

Data-warehousing of public genomic datasets

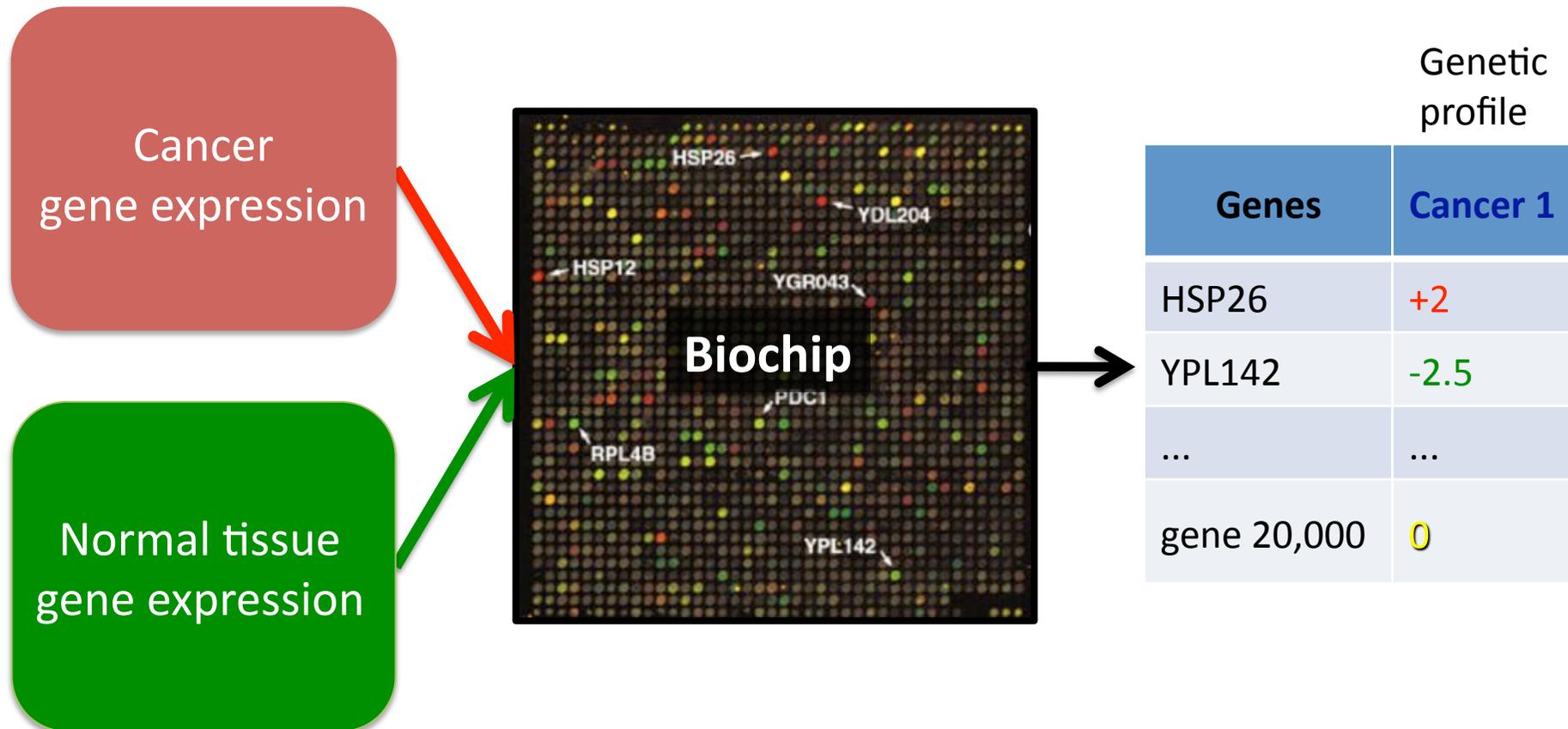
Current topics in computer science

David Weiss

dweiss@ulb.ac.be

Feb 9, 2011

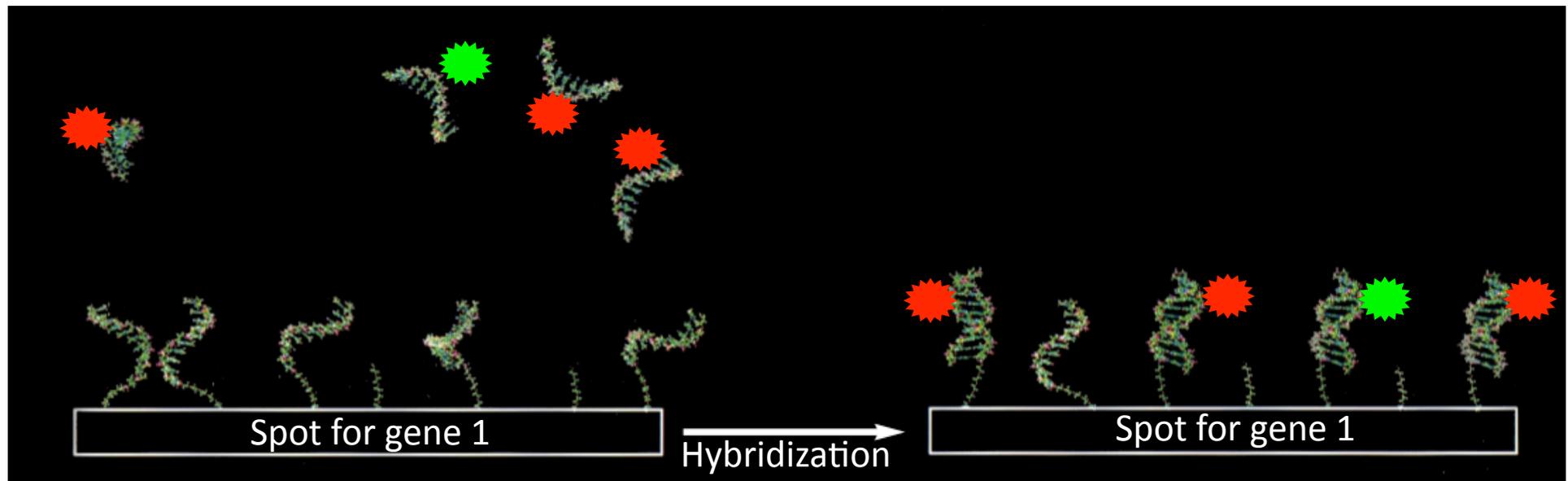
Personalized medicine: based on the whole genetic profile of each patient



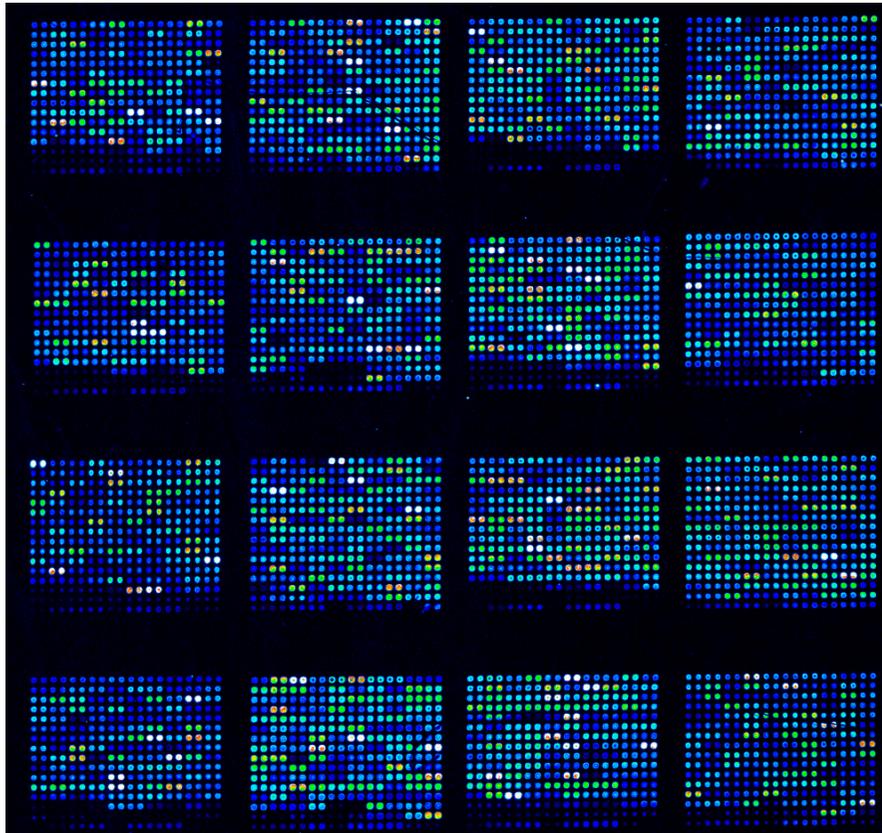
Profiling 100 cancer patients in a clinical trial
100 columns x 20,000 lines spreadsheet → Need to program

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



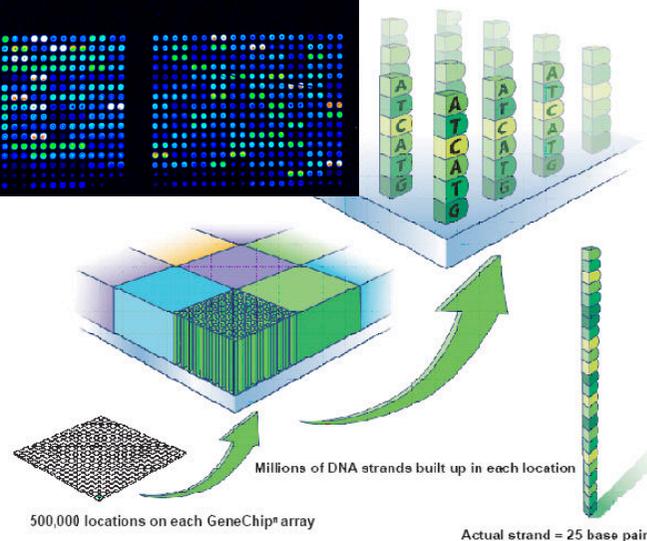
Single channel arrays have become the standard



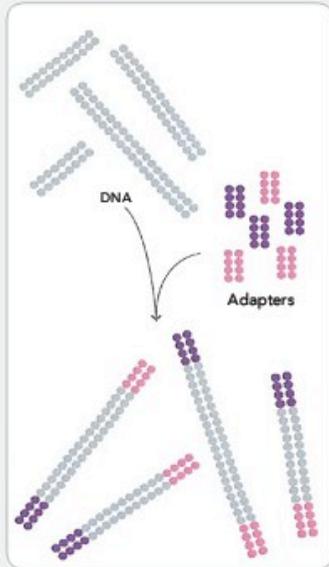
- Microarrays measure the mRNA activity of *all* the genes in a single experiment

- One can detect which genes show aberrant activation in diseases

- These may have diagnostic or therapeutic value

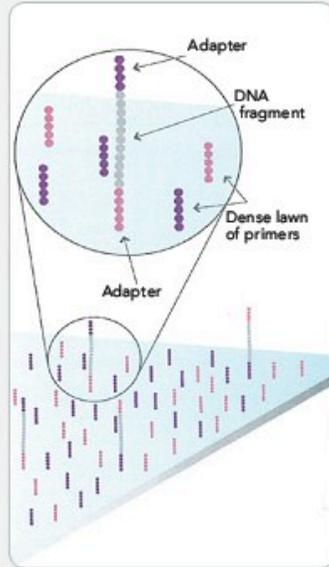


1. PREPARE GENOMIC DNA SAMPLE



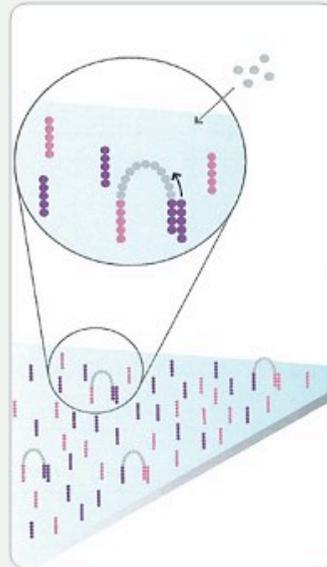
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



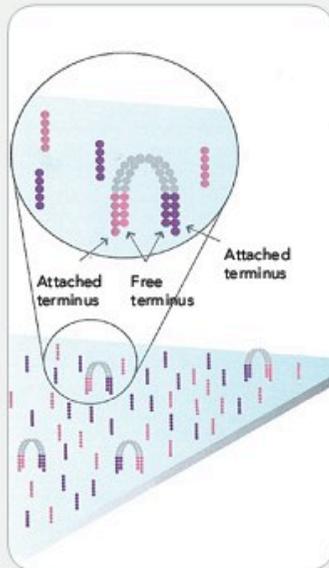
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION



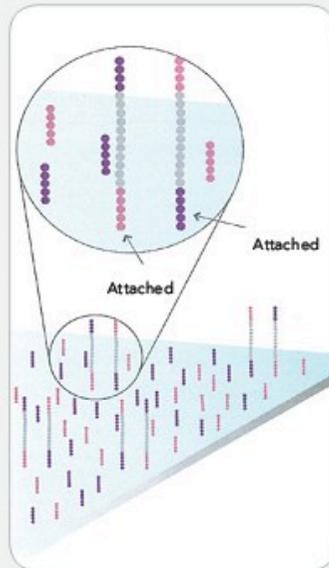
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED



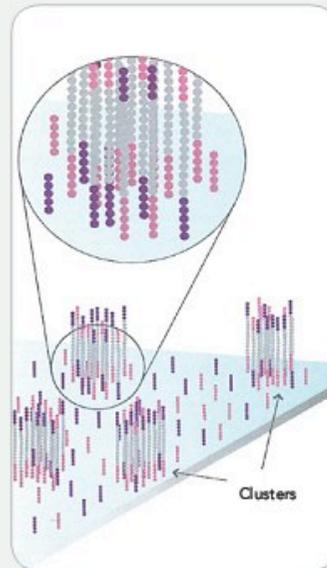
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

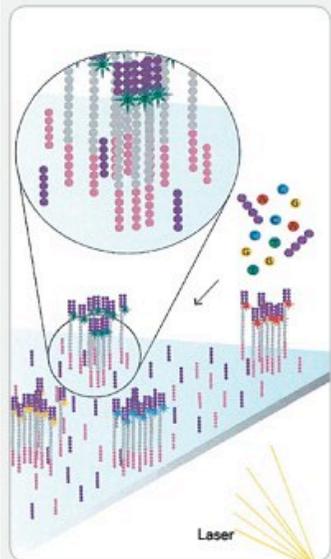


Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

De-Novo sequencing

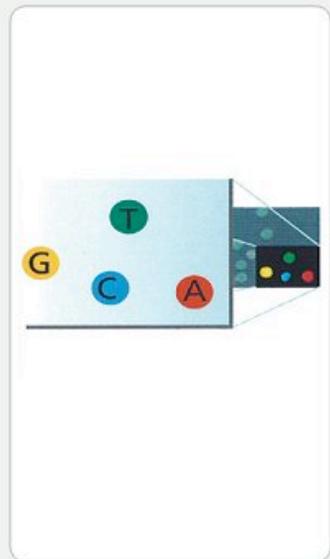
More powerful technology, and more challenging

7. DETERMINE FIRST BASE



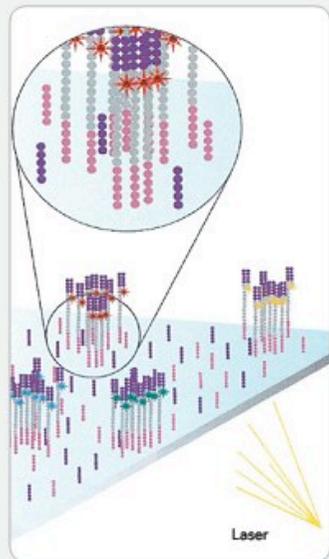
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



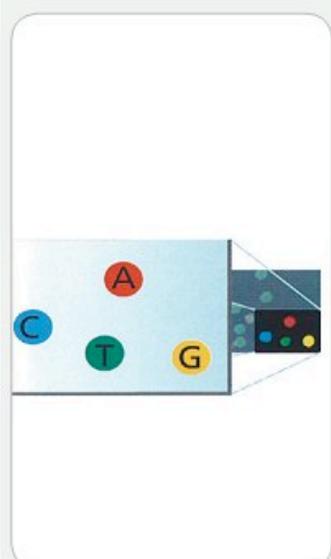
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



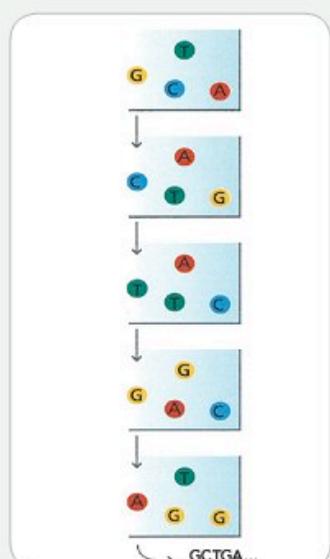
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



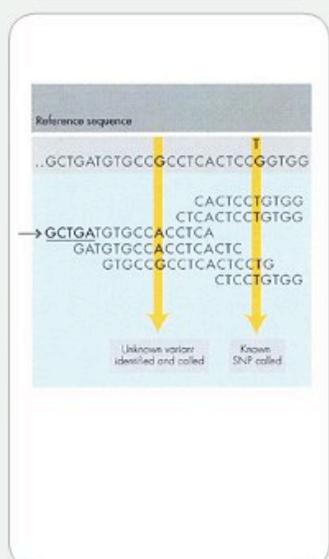
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA



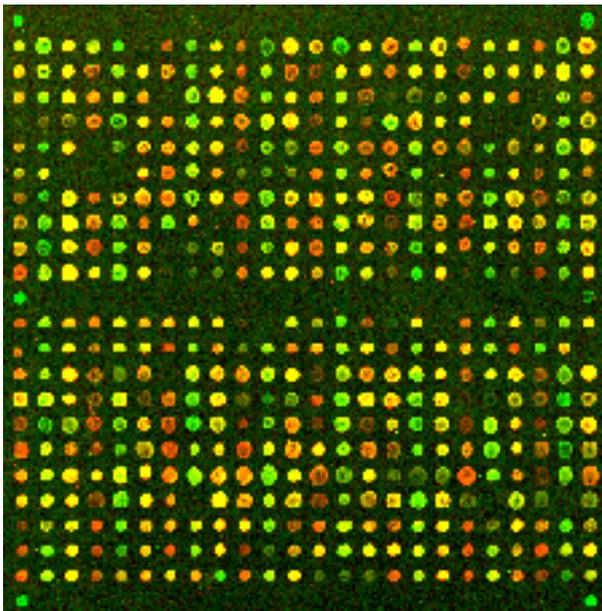
Align data, compare to a reference, and identify sequence differences.

De-Novo sequencing

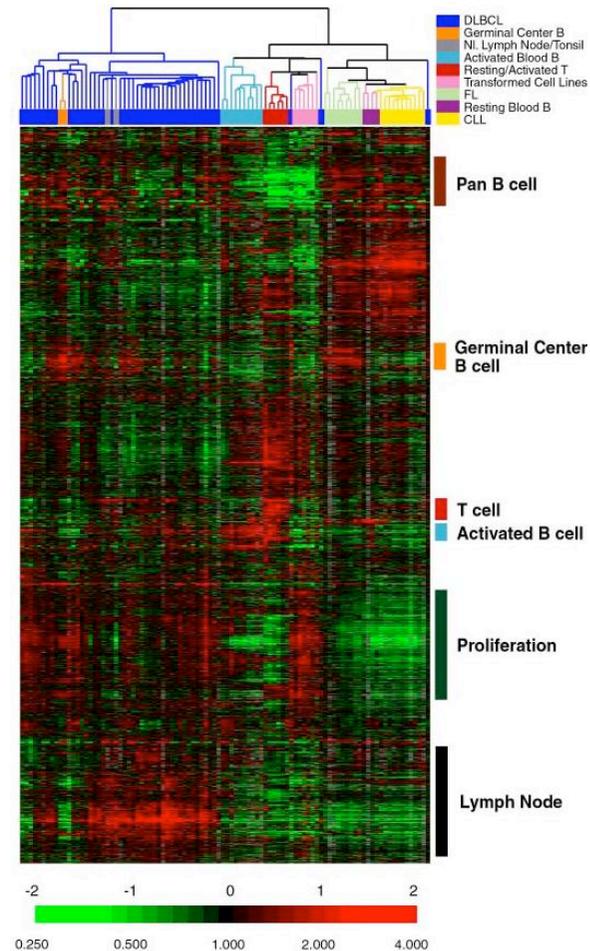
~1TB per sample generated

Pre-processing of data still subject of active research

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

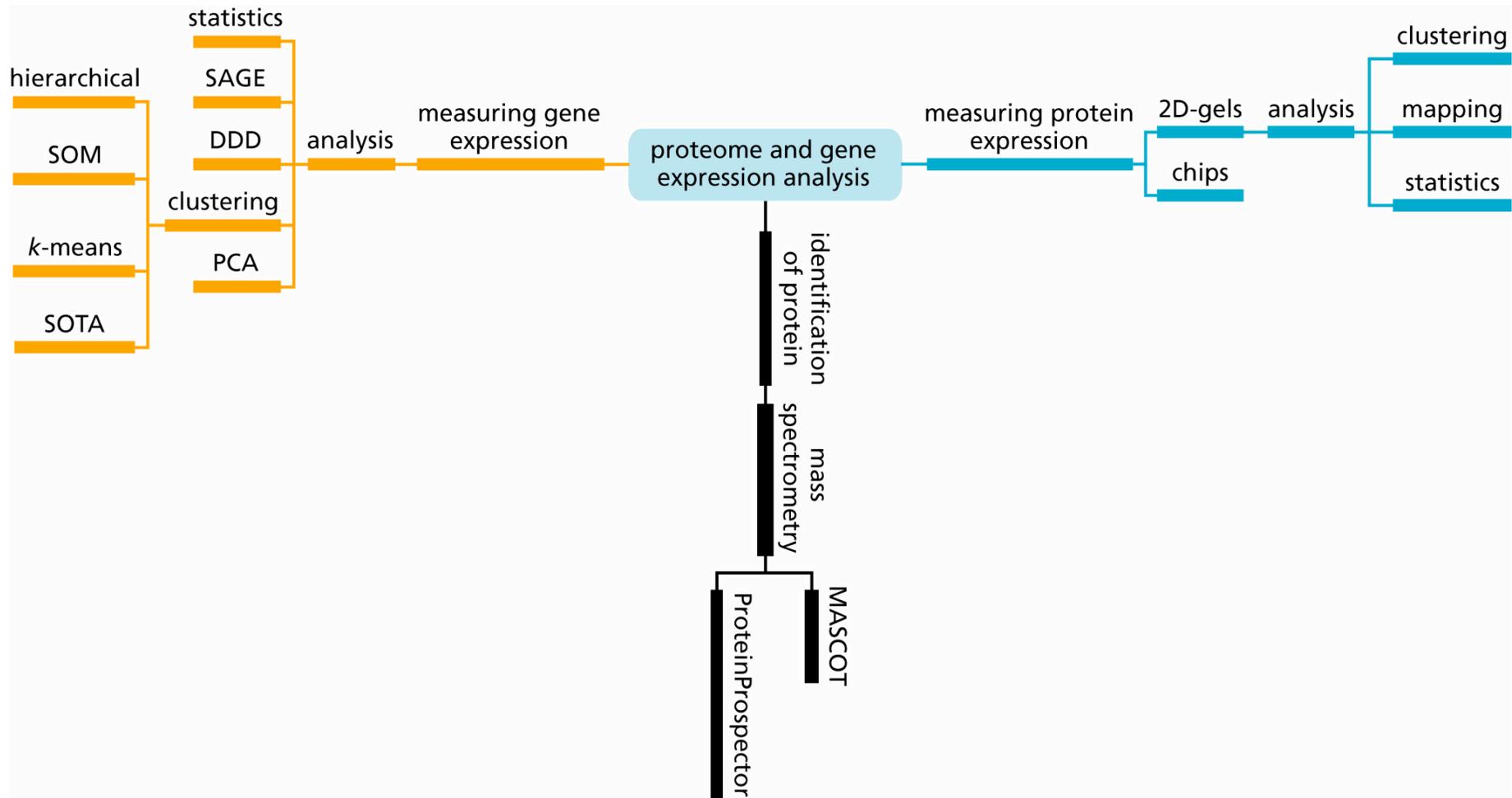


Microarray chip



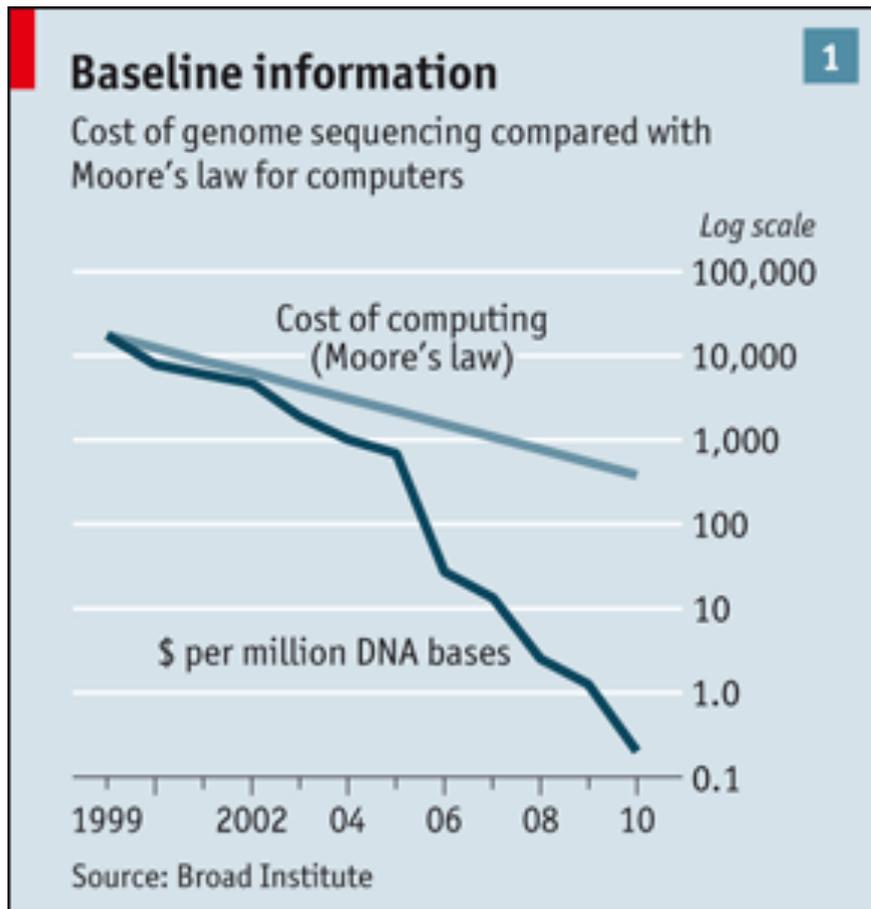
Nature. 2000 Feb 3;403(6769):503-11

Diverse analysis tools can be applied to genomic data



Zvelebil, MJ, Understanding bioinformatics, Garland Science 2008

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



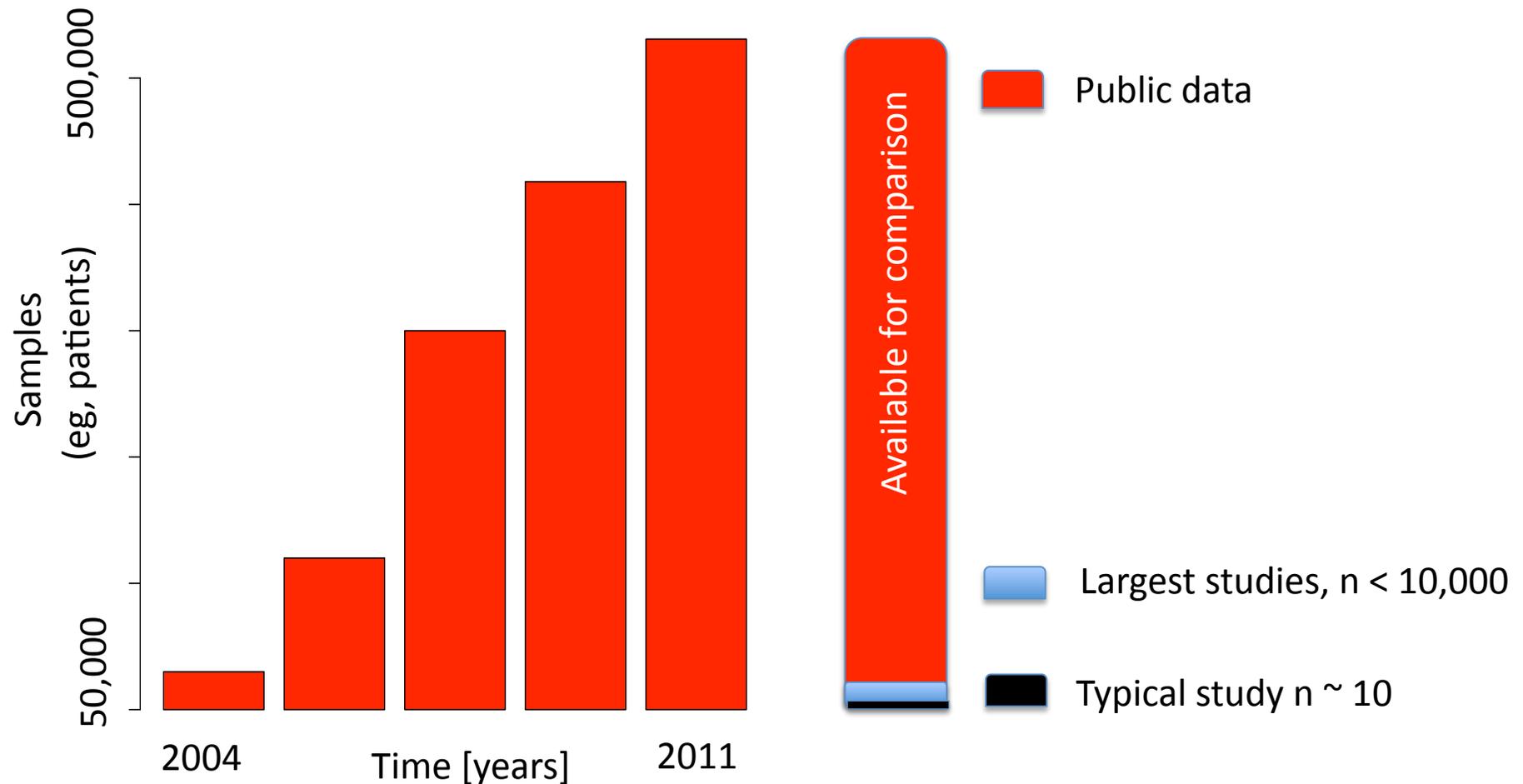
The Economist - 2010

- 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

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

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

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

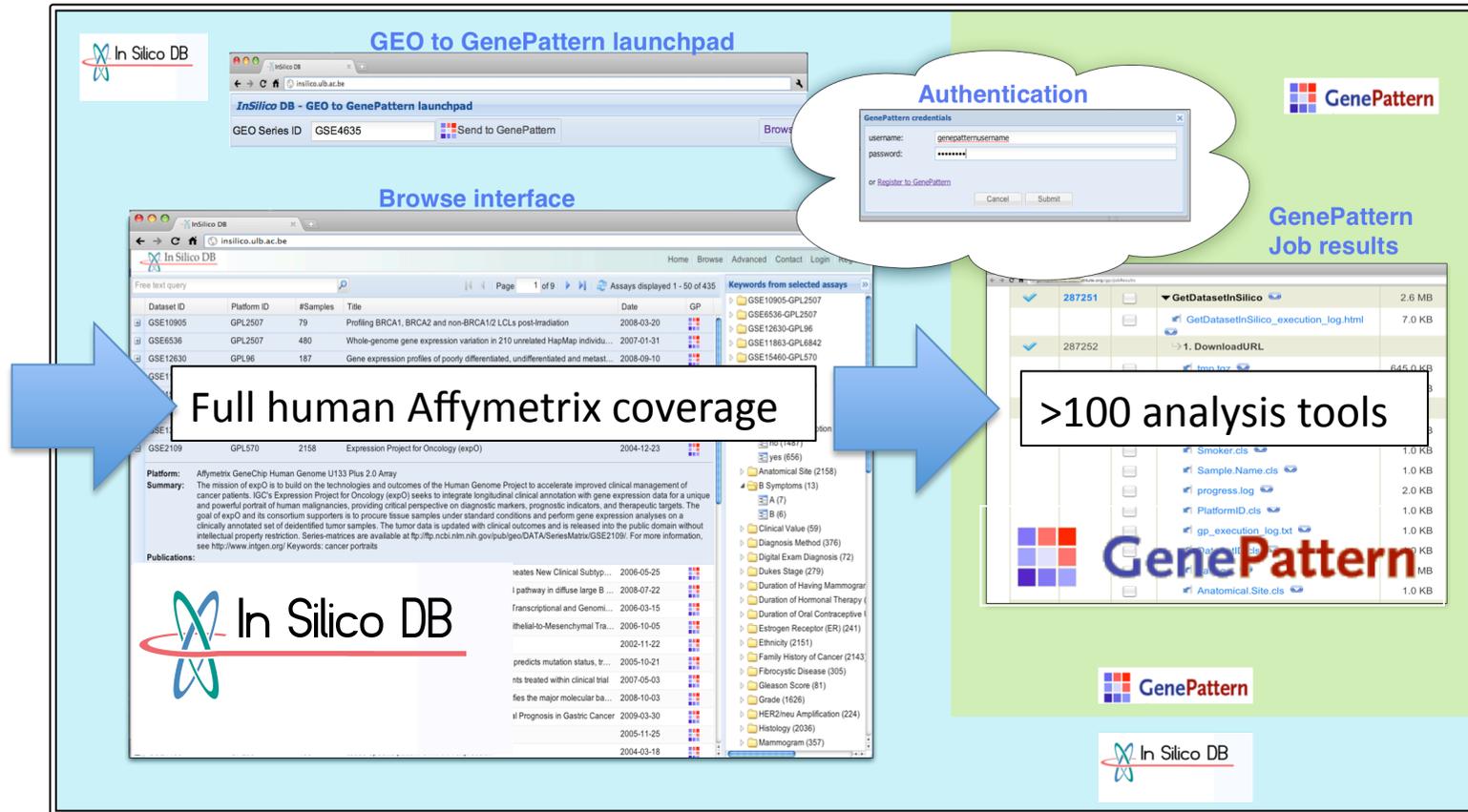
Success stories @ ULB: Biomedical progress through data-warehousing, comparing proprietary to public datasets

- 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, multivariate statistics, linear algebra, molecular biology

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

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

<http://insilico.ulb.ac.be>



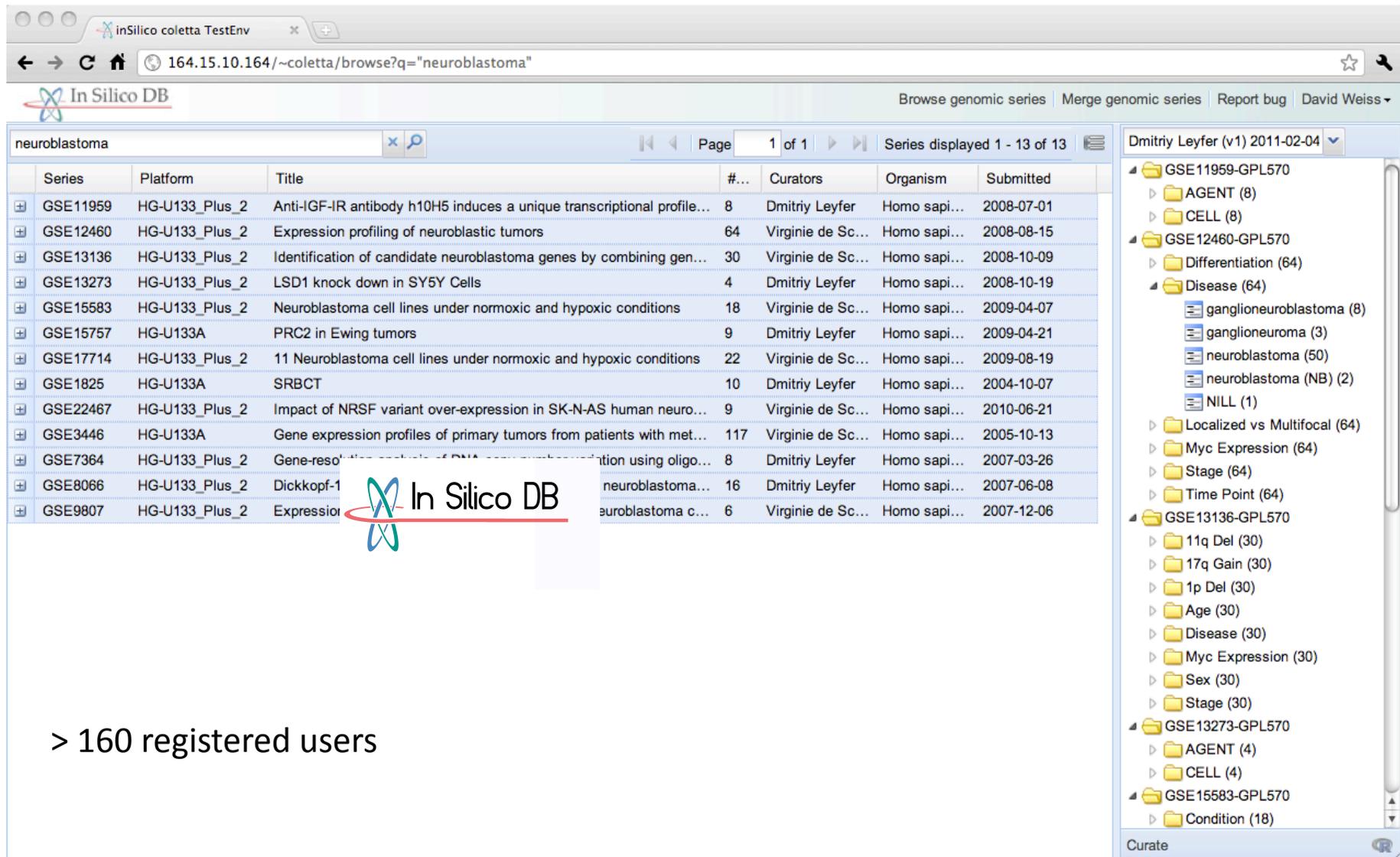
Private

160 registered users

InSilico DB web-app^{1/3}: Database browsing

<http://insilico.ulb.ac.be/browse>

- Biological samples information available in multiple versions



The screenshot shows the InSilico DB web application interface. The browser address bar displays the URL `164.15.10.164/~coletta/browse?q="neuroblastoma"`. The page title is "In Silico DB". The search results are displayed in a table with the following columns: Series, Platform, Title, #..., Curators, Organism, and Submitted. The table lists 13 series related to neuroblastoma. A sidebar on the right shows a hierarchical view of the database structure, including folders for "Dmitriy Leyfer (v1) 2011-02-04" and "GSE11959-GPL570".

Series	Platform	Title	#...	Curators	Organism	Submitted
GSE11959	HG-U133_Plus_2	Anti-IGF-IR antibody h10H5 induces a unique transcriptional profile...	8	Dmitriy Leyfer	Homo sapi...	2008-07-01
GSE12460	HG-U133_Plus_2	Expression profiling of neuroblastic tumors	64	Virginie de Sc...	Homo sapi...	2008-08-15
GSE13136	HG-U133_Plus_2	Identification of candidate neuroblastoma genes by combining gen...	30	Virginie de Sc...	Homo sapi...	2008-10-09
GSE13273	HG-U133_Plus_2	LSD1 knock down in SY5Y Cells	4	Dmitriy Leyfer	Homo sapi...	2008-10-19
GSE15583	HG-U133_Plus_2	Neuroblastoma cell lines under normoxic and hypoxic conditions	18	Virginie de Sc...	Homo sapi...	2009-04-07
GSE15757	HG-U133A	PRC2 in Ewing tumors	9	Dmitriy Leyfer	Homo sapi...	2009-04-21
GSE17714	HG-U133_Plus_2	11 Neuroblastoma cell lines under normoxic and hypoxic conditions	22	Virginie de Sc...	Homo sapi...	2009-08-19
GSE1825	HG-U133A	SRBCT	10	Dmitriy Leyfer	Homo sapi...	2004-10-07
GSE22467	HG-U133_Plus_2	Impact of NRSF variant over-expression in SK-N-AS human neuro...	9	Virginie de Sc...	Homo sapi...	2010-06-21
GSE3446	HG-U133A	Gene expression profiles of primary tumors from patients with met...	117	Virginie de Sc...	Homo sapi...	2005-10-13
GSE7364	HG-U133_Plus_2	Gene-resolution analysis of DNA copy number variation using oligo...	8	Dmitriy Leyfer	Homo sapi...	2007-03-26
GSE8066	HG-U133_Plus_2	Dickkopf-1	16	Dmitriy Leyfer	Homo sapi...	2007-06-08
GSE9807	HG-U133_Plus_2	Expression	6	Virginie de Sc...	Homo sapi...	2007-12-06

> 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

Curations from selected series

- ▲ MIXED
 - ▷ Platform
 - ▷ CELL
 - ▷ AGENT
 - ▲ Disease
 - ≡ NILL(148)
 - ≡ neuroblastoma(81)
 - ≡ ganglioneuroma(3)
 - ≡ ganglioneuroblastoma(8)
 - ≡ neuroblastoma (NB)(2)
 - ≡ primary neuroblastoma(29)
 - ≡ Sample 27 primary neuroblastoma(1)
 - ▷ Time Point
 - ▷ Myc Expression
 - ▷ Stage
 - ▷ Differentiation
 - ▲ Localized vs Multifocal
 - ≡ NILL(266)
 - ≡ localized(5)
 - ≡ multifocal(1)
 - ▷ Age
 - ▷ Sex
 - ▲ 17q Gain
 - ≡ NILL(242)
 - ≡ no(11)
 - ≡ yes(17)
 - ≡ mixed(2)
 - ▲ 11q Del
 - ≡ NILL(242)
 - ≡ no(13)
 - ≡ not known(10)
 - ≡ yes(7)
 - ▲ 1p Del
 - ≡ NILL(242)
 - ≡ no(18)



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

GenePattern
Curated biological samples information

Excel

Integr. Gen. Viewer

R/Bioconductor

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Or, your favourite viewer/analyser

Smoker

	No	Yes
angiogenin, ribonuclease, RNase A	Red	Yellow
melanoma antigen family D, 1	Red	Yellow
NAD(P)H dehydrogenase	Blue	Red
CAP, adenylate cyclase	Blue	Red
WAP four-disulfide core domain	Blue	Red
histone cluster 1	Blue	Red
homogentisate 1,2-oxygenase	Blue	Red
pirin (iron-binding)	Blue	Red
family with sequence homology to	Blue	Red
claudin 10	Blue	Red
transcobalamin II	Blue	Red
carboxyl reductase	Blue	Red
olase domain	Blue	Red
-related	Blue	Red
Inha (alpha)	Blue	Red
high gene	Blue	Red

Job Results table:

Status	Job	delete	Module Name
✓	229454	<input type="checkbox"/>	GetDatasetInSilico
✓	229455	<input type="checkbox"/>	1. DownloadURL
✓	229456	<input type="checkbox"/>	2. untar

Excel table:

NUMBER	SUB-ARR/GENE	log ratio	log ratio	log ratio
1	1	1	0.13685	1.196416
2	1	1	-0.02827	0.753543
3	1	1	removed	removed
4	1	1	removed	removed
5	1	1	removed	removed
6	1	1	removed	removed

IGV table:

DATA FILE	DATA TYPE	LINKING_ID	NAME	PARTICIPANT_ID	SAMPLE_ID	T/N	TUMOR TYPE
550001402...07011.A05							
550001402...07011.A05							
550001402...07011.A07							
550001402...07011.A07							
5500014026...607011.B05							
5500014026...607011.B05							
5500014026...607011.B07							
5500014026...607011.B07							
550001402...07011.C05							

Programming organization

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

Programming environment

- LAMP infrastructure
 - Linux
 - Apache
 - MySQL
 - PHP
- Scripting languages
- Web application framework
- JavaScript library
- Statistical programming: R/Bioconductor packages for genomics data analysis

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

Thank you