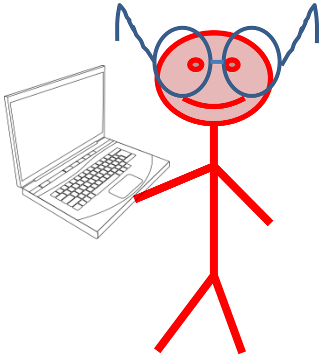
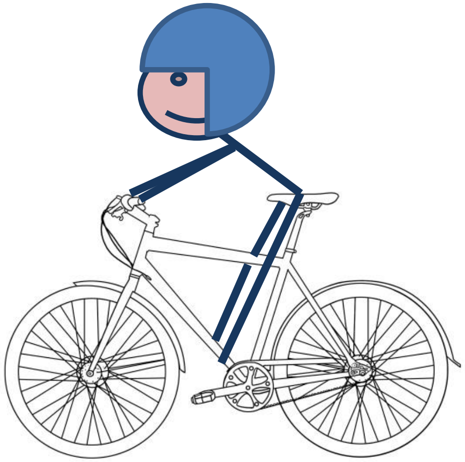
Genetic variation analysis: variant calling and annotations





Vincenza Colonna

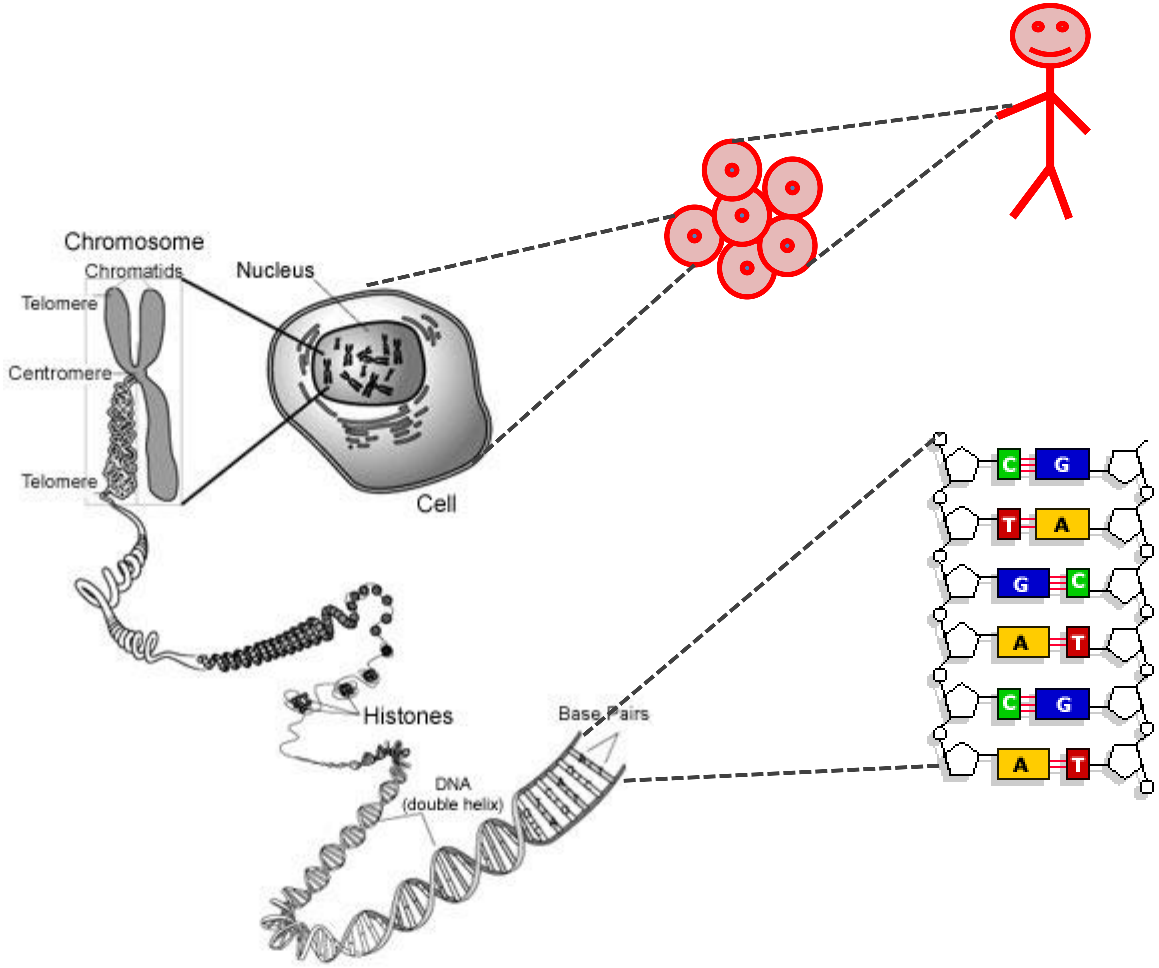
InterOmics Tutorial Day

14 Novembre 2013

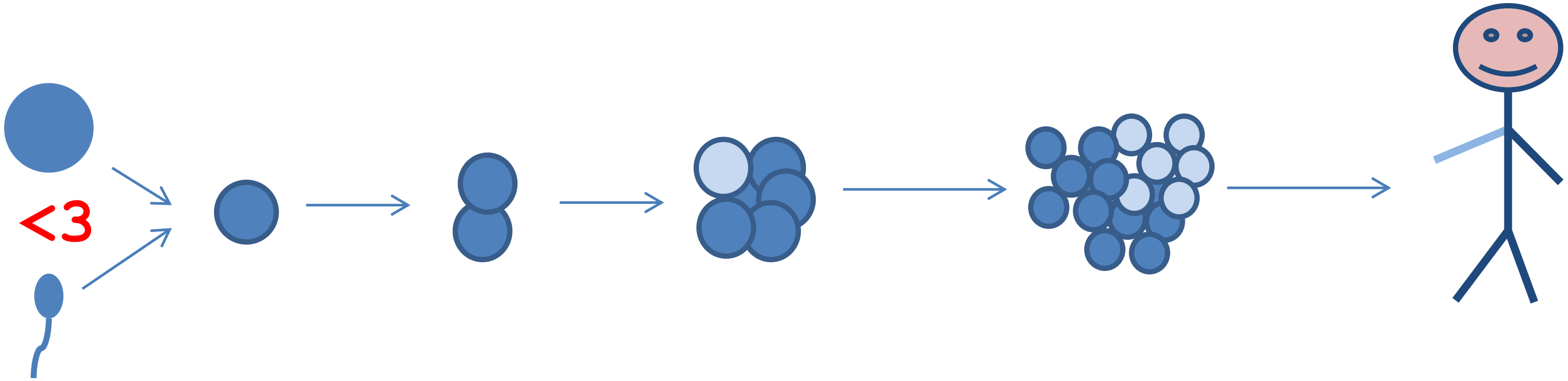
Area di Ricerca CNR, Via Castellino 111, Napoli

- Understanding the genomic variability in five minutes
- Few details on whole genome sequencing
- Variant detection – variant annotation
- Practical session

Where is the genome?

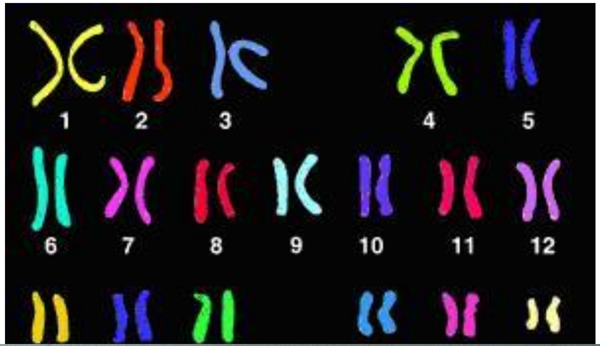


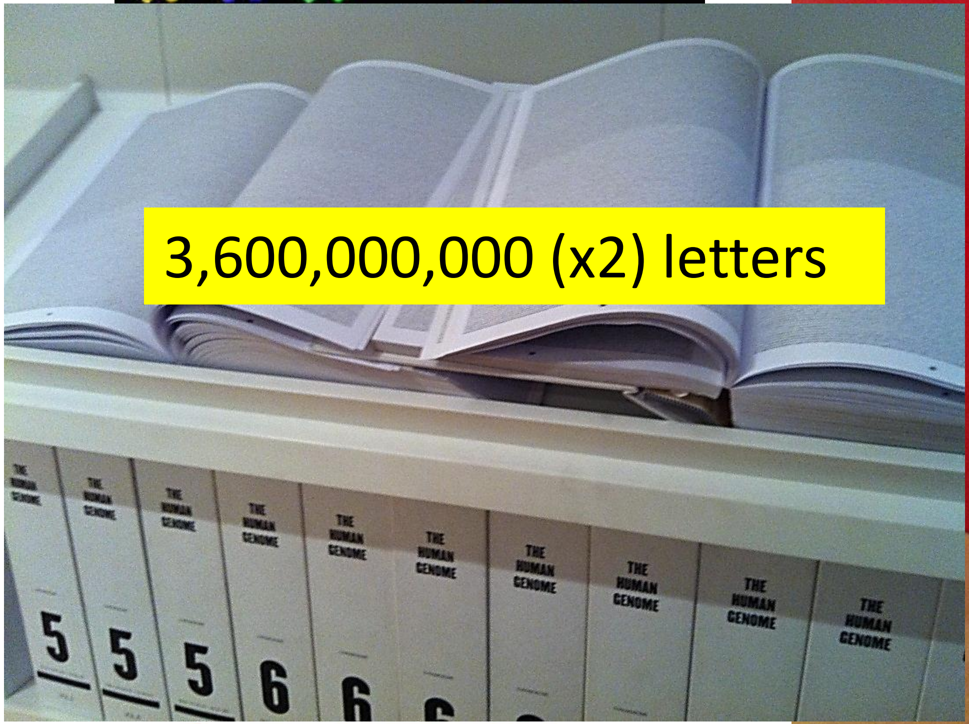
Does all the cells have the same genome in one organism?



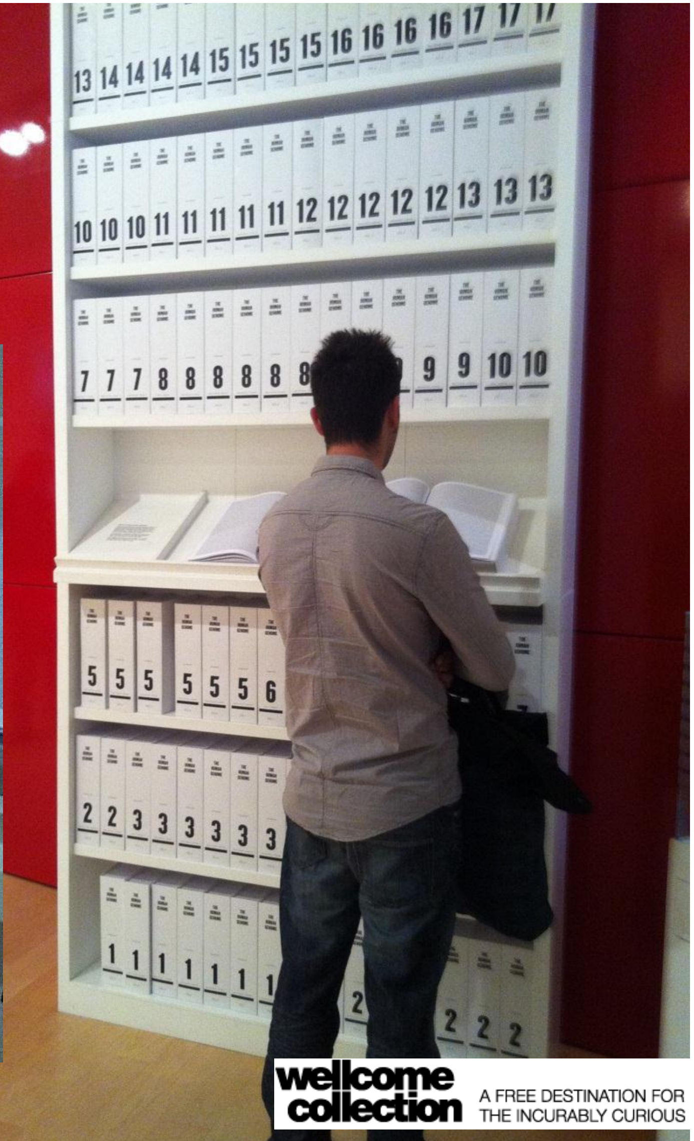
…well, mostly yes, but no…

How big is the human genome?

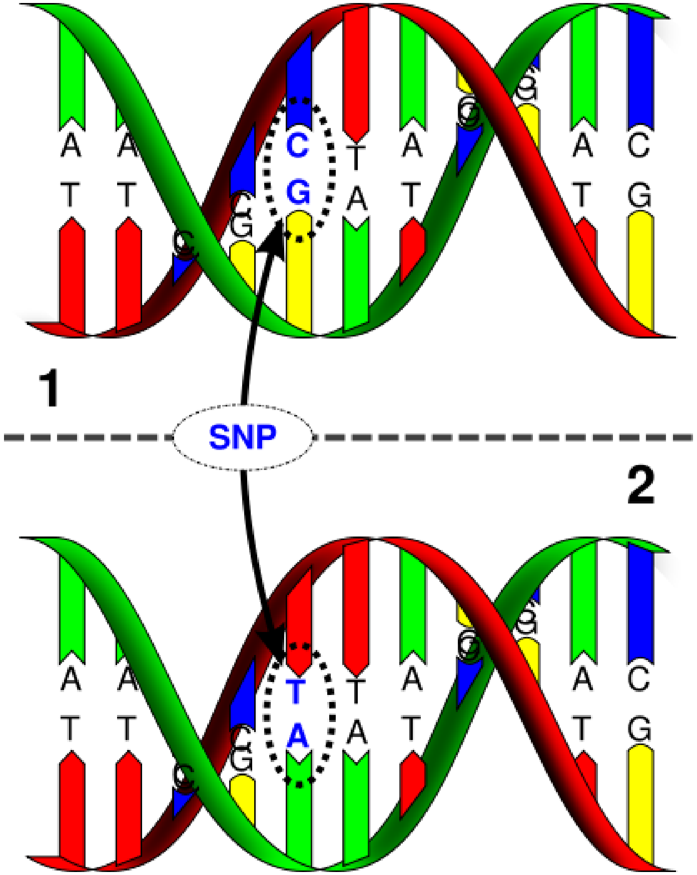




3,600,000,000 (x2) letters



Is the DNA sequence identical among all genomes?

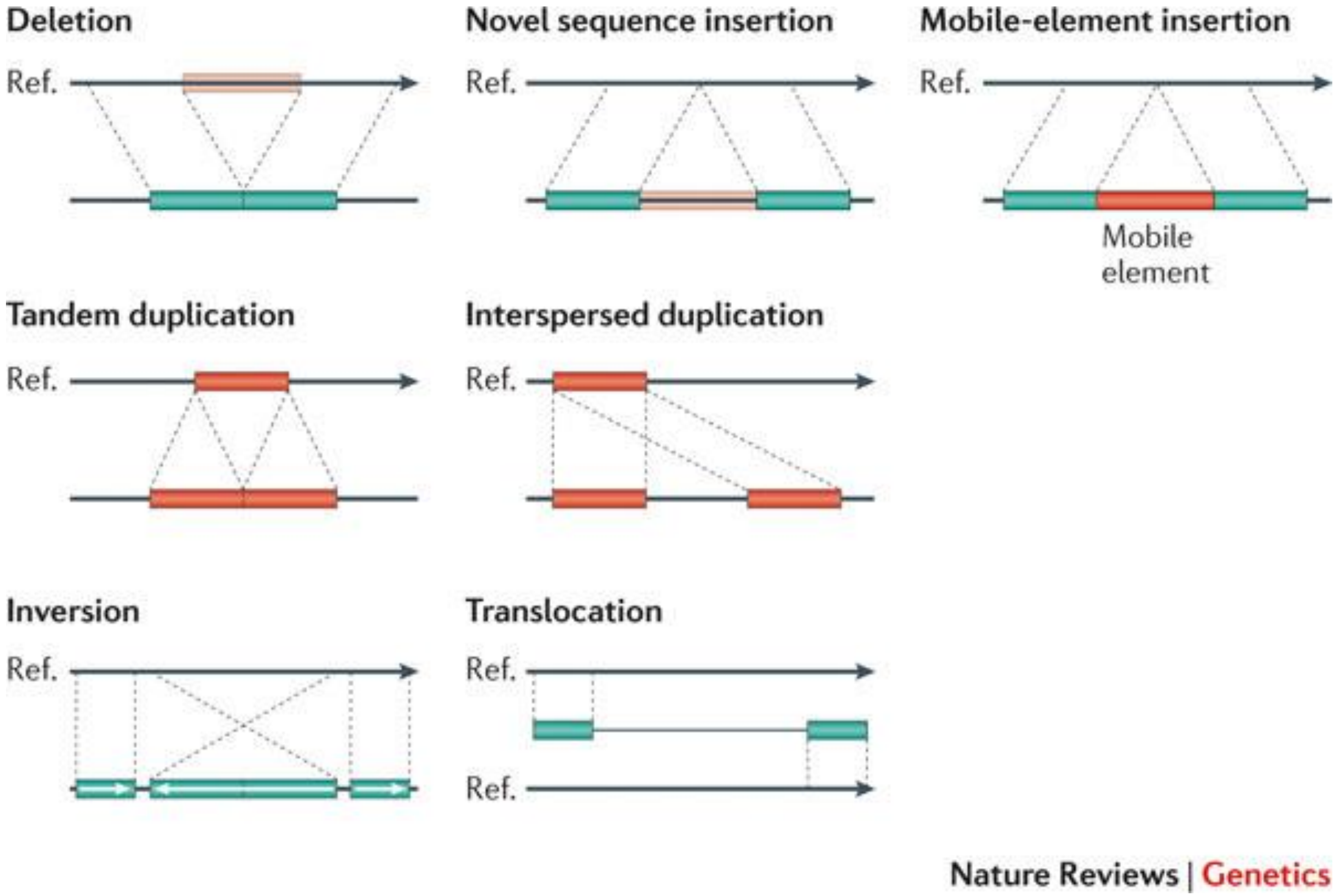


SNPs: Single letter changes

Indels: Small insertions and deletions

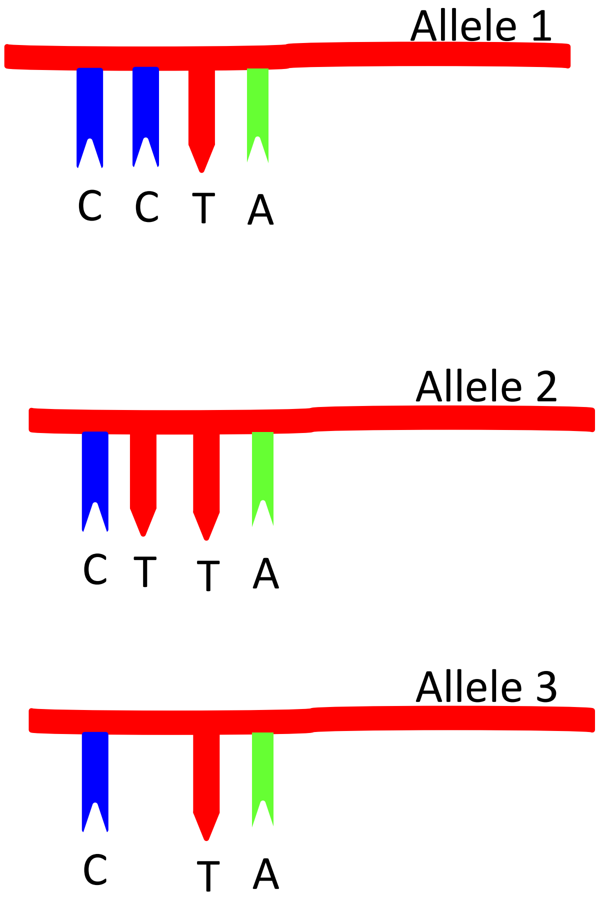
Structural variations: Large changes in the structure and copy number of chromosomes or part of them

Structural variants



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

Alleles or Variants



- Arise due to mutation

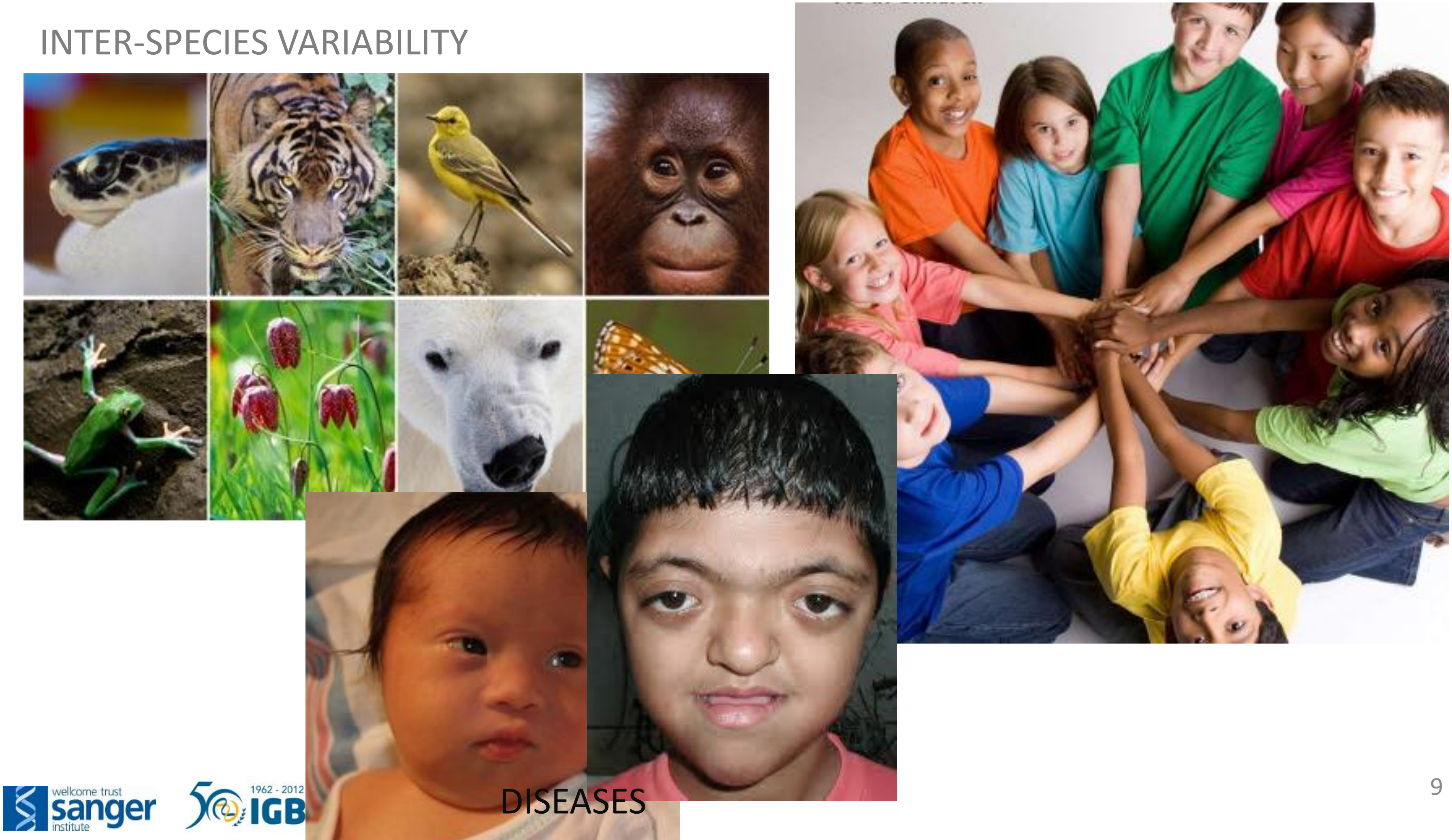
- Shuffled by recombination

- Diffused by migration

Which are the consequence of DNA differences?

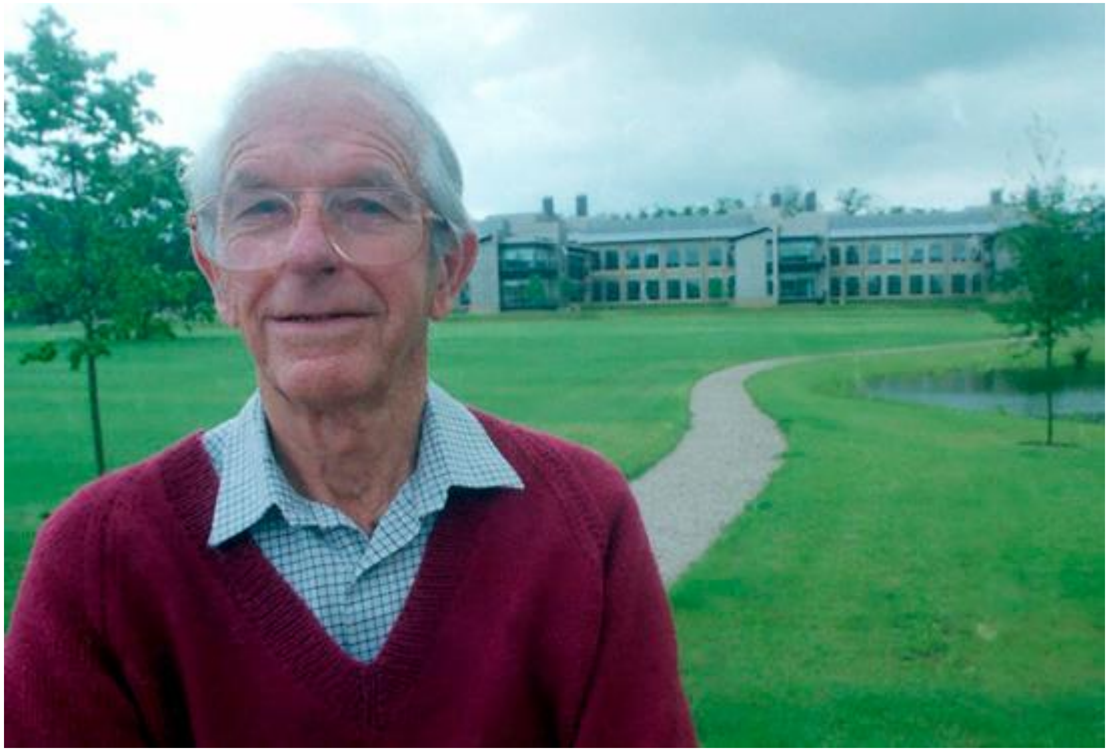
INTRA-SPECIES VARIABILITY

INTER-SPECIES VARIABILITY

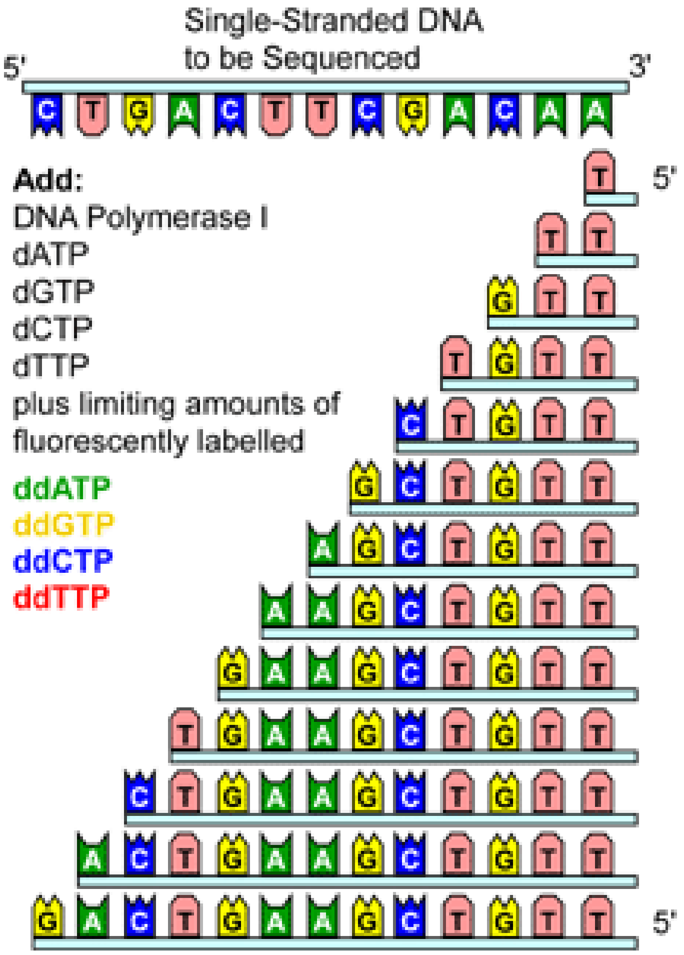


DISEASES

- Understanding the genomic variability in five minutes
- Few details on whole genome sequencing
- Variant detection – variant annotation
- Practical session

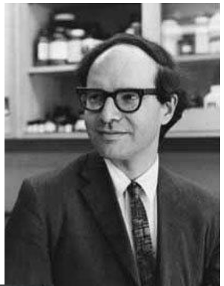


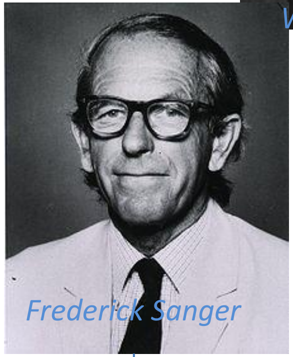
Chain Termination reaction

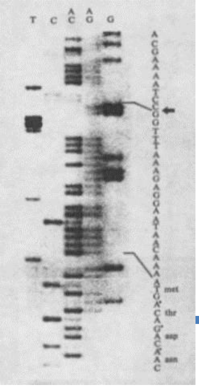


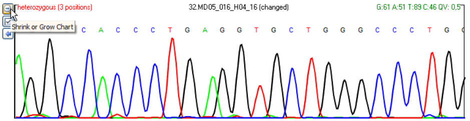
Sequencing technology evolution

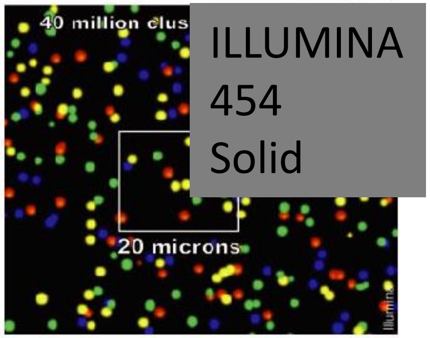
James Watson and Francis Crick


Walter Gilbert


Frederick Sanger




heterozygous (3 positions)
32.MD05_016_H04_16 (changed)
G:61 A:51 T:89 C:46 QV: 0,5
Shrink or Grow Chart
C A C C C T G A G G T G C T G G G C C C T G



ILLUMINA
454
Solid

1953

1977
Maxam-Gilbert
Sanger

1987
Chain termination

2007.....
High-throughput

“State of the art” technologies

Ion Torrent™ next-gen sequencing technology:

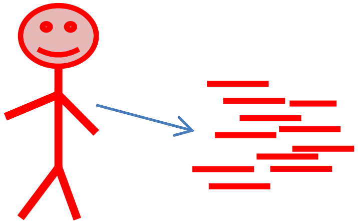
http://www.youtube.com/watch?v=MxkYa9XCvBQ

Pac Bioscience

http://www.youtube.com/watch?v=v8p4ph2MAvI

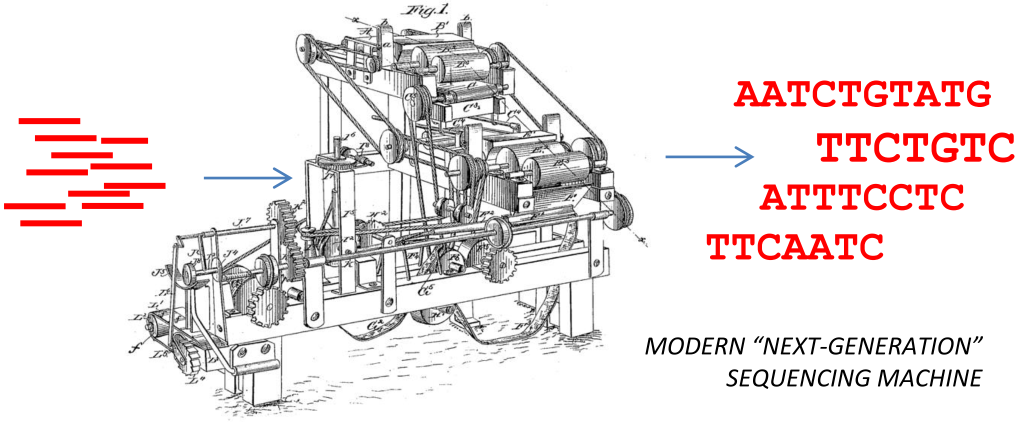
How do we ‘read’ whole genomes?

1. DNA is extracted from donors and fragmented



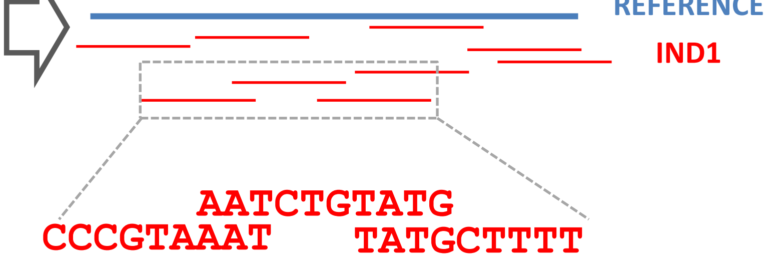
Many copies of the genome in fragments

2. DNA sequence is determined for each fragment

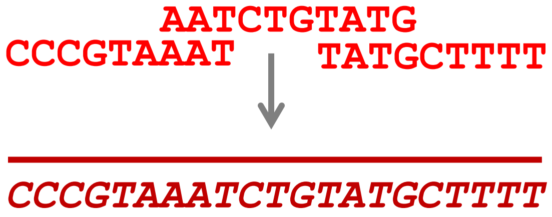


MODERN “NEXT-GENERATION” SEQUENCING MACHINE

3. Fragments are aligned against a reference sequence



4. Overlapping fragments are merged into a ‘consensus’ sequence



What do we sequence and for what?

✓ Variation → DNA-Seq <3

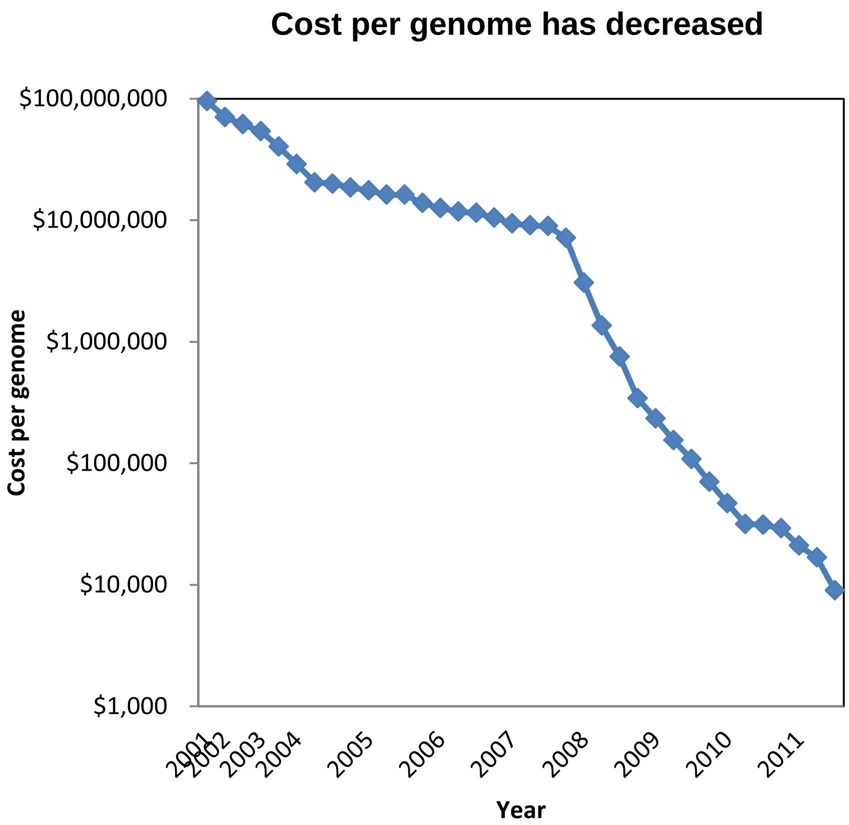
✓ Expression → RNA-seq

✓ Regulation → ChIP-seq

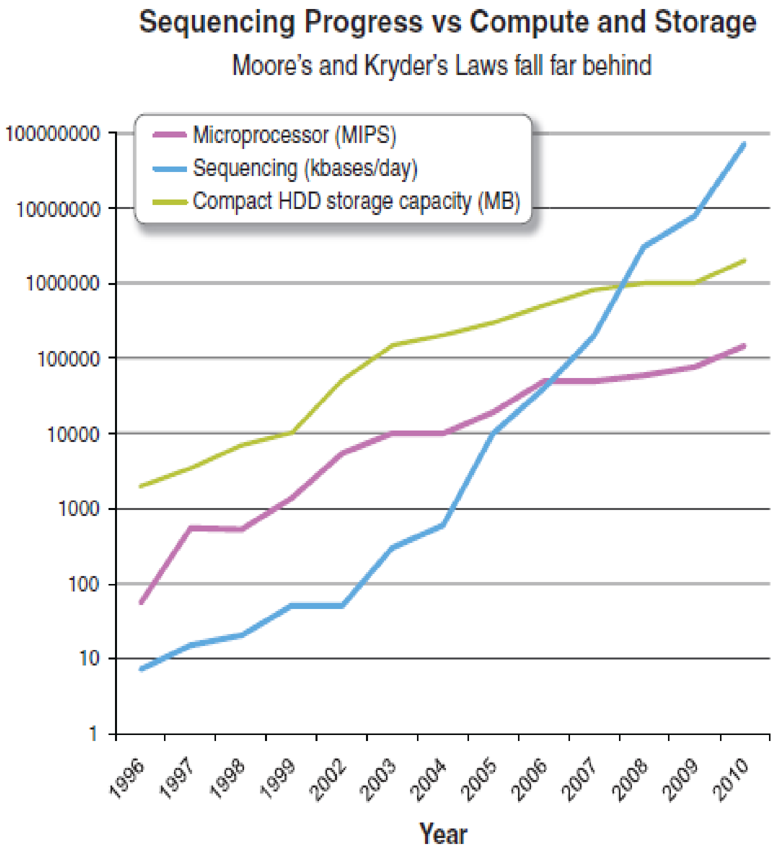
✓ Metagenomics → pooled DNA seq

✓ Non-model organisms…

Why DNA-seq is so exciting?



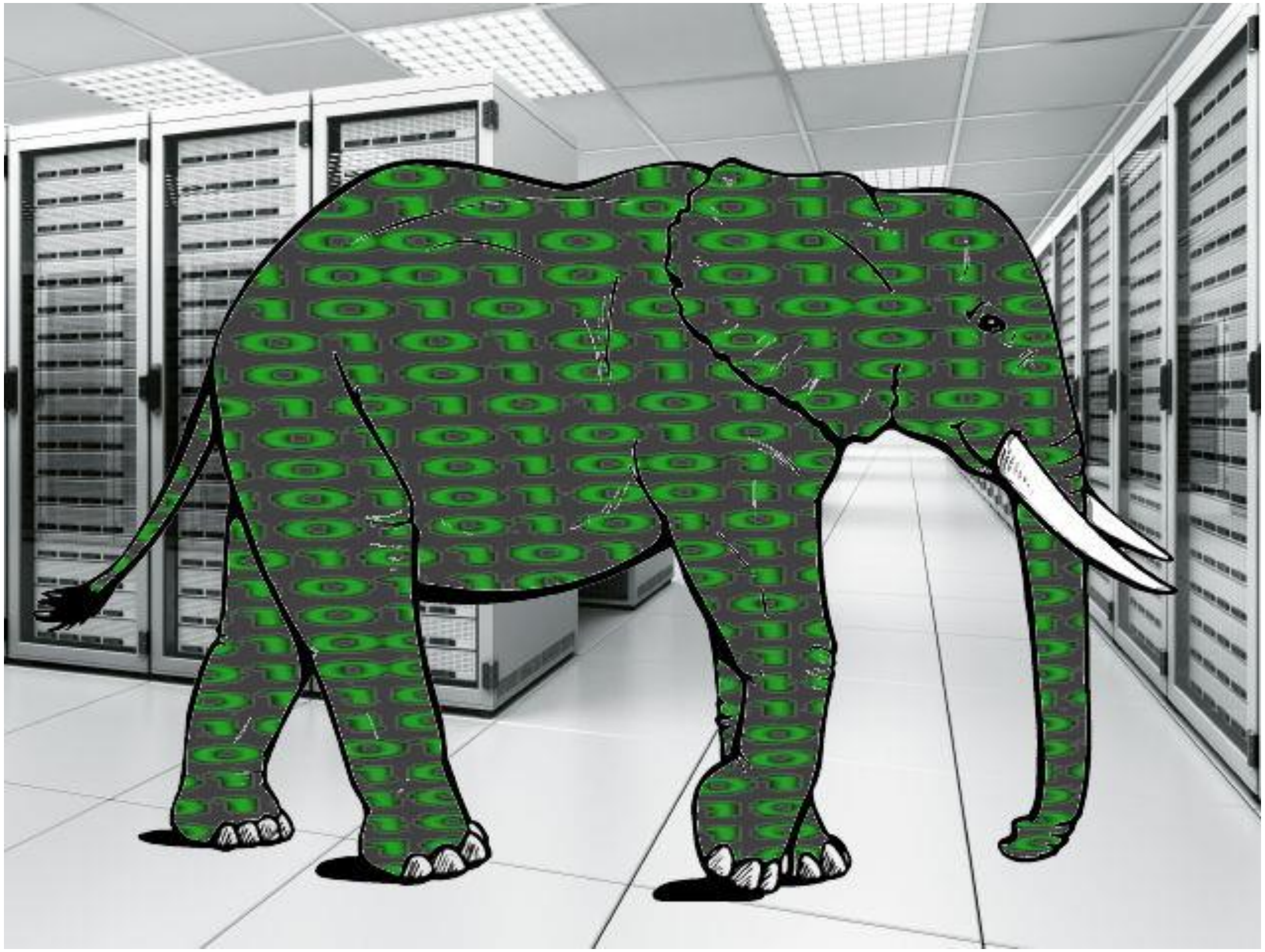
Adapted from NHGRI



Kahn (2011) *Science* **331**, 728-729

wellcome trust sanger institute

1962 - 2012 IGB Consiglio Nazionale delle Ricerche Institute of Genetics and Biophysics Adriano Buzzati-Traverso



Why DNA-seq is hard?

- Human genome is large:

  ~3 x 10^9 nucleotides per haploid genome

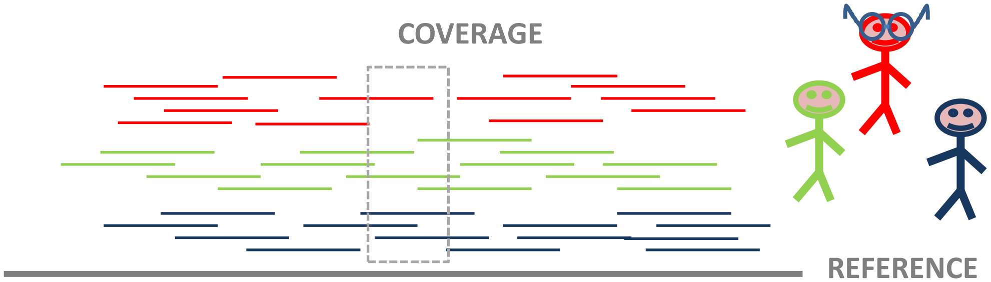
- Sequence reads are short:

  35 – ~1,500 bp, with ~1% errors

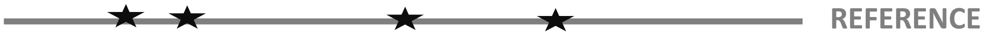
- Understanding the genomic variability in five minutes
- Few details on whole genome sequencing
- Variant detection – variant annotation
- Practical session

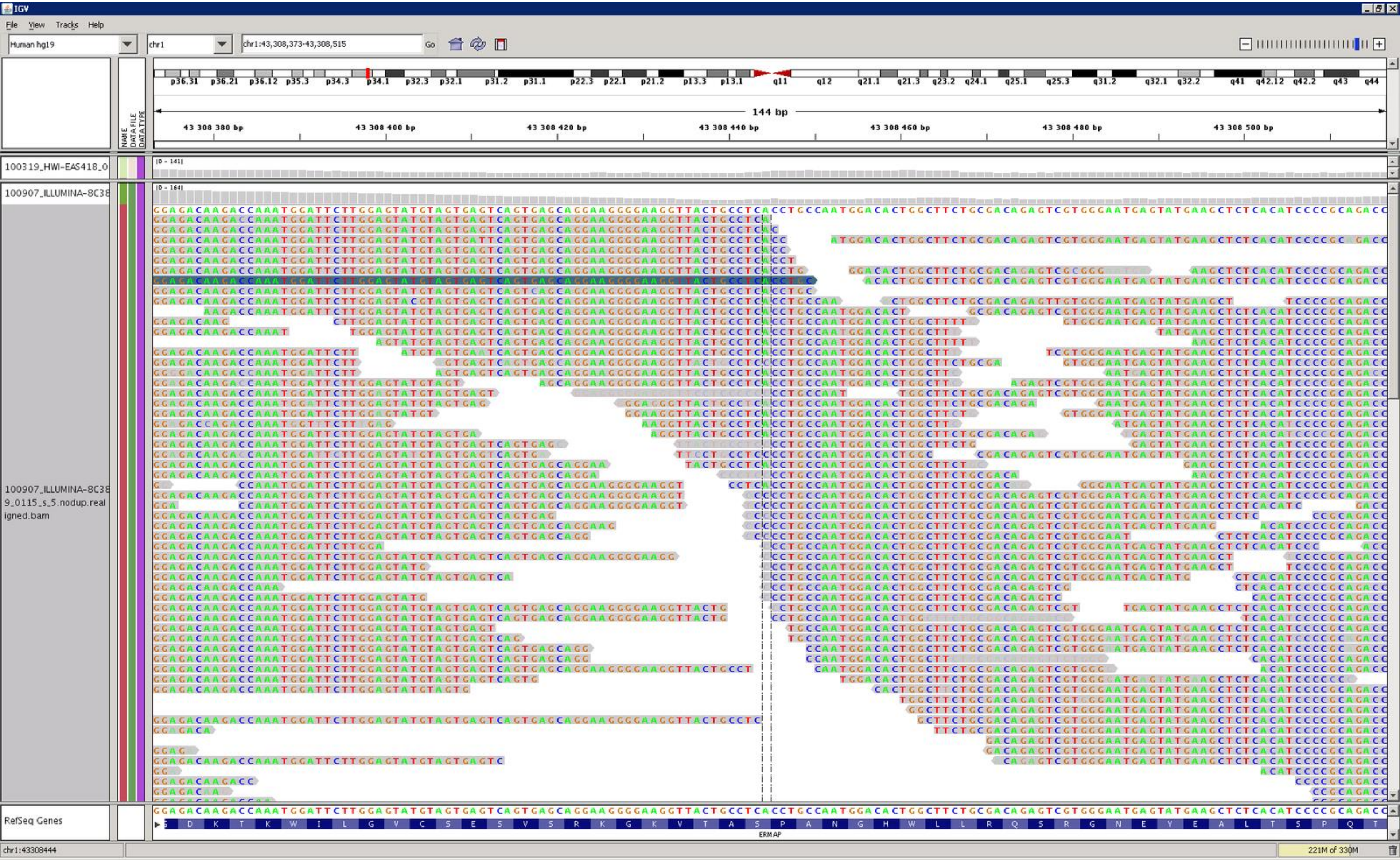
Variant discovery in a nutshell

1. Read mapping



2. Identification of variable sites





Human hg19

chr1

chr1:43,308,373-43,308,515

100319_HWI-EAS418_0

100907_ILLUMINA-8C38

100907_ILLUMINA-8C389_0115_s_5.nodup.realigned.bam

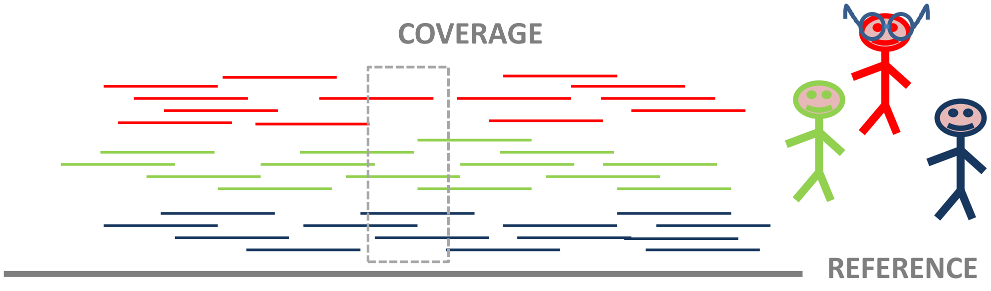
GGAGACAAGACCAAATGGATTCTTGGAGTATGTAGTGAGTCAGTGAGCAGGAAGGGGAAGGTTACTGCCTCACCTGCCAATGGACACTGGCTTCTGCGACAGAGTCGTGGGAATGAGTATGAAGCTCTCACATCCCCGCAGACC

chr1:43308444

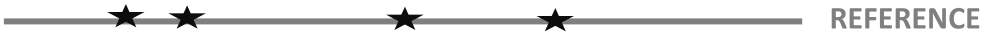
221M of 330M

Variant discovery in a nutshell

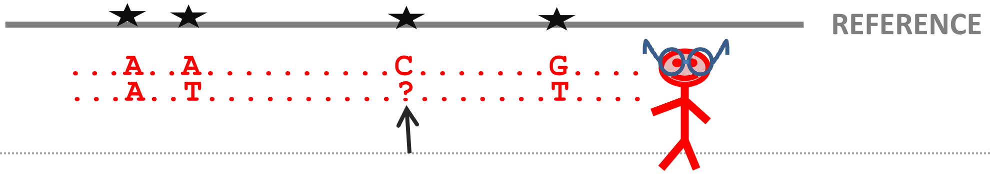
1. Read mapping



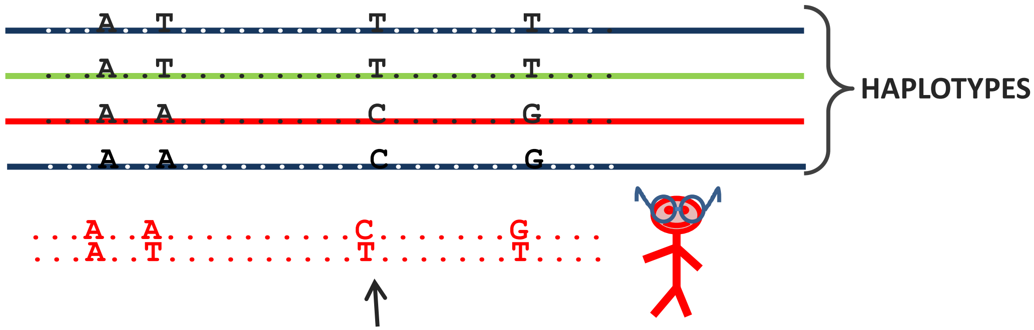
2. Identification of variable sites



3. Individual genotype calling


REFERENCE
...A..A............C........G....
...A..T............?........T....

4. Imputation


...A..T............T........T....
...A..T............T........T....
...A..A............C........G....
...A..A............C........G....
HAPLOTYPES
...A..A............C........G....
...A..T............T........T....

Variant calling algorithms

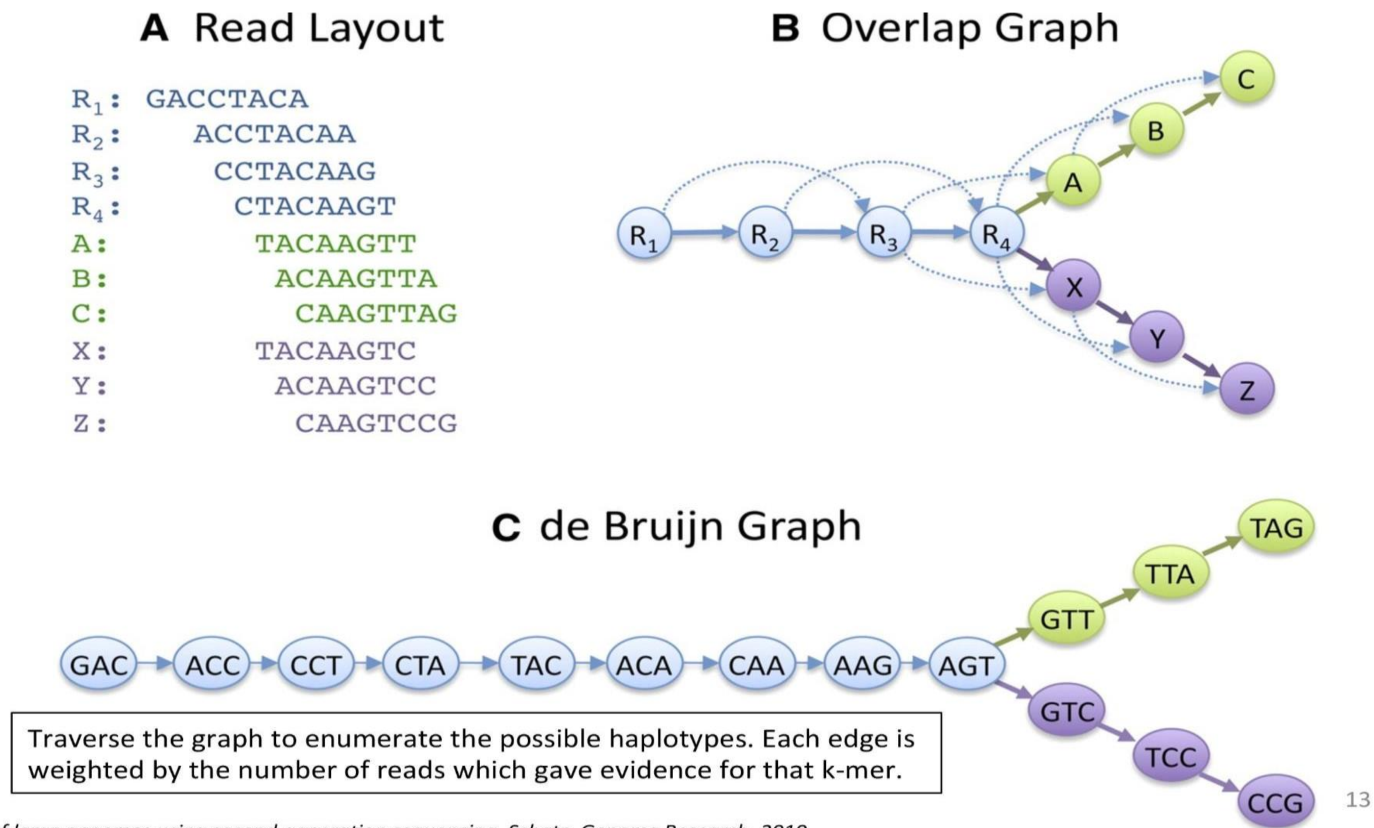
- Allele counting
- Probabilistic methods, e.g. Bayesian model
  - quantify statistical uncertainty
  - assign priors based on observed allele frequency of multiple samples
- Heuristic approach
  - based on thresholds for read depth, base quality, variant allele frequency, statistical significance

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

few examples

1. http://samtools.sourceforge.net/mpileup.shtml
2. https://github.com/ekg/freebayes
3. http://www.broadinstitute.org/gatk/

Discovering alleles using graphs (GATK HaplotypeCaller)



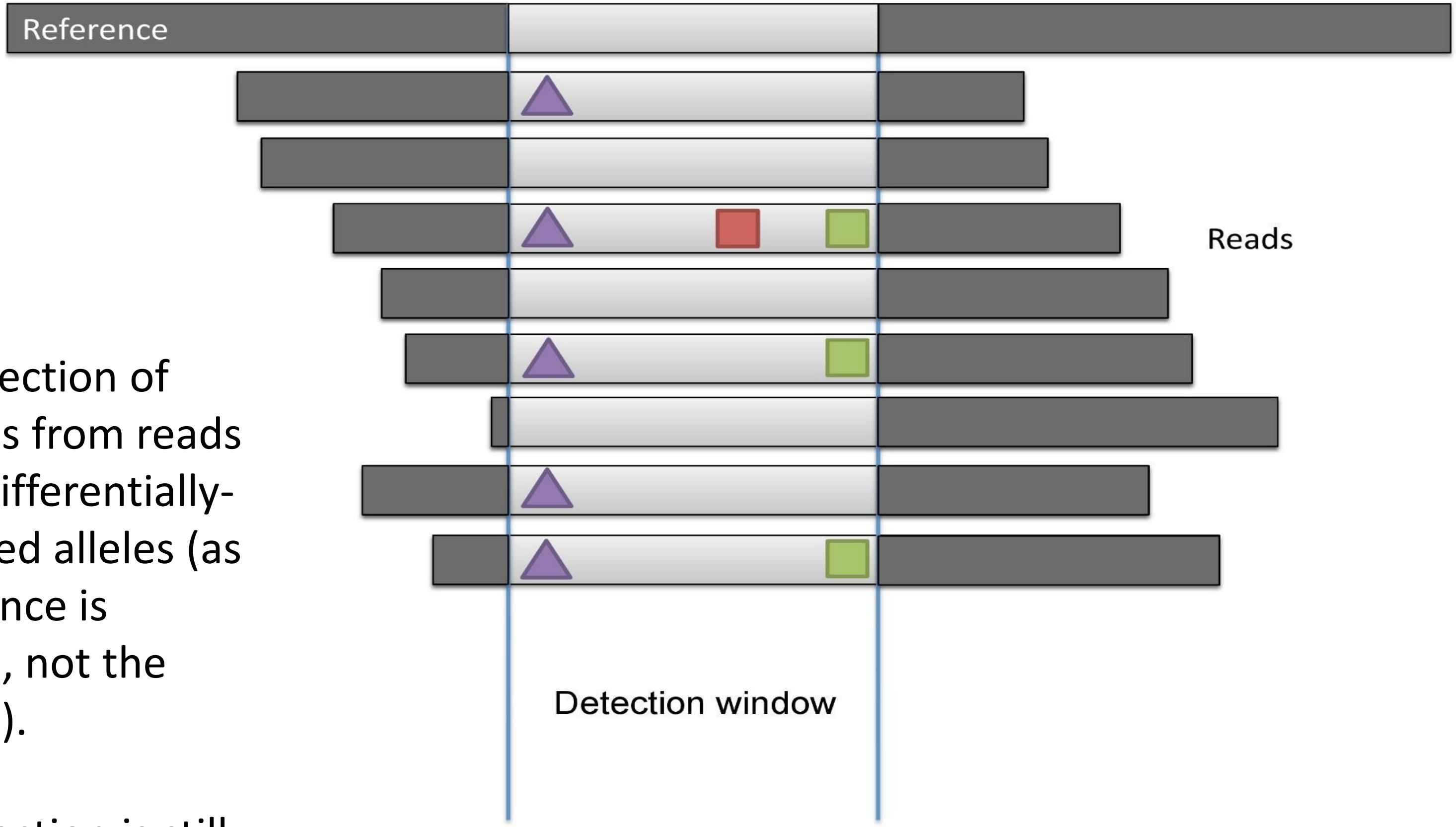
R_1:	GACCTACA
R_2:	ACCTACAA
R_3:	CCTACAAG
R_4:	CTACAAGT
A:	TACAAGTT
B:	ACAAGTTA
C:	CAAGTTAG
X:	TACAAGTC
Y:	ACAAGTCC
Z:	CAAGTCCG

Traverse the graph to enumerate the possible haplotypes. Each edge is weighted by the number of reads which gave evidence for that k-mer.

Assembly of large genomes using second-generation sequencing. Schatz. Genome Research. 2010.

Courtesy of Erik Garrison

Haplotype detection (FreeBayes)

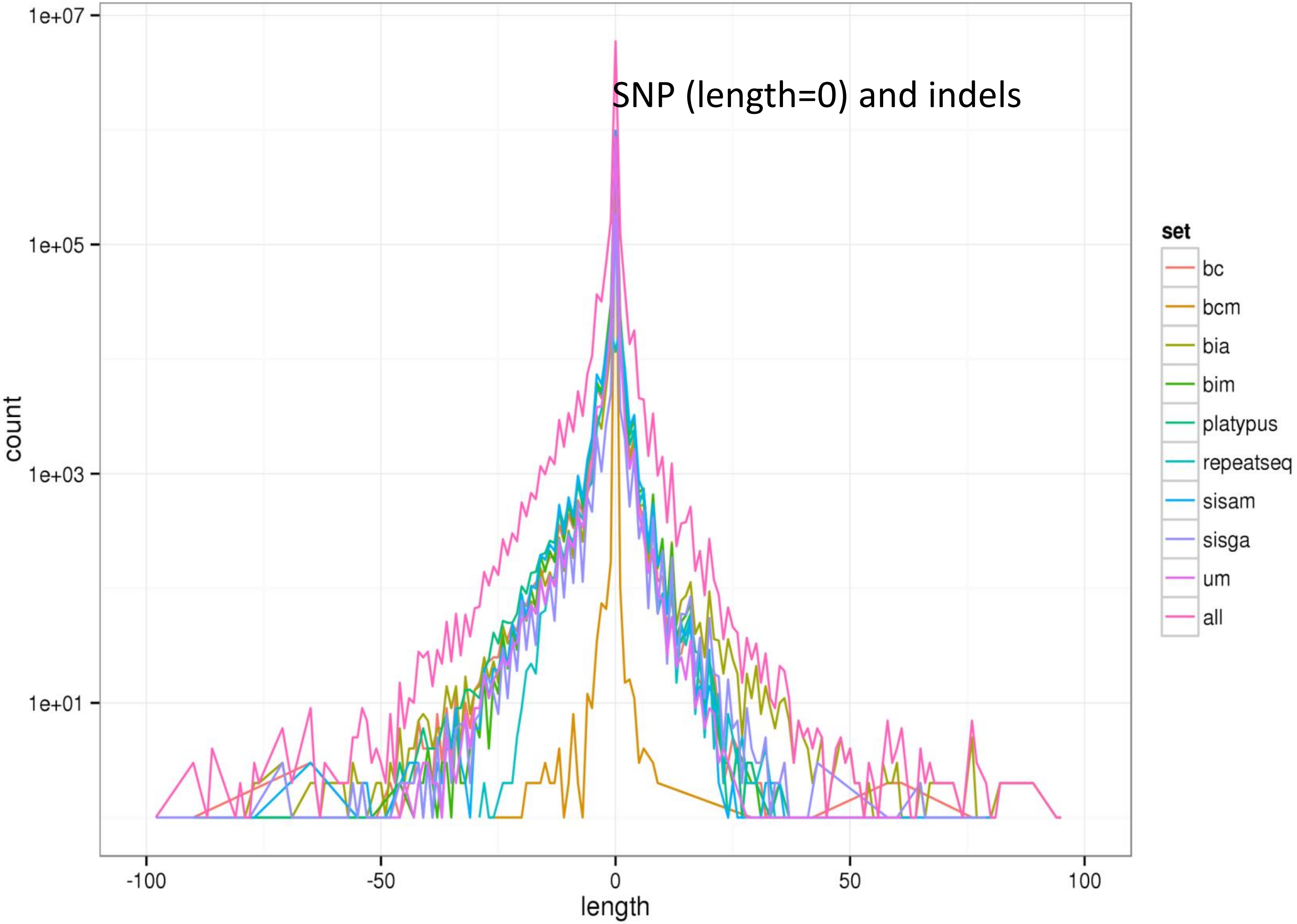


Direct detection of haplotypes from reads resolves differentially-represented alleles (as the sequence is compared, not the alignment).

Allele detection is still alignment-based.

Courtesy of Erik Garrison

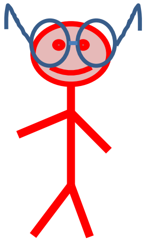
Length-frequency spectrum



SNP (length=0) and indels

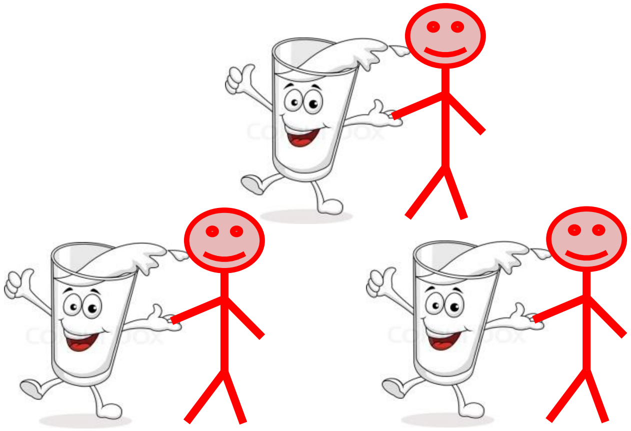
Courtesy of Erik Garrison

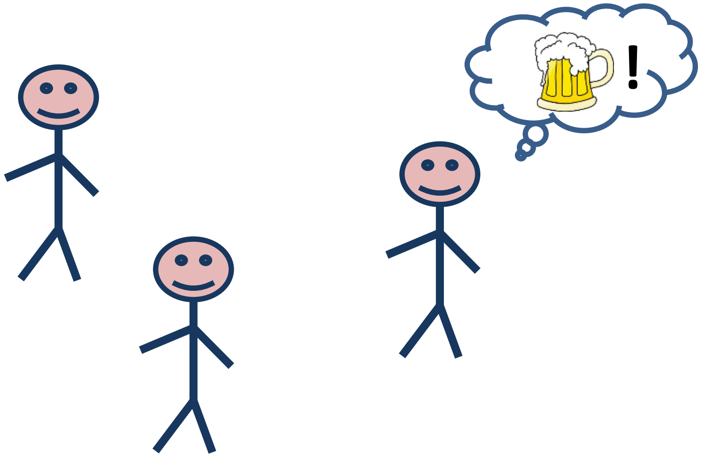
GREAT, WE HAVE A LIST OF VARIABLE SITES, WHAT WE DO NEXT?



Variant annotation!

Is there a functional consequence for a variant?





Chr 2 position 136608646: **T**

↓

Lactose tolerant

Chr 2 position 136608646: **C**

↓

Lactose intolerant

How many variable sites are expected in the human genome?

- Mutation rate is ~1.2 × 10^{-8} bp^{-1} $generation^{-1}$
- In every gamete ~ 30 bases mutate
- In a population of ~7 x 10^{9}, almost every possible genetic variant will be present

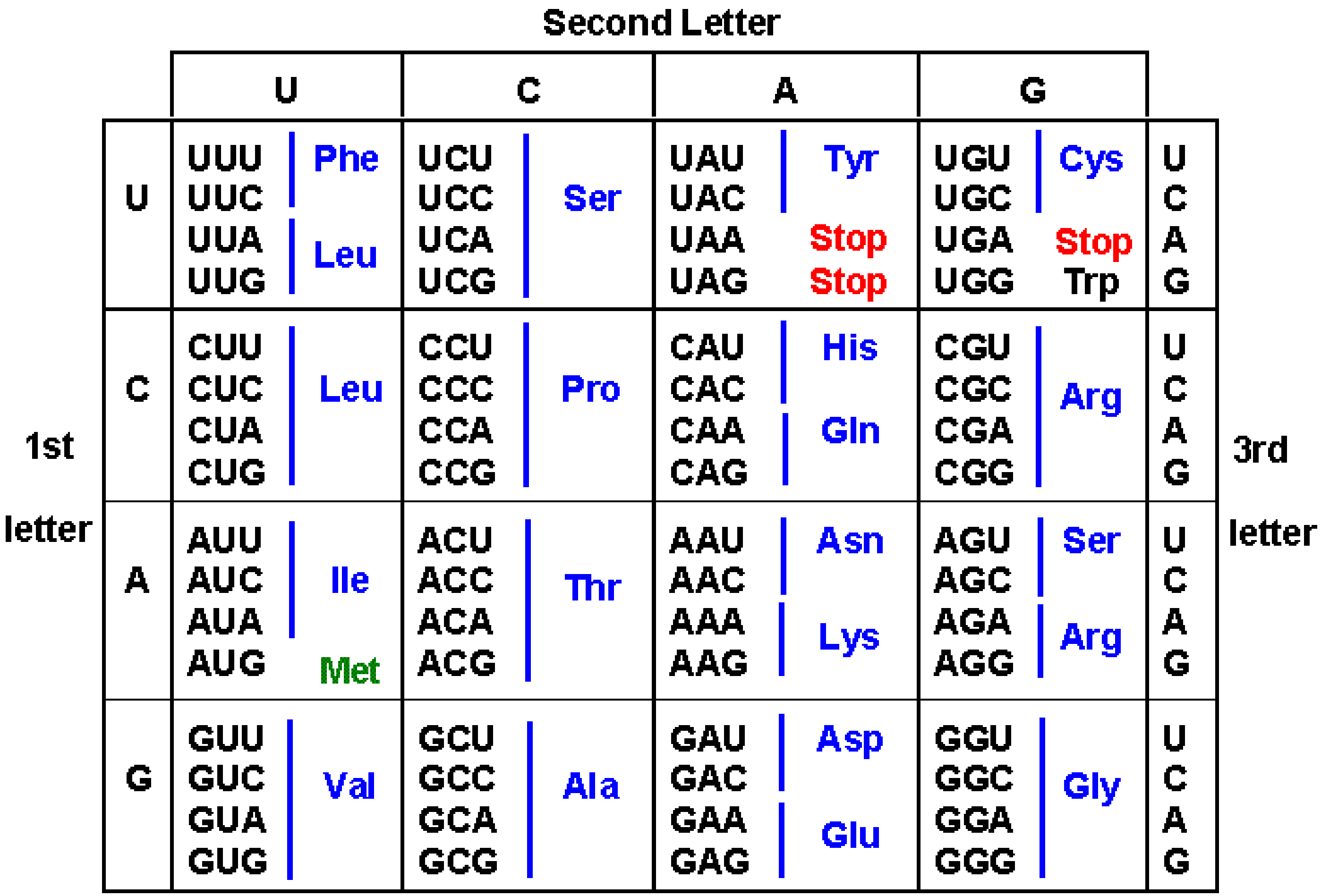
How many variable sites are observed in the human genome?

- Two human genomes typically differ at 3.5 millions of positions

- Thousand humans (two thousands genomes) typically have only 40 millions variable sites

→ Some sites can't change

Consequences in coding (2% of the genome)



	Second Letter								
	U		C		A		G		
1st letter U	UUU UUC UUA UUG	Phe Leu	UCU UCC UCA UCG	Ser	UAU UAC UAA UAG	Tyr Stop Stop	UGU UGC UGA UGG	Cys Stop Trp	U C A G
C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg	U C A G 3rd letter
A	AUU AUC AUA AUG	Ile Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg	U C A G
G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly	U C A G

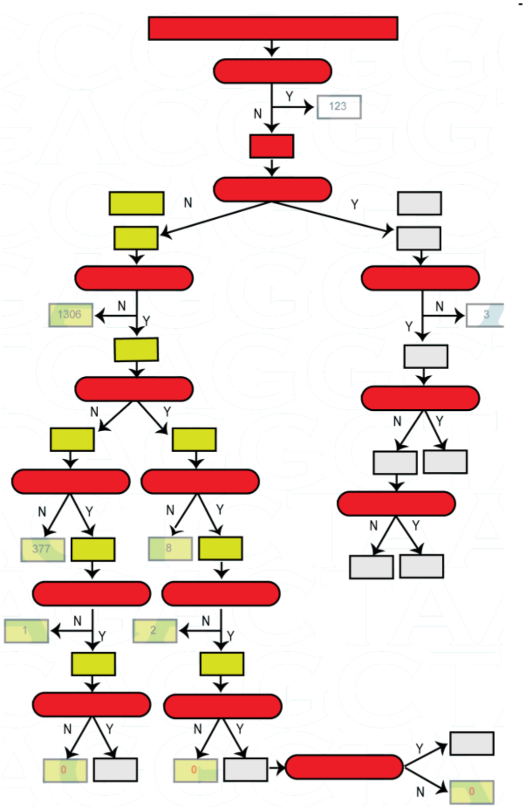
Consequences in non coding (98% of the genome)

- Transcription factor binding sites
- Promoters
- Enhancers
- Chromatin modifications
- ….

Consequences Prioritization

FunSeq

- http://funseq.gersteinlab.org/
- http://info.gersteinlab.org/FunSeq



E. Khurana, Y. Fu, V. Colonna, X. J. Mu, etal.. Integrative annotation of variants from 1092 humans: Application to cancer genomics. Science, 342(6154), 2013.

Few examples of software for annotation

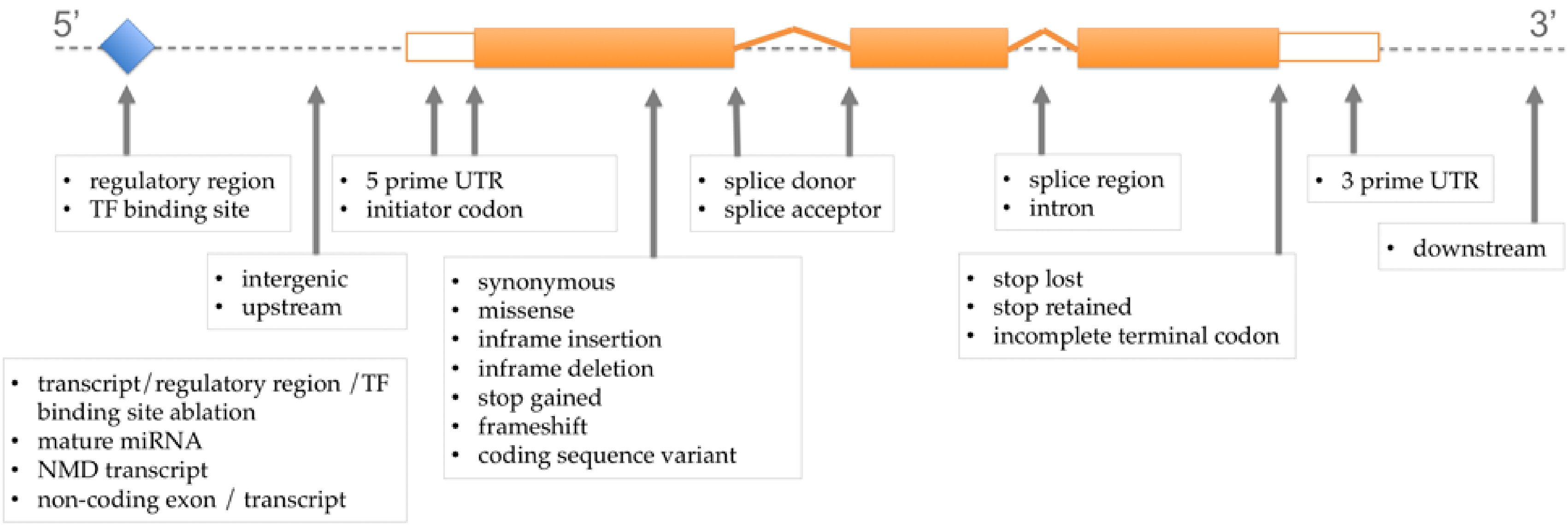
1. http://www.bioconductor.org/packages/2.13/bioc/html/VariantAnnotation.html

2. http://www.openbioinformatics.org/annovar/

3. http://vat.gersteinlab.org/www.openbioinformatics.org/annovar/

4. http://www.ensembl.org/info/docs/tools/vep/index.html <3

Sequence ontology terms



http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences

SIFT http://sift.bii.a-star.edu.sg/

Polyphen http://genetics.bwh.harvard.edu/pph2/

- Understanding the genomic variability in few minutes
- Few details on whole genome sequencing
- Variant detection – variant annotation
- Practical session

Thanks!

- ClaudiaR
- Mario
- Pasquale
- Valerio

WIFIGB

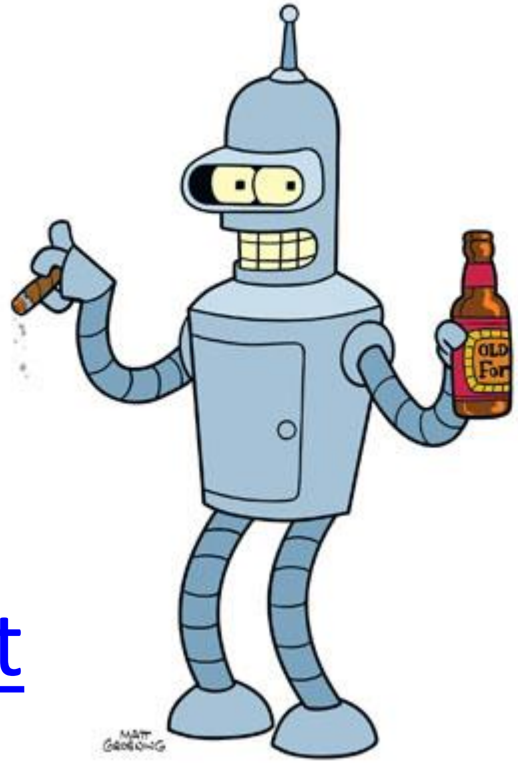
interomics

bioinformatica

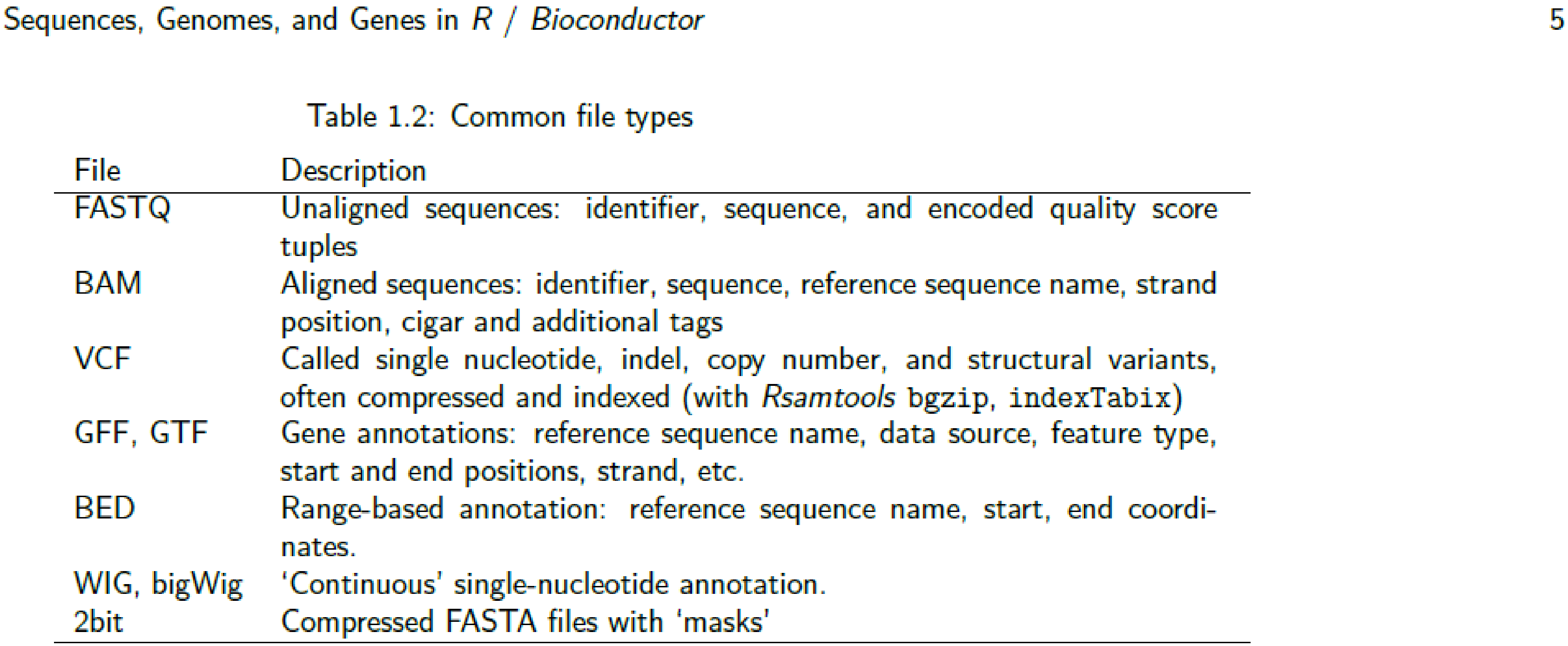
SSH

corso@bender.igb.cnr.it

bioinformatica



Sequence file types



Sequences, Genomes, and Genes in *R* / *Bioconductor* 5

Table 1.2: Common file types

File	Description
FASTQ	Unaligned sequences: identifier, sequence, and encoded quality score tuples
BAM	Aligned sequences: identifier, sequence, reference sequence name, strand position, cigar and additional tags
VCF	Called single nucleotide, indel, copy number, and structural variants, often compressed and indexed (with *Rsamtools* `bgzip`, `indexTabix`)
GFF, GTF	Gene annotations: reference sequence name, data source, feature type, start and end positions, strand, etc.
BED	Range-based annotation: reference sequence name, start, end coordinates.
WIG, bigWig	'Continuous' single-nucleotide annotation.
2bit	Compressed FASTA files with 'masks'

bgzip

- BAM files are compressed using a variant of GZIP (GNU ZIP), called BGZF (Blocked GNU Zip Format)

- BGZF is intended to improve on GZIP for random access.

tabix

- Generic indexer for TAB-delimited genome position files

- fast retrieval of sequence features from a big tab-delimited file

VCF, VCFtools, vcflib

- Poster
- http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40)
- Vcftools http://vcftools.sourceforge.net/
- Vcflib https://github.com/ekg/vcflib

Variant effect predictor

- http://www.ensembl.org/info/docs/tools/vep/index.html
- ENCODE

YES, BUT IN PRACTICE ?
ZZZ….

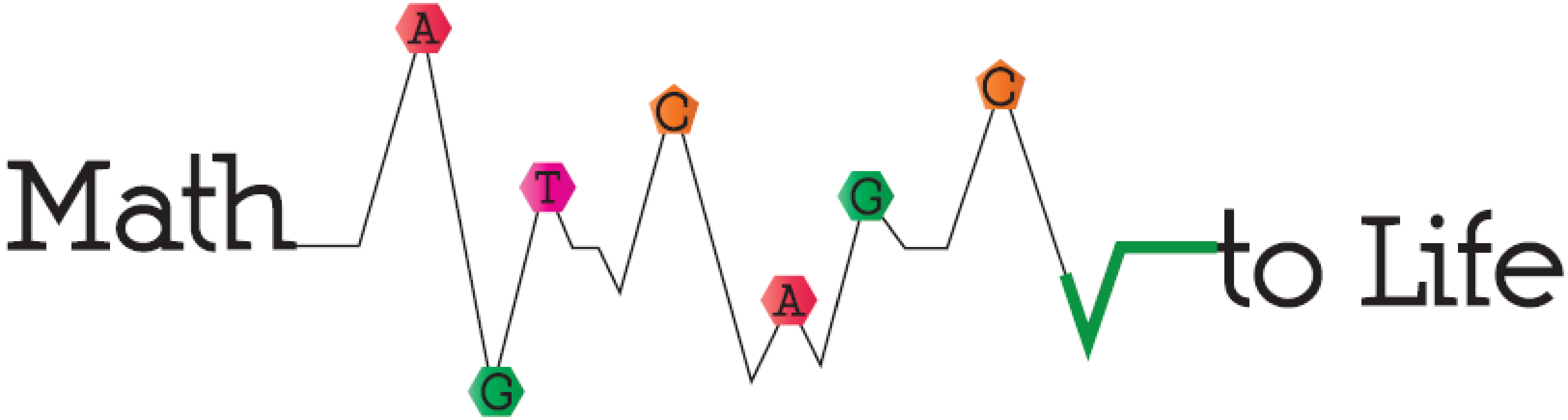
- I have sequenced a number of individuals and I want to know allele frequencies in a subset of them

- I want to download 1000Genomes vcf file to use as comparison with the samples of Asians that I am studying

- I have discovered some variants in my samples of patients and I would like to know if there are functional consequences related to them

- I have discovered a variant in my samples and I would like to know if Neanderthal had it

WORKSHOP ANNOUNCEMENT



Napoli May 2014