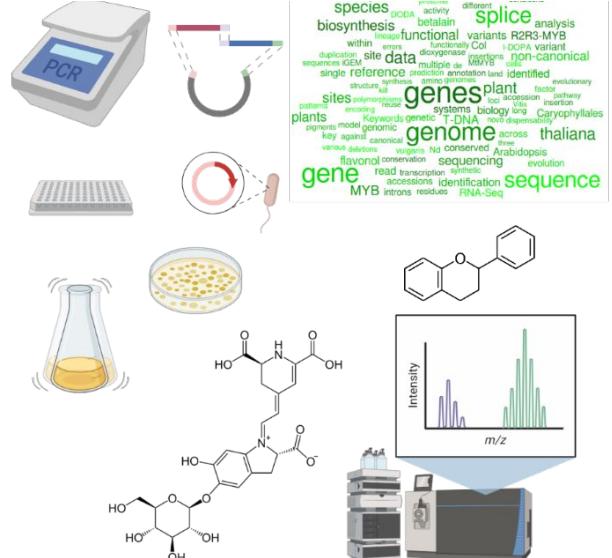
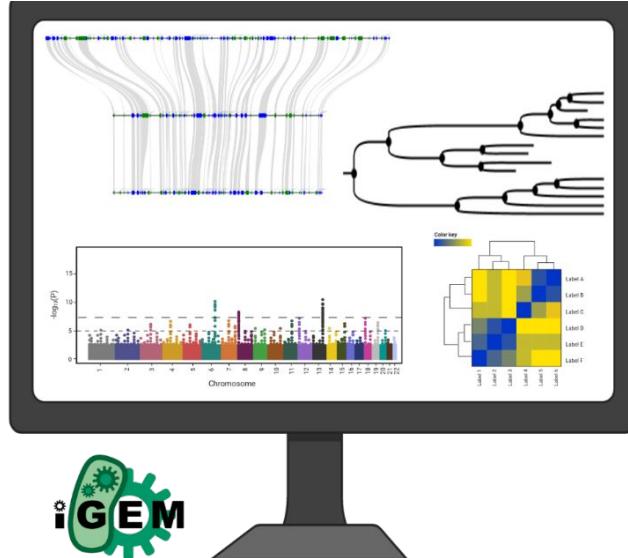
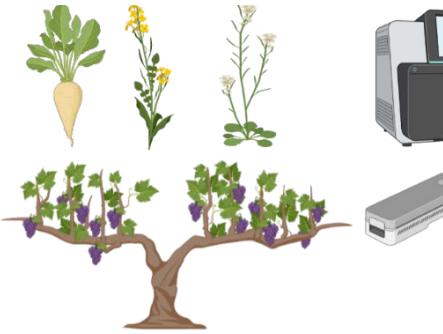




Technische  
Universität  
Braunschweig



# Genome Sequence Annotation

Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

# Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **GE31/MM12**
  - GitHub: <https://github.com/bpucker/teaching>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: [b.pucker\[a\]tu-braunschweig.de](mailto:b.pucker[a]tu-braunschweig.de)



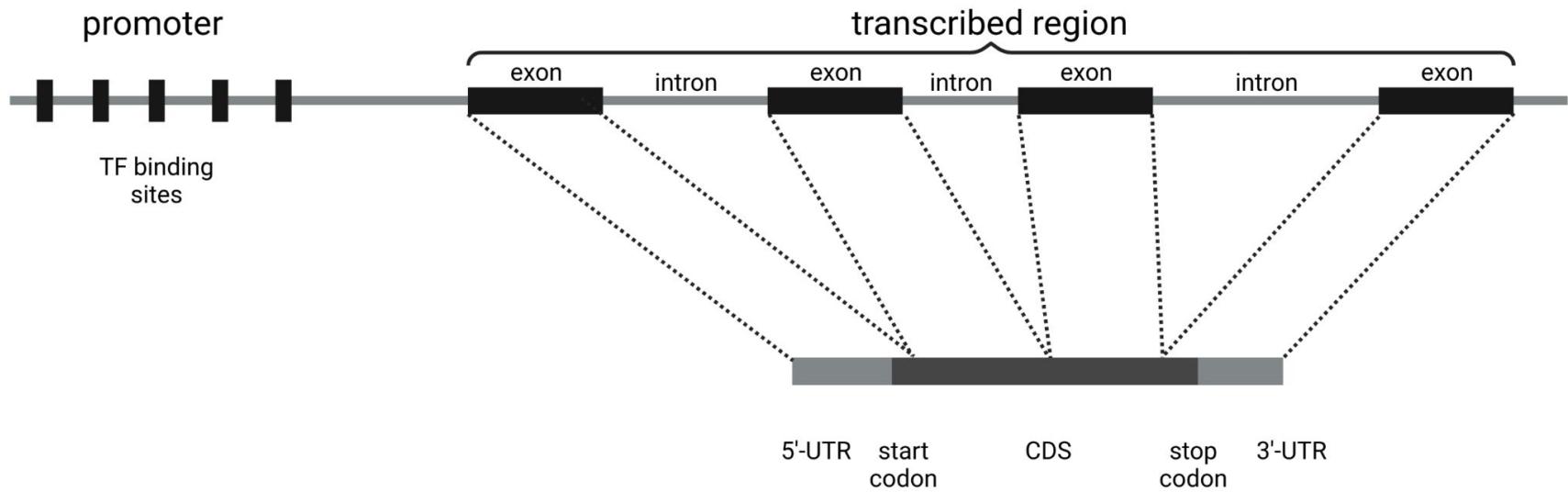
My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

# Finding genes in a genome sequence

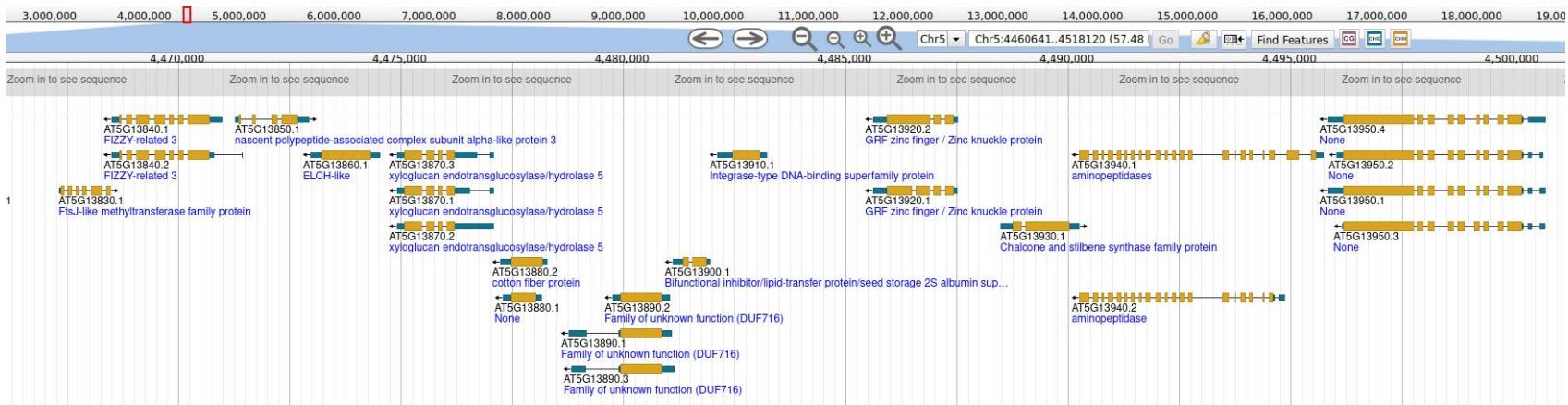
# Plant gene structure

CDS = (Protein) Coding Sequence

ORF = Open Reading Frame

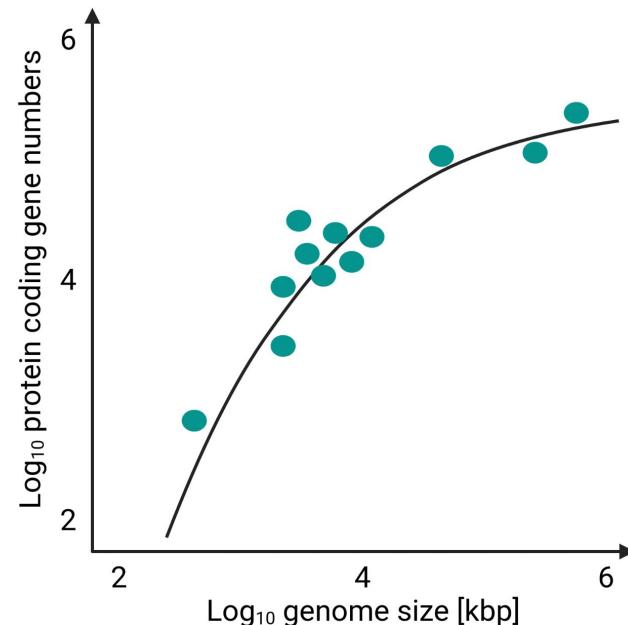


# Finding genes in a genome sequence



# Gene numbers

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

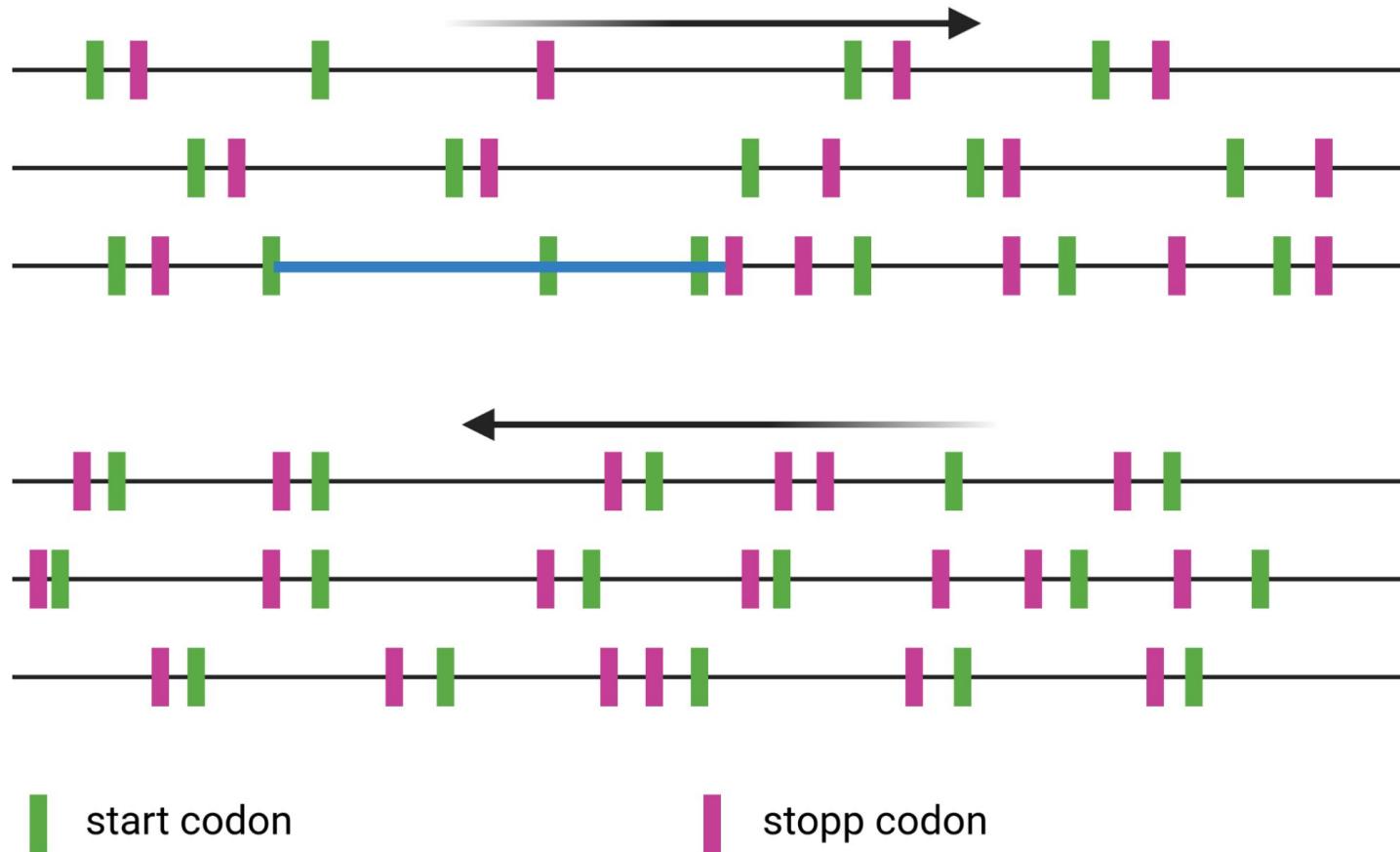


Pucker & Brockington, 2019: 10.1186/s12864-018-5360-z  
Michael, 2014: 10.1093/bfgp/elu005

# Repeat masking

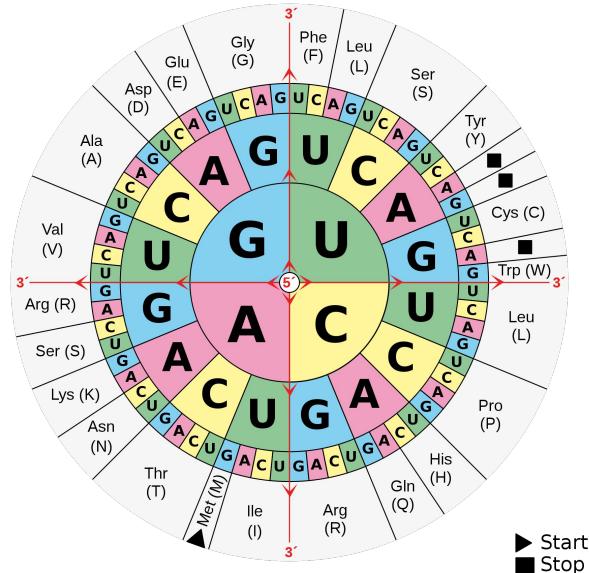
- Simple sequence repeats (SSR)
- Transposable elements (TE)
- Centromeric repeats (CEN)
- Telomeric repeats (TEL)

# 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):
  - $\frac{1}{10} * \frac{1}{10} * \frac{1}{2} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{2} * \frac{1}{10} * \frac{1}{10} = 2.7778 * 10^{-10}$



# Hidden Markov Models (3)

- Observation:

1625453666166152316342423315266261662656462624351354

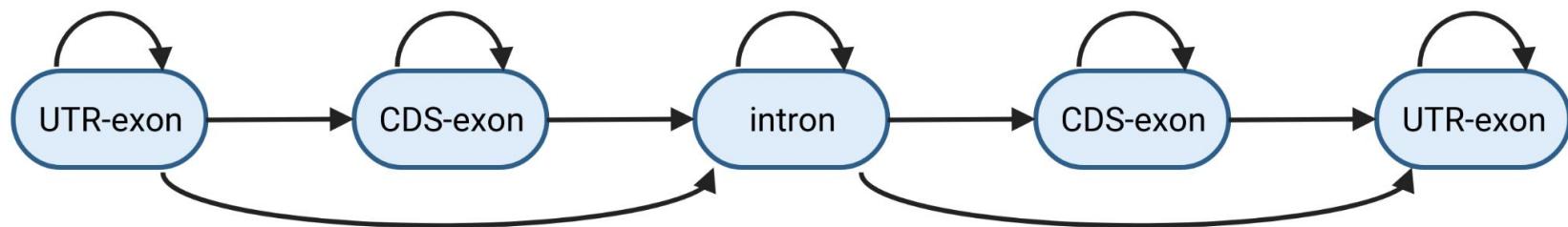
FFFFFFFFFFLLLLLFFFFFFFLLLLLFFFFFFFLLLLLFFFFFFF



- How does this fit to gene prediction?

# Hidden Markov Model for gene prediction

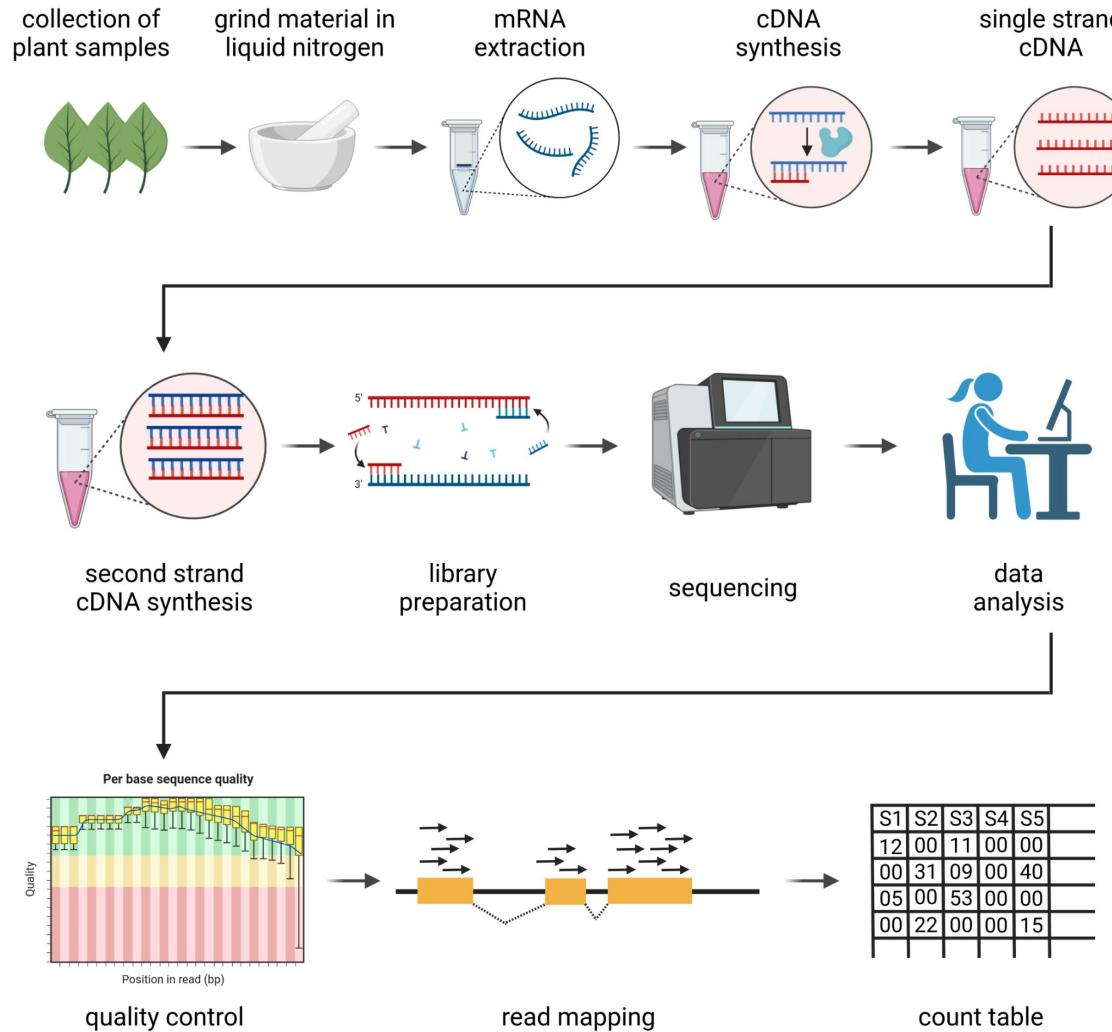
- Fair die = exon (GC-rich in plants)
- Loaded die = intron (AT-rich in plants)
- Switch between fair die and loaded die at intron/exon borders



# Parameters for *ab initio* gene prediction

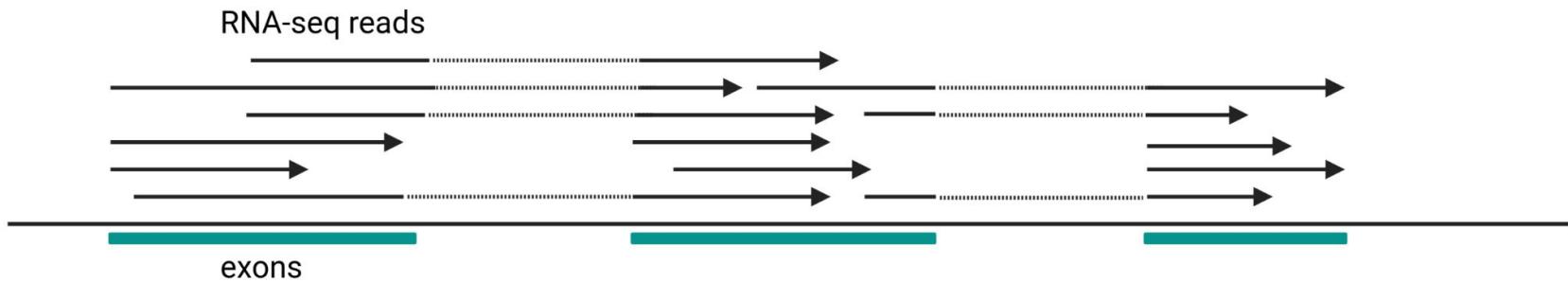
- Features: utr, exon, intron, intergenic region
- Composition of different features (codons)
- feature lengths
- Number of features per gene
- Possible splice sites

# RNA-seq



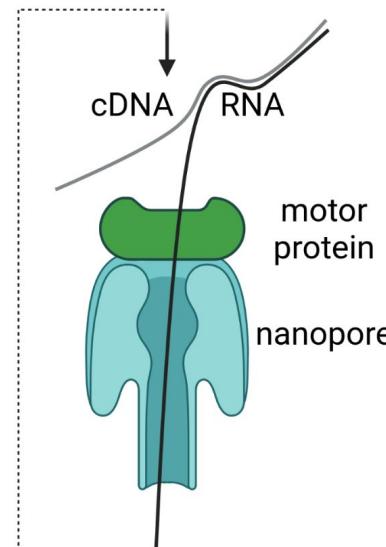
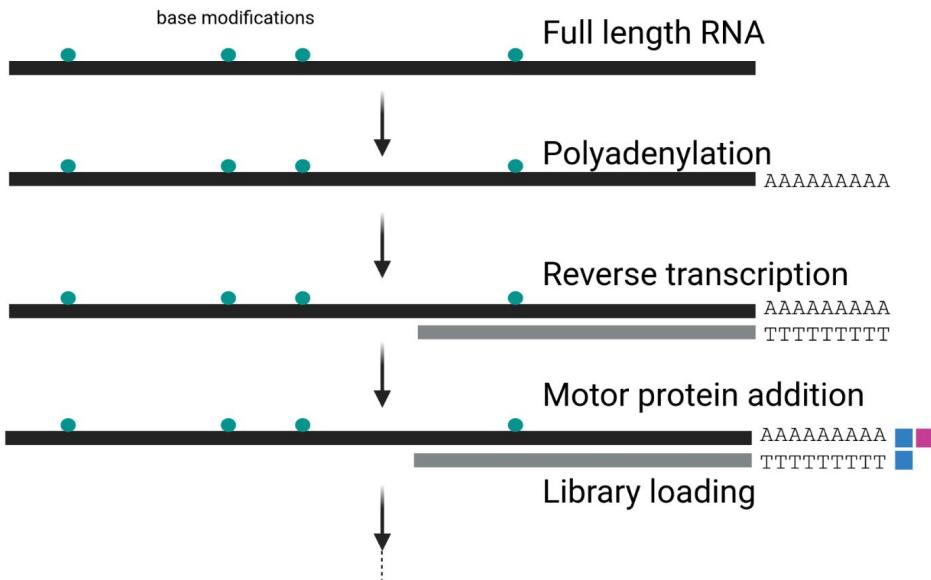
# RNA-seq hints

- 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)



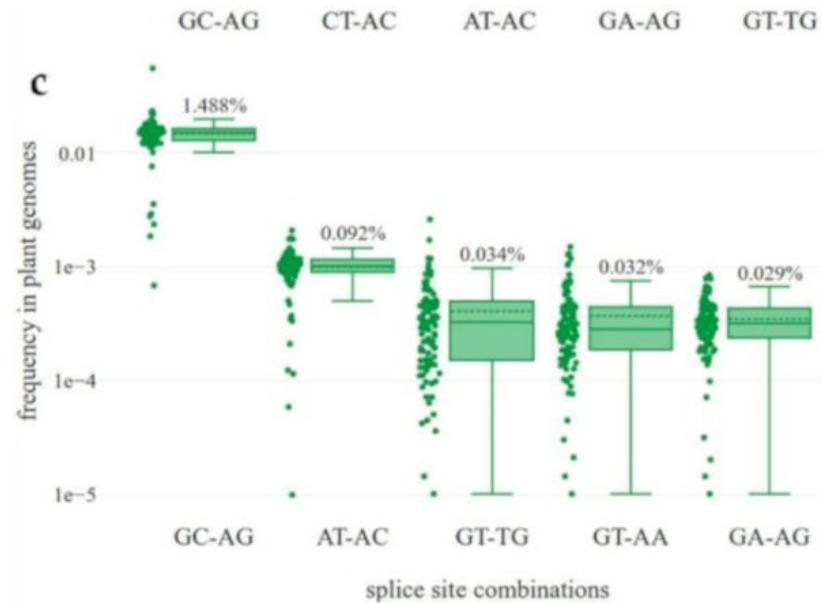
# Canonical splice sites

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

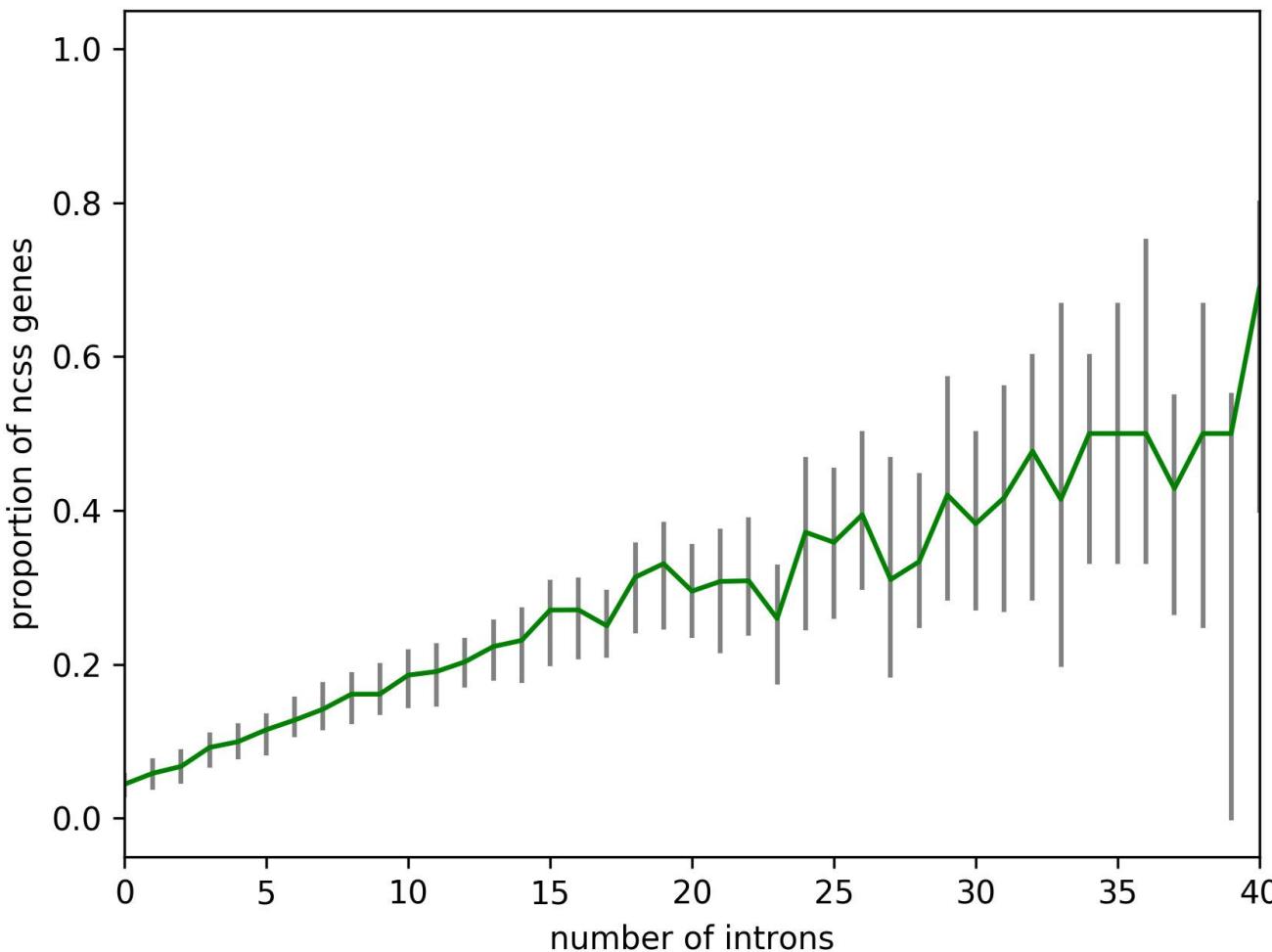


# Non-canonical splice sites

- 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 (exonrate)

- 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 : ATGGTGTGCTGGTGCCTTCTGGATGAGATCAGACAGGCTCAGAGAGCTGA : 56         4488762 : ATGGTGTGCTGGTGCCTTCTGGATGAGATCAGACAGGCTCAGAGAGCTGA : 4488817  57 : TGGACCTGCAGGCATCTGGCTATTGGCACTGCTAACCTGAGAACCATGTGCTTC : 112         4488818 : TGGACCTGCAGGCATCTGGCTATTGGCACTGCTAACCTGAGAACCATGTGCTTC : 4488873  113 : AGGGGGAGTACTCTGACTACTCTCCGCATACCAACAGTGAACACATGACCAC : 168         4488874 : AGGGGGAGTACTCTGACTACTCTCCGCATACCAACAGTGAACACATGACCAC : 4488929  169 : CTCAAGGAGAAAGTTCAAGCGCATGT >>> Target Intron 1 >>> GC : 195        +   +    86 bp +    4488930 : CTCAAGGAGAAAGTTCAAGCGCATGTgt.....agGC : 4489042  196 : GACAAGTCGACAATTGGAAACGTACATGCTACATCTGACGGAGGAATTCCCTCAAGGA : 251         4489043 : GACAAGTCGACAATTGGAAACGTACATGCTACATCTGACGGAGGAATTCCCTCAAGGA : 4489098  252 : AAACCCCACACATGTGCTTACATGGCTCTCTGGACACCCAGACAGGACATCG : 307         4489099 : AAACCCCACACATGTGCTTACATGGCTCTCTGGACACCCAGACAGGACATCG : 4489154  308 : TGGTGGTCAAGTCCCTAACGCTAGGCAAAGAAGCGGCAGTGAAGGCCATCAAGGAG : 363         4489155 : TGGTGGTCAAGTCCCTAACGCTAGGCAAAGAAGCGGCAGTGAAGGCCATCAAGGAG : 4489210	1 : MetValMetAlaGlyAlaSerSerLeuAspGluIleArgGlnAlaGlnArgAlaAs : 19         4488762 : MetValMetAlaGlyAlaSerSerLeuAspGluIleArgGlnAlaGlnArgAlaAs : 4488816  20 : pGlyProAlaGlyIleLeuAlaIleGlyThrAlaAsnProGluAsnHisValLeuG : 38         4488817 : pGlyProAlaGlyIleLeuAlaIleGlyThrAlaAsnProGluAsnHisValLeuG : 4488873  39 : lnAlaGluTyrProAspTyrPheArgIleThrAsnSerGluHisMetThrAsp : 56         4488874 : lnAlaGluTyrProAspTyrPheArgIleThrAsnSerGluHisMetThrAsp : 4488927  57 : LeuLysGluLysPheLysArgMet(C) >>> Target Intron 1 >>> : 65        +   +    86 bp +    4488928 : LeuLysGluLysPheLysArgMet(C)++...ag : 4489039  66 : {ys}AspLysSerThrIleArgLysArgHisLeuThrGluGluPheLeuL : 83 {   }   {   }{   }{   } 4489040 : {ys}AspLysSerThrIleArgLysArgHisLeuThrGluGluPheLeuL : 4489094  84 : ysGluAsnProHisMetCysAlaTyrMetAlaProSerLeuAspThrArgGlnAsp : 101         4489095 : ysGluAsnProHisMetCysAlaTyrMetAlaProSerLeuAspThrArgGlnAsp : 4489148  102 : IleValValValGluValProLysLeuGlyLysGluAlaIvaValLysAlaIleLy : 120         4489149 : IleValValValGluValProLysLeuGlyLysGluAlaIlaValLysAlaIleLy : 4489205 ATCGTGGTGTGAAAGTCCCTAACGCTAGGCAAAGAACGGCAGTGAAGGCCATCAA : 4489205
--	--



# 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/>)

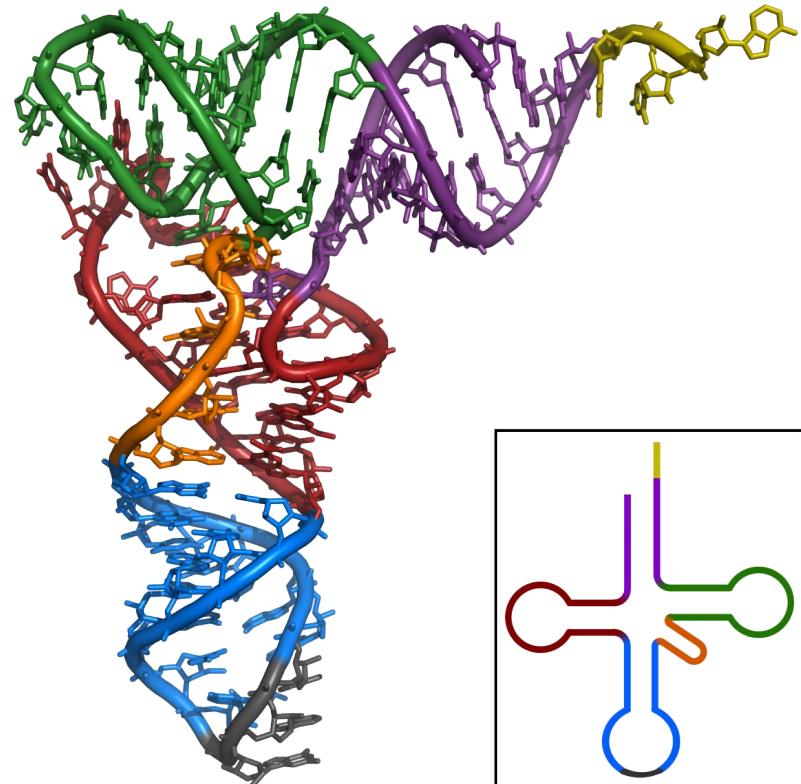
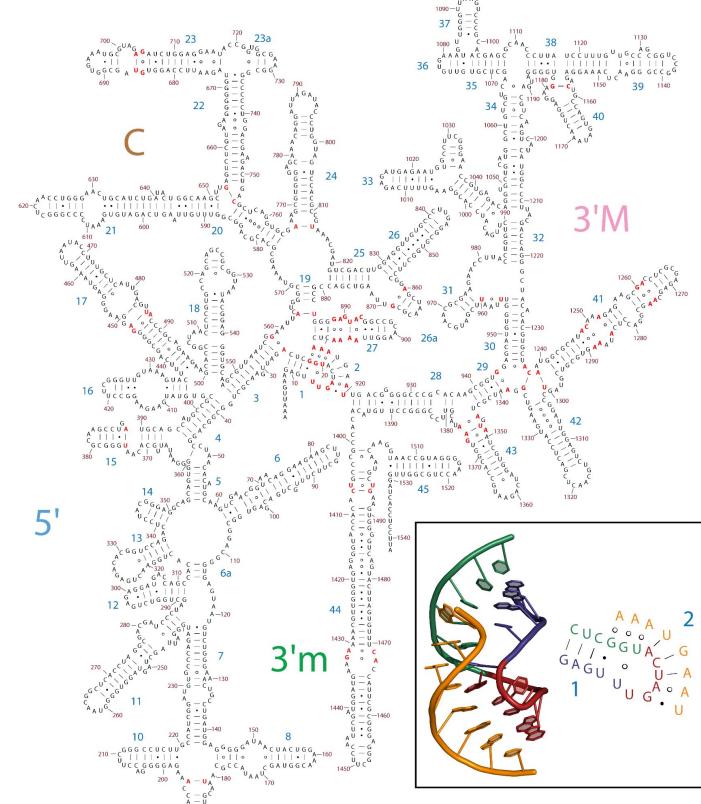


Figure source: [https://commons.wikimedia.org/wiki/File:TRNA-Phe\\_yeast\\_1ehz.png](https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png)

# Identification of rRNAs

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



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

The screenshot shows the main interface of the NONCODEV6 database. At the top, there is a search bar with placeholder text "Search a gene/transcript, eg. NONHSAG000148.2" and a red "Search" button. Below the search bar is a navigation bar with links: "Aliases", "Location", "Sequence", "Expression", "Orthologs", "Function", "Disease relation", and "RNA Structure".

The main content area features two circular navigation menus. The left menu, titled "Animals", contains links to "Browse DB", "Search", "Function", "Disease", "Conservation", "Genome", and "Id conversion". The right menu, titled "Plants", contains links to "Browse DB", "Search", "Function", "Conservation", "Genome", and "Id conversion". Both menus have a central image representing their respective organism groups.

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

# 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

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

# 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

- 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/SOOT>
  - 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

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

BLAST+ 2.13.0 is here!  
Starting with this release, we are including the blastn\_vdb and tblasn\_vdb executables in the BLAST+ distribution.

Thu, 17 Mar 2022 12:00:00 EST [More BLAST news...](#)

**Web BLAST**

Nucleotide BLAST  
nucleotide ▶ nucleotide

blastx  
translated nucleotide ▶ protein

tblastn  
protein ▶ translated nucleotide

Protein BLAST  
protein ▶ protein

**BLAST Genomes**

Enter organism common name, scientific name, or tax id  [Search](#)

Human Mouse Rat Microbes

**Standalone and API BLAST**

[Download BLAST](#)  
Get BLAST databases and executables

[Use BLAST API](#)  
Call BLAST from your application

[Use BLAST in the cloud](#)  
Start an instance at a cloud provider

**Specialized searches**

SmartBLAST  
Find proteins highly similar to your query

Primer-BLAST  
Design primers specific to your PCR template

Global Align  
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search  
Find conserved domains in your sequence

IgBLAST  
Search immunoglobulins and T cell receptor sequences

VecScreen  
Search sequences for vector contamination

CDART  
Find sequences with similar conserved domain architecture

Multiple Alignment  
Align sequences using domain and protein constraints

MOLE-BLAST  
Establish taxonomy for uncultured or environmental sequences



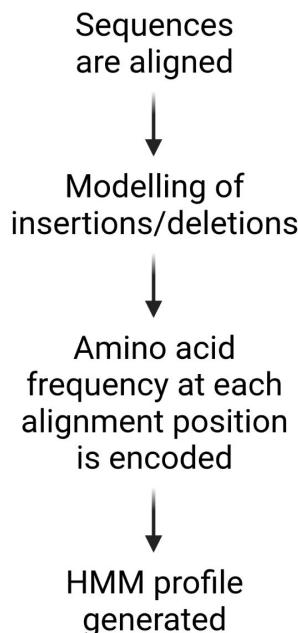
# Pfam: Protein family database

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

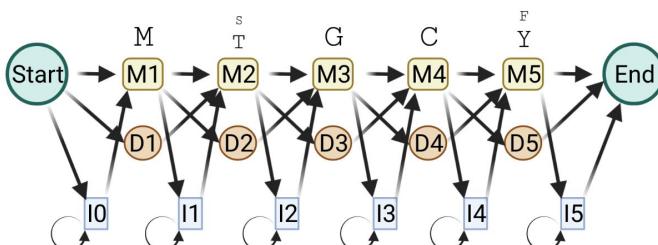
Pfam 35.0 (November 2021, 19632 entries)

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

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



seq1	MTGC - Y	2 = deletion
seq2	MSGC - F	5 = insertion
seq3	MTGC - Y	
seq4	M - GCAY	
	1 2 3 4 5 6	



## 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 type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

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

## Recent Pfam blog posts

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

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 [...]

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

Hot on the tail of our inclusion of the Baker group's trRosetta 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 Bileschi 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. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, 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

Pfam is an ELIXIR service [Read more](#)

Comments or questions on the site? Send a mail to [pfram-help@ebi.ac.uk](mailto:pfram-help@ebi.ac.uk)  
European Molecular Biology Laboratory

<http://pfam.xfam.org/>

Mistry et al., 2020: 10.1093/nar/gkaa913

# 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

## 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:KPYMF\_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

TIGRFAM       SFLD       Phobius       SignalP       SignalP\_EUK  
 SignalP\_GRAM\_POSITIVE       SignalP\_GRAM\_NEGATIVE       SUPERFAMILY       PANTHER       Gene3D  
 Hmmp       ProSiteProfiles       ProSitePatterns       Coils       SMART  
 CDD       PRINTS       Pfam       MobidBlite  
 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)

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

Tool Output Submission Details

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:KPYMF\\_HUMAN](http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5&sequence=uniprot:KPYMF_HUMAN)  
Jones et al., 2014: 10.1093/bioinformatics/btu031



Technische  
Universität  
Braunschweig

Boas Pucker| Plant Biotechnology & Bioinformatics | GE31-4 | 33

# 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
```

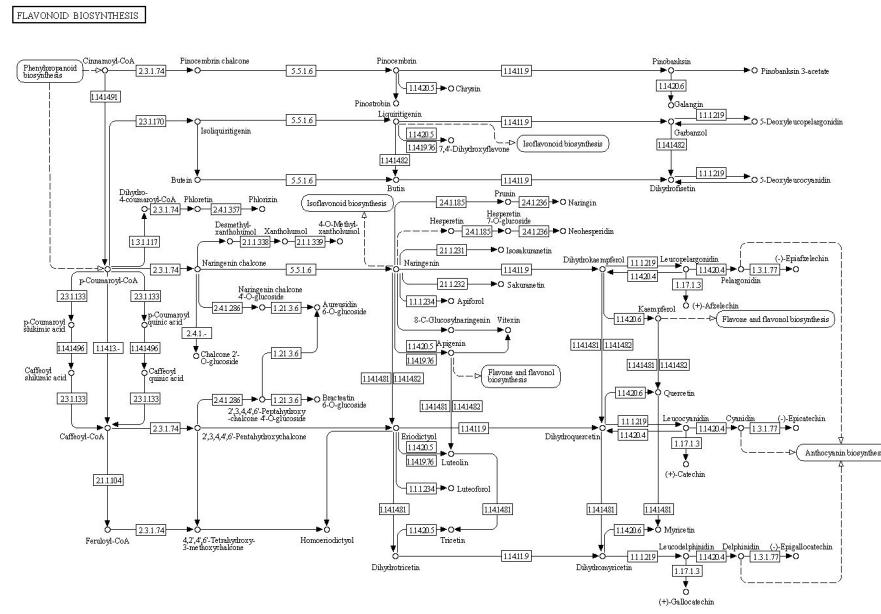
where INPUT\_FASTA is a fasta file containing the amino acid sequence for the search and SHOOT\_DB is the SHOOT database directory created using the steps above.

<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

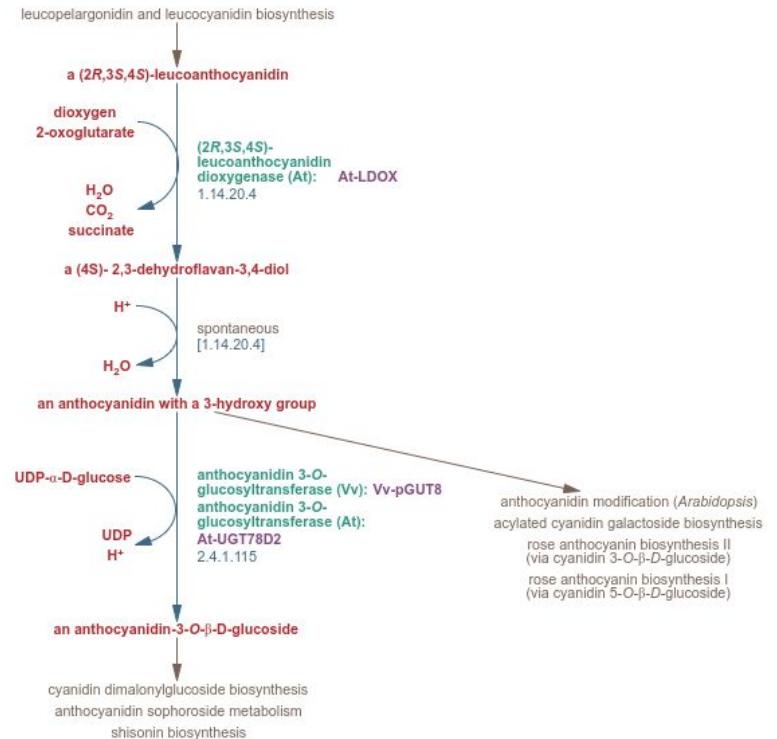
# Gene Ontology (GO)

- 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



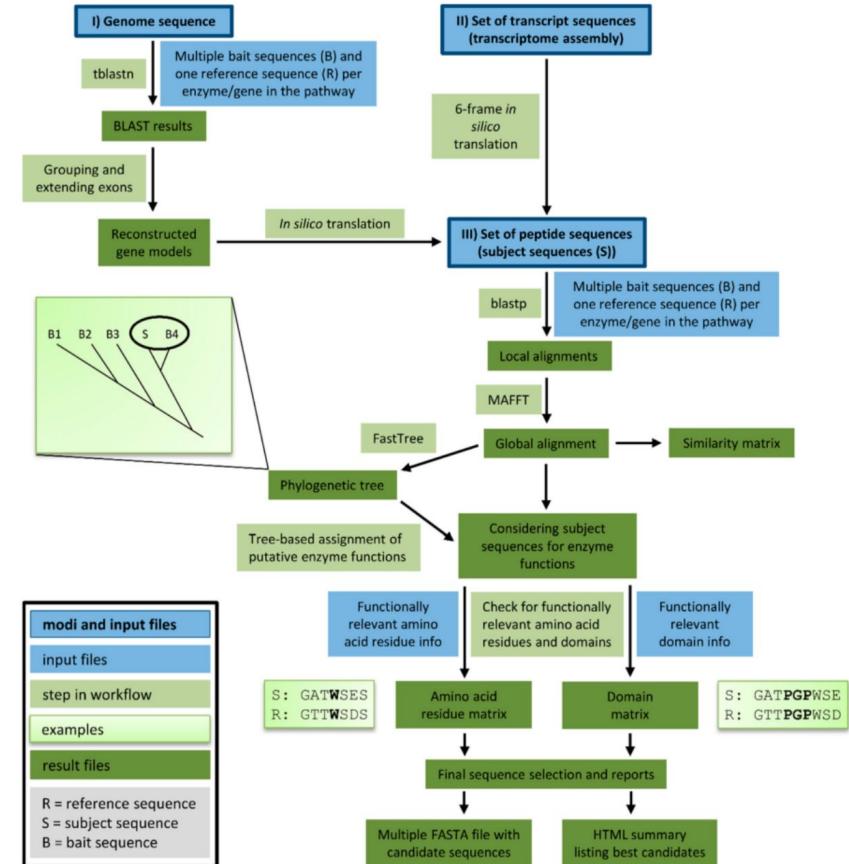
<https://metacyc.org/>

Karp et al., 2015: <https://doi.org/10.48550/arXiv.1510.03964>

Figure: <https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5125>

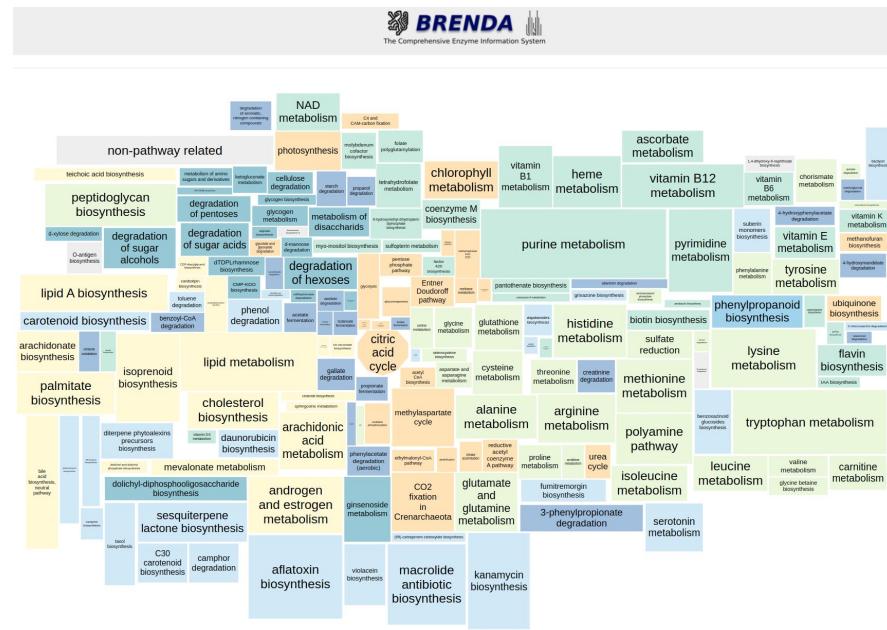
# KIPeS

- 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://pbb.bot.nat.tu-bs.de/KIPeS/>



# 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](http://www.brenda-enzymes.org)  
Chang et al., 2021: 10.1093/nar/gkaa1025

# 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
MIGPMPGMEGKLMLPGASPVGLEVVLVASTPILSDTSLCTPSFYHFLLLGPITSISNLIVRPFLLSSITVIYGRLCFFAFDFYVY
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRKGRPPKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHIPPTSIPSFNNPNNIHQSSSDRQMPP
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTAAATGDKKRTKARKETYSSYIYKVLKQVHPDTGISNRAMSILNSFVNDFERVATEASKLA
>TRINITY_DN10001_c0_g1_i2
MAKVGNPVIDETDGSVNEPESSEKNIEVSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREHNIDINNIKGTKNGRITKEDILNYV
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFKPGWKEFVRNSDLEGGDFLVNLVDKISYQVVIFDGTCACPDKLCPFSIMNPIFIQHLRNKIFLSKKEEIKLKGNRKVHSVNEN
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVVFVAFVTAGYPDSEETVDILLGLEAGGADIIIELGIPFTDPMVDGKTIQDANNVALENKIDISKCLSYVSESRAK
```

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

# Time for questions!



# Questions

1. What are the important features of a eukaryotic gene?
2. What are the important steps of an RNA-seq experiment?
3. Which hints can be used in a gene prediction process?
4. How would you perform a gene prediction?
5. What is the relevance of non-canonical splice sites?
6. Which types of features can be annotated in a genome sequence?
7. Which tools are available for the annotation of different gene types?
8. How can TEs be annotated?
9. Which tools can be used for the functional annotation of genes?