Enabling a Multivariate Strategy for Genotyping Quality Control as a Grid Service

Malovini A[1,3], Nuzzo A[2], Puca AA[3], Bellazzi R[1]

Motivation

The vast amount of molecular data generated by high-throughput techniques requires robust computational approaches already at the pre-processing stage, which is a heavily error-prone processe. In particular, quality control of genotyping data is based on parameters whose setting is often unclear and subjective, leading to a lack of reproducibility, and heavily affecting the final results. A formal approach is then needed for parameters tuning. As for other pre-processing tasks involved in high-throughput experiments, computational implementations greatly benefit from a High-Performance Computing infrastructure.

Methods

Experimental genotyping errors in Genome-wide association studies (GWAS) can lead to false positive findings and therefore to spurious associations. The Quality Control (QC) phase, which is needed to minimize the effects of this kind of errors, relies on filtering procedures aimed at identifying: i) individual samples with errors across multiple markers (problems with the DNA), and ii) SNPs yielding errors in multiple individuals (marker-affecting errors). Several criteria (genotyping rate, Hardy-Weinberg Equilibrium, samples heterozygote rate, minor allele frequency, genomic inflation factor, etc) can be used to evaluate the effect of the removal of SNPs and individuals, but the choice of the most appropriate threshold for this filtering is usually based on visual inspection of the data plots, looking for a tradeoff between losing samples or missing potentially associated SNPs. In order to make this process more reproducible we propose two strategies based on the Multi-Criteria Decision Making theory for setting appropriate genotyping call rate (CR) thresholds, with the final goal of maximizing the study power, which means removing as few individual samples and markers as possible, while minimizing the genotyping error rate. In the first strategy, called Simple Multi-Attribute Rating Technique (SMART) the decision maker is required to answer a pairwise comparison question about the relative importance of a set of QC criteria. The second strategy implements a different procedure for criteria weights assignment, based on direct elicitation of user preferences (D-MCDM) using a 0 to 10 scale. The best alternative for both strategies is the highest scored one. These methods are based on a comparison of different combinations of samples and SNP CR

[1] Department of Computer Engineering and Systems Science, University of Pavia, Italy [2] Centre for Tissue Engineering, University of Pavia, Italy [3] IRCCS Multimedica, Milan, Italy

thresholds, which even for a small GWAS (300K SNPs for few hundreds of patients) requires a large computational effort. Thus, a parallelization strategy has been studied and applied to the overall analysis process in order to increase computing performance on a Grid infrastructure and make it available in a reasonable time. The service module submits commands to the statistical programs R and Plink through an automated pipeline, exploiting the available high performance computing resources. The Grid portal providing this service is interfaced with two different environments: an IBM cluster based on the Platform LSF scheduler (http://www.platform.com) and the gLite (http://glite.web.cern.ch/glite) middleware. This environment has been chosen as it has already been developed and validated for microarray gene expression data and clinical data for survival analysis.

Results

A genetic association module has been integrated in the HPC platform, whose front-end, based on EnginFrame Grid Portal, provides users with customized Web interfaces, increasing application usability and productivity. We validated our methods on (i) a real dataset generated by an Arterial Hypertension GWAS on 734 cases and 486 controls genotyped using Illumina 317k SNPs and (ii) a larger simulated dataset. The results of the two strategies were comparable (best SMART alternative: “samples CR >95% and SNP CR >96%”; best D-MCDM alternative: “samples CR >95% and SNP CR >97%”). In particular, the two score profiles were very similar for samples with CR <95%, with comparable score profiles. For samples with CR >96% the interpretation of the two profiles is more complex: D-MCDM appears more “conservative”, penalizing stringent CR thresholds corresponding to a decrease in statistical power (the elicitation process of the criteria weights is done independently for each criteria), while SMART is able to take into account correlations between criteria and therefore it is related to smoother score functions. We also tested computational time requirements by simulating GWAS datasets with different sample sizes (1000, 2000, 3000 and 4000 samples) and marker densities (370K and 550K SNPs). The parallelization strategy leads to a decrease of one order of magnitude in computation time, from tens of hours required on a standard PC to about 15 minutes for a 500 cases-500 controls-370K SNPs dataset, and to about 2 hours for a 2000 cases-2000 controls-550K SNPs dataset.

Availability

http://ada.dist.unige.it:8080/enginframe/bioinf

Contact e-mail

angelo.nuzzo@unipv.it

Supplementary information

We gratefully acknowledge Dr. Livia Torterolo and Prof. Marco Fato from the University of Genoa for their support in deploying the algorithms on their Grid Platform