Computational challenges in genome wide association studies: data processing, variant annotation and epistasis

Pablo Cingolani

PhD.

School of Computer Science

McGill University

Montreal,Quebec

March 2015

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Pablo Cingolani 2015

CHAPTER 1
Conclusions

1.1 Contributions

In this report we show three steps involved in the analysis of sequencing data and identifying the links to disease. Each step is characterized by very different problems that need to be addressed.

i) The first step is to reduce large amounts of information generated by high throughput experiments into a manageable subset. In our case, it involves reducing the raw sequencing information to a variant call set, but it could be any other features to be analyzed (RNA expression, transcript structure, enrichment peaks, genome reference assembly, etc.). This is mainly done by mapping reads into a reference genome and then using variant call algorithms. This step is characterized by requiring fast parallel algorithms and usually, due to the amount of data involved, I/O can be one of the bottlenecks. Algorithm that work on "chunks of data" instead of the whole data-set are preferred, and in many cases exist, because it makes the problem trivial to parallelize. Usually several stages of these highly specialized algorithms are combined into a "data analysis pipeline". Programming data analysis pipelines is not trivial since it requires process coordinations, robustness, scalability and flexibility (data pipelines, particularly in research environments, tend to change often). Although many solutions are available (usually in the form of libraries), these tend to make pipeline programming cumbersome or create new programming paradigms thus introducing steep learning curves. In Chapter **??**, we solve problems related to pipeline programming in a

novel way by creating a new programming language, BDS, that simplifies the creation of robust, scalable and flexible data pipelines. Although the main goal was managing our sequencing data pipelines, BDS is a flexible datacenter-scale programming language that can be applied to many large data pipelines (a.k.a. Big Data problems).

ii) The second step in our data analysis, consists of functional annotations, prioritization and filtering. The main concern in the annotation step performing an adequate filtering of what should be considered relevant variants for our experiment from irrelevant ones. Functional annotation of genomic variants was until not long ago an unsolved problem and shortly after created SnpEff & SnpSift, they quickly became widely adopted by the research community. In Chapter **??** we describe the challenges of variant annotations and some of the solutions we implemented in our algorithms.

iii) Finally, in Chapter **??**, we analyse the problem of finding genetic links to complex disease. This is known to be a difficult problem affected by several hidden co-factors that bias the results (e.g. population structure). Furthermore there are unsolved problems, such as missing heritability, implying that genomic links to complex disease may not be found using traditional GWAS methodologies. We believe that alternative models that combine higher level information, may help to boost statistical significance.

iii.a) We were involved in two major projects on GWAS of type II diabetes using: a) cohorts of multi-ethnic unrelated individuals and b) family pedigrees. Results uncovered new genes linked to diabetes. Also, the studies indicate that one of the main hypothesis in the field, the "Rare variant hypothesis", might not hold strong.

iii.b) We propose a new methodology for addressing a difficult problem: detection of two interacting genomic loci that affect disease risk. Our models combine genotype information and co-evolutionary methods. We show that efficient algorithms make these studies computationally feasible, albeit using large computational resources, and we apply them to real data form type II diabetes sequencing study of over 13,000 individuals.

These three Chapters (three steps) complete our journey from "raw data" to "biological insight" trying to find the genetic causes of complex disease.

1.2 Future work

Here we propose several improvements, extensions and future lines of work for each of the methods developed in this thesis (some of them are currently being developed / explored):

BigDataScript. We are adding native support for new clusters and frameworks, such as LSF, Mesos, Kubertes as well as a *"Generic cluster"* API which allows the user to customize BigDataScript for any cluster or framework by encapsulating task management via user defined scripts. On the language specification side, we are exploring ways to add functional constructs such as `map`, `apply`, `filter` as well as support for *map/reduce* and *scatter/gather* which are convenient ways to define some problems in data pipeline programming. Finally, will be incorporating user defined data structures or a basic class mechanism (BDS currently supports maps and list).

Annotations. We are creating new annotation standard for VCF files, in an effort coordinated with the developer of other annotations tools (such as ANNOVAR, ENSEMBLs Variant effect predictor -VEP-, JAnnovar, etc.). We are actively contributing with the *"Global Alliance for genomic and Health"*

(GA4GH) in the variant annotation specification & API definitions. We plan to extend SnpEff's variant annotations capabilities to *haplotype-based* annotations, which means taking into account phasing information (either by phasing trios or "read phasing") to calculate compound variant effects (e.g. phased SNPs affecting the same codon or compensating frame shifts within the same DNA strand). Finally, we are using information theoretic analysis of splice sites from several species in order to improve splicing effect predictions.

GWAS Epistasis. As future work, we'd like to evaluate the possibility of incorporating contextual information, such as protein domain, in order to build more specific co-evolutionary models. Other improvements include further optimization of logistic regression and Bayes factor algorithms since any improvement greatly reduces computational times. We also plan to use our methods on even larger type II diabetes cohorts that are currently being sequenced. Finally, we are evaluating the possibility of incorporating higher order interactions by clustering genes from our variant-pairs analysis and then evaluate them in a joint analysis.

1.3 Perspectives

Genomic research for complex disease is trending towards larger and larger cohorts in order to improve statistical power. Some years ago, projects involving hundreds to a thousand individuals were common. To put this in perspective, thats the population of a village, or a small town. Nowadays, projects like the T2D consortia, sequence in the order of 20,000 people (i.e. the population of a large town). I am aware, through personal communications with other researchers, that projects being drafted for sequencing over 100,000 individuals (i.e. the population of a whole city). This quest for ever bigger sample sizes shows how elusive the genetic causes of complex diseases are.

The methods developed here aim to help in the processing of these huge datasets (BDS), annotate and prioritize the variants (SnpEff) before testing for significance. But also help in looking at these variants from another perspective (epistatic GWAS) than the traditional "single variant association" approach. It might be true that huge sample sizes are needed to uncover risk loci, but perhaps one of the reasons why traditional GWAS studies are not finding as many associations as expected is just that we they are looking in the wrong place. In science we must explore all possibilities.

References

[1] Marit Ackermann and Andreas Beyer. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS genetics*, 8(2):e1002463, 2012.

[2] D.J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[3] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715, 2004.

[4] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[5] Lukas Burger and Erik van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, 6(1):e1000633, 2010.

[6] Greg W Clark, Sharon H Ackerman, Elisabeth R Tillier, and Domenico L Gatti. Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments. *BMC bioinformatics*, 15(1):157, 2014.

[7] G.M. Clarke, C.A. Anderson, F.H. Pettersson, L.R. Cardon, A.P. Morris, and K.T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[8] Onofre Combarros, Mario Cortina-Borja, A David Smith, and Donald J Lehmann. Epistasis in sporadic alzheimer's disease. *Neurobiology of aging*, 30(9):1333–1349, 2009.

[9] Diabetes SAT2D Consortium, Diabetes MAT2D Consortium, Anubha Mahajan, Min Jin Go, Weihua Zhang, Jennifer E Below, Kyle J Gaulton, Teresa Ferreira, Momoko Horikoshi, Andrew D Johnson, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–244, 2014.

[10] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[11] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[12] Joseph Felsenstein and Joseph Felenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004.

[13] Hong Gao, Julie M Granka, and Marcus W Feldman. On the classification of epistatic interactions. *Genetics*, 184(3):827–837, 2010.

[14] Chern-Sing Goh, Andrew A Bogan, Marcin Joachimiak, Dirk Walther, and Fred E Cohen. Co-evolution of proteins with their interaction partners. *Journal of molecular biology*, 299(2):283–293, 2000.

[15] Steven N Goodman. Toward evidence-based medical statistics. 2: The bayes factor. *Annals of internal medicine*, 130(12):1005–1013, 1999.

[16] L Guariguata, DR Whiting, I Hambleton, J Beagley, U Linnenkamp, and JE Shaw. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice*, 103(2):137–149, 2014.

[17] Christopher A Hunter, Juswinder Singh, and Janet M Thornton. π-π interactions: The geometry and energetics of phenylalanine-phenylalanine interactions in proteins. *Journal of molecular biology*, 218(4):837–846, 1991.

[18] David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 2012.

[19] Luke Jostins, Adam P Levine, and Jeffrey C Barrett. Using genetic prediction from known complex disease loci to guide the design of next-generation sequencing experiments. *PloS one*, 8(10):e76328, 2013.

[20] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome browser database: 2014 update. *Nucleic acids research*, 42(D1):D764–D770, 2014.

[21] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[22] Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.

[23] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, page gkt1113, 2013.

[24] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[25] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.

[26] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[27] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

[28] et. al. McCarthy, Mark. Variation in protein-coding sequence and predisposition to type 2 diabetes. *To be published*, 2015.

[29] Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

[30] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[31] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segré, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

[32] Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, and Alfonso Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, 271(4):511–523, 1997.

[33] Patrick C Phillips. Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[34] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[35] Mario X Ruiz-González and Mario A Fares. Coevolution analyses illuminate the dependencies between amino acid sites in the chaperonin system groes-l. *BMC evolutionary biology*, 13(1):156, 2013.

[36] Clara M Santiveri and M Jiménez. Tryptophan residues: Scarce in proteins but strong stabilizers of β-hairpin peptides. *Peptide Science*, 94(6):779–790, 2010.

[37] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part i. experimental techniques and databases. *PLoS computational biology*, 3(3):e42, 2007.

[38] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS computational biology*, 3(4):e43, 2007.

[39] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[40] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[41] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, et al. An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90, 2009.

[42] Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic epidemiology*, 33(1):79–86, 2009.

[43] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[44] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[45] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[46] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 2011.

[47] Ziheng Yang. *Computational molecular evolution*, volume 284. Oxford University Press Oxford, 2006.

[48] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

[49] Jinying Zhao, Li Jin, and Momiao Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, 2006.

[50] O. Zuk, E. Hechter, S.R. Sunyaev, and E.S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[51] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.