QColors: An Algorithm for Conservative Viral Quasispecies Reconstruction from Short and Non-Contiguous Next Generation Sequencing Reads

Austin Huang*, Rami Kantor*, Allison DeLong†, Leeann Schreier*, and Sorin Istrail‡
* Division of Infectious Disease, † Center for Statistical Sciences, ‡ Department of Computer Science,
Brown University
Providence, RI, USA,
Email: austinh@alum.mit.edu

***Abstract*—Next generation sequencing technologies have been successfully applied to HIV-infected patients in order to obtain the mutational spectrum of heterogeneous viral populations within individuals, known as quasispecies. However, the metagenomics problem of quasispecies sequence reconstruction from next generation sequencing reads is not-yet widely applied in current practice and remains an emerging area of research. Furthermore, the majority of research methodology in HIV has focused on 454 sequencing, while many next-generation sequencing platforms are limited to shorter read lengths relative to 454 sequencing. Little work has been done in determining how best to address the read length limitations of other platforms.**

The approach described here incorporates graph representations of both read differences and read overlap to conservatively determine the regions of the sequence with sufficient variability to separate quasispecies sequences. Within these tractable regions of quasispecies inference, we use constraint programming to solve for an optimal quasispecies subsequence determination via vertex coloring of the conflict graph, a representation which also lends itself to data with non-contiguous reads such as paired-end sequencing.

We demonstrate the utility of the method by successfully applying it to simulations based on actual intra-patient clonal HIV-1 sequencing data.

***Index Terms*—HIV; quasispecies; virology; graph coloring; constraint programming;**

I. Introduction

Viral genome sequencing has been central to the study and treatment of HIV. The use of sanger Sequencing and resistance mutation profiles are now part of the standard of care in determining initial treatment regimens for new patients and optimizing changes upon treatment failure [1]–[3]. Databases of global sequences have been established [4]–[6] to serve as both basic science and clinical tools. In our own work we have used these databases to study the impact of different genetic backgrounds on the evolution of HIV-1 drug resistance [7]–[9]. As such, the field of HIV drug resistance is an ongoing case study of how sequencing data and molecular epidemiology can have a direct role in research and clinical practice.

However, the standard sequencing approach also has limitations. The high rate of viral turnover [10] and the error-prone viral genome replication process [11] lead to an intra-patient viral population that consists of genetically distinct subpopulations [12]–[14]. Thus the viral population is often described as a quasispecies [15]. Viral populations that circulate at low levels, termed minor variants, are clinically relevent because drug-resistant subpopulations may be selected upon treatment [16]–[18]. These minor variants are often undetectable using standard sequencing protocols. Specialized techniques such as clonal and single genome sequencing (SGS) have been developed to obtain sequences of minor variant subpopulations within patients [19]–[22]. SGS, in particular, was designed to address and minimize sequencing artifacts such as recombination of qusispecies sequences during PCR [19]. However, these methods require significant expertise and investment.

Next Generation Sequencing (NGS) is a term applied to a variety of recent sequencing platforms which sequence samples at high depth of coverage at relatively low monetary and training cost. The depth of coverage allows for the detection of minority variant subpopulations with prevalences $< 20\%$ [19]. Its application to HIV has been an active area of research in recent years [18], [23]–[28]. Nevertheless, in most research settings, NGS data have primarily been applied in a restricted manner - reads are mapped to an HIV-1 reference genome and the mutational composition of each sequence position is examined independently. Determination of mutational linkage and the reconstruction of the quasispecies population spectra remain active areas of research. Analysis tools to address these problems will aid in our ability to understand and interpret the vast amounts of data produced when applying these new sequencing technologies to HIV.

A. Reconstruction of Quasispecies Sequences

Any two virus particles with differing genomes may be considered as originating from genotypically distinct subpopulations. However, this view of quasispecies subpopulations is not especially useful in practice since it is relatively uncommon for viral sequencing to examine full genomes.

More commonly, the genotypes of interest are described by a subset of medically-relevant sequence positions. For example, studies of HIV drug resistance may focus on pol gene sequences. One may further project this set of genotypes

978-1-4577-1613-3/11/$26.00 ©2011 IEEE

into a smaller subset of sequence positions, such as the set of reverse transcriptase genotypes, or the set of genotypes with respect to a set of known drug resistance mutations. In general, the number of distinct genotypes in the subsequence is always fewer than the number of genotypes within the larger sequence.

The quasispecies reconstruction problem is to construct a representation of the spectrum of underlying subpopulation sequences from the partial information provided by sequencing reads. Sequencing reads represent a random sampling of a fragment of one of the underlying quasispecies sequences. Quasispecies reconstruction is closely related to haplotype assembly - the construction of haplotypes of a diploid organism from sequencing read fragments. A few approaches to quasispecies reconstruction have been proposed thus far [29]–[32], each of which has introduced mathematical representations which have provided new insights on the problem. Related problems have also arisen in genome assembly [33]–[39] and metagenomics [40]. The bipartition approach to haplotype assembly [41], [42] served as an inspiration for the approach taken in this paper.

Due to the complex error profile of next generation sequencing, error correction is a chief concern. Errors and biases arise for a multitude of reasons, including recombination during PCR, selection of primers, and sequence-specific errors [43]–[45]. Successful quasispecies reconstruction requires robust error correction in addition to reconstruction algorithms. Attempts to address these errors include both analysis approaches [23], [44] as well as tagging protocols [46].

Although robust error correction procedures are an important step in applying any method in practice, the focus of the current study is to present a different approach to representation and quasispecies reconstruction. The approach is influenced by practical motivations - conservative reconstruction in the presence of shorter reads and the need for representations which encompass non-contiguous reads. Non-contiguous reads include paired-end sequencing and some third generation sequencing technologies [47]. While the majority of next-generation sequencing research in the HIV field has focused on Roche 454 Sequencing, Illumina sequencing platforms are more widely available. One challenge in interpreting Illumina reads is that reads are generally shorter - ranging from 35bp contiguous reads to 2x150bp paired-end reads. By contrast, 454 sequencing reads currently can range from 400bp to 800bp. With shorter reads, there may be conserved regions of the genome which are not spanned by any read regardless of the depth of coverage. An objective of our approach is to conservatively determine where subpopulation sequence distinctions are well defined by the data and to reconstruct subsequences if reconstruction on the larger sequence is indeterminate and likely to introduce false recombinants. Our approach is to use a representation which reflects the relative tractability of different regions of the sequence and focus on reconstructions where there are data to support it.

II. Method

As suggested in the introduction, we adopt a slightly different problem formulation than previous approaches to quasispecies reconstruction:

QUASISPECIES SUBSEQUENCE RECONSTRUCTION PROBLEM. *Given a set of reads $r_1, ..., r_n$, find subsets of reads where quasispecies sequence distinctions can be inferred while minimizing the introduction of false recombinants. Within these read subsets, partition reads into a parsimonious generating set of subsequences.*

Our approach was motivated by the problem that short read lengths can render full-length reconstructions undecideable and a method which produces too many false recombinants will not be useful for researchers. Furthermore, there is a need for representations which allow for non-contiguous reads. A summary of our approach to this problem is outlined in Algorithm 1. The algorithm and simulations were implemented in C++ in conjunction with the Gecode constraint programming library [48]. Details of the method are described in the following sections.

Algorithm 1 QColors

Map reads to the reference sequence
Construct conflict $G_c = (V, E_c)$ and overlap $G_o = (V, E_o)$ graphs
Determine connected subgraphs of G_o and G_c using a Depth-first traversal
for each connected subgraph $G(V'_o, E'_o)$ in G_o **do**
  for each connected subgraph $G(V'_c, E'_c)$ in G_c **do**
    Define the neighborhood conflict graph, $G(V', E')$ with $V' = V'_o \cap V'_c$ and $E' = \{(v_i, v_j) \in E'_c : v_i \in V' \wedge v_j \in V'\}$
    Find a homomorphic reduction $G(V', E') \rightarrow H$
    Find maximal cliques of H
    Solve for the optimal coloring (QS sequence assignment) of H with clique and pairwise constraints
  end for
end for

A. Read Mapping

Reconstruction of quasispecies spectra differs from more generic metagenomics problem formulations [40] since a well-defined reference genome for all sequence fragments is available. A scanning analysis of the HXB2 HIV-1 reference genome [49] shows that 95% of reads can be uniquely mapped to the full genome using k-mers as short as 10bp. Since mapping is generally not a significant source of error for virus sequences due to their short genome length, mapping is assumed to be unambiguous for the purpose of these simulations.

With sufficiently high coverage depth, the identity of sequence characters which exhibit no variation across the quasispecies spectrum can be easily obtained directly from the mapped reads. Thus the remaining problem is to assign

reads with sequence variation to an appropriate quasispecies sequence.

B. The Overlap Graph and the Conflict Graph

We can represent relationships between reads as two complementary graphs - the overlap graph and the conflict graph.

A read conflict occurs when two reads have inconsistent sequences within an overlapping region. The conflict graph is defined as $G_c = (V, E_c)$ consisting of vertices, V, representing reads and edges, E_c, with pairs of vertices connected by an edge e_{ij} iff reads represented by vertices v_i and v_j overlap and conflict.

The overlap graph is defined as $G_o = (V, E_o)$ consisting of the same set of vertices V (again representing reads) and edges E_o which represent consistent (non-conflicting) relationships between overlapping reads. Pairs of vertices, v_i and v_j are connected by an edge e_{ij} iff reads represented by the vertices sufficiently overlap and do not conflict. An input parameter is used to determine the minimum number of overlapped positions between two reads for an edge to exist. This input parameter affects the conservativeness of the reconstruction procedure, and should be as high as the sequencing parameters (coverage, insert sizes) allow.

Unlike the overlap graph, there is no minimum number of overlapped positions required for a conflict graph edge because a conflict invalidates the possibility that two reads originate from the same quasispecies sequence, while agreement between reads does not prove that they originate from the same quasispecies sequence (roughly speaking, conflicts are deductive evidence while overlaps can be viewed as inductive or abductive evidence). While the objective here is to describe the reconstruction approach rather than error correction strategies, for experimentally-derived data, the error profile should be used to suggest a less-stringent conflict definition (e.g. more than one mutation between reads).

In the limit that reads span the entire sequence, G_c will be the graph complement implied by G_o, but this is not generally true if reads are shorter than the length of the sequence.

C. Reconstructing Quasispecies from Reads

Reconstructing quasispecies sequences from NGS reads is informed by both differences between reads (to distinguish sequences) and overlaps between reads (to extend sequences). In terms of the graph representations, quasispecies spectra can be inferred where connected subgraphs of the conflict and overlap graphs intersect.

We refer to these subgraphs as neighborhood conflict graphs, $G(V', E')$ (see Algorithm 1). For a constant depth of coverage, longer reads will lead to sequence regions which encompass larger spans of the sequence, while for shorter read lengths, inferences may only be made on pockets of variation within the genome.

Within the neighborhood conflict graph the objective is to partition the vertices into a minimal number of non-conflicting independent sets. This is equivalent to a vertex graph coloring problem, which is known to be NP-Hard [50], but with data

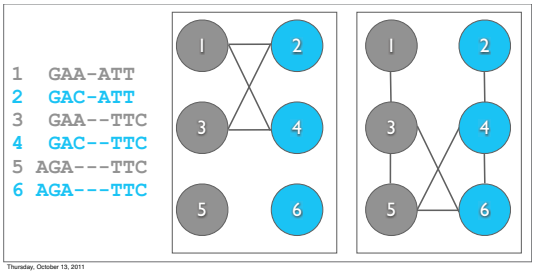

Fig. 1: A trivial "toy" example of 6 paired-end reads with inserts ranging from 1-3bp (left) of 2 quasispecies sequences (represented by blue and gray). The conflict graph (middle) and an overlap graph with an overlap threshold of 5 (right) are shown. Reads 1, 2, 3, and 4, define a neighborhood conflict graph for which 1 and 3 are assigned a single color and 2 and 4 are assigned a second color. Characters in reads 5 and 6 exhibit no conflicts, reflecting conserved positions which are included in all quasispecies sequences.

reduction and a constraint programming (CP) formulation, useful solutions can be obtained within acceptable computation times. This framework naturally permits paired-end and other non-contiguous reads, as the discontinuous character of the reads does not change the construction of the conflict and overlap graphs.

The set of vertices in this graph is further reduced by collapsing redundant reads. Redundancy is defined in terms of the graph data structures rather than the span of the reads. Due to the distances between variable positions in the sequence, it is common for reads to span different sequence positions, yet share the same set of edges. Groups of such vertices are collapsed into a single vertex to create a homomorphism of the neighborhood conflict graph. A proper coloring of a homomorphism is also a valid coloring of the original graph [51]. In the limit that the read length is the length of the sequence and coverage depth is high, this simple reduction will produce a complete graph in which each vertex corresponds to a distinct quasispecies sequence.

We use a simple Bron-Kerbosch algorithm with pivot vertices to identify maximal cliques in the graph [52]. This provides a lower bound to the chromatic number of the CP [51] and also introduces a distinctness constraint on vertices within cliques. The colors of the maximal clique are also assigned a fixed set of colors $1...s$ where s is the size of the maximal clique, reducing the domain of the search space by eliminating equivalent solutions. Constraint programming was implemented using C++ and the Gecode constraint programming library [48]. A branch and bound best solution search [53] was used with the number of colors as a cost function along with the following constraints:

- **C1** : The colorings of the maximal clique in H are fixed as $1, ..., s$, where s is the size of the maximal clique.
- **C2** : All colors of cliques in H are distinct
- **C3** : Colors of vertices connected by edges in H are distinct.

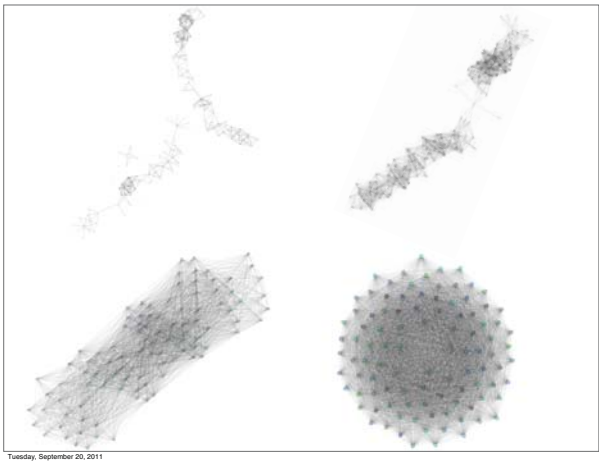

Fig. 2: Conflict graphs using read lengths of short (50 bp top left, 100 bp top right) and long (300 bp bottom left, 600 bp bottom right) contiguous reads sampled from Patient ID P00001 in [54]. Vertices in the graph represent reads while edges represent conflicting sequences between overlapping reads. Only a small number of samples was used to generate these graphs for the sake of clarity. Colors shown correspond to the underlying quasispecies sequences used to generate the graph.

Once a coloring of H is obtained, these quasispecies sequence assignments are propogated back to the reads modeled by H to obtain quasispecies assignments for each read. Conserved sequence positions are also added to generate model sequences (see Figure 3). Figure 1 illustrates a toy example of how reads encode conflict and overlap graphs.

D. Simulation and Evaluation

A read sampling simulator was written in C++ to evaluate the reconstruction. The simulation allows for either contiguous or paired-end fragments to be randomly sampled from an underlying set of known quasispecies sequences. Quasispecies sequences were obtained from a clonal sequencing study of the HIV-1 pol gene by Bacheler et al [54] available online at the Los Alamos HIV Database [55]. Initial simulations were performed to examine the qualitative characteristics of conflict graphs as read characteristics were varied.

Simulations of 10,000 150 x 2 bp paired-end reads with inserts varying uniformly from 0 to 100 bp were sampled from clonal sequences of two patients (patient ID P00001 and P00005) with relatively high number of clonal sequences available (49 and 42, respectively) [54]. An overlap threshold of 295 bp was used to construct the overlap graph. Reconstructed quasispecies subsequences were then compared to the generating set of quasispecies spectra to evaluate the reconstruction.

III. Results

To test the utility of this approach, clonal sequences [54] were used to simulate sampling and test quasispecies sequence reconstruction. These sequences consist of 984 bp from the pol gene region of the HIV-1 genome.

Figure 2 shows examples of conflict graphs using a small number of reads for clarity. Connected components in the conflict graph reflect sequence regions of pol for which variation distinguishes reads in different quasispecies sequences. In situations where reads are shorter than a conserved portion of the sequence, distinct connected components in the conflict graph will result (2 top left), independent of depth of coverage. Inferences can be made within local regions, but a larger reconstruction from read consistency alone will have degenerate solutions (many of which will be false recombinants) unless additional assumptions regarding long-range characteristics are incorporated.

Simulations of paired end reads were performed on clonal sequencing data of 2 patients from [54] having relatively larger numbers of quasispecies sequences available. Our method produces a set of conservative subsequence quasispecies, with reconstructed subsequences filling in a range of 878 to 911 of the 984 characters in the generating sequences. These reconstructed quasispecies subsequences are aggregated from read sets of varying sizes, from less than 10 to 410 reads.

Of the 36 quasispecies subsequences reconstructed for patient P00001, 32/36 (89%) represent subsequences of true quasispecies sequences, 16 of which map uniquely to a single true quasispecies sequence, while 4 reconstructed sequences did not correspond to a true quasispecies sequence. Of the 60 model quasispecies subsequences reconstructed for P00005, 54/60 (90%) represent subsequences of true quasispecies sequences, 36 of which map uniquely to a single true quasispecies sequence, while 7 of reconstructed sequences did not correspond to a true quasispecies sequence. Overall these results reflect a relatively small proportion of false positive sequences, and may be further improved by either longer read lengths or higher coverage coupled with a more stringent overlap threshold. Figure 3 shows a summary of the reconstruction pipeline.

These simulations and reconstructions were able to run on a desktop computer within 10 hours. We expect that larger-scale simulations and reconstructions at higher coverage should be tractable on an average computing cluster.

IV. Discussion

Here we have developed and tested a method which contributes three new perspectives to the quasispecies reconstruction problem: 1) A dual graph representation of the distinctness and relatedness between reads which is amenable to conservative interpretation of short reads and non-contiguous reads. This representation lends itself to non-contiguous reads (e.g. paired end or mate pair sequencing) because the algorithm does not use an ordering of the reads with respect to the genome, only the consistency/conflict between reads is utilized. 2) We suggest the intersections of the connected components of the overlap and conflict graphs as a local unit of data-supported inference. The goal of the approach is to infer quasispecies spectra within tractable domains even if

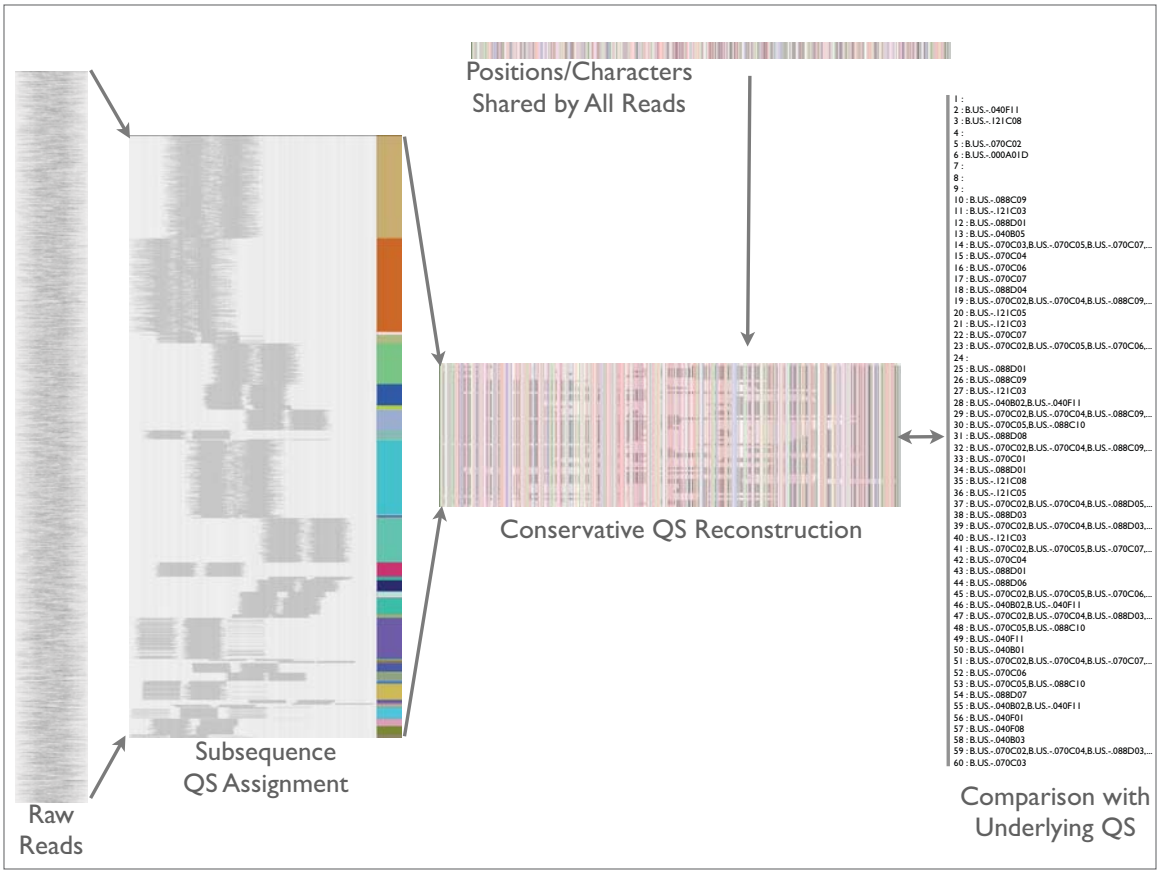

Fig. 3: The QS reconstruction pipeline can be seen as a data reduction which aims to limit false explanatory sequences. The process starts with raw reads (left). These are aggregated into tractable quasispecies subsequences supported by read conflicts and overlap, as discussed in the methods. Using the mapped reads, sequence positions which are perfectly conserved across reads (top) are also incorporated to construct an explanatory set of quasispecies subsequences (center, labeled "conservative quasispecies reconstruction", each row corresponds to the sequence obtained from a set of non-conflicting reads, columns correspond to sequence positions, and colors correspond to sequence characters - A=red, C=green, G=blue, T=white, undetermined=gray). Reconstruction is conservative in that the majority of these subsequences match at least one true underlying sequence (54/60 for P00005, shown in this figure). 36 of these quasispecies sequences contain sufficient information to map uniquely to an underlying quasispecies sequence.

short read lengths render full-length sequence reconstructions indeterminate. 3) We devise a method of data reduction and graph coloring using constraint programming to obtain optimal quasispecies sequence partitions within clusters of related reads and demonstrate its utility in making conservative sequence reconstructions using actual clonal sequence data.

Additional challenges remain before methods such as these will be widely adopted. Most importantly is to connect these methods to robust experimental and analytical error correction procedures [44], [46], [56]. Secondly, improving the performance of the constraint programming optimization will allow for denser, more extensive reconstructions. Consideration of additional/alternative constraints and cost functions could also improve the quality of the reconstruction. Alternatively, metaheurstics can also be considered in place of CP. Although the size of the current simulations may appear relatively small compared to the number of raw reads which can be obtained by NGS, error correction procedures and stringent overlap thresholds will also reduce the size of the read set by several orders of magnitude. This is especially true of paired-end and mate-pair designs with variable insert sizes where high coverage is required to obtain overlapping reads. Thirdly, other approaches have provided useful methods of estimating quasispecies sequence prevalences in addition to enumerating the set of quasispecies sequences [29]–[32], [44]. While the focus on conservative subsequence quasispecies spectra here is somewhat different from the previous approaches, generalization of the method into an appropriate form of population estimation will need to be examined. Finally, while the primary motivation for this work was to create a simple, easily-applied, practical tool, theoretical implications remain to be explored. Connections to graph coloring, refinement of the constraint programming algorithm, and more realistic

probabilistic modeling of the sampling and reconstruction processes will greatly improve our understanding of the quasispecies reconstruction problem. Further development of these and other quasispecies reconstruction approaches will add functionality to new sequencing technologies and aid in reducing large-scale sequencing data into a physiologically meaningful, interpretable form.

V. Conclusion

The quasispecies reconstruction problem remains an important challenge in utilizing next generation sequencing data for HIV. We have introduced a new representation and algorithm which can be used for conservative quasispecies sequence reconstructions using short read data and non-contiguous reads.

Acknowledgment

This work was supported by the National Institute of Allergy And Infectious Diseases at the National Institutes of Health grants number R01AI66922 for Dr. Huang, Dr. Kantor, Ms. DeLong, and Ms. Schreier, and P30AI042853 for Dr. Kantor and Ms. DeLong. Thanks to Derek Aguiar for useful discussions of constraint programming and Bjarni Halldorsson for early discussions on graph coloring extensions to haplotype phasing.

References

[1] V. A. Johnson, F. Brun-Vzinet, B. Clotet, H. F. Gnthard, D. R. Kuritzkes, D. Pillay, J. M. Schapiro, and D. D. Richman, "Update of the drug resistance mutations in HIV-1: december 2009," *Clin Infect Dis*, vol. 47, pp. 266–285, 2008.

[2] D. E. Bennett, R. J. Camacho, D. Otelea, D. R. Kuritzkes, H. Fleury, M. Kiuchi, W. Heneine, R. Kantor, M. R. Jordan, J. M. Schapiro *et al.*, "Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update," *PLoS One*, vol. 4, no. 3, 2009.

[3] P. A. Chan and R. Kantor, "Transmitted drug resistance in nonsubtype b HIV-1 infection," *HIV Therapy*, vol. 3, no. 5, pp. 447–465, 2009. [Online]. Available: http://www.futuremedicine.com/doi/abs/10.2217/hiv.09.30

[4] S. Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, "Human immunodeficiency virus reverse transcriptase and protease sequence database," *Nucleic acids research*, vol. 31, no. 1, p. 298, 2003.

[5] J. Zhou, N. Kumarasamy, R. Ditangco, A. Kamarulzaman, C. K. Lee, P. C. Li, N. I. Paton, P. Phanuphak, S. Pujari, and A. Vibhagool, "The TREAT asia HIV observational database: baseline and retrospective data," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 38, no. 2, p. 174, 2005.

[6] T. de Oliveira, R. W. Shafer, and C. Seebregts, "Public database for HIV drug resistance in southern africa," *Nature*, vol. 464, no. 7289, pp. 673–673, 2010.

[7] R. Kantor, D. A. Katzenstein, B. Efron, A. P. Carvalho, B. Wynhoven, P. Cane, J. Clarke, S. Sirivichayakul, M. A. Soares, J. Snoeck, C. Pillay, H. Rudich, R. Rodrigues, A. Holguin, K. Ariyoshi, M. B. Bouzas, P. Cahn, W. Sugiura, V. Soriano, L. F. Brigido, Z. Grossman, L. Morris, A. Vandamme, A. Tanuri, P. Phanuphak, J. N. Weber, D. Pillay, P. R. Harrigan, R. Camacho, J. M. Schapiro, and R. W. Shafer, "Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: Results of a global collaboration," *PLoS Med*, vol. 2, no. 4, p. e112, 2005. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pmed.0020112

[8] A. Huang, J. Hogan, S. Istrail, and R. Kantor, "Stratification by HIV-1 subtype does not eliminate systematic geographical variation," *Antiviral Therapy*, vol. 15 Supppl, no. 2, p. A161, 2010.

[9] A. Huang, J. W. Hogan, S. Istrail, A. DeLong, D. A. Katzenstein, and R. Kantor, "First-line antiretroviral treatment effects on resistance mutation prevalence in HIV-1 subtypes," *6th IAS Conference on HIV Pathogenesis, Treatment and Prevention*, 2011.

[10] D. D. Ho, A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz, "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection," *Nature*, vol. 373, no. 6510, pp. 123–126, 1995.

[11] L. M. Mansky, "Forward mutation rate of human immunodeficiency virus type 1 in a t lymphoid cell line*," *AIDS research and human retroviruses*, vol. 12, no. 4, pp. 307–314, 1996.

[12] M. Goodenow, T. Huet, W. Saurin, S. Kwok, J. Sninsky, and S. Wain-Hobson, "HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 2, no. 4, p. 344, 1989.

[13] J. M. Coffin, "HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy," *Science*, vol. 267, no. 5197, p. 483, 1995.

[14] O. G. Pybus and A. Rambaut, "Evolutionary analysis of the dynamics of viral infectious disease," *Nature Reviews Genetics*, vol. 10, no. 8, pp. 540–550, 2009.

[15] M. C. Boerlijst, S. Bonhoeffer, and M. A. Nowak, "Viral quasi-species and recombination," *Proceedings: Biological Sciences*, pp. 1577–1584, 1996.

[16] C. Briones, A. de Vicente, C. Molina-Pars, and E. Domingo, "Minority memory genomes can influence the evolution of HIV-1 quasispecies in vivo," *Gene*, vol. 384, pp. 129–138, 2006.

[17] C. Briones and E. Domingo, "Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications," *AIDS Rev*, vol. 10, p. 93109, 2008.

[18] A. M. N. Tsibris, B. Korber, R. Arnaout, C. Russ, C. C. Lo, T. Leitner, B. Gaschen, J. Theiler, R. Paredes, and Z. Su, "Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo," *PloS one*, vol. 4, no. 5, 2009.

[19] S. Palmer, M. Kearney, F. Maldarelli, E. K. Halvas, C. J. Bixby, H. Bazmi, D. Rock, J. Falloon, R. T. Davey, R. L. Dewar, J. A. Metcalf, S. Hammer, J. W. Mellors, and J. M. Coffin, "Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in Treatment-Experienced patients are missed by standard genotype analysis," *J. Clin. Microbiol.*, vol. 43, no. 1, pp. 406–413, 2005.

[20] M. Wirden, I. Malet, A. Derache, A. G. Marcelin, B. Roquebert, A. Simon, M. Kirstetter, L. M. Joubert, C. Katlama, and V. Calvez, "Clonal analyses of HIV quasispecies in patients harbouring plasma genotype with K65R mutation associated with thymidine analogue mutations or L74V substitution," *AIDS*, vol. 19, no. 6, p. 630, 2005.

[21] J. Lu, S. G. Deeks, R. Hoh, G. Beatty, B. A. Kuritzkes, J. N. Martin, and D. R. Kuritzkes, "Rapid emergence of enfuvirtide resistance in HIV-1-infected patients: results of a clonal analysis," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 43, no. 1, p. 60, 2006.

[22] S. Kassaye, E. Lee, R. Kantor, E. Johnston, M. Winters, L. Zijenah, P. Mateta, and D. Katzenstein, "Drug resistance in plasma and breast milk after single-dose nevirapine in subtype c HIV type 1: population and clonal sequence analysis," *AIDS research and human retroviruses*, vol. 23, no. 8, pp. 1055–1061, 2007.

[23] C. Wang, Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R. W. Shafer, "Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance," *Genome research*, vol. 17, no. 8, p. 1195, 2007.

[24] J. Archer, M. S. Braverman, B. E. Taillon, B. Desany, I. James, P. R. Harrigan, M. Lewis, and D. L. Robertson, "Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4-using HIV-1 with ultra-deep pyrosequencing," *AIDS (London, England)*, vol. 23, no. 10, p. 1209, 2009.

[25] N. Beerenwinkel and O. Zagordi, "Ultra-deep sequencing for the analysis of viral populations," *Current Opinion in Virology*, vol. In Press, Corrected Proof. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1879625711000629

[26] M. Lataillade, J. Chiarella, R. Yang, S. Schnittman, V. Wirtz, J. Uy, D. Seekins, M. Krystal, M. Mancini, and D. McGrath, "Prevalence and clinical significance of HIV drug resistance mutations by Ultra-Deep sequencing in Antiretroviral-Nave subjects in the CASTLE study," *PloS one*, vol. 5, no. 6, p. e10952, 2010.

[27] B. B. Simen, M. Braverman, I. Abbate, J. Aerssens, Y. Bidet, O. Bouchez, C. Gabriel, J. Izopet, H. Kessler, A. Radonic, K. Metzner, R. Paredes, P. Recordon-Pinson, J. Sakwa, G. Schmitz-Agheguian, and M. Daumer, "A multicentre collaborative study on HIV drug resistance testing using 454 massively parallel pyrosequencing," *Antiviral Therapy*, vol. 15 Supp, no. 2, p. A37, 2010.

[28] S. Margeridon-Thermet, N. S. Shulman, A. Ahmed, R. Shahriar, T. Liu, C. Wang, S. P. Holmes, F. Babrzadeh, B. Gharizadeh, and B. Hanczaruk, "Ultra-deep pyrosequencing of hepatitis b virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)treated patients and NRTI-naive patients," *Journal of Infectious Diseases*, vol. 199, no. 9, p. 1275, 2009.

[29] N. Eriksson, L. Pachter, Y. Mitsuya, S. Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel, "Viral population estimation using pyrosequencing," *PLoS Computational Biology*, vol. 4, no. 5, 2008.

[30] K. Westbrooks, I. Astrovskaya, D. Campo, Y. Khudyakov, P. Berman, and A. Zelikovsky, "HCV quasispecies assembly using network flows," in *Proceedings of the 4th international conference on Bioinformatics research and applications*. Springer-Verlag, 2008, pp. 159–170.

[31] I. A. Astrovskaya, "Inferring genomic sequences," 2011.

[32] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel, "ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data," *BMC bioinformatics*, vol. 12, no. 1, p. 119, 2011.

[33] M. Pop, S. L. Salzberg, and M. Shumway, "Genome sequence assembly: Algorithms and issues," *Computer*, pp. 47–54, 2002.

[34] S. Istrail, G. G. Sutton, L. Florea, A. L. Halpern, C. M. Mobarry, R. Lippert, B. Walenz, H. Shatkay, I. Dew, J. R. Miller *et al.*, "Whole-genome shotgun assembly and comparison of human genome assemblies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 7, p. 1916, 2004.

[35] R. M. Idury and M. S. Waterman, "A new algorithm for DNA sequence assembly," *Journal of Computational Biology*, vol. 2, no. 2, p. 291306, 1995.

[36] A. Sundquist, M. Ronaghi, H. Tang, P. Pevzner, and S. Batzoglou, "Whole-genome sequencing and assembly with high-throughput, short-read technologies," *PLoS One*, vol. 2, no. 5, p. e484, 2007.

[37] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de bruijn graphs," *Genome research*, vol. 18, no. 5, p. 821, 2008.

[38] P. A. Pevzner, H. Tang, and M. S. Waterman, "An eulerian path approach to DNA fragment assembly," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 17, p. 9748, 2001.

[39] M. J. Chaisson and P. A. Pevzner, "Short read fragment assembly of bacterial genomes," *Genome Research*, vol. 18, no. 2, p. 324, 2008.

[40] J. Laserson, V. Jojic, and D. Koller, "Genovo: De novo assembly for metagenomes," in *Research In Computational Molecular Biology*. Springer, 2010, pp. 341–356.

[41] R. Lippert, R. Schwartz, G. Lancia, and S. Istrail, "Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem," *Briefings in bioinformatics*, vol. 3, no. 1, p. 23, 2002.

[42] B. V. Halldrsson, D. Aguiar, and S. Istrail, "Haplotype phasing by Multi-Assembly of shared haplotypes: Phase-Dependent interactions between rare variants," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2011, p. 88.

[43] S. M. Willerth, H. A. M. Pedro, L. Pachter, L. M. Humeau, A. P. Arkin, D. V. Schaffer, and J. P. Vartanian, "Development of a low bias method for characterizing viral populations using next generation sequencing technology," *PloS one*, vol. 5, no. 10, pp. 255–264, 2010.

[44] O. Zagordi, R. Klein, M. Daumer, and N. Beerenwinkel, "Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies," *Nucleic Acids Research*, 2010.

[45] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, and H. Takahashi, "Sequence-specific error profile of illumina sequencers," *Nucleic Acids Research*, 2011.

[46] C. Jabara, C. Jones, J. Anderson, and R. Swanstrom, "Accurate sampling and deep sequencing HIV-1 protease using primer ID," 2011.

[47] A. Ritz, A. Bashir, and B. J. Raphael, "Structural variation analysis with strobe reads," *Bioinformatics*, vol. 26, no. 10, p. 1291, 2010.

[48] C. Schulte, M. Lagerkvist, and G. Tack, "Gecode," *Software download and online material at the website: http://www. gecode. org*.

[49] B. Korber, B. T. Foley, C. L. Kuiken, S. K. Pillai, and J. G. Sodroski, "Numbering positions in HIV relative to HXB2CG," *Human retroviruses and AIDS*, vol. 3, pp. 102–111, 1998.

[50] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*. WH Freeman & Co., 1979.

[51] T. R. Jensen, *Graph coloring problems*. John Wiley & Sons, 1994.

[52] C. Bron and J. Kerbosch, "Algorithm 457: finding all cliques of an undirected graph," *Communications of the ACM*, vol. 16, no. 9, pp. 575–577, 1973.

[53] C. Schulte, "Programming constraint inference engines," *Principles and Practice of Constraint Programming-CP97*, pp. 519–533, 1997.

[54] L. T. Bacheler, E. D. Anton, P. Kudish, D. Baker, J. Bunville, K. Krakowski, L. Bolling, M. Aujay, X. V. Wang, and D. Ellis, "Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy," *Antimicrobial agents and chemotherapy*, vol. 44, no. 9, p. 2475, 2000.

[55] B. T. Korber, C. Brander, B. F. Haynes, R. Koup, J. P. Moore, B. D. Walker, and D. I. Watkins, *HIV Molecular Immunology Compendium 2006/2007*. Los Alamos National Laboratory, 2007.

[56] O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel, "Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction," *Journal of Computational Biology*, vol. 17, no. 3, pp. 417–428, 2010.