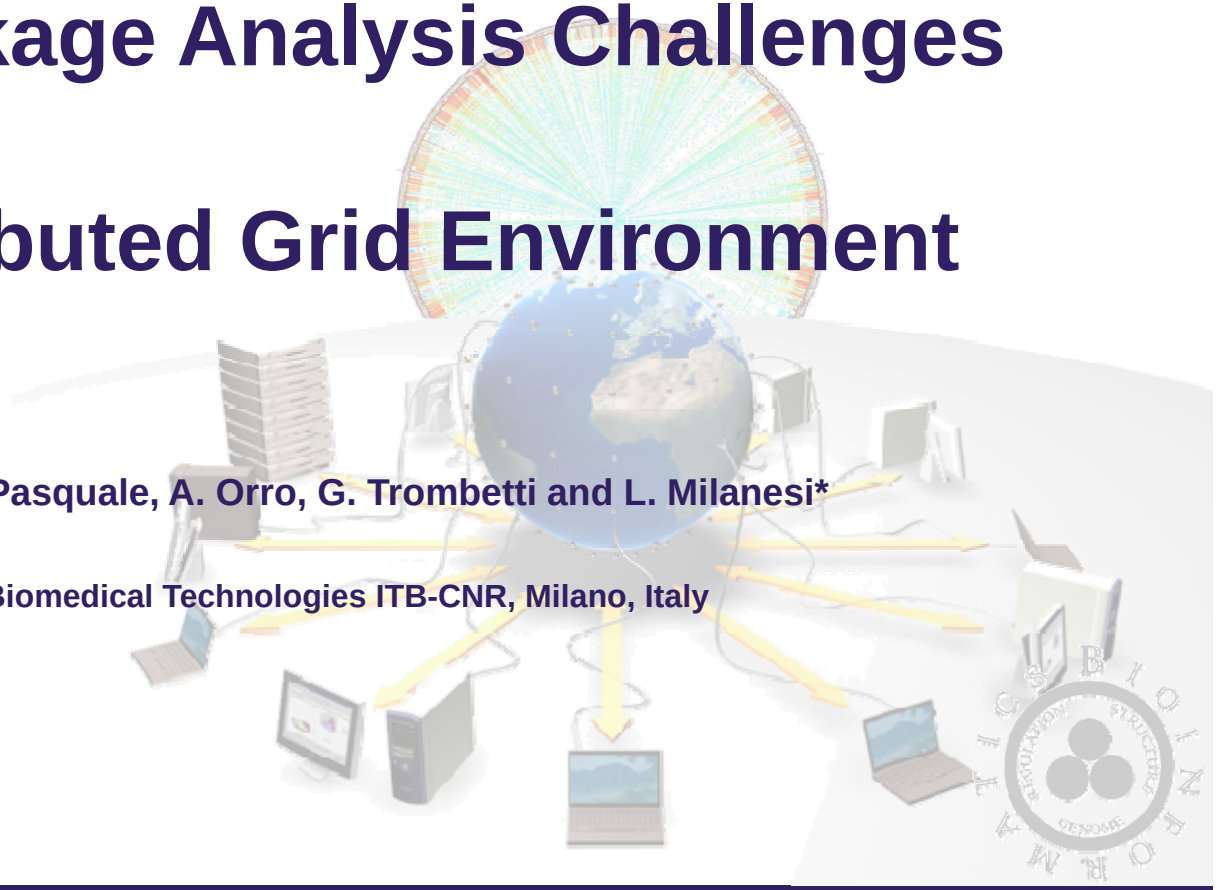
Bioinformatics

CNR

# Genetic Linkage Analysis Challenges On A Distributed Grid Environment

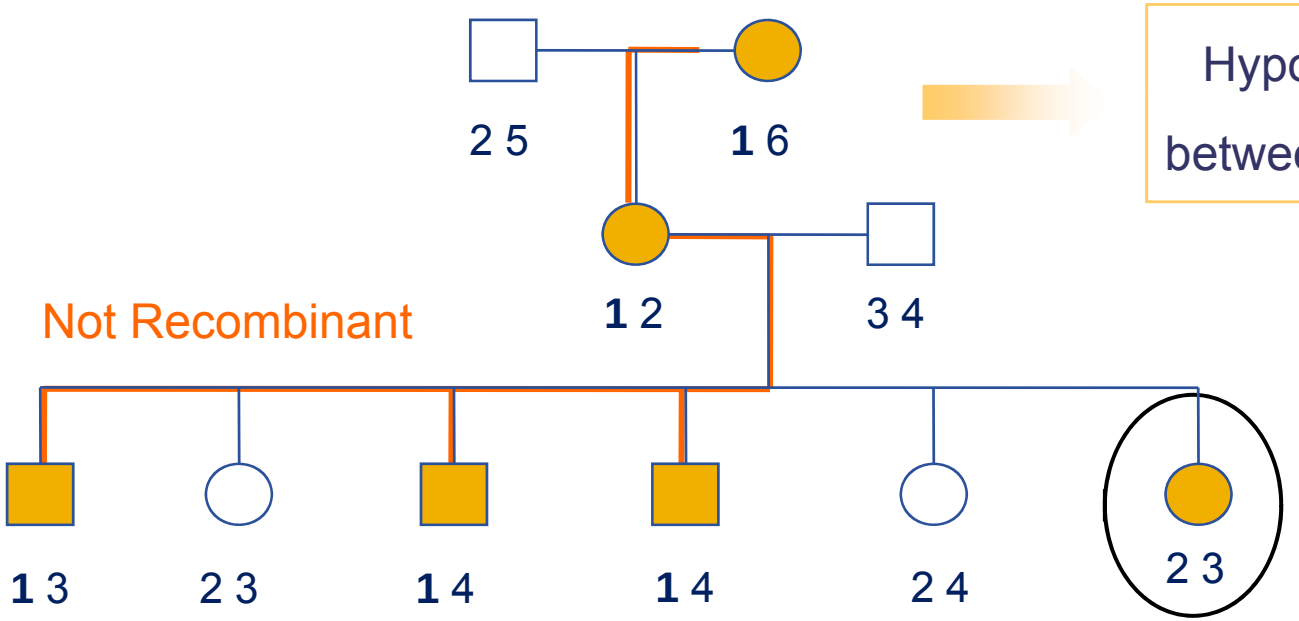


A. Calabria, D. Di Pasquale, A. Orro, G. Trombetti and L. Milanesi*

*Institute of Biomedical Technologies ITB-CNR, Milano, Italy

# Introduction

- The problem domain: Genetic Linkage Analysis
  - Pedigree example of recombination vs non rembination




Hypothesis H1 of linkage between allele 1 and disease

| | Healthy | Affected |
|---|---|---|
| Male | □ | ■ |
| Female | ○ | ● |

$$H_0 = L(1/2) = 0.5^6$$

$$H_1 = L(\theta) = (1-\theta)^5\theta$$

  - LOD Score Estimate

$$LOD = Z(\theta) = \log_{10}\frac{(1-\theta)^5\theta}{(0.5)^6}$$

# Objectives

- Compute Quantitative Computation of Linkage analysis with SNPs (biallelic markers)
  - Actual technologies for Chips collect more then 10.000 SNPs (whole genome)
  - Pedigrees are often large (more then 30 individuals)
  - Linkage analysis software are mostly not MPI or distributed
- Computational time and space on single CPU is not enough with these preconditions
- Need for distributed and high performance infrastructure and a system that enables linkage analysis with SNPs
  - Infrastructure: Grid technology is the best answare to distribute and improve efficiency
  - Application: a system which performs runnig linkage analysis challanges in grid environment adopting customizable workflows and user friendly access

# Methods

- **System's Design**
  - Logics to eneable distribution for grid environment
    - Choose linkage analysis software; ie: GenHunter
    - Split inputs (SNP or generic markers, and pedigree) into smaller sets having size smaller than bounds of the linkage analysis software chosen; ie: 370k SNP, 26 individuals → split SNP size into sets of 100, obtaining X jobs
    - Execute linkage analysis program N sets of the X jobs in parallel over Z working nodes
    - Monitor job's status, execution and outputs retrieving
  - Logics to ease access Grid technology
    - Create web access with standard technologies
    - Allow custom workflow creation of linkage analysis steps

# Methods

The system is designed in 3 different layers:

- the presentation layer

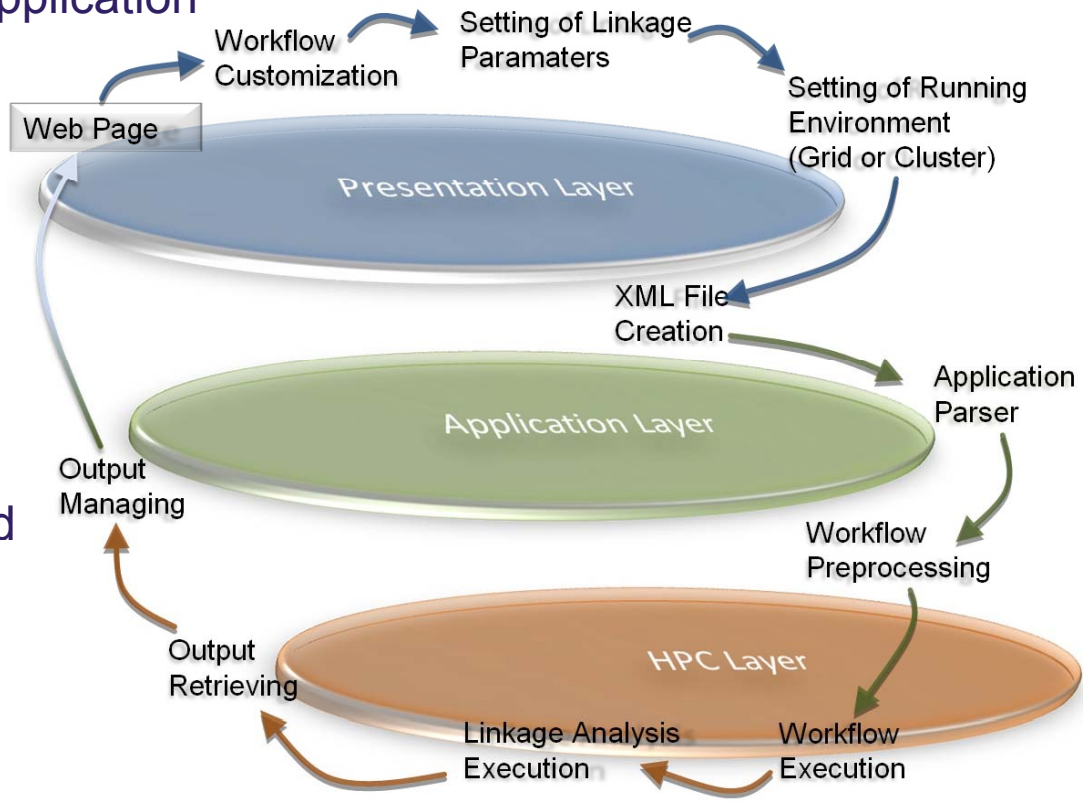
  where users interact with the application

- the application layer

  where are stored and run the logics of execution

- the HPC layer

  where interactions with the Grid middleware are managed



# Methods

- The HPC Layer

The workflow engine splits the workload into small jobs and distributes analysis tasks over the available resources

This is achieved by a software layer, called VNAS, built on top of the grid middleware which monitors each single grid process and ensures its elaboration success by managing the resubmission of failed jobs automatically .

When all tasks are computed the results are retrieved, merged and made available for downloading through the web interface.