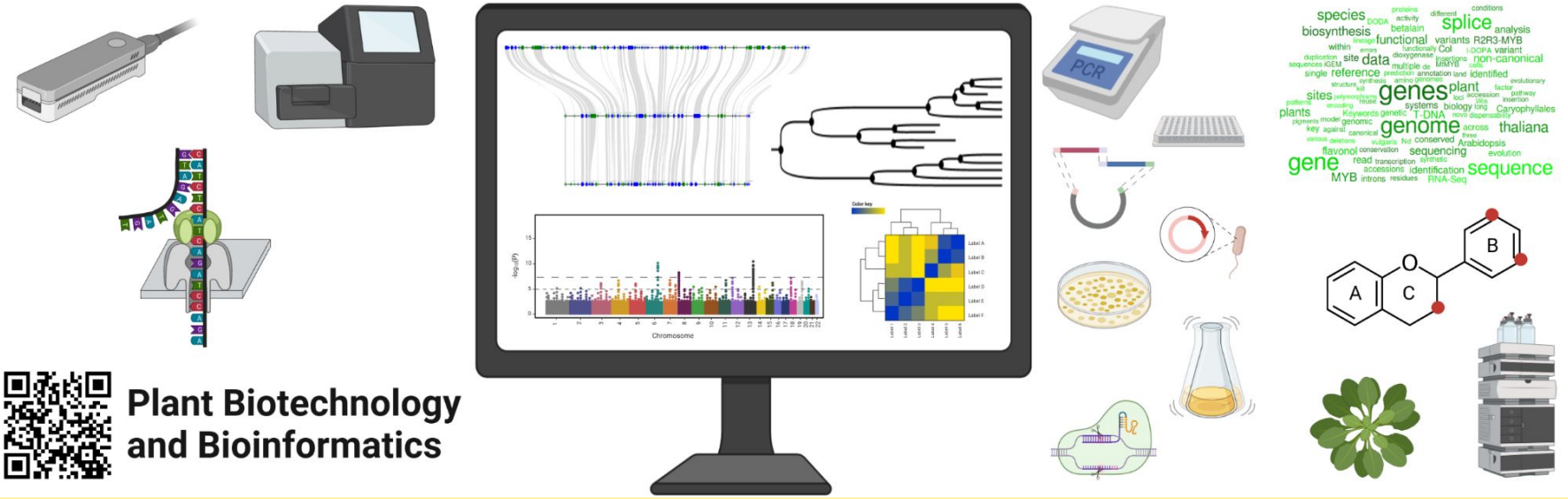
Technische Universität Braunschweig



Plant Biotechnology and Bioinformatics

Comparative genomics

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

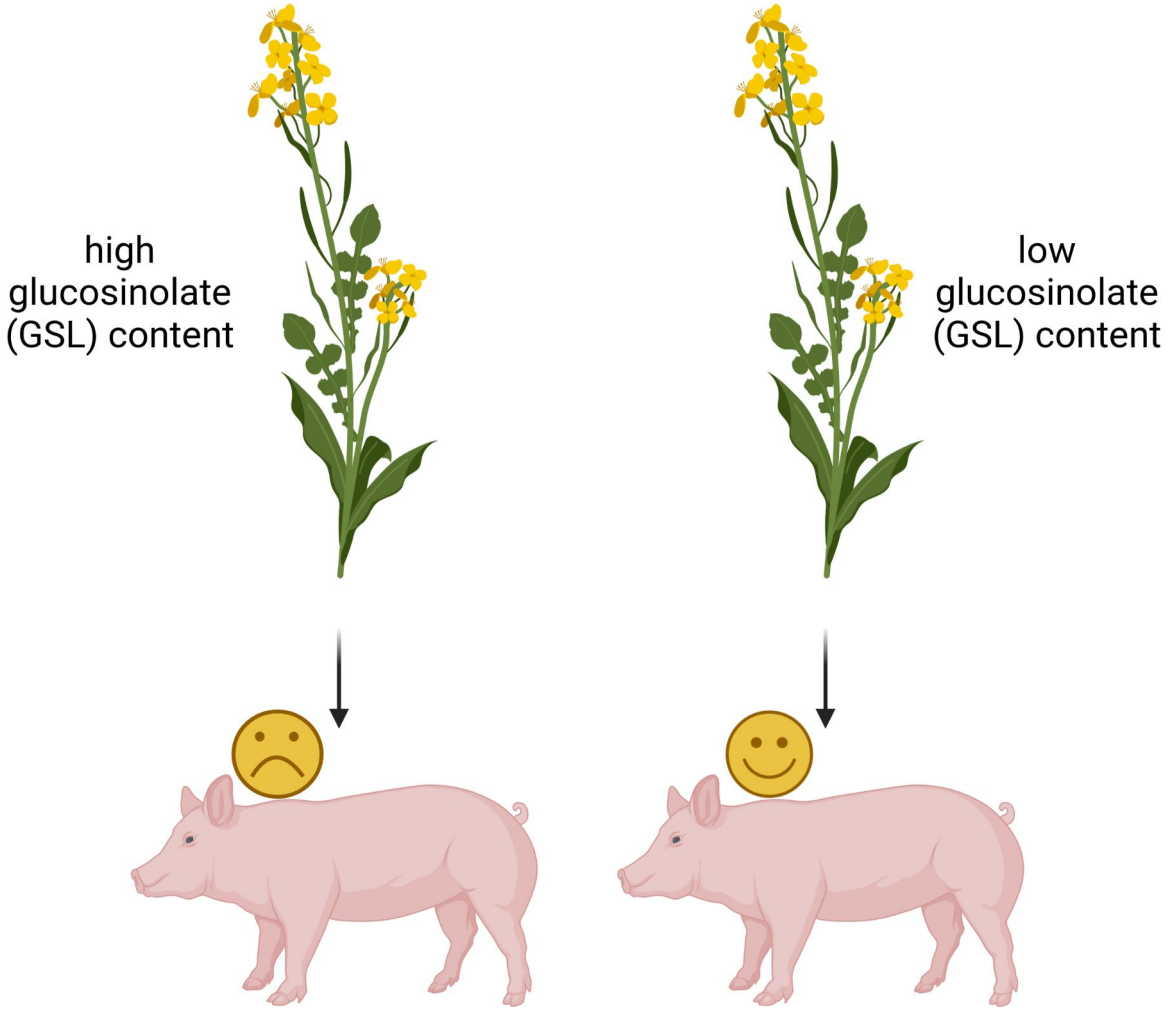
Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: Data Literacy in Genomics
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: b.pucker[a]tu-bs.de

Question: What determines differences between plants?



Intraspecific variation

- Different accessions of the same species have genomic differences
- Genotype determines phenotype (traits)
- Specific traits are relevant for breeding or biotechnological/agronomical applications

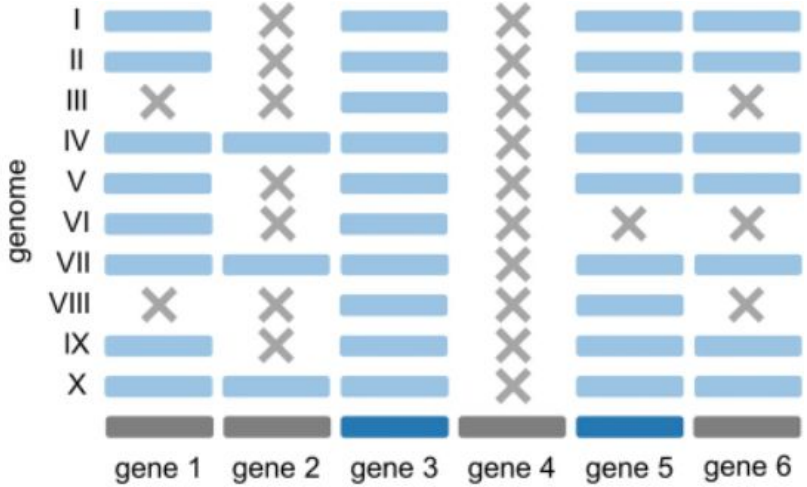
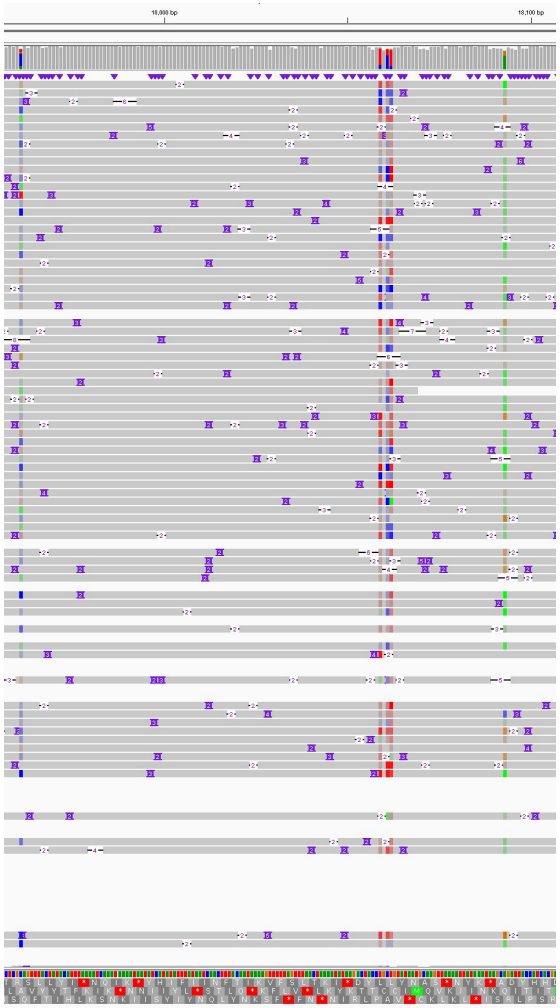


Figure source: 10.1186/s13007-021-00718-5

Differences between samples and reference

- Differences are identified against a reference genome sequence (classic approach)
- Reads of a sample are aligned against the reference (mapping)
- Differences are identified by tools (variant calling)



Read mapping

- Alignment of reads to a genome sequence
- Mapping needs to be fast at the cost of accuracy
- Long read mapping tools: minimap2, GraphMap, NGMLR
- Manual inspection of read mappings via Integrated Genomics Viewer (IGV)

SAM/BAM

- SAM = Sequence Alignment/Map format
- BAM = Binary Alignment/Map format (binary version of SAM)
- Another way to store read information: contains information from FASTA and FASTQ file (reads mapped to reference)

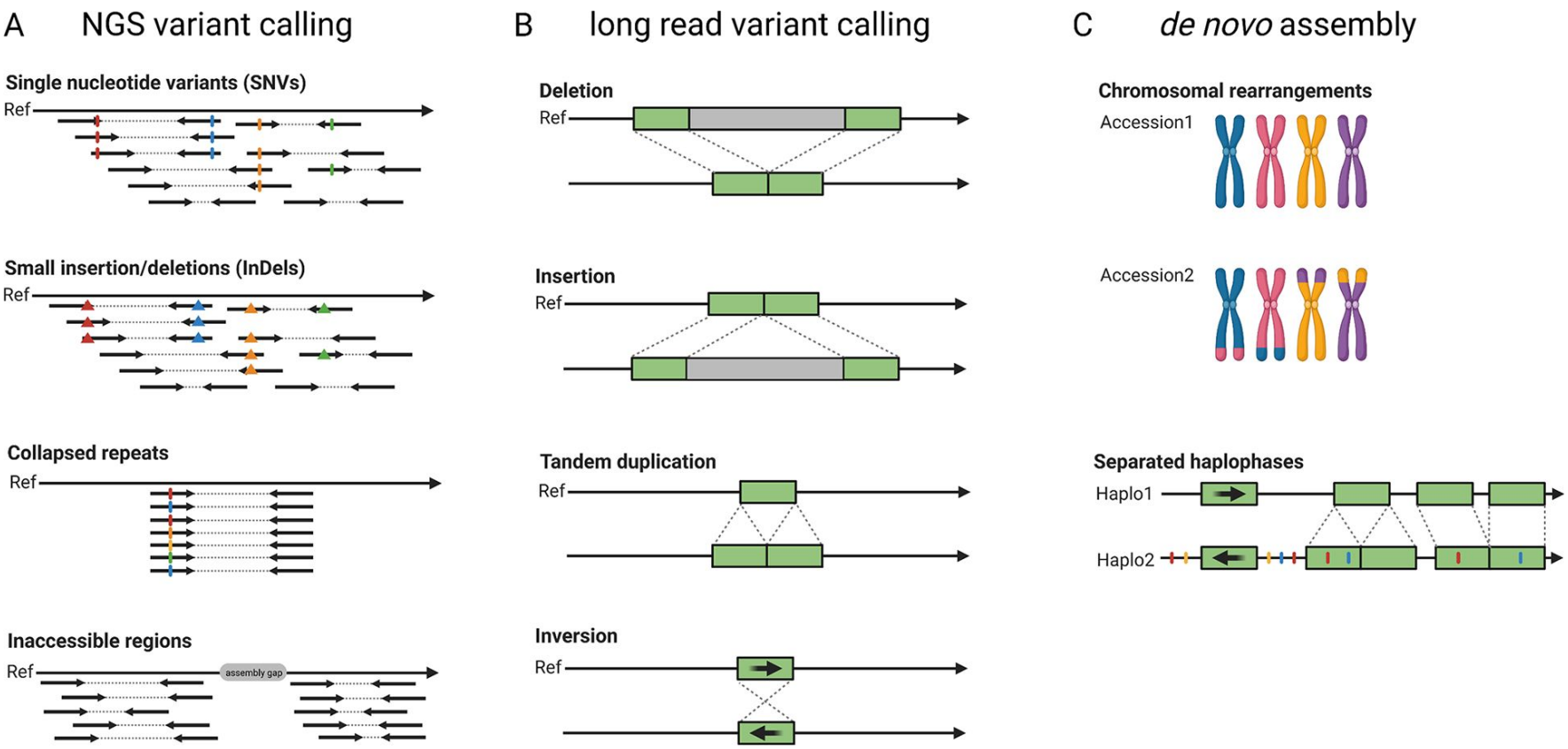
Variant calling

- Identification of sequence differences between reads and reference sequence
- Differences are listed in a specific file type: Variant Caller Format (VCF)
- ONT long reads are well suited for the identification of large structural variants

Types of sequence variants

- SNVs vs. SNPs
  - Single Nucleotide Variants = no information about their effect
  - Single Nucleotide Polymorphisms = no detrimental impact
- MNVs
  - Multiple Nucleotide Variants
- InDels
  - Insertions: additional bases in sample compared to reference
  - Deletions: loss of bases in sample compared to reference
- Inversions
  - Orientation of sequence differs between sample and reference
- Tandem duplications

Read mapping vs. de novo genome sequence assembly



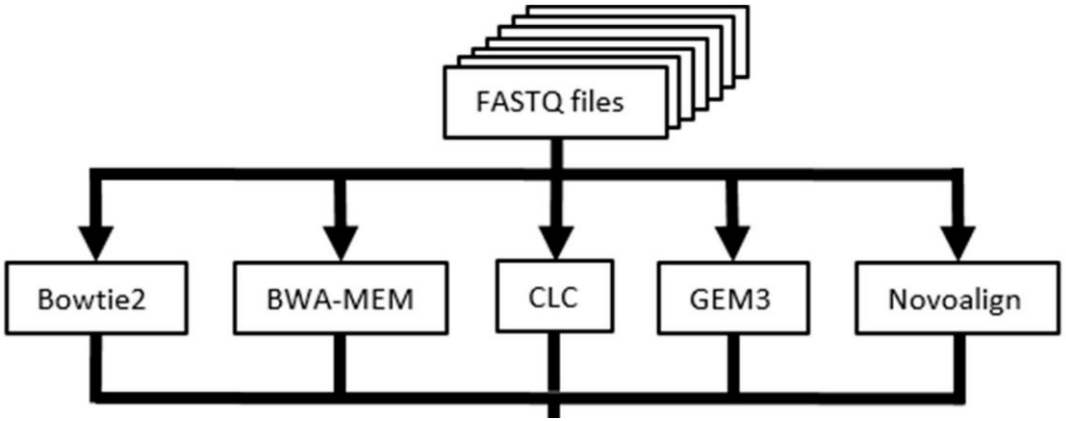
Pucker et al., 2022: 10.1017/qpb.2021.18

Challenges of read mappings

- Speed / computational costs
  - BLAST would be too slow
- Specific assignment of reads / repeats
  - (especially challenging in polyploid species)
- Splitting around InDels or across introns

Short read mappings

- Large number of tools for mappings
- DNA seq read mappers: BWA MEM, bowtie2
- RNA-seq read mappers (split read): STAR, HiSAT2



Li, 2013: 10.48550/arXiv.1303.3997
Langmead & Salzberg, 2012: 10.1038/nmeth.1923
Dobin et al., 2013: 10.1093/bioinformatics/bts635
Kim et al., 2015: 10.1038/nmeth.3317
Schilbert et al., 2020: 10.3390/plants9040439

Coverage analysis: genomic copy numbers

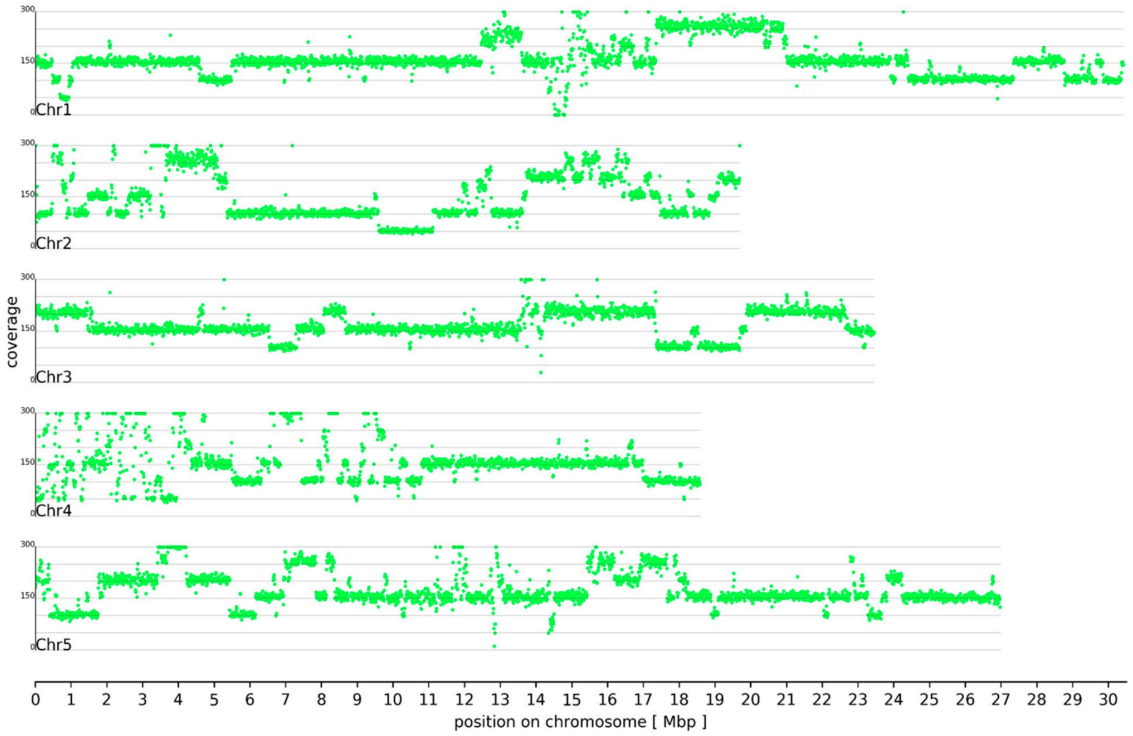


Figure source: 10.3390/genes10090671

Coverage analysis: segmental duplication

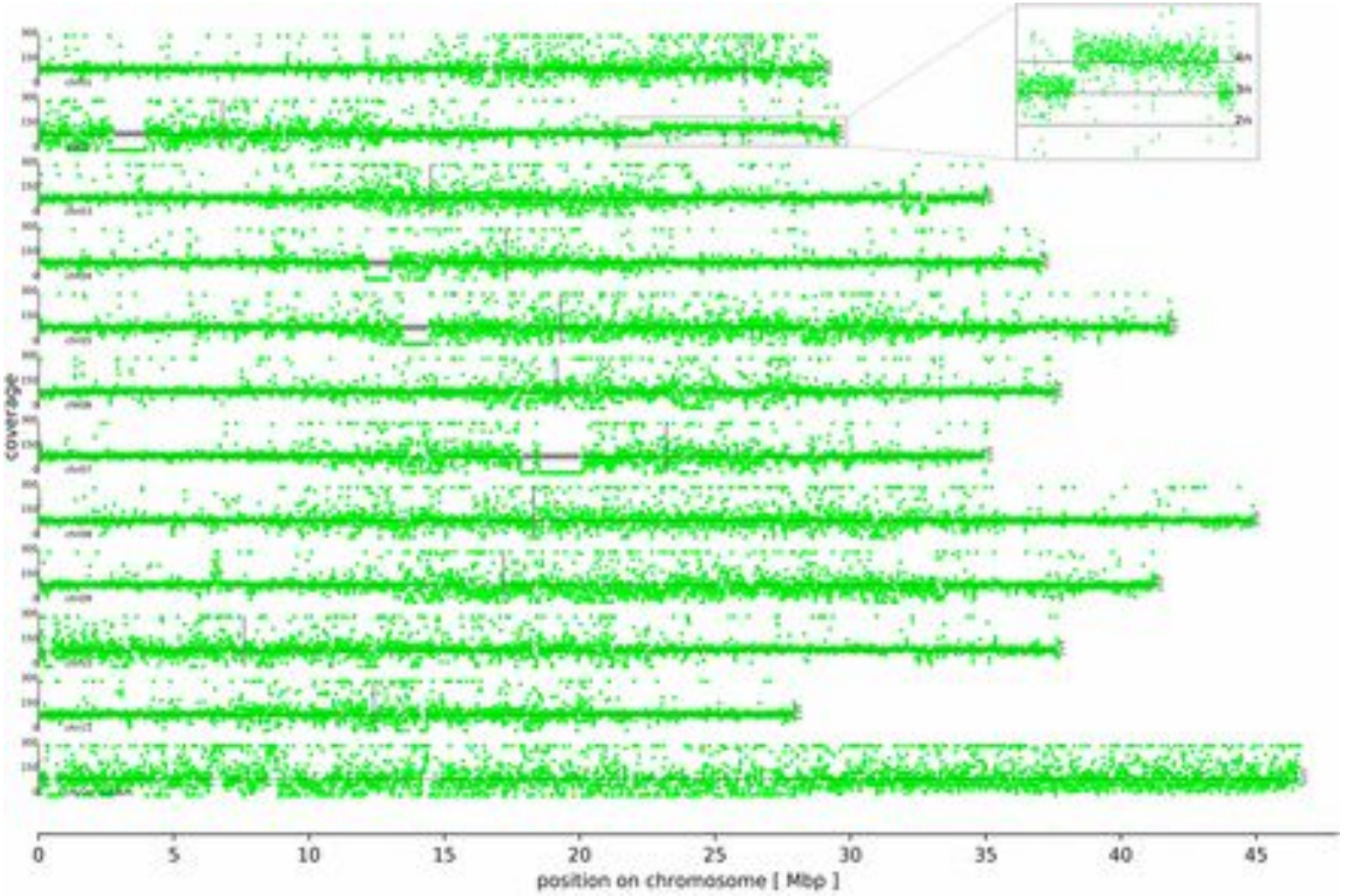
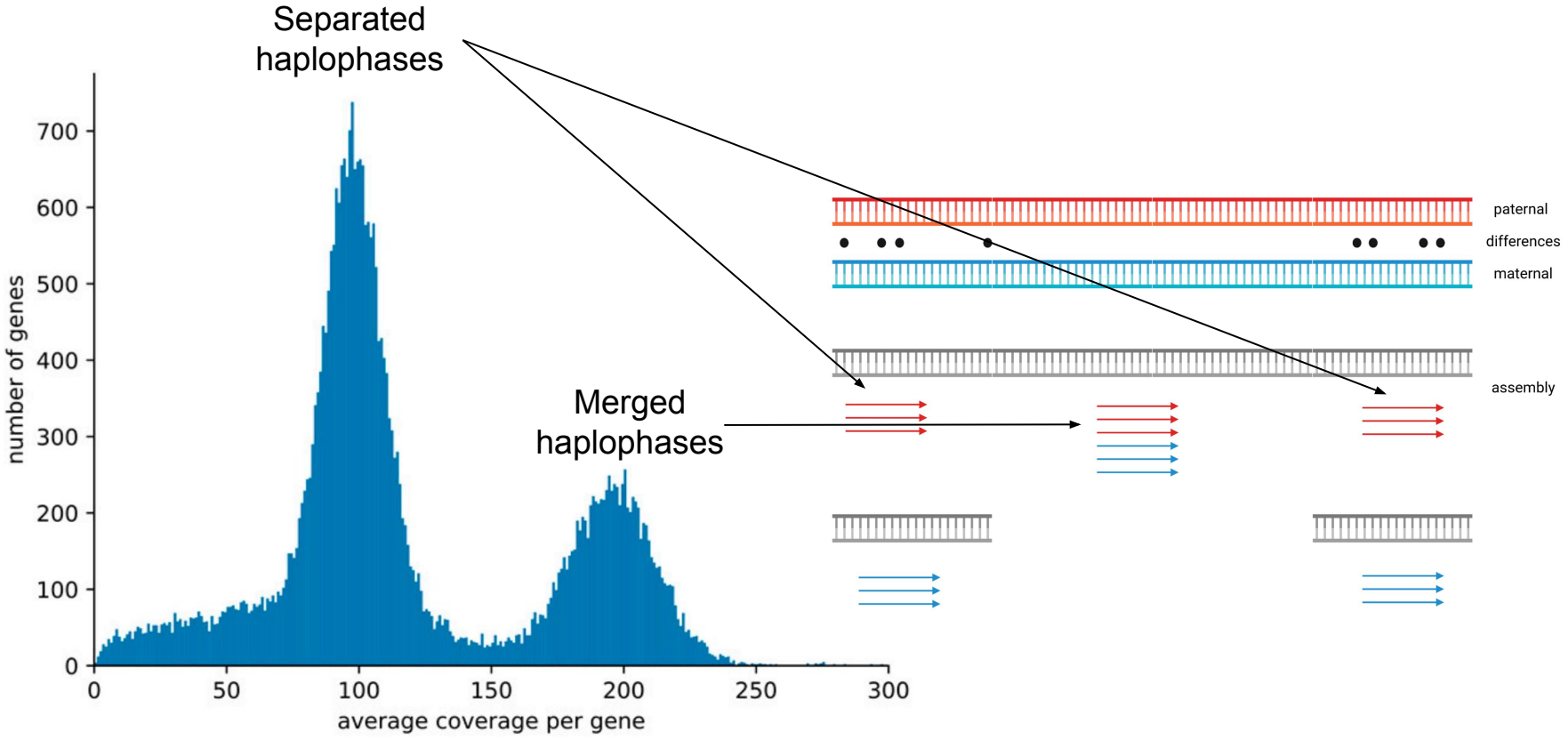
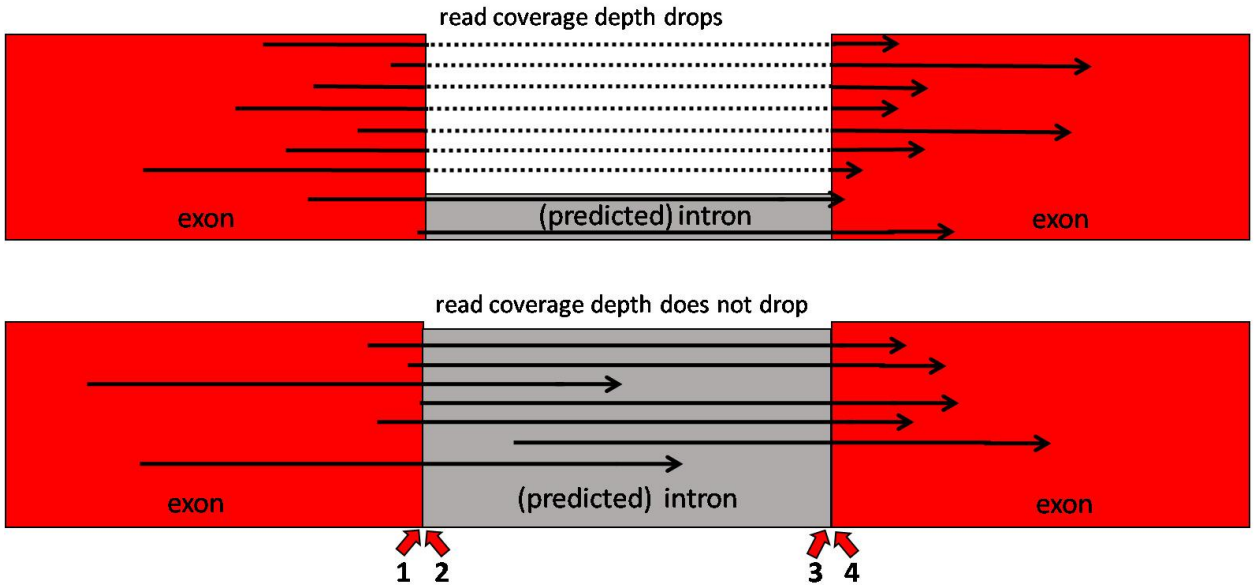


Figure source: 10.1534/g3.119.400847

Coverage analysis: haplotype phasing / ploidy



Coverage analysis: splice sites



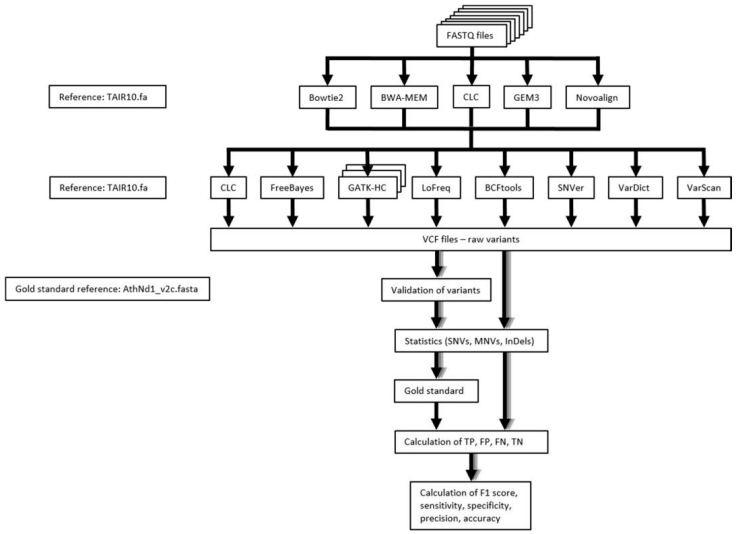
Pucker & Brockington, 2018: 10.1186/s12864-018-5360-z

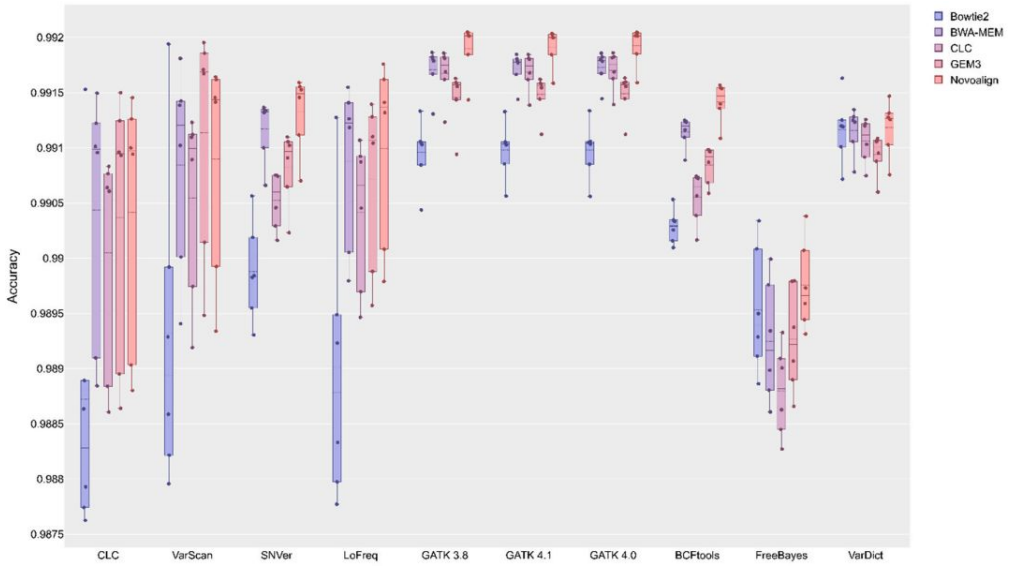
Short read variant callers

- Millions of differences between samples and reference
- Automatic detection with tools required
- Tools:
  - Genome Analysis Tool Kit (GATK): https://gatk.broadinstitute.org/hc/en-us
  - Samtools: http://www.htslib.org/
  - VarDict: https://anaconda.org/bioconda/vardict
  - VarScan: https://github.com/dkoboldt/varscan
  - SNVer: http://snver.sourceforge.net/
  - FreeBayes: https://github.com/freebayes/freebayes

McKenna et al., 2010: 10.1101/gr.107524.110
Li et al., 2009: 10.1093/bioinformatics/btp352
Lai et al., 2016: 10.1093/nar/gkw227
Koboldt et al., 2009: 10.1093/bioinformatics/btp373
Wei et al., 2011: 10.1093/nar/gkr599
Garrison & Marth, 2012: 10.48550/arXiv.1207.3907

Benchmarking of variant callers

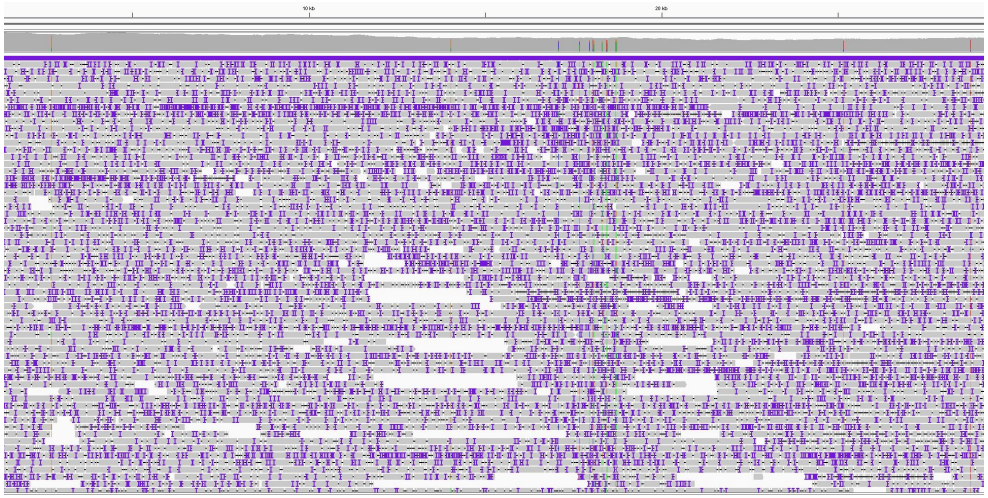




Schilbert et al., 2020: 10.3390/plants9040439

Long read mapping

- Noisy alignments due to high rate of sequencing errors
- InDels are particularly abundant (purple ticks)
- Rapid development of new tools
- Benchmarking studies are missing and would require frequent updates



Long read mapping tools

- Collection of long read tools: https://long-read-tools.org/
- Over 600 tools available
- Examples:
  - BWA-SW: http://bio-bwa.sourceforge.net/
  - Minimap2: https://github.com/lh3/minimap2
  - GraphMap: https://github.com/isovic/graphmap
  - NGMLR: https://github.com/philres/ngmlr
  - Winnowmap2: https://github.com/marbl/Winnowmap

Amarasinghe et al., 2021: 10.1093/gigascience/giab003
Li & Durbin, 2010: 10.1093/bioinformatics/btp698
Li, 2018: 10.1093/bioinformatics/bty191
Sovic, 2016: 10.1038/ncomms11307
Sedlazeck et al., 2018: 10.1038/s41592-018-0001-7
Jain et al., 2022: 10.1038/s41592-022-01457-8

Long read variant calling

- Rapid development of new tools and lack of benchmarking studies

- Tools:
  - SVIM2: https://github.com/eldariont/svim
  - Longshot: https://github.com/pjedge/longshot
  - NanoCaller: https://github.com/WGLab/NanoCaller
  - Sniffles2: https://github.com/fritzsedlazeck/Sniffles

Heller & Vingron, 2019: 10.1093/bioinformatics/btz041
Edge & Basal, 2019: 10.1038/s41467-019-12493-y
Ahsan et al., 2021: 10.1186/s13059-021-02472-2
Smolka et al., 2022: 10.1101/2022.04.04.487055

Variant Call Format (VCF)

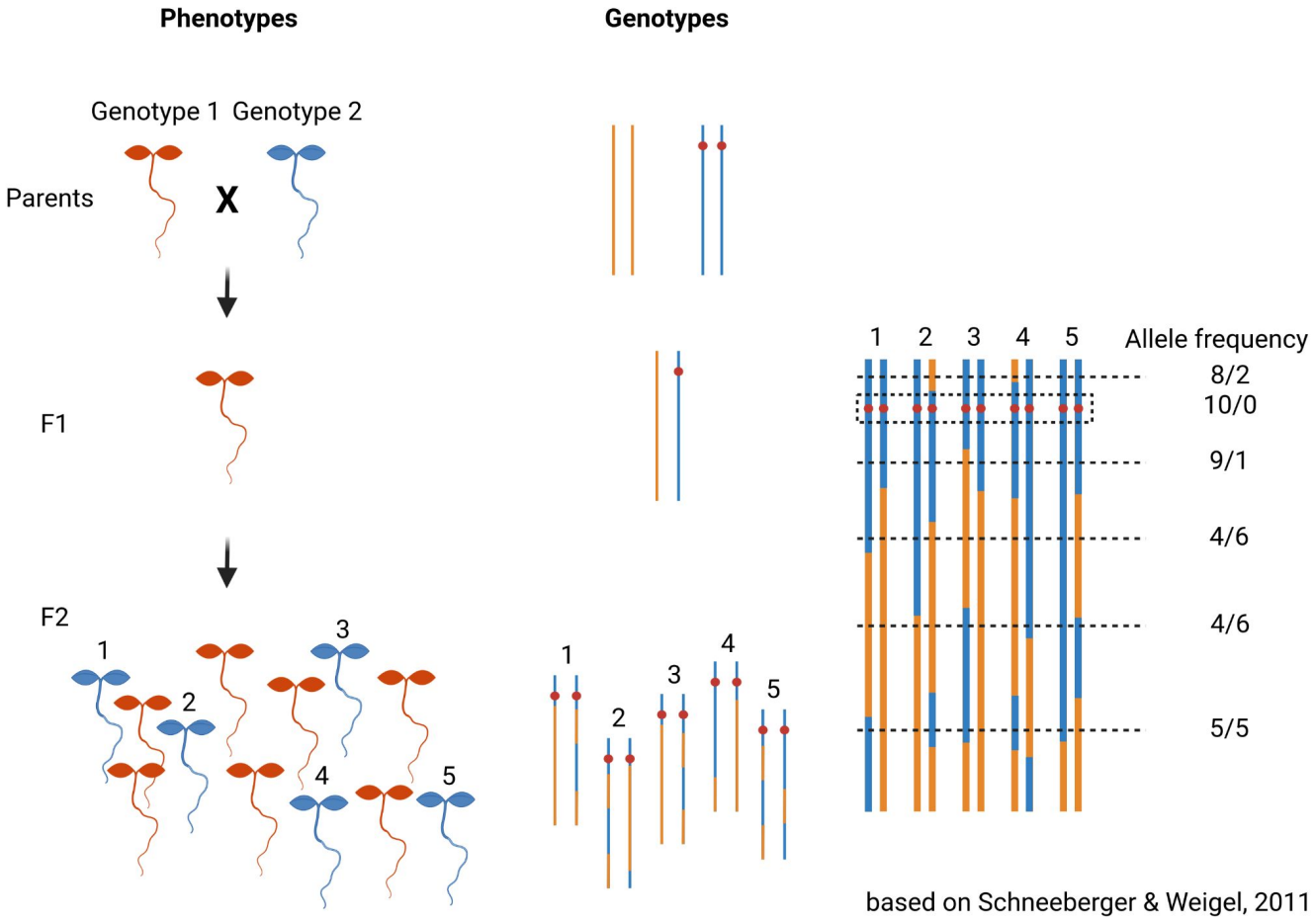
- Chromosome: name of sequence in the reference
- Position: position on the specified sequence
- ID (not relevant in plant biology): variants in humans have IDs
- Reference allele: nucleotide(s) in the reference sequence at the specified position
- Alternative allele: nucleotide(s) in the sample at the same position
- Quality: tool specific value that can be used for filtering
- Filter status: specifies filter name if variant was filtered out
- Info: collection of information
- Format: explains the fields in the following sample columns
- DATA_SET1, DATA_SET2, DATA_SET3, …: one column per sample

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	DATA_SET
R105	10929	.	A	G	58.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-1.242;DP=8;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=44.01;MQRankSum=-2.100;QD=7.33;ReadPosRankSum=0.619;SOR=0.307	GT:AD:DP:GQ:PL	0/1:6,2:8:66:66,0,230
R105	10944	.	A	G	58.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-1.068;DP=8;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=44.01;MQRankSum=-2.100;QD=7.33;ReadPosRankSum=-1.465;SOR=0.307	GT:AD:DP:GQ:PL	0/1:6,2:8:66:66,0,230
R105	10955	.	G	T	61.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-1.068;DP=7;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.16;MQRankSum=-1.981;QD=8.81;ReadPosRankSum=-1.465;SOR=0.446	GT:AD:DP:GQ:PL	0/1:5,2:7:69:69,0,204
R105	10962	.	A	G	61.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-1.068;DP=8;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=48.11;MQRankSum=-1.981;QD=8.81;ReadPosRankSum=-1.465;SOR=0.446	GT:AD:DP:GQ:PL	0/1:5,2:7:69:69,0,204
R105	13234	.	T	C	32.64	MQ_filter;QD_filter	AC=1;AF=0.500;AN=2;BaseQRankSum=1.062;DP=35;ExcessHet=3.0103;FS=2.782;MLEAC=1;MLEAF=0.500;MQ=37.70;MQRankSum=-1.106;QD=1.31;ReadPosRankSum=-0.209;SOR=1.911	GT:AD:DP:GQ:PL	0/1:22,3:25:40:40,0,706
R105	13260	.	T	G	72.64	MQ_filter	AC=1;AF=0.500;AN=2;BaseQRankSum=-0.172;DP=36;ExcessHet=3.0103;FS=11.698;MLEAC=1;MLEAF=0.500;MQ=38.68;MQRankSum=-1.349;QD=2.50;ReadPosRankSum=0.243;SOR=3.442	GT:AD:DP:GQ:PL	0/1:23,6:29:80:80,0,695
R105	13395	.	T	A	183.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=1.979;DP=19;ExcessHet=3.0103;FS=8.929;MLEAC=1;MLEAF=0.500;MQ=43.94;MQRankSum=-1.631;QD=9.67;ReadPosRankSum=1.186;SOR=3.126	GT:AD:DP:GQ:PL	0/1:13,6:19:99:191,0,413
R105	13538	.	T	C	61.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=1.359;DP=27;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.51;MQRankSum=-1.847;QD=2.47;ReadPosRankSum=1.188;SOR=0.760	GT:AD:DP:GQ:PL	0/1:21,4:25:69:69,0,660
R105	14511	.	C	T	39.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-0.563;DP=13;ExcessHet=3.0103;FS=2.522;MLEAC=1;MLEAF=0.500;MQ=43.42;MQRankSum=-2.771;QD=3.05;ReadPosRankSum=0.217;SOR=1.546	GT:AD:DP:GQ:PL	0/1:9,4:13:47:47,0,352
R105	14542	.	G	A	51.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-1.825;DP=15;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=42.30;MQRankSum=-2.843;QD=3.44;ReadPosRankSum=-0.919;SOR=1.022	GT:AD:DP:GQ:PL	0/1:10,5:15:59:59,0,390
R105	15681	.	A	G	116.64	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=1.897;DP=43;ExcessHet=3.0103;FS=4.187;MLEAC=1;MLEAF=0.500;MQ=40.84;MQRankSum=-2.008;QD=2.92;ReadPosRankSum=0.303;SOR=1.721	GT:AD:DP:GQ:PL	0/1:34,6:40:99:124,0,1161
R105	15712	.	C	T	70.64	MQ_filter	AC=1;AF=0.500;AN=2;BaseQRankSum=1.307;DP=35;ExcessHet=3.0103;FS=2.026;MLEAC=1;MLEAF=0.500;MQ=39.58;MQRankSum=-2.129;QD=2.08;ReadPosRankSum=-0.994;SOR=1.230	GT:AD:DP:GQ:PL	0/1:30,4:34:78:78,0,1247
R105	15719	.	C	T	64.64	MQ_filter;QD_filter	AC=1;AF=0.500;AN=2;BaseQRankSum=0.166;DP=37;ExcessHet=3.0103;FS=2.037;MLEAC=1;MLEAF=0.500;MQ=38.62;MQRankSum=-2.020;QD=1.80;ReadPosRankSum=-1.360;SOR=1.251	GT:AD:DP:GQ:PL	0/1:32,4:36:72:72,0,1310
R105	16037	.	C	A	50.64	MQ_filter	AC=1;AF=0.500;AN=2;BaseQRankSum=0.000;DP=7;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=34.75;MQRankSum=0.303;QD=7.23;ReadPosRankSum=-0.712;SOR=1.802	GT:AD:DP:GQ:PL	0/1:5,2:7:58:58,0,171
R105	16582	.	CATTGTAAG	C	58.60	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=0.712;DP=8;ExcessHet=3.0103;FS=3.680;MLEAC=1;MLEAF=0.500;MQ=38.22;MQRankSum=0.319;QD=7.33;ReadPosRankSum=0.366;SOR=2.236	GT:AD:DP:GQ:PL	0/1:6,2:8:66:66,0,246
R105	17718	.	A	G	79.64	MQ_filter	AC=1;AF=0.500;AN=2;BaseQRankSum=-0.722;DP=18;ExcessHet=3.0103;FS=7.656;MLEAC=1;MLEAF=0.500;MQ=29.01;MQRankSum=-0.858;QD=4.42;ReadPosRankSum=0.099;SOR=0.041	GT:AD:DP:GQ:PL	0/1:13,5:18:87:87,0,392
R105	37909	.	A	G	51.64	MQ_filter	AC=1;AF=0.500;AN=2;BaseQRankSum=-0.431;DP=9;ExcessHet=3.0103;FS=10.792;MLEAC=1;MLEAF=0.500;MQ=37.28;MQRankSum=0.549;QD=5.74;ReadPosRankSum=-0.549;SOR=2.206	GT:AD:DP:GQ:PL	0/1:7,2:9:59:59,0,254
R105	45724	.	C	CTTATA	61.60	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=0.210;DP=7;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=43.83;MQRankSum=0.967;QD=8.80;ReadPosRankSum=-0.210;SOR=0.446	GT:AD:DP:GQ:PL	0/1:5,2:7:69:69,0,204

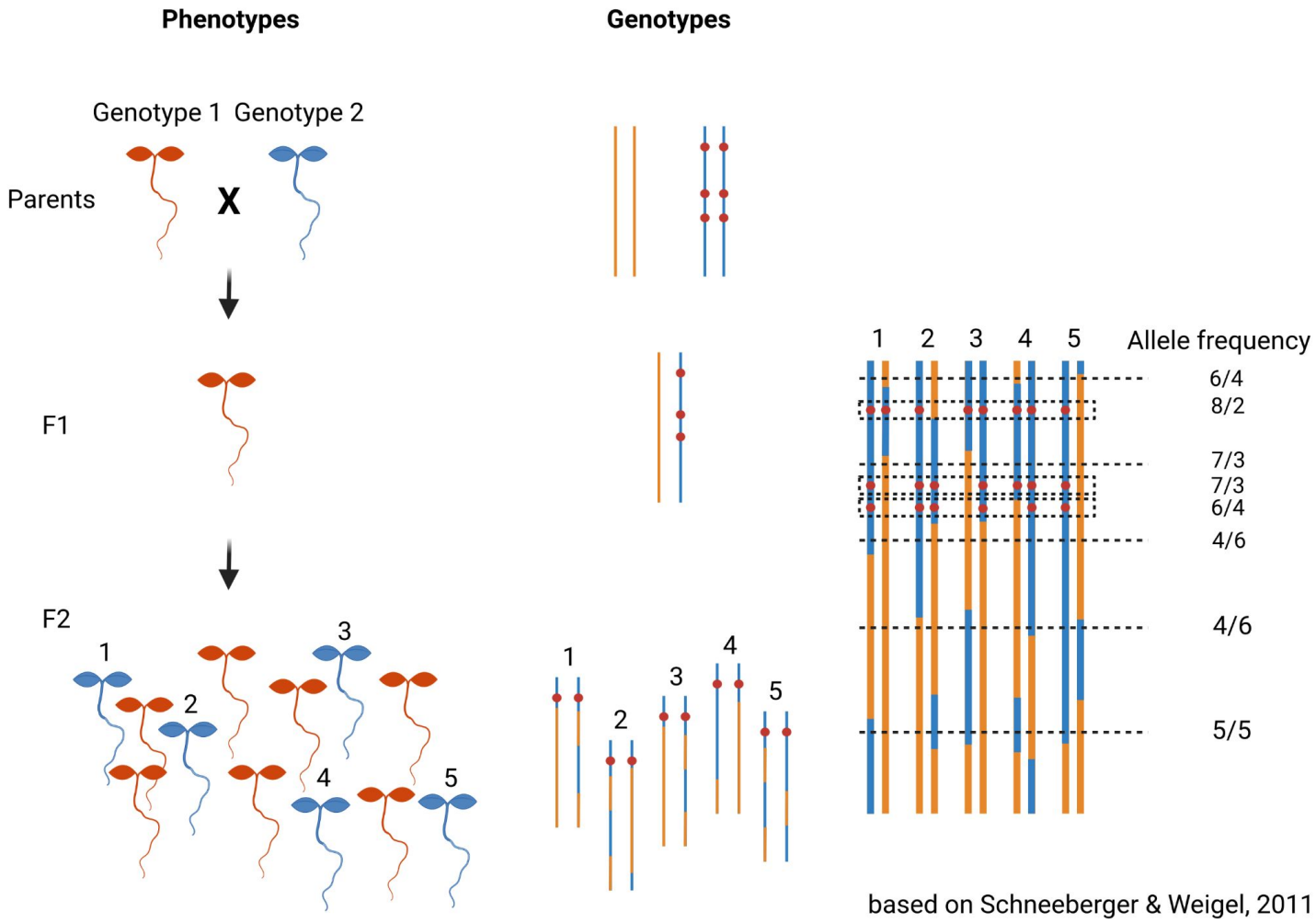
genomic Variant Call Format (gVCF)

- Generally similar to VCF file
- One entry for each position in the reference
- Not just variants, but also non-variants are listed

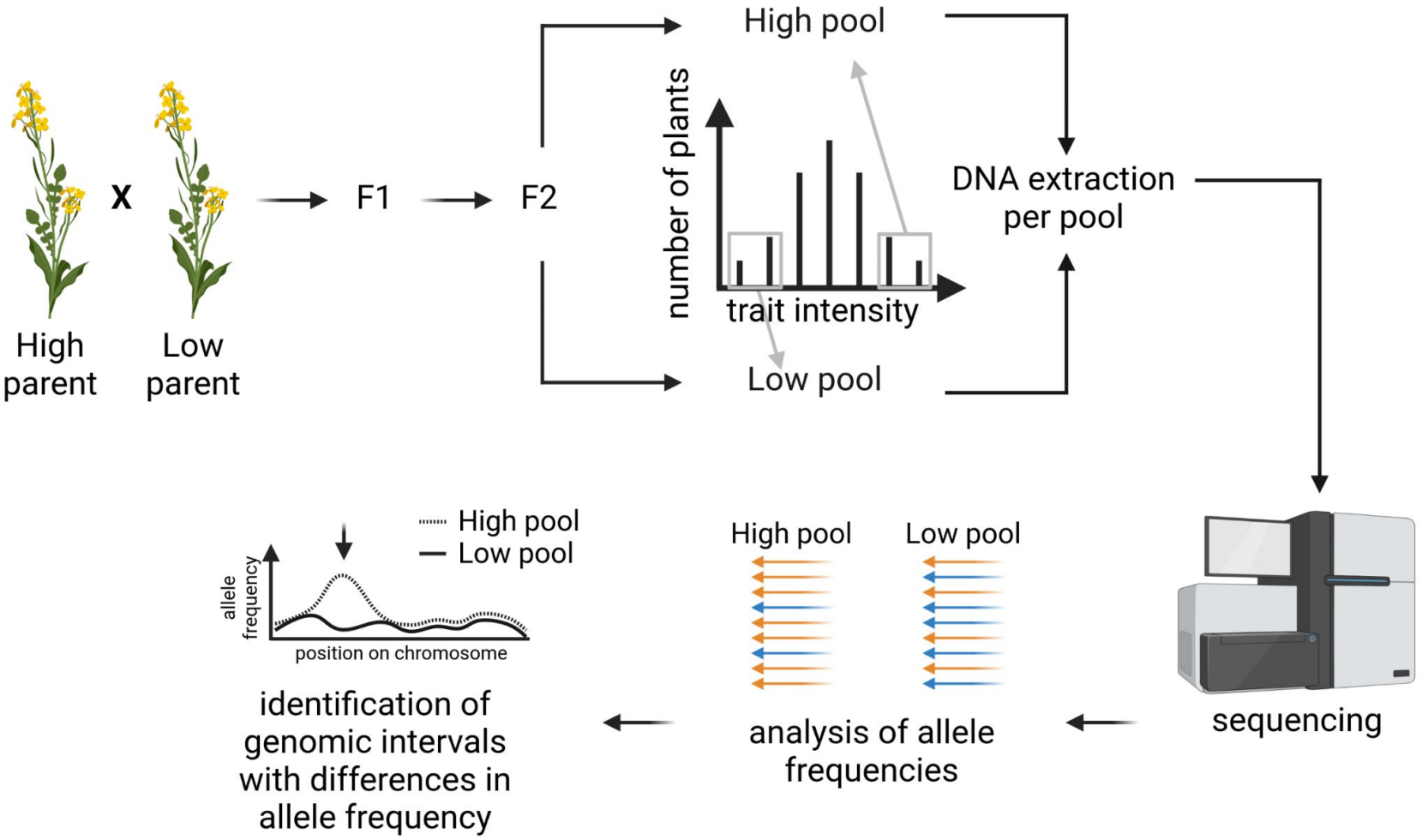
Mapping-by-sequencing (1)



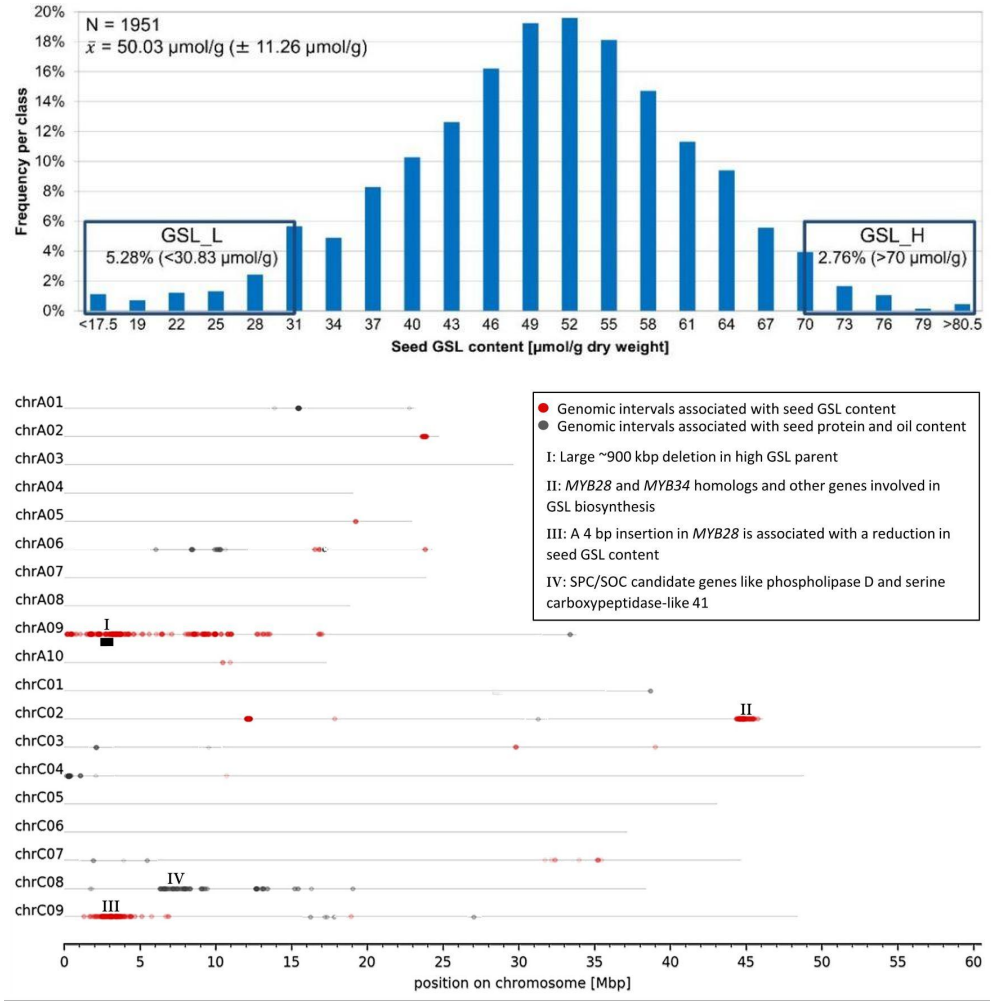
Mapping-by-sequencing (2)



Mapping-by-sequencing (3)



Mapping-by-sequencing (4)



Schilbert et al., 2022: 10.1101/2022.06.01.494149

Variant annotation

- Variants can have functional consequences
- Variants in coding sequences can change amino acids
- Variants outside CDS can have regulatory consequences

Variant Annotation - SnpEff

- High throughput annotation of SNVs and larger variants
- Reference sequence (FASTA) and structural annotation (GFF) required
- Freely available software: http://pcingola.github.io/SnpEff/
- Addition of one annotation field to VCF file

```
chrA01  179964  .  T  A  .  PASS  GSL_High,GSL_Low;ANN=A|stop_gained|HIGH|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.863T>A|p.Leu288*|863/1834|863/1834|288/610|   GT:AD:DP:GQ:PL
chrA01  179967  .  T  C  .  PASS  GSL_High,GSL_Low;ANN=C|missense_variant|MODERATE|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.866T>C|p.Phe289Ser|866/1834|866/1834|289/610|       GT:AD:DP:GQ:PL
chrA01  179979  .  C  T  .  PASS  GSL_Low;ANN=T|missense_variant|MODERATE|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.878C>T|p.Thr293Met|878/1834|878/1834|293/610|         GT:AD:DP:GQ:PL
chrA01  179988  .  A  T  .  PASS  GSL_Low;ANN=T|missense_variant|MODERATE|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.887A>T|p.Tyr296Phe|887/1834|887/1834|296/610|         GT:AD:DP:GQ:PL
chrA01  180003  .  G  A  .  PASS  GSL_Low;ANN=A|missense_variant|MODERATE|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.902G>A|p.Gly301Asp|902/1834|902/1834|301/610|         GT:AD:DP:GQ:PL
chrA01  180024  .  A  G  .  PASS  GSL_High,GSL_Low;ANN=G|missense_variant|MODERATE|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.923A>G|p.His308Arg|923/1834|923/1834|308/610|       GT:AD:DP:GQ:PL
chrA01  180039  .  T  C  .  PASS  GSL_High,GSL_Low;ANN=C|missense_variant|MODERATE|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.938T>C|p.Ile313Thr|938/1834|938/1834|313/610|       GT:AD:DP:GQ:PL
chrA01  180048  .  G  A  .  PASS  GSL_High,GSL_Low;ANN=A|stop_retained_variant|LOW|Gene_BnaA01g00340D|Gene_BnaA01g00340D|transcript|BnaA01g00340D|Coding|5/9|c.947G>A|p.Ter316Ter|947/1834|947/1834|316/610|       GT:AD:DP:GQ:PL
```

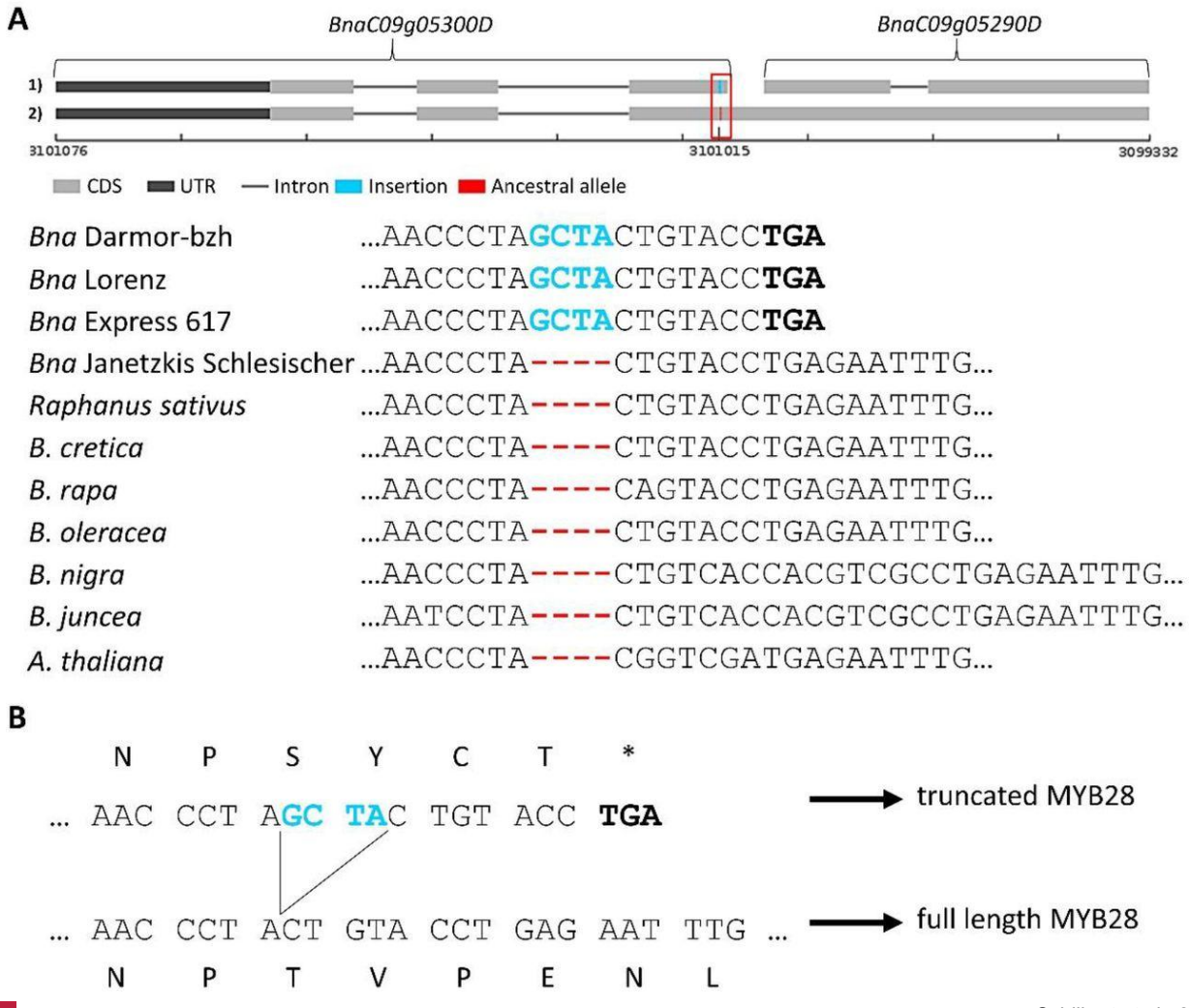
Cingolani et al., 2012: 10.4161/fly.19695

SnpEff predictions

Type	Meaning	Example
SNP	Single Nucleotide Polymorphism	Reference = ‘A’, Sample = ‘C’
Ins	Insertion	Reference = ‘A’, Sample = ‘AGT’
Del	Deletion	Reference = ‘AC’, Sample = ‘C’
MNP	Multiple Nucleotide Polymorphism	Reference = ‘ATA’, Sample = ‘GTC’
MIXED	MNP and InDel	Reference = ‘ATA’, Sample = ‘GTCAGT’

http://pcingola.github.io/SnpEff/se_introduction/

Example: Glucosinolate biosynthesis in rapeseed

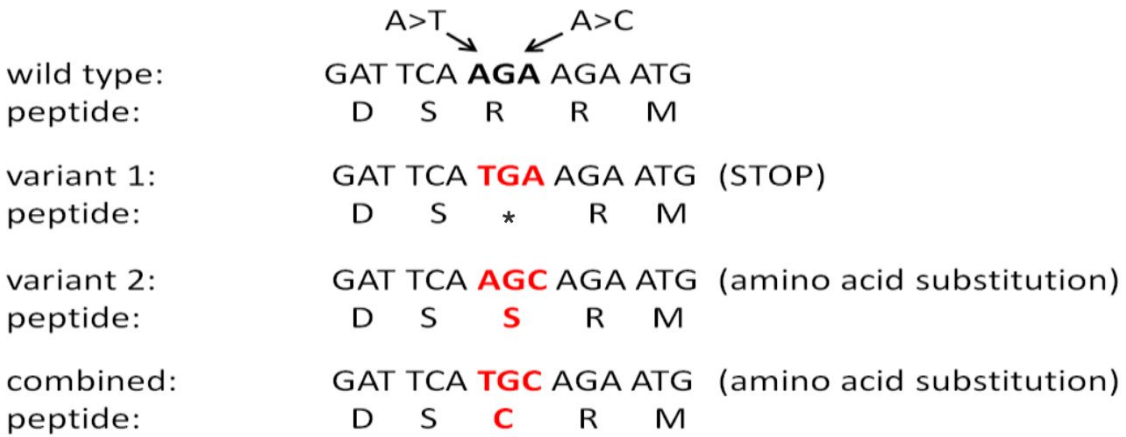


Schilbert et al., 2022: 10.1101/2022.06.01.494149

Is there something that SnpEff might miss?

Variant Annotation - NAVIP

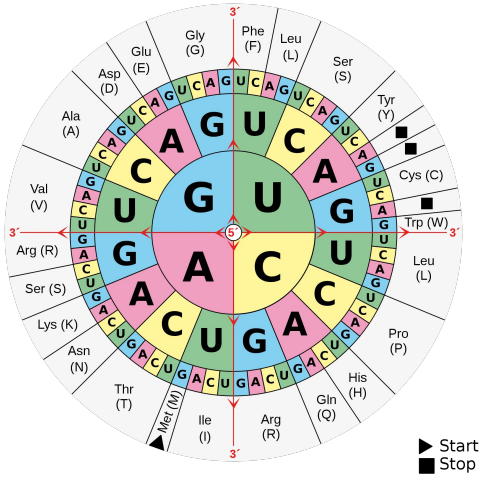
- NAVIP = Neighborhood-Aware Variant Impact Predictor
- Variants can have interacting effects e.g. compensate each other
- Considering multiple variants while predicting effects
- Tool freely available: https://github.com/bpucker/NAVIP

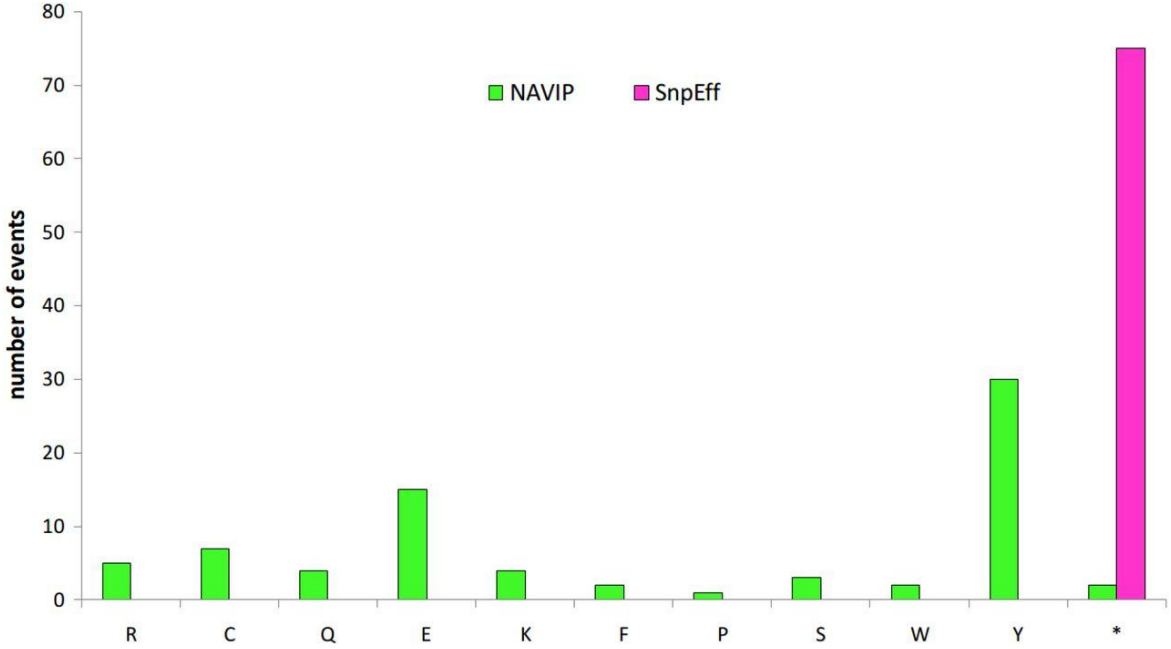


Baasner et al., 2019: 10.1101/596718

Connected SNVs

	A>T ↘ ↙ A>C	
wild type:	GAT TCA **AGA** AGA ATG	
peptide:	D S R R M	
variant 1:	GAT TCA **TGA** AGA ATG	(STOP)
peptide:	D S * R M	
variant 2:	GAT TCA **AGC** AGA ATG	(amino acid substitution)
peptide:	D S **S** R M	
combined:	GAT TCA **TGC** AGA ATG	(amino acid substitution)
peptide:	D S **C** R M	

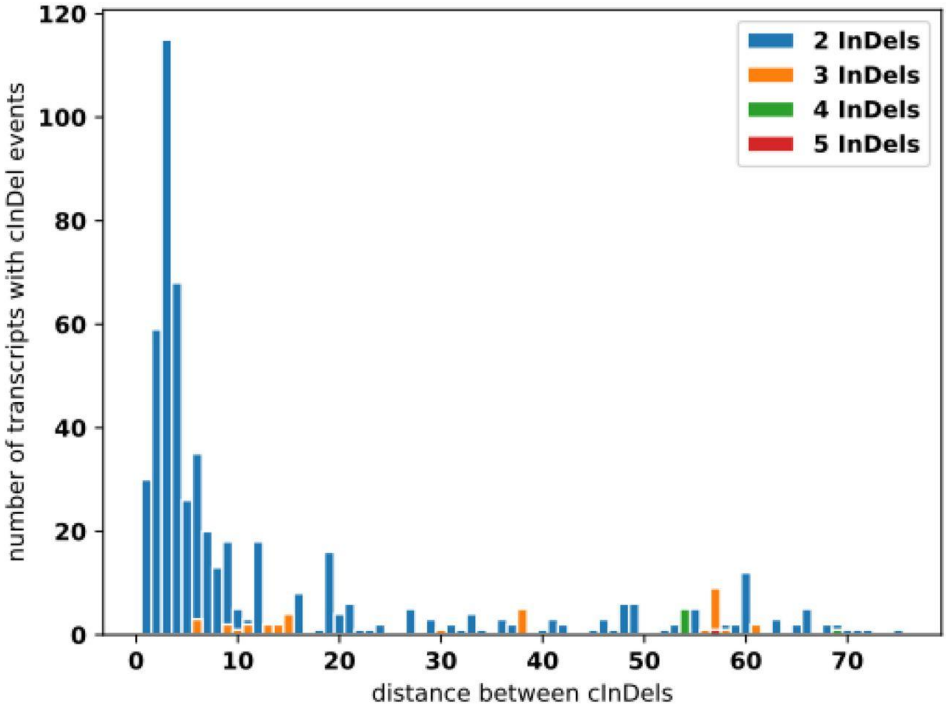




Baasner et al., 2019: 10.1101/596718

Compensating InDels

```
                  T>TA                C>CGC
wild type:        GTG TAT CTG CGC ATT
peptide:           V   Y   L   R   I

variant 1:        GTG TAT ACT GCG CAT T      (frameshift)
peptide:           V   Y   T   A   H

variant 2:        GTG TAT CTG CGC GCA TT     (frameshift)
peptide:           V   Y   L   R   A

combined:         GTG TAT ACT GCG CGC ATT
peptide:           V   Y   L   A   R   I
```



Baasner et al., 2019: 10.1101/596718

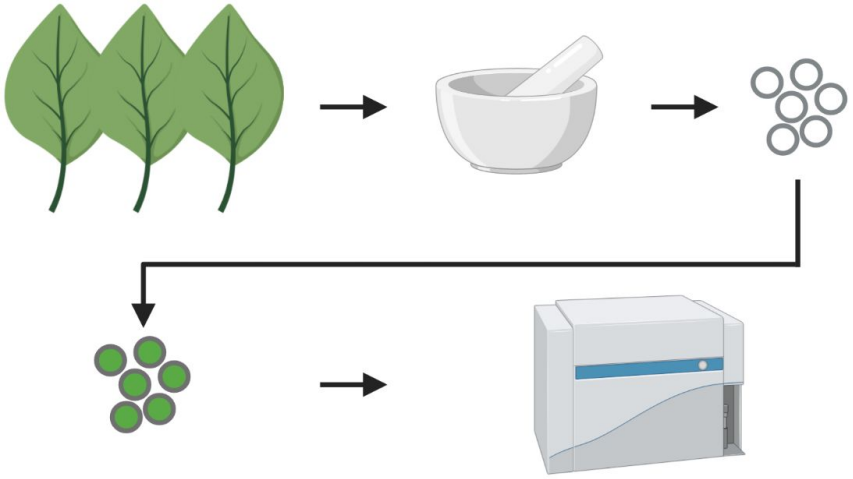
Comparative and evolutionary genomics

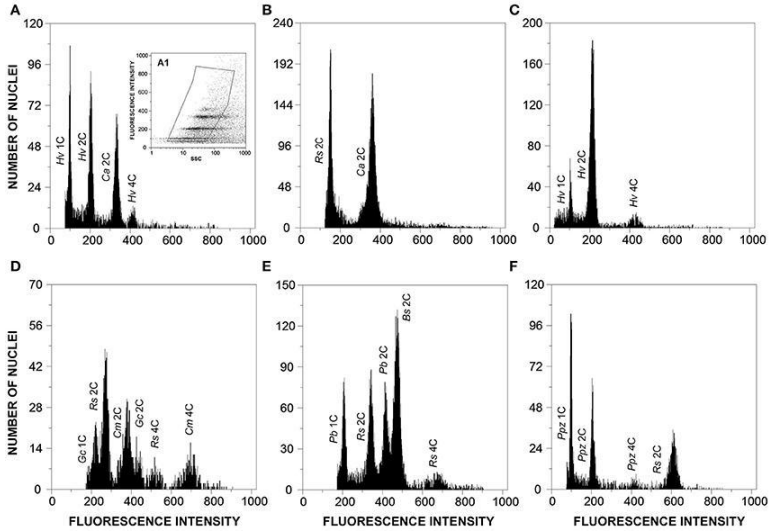
“Nothing in biology makes sense except in the light of evolution”
(who said this?)

- Compare genome sizes
- Compare chromosome numbers
- Compare gene sets
- Compare gene positions & chromosome structures

Genome size - 1

- Genome sequences of closely related species
- Flow cytometry is used to measure genome size biochemically
  - C-value
- Databases for plant genome sizes: https://cvalues.science.kew.org/



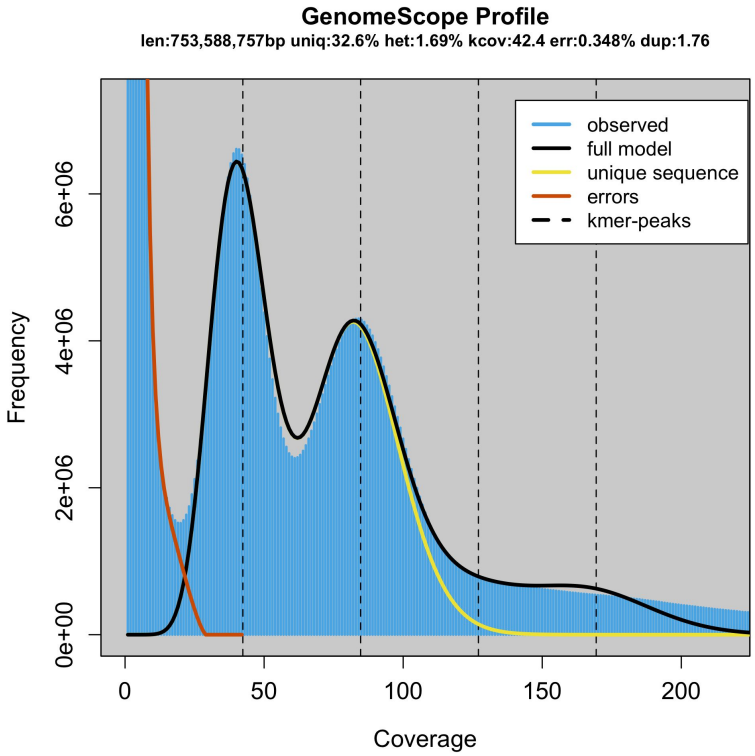


https://doi.org/10.3389/fpls.2014.00422

Genome size - 2

- Tools for genome size estimation based on short reads
  - K-mer-based: GenomeScope2, findGSE, gce

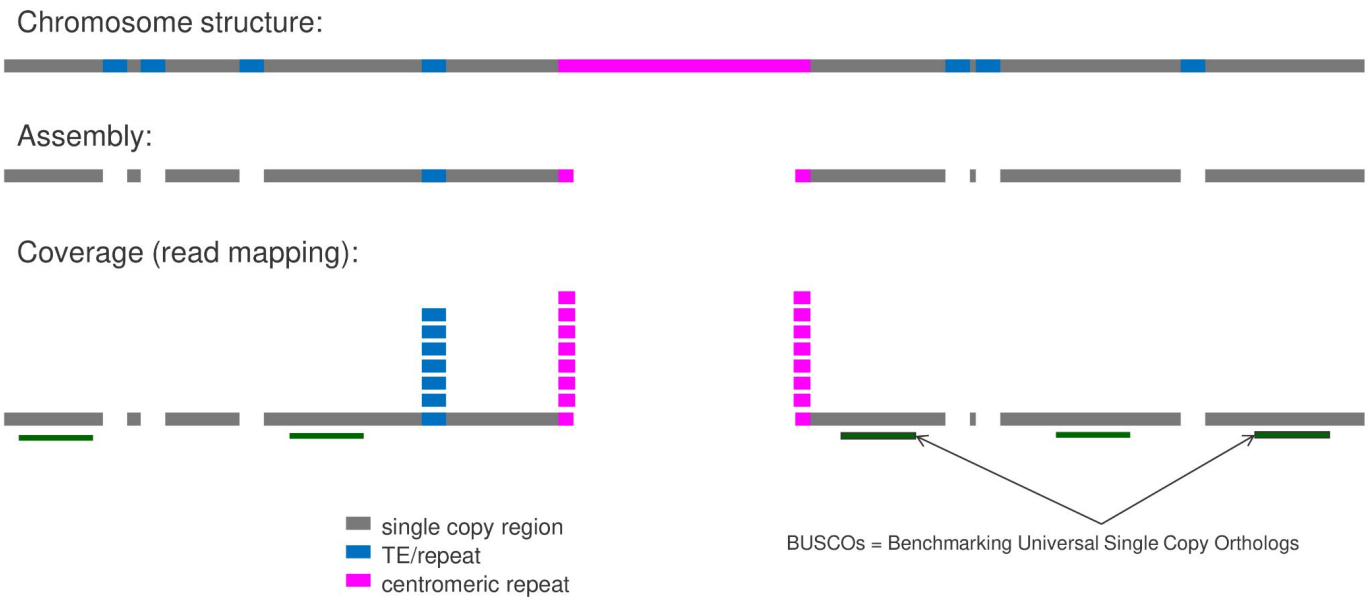
```
ACGAAGCCATAT

ACGAAGC

  CGAAGCC

    GAAGCCA

      AAGCCAT

        AGCCATAT
```



https://github.com/schatzlab/genomescope

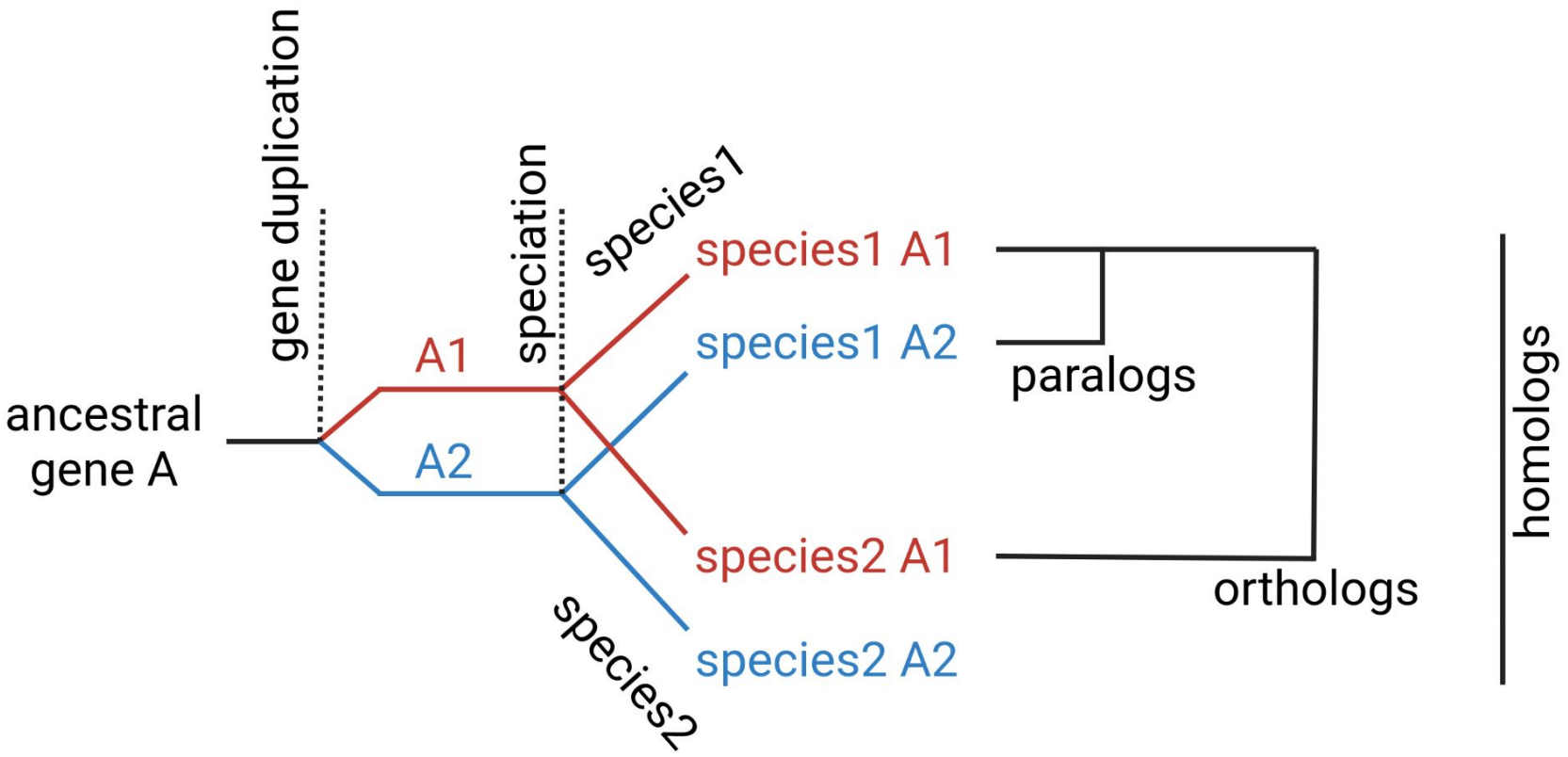
Genome size - 3

- Tools for genome size estimation based on short/long reads
  - Mapping-based: MGSE, Gnodes

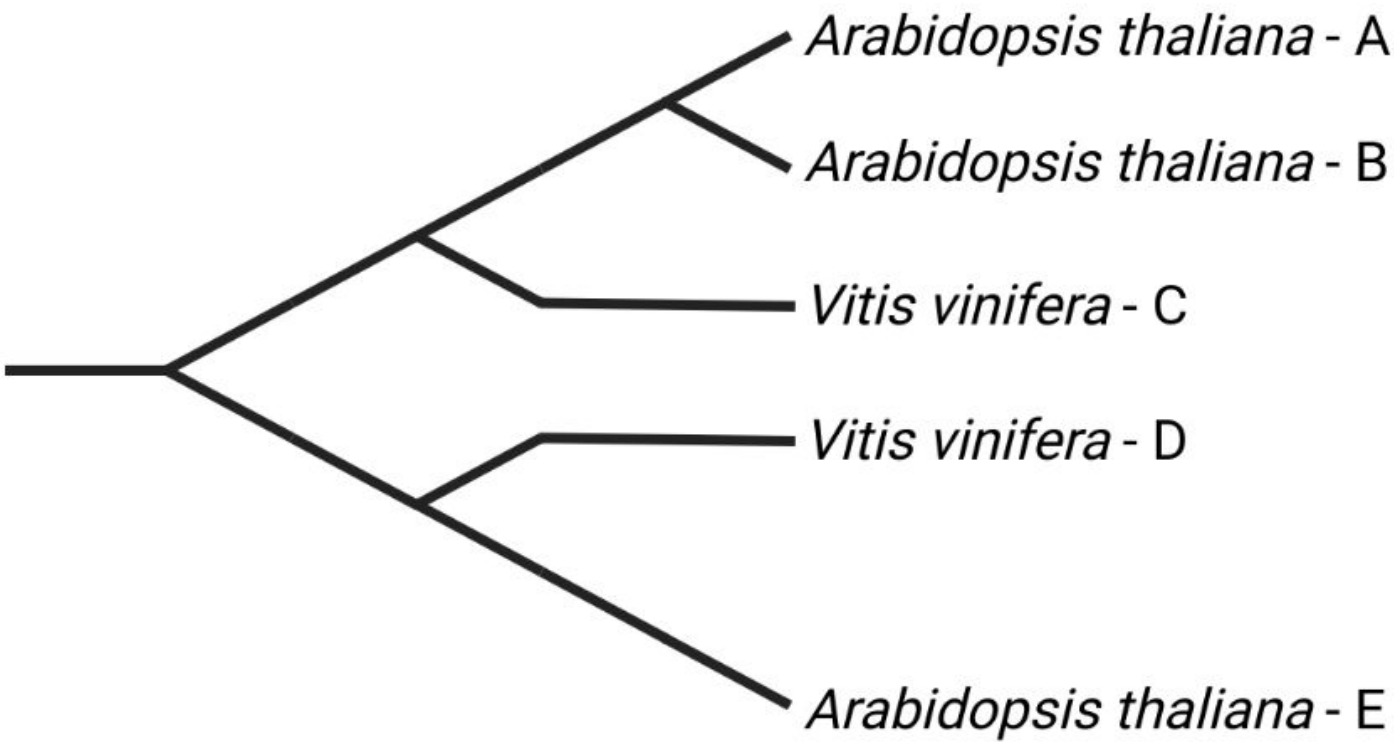


Pucker, 2019: 10.1101/607390

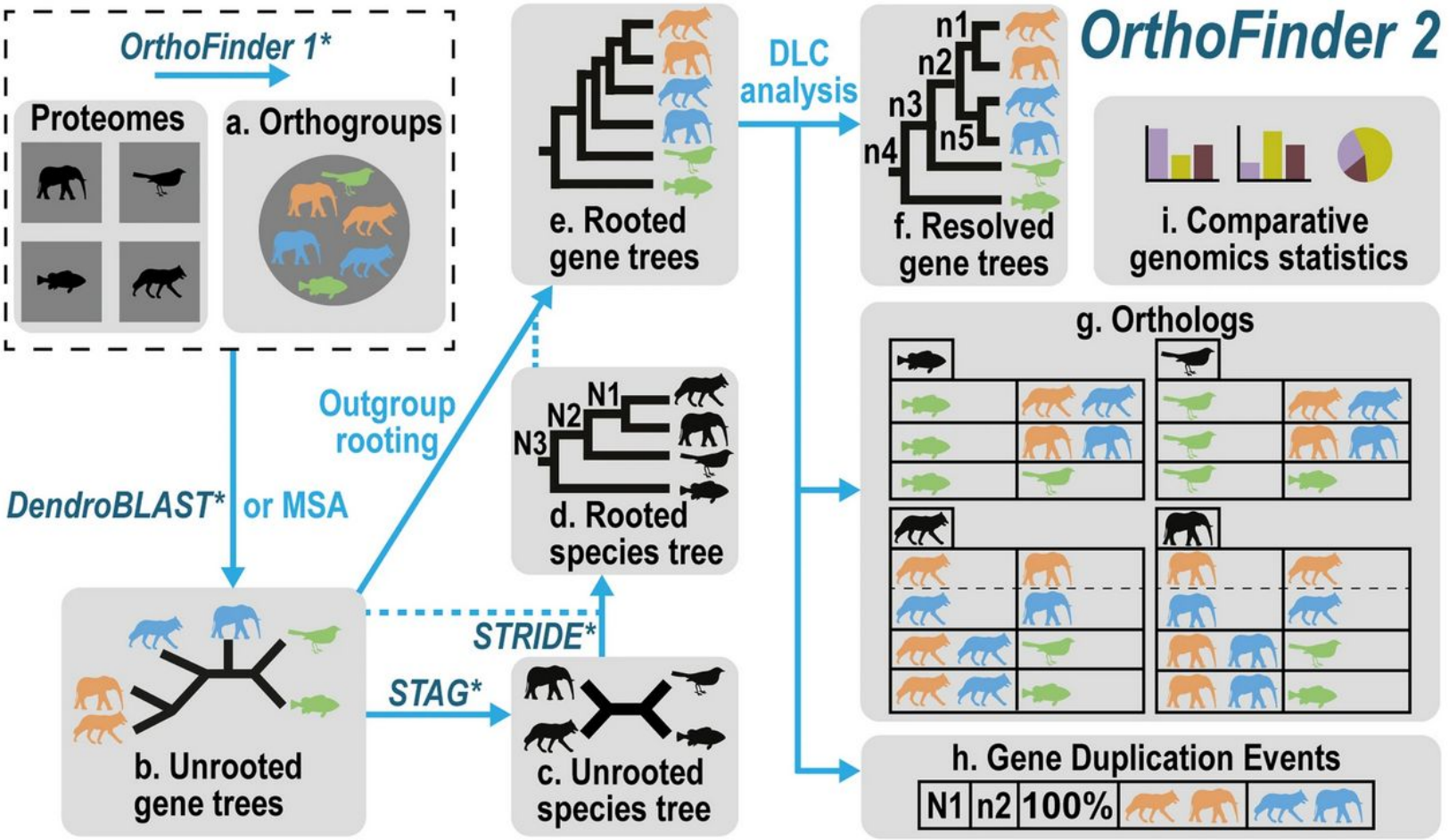
Paralogs & orthologs



Which are orthologs/paralogs?

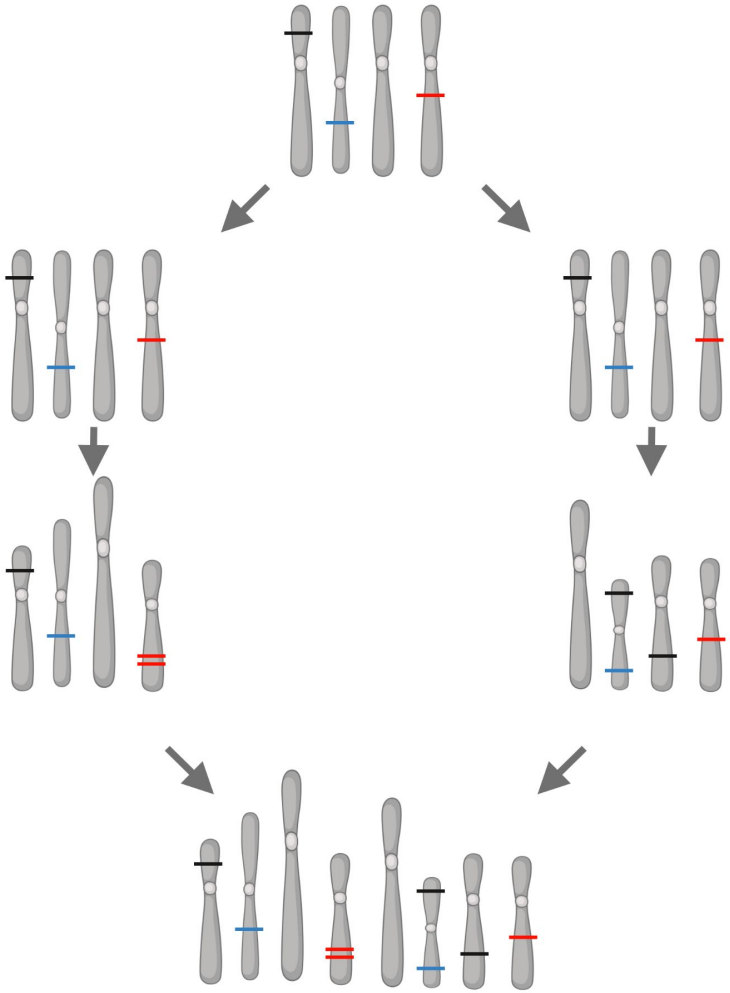


OrthoFinder2



Emms et al., 2019: 10.1186/s13059-019-1832-y

Homeologs



Same color indicates **orthologs**

Same color indicates **homeologs**

Speciation

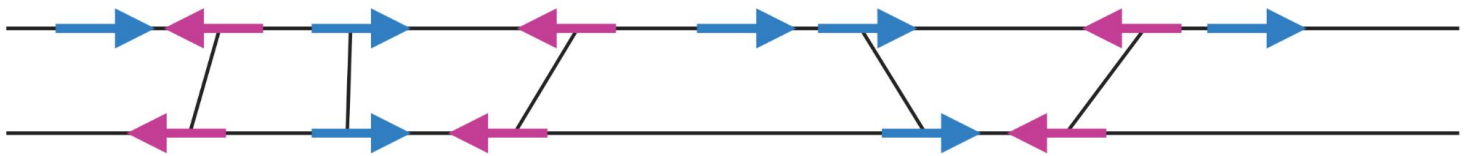
Gene duplications, translocations, and rearrangements

Hybridization leads to polyploidization

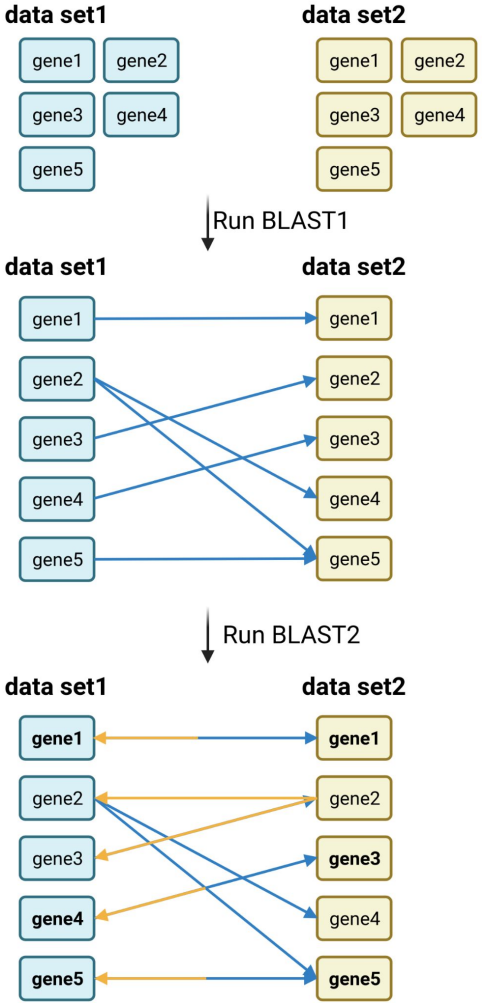
Glover et al., 2016: 10.1016/j.tplants.2016.02.005

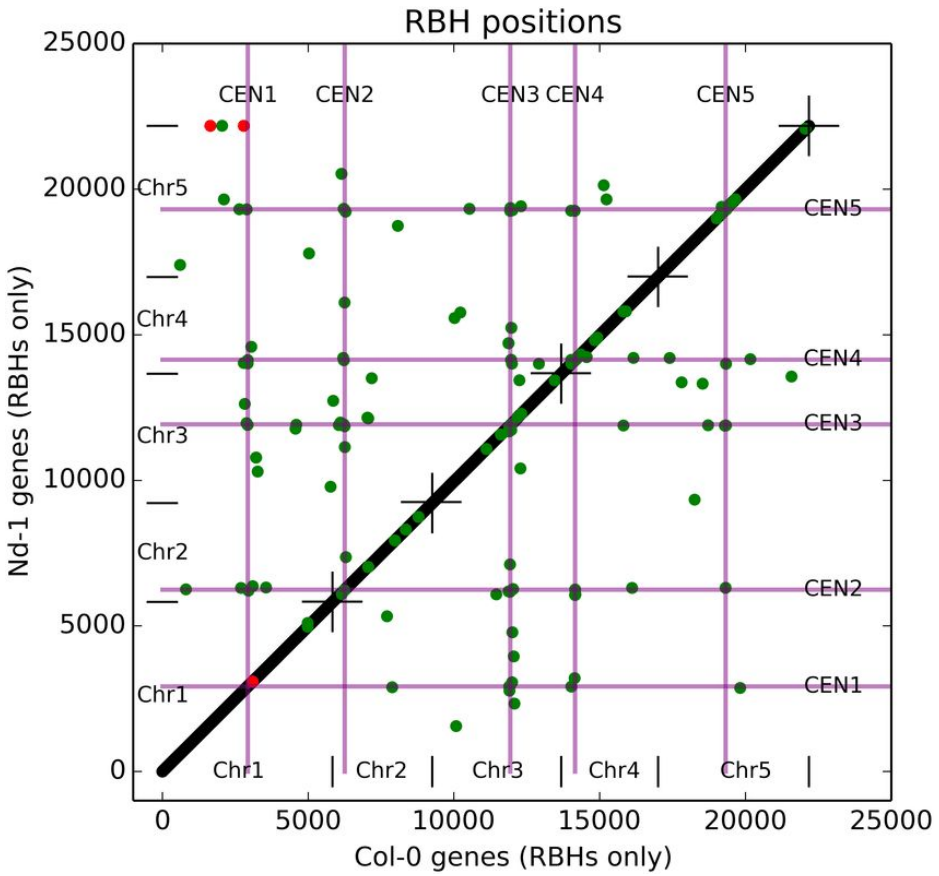
Synteny

Synteny = Same order of genes in different genomes



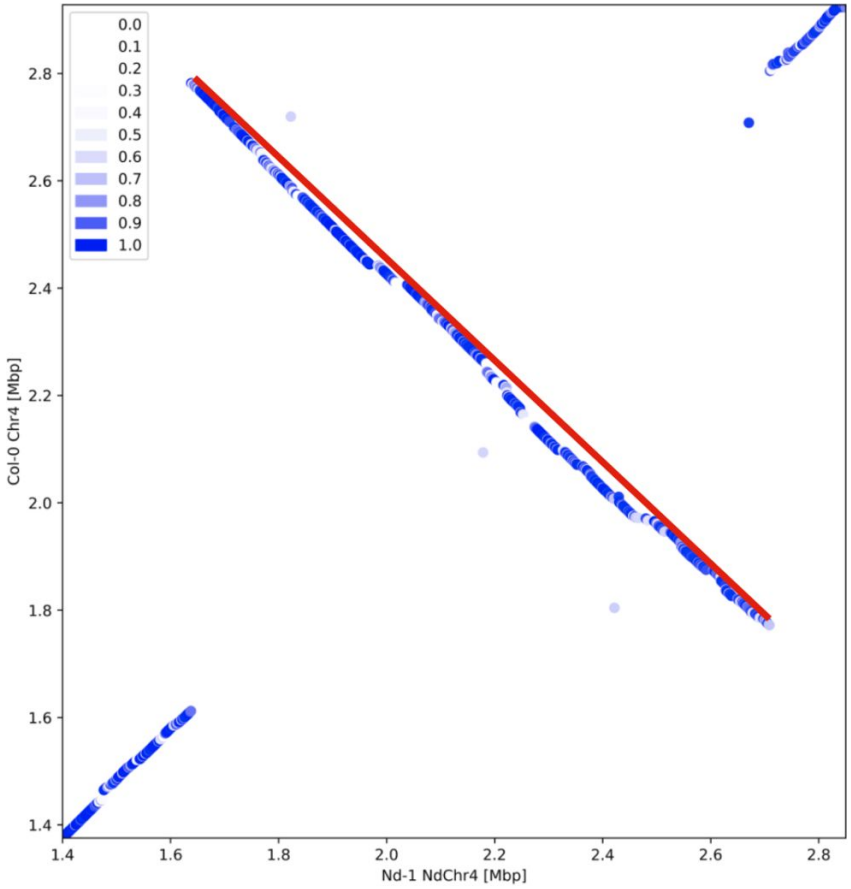
Reciprocal Best BLAST hits (RBHs)

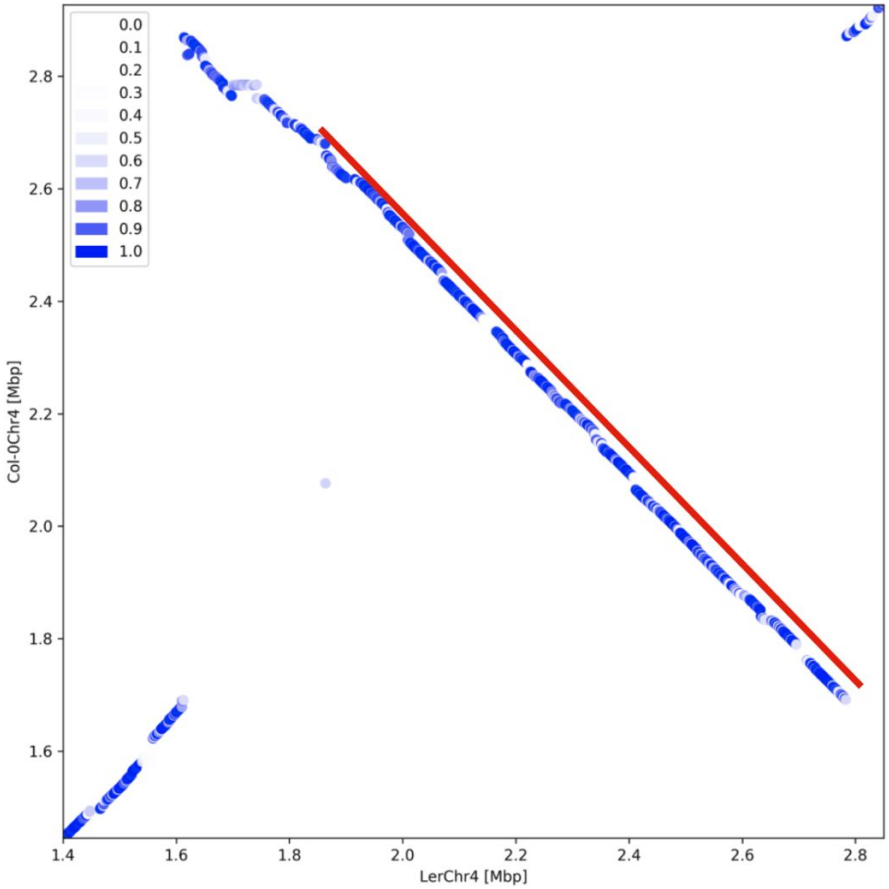




Pucker et al., 2016: 10.1371/journal.pone.0164321

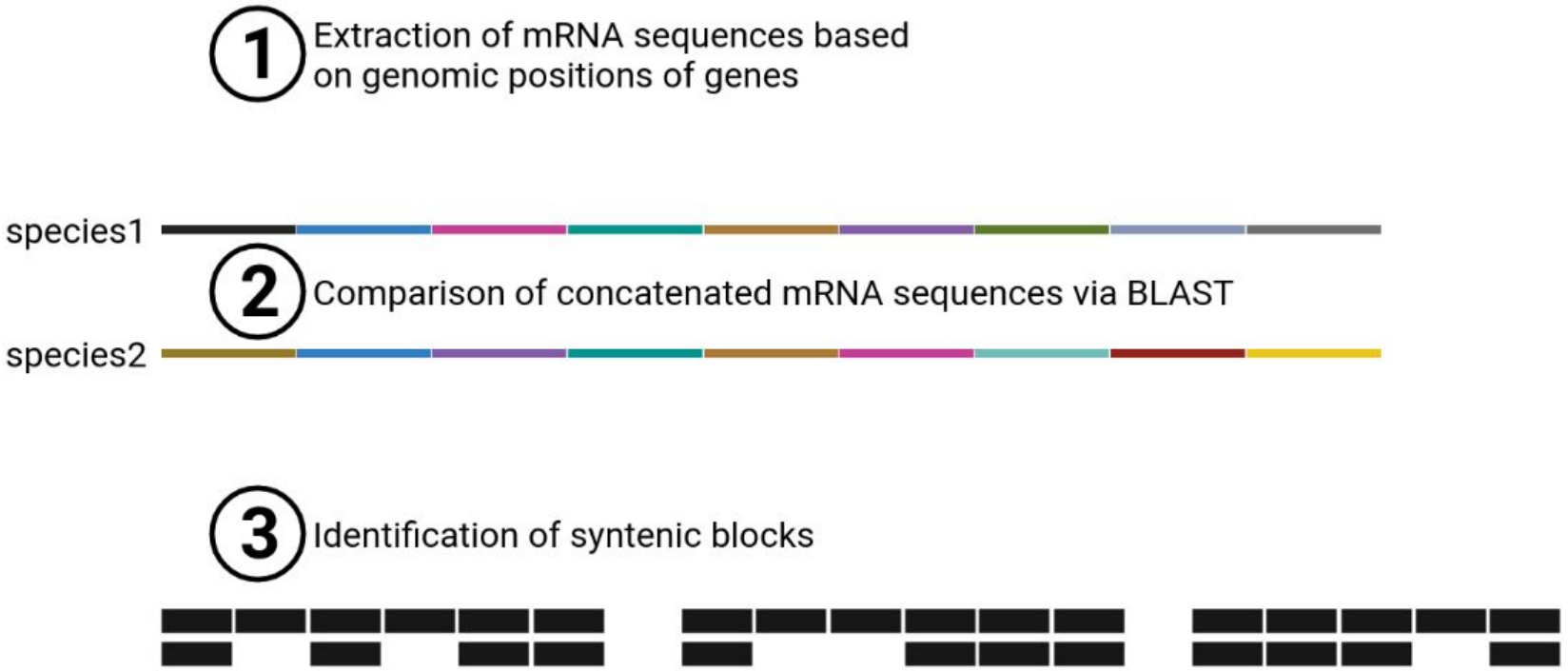
Dot plots





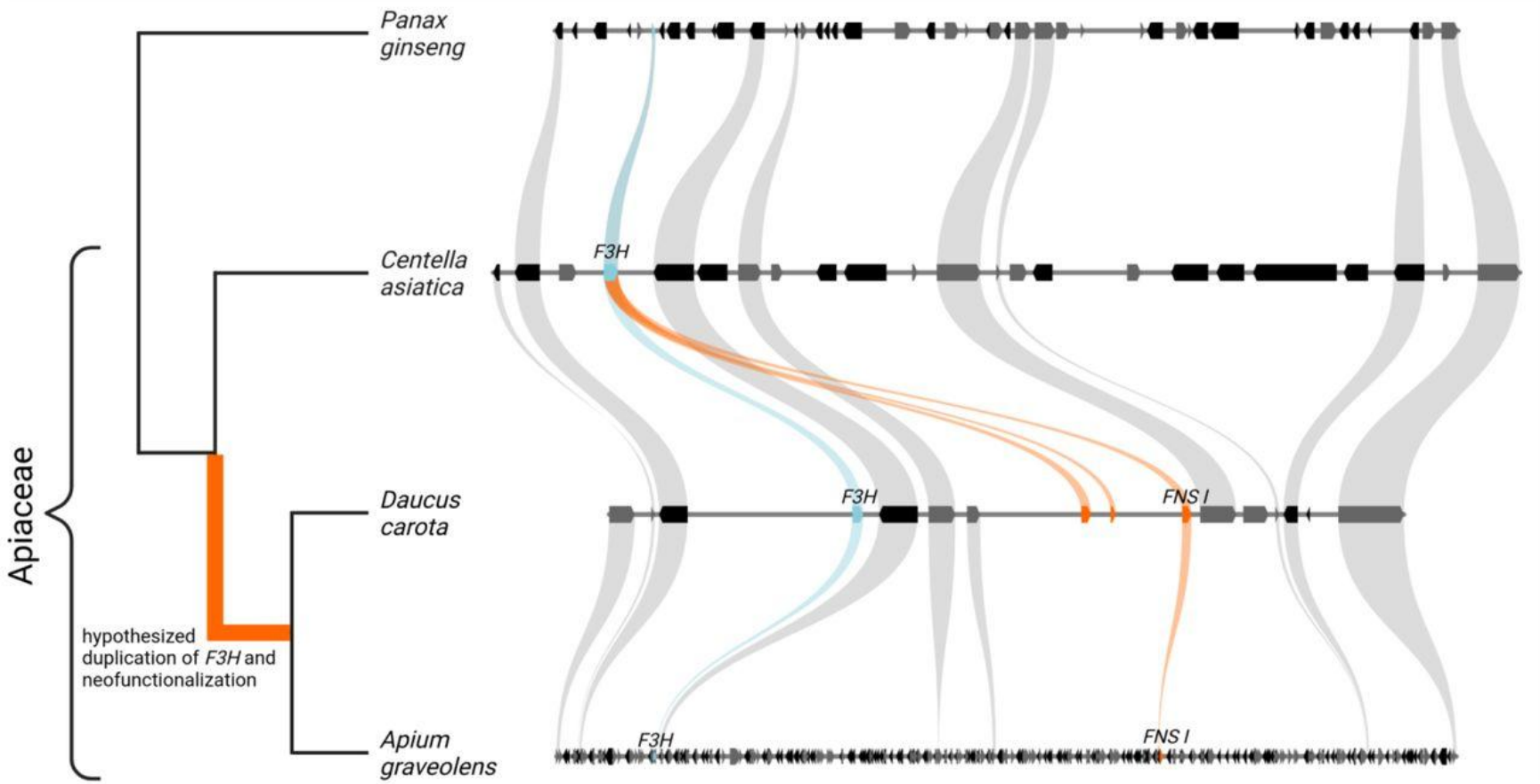
Pucker et al., 2019: 10.1371/journal.pone.0216233

JCVI/MCscan



https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)

Evolution of FNS I in the Apiaceae



Pucker & Iorizzo, 2022: 10.1101/2022.02.16.480750

Summary

- Read mapping & variant calling
- Mapping-by-sequencing
- Variant effect prediction
- Genome size estimation
- Phylogenetic relationships of genes
- Synteny

Time for questions!

Questions

1. What types of sequence variants can be detected?
2. How are sequence variants detected?
3. What can be inferred from read mapping-based coverage analyses?
4. What is a VCF file and what does it look like?
5. How does mapping-by-sequencing work?
6. What tools are available to predict the effects of sequence variants?
7. What is a paralog/ortholog?
8. What are homologs and homeologs?
9. What is synteny?