Programming for genomics

Pierre Neuvial

Fall 2015

Highlights of the course

How to use R for **genomic data analysis**

- Data analysis and visualisation
- Basic programming
- Implementing algorithms

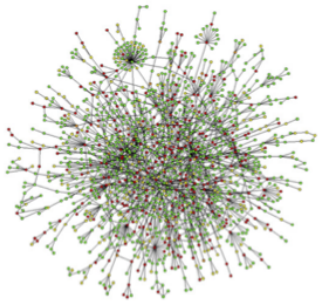
Running example: **segmenting genomic signals**

Other topics covered

- Interfacing R code with C/C++
- Comparison to other languages: Python, Julia
- Reproducible Research

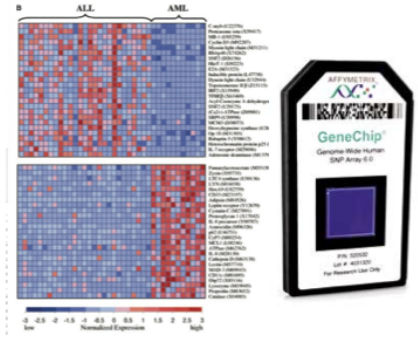
Genomic data are complex and heterogeneous

Source: slide from JP Vert

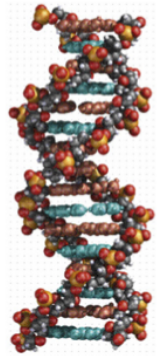


Interactome

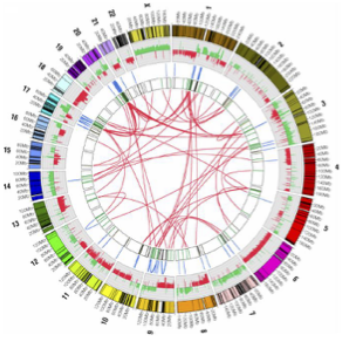
Publications



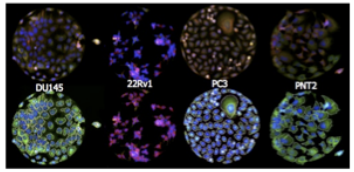
Transcriptome



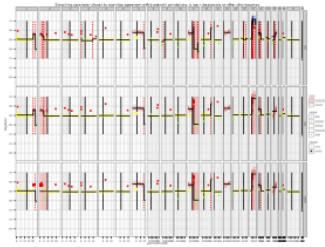
Genome



Mutations
Structural variations



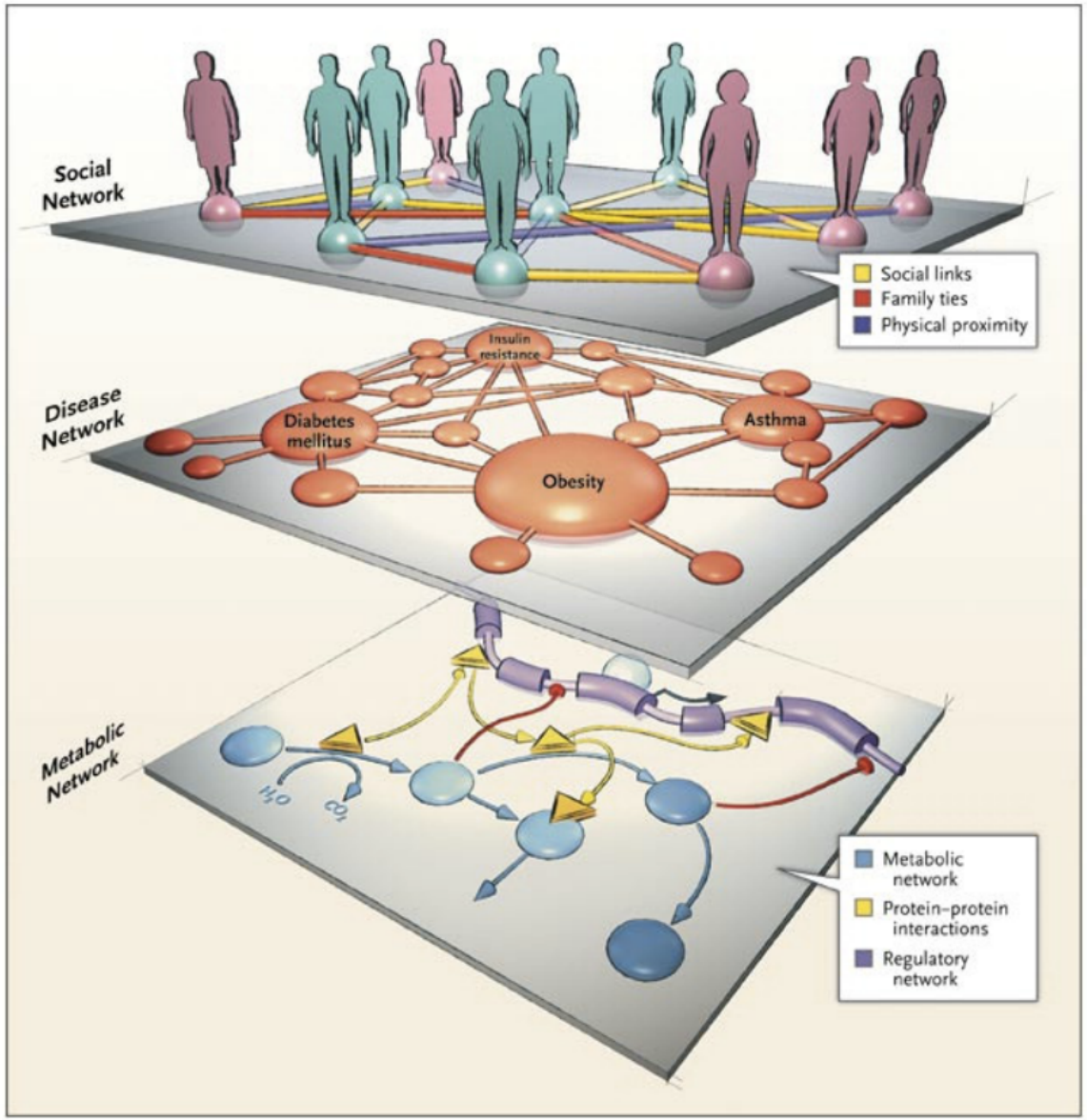
Phenome



Epigenome

Genomic data are structured

Source: Barabási, N Engl J Med 2007; 357:404-407}



The need for dedicated analysis methods and software

Characteristics of genomic data

- high-dimensionality: "$n \ll p$"

e.g. $n = 1,000$ patients and $p = 1,000,000$ genomic markers

---

- complexity and heterogeneity (various data sources)
- strong structuration (through the underlying biological reality)

Challenge: how best to combine

- mathematically-grounded models and methods
- computationally-efficient algorithms
- interpretable results

$=>$ statistical/computational/biological tradeoffs

Why R?

From the website `http://RevolutionAnalytics.com`:

"*During the last decade, the momentum coming from both academia and industry has lifted the R programming language to become the single most important tool for computational statistics, visualization and data science. Worldwide, millions of statisticians and data scientists use R to solve their most challenging problems in fields ranging from computational biology to quantitative marketing. R has become the most popular language for data science and an essential tool for Finance and analytics-driven companies such as Google, Facebook, and LinkedIn.*"

Why R?

. . . not only because I'm a statistician

Many R features are useful to genomic data analysis

- data management, I/O
- linear algebra
- descriptive statistics
- inferential statistics (tests, regression)
- programming (in plain R or interfaced with C++)
- graphics
- packages
- reproducible research

Of course R is not always the best tool!

"If all you have is a hammer, everything looks like a nail"

Some references and pointers

- See course of Julien Chiquet: "Programmation sous R": `http://julien.cremeriefamily.info/teachings_L3BI_ISV51.html`

R community blogs

- `http://www.inside-r.org`
- `http://www.r-statistics.com`
- `http://www.r-bloggers.com`

R package repositories dedicated to genomics

- `http://bioconductor.org`
- `http://aroma-project.org`