EXPLORING VIRUS SEQUENCE DIVERSITY USING VARIATION GRAPHS

Njagi Mwaniki

A research proposal submitted in partial fulfilment of the requirements for the award of a Master of Science Degree in Bioinformatics of Pwani University.

June 9, 2020.

1 Declaration

This proposal is my original work and has not been presented for a degree in any other university or any other award.

Moses Njagi Mwaniki
Date ..... Signature .....

I confirm that the work reported in this thesis was carried out by the candidate under my supervision.

Dr George Githinji
Date ..... Signature .....

Dr Pjotr Prins
Date ..... Signature .....

Prof James Nokes
Date ..... Signature .....

Abstract

There are several methods that are used to evaluate and describe sequence diversity, and these include Shannon's entropy, the number of polymorphic loci per kilobase (one thousand bases), and the nucleotide diversity statistic Pi (π). These methods compare nucleotide substitutions, insertions and deletions present in consensus sequences to describe and quantify sequence diversity from a given sample. However, such approaches are prone to underestimation of the actual diversity for example in cases of low abundance haplotypes given that consensus sequences are a mosaic of closely related haplotypes.

We propose to use variation graphs, that is, data structures that maintain available sequence variation from a sample or a collection of samples, to explore the sequence variation of respiratory syncytial virus (RSV), a single stranded, negative sense, enveloped RNA virus. This study will utilize sequence data from samples collected in a twenty-member household during the course of a household RSV outbreak.

We aim to create a pipeline for constructing a variation graph for describing virus diversity during the household outbreak. We will use this data to assess the utility of this pangenome in informing potential transmission events.

Contents

List of Figures

2 Introduction

2.1 Background Information

RSV was first isolated in Chimpanzees in 1956 and named Chimpanzee Coryza Agent (Morris, Blount, & Savage, 1956). A year later, in 1957, it was isolated in children from whom it had not been possible to isolate and renamed Respiratory Syncytial Virus (Beem, Wright, Hamre, Egerer, & Oehme, 1960; Chanock, Roizman, & Myers, 1957; Zlateva, Lemey, Moes, Vandamme, & Van Ranst, 2005).

During the first year of life, RSV is the most frequent cause of acute lower respiratory tract infection bringing infants between one and six months of age into the hospital with pneumonia, bronchitis (Stott & Taylor, 1985; Zlateva, Lemey, Vandamme, & Van Ranst, 2004; Borchers, Chang, Gershwin, & Gershwin, 2013), and otitis (Klein, Dollete, & Yolken, 1982) and significantly increases the prevalence of asthma amongst children who are hospitalized with RSV in infancy or early childhood (Borchers et al., 2013). Moreover, there is a marked correlation between the incidence of RSV in the community and the occurrence of sudden infant deaths in children above 3 months of age (Chanock et al., 1957). A simple upper respiratory illness in high-risk immunocompromised adults is no longer viewed as trivial (Whimbey & Ghosh, 2000). People with cardiopulmonary diseases and immunocompromised persons with bone marrow transplant patients prior to marrow engraftment are at highest risk for pneumonia and death (Morris et al., 1956). For cancer patients, the risks and benefits of administering intensive chemotherapy in the setting of a seemingly benign upper respiratory illness are now weighed heavily (Klein et al., 1982).

In animals, the virus is recognized as an important cause of Bovine Respiratory Disease (BRD) in Europe and the United States (Whimbey & Ghosh, 2000).

In terms of its epidemiology, it produces an annual epidemic of predominantly upper respiratory tract infections in children and healthy adults (Chanock et al., 1957) with re-infections occurring throughout life even in the presence of pre-existing antibodies (Sullender, Mufson, Anderson, & Wertz, 1991).

2.2 Problem Statement

Conventional methods of describing diversity involve comparing each sample against a given reference genome instead of comparing every sample against

every other sample (Paten, Novak, Eizenga, & Garrison, 2017). This comparison is non-transitive meaning that the way a sample varies from the reference does not expressly tell us how that sample varies from a separate sample.

2.3 Justification

A reference genome graph is a robust data structure for representing genome variation unlike the current approach where we compare a sample against a linear reference genome. Given that consensus sequences are a mosaic of haplotypes, we would like to make use of genome graph to disentangle the sequence diversity present in RSV sequences, and potentially other respiratory viruses.

2.4 Objectives

2.4.1 Main Objective

To construct an RSV variation graph from samples collected from a single household in the course of an RSV household outbreak.

2.4.2 Specific Objectives

1. To perform a review of existing genome graph tools used in constructing pangenome graphs.

2. To construct a variation graph from a set of samples collected from a single household during the course of an RSV epidemic.

3 Literature Review

3.1 RNA Viruses

RNA viral populations don't exist as a collection of organisms with a single genome but rather as a quasispecies (also called a mutant spectrum or a mutant cloud) where most of the biologically relevant variation observed in vivo is as a result of genetic variation, competitive selection and random events acting on multiple replicative units. These quasispecies dynamics have been used to explain the failure of monotherapy and synthetic antiviral vaccines but have opened new possibilities for antiviral interventions (Domingo, Sheldon, & Perales, 2012). These viruses have mutation rates up to a million times higher than their hosts; rates that are so high it is unlikely for a virus

to have an identical RNA molecule as its immediate progeny (Domingo et al., 2012). Negative selection controls this mutation rate and proof that it is not optimized by natural selection is that in some cases it leads to local extinction (Duffy, 2018).

3.1.1 Respiratory Syncytial Virus (RSV)

RSV is the major cause of acute lower respiratory tract infections associated with pneumonia, bronchitis (Borchers et al., 2013; Zlateva et al., 2004) and otitis (Klein et al., 1982) more frequently than any other agent and particularly in the first year of life (Stott & Taylor, 1985). As if that wasn't enough, RSV significantly increases the prevalence of asthma in children who are hospitalized with it and there is a marked correlation between the incidence of RSV in the community and the occurrence of sudden infant deaths as well as a third of cot deaths among children over 3 months of age (Chanock et al., 1957).

A simple upper respiratory illness in immunocompromised adults or the elderly is no longer viewed as trivial (Chanock et al., 1957; Whimbey & Ghosh, 2000). People with cardiopulmonary diseases and immunocompromised persons with bone marrow transplant patients prior to marrow engraftment are at highest risk for pneumonia and death (Morris et al., 1956). For cancer patients, the risks and benefits of administering intensive chemotherapy in the setting of a seemingly benign upper respiratory illness are now weighed heavily (Klein et al., 1982).

In animals, the RSV virus is also recognized as an important cause of Bovine Respiratory Disease (BRD) in Europe and the United States (Whimbey & Ghosh, 2000) being the most costly disease of beef cattle in North America (Griffin, 1997).

1. History The virus was first isolated in 1956 from Chimpanzees and named Chimpanzee Coryza Agent (Morris et al., 1956) after which it was isolated in children from whom it had not been possible to isolate and renamed RSV in 1957 then classified in order Mononegavirales, family Paramyxoviridae, subfamily Pneumovirinae, genus Pneumovirus (Chanock et al., 1957; Beem et al., 1960; Zlateva et al., 2005).

2. Epidemiology In older children and healthy adults, RSV presents in highly seasonal annual epidemics (Aamir et al., 2013; Al-Toum et al., 2006) of mild reinfections predominantly in the upper respiratory tract (Chanock et al., 1957) even in the presence of pre-existing antibodies (Cane, 2001; Sullender et al., 1991).

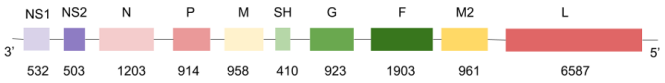

Figure 1: A schematic of RSV antisense RNA strand showing its 10 genes. The rectangles represent genes with the different shades of the same colour used to show similarity. The grey connectors are the intergenic regions. The numbers below are the estimated gene lengths. Adapted from (Nam & Ison, 2019)

The epidemics have been found to have a significant negative correlation with temperature and a significant positive correlation with relative humidity and rainfall (Al-Toum, Bdour, & Ayyash, 2006) and therefore crop up in the coldest months which naturally vary with latitude.

In temperate climates, RSV epidemics occur in the winter between December and February but peaking in January and February (Al-Toum et al., 2006) and are a major cause of winter mortality associated with 60-80% more deaths than influenza (Nicholson, 1996).

In tropical climates, epidemics occur during the rainy season (Al-Toum et al., 2006; Aamir, Alam, Sadia, Zaidi, & Kazi, 2013) but are also associated with religious festivals (Cane, 2001).

Serious disease is limited to the primary infection which occurs between six weeks and two years of age during the child's first or second epidemics (Cane, 2001) and can occur in the presence of maternally derived antibodies. However, infants with more severe illnesses were found to have lower levels of antibodies in serum collected near the onset of illness than did infants with milder illnesses (Glezen, Paredes, Allison, Taber, & Frank, 1981; Cane, 2001).

3. The Genetic Makeup of RSV

RSV, whose genome structure is shown above, is an enveloped virus with a nonsegmented negative-strand RNA genome of approximately 15,200 nucleotides containing 10 genes which code for 11 proteins whose order is 3k NS1, NS2, N, P, M1, SH, G, F, M2 (note that M2 codes for M2.1 and M2.2 proteins), and L with attenuation of transcription step-wise with distance from the 3k end (Cane, 2001).

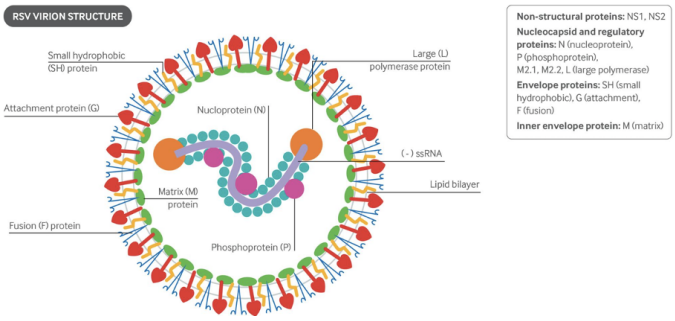

Figure 2: A schematic of the RSV capsid showing the lipid bilayer and most importantly the surface the F and G glycoproteins. From (Nam & Ison, 2019).

As shown in table 1 and figures 1 and 2, RSV has three surface glycoproteins: the small hydrophobic (SH) protein which may be non-structural, the fusion (F) protein which plays the main role in virus penetration, syncytium formation, and possibly can also mediate attachment and the attachment (G) glycoprotein which plays the main role in virus attachment. It has four nucleocapsid proteins including the nucleoprotein (N), phosphoprotein (P), M2-1(also designated 22K and sometimes considered a matrix protein) and polymerase (L). The M2 gene contains a second open reading frame encoding a protein (M2-2) which regulates transcription (Fearns & Collins, 1999). There is a single matrix protein, M1, which may mediate interactions between the nucleocapsid and envelope and the two non-structural proteins, NS1 and NS2 have recently been shown to antagonise the interferon-induced antiviral response (Fearns & Collins, 1999; Schlender, Bossert, Buchholz, & Conzelmann, 2000).

4. Groups of RSV RSV was initially divided into two antigenic groups A and B in 1966 by its reaction with panels of monoclonal antibodies particularly those directed against its P, F and G proteins (Coates, Alling, & Chanock, 1966). It is worth noting that only antibodies directed against the G and F proteins have been shown to be neutralising in vitro or protective in vivo (Cane, 2001).

   It was later demonstrated that the two groups are distinct at the genetic level (Johnson & Collins, 1988). The F and N proteins are highly conserved between the groups showing 91% and 96% amino acid simi-

larity, respectively (Johnson & Collins, 1988, 1989). In contrast, the G protein was found to be highly variable where the amino acid similarity of this protein between groups A and B was 53% (Johnson, Spriggs, Olmsted, & Collins, 1987; Zlateva et al., 2004).

Both groups are known to circulate within an epidemic (T. C. Peret, Hall, Schnabel, Golub, & Anderson, 1998) without any leading to the extinction of the other, although A tends to be more dominant in epidemics attributed to the higher variability among the A strains (T. C. Peret et al., 1998; Zlateva et al., 2005).

The sequence diversity of the G glycoprotein (the type II glycoprotein of 289–299 amino acids depending on the virus strain (Cane, 2001) coded by the G gene suggests that the two subgroups have evolved separately for a significant period of time with proof of RSV A's most recent common ancestor dating back as the early 1940s (Zlateva et al., 2004).

Because the F gene mutates at a much lower rate compared to the G gene it becomes an adequate vaccine target which is why we talk of RSV F vaccines (Anderson et al., 2013; Giersing et al., 2016). This lower rate of mutation also leads to consistent identification by antibodies and therefore the major neutralizing antibody response to RSV appears to be induced by the F protein (Olmsted et al., 1986).

Groups A and B are subdivided further into subgroups, as of 2012 there were 11 subgroups of RSV A: ON1, GA1–GA7, SAA1, NA1, and NA2 and 17 subgroups of RSV B: GB1–GB4, SAB1-SAB3, and BA1–BA10 (T. C. Peret et al., 1998; T. C. T. Peret et al., 2000; Venter, Madhi, Tiemessen, & Schoub, 2001; Trento et al., 2003; Shobugawa et al., 2009; Eshaghi et al., 2012; Aamir et al., 2013).

3.2 Graphs in Bioinformatics

Contemporary methods of representing a reference genome as a linear sequence of characters to represent bases (Dilthey, Cox, Iqbal, Nelson, & McVean, 2015) introduce a mapping bias towards alleles in the reference known as reference bias compared to the mapping of alternative alleles (Degner et al., 2009; Brandt et al., 2015).

This naturally leads to a need for a structure that can represent variation that is inherent in the genome. Other models can approach this structure with varying degrees of accuracy, but it is naturally represented as a graph in

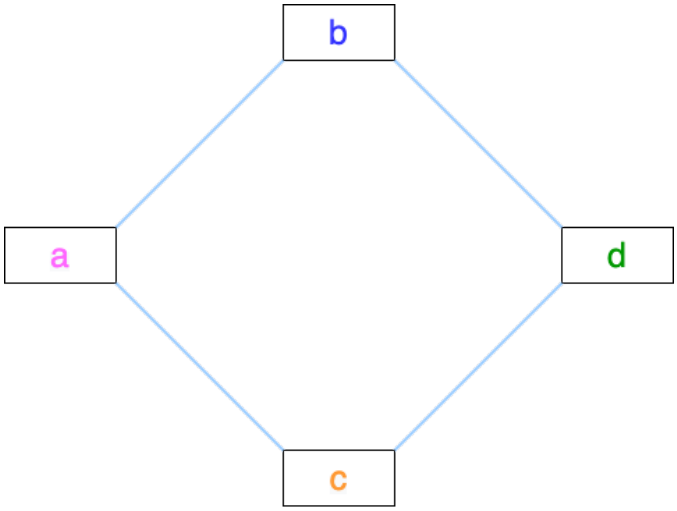

Figure 3: G is an undirected graph of four nodes a,b,c and d.

which the sequences themselves are implicitly encoded as walks in the graph (Paten, Novak, Eizenga, & Garrison, 2017).

3.3 Graph Theory

A graph is an object, or collection, of two sets, a vertex set and edge set. The vertex set is a finite non-empty set, to mean a graph must have at least one vertex. The edge set may be empty (Trudeau, 1993) and is used to present relationships between the vertices.

More formally, a graph G is an unordered pair (V(G), E(G)) consisting of a set V(G) of vertices and a set E(G), disjoint from V (G), of edges, together with an incidence function that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G (Bondy & Murty, 2011).

Graphs can be represented diagrammatically as shown below. G {{a c} {b d}}

H {{a c} {c d}}

3.3.1 Graph classifications

Graphs can be broken down into many classifications but in this case, we want to focus on simple versus multigraphs and directed versus undirected. A simple graph can only have one edge connecting two adjacent vertices

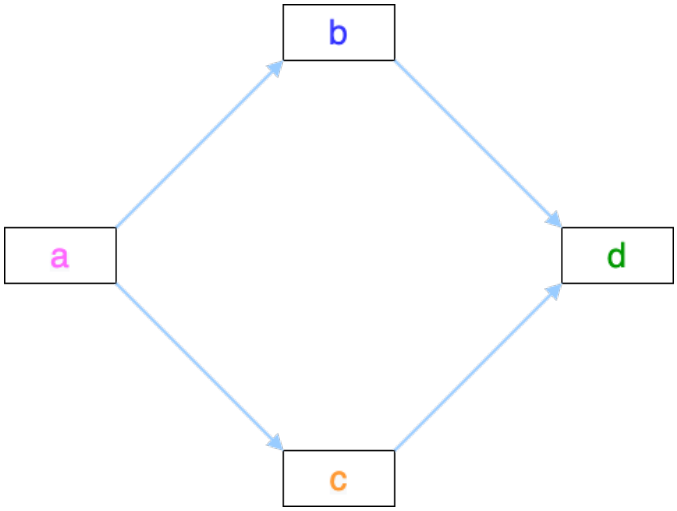

Figure 4: H is an undirected graph of notes a, b and c. Two vertices which are incident with a common edge are adjacent, as are two edges which are incident with a common vertex, and two distinct adjacent vertices are neighbours (Bondy & Murty, 2011).

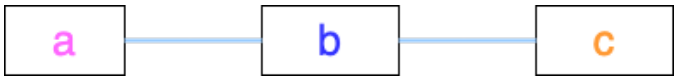

Figure 5: A simple graph showing only a single edge connecting any two nodes

while a multigraph is a graph in which two adjacent vertices are connected by more than one edge.

Simple Graph

Multigraph

Figure 3: (a) A simple graph showing only a single edge connecting any two nodes. (b) A multigraph where more than one edge can connect any two nodes. A directed graph also called a digraph is a graph in which the edges have direction.

Figure 4: A directed graph with the edges indicating direction. An undirected graph is one in which the edges do not have direction indicated on them.

Figure 5: An undirected graph where the edges have no indication of direction. A bidirected graph is one in which each edge has an independent orientation (Edmonds & Johnson, 2003). This is important for the represen-

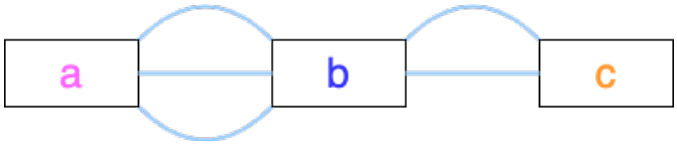

Figure 6: A multigraph where more than one edge can connect any two nodes.

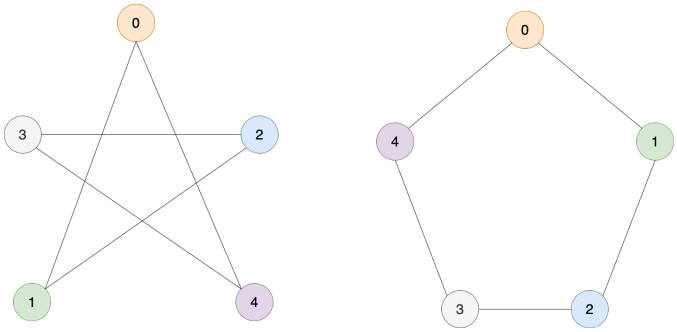

Figure 7: The two nodes are different visualizations of the same graph and therefore an isomorphism.

tation of strand, that is reading a DNA molecule in its forward or reverse complement orientation (Paten, Novak, Eizenga, & Garrison, 2017).

The degree of a vertex v in a graph G, is the number of edges of G incident with v (going in and out of v), each loop counting as two edges. In directed graphs, we have the concept of indegree and outdegree. The indegree refers to the numbers of head ends of the edges adjacent to a vertex and the outdegree is the number of tail ends of the edges adjacent to a vertex (Bondy & Murty, 2011). A vertex is even if its degree is an even number and odd otherwise (Trudeau, 1993).

An isomorphism is a relationship between two graphs such that the two graphs can be represented by identical diagrams (Bondy & Murty, 2011) whereas an automorphism of a graph is an isomorphism of the graph to itself as shown below.

3.3.2 Walks and paths

A path is a simple graph whose vertices can be arranged in a linear sequence in such a way that two vertices are adjacent if they are consecutive in the sequence, and are nonadjacent otherwise (Bondy & Murty, 2011).

A walk in a graph is a sequence A1 A2 A3 ... An of not necessarily distinct vertices in which A1 is joined by an edge to A2, A2 is joined by an edge to A3, ..., and An1 is joined by an edge to An. The walk A1 A2 A3 ... An is said to join A1 and An (Trudeau, 1993).

Therefore, a path is a graph, whereas a walk is a traversal of a graph.

An Euler or Eulerian walk is a walk that uses every edge in the graph exactly once.

A Hamiltonian walk is like an eulerian walk but for nodes and can be open or closed, an open hamilton walk is a walk that uses every vertex in the graph exactly once. A closed hamilton walk is a closed walk that uses the initial vertex exactly twice and all the other vertices in the graph exactly once (Trudeau, 1993).

3.4 Genome Graphs

A genome graph is a generic term that refers to the representation of a sequence or sequences or genetic material using graph-based methods implicitly or explicitly. Genome graphs are expected to lead to improvements in mapping reads, variant calling and haplotype determination (Paten, Novak, Eizenga, & Garrison, 2017).

Genome graphs are generally directed graphs and have different classifications, based on where the sequences are held within the graph, either on the edge or in the nodes.

These are vertex-labelled directed graphs, graphs whose nodes are labelled such that a directed walk can be interpreted as a DNA sequence, defined by the sequence of node labels along the walk and edge-labelled directed graphs in which case the nodes, rather than the edges, can be viewed as representing the intersection points between connected subsequences (Paten, Novak, Eizenga, & Garrison, 2017).

3.4.1 De Bruijn Graph

These are graphs used in the assembly of reads named after Dutch mathematician Nicolaas de Bruijn who became interested in the superstring problem: find a shortest circular superstring that contains all possible substrings of length k (k-mers) over a given alphabet which he solved using an eulerian walk over the k-mers (Compeau, Pevzner, & Tesler, 2011).

3.4.2 Sequence graph

A sequence graph is a bidirected graph in which each node is labelled with a nucleotide string a “sequence graph” (Paten, Novak, Eizenga, & Garrison, 2017). In this bidirected graph, the features of an edge indicate to which side of a node (sequence), 5’ or 3’, each end of the edge connects" (Novak et al., 2017).

3.4.3 Variation Graph

A variation graph is a graph where a complete walk along the graph represents a haplotype (Paten, Novak, Eizenga, & Garrison, 2017).

Many genome graphs don’t represent the concept of the strand, "reading a DNA molecule in its forward and reverse complement orientations". To express strandedness, directed graphs can be generalized to bidirected graphs (Edmonds & Johnson, 2003; Medvedev, Stanciu, & Brudno, 2009) in which each edge endpoint has an independent orientation, indicating whether the forward or the reverse complement strand of the attached node is to be visited when entering the node through that endpoint of the edge. Inversions, reverse tandem duplications, and arbitrarily complex rearrangements are expressible in the bidirected representation (Paten, Novak, Eizenga, & Garrison, 2017).

3.4.4 Population Reference Graphs (PRGs)

Population reference graphs are graphs that represent a population-wide genome combining multiple reference sequences and catalogues of variation (Dilthey et al., 2015). This concept may also be extended to represent, in our case, a virus mutant cloud.

3.4.5 Problems arising from graph-based reference models

1. Calling alleles at sites This involves declaring an allele at a given position, this position could span several nodes or edges in an undefined manner.

   A proposed way to describe their positions is via motif (Paten, Novak, Eizenga, & Garrison, 2017), patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks (Milo et al., 2002), called a superbubble in directed graph or an ultrabubble in bidirected graphs (Paten, Novak, Garrison, & Hickey, 2017).

Superbubbles and ultrabubbles are directed acyclic subgraphs that connect to the rest of the graph through one source node and one sink node (Paten, Novak, Garrison, & Hickey, 2017).

2. Non-trivial indexing and reference mapping We now need to use methods that are aware of alternative alleles to map reads to a graph reference (Paten, Novak, Eizenga, & Garrison, 2017). The indexing could be done through gbwt (Sirén, Garrison, Novak, Paten, & Durbin, 2018) could be achieved via partial order alignment gssw (Zhao, Lee, Garrison, & Marth, 2013).

3. Coordinate system A reference genome coordinate system is a system that uses coordinates to uniquely determine the positions of bases in the reference genome (Rand et al., 2017).

   An interesting problem introduced by graph-based reference structures is that it's no longer trivial to define a locus on the reference (Paten, Novak, Eizenga, & Garrison, 2017). The Computational Pan-Genomics Consortium (2016) however agreed on qualities that a coordinate system should have (Paten, Novak, Eizenga, & Garrison, 2017; Rand et al., 2017). A coordinate system should have: monotonicity genome graph coordinates of successive bases within a genome should be increasing, legibility coordinates should be compact and human interpretable, spatiality bases physically close together within a genome should have similar coordinates, vertical spatiality of bases that are allelic variants of one another (Rand et al., 2017). horizontal spatiality of bases that can appear together within a single molecule (Rand et al., 2017).

3.4.6 Mapping reads to a reference genome graph

Given that a genetic sequence is read in small pieces for short reads and much longer pieces for long reads, we need to find where in the genome a read comes from. Read mapping is the process of finding the position where the read came from in a reference sequence or graph (Novak et al., 2017).

1. Reference bias or reference allele bias Reference allele bias is the tendency to under-report data whose underlying DNA does not match a reference allele (Paten, Novak, Eizenga, & Garrison, 2017). Masking known SNP positions in the genome sequence can eliminate the reference bias but do not lead to more reliable results overall (Degner et al., 2009).

3.4.7 Variation Graphs in Virus Haplotype Detection and Quantification

Compared to eukaryotes, viruses have relatively short genomes and high mutation rates (Duffy, 2018) and RNA viruses exist as a quasi-species (Domingo et al., 2012). This gives rise to the need to deconvolute the individual haplotypes and quantify them.

There are a number of other tools for the assembly of haplotypes of virus quasispecies. These can be broadly categorized into reference-guided and reference-free. De novo approaches do not require any prior information, such as a reference genome or knowledge of the quasispecies composition. De novo approaches have been shown to have advantages over reference-guided reconstruction, since using a reference genome can induce significant biases (Baaijens, Stougie, & Schönhuth, 2020).

There exist methods for de novo, strain aware metagenomic assembly such as VG-flow (Baaijens et al., 2020) however which focus only on short-read data. VG-flow takes as input a next-generation sequencing (NGS) data set and a collection of strain-specific contigs assembled from the data and produces full-length haplotypes and corresponding abundance estimates (Baaijens et al., 2020).

References

Aamir, U. B., Alam, M. M., Sadia, H., Zaidi, S. S. Z., & Kazi, B. M. (2013, September). Molecular Characterization of Circulating Respiratory Syncytial Virus (RSV) Genotypes in Gilgit Baltistan Province of Pakistan during 2011-2012 Winter Season. *PLOS ONE*, *8*(9), e74018. doi: 10.1371/journal.pone.0074018

Al-Toum, R., Bdour, S., & Ayyash, H. (2006, August). Epidemiology and Clinical Characteristics of Respiratory Syncytial Virus Infections in Jordan. *Journal of Tropical Pediatrics*, *52*(4), 282–287. doi: 10.1093/tropej/fml002

Anderson, L. J., Dormitzer, P. R., Nokes, D. J., Rappuoli, R., Roca, A., & Graham, B. S. (2013, April). Strategic priorities for respiratory syncytial virus (RSV) vaccine development. *Vaccine*, *31*, B209-B215. doi: 10.1016/j.vaccine.2012.11.106

Baaijens, J. A., Stougie, L., & Schönhuth, A. (2020, February). Strain-aware assembly of genomes from mixed samples using flow variation graphs. *bioRxiv*, 645721. doi: 10.1101/645721

Beem, M., Wright, F. H., Hamre, D., Egerer, R., & Oehme, M. (1960, September). Association of the Chimpanzee Coryza Agent with Acute Respiratory Disease in Children. *New England Journal of Medicine*, *263*(11), 523–530. doi: 10.1056/NEJM196009152631101

Bondy, A., & Murty, U. S. R. (2011). *Graph Theory*. Springer London.

Borchers, A. T., Chang, C., Gershwin, M. E., & Gershwin, L. J. (2013, December). Respiratory Syncytial Virus—A Comprehensive Review. *Clinical Reviews in Allergy & Immunology*, *45*(3), 331–379. doi: 10.1007/s12016-013-8368-9

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015, March). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3: Genes|Genomes|Genetics*, *5*(5), 931–941. doi: 10.1534/g3.114.015784

Cane, P. A. (2001). Molecular epidemiology of respiratory syncytial virus. *Reviews in Medical Virology*, *11*(2), 103–116. doi: 10.1002/rmv.305

Chanock, R., Roizman, B., & Myers, R. (1957, November). Recovery from Infants with Respiratory Illness of a Virus Related to Chimpanzee Coryza Agent (CCA) Isolation, Properties and Characterization. *American Journal of Epidemiology*, *66*(3), 281–290. doi: 10.1093/oxfordjournals.aje.a119901

Coates, H. V., Alling, D. W., & Chanock, R. M. (1966, March). An Antigenic

Analysis of Respiratory Syncytial Virus Isolates by a Plaque Reduction Neutralization Test. *American Journal of Epidemiology*, *83*(2), 299–313. doi: 10.1093/oxfordjournals.aje.a120586

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011, November). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991. doi: 10.1038/nbt.2023

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009, December). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207–3212. doi: 10.1093/bioinformatics/btp579

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., & McVean, G. (2015, June). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, *47*(6), 682–688. doi: 10.1038/ng.3257

Domingo, E., Sheldon, J., & Perales, C. (2012, June). Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews*, *76*(2), 159–216. doi: 10.1128/MMBR.05023-11

Duffy, S. (2018, August). Why are RNA virus mutation rates so damn high? *PLOS Biology*, *16*(8), e3000003. doi: 10.1371/journal.pbio.3000003

Edmonds, J., & Johnson, E. L. (2003). Matching: A Well-Solved Class of Integer Linear Programs. In M. Jünger, G. Reinelt, & G. Rinaldi (Eds.), *Combinatorial Optimization — Eureka, You Shrink!: Papers Dedicated to Jack Edmonds 5th International Workshop Aussois, France, March 5–9, 2001 Revised Papers* (pp. 27–30). Berlin, Heidelberg: Springer. doi: 10.1007/3-540-36478-1_3

Eshaghi, A., Duvvuri, V. R., Lai, R., Nadarajah, J. T., Li, A., Patel, S. N., ... Gubbay, J. B. (2012, March). Genetic Variability of Human Respiratory Syncytial Virus A Strains Circulating in Ontario: A Novel Genotype with a 72 Nucleotide G Gene Duplication. *PLOS ONE*, *7*(3), e32807. doi: 10.1371/journal.pone.0032807

Fearns, R., & Collins, P. L. (1999, July). Role of the M2-1 Transcription Antitermination Protein of Respiratory Syncytial Virus in Sequential Transcription. *Journal of Virology*, *73*(7), 5852–5864.

Giersing, B. K., Modjarrad, K., Kaslow, D. C., Moorthy, V. S., Bavdekar, A., Cichutek, K., ... Smith, P. (2016, June). Report from the World Health Organization's Product Development for Vaccines Advisory Committee (PDVAC) meeting, Geneva, 7–9th Sep 2015. *Vaccine*, *34*(26), 2865–2869. doi: 10.1016/j.vaccine.2016.02.078

Glezen, W. P., Paredes, A., Allison, J. E., Taber, L. H., & Frank, A. L. (1981, May). Risk of respiratory syncytial virus infection for infants

from low-income families in relationship to age, sex, ethnic group, and maternal antibody level. *The Journal of Pediatrics*, *98*(5), 708–715. doi: 10.1016/S0022-3476(81)80829-3

Griffin, D. (1997, November). Economic impact associated with respiratory disease in beef cattle. *The Veterinary Clinics of North America. Food Animal Practice*, *13*(3), 367–377. doi: 10.1016/s0749-0720(15)30302-9

Johnson, P. R., & Collins, P. L. (1988). The Fusion Glycoproteins of Human Respiratory Syncytial Virus of Subgroups A and B: Sequence Conservation Provides a Structural Basis for Antigenic Relatedness. *Journal of General Virology,*, *69*(10), 2623–2628. doi: 10.1099/0022-1317-69-10-2623

Johnson, P. R., & Collins, P. L. (1989). The 1B (NS2), 1C (NS1) and N Proteins of Human Respiratory Syncytial Virus (RSV) of Antigenic Subgroups A and B: Sequence Conservation and Divergence within RSV Genomic RNA. *Journal of General Virology,*, *70*(6), 1539–1547. doi: 10.1099/0022-1317-70-6-1539

Johnson, P. R., Spriggs, M. K., Olmsted, R. A., & Collins, P. L. (1987, August). The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: Extensive sequence divergence between antigenically related proteins. *Proceedings of the National Academy of Sciences*, *84*(16), 5625–5629. doi: 10.1073/pnas.84.16.5625

Klein, B. S., Dollete, F. R., & Yolken, R. H. (1982, July). The role of respiratory syncytial virus and other viral pathogens in acute otitis media. *The Journal of Pediatrics*, *101*(1), 16–20. doi: 10.1016/S0022-3476(82)80172-8

Medvedev, P., Stanciu, M., & Brudno, M. (2009, November). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, *6*(11), S13-S20. doi: 10.1038/nmeth.1374

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002, October). Network Motifs: Simple Building Blocks of Complex Networks. *Science*, *298*(5594), 824–827. doi: 10.1126/science.298.5594.824

Morris, J. A., Blount, R. E., & Savage, R. E. (1956, July). Recovery of Cytopathogenic Agent from Chimpanzees with Goryza. *Proceedings of the Society for Experimental Biology and Medicine*, *92*(3), 544–549. doi: 10.3181/00379727-92-22538

Nicholson, K. G. (1996, February). Impact of influenza and respiratory syncytial virus on mortality in England and Wales from January 1975 to December 1990. *Epidemiology & Infection*, *116*(1), 51–63. doi:

10.1017/S0950268800058957

Novak, A. M., Hickey, G., Garrison, E., Blum, S., Connelly, A., Dilthey, A., ... Paten, B. (2017, January). *Genome Graphs* (Preprint). Bioinformatics. doi: 10.1101/101378

Olmsted, R. A., Elango, N., Prince, G. A., Murphy, B. R., Johnson, P. R., Moss, B., ... Collins, P. L. (1986, October). Expression of the F glycoprotein of respiratory syncytial virus by a recombinant vaccinia virus: Comparison of the individual contributions of the F and G glycoproteins to host immunity. *Proceedings of the National Academy of Sciences*, *83*(19), 7462–7466. doi: 10.1073/pnas.83.19.7462

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017, March). Genome graphs and the evolution of genome inference. *Genome Research*, gr.214155.116. doi: 10.1101/gr.214155.116

Paten, B., Novak, A. M., Garrison, E., & Hickey, G. (2017, January). Superbubbles, Ultrabubbles and Cacti. *bioRxiv*, 101493. doi: 10.1101/101493

Peret, T. C., Hall, C. B., Schnabel, K. C., Golub, J. A., & Anderson, L. J. (1998). Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. *Journal of General Virology,*, *79*(9), 2221–2229. doi: 10.1099/0022-1317-79-9-2221

Peret, T. C. T., Hall, C. B., Hammond, G. W., Piedra, P. A., Storch, G. A., Sullender, W. M., ... Anderson, L. J. (2000, June). Circulation Patterns of Group A and B Human Respiratory Syncytial Virus Genotypes in 5 Communities in North America. *The Journal of Infectious Diseases*, *181*(6), 1891–1896. doi: 10.1086/315508

Rand, K. D., Grytten, I., Nederbragt, A. J., Storvik, G. O., Glad, I. K., & Sandve, G. K. (2017, May). Coordinates and intervals in graph-based reference genomes. *BMC Bioinformatics*, *18*(1), 263. doi: 10.1186/s12859-017-1678-9

Schlender, J., Bossert, B., Buchholz, U., & Conzelmann, K.-K. (2000, September). Bovine Respiratory Syncytial Virus Nonstructural Proteins NS1 and NS2 Cooperatively Antagonize Alpha/Beta Interferon-Induced Antiviral Response. *Journal of Virology*, *74*(18), 8234–8242. doi: 10.1128/JVI.74.18.8234-8242.2000

Shobugawa, Y., Saito, R., Sano, Y., Zaraket, H., Suzuki, Y., Kumaki, A., ... Suzuki, H. (2009, August). Emerging Genotypes of Human Respiratory Syncytial Virus Subgroup A among Patients in Japan. *Journal of Clinical Microbiology*, *47*(8), 2475–2482. doi: 10.1128/JCM.00115-09

Sirén, J., Garrison, E., Novak, A. M., Paten, B., & Durbin, R. (2018, May).

Haplotype-aware graph indexes. *arXiv:1805.03834 [cs]*.

Stott, E. J., & Taylor, G. (1985, March). Respiratory syncytial virus. *Archives of Virology*, *84*(1), 1–52. doi: 10.1007/BF01310552

Sullender, W. M., Mufson, M. A., Anderson, L. J., & Wertz, G. W. (1991, October). Genetic diversity of the attachment protein of subgroup B respiratory syncytial viruses. *Journal of Virology*, *65*(10), 5425–5434.

Trento, A., Galiano, M., Videla, C., Carballal, G., García-Barreno, B., Melero, J. A., & Palomo, C. (2003). Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. *Journal of General Virology,*, *84*(11), 3115–3120. doi: 10.1099/vir.0.19357-0

Trudeau, R. J. (1993). *Introduction to Graph Theory*. Courier Corporation.

Venter, M., Madhi, S. A., Tiemessen, C. T., & Schoub, B. D. (2001). Genetic diversity and molecular epidemiology of respiratory syncytial virus over four consecutive seasons in South Africa: Identification of new subgroup A and B genotypesThe GenBank accession numbers of the sequences reported in this paper are AF348802–AF348826. *Journal of General Virology,*, *82*(9), 2117–2124. doi: 10.1099/0022-1317-82-9-2117

Whimbey, E., & Ghosh, S. (2000). Respiratory syncytial virus infections in immunocompromised adults. *Current clinical topics in infectious diseases*, *20*, 232–255.

Zhao, M., Lee, W.-P., Garrison, E. P., & Marth, G. T. (2013, December). SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. *PLOS ONE*, *8*(12), e82138. doi: 10.1371/journal.pone.0082138

Zlateva, K. T., Lemey, P., Moes, E., Vandamme, A.-M., & Van Ranst, M. (2005, July). Genetic Variability and Molecular Evolution of the Human Respiratory Syncytial Virus Subgroup B Attachment G Protein. *Journal of Virology*, *79*(14), 9157–9167. doi: 10.1128/JVI.79.14.9157-9167.2005

Zlateva, K. T., Lemey, P., Vandamme, A.-M., & Van Ranst, M. (2004, May). Molecular Evolution and Circulation Patterns of Human Respiratory Syncytial Virus Subgroup A: Positively Selected Sites in the Attachment G Glycoprotein. *Journal of Virology*, *78*(9), 4675–4683. doi: 10.1128/JVI.78.9.4675-4683.2004