



Session II. Bioconductor

Armando Reyes-Palomares

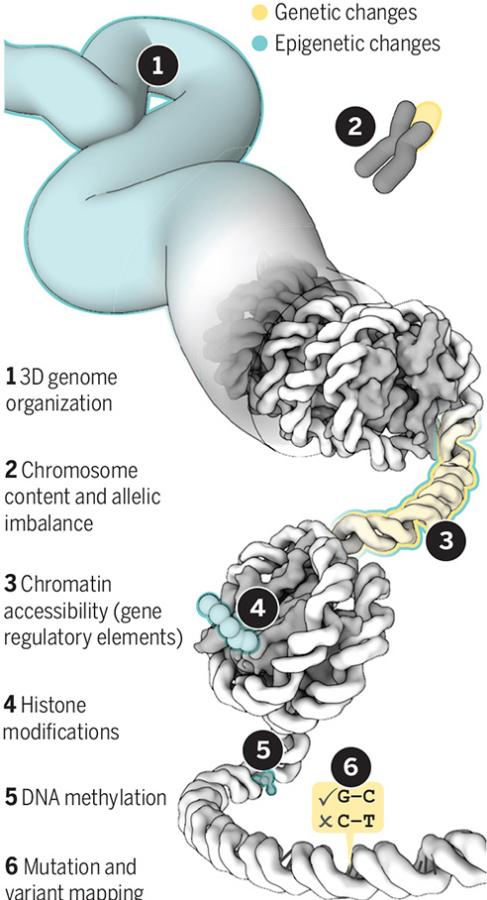
Department Biochemistry and Molecular Biology

Complutense University of Madrid

armandorp@ucm.es

Sequence and function of the cancer genome

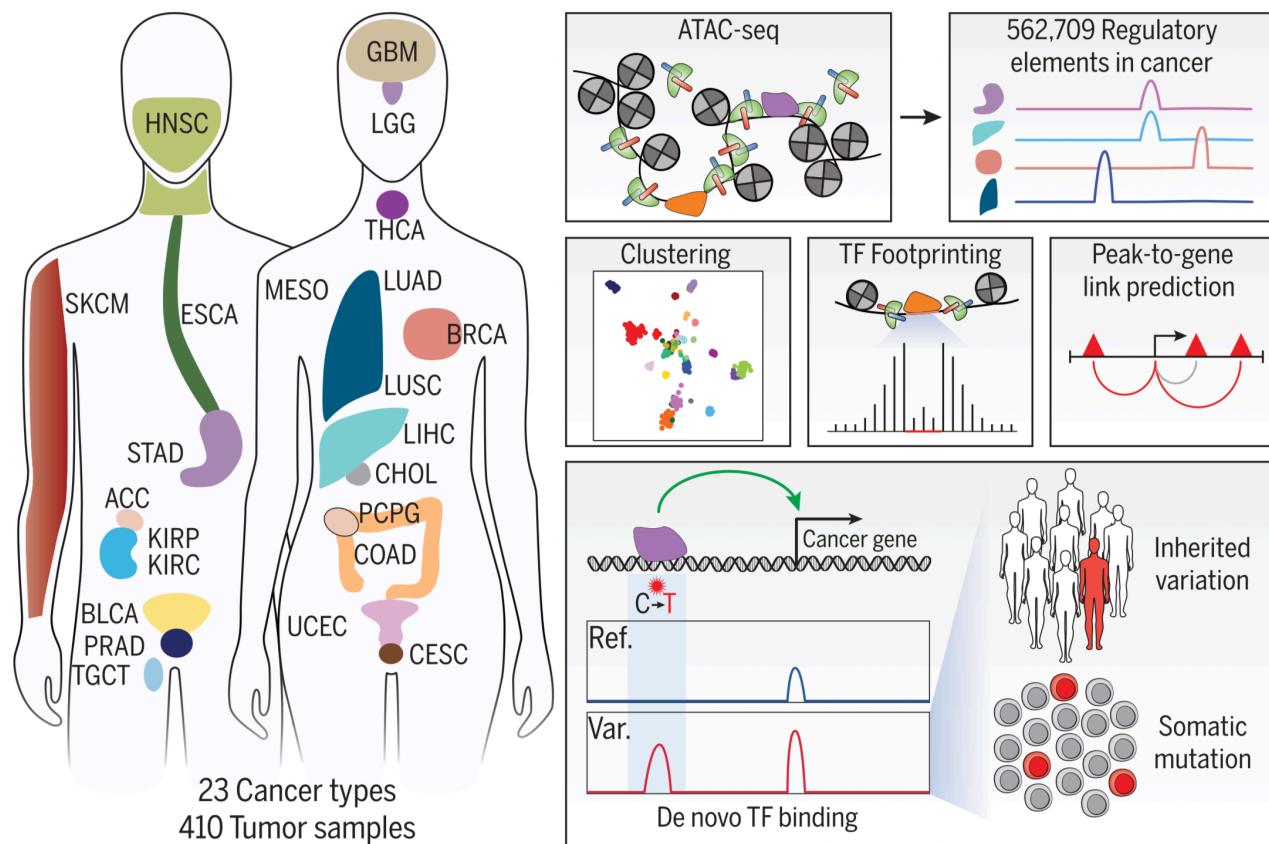
Multiomic mapping and comparison between genetic and epigenetic features are required for mechanistic understanding of cancer and for providing a "fingerprint" of the tumor. Multiomic analyses are likely to be important for cancer diagnosis and prediction of outcome, as well as for guiding treatment decisions and drug development.



RESEARCH ARTICLE SUMMARY

CANCER

The chromatin accessibility landscape of primary human cancers



Practical Sessions

1. Introduction Computational Genomics
2. Introduction to R part I
3. Introduction to R part II
4. Bioconductor: GenomicRanges and BiomaRt

Summarizing (R introduction)

1. R and RStudio
2. Operate in R
3. Make **R objects** and explore **data structures**
4. Useful and common commands in R
5. R plots & graphs
6. Merge function in R

Table of contents

BioConductor:

- BioMart
- GenomicRanges



Bioconductor (BioC) www.bioconductor.org

- Open source, free.
- Search for BioC packages.
 - Explore 3.10 version
 - DESeq2
- Courses, workflows and teaching.
- Strong efforts in documentation.

The screenshot shows the Bioconductor website (bioconductor.org) displayed in a web browser. The header features the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. The main content area includes a section about the EuroBioC 2018 meeting, an 'About Bioconductor' summary, and four large callout boxes: 'Install >', 'Learn >', 'Use >', and 'Develop >'. Each callout box lists specific resources or links related to its category.

EuroBioC 2018
The European Bioconductor meeting is on December 6 and 7, 2018, at the Technical University of Munich, Germany. The meeting is for biologists, bioinformaticians, statisticians, programmers and software engineers. The meeting aims to foster the exchange of technical expertise while keeping contributors up to speed with the latest developments in Bioconductor.

About Bioconductor
Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [1560 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

Install >
Get started with Bioconductor

- [Install Bioconductor](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn >
Master Bioconductor tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use >
Create bioinformatic solutions with Bioconductor

- [Software, Annotation, and Experiment packages](#)
- [Amazon Machine Image](#)
- [Latest release announcement](#)
- [Support site](#)

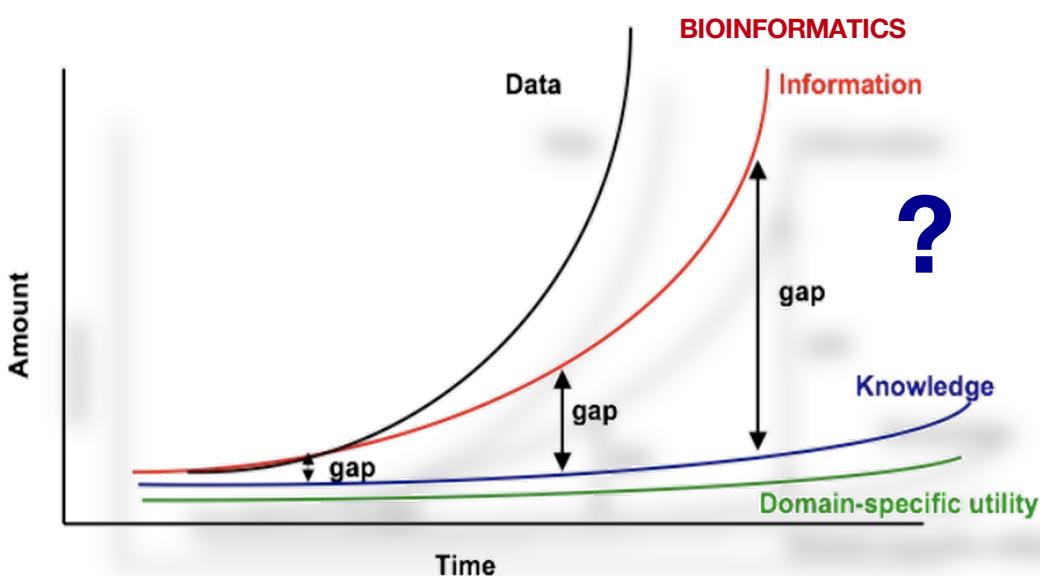
Develop >
Contribute to Bioconductor

- [Developer resources](#)
- [Use Bioc 'devel'](#)
- [Devel Software, Annotation and Experiment packages](#)
- [Package guidelines](#)
- [New package submission](#)
- [Git source control](#)
- [Build reports](#)

GENOMICS → MOLECULAR BIOLOGY

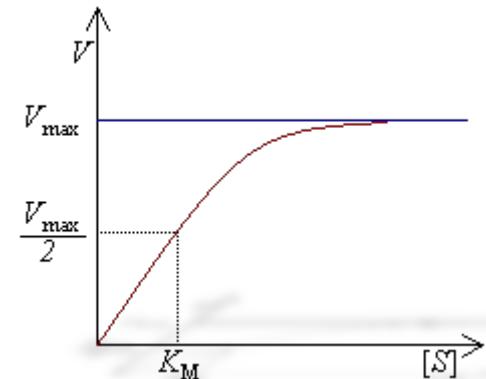
Revolutionary milestones: Recombinant DNA technology and NGS

DATA -> INFORMATION -> KNOWLEDGE -> APPLICATION



Analogy:

Data processing are saturated



Big Data problem:

- Multiple experimental data sets
 - Growing Imbalance

Repository of Data Resources at EMBL-EBI

Genes, genomes & variation

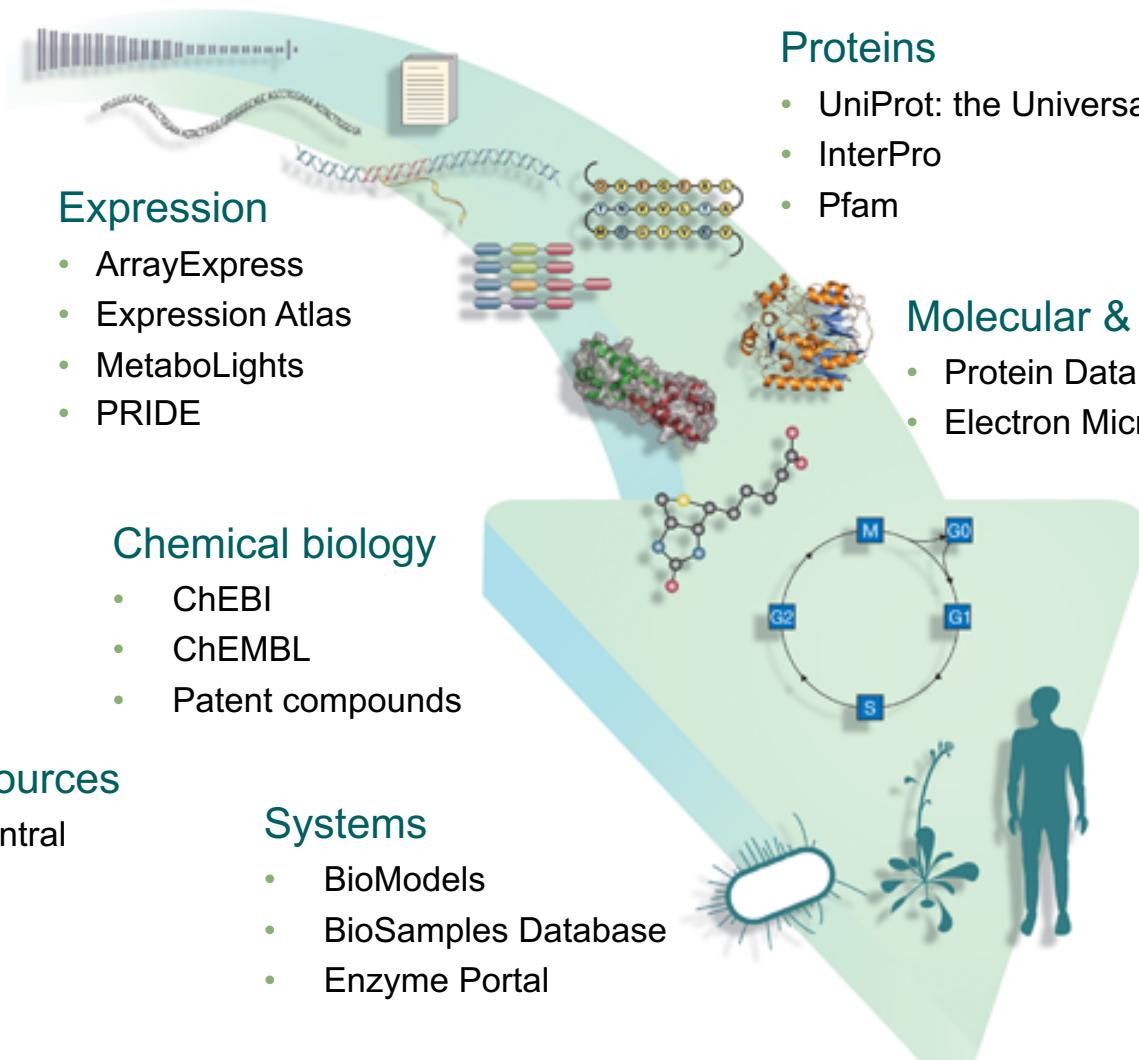
- European Nucleotide Archive (ENA)
- EBI Metagenomics

Ensembl

- Ensembl Genomes
- European Genome–phenome Archive
- Non-redundant patent sequence databases

Cross-domain resources

- Europe PubMed Central
- Gene Ontology



BioMart Data mining tool www.biomart.org

Screenshot of the BioMart website (www.biomart.org) showing the homepage layout and features.

The page includes a navigation bar with links to Home, Tools, Community, Publications, News, Credits, Documentation, Version 0.7, and Contact. A sidebar on the left highlights a "SPECIAL ISSUE DEDICATED TO BIOMART" from Nature Methods.

The main content area features a "BioMart" header, a news banner about updates to ensembl, ensembl genomes, and uniprot, and a "Download" button for version 0.9.

A central section describes BioMart as a community-driven project for unified access to research data. It highlights four key features:

- BROWSE DATA (highlighted with a red box)
- ID CONVERSION
- SEQUENCE RETRIEVAL
- ENRICHMENT ANALYSIS

To the right, there's a "JOIN OUR COMMUNITY" section with three steps:

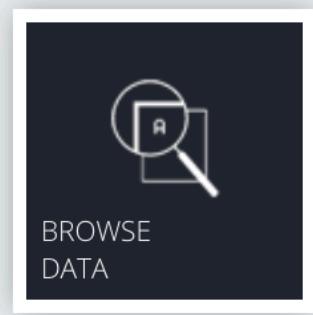
- Set up your own data source with a click of a button
- Expose your data to a world wide scientific community through BioMart Portal.
- Federate your local data with data from other community members

Buttons for "DOWNLOAD OUR SOFTWARE TO JOIN" and "DOWNLOAD NOW" are also present.

At the bottom, a yellow box contains a quote from Nature Methods and a "BROWSE DATA" link. To the right is a world map showing the distribution of BioMart databases across continents.

BioMart www.biomart.org

1



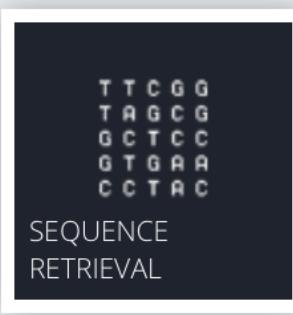
BROWSE
DATA

2



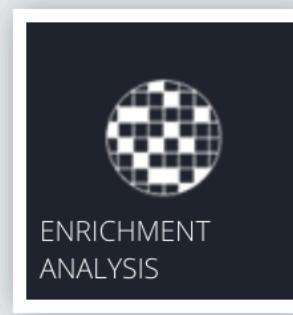
ID
CONVERSION

3



SEQUENCE
RETRIEVAL

4



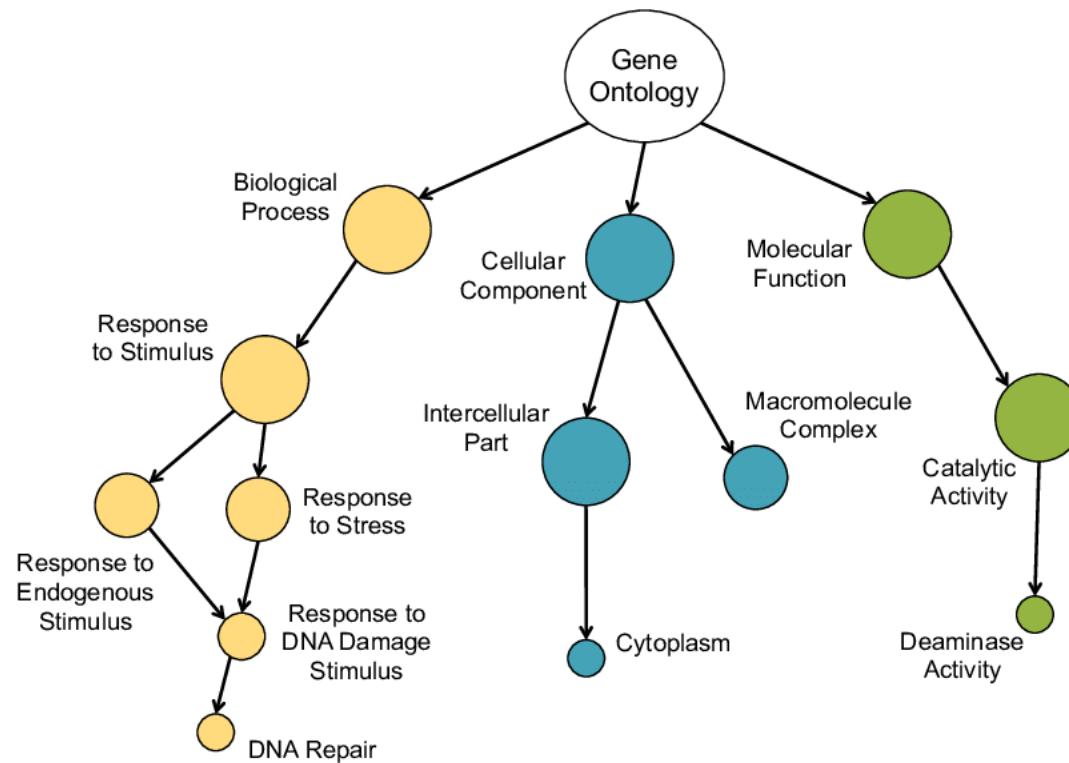
ENRICHMENT
ANALYSIS

BioMart integrated in ENSEMBL

The screenshot shows the Ensembl BioMart interface. At the top left is the Ensembl logo. The top right features a "Login/Register" link, a search bar with a dropdown menu set to "Search all species..." and a magnifying glass icon, and a "New" button. Below the header is a navigation bar with links: BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. The main content area has tabs for New, Count, and Results. A toolbar at the top of the main area includes URL, XML, Perl, and Help buttons. On the left, a sidebar titled "Dataset" shows "[None selected]" and a dropdown menu labeled "- CHOOSE DATABASE -". The main body of the page is currently empty.

3 Main Integrated Mart Databases

- **Ensembl Genes:** This mart contains the Ensembl gene set and allows you to retrieve Ensembl genes, transcripts and proteins as well as external references, microarrays, protein domains, structure, sequences, variants (*only variants mapped to Ensembl Transcripts*) and homology data.
- **Ensembl Variation:** This mart allows you to retrieve germline and somatic variants as well as germline and somatic structural variants. This mart also contains variants' phenotypes, citations, synonyms, consequences and flanking sequences; you can also retrieve Ensembl genes, transcripts, regulatory and motif features mapped to variants.
- **Ensembl Regulation:** This mart allows you to retrieve regulatory features, evidence and segments, miRNA target regions, binding motifs and other regulatory regions.

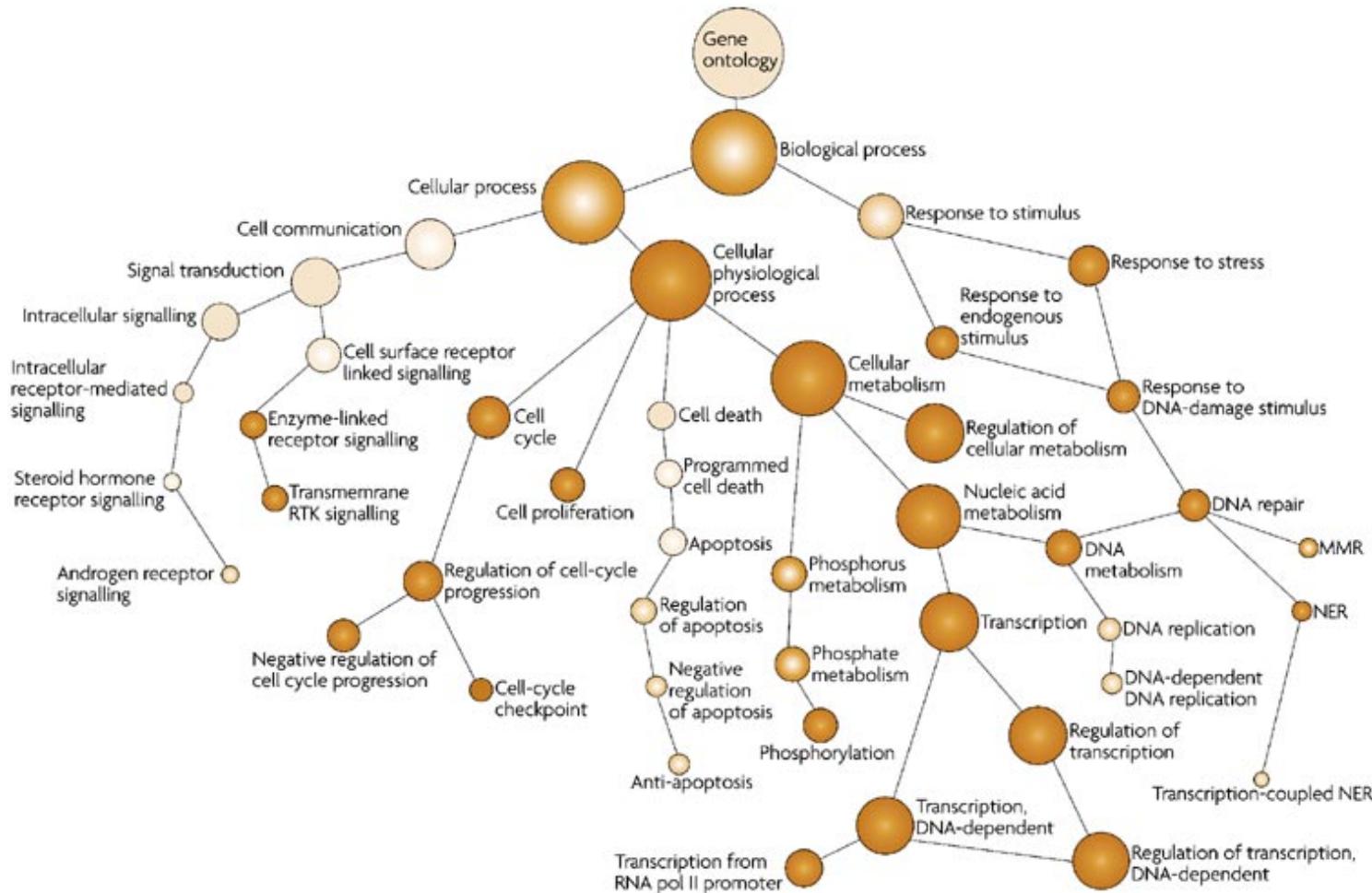




GENEONTOLOGY

Unifying Biology

geneontology.org



Types of inputs for genomic features:

1. Sequence → it's the reference
2. Genomic Coordinates → *Genome Reference Assembly*

```
>Homo sapiens, gene_ID_24349304323
GGCAGATTCCCCCTAGACCCGCCGCACCATGGTCAGGCATGC
CCCTCCTCATCGCTGGGACAGCCCAGAGGGTATAAACAGTGC
TGGAGGCTGGCGGGGCAGGCCAGCTGAGTCCTGAGCAGCAGC
CCAGCGCAGCCAC
```

Chromosome	Start	End
chr12	48366748	48398285

3. Database IDs → Useful for integrate information

Orthologs		
Species	Human	Mouse
Entrez	1280	12824
Ensembl	ENSG00000139219	ENSMUSG00000022483
UniProt	P02458	P28481

Some issues due to lack of consensus:

ENSEMBL ID	Entrez ID
ENSG000000XXXX	→ N.A.
ENSG000000YYYY	} → 14050
ENSG000000ZZZZ	

GenomicRanges Package

```
> gr = exons(TxDb.Hsapiens.UCSC.hg19.knownGene); gr
```

GRanges with 289969 ranges and 1 metadata column:

	seqnames	ranges	strand	exon_id
[1]	chr1	[11874, 12227]	+	1
[2]	chr1	[12595, 12721]	+	2
[3]	chr1	[12613, 12721]	+	3
...
[289967]	chrY	[59358329, 59359508]	-	277748
[289968]	chrY	[59360007, 59360115]	-	277749
[289969]	chrY	[59360501, 59360854]	-	277750

seqlengths:

	chr1	chr2 ...	chrUn_g1000249
	249250621	243199373 ...	38502

GRanges

```
length(gr); gr[1:5]  
seqnames(gr)  
start(gr)  
end(gr)  
width(gr)  
strand(gr)
```

DataFrame

```
mcols(gr)  
gr$exon_id
```

SqInfo

```
seqlevels(gr)  
seqlengths(gr)  
genome(gr)
```

GenomicRanges

An overview of functions for GRanges objects.

Intra range transformations
shift(), narrow(), resize(), flank()

Coverage and slicing
coverage(), slice()

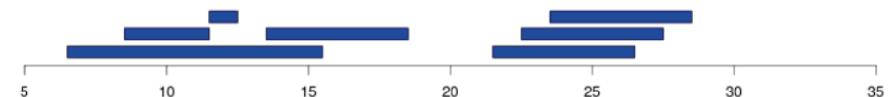
Inter range transformations
range(), reduce(), gaps(), disjoin()

Range-based set operations
union(), intersect(), setdiff(),
punion(), pintersect(), psetdiff(),
pgap()

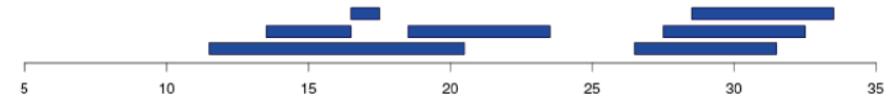
Finding/counting overlapping ranges
findOverlaps(), countOverlaps()

Finding the nearest range neighbor
nearest(), precede(), follow()

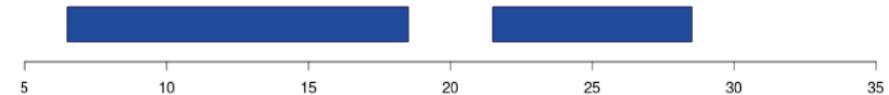
ir0



shift(ir0, 5)



reduce(ir0)



disjoin(ir0)

