Genome wide association studies of human complex diseases with EGEE

Alexandru Munteanu

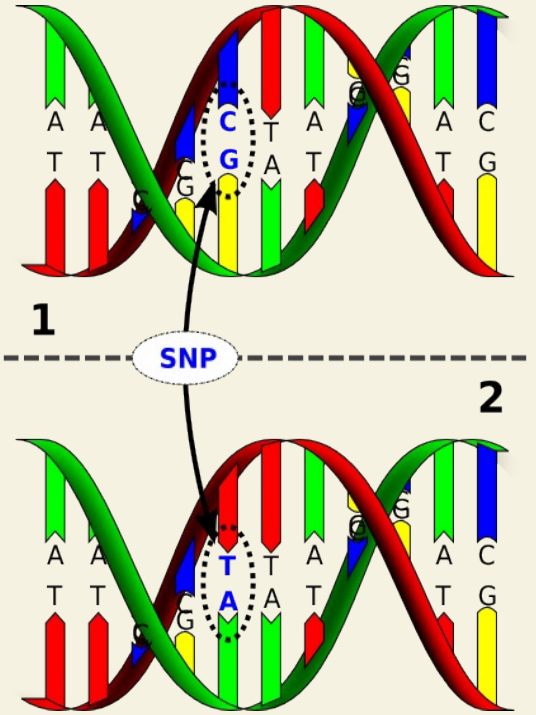
Inserm
Institut national
de la santé et de la recherche médicale

U525
Institut National de la Santé Et de la Recherche Médicale
Paris, France
munteanu@chups.jussieu.fr

Association studies

- Single Nucleotide Polymorphism (SNP)
  - a single genetic variation at a specific location on the genome

- Find genetic factors responsible for a disease :
  - compare frequency of SNPs between cases and controls
    - most commonly : each SNP is analysed at a time



source: wikipedia

Haplotype associations

~Haplotype :

- combination of SNPs on a chromosome
- more powerful approach than looking at each SNP separately

~Why ?

- several SNPs taken together could be responsible for the disease
- haplotypes better characterize the variable structure of the genome

THESIAS program

Testing Haplotype Effects In Association Studies

~Difficulty

- haplotypes must be inferred from the genotype
  - ~for individuals who are double heterozygotes (AC|GT) haplotypes cannot be deduced
    - example with the combination of SNPs A/C and G/T

	GG	GT	TT
AA	(AG,AG)	(AG,AT)	(AT,AT)
AC	(AG,CG)	(AG,CT) ***or*** (AT,CG)	(AT,CT)
CC	(CG,CG)	(CG,CT)	(CT,CT)

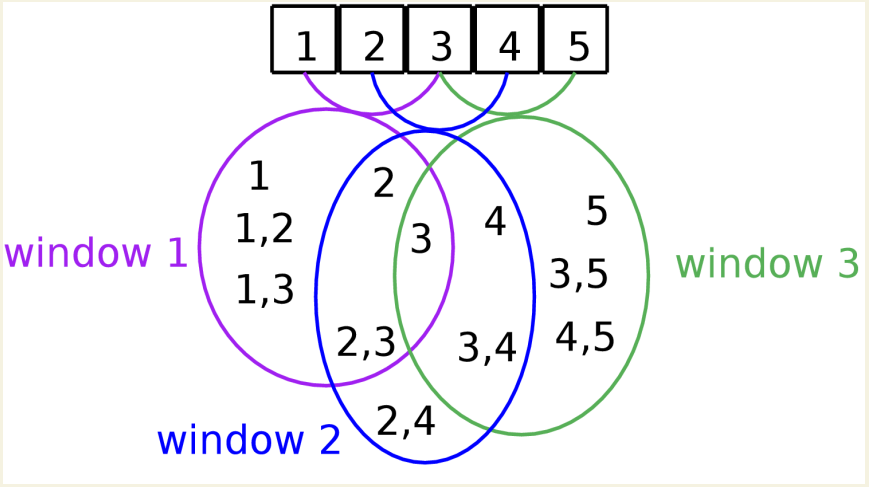
    - with more than 2 SNPs, the difficulty increases

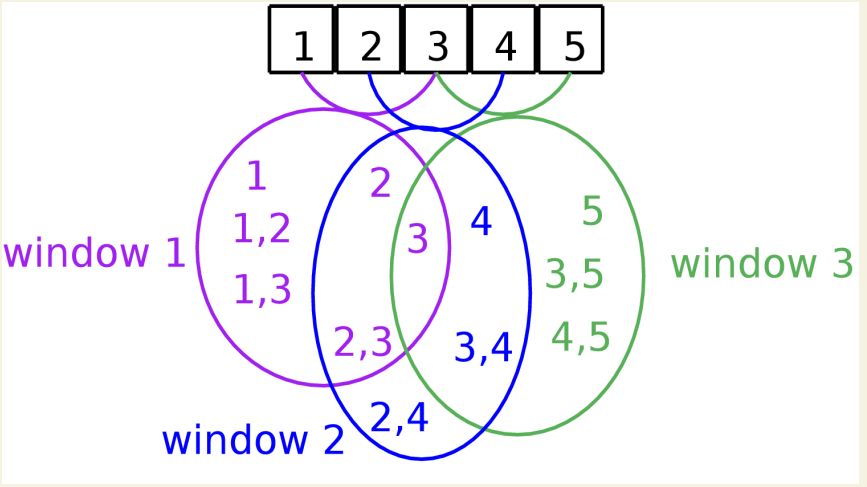
Complexity

- DNA chips
  - allow to genotype hundreds of thousands of SNPs across the genome
- Example with 8500 SNPs :
  - SNP only study
    - 8500 computations
  - haplotypes study
    - $C_{8500}^{1}+C_{8500}^{2}+\ldots+C_{8500}^{8499}+C_{8500}^{8500}=2^{8500}-1$ computations

Reducing complexity in THESIAS

- The sliding-window approach
  - haplotypes composed of SNPs located close to each other are expected to be more biologically meaningful





- The chosen combinations approach
  - keep only relevant combinations, by eliminating SNPs providing the same information

THESIAS job definition

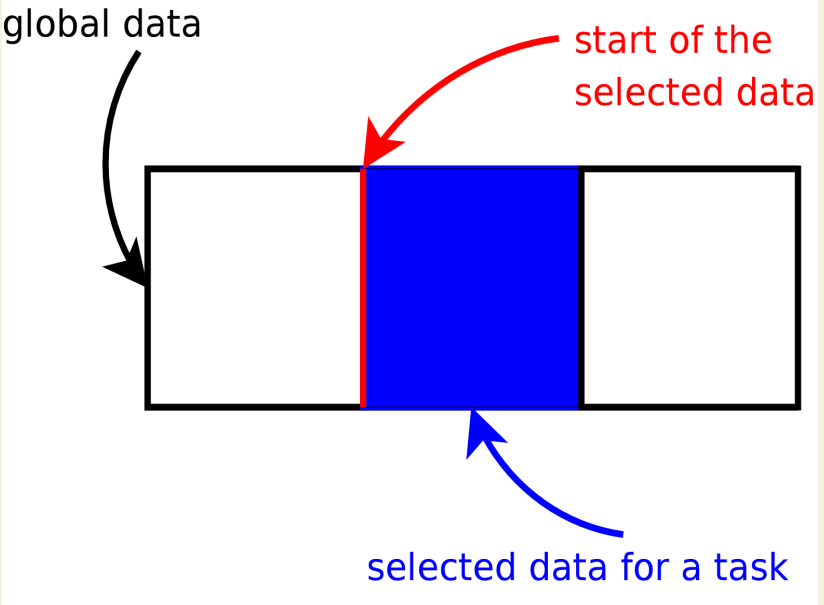
- A process on the grid
  - computes one block of SNPs combinations
  - defined by two arguments of THESIAS
    - start_id , end_id

1 job = one block of combinations

ID	Combination
1	(1,8)
2	(2,8)
3	(3,8)
4	(4,8)
5	(5,8)

THESIAS task definition

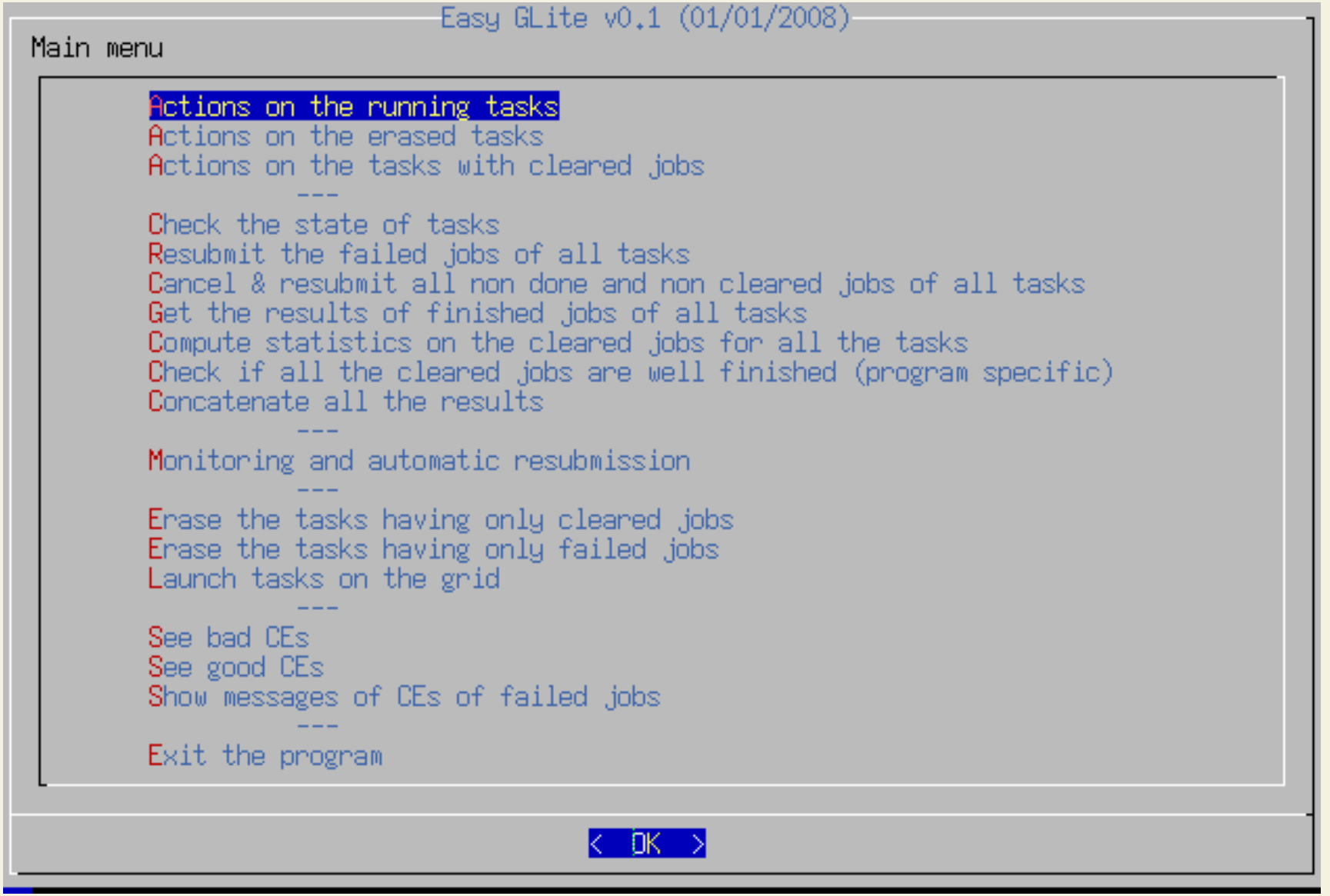
- Original set of SNPs splitted into regions
- Only the region containing the data needed for specific computations is sent on the grid
- Task definition
  - a set of blocks of SNPs combinations (a set of jobs) in a given region to be investigated



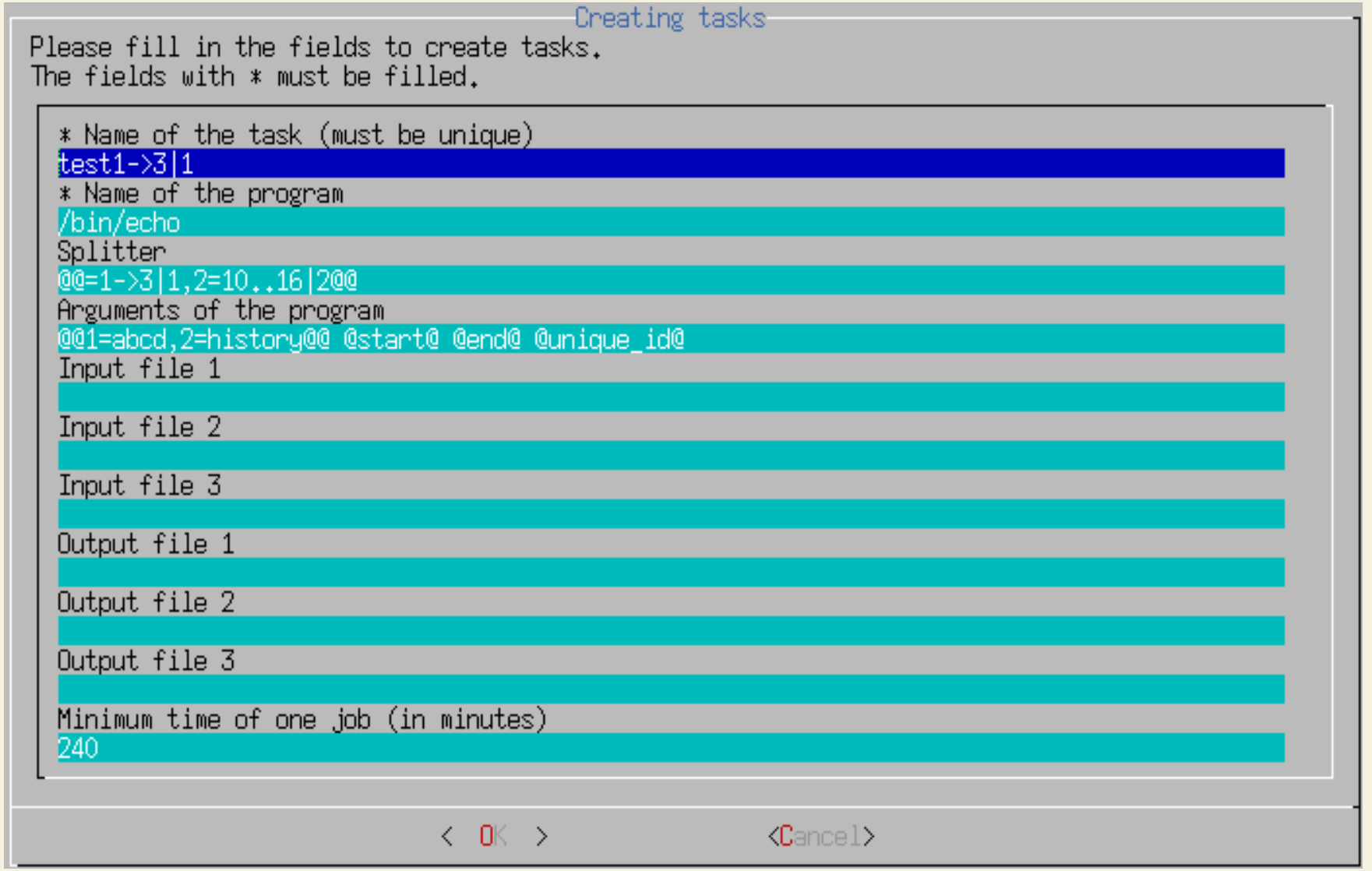
Easy-gLite

- General purpose and easy to use interface on top of the gLite UI (User Interface)
  - created to simplify batch jobs submissions
- Features
  - creating sets of tasks by filling only one form
  - manage jobs, tasks and all the tasks
  - automatic resubmission (“timeout” or failed jobs)
  - output retrieval of all (or some) completed jobs

Easy-gLite main menu



Easy-gLite submission form



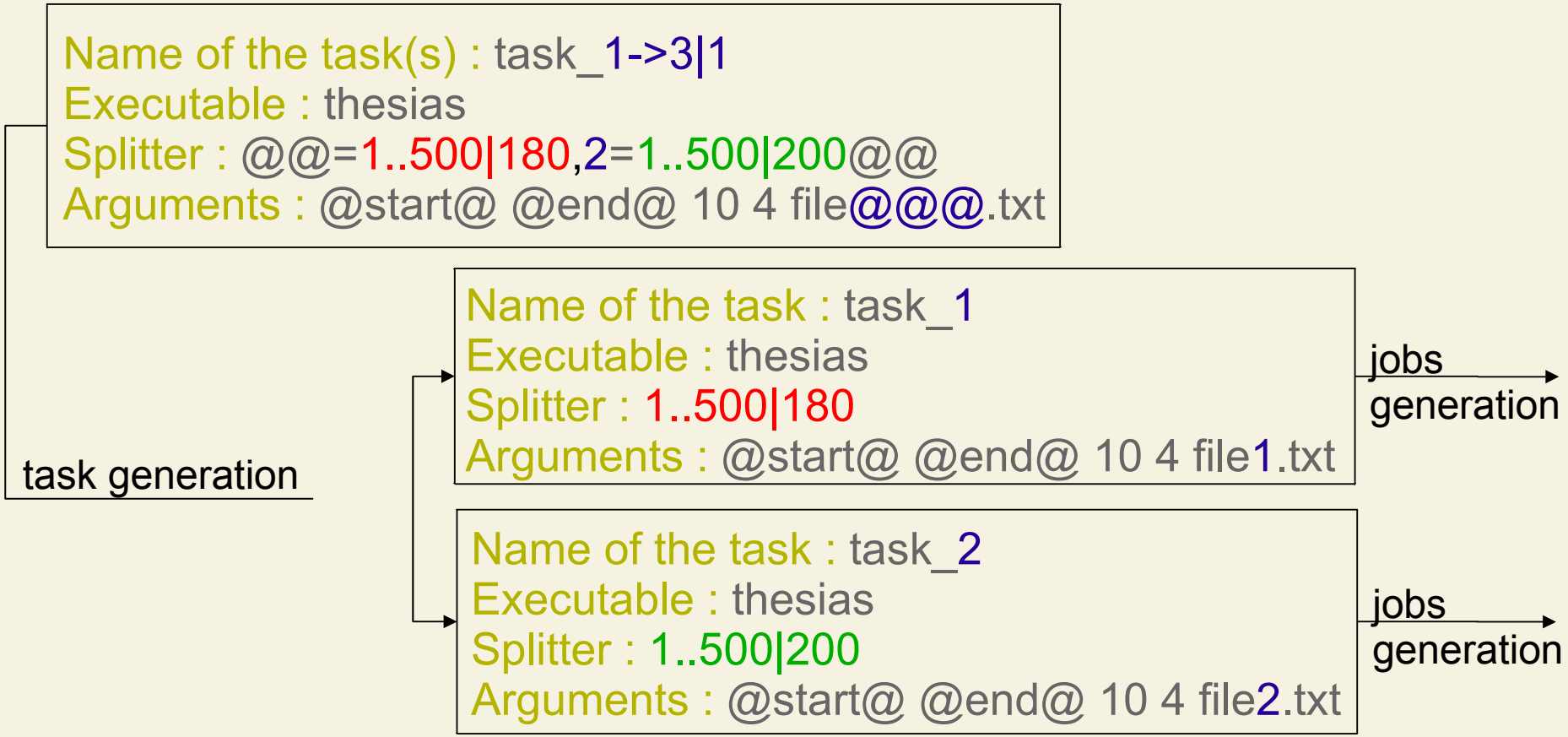
Easy-gLite submission form syntax

- Example of task creation with generation of arguments
  - the syntax “x..y|z” in the Splitter field

    - 3 jobs are generated with different parameters
- The syntax can eventually be improved by changing a small part of the code

Easy-gLite submission form syntax

- Example of a set of tasks creation
  - the “x->y|z” syntax in the “Name of the task” field



Other Easy-gLite features

- Logs of the jobs can be viewed or saved
- Reasons of failed jobs are stored
  - to analyse failures on different sites
- Consistency verification of the output files
  - with submission of jobs having bad output files
- Maintain a blacklist for Computing Elements (CE) where many jobs failed
  - in order to avoid unreliable CE for our jobs

Remarks

- Documentation
  - finished for both THESIAS and easy-gLite
- Release
  - the programs will soon be released

Results

- Proof of concept
  - over 1 million SNPs combinations of 8456 SNPs from a chromosome were analysed
  - several regions where haplotypes are associated with the disease have been identified
- EGEE performance
  - 3 days on EGEE while 2 years and a half on a single computer

Work in progress

- Genome wide analysis
  - ~5000 cases and controls with ~500000 SNPs

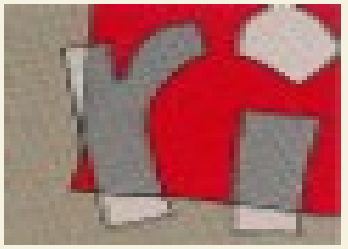
Credits

- François Cambien
- David Trégouët

Inserm U525

Institut national
de la santé et de la recherche médicale

LABORATOIRE
DE L'ACCÉLÉRATEUR
LINÉAIRE



Thank you

Questions ?