The use of genomics to understand human disease

Jonathan Pevsner, Ph.D.
Kennedy Krieger Institute

October25, 2016
eScience IEEE 2016

Outline

Introduction to genomics and human disease

Identifying a mutation causing a disease: Sturge-Weber

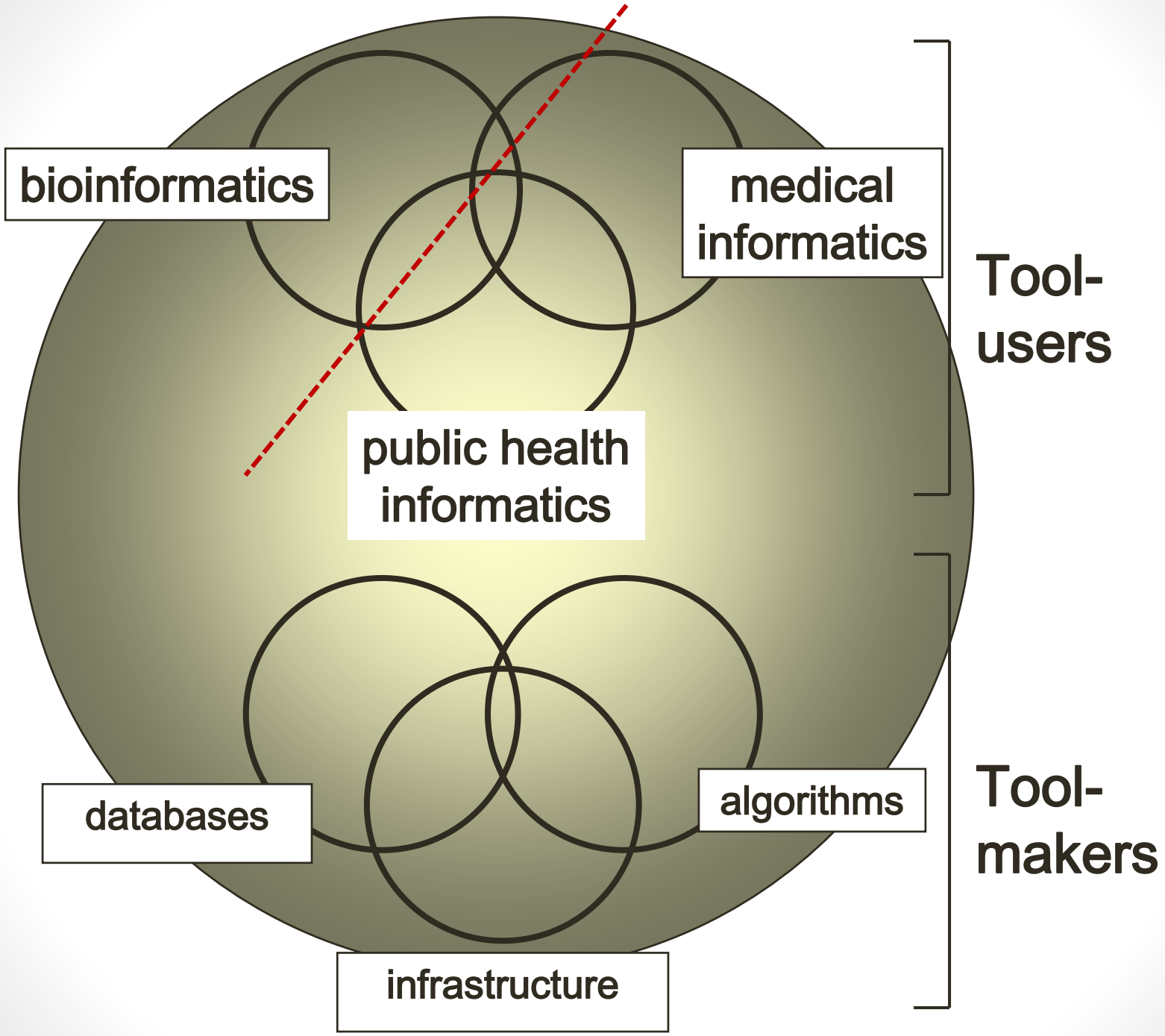
Genomic variation in autism spectrum disorder

Definitions of bioinformatics and genomics

- Bioinformatics is the interface of molecular biology and computer science.

- It is the analysis of proteins, genes and genomes using computer algorithms and computer databases.

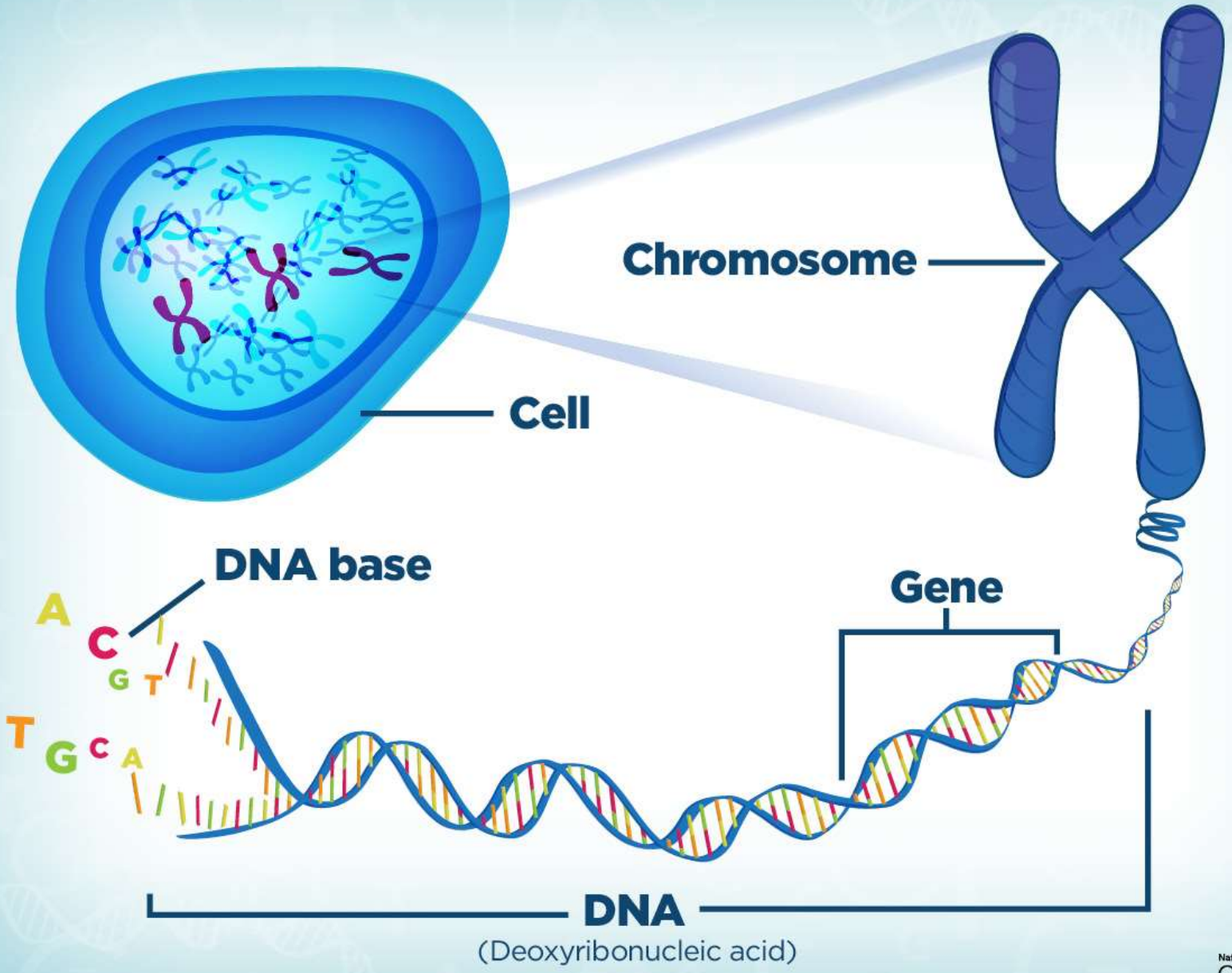
- Genomics is the analysis of genomes. The tools of bioinformatics are used to make sense of the quintillions of base pairs of DNA that are sequenced by genomics projects.


bioinformatics
medical informatics
public health informatics
databases
algorithms
infrastructure
Tool-users
Tool-makers

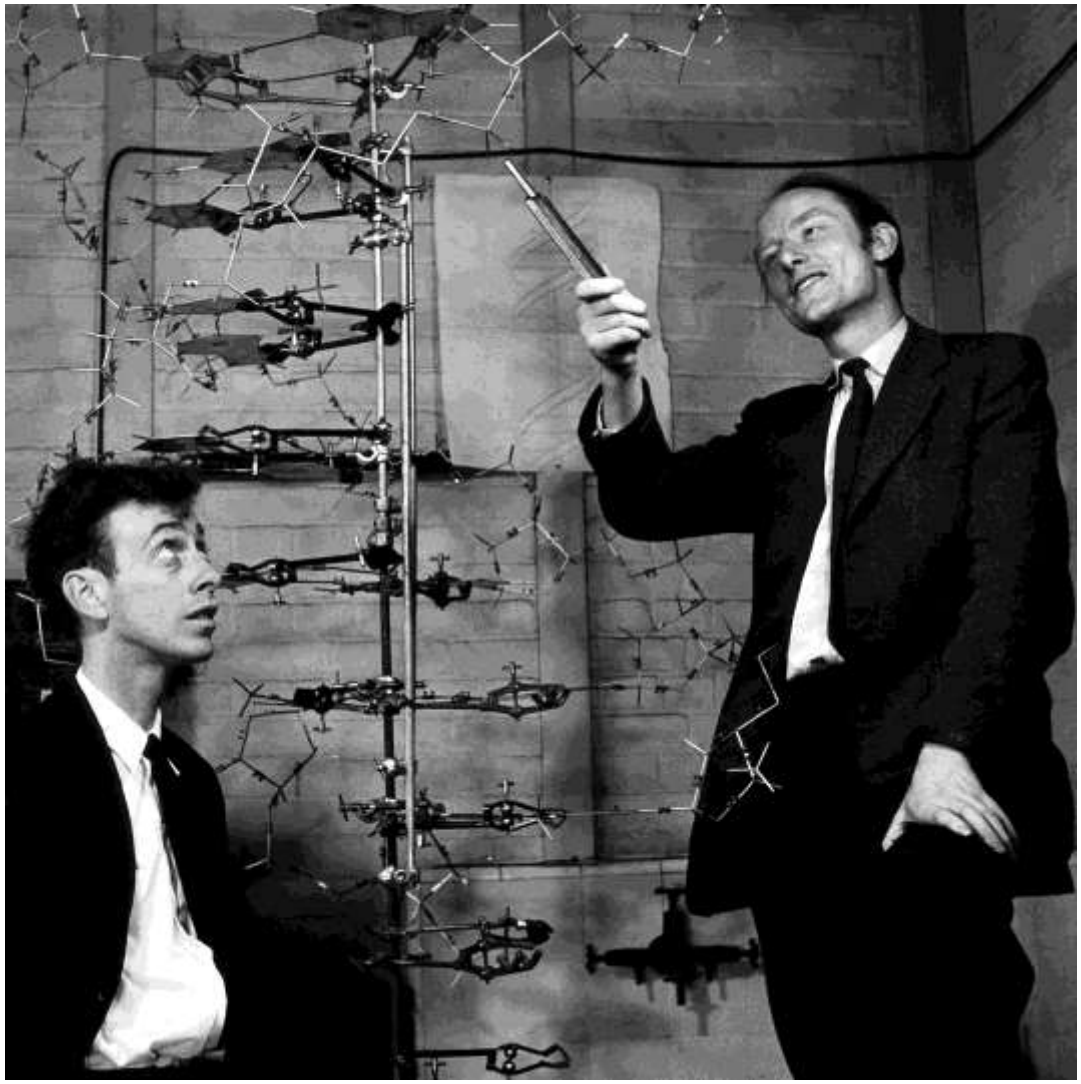
Genomes and genes

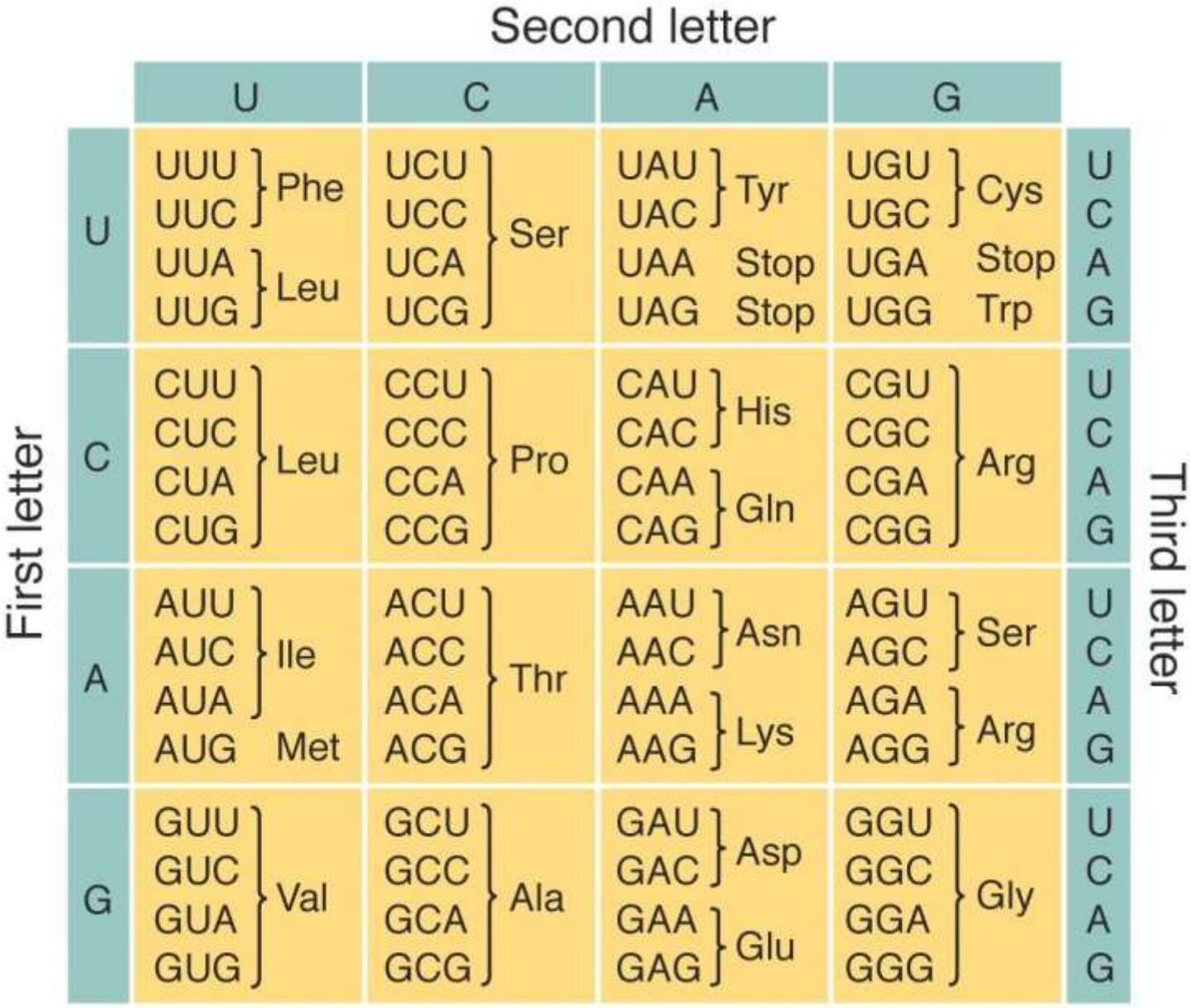
A genome is the collection of DNA that comprises an individual. The human genome is organized into 23 pairs of chromosomes (1-22, XX for girls, XY for boys).

Gene: Classically, a unit of hereditary information localized to a particular chromosome position and encoding one protein. It is a DNA sequence that makes RNA and that often then makes protein.

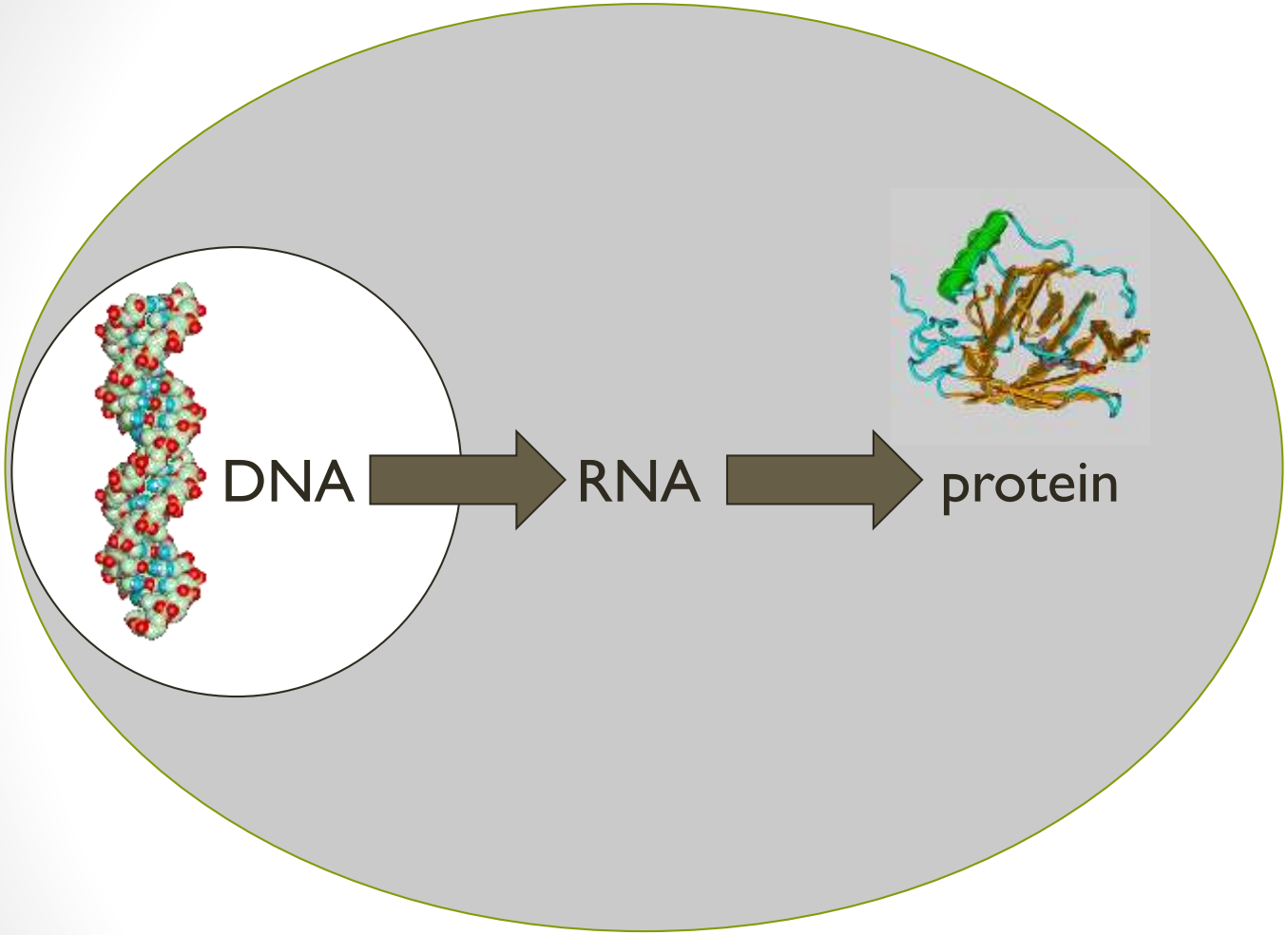


National Society of Genetic Counselors

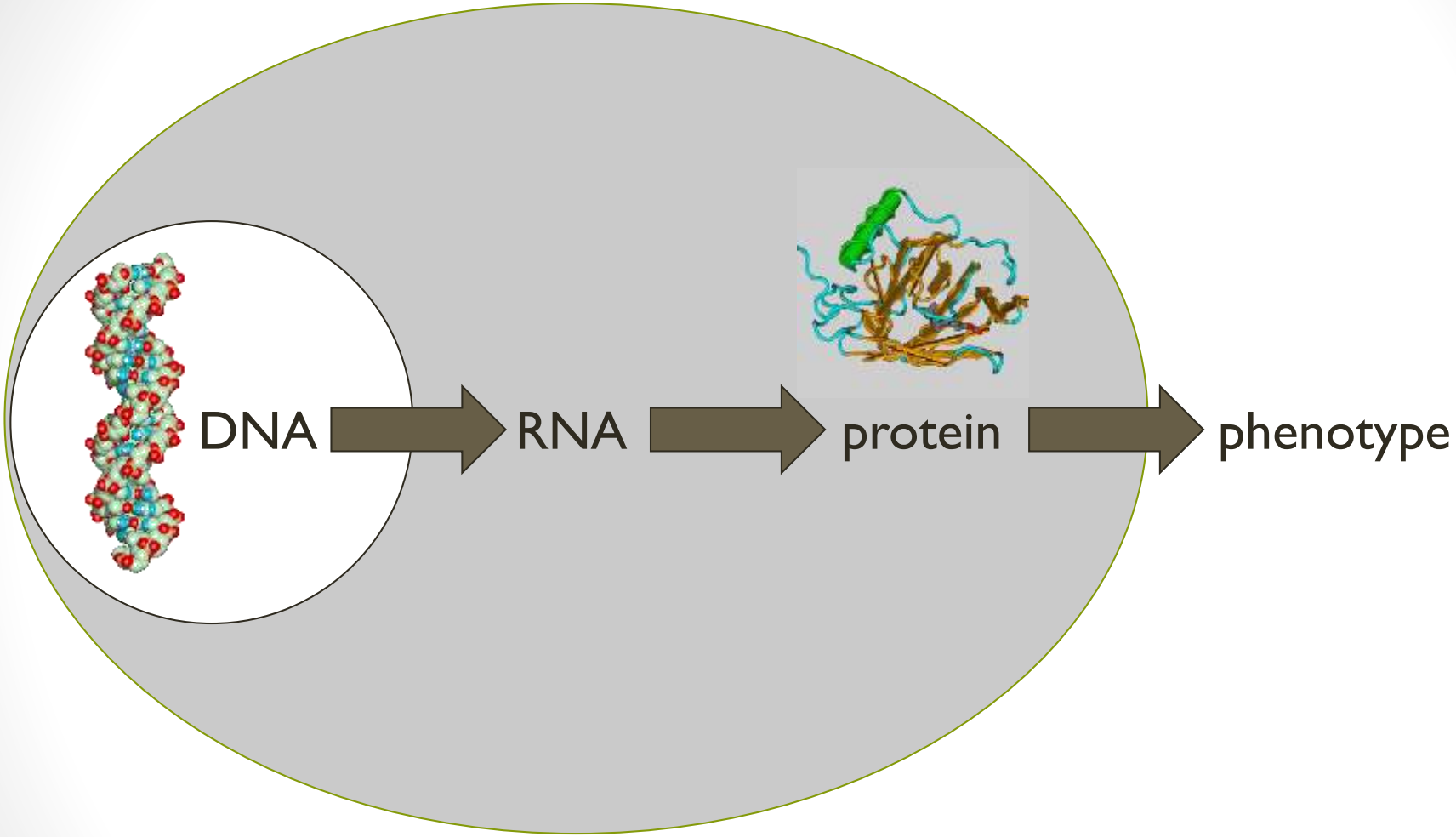



Second letter
U
C
A
G
First letter
Third letter
U
UUU Phe
UUC
UUA Leu
UUG
UCU Ser
UCC
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
C
CUU Leu
CUC
CUA
CUG
CCU Pro
CCC
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU Arg
CGC
CGA
CGG
A
AUU Ile
AUC
AUA
AUG Met
ACU Thr
ACC
ACA
ACG
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA Arg
AGG
G
GUU Val
GUC
GUA
GUG
GCU Ala
GCC
GCA
GCG
GAU Asp
GAC
GAA Glu
GAG
GGU Gly
GGC
GGA
GGG

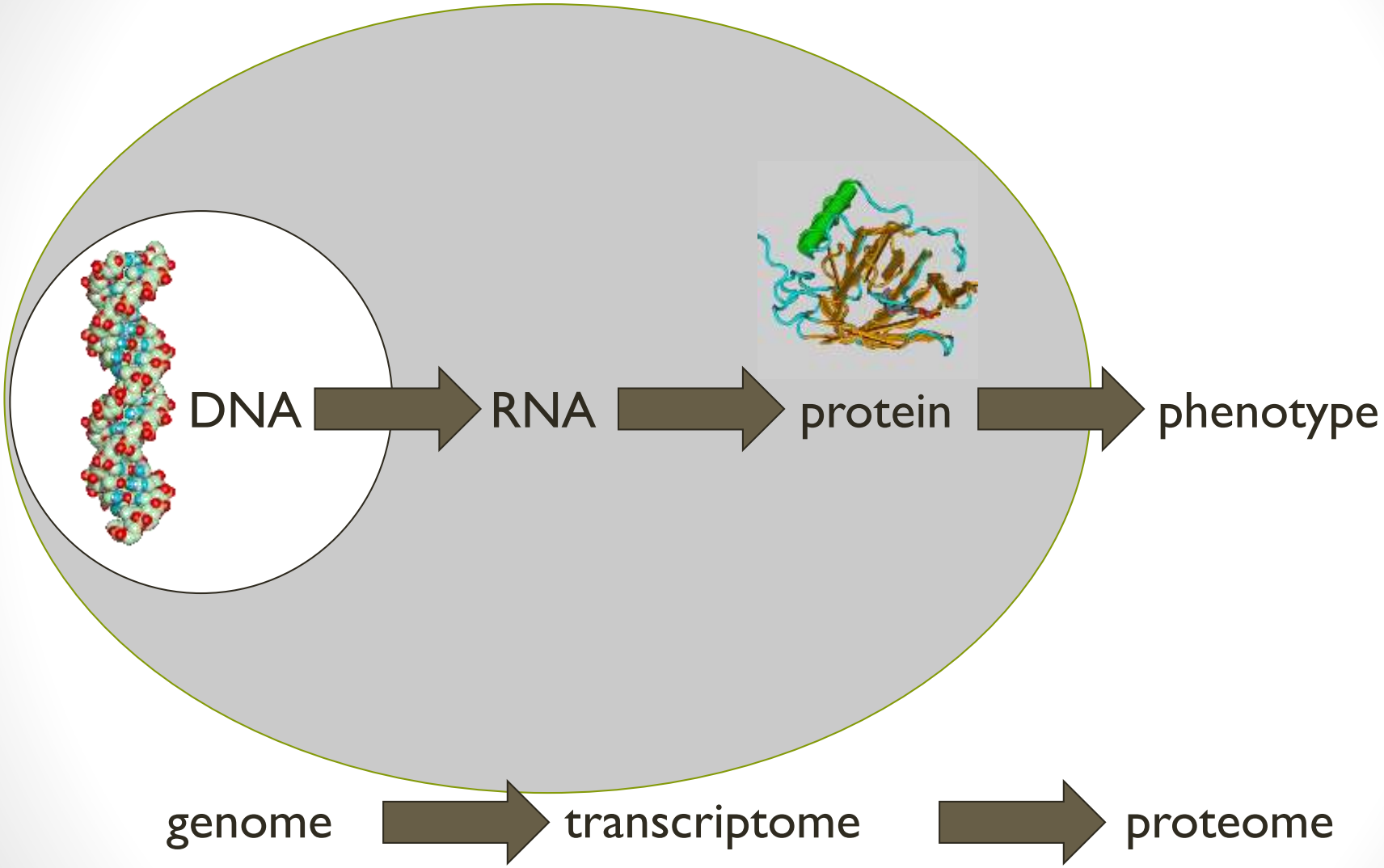
Central dogma of bioinformatics and genomics



Central dogma of bioinformatics and genomics

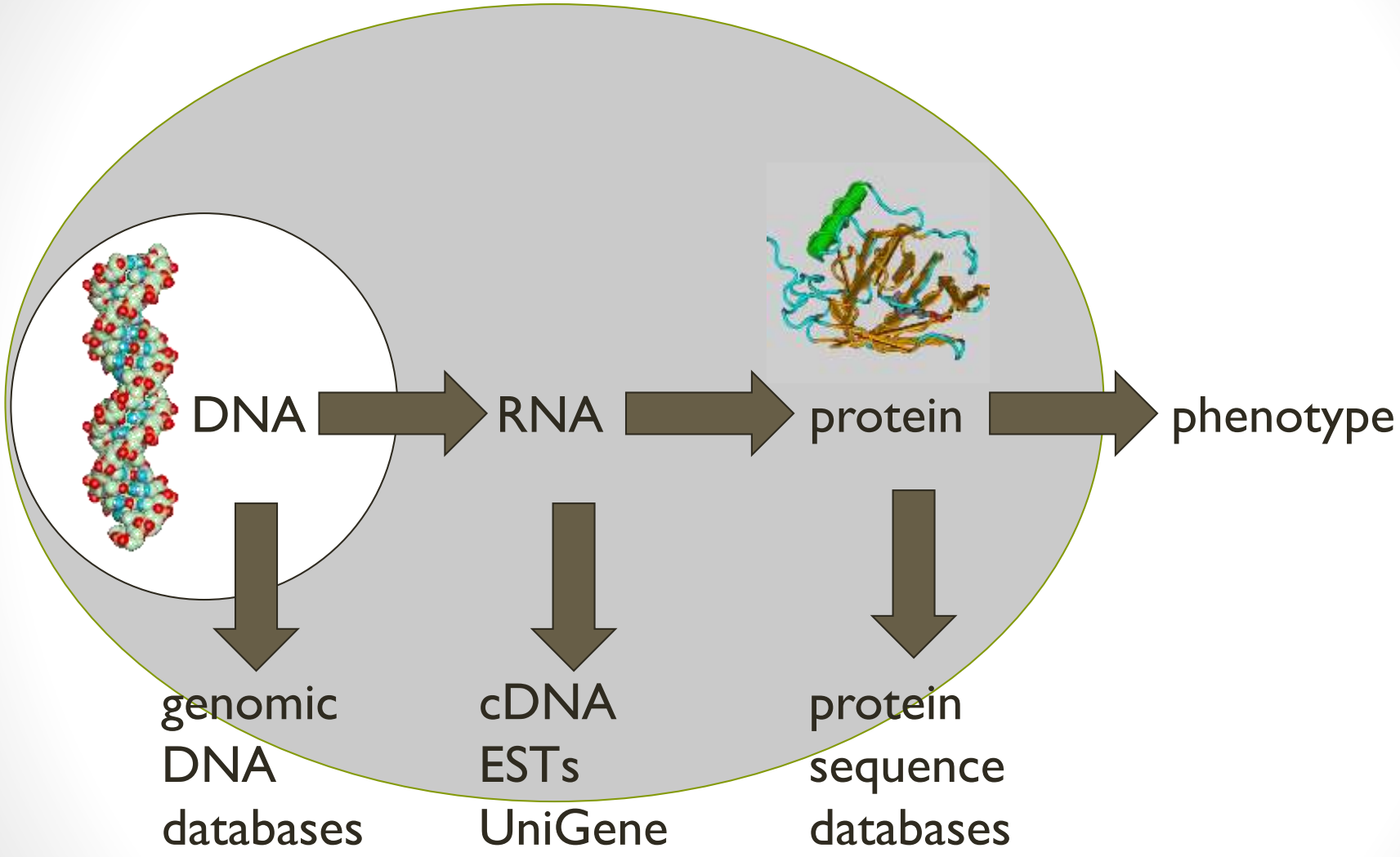


Central dogma of bioinformatics and genomics



Central dogma of bioinformatics and genomics

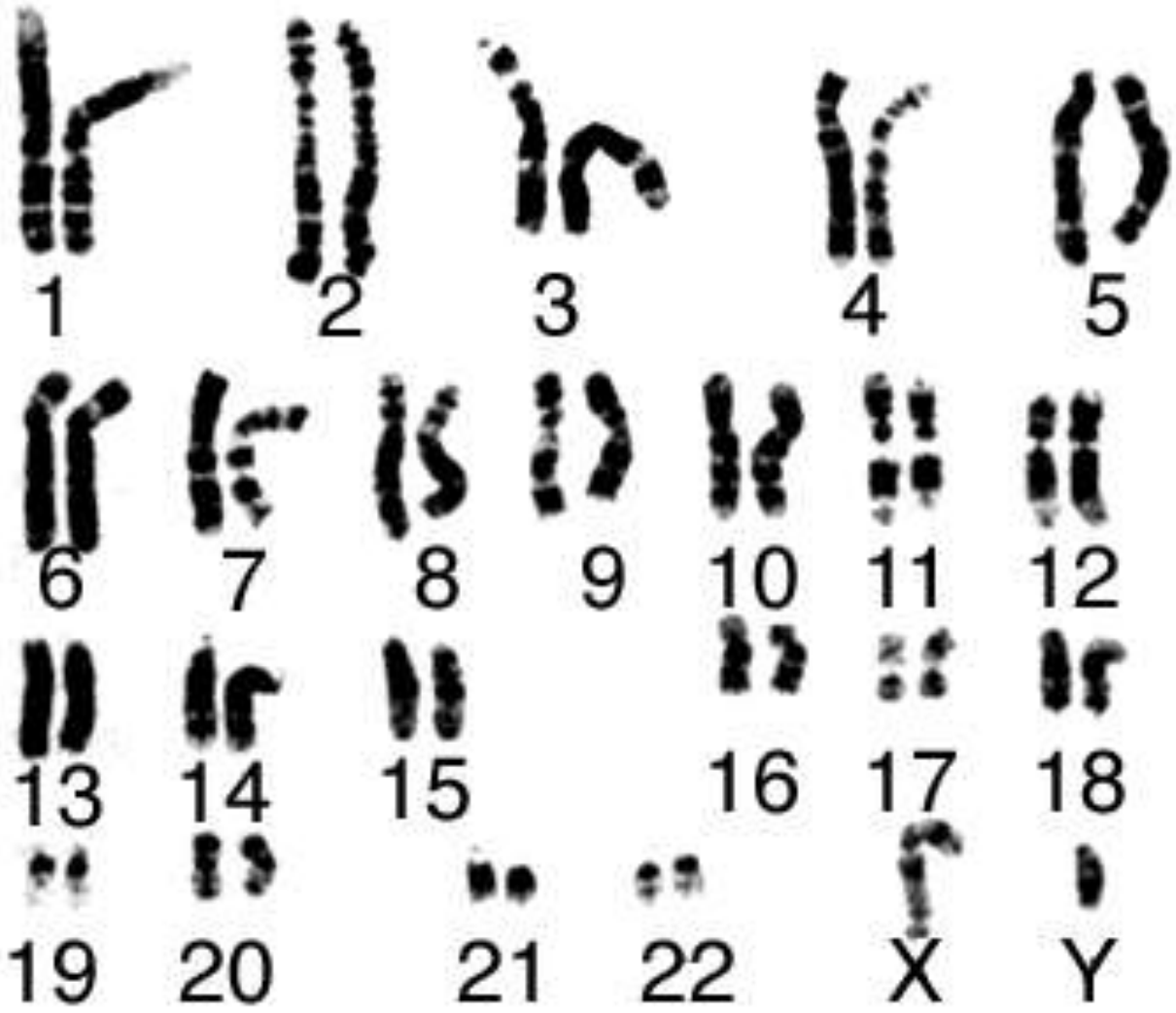
Central dogma of bioinformatics and genomics



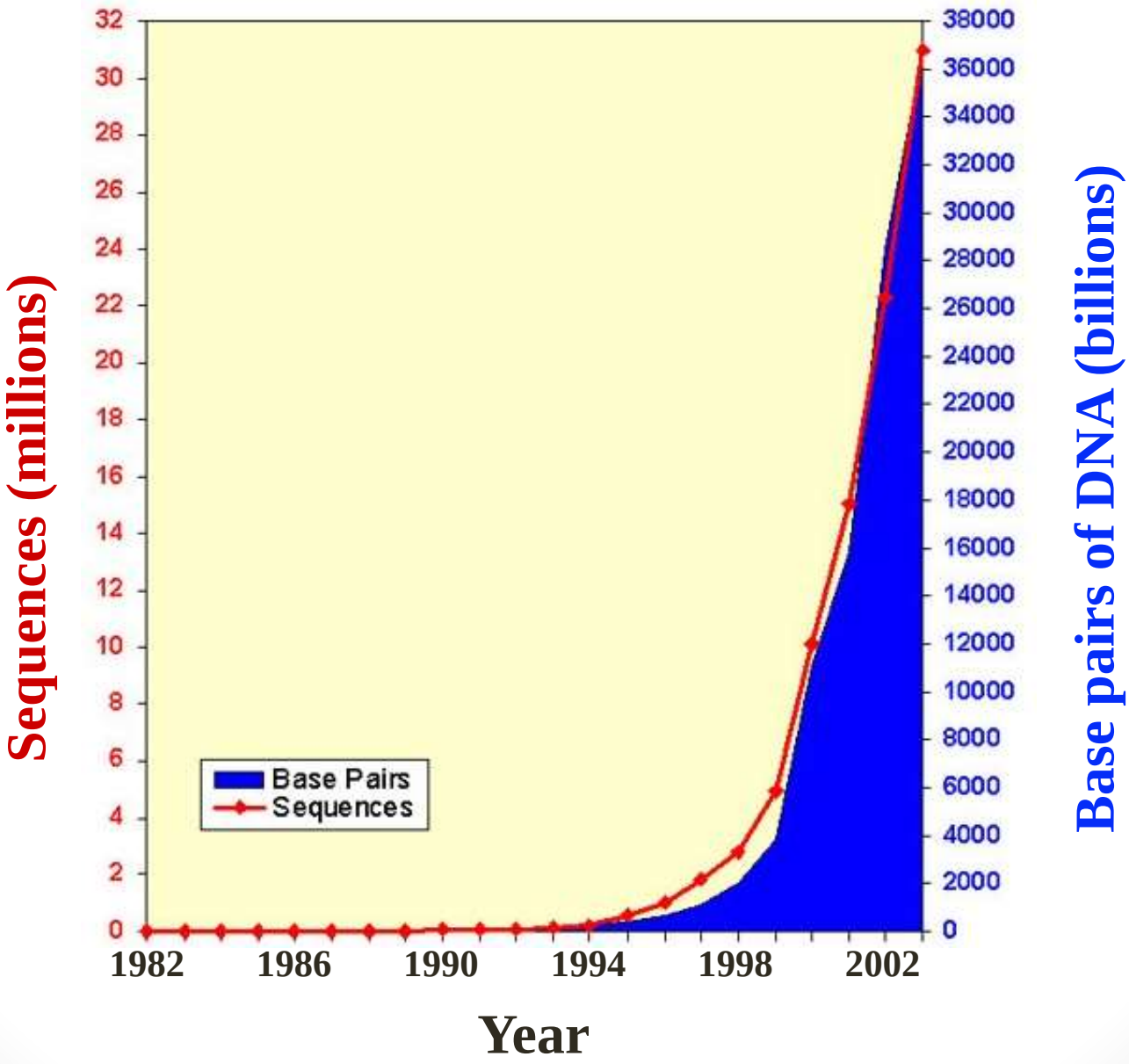
Genes are expressed at different times and places

Time of development

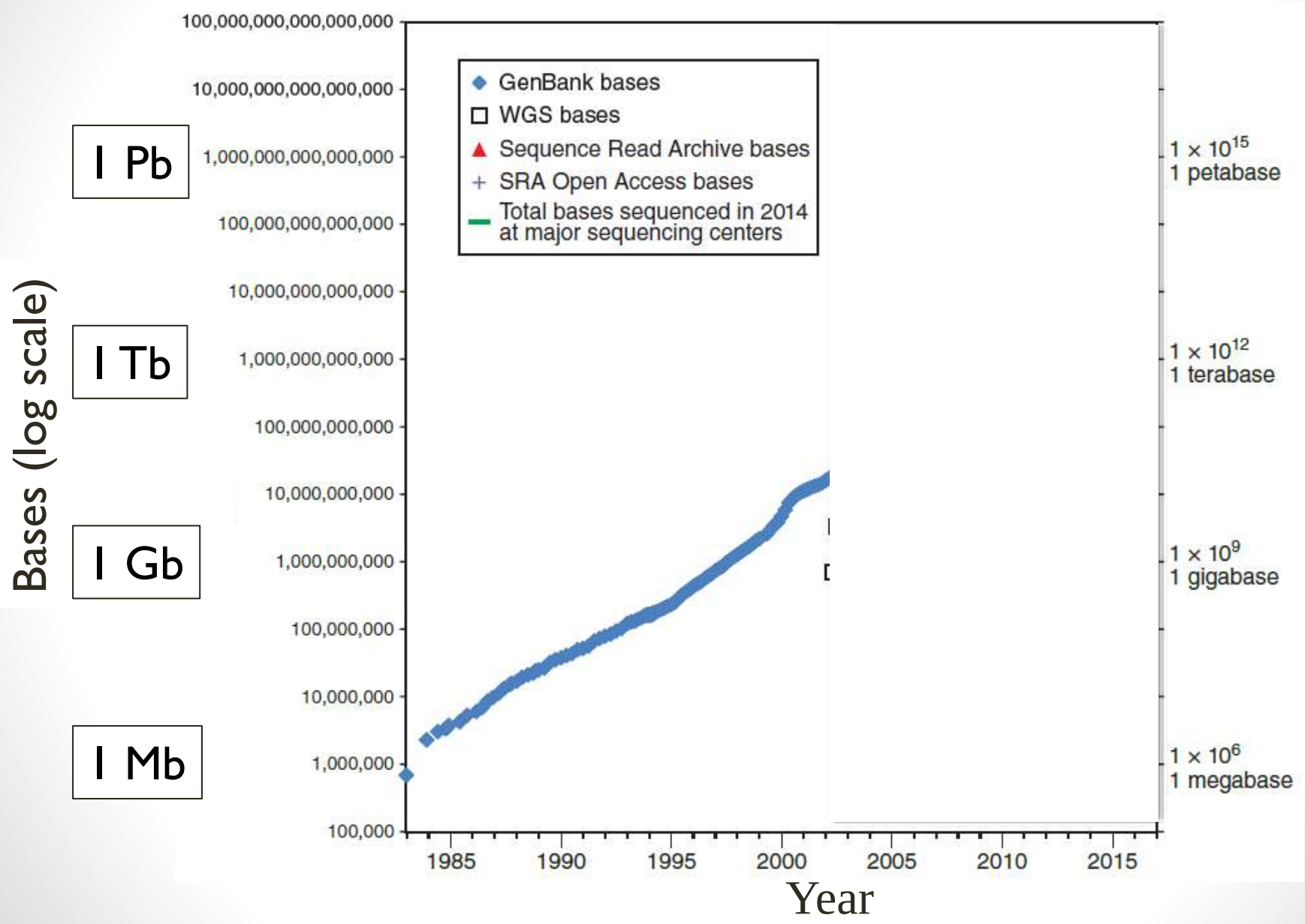
Body region, physiology, pharmacology, pathology


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

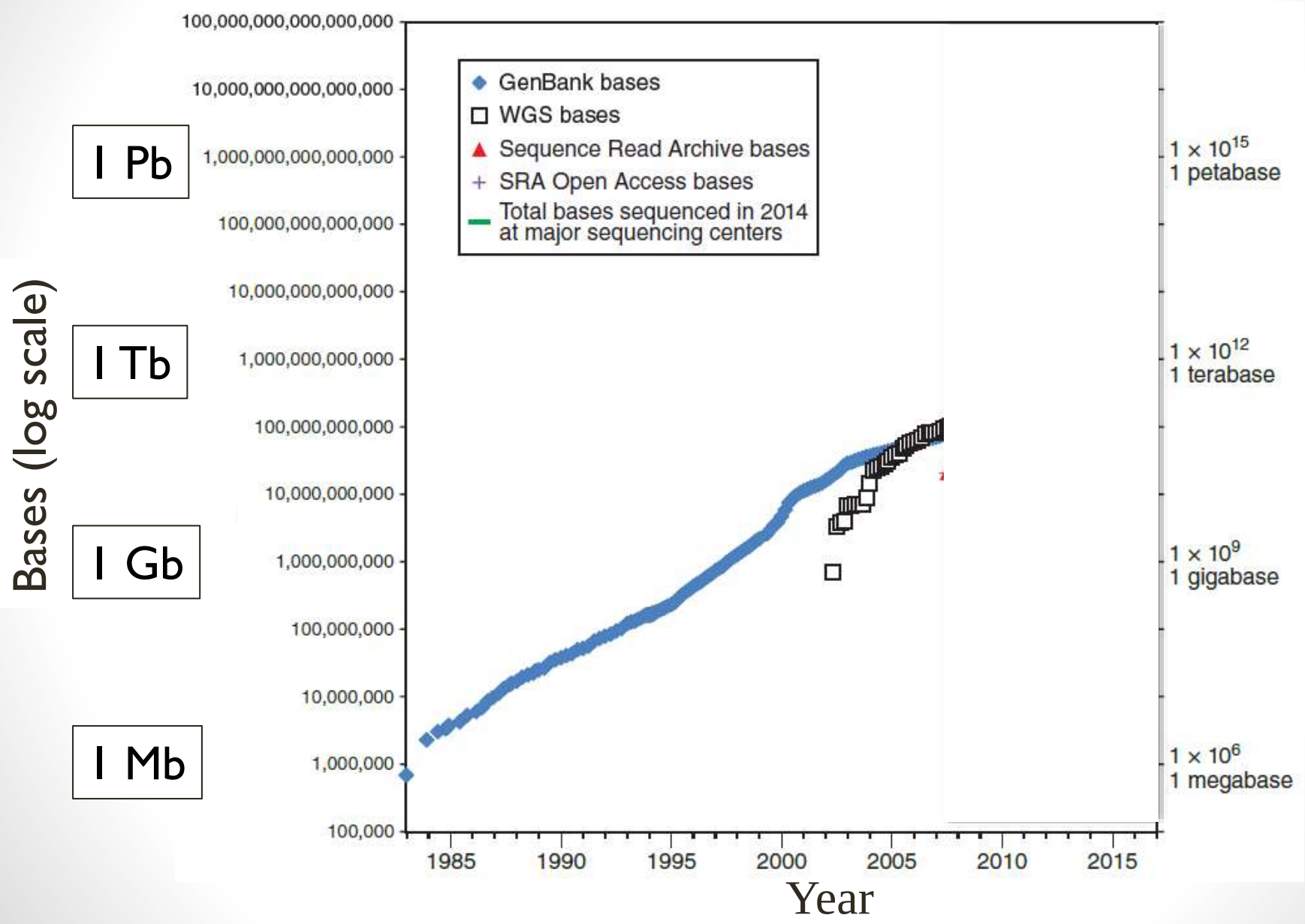
Growth of GenBank



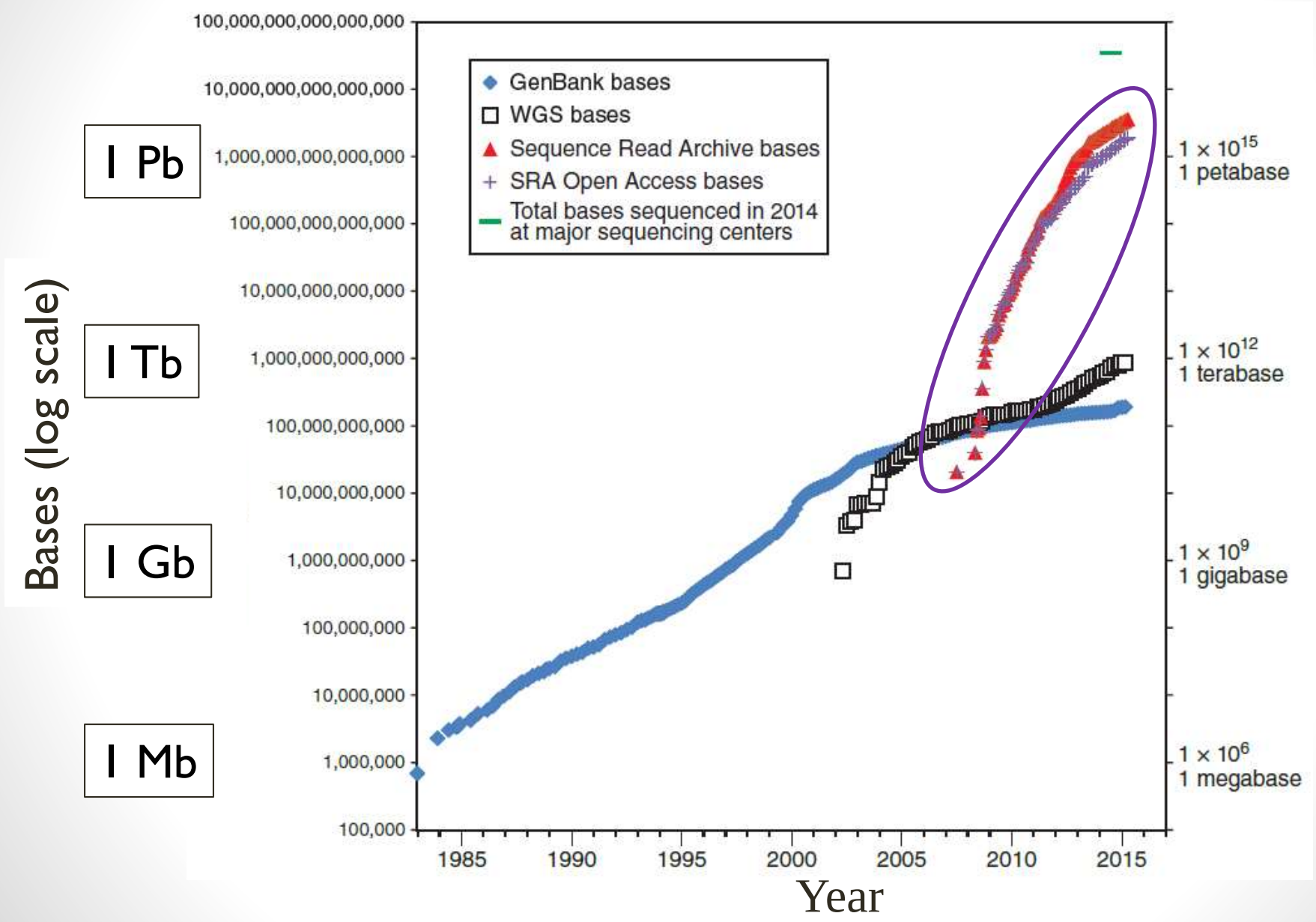
Growth of DNA sequence in repositories



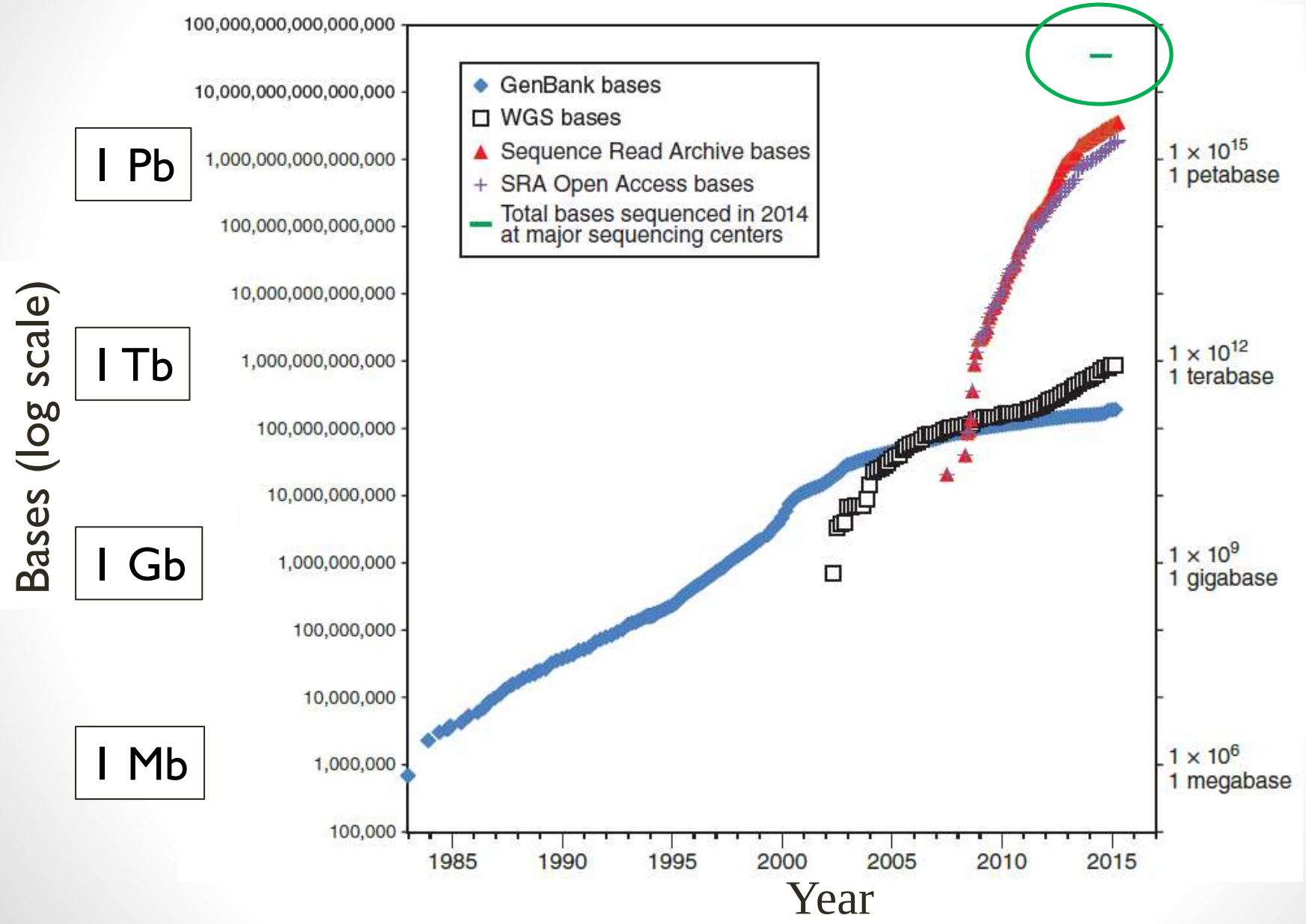
Growth of DNA sequence in repositories

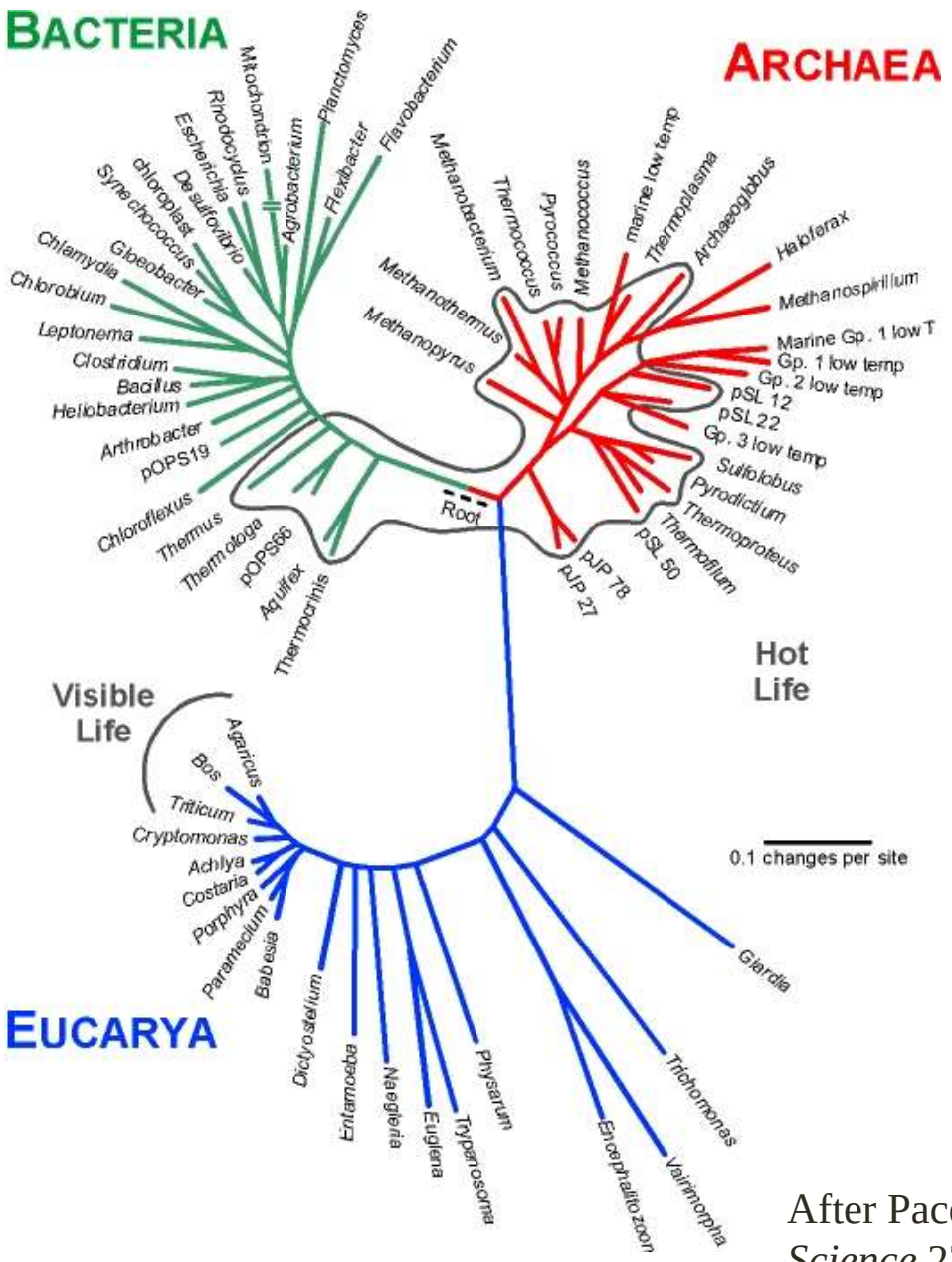


Growth of DNA sequence in repositories



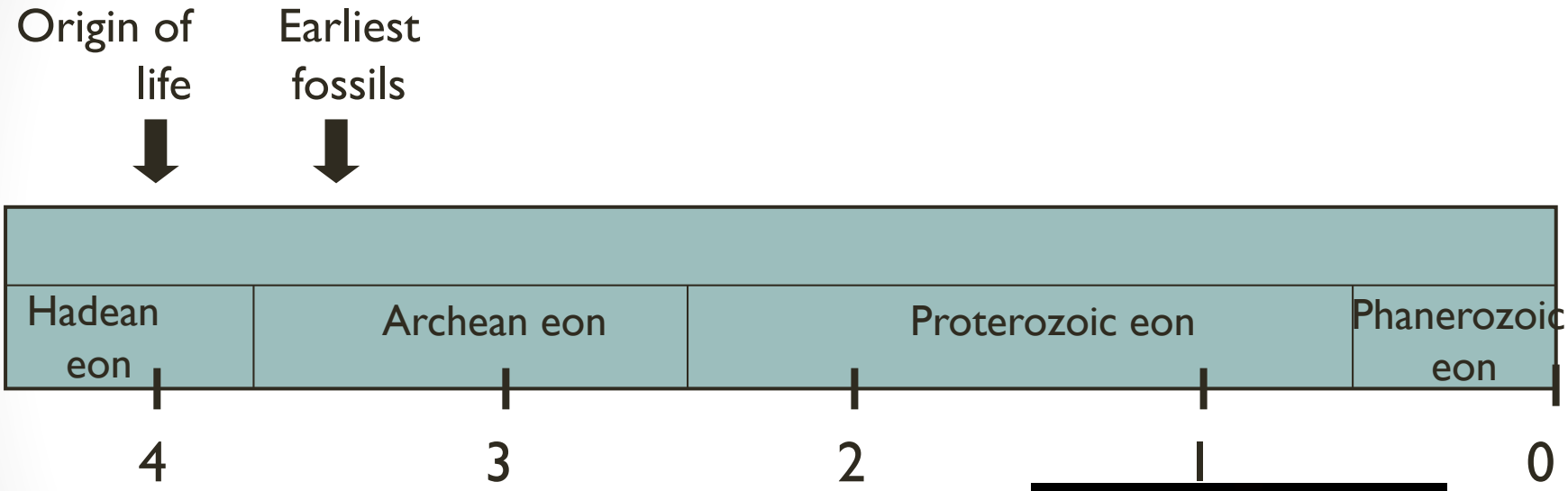
Growth of DNA sequence in repositories



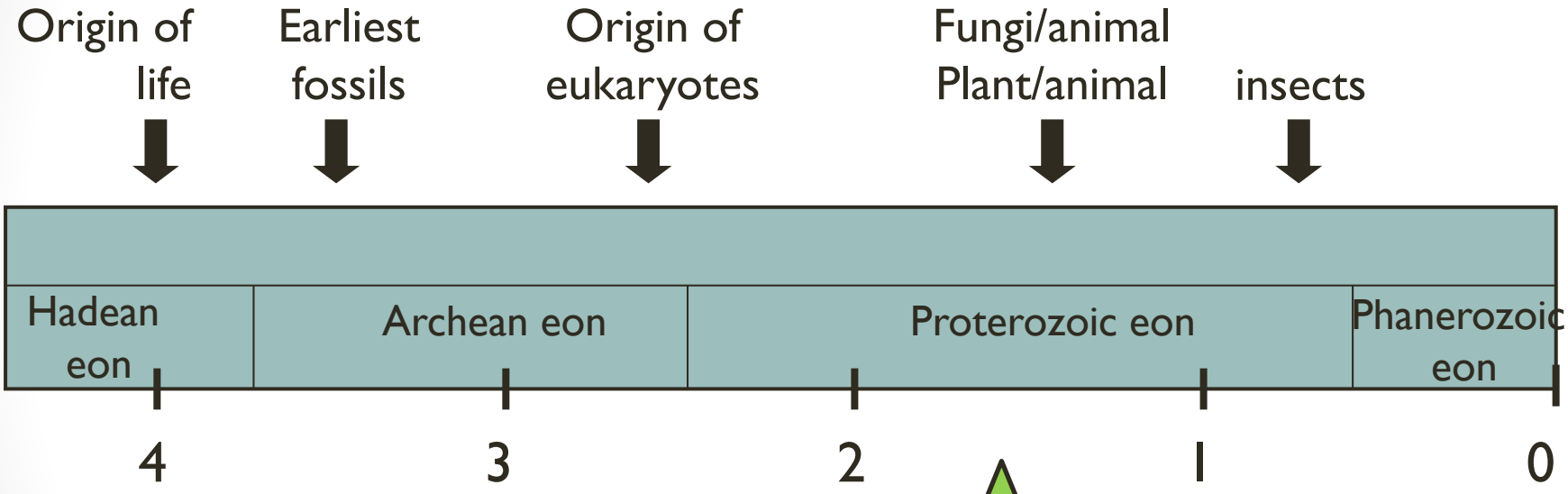


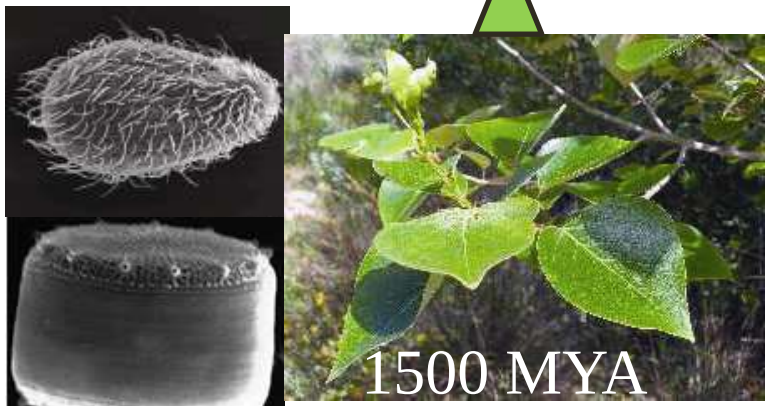
After Pace NR (1997)
Science 276:734

Billions of years ago (BYA)

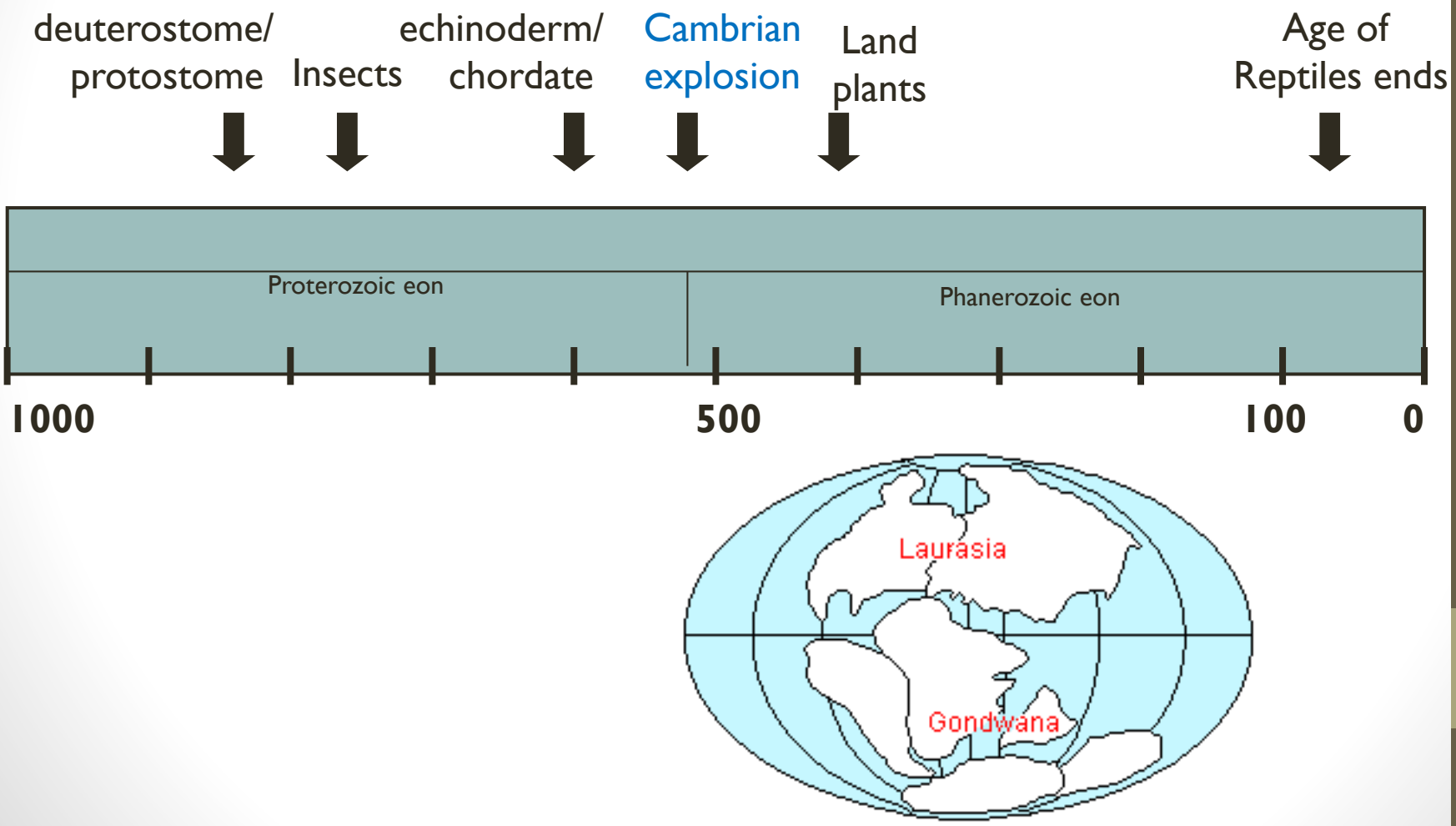


Billions of years ago (BYA)

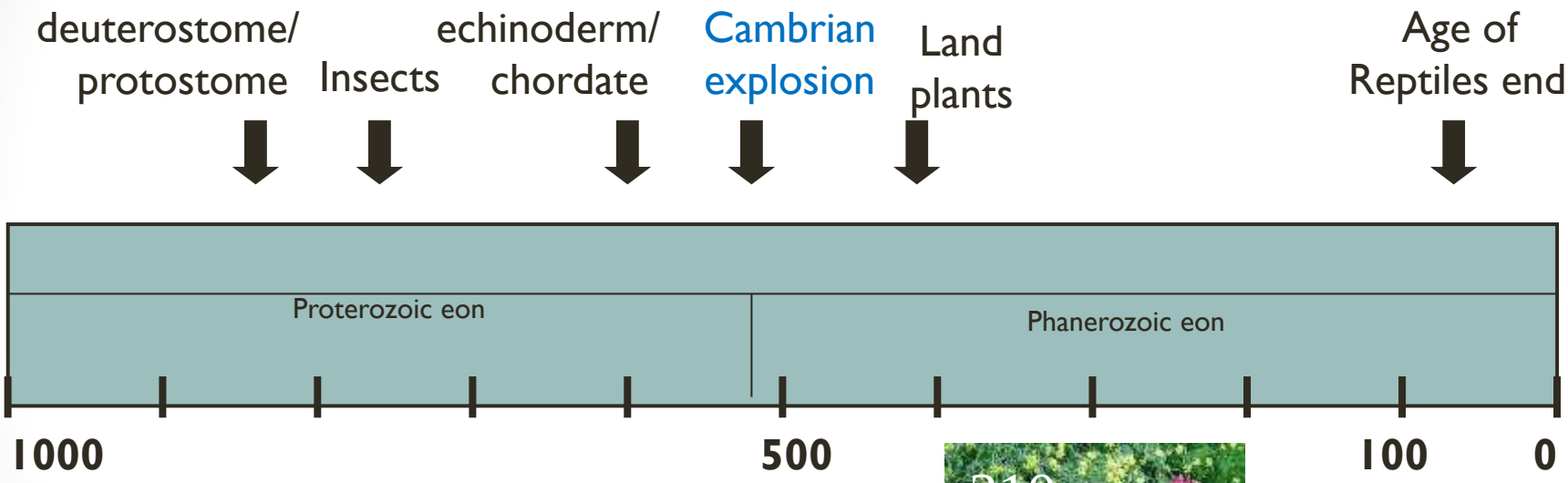


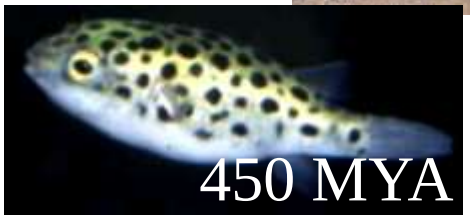


Millions of years ago (MYA)

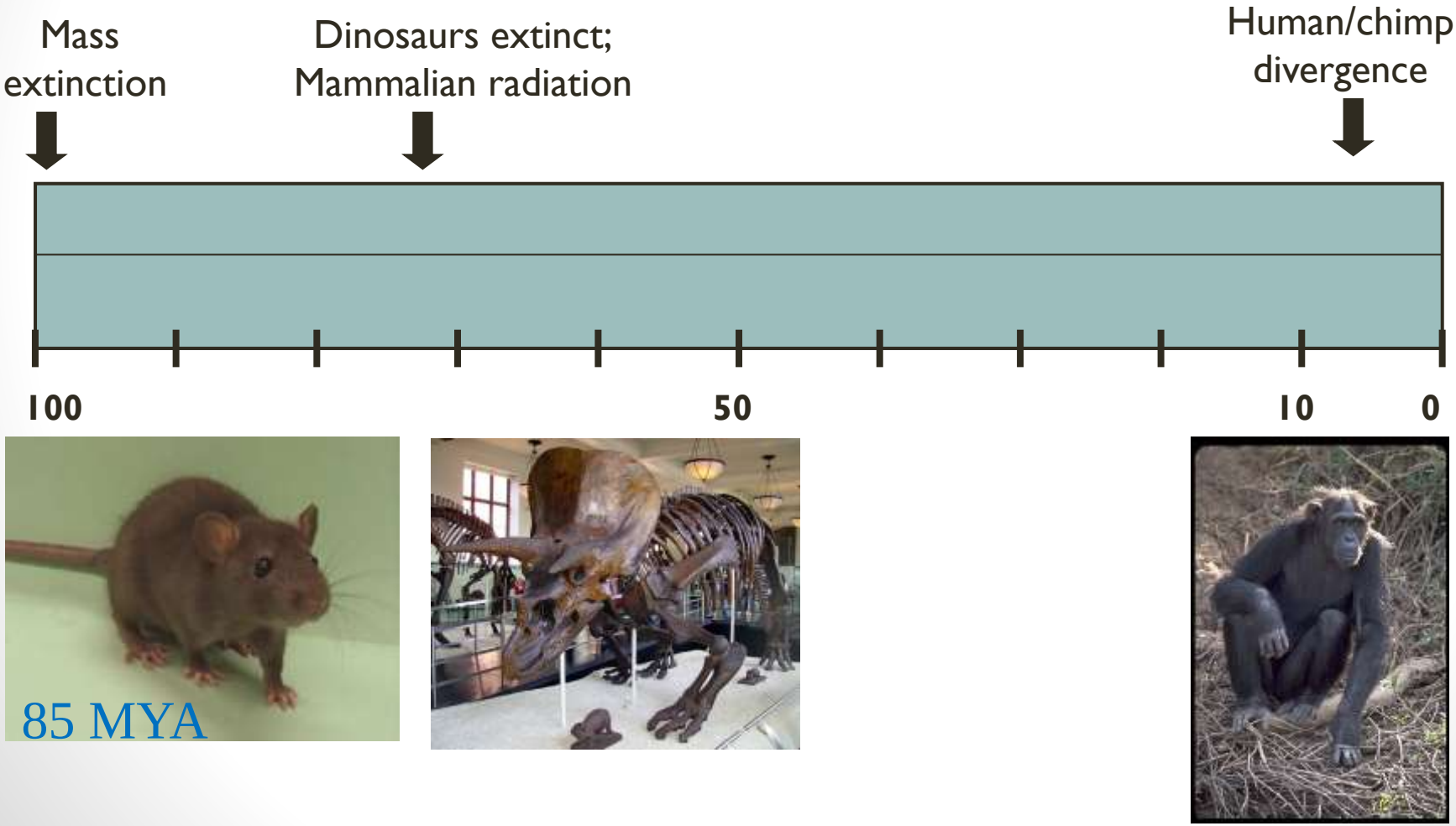


Millions of years ago (MYA)

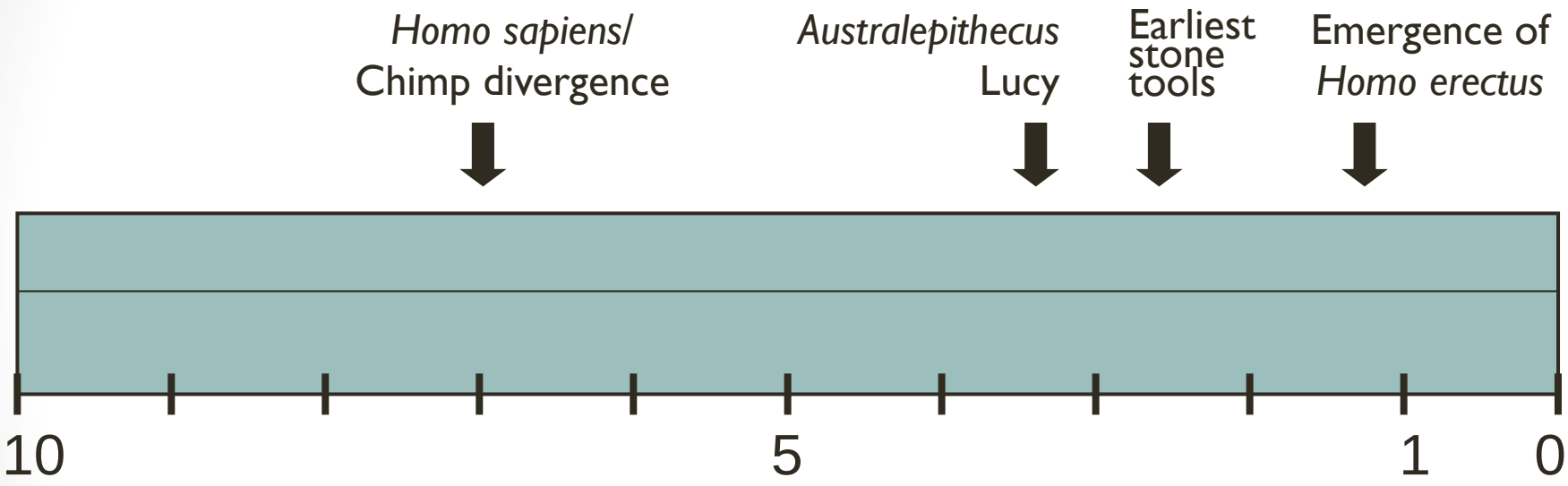




Millions of years ago (MYA)



Millions of years ago (MYA)



Years ago

Homo erectus emerges in Africa

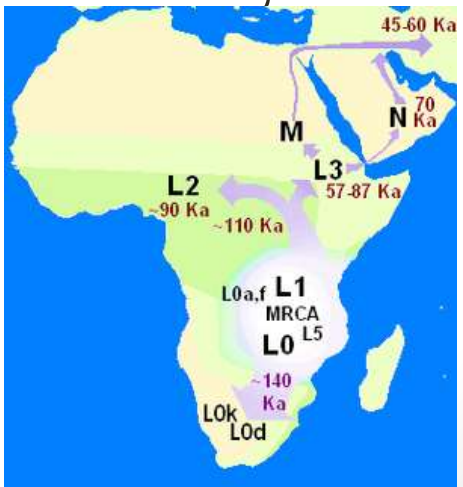
Mitochondrial Eve

1,000,000

500,000

100,000

0

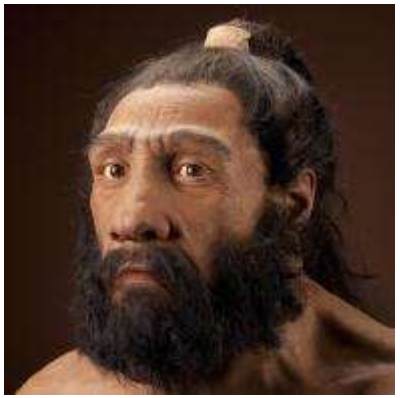


Years ago

Emergence of anatomically modern *H. sapiens*

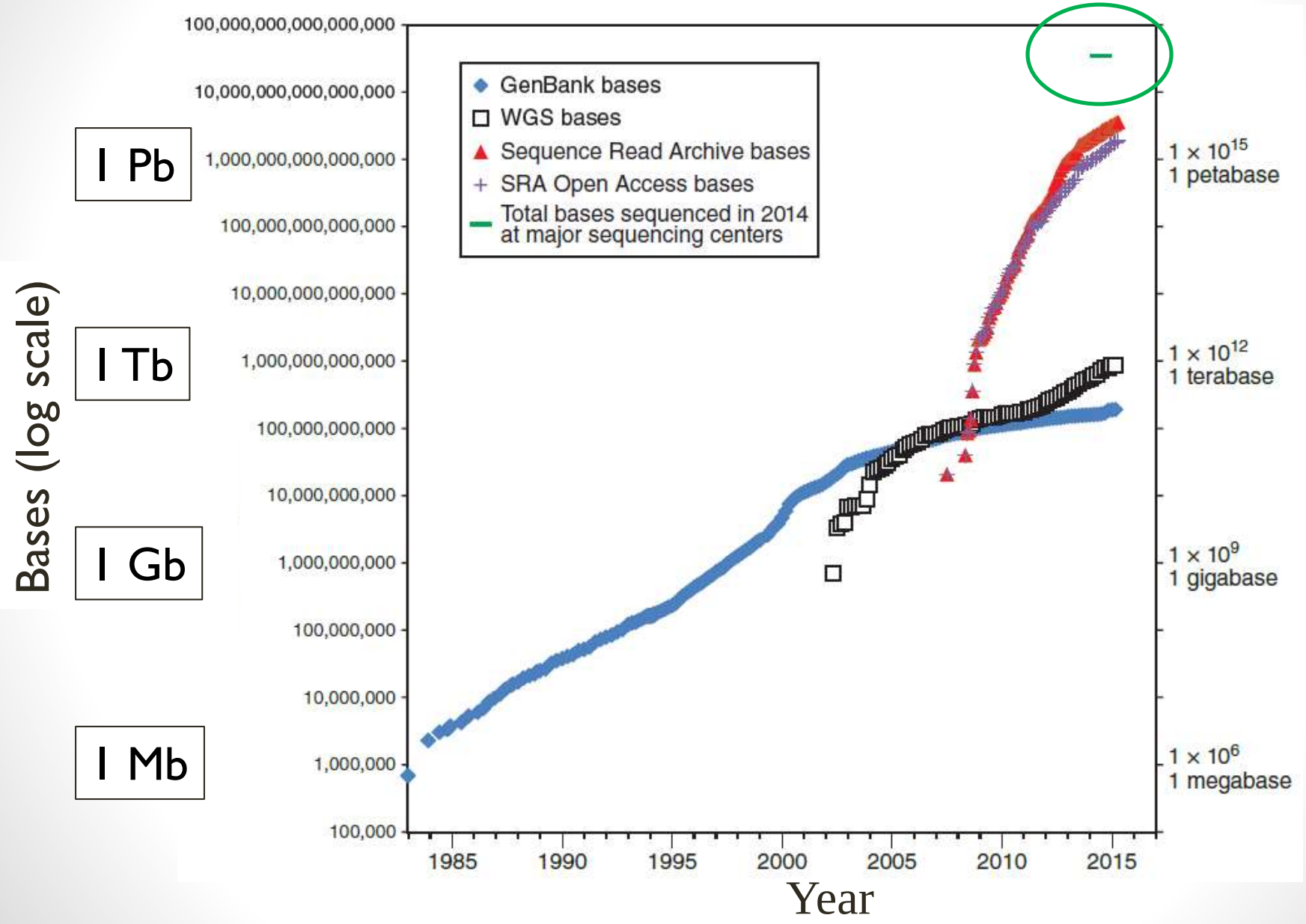
Neanderthal, Denisovan and *Homo erectus* disappear

100,000 50,000 10,000 0

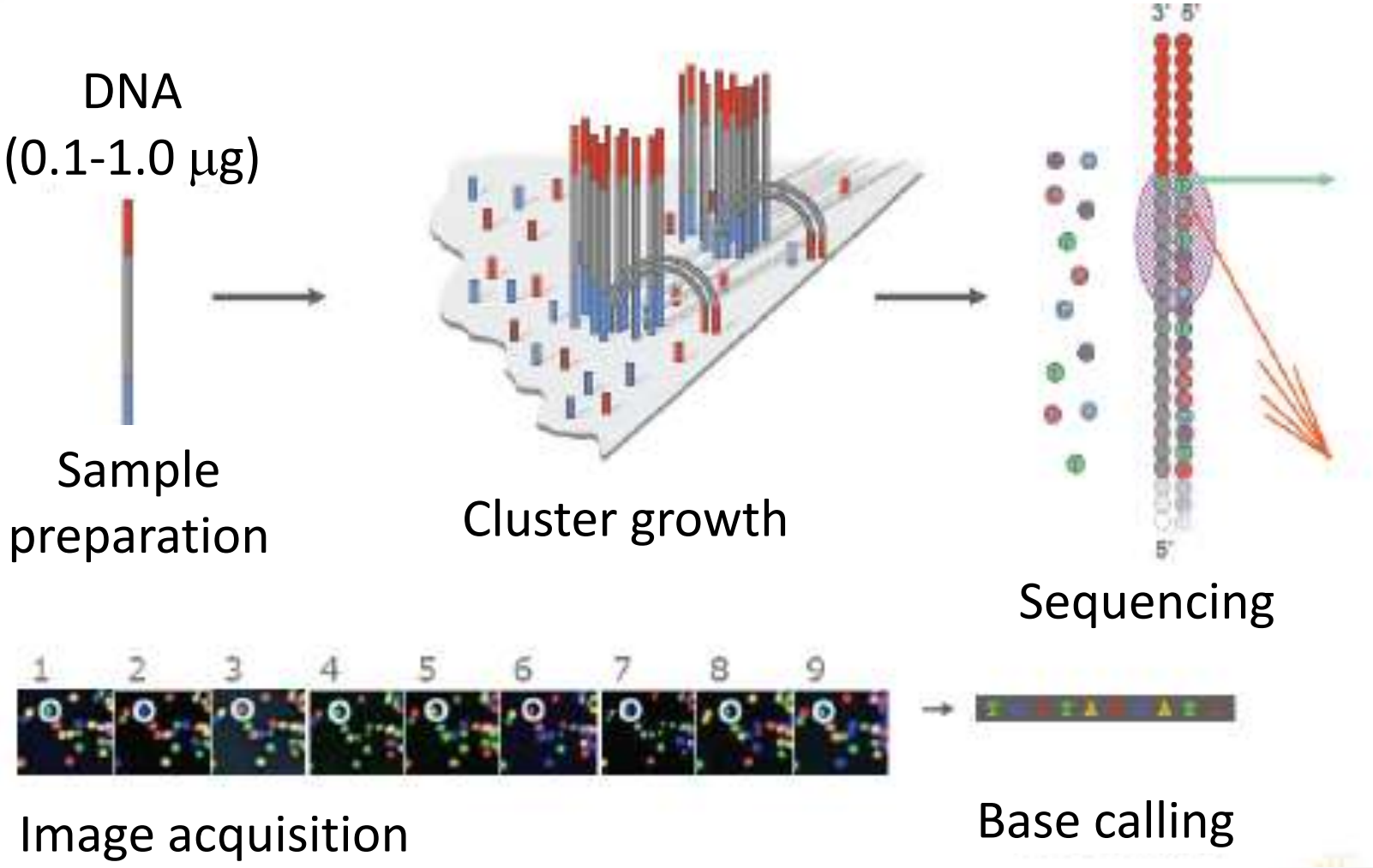


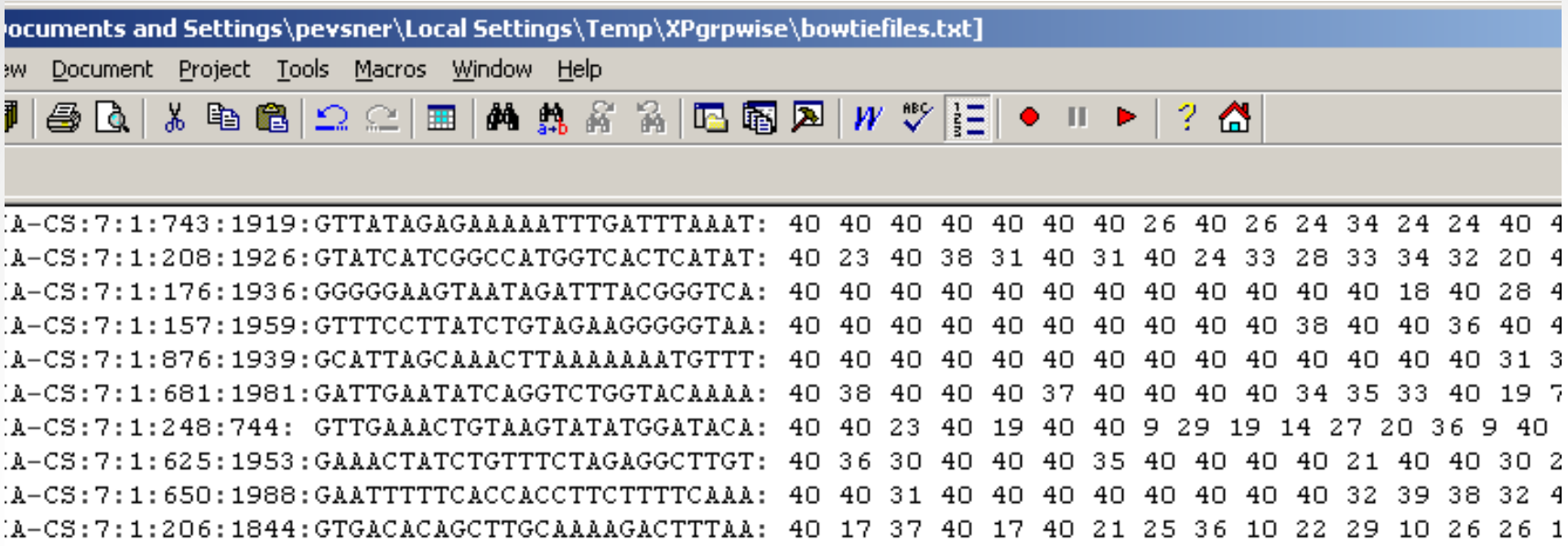
http://humanorigins.si.edu/

Growth of DNA sequence in repositories



Next-generation sequence technology: Illumina

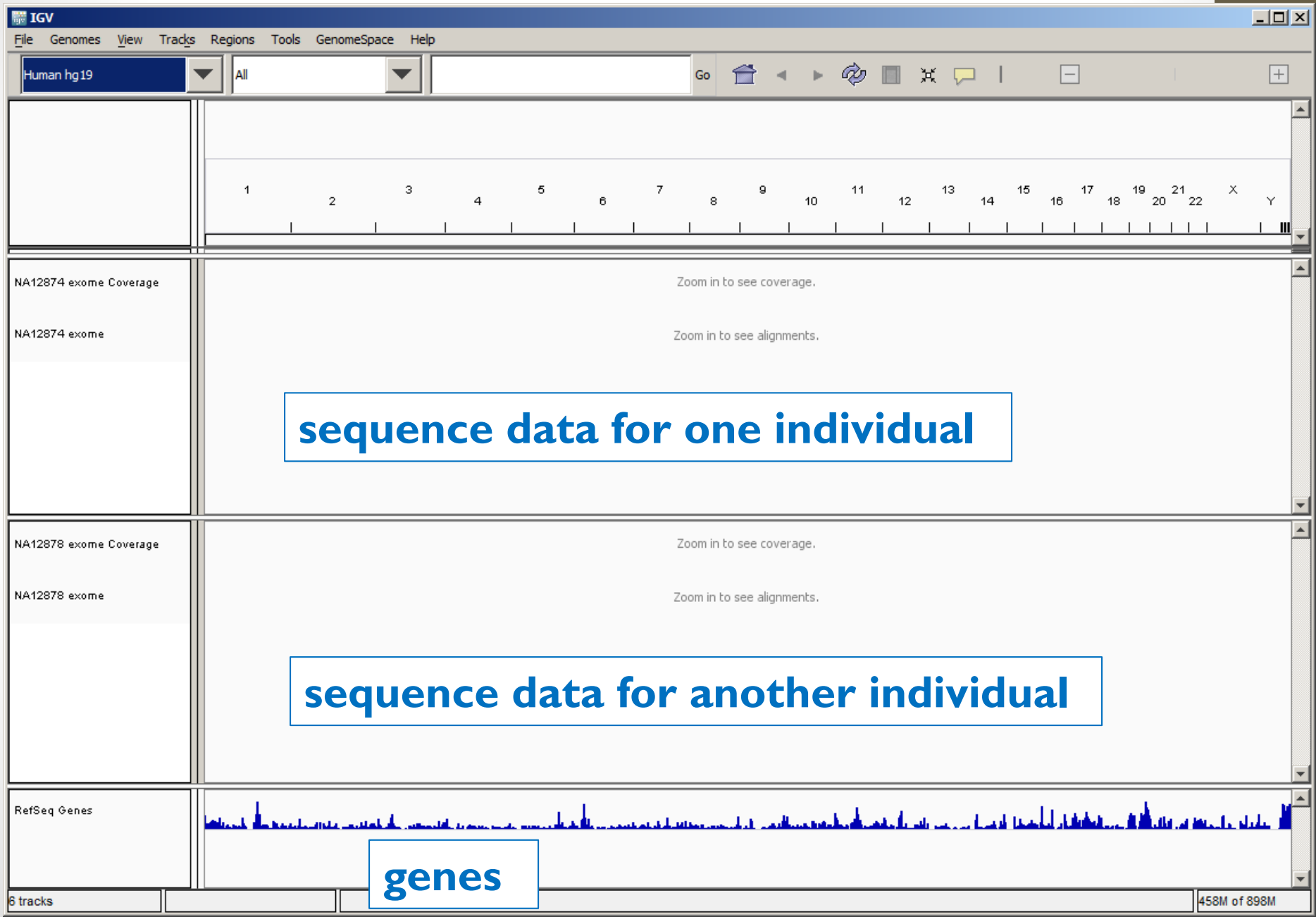




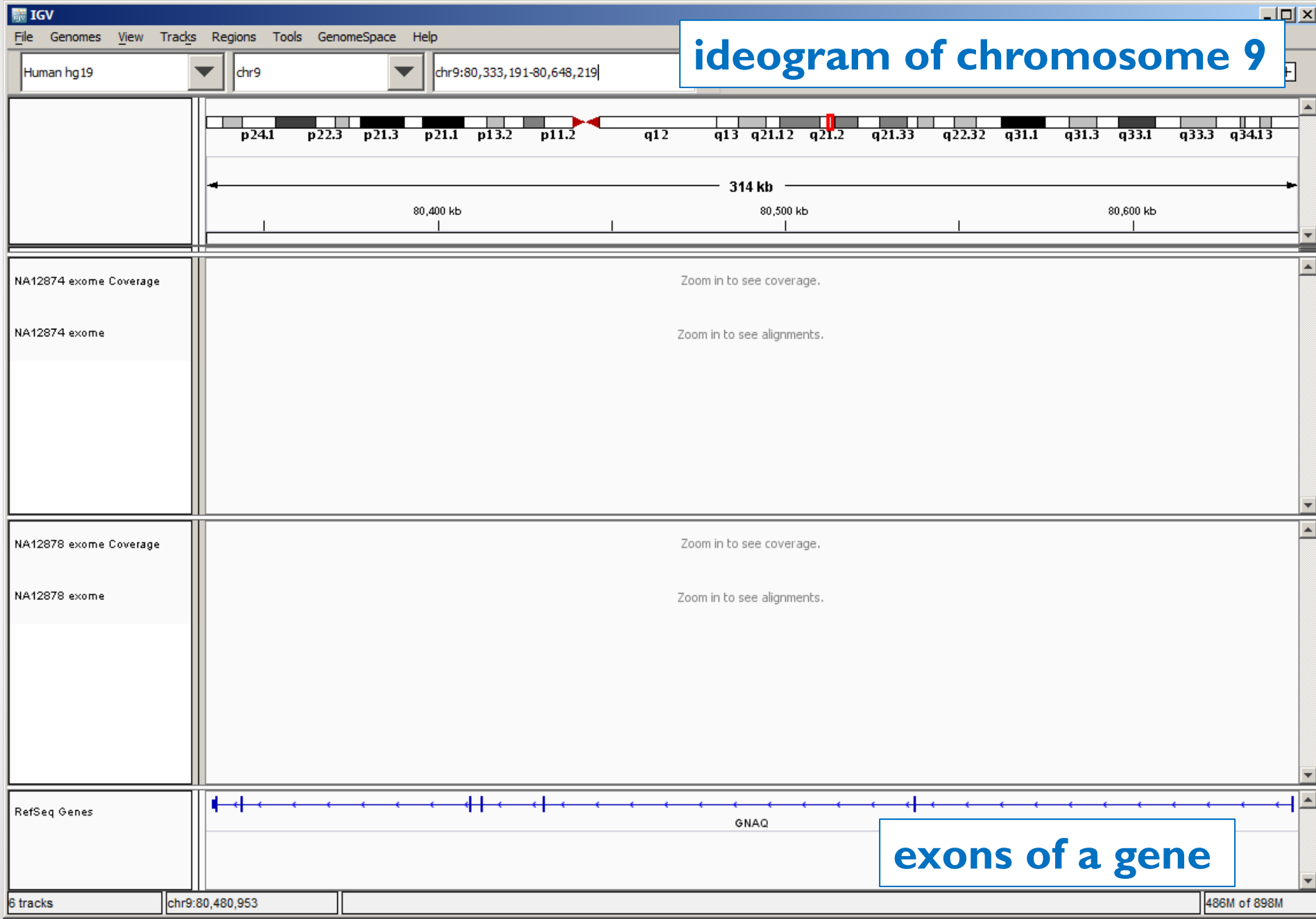
From Illumina:

raw sequence data includes short reads and quality scores

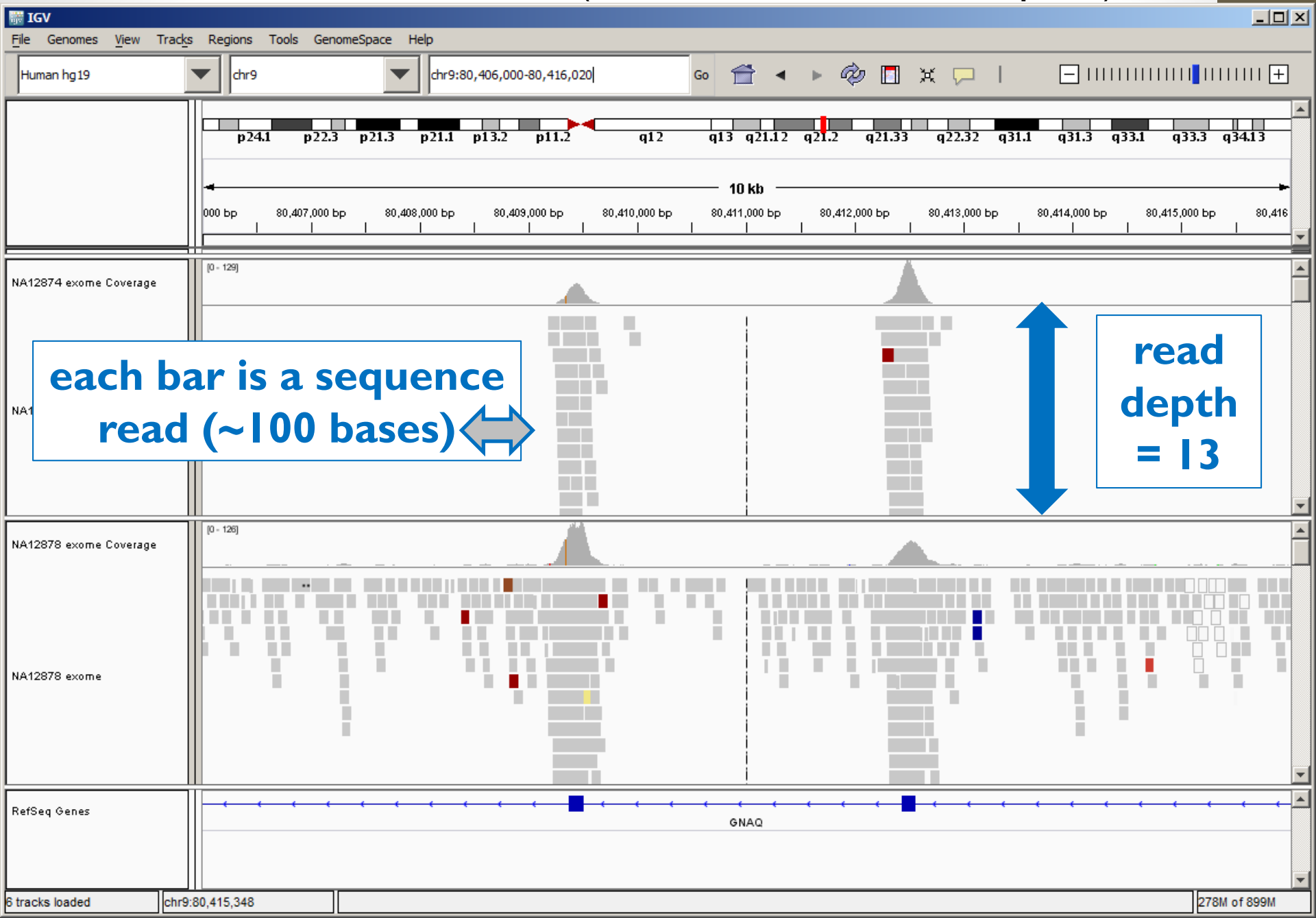
IGV view of the human genome (zoomed to 3 billion base pairs)



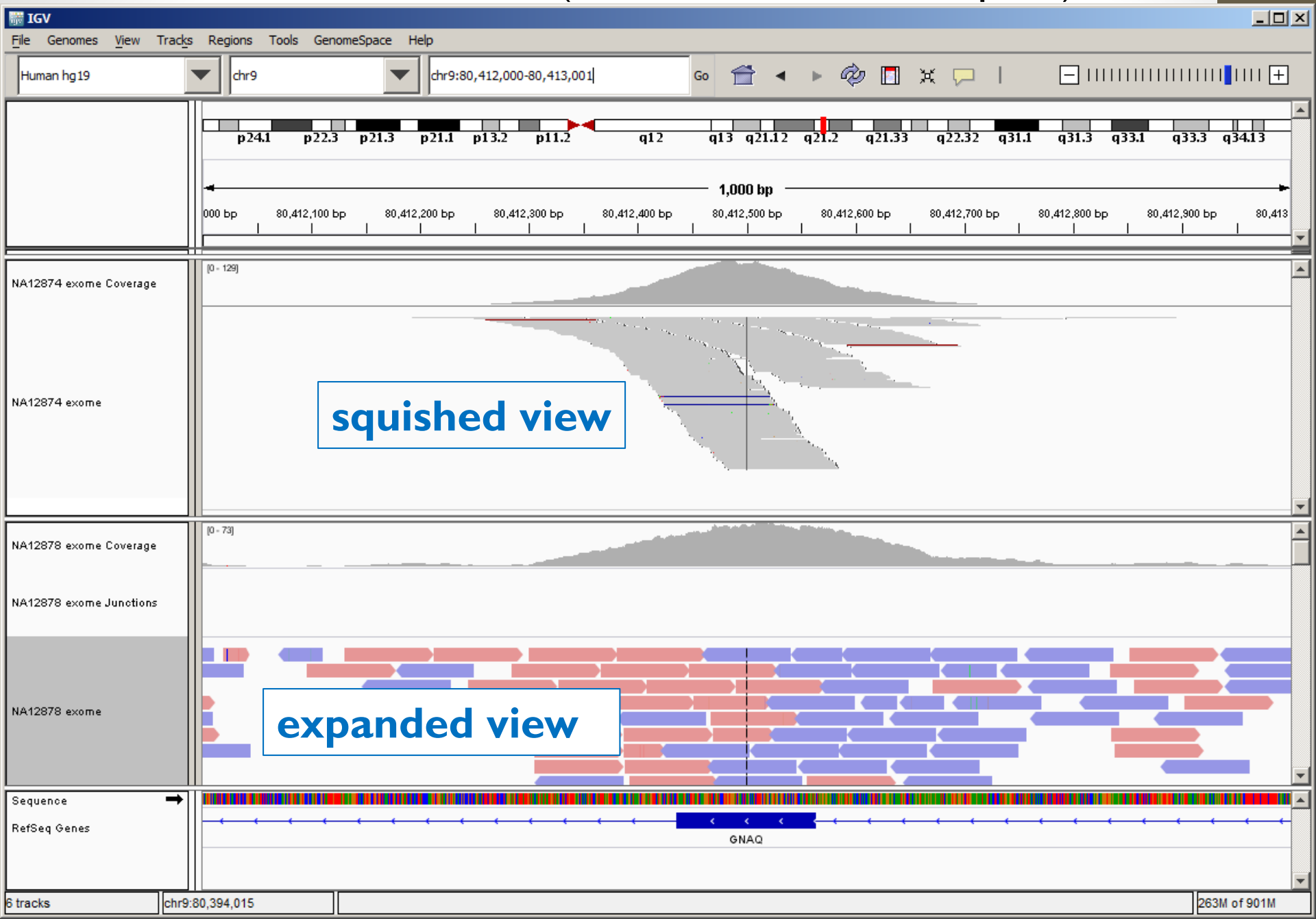
IGV view of one gene (zoomed to 300,000 base pairs)



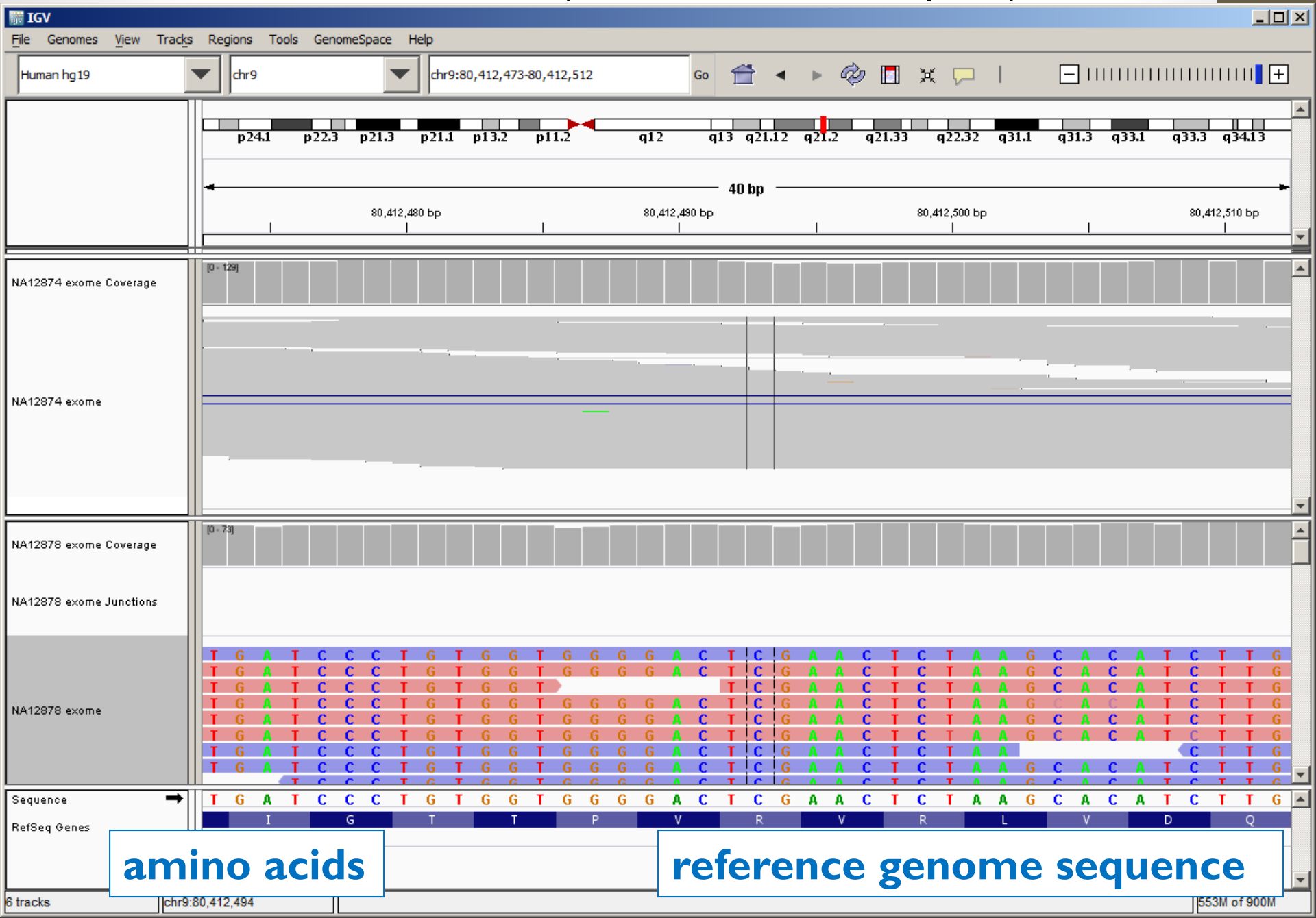
IGV view of two exons (zoomed to 10,000 base pairs)



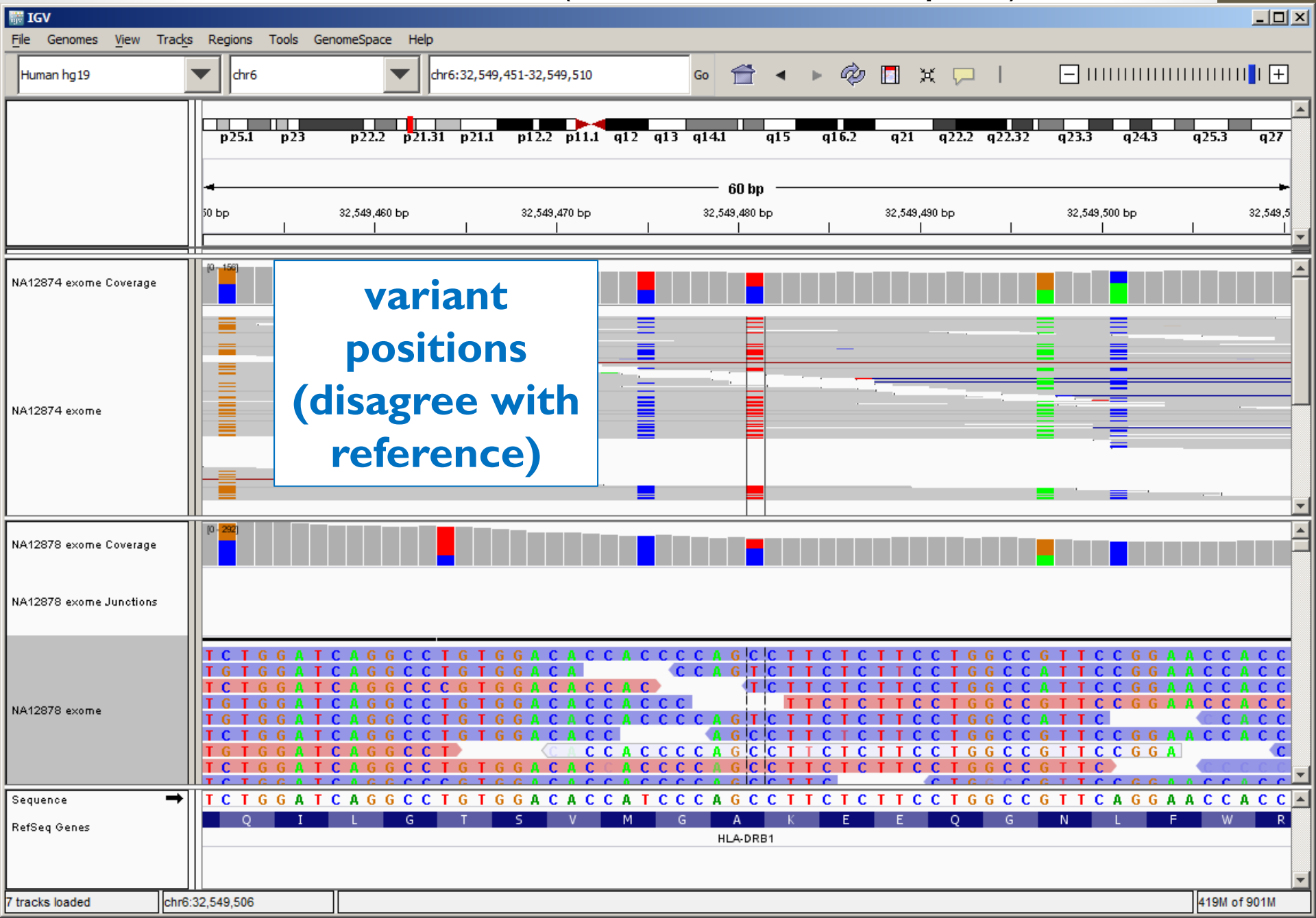
IGV view of one exon (zoomed to 1,000 base pairs)



IGV view of one exon (zoomed to 40 base pairs)



IGV view of one exon (zoomed to 60 base pairs)



Human genome sequencing

We currently obtain whole genome sequences at 30x to 50x depth of coverage. For a typical individual:

- 2.8 billion base pairs are sequenced x 30
  = 100 billion base pairs of DNA
- 3-4 million single nucleotide variants (SNVs)
- ~600,000 insertions/deletions (indels)
- Cost (research basis) is < $1500 per genome
- We try to sequence mother/father/child trios

Human genome sequencing: one purpose is to compare humans to animals

We want to understand what makes the human genome unique. We compare our genome to those of primates and other organisms across the tree of life.

This was a major goal of the Human Genome Project.

Phylogenetic footprinting

Fugu rubripes

Tetraodon nigroviridis

Gallus gallus

Canis familiaris

Rattus norvegicus

Mus musculus

Phylogenetic shadowing

Pan troglodytes

Population shadowing

Homo sapiens

Human genome sequencing: another purpose is to compare humans to each other

A second goal is to understand variation across human genomes. We compare genomes from different geographic (ethnic) groups. Currently we are in the process of sequencing >1 million genomes.

This is a major goal of the HapMap Project and the 1000 Genomes Project.

For Kennedy Krieger patients our goals are:

- improve diagnosis
- improve treatment
- offer genetic counseling (e.g. risk in siblings)

Human disease: a consequence of variation

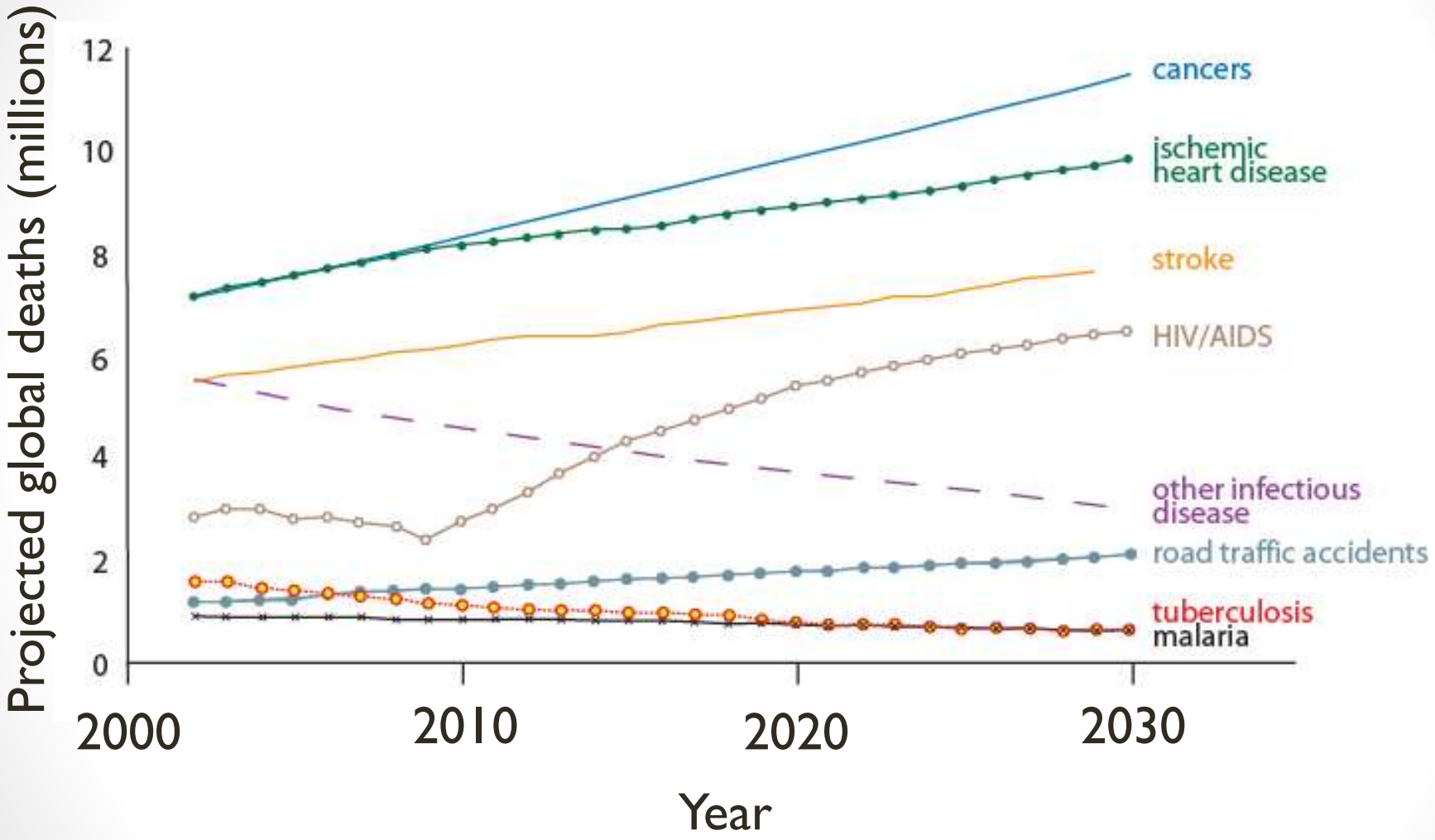
Genetic variation is responsible for the adaptive changes that underlie evolution.

Some changes improve the fitness of a species.
Other changes are maladaptive and represent disease.

Medical perspective: pathological condition.

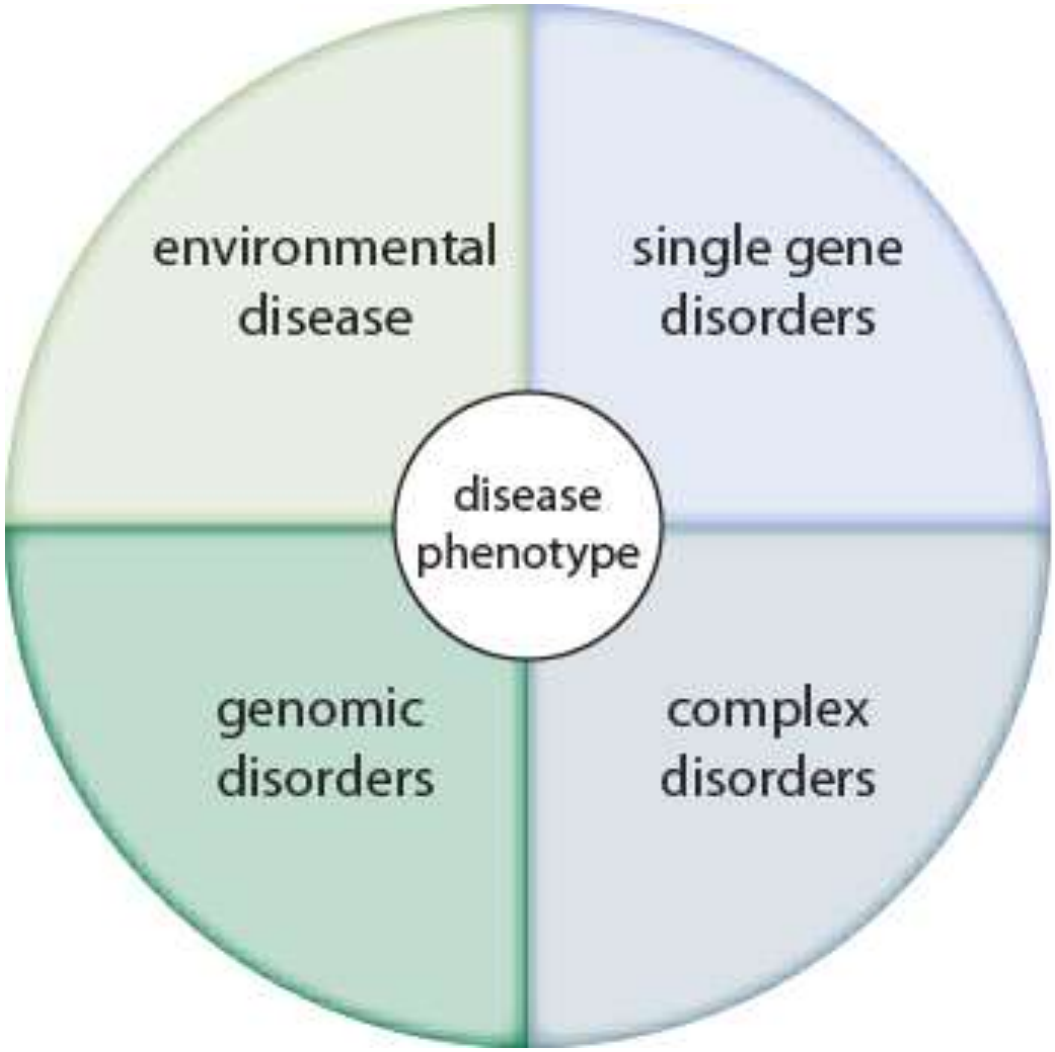
Molecular perspective: mutation and variation.

Projected global deaths (2002 to 2030)



http://www.who.int/whosis/whostat2007.pdf

Four broad causes of disease phenotypes



This chart is not to scale, and all the categories are interconnected. A genomic disorder could be caused by a deletion in which loss of a single gene has a key role (e.g. *RAI1* in Smith-Magenis syndrome)

Life is a relationship between molecules, not a property of any one molecule. So is therefore disease, which endangers life. While there are molecular diseases, there are no diseased molecules. At the level of the molecules we find only variations in structure and physicochemical properties. Likewise, at that level we rarely detect any criterion by virtue of which to place a given molecule "higher" or "lower" on the evolutionary scale. Human hemoglobin, although different to some extent from that of the horse, appears in no way more highly organized. Molecular disease and evolution are realities belonging to superior levels of biological integration. There they are found to be closely linked, with no sharp borderline between them. The mechanism of molecular disease represents one element of the mechanism of evolution. Even subjectively the two phenomena of disease and evolution may at times lead to identical experiences. The appearance of the concept of good and evil, interpreted by man as his painful expulsion from Paradise, was probably a molecular disease that turned out to be evolution. Subjectively, to evolve must most often have amounted to suffering from a disease. And these diseases were of course molecular.

Emile Zuckerkandl and Linus Pauling (1962)

Life is a relationship between molecules, not a property of any one molecule. So is therefore disease, which endangers life. While there are molecular diseases, there are no diseased molecules. At the level of the molecules we find only variations in structure and physicochemical properties. Likewise, at that level we rarely detect any criterion by virtue of which to place a given molecule "higher" or "lower" on the evolutionary scale. Human hemoglobin, although different to some extent from that of the horse, appears in no way more highly organized. Molecular disease and evolution are realities belonging to superior levels of biological integration. There they are found to be closely linked, with no sharp borderline between them. The mechanism of molecular disease represents one element of the mechanism of evolution. Even subjectively the two phenomena of disease and evolution may at times lead to identical experiences. The appearance of the concept of good and evil, interpreted by man as his painful expulsion from Paradise, was probably a molecular disease that turned out to be evolution. Subjectively, to evolve must most often have amounted to suffering from a disease. And these diseases were of course molecular.

Emile Zuckerkandl and Linus Pauling (1962)

Outline

Introduction to genomics and human disease

Identifying a mutation causing a disease: Sturge-Weber

Genomic variation in autism spectrum disorder

Sturge-Weber syndrome

A port-wine birthmark affects about 1:333 people.
It varies in size and location.

Sturge-Weber syndrome

A port-wine birthmark affects about 1:333 people.
It varies in size and location.

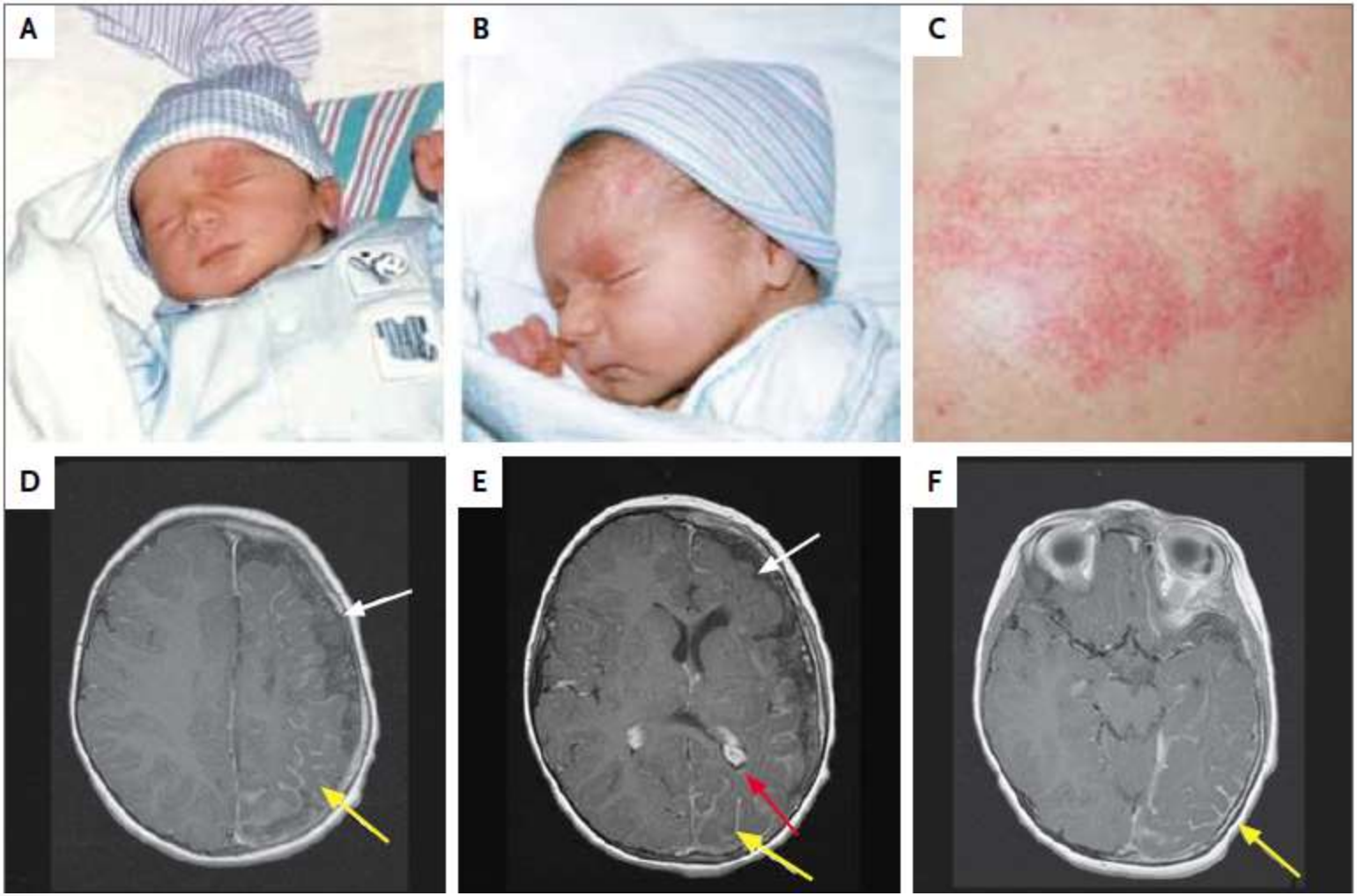
Sturge-Weber syndrome affects < 1:20,000 people.
It affects ~8% of individuals with a facial PW birthmark.

Sturge-Weber syndrome presentation

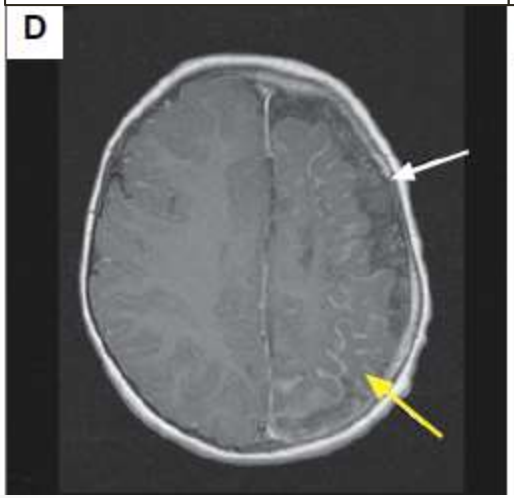
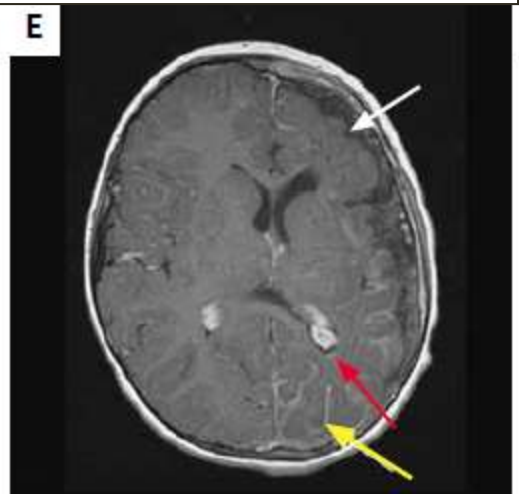
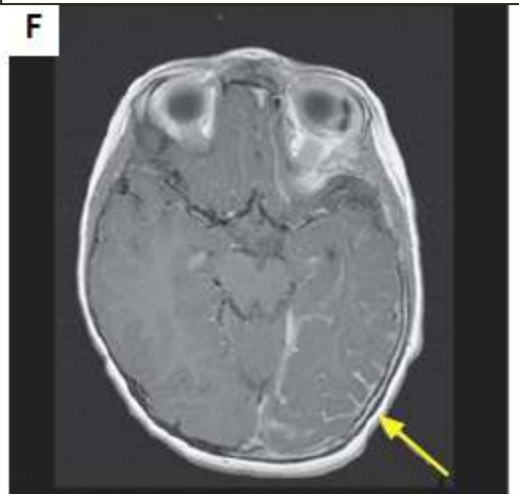
Features of SWS can be highly variable, and may include:

- Port-wine birthmark (facial cutaneous vascular malformation)
- Seizures
- Intellectual disability
- Abnormal capillary venous vessels in the leptomeninges of the brain and choroid
- Glaucoma
- Stroke

Sturge-Weber syndrome presentation



Sturge-Weber syndrome presentation

left hemispheric brain atrophy (white arrows)	enlarged left-sided choroid plexus (red)	left-sided hemispheric leptomeningeal enhancement (yellow)
		

Sturge-Weber syndrome: genetics

SWS appears to be sporadic (rather than familial)

In some studies, identical twins are discordant (consistent with a model of somatic mosaicism)

SWS: hypothesis of somatic mosaicism

Rudolf Happle (1987) speculated that a series of neurocutaneous disorders are caused by somatic mosaicism.

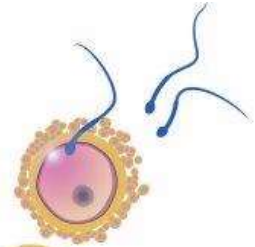
"A genetic concept is advanced to explain the origin of several sporadic syndromes characterized by a mosaic distribution of skin defects. It is postulated that these disorders are due to the action of a lethal gene surviving by mosaicism."

Somatic mosaic mutation

Somatic: changes occur in development (rather than being inherited).

Germline: perhaps an individual with such a mutation would not survive.

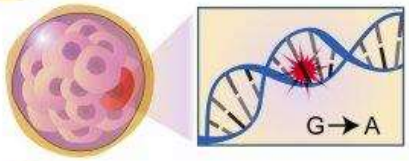
Mosaic: only part of the body is affected.



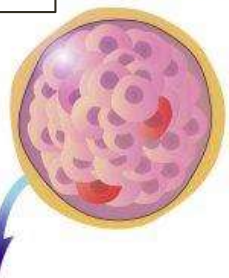
Fertilized egg (from which body's cells arise)

Fertilized egg divides, forms embryo

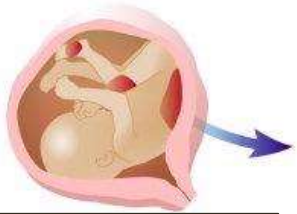
DNA in one cell becomes altered



G becomes A (in *AKT1* or in *GNAQ*)



As the cells in the embryo divide, both normal and mutant cells expand and affect development



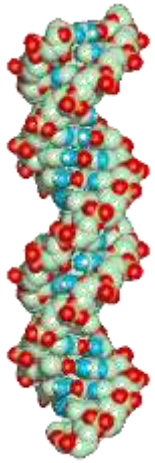
The baby's cells have normal or mutant gene

Some parts of the body grow differently than those with normal cells

http://www.genome.gov/dmd/index.cfm?node=Photos/Graphics

Strategy: sequence and compare two genomes from each patient (n=3 individuals)

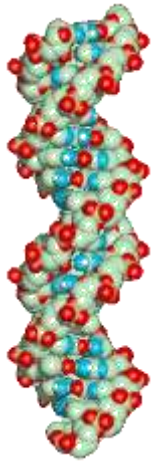
DNA from port-wine birthmark (presumed **affected**)



sequence the genome

Strategy: sequence and compare two genomes from each patient (n=3 individuals)

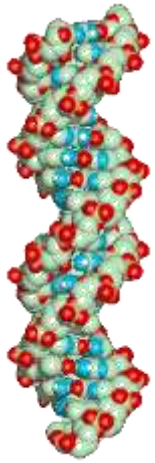
DNA from port-wine birthmark (presumed **affected**)



sequence the genome

compare

DNA from blood (presumed **unaffected**)



sequence the genome

Strategy: sequence and compare two genomes from each patient (n=3 individuals)

Each genome:

- ~3 billion bases of DNA
- Sequenced to 30x average depth of coverage, so 100 billion bases per genome
- A pair of genomes is compared (using a somatic variant caller)
- 100 GB raw data per genome
- Allow < 1 TB storage/genome

sequence the genome

compare

sequence the genome

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Sturge–Weber Syndrome and Port-Wine Stains Caused by Somatic Mutation in *GNAQ*

Matthew D. Shirley, Ph.D., Hao Tang, Ph.D., Carol J. Gallione, B.A., Joseph D. Baugher, Ph.D., Laurence P. Frelin, M.S., Bernard Cohen, M.D., Paula E. North, M.D., Ph.D., Douglas A. Marchuk, Ph.D., Anne M. Comi, M.D., and Jonathan Pevsner, Ph.D.

ABSTRACT

BACKGROUND

The Sturge–Weber syndrome is a sporadic congenital neurocutaneous disorder characterized by a port-wine stain affecting the skin in the distribution of the ophthalmic branch of the trigeminal nerve, abnormal capillary venous vessels in the leptomeninges of the brain and choroid, glaucoma, seizures, stroke, and intellectual disability. It has been hypothesized that somatic mosaic mutations disrupting vascular development cause both the Sturge–Weber syndrome and port-wine stains, and the severity and extent of presentation are determined by the developmental time point at which the mutations occurred. To date, no such mutation has been identified.

PMID: 23656586

Analysis of high confidence results with Strelka resulted in one candidate mutation

Step	somatic SNVs	somatic indels
Pre-filtered calls	24,848	1,646
Post-filtered calls	1,300	27
VAAST pre-filtered	83	NA
VAAST post-filtered	1	NA

All 27 somatic indels were in repetitive regions

We performed targeted sequencing of a portion of *GNAQ*.

In skin samples, almost all patients had the mutation.

The mutant allele frequency was 1% to about 18%.

Table 1. Somatic Mutation of *GNAQ* in Skin Samples.*

Patient No.	Mutation Present†	PWS	SWS	Mutant Allele Frequency‡	Total No. of Samples Assayed
				percent	
1	Yes	Yes	Yes	3.60	1
1	No	No	Yes	0.11	1
2	Yes	Yes	Yes	3.17	1
2	No	No	Yes	0.13	1
3	Yes	Yes	Yes	6.06–6.46	2
3	No	No	Yes	0.62–0.93	2
4	Yes	Yes	Yes	3.50–4.51	2
4	No	No	Yes	0.13–0.90	2
5	Yes	Yes	Yes	3.38	1
5	No	No	Yes	0.11	1
6	Yes	Yes	Yes	3.99	1
7	Yes	Yes	Yes	2.05–2.16	2
7	Yes	No	Yes	0.09–2.00	2
8	Yes	Yes	Yes	4.08	1
8	No	No	Yes	0.06	1
9	Yes	Yes	No	5.58	1
10	Yes	Yes	No	2.76	1
10	Yes	No	No	1.14	1
11	Yes	Yes	No	6.70	1
12	No	Yes	No	0.00	1
13	Yes	Yes	No	5.90	1
14	Yes	Yes	No	6.20	1
15	Yes	Yes	No	14.20	1
16	Yes	Yes	No	1.70	1
17	Yes	Yes	No	4.50	1
18	Yes	Yes	No	5.30	1
19	Yes	Yes	No	4.70	1
20	Yes	Yes	No	4.30	1
21	Yes	Yes	No	18.10	1
22	Yes	Yes	Yes	5.00	1

We performed targeted sequencing of a portion of *GNAQ*.

In skin samples, almost all patients had the mutation.

The mutant allele frequency was 1% to about 18%.

Table 1. Somatic Mutation of *GNAQ* in Skin Samples.*

Patient No.	Mutation Present†	PWS	SWS	Mutant Allele Frequency‡	Total No. of Samples Assayed
				percent	
1	✓ Yes	✓ Yes	Yes	3.60	1
1	No	No	Yes	0.11	1
2	✓ Yes	✓ Yes	Yes	3.17	1
2	No	No	Yes	0.13	1
3	✓ Yes	✓ Yes	Yes	6.06–6.46	2
3	No	No	Yes	0.62–0.93	2
4	✓ Yes	✓ Yes	Yes	3.50–4.51	2
4	No	No	Yes	0.13–0.90	2
5	✓ Yes	✓ Yes	Yes	3.38	1
5	No	No	Yes	0.11	1
6	✓ Yes	✓ Yes	Yes	3.99	1
7	✓ Yes	✓ Yes	Yes	2.05–2.16	2
7	Yes	No	Yes	0.09–2.00	2
8	✓ Yes	✓ Yes	Yes	4.08	1
8	No	No	Yes	0.06	1
9	✓ Yes	✓ Yes	No	5.58	1
10	✓ Yes	✓ Yes	No	2.76	1
10	Yes	No	No	1.14	1
11	✓ Yes	✓ Yes	No	6.70	1
12	No	Yes	No	0.00	1
13	Yes	Yes	No	5.90	1
14	Yes	Yes	No	6.20	1
15	Yes	Yes	No	14.20	1
16	Yes	Yes	No	1.70	1
17	Yes	Yes	No	4.50	1
18	Yes	Yes	No	5.30	1
19	Yes	Yes	No	4.70	1
20	Yes	Yes	No	4.30	1
21	Yes	Yes	No	18.10	1
22	Yes	Yes	Yes	5.00	1

In brain samples, most (not all) patients had a mutation.

Control brain samples: no mutation

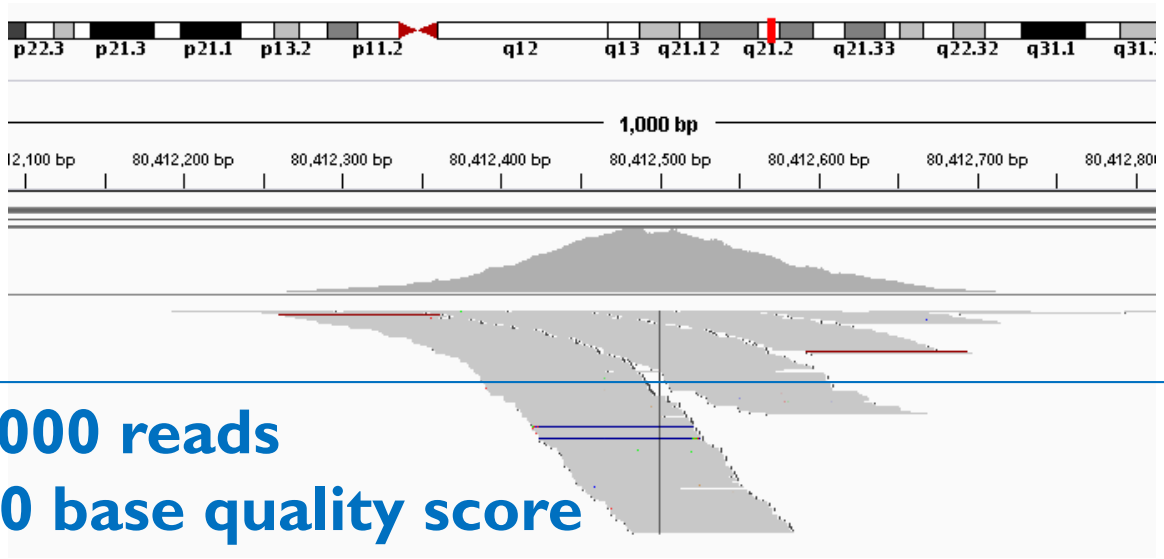
Table 2. Somatic Mutation of *GNAQ* in Brain-Tissue Samples.*

Patient No.	Mutation Present	SWS	Mutant Allele Frequency	Total No. of Samples Assayed
			percent	
7	✓ Yes	✓ Yes	5.57–5.63	2
23	✓ Yes	✓ Yes	5.56–5.78	2
24	✓ Yes	✓ Yes	2.67–3.51	2
25	No	✓ Yes	0.02–0.10	2
26	✓ Yes	✓ Yes	0.13–3.06	4
27	✓ Yes	✓ Yes	2.19–5.12	2
28	✓ Yes	✓ Yes	6.95–8.13	4
29	✓ Yes	✓ Yes	6.04–11.15	5
30	✓ Yes	✓ Yes	4.14	1
31	✓ Yes	✓ Yes	4.78	1
32	✓ Yes	✓ Yes	0.22–1.48	4
33	✓ Yes	✓ Yes	4.04–5.74	2
34	No	✓ Yes	0.05–0.12	2
35	✓ Yes	✓ Yes	0.05–1.51	7
36	✓ Yes	✓ Yes	0.35–6.03	5
37	✓ Yes	✓ Yes	5.74–6.49	2
38	No	✓ Yes	0.03–0.05	2
39	✓ Yes	✓ Yes	1.83	1
40	No	No	0.11	1
41	No	No	0.05	1
42	No	No	0.08	1
43	No	No	0.09	1
44	No	No	0.04	1
45	No	No	0.04	1
46	No	No, CCM	0.00	1
47	No	No, CCM	0.00	1
48	No	No, CCM	0.00	1
49	No	No, CCM	0.00	1

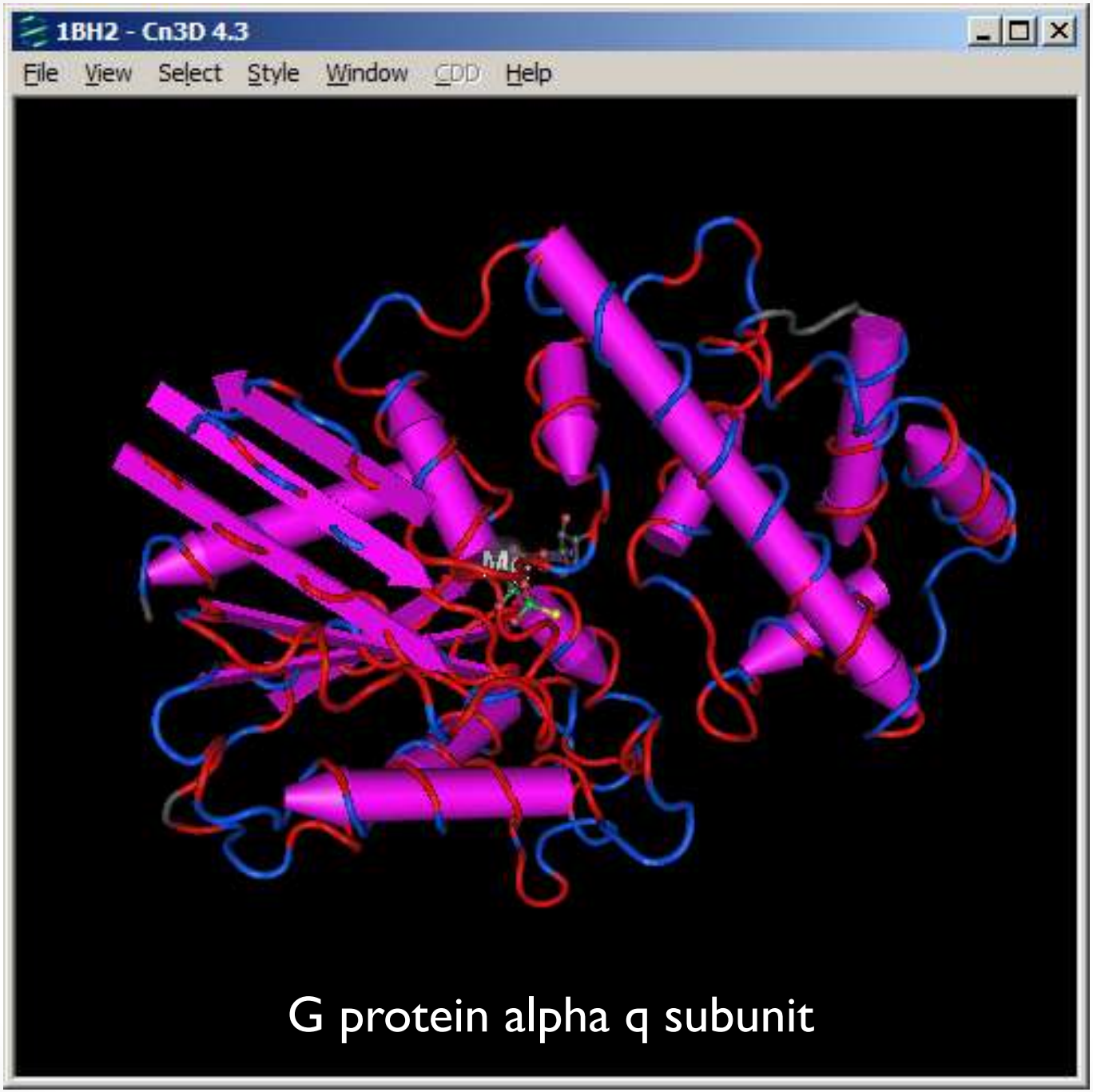
Targeted sequencing of a portion of *GNAQ* reveals mutations in SWS and PWS cases

# subjects	Tissue	SWS	GNAQ c.548 G->A	Detection
9	PWS	Yes	100%	Amplicon seq
7	Skin (non PWS)	Yes	14%	Amplicon seq
13	PWS	No	92%	Amplicon seq Primer extension
18	Brain	Yes	88%	Amplicon seq
6	Brain	No	0%	Amplicon seq
4	Brain	No: CCM	0%	Primer extension
669	Blood/LCL	N/A	0.7%	Exome seq

Amplicon sequencing: 13,000x median read depth
Exome sequencing (1KG project): 271x median read depth
Primer extension: SNaPshot assay (Doug Marchuk's lab)



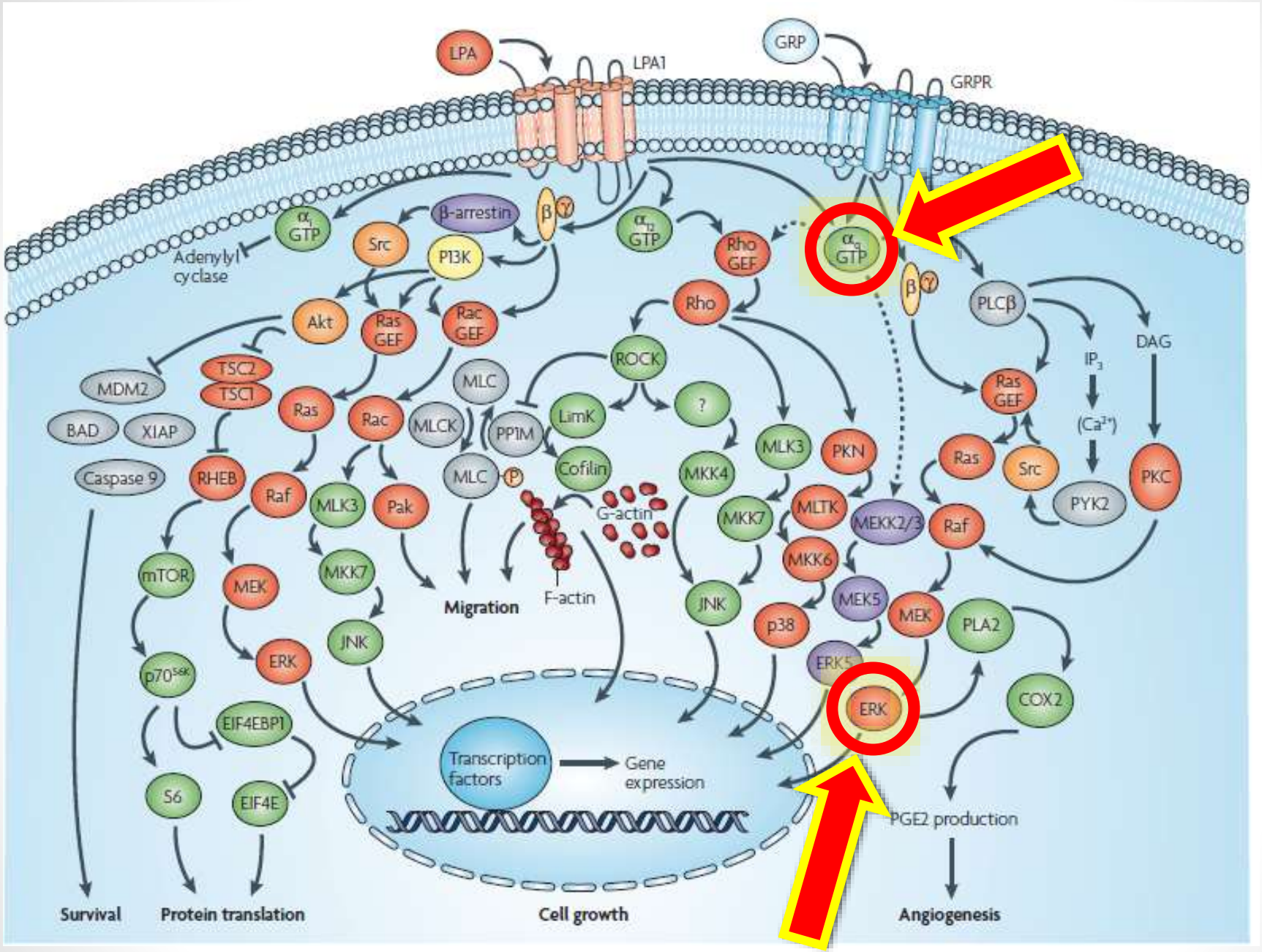
- 13,000 reads
- Q30 base quality score
- 1:1000 error rate
- Expect 13 errors in 13,000 reads
- If we see 10x the error rate, call a mutation
- Call mutation if we see 130 T bases per 13,000 normal bases



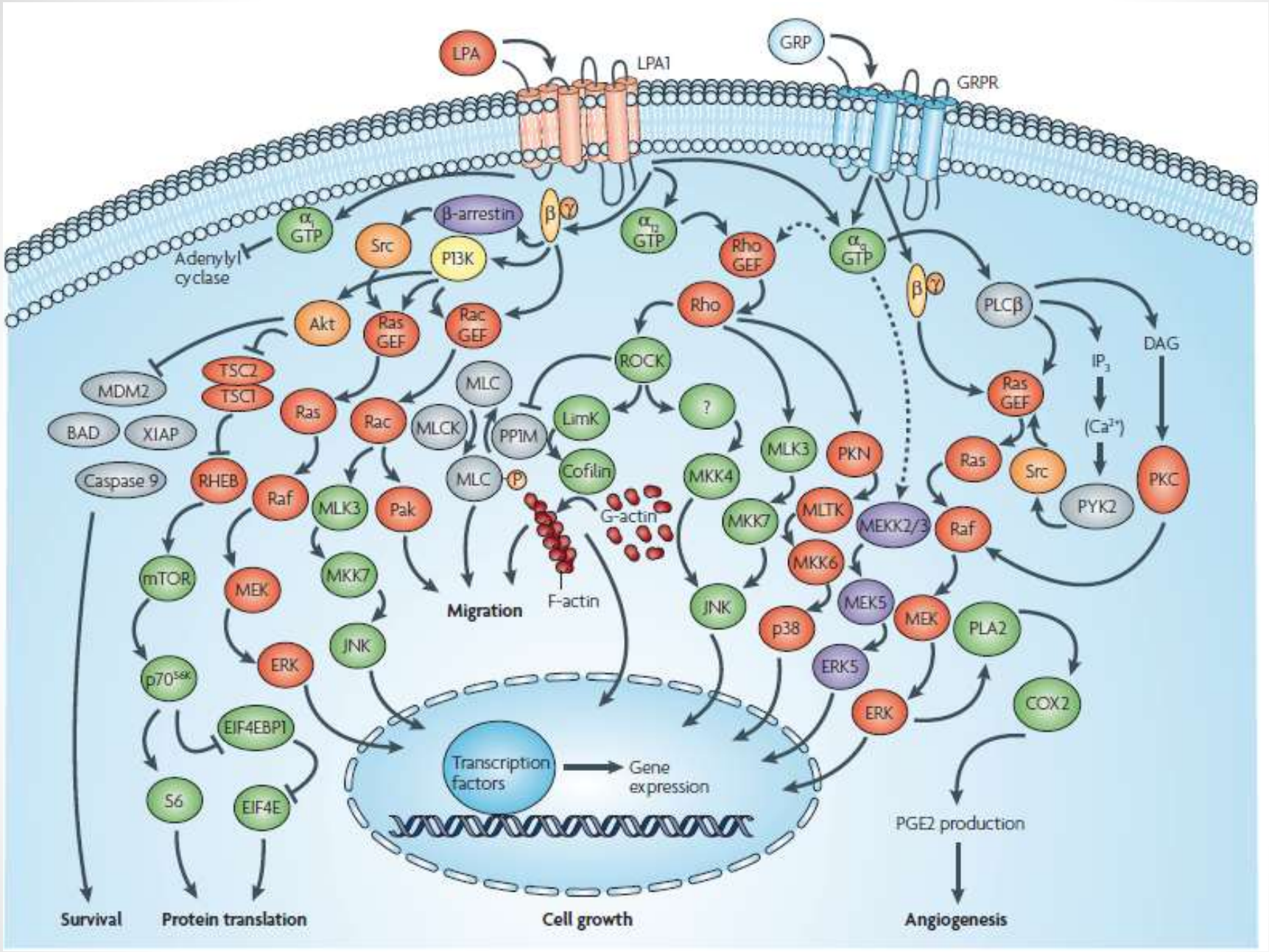
G protein alpha q subunit

R183Q: an activating mutation in $G\alpha q$

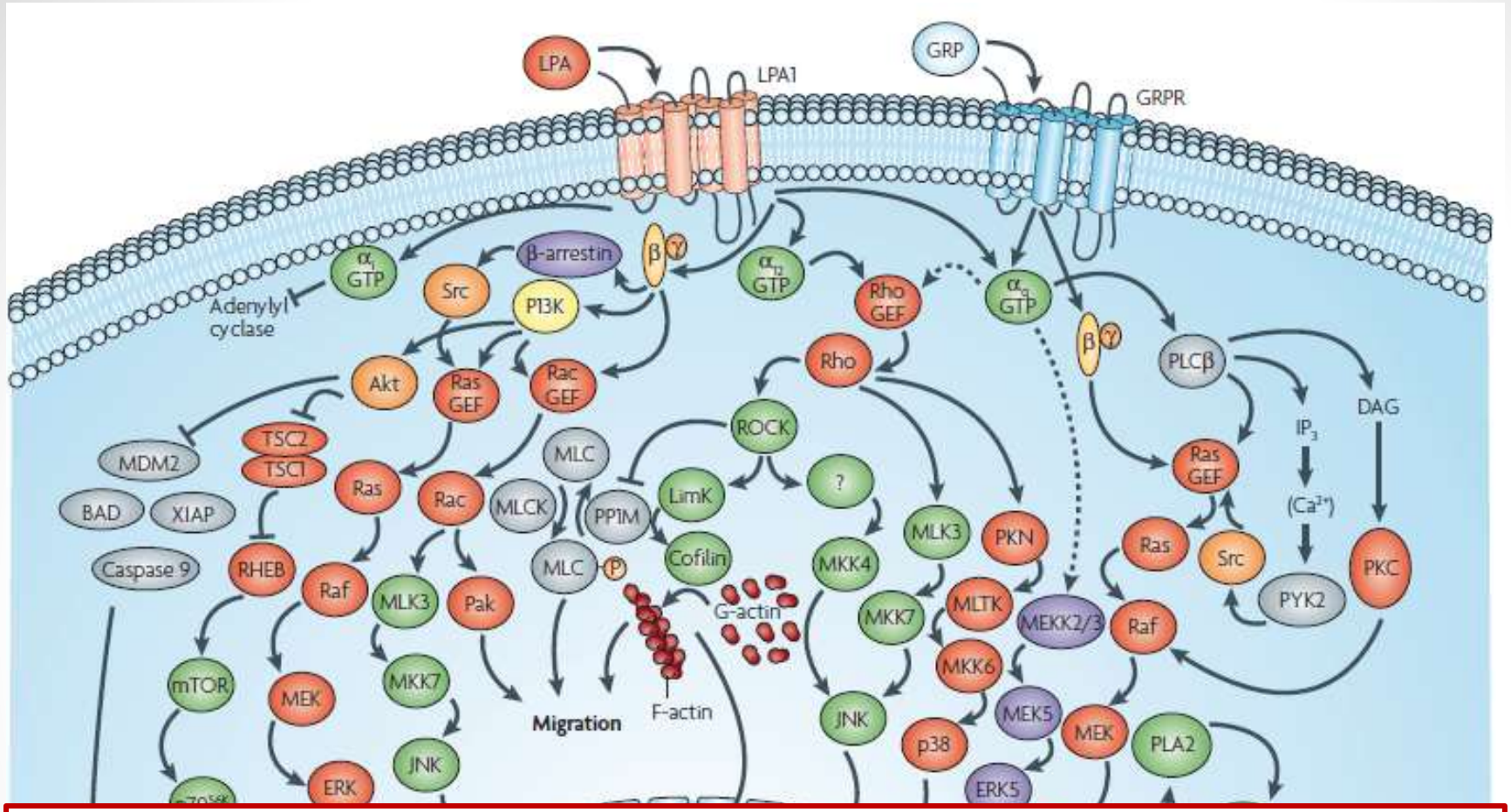
- In 2009 this identical mutation was described in uveal melanoma (a cancer involving melanocytes)
- The R183Q mutation occurs in 2-6% of these melanomas
- Another activating mutation (Q209L in $G\alpha q$) occurs in ~50% of uveal melanoma
- The mutation has been implicated in dermal hyper-pigmentation



2007 Dorsam and Gutkind



2007 Dorsam and Gutkind



Mutations in genes encoding many of these signaling proteins cause somatic, mosaic, and often neurocutaneous disorders.

TSC1, *TSC2*: tuberous sclerosis

NF1: neurofibromatosis

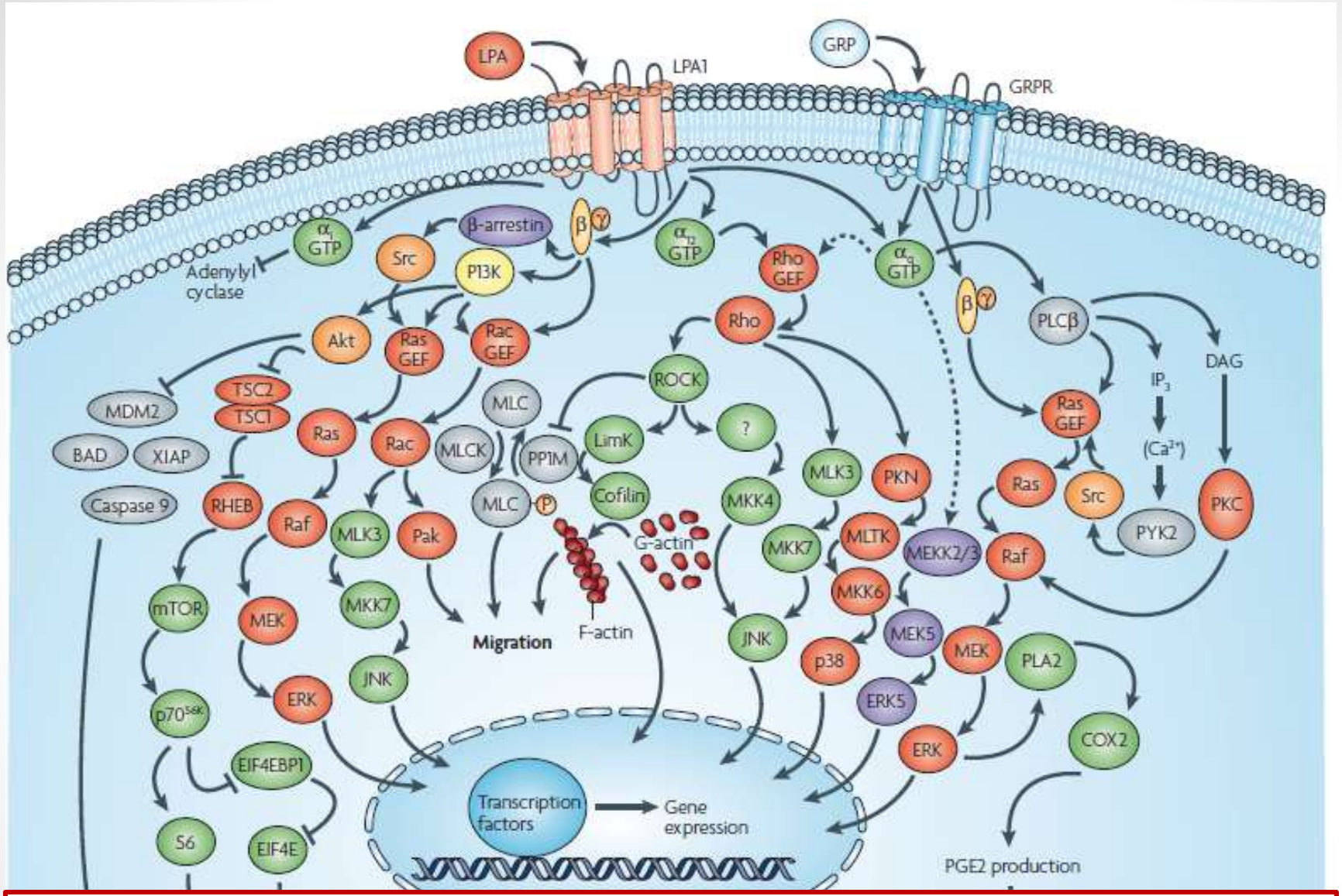
AKT1: Proteus syndrome

PI3K: CLOVE syndrome, hemimegalencephaly

GNAQ: Sturge-Weber

GNAS: McCune-Albright

RAS: epidermal nevi



Mutations in many of these genes cause cancer.

Tumor suppressors: *NF1*, *TSC1*, *TSC2*

Oncogenes: *RHEB*, *PIK3CA*, *RAS*, *GNAQ*, *RAF*, *MAP2K1*, *PKC*

Conclusions: Sturge-Weber syndrome

We identified mutations in the *GNAQ* gene as the main cause of Sturge-Weber syndrome and port-wine birthmarks.

Knowing the genetic cause of the disease offers us a direction to search for treatments (and cures).

The consequence of the *GNAQ* mutation is to activate a cellular pathway. We can test drugs for the ability to reduce this persistent activation.

The same strategies may apply to treating uveal melanoma.

Outline

Introduction to genomics and human disease

Identifying a mutation causing a disease: Sturge-Weber

Genomic variation in autism spectrum disorder

Autism spectrum disorder (ASD): diagnostic criteria

- Deficits in social communication and interaction
- Restricted and repetitive patterns of behavior, interests or activities

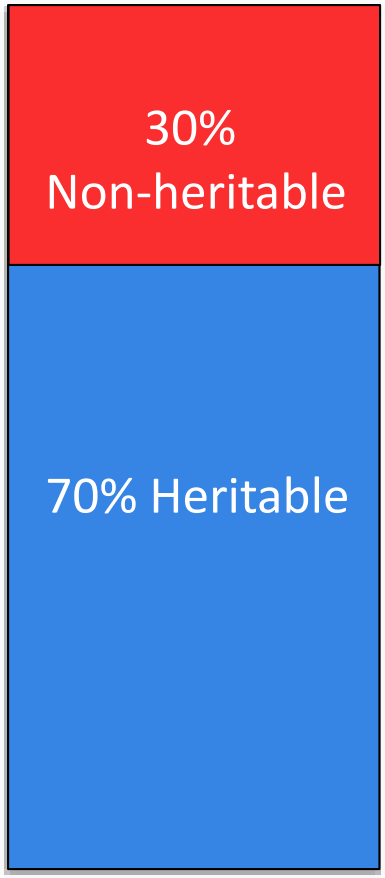
- Symptoms cause significant impairment of function
- Diagnosed in childhood
- Comorbidities: intellectual disability, seizure, developmental delay, self-injury

Causes of ASD

- Associated with syndromic disorders (12% of ASD cases)
  - Fragile X syndrome
  - Rett Syndrome
  - Tuberous sclerosis
- *de novo* CNVs (6% of simplex cases)
- *de novo* SNVs/Indels (21% of simplex cases)

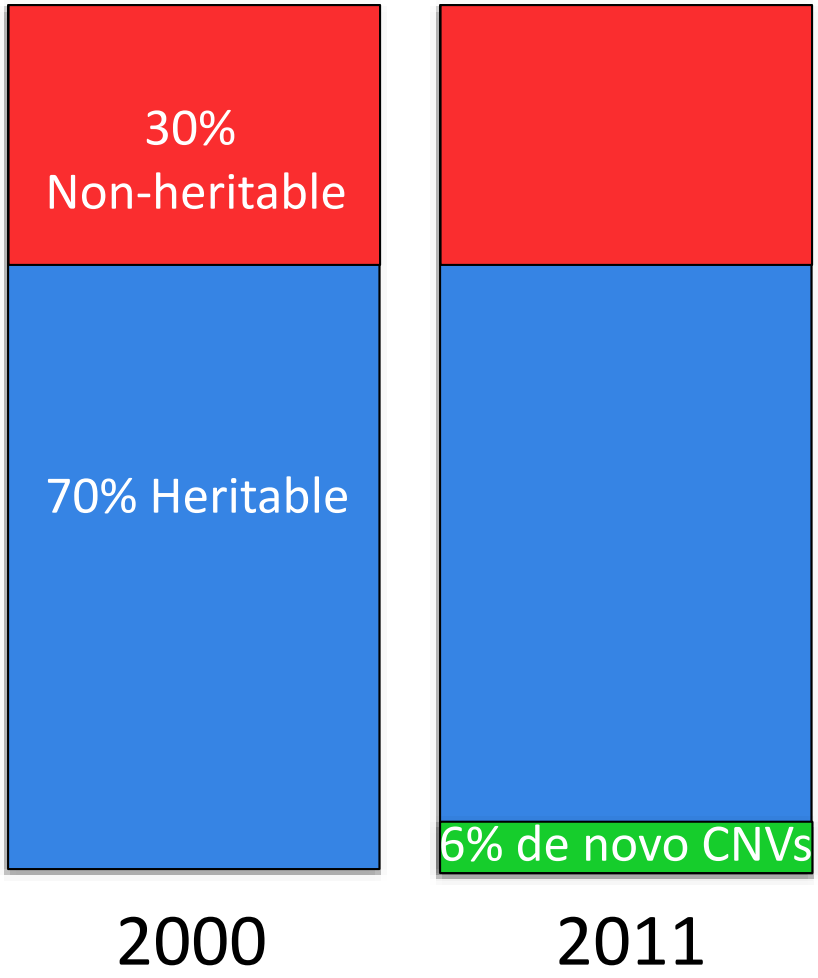
Heritability is the proportion of phenotypic variance due to genetic variance. For ASD, 50% to 90% heritability.

Understanding the genetic architecture of autism spectrum disorder

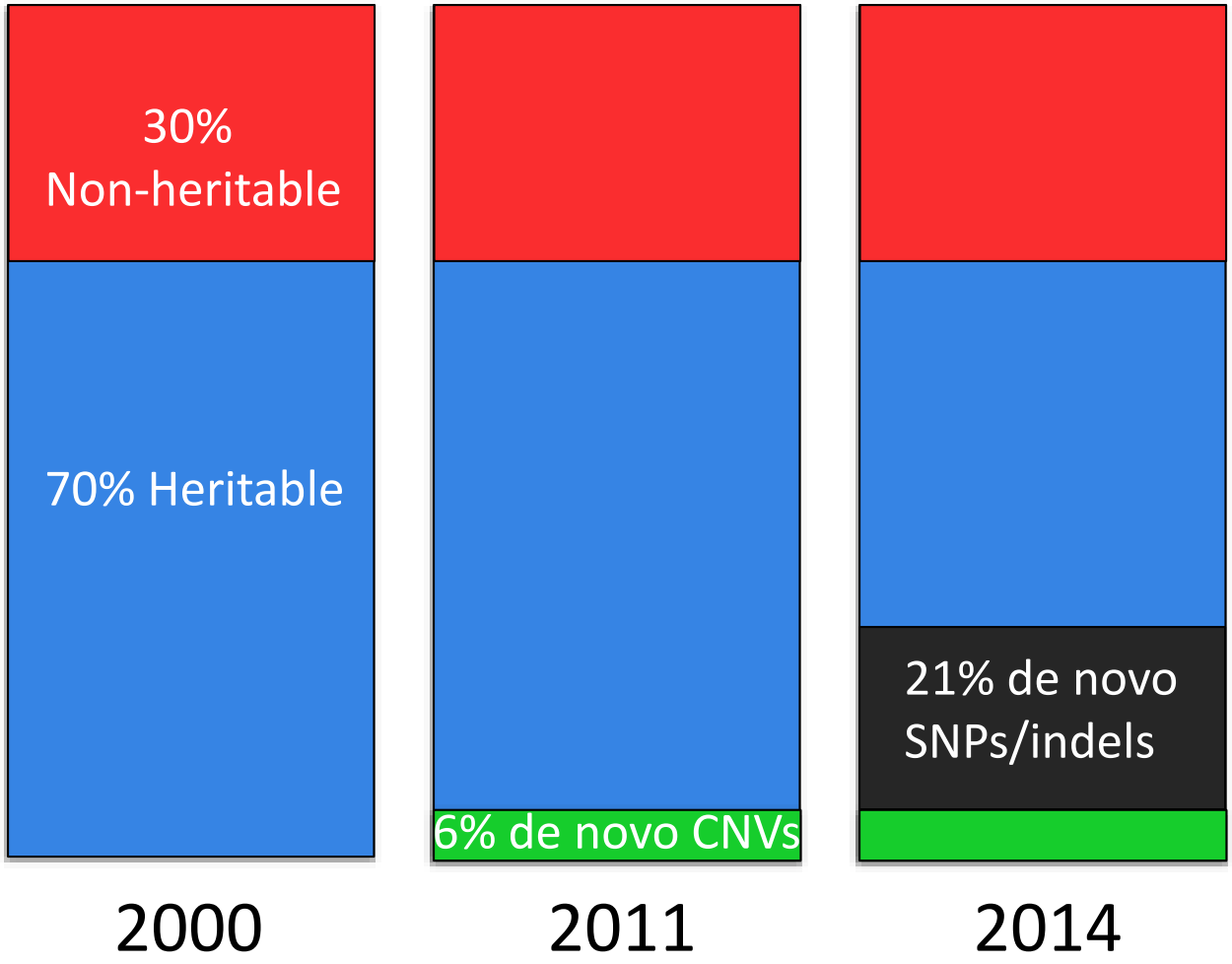


2000

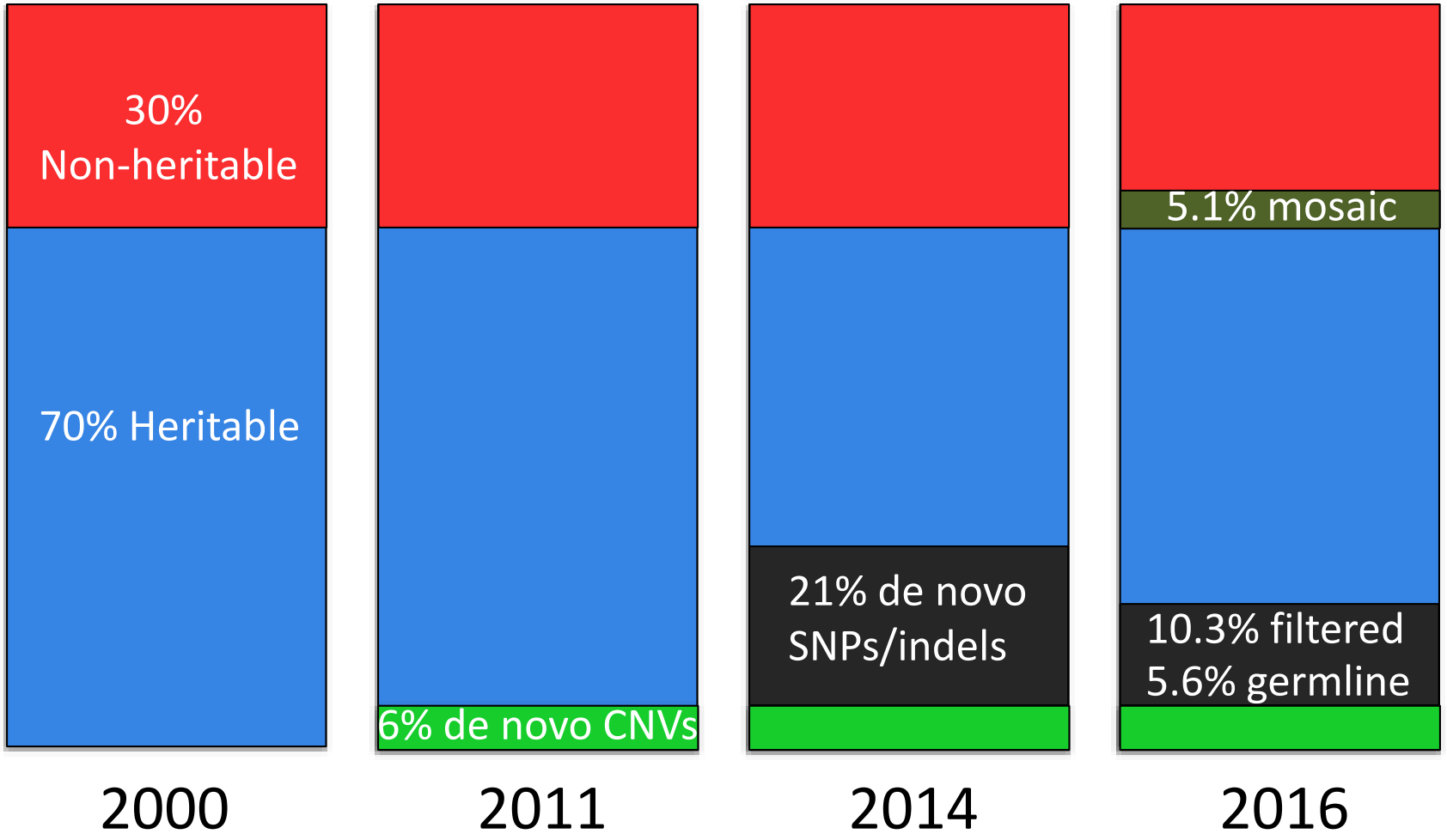
Understanding the genetic architecture of autism spectrum disorder



Understanding the genetic architecture of autism spectrum disorder



Understanding the genetic architecture of autism spectrum disorder



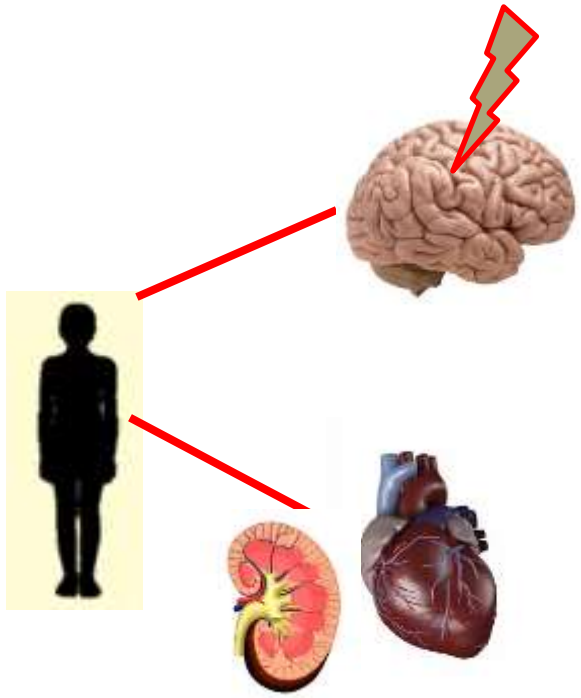
PLOS | GENETICS

RESEARCH ARTICLE

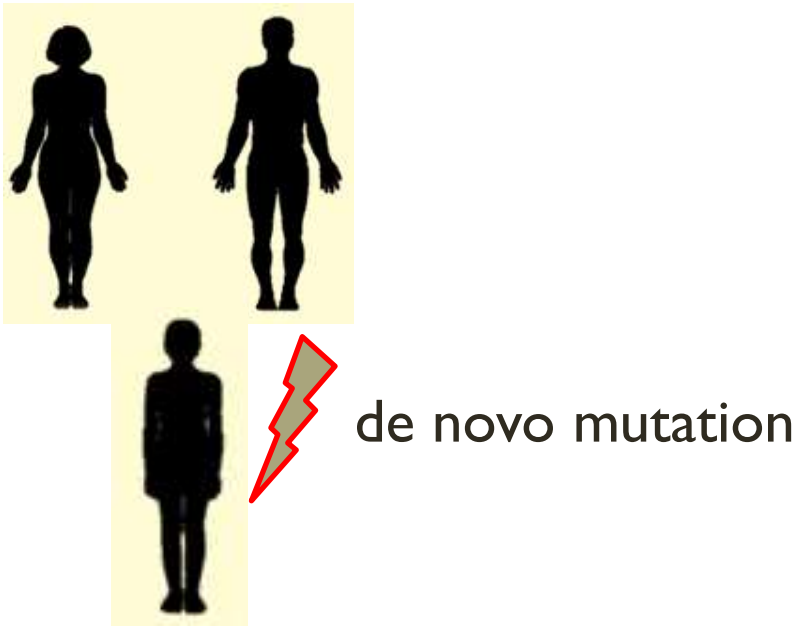
The Contribution of Mosaic Variants to Autism Spectrum Disorder

Donald Freed[1,2], Jonathan Pevsner[1,2,3]*

Somatic mosaic variation in autism



Somatic mosaic variation in autism



de novo mutation

Collections of genotype and phenotype data from individuals with ASD

- Patients at the Kennedy Krieger Institute (50 trios)
- Simons Simplex Collection (SSC)
- MSSNG Project

Large collections of genomic data (e.g. 10,000 genomes) are available to qualified researchers:
"the democratization of science."

Collections of genotype and phenotype data from individuals with ASD

- Patients at the Kennedy Krieger Institute (50 trios)
- Simons Simplex Collection (SSC)
- MSSNG Project

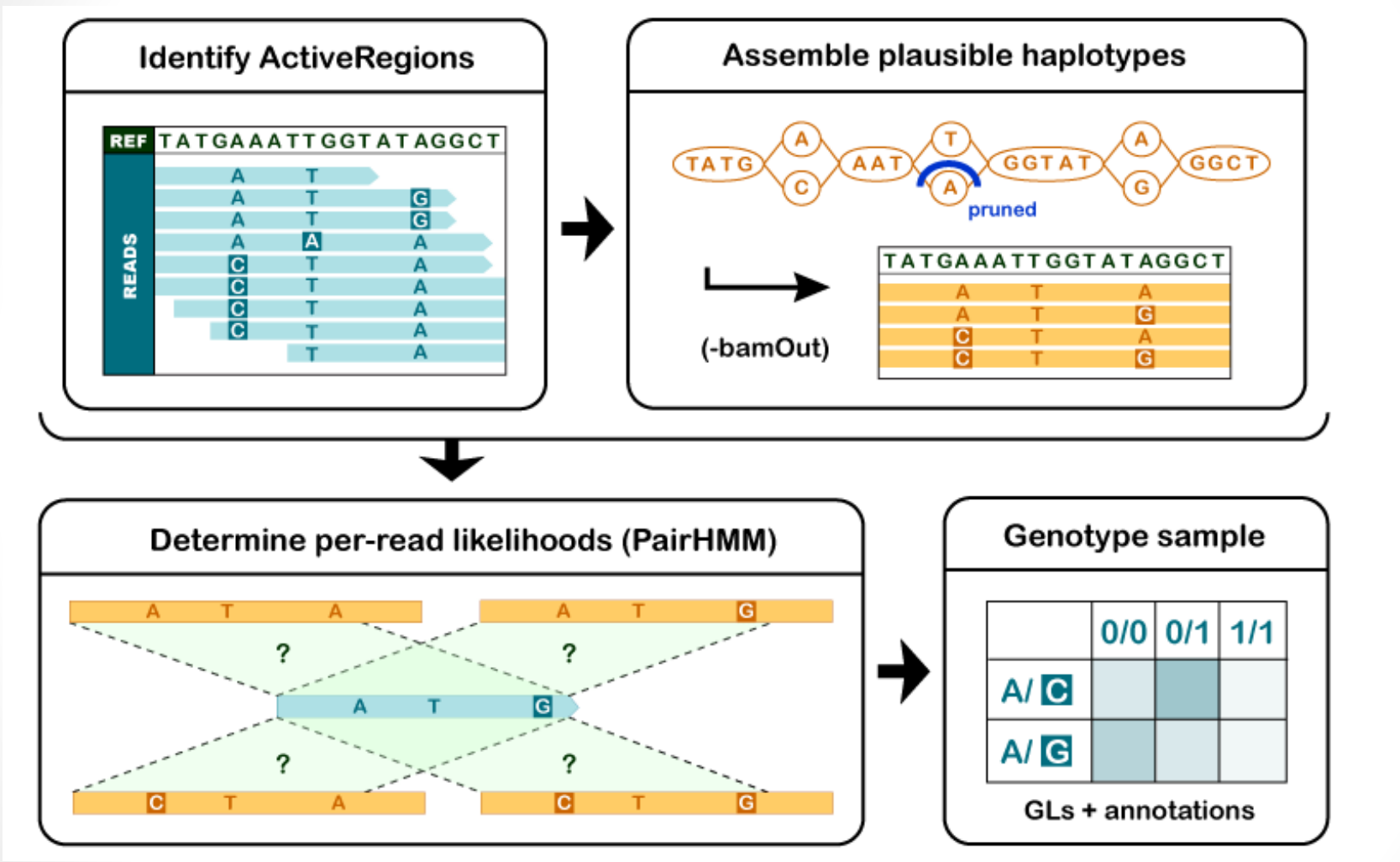
The Simons Simplex Collection (SSC)

- 8,938 individuals
  - 2,388 probands
  - 1,774 siblings
  - 4,776 parents
- Simplex autism diagnoses
- DNA purified from blood
- Whole-exome sequencing on an Illumina platform
- Aligned sequence data publicly available on NDAR / AWS

Methods overview: finding mosaic variants

- GATK pipeline (Genome Analysis Toolkit)
  - Variant calling
  - Genotyping
  - Variant Quality Score Recalibration
- Identification of *de novo* variants
- Variant effect annotation
- Identification of mosaic variants

Variant calling approach: GATK haplotype caller



https://software.broadinstitute.org/gatk/documentation/article?id=4148

Methods: Variant calling via cloud computing

- Amazon EC2 + S3
- Virtual machines
- StarCluster (EC2 toolkit)
- Common bioinformatics tools (e.g. samtools)
- Python applications, R

Methods: Variant calling via cloud computing

AWS S3

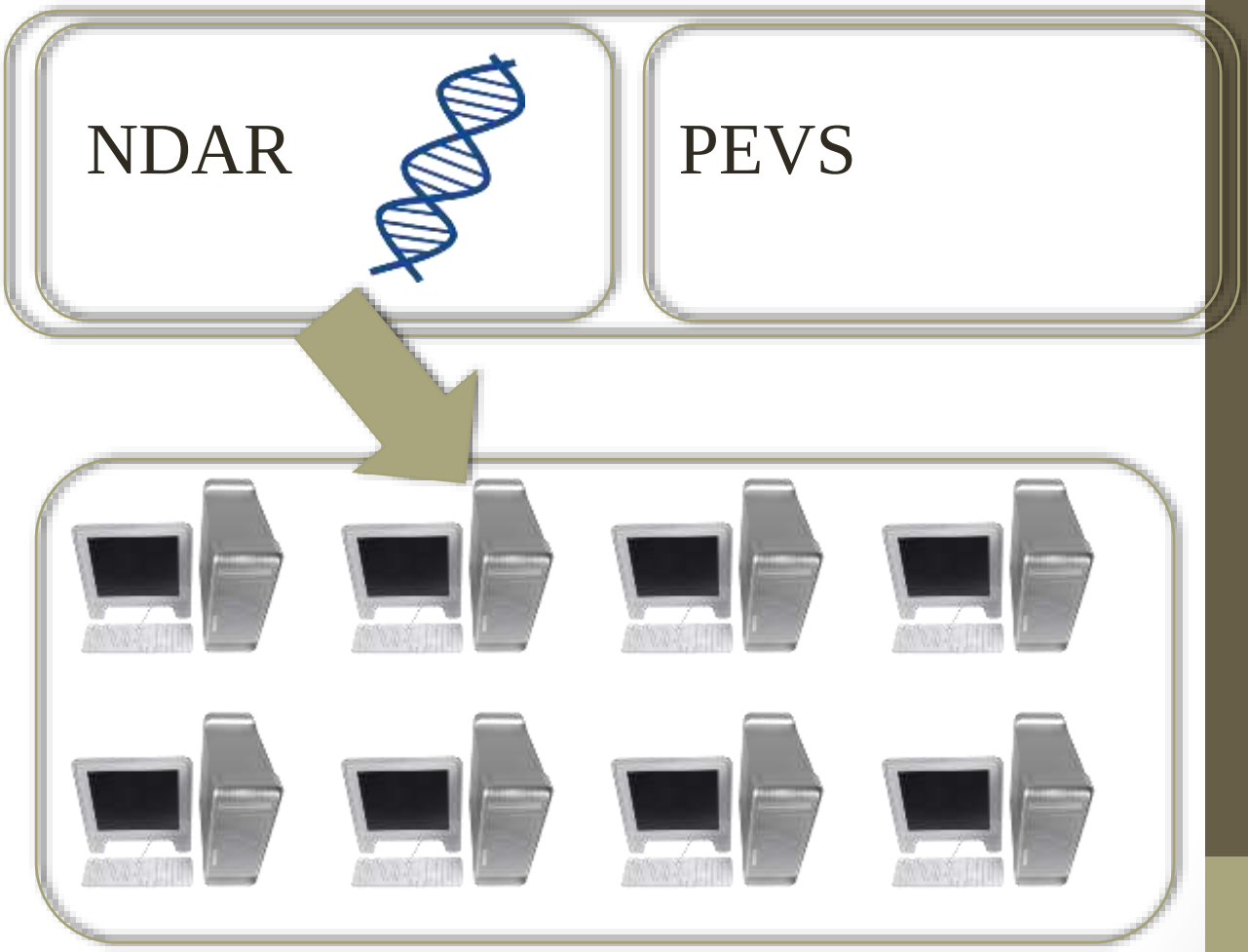
NDAR

PEVS

AWS EC2

Methods: Variant calling via cloud computing

AWS S3

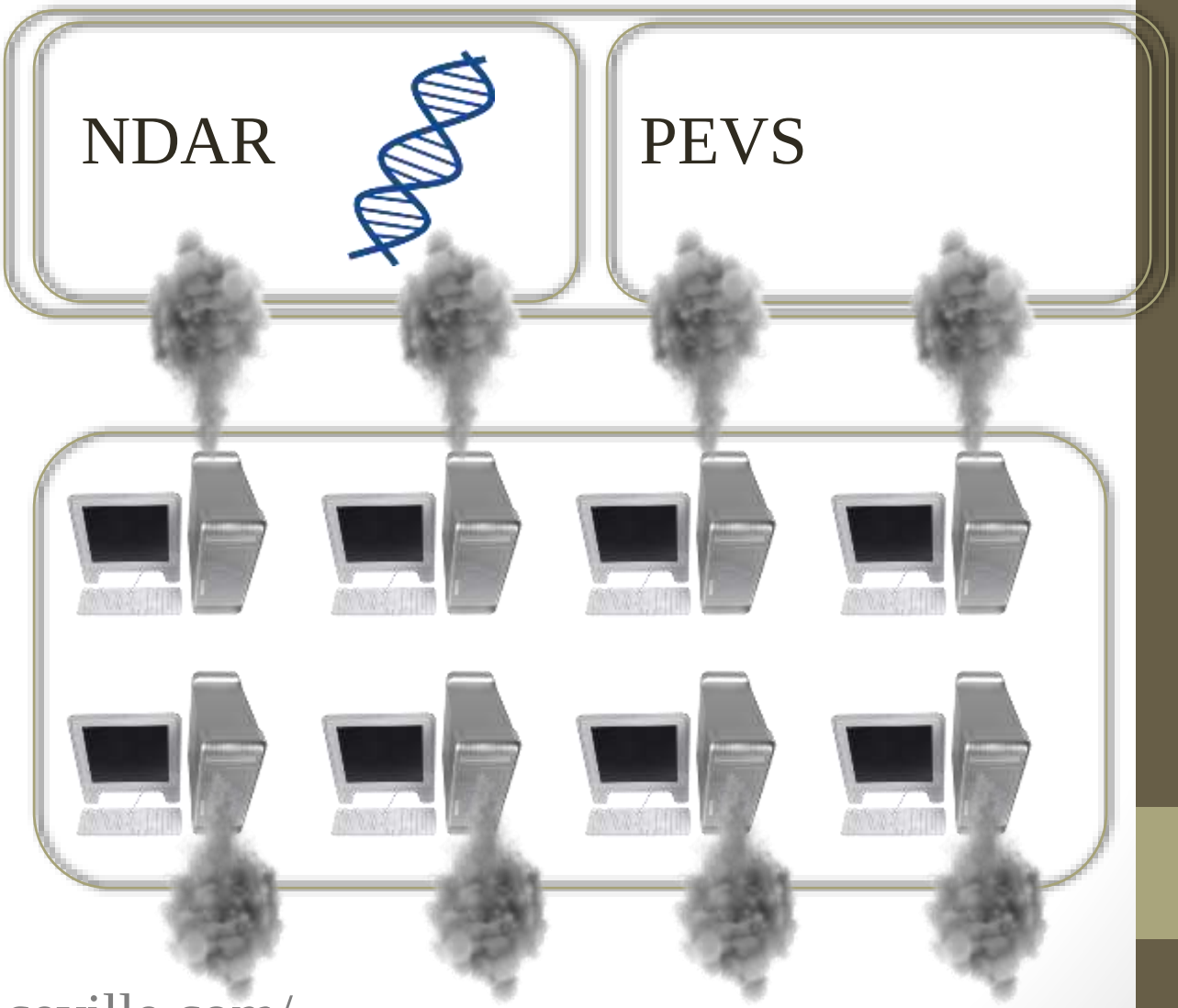


AWS EC2

Methods: Variant calling via cloud computing

AWS S3

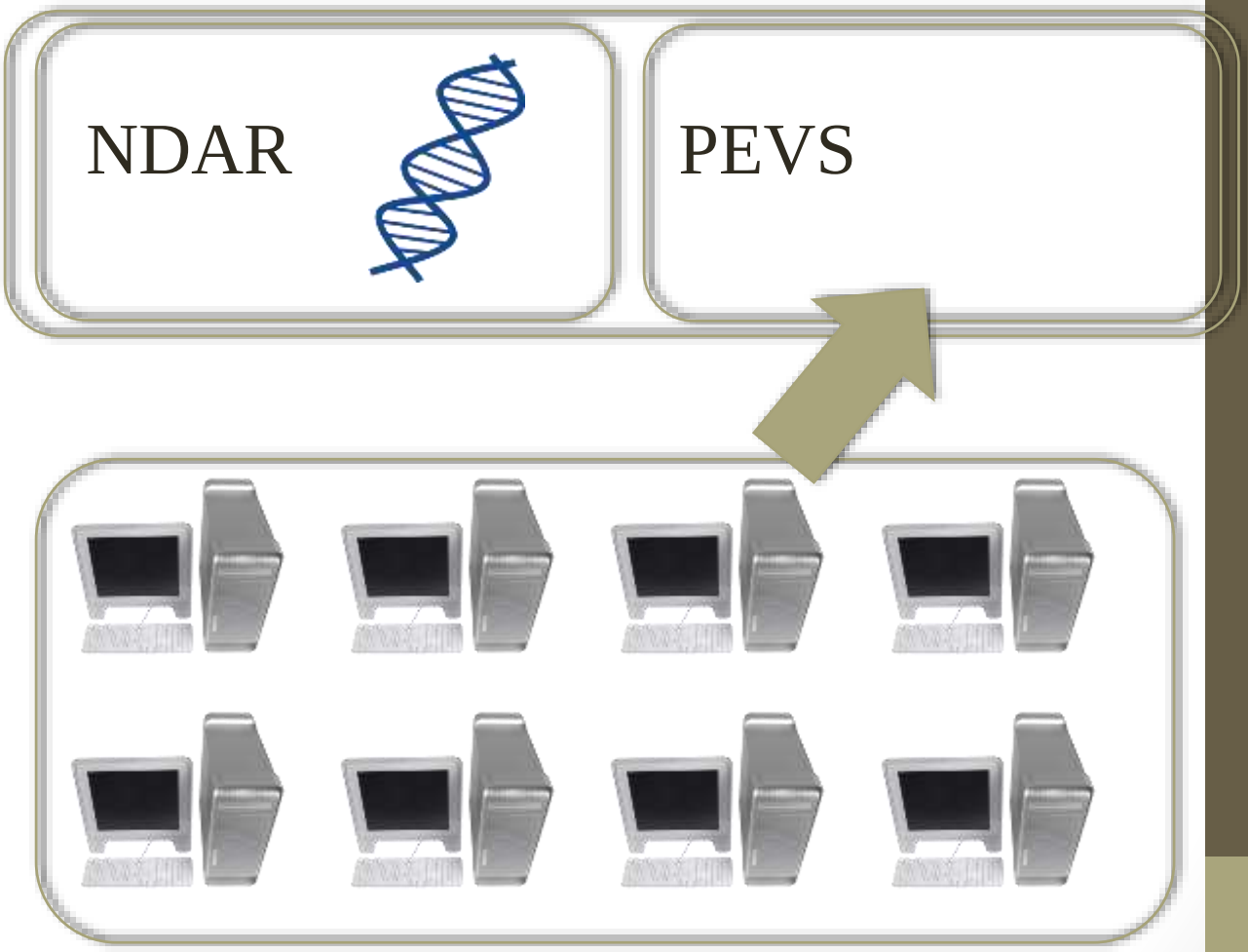
AWS
EC2



http://gallery.yopriceville.com/

Methods: Variant calling via cloud computing

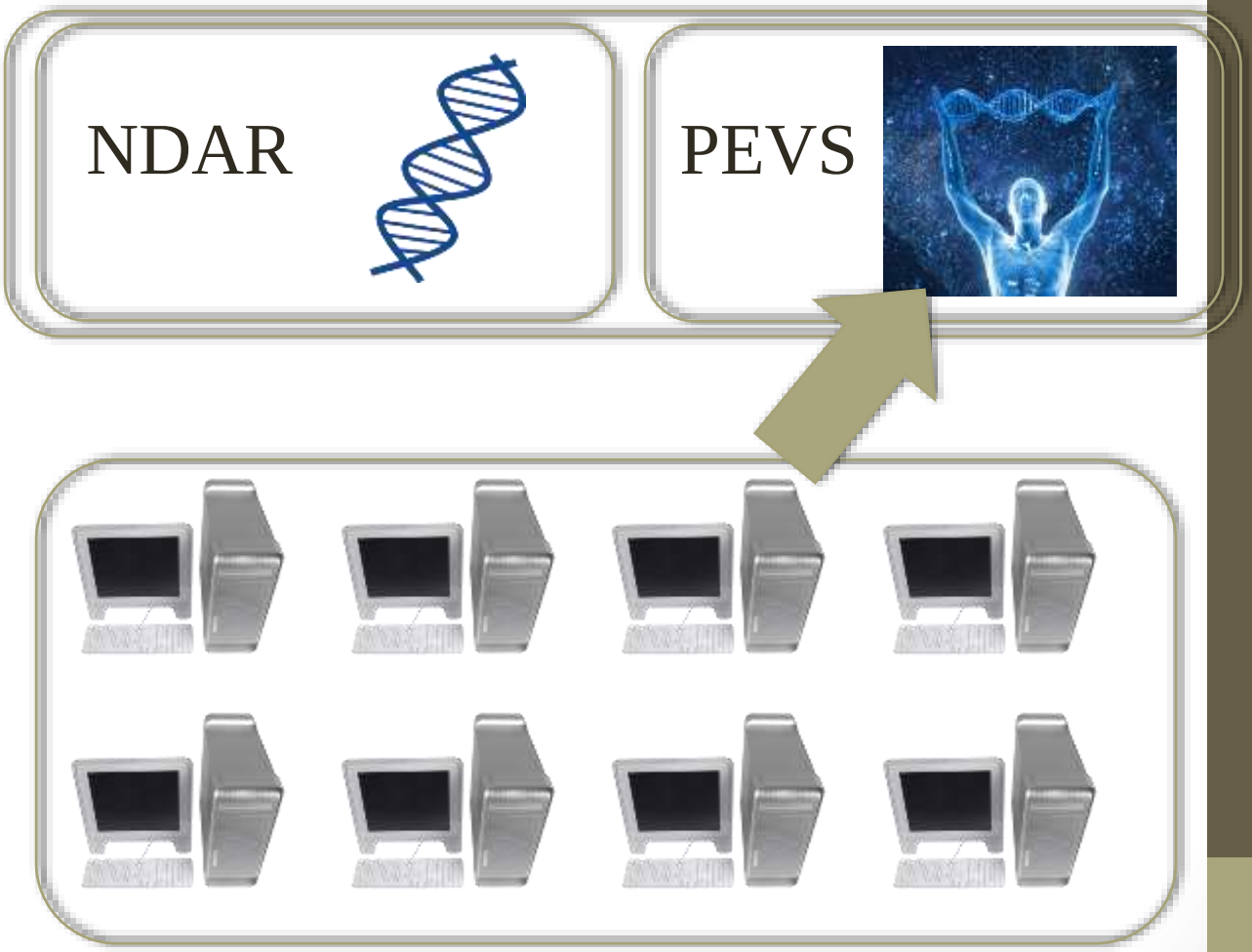
AWS S3



AWS
EC2

Methods: Variant calling via cloud computing

AWS S3



AWS EC2

http://www.livescience.com/topics/dna-genes

Methods overview: finding mosaic variants

- GATK pipeline
  - Variant calling (ploidy 5)
  - Genotyping
  - Variant Quality Score Recalibration
- Identification of *de novo* variants
- Variant effect annotation
- Identification of mosaic variants

Methods: Joint genotyping via cloud computing

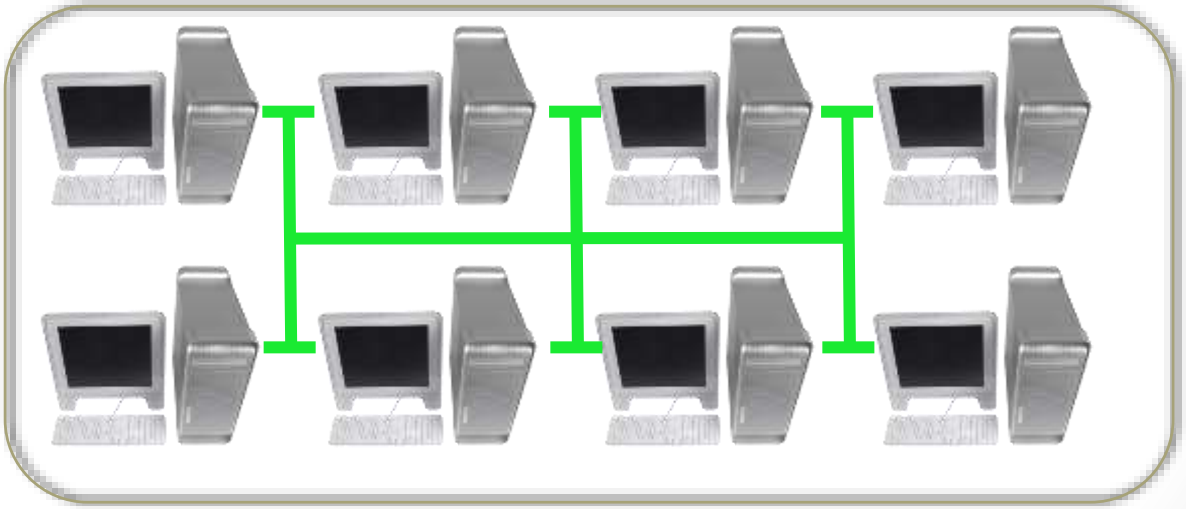
- Variants are called per sample (we want variant information across all samples)
- Joint genotyping assesses all samples in the cohort simultaneously
- Samples are re-assessed for the presence of variants

Methods: Joint genotyping via cloud computing

AWS S3

PEVS

AWS EC2



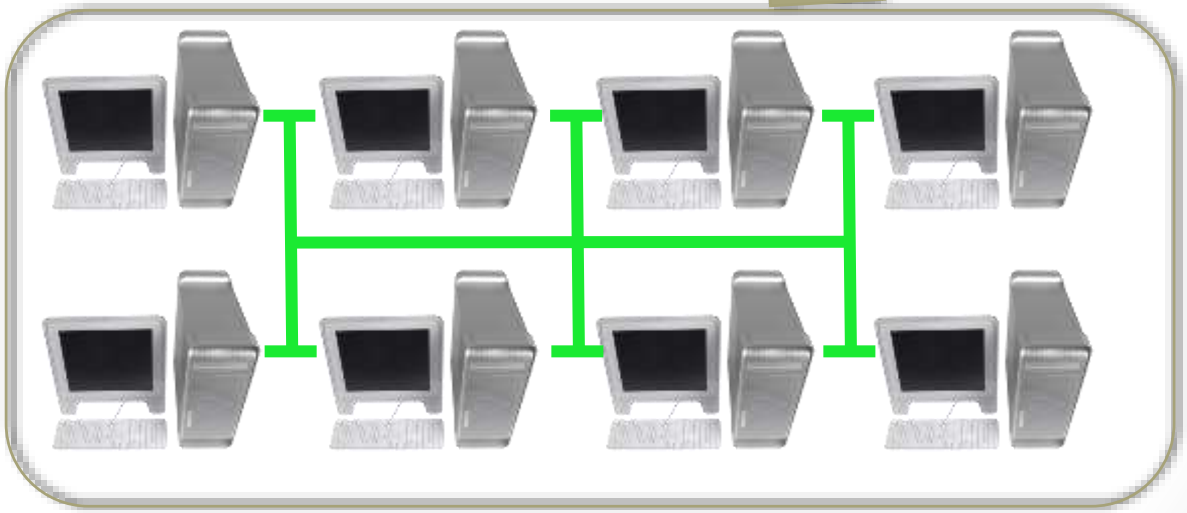
http://www.livescience.com/topics/dna-genes

Methods: Joint genotyping via cloud computing

AWS S3

PEVS

AWS EC2

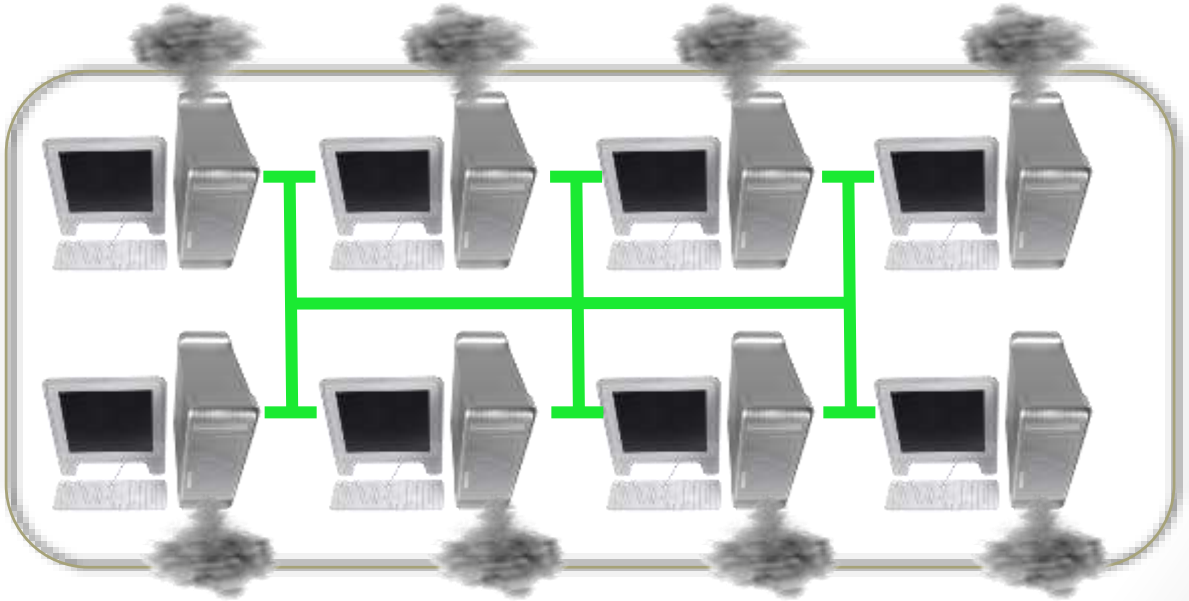


Methods: Joint genotyping via cloud computing

AWS S3

PEVS

AWS EC2

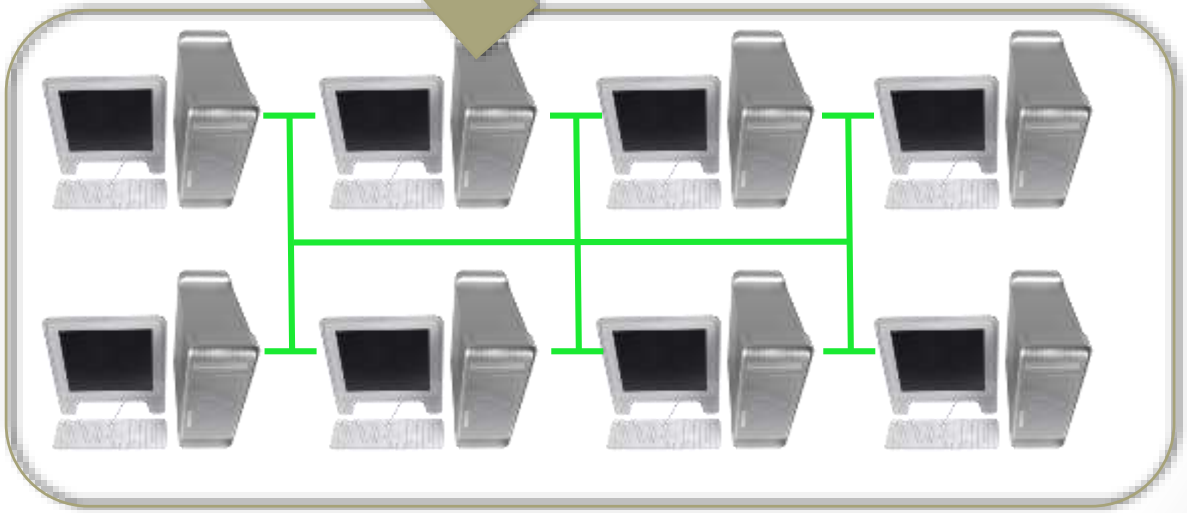


Methods: Joint genotyping via cloud computing

AWS S3

PEVS

AWS
EC2

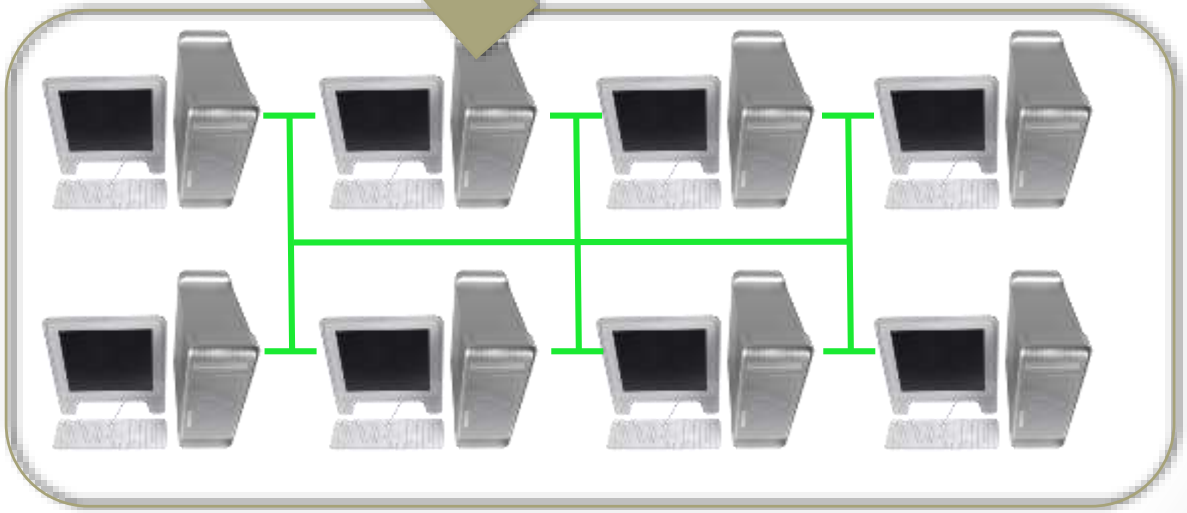


Methods: Joint genotyping via cloud computing

AWS S3

PEVS

AWS
EC2



Methods overview: finding mosaic variants

- GATK pipeline
  - Variant calling (ploidy 5)
  - Genotyping
  - Variant Quality Score Recalibration
- Identification of *de novo* variants
- Variant effect annotation
- Identification of mosaic variants

Variant Quality Score Recalibration

- Variant calling and genotyping are subject to systematic biases
- False positive variants due to these biases can be identified and filtered
  - Machine learning (Gaussian mixture model)
  - Known true positive (and false positive) variants
  - Set sensitivity thresholds

Methods overview: finding mosaic variants

- GATK pipeline
  - Variant calling (ploidy 5)
  - Genotyping
  - Variant Quality Score Recalibration
- Identification of *de novo* variants
- Variant effect annotation
- Identification of mosaic variants

Identification of *De Novo* Variants

- *De novo* variants are present in a child but not either parent

- Identified *de novo* variants using a hard-filter approach
  - Variant present in unrelated sample
  - Read depth (20x)
  - Minimum genotype quality (20)

Methods overview: finding mosaic variants

- GATK pipeline
  - Variant calling (ploidy 5)
  - Genotyping
  - Variant Quality Score Recalibration
- Identification of *de novo* variants
- Variant effect annotation
- Identification of mosaic variants

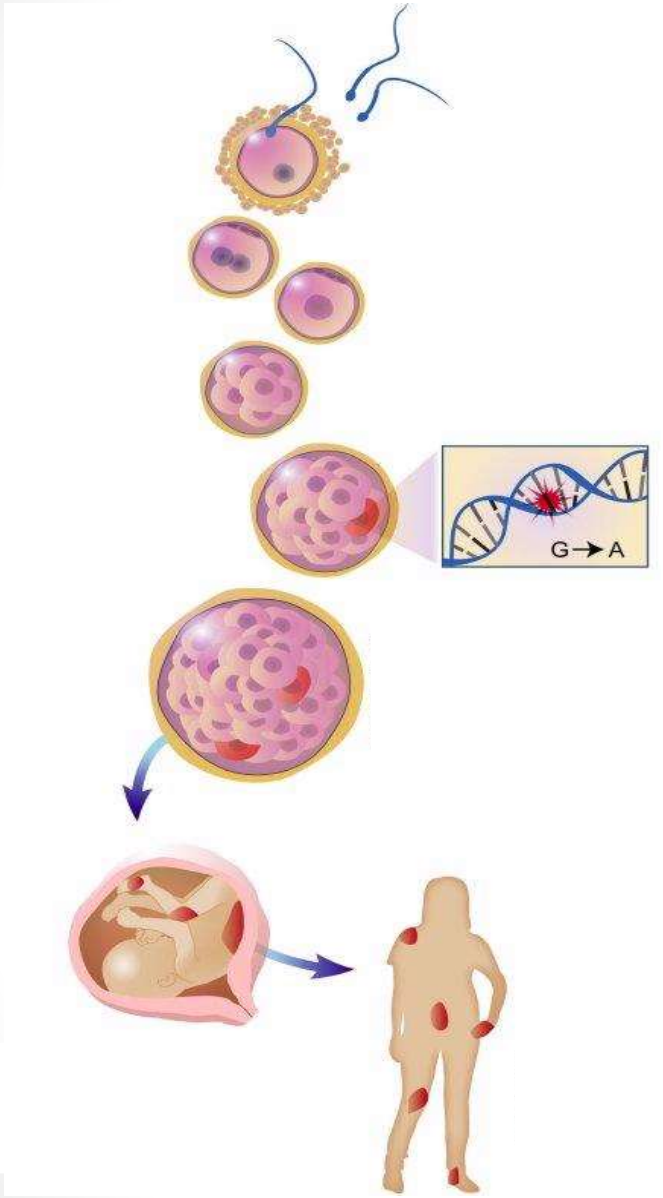
Variant Effect Annotation

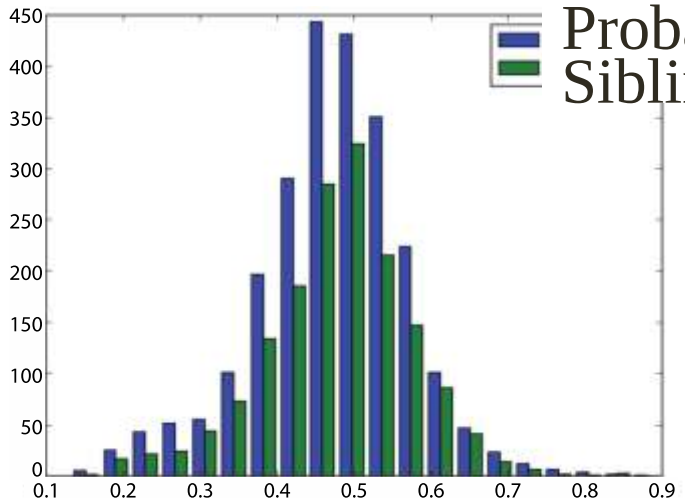
Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	102	1.119%
5_prime_UTR_premature_start_codon_gain_variant	8	0.088%
5_prime_UTR_variant	57	0.625%
disruptive_inframe_deletion	34	0.373%
disruptive_inframe_insertion	6	0.066%
downstream_gene_variant	771	8.458%
frameshift_variant	213	2.337%
inframe_deletion	24	0.263%
inframe_insertion	6	0.066%
intergenic_region	115	1.262%
intragenic_variant	47	0.516%
intron_variant	2,356	25.845%
missense_variant	2,687	29.476%
non_coding_exon_variant	69	0.757%
protein_protein_contact	235	2.578%
splice_acceptor_variant	54	0.592%
splice_donor_variant	62	0.68%
splice_region_variant	364	3.993%
start_lost	4	0.044%
stop_gained	159	1.744%
stop_lost	5	0.055%
synonymous_variant	1,172	12.857%
upstream_gene_variant	566	6.209%

Methods overview: finding mosaic variants

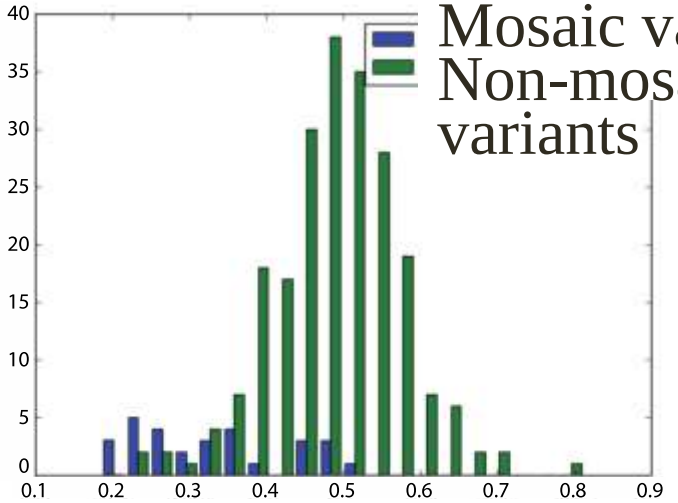
- GATK pipeline
  - Variant calling (ploidy 5)
  - Genotyping
  - Variant Quality Score Recalibration
- Identification of *de novo* variants
- Variant effect annotation
- Identification of mosaic variants

Identifying mosaic variants





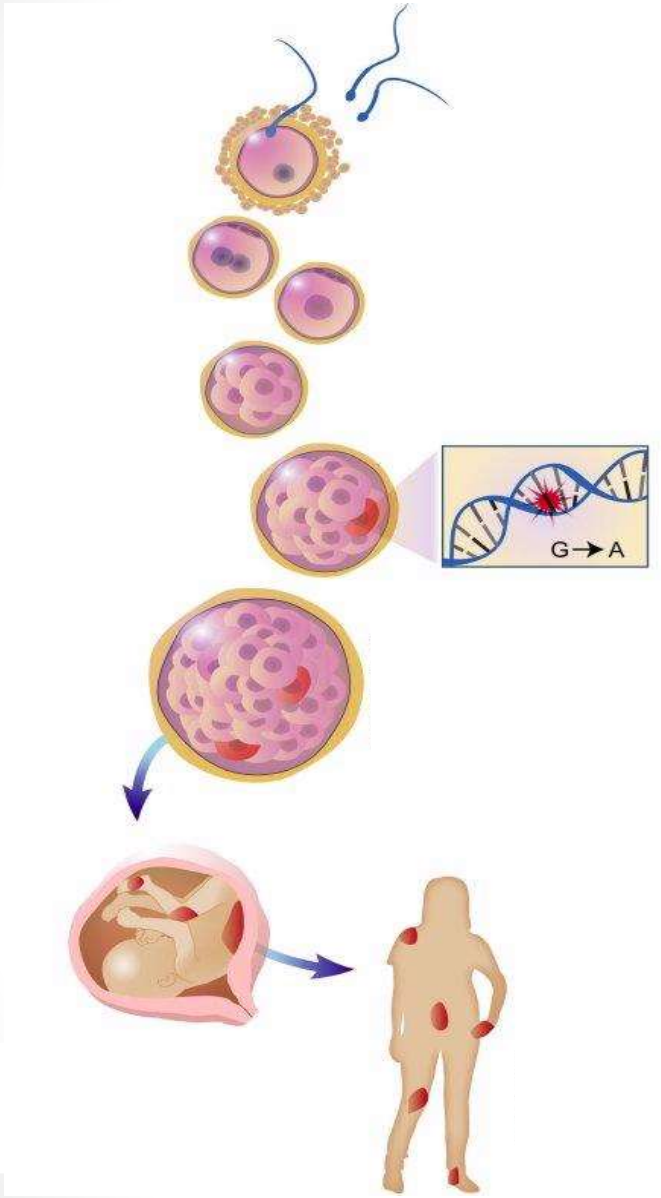
Alternate allele read frequency

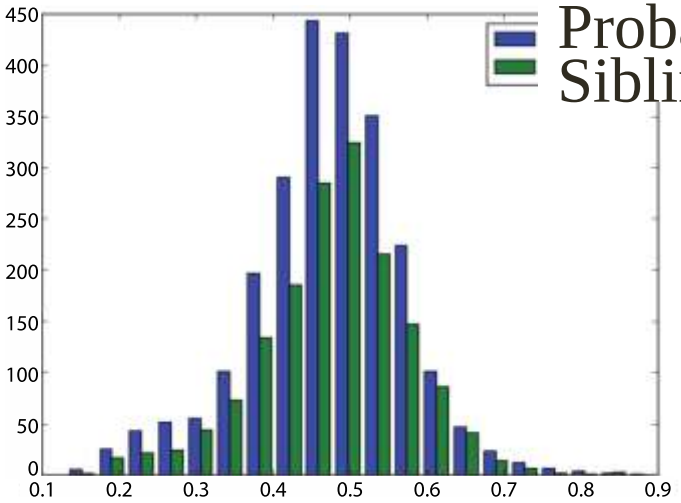


Alternate allele read frequency

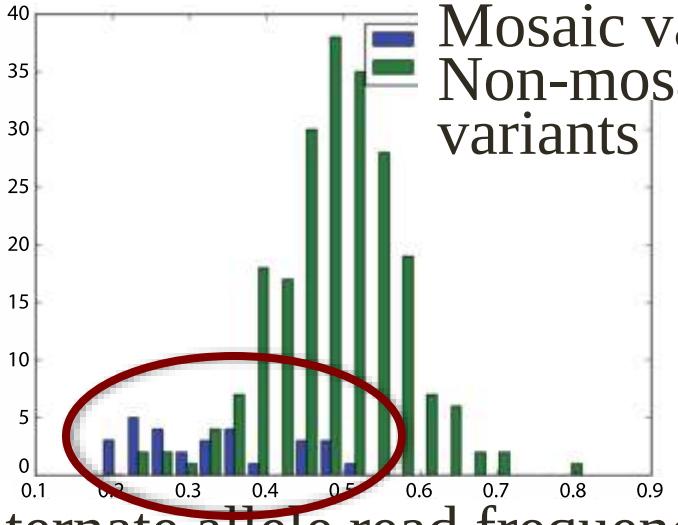
https://www.genome.gov/imagegallery/

Identifying mosaic variants





Alternate allele read frequency

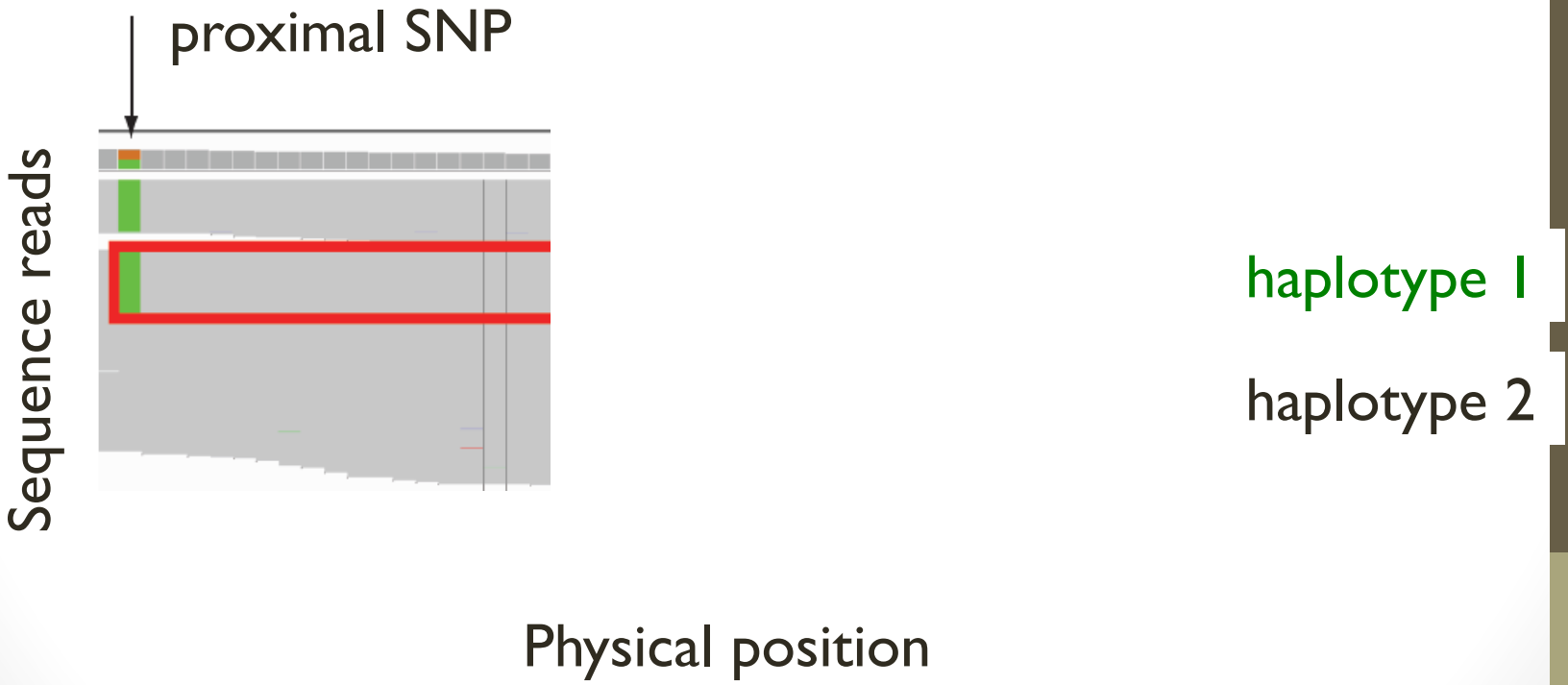


Alternate allele read frequency

https://www.genome.gov/imagegallery/

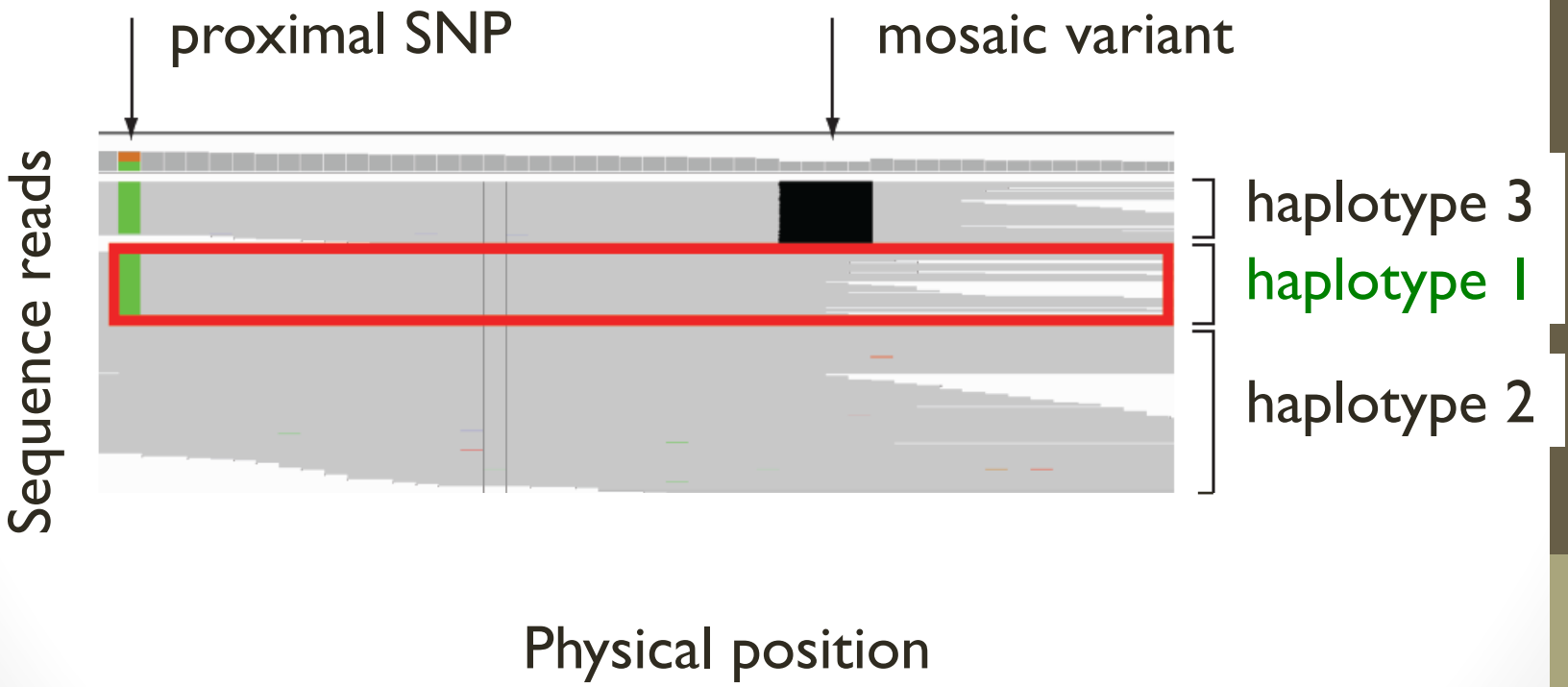
Validating mosaic variants by phasing

We developed a workflow to identify high quality candidates from sequence data. We also developed methods to validate somatic variants by phasing.



Validating mosaic variants by phasing

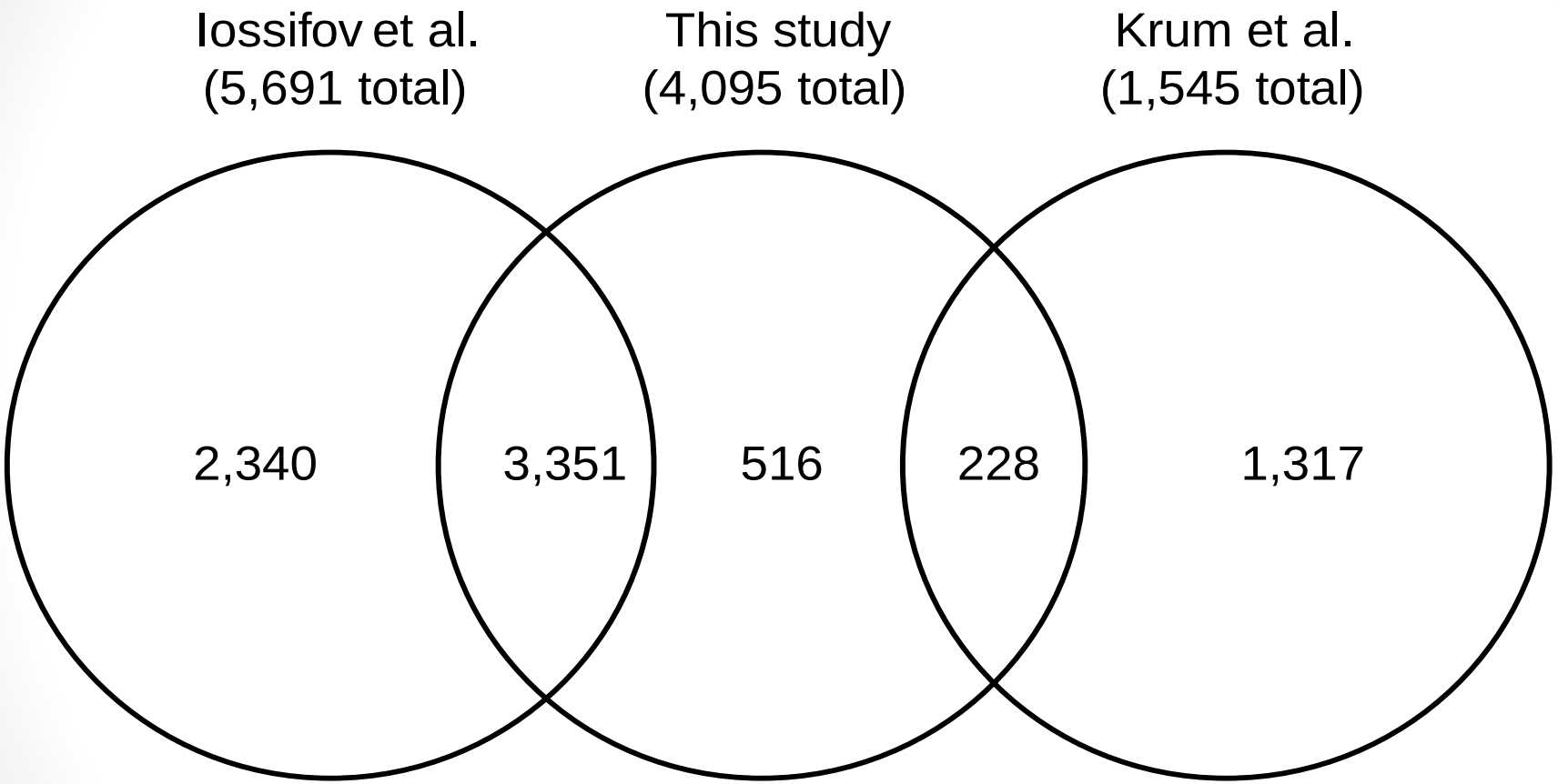
We developed a workflow to identify high quality candidates from sequence data. We also developed methods to validate somatic variants by phasing.



Identifying mosaic variants

- Binomial test
  - False discovery protection with FDR of 0.05
- Additional filters
  - Mosaic variants must be in Krumm or Iossifov
  - Mosaic variants must have AARF of < 0.34
- Callset metrics
  - 100% precision for variant presence
  - 85% precision for mosaic status

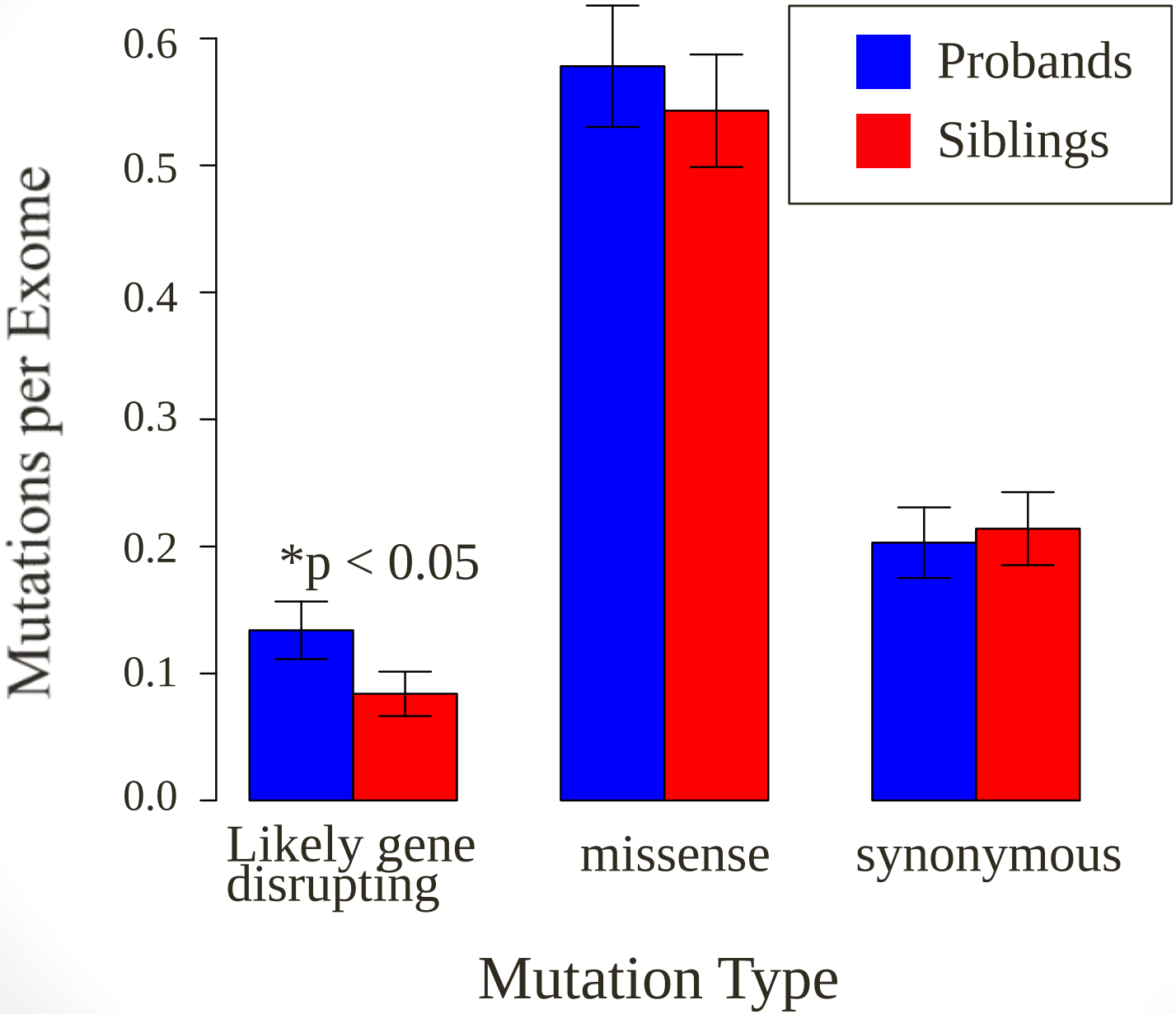
De novo calls: comparision two recent studies



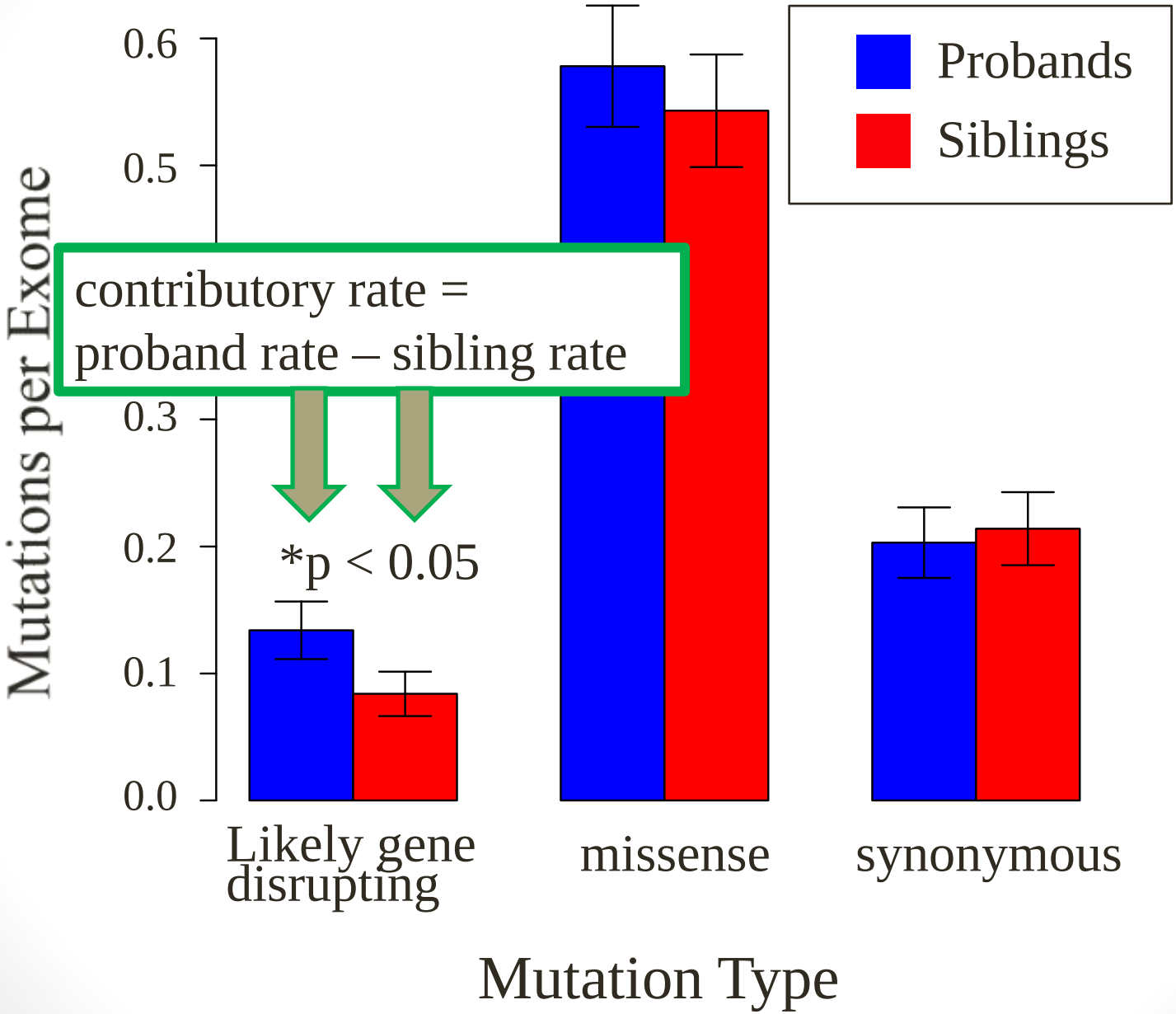
Analysis of mutation rates

- Compare probands and siblings within the same family
- Increased mutation burden indicates a "contributory" role in disease
  - Rate = number of mutations per exome
  - contributory rate = proband rate – sibling rate
  - % contributory = contributory rate / proband rate
- Only mutations at 40x sites in the trio
- Rates normalized to the entire capture target

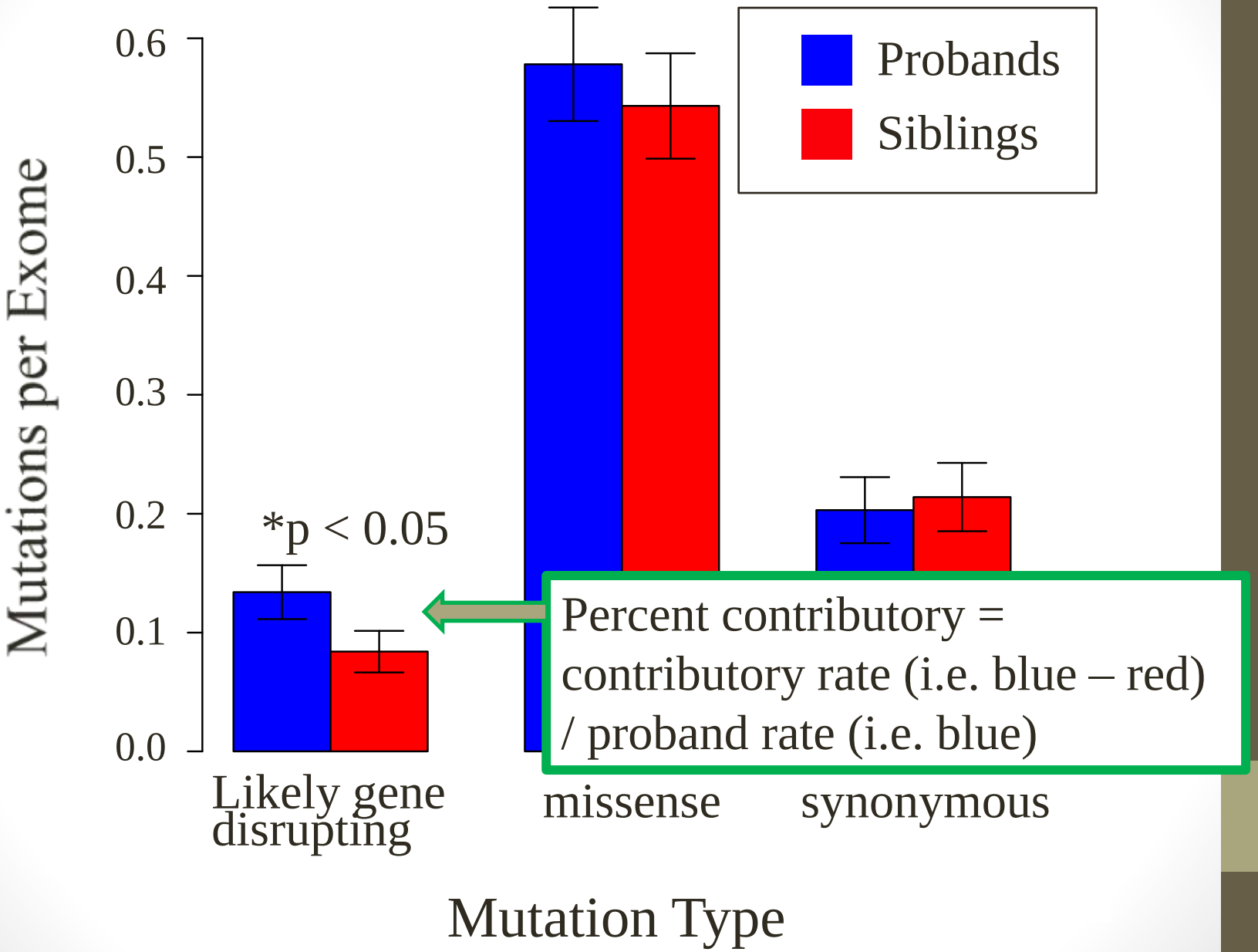
Rates of germline *de novo* mutation



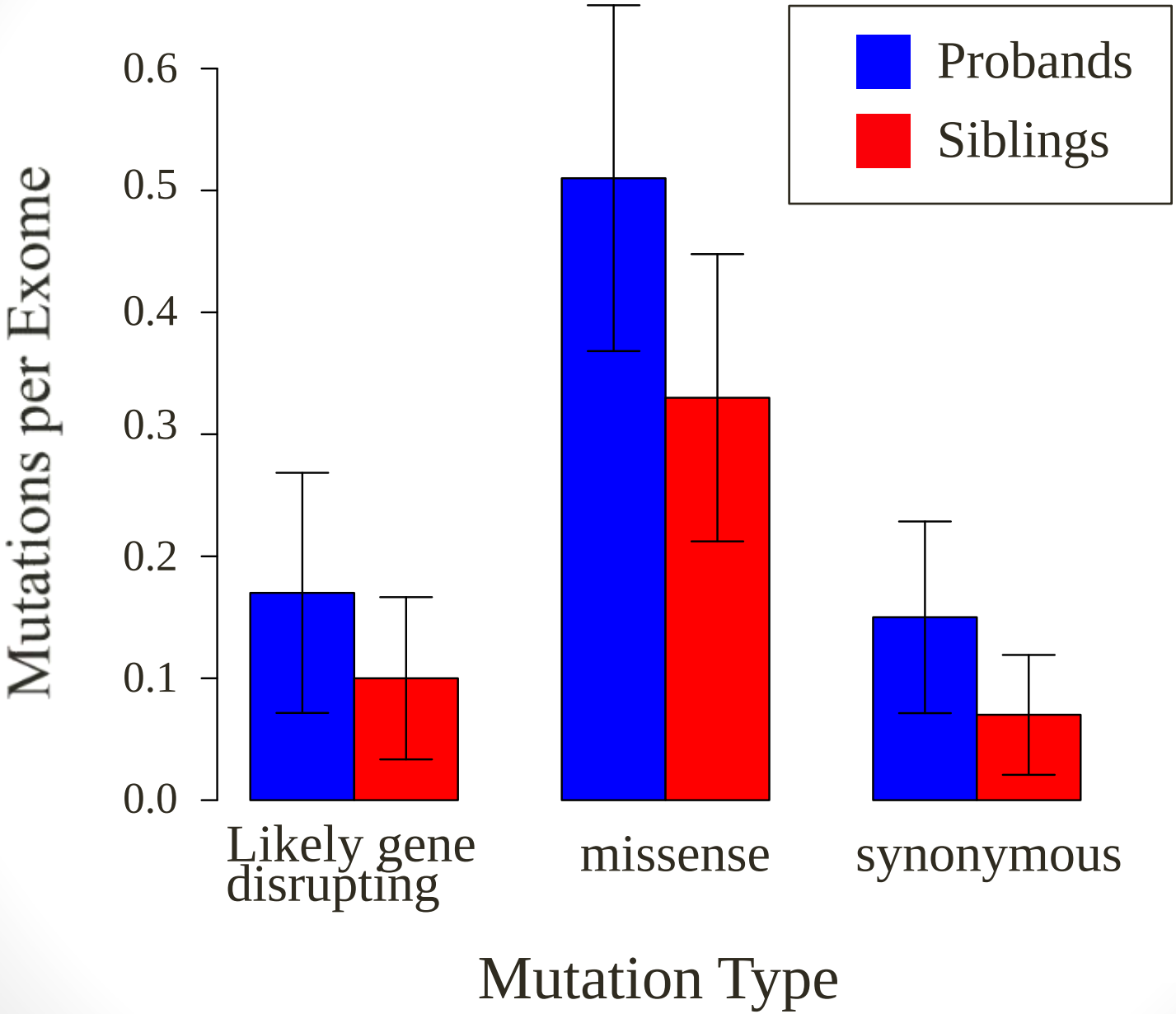
Rates of germline *de novo* mutation



Rates of germline *de novo* mutation



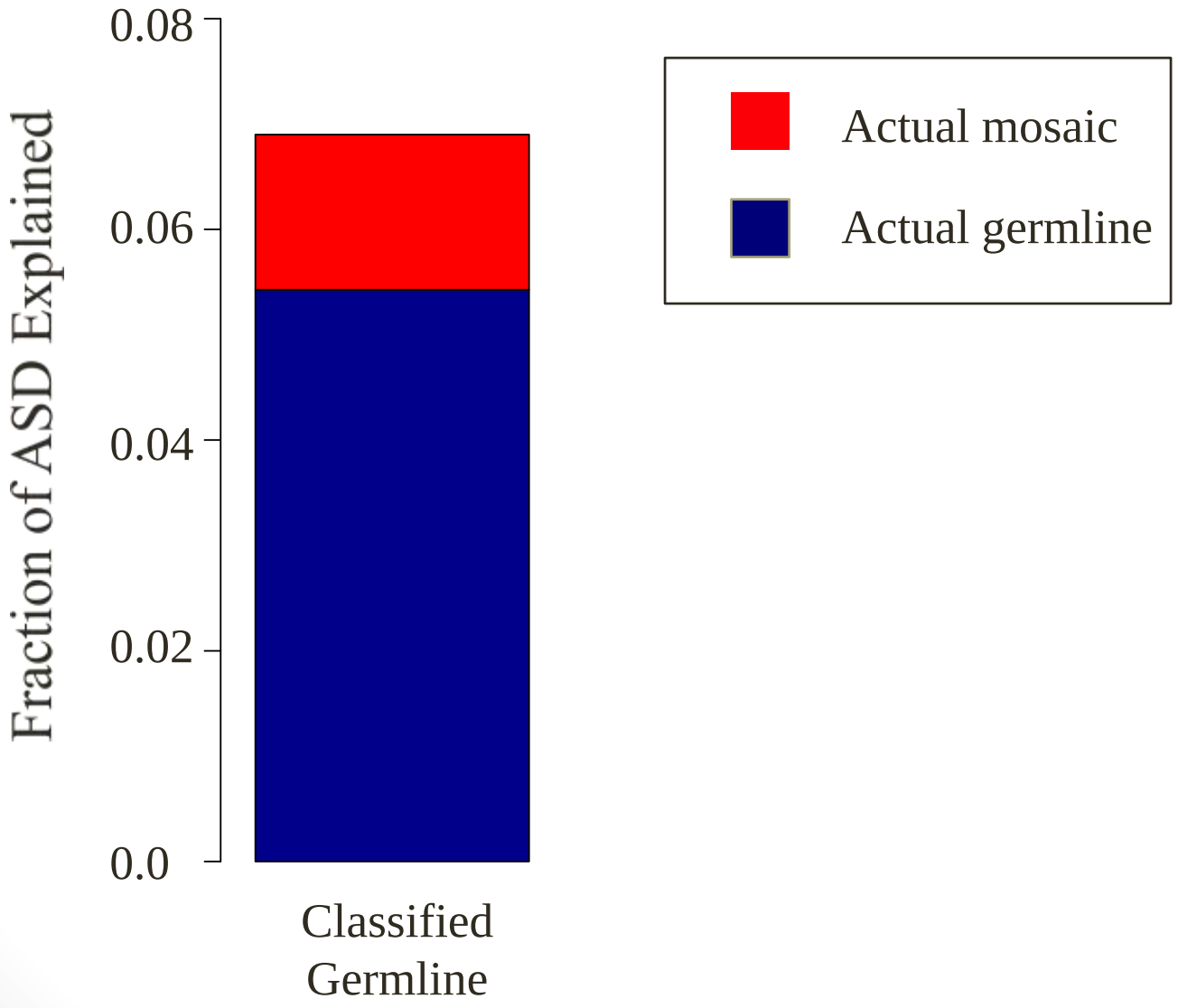
Rates of mosaic mutation



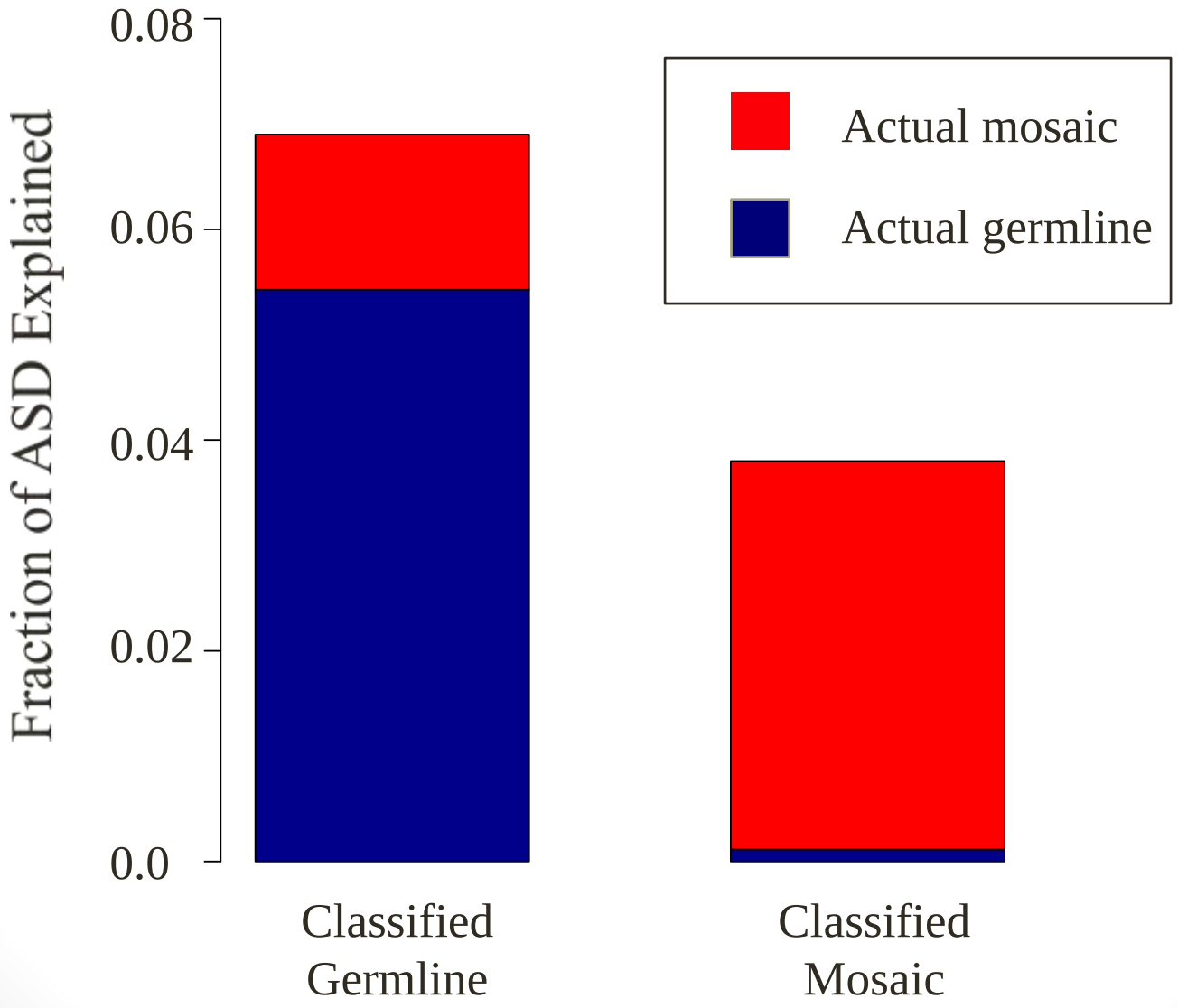
Modeling contributory variation: error rates

- Classified mosaic mutations are a mix of mosaic and germline *de novo* events
- Same for classified germline *de novo*
- What is the contribution of incorrectly classified variants?
- Model parameters
  - Errors in classification of mosaic status
  - Validation rates
  - Number of germline and mosaic mutations

Modeling contributory variation: error rates



Modeling contributory variation: error rates



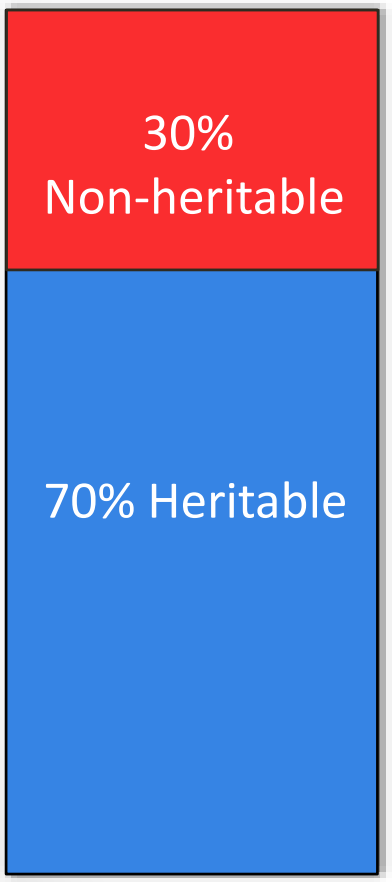
Modeling contributory variation: error rates

- The contribution of classified mosaic variants is primarily due to mosaic variation
- Some contribution of classified germline variants comes from mosaic variation

33% of mosaic variants contribute to 5.1% of ASD cases

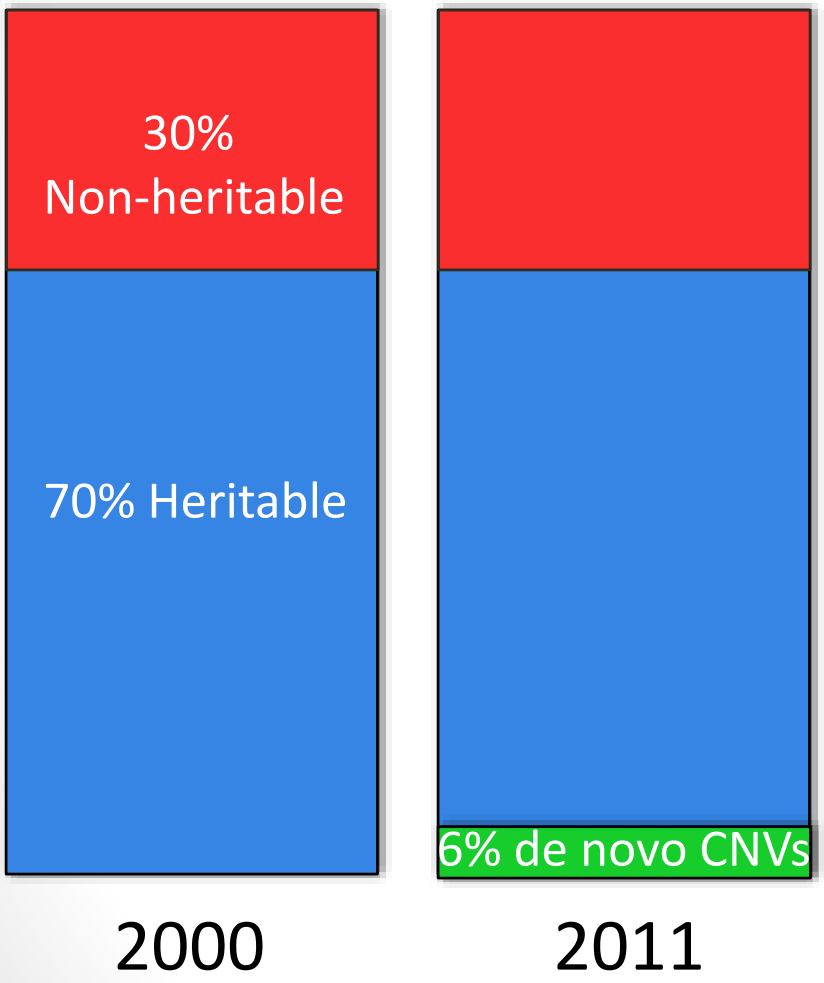
6% of germline variants contribute to 5.6% of ASD cases

Understanding the genetic architecture of autism spectrum disorder

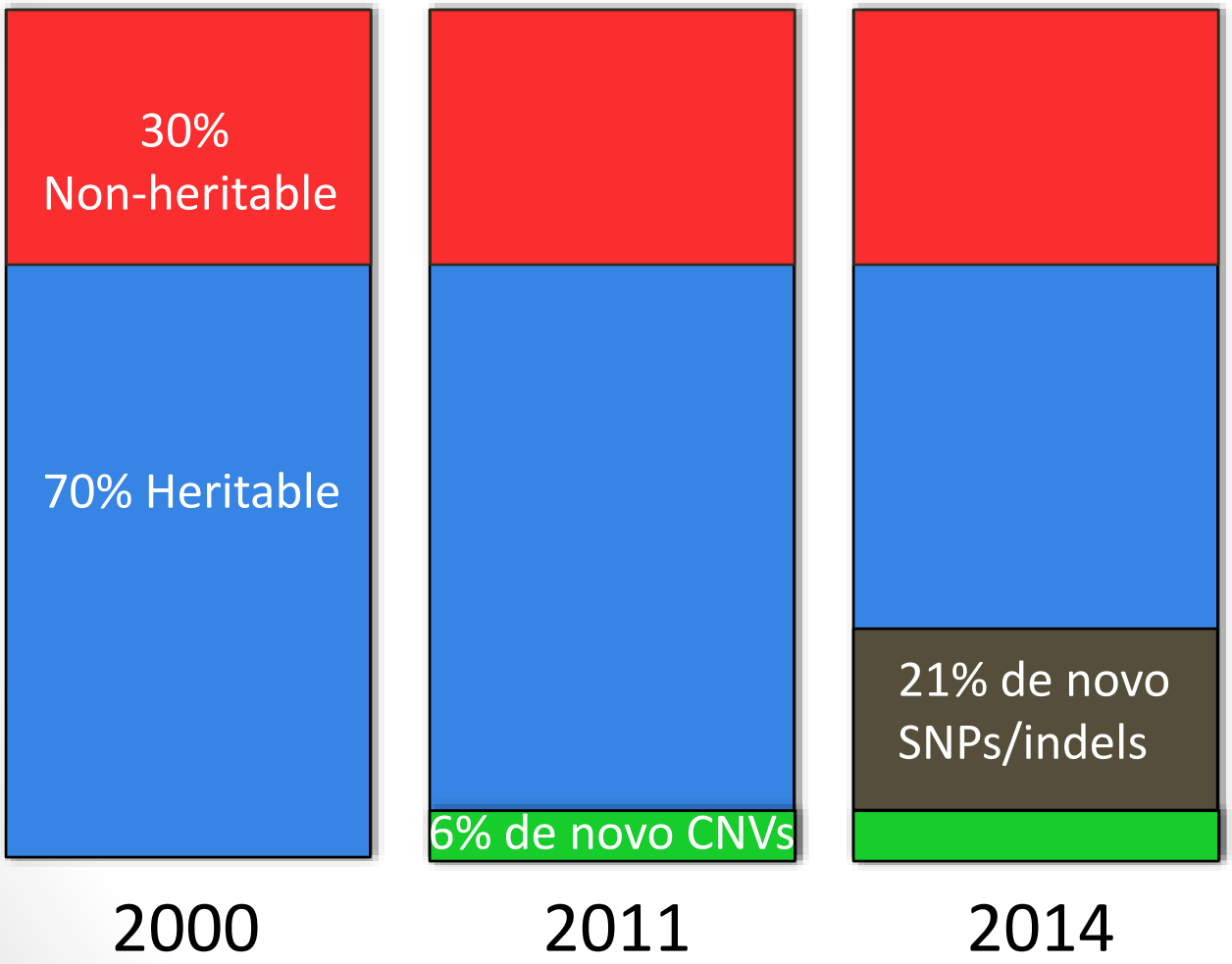


2000

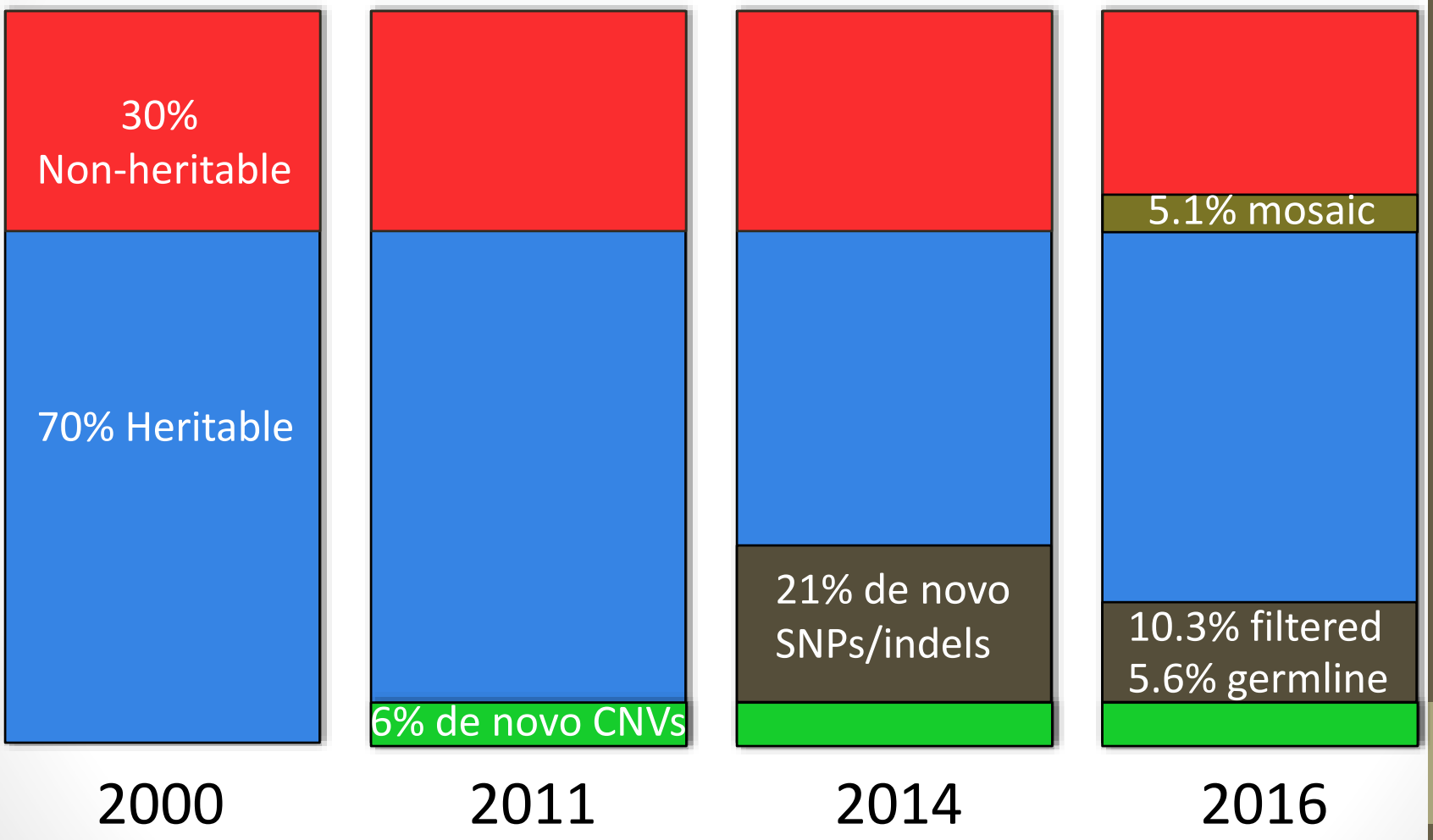
Understanding the genetic architecture of autism spectrum disorder



Understanding the genetic architecture of autism spectrum disorder



Understanding the genetic architecture of autism spectrum disorder



Conclusions

- We identified many mosaic mutations (221 of ~4000 de novo mutations, i.e. 5.4%).
- Mosaic mutations were significantly enriched in probands relative to siblings and contribute to ~5% of simplex autism spectrum disorder diagnoses.
- We did not detect mosaic variants in paired brain/heart samples, at our level of detection.
- Mosaic variation may contribute to other neuropsychiatric disorders.

Acknowledgments

Pevsner lab
Matt Shirley (now at Novartis)
Larry Frelin
Donald Freed (graduate student)

Alexis Norris, Jeremy Thorpe, Ike Adeshina, Kyra Feuer, McKinzie Garrison

Collaborators: Sturge-Weber syndrome
Anne Comi (KKI)
Doug Marchuk and Hao Tang, Carol Gallione (Duke)
Bernard Cohen (JHU) Paula North (Wisconsin)