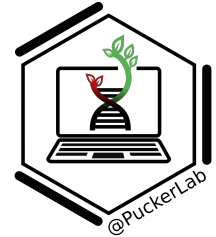
UNIVERSITÄT BONN



Prof. Dr. Boas Pucker

Long Read Genomics

- Generating a Structural & Functional Annotation

UNIVERSITÄT BONN

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

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

UNIVERSITÄT BONN

Repetition

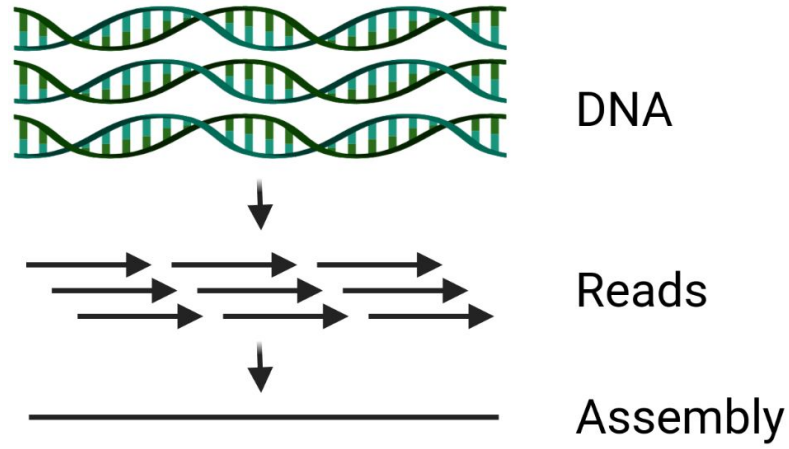
- Planning the sequencing experiment
- DNA extraction, quality control, library preparation
- Nanopore sequencing & real time monitoring
- Data analysis (overview)
- Nanopore sequencing applications
- Genome sequencing projects (examples)

UNIVERSITÄT BONN

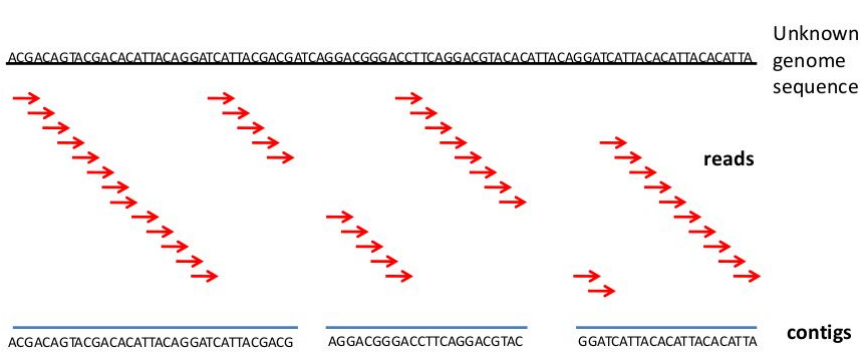
Genome Sequence Assembly

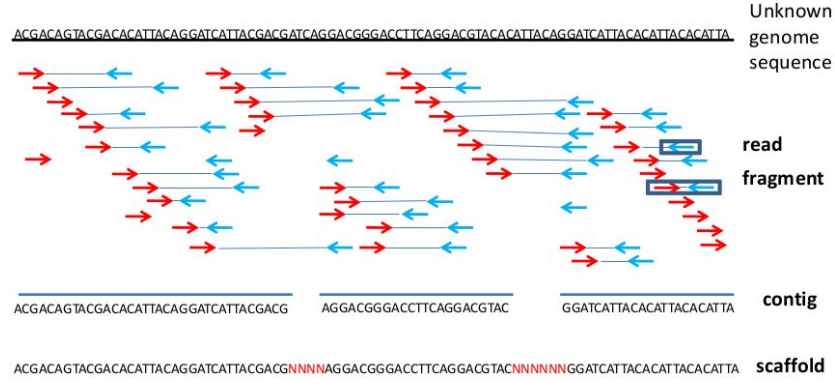
The assembly problem

- 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

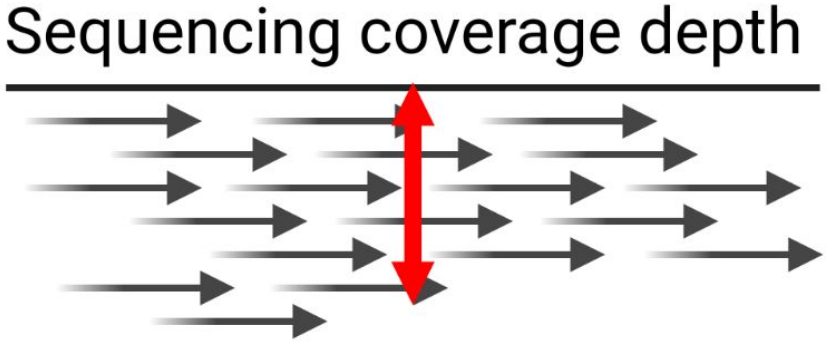


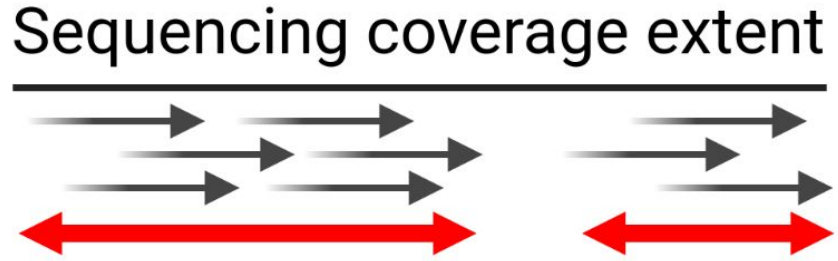


Pseudochromosomes = scaffolds representing an entire chromosome
Gaps = regions between contigs that are represented by Ns

UNIVERSITÄT BONN

Sequencing coverage depth vs. coverage extent





UNIVERSITÄT BONN

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 X L / G

- Coverage (depth) reflects total amount of sequencing data

- Coverage (depth) is very important parameter for sequencing projects

UNIVERSITÄT BONN

Sequencing coverage extent

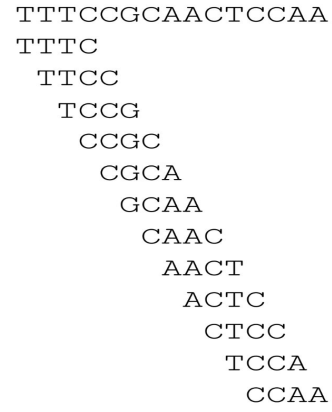
- Coverage extent = ratio of genome covered by at least one base
- Informative to calculate required sequencing depth for a project
- Coverage extent follows a Poisson distribution
- Calculation of coverage extent (c):
  - Non-coverage extent: $P(X=0) = e^{-C}$
  - Probability of coverage extent: $P(X>0) = 1 - e^{-C}$
- 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

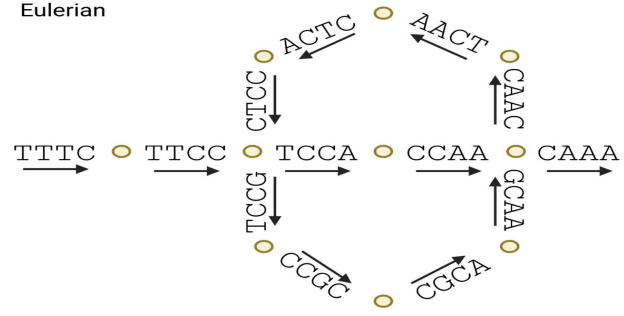
De Brujin graph (DBG)

- 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





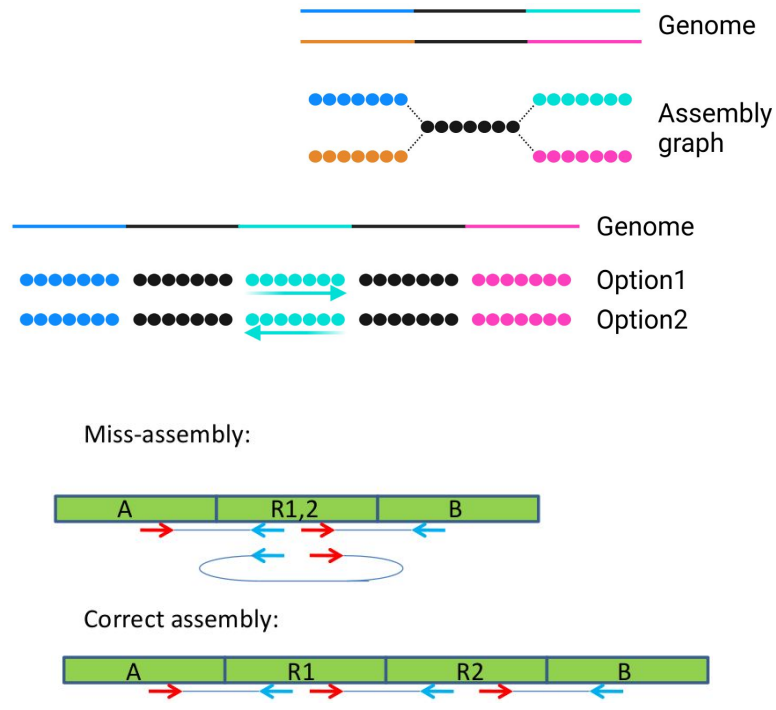
UNIVERSITÄT BONN

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

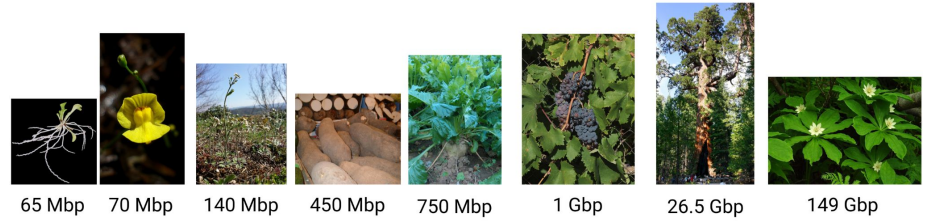
Assembly challenges

- Collapsed repeats
- Inversions
- Overstretched repeats

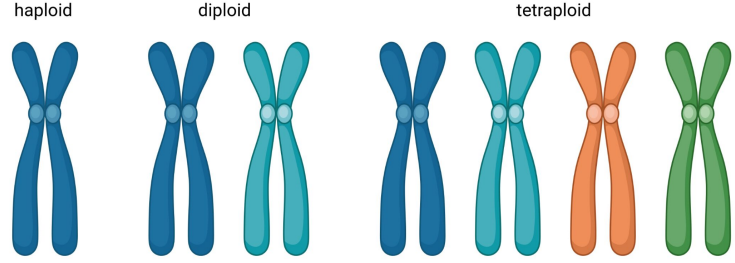


Assembly challenges (2)

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



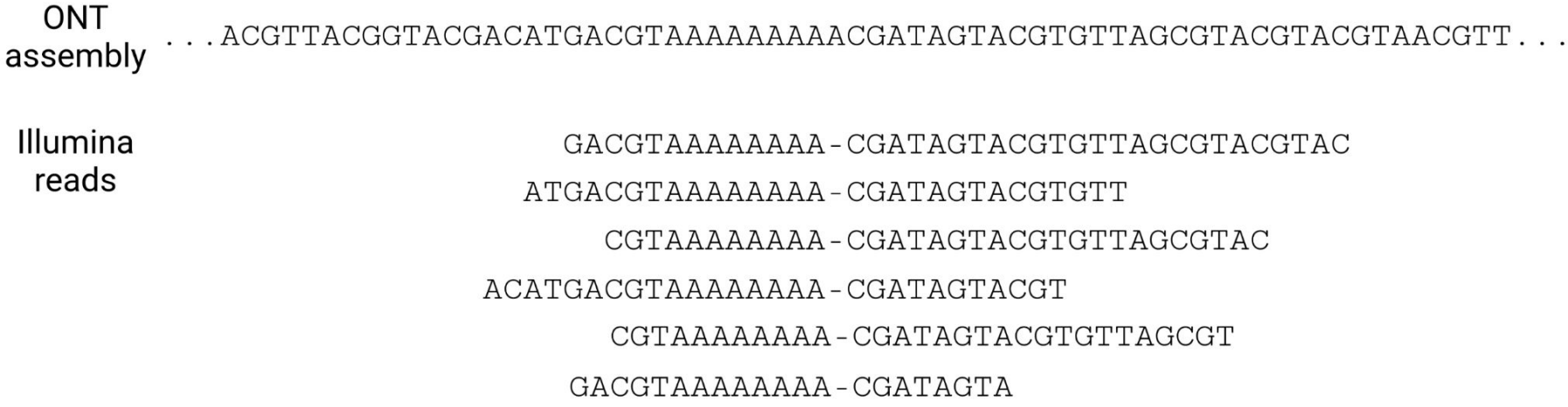
Genlisea violacea: https://commons.wikimedia.org/wiki/File:Genlisea_violacea_giant.jpg
Utricularia gibba: https://commons.wikimedia.org/wiki/File:Utricularia_gibba_flower_01.jpg
Arabidopsis thaliana: https://commons.wikimedia.org/wiki/File:Arabidopsis_thaliana_sl10.jpg
Dioscorea dumetorum: https://commons.wikimedia.org/wiki/File:YamsatBrixtonMarket.jpg
Beta vulgaris: https://de.wikipedia.org/wiki/Zuckerr%C3%BCbe#/media/Datei:Zuckerr%C3%BCbe.jpg
Vitis vinifera: https://commons.wikimedia.org/wiki/File:Vitis_vinifera_20120823.jpg
Sequoia giganteum: https://commons.wikimedia.org/wiki/File:Grizzly_Giant_Mariposa_Grove.jpg
Paris japonica: https://commons.wikimedia.org/wiki/File:Paris_japonica_Kinugasasou_in_Hakusan_2003_7_27.jpg



Kress et al., 2022: 10.1073/pnas.2115640118

UNIVERSITÄT BONN

Assembly polishing (outdated)



UNIVERSITÄT BONN

Assembly evaluation

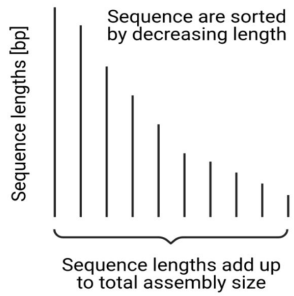
- Continuity: Does the assembly represent a genome in a small number of contigs?
- Completeness: Are all parts of the genome represented?
- Correctness: Is the assembly a correct representation of the genome?

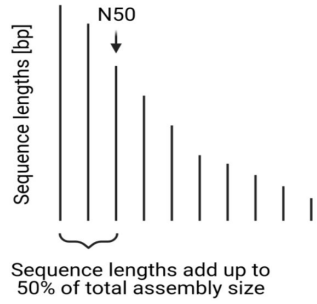
Evaluation: continuity

- 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



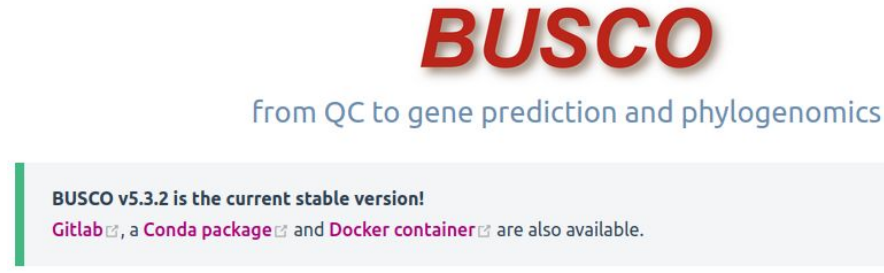


Pucker et al., 2019: 10.1371/journal.pone.0216233

UNIVERSITÄT BONN

Evaluation - completeness

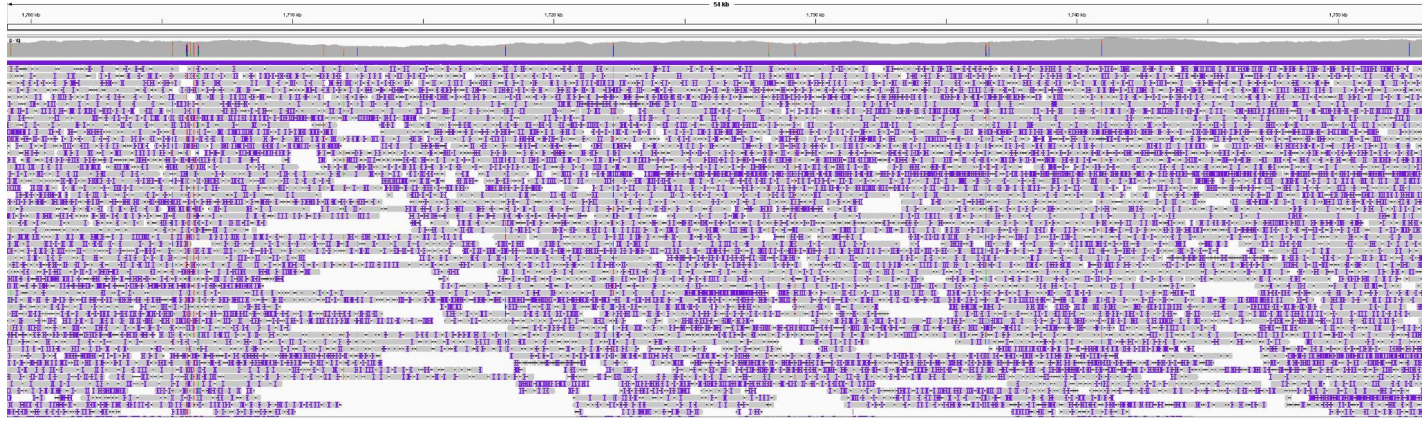
- Check assembly completeness: inspection for presence of conserved genes
- BUSCO = Benchmarking Universal Single-Copy Orthologue
- BUSCO genes are used to assess assembly completeness



UNIVERSITÄT BONN

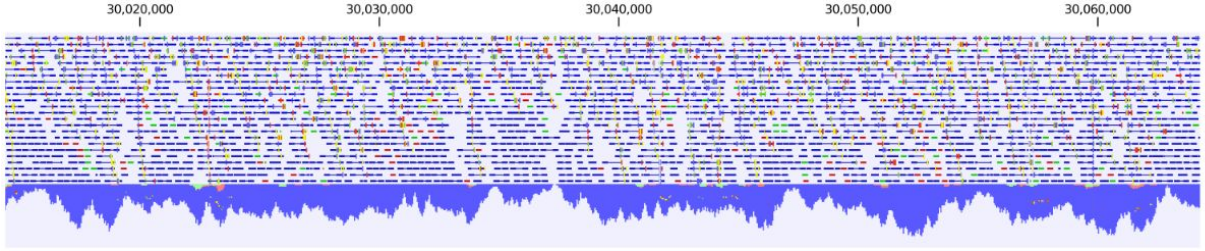
Evaluation - assembly correctness

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



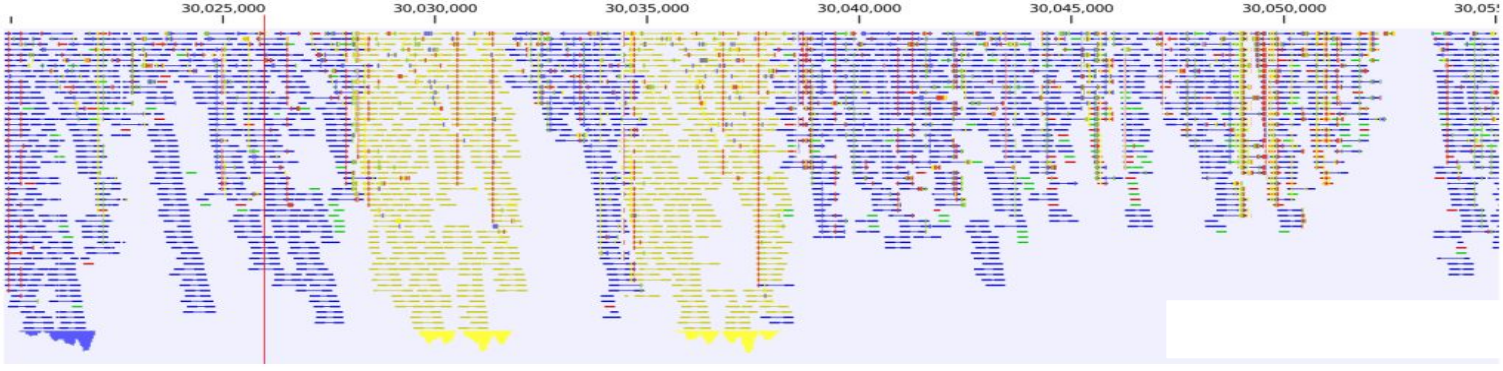
IGV: Thorvaldsdottir et al., 2013: 10.1093/bib/bbs017
REAPR: Hunt et al., 2013: 10.1186/gb-2013-14-5-r47
SQUAT: Yang et al., 2019: 10.1186/s12864-019-5445-3

Read mappings (1)

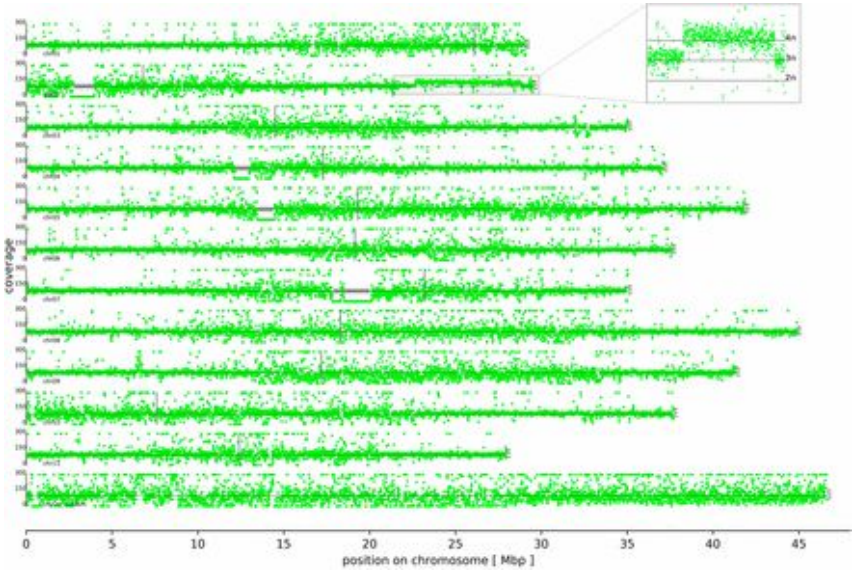


Read mapping of paired-end sequenced fragments (blue) to assembly

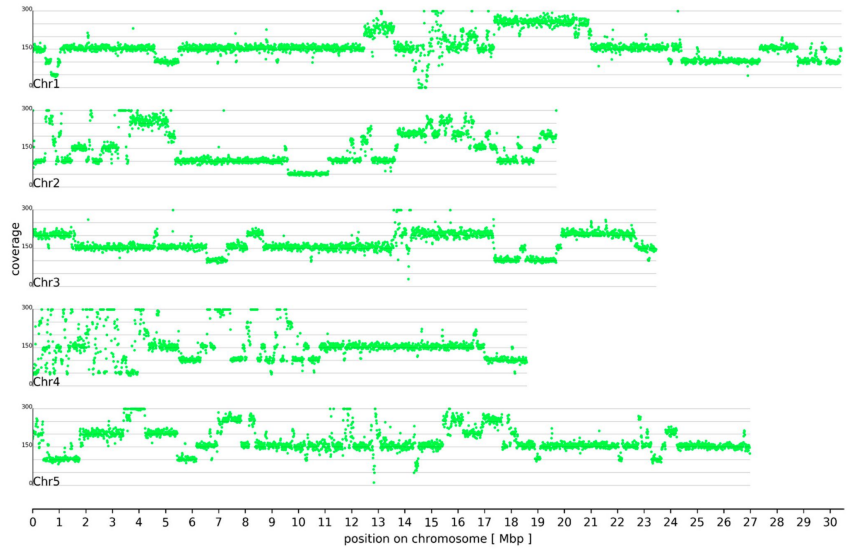
Coverage is too high to show all individual fragments at some positions



Read mappings (2)



Musa acuminata (banana) read mapping



Arabidopsis thaliana At7 read mapping

Pucker et al., 2019: 10.3390/genes10090671
Busche & Pucker et al., 2020: 10.1534/g3.119.400847

UNIVERSITÄT BONN

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

UNIVERSITÄT BONN

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

UNIVERSITÄT BONN

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

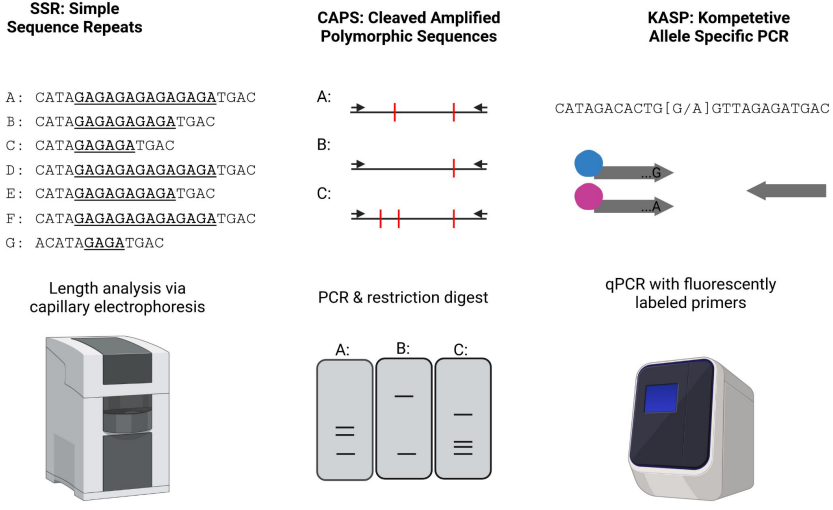
UNIVERSITÄT BONN

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)

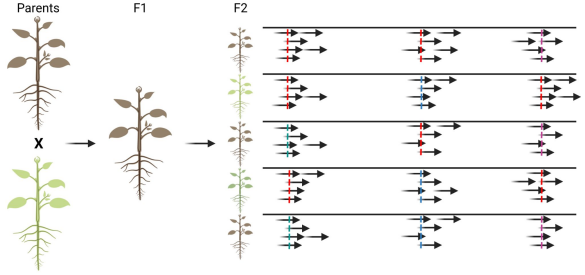
- Classical genetic markers: SSR, CAPS, KASP
  - SSR = Simple Sequence Repeats
  - CAPS = Cleaved Amplified Polymorphic Sequences
  - KASP = Kompetitive Allele Specific PCR



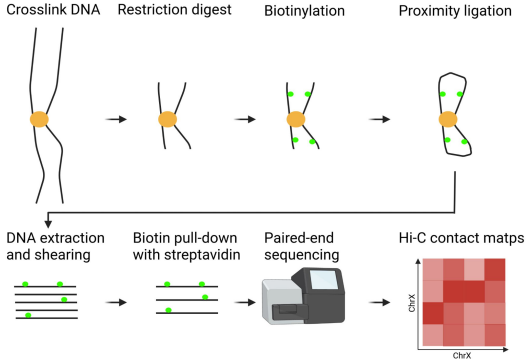
SSR: Holtgräwe et al., 2020: 10.3389/fpls.2020.00156
CAPS: Konieczny & Ausubel, 1993: 10.1046/j.1365-313x.1993.04020403.x
KASP: He et al., 2014: 10.1007/978-1-4939-0446-4_7

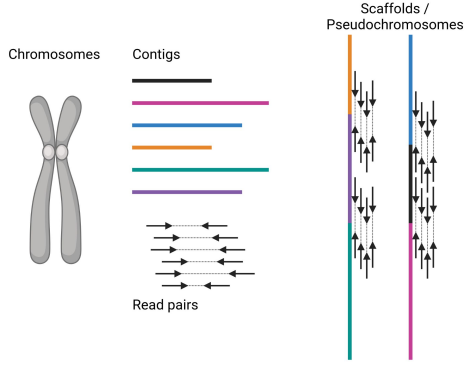
Integration of genetic linkage information (2)

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



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





UNIVERSITÄT BONN

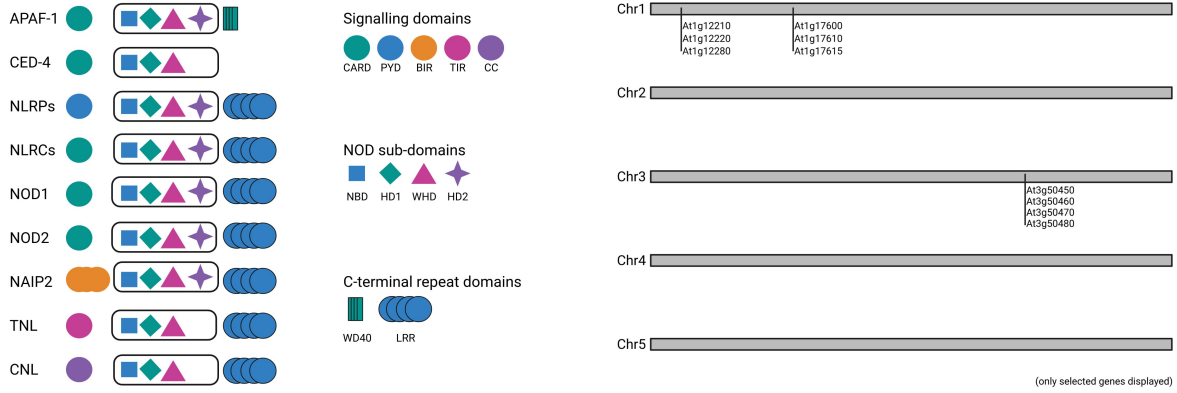
Checking challenging regions

- 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

UNIVERSITÄT BONN

NLRome

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



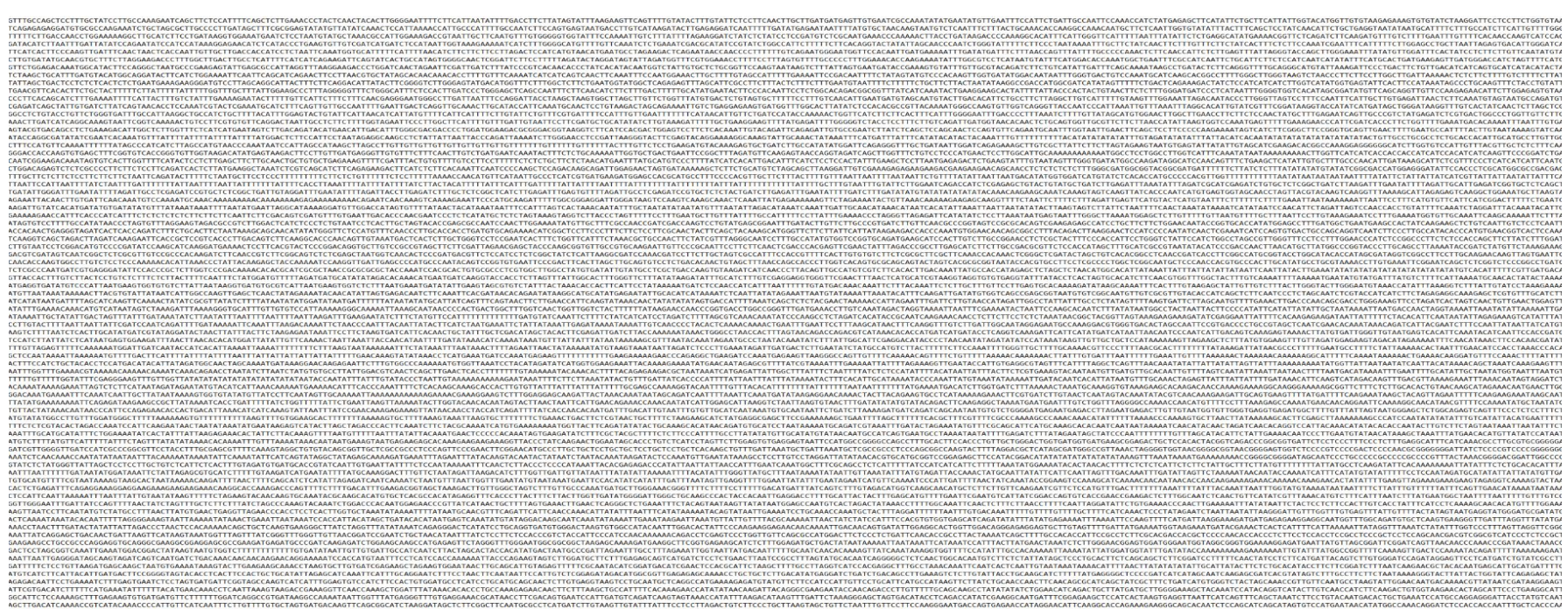
Van de Weyer et al., 2019: 10.1016/j.cell.2019.07.038

UNIVERSITÄT BONN

Structural annotation

Finding genes in a genome sequence

UNIVERSITÄT BONN

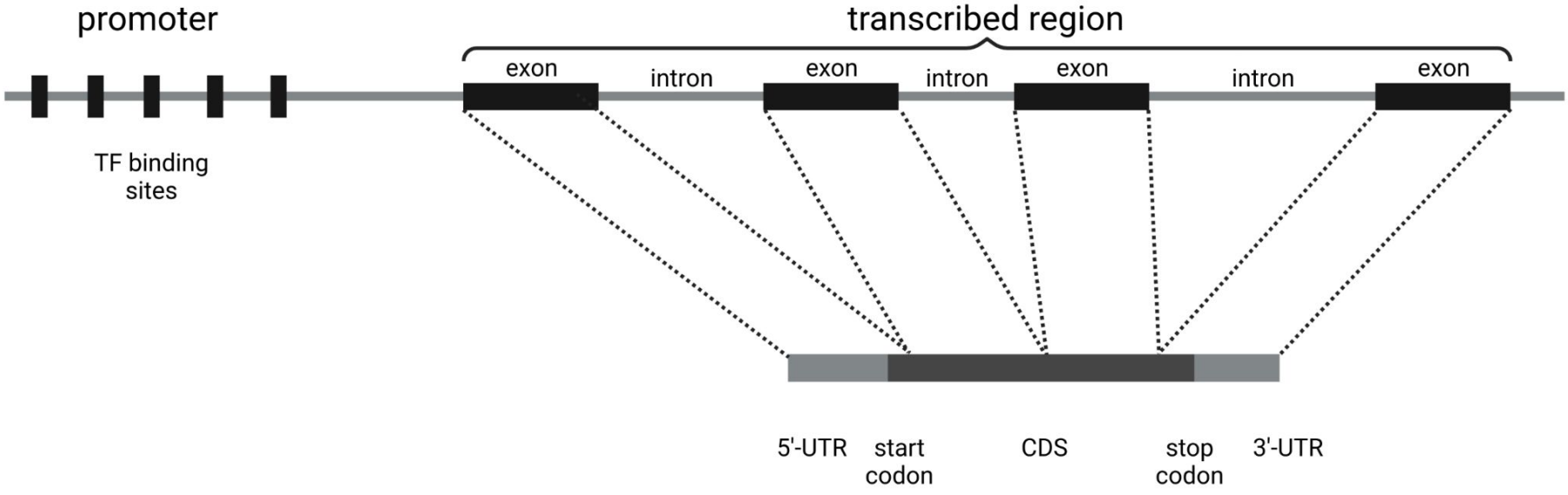


Siadjeu & Pucker et al., 2020: 10.3390/genes11030274

Plant gene structure

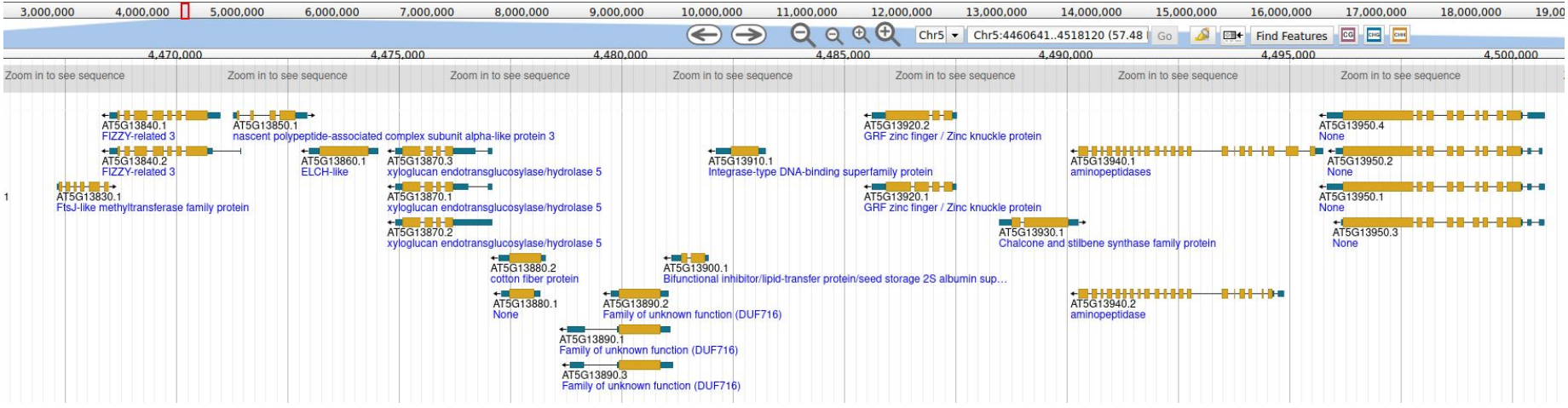
CDS = (Protein) Coding Sequence

ORF = Open Reading Frame

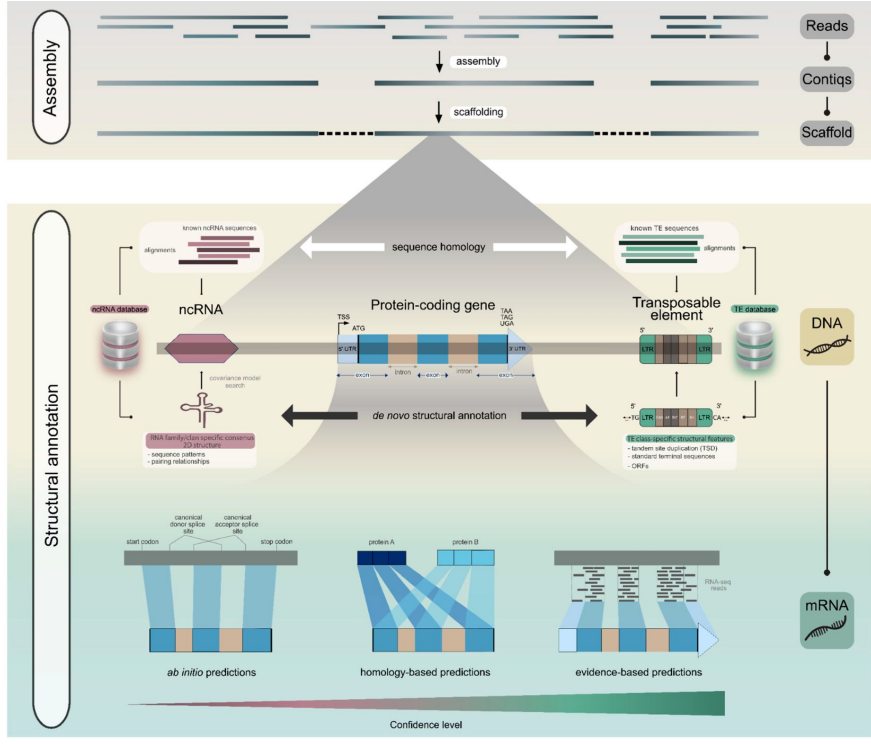


UNIVERSITÄT BONN

Finding genes in a genome sequence



Structural annotation (overview)

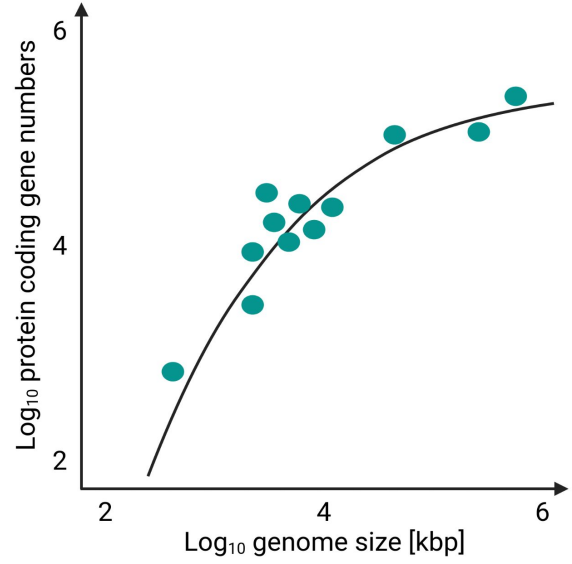


de Oliveira et al., 2025: 10.20944/preprints202508.1176.v2

Gene numbers

- Average number of genes in plants: 27200

- Gene number is not significantly correlated with genome size



Pucker & Brockington, 2019: 10.1186/s12864-018-5360-z
Michael, 2014: 10.1093/bfgp/elu005

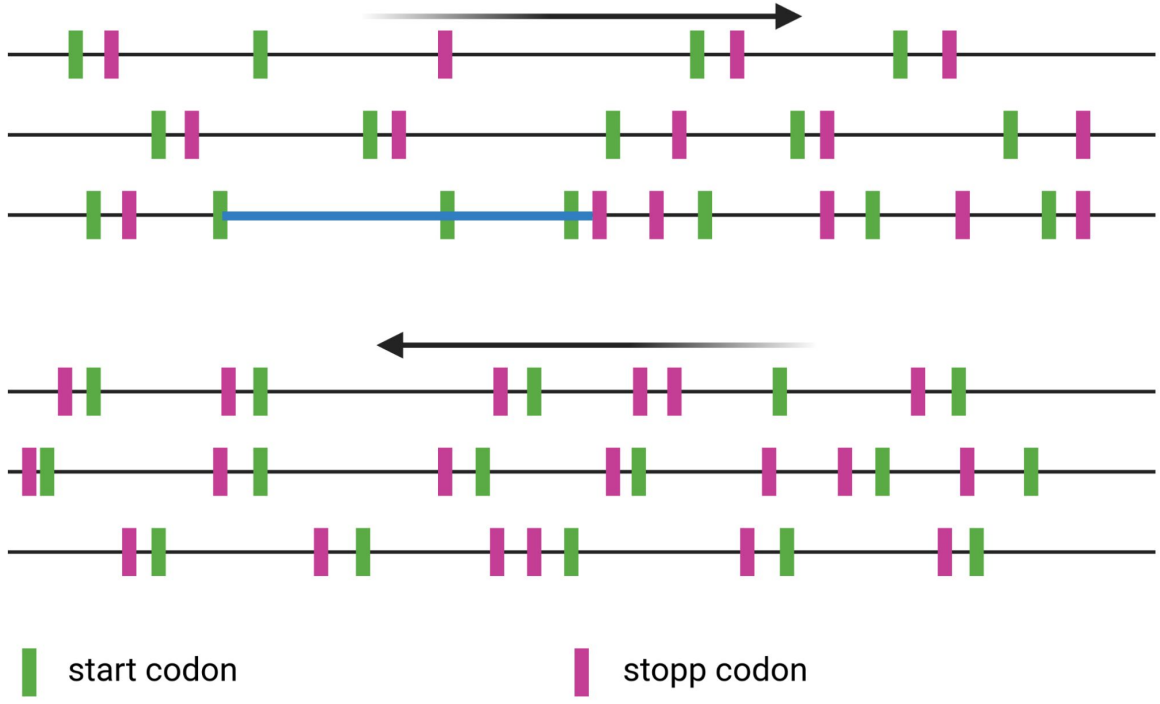
UNIVERSITÄT BONN

Repeat masking

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

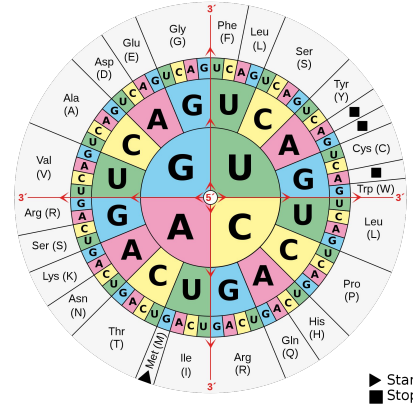
UNIVERSITÄT BONN

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



UNIVERSITÄT BONN

How does a Hidden Markov Model work?

UNIVERSITÄT BONN

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):
  - $(⅙)^{10} = 1.6538 * 10^{-8}$

- Probability using loaded die (L):
  - $^1/_{10} * ^1/_{10} * ½ * ^1/_{10} * ^1/_{10} * ^1/_{10} * ^1/_{10} * ½ * ^1/_{10} * ^1/_{10} = 2.7778 * 10^{-10}$

Hidden Markov Models (3)

- Observation:

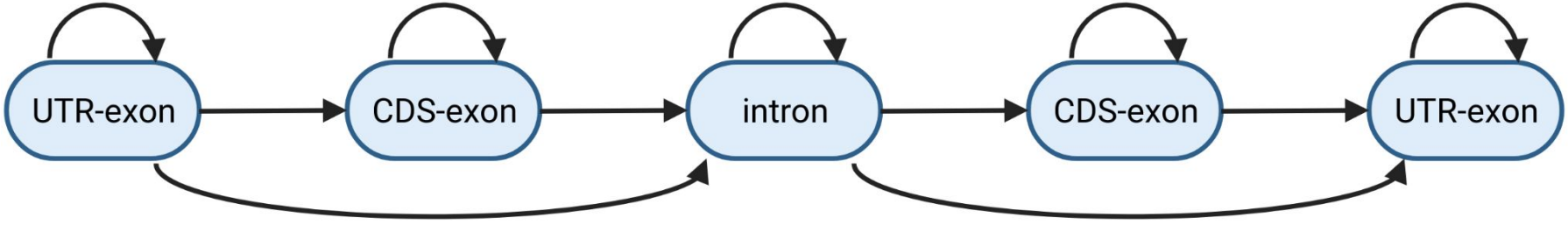
  1625453666166152316342423315266261662656462624351354

  FFFFFFFLLLLLLFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLFFFFFFFF

- 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



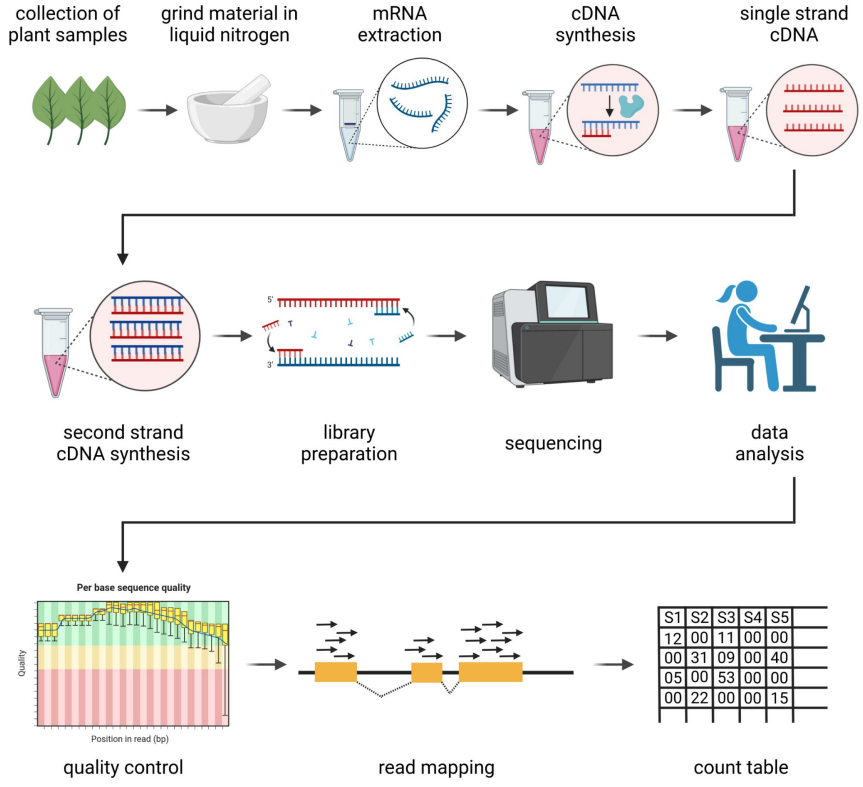
UNIVERSITÄT BONN

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

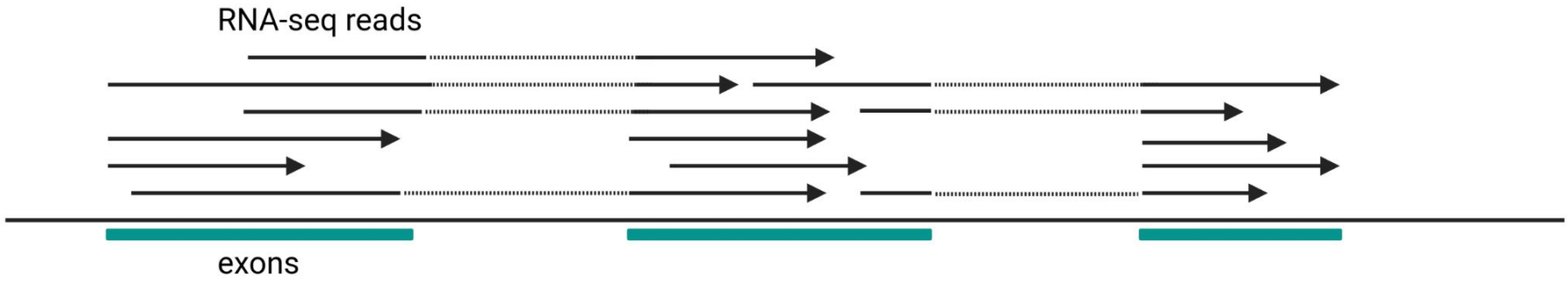
UNIVERSITÄT BONN

RNA-seq



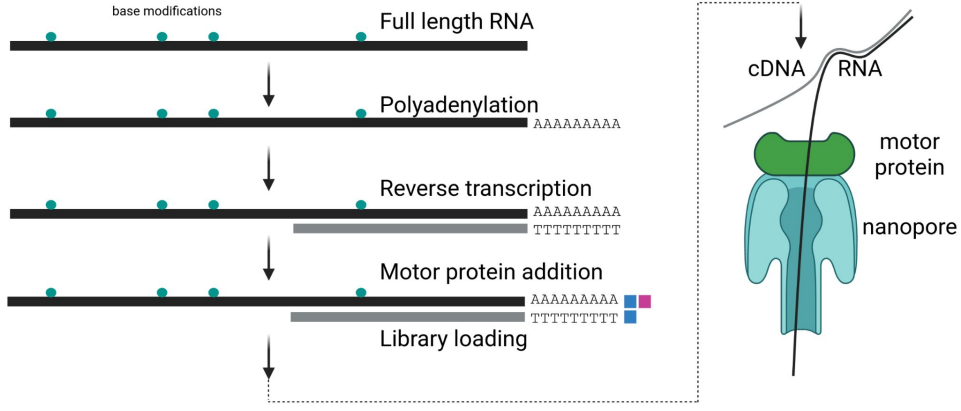
RNA-seq hints

- 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)



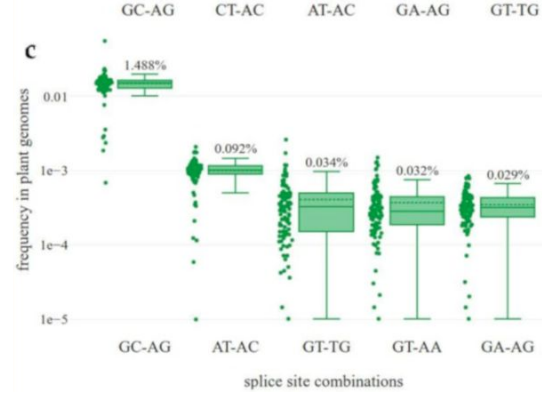
Canonical splice sites

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

UNIVERSITÄT BONN

Non-canonical splice sites

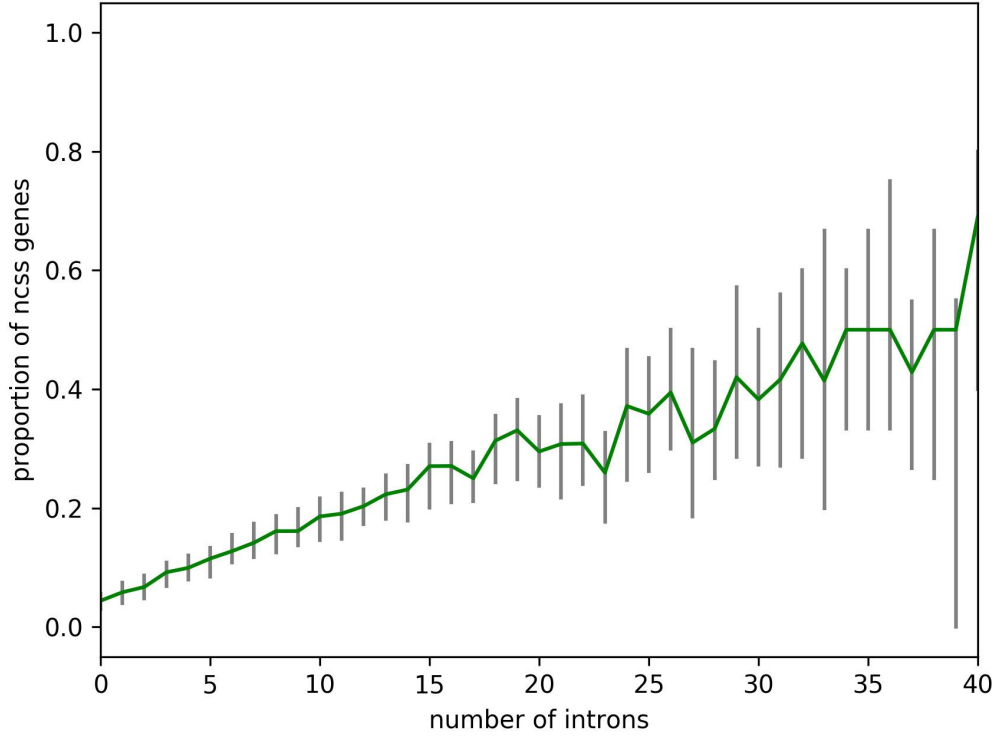
- Splice sites mark points for removal of introns
- Highly conserved to allow recognition by spliceosome
- Two types of spliceosomes: U2 and U12
- Canonical splice sites (98%): GT-AG
- Major non-canonical splice sites (1-2%): GC-AG, AT-AC
- Minor non-canonical splice sites (<1%): NN-NN



Frey & Pucker, 2020: 10.3390/cells9020458

UNIVERSITÄT BONN

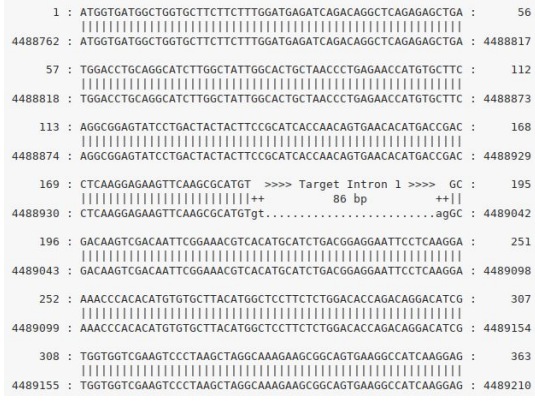
Relevance of non-canonical splice sites

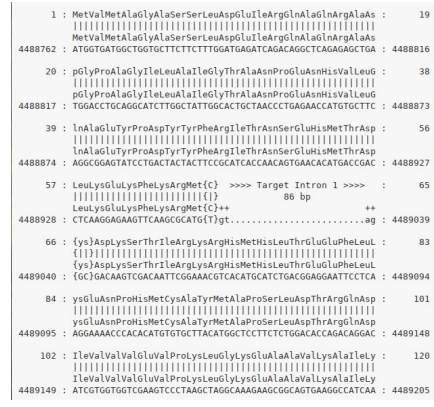


Pucker & Brockington, 2018: 10.1186/s12864-018-5360-z

Proteins as hints (exonerate)

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

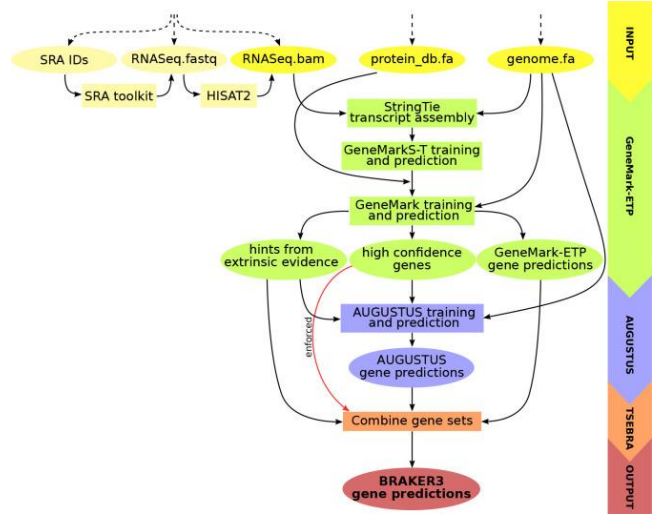




UNIVERSITÄT BONN

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



Gabriel et al., 2024: 10.1101/gr.278090.123

UNIVERSITÄT BONN

GeMoMa

- 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

Keilwagen et al., 2019: 10.1007/978-1-4939-9173-0_9

UNIVERSITÄT BONN

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

- NLRome: 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/)

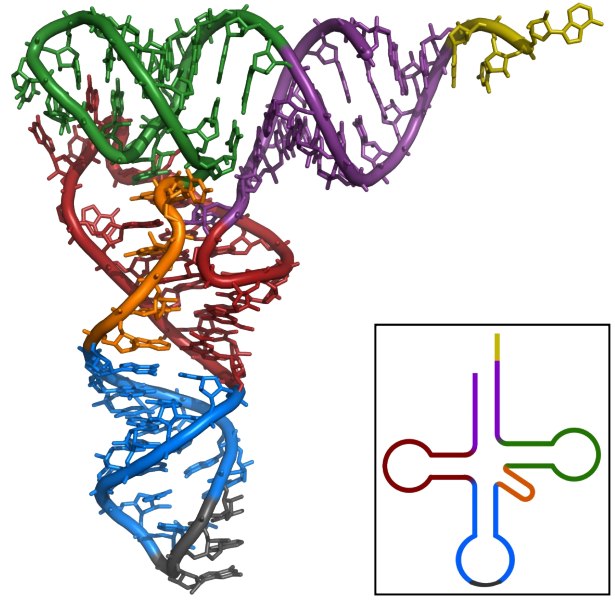
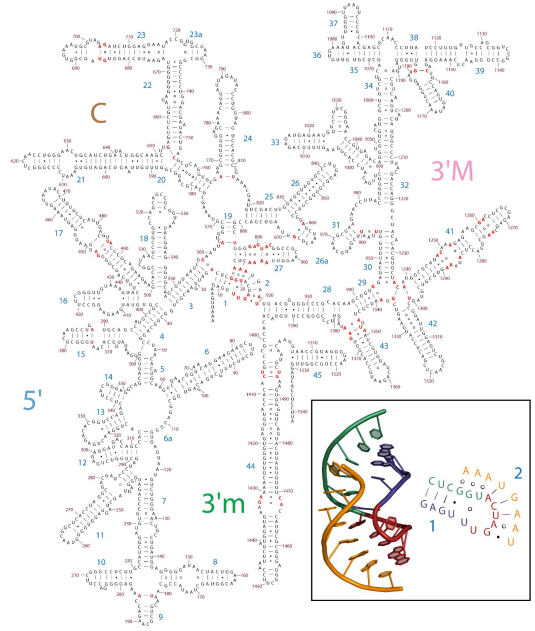


Figure source: https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png

Chan & Lowe, 2019: 10.1093/nar/gkab688

Identification of rRNAs

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

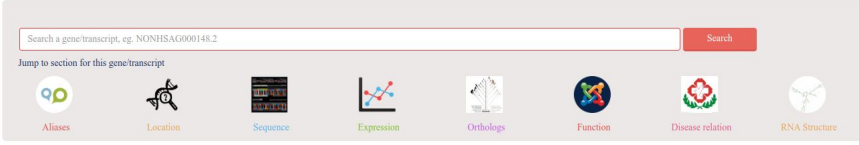


Kalvari et al. , 2017: 10.1093/nar/gkx1038
Lagesen et al., 2007: 10.1093/nar/gkm160
Petrov et al., 2014: 10.1371/journal.pone.0088222

Annotation of long non-coding RNAs (lncRNAs)

- NONCODEV6: long non-coding RNA database (http://www.noncode.org/)

- Identification based on sequence similarity



Zhao et al., 2020: 10.1093/nar/gkaa1046
http://www.noncode.org/

UNIVERSITÄT BONN

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

Wicker et al., 2007: 10.1038/nrg2165

UNIVERSITÄT BONN

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

Ou et al., 2019: 10.1186/s13059-019-1905-y
Flynn et al., 2020: 10.1073/pnas.1921046117
Tarailo-Graovac & Chen, 2009: 10.1002/0471250953.bi0410s25

UNIVERSITÄT BONN

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/)

Van Bel et al., 2022: 10.1093/nar/gkab1024

UNIVERSITÄT BONN

Functional annotation

UNIVERSITÄT BONN

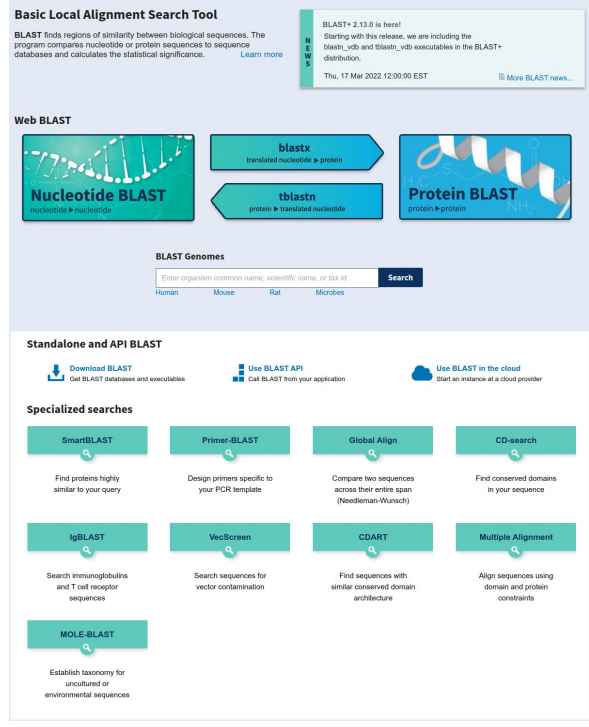
Functional annotation

- 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



https://blast.ncbi.nlm.nih.gov/Blast.cgi
Altschul et al., 1990: https://doi.org/10.1016/s0022-2836(05)80360-2

Pfam: Protein family database

- Assignment of protein functions based on Hidden Markov Models (HMMs)

- Sequences are screened based on HMM profile

Sequences are aligned

↓

Modelling of insertions/deletions

↓

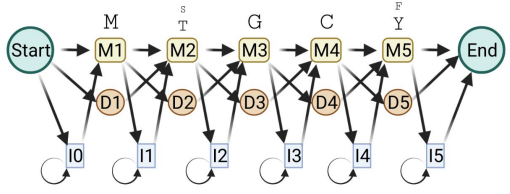
Amino acid frequency at each alignment position is encoded

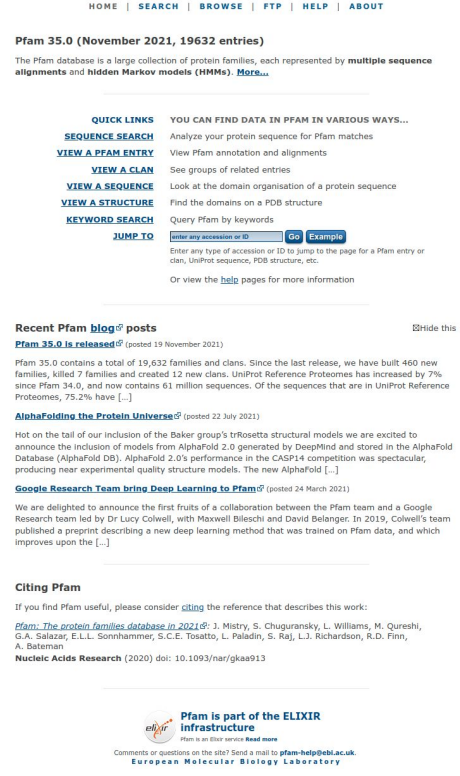
↓

HMM profile generated

```
seq1    MTGC-Y
seq2    MSGC-F
seq3    MTGC-Y
seq4    M-GCAY
        123456
```

2 = deletion
5 = insertion



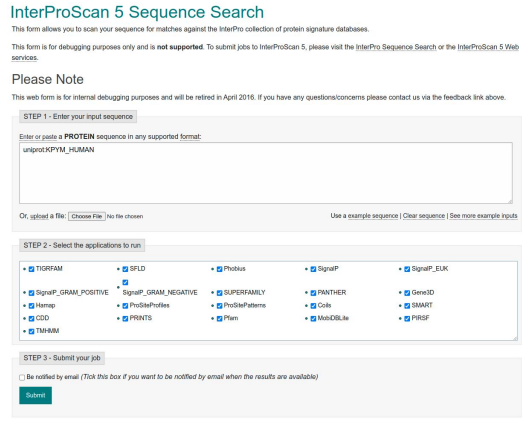


http://pfam.xfam.org/
Mistry et al., 2020: 10.1093/nar/gkaa913

UNIVERSITÄT BONN

InterProScan5

- 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



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   ba46b0f06bfbf1c91d161f0f91310f43   395   Gene3D   G3DSA:3.40.47.10          -   1     241   4.8E-101      T   20-03-2022
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   SUPERFAMILY   SSF53901          Thiolase-like   241   393   2.98E-51      T   20-03-20
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   SUPERFAMILY   SSF53901          Thiolase-like   10    237   2.38E-78      T   20-03-20
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   Pfam   PF02797 Chalcone and stilbene synthases, C-terminal domain   244   394   1.5E-71
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   PANTHER   PTHR11877:SF81   BNAA02G30320D PROTEIN   6   394   0.0   T   20-03-2022
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   Gene3D   G3DSA:3.40.47.10          -   242   395   9.7E-62   T   20-03-2022   IPR01603
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   PANTHER   PTHR11877   HYDROXYMETHYLGLUTARYL-COA SYNTHASE   6   394   0.0   T
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   CDD   cd00831 CHS_like   21   390   0.0   T   20-03-2022   -   -
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   PIRSF   PIRSF000451   PKS_III 7   394   0.0   T   20-03-2022   IPR011141
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   Pfam   PF00195 Chalcone and stilbene synthases, N-terminal domain   10   233   2.9E-119
sp|P13114|CHSY_ARATH   ba46b0f06bfbf1c91d161f0f91310f43   395   ProSitePatterns   PS00441 Chalcone and stilbene synthases active site.   161   177   -
```


http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5&sequence=uniprot:KPYM_HUMAN
Jones et al., 2014: 10.1093/bioinformatics/btu031

UNIVERSITÄT BONN

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/davidemms/OrthoFinder
   - Tutorials: https://davidemms.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

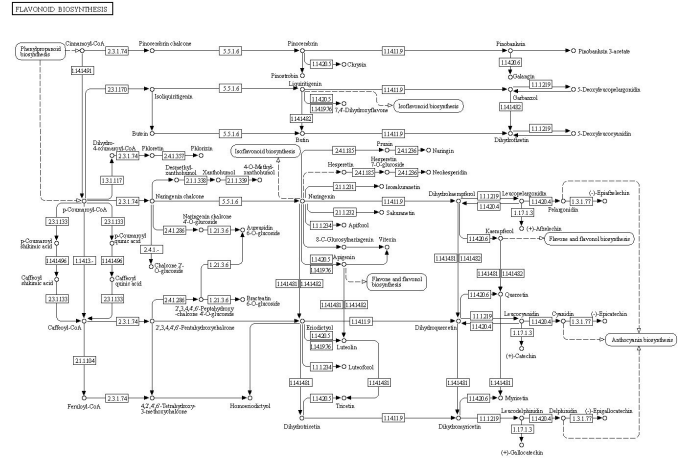
```
python shoot INPUT_FASTA SHOOT_DB
```

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.

https://github.com/davidemms/SHOOT

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

Kanehisa & Goto, 2000: https://doi.org/10.1093/nar/28.1.27
Kanehisa et al., 2022: https://doi.org/10.1002/pro.4172
https://www.genome.jp/kegg/

UNIVERSITÄT BONN

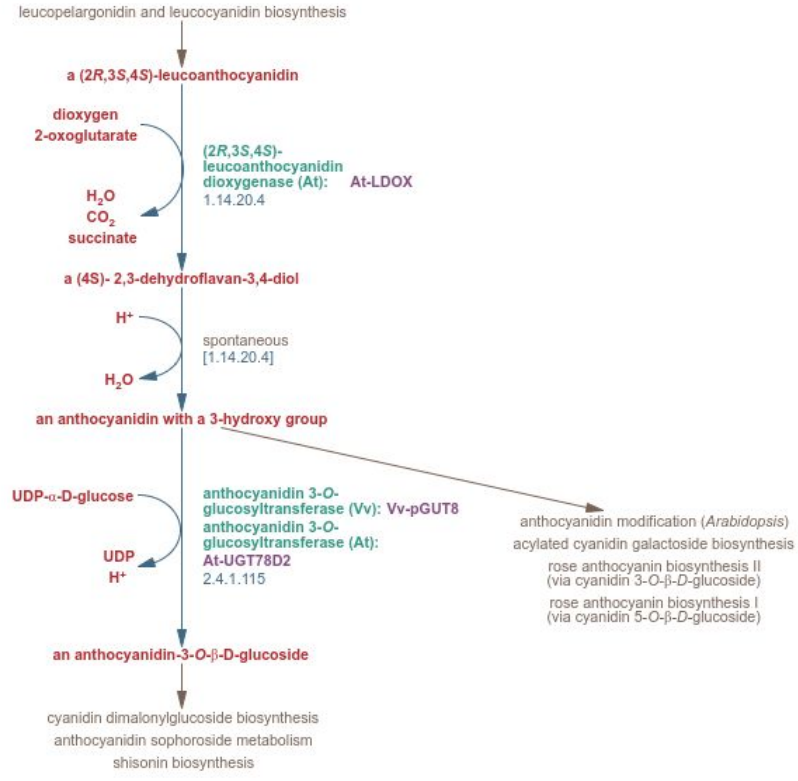
Gene Ontology (GO)

- 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)

http://amigo.geneontology.org/amigo/term/GO:0009813
The Gene Ontology Consortium, 2007: 10.1093/nar/gkm883

MetaCyc: Metabolite Encyclopedia

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

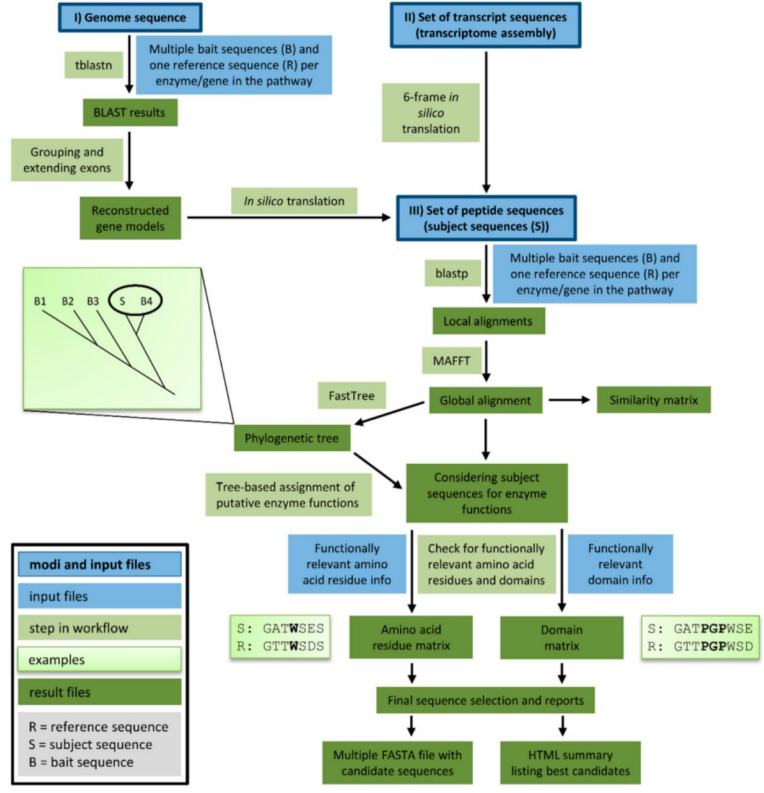


https://metacyc.org/
Karp et al., 2015: https://doi.org/10.48550/arXiv.1510.03964
Figure: https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5125

UNIVERSITÄT BONN

KIPEs

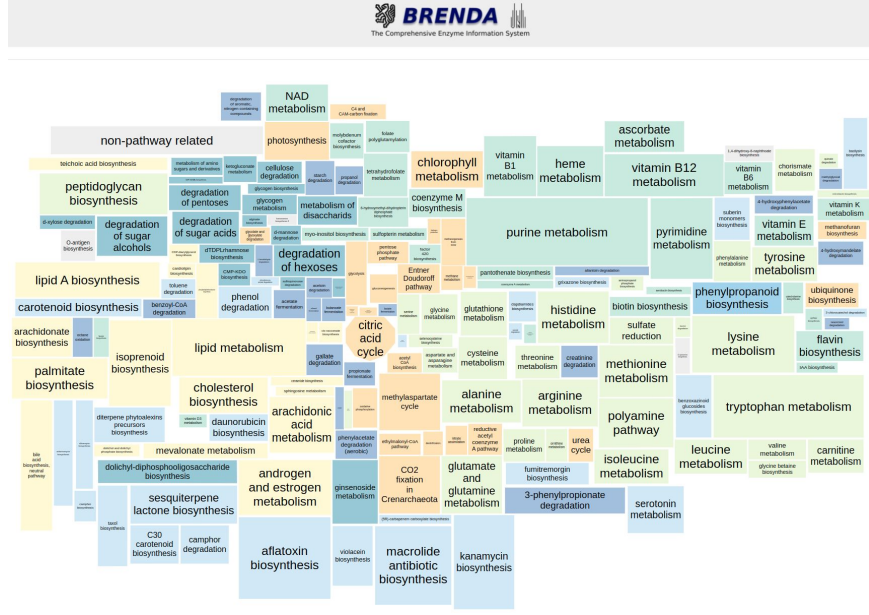
- Identification of all players in a biosynthesis pathway
- Dedicated to the identification of enzymes
- Functionally relevant amino acids need to be known
- Web server: http://www.pbb-tools.de



https://github.com/bpucker/KIPEs
Pucker et al., 2020: 10.3390/plants9091103

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, …)



www.brenda-enzymes.org
Chang et al., 2021: 10.1093/nar/gkaa1025

UNIVERSITÄT BONN

Mercator - functional annotation of protein sequences

- 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
MIGPMPGMEGKLMLPGASPVGLEVVLSLVVASTPILSDTSLCTPSFYHFLLLLGPITSISNLIVRPFFLTLLSSITVIYGRLCFFAFDFYVY
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRIKRKGRPPKKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHIPPTSIPSFNNPNNIHQSSSDRQMPP
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTAATGDKKKRTKARKETYSSYIYKVLKQVHPDTGISNRAMSILNSFVNDIFERVATEASKLA
>TRINITY_DN10001_c0_g1_i2
MAKVGNPIVDIETDGSVNEPESSEKNIEVSSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREHNIDINNIKGTGKNGRITKEDILNYV
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFFKPGWKEFVRSNDLEGGDFLVFNLVDKISYQVVIFDGTCCPKDLCPFSIMNPIFQHLRNKIFLSKKEEIKLKGNRKVHSVNEN
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFVAFVTAGYPDSEETVDILLGLEAGGADIIELGIPFTDPMVDGKTIQDANNVALENKIDISKCLSYVSESRAK
```


https://plabipd.de/portal/mercator-sequence-annotation
Lohse et al., 2014: 10.1111/pce.12231
Haak et al., 2018: 10.3389/fmolb.2018.00062

UNIVERSITÄT BONN

Summary

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

UNIVERSITÄT BONN

Time for questions!

UNIVERSITÄT BONN

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.

- 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.

- 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.