

Genome wide association studies of human complex diseases with EGEE

Inserm

Institut national
de la santé et de la recherche médicale

Alexandru Munteanu

U525

Institut National de la Santé Et de la Recherche Médicale
Paris, France

munteanu@chups.jussieu.fr

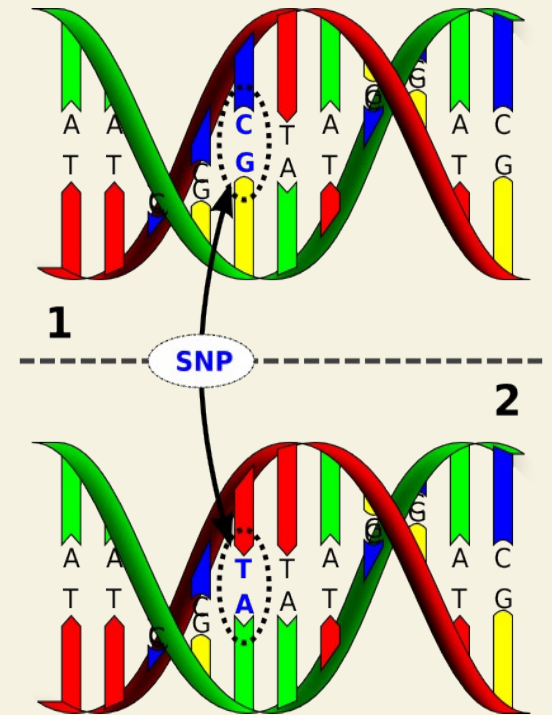
Association studies

~ 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) <i>or</i> (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 :

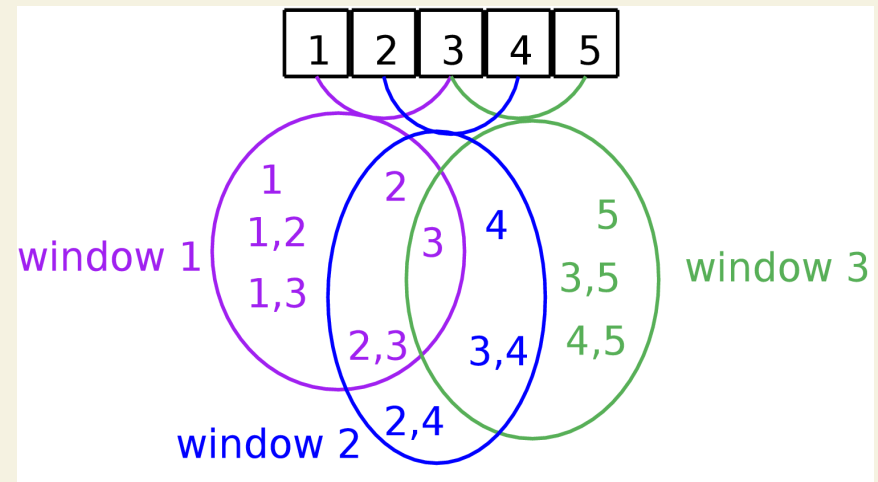
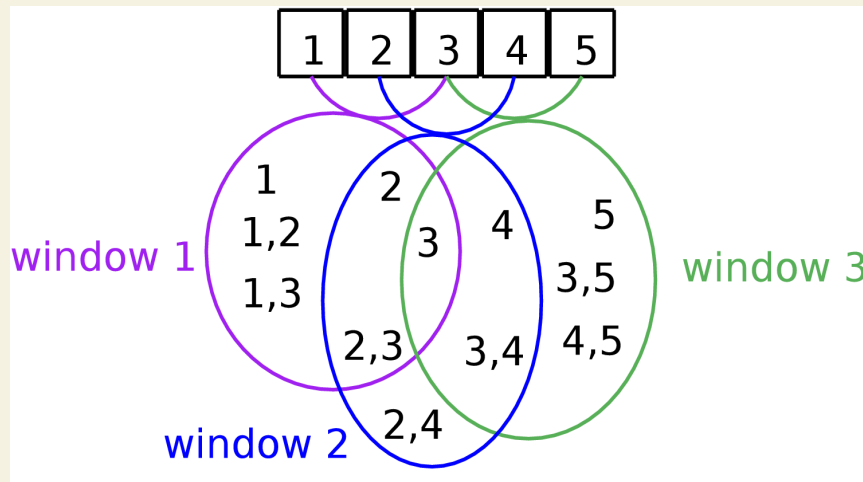
- SNP only study
 - * 8500 computations
- haplotypes study

$$\sim C_{8500}^1 + C_{8500}^2 + \dots + C_{8500}^{8499} + C_{8500}^{8500} = 2^{8500} - 1 \quad \text{computations}$$

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

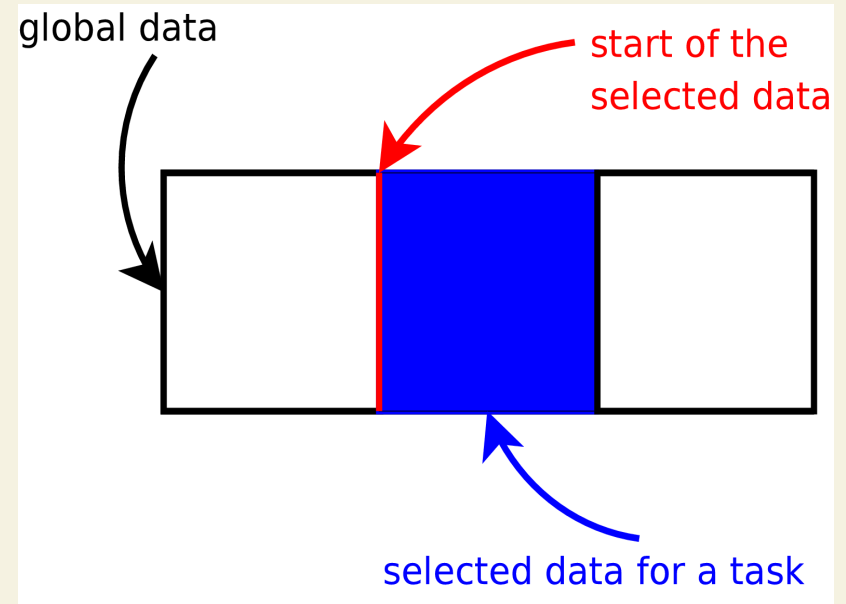
- computes one block of SNPs combinations
- defined by two arguments of THESIAS
 - ★ start_id , end_id

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

1 job
=
one block of
combinations

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 v0.1 (01/01/2008)
Main menu

Actions on the running tasks
Actions on the erased tasks
Actions on the tasks with cleared jobs
  ---
Check the state of tasks
Resubmit the failed jobs of all tasks
Cancel & resubmit all non done and non cleared jobs of all tasks
Get the results of finished jobs of all tasks
Compute statistics on the cleared jobs for all the tasks
Check if all the cleared jobs are well finished (program specific)
Concatenate all the results
  ---
Monitoring and automatic resubmission
  ---
Erase the tasks having only cleared jobs
Erase the tasks having only failed jobs
Launch tasks on the grid
  ---
See bad CEs
See good CEs
Show messages of CEs of failed jobs
  ---
Exit the program

< OK >
```

Easy-gLite submission form

Creating tasks

Please fill in the fields to create tasks.
The fields with * must be filled.

* Name of the task (must be unique)	test1->3 1
* Name of the program	/bin/echo
Splitter	
Arguments of the program	@@=1->3 1,2=10..16 2@@
Input file 1	@@1=abcd,2=history@@ @start@ @end@ @unique_id@
Input file 2	
Input file 3	
Output file 1	
Output file 2	
Output file 3	
Minimum time of one job (in minutes)	240

< OK > <Cancel>

Easy-gLite submission form syntax

~ Example of task creation with generation of arguments

- the syntax “x..y|z” in the Splitter field

Splitter : 3..9|2

Arguments : Hello @start@ **@end@**

jobs
generation →

Hello 3 **5**

Hello 6 **8**

Hello 9 **9**

★ 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

Name of the task(s) : task_1->3|1

Executable : thesias

Splitter : @@=1..500|180,2=1..500|200@@

Arguments : @start@ @end@ 10 4 file@@@.txt

task generation

Name of the task : task_1

Executable : thesias

Splitter : 1..500|180

Arguments : @start@ @end@ 10 4 file1.txt

jobs
generation

Name of the task : task_2

Executable : thesias

Splitter : 1..500|200

Arguments : @start@ @end@ 10 4 file2.txt

jobs
generation

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

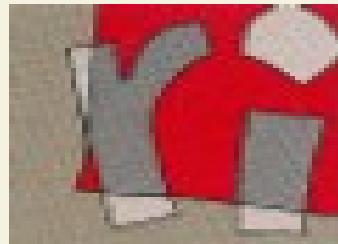
Work in progress

~Genome wide analysis

- ~5000 cases and controls with ~500000 SNPs

Credits

- ~ François Cambien
- ~ David Trégouët



Thank you

Questions ?