

Prof. Dr. Boas Pucker

Long Read Genomics

- Generating a Structural & Functional Annotation

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
 - GitHub: <https://github.com/bpucker/LRG>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: [pucker\[a\]uni-bonn.de](mailto:pucker[a]uni-bonn.de)



My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos.

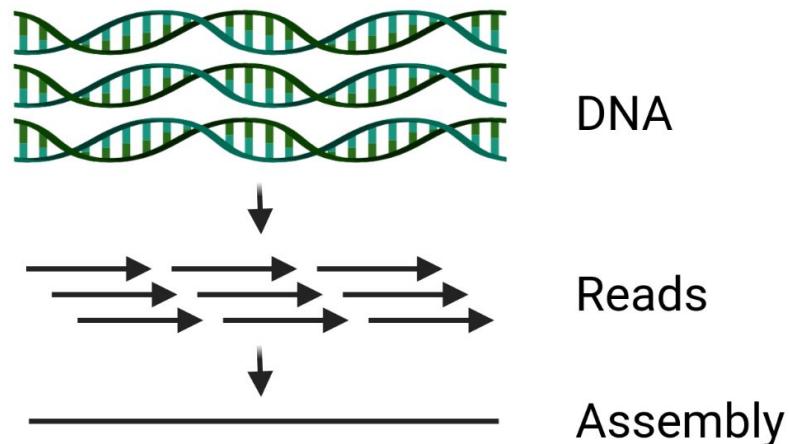
Repetition

- Planning the sequencing experiment
- DNA extraction, quality control, library preparation
- Nanopore sequencing & real time monitoring
- Data analysis (overview)
- Nanopore sequencing applications
- Genome sequencing projects (examples)

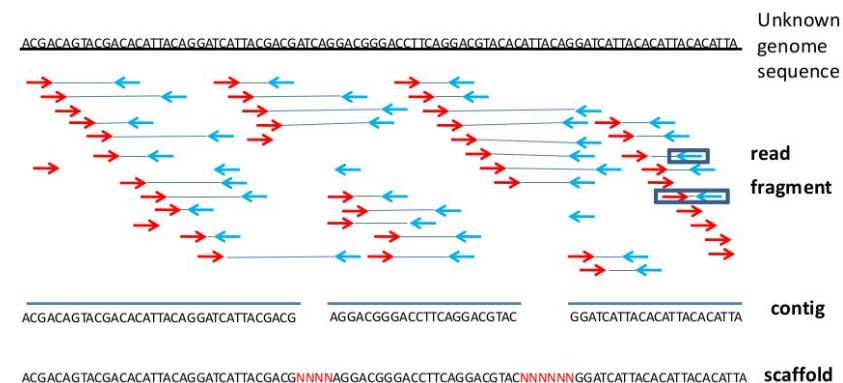
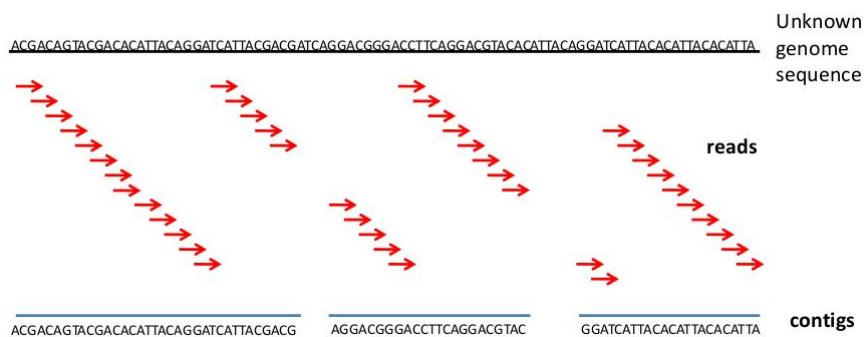
Genome Sequence Assembly

The assembly problem

- Reads are shorter than the chromosome
 - even long reads
- Multiple copies of the genome (DNA) exist that can be subjected to sequencing
- Assembly = putting sequence pieces together (finding the common string of all substrings)
- Genome = DNA in a cell
- Genome sequence = representation of the DNA in a cell



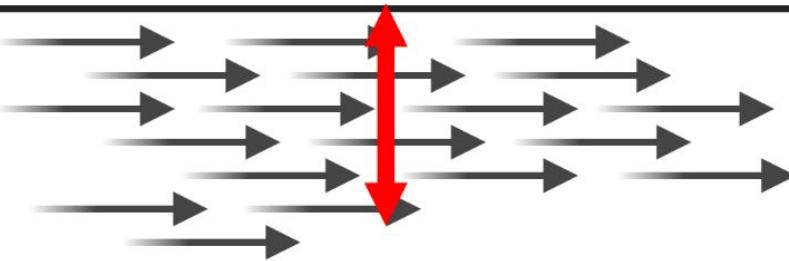
Contigs, scaffolds, pseudochromosomes



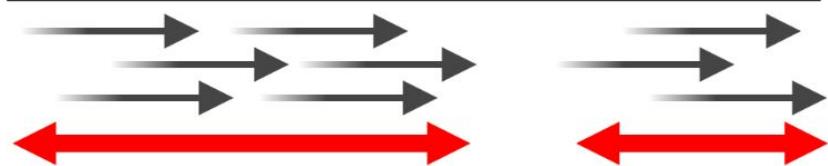
Pseudochromosomes = scaffolds representing an entire chromosome
Gaps = regions between contigs that are represented by Ns

Sequencing coverage depth vs. coverage extent

Sequencing coverage depth



Sequencing coverage extent



Sequencing coverage depth

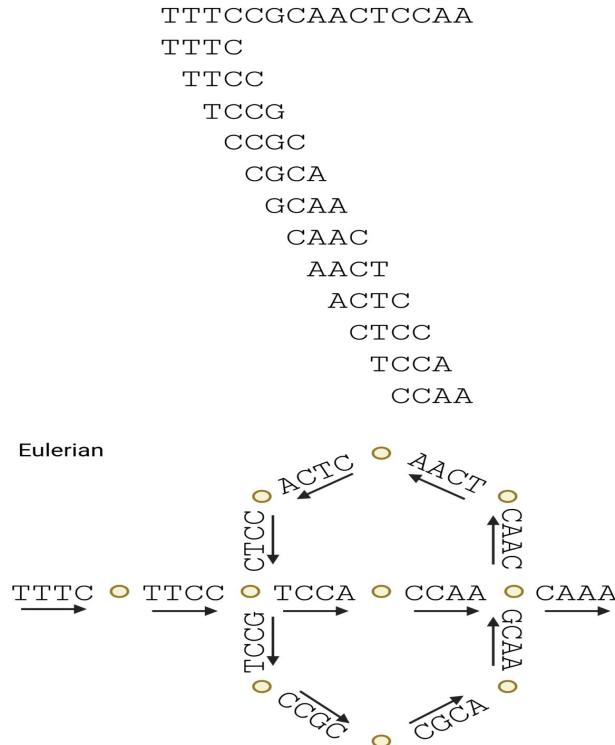
- Coverage depth = average number of times a given base is being sequenced
- Calculation:
 - N = number of reads
 - L = read length in base pairs
 - G = genome size in base pairs
 - Coverage depth $d = N \times L / G$
- Coverage (depth) reflects total amount of sequencing data
- Coverage (depth) is very important parameter for sequencing projects

Sequencing coverage extent

- Coverage extent = ratio of genome covered by at least one base
- Informative to calculate required sequencing depth for a project
- Coverage extent follows a Poisson distribution
- Calculation of coverage extent (c):
 - Non-coverage extent: $P(X=0) = e^{-c}$
 - Probability of coverage extent: $P(X>0) = 1 - e^{-c}$
- Complete coverage of genome requires $G \times e^{-c} < 1$
- Larger genomes required higher sequencing depth
- Real coverage extent is often suffering from sequencing bias

De Bruijn graph (DBG)

- 1) Reads are broken into smaller k-mers
 - 2) K-mers represented in a “De Bruijn graph”
 - 3) Inference of genome sequence from graph
 - Paradigm used in many assemblers: Velvet, ABySS, AllPath-LG, SOAPdenovo
 - Complexity:
 - Number of nodes and links equal to genome size
 - Eulerian path problem is easier to solve than Hamiltonian path problem

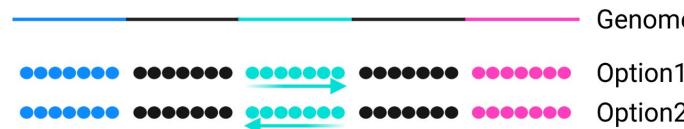
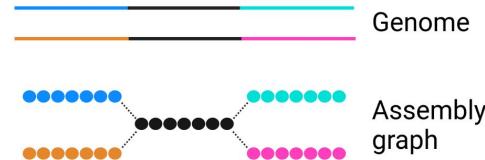


K-mer size

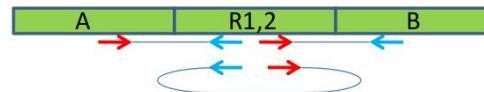
- Larger k-mers increase the assembly continuity by spanning repeats
- Larger k-mers are more sensitive to sequencing errors
- Removal of low abundance k-mers (caused by sequencing errors)
- K-mer size should be 0.5 to 0.75 of the read length
- Typical k-mer sizes:
 - 87 or 97 for 2x150bp reads

Assembly challenges

- Collapsed repeats
- Inversions
- Overstretched repeats



Miss-assembly:

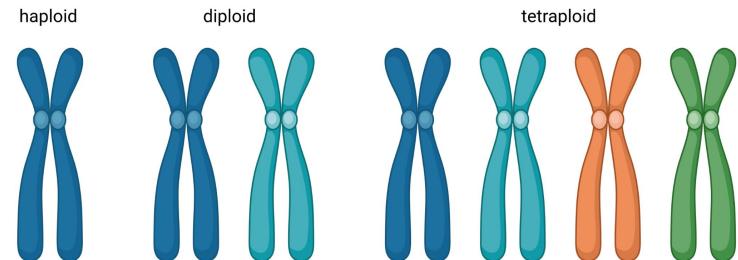
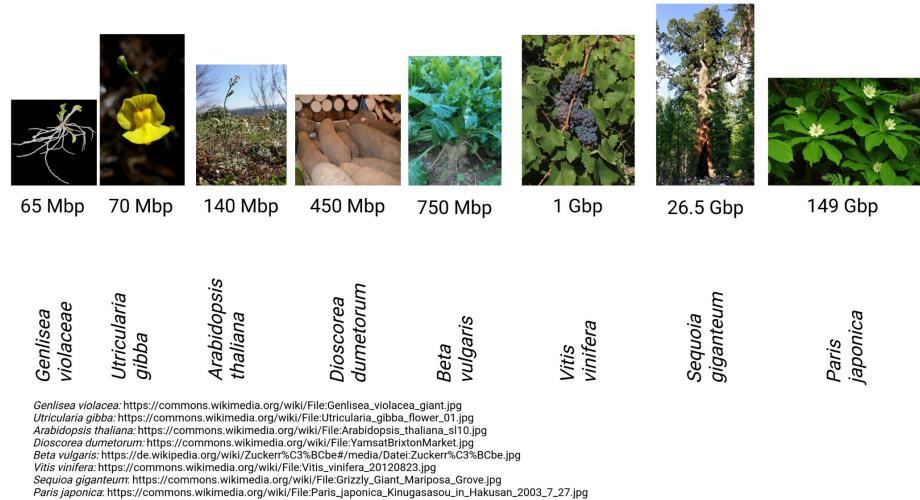


Correct assembly:



Assembly challenges (2)

- Genome size: variation from 65 Mbp to 149 Gbp
- Ploidy: haploid/diploid genomes are much easier to analyze than polyploid genomes



Assembly polishing (outdated)

ONT assembly . . . ACGTTACGGTACGACATGACGTAAAAAAAACGATAAGTACGTGTTAGCGTACGTACGTAAACGTT . . .

Illumina reads

GACGTAAAAAAA - CGATAGTACGTGTTAGCGTACGTAC
ATGACGTAAAAAAA - CGATAGTACGTGTT
CGTAAAAAAA - CGATAGTACGTGTTAGCGTAC
ACATGACGTAAAAAAA - CGATAGTACGT
CGTAAAAAAA - CGATAGTACGTGTTAGCGT
GACGTAAAAAAA - CGATAGTA

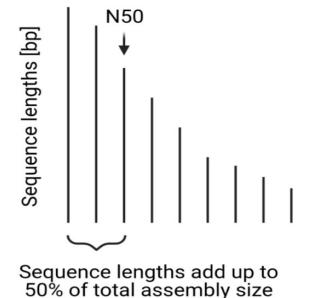
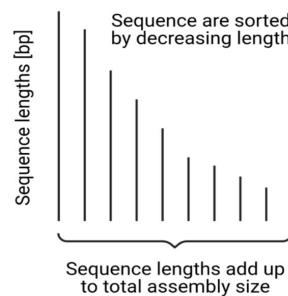
Assembly evaluation

- Continuity: Does the assembly represent a genome in a small number of contigs?
- Completeness: Are all parts of the genome represented?
- Correctness: Is the assembly a correct representation of the genome?

Evaluation: continuity

- Check assembly continuity: calculation based on sequences in FASTA file
- Number of contigs, assembly size, N50

assembler	Canu	FALCON	Miniasm	Flye
number of contigs	69	26	72	44
assembly size	123.5 Mbp	119.5 Mbp	120.2 Mbp	117 Mbp
maximal contig length	15.9 Mbp	15.9 Mbp	14.3 Mbp	14.9 Mbp
N50	13.4 Mbp	9.3 Mbp	8.6 Mbp	10.6 Mbp
N90	2.9 Mbp	2.8 Mbp	1.4 Mbp	2.5 Mbp



Evaluation - completeness

- Check assembly completeness: inspection for presence of conserved genes
- BUSCO = Benchmarking Universal Single-Copy Orthologue
- BUSCO genes are used to assess assembly completeness

BUSCO

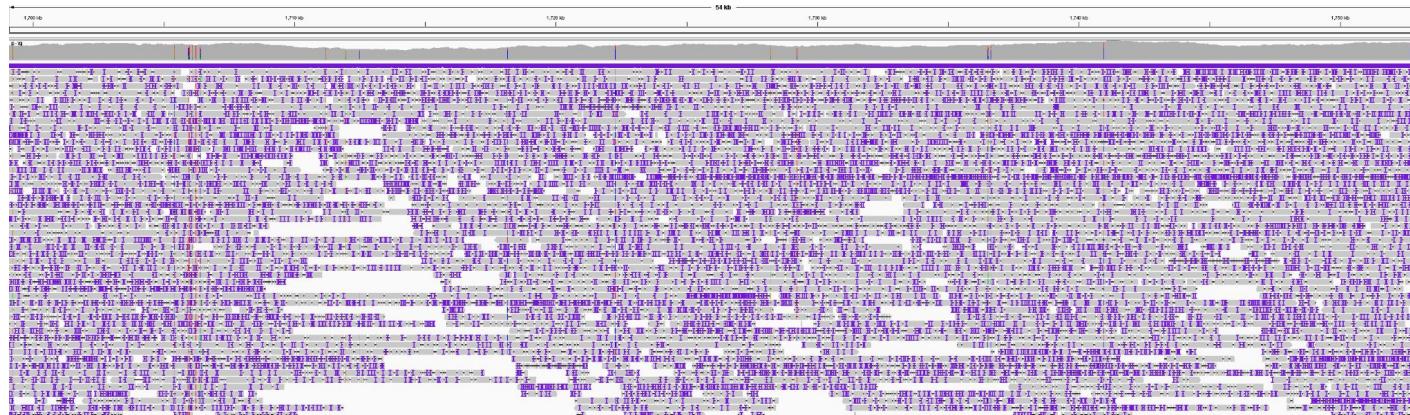
from QC to gene prediction and phylogenomics

BUSCO v5.3.2 is the current stable version!

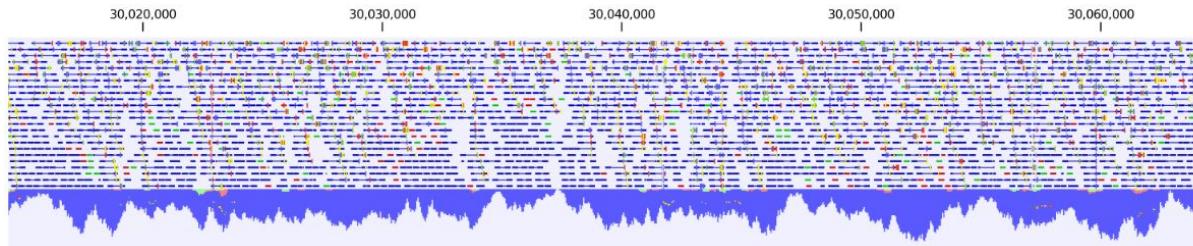
[Gitlab](#), a [Conda package](#) and [Docker container](#) are also available.

Evaluation - assembly correctness

- Check assembly correctness: analyses of read mappings
- Integrated Genomics Viewer (IGV) can visualize read mappings
- Tools: REAPR, SQUAT

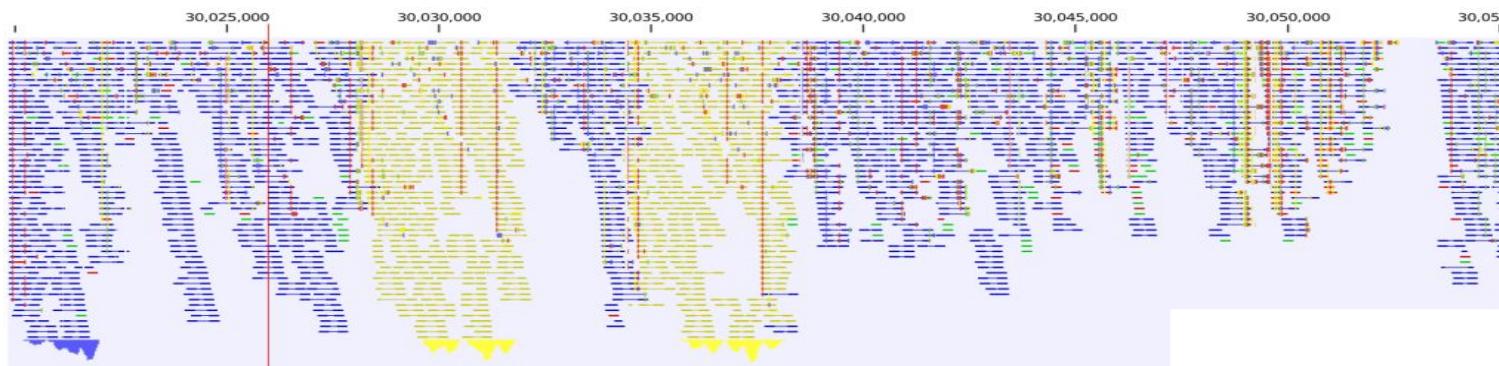


Read mappings (1)

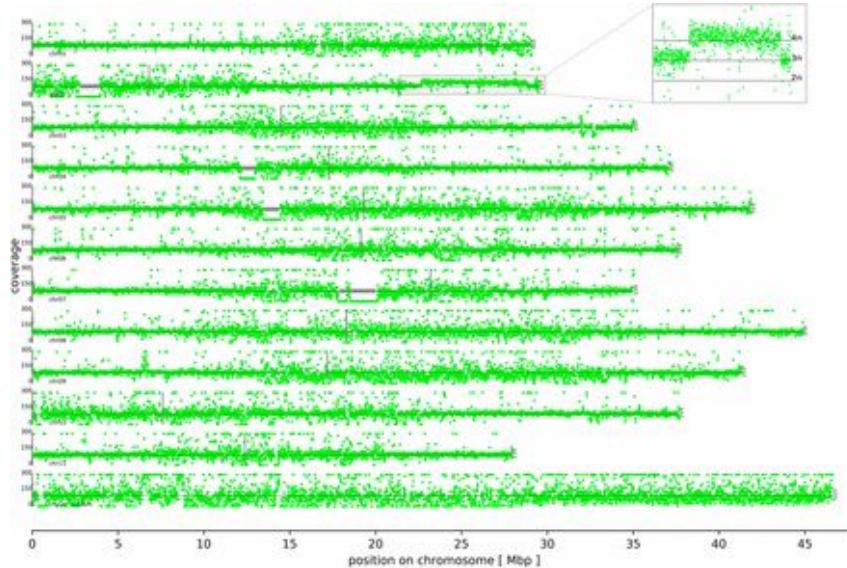


Read mapping of paired-end sequenced fragments (blue) to assembly

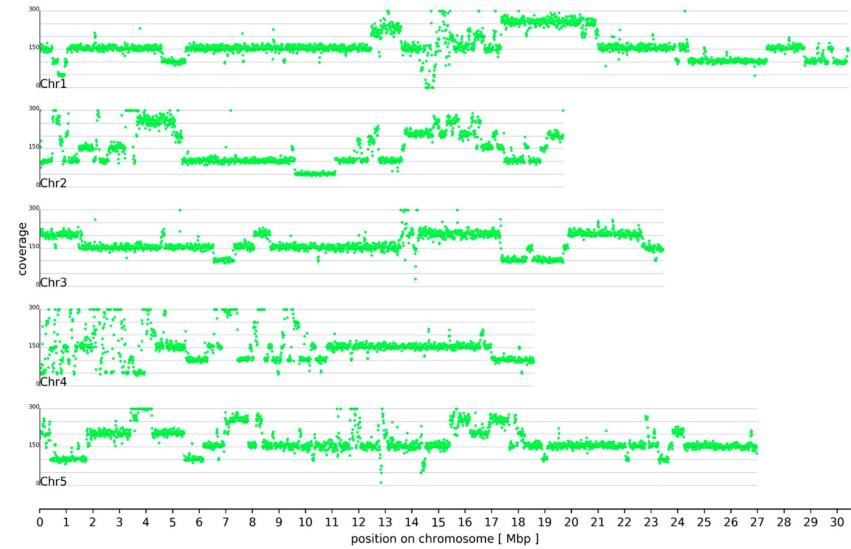
Coverage is too high to show all individual fragments at some positions



Read mappings (2)



Musa acuminata (banana) read mapping



Arabidopsis thaliana At7 read mapping

Advantages of long reads

- Span larger regions and enable assembly of repeats
- Specific mapping to repetitive regions possible
- Generation of larger contigs (no scaffolding)
- Contigs can represent entire chromosomes

Importance of error rate

- Long reads have been considered error-prone (noisy)
- Correction step is computationally extremely intense (HERRO)
 - Computation of all-vs-all alignments
- Direct assembly is possible with >99% raw read accuracy
- Higher accuracy allows to filter read overlaps more strictly

Long read assemblers

- NextDenovo2: <https://github.com/Nextomics/NextDenovo>
Hu et al., 2024: 10.1186/s13059-024-03252-4
- Verkko: <https://github.com/marbl/verkko>
Rautiainen et al., 2023: 10.1038/s41587-023-01662-6
- Hifiasm: <https://github.com/chhylp123/hifiasm>
Cheng et al., 2021: 10.1038/s41592-020-01056-5
- Shasta: <https://github.com/chanzuckerberg/shasta>
Shafin et al., 2020: 10.1038/s41587-020-0503-6

Important assembler features

- NextDenovo2: works well with R9 reads
- Verkko: can integrate Pore-C data for scaffolding
- Shasta: super fast and resource-efficient
- Hifiasm: excellent assembler for highly accurate long reads

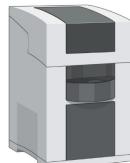
Integration of genetic linkage information (1)

- Classical genetic markers: SSR, CAPS, KASP
 - SSR = Simple Sequence Repeats
 - CAPS = Cleaved Amplified Polymorphic Sequences
 - KASP = Kompetitive Allele Specific PCR

SSR: Simple Sequence Repeats

A: CATAGAGAGAGAGAGAGATGAC
 B: CATAGAGAGAGAGATGAC
 C: CATAGAGAGATGAC
 D: CATAGAGAGAGAGAGAGATGAC
 E: CATAGAGAGAGAGATGAC
 F: CATAGAGAGAGAGAGAGATGAC
 G: ACATAGAGATGAC

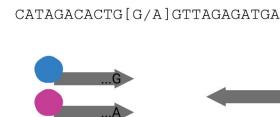
Length analysis via capillary electrophoresis



CAPS: Cleaved Amplified Polymorphic Sequences

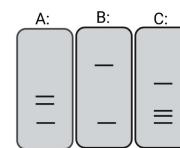
A:
 B:
 C:
 D:
 E:
 F:
 G:

KASP: Kompetitiv Allele Specific PCR



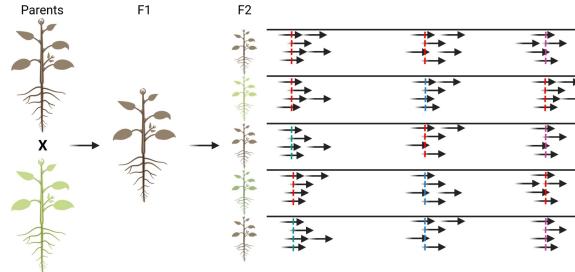
PCR with fluorescently labeled primers

Length analysis via capillary electrophoresis

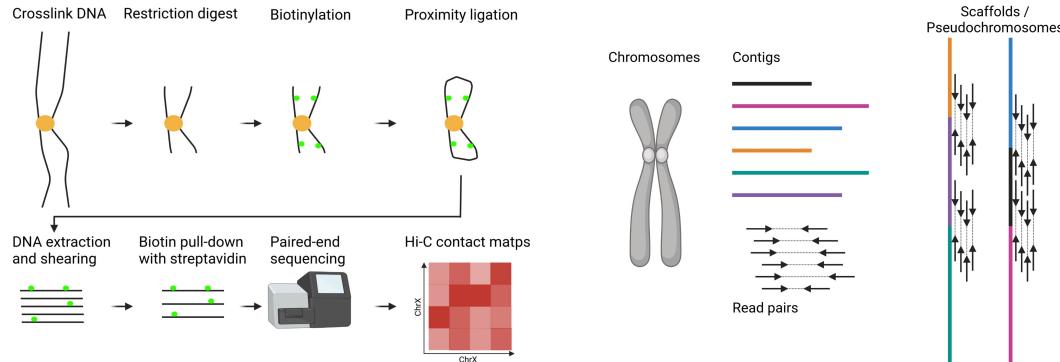


Integration of genetic linkage information (2)

- Genotyping-by-sequencing: SNPs inferred from sequencing data



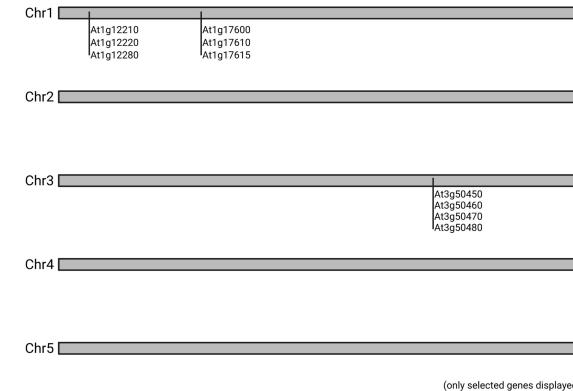
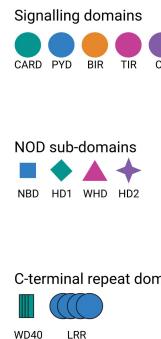
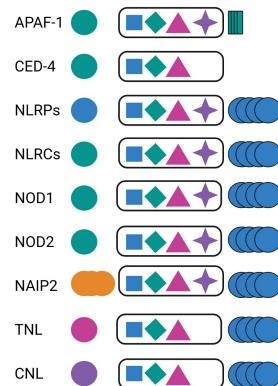
- Hi-C: chromatin interaction information for long range scaffolding



Checking challenging regions

- Resolving the centromeres and NORs are the last big challenges
- Centromeres = repeats in the middle of chromosomes
- Nucleolus Organizing Regions (NORs) = ribosomal RNA encoding repeats (rDNA)
- Checking presence of telomers at contig ends

- NLR genes (NL Rome) are often clustered in tandem repeat arrays
- NLR gene clusters are particularly tricky to assemble
- NL Rome used for benchmarking high quality assembly



Structural annotation



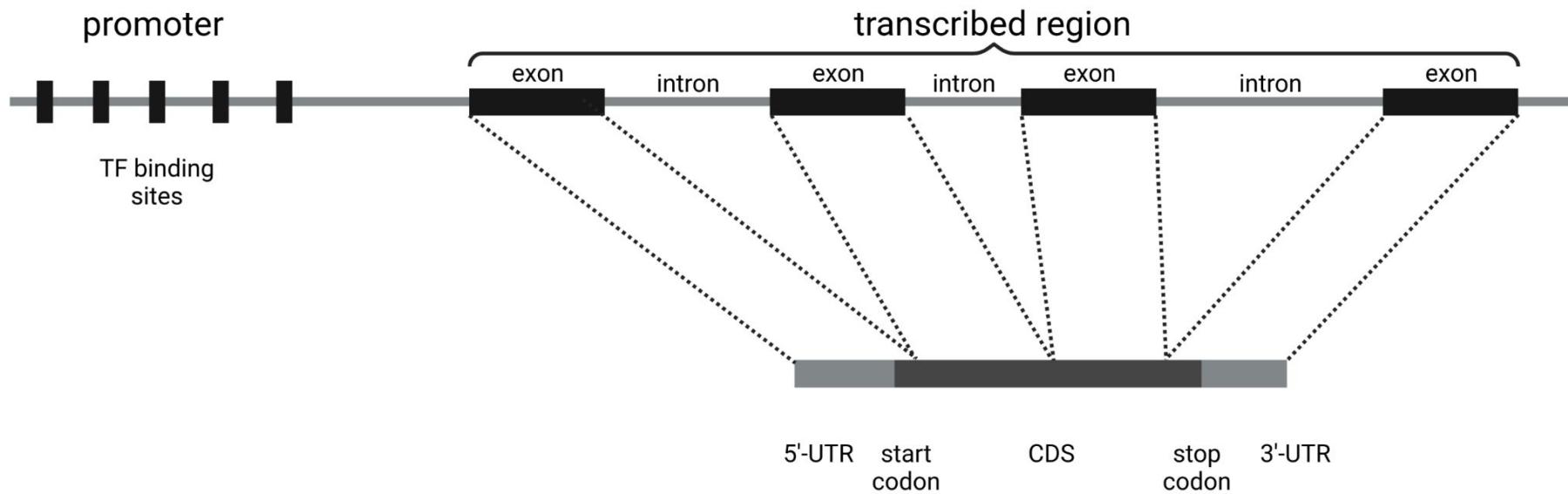
Finding genes in a genome sequence

UNIVERSITÄT BONN

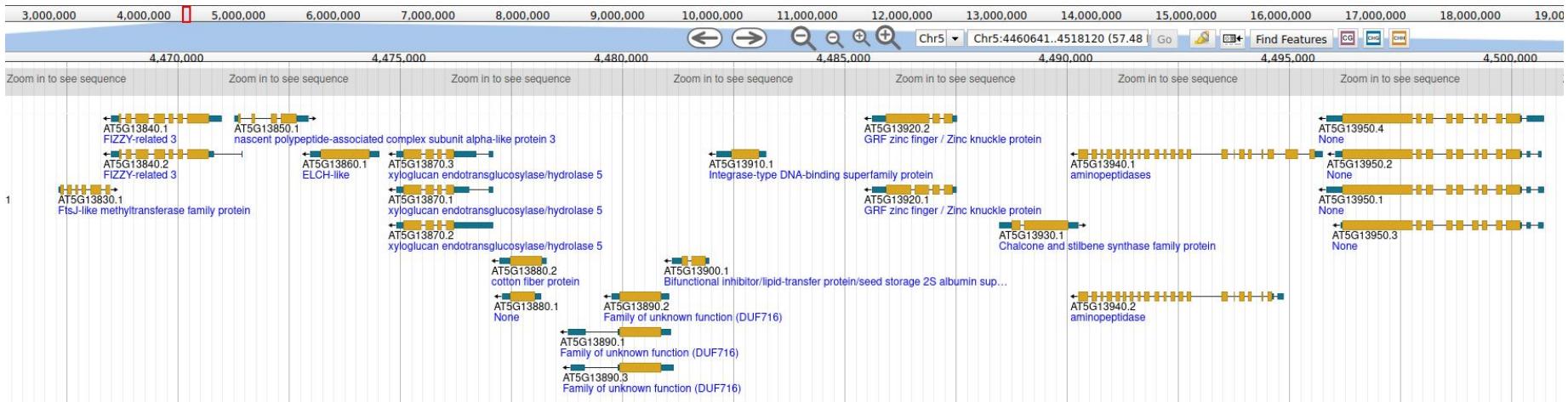
Plant gene structure

CDS = (Protein) Coding Sequence

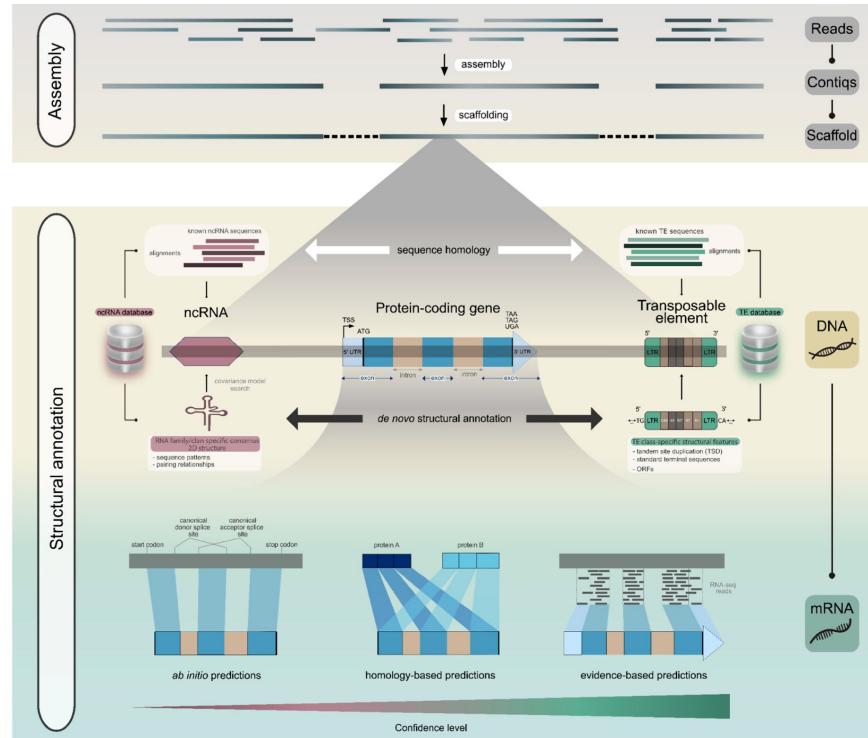
ORF = Open Reading Frame



Finding genes in a genome sequence

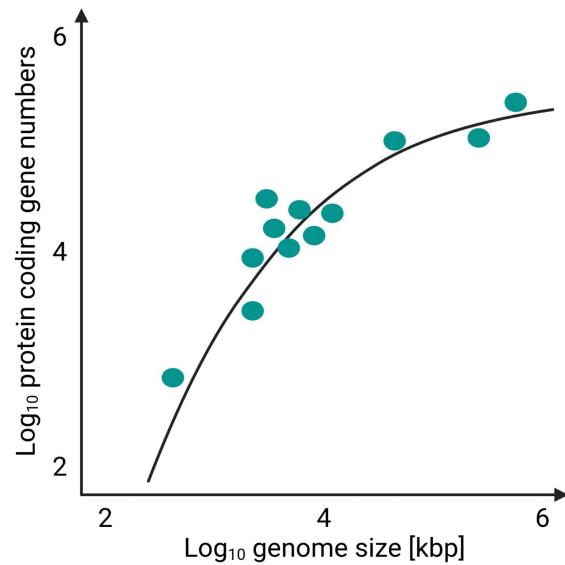


Structural annotation (overview)



Gene numbers

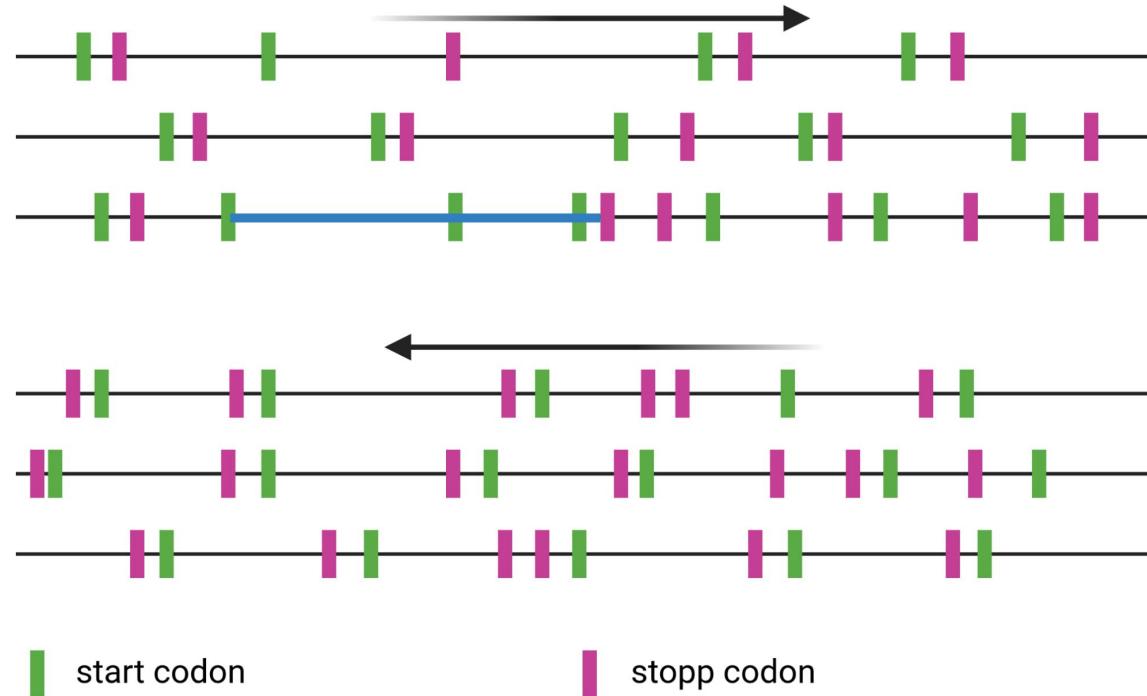
- Average number of genes in plants: 27200
- Gene number is not significantly correlated with genome size



Repeat masking

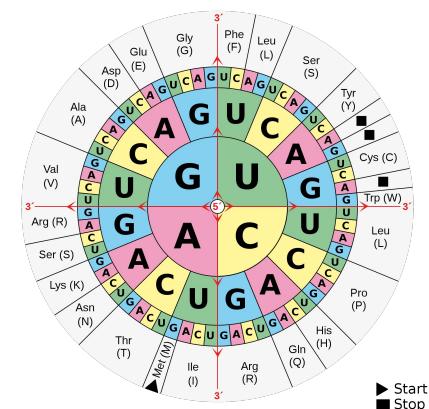
- Simple sequence repeats (SSR)
- Transposable elements (TE)
- Centromeric repeats (CEN)
- Telomeric repeats (TEL)

Finding ORFs



Codon usage

- Protein coding sequences have specific properties
- Codon usage is different between species i.e. different species prefer different codons for certain amino acids
- Rare codons slow down translation (useful between domains)
- Usage of dicodons (=hexamers)
- Codon usage can give additional CDS/ORF support



How does a Hidden Markov Model work?

Hidden Markov Models (1)

Result	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	1/2



Hidden Markov Models (2)

- Observation: 1-4-6-2-4-3-2-6-1-5
- Probability using fair die (F):
 - $(\frac{1}{6})^{10} = 1.6538 * 10^{-8}$
- Probability using loaded die (L):
 - $\frac{1}{10} * \frac{1}{10} * \frac{1}{2} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{2} * \frac{1}{10} * \frac{1}{10} = 2.7778 * 10^{-10}$



Hidden Markov Models (3)

- Observation:

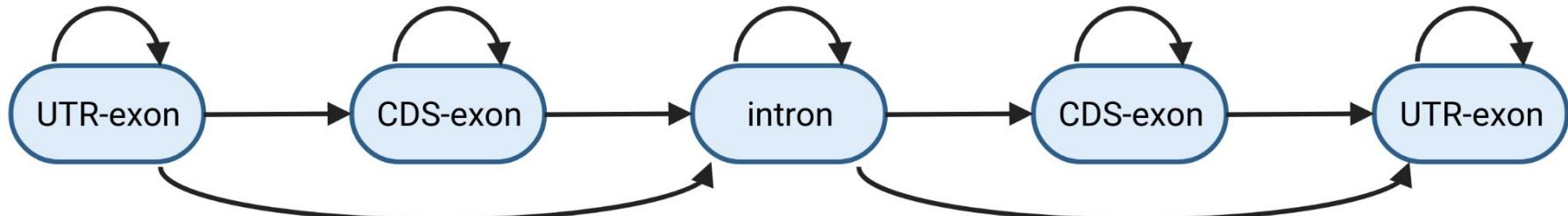
1625453666166152316342423315266261662656462624351354
FFFFFFFFFFLLLLLFFFFFFFFFLLLLLFFFLLLLLLFFFFFFFF



- How does this fit to gene prediction?

Hidden Markov Model for gene prediction

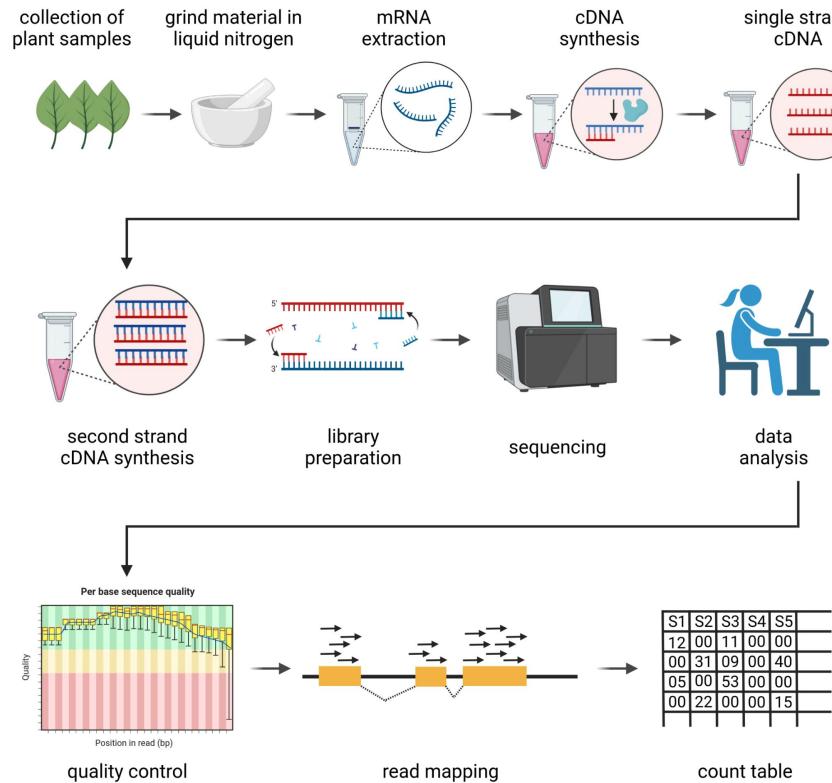
- Fair die = exon (GC-rich in plants)
- Loaded die = intron (AT-rich in plants)
- Switch between fair die and loaded die at intron/exon borders



Parameters for *ab initio* gene prediction

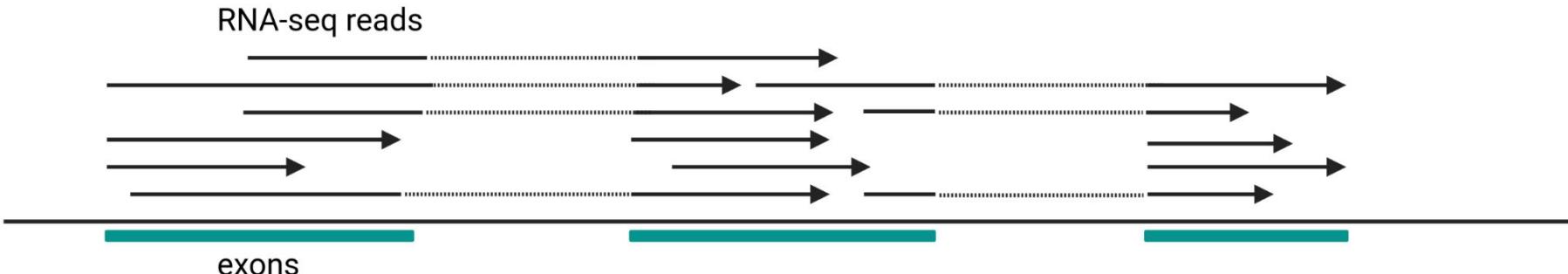
- Features: utr, exon, intron, intergenic region
- Composition of different features (dicodons)
- feature lengths
- Number of features per gene
- Possible splice sites

RNA-seq



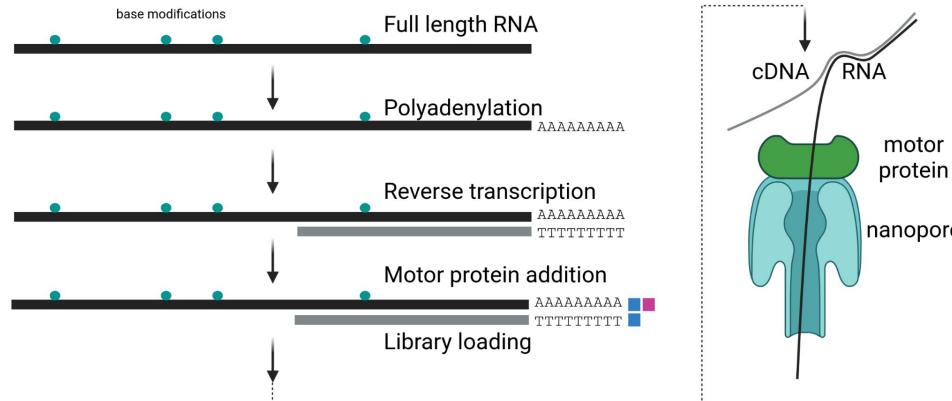
RNA-seq hints

- Aligned RNA-seq reads indicate exon positions
- Splitting of reads indicates intron positions
- CDS can be identified as ORF within the covered regions



Full length transcript sequences

- Capturing full length transcripts for RNA-seq (Illumina)
- Full length cDNA sequencing (PacBio, ONT)
- Direct RNA sequencing (ONT)



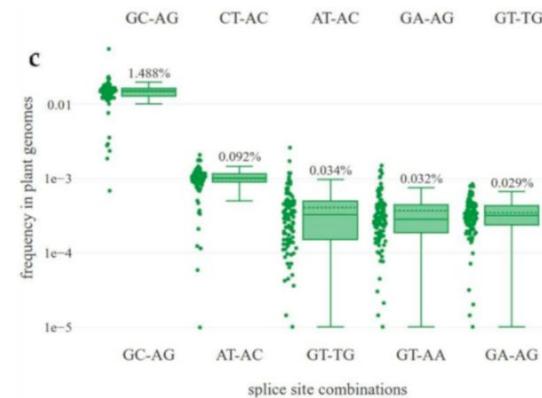
Canonical splice sites

- Spliceosome catalyzes the removal of introns from pre-mRNA
- Recognition of intron/exon borders required
- Splice sites are recognition sites at the intron borders

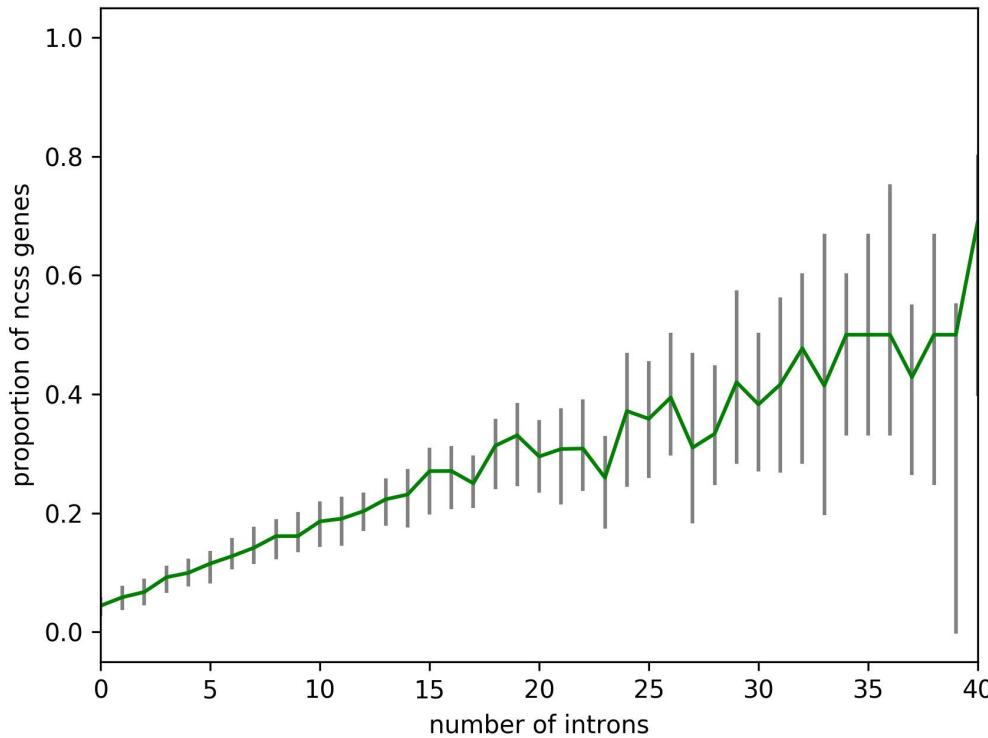


Non-canonical splice sites

- Splice sites mark points for removal of introns
- Highly conserved to allow recognition by spliceosome
- Two types of spliceosomes: U2 and U12
- Canonical splice sites (98%): GT-AG
- Major non-canonical splice sites (1-2%): GC-AG, AT-AC
- Minor non-canonical splice sites (<1%): NN-NN



Relevance of non-canonical splice sites



Proteins as hints (exonerate)

- Exonrate can align coding sequences/peptide sequences to DNA
- Intron positions are identified and intron information is given
- Exonrate can handle different types of splice sites

```

1 : ATGGTGATGGCTGGCTTCTTCTGGATGAGATCAGACAGGCTCAGAGAGCTGA : 56
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488762 : ATGGTGATGGCTGGCTTCTTCTGGATGAGATCAGACAGGCTCAGAGAGCTGA : 4488817

57 : TGGACCTGCAGGCATCTGGCTATTGGCACTGCTAACCCCTGAGAACCATGTGCTTC : 112
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488818 : TGGACCTGCAGGCATCTGGCTATTGGCACTGCTAACCCCTGAGAACCATGTGCTTC : 4488873

113 : AGGCCGGAGTATCCGTACTACTCTCGCATCACCAACAGTGAACACATGACCGAC : 168
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488874 : AGGCCGGAGTATCCGTACTACTCTCGCATCACCAACAGTGAACACATGACCGAC : 4488929

169 : CTCAAAGGAAAGTTCAAGGCATGT >>> Target Intron 1 >>> GC : 195
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488930 : CTCAAAGGAAAGTTCAAGGCATGTgt.....agGC : 4489042

196 : GACAAGTGCACAATTGGAAACGGTCACTGCACTGACGGAGGGATTCTCAAGGA : 251
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4489043 : GACAAGTGCACAATTGGAAACGGTCACTGCACTGACGGAGGGATTCTCAAGGA : 4489098

252 : AAACCACACATGGTGTCTTACATGGCTCTCTCTGGACACAGACAGGACATCG : 307
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4489099 : AAACCACACATGGTGTCTTACATGGCTCTCTCTGGACACAGACAGGACATCG : 4489154

308 : TGTTGGTGAAGGCCCTAACGTAAGGCAAAGAAGCGGCACTGAGGGCATCAAGGG : 363
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4489155 : TGTTGGTGAAGGCCCTAACGTAAGGCAAAGAAGCGGCACTGAGGGCATCAAGGG : 4489210

```

```

1 : MetValMetAlaGlyAlaSerSerLeuAspGluIleArgGlnAlaGlnArgAlaAs : 19
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488762 : ATGGTGATGGCTGGCTTCTTCTGGATGAGATCAGACAGGCTCAGAGAGCTGA : 4488816

20 : pglyProAlaGlyIleLeuAlaIleGlyThrAlaAsnProGluAsnHisLysValLeuG : 38
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488817 : TGGAACCTGCAGCATTCTTGCACTTGGACATCTACCTGGAGAACCATGTGCTTC : 4488873

39 : lIn1LaGluTyProlAspTyTyrPheArgIleThrAsnSerGluLysMetThrAsp : 56
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488874 : AGGCCGGAGTATCCGTACTACTCTCGCATCACCAACAGTGAACACATGACCGAC : 4488927

57 : LeuLysGluLysPhelysPheLysArgMet(C) >>> Target Intron 1 >>> : 65
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4488928 : CTCAAGGAGAAAGTTCAAGGCATGTgt.....ag : 4489039

66 : {ysAspLysSerThrIleArgLysArgLysMetLysLeuThrGluLysLeuLysLeuLys : 83
{|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4489040 : {ysAspLysSerThrIleArgLysArgLysMetLysLeuThrGluLysLeuLysLeuLys : 4489094

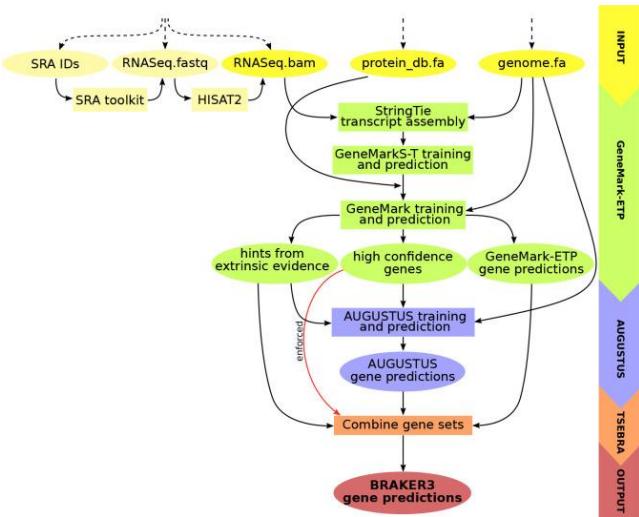
84 : ysGluLysSerLysMetCysAlaLysArgLysMetLysLeuThrGluLysLeuLysLeuLys : 101
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4489095 : AGGAAACCCACACATGGTGTCTTACATGGCTCTCTGGACACAGACAGGACATCG : 4489148

102 : IleLeuValValLysLeuLysLeuLysLeuLysLeuLysLeuLysLeuLysLeuLys : 120
|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:|
4489149 : ATCGTGTGGTGAAGGCCCTAACGTAAGGCAAAGAAGCGGCACTGAGGGCATCAAGGG : 4489205

```

BRAKER3

- Fully automated pipeline for predicting protein-coding genes in eukaryotic genomes
- Combines RNA-seq data and protein homology evidence for higher accuracy
- GeneMark-ETP (training), AUGUSTUS (prediction), TSEBRA (result selection)
- Inputs: soft-masked genome sequence, RNA-seq reads, and protein database
- Supports containerized (Docker/Singularity) runs



- GeMoMa (Gene Model Mapper) uses homologs from closely related species
- Transfers and refines intron-exon-structures from reference
- Can combine RNA-seq data with homology to improve accuracy
- Produces high-confidence gene models (GFF/GTF)
- Performs well when no transcriptome data is available

Annotation quality assessment

- BUSCO (Benchmarking Universal Single-Copy Orthologs)
 - Example: C:98%[S:85%,D13%],F:1.2%,M0.8%,n:1500
 - C = complete BUSCO genes
 - S = single copy BUSCO genes
 - D = duplicates BUSCO genes
 - F = fragmented BUSCO genes
 - M = missing BUSCO genes
 - n = number of BUSCO query sequences
- NL Rome: assessment of high continuity genome sequences

Identification of tRNAs

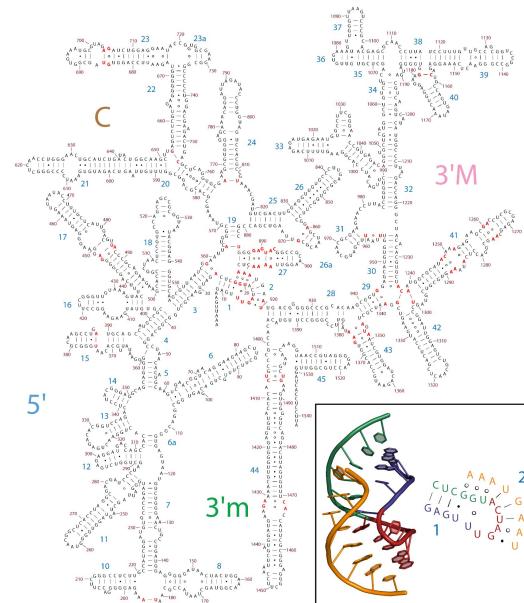
- tRNA genes have RNA gene products
- tRNAscan-SE2 is a dedicated tool for the annotation of tRNAs
- GtRNAdb contains the results of tRNAscan-SE (<http://gtrnadb.ucsc.edu/>)



Figure source: https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png

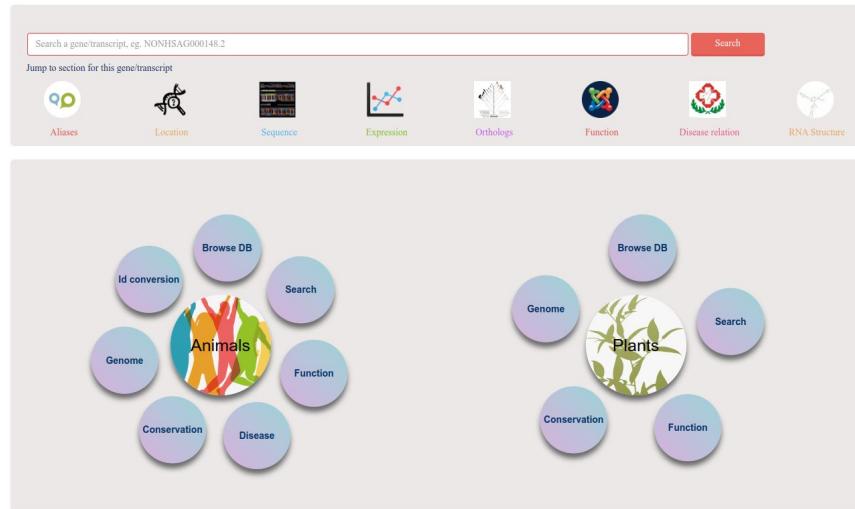
Identification of rRNAs

- Rfam: database of RNA genes (<http://rfam.xfam.org/>)
- RepeatMasker can identify rRNAs
- RNAmmer is dedicated tool for annotation of rRNAs



Annotation of long non-coding RNAs (lncRNAs)

- NONCODEV6: long non-coding RNA database (<http://www.noncode.org/>)
- Identification based on sequence similarity



Transposable elements (TEs)

- Transposons shape plant genomes (genome obesity)
- Systematics:
 - Class I (retrotransposons)
 - LTR: Copia, Gypsy, Bel-Pao, Retrovirus, ERV
 - DIRS: DIRS, Ngaro, VIPER
 - PLE: Penelope
 - LNE: R2, RTE, Jockey, L1, I
 - SINE: tRNA, 7SL, 5S
 - Class II (DNA transposons) - Subclass 1
 - TIR: Tc1-Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF-Harbinger, CACTA
 - Crypton
 - Class II (DNA transposons) - Subclass 2
 - Helitron
 - Maverick

Annotation of TEs

- Benchmarking study of TE annotation tools: <https://github.com/oushujun/EDTA>
- RepeatMasker: <https://www.repeatmasker.org/>
 - Screens genomic sequence for TEs
 - Soft/hard masking of genomic sequence
 - Dfam and Repbase are important databases
 - Dfam: open collection of TE sequences (<https://www.dfam.org/home>)
 - Repbase: subscription-based collection (<https://www.girinst.org/>)
 - Different search engines can be used
- RepeatModeler2: <http://www.repeatmasker.org/RepeatModeler>
 - Pipeline for discovery of TEs
- Extensive *de novo* TE Annotator (EDTA): <https://github.com/oushujun/EDTA>
 - Complex pipeline for TE annotation

Sharing structural annotation

- Species specific databases:
 - TAIR: *Arabidopsis thaliana*
 - BananGenomeHub: *Musa acuminata*
- EMBL/EBI: European Nucleotide Archive of the European Bioinformatics Institute (<https://www.ebi.ac.uk/ena/browser/home>)
- PLAZA (<https://bioinformatics.psb.ugent.be/plaza/>)
- Phytozome (<https://phytozome-next.jgi.doe.gov/>)

Functional annotation

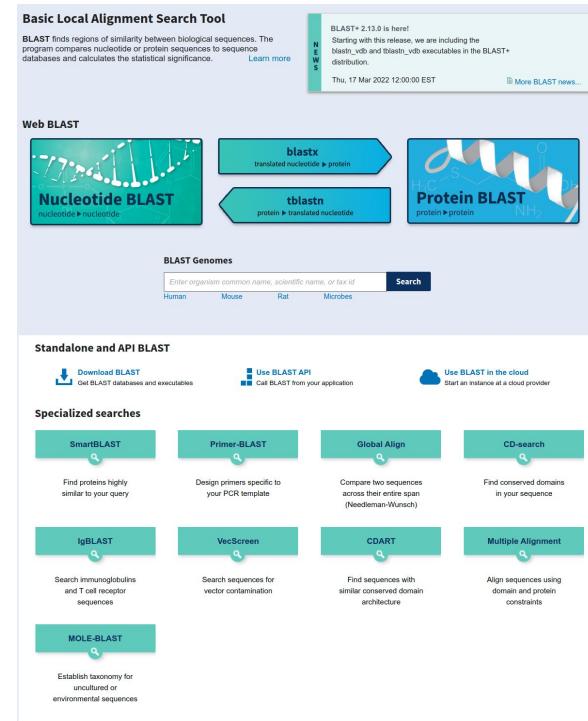
Functional annotation

- What is the function of a gene?
- Knockout experiments for all genes are time consuming and expensive
- Annotation transfer: orthologs are assumed to have the same function
- Tools:
 - BLAST: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
 - Pfam: <https://pfam.xfam.org/>
 - InterProScan5: <http://www.ebi.ac.uk/interpro/search/sequence/>
 - Shoot: <https://github.com/davidemms/SHOOT>
 - KEGG: <https://www.genome.jp/kegg/>
 - GO: <http://geneontology.org/>
 - MetaCyc: <https://metacyc.org/>
 - KIPES: <https://github.com/bpucker/KIPES>
 - BRENDA: <https://www.brenda-enzymes.org/>
 - Mercator: <https://plabipd.de/portal/mercator4>

Altschul et al., 1990: 10.1016/S0022-2836(05)80360-2
Mistry et al., 2021: 10.1093/nar/gkaa913
Jones et al., 2014: 10.1093/bioinformatics/btu031
Karp et al., 2002: 10.1093/nar/30.1.59
Emms & Kelly, 2022: 10.1186/s13059-022-02652-8
Kanehisa & Goto, 2000: 10.1093/nar/28.1.27
Ashburner et al., 2000: 10.1038/75556
Pucker et al., 2020: 10.3390/plants9091103
Schomburg et al., 2002: 10.1093/nar/30.1.47
Schwacke et al., 2019: 10.1016/j.molp.2019.01.003

BLAST: Basic Local Alignment Search Tool

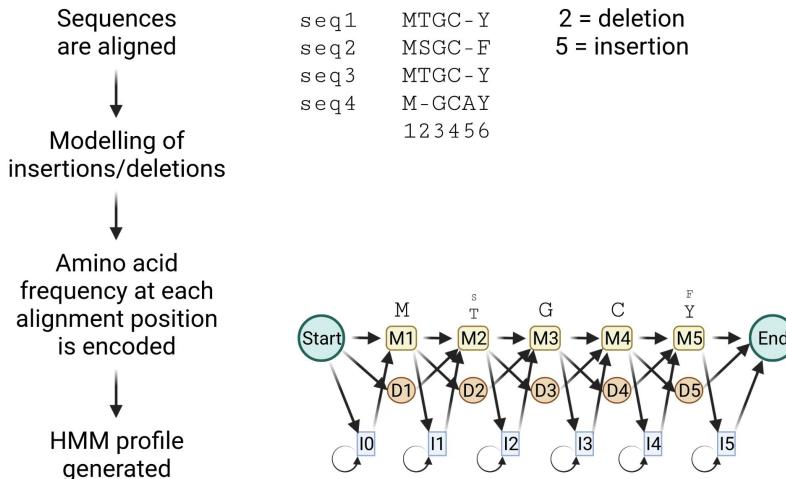
- Probably the most famous website of the NCBI
- Comparison of sequences against a large database
- Similar sequences are likely to have similar functions (ideally orthologs)
- Numerous variants of the initial BLASTn were developed



The screenshot shows the NCBI BLAST homepage. At the top right, there is a news banner about BLAST+ 2.13.0. Below it, the main navigation menu includes "Web BLAST", "Standalone and API BLAST", and "Specialized searches". Under "Web BLAST", there are links for "Nucleotide BLAST" (nucleotide → nucleotide), "blastx" (translated nucleotide → protein), and "tblastn" (protein → translated nucleotide). Under "Standalone and API BLAST", there are links for "Download BLAST", "Use BLAST API", and "Use BLAST in the cloud". Under "Specialized searches", there are nine boxes: SmartBLAST, Primer-BLAST, Global Align, CD-search, IgBLAST, VecScreen, CDART, Multiple Alignment, and MOLE-BLAST. Each box has a brief description of its function.

Pfam: Protein family database

- Assignment of protein functions based on Hidden Markov Models (HMMs)
- Sequences are screened based on HMM profile



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam 35.0 (November 2021, 19632 entries)

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). [More...](#)

QUICK LINKS YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

SEQUENCE SEARCH Analyze your protein sequence for Pfam matches

VIEW A PFAM ENTRY View Pfam annotation and alignments

VIEW A CLAN See groups of related entries

VIEW A SEQUENCE Look at the domain organisation of a protein sequence

VIEW A STRUCTURE Find the domains on a PDB structure

KEYWORD SEARCH Query Pfam by keywords

JUMP TO enter any accession or ID Go Example

Or view the [help](#) pages for more information

Recent Pfam blog posts

[Pfam 35.0 is released!](#) (posted 19 November 2021)

Hide this

Pfam 35.0 contains a total of 19,632 families and clans. Since the last release, we have built 460 new families, killed 7 families and created 12 new clans. UniProt Reference Proteomes has increased by 7% since Pfam 34.0, and now contains 61 million sequences. Of the sequences that are in UniProt Reference Proteomes, 75.2% have 1 [...]

[AlphaFolding the Protein Universe](#) (posted 22 July 2021)

Hot on the tail of our inclusion of the Baker group's Rosetta structural models we are excited to announce the inclusion of models from AlphaFold 2.0 generated by DeepMind and stored in the AlphaFold Database (AlphaFold DB). AlphaFold 2.0's performance in the CASP14 competition was spectacular, producing near experimental quality structure models. The new AlphaFold [...]

[Google Research Team bring Deep Learning to Pfam](#) (posted 24 March 2021)

We are delighted to announce the first fruits of a collaboration between the Pfam team and a Google Research team led by Dr Lucy Colwell, with Maxwell Bileshi and David Belanger. In 2019, Colwell's team published a preprint describing a new deep learning method that was trained on Pfam data, and which improves upon the [...]

Citing Pfam

If you find Pfam useful, please consider [citing](#) the reference that describes this work:

Pfam: The protein families database in 2021 (J. Mistry, S. Chugravsky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonhammer, S.C.E. Tosato, L. Paladini, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman, *Nucleic Acids Research* (2020) doi: 10.1093/nar/gkaa913

Pfam is part of the ELIXIR infrastructure
 ELIXIR is an ELI service. [Read more](#)
 Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
 European Molecular Biology Laboratory

- Screen of protein sequences against collection of protein signatures
- Allows the assignment of functional annotation terms
- Available as web service, but also as stand alone tool

InterProScan 5 Sequence Search

This form allows you to scan your sequence for matches against the InterPro collection of protein signature databases.

This form is for debugging purposes only and is not supported. To submit jobs to InterProScan 5, please visit the InterPro Sequence Search or the InterProScan 5 Web services.

Please Note

This web form is for internal debugging purposes and will be retired in April 2016. If you have any questions/concerns please contact us via the feedback link above.

STEP 1 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:
uniprot:KPYM_HUMAN

Or, upload a file: Choose file | No file chosen Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Select the applications to run

<input checked="" type="checkbox"/> TIGRGRAM	<input checked="" type="checkbox"/> SPLD	<input checked="" type="checkbox"/> Prokline	<input checked="" type="checkbox"/> SignalP	<input checked="" type="checkbox"/> SignalP_EUK
<input checked="" type="checkbox"/> SignalP_GRAM_POSITIVE	<input checked="" type="checkbox"/> SignalP_GRAM_NEGATIVE	<input checked="" type="checkbox"/> SUPERFAMILY	<input checked="" type="checkbox"/> PANTHER	<input checked="" type="checkbox"/> GeneID
<input checked="" type="checkbox"/> Hmmp	<input checked="" type="checkbox"/> ProSiteProfiles	<input checked="" type="checkbox"/> ProSitePatterns	<input checked="" type="checkbox"/> Coils	<input checked="" type="checkbox"/> SMART
<input checked="" type="checkbox"/> CDD	<input checked="" type="checkbox"/> PRINTS	<input checked="" type="checkbox"/> Pfam	<input checked="" type="checkbox"/> MotifDBlite	<input checked="" type="checkbox"/> PSIP
<input checked="" type="checkbox"/> TMHMM				

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

Results for job iprscan5-I20220320-102557-0980-87120812-p2m

Tool Output **Submission Details**

Download in XML format	Download in TSV format	Download in GFF3 format	Download in SVG format	Download HTML tarball file	Download in JSON format
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	Gene3D G3DSA:3.40.47.10	- 1 241 4.8E-101 T 20-03-2022	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	SUPERFAMILY SSF53901	Thiolase-like 241 393 2.98E-51 T 20-03-2022	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	SUPERFAMILY SSF53901	Thiolase-like 10 237 2.38E-78 T 20-03-2022	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	Pfam PF02797 Chalcone and stilbene synthases, C-terminal domain	244 394 1.5E-71	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	PANTHER PTHR11877:SF81 BNAA02G30320D PROTEIN	6 394 0.0 T 20-03-2022	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	Gene3D G3DSA:3.40.47.10	- 242 395 9.7E-62 T 20-03-2022 IPR01603	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	PANTHER PTHR11877 HYDROXYMETHYLGLUTARYL-COA SYNTHASE	6 394 0.0 T	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	CDD cd00831 CHS_like	21 390 0.0 T 20-03-2022	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	PIRSF PIRESF000451 PKS III	7 394 0.0 T 20-03-2022 IPR011141	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	Pfam PF00195 Chalcone and stilbene synthases, N-terminal domain	10 233 2.9E-119	
sp P13114 CHSY_ARATH ba46b0f06bfb1c91d161f0f91310f43	ba46b0f06bfb1c91d161f0f91310f43	395	ProSitePatterns PS00441 Chalcone and stilbene synthases active site.	161 177 -	

Shoot

- Initial search based on sequence similarity
- Phylogenetic relationships of sequences are considered based on a tree
- Universal tool, but computationally more intensive than a simple sequence similarity analysis

SHOOT.bio - the phylogenetic search engine

SHOOT is a phylogenetic alternative to BLAST. Instead of returning a list of similar sequences to a query sequence it returns a maximum likelihood phylogenetic tree with your query sequence embedded in it.

Try it out: <https://shoot.bio/>

Preprint: <https://www.biorxiv.org/content/10.1101/2021.09.01.458564>

Using the SHOOT command line tool

SHOOT allows you to search a protein sequence against a database of gene trees. It returns your gene grafted into the correct position within its corresponding gene tree.

Preparing a SHOOT phylogenetic database

0. Install dependencies:
 - Python libraries: ete3, sklearn, biopython
 - DIAMOND
 - MAFFT
 - EPA-ng & gappa (<https://github.com/lczech/gappa>)
 - Alternatively, IQ-TREE can be used instead of the combination EPA-ng + gappa
1. Run an OrthoFinder analysis on your chosen species, using the multiple sequence alignment option for tree inference, “-M msa”.
 - Paper: Emms, D.M., Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20, 238 (2019). <https://doi.org/10.1186/s13059-019-1832-y>
 - GitHub: <https://github.com/davideemms/OrthoFinder>
 - Tutorials: <https://davideemms.github.io/>
2. Run `python create_shoot_db.py RESULTS_DIRECTORY`, replacing “RESULTS_DIRECTORY” with the path to the OrthoFinder results directory from step 1.
3. Resolve polytomies (only necessary if using EPA-ng): `python bifurcating_trees.py RESULTS_DIRECTORY`

The OrthoFinder RESULTS_DIRECTORY is now a SHOOT database.

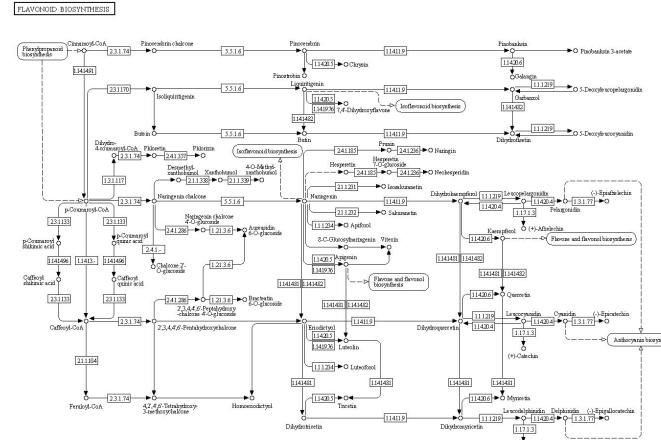
Running SHOOT

```
python shoot INPUT_FASTA SHOOT_DB
```

where INPUT_FASTA is a fasta file containing the amino acid sequence for the search and SHOOT_DB is the SHOOT database directory created using the steps above.

KEGG: Kyoto Encyclopedia of Genes and Genomes

- Maps of pathways showing the individual reactions with catalyzing enzymes
- Information about genomes and genes
- Chemical details about enzymes, substrates, and products
- KEGG is financed through a subscription model (for FTP download), but website is freely accessible



map00941 Flavonoid biosynthesis

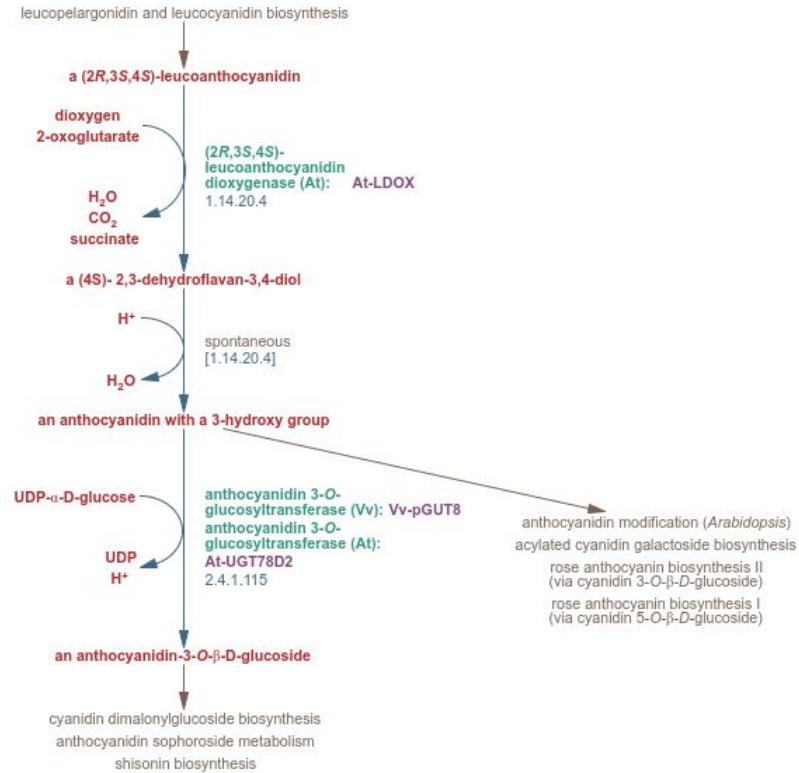
Gene Ontology (GO)

- Defined statements about the function of a gene (controlled vocabulary)
- Hierarchical structure
 - Example: 'metabolic process' > 'biosynthetic process' > ... > 'chalcone synthase'
- Supported by the Alliance of Genome Resources
- Connected with various other databases e.g. TAIR, FlyBase, Reactome, UniProt
- Machine readable to allow automatic processing
- Tools: Blast2GO and AmiGO
 - analyze the function of a sequence (web service and standalone)

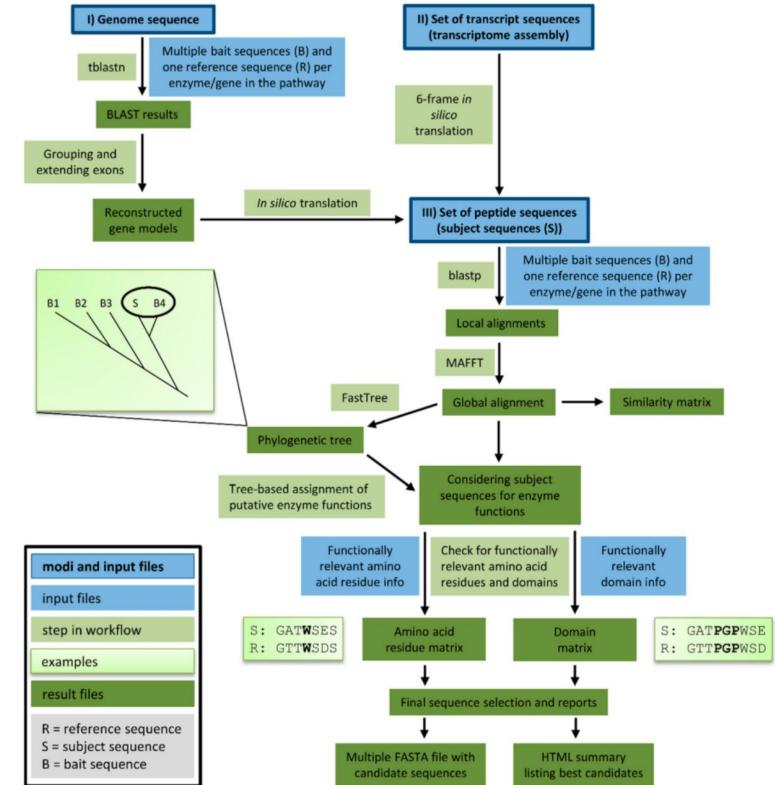
- GO:0008150 biological_process
- GO:0008152 metabolic process
- GO:0009058 biosynthetic process
- GO:0071704 organic substance metabolic process
- GO:0009812 flavonoid metabolic process
- GO:1901576 organic substance biosynthetic process
- GO:0009813 flavonoid biosynthetic process**
 - GO:0009718 anthocyanin-containing compound biosynthetic process
 - GO:0051551 aurone biosynthetic process
 - GO:0033485 cyanidin 3-O-glucoside biosynthetic process
 - GO:0033486 delphinidin 3-O-glucoside biosynthetic process
 - GO:0051553 flavone biosynthetic process
 - GO:0009716 flavonoid phytoalexin biosynthetic process
 - GO:0051557 leucoanthocyanidin biosynthetic process
 - GO:0009964 negative regulation of flavonoid biosynthetic process
 - GO:0033487 pelargonidin 3-O-glucoside biosynthetic process
 - GO:0009963 positive regulation of flavonoid biosynthetic process
 - GO:0009962 regulation of flavonoid biosynthetic process

MetaCyc: Metabolite Encyclopedia

- Integrates genomic data with functional annotation
- Visualization of pathway databases
- Shows intermediates and enzymes of biosynthesis pathways
- MetaFlux: flux-balance analysis

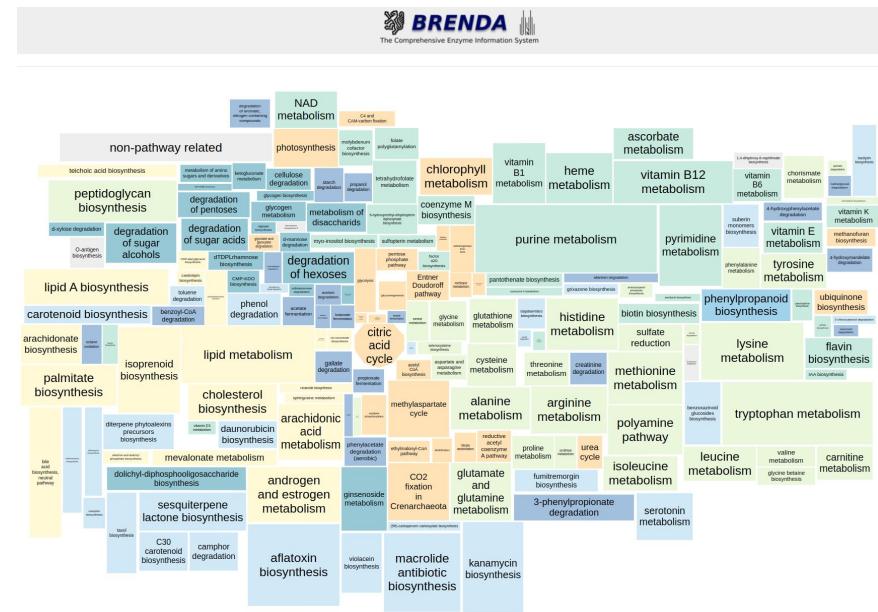


- Identification of all players in a biosynthesis pathway
- Dedicated to the identification of enzymes
- Functionally relevant amino acids need to be known
- Web server: <http://www.pbb-tools.de>



BRENDA: BRaunschweig ENzyme DAtabase

- Enzyme database hosted at TU Braunschweig (BRICS)
- Text and structure-based queries
- Visualization of pathways
- Manual curation of datasets
- Many details about enzyme properties (substrates, kinetics, mutants, ...)



- Online tool for the annotation of all protein sequences in a submitted FASTA file
- FASTA format:
 - Header line starts with ‘>’ followed by the sequence name
 - Header is followed by unrestricted number of sequence lines

```
>TRINITY_DN100013_c0_g1_i1
MIGPMPGMEGKMLPAGSPVGLEVVLSQLVVASTPILSDTSLCTPSFYHFLLLGPTSISNLIVRPFFLTLSSITVIYGRLCFFAFDFYVY
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQQNNQRRIKRKGRRPPKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHIPPTSIPSFNNPNNIHQSSSDRQMPP
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTATGDKKRTKARKETYSSYIYKVLKVHPDTGISNRAMSILNSFVNDIFERVATEASKLA
>TRINITY_DN10001_c0_g1_i2
MAKVGNPIVDIETDGSVNEPESSEKNIEVSSSSTQAESTNTTELLVNEKKAFSLATPAVRRVAREHNIDINNIKGTGKNGRITKEDILNYV
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFPKPGWKEFVRSNDLEGDFLVNLVDKISYQVVIFDGTCCPKDLCPPSIMNPIFQHLRNKIFLSKKEEIKLKGNRKVHSVNEN
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFAFVTAGYPDSEETVDILLGLEAGGADIELGIPFTDPMVDGKTIQDANNVALENKIDISKCLSYVSESRAK
```

Summary

- Genome sequence assembly
- Assembly quality evaluation (completeness, contiguity, correctness)
- Structural annotation (GeMoMa, BRAKER3)
 - ab initio
 - hint-based
- Functional annotation (KIPES, InterProScan5)

Time for questions!

Literature

- de Oliveira, J. A. V. S.; Choudhary, N.; Meckoni, S. N.; Nowak, M. S.; Hagedorn, M.; Pucker, B. (2025). Cookbook for Plant Genome Sequences. doi: [10.20944/preprints202508.1176.v2](https://doi.org/10.20944/preprints202508.1176.v2).
- Wolff, K.; Friedhoff, R.; Schwarzer, F.; Pucker, B. (2023). Data Literacy in Genome Research. *Journal of Integrative Bioinformatics*, 2023, pp. 20230033. doi: [10.1515/jib-2023-0033](https://doi.org/10.1515/jib-2023-0033).
- Pucker B, Irisarri I, de Vries J and Xu B (2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology*, 3, E5. doi: [10.1017/qpb.2021.18](https://doi.org/10.1017/qpb.2021.18).