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

Pablo Cingolani

PhD.

School of Computer Science - Bioinformatics

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

# CHAPTER 1
# Conclusions

## 1.1 Contributions

In this report we showed the 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 2, we solved the 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 3 we described the challenges of variant annotations and some of the solutions we implemented in our algorithms.

iii) Finally, in Chapter 4, we analyzed 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 proposed 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 26,000 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 or future lines of work for each of the methods developed in this thesis (some of them are currently being developed / explored):

BDS
(a) Native support for new clusters and frameworks (that now supported via "Generic cluster"): LSF, Mesos, Kubertes.
(b) Functional constructs: map, apply, filter. This allows for more compact and readable code.
(c) Richer data structures: BDS currently supports maps and list but does not support user defined structures.

SnpEff
(a) Creation of a new VCF annotation standard coordinated with the developer of other annotations tools (mainly ENSEMBLs VEP and ANNOVAR).
(b) GA4GH variant annotation specification & API definition.
(c) Haplotype effect predictions: Using phased (or "read phasing") to calculate compound variant effects (e.g. consecutive phased SNPs forming an MNP or two compensating frame shifts).

(d) Improved loss of functions predictions.

(e) Improved splice predictions using information theoretic analysis of splice sites from several species.

GWAS Epistasis

(a) Further optimization in logistic regression analysis:faster computations boosts program performance significantly.

(b) Analysis of context dependent Q2 matrices based on protein domains.

(c) Improved calculation of Bayesian priors.

### 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.