Tasks & Challenges in Data Analysis of Large Human Genome Sequences Data Sets

Fabio Macciardi
Department of Psychiatry & Human Behavior
Abnormal behavior reflects abnormal brain function

Genes:
- risk associated genotypes

Cells:
- molecular biology

Systems:
- abnormal information processing

Psychiatric Disorder
- complex functional interactions and emergent phenomena

Behavior:
- temperament

Cognition
- The Wisconsin Card Sorting Task
Sequencing Costs Plummeting

Cost per genome

The cost of sequencing fell dramatically a few years after high-throughput sequencing machines were introduced.

1,400 $
Exploding data volumes
The perfect neuroimaging-genetics-computation storm: collision of petabytes of data, millions of hardware devices and thousands of software tools

Ivo D. Dinov • Petros Petrosyan • Zhizhong Liu • Paul Eggert • Alen Zamanyan • Federica Torri • Fabio Macciardi • Sam Hobel • Seok Woo Moon • Young Hee Sung • ZhiGuo Jiang • Jennifer Labus • Florian Kurth • Cody Ashe-McNalley • Emeran Mayer • Paul M. Vespa • John D. Van Horn • Arthur W. Toga • for the Alzheimer’s Disease Neuroimaging Initiative

[.....] The volume, diversity and velocity of biomedical data are exponentially increasing providing **petabytes of new neuroimaging and genetics data every year**. At the same time, tens-of-thousands of computational algorithms are developed and reported in the literature along with thousands of software tools and services.

Users demand intuitive, quick and platform-agnostic access to data, software tools, and infrastructure from millions of hardware devices. This explosion of information, scientific techniques, computational models, and technological advances leads to enormous challenges in data analysis, evidence-based biomedical inference and reproducibility of findings.
huge flux of data

..and now what?
This is what we get
A view of the genome...

(... today)

RNASeq: Transcripts, splicing, expression profiling

Methylation

Histone Modifications

Promoter

CAGE Seq

Non-coding RNA

Non-coding RNA

Coding gene...

SNP

Insulator CTCF ChIP Seq

Enhancer

P300 ChiPSeq

[from E. Stupka - UCL]
Current challenges in BIG

• Thousands of sequenced genomes with BI data
• Storage volume = 1,000 genomes ~ 80 TB
• Working space = 1 genome -> ~ 1 TB of data with ZOT (Zillions of tiny) files

• Updating reference human genome(s) = 38 releases since 2001
• Genomic variants = ~ 4.50E+06/subj -> 4.50E+09
• Integrating annotation and functional genomics
• ......
More Data $\neq$ More Knowledge
Systems Genetics is the Future

Population Genomics
- Human Genome
- HapMap
- GWAS
- 1000 Genomes
- Sequencing

Functional Genomics
- DNA Microarrays
- Proteomics
- miRNA
- Methylation
- ENCODE

Technology

Bioinformatics

Systems Genetics

Yesterday  Today  Tomorrow

10 yrs