



Genomics and transcriptomics: from raw data to users

Chiara Colantuono
chiara.colantuono@szn.it

Extracting....

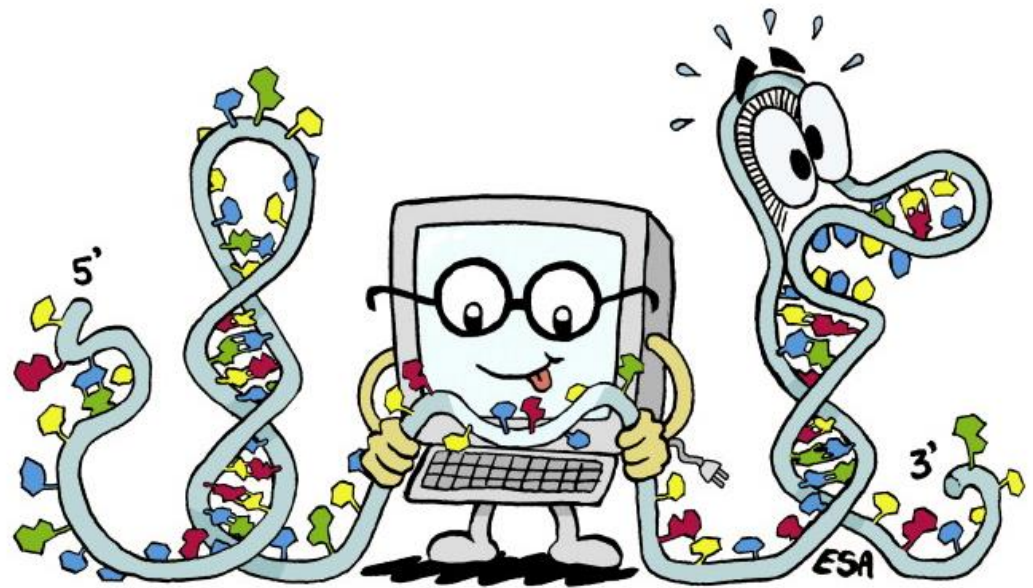
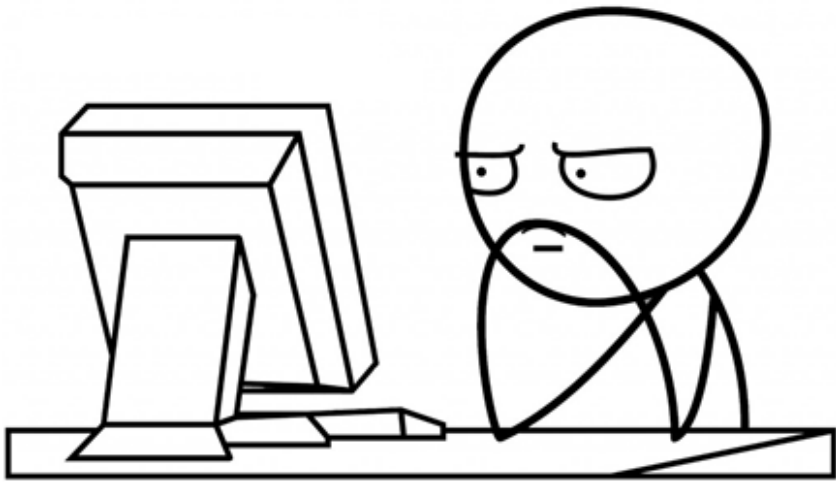


- Species
- Organisms
- Conditions
- Tissues
-

- DNA
- mRNA
- Total RNA
- Single cell
-



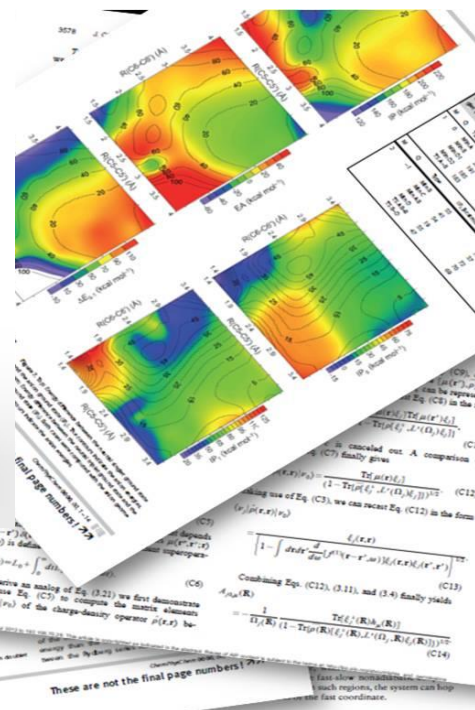
Waiting...



Sequencing results

You've

G



exchange com-
field are classic
of

such as those
who introduced many of the
as well as with
the early investigations, however.
Early in
nonpolar surfaces and two-state
transition probability computed by
the TSM method at then
the TSM method versions of the
programs, as we
always, the TSM
variable software
and Newton-X.

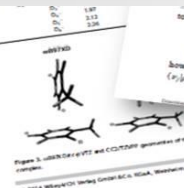


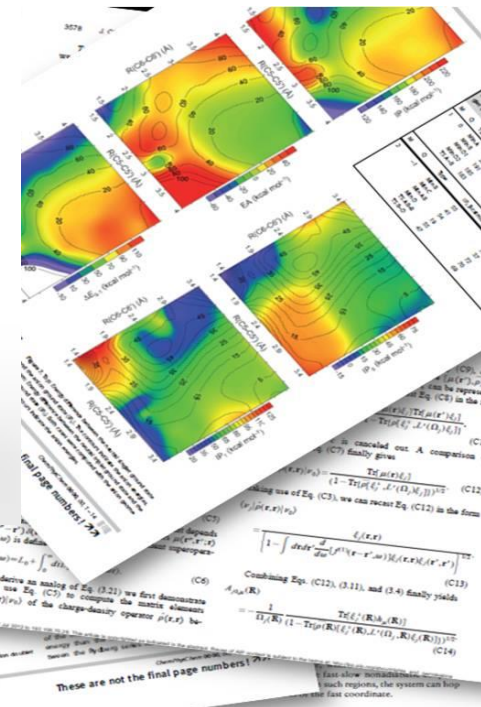
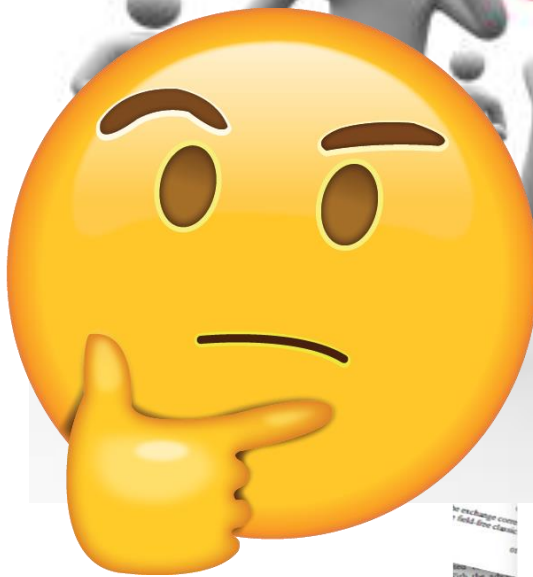
Figure 3. (a) TSM and (b) CVD/AFM
comparisons.

To derive an analog of Eq. (3.21) we first demonstrate
how to use Eq. (C7) to compute the matrix elements
 $\langle v_j | \rho(x, x') | v_k \rangle$ of the charge-density operator $\rho(x, x')$ be-

Combining Eqs. (C12), (3.11), and (3.4) finally yields

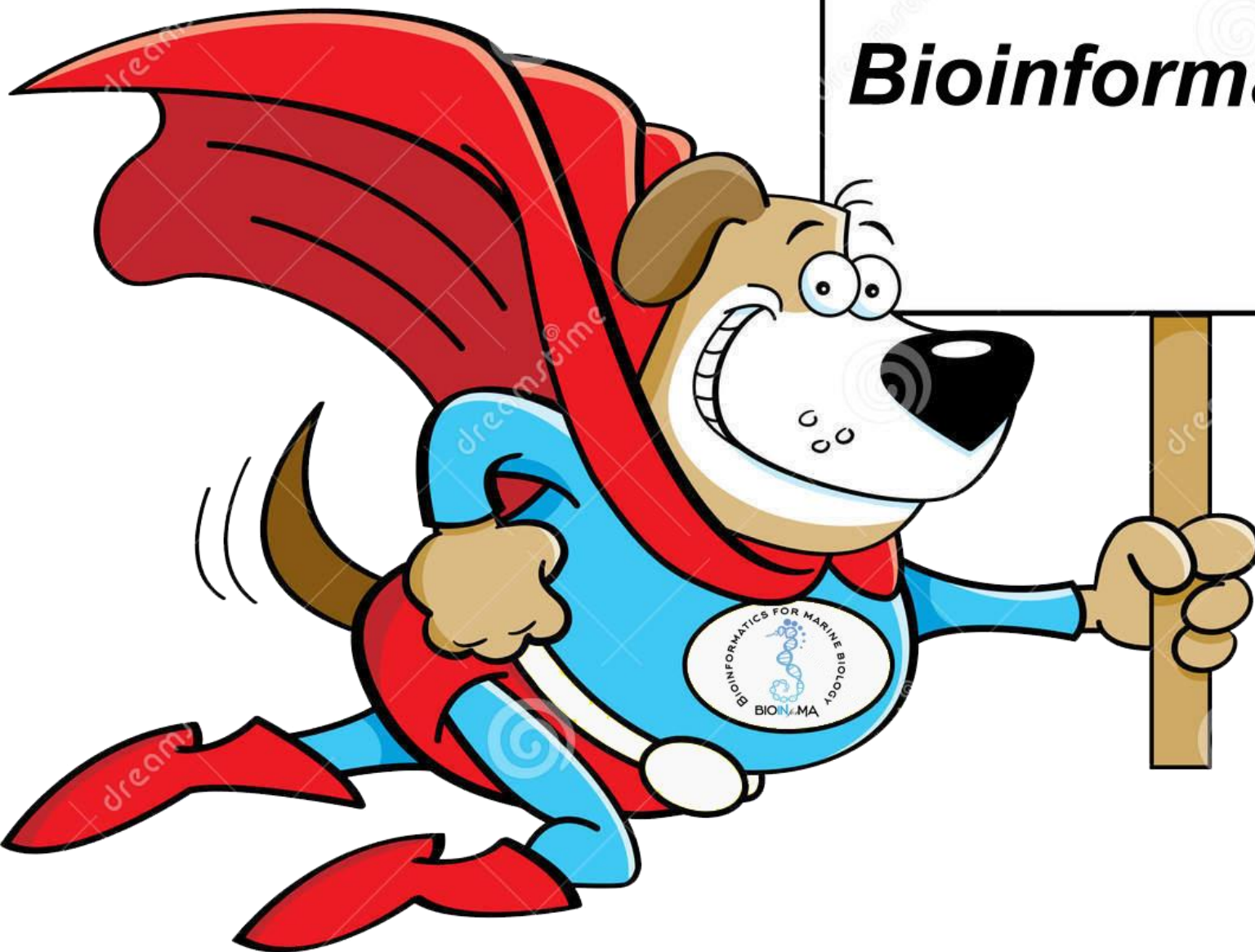
These are not the final page numbers!

You've
G



Sequencing results

[illegible]



Bioinformatician

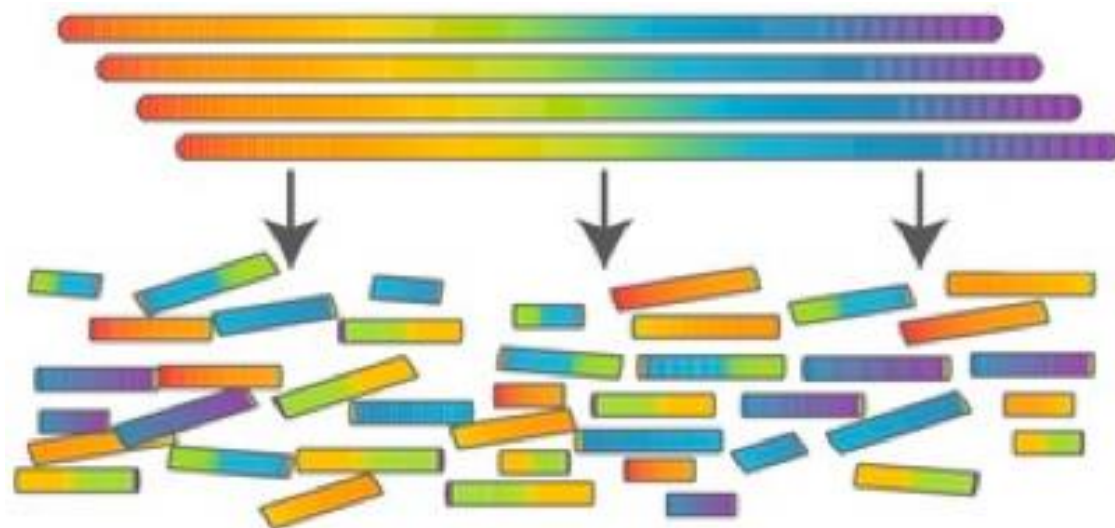


A word cloud of biological and genomic terms, arranged in a diamond shape. The words are written in a script font and color-coded: red, green, and blue. The terms include:

- Reads
- UTRs
- HeatMap
- HPCC
- GOenrich
- Genes
- Transposons
- Transcriptome
- Assembly
- Sequencing
- Structure
- Promoters
- Repeats
- Variant
- Genoma
- Function
- Annotation
- Adaptor
- Clean
- Exosome
- mRNA
- RNAseq

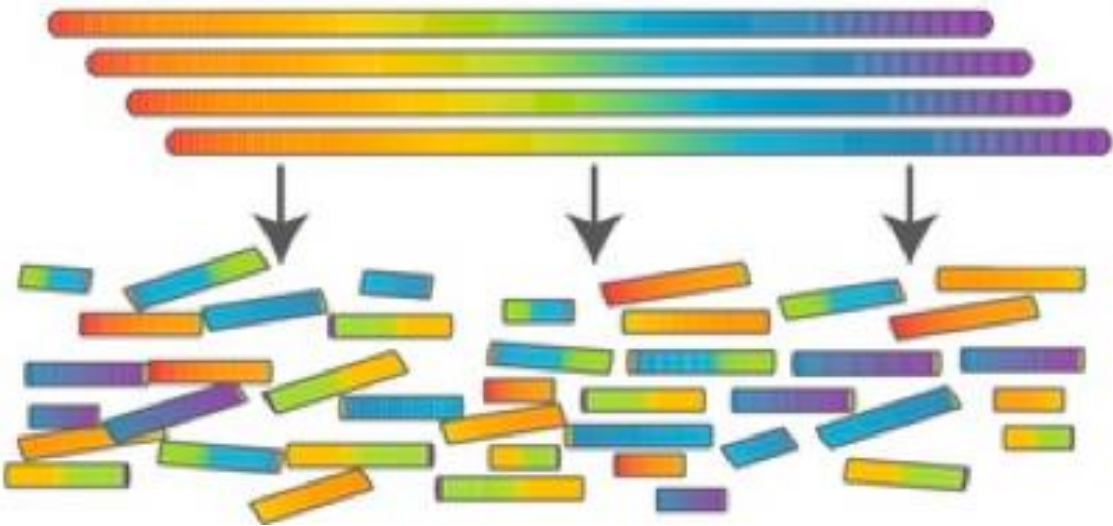


Reads Assembly



Genome/Transcriptome

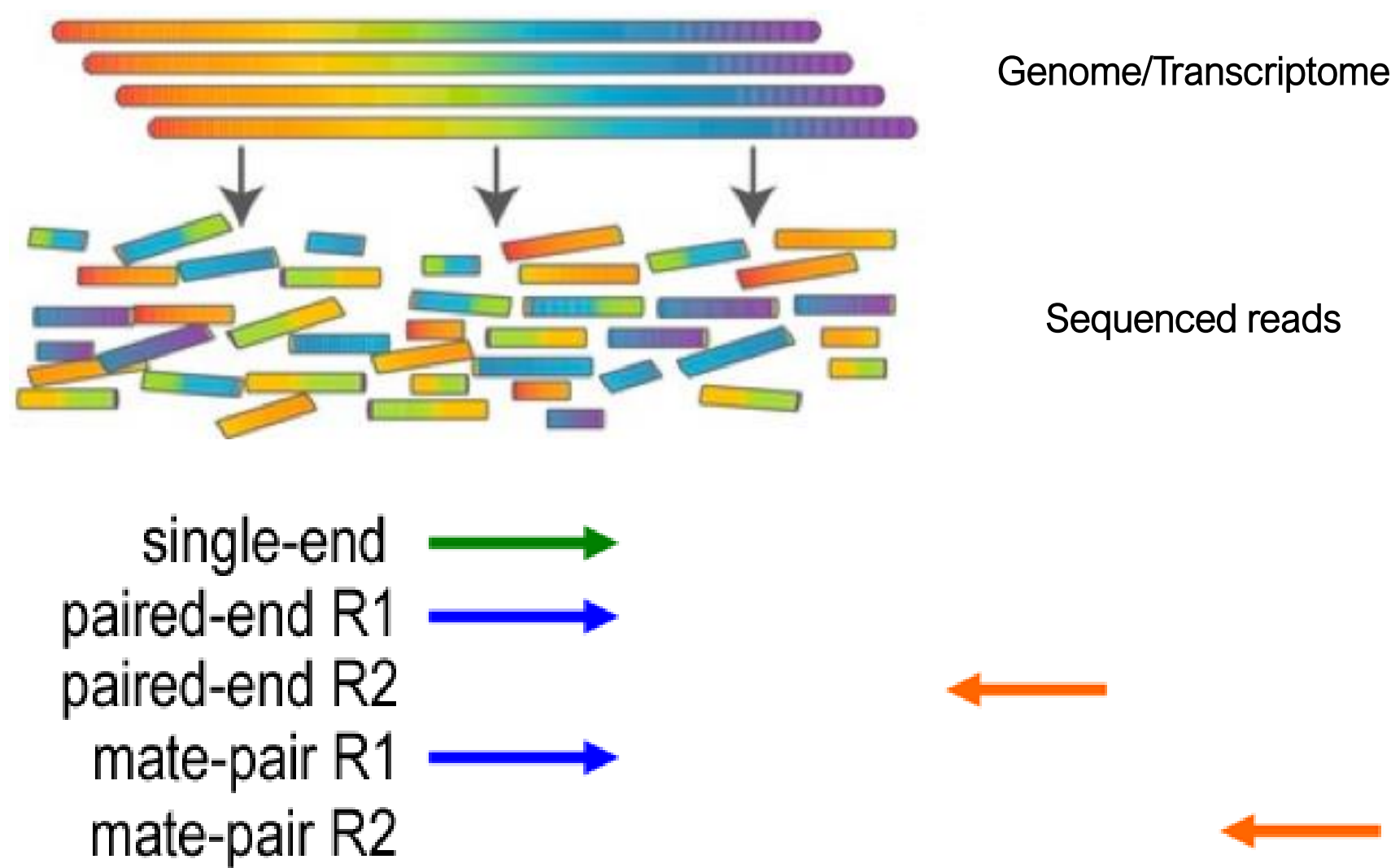
Sequenced reads



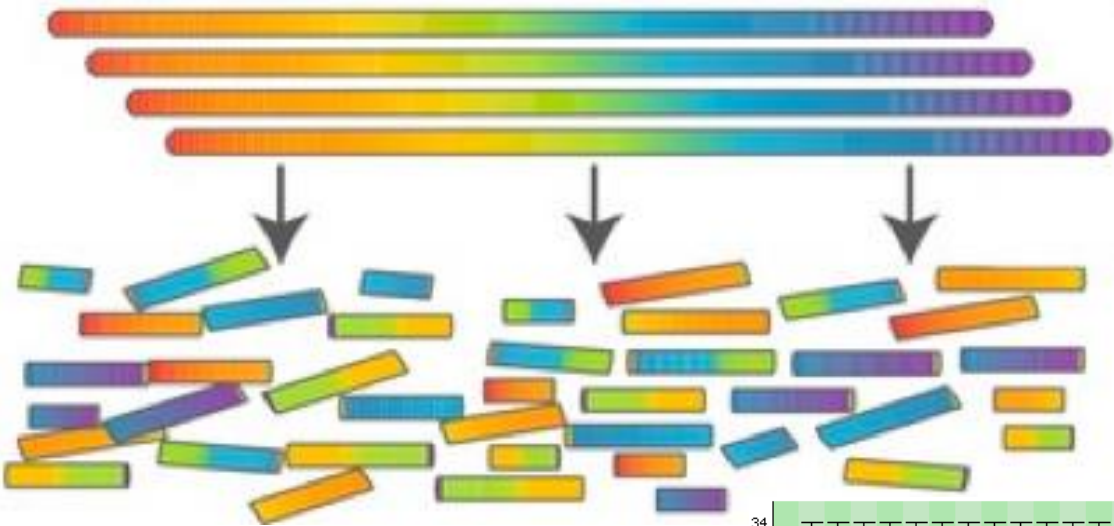
Genome/Transcriptome

Sequenced reads

Technology	Read length	Yield (Reads per Run)
Roche 454	700	~700 thousand
Illumina HiSeq	300	~300 billion
SOLiD	100	~200 billion
Ion Torrent	200	~60 billion
PacBio RS II	14,000	~47 thousand



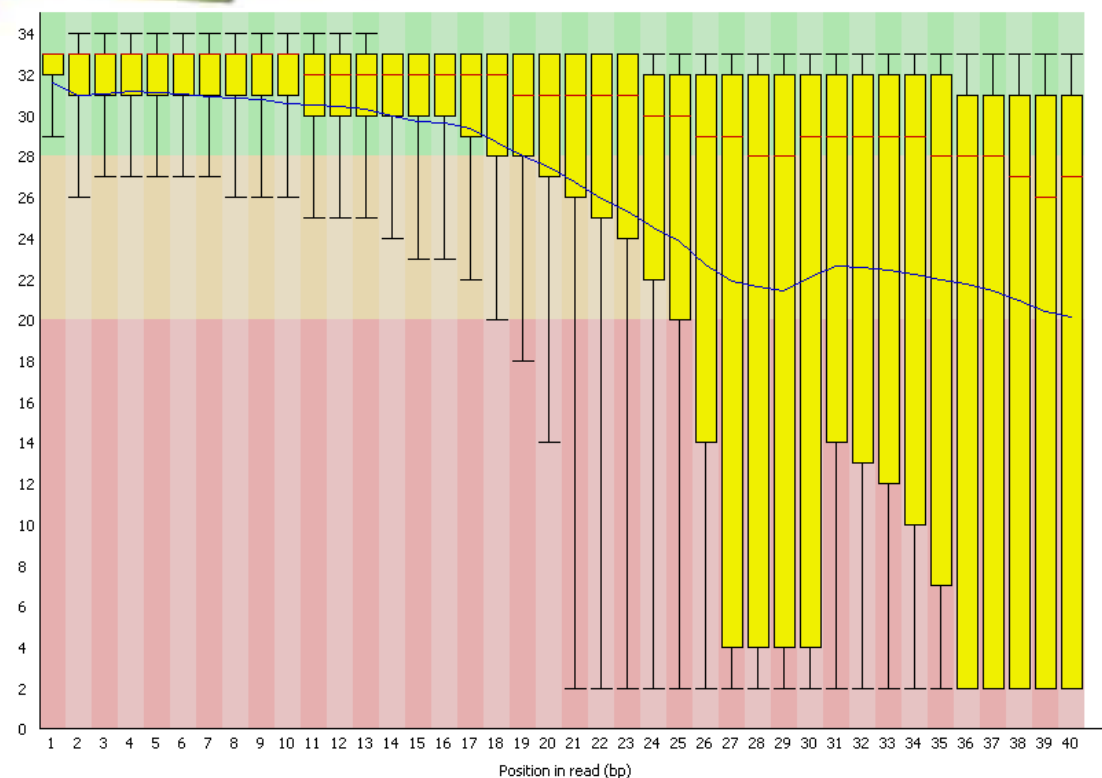
Reads Assembly



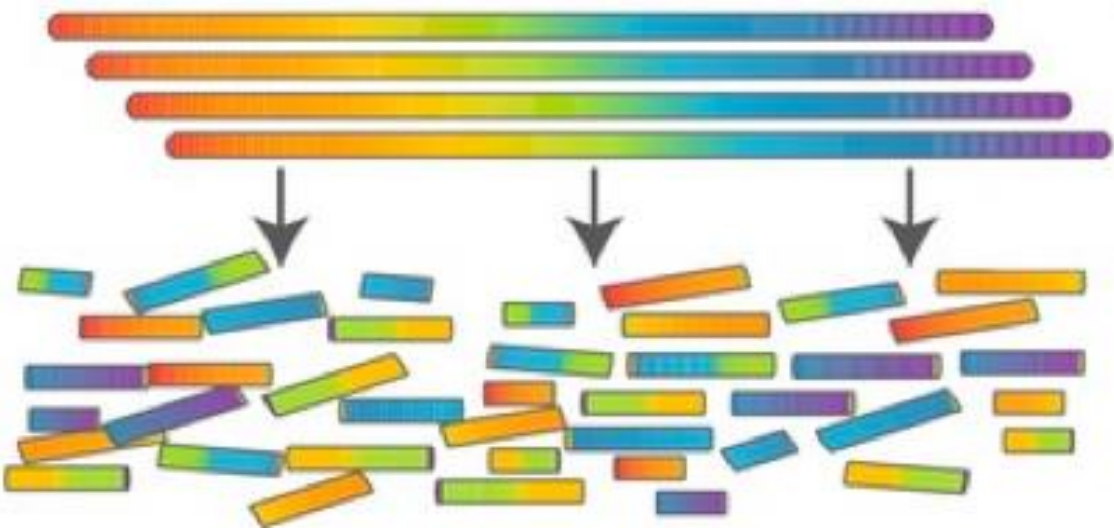
Genome/Transcriptome

Sequenced reads

Quality check

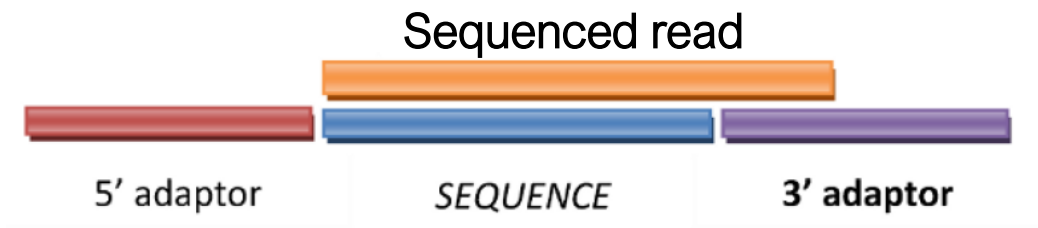
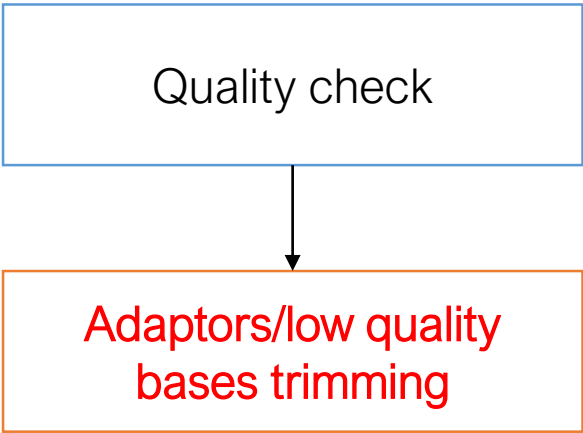


Reads Assembly

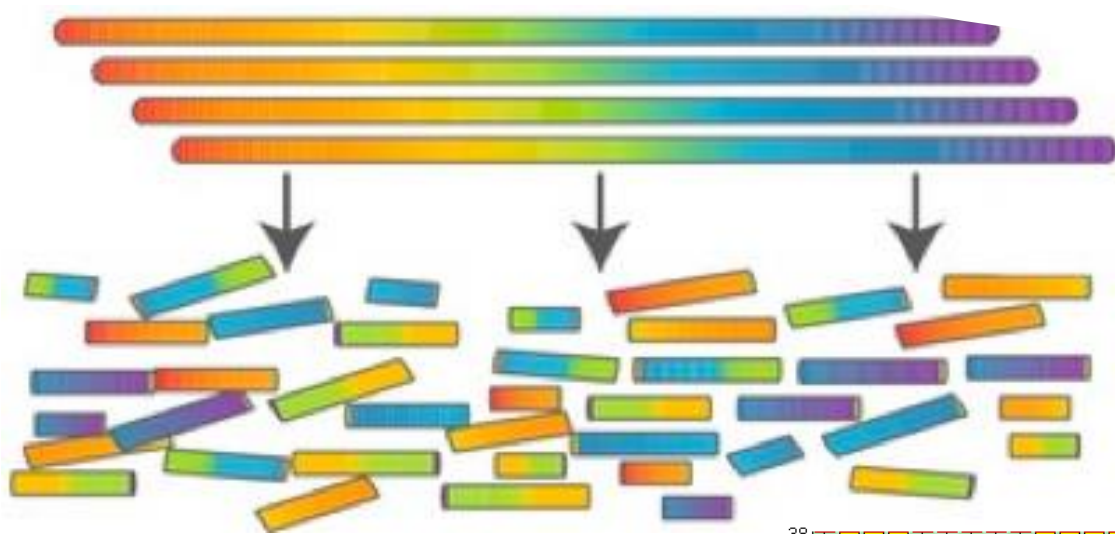


Genome/Transcriptome

Sequenced reads

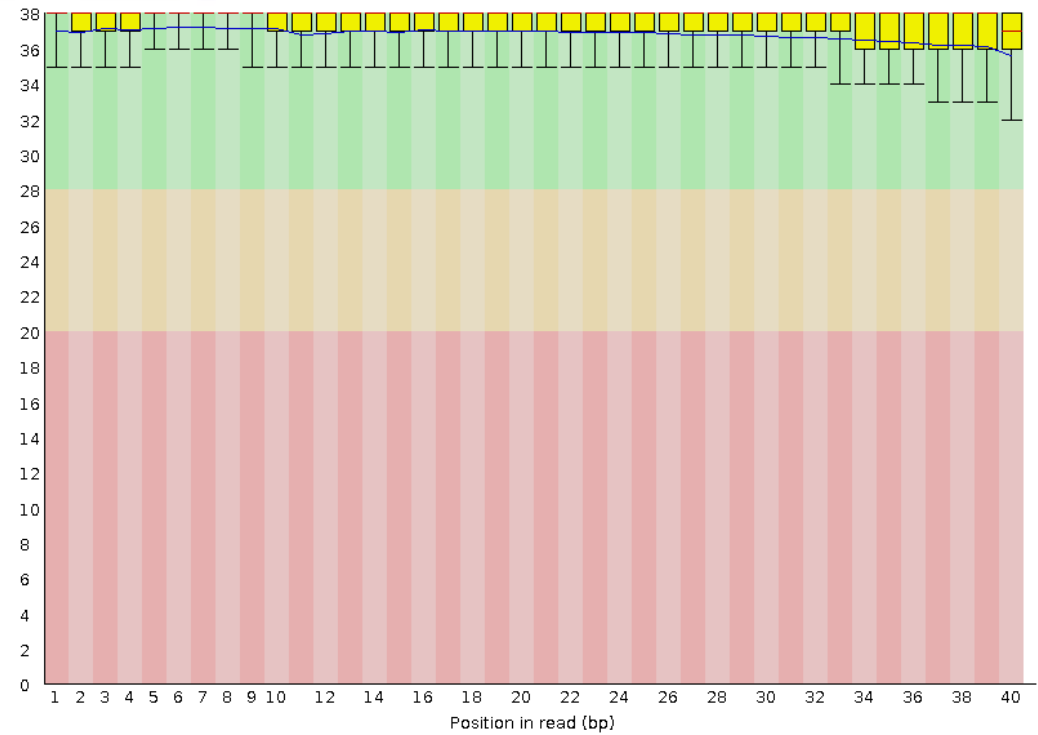
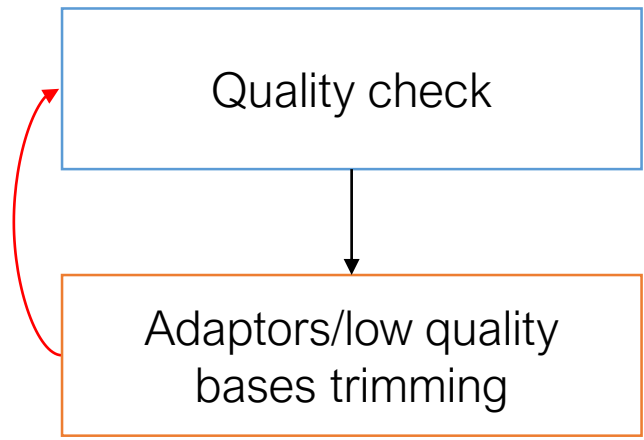


Reads Assembly

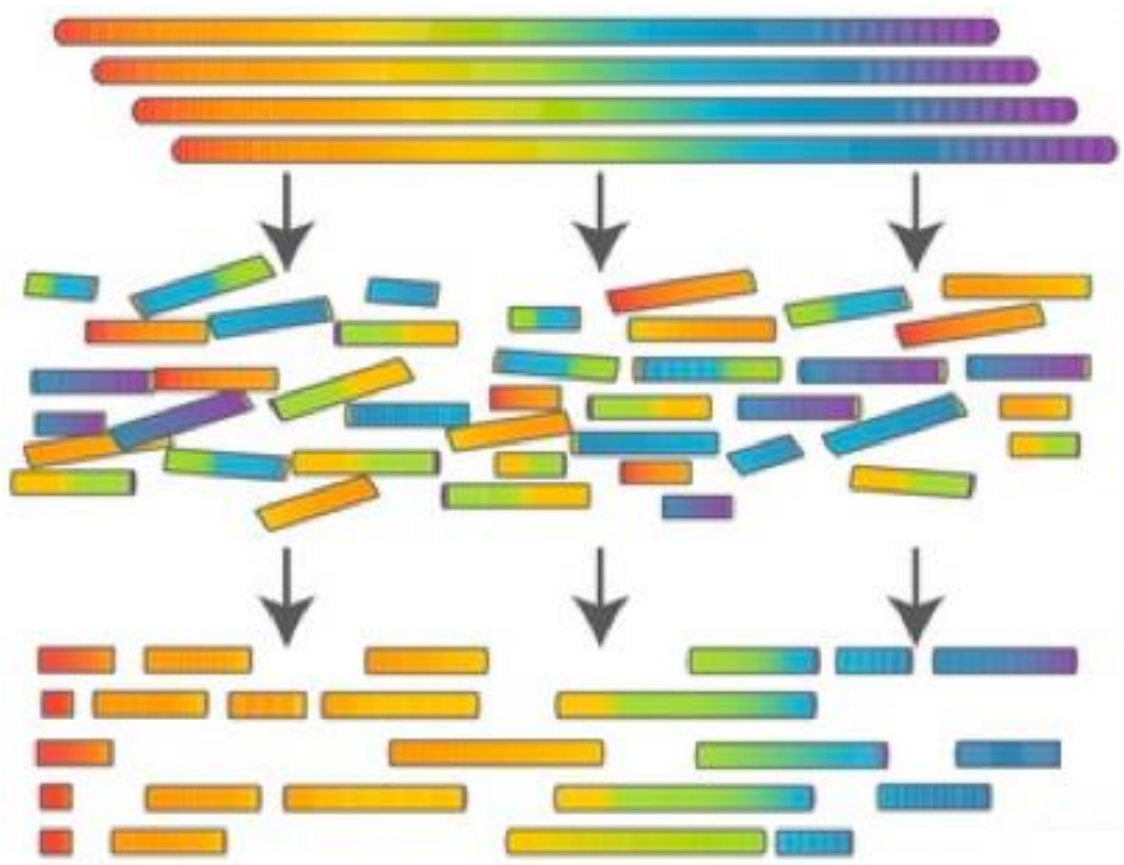


Genome/Transcriptome

Sequenced reads



Reads Assembly



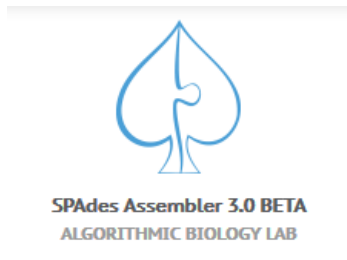
Genome/Transcriptome

Sequenced reads

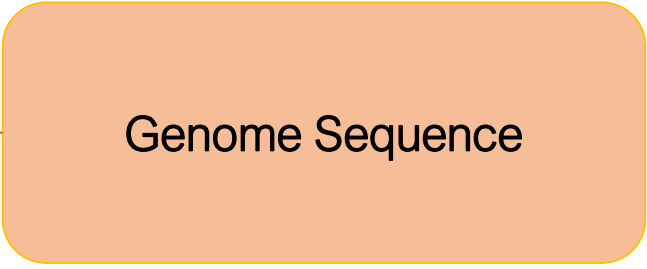
Path reconstruction by overlap

ATGTTCCGATTAGGAAACCTATCTGTAAGTGTTCATTTCAGTAAAAGGAGGAAA

Consensus







```
graph TD; A[Genome Sequence] --> B[Structural Annotation];
```

Genome Sequence

Structural Annotation

Genes structure prediction:

CDS;

UTRs;

exons/introns;

non-coding RNAs;

repeats;

.....

Genome Sequence

Structural Annotation

Ab initio

Tools scanning the sequence:
AUGUSTUS, Genefinder, Glimmer

....

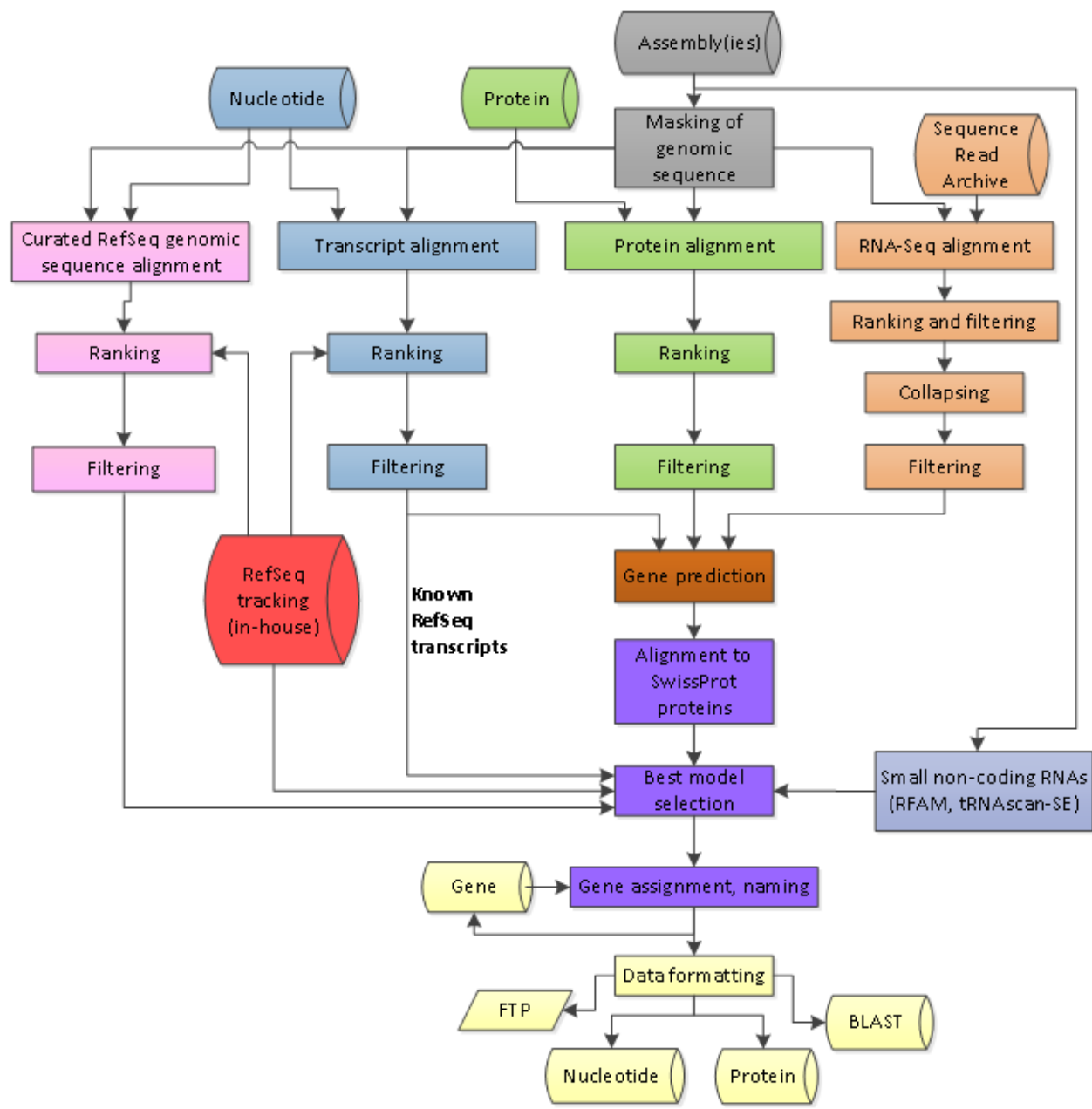
By similarity

Mapping cDNA/ESTs/repeats:
GenomeThreader, Infernal, gmap

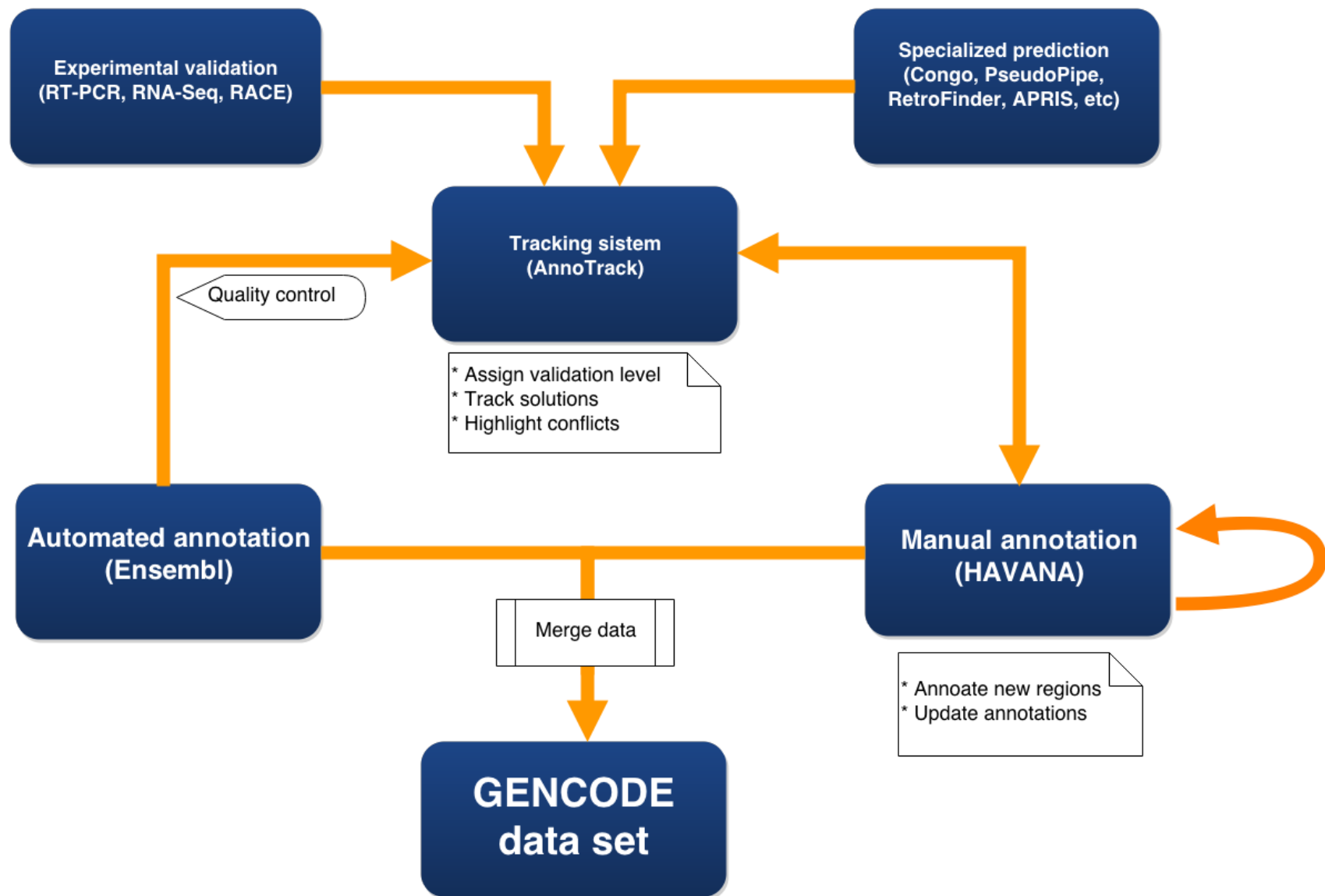
....

Genes structure prediction:
CDS;
UTRs;
exons/introns;
non-coding RNAs;
repeats;
.....

Genome Annotation: the RefSeq pipeline



Genome Annotation: the GENCODE pipeline



Ciona robusta

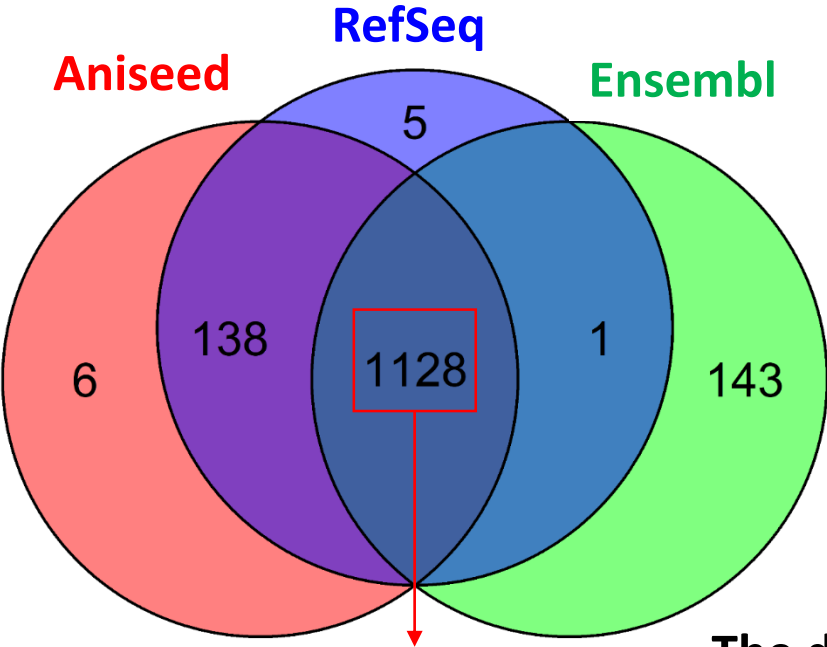


	ANISEED	REFSEQ	ENSEMBL
ANISEED	1272	1266	1128
REFSEQ	1266	1272	1129
ENSEMBL	1128	1129	1272

Number of genome scaffolds

ANISEED SPECIFIC	REFSEQ SPECIFIC	ENSEMBL SPECIFIC
chr4	chr4	chr1
chr9	chr9	chr2
chr11	chr11	chr3
chr12	chr12	chr4
chr13	chr13	chr5
MIT		chr6
		chr7
		chr8
		chr9
		chr10
		chr11
		chr12
		chr13
		chr14
		129

SCAFFOLDS

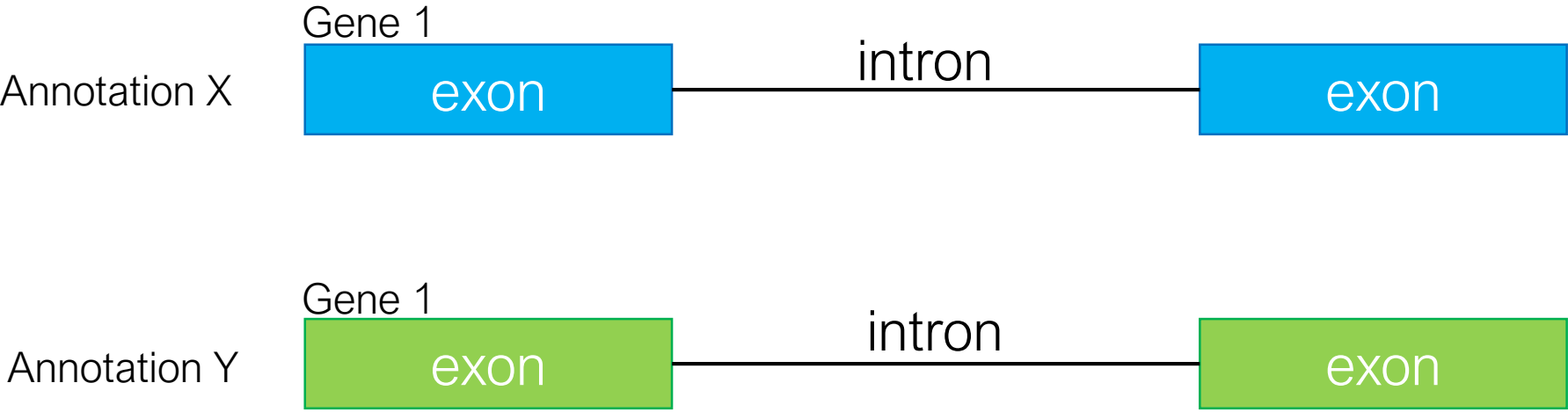


Number of scaffolds
in common among
all the assemblies

The differences among the scaffolds are mainly
inversions or a different number of undefined
nucleotides (Ns).

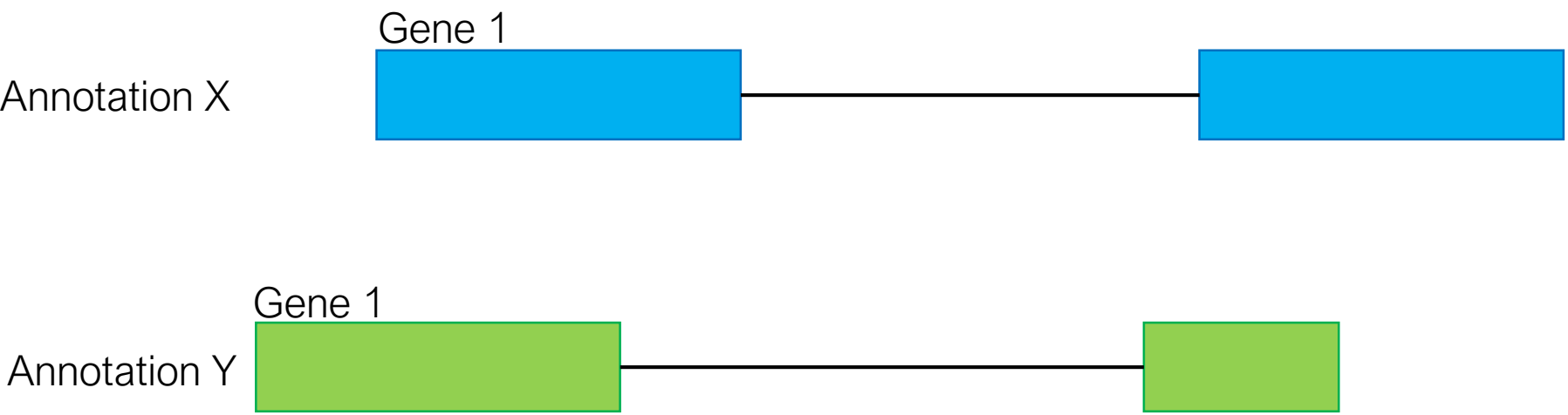
Identical exons start and end

Genome sequence



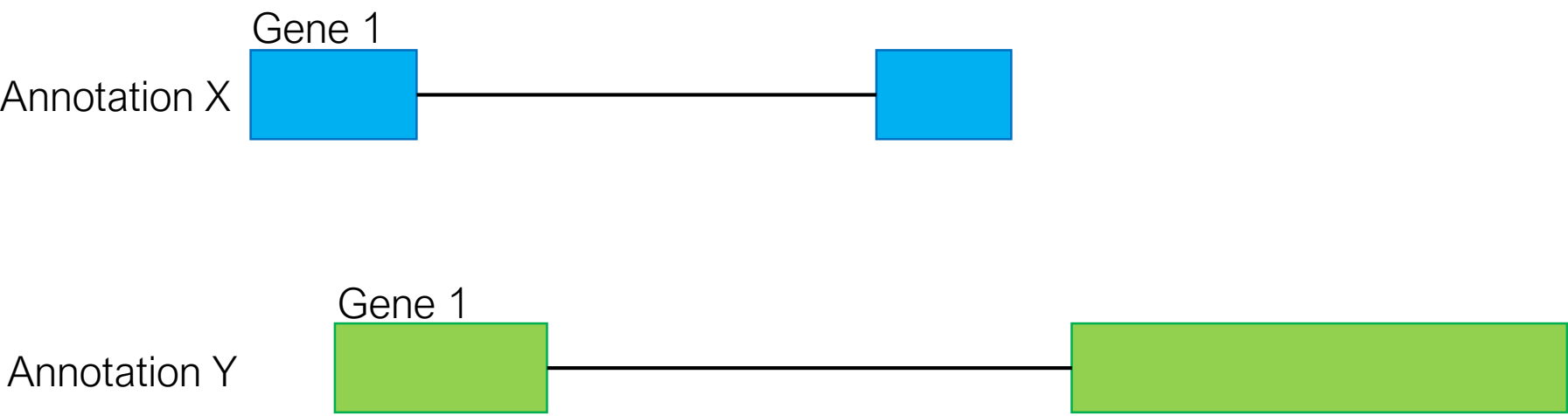
Genes overlapping but with a different gene structure

Genome sequence



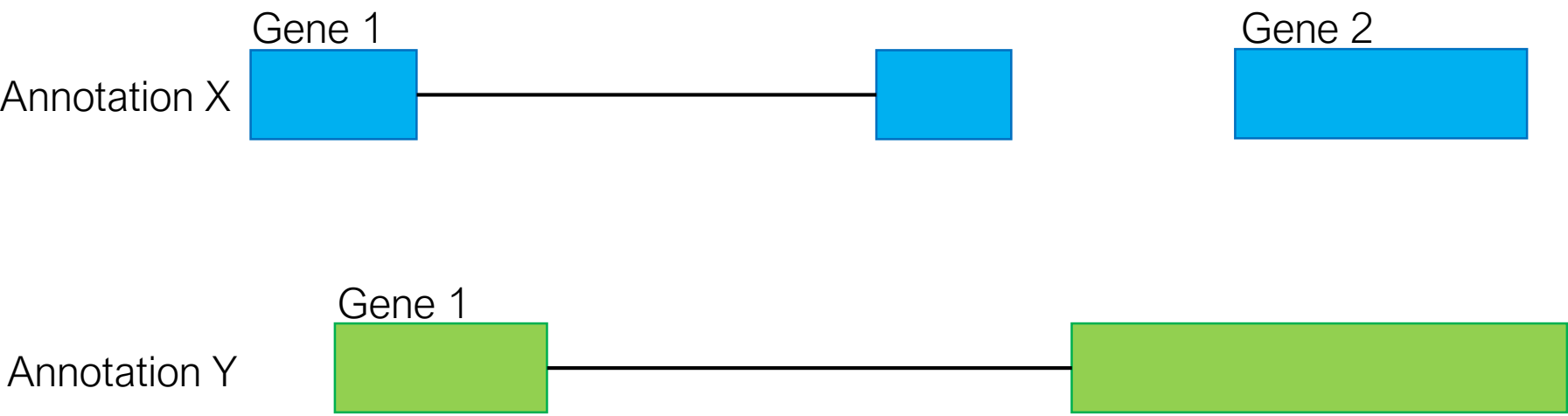
Genes overlapping but with a different gene structure

Genome sequence



Genes overlapping but with a different gene structure

Genome sequence



Not overlapping genes,
gene annotation specific

Genome sequence

Annotation X



Annotation Y



2

Acanthaster planci

<http://www.ncbi.nlm.nih.gov/>
<http://marinegenomics.oist.jp/>



Patiria miniata

<http://www.echinobase.org/>
<http://www.ncbi.nlm.nih.gov/>

2

2

Branchiostoma floridae

<http://www.ncbi.nlm.nih.gov/>
<http://genome.jgi.doe.gov/>



Phaeodactylum tricornutum

<http://protists.ensembl.org/>
<http://genome.jgi.doe.gov/>
<http://www.ncbi.nlm.nih.gov/>

3

3

Ciona robusta

<http://www.ensembl.org/>
<http://www.aniseed.cnrs.fr/aniseed/>
http://www.ncbi.nlm.nih.gov



Octopus bimaculoides

<http://www.metazome.net/>
<https://groups.oist.jp/>
<http://www.ncbi.nlm.nih.gov/>

3

3

Ciona savignyi

<http://www.ensembl.org/>
<http://www.aniseed.cnrs.fr/aniseed/>
http://www.ncbi.nlm.nih.gov



Saccoglossus kowalevskii

http://www.ncbi.nlm.nih.gov
<https://groups.oist.jp/>

2

3

Gasterosteus aculeatus

http://www.ncbi.nlm.nih.gov
http://www.ensembl.org
<http://sticklebrowser.stanford.edu/>



Strongylocentrotus purpuratus

<http://www.echinobase.org/>
<https://www.hgsc.bcm.edu/>
<http://www.ncbi.nlm.nih.gov/>

3

GENOMA: GENOMES FOR MARINE BIOLOGY

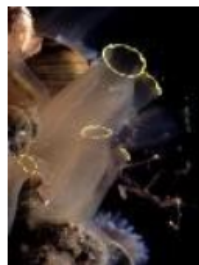


<http://bioinfo.szn.it/genoma/>

GENOMA: GENOMES FOR MARINE BIOLOGY

Species ▲	NCBI_latest_versions ◆	ENSEMBL_latest_versions ◆	JGI_latest_versions ◆	OTHERS_latest_versions ◆
Acanthaster planci	GCF_001949145.1	NA	NA	MarineGenomics: oki-cotsv1
Branchiostoma floridae	GCF_000003815.1_V2	NA	Branchiostoma_floridae_v2.0	NA
Ciona robusta	GCF_000224145.2_KH	GCA_000224145.1_KH	Assembly v2.0	Aniseed: KH12
Ciona savignyi	GCA_000149265.1	CSAV 2.0	NA	Aniseed: ENS81
Gasterosteus aculeatus	GCA_000180675.1	BROAD S1	NA	UCSC: gasAcu1
Octopus bimaculoides	GCF_001194135.1_V2.0	PRJNA270931	Obimaculoides_280	NA
Patiria miniata	GCA_000285935.1	NA	NA	Echinobase: V2
Phaeodactylum tricornutum	GCF_000150955.2	ASM15095v2	Phatr2	NA
Ptychodera flava	NA	NA	NA	OIST: 1426034866.unique_0.7_1e-10
Saccoglossus kowalevskii	GCF_000003605.2	NA	NA	OIST: V3
Strongylocentrotus purpuratus	GCF_000002235.4_Spur_4.2	Spur_3.1	NA	Echinobase: v4.2
Zostera marina	GCA_001185155.1_Zosma_marina.v.2.1	NA	NA	NA

[back to the GENOMA home page](#)



Ciona robusta genome portal

Genome information overview:

there are three different genome assemblies available:

- GCF_000224145.3 from NCBI;
- KH.92 from ENSEMBL;
- KHGene.2012 from Aniseed.

All the assemblies consist of 14 chromosomes plus the mitochondrion and 1257 unplaced scaffolds, having in total 1272 sequences. We compared the sequences of the assemblies, checking how many sequences are completely identical. The number of genomic elements in common (identical sequence) among the assemblies is reported in the following table:

	NCBI	ENSEMBL	ANISEED
NCBI	1272	1129	1266
ENSEMBL	1129	1272	1128
ANISEED	1166	1128	1272

The GENOMA platform

Gene content overview:

Search: Show entries

Type	Gene ID	Chromosome	Start	End	Strand	Annotation
RefSeq	XM_018815685.1	Chr01	2407	20500	-	regulator of G-protein signaling 22-like
ANISEED	KH.C1.286.v1.R.ND1-1	Chr01	2417	20402	-	ORF:[<1..4215] Homolog:[cDNA FLJ40080 fis clone TESTI2001795 highly similar to Homo sapiens regulator of G-protein signalling 22 (RGS22) mRNA]
ANISEED	KH.C1.286.v1.A.SL2-1	Chr01	2417	20413	-	ORF:[12..4226] Homolog:[cDNA FLJ40080 fis clone TESTI2001795 highly similar to Homo sapiens regulator of G-protein signalling 22 (RGS22) mRNA]
ENSEMBL	ENSCING00000016550	Chr01	4219	20394	-	n/a
ANISEED	KH.C1.560.v1.C.ND1-1	Chr01	14261	14503	+	ORF:[*90..245]
RefSeq	XM_002123150.2	Chr01	23507	52048	+	ileal sodium/bile acid cotransporter-like
ANISEED	KH.C1.907.v1.A.SL3-1	Chr01	23513	34214	+	ORF:[1..588] Homolog:[Solute carrier family 10 member 6]
ENSEMBL	ENSCING00000016539	Chr01	23518	34155	+	n/a
ANISEED	KH.C1.907.v1.R.ND1-1	Chr01	23530	34214	+	ORF:[<1..571] Homolog:[Solute carrier family 10 member 6]
ANISEED	KH.C1.907.v2.B.ND2-1	Chr01	24117	34214	+	ORF:[<1..965] Homolog:[similar to solute carrier family 10 member 2]
RefSeq	XM_018815677.1	Chr01	33654	60443	-	brefeldin A-inhibited guanine nucleotide-exchange protein 2 transcript variant X3
RefSeq	XM_018815674.1	Chr01	36633	60443	-	brefeldin A-inhibited guanine nucleotide-exchange protein 2 transcript variant X2
RefSeq	XM_009863811.2	Chr01	36633	60443	-	brefeldin A-inhibited guanine nucleotide-exchange

Download information:

Source	Genome Assembly (FASTA-big file)	Gene Annotation (GFF3)	Gene Annotation (CSV)	Gene Sequences (FASTA)	mRNA Sequences (FASTA)	Protein Sequences (FASTA)	INTERPRO results (CSV)
<u>NCBI</u>	GCF_000224145.3	GCF_000224145.3	download	download	download	download	download
<u>ENSEMBL</u>	KH.92	KH.92	download	download	download	download	download
<u>ANISEED</u>	KHGene.2012	KHGene.2012	download	download	download	download	download

Genome Assembly: FASTA file of all the chromosome/scaffolds/contigs;

Gene Annotation: GFF file containing information about all the annotated genes, mRNAs, UTRs, CDSs;

Gene Annotation: CSV file containing information ONLY about the annotated genes (Gene ID, chromosome, start, end, strand, functional annotation);

Gene Sequences: FASTA file of all the gene sequences;

mRNA Sequences: FASTA file of all the mRNA sequences;

Protein Sequences: FASTA file of all the protein sequences;

INTERPRO results: CSV file containing the INTERPROSCAN results (Gene ID, Interpro, PFAM, Panther, Pathways).

[back to the GENOMA home page](#)

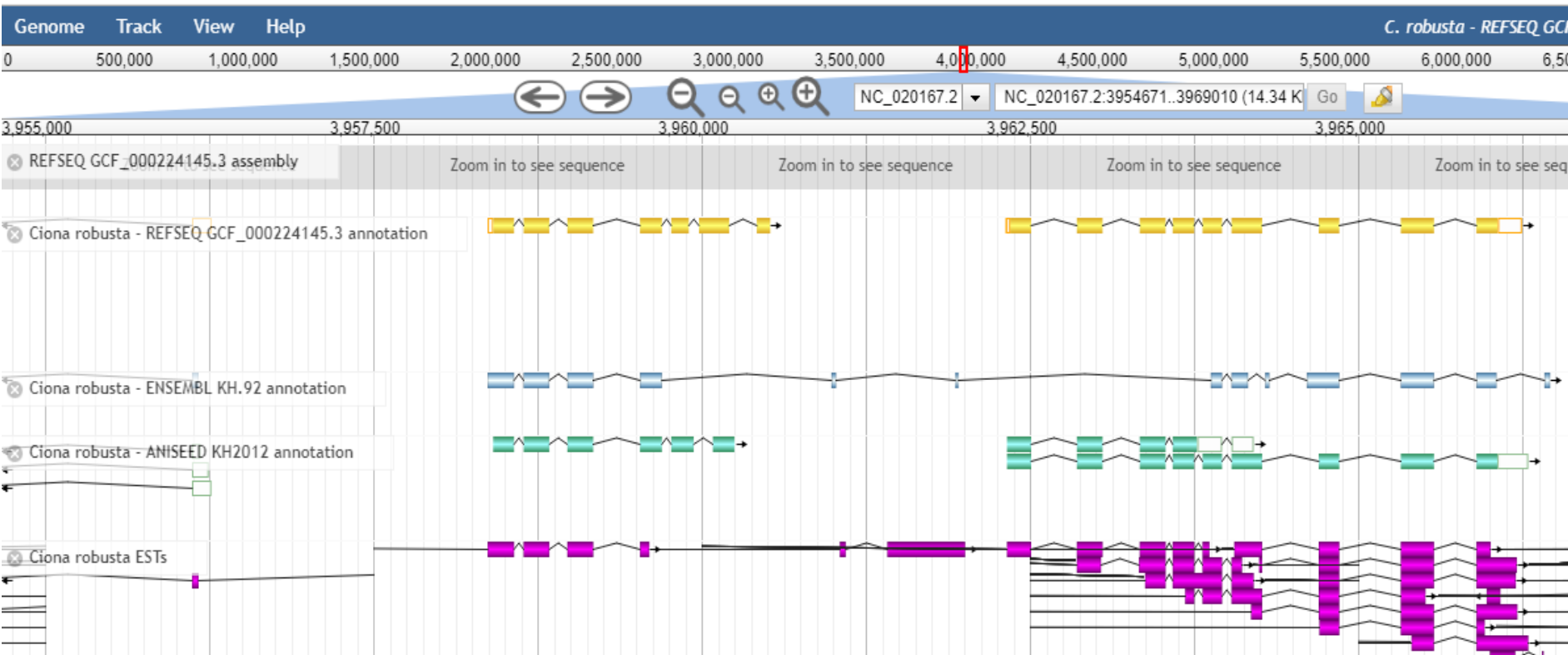


Ciona robusta genome portal

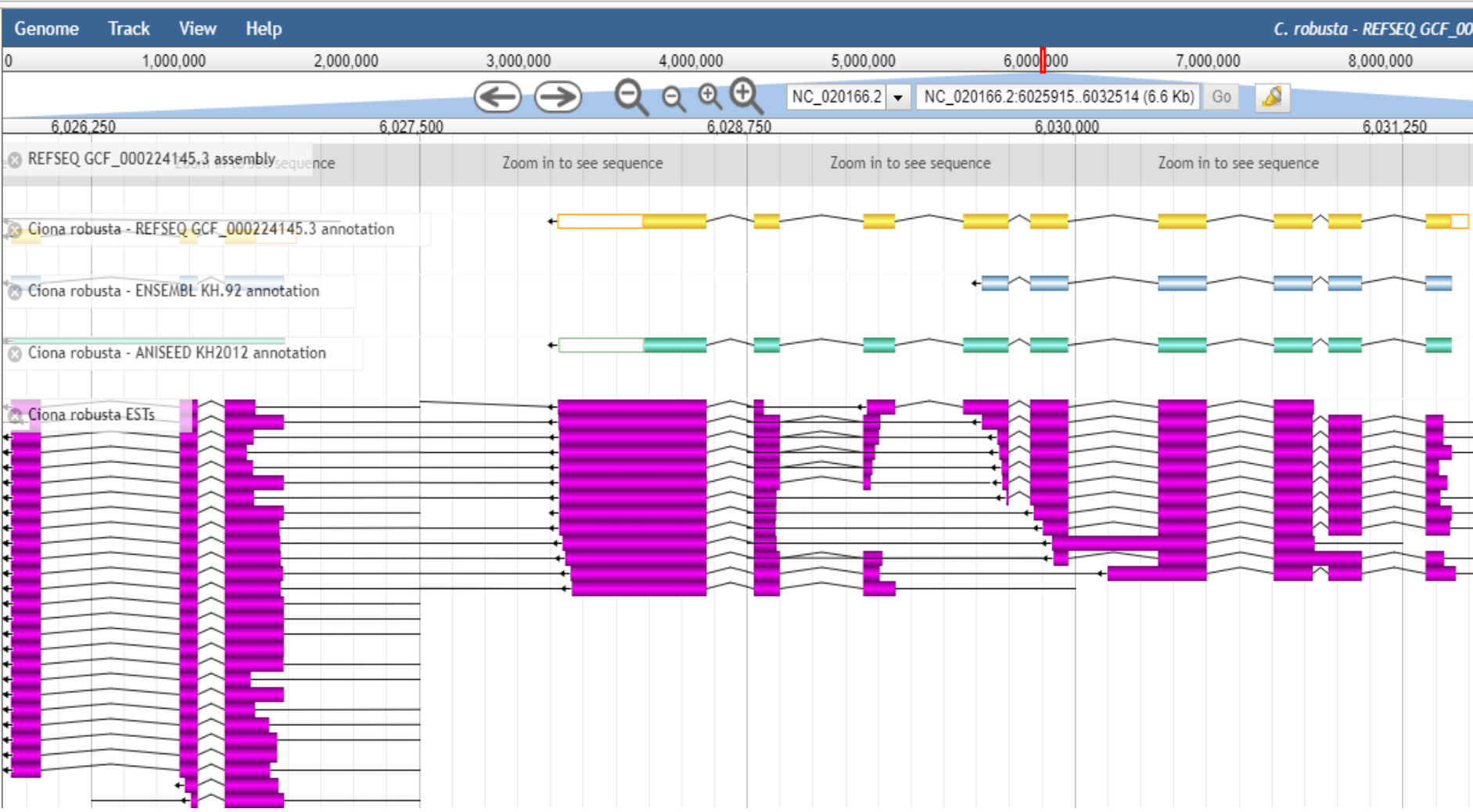
Browse the genome through the *JBrowse* interface



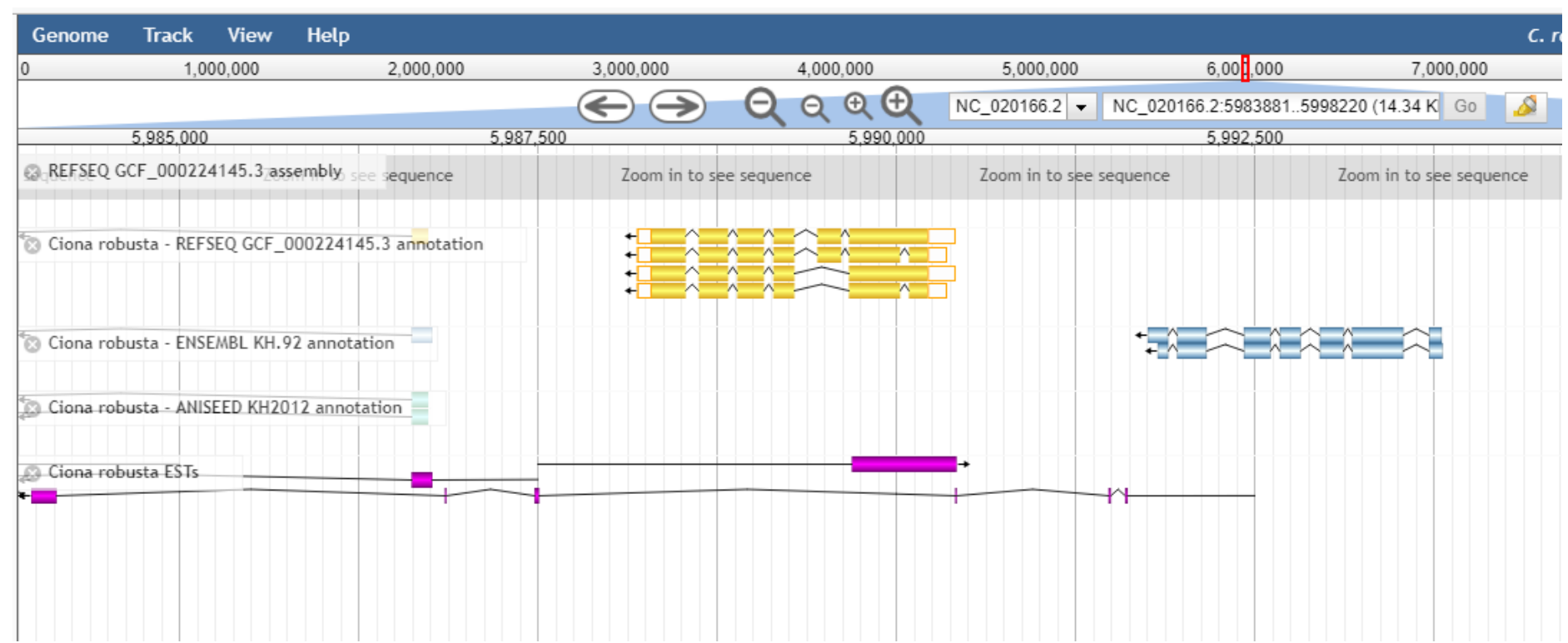
The GENOMA platform



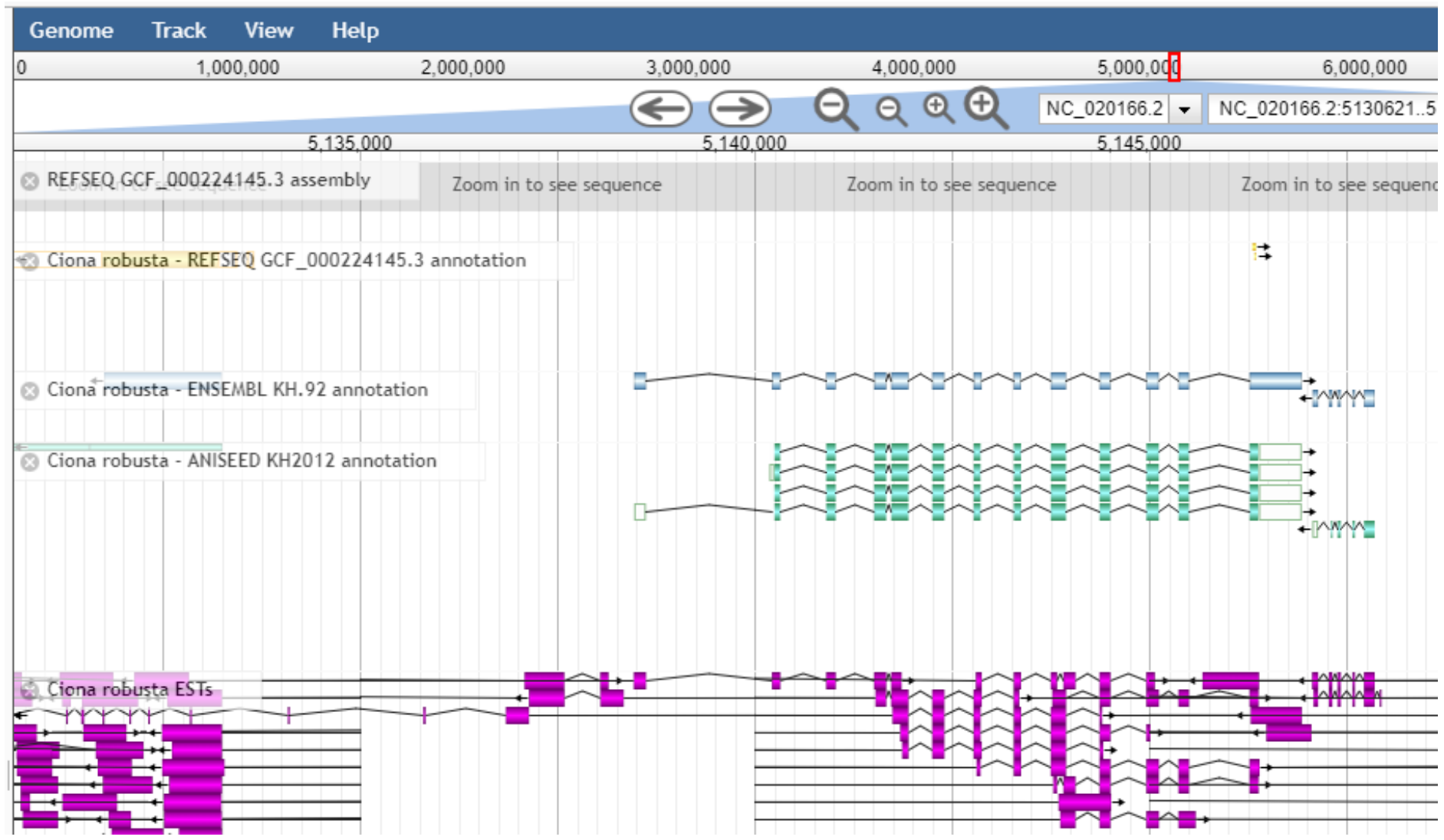
The GENOMA platform



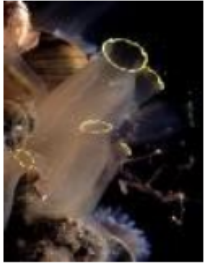
The GENOMA platform



The GENOMA platform



[back to the GENOMA home page](#)



Ciona robusta genome portal

Browse the genome through the [JBrowse](#) interface

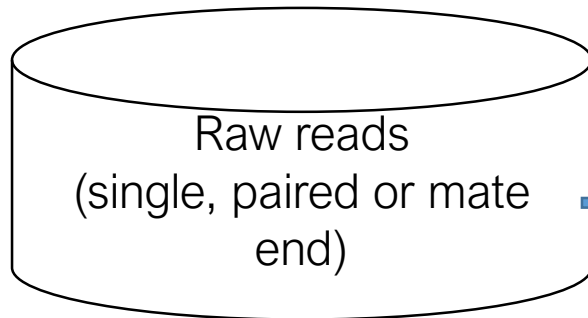
Search for specific elements through the [Query page](#)

Soon

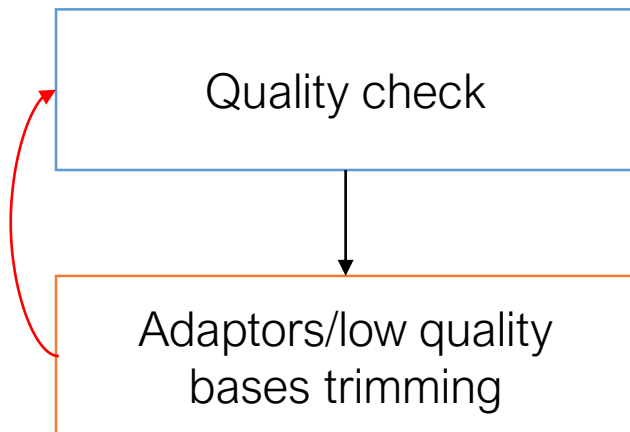


Experimental design
(conditions, stages...)

Sequencing

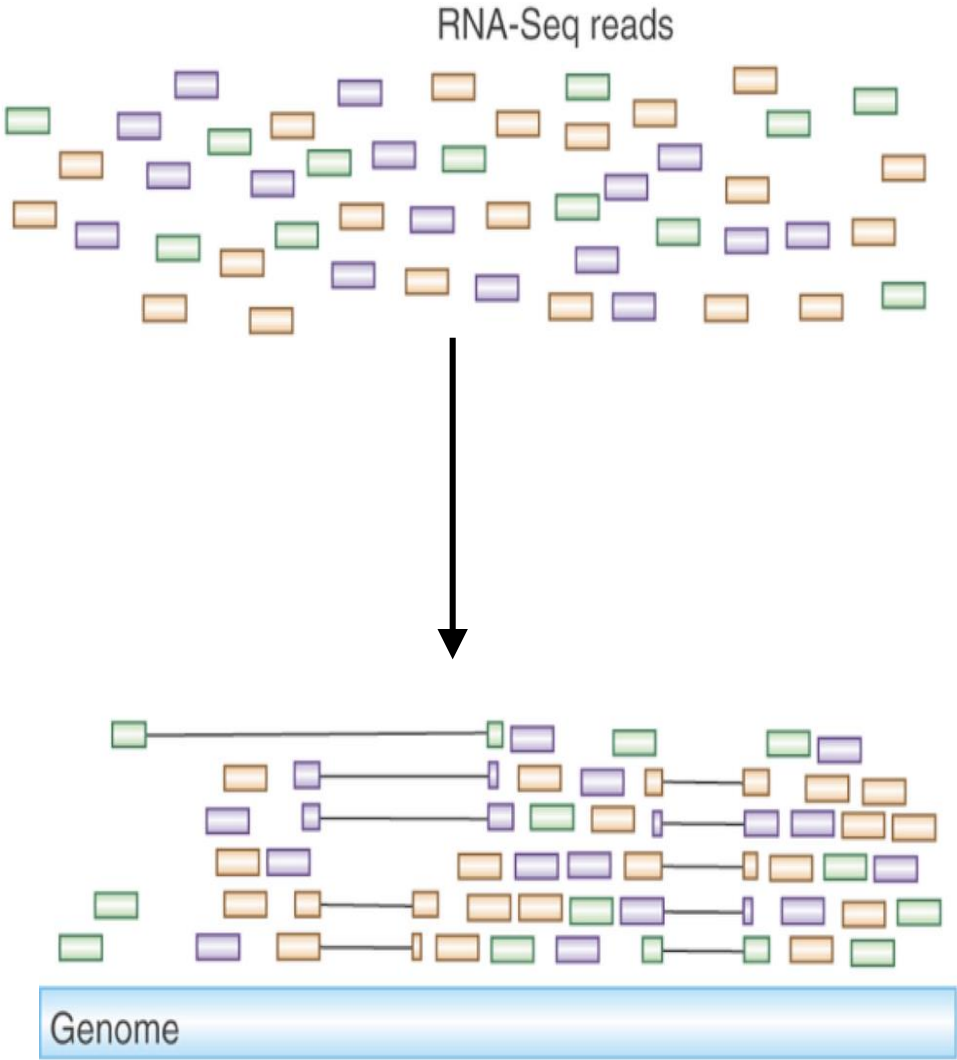


Data pre-processing



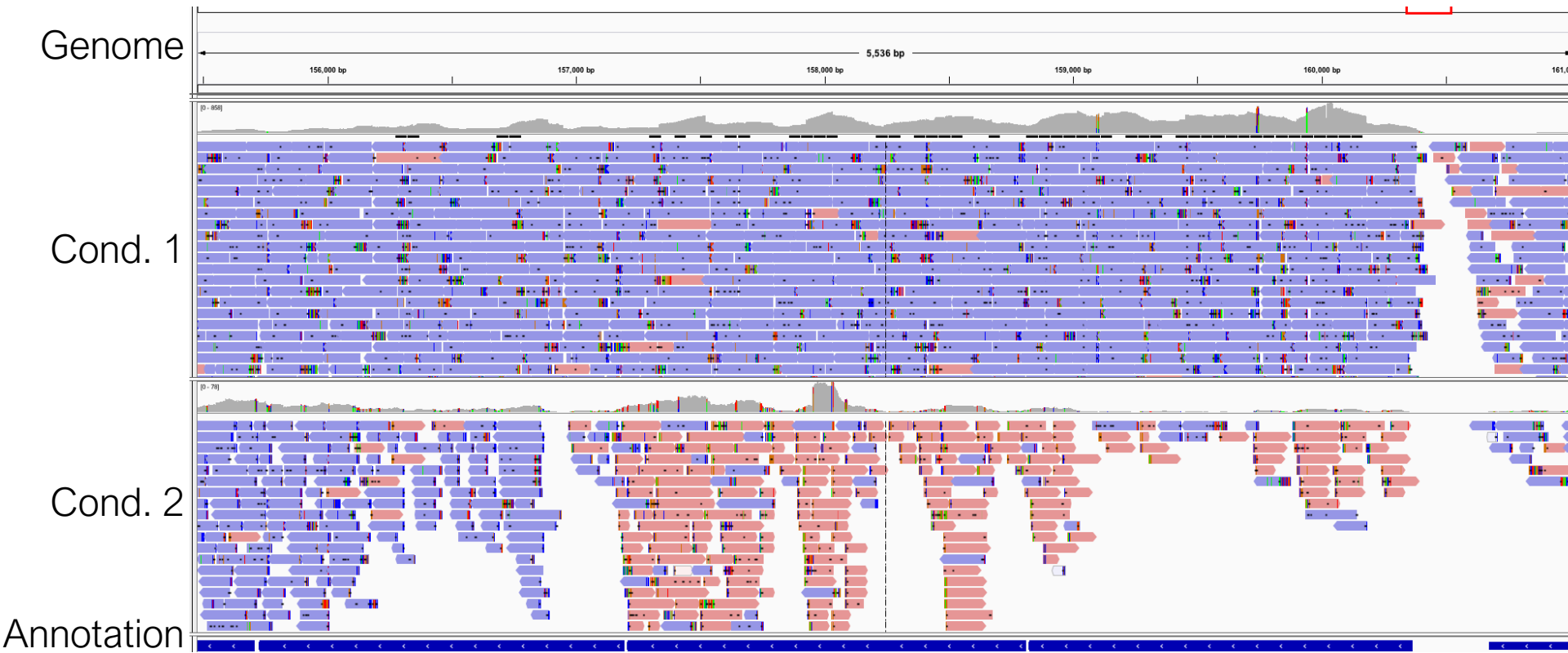
Data analyses

Reads mapping

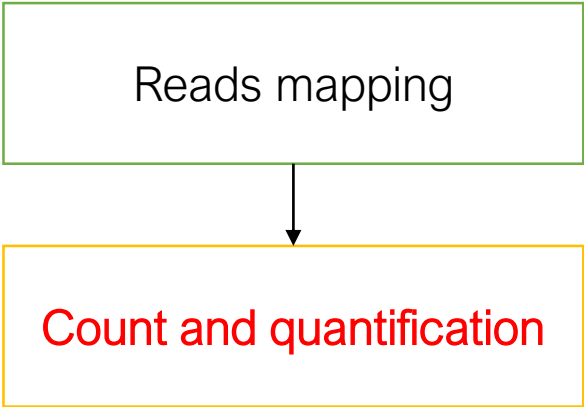


Data analyses

Reads mapping



Data analyses

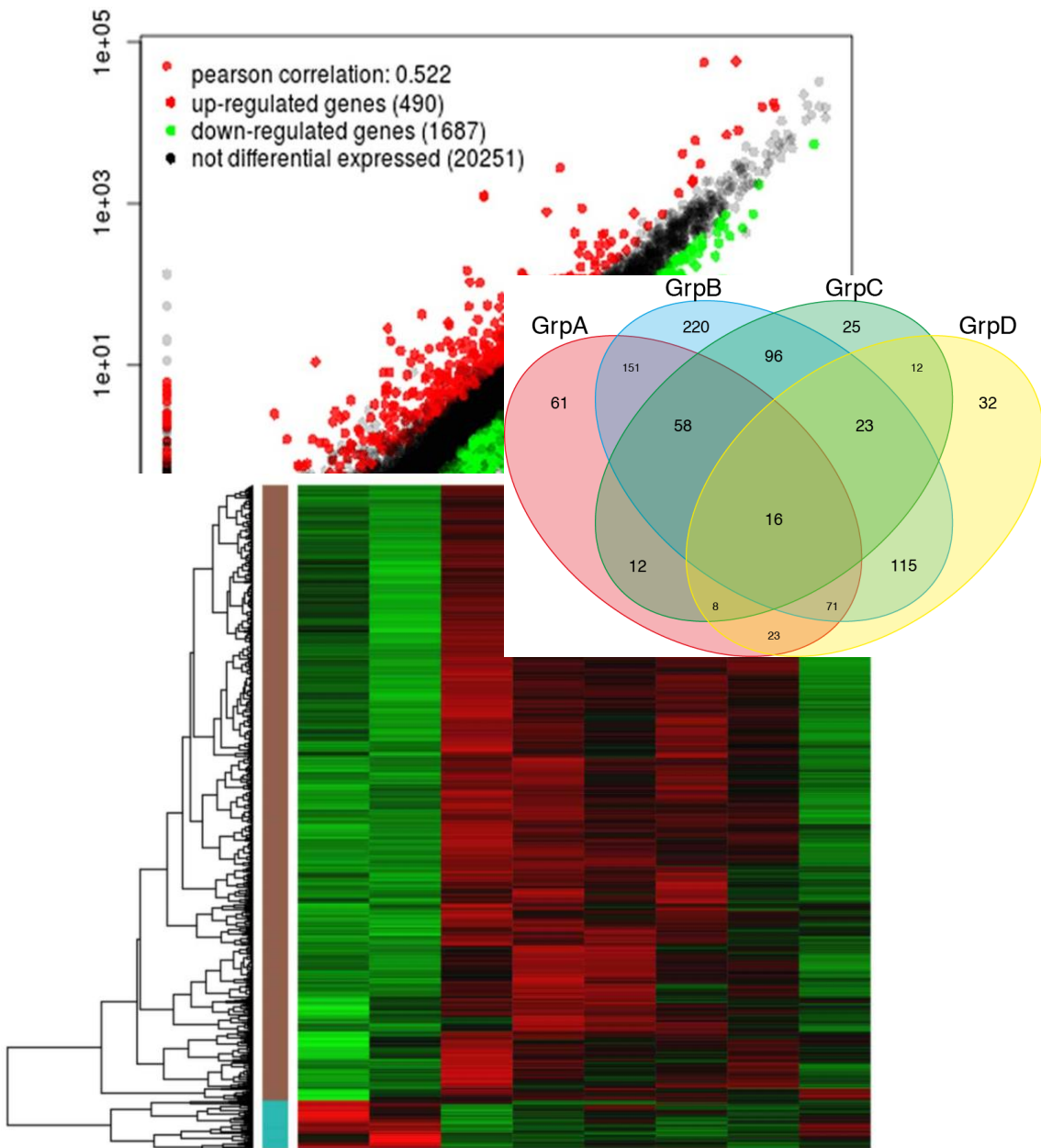
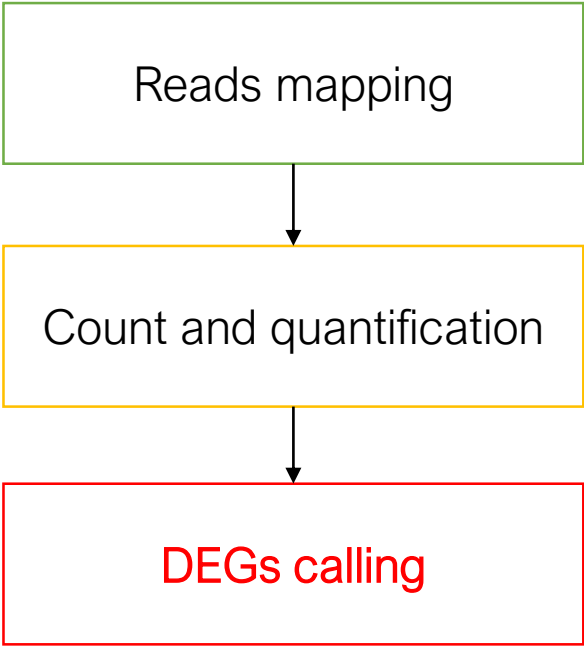


DESeq
edgeR
Cuffdiff

...

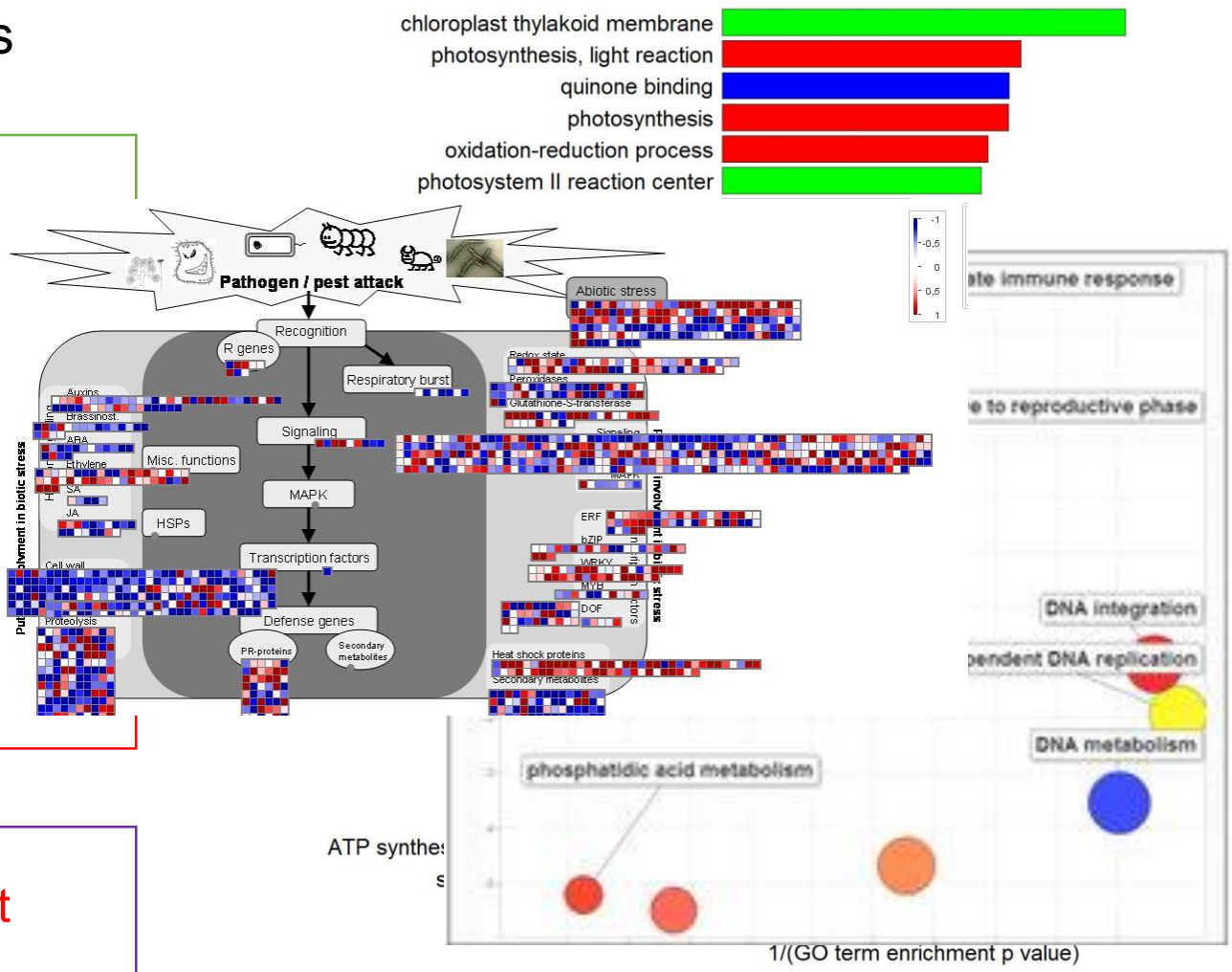
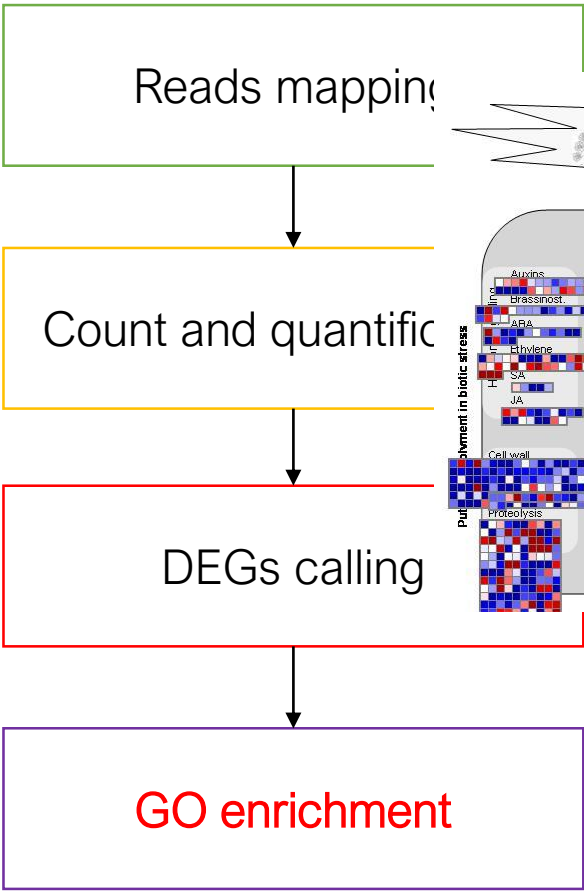
Geneid	C1	C2	C4	C5	C7	C8
AT1G01010.1	201	182	147	133	229	217
AT1G01020.1	224	218	229	206	187	167
AT1G01020.2	218	216	226	204	186	166
AT1G01030.1	75	94	47	49	69	64
AT1G01040.1	1017	1107	902	866	888	1083
AT1G01040.2	1001	1088	890	855	873	1072
AT1G01046.1	39	50	28	34	34	27
AT1G01050.1	763	810	732	685	695	805
AT1G01060.1	1442	1503	17	16	6716	9911
AT1G01060.2	1442	1505	17	16	6718	9913
AT1G01060.3	1357	1427	15	15	6321	9431
AT1G01060.4	1442	1503	17	16	6716	9911
AT1G01060.5	1376	1441	16	15	6401	9519
AT1G01070.1	192	160	117	104	73	93
AT1G01070.2	192	160	117	104	73	93
AT1G01073.1	0	0	0	0	0	1
AT1G01080.1	323	324	357	310	231	237
AT1G01080.2	324	324	357	310	231	237
AT1G01090.1	1735	1790	1752	1542	1698	1823
AT1G01100.1	1398	1562	1602	1464	1381	1559
AT1G01100.2	1398	1563	1603	1463	1381	1558
AT1G01100.3	1398	1563	1603	1463	1381	1558
AT1G01100.4	1398	1563	1603	1463	1381	1558

Data analyses



Differentially Expressed Genes

Data analyses



Genomics

- Genome Composition (GC/CpG etc content)
- Gene annotation structure quality check (via GFF)
- Gene-mRNA-protein sequences extraction from GFF
- GFF/GTF/BED files parsing and manipulation
- SAM file parsing and manipulation
- Genome sequence extraction
- Reads quality check and cleaning
- Reads mapping

Transcriptomics

- Reads quality check and cleaning
- Reads/ESTs mapping
- Reads count and RPKM calculation
- DEGs calling
- SAM file parsing and manipulation

...available soon

- Assembly & clustering
- Genome masking
- Preliminar genome gene structure annotation
- Preliminar genome repeats annotation
- Variant calling

<https://bioinfo.szn.it>

<https://ticketing.bioinfo.szn.it>

bioinforma@szn.it

