FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS
DEPARTAMENTO DE BIOTECNOLOGÍA Y TECNOLOGÍA ALIMENTARIA
UNIVERSIDAD ARGENTINA DE LA EMPRESA

Bioinformática

ANÁLISIS COMPUTACIONAL DE SECUENCIAS

Dr. Lucas L. Maldonado (PhD)

Lic. Biotechnologist and Molecular Biologist

Bioinformatics and genomics specialist

CONICET
Fac. de Medicina - UBA
Fac. de Ciencias Exactas y Naturales – UBA

lucamaldonado@uade.edu.ar
lmaldonado@fmed.uba.ar
lucas.l.maldonado@gmail.com.ar

Comparative Genomics

Comparative Genomics

- **What is comparative genomics?**
- Analyzing & comparing genetic material from different species to study:
    - Evolution
    - gene function
    - inherited disease
    - Phenotype differences
    - Anthropology and population genetics
    - Understand the uniqueness between different species
    - etc

Comparative Genomics

What is Comparative Genomics?

Comparative genomics is the analysis and comparison of genomes from different species.

- The purpose is to gain a better understanding of how species have evolved and to determine the function of genes and noncoding regions of the genome.
- DNA sequences that have been "conserved" (preserved in many different organisms over millions of years) - is an important step toward understanding the genome itself.
- It pinpoints genes that are essential to life and highlights genomic signals that control gene function across many species.
- Researchers have learned a great deal about the function of human genes by examining their counterparts in simpler model organisms such as the mouse.
- It involves the use of computer programs that can line up multiple genomes and look for regions of similarity among them.
- As DNA sequencing technology becomes more powerful and less expensive, comparative genomics is finding wider applications in agriculture, biotechnology, zoology and medicine

Comparative Genomics

The major principles of comparative genomics are:

- Common features of two organisms will often be encoded within the DNA that is conserved between the species.
- the DNA sequences encoding the proteins and RNAs responsible for functions that were conserved from the last common ancestor should be preserved in contemporary genome sequences.
- the DNA sequences controlling the expression of genes that are regulated similarly in two related species should also be conserved.
- sequences that encode (or control the expression of) proteins and RNAs responsible for differences between species will themselves be divergent.

Comparative Genomics

What are the comparative genome sizes of humans and other organisms being studied?

- What other genomes have been sequenced?

U.S. National Library of Medicine | NCBI National Center for Biotechnology Information

Genome > Genome Information by Organism

Organism name (com...

Overview (56847); Eukaryotes (13267); Prokaryotes (283957); Viruses (41502); Plasmids (24317); Organelles (17537)

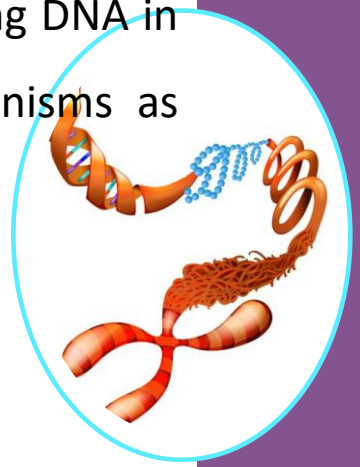
Choose Columns | Page 1 of 285 | 200

#	Organism Name	Organism Groups	Size(Mb)	Chromos
1	Glypta fumiferanae ichnovirus	Viruses;Other;Polydnaviridae	0.291597	105
2	Paralithodes platypus	Eukaryota;Animals;Other Animals	4805,18	99
3	Petromyzon marinus	Eukaryota;Animals;Fishes	1089,05	85
4	Entosphenus tridentatus	Eukaryota;Animals;Fishes	982.705	83
5	Eriocheir sinensis	Eukaryota;Animals;Other Animals	1272,14	72
6	Acipenser ruthenus	Eukaryota;Animals;Fishes	1830,5	60
7	Carassius auratus	Eukaryota;Animals;Fishes	1820,64	59
8	Hyposoter fugitivus ichnovirus	Viruses;Other;Polydnaviridae	0.246092	56
9	Chiloscyllium plagiosum	Eukaryota;Animals;Fishes	3776,55	51
10	Thymallus thymallus	Eukaryota;Animals;Fishes	1564,83	51
11	Cyprinus carpio	Eukaryota;Animals;Fishes	1713,66	50
12	Amblyraja radiata	Eukaryota;Animals;Fishes	2558,78	49
13	Pristis pectinata	Eukaryota;Animals;Fishes	2267,86	46
14	Catharus ustulatus	Eukaryota;Animals;Birds	1131,62	42
15	Bucorvus abyssinicus	Eukaryota;Animals;Birds	1132,6	41
16	Coregonus	Eukaryota;Animals;Fishes	2068,07	40
17	Lycaon pictus	Eukaryota;Animals;Mammals	2358,14	40
18	Salmo trutta	Eukaryota;Animals;Fishes	2371,88	40
19	Canis lupus familiaris	Eukaryota;Animals;Mammals	2544,13	39
20	Salvelinus	Eukaryota;Animals;Fishes	2169,55	39
21	Nyctibius grandis	Eukaryota;Animals;Birds	1256,41	38
22	Camelus dromedarius	Eukaryota;Animals;Mammals	2169,36	37
23	Camelus ferus	Eukaryota;Animals;Mammals	2087,09	37
24	Erithacus rubecula	Eukaryota;Animals;Birds	1086,74	37
25	Leishmania peruviana	Eukaryota;Protists;Kinetoplasts	32.9078	37
26	Oncorhynchus keta	Eukaryota;Animals;Fishes	1853,1	37
27	Aythya fuligula	Eukaryota;Animals;Birds	1127	36
28	Corvus moneduloides	Eukaryota;Animals;Birds	1112,73	36
29	Endotrypanum monterogeii	Eukaryota;Protists;Kinetoplasts	32.5204	36
30	Leishmania aethiopica	Eukaryota;Protists;Kinetoplasts	31.6308	36
31	Leishmania arabica	Eukaryota;Protists;Kinetoplasts	31.2691	36
32	Leishmania chagasi	Eukaryota;Protists;Kinetoplasts	31.925	36
33	Leishmania donovani	Eukaryota;Protists;Kinetoplasts	32.445	36

Comparative Genomics

What can be compared?

Genome researchers look at many different features when comparing genomes: sequence similarity, gene location, the length and number of coding regions (called exons) within genes, the amount of noncoding DNA in each genome, and highly conserved regions maintained in organisms as simple as bacteria and as complex as humans.



- Gene location
- Gene structure
  - Exon number
  - Exon lengths
  - Intron lengths
  - Sequence similarity
- Gene characteristics
  - Splice sites
  - Codon usage
  - Conserved synteny

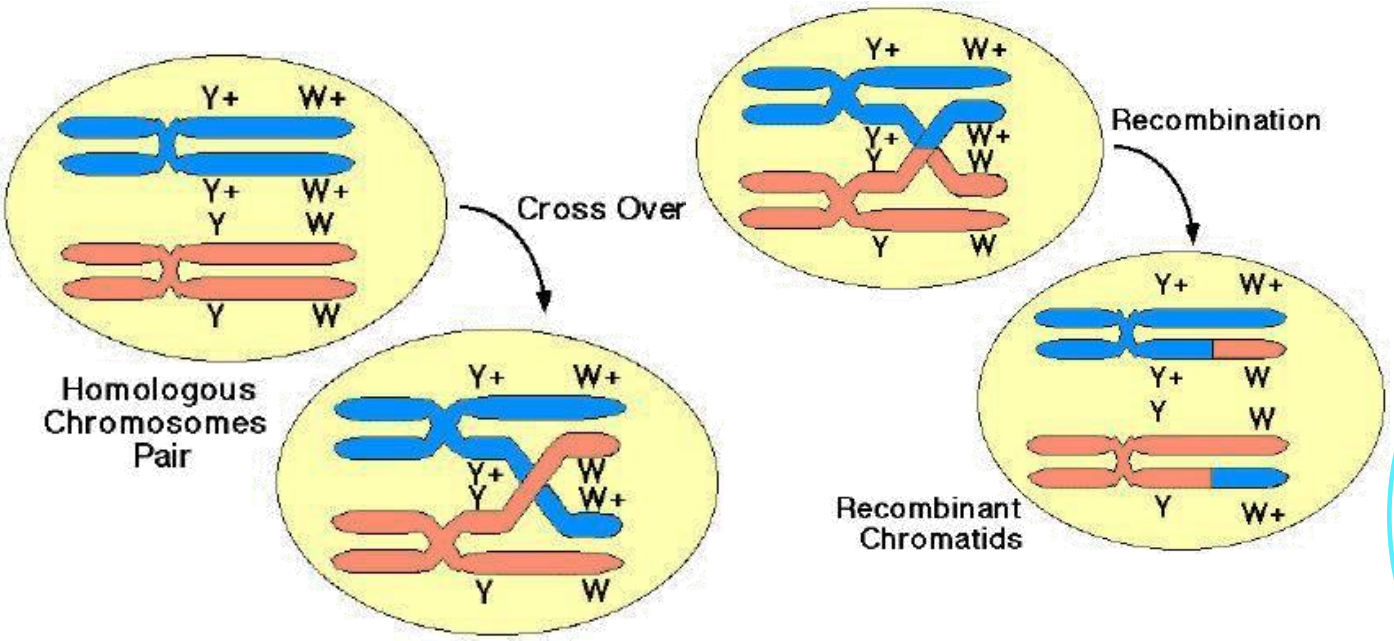
Important observations with regard to Gene Order

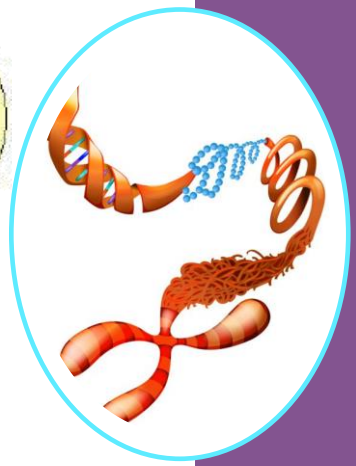
- Order is highly conserved in closely related species but gets changed by rearrangements
- With more evolutionary distance, no correspondence between the gene order of orthologous genes
- Group of genes having similar biochemical function tend to remain localized

Synteny

- Refers to regions of two genomes that show considerable similarity in terms of
  - Sequence conservation
  - conservation of the order of genes
- likely to be related by common descent.

Comparative Genomics



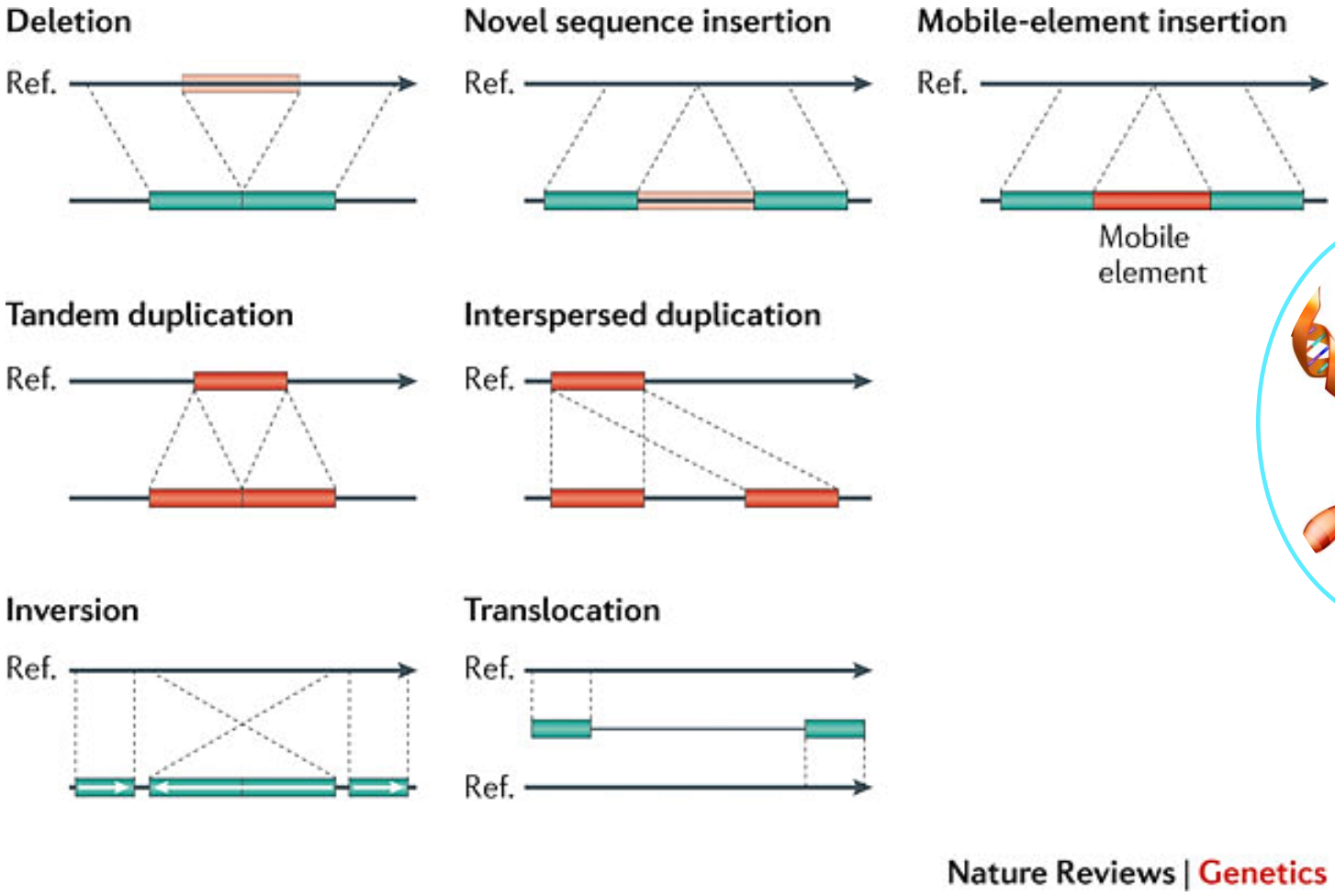


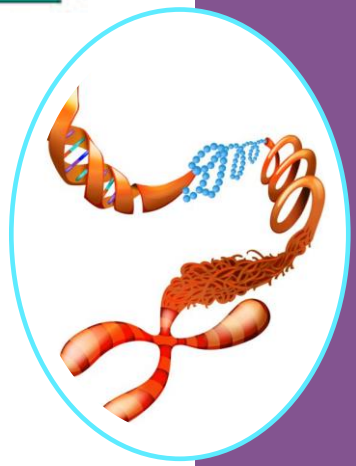
- This also can occur between chromosomes
- The longer the divergence time between 2 species, the more recombination has occurred
- 100 million years since human-mouse divergence
- 40 million years since rat-mouse divergence

Definitions are drawn from biology

- **SNP**: Single mutation surrounded by two matching regions
  - Regions of DNA where 2 sequences have diverged by more than one SNP
- **Large inserts:** regions inserted into one of the genomes
  - Sequence reversals, lateral gene transfer
- **Repeats:** the form of duplication that has occurred in either genome.
- **Tandem repeats:** regions of repeated DNA in immediate succession but with different copy number in different genomes.
  - A repeat can occur 2.5 times
- **Structural** variation (SV): a region of DNA approximately 1 kb and larger in size and can include inversions and balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as copy number variants (CNVs).

Classes of structural variation





Alkan, C. et al. Genome structural variation discovery and genotyping. Nature Reviews Genetics 12, 363-376 (2011).

Genome analyses in Bacteria

- Variation in
  - Genome size
  - GC content
  - Codon usage
  - Amino acid composition
  - Genome organisation
    - Single circular chromosomes
    - Linear chromosome + extra chromosomal elements

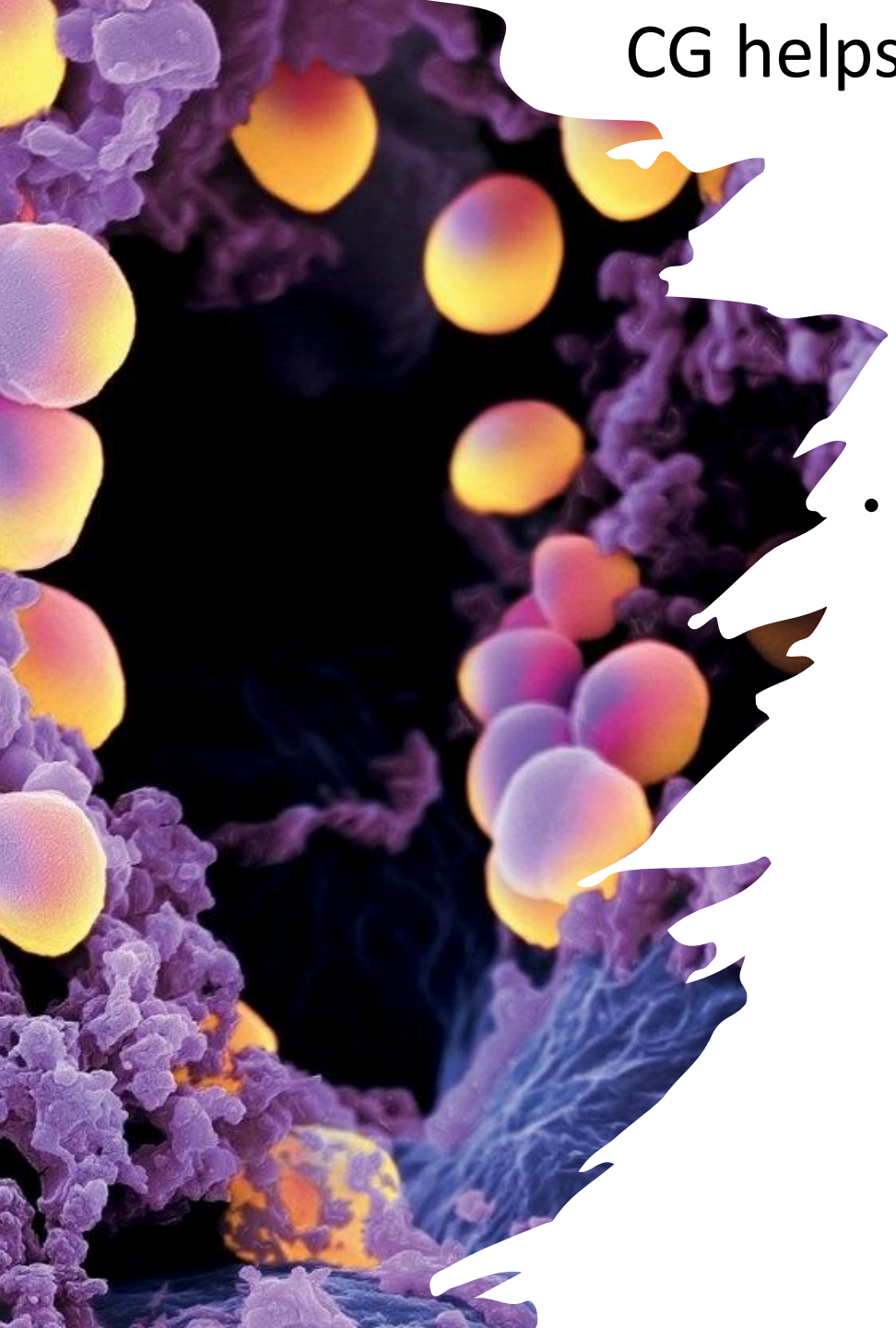
E. coli: 4.6Mbp
M. pneumoniae: 0.81Mbp
B. subtilis: 4.20Mbp

B. burgdorferi: 29%
M. tuberculosis: 68%

G, A, P, R: GC rich
I, F, Y, M, D: AT rich

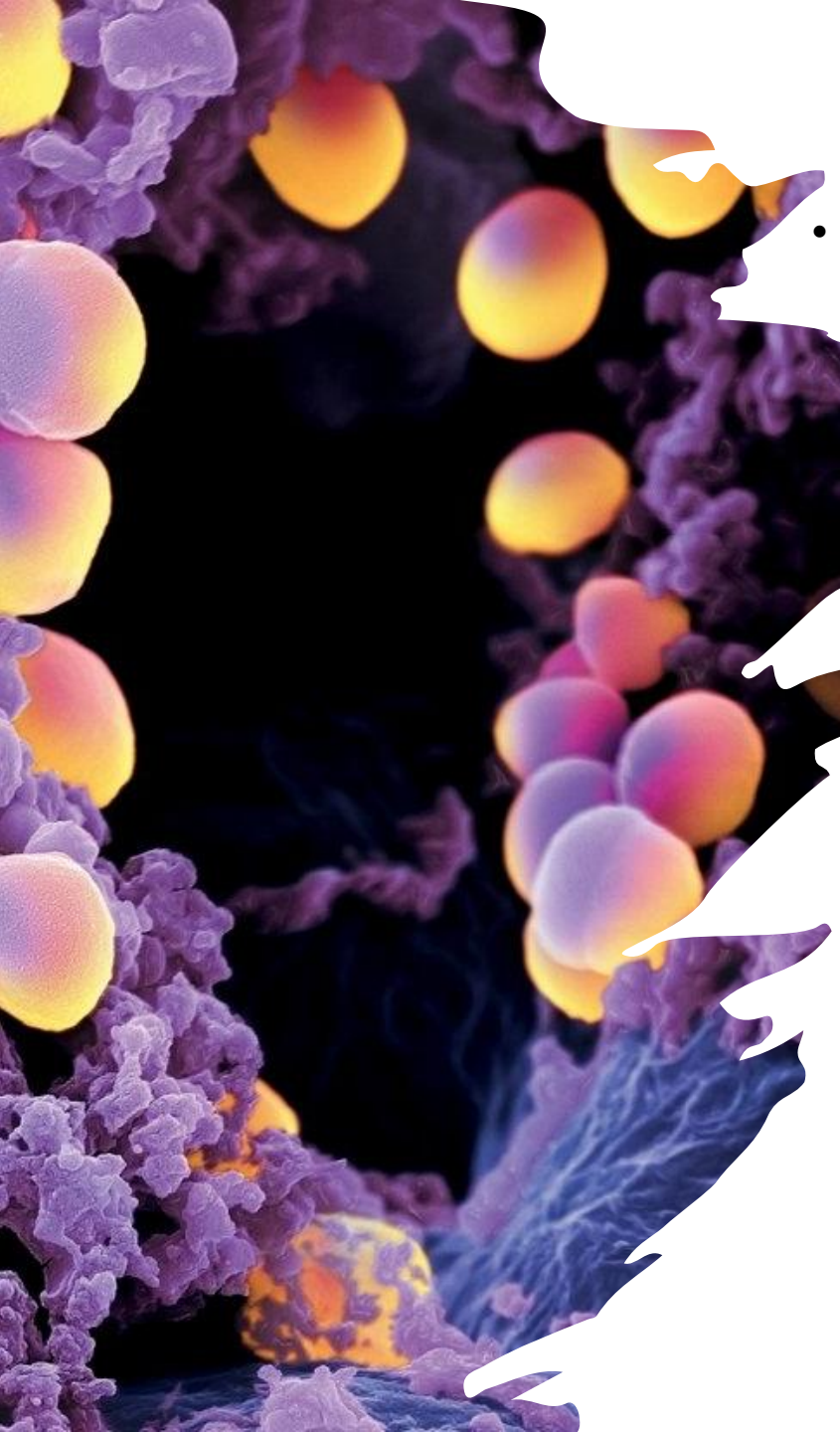
CG: Comparisons between genomes

- The stains of the same species
- The closely related species
- The distantly related species
  - List of Orthologs
  - Evolution of individual genes
  - Evolution of organisms



CG helps to ask some interesting questions

- Identification similarities/differences between genomes may allow us to understand :
  - How 2 organisms evolved?
  - Why certain bacteria cause diseases while others do not?
  - Identification and prioritization of drug targets

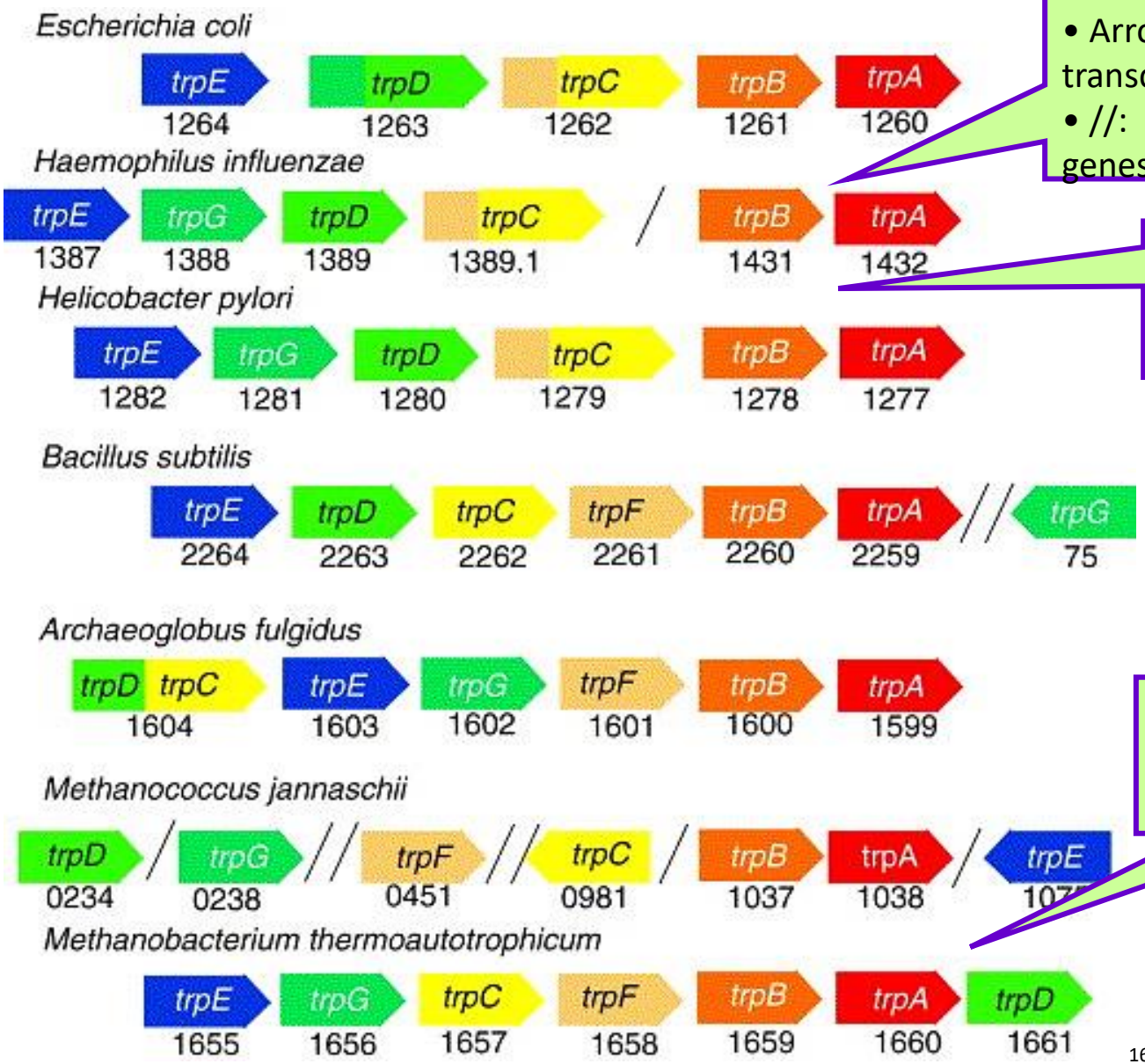


CG: Unit of comparison

- Unit of comparison: Gene/Genome
  - Number
  - Content (sequence)
  - Location (map position)
  - Gene Order
  - Gene Cluster (Genes that are part of a known metabolic pathway, are found to exist as a group)
  - Colinearity of gene order is referred as synteny
  - A conserved group of genes in the same order in two genomes as a syntenic groups or syntenic clusters
  - Translocation: movement of genomic part from one position to another

Comparison of the coding regions

Structure of tryptophan operon



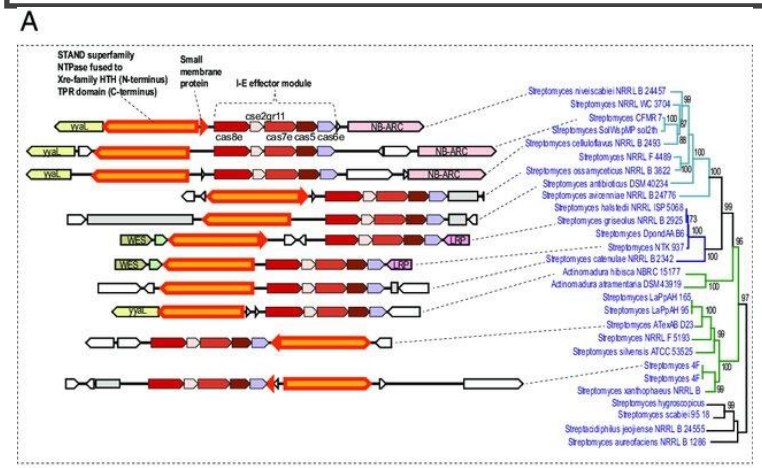
- Numbers: Gene number
- Arrows: Direction of transcription
- //: Dispersion of operon by 50 genes

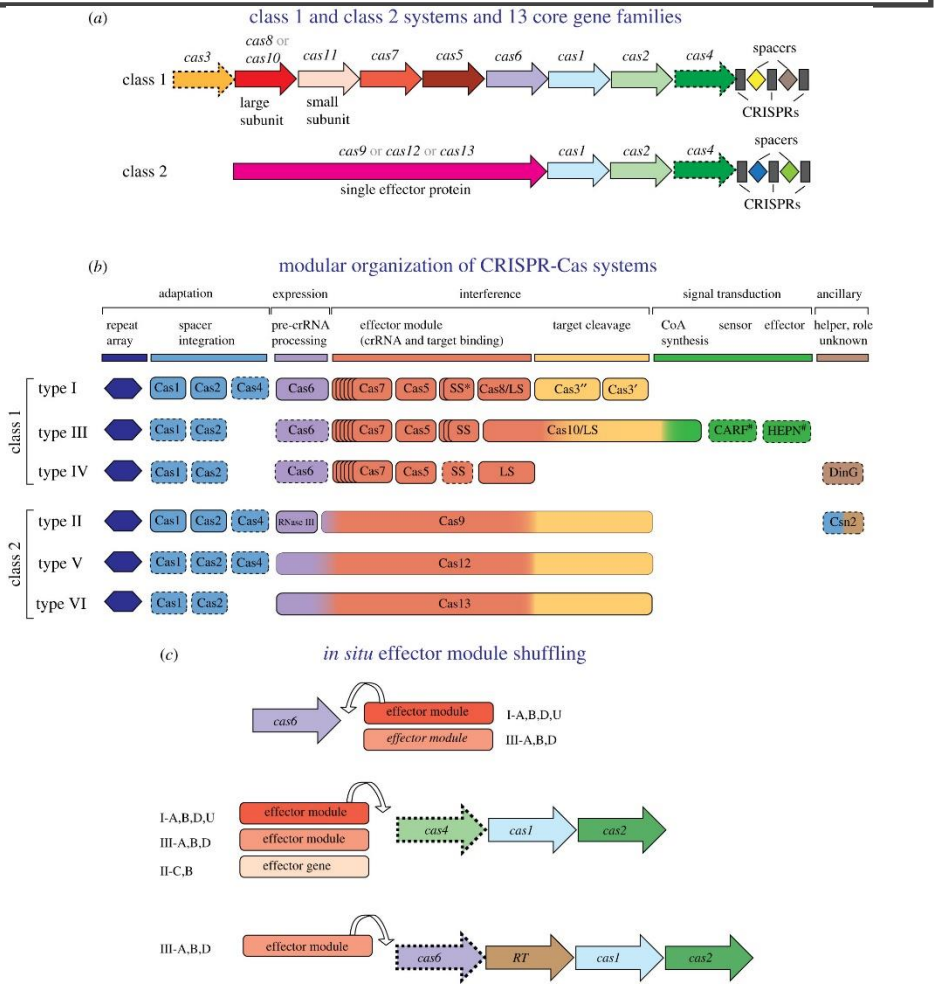
- Domain fusion
- trpD and trpG
- trpF and trpC

- trpB and trpA
- genetically linked separate genes

Dandekar et al., 1998

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated proteins)



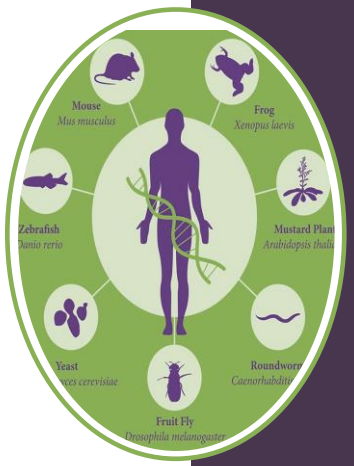


Comparative genomics in Eukaryotes

- Evolution – Charles Darwin (1838)

  ➡ Similarity between different species

  ➡ Model organisms



- A human shares 50% of his genes with a banana.

  How ?

  - Humans and bananas are multi-cellular
  - Other Similarities

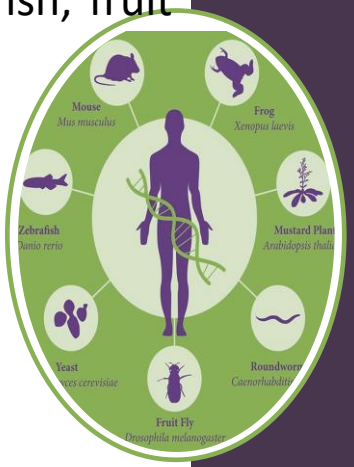
    Humans share 23% of their genes with Yeast

- Could banana be a good model organism ?

Comparative genomics in Eukaryotes

Model Organisms

In addition to the sequencing of the human genome, which was completed in 2003, scientists involved in the Human Genome Project sequenced the genomes of a number of model organisms that are commonly used as surrogates in studying human biology. These include the rat, puffer fish, fruit fly, sea squirt, roundworm, and the bacterium *Escherichia coli.*



Heavily Studied – used as examples for other species

Once it is studied enough – It is a good candidate

<u>Important requirements</u>:

- Size
- Generation Time (for genetic research)
- Manipulation (genetic and not)
- Little “Junk DNA” (easy for sequencing)
- Money

Comparative Genomics

Genome analyses in Eukaryotes

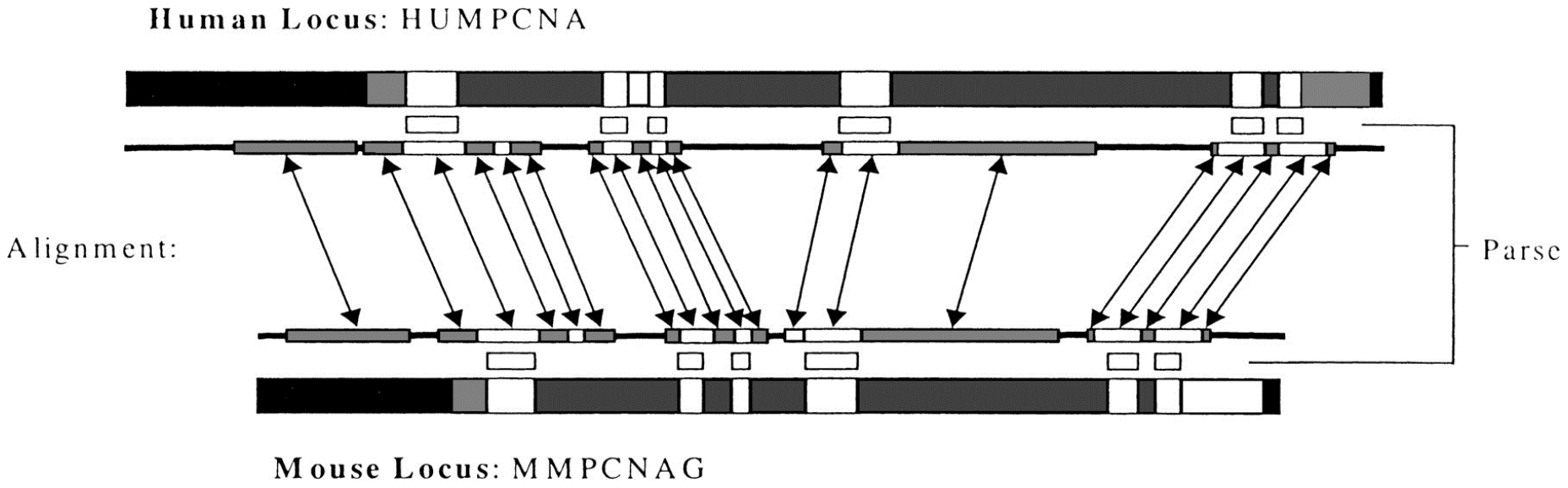


Figure 1 Regions of the human and mouse homologous genes: Coding exons (white), noncoding exons (gray}, introns (dark gray), and intergenic regions (black). Corresponding strong (white) and weak (gray) alignment regions of GLASS are shown connected with arrows. Dark lines connecting the alignment regions denote very weak or no alignment. The predicted coding regions of ROSETTA in human, and the corresponding regins in mouse, are shown (white) between the genes and the alignment regions.

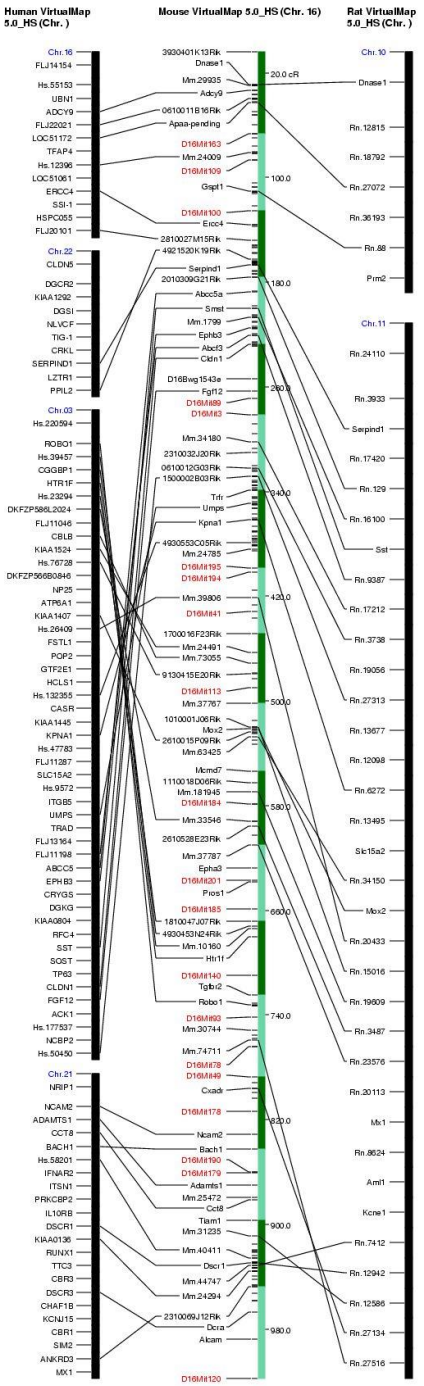
Comparative Genomics

Comparison of mouse chromosome 16 and the human genome

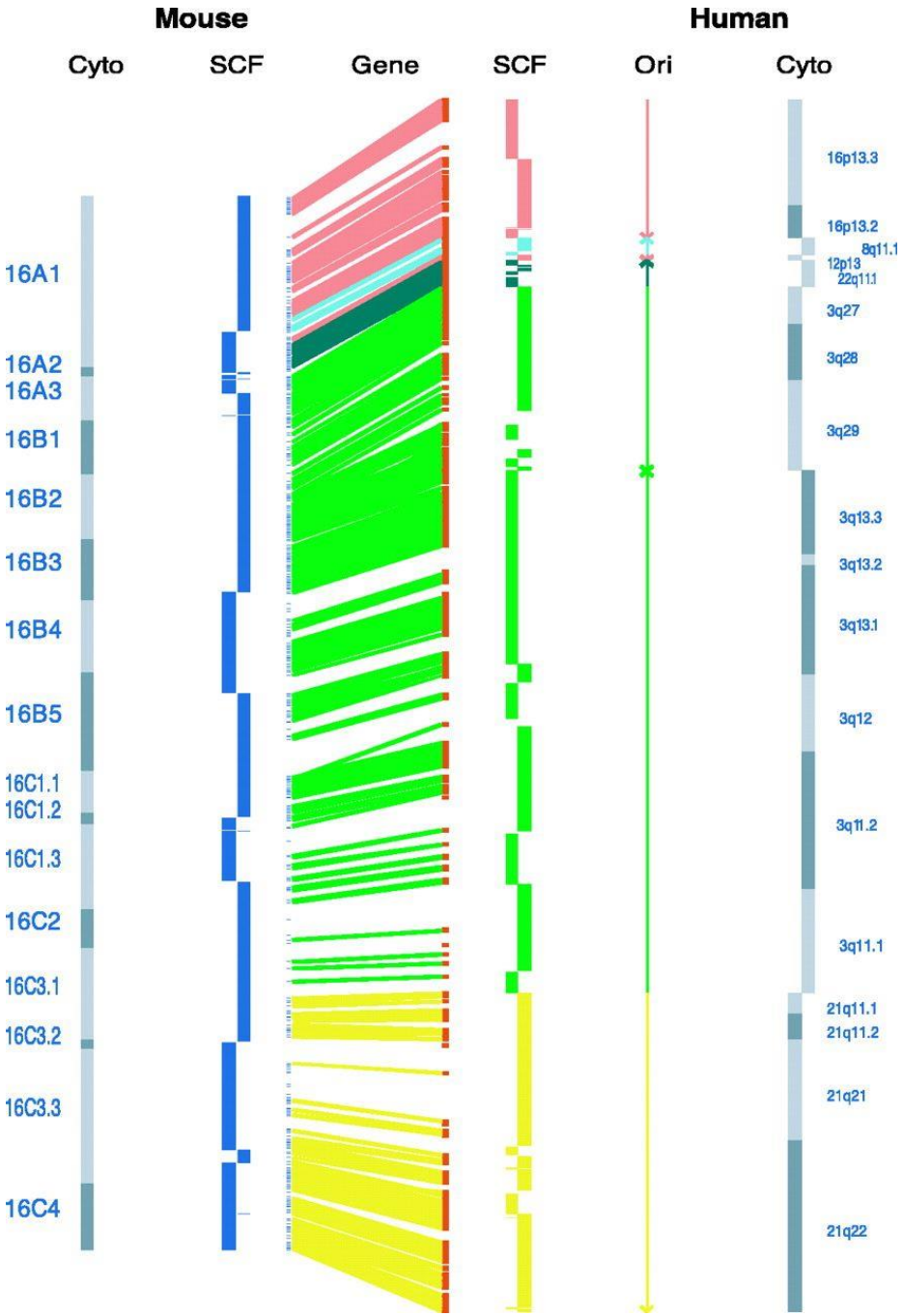
- **Mural et al., Science, 2002, 296:1661**
- Celera group
- Synteny with human chr.'s 3,8,12,16,21,22 and rat chr.'s 10,11

Q: Why more breakpoints in mouse-human than in mouse-rat?

Q: Why more conserved genes in human than in rat?



- Large regions conserved (1/3 of Mmu16 on Hsa16 & Hsa21; the rest in 5 other regions)
- Content of genes in regions preserved
- Order of genes preserved (only a couple exceptions)
- 99% of anchors conserved in order and orientation
- All 509 putative orthologues are consistent in their location and order
- About 2% of Mmu genes are unique to mouse (relative to human) – no homologue found for 14 of 731 genes

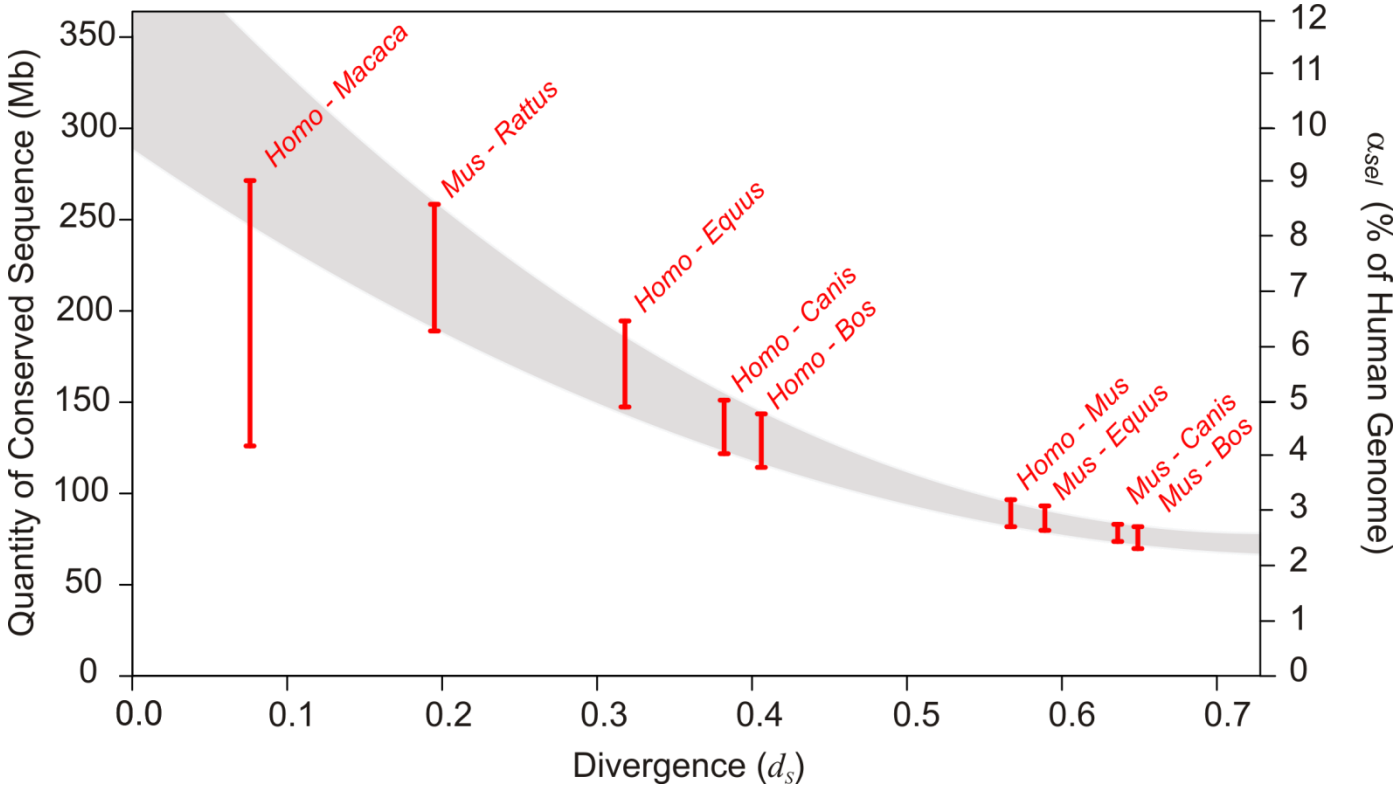


Conservation is often a good predictor of functionality

Orthologous human mouse genes have conserved exonic structure.

- 85% of the orhologous pairs have identical number of exons
- 91% of the orthologous exons have identical length
- 99.5% of the orthologous exons have identical phase
- there are a few cases of intron insertion/deletion (22)

Sequence conservation doesn't imply function conservation

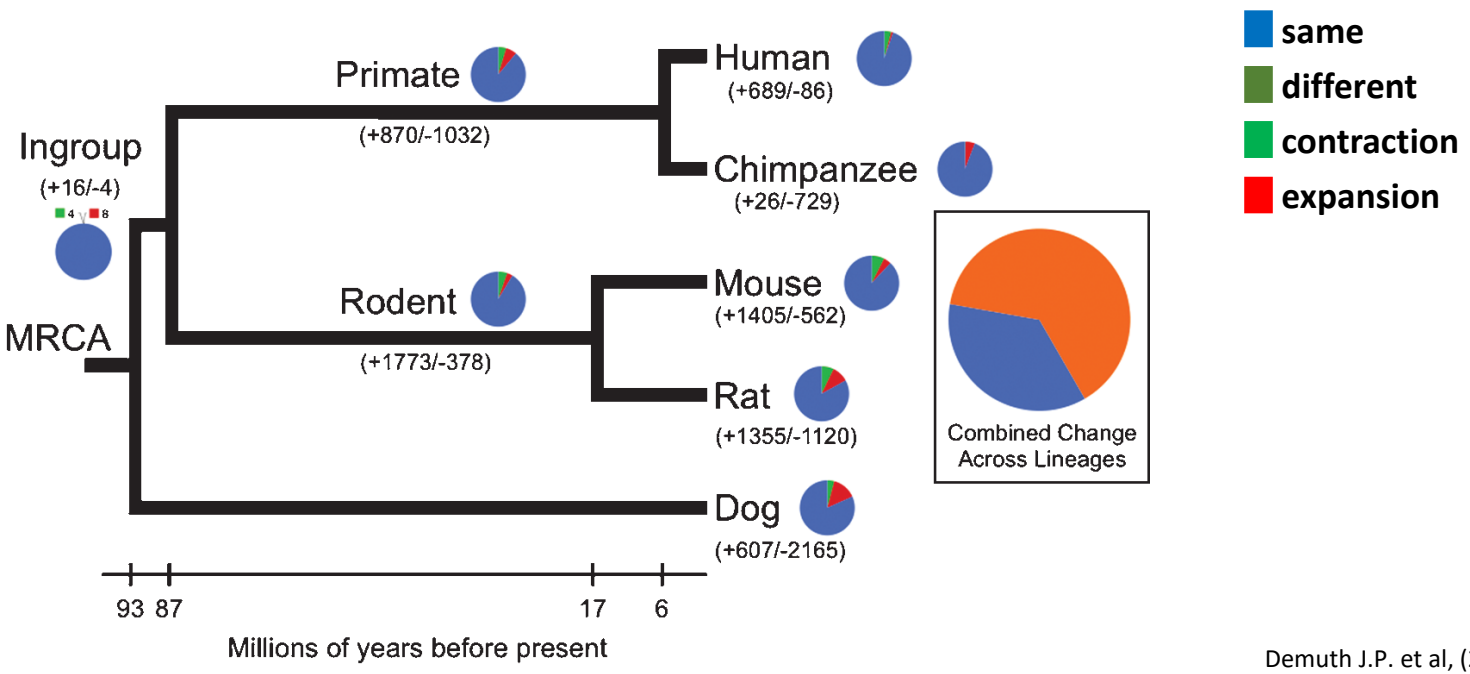


Massive turnover of functional sequence in mammalian genomes

Meader S et al. (2011)

Protein-coding genes and evolutions

Changes of protein coding repertoires and contributions to phenotypic differences

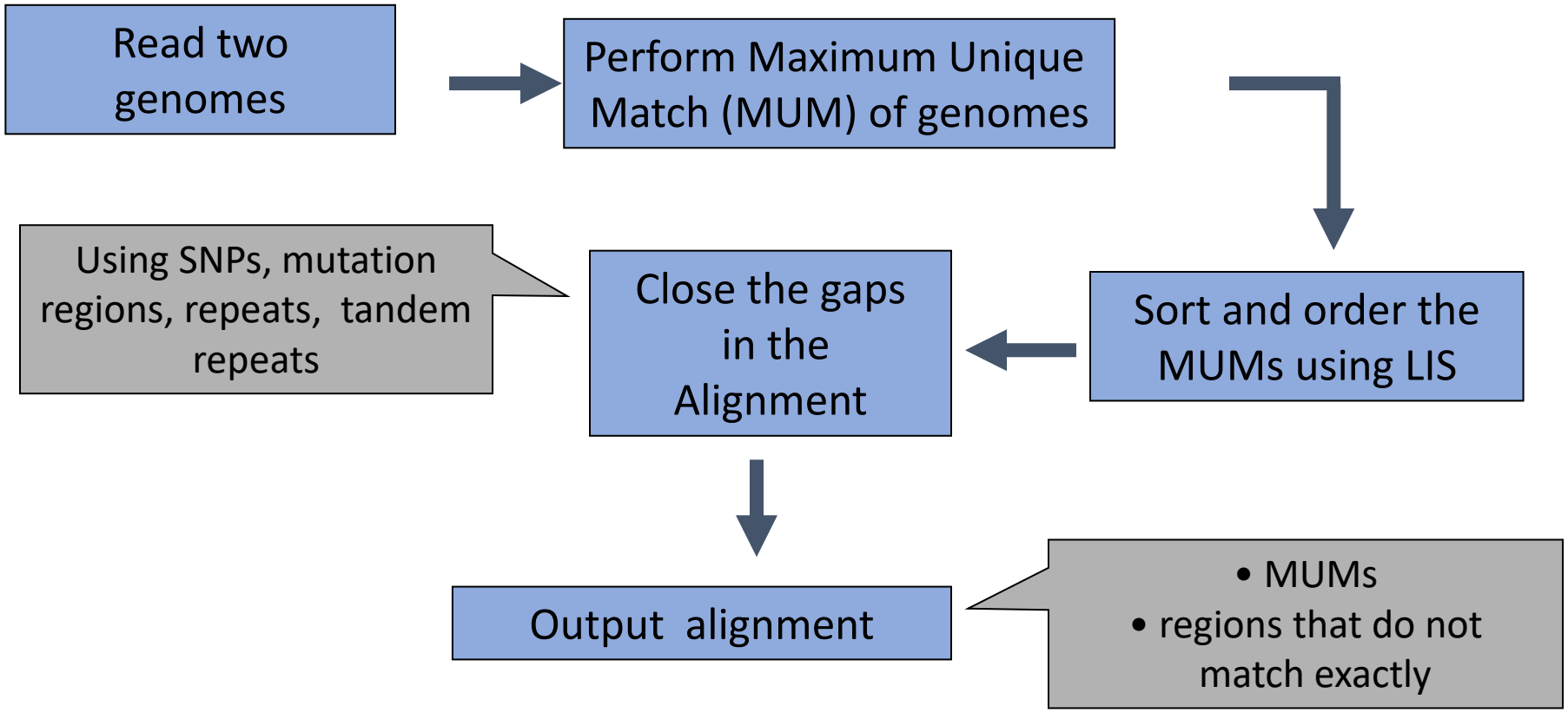


Demuth J.P. et al, (2006)

Comparative Genomics Tools

- BLAST2
- Comparisons and analyses at both
  - Nucleic acid and protein level
- MUMmer

MUMmer: Steps in the alignment process



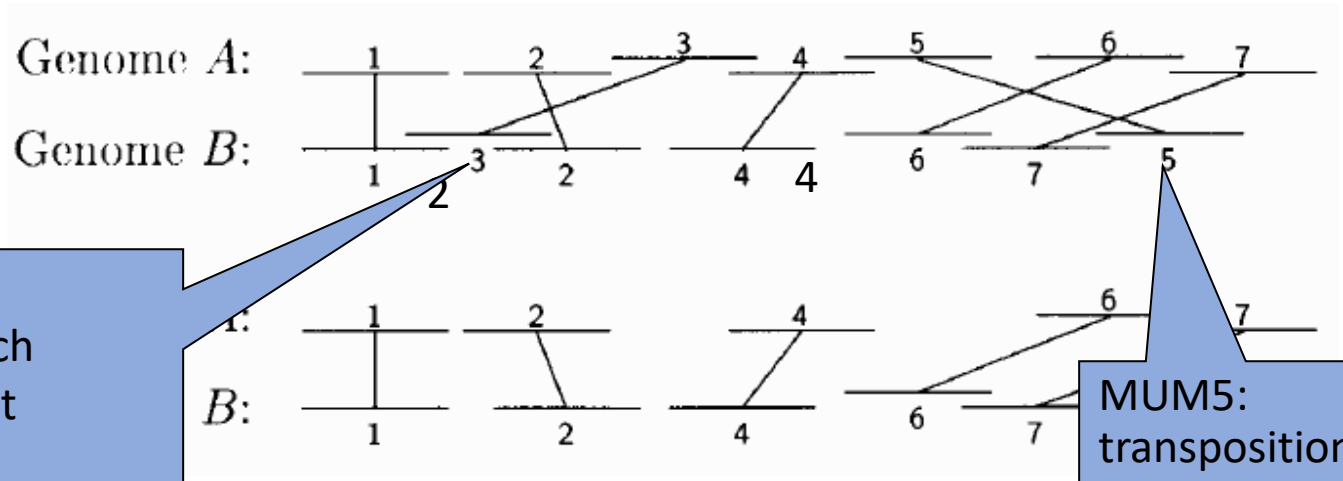
Genome1: ACTGATTACGTGAACTGGATCCA
Genome2: ACTCTAGGTGAAGTGATCCA

Genome1: ACTGATTACGTGAACTGGATCCA
Genome2: ACTCTAGGTGAAGTGATCCA

ACTGATTACGTGAACTGGATCCA
ACTC--TAGGTGAAGT-GATCCA

Sorting & ordering MUMs

- MUMs are sorted according to their position in Genome A
- The order of matching MUMs in Genome B is considered



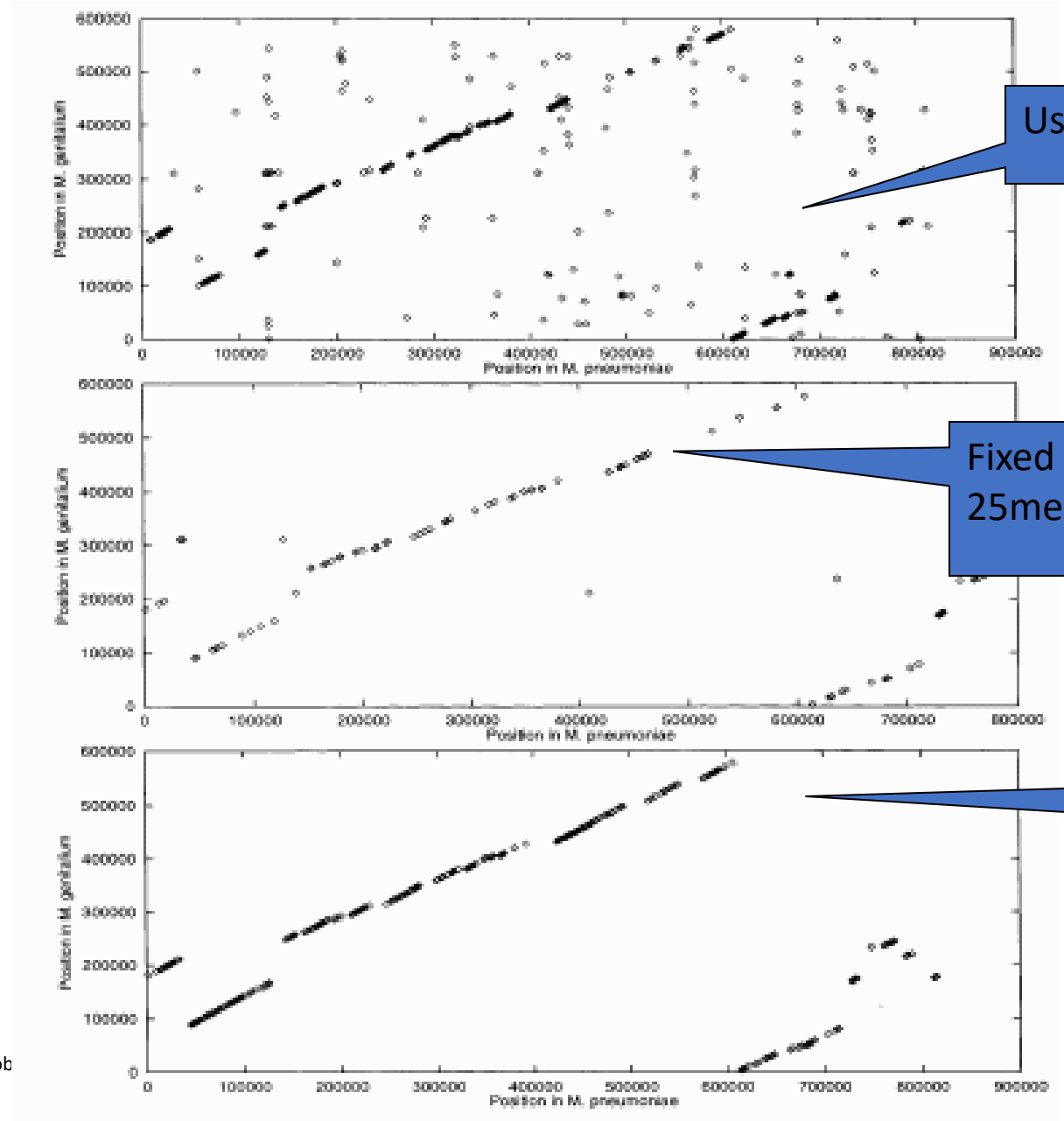
MUM3:
Random match
Inexact repeat

MUM5:
transposition

- LIS algorithm to locate longest set of MUMs which occur in ascending order in both genomes

Leads to Global MUM-alignment

Comparison of 2 Mycoplasma genomes



Using FASTA

Fixed length patterns: 25mers

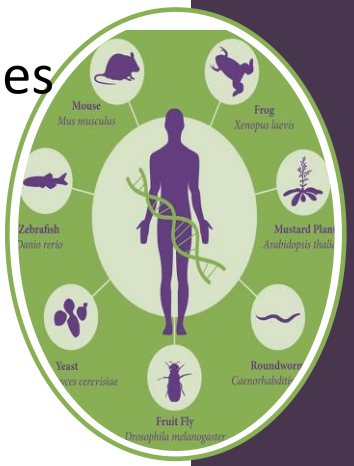
MUMmer

Comparative genomics in Eukaryotes

This paper describes:

A comparison between the genomes of 3 Eukaryotes:

Eukaryote – Cell has inner structures with membranes (nucleus)



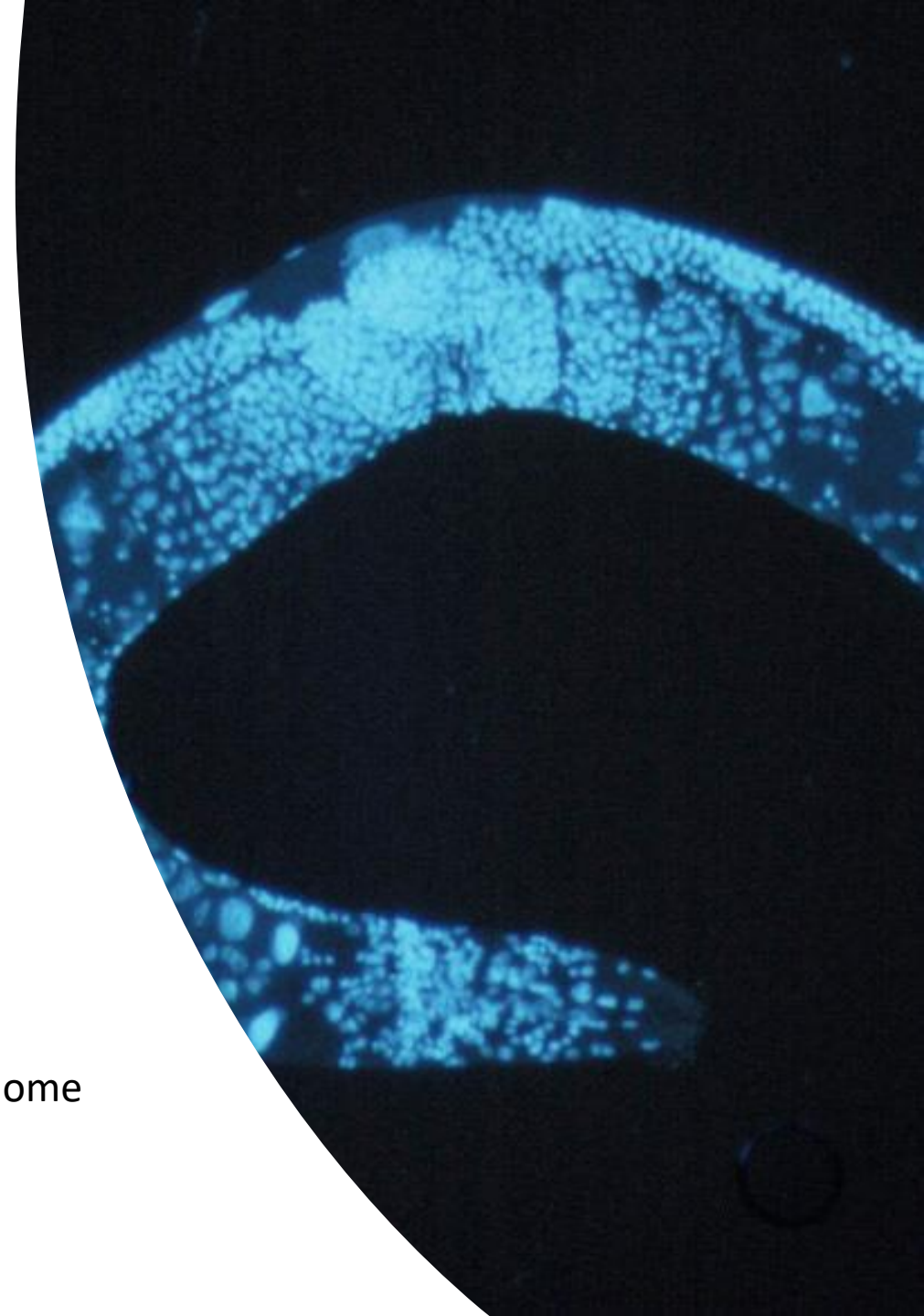
1) A fruit fly - Drosophila melanogaster
2) A worm – C. elegans
3) Yeast – S. cerevisiae

Drosophila melanogaster

- Popular model organism (for developmental biology)
- A trial for the human genome (sequenced at 2000)
- Easily induce mutations

Caenorhabditis elegans

- Transparent, 1-mm long
- Simple – 959 cells (300 neurons)
- Eat, sleep & have sex (or self-fertilize)
- Hermaphrodites – 99.95%, Males – 0.05%



Good as a model organism for:

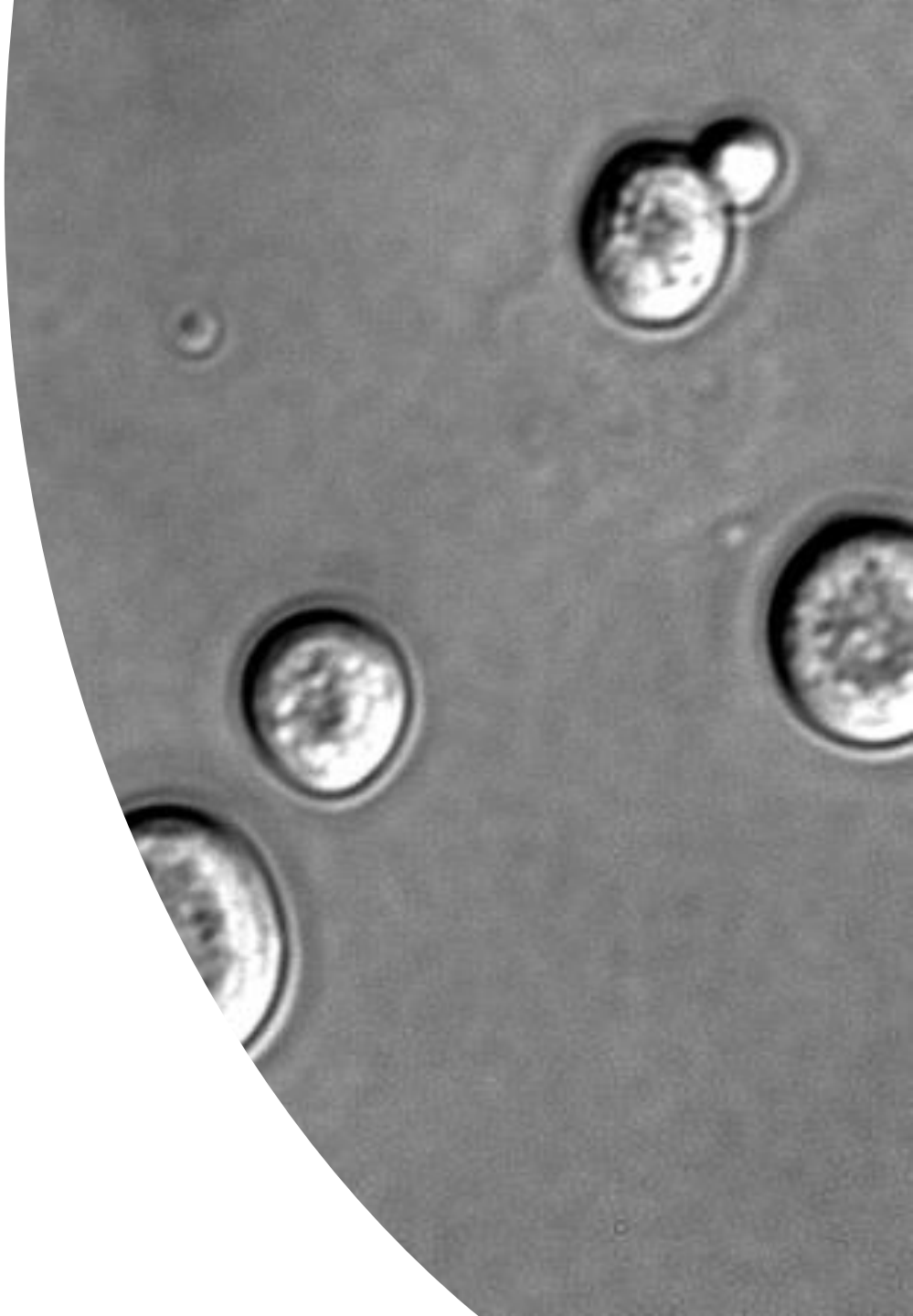
Genetics: First multi-cellular sequenced genome

Developmental biology: cell fate mapping

Neurobiology: neurons connectivity map

Saccharomyces cerevisiae

- Also called Baker's yeast
- Single-celled
- Diameter: 5-10 µ
- Popular model organism
- Simplest Eukaryote
- First Eukaryotic sequenced genome



The 1st comparison

- Instead of counting genes - count gene families
- What are gene families ?

Sets of paralogs

Paralogs = highly similar proteins in the same genome

Similar functionality – but not always

- Remark: proteins = genes

Findings

	H. influenzae	Yeast	Fly	worm
Total # of genes	1700	6200	13,600	18,400
# of gene families	1400	4400	8100	9400
# of duplicates	300	1800	5,500	9000

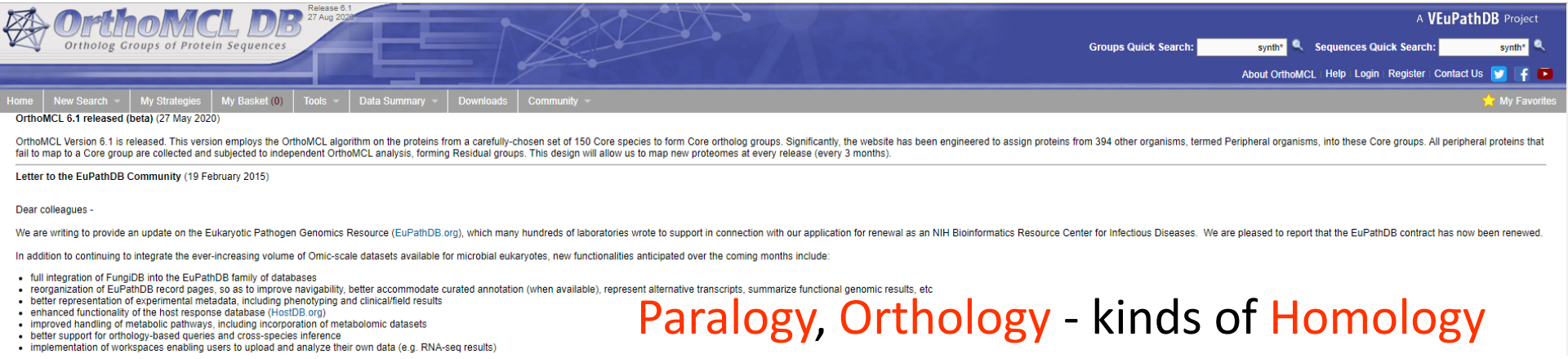
- Size of a family: one or more
- No. of families – not a good measure for complexity

The 2nd comparison

- Pool genes of large families of 3 species:
  - For each protein – search for orthologs
  - Orthologs = Similar proteins in other species
- Among families found in flies and worms (but not yeast):

  Responsible for multi-cellular development
- Among families found only in flies:

  Responsible for immune response and fly specific

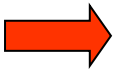


Paralogy, Orthology - kinds of Homology

The 3rd comparison

- Compare all genes of three species with length limitation (80% of length)

- 20% of the fly appear in worm and yeast



They perform functions common to all eukaryotic cells

The 4th comparison

- Compare all genes of three species to mammalian sequences

  (without length limitation)

- 50% of the fly proteins appear in mammals

- 36% of the worm proteins appear in mammals

  ➡ Fly is closer to mammals

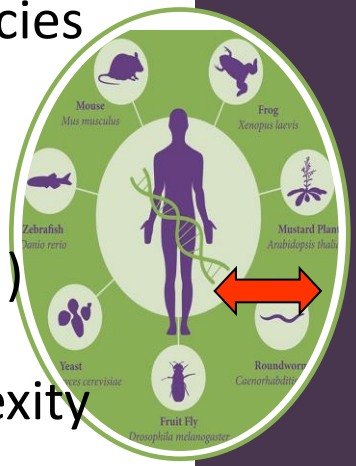
- Most of mammalian sequences used here were short

  ➡ The similarities reflect conserved domains

Comparative genomics in Eukaryotes

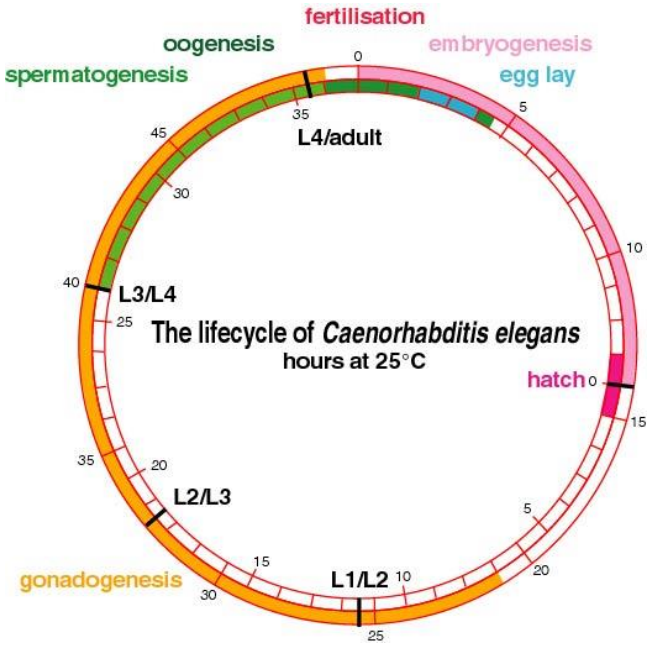
To conclude:

- Significant similarity between genomes of ”distant” species (Man – Yeast 23%)
- Similarity increases for taxonomically close species ( )
- No. of genes or gene families – bad measure for complexity
- Why ? More information that is not encoded in the genome (Protein interactions – e.g. physical proximity of genes)
- How to define complexity ?



The genome of the nematode *Caenorhabditis elegans*

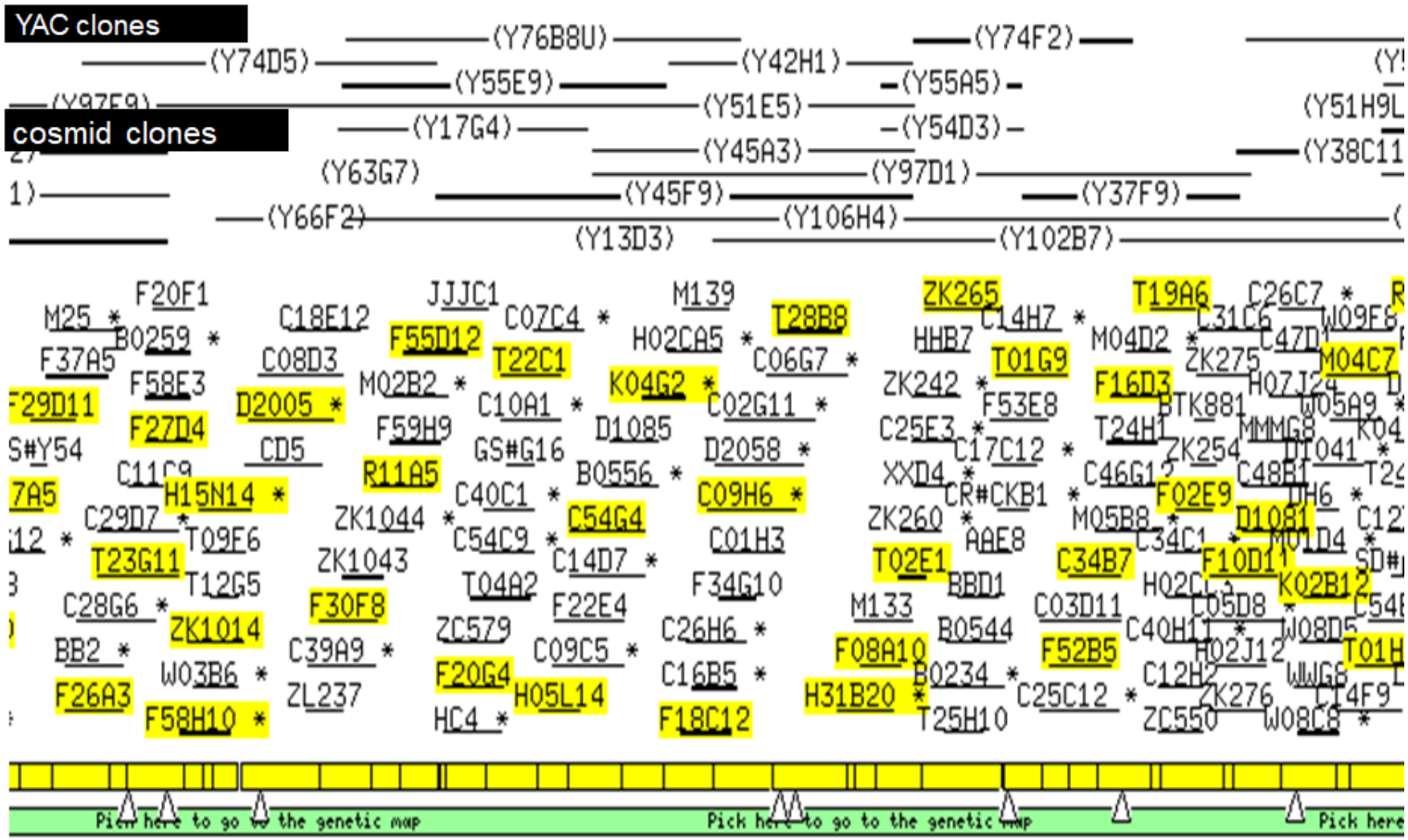
In common with other nematodes, *C. elegans* develops through four larval stages (also called juveniles in some nematode literature) which are separated by moults. The lifecycle takes about 3 days at 20 deg.C



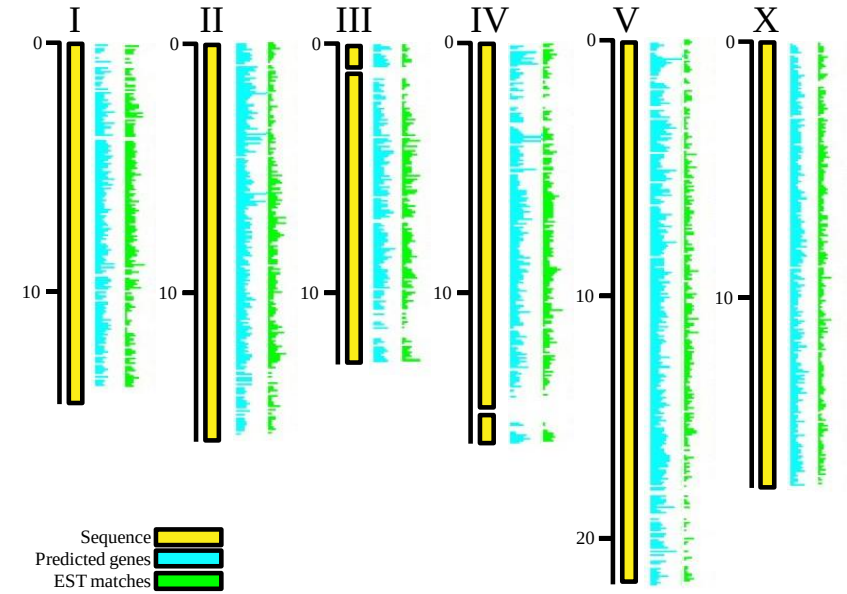
C. elegans genomeproject

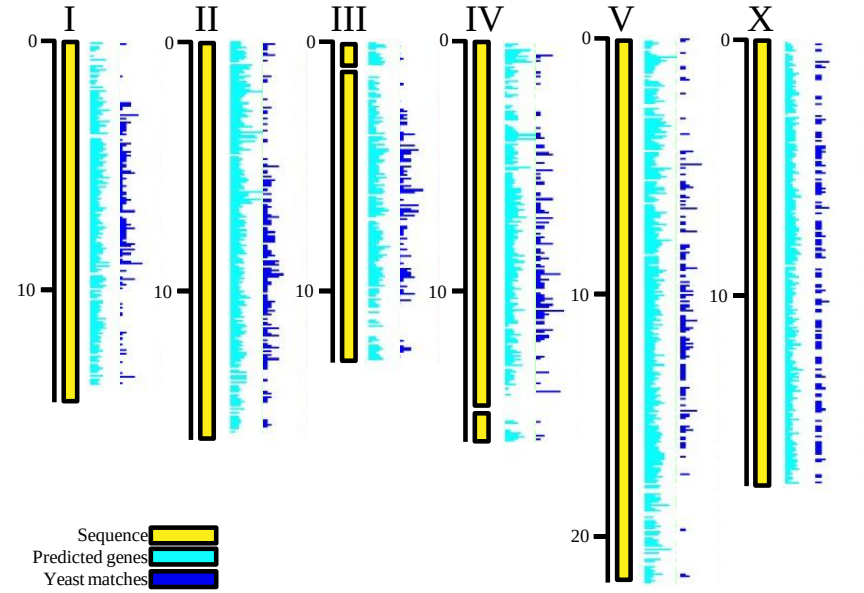
The completed genome sequence was derived from a PHYSICAL MAP anchored on the GENETIC MAP. From the physical map the teams selected

- 2527 cosmids
- 113 fosmids
- 257 YACs
- 44 long range PCR products

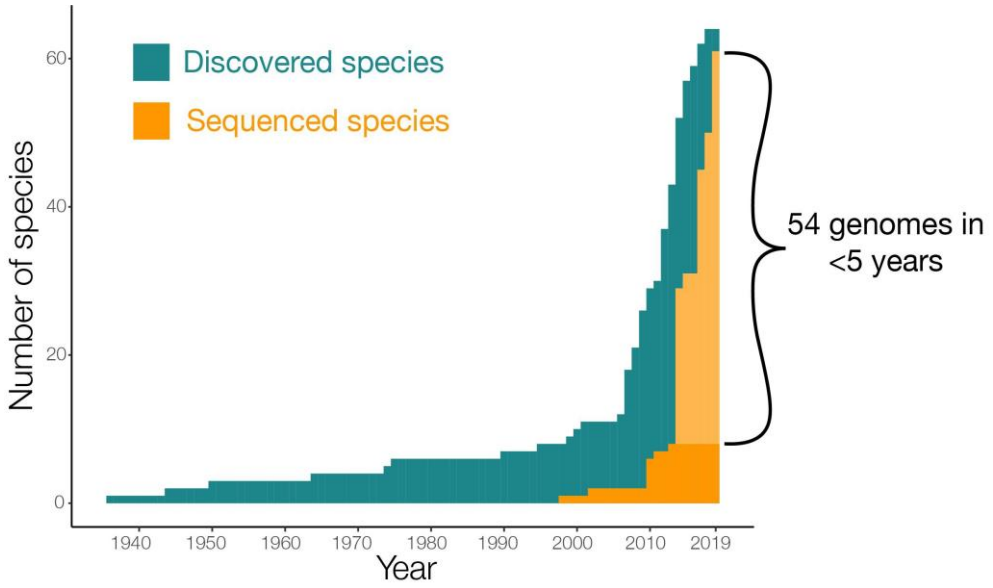


The *Caenorhabditis* Genomes Project

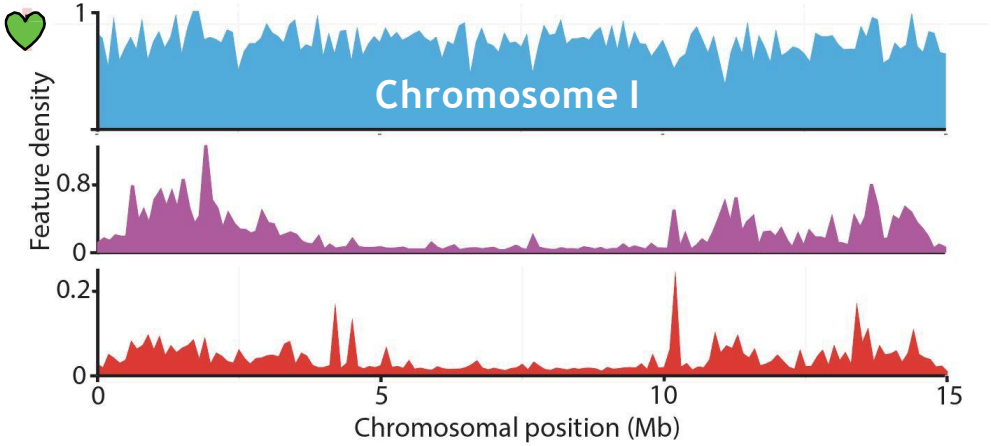




Sequence the genomes of all *Caenorhabditis* species currently in culture



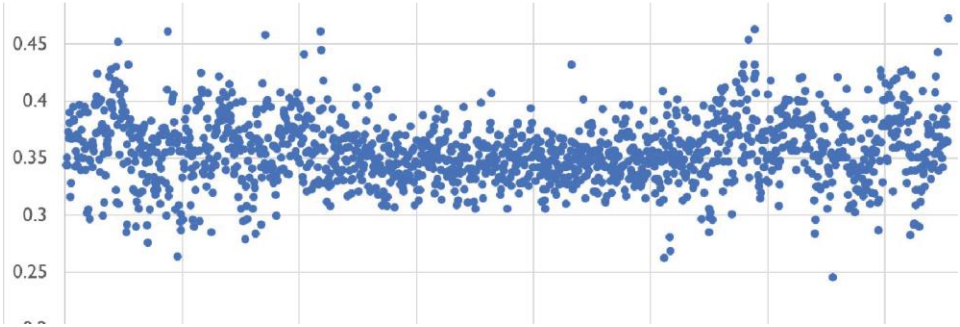
The C. elegans genome has remarkable long-range structure
How did this striking pattern arise? How is it maintained?



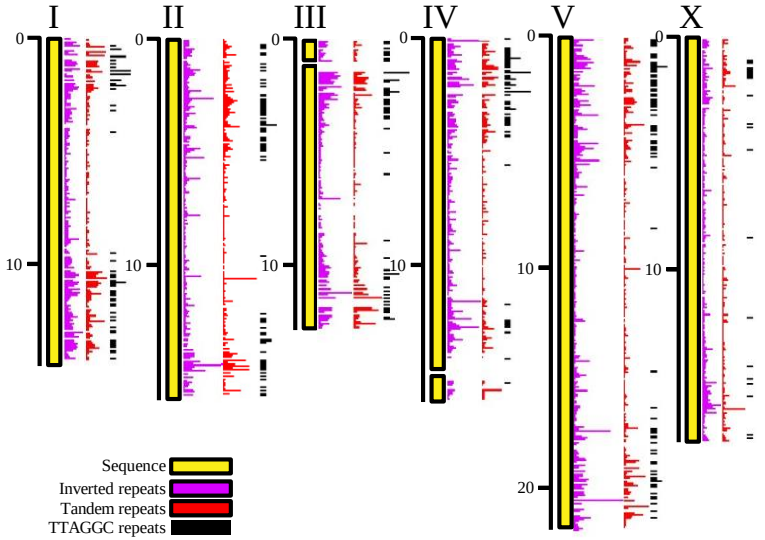
Genes

Repeats

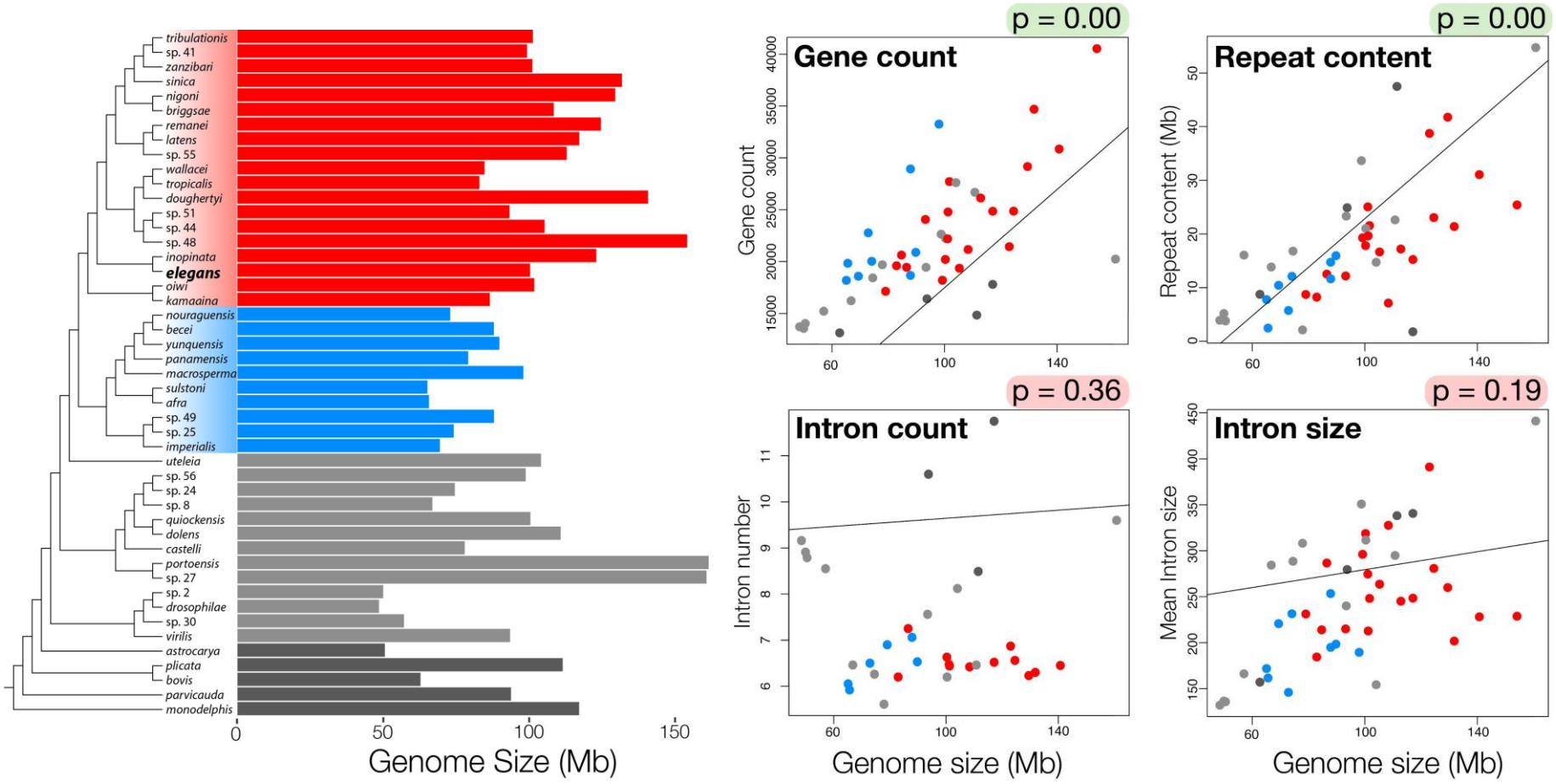
Inverted repeats



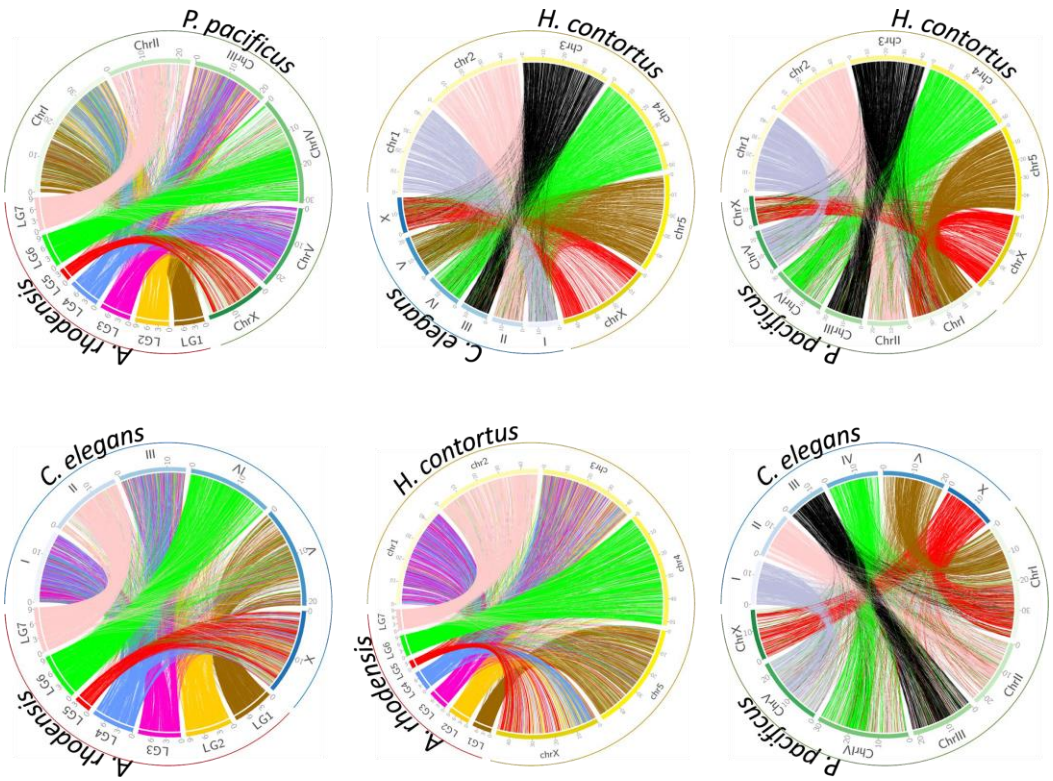
GC proportion



Extensive variation in genome size



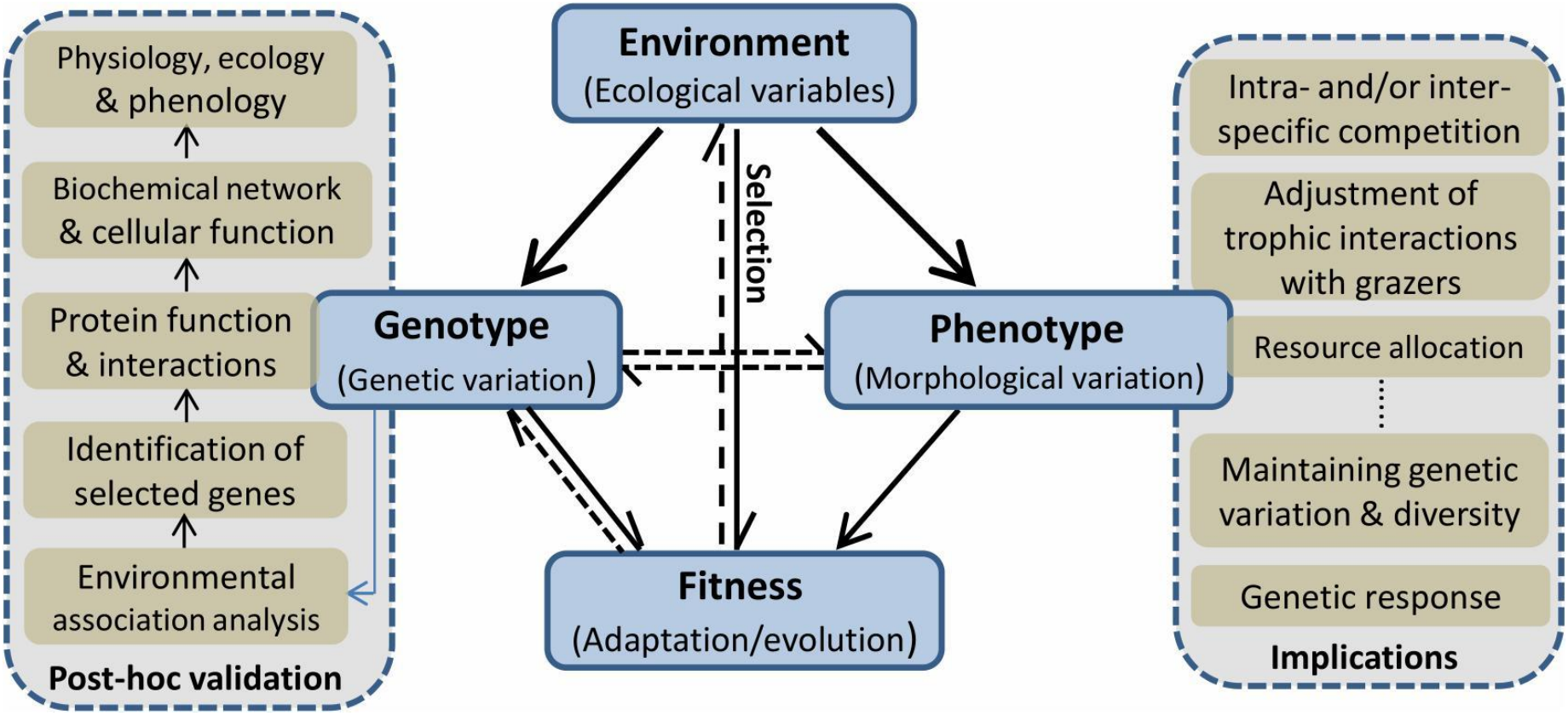
Linkage conservation across Rhabditina



Some linkage groups appear **conserved** across *Rhabditine* evolution.

Other linkage groups appear to have participated in **breakage and fusion events**.

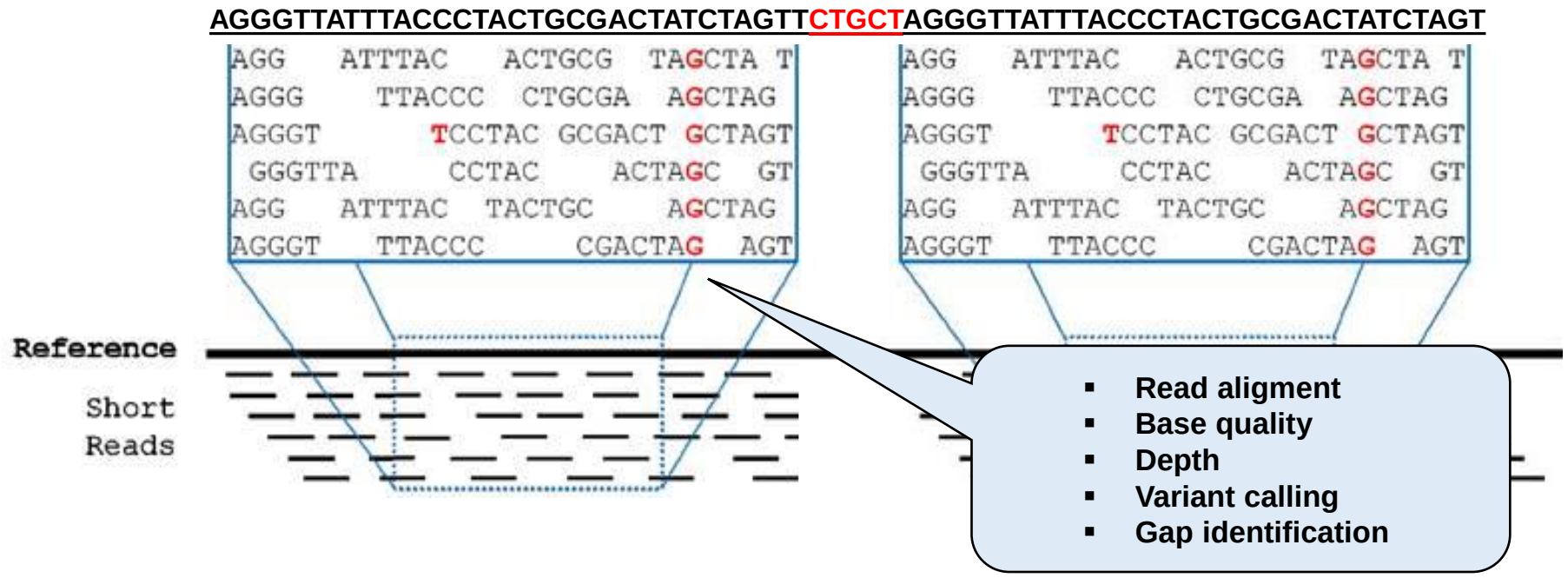
Relationships among environment, phenotype, and genotype

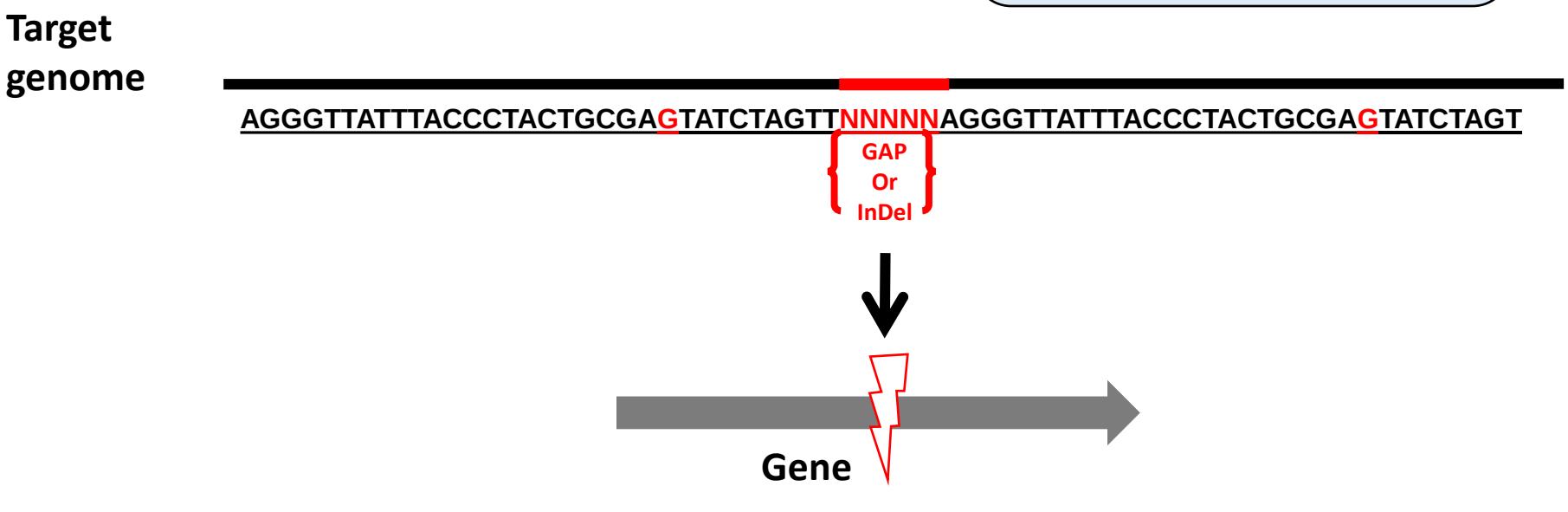


the evolved plastic phenotypes have wide ecological implications, such as the overwhelming competition to the rival, the adjustment of community balance, and resource allocation under a harsh environment.

https://www.frontiersin.org/articles/10.3389/fmars.2020.545102/full

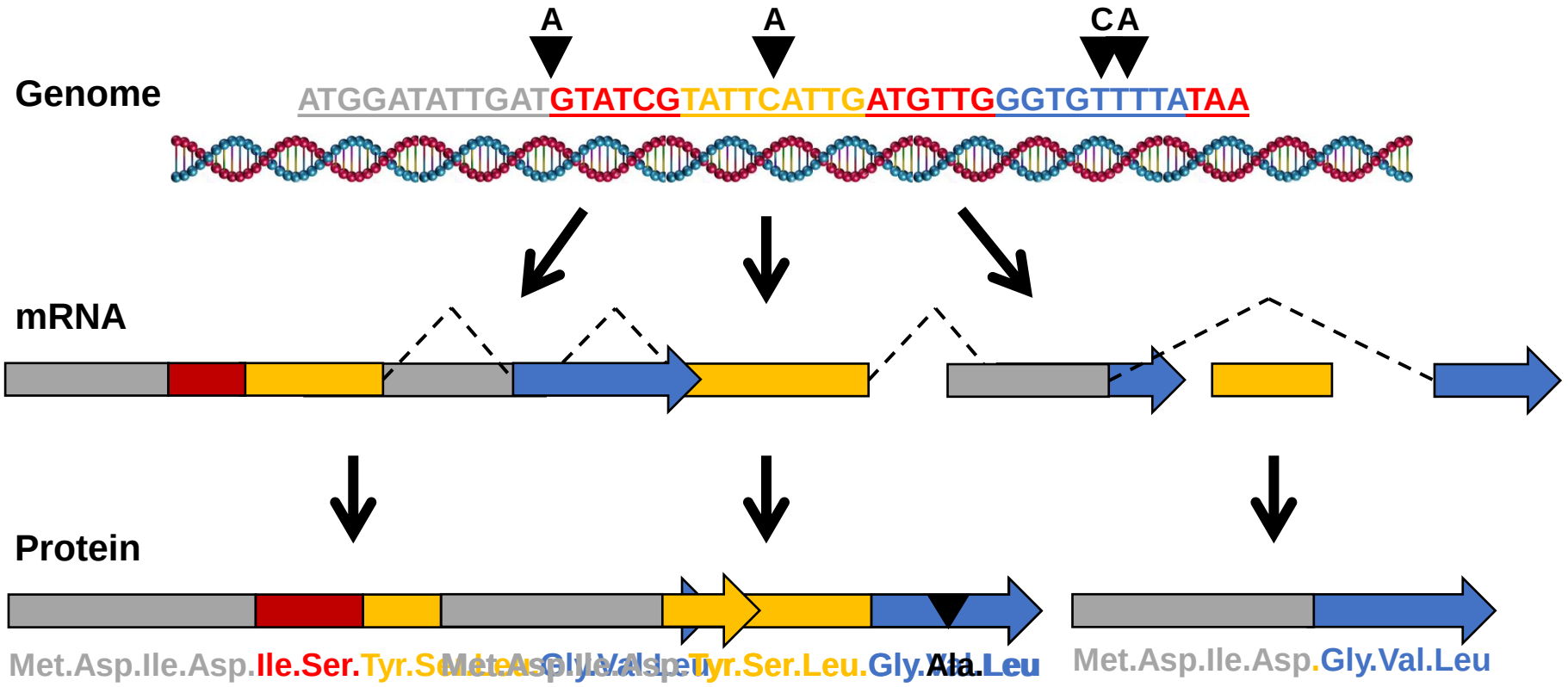
Genetic variant análisis: Mapping reads back to the genomes





GENETIC VARIANT ANALYSIS

SNPs: single nucleotide polimorphism, is the variation in the DNA sequence caused by the substitution of a single base (A,T, C, o G) in a particular loci in a genome



Synonymous substitution: No amino acid change

Alternative splicing :

Longer protein **Shorter protein**

Is it functional?

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data

SnpEff & SnpSift

Genomic variant annotations and functional effect prediction toolbox.

Download SnpEff

Latest version 5.0 (2020-08-09)

Requires Java 1.12

bcftools

genomic variant calling and manipulation of VCF/BCF files

VCFtools

Home | Sourceforge page | Examples & Documentation | Downloads

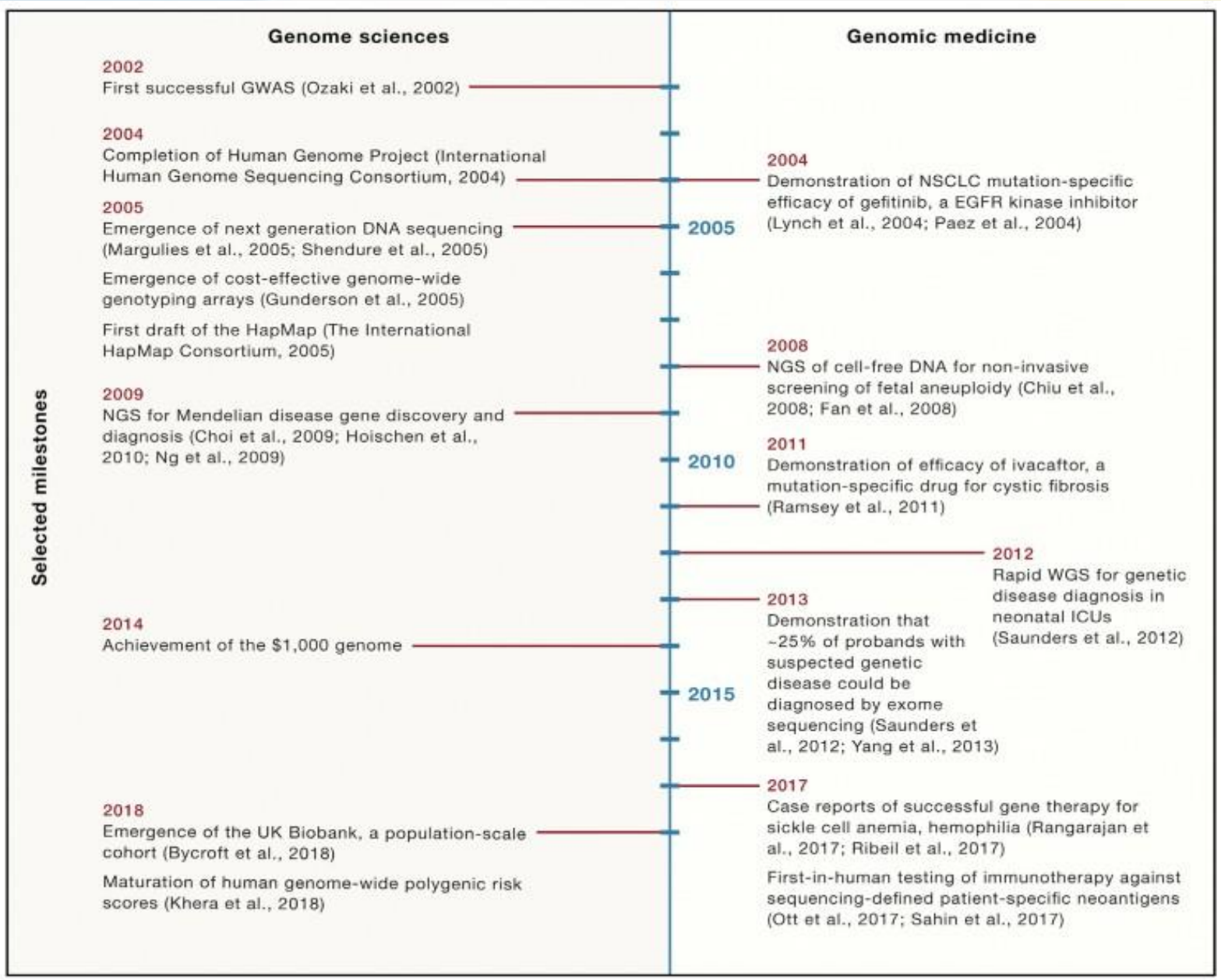
As of July 2015, the VCFtools project has been moved to github! Please visit the new website here:
vcftools.github.io

The human genome

2001

3 billion bases

Breve Historia de la Genomica



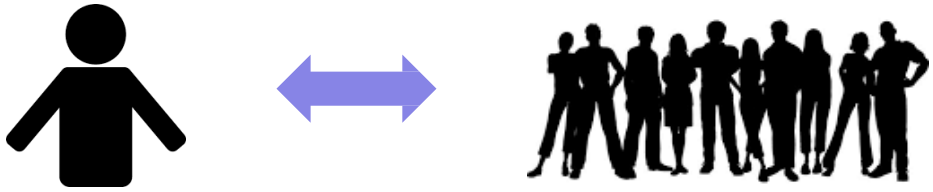
Human genomic variation

AGTGGAGCTGGGAAAGCAGCTGGCTAA
GAAAATAGAGCCTGAGCTTGATGGCAGT
GCTCAAGTGACCTCTCACGACGCTTCT

3 billion sites in the human genome

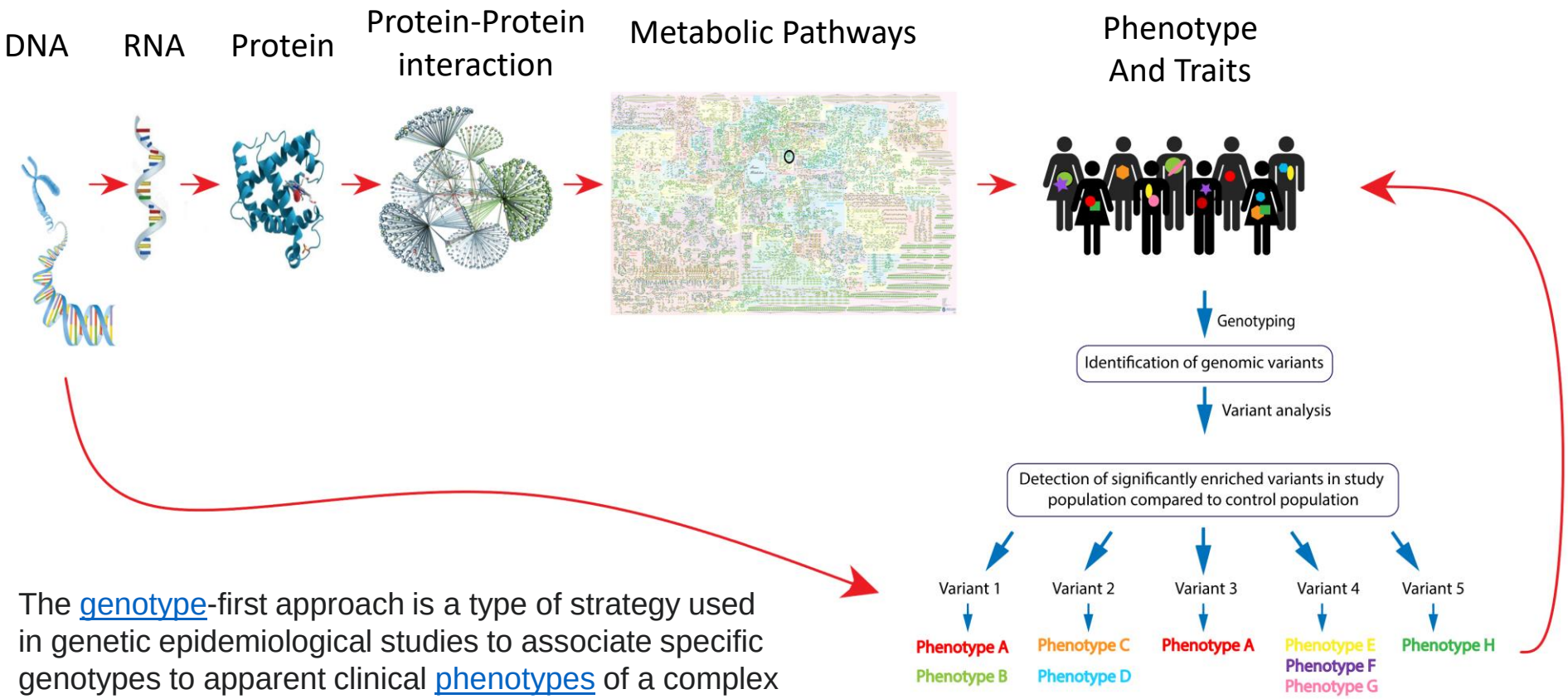
Humans share 99.5% DNA with any other human

We share commonly variant sites and most of these are *biallelic*



AGCTGGCTAAGAAAATAGAGCCTGAGCT

Genotype to Phenotype

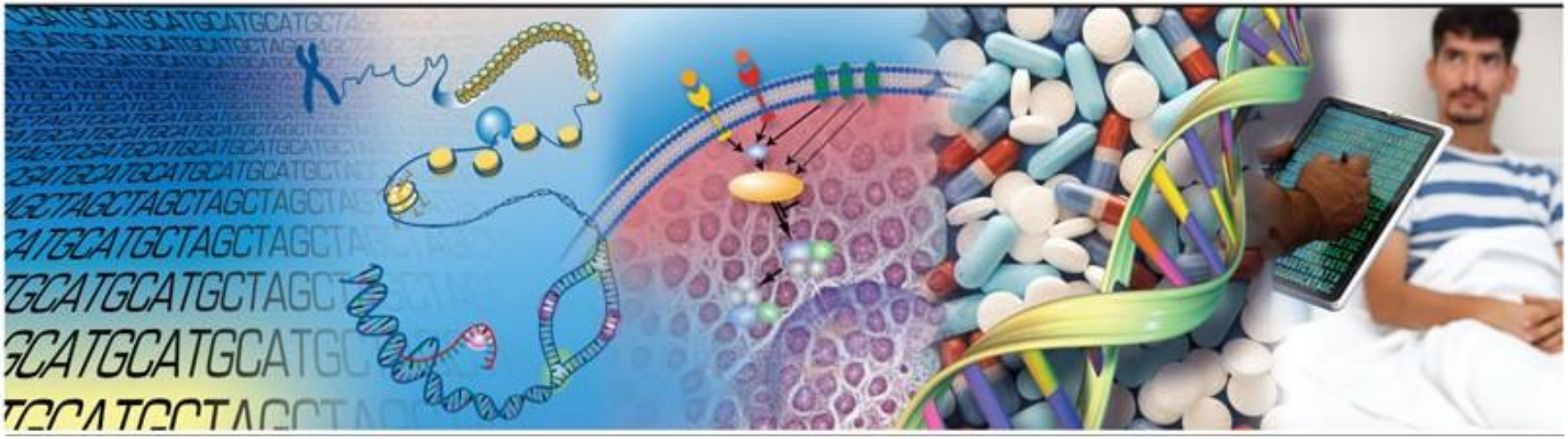


The genotype-first approach is a type of strategy used in genetic epidemiological studies to associate specific genotypes to apparent clinical phenotypes of a complex disease or trait.

“phenotype-first”, the traditional strategy that has been guiding genome-wide association studies (GWAS) so far, this approach characterizes individuals first by a statistically common genotype based on molecular tests prior to clinical phenotypic classification. This method of grouping leads to patient evaluations based on a shared genetic etiology for the observed phenotypes, regardless of their suspected diagnosis. Thus, this approach can prevent initial phenotypic bias and allow for identification of genes that pose a significant contribution to the disease etiology.

Genómica Clínica

Disciplina “emergente” que busca trasladar los hallazgos y métodos de la Genómica (estudios del genoma humano y asociados) a la práctica clínica en sus tres aspectos: Prevención, Diagnóstico y Tratamiento



Descifrar el Genoma

Comprender el Genoma

Vincular Genoma y Enfermedades

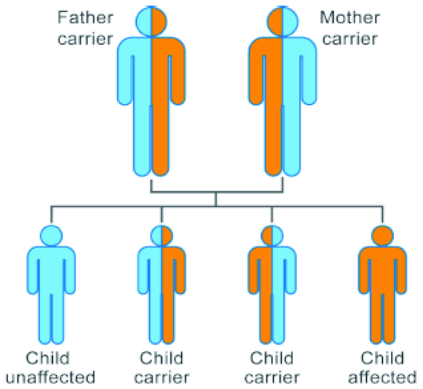
Mejorar el tratamiento y diagnostico

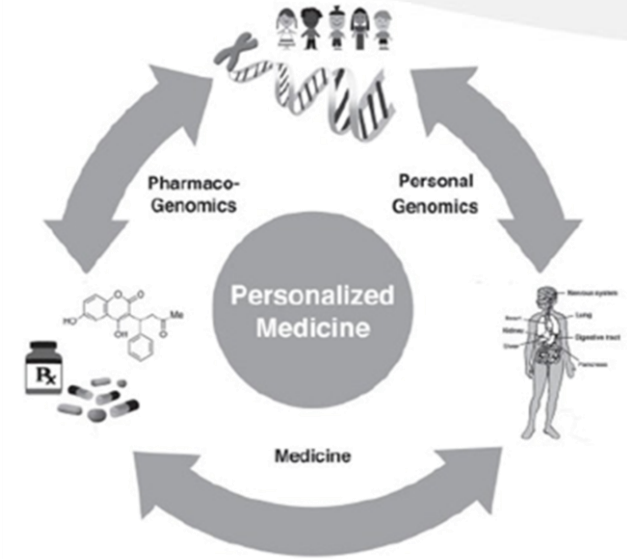
Mejorar la efectividad de los sistemas de salud

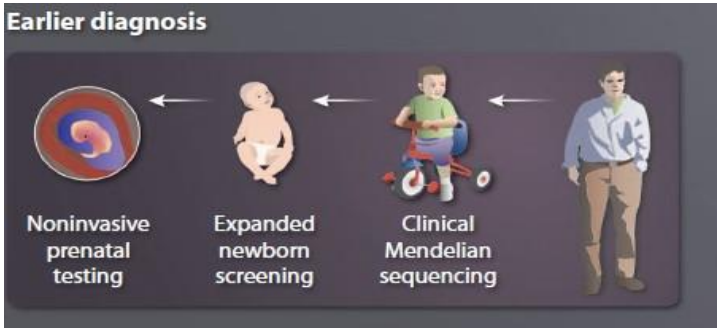
Medicina Personalizada

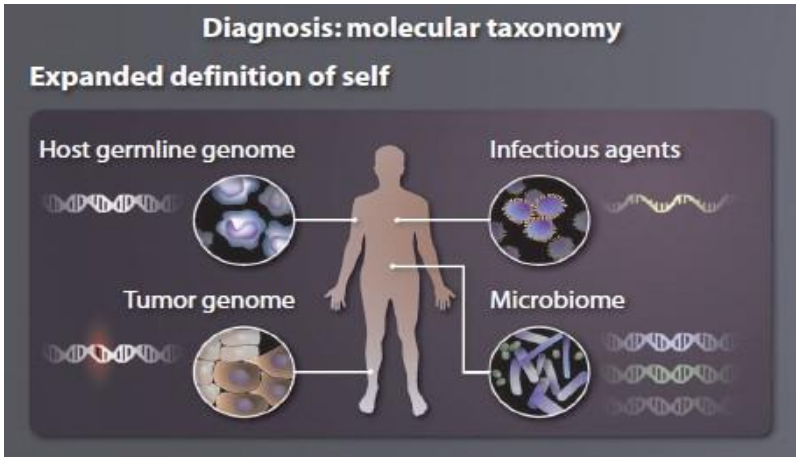
Carrier Status

You are a carrier for 0 conditions







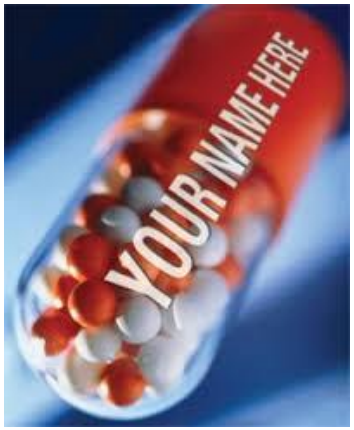


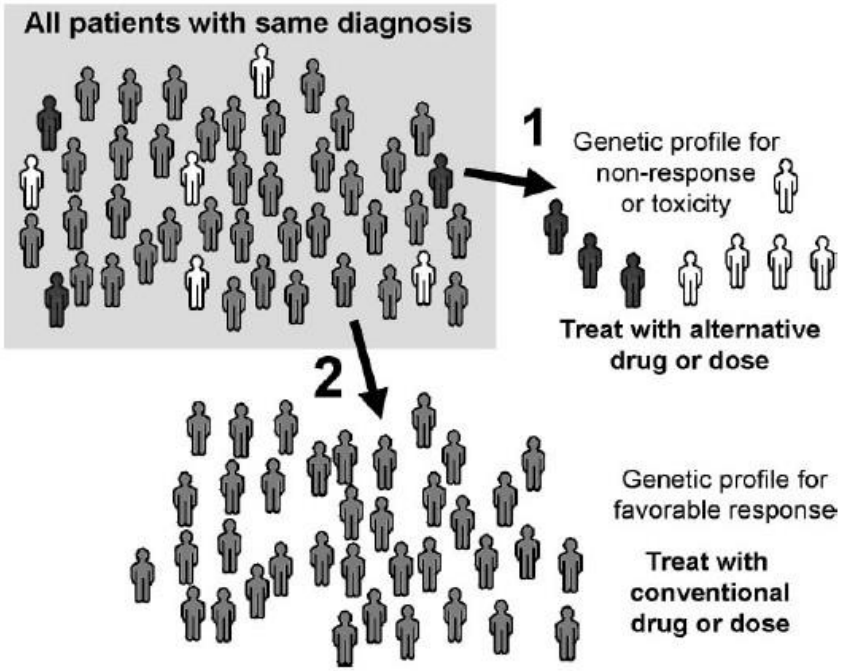
ScienceTranslationalMedicine 12 June 2013 Vol 5 Issue 189 189sr4

Aplicaciones..

Tratamiento!

- “darle la droga correcta la paciente correcto en la dosis correcta”

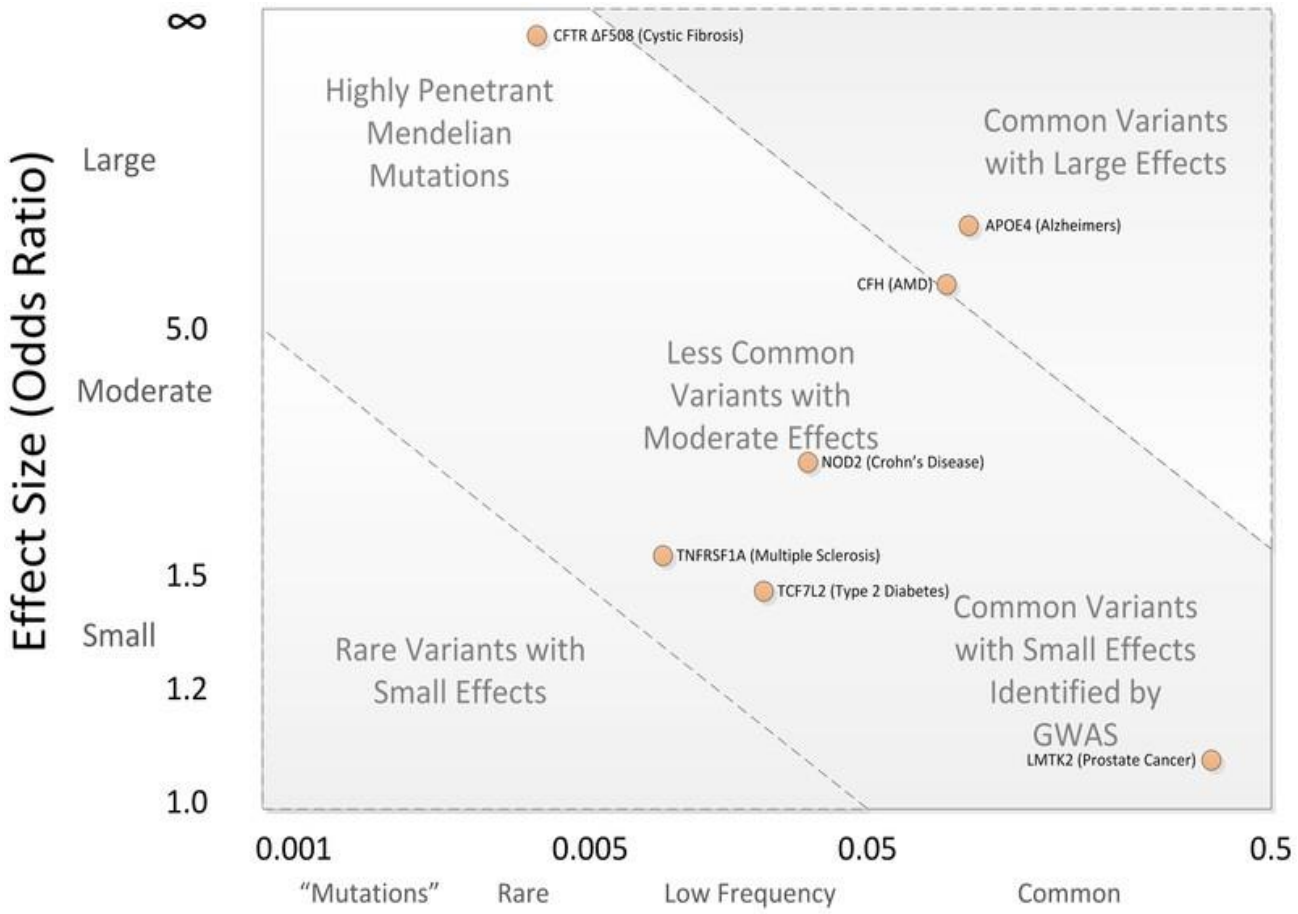




ScienceTranslationalMedicine 12 June 2013 Vol 5 Issue 189 189sr4

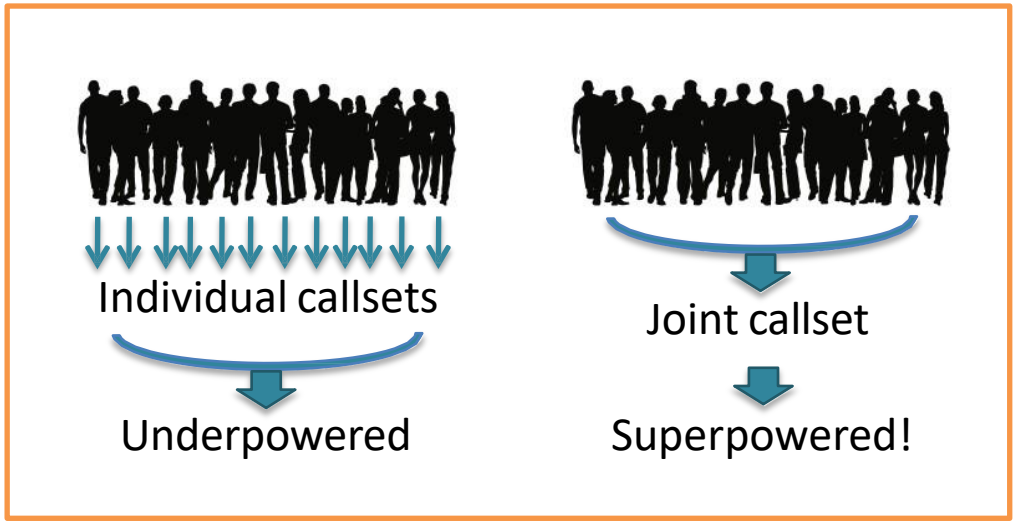
¿Cuál es la contribución de la genética a nuestra salud?

El Odds Ratio mide la probabilidad de riesgo relativa de contraer la enfermedad al portar la variante



Joint analysis empowers discovery

- Single genome in isolation: almost never useful
- Family or **population** data add valuable information
  - rarity of variants
  - *de novo* mutations
  - ethnic background



Human genome reference build GRCh38

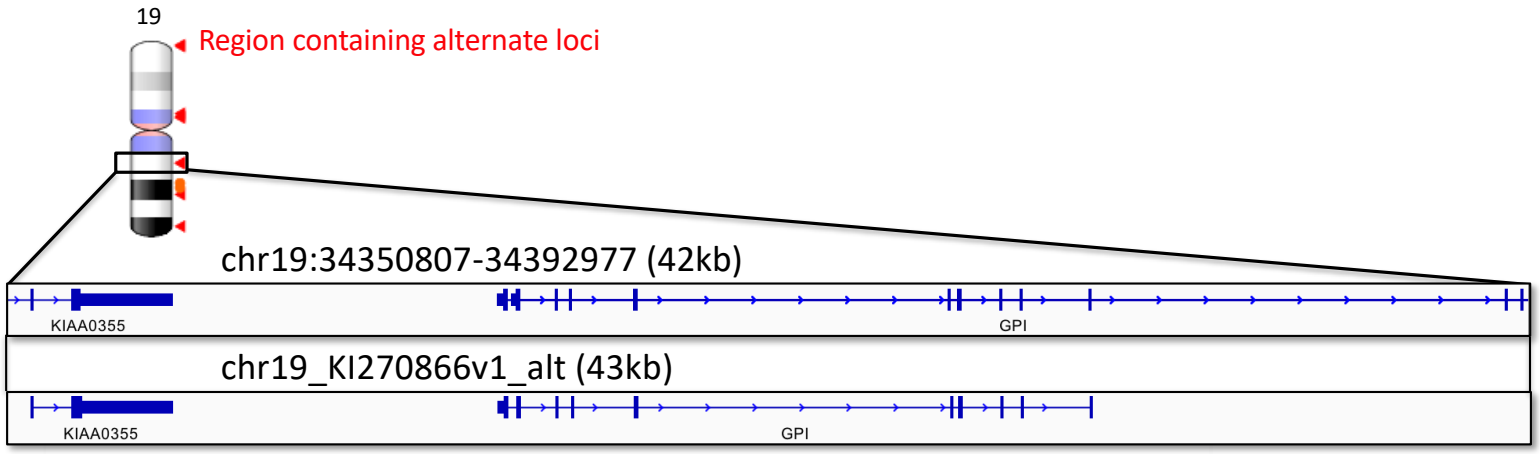
NCBI Build 34/hg16 (2003)

Iterative refinement

GRCh38 (2013)

Improvements to the reference aim to:

- Better account for population diversity
- Enable detection of complex variants
- Reflect copy number & pseudo autosomal regions
- Quarantine low complexity & extraneous sequences



** Image from GRC <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/> modified by S.H. Lee 2016*

Types of Genetic Variation

Genetic Diseases are driven by genomic alterations like:

- Single Nucleotide Aberrations
  - Single Nucleotide Polymorphisms (**SNP**s) - mutations shared amongst a population
  - Single Nucleotide Variations (**SNV**s) - private mutations
- Short Insertions or Deletions (**indels**)
- Copy Number Variations (**CNVs**)
- Larger Structural Variations (**SVs**)

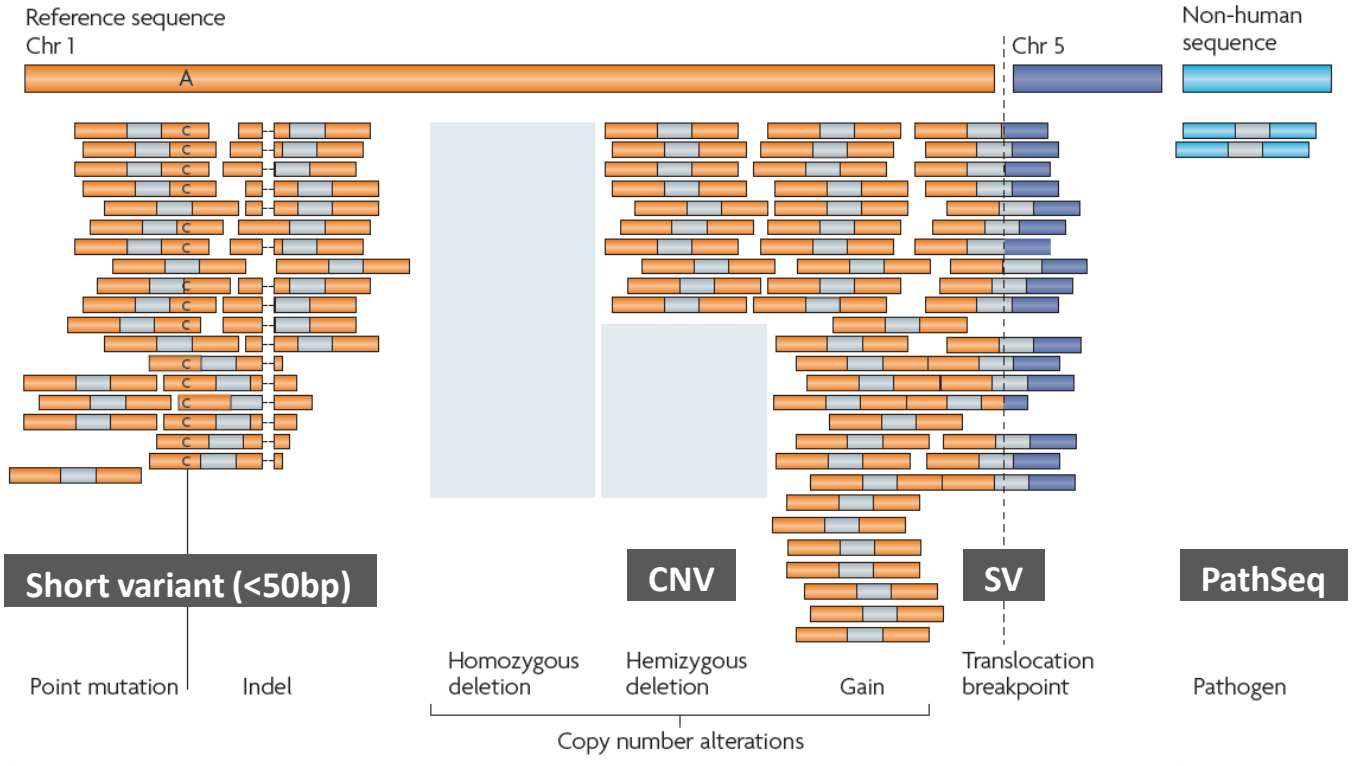
SNPs vs. SNVs

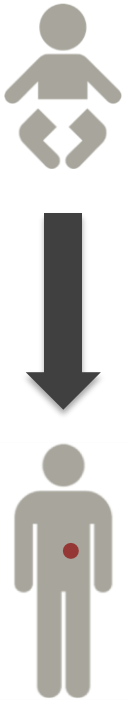
Both are aberrations at a single nucleotide

- **SNP**
  - Aberration expected at the position for any member in the species (well-characterized)
  - Occur in population at some frequency so expected at a given locus
  - Validated in population
  - Catalogued in dbSNP (http://www.ncbi.nlm.nih.gov/snp)
- **SNV**
  - Aberration seen in only one individual (not well characterized)
  - Occur at low frequency so not common
  - Not validated in population

Really a matter of frequency of occurrence

Different types of genomic variants

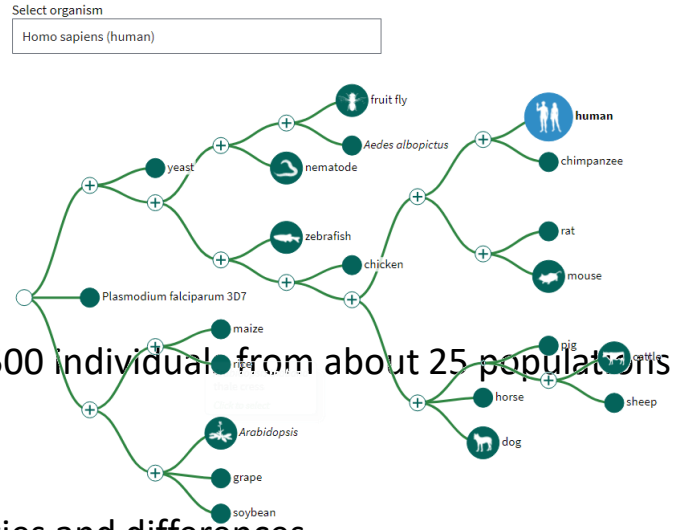


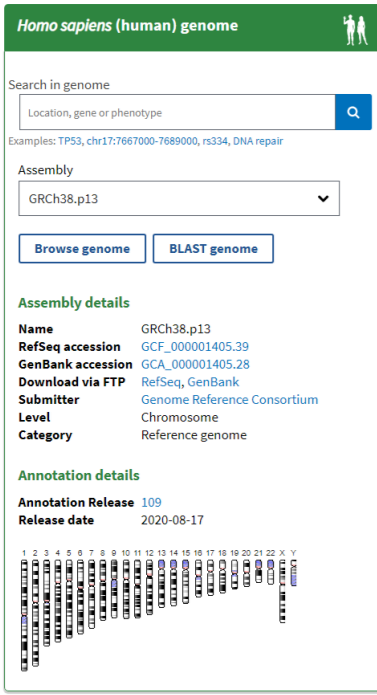


Germline

Somatic

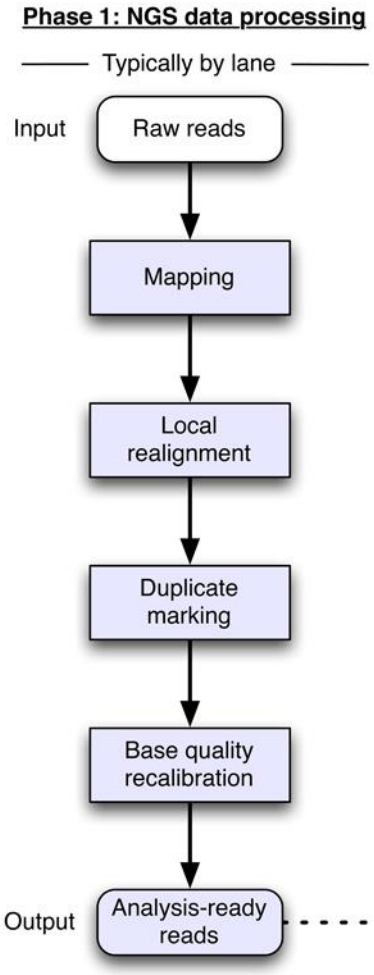
Databases
Catalogs of human genetic variation





- **The 1000 Genomes Project**
  - http://www.1000genomes.org/
  - SNPs and structural variants from 2500 individuals from about 25 populations
- **HapMap**
  - http://hapmap.ncbi.nlm.nih.gov/
  - identify and catalog genetic similarities and differences
- **dbSNP**
  - http://www.ncbi.nlm.nih.gov/snp/
  - Database of SNPs and multiple small-scale variations
- **COSMIC**
  - http://www.sanger.ac.uk/genetics/CGP/cosmic/
  - Catalog of Somatic Mutations in Cancer
- **TCGA**
  - http://cancergenome.nih.gov/
  - The Cancer Genome Atlas researchers are mapping the genetic changes in 20 selected cancers
- **ClinVar**
  - http://www.ncbi.nlm.nih.gov/clinvar/
  - aggregates information about sequence variation and its relationship to human health

A framework for variation discovery



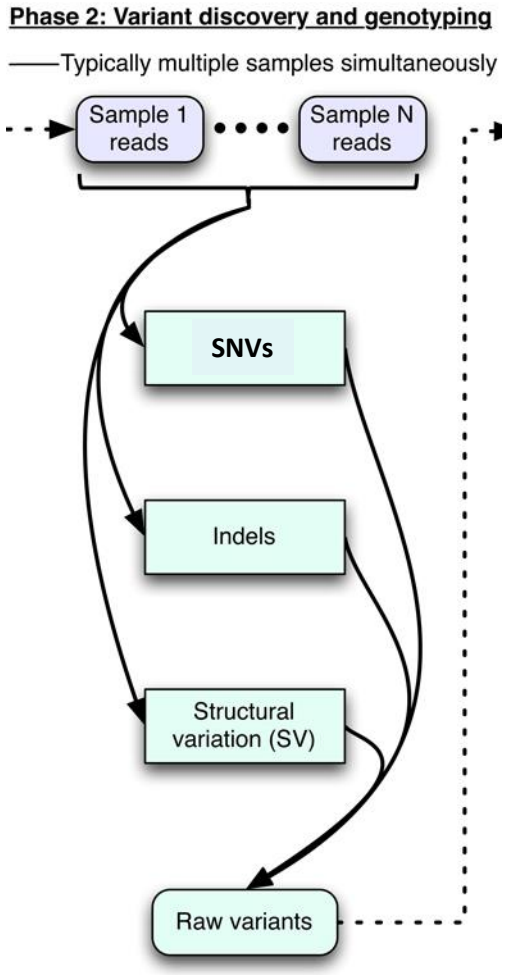
Phase 1: Mapping

- Place reads with an initial alignment on the reference genome using mapping algorithms
- Refine initial alignments
  - local realignment around indels
  - molecular duplicates are eliminated
- Generate the technology-independent SAM/BAM alignment map format

Accurate mapping crucial for variation discovery

DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

A framework for variation discovery

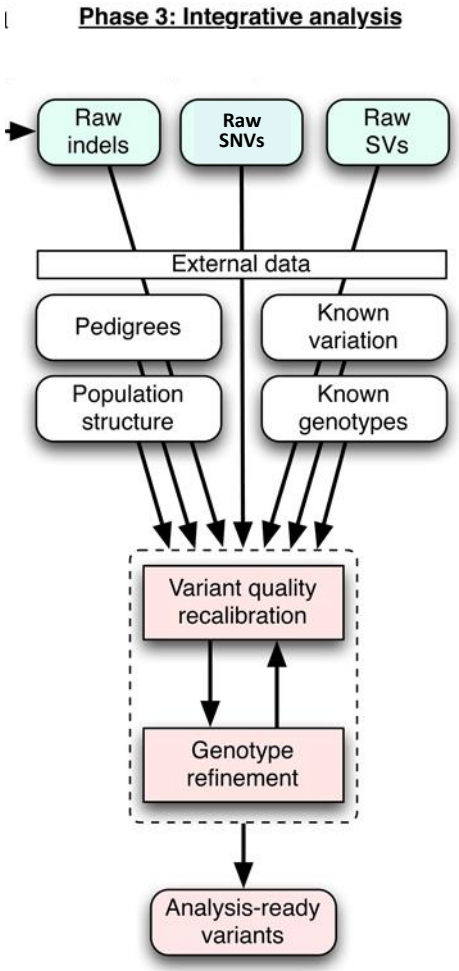


Phase 2: Discovery of raw variants

- Analysis-ready SAM/BAM files are analyzed to discover all sites with statistical evidence for an alternate allele present among the samples
- SNPs, SNVs, short indels, and SVs

DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

A framework for variation discovery

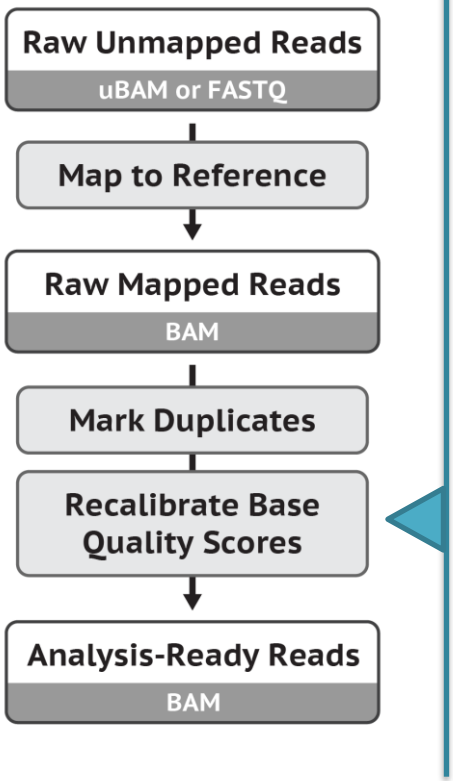


Phase 3: Discovery of analysis-ready variants

- technical covariates, known sites of variation, genotypes for individuals, linkage disequilibrium, and family and population structure are integrated with the raw variant calls from Phase 2 to separate true polymorphic sites from machine artifacts
- at these sites high-quality genotypes are determined for all samples

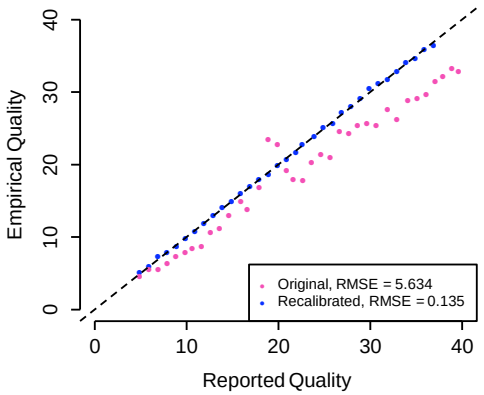
DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

Step 3: Base Recalibration (BQSR) corrects for machine errors

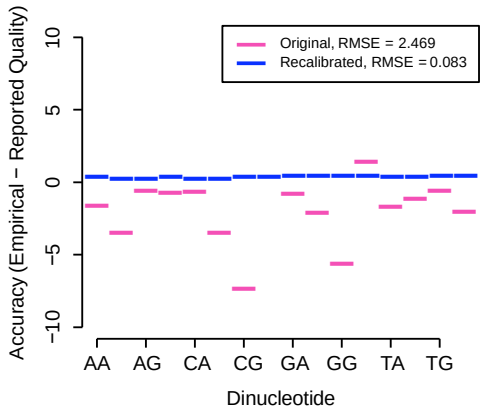


- Sequencers make systematic errors in base quality scores
- Sequencer quality cannot include PCR-based errors
- BQSR corrects the quality scores (not the bases)

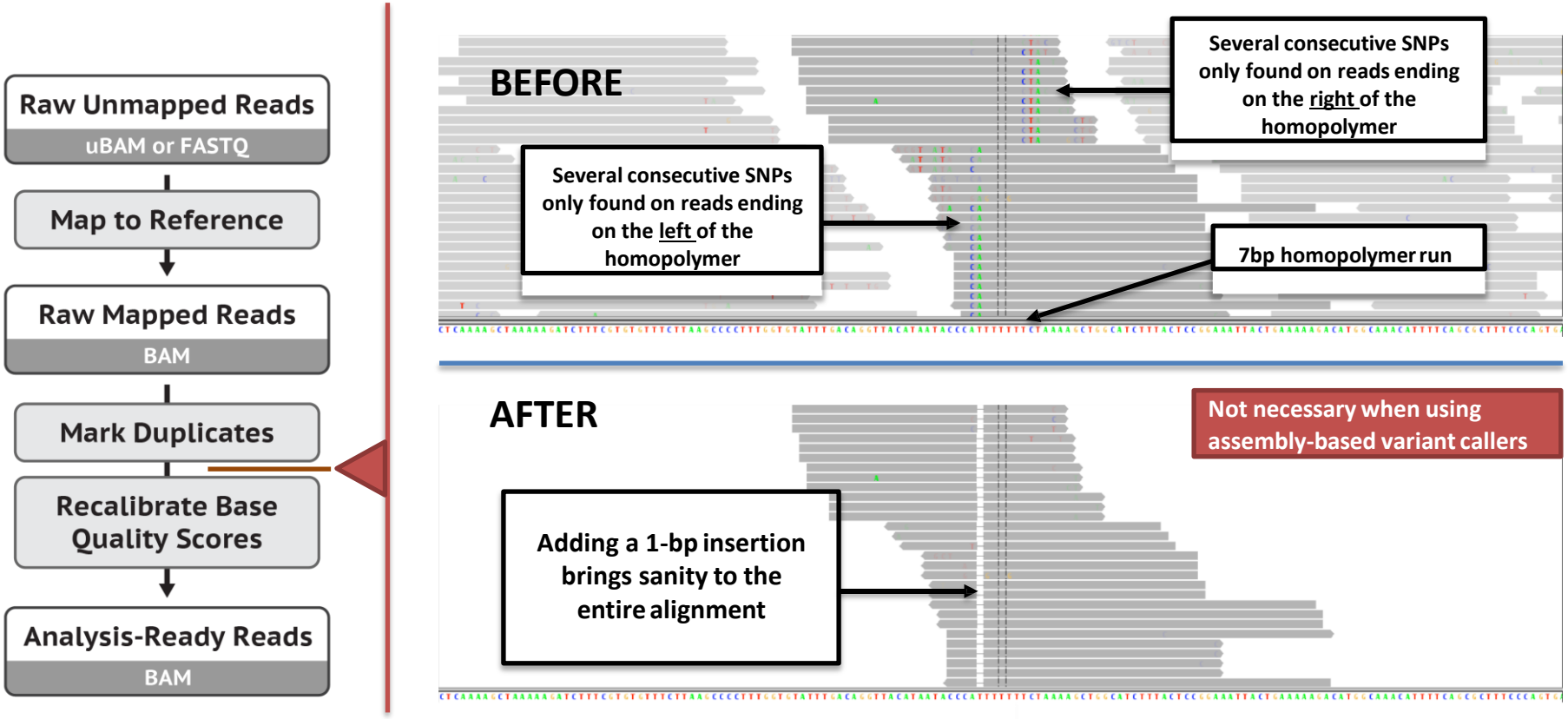
Overall effect of bias



Bias on nucleotide context

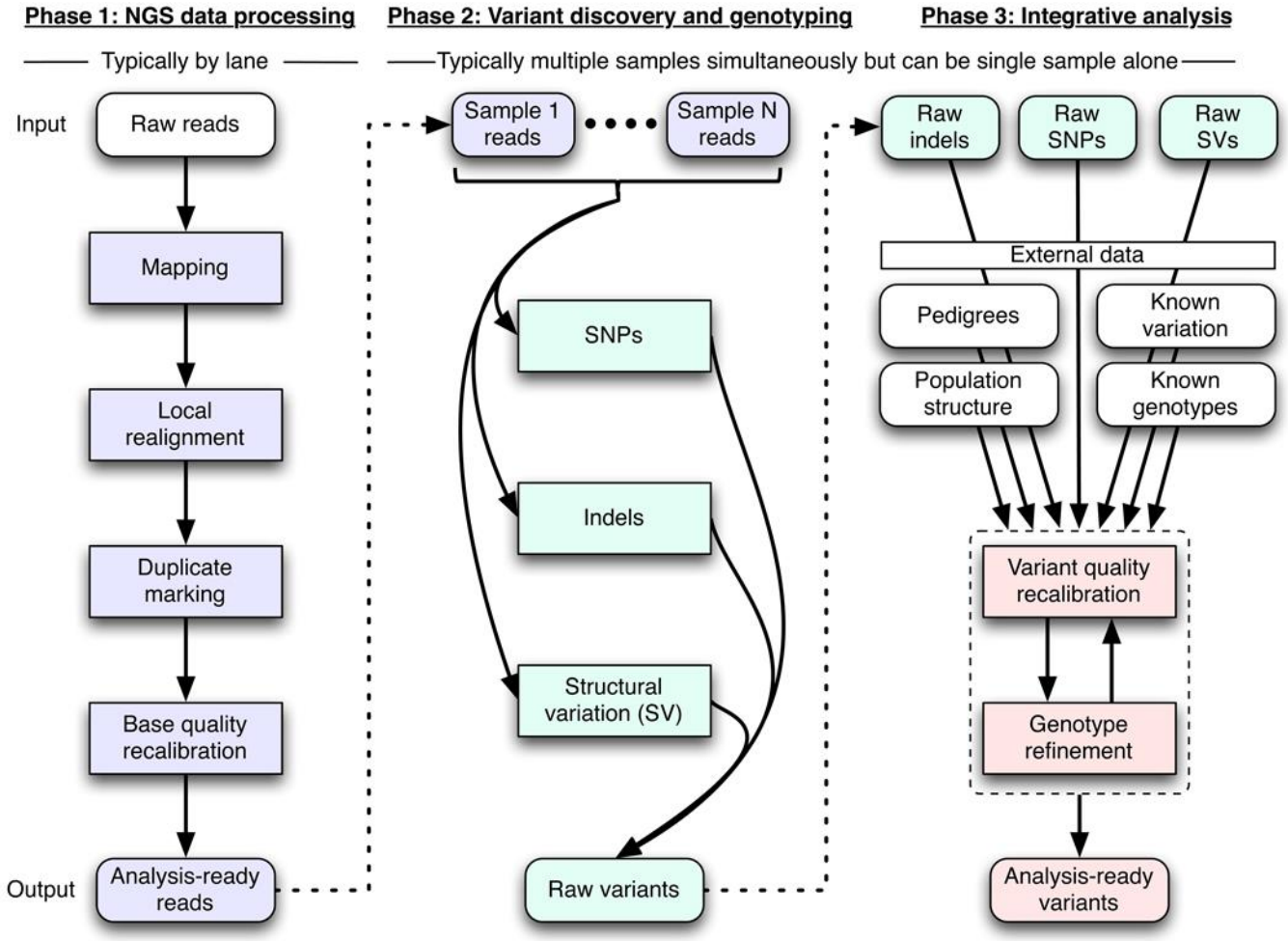


DEPRECATED: Local realignment around indels



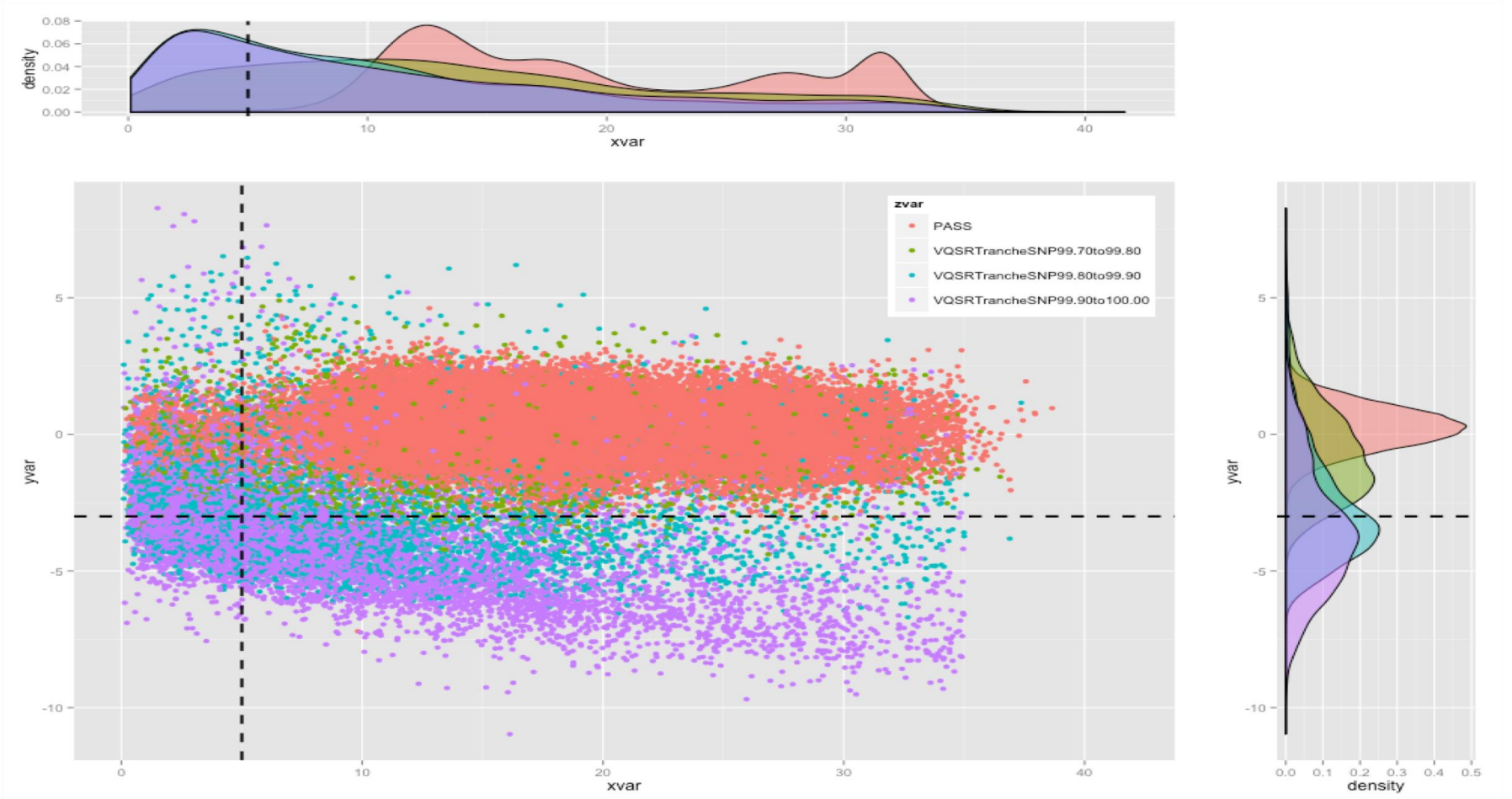
** For implications, see <https://gatkforums.broadinstitute.org/gatk/discussion/7847/changing-workflows-around-calling-snps-and-indels>*

A framework for variation discovery

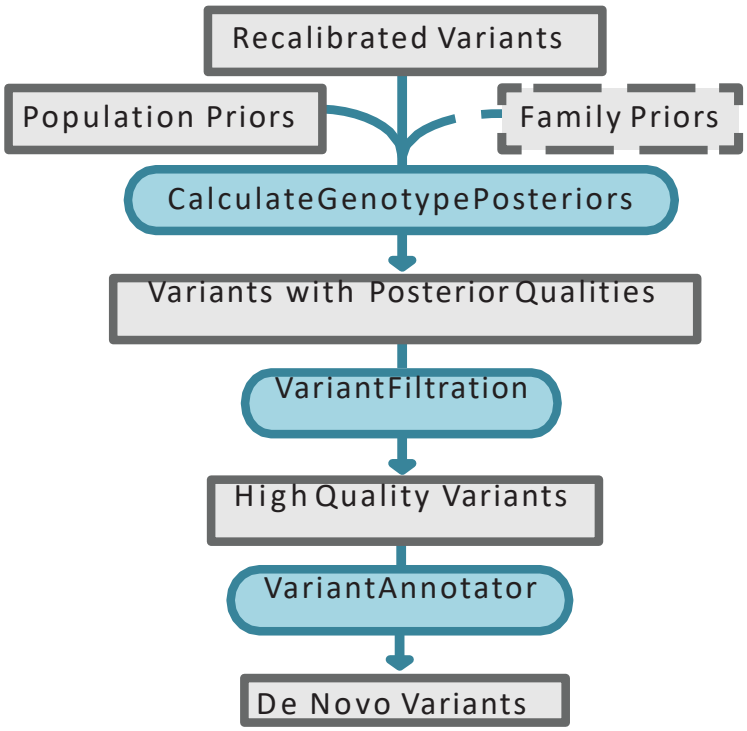


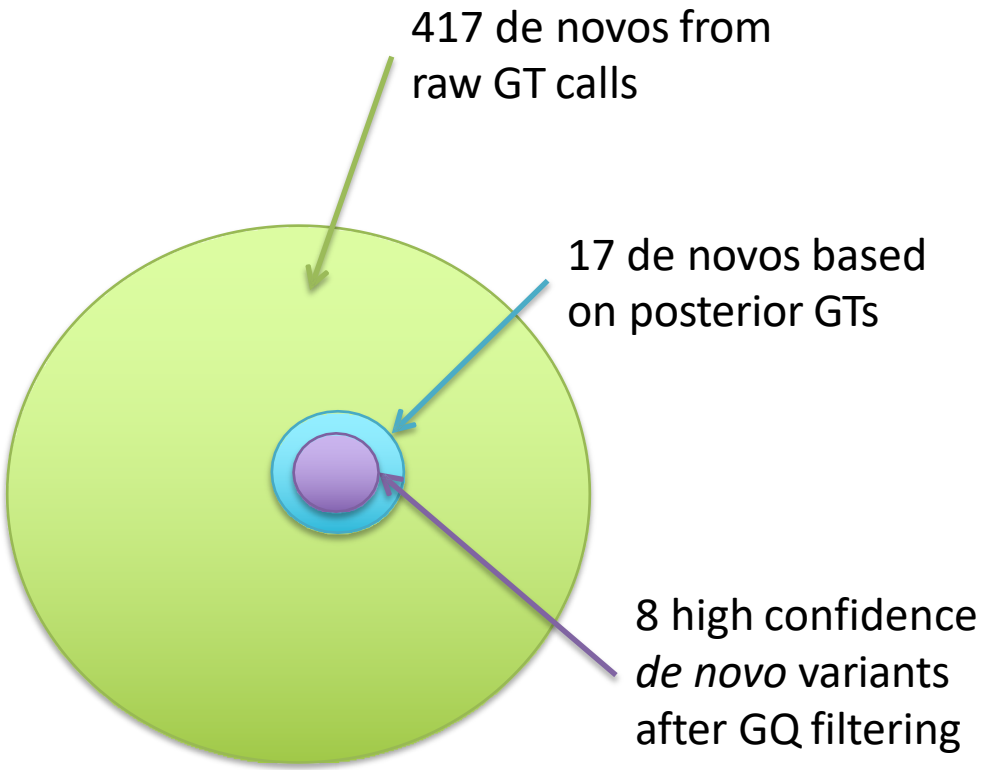
DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

Variant filtering reduces false positives



Genotype refinement improves GT quality and *de novo* calls





Some variant callers

Name	Category	Tumor/Normal Pairs	Metric	Reference
JointSNVMix (Fisher)	Allele Counting	Yes	Somatic probability	Roth, A. et al. (2012)
SomaticSniper	Heuristic	Yes	Somatic Score	Larson, D.E. et al. (2012)
VarScan2	Heuristic with allele counting	Yes	Somatic p-value	Koboldt, D. et al. (2012)
GATK UnifiedGenotyper	Bayesian	No	Phred QUAL	DePristo, M.A. et al. (2011)
Strelka	Bayesian	Yes	Somatic probability	Saunders, C.T. et al. (2012)
MuTect	Bayesian	Yes	Log odds score (LOD)	Cibulskis, K. et al. (2013)

VCF (Variant Call Format) is a standard file format for representing variant calls

Roth, A. et al. JointSNVMix : A Probabilistic Model For Accurate Detection Of Somatic Mutations In Normal/Tumour Paired Next Generation Sequencing Data. Bioinformatics (2012).

Larson, D.E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 28(3):311-7 (2012).

Koboldt, D. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22(3):568-76. doi: 10.1101/gr.129684.111 (2012).

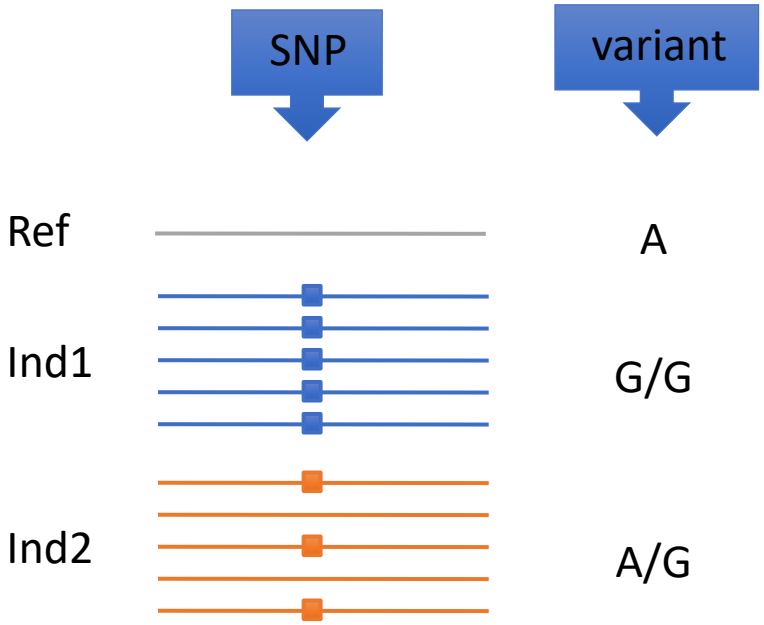
DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

Saunders, C.T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28(14):1811-7. doi : 10.1093/bioinformatics/bts271 (2012).

Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 31(3):213-9. doi : 10.1038/nbt.2514 (2013).

Variant calling methods

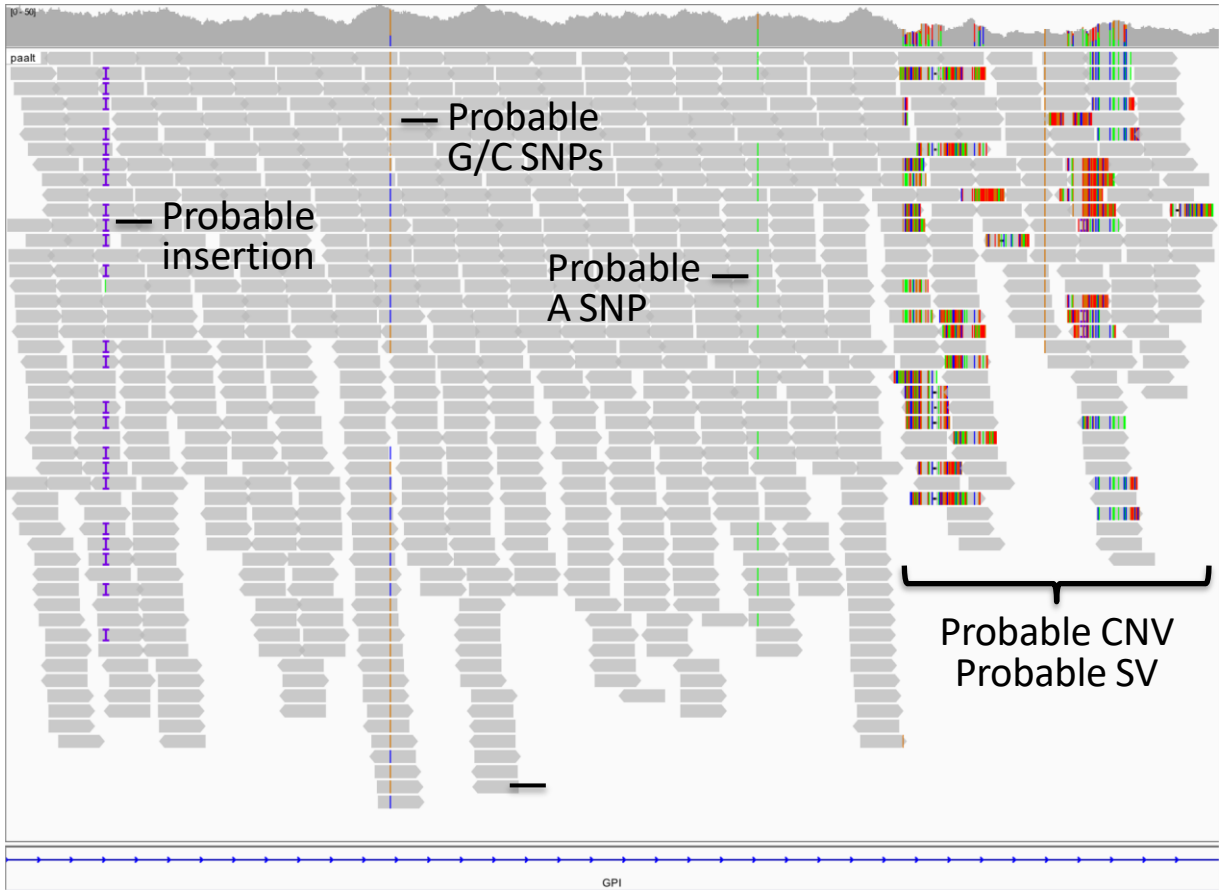
- > 15 different algorithms
- Three categories
  - Allele counting
  - Probabilistic methods, e.g. Bayesian model
    - to quantify statistical uncertainty
    - Assign priors based on observed allele frequency of multiple samples
  - Heuristic approach
    - Based on thresholds for read depth, base quality, variant allele frequency, statistical significance



Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011 Jun;12(6):443-51. PMID: 21587300.

http://seqanswers.com/wiki/Software/list

What variants look like in a genome browser



Top track: Depth of coverage

Non-reference bases are colored; reference bases are grey

A C G T

Reference genome

Individual aligned read

http://software.broadinstitute.org/gatk/

Variants are reported in VCF (Variant Call Format)

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM POS    ID         REF ALT   QUAL   FILTER INFO                FORMAT      NA00001    NA00002    NA00003
20    14370    rs6054257 G   A     29     PASS   DP=14;AF=0.5   GT:GQ:DP    0/0:48:1   1/0:48:8   1/1:43:5
20    1230237  .         T   .     47     PASS   DP=13          GT:GQ:DP    0/0:54:7   0/0:48:4   0/0:61:2
20    1234567  .         GT  G     50     PASS   DP=9           GT:GQ:DP    0/1:35:4   0/2:17:2   1/1:40:3
```

Header

Records

From format specification in https://samtools.github.io/hts-specs/VCFv4.2.pdf

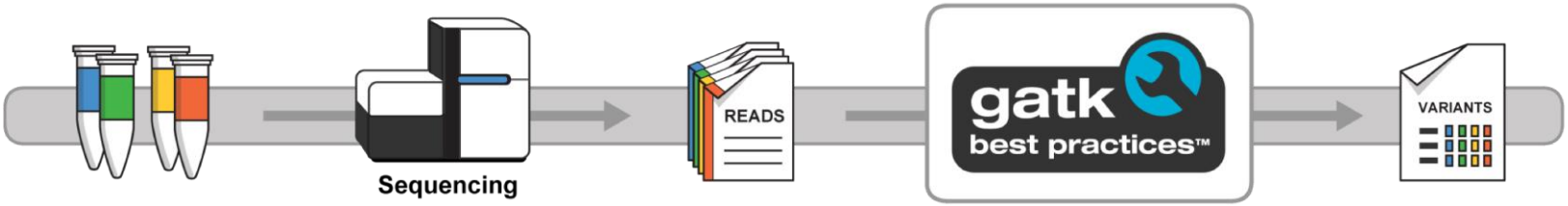
VCF format supports CNVs and SVs

```
...
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly fil
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
###INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM POS ID    REF   ALT   QUAL       FILTER     INFO       FORMAT     NA00001
1 2827694  rs2376870 CGTGGATGCGGGGAC  C     . PASS SVTYPE=DEL;END=2827708;HOMLEN=1;HOMSEQ=G;SVLEN=-14 GT:GQ
2  321682        .     T               <DEL> 6 PASS SVTYPE=DEL;END=321887;SVLEN=-205;CIPOS=-56,20;CIEND=-10,6
3 12665100       .     A               <DUP> 14 PASS SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND
```

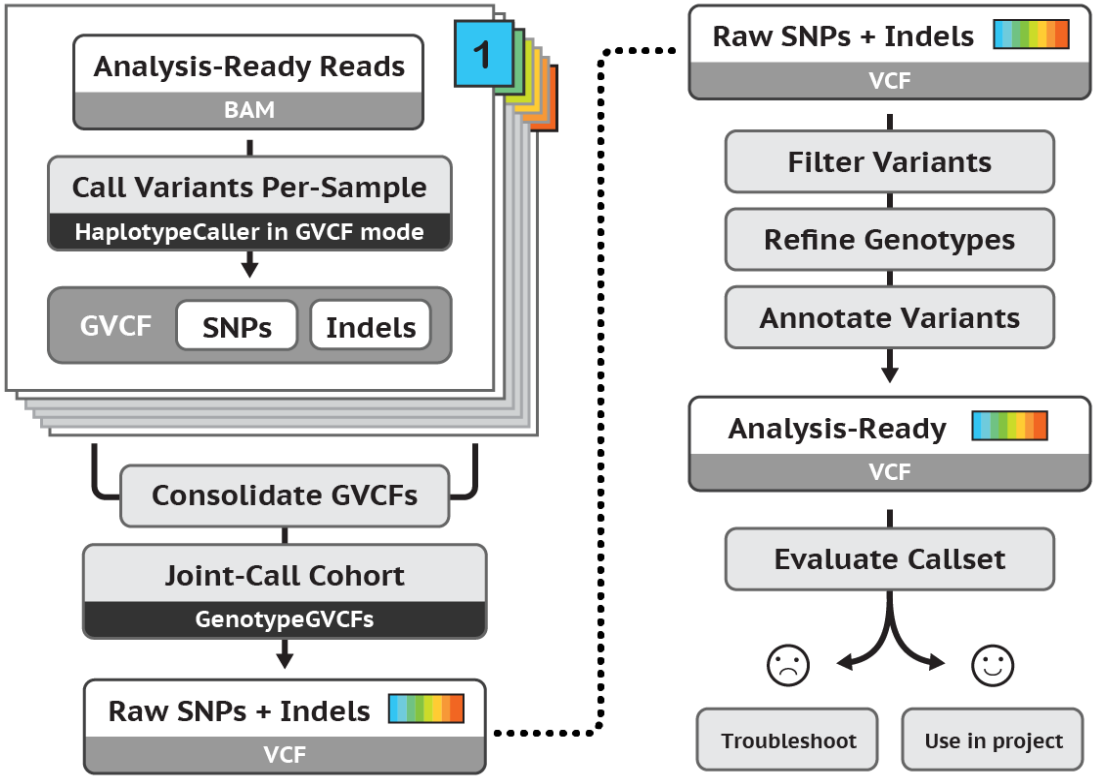
Symbolic allele in angle-bracketed ID

From format specification at https://samtools.github.io/hts-specs/VCFv4.2.pdf

Workflows for all major variant classes

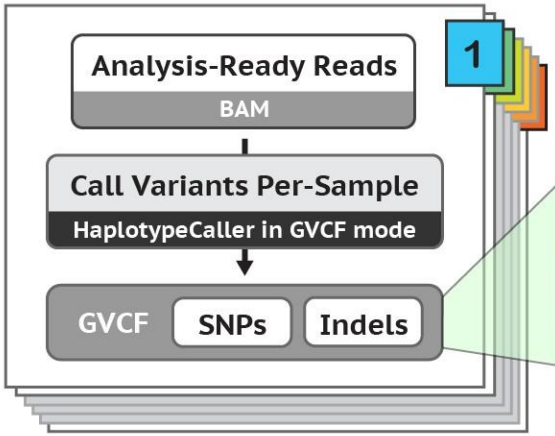


	GERMLINE	SOMATIC
SNPs & INDELs	**HaplotypeCaller GVCF**	**Mutect2**
Copy Number	**GATK gCNV**	**GATK CNV + aCNV**
Structure Variation	**GATK SVDiscovery** (beta)	(planned)



https://software.broadinstitute.org/gatk/best-practices

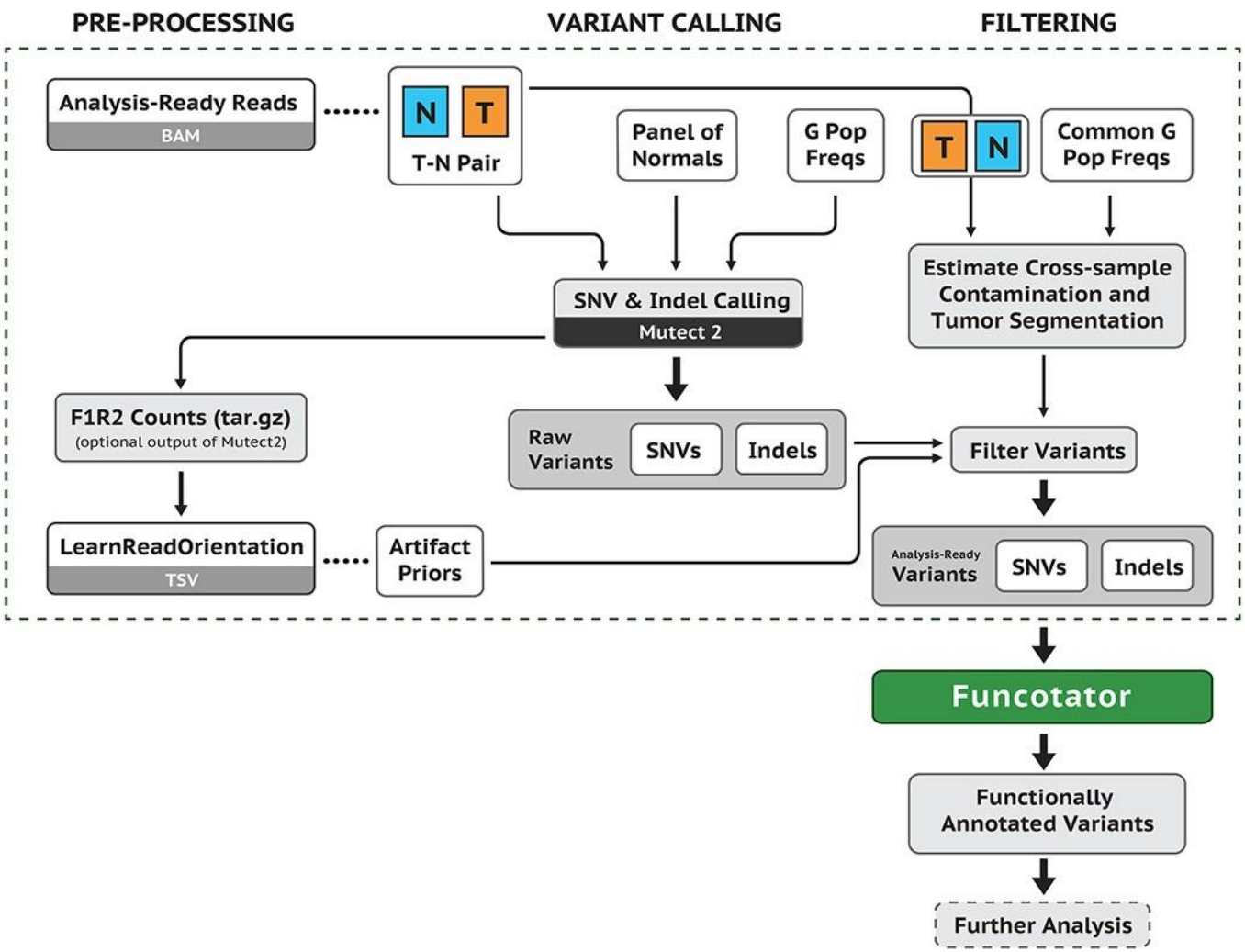
From per-sample GVCFs to final multi-sample VCF



Generate per-sample Genomic VCFs (GVCFs)
then joint-call across all samples -> final VCF

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 . A <NON_REF> . . END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 . C T,<NON_REF> 326.77 . BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB 0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 . A <NON_REF> . . END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

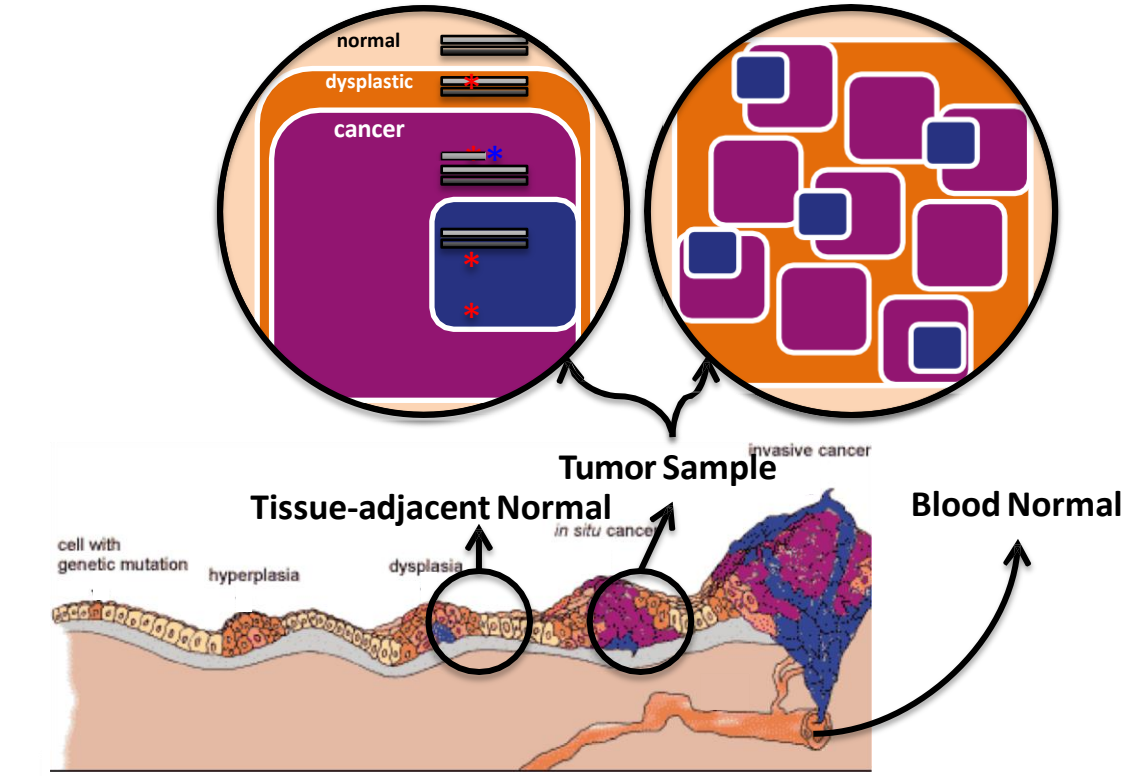
```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12877 NA12878 NA12882
20 10000117 . C T 1606.16 . AC=4;AF=0.667;AN=6;BaseQRankSum=1.66;ClippingRankSum=0.340;DP=85;
FS=5.718;MLEAC=4;MLEAF=0.667;MQ=60.36;MQRankSum=1.45;QD=18.90;ReadPosRankSum=1.62;SOR=1.503
GT:AD:DP:GQ:PL 0/1:17,15:32:99:399,0,439 0/1:11,12:23:99:291,0,292 1/1:0,30:30:90:948,90,0
20 10000211 . C T 1765.16 . AC=4;AF=0.667;AN=6;BaseQRankSum=2.34;ClippingRankSum=-1.147e+00;
DP=97;FS=0.809;MLEAC=4;MLEAF=0.667;MQ=60.00;MQRankSum=1.21;QD=18.58;ReadPosRankSum=0.152;SOR=0.831
GT:AD:DP:GQ:PL 0/1:13,10:23:99:243,0,341 0/1:21,14:35:99:355,0,526 1/1:0,37:37:99:1199,111,0
20 10000439 . T G 1982.13 . AC=5;AF=0.833;AN=6;BaseQRankSum=1.31;ClippingRankSum=0.549;DP=103;
FS=0.000;MLEAC=5;MLEAF=0.833;MQ=60.00;MQRankSum=0.972;QD=19.82;ReadPosRankSum=1.56;SOR=0.839
GT:AD:DP:GQ:PL 0/1:18,12:30:99:208,0,455 1/1:0,29:29:86:795,86,0 1/1:1,40:41:99:1010,110,0
```



Genetic variants DB

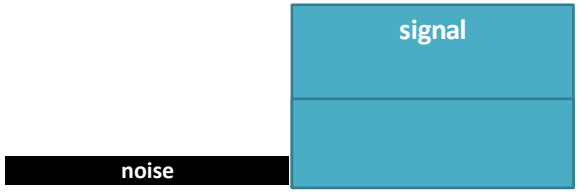
http://software.broadinstitute.org/gatk/

KEY CHALLENGES : tumor heterogeneity and contamination



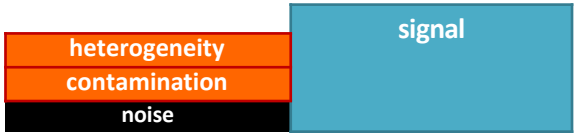
Adapted from https://science.education.nih.gov/supplements/nih1/cancer/guide/understanding1.html

Expectation for germline variants



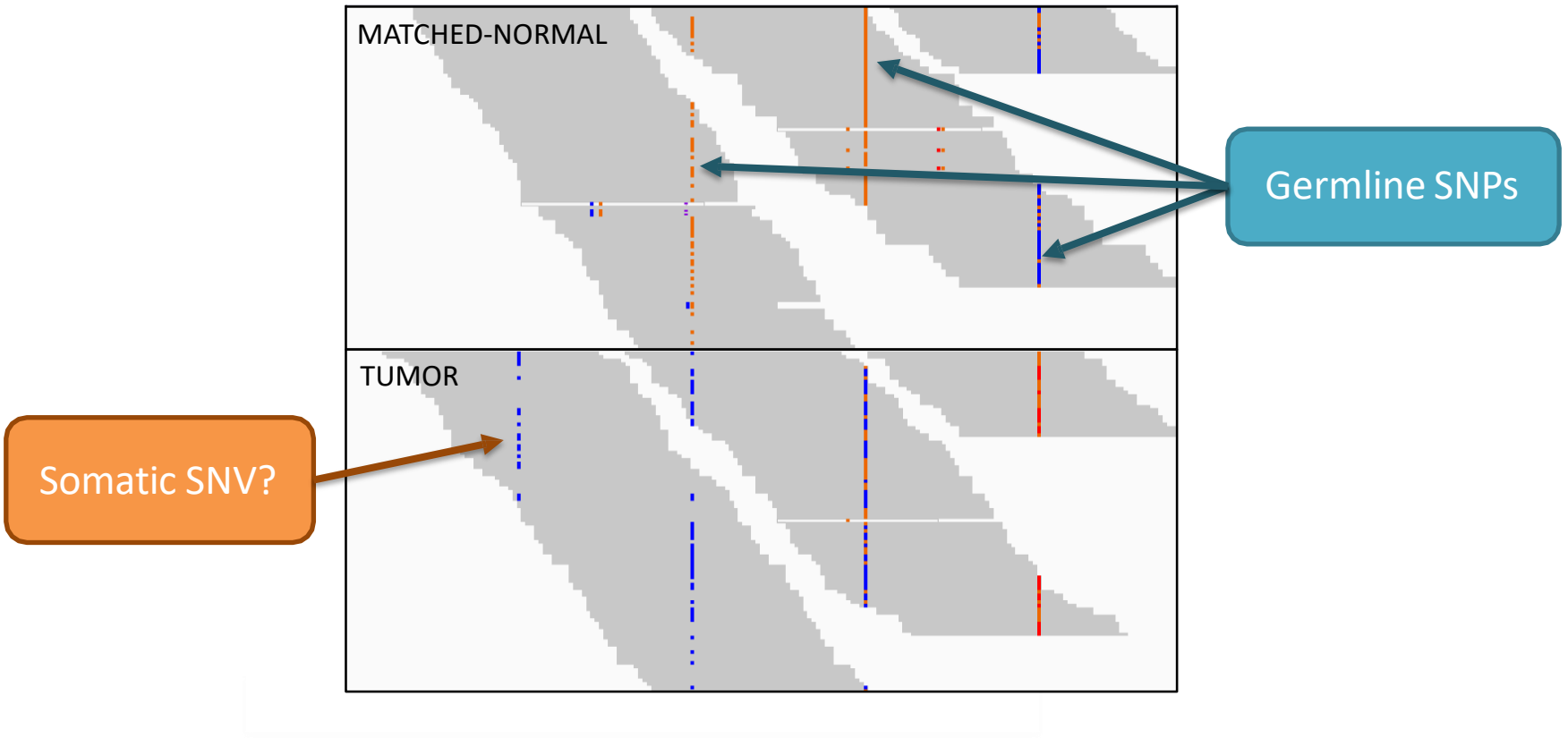
+ AF expected to follow ploidy

Expectation for somatic variants



+ no reliance on ploidy for AF

Subtracting germline variants with a matched normal



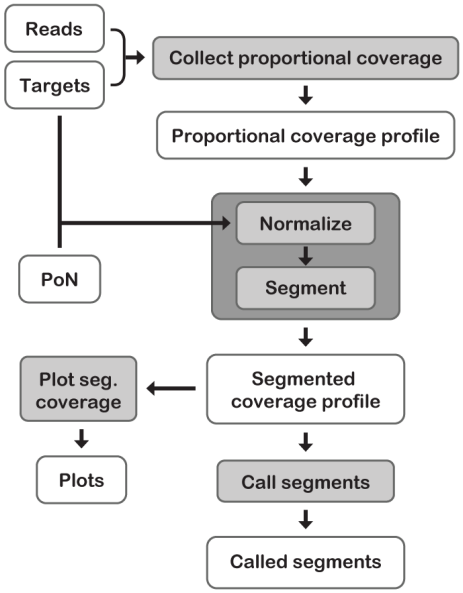
Challenges of accurate somatic variant calling

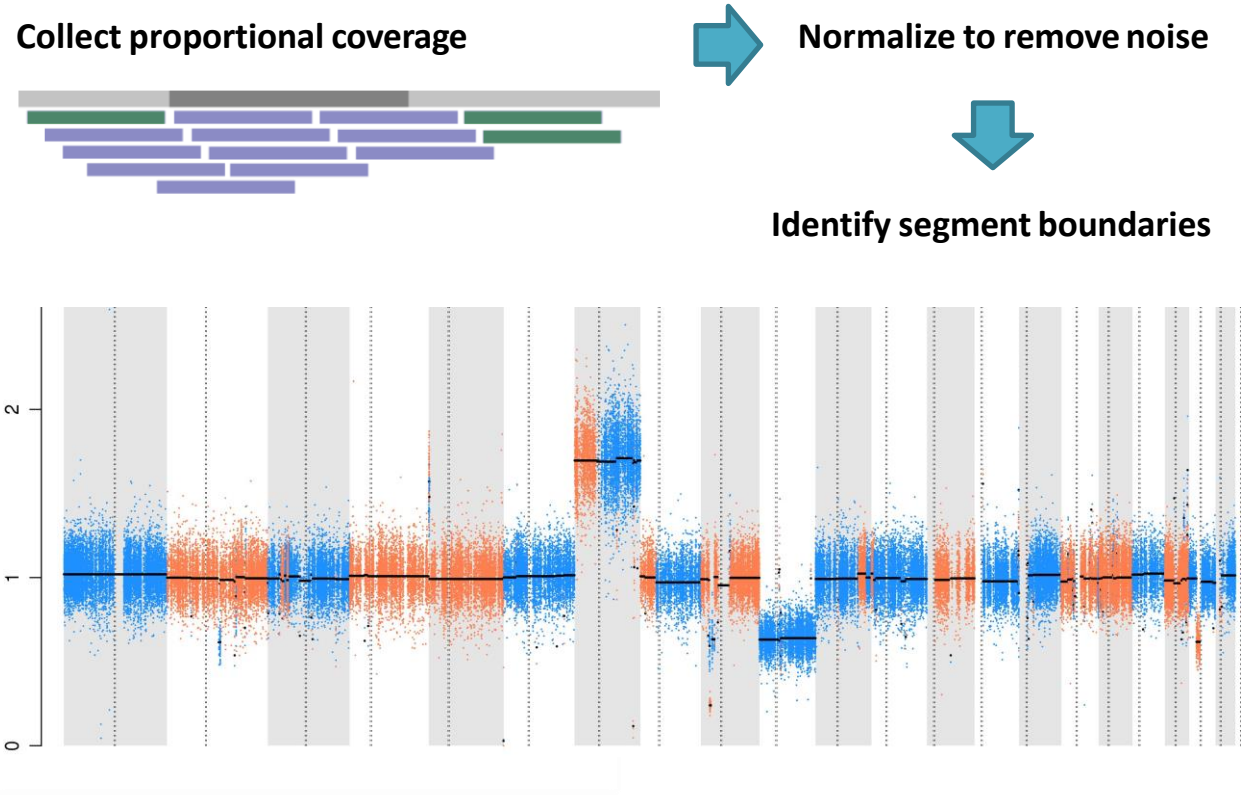
Not as simple as identifying sites with a variant allele in the tumor not present in the normal

- Artifacts from PCR amplification or targeted (exome) capture
- Machine sequencing errors
- Incorrect local alignment of reads
- Tumor heterogeneity
- Tumor-normal cross-contamination

Copy number variants cause coverage imbalance

Counting and normalizing coverage to identify alterations





Germline Structural Variants (SVs) are especially hairy beasts

Any variant that affects 50bp or more of sequence