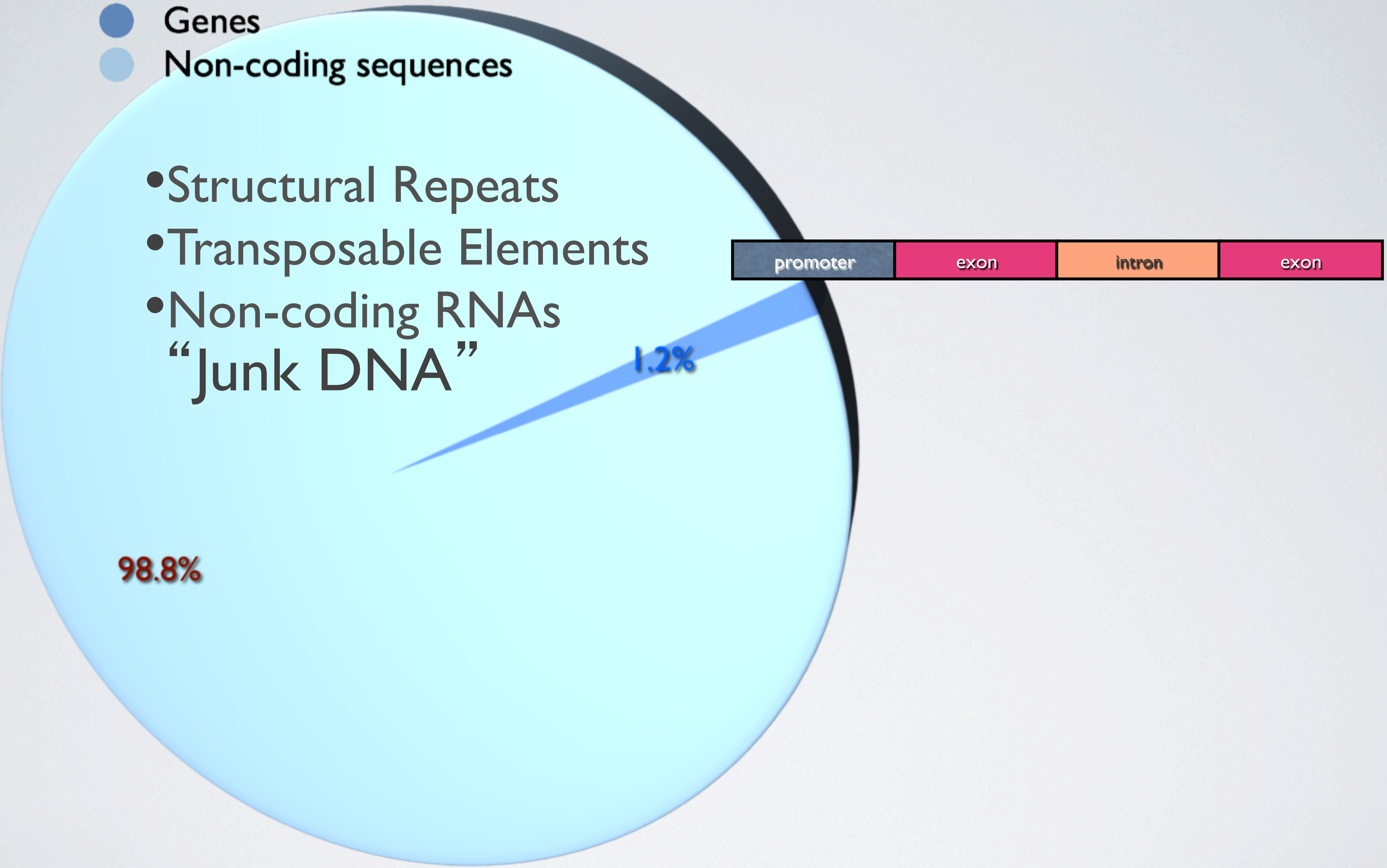
Approaches to Gene Discovery

Bruce R. Korf, MD, PhD

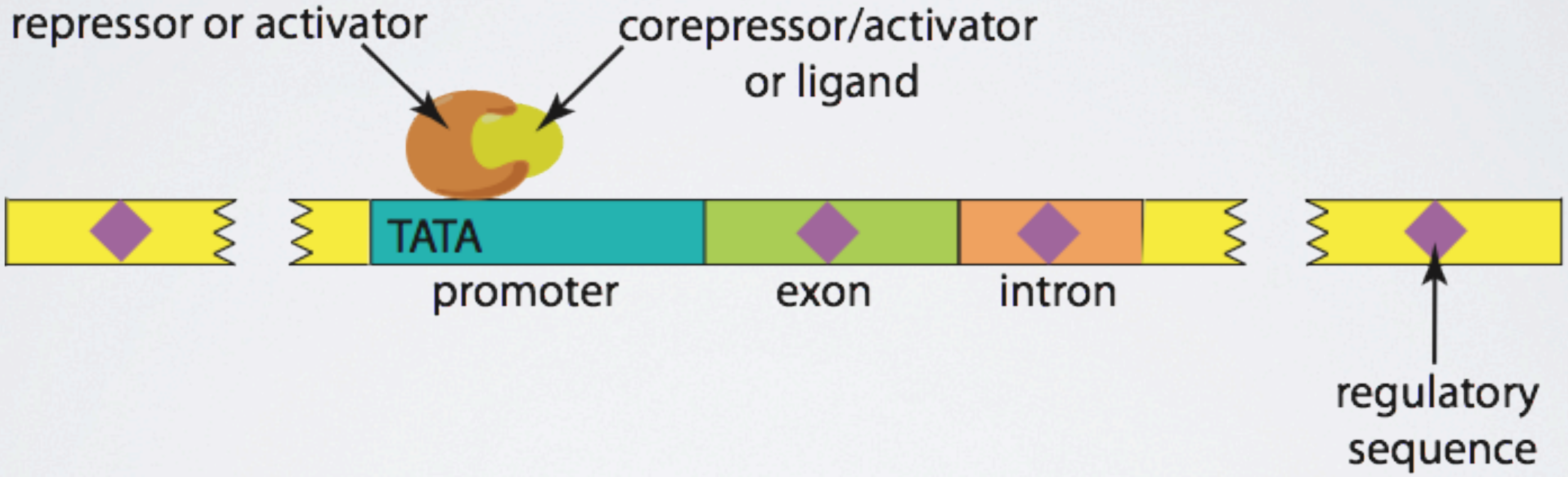
- The Human Genome
- Genetic Variation
- Gene Identification



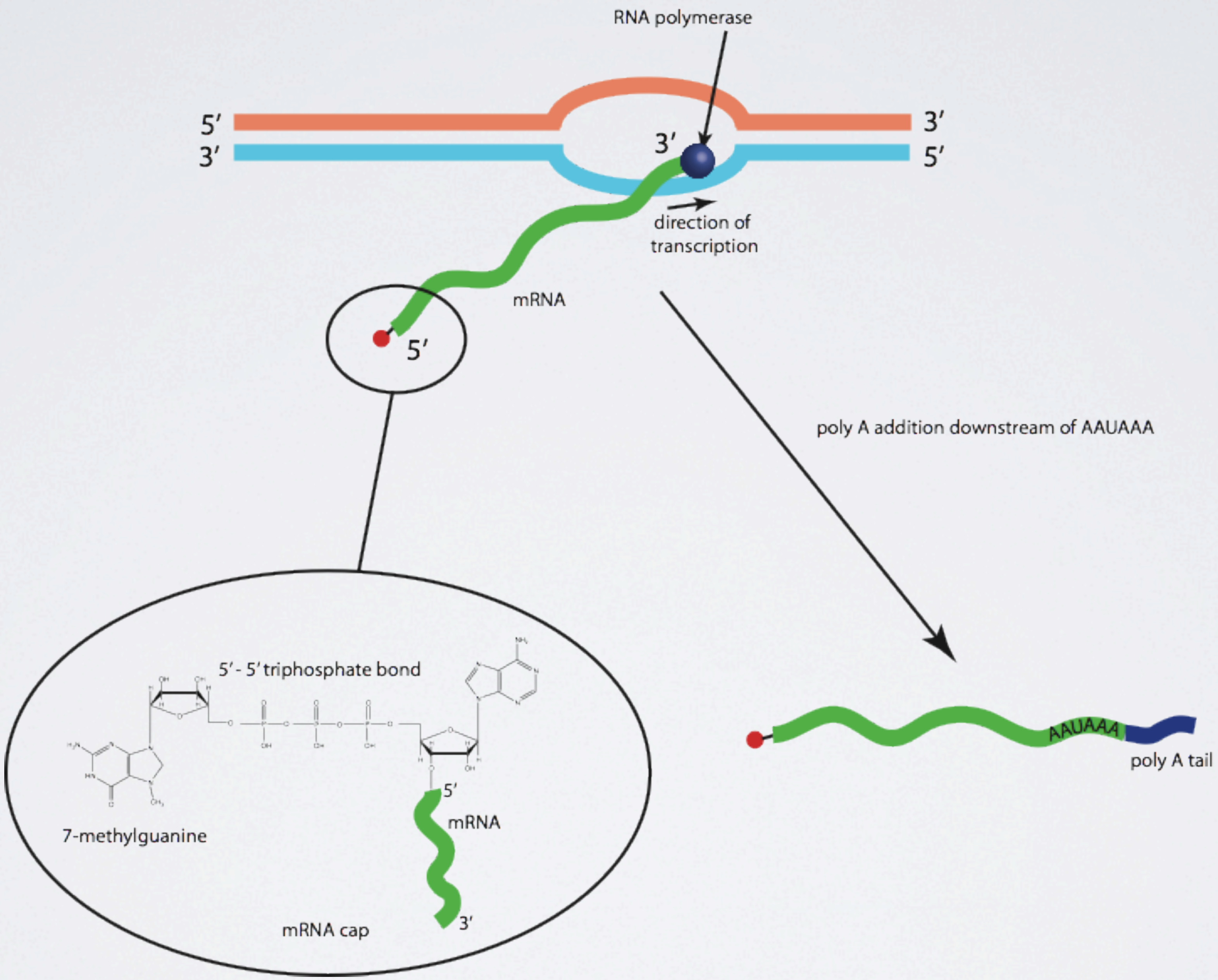
- Structural Repeats
- Transposable Elements
- Non-coding RNAs

"Junk DNA"

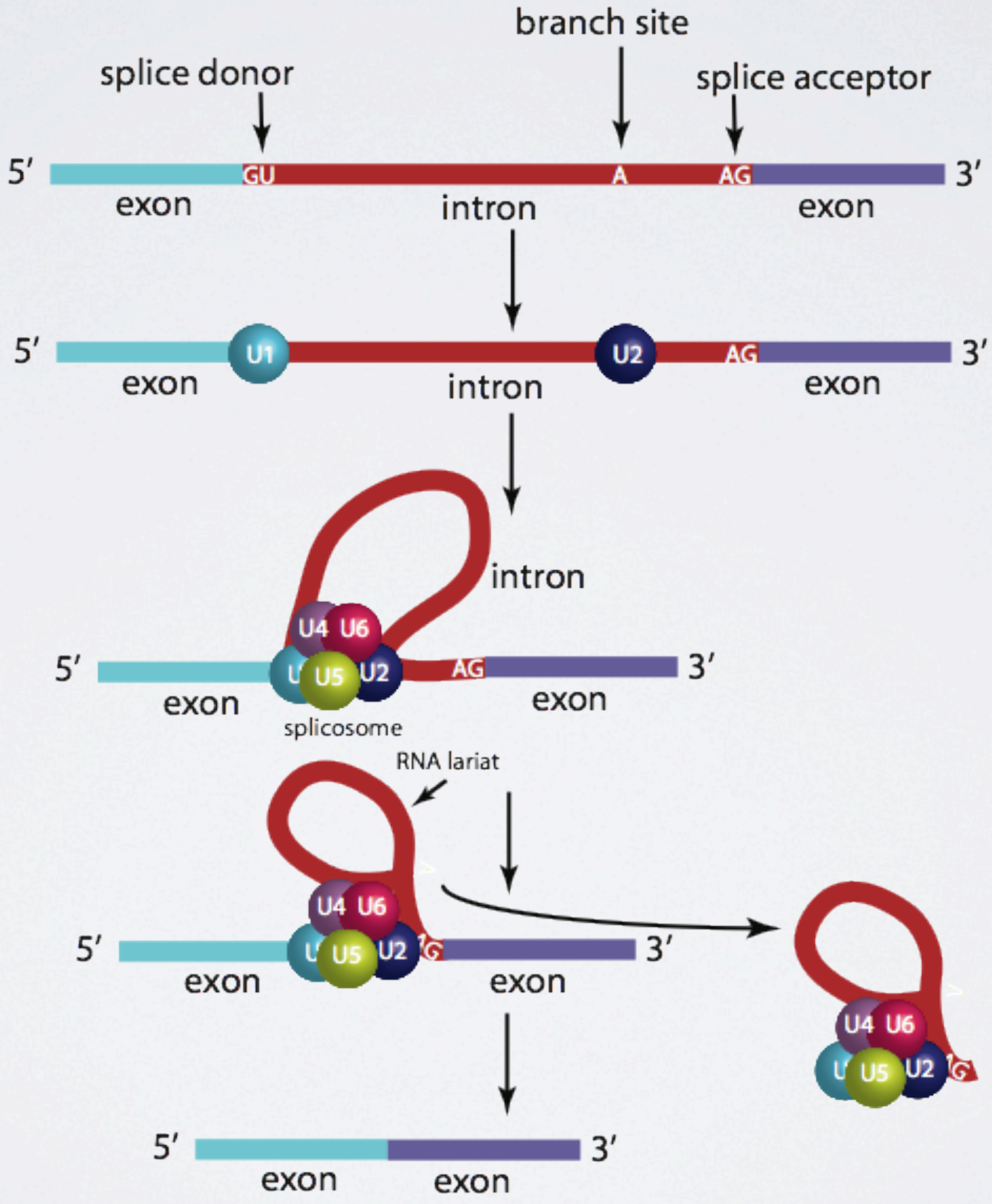
Gene Regulation



Transcription

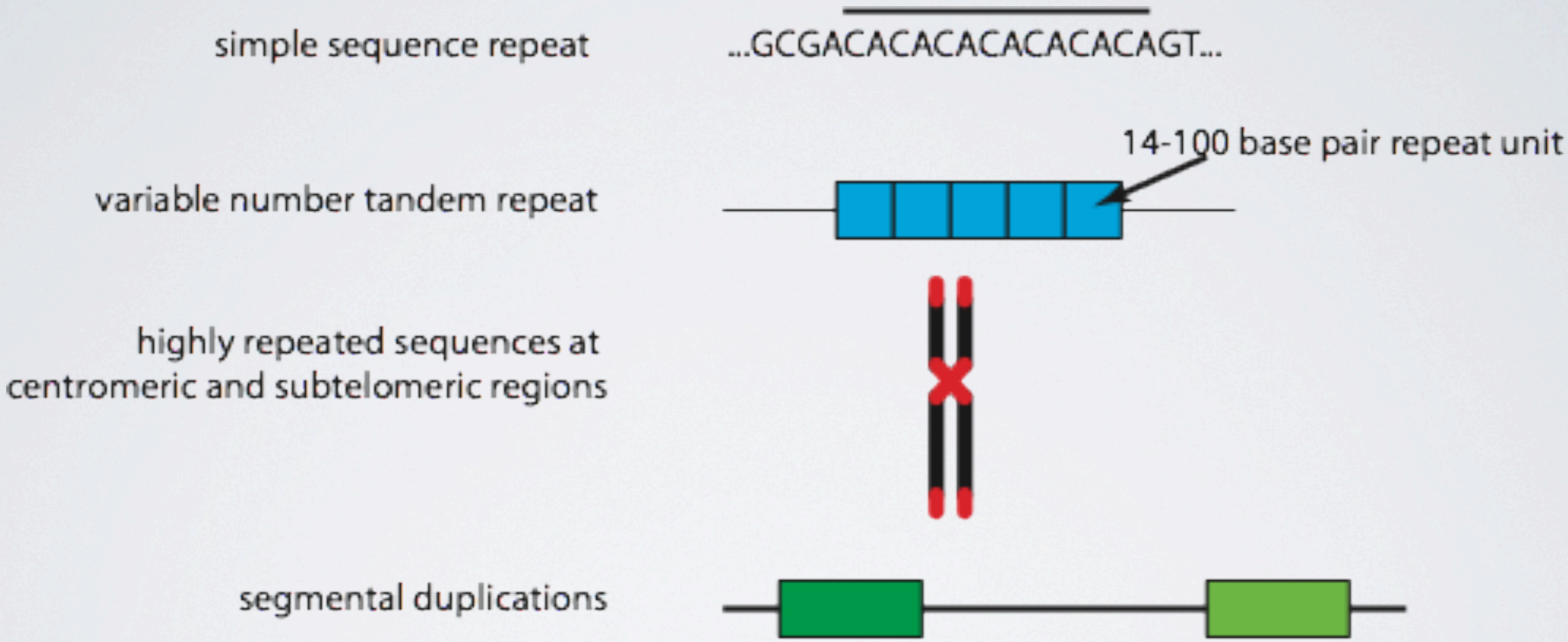

RNA polymerase
5′
3′
3′
5′
3′
direction of transcription
mRNA
5′
poly A addition downstream of AAUAAA
5′ - 5′ triphosphate bond
7-methylguanine
5′
mRNA
3′
mRNA cap
AAUAAA
poly A tail

Splicing



branch site
splice donor
splice acceptor
5′
GU
A
AG
3′
exon
intron
exon
5′
U1
U2
AG
3′
exon
intron
exon
intron
U4
U6
5′
U
U5
U2
AG
3′
exon
exon
splicosome
RNA lariat
U4
U6
5′
U
U5
U2
3′
exon
exon
U4
U6
U
U5
U2
5′
3′
exon
exon

Repeated Sequences



simple sequence repeat

...GCGACACACACACACACAGT...

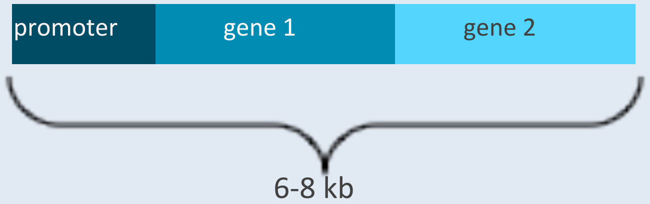
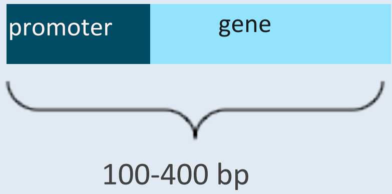
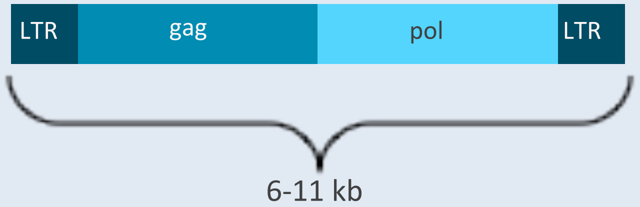
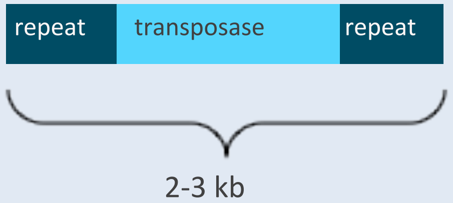
variable number tandem repeat

14-100 base pair repeat unit

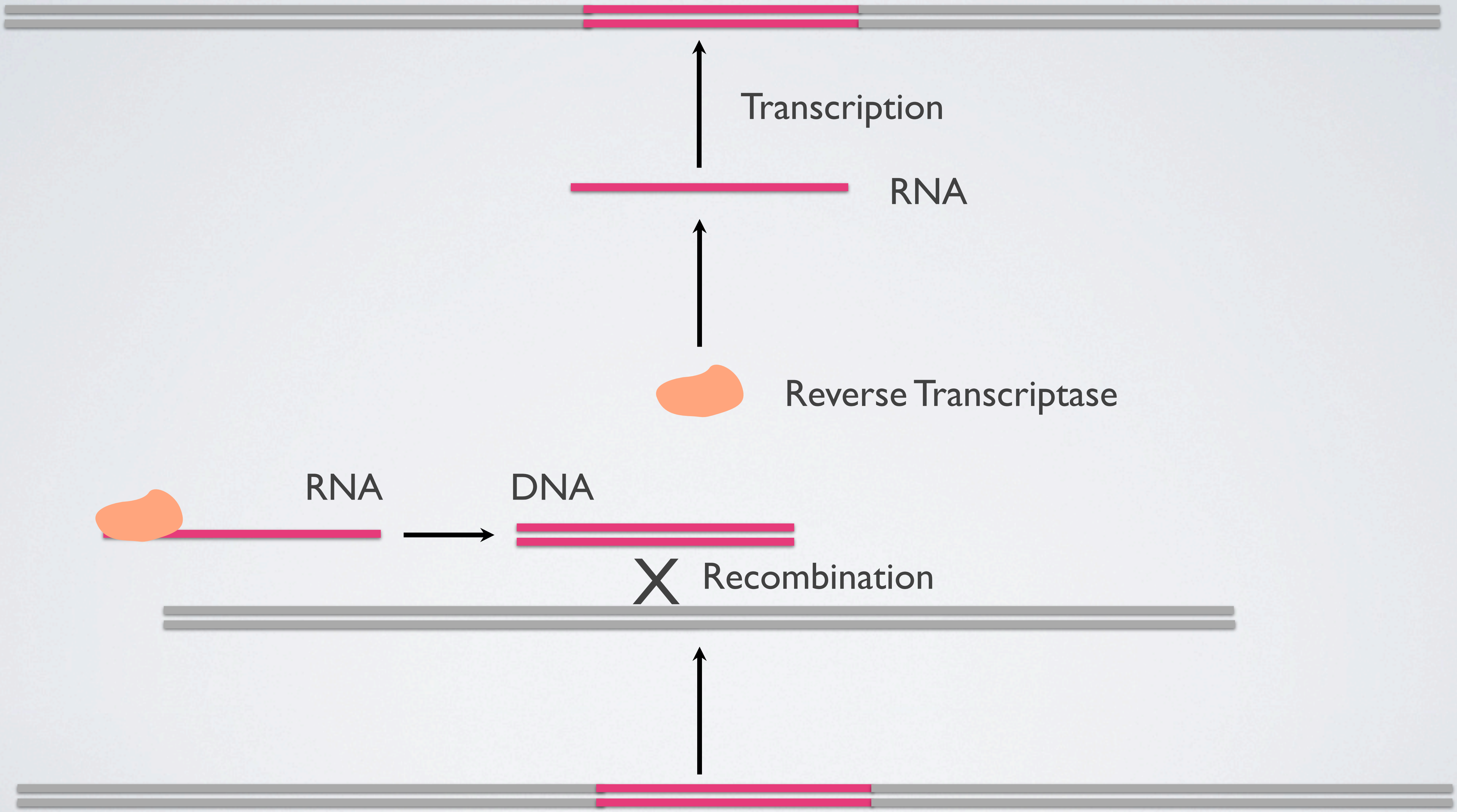
highly repeated sequences at centromeric and subtelomeric regions

segmental duplications

Transposable Genetic Elements

Type	Structure	Copy Number	Percent
LINE	 promoter gene 1 gene 2 6-8 kb	850,000	21
SINE	 promoter gene 100-400 bp	1,500,000	13
Retroviral-like	 LTR gag pol LTR 6-11 kb	450,000	8
Transposon	 repeat transposase repeat 2-3 kb	300,000	3

LINE “Life Cycle”



ENCODE project

- annotated 20,687 protein-encoding genes
- average 6.3 alternatively spliced isoforms per gene
- 8,801 small RNAs; 9,640 long non-coding transcripts
- >80% genome transcribed in some cell type
- >400,000 enhancers and 70,000 promoters

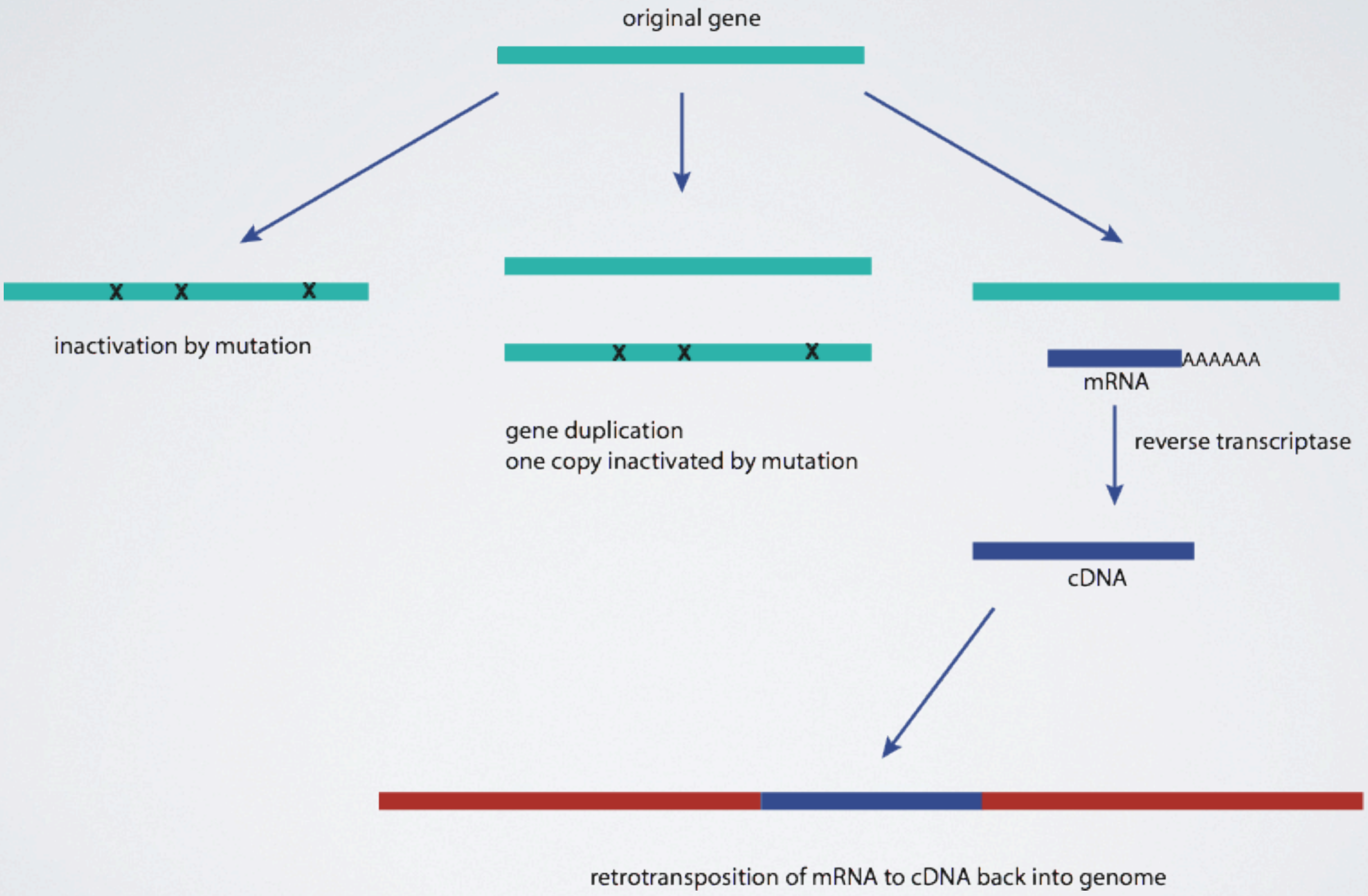
Non-Coding RNAs

tRNA	transfer RNA	protein synthesis
rRNA	ribosomal RNA	protein synthesis
snRNA	small nuclear RNA	splicing
snoRNA	small nucleolar RNA	RNA modification
miRNA	micro RNA	gene regulation
siRNA	small interfering RNA	viral defense
lncRNA	long non-coding RNA	gene regulation/unknown

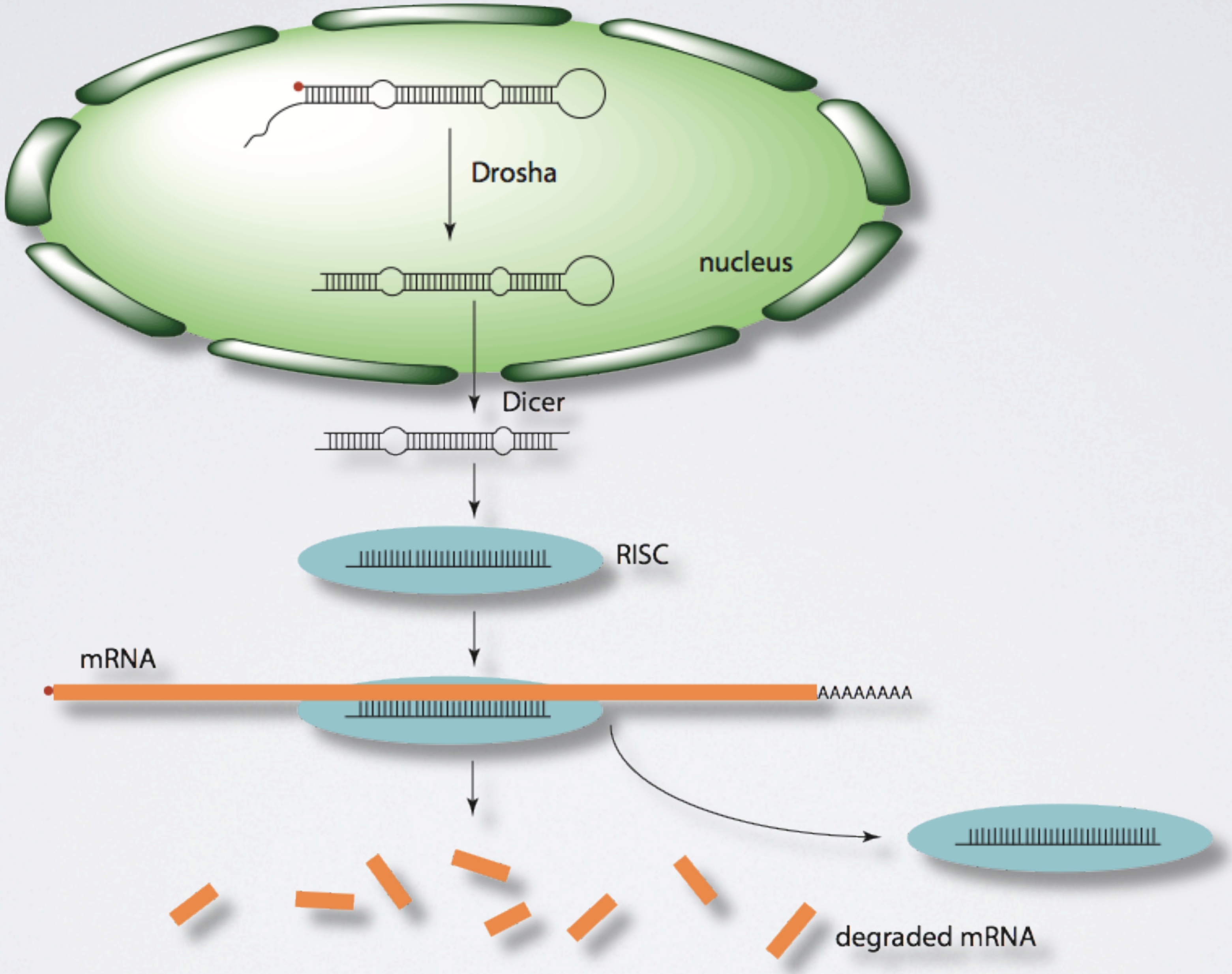
Long Non-Coding RNAs

- antisense
- intergenic
- sense overlapping
- sense intronic
- processed transcript

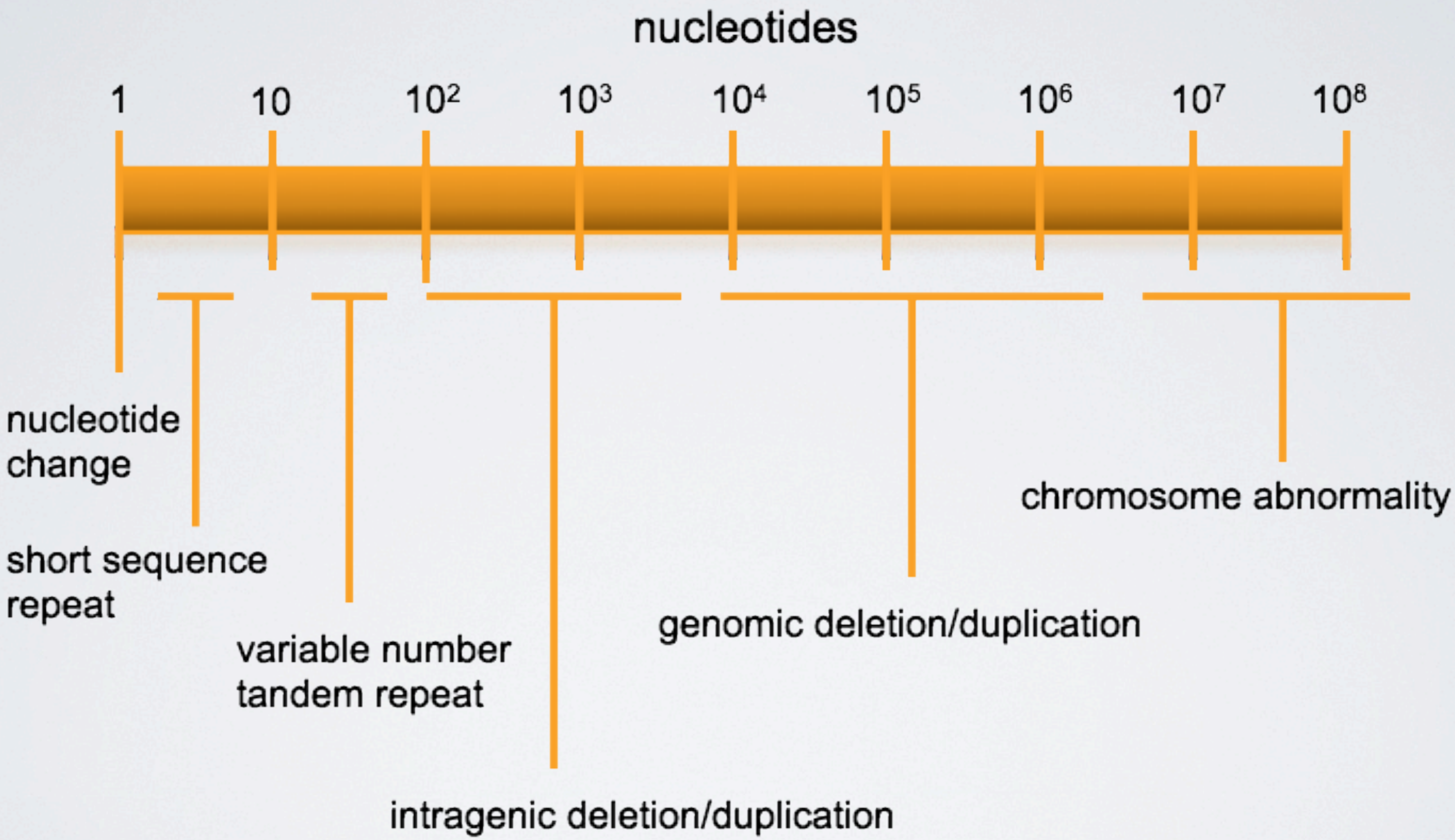
Pseudogenes



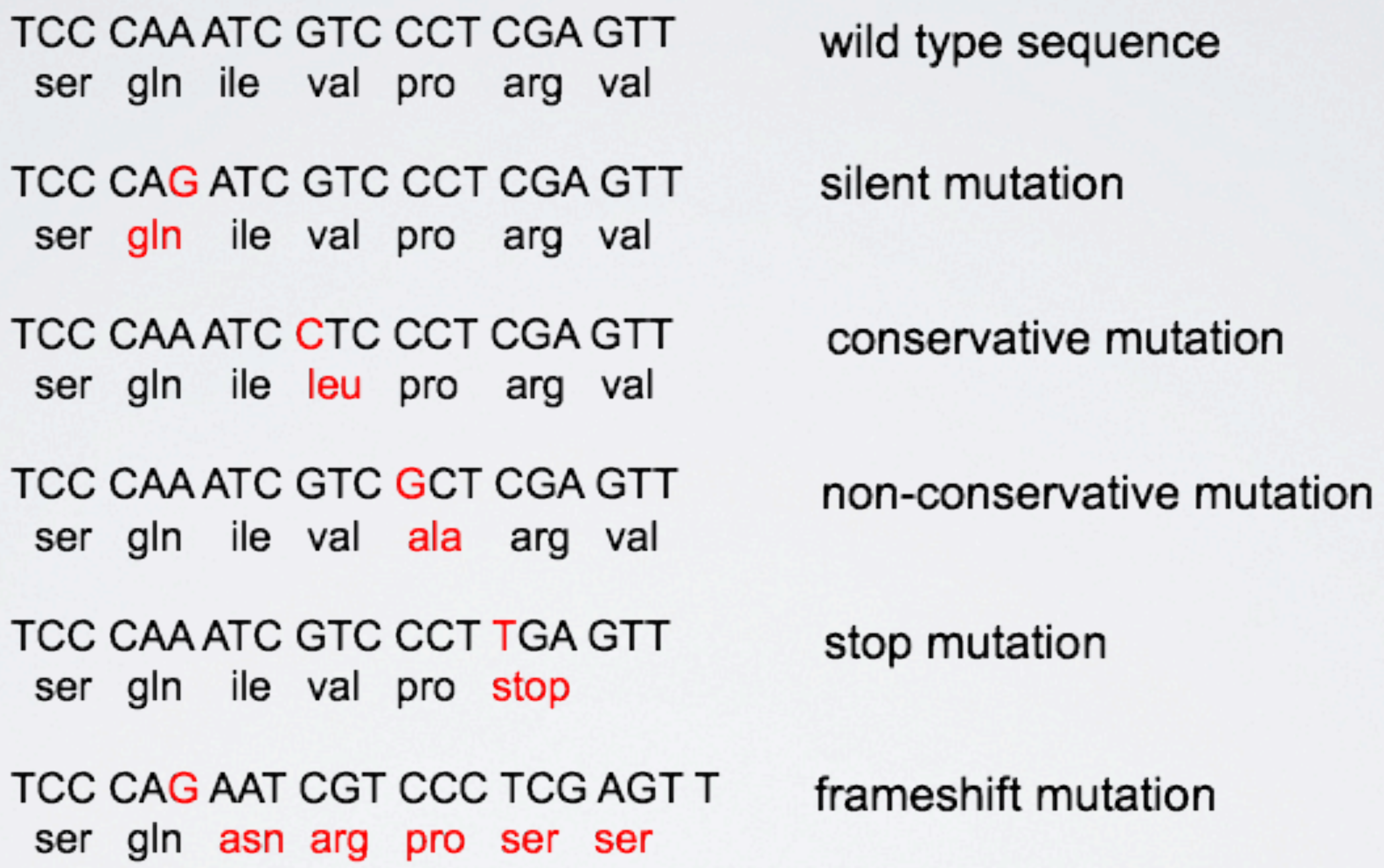
MicroRNA



Genetic Variation

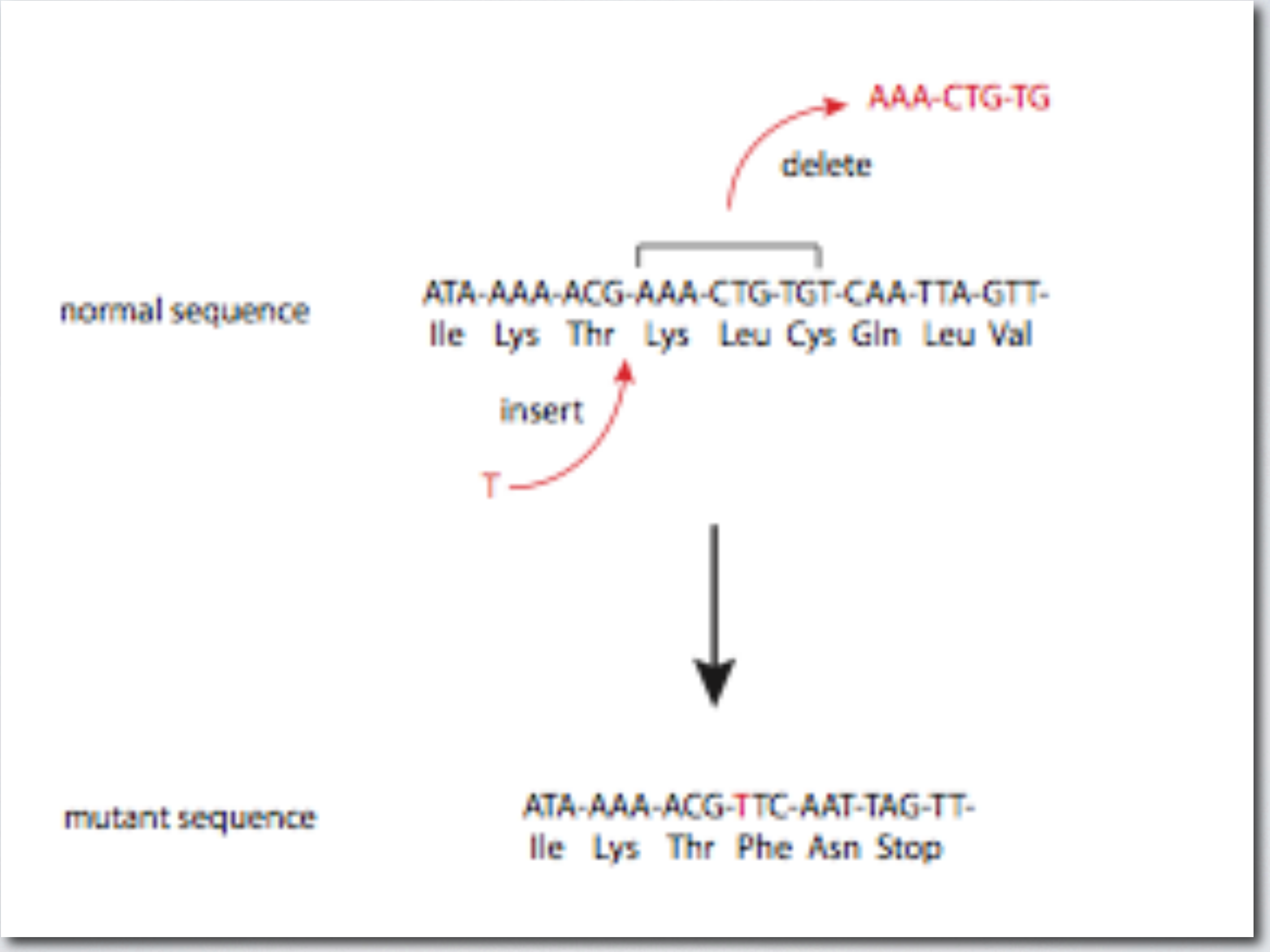


Point Mutations

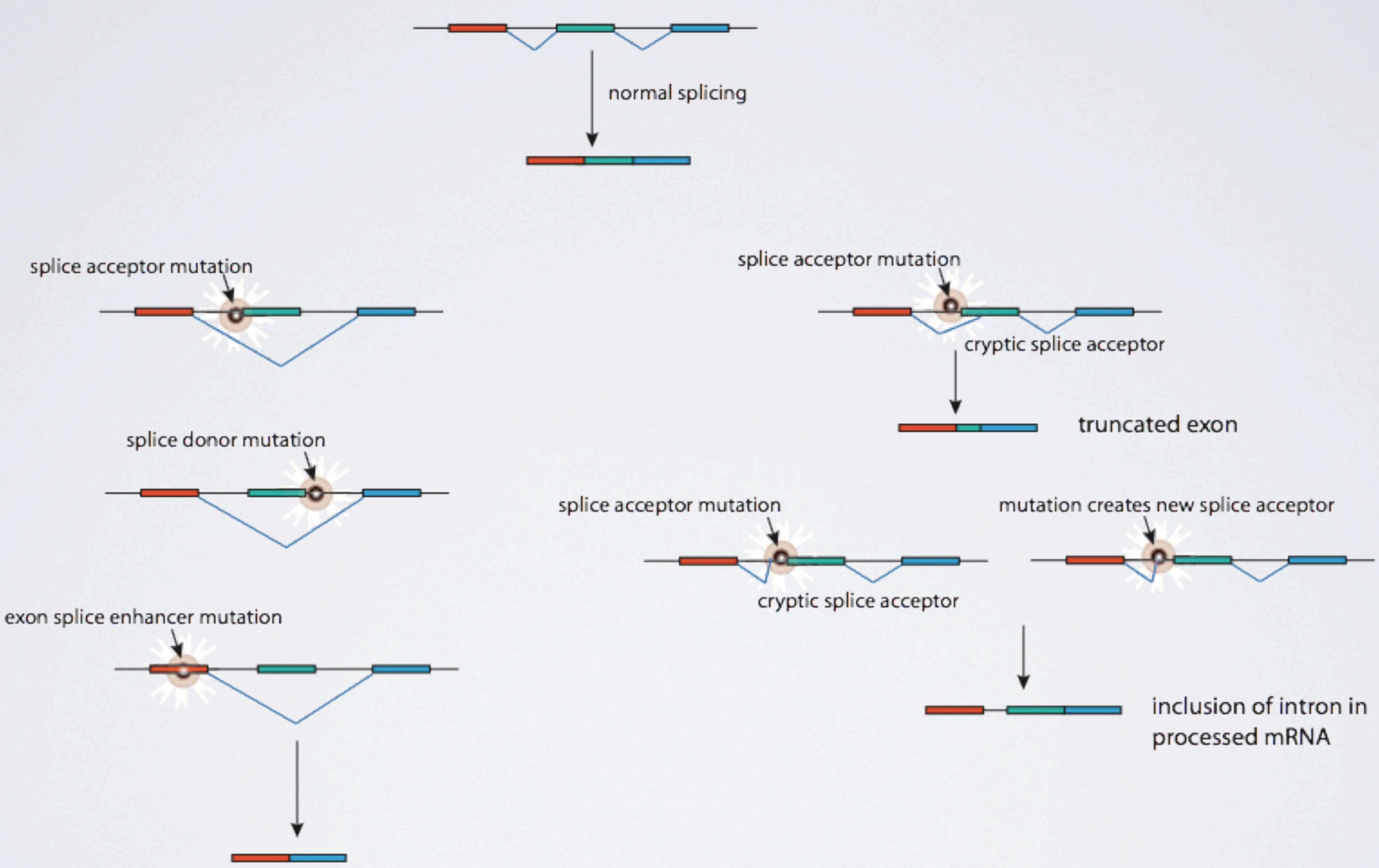


TCC CAA ATC GTC CCT CGA GTT ser gln ile val pro arg val	wild type sequence
TCC CAG ATC GTC CCT CGA GTT ser gln ile val pro arg val	silent mutation
TCC CAA ATC CTC CCT CGA GTT ser gln ile leu pro arg val	conservative mutation
TCC CAA ATC GTC GCT CGA GTT ser gln ile val ala arg val	non-conservative mutation
TCC CAA ATC GTC CCT TGA GTT ser gln ile val pro stop	stop mutation
TCC CAG AAT CGT CCC TCG AGT T ser gln asn arg pro ser ser	frameshift mutation

Indel


AAA-CTG-TG
delete
normal sequence
ATA-AAA-ACG-AAA-CTG-TGT-CAA-TTA-GTT-
Ile Lys Thr Lys Leu Cys Gln Leu Val
insert
T
mutant sequence
ATA-AAA-ACG-TTC-AAT-TAG-TT-
Ile Lys Thr Phe Asn Stop

Splicing Mutations



normal splicing

splice acceptor mutation

splice donor mutation

exon splice enhancer mutation

splice acceptor mutation

cryptic splice acceptor

truncated exon

splice acceptor mutation

cryptic splice acceptor

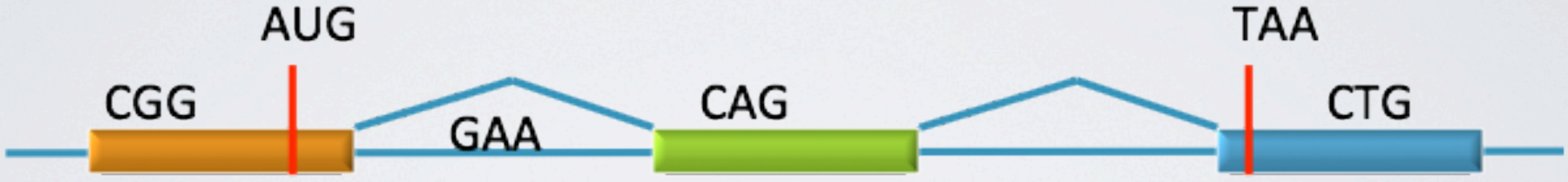
mutation creates new splice acceptor

inclusion of intron in processed mRNA

exon skip mutations

cryptic splice acceptor (or donor) mutations

Triplet Repeat Expansions



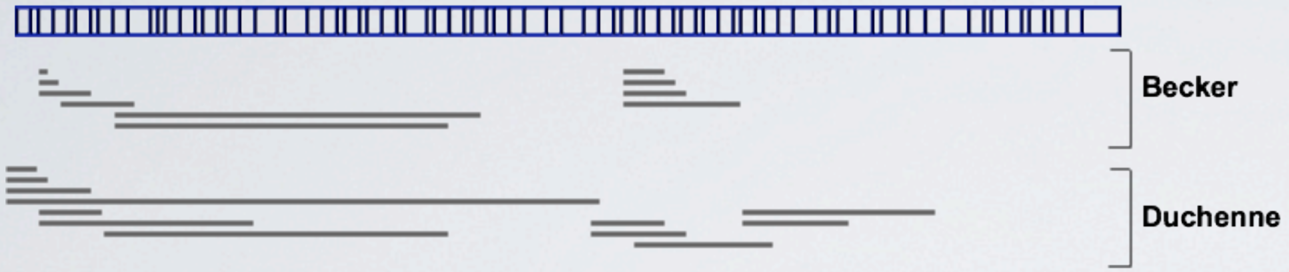
Fragile X syndrome

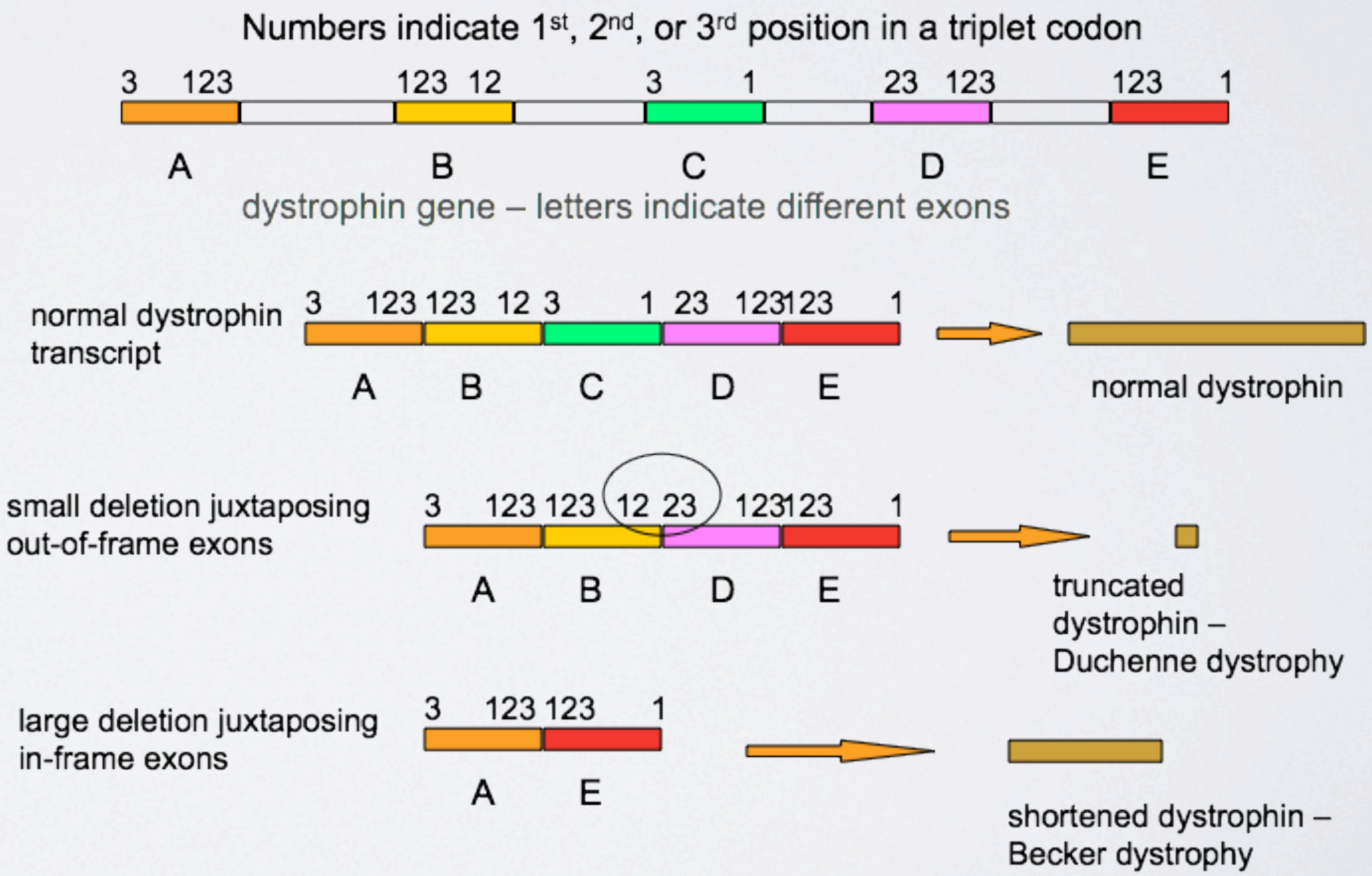
Friedreich ataxia

Spinocerebellar ataxias
Huntington disease
DRP atrophy
Spinal & bulbar atrophy
Machado-Joseph disease

myotonic dystrophy

Multiexon Deletion




Numbers indicate 1st, 2nd, or 3rd position in a triplet codon

dystrophin gene – letters indicate different exons

normal dystrophin transcript

normal dystrophin

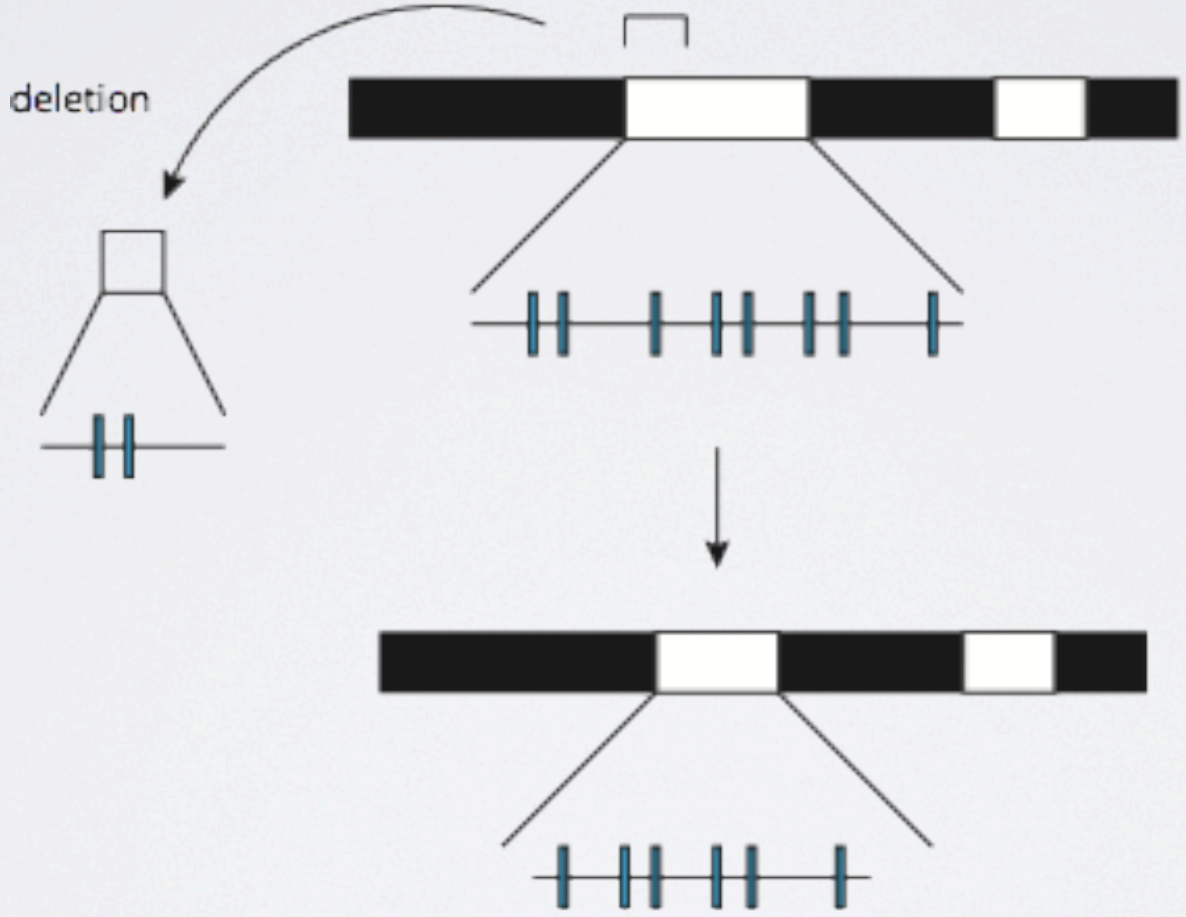
small deletion juxtaposing out-of-frame exons

truncated dystrophin – Duchenne dystrophy

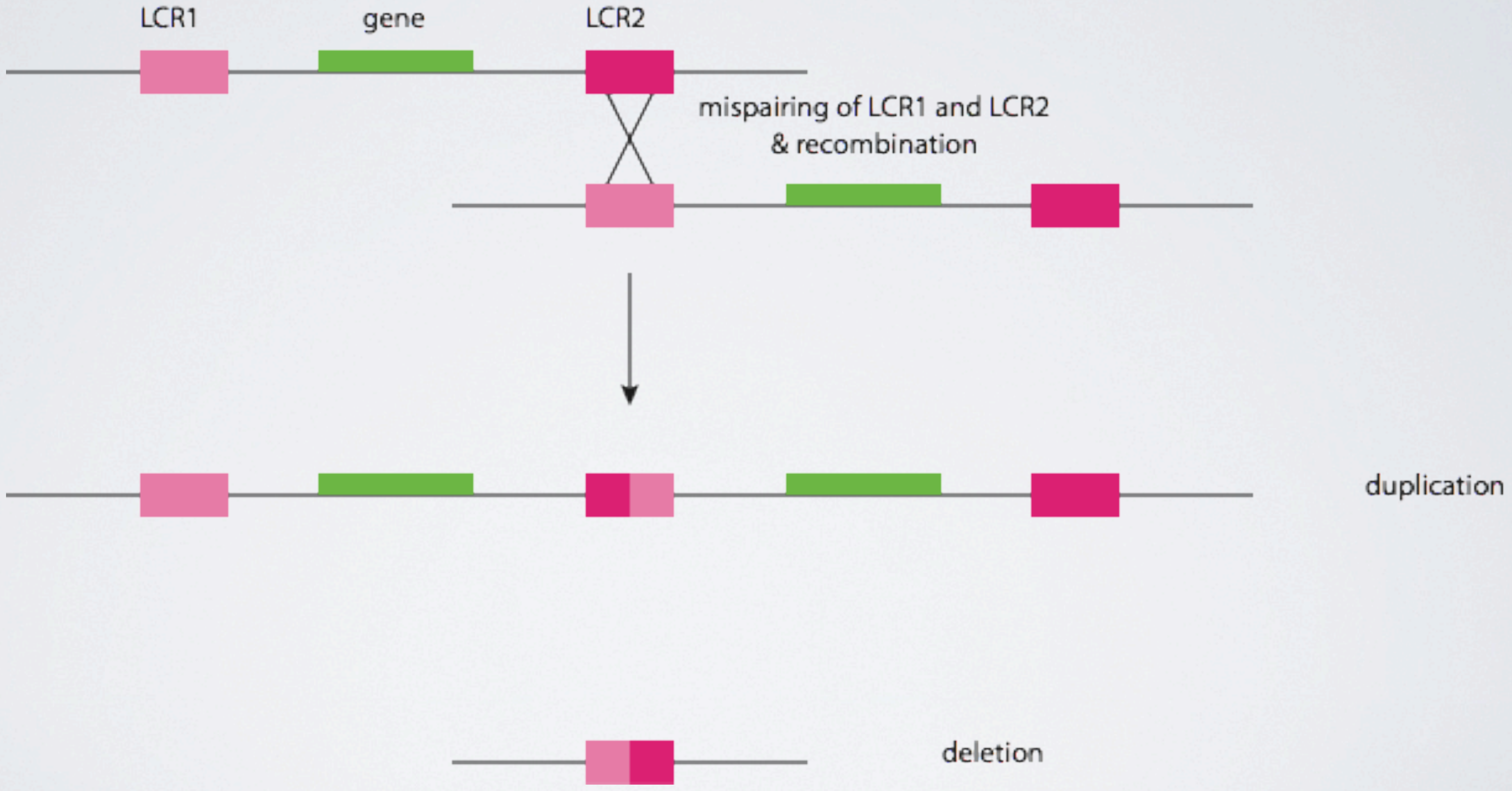
large deletion juxtaposing in-frame exons

shortened dystrophin – Becker dystrophy

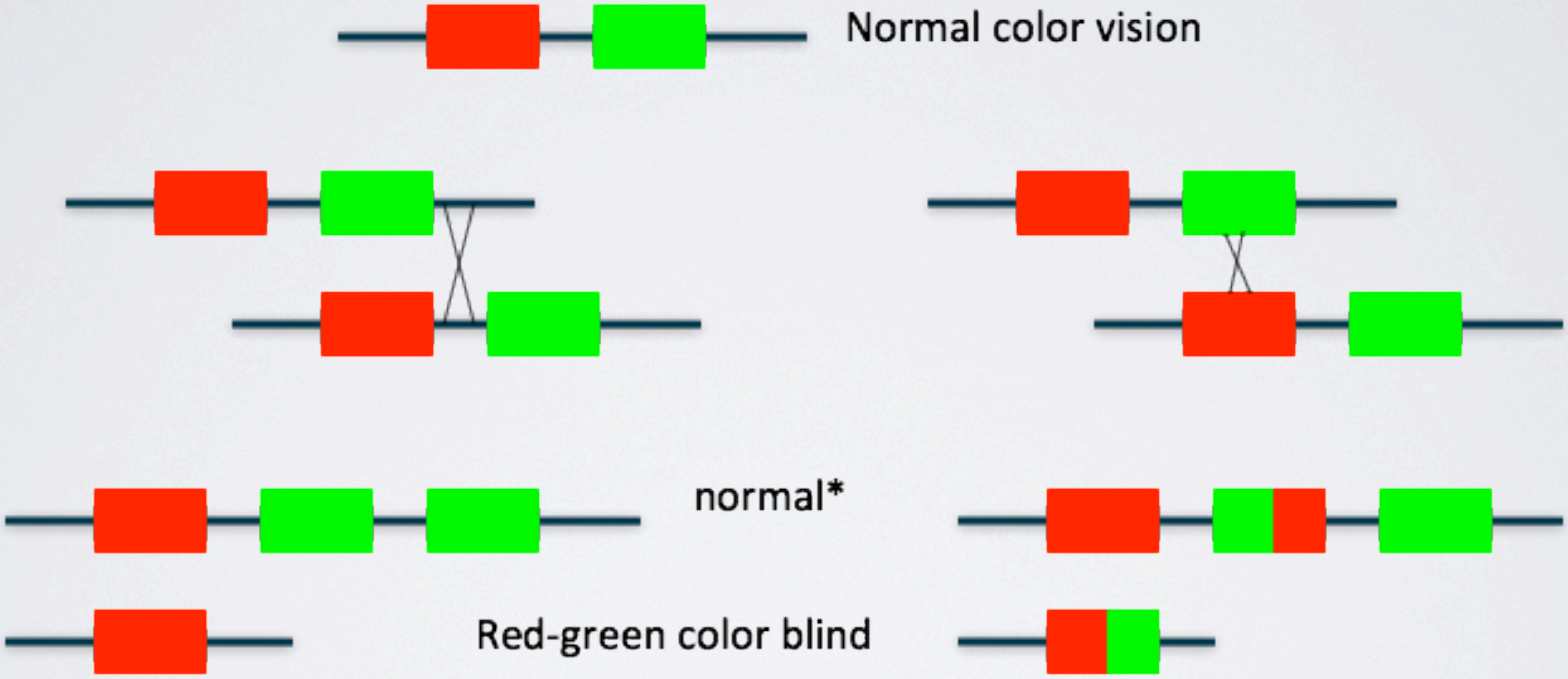
Chromosome Microdeletion



LCR Mispairing



Red-Green Color Blindness



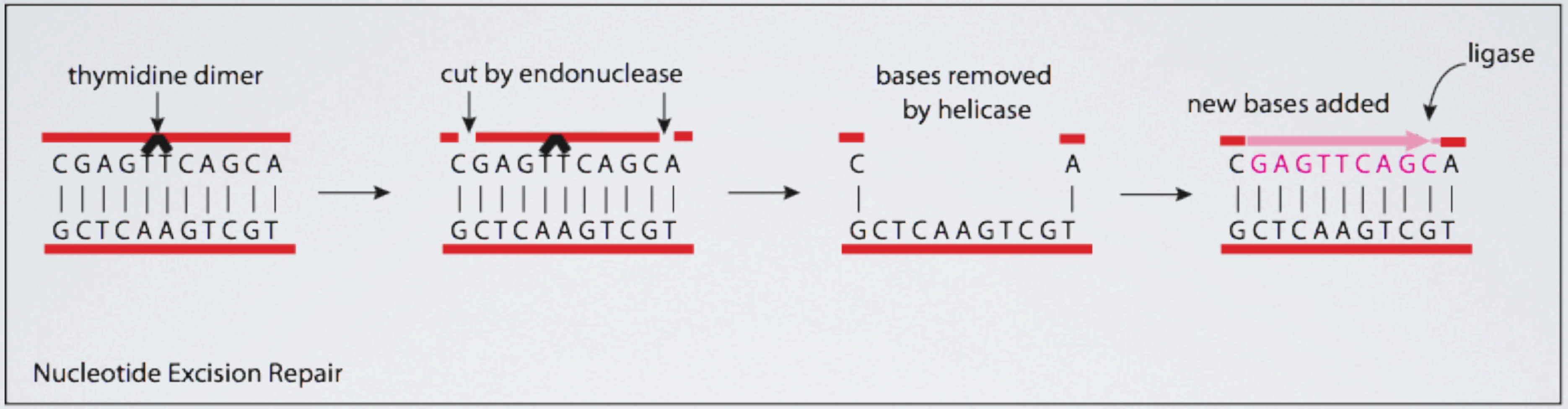
Normal color vision

normal*

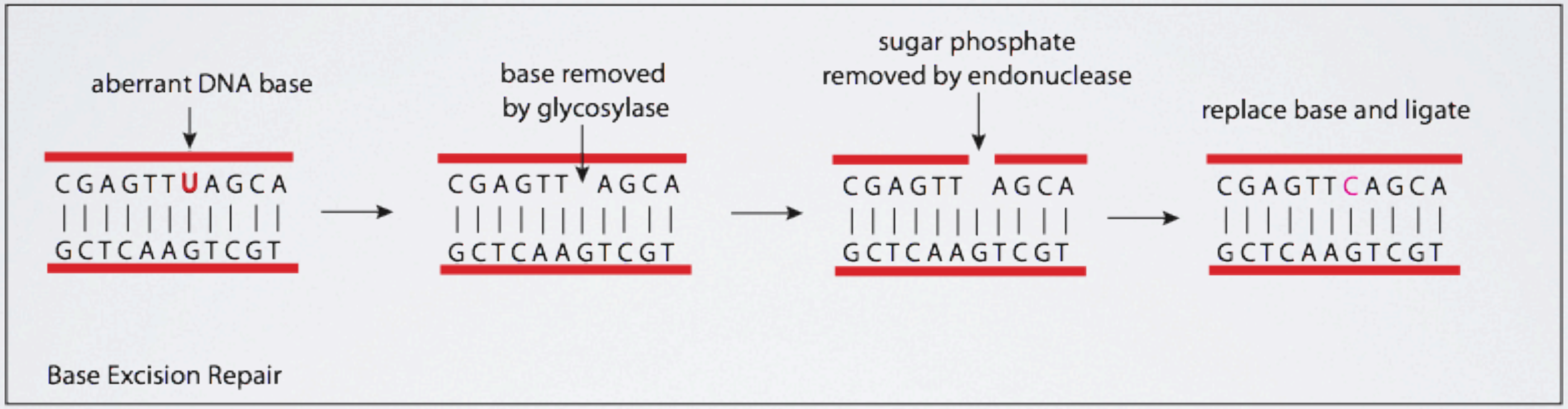
Red-green color blind

* Color vision may be abnormal if green gene not expressed

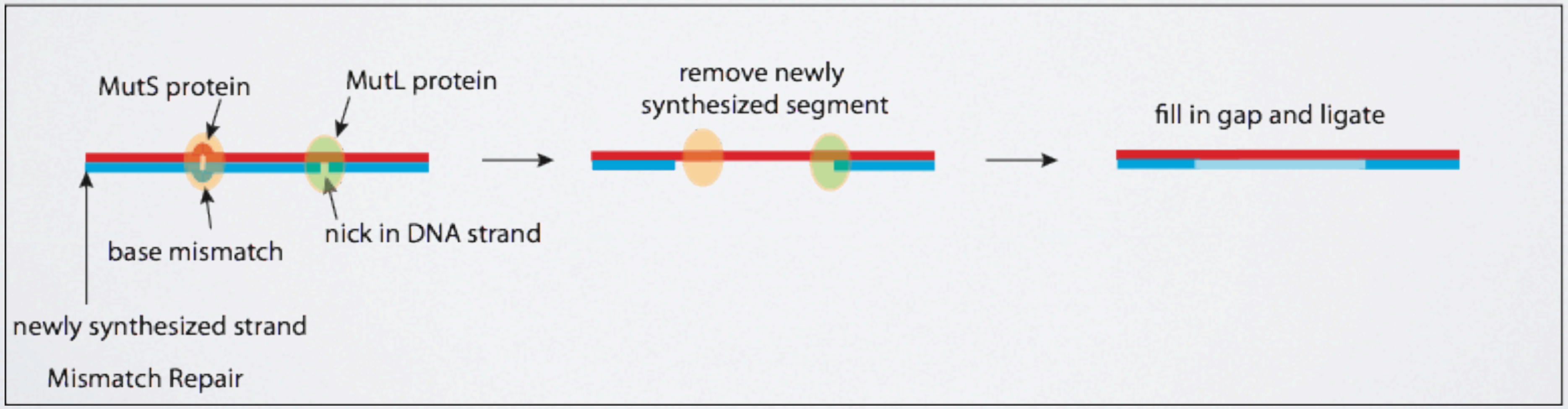
DNA Repair



Nucleotide Excision Repair

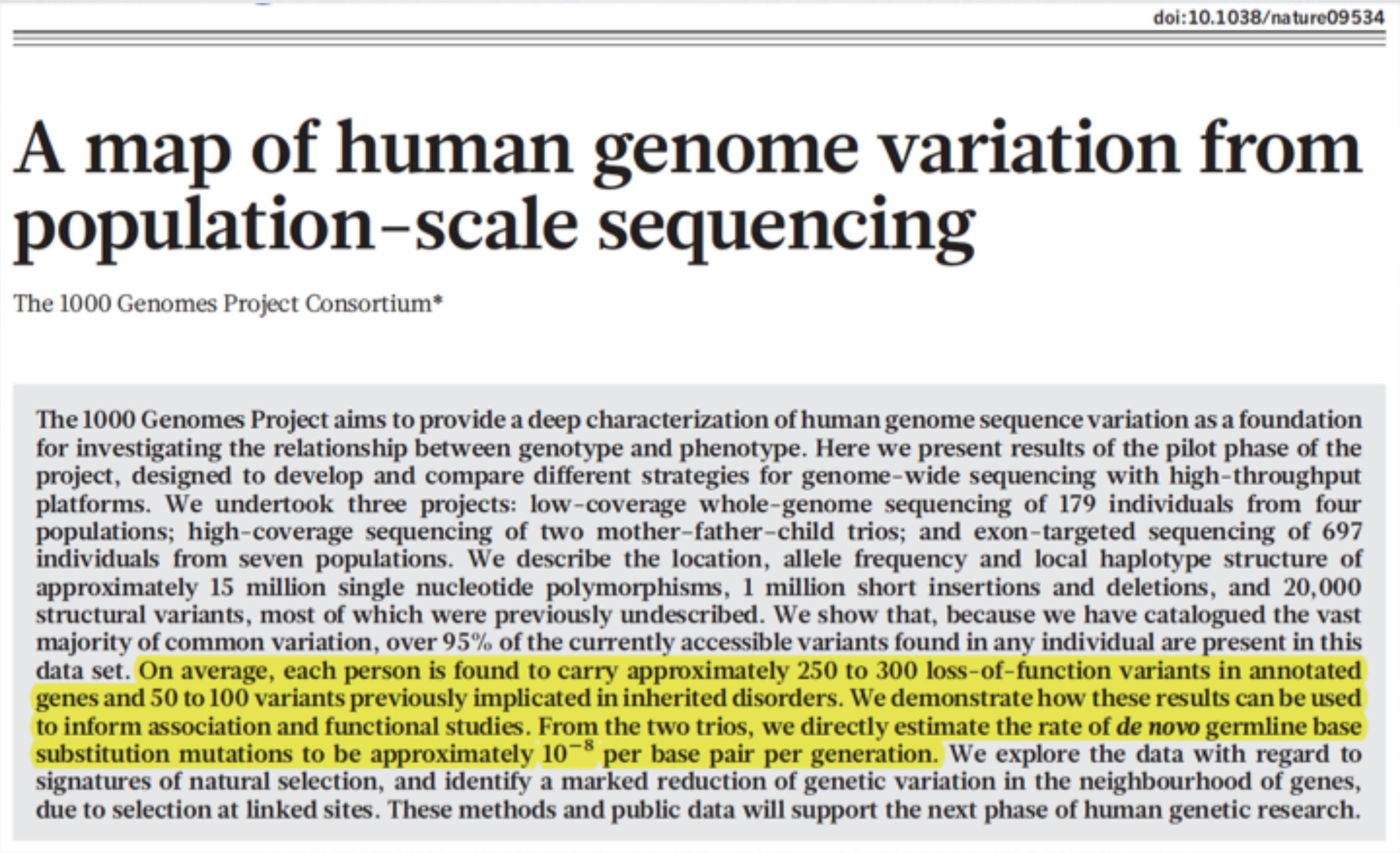


Base Excision Repair



Mismatch Repair

Frequency of Mutation



doi:10.1038/nature09534

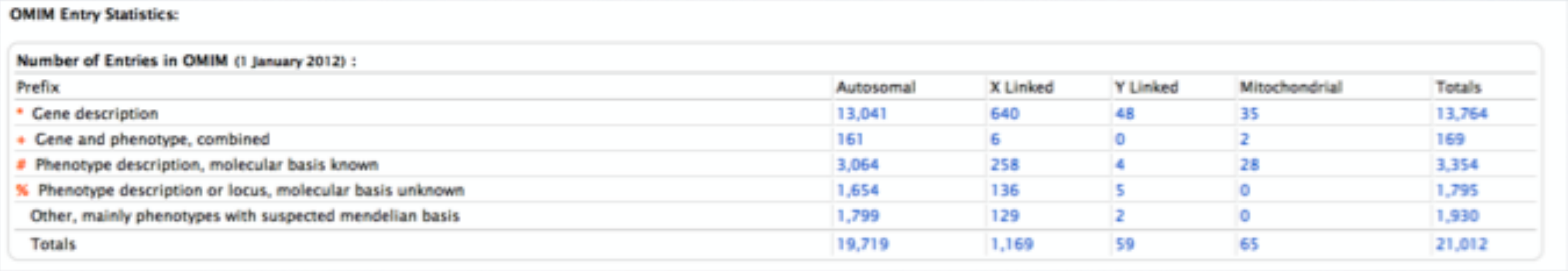
A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

If there are 10^8 sperm per ejaculate, in principle every base could be mutated in at least one sperm cell and each germ cell has around 10 mutations

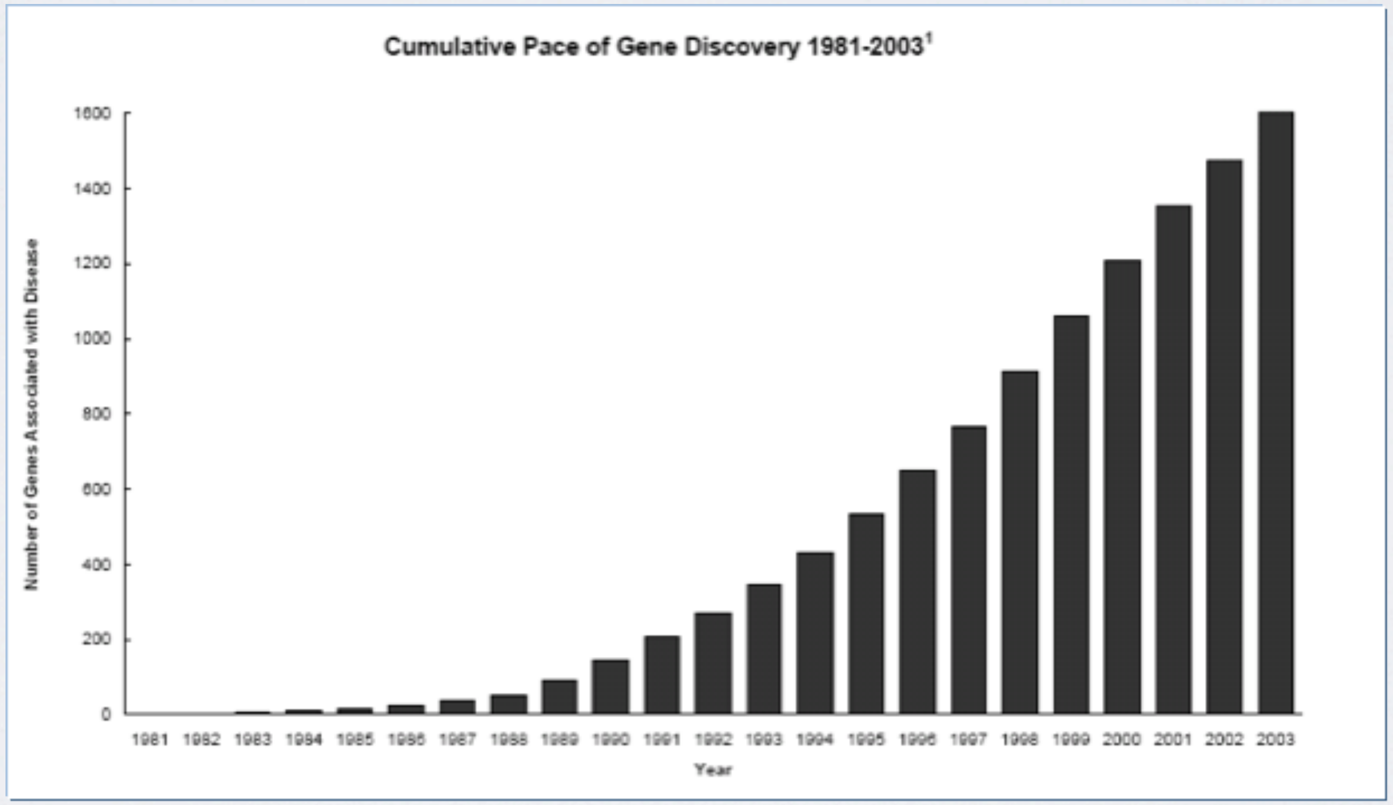
Human Mendelian Phenotypes



OMIM Entry Statistics:

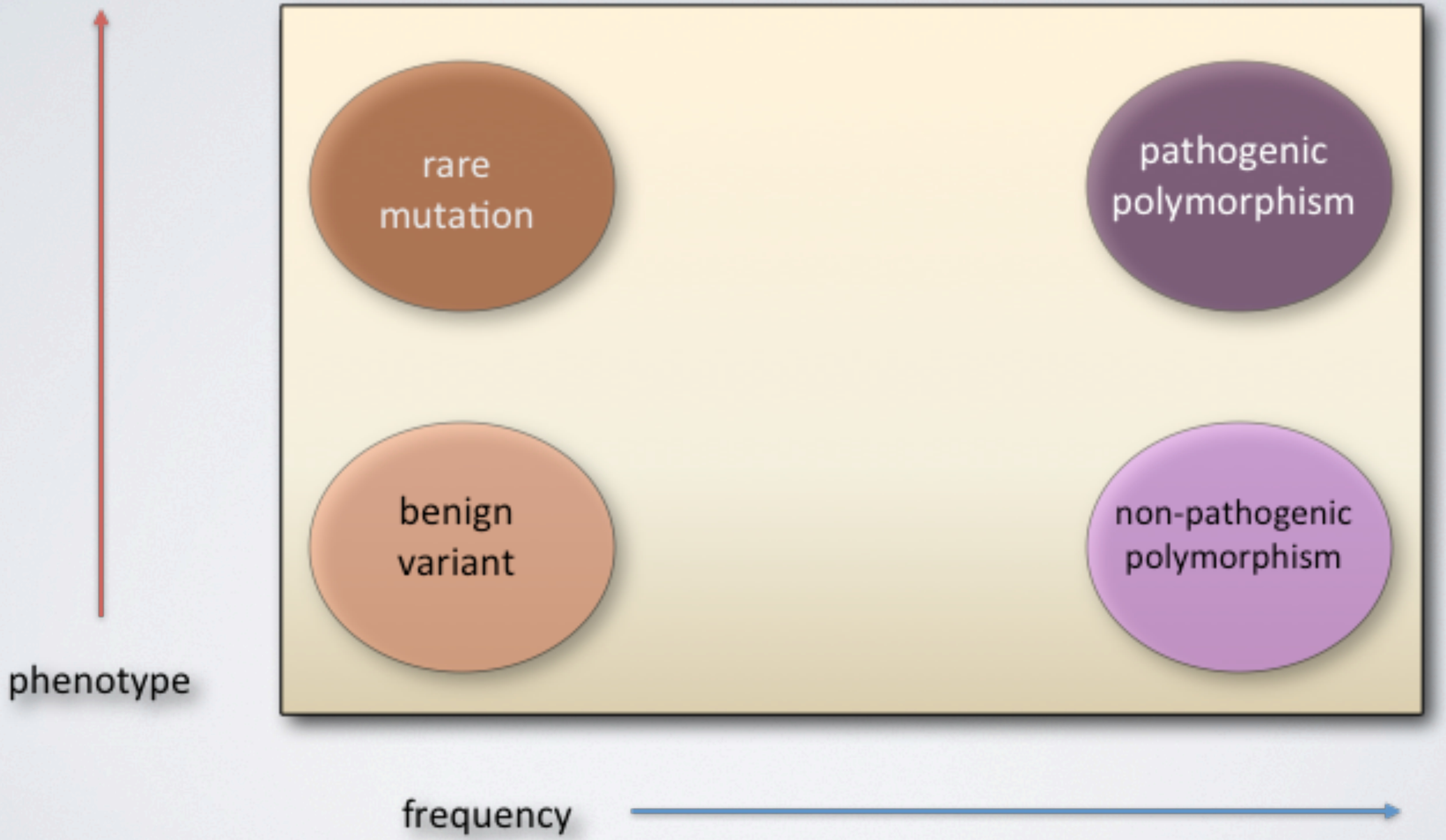
Number of Entries in OMIM (1 January 2012) :

Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
* Gene description	13,041	640	48	35	13,764
+ Gene and phenotype, combined	161	6	0	2	169
# Phenotype description, molecular basis known	3,064	258	4	28	3,354
% Phenotype description or locus, molecular basis unknown	1,654	136	5	0	1,795
Other, mainly phenotypes with suspected mendelian basis	1,799	129	2	0	1,930
Totals	19,719	1,169	59	65	21,012

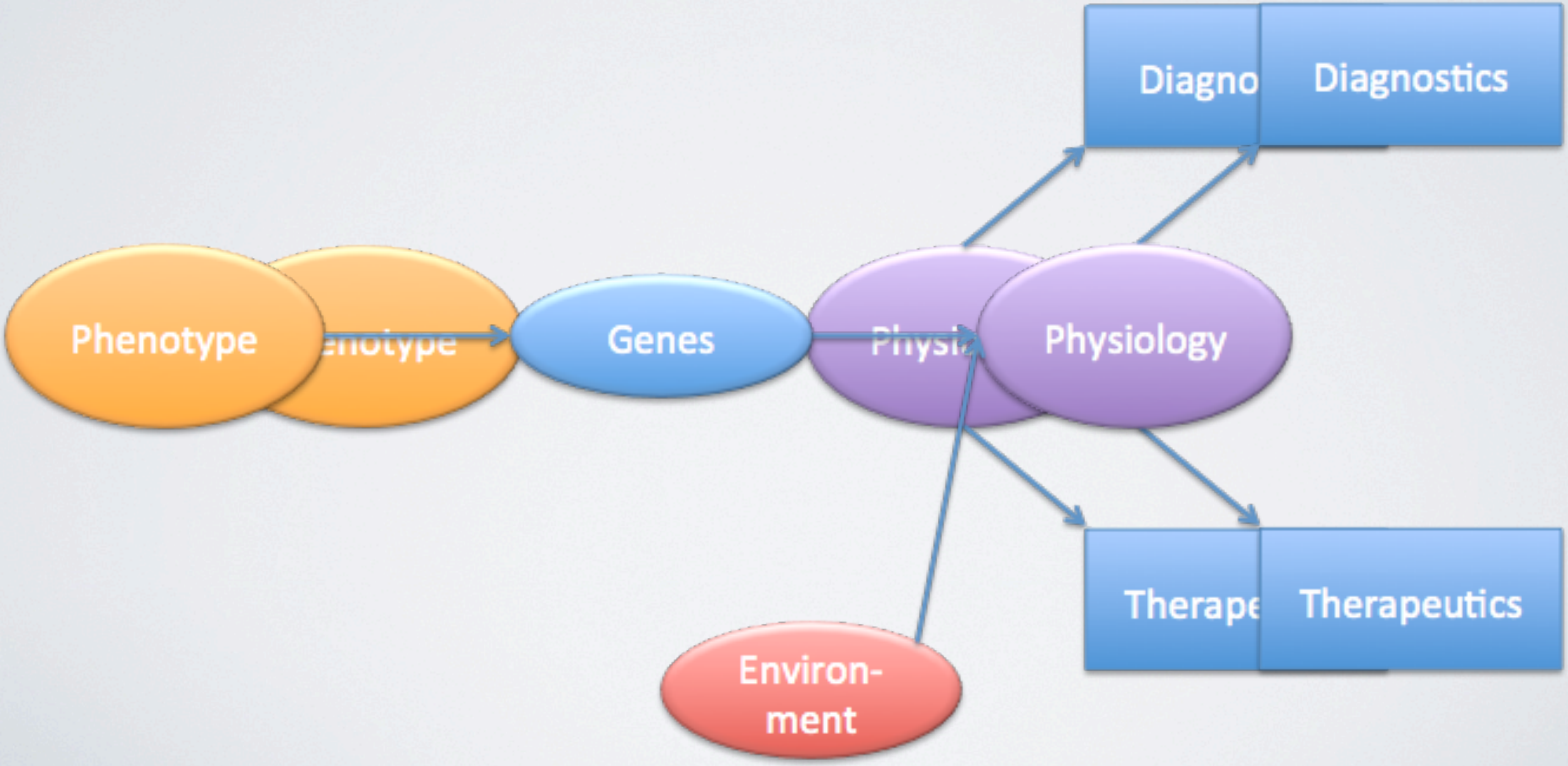


http://www.genome.gov/Pages/News/PaceofDiseaseGeneDiscovery.pdf

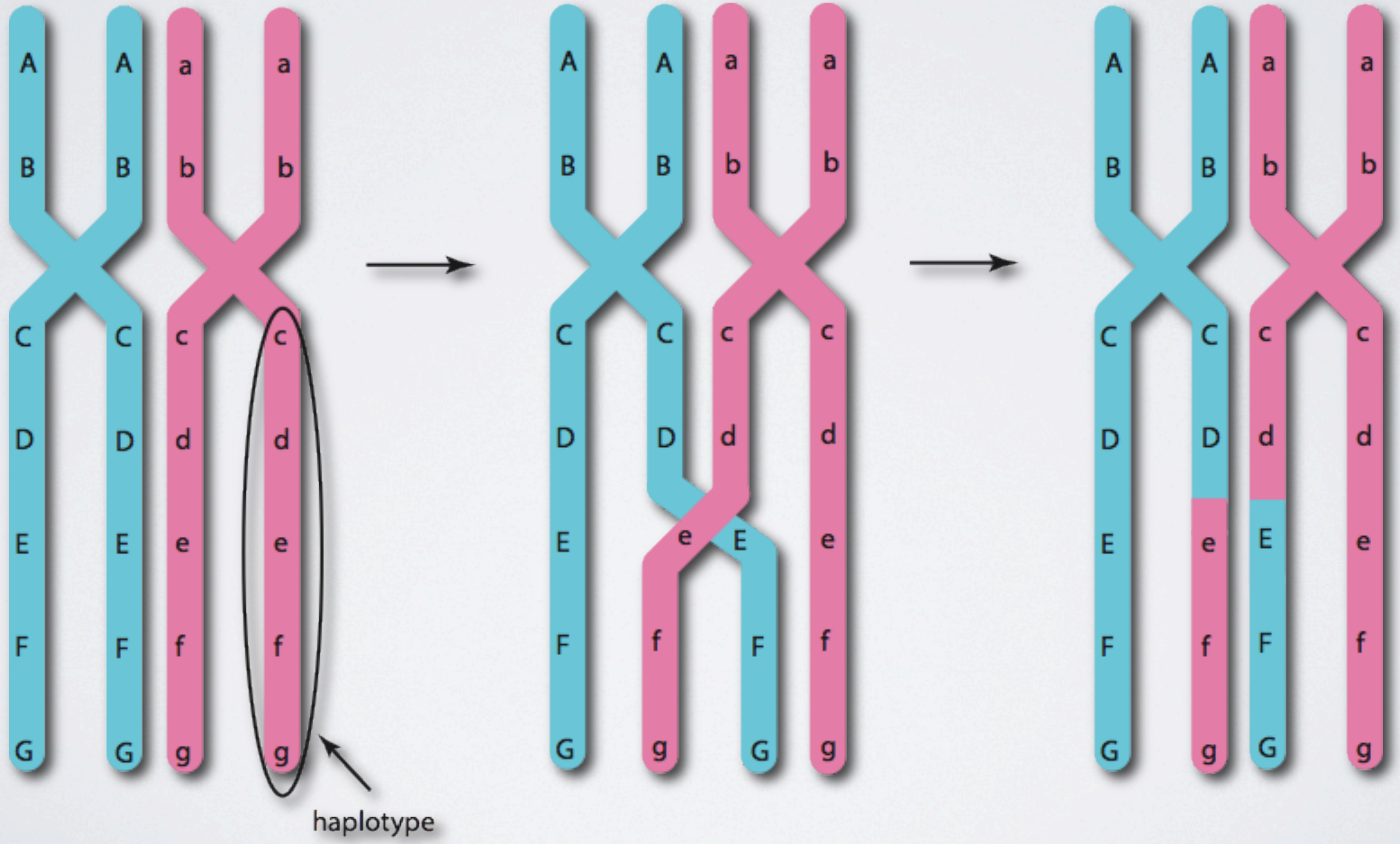
Types of Variants



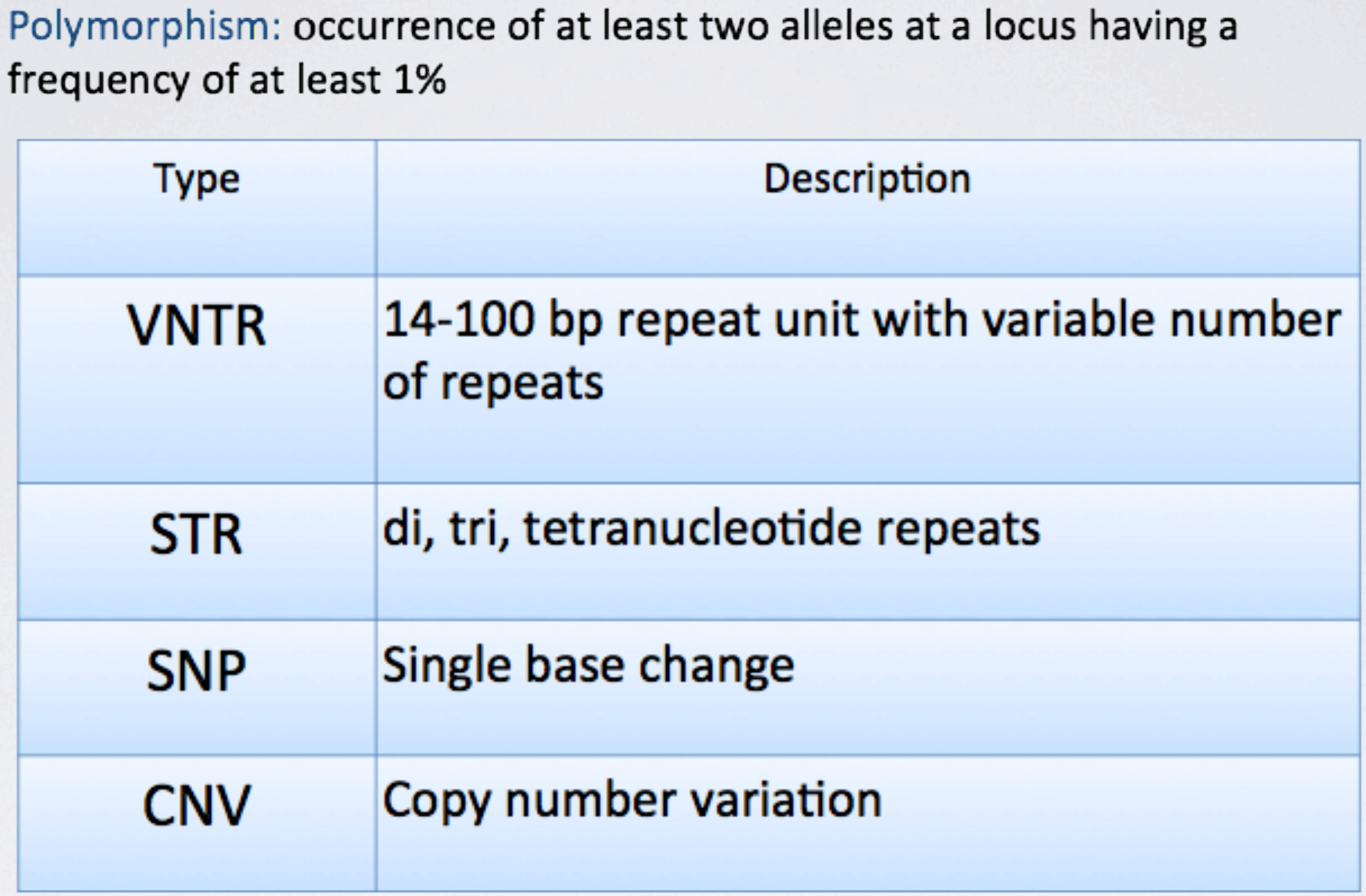
Approach to Genetic Disorders



Genetic Linkage



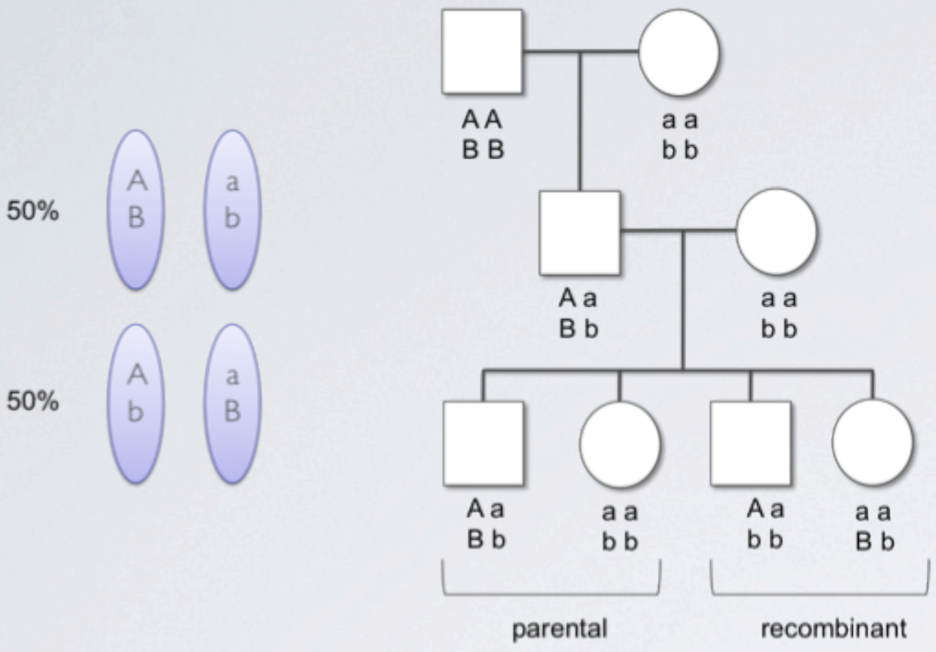
Polymorphism



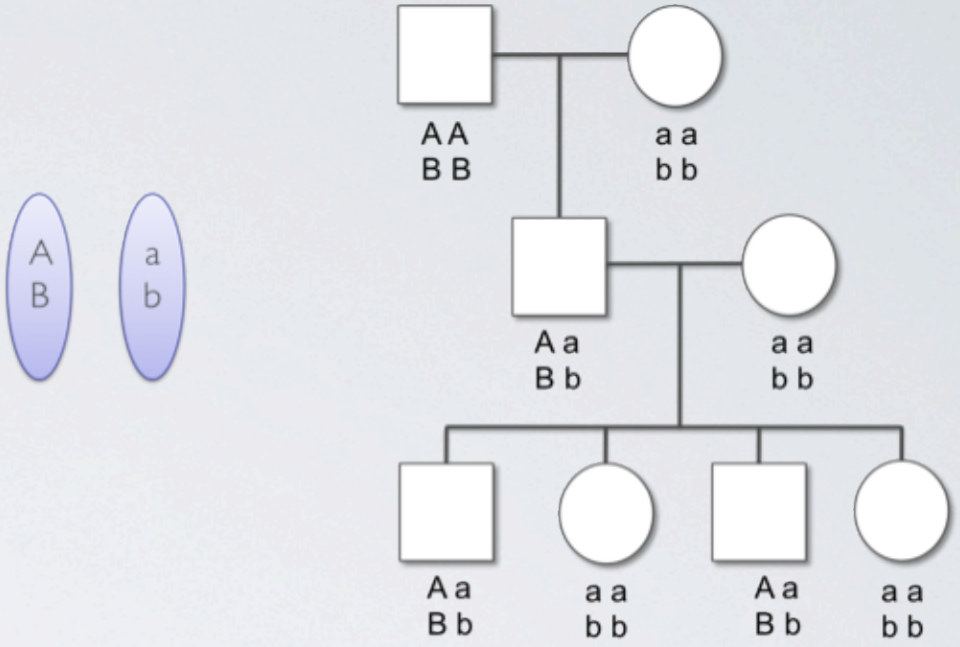
Polymorphism: occurrence of at least two alleles at a locus having a frequency of at least 1%

Type	Description
VNTR	14-100 bp repeat unit with variable number of repeats
STR	di, tri, tetranucleotide repeats
SNP	Single base change
CNV	Copy number variation

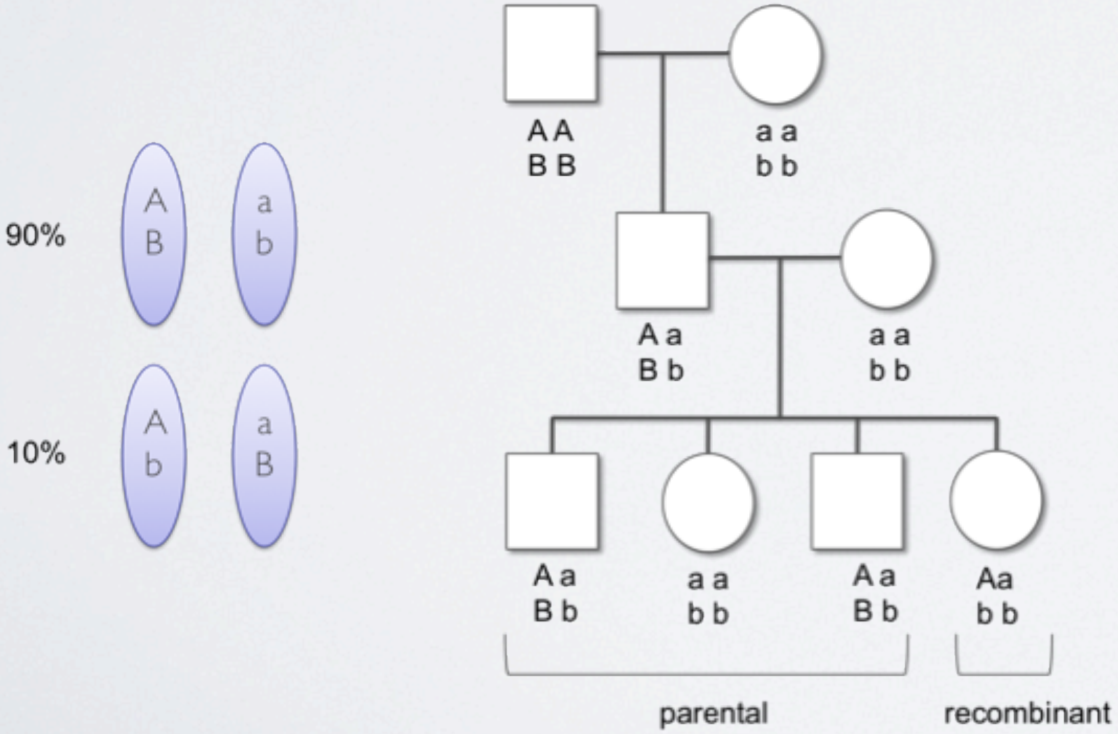
Linkage



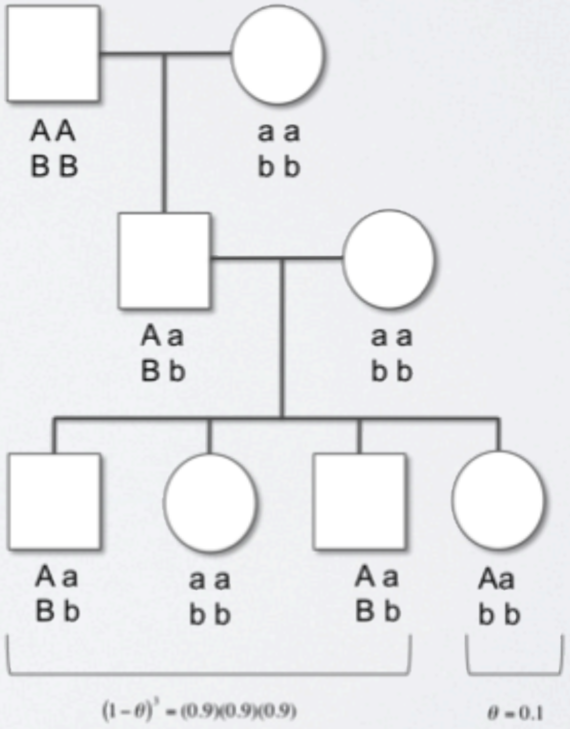
Independent Assortment



Complete Linkage



10% Recombination

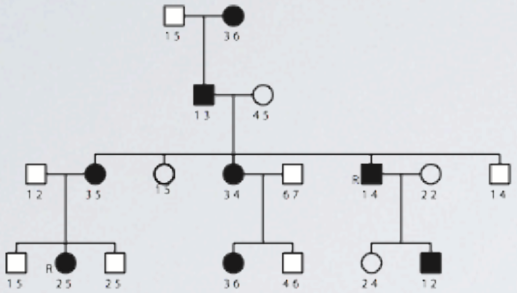


odds ratio = $\dfrac{(1-\theta)^n(\theta)^r}{(1/2)^{n+r}}$

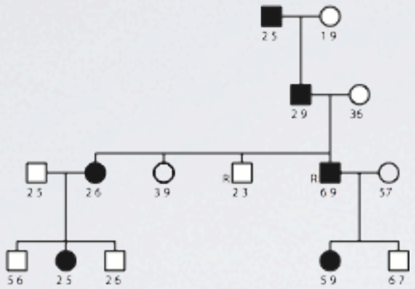
n = number non-recombinants
r = number recombinants

Likelihood Ratio

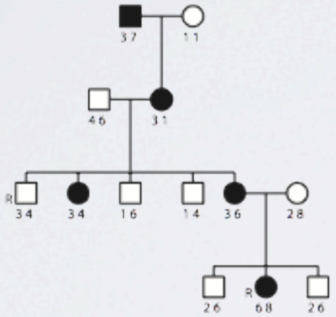
LOD Analysis



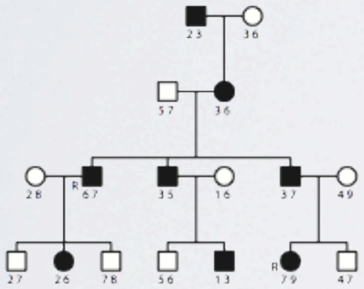
Family 1



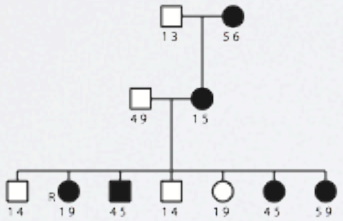
Family 2



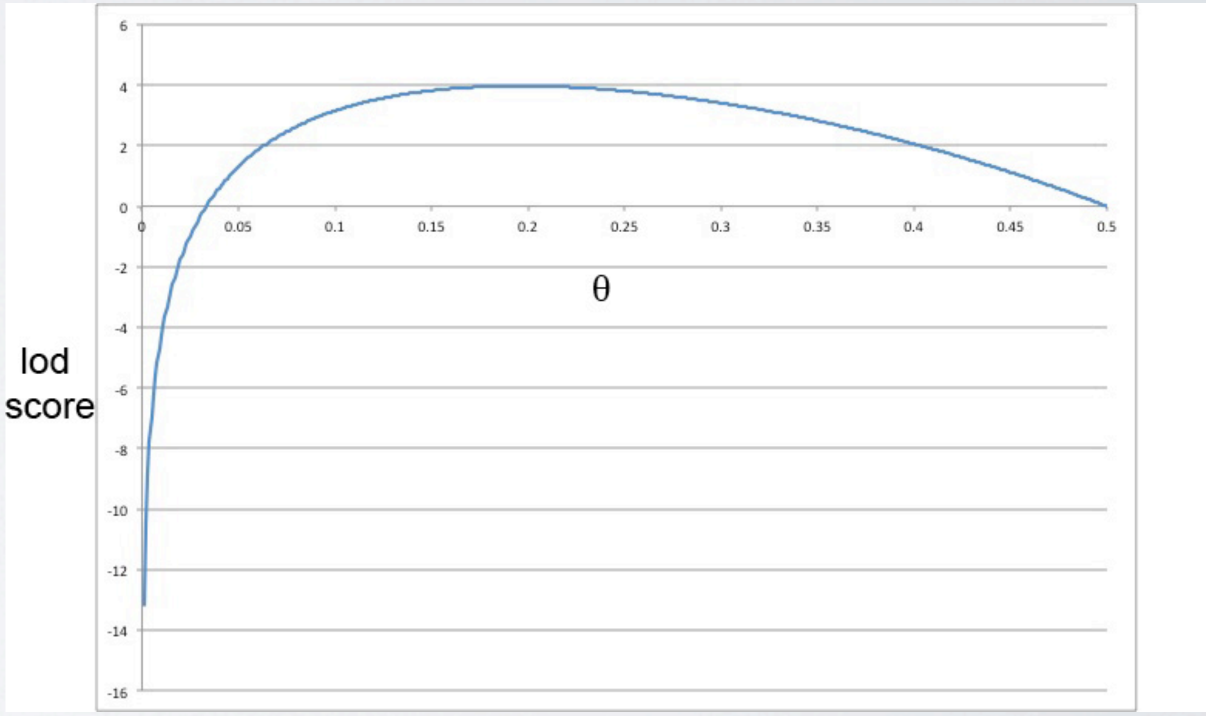
Family 3

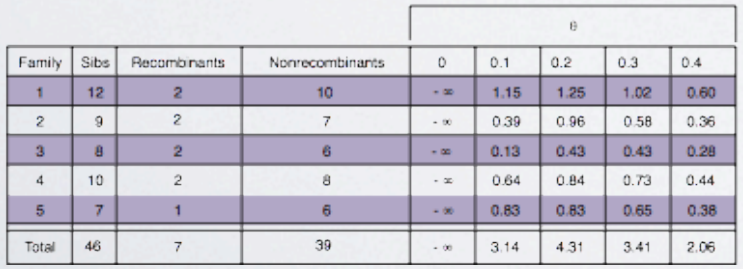


Family 4



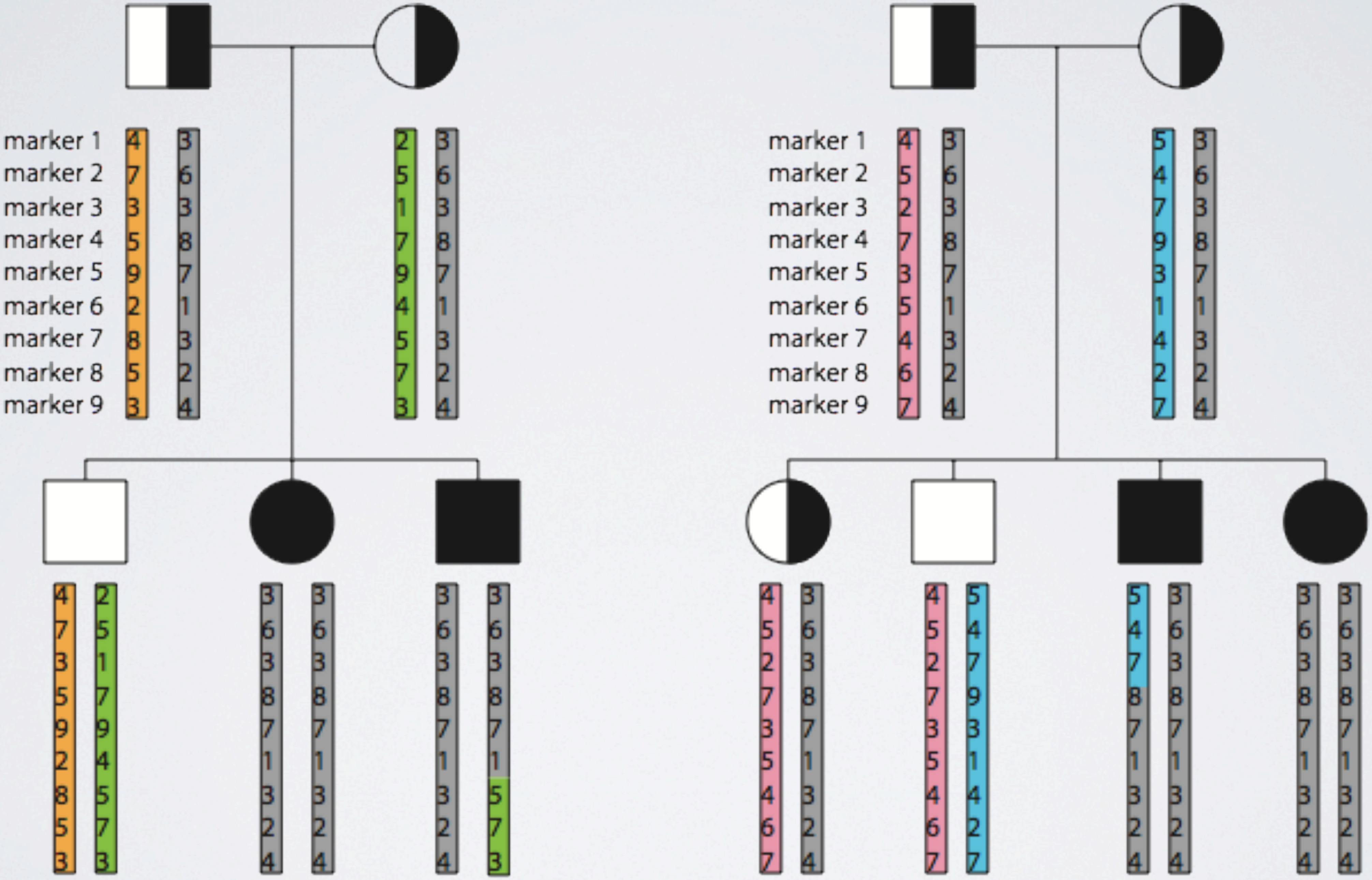
Family 5



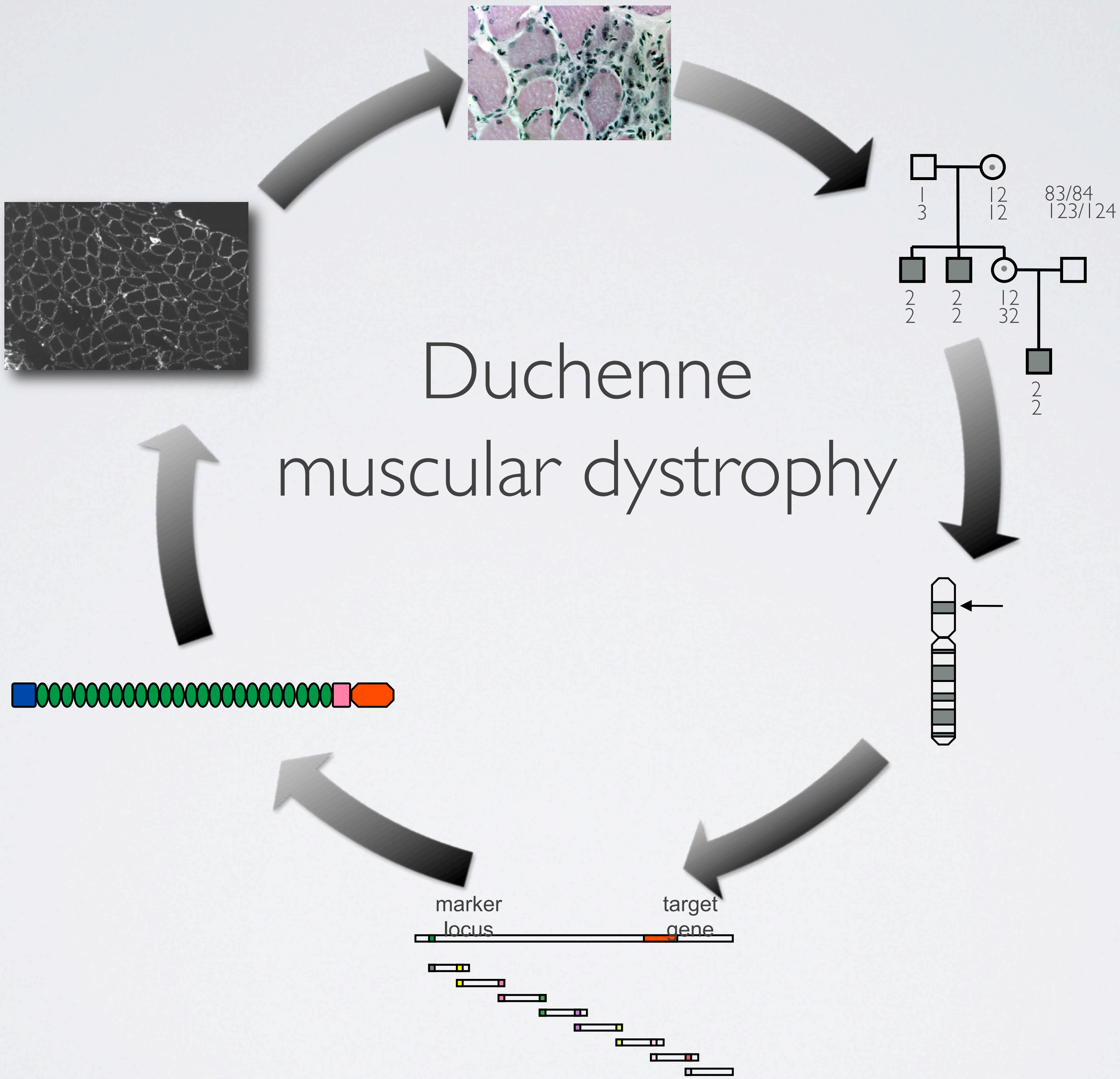


				θ				
Family	Sibs	Recombinants	Nonrecombinants	0	0.1	0.2	0.3	0.4
1	12	2	10	- ∞	1.15	1.25	1.02	0.60
2	9	2	7	- ∞	0.39	0.96	0.58	0.36
3	8	2	6	- ∞	0.13	0.43	0.43	0.28
4	10	2	8	- ∞	0.64	0.84	0.73	0.44
5	7	1	6	- ∞	0.83	0.83	0.65	0.38
Total	46	7	39	- ∞	3.14	4.31	3.41	2.06

Haplotype Analysis

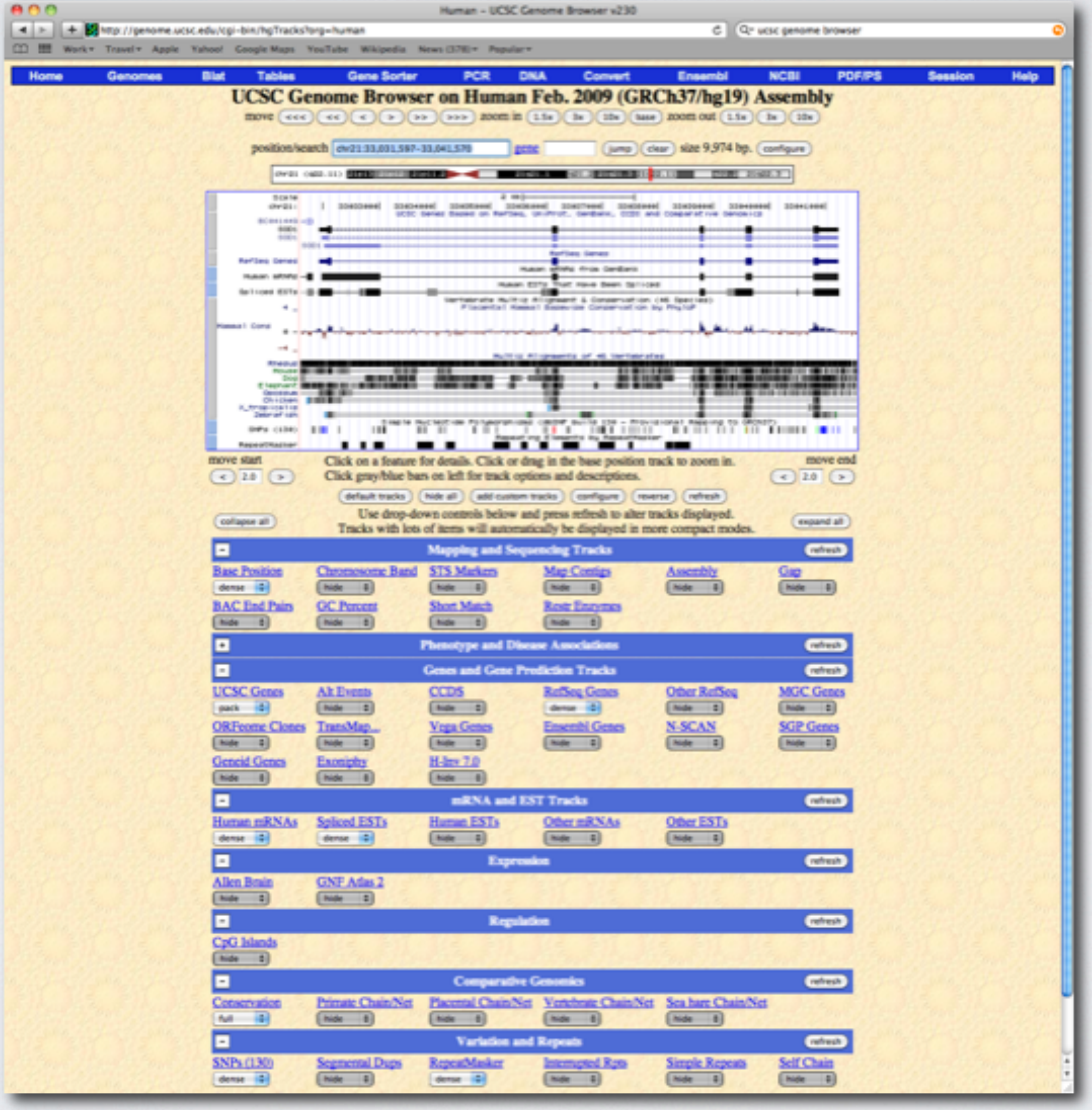


Positional Cloning



Duchenne muscular dystrophy

Genome Browser


Human – UCSC Genome Browser v230
http://genome.ucsc.edu/cgi-bin/hgTracks?org=human
ucsc genome browser
Work Travel Apple Yahoo! Google Maps YouTube Wikipedia News (378) Popular
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump clear configure
default tracks hide all add custom tracks configure reverse refresh
collapse all expand all
refresh

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

position/search chr21:33,031,597-33,041,570 gene size 9,974 bp.

move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in.
Click gray/blue bars on left for track options and descriptions.

move end < 2.0 >

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

Base Position: dense
Chromosome Band: hide
STS Markers: hide
Map Contigs: hide
Assembly: hide
Gap: hide
BAC End Pairs: hide
GC Percent: hide
Short Match: hide
Restr Enzymes: hide

Phenotype and Disease Associations

Genes and Gene Prediction Tracks

UCSC Genes: pack
Alt Events: hide
CCDS: hide
RefSeq Genes: dense
Other RefSeq: hide
MGC Genes: hide
ORFeome Clones: hide
TransMap...: hide
Vega Genes: hide
Ensembl Genes: hide
N-SCAN: hide
SGP Genes: hide
Geneid Genes: hide
Exoniphy: hide
H-Inv 7.0: hide

mRNA and EST Tracks

Human mRNAs: dense
Spliced ESTs: dense
Human ESTs: hide
Other mRNAs: hide
Other ESTs: hide

Expression

Allen Brain: hide
GNF Atlas 2: hide

Regulation

CpG Islands: hide

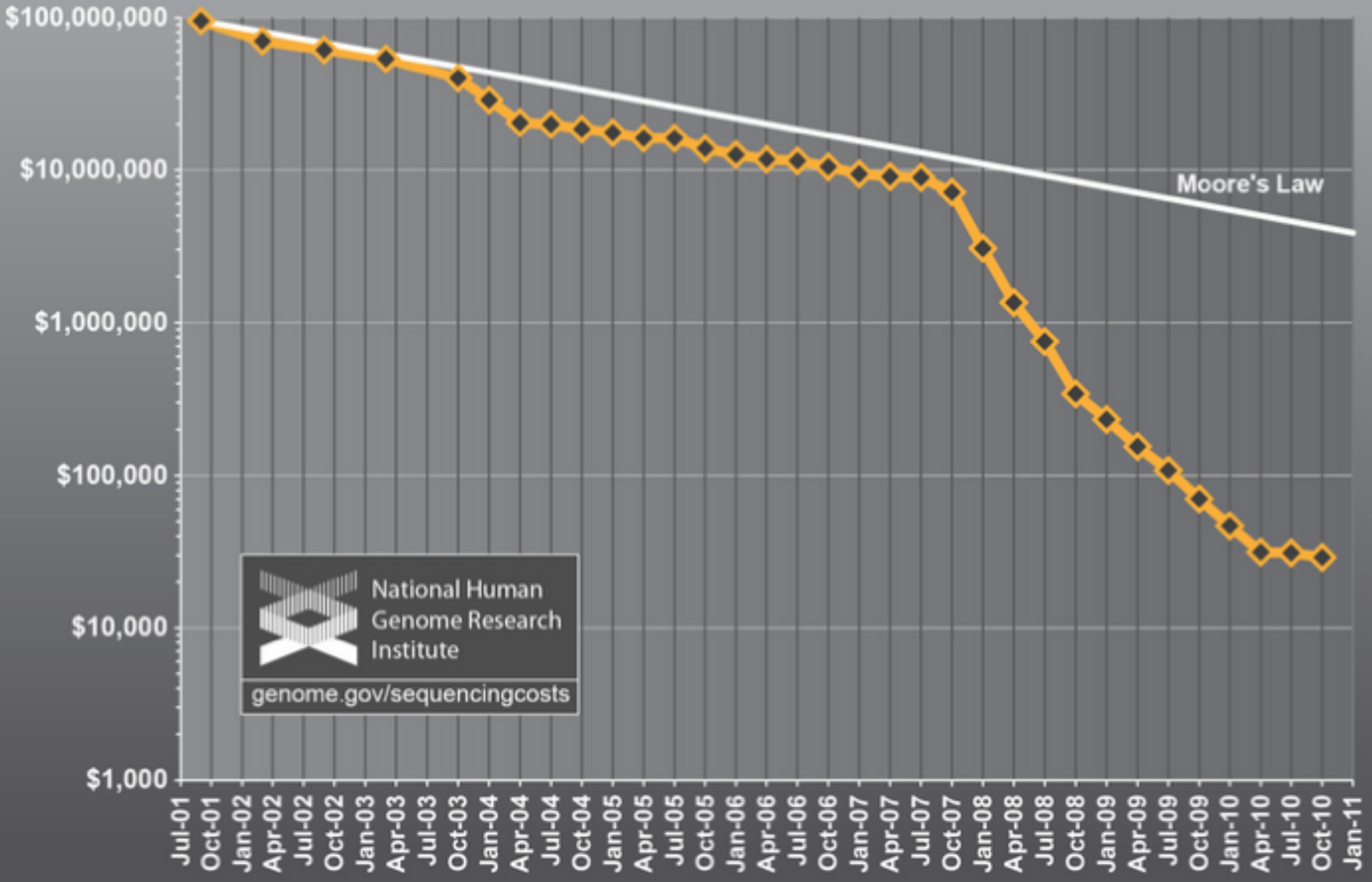
Comparative Genomics

Conservation: full
Primate Chain/Net: hide
Placental Chain/Net: hide
Vertebrate Chain/Net: hide
Sea hare Chain/Net: hide

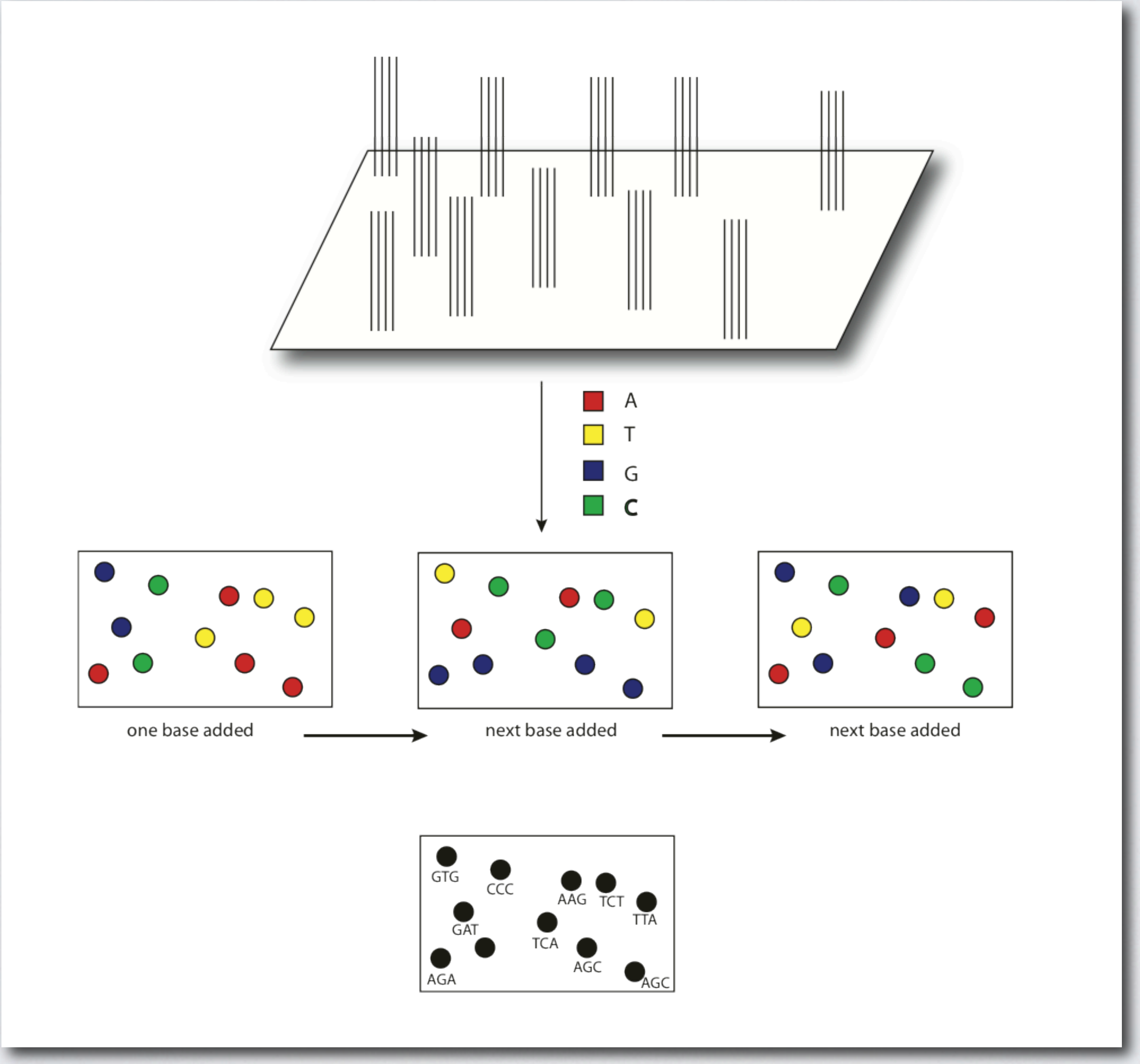
Variation and Repeats

SNPs (130): dense
Segmental Dups: hide
RepeatMasker: dense
Interrupted Rpts: hide
Simple Repeats: hide
Self Chain: hide

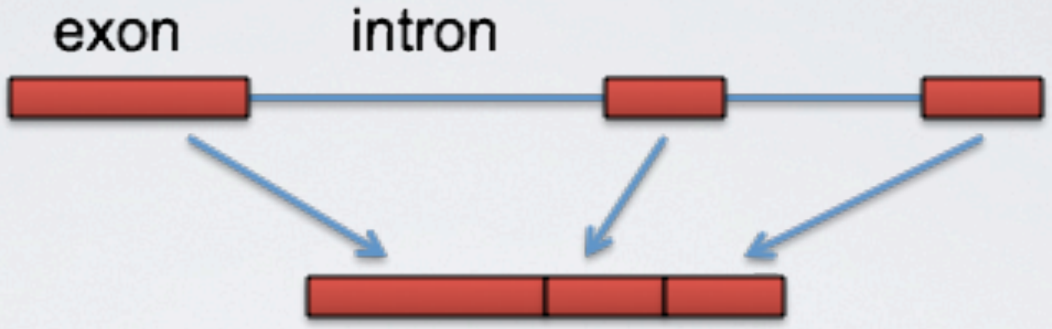
Cost per Genome

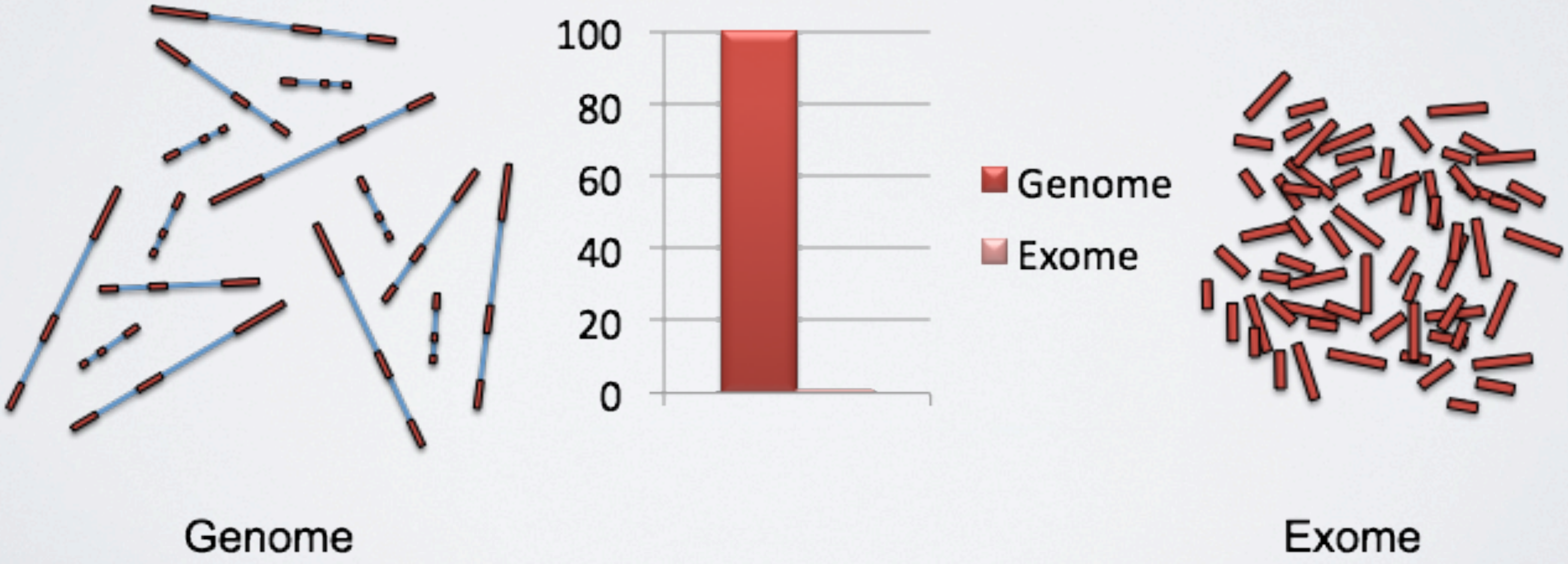


Massively Parallel Sequencing

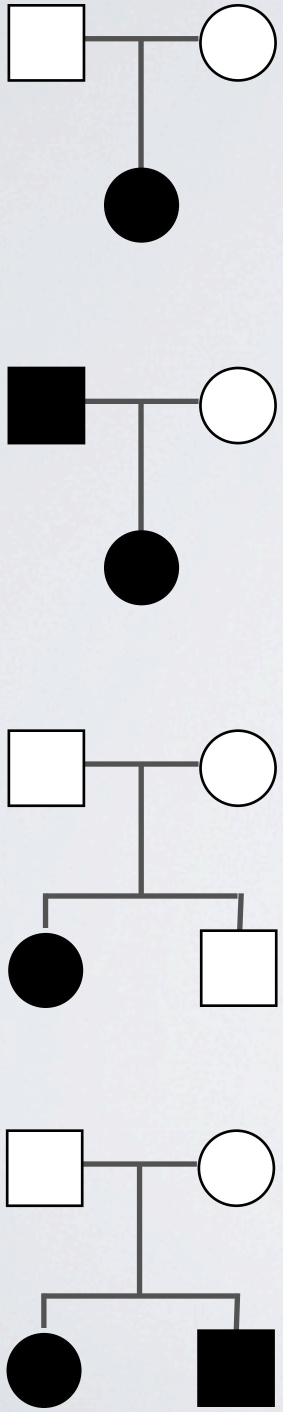


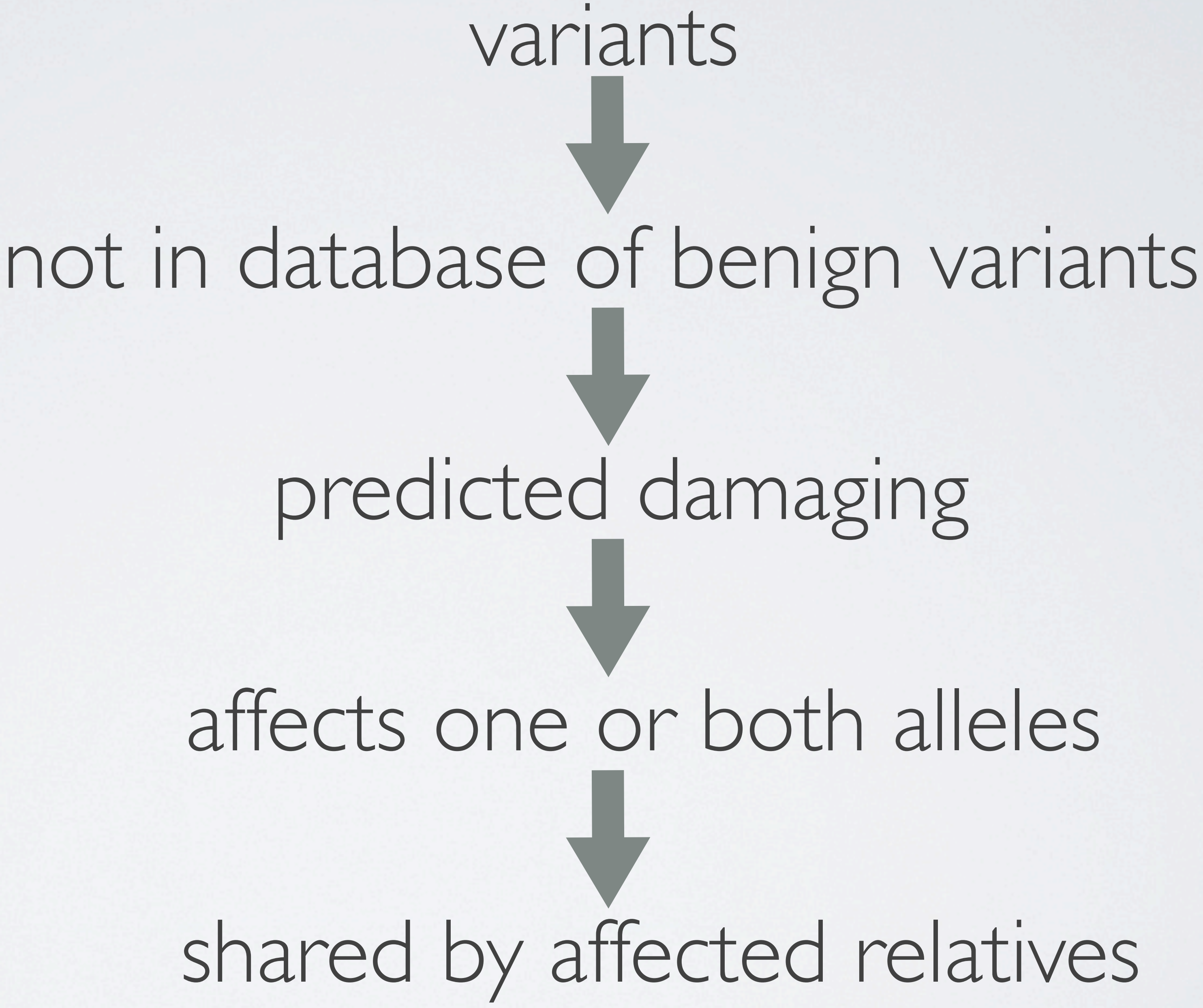
Exome vs. Genome Sequencing





Gene discovery





variants

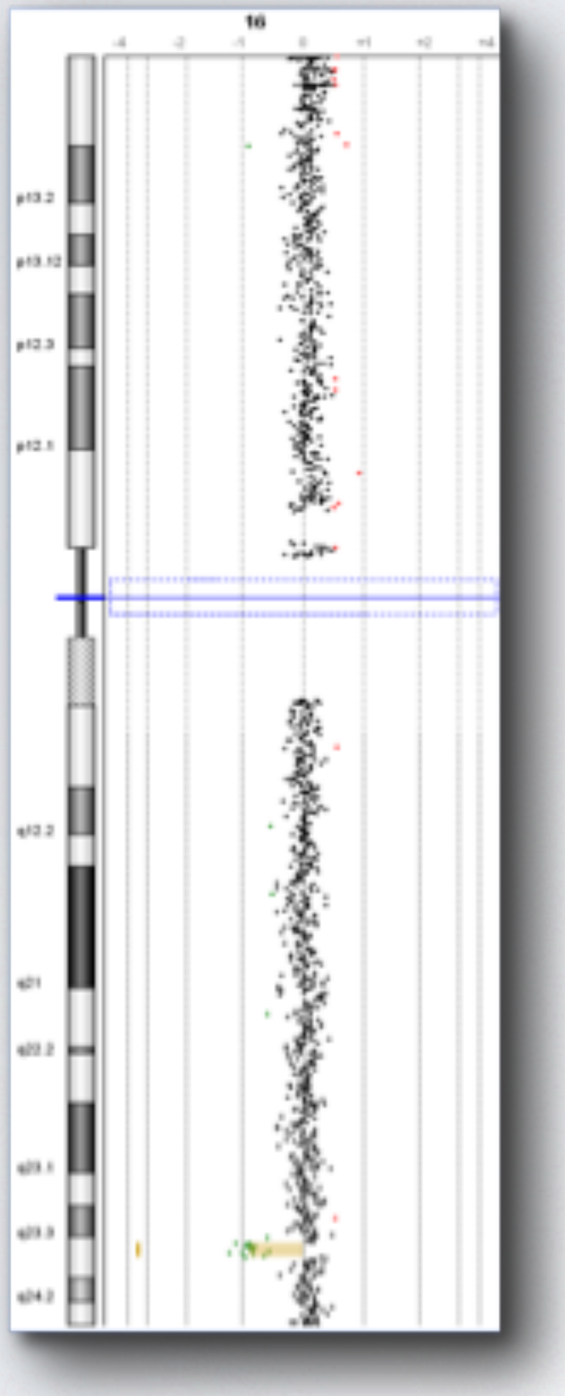
not in database of benign variants

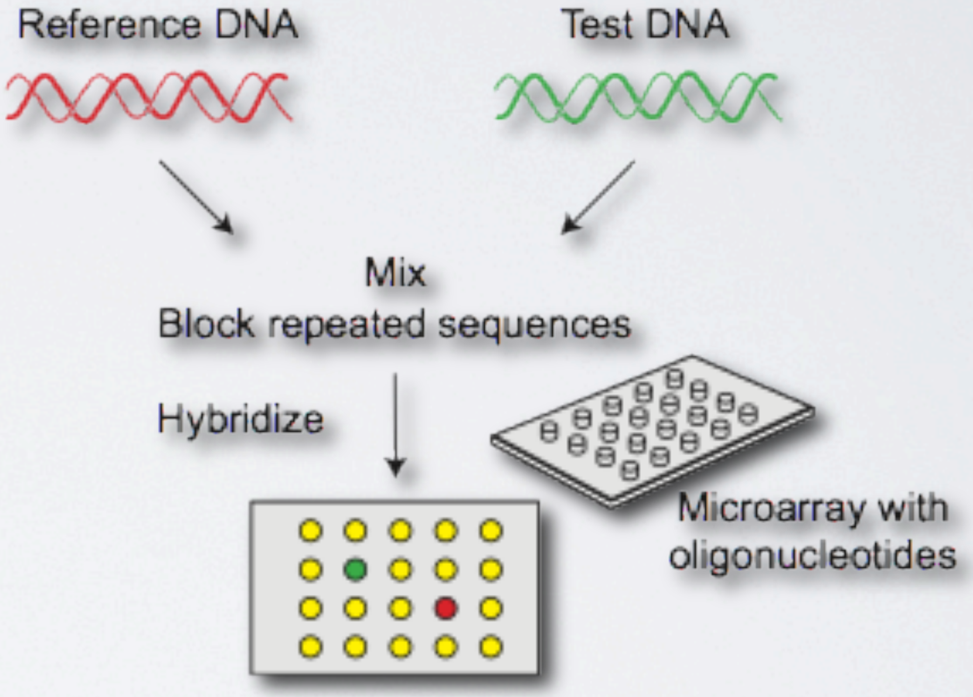
predicted damaging

affects one or both alleles

shared by affected relatives

Cytogenomics





Reference DNA

Test DNA

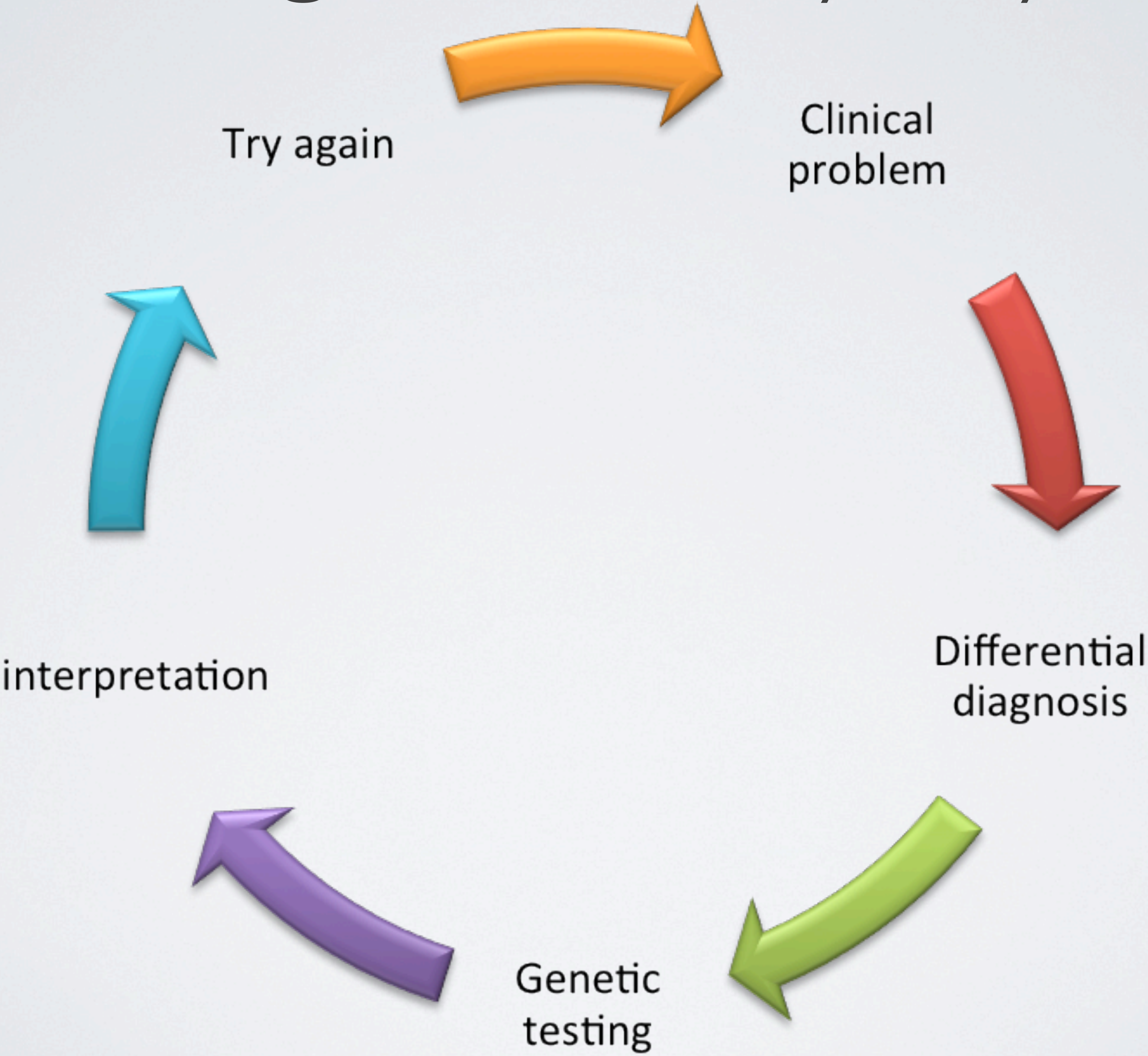
Mix

Block repeated sequences

Hybridize

Microarray with oligonucleotides

Diagnostic Odyssey



Clinical problem

Differential diagnosis

Genetic testing

interpretation

Try again