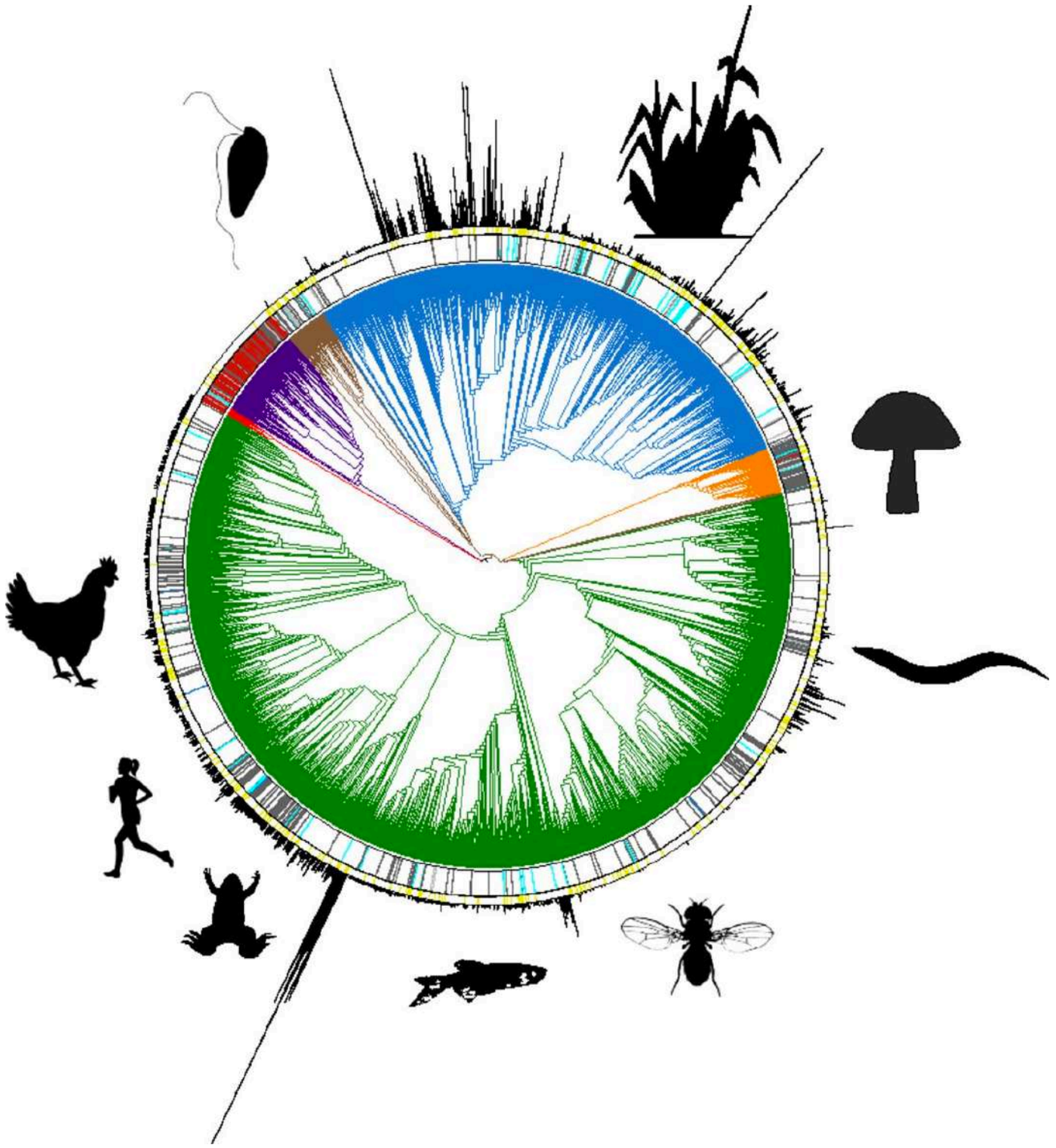
COMPARATIVE GENOMICS:
CHANGE OF GENOME ARCHITECTURES DURING ANIMAL EVOLUTION



CONTACTS

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Andrea Luchetti

Associate Professor

Department of Biological, Geological, and Environmental Sciences

Academic discipline: BIO/05 Zoology

Short Bio

Evolutionary biologist and zoologist, my research activity focuses on molecular phylogeny and phylogeography, comparative genomics and the evolution of transposable elements in the animal genome. Go to the Curriculum vitae

Contacts

E-mail: andrea.luchetti@unibo.it

Tel: +39 051 20 9 4165

Aims of the Talk

A <u>VERY</u> short (and incomplete) introduction to:

- Sequencing technologies
- Sequencing approaches/strategies
- Technical/computational challenges
- Data analysis pipelines
- Genome anatomy
- Genome comparison

Genomics Timeline

- 1975 – Sanger sequencing
- 1995 – 1st genome sequenced: *Haemophilus influenzae*
- 1996 – 1st eukaryote sequenced: *Saccharomyces cerevisiae*
- 1996 – 1st archaeon sequenced: *Metanococcus janneschii* (sampled @ 2600m of depth in the Pacific Ocean)
- 1998 – 1st metazoan sequenced: *Caenorhabditis elegans*
- 1999 – *Drosophila melanogaster genome sequenced*
- 2000 – 1st draft of the human genome completed
- 2001 – Publication of the human genome
- 2002 – *Mus musculus* sequenced
- 2004 – *Rattus norvegicus* sequenced
- 2005 – Chimpanzee genome sequenced
- 2008 – 1KGP starts
- 2009 – Genome 10K project established
- 2011 – i5k Project established
- 2013 – 1st GIGA workshop
- 2017 – Vertebrate Genome Project started
- 2018 – Earth BioGenome project launched

Sequencing Technology Timeline

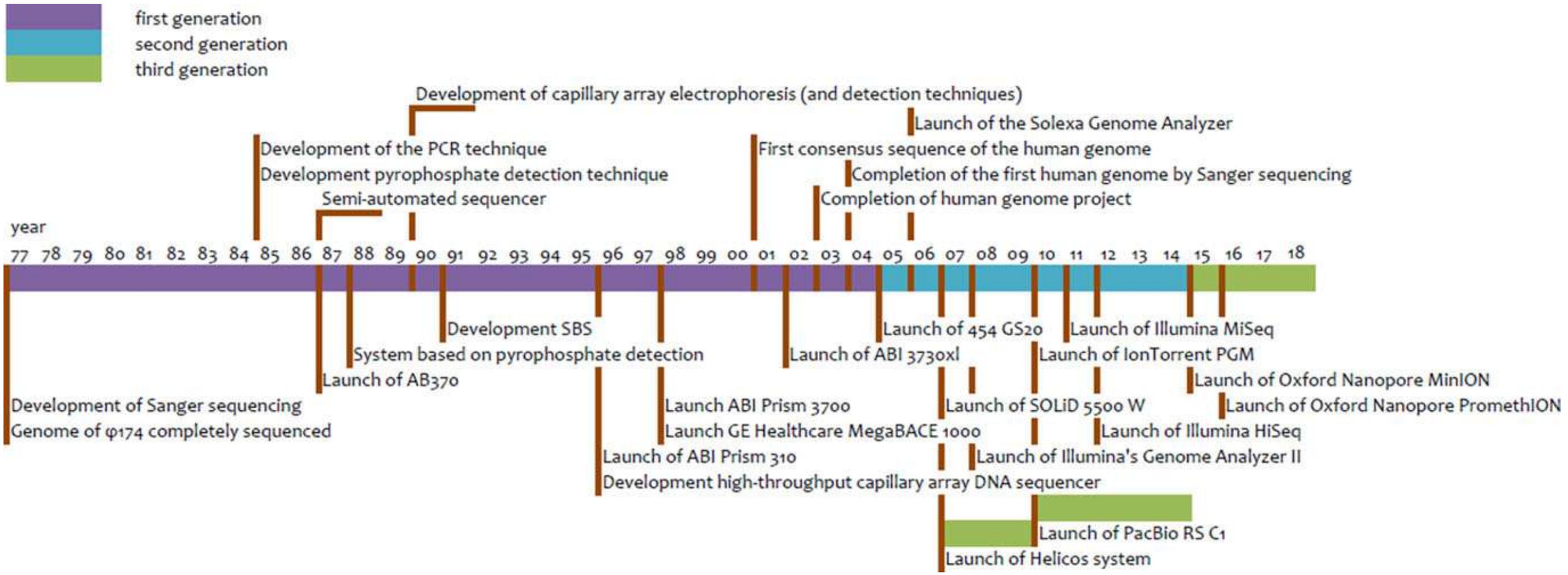


Table 1. Characteristics of the different sequencing techniques (first, second, and third generation). The read length range covers the read length at the introduction of the technique till the read length which can be obtained nowadays. Throughput and run time are for the machine that has currently the highest capacity. Throughput, reads and sequence data; AB, Applied Biosystems™; em-PCR; LT, Life Technologies; SBS, sequencing by synthesis [1, 3, 5, 7–9, 15–17]

Generation	Method	Launch	Technique	Read length (nt)	Throughput and run time	Comments
I	Sanger	1977	Cloning/chain termination	25–1200	96, 84 Kb, 2 h	First commercialized by AB (now LT)
II	454	2005	em-PCR/SBS/pyrosequencing	100–1000	1 million, 0.7 Gb, 24 h	Purchased by Roche in 2007
	Solexa/HiSeq®/MiSeq®	2006	Bridge PCR/SBS/reverse termination	36–300	6 billion, 1.8 Tb, several days	Solexa purchased by Illumina® in 2007
	SOLiD®	2007	em-PCR/ligation/probes	35–75	6 billion, 320 Gb, 1–2 weeks	Purchased by AB in 2006 (now LT)
	Ion Torrent™	2010	em-PCR/ion-sensitive SBS/pH change	200–400	60–80 million, 50 Gb, 2 h	Purchased by LT in 2010
III	PacBio®	2010	SMRT®/ZMW wells	8000–20000	350000, 7Gb, 0.5–6 h	
	(Oxford) nanopore	2014	Ion current shift	9545–200000	100000, 2–4 Tb up to 48 h	

GENOME 10K©

BAT 1K

GLOBAL INVERTEBRATE GENOMICS ALLIANCE

i5K

EARTH BIOGENOME PROJECT

sequencing life for the future of life

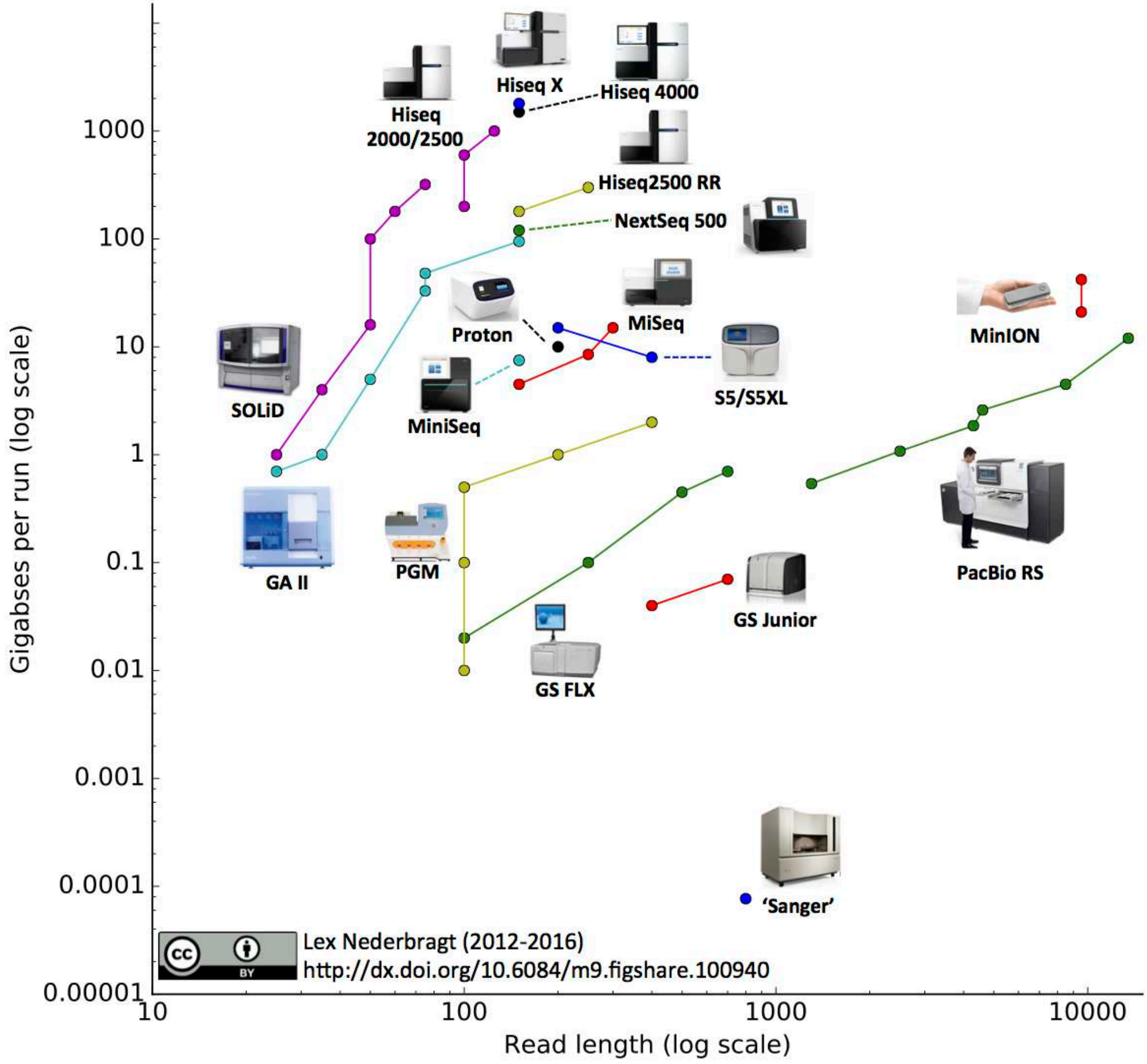
B10K

VGP

VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

Sequencing Technologies: Overview



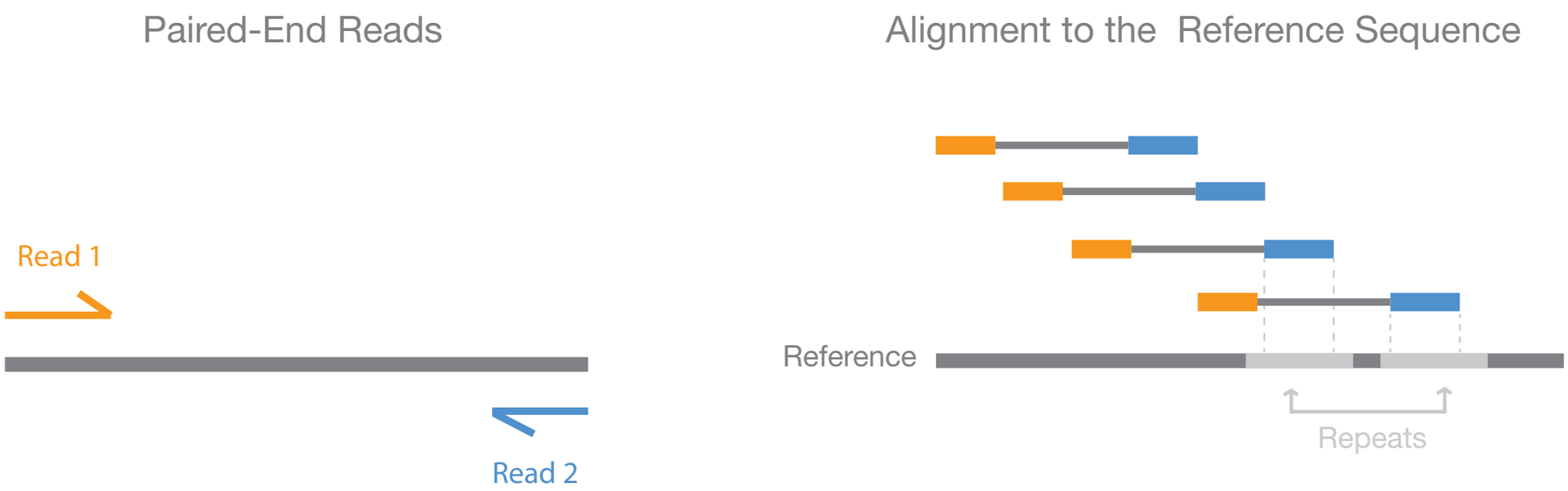
https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/

Sequencing Modes

- Single-end sequencing

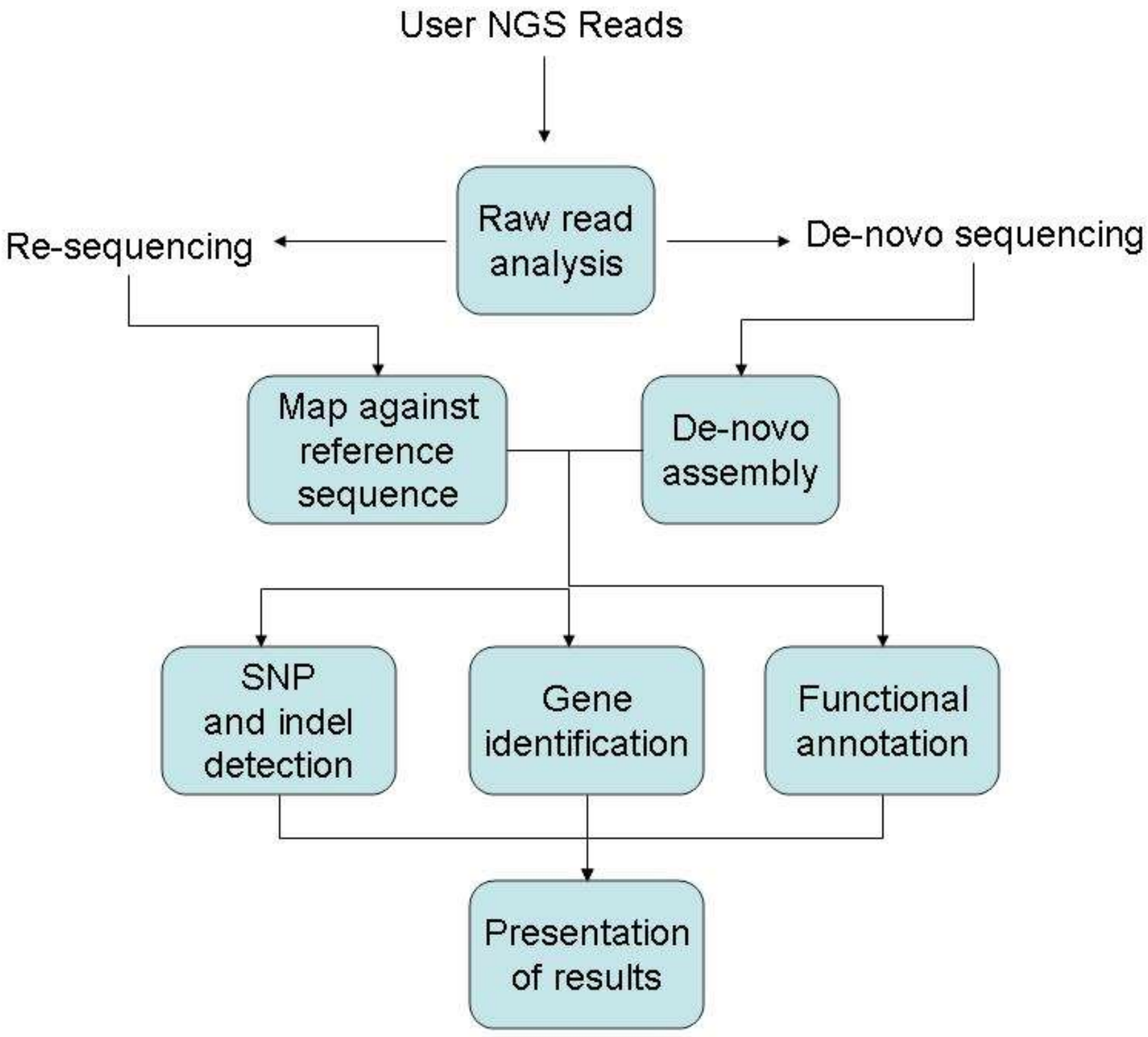
- Paired-end sequencing

Paired-End Reads

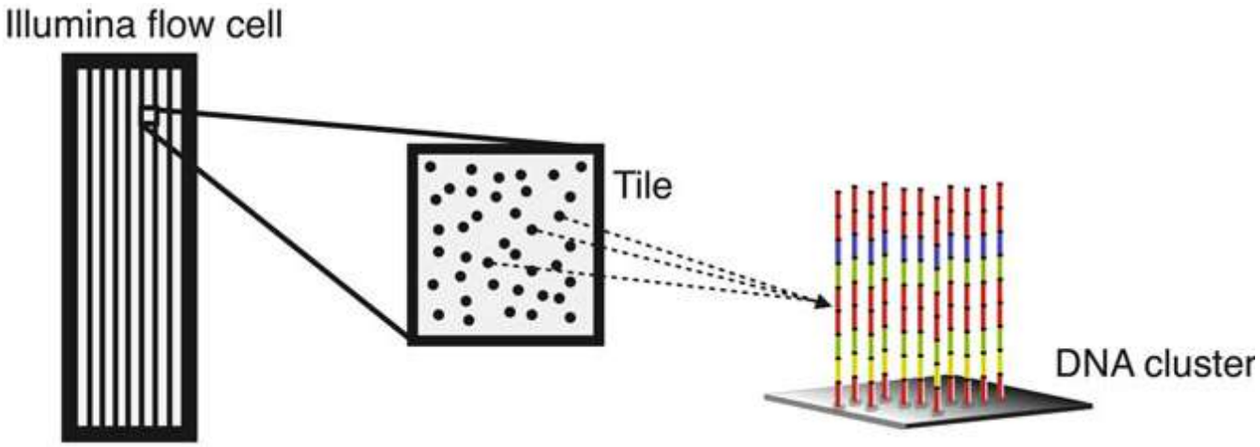


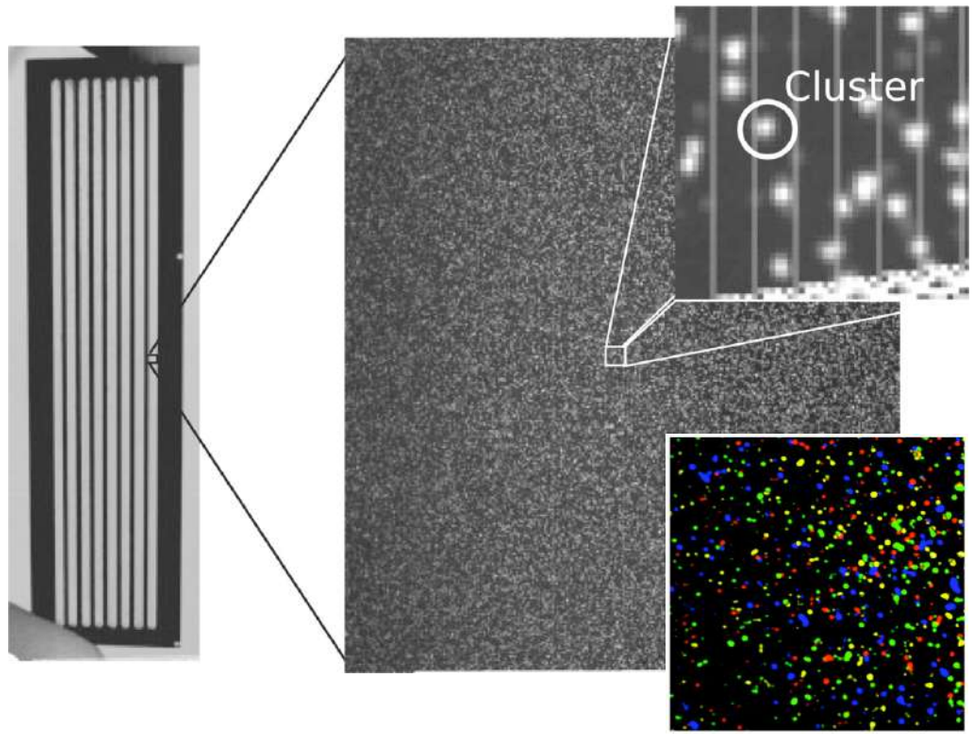
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

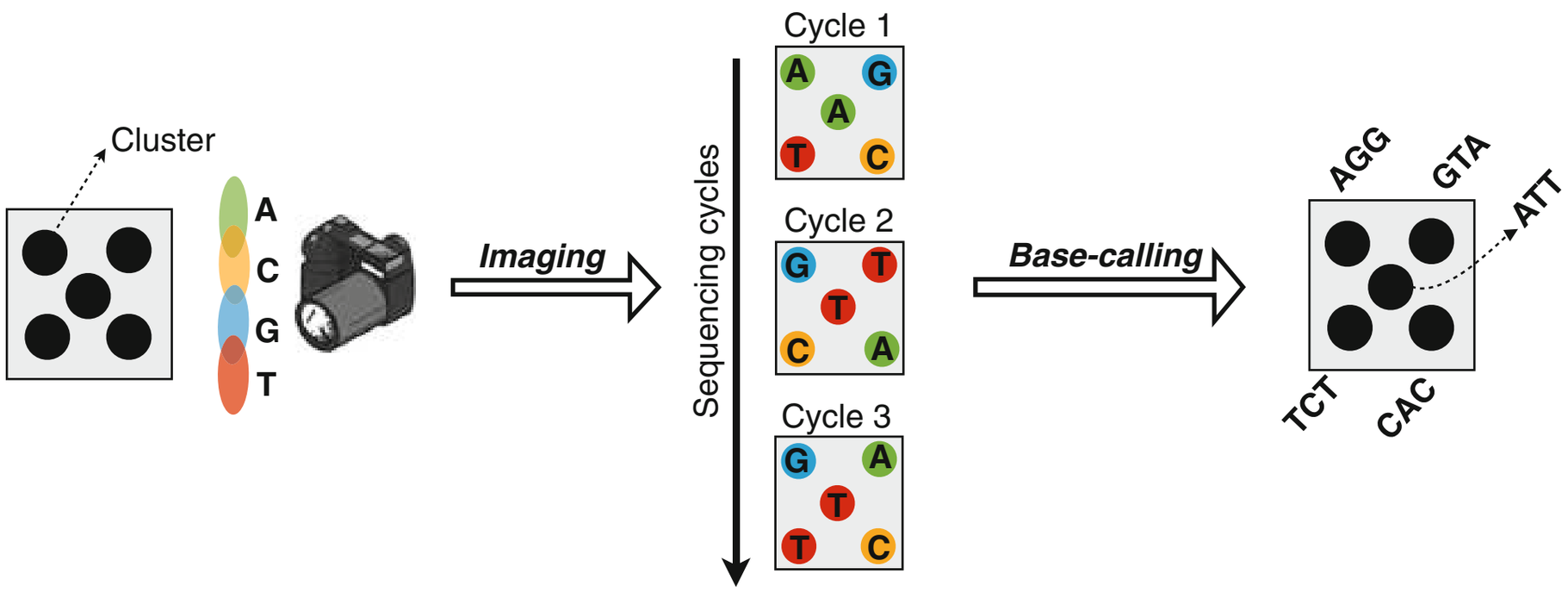
MPS Workflow



Output: Imaging/Base Calling



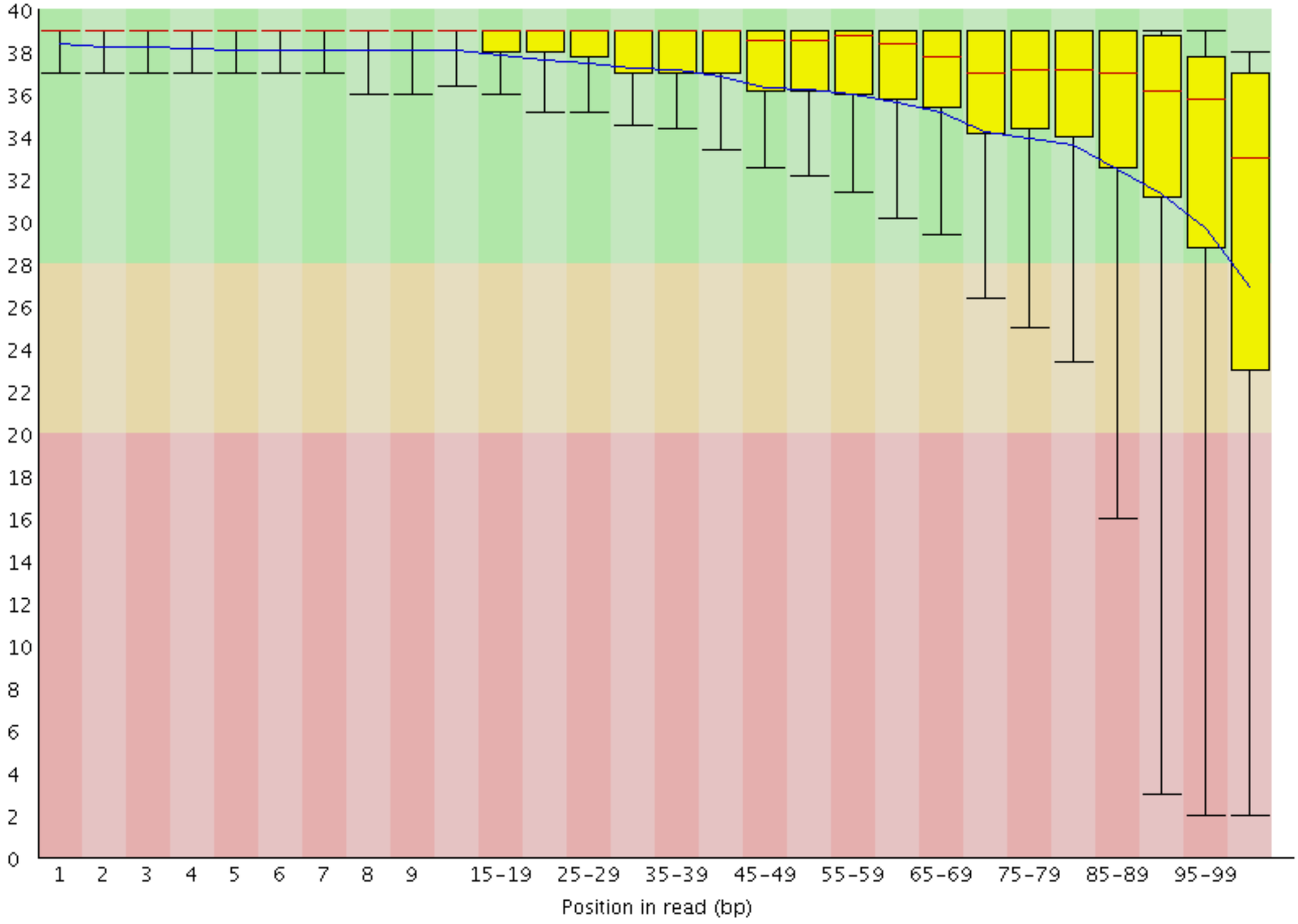




Output: Fastq Files

```
@HWI-ST867_0034:2:1101:1250:2103#CGATGT/1
TGTCCGGCCGCACACATCTAGAGCATCCAGGTCAGCGATCAGCCCCGAGAGCACACAGGATAAGTGCCTGCAGGTCTCATTGTTGTTCACACCCATCAGA
+HWI-ST867_0034:2:1101:1250:2103#CGATGT/1
fffeffefcfaeffffeffffcffff_ffeffffff\deeUaedeeedffffeffd_ce]dfbeffead\ad`ded_effdffddcdffffdcdd^dc
@HWI-ST867_0034:2:1101:1220:2108#CGANGT/1
GAAAGCAGACCTATAAAATATCACATAGTAAGATATTTATATTTAGATTTTTCTTATTTAGAATCTTCATCTGTAATGTATGATTTTGAAAATTAATTCT
+HWI-ST867_0034:2:1101:1220:2108#CGANGT/1
ggggggggegggggggggggggggeeegegegggggaedgggggggggegggggeggegggggegegggcdgfffffgggggfggegggdagefebf
@HWI-ST867_0034:2:1101:1204:2134#CGANGT/1
TGCCATAAAATCTCAGGATTTCAAAGATACTGCTGGGCATTCAGTGACTGCAGAAGGGTTTAGTATTCAGGATCTAGTCCTCGGTACAAGCATTGAGCAG
+HWI-ST867_0034:2:1101:1204:2134#CGANGT/1
ggggggggggggggfgggggggggggggggggggggfggggfggfggggeggggggdggegdggggege^egegfgggffffdeefcceeeeegggae
@HWI-ST867_0034:2:1101:1245:2141#CGATGT/1
AGGAATACAAATTCAGCTTATTGAAGTATTGTCATGTAAGGAAAGATATTTTAATAAAATTGCTCTTCCATCAGTTCCTGTTACTAAGGGACTAAAGCTA
+HWI-ST867_0034:2:1101:1245:2141#CGATGT/1
ggggggggggggggggggggggfgggegggggggggggegggggggggfgggggggggggggggagggggggfgg`dffffdgggggggggffggeggggf_
@HWI-ST867_0034:2:1101:1189:2205#CGANGT/1
TAATGTAATTCATGTGTATTTCAATGCCTTTTTTTTTTCTTTTTTTTGGAAATATTTTTCGTAGACGTTGGAACTTGCCAGGAATCCAGCAATGATGCAGG
+HWI-ST867_0034:2:1101:1189:2205#CGANGT/1
ccccc_eeeefacff_a_acfffffffffdfffffffff_ffffffdd_fffebffefe^edadfebdcbZYcc`dddb__bbc]d_ddaeefffeYce
@HWI-ST867_0034:2:1101:1187:2232#CGANGT/1
CAGGAACTGACTACTACTAATTCTGATAAGGCTTCCTTTTAGCTTAATTTCTGAACCTCTTTCATGGGGTTGTTAGAATTAGCTAATGATTGCAAGGCTT
+HWI-ST867_0034:2:1101:1187:2232#CGANGT/1
feffdegggegegggggggggggggggggggggeggggeeggegeggegecUaddddgggegcbee_ccbebggggfgggggefge]egfdcggggd
...
...
...
...millions and millions of lines...
```

Per base sequence quality



De novo Assembly

Multiple (Unsequenced) Genome Copies

Read Generation

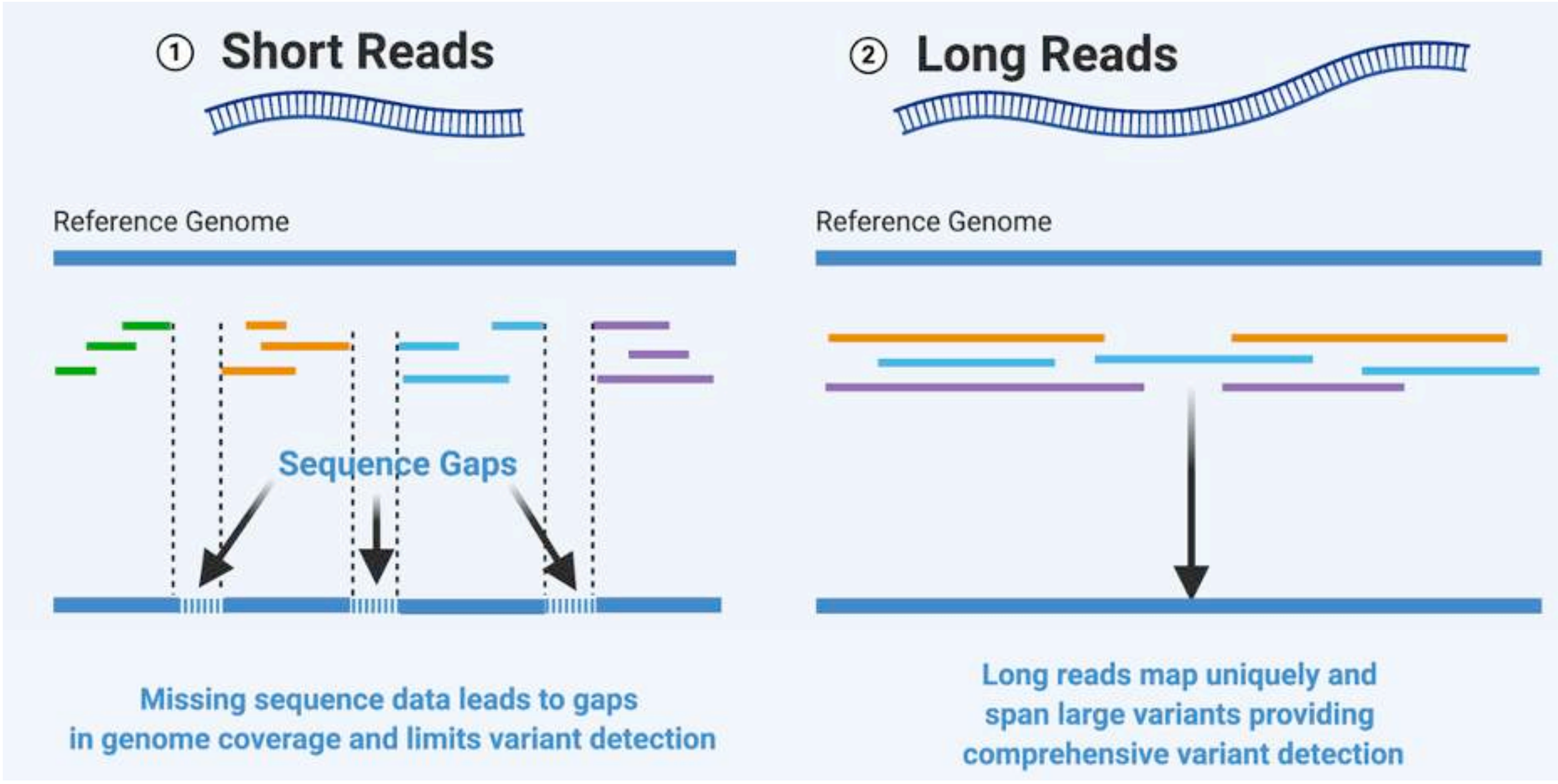
Reads

Fragment Assembly

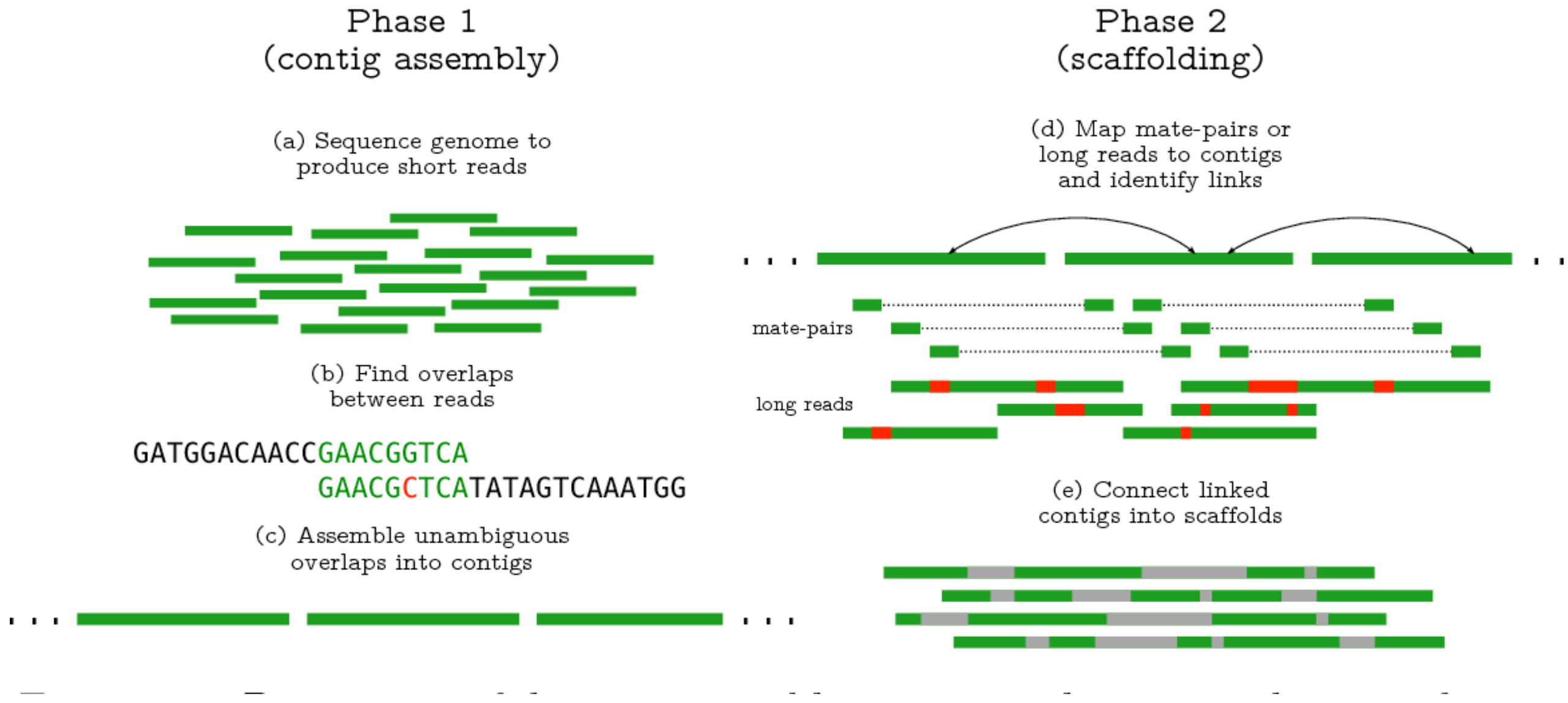
Sequenced Genome

…GGCATGCGTCAGAAACTATCATAGCTAGATCGTACGTAGCC…

De novo Assembly



De novo Assembly



Assembly Stats: N50, N90

- **N50** statistic defines assembly quality in terms of contiguity. Given a set of contigs, **the N50 is defined as the sequence length of the shortest contig at 50% of the total genome length.** It can be thought of as the point of half of the mass of the distribution.

- **N50** can be described as **a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value.**

- The **N90** statistic is the length for which the collection of all contigs of that length or longer contains at least 90% of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter contains at least 10% of the sum of the lengths of all contigs.

- Example:

# Scaffolds	Gsize	N50	N50_scaffold#	N90	N90_scaffold#
301516	2021935850	12068	46019	2414	170482

  - In this case, **50% of the assembly is represented by 46,019 scaffolds with length ≥ 12068bp.** 90% of the assembly is represented by 170,482 scaffolds with length ≥ 2414bp.

Assembly Quality

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics group

SIB

CEGG Home | OrthoDB *v9* | BUSCO *v3*

BUSCO v3

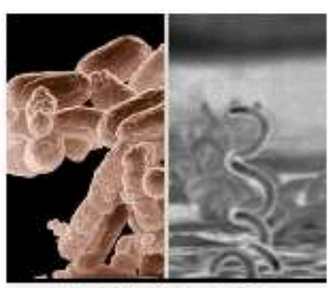
Assessing genome assembly and annotation completeness with
Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

About BUSCO

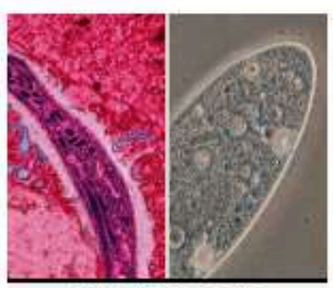
BUSCO *v3* provides quantitative measures for the assessment of genome assembly, gene set, and transcriptome completeness, based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB *v9*.

BUSCO assessments are implemented in open-source software, with a large selection of lineage-specific sets of Benchmarking Universal Single-Copy Orthologs. These conserved orthologs are ideal candidates for large-scale phylogenomics studies, and the annotated BUSCO gene models built during genome assessments provide a comprehensive gene predictor training set for use as part of genome annotation pipelines.

Datasets

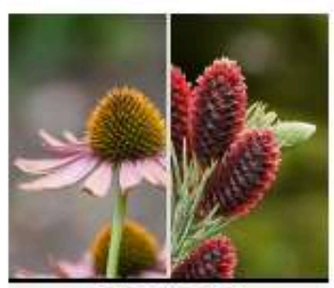

Bacteria sets


Eukaryota sets


Protists sets


Metazoa sets


Fungi sets


Plants set

Annotation

1. Sequence Similarity:

   - Nucleotide sequence (e.g. blastn)
   - Amino acid sequence (e.g. blastp)
   - HMM profile (es. hmmer)
   - Signatures (es. InterProScan)

2. Assignment of **GO Terms**.

Gene ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.

The project aims to:

1) maintain and develop its controlled vocabulary of gene and gene product attributes;
2) annotate genes and gene products, and assimilate and disseminate annotation data;
3) provide tools for easy access to all aspects of the data provided by the project, and to enable functional interpretation of experimental data using the GO, for example via enrichment analysis.

In computer science and information science, **an ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many or all domains**.

–Wikipedia

Annotation

The ontology covers three domains:

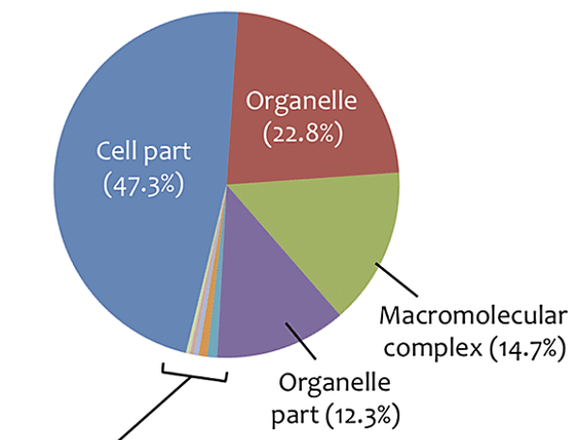
- **Cellular component**, the parts of a cell or its extracellular environment;
- **Molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- **Biological process**, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Each GO term within the ontology has a term name, which may be a word or string of words; a unique alphanumeric identifier; a definition with cited sources; and a namespace indicating the domain to which it belongs.
Terms may also have synonyms, which are classed as being exactly equivalent to the term name, broader, narrower, or related; references to equivalent concepts in other databases; and comments on term meaning or usage.
The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains.
The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.

Annotation

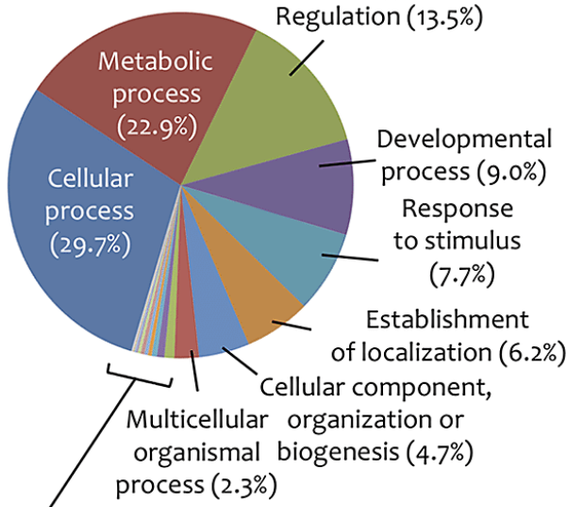
```
id:         GO:0000016
name:       lactase activity
namespace:  molecular_function
def:        "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]
synonym:    "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym:    "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref:       EC:3.2.1.108
xref:       MetaCyc:LACTASE-RXN
xref:       Reactome:20536
is_a:       GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds
```

Cellular Component



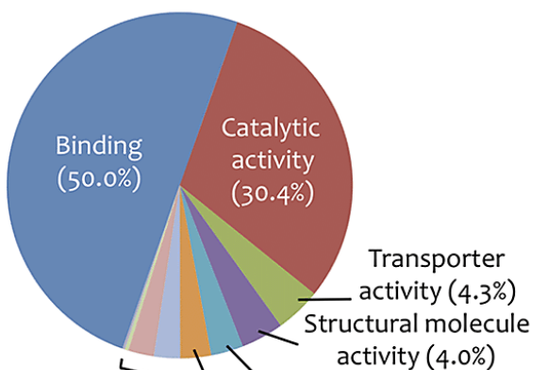
3.0%
- Extracellular region part (0.9%)
- Cell junction (0.8%)
- Membrane-enclosed lumen (0.5%)
- Synapse (0.4%)
- Synapse part (0.4%)

Biological Process



4.0%
- Locomotion (0.9%)
- Biological adhesion (0.7%)
- Death (0.5%)
- Growth (0.4%)
- Cell proliferation (0.3%)
- Localization (0.3%)
- Reproductive process (0.3%)
- Immune system process (0.2%)
- Pigmentation (0.2%)
- Multi-organism process (0.1%)
- Signaling (0.1%)
- Rhythmic process (0.03%)

Molecular Function



5.4%
- Molecular transducer activity (2.4%)
- Enzyme regulator activity (2.4%)
- Protein binding transcription factor activity (0.4%)
- Antioxidant activity (0.2%)

Why Sequencing Any Genome?

- Gain better understanding of biology
- Gain knowledge about genetic variation
- Allow the comparisons between taxa (understand evolution)
  - How much/How genes vary across taxa?
  - How much/How genome architecture changes?
  - How much/How gene content changes?
  - Reconstruct phylogenetic relationships and evolution of taxa
  - Reconstruct the evolutionary history of a species
- Biomedical/veterinary applications
- Forensic applications

Some MPS Applications

- **FUNCTIONAL GENOMICS**
  Understanding the function of genes and other genetic elements in a genome

- **COMPARATIVE GENOMICS**
  Comparing genomes of different organisms

- **POPULATION GENOMICS**
  Large-scale comparison of DNA sequences in populations

- **METAGENOMICS**
  Study of genetic material directly from environmental samples

- **CONSERVATION GENOMICS**
  Application of genomic analysis to the preservation of the viability of populations and the biodiversity of living organisms

Functional Genomics

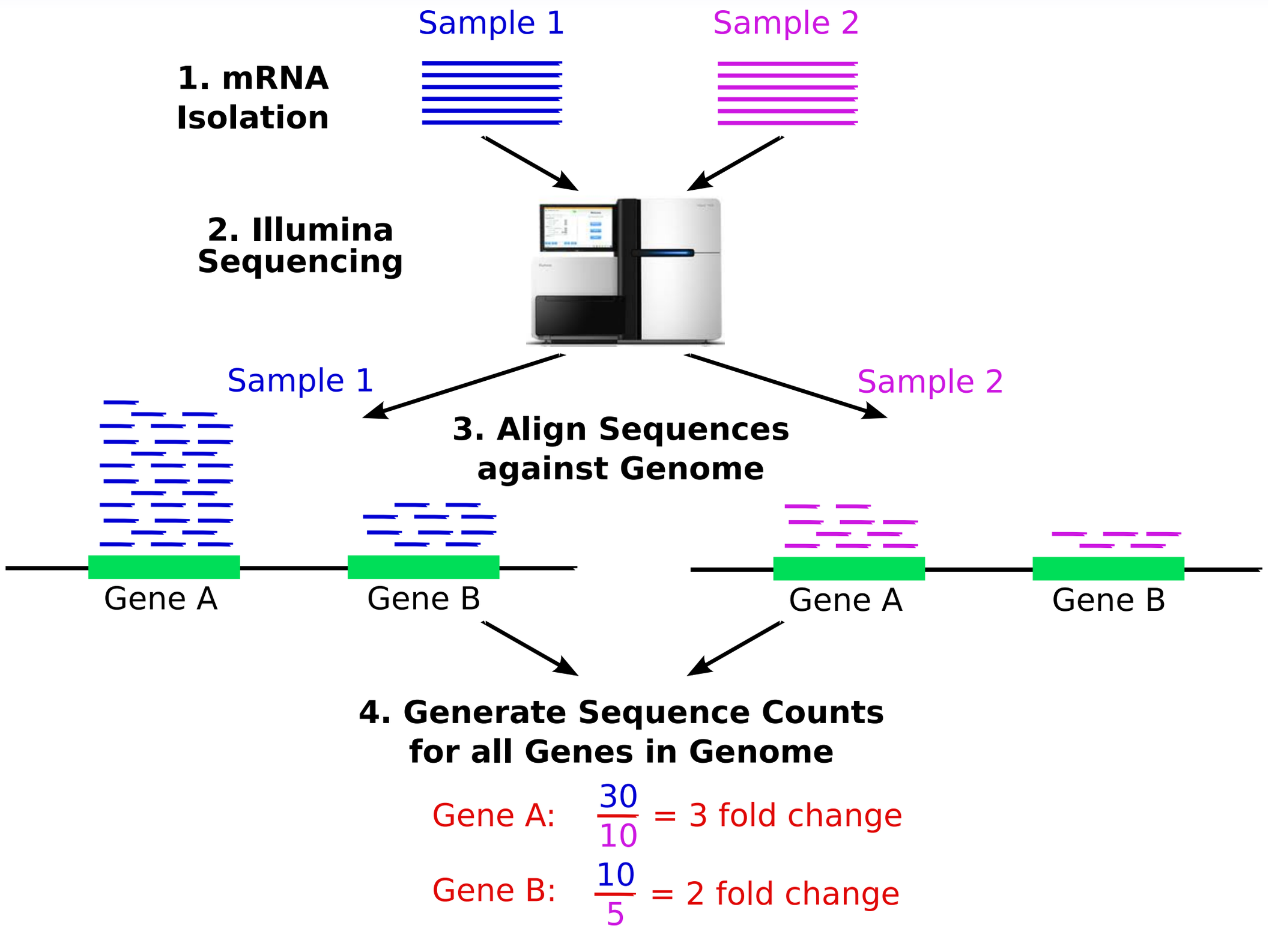
Use of the vast wealth of data given by genomic and transcriptomic projects to describe gene functions and interactions.

Focuses on the dynamic aspects such as gene transcription, translation, regulation of gene expression, and protein–protein interactions, as opposed to the static aspects of the genomic information such as DNA sequence or structures.

Attempts to answer questions about the function of genetic elemets at the levels of genes, RNA transcripts, and protein products.

A key characteristic of functional genomics studies is their genome-wide approach to these questions, generally involving high-throughput methods rather than a more traditional “gene-by-gene” approach.

RNA-Seq Technology



De Novo Assembly of the Manila Clam *Ruditapes philippinarum* Transcriptome Provides New Insights into Expression Bias, Mitochondrial Doubly Uniparental Inheritance and Sex Determination

Fabrizio Ghiselli,[1,*] Liliana Milani,[1] Peter L. Chang,[2] Dennis Hedgecock,[3] Jonathan P. Davis,[4] Sergey V. Nuzhdin,[2] and Marco Passamonti[1]

[1]Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Bologna, Italy

[2]Program in Molecular and Computational Biology, Department of Biological Sciences, University of Southern California

[3]Marine and Environmental Biology Section, Department of Biological Sciences, University of Southern California

[4]Taylor Shellfish Farms, Quilcene, Washington

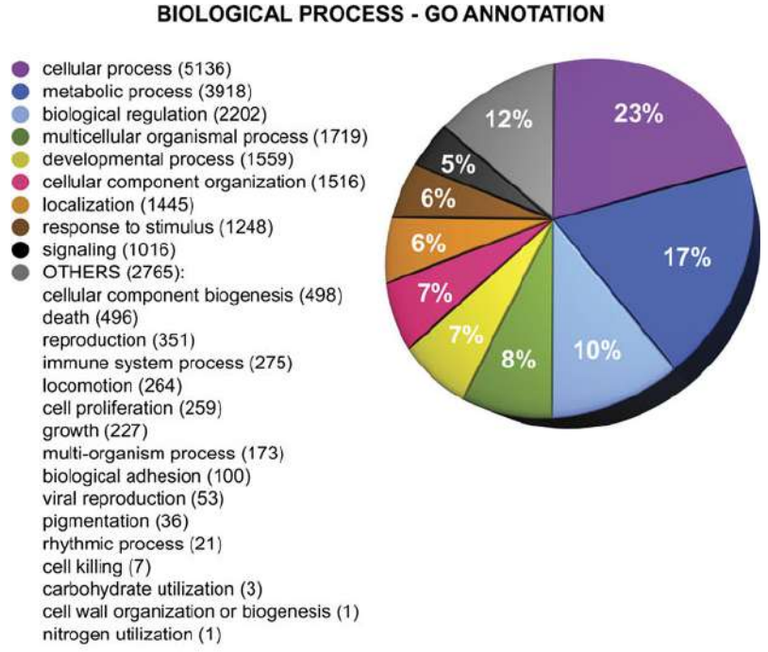
***Corresponding author:** E-mail: fabrizio.ghiselli@unibo.it.

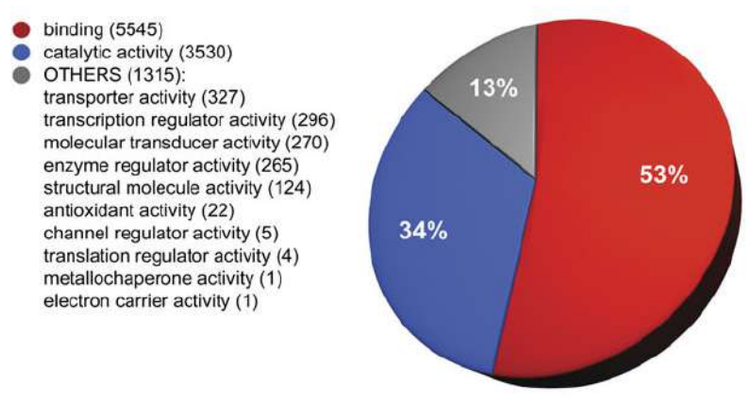
Associate editor: Richard Thomas

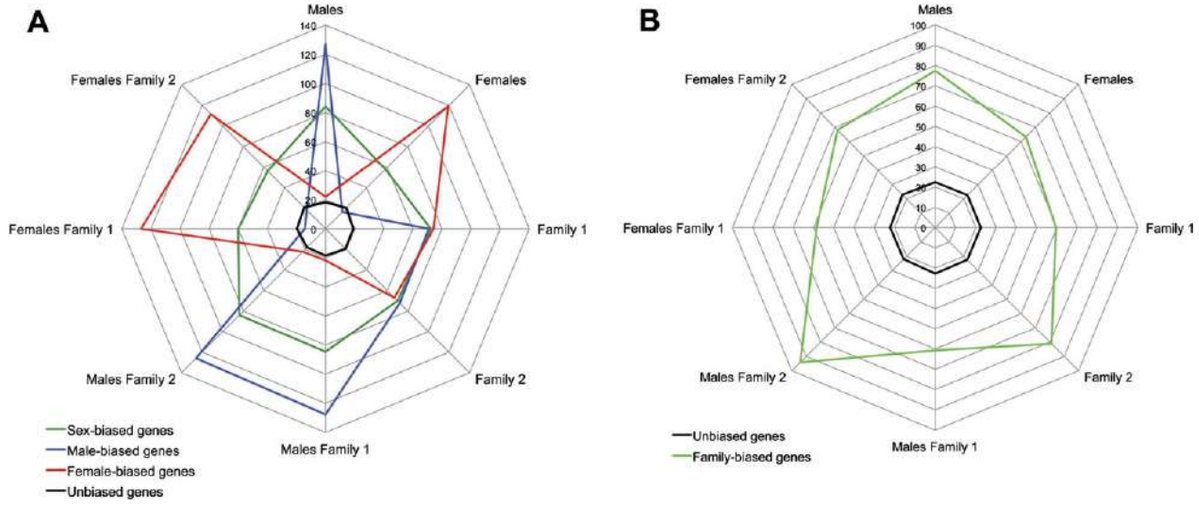
Abstract

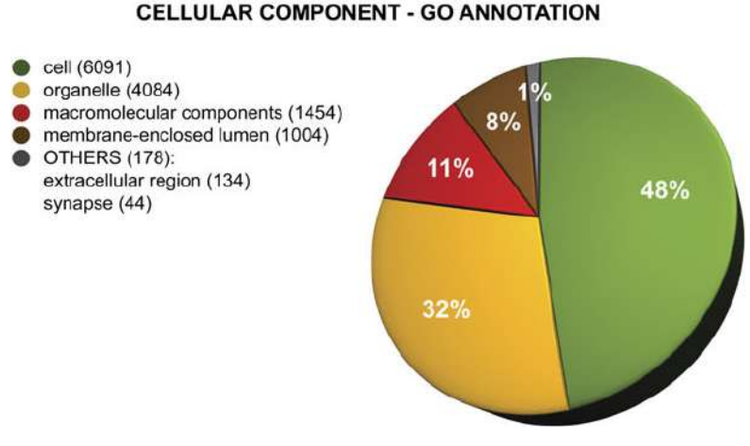
Males and females share the same genome, thus, phenotypic divergence requires differential gene expression and sex-specific regulation. Accordingly, the analysis of expression patterns is pivotal to the understanding of sex determination mechanisms. Many bivalves are stable gonochoric species, but the mechanism of gonad sexualization and the genes involved are still unknown. Moreover, during the period of sexual rest, a gonad is not present and sex cannot be determined. A mechanism associated with germ line differentiation in some bivalves, including the Manila clam *Ruditapes philippinarum*, is the doubly uniparental inheritance (DUI) of mitochondria, a variation of strict maternal inheritance. Two mitochondrial lineages are present, one transmitted through eggs and the other through sperm, as well as a mother-dependent sex bias of the progeny. We produced a de novo annotation of 17,186 transcripts from *R. philippinarum* and compared the transcriptomes of males and females and identified 1,575 genes with strong sex-specific expression and 166 sex-specific single nucleotide polymorphisms, obtaining preliminary information about genes that could be involved in sex determination. Then we compared the transcriptomes between a family producing predominantly females and a family producing predominantly males to identify candidate genes involved in regulation of sex-specific aspects of DUI system, finding a relationship between sex bias and differential expression of several ubiquitination genes. In mammalian embryos, sperm mitochondria are degraded by ubiquitination. A modification of this mechanism is hypothesized to be responsible for the retention of sperm mitochondria in male embryos of DUI species. Ubiquitination can additionally regulate gene expression, playing a role in sex determination of several animals. These data enable us to develop a model that incorporates both the DUI literature and our new findings.

Key words: *Ruditapes philippinarum*, de novo, transcriptome, doubly uniparental inheritance, sex bias, sex determination.









Comparative Transcriptomics in Two Bivalve Species Offers Different Perspectives on the Evolution of Sex-Biased Genes

Fabrizio Ghiselli[1,*,†], Mariangela Iannello[1,†], Guglielmo Puccio[1], Peter L. Chang[2], Federico Plazzi[1], Sergey V. Nuzhdin[2,†], and Marco Passamonti[1,†]

[1]Department of Biological, Geological, and Environmental Sciences, University of Bologna, Italy

[2]Program in Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, USA

†These authors contributed equally to this work.

*Corresponding author: E-mail:fabrizio.ghiselli@unibo.it.

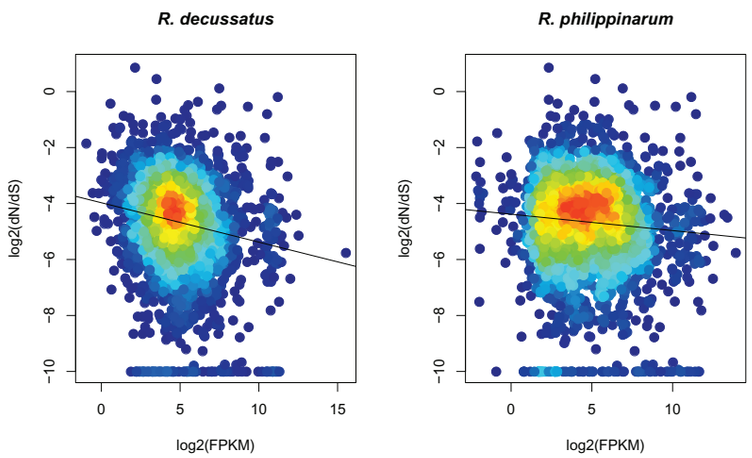
Accepted: April 19, 2018

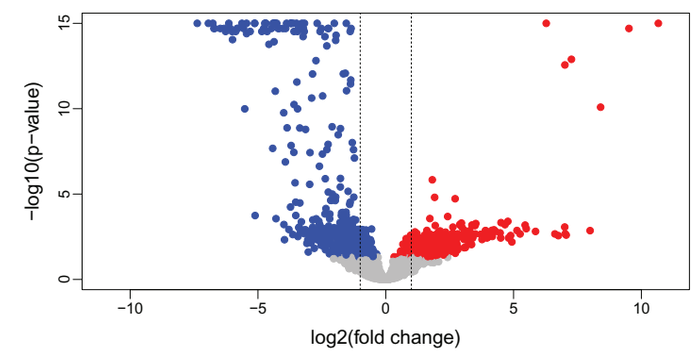
Data deposition: This project has been deposited at NCBI BioProject under the accessions PRJNA170478 and PRJNA68513; and on figshare, at the link: https://doi.org/10.6084/m9.figshare.5398618.v1

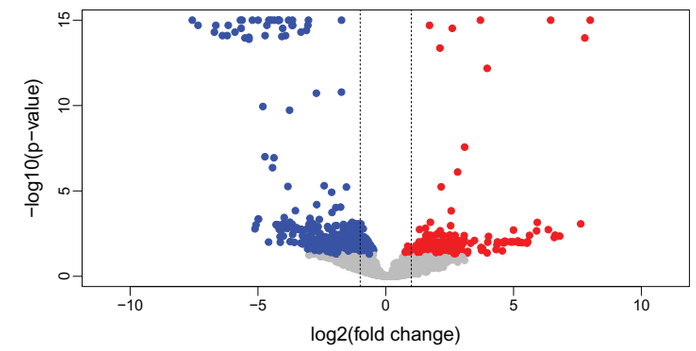
Abstract

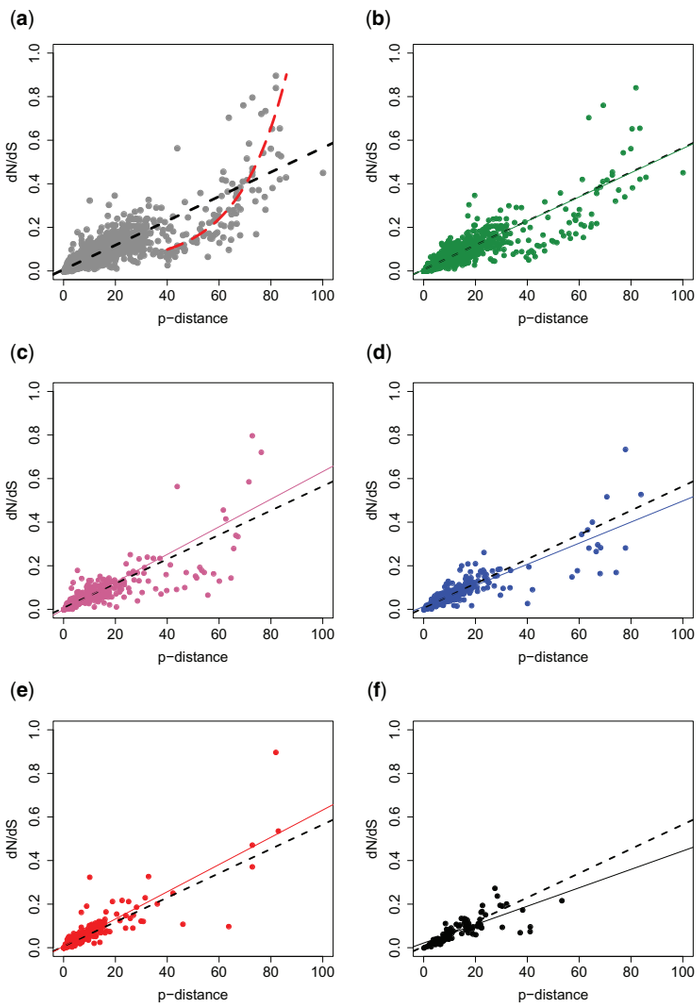
Comparative genomics has become a central tool for evolutionary biology, and a better knowledge of understudied taxa represents the foundation for future work. In this study, we characterized the transcriptome of male and female mature gonads in the European clam *Ruditapes decussatus*, compared with that in the Manila clam *Ruditapes philippinarum* providing, for the first time in bivalves, information about transcription dynamics and sequence evolution of sex-biased genes. In both the species, we found a relatively low number of sex-biased genes (1,284, corresponding to 41.3% of the orthologous genes between the two species), probably due to the absence of sexual dimorphism, and the transcriptional bias is maintained in only 33% of the orthologs. The d*N*/d*S* is generally low, indicating purifying selection, with genes where the female-biased transcription is maintained between the two species showing a significantly higher d*N*/d*S*. Genes involved in embryo development, cell proliferation, and maintenance of genome stability show a faster sequence evolution. Finally, we report a lack of clear correlation between transcription level and evolutionary rate in these species, in contrast with studies that reported a negative correlation. We discuss such discrepancy and call into question some methodological approaches and rationales generally used in this type of comparative studies.

Key words: RNA-Seq, transcription level, evolutionary rate, gametogenesis, embryo development, E–R correlation.

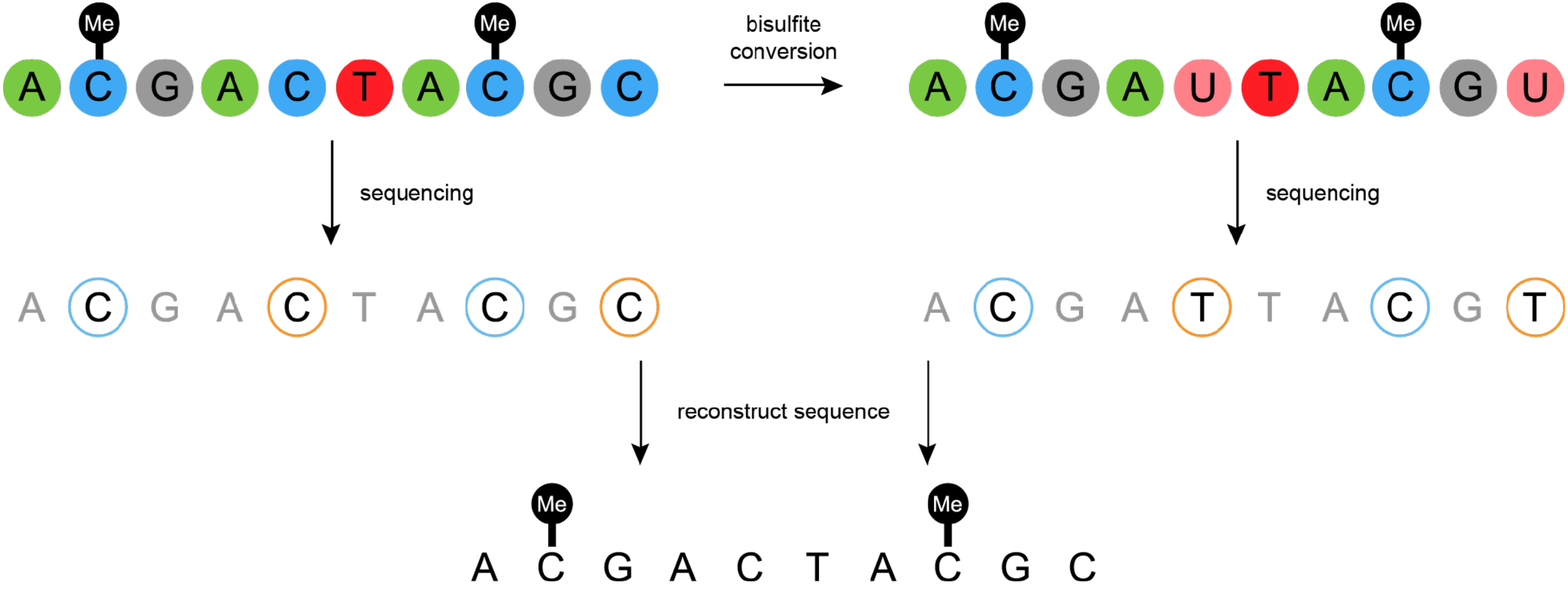




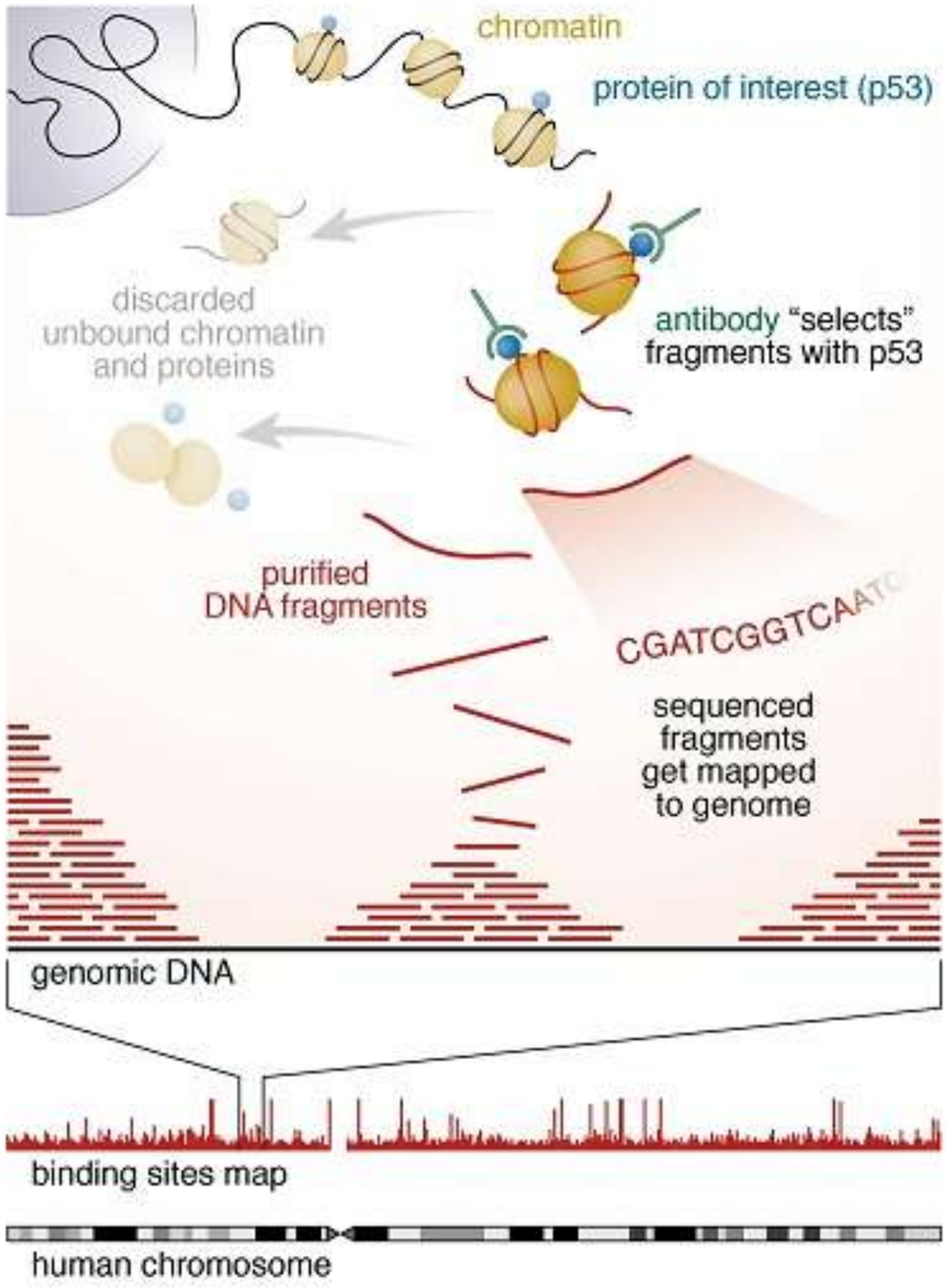




Epigenomics: Bisulfite Sequencing



ChIP-Seq


chromatin
protein of interest (p53)
discarded
unbound chromatin
and proteins
antibody "selects"
fragments with p53
purified
DNA fragments
CGATCGGTCAATC
sequenced
fragments
get mapped
to genome
genomic DNA
binding sites map
human chromosome

Comparative Genomics

Compares the genomic features (DNA sequence, genes, gene order, regulatory sequences, and other genomic structural landmarks) of different organisms.

Whole or large parts of genomes resulting from genome projects are compared to study basic biological similarities and differences as well as evolutionary relationships between organisms.

The major principle of comparative genomics is that common features of two organisms will often be encoded within the DNA that is evolutionarily conserved between them.

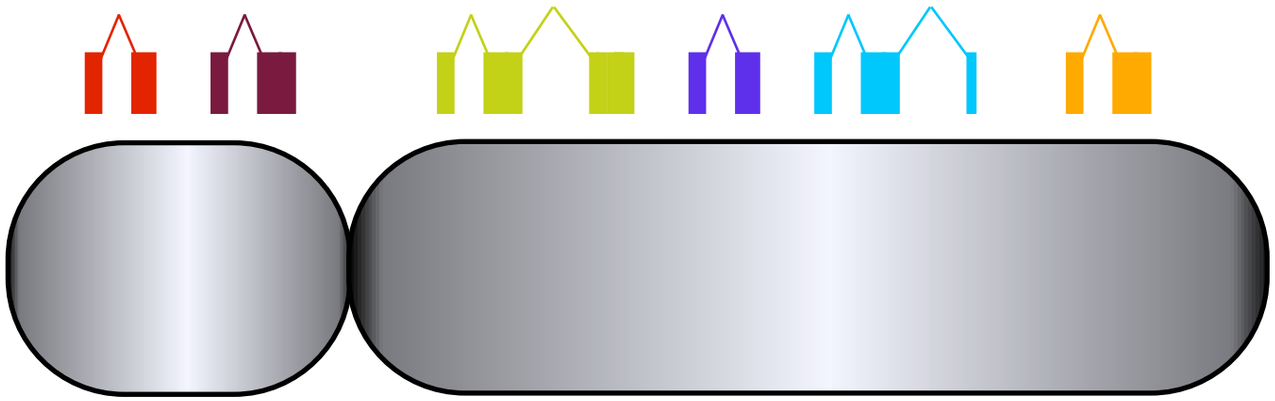
Comparative genomic approaches start with making some form of alignment of genome sequences and looking for orthologous sequences (sequences that share a common ancestry) in the aligned genomes and checking to what extent those sequences are conserved. Based on these, genome and molecular evolution are inferred and this may in turn be put in the context of, for example, phenotypic evolution or population genetics.

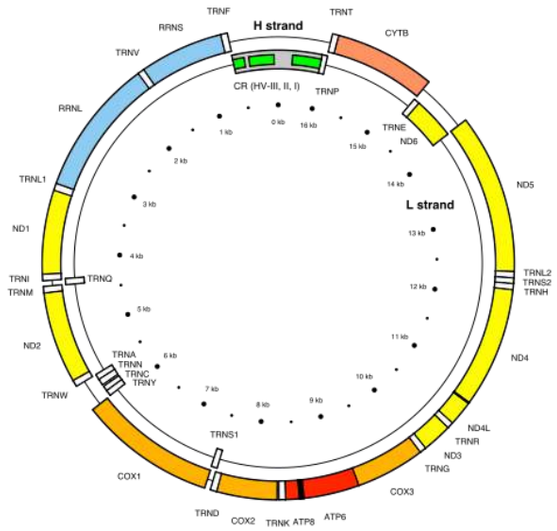
Genome anatomy

Genome

(from Wikipedia, the free encyclopedia)

In the fields of molecular biology and genetics, a genome is all genetic information of an organism.[1] It consists of nucleotide sequences of DNA (or RNA in RNA viruses). The genome includes both the genes (the coding regions) and the noncoding DNA, as well as mitochondrial DNA and chloroplast DNA.

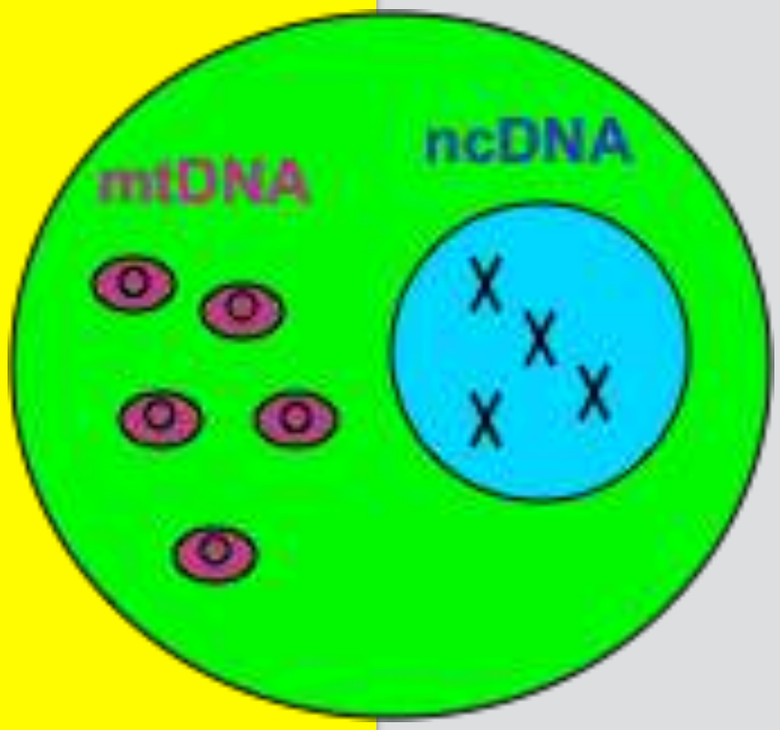




therefore the genome is plenty of genes...

Genome anatomy

mtDNA *vs* nDNA



more variable

no recombination (?)

haploid

compact genome

half of the story!!!

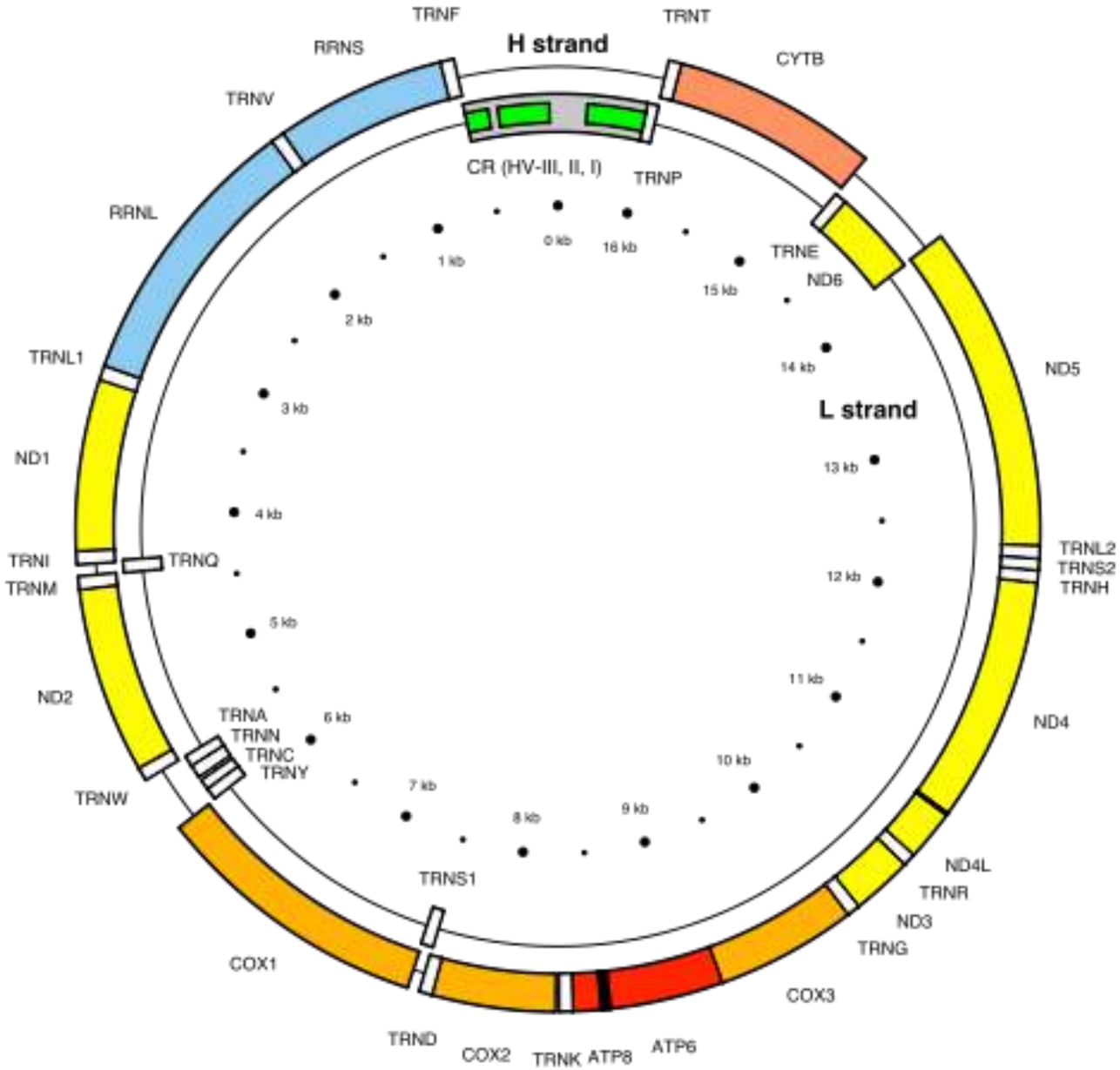
less variable

recombination

di(n)ploid

a mess!

Mitochondrial genome anatomy

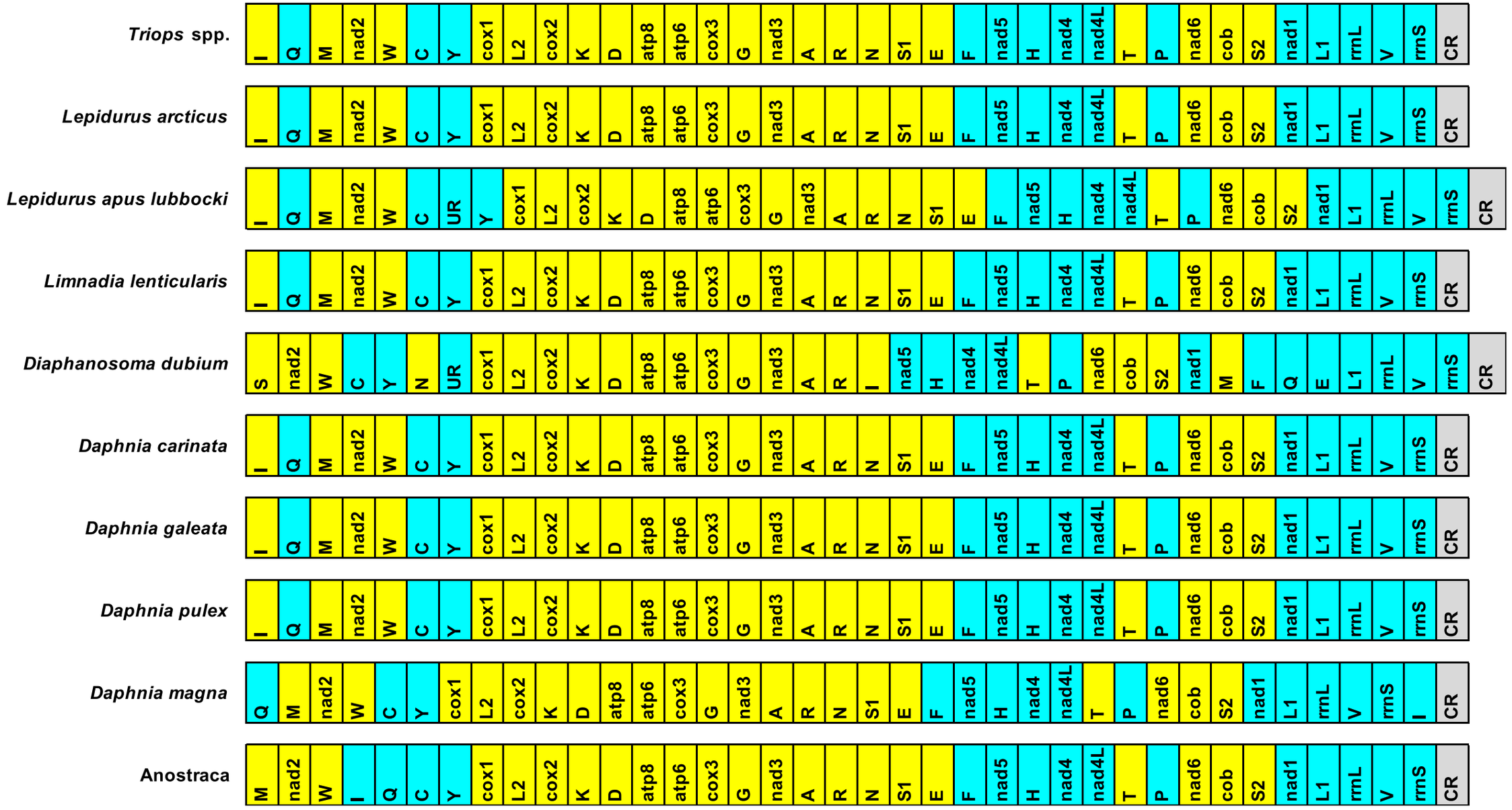


Metazoan mtDNA
12-20 kb

13 Protein Coding Genes
2 rRNA genes
22 tRNA genes
Control Region

Mitochondrial genome comparison:

gene order



Nuclear genome anatomy

articles

Initial sequencing and analysis of the human genome

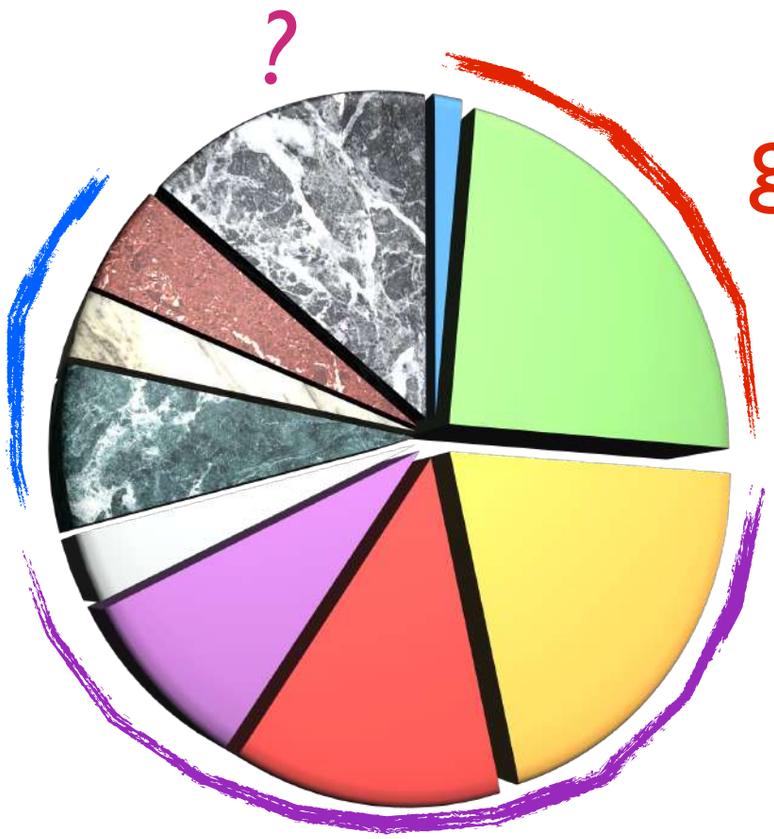
International Human Genome Sequencing Consortium*

genome size: ~3Gbp

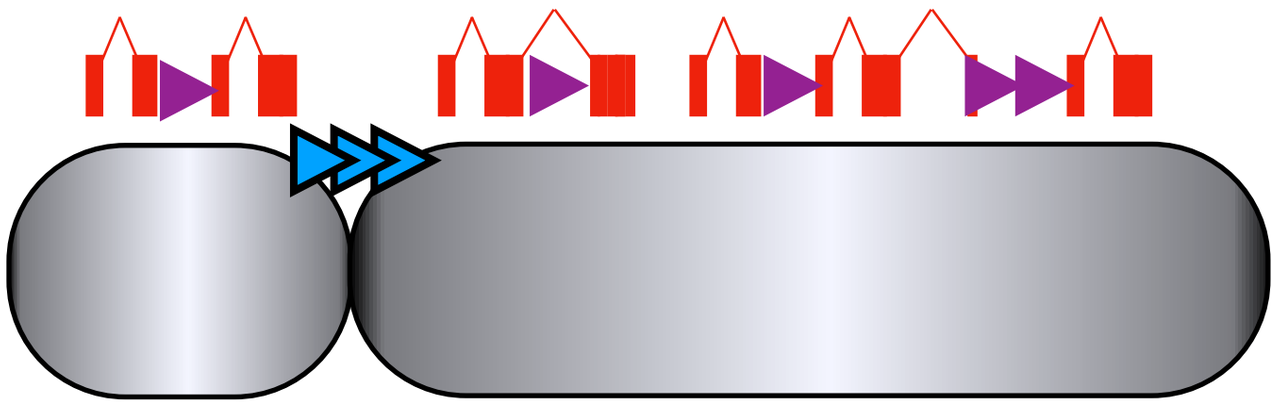
?

genes

tandem reps



mobile DNAs

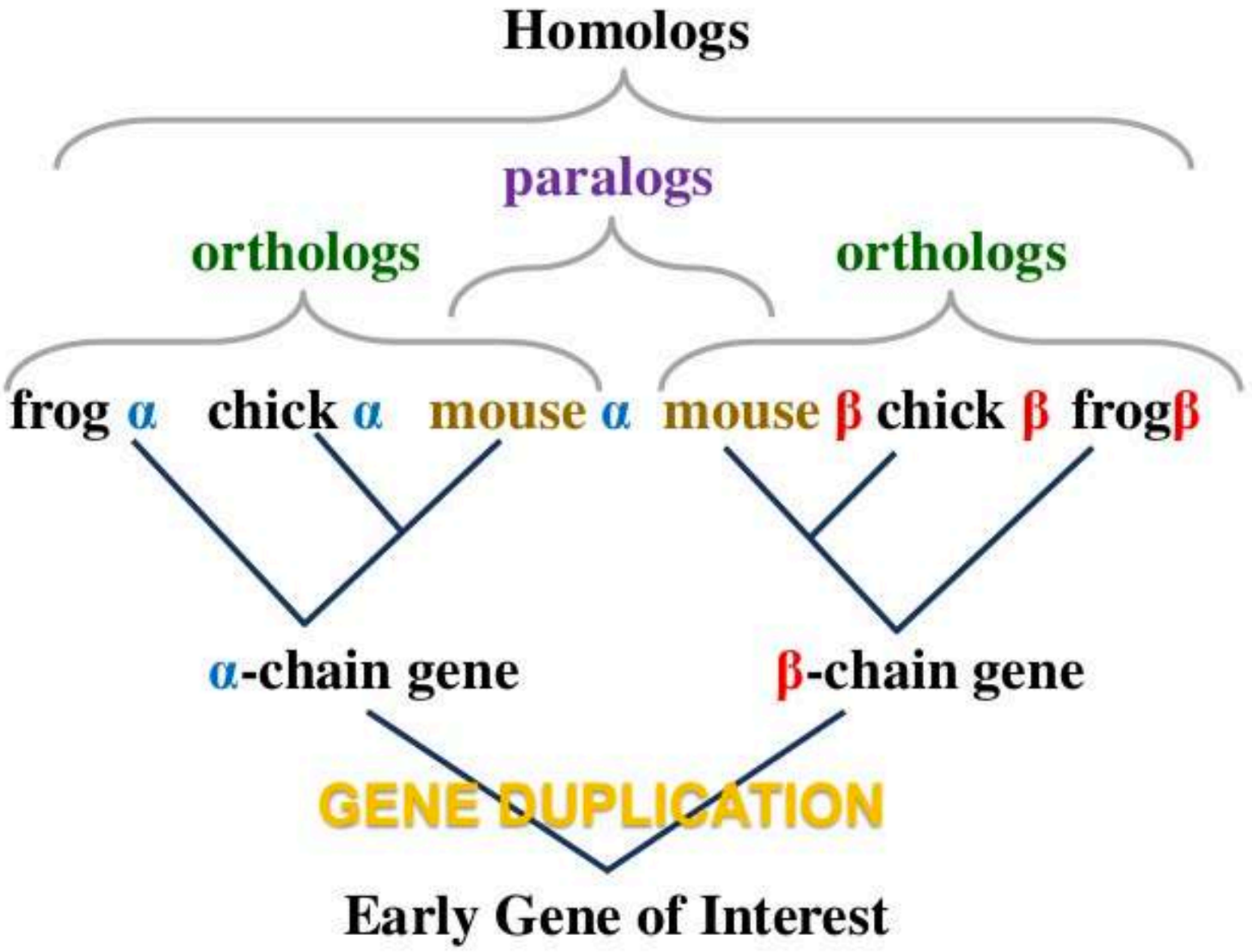


exon reg region line sine LTR DNA sat msat segm dupl other

Genome anatomy

Assembly size	811,852,226 bp
Number of scaffold (N50)	63,275 (30,597 bp)
Number of contigs (N50)	135,150 (10,077 bp)
Mapping rate	98%
k-mer completeness	97%
G+C content	39.6%
BUSCO (N=1,013)	C:84.2%[S:83.1%,D:1.1%],F:12.9%,M:2.9%
N. of predicted genes	17.407
N. of highly supported genes (AED < 0.5)	16.535
Proportion of repeats coverage	44.53%
LINE	13.12%
SINE	9.96%
LTR	4.33%
DNA	12.67%
MITE	3.74%

Types of Homology



"The Ortholog Conjecture"

According to the "**ortholog conjecture**", or standard model of phylogenomics, protein function changes rapidly after duplication, leading to paralogs with different functions, while orthologs retain the ancestral function

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs

Adrian M. Altenhoff[1,2], Romain A. Studer[2,3,4], Marc Robinson-Rechavi[2,3], Christophe Dessimoz[1,2,5]*

1 ETH Zurich, Department of Computer Science, Zürich, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, **4** Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, United Kingdom, **5** EMBL-European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

Abstract

The function of most proteins is not determined experimentally, but is extrapolated from homologs. According to the ''ortholog conjecture'', or standard model of phylogenomics, protein function changes rapidly after duplication, leading to paralogs with different functions, while orthologs retain the ancestral function. We report here that a comparison of experimentally supported functional annotations among homologs from 13 genomes mostly supports this model. We show that to analyze GO annotation effectively, several confounding factors need to be controlled: authorship bias, variation of GO term frequency among species, variation of background similarity among species pairs, and propagated annotation bias. After controlling for these biases, we observe that orthologs have generally more similar functional annotations than paralogs. This is especially strong for sub-cellular localization. We observe only a weak decrease in functional similarity with increasing sequence divergence. These findings hold over a large diversity of species; notably orthologs from model organisms such as *E. coli*, yeast or mouse have conserved function with human proteins.

Citation: Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLoS Comput Biol 8(5): e1002514. doi:10.1371/journal.pcbi.1002514

Editor: Jonathan A. Eisen, University of California Davis, United States of America

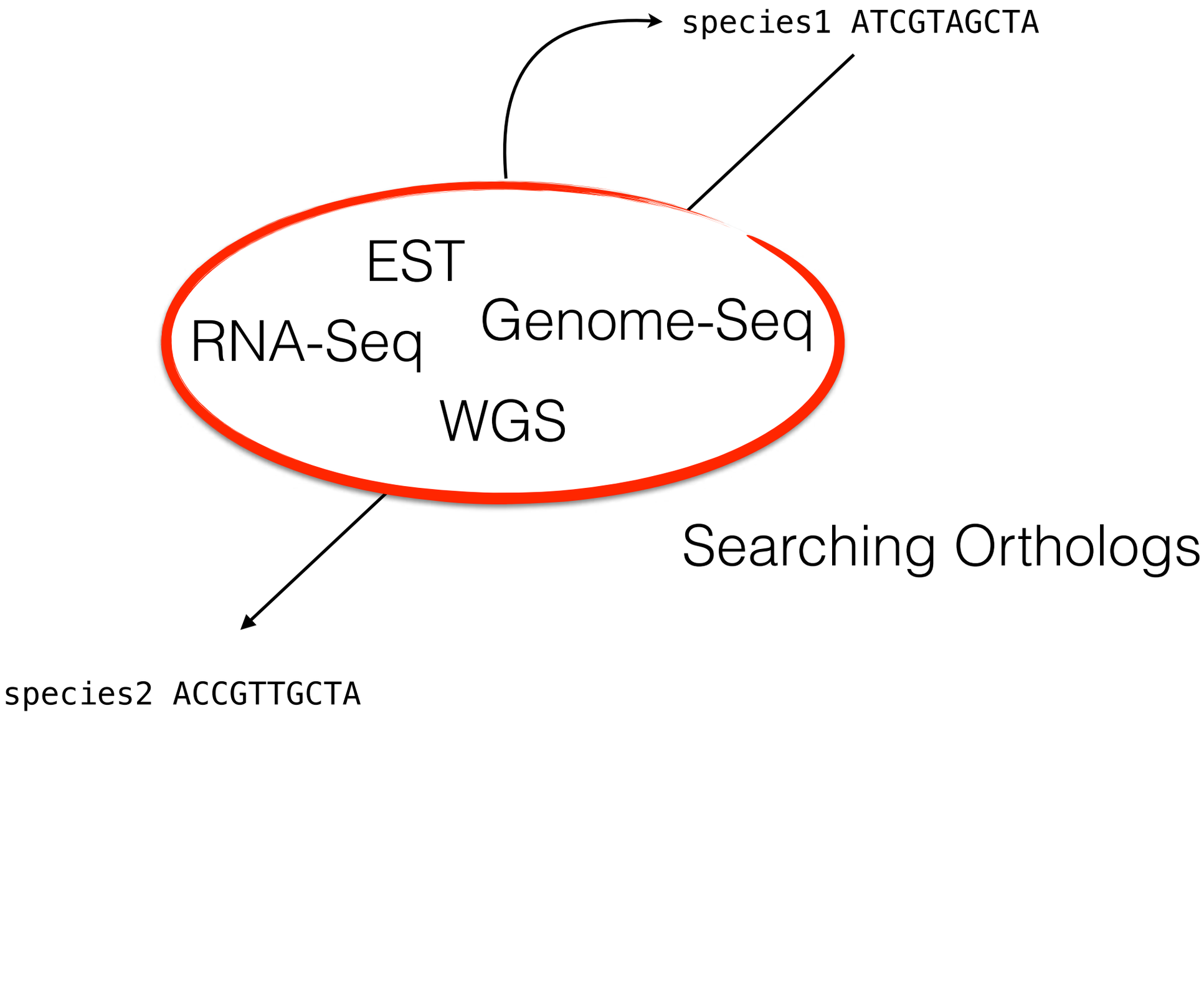
Received October 25, 2011; **Accepted** March 26, 2012; **Published** May 17, 2012

Funding: RAS acknowledges funding from the Fondation du 450ème anniversaire de l'Université de Lausanne and Swiss National Science Foundation grants 132476 and 136477. MR-R acknowledges funding from Etat de Vaud and Swiss National Science Foundation grant 133011. CD is supported by a fellowship from the Swiss National Science Foundation (grant 136461). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

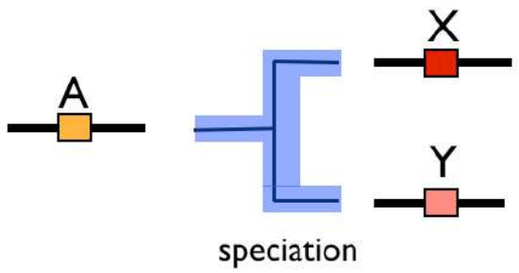
Competing Interests: The authors have declared that no competing interests exist.

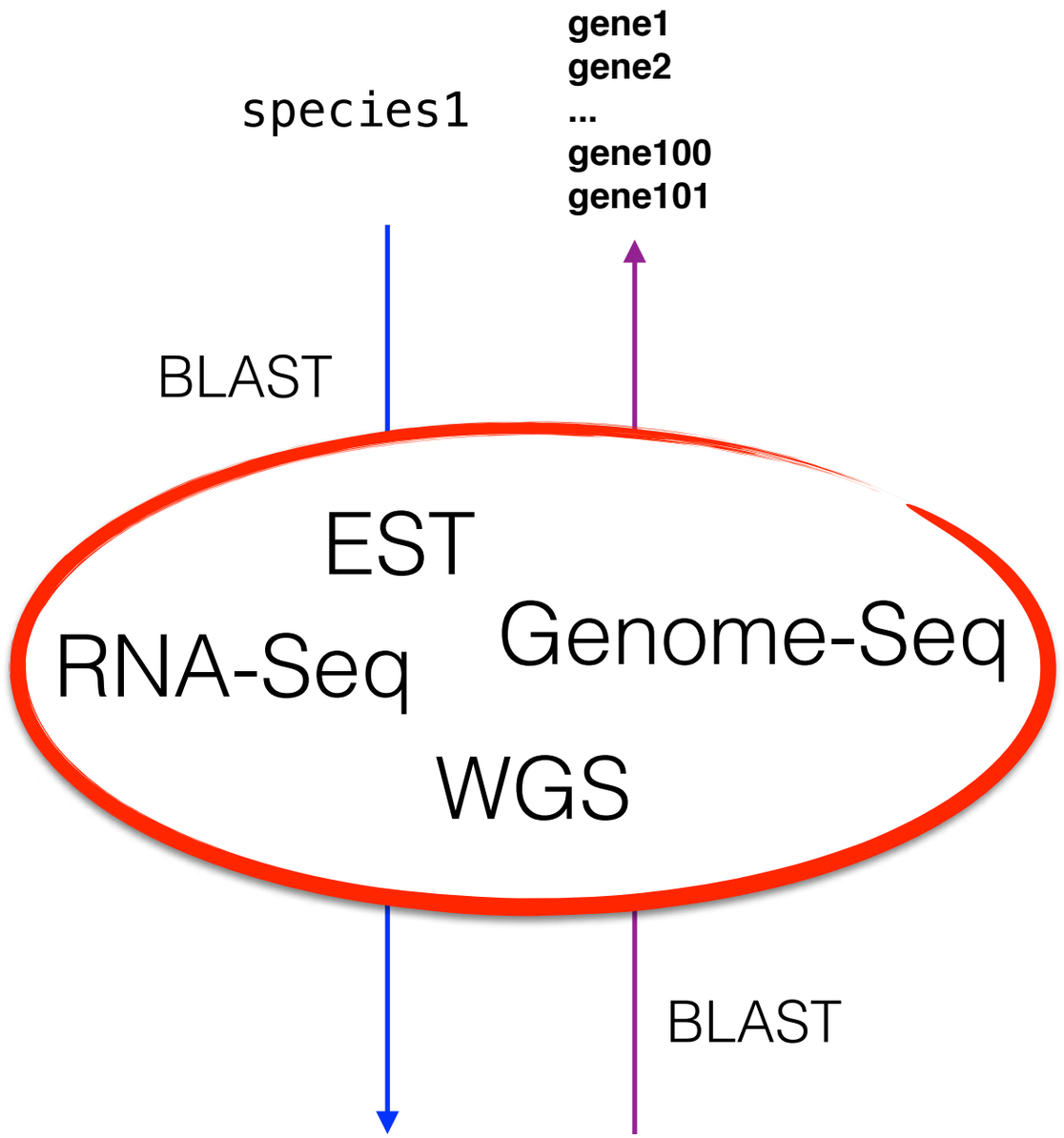
* E-mail: dessimoz@ebi.ac.uk

Gene orthology/paralogy



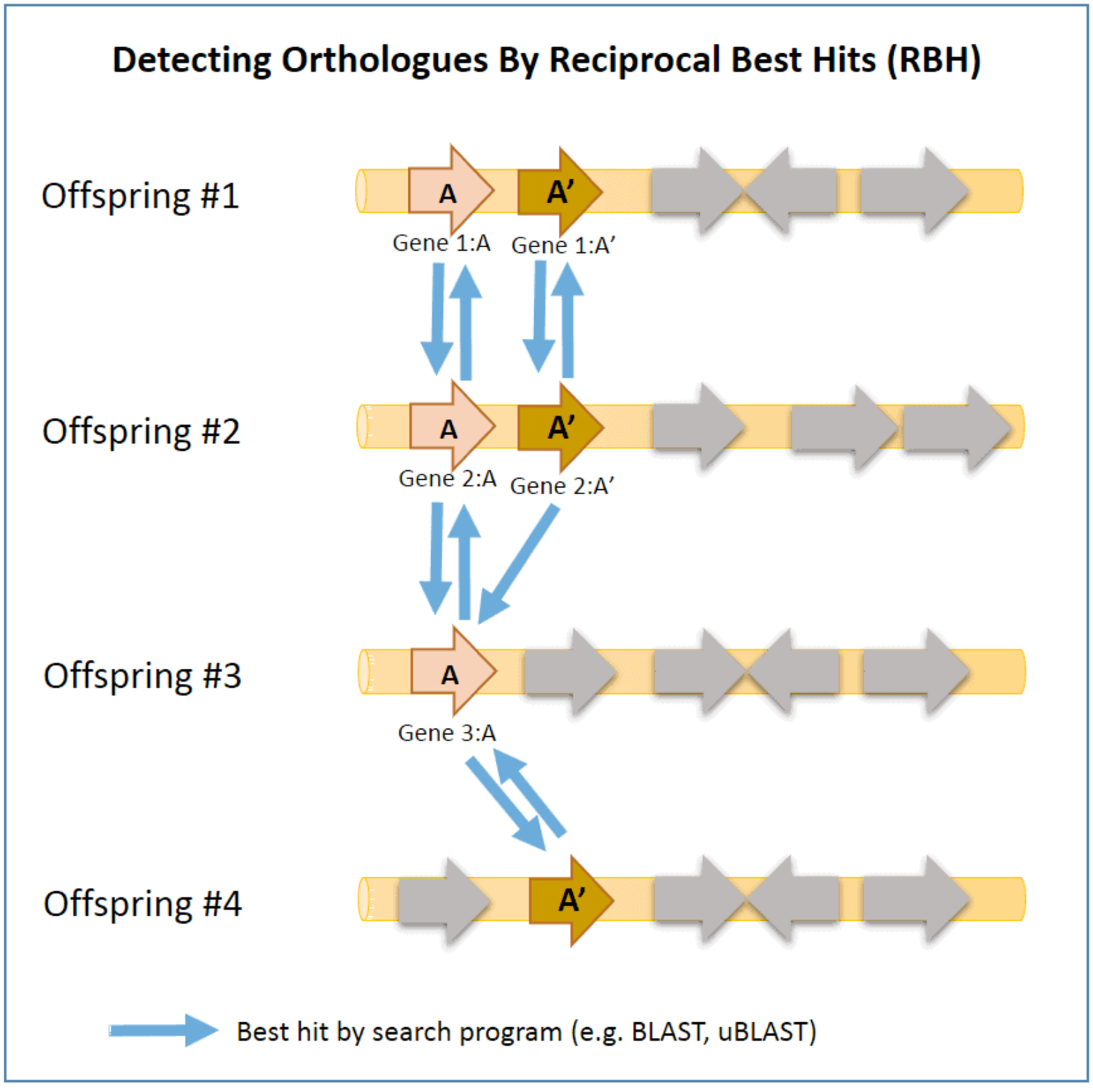
Searching Orthologs





The simplest way:
reciprocal BLAST hit

Gene orthology/paralogy



RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Draft genomes and genomic divergence of two *Lepidurus* tadpole shrimp species (Crustacea, Branchiopoda, Notostraca)

Castrense Savojardo[1,†] | Andrea Luchetti[2,†] | Pier Luigi Martelli[1] | Rita Casadio[1] | Barbara Mantovani[2]

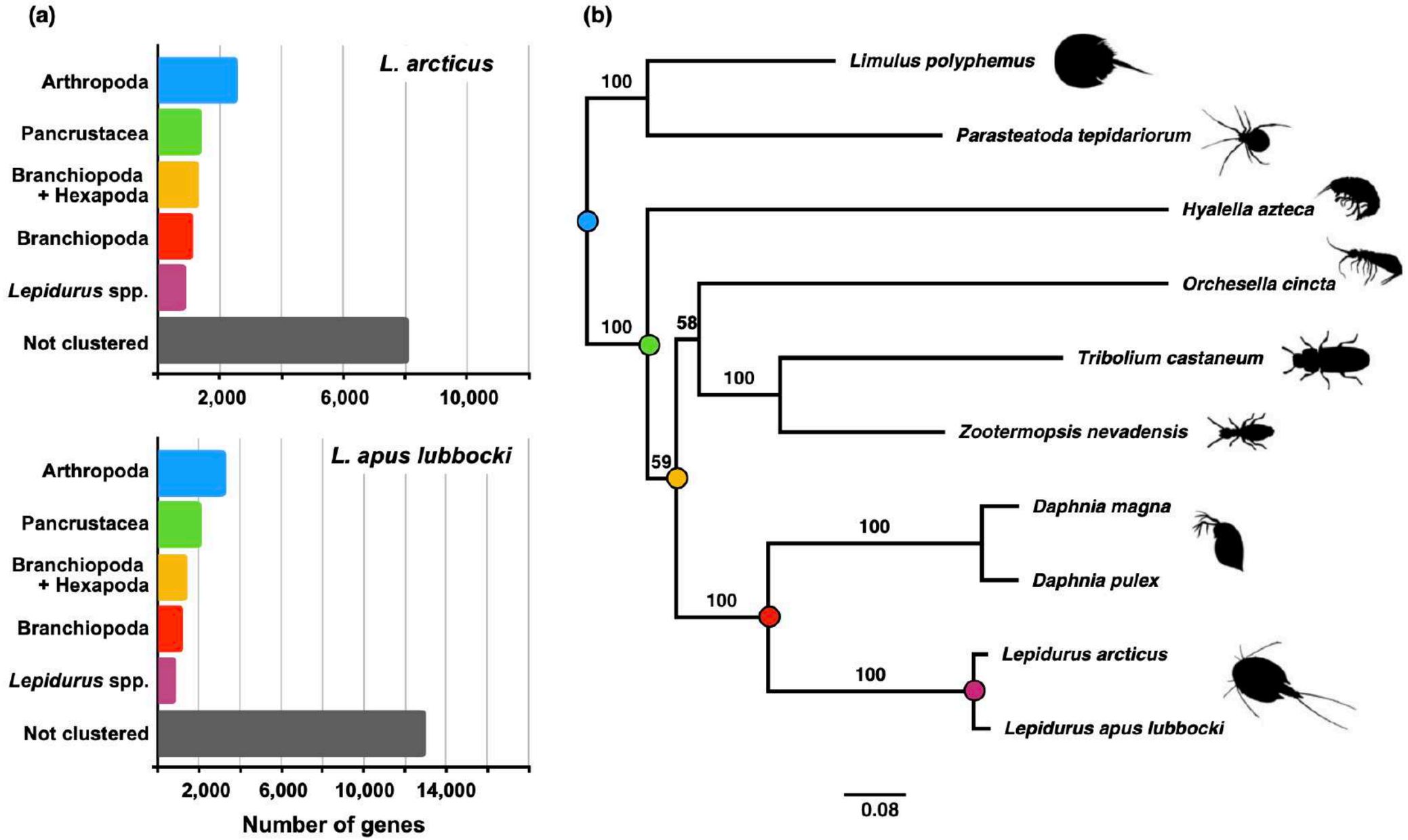


FIGURE 2 Orthologous gene analysis. (a) Taxonomic distribution of orthologous genes. (b) Maximum-likelihood tree (−ln*L* = 1,502,204.92) built on 432 orthologous proteins. The colour of squares at nodes corresponds to the taxonomic colour codes of the bins in the panel (a) bar plots; numbers above branches indicate bootstrap values [Colour figure can be viewed at wileyonlinelibrary.com]

nature ecology & evolution

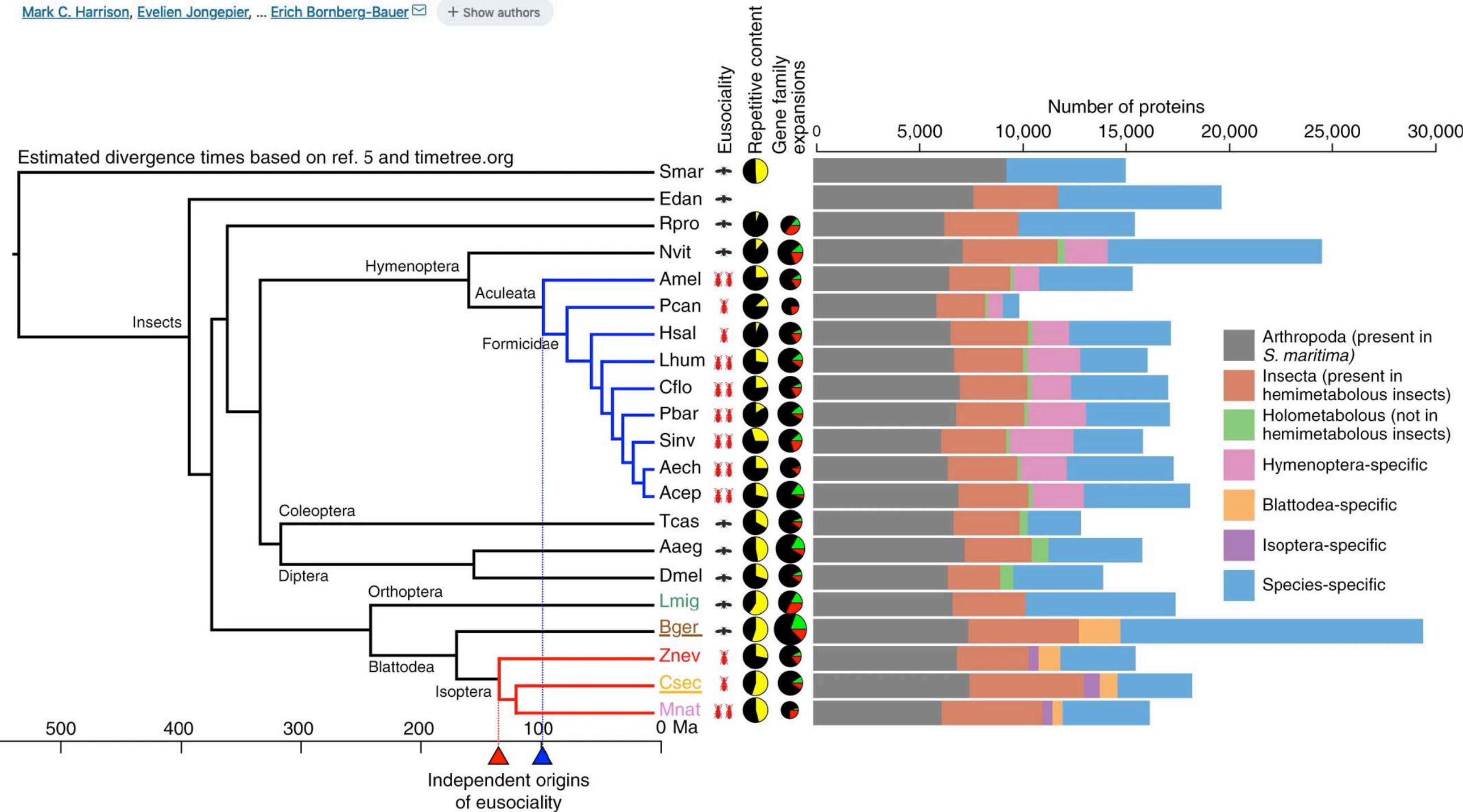
Explore content · About the journal · Publish with us

nature > nature ecology & evolution > articles > article

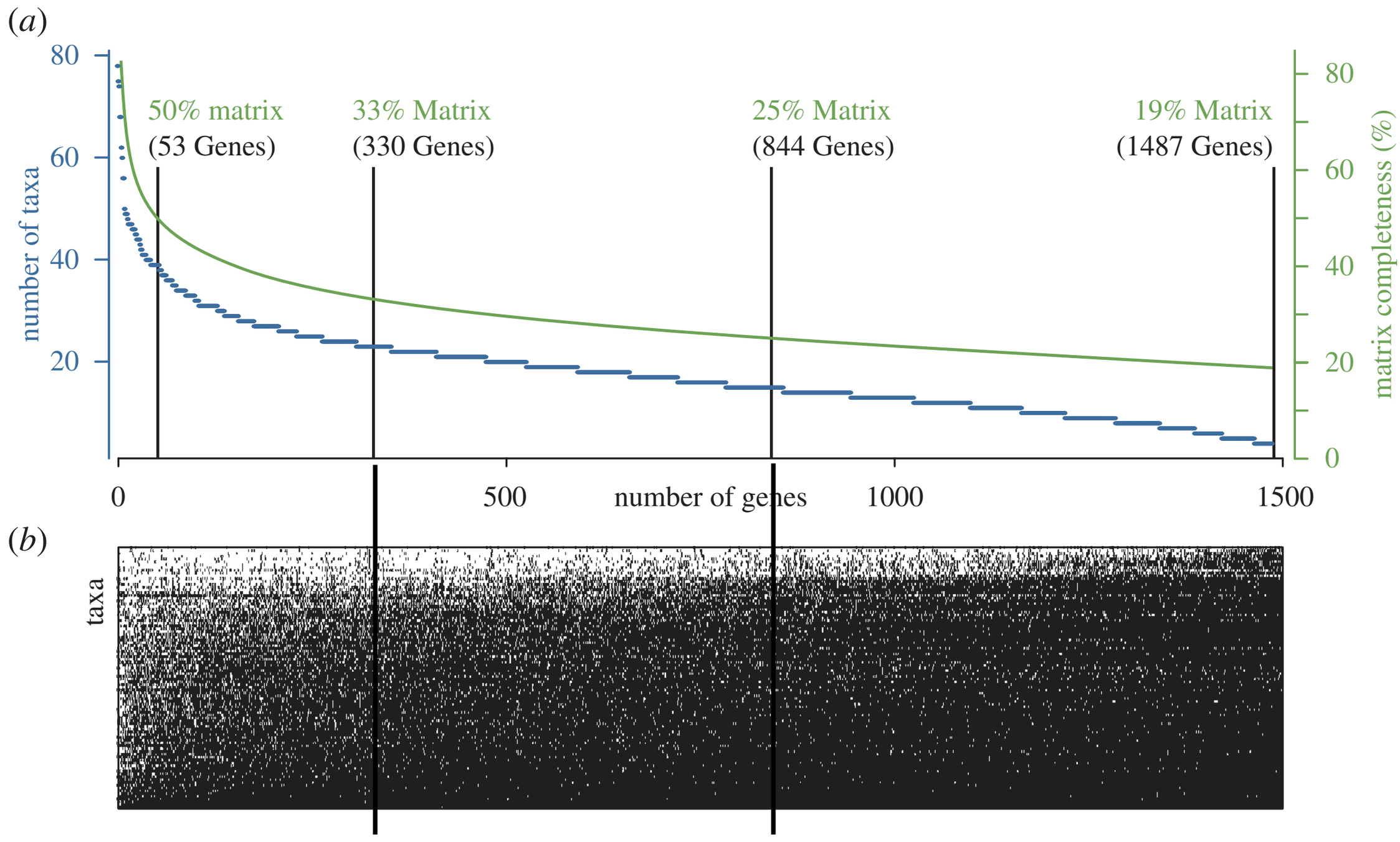
Article | Open Access | Published: 05 February 2018

Hemimetabolous genomes reveal molecular basis of termite eusociality

Mark C. Harrison, Evelien Jongepier, ... Erich Bornberg-Bauer + Show authors

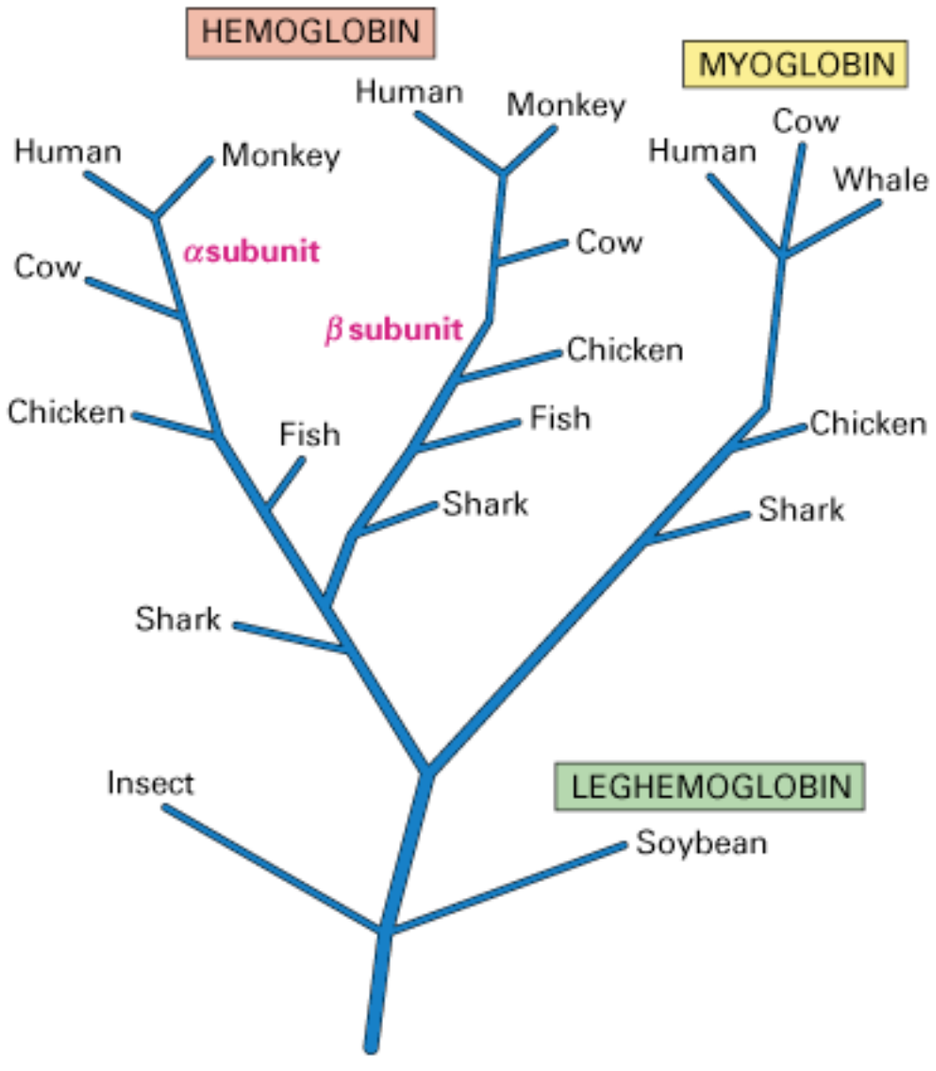


Phylogenomic matrices



Evolution of gene families

A **gene family** is a set of several similar genes, formed by duplication of a single original gene, and generally with similar biochemical functions. One such family are the genes for human hemoglobin subunits



Contents lists available at ScienceDirect

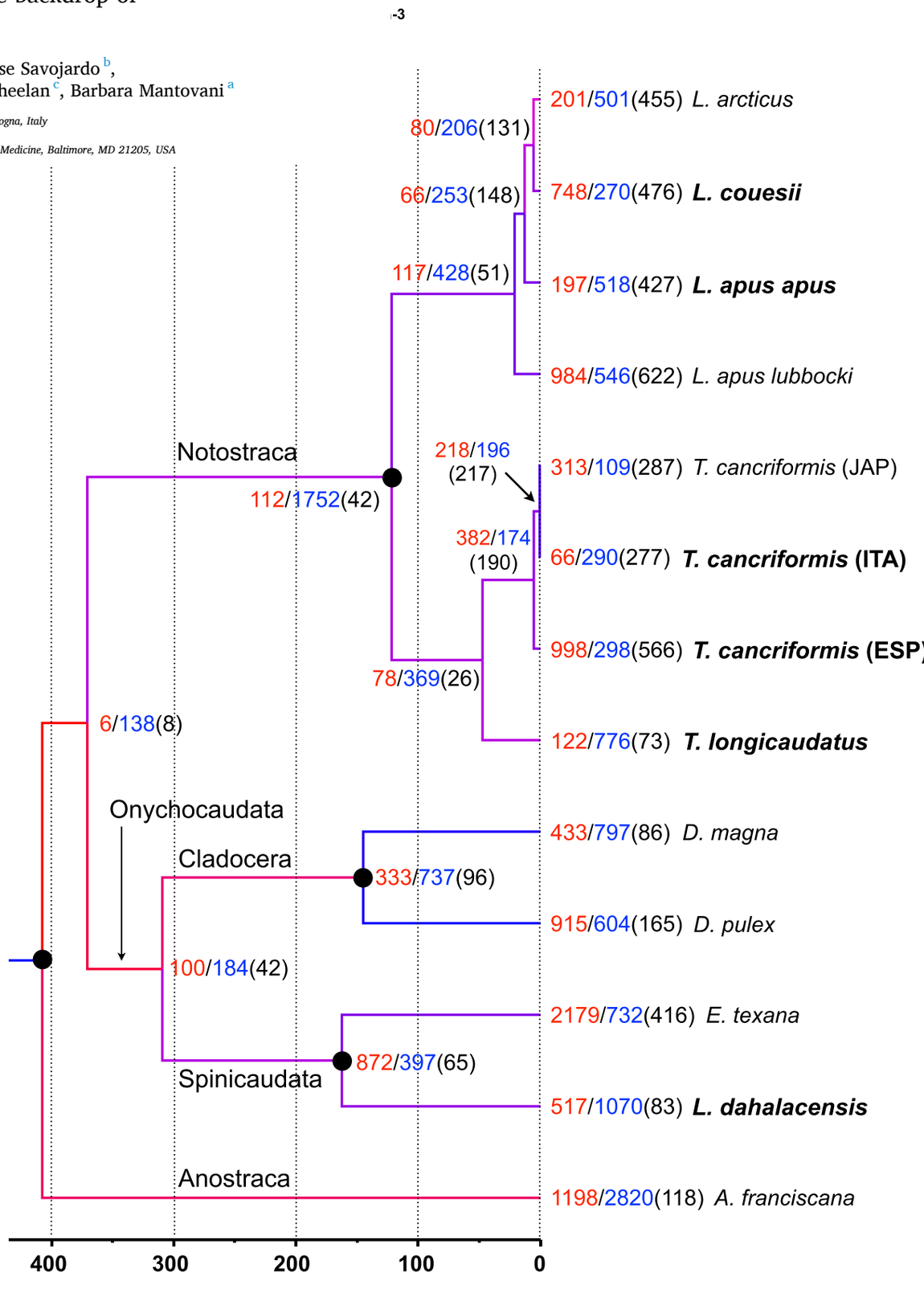
Genomics

journal homepage: www.elsevier.com/locate/ygeno

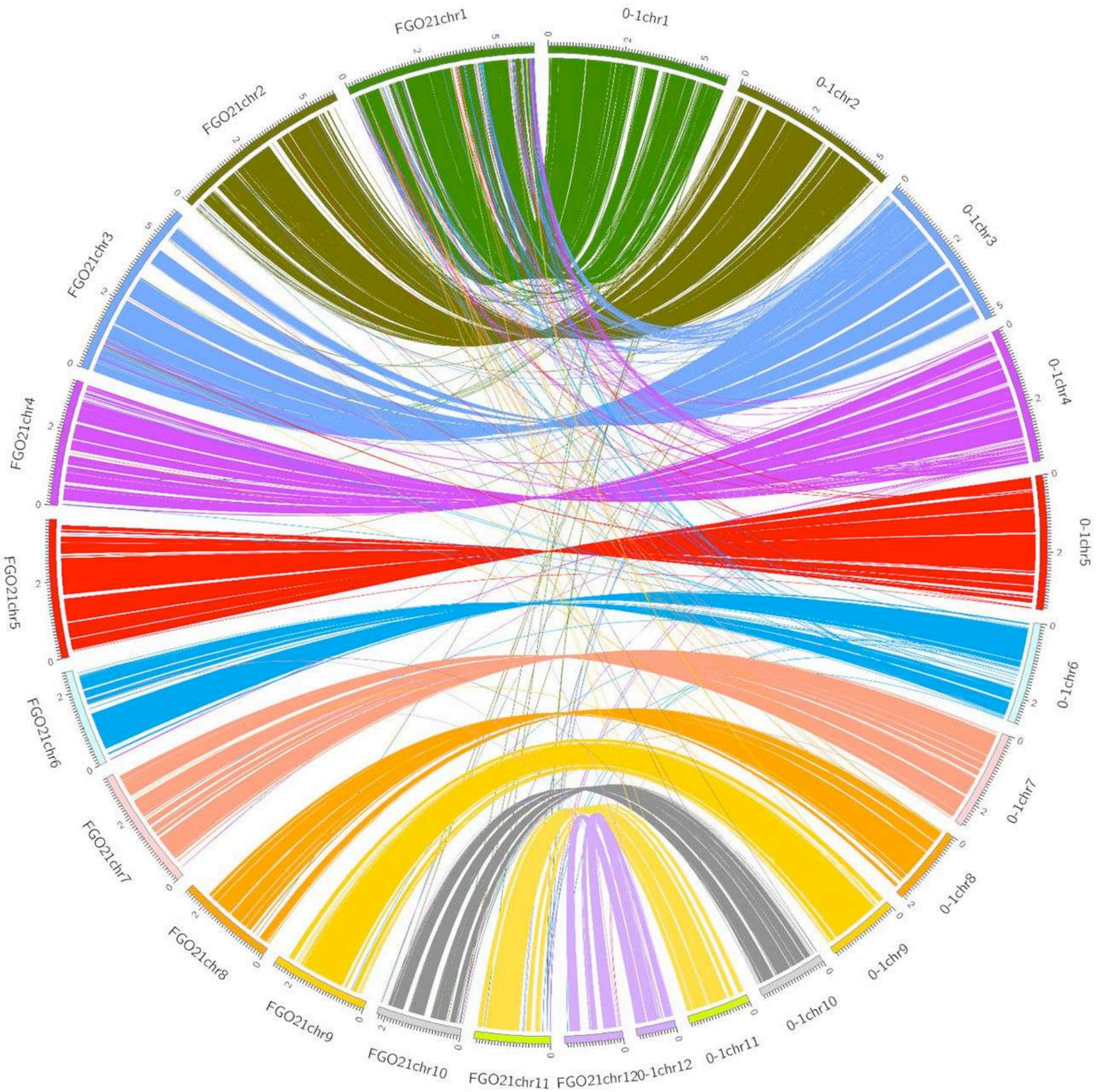
Comparative genomics of tadpole shrimps (Crustacea, Branchiopoda, Notostraca): Dynamic genome evolution against the backdrop of morphological stasis

Andrea Luchetti [a,*], Giobbe Forni [a,1], Jacopo Martelossi [a], Castrense Savojardo [b], Pier Luigi Martelli [b], Rita Casadio [b], Alyza M. Skaist [c], Sarah J. Wheelan [c], Barbara Mantovani [a]

[a] Department of Biological, Geological and Environmental Sciences, University of Bologna, via Selmi 3, 40126 Bologna, Italy
[b] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy
[c] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

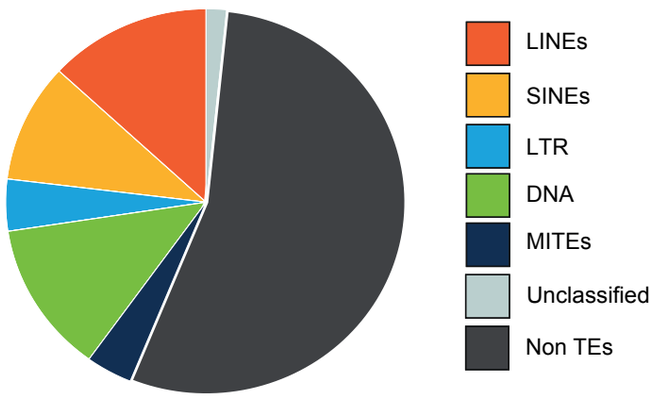


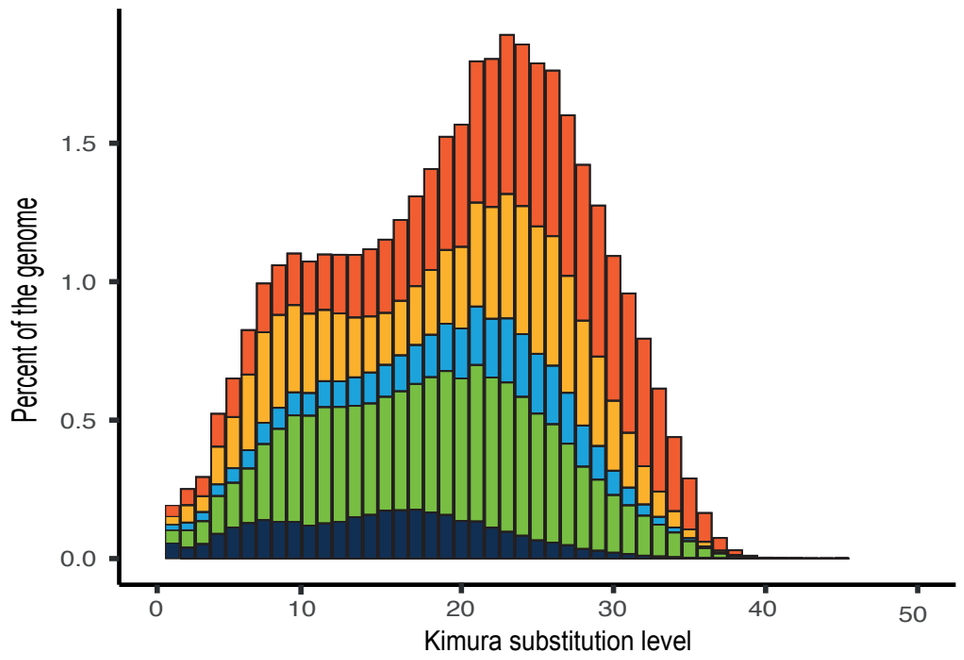
Gene synteny



Evolution of Transposable Elements

Proportion of repeats coverage	44.53%
LINE	13.12%
SINE	9.96%
LTR	4.33%
DNA	12.67%
MITE	3.74%

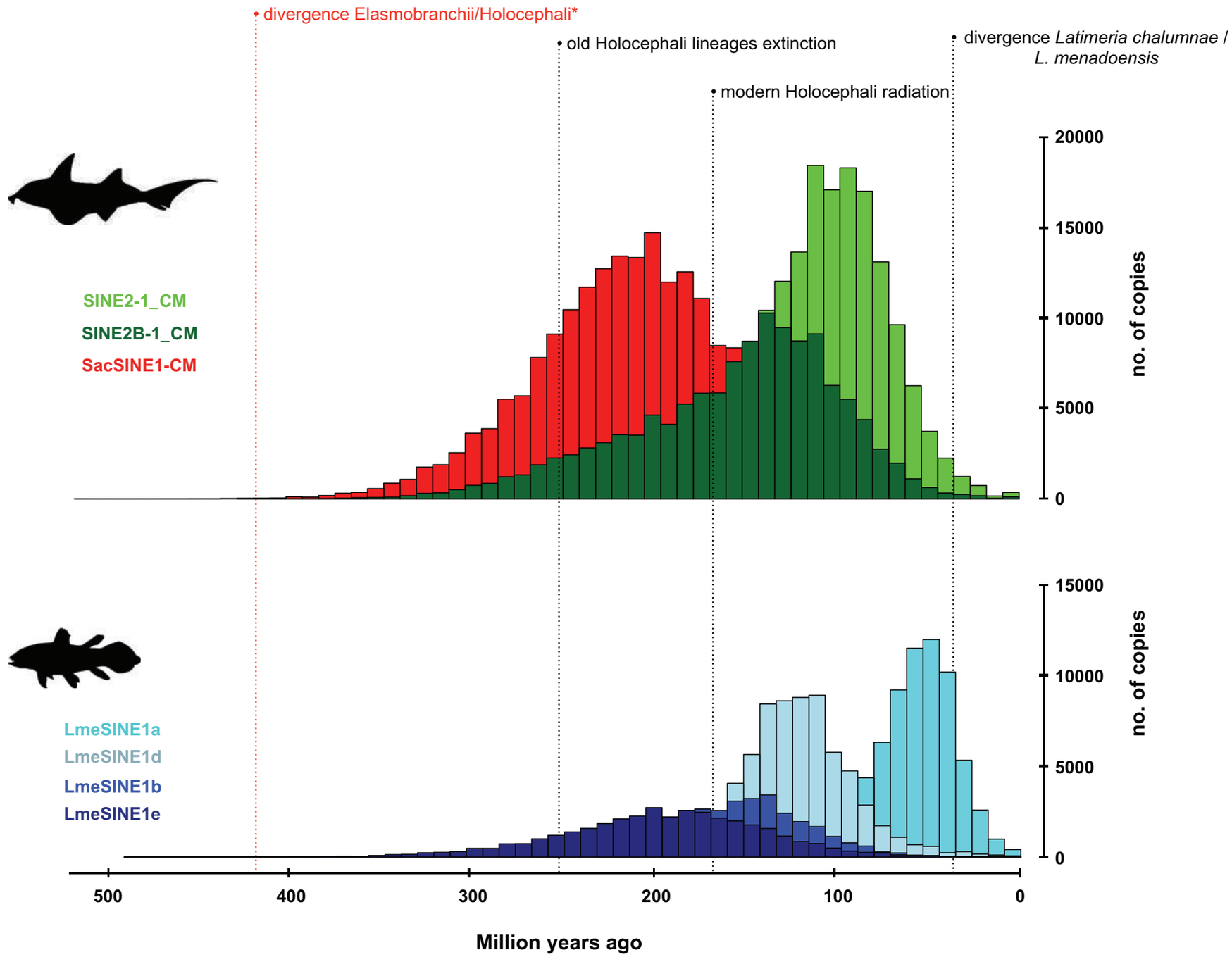




Evolution of Two Short Interspersed Elements in *Callorhinchus milii* (Chondrichthyes, Holocephali) and Related Elements in Sharks and the Coelacanth

Andrea Luchetti*, Federico Plazzi, and Barbara Mantovani

Dipartimento di Scienze Biologiche, Geologiche e Ambientali - Università di Bologna, Italy



Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Comparative genomics of tadpole shrimps (Crustacea, Branchiopoda, Notostraca): Dynamic genome evolution against the backdrop of morphological stasis

Andrea Luchetti [a,*], Giobbe Forni [a,1], Jacopo Martelossi [a], Castrense Savojardo [b], Pier Luigi Martelli [b], Rita Casadio [b], Alyza M. Skaist [c], Sarah J. Wheelan [c], Barbara Mantovani [a]

[a] Department of Biological, Geological and Environmental Sciences, University of Bologna, via Selmi 3, 40126 Bologna, Italy
[b] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy
[c] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA