PanTools update

March 2021

WAGENINGEN
UNIVERSITY & RESEARCH

PanTools v3.1 has been released

- Code
  http://git.wur.nl/bioinformatics/pantools
- Manual/tutorial
  http://www.bioinformatics.nl/pangenomics/manual/

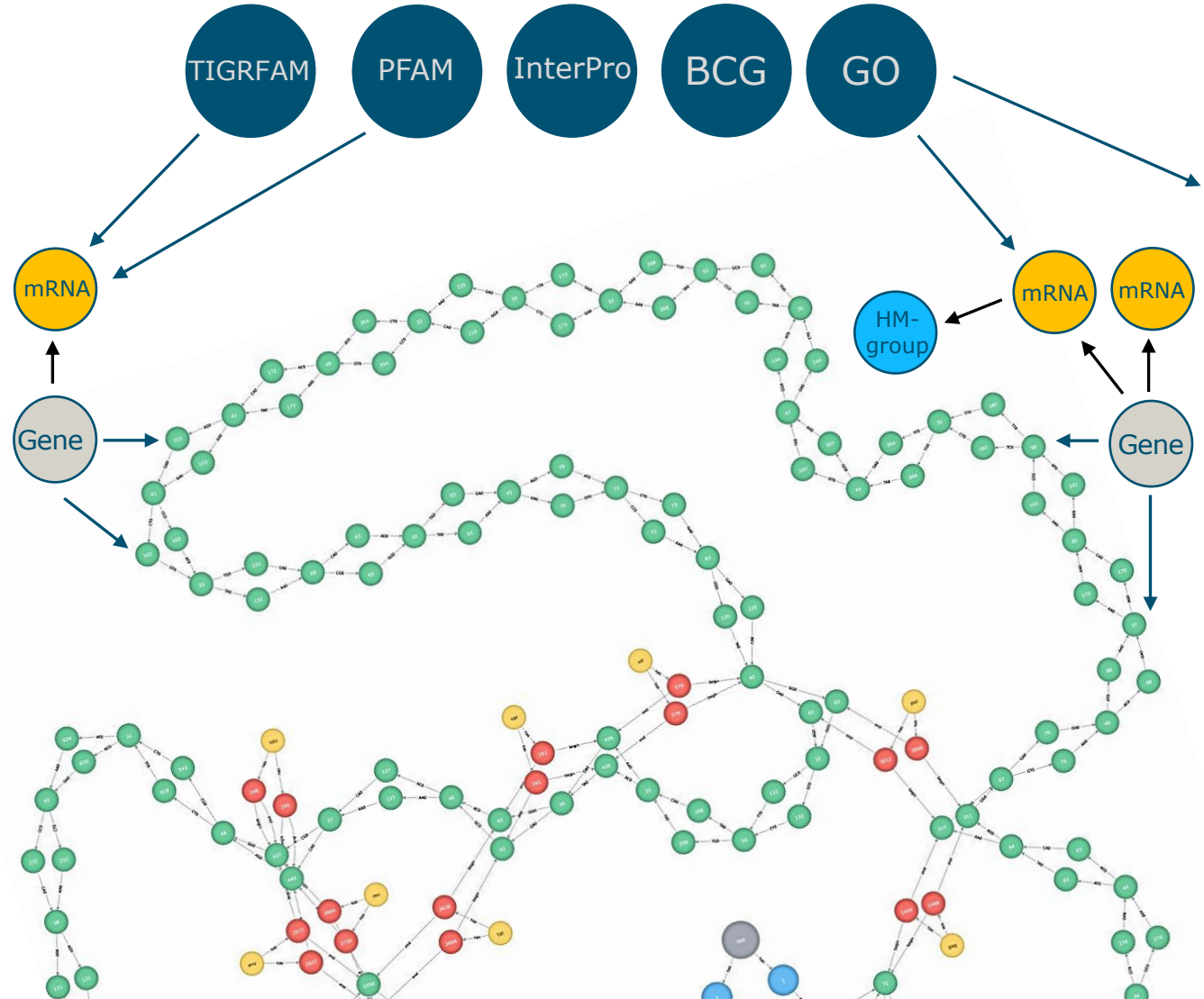
- Paper submitted to BMC genomics
  The *Pectobacterium* pangenome, with a focus on *Pectobacterium brasiliense*, shows a robust core and extensive exchange of genes from a shared gene pool

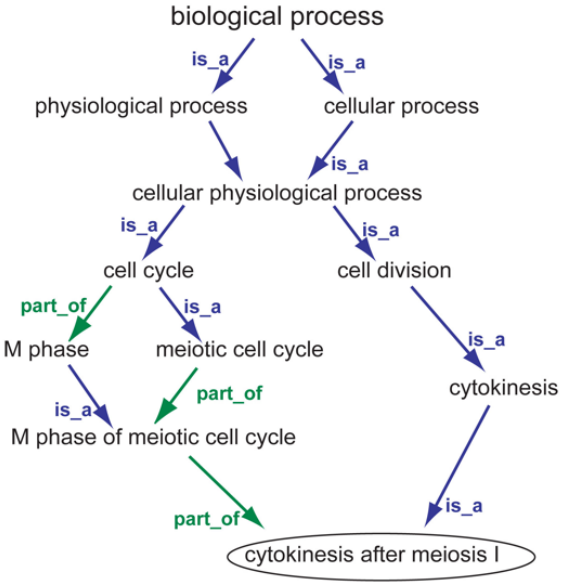
PanTools versions

http://git.wur.nl/bioinformatics/pantools

- PanTools v1
  - Construction, annotation, homology grouping
- PanTools v2
  - Read mapping (resequencing data)
- PanTools v3
  - More annotations, optimal homology grouping, gene-level analyses, pangenome size, phylogenetics

Functional annotations



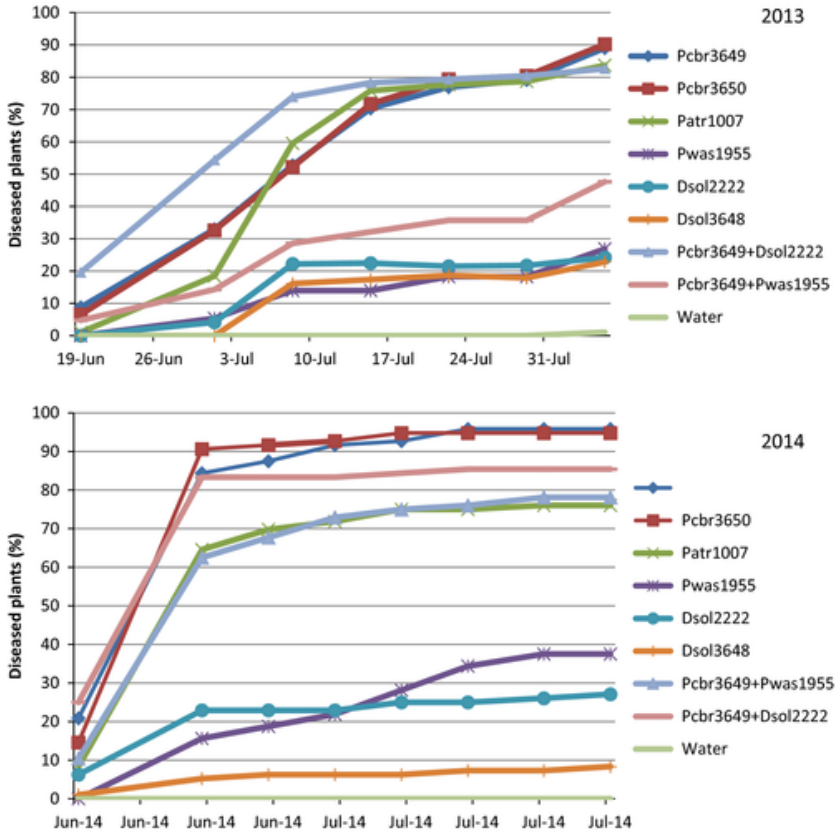


GENEONTOLOGY
Unifying Biology

Phobius
SignalP

Phenotype incorporation

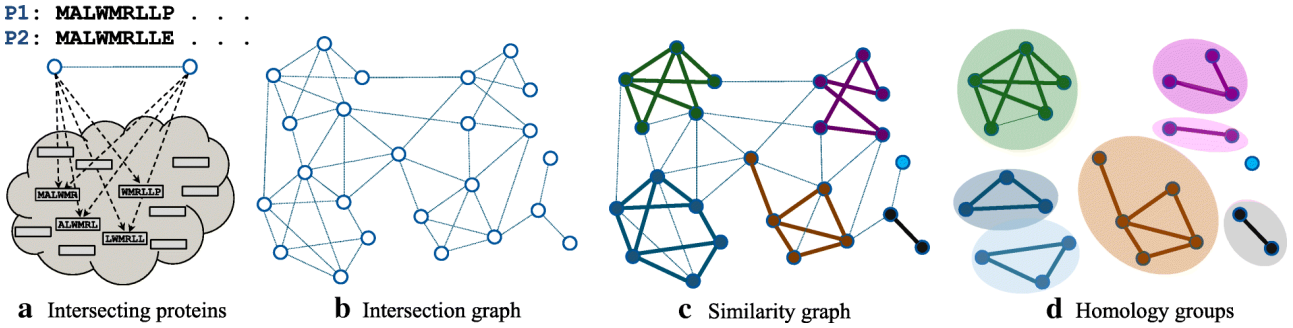
- Species
- Country/region
- Color
- Host
- Virulence
- Flowering time
- …



Van der Wolf, et al. (2016)

Optimal homology grouping

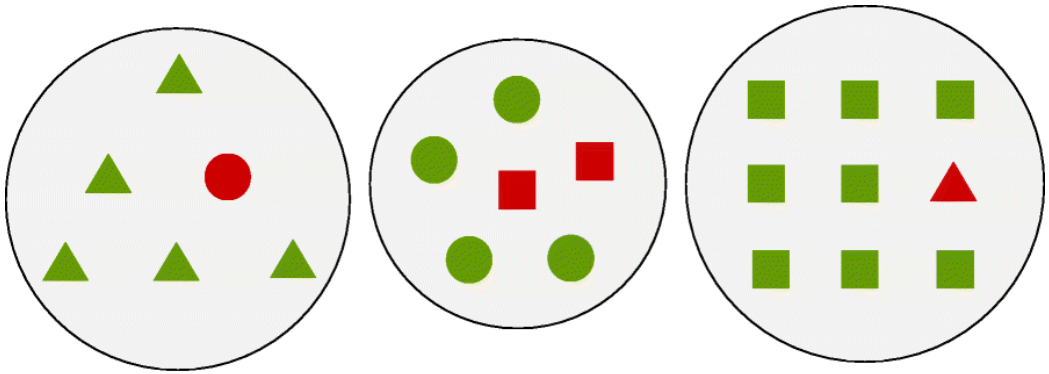
- Four settings to influence clustering
  - Intersection rate
  - Sequence similarity
  - MCL inflation
  - Contrast factor
- Precooked settings from D1 (strict) to D8 (relaxed)
- Which setting is optimal for my data set?



WAGENINGEN
UNIVERSITY & RESEARCH

Optimal homology grouping

- Use ‘complete’ BUSCO’s to validate grouping
- Calculate recall, precision, and F-score



	△	○	□	
tp	5	4	8	**TP = 17**
fn	1	1	2	**FN = 4**
fp	1	2	1	**FP = 4**

Efficient inference of homologs in large eukaryotic pan-proteomes

Siavash Sheikhizadeh Anari, Dick de Ridder, M. Eric Schranz & Sandra Smit

BMC Bioinformatics 19, Article number: 340 (2018) | Download Citation

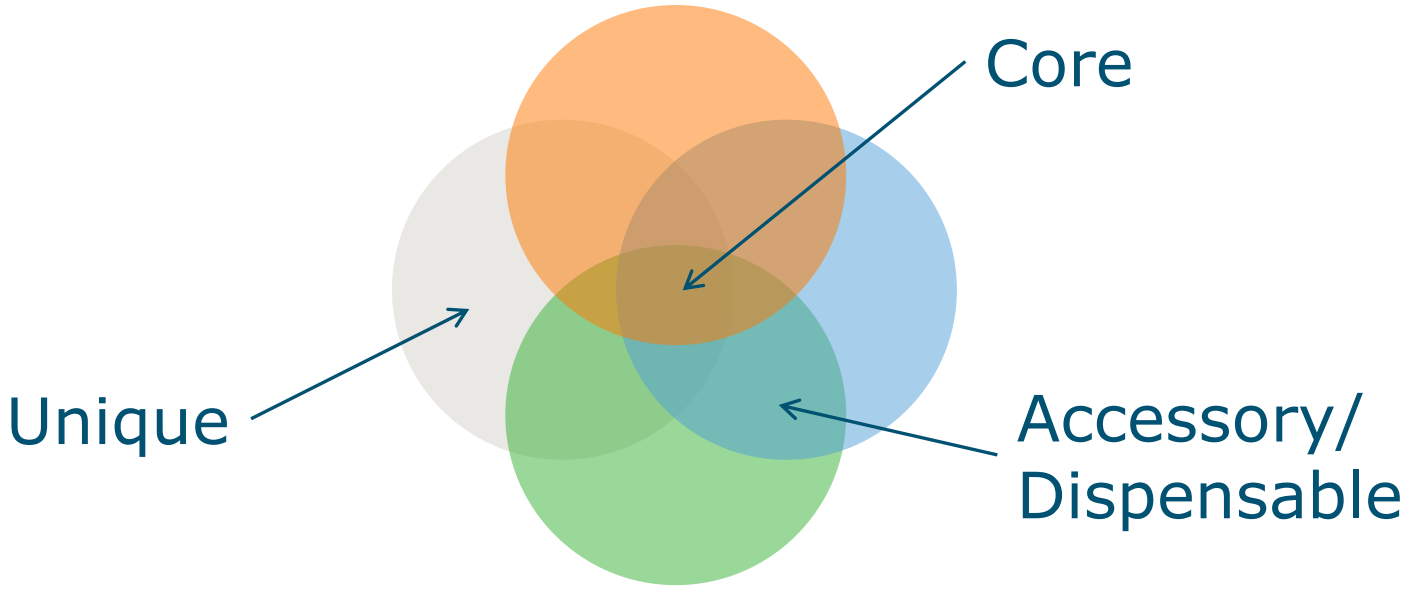
Optimal homology grouping

- 670 BUSCOs * 204 genomes: 136.680 genes
- Recall: TP/(TP+FN)
- Precision: TP/(TP+FP)
- F-score: 2×(Recall×Precision)/(Recall+Precision)

	Single copy groups	Correct groups (670)
D1 - 95%	816	395
D2 - 85%	1609	630
D3 - 75%	1687	638
D4 - 65%	1694	640
D5 - 55%	1681	639
D6 - 45%	1653	635
D7 - 35%	1582	619
D8 - 25%	1480	608

TP	FP	FN	Recall	Precision	F-score
132.929	14	3751	0,97256	0,9999	0,9860
136.473	24	207	0,99849	0,9998	0,9992
136.642	35	38	0,99972	0,9997	0,9997
136.665	38	15	0,99989	0,9997	0,9998
136.665	42	15	0,99989	0,9997	0,9998
136.675	243	5	0,99996	0,9982	0,9991
136.677	1515	4	0,99997	0,9890	0,9945
136.676	7885	4	0,99997	0,9455	0,9719

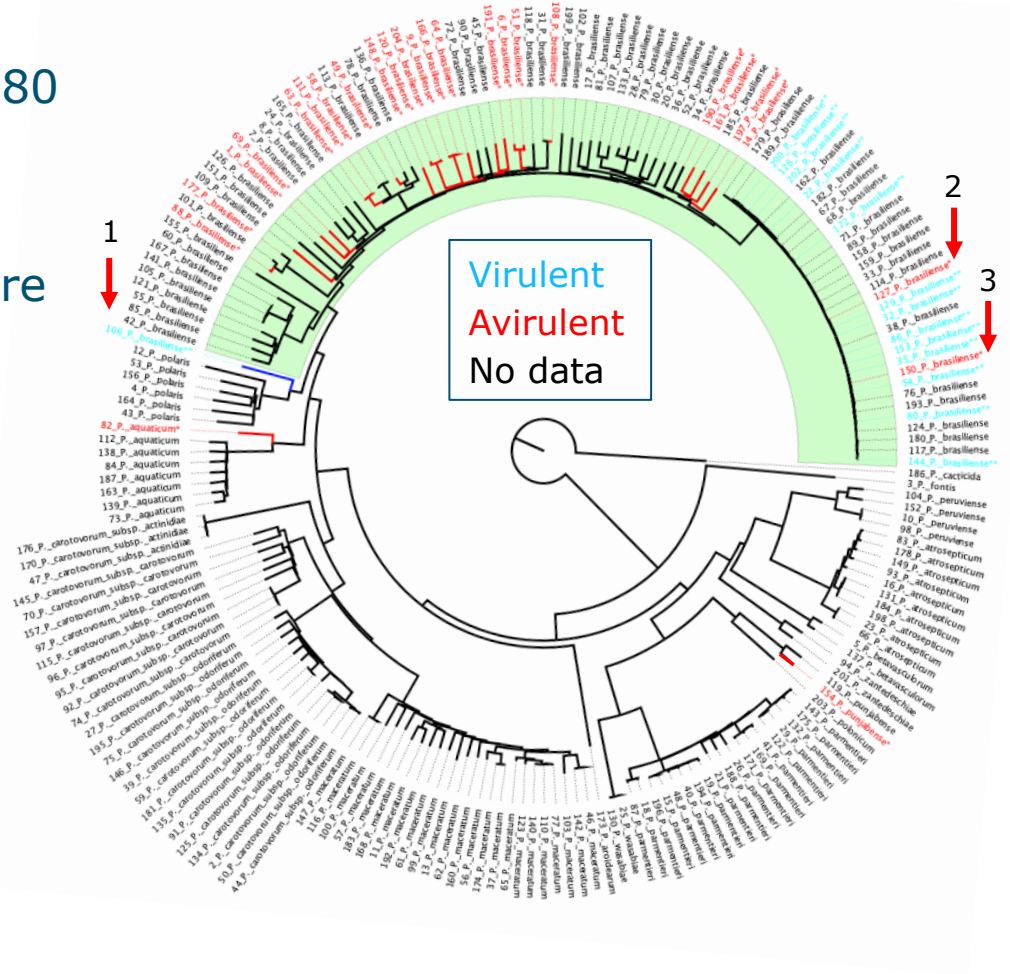
Gene-level analysis



Homology group	*P. atrosepticum*					*P.wasabiae*			*P. odoriferum*				**Gene definition**
	A1	A2	A3	A4	A5	W1	W2	W3	O1	O2	O3	O4	
1	1	3	1	1	2	1	1	1	1	1	1	1	Core
2	1	0	1	1	1	1	1	0	0	0	1	1	Dispensable
3	0	0	1	0	0	0	0	0	0	0	0	0	Unique
4	1	1	1	2	1	0	0	0	0	0	0	0	Species specific

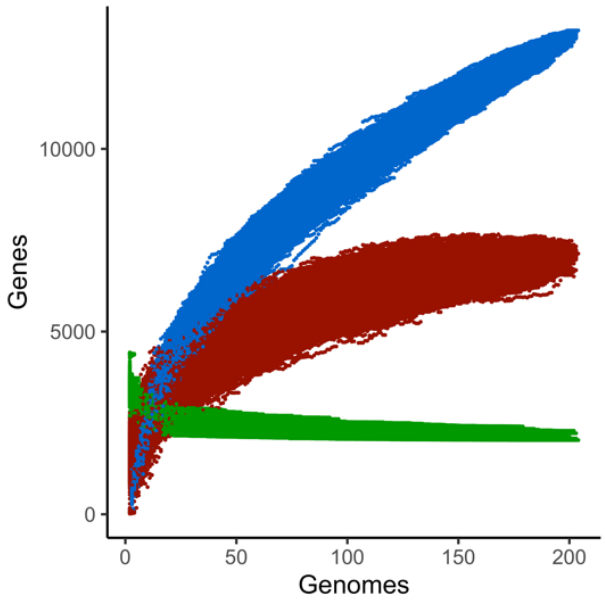
Gene classification: virulence

- Genes associated to phenotype: 80
- Most are colocalized: 66-71
- 6 clusters of minimum 3 genes are conserved in the 15 genomes
  - 4x 3 genes
  - 1x 4 genes
  - 1x 6 genes



The pangenome size

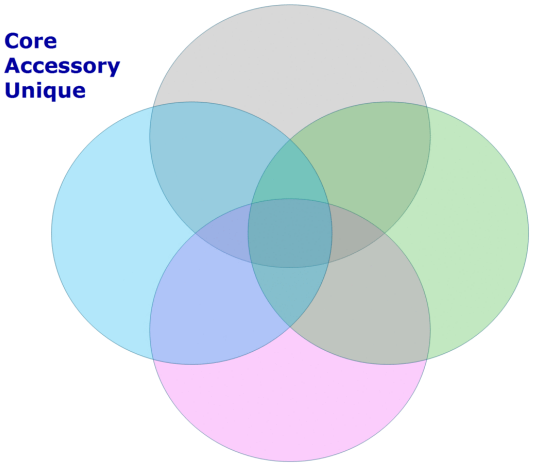
- Each iteration is a random combination of X genomes
- Pectobacterium
  - 22.403 homology groups (genes)
  - 10.000 iterations



Accessory

Unique

Core



Integrated phylogeny methods

- Automated MLSA (ML)
- ANI distance (NJ)
- K-mer distance (NJ)
- SNPs from single copy orthologs (NJ/ML)

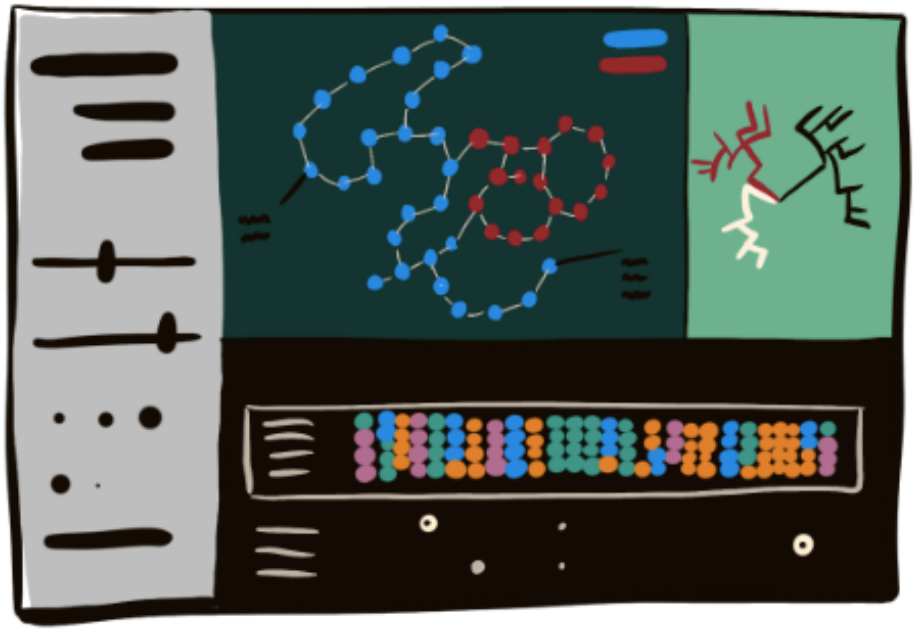
- Shared gene distance (NJ)

ML - Maximum likelihood

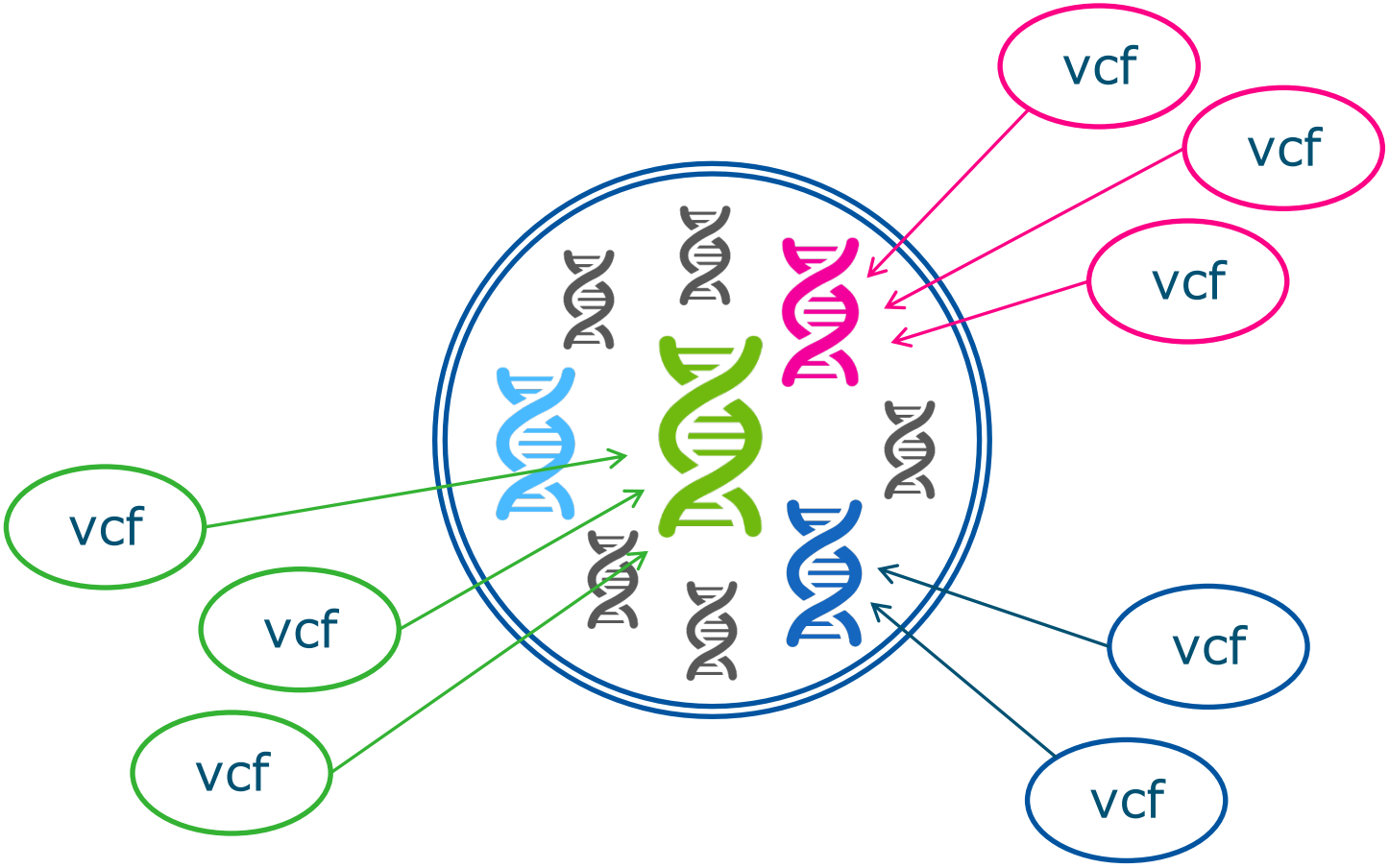
NJ - Neighbor joining

Progress related to VAPP

- Variant exploration
- Synteny analysis
- Scalability



Variant exploration (thesis Edwin)



Variant exploration (thesis Edwin)

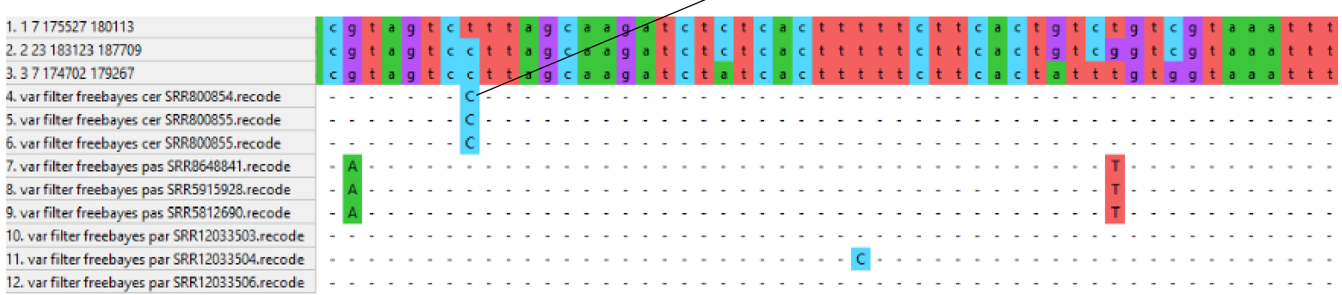
A.

2 23 183123 187709:	c	g	t	a	g	t	c	c	t	t	a	g	c	a	a	g	a	t	c	t	c	t	c	a	c	t	t	t	t	t	c	t	t	c	a	c	t	g	t	c	g
3 7 174702 179267:	c	g	t	a	g	t	c	c	t	t	a	g	c	a	a	g	a	t	c	t	a	t	c	a	c	t	t	t	t	t	c	t	t	c	a	c	t	a	t	t	t
1 7 175527 180113:	c	g	t	a	g	t	c	t	t	t	a	g	c	a	a	g	a	t	c	t	c	t	c	a	c	t	t	t	t	t	c	t	t	c	a	c	t	g	t	c	t
A:	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C:	-	-	-	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-
T:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
G:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GAATTCGAATCAAAT:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AGCG:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AG:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TGCA:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CTCC:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ATTTG:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CTTC:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

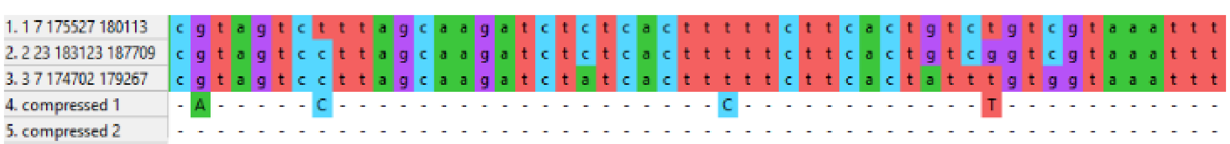
B.

Ref_Base	Variant_Base	Original_value	Aligned
A	G	175590	63
A	G	175787	260
G	A	175866	339
A	G	175871	344
C	G	175893	366
T	C	176107	580
A	G	176415	888
A	C	177152	1625
C	T	177165	1638
T	C	177524	1997
C	T	177632	2105
T	C	177687	2160
C	T	177783	2256
A	G	177876	2349
T	C	177882	2355
G	T	178052	2525
G	A	178133	2606
C	T	178179	2652
C	T	178295	2768
G	A	178402	2875
T	C	178509	2982
G	A	179910	4383
A	G	180027	4500

C.

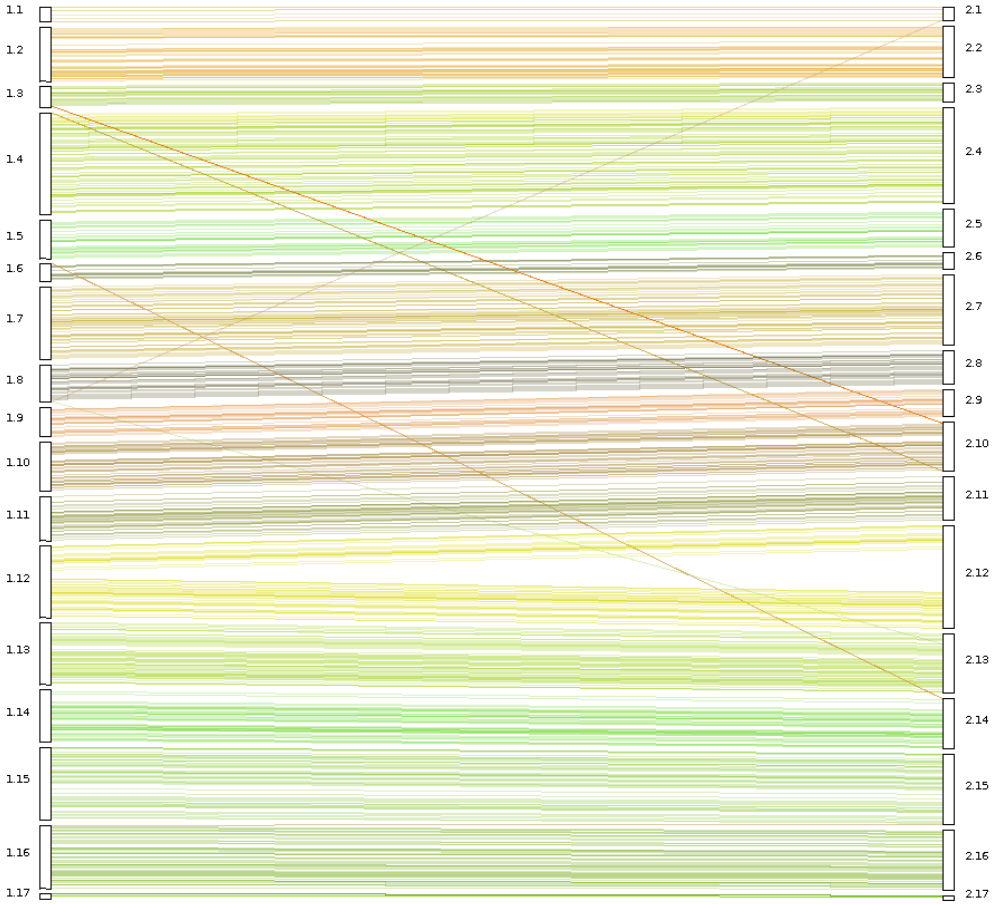


D.



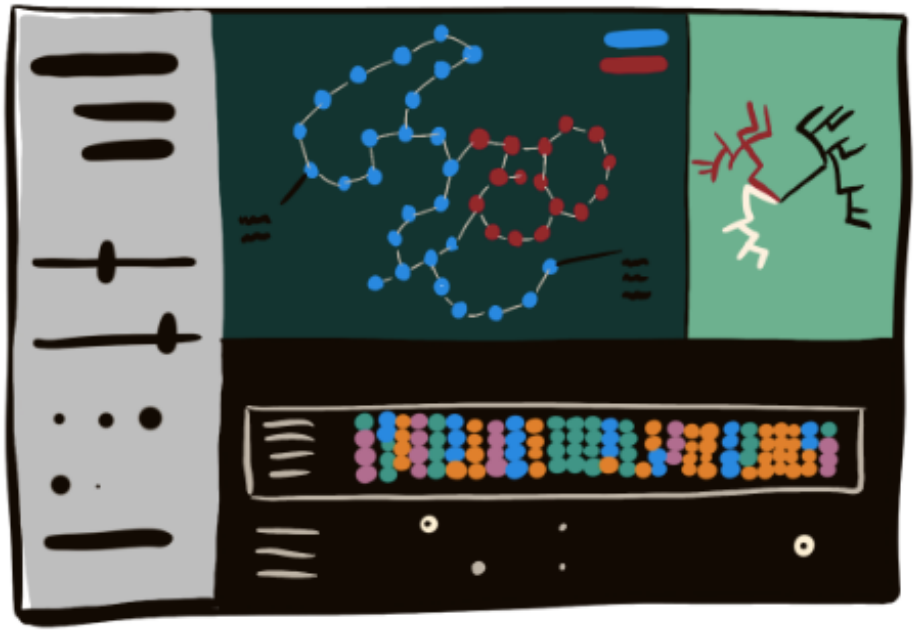
Synteny (thesis Roel)

- Conserved gene order
  - 2, 3, …, 100 species
- Algorithm make use of homolgy groups and gene location



Scaling up pangenomics for plant breeding

Next:
Visual analytics for plant pangenomes



Living up to the expectations:
"A core part of our work is how we might provide visualisations of variation for breeders, which is a particular strength of your tools."