

GRID ENABLING FRAMEWORK FOR POPULATION SNPS GENETIC LINKAGE ANALYSIS



L. MILANESI, A. CALABRIA, D. DI PASQUALE, M. GNOCCHI, G. TROMBETTI and A. ORRO
 Istituto di Tecnologie Biomediche - Consiglio Nazionale delle Ricerche, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy

luciano.milanesi@itb.cnr.it



Motivation

The Genetic Linkage Analysis of SNPs (Single Nucleotide Polymorphism) markers enables to discover the genetic correlation in complex diseases following their transmission through family generations. The major algorithms proposed in literature require high efforts in terms of computational power and memory requirements, making large data sets very hard to be analyzed on a single CPU. The aim of the present work is to enable the use of the EGEE Grid Infrastructure for the execution of linkage analysis on very large SNPs data sets, setting up a web tool for achieving a Whole-Genome Linkage Analysis. Test cases have been performed with 10.000 to 1 million SNPs per Chip, as well as an application study on cardiac conduction disorders.

Implementation

The Genetic Linkage Analysis method is based on the calculation of the LOD score, a test parameter defined as the Logarithm Of Odds ratio between hypothesis of linkage versus the null one. To avoid the computational limits for the linkage algorithms, we adopt a heuristic method, splitting input data and running independent analyses of consecutive subsets of the genome, merging then together the results. The approach was verified with benchmark analysis run also in a single run on a Cluster: the major signal peaks of test datasets are well represented in our system's results (an example in Fig 1). The system is designed in three different layers: in the presentation layer a web based user interface helps users to setup, launch and monitor linkage analysis and finally to manage the results. The application layer preprocess and splits input data and prepares the Grid jobs; the HPC layer submits the jobs to the Grid middleware, featuring automatic resubmission management and corruption-checked data transfers through the implementation of the original software VNAS. When all tasks are computed, the results are retrieved, merged and made available through the web interface.

Performances

To evaluate the effective performance of the system, more than 25.000 jobs, including real data analysis and test calculations of different duration, has been launched through the resources of the Virtual Organization Biomed into the EGEE Grid. Analysis on pedigrees composed by 38 subjects, including individuals genotyped with different genotyping chips of markers from 10k up to 1 million of SNPs each, were submitted as job collections (also called *challenges*) and the resulting estimate of the infrastructure performances was compared to those of a single 2 GHz CPU workstation and of a Cluster composed by 280 CPU cores (Fig. 2). Comparing the results of the different computation

infrastructures it can be seen that distributed analysis pipelines with big datasets, i.e. with a high number of linkage variables, achieved a speedup up to more than 72x compared to a mid range dual-core 2 GHz CPU execution (Tab. 1). Considering markers chips greater than 100 k, the advantage of the distributed architecture gets proportionally bigger compared to the single CPU, due to the difference between linear increase of computational time for the sequential run and the saturation trend of the parallelized data flow obtained distributing the workload on the computing elements. Analyzing the distribution of the total execution times for a pool of 10.500 jobs with a running time of about 3 hours (Fig. 3), it should be noticed that more than half of the jobs are completed in less than 10% of the total completion

| Genotyping Chip (# of SNPs) | Jobs (# h) | Computational Cost (hours) | | | Speedup relative to the sequential execution | |
|-----------------------------|------------|----------------------------|---------|------|----------------------------------------------|------|
| | | Single 2GHz CPU | Cluster | Grid | Cluster | Grid |
| 10 k | 6 | 33 | 8 | 19 | 4.1 | 1.7 |
| 66 k | 35 | 220 | 9.5 | 28 | 23.2 | 7.9 |
| 100 k | 60 | 333 | 10 | 30 | 33.3 | 11.1 |
| 317 k | 172 | 1056 | 13 | 38 | 81.2 | 27.8 |
| 370 k | 206 | 1233 | 15 | 39 | 82.2 | 31.6 |
| 500 k | 278 | 1665 | 16 | 42 | 104.1 | 39.6 |
| 670 k | 373 | 2233 | 18 | 42 | 124.1 | 53.2 |
| 1 M | 556 | 3332 | 20 | 46 | 166.6 | 72.4 |

Table 1 - Infrastructures performances. From left to right: the challenges characteristics, their durations and the speedup relative to the sequential run.

time of the challenge they belong to: tails of few long lasting jobs affect in a heavy manner the overall performances. The big variability of the queuing system of the Grid environment may be also noticed examining the central graph in the right part of Fig. 3: the Figure shows the components in time of the average life span of a Grid job launched with our system.

Atrial Flutter Test Case

To confirm the effectiveness of the proposed approach, an analysis on the Atrial Flutter, a monogenic disease that concerns to cardiac conduction disorders, was performed. The patients of an isolated family were genotyped using the Infinium II Assay-HumanHap BeadChip 370k SNPs and the resulting dataset, after some preprocessing steps (search for Mendelian inconsistencies, markers filtering and linkage disequilibrium modeling), was processed with the proposed pipeline for multipoint parametric and nonparametric linkage analysis; heterogeneity LOD scores (HLOD) under dominant model, NPLOD scores and LOD score for individual families were computed. Candidate linkage regions were defined as those with NPLOD scores with associated p-values < 0.0001 and maximum LOD Score. The splitting procedure gave 305 files, with 40 SNPs: each job had a running time on a 2.8 GHz CPU of about 1.20 hours. The analysis was carried out in about 36 hours, showing a real benefit on using a distributed environment like the Grid. Results identify two candidate regions within the chromosomes 1 and 19, as showed in Fig. 4: a sequencing activity may confirm the association results.

Conclusions

We have developed the Genetic Linkage Analysis application in GRID in order to enable the user to perform large scale genetic genotype analysis over a EGEE distributed computational infrastructure.

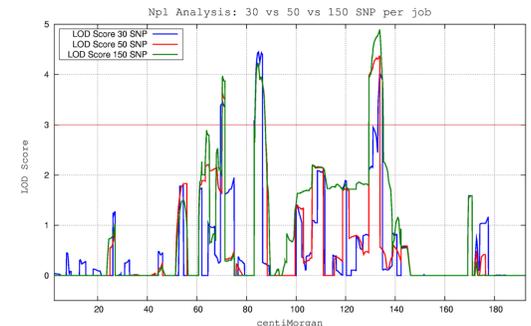


Figure 1. Methodology test comparisons and results plotting. In each run of the three execution from differently formatted input files (30 SNPs per job against 50 markers and 150 SNPs), it has been observed a similar behavior: all peaks for the significant values, Lod score greater than 3, have been retrieved from each analysis, even with an approximation error decreasing with the number of SNPs per job. The approximation introduced by splitting the data and merging the results is, therefore, acceptable.

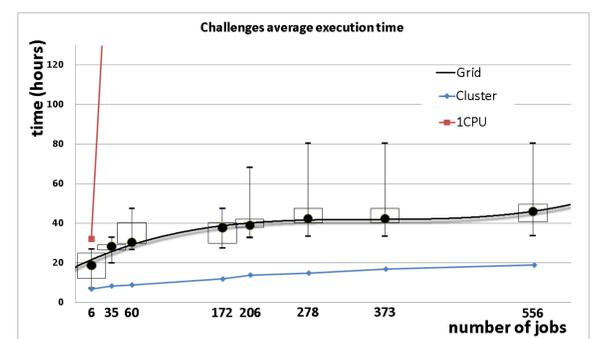


Figure 2 - Challenges average duration. Data derived from different genotyping chips were analyzed using 3 computational infrastructures: our Grid-based system, a 70 nodes and 280 cores Cluster and a single 2GHz CPU Work-station. The overall lasting time of 96 challenges (12 for each size composed by 10 to 556 jobs, with runtime of around 6 hours each) were considered to obtain this graph: boxes represent 25th and 75th percentile, whiskers represent minimum and maximum, dots indicate the median values. Red line shows performances of the single 2 GHz CPU, blue line is for Cluster.

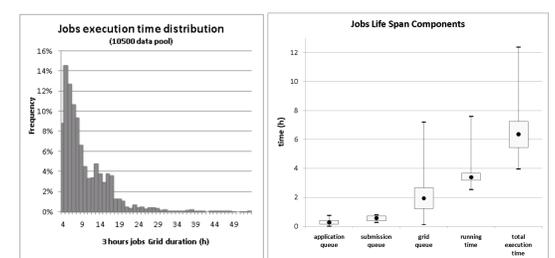


Figure 3 - Left: the distribution of total execution times for a pool of 10.500 test jobs with a running time of about 3 hours: the duration of each job is normalized with the total completion time of the relative challenge. Right: the duration of each component of the life cycle of the same pool of jobs. Boxes represent 25th and 75th percentile, whiskers represent minimum and maximum, dots indicate the median values.

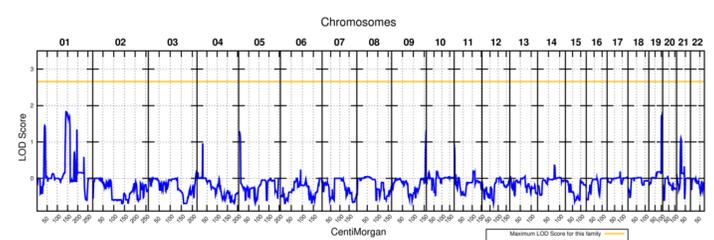


Figure 4 - The Lod Score of the whole genome analysis for Atrial Flutter adopting the heuristic approach on distributed grid environment. The peaks in the LOD score signal identify regions within the chromosomes 1 and 19 candidate for being linked to the disease.