Data-warehousing of public genomic datasets

Current topics in computer science

David Weiss

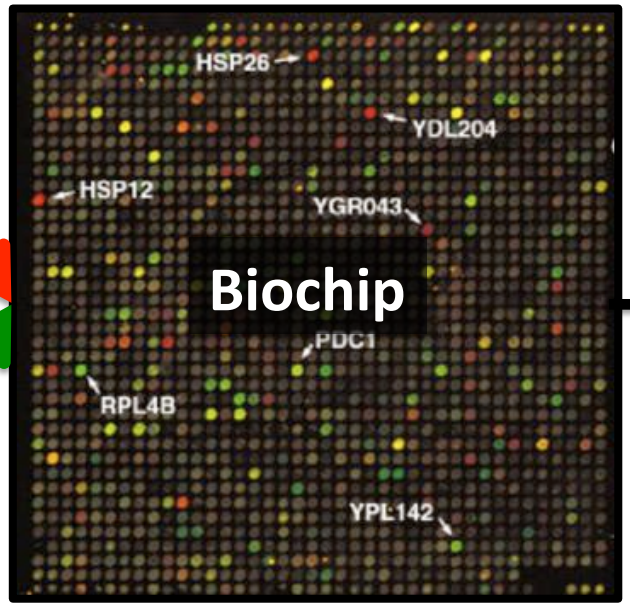
dweiss@ulb.ac.be

Feb 9, 2011

Personalized medicine: based on the whole genetic profile of each patient

Cancer gene expression

Normal tissue gene expression



Genetic profile

Genes	Cancer 1
HSP26	+2
YPL142	-2.5
…	…
gene 20,000	0

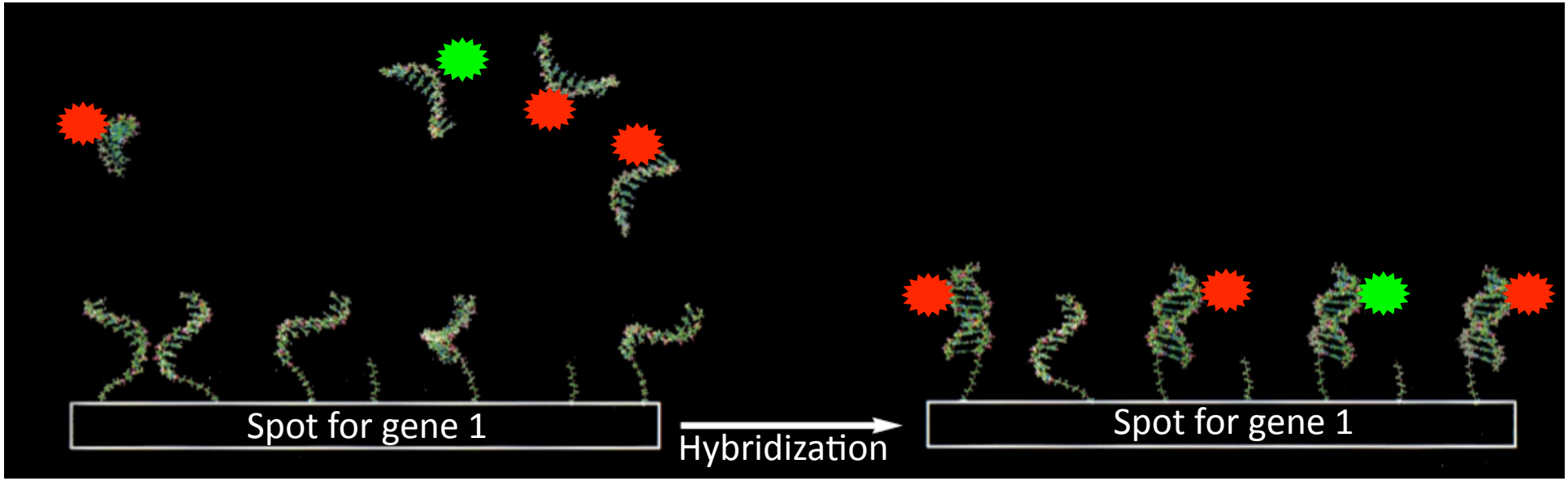
Profiling 100 cancer patients in a clinical trial

100 columns x 20,000 lines spreadsheet ➔ Need to program

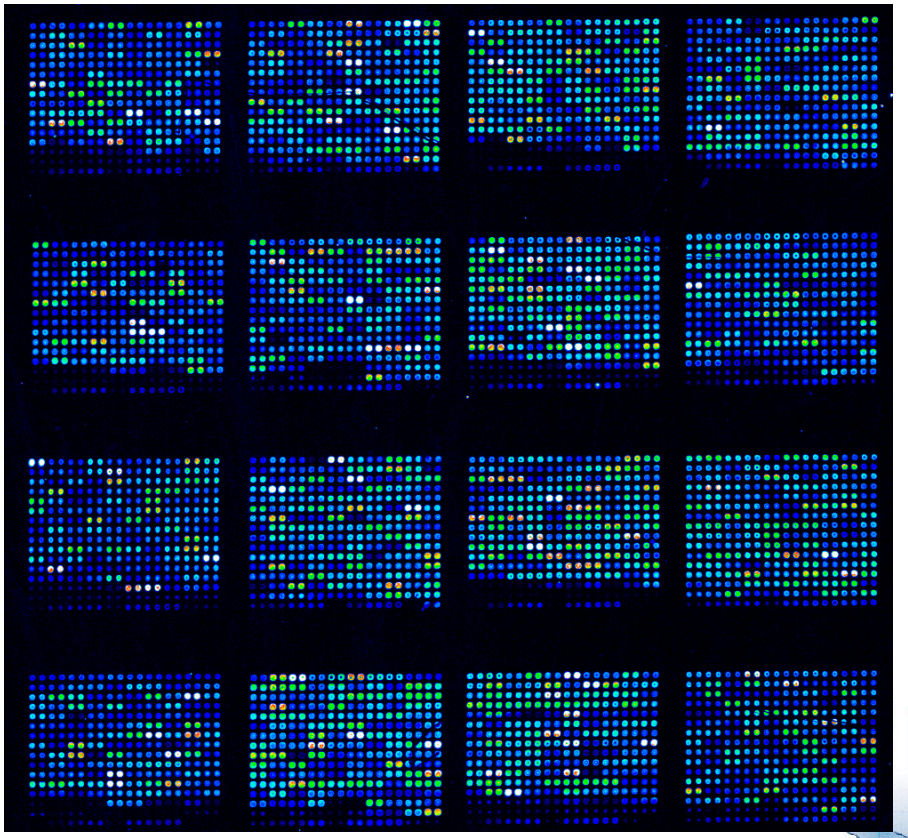
David Weiss - dweiss@ulb.ac.be

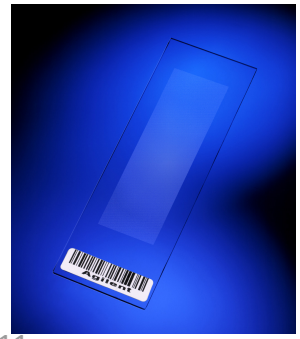
Microarrays apply complementarity for sequence detection in a spatially parallel way

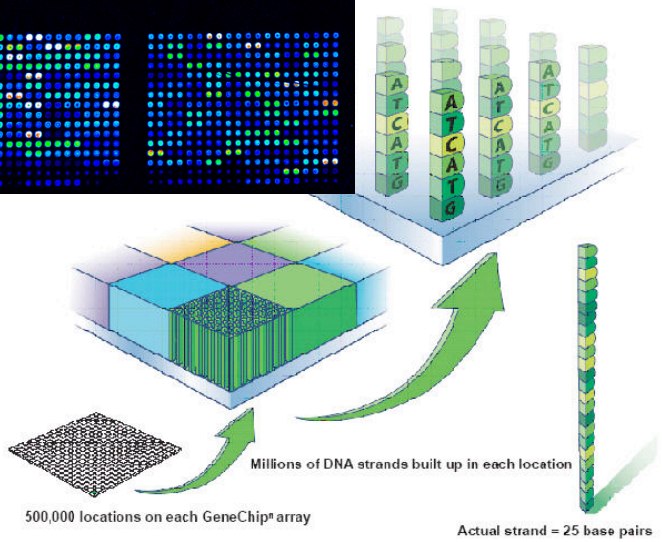
The sequence corresponding to one gene is **fixed** on a location or *spot* of a microarray



Single channel arrays have become the standard

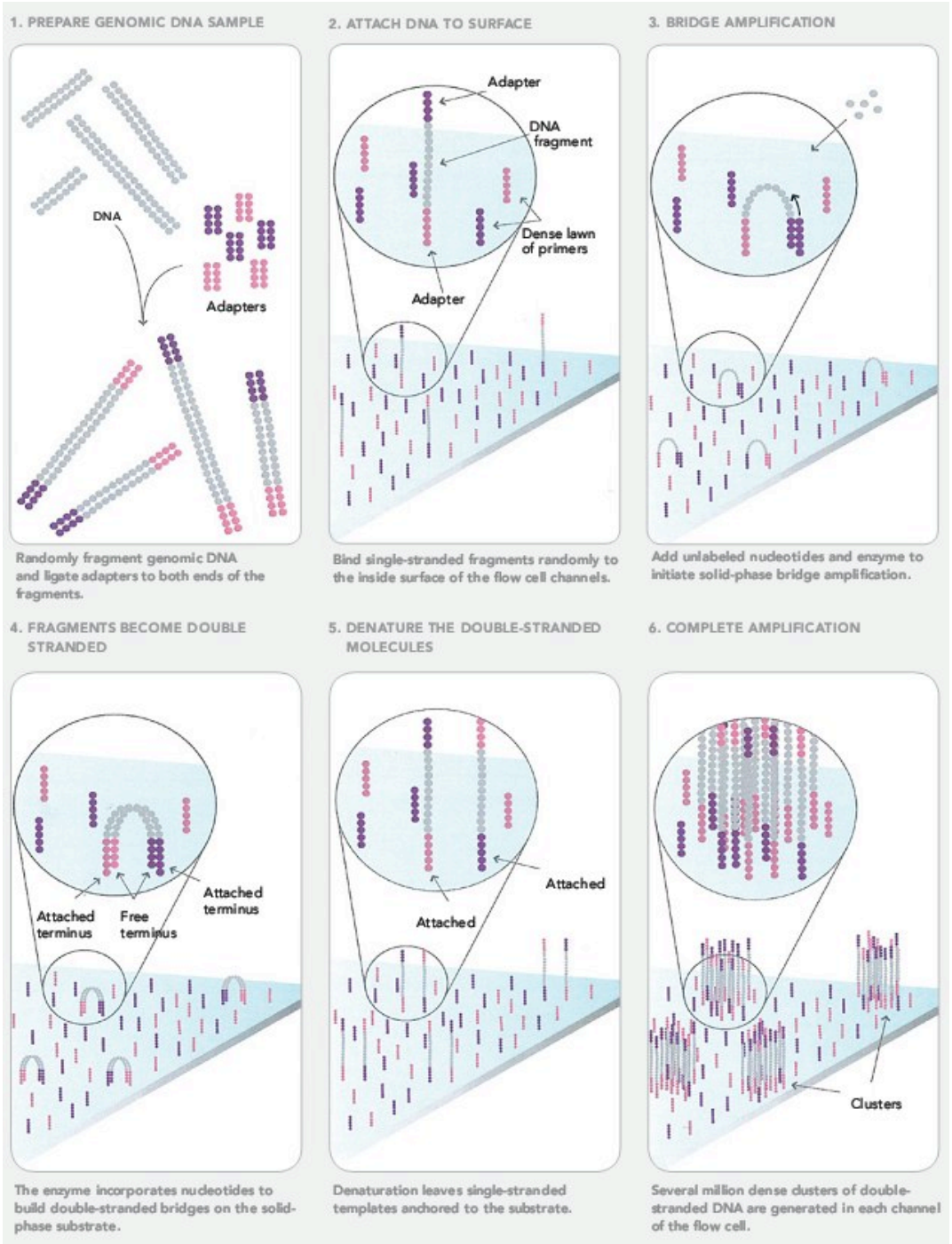






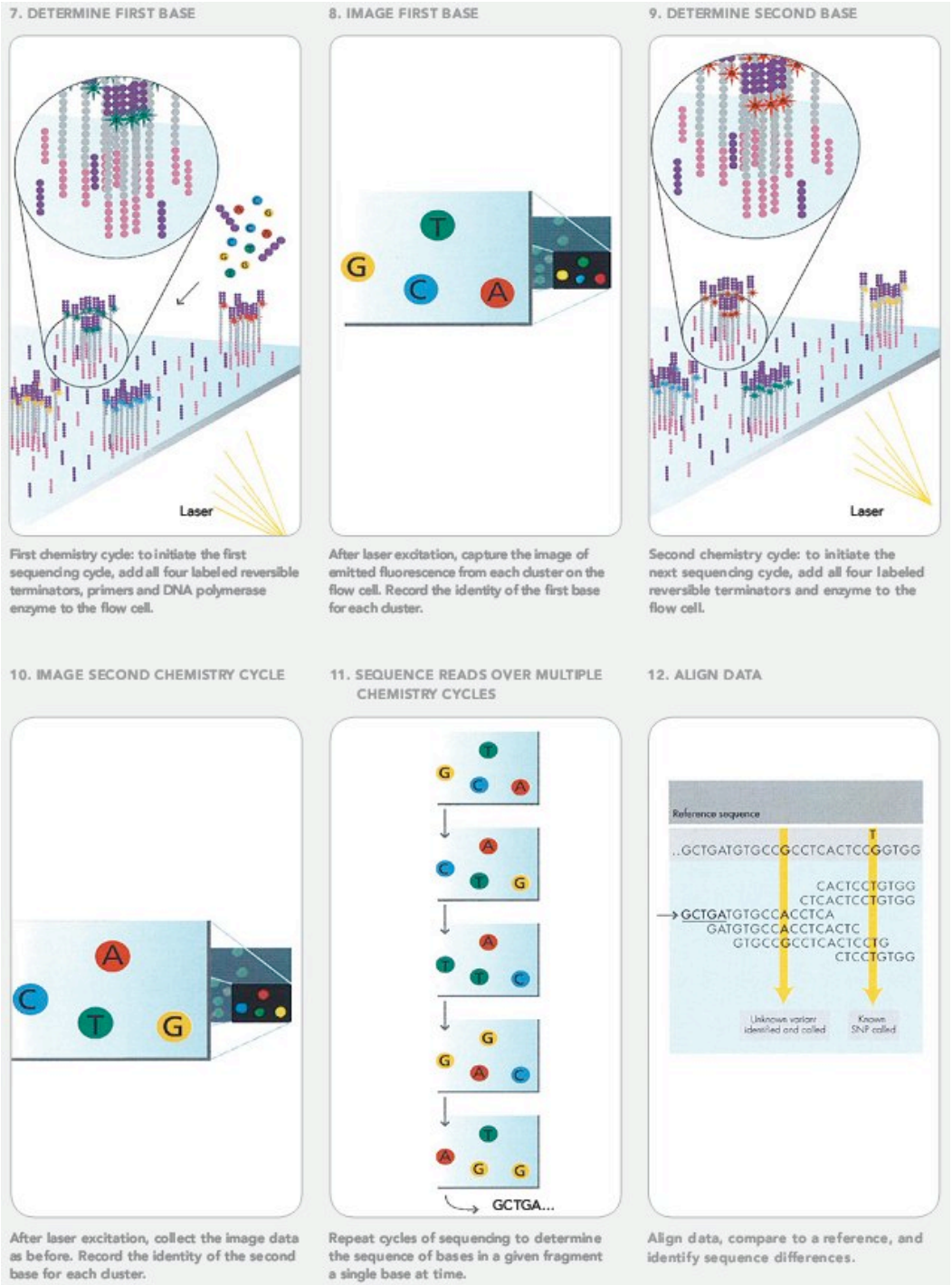
- Microarrays measure the mRNA activity of *all* the genes in a single experiment
- One can detect which genes show aberrant activation in diseases
- These may have diagnostic or therapeutic value

David Weiss - dweiss@ulb.ac.be



De-Novo sequencing

More powerful tehnology, and more challenging

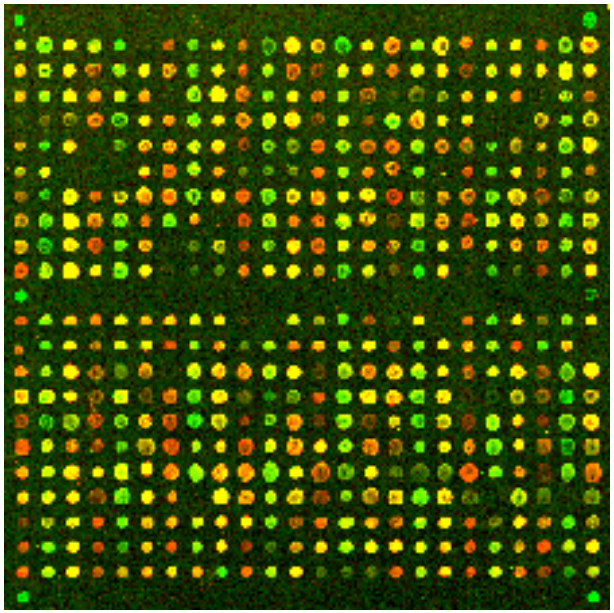


De-Novo sequencing

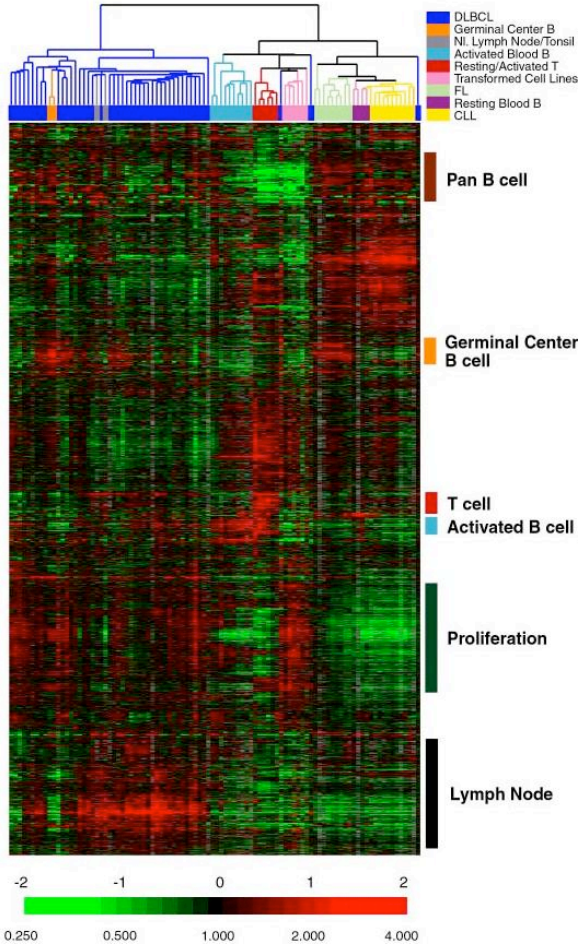
~1TB per sample generated

Pre-processing of data still subject of active research

Machine learning methods can be applied to genome output: class discovery and class prediction



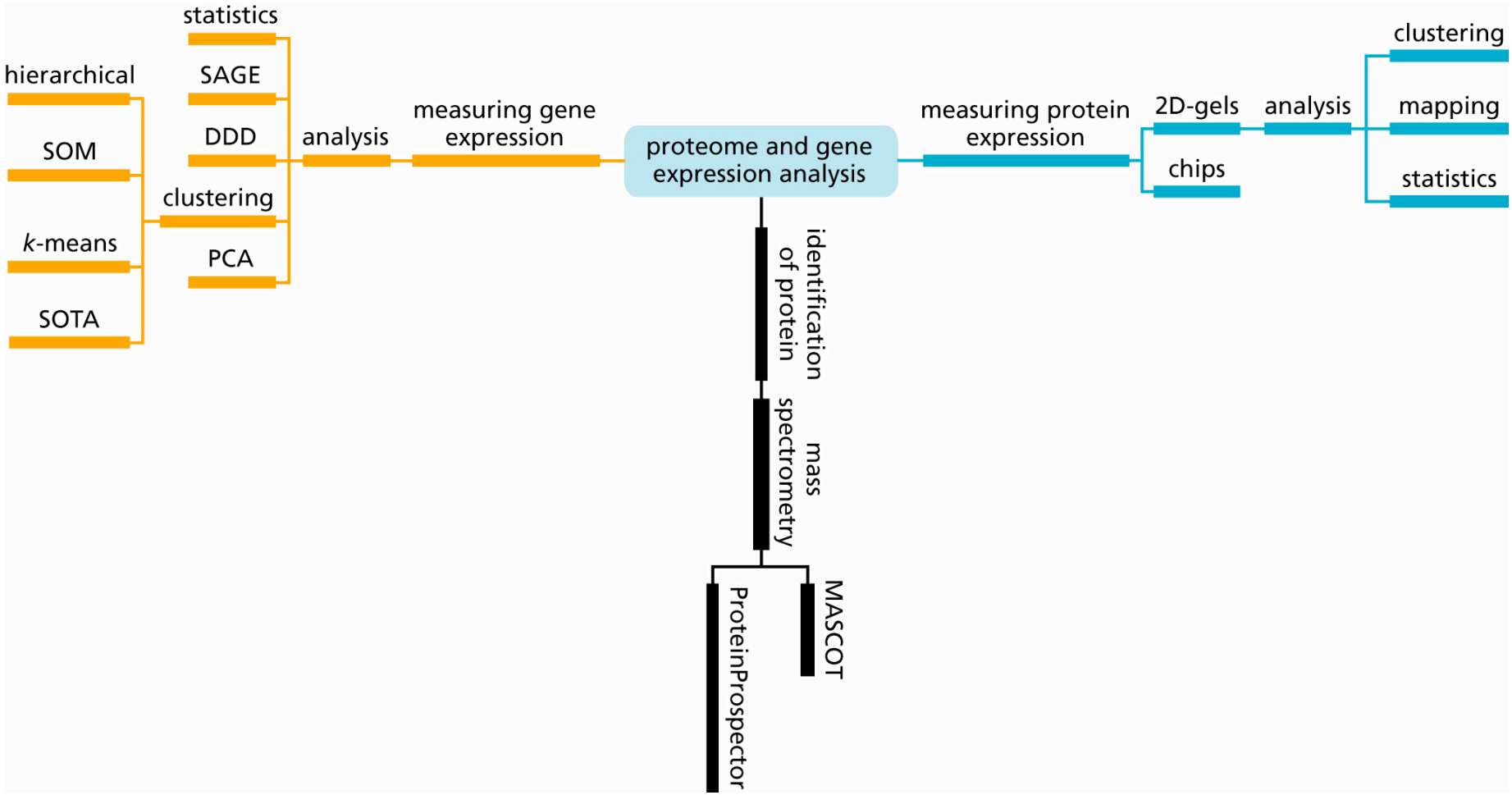
Microarray chip



Nature. 2000 Feb 3;403(6769):503-11

David Weiss - dweiss@ulb.ac.be

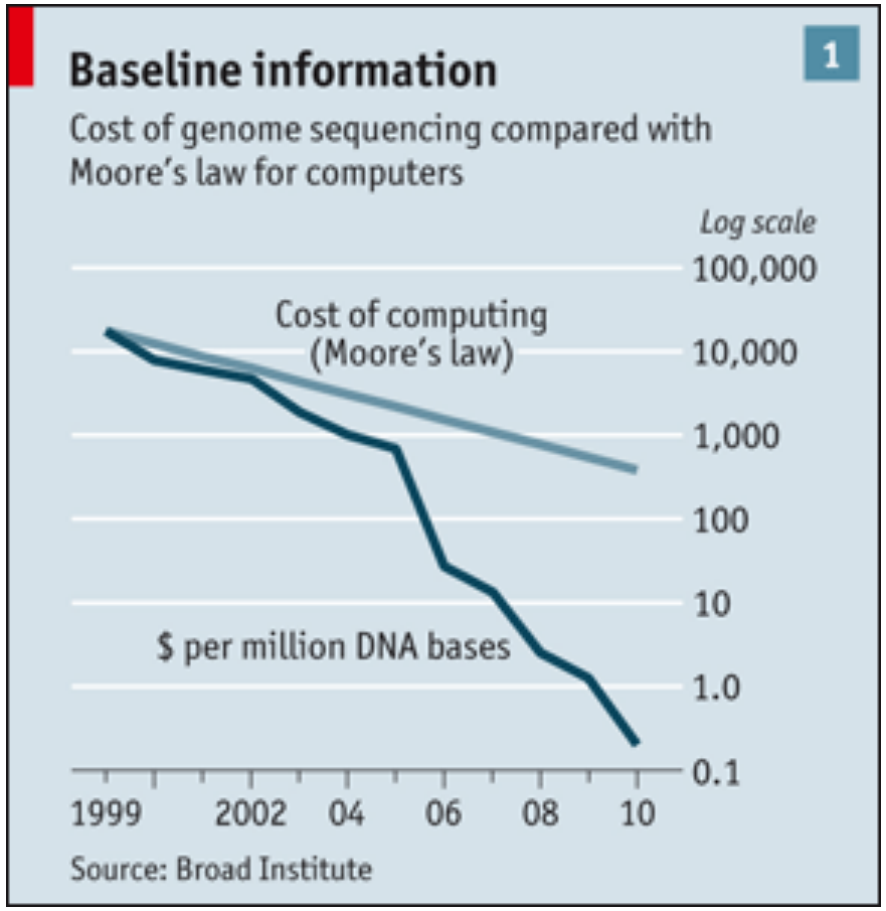
Diverse analysis tools can be applied to genomic data



Zvelebil, MJ, Understanding bioinformatics, Garland Science 2008

David Weiss - dweiss@ulb.ac.be

Genome-wide measurements approaching mainstream but, how can these data be accessed and compared?



The Economist - 2010

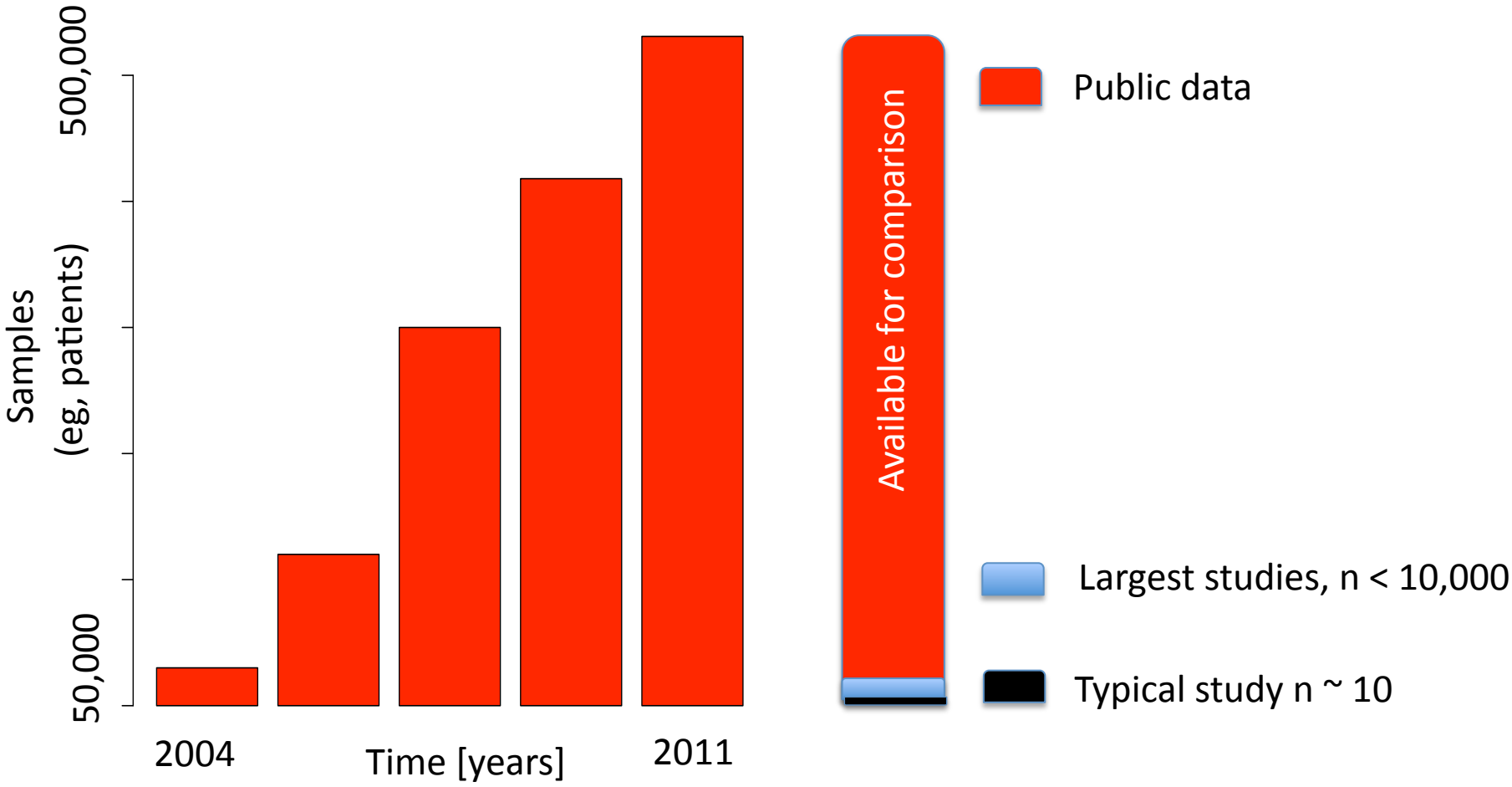
- Many labs (anyone) can obtain genome scans
  - Sequencing a human costs ~$10,000 (There is a project to sequence all the US army: 1.4 million profiles)
  - Used in clinical trials
  - Inevitably every doctor's visit will include a genome scan

But how to access and compare these data?

  - There is no standard for data storage and exchange
  - The data are thus fragmented and not interoperable
  - Training resources to handle data is very expensive and time consuming

David Weiss - dweiss@ulb.ac.be

Accumulating public data is a basis for comparison and interpretation but no tools exist to do so easily



No easy way to compare proprietary and public data!

Up to weeks of work for manual compilation by computational biologist

David Weiss - dweiss@ulb.ac.be

The existing primary databases are either insufficiently structured or incomplete

- European Bioinformatics Institute's ArrayExpress
  - MIAME standard, article cited 2,307 times, but
  - Not widely adopted
- National Center for Biotechnology Information USA, GEO database
  - No standard enforcement
  - Data accumulates but is unstructured

- Example: GSE2109, E-GEOD-12630

David Weiss - dweiss@ulb.ac.be

Clinical, zoological and experimental vocabularies are hard to manage

- No single standard ontology
- Two approaches:
  - Top-down
  - Bottom-up
- Types of variables: continuous, ordered/unordered categorical
- Units, hierarchies
- Format neutrality

David Weiss - dweiss@ulb.ac.be

Success stories @ ULB: Biomedical progress through data-warehousing, comparing proprietary to public datasets

- Combining datasets from different private and public sources:
  - Thyroid cancer and experimental models compendium

    Discovery of new cancer mechanism: repression of negative feedbacks in oncogenic pathway cAMP:

    Proc Natl Acad Sci U S A. 2006 Jan (cover)

  - ULB colleague from Bordet Institute, B. Haibe-Kains: Breast cancer compendium: Several high-level publications, clinical applications

- Skills needed: software engineering, multivatriate statistics, linear algebra, molecular biology

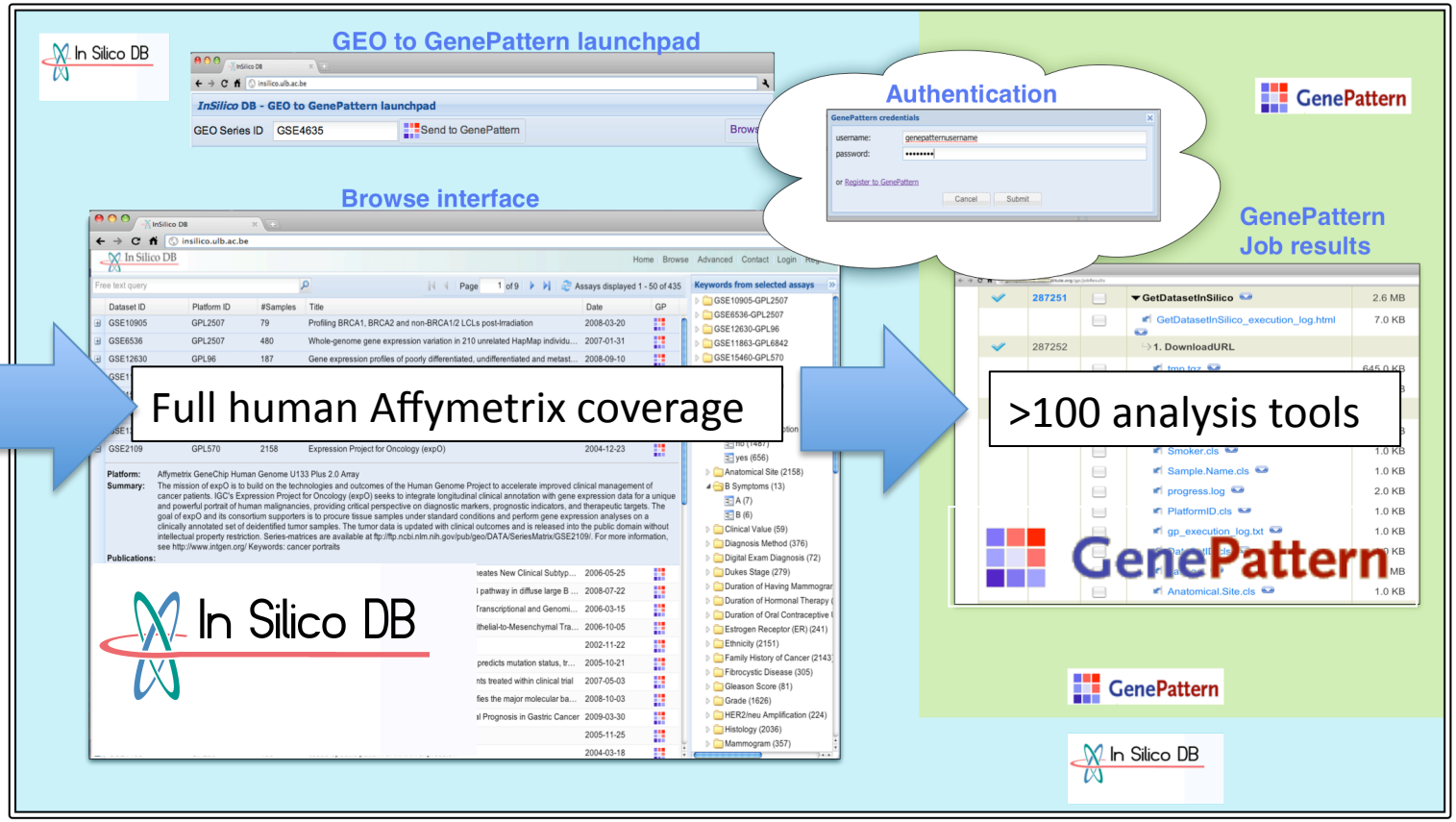
=> Develop infrastructure to facilitate access to public datasets: InSilico DB

David Weiss - dweiss@ulb.ac.be

InSilico DB aggregates content and serves it through secure integration with bioinformatics AI tools

http://insilico.ulb.ac.be

Gene Expression Omnibus



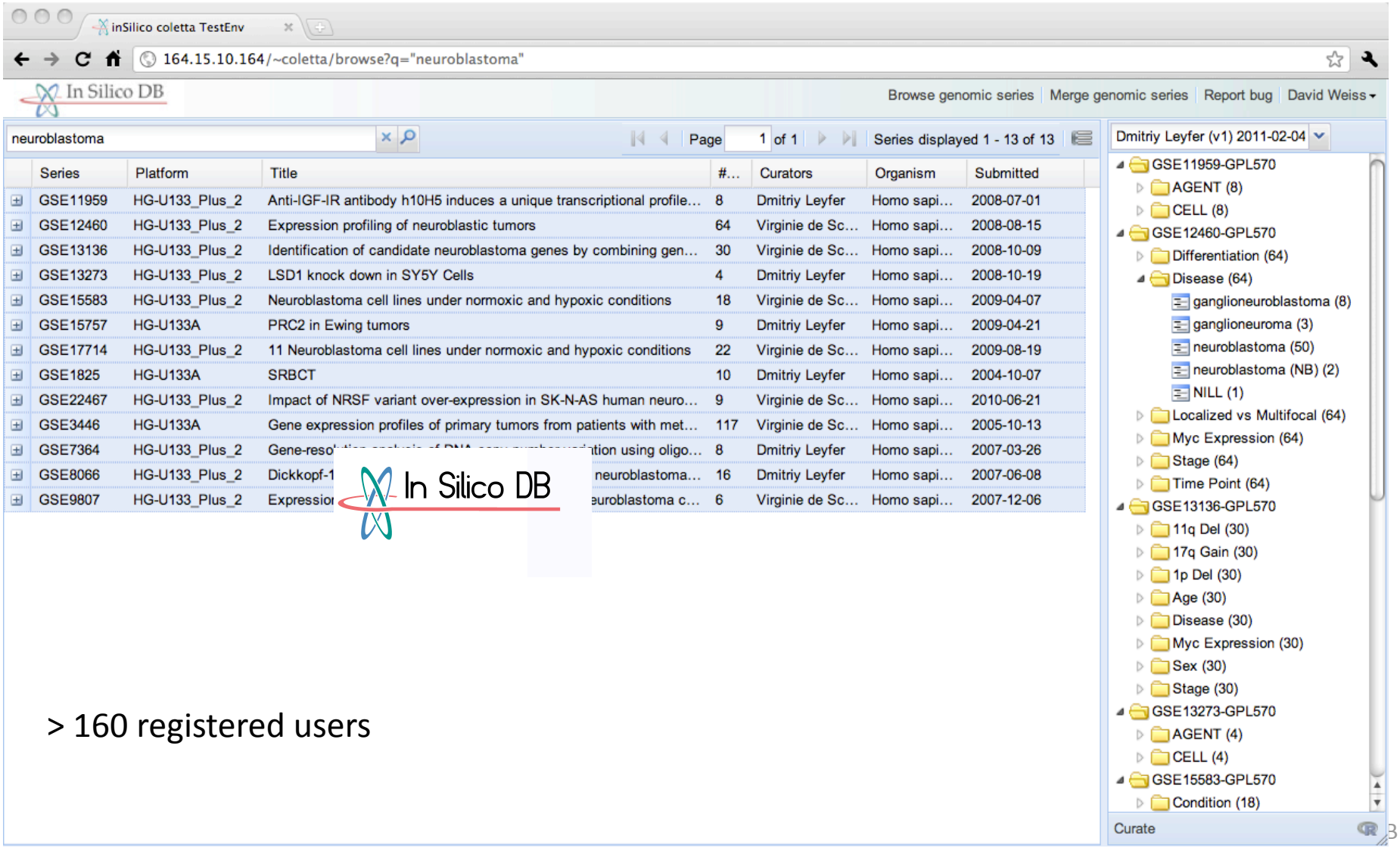
Private

160 registered users

David Weiss - dweiss@ulb.ac.be

InSilico DB web-app[1/3]: Database browsing http://insilico.ulb.ac.be/browse

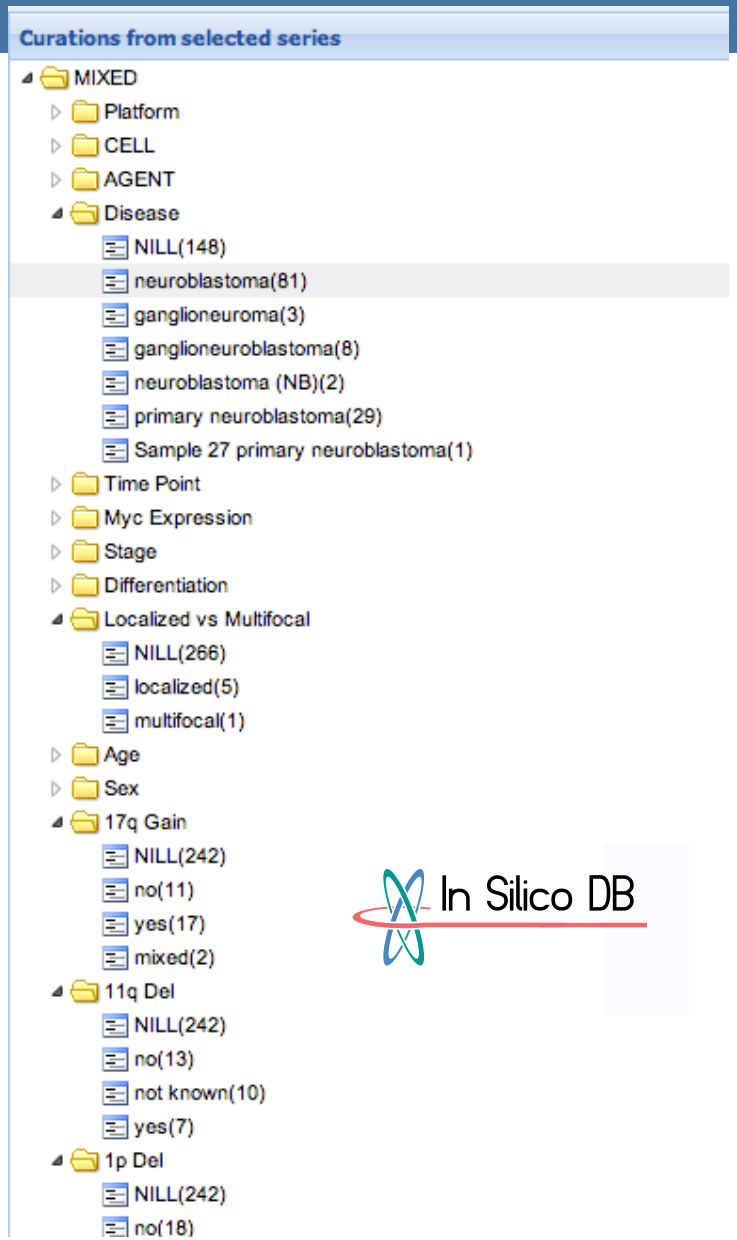
- Biological samples information available in multiple versions



> 160 registered users

InSilico DB web-app[2/3]: Datasets Merging

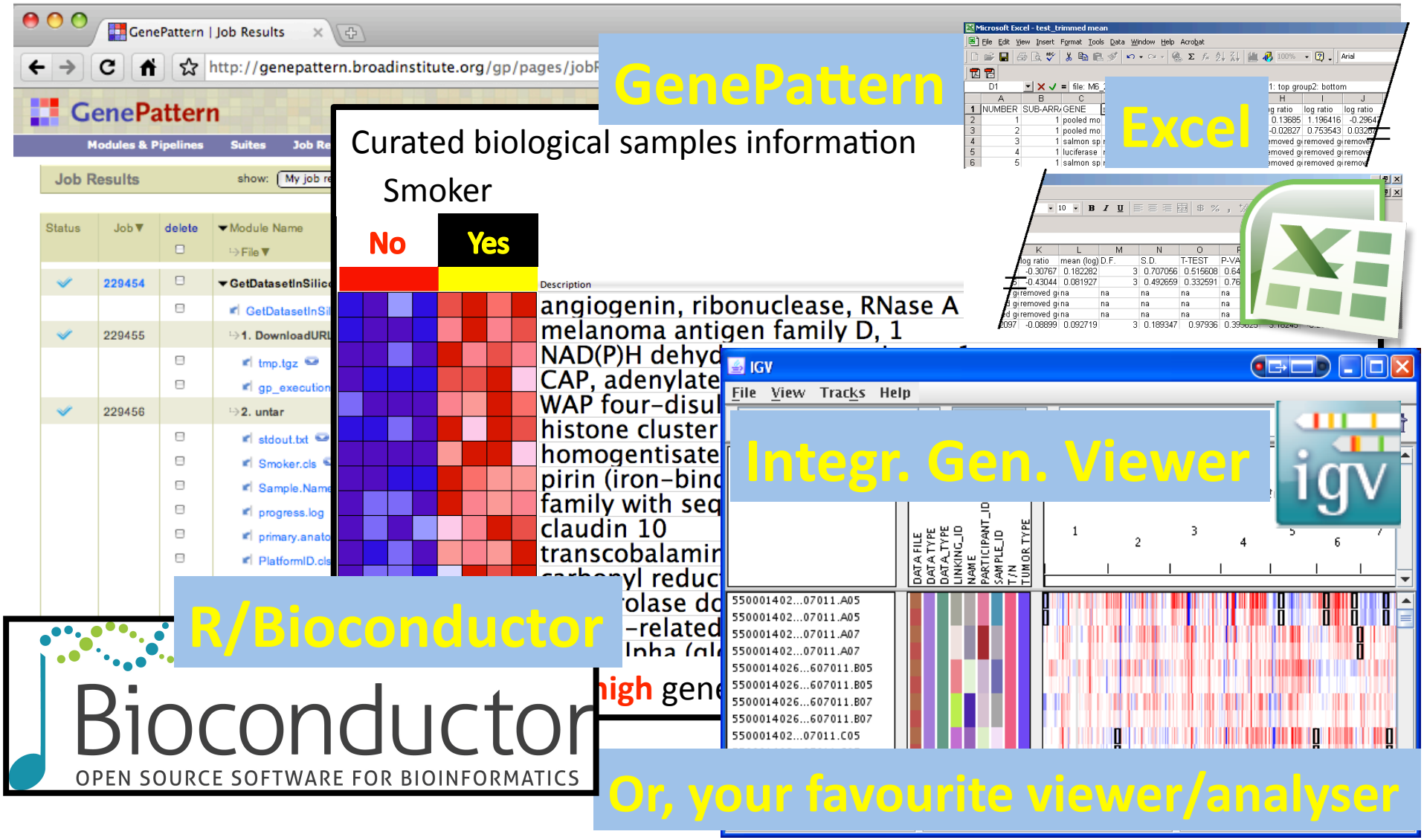
- By-request:
  - Comparing public and proprietary data, incl. Next-Generation Sequencing
  - Possible to produce on-demand compendia on e.g., lung cancer, neuroblastoma, etc.
  - Merging of studies with quality control and state-of-the-art bias-corrective algorithms



David Weiss - dweiss@ulb.ac.be

InSilico DB web-app[3/3]: Easy export to multiple visualization and analysis platforms

User's datasets available through e.g., GenePattern (MIT/Harvard), >100 modules available



Programming organization

- Large-scale data and computation
- Hands-on programming knowledge needed
- Design is the common language
- Agile methods
- Pair programming

David Weiss - dweiss@ulb.ac.be

Programming environment

- LAMP infrastructure
  - Linux
  - Apache
  - MySQL
  - PHP

- Scripting languages
- Web application framework
- JavaScript library
- Statistical programming: R/Bioconductor packages for genomics data analysis

David Weiss - dweiss@ulb.ac.be

Programming Team

- R&D:
  - 4 Software engineers – 2 PhDs
  - 1 Biocurator
  - 2 data analysts

- **Looking for:**
  - Students
    - Undergraduate thesis/ PhD candidates
    - Student jobs
    - Skills: good programmer/fast-learning

David Weiss - dweiss@ulb.ac.be

Thank you

David Weiss - dweiss@ulb.ac.be