Single Nucleotide Polymorphisms (SNPs), population genetics and human genetics

Biosciences 741: Genomics

Fall, 2013

Week 4

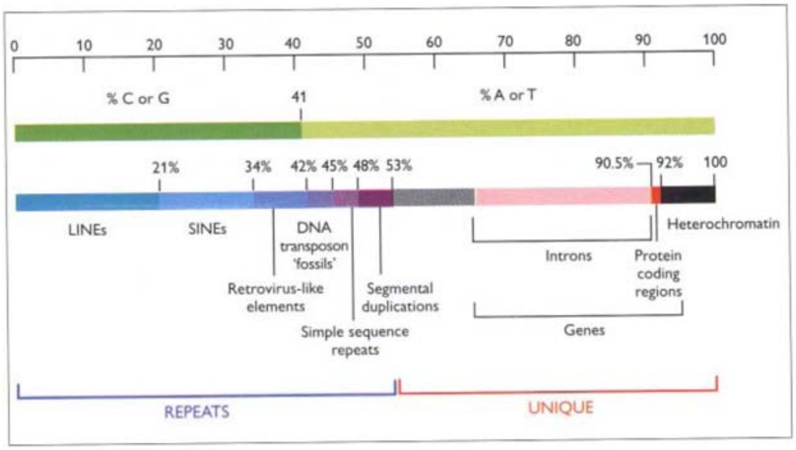
Human Genetic Diversity

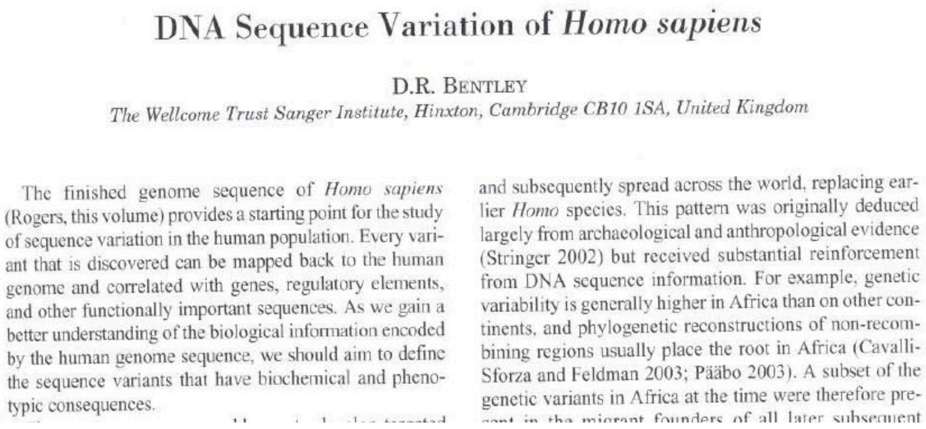
- 90% of human genetic polymorphisms are caused by SNPS; the remaining polymorphisms are structural variants including insertions, deletions, and so on.
- Types of SNPs (Bentley)
- Linkage disequilibrium
- The neutral theory of molecular evolution
- Techniques used to map human traits & score SNPs
- Structural variants & medical applications

Types of SNPs

- Coding vs. non-coding
- Synonymous vs. non-synonymous
- Transitions vs. transversions
- Functional vs. non-functional
- Mutations vs. polymorphisms
- Substitutions vs. polymorphisms

Content of the human genome





DNA Sequence Variation of *Homo sapiens*

D.R. Bentley
The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom

The finished genome sequence of *Homo sapiens* (Rogers, this volume) provides a starting point for the study of sequence variation in the human population. Every variant that is discovered can be mapped back to the human genome and correlated with genes, regulatory elements, and other functionally important sequences. As we gain a better understanding of the biological information encoded by the human genome sequence, we should aim to define the sequence variants that have biochemical and phenotypic consequences.

[illegible] also targeted

and subsequently spread across the world, replacing earlier *Homo* species. This pattern was originally deduced largely from archaeological and anthropological evidence (Stringer 2002) but received substantial reinforcement from DNA sequence information. For example, genetic variability is generally higher in Africa than on other continents, and phylogenetic reconstructions of non-recombining regions usually place the root in Africa (Cavalli-Sforza and Feldman 2003; Pääbo 2003). A subset of the genetic variants in Africa at the time were therefore present in the migrant founders of all later subsequent

Human SNPs in protein-coding sequences

- Protein-coding sequences account for ~1.5% of the genome, but only ~0.9% of the SNPs (the search is limited to minor allele frequency > 1%). In other words, roughly half of all SNPs in protein-coding sequences have been eliminated by natural selection.
- Protein-coding SNPs in the human population are ~50% synonymous, which suggests that roughly half of the nonsynonymous SNPs have also been eliminated by natural selection.
- Nonsynonymous SNPs in the human population are 66% conservative, and 34% nonconservative, which again suggests that the majority of the nonconservative SNPs have been eliminated by natural selection.
- Taken together, these considerations show that many of the protein-coding SNPs remaining in the human population are not functionally important. Others are known to cause > 2,000 human genetic diseases.

Human SNPs in noncoding sequences

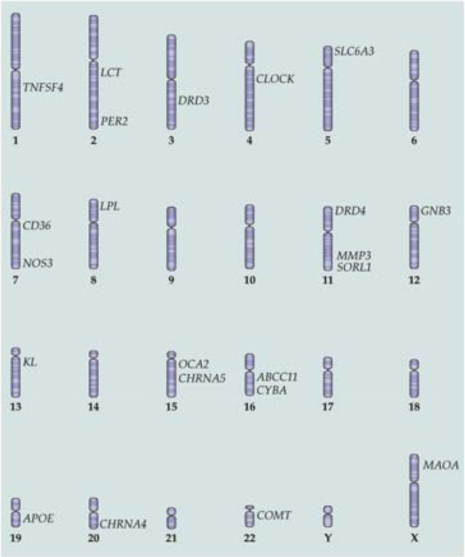
- Sequence comparisons with other species indicate that ~5-10% of the human genome is under natural selection for a conserved function.

- As the majority of these conserved sequences do not encode a protein sequence, it follows that the majority of functional SNPs are likely to be in noncoding DNA.

- Because many of these functional SNPs have been eliminated by natural selection, it follows that < 5% of SNPs in the human genome are likely to be functional.

- These (rare) functional SNPs may be identified (in part) by sequence comparisons (to identify conserved sequences).

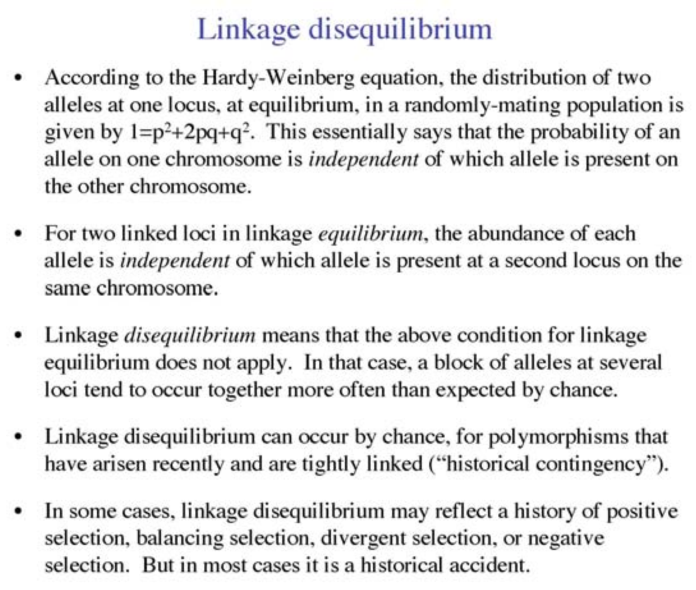
Figure 3.1 Location of polymorphisms that have been associated with various traits in J. C. Venter



A PRIMER OF GENOME SCIENCE 3e, Figure 3.1

Human Genetic Diversity

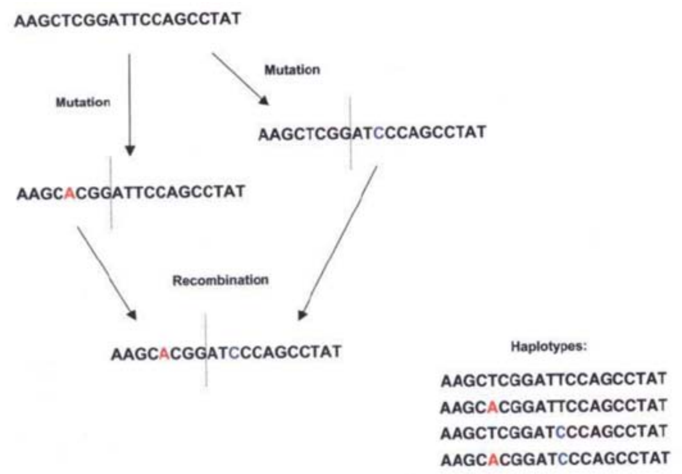
- 90% of human genetic polymorphisms are caused by SNPS; the remaining polymorphisms are structural variants including insertions, deletions, and so on.
- Types of SNPs (Bentley)
- Linkage disequilibrium
- The neutral theory of molecular evolution
- Techniques used to map human traits & score SNPs
- Structural variants & medical applications



Linkage disequilibrium

- According to the Hardy-Weinberg equation, the distribution of two alleles at one locus, at equilibrium, in a randomly-mating population is given by $1=p^2+2pq+q^2$. This essentially says that the probability of an allele on one chromosome is *independent* of which allele is present on the other chromosome.
- For two linked loci in linkage *equilibrium*, the abundance of each allele is *independent* of which allele is present at a second locus on the same chromosome.
- Linkage *disequilibrium* means that the above condition for linkage equilibrium does not apply. In that case, a block of alleles at several loci tend to occur together more often than expected by chance.
- Linkage disequilibrium can occur by chance, for polymorphisms that have arisen recently and are tightly linked ("historical contingency").
- In some cases, linkage disequilibrium may reflect a history of positive selection, balancing selection, divergent selection, or negative selection. But in most cases it is a historical accident.

Mutation and recombination produce haplotypes



AAGCTCGGATTCCAGCCTAT

Mutation

Mutation

AAGCTCGGATCCCAGCCTAT

AAGCACGGATTCCAGCCTAT

Recombination

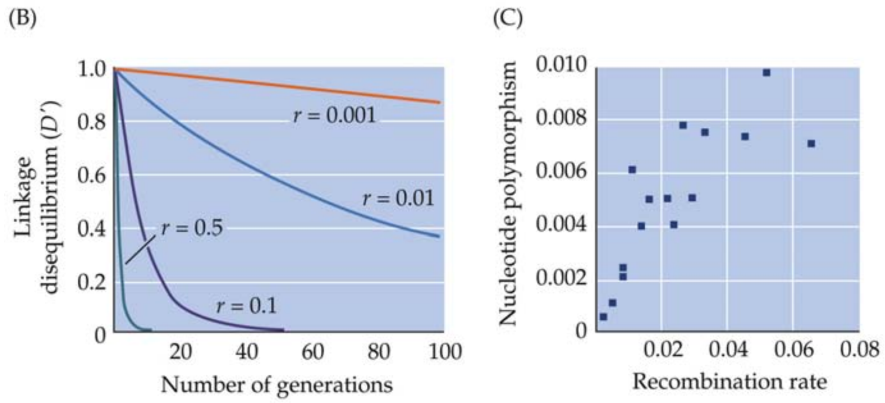
AAGCACGGATCCCAGCCTAT

Haplotypes:

AAGCTCGGATTCCAGCCTAT
AAGCACGGATTCCAGCCTAT
AAGCTCGGATCCCAGCCTAT
AAGCACGGATCCCAGCCTAT

Figure 1. Origin of sequence variation. Sequence variation arises by mutation (*colored bases*) and by recombination (*dotted lines*). These processes give rise to individual haplotypes (listed on the right) that coexist in the population.

Cold Spring Harbor Symposia on Quantitative Biology, Volume LXVIII. © 2003 Cold Spring Harbor Laboratory Press 0-87969-709-1/04. 55



***A PRIMER OF GENOME SCIENCE 3e*, Figure 3.2 (Part 2)**

© 2009 Sinauer Associates, Inc.

D is a simple estimate of linkage disequilibrium (the "linkage disequilibrium coefficient")

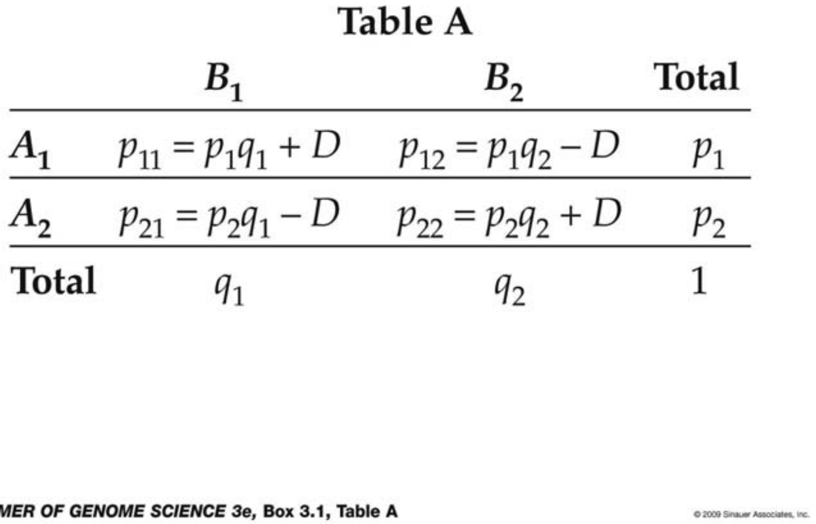


Table A

	B_1	B_2	Total
A_1	$p_{11} = p_1q_1 + D$	$p_{12} = p_1q_2 - D$	p_1
A_2	$p_{21} = p_2q_1 - D$	$p_{22} = p_2q_2 + D$	p_2
Total	q_1	q_2	1

A PRIMER OF GENOME SCIENCE 3e, Box 3.1, Table A

Chromosomal regions of high linkage disequilibrium correspond to chromosomal regions of low recombination frequency

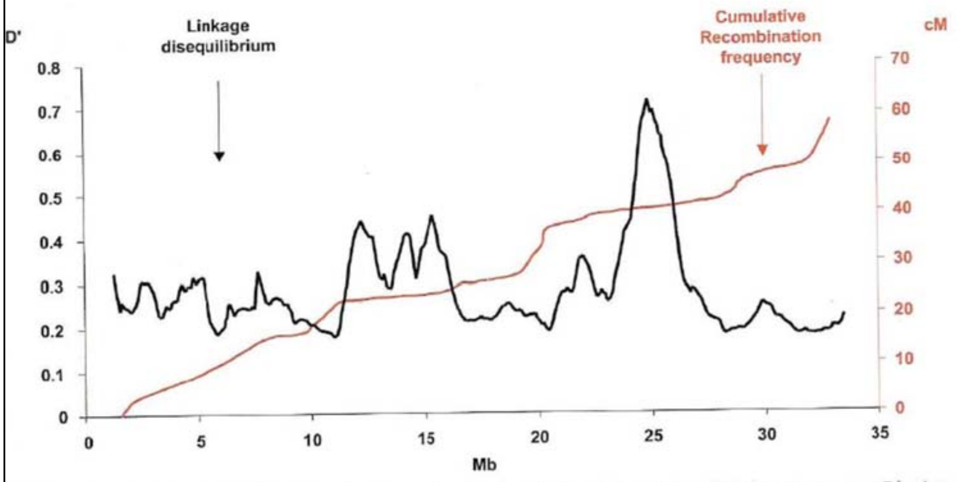


Figure 2. Linkage disequilibrium (LD) and meiotic recombination: Chromosome 22. The LD profile is based on average D' values in sliding windows (see text). LD and cumulative recombination frequency are plotted relative to physical distance along the chromosome, with the telomere of the long arm on the right of the figure.

The number and length of apparent haplotype blocks depend on the spacing between SNP markers used

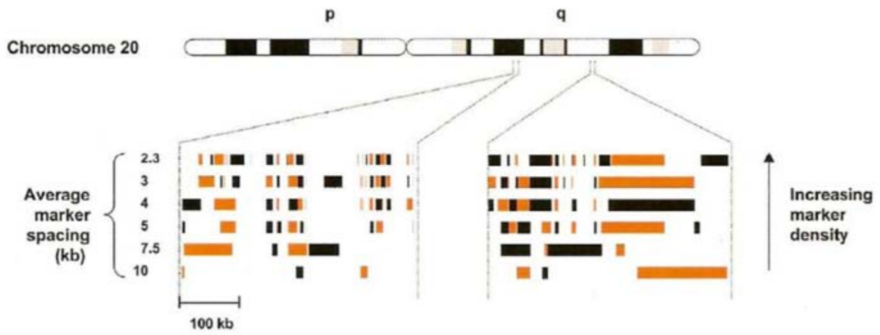
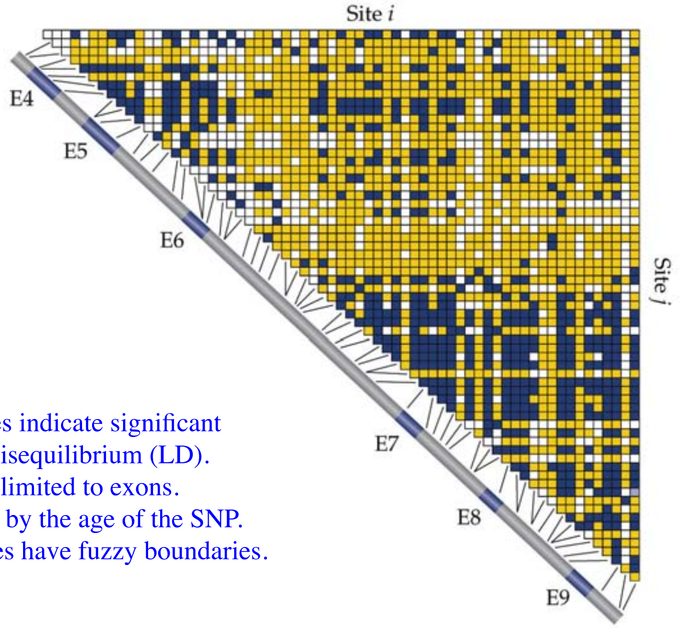


Figure 4. LD analysis of Chromosome 20. Haplotype blocks (*red* and *black boxes*) were computed from LD data on Chromosome 20 and are shown for two regions, one each of high and low overall LD. The analysis is repeated using data from different densities of SNPs (average marker spacings in each analysis are listed on the left of the figure, and increasing SNP density is indicated by the vertical arrow).

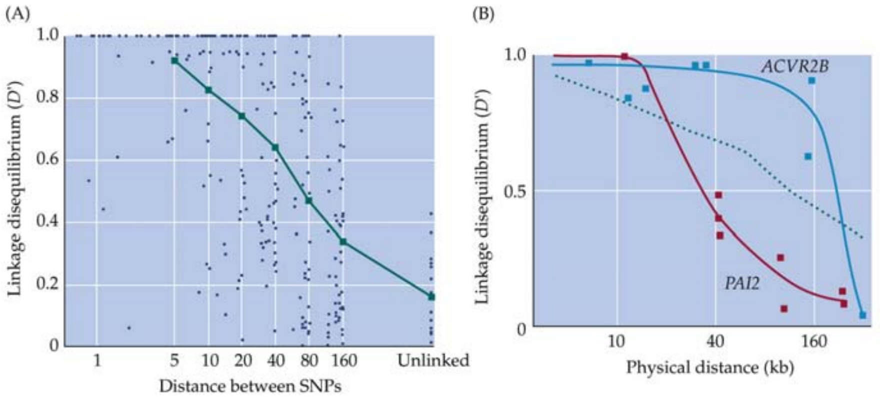
Distribution of linkage disequilibrium across the *lipoprotein lipase* gene



Blue boxes indicate significant
Linkage disequilibrium (LD).
LD is not limited to exons.
LD varies by the age of the SNP.
Haplotypes have fuzzy boundaries.

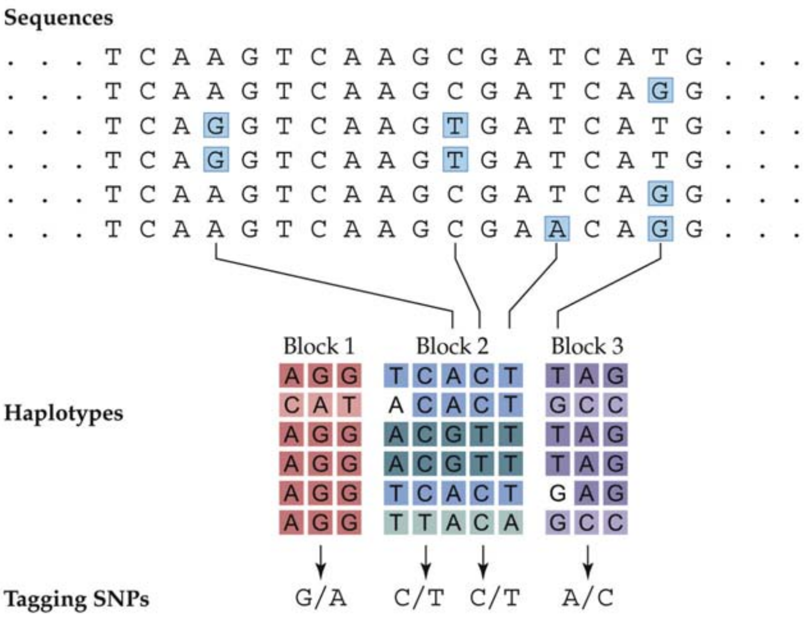
***A PRIMER OF GENOME SCIENCE 3e*, Figure 3.3**

Figure 3.4 Distribution of linkage disequilibrium in the human genome



A PRIMER OF GENOME SCIENCE 3e, Figure 3.4

Tagging SNPs are used to define most of the variation in a haplotype.



Sequences

. . . T C A A G T C A A G C G A T C A T G . . .
. . . T C A A G T C A A G C G A T C A G G . . .
. . . T C A G G T C A A G T G A T C A T G . . .
. . . T C A G G T C A A G T G A T C A T G . . .
. . . T C A A G T C A A G C G A T C A G G . . .
. . . T C A A G T C A A G C G A A C A G G . . .

Haplotypes

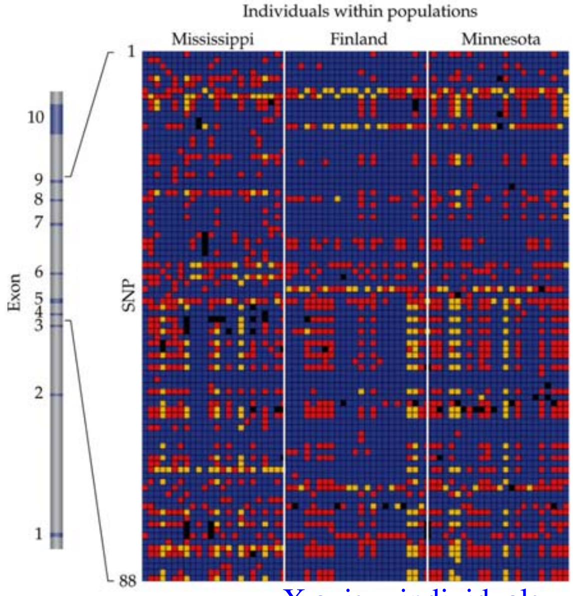
Block 1	Block 2	Block 3
A G G	T C A C T	T A G
C A T	A C A C T	G C C
A G G	A C G T T	T A G
A G G	A C G T T	T A G
A G G	T C A C T	G A G
A G G	T T A C A	G C C

Tagging SNPs

G/A C/T C/T A/C

A PRIMER OF GENOME SCIENCE 3e, Figure 3.5

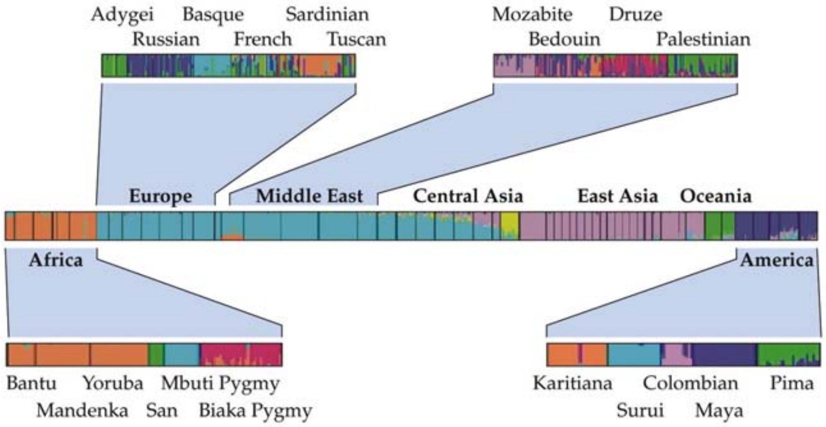
Figure 3.6 Haplotype structure in the human *lipoprotein lipase* gene (homozygotes for the common allele in blue, heterozygotes red, homozygotes for the rare allele in yellow). Two main haplotypes.



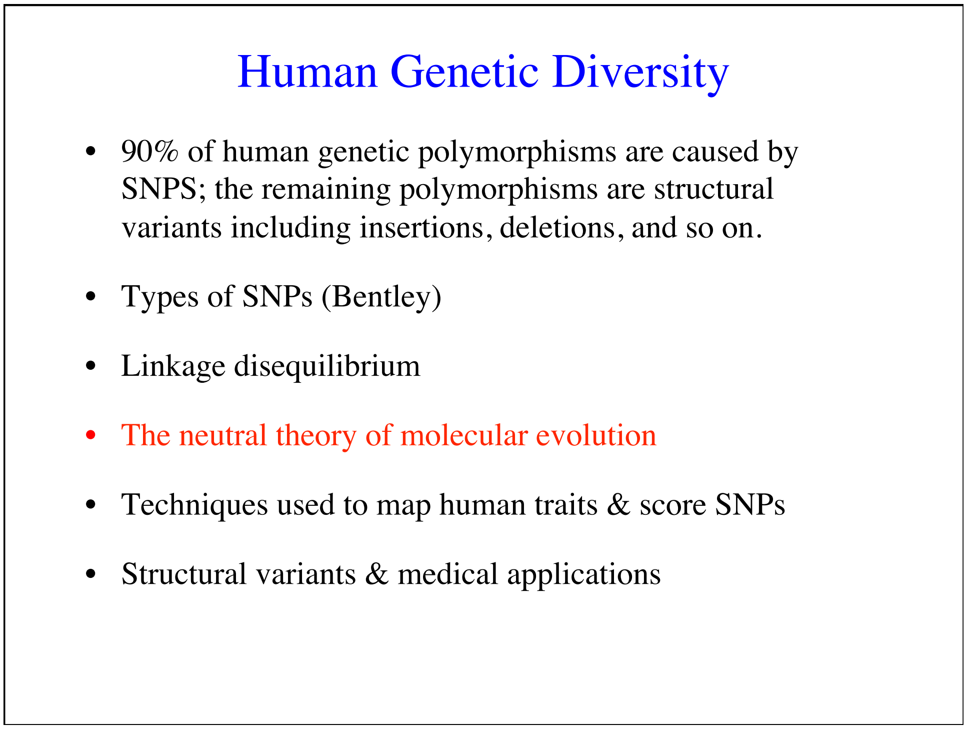
A PRIMER OF GENOME SCIENCE 3e, Figure 3.6

X axis = individuals

Figure 3.7 Human diversity and population structure

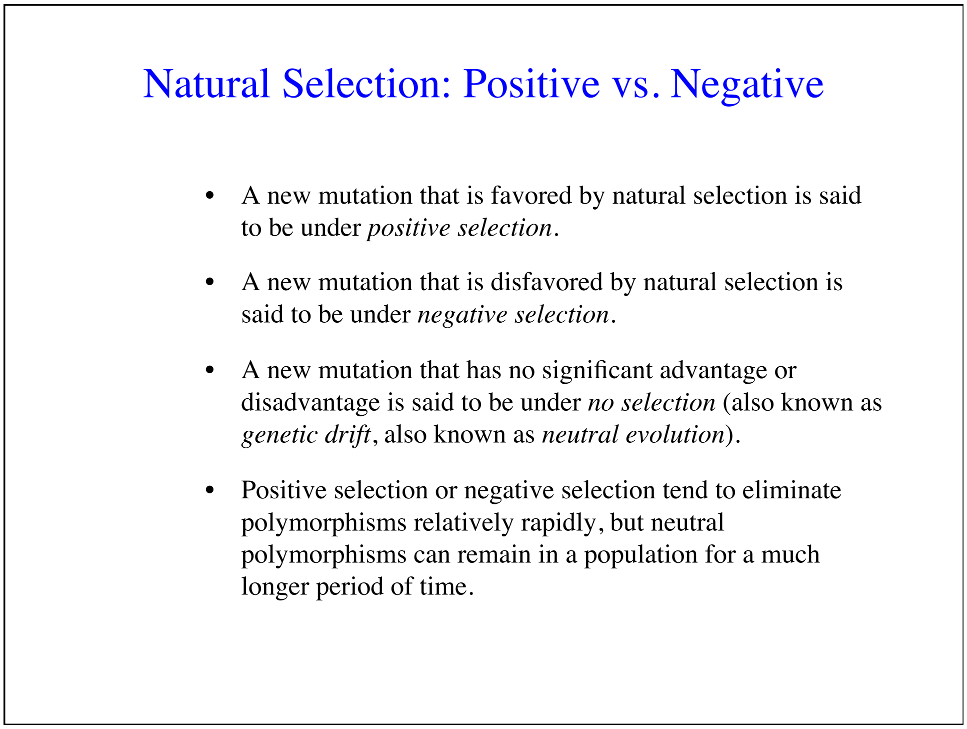


A PRIMER OF GENOME SCIENCE 3e, Figure 3.7



Human Genetic Diversity

- 90% of human genetic polymorphisms are caused by SNPS; the remaining polymorphisms are structural variants including insertions, deletions, and so on.
- Types of SNPs (Bentley)
- Linkage disequilibrium
- The neutral theory of molecular evolution
- Techniques used to map human traits & score SNPs
- Structural variants & medical applications

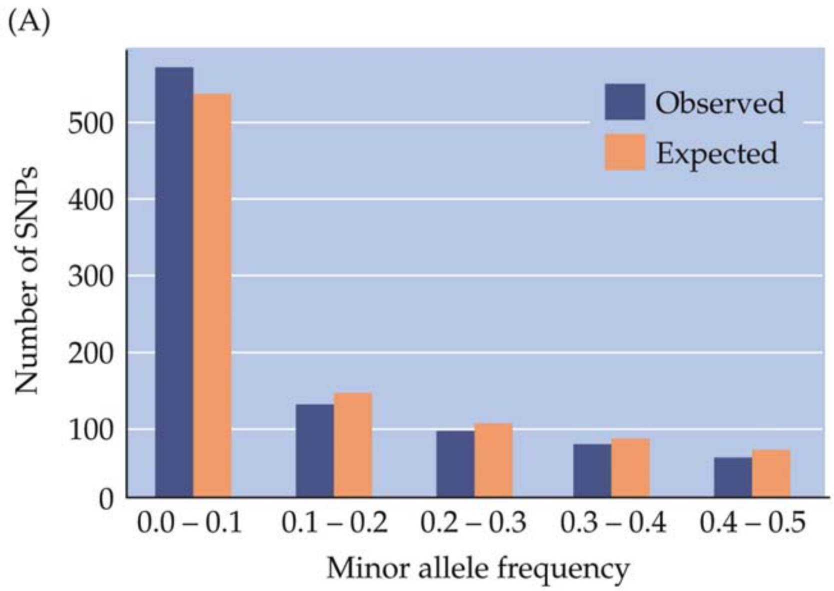


Natural Selection: Positive vs. Negative

- A new mutation that is favored by natural selection is said to be under *positive selection*.
- A new mutation that is disfavored by natural selection is said to be under *negative selection*.
- A new mutation that has no significant advantage or disadvantage is said to be under *no selection* (also known as *genetic drift*, also known as *neutral evolution*).
- Positive selection or negative selection tend to eliminate polymorphisms relatively rapidly, but neutral polymorphisms can remain in a population for a much longer period of time.

The neutral theory of molecular evolution

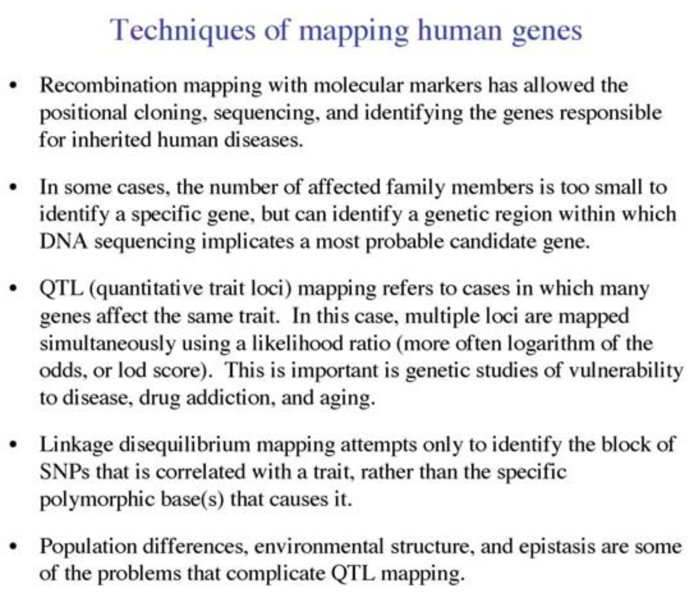
- The neutral theory postulates that the majority of all SNPs do not confer a significant selective advantage or disadvantage.
- Under the neutral theory, the majority of evolutionary sequence change is caused by random fluctuations in allele frequencies, which eventually cause particular SNPs to become homozygous throughout a population and thus "fixed" in the species.
- The neutral theory has been successful in many cases, particularly in explaining why there are so many noncoding and synonymous SNPs.
- However, the neutral theory does not apply to every case, particularly non-synonymous SNPs and promoter SNPs.
- There are many tests of the neutral theory - the simplest is the McDonald-Kreitman test, based on a 2x2 contingency test of synonymous & non-synonymous polymorphisms within a species, vs. synonymous & non-synonymous substitutions between two species.



***A PRIMER OF GENOME SCIENCE 3e*, Figure 3.2 (Part 1)**

Human Genetic Diversity

- 90% of human genetic polymorphisms are caused by SNPS; the remaining polymorphisms are structural variants including insertions, deletions, and so on.
- Types of SNPs (Bentley)
- Linkage disequilibrium
- The neutral theory of molecular evolution
- Techniques used to map human traits & score SNPs
- Structural variants & medical applications



Techniques of mapping human genes

- Recombination mapping with molecular markers has allowed the positional cloning, sequencing, and identifying the genes responsible for inherited human diseases.
- In some cases, the number of affected family members is too small to identify a specific gene, but can identify a genetic region within which DNA sequencing implicates a most probable candidate gene.
- QTL (quantitative trait loci) mapping refers to cases in which many genes affect the same trait. In this case, multiple loci are mapped simultaneously using a likelihood ratio (more often logarithm of the odds, or lod score). This is important is genetic studies of vulnerability to disease, drug addiction, and aging.
- Linkage disequilibrium mapping attempts only to identify the block of SNPs that is correlated with a trait, rather than the specific polymorphic base(s) that causes it.
- Population differences, environmental structure, and epistasis are some of the problems that complicate QTL mapping.

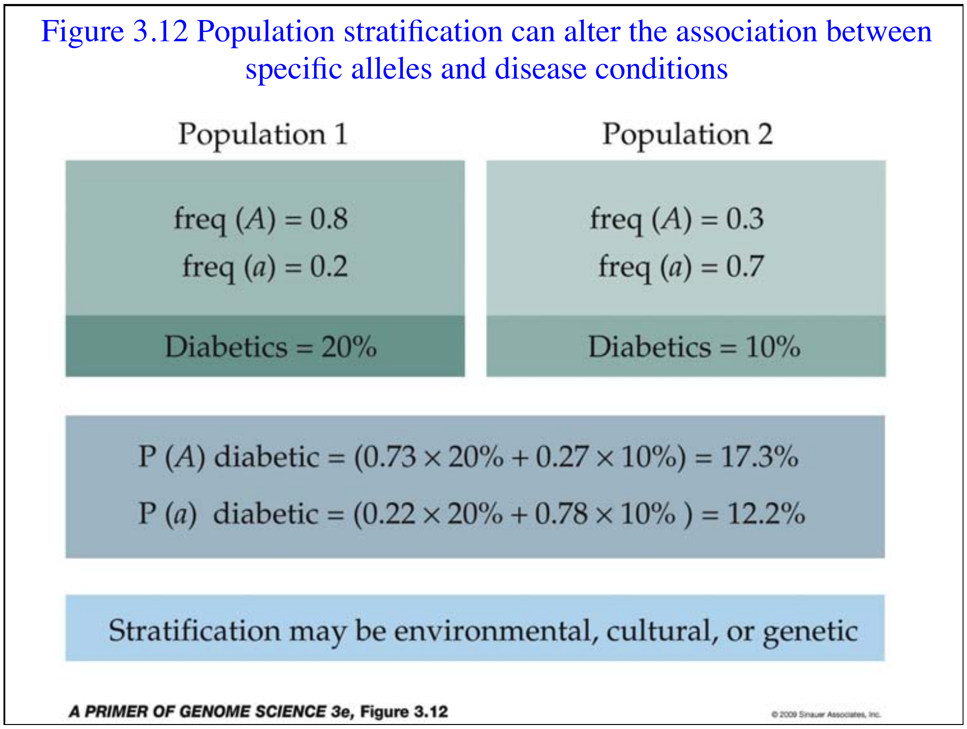


Figure 3.12 Population stratification can alter the association between specific alleles and disease conditions

Population 1	Population 2
freq (A) = 0.8 freq (a) = 0.2	freq (A) = 0.3 freq (a) = 0.7
Diabetics = 20%	Diabetics = 10%

P (A) diabetic = (0.73 × 20% + 0.27 × 10%) = 17.3%

P (a) diabetic = (0.22 × 20% + 0.78 × 10%) = 12.2%

Stratification may be environmental, cultural, or genetic

A PRIMER OF GENOME SCIENCE 3e, Figure 3.12

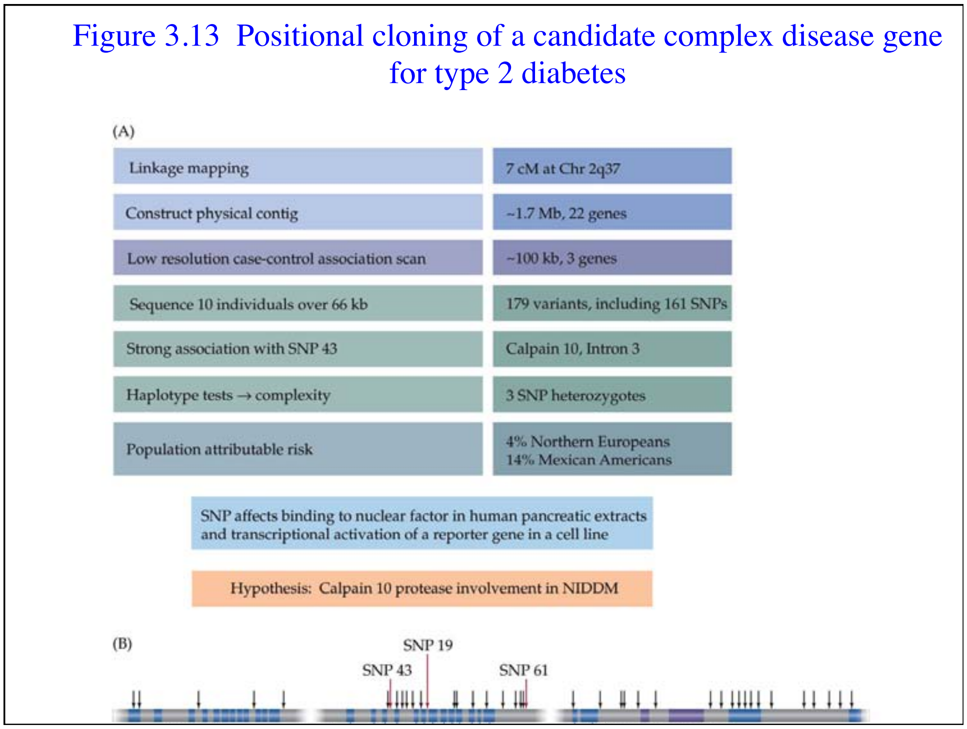


Figure 3.13 Positional cloning of a candidate complex disease gene for type 2 diabetes

(A)

Linkage mapping	7 cM at Chr 2q37
Construct physical contig	~1.7 Mb, 22 genes
Low resolution case-control association scan	~100 kb, 3 genes
Sequence 10 individuals over 66 kb	179 variants, including 161 SNPs
Strong association with SNP 43	Calpain 10, Intron 3
Haplotype tests → complexity	3 SNP heterozygotes
Population attributable risk	4% Northern Europeans 14% Mexican Americans

SNP affects binding to nuclear factor in human pancreatic extracts and transcriptional activation of a reporter gene in a cell line

Hypothesis: Calpain 10 protease involvement in NIDDM

(B)

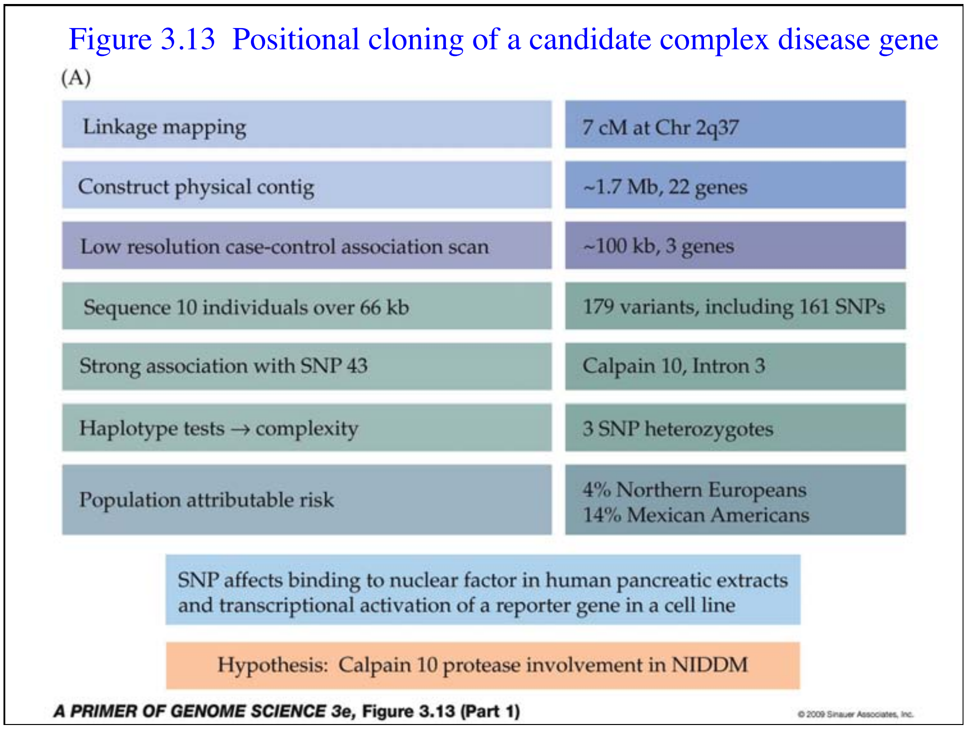


Figure 3.13 Positional cloning of a candidate complex disease gene

(A)

Linkage mapping	7 cM at Chr 2q37
Construct physical contig	~1.7 Mb, 22 genes
Low resolution case-control association scan	~100 kb, 3 genes
Sequence 10 individuals over 66 kb	179 variants, including 161 SNPs
Strong association with SNP 43	Calpain 10, Intron 3
Haplotype tests → complexity	3 SNP heterozygotes
Population attributable risk	4% Northern Europeans 14% Mexican Americans

SNP affects binding to nuclear factor in human pancreatic extracts and transcriptional activation of a reporter gene in a cell line

Hypothesis: Calpain 10 protease involvement in NIDDM

***A PRIMER OF GENOME SCIENCE 3e*, Figure 3.13 (Part 1)**

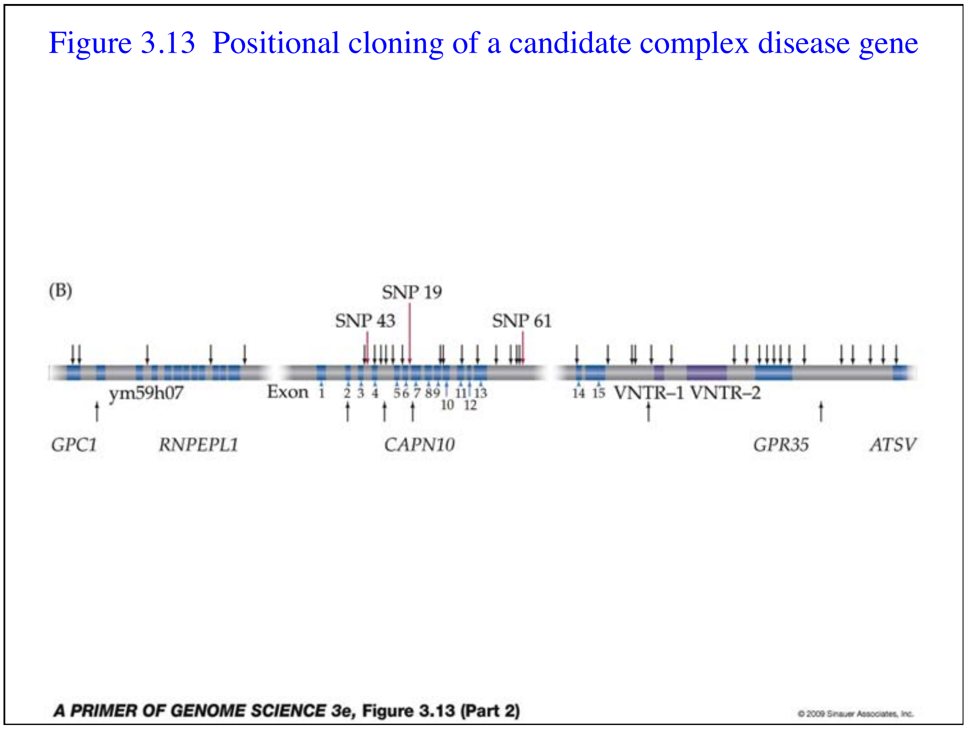


Figure 3.13 Positional cloning of a candidate complex disease gene

***A PRIMER OF GENOME SCIENCE 3e*, Figure 3.13 (Part 2)**

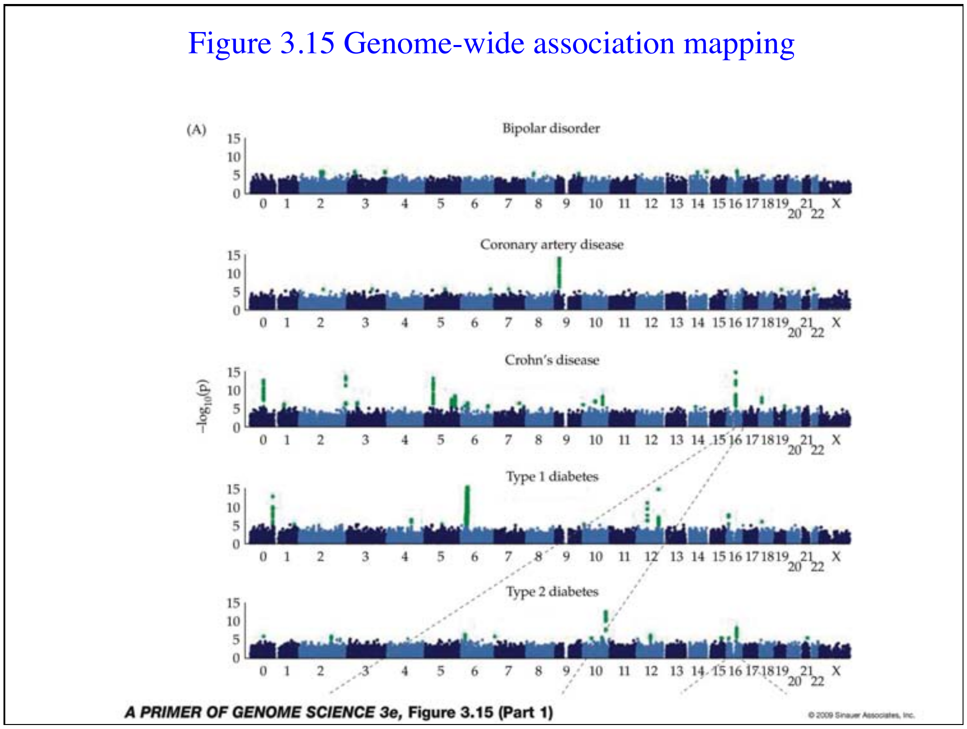
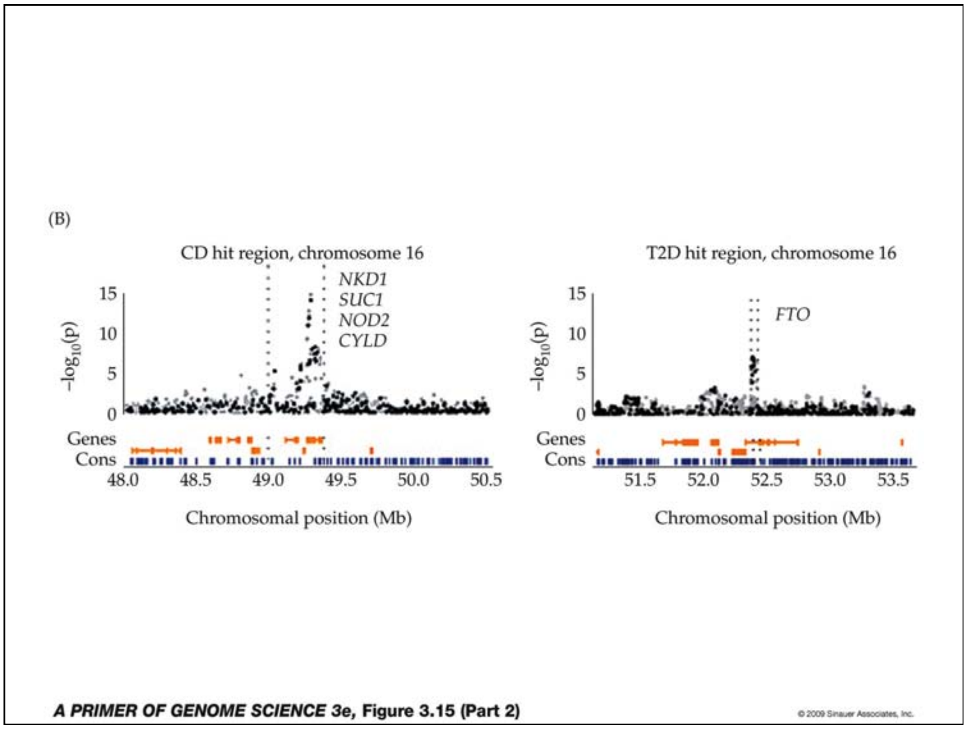


Figure 3.15 Genome-wide association mapping

A PRIMER OF GENOME SCIENCE 3e, Figure 3.15 (Part 1)

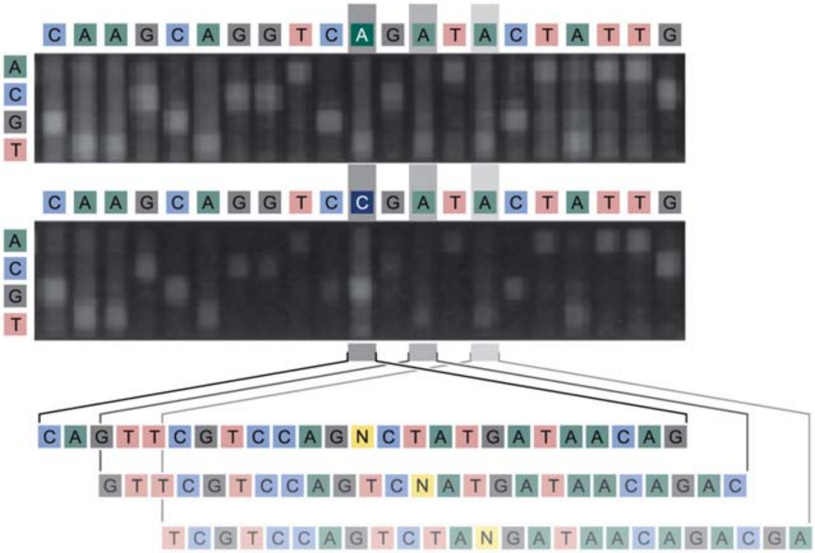


A PRIMER OF GENOME SCIENCE 3e, Figure 3.15 (Part 2)

SNP methods

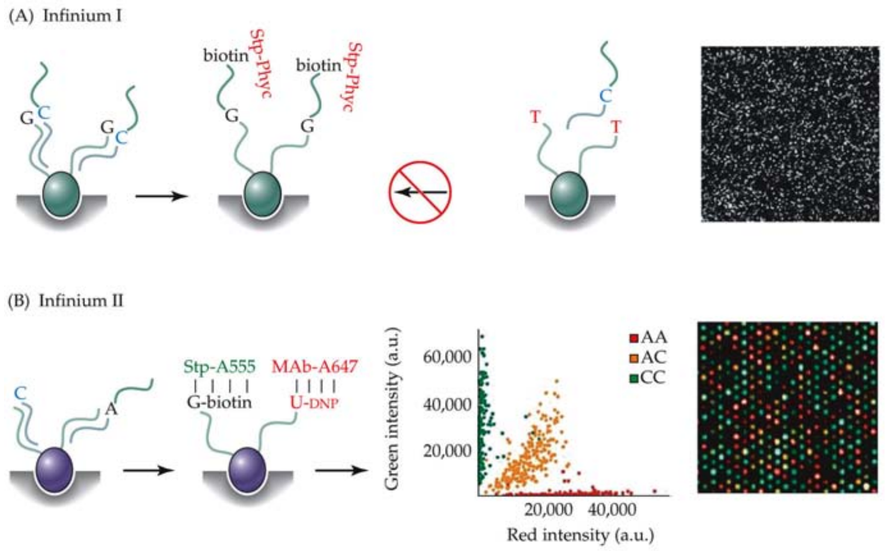
- Computational - millions of SNPs, along with their frequencies in the major ethnic groups, are already known. Thus it is possible to “choose your SNP” in this way.
- Allele-specific oligonucleotide hybridization - high throughput but can have data normalization problems.
- Illumina bead assay: add one base with fluorescent tag to a microarray. Very high throughput, very good accuracy.
- Pyrosequencing - add one base at a time, measure light flash, wash out and cycle with another base. Similar to above but more complicated.
- RT-PCR methods (SYBR green, Taq Man, etc) cost-effective, accurate, and flexible. Not high-throughput.
- RFLP electrophoresis methods - require that a restriction site exist or be engineered. Accuracy is questionable, because of variable PCR yields and incomplete restriction digests.
- Mass spectrometry - moderate throughput, excellent accuracy.
- DNA sequencing - expensive, slow, accurate. The only good way to map haplotype blocks (phasing).

Figure 3.16 Sequencing by hybridization - similar to Affymetrix SNP chips


C A A G C A G G T C A G A T A C T A T T G
A
C
G
T
C A A G C A G G T C C G A T A C T A T T G
A
C
G
T
C A G T T C G T C C A G N C T A T G A T A A C A G
G T T C G T C C A G T C N A T G A T A A C A G A C
T C G T C C A G T C T A N G A T A A C A G A C G A

A PRIMER OF GENOME SCIENCE 3e, Figure 3.16

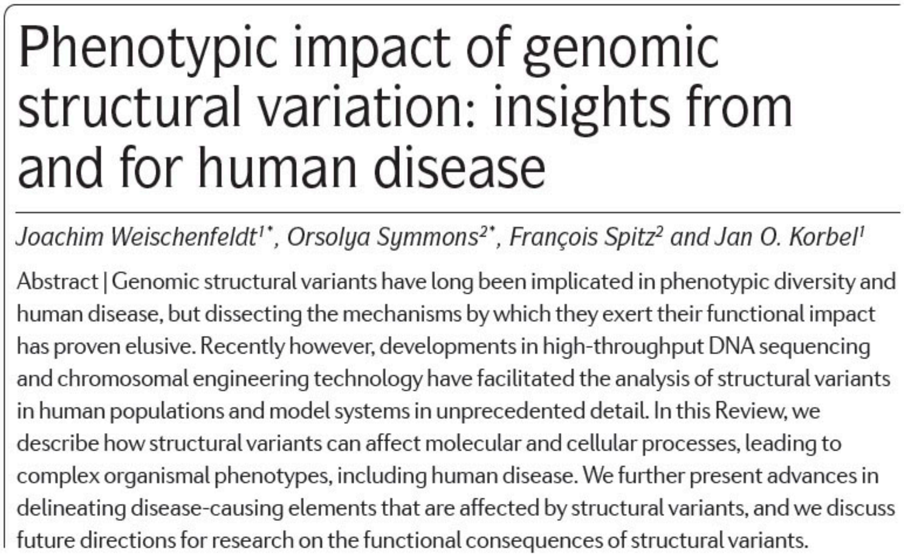
Figure 3.17 The Illumina Infinium I and II genotyping assays



A PRIMER OF GENOME SCIENCE 3e, **Figure 3.17**

Human Genetic Diversity

- 90% of human genetic polymorphisms are caused by SNPS; the remaining polymorphisms are structural variants including insertions, deletions, and so on.
- Types of SNPs (Bentley)
- Linkage disequilibrium
- The neutral theory of molecular evolution
- Techniques used to map human traits & score SNPs
- Structural variants & medical applications

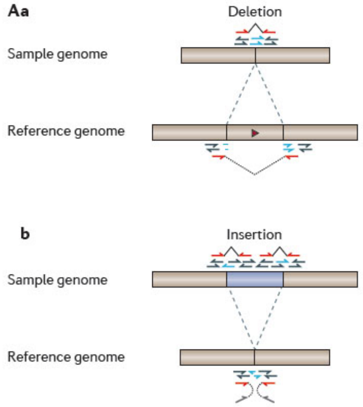


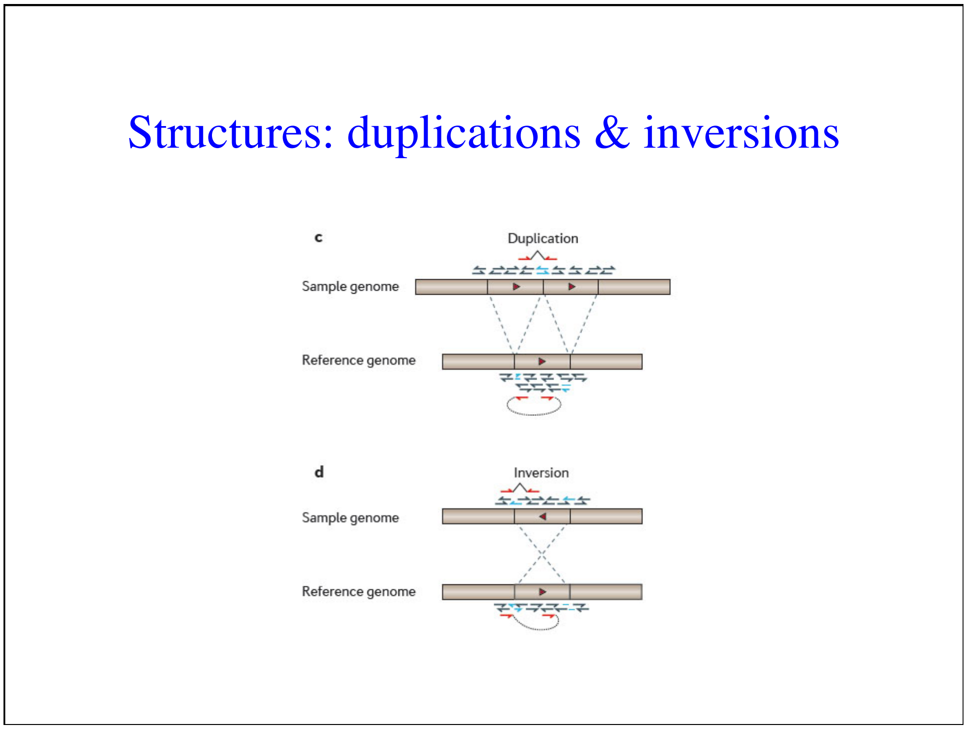
Phenotypic impact of genomic structural variation: insights from and for human disease

Joachim Weischenfeldt[1], Orsolya Symmons[2]*, François Spitz[2] and Jan O. Korbel[1]*

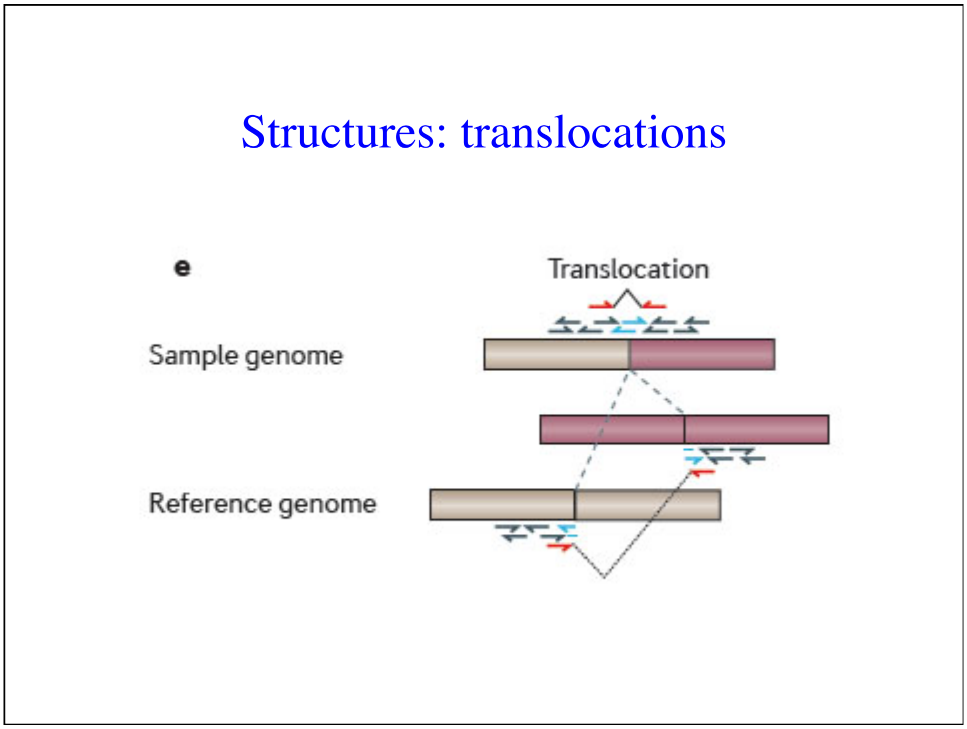
Abstract | Genomic structural variants have long been implicated in phenotypic diversity and human disease, but dissecting the mechanisms by which they exert their functional impact has proven elusive. Recently however, developments in high-throughput DNA sequencing and chromosomal engineering technology have facilitated the analysis of structural variants in human populations and model systems in unprecedented detail. In this Review, we describe how structural variants can affect molecular and cellular processes, leading to complex organismal phenotypes, including human disease. We further present advances in delineating disease-causing elements that are affected by structural variants, and we discuss future directions for research on the functional consequences of structural variants.

Structures: insertions & deletions

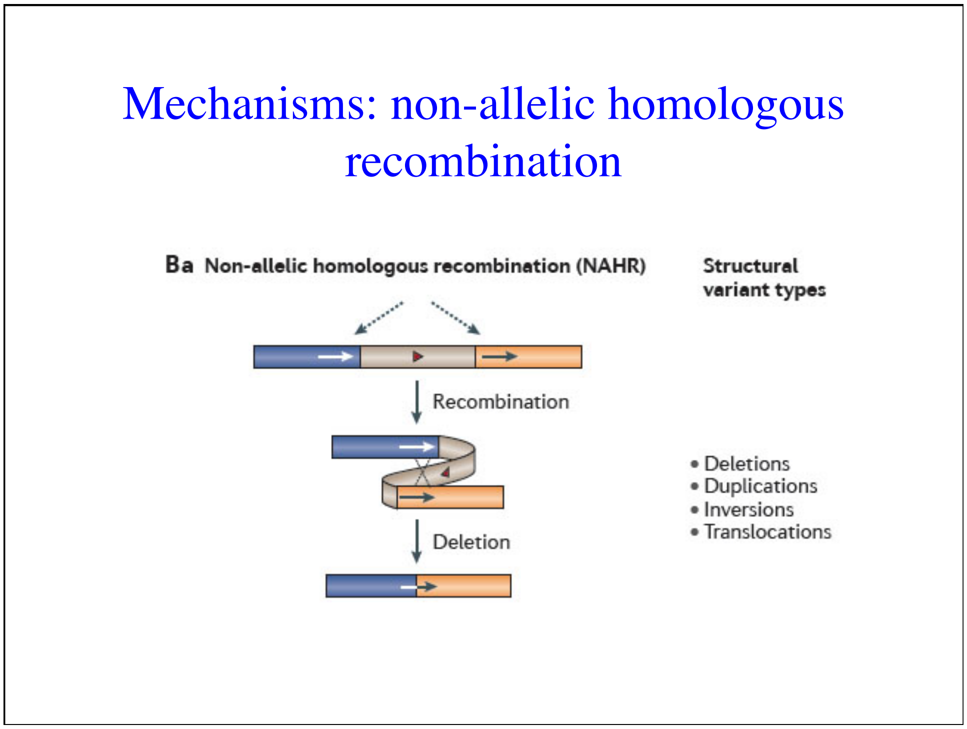




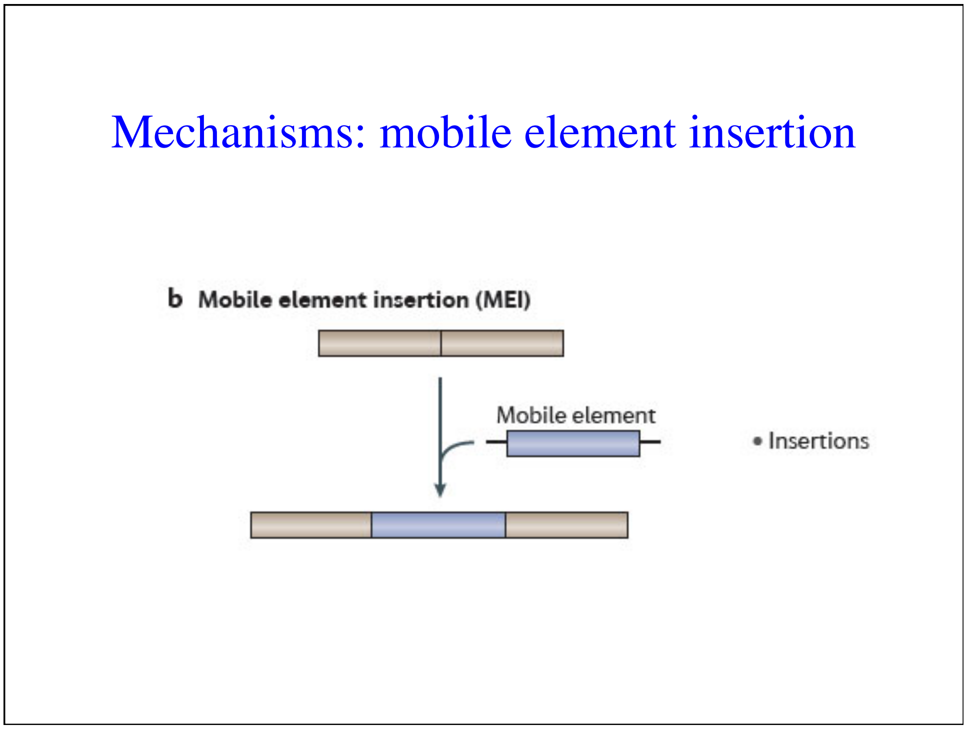
Structures: duplications & inversions



Structures: translocations



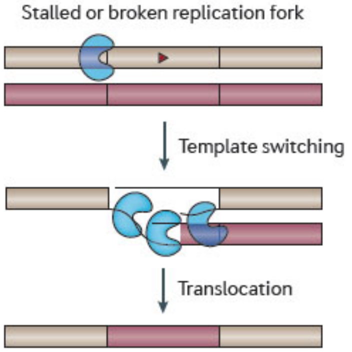
Mechanisms: non-allelic homologous recombination



Mechanisms: mobile element insertion

Mechanisms: replication template switching

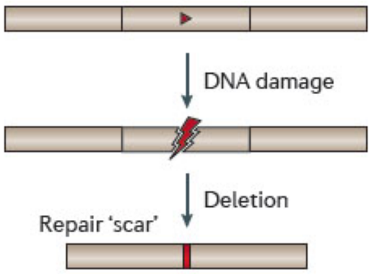
c Replication-based template switching (FoSTeS or MMBIR)



- Deletions
- Duplications
- Inversions
- Translocations
- Complex structural variants

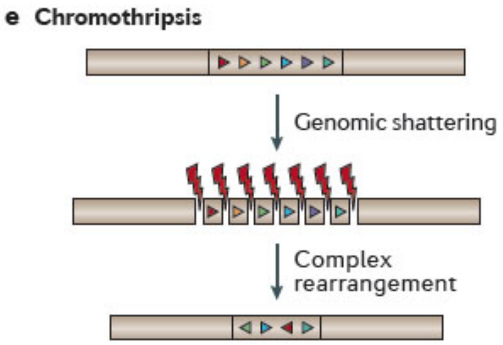
Mechanisms: non-homologous end joining

d Non-homologous end joining (NHEJ)



- Deletions
- Duplications
- Translocations

Mechanisms: chromosome shattering (chromothripsis)



- Deletions
- Inversions
- Translocations
- Complex structural variants

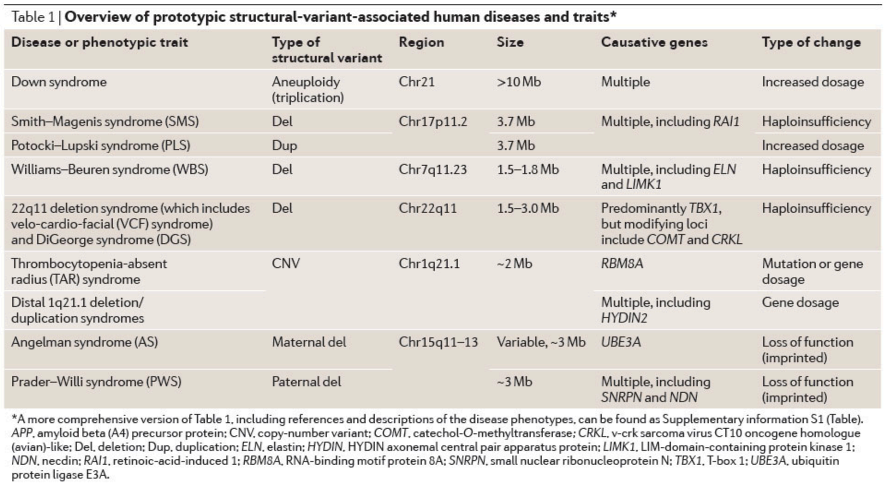
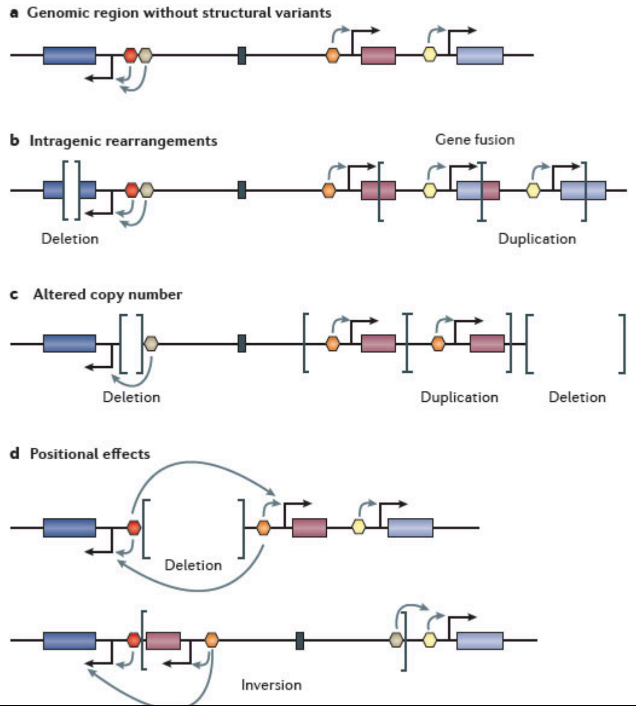
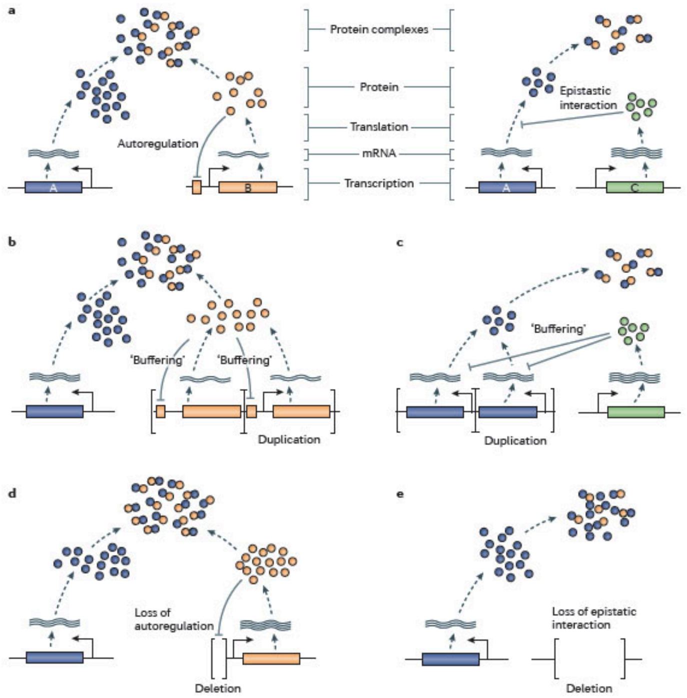


Table 1 | **Overview of prototypic structural-variant-associated human diseases and traits***

Disease or phenotypic trait	Type of structural variant	Region	Size	Causative genes	Type of change
Down syndrome	Aneuploidy (triplication)	Chr21	>10 Mb	Multiple	Increased dosage
Smith–Magenis syndrome (SMS)	Del	Chr17p11.2	3.7 Mb	Multiple, including *RAI1*	Haploinsufficiency
Potocki–Lupski syndrome (PLS)	Dup		3.7 Mb		Increased dosage
Williams–Beuren syndrome (WBS)	Del	Chr7q11.23	1.5–1.8 Mb	Multiple, including *ELN* and *LIMK1*	Haploinsufficiency
22q11 deletion syndrome (which includes velo-cardio-facial (VCF) syndrome) and DiGeorge syndrome (DGS)	Del	Chr22q11	1.5–3.0 Mb	Predominantly *TBX1*, but modifying loci include *COMT* and *CRKL*	Haploinsufficiency
Thrombocytopenia-absent radius (TAR) syndrome	CNV	Chr1q21.1	~2 Mb	*RBM8A*	Mutation or gene dosage
Distal 1q21.1 deletion/ duplication syndromes				Multiple, including *HYDIN2*	Gene dosage
Angelman syndrome (AS)	Maternal del	Chr15q11–13	Variable, ~3 Mb	*UBE3A*	Loss of function (imprinted)
Prader–Willi syndrome (PWS)	Paternal del		~3 Mb	Multiple, including *SNRPN* and *NDN*	Loss of function (imprinted)

*A more comprehensive version of Table 1, including references and descriptions of the disease phenotypes, can be found as Supplementary information S1 (Table). *APP*, amyloid beta (A4) precursor protein; CNV, copy-number variant; *COMT*, catechol-*O*-methyltransferase; *CRKL*, v-crk sarcoma virus CT10 oncogene homologue (avian)-like; Del, deletion; Dup, duplication; *ELN*, elastin; *HYDIN*, HYDIN axonemal central pair apparatus protein; *LIMK1*, LIM-domain-containing protein kinase 1; *NDN*, necdin; *RAI1*, retinoic-acid-induced 1; *RBM8A*, RNA-binding motif protein 8A; *SNRPN*, small nuclear ribonucleoprotein N; *TBX1*, T-box 1; *UBE3A*, ubiquitin protein ligase E3A.


a Genomic region without structural variants
b Intragenic rearrangements
Gene fusion
Deletion
Duplication
c Altered copy number
Deletion
Duplication
Deletion
d Positional effects
Deletion
Inversion


a
Protein complexes
Protein
Translation
mRNA
Transcription
Autoregulation
A
B
Epistastic interaction
A
C
b
'Buffering'
'Buffering'
Duplication
c
'Buffering'
Duplication
d
Loss of autoregulation
Deletion
e
Loss of epistatic interaction
Deletion

Discussion Questions - week 4

- Discuss several reasons why the length of haplotypes (blocks of linkage disequilibrium) would be expected to vary between human populations. Would you expect them to be longer (on average) in Europe or in Africa? Why? How could this be advantageous (or disadvantageous) for finding human disease genes?
- Discuss the evidence for negative selection acting on human protein-coding SNPs, in terms of the observed numbers of coding vs. noncoding SNPs, synonymous vs. nonsynonymous SNPs, and conservative nonsynonymous vs. nonconservative nonsynonymous SNPs.
- Why are base substitutions (SNPs) about 10 times more common in the genome than insertions and deletions (indels)? Does it follow that most functional human genetic diversity is caused by single base substitutions?
- What are some of the methods that could be used to identify a functional SNP (as opposed to linked but nonfunctional SNPs)?
- Why is linkage disequilibrium so important for population and quantitative genetic analysis? Why is it essential that SNP association studies be replicated? Does failure to replicate a finding mean that the original study was incorrect?

Discussion Questions - week 4 (continued)

- Discuss some of the mechanisms by which structural variations can cause human disease.
- A critical issue in analyzing copy number variations (CNVs) in the human genome is the extent to which specific genes are haploinsufficient, vs. dosage-sensitive, vs. neither. Define these terms. Which is more common? Why?