# High-troughput technologies, omics data and analysis tools

**Irene Pérez Díez**

Unidad de Bioinformática y Bioestadística

18 de Junio de 2019

Unidad de Bioinformática y Bioestadística

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

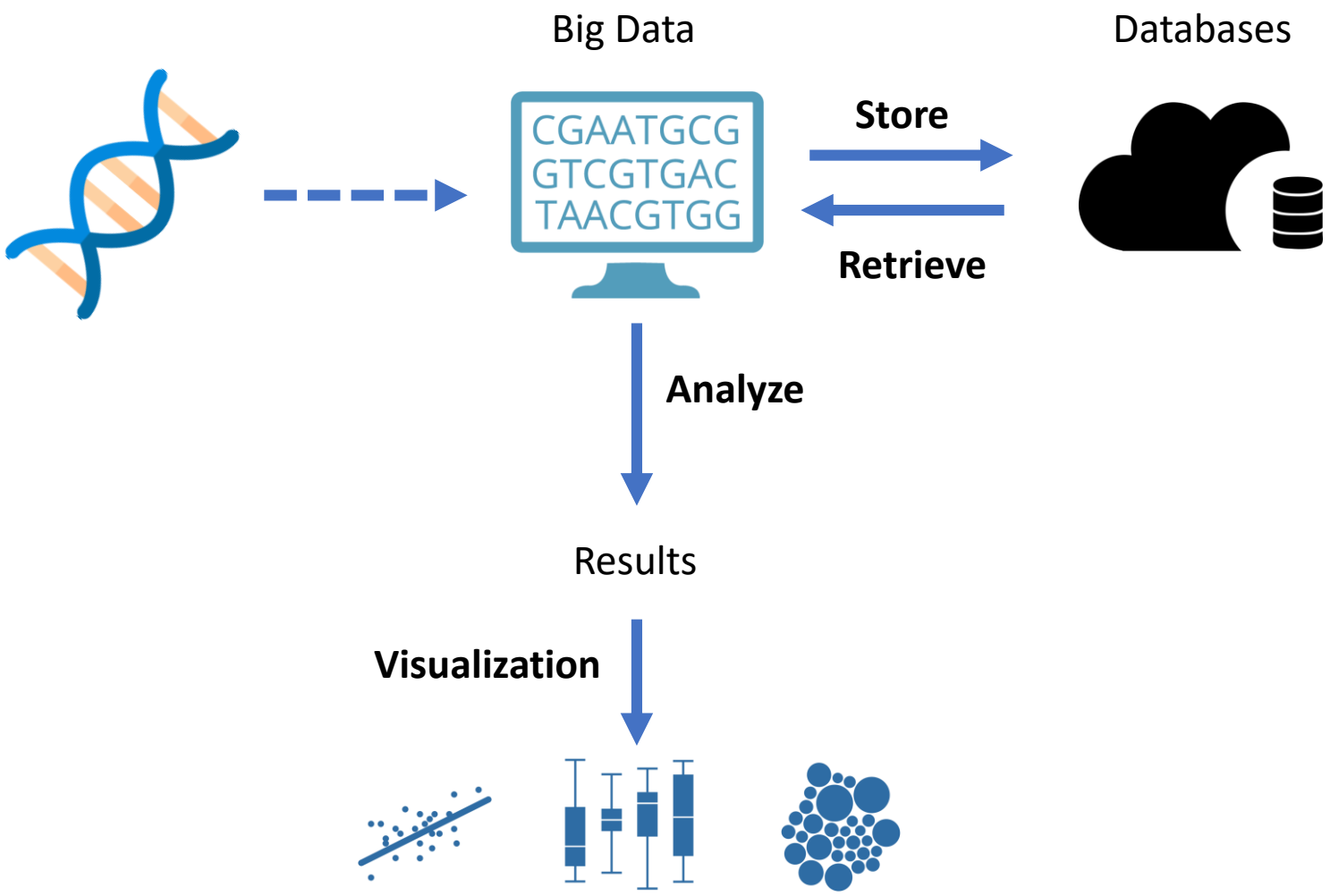
CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

WODA
WEB-BASED OMICS DATA ANALYSIS

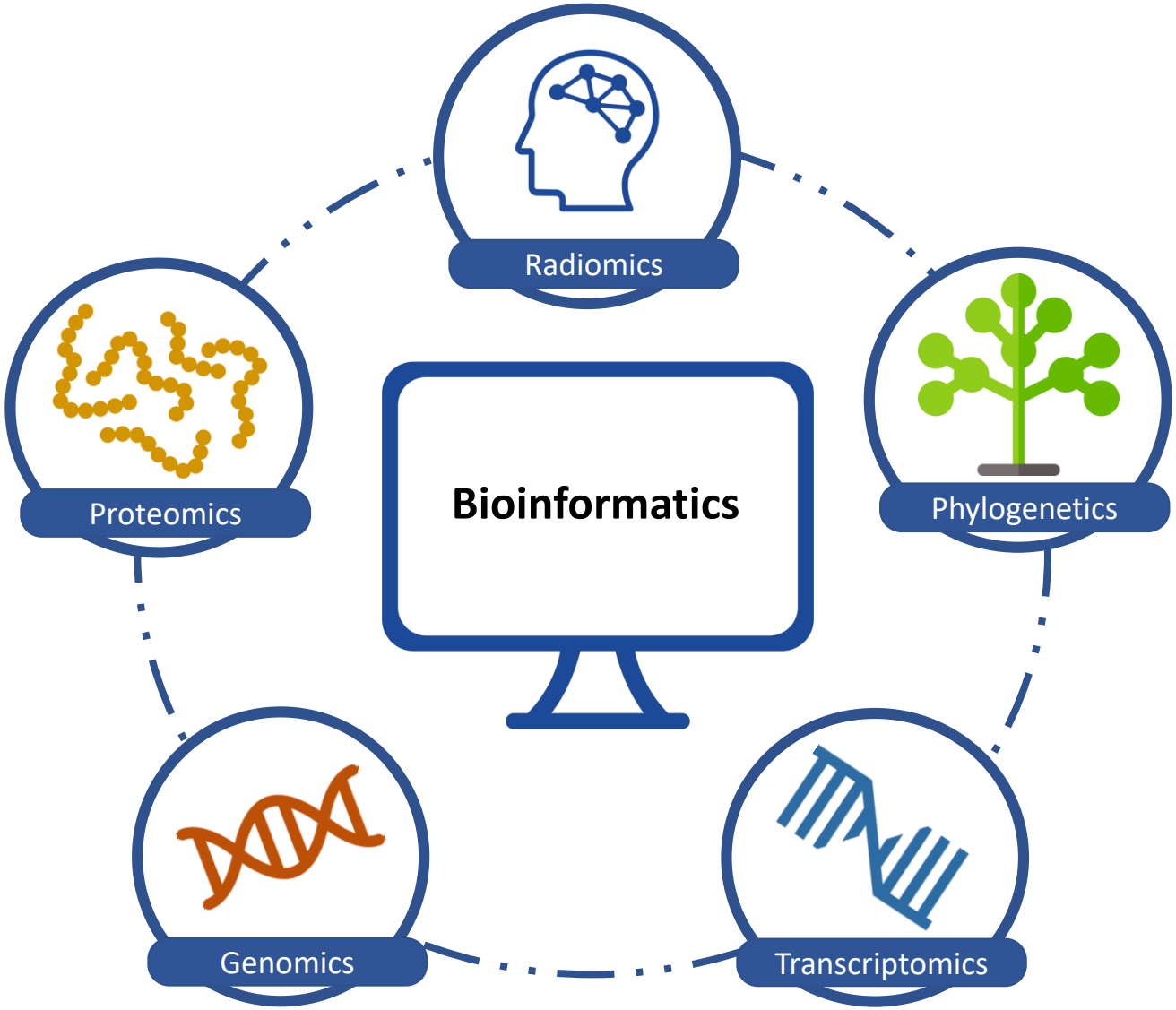
# Outline

- Bioinformatics
- Omics and high-throughput technologies
- Workflow and tools
- Databases

# What is bioinformatics?

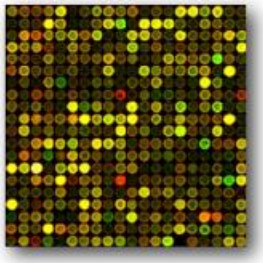


# What is bioinformatics?



# NGS technologies

NGS / High-throughput



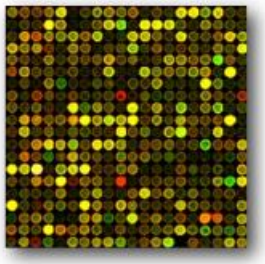
| Sanger DNA Sequencing | Microarrays | 2nd generation sequencing | 3rd generation & single-molecule sequencing |
|---|---|---|---|
| Since 1977 | Since mid-1990s | Since 2007 | Since 2010 |

# NGS technologies

**Sanger**

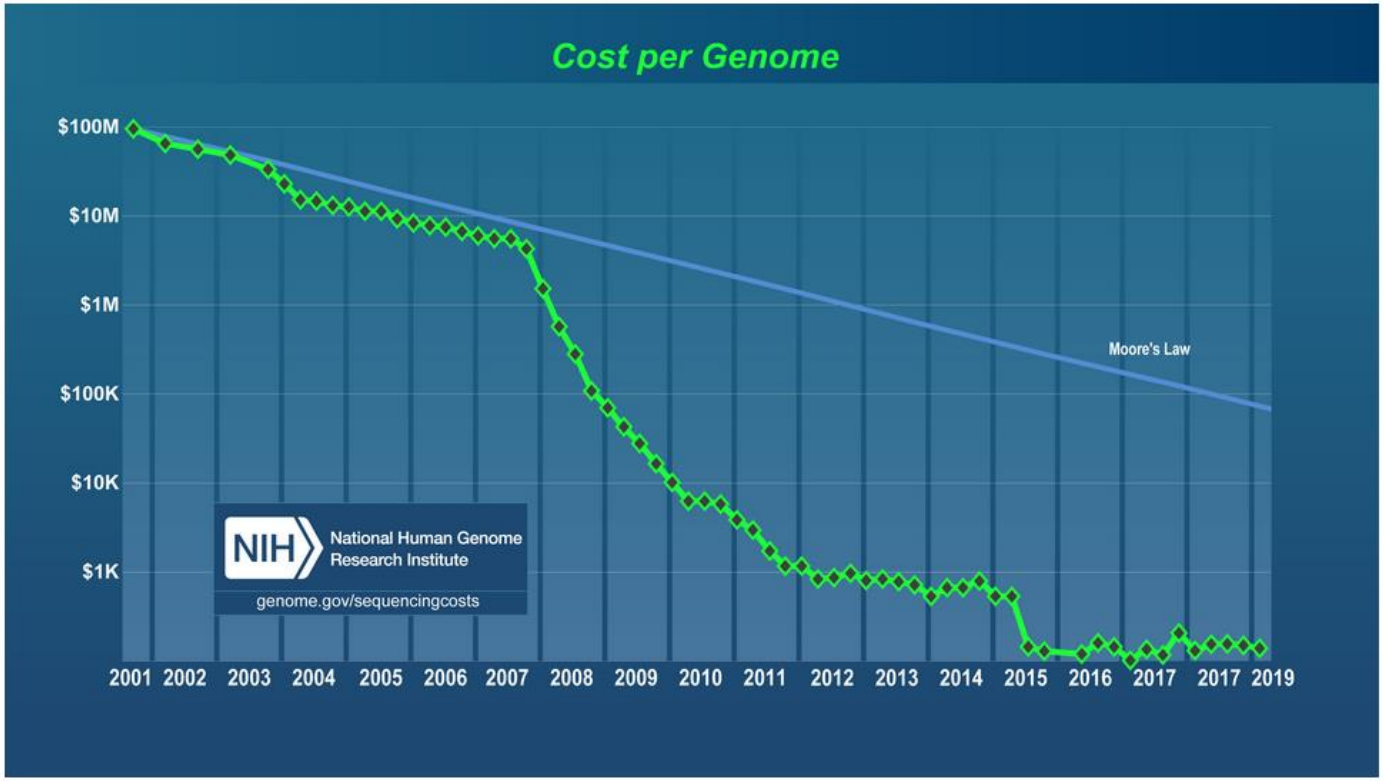
- *De novo* sequencing
- Targeted DNA sequencing
- NGS sequencing validation
- Microbial sequencing
- Mitochondiral sequencing

**Microarray**



- Comparative genomics
- Gene expresión profiling
- Clinical diagnostics
- Metylation analysis

# NGS technologies



# NGS technologies

| Platform (Company) | Chemistry | Read length (bp) | No. reads | Raw error rate (%) | Applications |
|---|---|---|---|---|---|
| 454 (Roche) | Pyro-sequencing | 700 | $1x10^6$ | 1 | Bacterial and viral genomes, multiplex PCR, validation of point mutations, targeted somatic-mutation detection |
| HiSeq (Illumina) | Synthesis | 150x2 | $5x10^9$ | 0,8 | Complex genomes (human, mouse and plants) and genome-wide NGS, RNA-seq, hybrid capture or multiplex-PCR, somatic-mutation detection, forensics, noninvasive prenatal testing |
| MiSeq (Illumina) | Synthesis | 300x2 | $3x10^8$ | 0,8 | |
| SOLiD (Thermofisher) | Ligation | 50 | $1x10^9$ | 0,01 | Complex genomes and genome-wide NGS, RNA-seq, hybrid capture or multiplex-PCR, somatic-mutation detection |
| Ion Torrent (Thermofisher) | Synthesis | 200-400 | $6x10^7$ | 1,7 | Multiplex-PCR, microbiology and infectious diseases, somatic-mutation detection, validation of point mutations |
| 3rd generation | | | | | |
| SMRT (Pac Bio) | Real-time SMS | > 10,000 | $1x10^6$ | 12,9 | Complex genomes, microbiology and infectious-disease genomes, transcript-fusion detection, methylation detection |
| MinION PromethION (Oxford Nanopore) | Real-time SMS | > 5000 | $6x10^4$ | 34 | Pathogen surveillance, targeted mutation detection, metagenomics, bacterial and viral genomes |

Jerzy K. Kulski (January 14th 2016). Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications, Next Generation Sequencing - Advances, Applications and Challenges, Jerzy K Kulski, IntechOpen, DOI: 10.5772/61964

# NGS chemistry overview

A. Library Preparation

B. Cluster Amplification

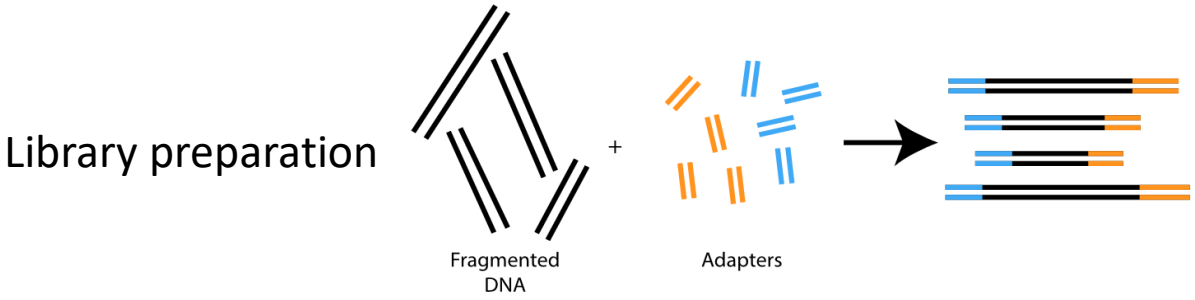
- Bridge PCR
- Emulsion PCR

C. Sequencing

- Pyrosequencing
- Sequencing by synthesis
- Ion semiconductor sequencing
- Sequencing by ligation
- Real-time SMS  No PCR!!

D. Alignment/mapping and Data Analysis

# Types of libraries

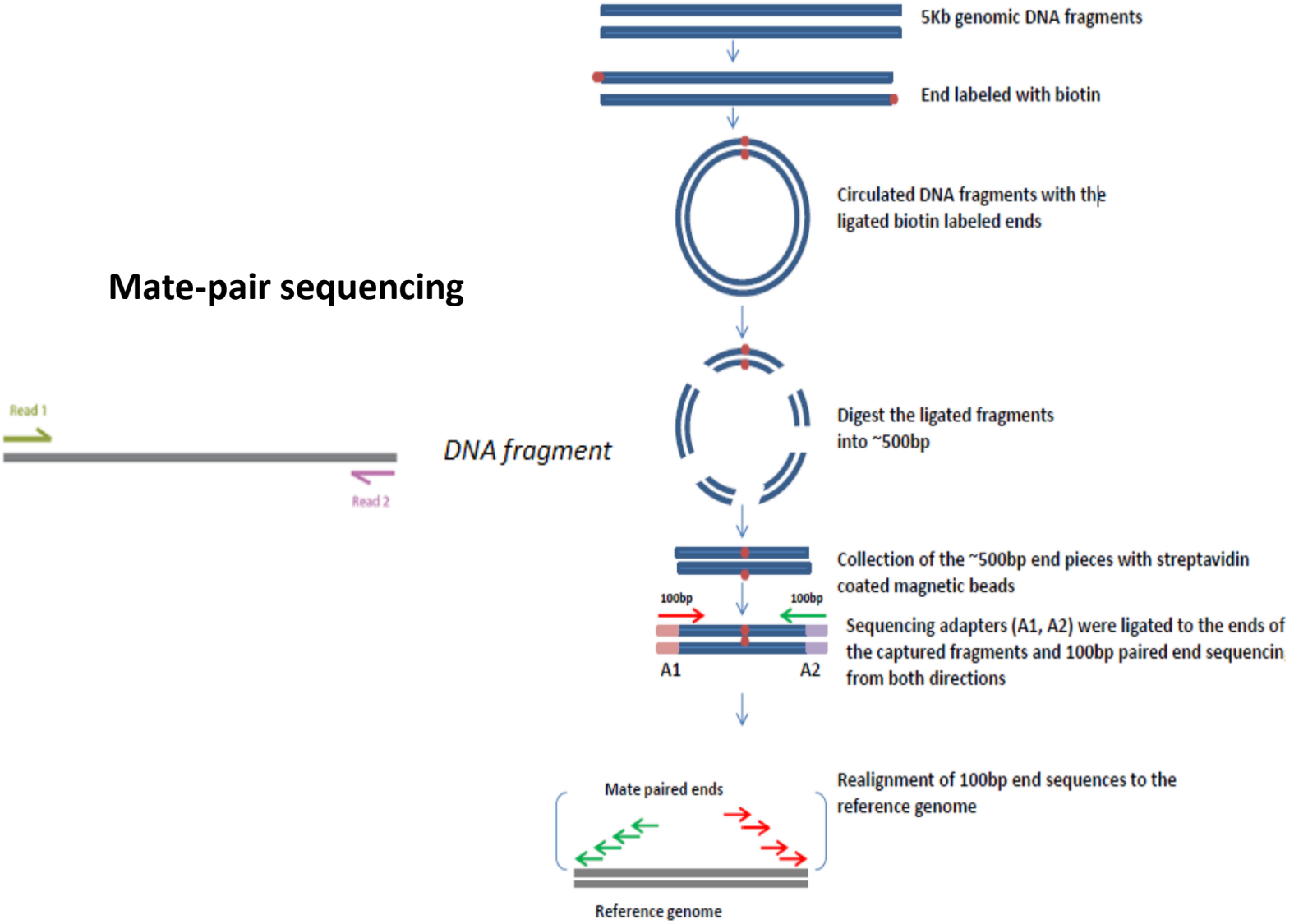



**Single-end sequencing**

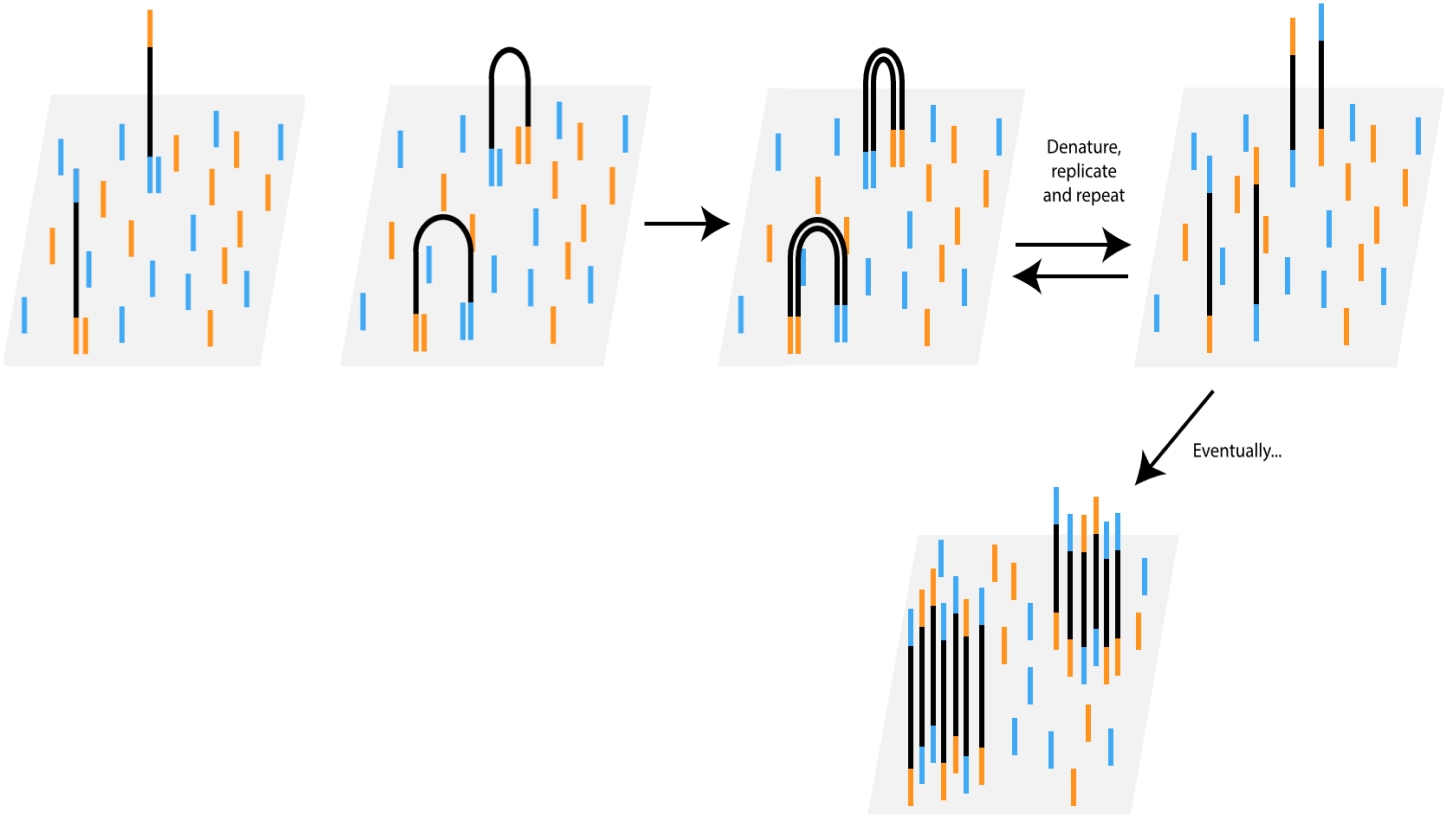

**Paired-end sequencing**

# Types of libraries

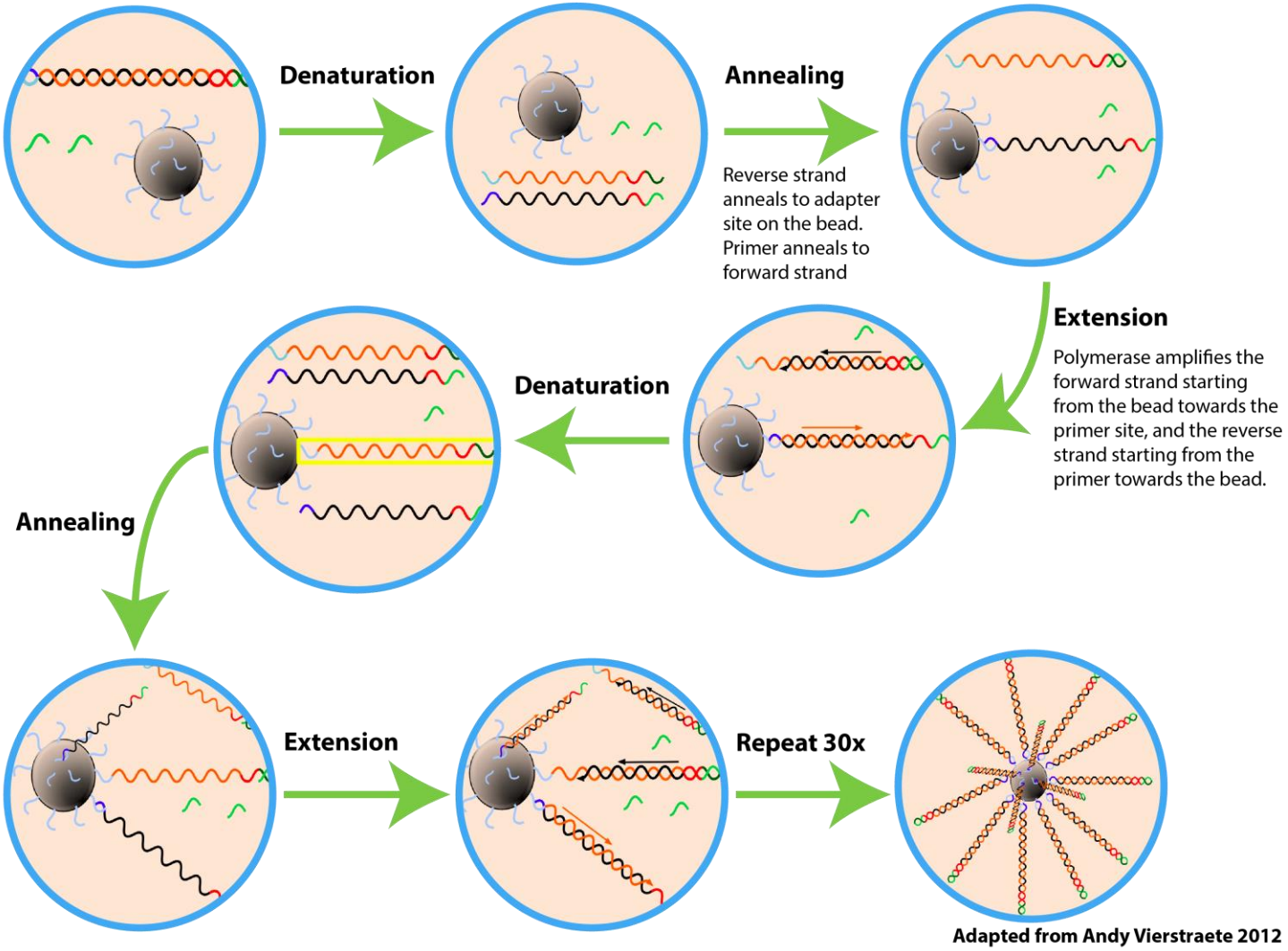
**Mate-pair sequencing**



# Bridge PCR

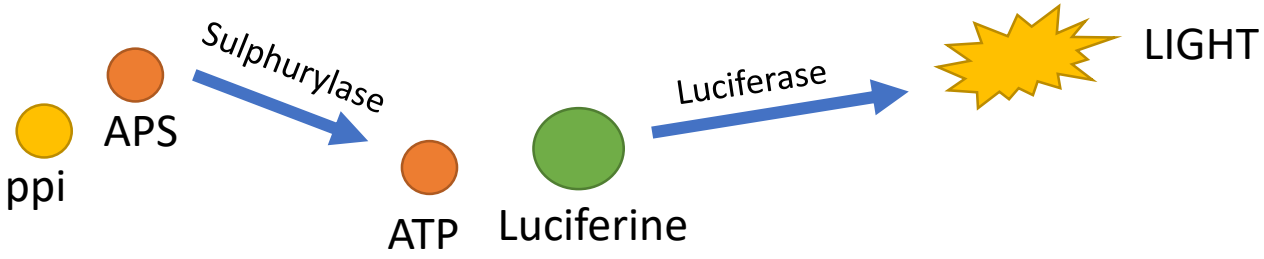


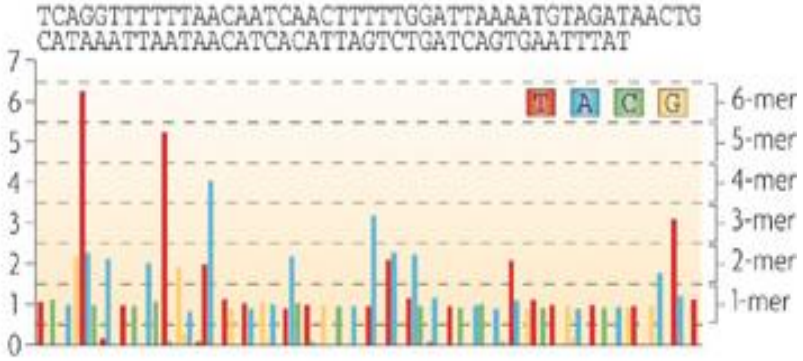
# Emulsion PCR



Roche

# Pyrosequencing

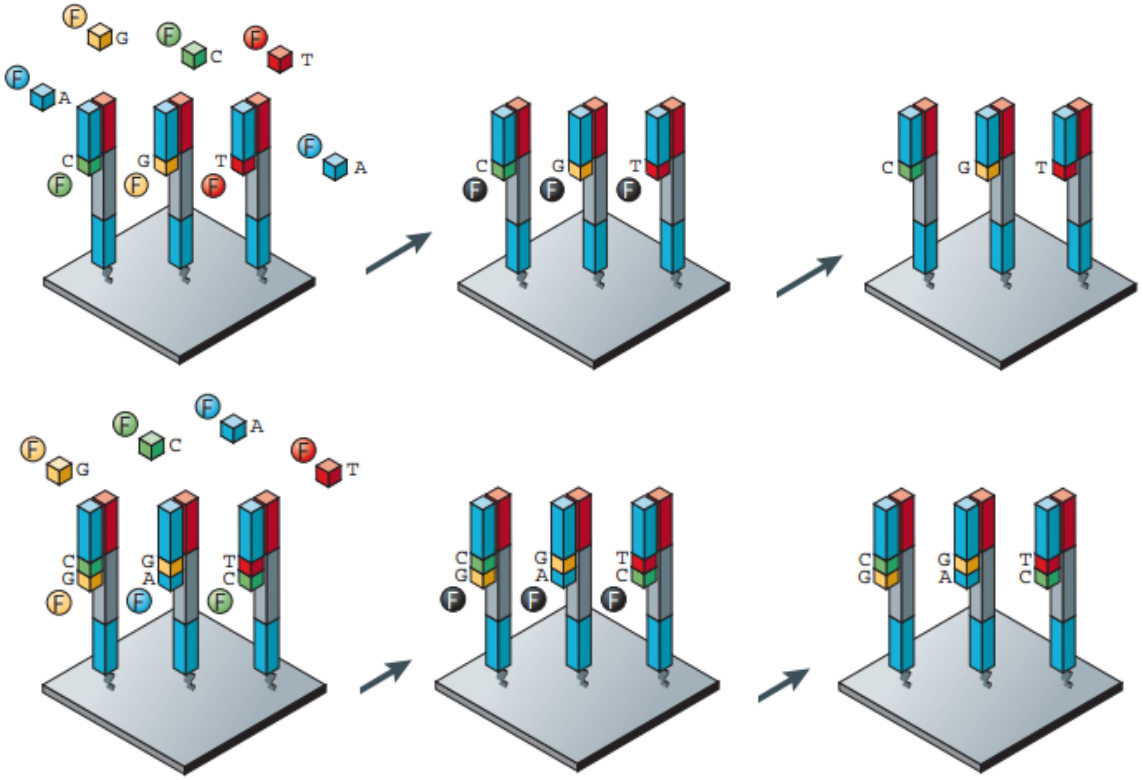



- Large read lenghts generation
- High reagent cost
- High error rate over strings of 6+ homopolymer

Illumina

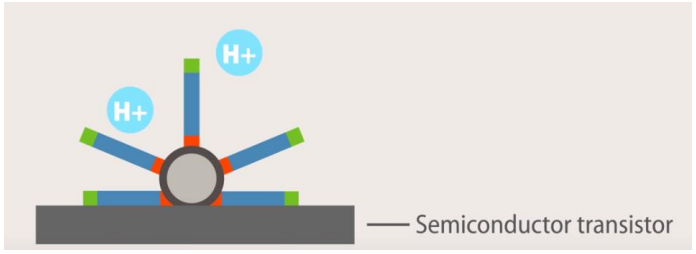
# Sequencing by synthesis

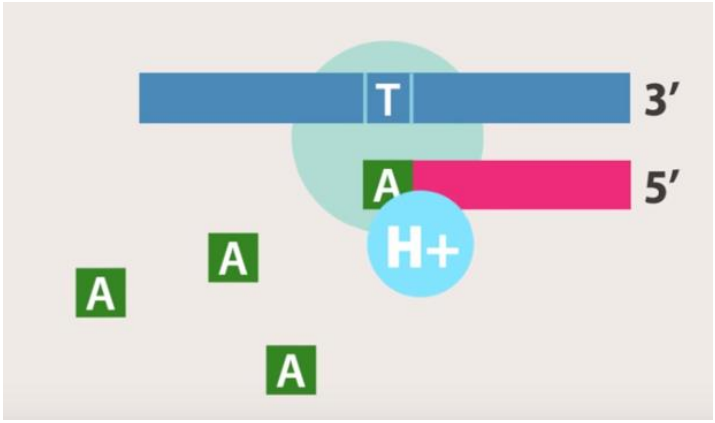


- Overcomes homopolymer issue due to terminated nucleotides
- Increased error rate with read length

Ion Torrent - Thermofisher

# Ion semiconductor

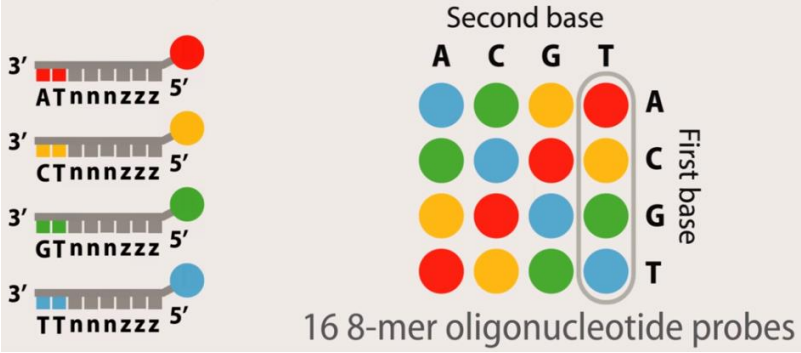


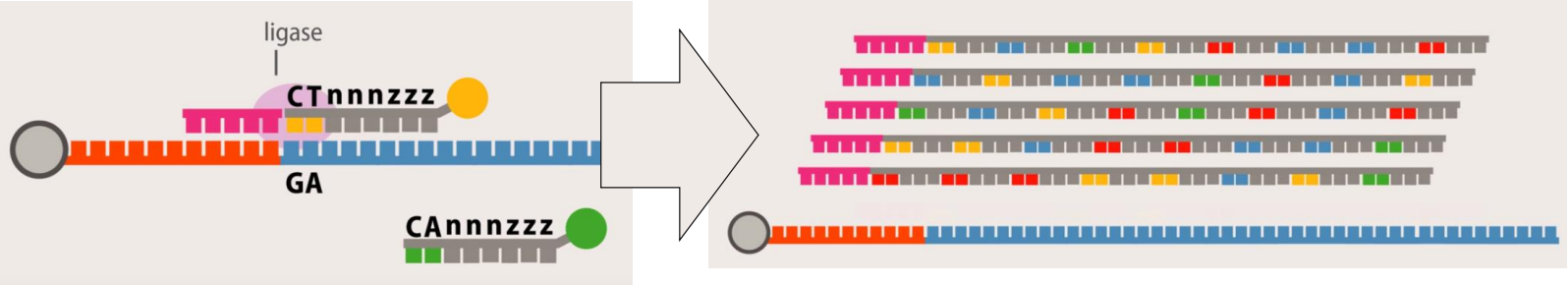




- Similar to pyrosequencing, but measures the release of H+ instead of pyrophosphate
- More cost-effective and time-efficient

SOLiD- Thermofisher

# Sequencing by ligation

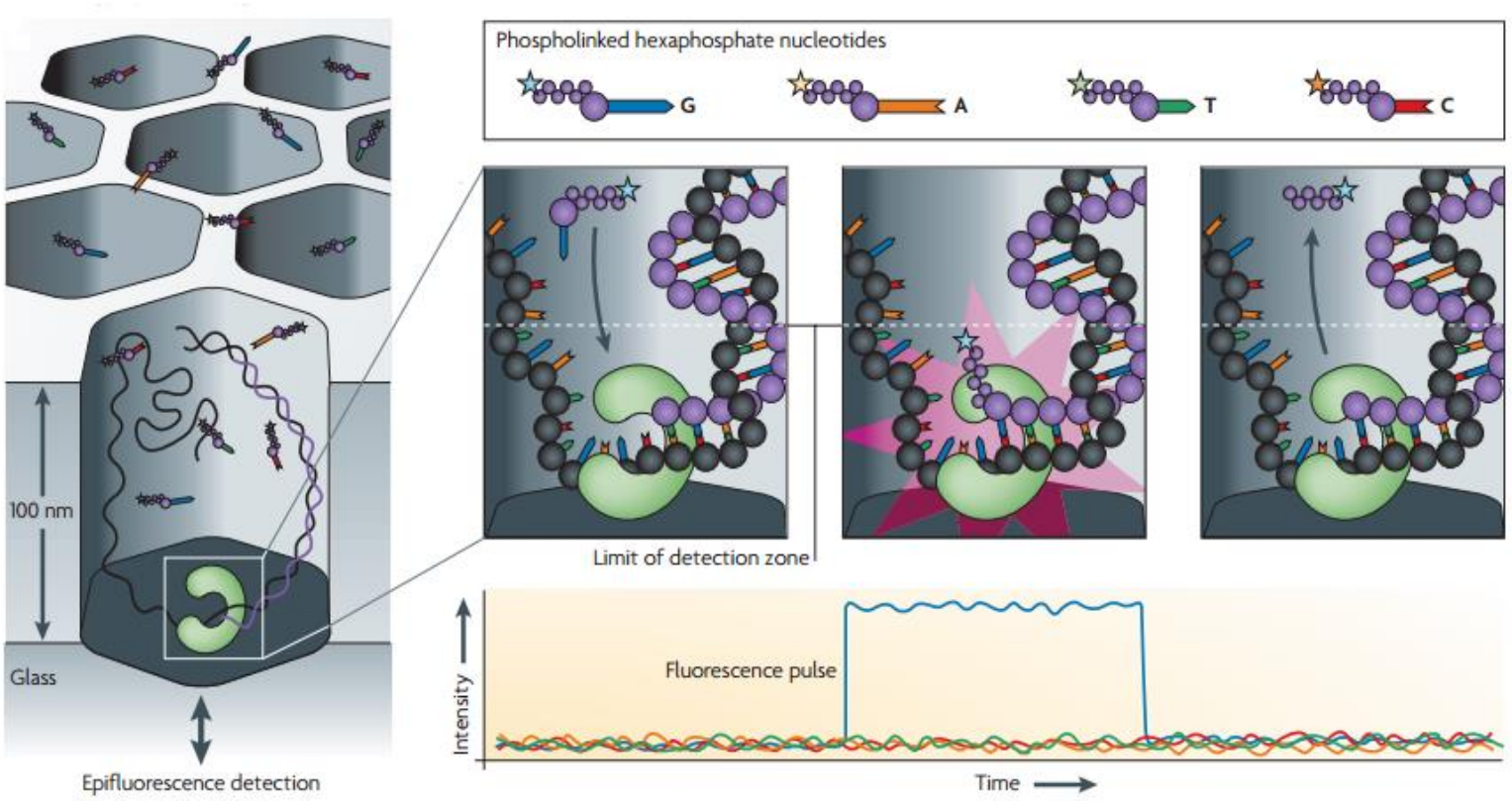



- Oligonucleotide probes used rather than DNA polymerase
- Very short read lengths

Pac Bio

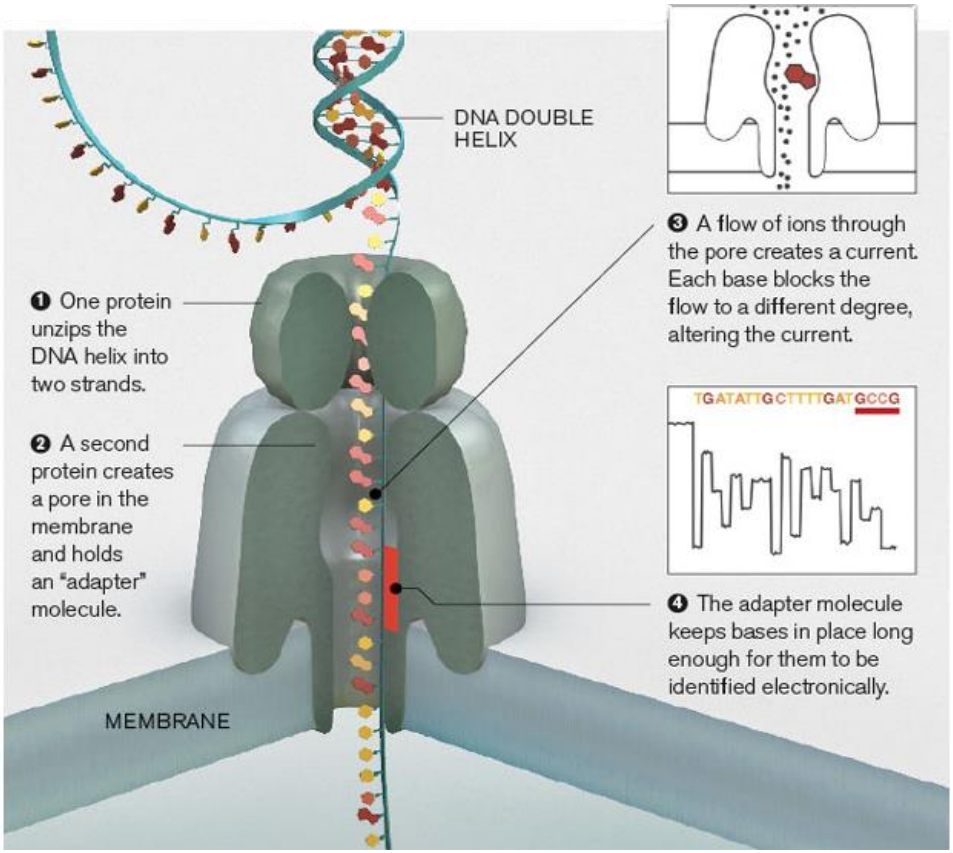
# Real-time SMS




- Non-stop sequencing, no need to “wash and scan”
- DNApol fixed at the bottom of the well, the laser detector aims at the active site

Oxford Nanopore

# Real-time SMS




- Non-stop sequencing, no need to "wash and scan"
- High error rate, this technology is still improving

# Applications

Whole Genome Sequencing (WGS)

Variant Calling

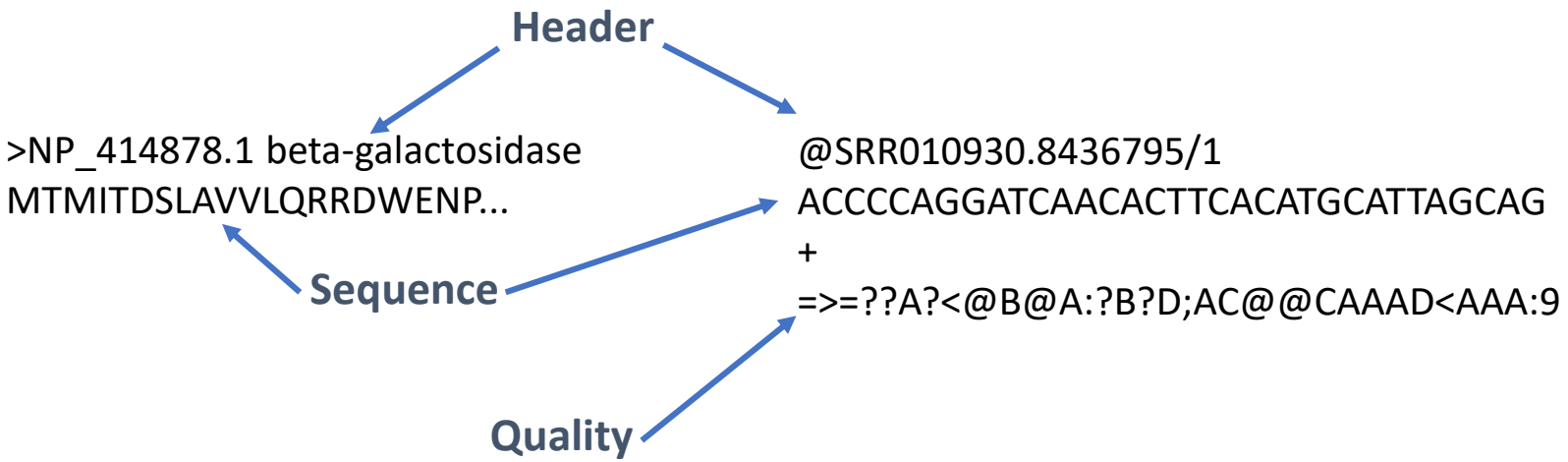
RNA-seq

De novo sequencing

Metagenomics

# FASTA - FASTQ

- Fasta
  - .fasta
  - .fa
- Fastq
  - .fastq
  - .fq



# PHRED quality

- Encodes the probability of an erroneous call
  - Phred quality score $Q = -10 \log_{10} P$
  - Base-calling error probabilities $P = 10^{-Q/10}$
  - ASCII encoding

Example: call with Q = 30 has error probability $P = 10^{-3}$ = 1 in 1000

| Encoding | ! | “ | # | $ | % | & | ‘ | ( | ) | * | + | , | - | . | / | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q score | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |

# Computing requirements

Conditioned data center (server roms)
Computing cluster
Many computer nodes (servers)
High performance and storage capacity
Fast networks
Sysadmins and developers

# Computational infrastructure

- 44 computer nodes
- 600 CPUs
- 11 TB RAM memory
- 1 PB storage capacity

Public resource devoted to the storage and analysis of data coming from:

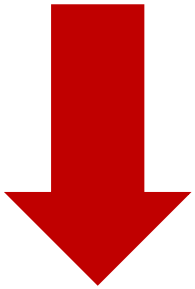
1. Precision medicine data
2. Bioimage
3. Biomedical data

The use of the cluster is open to public institutions and research groups in the Comunitat Valenciana working in biomedical fields.

ubb@cipf.es

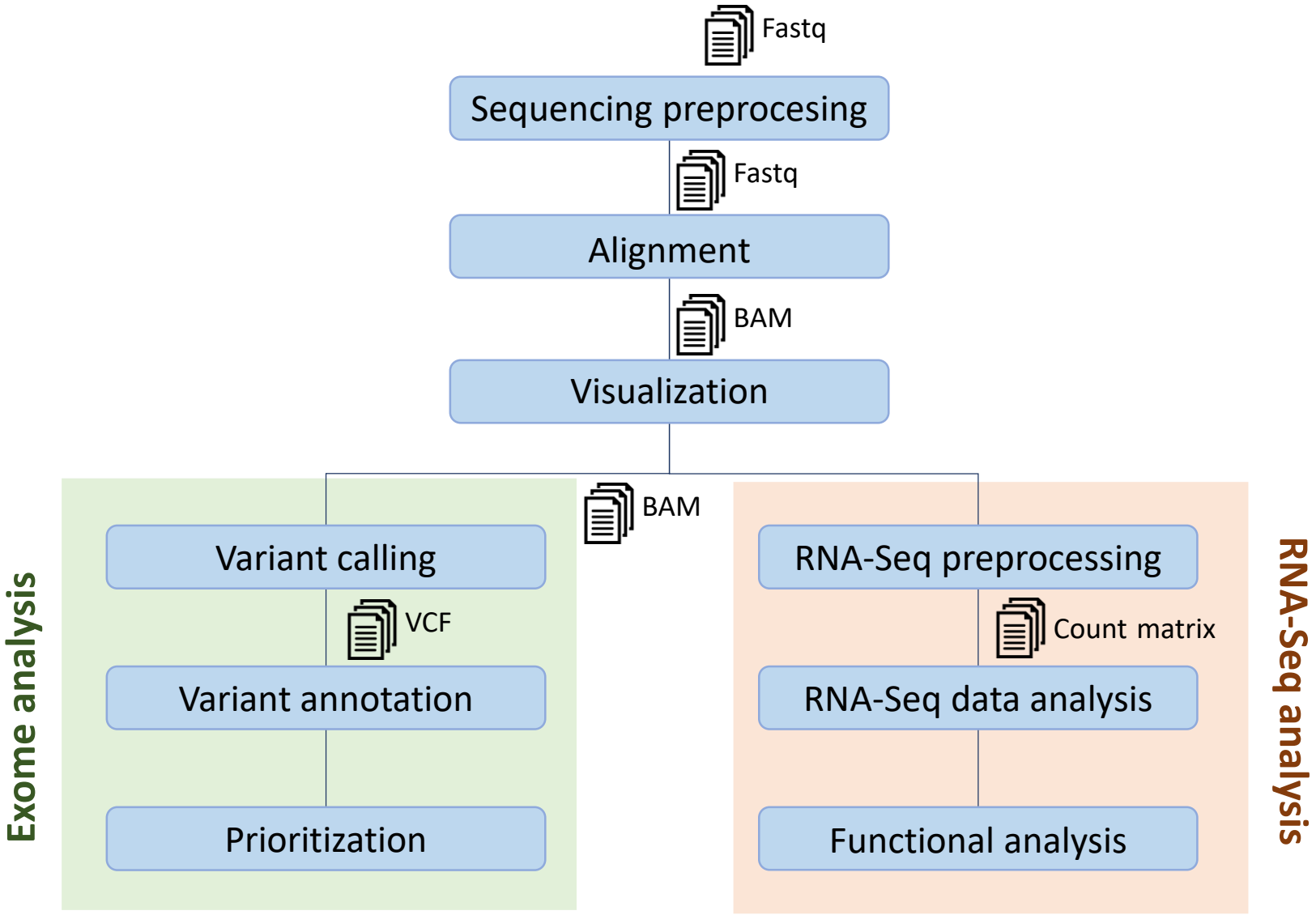
# Cloud computing

- Flexibility
- You pay what you use
- Don’t need to mantain a data center

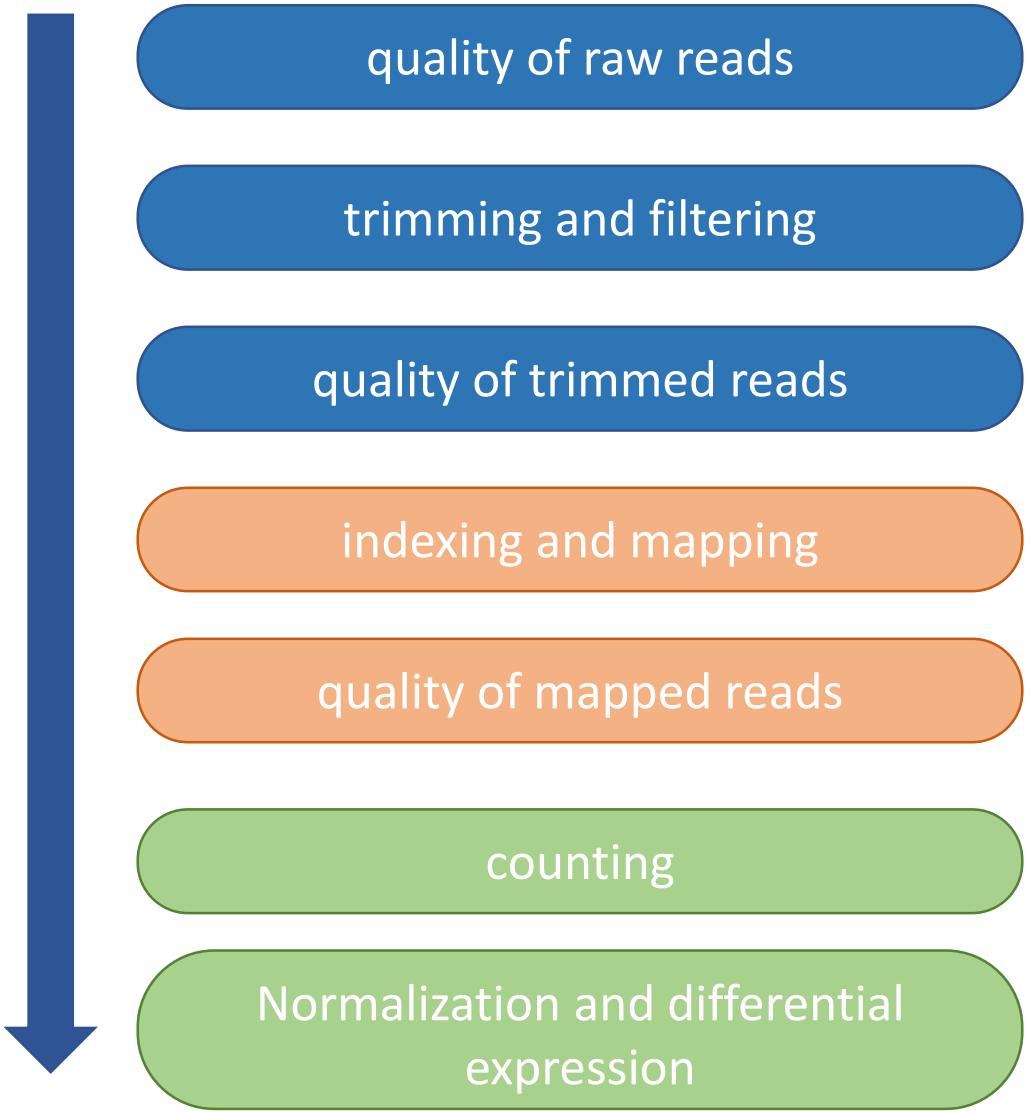


- Transfer datasets through the internet is slow
- Lower performance
- Privacy and security concerns
- More expensive for big and long term projects

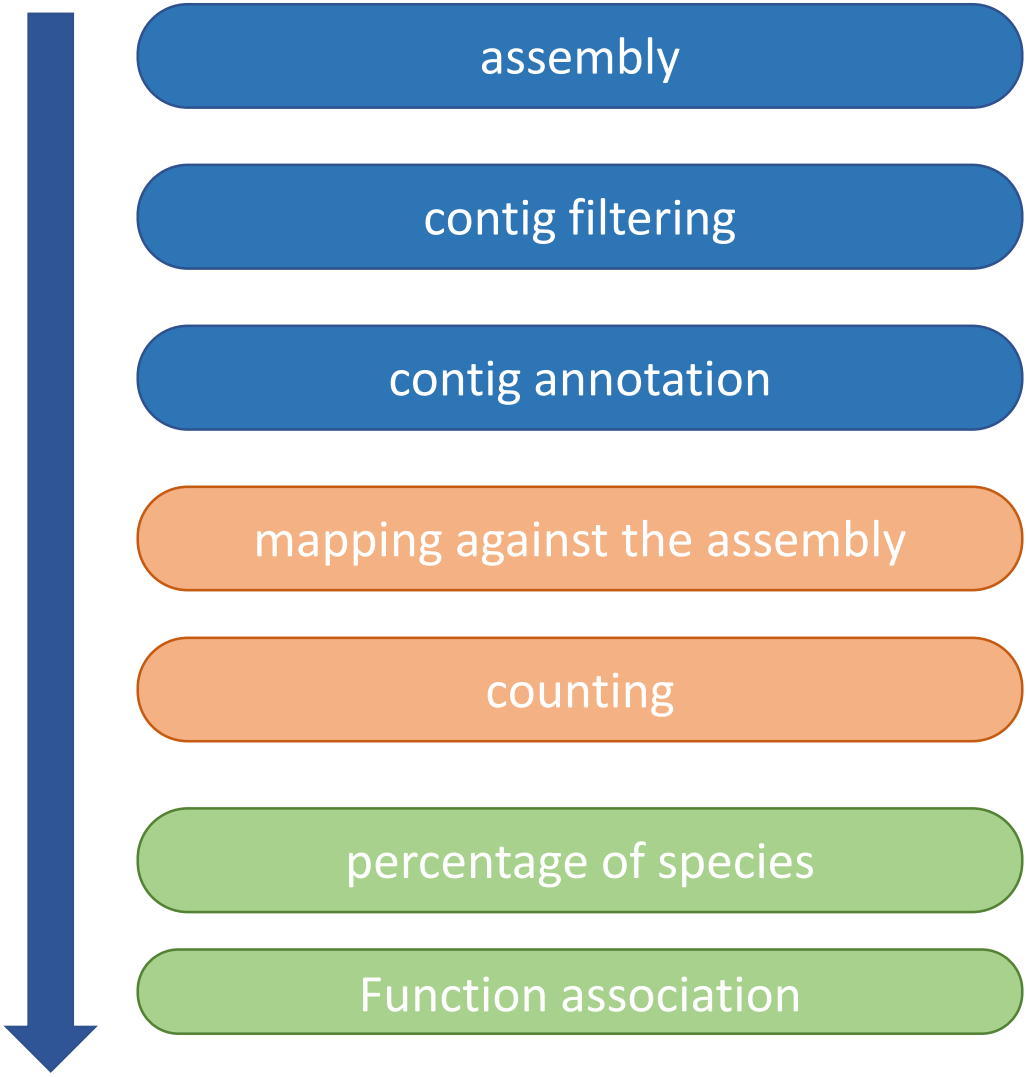
# Basic workflow



# RNAseq pipeline



# Metagenomics pipeline



# NGS tools

| | |
|---|---|
| **FastQC**<br>https://www.bioinformatics.babraham.ac.uk/projects/fastqc/<br>Quality control | **Blast2GO**<br>https://www.blast2go.com/<br>Functional annotation and analysis |
| **Cutadapt**<br>https://cutadapt.readthedocs.io/en/stable/<br>Trimming: remove adaptors and other sequences | **Samtools**<br>http://www.htslib.org/<br>Work with SAM/BAM/CRAM files |
| **Bowtie2**<br>http://bowtie-bio.sourceforge.net/bowtie2/index.shtml<br>Alignment | **Vcftools**<br>https://vcftools.github.io/index.html<br>Work with VCF files |
| **Bwa**<br>http://bio-bwa.sourceforge.net/<br>Alignment | **GATK**<br>https://software.broadinstitute.org/gatk/<br>From variant discovery to metagenomics |
| **STAR**<br>https://github.com/alexdobin/STAR<br>RNA-seq aligner | **HISAT2**<br>http://ccb.jhu.edu/software/hisat2/index.shtml<br>RNA / DNA aligner |

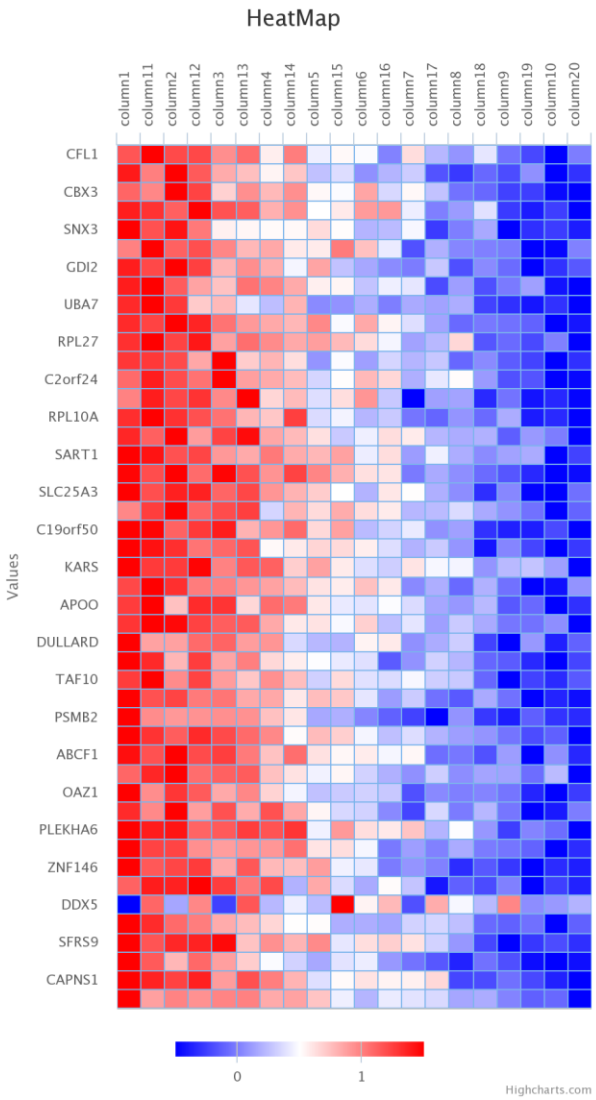
# NGS tools

| | |
|---|---|
| **Cufflinks**<br>http://cole-trapnell-lab.github.io/cufflinks/<br><br>Transcriptome assembly and differential expression | **Mothur**<br>https://www.mothur.org/<br><br>Microbial ecology toolbox |
| **ABySS**<br>https://github.com/bcgsc/abyss<br><br>*de novo* sequence assembler (large genomes) | **Bismark**<br>https://www.bioinformatics.babraham.ac.uk/projects/bismark/<br><br>Bisulfite converted sequence reads – cytosine methylation |
| **SPAdes**<br>http://cab.spbu.ru/software/spades/<br><br>Genome assembler (small genomes) | **BLAST**<br>https://blast.ncbi.nlm.nih.gov/Blast.cgi<br><br>Alignment |
| **GLIMMER**<br>https://ccb.jhu.edu/software/glimmer/<br><br>Gene predictor - microbial DNA | **Augustus**<br>http://bioinf.uni-greifswald.de/augustus/<br><br>Gene predictor – eukaryotic DNA |
| **IGV**<br>https://software.broadinstitute.org/software/igv/<br><br>Genome visualization | **Quiime2**<br>https://qiime2.org/<br><br>Microbiome bioinformatics platform |

# NGS tools


Babelomics 5
log in
sign up
Babelomics 5
GENE EXPRESSION, GENOME VARIATION AND FUNCTIONAL PROFILING ANALYSIS SUITE
Try it now
Note
Web optimized for Chrome. Only modern web browsers are fully supported, these include Chrome 36+, Firefox 36+, Safari 8+ and Opera 24+.
For teaching activities with Babelomics we recommend you to use:
courses.babelomics.org
BABELOMICS: developed by the Computational Genomics Department
bioinfo.cipf.es: babelomics@cipf.es
Principe Felipe Research Center

# NGS tools

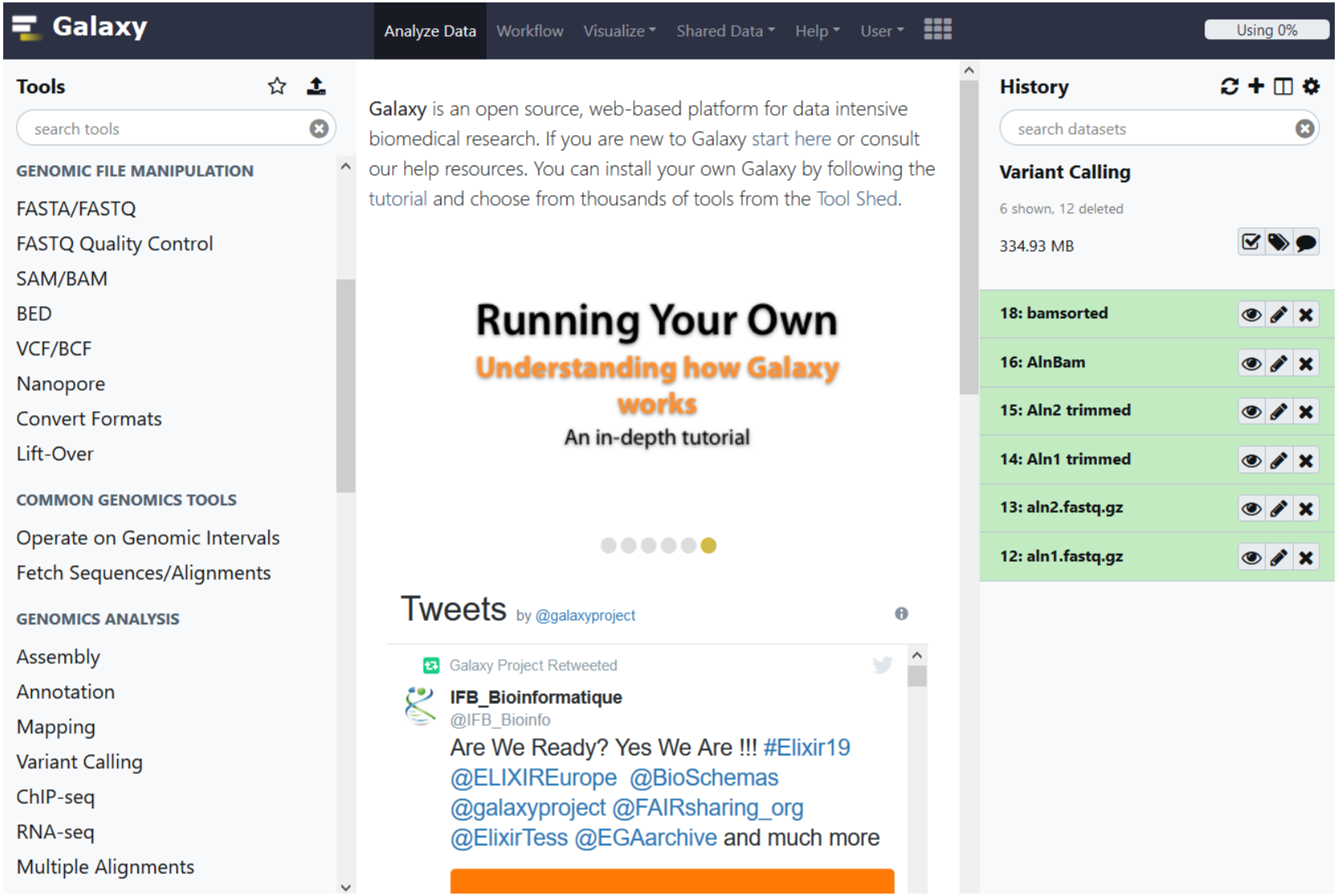



| #NAMES | statistic | coefs. | p-value | adj. p-value |
|---|---|---|---|---|
| CFL1 | 3.16 | 1.05 | 0.0016 | 0.019 |
| CCL5 | 3.12 | 1.04 | 0.0018 | 0.019 |
| CBX3 | 3.04 | 1.06 | 0.0024 | 0.019 |
| RPL19 | 2.99 | 1.02 | 0.0028 | 0.019 |
| SNX3 | 2.95 | 0.85 | 0.0032 | 0.019 |
| JTB | 2.95 | 0.94 | 0.0032 | 0.019 |
| GDI2 | 2.93 | 1.04 | 0.0034 | 0.019 |
| RPS24 | 2.93 | 0.96 | 0.0034 | 0.019 |
| UBA7 | 2.92 | 0.91 | 0.0034 | 0.019 |
| MYST2 | 2.9 | 1.11 | 0.0037 | 0.019 |

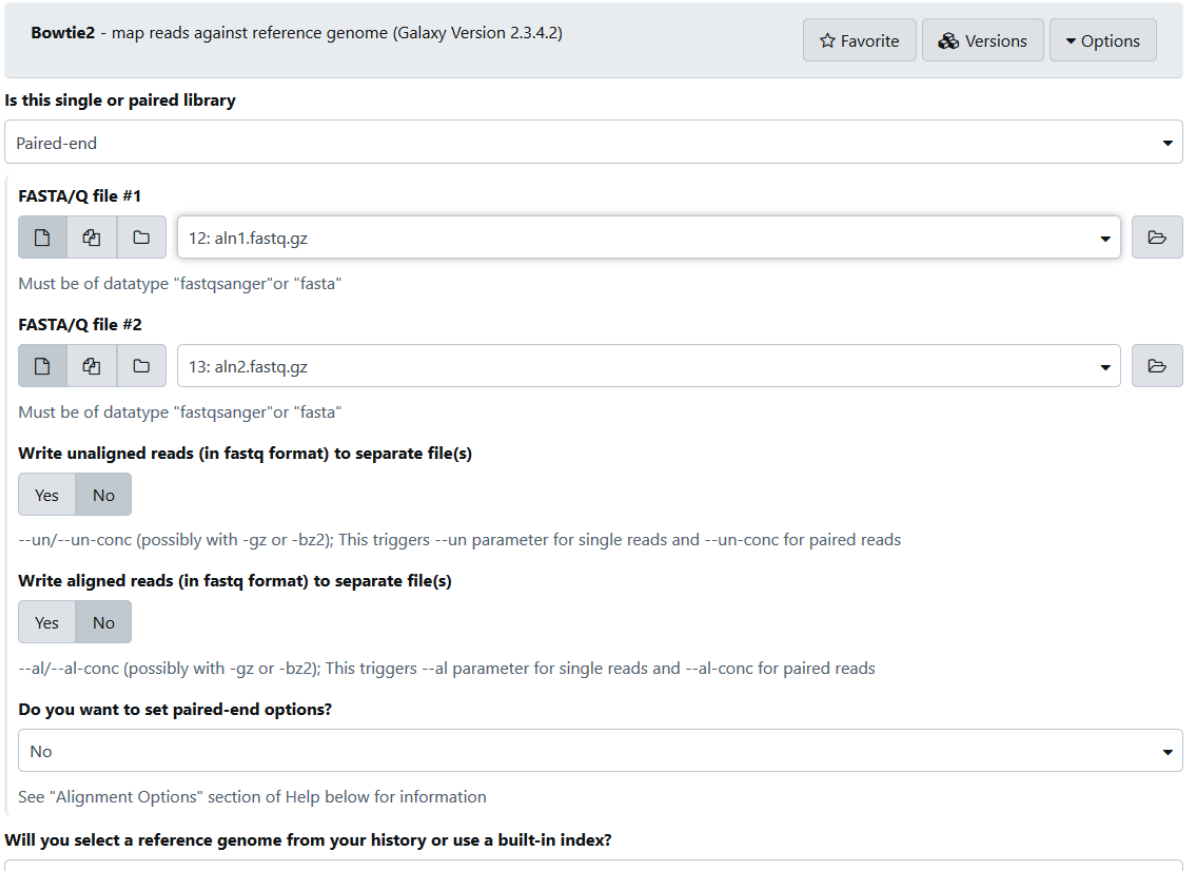
46 Results ‹ 1 of 5 ›

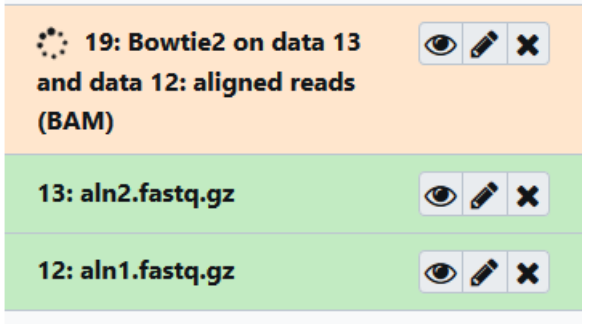
| Term | Term size | Term size(in genome) | annotated_genes lists | converged ids list | lor | adj_pvalue |
|---|---|---|---|---|---|---|
| positive regulation of developmental process(GO:0051094) | 8 | 1937 | THRA SART1 PAX8 RHOA EPHB3 BAD | true | -0.51 | 0.049 |
| negative regulation of cellular biosynthetic process(GO:0031327) | 11 | 2778 | THRA CBX3 NONO TARDBP RPS27A KHDRBS1 | true | -0.52 | 0.024 |
| heterocycle metabolic process(GO:0046483) | 64 | 20258 | RPL18 ABCF1 RPL17 CHURC1 THRA RPL19 | true | -0.4 | 0.02 |

# NGS tools


Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 0%
Tools
search tools
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Sequences/Alignments
GENOMICS ANALYSIS
Assembly
Annotation
Mapping
Variant Calling
ChIP-seq
RNA-seq
Multiple Alignments
Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.
Running Your Own
Understanding how Galaxy works
An in-depth tutorial
Tweets by @galaxyproject
Galaxy Project Retweeted
IFB_Bioinformatique
@IFB_Bioinfo
Are We Ready? Yes We Are !!! #Elixir19 @ELIXIREurope @BioSchemas @galaxyproject @FAIRsharing_org @ElixirTess @EGAarchive and much more
History
search datasets
Variant Calling
6 shown, 12 deleted
334.93 MB
18: bamsorted
16: AlnBam
15: Aln2 trimmed
14: Aln1 trimmed
13: aln2.fastq.gz
12: aln1.fastq.gz

# NGS tools


Bowtie2 - map reads against reference genome (Galaxy Version 2.3.4.2)
Favorite
Versions
Options
Is this single or paired library
Paired-end
FASTA/Q file #1
12: aln1.fastq.gz
Must be of datatype "fastqsanger"or "fasta"
FASTA/Q file #2
13: aln2.fastq.gz
Must be of datatype "fastqsanger"or "fasta"
Write unaligned reads (in fastq format) to separate file(s)
Yes
No
--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads
Write aligned reads (in fastq format) to separate file(s)
Yes
No
--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads
Do you want to set paired-end options?
No
See "Alignment Options" section of Help below for information
Will you select a reference genome from your history or use a built-in index?

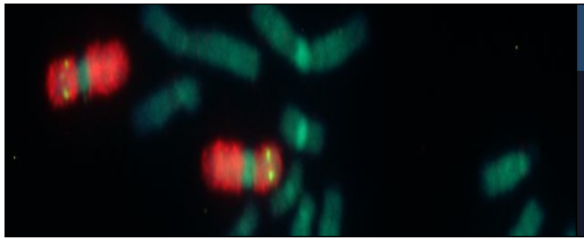

19: Bowtie2 on data 13 and data 12: aligned reads (BAM)
13: aln2.fastq.gz
12: aln1.fastq.gz

# Databases

NCBI Resources How To Sign in to NCBI

Gene Gene BRCA2 Search

Advanced Help



## Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

### Using Gene

- Gene Quick Start
- FAQ
- Download/FTP
- RefSeq Mailing List
- Gene News
- Factsheet

### Gene Tools

- Submit GeneRIFs
- Submit Correction
- Statistics
- BLAST
- Genome Workbench
- Splign

### Other Resources

- HomoloGene
- OMIM
- RefSeq
- RefSeqGene
- UniGene
- Protein Clusters

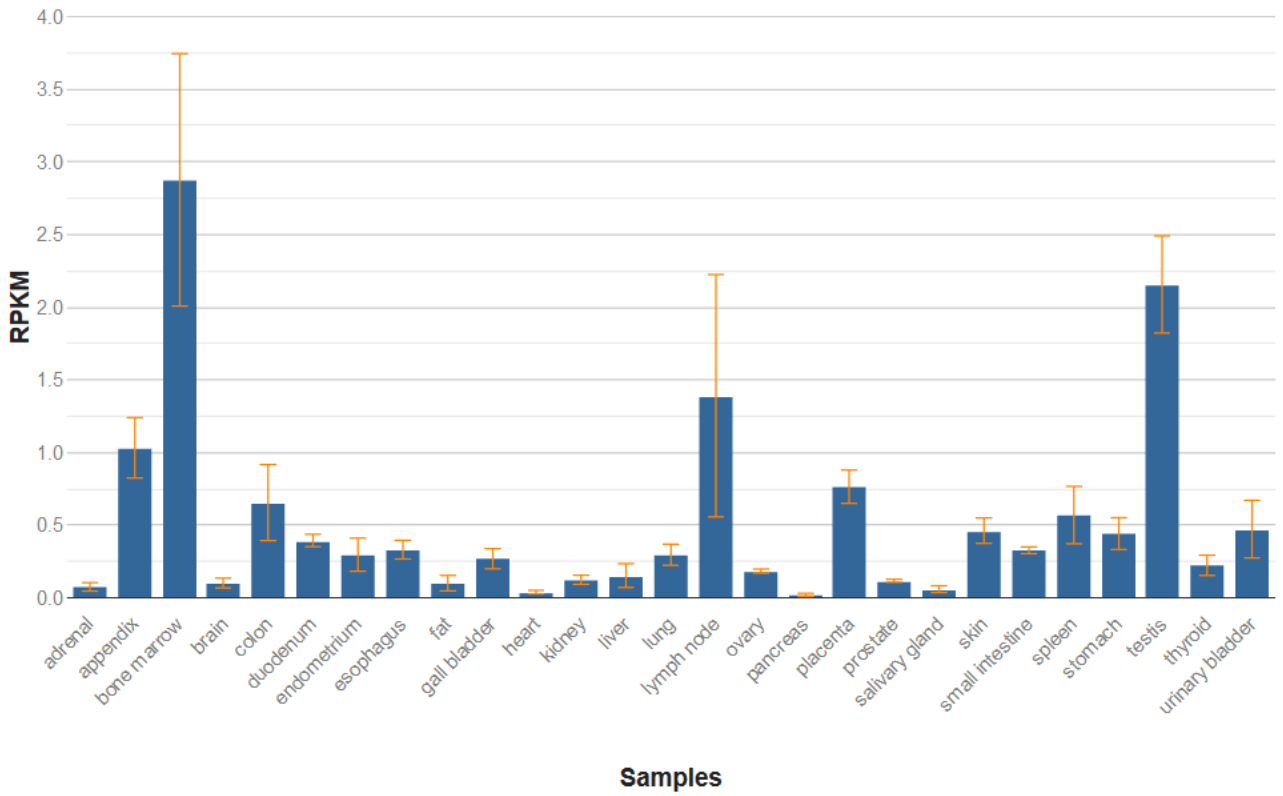
# Databases

## Summary

| | |
|---|---|
| **Official Symbol** | BRCA2 provided by HGNC |
| **Official Full Name** | BRCA2 DNA repair associated provided by HGNC |
| **Primary source** | HGNC:HGNC:1101 |
| **See related** | Ensembl:ENSG00000139618 MIM:600185 |
| **Gene type** | protein coding |
| **RefSeq status** | REVIEWED |
| **Organism** | Homo sapiens |
| **Lineage** | Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo |
| **Also known as** | FAD; FACD; FAD1; GLM3; BRCC2; FANCD; PNCA2; FANCD1; XRCC11; BROVCA2 |
| **Summary** | Inherited mutations in BRCA1 and this gene, BRCA2, confer increased lifetime risk of developing breast or ovarian cancer. Both BRCA1 and BRCA2 are involved in maintenance of genome stability, specifically the homologous recombination pathway for double-strand DNA repair. The BRCA2 protein contains several copies of a 70 aa motif called the BRC motif, and these motifs mediate binding to the RAD51 recombinase which functions in DNA repair. BRCA2 is considered a tumor suppressor gene, as tumors with BRCA2 mutations generally exhibit loss of heterozygosity (LOH) of the wild-type allele. [provided by RefSeq, Dec 2008] |
| **Expression** | Broad expression in bone marrow (RPKM 2.9), testis (RPKM 2.2) and 17 other tissues See more |
| **Orthologs** | mouse all |

# Databases

**Expression**




**Genomic context**

**Bibliography**

**Variation**

# Databases


BIER
Collaborative Spanish Variant Server
BIER
CSVS
Start
Overview
Welcome to the Collaborative Spanish Variant Server. CSVS was created to provide information about the variability of the Spanish population to the scientific/medical community. It is useful for filtering polymorphisms and local variations in the process of prioritizing candidate disease genes. CSVS currently stores information on 1644 unrelated Spanish individuals. We accept submissions from WES or WGS.
Supported by
ciberer BIER
ciberer
Fundación Progreso y Salud
CONSEJERÍA DE SALUD
elixir
CSVS: created by Clinical Bioinformatics Area
Fundación Progreso y Salud
2015-2017

# Databases

**Position**

Chromosomal location:

1:1-1000000,2:1-1000000

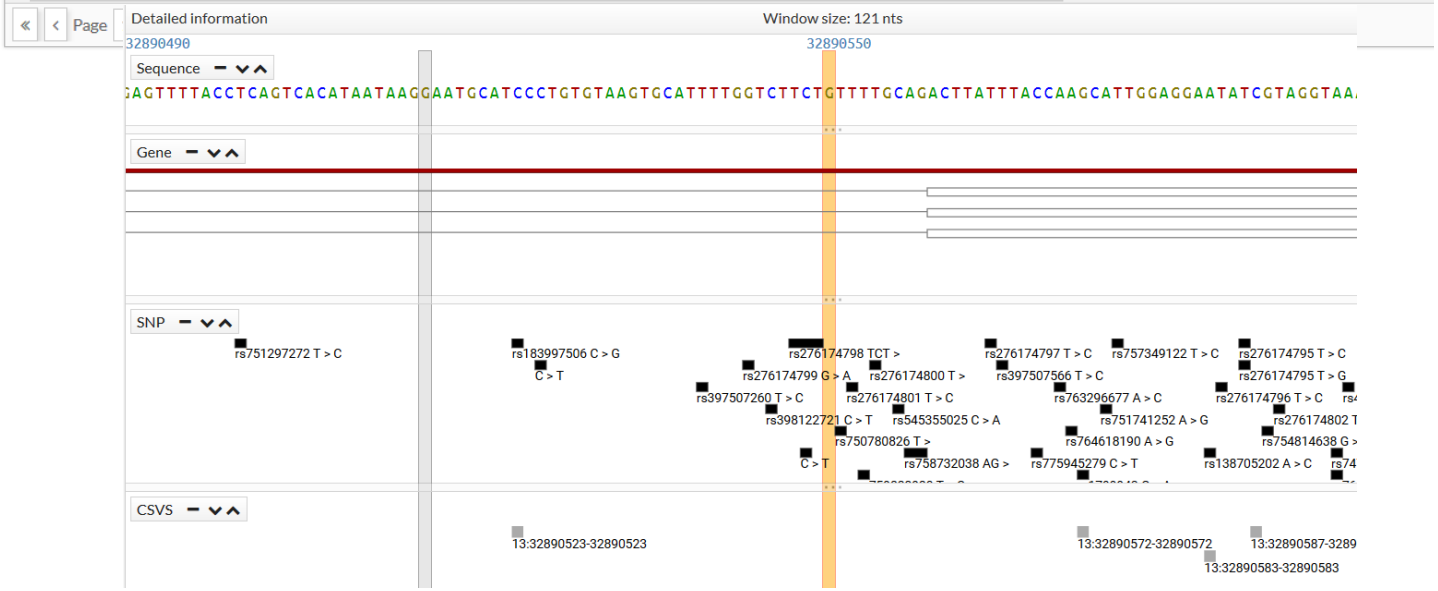
Gene:

BRCA2

**Subpopulations**

- MGP (267 healthy controls)
- IBS (107 Spanish individuals from 1000genomes)
- Healthy controls
- I Certain infectious and parasitic diseases
- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
- V Mental and behavioural

| Chr | Position | Alleles | Gene | Id | MAF | | | | | | | 1000G MAF (phase 3) | | ExAC | ESP 6500 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Genotypes | | | | Freq. | | | ALL | EUR | ALL | ALL | Eur. An |
| | | | | | 0/0 | 0/1 | 1/1 | ./. | 0 Freq | 1 Freq | MAF | | | | | |
| 13 | 32889792 | A>G | ZAR1L,BRCA2 | rs206118 | 1519 | 5 | 2 | 56 | 0.997 | 0.003 | 0.003 | 0.149 | 0.193 | . | . | . |
| 13 | 32889968 | G>A | ZAR1L,BRCA2 | rs206119 | 1581 | 0 | 1 | 0 | 0.999 | 0.001 | 0.001 | 0.259 | 0.216 | . | . | . |
| 13 | 32890404 | A>G | BRCA2,ZAR1L | . | 1581 | 1 | 0 | 0 | 1 | 0 | 0 | . | . | . | . | . |
| 13 | 32890523 | C>G | ZAR1L,BRCA2 | rs183997506 | 1579 | 3 | 0 | 0 | 0.999 | 0.001 | 0.001 | 0.000 | . | . | . | . |
| 13 | 32890572 | G>A | ZAR1L,BRCA2 | rs1799943 | 920 | 547 | 100 | 15 | 0.762 | 0.238 | 0.238 | 0.209 | 0.216 | 0.247 | 0.209 | 0.263 |
| 13 | 32890583 | A>C | BRCA2,ZAR1L | rs138705202 | 1581 | 1 | 0 | 0 | 1 | 0 | 0 | 0.001 | . | 0.000 | 0.000 | . |
| 13 | 32890587 | C>T | BRCA2,ZAR1L | rs76874770 | 1643 | 1 | 0 | 0 | 1 | 0 | 0 | 0.004 | . | 0.002 | 0.006 | 0.000 |
| 13 | 32890629 | T>A | BRCA2 | . | 1643 | 0 | 0 | 1 | 1 | 0 | 0 | . | . | . | . | . |
| 13 | 32890726 | T>G | BRCA2,ZAR1L | rs11571574 | 1436 | 1 | 0 | 145 | 1 | 0 | 0 | 0.003 | 0.002 | . | 0.004 | 0.004 |
| 13 | 32893197 | AT>A,ATT | BRCA2,ZAR1L | . | 1562 | 14 | 0 | 6 | 0.996 | 0.004 | 0.004 | . | . | . | . | . |



# Databases

**dbSNP** Short Genetic Variations

Search for rs | Search

Example: rs268

## Reference SNP (rs) Report

Download

Switch to classic site

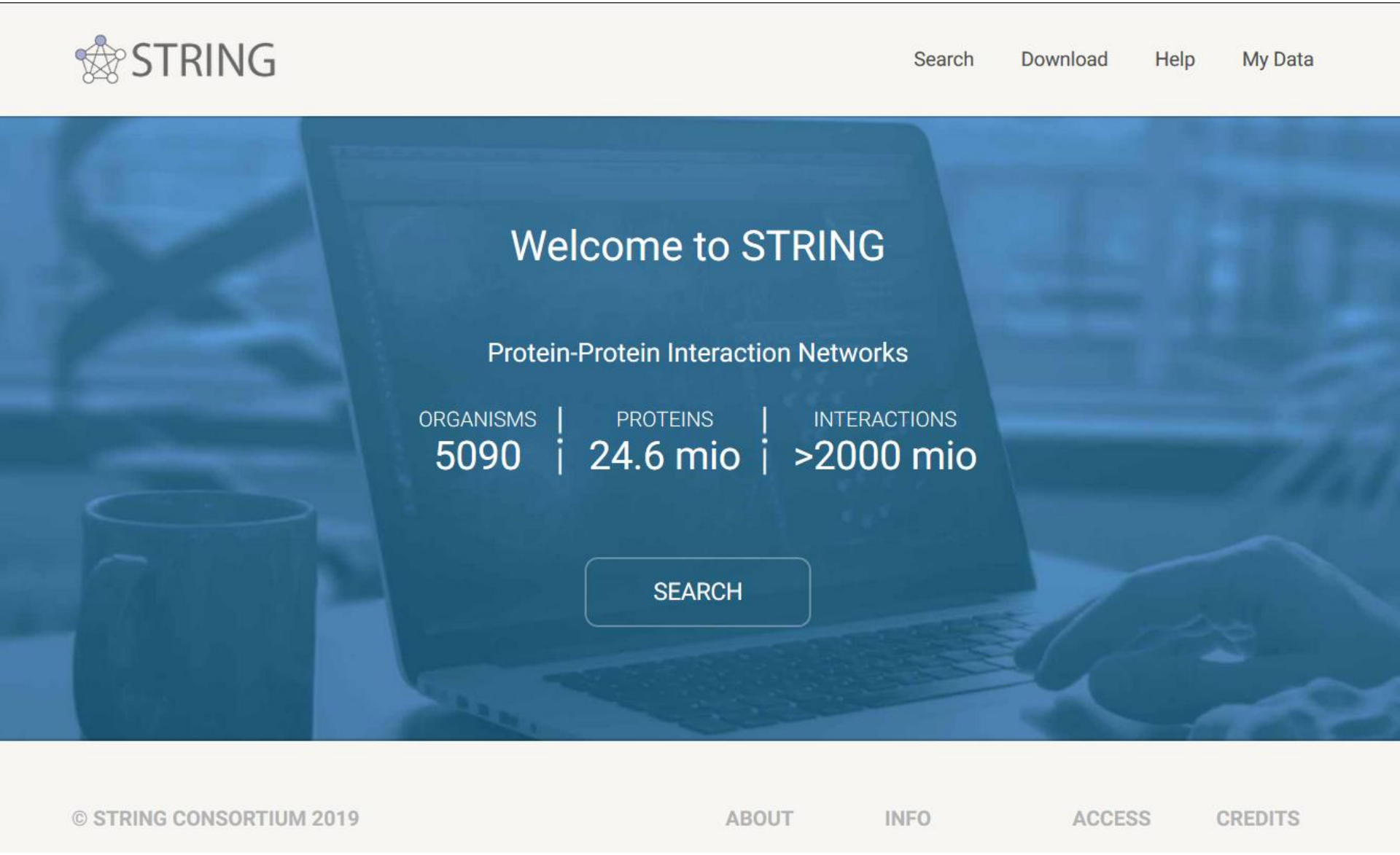
**rs757349122**

**Current Build** 152
**Released** October 2, 2018

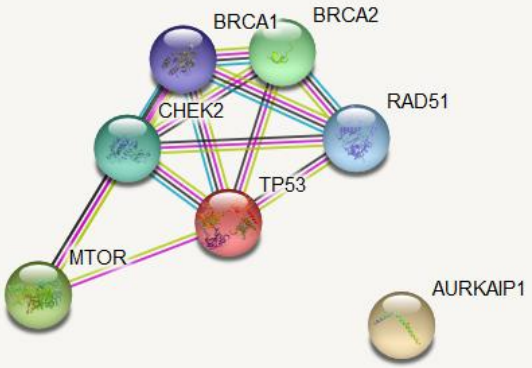
| | | | |
|---|---|---|---|
| **Organism** | *Homo sapiens* | **Clinical Significance** | Reported in ClinVar |
| **Position** | chr13:32316438 (GRCh38.p12) | **Gene : Consequence** | BRCA2 : 5 Prime UTR Variant |
| **Alleles** | T>C | **Publications** | 0 citations |
| **Variation Type** | SNV Single Nucleotide Variation | **Genomic View** | See rs on genome |
| **Frequency** | C=0.00001 (2/245764, GnomAD)<br>C=0.00001 (1/118930, ExAC) | | |

Variant Details | **Clinical Significance** | Frequency | Aliases | Submissions | History | Publications

# Databases


STRING
Search
Download
Help
My Data
Welcome to STRING
Protein-Protein Interaction Networks
ORGANISMS
5090
PROTEINS
24.6 mio
INTERACTIONS
>2000 mio
SEARCH
© STRING CONSORTIUM 2019
ABOUT
INFO
ACCESS
CREDITS

# Databases




Viewers › | Legend ⌄ | Settings › | Σ Analysis › | Exports › | Clusters › | More | Less

**Nodes:**

Network nodes represent proteins

*splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.*

Node Color

*colored nodes:*
*query proteins and first shell of interactors*

*white nodes:*
*second shell of interactors*

Node Content

*empty nodes:*
*proteins of unknown 3D structure*

*filled nodes:*
*some 3D structure is known or predicted*

**Edges:**

Edges represent protein-protein associations

*associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.*

Known Interactions

*from curated databases*

*experimentally determined*

Predicted Interactions

*gene neighborhood*

*gene fusions*

*gene co-occurrence*

Others

*textmining*

*co-expression*

*protein homology*

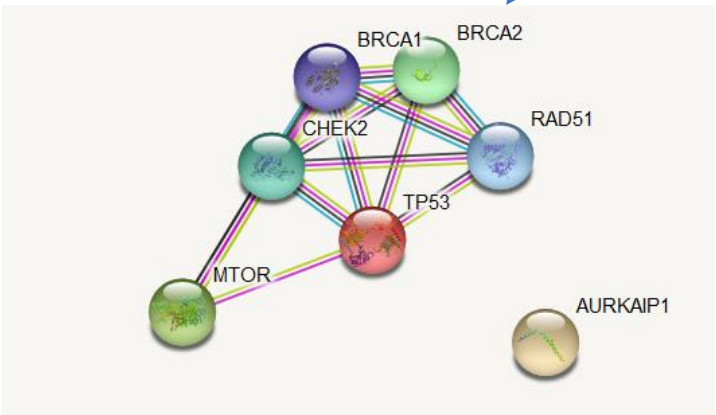
# Databases

**BRCA2**

**Information**
Breast cancer type 2 susceptibility protein; Involved in double-strand break repair and/or homologous recombination. Binds RAD51 and potentiates recombinational DNA repair by promoting assembly of RAD51 onto single-stranded DNA (ssDNA). Acts by targeting RAD51 to ssDNA over double-stranded DNA, enabling RAD51 to displace replication protein-A (RPA) from ssDNA and stabilizing RAD51-ssDNA filaments by blocking ATP hydrolysis. Part of a PALB2-scaffolded HR complex containing RAD51C and which is thought to play a role in DNA repair by HR. May participate in S phase checkpoint activation. B [...]

Identifier: ENSP00000369497, BRCA2
Organism: Homo sapiens

**Actions**
- re-center network on this node
- remove this node from input nodes
- show protein sequence
- homologs among STRING organisms
- Pathways, Functions, Resources (GeneCards)

# Databases

**Interaction**

RAD51 *[ENSP00000372088]*

DNA repair protein RAD51 homolog 1; Fanconi anemia complementation groups

↔

BRCA2 *[ENSP00000369497]*

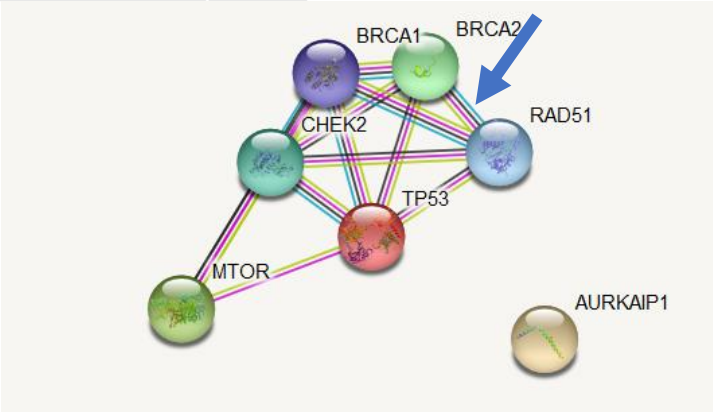
Breast cancer type 2 susceptibility protein; Involved in double-strand break repair and/or homologous recombination. Binds RAD51 and potentiates recombinational DNA repair by promoting assembly of RAD51 onto single-stranded DNA (ssDNA). Acts by targeting RAD51 to ssDNA over double-stranded DNA, enabling RAD51 to displace replication protein-A (RPA) from ssDNA and stabilizing RAD51-ssDNA filaments by blocking ATP hydrolysis. Part of a PALB2-scaffolded HR complex containing RAD51C and which is thought to play a role in DNA repair by HR. May participate in S phase checkpoint activation. B [...]

**Evidence suggesting a functional link:**

| | | |
|---|---|---|
| Neighborhood in the Genome: | none / insignificant. | |
| Gene Fusions: | none / insignificant. | |
| Cooccurence Across Genomes: | none / insignificant. | |
| Co-Expression: | yes (score 0.137). In addition, putative homologs are coexpressed in other species (score 0.088). | Show |
| Experimental/Biochemical Data: | yes (score 0.689). In addition, putative homologs were found interacting in other species (score 0.741). | Show |
| Association in Curated Databases: | yes (score 0.900). | Show |
| Co-Mentioned in PubMed Abstracts: | yes (score 0.956). In addition, putative homologs are mentioned together in other species (score 0.487). | Show |
| Combined Score: | 0.999 | |

**Predictions for specific actions:**

| | | |
|---|---|---|
| Activation: | yes (score: 0.741) | Show |
| Binding: | yes (score: 0.972) | Show |
| Catalysis: | yes (score: 0.900) | Show |
| Post-translational modification: | yes (score: 0.637) | Show |
| Reaction: | yes (score: 0.964) | Show |



# Databases

**Functional enrichments in your network**

**Biological Process (GO)**

| *GO-term* | *description* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| GO:0006978 | DNA damage response, signal transduction by p53 class m... | 4 of 17 | 4.07e-08 |
| GO:0010212 | response to ionizing radiation | 5 of 149 | 2.33e-07 |
| GO:0009314 | response to radiation | 6 of 425 | 2.33e-07 |
| GO:0006281 | DNA repair | 6 of 491 | 4.25e-07 |
| GO:0051726 | regulation of cell cycle | 7 of 1129 | 4.33e-07 |
| | | | *(more ...)* |

**Molecular Function (GO)**

| *GO-term* | *description* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| GO:0019899 | enzyme binding | 6 of 2197 | 0.0018 |
| GO:0140296 | general transcription initiation factor binding | 2 of 39 | 0.0064 |
| GO:0042802 | identical protein binding | 5 of 1754 | 0.0064 |
| GO:0031625 | ubiquitin protein ligase binding | 3 of 311 | 0.0064 |
| GO:0003684 | damaged DNA binding | 2 of 64 | 0.0064 |
| | | | *(more ...)* |

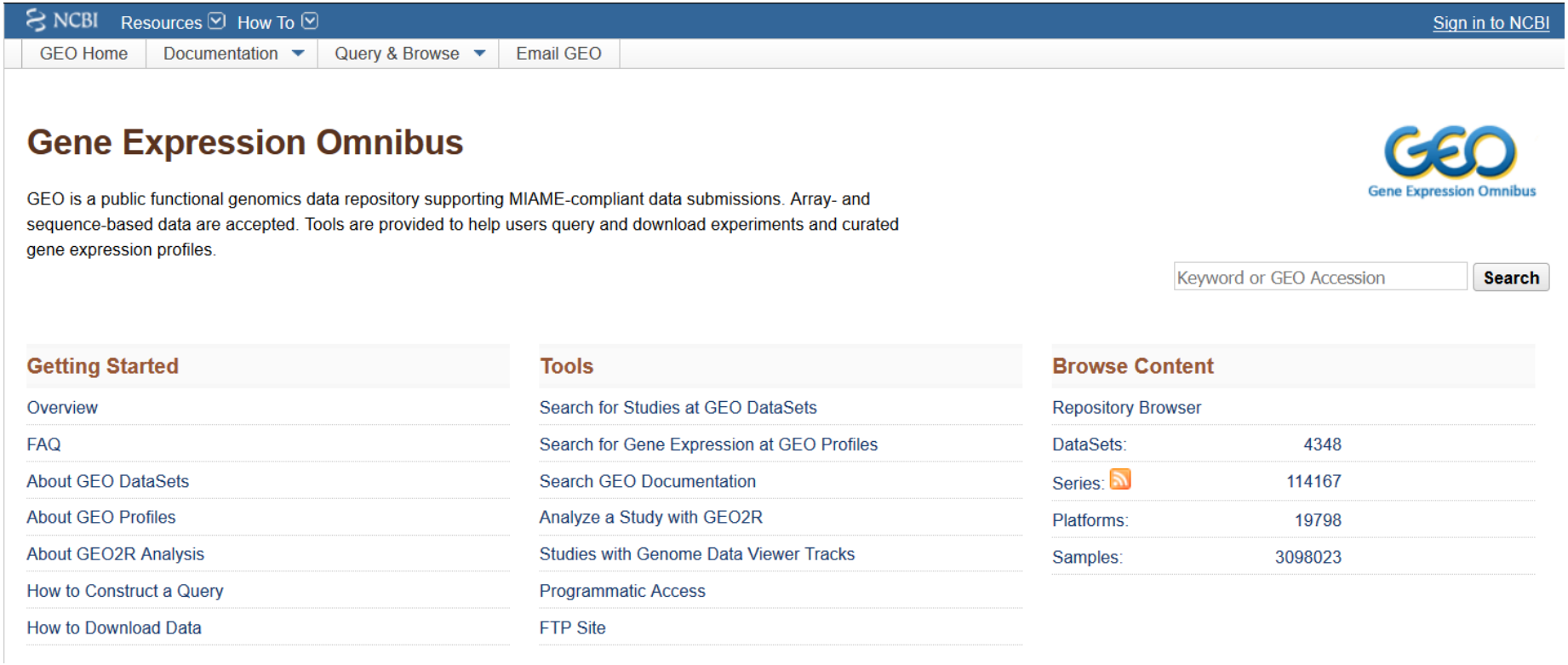
**Cellular Component (GO)**

| *GO-term* | *description* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| GO:0016605 | PML body | 4 of 98 | 2.79e-06 |
| GO:0000800 | lateral element | 3 of 16 | 2.79e-06 |
| GO:0005654 | nucleoplasm | 7 of 3446 | 0.00010 |
| GO:0044454 | nuclear chromosome part | 4 of 480 | 0.00020 |
| GO:0000784 | nuclear chromosome, telomeric region | 2 of 51 | 0.0010 |
| | | | *(more ...)* |

**Reference publications**

| *publication* | *(year) title* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| PMID:26599019 | (2015) Metformin Radiosensitizes p53-Deficient Colorectal ... | 6 of 15 | 8.61e-13 |
| PMID:29113422 | (2017) Novel DNA targeted therapies for head and neck can... | 6 of 34 | 1.54e-11 |
| PMID:28825622 | (2017) FANCD2 and DNA Damage. | 6 of 32 | 1.54e-11 |
| PMID:24063014 | (2013) New hypothesis on pathogenesis of ovarian cancer l... | 6 of 30 | 1.54e-11 |
| PMID:23137030 | (2012) DNA damage signaling assessed in individual cells i... | 6 of 31 | 1.54e-11 |

# Databases


NCBI
Resources
How To
Sign in to NCBI
GEO Home
Documentation
Query & Browse
Email GEO
Gene Expression Omnibus
GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.
Gene Expression Omnibus
Keyword or GEO Accession
Search
Getting Started
Overview
FAQ
About GEO DataSets
About GEO Profiles
About GEO2R Analysis
How to Construct a Query
How to Download Data
Tools
Search for Studies at GEO DataSets
Search for Gene Expression at GEO Profiles
Search GEO Documentation
Analyze a Study with GEO2R
Studies with Genome Data Viewer Tracks
Programmatic Access
FTP Site
Browse Content
Repository Browser
DataSets: 4348
Series: 114167
Platforms: 19798
Samples: 3098023

# Databases

GEO DataSets

☐ **BRCA2** abrogation triggers innate immune responses potentiated by treatment with PARP inhibitors

5. (Submitter supplied) Heterozygous germline mutations in **BRCA2** predispose to breast and ovarian cancer. Contrary to non-cancerous cells, where **BRCA2** deletion causes cell cycle arrest or cell death, **BRCA2** inactivation in tumors is associated with uncontrolled cell proliferation. We set out to investigate this conundrum by exploring modalities of cell adaptation to loss of **BRCA2** and focused on genome-wide transcriptome alterations. more...

Organism: **Homo sapiens**
Type: Expression profiling by high throughput sequencing
Platform: GPL20301 48 Samples
Download data: TXT
Series Accession: GSE123631 ID: 200123631
SRA Run Selector

☐ Breast tumor subtypes correlate with prognosis

6. (Submitter supplied) To advance in our understanding of the biological pathways involved in **breast cancer** tumor progression we have analyzed a set of breast tumor biopsies in order to identify the genomic pathways in which tumor may develop. With this objective, a cDNA microarray platform containing 800 genes was constructed. These genes were chosen because they are in several representatives signaling pathways, namely estrogen and progesterone receptor related pathways, cell cycle, DNA repair, chromatin remodeling, cell proliferation, apoptosis, cell adhesion, cell invasion and angiogenesis. more...

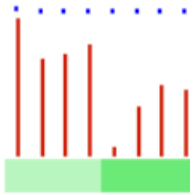
Organism: **Homo sapiens**
Type: Expression profiling by array
Platform: GPL5953 111 Samples
Download data: GPR
Series Accession: GSE18908 ID: 200018908
Analyze with GEO2R

# Databases

## GEO Profiles

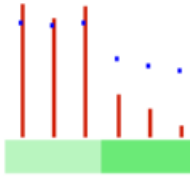
☐ 11. Bccip - Niacin-bound chromium effect on the subcutaneous fat tissue of a model for type 2 diabetes with obesity

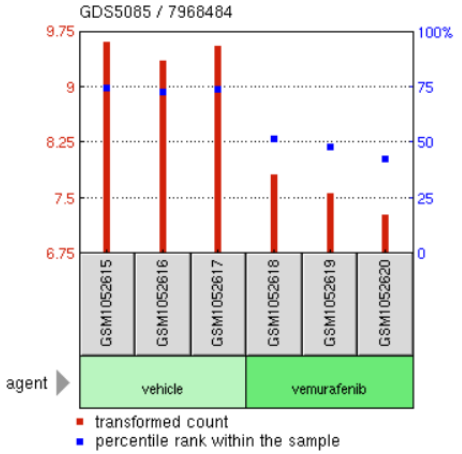
Annotation: Bccip, **BRCA2** and CDKN1A interacting protein
Organism: Mus musculus
Reporter: GPL1261, 1448542_at (ID_REF), GDS2605, 66165 (Gene ID), NM_025392
DataSet type: Expression profiling by array, count, 8 samples
ID: 35750542
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors



☐ 12. **BRCA2** - Oncogenic BRAF harboring melanoma cell line response to BRAF inhibition

Annotation: **BRCA2**, **BRCA2**, DNA repair associated
Organism: Homo sapiens
Reporter: GPL6244, 7968484 (ID_REF), GDS5085, NM_000059, DQ897648, U43746, chr13:32889617-32973809 (SPOT ID)
DataSet type: Expression profiling by array, transformed count, 6 samples
ID: 112035057
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors





# THANKS!

# BAM/SAM

@HD – Header

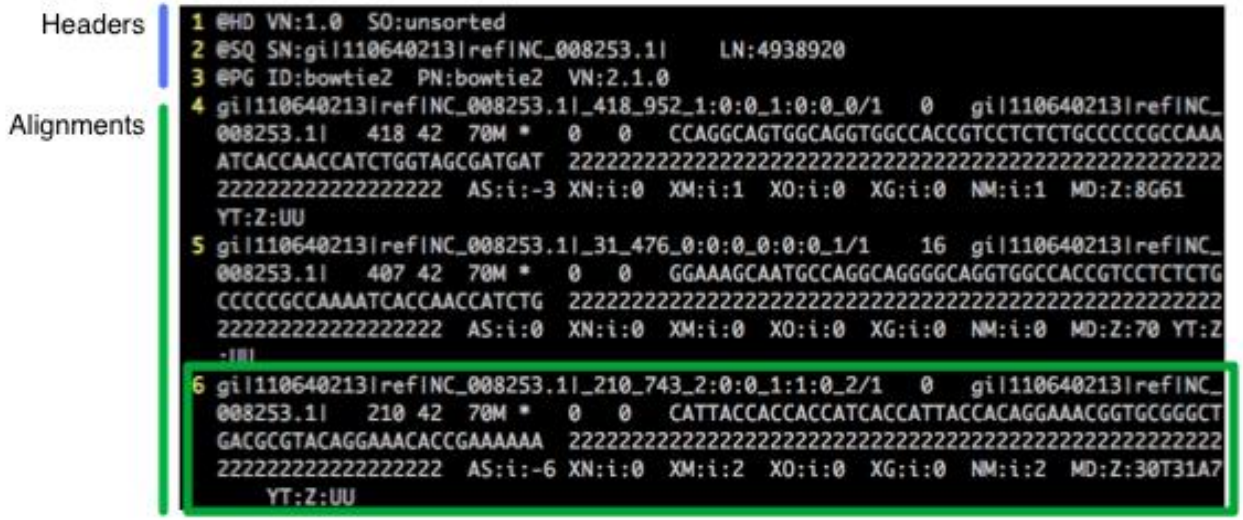
- VN:1.0 – SAM format version
- SO: unsorted – sorting

@SQ – Reference sequence dictionary

- SN: … – Sequence name
- LN: … – Length

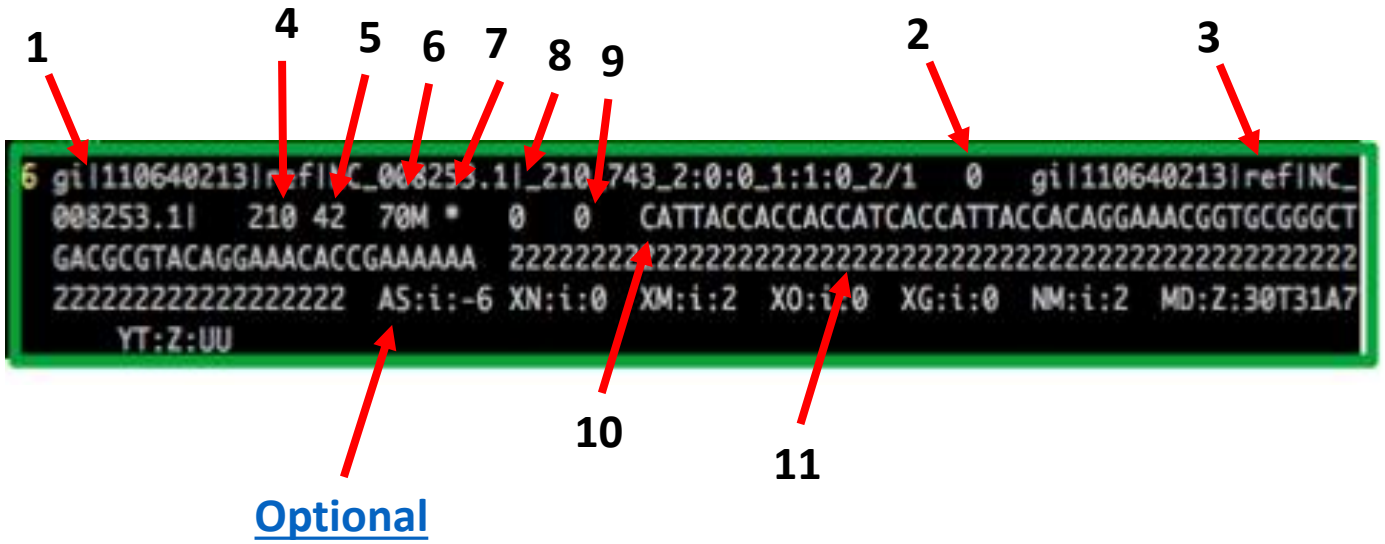
@PG – Program

- ID: … - ID of the program
- PN: … - Program name
- VN: … - Version



# BAM/SAM

1. Query Name – Read name
2. FLAG – Alingment properties
3. RNAME – Reference sequence name
4. POS – Alingment start position
5. MAPQ – Alingment quality
6. CIGAR – Alignment operations
7. RNEXT – Mate sequence name
8. PNEXT – Mate alignment position
9. TLEN – Template length
10. SEQ – Read sequence
11. Qual – Read quality (FASTQ like)



# VCF

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G    A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T    A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T    .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC  G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```