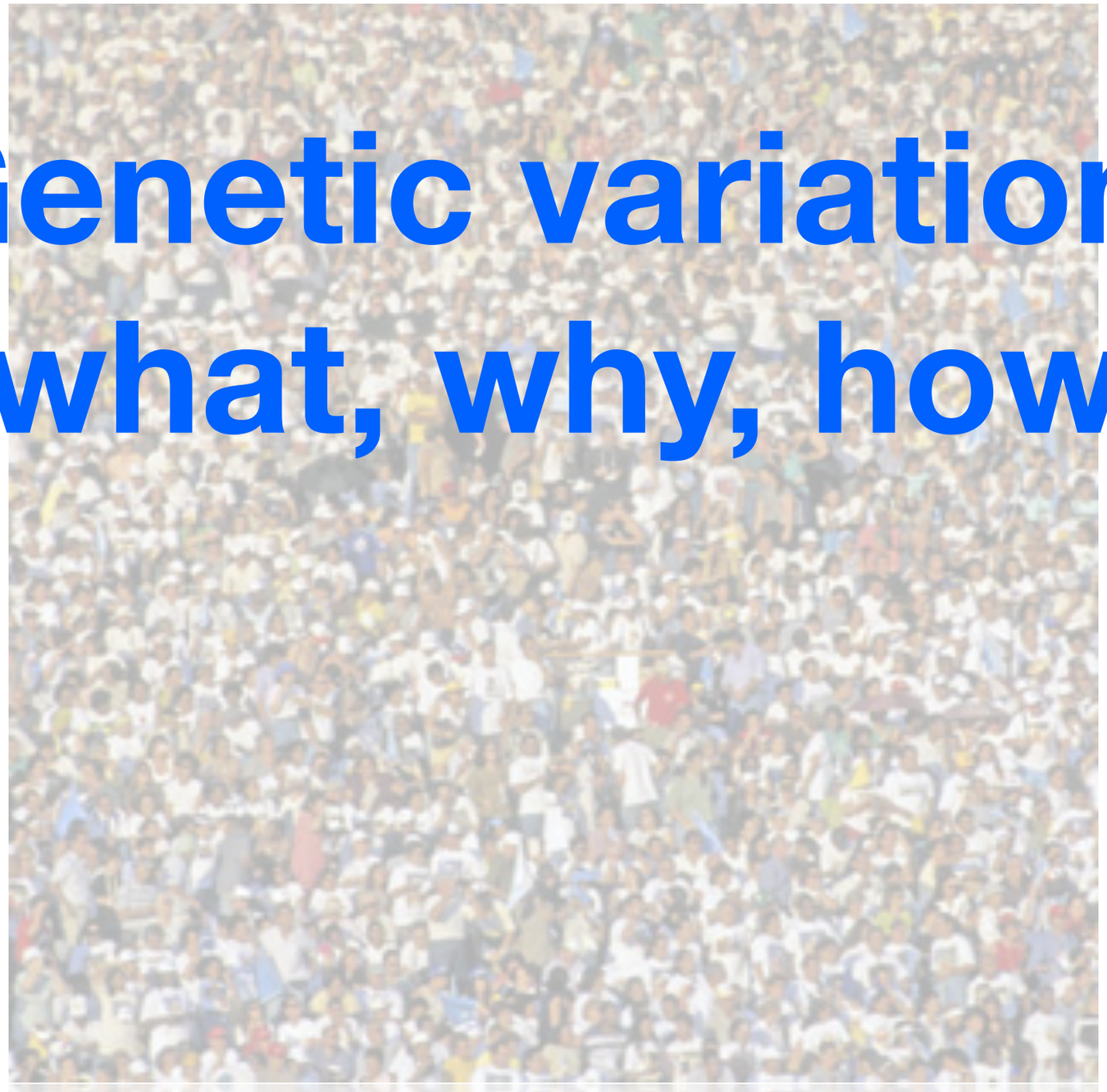
Genetic variation: what, why, how



Aaron Quinlan
BIMS 6000
04-Sept-2013
quinlanlab.org | arq5x@virginia.edu

Goals

- Understand the origins of genetic mutation
- Be familiar with how mutations become *polymorphic*
- Understand the processes of genetic drift and selection.
- Recognize the different types of genetic variation and their spectrum of functional consequence.

What is genetic variation?

- Differences in DNA content or structure among individuals
- Any two individuals have ~99.5% identical DNA.
- But the human genome is big - each haploid set of 23 chromosomes has 3 billion nucleotides.
- The details matter.

~98-99% identical DNA

~99.5% identical DNA

CGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTC
GTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAGCCAGACTTCCCGTGTCCTTTCC
ACCGGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGG
GTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTG
TTGTCTGCATGTAACTTAATACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCT
AGGCCAGTAAGTAGTGCTTGTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCT
GCTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGG
AGAGCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTGGCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCT
CTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGCCGCAGGCACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGAT
GTGTGGGCCCACCGGCCCCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCAGGTCTCCAGAG
CTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGA
GTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCC
TGTGATGTCATCTGGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAAGGCCTGGCAGAGTCTTTCCCAGGGA
AAGCTACAAGCAGCAAACAGTCTGCATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCAAGAAAGGCTCTGG
TGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAGG
GGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGAGGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTA
ACTCAAAGCCTTAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACCAGTGAT
ACACCCGGCACCCTGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGC
TGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCCAGT

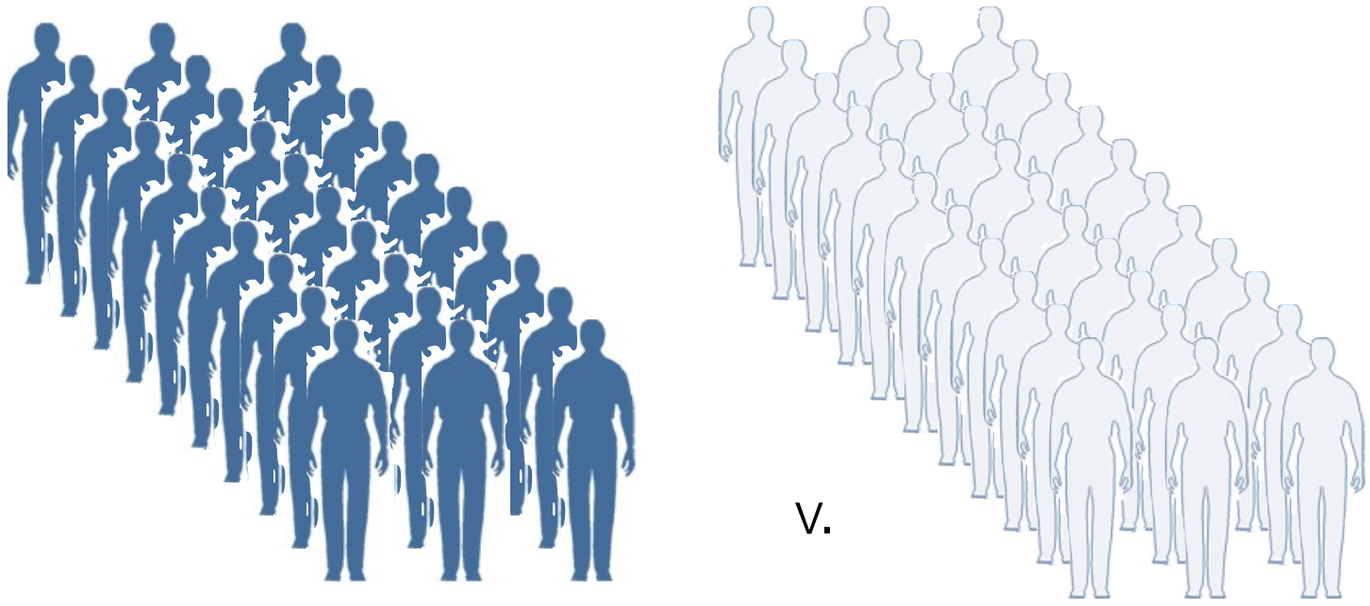
TTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTG
TTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCA
CAGGCATTAGTGCCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGGGTCCTGGCCATCCGTGAGATCTTCCCAGGTGTGCCGT
TTTCTCTGGAAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCCATGGAGCACAGGCAGACAGAAGTCCCCGCCC
CAGCTGTGTGGCCTCAAGCCAGCCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGCTGGTTCCGTCACCCCCTCCCAAGGAAGTAGG
TCTGAGCAGCTTGTCCTGGCTGTGTCCATGTCAGAGCAACGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAGCCCCCTACGA
TTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCCT
CCGGGCCCCTCACCAGCCCCAGGTCCTTTCCCAGAGATGCCTGGAGGGAAAAGGCTGAGTGAGGGTGGTTGGTGGGAAACCCTGGTTC
CCCCAGCCCCCGGAGACTTAAATACAGGAAGAAAAAGGCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCTGCCTC
CTACCCTTGCGCCTCATGACCGGAGCCATAGCCCAGGCAGGAGGGCTGAGGACCTCTGGTGGCGGCCCAGGGCTTCCAGCATGTGCCC
TAGGGGAAGCAGGGGCCAGCTGGCAAGAGCAGGGGGTGGGCAGAAAGCACCCGGTGGACTCAGGGCTGGAGGGGAGGAGGCGATCTTG
CCCAAGGCCCTCCGACTGCAAGCTCCAGGGCCCGCTCACCTTGCTCCTGCTCCTTCTGCTGCTGCTTCTCCAGCTTTCGCTCCTTCAT
GCTGCGCAGCTTGGCCTTGCCGATGCCCCCAGCTTGGCGGATGGACTCTAGCAGAGTGGCCAGCCACCGGAGGGGTCAACCACTTCCC
TGGGAGCTCCCTGGACTGGAGCCGGGAGGTGGGGAACAGGGCAAGGAGGAAAGGCTGCTCAGGCAGGGCTGGGGAAGCTTACTGTGTC
CAAGAGCCTGCTGGGAGGGAAGTCACCTCCCCTCAAACGAGGAGCCCTGCGCTGGGGAGGCCGGACCTTTGGAGACTGTGTGTGGGGG
CCTGGGCACTGACTTCTGCAACCACCTGAGCGCGGGCATCCTGTGTGCAGATACTCCCTGCTTCCTCTCTAGCCCCCACCCTGCAGAG
CTGGACCCCTGAGCTAGCCATGCTCTGACAGTCTCAGTTGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGG
TTACACTGGGAGACACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTGTTATTAGACCCCTTCTTTCCATTGGTTTAATT

CGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTC
GTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAGCCAGACTTCCCGTGTCCTTTCC
ACCGGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGG
GTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTG
TTGTCTGCATGTAACTTAATACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCT
AGGCCAGTAAGTAGTGCTTGTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCT
GCTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGG
AGAGCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTGGCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCT
CTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGCCGCAGGCACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGAT
GTGTGGGCCCACCGGCCCCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCAGGTCTCCAGAG
CTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGA
GTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCC
TGTGATGTCATCTGGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAAGGCCTGGCAGAGTCTTTCCCAGGGA
AAGCTACAAGCAGCAAACAGTCTGCATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCAAGAAAGGCTCTGG
TGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAGG
GGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGAGGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTA
ACTCAAAGCCTTAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACCAGTGAT
ACACCCGGCACCCTGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGC
TGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCCAGT

TTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTG
TTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCA
CAGGCATTAGTGCCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGGGTCCTGGCCATCCGTGAGATCTTCCCAGGTGTGCCGT
TTTCTCTGGAAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCCATGGAGCACAGGCAGACAGAAGTCCCCGCCC
CAGCTGTGTGGCCTCAAGCCAGCCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGCTGGTTCCGTCACCCCCTCCCAAGGAAGTAGG
TCTGAGCAGCTTGTCCTGGCTGTGTCCATGTCAGAGCAACGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAGCCCCCTACGA
TTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCCT
CCGGGCCCCTCACCAGCCCCAGGTCCTTTCCCAGAGATGCCTGGAGGGAAAAGGCTGAGTGAGGGTGGTTGGTGGGAAACCCTGGTTC
CCCCAGCCCCCGGAGACTTAAATACAGGAAGAAAAAGGCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCTGCCTC
CTACCCTTGCGCCTCATGACCGGAGCCATAGCCCAGGCAGGAGGGCTGAGGACCTCTGGTGGCGGCCCAGGGCTTCCAGCATGTGCCC
TAGGGGAAGCAGGGGCCAGCTGGCAAGAGCAGGGGGTGGGCAGAAAGCACCCGGTGGACTCAGGGCTGGAGGGGAGGAGGCGATCTTG
CCCAAGGCCCTCCGACTGCAAGCTCCAGGGCCCGCTCACCTTGCTCCTGCTCCTTCTGCTGCTGCTTCTCCAGCTTTCGCTCCTTCAT
GCTGCGCAGCTTGGCCTTGCCGATGCCCCCAGCTTGGCGGATGGACTCTAGCAGAGTGGCCAGCCACCGGAGGGGTCAACCACTTCCC
TGGGAGCTCCCTGGACTGGAGCCGGGAGGTGGGGAACAGGGCAAGGAGGAAAGGCTGCTCAGGCAGGGCTGGGGAAGCTTACTGTGTC
CAAGAGCCTGCTGGGAGGGAAGTCACCTCCCCTCAAACGAGGAGCCCTGCGCTGGGGAGGCCGGACCTTTGGAGACTGTGTGTGGGGG
CCTGGGCACTGACTTCTGCAACCACCTGAGCGCGGGCATCCTGTGTGCAGATACTCCCTGCTTCCTCTCTAGCCCCCACCCTGCAGAG
CTGGACCCCTGAGCTAGCCATGCTCTGACAGTCTCAGTTGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGG
TTACACTGGGAGACACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTGTTATTAGACCCCTTCTTTCCATTGGTTTAATT

Why do we care?

- Understand the relationship between genotype and phenotype.



v.

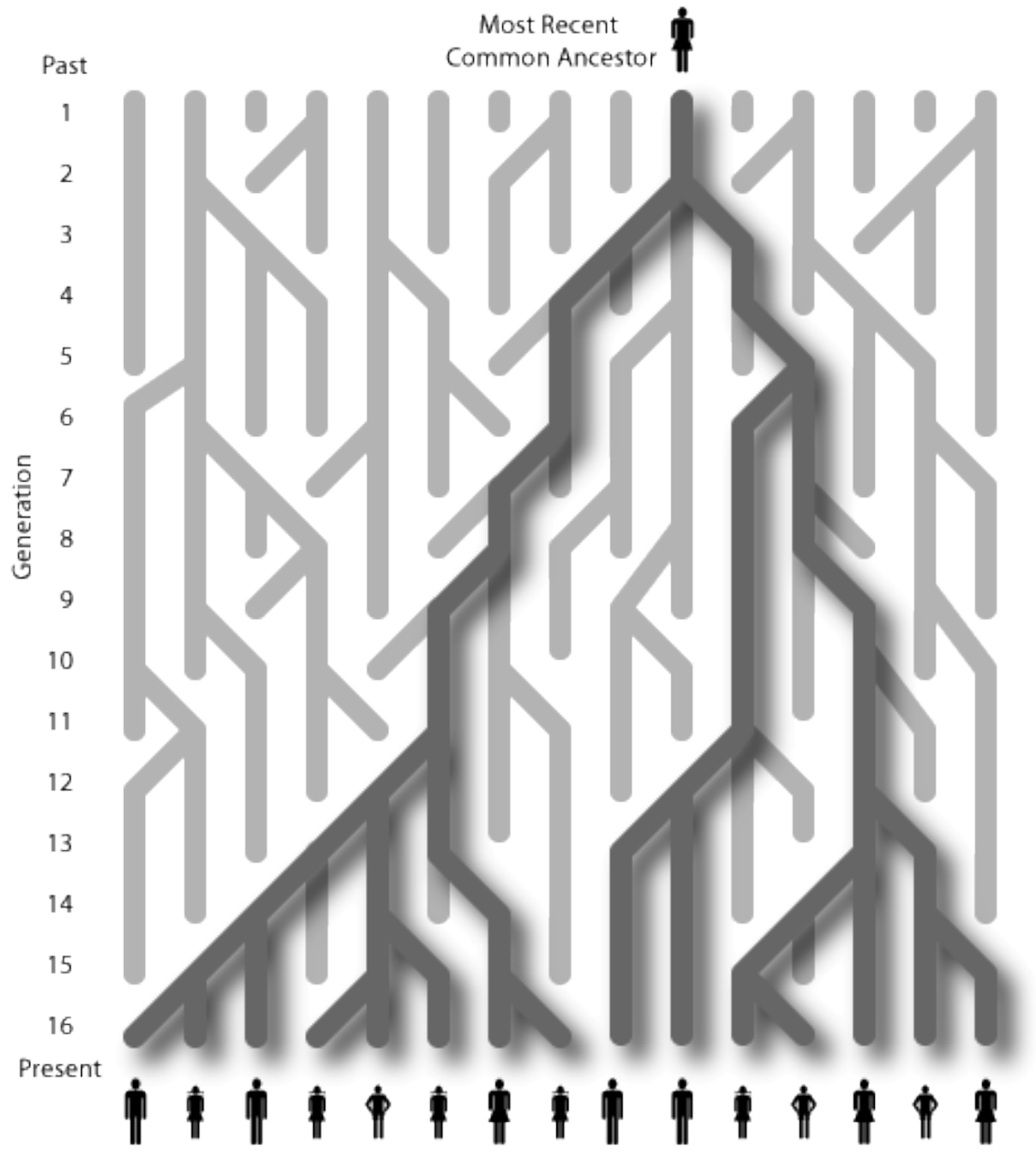
Cases
(have disease)

Controls
(no disease)

Complex diseases
(multiple genes contribute to risk)

Why do we care?

- Bread crumbs of evolution



http://www.fos.auckland.ac.nz/~howardross/RecreateTheResearch/RTR/images/Coalescence.png

Why do we care?

- How, when, where does our genome evolve?

Types of genetic variation

ctc**c**gag
ctc**t**gag

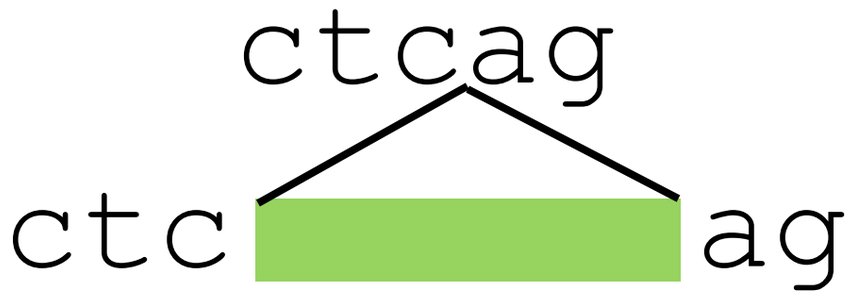
Single-nucleotide polymorphisms (**SNPs**)

“spelling mistakes”

ctc**--**ag
ctc**tg**ag

Insertion-deletion polymorphisms (**INDELs**)

“extra or missing letters”



Structural variants (**SVs**)

“extra, missing or reordered chapters”

Properties of genetic variation

	Single-nucleotide (**SNPs**)	Insertion-deletions (**INDELs**)	Structural variants (**SVs**)
	ctc**c**gag ctc**t**gag	ctc**--**ag ctc**tg**ag	ctcag ctc ag
Size	1bp	1-100bp	100bp-1Mb+
Frequency	3 million / genome	300K / genome	3,000 / genome
Detection Difficulty	Easy	Medium	Hard

How different are we?

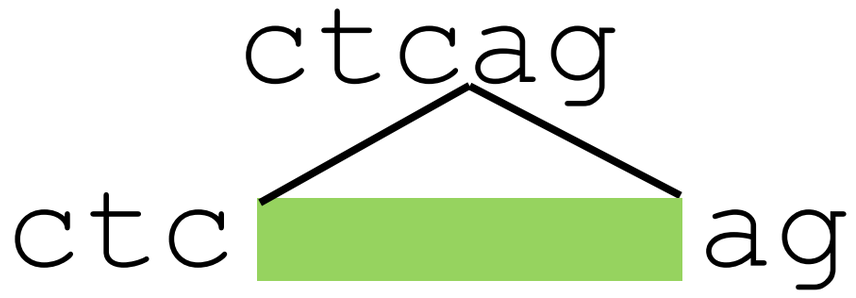
Single-nucleotide (**SNPs**)

```
ctccgag
ctctgag
```

Insertion-deletions (**INDELs**)

```
ctc--ag
ctctgag
```

Structural variants (**SVs**)



How different are we?

The genomes of any two humans are ~99.5% identical

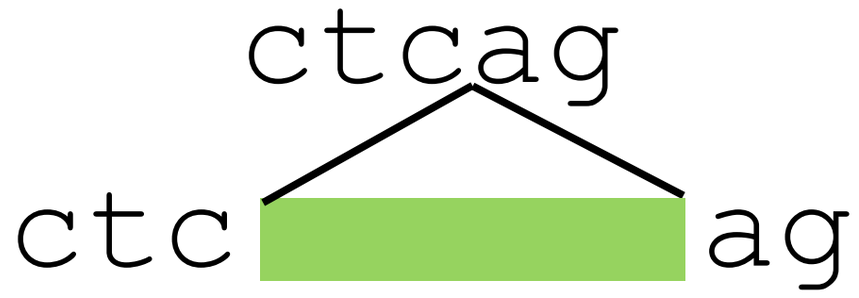
Single-nucleotide (**SNPs**)

ctccgag
ctctgag

Insertion-deletions (**INDELs**)

ctc--ag
ctctgag

Structural variants (**SVs**)



How different are we?

The genomes of any two humans are ~99.5% identical

But, the genome is big.

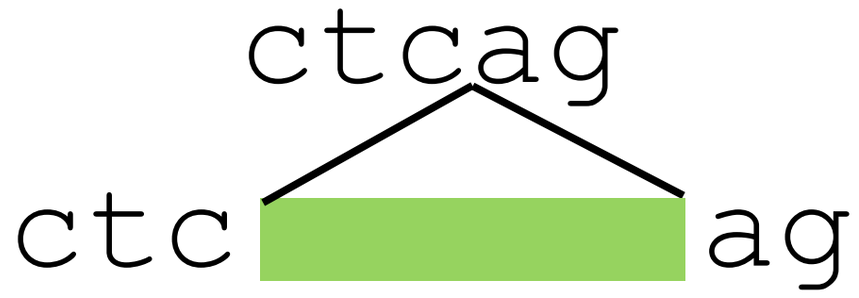
Single-nucleotide (**SNPs**)

```
ctccgag
ctctgag
```

Insertion-deletions (**INDELs**)

```
ctc--ag
ctctgag
```

Structural variants (**SVs**)



How different are we?

The genomes of any two humans are ~99.5% identical

But, the genome is big.

~21,000,000 different base pairs.

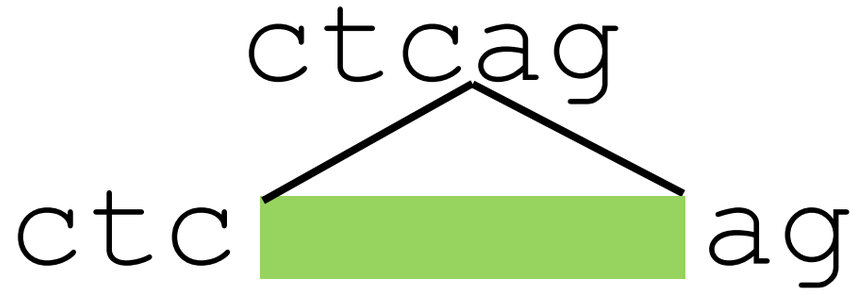
Single-nucleotide (**SNPs**)

```
ctccgag
ctctgag
```

Insertion-deletions (**INDELs**)

```
ctc--ag
ctctgag
```

Structural variants (**SVs**)



It all starts with mutation

```
acctccgagta    acctccgagta
acctccgagta    acctccgagta
acctccgagta    acctccgagta
acctccgagta    acctccgagta
acctccgagta    acctccgagta
```

a toy population of 10 identical chromosomes

Mutation creates genetic diversity

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctcTgagta

mutation:
private to this chromosome

From mutation to polymorphism

acctccgagta

acctccgagta

acctcTgagta

acctccgagta

acctcTgagta

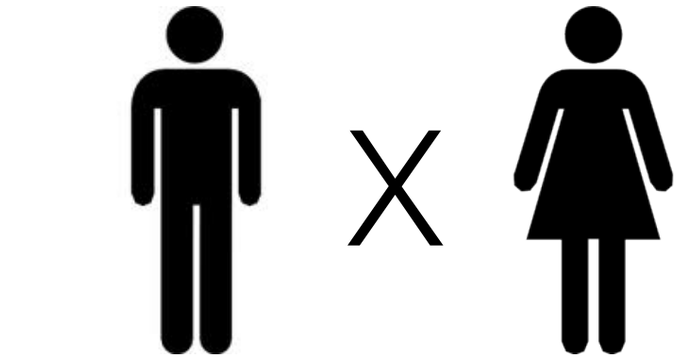
acctcTgagta

acctccgagta

acctccgagta

acctcTgagta

de novo mutation: the birth of new variation



♂ctccgag ♀ctccgag
♂ctccgag ♀ctccgag

***Example**: Mom and dad are homozygous for the same alleles.*

New mutation occurs in father's or mother's germ cell

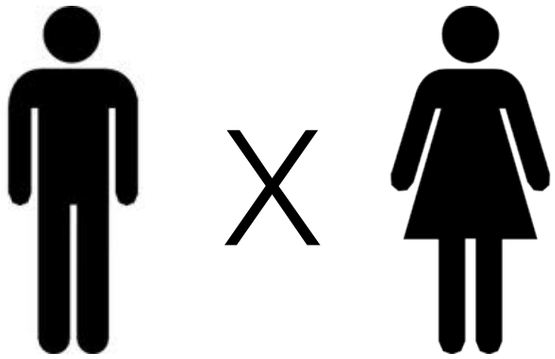
♂ctccgag → ♂ctccgag

***Note**: This is a derivative chromosome of the one the father inherited from hist parents. The mutation occurred in his gamete (sperm) and was passed on to the child.*

♂ ctctgag
♀ ctccgag

Kid is heterozygous owing to *de novo mutation.*
(C/T)

How existing (germline) variation is inherited



ctccgag
ctctgag

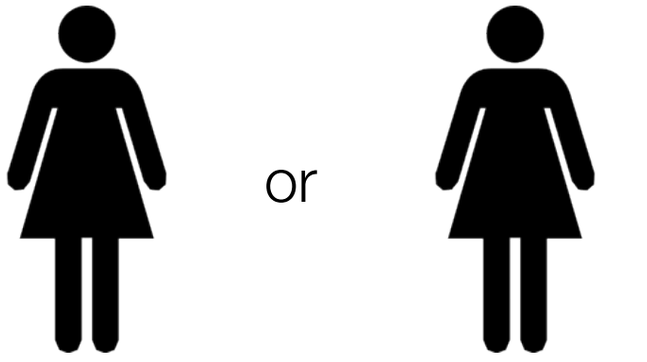
ctccgag
ctctgag

***Example**: Mom and dad are heterozygous; that is, the zygote from which they developed was comprised of a sperm and egg with two different alleles*

♂ ♀ denotes from which parent the chrom. was inherited

♂ ctccgag
♀ ctccgag

Kid is homozygous (C/C)



♂ ctccgag
♀ ctctgag

or

♀ ctccgag
♂ ctctgag

Kid is heterozygous (C/T)

♂ ctctgag
♀ ctctgag

Kid is homozygous (T/T)

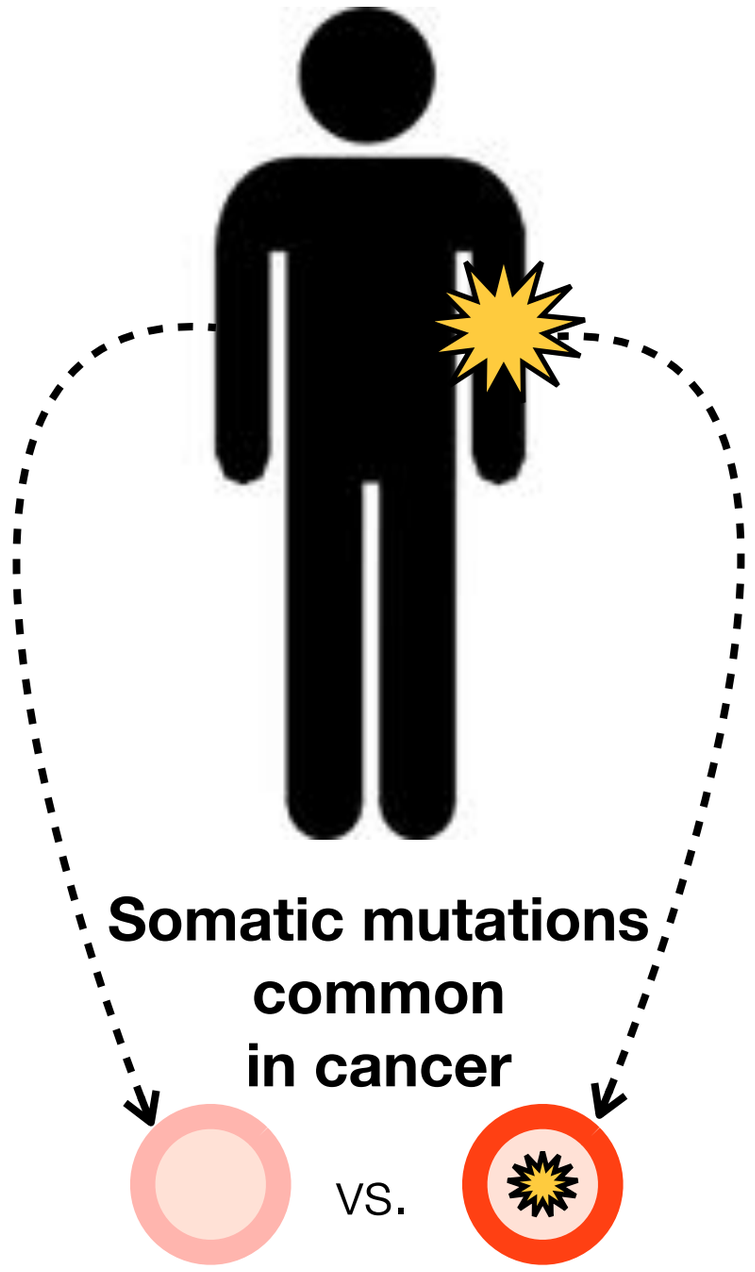
somatic mutations

Germline mutation

- occur in sperm or egg.
- are heritable

Somatic mutation

- non-germline tissues.
- are not heritable



compare DNA from cancer cells to healthy cells from same individual

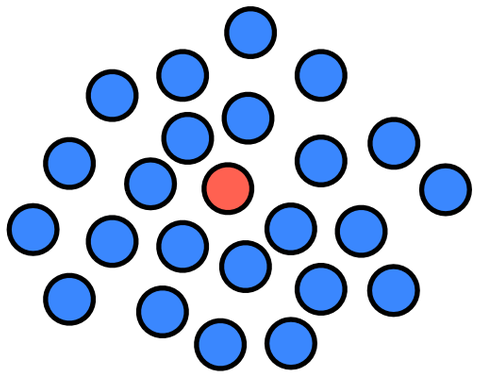
Selection and genetic drift

All other chromosomes

acctccgagta

acctctgagta

Chromosome with new allele



What if the mutation is:

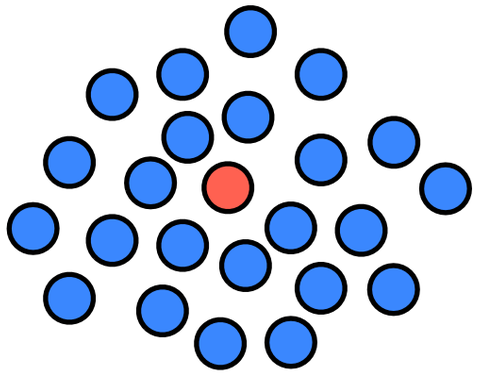
Selection and genetic drift

All other chromosomes

acctc**c**gagta

acctc**t**gagta

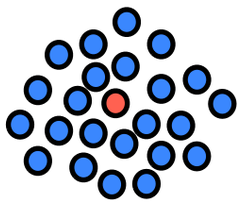
Chromosome with new allele

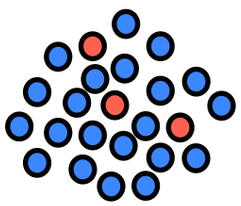


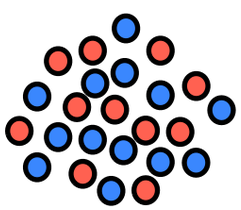
What if the mutation is:

beneficial

Time







Positive
selection

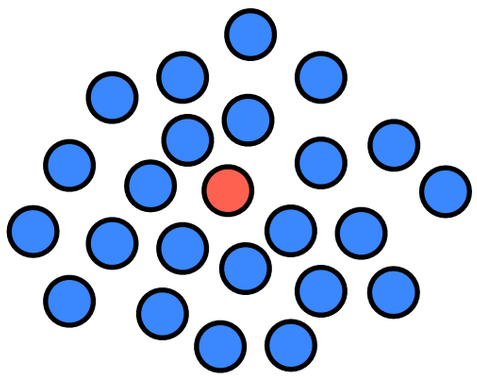
Selection and genetic drift

All other chromosomes

acctc**c**gagta

acctc**t**gagta

Chromosome with new allele

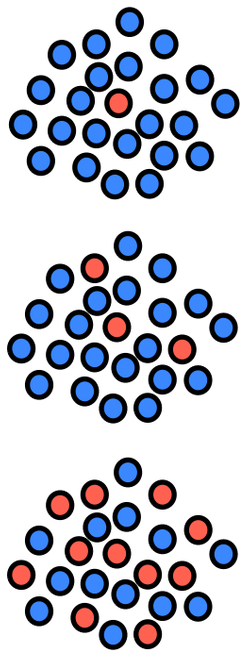


What if the mutation is:

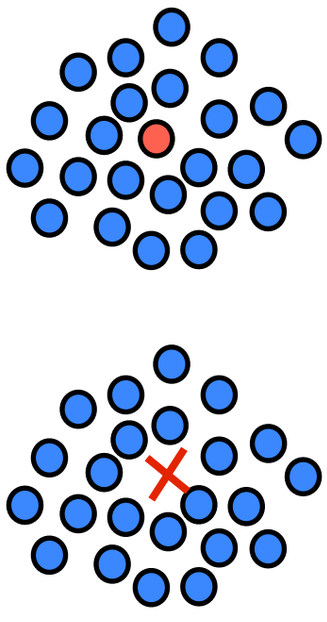
beneficial

deleterious

Time



Positive selection



Negative selection

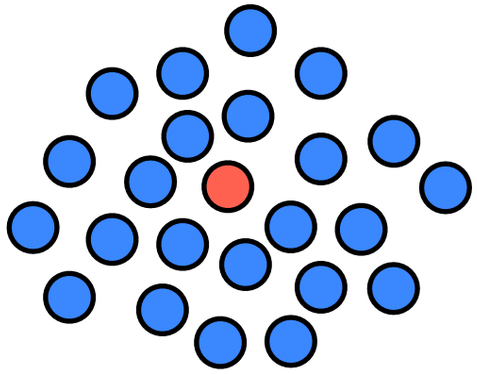
Selection and genetic drift

All other chromosomes

acctc**c**gagta

acctc**t**gagta

Chromosome with new allele



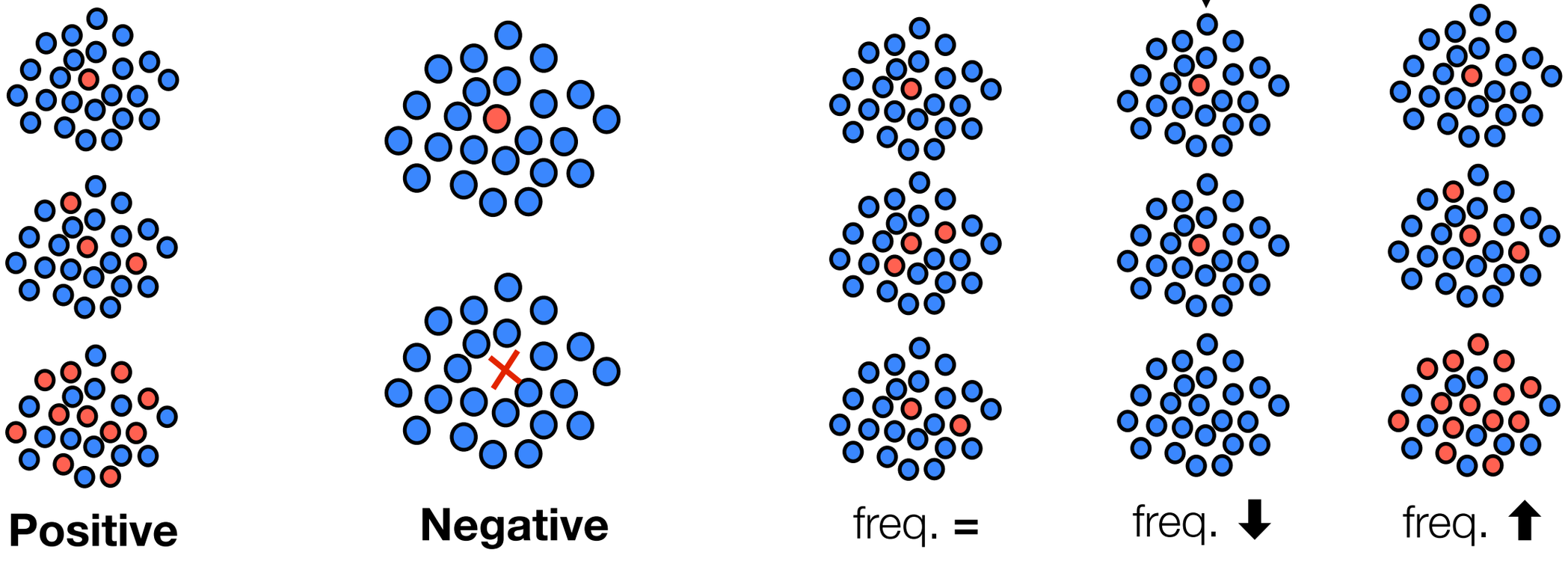
What if the mutation is:

beneficial

deleterious

neutral

Time



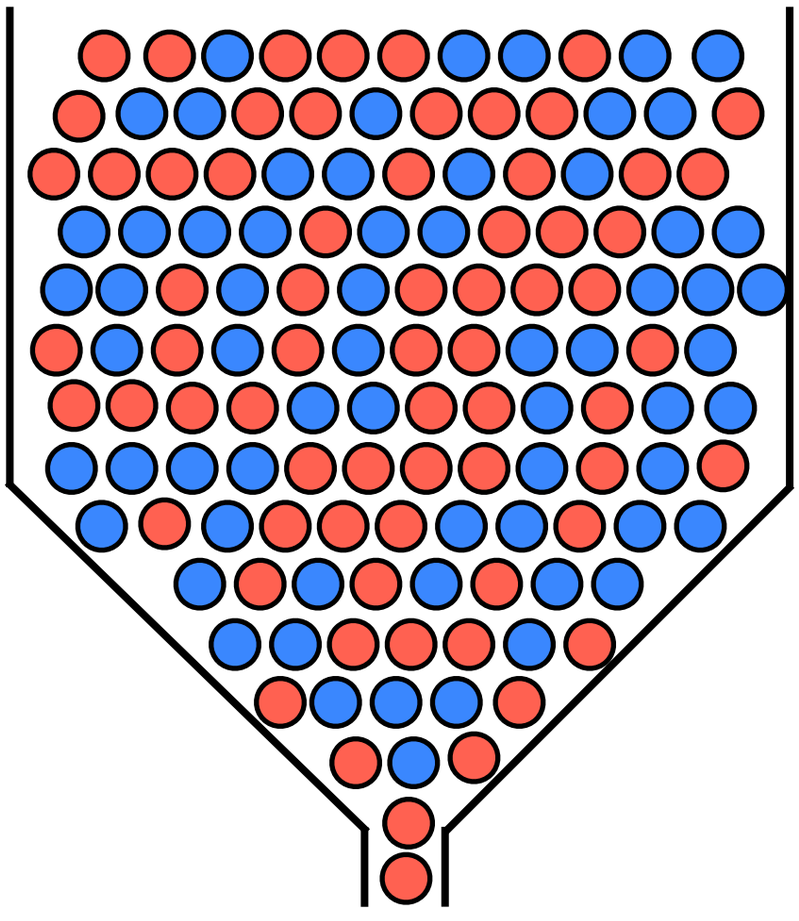
Positive selection

Negative selection

Genetic “drift” - random process

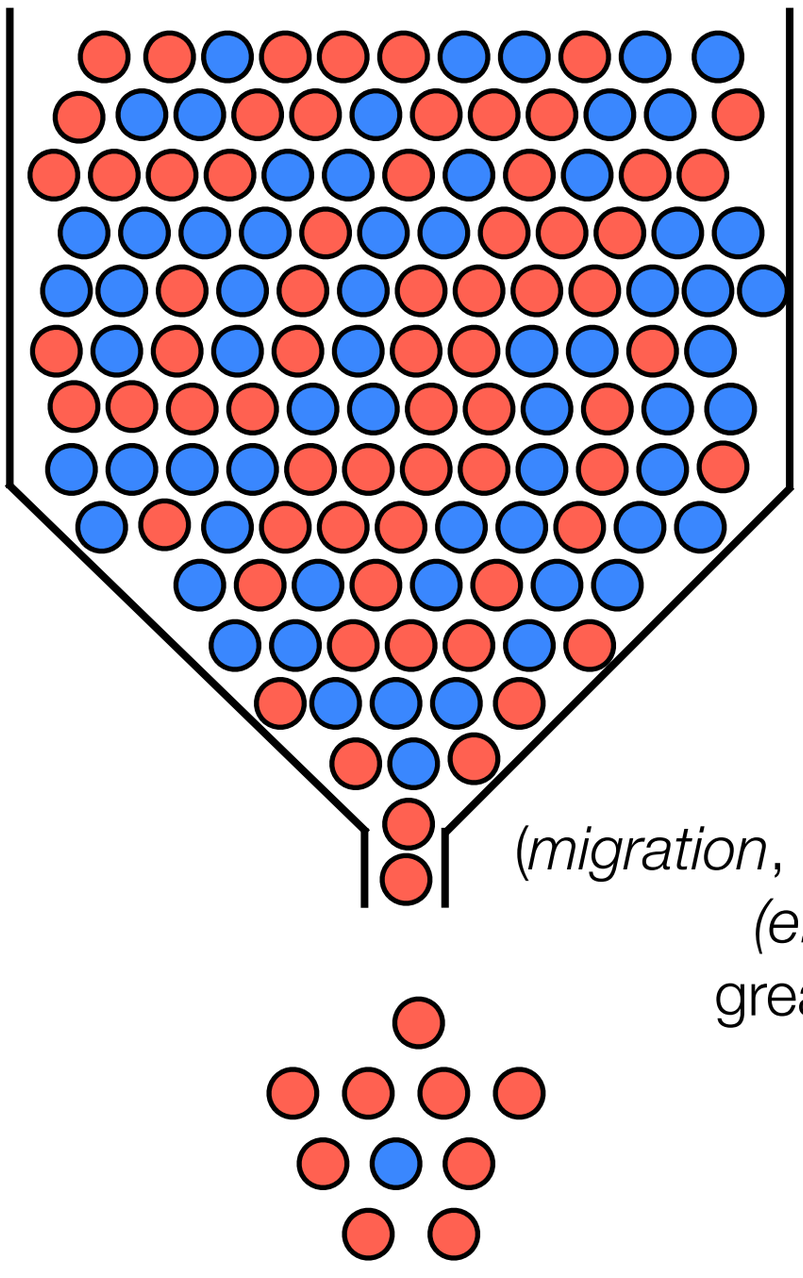
Genetic bottleneck

"Mixed" population



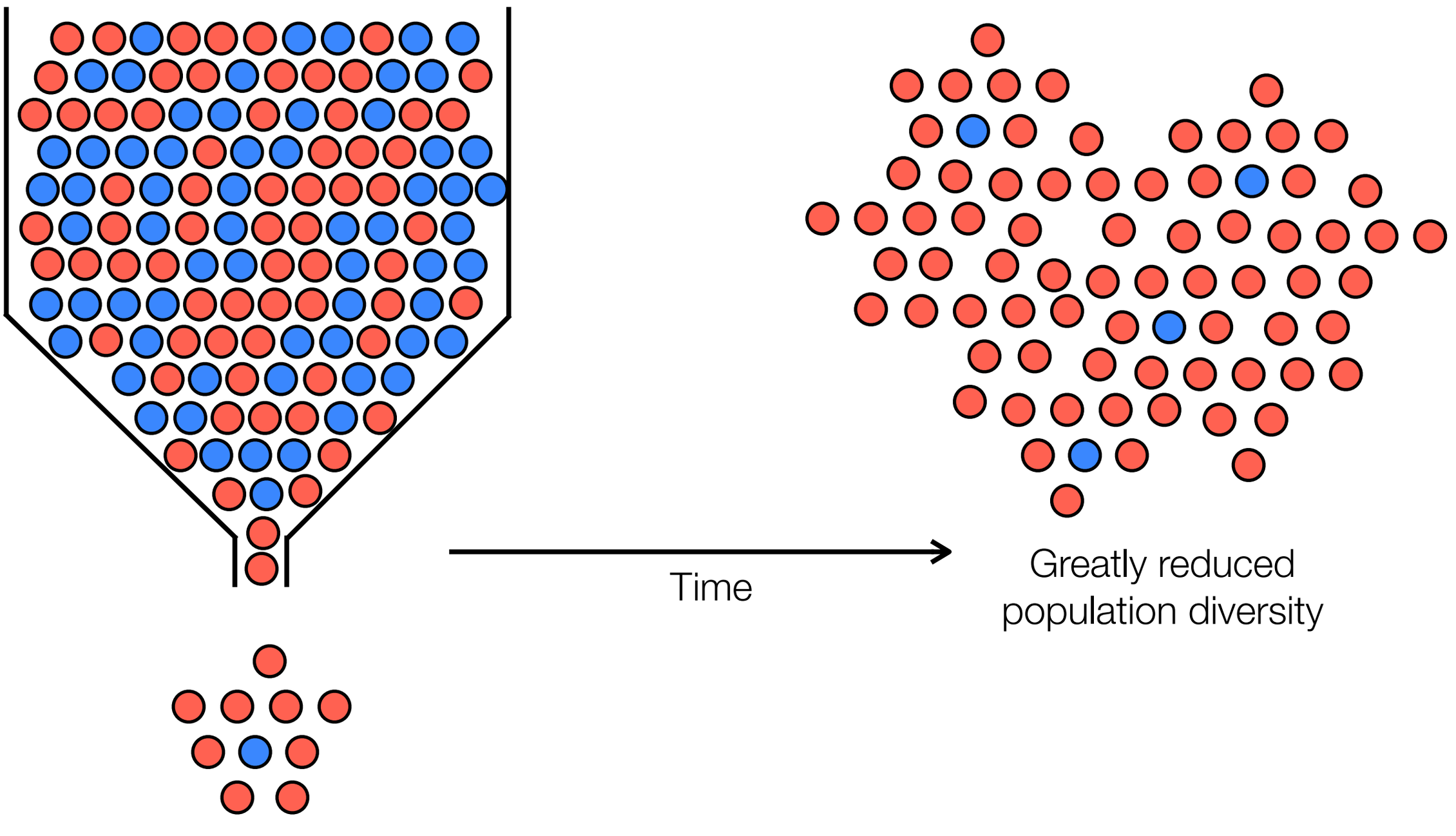
Genetic bottleneck

"Mixed" population



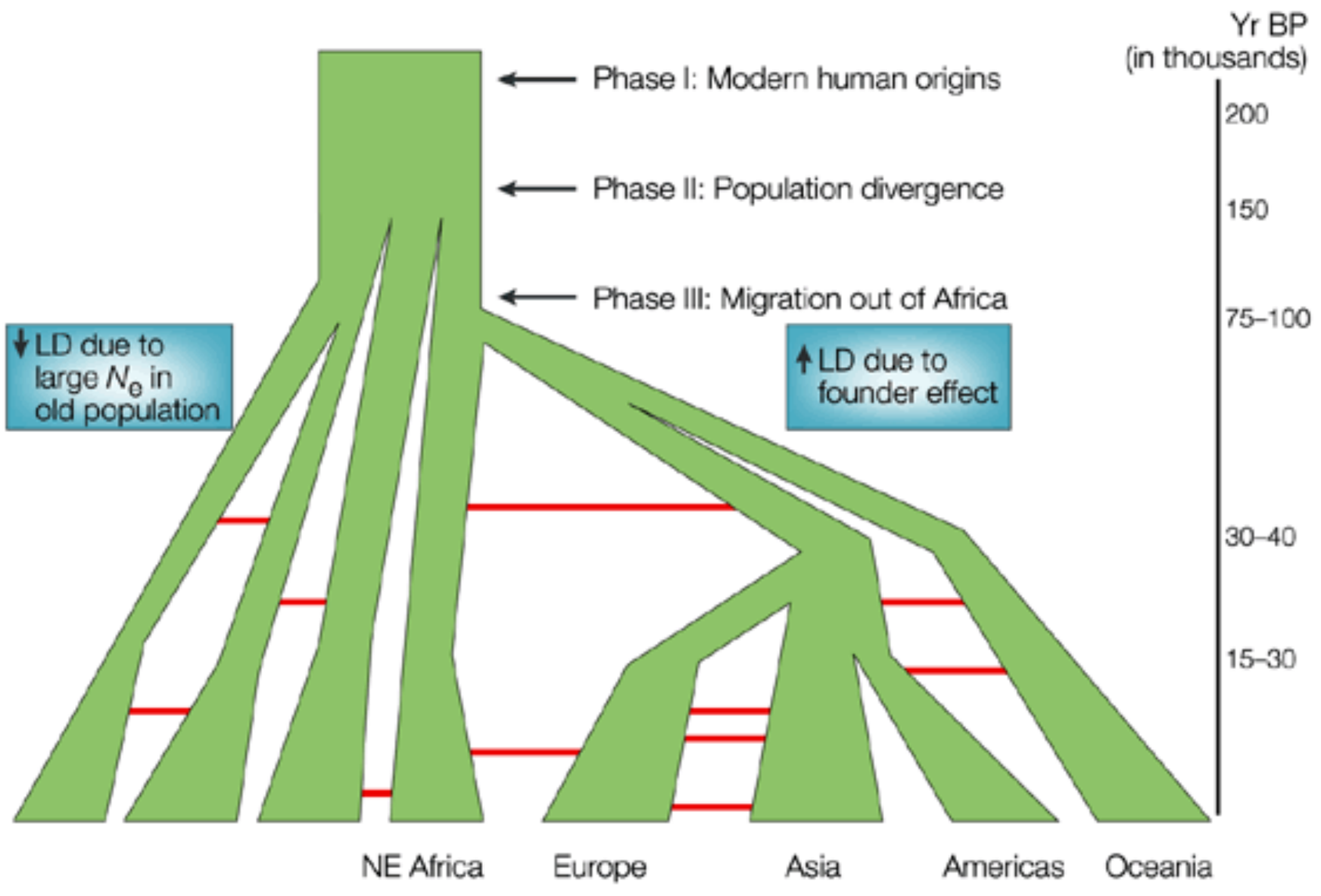
Bottleneck event
(*migration*, *war*, *disease, nearly extinct animals*
(e.g., elephant seals in Baja)
greatly reduces population size

Genetic bottleneck



Out of Africa Theory

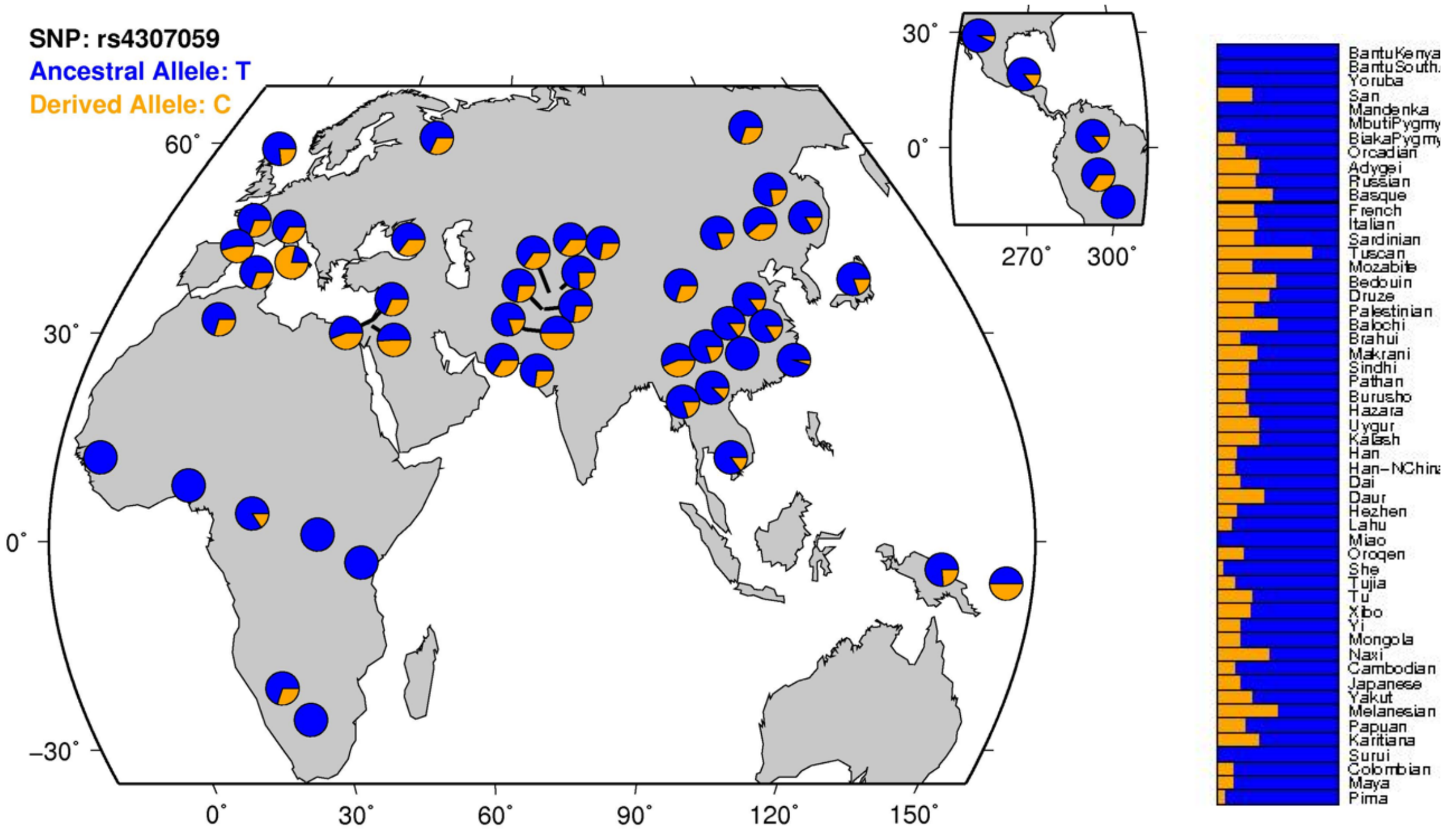
bottlenecks & reduced diversity



Nature Reviews Genetics **3**, 611-621 (August 2002)

A founder effect.

Population stratification



What do allele frequencies tell us?

- Sequencing many individuals allows us to look at genetic variation and their frequencies at a moment in time.
- But, **BUT!**, the frequency of a variant is in the population is just a snapshot in time
    - e.g., a frame in a movie of selection, genetic drift.
- For example, a variant may be **rare** b/c it just occurred, b/c of purifying selection, or b/c of genetic drift.
- Or, a variant may be **common** because it is beneficial or because of genetic drift.

A map of human genetic variation

1000 Genomes
A Deep Catalog of Human Genetic Variation

Sequencing is dirt cheap.
Let's catalog ~all genetic variation!

ARTICLE

doi:10.1038/nature09534

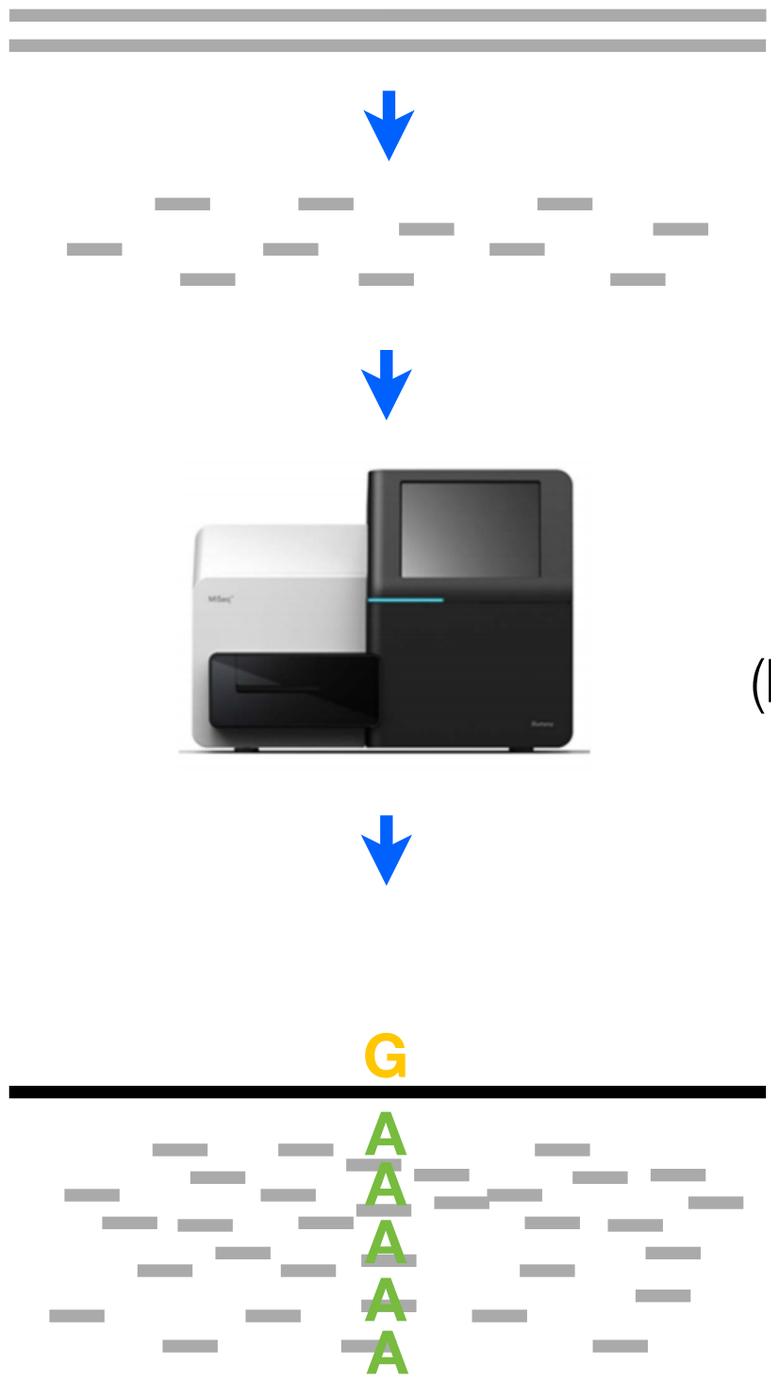
A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

More on this in the afternoon...

Detecting genetic variation

It all starts with DNA sequencing



Sequence alignment is non-trivial and the strategy depends upon the type of genetic variation you are trying to detect

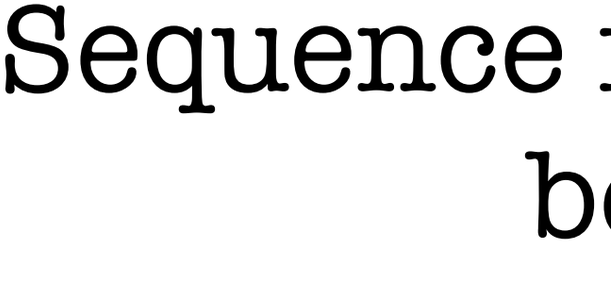
Align DNA to a reference genome. Comparing sample DNA to reference reveals genetic differences.

Important:
DNA sequencers make mistakes.

Sequence mapping: where does the read belong in the genome?

d. the genome is big and repetitive.

e. DNA sequencing yields short sequences (<100 characters)

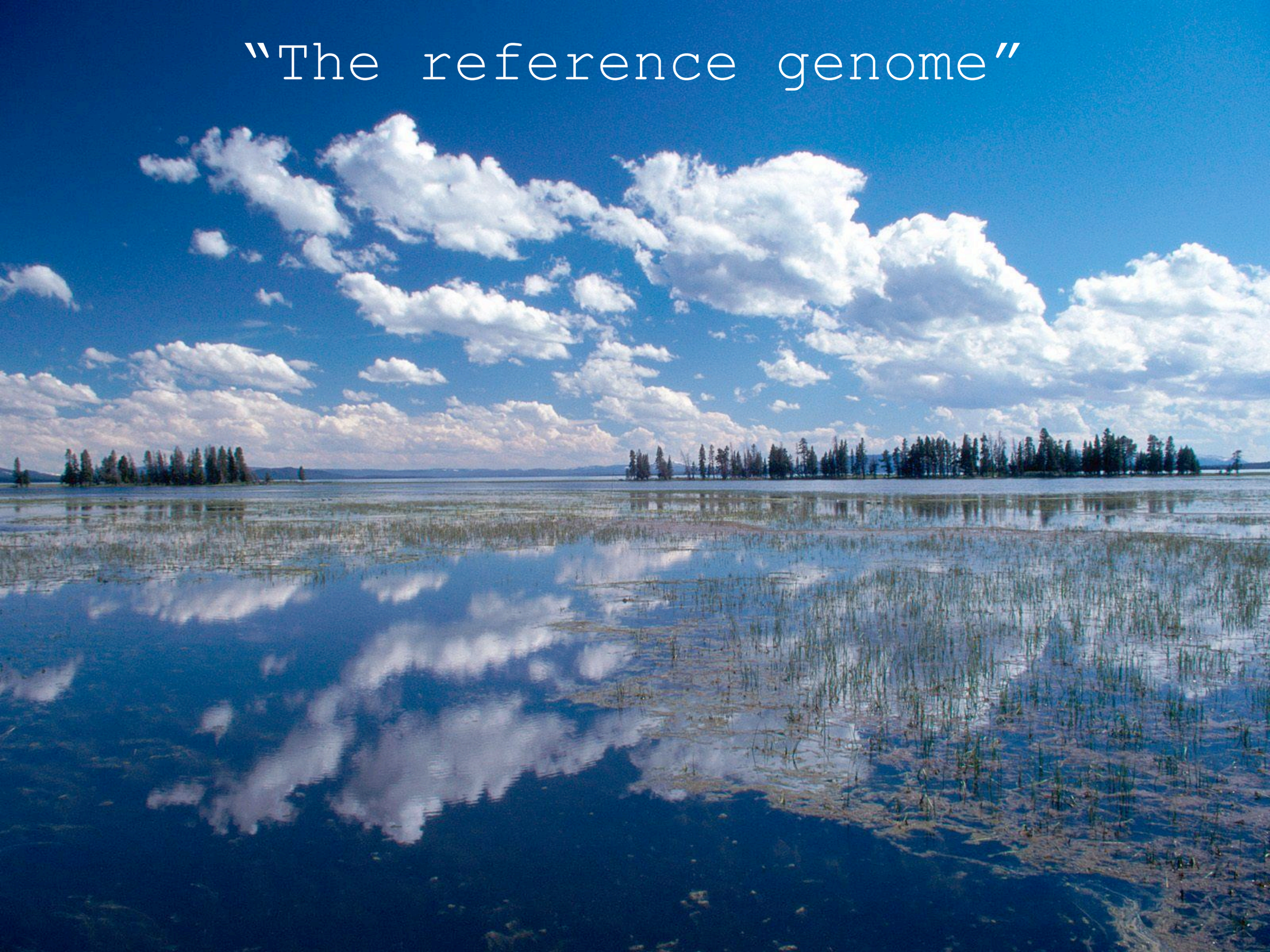


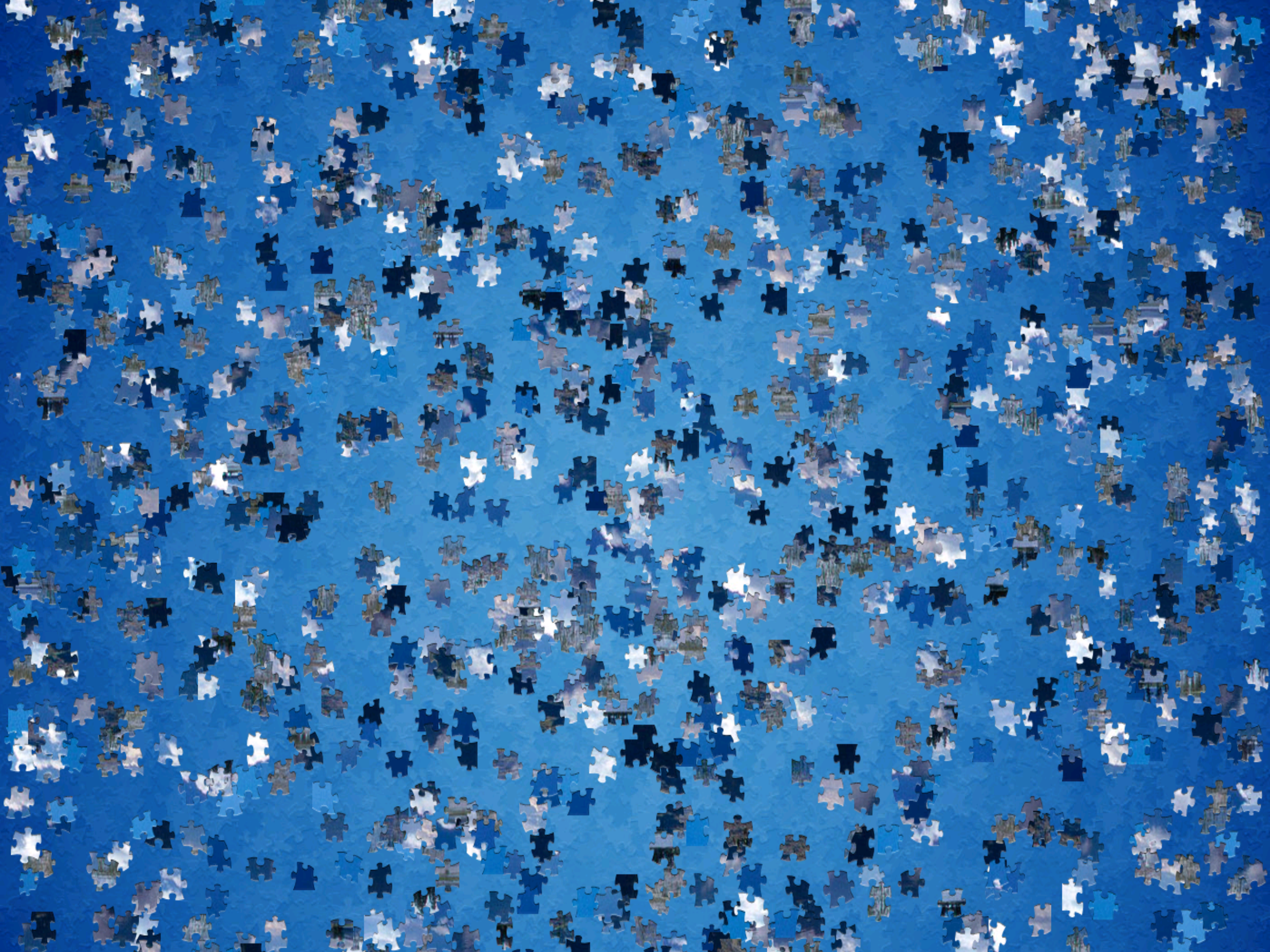
cctaaccct
acccctgg
acccctgg
tcccctgg
cctaaccct
cctaaccct
ggccctgg
acccctgg
acccctgg
actaaccct
cctaaccct
cctaaccct
acccctgg
acccctgg
cctaaccct
cctaaccct

Where the small subsequences belong?

How do we “stitch” them together to make a complete genome?

"The reference genome"





Sequence alignment

Smith-Waterman, Needleman-Wunsch

Reference `cgggtatccaa`

Read `ccctaggtccca`

What is the best alignment?

Reference cgggtatccaa

Read ccctaggtccca

Reference cgggta--t-ccaa

Read ccc-taggtccc-a

Reference cgggtatccaa

Read ccctaggtccca

Reference cgggta--t-ccaa

Read ccc-taggtccc-a

Reference cgggta---tccaa

Read cc--ctaggtccca

Reference cgggtatccaa

Read ccctaggtccca

Reference cgggta--t-ccaa

Read ccc-taggtccc-a

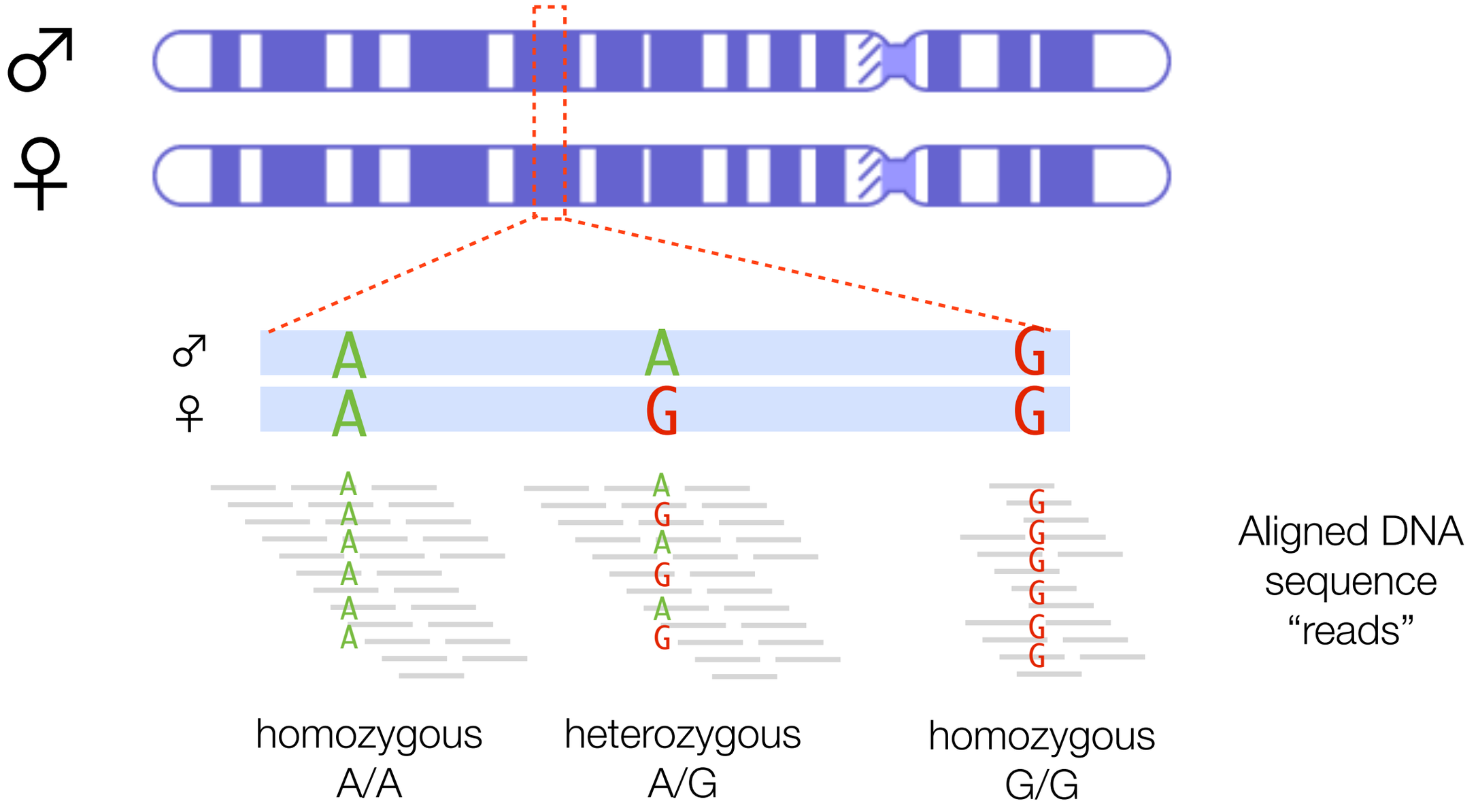
Reference cgggta---tccaa

Read cc--ctaggtccca

Reference c-gggta--tccaa

Read cc--ctaggtccca

Recall: we are diploid.

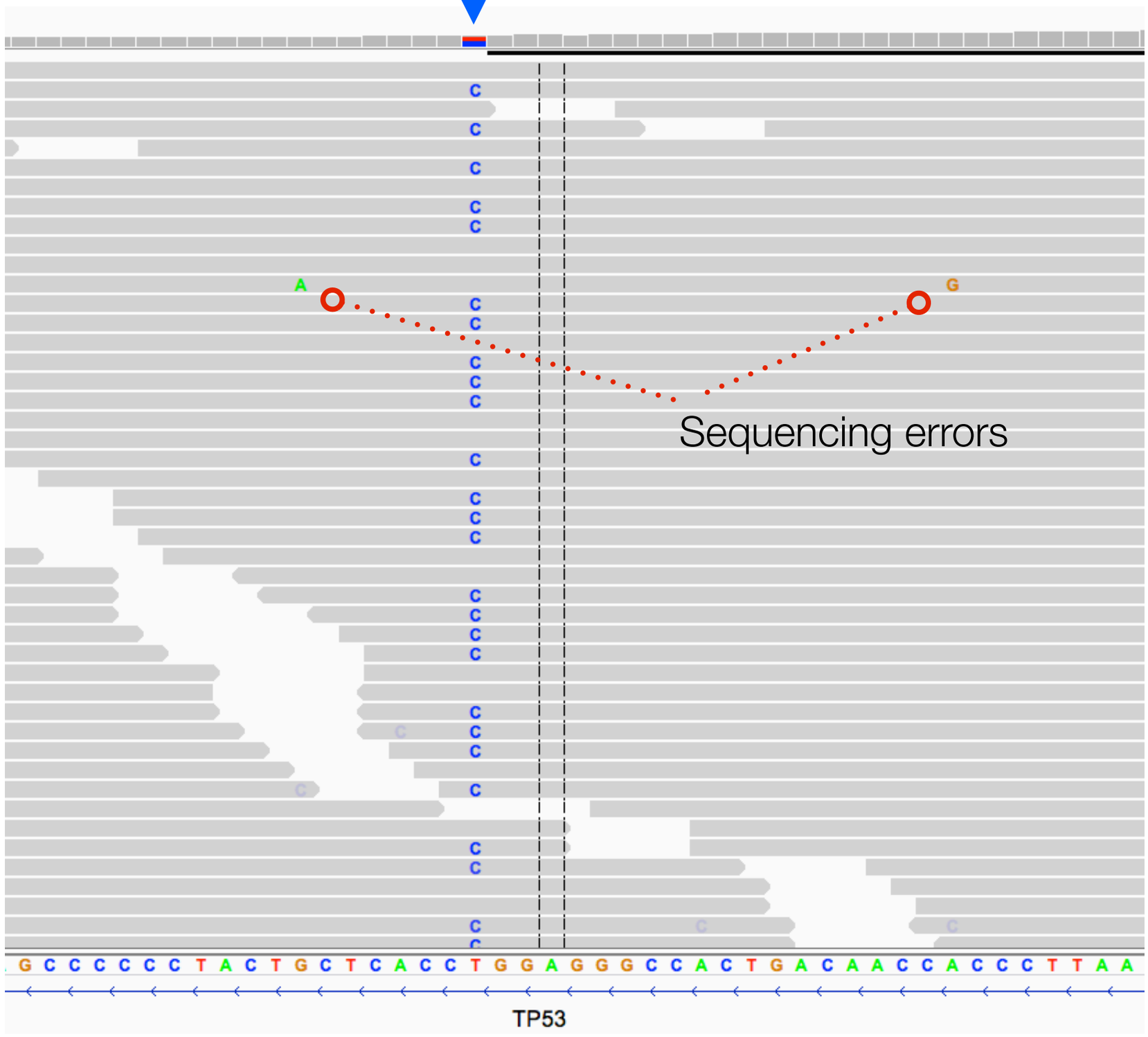


Single-nucleotide polymorphisms (SNPs)

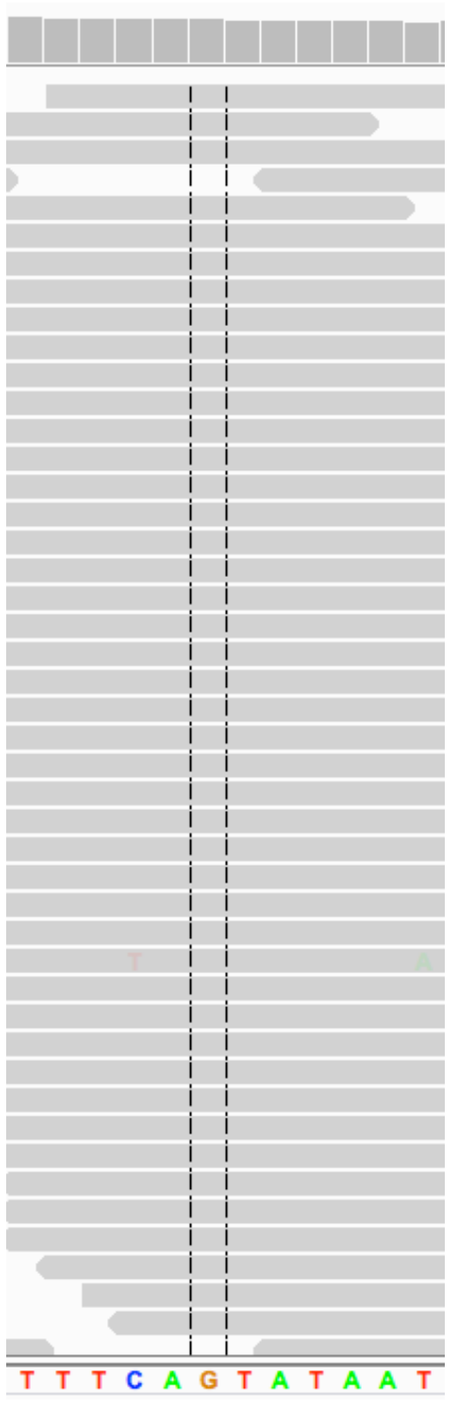
Single-nucleotide polymorphisms (SNPs)

True variation. # of T = 19

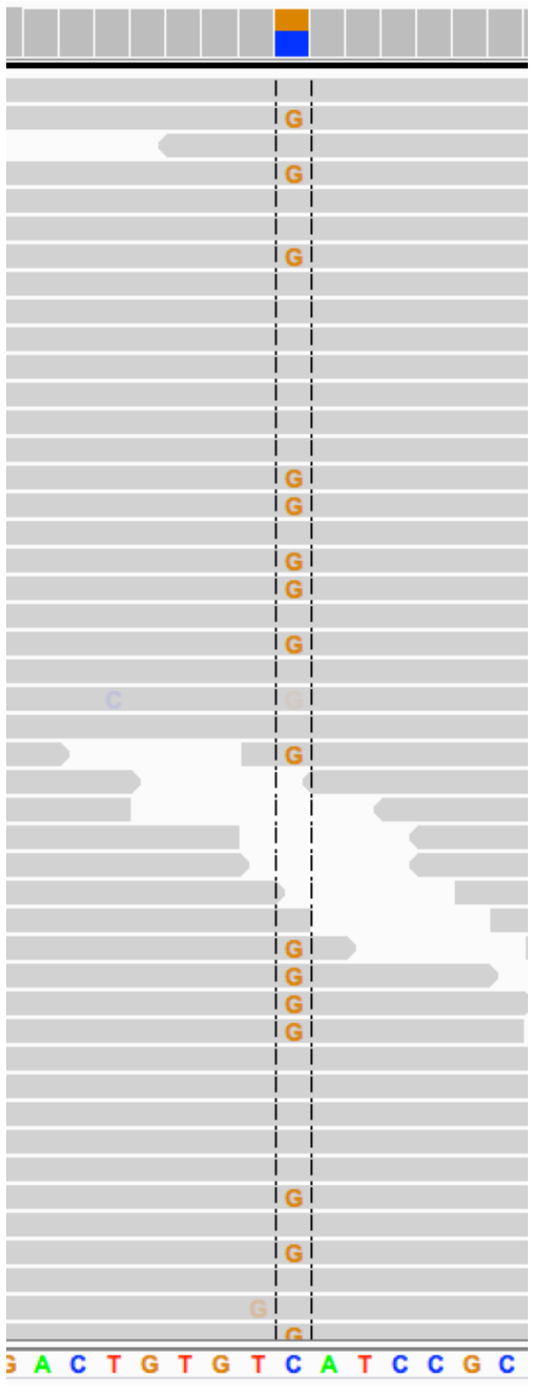
Homoyzgous or heterozygous? # of C = 25



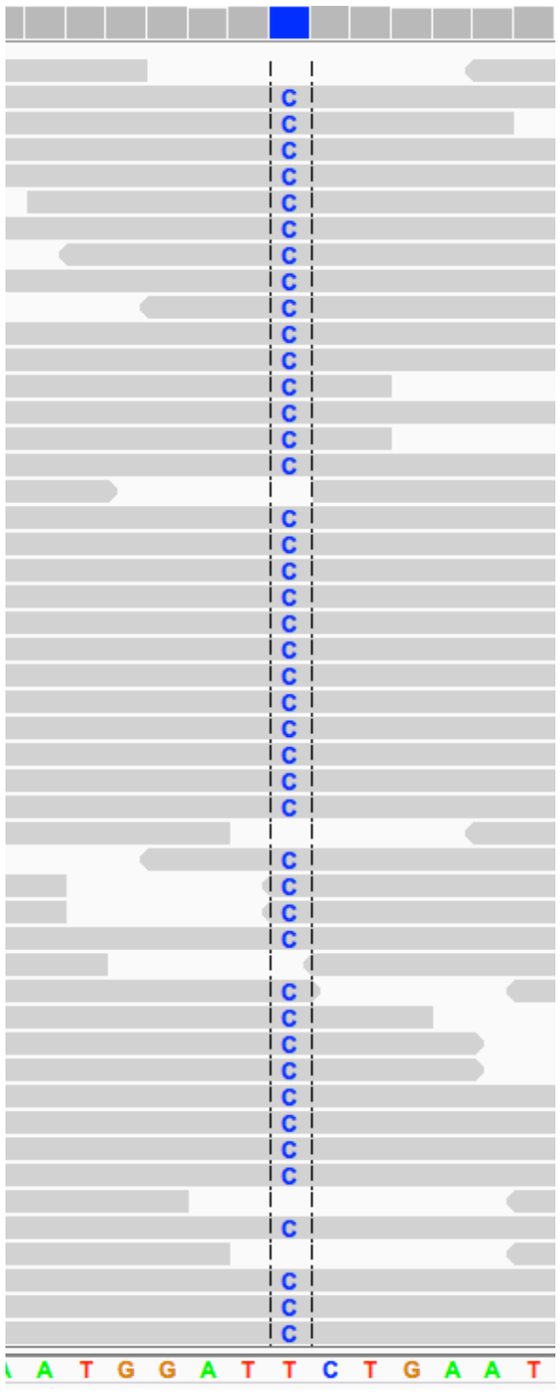
Different SNP genotypes



Homozygous for reference
(i.e., both chroms same as ref)

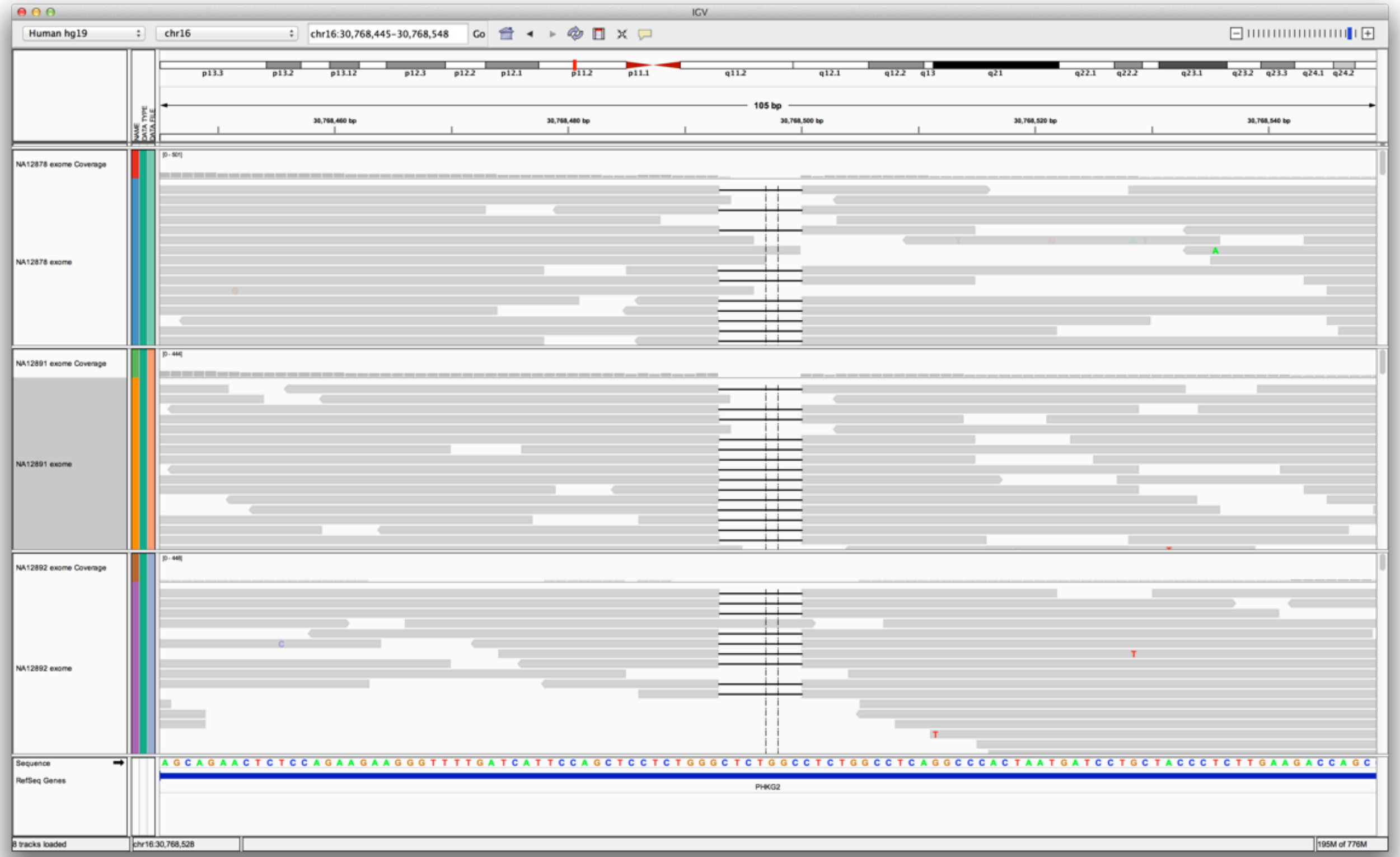


Heterozygous
(i.e., 1 chrom same as ref, 1 diff.)

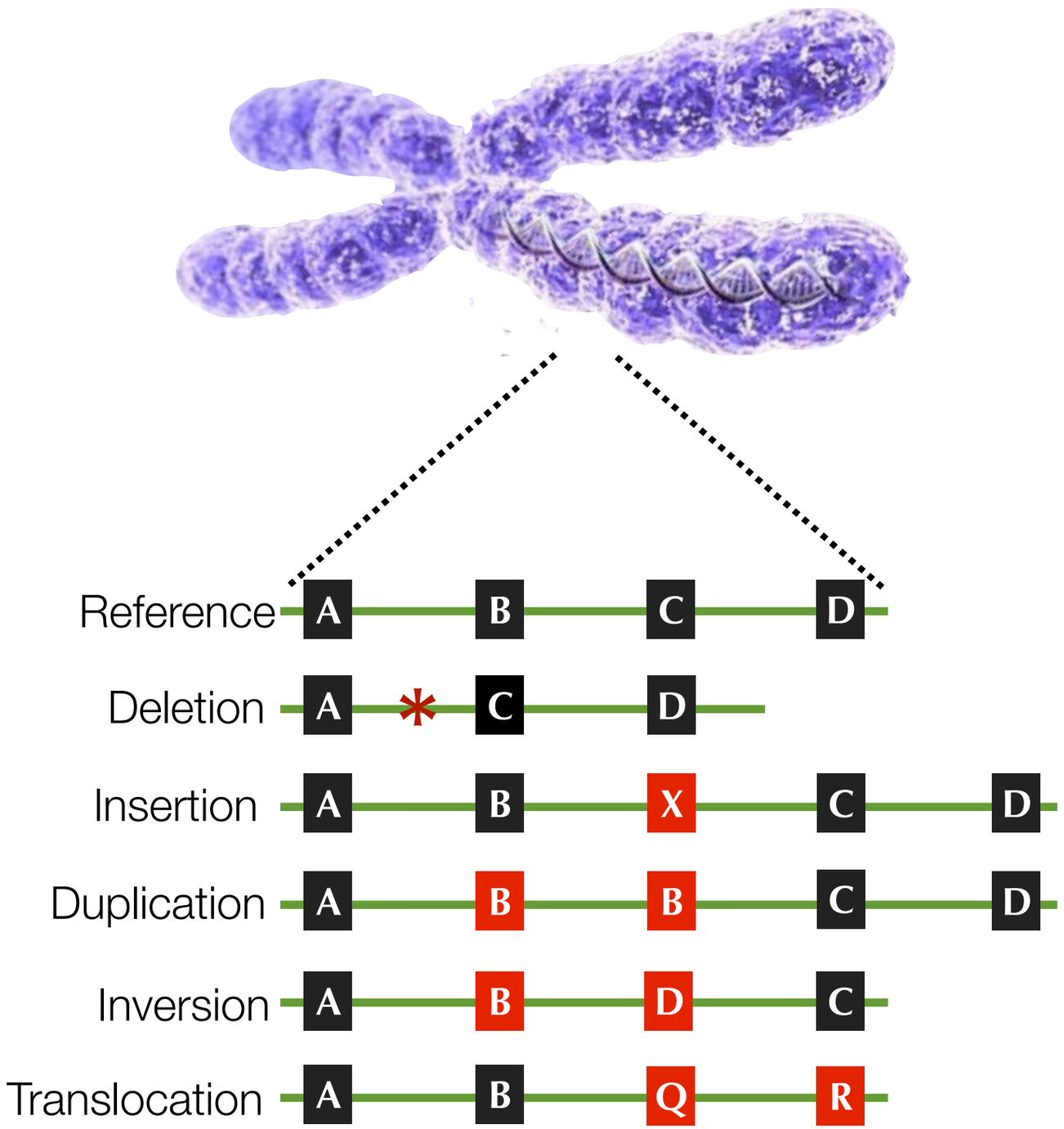


Homozygous for reference
(i.e., both chroms diff than ref)

Insertion-deletion polymorphisms (INDELs)

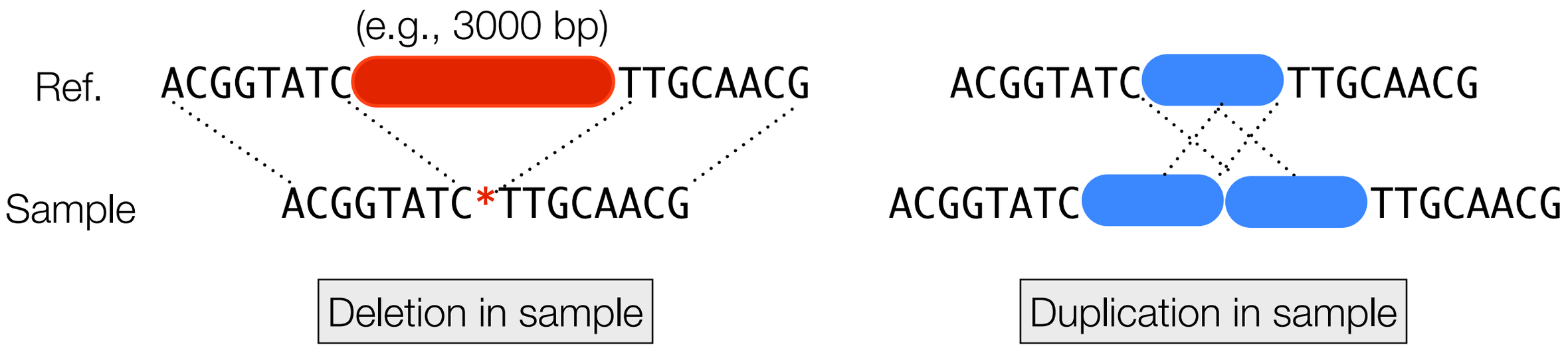


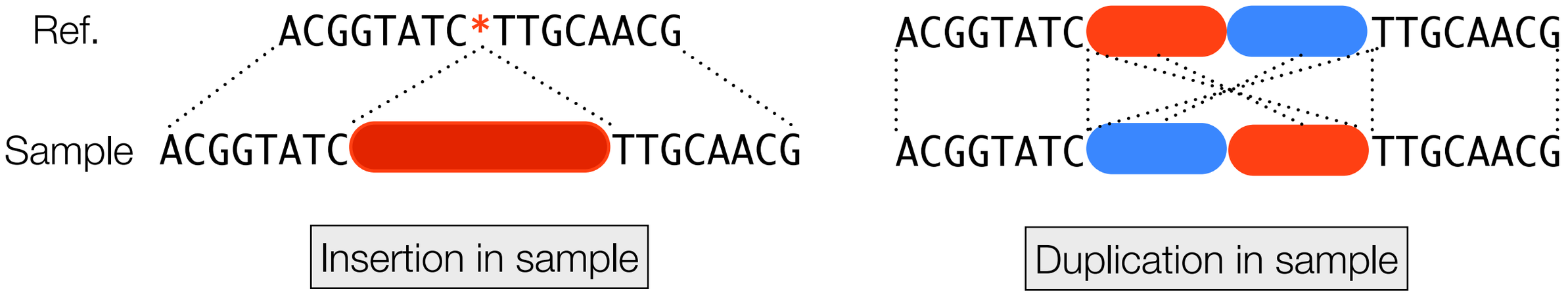
Structural Variation



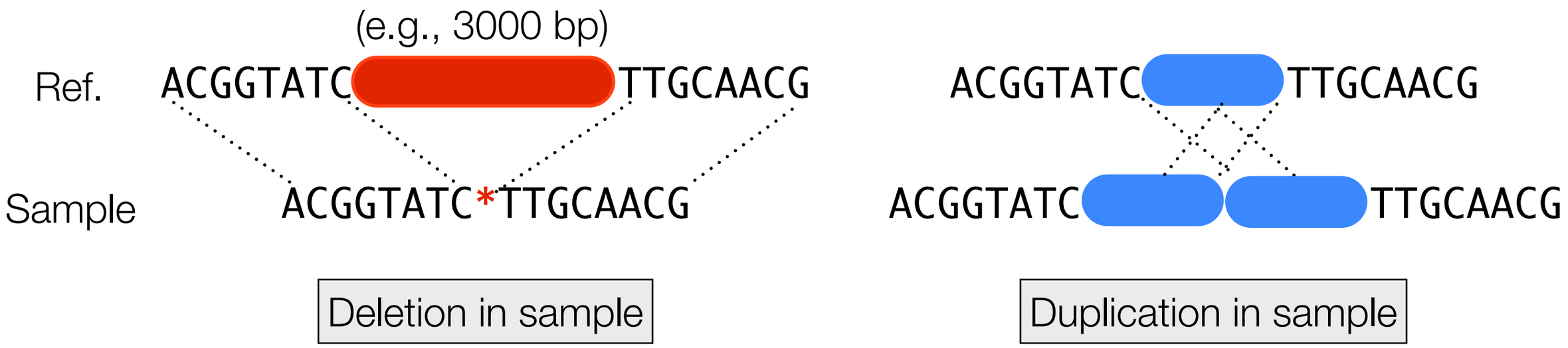
- large (>1kb) differences that affect the copy number, orientation, or location of genomic segments
- Common in mammalian genomes (thousands between two people)
- A hallmark of cancer
- A major cause of spontaneous disease
- more are functional than SNPs
- very challenging to identify

SV Alignment structures



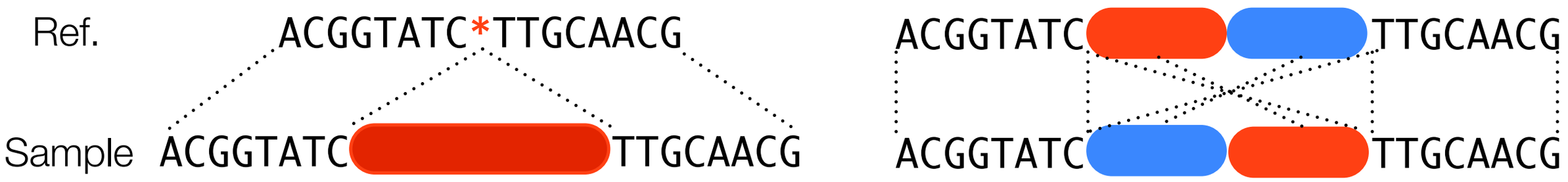


SV patterns



Deletion in sample

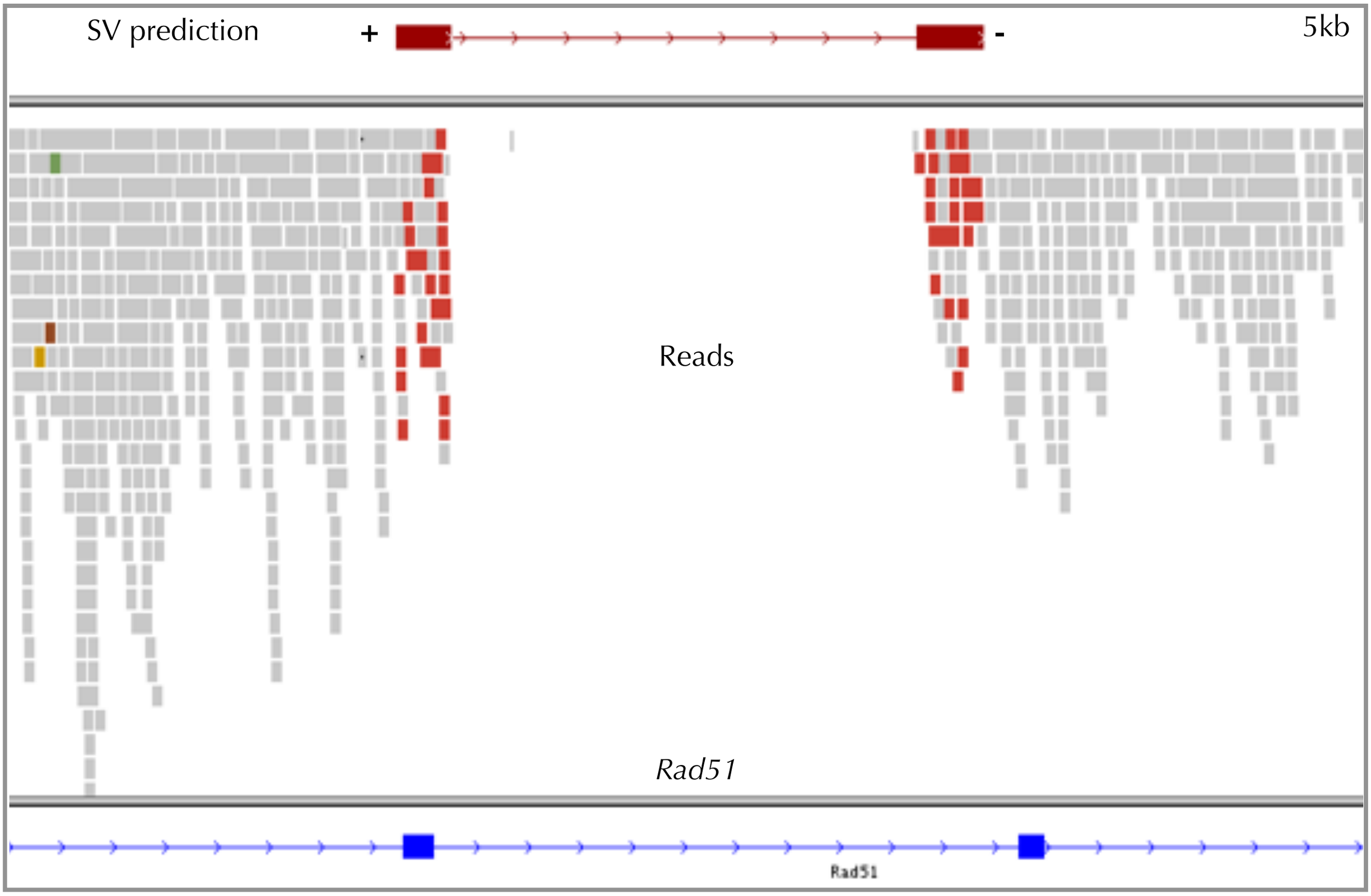
Duplication in sample



Insertion in sample

Duplication in sample

A real deletion detected with sequencing



Functional consequences

Impact sometimes hard to predict.

synonymous (silent)

	L	Q	T
Normal	ctg	cag	act
Mutated	ctg	caa	act

non-synonymous (missense)

	L	Q	T
Normal	ctg	cag	act
Mutated	ctg	cgg	act

stop-gain (nonsense)

	L	Q	T
Normal	ctg	cag	act
Mutated	ctg	tag	act

L	STOP	T

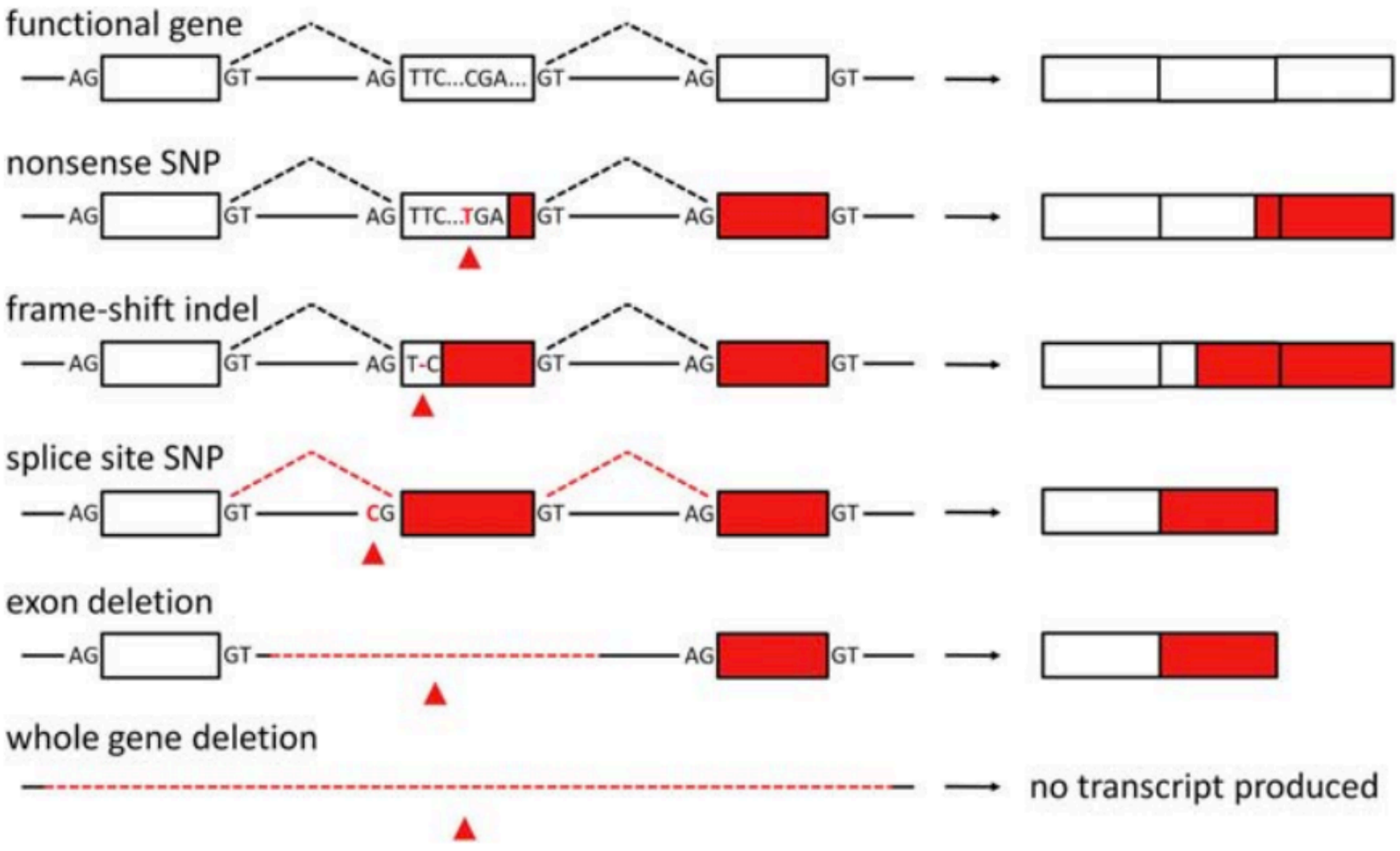
stop-loss

	L	STOP	T
Normal	ctg	tag	act
Mutated	ctg	cag	act

L	Q	T

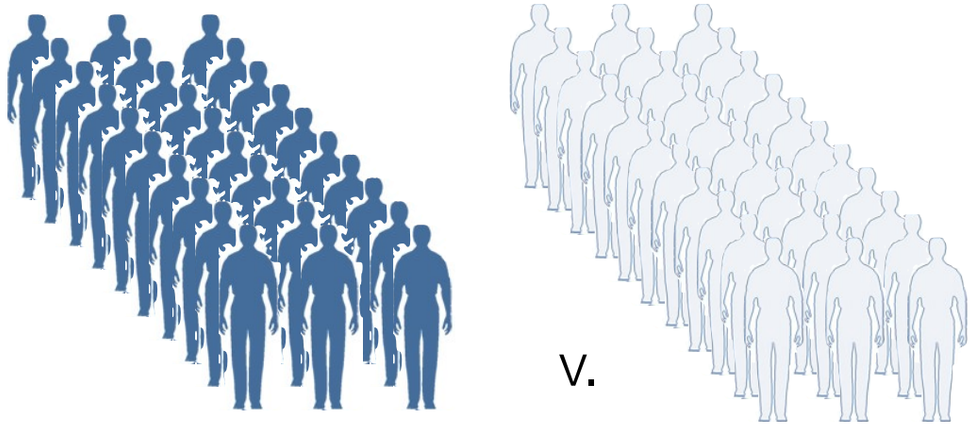
http://www.ebi.ac.uk/2can/disease/genes5.html

Loss of function mutations



MacArthur et al, Science. 2012 February 17; 335(6070): 823–828.

Interpreting relevance to disease.



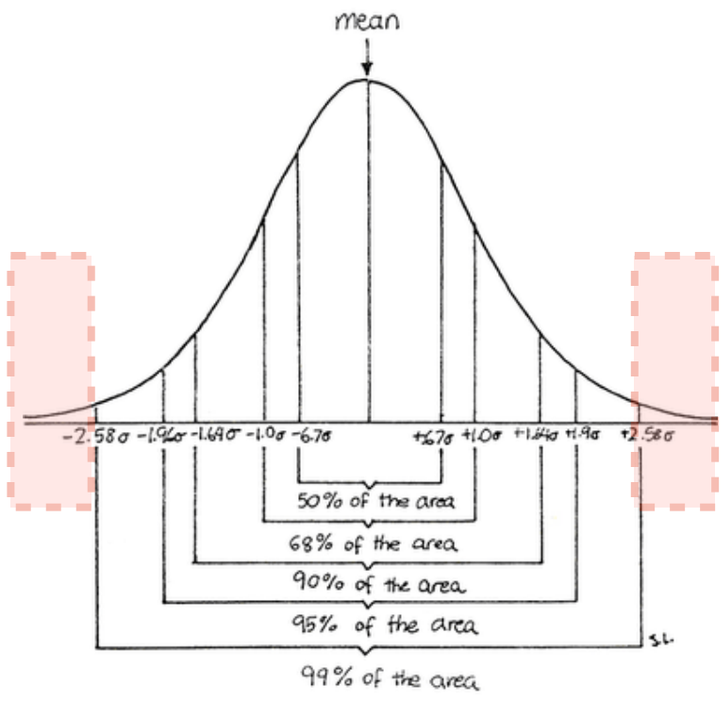
Cases
(have disease)

v.

Controls
(no disease)

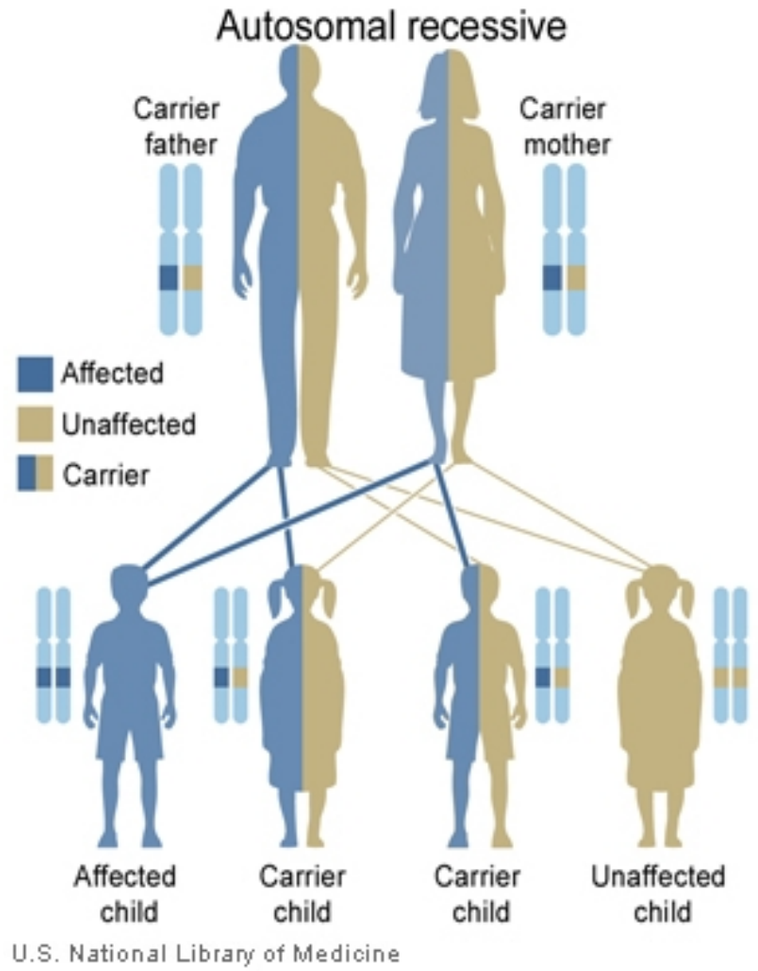
Complex diseases
(multiple genes contribute to risk)

e.g.,
Low LDL



e.g.,
High LDL

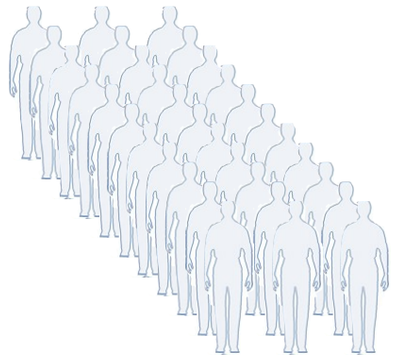
Sampling individuals at extremes of trait distribution



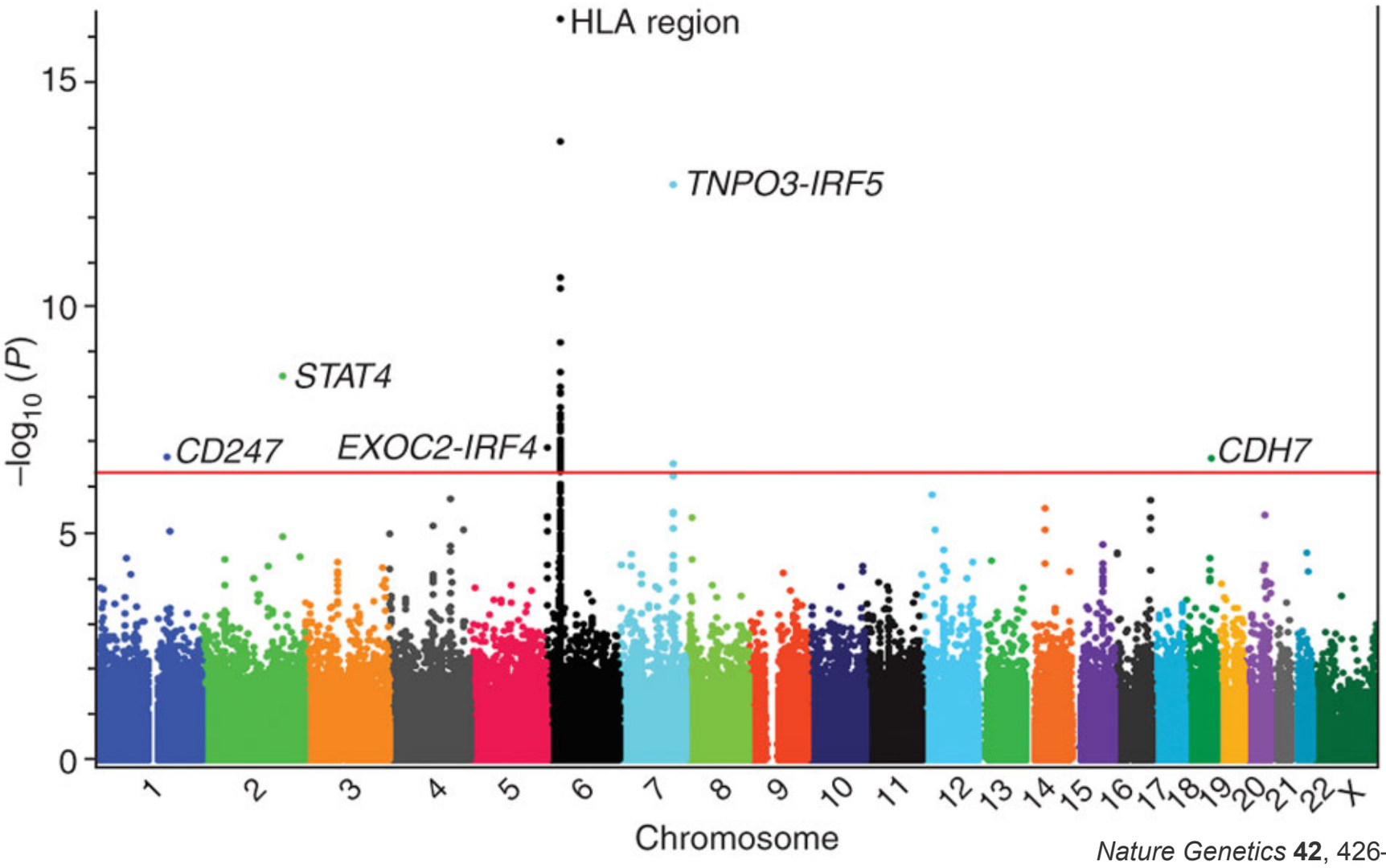
Single-gene diseases

Genome-wide association studies

Cases
(have disease)

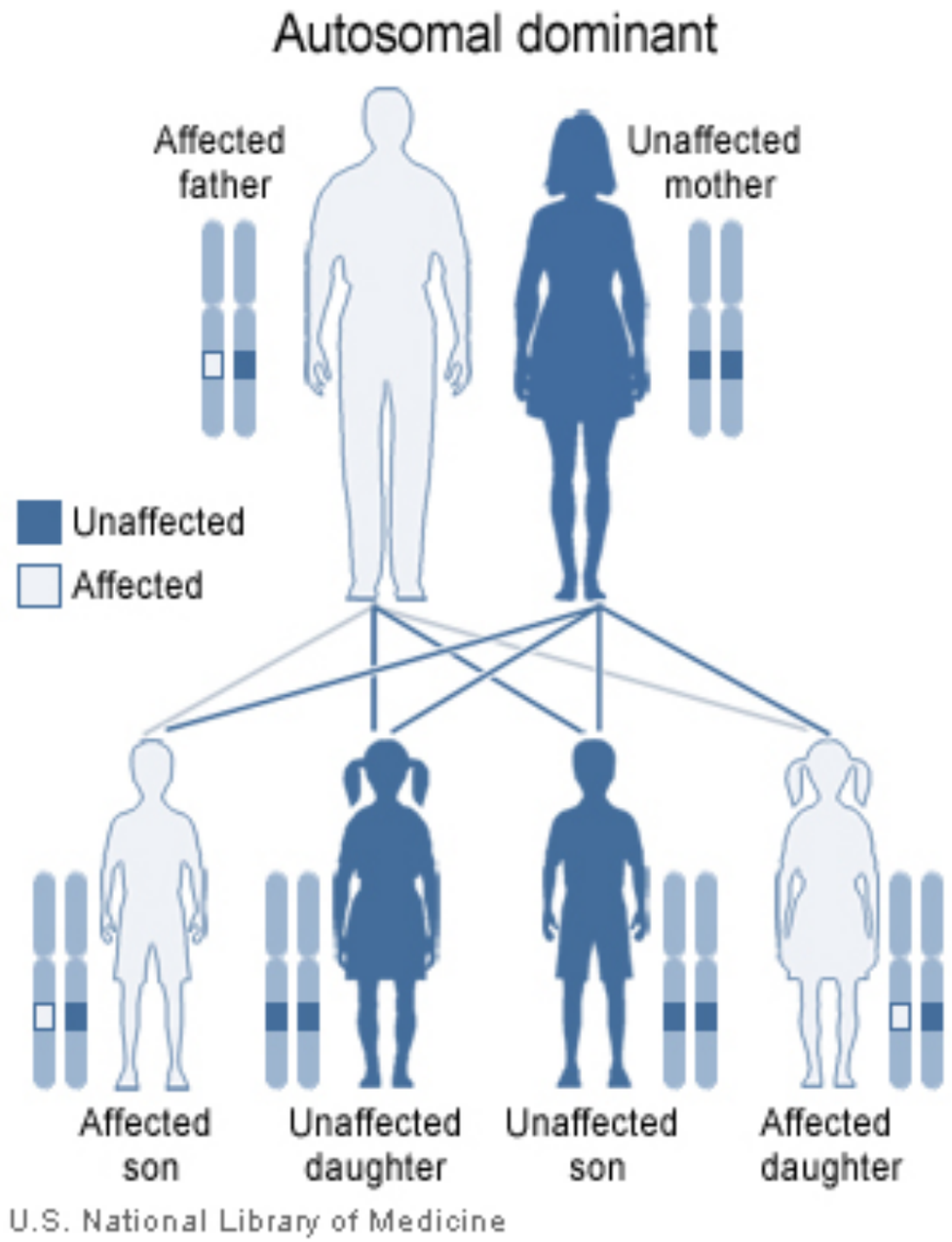


Controls
(no disease)

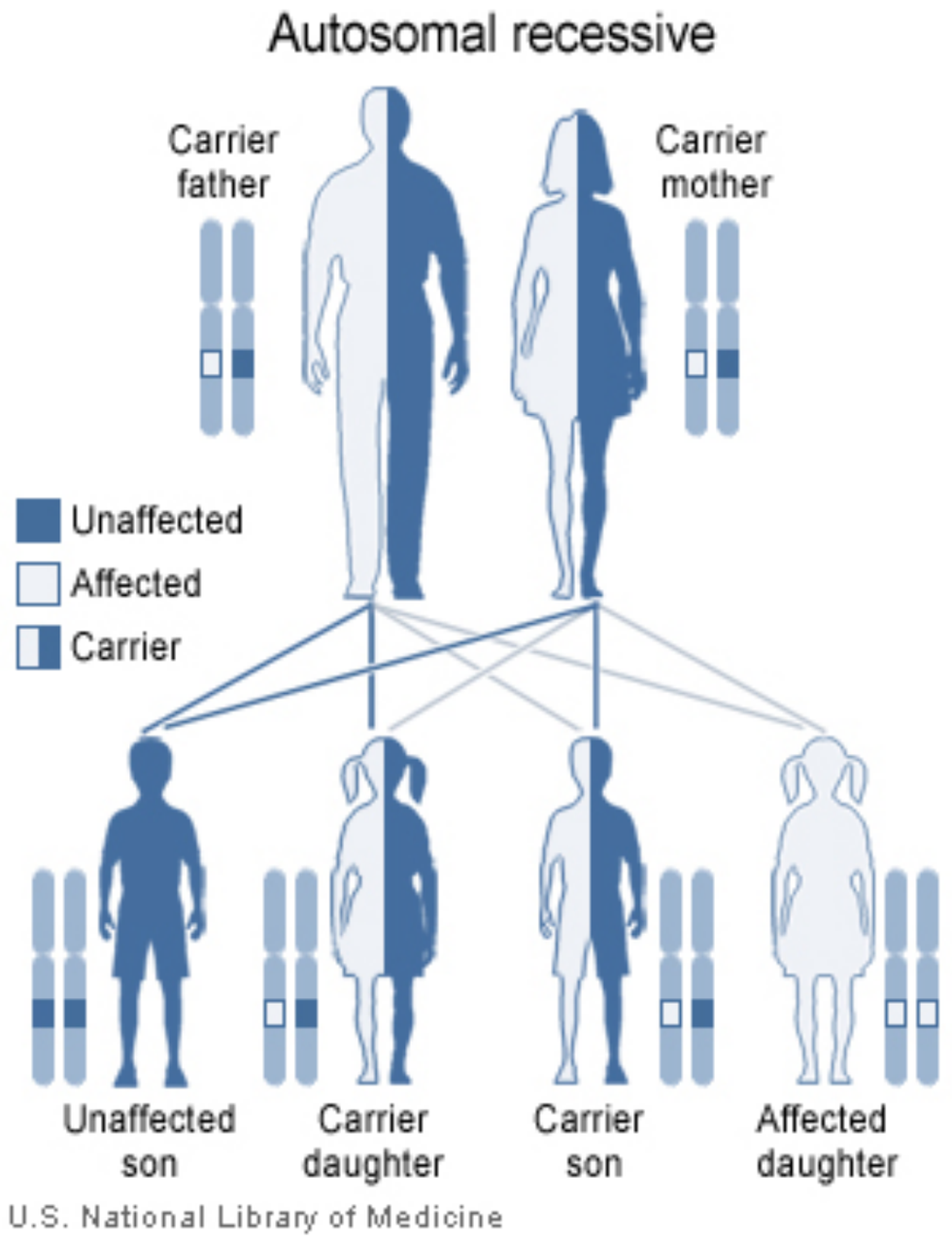


Nature Genetics **42**, 426–429 (2010) doi:10.1038/ng.565

Single gene diseases (autosomal)

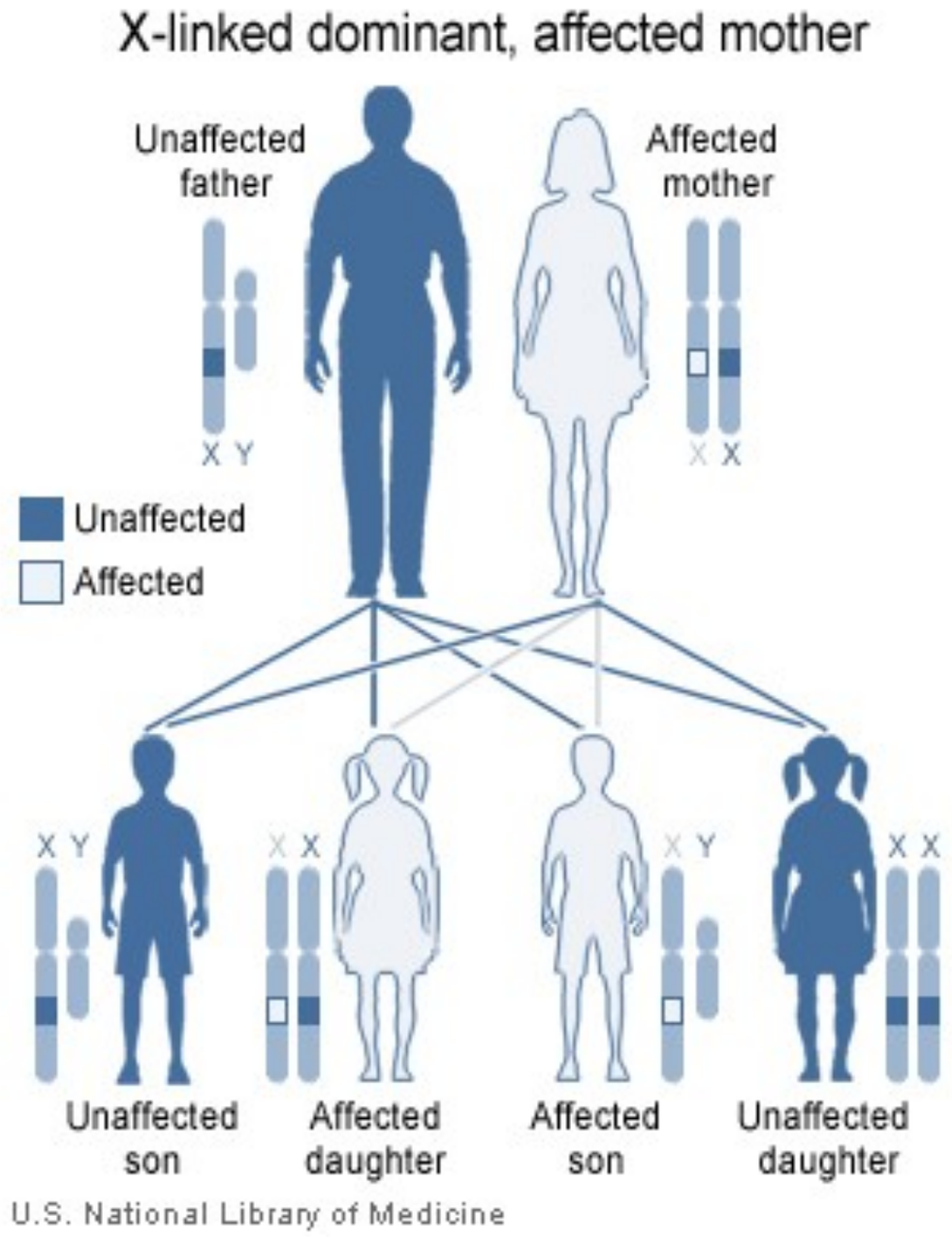


e.g., Huntington's disease

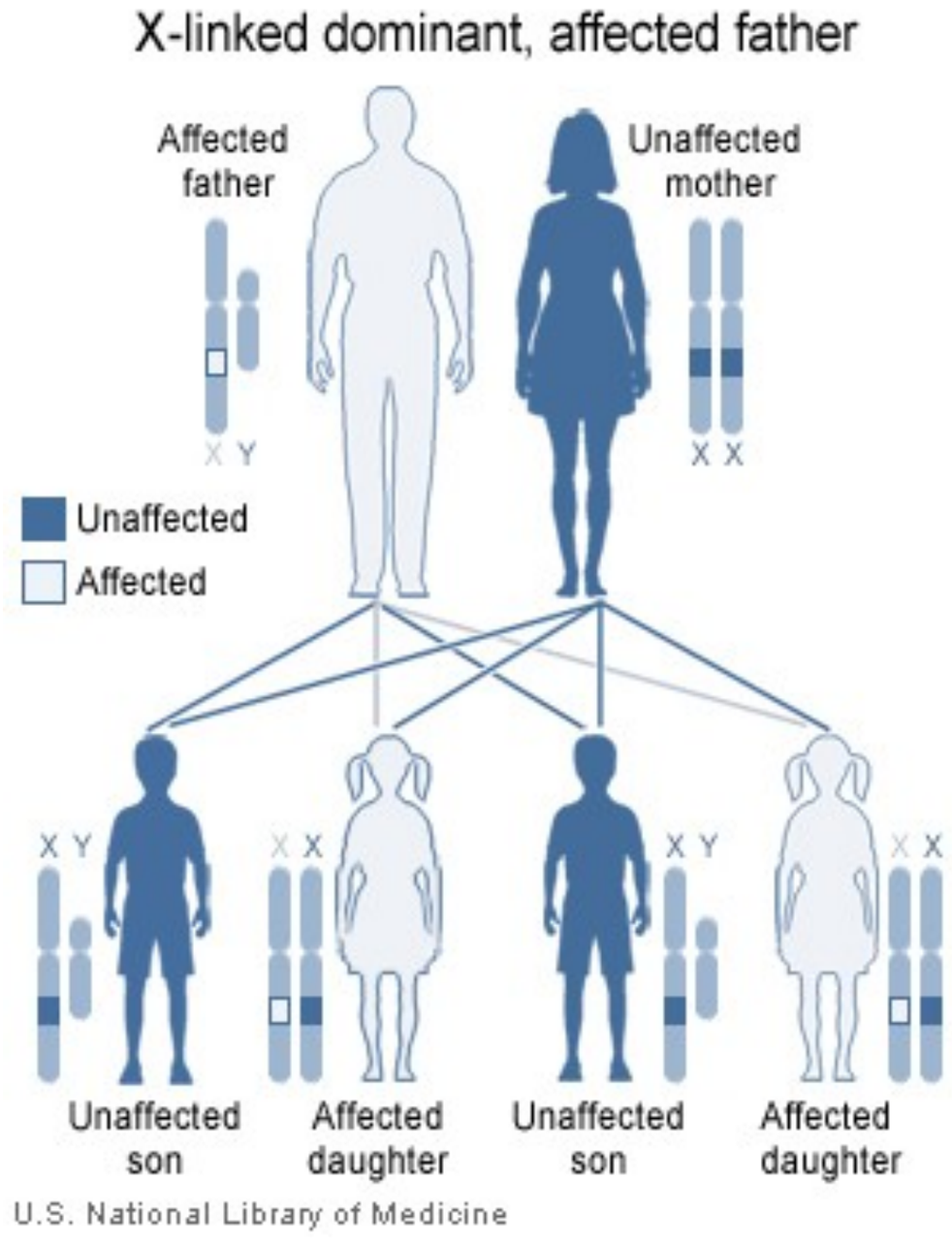


e.g., cystic fibrosis

Single gene diseases (X-linked dominant)



Both sons and daughters can be affected.



Affected dads cannot pass on disease to sons

Single gene diseases (X-linked recessive)