EGEE'08 Conference
22-26 September 2008
Harbiye Askeri Museum, Istanbul - Turkey

Grid based genetic population analysis challenges for Genetics Linkage Analysis of SNPs.

Luciano Milanesi, Gabriele Trombetti, Andrea Calabria, Davide Di Pasquale, Matteo Gnocchi, Alessandro Orro.

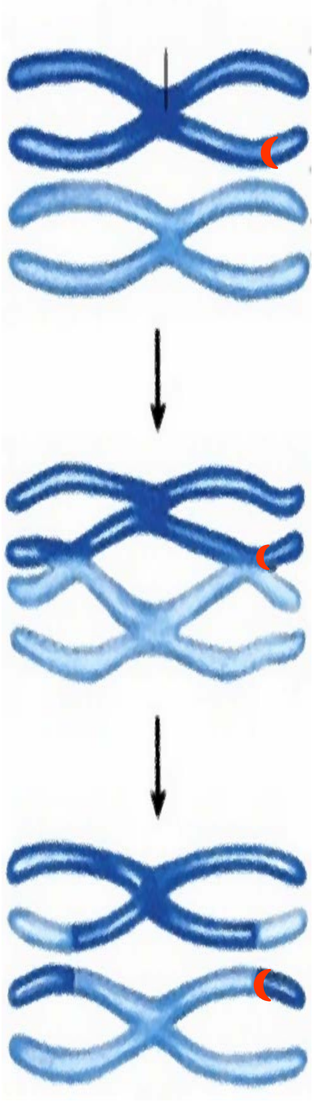
Institute of Biomedical Technologies ITB-CNR,
via F.lli Cervi 93, I-20090 Segrate (Milano), Italy
luciano.milanesi @itb.cnr.it

www.eu-egee.org

e-infrastructure

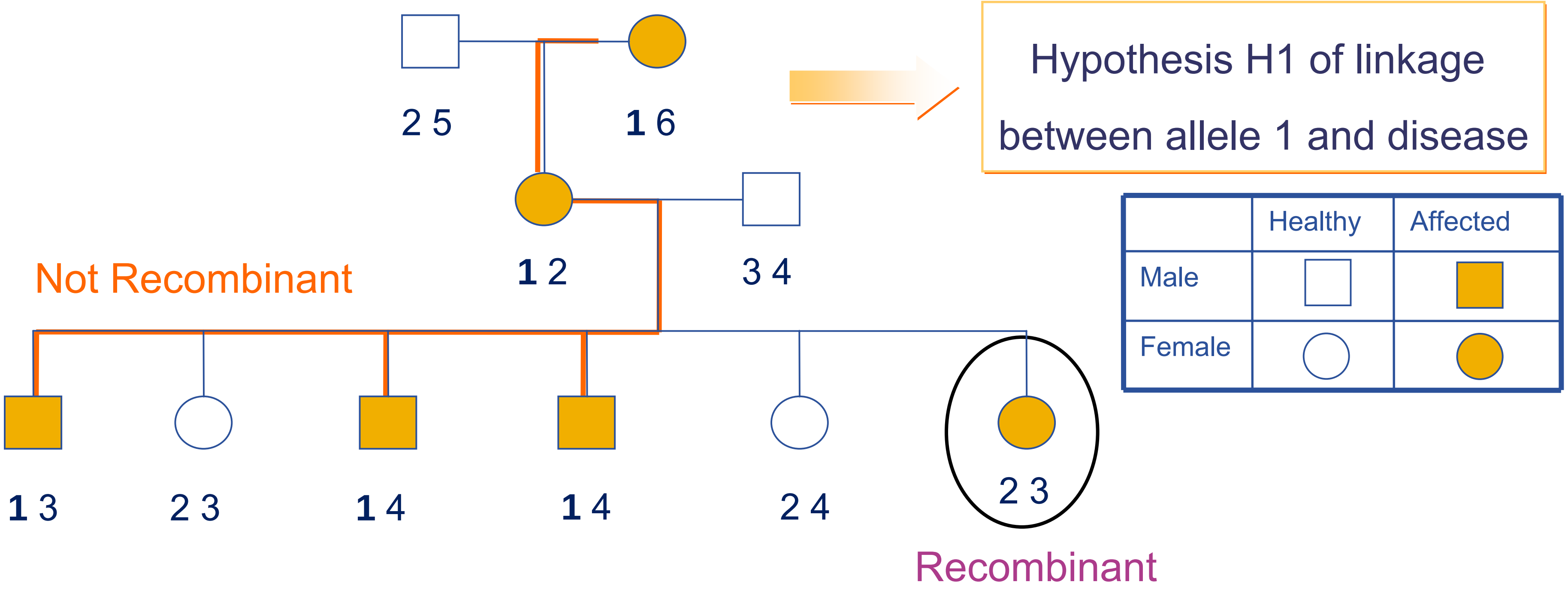
Introduction

- The human genome is very large and contains many thousands of genes.
- Traditionally, the search for a disease gene begins with Linkage Analysis that is a statistical method used to identify the location on a chromosome of a given gene involved in a disease relative to a known location of chromosome markers.
- Usually markers are Quantitative Traits Loci, Microsatellites, and their number is limited on the linkage analysis. Recently, also **SNPs** can be adopted as **biallelic markers**.
- New technologies on chips (for example Illumina) have developed SNP genotyping array, from 10000 SNPs to more than 1 million.



Introduction

- Genetic Linkage Analysis
  - Pedigree example of recombination vs non recombination



Hypothesis H1 of linkage between allele 1 and disease

	Healthy	Affected
Male	□	■
Female	○	●

  - LOD Score Estimate

Introduction

- The Linkage Analysis Problem is NP-hard

Algorithms (most important)	**Applications**	**Computational Bounds**
Elston-Stewart	Linkage, *SLink*, Fastlink, Vitesse, Mendel	n° loci: ~8 n° subjects: > 50
Lander-Green	*GeneHunter*, Allegro, Merlin	n° loci: > 20 n° subjects: ~20
Bayesian Networks	Superlink	n° loci: nr n° subjects: nr

Algorithms	**Increasing computational time by**		
	Individuals	**Loci**	**Time**
Elston-Stewart	Linear	Exponential	$O((m2^n)^p)$
Lander-Green	Exponential	Linear	$O(m2^{4p})$
Bayesian Networks	Linear	Linear	nr

Objectives

- Execute quantitative computation of Linkage analysis with SNPs (biallelic markers)
  - Actual technologies for Chips collect up to **1.000.000** 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** can be a possible answare 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**
  - Our system relies on the grid middleware for low level interactions with the hardware resources
  - 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
    - Create a workflow for the linkage analysis steps

Methods

The system is designed in 3 different layers:

- **the presentation layer**

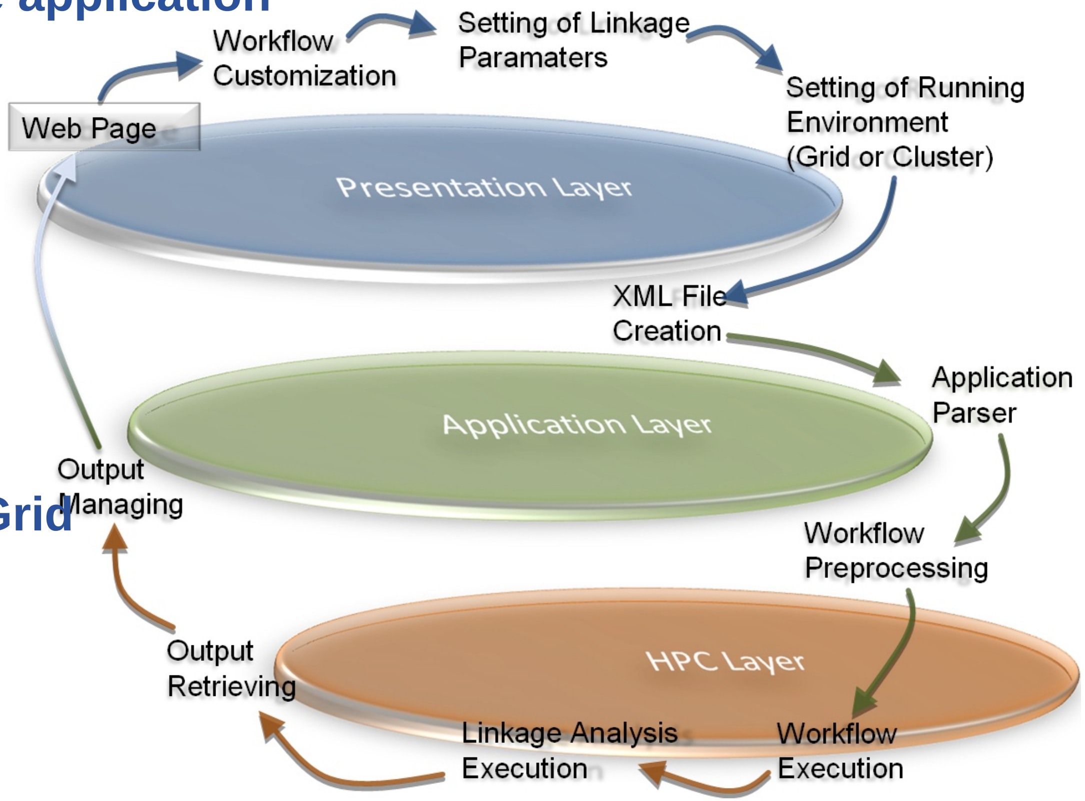
  where users interact with the application

- **the application layer**

  where are stored and run the logics of execution

- **the Grid layer**

  where interactions with the Grid middleware are managed



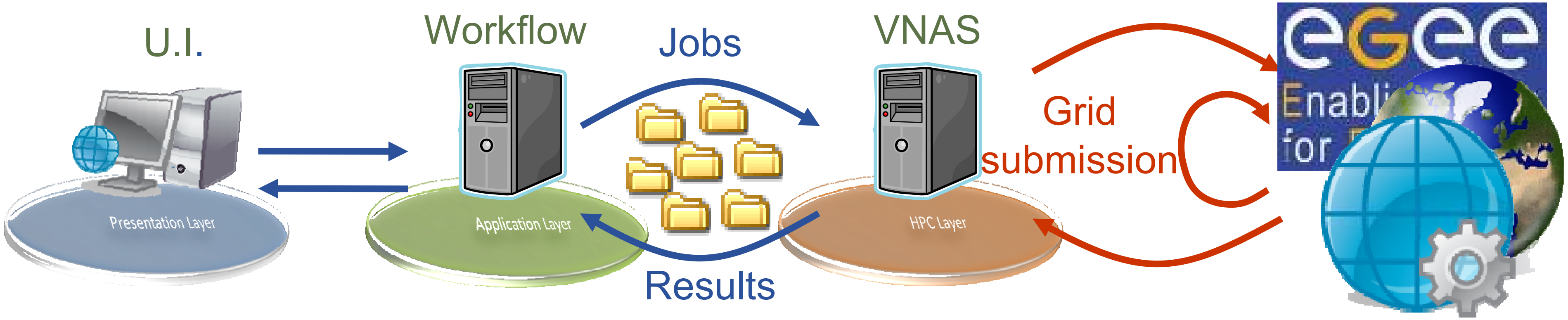
Methods

- **The Grid 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.

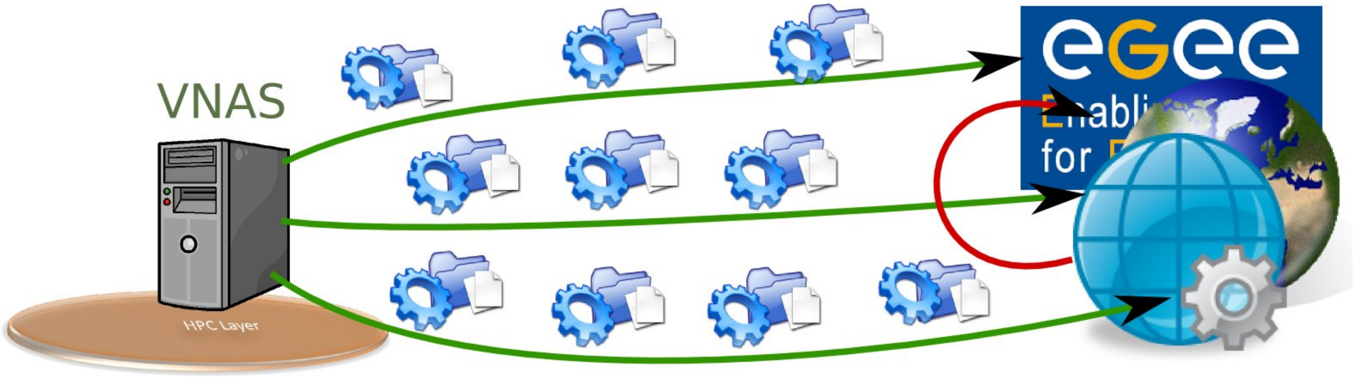


Methods

- **VNAS**

VNAS is an interface providing an abstraction over the Grid middleware

- **Vnas manages job submission over the Grid, monitoring of jobs, fetching of results.**
- **Reduces the complexity in writing Grid applications and pipelines**
- **Reduces Grid overhead, increases throughput**

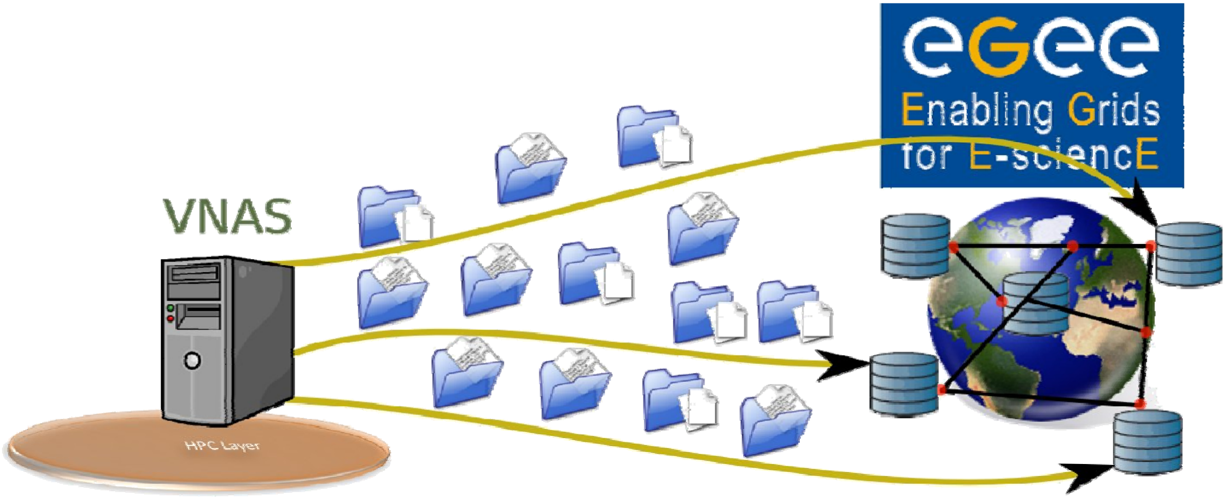


Methods

- **VNAS**

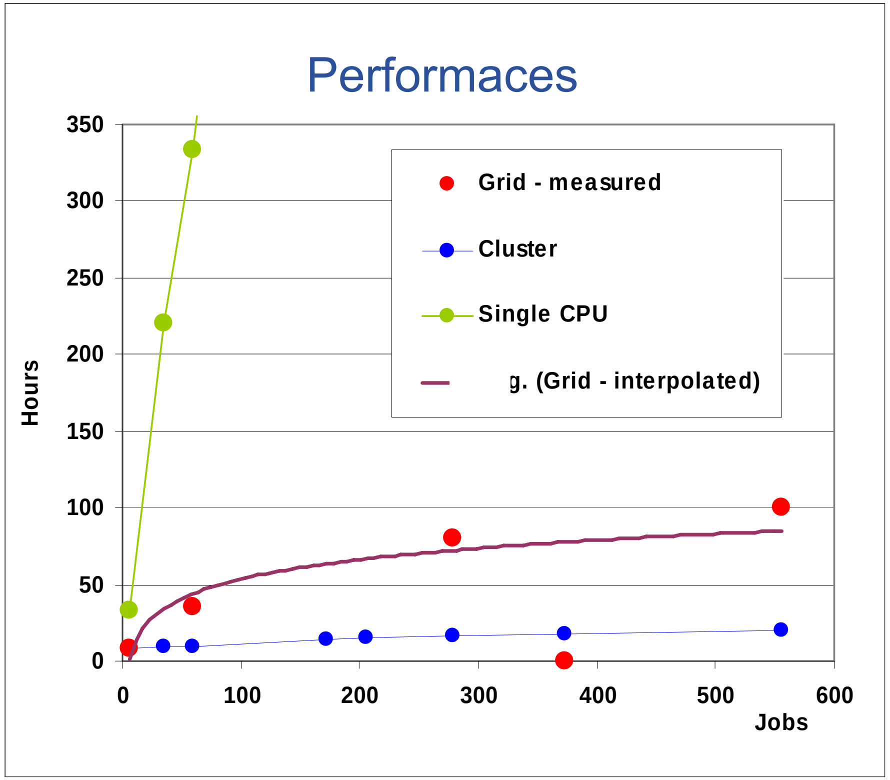
Provides abstraction of the Grid's storage system (Storage Elements)

- **identifies files needed for the job to run on the Grid WNs**
- **identifies files already uploaded to the Grid SEs and re-uses them**
- **reduces Grid bandwidth usage**
- **garbage-collects files from SEs after N days of no-use**
- **prevents user's leftovers on Storage Elements**



Results

Illumina Chip	# Runs [50 SNP]	# Jobs [6 h]	Comput. Cost (time)		
			Single CPU	Cluster (70 nodes) (280 CPUs)	Grid *interp.
10 k	200	6	33 h	8 h	8 h
66 k	1320	35	220 h	9.5 h	*30 h
100 k	2000	60	333 h	10 h	35 h
317 k	6340	172	1056 h	13 h	*72 h
370 k	7400	206	1233 h	15 h	*75 h
500 k	10000	278	1665 h	16 h	80 h
670 k	13400	373	2233 h	18 h	*87 h
1 M	20000	556	3332 h	20 h	100 h



- This approach is mostly useful in high-end challenges, where Grid overheads are less affecting overall execution times compared to single CPU performances. Only very small challenges may show higher efficiency when run in a single CPU workstation.

Results: Case Study 1

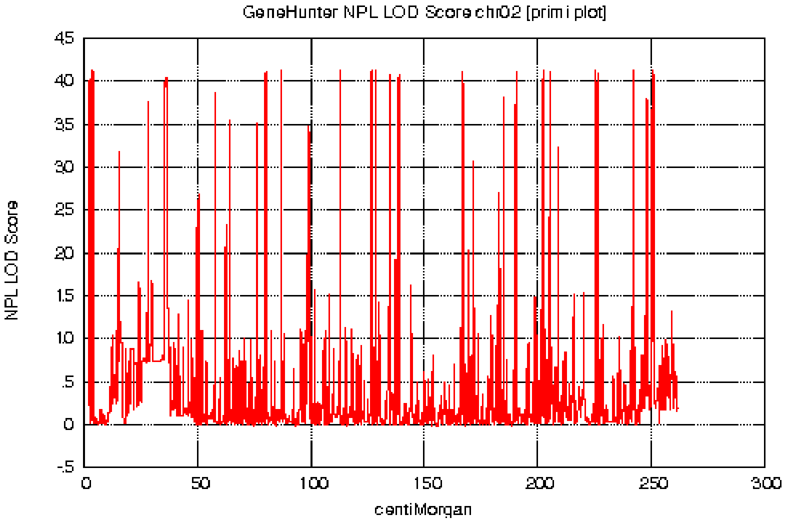
- Chip: Illumina 317.000 SNPs
- Analysis: whole genome (23 chromosomes)
- Software: Genehunter (patients: 32)
- Run time per job: 10 mins
- SNPs per job: 80
- Total jobs:
  - number of runs of Genehunter: 3962
  - number of hours each Computing Element: 6 hours -> 36 runs each Grid Job execution
  - number of Grid Jobs: 111

- Results are presented joined per chromosome

Results: Case Study 1

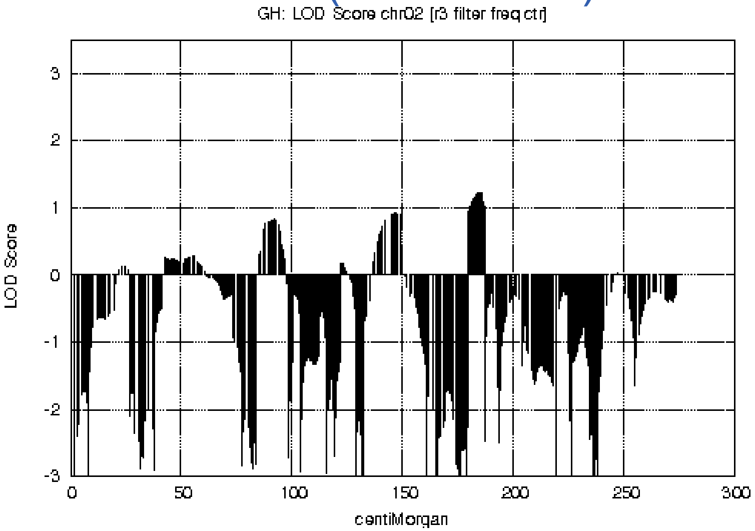
LOD score function for whole Chromosomes 2 and 6 (317k SNPs)

Chrom. 2



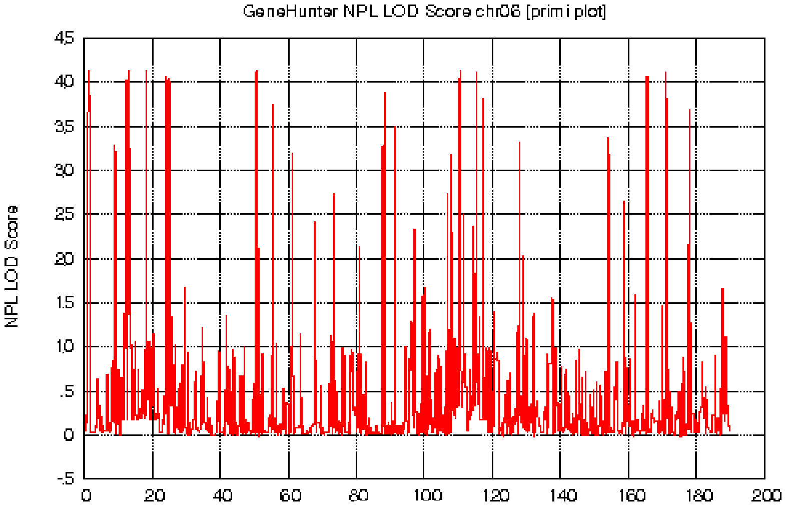
Non Parametric Linkage Analysis

Chrom. 2



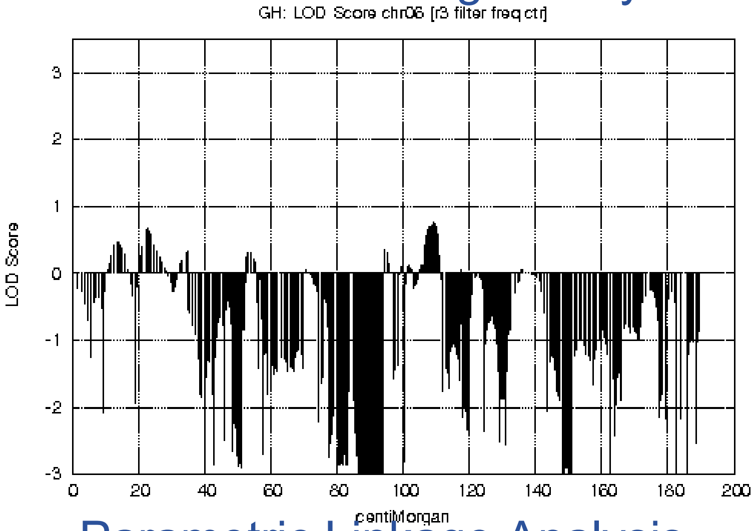
Parametric Linkage Analysis

Chrom. 6



Non Parametric Linkage Analysis

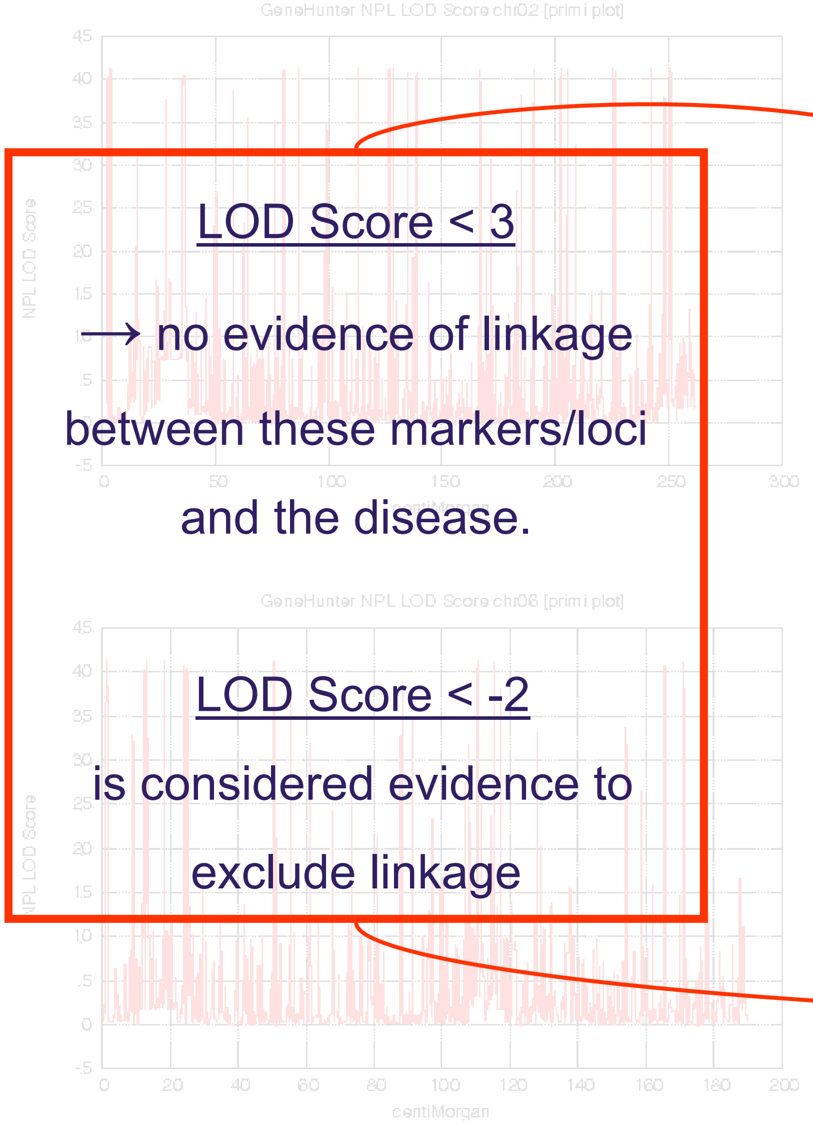
Chrom. 6



Parametric Linkage Analysis

Results: Case Study 1

LOD score function for whole Chromosomes 2 and 6 (317k SNPs)

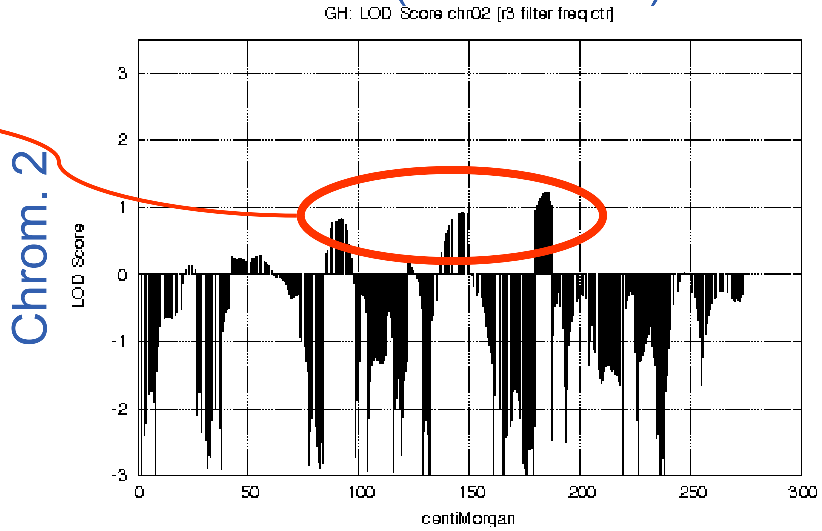


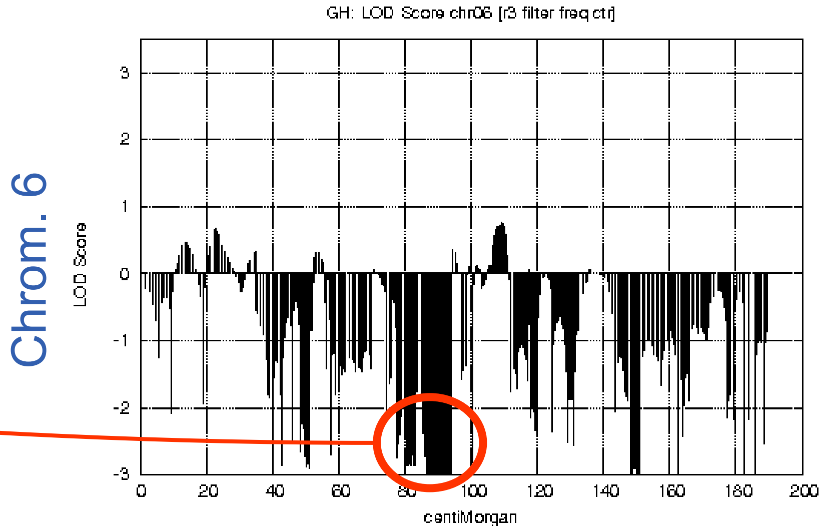
LOD Score < 3

→ no evidence of linkage between these markers/loci and the disease.

LOD Score < -2

is considered evidence to exclude linkage





Results: Case Study 2

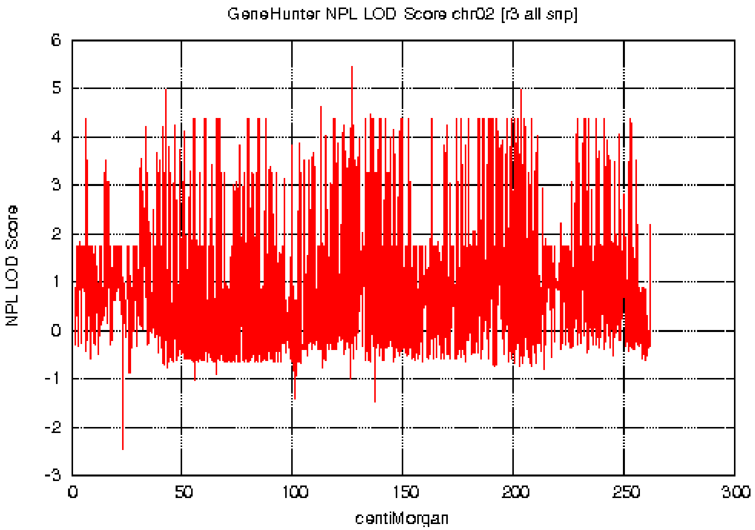
- Chip: Illumina 1 Million SNPs
- Analysis: whole genome (23 chromosomes)
- Software: Genehunter (patients: 24)
- Run time per job: 13 mins
- SNPs per job: 80
- Total jobs:
  - number of runs of Genehunter: 12500
  - number of hours each Computing Element: 6 hours -> 36 runs each Grid Job execution
  - number of Grid Jobs: 348

- Results are presented joined per chromosome

Results: Case Study 2

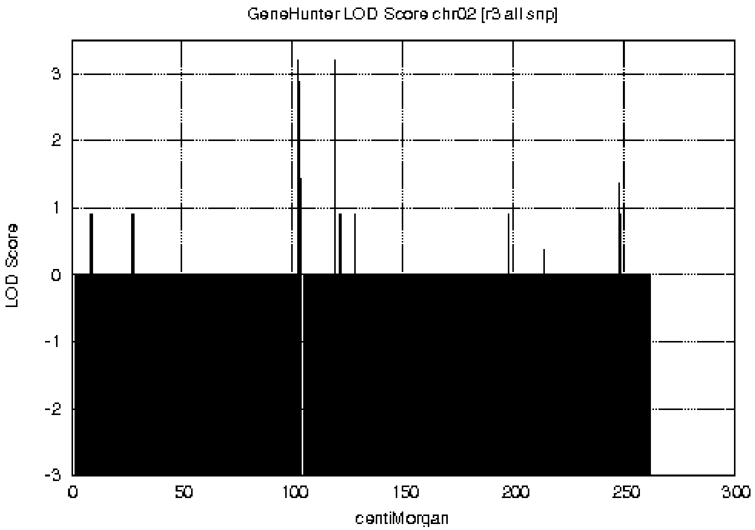
LOD score function for whole Chromosomes 2 and 17 (1M SNPs)

Chrom. 2



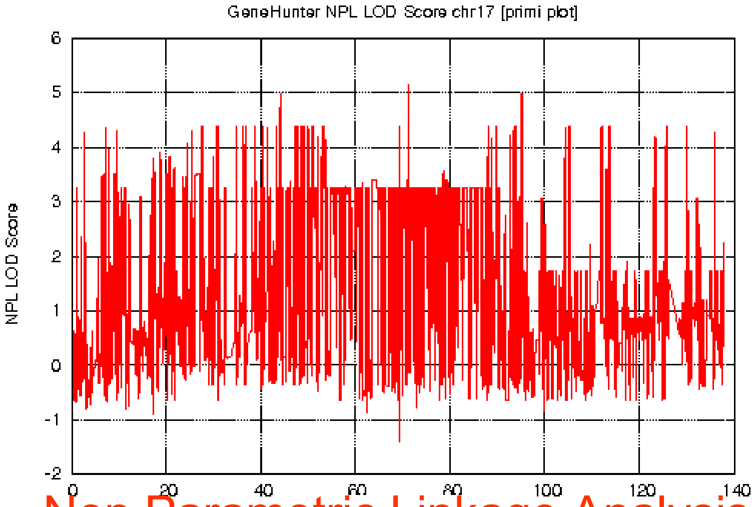
Non Parametric Linkage Analysis

Chrom. 2



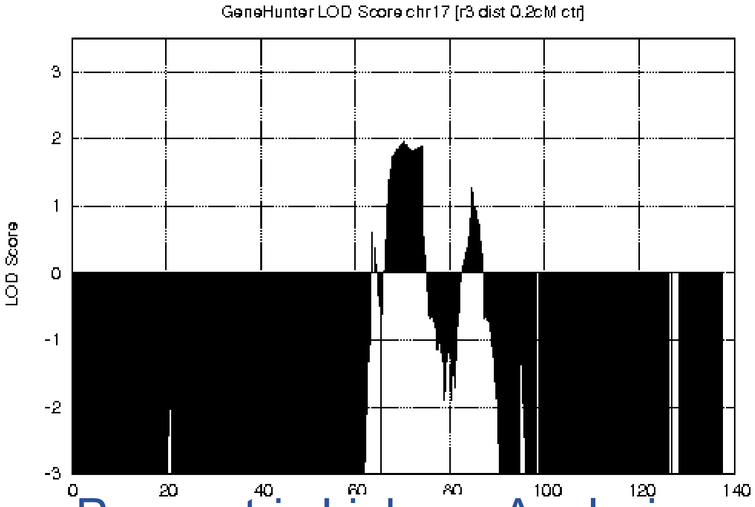
Parametric Linkage Analysis

Chrom. 17



Non Parametric Linkage Analysis

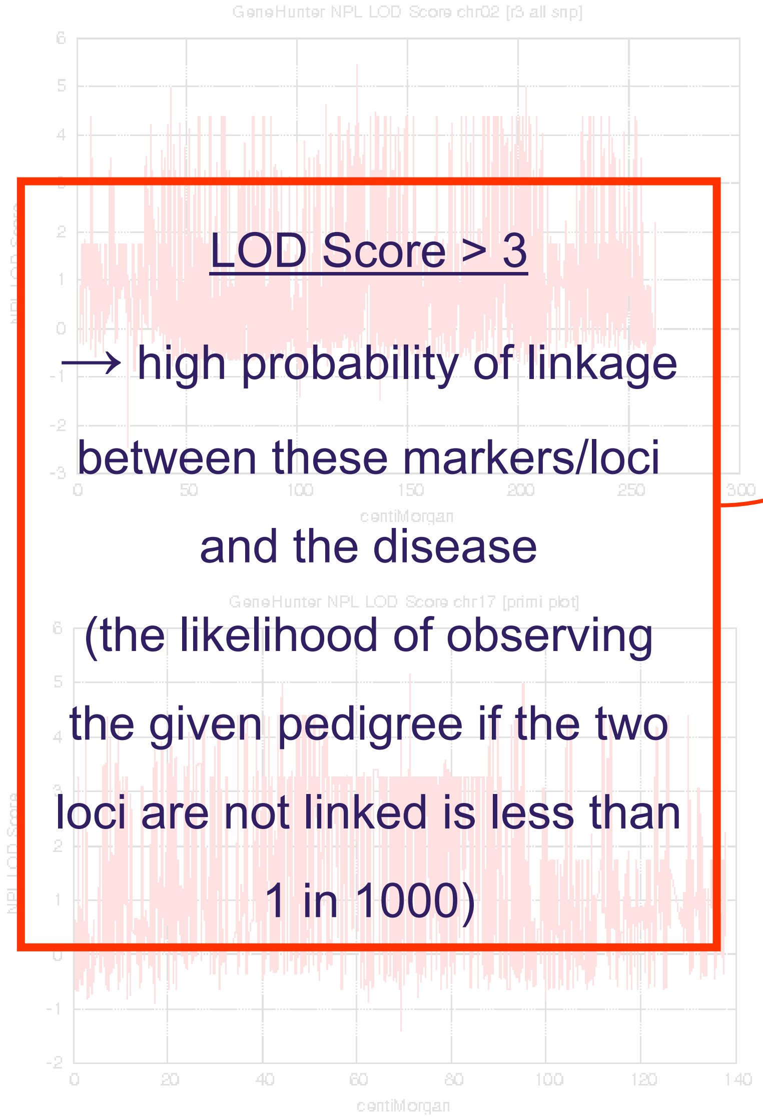
Chrom. 17



Parametric Linkage Analysis

Results: Case Study 2

LOD score function for whole Chromosomes 2 and 17 (1M SNPs)

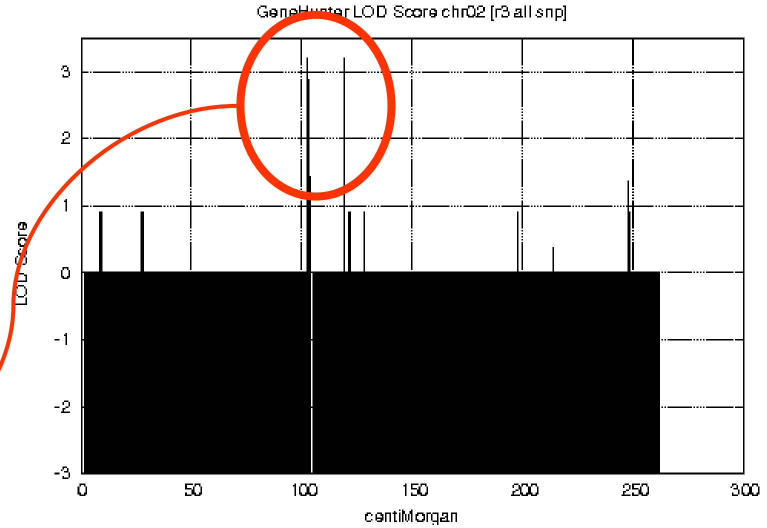


LOD Score > 3

→ high probability of linkage between these markers/loci and the disease

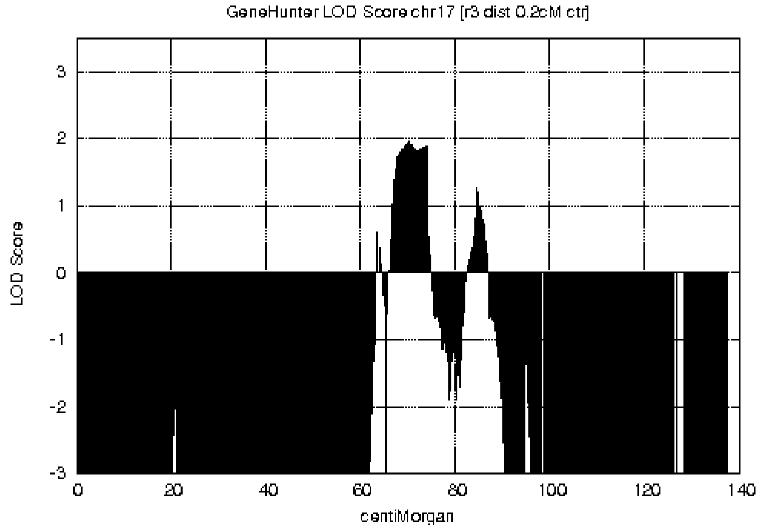
(the likelihood of observing the given pedigree if the two loci are not linked is less than 1 in 1000)

Chrom. 2



Parametric Linkage Analysis

Chrom. 17



Summary

- **This application enables the user to launch genetic linkage analysis calculations for medium to large challenges over a distributed computational infrastructure like the EGEE Grid. It offers:**
  - a parallel processing of the pipeline tasks;
  - A user interface that provides an easier approach to linkage analysis software;
  - A reliable software layer that manages low-level interactions with the distributed computing elements.

Acknowledgments

- **This work has been supported by the EGEE III, CNR-BIOINFORMATICS and by the LITBIO, ITALBIONET FIRB-MIUR projects.**