Session II. Bioconductor

Armando Reyes-Palomares

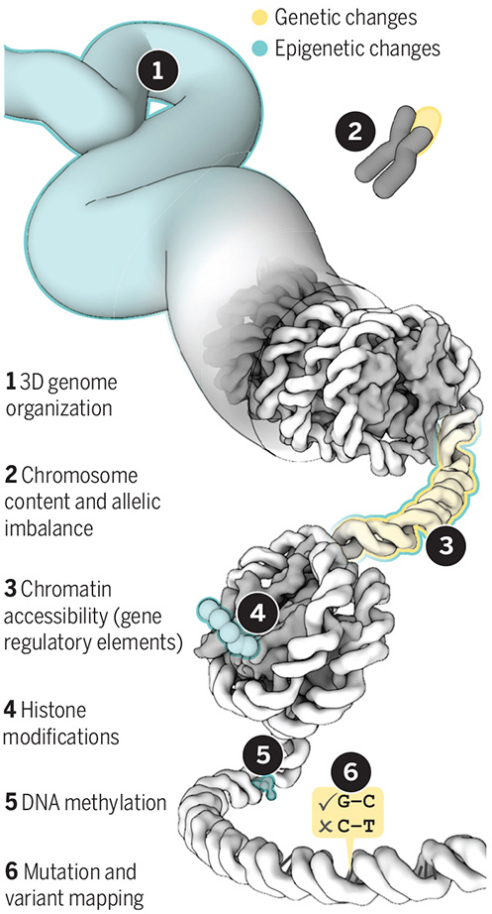
Department Biochemistry and Molecular Biology

Complutense University of Madrid

armandorp@ucm.es

Sequence and function of the cancer genome

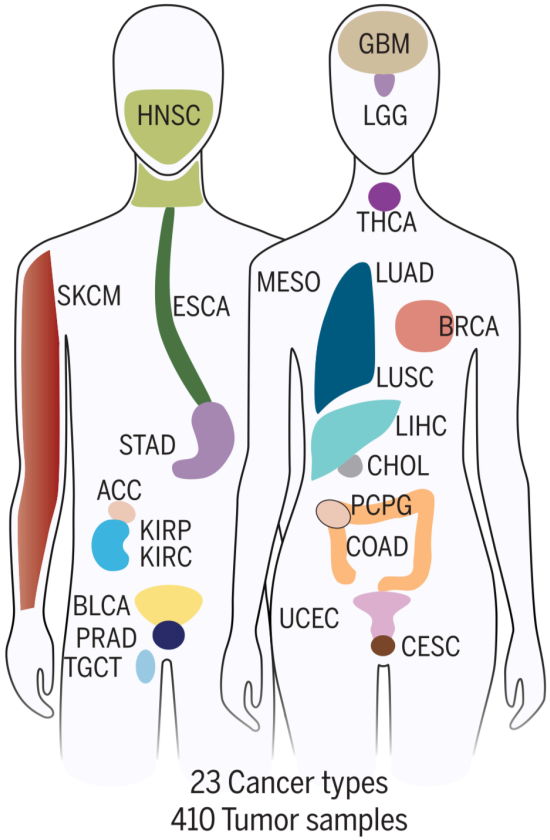
Multiomic mapping and comparison between genetic and epigenetic features are required for mechanistic understanding of cancer and for providing a "fingerprint" of the tumor. Multiomic analyses are likely to be important for cancer diagnosis and prediction of outcome, as well as for guiding treatment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

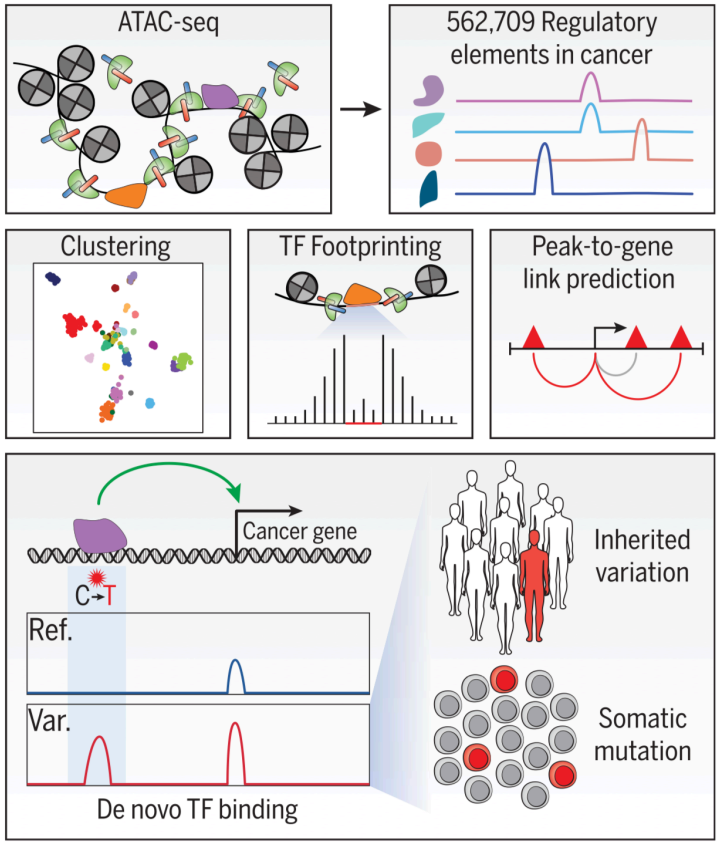


RESEARCH ARTICLE SUMMARY

CANCER

The chromatin accessibility landscape of primary human cancers





UNIVERSIDAD COMPLUTENSE MADRID

armandorp@ucm.es

Practical Sessions

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

armandorp@ucm.es

Summarizing (R introduction)

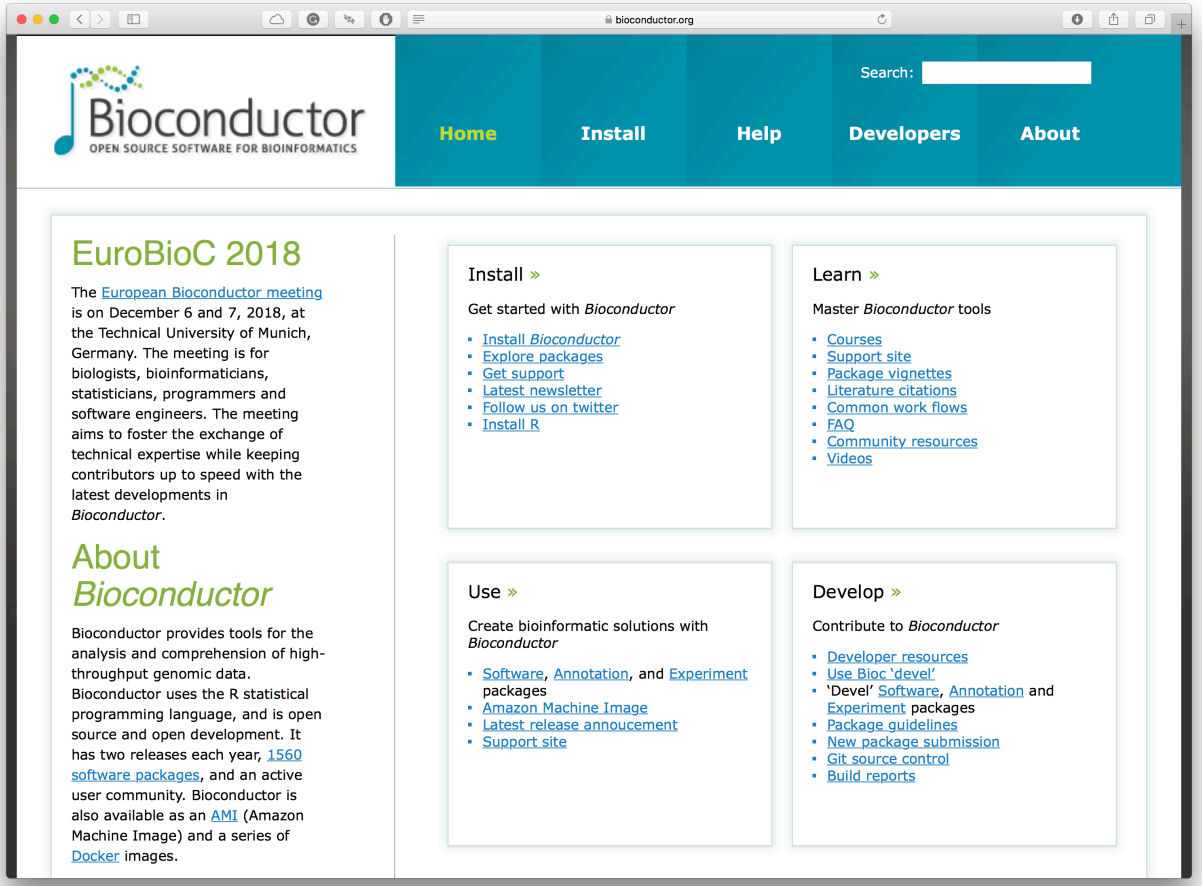
1. R and RStudio
2. Operate in R
3. Make **R objects** and explore **data structures**
4. Useful and common commands in R
5. R plots & graphs
6. Merge function in R

armandorp@ucm.es

Table of contents

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.



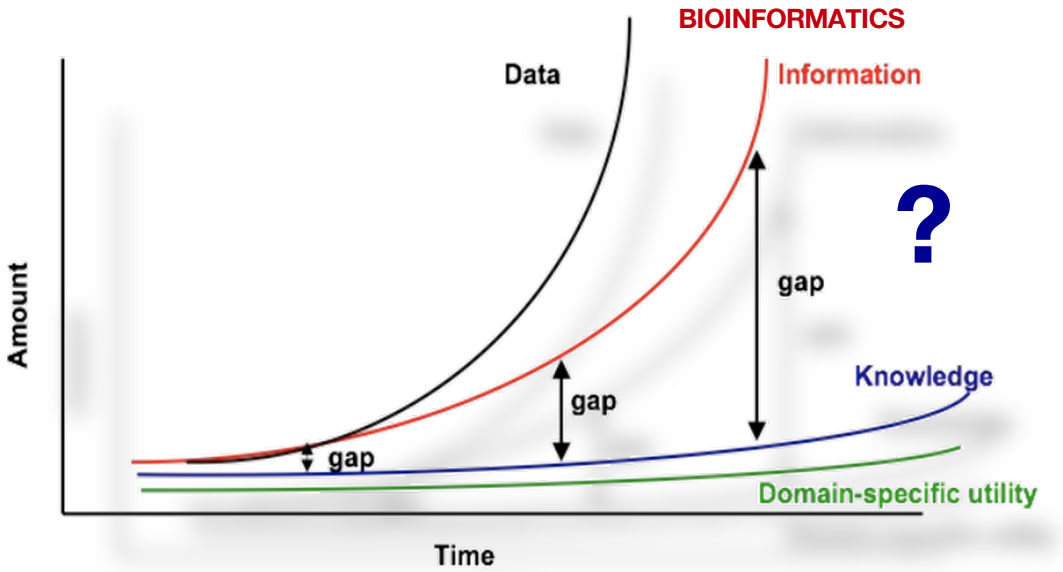
UNIVERSIDAD COMPLUTENSE MADRID

armandorp@ucm.es

Genomics → Molecular Biology

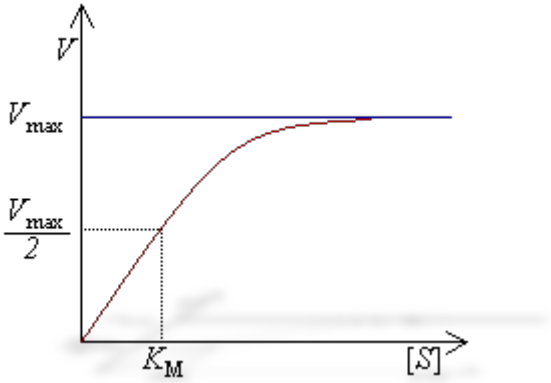
Revolutionary milestones: Recombinant DNA technology and NGS

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

Expression

- ArrayExpress
- Expression Atlas
- MetaboLights
- PRIDE

Chemical biology

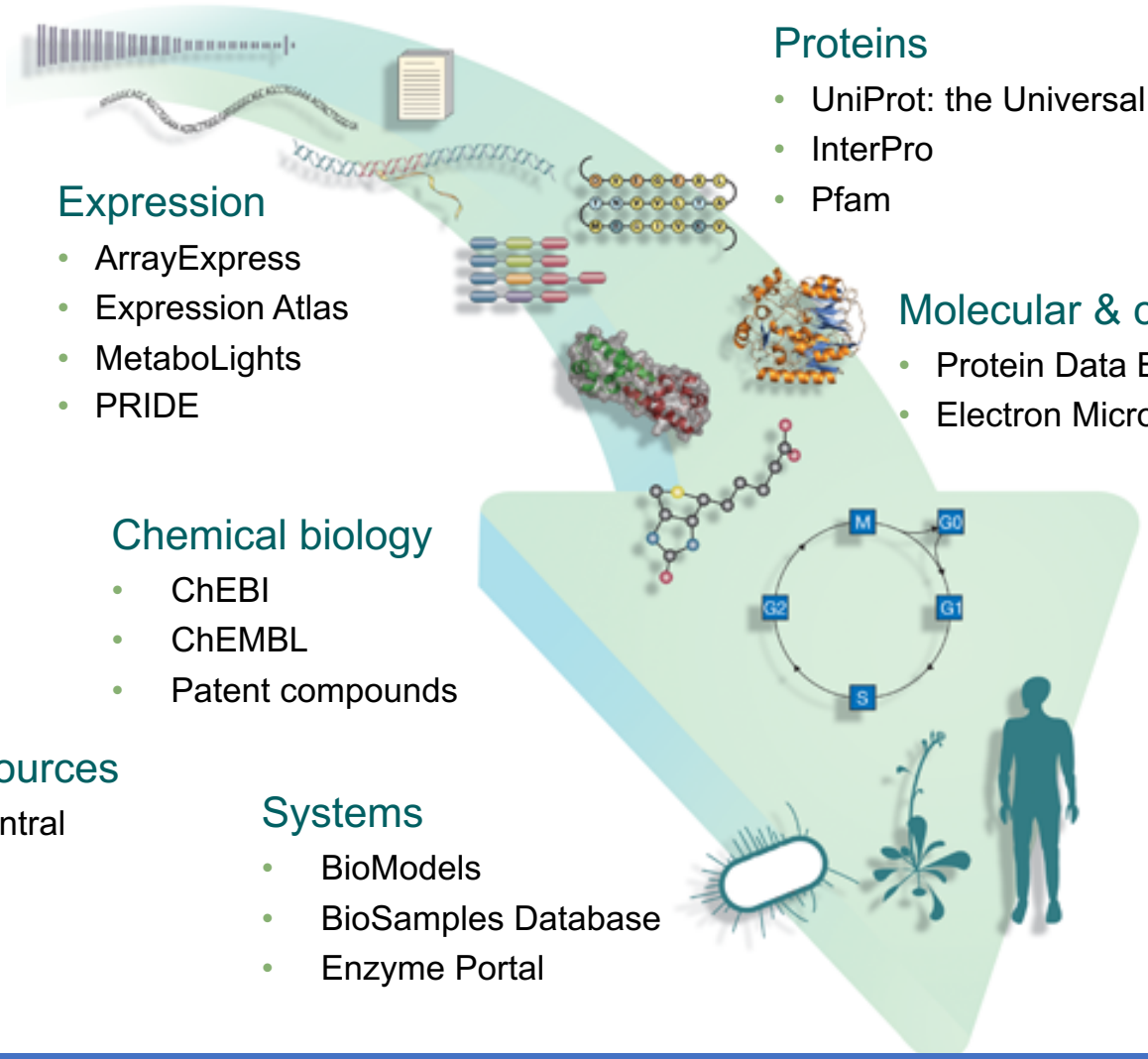
- ChEBI
- ChEMBL
- Patent compounds

Cross-domain resources

- Europe PubMed Central
- Gene Ontology

Systems

- BioModels
- BioSamples Database
- Enzyme Portal



Proteins

- UniProt: the Universal Protein Resource
- InterPro
- Pfam

Molecular & cellular structure

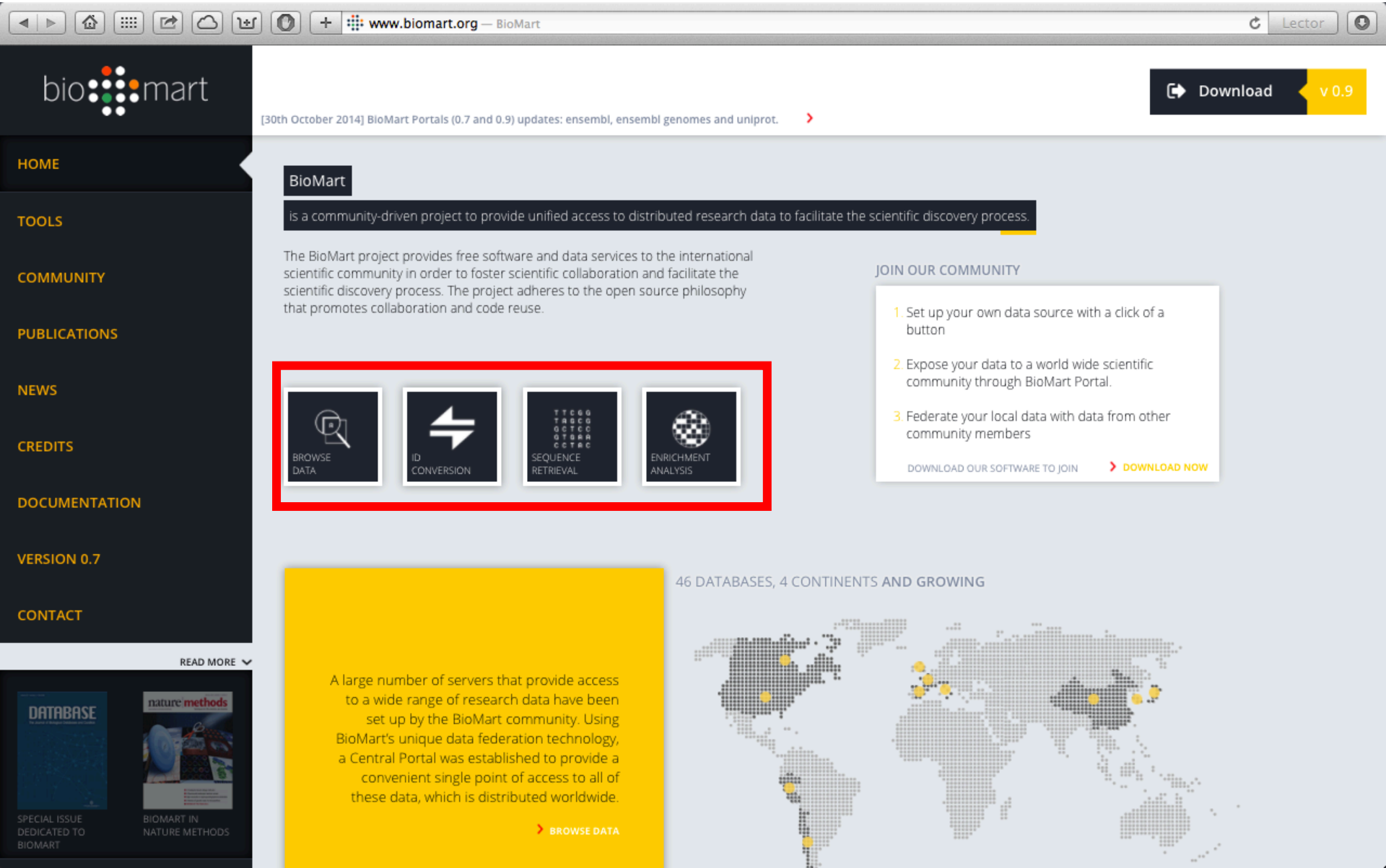
- Protein Data Bank in Europe
- Electron Microscopy Data Bank

Reactions, interactions & pathways

- IntAct
- Reactome

armandorp@ucm.es

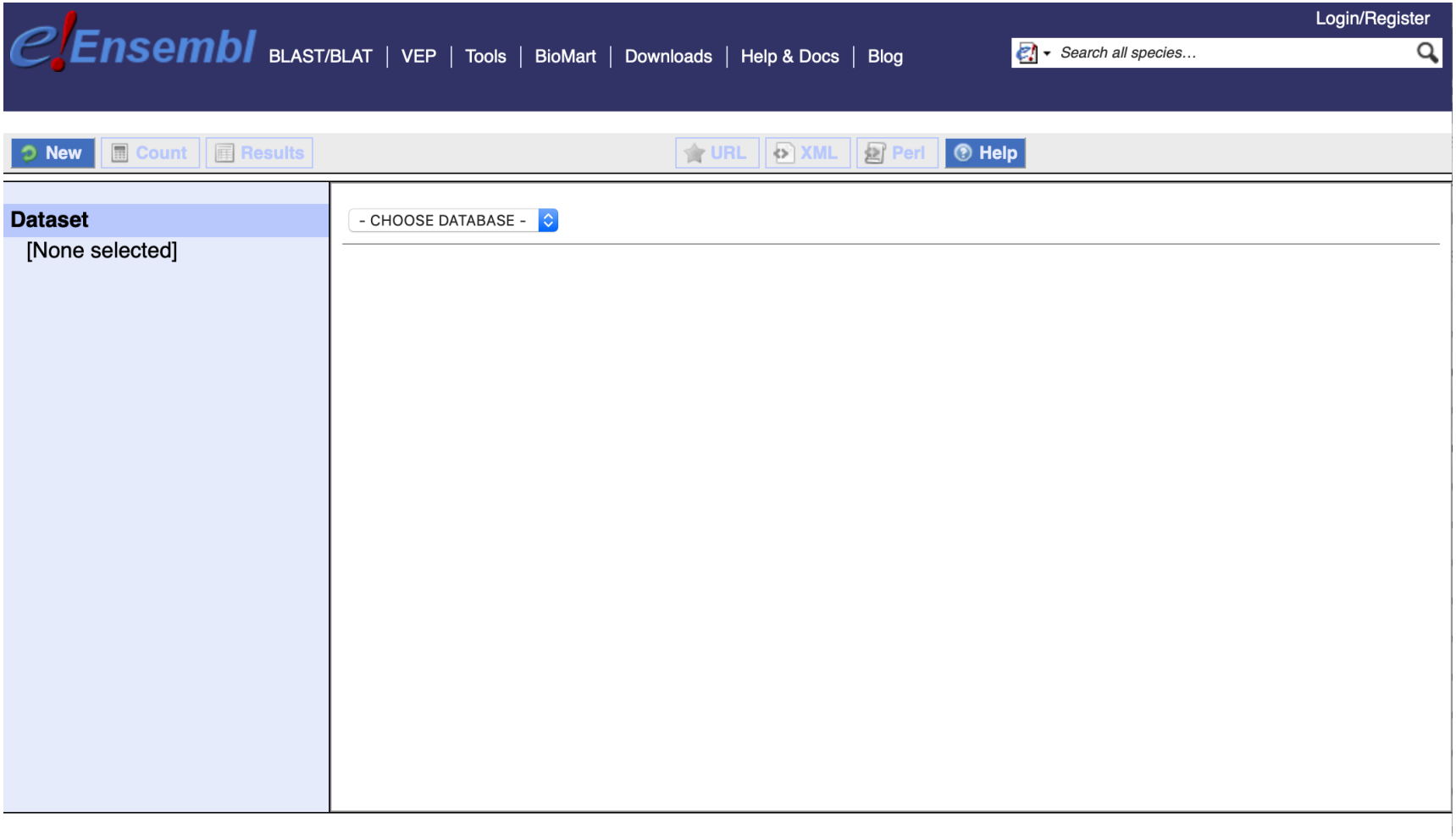
BioMart Data mining tool www.biomart.org



BioMart www.biomart.org

1 2 3 4

BioMart integrated in ENSEMBL

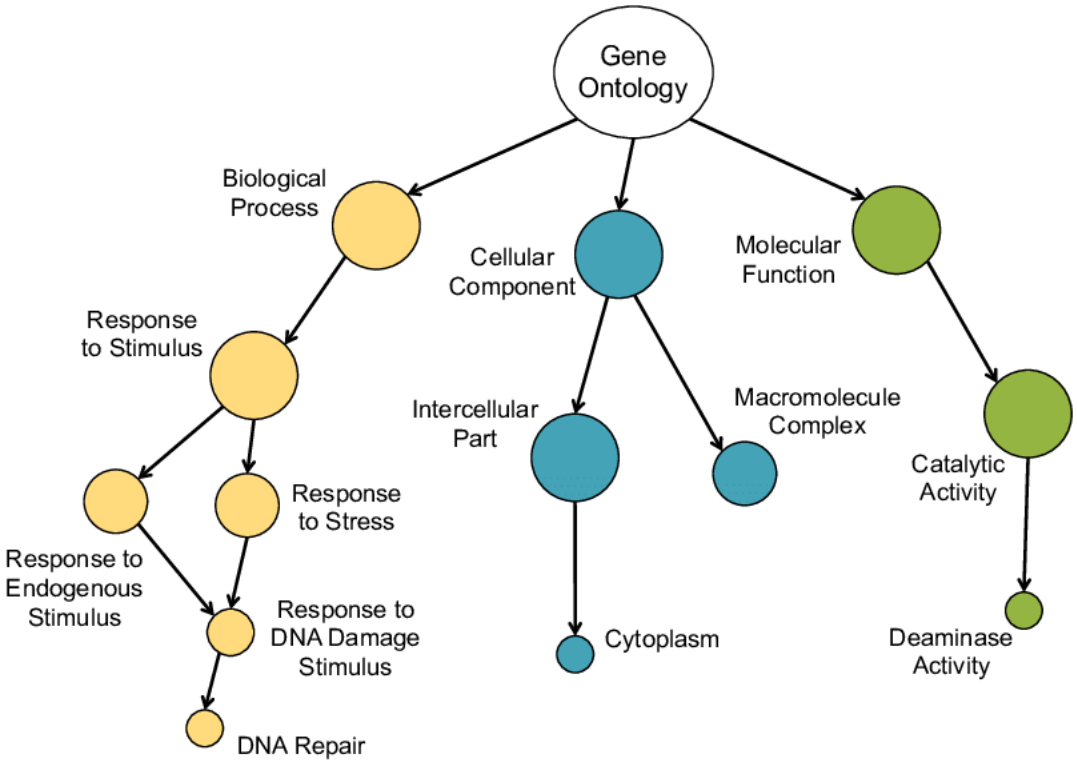


3 Main Integrated Mart Databases

- **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, synomyms, 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.

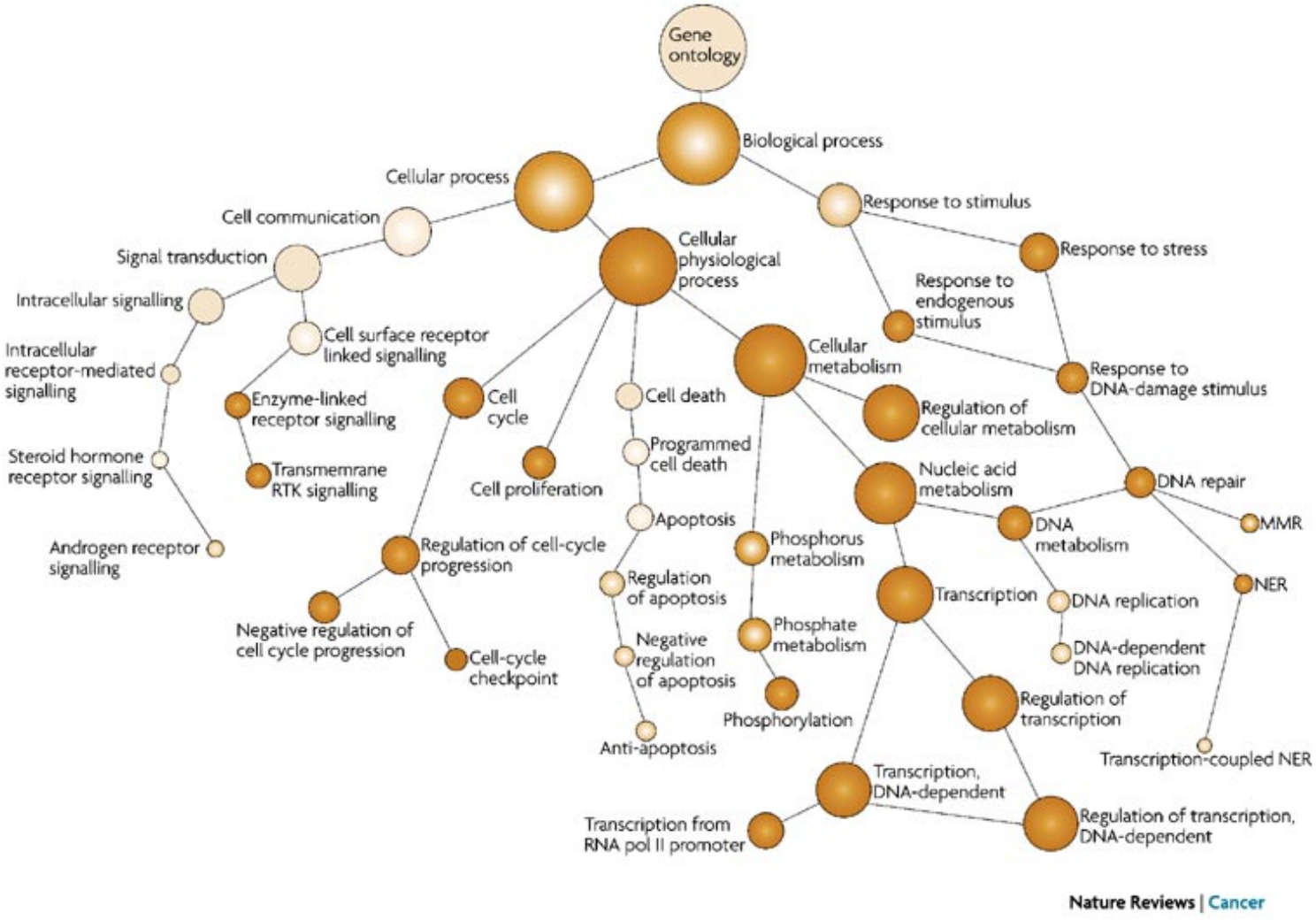
GENEONTOLOGY
Unifying Biology

geneontology.org



GENEONTOLOGY
Unifying Biology

geneontology.org



Types of inputs for genomic features:

>Homo sapiens, gene_ID_24349304323
GGCAGATTCCCCCTAGACCCGCCCGCACCATGGTCAGGCATGC
CCCTCCTCATCGCTGGGCACAGCCCAGAGGGTATAAACAGTGC
TGGAGGCTGGCGGGGCAGGCCAGCTGAGTCCTGAGCAGCAGC
CCAGCGCAGCCAC

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

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	14050

armandorp@ucm.es

GenomicRanges Package

```
> gr = exons(TxDb.Hsapiens.UCSC.hg19.knownGene); gr
GRanges with 289969 ranges and 1 metadata column:
            seqnames              ranges strand |   exon_id
               <Rle>           <IRanges>  <Rle> | <integer>
        [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_gl000249
            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

Seqinfo
- seqlevels(gr)
- seqlengths(gr)
- genome(gr)

armandorp@ucm.es

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 neighborn** nearest(), precede(), follow()