How to find and interpret genomic variants in Next Generation Sequencing data

Sophia Derdak

Barcelona, May 3rd 2016

Next generation sequencing (previously: Second generation sequencing)

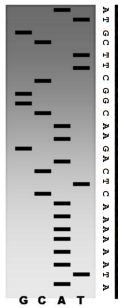
“next”?

“after” Sanger...

Genome sequencing milestones

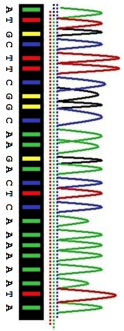
1977

Sanger Sequencing



1987

Sanger Sequencing using fluorescent dyes



2000

Sequencing-by-synthesis

2003

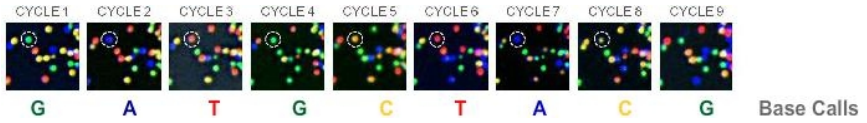
Human Genome Project completed

2006

Solexa Genome Analyzer launched

2009

Illumina Genome Analyzer IIx



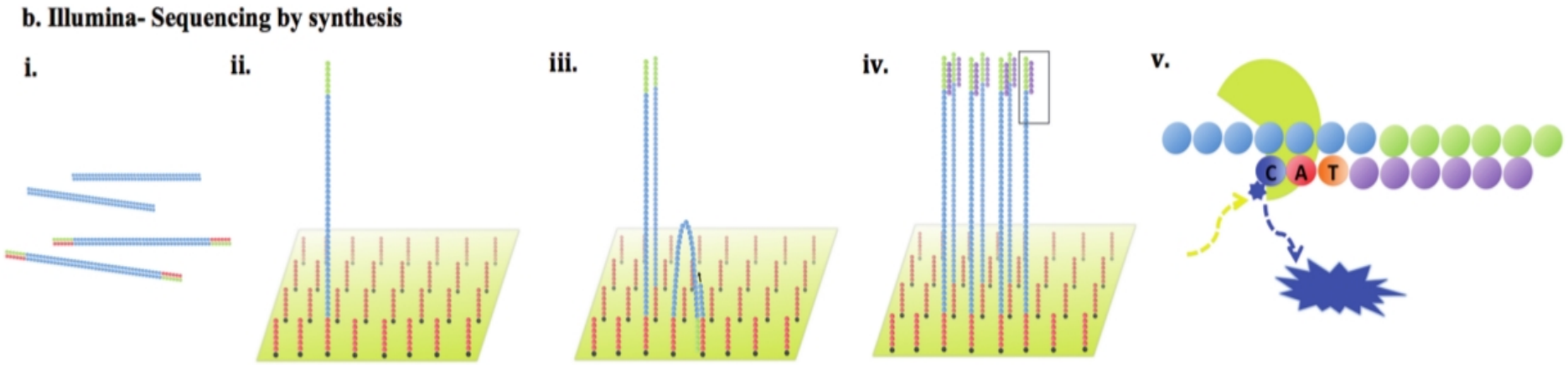
2010

Illumina HiSeq launched

COMING UP:

New machines for
- higher throughput
- clinical applications
- single cell/ single molecule capacity

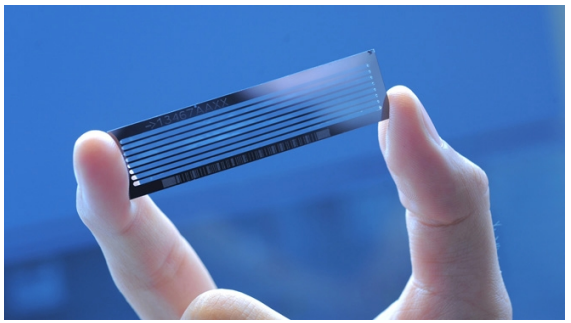
Illumina sequencing



INPUT:
- whole genome in fragments
- optional: selection of coding regions (“exome”)

SCAFFOLD:
- flowcell
- no beads
- no microwells

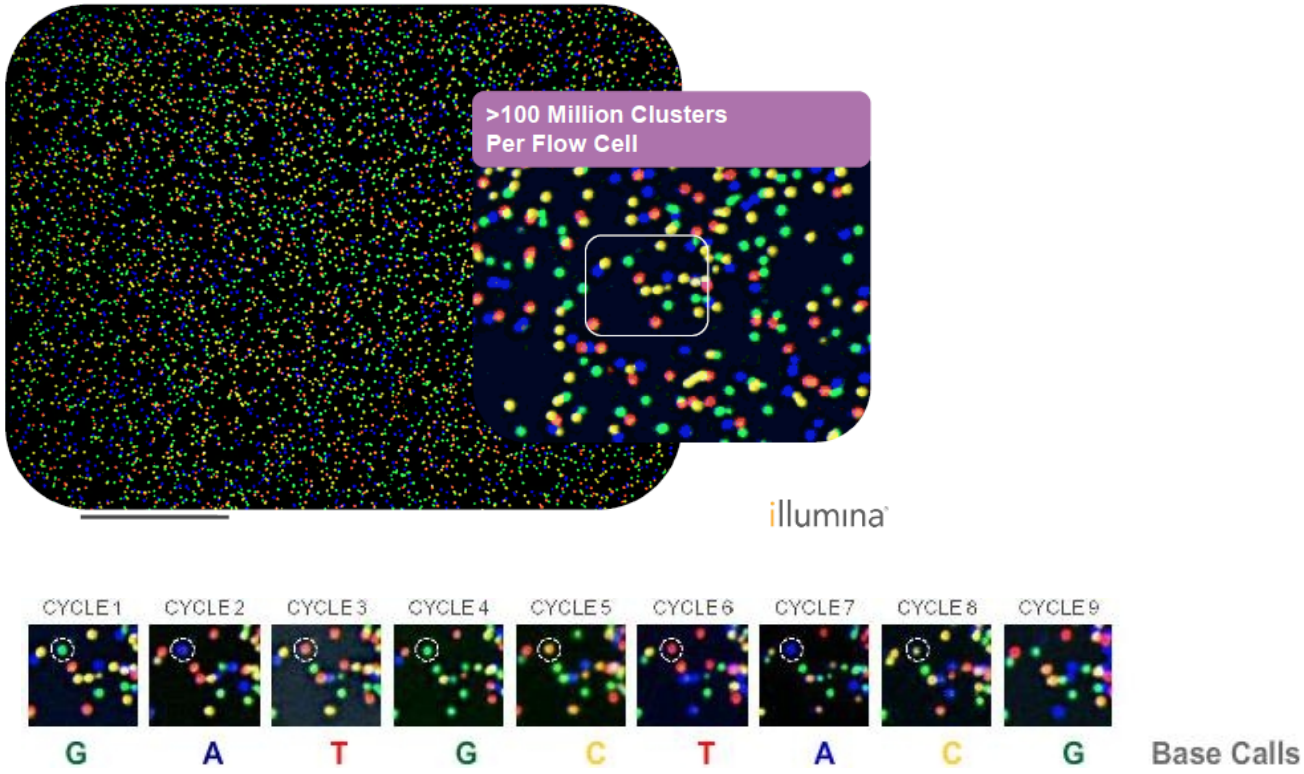
READOUT:
- fluorescent, base-by-base



Nguyen L, Burnett L. Clin Biochem Rev 2014: *Automation of Molecular-Based Analyses: A Primer on Massively Parallel Sequencing.*

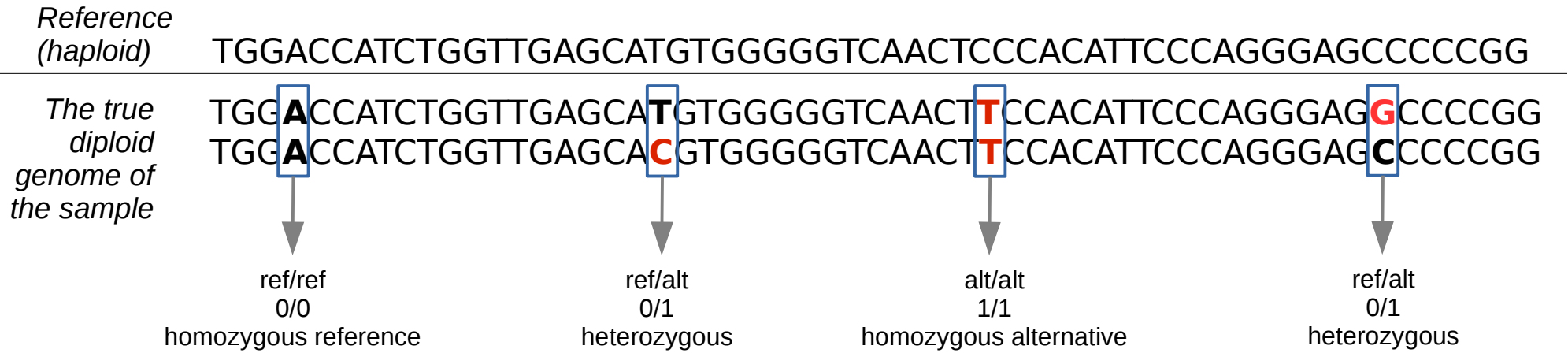
From flowcell to computer: Base calling

- The sequence of colors read for each cluster in each cycle are translated to nucleotide sequence



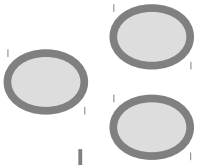
What are genomic variants?

Identification of genetic differences in comparison to a reference



~3.700.000 variant positions / 3.200.000.000 base position genome

>99% of the genomic positions are **not** variant positions



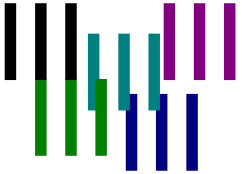
Biological material

- pool of cells
- 2 homologous chromosomes per cell

DNA sample

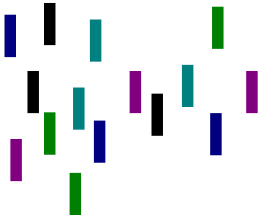
DNA fragments
- overlapping

PCR



amplified DNA fragments

dilution



load on sequencer

raw read data

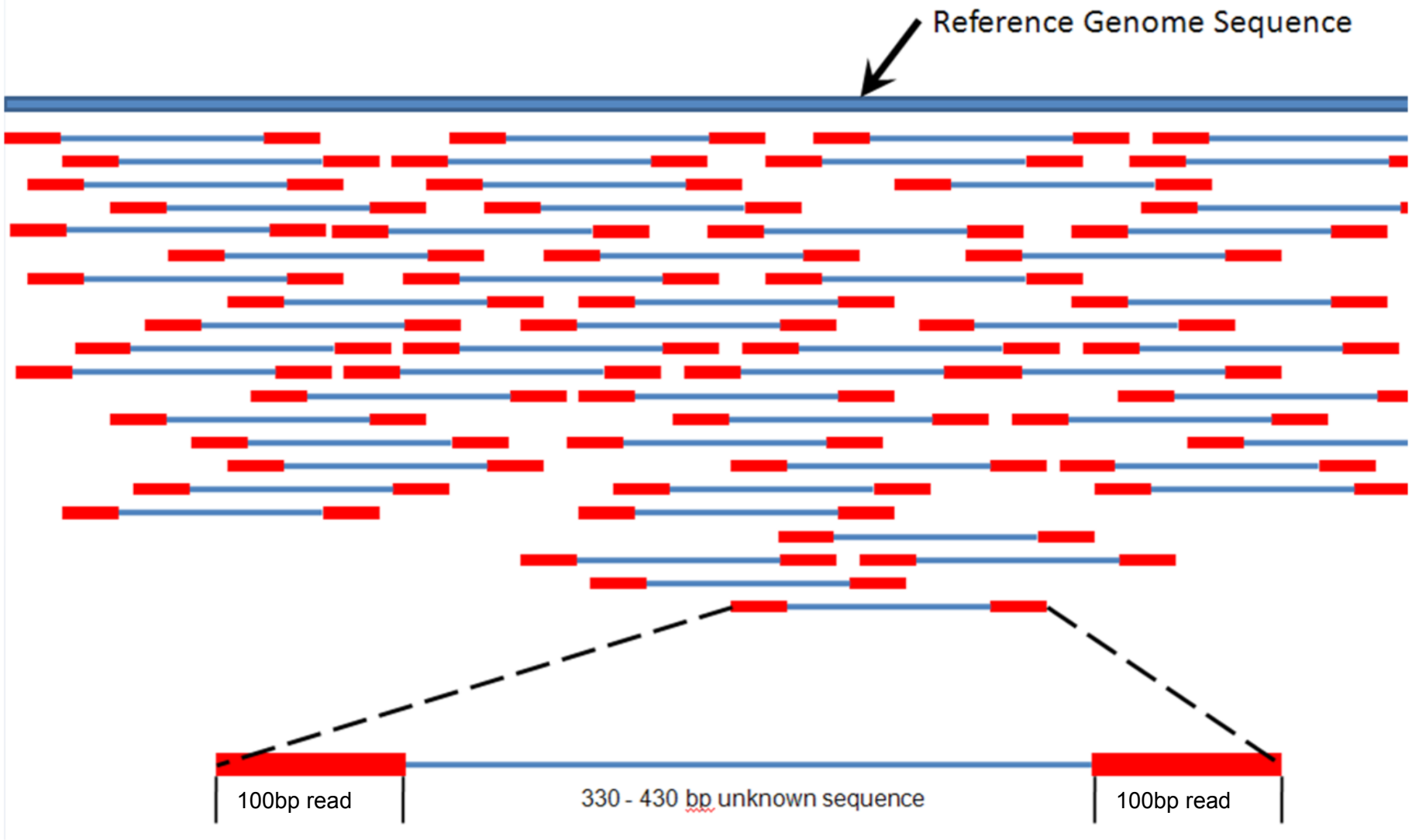
alignments

reference sequence

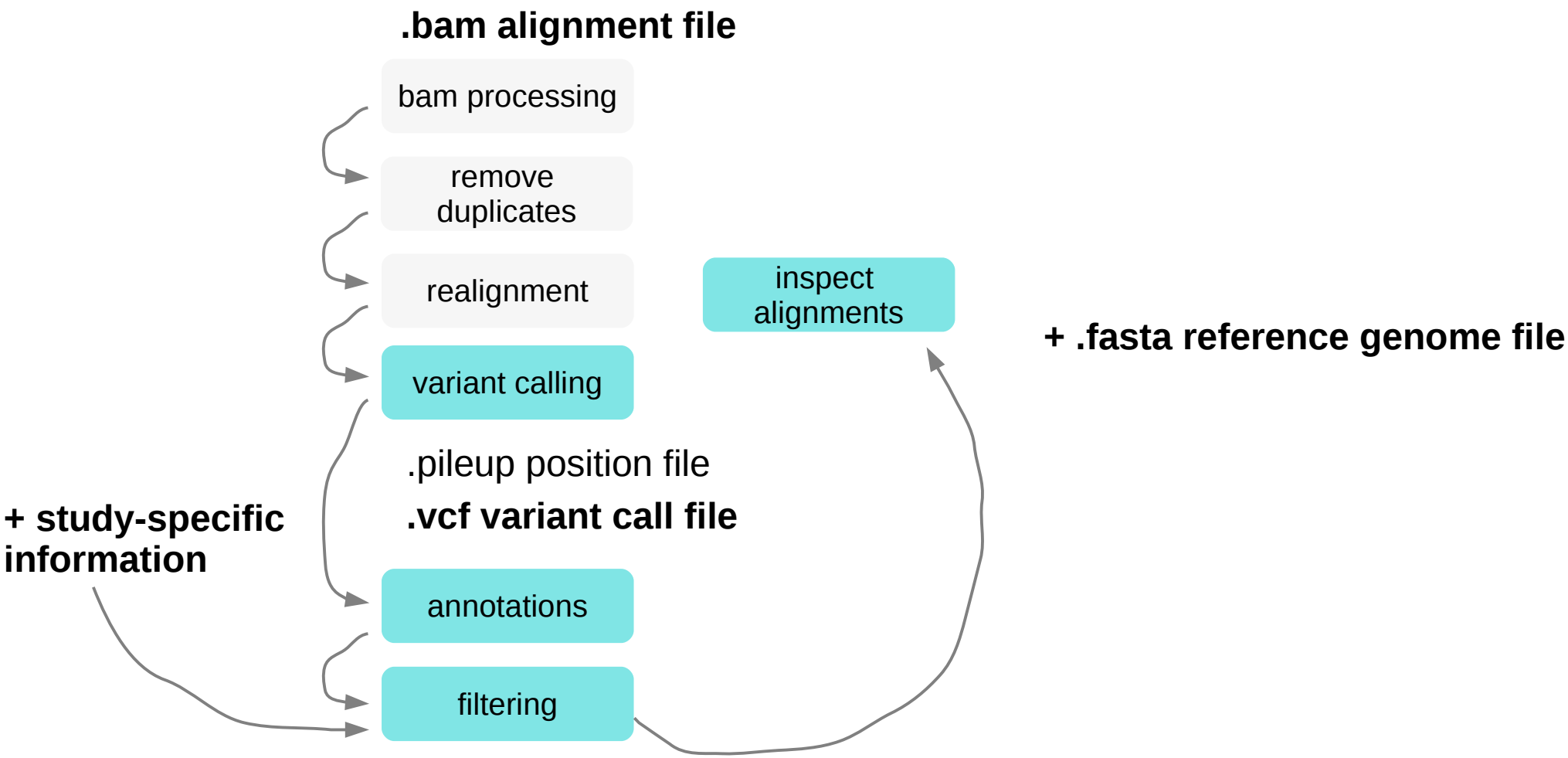
aligned read pairs

variant calling

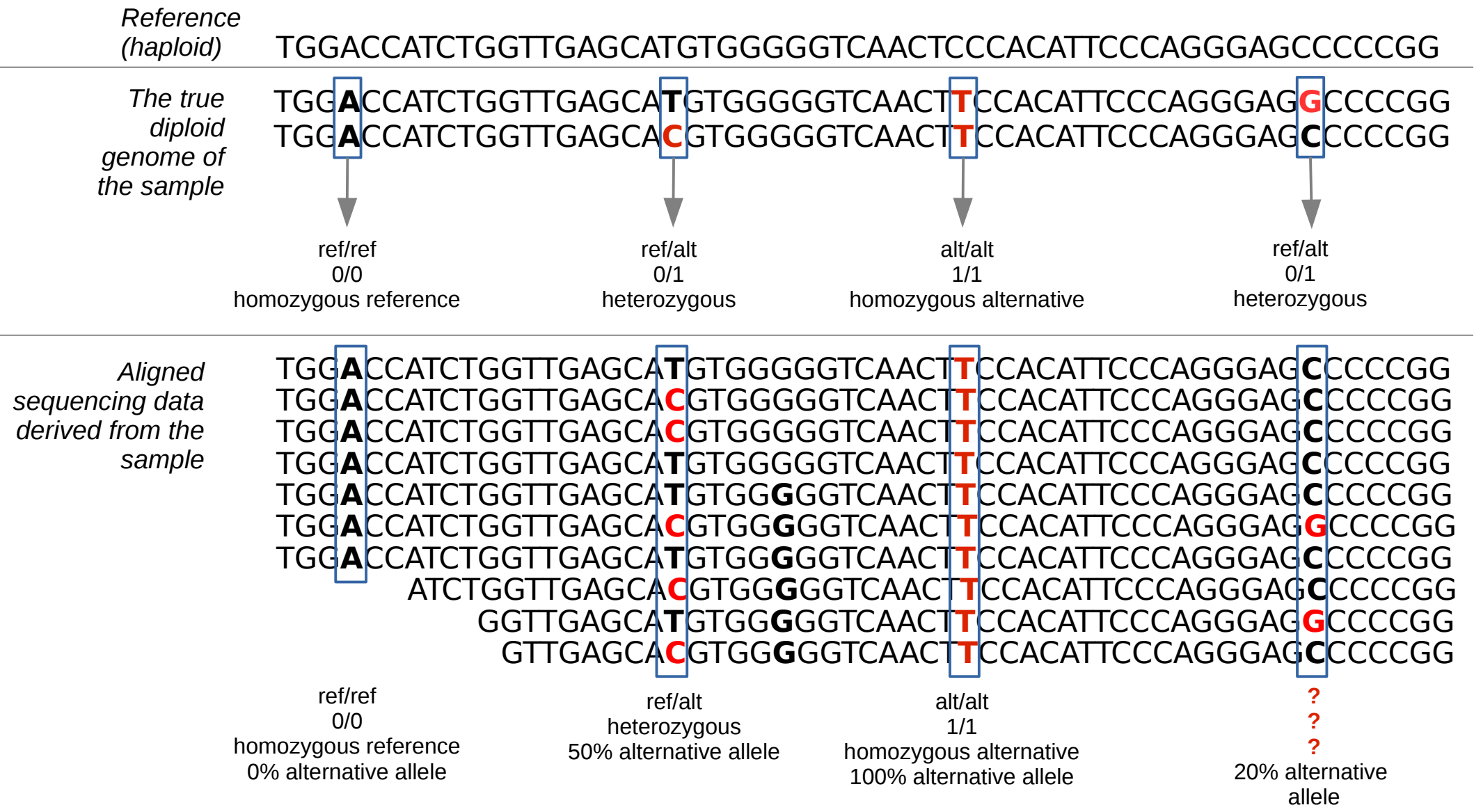
Mapping of reads to the reference sequence



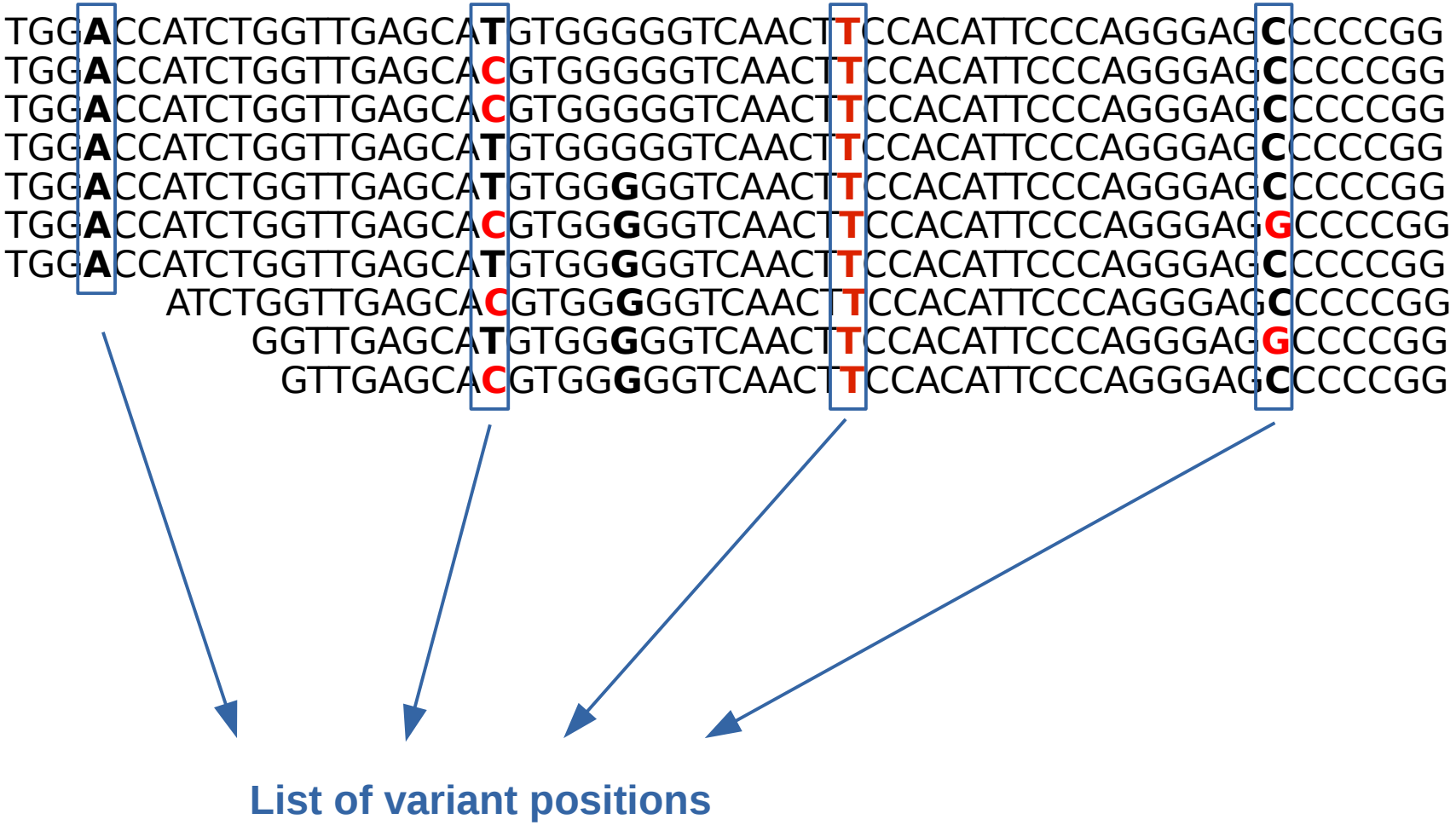
(adapted from wikipedia)


.bam alignment file
bam processing
remove duplicates
realignment
variant calling
inspect alignments
+ .fasta reference genome file
.pileup position file
.vcf variant call file
+ study-specific information
annotations
filtering

Identification of genetic differences in comparison to a reference

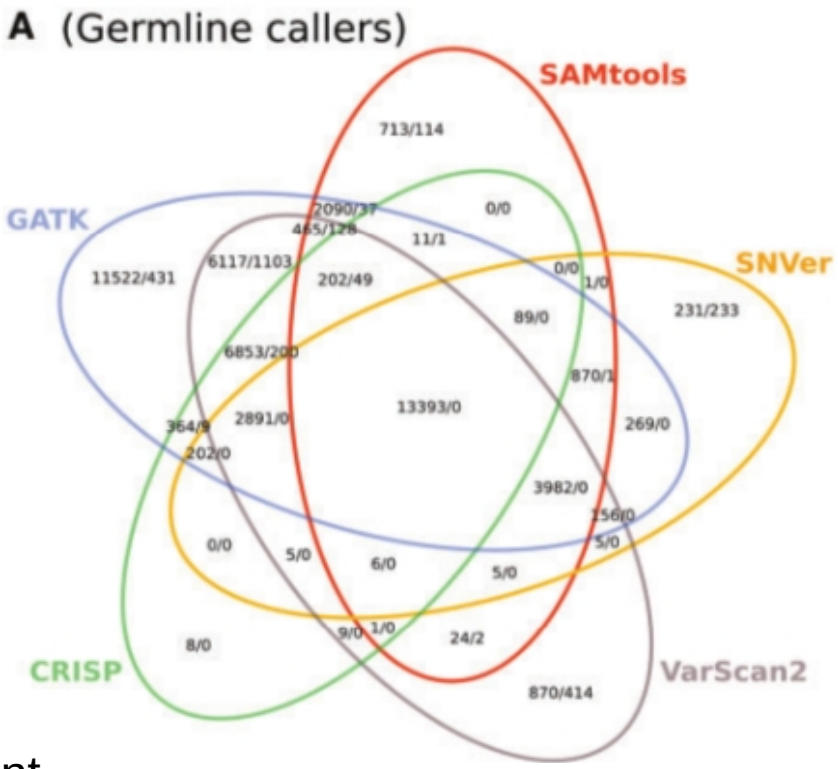


Aligned sequencing data derived from the sample



Altmann A et al. Hum Genet 2012: A beginners guide to SNP calling from high-throughput *DNA-sequencing data.*

- samtools + bcftools (Sanger Institute, UK, and Broad Institute, US)
- Genome Analysis Tool Kit (GATK) (Broad Institute, US)
- VarScan (Washington University)
- Platypus (Welcome Trust Center, UK)
- freebayes (Boston College, US)



Keep in mind that different software use different algorithms and thresholds and results may vary **A LOT**.

Pabinger S et al. Briefings in Bioinformatics 2013: *A survey of tools for variant analysis of next-generation genome sequencing data.*

The coverage

represents the number of times a base of the sample genome (or target region) is read during sequencing.

A higher coverage provides higher power for data analysis.

How to get a higher coverage:

- mainly by loading more sequencing units (indexes, lanes, entire flowcells) with the same library preparation

Typical coverage numbers (in CNAG projects):

- whole genome: 30x
- exome: 50-100x
- custom gene panel capture: >1000x

"I believe that we do not know anything for certain, but everything probably."
Christiaan Huygens

- base calling (base qualities in the fastq files)
- contig order in the reference assembly
- reference sequence (not yet...)
- **read alignment (mapping quality)**
- **variant position (variant and genotype quality)**

Plato, ~400 BC

- p-values
- probability likelihoods
- PHRED scores

raw vcf file (“all variants”)

mostly experiment-independent
technical and quality filtering (well-covered positions with confident alternative allele)

filtered vcf file (“good quality variants”)

CHR	POS	REF	ALT	GT
1	148588972	G	C	0/1
1	154284894	A	G	0/1
1	203923829	A	G	0/1
1	243329075	T	C	0/1
2	102968362	T	C	0/1
2	122096456	G	A	0/1
2	242612151	C	T	1/1
3	56591283	TAAGCAGGGG	TAAGCAGGGGGAAGCAGGGG	0/1
4	146297387	CAAAAAAAA	CAAAAAAAAAA	0/1
6	116263181	T	C	1/1
8	96070181	T	C	1/1
9	129831659	T	A	1/1
9	131454120	C	T	1/1
10	29834095	G	A	1/1
11	18159254	A	G	1/1
11	35274829	A	G	0/1
12	21628320	C	T	0/1
15	42619508	C	T	1/1
15	75336729	A	G	0/1
16	84035844	T	C	0/1

CHR chromosome

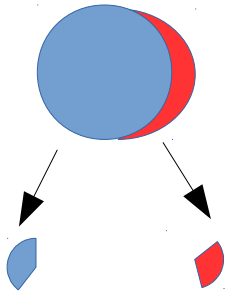
POS position on the chromosome

REF sequence in the reference genome

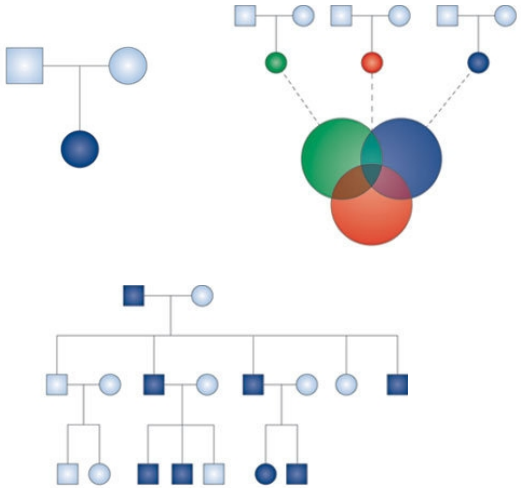
ALT alternative sequence detected in the sample

GT genotype in the (diploid) sample

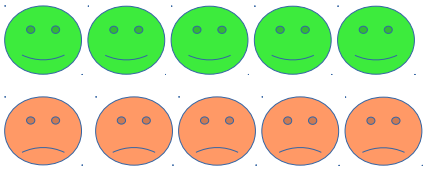
Somatic variants



Inheritance
De novo variants



Affected vs. control group



Compare two samples of the same individual (e.g. tumor-normal)

vcf file (“good quality variants - all genotypes”)

Definition of “somatic variant”, consider sample purity information

Select variants with genotype 0/0 in the normal and 0/1 in the tumor sample

Additionally, select alternative allele frequency thresholds for normal and tumor sample using AC

filtered vcf file (“somatic variants”)

CHR	POS	REF	ALT	GT_normal	AC_normal	GT_tumor	AC_tumor
1	1421916	T	C	0/0	37,1	0/1	44,7
1	179528803	A	C	0/0	53,0	0/1	53,16
1	59096853	C	T	0/0	19,1	0/1	21,7
2	132236963	C	T	0/0	11,0	0/1	21,5
2	166756497	T	A	0/0	20,1	0/1	23,8
3	53910122	G	A	0/0	28,0	0/1	29,7
7	151962062	A	G	0/0	12,0	0/1	19,5
9	136083801	A	G	0/0	10,0	0/1	16,5
11	89407177	C	T	0/0	15,0	0/1	31,6
12	129298780	G	A	0/0	19,1	0/1	30,6
15	20454042	A	T	0/0	22,1	0/1	24,5
15	23113683	T	C	0/0	12,0	0/1	4,6
17	15468718	G	A	0/0	11,0	0/1	16,5
17	15468728	T	C	0/0	11,0	0/1	19,5
17	36365191	A	T	0/0	11,0	0/1	18,4
17	66195635	T	G	0/0	12,0	0/1	12,5
19	43783125	C	A	0/0	10,0	0/1	24,5
19	43783146	A	T	0/0	13,0	0/1	29,8
20	29628070	T	C	0/0	13,0	0/1	22,4
21	11181025	G	A	0/0	39,0	0/1	36,10

CHR chromosome

POS position on the chromosome

REF sequence in the reference genome

ALT alternative sequence detected in the sample

GT genotype in the (diploid) sample, per sample

AC allele count, number of (ref, alt) bases, per sample

vcf file (“good quality variants - all genotypes”)

Apply model of inheritance: e.g. autosomal recessive

Select variants with genotype 0/1 in the parents and 1/1 in the daughter

filtered vcf file (“recessively inherited variants”)

CHR	POS	REF	ALT	GT_daughter	GT_father	GT_mother
1	200827638	A	G	1/1	0/1	0/1
1	22158157	A	G	1/1	0/1	0/1
2	171256597	A	C	1/1	0/1	0/1
2	208976955	A	C	1/1	0/1	0/1
4	48496368	A	G	1/1	0/1	0/1
7	14017007	C	T	1/1	0/1	0/1
8	143310815	G	A	1/1	0/1	0/1
8	41517860	G	A	1/1	0/1	0/1
11	47437403	C	T	1/1	0/1	0/1
11	59837097	C	T	1/1	0/1	0/1
12	9833628	C	T	1/1	0/1	0/1
13	36699762	G	A	1/1	0/1	0/1
13	52523808	C	T	1/1	0/1	0/1
14	38256944	T	C	1/1	0/1	0/1
15	79026001	C	A	1/1	0/1	0/1
16	10788129	G	T	1/1	0/1	0/1
16	1498197	A	G	1/1	0/1	0/1
19	49640002	G	T	1/1	0/1	0/1
20	10026357	T	C	1/1	0/1	0/1
22	23657980	G	A	1/1	0/1	0/1

CHR chromosome

POS position on the chromosome

REF sequence in the reference genome

ALT alternative sequence detected in the sample

GT genotype in the (diploid) sample, per sample

vcf file (“good quality variants - all genotypes”)

Apply model of inheritance: e.g. de-novo

Select variants with genotype 0/0 in the parents and 0/1 in the daughter

filtered vcf file (“de-novo variants”)

CHR	POS	REF	ALT	GT_daughter	GT_father	GT_mother
1	13365778	A	G	0/1	0/0	0/0
1	144676632	T	C	0/1	0/0	0/0
1	144853029	C	G	0/1	0/0	0/0
1	16891333	C	T	0/1	0/0	0/0
3	143697451	T	G	0/1	0/0	0/0
3	72311749	G	T	0/1	0/0	0/0
3	9057481	C	A	0/1	0/0	0/0
5	39002519	C	T	0/1	0/0	0/0
6	33060143	A	G	0/1	0/0	0/0
6	37845185	C	A	0/1	0/0	0/0
7	1586741	G	T	0/1	0/0	0/0
8	49987965	A	C	0/1	0/0	0/0
9	39888209	C	A	0/1	0/0	0/0
12	92562268	G	T	0/1	0/0	0/0
12	94034134	G	T	0/1	0/0	0/0
13	19042019	G	T	0/1	0/0	0/0
15	84855648	C	A	0/1	0/0	0/0
16	46427389	T	C	0/1	0/0	0/0
17	10550780	A	G	0/1	0/0	0/0
22	20643742	C	A	0/1	0/0	0/0

CHR chromosome

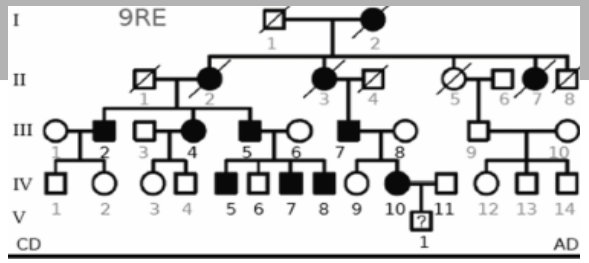
POS position on the chromosome

REF sequence in the reference genome

ALT alternative sequence detected in the sample

GT genotype in the (diploid) sample, per sample

… a real world success story of finding the causative variant

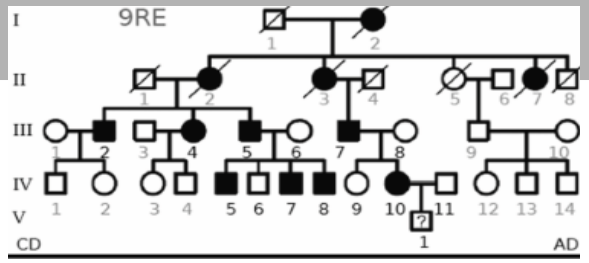


chr17

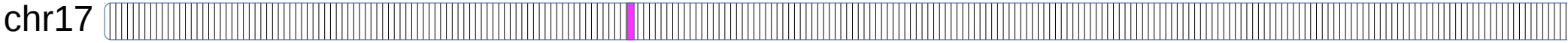
Discard variants because:

- they have low technical quality
- they are known polymorphisms
- they do not have a protein-coding effect

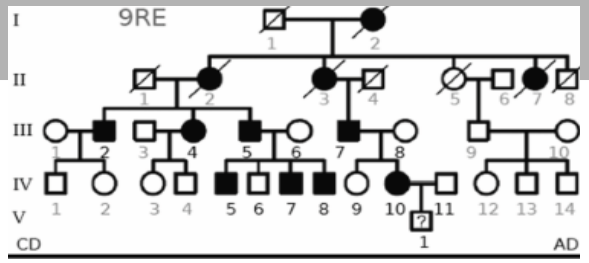
de Castro-Miró M et al. PLOS One 2014: *Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies.*



chr17:7918347, T>C, 0/1

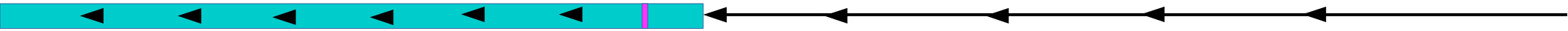


de Castro-Miró M et al. PLOS One 2014: *Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies.*



chr17:7918347, T>C, 0/1

chr17



annotations at gene level

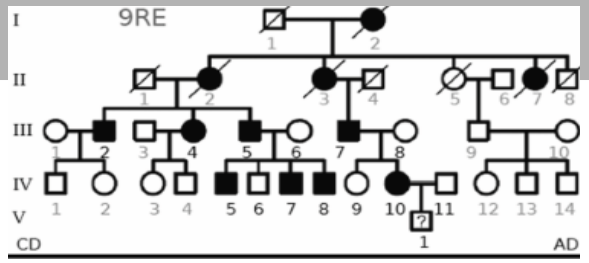
GUCY2D

Retina

- ENSEMBL Functional annotations: genes, transcripts, coding sequences
- UCSC genome browser, GeneCards... Tissue specificity of gene function

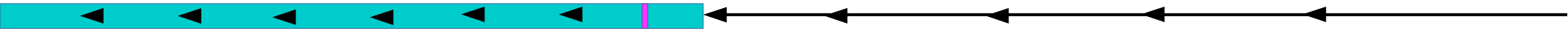
de Castro-Miró M et al. PLOS One 2014: *Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies.*

Causative variant for inherited retinal dystrophy?



chr17:7918347, T>C, 0/1

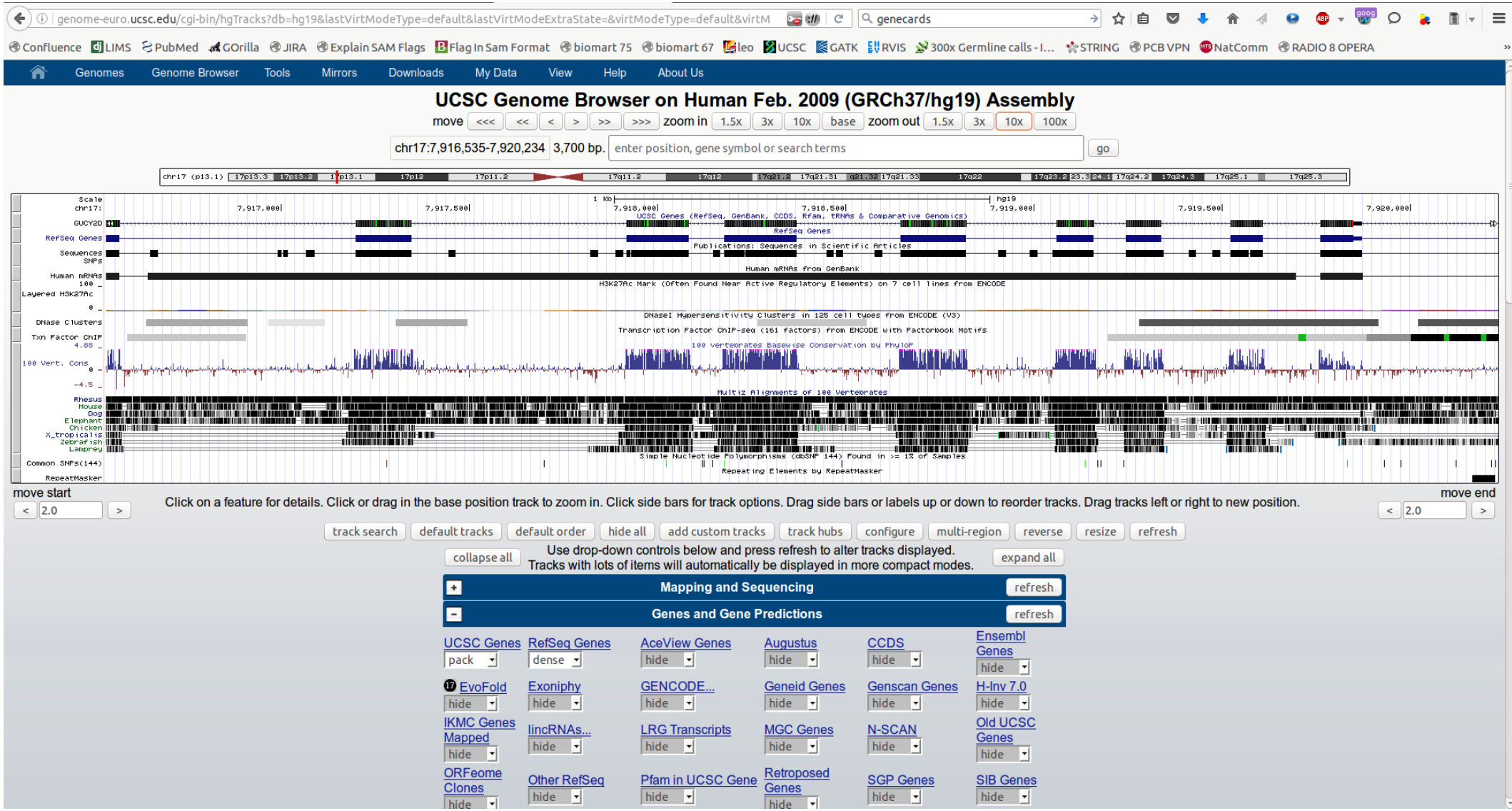
chr17



annotations at gene level	GUCY2D	• ENSEMBL Functional annotations: genes, transcripts, coding sequences
	Retina	• UCSC genome browser, GeneCards... Tissue specificity of gene function
annotations at position level	c.2747T>C p.I916T	• base change • amino acid change
	Variant not annotated	• ExAC (> 60.000 exomes) general population frequency
	damaging probably damaging	• Deleteriousness predictions: · SIFT · PolyPhen2 · CADD

de Castro-Miró M et al. PLOS One 2014: *Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies.*

Look up a gene in a Genome Browser:



Variants **inside candidate genes or genomic regions** are interesting variants

HGMD:: Human Gene Mutation Database (Cardiff University and Biobase GmbH)

OMIM :: Online Mendelian Inheritance in Man (John Hopkins University)

Orphanet :: The portal for rare diseases and orphan drugs (INSERM, France)

ClinVar :: Information about relationships among variation and human health (NCBI)

Disease-specific databases and publications (e.g. COSMIC database for cancer)

Genetic linkage studies

→ Helpful, when studying a case with a previously described disease phenotype

The OMIM database is available and may be queried at: http://omim.org/

The Orphanet database is available at: http://www.orpha.net/consor/cgi-bin/index.php

ClinVar is available at: http://www.ncbi.nlm.nih.gov/clinvar/

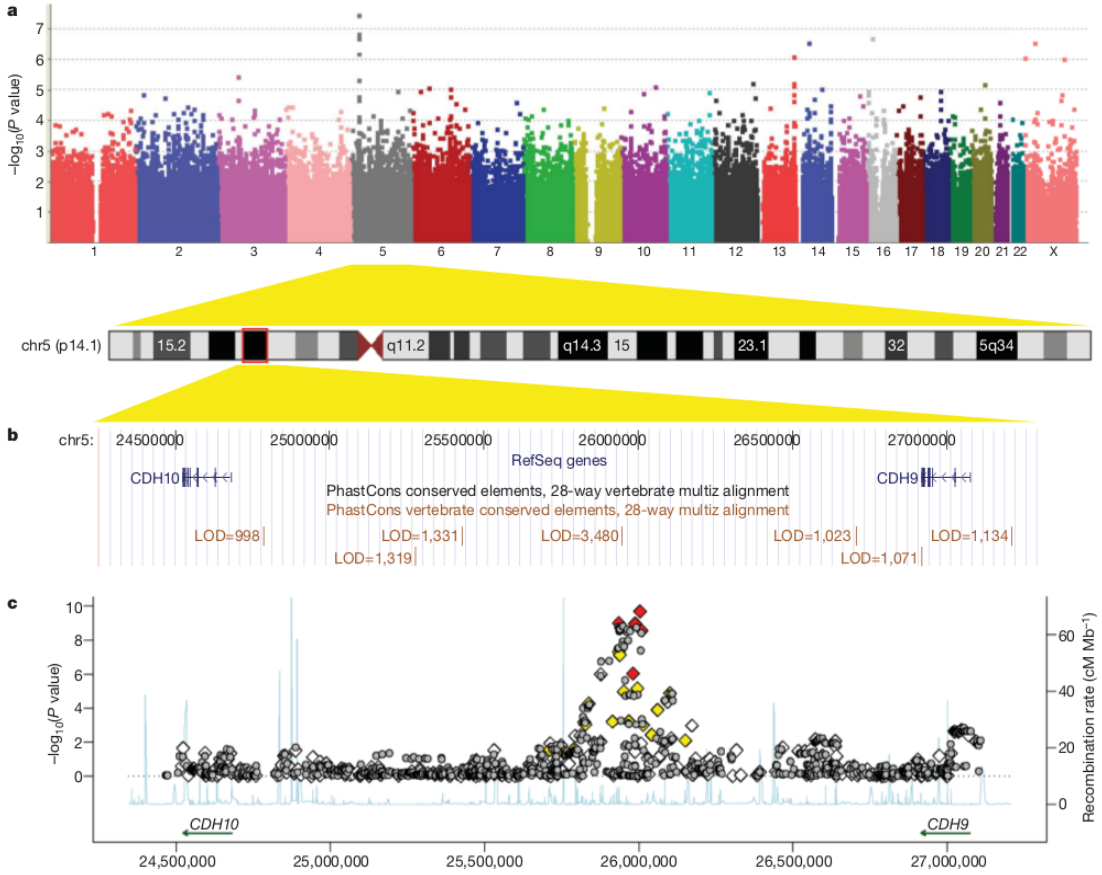
The COSMIC Catalogue for somatic mutations in cancer is available at: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/

What else can genomic variants tell us?

… more complex than coding effect and inheritance

One of the methods to assess complex disease is GWAS – Genome Wide Association Studies.

- Look for genetic polymorphisms (not necessarily coding!) that associate with the trait
- in 1000's of samples: cases and controls, perform statistical tests
- Results can be single position or hotspot region around a position: “Manhattan plots”



e.g. Autism spectrum disorders

← highly associated polymorphisms on chr5

← zoom in

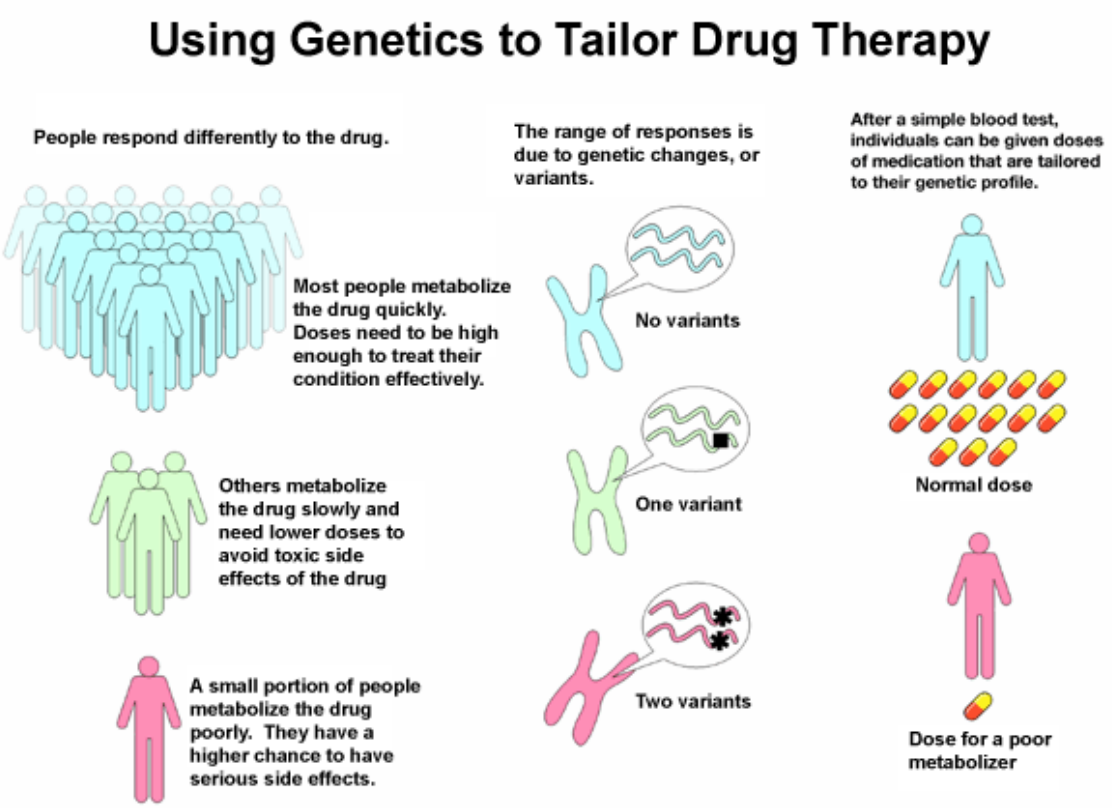
← the hotspot is in the intergenic region

← close-up of the hotspot

Wang K et al. Nature 2009: *Common genetic variants on 5p14.1 associate with autism spectrum disorders.*

Pharmacogenomics is an emerging field that combines genetics with pharmacokinetics and pharmacodynamics of drugs.

- to understand genetic polymorphisms among patients
- to study the effect of these polymorphisms on the activity of the enzyme metabolizing the drug
- to develop more accurate drug dosing in order to avoid intoxication or insufficient drug action.



drug	*Genes with variants affecting drug action*
Warfarin (inhibitor of blood coagulation)	VKORC1 and CYP2C9
Irinotecan (cancer)	UGT1A1
Thiopurine drugs (autoimmune disorders)	TPMT and ITPA

Lee JW et al. Clin Genet 2014: *The emerging era of pharmacogenomics: current successes, future potential,and challenges.*

- originally coined in the field of radiology

A clinically relevant incidental DNA variation can be defined as a verified DNA variation that has a proven medically relevant phenotype not directly related to the condition being studied for research.

It is an unforeseen clinical finding relevant to the individual research participant involved (and possibly to the family of the participant).

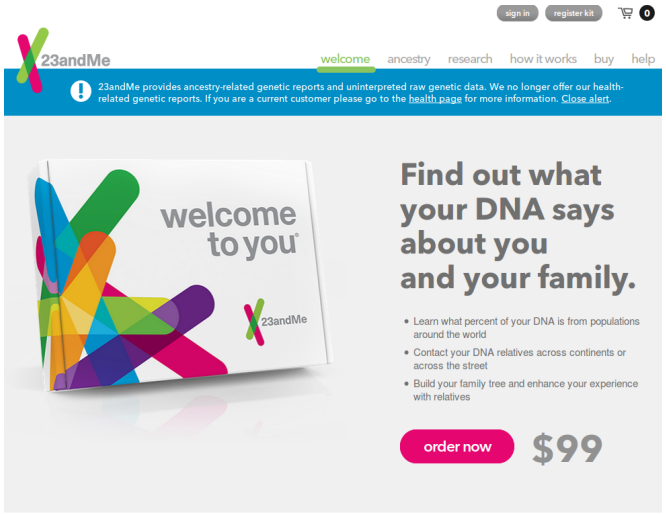
- to be discussed in the field of bioethics

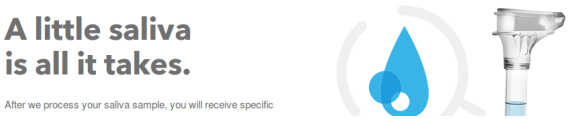
Should the participant (or the participant's physician) be informed about the incidental finding?

Does it make a difference whether the incidentally discovered genetic variant points at a disease with a therapy available or not?

Properly informed consent for the study participants must explain the possibility of finding an incidental DNA variation (especially in whole genome sequencing).

Krier JB and Green RC. Curr Protoc Hum Genet. 2013: *Management of Incidental Findings in Clinical Genomic Sequencing.*

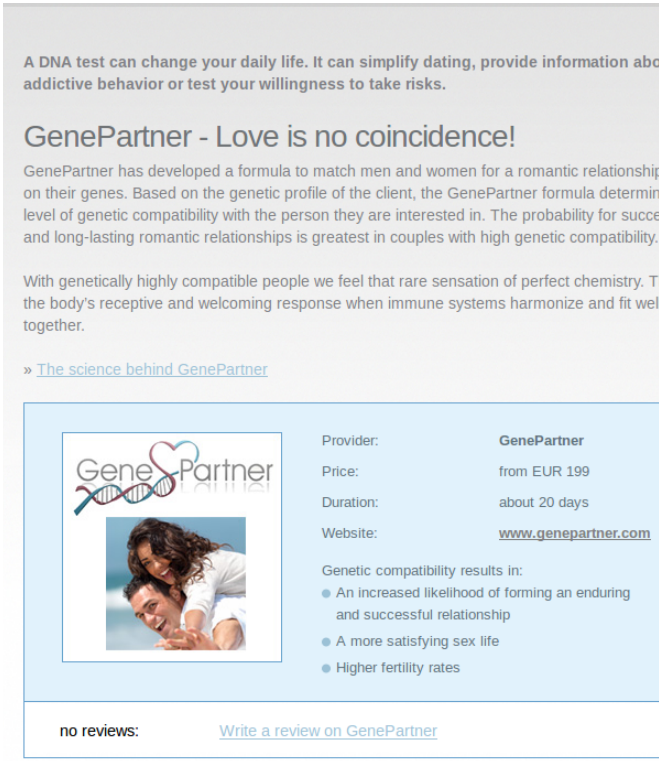




~~Health reports~~

- ancestry-related genetic reports
- uninterpreted raw genetic data
- oddities:

Does fresh cilantro taste like soap to you?
Yes
No
Not sure



Eriksson N et al. arXiv 2012: *A genetic variant near olfactory receptor genes influences cilantro preference.*

5th CNAG Symposium on Genome Research: Single Cell Studies

19th May 2016
Auditori Antoni Caparrós
Torre D - Parc Científic de Barcelona

Free registration at:
www.cnag.eu

Speakers

Heather Lee, Babraham Institute
Christian Conrad, DKFZ & University of Heidelberg
Salvador Aznar-Benitah, Institute for Research in Biomedicine
Thomas Graf, Centre for Genomic Regulation
Eduard Batlle, Institute for Research in Biomedicine
Ramon Massana, Institute of Marine Sciences (CSIC)
Holger Heyn, Centro Nacional de Análisis Genómico

---

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation