Fishing Proteins in an Ocean of ORFs:
Neoblast-specific Functional Screening in *Schmidtea mediterranea*

G. Rodríguez-Esteban E. Fernández-Taboada E. Saló J. F. Abril

Planarian Lab, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona **&** Institute of Biomedicine of the Universitat de Barcelona (IBUB).

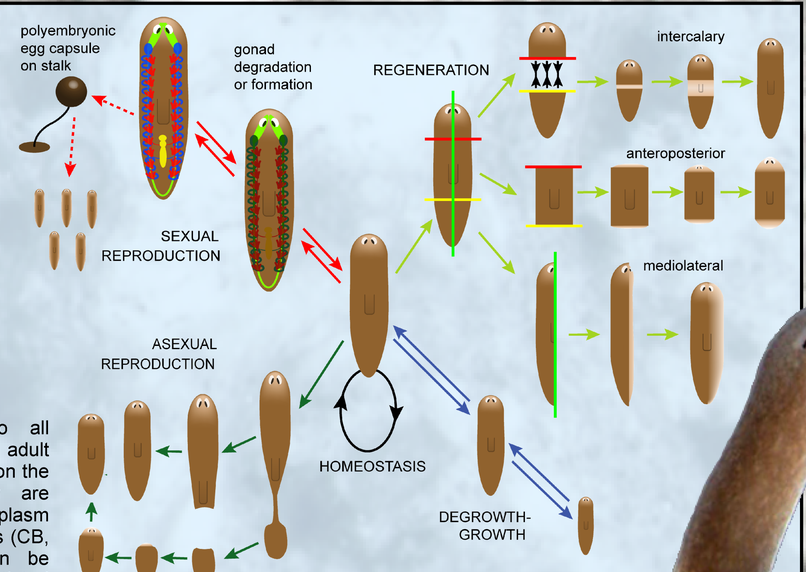
Summary

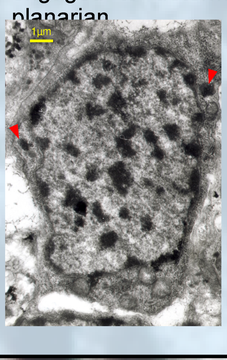
Recent metagenomics projects have proved how far we are of a complete catalog of protein functions yet. Despite having a large number of sequences, current nucleotide and protein databases cannot assist us to find species-specific functional sequences or completely novel undescribed ones. Proteins being expressed under determined experimental conditions can be detected by different proteomic approaches, including mass spectrometry. Using this information to define their genomic locations and to detect novel functions can be challenging, specially when the underlying genome is partially assembled, or not at all.

We have integrated the flatworm *Schmidtea mediterranea* genomic shotgun-traces with mass-spectrometry peptide data, in order to provide sets of putative proteins, containing both known and novel sequences, that can be experimentally validated. The protein set of control and irradiated planarians was compared. Irradiation affects cells that are actively replicating its DNA. Thus, it depletes the animal from neoblasts, the cells involved in cell turnover and regeneration, which are the only proliferating cells in this organism. We aimed to find proteins being expressed differentially at an undifferentiated stage or under a DNA-damage repairing scenario. We discuss here the computational protocol, the results on different datasets of open reading frame sequences, as well as some experimental results validating this approach.

Model Organism

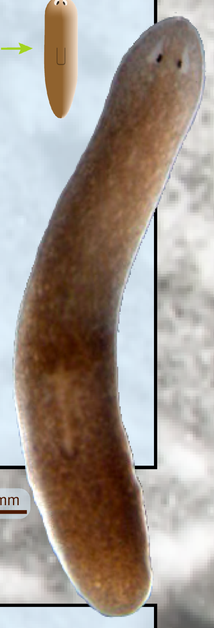
Planarians possess remarkable regenerative capabilities which are summarized on the right Figure. They also have a constant pool of stem cells, about 30% of organism cells, that are named **neoblasts**. Their use in stem cell and regeneration research is both appropriate and lacking in ethical problems. The planarian offers a model in which stem cells can be studied in their natural environment. Furthermore, they are easy to breed in the laboratory and amenable to molecular techniques [HT2008]. A set of techniques involving *in situ* hybridization and RNA interference (RNAi) are available nowadays to locate the expression pattern of a gene. A whole proteomics approach was engaged to determine neoblast-specific functions on planarian.



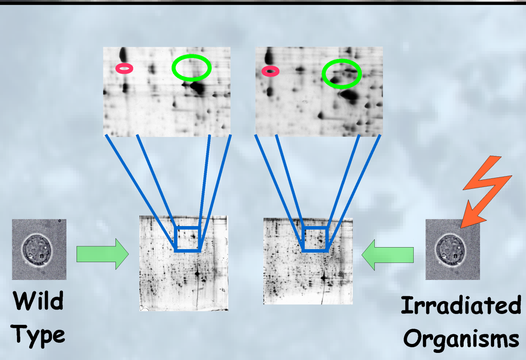


Planarian neoblasts can give rise to all differentiated cell types present in the adult organism [B1989; NSA2000]. As illustrated on the left electron microscope picture, they are characterized by a large nucleus vs cytoplasm ratio and the presence of chromatoid bodies (CB, red arrows). The presence of CB can be considered as a labeling for undifferentiated neoblasts. It seems that CB are progressively lost along the process of differentiation [H2007].
These CB are complexes mainly formed by proteins and RNA, and seem to be somehow independent of transcriptional dynamics, acting as an "expression reservoir".

As neoblast chromatoid bodies contain a pool of silenced transcripts, a proteomics approach seems to be most suitable than other "omics" technologies to capture the set of protein functions that are specifically involved in regeneration processes.

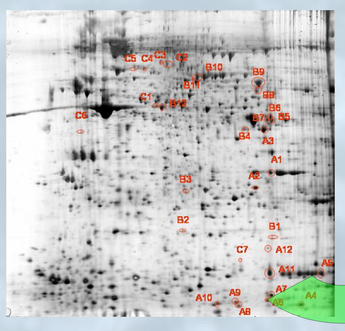


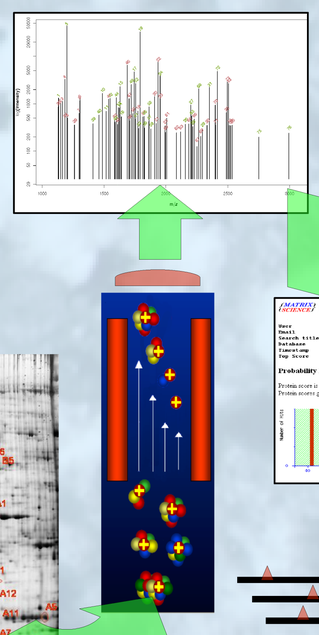
Proteomics & Bioinformatics

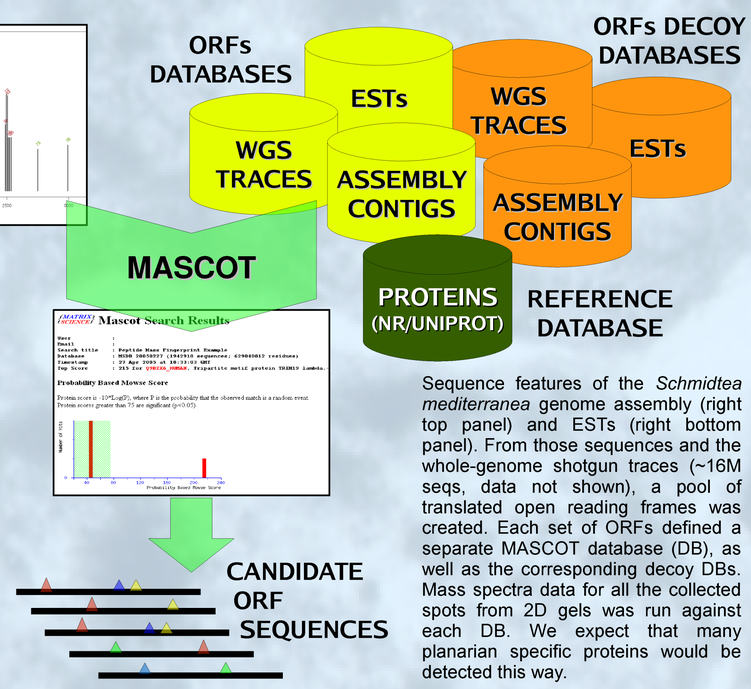


Top: Protein extracts from control animals and irradiated ones (10krads), were analysed using 2-D electrophoretic gels.

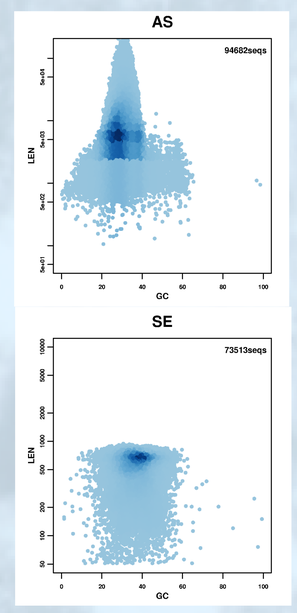
Right: a silver stained 2-D gel of a whole wild type planarian proteome. In red the spots not present in irradiated planarian proteomic profile. Those spots were extracted and analysed by mass spectrometry.
MASCOT was run on the resulting peak lists against a set of different databases.

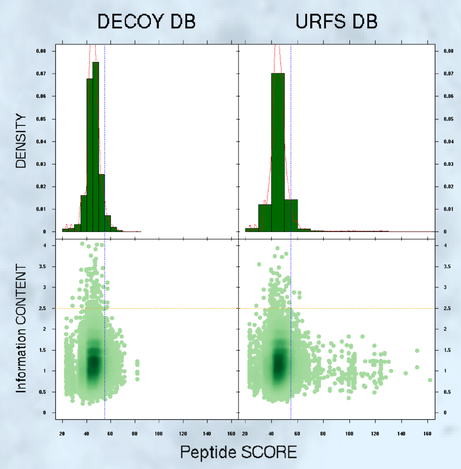






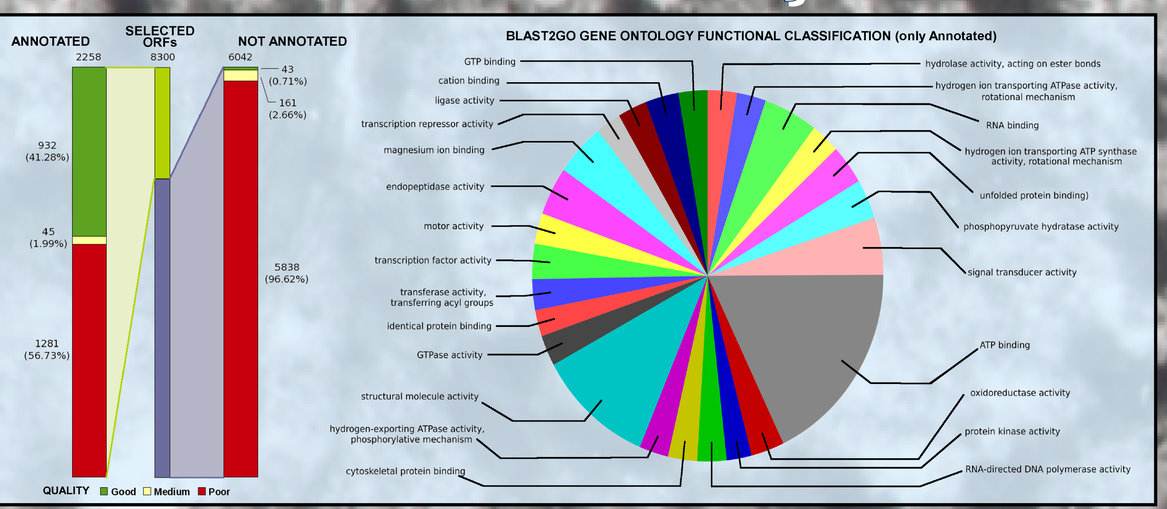
Sequence features of the *Schmidtea mediterranea* genome assembly (right top panel) and ESTs (right bottom panel). From those sequences and the whole-genome shotgun traces (~16M seqs, data not shown), a pool of translated open reading frames was created. Each set of ORFs defined a separate MASCOT database (DB), as well as the corresponding decoy DBs. Mass spectra data for all the collected spots from 2D gels was run against each DB. We expect that many planarian specific proteins would be detected this way.





The chance for making false matches increases when large-sized databases, with millions of protein sequences, are used to detect a wide variety of possible candidate proteins in a sample [R2004]. To assess the significance of the peptide hits found by MASCOT when using our ORF dataset, a decoy database was built by reversing all the ORF sequences [EG2007].

Functional Analysis



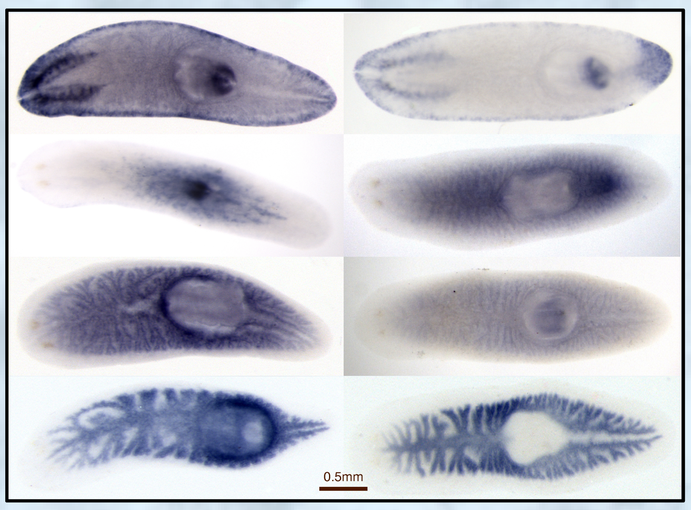
Experimental Follow-up

2H2 type family protein
cephalic ganglia

Unknown protein
parenchyma

MSX-2 interacting nuclear target protein
digestive system

Actin
digestive system

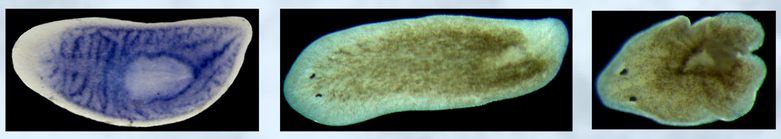


TNF receptor-associated factor 3
cephalic ganglia and tail

TBP-associated factor RNA polymerase 1-like
parenchyma

Homeobox protein six1
parenchyma / digestive

Glutathione S-transferase
digestive system



***Smed*-LSm:** left) *in situ* showing typical neoblast pattern; middle) control tail regenerant at day 10; and right) LSm RNAi inhibits tail regeneration (day 10) [FT2009].

Conclusions

- Planarian is becoming a model organism to understand the molecular mechanism shaping regeneration processes, thanks to a stem cell type known as neoblast. Despite efforts have been focussed to implement trangenesis or cellular cultures, those techniques are not available yet. However, RNAi protocols work well. Transcriptomics and proteomics approaches, like the work presented in this poster, are currently being deployed.
- Our data shows that it is possible to overcome some limitations found on our model species. The two major issues are due to: 1) its intrinsic genomic properties, say here high polymorphism rates, AT rich sequences, many novel repeats, current assembly version too much fragmented; and 2) the amount of sample we can gather due to the small size of the specimens. The later can be an issue to perform tandem-MS, from which we could recover amino acid sequences for each peak.
- Using the translated Open Reading Frames (ORFs), derived from whole genome shotgun data, we have developed a Bioinformatics protocol that allows us to retrieve functional candidate fragments overcoming many of the limitations already mentioned. As those ORFs are linked in our database to the corresponding nucleotide sequences from they are translated, this facilitates posterior functional analyses at different levels: 1) improving localization using more specific primers; 2) mapping the expression of the corresponding mRNAs by *in situ* staining; and 3) facilitating the design of knock-down experiments by RNAi to understand, for instance, the role of a given protein in regeneration.
- At experimental level, shorter ORF sequences, less of 300bp, yield worse results than larger ones when designing probes for *in situ*. As PCR validation of primers was positive for most of them, and the posterior sequencing of the corresponding band returned the same sequence as the one used to design the primers, we are going to assemble the ORFs selected by our protocol when possible; we expect to recover larger protein fragments for those cases.

References & Acknowledgements

[B1989] Baguñà et al. "Effects of activators and antagonists of the neuropeptides substance P and substance K on cell proliferation in planarians", *Int J Dev Biol*, 33, 261-6, 1989.
[NSA2000] Newmark and Sánchez-Alvarado. "Bromodeoxyuridine specifically labels the regenerati-ve stem cells of planarians", *Dev Biol*, 220, 142-53, 2000.
[R2004] Resing et al. "Improving reproducibility and sensitivity in identifying human proteins by shot-gun proteomics", *Anal Chem*, 76, 3556-68, 2004.
[EG2007] Elias and Gygi. "Target decoy search strategy for increased confidence in large-scale pro-tein identifications by mass spectrometry", *Nat Methods*, 4, 207-14, 2007.
[H2007] Higuchi et al. "Characterization and categorization of fluorescence activated cell sorted pla-narian stem cells by ultrastructural analysis", *Dev Growth Differ*, 49, 571-81, 2007.
[HT2008] Handberg-Thorsager et al. "Stem cells and regeneration in planarians", *Frontiers in Biosciences*, 13, 6374-6394, 2008.
[FT2009] Fernández-Taboada et al. "A proteomics approach to unravel the molecular nature of plan-arian stem cells: the Neoblasts", *submitted*, 2009

We would like to thank Dra. Eliandre Oliveira and all components of the proteomic facilities in the Parc Cientific de Barcelona for their help in the proteomic work and analyses.
This work was supported by grants BCM-2002-03992 and BFU-2005-00422 from the Ministerio de Educación y Ciencia (Spain), and grant 2005SGR00769 from AGAUR (Generalitat de Catalunya, Spain). G.R.E. and E.F.T. received an FPI fellowship from the Ministerio de Ciencia e Innovación (formerly known as Ministerio de Ciencia y Cultura).

GOBIERNO DE ESPAÑA · MINISTERIO DE CIENCIA E INNOVACIÓN · Generalitat de Catalunya · UNIVERSITAT DE BARCELONA · I→B=U→B

http://planarian.bio.ub.es
gustavorodriguez@ub.edu