Genomics & Bioinformatics

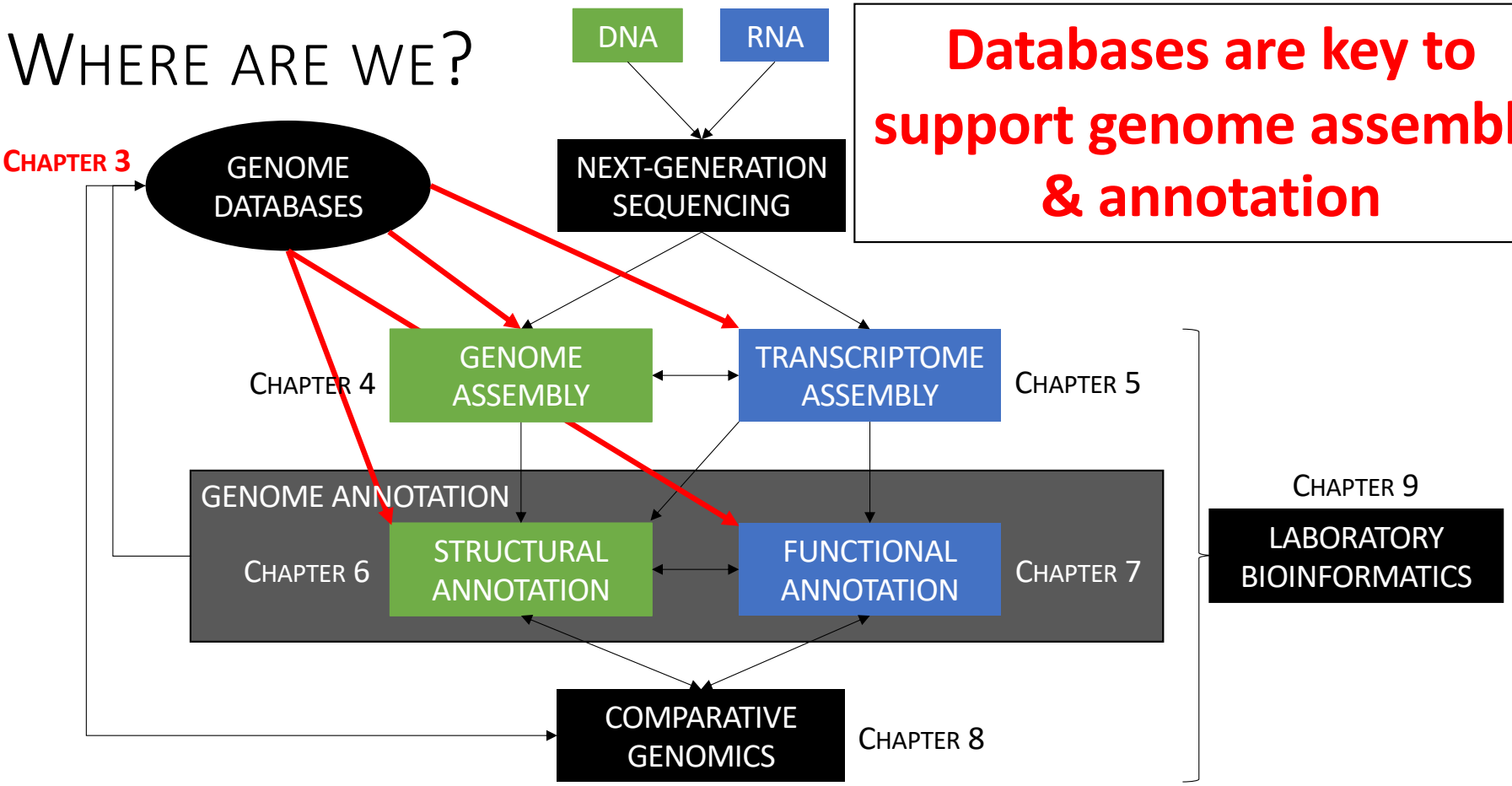
Chapter 3 – Genomes databases

BIOL 497, 597

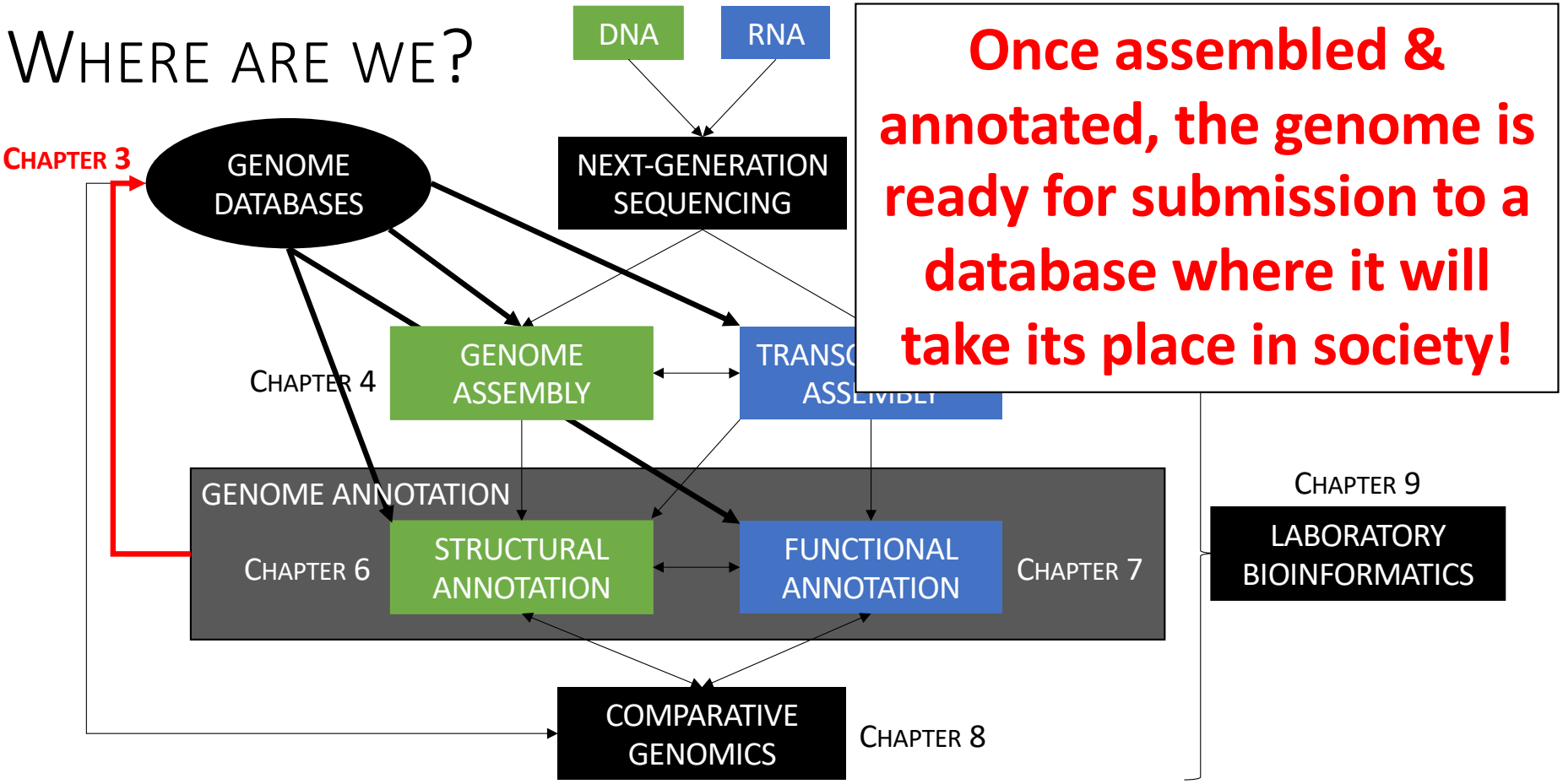
Boise State University

Spring 2022

WHERE ARE WE?

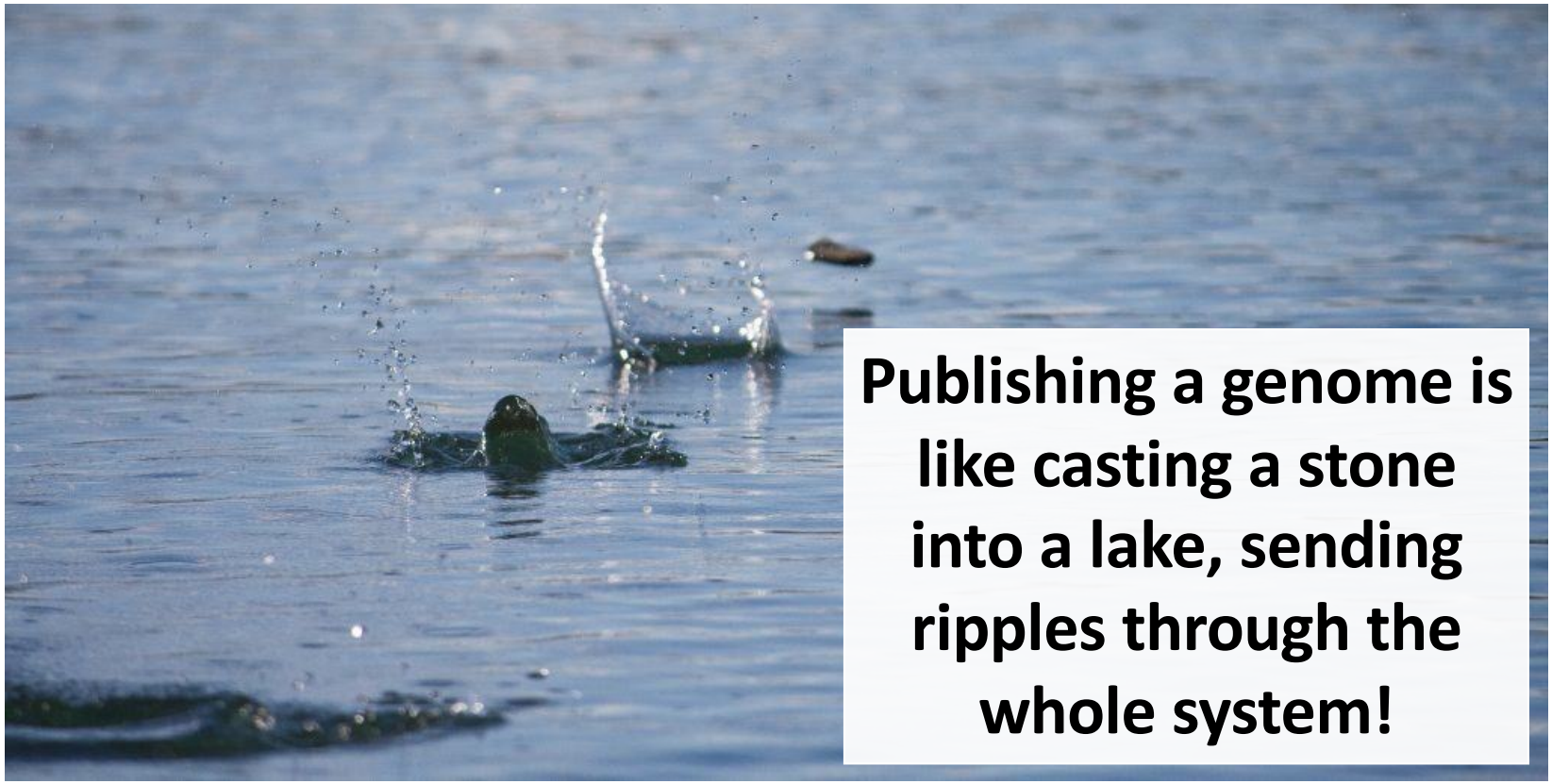

DNA
RNA
Databases are key to support genome assembly & annotation
CHAPTER 3
GENOME DATABASES
NEXT-GENERATION SEQUENCING
CHAPTER 4
GENOME ASSEMBLY
TRANSCRIPTOME ASSEMBLY
CHAPTER 5
GENOME ANNOTATION
CHAPTER 6
STRUCTURAL ANNOTATION
FUNCTIONAL ANNOTATION
CHAPTER 7
CHAPTER 9
LABORATORY BIOINFORMATICS
COMPARATIVE GENOMICS
CHAPTER 8

WHERE ARE WE?

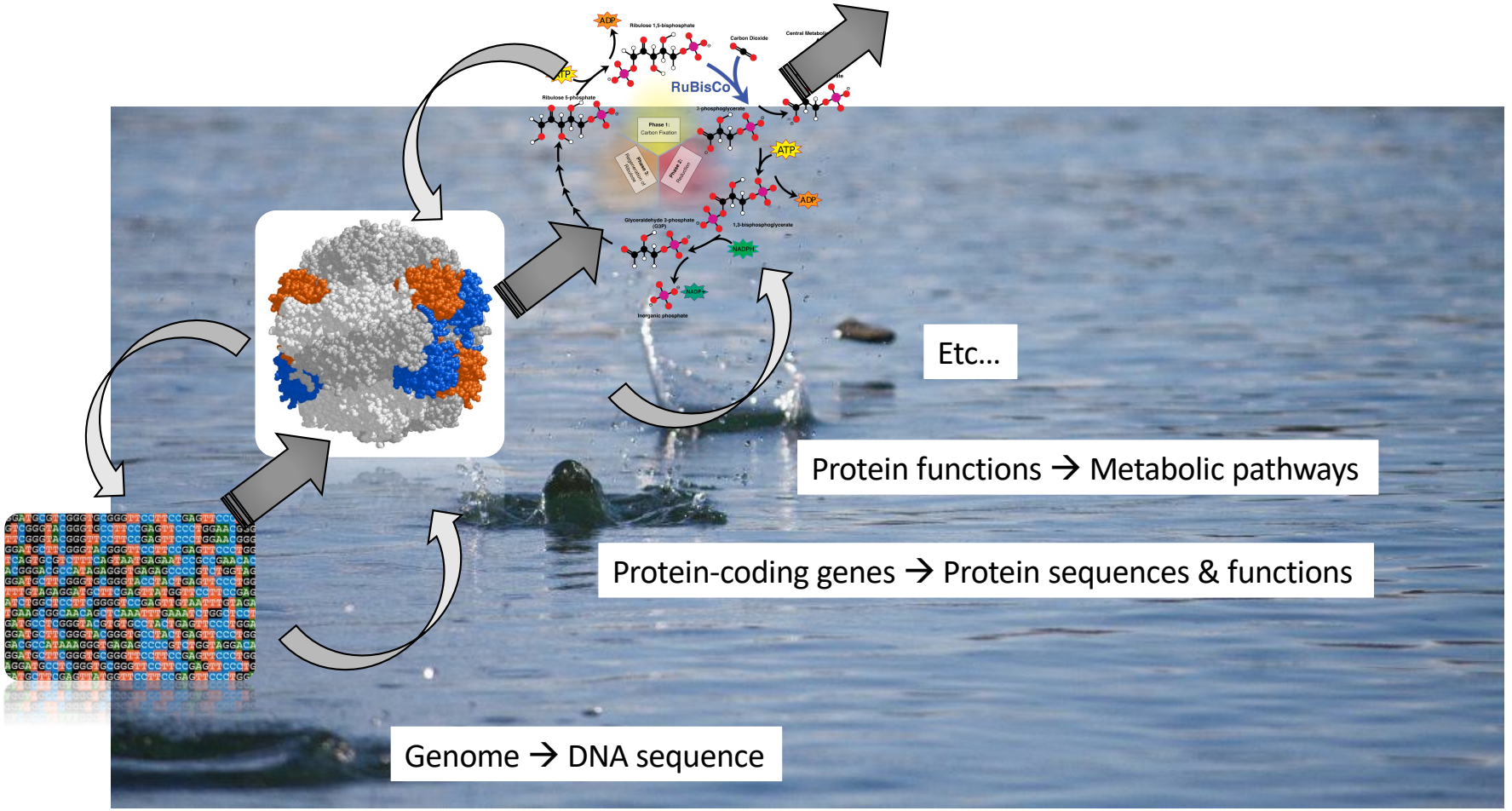


Once assembled & annotated, the genome is ready for submission to a database where it will take its place in society!

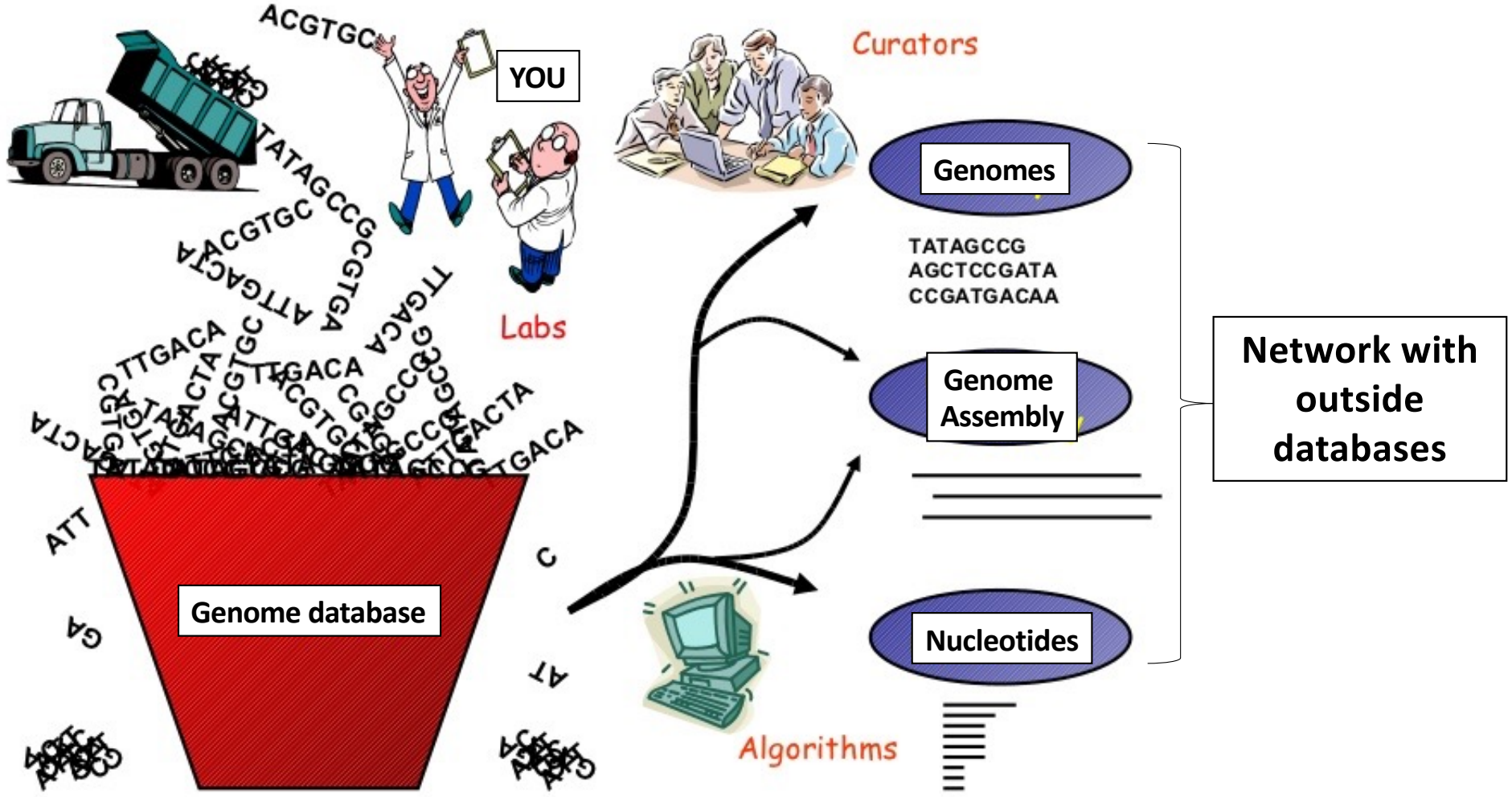
Databases form an interlocking network



DATABASES FORM AN INTERLOCKING NETWORK



GENOME DATABASES – OVERVIEW OF WORKFLOW



REQUIREMENTS OF GENOME DATABASES

Institutions in charge of molecular databases develop tools to:

- **Harvest and curate data** (plus annotations) – that is, check both for accuracy and format – and distribute them.
- **Track and archive data** so that they do not get lost.
- **Record provenance and other information on samples** (e.g. location, vouchers, tissue type, taxonomy).
- **Provide links from the data to relevant items** in other databases, including bibliographical libraries (e.g. PubMed).
- **Provide information retrieval and analysis software** to support research: recovery of selected data and calculations with them (e.g. SRA toolkit, BLAST).
- **Provide documentation and tutorials**.
- Keep up with scientific advances in both biology and informatics.
- Be responsive to users' needs.

GENOME DATABASES – MAJOR TYPES

- Nucleic acid sequences databases
- Protein sequences databases
- Gene ontology databases
- Metabolic pathways databases
- Specialized annotated genomes portals

GENOME DATABASES – MAJOR TYPES

- Nucleic acid sequences databases (In class)
- Protein sequences databases
- Gene ontology databases
- Metabolic pathways databases
- Specialized annotated genomes portals.

Mini-Report 2

NUCLEIC ACID SEQUENCES DATABASES

The International Nucleotide Sequence Database Collaboration (INSDC) is a partnership between 3 DNA seq. databases:

- ✓**DDBJ:** DNA Data Bank of Japan.
- ✓**EMBL-EBI:** European Bioinformatics Institute.
- ✓**NCBI:** National Center for Biotechnology Information (USA).

NUCLEIC ACID SEQUENCES DATABASES

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

NEXT GENERATION READS - SRA

- Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries.
- The SRA stores raw sequencing data (WGS, RNA-Seq) and alignment information from high-throughput sequencing platforms.
- NCBI SRA Toolkit allows to remotely download SRA files.
- We will learn protocols to download SRA files in class (Chapter 4).

WGS of Apostasia shenzhenica: 180 insert size (SRR5759389)

Metadata | Analysis (alpha) | Reads | Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR5759389	84.1M	15.1Gbp	11.3G	35.5%	2017-06-27	public

This run has 2 reads per spot:

L=90, 100%	L=90, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX2959224 to BLAST	Apostasia180	Illumina	WGS	GENOMIC	PCR	PAIRED

Design:

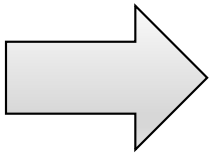
180 insert size library on Illumina

Biosample	Sample Description	Organism	Links
SAMN04453324 (SRS2316248)		Apostasia shenzhenica	• PRJNA310678 [Apostasia shenzhenica isolate:ASH160606] • The Apostasia genome and the evolution of orchids.

Bioproject	SRA Study	Title
PRJNA310678	SRP109877	Apostasia shenzhenica isolate:ASH160606 Genome sequencing and assembly

Show abstract

NUCLEIC ACID SEQUENCES DATABASES

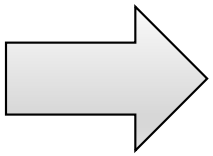


Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

Capillary reads – Trace archives

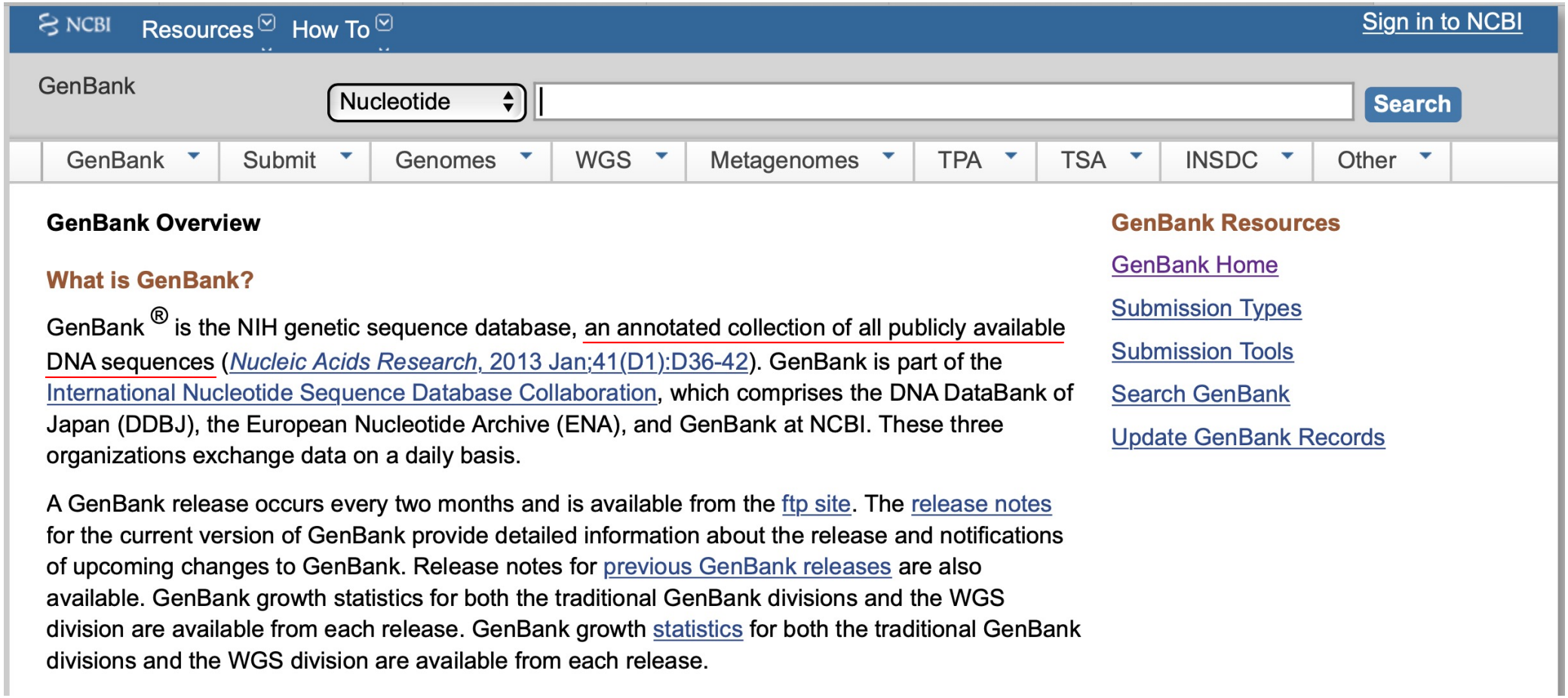
Trace Archive serves as repository of sequencing data from capillary platforms e.g. Applied Biosystems ABI 3730 (= Sanger sequencing).

NUCLEIC ACID SEQUENCES DATABASES



Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

ANNOTATED SEQUENCES – GENBANK



Annotated sequences – GenBank

There are several ways to search and retrieve data from GenBank:

a. Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.
   - Entrez is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).

b. Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool).

Annotated sequences – GenBank

There are several ways to search and retrieve data from GenBank:

c. Search, link, and download sequences programmatically using NCBI e-utilities.

d. The ASN.1 and flatfile formats are available at **NCBI's anonymous FTP server**: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and ftp://ftp.ncbi.nlm.nih.gov/genbank.

ANNOTATED SEQUENCES – GENBANK

GenBank Flat File format

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
```

```
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                     DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                     DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                     NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
```

```
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
     1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
     1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
     1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
     1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
     1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
     1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag
     1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag ttaactacct
     1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca tatgacttac
     1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaattg ggttctataa
     1561 acttattgga tgctccagac tgggtggcat tagataatgc taccatttcc gggtctgtcc
     1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc atttatgata
     1681 cttatggtga tgtgatttat ttcaacttcg aagttgtctc cacaacggat ttgtttgcca
     1741 ttagttctct tcccaatatt aacgctacaa ggggtgaatg gttctcctac tatttttgc
     1801 cttctcagtt tacagactac gtgaatacaa acgtttcatt agagtttact aattcaagcc
     1861 aagaccatga ctgggtgaaa ttccaatcat ctaatttaac attagctgga gaagtgccca
     1921 agaatttcga caagctttca ttaggtttga aagcgaacca aggttcacaa tctcaagagc
```

NUCLEIC ACID SEQUENCES DATABASES

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

SAMPLES – BIOSAMPLES

The BioSample database contains descriptions of biological source materials used in experimental assays.

Plant sample from Apostasia shenzhenica

Identifiers	BioSample: SAMN07191733; Sample name: Apostasia shenzhenica tuber; SRA: SRS2300266
Organism	Apostasia shenzhenica cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; Mesangiospermae; Liliopsida; Petrosaviidae; Asparagales; Orchidaceae; Apostasioideae; Apostasia
Package	Plant; version 1.0
Attributes	**isolate** wild Apostasia shenzhenica **development stage** reproductive growth **geographic location** China: Shenzhen **tissue** tuber
BioProject	PRJNA310678 Apostasia shenzhenica isolate:ASH160606 Retrieve all samples from this project
Submission	The National Orchid Conservation Center of China, Zhongjian Liu; 2017-06-04

Accession: SAMN07191733 ID: 7191733
BioProject SRA

NUCLEIC ACID SEQUENCES DATABASES

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

STUDIES – BIOPROJECT

- A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium.
- A BioProject record provides users a single place to find links to the diverse data types generated for that project.

Apostasia shenzhenica isolate:ASH160606

Accession: PRJNA310678 ID: 310678

Apostasia shenzhenica isolate:ASH160606 Genome sequencing

Apostasia shenzhenica has a karyotype of 2N=2X=68 with uniform small-size chromosomes . More...

Accession	PRJNA310678
Data Type	Genome sequencing
Scope	Monoisolate
Organism	**Apostasia shenzhenica** [Taxonomy ID: 1088818] Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Orchidaceae; Apostasioideae; Apostasia; Apostasia shenzhenica
Publications	Zhang GQ *et al.*, "The Apostasia genome and the evolution of orchids.", *Nature*, 2017 Sep 13;549(7672):379-383
Submission	Registration date: 2-Feb-2016 **Shenzhen Key Laboratory for Orchid Conservation and UtilizationThe National Orchid Conservation Center of China**
Relevance	Evolution

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	2986
WGS master	1
SRA Experiments	45
Protein Sequences	21743
PUBLICATIONS	
PubMed	1
OTHER DATASETS	
BioSample	38

▾ SRA Data Details

Parameter	Value
Data volume, Gbases	467
Data volume, Tbytes	0.24

ADDITIONAL USEFUL DATABASES – NCBI

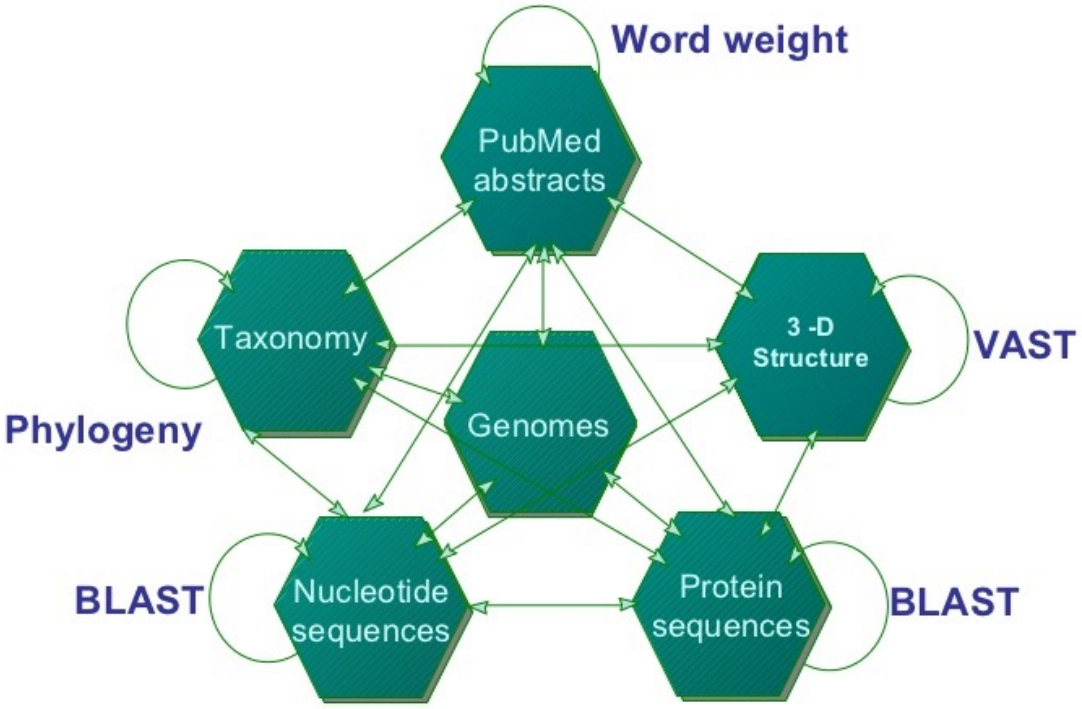
NCBI (via Entrez platform) allows accessing 39 databases:

NCBI Resources How To Sign in to NCBI

Search NCBI databases

Literature	
Books	books and reports
MeSH	ontology used for PubMed indexing
NLM Catalog	books, journals and more in the NLM Collections
PubMed	scientific & medical abstracts/citations
PubMed Central	full-text journal articles
Health	
ClinVar	human variations of clinical significance
dbGaP	genotype/phenotype interaction studies
GTR	genetic testing registry
MedGen	medical genetics literature and links
OMIM	online mendelian inheritance in man
PubMed Health	clinical effectiveness, disease and drug reports
Genomes	
Assembly	genome assembly information
BioCollections	museum, herbaria, and other biorepository collections
BioProject	biological projects providing data to NCBI
BioSample	descriptions of biological source materials
Clone	genomic and cDNA clones
dbVar	genome structural variation studies
Genome	genome sequencing projects by organism
GSS	genome survey sequences
Nucleotide	DNA and RNA sequences
Probe	sequence-based probes and primers
SNP	short genetic variations
SRA	high-throughput DNA and RNA sequence read archive
Taxonomy	taxonomic classification and nomenclature catalog

Genes	
EST	
Gene	
GEO DataSets	
GEO Profiles	
HomoloGene	
PopSet	
UniGene	
Proteins	
Conserved Domains	
Protein	
Protein Clusters	
Structure	
Chemicals	
BioSystems	
PubChem BioAssay	
PubChem Compound	
PubChem Substance	



ASSEMBLY

- The new Assembly Archive at NCBI is a **repository of fully and partially complete genomic assemblies** that exists in association with sequence submissions in GenBank and trace submissions in the NCBI Trace Archive.
- The repository provides users with the **ability to access and evaluate the assemblies** from which finished genomic nucleotide sequence has been derived.
- Many benefits accrue to users of this data, including for example the **ability to determine that a spurious frame shift has occurred, or that a putative SNP is not well supported by adequate coverage**.

ASSEMBLY

NCBI Resources How To Sign in to NCBI

Sequence Set Browser Show help

Project: PEFY01 Search List of all Projects

PEFY00000000.1 Apostasia shenzhenica

Master | Contigs | Download | History

# of Contigs:	12,380
# of Proteins:	21,743
# of Scaffolds/Chrs:	2,985
Total length:	322,899,837 bp
BioProject:	PRJNA310678
BioSample:	SAMN04453324
Keywords:	WGS
Annotation:	Scaffolds
Organism:	Apostasia shenzhenica – show lineage
Biosource:	/country = China: Shenzhen /ecotype = Shenzhen /isolate = ASH160606 /mol_type = genomic /tissue_type = stem; leaf
WGS:	PEFY01000001:PEFY01012380
Scaffolds:	KZ451883:KZ454867 2,985 scaffolds, 21,743 proteins, total length is 348,733,136 bases
Reference:	The Apostasia genome and the evolution of orchids : Nature 549 (7672), 379-383 (2017) – show 35 authors
Submission:	Submitted (25-OCT-2017) Shenzhen Key Laboratory for Orchid Conservation and Utilization, The National Orchid Conservation Center of China, Wangtong Road, Shenzhen 518114, China – Liu,Z.-J.

The Apostasia shenzhenica whole genome shotgun (WGS) project has the project accession PEFY00000000. This version of the project (01) has the accession number PEFY01000000, and consists of sequences PEFY01000001-PEFY01012380.

```
##Genome-Assembly-Data-START##
Assembly Method       : AllPaths v. 49292; Pbjelly v. 14.1; fragscaff v. 140324
Genome Representation : Full
Expected Final Version: Yes
Genome Coverage       : 455.49x
Sequencing Technology : Illumina; PacBio
##Genome-Assembly-Data-END##
```

GENOME

- This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.
- You can download genomes directly using this ftp or browse by organisms.

GENOME

Organism Overview ID: 32002

Acanthisitta chloris (rifleman)

Acanthisitta chloris overview

Lineage: Eukaryota[2661]; Metazoa[876]; Chordata[376]; Craniata[368]; Vertebrata[368]; Euteleostomi[362]; Archelosauria[104]; Archosauria[99]; Dinosauria[95]; Saurischia[95]; Theropoda[95]; Coelurosauria[95]; Aves[95]; Neognathae[92]; Passeriformes[23]; Acanthisittidae[1]; Acanthisitta[1]; Acanthisitta chloris[1]

The rifleman (*Acanthisitta chloris*) is a small suboscine passerine bird native to New Zeeland. Perching birds (i.e., passerines) are usually divided into the songbirds (oscines) and non-songbirds (suboscines) based, in part, on differences in the anatomy of the tracheal structures used to produce sound. The rifleman is thought to belong to an More...

Summary

Submitter: BGI
Assembly level: Scaffold
Assembly: GCA_000695815.1 ASM69581v1 **scaffolds:** 53,875 **contigs:** 120,312 **N50:** 20,602 **L50:** 14,656
BioProjects: PRJNA253841, PRJNA212877
Whole Genome Shotgun (WGS): INSDC: JJRS00000000.1
Statistics: total length (Mb): 1035.88
protein count: 16077
GC%: 41.6
NCBI Annotation Release: 100

Publications

1. Comparative genomics reveals insights into avian genome evolution and adaptation. Zhang G, et al. Science 2014 Dec 12
2. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Lowe TM, et al. Nucleic Acids Res 1997 Mar 1

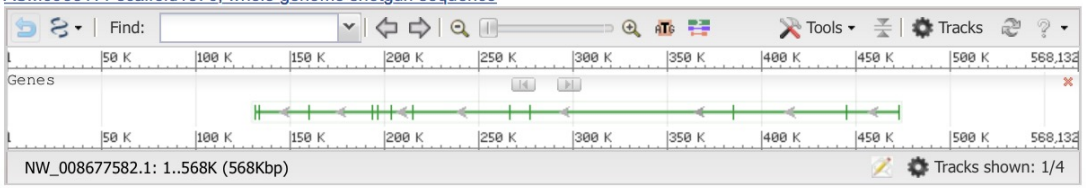
Genome Assembly Annotation

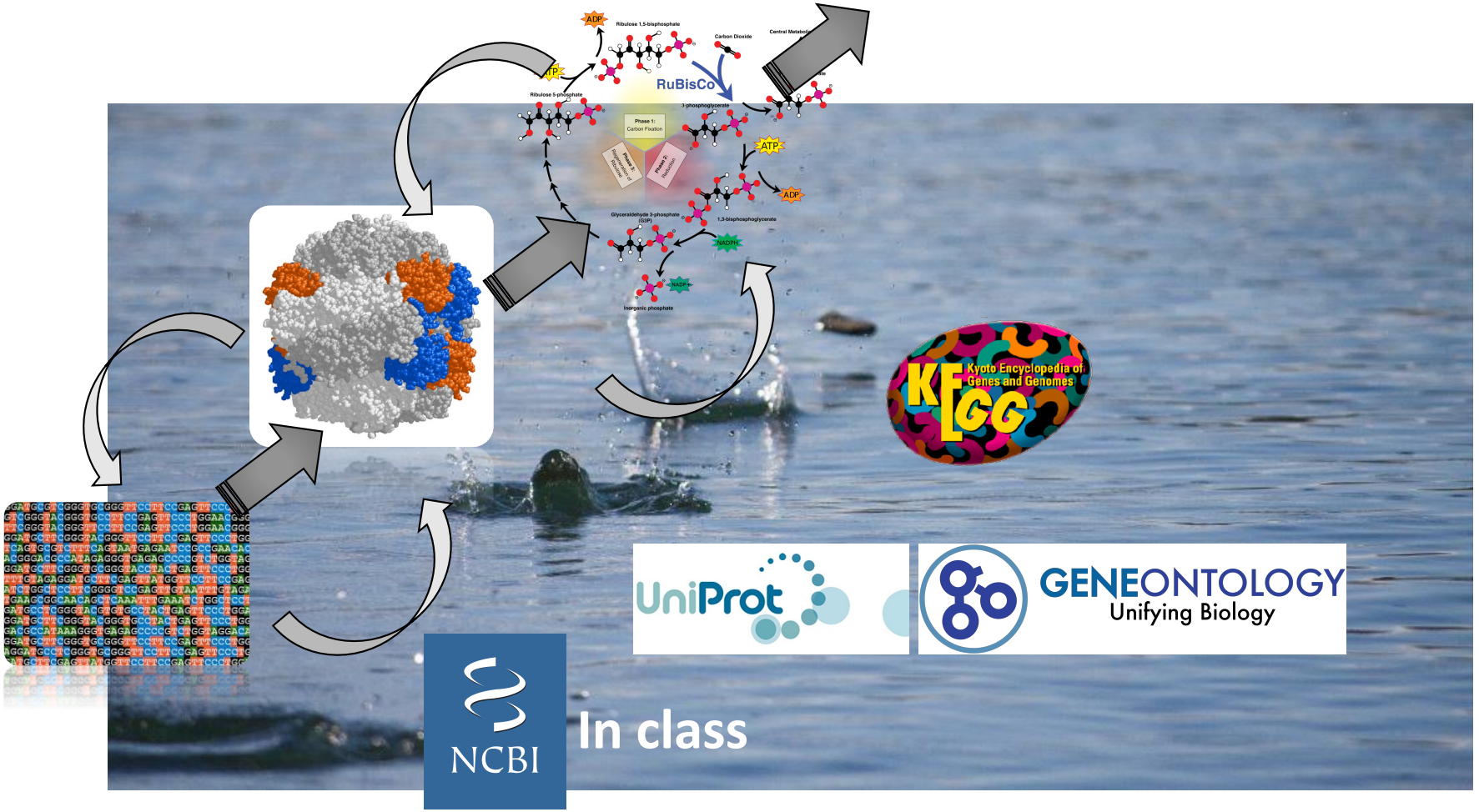
Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	tRNA	Other RNA	Gene	Pseudogene
		master WGS	-	JJRS00000000.1	1,031.1	41.6	2,369	-	-	2,713	146

Genome Region

Acanthisitta chloris isolate BGI_N310 unplaced genomic scaffold, ASM69581v1 scaffold1876, whole genome shotgun sequence

Go to nucleotide: Graphics FASTA GenBank




ADP
Ribulose 1,5-bisphosphate
Carbon Dioxide
ATP
RuBisCo
Ribulose 5-phosphate
3-phosphoglycerate
Phase 1:
Carbon Fixation
Phase 3:
Regeneration of Rubisco
Phase 2:
Reduction
ATP
ADP
Glyceraldehyde 3-phosphate (G3P)
1,3-bisphosphoglycerate
NADPH
NADP+
Inorganic phosphate
KEGG
Kyoto Encyclopedia of Genes and Genomes
UniProt
GENEONTOLOGY
Unifying Biology
NCBI
In class