3.4. *Phasing Components in the Fragment Conflict Graph*

Even with error-free data we aren't guaranteed to be able to assemble and phase the data. *Long runs of homozygosity form disconnected components in the fragment conflict graph.* Runs of homozygosity, which are paradoxically simple to phase, cause problems when assembling haplotypes. If the run of homozygosity is longer than the mate pair length no read can connect the two components as there wont be any conflicts in homozygous regions (Fig. 2). The more connected the graph is, the easier it is to phase because you have to eventually phase the shores of each component into two haplotypes. The number of valid haplotype phasings may therefore be large once the haplotypes of each individual have been assembled; if the haplotype assembly of a single individual consists of k disconnected bipartite components then there are 2^{k-1} unique ways to map the shores to haplotypes. Varying the mate pair read length, increasing the read length, adding coverage, or adding more individuals who may share a haplotype IBD help connect components together.

Fragments from haplotypes that are identical by descent can be considered when constructing bipartitions for both individuals. If two components need to be phased and one haplotype is shared then we'd expect the shared haplotype to have twice the coverage of the non-shared haplotype in both components, thus we phase the two shores with greater coverage from different components together. For example, Fig. 2 shows fragments from two haplotypes of two individuals one of which (10000001) is shared. The phasing of the two components is ambiguous but we know that the shared haplotype is likely to have approximately 50% more coverage. Therefore, it is more likely to phase the components such that we maximize the difference of cardinality between the phasings. For Fig. 2 the first phasing (10000001/00000000) yields $|6-3| = 3$ while the second phasing (10000000/00000001) yields $|5-4| = 1$. When phasing disconnected components where sharing is not known, the resulting phasing should try to minimize the difference of cardinality in the overall phasing.

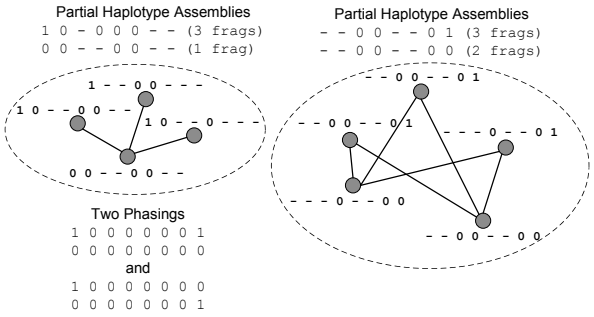

Fig. 2. Fragment conflict graph separated by a run of homozygosity. We assume the maximum distance between fragments is 2 SNPs.

4. Results on Simulated Data

We ran simulations on individuals of the CEU and JPT populations from HapMap[c]. First we sampled individuals randomly and then isolated a subset of the haplotype (30 SNPs for

[c]CEU denotes Utah residents with Northern and Western European ancestry and JPT denotes Japanese individuals from Tokyo, Japan.

visualization purposes). We placed the SNPs from the phased HapMap haplotypes a uniform distance from each other (500bp). Genome length is calculated by *number of SNPs* $\times$ *distance between SNPs*. The distance between sequence reads is calculated using a Poisson distribution and is varied under different models because most NGS technologies are capable of varying the distances between reads (e.g. Solexa or SOLiD). The average read length and coverage are also varied. Figures 3 and 4 show simulations on two unrelated individuals (one from CEU colored green, one from JPT colored red) while Fig. 5 shows simulations from two related individuals.

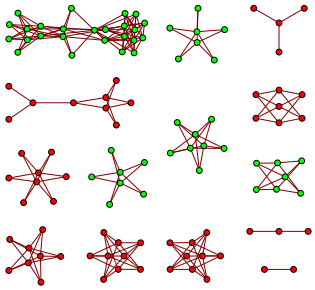

Fig. 3. Fragment conflict graph for unrelated individuals with coverage $c = 4$, read length $L = 50$, and distance between reads is Poisson with $\lambda = 2000$. Green vertices denote fragments originating from the CEU individual.

Figure 3 has many disconnected components due to the low probability of a good-read and regular distance between reads. Figure 4 shows the effect of changing the read length, coverage, and distance between reads. Read length, coverage, and variation of mate pair length correlate strongly with connectivity of the fragment conflict graph. In Fig. 5 two related individuals are shown with the same parameters used in Fig 3. It is clear the more sharing existing in the population, the easier it is to assemble and phase the data.

We also used our haplotype assembly simulator to test the accuracy and scalability of our minimum fragment removal heuristic. The first dataset we tested is the same 30 SNP segment from the HapMap CEU individual; the second dataset is a Hudson simulated chromosome of length 3434 SNPs. We decided to use the ratio of the number of erroneous fragments removed to the number of non-erroneous fragments removed as our metric. After the fragment conflict graph is generated, it may be advantageous to remove non-erroneous fragments to minimize our objective function. Nevertheless, this ratio is a good indicator of the quality of the output. For 1000 runs of the 30 SNP dataset we observed an overall ratio of 6.73; and for 100 runs of the 3434 SNP dataset we observed a ratio of 5.72. Further improvements to this type of algorithmic strategy for this problem is the subject of future work.

We've presented statistical estimates of coverage needed to cover a percentage of SNPs on a genome. These estimates could provide valuable insight when deciding sequence coverage per individual in association studies employing NGS technology. We've suggested a practical algorithmic strategy that exploits the high coverage possible with next-generation sequencing technology and the structure of errors in the fragment conflict graph. This algorithm produces promising results on the simulated fragment conflict graphs. We have presented an algorithm

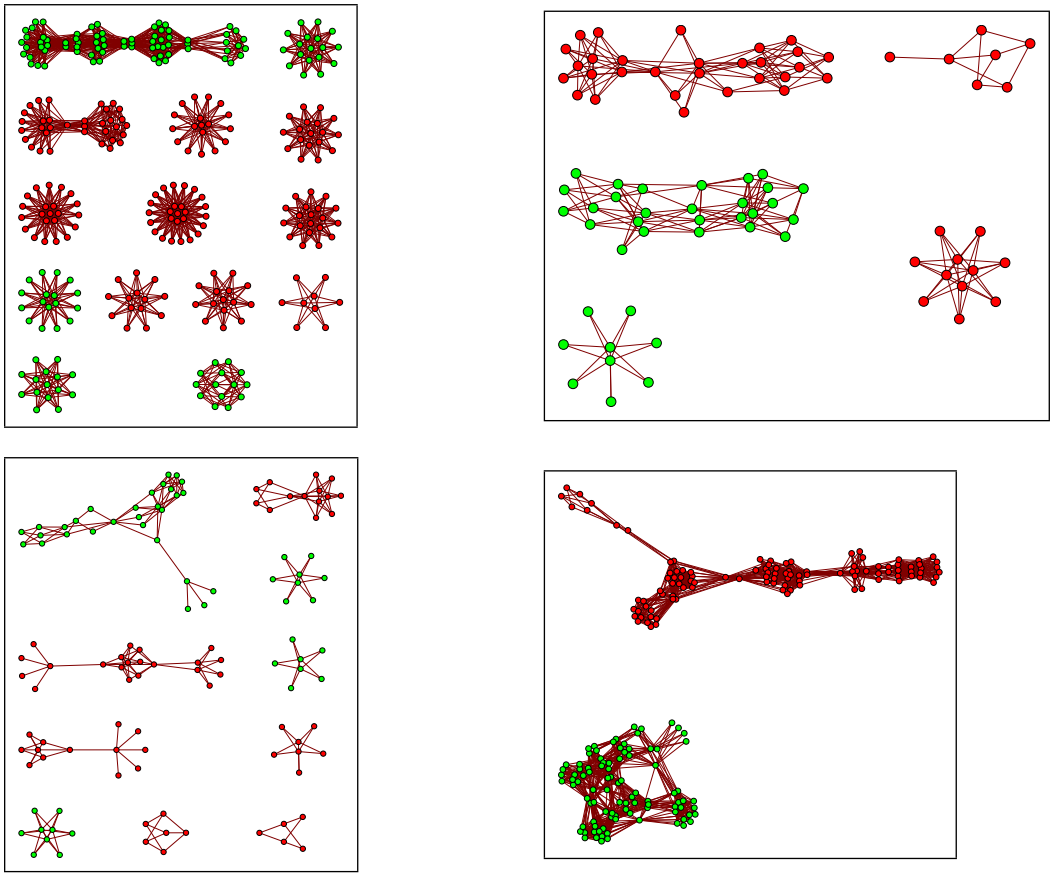

Fig. 4. Fragment conflict graph for two unrelated individuals. Green vertices denote fragments originating from the CEU individual. The baseline for each graph is: Read length $L = 50$; Coverage $c = 4$; distance between reads is Poisson with $\lambda = 2000$. From bottom left clockwise: (1) Distance between reads is Poisson with $\lambda = \{1000, 2000, 5000, 10000\}$ which is selected uniformly at random. (2) Coverage is changed to $c = 10$. (3) Read length is changed to $L = 1000$. (4) Coverage is $c = 10$, read length is $L = 1000$, and distance between reads is varied from $\{1000, 2000, 5000, 10000\}$.

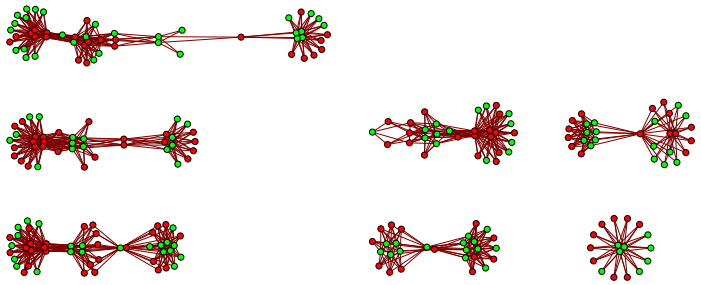

Fig. 5. Fragment conflict graph for two individuals sharing one haplotype. Green and red vertices denote fragments originating from different individuals. Read length $L = 50$. Coverage $c = 4$. The distance between reads is Poisson with $\lambda = 2000$.

for finding and exploiting haplotype sharing in the fragment conflict graph to enable the reliable phasing of disconnected components. We've also shown through simulation how various genomic and experimental parameters impact the quality of the haplotype assembly.

5. Acknowledgments

We thank the National Science Foundation for their support of this research.

References

1. B. V. Halldórsson, D. Aguiar, R. Tarpine and S. Istrail, *International Conference on Research in Computational Molecular Biology (RECOMB)* **6044**, 158 (2010).
2. B. V. Halldórsson *et al.*, *Lecture Notes in Computer Science* (2004).
3. R. Sharan, B. V. Halldórsson and S. Istrail, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3**, 303 (2006).
4. R. Lippert, R. Schwartz, G. Lancia and S. Istrail, *Brief Bioinform* **3**, 23 (March 2002).
5. G. Lancia, V. Bafna, S. Istrail, R. Lippert and R. Schwartz, *Proceedings of the 3rd European Symposium on Algorithms, (EAS01) Springer Lecture Notes in Computer Science* **2161**, 182 (2001).
6. S. Istrail *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 1916 (2004).
7. J. C. Venter, *Nature* **464**, 676 (April 2010).
8. R. Sladek *et al.*, *Nature* **445**, 881 (February 2007).
9. *Nature* **447**, 661 (June 2007).
10. R. J. Klein *et al.*, *Science* **308**, 385 (April 2005).
11. D. A. Hafler *et al.*, *N Engl J Med* **357**, 851 (2007).
12. N. J. Samani *et al.*, *N Engl J Med* **357**, 443 (August 2007).
13. M. N. Weedon *et al.*, *Nature Genetics* **40**, 575 (April 2008).
14. E. E. Eichler *et al.*, *Nature reviews. Genetics* **11**, 446 (June 2010).
15. J. McClellan and M.-C. King, *Cell* **141**, 210 (April 2010).
16. T. Walsh and M.-C. King, *Cancer Cell* **11**, 103 (February 2007).
17. J. C. Cohen *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1810 (February 2006).
18. A. A. Dror and K. B. Avraham, *Annual Review of Genetics* **43**, 411 (2009).
19. H. Stefansson, D. Rujescu, S. Cichon, O. P. H. Pietilainen *et al.*, *Nature* **455**, 232 (Sep 2008).
20. J. Sebat *et al.*, *Science* **316**, 445 (April 2007).
21. N. Siva, *Nature biotechnology* **26**, p. 256 (March 2008).
22. Wellcome trust launches effort to sequence 10,000 human genomes `http://www.genomeweb.com//node/943774?hq_e=el&hq_m=751896&hq_l=1&hq_v=672d790b6d`.
23. R. Schwartz, *Communications in Information and Systems* **10**, 23 (2010).
24. E. S. Lander and M. S. Waterman, *Genomics* **2**, 231 (April 1988).
25. V. Bansal, A. L. Halpern, N. Axelrod and V. Bafna, *Genome Research* **18**, 1336 (August 2008).
26. D. He *et al.*, *Bioinformatics* **26**, i183 (June 2010).
27. V. Bansal and V. Bafna, *Bioinformatics* **24**, i153 (August 2008).
28. Y. Wang *et al.*, *Computational Biology and Chemistry* **31**, 288 (August 2007).
29. J. Wang, M. Xie and J. Chen, *Algorithmica* **56**, 283 (March 2010).
30. S. Kang, I. Jeong, M. Choi and H. Lim, *Lecture Notes in Computer Science* **5059/2008**, 45 (2008).
31. S. Purcell *et al.*, *American journal of human genetics* **81**, 559 (September 2007).
32. S. R. Browning and B. L. Browning, *American journal of human genetics* **86**, 526 (April 2010).
33. Z. Li, L. Wu, Y. Zhao and X. Zhang, *Acta Mathematicae Applicatae Sinica* **22**, 405 (2006).
34. R. R. Hudson, *Bioinformatics* **18**, 337 (2002).
35. A. Kong, G. Masson, M. L. Frigge *et al.*, *Nat Genet* **40**, 1068 (Sep 2008).
36. M. J. Minichiello and R. Durbin, *Am J Hum Genet.* **79**, 910 (Nov 2006).