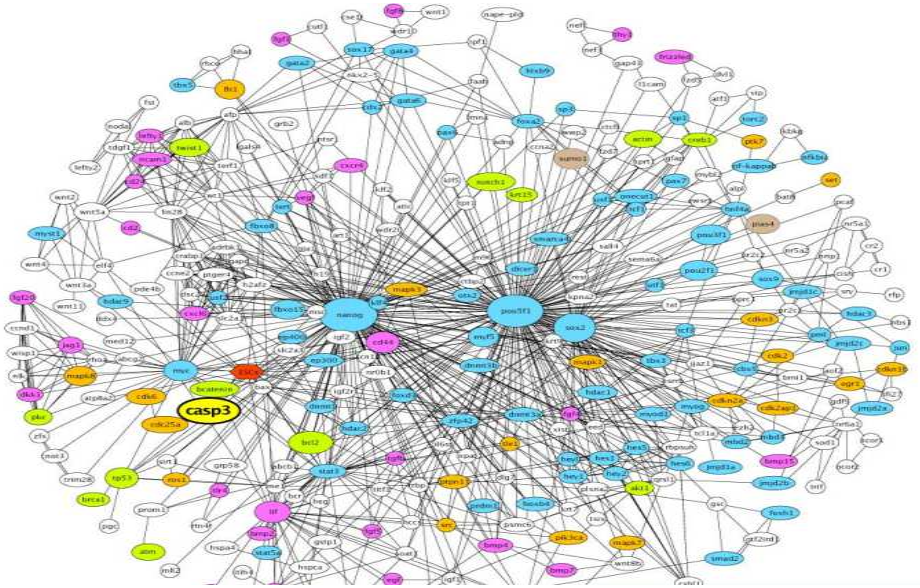
Un (bref) aperçu des méthodes et outils de fouilles et de visualisation de données « omics »



Workshop « Protéomique & Maladies rares »

25th September 2012, Paris

yves.vandenbrouck@cea.fr

CEA Grenoble – iRTSV – BGE (U1038)

iRTSV- Biologie à Grande Echelle

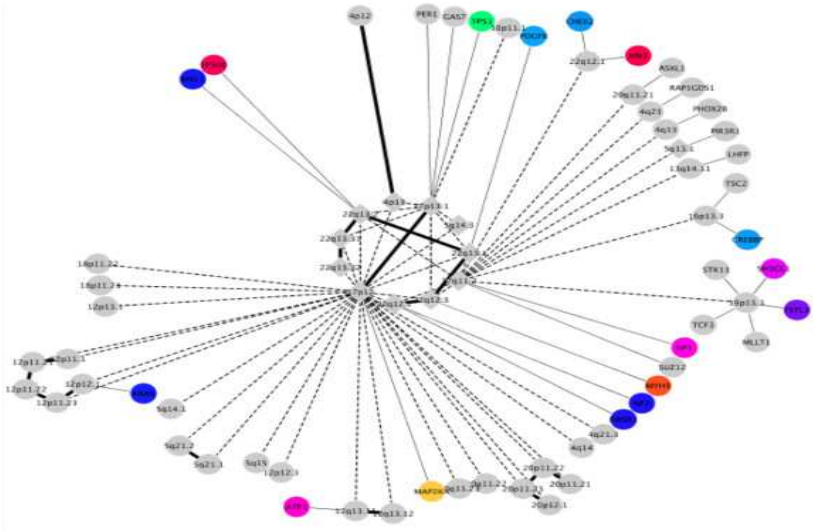
Inserm

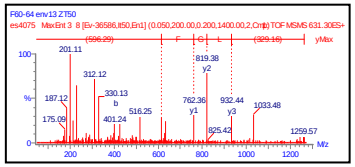
UNIVERSITE JOSEPH FOURIER

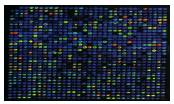
cea

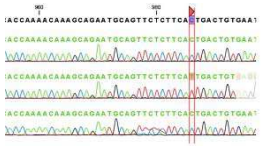
A pressing need for:

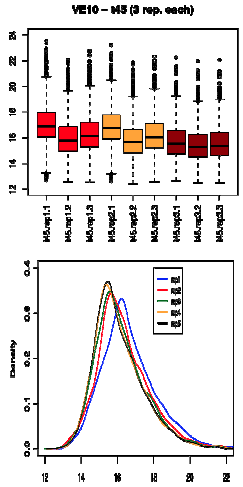
- **analyzing**
- **integration**
- **visualization**
- browsing
- querying
- editing (annotation)

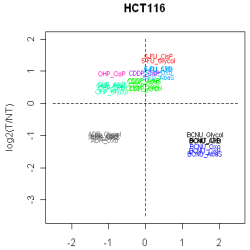


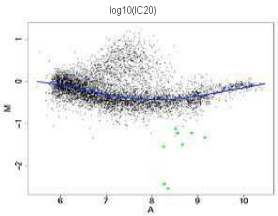


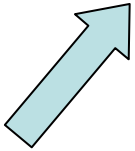












Data mining: more than a field…!

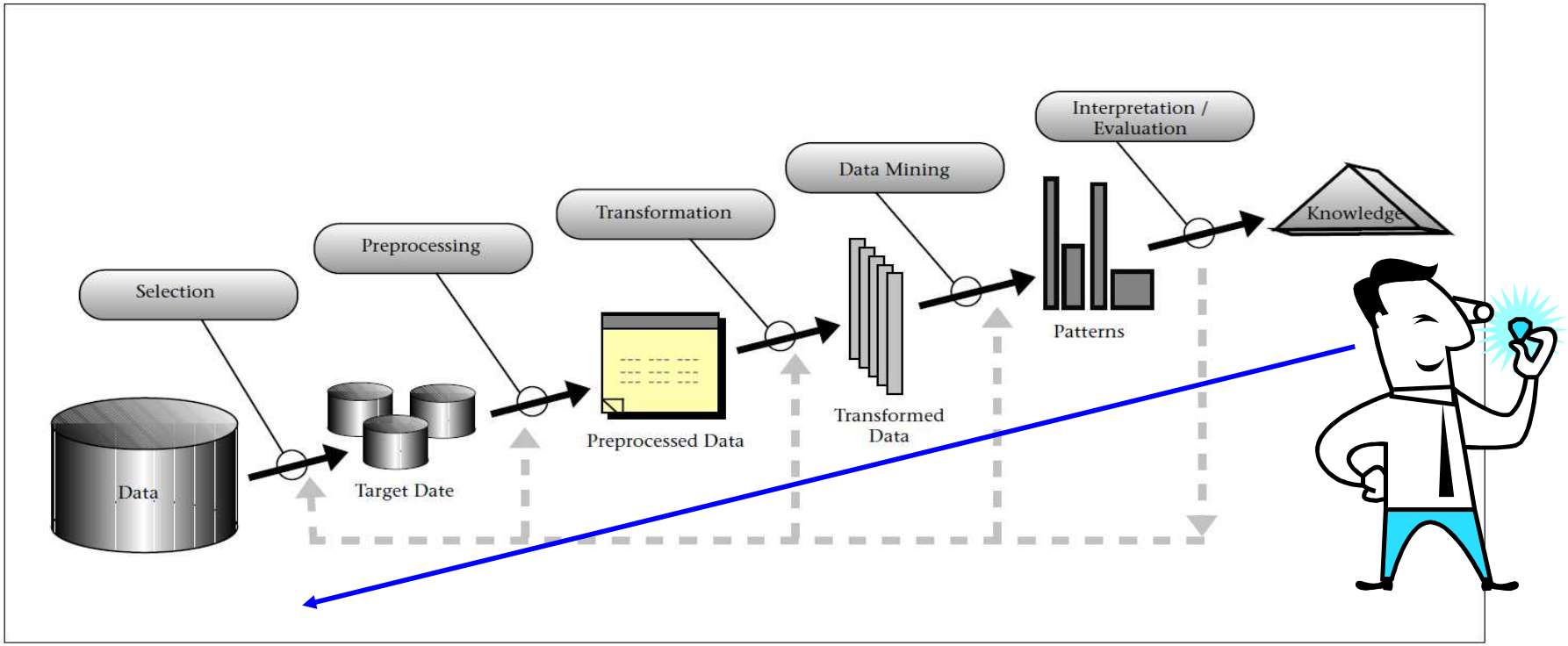
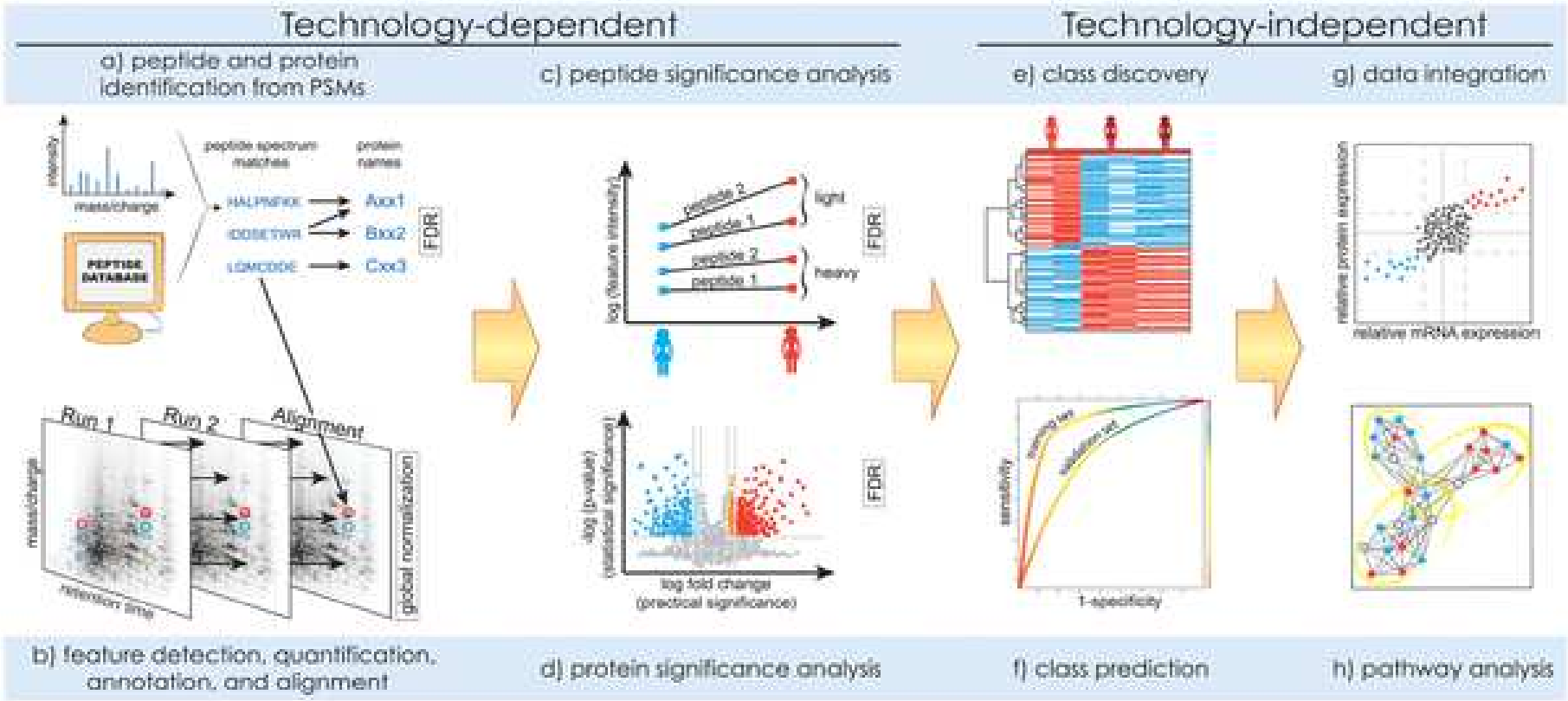


Figure 1. An Overview of the Steps That Compose the KDD Process.

From Fayyad, 1996

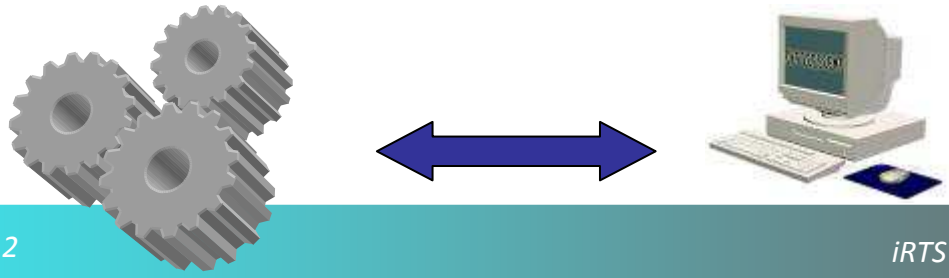
Data mining in the context of MS-based proteomics analysis



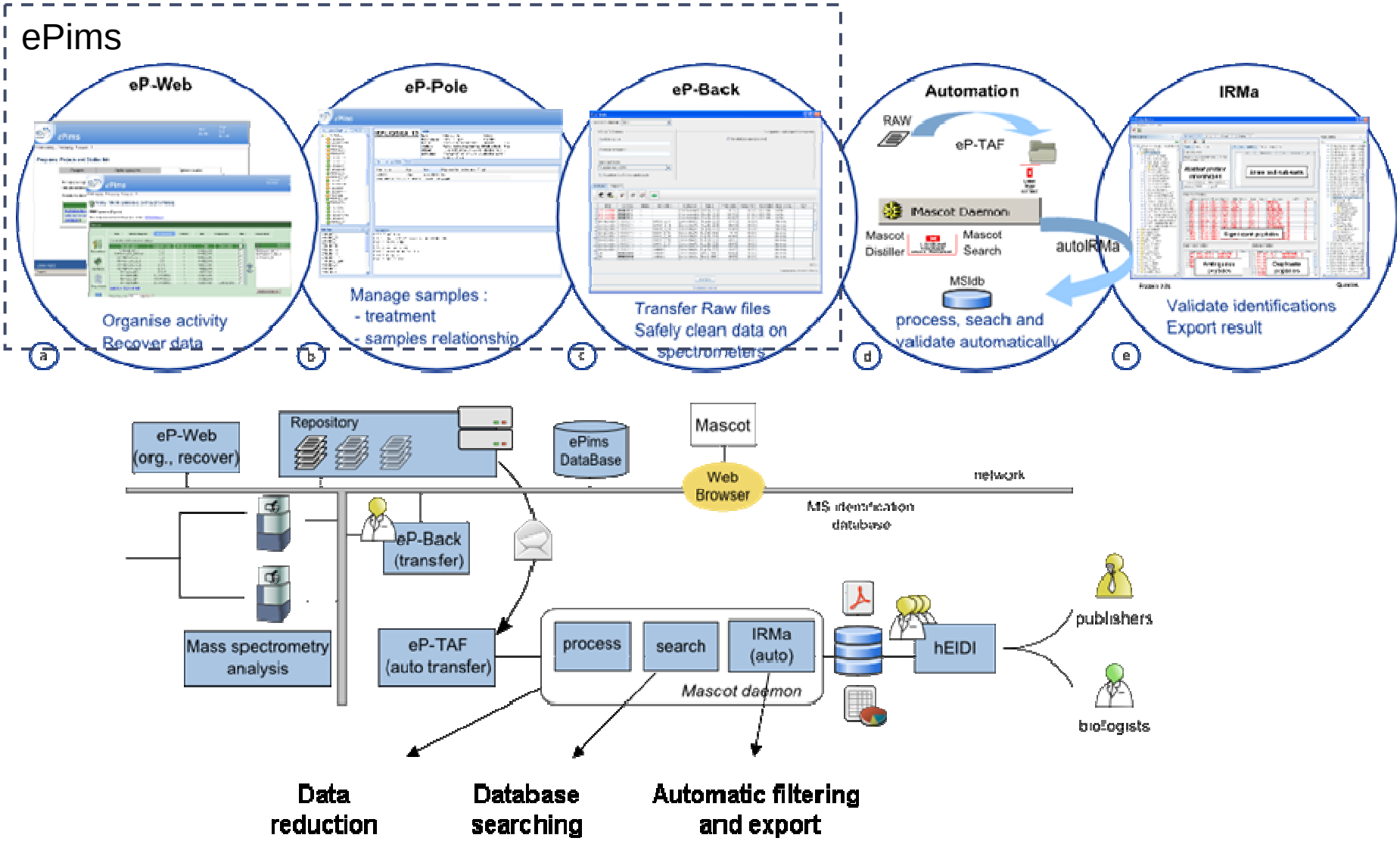
Käll L, Vitek O. 2011 PLoS Comput Biol

Current challenges in software solutions for MS-based quantitative proteomics (Cappadona et al, 2012)

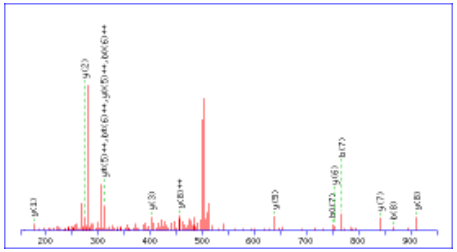
- Software usability
- Data reduction (MS and MS/MS data) => peptide <m/z, Rt, I>
- Feature detection (desiotoping, isobaric interference…)
- Noise rejection (random, chemical, contaminants)
- Retention time alignment (time shifts)
- Peptide identification (FDR, peptide modifications) => search engines and decoy strategies
- Normalization of peptide abundances
- Protein inference (crucial for accurate quantification)
- Protein quantification
- Statistical significance analysis and data mining



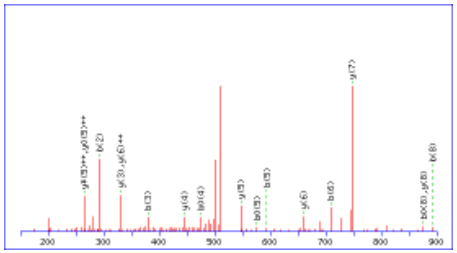
EDyP lab IT setup: workflow and automation



Identification : spectrum – peptide correspondence

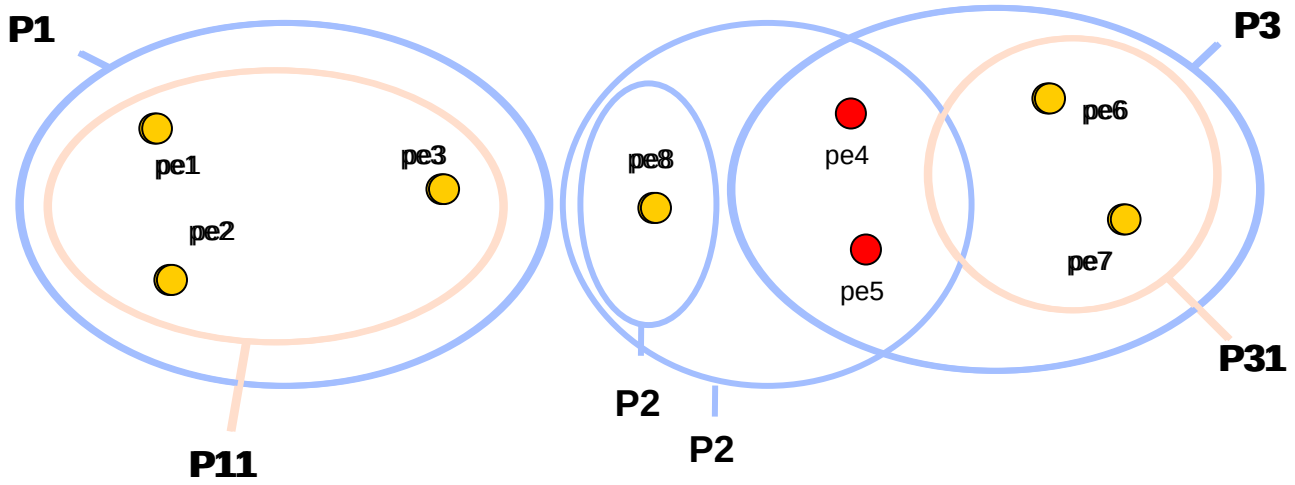


519.79	1037.565448	1039.541016	-1.975568	1	40.12	1	KASDVHEVR
519.79	1037.565448	1039.424011	-1.858563	0	14.92	2	MTTDEGTGGR
519.79	1037.565448	1039.475632	-1.910184	0	14.07	3	EGSFMSLNR
519.79	1037.565448	1035.618881	1.946567	1	10.8	4	RLQHLLEK
519.79	1037.565448	1038.480392	-0.914944	0	10.55	4	FALACNASDK
519.79	1037.565448	1039.52327	-1.957822	0	10.4	4	ADICVHLNR
519.79	1037.565448	1035.618866	1.946582	1	9.94	7	IRGLPPEVR
519.79	1037.565448	1037.623291	-0.057843	1	9.44	8	IISPKVEPR
519.79	1037.565448	1037.525375	0.040073	0	9.3	8	LHPATETGGR
519.79	1037.565448	1038.654922	-1.089474	2	8.91	10	KLVDKAPLR



YESLTDPSK
AMKYLASKK
QNHLASLEK
QYSSPAGDSK
GFTGMLQSR
FCKDVVSNK
MKCLGQSKK
HLKEGARTK
YTNLMRPK
ASEGSDSGSDK

KPYM_HUMAN
Q5T9W7_HUMAN
Q5CAQ6_HUMAN
ENOA_HUMAN
Q53G99_HUMAN
PLSL_HUMAN
TBB2_HUMAN
PGK1_HUMAN
G3P2_HUMAN
PDIA1_HUMAN
TBA6_HUMAN

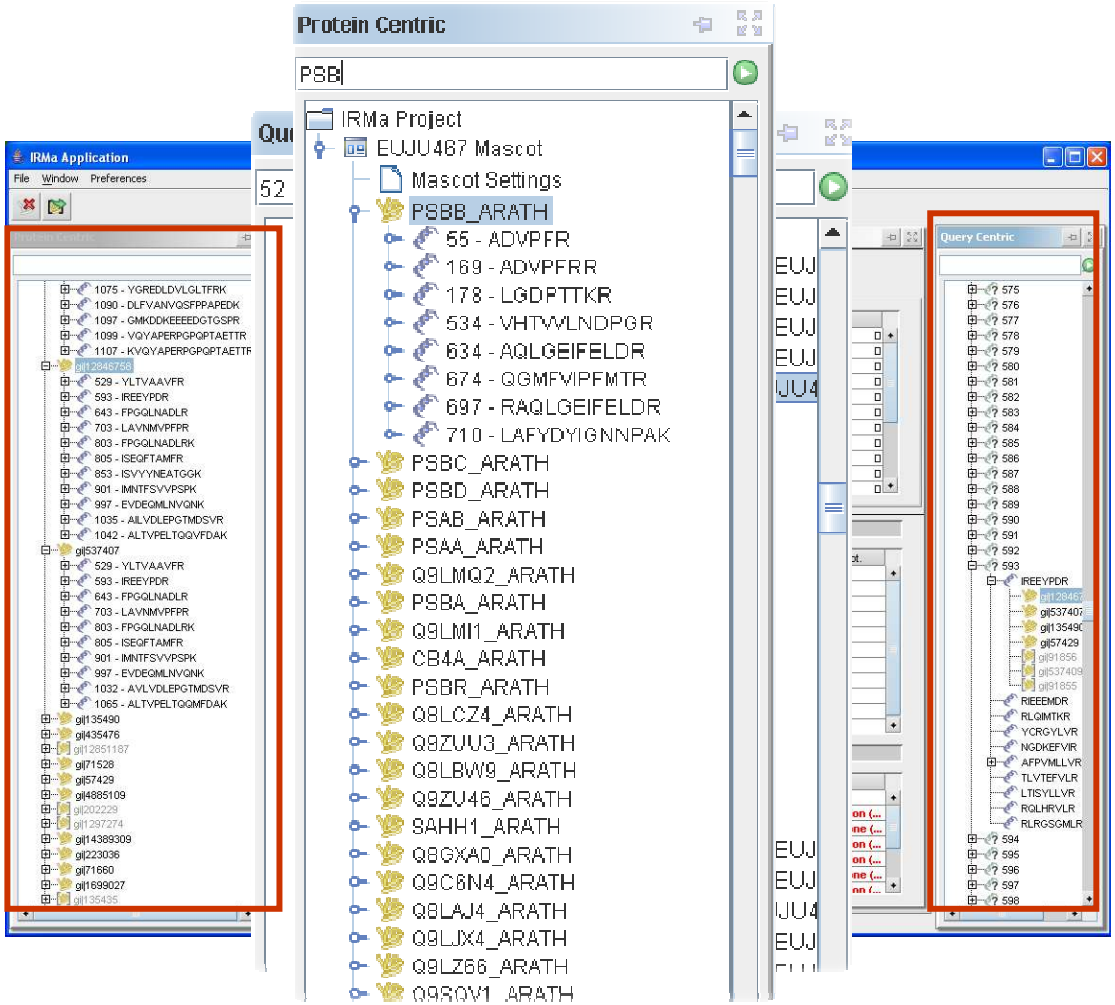


From peptides identification to protein validation: IRMa

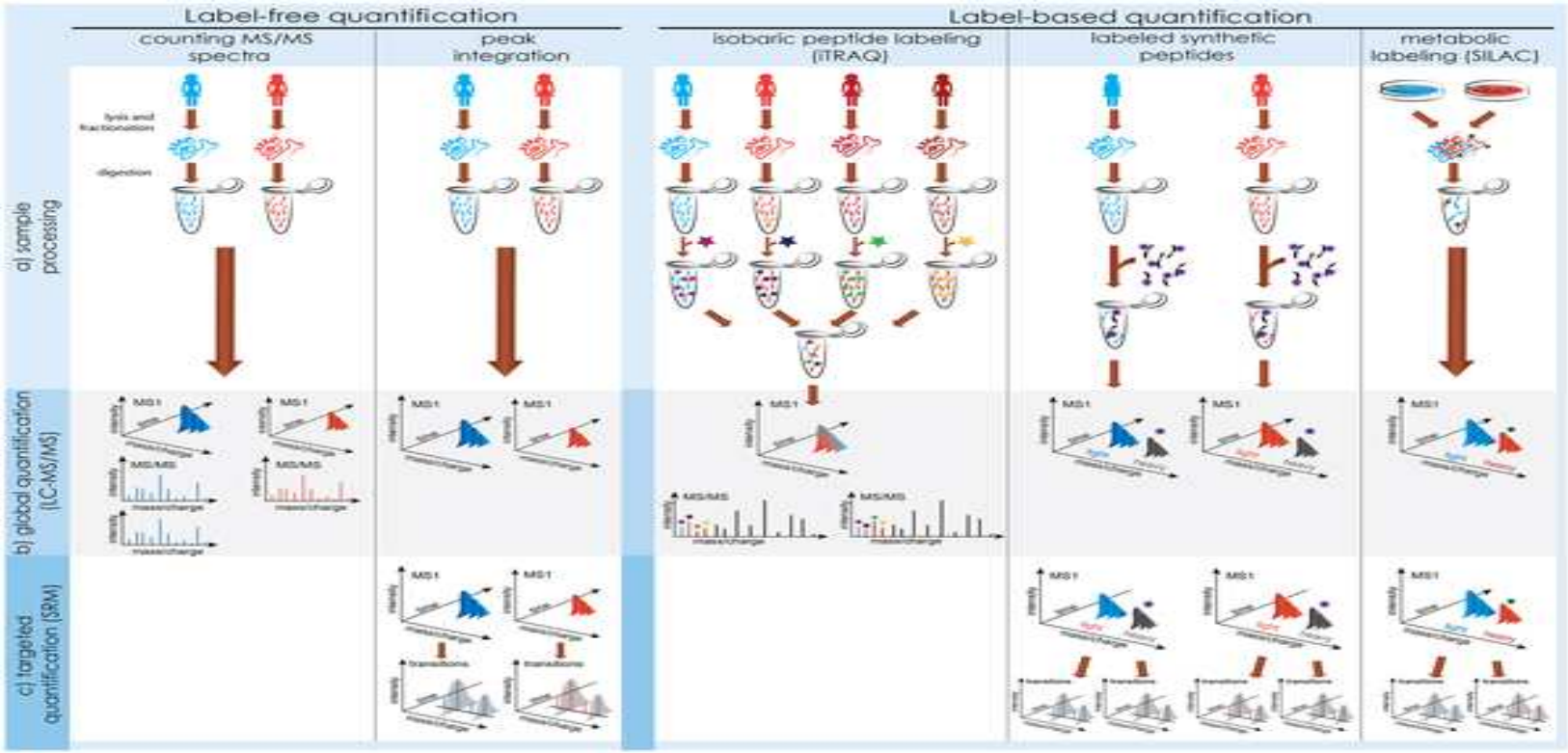
- **Filtering rules**
  - **Score/rank**
  - **Mass tolerance**
- **Peptides groups**
  - **Significant, duplicated, ambiguous**
- **Expertise**
  - **Result visualisation**
  - **Spectra list**
  - **Proteins list**
  - **spectrum/peptide corr.**
- **FDR computation**
- **Open-source software**

  http://www.grenoble.prabi.fr/protehome

Dupierris et al. Bioinformatics, 2009



Mass spectrometry–based measurements



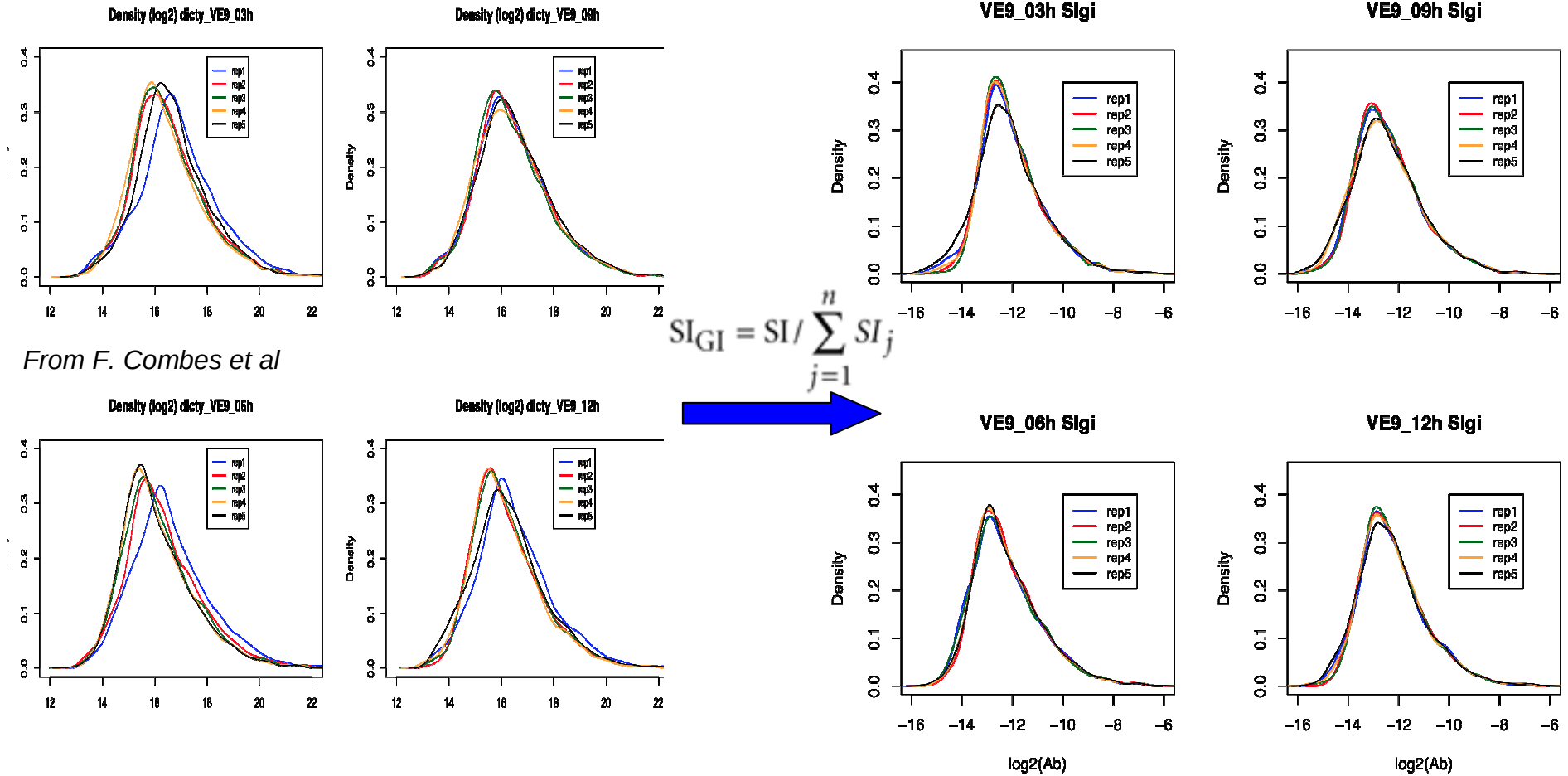
Käll L, Vitek O. 2011 PLoS Comput Biol

Spectral counting (# of MS/MS identifications assigned to protein)
vs.
intensity-based (summarizing signals from spectral peaks)

A diverse range of methods has been proposed and there is no single generally accepted procedure today. => Depending on your sample & your biological question, expertise is strongly required

Data preprocessing for quantification: role of normalization

Changes in relative peptides abundance may reflect not just true biological differences but also systematic bias and random noise



From F. Combes et al

Samples-based (Global normalization) or control-based (e.g. spiked-in internal standards)

Peptide and protein significance analysis:

Statistical analysis and experimental design are closely intertwined!

	2 groups	>2 groups
Independant samples	t-test Mann-Whitney	ANOVA Kruskall-Wallis
Non-independant (paired samples)	t-student Wilcoxon	ANOVA Friedman

ANOVA Assumptions and good practice:

- Independence of samples
- Normality (not essential; transformation)
- Variance homogeneity (=homoscedasticity); essential but rarely justified
- ANOVA does not maintain a significance level (p-val) against comparisons in multiple dimensions => p-value adjustements are required (e.g. FDR)

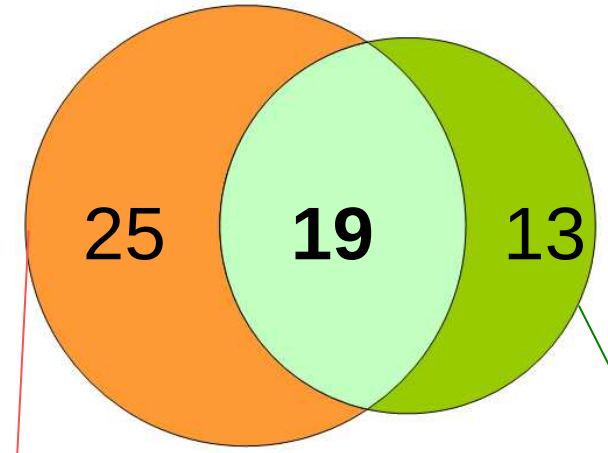
The DECanBlo project: bladder cancer candidates biomarkers discovery (FP7 EU funded coord. J. Garin)

98 patients in discovery sub-cohort
Label-free LC-MS analysis

Association of 2 statistical methods (peptide and protein-centric)

Healthy *vs.* controls

C. Masselon et al., HUPO 2011



Spectral Index permutations

Fu et al., 2008, JPR

Mixed-ANOVA

Clough et al., 2009, JPR

Data mining : various methods

Descriptive (unsupervised) => discovery

- factorial analysis, automatic classification (clustering), association rules…

Predictive (supervised) => inference

- qualitative variable

  discriminant analysis (logistic regression), decision trees, neural networks…

- quantitative variable

  linear regression (simple & multiple), ANOVA, MANOVA, ANCOVA, GLM, decision tree…

How to choose the right method?
Two criteria: accuracy - robustness

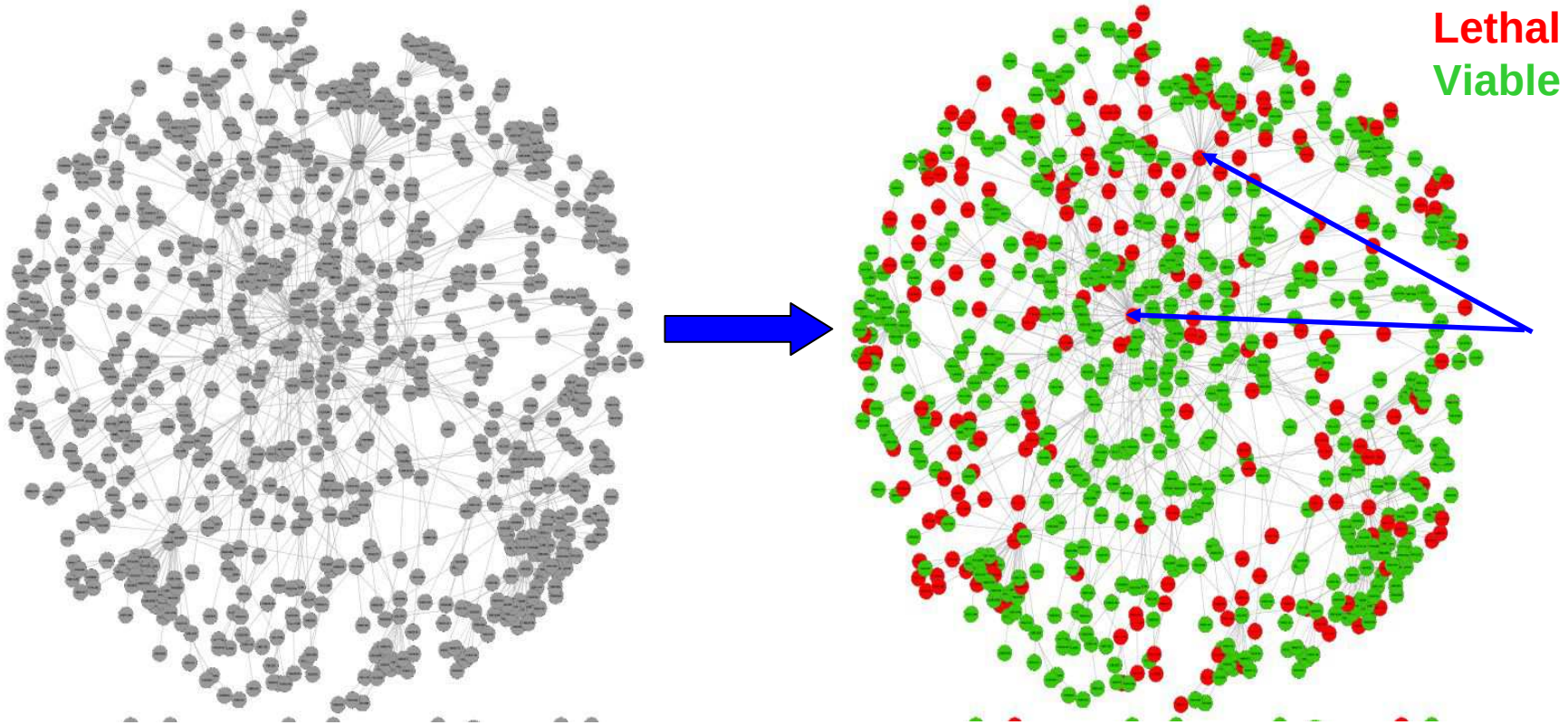
The challenge of defining « valid » proteomic biomarkers

- Is the change (frequency or abundance) of a certain molecule observed in a proteomics study of disease, the result of the disease, or does it merely reflect an artefact due to technical variability in the pre-analytical steps or in the analysis, biological variability, or bias introduced in the study (e.g. due to lifestyle, age, and gender)?
- How should we estimate the number of samples required for the definition of likely valid biomarkers?
- Which algorithms can be employed to combine biomarkers into a multi-marker classifier, and how can the validity of a multi-marker classifier be assessed? Is validation in an independent test set necessary?

Dakna et al, 2010 BMC bioinformatics

Data integration and visualization

Protein → phenotype: are essential proteins more connected ?



Jeong et al, 2001, Nature revisited by C. Hermann (TAGC, Luminy, FR)

Network analysis platform

TABLE 1 | Comparison of network analysis platforms.

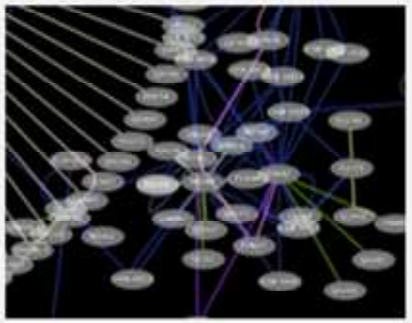
Feature	CY	GM	VA	OS	CD	AR	IN	GG	PI	PR	BL	PA
Free for academic use	√	√	√	√	√				√	√	√	√
Free for commercial use	√	√	√		√				√	√	√	
Open source	√	√							√	√	√	
Curated pathway/network content		√		√		√	√	√				
Standard file format support	√		√		√				√	√		√
User-defined networks/pathways	√	√	√	√	√	√	√	√	√	√	√	√
Functionality to infer new pathways	√		√			√		√	√			
GO/pathway enrichment analysis	√	√	√				√	√				
Automated graph layout	√		√	√	√	√	√	√		√	√	√
Complex criteria for visual properties		√				√	√	√		√	√	√
Multiple visual styles	√		√	√		√	√			√		
Advanced node selection	√		√	√		√	√	√	√	√	√	√
Customizable gene/protein database		√	√			√		√	√			
Rich graphical annotation		√	√				√	√				√
Statistical network analysis	√		√				√	√	√		√	
Extensible functionality: plugins or API	√		√		√	√	√	√	√			
Quantitative pathway simulation					√	√						

CY, Cytoscape[31]; GM, GenMAPP[26]; VA, VisANT[24]; OS, Osprey[23] (http://biodata.mshri.on.ca/osprey/); CD, CellDesigner[25]; AR, Ariadne Genomics Pathway Studio; IN, Ingenuity Pathways Analysis; GG, GeneGo; PI, PIANA (http://sbi.imim.es/piana/); PR, ProViz (http://cbi.labri.fr/eng/proviz.htm); BL, BioLayout; PA, PATIKA.

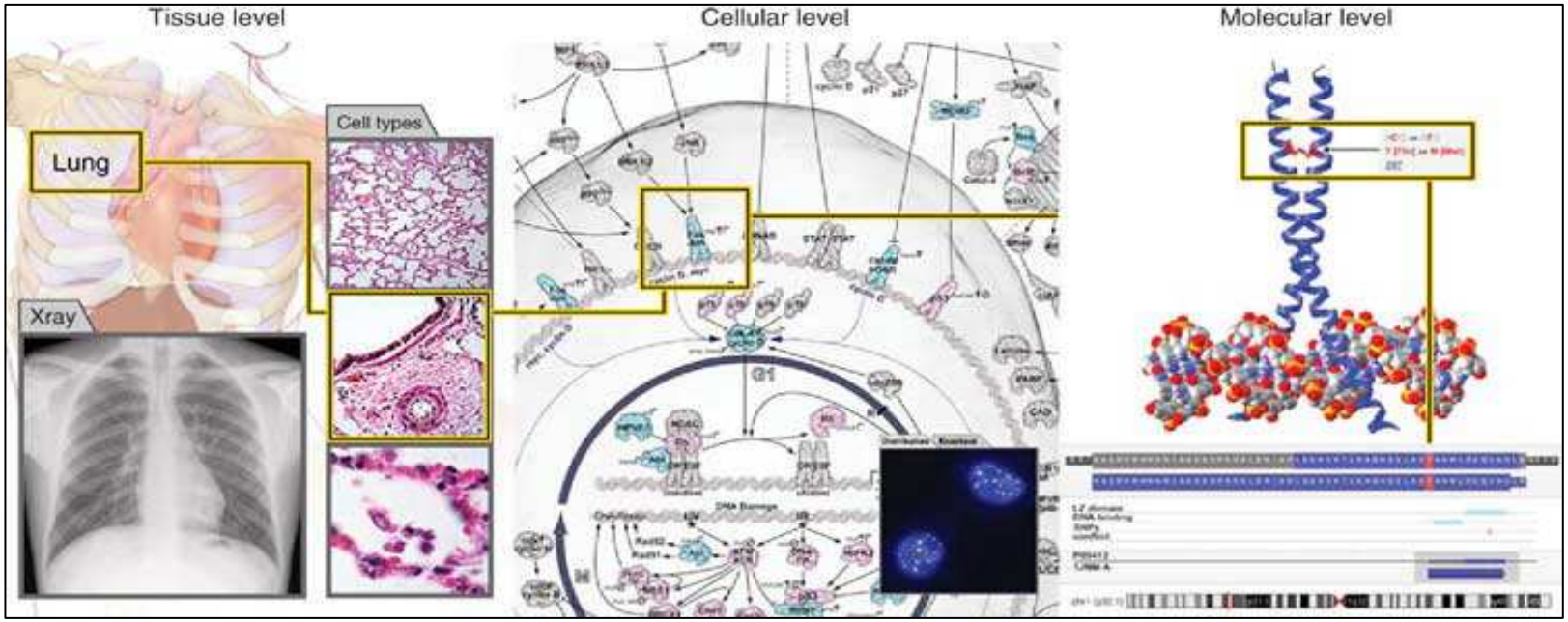
Cline et al, Nature Protocols 2007

What is Cytoscape?

Cytoscape is an open source bioinformatics software platform for ***visualizing*** molecular interaction networks and biological pathways and ***integrating*** these networks with annotations, gene expression profiles and other state data. Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization. Cytoscape *core* distribution provides a basic set of features for data integration and visualization. Additional features are available as ***plugins***. Plugins are available for network and molecular profiling analyses, new layouts, additional file format support, scripting, and connection with databases. Plugins may be developed by anyone using the Cytoscape open API based on Java™ technology and plugin community development is encouraged. Most of the plugins are freely available.



Visualizing biological data—now and in the future: possible integrated visualization environment



O'Donoghue Nat.Meth.S7, S2 - S4 (2010)

Many challenges:

- Large and complex dataset
- Multiscale representation and navigation
- Visual analytics
- Innovative representation…

=> « the revolution in biological data visualization hasn't yet started! »

Take home message

- « A good statistical analysis will not correct a poor experimental design » (O. vitek)

- « Valid proteomic biomarkers for diagnosis and prognosis only can be defined by applying proper statistical data mining procedures ». => “Embedded statisticians” from the beginninng

- Data mining techniques: apply the « simplest model » rule; assessment in an independent dataset is essential (or cross-validation)

- Omics data are largely undermined (unified repositories, knowledge-base, workbench user-friendly, a room for the improvement of methods for optimal design and data mining…)

Acknowledgements:

- EDyP lab (C. Bruley) part of the BGE unit (U1038 – J. Garin)

- IBISA Proteomics platform (ISO9001 certified): yohann.coute@cea.fr
- PROteomics French Infrastructure (PROFI)
- Member of PRIME-XS (FP7, European Proteomics Research Infrastructure)

Further reading and useful links

- Bibliography:

-Cappadona S, Baker PR, Cutillas PR, Heck AJ, van Breukelen B. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Amino Acids. 2012. 43:1087-108.

- Käll L, Vitek O. Computational mass spectrometry-based proteomics. PLoS Comput Biol. 2011. 7:e1002277.

- Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics. 2003. 19:1484-91

- Special issue in Nature Methods, March 2010, Volume 7 No 3s ppS1-S68

- Open-source system for data mining tools:
  - R community: http://cran.r-project.org/
  - Bioconducor: http://www.bioconductor.org/
  - RapidMiner: http://rapid-i.com/content/view/181/196/

- « Omics » data visualization, integration:
  - Cytoscape: http://www.cytoscape.org