Inferring Viral Transmission Using Closely Related Virus Genomes

Mauriana Pesaresi Seminar Series 2021/2022

Njagi Mwaniki

Fri 8 April 2022

Introduction

Problem Statement

What

- Determine who infects whom within a household

Why

- Better prepare for imminent viruses
- Understand seasonality of viruses
- Protect the vulnerable

How

- Sample each individual often
  - nasopharyngeal swabs twice a week
- Perform an all vs all comparison of virus genomes
  - Subtext: genome comparison is not all vs all

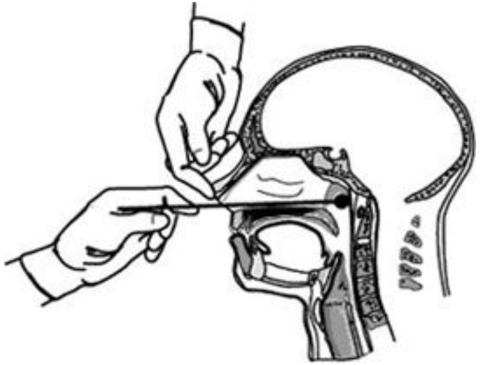
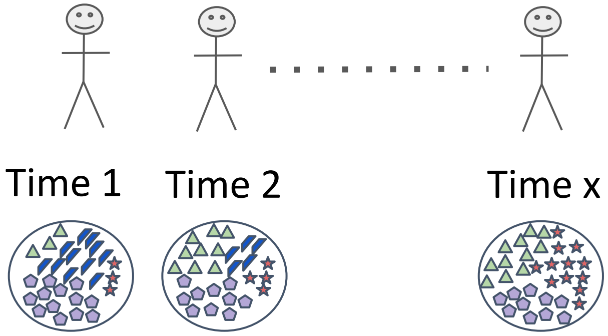


Image credit: cdc.gov - https://www.cdc.gov/pertussis/clinical/diagnostic-testing/specimen-collection.html, Public Domain, https://commons.wikimedia.org/w/index.php?curid=38360525

Virus Quasispecies

A viral infection is composed of multiple variants that change over time

within an individual

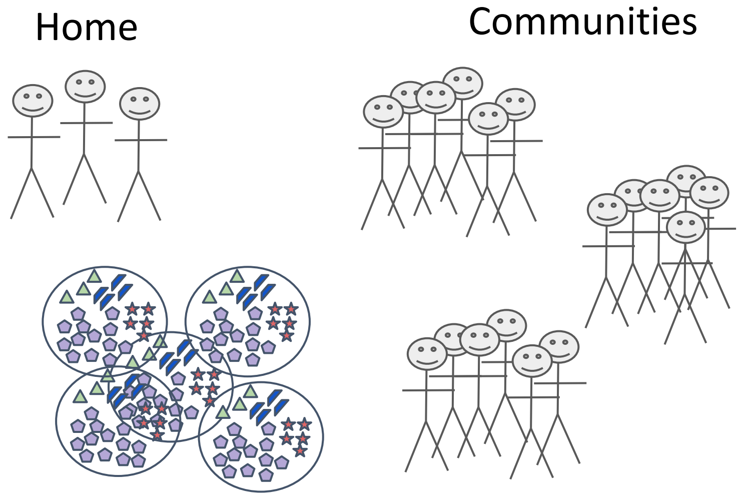


Time 1

Time 2

Time x

between individuals



Home

Communities

Seasonality of Viruses

Seasonality

- Predictable pattern in infections based on annual changes in the environment
- Expected in viruses transmitted through respiratory droplets

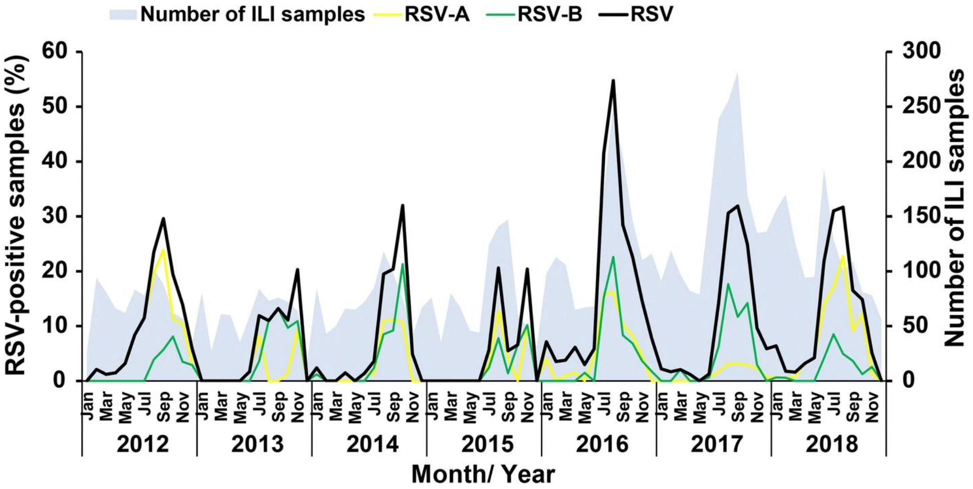
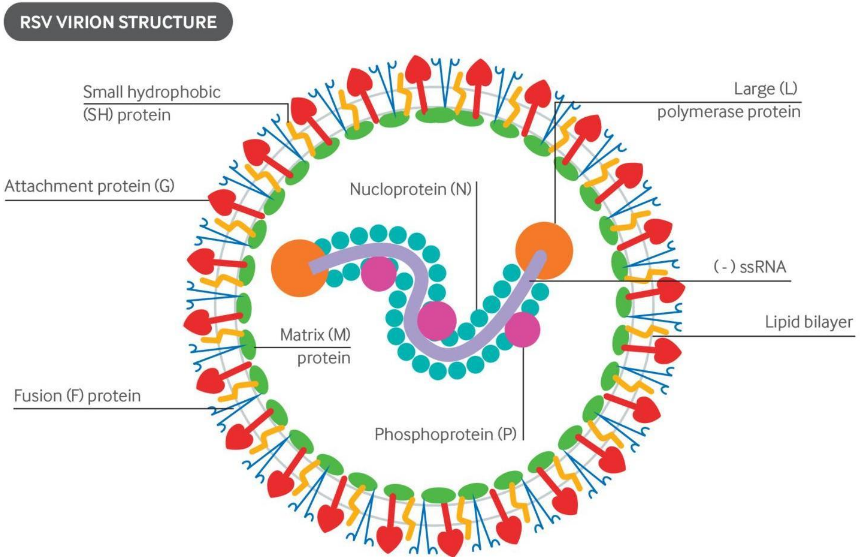


Image credit (Thongpan et al., 2020)

Viruses (RSV)



Non-structural proteins: NS1, NS2

Nucleocapsid and regulatory proteins: N (nucleoprotein), P (phosphoprotein), M2.1, M2.2, L (large polymerase)

Envelope proteins: SH (small hydrophobic), G (attachment), F (fusion)

Inner envelope protein: M (matrix)

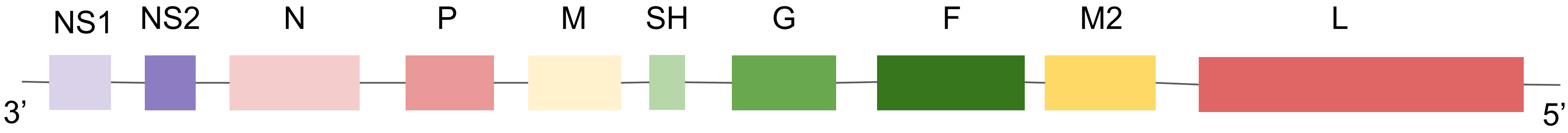
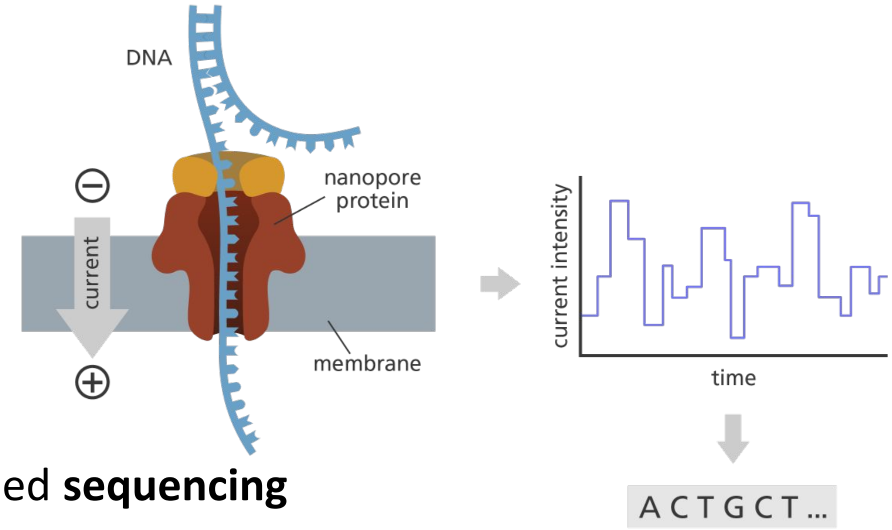


Image credit and genome adapted from from (Nam & Ison, 2019).

What is DNA data?

DNA is a chain (strand) of molecules

- Adenine (A)
- Guanine (G)
- Thymine (T)
- Cytosine (C)



DNA data is extracted through a process called **sequencing**

- modern methods run a DNA strand through a pore

Result is a **string** of A, T, C and Gs

Image credit: Genome Research Limited.

Pan-genome

A collection of many genomes

- Pan (many) genome (genetic material of an organism)

Map newly sequenced samples to a pangenome to determine similarity

- Map: find the closest genome to a newly sampled genome

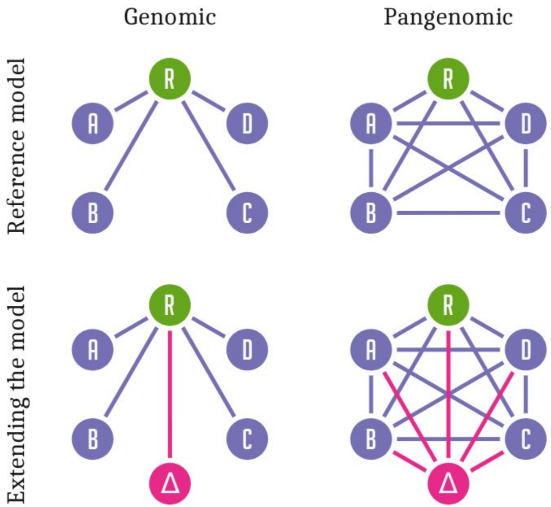
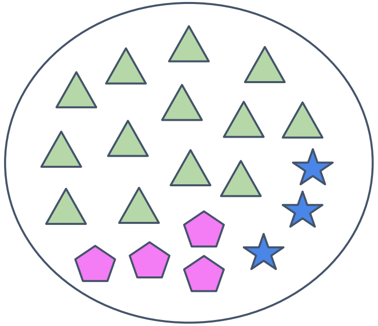


Image credit: (Eizenga et al., 2020)

Comparing Genomes

Linear Reference Genome

A reference genome is a consensus of the most abundant sequence and is traditionally linear



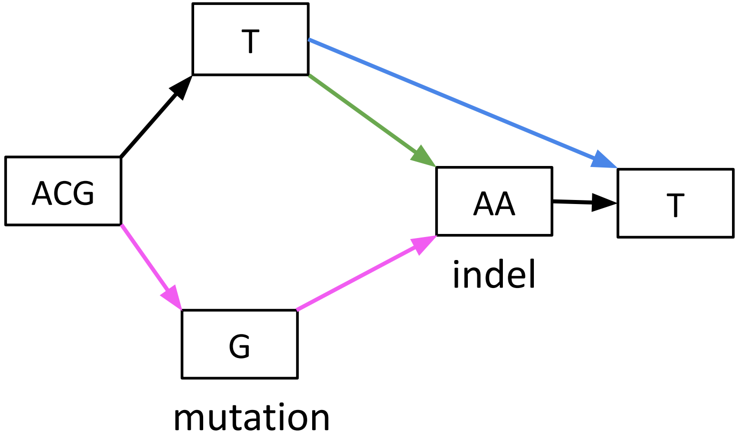
	1	2	3	4	5	6	7
Variant 1	A	C	G	T	A	A	T
Variant 2	A	C	G	T	-	-	T
Variant 3	A	C	G	G	A	A	T
Consensus	A	C	G	T	A	A	T

Pangenomic (Graphical) Reference Genome

The same consensus from the previous slide

	1	2	3	4	5	6	7
Variant 1	A	C	G	T	A	A	T
Variant 2	A	C	G	T	-	-	T
Variant 3	A	C	G	G	A	A	T
Consensus	A	C	G	T	A	A	T

The three sequences presented as a graph



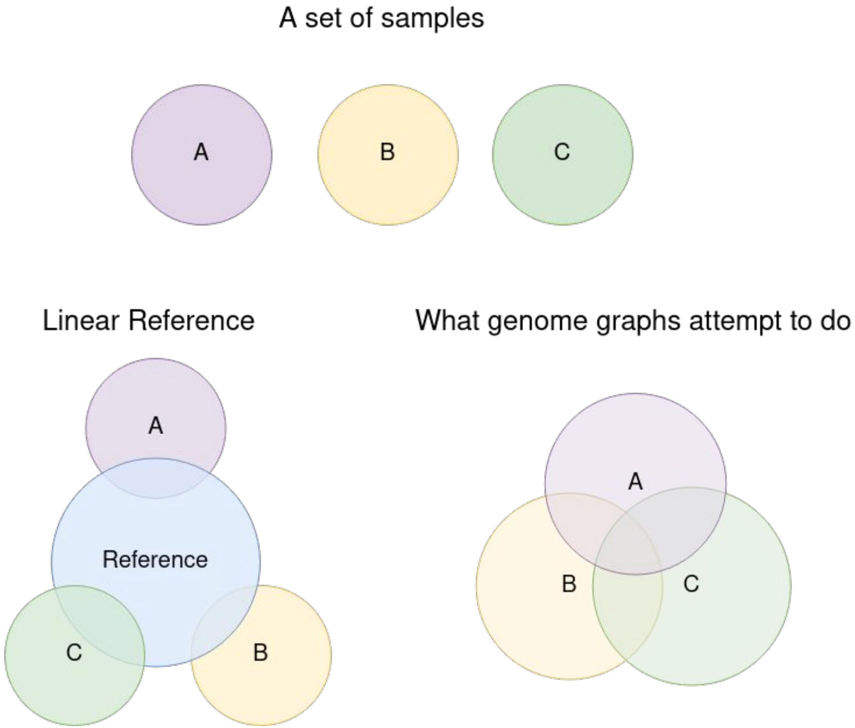
Genomic Sample Comparison

Genomic sample similarity or difference is based on how samples compare against a reference

Linear reference

- less precise
- reference bias
- non-transitive (1 < 2 < 3 therefore 1 < 3)

Increase precision by comparing against a graphical reference



A set of samples

Linear Reference

What genome graphs attempt to do

Alignment, Mapping and Variant Calling

Align (edit distance)

- maximize similarity between two strings
  - Reward matches
  - Penalize mismatches, gaps, and gap extensions

Map

- find the **location** of a query in a text (reference)

Variant calling

- assemble reads into contigs
- map reads to a reference
- a variant is called when the number of reads with differences against the reference meets a set threshold

Align DROWN and GOWN

D	R	O	W	N
G	-	O	W	N

Align TTAAT and TT

T	T	A	A	T
T	T			
	T	T		
	T	-	-	T

Mapping Against a Linear Reference

Mapping 3 reads to a linear reference

- TAAT
- TT
- GAAT

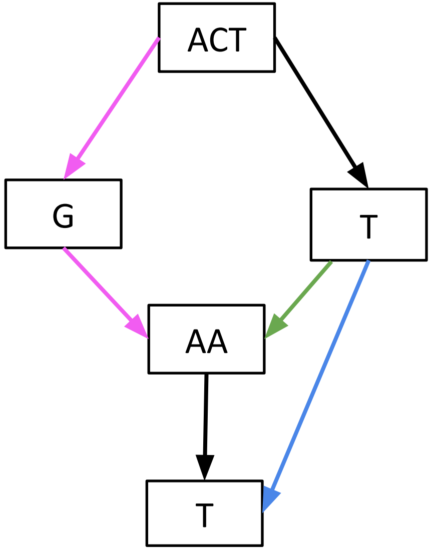
Linear reference	A	C	G	T	A	A	T
Read 1				T	A	A	T
Read 2			T	T			
Read 3				G	A	A	T

A	C	T	T	A	A	T	Variant 1
A	C	G	T	T			Variant 2
A	C	T	G	A	A	T	Variant 3

Mapping Against a Graphical Reference

Mapping 3 reads to a graphical reference

- TAAT
- TT
- GAAT



	Graphical Reference					
1	A					
2	C					
3	G					
4	T	G		T	T	G
5	A		-	A	-	A
6	A		-	A	-	A
7	T			T	T	T

Linear vs Graphical Reference Mapping

Compare the following sequences to a reference:

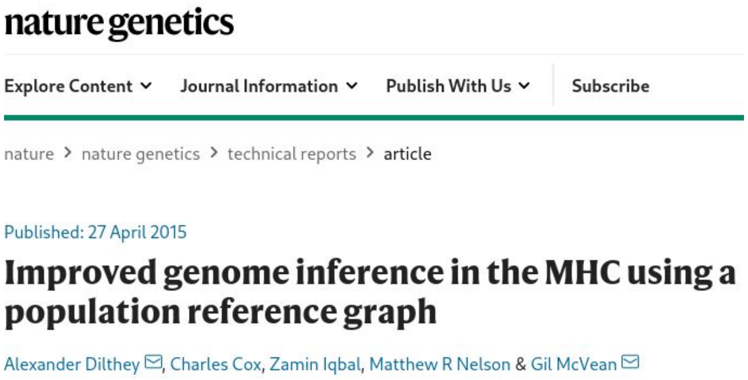
- TT
- TAAT
- GAAT

	1	2	3	4	5	6	7
Linear Reference	A	C	G	T	A	A	T
Read 1				T	A	A	T
Read 2			T	T			
Read 3				G	A	A	T

			Graphical Reference			
					A	1
					C	2
					T	3
T	T	G		G	T	4
A	-	A	-		A	5
A	-	A	-		A	6
T	T	T			T	7

Reference Bias

An increased likelihood of tools and methods used for read mapping to fail to identify variation and over-report variants present in the reference compared to the variants that are not present in the reference.



nature genetics

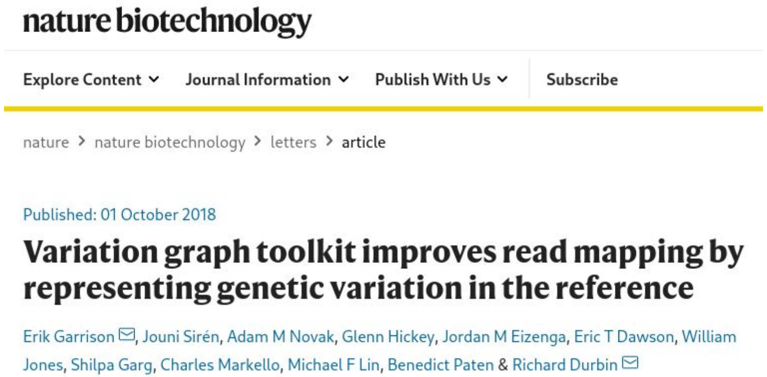
Explore Content ⌄ Journal Information ⌄ Publish With Us ⌄ Subscribe

nature › nature genetics › technical reports › article

Published: 27 April 2015

Improved genome inference in the MHC using a population reference graph

Alexander Dilthey ✉, Charles Cox, Zamin Iqbal, Matthew R Nelson & Gil McVean ✉



nature biotechnology

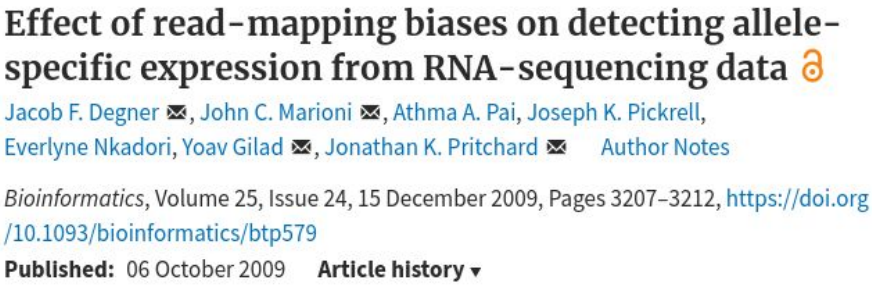
Explore Content ⌄ Journal Information ⌄ Publish With Us ⌄ Subscribe

nature › nature biotechnology › letters › article

Published: 01 October 2018

Variation graph toolkit improves read mapping by representing genetic variation in the reference

Erik Garrison ✉, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten & Richard Durbin ✉



Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data

Jacob F. Degner ✉, John C. Marioni ✉, Athma A. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad ✉, Jonathan K. Pritchard ✉ Author Notes

Bioinformatics, Volume 25, Issue 24, 15 December 2009, Pages 3207–3212, https://doi.org/10.1093/bioinformatics/btp579

Published: 06 October 2009 **Article history** ▾

Analysis

Coverage

How many substrings match a section of a genome

- Caused by similarity

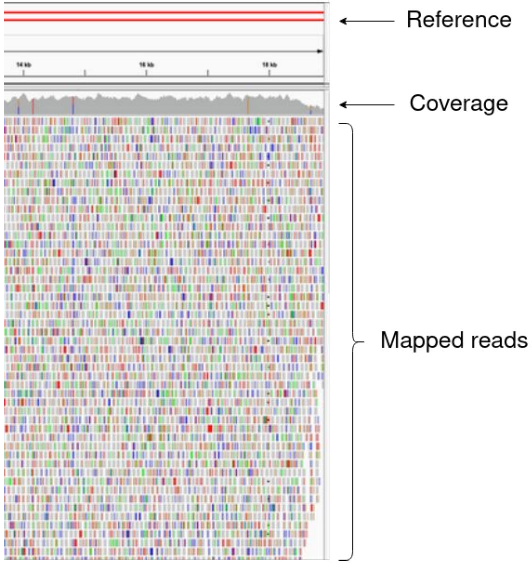
Find ***aa***, ***bb***, ***aa***, ***baa***, ***aba***, ***baba*** in ***abbaababa*** exactly

0	1	2	3	4	5	6	7	8
a	b	b	a	a	b	a	b	a
						a	b	a
					b	a	b	a
			a	a				
	b	b						
		b	a	a				
				a	b	a		

Linear vs Graphical Reference Mapping

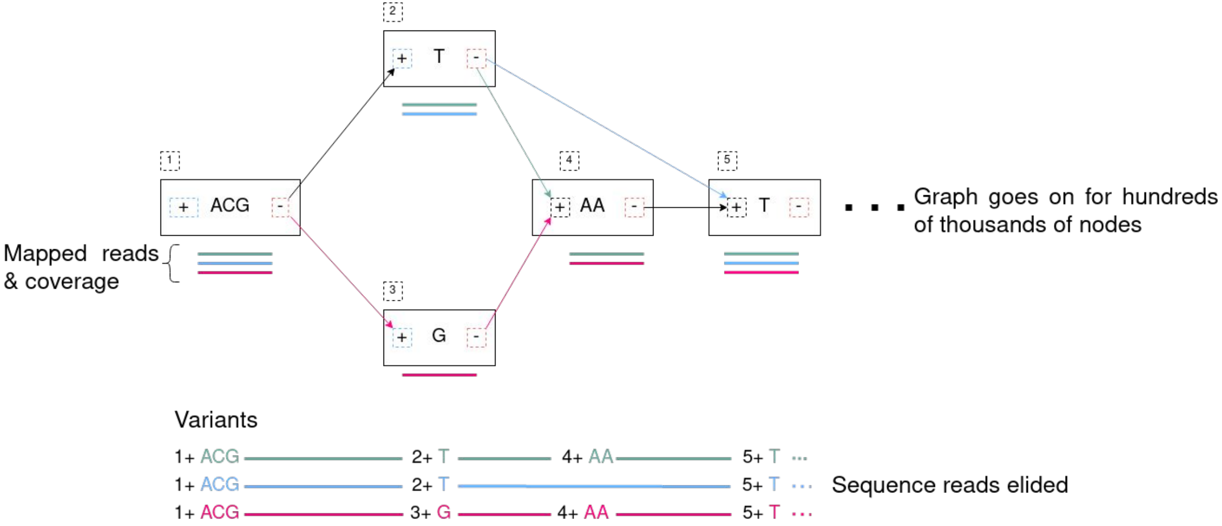
Linear Reference

Reads stack up below the loci they map against



Graphical Reference

Reads stack up below the nodes whose sequences they map against



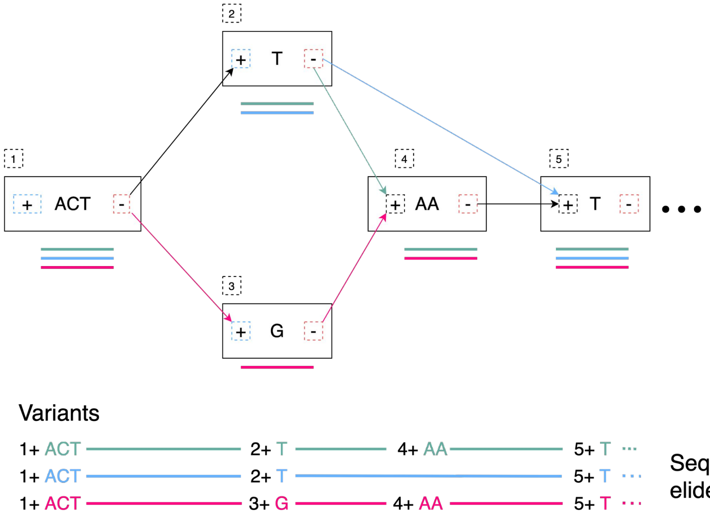
. . . Graph goes on for hundreds of thousands of nodes

Sequence reads elided

Coverage Vector and Coverage Statistics

Graphical Reference

Reads stack up below the nodes whose sequences they map against



Graph nodes elided

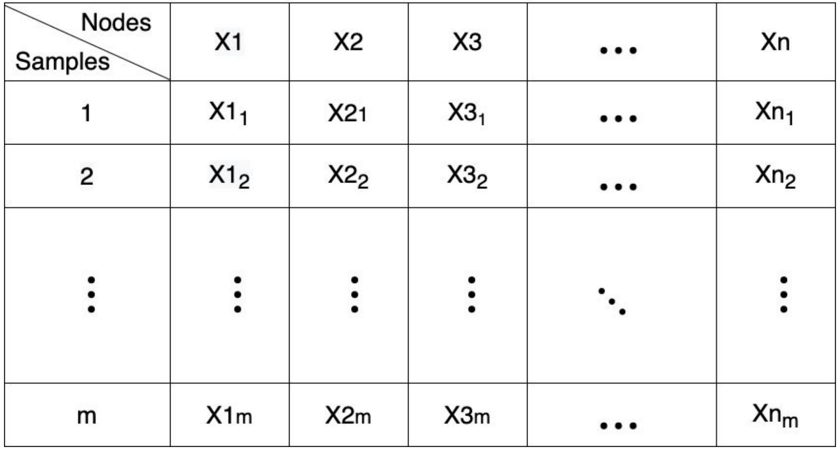
Variants

1+ ACT 2+ T 4+ AA 5+ T ···

1+ ACT 2+ T 5+ T ···

1+ ACT 3+ G 4+ AA 5+ T ···

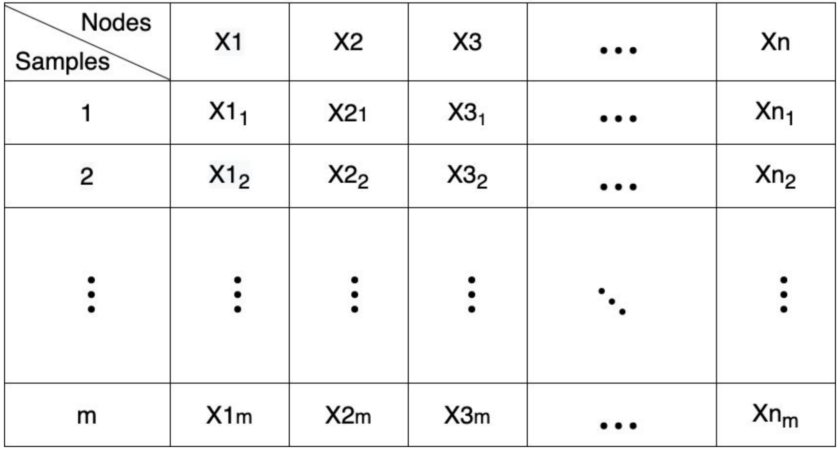
Sequence reads elided



Nodes / Samples	X1	X2	X3	• • •	Xn
1	$X1_1$	$X2_1$	$X3_1$	• • •	Xn_1
2	$X1_2$	$X2_2$	$X3_2$	• • •	Xn_2
⋮	⋮	⋮	⋮	⋱	⋮
m	X1m	X2m	X3m	• • •	Xn_m

Pairwise Distances

Coverage Statistics



Nodes Samples	X1	X2	X3	…	Xn
1	$X1_1$	$X2_1$	$X3_1$	…	Xn_1
2	$X1_2$	$X2_2$	$X3_2$	…	Xn_2
⋮	⋮	⋮	⋮	⋱	⋮
m	$X1_m$	$X2_m$	$X3_m$	…	Xn_m

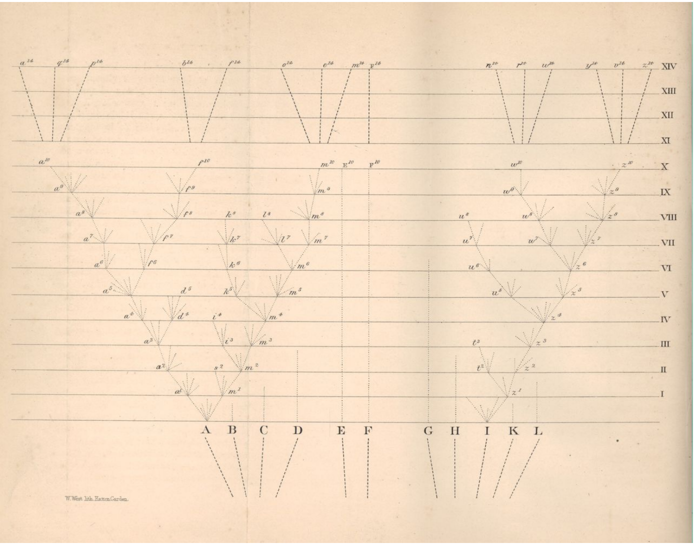
Pairwise Distances (Euclidean)

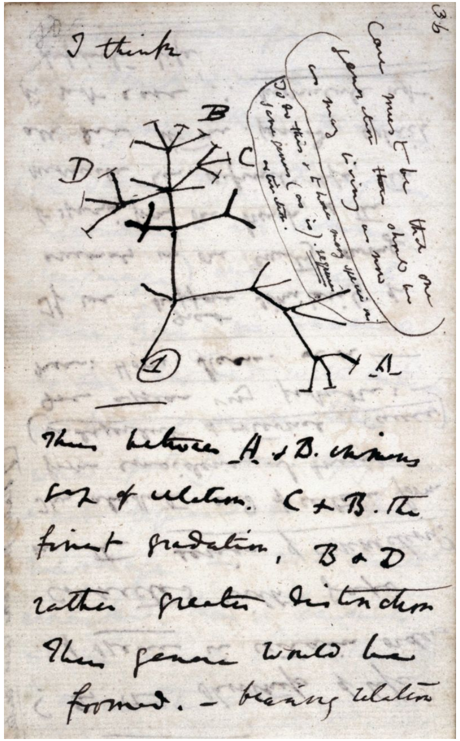
$$\sqrt{(x1_1 - x1_2)^2 + (x2_1 - x2_2)^2 + \ldots + (xn_1 - xn_2)^2}$$

Phylogenetics and Phylogenetic Trees

Evolutionary history and relationships between or within groups of organisms

- Analog: the tree of life





I think

Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is). requires extinction.

Thus between A & B. immense gap of relation. C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed. – bearing relation

Image credit: wikipedia

Phylogenetics and Phylogenetic Trees

Evolutionary distance

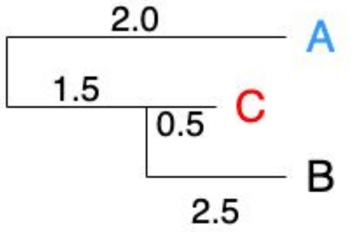
- A measure of genetic difference (mutation)

A & C are more similar

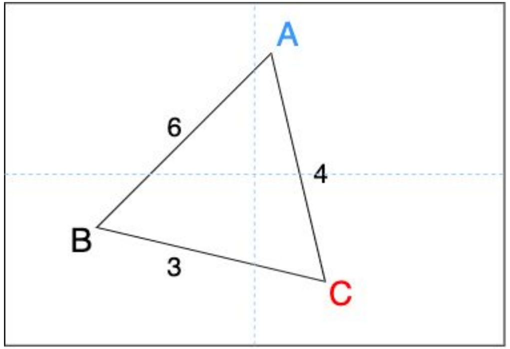
B & C are more similar

A & B are further apart from each other but closer to C

A to C	2.0+1.5+0.5 = 4
A to B	2.0+1.5+2.5 = 6
C to B	0.5+2.5 = 3



Neighbour Joining

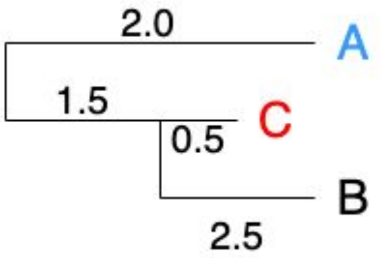


	A	B	B
A	0	6	4
B	6	0	3
C	4	3	0

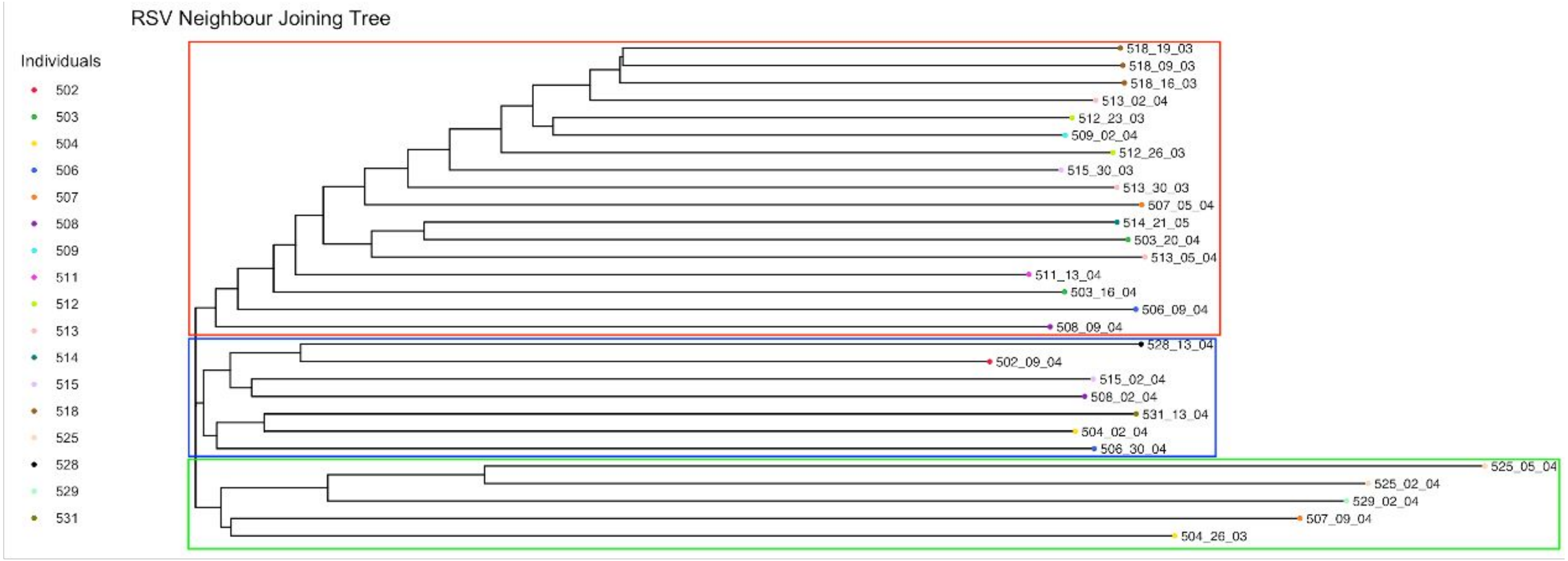
A to C 2.0+1.5+0.5 = 4

A to B 2.0+1.5+2.5 = 6

C to B 0.5+2.5 = 3



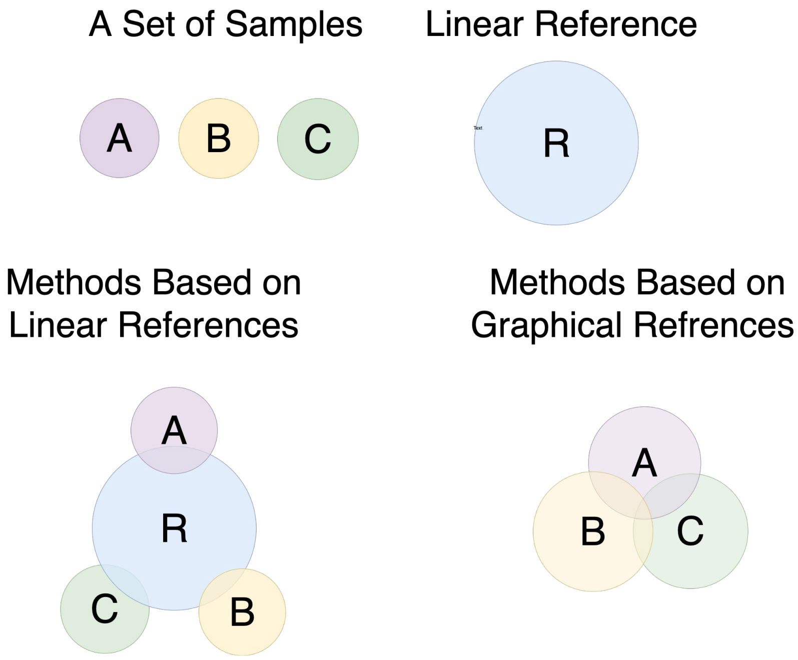
Neighbour Joining (Example)



Conclusion and Further Work

Conclusion

- A pangenomic reference can be used for sample comparison
- It is possible to perform all versus all sample comparison by comparing the coverage of reads to a graph



A Set of Samples

Linear Reference

Methods Based on Linear References

Methods Based on Graphical Refrences

Conclusion

- The linear reference loses information
- A pan-genome representation is comprehensive
- By comparing new samples to a pan-genome we can
  - compare samples better
  - infer transmission with a higher degree of certainty
  - better prepare for imminent variants

Conclusion

Assumptions

- Sample relatedness implies potential transmission
- Minimum evolution (neighbour joining assumes minimum evolution)
- Bases have an equal probability of substitution (JC69 model)
- Coverage under each node is equally informative

Further work

- Extend the approach to other households
- Extend the approach to more sparse testing time periods
- Use more robust phylogenetic methods than neighbour joining

Acknowledgements

- Dr George Githinji
- Dr Pjotr Prins
- Prof James Nokes

- Prof Erik Garrison
- Prof Santie de Villiers

- Prof Sam Kinyanjui
- Dr Dorcas Mbuvi
- Liz Murabu
- Rita Baya
- Florence Kirimi

Pwani UNIVERSITY

IDeAL
Initiative to Develop African Research Leaders

KEMRI | Wellcome Trust