Layering for Genomics: iDASH Report January

Vineet Bafna* George Varghese*

1 January, 2011 update

- We propose continued collaborations with School of Medicine on the development of tools for analyzing genomic data.
  - *In collaboartion with Prof. Gabriel Haddad, we published a study of hypoxia tolerance in Drososphila (PNAS ePUB January 24, 2011), and are negotiating with BGI for sequencing of human highlander populations*[8].
  - *We published a novel approach to haplotyping using Probe sequencing technologies from Pacific Biosciences*[7].
- We propose to develop and publish a layered abstraction of specific software modules that process, map, compress and query the donor data for cataloging variations, with precise descriptions of the interfaces between modules. We hope our layering will provide a context for different software generated by the larger research community, and help with Life Tech's goal of standardizing tools for genome sequences.
  - *In progress. We are starting by implementing specific queries.*
- We propose a prototype implementation of two specific software layers. First, we will develop a compression layer, extending the ideas presented in Kozanitis et al., 2010. The second is an *evidence layer (EL)*. The EL is a collection of APIs that efficiently retrieve *all* raw data relevant to inferring specific variations. Note that our proposal separates out an evidence layer (EL) from a separate inference layer (IL), two layers that are commonly intertwined in existing software.
  - *We have developed a query for interrogating large structural deletions that runs on 40X human genome data in less than 60 minutes to identify all possible deletions.*

Previous publications[1;2;3;5;4;6?] , and helps guide the way genomes are being queried and analyzed.

As a first step towards efficient querying of genomic data, we have developed a tool for genome compression[5]. We have initated collaborations with Life Technologies, and Illumina, which will provide us with valuable data and access to end-users.

References

[1] G. Bhatia, V. Bansal, O. Harismendy, N. J. Schork, E. J. Topol, K. Frazer, and V. Bafna. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput. Biol.*, 6:e1000954, 2010.

[2] D. Brinza, M. Schultz, G. Tesler, and V. Bafna. RAPID detection of gene-gene interactions in genome-wide association studies. *Bioinformatics*, 26:2856–2862, Nov 2010.

[3] B. Dost, C. Wu, A. Su, and V. Bafna. TCLUST: A Fast Method for Clustering Genome-Scale Expression Data. *IEEE/ACM Trans Comput Biol Bioinform*, Apr 2010.

[4] O. Harismendy, G. Bhatia, N.G. Rahim, V. Bansal, M. Nakano, X. Scott, M. Wang, C. Dib, E. Turlotte, N. Heintzman, S. Murray, J.-F. Deleuze, J.C. Sipe, B. Ren, V. Bafna, E. J. Topol, and K. A. Frazer. Deciphering the genetic basis of common diseases by integrated functional annotation of common and rare variants. *Genome Biology*, 11:I8, October 2010.

[5] C. Kozanitis, C. Saunders, S. Kriglyak, V. Bafna, and G. Varghese. Compressing genomic sequence fragments using SlimGene. In *Proceedings of the Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 310–424, 2010. (Accepted for publication in Jnl. of Computational Biology).

*CSE 4218, Univ. California, San Diego. {vbafna,gvarghese}@cs.ucsd.edu

[6] S. Levy, G. Sutton, P.C. Ng, L. Feuk, A.L. Halpern, B.P. Walenz, N. Axelrod, J. Huang, E.F. Kirkness, G. Denisov, Y. Lin, J.R. MacDonald, A.W. Pang, M. Shago, T.B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S.A. Kravitz, D.A. Busam, K.Y. Beeson, T.C. McIntosh, K.A. Remington, J.F. Abril, J. Gill, J. Borman, Y.H. Rogers, M.E. Frazier, S.W. Scherer, R.L. Strausberg, and J.C. Venter. The diploid genome sequence of an individual human. *PLoS Biol.*, 5:e254, Sep 2007.

[7] Christine Lo, Ali Bashir, Vikas Bansal, and Vineet Bafna. Strobe sequence design for haplotype assembly. *BMC Bioinformatics*, 12(Suppl 1):S24, 2011.

[8] D. Zhou, N. Udpa, M. Gersten, D. Visk, A. Bashir, J. Xue, K. Frazer, J. Posakony, S. Subramaniam, V. Bafna, and G. Haddad. Experimental selection of hypoxia-tolerant Drosophila melanogaster: Major role for Notch activation. PNAS (ePUB 2011).