Agriculture and Agri-Food Canada

Draft genome assemblies and the pan-genome of *Pyrenophora tritici-repentis*, the causal agent of the wheat disease tan spot

Ryan Gourlie[1], Rodrigo Ortega Polo[1], Kaveh Ghanbarnia[1], Mohamed Hafez[1], Raja Ragupathy[1], Fouad Daayf[2], Stephen Strelkov[3], and Reem Aboukhaddour[1]*

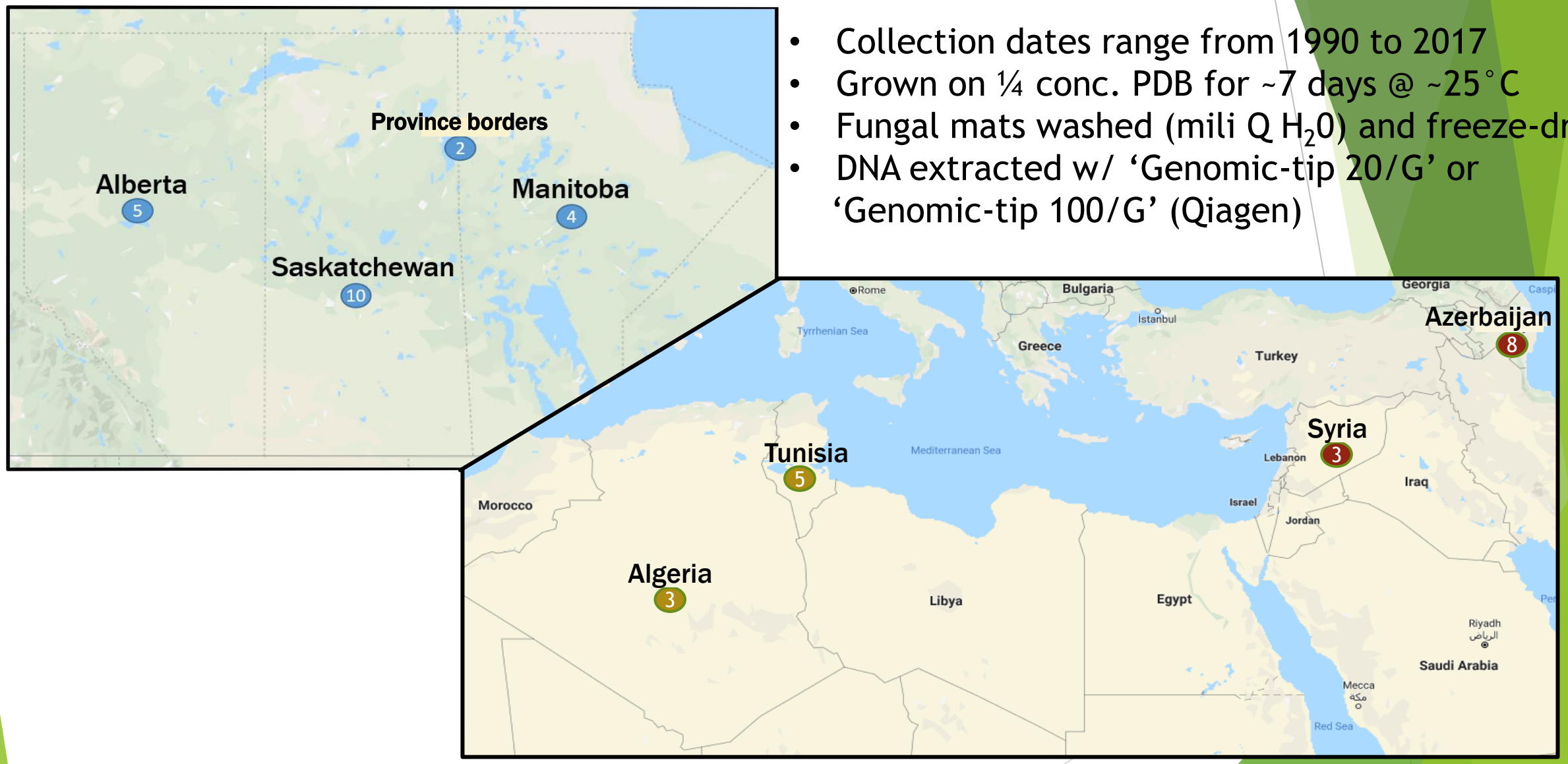
[1]*Agriculture and Agri-Food Canada, Lethbridge Research and Development Center, Lethbridge, AB, Canada*
[2]*University of Manitoba, Department of Plant Science, Winnipeg, MB, Canada*
[3]*University of Alberta, Department of Agricultural, Food and Nutritional Science, Edmonton, AB, Canada*

Introduction

- *Pyrenophora tritici-repentis* (Ptr) is a destructive foliar pathogen of wheat worldwide.
- Eight races have been established based on their ability to produce combinations of three necrotrophic effectors (i.e. host-selective toxins).
- Objective: sequence large number of genomes and create high quality de novo assemblies to answer a broad set of research questions.

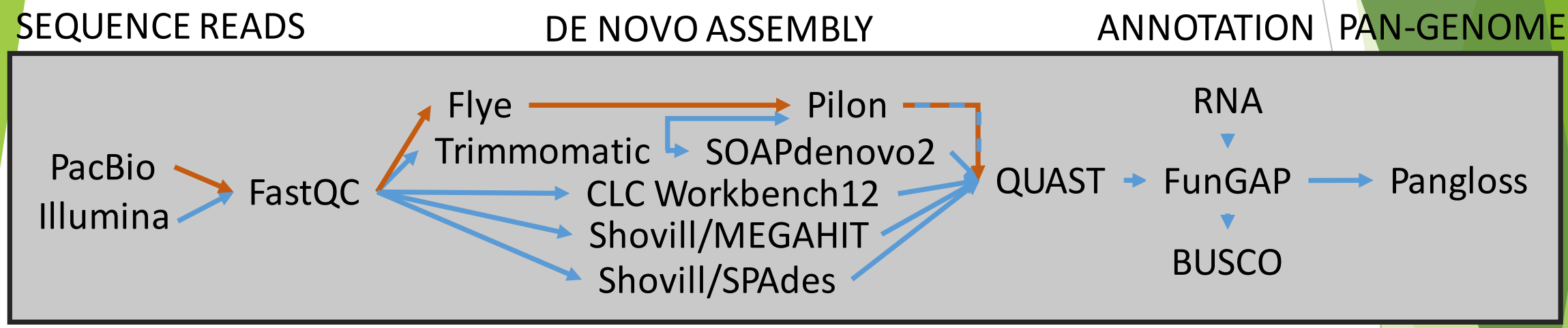
Race	ToxA	ToxB	ToxC	Number of isolates sequenced
1	+	-	+	10
2	+	-	-	6
3	-	-	+	4
4	-	-	-	3
5	-	+	-	8
6	-	+	+	3
7	+	+	-	2
8	+	+	+	3
novel				1
			Total	40

Number of Ptr isolates by location

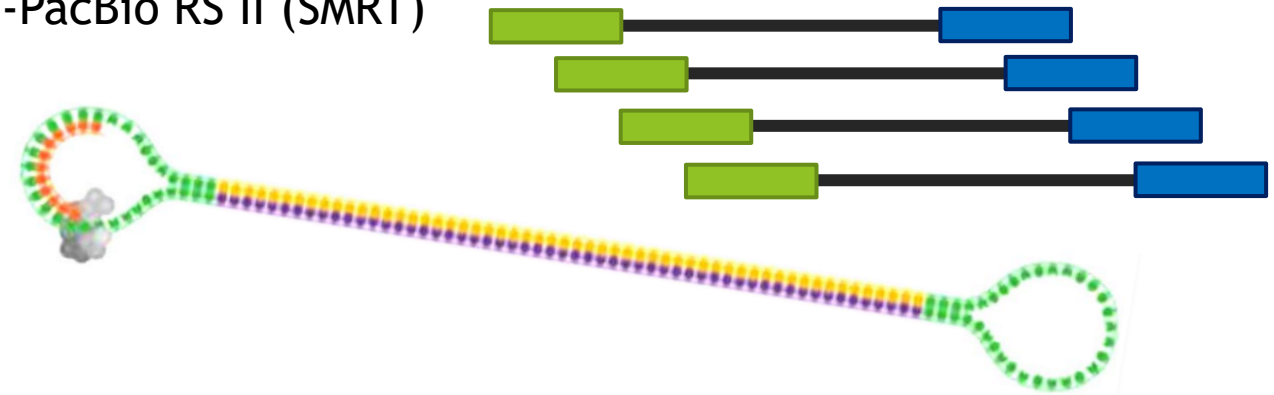


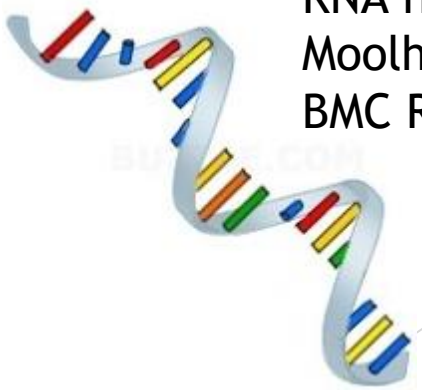
- Collection dates range from 1990 to 2017
- Grown on ¼ conc. PDB for ~7 days @ ~25°C
- Fungal mats washed (mili Q H_20) and freeze-dried
- DNA extracted w/ 'Genomic-tip 20/G' or 'Genomic-tip 100/G' (Qiagen)

Overview of Ptr genomics pipeline



Sequenced at Genome Quebec
-Illumina Hiseq X (shotgun; 150 bp paired-end)
-PacBio RS II (SMRT)

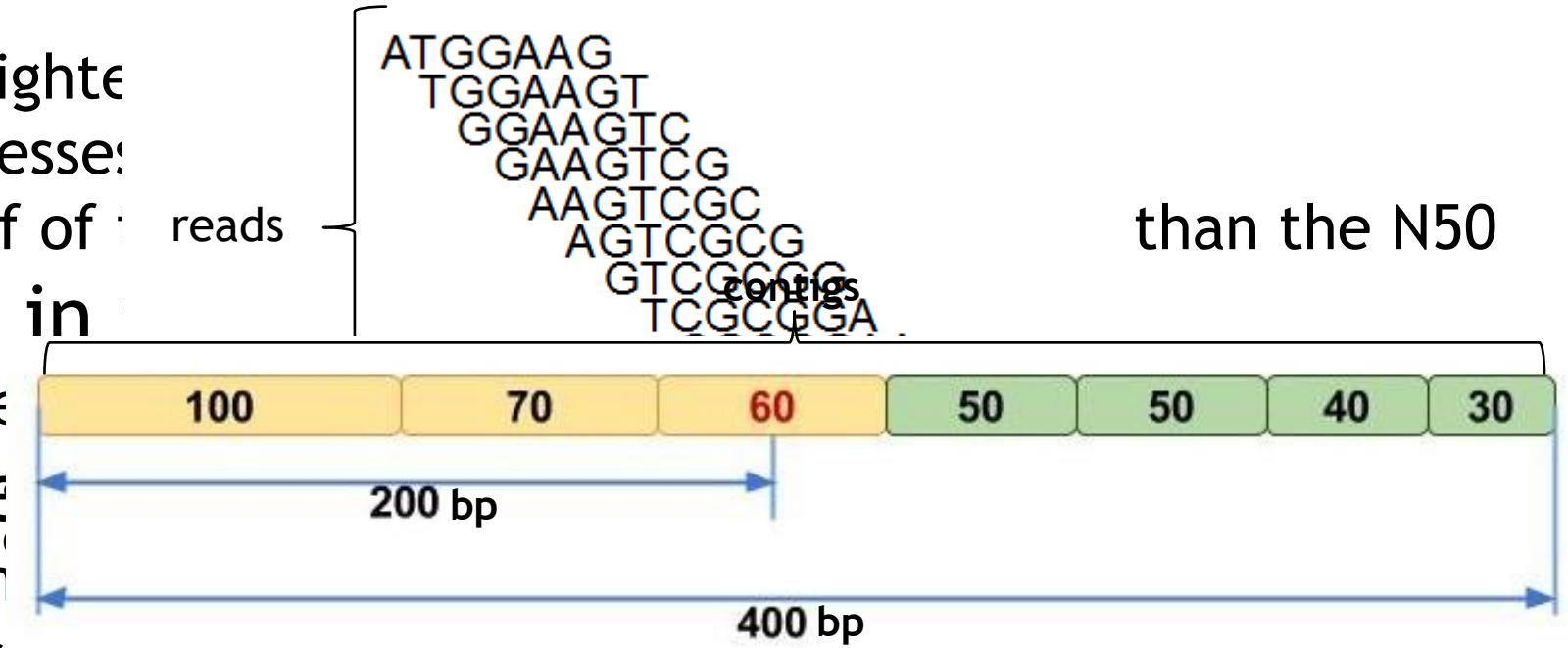




RNA from:
Moolhuijen et al., 2018.
BMC Research Notes, 11(1), 907-909

How to measure the quality of an assembly?

- Number of contigs
  - How many contiguous sequences were assembled?
  - Fewest contigs possible = number of chromosomes
- N50
  - weighte
  - assesse
  - half of
- Gaps in
  - inse
- Perce
  - Ber
  - Ascomycota set has 1,315 genes all "should" be present in assembly



than the N50

difficult seq)

In this 'assembly' the N50 is **60**

Assembly statistics for five Ptr isolates

Isolate	Location	Race	Tox	Assembler	QUAST			FunGAP	BUSCO
					N50 (bp)	contigs	N's	genes	%
90-2	Alberta/Sask	4	-	CLC Workbench12	170,424	2,220	~27K	13,005	99.4
90-2	Alberta/Sask	4	-	MEGAHIT (Shovill)	87,309	37,952	0	13,112	99.5
90-2	Alberta/Sask	4	-	SPAdes (Shovill)	287,769	3,872	0	13,011	99.5
90-2	Alberta/Sask	4	-	SOAPdenovo2	294,236	9,296	~16K	12,976	99.5
AB88-2	Alberta	2	A	CLC Workbench12	69,421	2,782	~18K	13,086	99.6
AB88-2	Alberta	2	A	MEGAHIT (Shovill)	46,933	34,163	0	13,045	99.7
AB88-2	Alberta	2	A	SPAdes (Shovill)	80,901	6,504	0	13,010	99.5
AB88-2	Alberta	2	A	SOAPdenovo2	82,444	4,374	~10K	12,885	99.5
ASC1	Manitoba	1	AC	CLC Workbench12	64,418	2,859	~10K	13,004	99.4
ASC1	Manitoba	1	AC	MEGAHIT (Shovill)	43,607	32,968	0	13,089	99.4
ASC1	Manitoba	1	AC	SPAdes (Shovill)	78,535	6,518	0	13,089	99.5
ASC1	Manitoba	1	AC	SOAPdenovo2	92,353	4,412	~15K	11,430	89.4
AZ35-5	Azerbaijan	5	B	CLC Workbench12	66,400	2,974	~25K	13,248	99.6
AZ35-5	Azerbaijan	5	B	MEGAHIT (Shovill)	43,709	38,454	0	Running	
AZ35-5	Azerbaijan	5	B	SPAdes (Shovill)	77,908	7,229	0	13,214	99.6
AZ35-5	Azerbaijan	5	B	SOAPdenovo2	79,529	5,661	~10K	13,127	99.6
I72-1	Syria	3	C	CLC Workbench12	54,919	2,663	~23K	12,893	99.6
I72-1	Syria	3	C	MEGAHIT (Shovill)	40,319	24,387	0	12,948	99.5
I72-1	Syria	3	C	SPAdes (Shovill)	63,650	6,744	0	12,886	99.6
I72-1	Syria	3	C	SOAPdenovo2	65,951	4,573	~11K	12,904	99.6

Proprietary software 'black box'

High contigs with MEGAHIT

Poor BUSCO annotation with SOAP

Gap inserts by CLC and SOAP

SPAdes assemblies

- Overall, SPAdes assemblies were of consistent quality across all isolates
- 90-2 had particularly good assembly
- 92-171 had a poor assembly
- Both possibly due to DNA quality

- Average genome size: 34.08 MB
- Average gene count: 13,141

- Two isolates have abnormally high gene count: I36-1 and T181-1
- Isolates may have supernumerary chromosomes, bacterial/yeast contam, horizontal gene transfers, other?

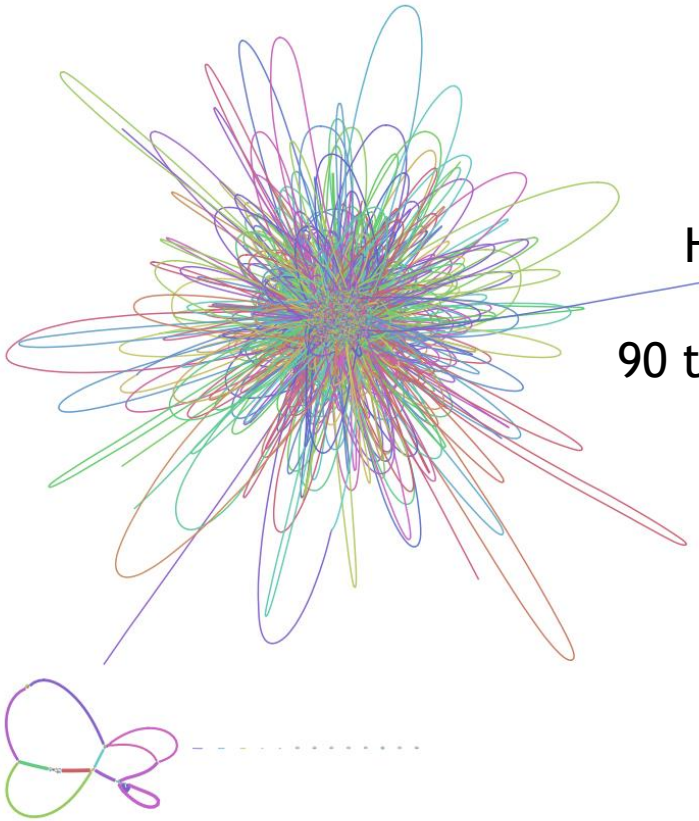
- I36-1, T181-1, and 92-171 **temporarily omitted** from downstream analysis until issues are resolved

Kraken2
Search raw reads for bacterial, viral, archea, protozoan DNA markers

Isolate	Race	HST	Year	Location	Size (MB)	Contigs	N50	Genes predicted by FunGAP
ASC1	1	AC	1990	Manitoba	33.77	6,518	78,535	13,089
I33-1	1	AC	2001	Azerbaijan	34.00	6,836	77,286	13,209
L3-1	1	AC	2016	Alberta	33.90	6,714	79,026	13,123
L4-1	1	AC	2016	Alberta	33.76	6,437	77,048	13,098
SW20-7	1	AC	2016	Saskatchewan	33.98	6,639	82,765	13,135
SW2-1	1	AC	2016	Saskatchewan	34.06	6,847	77,179	13,169
SW21-1	1	AC	2016	Saskatchewan	33.64	6,318	78,016	13,034
SW21-7	1	AC	2016	Saskatchewan	33.94	6,763	79,547	13,071
SW21-8	1	AC	2016	Saskatchewan	33.93	6,705	80,586	13,085
SW7-5	1	AC	2016	Saskatchewan	34.61	6,570	78,961	13,516
86-124	2	A	1990	Manitoba	33.86	6,717	80,673	13,055
AB88-2	2	A	2010	Alberta	33.84	6,504	80,901	13,010
L2-1	2	A	2016	Alberta	33.88	6,947	78,054	13,160
SW1-2	2	A	2016	Saskatchewan	34.05	6,701	77,323	13,060
SW15-1	2	A	2016	Saskatchewan	33.87	6,296	83,342	13,073
T132-2	2	A	2017	Tunisia	34.43	6,456	64,940	12,839
331-2	3	C	?	Manitoba	33.41	6,859	61,969	12,851
D308	3	C	1990	Manitoba	33.30	6,805	64,406	12,876
SC29-1	3	C	1999	Saskatchewan	33.21	6,558	65,638	12,898
SW21-5	3	C	2016	Saskatchewan	33.66	6,629	66,166	12,981
90-2	4	absent	2016	Alberta/Saskatchewan	34.62	3,872	287,769	13,011
				Alberta	35.76	7,953	137,311	13,101
				Tunisia	34.17	6,399	66,171	12,914
				Canada	34.34	15,018	48,800	13,276
				Algeria	33.36	6,113	81,155	12,943
				Algeria	34.44	8,094	78,975	13,176
				erbaijan	34.19	7,229	77,908	13,214
				erbaijan	33.26	6,569	64,475	12,875
				erbaijan	33.36	6,328	63,573	12,981
				erbaijan	33.25	6,582	65,477	12,881
I36-1	5	B	2001	Azerbaijan	36.97	6,657	74,120	15,130
AlgH1	6	BC	1995	Algeria	33.71	7,052	65,462	13,029
I72-1	6	BC	2001	Syria	33.32	6,744	63,650	12,886
I72-7	6	BC	2001	Syria	33.32	6,843	64,876	12,916
T176-2	7	AB	2017	Tunisia	34.74	6,978	63,367	13,088
T181-1	7	AB	2017	Tunisia	38.29	6,395	71,644	15,018
I34-1	8	ABC	2001	Azerbaijan	33.41	6,409	66,757	12,877
I35-18	8	ABC	2001	Azerbaijan	33.81	6,716	74,243	13,147
I73-1	8	ABC	2001	Syria	33.62	6,698	65,691	13,013
T128-1	atypical	B	2017	Tunisia	34.15	6,122	64,503	12,819

Hybrid assemblies using PacBio and Illumina reads

Assembly graph Illumina reads (SPAdes)



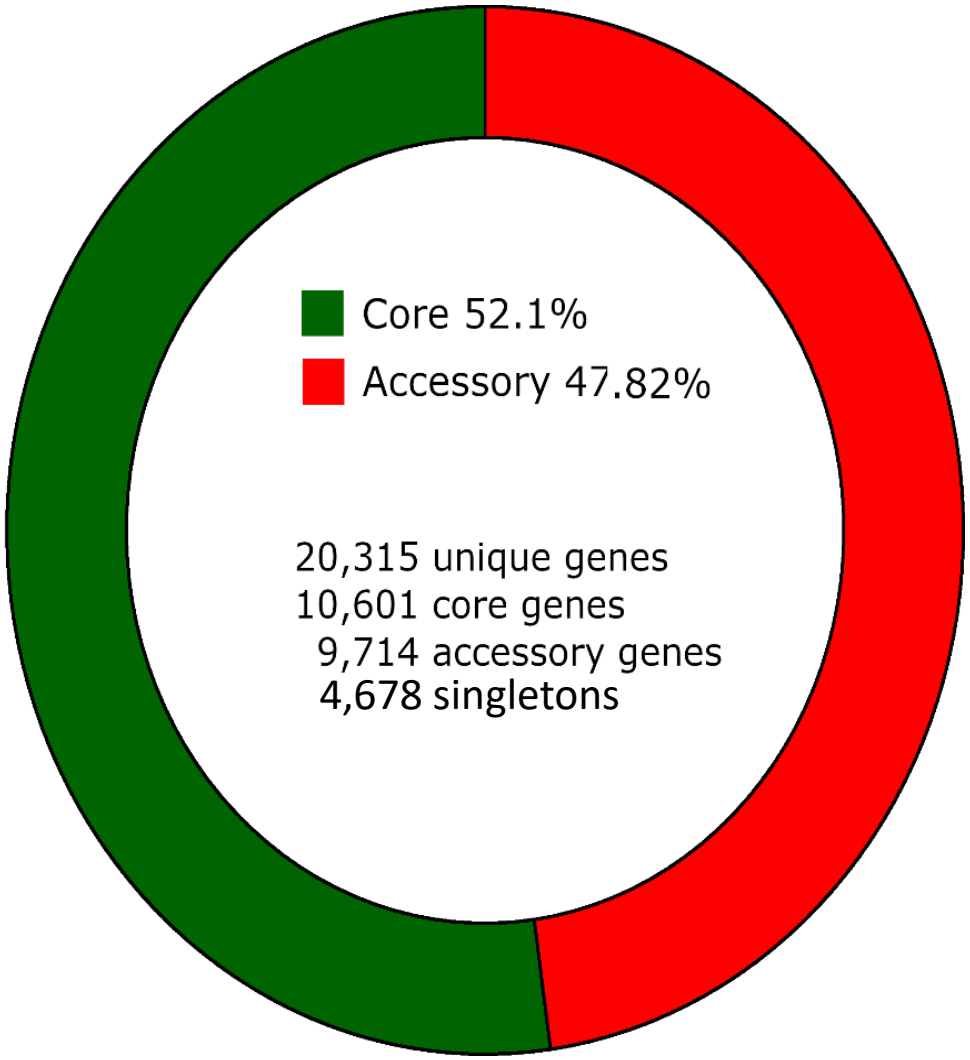
Mitochondrial DNA?

Huge difference in N50
(i.e. contig lengths)
90 thousand vs 3.6 million bp

Assembly graph PacBio reads (Flye)

Two isolates near chrm resolution
-70 contigs and 39 contigs
-Attempting to refine further

Pangenome of Ptr (38 isolates)

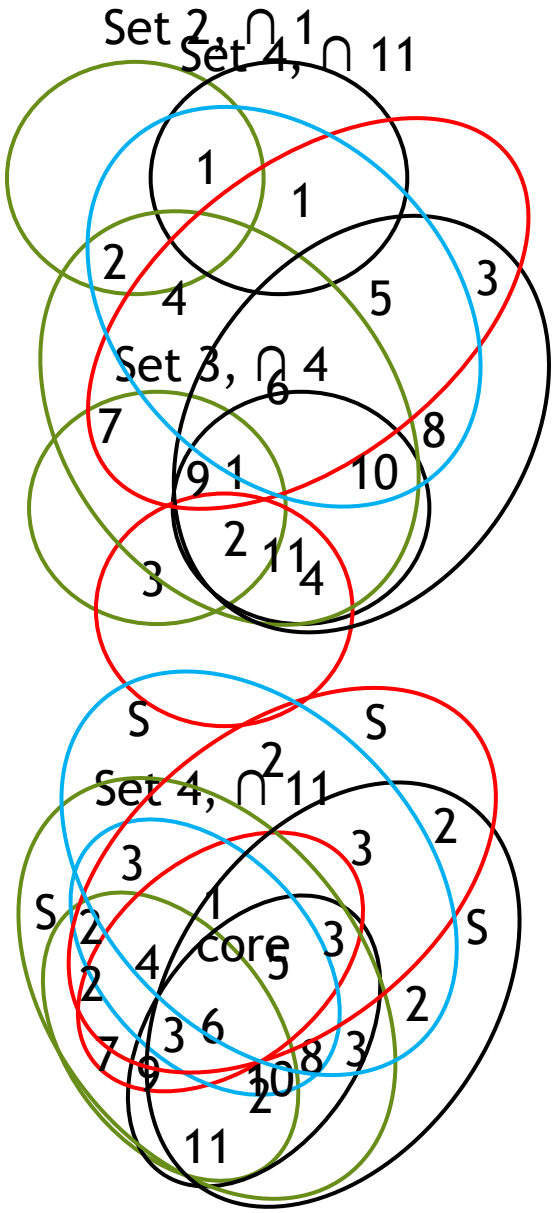


- Core = genes present in all isolates
- Accessory = genes present in some isolates
- Singletons = genes present in one isolate

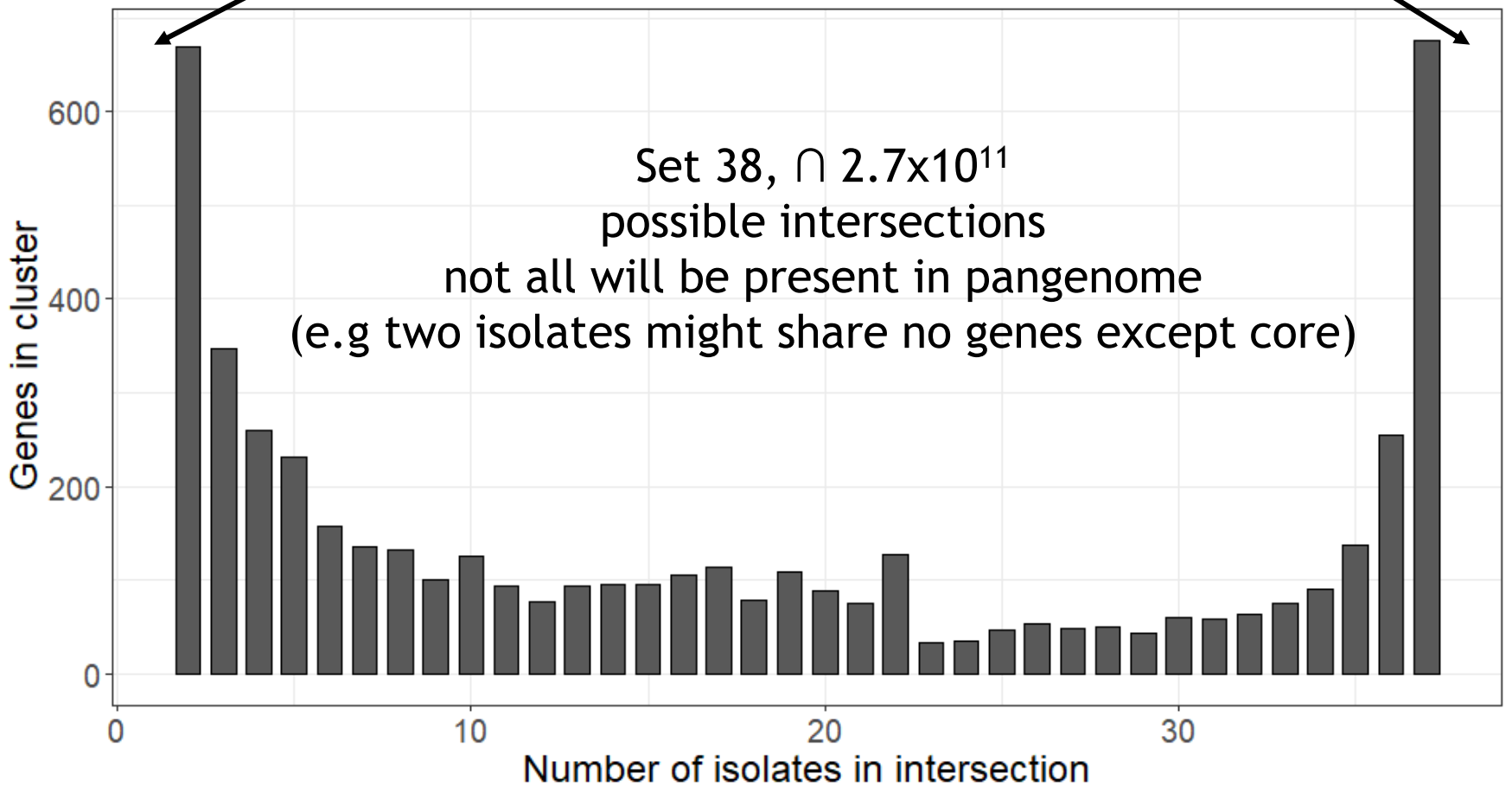
- Large accessory genome and singleton count
  - G9-4, 90-2, and T126-1 (race 4 non-path) contribute 2,049 genes (21%) to accessory
  - Of those genes, 95% are singletons and none were present in all three, suggesting unique survival strategies for non-pathogenic strains

  - 39 genes unique to isolates containing ToxA (22)
  - 1 gene unique to isolates containing ToxB (13)
  - 0 genes unique to isolates phenotyped as ToxC
  - 2 genes unique to race 3 [C] isolates (4)
  - 1 gene unique to race 8 [ABC] isolates (3)
  - 42 genes unique to novel isolate T128-1

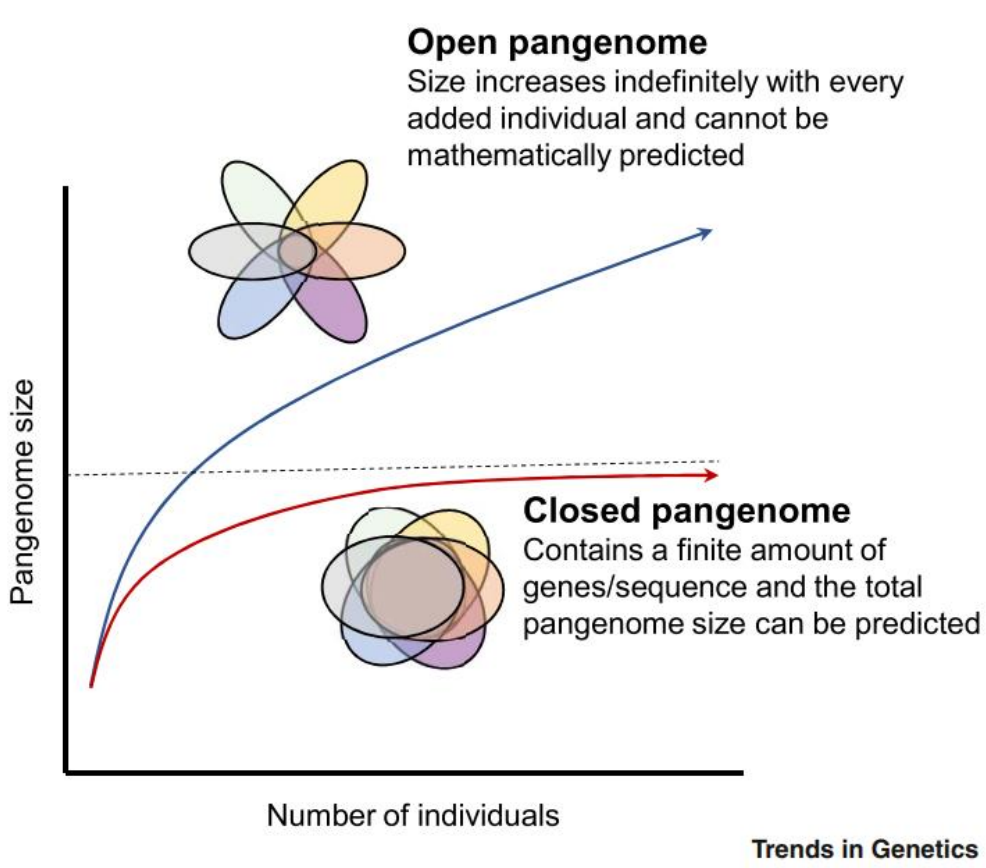
Pangenome of Ptr (38 isolates)

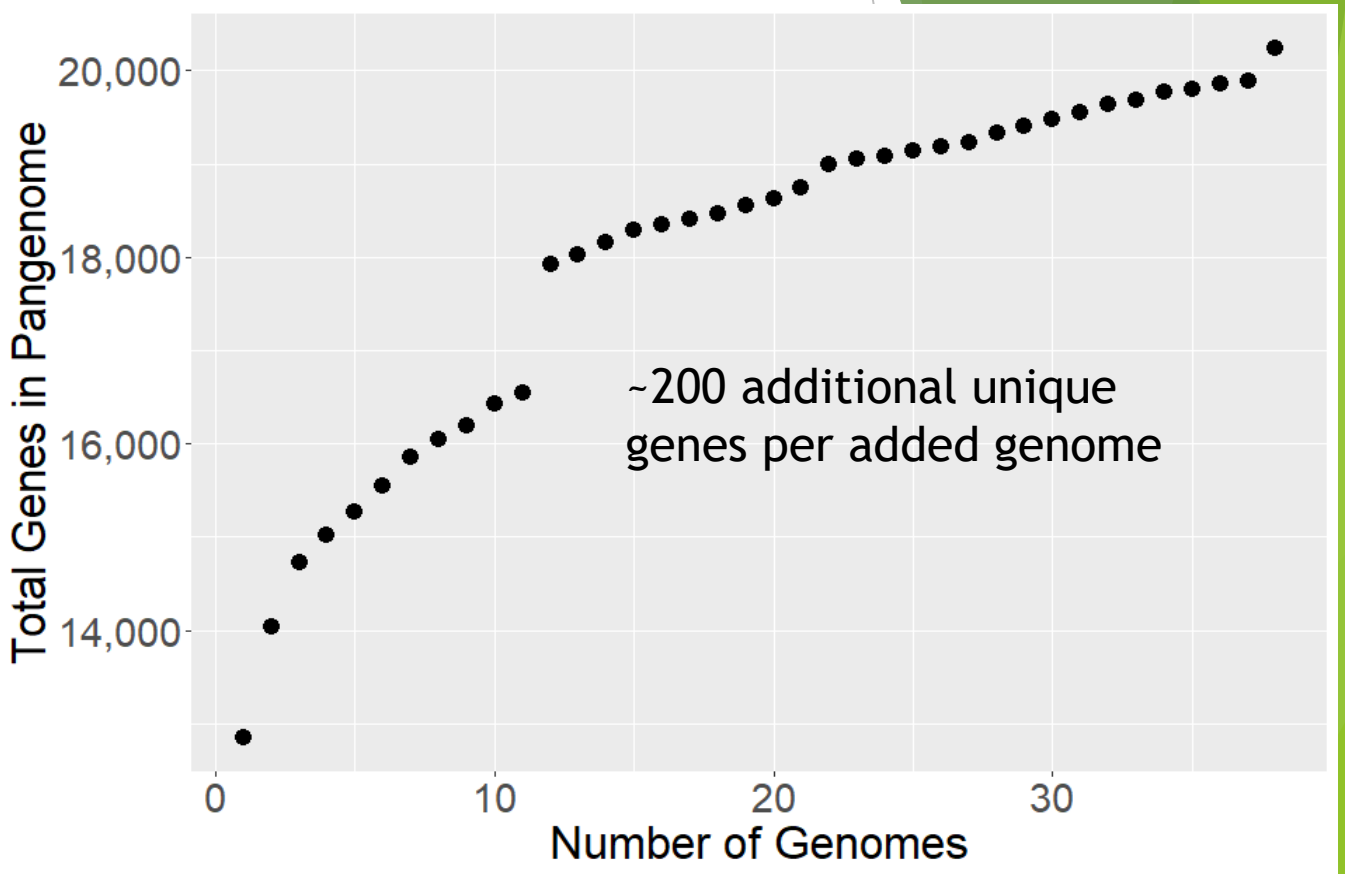


singletons and core not shown



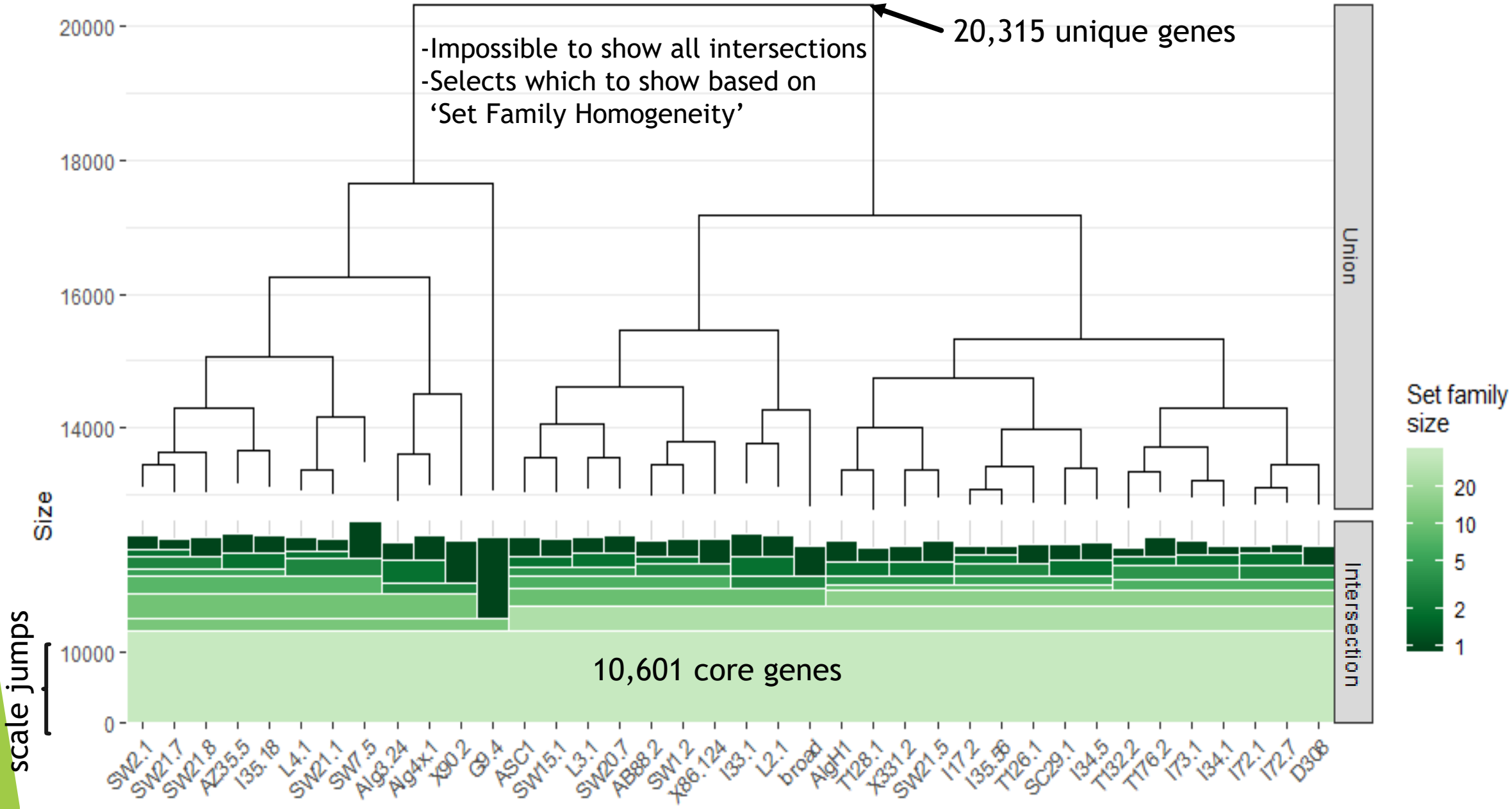
Pangenome of Ptr (38 isolates)



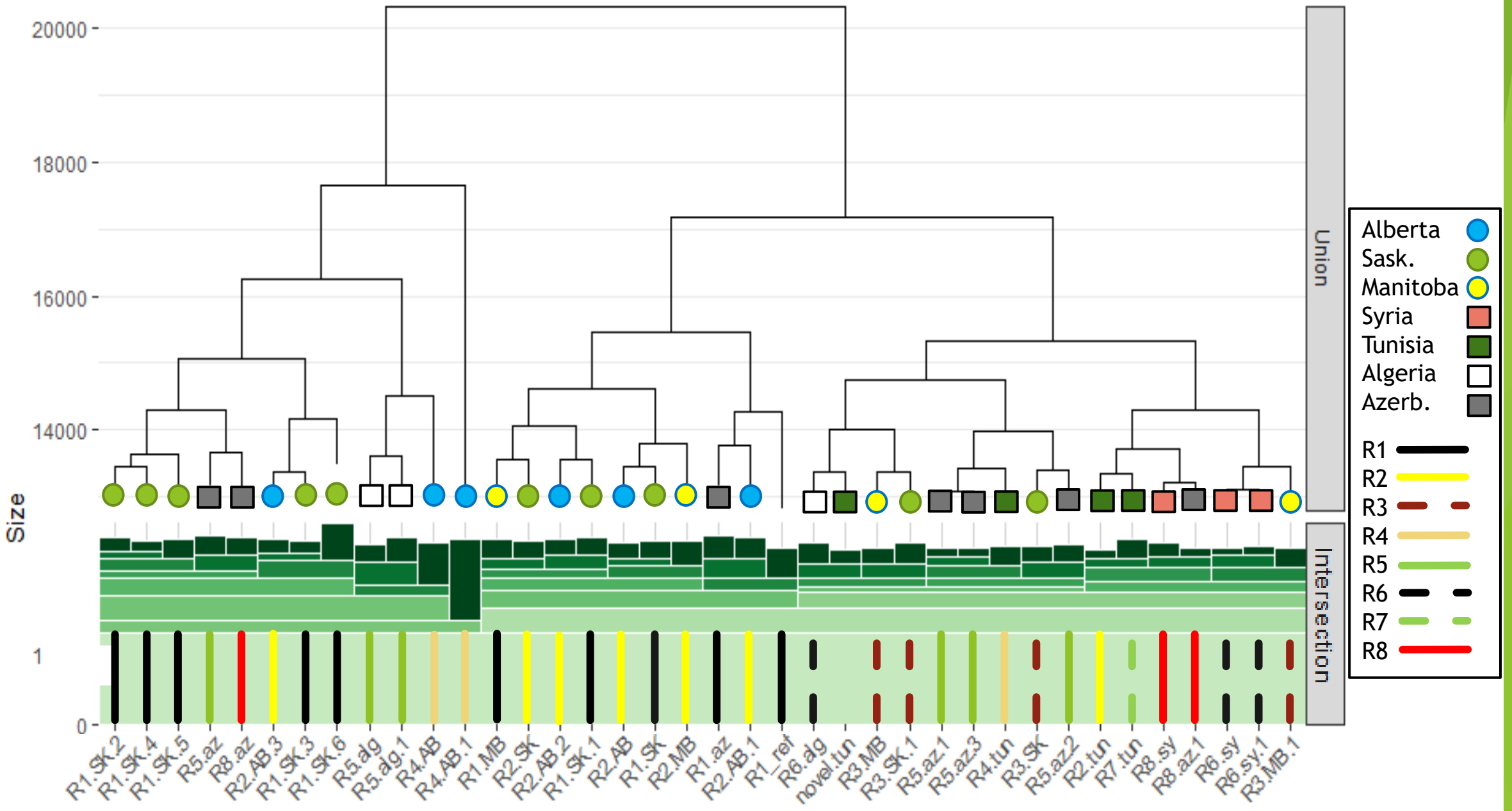


- Open pangenomes continually expand their total gene set
  - Mixed microbial communities
  - Multiple methods of exchanging genetic material
  - Wide-spread/cosmopolitan (i.e. not isolated population)

Hierarchical set clustering based on gene presence/absence

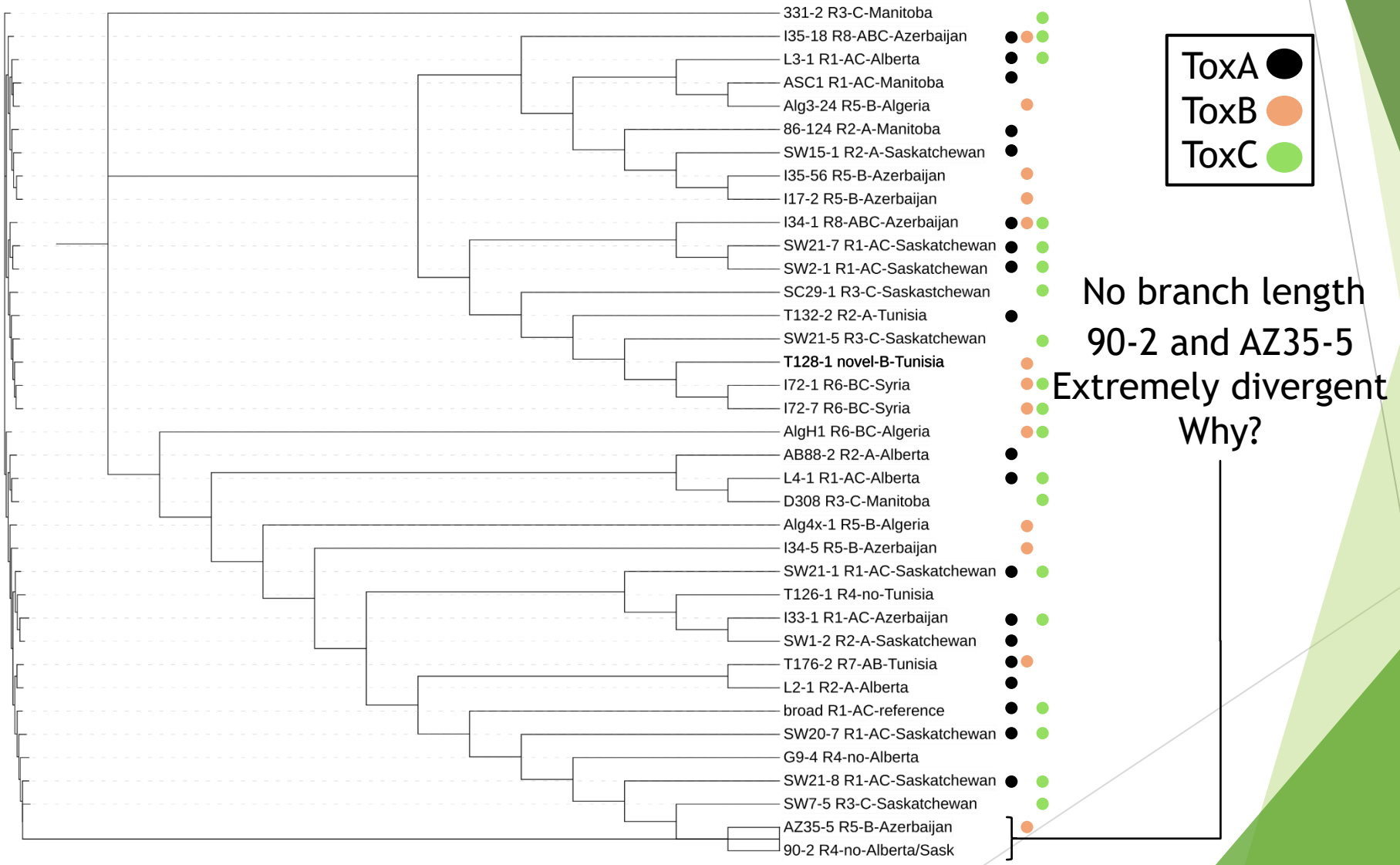


Hierarchical set clustering based on gene presence/absence



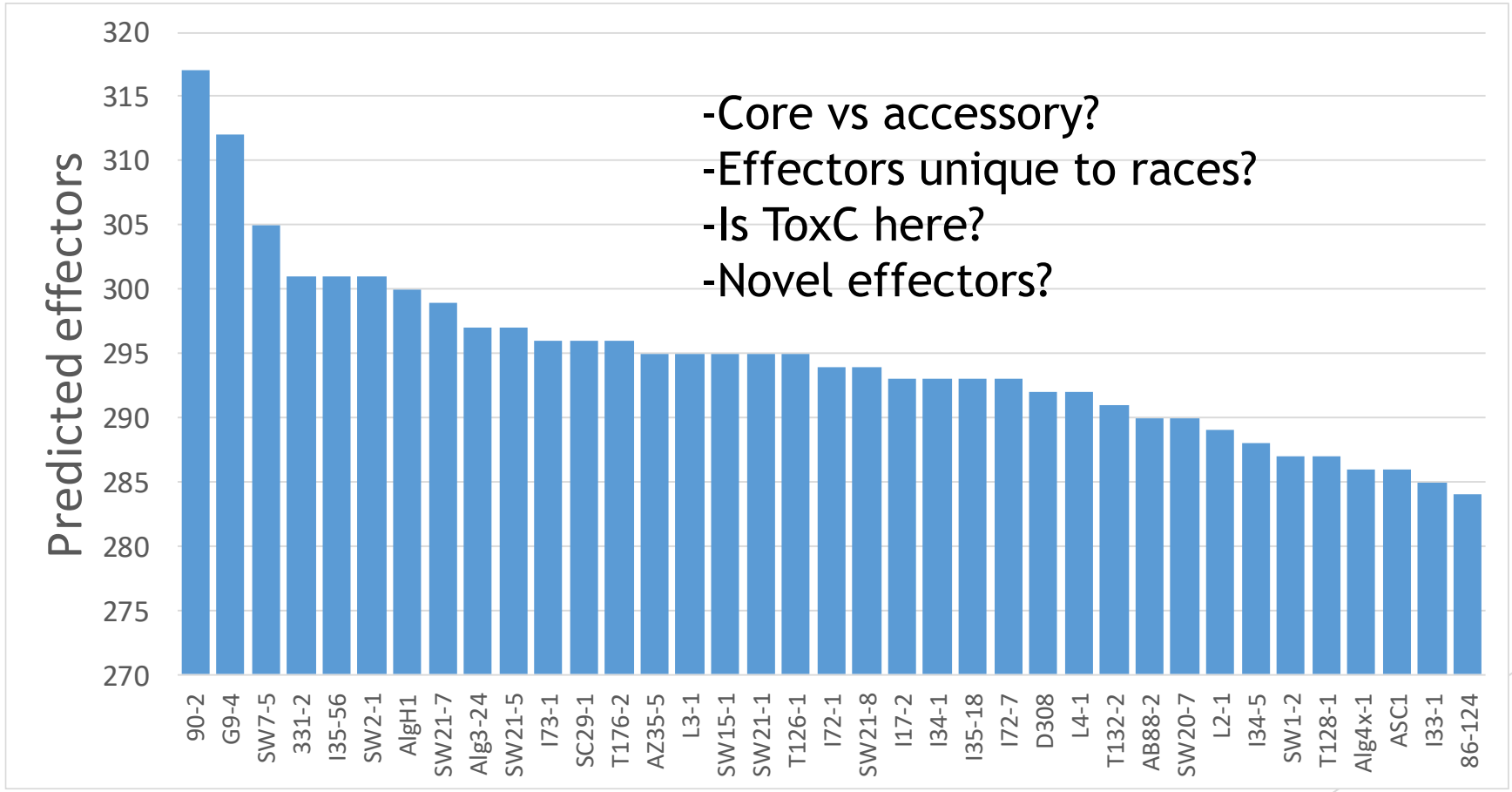
Core genome phylogeny

-For each of 10,601 core genes performed individual alignments (MUSCLE)
-Concatenated all alignments and used for Maximum Likelihood phylogeny (RAxML)



Identification of effector candidates

-Gene sets checked for presence of signal peptides and transmembrane domains (Phobius)
-Filtered for genes with SP but no TD
-Machine learning algorithm trained to identify effectors (EffectorP)
-Functional annotations from Pfam database for candidates



Remaining analysis

- Finalize hybrid assemblies (ideally to chromosome level)
- Refine phylogenetic analysis
  - Combine hierarchical sets with core-phylogeny
  - Cluster by race? Location? HST presence?
  - Other interesting intersections
- Chromosomal locations of core and accessory genes
- Chromosomal rearrangements
- Supernumerary chromosomes
- Refine effector search
- Transposable element content
- CRE's [i.e. non-coding region analysis]
- Gene regulatory network (especially for accessory genes)
- Horizontal gene transfers?
- More!!
- Manuscripts

Acknowledgements

- Therese Despins
- HPC Biocluster Team
- Local IT Department

Agriculture and Agri-Food Canada

SaskWheat DEVELOPMENT COMMISSION

Agriculture Funding Consortium

Alberta Wheat COMMISSION

Author contributions

RG – SOAPdenovo2, MEGAHIT, and SPAdes assemblies; annotations; pan-genome; phylo; effector

ROP – SPAdes assemblies; hybrid assemblies; technical support

KG – conceived project; DNA extraction; CLC assemblies

MH – DNA extraction

RR – will help with analysis downstream

SS – isolates

FD – isolates

RA – conceived project; oversight; funding

References

SPAdes: Bankevich et al., 2012. *Journal of Computation Biology*, 19(5), 455-477
Shovill: Seemann, 2019. github.com/tseemann/shovill
MEGAHIT: Li et al., 2015. *Bioinformatics*, 31(10), 1674-1676
SOAPdenovo2: Luo et al., 2012. *Gigascience*, 1(1), 18
CLC: Qiagen, 2018
Fungap: Min et al., 2017. Bioinformatics 33(18), 2936-2937
Flye: Kolmogorov et al., 2019. *Nature Biotechnology* 37(5), 540–546;
Lin et al., 2016. *Proceedings of the National Academy of Sciences*, 113(52), E8396-E8405
Pylon: Walker et al., 2014. *PloS One*, 9(11), e112963
RNA: Moolhuijen et al., 2018. *BMC Research Notes*, 11(1), 907-909
BUSCO: Simão et al., 2015. *Bioinformatics* 31(19), 3210-3212
Pangloss: McCarthy & Fitzpatrick, 2019. *Genes* 10(7), 521
Broad: Manning et al., 2013. *G3* 3(1), 41-63
Hierarchical Sets: Pedersen 2016. *github.com/thomasp85/hierarchicalSets*
Phobius: Käll et al., 2004. *Journal of Molecular Biology 338(5), 1027-1036*
EffectorP: Sperschneider et al., 2018. *Molecular Plant Pathology*.
MUSCLE: Edgar, 2004. *Nucleic Acid Research 32(5), 1792-1797*
RAxML: Stamatakis, 2014. *Bioinformatics 30(9), 1312-1313*
Kraken2: Wood et al., 2019. *Genome Biology 20*