



Network approaches to Genome-Wide Association studies

Daniel Remondini
Dip. Fisica, INFN & CIG (UniBO)
daniel.remondini@unibo.it



GWA & complex traits

What is a complex trait (disease)?

What is a GWA study?



Complex trait

A *phenotypic* characteristic (trait) determined by the *interaction* of many different genes (and allelic variants) or more gene-environment interactions. A complex trait follows *non-mendelian* models of inheritance.

Examples: eye color, obesity, asthma, *longevity* (?)



GWA study

A *genotyping* of allelic variants along the entire genome in a family-based or case-control study with an association design:

- 1) for population/evolutionary genetics studies
- 2) for complex trait disease determination



Polymorphism – SNP

SINGLE NUCLEOTIDE POL. (SNP)

CTGACAAACCTCCAGGGAGGTCCCAAGGATGAG
CAGCAGATCTGCCATGGGGAAATCAACCACATG
CAGCAAGAACC**T/A**CTGGGGCAGCGGCTCCCA
AAGAACAATG**T/C**GGGCTTCACAACGCCGGTG
CAGACCGGGCAGCGGGGAACCCTGTCTGCCATC
ACGTCAGCCTGGAAACAGAG**G/C**AAAACAGGCCT
GAGGAAGATGCCTGCAACACCCTGTAAAAGGAA
AAGGCAACAAGTTCTTATTAAAGTCTCAAACATTC
TCCCCAAAAAAGGTGAGAGAGGGATCATGAAC
CCCCATGTACCCAGCT



Haplotypes: allele combinations

Haplotypes			
Haplotype 1	C	T	C
Haplotype 2	T	T	A
Haplotype 3	C	G	T
Haplotype 4	T	C	A

Tag SNPs

A / G	T / C	C / G
---------------------------	---------------------------	---------------------------

SNPs are in general not independent



Measures: Disequilibrium D

A measure of *association* between SNPs appearance.

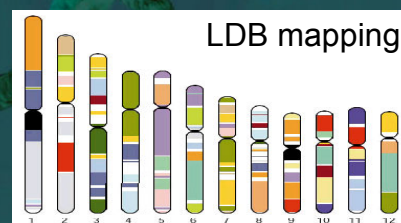
Example:

$$D_{(12)} = P(S_1, S_2) - P(S_1) \cdot P(S_2)$$

Linkage Blocks – Association

LDB (Linkage Disequilibrium Block) = group of neighbouring loci on a chromosome with null or scarce evidence of recombination

$$D' = \frac{D}{Norm} \approx 1$$



Average block size: 2-20Kb

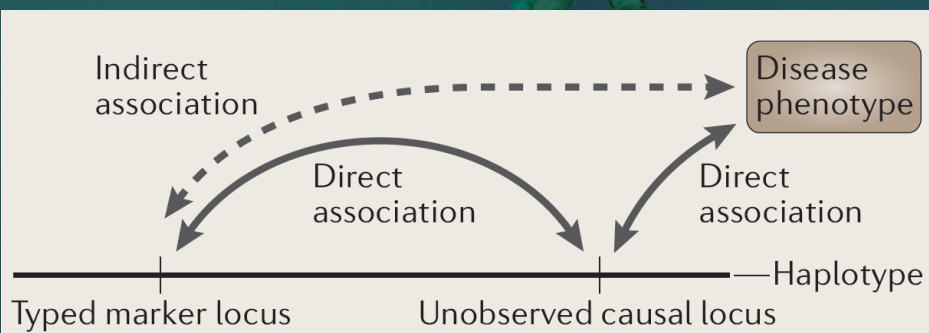


Tagging

choosing an appropriate set of marker SNPs (possibly at least one for each LDB) able to span the whole genome area of study

Association studies

direct association between marker SNPs and disease leads to *indirect* association between (unobserved) causal locus & disease



Some numbers

Human genome:
about 20000 genes
~3Gb (Gigabases, i.e. potential SNPs!)

Large-scale high-frequency SNP scanning:
~20 SNP/gene:
= 400000-1000000 SNP/sample



GWA studies: challenges

Very high dimensionality:
 $(10^5-10^6)^2$ LD arrays!

Computationally demanding

Low Signal/Noise ratio



Problems for GWA

Complex traits can be characterized by **several SNPs weakly** associated with disease

Single-SNP statistics may **not** work

BUT

Looking for all possible combinations is computationally **unfeasible**



Improving analysis

Consider SNP association structure to various degrees:

1 - external superimposed structures

2 - internal dataset structure



SNP association structure 1

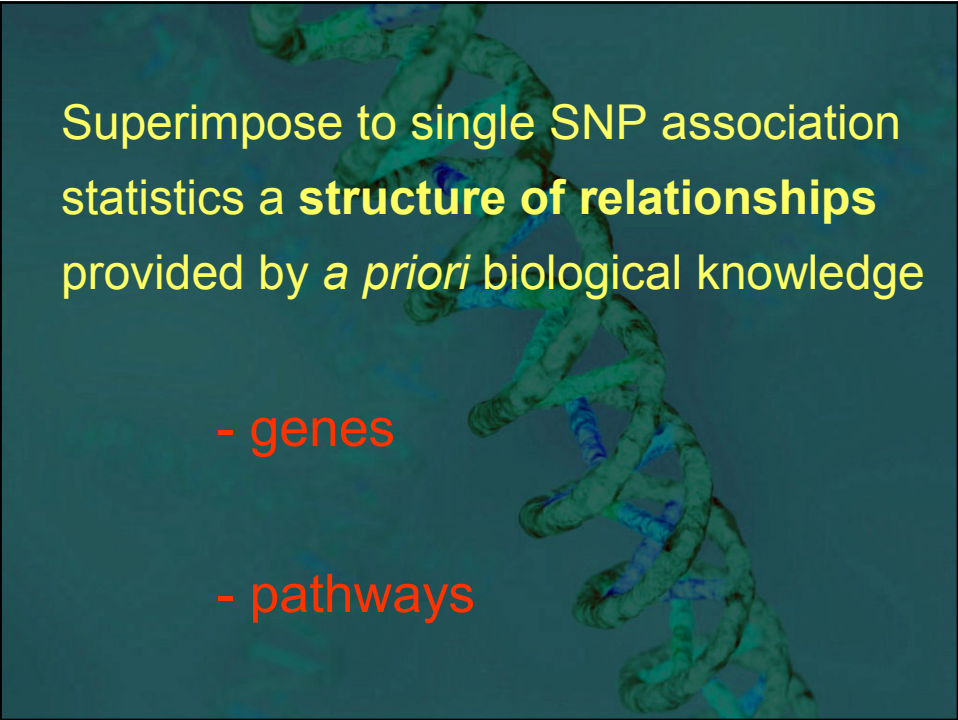
A priori biological knowledge:

- genes (mostly *multiples* of LDB)
- biological pathways (gene-gene functional interaction)



Strategy #1:

**Embedding additional
biological knowledge**



Superimpose to single SNP association statistics a **structure of relationships** provided by *a priori* biological knowledge

- genes

- pathways

Example: gene expression

In a case/control design experiment, which genes/functions are involved in disease?

Main differences with SNPs studies:

- 40000 genes “only”
- no metrics, only topological structure
- clear biological role of each node

Statistical significance



Statistical significance is **not** necessarily related to biological significance because:

- complex (multi SNP) associations often show slight variations in expression (low LD)
-> low statistical significance

Introducing biological knowledge



Significance *must* also rely on the **known role** that the gene has in cellular mechanisms

Single element statistical analysis *must* be integrated with a priori higher level biological knowledge

-> known biological pathways (e.g. KEGG)

-> genetic association background structure

Improvements

Robustness of statistical analysis is increased (multiple genes/SNP counting for each pathway)

Single-element significance conditions can be **weakened**, followed by **higher-level** significance filtering

A priori knowledge can be exploited for **further analysis/comparison**

Kegg based significance analysis

Reconstructing networks of pathways via significance analysis of their intersections.

Mirko Francesconi^{2,7}, D. Remondini^{2,7}, N. Neretti^{1,2}, J. M. Sedivy⁴, L N Cooper^{1,5}, E. Verondini^{2,3}, L. Milanesi⁶, G. C. Castellani^{1,2,7,*}

¹Institute for Brain and Neural Systems, Brown University, Providence RI USA

²Centro Interdipartimentale "L. Galvani", Università di Bologna, Bologna, IT

³Department of Physics, Università di Bologna, Bologna, IT

⁴Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence RI USA

⁵Department of Physics, Brown University, Providence RI USA

⁶Istituto di Tecnologie Biomediche (ITB) CNR, Milano, IT

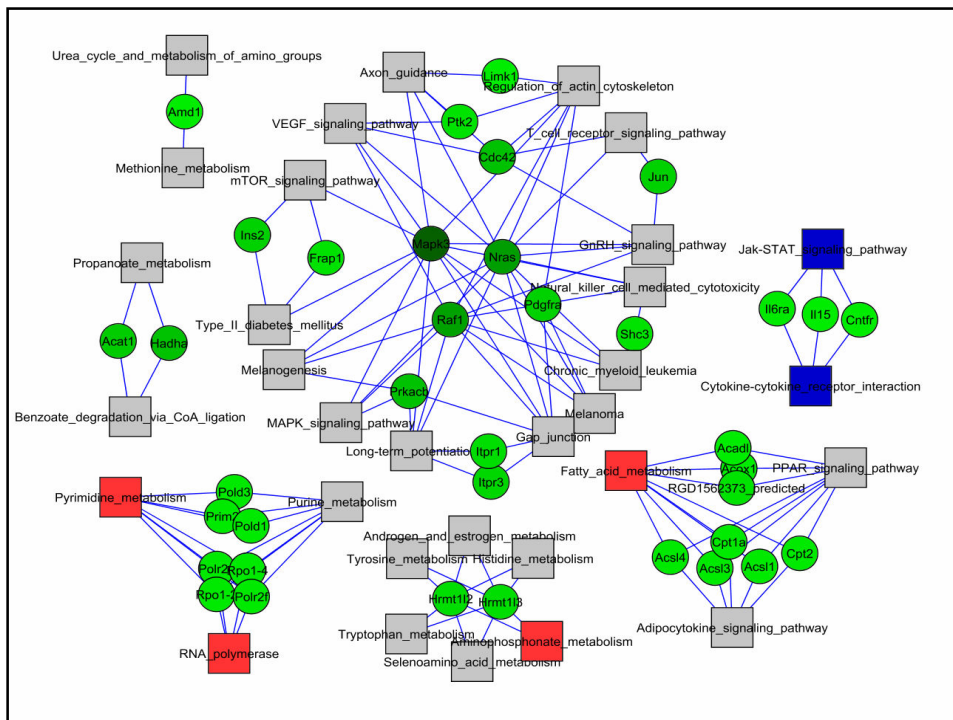
⁷DIMORFIPA, Università di Bologna, Bologna, IT

BMC Bioinformatics 2008

Pathway network analysis

Given statistically significant nodes and edges, the **significant pathway network** can be reconstructed.

Edges and nodes can be **ranked** based on their **centrality** in the network (e.g. connectivity degree or **betweenness**)



SNP association structure 2

Consider LD for control (and case) groups

High linkage, small distances:

Linkage Disequilibrium Blocks (LDB)

Smaller linkage, unknown (large) distance:

Higher structures (*chromatin?*)

Complex traits

...and mostly **Noise**

Network approach

LD matrix = Weighted undirected network
with mixed metrics/topology informations

NODES: SNPs

LINKS: LD values



Network approach

Network structure:

Closest neighbours: LDBs

Far neighbours (in ascending order):
genes -> regions -> chromatin (?) ->
complex traits



SNP-based data

Background network: hereditary genetic associations (LDB & more) from **control dataset**

Case network: LD matrix performed over **case dataset**

Node attributes: P-value of associations with disease (case vs. control χ^2)



Rationale

Complex traits should be embedded in **network relationships**, differing from LDBs, other inherited structures and noise (e.g. E-R network structure)



Strategy #2:
exploiting intrinsic
network
topological/metric
structure

GWA studies: working hypothesis

How can a complex profile composed of multiple (weak) associations be revealed?

Relationships between associations should be reflected into the **network structure**

GWA studies: working hypothesis

- 1) characterization of LD network structure (LDBs and other modules/cliques/communities)
- 2) SNPs classification by a network parameter *signature*
- 3) background vs. case network

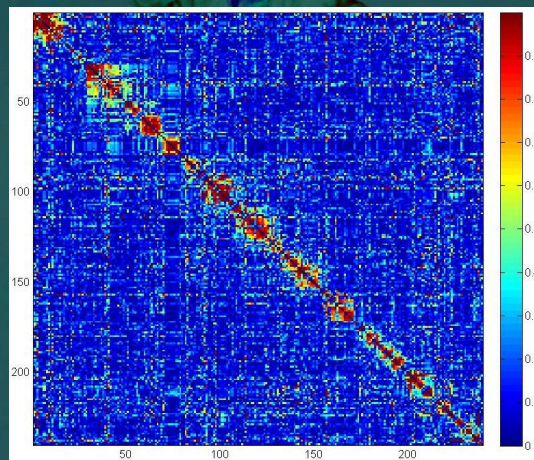
Case study dataset: GEHA

EU Project: GENetics of Healthy Ageing
AIM: any complex traits related to ageing?

DATASET: High-density region-focused
(~240 sites) mapping on specific
chromosome areas (not Genome-Wide)
- 307 centenarian vs. 352 control PBMCs

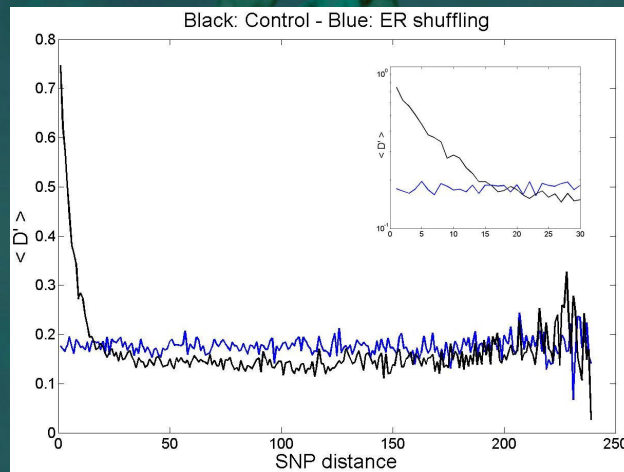


LD background network



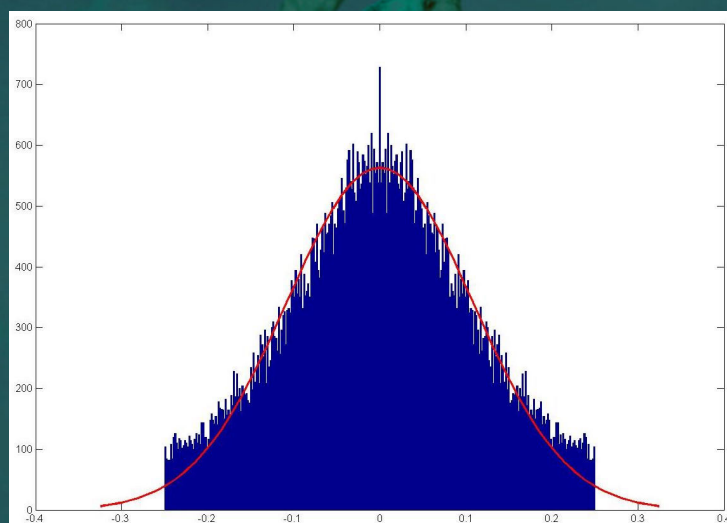
Undirected, weighted [0:1], fully connected network

D' vs node distance



Significant (Bonferroni) difference between Control and ER up to 12 SNPs distance (LDB area?)

D' histogram & noise



Low values follow gaussian (noise) distribution



D' values

High value & small distance: LDB

Low value ($<0.2-0.3$): noise

Intermediate values:
noise + long-range structure



D' network 1

Set Low values (<0.25) = 0

Topological network: main node parameters

Connectivity Degree K

Betweenness Centrality BC

Network parameters

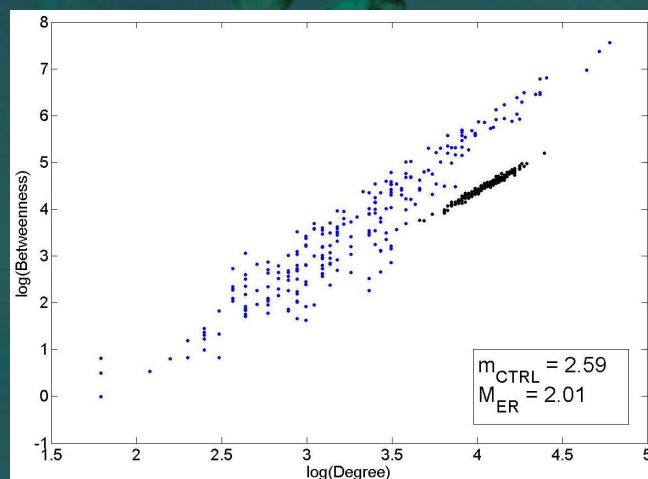
Nodes = N ; Links = M :

“Centralities”: $O(N^3)$

Random Walk Betweenness $O(N \cdot M^2)$

Cliques Search: exponential

D' network structure



For ER random networks: $BC = K^2$ ($m=2$)
-> Background network is NOT a ER network



D' network 2

- find LDBs as **communities**
- **remove** LDB structure



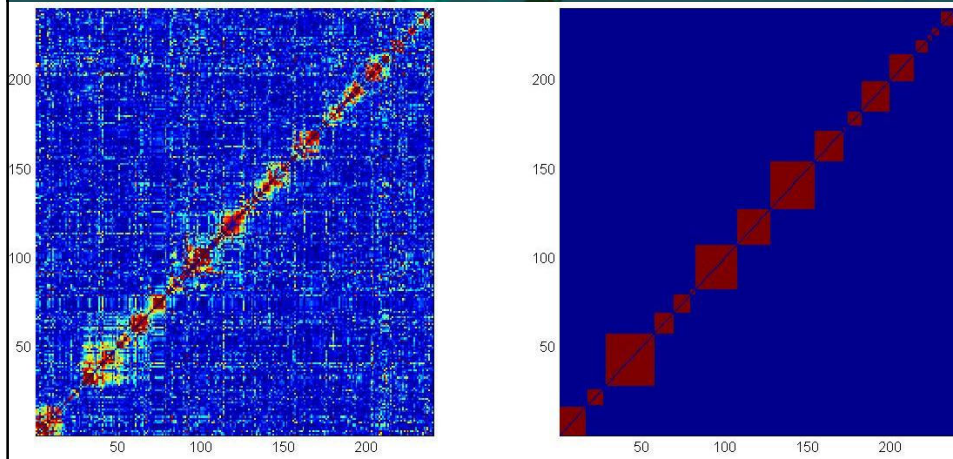
D' network 2

A novel algorithm have been written for LDB finding starting from full LD adjacency matrix

It can extract:

- LD blocks
- genes (?)
- ...? (to be tested)

D' network 2



NOTE: LDB definition is not unique...

Finding LDB

- compare algorithm with common LDB definitions [e.g. *Gabriel Science 02*]
- tune algorithm parameters to find LDB & higher structures (e.g. genes, chromatin/related regions)

Association

- embed significant SNPs into this network structure for topological characterization
- associate significant SNPs between them (emergence) and search for other significant (not Pval) ones

Research Team

CIG Bologna
ITB CNR Milano

Prof. Gastone Castellani
Dott. Mirko Francesconi
Dott. Francesco Lescai
Prof. Luciano Milanesi
Prof. Claudio Franceschi