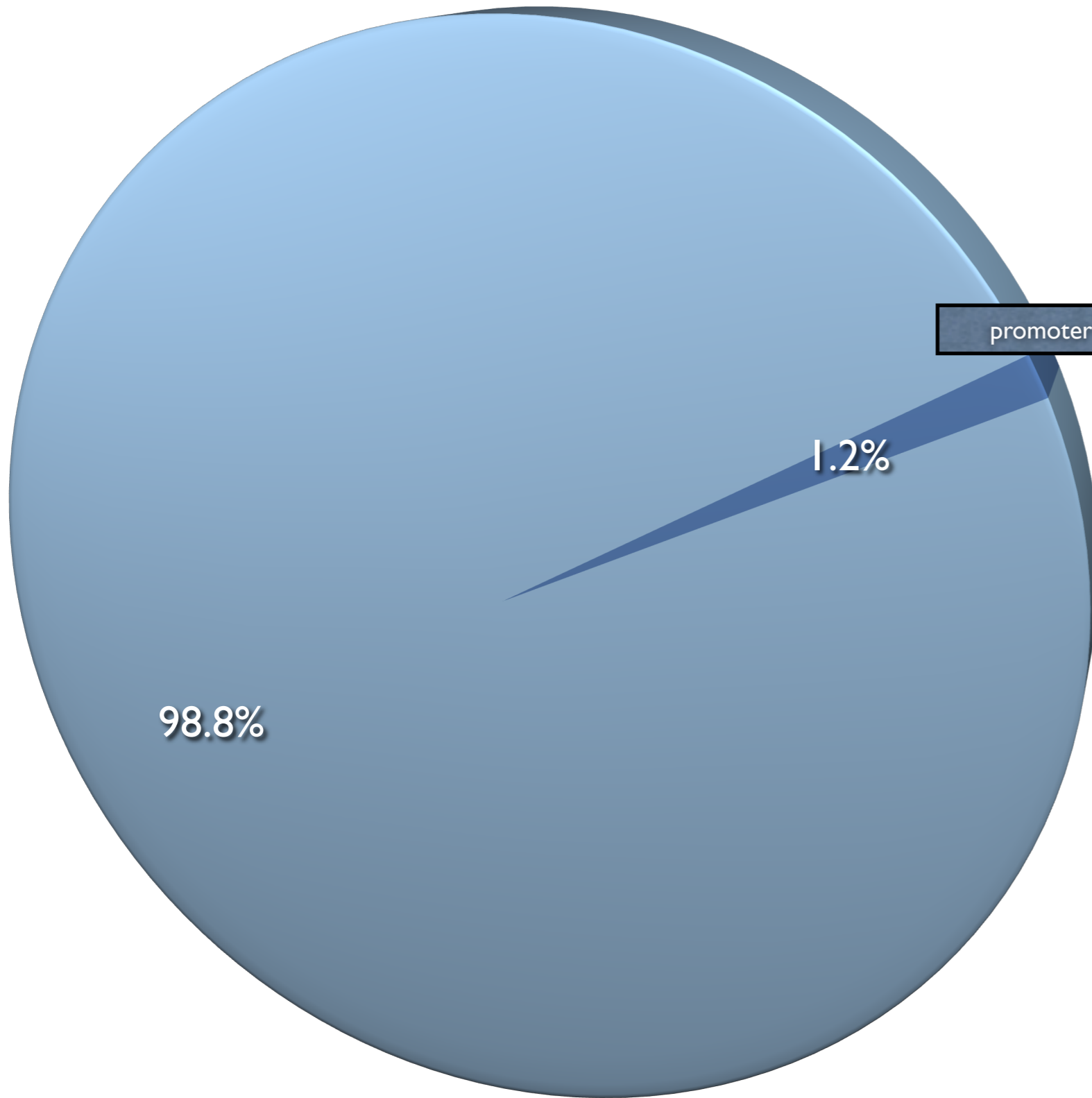
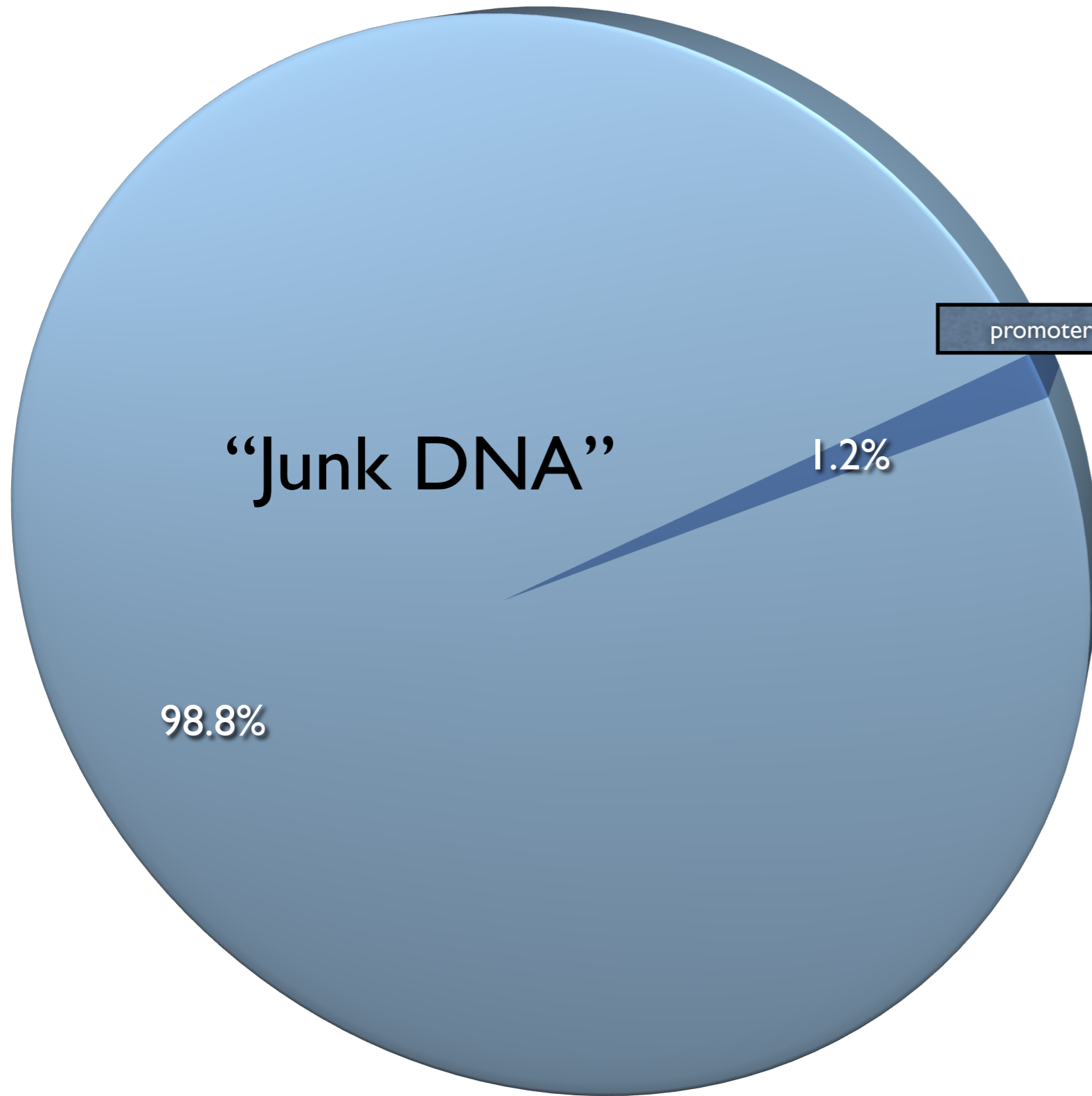


Approaches to Gene Discovery

Bruce R. Korf, MD, PhD

- The Human Genome
- Genetic Variation
- Gene Identification





"Junk DNA"

1.2%

98.8%

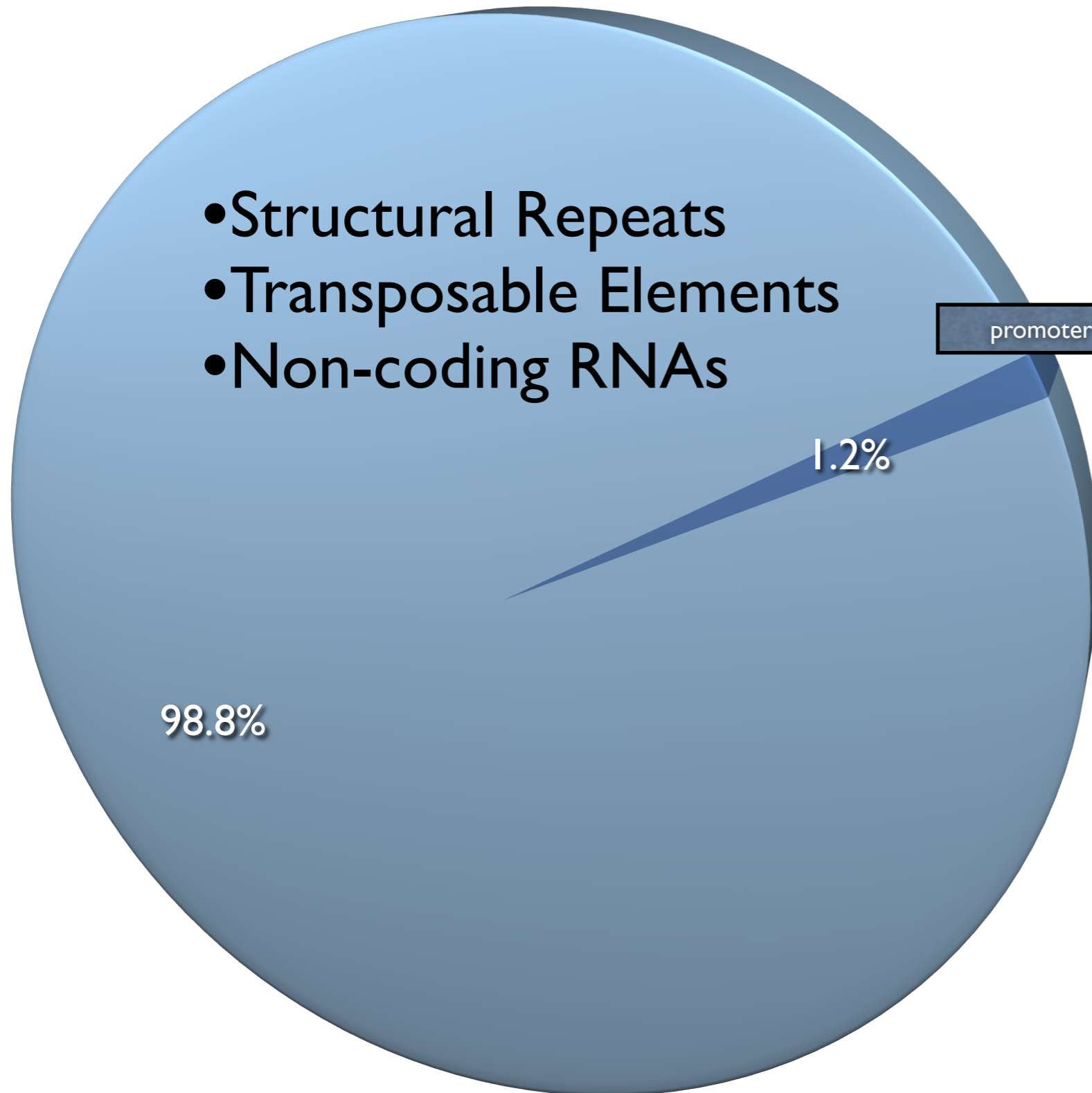
promoter

exon

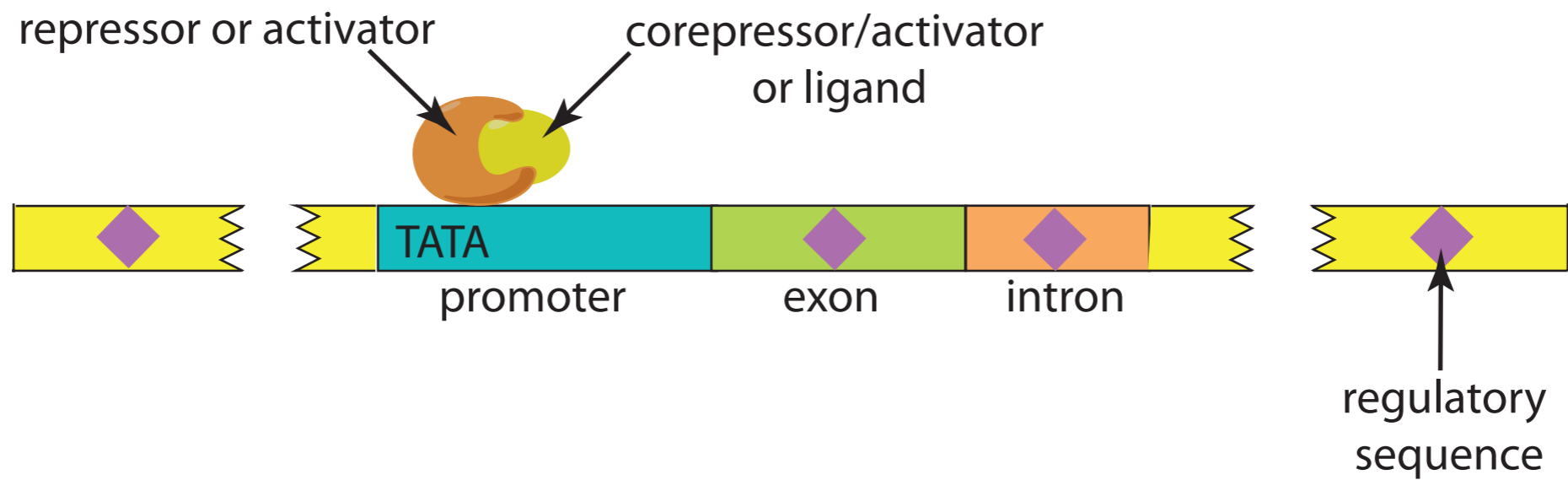
intron

exon

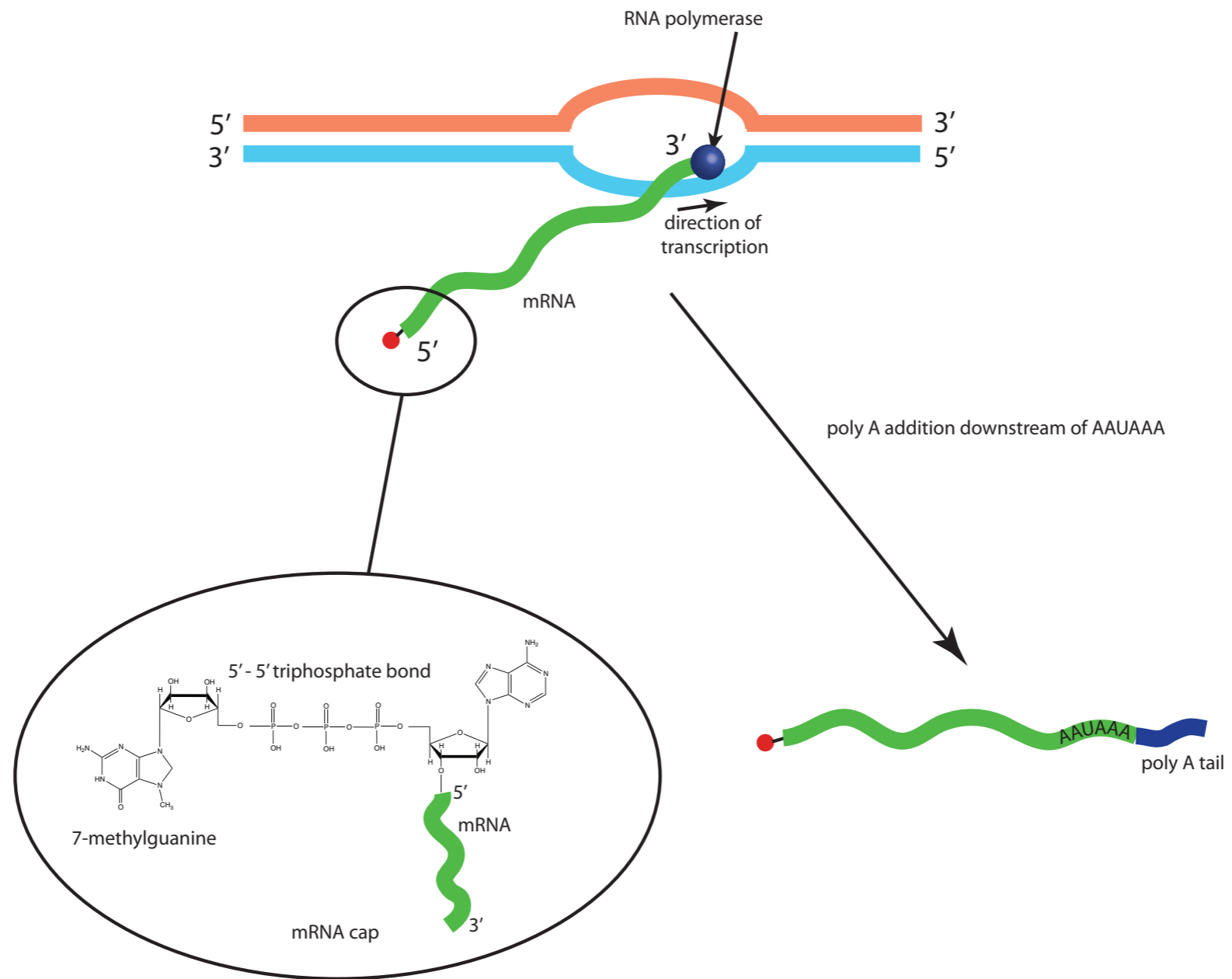
- Structural Repeats
- Transposable Elements
- Non-coding RNAs



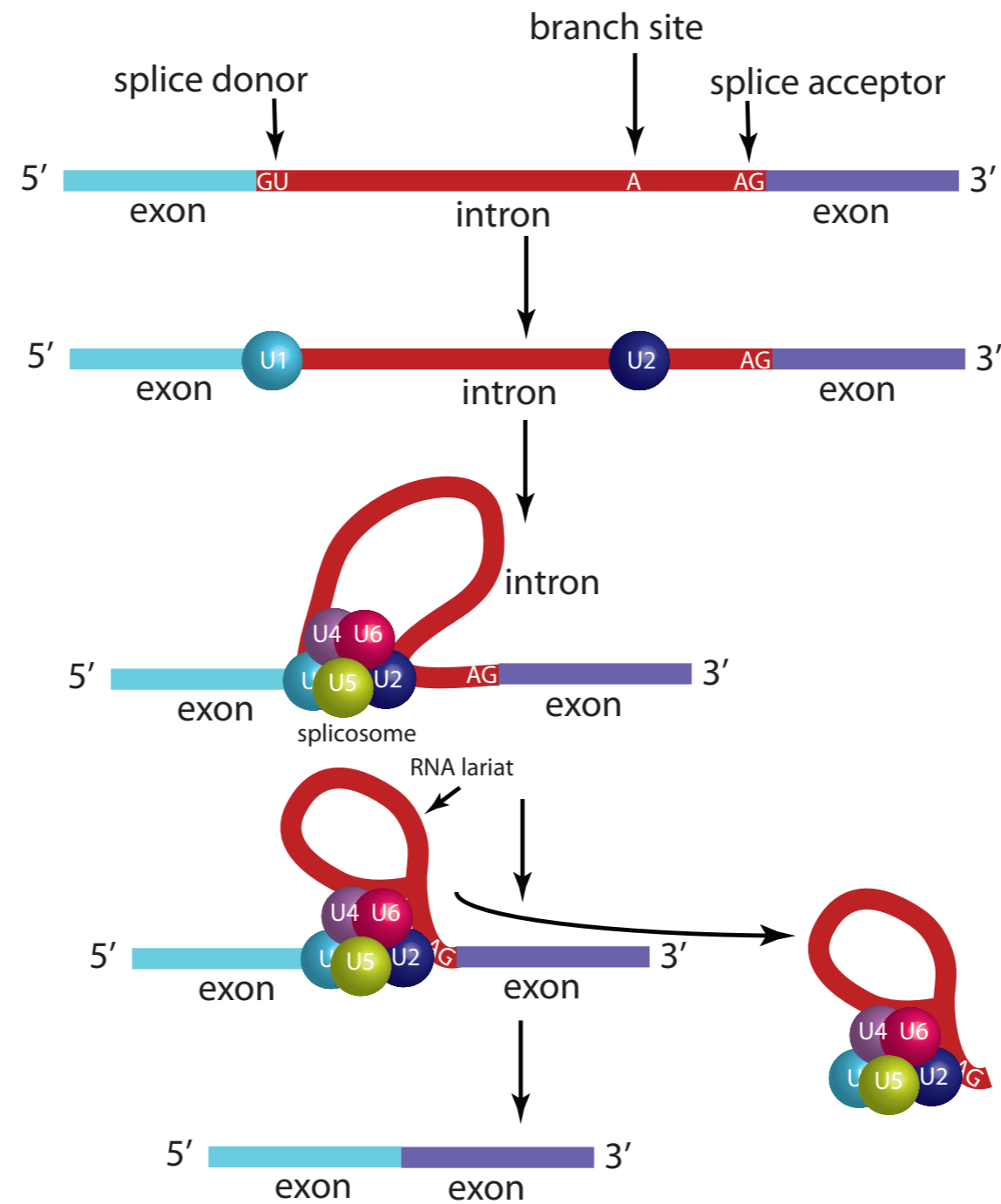
Gene Regulation



Transcription



Splicing

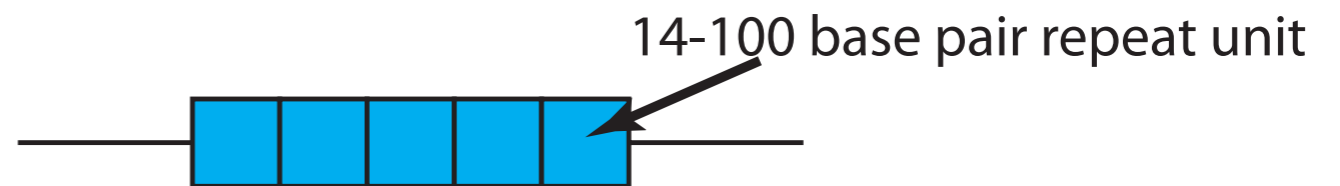


Repeated Sequences

simple sequence repeat

...GCGACACACACACACAGT...

variable number tandem repeat



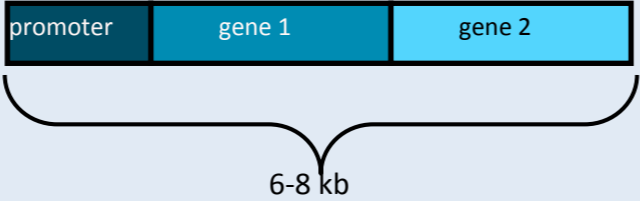
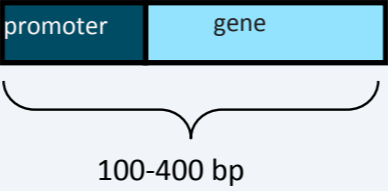
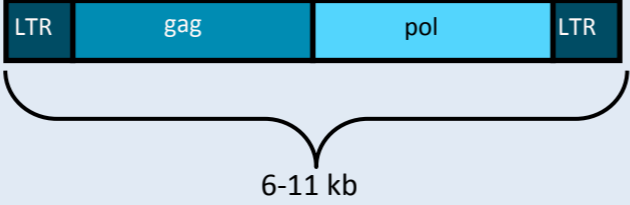
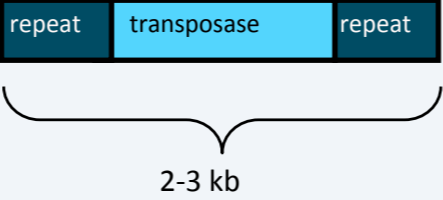
highly repeated sequences at centromeric and subtelomeric regions



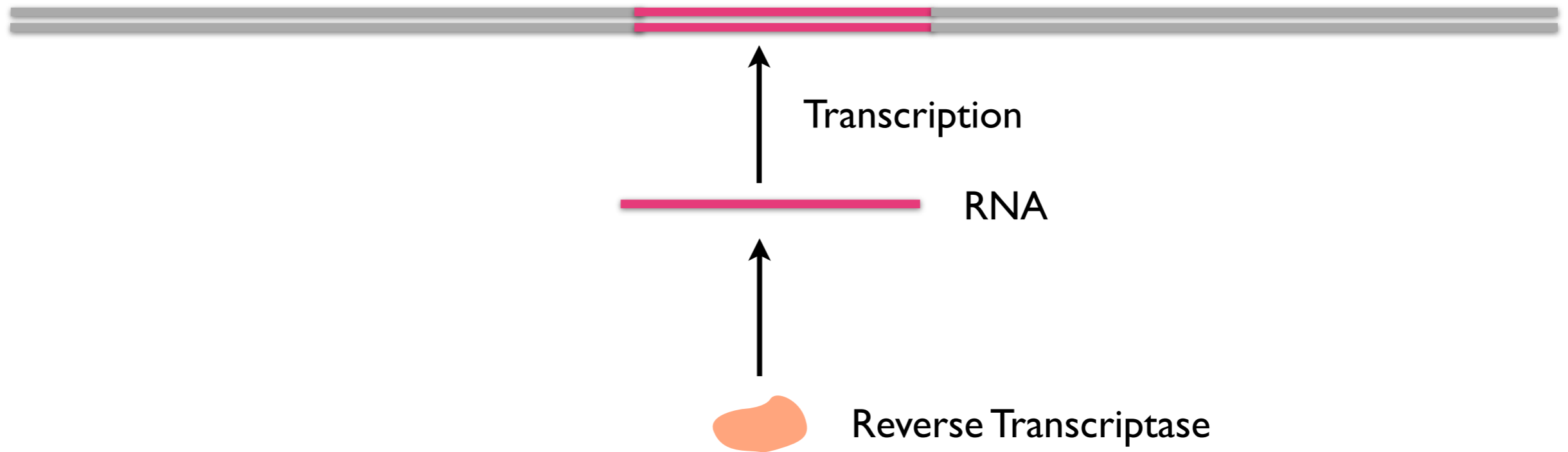
segmental duplications



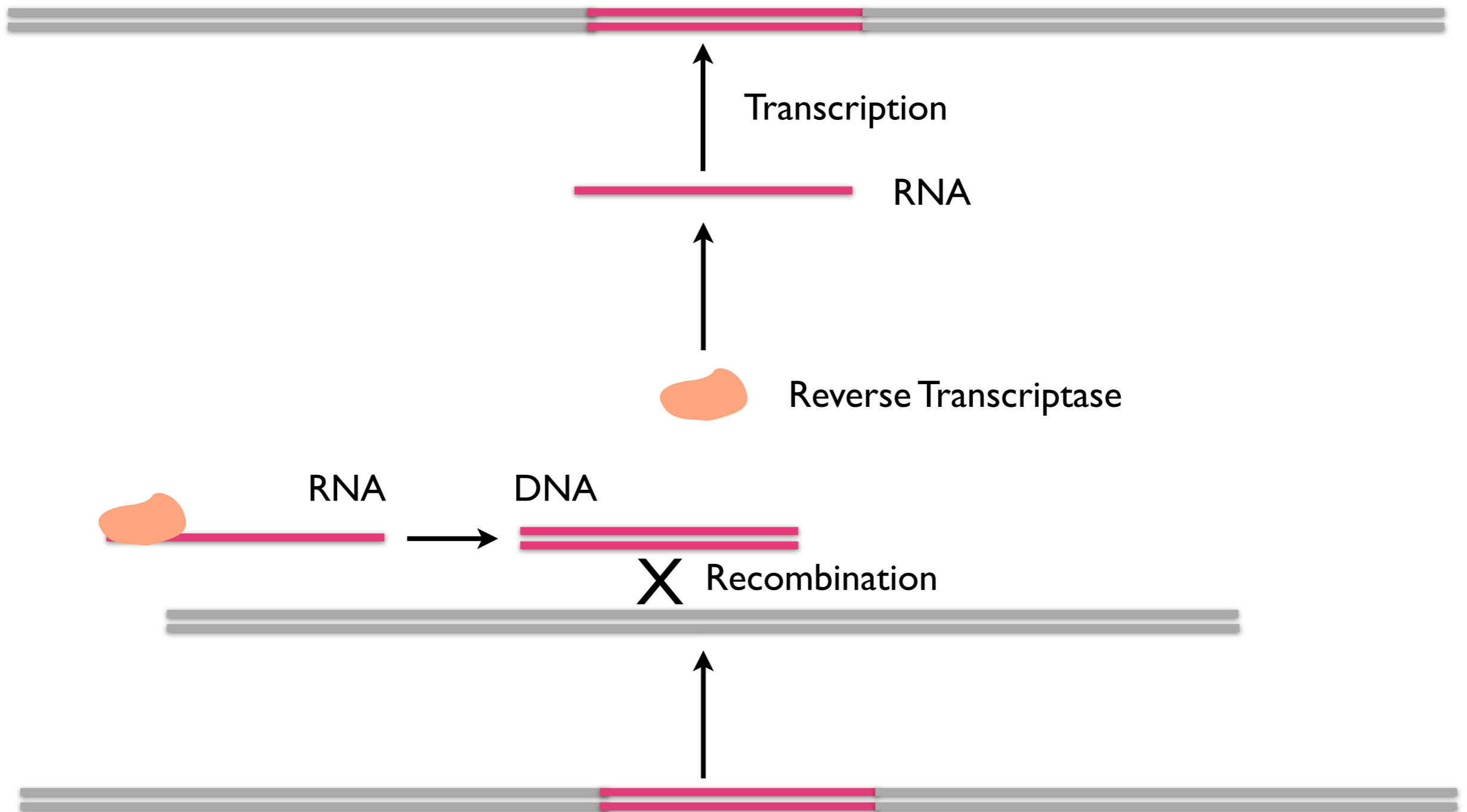
Transposable Genetic Elements

Type	Structure	Copy Number	Percent
LINE		850,000	21
SINE		1,500,000	13
Retroviral-like		450,000	8
Transposon		300,000	3

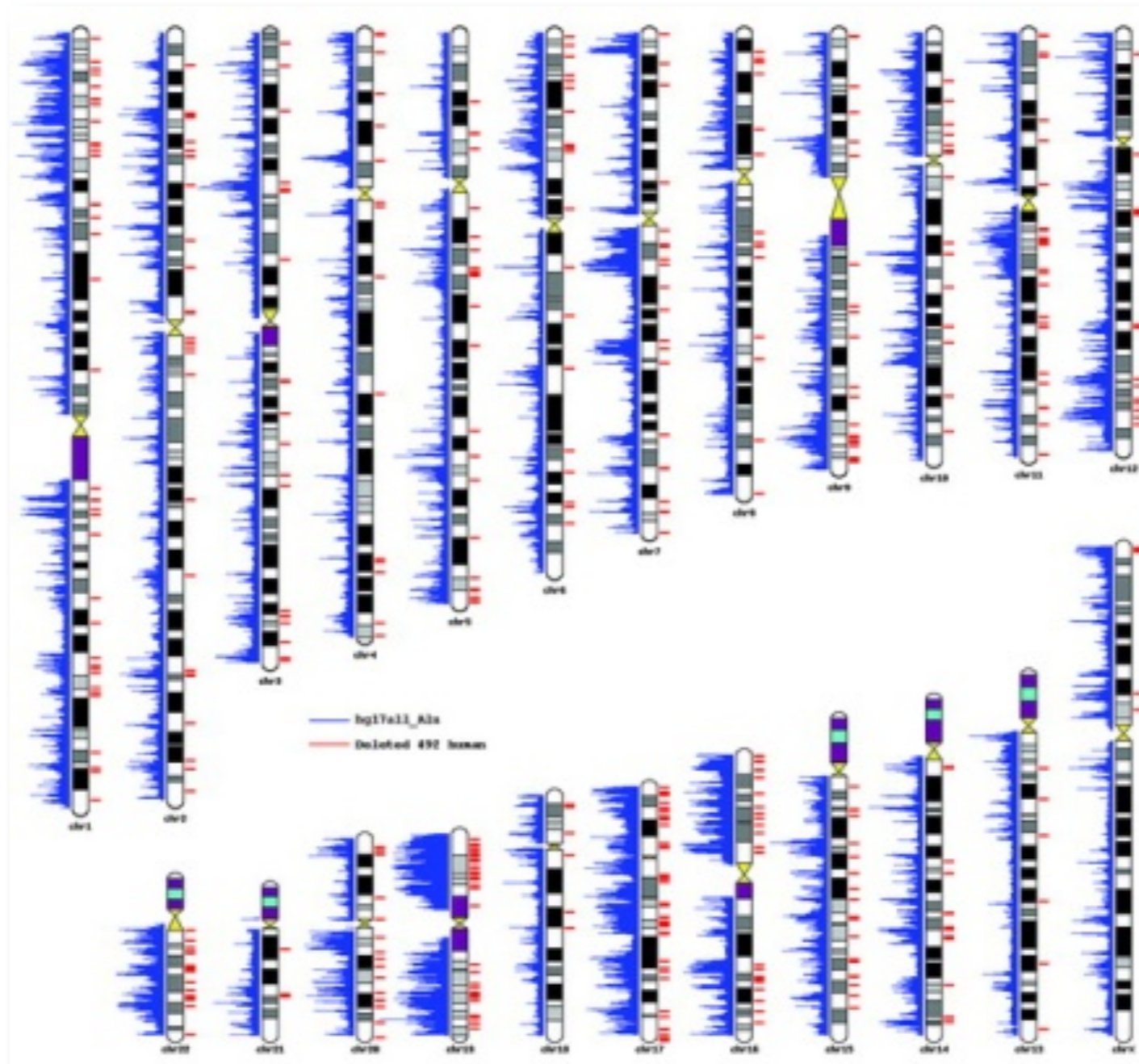
LINE “Life Cycle”



LINE “Life Cycle”



Alu Sequences



Sen SK, et al. *Am. J. Hum. Genet.*, 79:41-53, 2006

ENCODE Project

The screenshot shows the ENCODE Project website interface. At the top, there is a navigation bar with the 'nature' logo in red and 'ENCODE' in white. Below the navigation bar, there are links for 'Home', 'Research', 'Threads', 'Additional Research', 'News and Comment', 'About', and 'Sponsor'. The main content area is divided into two sections. On the left, there is a 'nature ENCODE explorer' section with a 'THREADS' heading. It features a vertical sequence of 13 numbered circles (01 to 13) arranged in a slightly curved path, each with a different color. On the right, there is a 'PRODUCED WITH SUPPORT FROM illumina' logo. Below the logo, there is a welcome message: 'Welcome to the nature ENCODE explorer'. The message continues: 'Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.' and ends with 'Select a thread to start'.

nature | **ENCODE**

Home | Research | Threads | Additional Research | News and Comment | About | Sponsor

nature **ENCODE** explorer

THREADS

01
02
03
04
05
06
07
08
09
10
11
12
13

PRODUCED WITH SUPPORT FROM **illumina**

Welcome to the **nature** **ENCODE** explorer

Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.

Select a thread to start

ENCODE Findings

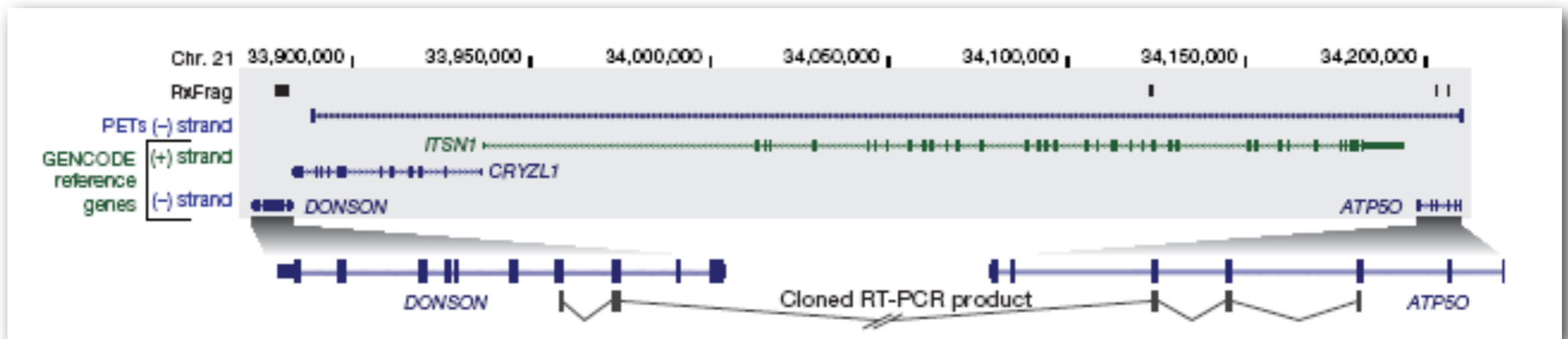
- annotated 20,687 protein-encoding genes
- average 6.3 alternatively spliced isoforms per gene
- 8,801 small RNAs; 9,640 long non-coding transcripts
- >80% genome transcribed in some cell type
- >400,000 enhancers and 70,000 promoters

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

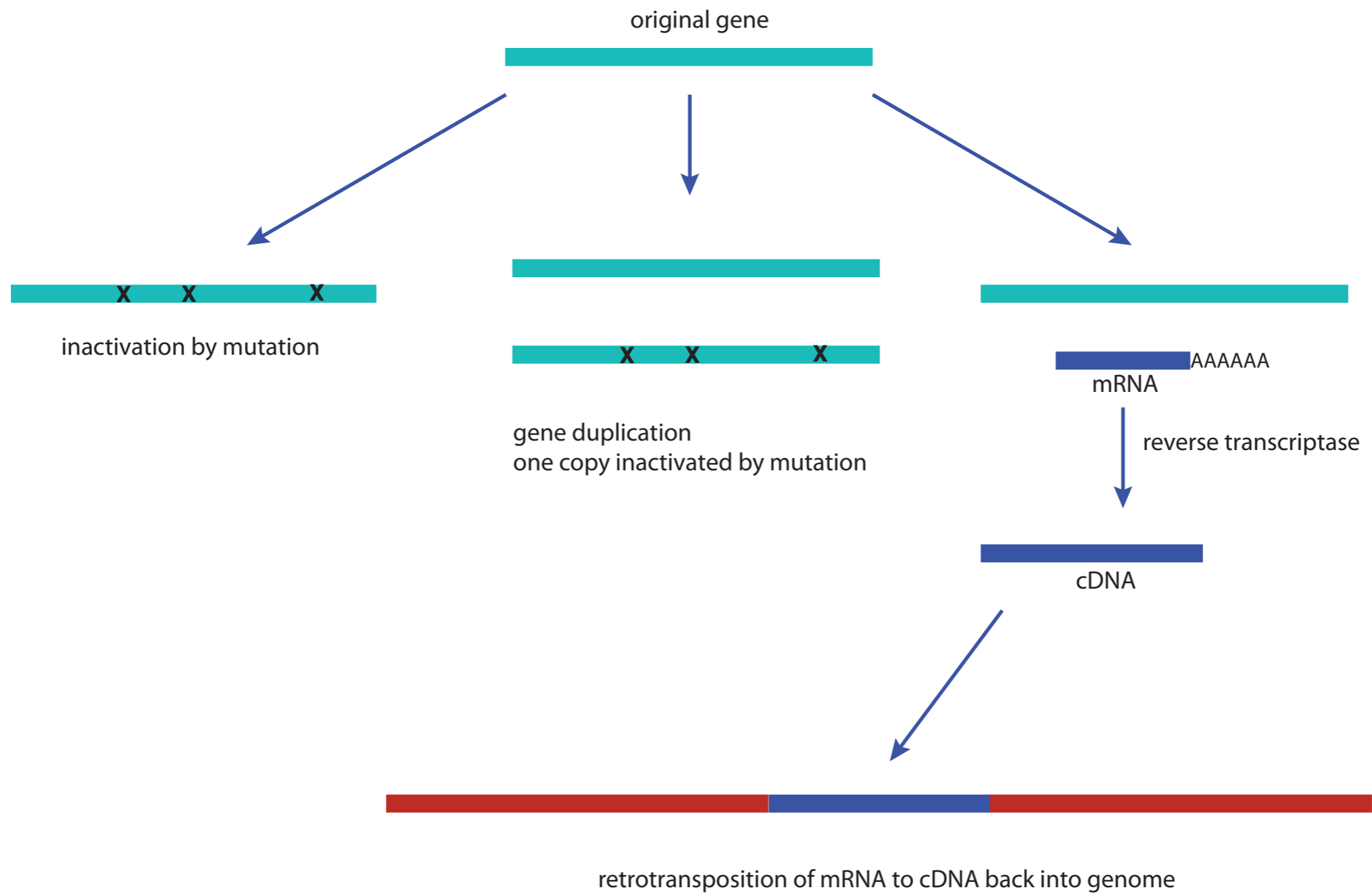
Long Non-Coding RNAs

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

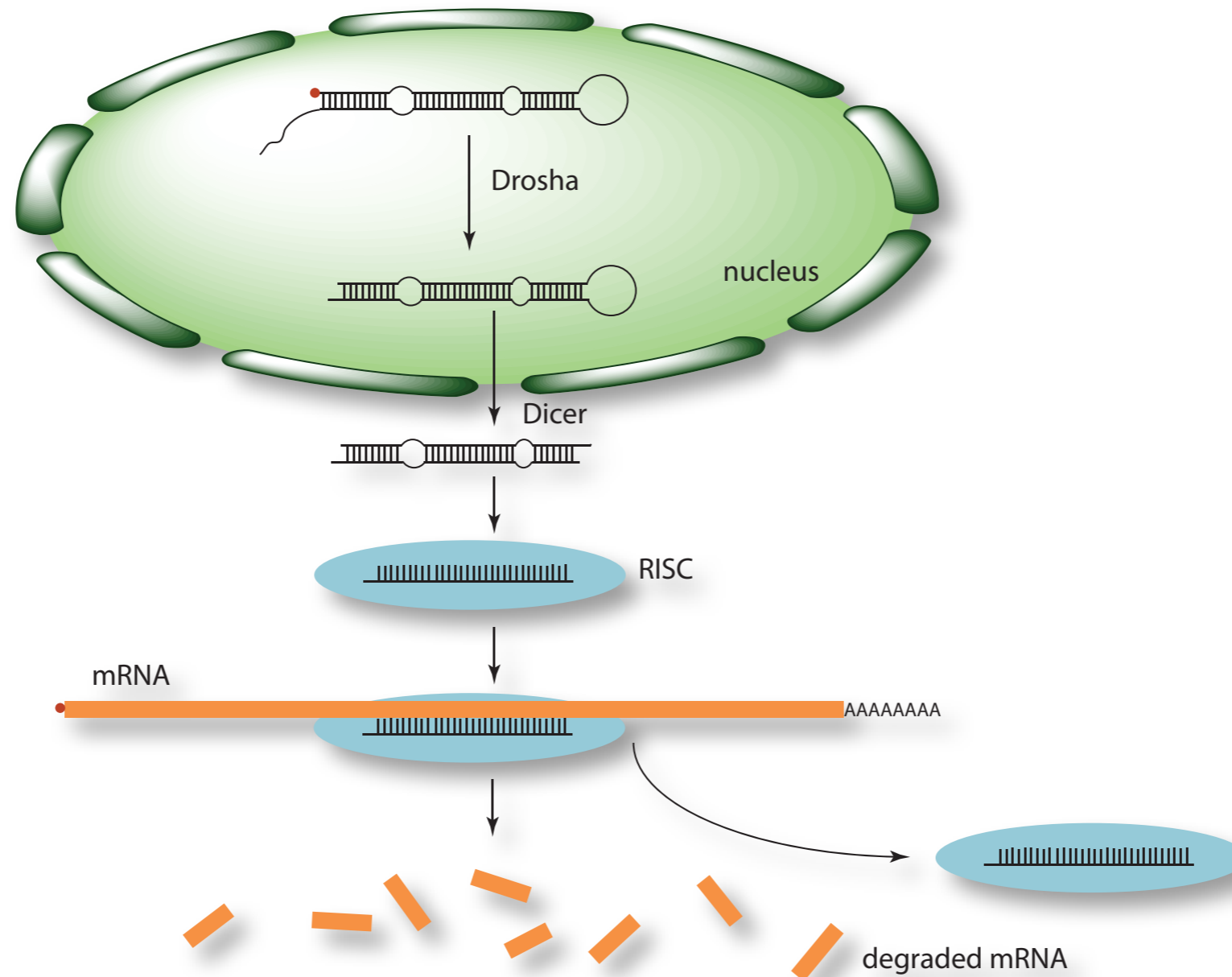


Nature 447:799, 2007

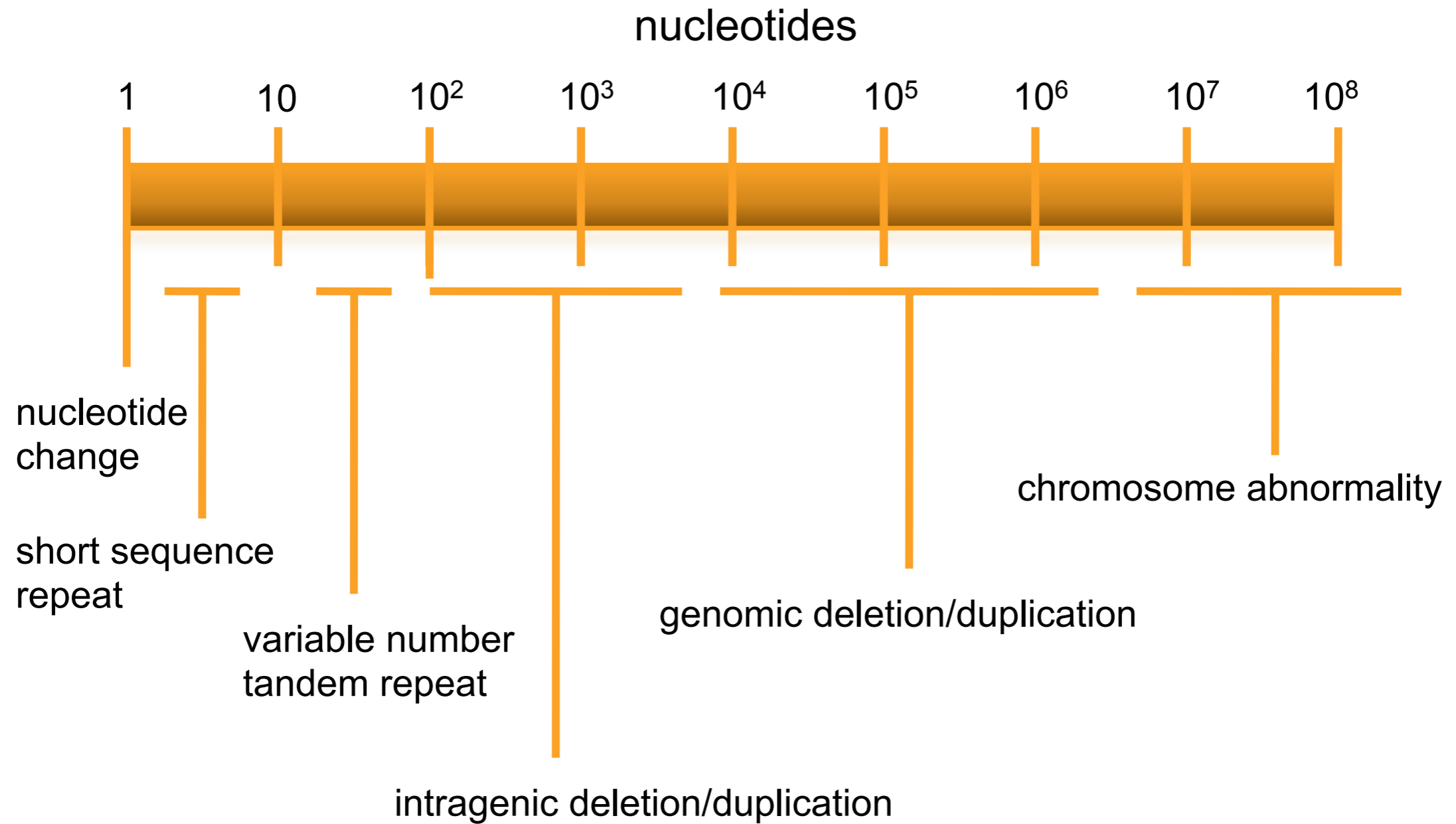
Pseudogenes



MicroRNA



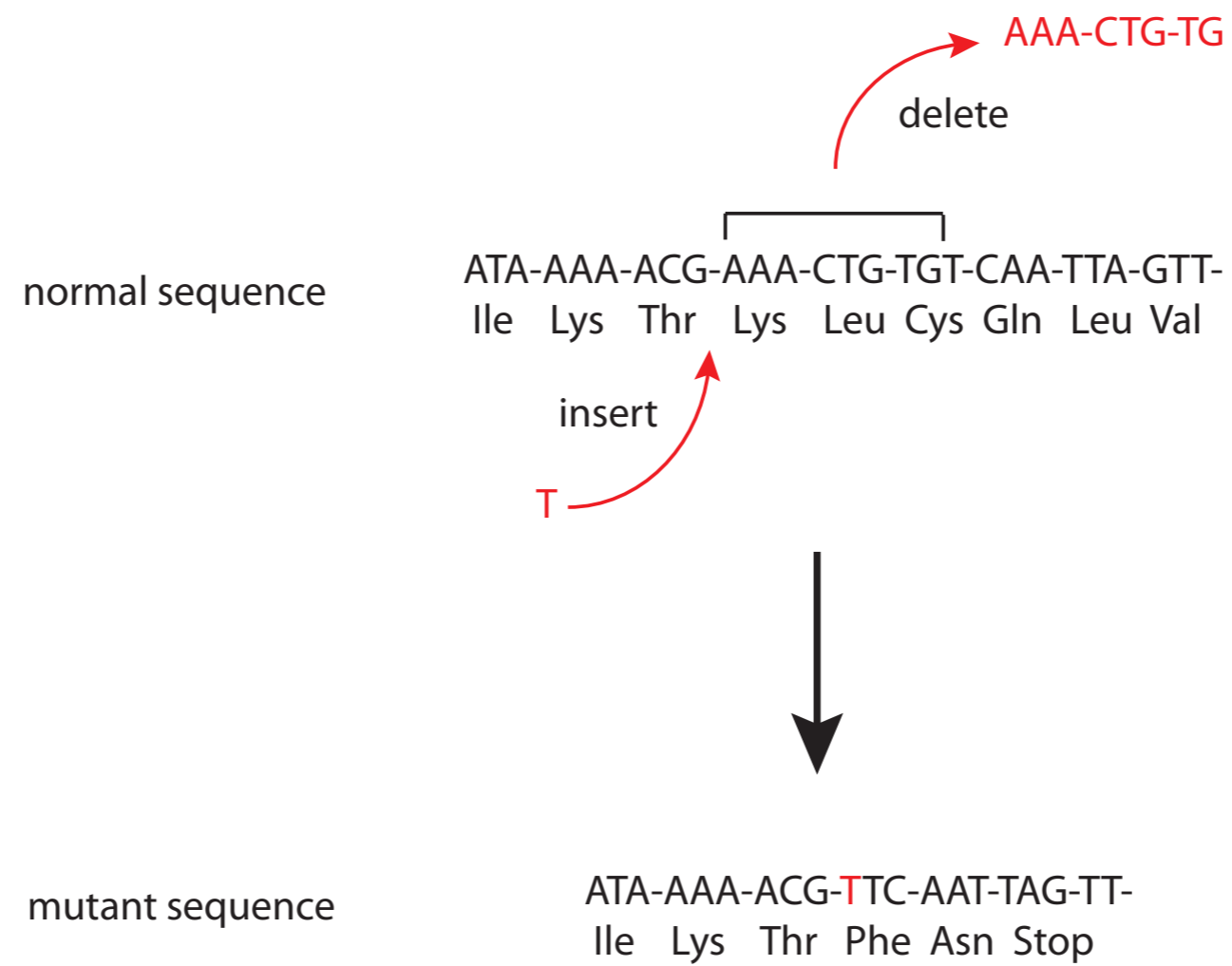
Genetic Variation



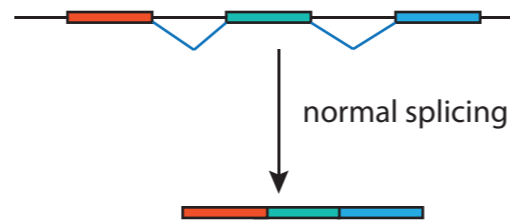
Point Mutations

TCC CAAATC 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 CAAATC CTC CCT CGA GTT ser gln ile leu pro arg val	conservative mutation
TCC CAAATC GTC GCT CGA GTT ser gln ile val ala arg val	non-conservative mutation
TCC CAAATC 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

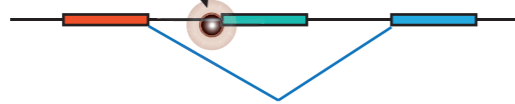
Indel



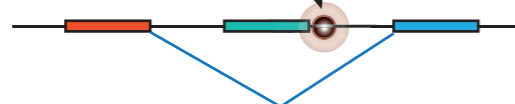
Splicing Mutations



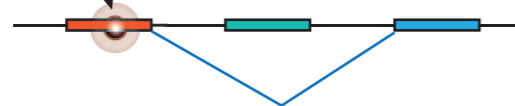
splice acceptor mutation



splice donor mutation



exon splice enhancer mutation



exon skip mutations

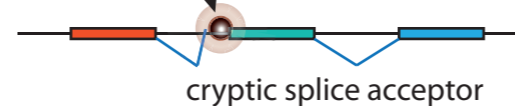
splice acceptor mutation



cryptic splice acceptor

truncated exon

splice acceptor mutation



cryptic splice acceptor

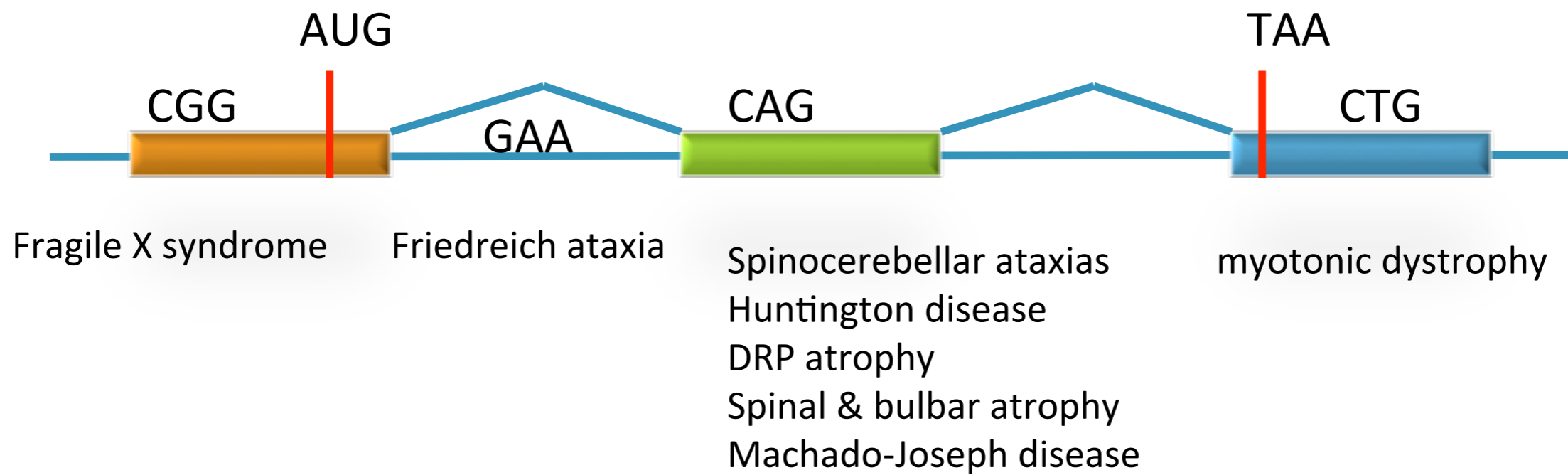
mutation creates new splice acceptor



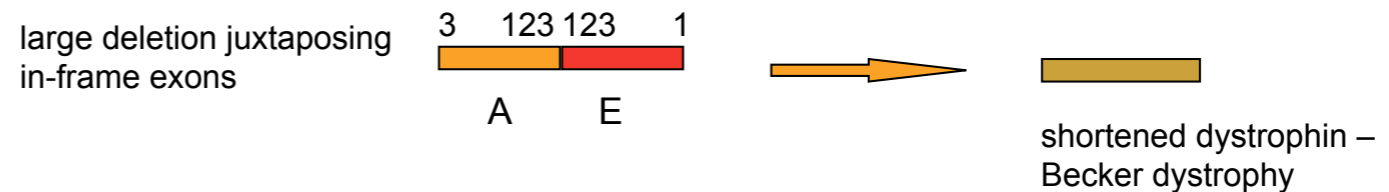
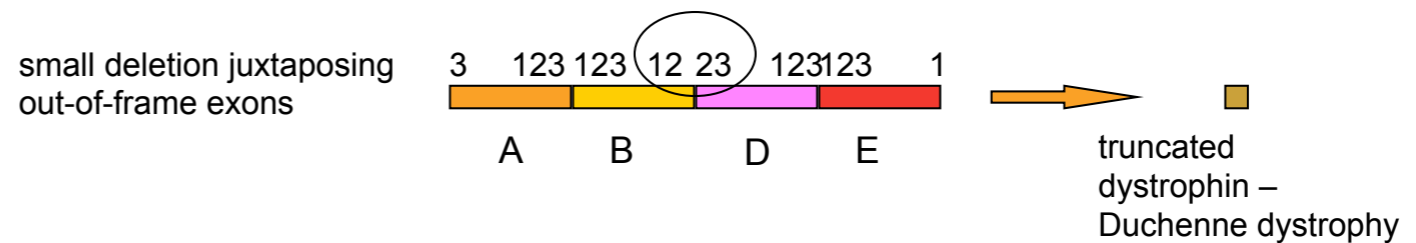
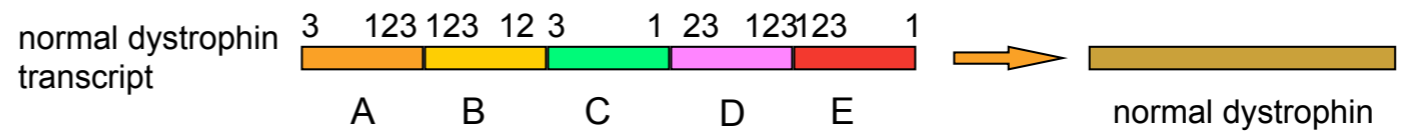
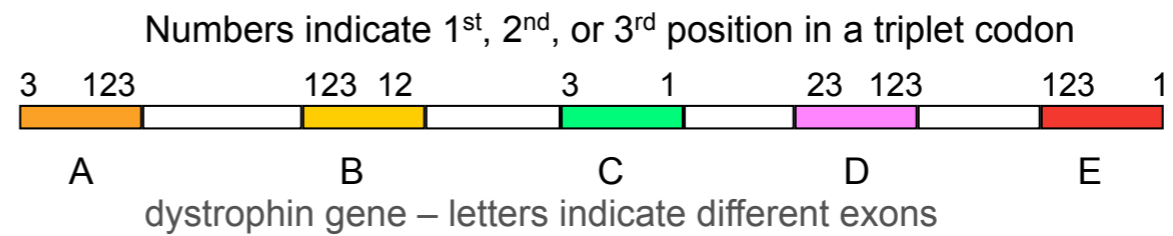
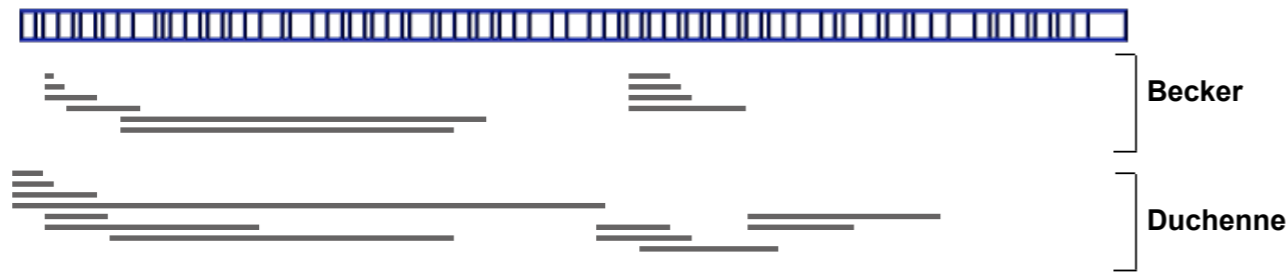
inclusion of intron in processed mRNA

cryptic splice acceptor (or donor) mutations

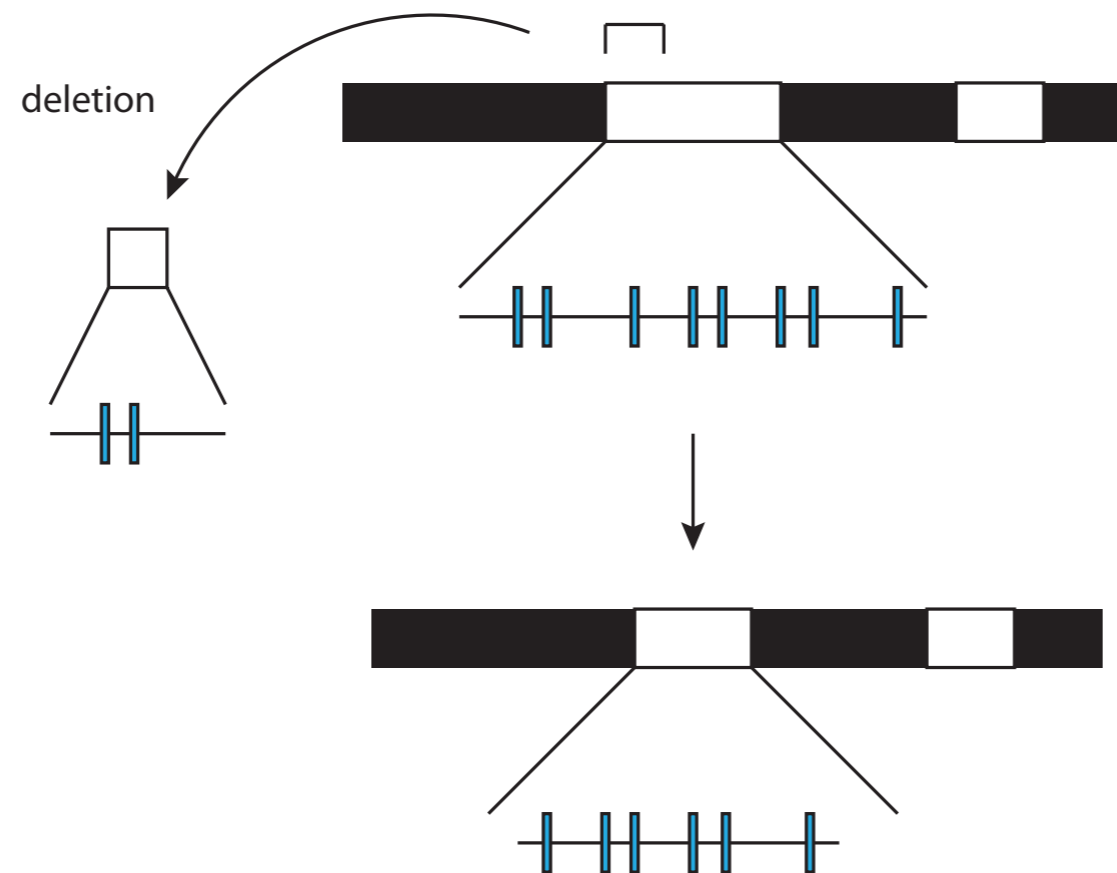
Triplet Repeat Expansions



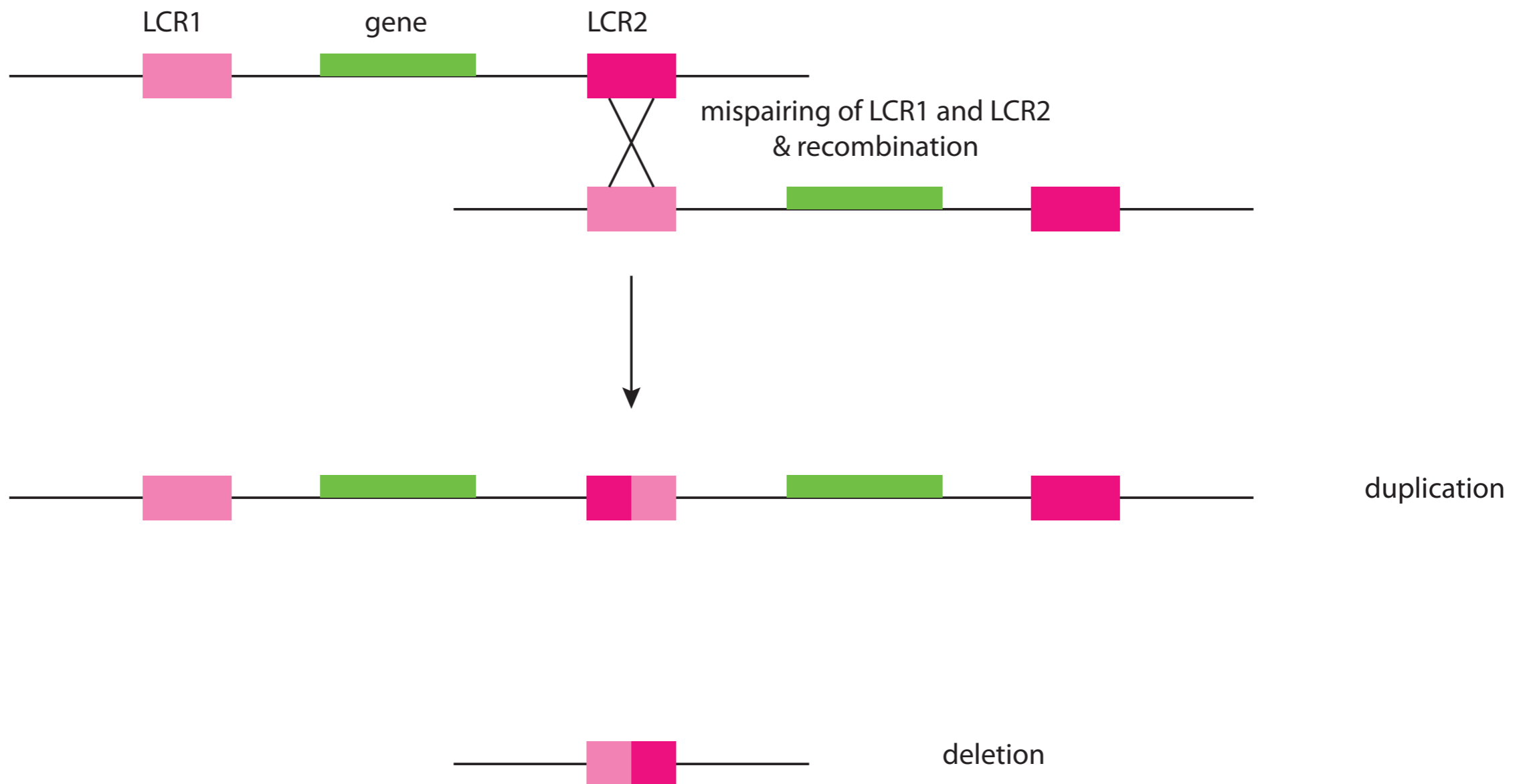
Multiexon Deletion



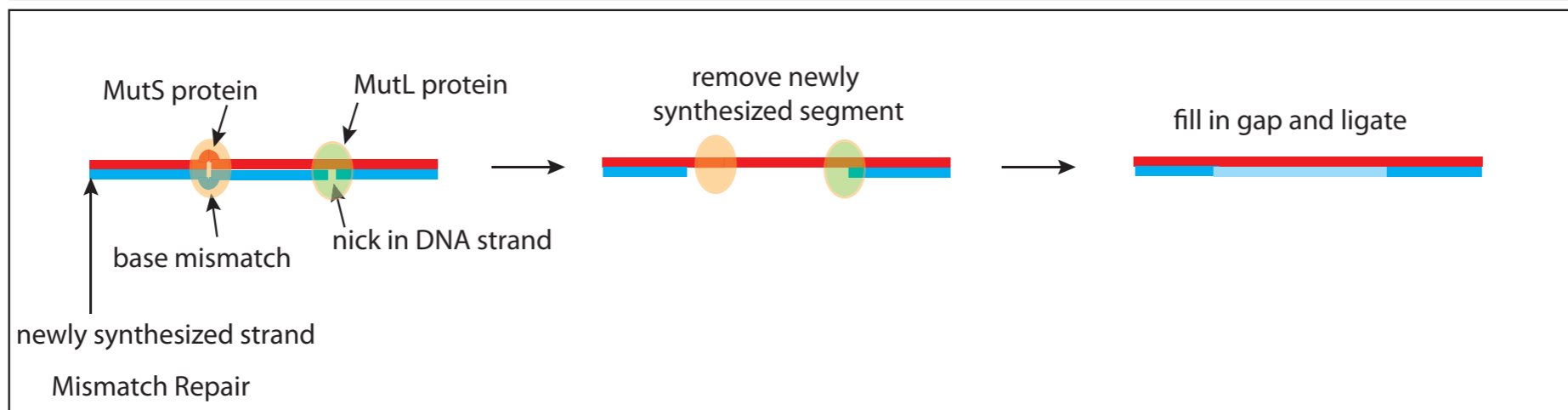
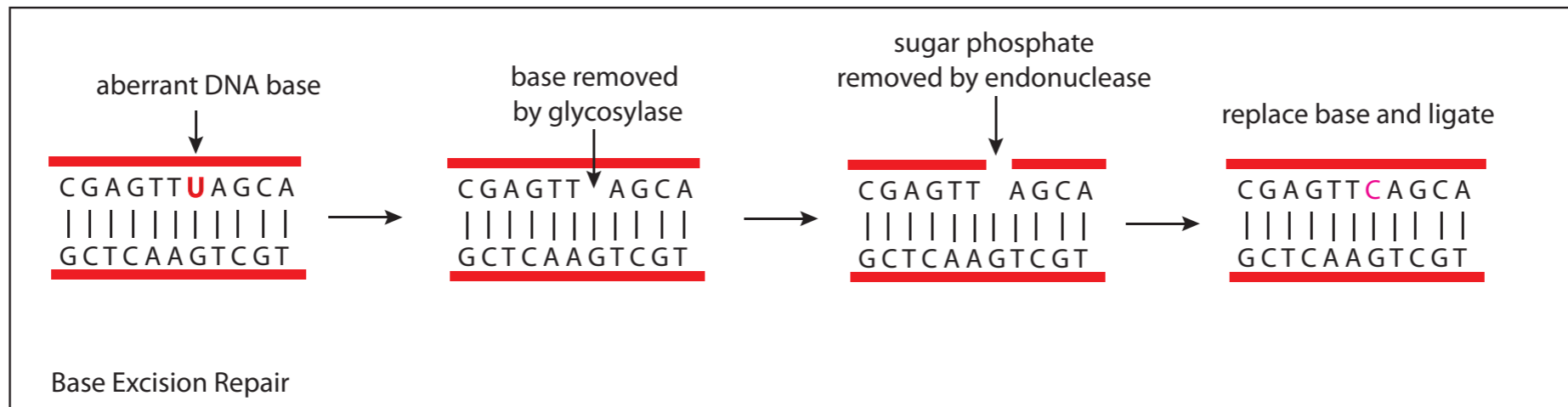
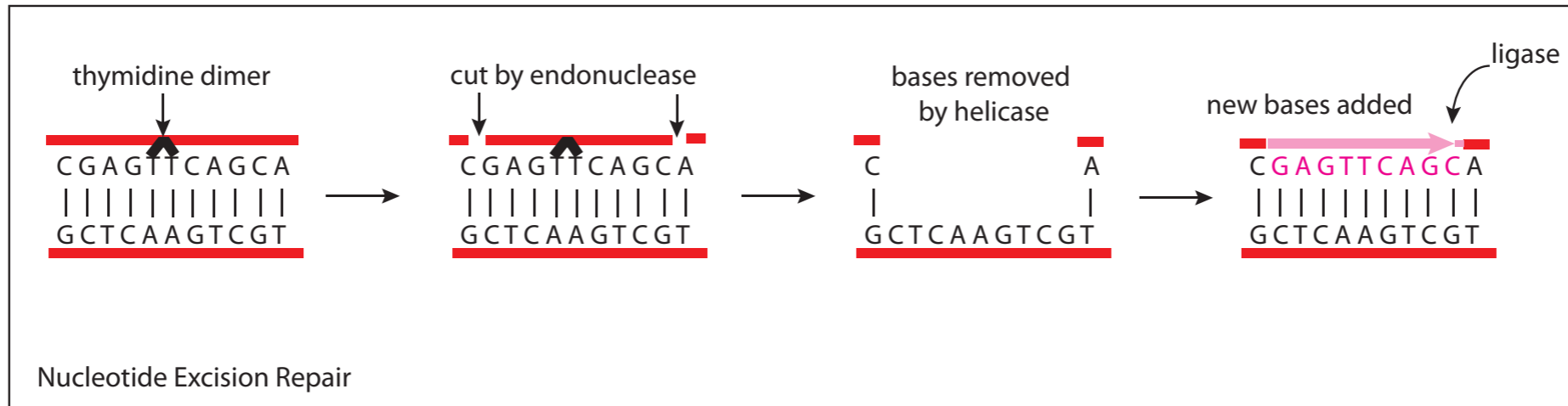
Chromosome Microdeletion



LCR Mispairing



DNA 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.

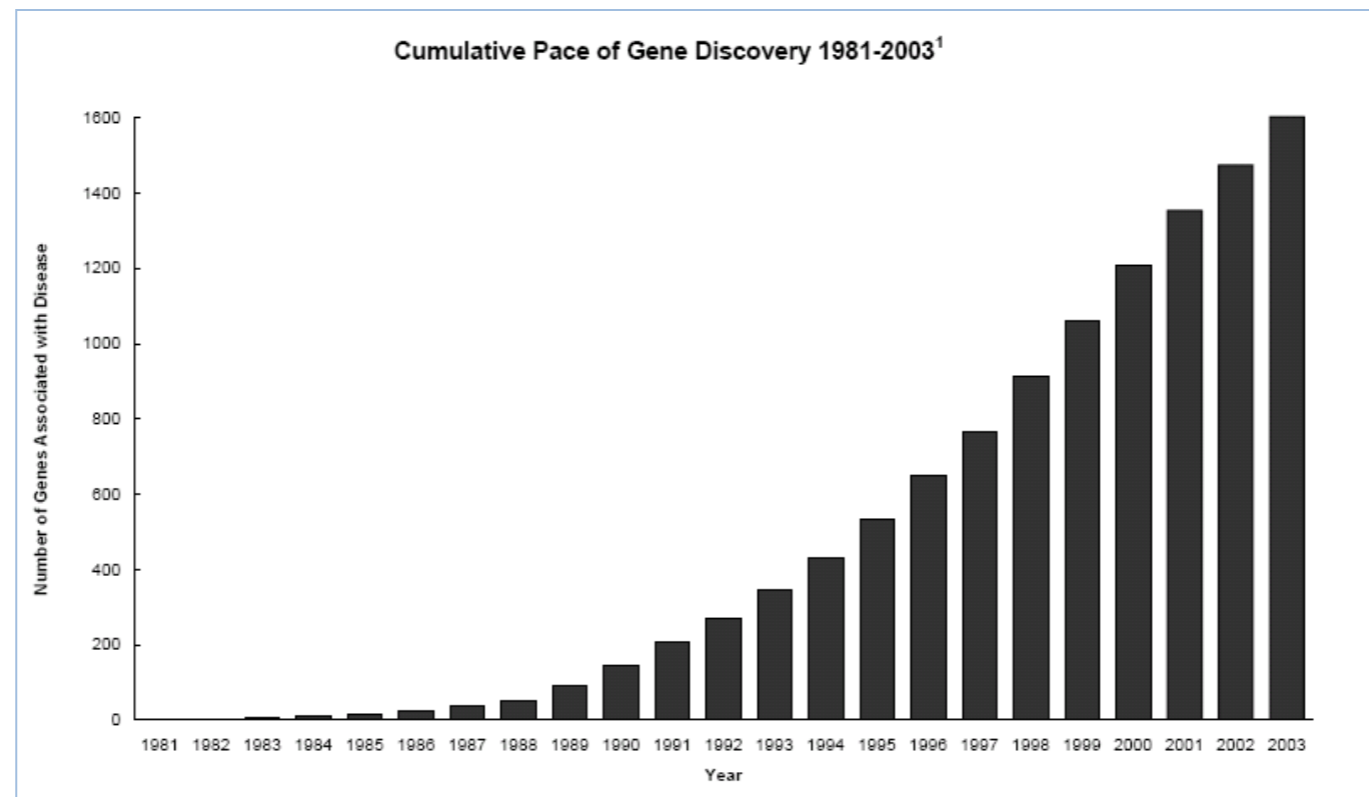
If there are 10^8 sperm per ejaculate, in principle every base could be mutated in at least one sperm cell and each germ cell has around 10 mutations

Human Mendelian Phenotypes

OMIM Entry Statistics:

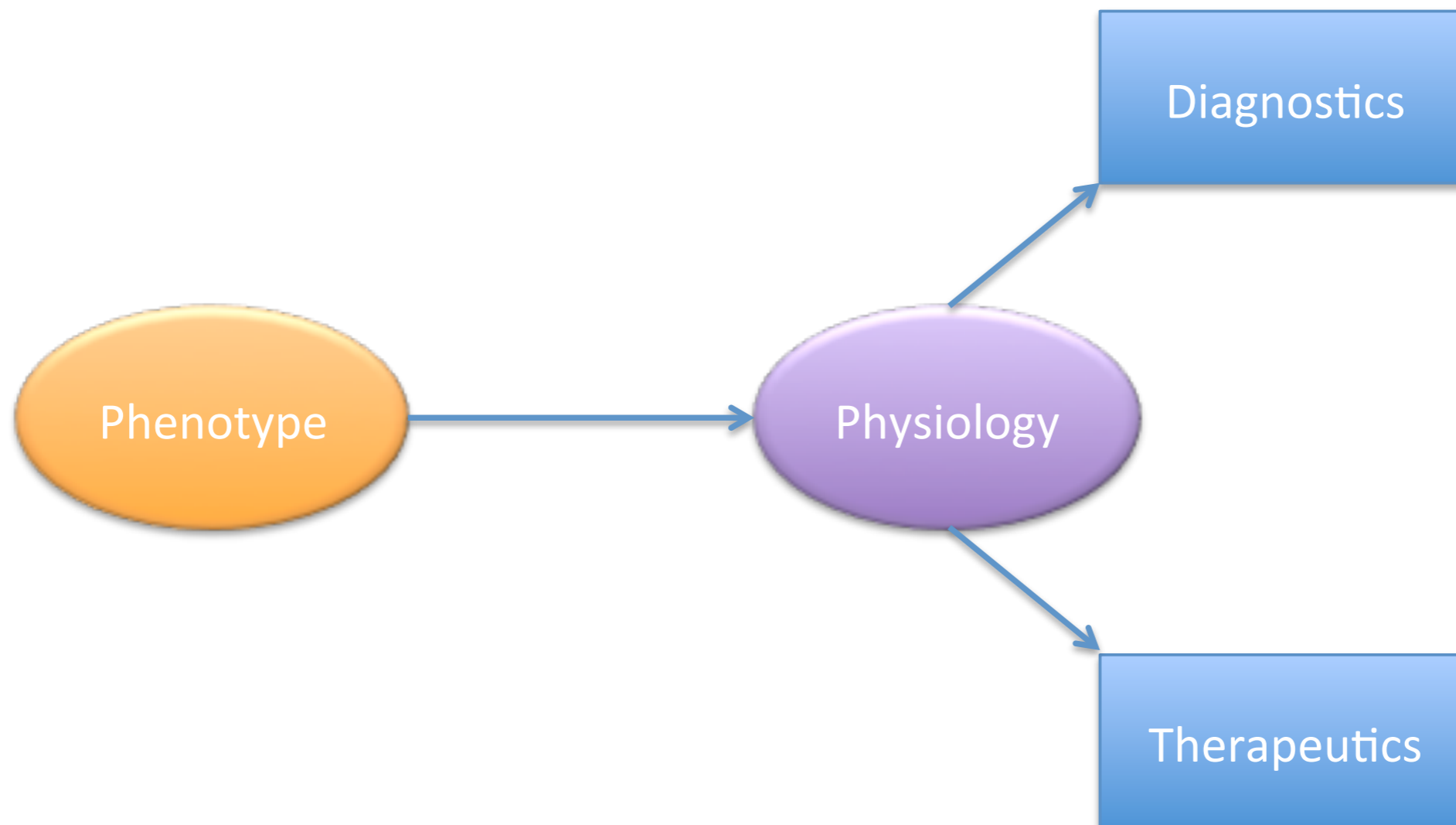
Number of Entries in OMIM (1 January 2012) :

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

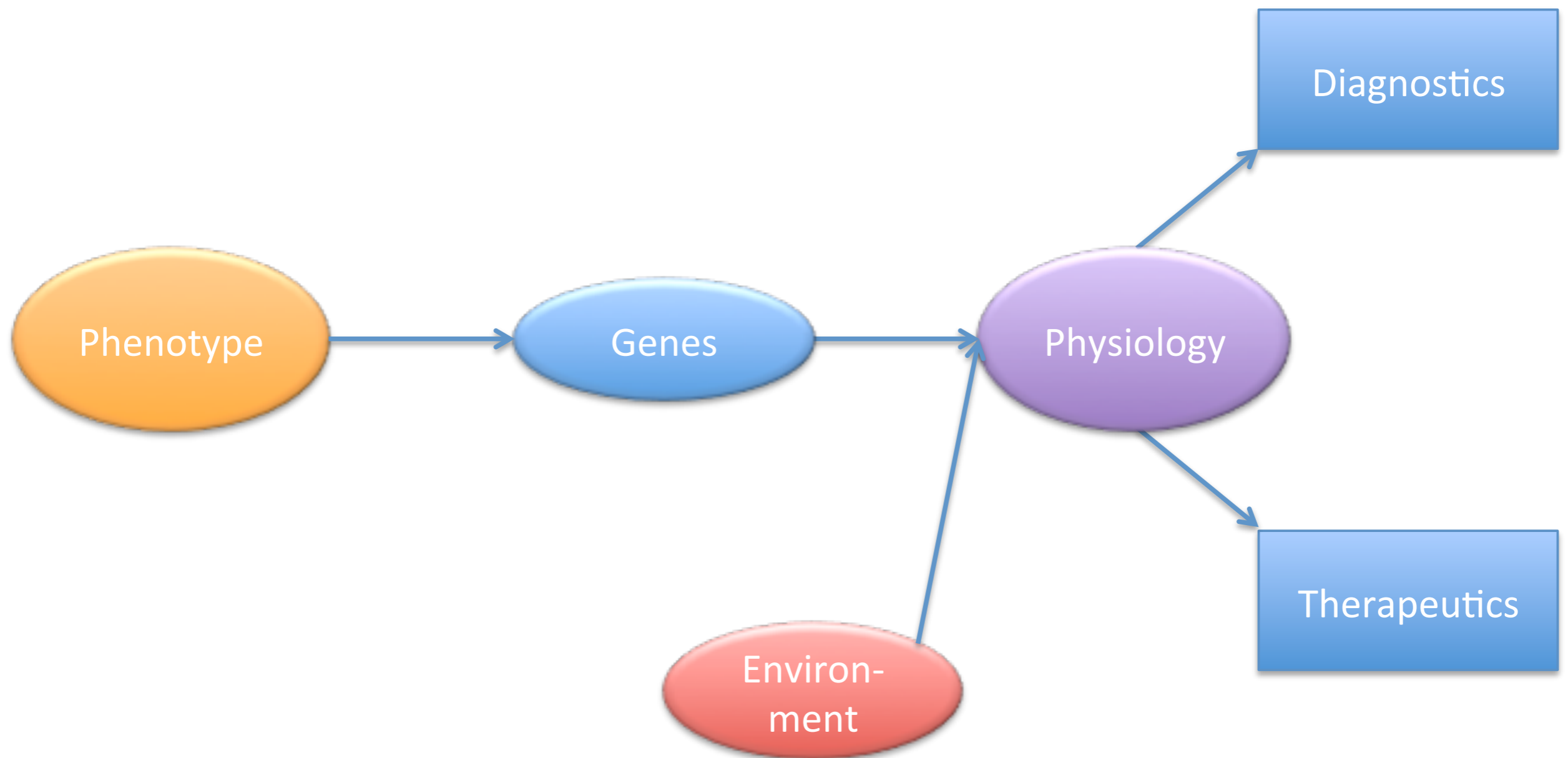


<http://www.genome.gov/Pages/News/PaceofDiseaseGeneDiscovery.pdf>

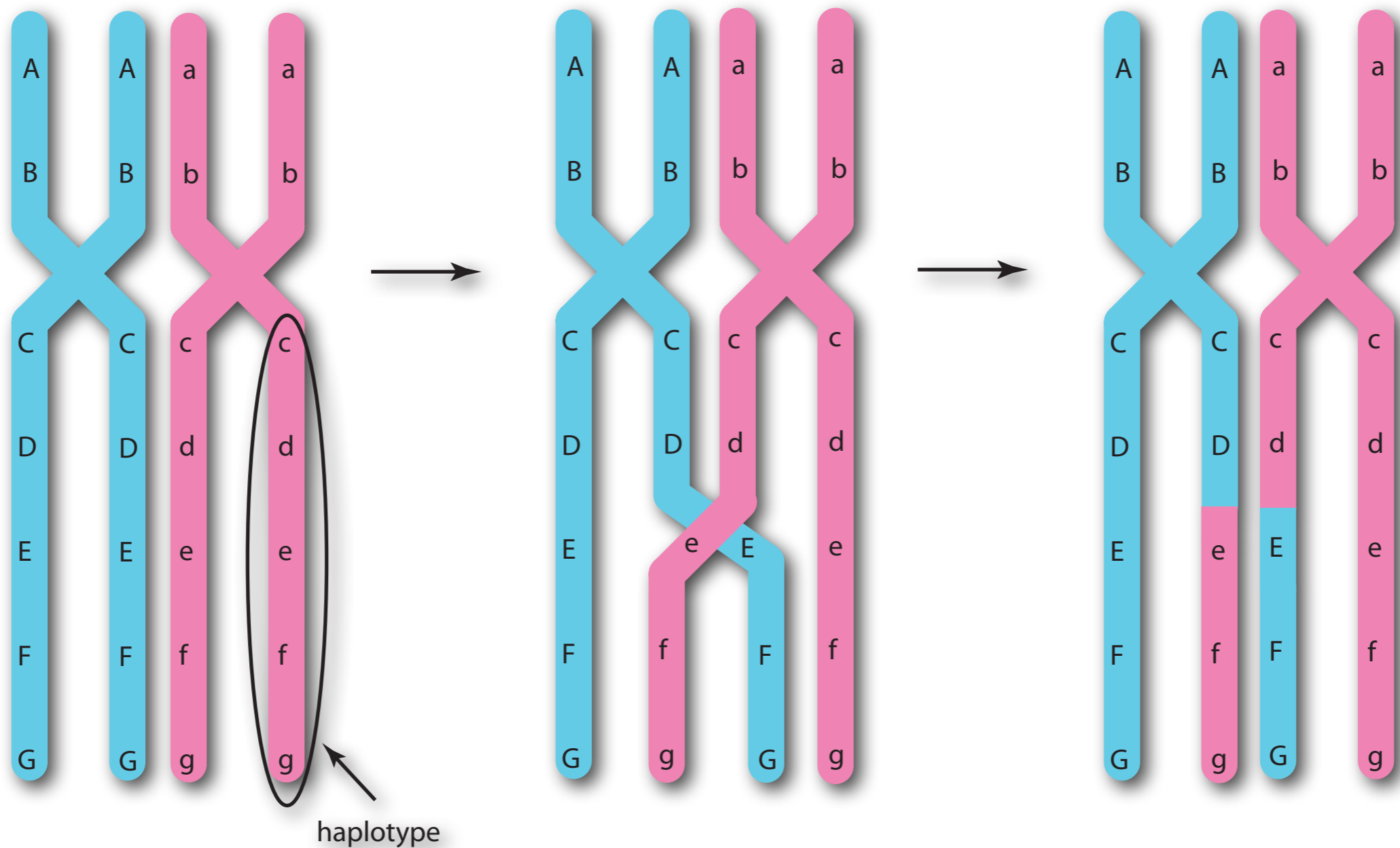
Approach to Genetic Disorders



Approach to Genetic Disorders



Genetic Linkage

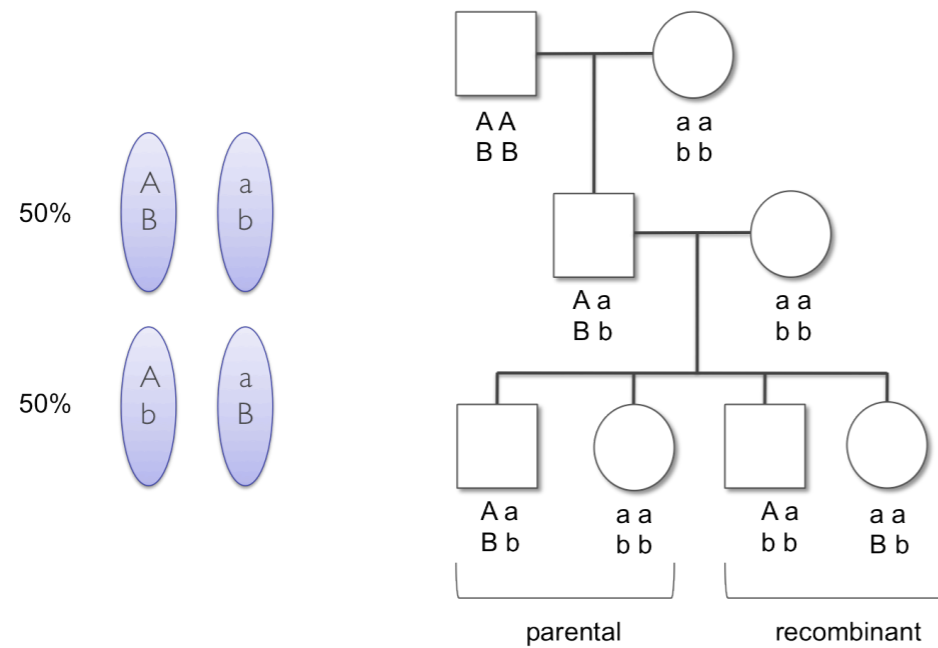


Polymorphism

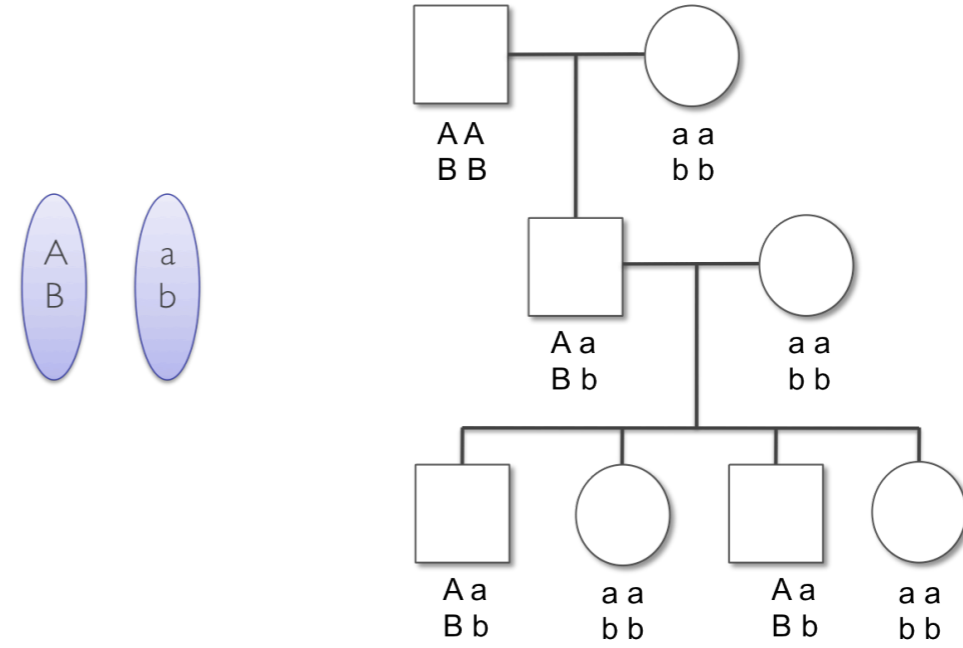
Polymorphism: occurrence of at least two alleles at a locus having a frequency of at least 1%

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

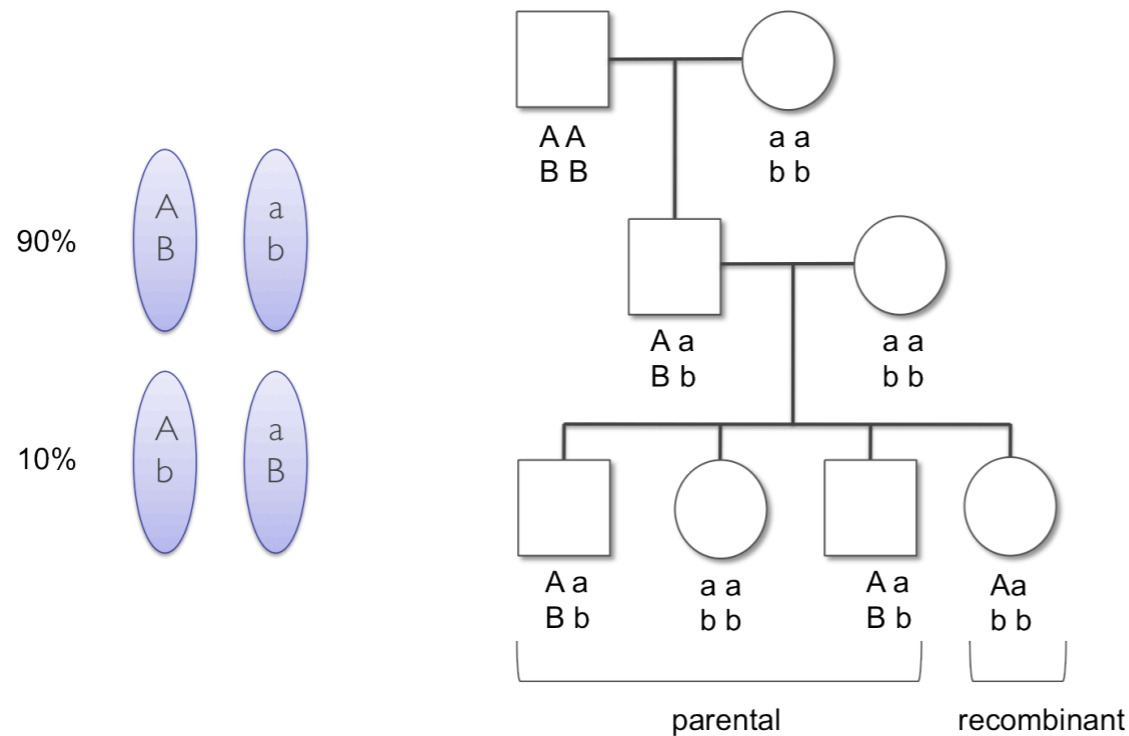
Linkage



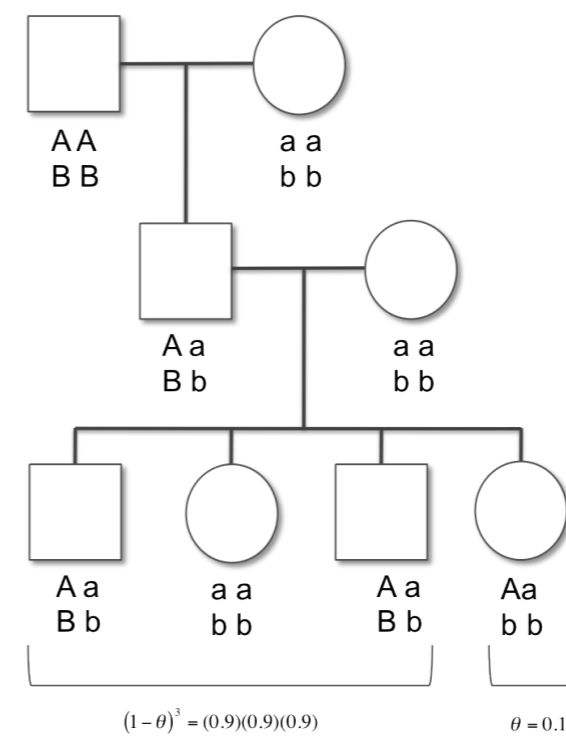
Independent Assortment



Complete Linkage



10% Recombination

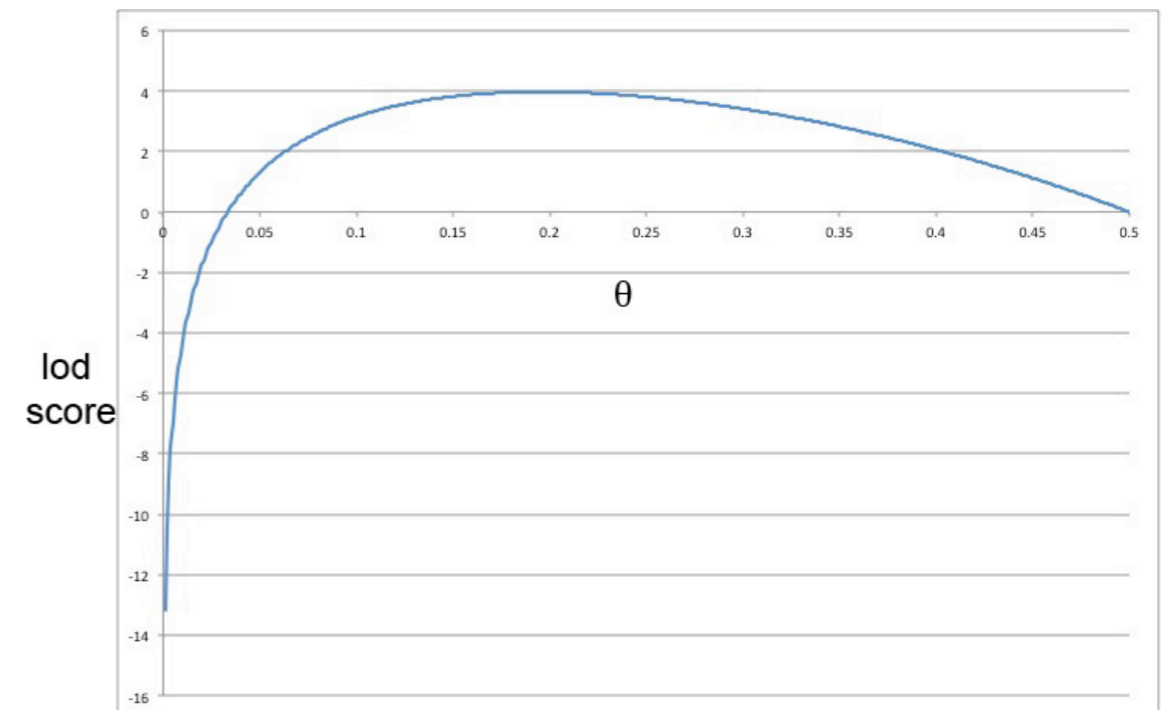
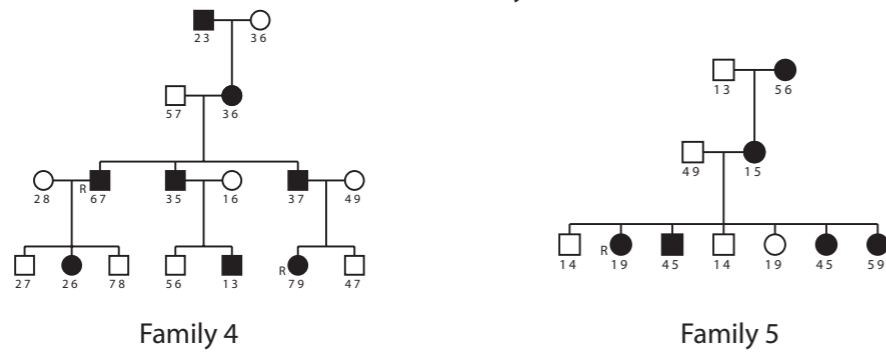
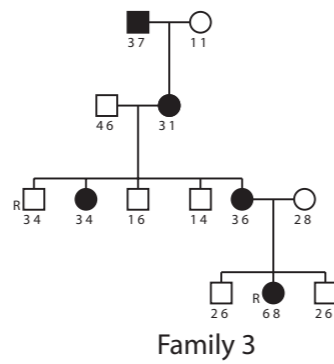
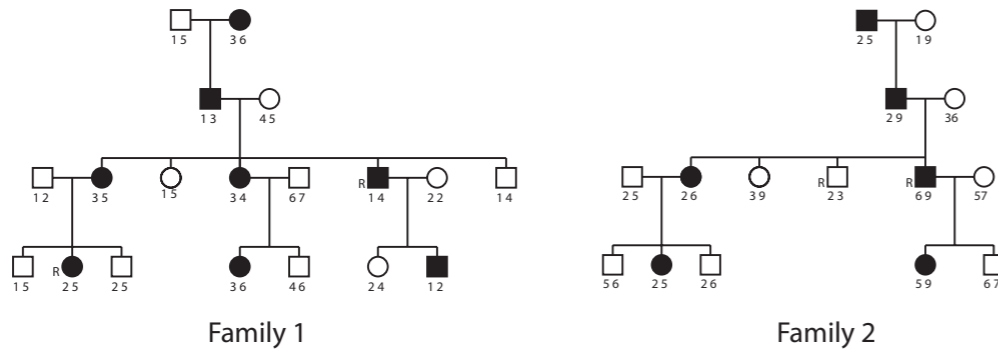


Likelihood Ratio

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

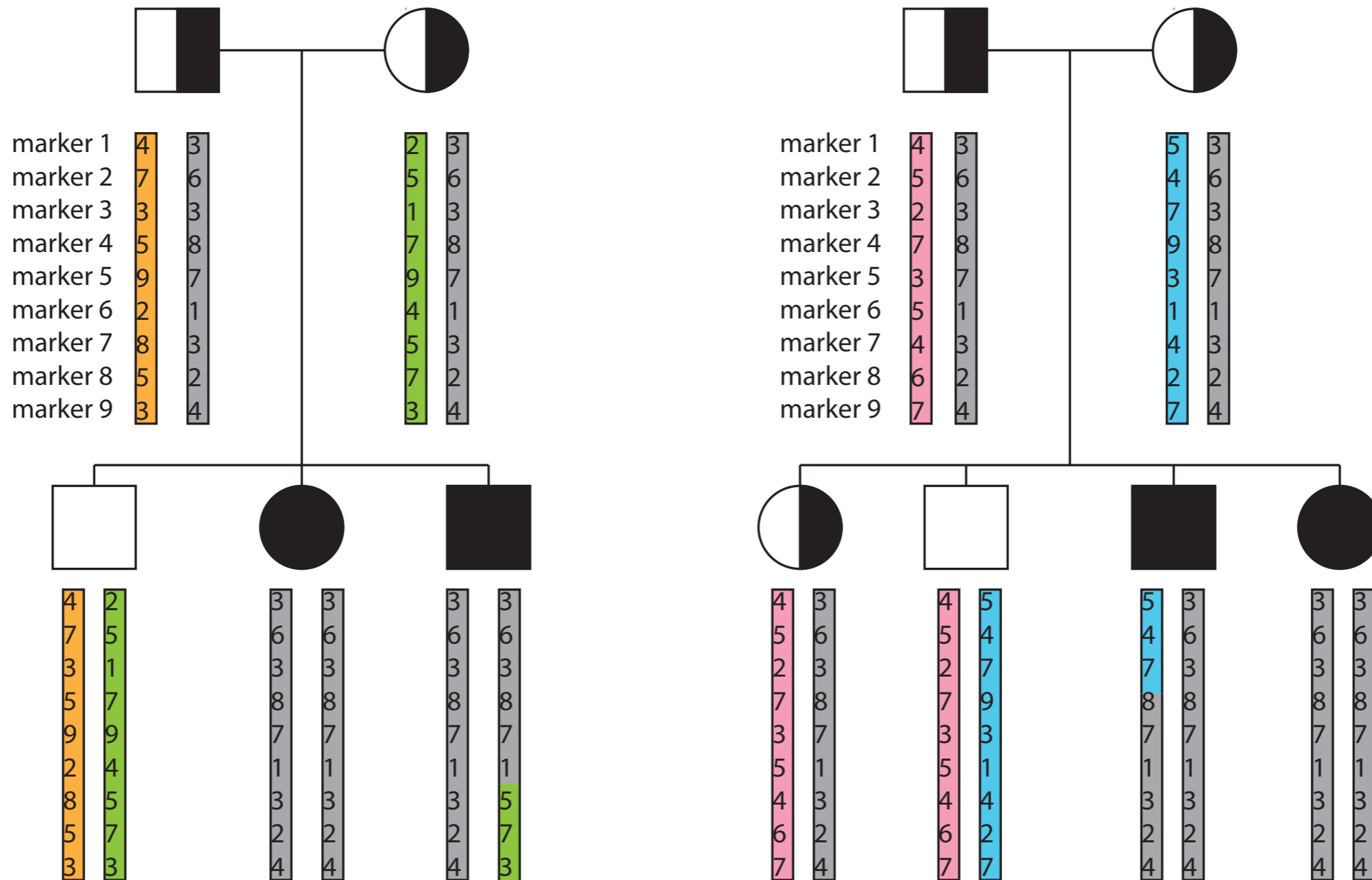
n = number non-recombinants
r = number recombinants

LOD Analysis

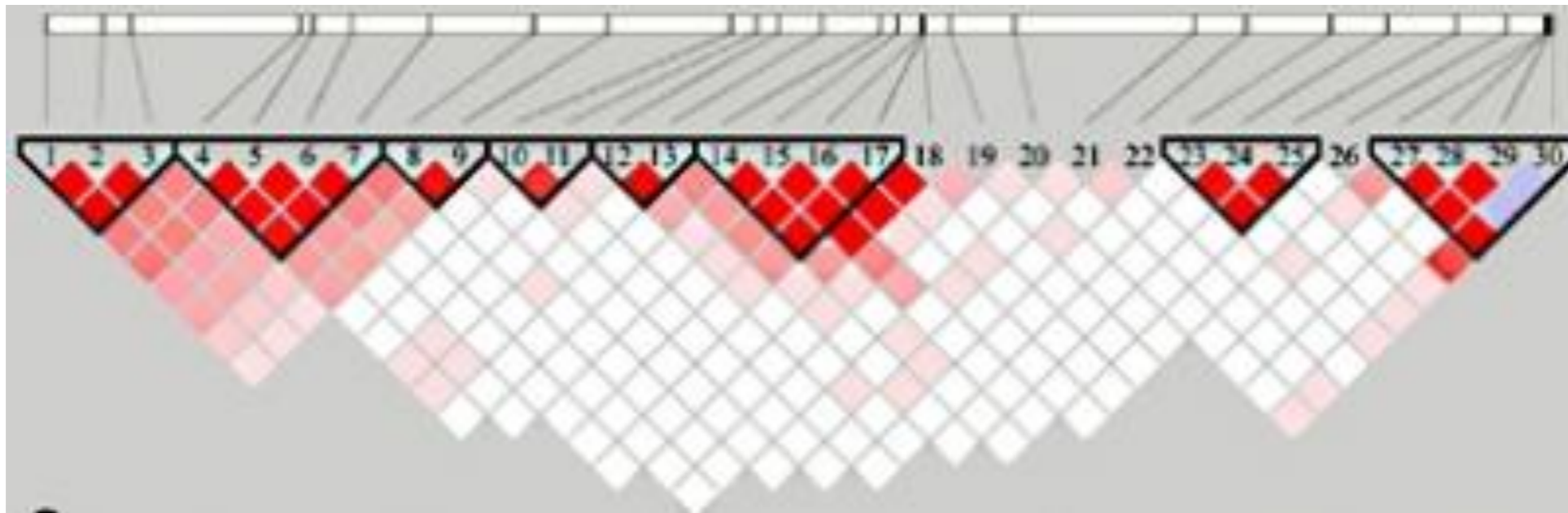


Family	Sibs	Recombinants	Nonrecombinants	θ				
				0	0.1	0.2	0.3	0.4
1	12	2	10	$-\infty$	1.15	1.25	1.02	0.60
2	9	2	7	$-\infty$	0.39	0.96	0.58	0.36
3	8	2	6	$-\infty$	0.13	0.43	0.43	0.28
4	10	2	8	$-\infty$	0.64	0.84	0.73	0.44
5	7	1	6	$-\infty$	0.83	0.83	0.65	0.38
Total	46	7	39	$-\infty$	3.14	4.31	3.41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

Haplotype Analysis

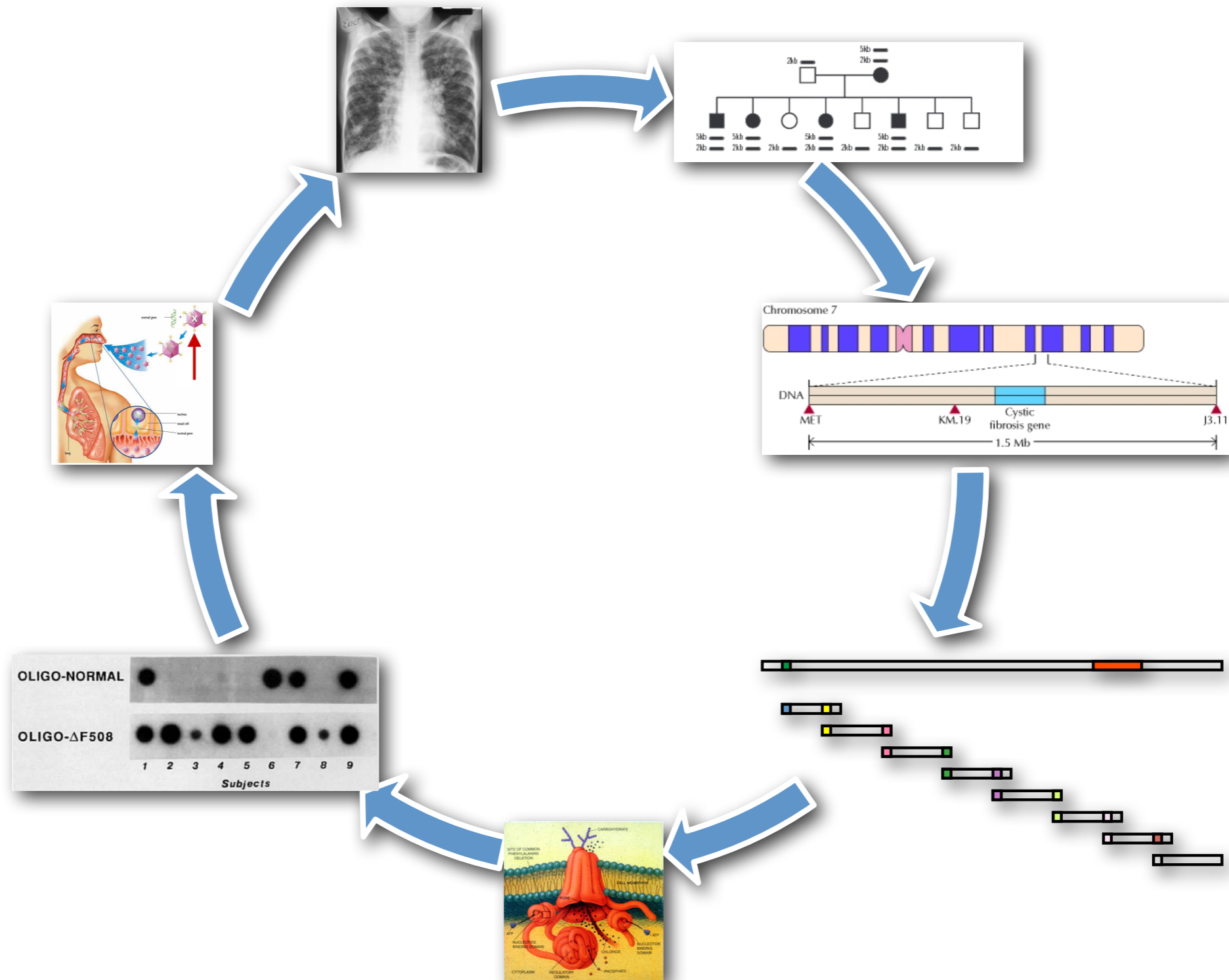


Linkage Disequilibrium



<http://estrip.org/articles/read/tinypliny/44920>

Positional Cloning



Genome Browser

Human - UCSC Genome Browser v230

http://genome.ucsc.edu/cgi-bin/hgTracks?org=human

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

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

chr21 (q22.11) 21p13 21p12 21p11.2 21q21.1 21q21.2 21q21.3 21q22.1 21q22.2 21q22.3

Scale 2 kb

chr21: 33033000 | 33034000 | 33035000 | 33036000 | 33037000 | 33038000 | 33039000 | 33040000 | 33041000

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

RefSeq Genes

Human mRNAs from GenBank

Human ESTs That Have Been Spliced

Spliced ESTs

Vertebrate Multiz Alignment & Conservation (46 Species)

Placental Mammal Basewise Conservation by PhyloP

Mammal Cons

Multiz Alignments of 46 Vertebrates

Rhesus Mouse Dog Elephant Opossum Chicken X_tropicalis Zebrafish

Simple Nucleotide Polymorphisms (dbSNP build 130 - Provisional Mapping to GRCh37)

SNPs (130)

Repeating Elements by RepeatMasker

move start Click on a feature for details. Click or drag in the base position track to zoom in. move end

Click gray/blue bars on left for track options and descriptions.

default tracks hide all add custom tracks configure reverse refresh

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.

collapse all expand all

Mapping and Sequencing Tracks refresh

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

Phenotype and Disease Associations refresh

Genes and Gene Prediction Tracks refresh

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

mRNA and EST Tracks refresh

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

Expression refresh

Allen Brain	GNF Atlas 2
hide	hide

Regulation refresh

CpG Islands

hide

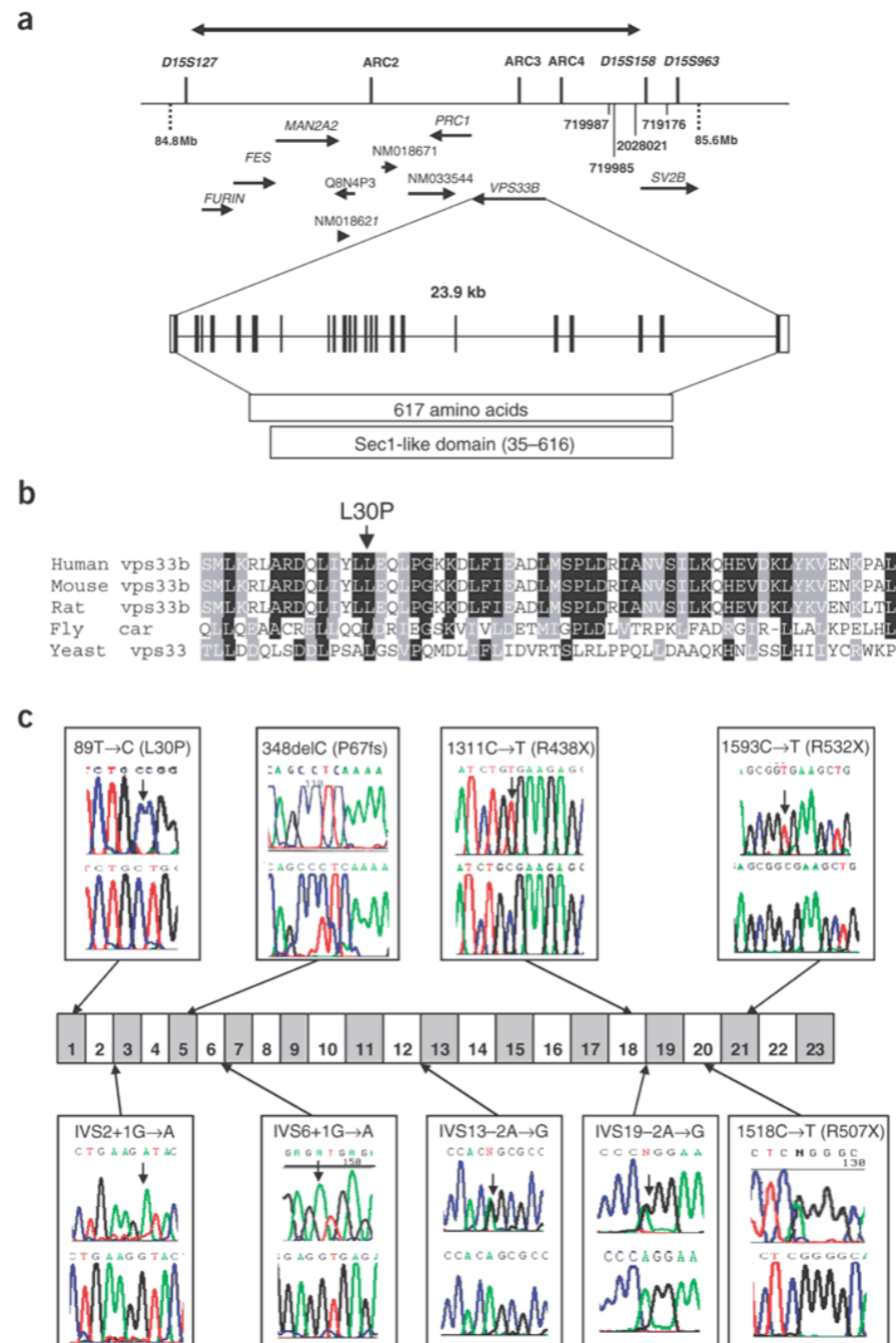
Comparative Genomics refresh

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

Variation and Repeats refresh

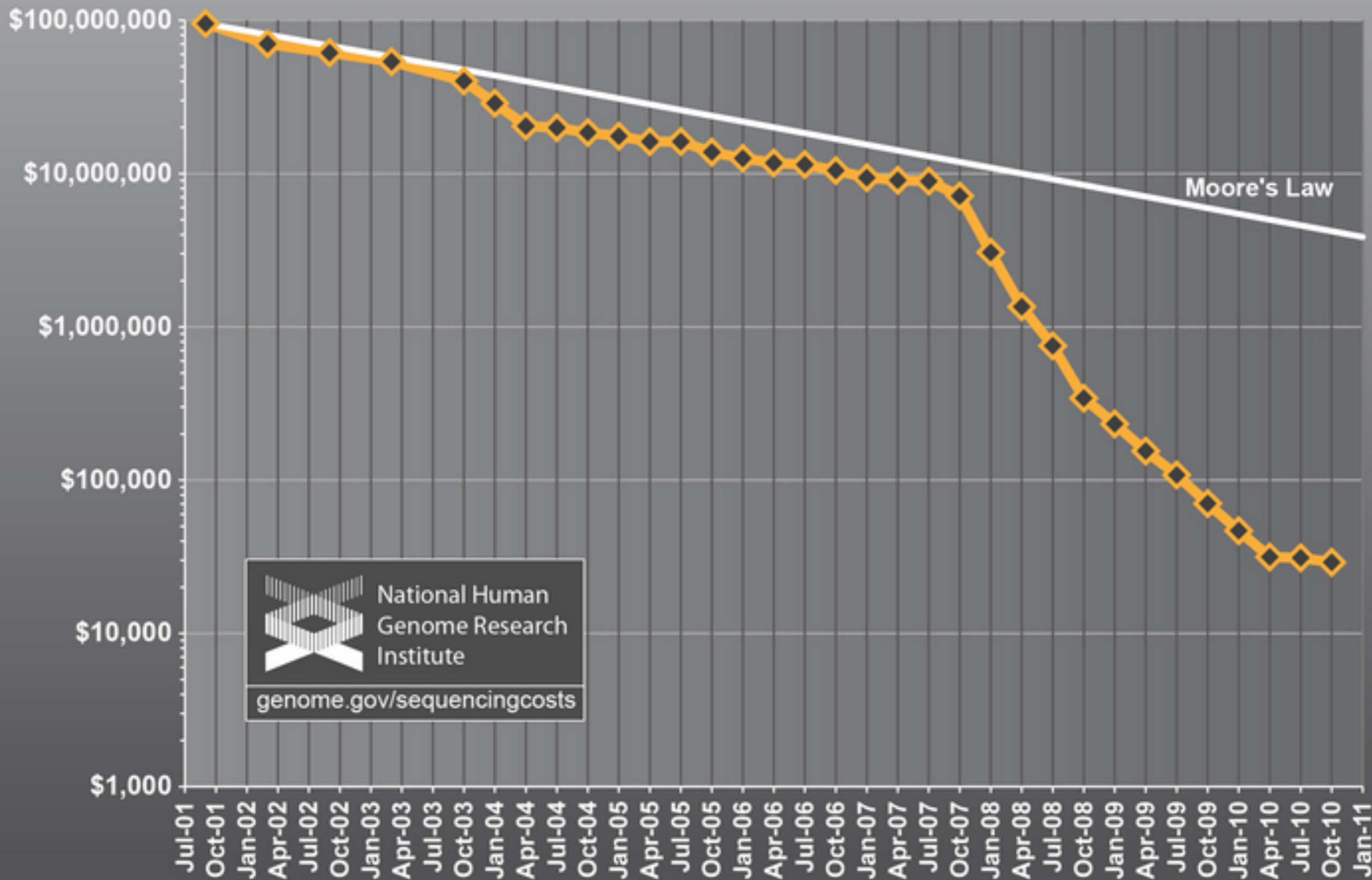
SNPs (130)	Segmental Dups	RepeatMasker	Interrupted Rpts	Simple Repeats	Self Chain
dense	hide	dense	hide	hide	hide

Candidate Genes



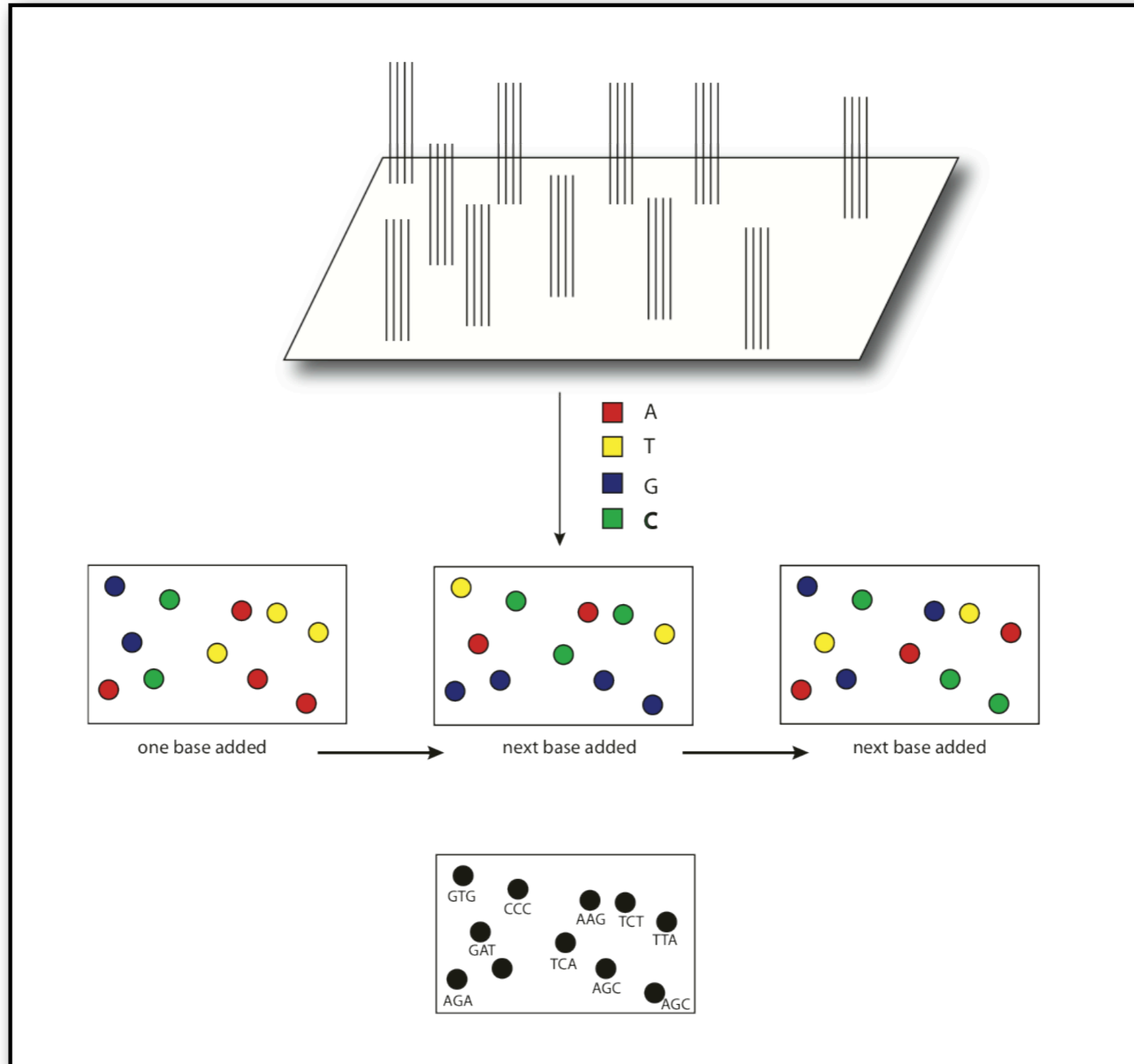
Nature Genetics **36**, 400 - 404 (2004)

Cost per Genome

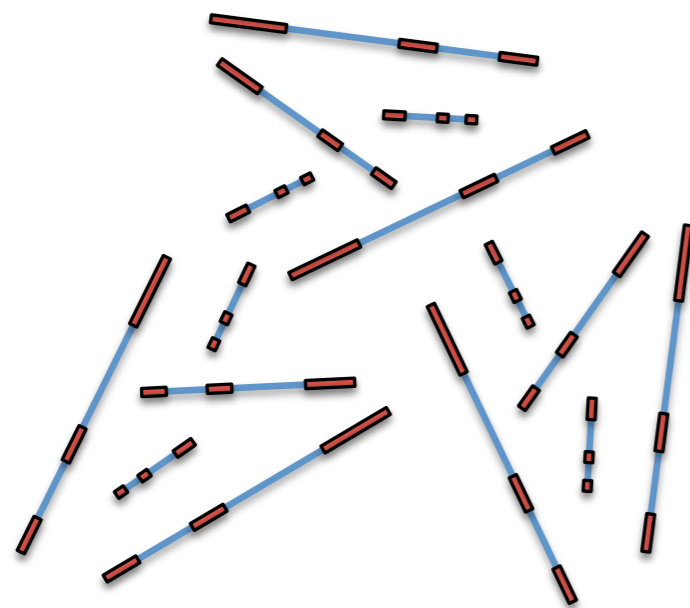
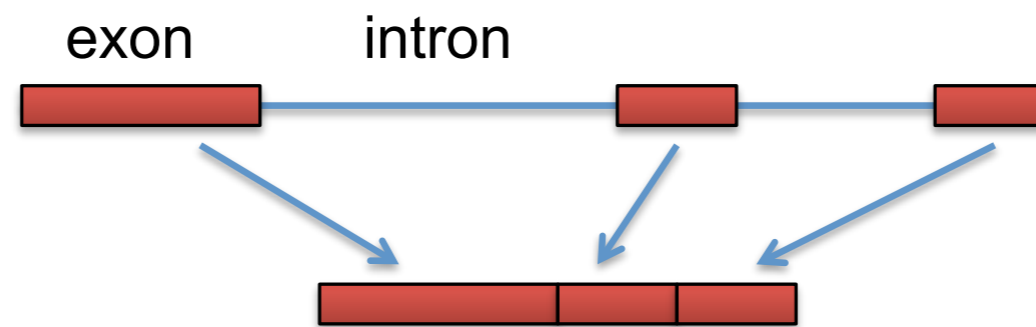


 National Human
Genome Research
Institute
genome.gov/sequencingcosts

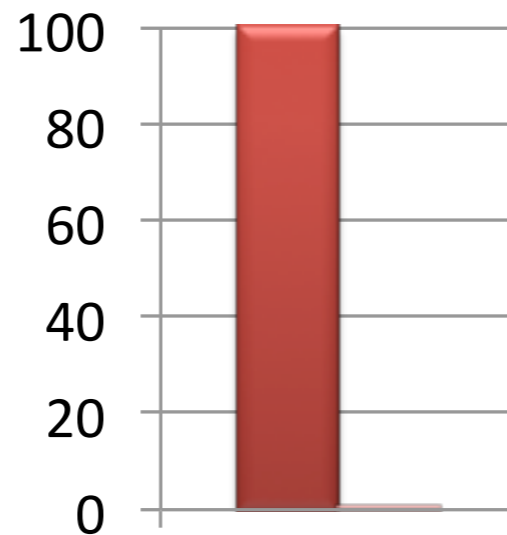
Massively Parallel Sequencing



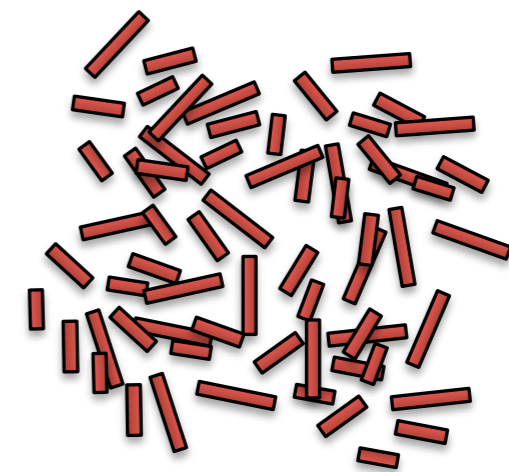
Exome vs. Genome Sequencing



Genome



■ Genome
■ Exome



Exome

Gene Discovery

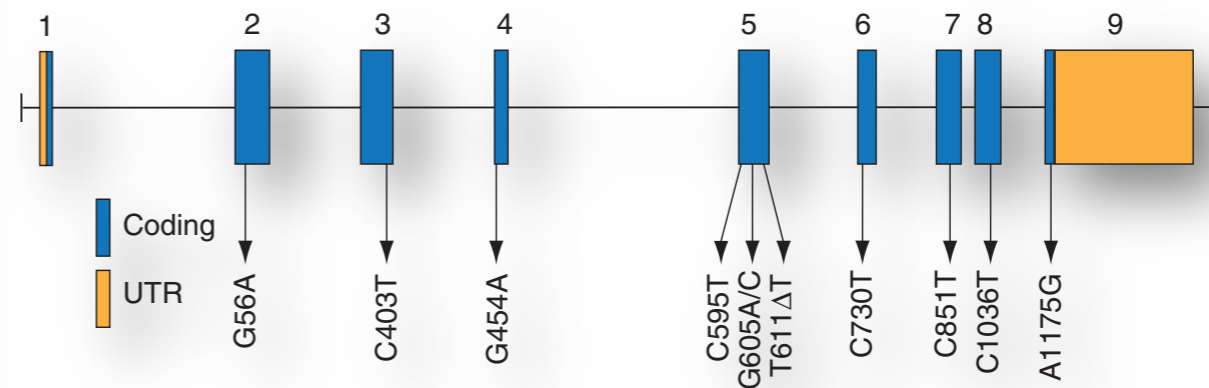
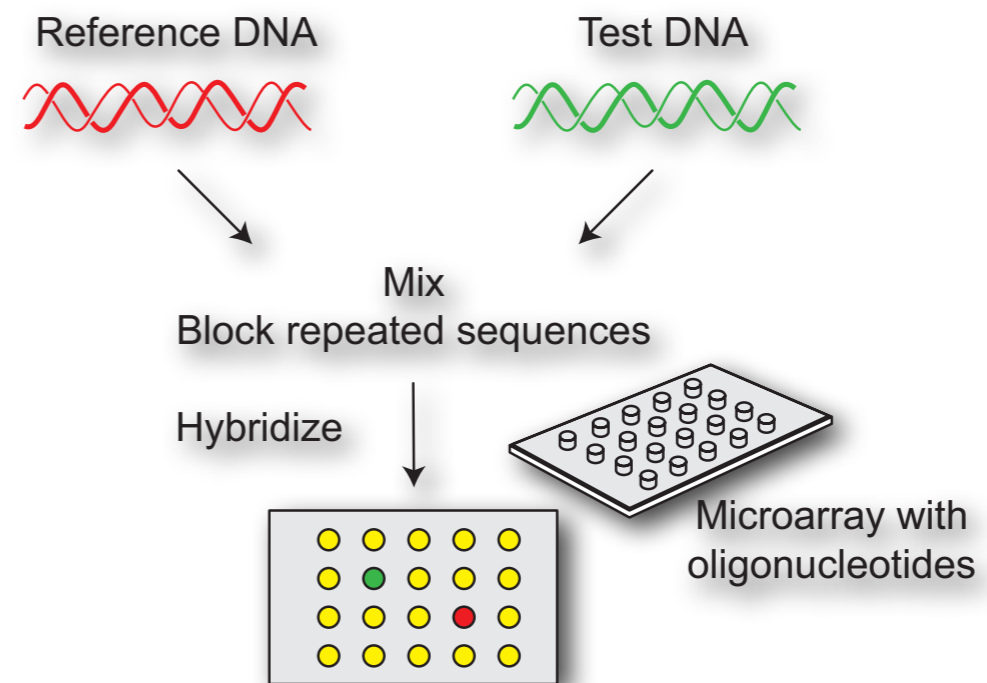
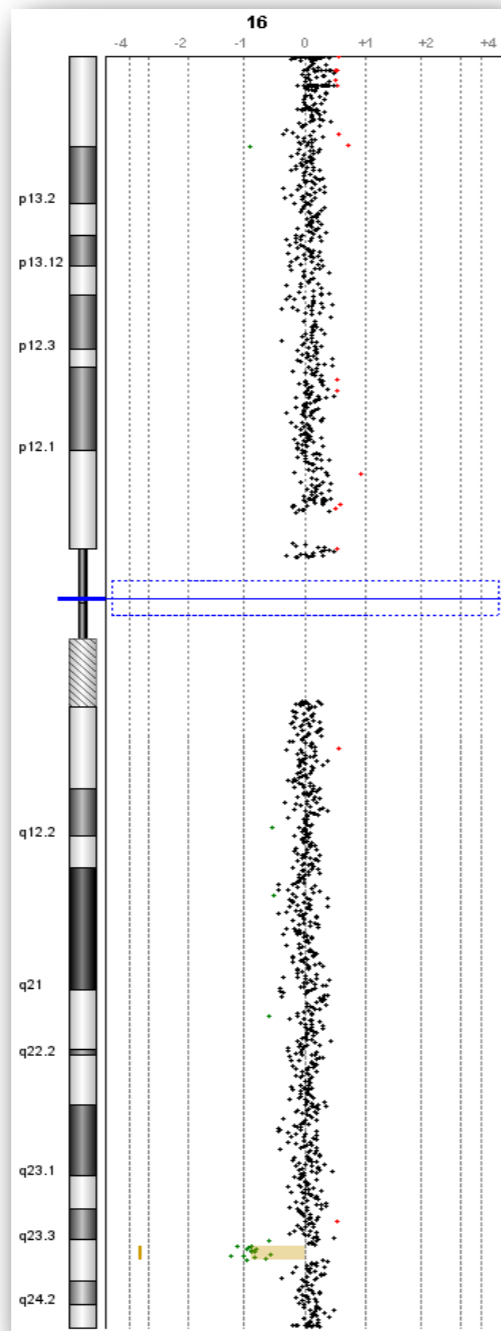


Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

Filter	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)		Kindreds 1+2		Kindreds 1+2+3	
	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362	3,099	1,810	2,654	1,525
Not in dbSNP129	641	102	647	114	369	53	105	25	63	21
Not in HapMap 8	898	123	923	128	506	46	117	7	38	4
Not in either	456	31	464	33	228	9	26	1*	8	1*
Predicted damaging	204	6	204	12	83	1	5	0	2	0

Ng, S., et al. Nature Genetics 2010;42:30

Cytogenomics



Mendelian Disorders Sequencing Centers

Mendelian Disorders Sequencing Centers

- [Program Rationale](#)
- [Grantees of the Program](#)
- [Program Contacts](#)

Program Rationale

Discovering the genes and genetic variants underlying human Mendelian disorders is of significant biomedical relevance. The knowledge of those variants, which are rare and highly penetrant, will facilitate rapid and accurate diagnosis of Mendelian disorders and might lead to new therapeutic approaches. This knowledge can also lead to insight about the common or more complex phenotypes that involve similar genes, pathways, and phenotypes. In the long run, a comprehensive collection of rare and highly penetrant variants would be a highly valuable resource for understanding basic human genetics and would identify entry points into fundamental developmental and physiological pathways.

While the genetic basis of more than 3000 Mendelian disorders has been determined so far, the genetic basis remains to be determined for a larger number of confirmed or suspected Mendelian disorders. Recent advances in genome technology and computational methods have made it possible to identify the genetic basis of Mendelian disorders using genome-wide approaches in a more rapid and cost-effective way than linkage mapping and candidate gene approaches.

The Mendelian Disorders Genome Centers Program aims to contribute to the discovery of the genetic basis of most Mendelian disorders in two main ways. The first is to use genome-wide sequencing and other genomic approaches to discover the genetic basis underlying as many disorders and health-related traits as possible, spanning the various Mendelian inheritance patterns, during the funding period. The second is to build a better foundation for elucidating the genetic basis of Mendelian disorders by 1) establishing and disseminating information about effective approaches to the identification of the causative genetic variants, and gaining insight about the overall tractability of Mendelian disorders to state-of-the-art genomic approaches, and 2) compiling a comprehensive list of available human samples of Mendelian disorders and other health-related Mendelian traits as a public resource to help coordinate genetic variant discovery activities that will be carried out by many groups.

[Top of page](#)

Grantees of the Program

The currently funded centers are:

- University of Washington Center for Mendelian Genomics
- Yale Center for Mendelian Disorders
- Baylor-Johns Hopkins Center for Mendelian Genetics

In addition to these centers, the Genome Sequencing and Analysis Centers also carry out efforts to discover the genetic basis of Mendelian disorders (see above).

[Top of page](#)

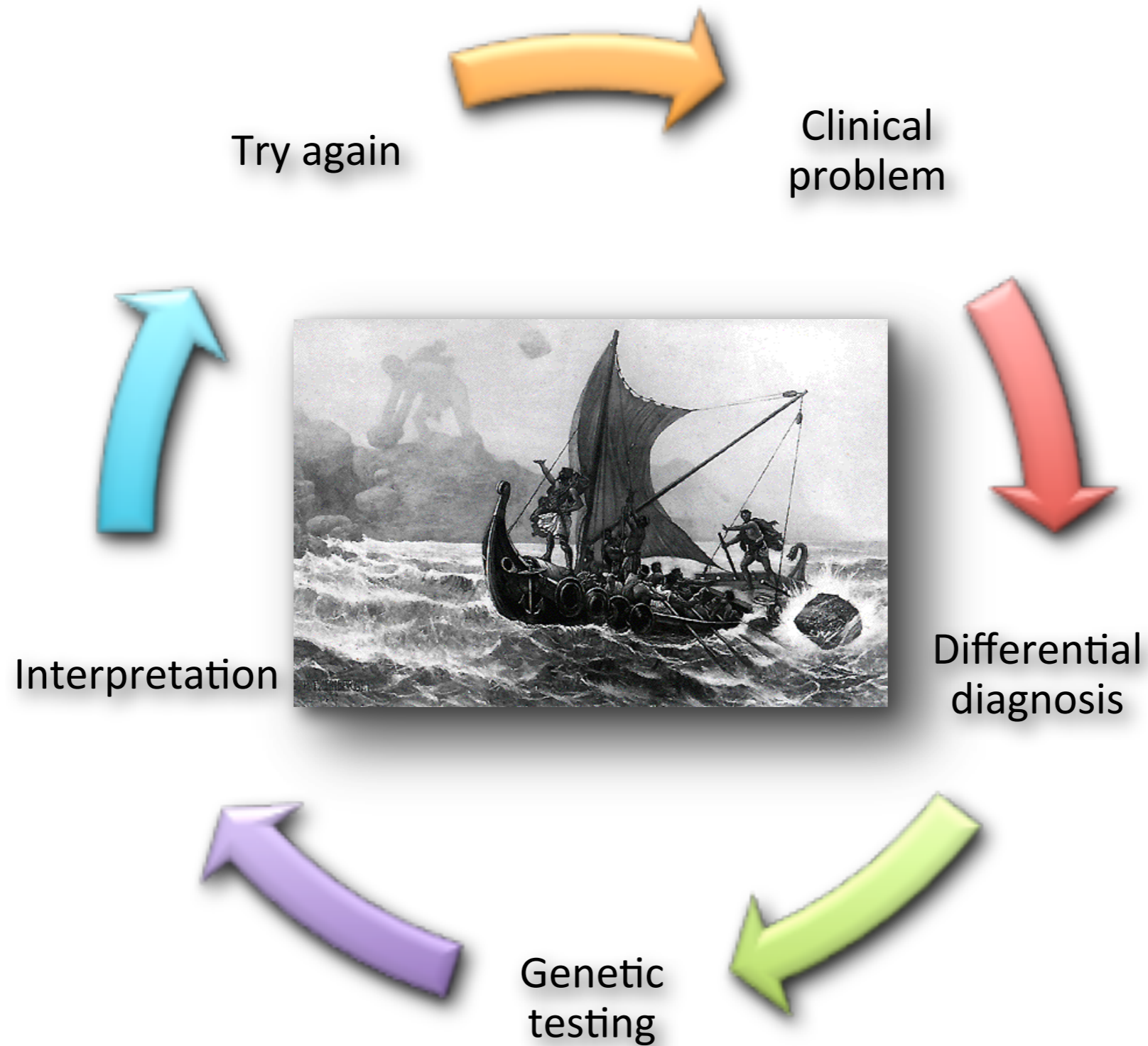
Program Contacts

For general inquiries about the program, please contact:

Lu Wang, Ph.D.
Program Director
E-mail: wanglu@mail.nih.gov

If you wish to provide samples with confirmed or suspected Mendelian disorders or traits for the Mendelian Disorders Genome Centers to study, please contact the Coordination Site of the Program at gmendel@uw.edu. The Program will decide on the feasibility and priority of sequencing these samples.

Diagnostic Odyssey



Genomic Diagnosis

NT5E Mutations and Arterial Calcifications

Cynthia St. Hilaire, Ph.D., Shira G. Ziegler, B.A., Thomas C. Markello, M.D., Ph.D.,
Alfredo Brusco, Ph.D., Catherine Groden, M.S., Fred Gill, M.D.,
Hannah Carlson-Donohoe, B.A., Robert J. Lederman, M.D.,
Marcus Y. Chen, M.D., Dan Yang, M.D., Ph.D., Michael P. Siegenthaler, M.D.,
Carlo Arduino, M.D., Cecilia Mancini, M.Sc., Bernard Freudenthal, M.D.,
Horia C. Stanescu, M.D., Anselm A. Zdebik, M.D., Ph.D.,
R. Krishna Chaganti, M.D., Robert L. Nussbaum, M.D., Robert Kleta, M.D., Ph.D.,
William A. Gahl, M.D., Ph.D., and Manfred Boehm, M.D.

N ENGL J MED 364:5 NEJM.ORG FEBRUARY 3, 2011

Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease

Elizabeth A. Worthey, PhD^{1,2}, Alan N. Mayer, MD, PhD^{2,3}, Grant D. Syverson, MD²,
Daniel Helbling, BSc¹, Benedetta B. Bonacci, MSc², Brennan Decker, BSc¹, Jaime M. Serpe, BSc²,
Trivikram Dasu, PhD², Michael R. Tschannen, BSc¹, Regan L. Veith, MSc², Monica J. Basehore, PhD⁴,
Ulrich Broeckel, MD, PhD^{1,2,3}, Aoy Tomita-Mitchell, PhD^{1,2,3}, Marjorie J. Arca, MD^{3,5},
James T. Casper, MD^{2,3}, David A. Margolis, MD^{2,3}, David P. Bick, MD^{1,2,3}, Martin J. Hessner, PhD^{1,2},
John M. Routes, MD^{2,3}, James W. Verbsky, MD, PhD^{2,3}, Howard J. Jacob, PhD^{1,2,3,6},
and David P. Dimmock, MD^{1,2,3}

Genetics in Medicine • Volume 13, Number 3, March 2011

