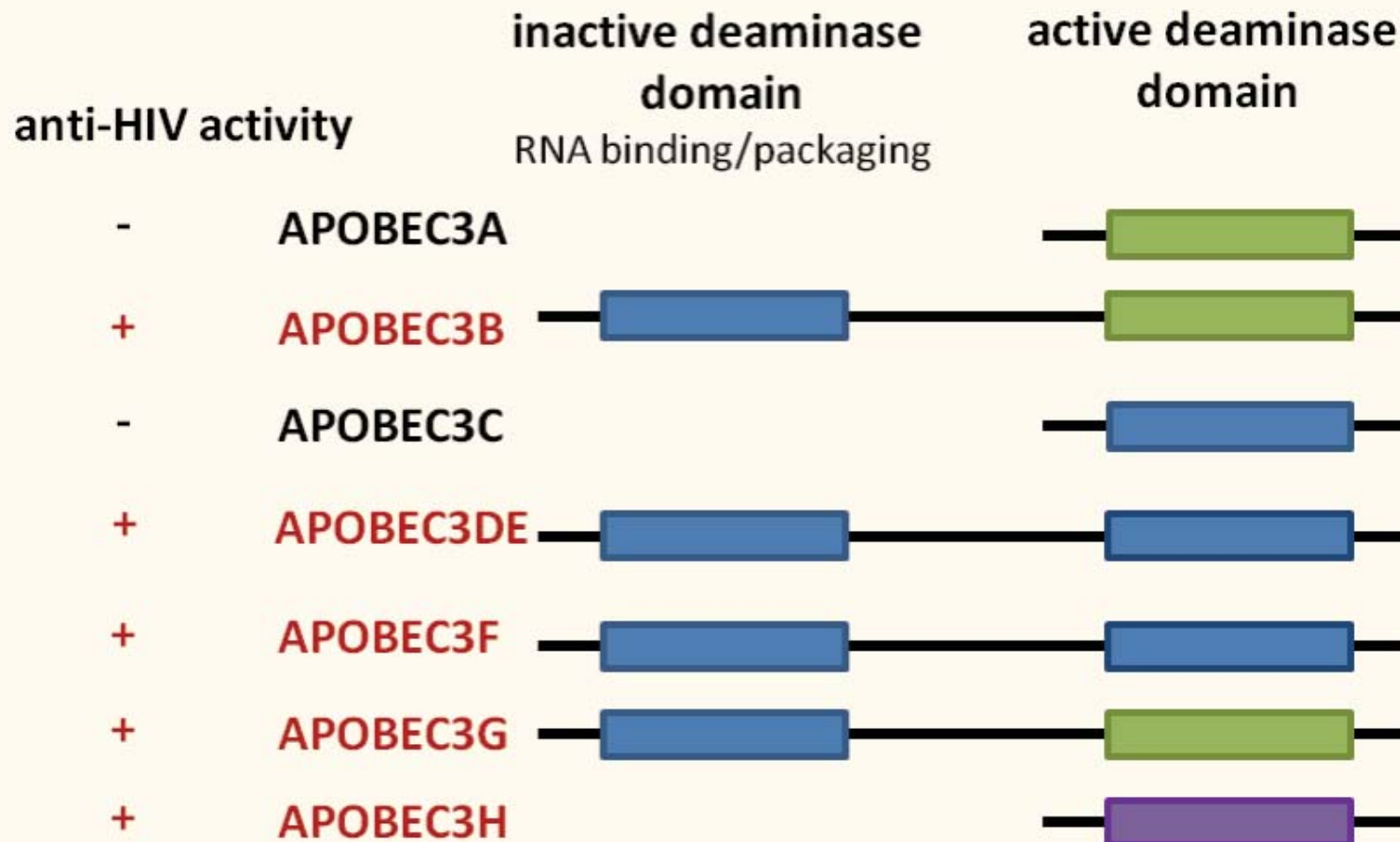
APOBEC3 Antiviral Activity



HIV-1 Vif targets APOBEC3 for proteasomal degradation



APOBEC3 cytidine deaminases



Mechanisms driving HIV-1 diversification

- Error prone nature of HIV-1 reverse transcriptase
 - Lack of proofreading due to absence of exonuclease activity
 - Misincorporation rate 1:10,000 (~1 mutation/genome)
 - *Incremental accumulation of changes*
- Recombination
 - Dually infected cells produce “chimeric” progenies
 - Recent data suggest that is common *in vivo*
 - *Simultaneous introduction of large numbers of changes*
- Deamination by DNA editing enzymes can shape circulating viruses and favor escape from antiretroviral drugs / immune control. APOBEC3, a defense mechanism, might help virus escape anti-retroviral drugs through increased mutation.

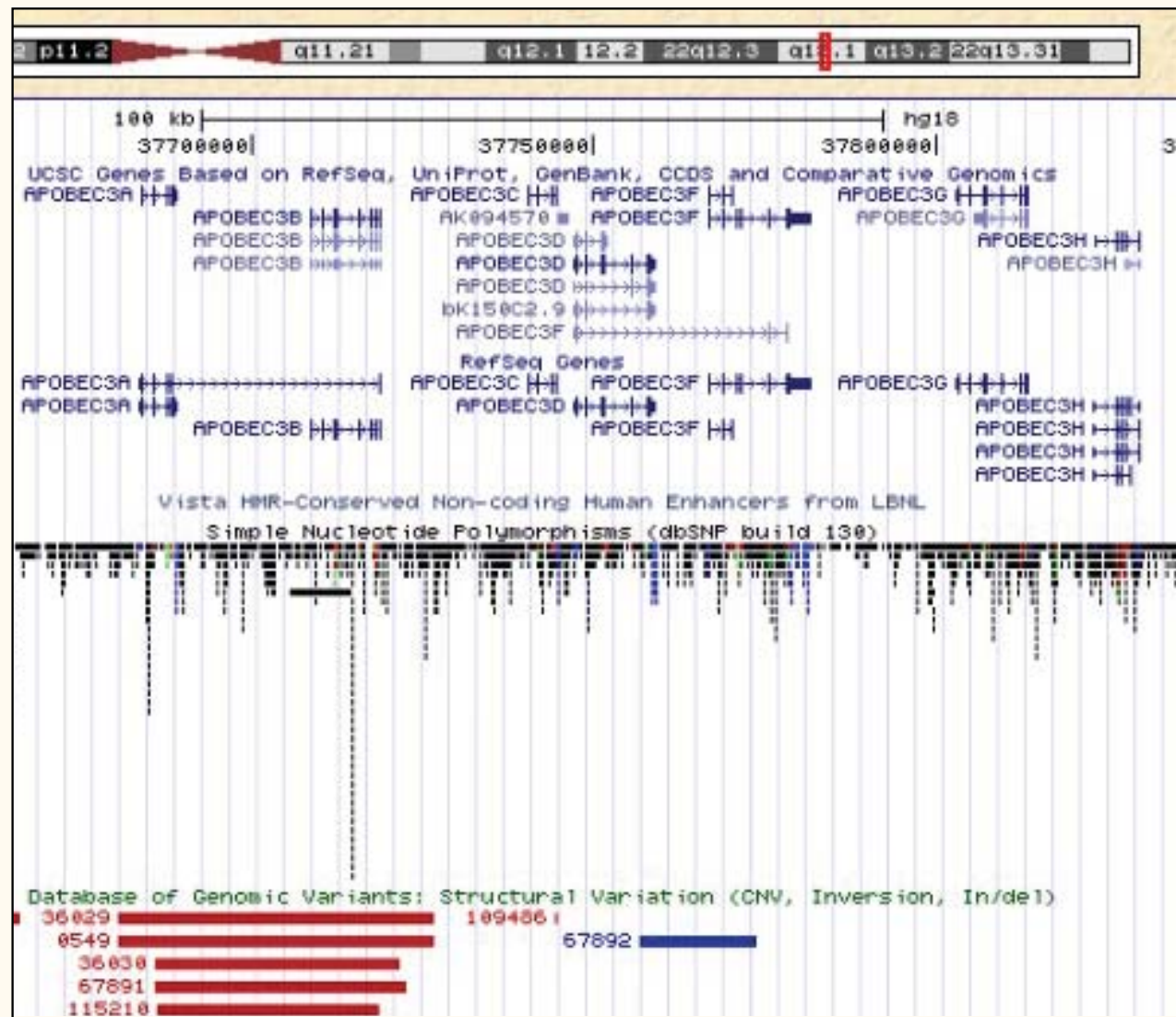
The APOBEC3 locus is polymorphic among human populations

7 deaminases (A3A-A3H)

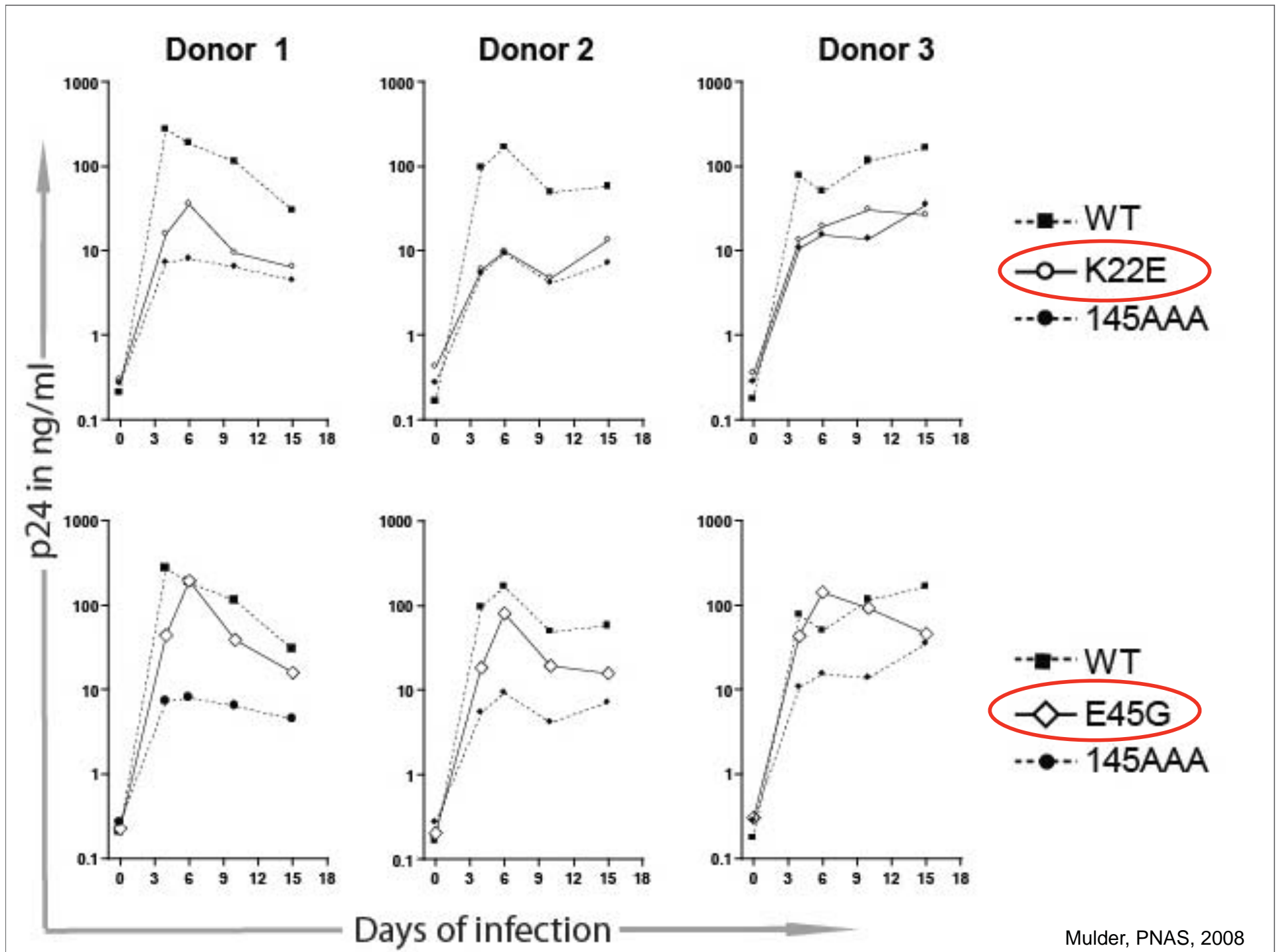
- Splice variants

- SNPs

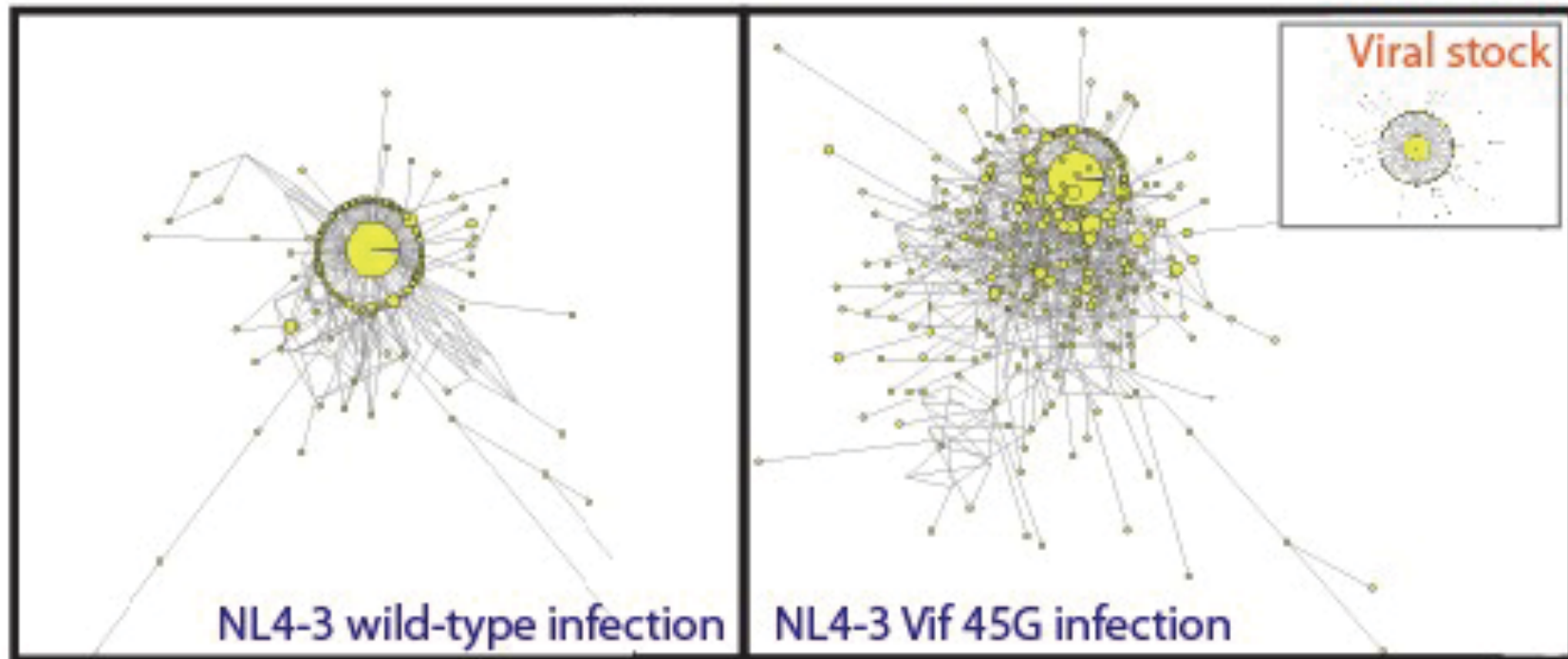
- CNVs



Using sequencing to study viral diversity and evolution



Partial neutralization of A3G drives viral diversity (454 NGS)



NL4-3 wild-type Vif

	A	T	G	C
A		6	26	3
T	5		1	22
G	532	13		1
C	7	16	0	

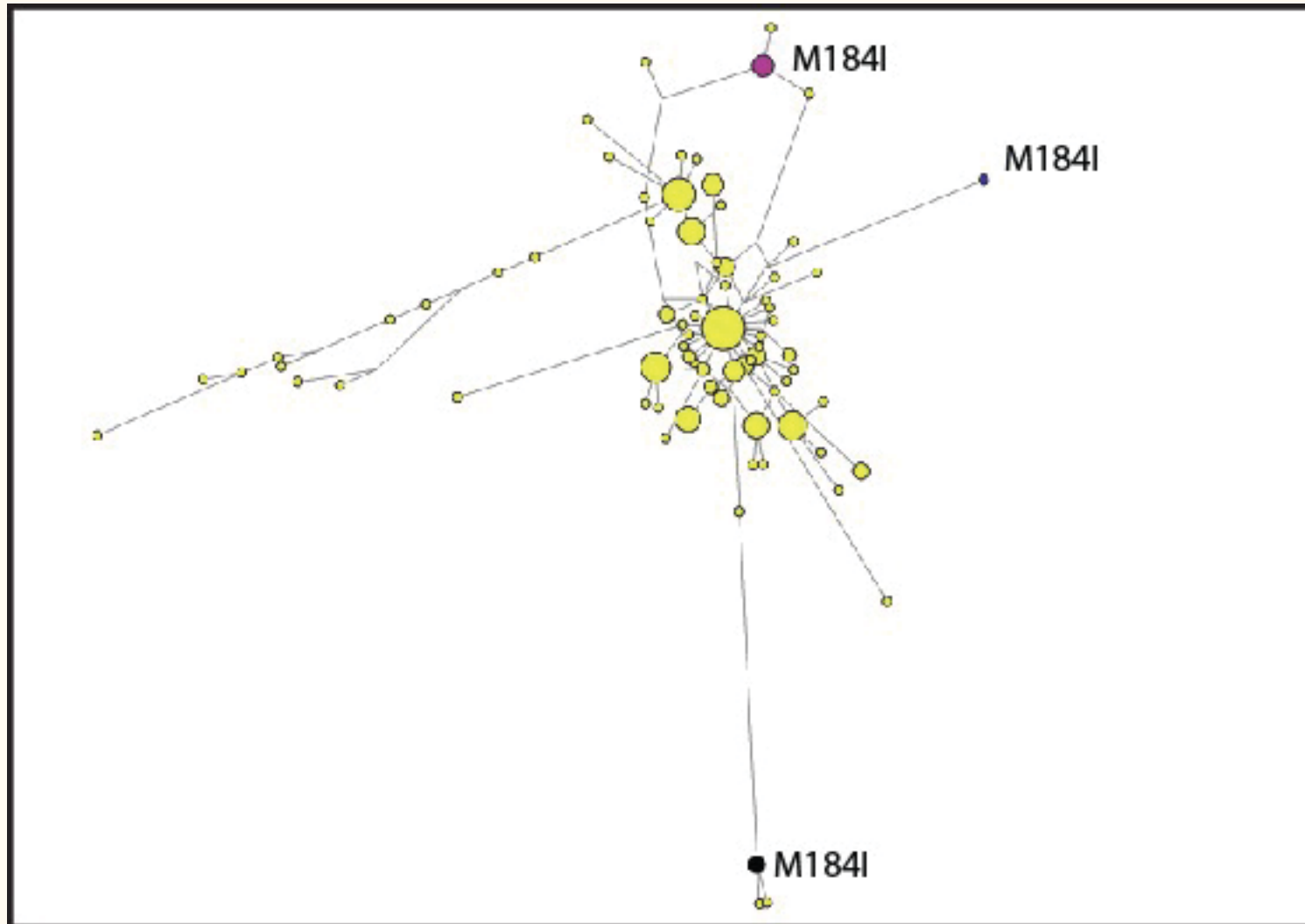
1,345,920 nucleotides/
4206 reads

NL4-3 Vif mutant 45G

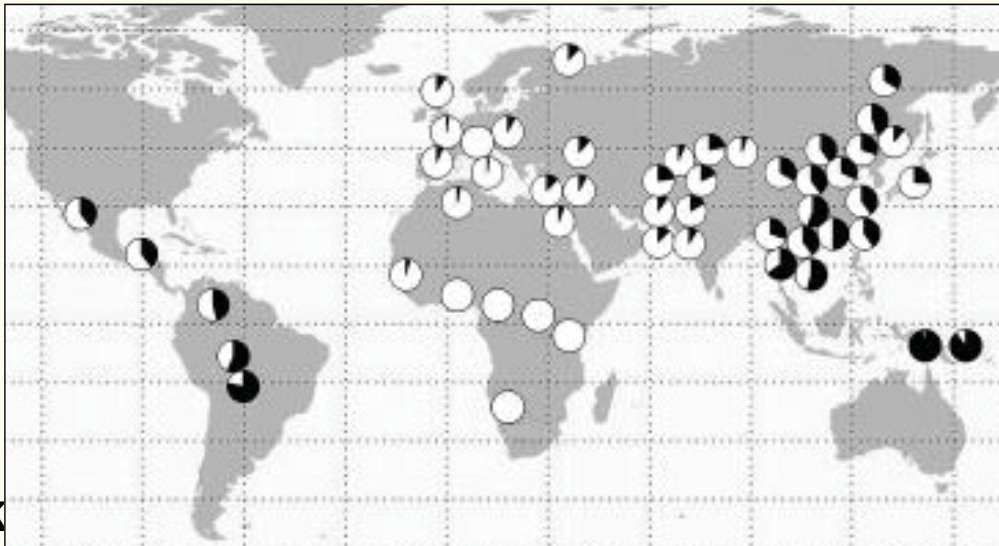
	A	T	G	C
A		7	16	0
T	1		2	4
G	2324	11		0
C	6	20	0	

744,640 nucleotides/
2327 reads

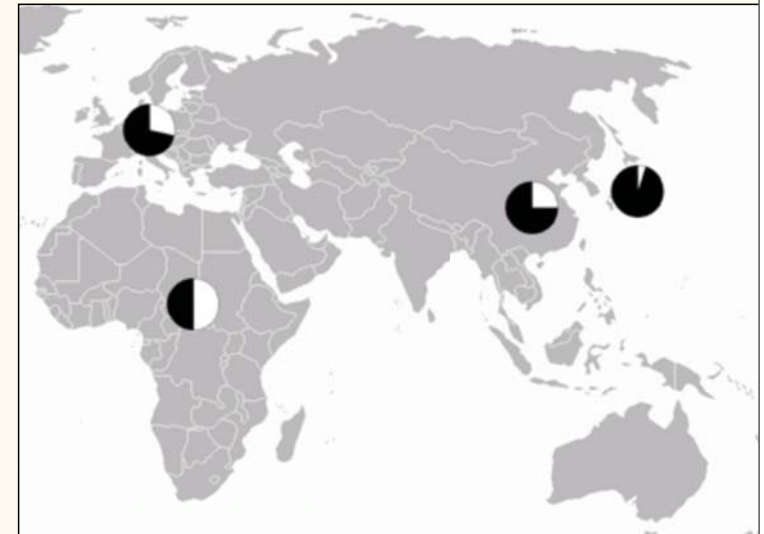
Viral diversity and RT M184I in a chronically infected patient (454 NGS)



Distribution of A3B deletion and A3H hapl among human populations



● - A3B
○ + A3B



● Inactive A3H
○ Active A3H

Conclusions: HIV & APOBEC3 interactions

- APOBEC3 locus is highly polymorphic (+/- 7 deaminases, SNPs, CNVs, splice variants)
- SNPs and CNVs may affect expression and antiviral activity of APOBEC3 variants
- HIV Vif alleles may adapt to APOBEC3 variants
- More Vif genotype-to-phenotype analyses are needed to develop algorithms that allow predictions of anti-APOBEC3 activity
- Some individuals could harbor less effective & protective APOBEC3 repertoires

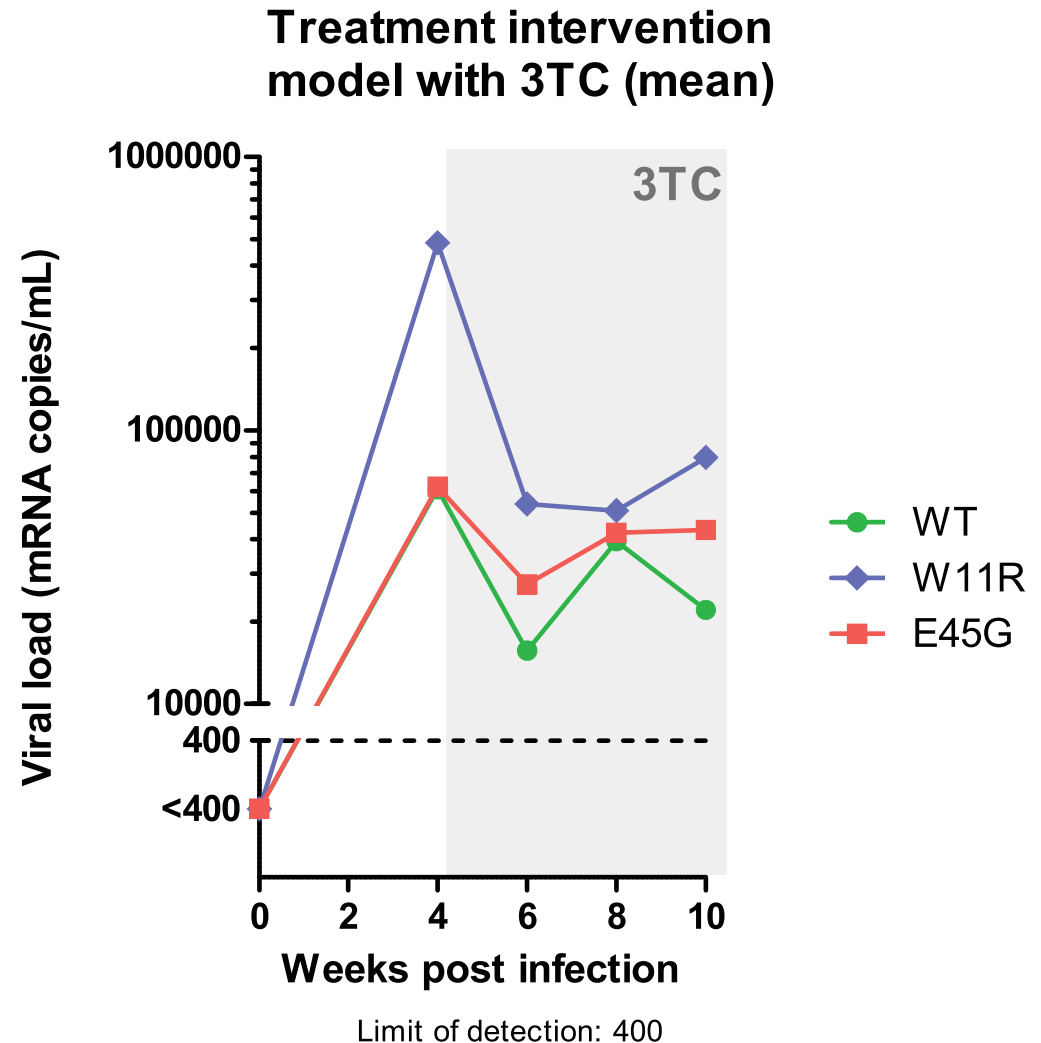
Conclusions: HTLV-1 & APOBEC3

- A3A, A3B and A3H-II restrict HTLV-1
- A3A and A3B require catalytic activity but A3H hapII restricts through an editing independent mechanism
- Several independently mutated HTLV-1 proviruses suggest that A3A, A3B and A3G could be active *in vivo*
- APOBEC3 repertoire could contribute to the geographic distribution of HTLV-1

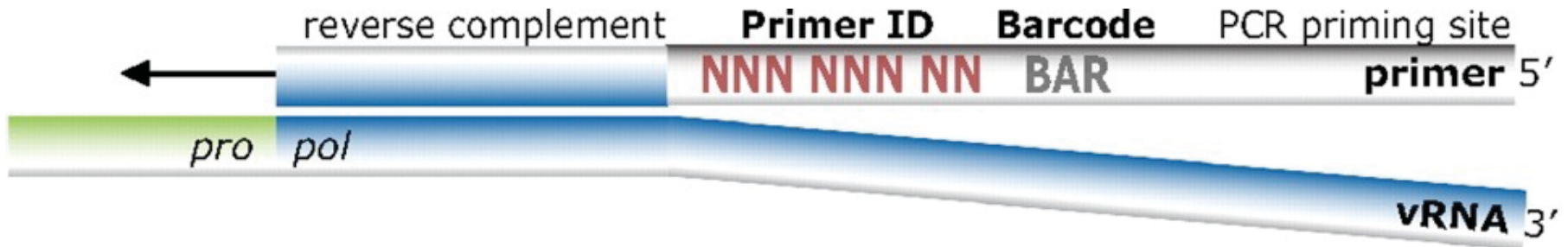
Consequences of APOBEC3 driven viral evolution in vivo: Our Deep sequencing Approach

- 46 plasma samples from HIV infected humanized mice
 - extract viral RNA,
 - Make cDNA with ID primer
 - Run 1st Round PCR (20-24 cycles)
 - Run 2nd Round PCR (27 cycles)

NGS with two different platforms (454 and MiSeq Illumina)



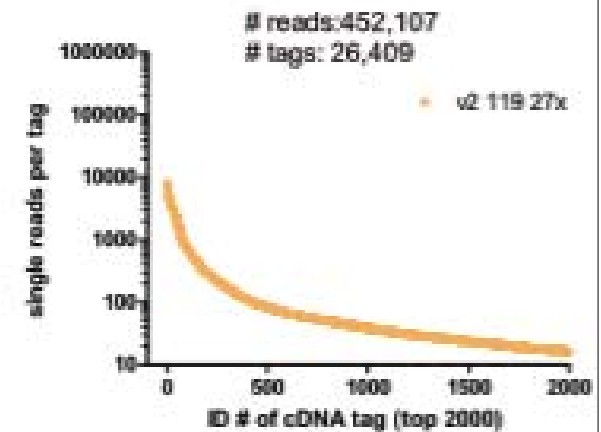
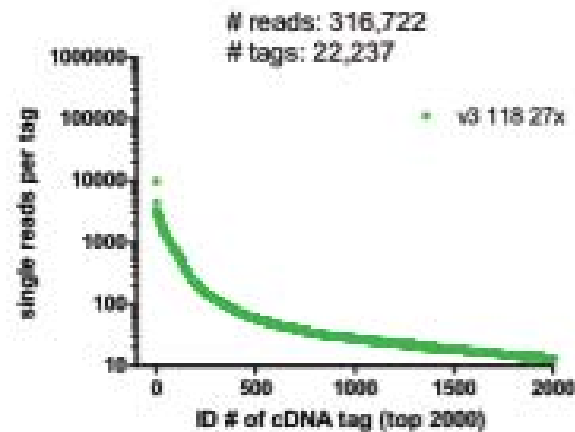
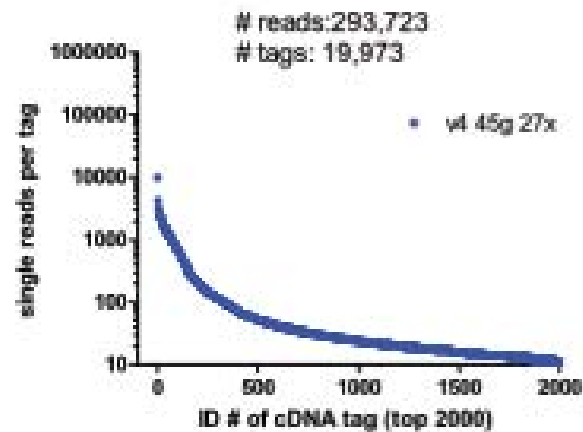
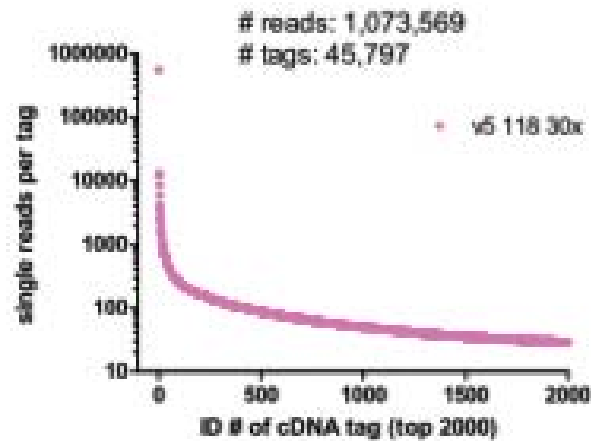
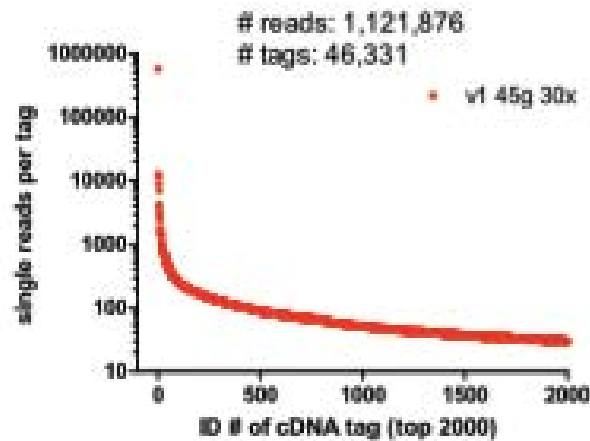
Tagging viral RNA templates with a Primer ID before PCR amplification and sequencing allows for direct removal of artifactual errors and identifies resampling



Raw sequence reads	Primer ID	Barcode
	CATAATAC	TAG
	CATAATAC	TAG
	CATAATAC	TAG
	CATAATAC	TAG
	CATAATAC	TAG
	CATAATAC	TAG
<hr/>		
Consensus sequence	CATAATAC	TAG

Number of Reads and cDNA tags per sample

3x10⁶ reads
per MiSeq run

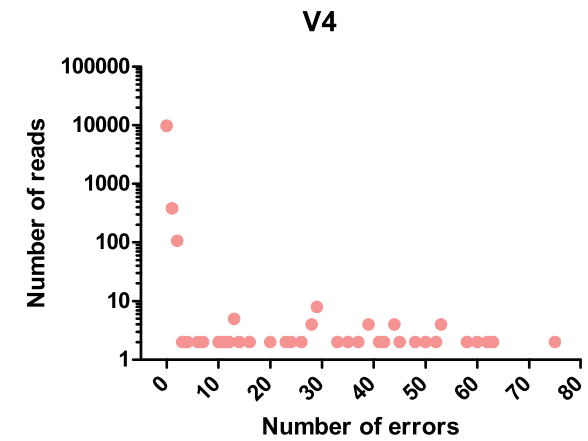
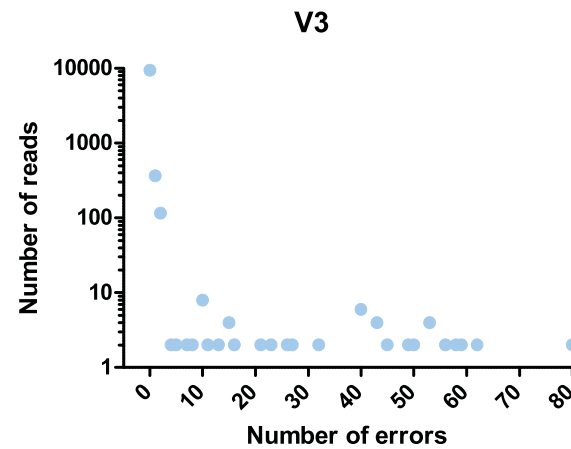
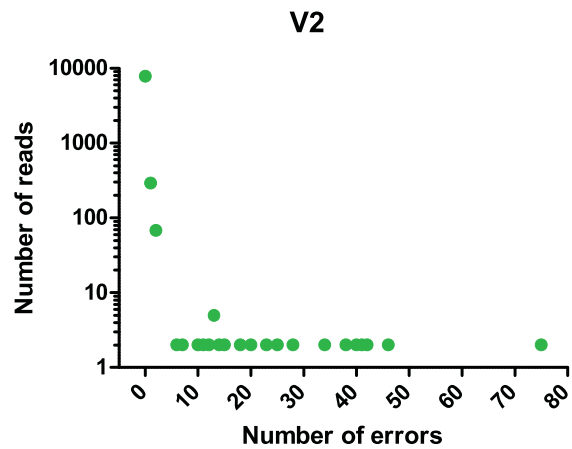
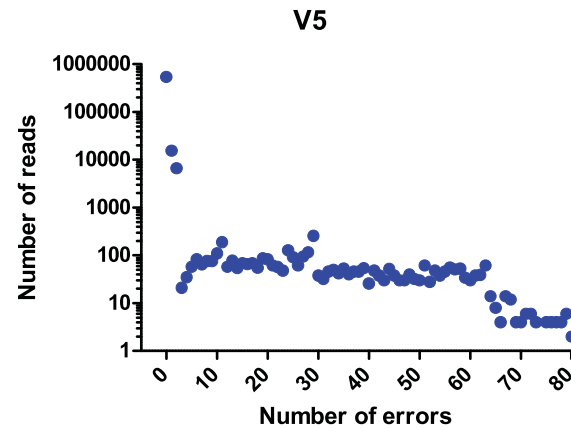
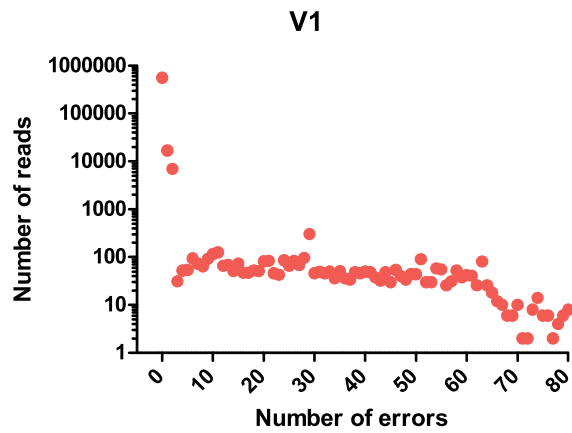


Generate consensus sequence for each cDNA tag

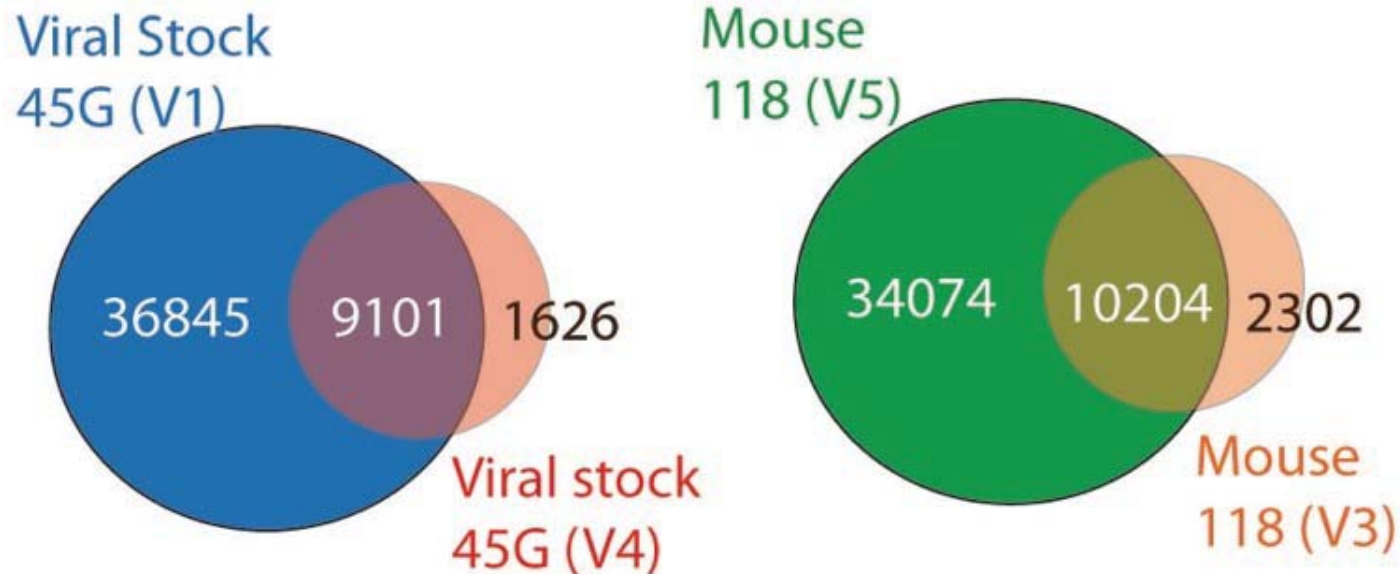
Numbers of errors per cDNA



Exclude cDNAs with more than 5 errors



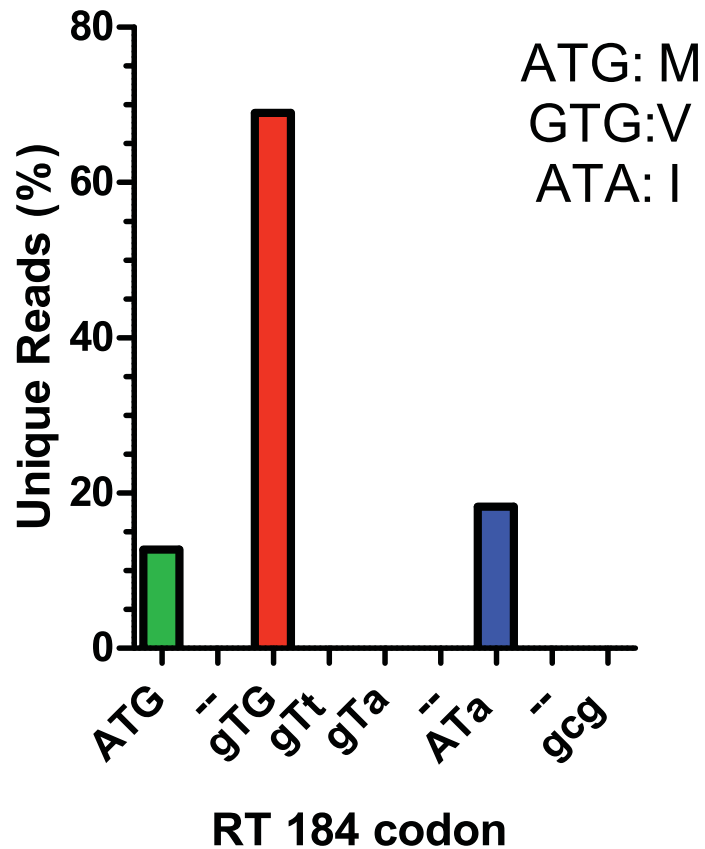
cDNA tag distribution in separate 2nd Round PCR samples (same 1st Round)



	private	shared	total		private %	shared %
V1 private	36845	9101	45946	V1	80	20
V4 private	1626	9101	10727	V4	15	85
V5 private	34074	10204	44278	V5	77	23
V3 private	2302	10204	12506	V3	18	82

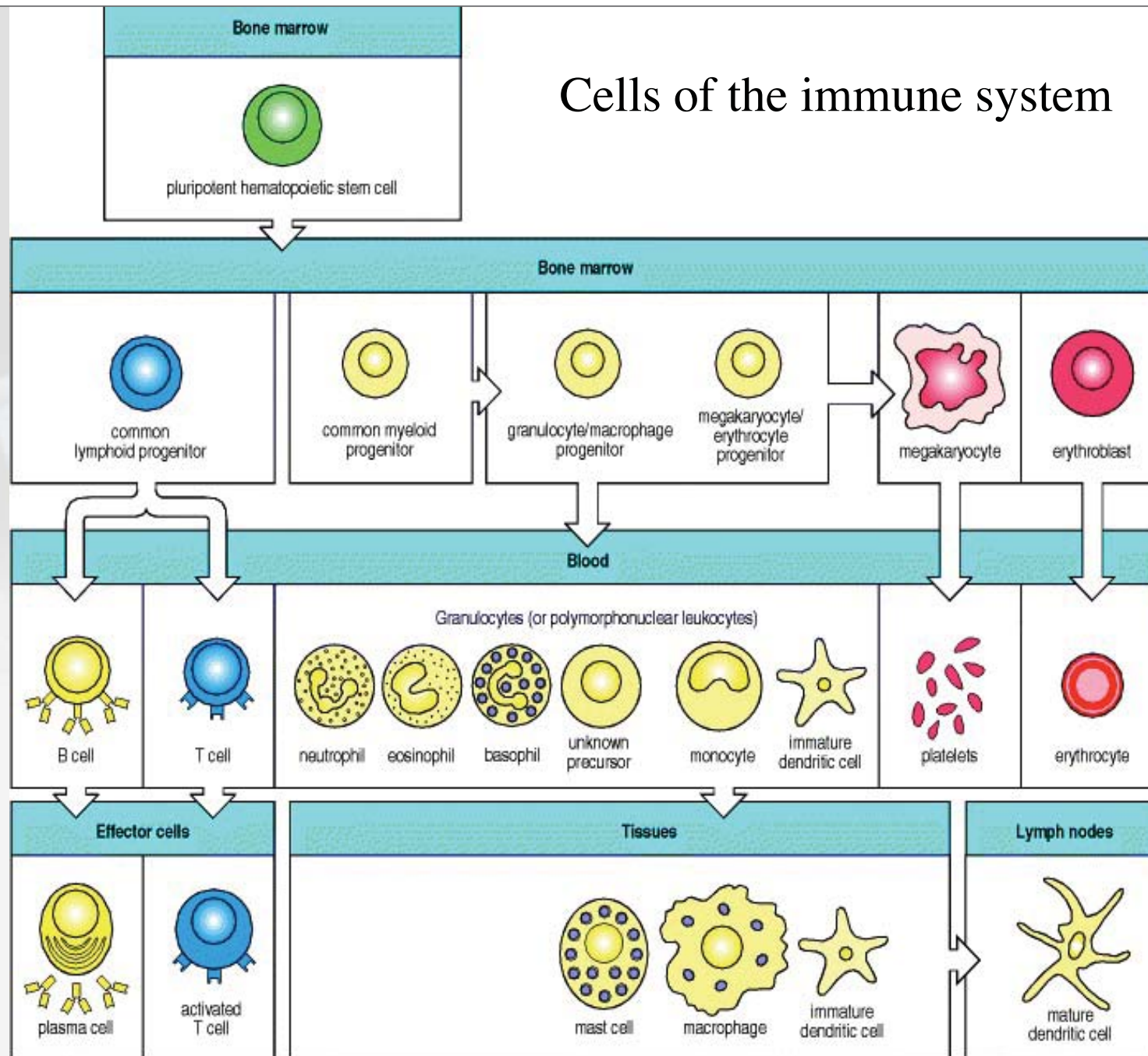
Mouse 118, time point 4 illumina

118-T4 nested (TM4) Relative distribution

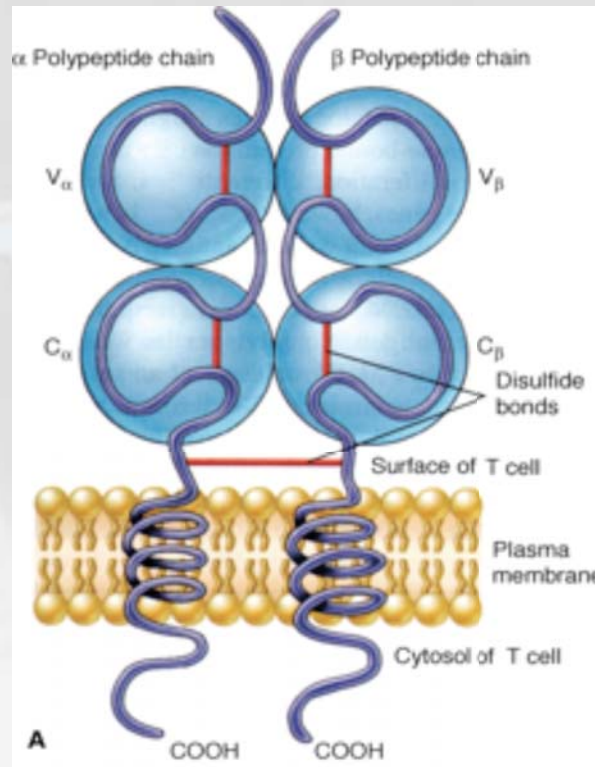


		ATG 12.7%	
V	gTG 68.9%	I	ATA 18.2%
L	cTG 0.002%	I	ATC 0%
L	tTG 0.002%	I	ATT 0.002%
		V	gTa 0.044%
		V	gTt 0.044%
		V	gTc 0%

Cells of the immune system



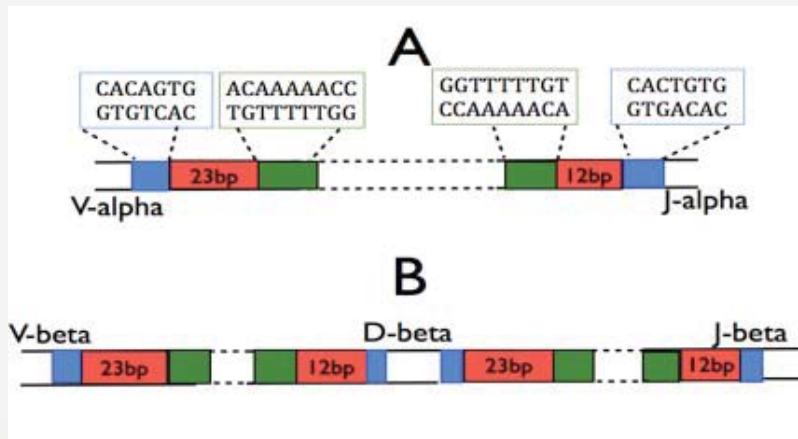
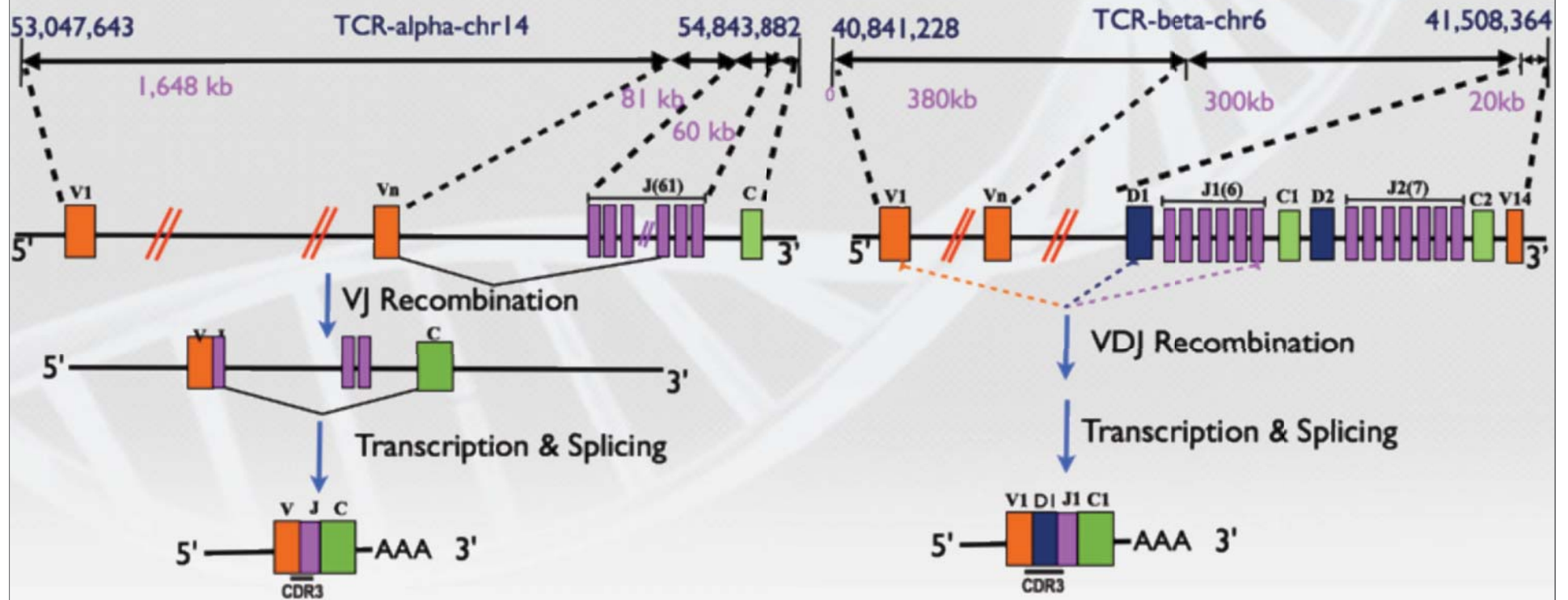
The T cell Receptor



Structure of the T cell receptor.
Adapted from Sherris Medical
Microbiology, 5th edition by
Ryan KJ & Ray CG.



VDJ recombination



The 12-23 rule

RAG (recombination activating genes)


Recombination signal sequences (RSS)



Why study the repertoire

- The repertoire is extraordinarily diverse (more combinations than the number of galaxies in the universe 10^{12})
- The repertoire is highly plastic.
- The repertoire can serve as a biomarker in cases of infection, transplant, autoimmunity, allergy and potentially other medical conditions
- Very elementary questions still remain unanswered.
- Large scale monitoring of the immune system can open a path for patient-individual medicine.





Notes for material worked out on the board



- Assembly of reads
 - Since reads are generated from randomly fragmented pieces, they can be assembled back into the original fragments, provided there is sufficient coverage (amount of sampling per position) and the sequence is non-repetitive (information content is high)
 - Assemblies are often partial, due to inadequate coverage across portions of the genome
- Problem -- Assemble N-mers (sub-sequences of length N) into a minimal string that contains all the sequenced N-mers. Higher N's ensure unique assemblies.



Toy model

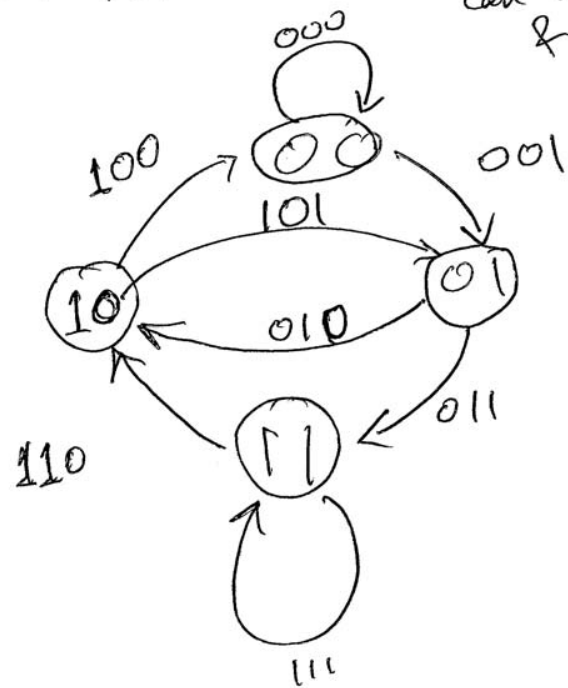
- Assemble a set of 3-mers consisting of 0's and 1's to a minimal string that contains all the 3-mers. This can then be generalized to any N-mer made up of A's, C's, G's and T's.
- Write down all possible 2-mers, which form the vertices of a graph we are going to draw, called the **de-Bruijn** graph. The directed edges are the 3-mers that have the first vertex as a prefix and the second vertex as a suffix, so not all pairs of vertices can be joined together. The process is shown on the next slide.



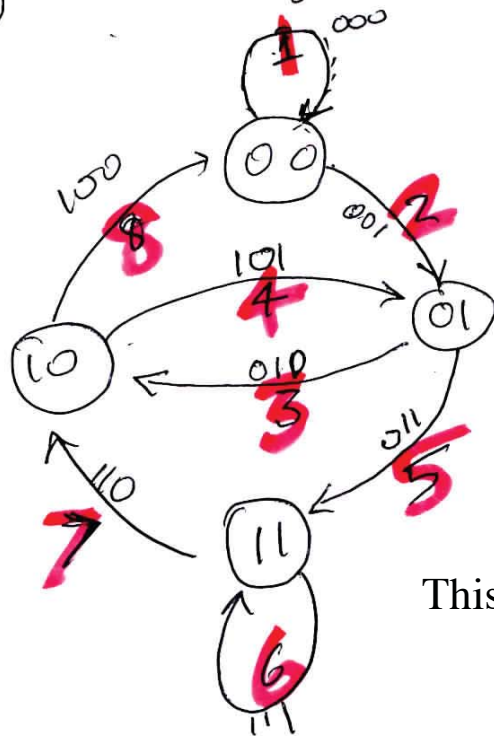
00
 10 01
 11

00 $\xrightarrow{001}$ 01 allowed

01 $\not\rightarrow$ 00 Not allowed (as no 3-mer
 can have 01 as prefix
 & 00 as suffix)



Find path through all vertices going through each edge only once (Eulerian path)



This problem of finding an Eulerian path through this graph is computationally easy to solve.

the 3-mers in order of traversal

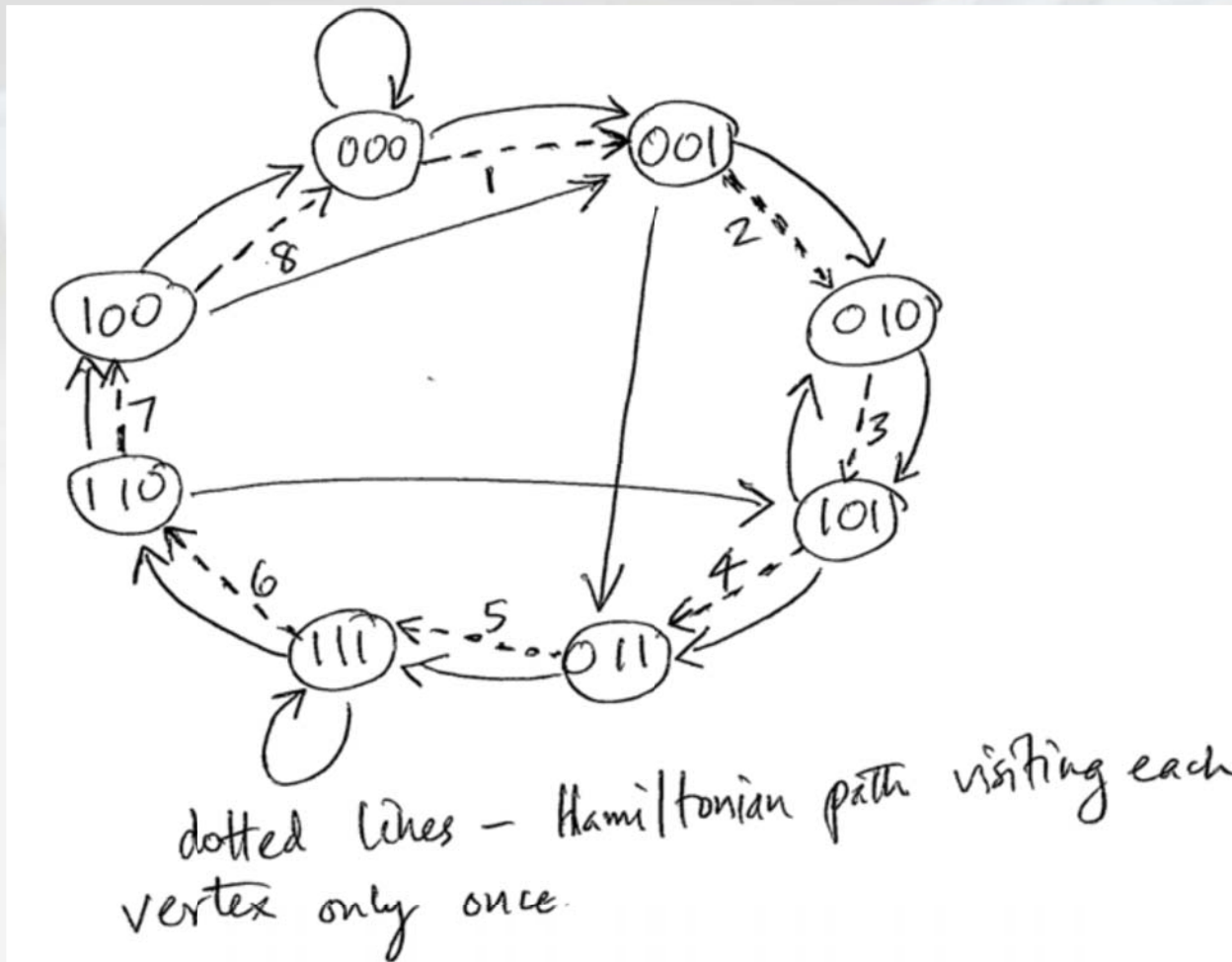
in order 000, 001, 010, 101, 011, 111, 110, 100

the first letters of the 3-mers form the superstring

contains all 3-mers



The dual to the Eulerian problem is the one to find a Hamiltonian path through a dual graph, where edges are replaced with vertices and vice-versa, and each vertex is visited only once. This problem is hard to solve.



Alignments

Two strings S (ACGTAC) & T (AGGTAC)
 The Needleman-Wunsch matrix

S →	x	A	C	G	T	A	C
T ↓	x						
	A	↙	↓				
	G						
	G						
	T						
	A						
	C						

A scoring matrix decides scores for matches, mismatches and gaps. This is critical in determining optimal alignments.

Any alignment is a path through this matrix. To efficiently determine the path, each cell is given a score, based on the 3 ways the cell can be reached as shown in the figure.

This approach of solving a big problem by solving small local problems is called Dynamic Programming.

- horizontal lines - Add letter from S, gap from T and penalty of gap.
- ↓ vertical lines - Add letter from T, gap from S, & penalty of gap.
- ↙ diagonal line - Add letter from T & S and a match or mismatch score depending on letters in T and S.



The score at a cell is the sum of the score at origin cell + value of the path (gap or mismatch or match). ~~The~~ The score is made zero if its best value goes below zero.

The best path starting from cell with highest score & tracing back till a zero or end of matrix is reached. This is an optimal local alignment & The algorithm is called Smith - Waterman - Gotoh.



Gap penalties

if $g(n)$ is gap penalty for a gap of “ n ” letters, then a realistic constraint is for the gap to be a convex function of n , that is,

$$g(n+1)-g(n) \leq g(n)-g(n-1)$$

which means it costs less to extend the gap by one, than it did to add the last gap.



Distances between sequences

- Hamming or edit distance. After aligning sequences, differences between them are counted to get a distance.
- A single change at a position can hide multiple ways in which the final state could have been reached,
- $A \rightarrow B$ or $A \rightarrow C \rightarrow B$ or $A \rightarrow C \rightarrow D \rightarrow B$ etc. longer cycles are less probable than shorter ones, but they all contribute.



Jukes-Cantor distance

Given two sequences of equal length which align to each other, what is the correct definition of distance, allowing for the multiple changes that might have occurred. Jukes-Cantor distance is one possible definition

if, α is probability of a base mutating into another

then $P_A(t)$ is prob. of A at a location at the next instant,

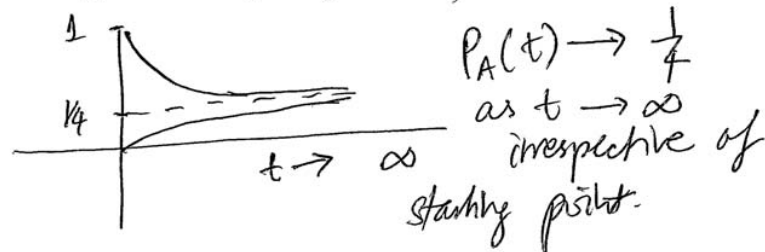
$$P_A(1) = (1 - 3\alpha) P_A(0) \quad \left[\begin{array}{l} \text{can mutate into} \\ 3 \text{ alternates.} \end{array} \right]$$

at the next instant,

$$P_A(2) = (1 - 3\alpha) P_A(1) + \alpha (1 - P_A(1))$$

We can convert the difference equation into a differential equation, & solve to get

$$P_A(t) = \frac{1}{4} + \left(P_A(0) - \frac{1}{4} \right) e^{-4\alpha t}$$



~~If~~ The chance of an identity at a position in the two sequences is $I(t)$

$$I(t) = P_{AA}^2(t) + P_{AT}^2(t) + P_{AC}^2(t) + P_{AG}^2(t)$$

$$= \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$$

$$1 - I(t) = p = \frac{3}{4} (1 - e^{-8\alpha t})$$

(proportion of differences between sequences)

$$8\alpha t = -\ln\left(1 - \frac{4}{3}p\right)$$

$K = 6\alpha t = \#$ of subs between them

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right)$$

Jukes-Cantor distance between sequences.



Questions ?

Contact me at ravi.mssm@gmail.com for criticisms, comments and questions.

I have tried to give a quick overview in lectures 1 and 2 of the broad ideas in genomics/ bioinformatics.

I have attached to this presentation, slides containing notes, for some material I worked out on the board, as well as some extra population genetics material that I did not have time to get into.

