# Gene Function Prediction
# &
# Whole Genome Alignment.

**Qi Sun, Cheng Zou**

**Bioinformatics Facility**

**Cornell University**

# Given a protein, how to predict its function?

>unknown_protein

MEPSSETGMDPPLSQETFEDLWSLLPDPLQTVTCRLDNLSEFPDYPLAADMSVLQEGLMGNAVPTVTSCA
PSTDDYAGKYGLQLDFQQNGTAKSVTCTYSPELNKLFCQLAKTCPLLVRVESPPPRGSILRATAVYKKSE
HVAEVVKRCPHHERSVEPGEDAAPPSHLMRVEGNLQAYYMEDVNSGRHSVCVPYEGPQVGTECTTVLYNY
MCNSSCMGGMNRRPILTIITLETPQGLLLGRRCFEVRVCACPGRDRRTEEDNYTKKRGLKPSGKRELAHP
PSSEPPLPKKRLVVDDDEEIFTLRIKGRSRYEMIKKLNDALELQESLDQQKVTIKCRKCRDEIKPKKGKK
LLVKDEQPDSE

## BLAST

**mutant tumor protein p53 [Homo sapiens]**

**Sequence ID:** AYE20623.1 **Length:** 393 **Number of Matches:** 1

**Range 1: 3 to 393** GenPept Graphics ▼ Next Match ▲ Previou

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 345 bits(884) | 4e-116 | Compositional matrix adjust. | 203/399(51%) | 252/399(63%) | 47/399(11%) |

```
Query  2    EPSSETGMDPPLSQETFEDLWSLLPD-----PLQTVTCRLDNLSEFPD---YPLAADMSV  53
            EP S+  ++PPLSQETF DLW LLP+     PL +    +D+L   PD        D
Sbjct  3    EPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQA--MDDLMLSPDDIEQWFTEDPGP  60

Query  54   LQEGLMGNAVPT--------------------VTSCAPSTDDYAGKYGLQLDFQQNGTAK  93
             +   M  A P                     ++S  PS   Y G YG +L F  +GTAK
Sbjct  61   DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK  120

Query  94   SVTCTYSPELNKLFCQLAKTCPLLVRVESPPPRGSILRATAVYKKSEHVAEVVKRCPHHE  153
            SVTCTYSP LNK+FCQLAKTCP+ + V+S PP G+ +RA A+YK+S+H+ EVV+RCPHHE
Sbjct  121  SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE  180

Query  154  RSVEPGEDAAPPSHLMRVEGNLQAYYMEDVNSGRHSVCVPYEGPQVGTECTTVLYNYMCN  213
            R  +  +  APP HL+RVEGNL+  Y++D N+ RHSV VPYE P+VG++CTT+ YNYMCN
Sbjct  181  RCSD-SDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCN  239

Query  214  SSCMGGMNRRPILTIITLETPQGLLLGRRCFEVRVCACPGRDRRTEEDNYTKK----RGL  269
            SSCMGGMNRRPILTIITLE   G LLGR  FEVRVCACPGRDRRTEE+N+ KK      L
Sbjct  240  SSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENFRKKGEPHHEL  299

Query  270  KP-SGKRELAHPPSSEPPLPKKRLVVDDDEEIFTLRIKGRSRYEMIKKLNDALELQESLD  328
             P S KR L +  SS P   KK L    D E FTL+I+GR R+EM ++LN+ALEL+++
Sbjct  300  PPGSTKRALPNNTSSSPQPKKKPL----DGEYFTLQIRGRERFEMFRELNEALELKDA-Q  354

Query  329  QQKVTIKCRKCRDEIKPKKG------KKLLVKDEQPDSE  361
              K     R     +K KKG      KKL+ K E PDS+
Sbjct  355  AGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD  393
```

## HMM

| Domain | Description | Evalue |
|---|---|---|
| P53_TAD | P53 transactivation motif | 3.80E-10 |
| P53 | P53 DNA-binding domain | 2.70E-59 |
| P53_tetramer | P53 tetramerisation motif | 1.30E-17 |

# BLAST Databases

## NCBI

### Genbank

-user submitted

### Refseq

-NCBI curated

Commands to use:

https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete

## EBI

Uniref100
-all proteins

Uniref90
-collapse proteins 90% identical

Uniref50
-collapse proteins 50% identical

---

Ensembl
Plant Ensembl

For individual genome

## Others

**Institute specific**

JGI Genome Portal

**Species specific**

Flybase (fruit fly)
Wormbase (C. elegans)
SGD (yeast)
TAIR (Arabidopsis)

* Mostly deposited to NCBI or EBI

Using BioMart in Ensembl or Plant.Ensembl for gene ID conversion

# NCBI Honeybee Assembly Record

## Index of /genomes/all/GCF/003/254/395/GCF_

| Name | Last modified | Size |
|---|---|---|
| Parent Directory | | - |
| Annotation_comparison/ | 2019-08-07 20:38 | - |
| Evidence_alignments/ | 2019-08-07 20:38 | - |
| GCF_003254395.2_Amel_HAv3.1_assembly_structure/ | 2019-08-07 20:37 | - |
| Gnomon_models/ | 2019-08-07 20:38 | - |
| RefSeq_transcripts_alignments/ | 2019-08-07 20:38 | - |
| Apis_mellifera_AR104_annotation_report.xml | 2019-08-07 20:37 | 108K |
| GCF_003254395.2_Amel_HAv3.1_assembly_report.txt | 2019-08-07 20:37 | 20K |
| GCF_003254395.2_Amel_HAv3.1_assembly_stats.txt | 2019-08-07 20:37 | 18K |
| GCF_003254395.2_Amel_HAv3.1_cds_from_genomic.fna.gz | 2019-08-07 20:37 | 9.0M |
| GCF_003254395.2_Amel_HAv3.1_feature_count.txt.gz | 2019-08-07 20:37 | 476 |
| GCF_003254395.2_Amel_HAv3.1_feature_table.txt.gz | 2019-08-07 20:37 | 1.4M |
| GCF_003254395.2_Amel_HAv3.1_genomic.fna.gz | 2019-08-07 20:37 | 66M |
| GCF_003254395.2_Amel_HAv3.1_genomic.gbff.gz | 2019-08-07 20:38 | 95M |
| GCF_003254395.2_Amel_HAv3.1_genomic.gff.gz | 2019-08-07 20:38 | 6.4M |
| GCF_003254395.2_Amel_HAv3.1_genomic.gtf.gz | 2019-08-07 20:38 | 5.9M |
| GCF_003254395.2_Amel_HAv3.1_genomic_gaps.txt.gz | 2019-08-07 20:38 | 770 |
| GCF_003254395.2_Amel_HAv3.1_protein.faa.gz | 2019-08-07 20:38 | 6.1M |
| GCF_003254395.2_Amel_HAv3.1_protein.gpff.gz | 2019-08-07 20:38 | 14M |
| GCF_003254395.2_Amel_HAv3.1_pseudo_without_product.fna.gz | 2019-08-07 20:38 | 39K |
| GCF_003254395.2_Amel_HAv3.1_rm.out.gz | 2019-08-07 20:38 | 6.9M |
| GCF_003254395.2_Amel_HAv3.1_rm.run | 2019-08-07 20:38 | 888 |
| GCF_003254395.2_Amel_HAv3.1_rna.fna.gz | 2019-08-07 20:38 | 19M |
| GCF_003254395.2_Amel_HAv3.1_rna.gbff.gz | 2019-08-07 20:38 | 50M |
| GCF_003254395.2_Amel_HAv3.1_rna_from_genomic.fna.gz | 2019-08-07 20:38 | 17M |
| GCF_003254395.2_Amel_HAv3.1_translated_cds.faa.gz | 2019-08-07 20:38 | 5.6M |
| README.txt | 2020-09-02 16:26 | 43K |
| README_Apis_mellifera_annotation_release_104 | 2019-08-07 20:37 | 712 |
| annotation_hashes.txt | 2019-08-07 22:35 | 410 |
| assembly_status.txt | 2020-09-24 17:09 | 14 |
| md5checksums.txt | 2019-08-07 22:35 | 19K |

## Genome

*_genomic.fna.gz

## cDNA

*_cds_from_genomic.fna.gz

## Proteins

*_protein.faa.gz

*_translated_cds.faa.gz

* Protein sequences in *_translated_cds.faa.gz is better correlated with cDNA

## GFF and GTF

Coordinates file

# How to describe the function of a gene?

**Free text:** Each gene is associated with a sentence and/or a paragraph

| Gene ID | Gene description |
|---|---|
| GRMZM2G002950 | Putative leucine-rich repeat receptor-like protein kinase family |
| GRMZM2G006470 | Uncharacterized protein |
| GRMZM2G014376 | Shikimate dehydrogenase; Uncharacterized protein |
| GRMZM2G015238 | Prolyl endopeptidase |
| GRMZM2G022283 | Uncharacterized protein |

# How to describe the function of a gene?

**Gene Ontology:** A set of controlled vocabulary

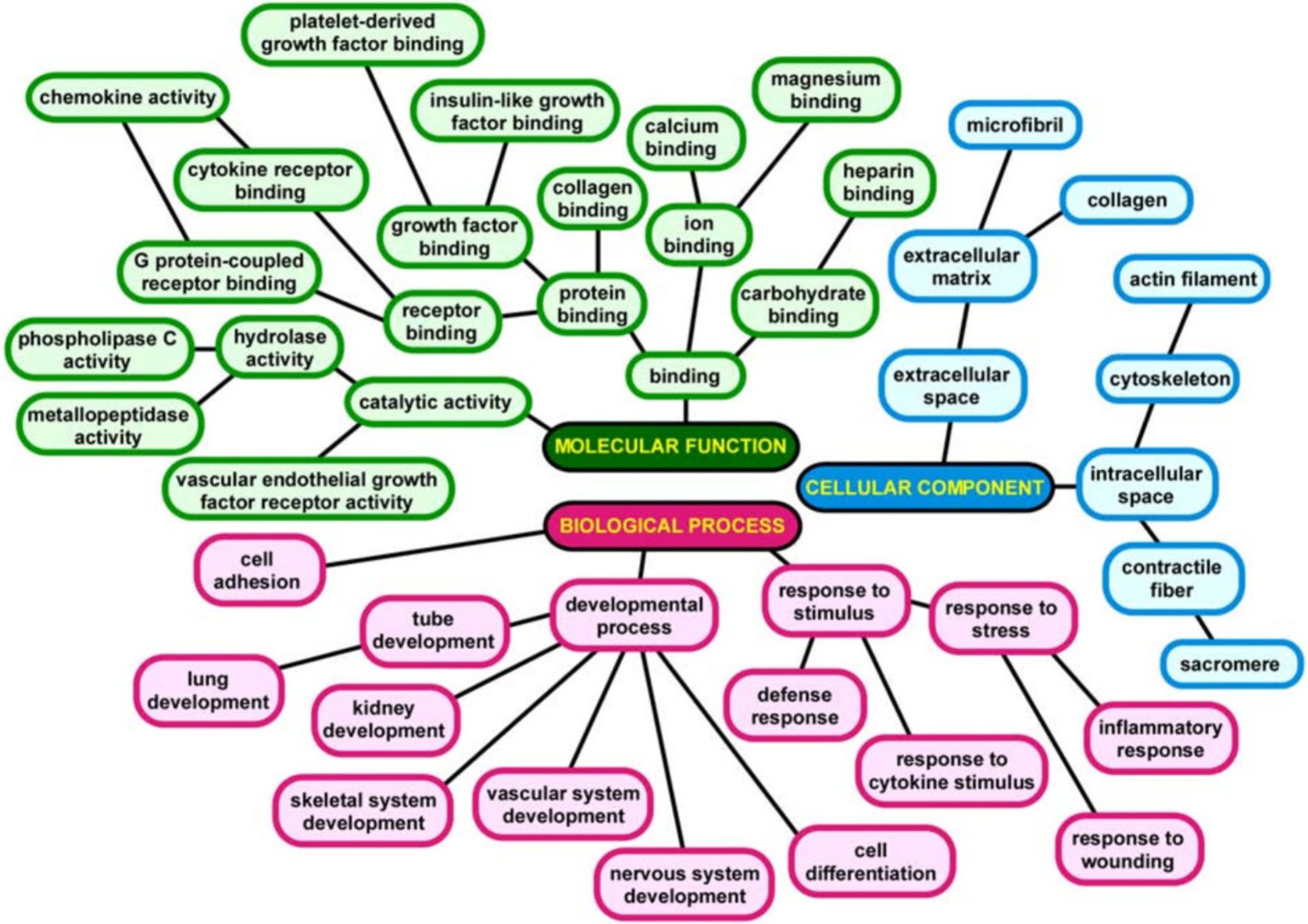
metabolic process

| Gene | GO |
|---|---|
| GRMZM5G888620 | GO:0003674 |
| GRMZM5G888620 | GO:0008150 |
| GRMZM5G888620 | GO:0008152 |
| GRMZM5G888620 | GO:0016757 |
| GRMZM5G888620 | GO:0016758 |
| GRMZM2G133073 | GO:0003674 |
| GRMZM2G133073 | GO:0016746 |

A gene can be associated with multiple GO

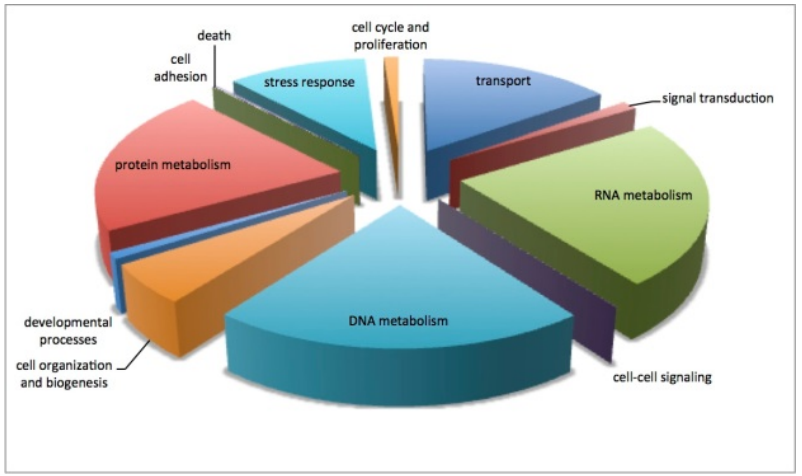
transferase activity, transferring acyl groups

# Hierarchical structure of the gene ontology terms

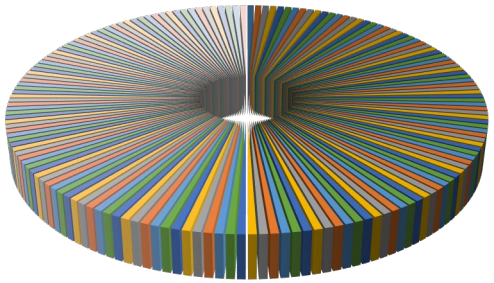


# GO SLIM How to make a pie chart?

Gene compositions in a species




A pie chart with too many categories



**GO category distribution**

## Collapse to GO SLIM


**To download premade GO Slim:**
http://geneontology.org/docs/download-ontology/#subsets
**Create your own GO Slim with OBO-Edit:**
http://oboedit.org/

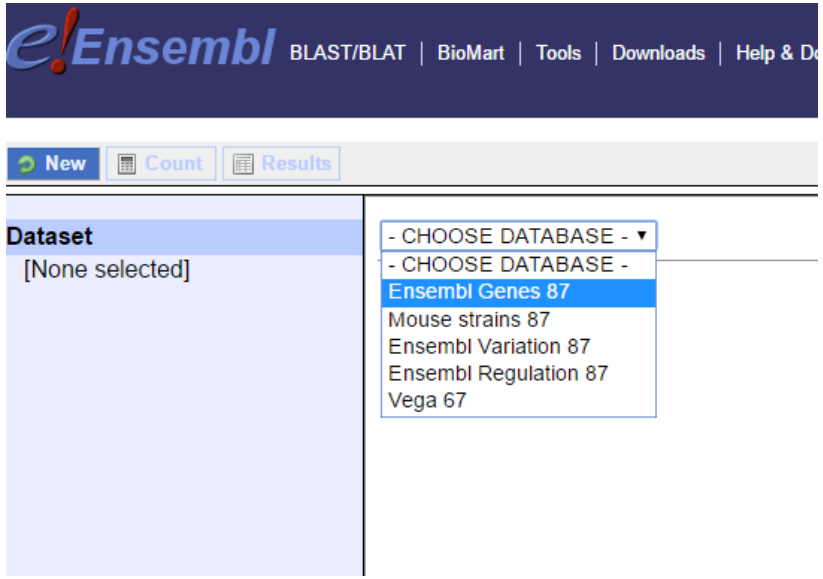
# How to attach Gene Ontology terms to each gene?

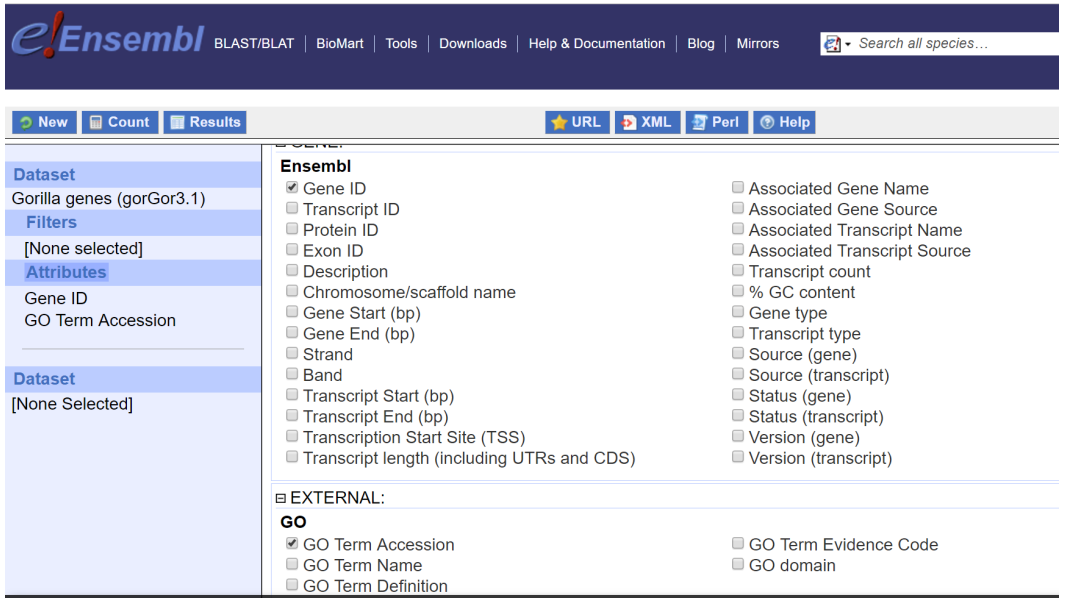
| | | | |
|---|---|---|---|
| GRMZM2G035341 | molecular_function | GO:0008270 | zinc ion binding |
| GRMZM2G035341 | molecular_function | GO:0046872 | metal ion binding |
| GRMZM2G035341 | cellular_component | GO:0005622 | intracellular |
| GRMZM2G035341 | cellular_component | GO:0019005 | SCF ubiquitin ligase complex |
| GRMZM2G035341 | biological_process | GO:0009733 | response to auxin |
| | | | |
| GRMZM2G047813 | molecular_function | GO:0003677 | DNA binding |
| GRMZM2G047813 | cellular_component | GO:0005634 | nucleus |
| GRMZM2G047813 | cellular_component | GO:0005694 | chromosome |
| GRMZM2G047813 | biological_process | GO:0006259 | DNA metabolic process |
| GRMZM2G047813 | biological_process | GO:0034641 | cellular nitrogen compound metabolic process |

# For model organisms, you can download GO annotation from Ensembl BioMart:

Animal genomes: http://www.ensembl.org

Plant genomes: http://plants.ensembl.org



# Non model organism

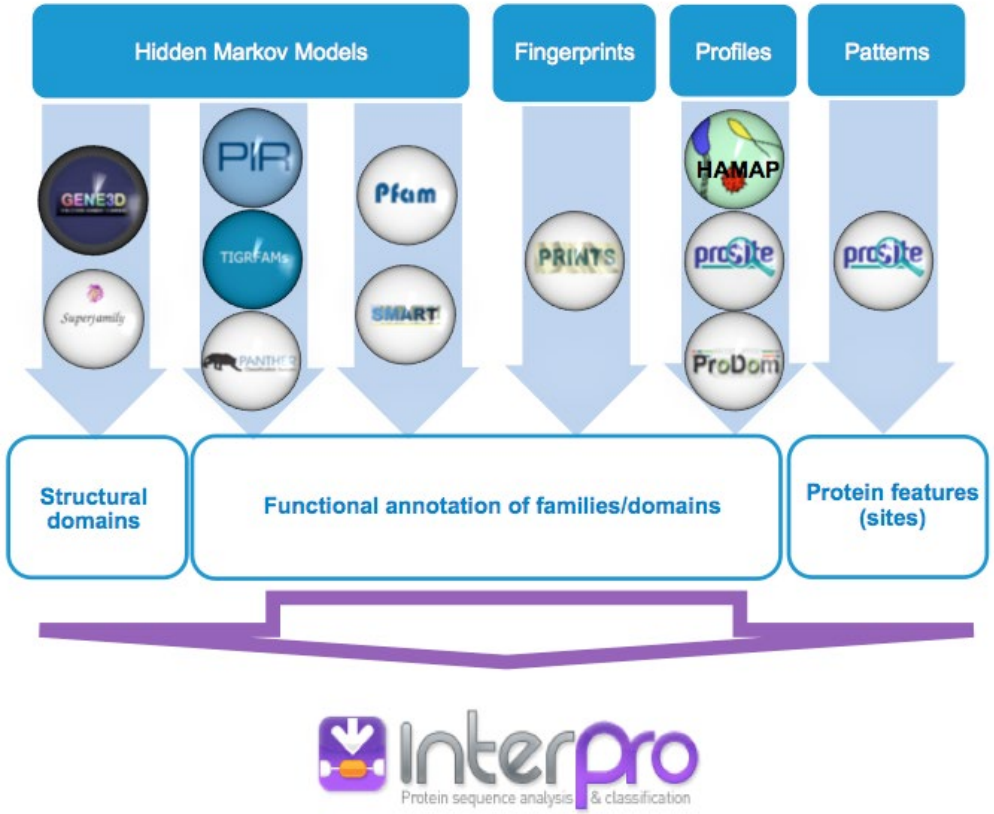
**InterProScan** (HMM + BLAST) *Free*

**BLAST2GO** (BLAST) *License required*

* BioHPC has a single-machine license on cbsumm10. Command line only, no GUI.

# InterProScan integrate many software results to produce gene annotation: Interpro ID & GO

To run InterproScan on BioHPC, follow the instructions on this page:
https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=87#c



# BLAST2GO: a fast annotation pipeline based on BLAST

Three steps (https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=87#c )

1. BLAST (Run Diamond on BioHPC gen2 machine)
2. InterproScan (optional, also on a gen2 machine)
3. GO assignment (Run on cbsumm10, require license)

**Diamond command:**

```
diamond blastp \
    --db uniref90 \
    --query mygenome.protein.fasta \
    --outfmt 5 \
    --max-target-seqs 100 \
    --max-hsps 1  \
    --evalue 1e-10 \
    -t ./ \
    --block-size 10
    --index-chunks 1
    -o blastresults.xml
```

Uniref90 as db

Output xml file

Directory to hold temporary files. "-t" is the same as "--tmpdir". "./" means current directory.

Tune performance and memory usage. The setting here is to maximize the performance.

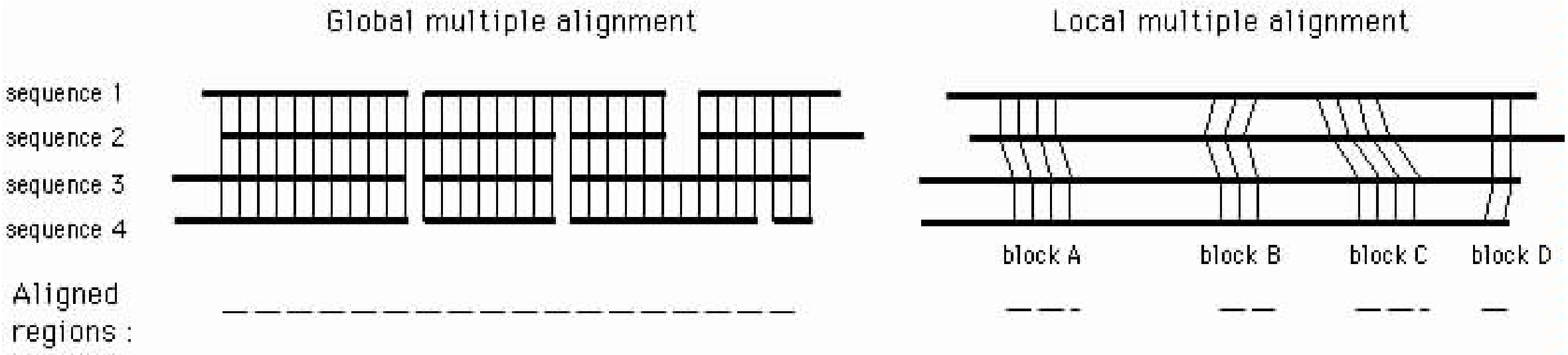
Output many target proteins in case some best hits do not have GO in BLAST2GO database

So that a target gene with multiple HSPs only show up once in result file.

More stringent e-value would lead to less but more reliable GO annotations.

# Whole Genome Alignment

# Genome alignment is based on local alignment




**end-to-end alignment**

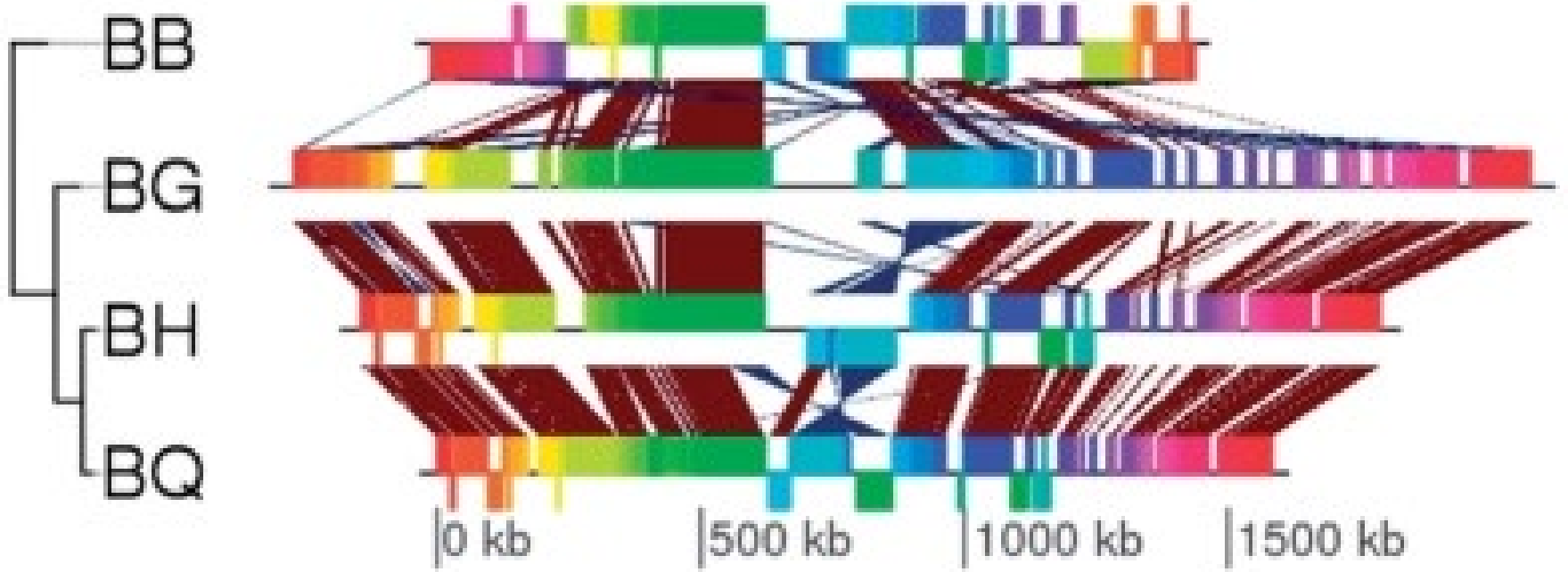
**find local region with highest similarity**

# Application of Genome Alignment

- Identify SV among the genome
- Chromosome rearrangement
- Syntenic  orthologue
- Identify core genome and dispensable genome
- Find evolutionary footprints
- Coordinates lifting between Genome version

# Three types of plots illustrating genome alignment

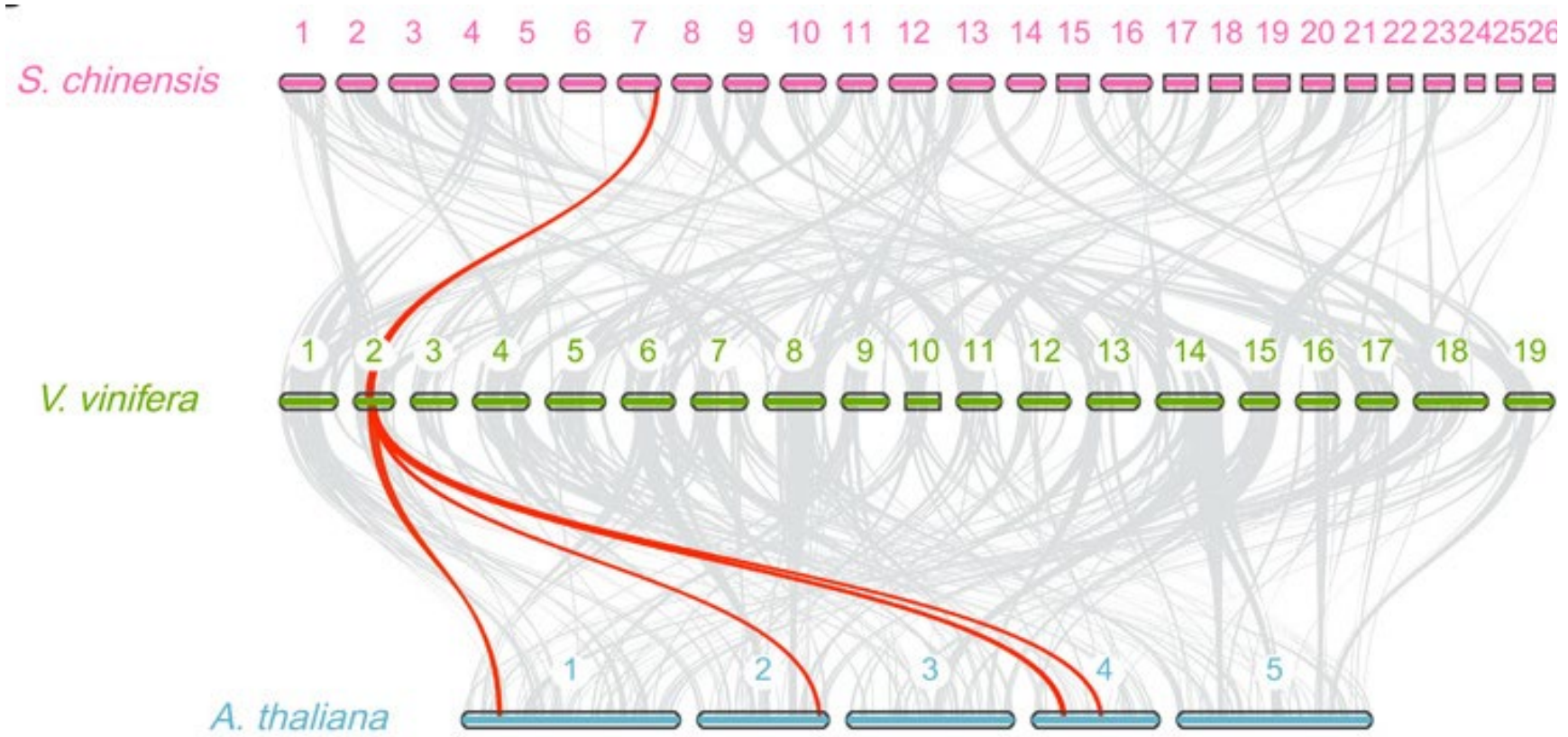
**1. linear blocks plot**




Guy L, Roat Kultima J, Andersson S G E. genoPlotR: comparative gene and genome visualization in R[J]. Bioinformatics, 2010, 26(18): 2334-2335.

# Three types of plots illustrating genome alignment

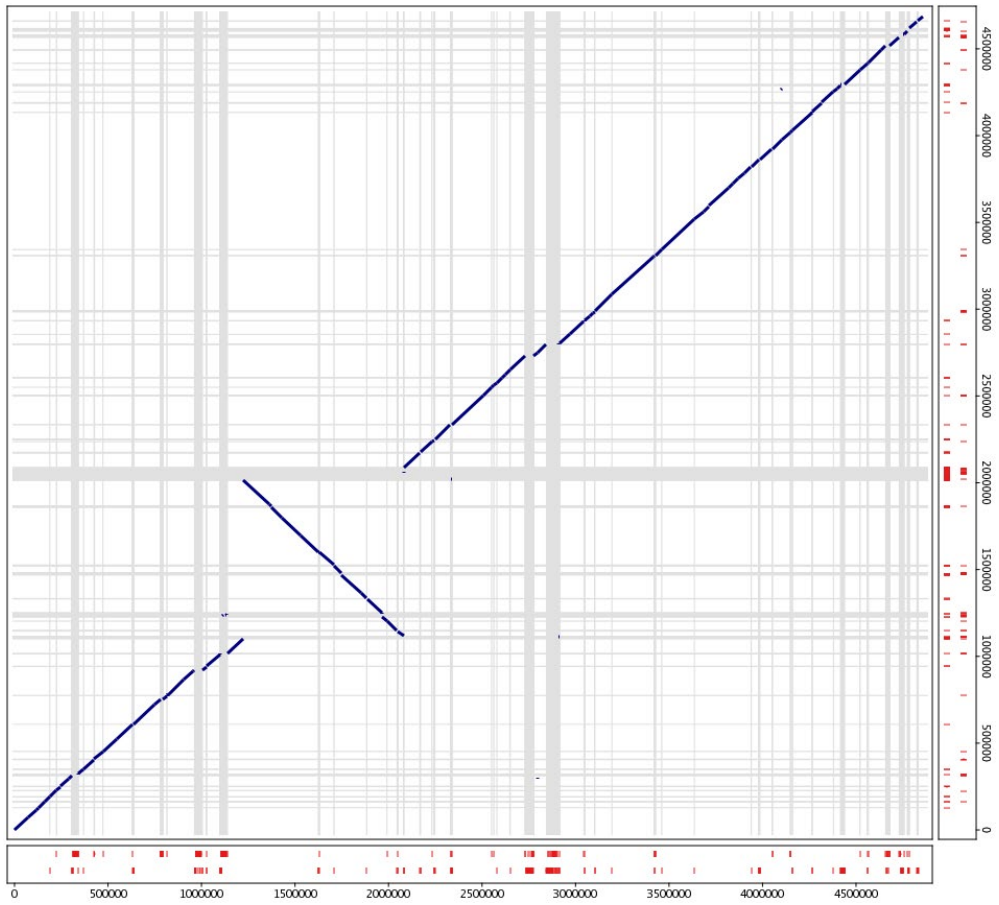
**1. linear blocks plot**




Sturtevant D, Lu S, Zhou Z W, et al. The genome of jojoba (Simmondsia chinensis): A taxonomically isolated species that directs wax ester accumulation in its seeds[J]. Science Advances, 2020, 6(11): eaay3240

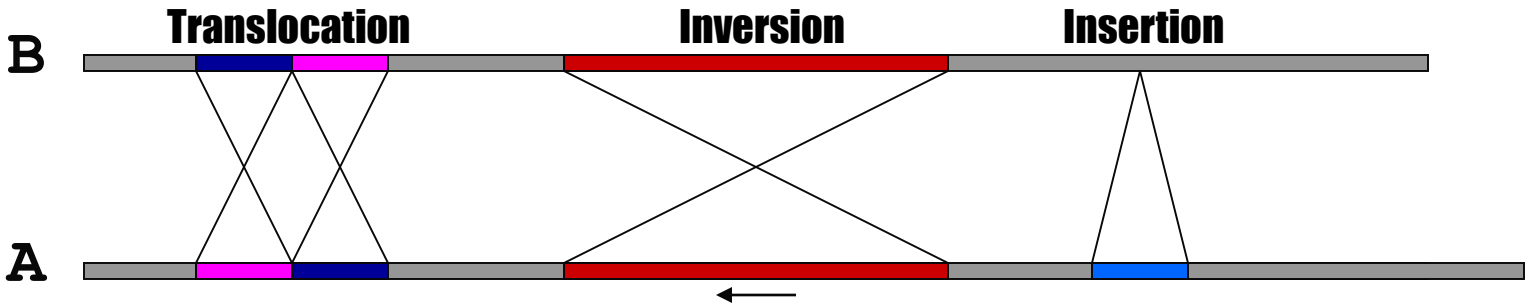
# Three types of plot illustrating genome alignment

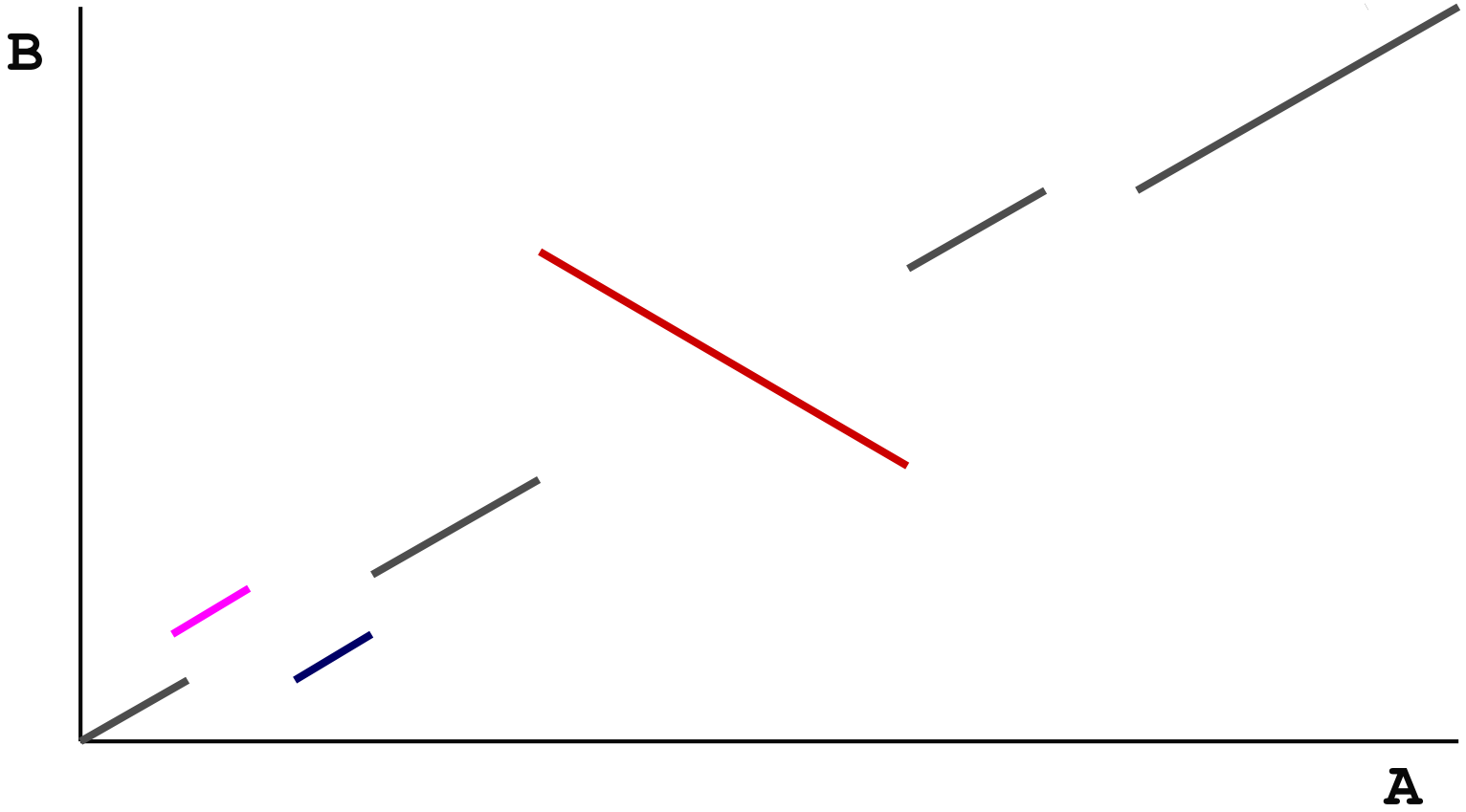
## 2. dotplot




With an alignment dot plot

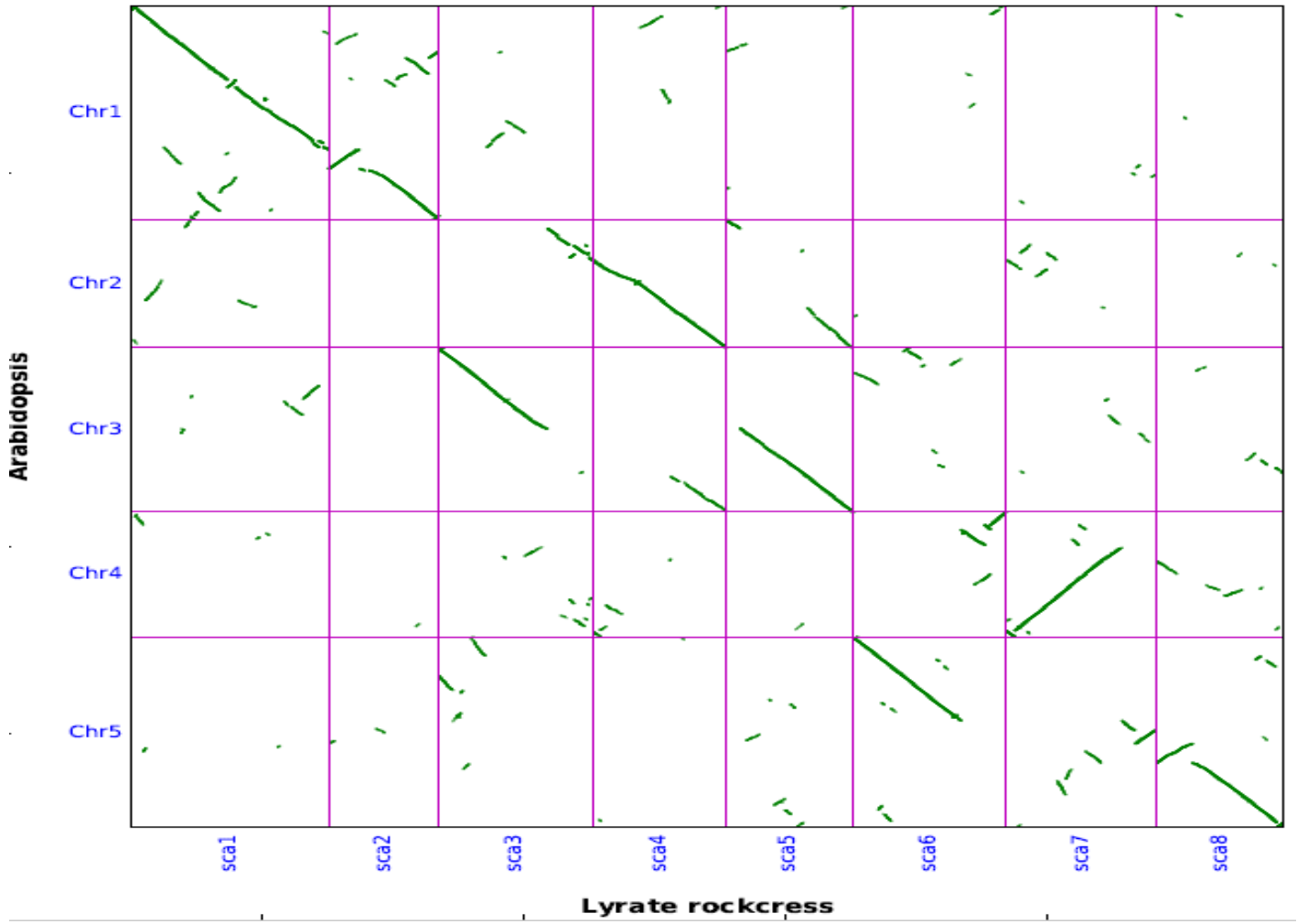
- *N* x *M* matrix
  - Let *i* = position in genome *A*
  - Let *j* = position in genome *B*
  - Fill cell *(i,j)* if $A_i$ shows similarity to $B_j$
- A perfect alignment between *A* and *B* would completely fill the positive diagonal



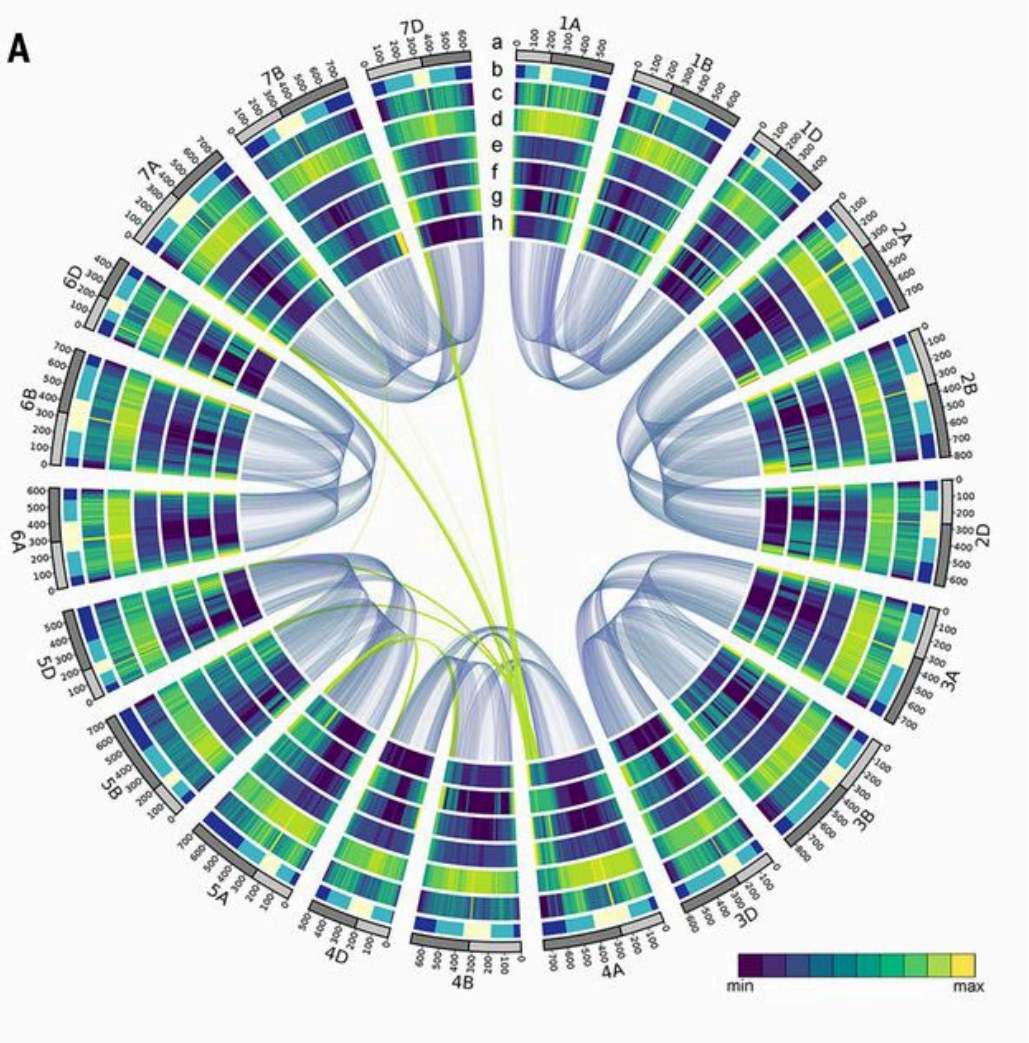

http://mummer.sourceforge.net/manual/AlignmentTypes.pdf

# Dotplot between Arabidopsis and Lyrata



# Three types of plot illustrating genome alignment

## 3. circos plot




The International Wheat Genome Sequencing Consortium (IWGSC) et al. Science 2018;361:eaar7191

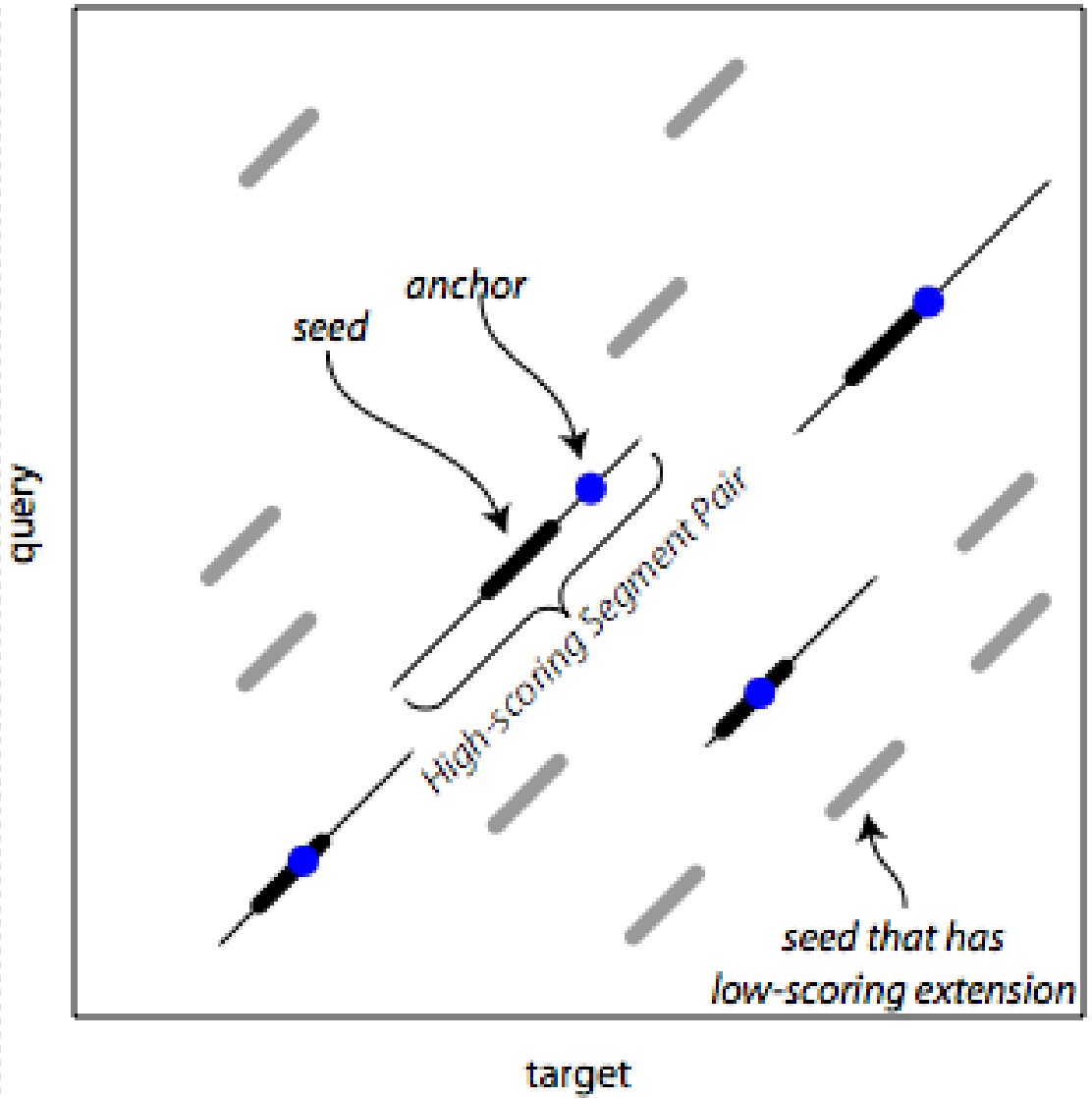
# How to build a genome alignment?

# HSP(high-scoring segment pairs), Chain, Net

- Starts with a local alignment

Seeds are short near-matches between the target and query sequences, where "short" typically means less than 20 bp

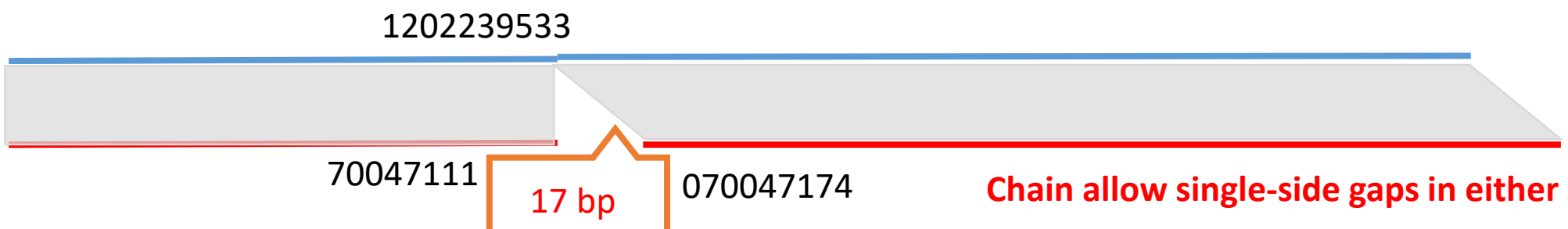
--seed=match12
--seed=12of19



# Chaining – joining the HSP(high-scoring segment pairs)

```
102239156 ggggtaggaaaggcccccaggagcacggggcctgagtcagtcagcaggcg 102239205
<<<<<<<<< |||||||||||||||||||| |||| |||| ||||||||||||||||   | <<<<<<<<<
070048013 ggggtaggaaaggcccccagaagcatgggggctgagtcagtcagcaaatg 070047964

102239206 cgtaagtcctgggcacatggg-gcaggttagggcacatctttcttctcag 102239254
<<<<<<<<< |||||| ||||||||||  || ||||||||||||||||    |||||||| <<<<<<<<<
070047963 cgtaaggcctgggcacagaggagcaggttagggcacattcaccttctcag 070047914

102239255 gattctgtggcttctcctttgggggagggagagaacatcttggagagcct 102239304
<<<<<<<<< |||||||||||| |  |  ||| | ||||||||| |||||||| |   |  <<<<<<<<<
070047913 gattctgtggctccctcactggagaagggagagagcatcttgggggtgca 070047864

102239305 agtccaaga---gctaaggcagag---agaggtt----agagcccctg-t 102239343
<<<<<<<<< | |||||||   || | |||||||   |||| |     |||||||||    <<<<<<<<<
070047863 attccaagaacagcgagggcagaggttagagatggctgagagcccctaac 070047814

102239344 cttaggaggcatcacaacaggcagcaacaactttg--gaaagctggatga 102239391
<<<<<<<<< || |||||||| ||||| |||||||||||||| ||  ||||||| ||||| <<<<<<<<<
070047813 ctgaggaggcaccacaataggcagcaacaactgtgtggaaagctagatga 070047764

102239392 actggtcagtagcagggaatggaggcgagcaccgggttggcctcttagg- 102239440
<<<<<<<<< |||||||||||||  |  |||||  | | |||| || |||||||||| ||  <<<<<<<<<
070047763 actggtcagtagcggaaaatgggaggggcactgg-ttggcctcttgggg 070047715

102239441 -------taaacctggcttagttgaactcatg-gaatacttggtattccc 102239482
<<<<<<<<<         || | |||||| | ||| |||||| ||||||  ||| ||||| <<<<<<<<<
070047714 aggggtccaaccttggcttggatgagctcatgagaatacccggtgttccc 070047665

102239483 aagcagagtggggtggggcccaaagccccttccctgtgtacctccttaa 102239532
<<<<<<<<< |||||||| ||||  | |||||||||||||||| |||    ||||||||| <<<<<<<<<
070047664 aagcagagcggggcagagcccaaagccccttctctgcaggcctccttaa 070047615

102239533 g 102239533
<<<<<<<<< | <<<<<<<<<
070047614 g 070047614
```

```
102239534 gaataaaaggcattcagggagtttccaggcaaggggtgccagaattagtc 102239583
<<<<<<<<< ||  ||||||||||||||||   || || |||||||||||||||||||| <<<<<<<<<
070047597 gagaaaaaggcattcagggagccccctggaaaggggtgccagaattagtc 070047548

102239584 cttaaggcacagctgggggcagacaaggcgccaaggcacaattggtgggg 102239633
<<<<<<<<< |||||||||||||||| | |||||||| |||||||||| |||||||| <<<<<<<<<
070047547 cttaaggcacagctgggg-cggacaaggcaccaaggcacaactggtgggg 070047499

102239634 gggcaaagggatagcctccaagctgagtgccagggtcacaagaggatgca 102239683
<<<<<<<<< | |   ||||  |||||||||| |||||||||||||||||||| ||| <<<<<<<<<
070047498 gagggcaggggcagcctccaagccgagtgccagggtcacaagaggacgca 070047449

102239684 ggaccgcccacgctttatcggtgttgggttaagcaccgcccggacagcct 102239733
<<<<<<<<< |||||| || ||||| ||||| || ||||| ||||| |||||||||||| <<<<<<<<<
070047448 ggaccgtccccgcttgatcggagtggggttgagcacagcccggacagcct 070047399

102239734 ccgcaaacacctccttgacgccgtcttgttgcagtgctgagcactcgagg 102239783
<<<<<<<<< | || ||||| |||||||| || || |||||||| |||||||| |||||| <<<<<<<<<
070047398 cggcgaacacttccttgacaccatcctgttgcagggctgagcattcgagg 070047349

102239784 tagcgcacagcatggatctgcttggccagtgcctggccctgctgtggagt 102239833
<<<<<<<<< |||||||||||| |||||||||||||||||||||||||||||||| || || <<<<<<<<<
070047348 tagcgcacagcgtggatctgcttggccagtgcctggccctgctgcggtgt 070047299

102239834 gatgggcgcttgaccctgctccttgaggcgccgtagggtatcaggctggg 102239883
<<<<<<<<< ||||||||| || |||||||||||||||||||||||||| | |||||||| <<<<<<<<<
070047298 gatgggcgcctggccctgctccttgaggcgccgtagggtgtaaggctggg 070047249

102239884 ctctcaggtccttcttggtacccaccaggaggataggcacgtcagggcag 102239933
<<<<<<<<< |||||||||||||||||| |||||||| ||||| || || ||||||||| <<<<<<<<<
070047248 ctctcaggtccttcttggtgcccaccagcaggatgggtacatcagggcag 070047199

102239934 tggtgacaaacctctgggtgccatttgtgcctcacgttctcataggaggg 102239983
<<<<<<<<< ||||| || |||||||| ||||| ||||| | |||||||||||||| || <<<<<<<<<
070047198 tggtggcacacctctggatgccacttgtgtcgcacgttctcataggaagg 070047149
```

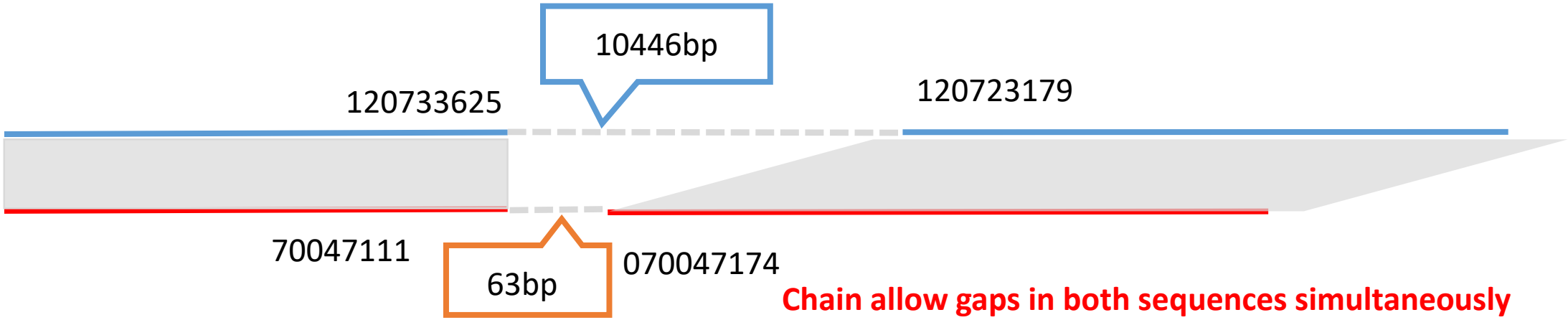



**Chain allow single-side gaps in either sequence**

# Chaining the HSP(high-scoring segment pairs)

```
120722510 ttcgacaactactcagccaacgtgatggtggatgggaagccagttaacct 120722559
>>>>>>>>> |||||||| |||   ||  |      || || |||    |||| ||||| >>>>>>>>>
070046996 ttcgacaattacagcgcgcagagcgcagttgacgggcgcacagtgaacct 070047045

120722560 ggggctgtgggacactgctggccaggaagactacgataggcttcggccac 120722609
>>>>>>>>> |   |||||||||||||| |||||||| || || ||  | || ||  ||| >>>>>>>>>
070047046 gaacctgtgggacactgcgggccaggaggagtatgaccgcctccgcacac 070047095

120722610 tctcctatcctcaaac 120722625
>>>>>>>>> ||||||| ||||| || >>>>>>>>>
070047096 tctcctaccctcagac 070047111
```

```
120723179 agtggtacccagaggtgcggcaccactgcccacacacacccatccttctg 120723228
>>>>>>>>> ||||| | ||||||||| | |||||||||||  |   ||||||||| ||| >>>>>>>>>
070047174 agtggcatccagaggtgtgccaccactgccctgatgtacccatcctgctg 070047223

120723229 gtgggcaccaagctggacctccgtgatgacaaggatacgattgaacggct 120723278
>>>>>>>>> ||||||||||||  ||||||  | |   |     | ||  |    || || >>>>>>>>>
070047224 gtgggcaccaagaaggacctgagagcccagccttacaccctacggcgcct 070047273

120723279 gcgggacaagaagctggcacccataacctacccccaaggcctggccatgg 120723328
>>>>>>>>>    |||  ||   | ||| ||||| ||    |  || |||| |||  ||| >>>>>>>>>
070047274 caaggagcagggccaggcgcccatcacaccgcagcagggccaggcactgg 070047323

120723329 cc 120723330
>>>>>>>>> || >>>>>>>>>
070047324 cc 070047325
```




**Chain allow gaps in both sequences simultaneously**

# Chaining the HSP(high-scoring segment pairs)

- chr14:70,046,658-70,048,013  1,356 bp **Rhesus Oct. 2010**


**Rhesus position:** chr14:54942051-89328298 size: 34386248
**Strand:** -
**Mouse position:** chr7:87110277-116542127 size: 29431851
**Chain ID:** 18
**Score:** 249021636  **Approximate Score within browser window:** 90214

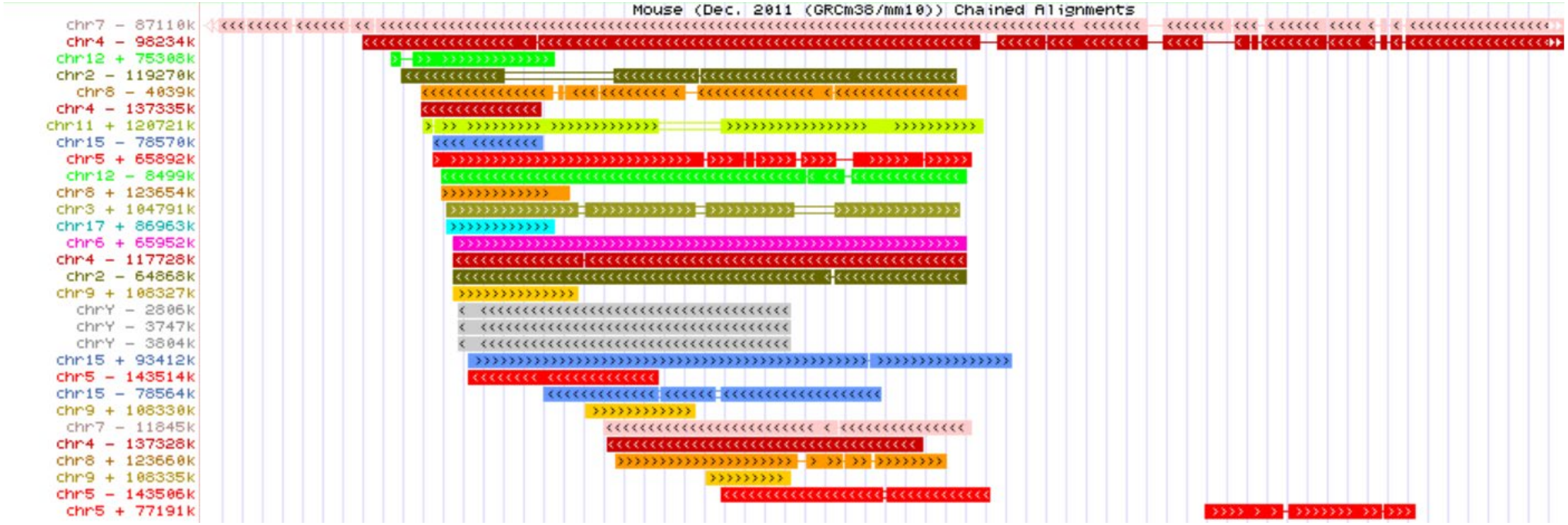
In total, this chain contain 398 HSPs

**Rhesus position:** chr14:70046817-70048041 size: 1225
**Strand:** -
**Mouse position:** chr4:98234213-98235361 size: 1149
**Chain ID:** 19415
**Score:** 71386  **Approximate Score within browser window:** 68730

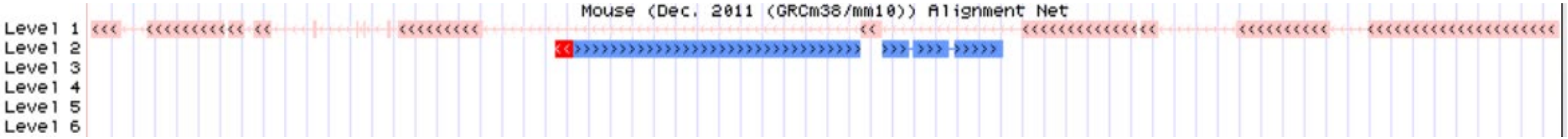
In total, this chain contain 24 HSPs

# Many Chains can map to the same region

- chr14:70,046,658-70,048,013  1,356 bp **Rhesus Oct. 2010**



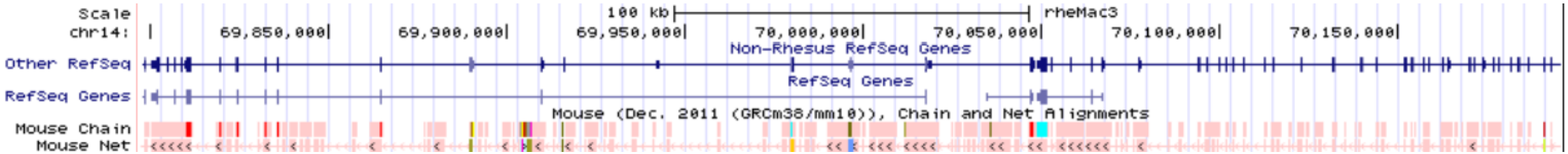
# A **net** is a hierarchical collection of chains




The highest-scoring non-overlapping chains on top, and their gaps filled in where possible by lower-scoring chains, which in turn may have gaps filled in by lower-level chains and so on. Maximum seven level are allowed.

# Chain and Net are enriched in exons

chr14:69,797,131-70,196,130 **Rhesus Oct. 2010** 399,000 bp.



# Software and pipeline

- UCSC Blastz/Lastz, chain, netting pipeline
- MUMer (Maximal Unique Matcher (MUM)
  occurs only once in both sequences

- Minimap2 using minimizer to save sequence information

- MCScanX

# UCSC Blastz/Lastz, chain, netting pipeline

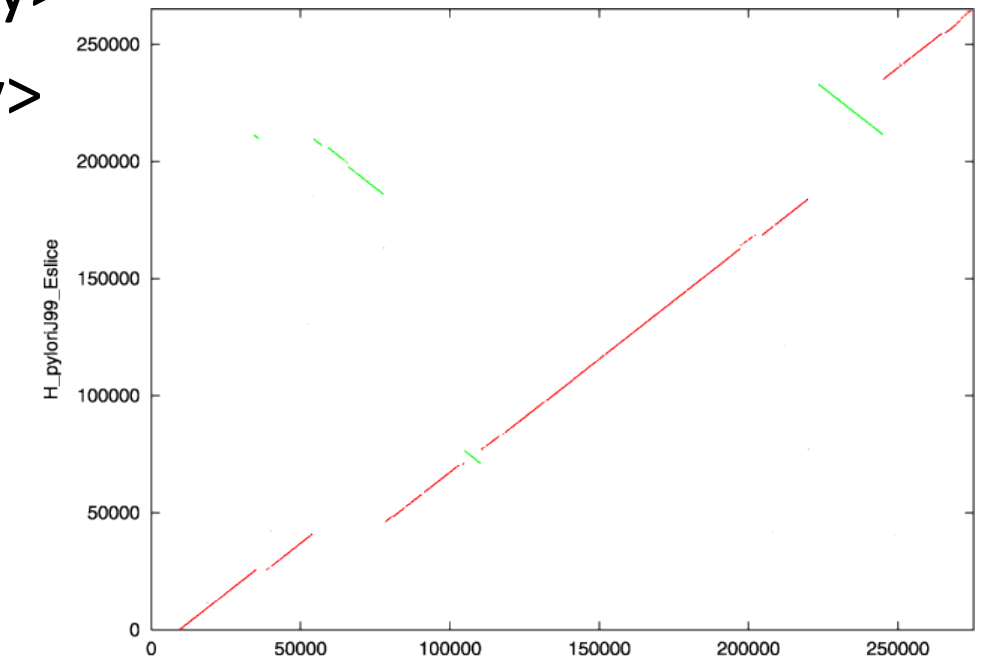
- Step 0: Both genomes have to be repeatmasked and masked
  Tandem Repeat Finder (trf) first
- Step 1: Alignments with Blastz /lastz, an improved version of blastz
- Step 2: Chaining
- Step 3: Netting

UCSC tool chain command: DoBlastzChainNet.pl can now perform this entire sequence

# MUMmer4

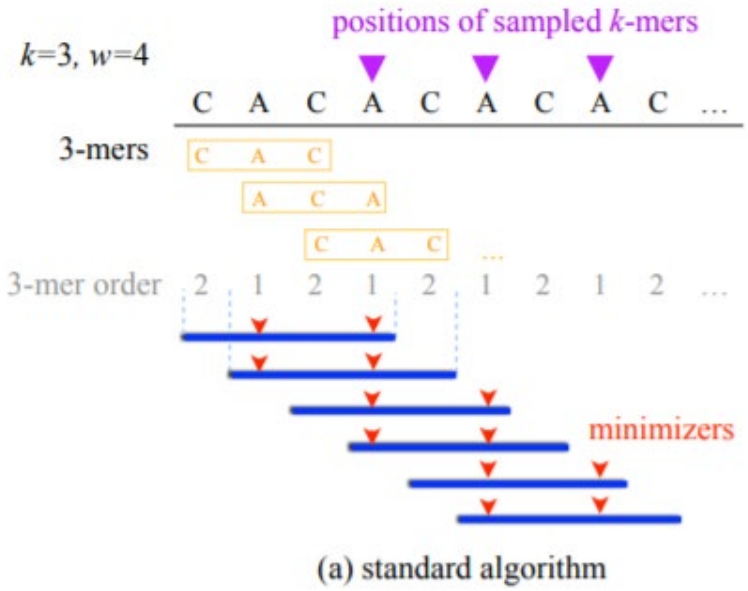
**MUMmer4: Multithreaded, whole genome, align long reads, Comparative Scaffolding**

- ./nucmer -p <prefix> <reference> <query>
- ./promer -p <prefix> <reference> <query>
- ./dnadiff [options] <reference> <query>
- ./mummerplot



# Minimap2:A versatile pairwise aligner

Using Minimizer to representing sequences



(a) standard algorithm

./minimap2 -ax map-pb ref.fa pacbio.fq.gz > aln.sam # PacBio genomic reads

./minimap2 -ax map-ont ref.fa ont.fq.gz > aln.sam # Oxford Nanopore genomic reads

./minimap2 -ax asm20 ref.fa pacbio-ccs.fq.gz > aln.sam # PacBio CCS genomic reads

./minimap2 -ax sr ref.fa read1.fa read2.fa > aln.sam # short genomic paired-end reads

./minimap2 -ax splice ref.fa rna-reads.fa > aln.sam # spliced long reads (strand unknown)

./minimap2 -ax splice -uf -k14 ref.fa reads.fa > aln.sam # noisy Nanopore Direct RNA-seq

./minimap2 -ax splice:hq -uf ref.fa query.fa > aln.sam # Final PacBio Iso-seq or traditional cDNA

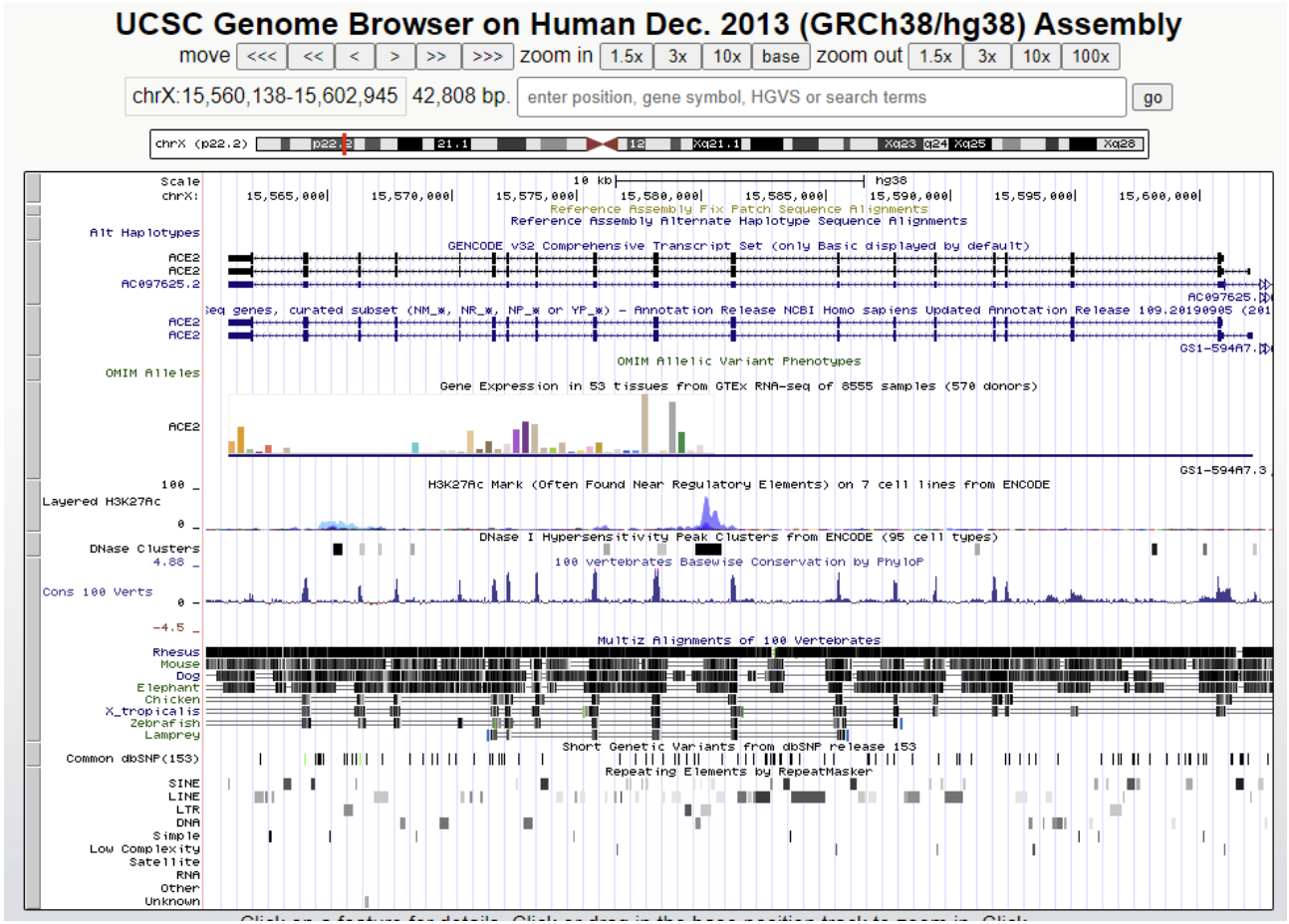
**./minimap2 -cx asm5 asm1.fa asm2.fa > aln.paf # intra-species asm-to-asm alignment**

./minimap2 -x ava-pb reads.fa reads.fa > overlaps.paf # PacBio read overlap

./minimap2 -x ava-ont reads.fa reads.fa > overlaps.paf # Nanopore read overlap

# pre-existing genome alignment database

- http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=cons100way



PGDD

PLANT GENOME DUPLICATION DATABASE

**While PGDD is no longer funded and therefore is not being regularly updated, we intend to keep it online for as long as resources permit and interest warrants. Periodic updates may occur but cannot be assured. Note the last update of PGDD was in 2014.**

PGDD is a public database to identify and catalog plant genes in terms of intragenome or cross-genome syntenic relationships. Current efforts focus on flowering plants with available whole genome sequences (preferrably assembled pseudomolecules with ordered gene models).

## DATA SOURCES

**Plant genomes in this database (47 genomes)**

| Species name | Common name | Version version | Gene number | Access | Reference |
|---|---|---|---|---|---|
| *Actinidia chinensis* | Kiwifruit | May 2013 | 32,670 | KGD | *Nature Communications* |
| *Amborella trichopoda* | Amborella | Version 1.0 | 26,846 | AGD | *Science* |
| *Arabidopsis lyrata* | Lyrate rockcress | Version 1.0 | 32,670 | JGI | *Nature Genetics* |
| *Arabidopsis thaliana* | Arabidopsis | TAIR10 | 27,416 | TAIR | *Nature* |
| *Brachypodium distachyon* | Purple false brome | Version 2.1 | 31,694 | JGI | *Nature* |
| *Brassica oleracea* | Kale | Version 2.1 | 59,225 | BRAD | *Genome Biology* |
| *Brassica rapa* | Chinese cabbage | Version 1.3 | 40,492 | BRAD | *Nature Genetics* |
| *Beta vulgaris* | Sugar beet | RefBeet-1.1 | 27,421 | BVR | *Nature* |
| *Cajanus cajan* | Pigeonpea | Nov 2011 | 48,680 | IIPG | *Nature Biotechnology* |
| *Capsella rubella* | Capsella | Version 1.0 | 26,521 | JGI | *Nature Genetics* |
| *Capsicum annuum* | Hot pepper | Version 1.55 | 34,899 | PepperGenomeDB | *Nature Genetics* |
| *Carica papaya* | Papaya | ASGPBv0.4 | 24,782 | Hawaii | *Nature* |
| *Chlamydomonas reinhardtii* | Green algae | Version 5.5 | 17,741 | JGI | *Science* |
| *Cicer arietinum* | Chickpea | Version 1.0 | 28,269 | LIS | *Nature Biotechnology* |
| *Citrullus lanatus* | Watermelon | Version 1.0 | 23,440 | ICUGI | *Nature Genetics* |
| *Citrus sinensis* | Sweet orange | Version 1.1 | 25,379 | CSAP | *Nature Genetics* |
| *Cucumis sativus* | Cucumber | Version 1.0 | 21,491 | JGI | *Nature Genetics* |
| *Eucalyptus grandis* | Eucalyptus | Version 1.1 | 36,376 | JGI | *Nature* |
| *Elaeis guineensis* | Oil palm | Version 2.0 | 30,752 | MPOB | *Nature* |
| *Fragaria vesca* | Strawberry | Version 1.1 | 32,831 | PFR | *Nature Genetics* |
| *Glycine max* | Soybean | Wm82.a2.v1 | 56,044 | JGI | *Nature* |

HOME
DOT PLOT
LOCUS SEARCH
MAP VIEW
DOWNLOADS
ABOUT
CONTACT
MCSCAN

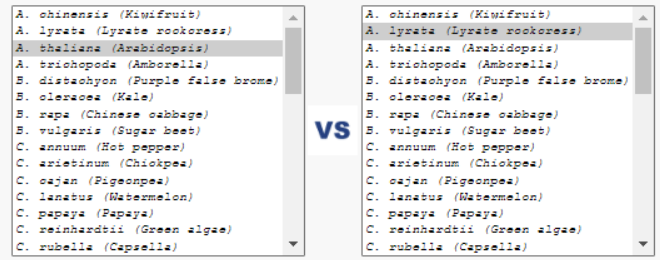
TOOLS FOR SYNTENY

CoGe
VISTA
phytozome
Plaza
Ensembl

PGDD

PLANT GENOME DUPLICATION DATABASE

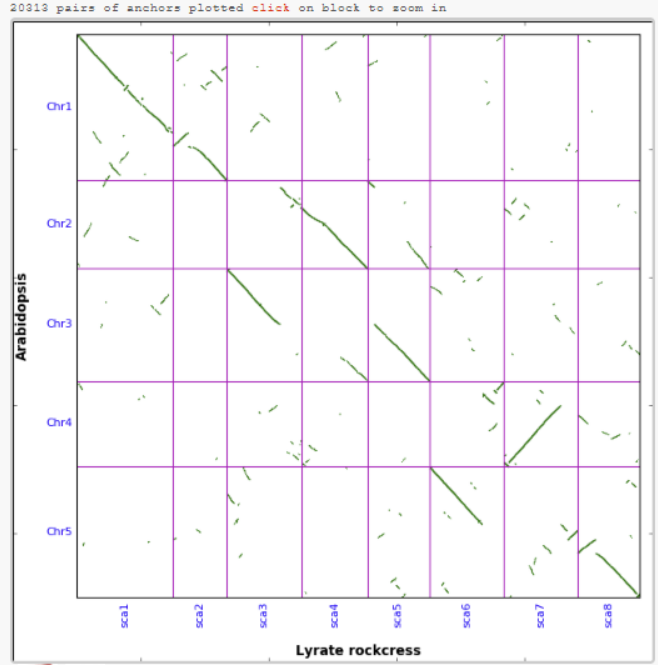
## VISUALIZING SYNTENIC BLOCKS




Ks filter: between 0.5 and 1.0 *(use Ks button below to identify the range)*
Plot with base-pair distance *(default: gene ranks)*
Show boundary lines
- Minimum length of a scaffold to display Mbp

Dotplot Reset Ks distribution

20313 pairs of anchors plotted click on block to zoom in




PGML HOMEPAGE
HOME
DOT PLOT
LOCUS SEARCH
MAP VIEW
DOWNLOADS
ABOUT
CONTACT
MCSCAN

TOOLS FOR SYNTENY

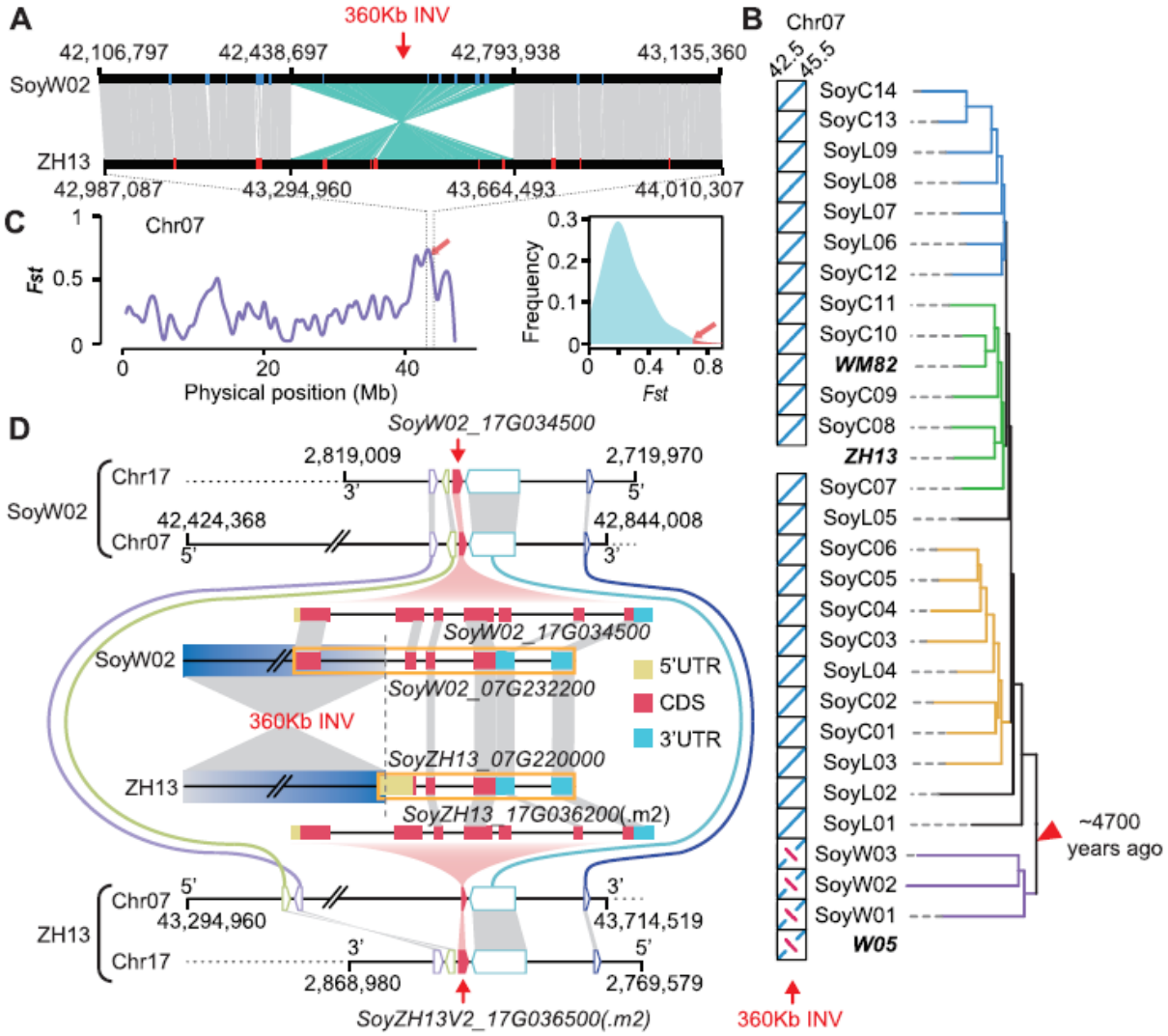
CoGe
VISTA
phytozome
Plaza
Ensembl

# Application of Genome Alignment

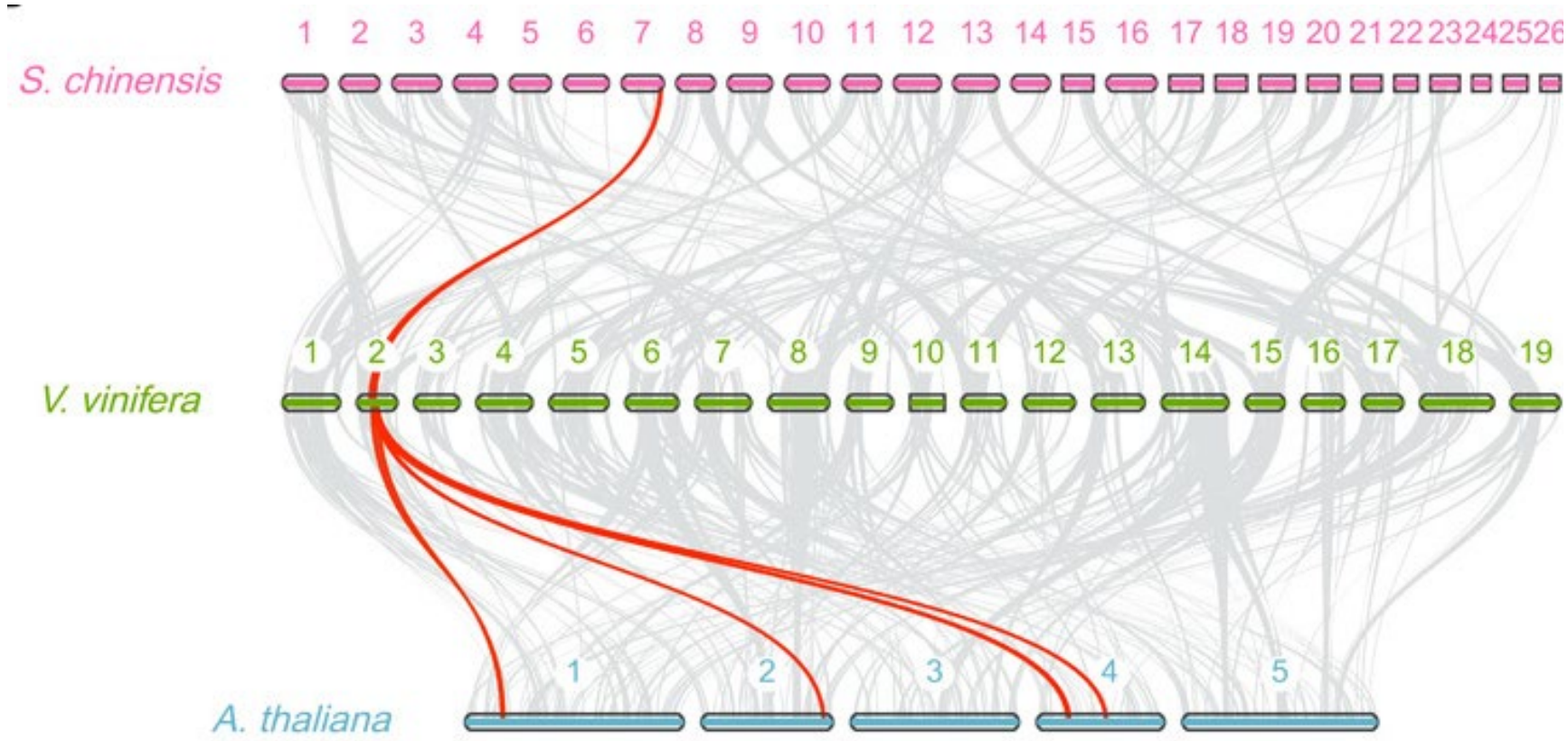
- Identify SV among the genome
- chromosome rearrangement
- Syntenic  orthologue
- Identify core genome and dispensable genome
- Find evolutionary footprints
- Coordinates lifting between Genome version

# Identify SV among the genome

An Inversion on Chromosome 07 between Wild Soybean and Cultivated Soybean Genomes Is Associated with Soybean Domestication



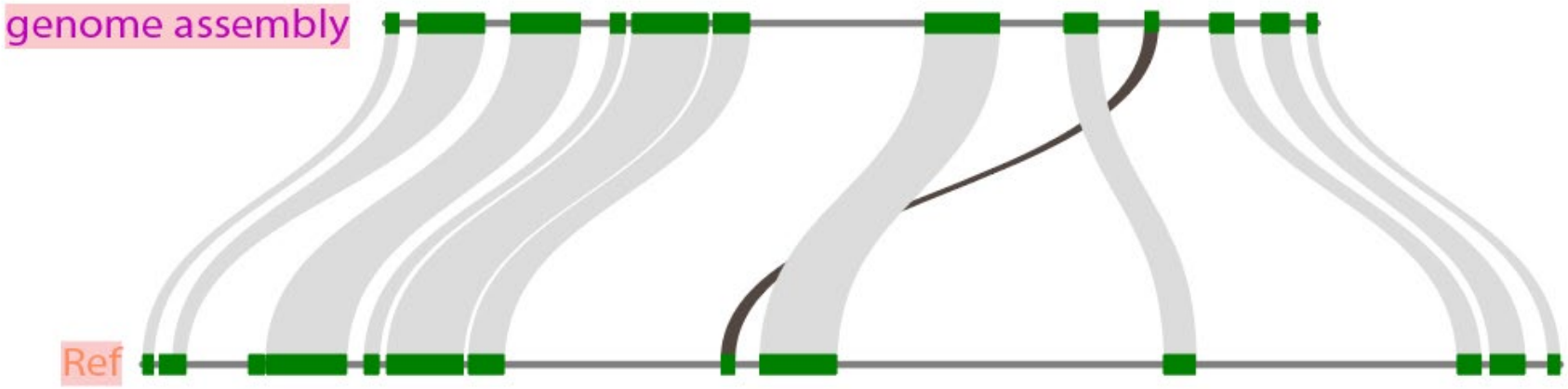
# chromosome rearrangement



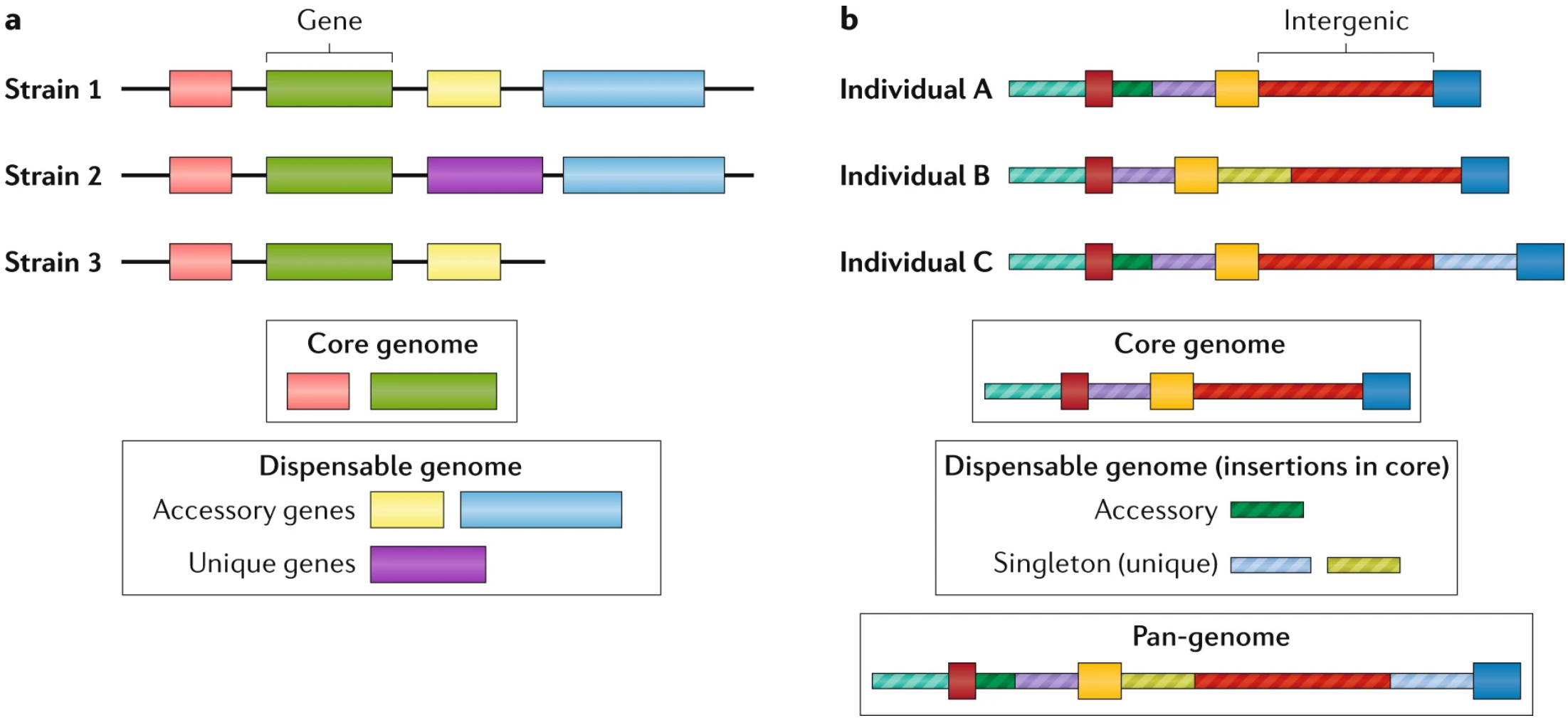

Sturtevant D, Lu S, Zhou Z W, et al. The genome of jojoba (Simmondsia chinensis): A taxonomically isolated species that directs wax ester accumulation in its seeds[J]. Science Advances, 2020, 6(11): eaay3240

# Syntenic orthologue

**Syntenic genome alignment**

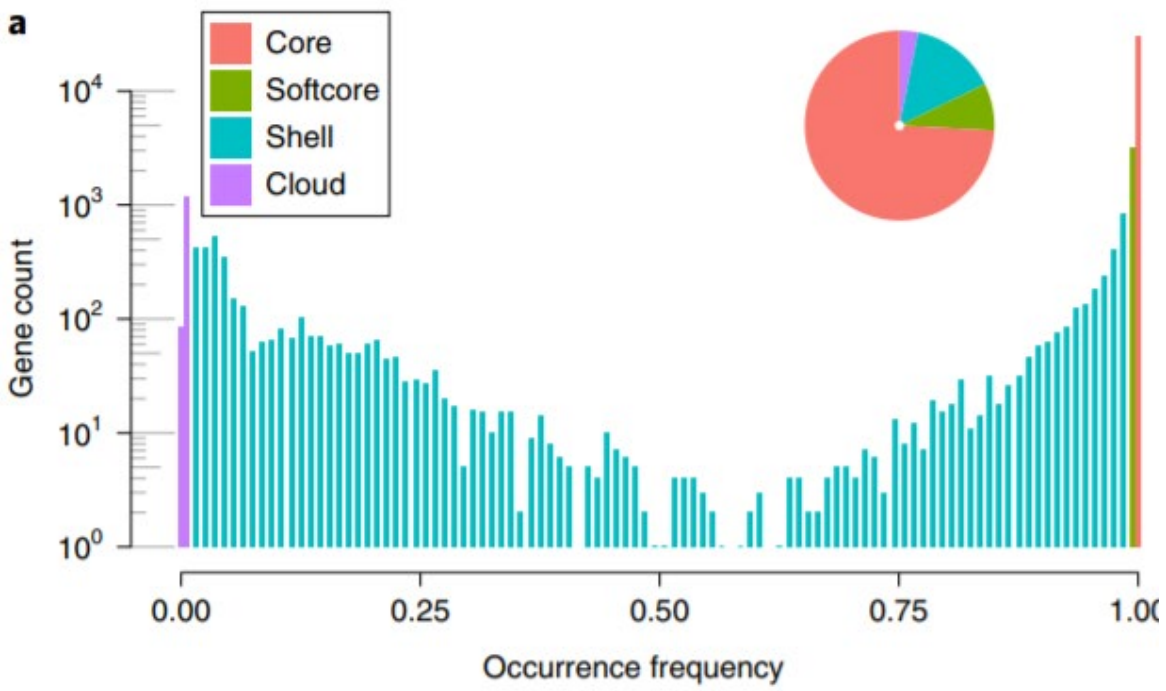


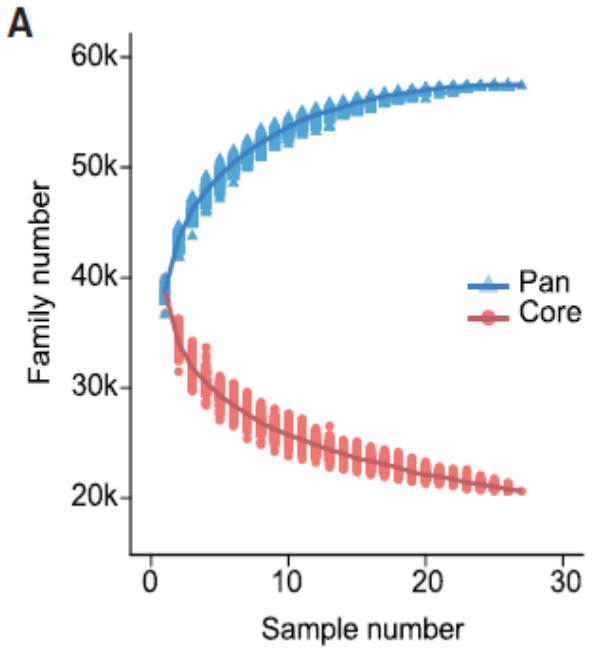
# Identify core genome and dispensable genome

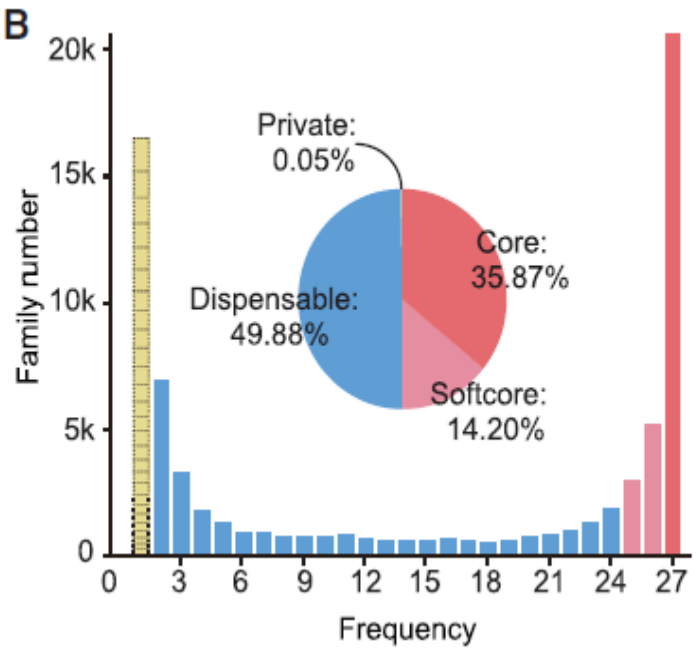



10.1038/s41576-020-0210-7

# Core and softcore genome



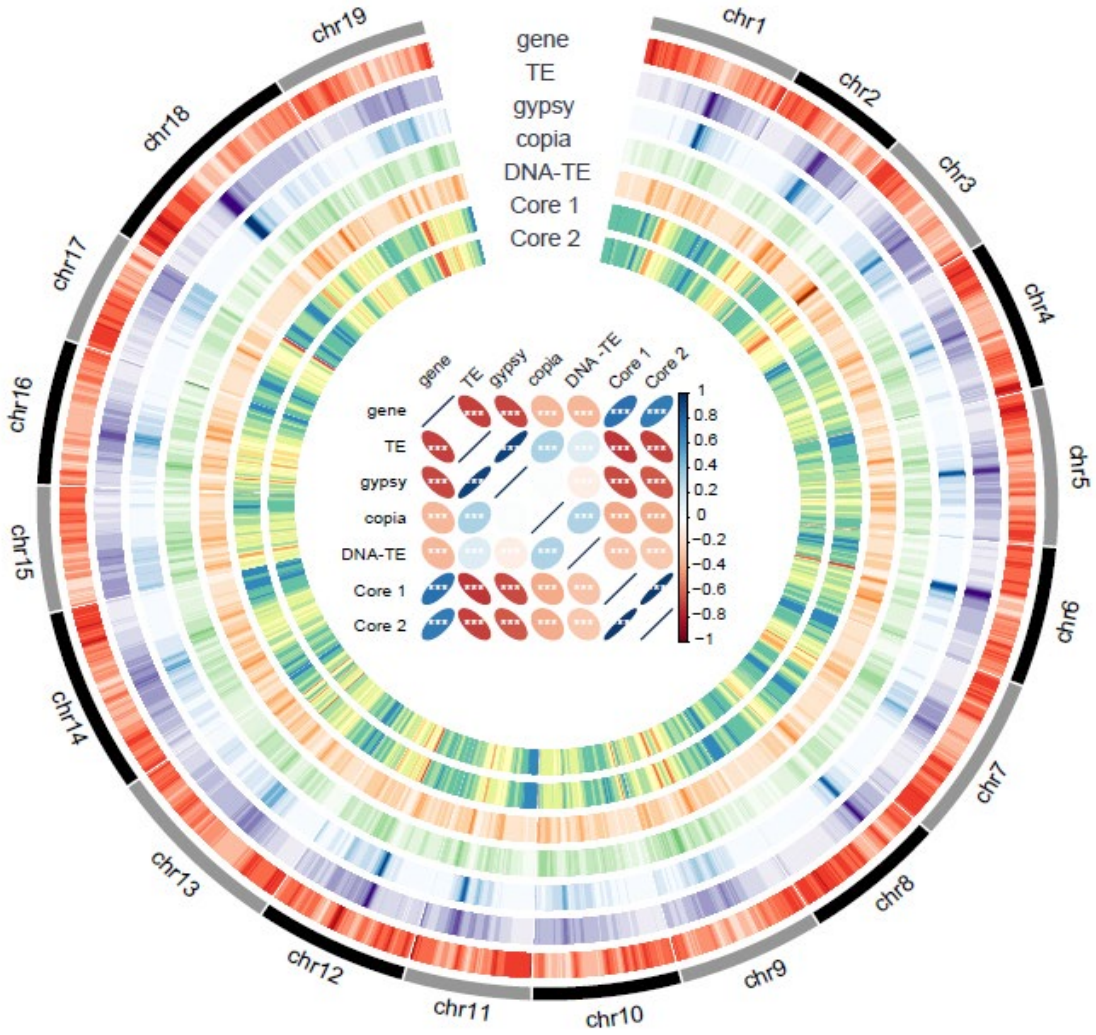

**725 tomato accessions**

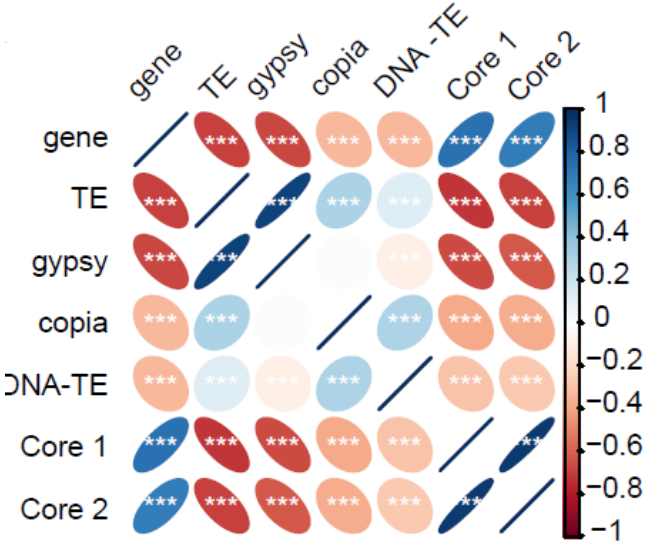
(10.1038/s41588-019-0410-2)

- 2,898 soybean accessions WGS
- De Novo Genome Assembly and Annotation of 26 Soybean Accessions

(10.1016/j.cell.2020.05.023)

# Correlation between the core genome and other genome features



# Find evolutionary footprints