



# EMBnet.journal

Volume 18  
Supplement A  
April 2012

**BITS 2012 - Ninth Annual Meeting  
of the Bioinformatics Italian Society  
Meeting Abstracts  
2-4 May 2012, Catania, Italy**

# Editorial

The Bioinformatics Italian Society (BITS), in collaboration with the Department of Molecular and Clinical Biomedicine and the Department of Mathematics and Computer Science of the University of Catania (Cittadella Universitaria), Italy, held its 9th Annual Meeting from May 2-4, 2012.

The event featured a range of lively and stimulating scientific sessions spanning Next-Generation Sequencing and 'omics disciplines (including genomics, pharmacogenomics, transcriptomics and metagenomics), systems biology and molecular evolution, protein structure and function, algorithms for bioinformatics, biobanks and biological databases.

While the full proceedings of the meeting will appear in BMC Bioinformatics, EMBnet.journal is very pleased to publish abstracts from the major bioinformatics event of the Italian research community.

For this special issue (18.A), the selection of abstracts was overseen by the Conference Scientific Committee, while the layout and logistics were organised by the EMBnet.Journal Editorial Team. For future conferences, our Online Journal System (OJS) can also be used for receiving, archiving and managing the full review process – note that we recently also published the proceedings of the First Scientific Meeting of the COST Action, SeqAhead (issue 17.B). We therefore welcome other societies networks to consider doing the same, and encourage interested parties to contact members of the Editorial Board.

*EMBnet.journal Editorial Board*

Cover picture: A spectacular eruption, Mt. Etna volcano, Catania, Sicily, 2012. [© Andrea Rapisarda, <http://rapis60.redbubble.com/>]

# Contents

Editorial .....	2
The Bioinformatics Italian Society .....	3
Preface - BITS 2012: Ninth Annual Meeting of the Bioinformatics Italian Society .....	4
Conference Programme .....	8
Keynote Lectures.....	12
Oral Presentations .....	16
Posters.....	63

## EMBnet.journal Executive Editorial Board

**Erik Bongcam-Rudloff**, Department of Animal Breeding and Genetics, SLU, SE,  
[erik.bongcam@slu.se](mailto:erik.bongcam@slu.se)

**Teresa K. Attwood**, Faculty of Life Sciences and School of Computer Sciences, University of Manchester, UK,  
[teresa.k.attwood@manchester.ac.uk](mailto:teresa.k.attwood@manchester.ac.uk)

**Domenica D'Elia**, Institute for Biomedical Technologies, CNR, Bari, IT,  
[domenica.delia@ba.itb.cnr.it](mailto:domenica.delia@ba.itb.cnr.it)

**Andreas Gisel**, Institute for Biomedical Technologies, CNR, Bari, IT,  
[andreas.gisel@ba.itb.cnr.it](mailto:andreas.gisel@ba.itb.cnr.it)

**Laurent Falquet**, Swiss Institute of Bioinformatics, Génopode, Lausanne, CH,  
[Laurent.Falquet@isb-sib.ch](mailto:Laurent.Falquet@isb-sib.ch)

**Pedro Fernandes**, Instituto Gulbenkian. PT,  
[pfern@igc.gulbenkian.pt](mailto:pfern@igc.gulbenkian.pt)

**Lubos Klucar**, Institute of Molecular Biology, SAS Bratislava, SK,  
[klucar@EMBnet.sk](mailto:klucar@EMBnet.sk)

**Martin Norling**, Swedish University of Agriculture, SLU, Uppsala, SE,  
[martin.norling@slu.se](mailto:martin.norling@slu.se)

## The Bioinformatics Italian Society



**Manuela Helmer-Citterich<sup>1</sup>, Paolo Romano<sup>2</sup>**✉

<sup>1</sup>Department of Biology, University of Tor Vergata, Rome, Italy

<sup>2</sup>IRCCS San Martino University Hospital - IST National Cancer Research Institute, Genova, Italy

The Bioinformatics Italian Society (BITS) is a non-profit scientific association grounded on 19 June 2003, to gather scientists with interests in the field of Bioinformatics, intended as multidisciplinary science studying biological problems at the molecular level by using informatics and computational methods. The Society has now about 230 members and aims at overcoming 250 in 2012.

Founding members were Rita Casadio, Manuela Helmer-Citterich, Giancarlo Mauri, Luciano Milanese, Graziano Pesole, Cecilia Saccone, and Giorgio Valle. The first elected President was Cecilia Saccone who was in charge until April 2006. Graziano Pesole was then elected President in 2006. Finally, Manuela Helmer Citterich, who had been BITS Secretary since the beginning, was elected President in 2010.

The main aim of the Society, which is a Regional group of the International Society for Computational Biology (ISCB), is the fostering of Bioinformatics in Italy. Its activities include the organization of an annual scientific meeting, the maintenance of a web site and of a mailing list for the distribution of news of interest for the involved community of researchers, the coordination of educational initiatives in Italy, from bachelor to PhD degrees, the coordination of research activities among members, and the promotion of the participation of Italian researchers, both senior and junior, to international events and projects of relevance.

The Society has always devoted a special attention to the needs of young researchers. In

2007, it granted the BITS prize for best bioinformatics PhD thesis, and in 2008 it granted the BITS prize for best bioinformatics research paper from a young researcher. Since 2010, also thanks to the "Preparata" Foundation, it is managing Travel Grants for the participation of young researchers in the BITS Annual Meeting. This initiative already allowed some 50 researchers to attend past and current meetings.

The Society is also paying attention to scientific events at an International level. It awarded three travel grants to young Italian researchers wishing to attend the ISMB/ECCB 2011 Joint Conference in Vienna and is now awarding three more grants for attending the ECCB 2012 Conference in Basel.

BITS is offering its patronage and organization support to scientific events of an international interest which are organized by its members. In the last three years, about 10 such events, including workshops, meetings and courses, were sponsored by BITS.

### BITS Steering Committee (in charge until BITS 2012)

#### President

**Manuela Helmer-Citterich**

University of Rome Tor Vergata, Rome

#### Councillors

**Paolo Romano** (Secretary)

IRCCS San Martino IST, Genoa

**Sabino Liuni** (Treasurer)

CNR-Institute for Biomedical Technologies, Bari

**Gianni Cesareni**

University of Rome Tor Vergata, Rome

**Domenica D'Elia**

CNR-Institute for Biomedical Technologies, Bari

**Angelo Facchiano**

CNR-Institute of Food Science, Avellino

**Giorgio Valle**

University of Padua, Padua

### Contact information

Official BITS email address: [bits@bioinformatics.it](mailto:bits@bioinformatics.it)

BITS Web site: <http://www.bioinformatics.it/>

Web manager: [bits\\_info@bioinformatics.it](mailto:bits_info@bioinformatics.it)

## Preface

### BITS 2012: Ninth Annual Meeting of the Bioinformatics Italian Society



**Alfredo Ferro<sup>1</sup>, Rosalba Giugno<sup>1</sup>, Alfredo Pulvirenti<sup>1</sup>, Domenica D'Elia<sup>2✉</sup>, Manuela Helmer-Citterich<sup>3</sup>, Paolo Romano<sup>4</sup>**

<sup>1</sup>Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy

<sup>2</sup>CNR, Institute for Biomedical Technologies, Bari, Italy

<sup>3</sup>Department of Biology, University of Tor Vergata, Rome, Italy

<sup>4</sup>IRCCS San Martino University Hospital - IST National Cancer Research Institute, Genova, Italy

The Annual Meeting of the Bioinformatics Italian Society (BITS) is an important event for scientists and professionals working in the field of Bioinformatics or using the software tools developed within this scientific discipline. The meeting provides a thorough view of how and in which directions the Italian scientific community is investing financial and human resources, and an opportunity to establish and strengthen scientific collaborations.

The first gathering of Italian scientists involved with bioinformatics was held in Rome in 1999. The next three meetings, in 2000, 2001 and 2002, took place in the Certosa di Pontignano, in the neighbourhood of Siena. In the fourth meeting, held again in Rome, the Italian Bioinformatics Society was grounded. Since then, the meetings of Padua (2004), Milan (2005), Bologna (2006), and Naples (2007) were characterized by an increasing number of participants and a growing quality of the works presented. In 2008, the meeting was held in Cagliari, in conjunction with the European

Conference of Computational Biology (ECCB), as a further demonstration of the relevance that it was achieving. In the last three years the meeting has been held in Genoa (2009), Bari (2010), and Pisa (2011).

The Bioinformatics Italian Society, the University of Catania, and the J. T. Schwartz International School for Scientific Research have supported the BITS 2012 meeting. It takes place in Catania, Sicily, from the 2nd of May to the 4th of May. The meeting is kindly hosted by the Department of Mathematics and Computer Science of Catania University. The University of Catania was founded in 1434 and among its famous scholars it had Napoleone Ferrara, molecular biologist, winner of the 2010 Lasker-DeBaakey Clinical Medical Research Award.

The high quality of BITS 2012 is witnessed by the number of applications for oral presentation (45), of which only 33 could be selected and scheduled during the three days of the meeting. Selection of best contributions for oral presentation was carried out by the Scientific Committee after a well established peer-review procedure based on two reviews and scores per paper. Overall, 26 referees were involved in the selection of oral communications, both BITS members and non members.

Accepted contributions, whose authors have released the copyright licence agreement to the EMBnet.journal and are presenting their work at the conference, are published in this EMBnet.journal Supplement. The total number, including oral presentations and posters, was 89, divided in 14 thematic areas: Systems Biology, Genomics, Comparative genomics, Molecular evolution, Metagenomics, Next Generation Sequencing, Transcriptomics, Pharmacogenomics, Proteins structure and function, Proteomics, Algorithms for Bioinformatics, Biobanks, Biological Databases, and Technological track.

The number of contributions in each area, shown in Table 1, provides an interesting perspective of the kind of applications the Italian bioinformatics community is presently most involved in.

We are grateful to all the lecturers and co-authors for their excellent contributions, to the Steering Committee of the Bioinformatics Italian Society, to the referees, and to all the people, Institutes, projects, companies, and local authorities that supported this event and contributed to make it successful.

Moreover, we are especially grateful to the keynote speakers for accepting our invitation to contribute to the meeting: Prof. Charles E. Lawrence, Brown University (US); Prof. Eugene Myers, HHMI Janelia Farm Research Campus (US) and Max Planck Institute for Cellular Molecular Biology and Genetics (DE); Dr. Ileana Zucchi, CNR-Institute for Biomedical Technologies (IT).

Their outstanding, well recognized scientific value and their participation to the meeting confirm the high respect that the Italian Bioinformatics Society receives from the international scientific community.

Table 1: number of oral communications and posters presented at the BITS 2012 meeting grouped by scientific topic. OC=oral contribution; PC=poster contribution

Thematic Areas	OC	PC
Systems Biology	6	5
Genomics	3	3
Comparative genomics	0	3
Molecular evolution	1	2
Metagenomics	1	1
Next Generation Sequencing	5	10
Transcriptomics	2	4
Pharmacogenomics	1	0
Proteins structure and function	1	12
Proteomics	0	1
Algorithms for Bioinformatics	6	7
Biobanks	1	0
Biological Databases	3	7
Technological track	3	0

## Speakers

### Oral presentations

Bicciato S.	Gherardini P.F.
Calogero R.	Liberti S.
Calviello L.	Mazza T.
Canakoglu A.	Micale G.
Cangelosi D.	Pappalardo A.M.
Cannataro M.	Pendino V.
Caselle M.	Pio G.
Castrignanò T.	Pohl M.
Colantoni A.	Pozzoli U.
Cordero F.	Romano C.
D'Antonio M.	Rossi E.
Di Domenico T.	Sinha S.
Ferrarini A.	Stupka E.
Frasca M.	Valentini G.
Gaido L.	Vicedomini R.

### Tutorial

Micale G.

### Keynote Speakers

Charles E. Lawrence  
Eugene W. Myers, G. Preparata Lecture  
Ileana Zucchi, Dulbecco Lecture

## Editors

<b>Domenica</b> <b>Alfredo</b>	<b>D'Elia</b> <b>Ferro</b>	CNR - Institute for Biomedical Technologies, Bari (IT) Department of Clinical and Molecular Biomedicine, University of Catania, Catania (IT)
<b>Rosalba</b>	<b>Giugno</b>	Department of Clinical and Molecular Biomedicine, University of Catania, Catania (IT)
<b>Manuela</b> <b>Alfredo</b>	<b>Helmer-Citterich</b> <b>Pulvirenti</b>	University of Rome Tor Vergata, Rome (IT) Department of Clinical and Molecular Biomedicine, University of Catania, Catania (IT)
<b>Paolo</b>	<b>Romano</b>	IRCCS San Martino IST, Genoa (IT)

## Sponsors and Supporters



[Bioinformatics Italian Society<sup>1</sup>](http://www.bioinformatics.it/)



Associazione per la Fondazione  
Preparata



[J.T. Schwartz International  
School for Scientific Research<sup>2</sup>](http://lipari.dmi.unict.it/)



[University of Catania<sup>3</sup>](http://www.unict.it/)

<sup>1</sup> <http://www.bioinformatics.it/>

<sup>2</sup> <http://lipari.dmi.unict.it/>

<sup>3</sup> <http://www.unict.it/>

## Conference Committees

### Scientific committee

<b>Gianni</b>	<b>Cesareni</b>	University of Rome Tor Vergata, Rome (IT)
<b>Domenica</b>	<b>D'Elia</b>	CNR – Institute for Biomedical Technologies, Bari (IT)
<b>Angelo</b>	<b>Facchiano</b>	CNR - Institute of Food Science, Avellino (IT)
<b>Alfredo</b>	<b>Ferro</b>	University of Catania, Catania (IT)
<b>Rosalba</b>	<b>Giugno</b>	University of Catania, Catania (IT)
<b>Manuela</b>	<b>Helmer-Citterich</b>	University of Rome Tor Vergata, Rome (IT)
<b>Sabino</b>	<b>Liuni</b>	CNR – Institute for Biomedical Technologies, Bari (IT)
<b>Alfredo</b>	<b>Pulvirenti</b>	University of Catania, Catania (IT)
<b>Paolo</b>	<b>Romano</b>	IRCCS San Martino IST, Genoa (IT)
<b>Giorgio</b>	<b>Valle</b>	University of Padua, Padua (IT)

### Organising Committee

<b>Alfredo</b>	<b>Ferro</b>	University of Catania, Catania (IT)
<b>Rosalba</b>	<b>Giugno</b>	University of Catania, Catania (IT)
<b>Alfredo</b>	<b>Pulvirenti</b>	University of Catania, Catania (IT)

## Conference Programme

---

# **BITS 2012 - Ninth Annual Meeting of the Bioinformatics Italian Society**

**2-4 May 2012, Catania, Italy**

## Conference Programme

Wednesday		May 2
13.00	14.00	Registration and poster hang-up
14.00	14.20	Welcome and Introduction
14.20	15.10	<b>Keynote Lecture</b> (Dulbecco Lecture) <i>Dr. Ileana Zucchi</i> <b>Renato Dulbecco: genes, cancer and epigenomics</b>
15.10	16.10	<b>Session 1 (first part): Next Generation Sequencing, Pharmacogenomics</b>
15.10	15.30	<i>Giugno R, Abate F, Bombieri N, Delledonne M, Ferrarini A, Ficarra E, Pulvirenti A, Acquaviva A</i> <b>Integrated cloud environment for characterization of genotype specific transcriptome from next generation sequencing data</b>
15.30	15.50	<i>Pozzoli U</i> <b>Next Generation Programming: software tools for NGS tertiary analysis</b>
15.50	16.10	<i>D'Antonio M, D'Onorio De Meo P, Pesole G, Castrignanò T</i> <b>Building an optimized pipeline for whole-exome sequencing</b>
16.10	16.30	Coffee Break
16.30	17.30	<b>Session 1 (second part): Next Generation Sequencing, Pharmacogenomics</b>
16.30	16.50	<i>Castellana S, Mazza T</i> <b>On the impact of short-reads quality on variants detection</b>
16.50	17.10	<i>Conte I, Migliore C, Merella S, Avellino R, Marco-Ferreres R, Carrella S, Emmett W, Sanges R, Bockett N, Van Heel D, Meroni G, Bovolenta P, Banfi S, Stupka E</i> <b>An RNA-seq-based approach highlights a role of miR-204 in axon guidance in vertebrates</b>
17.10	17.30	<i>Visconti A, Cordero F, Calogero RA</i> <b>Improving biomarker discovering for chemosensitivity prediction using an integrated approach</b>
17.30	20.00	<b>BITS General Assembly</b>
Thursday		May 3
8.30	9.00	Poster hang-up
9.00	9.50	<b>Keynote Lecture</b> (G. Preparata Lecture) <i>Prof. Eugene W. Myers</i> <b>On bioimage informatics and decoding genomes</b>
9.50	10.50	Coffee Break and Poster Session
10.50	12.50	<b>Session 2: System Biology, Molecular Evolution</b>
10.50	11.10	<i>Beka S, te Boekhorst I, Abnizova I</i> <b>The location of T1 diabetes associated SNPs in regulatory regions</b>
11.10	11.30	<i>Calviello L, Stano P, Mavelli F, Luisi PL, Marangoni R</i> <b>Quasi-cellular Systems: stochastic simulation analysis at nanoscale range</b>
11.30	11.50	<i>Cordero F, Gribaudo M, Manini D</i> <b>An analytical spatially-based approach to study cancer cell population evolutions</b>

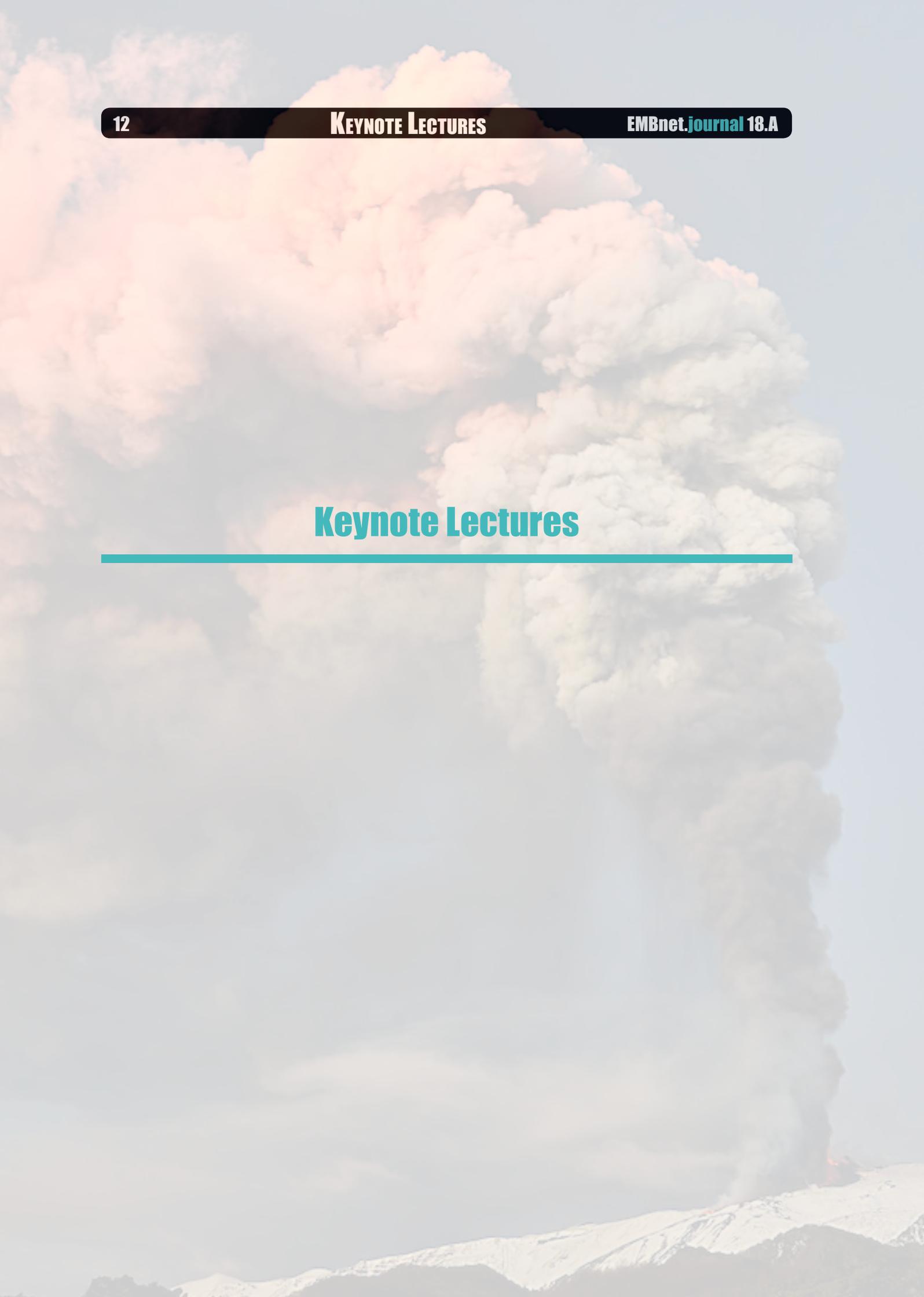
11.50	12.10	<i>Gherardini PE, Sacco F, Paoluzi S, Saez-Rodriguez J, Helmer-Citterich M, Ragnini-Wilson A, Castagnoli L, Cesareni G</i> <b>A combined computational/experimental strategy to map phosphatases on growth pathways</b>
12.10	12.30	<i>Pendino V, Ciccarelli FD</i> <b>microRNAs regulate the dosage of duplicated genes</b>
12.30	12.50	<i>Re M, Mesiti M, Valentini G</i> <b>Drug repositioning through pharmacological spaces integration based on networks projections</b>
12.50	13.10	<i>Guzzi PH, Cannataro M</i> <b>Cyto-Sevis: semantic similarity-based visualisation of protein interaction networks</b>
13.10	14.10	<i>Lunch Break</i>
14.10	16.10	<b>Session 3: Genomics, Transcriptomics, Proteins Structure and Function</b>
14.10	14.30	<i>Carrara M, Calogero R</i> <b>Digging in the RNA-seq garbage: evaluating the characteristics of unmapped RNA-seq reads in normal tissues</b>
14.30	14.50	<i>Colantoni A, Ferrè F, Helmer-Citterich M</i> <b>Alternative splicing as regulator of protein-protein interactions</b>
14.50	15.10	<i>Grützmann K, Szafranski K, Pohl M, Voigt K, Petzold A, Schuster S</i> <b>Fungal alternative splicing associates with higher cellular complexity and virulence</b>
15.10	15.30	<i>Romano C, Buffa P, Pandini A, Massimino M, Tirrò E, Manzella L, Fraternali F, Vigneri P</i> <b>Computational and experimental characterization of critical amino acidic residues in the BCR-ABL kinase domain explaining TKIS resistance in patients with chronic myeloid leukemia</b>
15.30	15.50	<i>Sinha S, Iannelli F, Collino A, Ghisletti S, Natoli G, Ciccarelli FD</i> <b>CNV analysis of inflammation driven hepatocellular carcinoma</b>
15.50	16.10	<i>Testori A, Caizzi L, Cutrupi S, Friard O, De Bortoli M, Corà D, Caselle M</i> <b>The role of transposable elements in shaping the combinatorial interaction of transcription factors</b>
16.10	19.30	<b>Social Tour of Catania</b>
19.30	23.00	<b>Social Dinner</b>

Friday		May 4
8.15	8.30	<i>Poster hang-up</i>
8.30	9.20	<b>Keynote Lecture</b> <i>Prof. Charles E. Lawrence</i> <b>Statistical inference in high dimensional spaces of genomics: an RNA structural example</b>
9.20	10.20	<b>Session 4 (first part): Algorithms for Bioinformatics, Metagenomics</b>
9.20	9.40	<i>Cangelosi D, Muselli M, Blengio F, Versteeg R, Eggert A, Schramm A, Garaventa A, Gambini C, Varesio L</i> <b>Translation of a robust, biology-driven, prognostic classifier of cancer patients outcome into clinically relevant rules</b>

9.40	10.00	<i>Frasca M, Bertoni A, Valentini G</i> <b>Regularized network-based algorithm for predicting gene functions with high-imbalanced data</b>
10.00	10.20	<i>Pio G, Ceci M, D'Elia D, Loglisci C, Malerba D</i> <b>A novel biclustering algorithm for the discovery of meaningful biological correlations between miRNAs and mRNAs</b>
10.20	11.20	<i>Coffee Break and Poster Session</i>
11.20	12.40	<b>Session 4 (second part): Algorithms for Bioinformatics, Metagenomics</b>
11.20	11.40	<i>Malagoli Tagliazucchi G, Miccio A, Cavazza A Poletti V, Peano C, De Bellis G, Mavilio F, Biccato S</i> <b>A bioinformatics framework for the identification of active regulatory elements through the integrative analysis of high-throughput genomic data</b>
11.40	12.00	<i>Micale G, Pulvirenti A, Giugno R, Ferro A</i> <b>A greedy and stochastic algorithm for multiple local alignment of interaction networks</b>
12.00	12.20	<i>Policriti A, Scalabrin S, Vezzi F, Vicedomini R</i> <b>GAM: Genomic Assemblies Merger</b>
12.20	12.40	<i>D'Antonio M, Paoletti D, Santamaria M, Castrignanò T, Pesole G</i> <b>SARMA: a web resource for species assignment of high-throughput sequencing reads from metagenomics analysis</b>
12.40	13.40	<i>Lunch Break</i>
13.40	15.00	<b>Session 5: Biobanks and Biological Databases</b>
13.40	14.00	<i>Di Domenico T, Walsh I, Martin A, Tosatto S</i> <b>MobiDB: a comprehensive database of intrinsic protein disorder annotations</b>
14.00	14.20	<i>Canakoglu A, Gangi P, Gennaro S, Masseroli M</i> <b>Identification of gene annotations and interactions and protein-protein interaction associated disorders through data integration</b>
14.20	14.40	<i>Liberti S, Calderone A, Sacco F, Perfetto L, Iannuccelli M, Panni S, Santonico E, Palma A, Nardoza AP, Castagnoli L, Cesareni G</i> <b>HUPHO: the human phosphatase portal</b>
14.40	15.00	<i>Pappalardo AM, Guarino F, Messina A, Pulvirenti A, Giugno R, Ferro A, De Pinto V</i> <b>A knowledge base for fish and fishery products</b>
15.00	15.45	<b>Session 6: Technological track</b>
15.00	15.15	<i>Guzzi PH, Cannataro M</i> <b>Micro-Analyzer: a tool for automatic pre-processing of multiple affymetrix arrays</b>
15.15	15.30	<i>Gaido L, Bencivenni M, Cesini D, Donvito G, Veronesi P</i> <b>IGI grid services for the bioinformatics community</b>
15.30	15.45	<i>Falciano F, Rossi E</i> <b>FERMI: the most powerful computational resource for Italian scientists</b>
15.45	16.00	<b>Presentation of BITS 2013 and Farewell</b>
16:00	17:00	<b>Tutorial</b>
16.00	17.00	<i>Micale G.</i> <b>Biological Network Alignment</b>

# Keynote Lectures

---



## Renato Dulbecco: genes, cancer and epigenomics



**Ileana Zucchi**

CNR, Institute for Biomedical Technologies, Milan, Italy

In 1986 Dr. Renato Dulbecco was one of the earliest proponents for sequencing of the human genome in order to guide human medical and disease research. He proposed that in the near future, the harmful effects of diseases could be treatable or even curable when the genes and the mutations that the genes harbored could be identified with certainty. This proposal commonly referred as the “human genome project” was based on concepts and ideas that he developed earlier.

In 1975 during the Nobel Prize acceptance awards ceremony, Dr. Dulbecco pioneered the idea of a cancer genome and transcriptome initiative proposing, “cancer is a disease of our genes” and suggesting that the cancer genome and cancer transcriptome would be generated by comparing the cancer cells with their corresponding healthy cells derived from the same patient. The nascent ideas generated from these sequencing proposals were the foundation of his

two primary research focuses developed later. These were the identification of genes and pathways associated with normal cell differentiation compared to genes and pathways associated with abnormal cell regulation that resulted in cancer.

To understand cancer, he suggested one had to understand the normal differentiation process of stem cells, and in particular the genes that govern cell differentiation and tissue development. Stem cells have specific properties that are critical for cancer development and progression, such as significant potential for cellular proliferation and the property of chromatin remodeling, attributes required for cancer cells to adapt, evolve and modify themselves.

The characterization of normal stem cells and the origin of cancer stem cells were the focus of his research spanning almost two decades at the CNR-ITB.

## On bioimage informatics and decoding genomes



**Eugene W. Myers**<sup>1,2</sup>

<sup>1</sup>HHMI Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, United States

<sup>2</sup>Director, Max Planck Institute for Molecular Cell Biology and Genetics (June 2012), Dresden, Germany

We are now at a time when we can systematically alter animals genetically so that any given protein or its expression can be observed in their cells. Combined with new modalities of light microscopy, this allows us to observe molecular mechanisms within the cell, observe the developmental trajectory of growing organs, and to map the cellular anatomy of organisms and organs such as the brain, the heart, or the stem of a plant. All this increasingly requires computation to either extract information or to quantitatively measure an effect in the vast sea of images produced by such explorations. This is creating the growing sub-field of bioimage informatics.

In this talk we introduce the subject to the non-expert with a series of examples from the

work of my group. These include (1) the biophysics of cell division, (2) studies of gene expression in individual cells within the worm *C. elegans*, (3) tracking whiskers in a behaving mouse, and (4) a detailed reconstruction of a fly's brain including the patterning of its development. As time permits, I will give a sample of some of the interesting methods that have arisen from pursuing these projects. Among the highlights are (a) a deformable registration method that reveals glomureli in a consensus built over 100's of brains that cannot be seen in an individual brain, and (b) a neuron tracing algorithm based on Dijkstra's classic shortest path algorithm.

## Statistical inference in high dimensional spaces of genomics: an RNA structural example



**Charles E. Lawrence**

Division of Applied Mathematics, Center for Computational Molecular Biology, Brown University, Providence (RI), United States

The emergence of genome scale data sets leads to increasingly more precise parameter estimates that are ideally suited for maximum likelihood methods and other highest scoring procedures, when the number of unknowns is modest. However, paradoxically just the opposite is becoming increasingly common in genomics. This paradox has emerged because these technologies have simultaneously opened opportunities to draw inferences on previously unanswerable high dimensional questions. In this regime the curse of dimensionality not only denies frequentist methods including maximum likelihood estimation of all their asymptotic advantages, but also often makes these estimates at best misleading if not downright wrong. However, ensemble based Bayesian inferences do not suffer from these afflictions, as they recognize that drawing inferences is an inherently uncertain process and employ the laws of probability to address this uncertainty. This talk will briefly introduce the ideas probabilistic statistical inference using the following example of RNA secondary structure prediction. RNA secondary structures play a crucial role in the function of many RNAs, and structural features are often essential to their interaction with other cellular components. But as we show the

Boltzmann weighted space of RNA secondary structures can be very complex. Here we present a new algorithm, RGibbs, to identify RNA motifs in longer unaligned sequence, and predict consensus secondary structures for using the blocked Gibbs sampler, which has theoretical advantage in convergence time. This algorithm iteratively samples from the conditional probability distributions  $P(\text{Structure} \mid \text{Alignment})$  and  $P(\text{Alignment} \mid \text{Structure})$ . We illustrate how these probabilistically drawn samples can characterize these potentially complex spaces using hierarchical clustering method to characterize the shape of the posterior space,  $\gamma$ -centroid estimator to generate a prediction from sampled structures, and credibility limits to characterize the uncertainty. An analysis of 17 RNA families shows substantially improved structural prediction based on PPV-SEN curves comparisons, compactness of sampled structures around their ensemble centroids, at least eleven families with well separated clusters. The fact that the distances between the references structures and the centroid structures were large compared to the variation among structures within an ensemble raises questions the aptness of the term maximum expected accuracy estimator.

# Oral Presentations

---



## Integrated cloud environment for characterization of genotype specific transcriptome from next generation sequencing data

R. Giugno<sup>1</sup>, F. Abate<sup>2</sup>, N. Bombieri<sup>3</sup>, M. Delledonne<sup>4</sup>, A. Ferrarini<sup>4</sup>, E. Ficarra<sup>2</sup>, A. Pulvirenti<sup>1</sup>, A. Acquaviva<sup>2</sup>✉

<sup>1</sup>Dipartimento di Biomedicina Clinica e Molecolare, Università di Catania, Italy

<sup>2</sup>Dipartimento di Automatica ed Informatica, Politecnico di Torino, Italy

<sup>3</sup>Dipartimento di Informatica, Università di Verona, Italy

<sup>4</sup>Dipartimento di Biotecnologie, Università di Verona, Italy

### Motivations

Recent data coming from the comparison of genomes of different individuals in human species and of different genotypes in plants has led interesting findings about the differences among individuals, ecotypes or genotypes. Cross-species conservation analysis revealed that many of the genes potentially encoded by novel sequences are conserved across a number of mammal and might be biologically functional and thus may be related to differences in gene networks between human individuals. This strongly suggests that genetics and transcriptomics must be performed in the context of individual genomes.

NGS technologies provide for the first time the opportunity to study the complexity of individual-specific sequences. However a full genome assembly still presents problems due to highly repetitive sequences which cannot be easily solved with current technologies.

### Methods

The first step in our workflow is de novo assembly based on de bruijn graph assembly plus an error detection and correction step based on comparison with datasets of annotated proteins. This has been implemented in order to overcome limitations of current assembly methods which

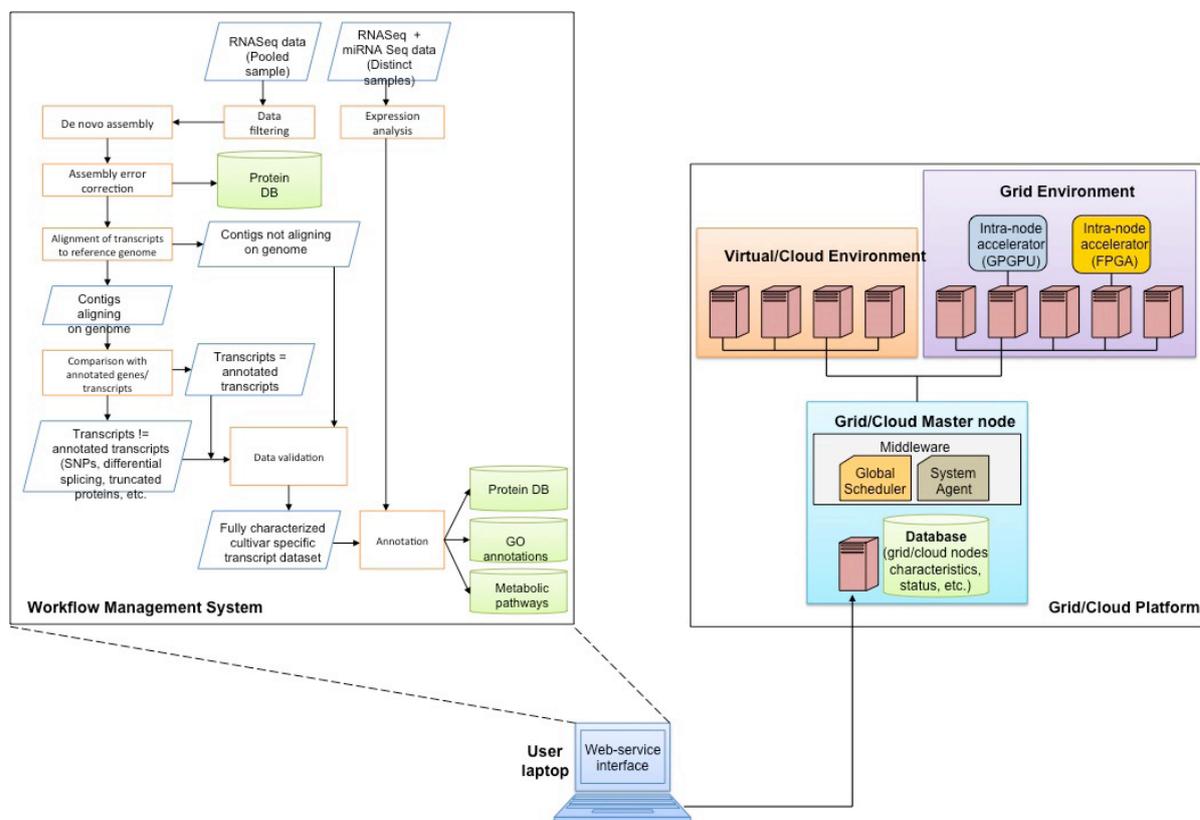


Figure 1. System architecture.

rely uniquely on sequence data and thus they do not prevent frameshift or overassembly errors. The platform determines if the new genes and transcript isoforms are potentially functional and if mutations disrupting the functionality of the original gene models present in the reference genome are compensated by the new isoform. Those data are integrated and linked to expression profiles, annotation functions and network data. This allows determining if metabolic pathways are affected or modified by the expression of transcripts alternative to those expressed in the reference genotype or by the expression of novel genes. On the algorithmic viewpoint, innovative approaches contributing to efficiently carry out the comparison of reconstructed transcriptomes with reference genome and quantify the transcriptome and proteome diversity will be proposed based on: (i) Machine learning techniques to genome reassembling; (ii) Functional enrichment based on non parametric statistical tests; (iii) Gene similarity based on common miRNA targeting and RNA editing function; (iv) Probabilistic generative models for network analysis. On the computational viewpoint, we propose an innovative infrastructure, based on grid/cloud computing and efficient intra-node accelerators (i.e., GP-GPUs and FPGAs). Since complex analysis pipeline made of several stages are characterized by heterogeneous computational

requirements, we developed a middleware infrastructure where specific schedulers and task migration agents will orchestrate task allocation both across nodes and within nodes. The orchestration will be performed by matching application computational kernels characteristics (obtained through off-line profiling) with computational capabilities of nodes. Moreover, since transcriptome reconstruction requires the capability of processing many biological samples for statistical and comparative reasons and current frameworks are not optimized for multi-sample analysis, rather they run various samples sequentially, we designed techniques for efficient sample-level allocation on computational nodes. See Figure 1 for a description of the platform.

### Results

The solution we propose here improves the existing solutions in the following two directions. First, efficient algorithms are applied for genome reconstruction and identification. Second, these algorithms are implemented in a pipeline analysis framework, where the processing of multiple samples is optimized to better exploit computational resources. The infrastructure makes possible for bioinformaticians, through a web service interface, to build workflows and execute them on a grid/cloud computing platform in a easy to use and programming-friendly environment.

## Next Generation Programming: software tools for NGS tertiary analysis

**U. Pozzoli**

Bioinformatics Lab, Scientific Institute I.R.C.C.S E. Medea, Bosisio Parini, Italy

### Motivations

NGS experiments produce a huge amount of raw data and great efforts are currently ongoing to develop algorithms and workflows to manage this data-flood and to produce reliable results. Computational tools for primary and secondary analysis are actually part of and evolving along with sequencing technologies. Still, the amount of results (i.e. genomic variations calls) is considerable and their interpretation far from an easy task. This is the most complex, experiment-specific and time-consuming phase of the NGS data analysis pipeline. In most laboratories, with Sanger sequencing, few Kb sequences are analyzed at a time and most of the functional analysis are performed manually, filtering out known SNPs, and using annotations from genome browsers to obtain hints about the possible functional impact of candidate genomic variations. Given their relatively complex usage, computational tools are rarely used to predict variations effect on biological function. Lab people tend to apply a similar approach for the analysis of NGS results and this leads to a marked preference for small targeted experiments. Targeting of small regions can be the proper approach when sequencing of small regions in a high number of samples is needed; nevertheless by comparing target enrichment and individual genome sequencing costs it is evident that this is going to change. The role of computational methods is going to be pivotal in the interpretation of NGS experiments. Despite the great variety of extremely useful algorithms their software implementation is lacking the characteristic that can make them readily applicable to NGS results. The vast majority of them is sequence driven and analyze sequences provided by the user which has to manually deal both with variations and annotation. This is inefficient and time consuming. Workflow management platforms can partially overcome these limitation but they still lack efficiency and often require adapter software layers to incorporate new algorithms. A set of software objects able to represent genomic annotation, computational features and varia-

tions is therefore needed to implement software readily suitable for NGS results.

### Methods

Based on this premises we extended the previously described GeCo++ C++ Library [1] by adding a class, namely gGenotype that can describe genotypes for one or more samples deriving both from sequencing and genotyping experiments. The class serves as an interface between the software and a properly defined relational database containing genotype information in space efficient way (i.e. only variations from some reference are inserted). Any NGS sequencing experiment result set can be stored in the database and the results accessed from any software using the gGenotype class. Another class (gElement) gives a representation of a genomic element (namely a transcript, a set of gene with isoforms etc) that can be enriched with computed features. Variations can then being applied to the element obtaining a "mutated" version for which any feature is recalculated only where needed. Furthermore we developed a set of tools to obtain annotations from the most used public databases (i.e. UCSC Genome Browser and Ensembl), to read genotype information from vcf files and to interface the software with the sequencing database. We also developed a simple but effective application framework to write applications that can easily communicate (i.e. no need for format conversions) and being called remotely using http as a transfer protocol.

### Results

We briefly present here the principles of GeCo++ library usage and, more extensively, two application examples: one in the field of population genetics and the other one to predict the effects of mutations on Transcription Factor Binding Sites.

### Availability

<http://gecolibrary.sourceforge.net/>.

### References

1. Cereda M, Sironi M, Cavalleri M, Pozzoli U: GeCo++: a C++ library for genomic features computation and annotation in the presence of variants. *Bioinformatics* 2011, 27(9):1313-5

## Building an optimized pipeline for whole-exome sequencing

M. D'Antonio<sup>1</sup>, P. D'Onorio De Meo<sup>2</sup>, B. Elmi<sup>1</sup>, N. Sanna<sup>2</sup>, G. Pesole<sup>3,4</sup>, T. Castrignanò<sup>2</sup>✉

<sup>1</sup>Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli Studi di Bari, Bari, Italy

<sup>2</sup>CASPUR, Consorzio interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy

<sup>3</sup>Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli Studi di Bari, Bari, Italy

<sup>4</sup>Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

### Motivations

Managing the huge amount of data produced by NGS platforms requires non trivial IT skills. Furthermore the wide list of freely available analytical tools for NGS data analysis makes difficult to choose easily the pipeline components. An additional layer of complexity is due to the need of integrate all steps in a single analysis: the best tool for a specific purpose could be incompatible with other tools of the pipeline, i.e. a tool performing a statistical calculation when raw data are required for the next step. In case of a whole exome analysis building effective pipelines that relate variants to their samples and controls, annotate them from multiple sources requires a large customization effort.

### Methods

Our proposed pipeline walks through several steps to perform a full analysis. 1) Before mapping the short-reads against the reference genome, a pre-process is necessary. FastQ files should be checked for integrity and cleaned up from any unwanted symbols that can alter any NGS tool behavior. Quality checks can be pursued with tools like FastQC to ensure that sequences provided reach the minimum level of mean quality necessary for a complete analysis. 2) Alignment is usually performed with BWA [1], which is capable of finding gaps. It results to be a good compromise between speed and accuracy. When there are known problems in the sequence provided, e.g. FastQC outlines a poor quality in the last or first bases sequenced, other tools can perform a more sensible alignment at a lower speed. 3) BWA provides mapping results in SAM format [2]. This is the most widely used format for alignment output. This text-based format should be converted into its binary equivalent BAM format through the SAMtools; BAM can be indexed and sorted to enable faster operations at subsequent steps. 4) Before searching any variant in mapping binary data, some other editing are required to prevent

artifacts in results. Quality recalibration is required to refine some oddness caused by sequencing and mapping on quality scores. Duplicates are in most of the case result of PCR amplification and should be avoided as they lead to false positives. A re-alignment around known indels position should be also carried on to delete other artifacts. 5) Single Nucleotide Polymorphism (SNP, a single nucleotide occurring in one member is replaced by another nucleotide in the other member) and Deletion-Insertion Polymorphism (DIP, refers to the fact that a short nucleotide sequence in one member is omitted in the other member) can be now called from the mapping data obtained from the previous 4 steps. 6) SNP and DIP obtained have various score to consider to ensure a minimum depth coverage and quality score in order to remove any false positive in the list. 7) When dealing with multiple WES data lanes, the usual scenario is a combination of affected/unaffected tissue samples. In this case a critical information is about the haplotype phasing, which allow discovering complex heterozygous or homologous mutations. 8) The last critical aspect of variants calling is to associate as many annotation as possible to the variant list i.e. annotation stored in database like dbSNP, 1000genomes, etc. After these steps data can be saved into custom databases to allow cross-linking and intersections, statistics and much more.

### Results

We have tested different freely available algorithms used at the alignment and post alignment stage and integrated them with custom-build scripts to provide the most suitable and complete combination to create significant whole exome dataset results. Hence, we have customized the whole-exome data analysis pipeline to preferentially held true variants by minimizing the incidence of false positives and providing the benchmarks for the best choice of right analytical tools.

## References

1. Li, H, Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009 Jul 15;25(14):1754-60
2. Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G, Durbin, R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078-9

## On the impact of short-reads quality on variants detection

S. Castellana<sup>✉</sup>, T. Mazza

IRCCS Casa Sollievo della Sofferenza - Mendel, Italy

### Motivations

Next Generation Sequencing technologies have greatly improved our ability to detect genetic variations within organisms. In particular, they have been recently applied with great success to the discovery of mutations linked to rare mendelian diseases. The goodness of the results has been showed to be tightly dependent on two as crucial as systematic family of errors: those introduced by the particular sequencing chemistry and those due to the library preparation procedure. As a consequence, it is rather important to always assess the quality of the sequencing products before proceeding to carry out any downstream analysis, especially when errors, even if rare, might be decisive to the fulfillment of the overall objective. In the context of variants discovery, errors are deleterious. We have then implemented an accurate bioinformatics pipeline, which particularly applies to SOLiD raw data, and that aims to reduce the occurrences of false positive mutation calls. It consists of several sequential steps: SOLiD short reads quality assessment, dataset filtering, mapping, exome coverage analysis and variants discovery.

### Methods

Four exon-enriched genomic libraries, belonging to 3 unrelated patients with a severe neurogenetic disease and one control sample, were sequenced by the SOLiD 4 sequencing platform (Applied Biosystems, part of Life Technology, Foster, USA). Site-specific and global reads quality were analyzed by the QV<sub>2</sub> assessment tool [1] and by means of custom scripts. Then, because of the assessed qualities, color-space sequence files were filtered according to different threshold values. Afterwards, filtered and original sequence libraries were mapped, using the LifeScope Genome Analysis Software. Aligned reads were recalibrated through the GATK (The Genome Analysis ToolKit). Performances of mapping were calculated by custom scripts, while depth of coverage for each target site and target exon was calculated by means of the Bedtools Package and some R custom scripts. Variants were de-

tected by the GATK and, later, cross-checked by Samtools and diBayes. Therefore, they were annotated by ANNOVAR. We looked for novel candidate mutations by searching our variants in the dbSNP (ver. 1.35) repository and setting a minimum allele frequency threshold to 0.02. For two out of the four libraries, we could check the resulting variants against those of as many SNP arrays. We could then assess the performance of our filtering strategies and do some statistics on the number of false/true positive variants call.

### Results

The quality of these SOLiD 4 raw data has been found to be generally low. Accuracy of color calls decreased gradually along the 50 bp fragments length, and resulted to be under 20 already around the 20th-25th read position. As a confirmation, even the application of the most relaxed filtering strategy caused a dramatic reduction of mappable reads: indeed, the four raw datasets ranged from 50 to 80 Mb, but only the 35-65% of sequence data passed the filters. Reason for this seems to be linked to SOLiD 4 sequencing chemistry rather than to the library preparation procedure. For filtered and unfiltered libraries, the median site-specific coverage ranged from 12 to 32, with the 50%-80% of sites having a depth of coverage higher than 10. Regarding the entire set of target exons (over than 200000), from 4% to 10% of them were systematically skipped or poorly covered by the sequencing process. Regarding the variant calling task, a large number of low quality SNPs and Indels has been detected on the four raw datasets, while from 5000 to 15000 variations have been found among the filter-passing datasets (going from the more stringent to the more relaxed filtering criteria). Because of the recessive mode of inheritance of the mendelian diseases under examination, we searched for homozygous mutations in exons and splice sites, finding different pools of non-synonymous, stop-gain, stop loss and splicing mutations within each filtered library. In addition, low ratios of transition vs. transversion and non-synonymous vs. synonymous variants clearly

confirmed the general low sequencing quality of the four experiments. Comparison with SNP array data finally helped to determine the best filtering configuration and the variants set with minimal impact of false positive calls. As a side effect, the presented pipeline produced some detailed reports about the SOLiD read quality, mapping accuracy, coverage of target regions and variants detection. It implemented some well known

best practices and custom methodologies and, even in case of bad sequencing outputs (as the above-mentioned cases), tried to extract reliable information and present them to biologists and clinicians.

### References

1. Sasson A, Michael TP (2010) Filtering error from SOLiD Output, *Bioinformatics* 26, 849-850.

## Improving biomarker discovering for chemosensitivity prediction using an integrated approach

A. Visconti<sup>1</sup>, F. Cordero<sup>1</sup>, R.A. Calogero<sup>2</sup>✉

<sup>1</sup>Department of Computer Science, University of Torino, Torino, Italy

<sup>2</sup>Molecular Biotechnology Center, University of Torino, Torino, Italy

### Motivations

Recently, high-throughput techniques have been successfully used to investigate different aspects of the cell behaviour, opening new perspectives in fields such as molecular biology, medicine, and pharmacology. These advancements have led to the arising of new areas of research such as pharmacogenomics: the discipline studying the influence of genetic variations on drug response having the objective of optimizing drug therapies so to ensure the maximum of efficacy and the minimum of side effects. A crucial step in pharmacogenomics is the discovery of genes (biomarkers) that are responsible for drug responses. By using each gene as a 'feature' and the drug response as the 'class' to be predicted, the problem can be casted as the one of 'finding the set of features that allows the best prediction of the class'. This is a known machine learning task known as the 'feature selection problem'. Feature selection algorithms are usually classified into two categories: filter and wrapper approaches. In literature, several filter methods have been described for the task of detecting biomarkers. However, the kind of statistical analysis carried on by filter approaches cannot capture interactions among genes. Wrapper methods use the prediction performances of a given machine learning approach to assess the usefulness of a subset of genes. They consider reciprocity among genes and obtain remarkable performances. Nevertheless, the information about genes interactions in biological pathways is still missed.

### Methods

In this work we propose an integrated approach to detect biomarkers liable for cell lines respons-

es to drug administration. Specifically, our approach integrates: i) a filter and a wrapper technique for biomarker discovering, and ii) different sources of knowledge, namely transcriptional profiles, drugs activity, and pathways interactions. The proposed approach is composed by two steps. In the first step, we apply a filter method to identify a set candidate biomarkers. This is a pre-processing method exploiting a priori knowledge about gene expression levels and gene interactions. In detail, we identify differential expressed genes using the Rank Product methodology. Then, we perform a pathway analysis for extracting genes (i.e. network hubs and bottlenecks) that are likely to be responsible for the measured differential gene expression levels. In the second step, we utilize a wrapper approach to single out biomarkers. To this purpose, we use multiple runs of a genetic algorithm to assess the importance of each candidate biomarker.

### Results

We use the NCI60 DNA panel to test our approach. It consists in an in vitro screening of several chemical compounds over 60 human cancer cell-lines. We select 118 drugs whose mechanisms of action are known. To assess the quality of the proposed approach, we compare the obtained accuracies to those of a rough wrapper technique, namely Random Forests. Random Forests has been considered standard 'tool-box of methods' for class prediction and gene selection with microarray data. We outperform Random Forests approach. Finally, we analyze the extracted biomarkers by using the 'Ingenuity systems' showing that they are strictly related to the targets of administrated drugs.

## The location of T1 diabetes associated SNPs in regulatory regions

S. Beka<sup>1</sup>, I. te Boekhorst<sup>1</sup> , I. Abnizova<sup>2</sup>

<sup>1</sup>School of Computer Science, University of Hertfordshire, Hatfield, United Kingdom

<sup>2</sup>Wellcome Trust Sanger Institute, Hinxton, United Kingdom

### Motivations

Although many association studies on complex diseases focus on variation in coding DNA, recent research shows increasing evidence that the cause of such diseases should be sought in the regulation of gene activity. Rather than studying mutations in genes coding for transcription factors, my work focuses on genetic variants (SNPs) in regulatory modules (TFBS, enhancers, promoters, or other genic locations likely to be involved in gene regulation such as UTR, introns and splice junctions) that are in Type 1 Diabetes (T1D) susceptibility regions of the human genome. The specific research question is: are SNPs associated with T1D more likely to occur in (putative) regulatory regions than other SNPs found in T1D susceptible regions? In addition: Because genes may overlap and/or occurrence in multiple transcripts, one and the same SNP may be associated with more than one genic location and affect more than one functional region (e.g. is a mutation in as well a coding region as a regulatory region). Are SNPs associated with a complex disease such as T1D more likely to be of this kind?

### Methods

An extensive search in the databases ENSEMBL and T1Dbase was conducted to collect information all SNPs in T1D susceptible regions [coordinates on genome, type of variant (mutant allele, wildtype allele, transversion or transition, synonymous or non-synonymous), transcriptID, intra-genic region (up/downstream, exon, intron, splice junction, UTR, type of (micro)RNA), type of gene (coding/pseudo/microRNA) and associated diseases] A statistical analysis was performed to assess possible associations between the status of a SNP (associated/not-associated with T1D) on the one hand and features characterising the (intra/inter-) genic position on the other hand.

### Results

1) SNPs associated with T1D are more likely to occur in regulatory regions than those that are not associated with T1D 2) SNPs associated with T1D occur more often in multiple genic locations than those that occur in only one genic location. They appear to be relatively over-abundant in transcription factor binding sites, introns and splice-sites.

## Quasi-cellular systems: stochastic simulation analysis at nanoscale range

L. Calviello<sup>1</sup>✉, P. Stano<sup>2</sup>, F. Mavelli<sup>3</sup>, P.L. Luisi<sup>2</sup>, R. Marangoni<sup>1,4</sup>

<sup>1</sup>Department of Computer Science, University of Pisa, Pisa, Italy

<sup>2</sup>Department of Biochemistry, University of Rome III, Rome, Italy

<sup>3</sup>Department of Chemistry, University of Bari, Bari, Italy

<sup>4</sup>CNR - Institute of Biophysics, Pisa, Italy

### Motivations

The artificial creation of the simplest forms of life (minimal cells) is a challenging aspect in modern synthetic biology. Quasi/cellular systems able to produce proteins directly from DNA can be created by encapsulating a cell-free transcription/translation system (PURESYSTEM) in microemulsion droplets and liposomes ( $10^{-5}$  -  $10^{-7}$  m of diameter). It is possible to detect the overall protein production inside these compartments using DNA encoding for GFP and monitoring the fluorescence emission over time. The entrapment of solutes in microemulsion droplets and liposomes is a complex process that creates a population of compartments with different internal compositions of molecular species, which affects the final protein production. A complete understanding of the distribution of solutes inside the different compartments and on its effect on the course of internal reactions are two relevant and still open issues in the field. Stochastic simulation is a valuable tool in the study of biochemical reaction at nanoscale range; QDC (Quick Direct-Method Controlled), a stochastic simulation software which uses the well-known Gillespie's SSA algorithm, was used; a suitable model reproducing the PURESYSTEM reactions network was hence created, with the aim to describe how the different composition of species affects the overall translation process, thus trying to infer the internal composition of each microcompartment from its observable fluorescent signal emission.

### Methods

In order to understand how the protein production is affected by the reactants concentration, we first generate experimental data by combining different amounts of DNA, enzymes, translation factors and consumable. Consequently, a set of fluorescence vs. time curves were gener-

ated. Next, the pre-existing translation model was improved to describe in detail a coupled transcription/translation system with simultaneous elongation events on the same molecule. The dynamical coupling between the transcription and translation systems was assessed using logical formulations allowed in QDC's syntax, thus creating sequentially dependent processes in the concurrent-only environment of Gillespie's algorithm. Stochastic simulations were performed in order to globally fit, by sigmoid curves ( $R^2 > 0.98$ ), the entire experimental dataset for protein production. The comparison of the parameters estimates for the different inputs showed how protein production is strongly affected by enzymes concentration.

### Results

Different kinetic parameters were considered, such as the final plateau of production and the initial time required to release the first complete proteins from elongating ribosomes. Further analysis demonstrates how the transcription process plays a fundamental role, determining the rate of GTP depletion from the system, hence preventing ulterior elongation of peptide chains. To the best of our knowledge, the present work is the first one describing in detail the stochastic behavior of the PURESYSTEM™. The developed model allows us to scale the system down to compartment diameters where anomalous entrapment phenomena are presumed. Thanks to our results, an experimental approach is now possible, aimed at recording the GFP production kinetics in very small microemulsion droplets or liposomes, and inferring, by using the simulation as a hypotheses test benchmark, the internal solutes distribution, and shed light on the still unknown forces driving the entrapment phenomenon.

## An analytical spatially-based approach to study cancer cell population evolutions

F. Cordero<sup>1</sup>✉, M. Gribaudo<sup>2</sup>, D. Manini<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Torino, Torino, Italy

<sup>2</sup>Department of Electronics and Computer Science, Polytechnic of Milano, Milano, Italy

### Motivations

The evolution of many biological dynamics is strongly related to the spatial composition of the tissues where it takes place. It becomes of paramount importance, not only the knowledge of the type and the number of different biological entities involved, but also their position and their distribution in space. Spatial techniques are particularly useful to investigate the mechanisms affecting tumor growth and maintenance upon vaccination or drug treatment. However not many techniques have been studied to tackle the spatial definition of the environment in addition to the densities of the various elements. This is mainly due to the difficulties of modeling a spatial distributed system using conventional techniques.

### Methods

In this work we present a new approach, based on an evolution of the mean field analysis that exploits some of the features of the “Markovian agents” performance evaluation formalisms. This new technique allows to study the system evolution through a sound analytical treatment of the spatial dependency. In particular, we focus on the study of tumor progression, showing that the growth and progression of many cancers are driven by small groups of Cancer Stem Cells (CSCs). The CSC tumor model presents a hierarchically structured organization similar to that found in normal tissues: CSCs are self-renewing, capable of tissue regeneration and of giving rise to non-CSCs, the latter being more differentiated and largely lacking in tissue-regeneration ability. Considering this hierarchical structure, the response to drug treatments on the several cell populations that compose the cancer will produce different effects depending on the characteristics of the cancer subpopulation. For this reason, the CSCs are believed to be the cause of failure of conventional therapies, since effective drug treatments able to eliminate all the

CSCs, hence to avoid relapses, are not easy to find. Based upon this model we study the cancer progression, the drug or DNA-vaccination effects, that locally administered, at predefined time periods, can slow down the expansion of the tumor. The considered tissue is divided in small areas, each characterized by its own parameter. The evolution in one area can affect the neighbor: this allows modeling the propagation of both cancer growth or the therapeutic effects of drugs or vaccination. Our approach will aggregate similar cellular population, using differential equations, rather than modeling all the involved entities separately, in order to reduce the model complexity.

### Results

We expect to model the evolution of the cancer in a volume, taking into account the total number of the different cellular population involved and their spatial distribution. The results will be compared with similar approaches in the literature based on cell automata. We show the spatial behavior of the well known therapy, and how the framework proposed could be a basic step for an optimization algorithm that will determine the best time instants in which administer the drug or vaccination, and the best locations where it should be inoculated. The solution of the model can support the in vitro/in vivo experiments for testing new biological hypothesis. Furthermore, our approach can be extended to consider the immunological tumor micro-environment by adding more details on the immunological system involved on the drug/vaccination treatments. This might be particularly interesting in the area of combined treatment development. Tumor vaccination alone is not sufficient to eradicate the disease, but combined with other immuno-pharmacological treatments, affecting the CSC differentiation rate might represent an interesting approach in the area of tertiary cancer prevention.

## A combined computational/experimental strategy to map phosphatases on growth pathways

P.F. Gherardini<sup>1</sup>, F. Sacco<sup>1</sup>, S. Paoluzi<sup>1</sup>, J. Saez-Rodriguez<sup>2</sup>, M. Helmer-Citterich<sup>1</sup>, A. Ragnini-Wilson<sup>1,3</sup>, L. Castagnoli<sup>1</sup>, G. Cesareni<sup>1,4</sup>✉

<sup>1</sup>Department of Biology, University of Rome "Tor Vergata", Italy

<sup>2</sup>EMBL-EBI, Hinxton, UK and EMBL-Genome Biology Unit, Heidelberg, Germany

<sup>3</sup>High-throughput Microscopy facility; Department of Translational and Cellular Pharmacology, Consorzio Mario Negri Sud, SM. Imbaro, Italy

<sup>4</sup>IRCCS Fondazione Santa Lucia, Rome, Italy

### Motivations

Large-scale screenings allow linking the function of poorly characterized genes to phenotypic readouts. According to this strategy, genes are associated to a function of interest if the alteration of their expression perturbs the phenotypic readouts. However, given the intricacy of the cell regulatory network, such screenings often provide mostly qualitative results, as it is difficult to identify the molecular mechanisms underlying the observed phenotype. In recent years computational modeling has emerged as a powerful tool to investigate biological signaling. In general these models are necessarily limited in size due to their complexity. Conversely large-scale perturbation screenings provide a higher-level view of a larger number of proteins. In this work we aim to bridge the gap between these two worlds in order to obtain a higher-detail mapping of gene products onto complex pathways on a large scale. This strategy was applied to map the poorly characterized family of human phosphatases on pathways related to cell growth.

### Methods

The strategy we have developed is based on two complementary datasets, which are conceptually linked by "sentinel proteins". These are defined as a number of molecular readouts that define the state of the cell, given specific experimental conditions. The datasets we used are: 1) A detailed mechanistic model describing a pathway of interest. The model includes as entities the sentinel proteins. This is constructed starting from low/medium-throughput experiments. 2) A high-throughput perturbation screening where the molecular readout is defined by the sentinel proteins. The core of the strategy is to use the signaling model to simulate the result of up/down regulating each protein in the signaling pathway. Each perturbation results in a

predicted cell state defined as the calculated activity of the sentinel proteins. By matching this signature with the experimental profile obtained in the high-throughput screening it is possible to infer the target and effect (activating/inhibitory) of each protein screened, thus defining its "entry-point" in the network. For instance, if the silencing of a gene results in the same cell state obtained when protein A is up-regulated in the simulation, we can infer that the gene modulates the activity of protein A. The biological modeling was performed using CellNetOptimizer. This software allows the construction of boolean logic models which are represented as signed direct graphs, representing activating/inhibitory relationship between proteins. CNO then optimizes the topology of the model against a dataset of experimental data in order to remove connections that are not relevant in the specific cell system and to integrate using AND/OR logic gates multiple stimuli acting on the same protein. The optimization procedure is repeated 1000 times both because of its stochastic nature and also because the training data are generally not sufficient to fully constrain the model. By performing subsequent simulations on this family of 1000 models it is possible to average out the inconsistencies present in any single model. Moreover by this approach one obtains quantitative predictions even though a given node can only be on (1) or off (0) in any single model. We also extended this strict boolean approach to allow the simulation of three different states for a protein, i.e. control, up and down regulation.

### Results

We assembled from the literature a network describing cell growth pathways. This model includes 34 species and 59 stimulatory or inhibitory interactions and was optimized using CNO against a dataset of experimental data ob-

tained in HeLa cells. The model also includes five sentinel proteins, whose activation status defines the "cell state". We used the results of a siRNA screening of the human phosphatase family that yielded 58 proteins whose silencing affect the activity of one of the sentinel proteins. The interference of 35 of the 58 phosphatase hits (60%) results in a profile that matches one of those inferred by inactivating or activating in silico one of the nodes of the model. The correctness of this mapping, and thus the predictive capabilities of the model, was demonstrated in a number of experiments. In particular one experiment confirmed that the over-expression of four

phosphatases, which were mapped to different positions of the signaling network (i.e. upstream and downstream of AKT and ERK), differentially affect the activity of two readouts (RAF1 and AKT activation) that were not considered in the mapping procedure. In conclusion in this study we developed a novel strategy to map perturbations screening onto complex pathways. The proposed mapping strategy is general and could be used in combination with the results of such large screenings to achieve a more detailed mechanistic description of the molecular mechanisms by which genes or small molecules determine phenotype modulation.

## Drug repositioning through pharmacological spaces integration based on networks projections

M. Re<sup>1</sup>, M. Mesiti<sup>2</sup>, G. Valentini<sup>1</sup> ✉

<sup>1</sup>Department of Computer Science, University of Milano, Italy

<sup>2</sup>Department of Informatics and Communication, University of Milano, Italy

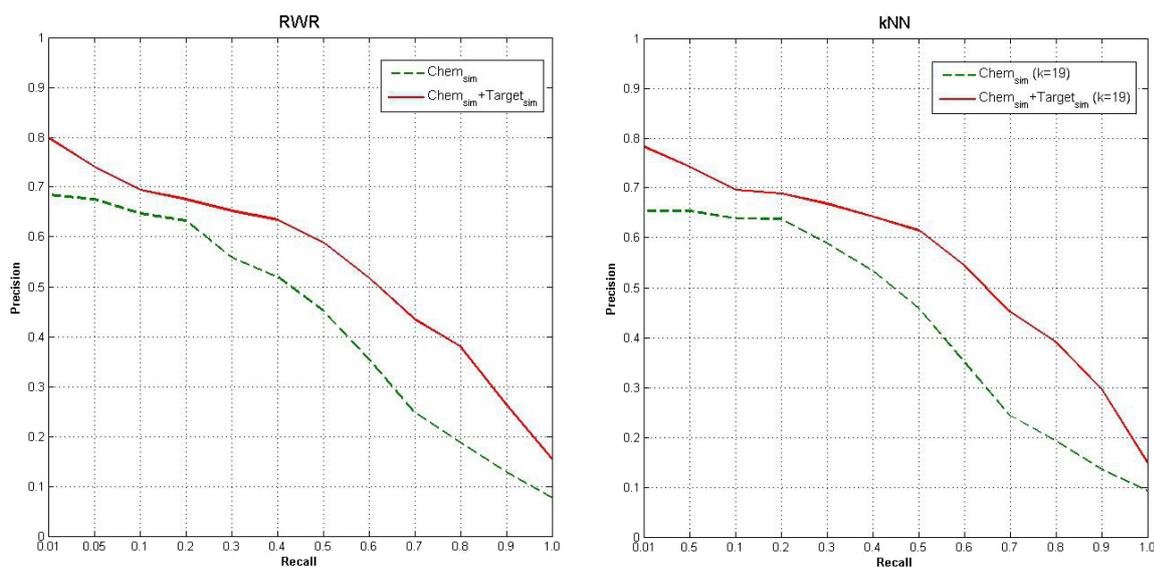
### Motivations

Drug development is a costly and failure-prone process and, in recent years, pharmaceutical industry has experienced a difficult period whereby productivity has not kept pace with increases in research and development costs. As a consequence, quite recently research efforts focused on a novel paradigm for drug development, named drug repurposing, to discover novel pharmacological applications of existing drugs. Computational approaches for drug repositioning focused mainly on small-scale applications, such as the analysis of specific classes of drugs or drugs for specific diseases. Large-scale applications, involving a relatively large number of drugs and diseases, count only a few examples. Despite the availability of many drug repositioning methods, they all suffer from a serious limitation: the inference task is performed in an inhomogeneous similarity space induced by the relationships existing between drugs and a second type of entity (e.g. disease, target, or ligand set), thus making difficult the integration

of multiple sources of biomolecular and chemical data into a homogeneous pharmacological space.

### Methods

To overcome this limitation we propose a general framework based on bipartite networks projections for the construction of homogeneous pharmacological spaces. The nature of these network structured projected spaces allows the application of prediction algorithms to homogeneous pharmacological spaces and improves the integration of different chemical, biomolecular and clinical sources of information. At the core of the proposed approach there is the notion of homogeneous pharmacological similarity space defined as a collection of similarities between drugs induced by common relationships between drugs and a second type of suitable entities (i.e. drug-protein target). The reconciliation between these heterogeneous similarity spaces is performed by means of a network projection operation enabling the reduction of a network composed by two types of nodes (i.e. drugs and



**Fig.1** : Precisions at fixed recall levels, with the chemical similarity and chemical/target similarity pharmacological networks. Random Walks with Restart (RWR), our proposed method using the kNN score (kNN). Results are averaged across all the considered therapeutic DrugBank classes.

drug targets) to a network composed only by drugs. A key feature of the proposed framework is its ability to integrate networks of different sizes, enabling the combination of both high and low coverage networks and resulting into a progressively enriched pharmacological similarity network. We also propose a novel and very fast kernelized semi-supervised network based method for ranking drugs according to their likelihood to belong to a given therapeutic category.

### Results

We evaluated the proposed approach by integrating two pharmacological similarity spaces accounting, respectively, for chemical similarity and drug-targets interaction similarity, in order to rank about 1300 U.S. Food and Drug Administration (FDA) approved drugs according to DrugBank 3.0 therapeutic categories. The experimental setup is based on a canoni-

cal 5-fold cross validation scheme repeated 10 times. The analysis of the precision at fixed recall levels (see Figure), shows that the integration of pharmacological spaces constructed through the proposed network projections significantly enhances the results obtained with different network-based ranking methods. Moreover our proposed kernelized semi-supervised method for ranking drugs according to a given therapeutic category is at least comparable in terms of AUC and precision at fixed recall, and orders of magnitude faster than state-of-the-art ranking methods. Despite a thorough analysis of the results relative to each therapeutic category is out of the scope of this preliminary investigation, the analysis of the top ranked false positives predicted in three drug categories shows that our proposed approach can be successfully applied to discover potential drug candidates for novel therapeutic indications.

## Cyto-Sevis: semantic similarity-based visualisation of protein interaction networks

P.H. Guzzi<sup>✉</sup>, M. Cannataro

Department of Medical and Surgical Sciences, University "Magna Graecia" of Catanzaro, Catanzaro, Italy

### Motivations

The analysis of the whole set of molecular interactions in an organism, often referred to as interaction networks, is becoming an important research area. The recent trend in this area is represented by the integration of raw proteomics data with biological knowledge, usually represented as terms and relations into different ontologies (e.g. Gene Ontology). Usually raw proteomics data are enriched by such biological knowledge by annotating them with terms extracted by the ontologies. The use of such annotations for the analysis of protein data is a relatively novel research area that is currently becoming more and more central in research. A main area is represented by the definition of the similarity among genes and proteins on the basis of the annotating terms, also referred to as semantic similarity analysis of protein data. Semantic similarity measure refers to a set of techniques used to evaluate the similarity of two or more terms belonging to the same ontologies. Consequently, they may be used to evaluate the similarity of two genes or proteins starting from the terms extracted from the same ontology used to annotate them. This theoretical framework may be used also for the development of novel visualization techniques for protein interaction networks based on the semantic similarities of proteins [1]. In this scenario three kind of bioinformatics software are appearing: (i) tools for automatic annotation of proteomic data with Gene Ontology terms yielding to annotated protein interaction databases; (ii) tools for querying and filtering in a semantic way such annotated databases; (iii) tools for semantic visualization of annotated protein interaction networks. The work presented here belongs to the third class [2-3].

### Methods

Cytoscape is a tool for visualizing and analyzing interaction networks, based on an extensible architecture. Cytoscape offers basic visualization techniques where a distinction among nodes

(proteins) or edges (interactions) belonging to different classes can be obtained in a manual way by explicitly setting their colors. Moreover, recent Cytoscape plug-ins are involved with Gene Ontology terms and interaction networks. BINGO determines which GO categories are statistically overrepresented in a set of genes of a sub-graph of a biological network, while GOLORIZE uses GO categories to guide the network graph layout process and to emphasize the biological function of the nodes. Nevertheless Cytoscape currently does not offer directly or through additional plug-ins, the capabilities to visualize networks in a semantic similarity space.

### Results

CytoSevis is a Cytoscape plugin that is able to visualize protein interaction networks in a semantic similarity space. Based on an intuitive interface it is able to load a network from Cytoscape, to load the semantic similarities provided as a separate file, and to visualize resulting coloured network into Cytoscape. The main contributions of the proposed plugin are the following: (i) we exploit semantic similarity analysis, that is currently one of the main developing research area in protein interaction networks; (ii) CytoSevis provides a graphical user interface and enables the visualisation of networks in such a space; (iii) we provide a cross-platform Java-based CytoSevis distribution running on Windows/Linux/macOS; (iv) we provide semantic similarity data on a separate web site. CytoSevis for Windows, Linux and Mac OSX platforms is available under GPL license. User can download it from the Cytoscape web site (<http://cytoscape.org>). Semantic similarities measures are available on the CytoSevis web site (<http://bioingegneria.unicz.it/~guzzi/ss/>).

### Availability

<http://cytoscape.org/>

## References

1. Guzzi P, Mina M. Investigating bias in semantic similarity measures for analysis of protein interactions. In: Proceedings of 1st International Workshop on Pattern Recognition in Proteomics, Structural Biology and Bioinformatics (PR PS BB 2011) published in Nuovo Cimento. Ravenna, 13th September 2011.
2. Guzzi PH, Milano M, Veltri P, Cannataro M. Semantic similarities as features of protein complexes, accepted in 4th IWBNA International Workshop on Biomolecular Network Analysis, in conjunction with IEEE BIBM International Conference on Bioinformatics and Biomedicine, Atlanta 11-13 November 2011.
3. Hiram P, Cuzzi M, Mina M, Guerra C, Cannataro M. Semantic Similarity Measures: Assessment with biological features and Issues, to appear in Briefings In Bioinformatics Oxford Journal 10.1093/bib/BBR066

## Digging in the RNA-seq garbage: evaluating the characteristics of unmapped RNA-seq reads in normal tissues

M. Carrara✉, R. Calogero

Molecular Biotechnology Center, University of Torino, Torino, Italy

### Motivations

RNA-seq has the potential to discover genes created by complex chromosomal rearrangements. 'Fusion' genes formed by the breakage and re-joining of two different chromosomes have repeatedly been implicated in the development of cancer. However, due to the heterogeneous nature of solid tumors, pathological fusions can be confounded with read-through fusions across adjacent genes in the genome, or transcription-induced chimeras (TICs) present in normal tissues.

### Methods

The pipe-line used to detect TICs is the following:

- Removing PE reads associated to transcripts (Bowtie, followed by SHRIMP on everything not mapped by Bowtie);
- removing reads mapping to spiked-in PhiX control;

- trimming linkers;
- detecting TICs with single end mapping (SHRIMP).

### Results

We detected TICs events in 16 normal tissue samples. The information derived by our analysis indicates the presence of significant number of TICs in normal tissues. Some of these TICs can be erroneously associated to cancer development. In some cases TICs expression is corroborated by large number of reads and their expression is significantly spread over different tissues. Our data highlight that TICs could be erroneously associated to cancer aberrations, if RNA-seq analysis is only evaluated in tumor samples, not considering the parallel analysis of normal tissue samples associated to the tumor.

## Alternative splicing as regulator of protein-protein interactions

A. Colantoni, F. Ferrè✉, M. Helmer-Citterich

Department of Biology, University of Tor Vergata, Rome, Italy

### Motivations

In recent years, the role of alternative splicing in protein function regulation has been widely investigated. Examples have been reported in which alternative splicing modulates the interaction between two proteins by removing interacting regions. However, despite the large amount of data available about alternative splicing isoforms and protein-protein interactions, only a few systematic studies have been carried to assess how much widespread and general is this form of regulation. Some of these works lack specificity, while others are based on a relatively small number of cases; in addition, they show conflicting results. In this work, we have created a non-redundant residue-level dataset of protein-protein interfaces derived from the three dimensional structure of human protein complexes in order to determine whether the removal of protein-protein interfaces via alternative splicing in the human proteome could have evolved as a form of interaction regulation.

### Methods

To create a non-redundant dataset of human protein-protein interfaces, we first obtained from the Protein Data Bank (PDB) all the Biological Units containing at least two human protein chains that could be assigned to Uniprot entries according to SIFTS. From this dataset we extracted all the protein-protein interfaces (defined as sufficiently large sets residues involved in Van der Waals contacts) and filtered out those that were not present in the Asymmetric Unit. Using this procedure, we obtained a set of about 12100 protein-protein interfaces, from which, we derived a non-redundant dataset of nearly 2300 interfaces, using a multi-level clustering procedure based on sequence identity between chains, evaluated using the BLASTClust software (<ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>), and on a simple measure of the similarity between two interfaces. For each interface, we mapped the interacting residues to each Ensembl splicing

isoform (if present) of the gene encoding it using residue genomic coordinates retrieved using Biomart, excluding those isoforms which undergo nonsense mediated decay. To evaluate whether splicing has a significant effect on the availability of the interface in the mature protein product, we compared the rate of splicing-mediated interface deletion with that of random controls using a chi-square test. Controls were created picking groups of residues that shared the same dispersion in the sequence with the corresponding interface. An interface, or a control, was considered to be removed from a splicing isoform if the corresponding protein product did not contain at least a certain fraction of its residues: we repeated the procedure using different values of this cutoff. This general procedure was differently adapted for hetero- and homo-dimeric interfaces, taking into account their different nature that might also imply a different evolution of their regulation.

### Results

Initial results showed that, if we consider all the isoforms, there isn't any statistically significant difference between the frequency of removal of real interfaces and random controls, both for heterodimeric and homodimeric interactions. However, discarding all the isoforms that determine a large change in the sequence of the interacting protein (which are not likely to regulate specific functions of that protein), we detected a significant tendency to avoid removal of small fractions of the homodimeric interfaces, suggesting that interaction interfaces are protected from being partially spliced and are more often completely included or removed in the mature transcript. This effect was also seen for compact heterodimeric interfaces. This effect is mainly ascribed to interfaces that are small and whose residues are relatively less scattered on the sequence, suggesting that alternative splicing is less likely to have evolved as regulative agent for large interfaces having more dispersed residues.

## Fungal alternative splicing associates with higher cellular complexity and virulence

K. Grützmann<sup>1</sup>, K. Szafranski<sup>2</sup>, M. Pohl<sup>1</sup>, K. Voigt<sup>3</sup>, A. Petzold<sup>2</sup>, S. Schuster<sup>1</sup>✉

<sup>1</sup>Department of Bioinformatics, Friedrich Schiller University, Jena, Germany

<sup>2</sup>Genome Analysis, Leibniz Institute for Age Research, Fritz Lipmann Institute, Jena, Germany

<sup>3</sup>Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute, Jena, Germany

### Motivations

Alternative splicing (AS) is a cellular process that increases a cell's coding capacity from a limited set of genes. Although AS is common in higher plants and animals, its prevalence and abundance in other eukaryotes is mostly unknown. Especially in fungi the involvement of AS in gene expression is of great interest, as many fungal species are human- and plant-pathogenic.

### Methods

We present a genome-wide comparative study of alternative splicing in 28 fungi from the three phyla Ascomycota, Basidiomycota and Mucoromycotina, based on spliced alignments of transcripts. We apply a sophisticated random sampling strategy to accurately estimate per-gene AS rates of each species.

### Results

We show that our method yields estimates that are independent of available transcript data

amounts. Our analysis reveals that a greater fraction of fungal genes than previously expected is alternatively spliced. On average over all fungi with sufficient data, 6.4% of the genes are affected by AS, with *Cryptococcus neoformans* and *Coccidioides immitis* showing extraordinary rates of 18% and 13%, respectively. On average, the investigated Basidiomycota show higher rates of AS associated genes (8.6%) than the Ascomycota (7.2%, excluding yeasts). We find that fungi with more complex cellular structures and a younger evolutionary age, show higher AS rates and more diverse categories of AS involved genes. Thus, we speculate that AS could facilitate higher cellular complexity in fungi. Furthermore, AS affects genes essential for a pathogenic lifestyle, particularly genes associated with cell rescue and dimorphic switching. We hypothesize that AS is a crucial component of gene regulation and fosters fungal virulence.

## Computational and experimental characterization of critical amino acidic residues in the BCR-ABL kinase domain explaining TKIS resistance in patients with chronic myeloid leukemia

C. Romano, P. Buffa, A. Pandini, M. Massimino, E. Tirrò, L. Manzella, F. Fraternali, P. Vigneri✉

Department of Clinical and Molecular Bio-Medicine, University of Catania, Catania, Italy

### Motivations

Suppression of BCR-ABL1 catalytic activity by the Tyrosine Kinase Inhibitor (TKI) Imatinib Mesylate (IM) has dramatically improved the natural history of Chronic Myeloid Leukemia (CML) ushering the era of molecular targeted therapy. Despite the unparalleled results achieved by IM, 30% of CML patients become resistant to the compound, mostly because of point mutations that interfere with drug binding. Thus, the need for a structural characterization of the interactions between the BCR-ABL kinase domain and different TKIs, that will define the mechanisms allowing BCR-ABL mutants to avoid kinase inhibition by most of these drugs.

### Methods

Point mutations affecting BCR-ABL Kinase domain are not randomly distributed in the BCR-ABL kinase sequence. Indeed, only one (T315) of five residues critical for IM interaction is the object of amino acidic substitutions in IM-resistant patients. We have computationally investigated why these five amino acids (E286, T315, M318, I360, D381) are critical for IM binding. Two of them (T315 and M318) are maintained in binding the second generation (2G) TKI Dasatinib (DAS), while E286, M318, I360 and D381 are required for the interaction with the third generation (3G) inhibitor Ponatinib (PON). We generated tagged-BCR-ABL constructs displaying conservative or non-conservative mutations in each of these amino acidic residues. Every BCR-ABL mutant was lentivirally transduced in Ba/F3 cells and evaluated for expression, catalytic activity, transforming potential and response to TKI treatment. We then performed 50ns Molecular Dynamics (MD) simulations of the mutants showing catalytic activity in a simulated aqueous environment using the GROMACS package with AMBER force field (ffamber99sb), obtaining 50ns of data collection for each system. We also performed MD simulations of ABL KD in complex with the two approved

therapeutic inhibitors, IM and DAS, and with PON. These simulations were aimed at highlighting if and how the mutations influenced inhibitor binding.

### Results

Combining computational and biological approaches we demonstrate that, with the exception of T315, the four remaining amino acids critical for TKI interaction are pivotal to preserve both BCR-ABL kinase activity and oncogenic potential. The conservative mutation I360T was the only substitution capable of maintaining BCR-ABL kinase activity and transforming potential. However, BCR-ABL I360T remained sensitive to TKI treatment. We also employed Molecular Dynamics (MD) simulations to probe molecular motions at an atomic scale and investigate the dynamical features of the mutants that could not be extracted from static structures. The I360T mutation clearly induces a displacement of the C-helix, taking away E286 from the catalytic pocket. This residue is critical for IM binding. To establish if this is a disadvantageous condition for IM stability, we docked this drug inside the pocket of the structure performing further 50ns of MD simulation. Preliminary results suggest that not only the C-helix moves back towards the catalytic pocket (maybe under the electrostatic field generated by the drug) but also that IM restores three of the five original h-bond interactions guarantying IM binding. This result seems to be in agreement with our experimental data showing that BCR-ABL I360T is inhibited by IM. Analysing the interactions between the BCR-ABL kinase domain and different TKIs, our data confirm the efficacy of 3G inhibitor Ponatinib on CML patients failing IM and 2G TKIs, since the interaction with the T315 residue (the only one among five susceptible of mutation) is lost. These findings may be extended to further protein kinases involved in solid and hematologic malignancies, thus contributing to the design of additional TKIs.

## The role of transposable elements in shaping the combinatorial interaction of transcription factors

A. Testori<sup>1</sup>, L. Caizzi<sup>2</sup>, S. Cutrupi<sup>3</sup>, O. Friard<sup>3</sup>, M. De Bortoli<sup>1</sup>, D. Corà<sup>1</sup>, M. Caselle<sup>4</sup>✉

<sup>1</sup>Center for Molecular Systems Biology, University of Turin, Dept. Oncological Sciences, SP142, Candiolo, Italy

<sup>2</sup>Center for Molecular Systems Biology, University of Turin, Bioindustry Park Silvano Fumero, Colletterto Giacosa, Italy

<sup>3</sup>Center for Molecular Systems Biology and Dept. of Life Sciences and Systems Biology, University of Turin, Italy

<sup>4</sup>Center for Molecular Systems Biology and Dept. of Physics v. P. Giuria 1 -10125 Turin, University of Turin, Italy

### Motivations

In the last few years several studies showed that transposable elements (TEs) in the human genome are significantly associated with transcription factor binding sites (TFBSs) and that in several cases their expansion within the genome led to a substantial rewiring of the regulatory network. Here we suggest another possible role played by TEs in the evolution of the regulatory networks. We discuss a set of evidences supporting the idea that the evolution of particular patterns on combinatorial interactions among Transcription Factors (TFs) was mediated and supported by the expansion of specific classes of TEs in the human genome.

### Methods

We tested our conjecture looking, as a case study, at the binding of Estrogen Receptor alpha (ERalpha) to DNA using two chromatin immunoprecipitation sequencing (ChIP-seq) public datasets on MCF7 cell lines corresponding to different modalities of exposure to estrogen. On these datasets we performed a genome-wide analysis of Transposable Elements overlapping ChIP-seq binding peaks and for each of these sequences we performed a genome-wide scan for putative TFBSs employing canonical Positional Weight Matrices (PWMs).

### Results

We found a remarkable enrichment of a few well defined types and classes of transposable elements overlapping the ChIP-seq peaks in our two datasets. Among these enriched TEs a prominent role was played by MIR (Mammalian Interspersed Repeats) transposons. These TEs underwent a dramatic expansion at the beginning of the mammalian radiation and then stabilized. We conjecture that the special affinity of ERalpha for the MIR class of TEs could be at the origin of the important role which ERalpha assumed in mammals. Then, looking at the results of the scan for TFBSs, we found strong enrichment and correlated presence of a few specific TFs. In several cases these TFs were known cofactors of ERalpha, thus supporting the idea of a co-regulatory role of Transcription Factors located within the same TE. Most of these correlations turned out to be strictly associated to specific classes of TEs thus suggesting the presence of a well defined "transposon code" within the regulatory network. Altogether our results support the idea that transposition events, besides rewiring the network, also played a central role in the emergence and success of combinatorial gene regulation in complex eukaryotes and that the evolution of specific combinations of TFs interactions was actually mediated and driven by the expansion of a few specific classes of Transposable Elements.

## Translation of a robust, biology-driven, prognostic classifier of cancer patients' outcome into clinically relevant rules

D. Cangelosi<sup>1</sup>✉, M. Muselli<sup>2</sup>, F. Blengio<sup>1</sup>, R. Versteeg<sup>3</sup>, A. Eggert<sup>4</sup>, A. Schramm<sup>4</sup>, A. Garaventa<sup>1</sup>, C. Gambini<sup>1</sup>, L. Varesio<sup>1</sup>

<sup>1</sup>Laboratory of Molecular Biology, G. Gaslini Institute, Genoa, Italy

<sup>2</sup>Institute of Electronics, Computer and Telecommunication Engineering, Italian National Research Council, Genoa, Italy

<sup>3</sup>Department of Human Genetics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

<sup>4</sup>Department of Pediatric Oncology and Hematology, University Children's Hospital Essen, Essen, Germany

### Motivations

Cancer patient's outcome is written in part in the gene expression profile of the tumor. Clinical bioinformatic is instrumental in extracting the relevant gene clusters (signatures), the prediction models and the rules for patients stratification and clinical decision making. Hypoxia, a condition of low oxygen tension occurring in poorly vascularized tissues, has a profound effects on tumor growth and resistance to therapy. We utilized a novel biology driven approach coupled with appropriate feature selection [1] to identify the signature of hypoxic neuroblastoma cells (NB-Hypo) that stratifies neuroblastoma patients in good and poor outcome [2]. In the present work, we develop and validate a robust classifier to predict neuroblastoma patients' outcome on the bases of tumor hypoxia and we identify the most informative clinically relevant rules that combine classical risk factors and NB-hypo prognostic signature.

### Methods

Gene expression profiles of 182 neuroblastoma tumors were used to develop and validate our prediction models. Validation was performed either by leave one out cross validation or 66% split of the dataset in training and testing. We utilized a Multi Layer Perceptron (MLP), a feedforward Artificial Neural Network classifier, to predict patients' outcome (alive or dead 5 years after diagnosis). Integrated analysis of classical risk factors and prognostic NB-hypo signature utilized either C4.5, an efficient algorithm used to generate decision tree or RuleX 2.0, a software suite capable of building Intelligent Learning Machines (ILM) through Shadow Clustering (SC) [3]. The classifiers were trained and tested on the expression values of the 62 probsets constituting NB-hypo signature. Furthermore, in some instances NB-hypo was condensed into a single binary attribute by

means of an unsupervised k-means clustering of the patients' cohort based on the NB-hypo, 62 probsets expression values.

### Results

The NB-hypo classifier based on MLP predicted the outcome with an accuracy of 87% and was able to correctly predict poor outcome of all patients in the low-intermediate risk classes. The accuracy increased when NB-hypo classifier predicted the hypoxic state of neuroblastoma tumors even when it was not associated with poor outcome. Thus, NB-hypo classifier, while probing the hypoxic status of the tumor, is a new and robust predictor of neuroblastoma patient's outcome with very low error rate that decreases to negligible levels in localized tumors. Clinicians utilize established risk factors ( tumor stage, amplification of the MYCN oncogene or age at diagnosis) for prognosis and treatment choice. We investigated whether NB-hypo could be successfully added to the classical decision process improving risk definition and whether the relevant rules could be expressed in a clinically applicable form. We choose to utilize decision tree and ILM algorithms to facilitate the extraction of clinically applicable rules. Decision tree analysis revealed interesting relationships among risk factors and pointed to NB-hypo as the key element in identifying poor prognosis patients in stage 3 neuroblastoma. However, the divide and conquer approach employed by decision tree was unsatisfactory for a global study of these variables because of the excessive fragmentation of the database into small, poorly indicative groups of patients. In contrast, the covering algorithm adopted by SC identified a set of 11 rules each with a coverage greater than 30% and with less than 0.1% error. NB-hypo was integral part of these rules either as representative probsets or as a single binary attribute. Interestingly, the algo-

rithm divided the expression values of individual probsets in broad (generally two), statistically different, categories corresponding to clearly identifiable low and high expression levels. This conversion was critical for counteracting the variability microarray experiment data.

The importance of NB-hypo for outcome prediction of low risk neuroblastoma patients was confirmed. Moreover, we established rules for outcome prediction of stage 4 neuroblastoma patients that are very heterogeneous and difficult to stratify. In summary, we found that biology-based gene expression signatures and machine learning lead to patients outcome prediction and that appropriate clinical bioinformatic approaches can extract relevant rules translatable to the clinical setting. This approach was devel-

oped for neuroblastoma tumors but the rationale and methodology can be applied successfully to other types of cancer.

## References

1. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L. The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. *BMC Genomics*. 2009 Oct 15;10:474.
2. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Versteeg R, Caron HN, Molenaar JJ, Ora I, Eva A, Puppo M, Varesio L. A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients. *Mol Cancer*. 2010 Jul 12;9:185.
3. Muselli M and Ferrari E. Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction. *IEEE Trans. on Knowl. and Data Eng.* 23, 1, 2011.

## Regularized network-based algorithm for predicting gene functions with high-imbalanced data

M. Frasca✉, A. Bertoni, G. Valentini

Department of Computer Science, University of Milan, Italy

### Motivations

The gene function prediction problem is a real-world problem consisting in finding new bio-molecular functions of genes/gene products and characterized by hundreds or thousands of functional classes structured according to a predefined hierarchy.

This problem can be formalized as a semi-supervised multi-class, multi-label classification problem where the biological functions of new

genes can be predicted by exploiting their connections with genes whose biological functions are known.

Many different approaches have been proposed to address this problem, including “guilt-by-association” [1], “label propagation” [2], module-assisted techniques [3], SVMs [4]. Nevertheless, these methods usually suffer a decay in performance when input data are highly unbalanced, that is positive examples are sig-

Dataset	F-score		Method
	Level 4	Level 5	
Expr	0,052	0,033	COSNet
	0,095	0,071	R-COSNet
	0,043	0,017	LP-Zhu
	0,032	0,015	SVM
PPI-BG	0,363	0,281	COSNet
	0,370	0,297	R-COSNet
	0,292	0,268	LP-Zhu
	0,132	0,130	SVM
Pfam	0,349	0,258	COSNet
	0,350	0,283	R-COSNet
	0,268	0,212	LP-Zhu
	0,051	0,028	SVM
Sim-SW	0,323	0,265	COSNet
	0,333	0,279	R-COSNet
	0,254	0,239	LP-Zhu
	0,050	0,023	SVM

Figure 1. Average F-score across levels 4 and 5 of the FunCat classes using four data sources of *S.cerevisiae* organism: Pfam (protein domain data obtained from the Pfam data base), Expr (gene expression data from Spellman and Gasch experiments), PPI-BG (protein-protein interaction data obtained from the BioGRID databases), and Sim-SW (sequence similarities obtained through Smith Waterman algorithm).

nificantly less than negatives. This scenario characterizes in particular the most specific classes of the ontology, which are the classes more far from the root classes and that better describe the functions of genes.

### Methods

To address these items, we propose a regularization of a Hopfield-based cost-sensitive algorithm, COSNet, recently proposed to predict gene functions [5]. This algorithm, although designed to manage the imbalance in labeled data, tends to predict an excessively high proportion of positives when data are particularly unbalanced (that is in particular on most specific classes). By adding a term to the energy function of the network, we are able in modifying the dynamics in order to prevent the number of positives becomes too large. This energy term is minimized when the proportion of positive neurons (current positive rate) resembles the rate of positive labels in the training set (expected positive rate). The higher the difference between current and expected positive rates, the more the penalty to the energy function. We call this regularized version R-COSNet.

### Results

We tested R-COSNet on the prediction of yeast genes, by using four different data sets and the classes of the FunCat ontology [6]. This ontology is structured in forest of trees, in which each node

belong to one of the six levels of specificity. Level 1 refers to the root nodes, level  $i$  to nodes at distance  $i$  from the root. The considered classes are those with at least 20 positives and are spanned from level 1 to level 5. We compared our methods with a label propagation algorithm, LP-Zhu [2], and Support Vector Machine (SVM) with probabilistic output [4].

In Figure 1 we report the results in terms of F-score averaged across the functional classes belonging to the level 4 and level 5 of the hierarchy.

### References

1. Oliver, S. Guilt-by-association goes global. *Nature* 2000, 403: 601-603.
2. Zhu, X, Ghahramani, Z, and Lafferty, J. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003*, 912-919.
3. Sharan, R, Ulitsky, I, and Shamir, R. Network-based prediction of protein function. *Molecular Systems Biology* 2007, 3:88.
4. Lin, HT, Lin, CJ, Weng, R. A note on platt's probabilistic outputs for support vector machines. *Machine Learning* 2007, 68(3): 267-276.
5. Bertoni, A, Frasca, M, Valentini, G. Cosnet: A cost sensitive neural network for semi-supervised learning in graphs. *ECML/PKDD (1) 2011, Lecture Notes in Computer Science*, 6911: 219-234.
6. Ruepp, A, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research* 2004, 32(18): 5539-5545.

## A novel biclustering algorithm for the discovery of meaningful biological correlations between miRNAs and mRNAs

G. Pio<sup>1</sup>, M. Ceci<sup>1</sup>, D. D'Elia<sup>2</sup>✉, C. Loglisci<sup>1</sup>, D. Malerba

<sup>1</sup>Department of Computer Science - University of Bari, Bari, Italy

<sup>2</sup>CNR, Institute for Biomedical Technologies, Bari, Italy

### Motivations

microRNAs (miRNAs) are post-transcriptional regulators which represent one of the major regulatory gene families in animals, plants and viruses and that plays a key role in almost all main cellular processes. The computational prediction of miRNA target genes is important for the functional annotation of genomes and, on the other side, functional annotation of target genes can be of great help in suggesting specific biological functions of miRNAs [1]. This work aims to contribute to the elucidation of miRNAs role in the regulation of gene expression, by proposing a method for the hierarchical and overlapping biclustering of miRNAs and target messenger RNAs (mRNAs). The method allows to discover possible miRNA:mRNA functional relationships, at different granularity levels, in large datasets produced by miRNA target site prediction algorithms, thus reducing the impact of noise on the significance of the resulting biclusters.

### Methods

In order to properly work on miRNA:mRNA interactions, three important issues have to be considered. In particular, extracted biclusters should be: i) possibly overlapping, because miRNAs can be involved in different regulatory networks; ii) exhaustive, i.e. each miRNA or mRNA should belong to at least one bicluster, thus preventing the loss of possible co-regulations; iii) hierarchically organised, thus facilitating biological interpretability of results even when a high number of biclusters is extracted from large miRNA:mRNA datasets.

We propose an algorithm for the efficient discovery of overlapping, exhaustive and hierarchically organised biclusters. Our algorithm effectively deals with a kind of "relational" imbalance (i.e. miRNAs and mRNAs participate with significantly different cardinalities in the interactions). Moreover, it combines the notions of bicluster separability and bicluster distance in order to

extract significant biclusters according to both density and distance-based criteria.

The performance of our method is evaluated on miRNAs target predictions in the human genome dataset extracted by miRNAMap 2.0 [2]. The strength of the miRNA:mRNA interactions is estimated on the basis of predictions provided by miRanda, RNAHybrid and TargetScan algorithms. It is computed as a linear combination of three criteria: 1) the number of algorithms which predict the miRNA target site; 2) the number of miRNA target sites found in the same UTR region; 3) the accessibility of the target site. Weights of the linear combination are selected such that criterion 1 dominates over the other two, while criterion 3, *ceteris paribus* on criterion 1, dominates over criterion 2.

### Results

The performance of our method is evaluated in terms of execution time, bicluster compactness (the intra-bicluster cohesion) and bicluster co-regulation (the ability to group together miRNAs that target the same mRNAs). A comparative analysis shows that our method is able to extract a smaller number of (hierarchically organised) biclusters, with higher compactness and co-regulation values than ROCC [3]. Execution times of our method and ROCC are comparable. The significance of the extracted hierarchies is evaluated in terms of the F-Measure, on a set of synthetic datasets generated at different levels of noise. The analysis reveals that the hierarchy structure is almost correctly discovered, even for high levels of noise.

The effectiveness of the algorithm in extracting biologically related biclusters is tested on the basis of: i) classification of biclustered miRNAs in the same miRNA family or gene cluster; ii) validated functional associations of biclustered miRNAs reported in the literature and in major web specialised resources, iii) GO classification and functional clustering of biclustered mRNAs [4]. Results show that the proposed algorithm allows

to extract a relatively small number of biclusters that preserve both compactness and co-regulation. Biclusters extracted by our method represent meaningful biological correlations between miRNAs and mRNAs.

#### Availability

<http://dl.dropbox.com/u/66737165/biclustering/index.html>

#### Competing interests statement

Domenica D'Elia is on the Editorial Board of the EMBnet.journal

## References

1. Grun D, Wang YL, Langenberge D, Gunsalus KC, Rajewsky N (2005) microRNA Target Predictions across Seven *Drosophila* Species and Comparison to Mammalian Targets. *PLoS Comput Biol.* 1(1)
2. Hsu SD, Chu CH, Tsou AP, Chen SJ, Chen HC, Hsu PW, Wong YH., Chen YH, Chen GH, and Huang HD (2008) miRNAMap2.0: genomic maps of microRNAs in meta-zoan genomes, *Nucleic Acids Res.* 36
3. Deodhar M, Gupta G, Ghosh J, Cho H, Dhillon IS (2009) A scalable framework for discovering coherent coclusters in noisy data, *ICML*
4. Huang DW, Sherman BT and Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources, *Nature Protoc.* 1(4), 44-57

## A bioinformatics framework for the identification of active regulatory elements through the integrative analysis of high-throughput genomic data

D. Malagoli Tagliacruzchi<sup>1</sup>, A. Miccio<sup>1</sup>, A. Cavazza<sup>1</sup>, V. Poletti<sup>1</sup>, C. Peano<sup>2</sup>, G. De Bellis<sup>2</sup>, F. Mavillo<sup>2</sup>, S. Bicciato<sup>1</sup>✉

<sup>1</sup>Center for Genome Research, University of Modena and Reggio Emilia, Modena, Italy

<sup>2</sup>Institute of Biomedical Technologies, CNR, Milano, Italy

### Motivations

High-throughput technologies as microarray, Cap Analysis of Gene Expression (CAGE), and chromatin immunoprecipitation (ChIP), coupled with next generation sequencing (NGS) allow the identification of the molecular mechanisms regulating DNA transcription. In particular, CAGE and chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) are particularly suited to analyze transcriptional regulation and DNA methylation and histone modification patterns, while gene expression microarrays allow assessing transcriptional patterns. All these technologies produce data that are highly informative per se, but merging and integrating the various types of information would help answering many long-standing questions related to fundamental mechanisms of gene regulation and genome utilization. Data integration can be addressed using three main approaches, i.e., reduction of data complexity, unsupervised integration, and supervised integration [1]. Although these methods proved their efficiency in the analysis of single types of data, no bioinformatics tool integrates them all in a unified pipeline. In this context, we present a bioinformatics framework for the integrative analysis of CAGE, ChIP-Seq, and microarray data and the genome-wide identification of active regulatory elements. The computational workflow comprises three steps, i.e., i) stand-alone analysis of CAGE, ChIP-Seq, and microarray data; ii) construction of a unified data structure where results from CAGE, ChIP-Seq, and microarray analysis converge to generate integrated patterns of genomic signals; and iii) analysis of the integrated patterns to identify active regulatory elements.

### Methods

CAGE data have been analyzed using standard methods [2]. ChIP-seq peak calling was obtained using SICER [3]. Microarray data have

been processed using PREDA for the detection of chromosomal patterns of gene expression [4]. Results from stand-alone analyses have been integrated in a unique matrix based on the UCSC BED file structure and on the chromosomal coordinate of genomic elements. The analysis of the integrated matrix has been performed using an ad-hoc procedure coded in R for the identification of i) enhancers, ii) non-coding RNA, and iii) promoters. The algorithm comprises three steps: in the first, CAGE and ChIP-seq peaks are merged to determine which histone modifications are present in correspondence of a CAGE peak. CAGE peaks that fall within H3K4me1 regions are classified as associated with a putative enhancer, while peaks that fall into both H3K4me1 and H3K4me3 regions are classified as associated with a putative promoter. In the second step, the algorithm verifies if there exists bi-directional transcription between two consecutive CAGE peaks located on opposite strands. Briefly, given any two peaks on opposite strands, an R function calculates the distance between them and labels as bi-directionally transcribed those peaks lying at a distance  $\leq 1$  kb. All other peaks are marked as mono-directionally transcribed. In the last step, the presence of putative enhancers and promoters is linked to the local pattern of gene expression.

### Results

The pipeline has been applied for the genome-wide characterization of mono-directionally and bi-directionally transcribed promoters and enhancers involved in self-renewal, commitment, and differentiation of human stem/progenitor cells and their progeny. Specifically, we analyzed data from cord-blood derived hematopoietic stem cells (HSCs) and lineage-restricted erythroblasts and myelomonocytes identifying specific regulatory regions differentially engaged during HSC differentiation.

## Acknowledgements

This work is supported by FIRB – Programma “Futuro in Ricerca” 2010 (RBF100S4G) and Progetto Bandiera Epigenomica.

## References

1. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet.* 2010 Jul;11(7):476-86
2. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P. CAGE: cap analysis of gene expression. *Nat Methods.* 2006 Mar;3(3):211-22
3. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics.* 2009 Aug 1; 25(15):1952-8
4. Ferrari F, Solari A, Battaglia C, Bicciato S. PREDA: an R-package to identify regional variations in genomic data. *Bioinformatics.* 2011 Sep 1;27(17):2446-7

## A greedy and stochastic algorithm for multiple local alignment of interaction networks

G. Micale<sup>1</sup>, A. Pulvirenti<sup>2</sup>✉, R. Giugno<sup>2</sup>, A. Ferro<sup>2</sup>

<sup>1</sup>Department of Computer Science University of Pisa, Pisa, Italy

<sup>2</sup>Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy

### Motivations

A central problem in biological network analysis is the Local Network Alignment. The aim is to detect conserved subnetworks or complexes of proteins, across two or more species, which are involved in processes or functions. This allows to predict either new interactions or the functions of unknown proteins. Since the problem of finding conserved subnetworks in a set of networks is related to subgraph isomorphism, which is known to be NP-Hard, several heuristics have been proposed. These include PathBlast [1] and MaWISH [2] for pairwise local alignment and NetworkBlast-M [3] and Graemlin [4] for the multiple case. Although NetworkBlast-M has been proved to be the most efficient and accurate method, it is able to find significant conserved complexes composed by no more than 15 proteins. Here we introduce GASOLINE (Greedy And Stochastic algorithm for Optimal Local alignment of Interaction Networks) an algorithm based on Gibbs Sampling [5] in connection to a seed-extend approach to search for significant conserved complexes of any size.

### Methods

The algorithm consists of two main phases. In the first phase, we look for ortholog proteins across the networks. We call these proteins seeds of the suboptimal pattern. In the second phase, called iterative phase, we extend each seed, by adding one adjacent node. Here, through a stochastic process based on Gibbs Sampling, we choose a node among a set of randomly picked adjacent ones. The chosen nodes will be those that maximize similarity among the N extended seeds. We repeat the iterative phase until we obtain a set of N conserved subgraphs each of size W. These N subgraphs represent our final alignment. The topological density and the conservation of a complex are measured through an Index of Density and Structural Conservation (hereafter IDSC). The IDSC score ranges from 0 to 1 and it is dynamically computed during the iterative

phase. The iteration will terminate when IDSC is above a fixed threshold. This allows the removal of parameter W producing a set of highly dense and conserved complexes of different sizes.

### Results

GASOLINE has been tested on 10 microbial PPI networks, taken from Graemlin [4] and on 6 eukaryotic PPI networks, taken from STRING database [6]. The number of proteins in microbial networks ranges from 1,000 to 7,000 with the amount of interactions ranging from 13,000 to 230,000, whereas the size of eukaryotic networks ranges from 6,000 to 12,500 proteins and from 26,000 to 166,000 edges. 2000 of execution of GASOLINE have been performed. As output we consider the best distinct (not overlapping) complexes, with respect to size and IDSC score. In our experiments we selected complexes of at least 5 proteins with IDSC score > 0.7. All tests have been performed on an Intel Core i5-2500 3.30Ghz CPU with 4 GB RAM.

The complexes computed by the algorithm have been then validated by annotating their proteins with GO categories. For the microbial networks, annotations have been downloaded from DAVID [7,8], while eukaryotic networks proteins have been annotated using BioDBNet [9]. Significant conserved categories have been obtained by computing a p-value (< 0.0001), based on hypergeometric distribution. A GO category has been considered conserved when it resulted significant in at least N-1 species, where N is the number of aligned networks.

The executions of GASOLINE on microbial networks revealed the existence of a big conserved complex of 40 proteins, forming the large and small subunits of ribosome, with IDSC equals to 0.755 (see Tab. 1). As for the 6 eukaryotic networks, 15 conserved complexes have been found by GASOLINE with IDSC greater than 0.75. They are listed in Tab. 2, with their IDSC and the number of GO categories enriched.

Significant GO categories	GASOLINE (Size = 40)	NetworkBlast-M (Size = 15)
GO:0003735 (structural constituent of ribosome)	$1.887 \times 10^{-16}$	$1.11 \times 10^{-16}$
GO:0005198 (structural molecule activity)	$1.776 \times 10^{-16}$	$9.992 \times 10^{-17}$
GO:0003723 (RNA binding)	$1.665 \times 10^{-16}$	$1.776 \times 10^{-16}$
GO:0019843 (rRNA binding)	$1.443 \times 10^{-16}$	$7.771 \times 10^{-17}$
GO:0006412 (translation)	$8.882 \times 10^{-17}$	$5.329 \times 10^{-16}$

**Table. 1: Significant GO for microbial networks in GASOLINE and NetworkBlast-M**

Complex	Size	IDSC	# GO categories enriched
Small and large subunit of ribosomes	43	0.716	7
Proteasome	32	0.847	10
Spliceosome	26	0.701	5
DNA-directed RNA polymerase	19	0.789	9
Small subunit (SSU) processome	15	0.832	2
Chaperonin-containing T-complex	13	0.737	4
V-ATPase	11	0.78	11
Exosome (RNase complex)	10	0.878	5
Replication fork protection	7	0.984	4
DNA replication factor C	7	0.928	5
Mitochondrial respiratory chain complex III	7	0.889	4
Arp2/3 protein complex	7	0.722	3
Translation initiation factor 2/2B	6	0.855	4
Cdc73/Paf1 complex	6	0.8	1
Endosomal sorting complex required for transport (ESCRT-III)	5	0.967	1

**Table. 2: Conserved complexes found by GASOLINE in the 6 eukaryotic networks**

## References

1. Kelley BP, Sharan R, Karp R, Sittler ET, Root DE, Stockwell BR, and Ideker T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A* 100, 11394-9, 2003.
2. Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, and Grama A, Pairwise alignment of protein interaction networks, *Journal of Computational Biology*, 13(2), 182-199, 2006
3. Kalaev M, Bafna V, and Sharan R. Fast and accurate alignment of multiple protein networks. *Journal of computational biology*, 16(8), 989–999, 2009.
4. Flannick J, Novak A, Srinivasan B, McAdams H, and Batzoglou S. Graemlin: general and robust alignment of multiple large interaction networks. *Genome research*, 16(9), 1169, 2006
5. Geman S, Geman D. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (6): 721–741, 1984
6. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguéz P, Doerks T, Stark M, Müller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39 (Database issue):D561-8. Epub 2010 Nov 2, 2011.
7. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* 4(1):44-57, 2009.
8. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37(1):1-13, 2009
9. Mudunuri U, Che A, Yi M, and Stephens RM. BioDBnet: the biological database network. *Bioinformatics.* 25(4): 555–556, 2009

## GAM: Genomic Assemblies Merger

A. Policriti<sup>1,2</sup>, S. Scalabrin<sup>1</sup>, F. Vezzi<sup>3</sup>, R. Vicedomini<sup>2</sup>✉

<sup>1</sup>Applied Genomics Institute, Udine, Italy

<sup>2</sup>DIMI, University of Udine, Udine, Italy

<sup>3</sup>KTH: Royal Institute of Technology, SciLife Lab Stockholm, Sweden

### Motivations

In the last 3 years more than 20 assemblers have been proposed to tackle the hard task of assembling. Recent evaluation efforts (Assemblathon 1 and GAGE) demonstrated that none of these tools clearly outperforms the others. However, results clearly show that some assemblers perform better than others on specific regions and statistics while poorly performing on other regions and evaluation measures.

With this picture in mind we developed GAM (Genomic Assemblies Merger) whose primary goal is to merge two or more assemblies in order to obtain a more contiguous one. Moreover, as a by-product of the merging step, GAM is able to correct mis-assemblies.

GAM does not need global alignment between contigs, making it unique among other Assembly Reconciliation tools. In this way a computationally expensive alignment is avoided, and paralog sequences (likely to create false connection among contigs) do not represent a problem.

GAM procedure is based only on the information coming from reads used in the assembling phases, and it can be used even on assemblies obtained with different datasets.

### Methods

Let us concentrate on the merging of two assemblies, dubbed M and S. As a preprocessing step, that is an almost mandatory analysis, reads (or a subset of them) used in the assembling phase are aligned against M and S using a SAM-compatible aligner (e.g., BWA, rNA).

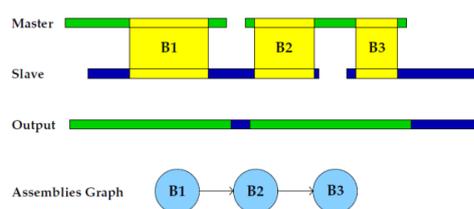
GAM takes as input M, S and the two SAM files produced in the preprocessing step. The main idea is to identify fragments belonging to M and S having high similarity. For this purpose, GAM identifies regions, named blocks, belonging to M and S that share an high enough amount of reads (i.e. regions sharing the same aligned reads).

After all blocks are identified the Assembly Graph (AG) is built: each node corresponds to a block and a directed edge connects block A to block B if the first precedes the second in either M or S (see Fig.1).

Once AG is available, the merging phase can start. As a first step GAM identifies genomic regions in which assemblies contradict each other (loops, bifurcations, etc.). These areas represent potential inconsistencies between the two sequences. We chose to be as much conservative as possible electing (for example) M to be the Master assembly: all its contigs are supposed to be correct and cannot be contradicted. S becomes the Slave and everywhere an inconsistency is found, M is preferred to S.

After the identification and the resolution of problematic regions, GAM visits the simplified graph, merges contigs accordingly to blocks and edges in AG (each merging phase is performed using a Smith-Waterman algorithm variant) and finally outputs the new improved assembly.

GAM is not only limited to contigs, it can also work with scaffolds, filling the N's inserted by an assembler and not by the other.



Genome	Tool	Length	Contigs	L50 (bp)
Olea (chloroplast)	CLC-III	127,942 bp	10	16,215
	CLC-454	128,572 bp	9	15,993
	GAM	130,101 bp	3	112,156
Populus trichocarpa	CLC	339,551 Kbp	104,432	6,130
	ABYSS	526,633 Kbp	88,193	11,768
	GAM	441,133 Kbp	83,978	14,407
Boa constrictor	CLC	1,363 Mbp	373,909	7,716
	ABYSS	1,730 Mbp	7,042,239	2,348
	GAM	1,372 Mbp	367,060	8,031

Figure 1

### Results

GAM has been tested on several real datasets, in particular on *Olea*'s chloroplast (241X Illumina paired reads and 21X 454 paired reads), *Populus trichocarpa* (82X Illumina paired reads), *boa constrictor* (40X Illumina paired reads). Illumina reads have average length of 100 bp and insert size of 500 bp. All tests have been performed on a computer equipped with 8 cores and 32GB RAM. ABySS and CLC were selected as assemblers. Results are summarized in Fig. 1.

*Olea*'s chloroplast has been used as a proof of concept experiment. The presence of a refer-

ence sequence allowed GAM's output validation (using *dnadiff*). Two assemblies were obtained with CLC using Illumina and 454 data. GAM was used to merge them. Figure 1 shows how GAM assembly is not only more contiguous but also more correct: while Master (CLC-Illumina) and Slave (CLC-454) have 58 and 39 suspicious regions respectively, GAM has only 14 of those.

On *Populus trichocarpa* and *Boa constrictor*, CLC assemblies were used as master due to their better contiguity. In both cases assemblies returned by GAM were more contiguous (see Fig. 1).

## SARMA: a web resource for species assignment of high-throughput sequencing reads from metagenomics analysis

M. D'Antonio<sup>1</sup>✉, D. Paoletti<sup>2</sup>, M. Santamaria<sup>3</sup>, T. Castrignanò<sup>2</sup>, G. Pesole<sup>1,4</sup>

<sup>1</sup>Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli Studi di Bari, Bari, Italy

<sup>2</sup>CASPUR, Consorzio interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy

<sup>3</sup>Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

### Motivations

The exceptional development of next generation sequencing platforms (NGS) has opened unprecedented possibilities for the comprehensive investigation of entire microbial and viruses communities at taxonomic and functional level in environmental and clinical samples. In the latter case, the characterization of human microbiome is a crucial achievement to fully understand the regulation of gene expression in physiological and pathological conditions. The computational approach to viral and microbial discovery is based on the premise that the assayed tissues contain a small viral and microbial component that can be detected and analyzed after the subtraction of the large excess of human sequences. Rough subtractive processes can however lead to problems due to regions in the host genome derived from previous or ancient infections (a typical example is the presence of endogenous retrovirus).

### Methods

To address this issue we developed a user-friendly web resource, named SARMA (Species Assignment of Reads from Metagenomic Analysis), using as input one or more metagenomic datasets. The SARMA workflow serially subtracts input reads mapping to human sequences using both ultra fast (i.e. BWA) and sensitive (i.e. BLASTN) read aligners. To avoid false positives in this phase we utilize a custom built human genome obtained filtering out all regions derived from viruses and bacteria. After the subtractive process, any remaining unmapped reads may represent candidate non-human

microbial-derived species. Only top alignments fulfilling appropriate thresholds are considered with reads inheriting the taxonomy attributes of the corresponding aligned sequence. A statistical methodology, taking into account the alignment quality, is then adopted for species assignment, allowing also for weighted partitioned assignments. Unassigned reads, possibly deriving from novel organisms, can be then assigned to higher taxonomic ranks using other publicly available tools such as MEGAN. SARMA provides a WEB interface allowing job submission and results browsing. The SARMA output consists of several tabular and tree-based visualizations of the results allowing users to browse their data from high-level summaries down to the more detailed views. Reads shared between host and other organisms are highlighted by crosslinks. Results deriving from two or more datasets can be differentially compared to identify over- and under-represented species.

### Results

SARMA has been tested using a public dataset of Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma [1]. A sample manually contaminated with a known family of virus should be a perfect training set for a resource like SARMA. The aim of this test were, of course, determining the presence of expected viruses.

### References

1. Feng, H, Shuda, M, Chang, Y, and Moore, PS. Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma. *Science* 2008, 319 (5866): 1096-1100. doi: [10.1126/science.1152586](https://doi.org/10.1126/science.1152586)

## Identification of gene annotations and interactions and protein-protein interaction associated disorders through data integration

A. Canakoglu, P. Gangi, S. Gennaro, M. Masseroli✉

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy

### Motivations

Available biomolecular annotations are very valuable, but dispersed and far from being complete. High quality integration of scattered annotation data and reliable identification of new annotations can greatly support unveiling new biomedical knowledge. Biomolecular interactions are one of the main objectives of biomolecular studies, due to the understanding that biological processes are mainly driven by interactions among biomolecular entities, such as proteins and DNA. New powerful high-throughput experimental techniques are providing numerous protein-protein interaction data; they are being collected, together with computational results, in several different databases, which include IntAct, BioGrid, BIND, DIP, HPRD and MINT. Yet, they generally do not include phenotypic or even functional or structural information about the interactors, which in many cases are available in other databases. In particular, no information is available about the association of protein-protein interactions with genetic disorders. This creates the need to integrate the sparsely available data in order to enrich the identified interactions with additional evidence, support their biological interpretation and identify their involvement in inherited pathologies. In addition, integration of gene and protein annotations and protein-protein interaction data provides the base to infer new gene annotations and interaction networks.

### Methods

We developed an automatic association inference method, based on the transitive closure concept, and applied it on the data from several distributed sources integrated in our Genomic and Proteomic Data Warehouse (GPDW). In particular, by leveraging protein-protein interaction data, provided by the IntAct and MINT databases, and protein encoding gene data from the Entrez Gene database, we inferred gene interaction networks. In addition, by taking advantage of genetic disorder and phenotype data provided by the OMIM database, we inferred asso-

ciations between proteins and genetic disorders and their phenotypes. Then, in order to identify genetic disorders possibly associated with protein-protein interactions, we looked for those interacting proteins that resulted associated with the same genetic disorder.

### Results

Our GPDW currently includes 46,154 human protein-protein interactions (out of the 254,048 protein-protein interactions contained), which involve 12,178 different human proteins (out of the 326,766 human proteins in the GPDW) that are encoded by 11,232 different human genes. By applying the above described method, we identified 1,130 gene networks and found 1,136 human protein-protein interactions associated with 628 genetic disorders (6% of all genetic disorders in the GPDW), which are related to 86 clinical synopses (87% of all clinical synopses in the GPDW) and 3,481 phenotypes (10% of the total phenotypes in the GPDW). Among others, we found four interacting proteins (AHSA1\_HUMAN, CFTR\_HUMAN, DERL1\_HUMAN, and RNF5\_HUMAN) whose encoding genes are known to be associated with cystic fibrosis, as well as other 43 genes. Mutations of the CFTR human gene are known to be directly involved in different grades and manifestations of cystic fibrosis. The found associations of the AHSA1\_HUMAN, CFTR\_HUMAN, DERL1\_HUMAN, and RNF5\_HUMAN interacting proteins with cystic fibrosis could suggest that some types of this multi-variant disorder may be associated with defects in the interactions between these proteins. Possibly, different CFTR\_HUMAN protein mutations could alter its functional interaction with one or more of the AHSA1\_HUMAN, DERL1\_HUMAN and RNF5\_HUMAN proteins. If this would be proven, such finding would also suggest, as a possible disease treatment strategy, the engineering of a synthetic protein interacting with the mutated CFTR\_HUMAN protein and similar in function to the one of the AHSA1\_HUMAN, DERL1\_HUMAN or RNF5\_HUMAN proteins whose interactions with the mutated CFTR\_HUMAN pro-

tein result altered. The above discussed findings demonstrate the importance of the transitive closure based inference method developed, as well as of the data integration approach implemented in the GPDW and the relevance of the comprehensive data there integrated. The GPDW constitutes the backend of a Genomic

and Proteomic Knowledge Base (GPKB) that is publicly available at <http://www.bioinformatics.dei.polimi.it/GPKB/> through a prototype easy-to-use and efficient Web interface.

**Availability**

<http://www.bioinformatics.dei.polimi.it/GPKB/>

## HUPHO: the human phosphatase portal

S. Liberti<sup>1</sup>✉, A. Calderone<sup>1</sup>, F. Sacco<sup>1</sup>, L. Perfetto<sup>1</sup>, M. Iannuccelli<sup>1</sup>, S. Panni<sup>2</sup>, E. Santonico<sup>1</sup>, A. Palma<sup>1</sup>, A.P. Nardoza<sup>1</sup>, L. Castagnoli<sup>1</sup>, G. Cesareni<sup>3</sup>

<sup>1</sup>Department of Biology, University of Rome Tor Vergata, Rome, Italy

<sup>2</sup>Department of Biology, University of Rome Tor Vergata, Rome, Department of Cell Biology, University of Calabria, Rende, Italy

<sup>3</sup>Department of Biology, University of Rome Tor Vergata, Rome, Research Institute Fondazione Santa Lucia, Rome, Italy

### Motivations

Phosphatases, together with kinases, contribute to the regulation of protein phosphorylation homeostasis in the cell. Phosphorylation is a key post translational modification underlying the regulation of many cellular processes. Thus a comprehensive picture of phosphatase function and the identification of their target substrates would aid a systematic approach to a mechanistic and holistic description of cell signalling. In this paper we report a web site designed to facilitate the retrieval of information about human protein phosphatases.

### Methods

We developed a web server, named HuPho (Human Phosphatases), and a search engine to recover information that has been annotated in several publicly available web resources. From the web site interface, scientists can dynamically access a number to different repositories where public data about protein domain composition, aminoacidic sequence, 3D structure,

protein-protein interactions, pathways annotation, annotation for disease and more are stored. The protein-protein interaction data have been further explored and integrated to identify phosphoproteins enzymatic substrates. The utilization of a programmatic access to the data sources avoid updating issues on data, that as a result are always aligned to the latest release.

### Results

We provided the scientific community with a resource that can be easily interrogated to obtain structural and functional information about human protein phosphatases. Much of the information is retrieved by interrogating the web services of publicly available repositories. This information is integrated with data specifically curated by our group and stored in our internal database. Finally the information is organized to allow the user to browse the knowledge through a single pane of glass.

### Availability

<http://hupho.uniroma2.it/>

## A Knowledge Base for fish and fishery products

A.M. Pappalardo<sup>1</sup>, F. Guarino<sup>1</sup>, A. Messina<sup>1</sup>, A. Pulvirenti<sup>2</sup>, R. Giugno<sup>2</sup>, A. Ferro<sup>2</sup>✉, V. De Pinto<sup>1</sup>

<sup>1</sup>Dipartimento di Sc. Biologiche, G. e A., Università di Catania e INBB, Istituto Nazionale Biostrutture e Biosistemi, Roma, Italy

<sup>2</sup>Dipartimento di Biomedicina Clinica e Molecolare, Università di Catania, Italy

### Motivations

The species subject of our work belong to different orders of Teleosts, one of which is the order of Clupeiformes consisting in two large families of commercial interest (Clupeidae and Engraulidae), which include the small fish commonly known, respectively, for herring, sardines, shads, and sprats and anchovies. Their molecular identification is useful because a common fraud is their species replacement. In addition, because of globalization, it is possible to find in our markets also other species of anchovy typical of different distribution areas, such as Japanese anchovy (*Engraulis japonicus*), the Peruvian Anchovy (*E. ringens*), the Atlantic Ali (*E. anchoita*) and the Californian anchovy (*E. mordax*). Due to the differentiation of the processed foodstuffs that characterizes the market, it is essential to develop tools for unequivocal and quick identification of the species present in the market even when morphological identification is no longer possible.

### Methods

DNA barcoding consists of the utilization of short DNA sequences to identify organisms, in particular the species. The quest for a genetic marker useful to determine unambiguously the species is still a matter of debate. For animals, the best candidate to this role has been proposed to be the cytochrome oxidase I (COI) gene. In vertebrates the mtCOI gene is 1545 bp long, a region of 648 bp close to the beginning of the translated sequence is properly said to be the "barcode". We applied the DNA barcoding technologies, upon a segment of COI to compare fish belonging to Clupeiformes order and the analysis of 5'Dloop sequences to stock identification [1]. The fish was used for extracting genomic DNA, amplifying and sequencing. The sequences obtained were aligned using ClustalX. Relationships among the sequence obtained were examined using Neighbour-joining (NJ) and Bayesian analyses. The NJ tree was constructed using pairwise

distances calculated following the application of Kimura's two-parameter (K2P) correction for multiple substitutions in MEGA v. 4.0. The robustness of internal branches of distance was estimated by bootstrapping with 1000 replicates. Modeltest v. 3.06 was used to select the most appropriate models of sequence evolution for the Bayesian analysis that was implemented in MrBayes v. 3.0 using a Metropolis-coupled, Markov Chain Monte Carlo (MCMC) sampling approach. The DNA Barcoding procedure produces a whole catalogue of I.D. of individual species or populations of fish. A key service for the utilization of analytical data is the organization of a BioBank collecting biological samples, DNA samples, genomic sequences connected also to a Knowledge base that represents sequence collected and is capable to interact with remote databases. The knowledge base will be equipped with a data mining module providing advanced tools to automatically extract new knowledge by highlighting predictive patterns of interest [2].

### Results

Our results allow the molecular identification of the fish (through the COI) also in its processed product, and in some case the identification of the geographical origin of the specific fish (through the 5'Dloop). Genetic analysis of the species of interest, are conducted through the innovative technique of DNA Barcoding alone or in combination with the high-resolution melting analysis (HRM) analysis, called Bar-HRM (High Resolution Melting-barcode DNA). This will allow to increase the knowledge of species not yet well studied, but of considerable importance for the Mediterranean area. The genetic analysis of the species of interest are exploited not only for the recognition of species-specific sequences, i.e. the ability to distinguish between two organisms belonging to different species, but also for determining the provenance of the fish (characterization of fish stock).

## References

1. A.M. Pappalardo, F. Guarino, S. Reina, A. Messina and V. De Pinto Geographically widespread swordfish barcode stock identification: a case study of its application (2011) PLOS ONE, doi: [10.1371/journal.pone.0025516](https://doi.org/10.1371/journal.pone.0025516)
2. A. Laganà, S. Forte, A. Giudice, M. R. Arena, P. L. Puglisi, R. Giugno, A. Pulvirenti, D. Shasha, Ferro A. (2009). miRò: a miRNA knowledge base. DATABASE, vol. Vol. 2009, bap008; p. \*, ISSN: 1758-0463, doi: [10.1093/database/bap008](https://doi.org/10.1093/database/bap008)

## Micro-Analyzer: a tool for automatic pre-processing of multiple Affymetrix arrays

P.H. Guzzi<sup>✉</sup>, M. Cannataro

Department of Medical and Surgical Sciences, University "Magna Græcia" of Catanzaro, Catanzaro, Italy

### Motivations

A current trend in genomics is the investigation of cell mechanism using different technologies in order to explain the relationship among genes, molecular processes and diseases on a different scale. For instance, the combined use of expression arrays and SNP arrays has been demonstrated as an effective instrument in clinical practices [1,3,4]. Consequently, in a single experiment different kind of microarrays may be used, resulting in the production of different types of binary data (images and raw data). The analysis of microarray data requires an initial preprocessing phase of raw data that makes them suitable for use on existing platforms, such as the TIGR M4 Suite. An additional challenge to be faced by emerging data analysis platforms is the ability to treat in a combined way such different microarray data coupled with clinical data. In fact resulting integrated data may include both numerical and symbolic data (e.g. gene expression and SNPs regarding molecular data), as well as temporal data (e.g. the response to a drug, time to progression, survival rate, etc., regarding clinical data). Raw data preprocessing is a crucial step in analysis but is often performed in a manual and error prone way using different software tools. Thus novel, platform independent, and possibly open source tools enabling the semi-automatic preprocessing and annotation of different microarray data are needed.

### Methods

The paper presents Micro-Analyzer (Microarray Cel file Summarizer), a cross-platform tool for the automatic normalization, summarization and annotation of Affymetrix expression and SNP data binary data. It represents the evolution of the  $\mu$ -CS tool, extending the preprocessing to SNP arrays that were not allowed in  $\mu$ -CS [2]. Using the tools made available by Affymetrix (e.g. apt-summarize and apt-genotype), the user needs to download from the Affymetrix web site the right preprocessing and annotation libraries, then needs to manually invoke such tools to obtain

preprocessed data and then has to import them into an external data analysis tools, e.g. TMEV. This approach presents numerous drawbacks, among those the need to manually perform all these tasks and the possibility to use the wrong or older libraries, obtaining wrong results, finally, data must be manually imported into analysis tools. To reduce such drawbacks we propose Micro-Analyzer. Micro-Analyzer is based on a client-server architecture. The Micro-Analyzer client is provided as a Java standalone tool and enables users to read, preprocess and analyse binary microarray data (gene expression and SNPs). It avoids: (i) the manual invocation of external tools (e.g. the Affymetrix Power Tools), (ii) the manual loading of preprocessing libraries, and (iii) the management of intermediate files. The Micro-Analyzer server automatically updates the references to the summarization and annotation libraries, hiding to the user the location of libraries and automatizing the process of updating such libraries when new versions of the microarray are released. By using Micro-Analyzer the user may preprocess both data using a single tool, retaining the advantage of storing in a single way both preprocessing results and metadata.

### Results

Micro-Analyzer users can directly manage Affymetrix binary data without worrying about locating and invoking the proper preprocessing tools and chip-specific libraries. Moreover, users of the Micro-Analyzer tool can load the preprocessed data directly into the well-known TM4 platform, extending in such a way even the TM4 capabilities. Consequently, Micro Analyzer offers the following advantages: (i) it reduces possible errors in the preprocessing and further analysis phases, e.g. due to the incorrect choice of parameters or due to the use of old libraries, (ii) it enables the centralized pre-processing of different arrays, (iii) it may enhance the quality of further analysis by storing the information about the preprocessing steps.

**Availability**

<http://sourceforge.net/projects/microanalyzer/>

**References**

1. André Koschmieder, Karin Zimmermann, Silke Trüßl, Thomas Stoltmann, Ulf Leser Tools for managing and analyzing microarray data Briefings in Bioinformatics, Vol. 13, No. 1. (01 January 2012), pp. 46-60, doi:10.1093/bib/bbr010.
2. Pietro Hiram Guzzi, Mario Cannataro: mu-CS: An extension of the TM4 platform to manage Affymetrix binary data. BMC Bioinformatics 11: 315 (2010)
3. [www.affymetrix.com](http://www.affymetrix.com)
4. Walker BA, Leone PE, Jenner MW, Li C, Gonzalez D, Johnson DC, Ross FM, Davies FE, Morgan GJ.

## IGI grid services for the bioinformatics community

L. Gaido<sup>1,2</sup>✉, M. Bencivenni<sup>2,3</sup>, D. Cesini<sup>2,3</sup>, G. Donvito<sup>2,4</sup>, P. Veronesi<sup>2,3</sup>

<sup>1</sup>INFN Torino, Italy

<sup>2</sup>IGI, Italy

<sup>3</sup>INFN CNAF, Italy

<sup>4</sup>INFN Bari

### Motivations

In the last decade many projects related to grids have been carried out in Europe at both national and international levels with an important economic contribution by the European Commission. The grid middleware developed within these projects has been deployed into the European grid infrastructure (EGI) made by more than 350 sites all over Europe. The Italian Grid Infrastructure (IGI) is part of EGI and is one of the most important and widest national grid infrastructures in Europe since it provides about 33000 CPU cores, 17 PB of disk space and 9 PB of tape capacity spread over more than 50 sites. Although the grid infrastructure has been initially built according to the needs of a few scientific communities (high energy physics, earth observation and Bioinformatics among the others), it has been gradually evolving in order to provide grid services to a wider and wider user base. Several scientific communities are observing that as the instruments become more and more powerful the need for storage and computing is increasing day by day. This will also increase the number of users that could benefit from a geographical distributed computing grid infrastructure.

### Methods

In order to support new communities (users and resource providers), various activities have been started. Training has been considered very important, so tutorial for users and grid administrators are regularly organized. In addition great effort has been devoted to understanding the user needs, by defining appropriate use cases, and to supporting the user communities to port their applications on the grid environment (the so-called "gridification"). An important effort is also spent in developing new tools that could make the interaction between the final users and the grid as easy as possible. In particular web tools to submit jobs to the grid infrastructure have been deployed and used by some bioinformatics com-

munities. In order to address the needs of users relying on high-level tools like Workflow managers (e.g. Taverna and other similar tools), a front-end web service has been developed. This web interface could be used as a bridge towards the EGI/IGI grid infrastructure. The IGI community is also providing a service that allows the exploitation of Relational Databases over the grid infrastructure, assuring a high level of security and privacy.

### Results

The usage of the standard EGI/IGI resources and services, together with the high level services that IGI is providing on top of the grid, has provided the end users with the capability of carrying out their high demanding computing activities in an easy and reliable way. In the past years, indeed, IGI has supported several bioinformatics communities to "gridify" many different applications such as: ASPic, PAML, MrBayes, CSTGrid, DNAFan, BLAST, BayeSSC, FT-Comar, Muscle, Gene Ontology DB analysis, ABCtoolbox, EMBOSS, Bowtie, SAMtools, Illumina Solexa data processing, AmpliconNoise, BioPython, HMMER. As a result the CPU consumed by various Italian bioinformatics groups on the IGI grid infrastructure has exceeded the 10 years in a few days of activity, thus hugely reducing the overall time needed for the execution of the jobs. Thanks to IGI the bioinformatics users have carried out their analysis in an easy and transparent way, both through simple web interfaces and through complex Workflow managers, reducing the time needed to get the results of about 2-3 orders of magnitude. The IGI infrastructure has also been exploited by the Computational Biology group of the Bologna University, in the frame of DUCK (Distributed Unified Computing for Knowledge, a collaboration between multidisciplinary Academic and Research Institutions located in the Emilia Romagna region), to run protein annotation application based on the Bologna Annotation Resource (BAR) method, and to perform massively parallel genome sequencing including about 18 millions of protein

sequences. More than 150 computing nodes in the IGI grid infrastructure have been used, successfully dropping the computational times if compared to the computing resources of the local cluster available to the group. The protein annotation application reached a dropping factor of 120, i. e. the computation was performed in few weeks instead of years. Data management facilities offered by the Grid were also exploited to easily handle input and output files. To provide

an easy-to-use service to the user communities IGI is developing a web portal that will hide the complexity of the authentication/authorization mechanisms and will also integrate the computing frameworks needed by the different user communities. A prototype of this portal can be easily set up for the bioinformatics community.

**Availability**

<http://www.italiagrid.it/>

## FERMI: the most powerful computational resource for Italian scientists

F. Falciano, E. Rossi 

Supercomputing Applications and Innovation Department, CINECA, Casalecchio di Reno (BO), Italy

By the last quarter of 2012 CINECA, the largest super-computing center in IT and one of the most important worldwide, will provide to all the scientific community one of the world fastest super-computer, the IBM BlueGene-Q, named FERMI to honor the famous Italian physicist. FERMI with its advanced technology and its 2.1 PFlop peak perform, will allow the Italian scientists for the fastest and most advanced technology available worldwide for scientific calculations. CINECA provides High Performance Computing (HPC) infrastructure for Italian and European researchers and operates in the technological transfer sector through high performance scientific computing, the management and development of networks and web based services, and the development of complex information systems for treating large

amounts of data. Moreover CINECA has been appointed by the Italian Ministry of Research and University to represent IT in the European HPC research infrastructure (PRACE). The aim of the HPC Department is to develop and promote technical and scientific services for the Italian and European research community. It constantly invests in Supercomputing technologies, thus being able to provide researchers with some of the hardware resources among the most powerful and energy efficient in Europe. Twice a year CINECA will directly award in excess of 40 million processor core hours, to ensure an adequate supply to scientists and engineers for HPC-related research, and thanks to the new supercomputer the core hours awarded will rise above 1 billion.

# Posters

---



## Devising and experimenting correlation-based metrics for evaluating the effectiveness of input encoding techniques in prediction tasks

G. Armano, E. Tamponi✉

Department of Electrical and Electronic Engineering, University of Cagliari, Italy

### Motivations

Defining an optimal encoding for input data is fundamental to achieve high performances in prediction tasks. Its main responsibility is to transform input data to a format suitable for the classification algorithm. The selection of the best encoding is typically done by resorting to the knowledge of a human expert, entrusted with extracting the features that s/he deems useful for the task. In some cases, the evaluation of an encoding can be performed by heavyweight tests, where most of the computational effort is spent to train classifiers. Furthermore, these tests may introduce a bias due to many factors, which depend on the adopted learning technique rather than on the encoding under analysis. To overcome these problems, we propose to investigate the correlation between the input variables (whose actual values depend on the adopted encoding technique) and the output variables (which encode the labels associated with each input). According to this insight, we devised and experimented some "correlation-based metrics" aimed at evaluating the encodings used for prediction tasks. The availability of efficient metrics would be particularly useful in domains where the selection of the right encoding is crucial, such as in the prediction of biomedical data. For example, in the prediction of protein secondary structure, different encodings greatly affect the overall performance of a system. In particular, multiple alignments of amino acid

sequences bring an exceptional performance increase with respect to one-hot encoding.

### Methods

We started our research by finding suitable methods for evaluating the input-output correlation. An obvious starting point was the Pearson product-moment correlation coefficient, but we found many more algorithms that resulted more suitable for our needs. In particular, we tested the novel distance correlation coefficient and the generalized correlation ratio. A fundamental part of the metrics is to extract a "synthetic value" from the correlation matrices obtained during the input-output correlation analysis. To evaluate our system, we selected a particularly difficult domain, the prediction of protein secondary structure. We measured the performance of 30 input encodings using our metrics. For each encoding, we compared the performance predicted by our metrics with the effective results obtained by actual prediction systems. To implement and test the system, we used the GAME framework.

### Results

The tests showed that our metrics predicted the performance of actual prediction systems with high accuracy. In particular, the metrics based on the generalized correlation ratio resulted very effective and fast (more than 1000 times faster than traditional performance tests).

### Availability

<http://iasc2.diee.unica.it/GAME/>

## Core algorithms to search in biological structured data

V. Bonnici<sup>1</sup>, R. Giugno<sup>2</sup>✉, A. Pulvirenti<sup>2</sup>, D. Shasha<sup>3</sup>, A. Ferro<sup>2</sup>

<sup>1</sup>Dept. Computer Science, University of Verona, Italy

<sup>2</sup>Dept. Clinical and Molecular Biomedicine, University of Catania, Italy

<sup>3</sup>Courant Institute of Mathematical Sciences, New York University, United States

### Motivations

The graph is a data structure to represent biological data ranging from molecules and proteins to biological networks and metabolic pathways. Working on those data involves mainly applying graph isomorphism algorithms. Those algorithms are computationally hard and their efficiency may depend upon the input graphs. We are building a library, SubGraphLib, of the most popular searching algorithms and benchmarks highlighting drawbacks, advantages, and best performance input cases for each method. A novel approach to find all occurrences of a query subgraph in a target graph is also proposed. This new method applies a search strategy which significantly reduces the search space without using any complex pruning rule. Results show a significant reduction of the running time with respect to other methods together with a scalable memory requirement.

### Methods

The best known algorithms to solve the subgraph isomorphism problem are the ones proposed by Ullmann [1] and by Cordella et al. [2] (VF2), which make use of backtracking algorithms in conjunction with some filtering rules to prune branches of the search space represented as a tree. The nodes of the tree denote pairs of matched vertices of the query and the target graphs, respectively. During the visit, the isomorphism conditions are applied to verify the partial matches. The algorithm in [1] modeled the graph isomorphism problem also as a constraint satisfaction problem (CSP). A CSP is defined by a set of variables and a set of constraints among them. To each variable a set of possible values, called domain, is associated. The solution of a given CSP problem is an assignment of values to all variables such that all constraints are satisfied. More recently, Solnon [3] published a method, LAD, for propagating global neighborhood constraints together with a generalized arc consistency. Ullmann [4] proposed a new method, FocusSearch, based on bitvector representation of domains, to deal

with parallel operations. In FocusSearch, domain reduction is not applied until convergence is achieved. The search phase is preceded by two steps based on vertex invariants and local AllDifferent constraints [3,4]. Search strategy is established by a static instantiation sequence based on the number of future branches. Our newly proposed algorithm, called CoreGraph, is based on a new search strategy which builds a static instantiation sequence of the query node. CoreGraph does not deal with complex filtering rule or domains. The basic idea for the construction of the search sequence is to maximize the number of branches to preceding nodes in the sequence. The sequence is recursively generated by adding those neighbors maximizing a score function. The score of each candidate node is assigned taking into account its degree, the number of its edges leading to nodes in the sequence and to their neighbors. Notice that, CoreGraph applies those filtering rules only to the query graph. Concerning the target graphs, the only information CoreGraph uses for pruning is node degree. Finally, since the search strategy does not give priority to more dense parts of the target graphs it results efficient in a large variety of query and target graphs.

### Results

SubGraphLib contains the original implementation of VF2, LAD, and CoreGraph and a new implementation of FocusSearch in C++ (which is originally distributed in modula2). All algorithms have been compared on benchmarks such as synthetic unlabeled graphs, molecules, and biological networks. CoreGraph and FocusSearch in all cases outperform the other algorithms in terms of execution time. In most benchmarks, CoreGraph outperforms also FocusSearch. FocusSearch results particularly efficient on regular graphs having a mesh structure. However, since FocusSearch uses initial domains to avoid label comparisons, the memory requirements do not scale with respect to graphs size. On the

other hand, CoreGraph maintains a low memory profile.

## References

1. Ullmann, J. R. 1976. An algorithm for subgraph isomorphism. *J. ACM* 23, 1, 31-42.
2. Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. 2004. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 10, 1367-1372.
3. Solnon, C. 2010. AllDifferent-based Filtering for subgraph isomorphism. *Artif. Intell.* 174, 12-13, 850-864.
4. Ullmann, J. R.. 2011. Bit-vector algorithms for binary constraint satisfaction and subgraph isomorphism. *J. Exp. Algorithmics* 15, Article 1.6 February 2011.

## Truncated SVD best rank choice through ROC curves for genomic annotation prediction

D. Chicco , M. Masseroli

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

### Motivations

Correct interpretation of many biological experiments is currently based on consistency of biomolecular annotation databases. Such databases are very widespread and very useful for the scientific community, but, unfortunately, incomplete by definition. To support and quicken their time consuming curation procedure, and to improve their consistence, computational methods that supply a ranked list of predicted annotations are hence extremely useful. We depart from a previous work on the prediction of Gene Ontology (GO) annotations, based on the truncated Singular Value Decomposition (SVD) of the annotation matrix, where the truncation level  $k$  of the input matrix is a keypoint in obtaining both best biomolecular annotation predictions and best performance. Here we propose a method that chooses this truncation level by computing and evaluating the Area Under the Curve (AUC) of different Receiver Operating Characteristic (ROC) curves.

### Methods

Let the matrix  $A(i,j)$ , with  $m$  rows (genes) and  $n$  columns (annotation terms), represent all annotations of a specific controlled vocabulary for a given organism. The entry  $A(i,j)=1$  if gene  $i$  is annotated to term  $j$  (or descendant), 0 otherwise. The annotation prediction is performed by computing a reduced rank approximation  $A_k$  of the matrix  $A$ , by means of the SVD.  $A_k$  contains real value entries related to the likelihood that gene  $i$  shall be annotated to term  $j$ . For a defined threshold  $t$ , if  $A_k(i,j)>t$ , gene  $i$  is predicted to be annotated to term  $j$  and, if  $A(i,j)\leq 0$ , a new annotation is suggested (AP). Conversely, if  $A(i,j)>0$  &  $A_k(i,j)\leq t$ , an existing annotation is suggested as semantically inconsistent with the available data (AR).

The method core is the truncation level  $k$ , which defines the size of the submatrix used by the algorithm to compute the SVD. For any considered truncation value, our greedy algorithm generates a ROC curve drawing the AR rate (1.0 - Sensitivity) vs the AP rate (1.0 - Specificity), and computes the ROC AUC. If  $p$  is the maximum rank

of  $A$ , where  $p=\min(m,n)$ , and  $r\leq p$  is the number of non-zero singular values along the diagonal of Sigma matrix, the best truncation value is in the  $[1;r]$  interval. To avoid performing the SVD and ROC analysis for every integer value in  $[1;r]$  we sample within this interval  $q$  values that could be used as adequate truncation values. To obtain the best sampling, we study the distribution of the AUC values for different truncation levels, for a sample dataset (i.e. organism *Gallus gallus*, GO Biological Process). First, we exclude first and last 10% values, to avoid taking levels that, during SVD reconstruction, would consider too few or too many non-zero singular values of  $A$ . By analyzing gradient variations in AUCs distribution function, we sample  $q$  truncation values, inside the above range. We consider every  $q_i$  as a new SVD truncation value, and compute the  $AUC_{q_i}$  of the corresponding ROC $_{q_i}$  curve. Finally, we take  $\min(AUC_{q_i})$  as the best  $q_i$  truncation value.

### Results

For evaluation, we use old GO annotations of *Gallus gallus* and *Bos taurus* genes available on July 2009 in an old version of GO Annotation databases (<http://geneontology.org>). By analyzing *Gallus gallus* annotations between genes and Biological process (BP) (8,731 annotations; 275 genes; 610 BP terms), the algorithm suggests  $k=77$  as the best truncation level for SVD. This level led to a ROC curve having  $AUC=40.27\%$ , while the 2nd best value, 59, led to  $AUC=40.46\%$ . From the 8,731 input annotations, with  $t=0.4$ , the SVD method with value 77 predicted 44 AP annotations. Out of these, 28 (63.63%) turned out to be present among the 27 month newer GO annotations in a more recent GO database version (Oct-2011); these 28 APs included 14 annotations (50%) with GO evidence different from IEA-ND. On the other hand, the 2nd best value, 59, led to worst results: same number of 44 APs, but just 14 of these (31.81%) were present among the newer GO annotations considered. Other truncation values, related to higher AUC values, led to even worst prediction results.

### Erratum.

This is modified version replaced on 22 May 2012. The Editor guarantes the scientific integrity of the abstract content.

## New metaheuristics approaches for biclustering of gene expression data

F. Musacchia<sup>1</sup> ✉, A. Marabotti<sup>2</sup>, A. Facchiano<sup>3</sup>, L. Milanese<sup>2</sup>, P. Festa<sup>1</sup>

<sup>1</sup>Dipartimento di Matematica e Applicazioni "R. Caccioppoli", Università Federico II, Napoli, Italy

<sup>2</sup>Istituto di Tecnologie Biomediche - CNR, Milano, Italy

<sup>3</sup>Istituto di Scienze dell' Alimentazione - CNR, Avellino, Italy

### Motivations

Biclustering or simultaneous clustering of both genes and conditions have generated considerable interest over the past few decades, particularly related to the analysis of high-dimensional gene expression data in information retrieval, knowledge discovery, and data mining [1]. Given a gene expression data matrix, a bicluster is a submatrix of genes and conditions that exhibits a high correlation of expression activity across both rows and columns. The problem of locating the most significant bicluster has been shown to be NP-complete. Therefore, given the inner "intractability" of the problem from a computational point of view, to efficiently find good solutions in a reasonable running times we have designed and implemented several different metaheuristics based on a GRASP framework.

### Methods

All the procedures are based on a GRASP (Greedy Randomized Adaptive Search Procedure) framework [2]. A GRASP is a randomized multistart iterative metaheuristic consisting of two phases: a construction phase and a local search phase. The construction phase builds iteratively a feasible solution in a greedy randomized fashion. Once a feasible solution is obtained, a local search procedure attempts to improve it by producing a locally optimal solution with respect to suitable defined neighborhood structure. The construction and the local search phases are repeatedly applied. The best local optimum solution found is returned as an approximation of the optimal one. We designed and implemented several different GRASP-based metaheuristics that differ in both construction and local search phase. In the construction phase, one implements a k-means and a second one a greedy randomized procedure inspired by a minimum spanning tree of a suitable weighted graph. Then, two types of lo-

cal searches have been implemented: one has been already proposed [3] and a second one is an Iterated Local Search [4]. All the designed algorithms have been tested and compared using the lymphoma dataset [5].

### Results

Both the solution construction procedure and the local search procedure use a gain function that combines the mean squared residue [1], the row variance, and the size of the bicluster. Experimental results on Lymphoma microarray data set [5] are promising indicating that the methods are able to find significant biclusters, also from a biological point of view.

### Acknowledgements

A.M. and L.M. are supported by MIUR FIRB ITALBIONET (RBPR05ZK2Z and RBIN064YAT\_003). The work has been made in the frame of the Flagship Project InterOmics.

### References

1. Y.Cheng and G. Church. (2000) Biclustering of Expression Data, Proc. Int. Conf. Intell. Syst. Mod. Biol., 93-103.
2. T.A. Feo and M. G.C. Resende (1995) Greedy Randomized Adaptive Search Procedures, J. Global Optim., 6, 109-134.
3. F. Musacchia, A. Marabotti, A. Facchiano, L. Milanese, and P. Festa (2011) Biclustering of gene expression data based on GRASP-like algorithms. BITS2011, ISBN 978-884673069-5, pp. 100-101.
4. W.Ayadi, M. Elloumi and J.-K. Hao (2010) Iterated Local Search for Biclustering of Microarray Data. Lect. Notes Comput. Sci., 6282, 219-229
5. A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Jr Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403, 503-511

## Identification, reconstruction and validation of insertions in resequencing projects with GapFiller

F. Nadalin , F. Vezzi, S. Scalabrin, A. Policriti

Istituto di Genomica Applicata (IGA), Italy

### Motivations

A difficult problem in resequencing projects is the identification, reconstruction, and validation of inserted regions within an unknown genome, with respect to a reference one. In most organisms, many structural variants are found in highly repetitive regions of the genome, making their identification difficult. Recent studies demonstrate the feasibility of detecting structural variants using next-generation, paired-end sequencing reads. Reconstructing completely or at least both ends straddling the inserted sequence is a first validation step. Both reconstruction and validation are still weak in existing tools.

### Methods

GapFiller is a tool developed to fill with a single sequence the gap between paired reads produced by NGS technologies. It is based on a seed-and-extend scheme and it implements techniques to avoid errors in produced contigs. A contig spanning the whole gap is dubbed certified if the mate of the seed read is found. As a matter of fact, GapFiller can be applied to whatever pair of sequences known to lay at an estimated distance, as long as a set of uniformly distributed short reads are provided as input to fill the gap. Our pipeline to reconstruct and validate insertions in a resequenced individual is divided in two main phases: the first one consists in constructing contigs straddling the borders of putative insertions, the second one in filling the gap between them. Before running the pipeline we need to determine locations of putative insertions and to extract the reads aligning around them. More specifically, if insertions in organism A with respect to organism B are to be investigated, we first extract locations of putative insertions, using a tool designed for this purpose (i.e., BreakDancer). Then the reads of A are aligned against the reference B and those mapping next to insertions are extracted, as well as their, possibly unmapped, mates, with the proper orientation. GapFiller is then run twice: a first time to fill the gap between the extracted paired reads in

order to reconstruct the borders of each insertion, and a second time to reconstruct the sequence between the certified contigs produced. In the latter phase we treat contigs on the left and on the right side of an insertion as if they were the two pairs of a paired read, respectively, and GapFiller's output will be a super-contig. The event that the super-contig obtained starting from the left contig finally matches against the right one, represents an evidence that we have reconstructed the desired sequence. Clearly, the level of confidence is a function of the number of super-contigs for each (putative) insertion. Moreover, as an important byproduct, we have the assembly of the missing sequence.

### Results

In order to check correctness we tested our pipeline on a real dataset, consisting of a 30x coverage of paired reads from the *Vitis vinifera* variety PN40024 (485Mbp), for which the reference genome is known. Using BreakDancer we extracted pairs of coordinates on the reference corresponding to deletions on the resequenced variety Sangiovese. This way we simulated insertions in PN40024, with the advantage of being able to check if the sequences assembled by GapFiller were correct, by simply aligning them against the reference. As a preliminar validation step, for each sequence *S* identifying a putative insertion we computed the maximal tails of *S* covered by the (certified) contigs produced in the first phase, i.e. we consider only the contigs consisting of paired reads whose gap has been successfully filled. In particular, we identified 800 putative insertions for which we were able to correctly assemble at least 200bp on both tails. The second phase is more difficult as we try to reconstruct (probably highly) repetitive regions. For a few pairs of left and right contigs we were able to entirely reconstruct the inserted sequence, with respect to the reference genome PN40024. However, this point requires a deeper analysis and we are currently working on the improvement of the final step of our pipeline.

## CytokineDB and CytReD@CROM

S. Costantini<sup>1</sup>✉, A. Sharma<sup>2</sup>, R. Raucci<sup>2</sup>, F. Capone<sup>1</sup>, M. Miele<sup>2</sup>, E. Guerriero<sup>1</sup>, G. Castello<sup>1</sup>, M. Di Stasio<sup>3</sup>, G. Colonna<sup>2</sup>

<sup>1</sup>INT Pascale-Centro Ricerche Oncologiche Mercogliano, Mercogliano, Italy

<sup>2</sup>Dipartimento di Biochimica e Biofisica and Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, Italy

<sup>3</sup>Istituto di Scienze dell'Alimentazione-CNR, Avellino, Italy

### Motivations

The cytokines family is composed by many proteins that need to bind to specific receptors on the cell surface to perform their biological function. This binding can stimulate both the expression of receptors for cytokines and the production of other cytokines that in turn act on other target cells. On the whole, the totality of the cytokines and of their interactions in and around biological cells is defined with the "cytokinome" term. Often these molecules are involved in cancer-related chronic inflammation and play a pivotal role in promoting tumorigenesis and metastatic processes. Therefore, we have developed i) CytokineDB that is an annotated database that collects biological information regarding the cytokines family in human and ii) CytReD (Cytokine Receptor Database) that collects biological information regarding the human cytokine receptor families and their related ligands and can be used by researchers as well as physicians and clinicians to identify what cytokines are reported in the literature as significant in a given disease.

### Methods

Some databases were used to collect gene and protein data regarding the human cytokine family: Pubmed and OMIM for biological activity, ENSEMBL for gene records, SRS retrieval system for searching DNA and protein sequences, PDB for three-dimensional structures, and FoldIndex for the prediction of disorder propensity of the cytokine receptors. Search form of CytokineDB is based on a CGI script written in PERL language. CytReD is developed using a dynamic content management system Drupal version 6.17 and scripting language PHP version 5.3.2. JMOL visualization package is embedded for cytokine receptors three dimensional structure in detailed information page.

### Results

In CytokineDB the human cytokine family was subdivided in 12 sub-families and the user can click on the image near to the name of each cy-

tokine family and have a short description of the structures of these families. In each subgroup, all the known cytokines were inserted. The user can select a sub-family and choose the protein of which want have information. The output page for each cytokine reports the cell type, where the protein is located, Entrez Gene, the target receptors and cells, the main effects, the description of biological activity, the references, EMBL code, Ensembl protein \_ coding gene, the number of transcripts, the number and the code of associated peptides, the number and the code of exons, the chromosome location, CCDS, RefSeq DNA, RefSeq peptide, Protein ID, UniProt code, Sequence isoforms, amino acid sequence, PDB code, CATH and SCOP classifications, and structural features. In CytReD the user can search by selecting from four options: cytokine receptor name, ligands, cytokine Family or disease. The result page is divided into description, sequences and accession codes and other biological information along with the references. Description provides information about the biological description, the name of ligands linked to CytokineDB database, the related synonyms, cell type on which the protein is expressed, the role and the diseases in which it is involved linked to PharmaGKB database. The sequences and accession codes section contains the nucleotide sequence, protein sequence and its isoforms. In particular, the accession codes are linked to the other important databases like Uniprot, Protein Databank, Entrez gene. In the biological information section there are information related to chromosome localization, Ensemble protein coding gene, the codes of transcripts, associated peptides, references, CATH and SCOP classifications, and quantitative analysis like number of residues, isoelectric field, number of positively and negatively charged residues, unfoldability index, and mean hydropathicity value(GRAVY). CytokineDB and CytReD are part of a broader project aimed to develop tools and portals able to be useful supports for a reliable predictive medicine.

## VIRES: visualization and identification of A-to-I RNA editing sites in genomic sequences

R. Distefano<sup>1</sup>, G. Nigita<sup>1</sup>, V. Macca<sup>1</sup>, R. Giugno<sup>2</sup>, A. Pulvirenti<sup>2</sup>✉, A. Ferro<sup>2</sup>

<sup>1</sup>Dipartimento di Matematica e Informatica Università degli Studi di Catania, Catania, Italy

<sup>2</sup>Dipartimento di Biomedicina Clinica e Molecolare, Università degli Studi di Catania, Italy

### Motivations

RNA Editing is a type of post-transcriptional modification that takes place in the eukaryotes and represents one of the last frontiers of molecular biology. It alters the sequence of primary RNA transcripts by deleting, inserting or modifying residues. Several forms of RNA editing have been discovered including A-to-I, C-to-U, U-to-C and G-to-A editing. A-to-I RNA editing (Adenosine-to-Inosine) is the most frequent and common post-transcriptional modification, where adenosine (A) deamination produces its conversion into inosine (I), which in turn is interpreted by the machinery translation and splicing as guanosine (G) and so this causes the change of the RNA sequence. This biological phenomenon is catalyzed by members of the Adenosine Deaminase Acting on RNA (ADAR) family of enzymes and occurs only on dsRNA structures. Thus, A-to-I editing changes RNA molecules in various ways including: the translation of its codons; the creation and/or destruction of splicing sites; the micro-RNA/mRNA binding. Therefore, it is not surprising that the malfunction of the editing machinery has been implicated in various human diseases. In the last years, the application of global approaches to the study of A-to-I editing, including high throughput sequencing and bioinformatics, has led to important advances. However, in spite of enormous efforts, the real biological functioning of this phenomenon remains unknown. In this work, starting from genomic sequences, given as input, we present a bioinformatics approach to discover and visualize A-to-I editing sites.

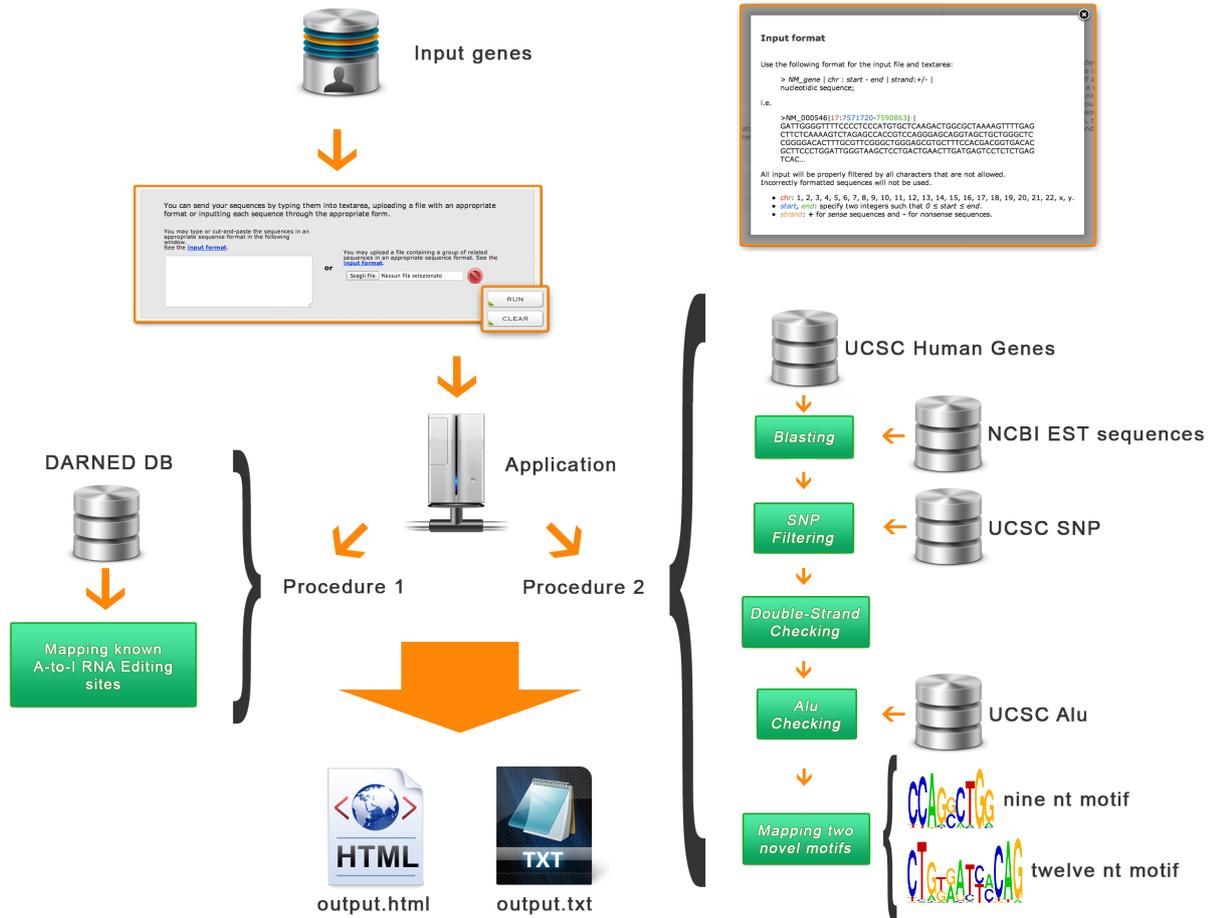
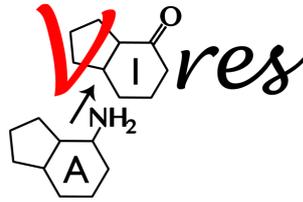
### Methods

VIRES is a web-based tool that maps newly predicted and known A-to-I editing site in genomic sequences. The tool is equipped with a user-friendly interface allowing users to analyze

genomic sequences in order to identify candidate A-to-I editing sites. VIRES action can be subdivided in two different tasks: the identification of known editing sites and the mapping of new ones. The system highlights the known editing events falling into the input sequences by searching the genomic positions of sequences in the DARNED dataset containing approximately 42,000 human genome loci corresponding to validated A-to-I RNA editing sites. In the second phase, we search for new putative editing events in these sequences. This task is performed by analyzing more than 38,000 Human genes (build GRCh37/hg19). In more details, the searching of new editing sites can be divided into the following steps. First, we execute BLAST between Human genes and EST (Expressed Sequence Tag) sequences selecting only the alignments which contains at least one A-G mismatch between genomic and EST sequence. Next, we remove all mismatches that are SNPs (Single Nucleotide Polymorphisms). Finally, we verify if this candidate edited site belongs to a double strand region. Once we identify the candidate A-to-I editing sites, we add the information for the functional enrichment including the presence of repetitive elements, the ESTs sequences with the candidate editing sites, and the location of two novel motifs characterizing A-to-I editing events (CCAG[G|C]CTGG and CTG[T|G][G|A]AT[C|T][A|C]CAG) in flanking regions of putative editing sites. The user can choose whether to download this information in a text or xml file.

### Results

VIRES is an easy to use bioinformatic tool that allows the in-silico-identification of putative and validated A-to-I RNA editing sites. It is built on top of several knowledge bases such as DARNED, EST, SNPs.



NOTE: you can see results obtained by clicking light blue boxes.

Seq #/1	Gene name: NM_017871	Chr: 1	Start (bp): 124095	End (bp): 125044	Strand:
88481	1	ACTCTGCTGATGATCAACAGCAGCGGCTCCGCTCCGAGAGCGGCGAGAGATCTCTGATGATAGATGACG	75		
88482	2	GGTCTGCTGATGATGATCAACAGCAGCGGCTCCGCTCCGAGAGCGGCGAGAGATCTCTGATGATAGATGACG	76		
88483	8	CTCTGATGATGATGATCAACAGCAGCGGCTCCGCTCCGAGAGCGGCGAGAGATCTCTGATGATAGATGACG	108		
88484	8	CTCTGATGATGATGATCAACAGCAGCGGCTCCGCTCCGAGAGCGGCGAGAGATCTCTGATGATAGATGACG	109		
88485	141	CTCTGATGATGATGATCAACAGCAGCGGCTCCGCTCCGAGAGCGGCGAGAGATCTCTGATGATAGATGACG	236		
88486	141	CTCTGATGATGATGATCAACAGCAGCGGCTCCGCTCCGAGAGCGGCGAGAGATCTCTGATGATAGATGACG	237		
88487	224	GG	626		
88488	224	GG	627		
88489	422	TAGAATTCTTAAGCCCTCCGACACTTCTCTTATTATATATCTCTCTTCTTCTTCTTCTTCTTCTTCTTCTT	636		
88490	422	TAGAATTCTTAAGCCCTCCGACACTTCTCTTATTATATATCTCTCTTCTTCTTCTTCTTCTTCTTCTTCTT	637		
88491	440	ATTTTCTACTT	647		
88492	440	ATTTTCTACTT	648		
88493	441	CAATTATTCT	658		
88494	441	CAATTATTCT	659		
88495	441	CAATTATTCT	660		
88496	441	CAATTATTCT	661		
88497	441	CAATTATTCT	662		
88498	441	CAATTATTCT	663		
88499	441	CAATTATTCT	664		
88500	441	CAATTATTCT	665		
88501	441	CAATTATTCT	666		
88502	441	CAATTATTCT	667		
88503	441	CAATTATTCT	668		
88504	441	CAATTATTCT	669		
88505	441	CAATTATTCT	670		
88506	441	CAATTATTCT	671		
88507	441	CAATTATTCT	672		
88508	441	CAATTATTCT	673		
88509	441	CAATTATTCT	674		
88510	441	CAATTATTCT	675		
88511	441	CAATTATTCT	676		
88512	441	CAATTATTCT	677		
88513	441	CAATTATTCT	678		
88514	441	CAATTATTCT	679		
88515	441	CAATTATTCT	680		
88516	441	CAATTATTCT	681		
88517	441	CAATTATTCT	682		
88518	441	CAATTATTCT	683		
88519	441	CAATTATTCT	684		
88520	441	CAATTATTCT	685		

**Chromosome: chr1**

**Strand: -**

**Double Strand: yes**

**Absolute position: 125134**

**Relative position: 663**

**Motives (the nearest motives to RNA Editing site)**

Start 995 - Motive - strand - chromosome

9motif 125134 CCGACCTG - chr1

12motif 125136 CCGACCTGAC - chr1

**EST sequences**

EST	Range (start-end)	Name	Class	Family
CG222773	125134-125135	Alu	SINE	Alu
CG222766	125134-125135	Alu	SINE	Alu

**Repetitive Regions**

EST	Range (start-end)	Name	Class	Family
CG222773	125134-125135	Alu	SINE	Alu
CG222766	125134-125135	Alu	SINE	Alu

A fragment of output.html page

Figure 1. Vires usage workflow and architecture.

## SpliceAid-F: a database of human splicing factors and their RNA binding sites

M. Giulietti<sup>1</sup>, F. Piva<sup>2</sup>, M. D'Antonio<sup>3</sup>, P. D'Onorio De Meo<sup>3</sup>, T. Castrignanò<sup>3</sup>, G. Pavesi<sup>4</sup>, G. Pesole<sup>5</sup>✉

<sup>1</sup>Department of Biosciences, Biotechnology and Pharmacological Sciences, University of Bari, Bari, Italy

<sup>2</sup>Department of Specialized Clinical Sciences and Odontostomatology, Polytechnic University of Marche, Ancona, Italy

<sup>3</sup>Consorzio per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy

<sup>4</sup>Department of Biomolecular Sciences and Biotechnology, University of Milan, Milan, Italy

<sup>5</sup>CNR - Institute of Biomembrane and Bioenergetics, Bari, Italy

### Motivations

The pre-mRNA sequences of eukaryotes harbour multiple information layers in addition to the one specifying the amino acid sequence. Indeed, specific signals regulating the splicing process, RNA folding, capping, polyadenylation, stability and nuclear export, overlap each other and with the coding information. Mutagenesis experiments have highlighted that splicing regulatory elements are scattered in the entire pre-mRNA sequence and that all nucleotide positions are potentially involved in the generation of the specific splicing pattern through the specific interaction with trans-acting RNA-binding proteins. In particular, the identification of cis-regulatory motifs in exonic regions, such as exonic splicing enhancers (ESEs) or silencers (ESSs), superimposed on the coding sequence of the gene can be driven by the observation of purifying selection occurring at synonymous codons [5]. The coordinated binding of combinations of regulatory proteins to their binding sites modulate the expression of specific transcript isoforms in a cell/tissue type-, development stage-, disease- and/or other condition-specific manners and may also promote or repress the formation of the spliceosome, the large (~60S) RNA-protein machinery that catalyzes intron removal. A growing list of mammalian protein factors involved in splicing regulation and their target sites in the pre-mRNA has been reported in the literature [2]. In order to establish a curated and retrievable repository of splicing regulatory factors and target sites for human genes we have recently created SpliceAid [4] that can also be used to find putative regulatory motifs in user submitted sequences. As a further evolution of SpliceAid, we present here SpliceAid-F, a database compilation of splicing regulatory factors and their experimentally validated target RNAs extracted from an exhaustive hand-curated literature search.

### Methods

For each known splicing factor, cross-links to gene (NCBI Entrez) and protein (Uniprot) IDs, as well as information on the structure of the RNA binding domain, the protein-defect associated disease (MIM), and the interacting proteins (from STRING and IntAct) have been collected. Moreover, we have extracted from the literature the relevant information on the genome coordinates of RNA binding sites (or experimentally validated non-binding sites), type of binding assay, gene information and the context-specific splicing effects of splicing factor binding in terms of exonization or intronization. Furthermore, binding site information have been also reported in the transcript view of ASPicDB [1,3].

### Results

SpliceAid-F, currently collects 68 records for splicing regulatory factors and 2489 records for their related binding sites and can be retrieved through a web interface. SpliceAid-F collects in a unique resource heterogeneous information about splicing regulatory proteins, related RNA binding sites and context-specific activity of their interaction. Our database may be a useful resource to retrieve and visualize experimentally known splicing factor binding sites in a gene and to investigate their context where additional binding sites may establish a potential competition or co-regulation. All these information may help to explain the observed splicing pattern as well as the effect of mutations, which if located in functional regulatory elements may generate an aberrant and possibly pathological splicing pattern. This resource can also be useful to develop a new generation of prediction software taking into account all the splicing regulatory element and so allowing to attain a more accurate prediction of splicing patterns.

### Availability

<http://www.caspur.it/SpliceAidF>

## References

1. Castrignano T, D'Antonio M, Anselmo A, Carrabino D, D'Onorio De Meo A et al. (2008) ASPicDB: a database resource for alternative splicing analysis. *Bioinformatics* 24(10):1300-1304.
2. Gabut M, Chaudhry S, Blencowe BJ (2008) SnapShot: The splicing regulatory machinery. *Cell* 133(1): 192e191.
3. Martelli PL, D'Antonio M, Bonizzoni P, Castrignano T, D'Erchia AM, D'Onorio De Meo P et al. (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res.* 39(Database issue):D80-85.
4. Piva F, Giulietti M, Nocchi L, Principato G (2009) SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans. *Bioinformatics* 25(9):1211-1213.
5. Schattner P, Diekhans M (2006) Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Res* 34(6):1700-1710

## An integrated system for mining relations among microRNAs, drugs and phenotypes

A. Pulvirenti<sup>\*1</sup>, R. Giugno<sup>\*1</sup>, S. Di Bella<sup>2</sup>, G. Nigita<sup>2</sup>, V. Macca<sup>2</sup>, A. Giummarra<sup>2</sup>, D. Garofalo<sup>2</sup>, G. Caruso<sup>2</sup>, V. Bonnici<sup>3</sup>, A. Ferro<sup>\*1</sup>✉

\*these authors contributed equally

<sup>1</sup>Dipartimento di Biomedicina Clinica e Molecolare, Università di Catania, Catania, Italy

<sup>2</sup>Dipartimento di Matematica e Informatica, Università di Catania, Italy

<sup>3</sup>Dipartimento di Informatica, Università di Verona, Verona, Italy

### Motivations

Interactions among genes together with their expression level determine a phenotype in a given organism. Genetic origin of a disease is often discovered once its phenotype has been clearly defined. Human phenotypes can be easily described by their observable features. Most human phenotypes can be found in the database OMIM (Online Mendelian Inheritance in Man). Van Driel et al. proposed a text mining technique to be applied to OMIM in order to find phenotypes similarity. On the other hand modern studies have shown that genetic variations may affect

drug response. Moreover, drugs similarity may be used to predict drug-diseases relations. The aim of such research area is to computationally predict drugs response and their side effects. Indeed this information may be used to reduce new drugs development cost. It is well known that microRNAs (miRNAs) expressions strongly affect phenotypes sometimes causing diseases. Consequently, miRNAs may provide insights on drug response, for example, by analysing their regulating interaction with drug targets. Simple associations miRNA-disease through their target genes can be obtained querying a knowledge

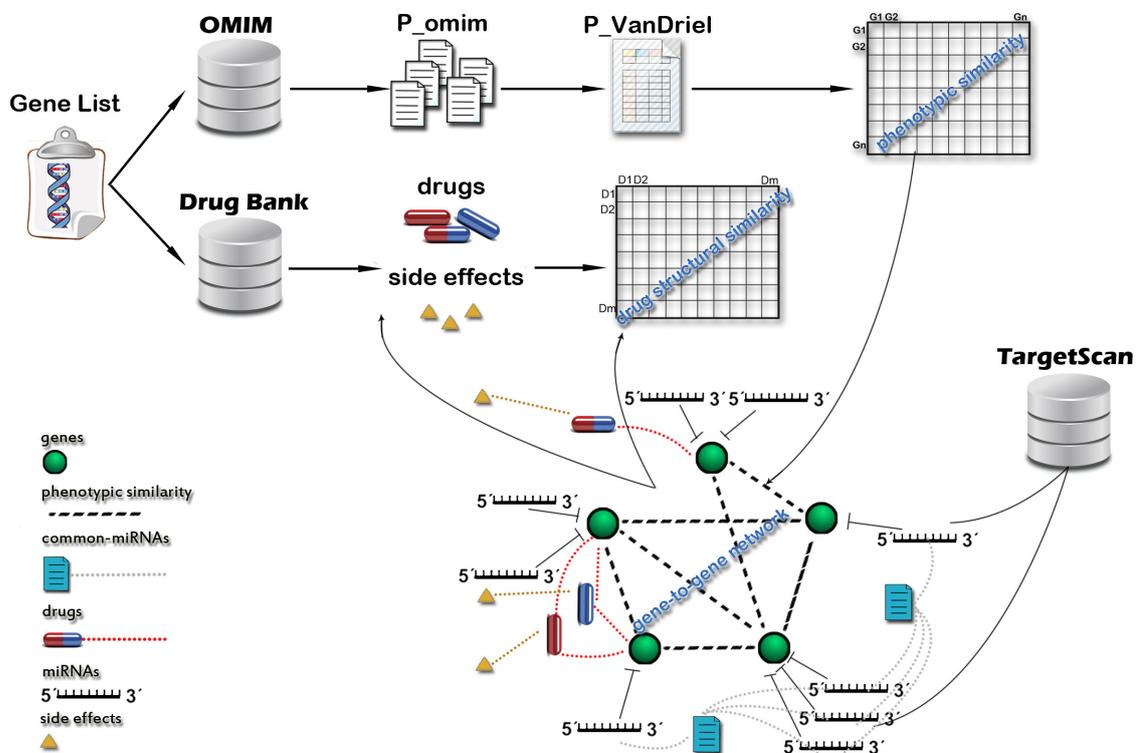


Figure 1. Data mining workflow.

base such as miRò. The subject of the present study is the design of an environment in which microRNAs-drugs-phenotypes relationships can be deduced using more advanced network-based techniques.

### Methods

Starting from matrices of phenotypic and drugs similarities, the tool provides users an interface to establish those miRNAs, known to be responsible for the regulation of genes involved in a disease, associated with other diseases showing similar phenotypes. Given a list of genes, e.g. those composing a metabolic pathway, we lead a phenotypic and miRNAs-based analysis by calculating distances among gene pairs. The phenotypic analysis consists of two steps. In the first step, we extract for each pairs of genes the phenotypes from OMIM. Then, using Van Driel et al., for each phenotype, we select the one hundred most similar to it. This allows the construction of a

complete list of phenotypes associated to each genes pair. In the second step, we calculate the phenotypic distance matrix between any two genes according to the their common phenotypes. This can be viewed also as a weighted gene network. Applying TargetScan (or any other targeting program) to the endpoints of each edge in the network, we build a list of miRNAs. This yields a list of miRNA-phenotype associations. Genes in the networks are also correlated with drugs targeting them together with their side effects through drug-bank repositories. Drugs structural similarity is computed by Tanimoto technique.

### Results

We propose a software tool that using gene-phenotypes association techniques in connection with gene network-based correlation algorithms allows the identification of statistically significant microRNA-phenotype-drug associations.

## A comprehensive 16 loci-based DNA fingerprinting dataset of a broad tumor cell line panel for cancer research

A. Somaschini, E. Scacheri, A. Nuzzo, N. Amboldi, D. Ballinari, G. Ukmar, A. Isacchi, R. Bosotti 

Business Unit Oncology, Nerviano Medical Sciences, Milan, Italy

### Motivations

Tumor cell lines are widely used as in vitro models and as screening tools in cancer research. A significant portion of cell lines have been reported to be misidentified, potentially generating misleading data interpretation. Correct assessment of cell line genetic identity is therefore critical to cancer biology studies as well as to drug discovery. Several methods may be applied to authenticate cell lines, but DNA fingerprinting based on the detection of Short Tandem Repeats (STR) has emerged as the standard approach. STR profiles for several commercial cell lines can be retrieved from sparse literature reports or from vendor databases, however these are mainly based on the analysis of 8 loci, except for the NCI60 panel, on which 16 loci were profiled. Inconsistencies can be found in allele designation among the different authors, mainly when more than two alleles per locus are present, as is frequently found in tumor cell lines. A resource to facilitate literature interrogation is represented by the Cell Line Integrated Molecular Authentication database (CLIMA), but to date there are no reported comprehensive databases containing 16 loci profile sets generated in parallel. Here we disclose the in-house generation of a homogeneous 16 loci dataset, reporting the STR profiles for a panel of about 300 tumor cell lines, representative of diverse solid and circulating tumor types. Our intent was to facilitate the management of the internal cell bank, by assessing the identity and stability over time of the tumor cell lines used for research, with the highest accuracy in discrimination and with an optimized score for profile similarity checking. The STR database can be linked to Nerviano Medical Sciences (NMS) internal database, which contains information on cell line origin, morphology, growth conditions, as well as known somatic mutations (from the Wellcome Trust Sanger Institute Cosmic database), thus making it a comprehensive integrated platform for cell line utilization in drug discovery. Our plan is to make the STR database available to the scientific community.

### Methods

DNA fingerprinting is based on the simultaneous amplification of highly polymorphic STR sites, which are short DNA sequences with a varying number of repeats in each cell line. Genetic abnormalities in tumor cell lines can result in more than two alleles at each locus, which complicates the analysis. Most STR profiles reported so far are based on the analysis of a limited number of STR loci, usually 8. The probability for two individuals to share the same STR profile is estimated to range from  $\sim 10^{-8}$  for 8 loci to  $\sim 10^{-17}$  for 16 loci. A 16 loci profile can be beneficial in cancer research, due to the intrinsic genetic instability of tumor cells that may result in allele acquisition or loss. In the current study we performed a 16-loci STR fingerprinting analysis on a panel of about 300 commercially available cancer cell lines. DNA was prepared directly from frozen cells, using NucleoSpin Tissue (Macherey-Nagel), and then analysed using AmpFISTR Identifier Plus PCR Amplification kit (Applied Biosystems), that amplifies simultaneously 15 tetranucleotide repeat loci and the amelogenin gender marker. For a more accurate comparison of STR profiles in different cell lines, we have introduced a modification to the standard calculation of similarity scores to account for partial allele identity at each locus, allowing a more precise definition of the level of similarity, which is particularly relevant when multiple alleles are present. This is obtained by applying the following formula that calculates the number of identical alleles at each locus, divided by the total number of alleles: given a cell line A and a cell line B, the score is defined as the identified sum of similarities at each  $i$ -th locus  $s(i)$  normalized by the number of not empty loci, where  $s(i) = 2 \times [\text{alleles}(A) = \text{alleles}(B)] / [\#\text{alleles}(A) + \#\text{alleles}(B)]$ . Based on a preliminary sensitivity analysis, the similarity score threshold to classify two cell lines as identical was set at 80%.

### Results

A DNA fingerprinting analysis based on 16 loci was performed on a panel of about 300 widely

used tumor cell lines, purchased at NMS over several years from diverse providers. Using an established cut off of 80%, we found that all cell lines were correctly classified, suggesting that the cell line panel is genetically stable over time. As expected, an overall pairwise analysis of all the different cell line profiles revealed a generally low similarity degree (less than 50%), with no evi-

dence for any similarity bias due to tumor type. The main outcome of this study is the availability of a large 16 loci STR dataset of tumor cell lines, which was integrated as a fundamental part of the NMS Cell Bank database. The STR dataset can be easily interrogated and distributed for use as a reliable reference for cancer research.

## An interactive tool enabling a comparative analysis of STR profiles

G. Ukmar, R. Bosotti, A. Somaschini, J. Malysko, L. Radrizzani, G. Masetti, E. Scacheri, A. Isacchi, A. Nuzzo 

Business Unit Oncology, Nerviano Medical Sciences, Milan, Italy

### Motivations

Cell line misidentification is a critical issue in molecular experiments that use normal and/or tumor cell lines as tools for disease characterization and pharmaceutical treatments. Thus, confirmation of cell line genetic identities is crucial to validate the obtained results. DNA fingerprinting of Short Tandem Repeats (STR) has become a standard technique used to identify the unique genetic profile of a cell line, based on the comparison of its microsatellite loci pattern with a known profile. The efficiency of this approach depends on the number of loci analysed, as well as on the existence of a large reference dataset. We applied a large scale STR characterization of 16 loci to a panel of about 300 commercially available tumor cell lines, generating an unprecedented dataset with uniform characterization. In order to maximize its exploitation by the scientific community, we developed an interactive software tool which allows to easily perform the automated comparison of the STR profile obtained for a cell line of interest against Nerviano Medical Sciences cell line database, facilitating the identification of the cell line and the potential discovery of unreported similarities.

### Methods

We designed and developed an easy-to-use software tool which allows to quickly retrieve STR profiles based on the degree of similarity to the input profile of the cell line of interest. The tool is built on a dataset which includes about 300 commercially available human tumor cell lines, profiled in house. The panel is representative of diverse tumor tissue types. Users can insert data relative to a cell line of interest specifying either the cell line name or its STR profile. Then a query is run against the dataset in order to compute a similarity score versus each cell line and to generate the final summary table. The similarity

is computed using a score that we implemented in order to account for multiallelic values at individual loci, which may be easily found in aberrant tumor cell line genomes. The similarity threshold by which two cell lines are considered identical is set at 80%. This cut-off value has been defined through a sensitivity analysis of available profiles. The tool has been developed using the Java programming language, which makes it easily portable on different platforms. The Graphical User Interface is composed of four main sections: i) the "Cell line" identification section by which the user may enter a specific identifier name or choose a cell line from the underlying dataset; ii) the "Loci" input pane, where the user can enter allelic values for each locus, or automatically retrieve loci values for an available cell line; iii) the "Command" pane containing the search task launcher button; iv) the "Result table" section, which reports all cell line matches found with a similarity score of at least 80%.

### Results

The tool has been implemented and used to perform a complete characterization of the STR profiles of a 300 tumor cell line panel. Query automation allowed to calibrate optimal settings for the parameters involved in cell identification. In particular, the optimal threshold cut-off value has been identified through a sensitivity analysis using the available profiles. Moreover, the similarity value that we computed overcomes critical aspects of other adopted scores, which usually either do not take into account or provide inconsistent values for multiallelic loci. The availability of a portable tool will allow bench scientists to have an immediate authentication of the cell line of interest, which is nowadays a mandatory requirement for paper submission to most scientific journals.

## OsteoChondroDB: a database about biomolecular chondral- bone development in physiological and diseased conditions

F. Viti<sup>✉</sup>, I. Merelli, L. Milanese

Institute for Biomedical Technologies, CNR, Segrate, Italy

### Motivations

Current researches on osteochondral tissue focus on understanding the biomolecular mechanisms of bone/cartilage development [1,2], which helps in understanding the onset of the related genetics diseases [3] and prompt the development of new approaches for tissue engineering [4]. Nowadays, regulation of tissue formation and remodelling is not completely understood, especially when dealing with the process of differentiation into bone and cartilage or with tissues affected by complex pathologies. Concerning the biomolecular mechanisms of bone formation, to authors' knowledge the only data collection existing is the Skeletal Gene Database [5], a list of genes involved in the bone metabolism accompanied by PubMed [6] references to scientific papers. This database consists of a pdf file, which describes and annotates many genes involved in osteochondral development in a simple, but static way. Other data are sparse in literature, although a rationalization of the available information about bone pathologies has been done by OMIM [7]. In order to overcome this limitation, authors designed a database, the OsteoChondroDB, by employing a vertical data integration strategy to connect tissue specific information referred to diverse biomolecular levels. In particular, the database stores information about bone development in physiological conditions together with data about osteochondral pathologies, which helps in highlighting pathways of differentiation and tissue maintaining. The resource aims at collecting and organizing data, to facilitate mining of active components in bone tissue cells. The resource is intended to be a reference knowledge base for research studies about the genetics of bone and cartilage pathologies, with the aim of improving the knowledge about physiological pathways involved in the development of this tissue. Moreover it represents a support for tissue engineering, to identify always better methods to grow cells on biomaterial scaffolds, and for new therapies identifica-

tion, proposing molecules as possible targets for drug treatments of bone diseases.

### Methods

The developed resource relies on MySQL. The database presents a snowflake schema, with the central table collecting genes involved in bone metabolism and related genetic pathologies, together with literature references. Genes have been identified manually from literature, which guarantees a high reliability of data. To promote the real comprehension of biomolecular mechanisms, data are accompanied with annotations and metadata: Single Nucleotide Polymorphisms [8] occurring in the listed genes and flanking regions, gene expression profile (from Gene Expression Atlas [9]), microRNAs plausibly targeting the gene transcripts (from myMIR site [10]), gene products (as list of mRNAs sequences from RefSeq [11]), functional domains (from InterPro [12]) and structural models from Protein Data Bank [13]. The most important aspect of collected data regards proteins interactions (PPI), from BioGRID [14], and biomolecular pathways, from KEGG [15] and Reactome [16]; this information can be exploited to create PPIs networks, based on the shortest paths, to help identifying novel hypothetical sub-pathways or extending existing pathways.

### Results

The OsteoChondroDB site provides a query system to access and visualise maintained data in different ways. The most intuitive mode is by gene or protein name: the gene profile is shown, together with the osteochondral developmental pathway or bone pathology where the selected gene is involved. Concerning osteochondral development, our database reports many genes that are known to intervene in bone development: the BMP family [17], the collagen family [18], the fibroblast growth factor [19]. Nevertheless, the study of complex mechanisms needs a deeper level of data integration, in particular the analysis of the bone and chondral pathologies

plays a critical role for understanding molecular mechanisms and regulative interactions. Among the pathologies considered in our database we can list: osteoporosis [20], osteogenesis imperfecta [21], osteopetrosis [22], osteoarthritis [23] and juvenile Paget's disease [24]. This represents a crucial aspect of the OsteoChondroDB, since no other resource tries to structure bone related biomolecular data relying on healthy and pathological conditions. In case of osteoporosis, for example, characterized by reduced bone mineral density and successive increased fracture risk, involved genes include TNFRSF11A, CSF1, OPTN, and TM7SF4, known to intervene in regulating osteoclast metabolism. Finally it is possible to retrieve maintained data on the basis of the type of cell or the localisation in the cell environment. In conclusion, the developed platform is a biomolecular knowledge base for normal and diseased osteochondral tissue analysis, and represents a potential support in 'omics' research, tissue engineering and drug discovery.

## References

- Hattori T, Müller C, Gebhard S et al., SOX9 is a major negative regulator of cartilage vascularization, bone marrow formation and endochondral ossification. *Development*. 2010 Mar;137(6):901-11.
- Zhang C, Transcriptional regulation of bone formation by the osteoblast-specific transcription factor Osx. *J Orthop Surg Res*. 2010 Jun 15;5:3
- Wang M, Shen J, Jin H et al., Recent progress in understanding molecular mechanisms of cartilage degeneration during osteoarthritis. *Ann N Y Acad Sci*. 2011 Dec;1240:61-9.
- Jiang J, Fan CY, Zeng BF, Experimental Construction of BMP2 and VEGF Gene Modified Tissue Engineering Bone in Vitro. *Int J Mol Sci*. 2011;12(3):1744-5
- Ho NC, Jia L, Driscoll CC et al., A skeletal gene database. *J Bone Miner Res*. 2000 Nov;15(11):2095-122
- PubMed [<http://www.ncbi.nlm.nih.gov/pubmed>]
- McKusick VA, Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 1998 (12th edition).
- dbSNP [<http://www.ncbi.nlm.nih.gov/projects/SNP/>]
- Parkinson H, Kapushesky M, Kolesnikov N, et al., ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 2009, 37(Database issue):D868-D872.
- Corrada D, Viti F, Merelli I et al., myMIR: a genome-wide microRNA targets identification and annotation tool. *Brief Bioinform*. 2011 Nov;12(6):588-600.
- Sayers EW, Barrett T, Benson DA et al., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 2009, 37:D5-15.
- Hunter S, Apweiler R, Attwood TK et al., InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, 37:D211-D215.
- Berman H, Henrick K, Nakamura H et al., The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, 2007, 35:D301-D303.
- Breitkreutz BJ, Stark C, Reguly T et al., The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res*, 2008, 36:D637-D640.
- Aoki-Kinoshita KF, Kanehisa M, Gene annotation and pathway mapping in KEGG. *Methods Mol Biol* 2007, 396:71-91.
- Matthews L, Gopinath G, Gillespie M et al.: Reactome knowledgebase of human biological pathways and processes, *Nucleic Acids Res* 2009, 37(Database issue):D619-D622.
- Samee M, Kasugai S, Kondo H et al., Bone morphogenetic protein-2 (BMP-2) and vascular endothelial growth factor (VEGF) transfection to human periosteal cells enhances osteoblast differentiation and bone formation. *J Pharmacol Sci*. 2008 Sep;108(1):18-31.
- Perrier E, Ronzière MC, Bareille R et al., Analysis of collagen expression during chondrogenic induction of human bone marrow mesenchymal stem cells. *Biotechnol Lett*. 2011 Oct;33(10):2091-101.
- Lin JM, Callon KE, Lin JS et al., Actions of fibroblast growth factor-8 in bone cells in vitro. *Am J Physiol Endocrinol Metab*. 2009 Jul;297(1):E142-50.
- Huang QY, Kung AW, Genetics of osteoporosis. *Mol Genet Metab*. 2006 Aug;88(4):295-306.
- Forlino A, Cabral WA, Barnes AM et al., New perspectives on osteogenesis imperfecta. *Nat Rev Endocrinol*. 2011 Jun 14;7(9):540-57.
- Del Fattore A, Fornari R, Van Wesenbeeck L et al., A new heterozygous mutation (R714C) of the osteopetrosis gene, pleckstrin homolog domain containing family M (with run domain) member 1 (PLEKHM1), impairs vesicular acidification and increases TRACP secretion in osteoclasts. *J Bone Miner Res*. 2008 Mar;23(3):380-91.
- Tchetina EV, Developmental Mechanisms in Articular Cartilage Degradation in Osteoarthritis. *Arthritis*, Volume 2011 (2011), Article ID 683970, 16 pages
- Ralston SH, Juvenile Paget's disease, familial expansile osteolysis and other genetic osteolytic disorders. *Best Pract Res Clin Rheumatol*. 2008 Mar;22(1):101-11.

## NumtS footsteps from the past: an interbreed between eukaryotic nucleus and the mitochondrion

F.M. Calabrese✉, D.L. Balacco, D. Simone, M. Attimonelli

Department of Biosciences, Biotechnologies and Pharmacological Sciences, University of Bari, Bari, Italy

### Motivations

NumtS, the Nuclear sequences of mitochondrial origin, populates eukaryotic genomes more or less abundantly [1,2]. Mitochondrion gene order varies between lineages and its conservation has been observed among neighbour clades. The availability of complete and draft eukaryotic nuclear genomes encourages us to produce NumtS tracks (loadable on UCSC genome browser), a suitable tool for the comparison of NumtS pattern among species. Analyses have been performed on 21 different species ascribable to any of the clades reported in the UCSC gateway. The inter-species comparisons led to species-specific and shared NumtS sets. In order to infer where a shared NumtS locus has been fixed during evolution and to test if a NumtS set definition may be influenced by the mitochondrial genome used in the application of the *in silico* hybridisation nuclear DNA vs mitochondrial DNA, we have carried out crossed analyses of each nuclear genome versus any of the other mitochondrion genomes within our species dataset.

### Methods

Blastn software [3] has been used to obtain the NumtS compilations then implemented, using python scripts, as tracks at the UCSC Genome browser. Galaxy suite has allowed us to intersect and extract data. SAMtools [4] have been used for BAM tracks and in manipulating alignments in the SAM format. The alignments and the mitochondrial gene orders have been checked by using Mauve software and by manual inspection.

The matrix reporting the content of the NumtS dataset resulting from crossed analyses has displayed differences in NumtS content depending on the mitochondrion genome used. Overall, among monotremes, platypus mitochondrion versus other nuclear genomes (*Ornithoryncus anatinus*) has shown the widest and the most recurrent NumtS number with the smallest standard deviation. Otherwise, *Caenorhabditis elegans* mitochondrion has led to a higher NumtS number among some mammals despite low values in the other species (itself included).

### Results

Phylogenetic inferences will be reported based on inter species gene order and gene sequence conservation, nuclear genome size and other elements that will contribute to infer when and how NumtS have reached the nuclear genome.

### References

1. Calabrese FM, Simone D, Attimonelli M. Primates and Mouse NumtS in the UCSC Genome Browser BMC bioinformatics (in press)
2. Simone D, Calabrese FM, Lang M, Gasparre G, Attimonelli M (2011) The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. BMC Genomics 12, 517.
3. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-Mouse Alignments with BLASTZ. Genome Res. 13(1), 103-7.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) 1000 Genome Project Data Processing Subgroup: The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25, 2078-9.

## In silico prediction of virus-derived small interfering RNAs and their putative host messenger targets in *Solanum lycopersicum* infected by different potato virus Y isolates

D. Catalano, F. Cillo, M. Finetti-Sialer✉

Istituto di Genetica Vegetale, Consiglio Nazionale delle Ricerche, Bari, Italy

### Motivations

RNA silencing, or post-transcriptional gene silencing (PTGS), is a conserved mechanism in a broad range of eukaryotes. In plants, PTGS acts as an antiviral system and a successful virus infection requires suppression or evasion of the induced silencing response. Small interfering RNAs (siRNAs) accumulate in plants infected with RNA and DNA viruses and provide specificity to this RNA-mediated immune system. High-throughput sequencing has contributed to expanding our previously knowledge of siRNA populations better describing their abundance, complexity and diversity in infected tissues. It is now known that siRNAs from virus-infected plants are extraordinarily abundant and diverse, and are widespread in near saturation at any region of either positive and negative genomic RNAs. However, certain regions of viral genomes ("hot spots") are usually more represented than others in sequenced siRNA populations. A gene involved in chlorophyll biosynthesis has been shown targeted by a siRNA derived from viral satellite RNA, revealing PTGS mechanism at basis of the symptom. Potato virus Y (PVY) is the type species of Potyvirus, a genus of agricultural importance belonging to the largest plant virus family, Potyviridae. The potyviral genome is a single-stranded, positive-sense RNA of ca. 10 kilobases (kb). PVYc-to and PVY-SON41 are two isolates of PVY that infect solanaceous hosts. While PVYc-to induces severe leaf distortion in different cultivars of tomato (*Solanum lycopersicum*), PVY-SON41 produces in the same host only a mild mosaic, followed by recovery. In order to elucidate the molecular mechanism underlying the different symptoms induced by PVYc-to and PVY-SON41 infections on tomato, we set up an in silico approach, mining genomic regions of PVY isolates and looking at possible RNA-based mechanisms where siRNA putatively generated from the PVY genome could target and suppress accumulation of host messenger RNA (mRNA), leading to dysfunctional biological

processes that could explain different disease phenotypes.

### Methods

A Perl script was used to extract the complete datasets of 21-mers from isolates PVYc-to and PVY-SON41 complete genomes, leading the scanning of the complete sequence, shifted by one base at the time. MySQL was used for identification of the 21-mers shared and unique between the two viral genomes, highlighting sequence differences that could be at the base of diverse induced symptoms. The data obtained in the previous step were used to build the genomic map of identical regions, by fancyGene. A BLAST analysis was conducted with the identified 21-mer dataset, considering only the dissimilar sequences between the two isolates (blastn-short identity > 94%, max 2 mismatch or gap, alignment length 21 bp). The 21-mer were used as query and the tomato mRNA database (Solgenomics, release ITAG2.3) was used as target. The results were further used as input in a gene ontology analysis, through Blast2GO.

### Results

Despite the high identity (> 91%) shared by the two viral genomes, the 21-mer dataset produced a high degree of variation, as the analyses showed 1750 identical and 9671 and 9680 dissimilar sets of 21-mer, for PVYc-to and PVY-SON41, respectively. Data showed that 315 and 381 tomato mRNA were complementary to, and thus possible targets of, the in silico-generated siRNAs deduced from of PVYc-to and PVY-SON41 sequences, respectively. The GO analysis showed different putative gene targets, involved in several metabolic pathways indicating specific targets for each isolate. Of particular interest, for instance, are putative target mRNAs of transcription factors with a known role concerning leaf development and symmetry. These genes, if differentially modulated during the infections of the two different PVY isolates, could cause the

leaf malformations observed as major differential symptoms between the two PVY isolates. To validate the bioinformatic approach, target genes with highly significant complementarity

have been selected from the in silico data for experimental validation in tomato plants mechanically infected by virus isolates.

## Plasmodium phylogenetic profile: an assessment of a predictive tool for protein-protein interactions

G. Sferra<sup>1</sup>✉, D. Santoni<sup>2</sup>, E. Pizzi<sup>1</sup>

<sup>1</sup>Istituto Superiore di Sanità, Roma, Italy

<sup>2</sup>Istituto di Analisi dei Sistemi ed Informatica, CNR, Italy

### Motivations

Prediction of protein-protein interactions (PPIs) is a crucial goal for bioinformatics and the increasing availability of sequenced genomes support this challenge. One of the computational tools mostly used for this aim is the phylogenetic profiling, based on the co-conserved proteins identification by local alignments. The master idea is that co-evolving proteins share the same phylogenetic profile and can be grouped functionally. The field of parasite biology received a powerful improvement from post-genomics data. Several approaches have been utilized to predict PPIs for *Plasmodium falciparum*, including phylogenetic profiling. In contrast, no information about this is available in the case of *Plasmodium berghei*, even though for this parasite a huge amount of data is now available especially for mutant phenotypes.

### Methods

A new strategy has been developed to derive phylogenetic profiling. The critical steps of this strategy are: 1) genomes selection; 2) global vs

local alignments comparison; 3) mutual information vs correlation coefficients calculation. Agreeing with specific criteria, 774 reference organisms, on 1133 available on EggNOG database (January 2012), were included in the study, a global alignment algorithm, over the mostly used local one, was used to perform proteins identification across the genomes, the Mutual Information and the Correlation Coefficient were calculated and the results were compared. *Escherichia coli* K12 before, suitable for the assessment of the method, and *P. berghei* later, were used as target organisms.

### Results

This analysis offers a new bioinformatical strategy to derive phylogenetic profile of an organism, highlighting on guidelines for the genomes selection, on the performance of different alignment algorithms and mathematical procedures. Moreover, strongly improve the knowledge about *P. berghei* and offers a new tool for evolution understanding and functional grouping of the proteins of this important biological model.

## Role of alternative splicing in modulating protein-protein interactions

V. Bianchi<sup>✉</sup>, A. Colantoni, M. Helmer-Citterich, F. Ferrè

Department of Biology, University of Tor Vergata, Rome, Italy

### Motivations

Alternative splicing (AS) permits the synthesis of multiple transcript variants from a single gene increasing the diversity of proteins encoded by a genome. Through the use of recent high-throughput sequencing technologies, it has been demonstrated that approximately 95% of multi-exon genes undergo AS in panels of human tissues [5], shaping the expressed transcriptome in various ways, from effectively turning off gene expression by the inclusion of early stop codons in the sequence, to subtle changes in protein function [1]. In addition, new data suggest that aberrant mRNAs generated through the AS machinery and their protein products have unique characteristics that confer new properties to cancer cells [2,3]. In this context it becomes crucial to link together heterogeneous data from different sources such as domain composition, protein structures, gene-interaction networks, in order to better understand AS mechanism and regulation, and its effects on protein products. We aim at a detailed analysis of how AS can modulate protein interactions by the differential expression of isoforms encoding or not for the interaction interfaces.

### Methods

We analyzed RNA-seq data from two experiments, a panel of 9 human tissues [6] and a panel of 16 human tissues (Illumina BodyMap 2.0), the former downloaded from the Gene Expression Omnibus repository (GEO identifier: GSE12946), the latter from the ArrayExpress Archive database (ID: E-MTAB-513). We used the Tuxedo suite (bowtie, tophat, cufflinks) to map RNA-seq reads on the reference human genome (the hg19 assembly) up to two mismatches [4] and evaluate tissue-specific expression level for each isoform annotated in the Ensembl database (release 65). In a recent work of our group, we identified all human hetero-dimeric interactions solved by X-ray crystallography present in the Protein Data Bank, and whose residues are involved in the formation of the protein-protein interface using distance and energy criteria. Such residues were

mapped to the hg19 human genome assembly to establish how many interface residues are part of each splicing isoform of a gene.

### Results

We calculated that a considerable amount of human genes, about 24%, for which an interaction is known at the molecular level in the PDB, encode for at least one isoform where all interface residues are lost due to an alternative splicing event. We computed splice isoform expression levels in all tissues under analysis, and draw tissue-level interaction maps based on the expression of splice variants encoding or losing the interface residues, detecting that in many cases (and in fractions that differ in different tissues), even if two binding partner genes are actively expressed, the usage of splicing isoforms not encoding for the interface residues prevents the interaction. The distribution of the number of tissues that lose the interface highlights a large number of ubiquitous interfaces (214 pairs out of 620) and a smaller number of tissue-specific interfaces, which are lost in all except one tissue (32 pairs). The functional characterization made by enrichment in GO terms confirmed the characteristics of the ubiquitous and tissue-specific interactions, the former being mostly involved in general metabolic pathways, the latter involved in tissue-specific pathways. Our results indicate that AS is a powerful modulator of protein interactions, and that splicing isoform usage is finely tuned to allow or prevent specific interactions.

### References

1. David CJ, Manley JL: The search for alternative splicing regulators: new approaches offer a path to a splicing code. *Genes and development* 2008, 22:279.
2. Fackenthal JD, Godley LA: Aberrant RNA splicing and its functional consequences in cancer cells. *Disease models and mechanisms*, 1:37-42.
3. Kim E, Goren A, Ast G: Insights into the connection between cancer and alternative splicing. *Trends in genetics*: TIG 2008, 24:7-10.
4. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 2009, 10:R25.

5. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* 2008, 40:1413-5.
6. Wang ET, Sandberg R, Luo S, et al.: Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* 2008, 456:470.

## GC content dependency of open reading frame prediction

M. Pohl<sup>1</sup> ✉, G. Theißen<sup>2</sup>, S. Schuster<sup>1</sup>

<sup>1</sup>Department of Bioinformatics, Friedrich Schiller University, Jena, Germany

<sup>2</sup>Department of Genetics, Friedrich Schiller University, Jena, Germany

### Motivations

A frequently used approach for detecting potential coding regions is to search for stop codons. In the standard genetic code 3 out of 64 trinucleotides are stop codons. Hence, in random or non-coding DNA one can expect every 21st trinucleotide to be equivalent to a stop codon. In contrast, the open reading frames (ORFs) of most protein coding genes are considerably longer. Thus, the stop codon frequency in coding sequences deviates from the background frequency of the corresponding trinucleotides. This has been utilized for gene prediction, in particular, in detecting ORFs. Traditional methods based on stop codon frequency are based on the assumption that the GC content is about 50 %. However, many genomes show significant deviations from that value.

### Methods and Results

With the presented method we can describe the effects of GC content on the selection of appropriate length thresholds of ORFs. Conversely, for a given length threshold, we can calculate the probability of observing it in a random sequence. Thus, we can derive the maximum GC content for which ORF length is practicable as a feature for gene prediction methods and the resulting false positive rates. A rough estimate for an upper limit is a GC content of 80 %. This estimate can be made more precise by including further parameters and by taking into account start codons as well. We demonstrate the feasibility of this method by applying it to the genomes of the bacteria *Rickettsia prowazekii*, *Escherichia coli* and *Caulobacter crescentus*, exemplifying the effect of GC content variations according to our predictions.

## Whole genome sequencing, assembly and annotation of *Halomonas smyrnensis*, a levan producing halophilic bacterium

E. Sogutcu<sup>1</sup>✉, Z. Emrence<sup>2</sup>, M. Arikan<sup>2</sup>, A. Cakiris<sup>2</sup>, N. Abaci<sup>2</sup>, E. Toksoy Oner<sup>1</sup>, D. Ustek<sup>2</sup>, K.Y. Arga<sup>1</sup>

<sup>1</sup>Department of Bioengineering, Marmara University, Goztepe, Istanbul, Turkey

<sup>2</sup>Department of Genetics, Institute for Experimental Medicine, Istanbul University, Istanbul, Turkey

### Motivations

Next-generation sequencing technologies have been available in the relatively short time frame providing a wealth of data and specific information that cannot be obtained by other experimental approaches. In systems biology research, the microbial genome sequence is the starting point for detailed analysis of identifying gene-protein associations and metabolic reconstruction. For this purpose, we performed whole genome sequence analysis of *Halomonas smyrnensis*, which has been reported as a high level levan-producer halophilic bacterium for the first time by our research group. Levan is a linear fructose polymer and has many potential uses in foods, feeds, cosmetics, and the pharmaceutical and chemical industries.

### Methods

A hybrid strategy is performed both in sequencing and assembly process. The genome sequence of *Halomonas smyrnensis* has obtained using two independent experimentation: Roche

454 GS FLX+ System and Ion Torrent Sequencer. Reads have de novo assembled into scaffolds using the 454 Newbler, CLC Genomics Workbench 5.0.1 and Geneious Pro 5.5.6 assembler software. The draft genome has structured with assembled contigs that have ordered by CONTIGuator bacterial genome finishing tool. Thereafter we have performed draft genome annotation via RAST annotation server.

### Results

The Roche 454 GS FLX+ System (1,442,441 reads with an average length of 495.14 bp) and Ion Torrent Sequencer System (359,558 reads with an average length of 122.42 bp) resulted with 260x coverage. As a result of the de novo assembly process, 105 high-quality contigs with size greater than 8000 bp were obtained. Among those, 42 contigs were ordered and structured with a genome size of 4,242,280 bp (G+C content 67.8%). The annotation of draft genome ended up with 3845 coding sequences and 96 RNAs.

## ITSoneDB: a specialized ITS1 database for amplicon-based metagenomic characterization of environmental fungal communities

B. Fosso<sup>1</sup>✉, M. Santamaria<sup>2</sup>, A. Consiglio<sup>3</sup>, G. De Caro<sup>3</sup>, G. Grillo<sup>3</sup>, F. Licciuli<sup>3</sup>, S. Liuni<sup>3</sup>, M. Marzano<sup>2</sup>, G. Pesole<sup>1</sup>

<sup>1</sup>Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli Studi di Bari "A. Moro", Bari, Italy

<sup>2</sup>Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

<sup>3</sup>Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy

### Motivations

Metagenomics is experiencing an explosive improvement from the advent of high-throughput next-generation sequencing (NGS) technologies which allows an unprecedented large-scale identification of microorganisms living in almost every environment. In particular, the use of amplicon-based metagenomic approach to explore the diversity of fungal environmental communities is increasingly expanding. At the species level, a number of studies have used the non-conserved internal transcribed spacers (ITS) 1 and 2 of the ribosomal RNA genes cluster as genetic markers to explore the fungal taxonomic diversity. Particularly, ITS1 is gaining an increasing popularity as better discriminating species marker in Fungi because of its higher variability compared to ITS2. Starting from the total DNA extracted from any environmental sample, this locus can be easily amplified with taxonomically universal primers and sequenced by means of high-throughput next generation platforms. Reference databases and robust supporting taxonomies are crucial in assigning phylogenetic affiliation to the huge amount of produced sequences. Even if a large number of ITS1 sequences are collected in public databases, a specialized resource focused particularly on this region, where sequences identity, boundaries and taxonomic assignment are validated, is still needed at present. In this work we present ITSoneDB, a new comprehensive collection of ITS1 sequences belonging to Fungi Kingdom.

### Methods

ITSoneDB has been generated and populated using a multi-step Python workflow. In the first step the ribosomal RNA gene cluster sequences of Fungi including the target ITS1 region were retrieved from Genbank. Then, ITS1 start and end boundaries were extracted from the Features Tables annotations, if available. In order to infer, validate and, eventually, redesign the ITS1 location, Hidden Markov Model (HMM) profiles of flanking genes for 18S and 5.8S ribosomal RNA, generated from their reference alignments stored in RFAM database, were mapped on the entire collection of retrieved nucleotide sequences, by means of the *hmmsearch* tool from HMMER 3.0 package.

### Results

At present, ITSoneDB includes 405,433 taxonomically arranged sequence entries provided with ITS1 both start and end positions defined by GenBank annotations and/or HMM based method. ITSoneDB front-end is a JAVA platform-based website for data browsing and downloading. The database can be queried by species or taxon name, GenBank accession ID or by "expanding" the target rank on a detailed fungal taxonomical tree. The complete ITS1 sequences dataset collected in ITSoneDB is available in Fasta format and the users can extract and locally save all or selected queried ITS1 sequences for further analysis.

### Availability

<http://itsonedb.ba.itb.cnr.it/>

## Towards an integrated resource for the study of population and disease associated variability of the human mitochondrial genome

M.A. Diroma , M. Attimonelli

Department of Biosciences, Biotechnologies and Pharmacological Sciences, University of Bari, Italy

### Motivations

Thanks to the development of valid DNA sequencing technologies and statistical methods of analysis, nowadays we have a great amount of human mtDNA data. The mitochondrion genome presents a high rate of variability; the turning point of a polymorphism into a mutation is linked to heteroplasmy. A variability analysis of sequenced human mtDNAs has been performed by comparing published sequences with the reference sequence, rCRS, to detect and to characterize polymorphic positions in order to recognize relationships between i. variability values and haplogroups patterns, and ii. variability and pathogenicity.

### Methods

Different web resources, such as Phylotree [1] and MITOMAP [2] besides GenBank and Pubmed, have allowed a variability analysis that has been carried out on mitochondrion genomes of healthy and pathologic individuals within HmtDB [3] where an evaluation of site-specific nucleotide and aminoacid variability is implemented, SiteVar [4]. Pathogenicity analysis has been possible by applying i. Polyphen2 [5] which compares wild type and variant alleles based on the aminoacidic conservation observed in proteins multialignment, and ii. SNPs&GO [6], a method based on support vector machines, for the prediction of functional effects of human non synonymous SNPs on mitochondrial proteins.

### Results

The usage of SiteVar on continent-specific datasets has supported the knowledge of specific variability values for each ethnic group leading to the classification of haplogroups and to the characterization of potentially pathogenic mutations, even if there is no completely agreement among methods of prediction of pathogenicity yet. Among 5902 variable sites on the entire mito-

chondrial genome a half fits with those reported in Phylotree and 67 show the highest variability values. Patients present a considerable increase of low variability positions which could be specific mutations associated to a specific pathology while a lot of somatic mutations corresponds to polymorphisms in healthy individuals defining a specific haplogroup, suggesting that these mutations may be not necessarily associated to a pathology. Conclusion: The available results suggest that the integration of the abovementioned web resources, Phylotree, MITOMAP and HmtDB, could surely add values to the knowledge concerning human mitochondrial DNA, population histories and mitochondrion associated diseases. This is a message to the Bioinformatics community to activate interdisciplinary collaborations.

### References

1. van Oven, M. and Kayser, M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, 30(2): E386-E394.
2. Ruiz-Pesini, E., Lott, M.T., Procaccio, V., Poole, J., Brandon, M.C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P. and Wallace, D.C. (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Research*, 35 (Database issue):D823-D828.
3. Rubino, F., Piredda, R., Calabrese, F.M., Simone, D., Lang, M., Calabrese, C., Petruzzella, V., Tommaseo-Ponzetta, M., Gasparre, G. and Attimonelli, M. (2012) HmtDB, a genomic resource for mitochondrion-based human variability studies. *Nucleic Acids Res.*, 40(D1), D1150-D1159.
4. Pesole, G. And Saccone, C. (2001) A novel method for estimating substitution rate variation among sites in a large dataset of homologous DNA sequences. *Genetics*, 157, 859-865.
5. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010), *Nat. Methods*, 7(4), 248-249.
6. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L. and Casadio, R. (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation*, 30(8), 1237-1244.

## Analysis of barcode sequences by means of compression-based methods

M. La Rosa , A. Fiannaca, R. Rizzo, A. Urso

ICAR-CNR, National Research Council of IT, Palermo, Italy

### Motivations

The key idea of DNA barcode initiative is to identify, for each group of species belonging to different kingdoms of life, a short DNA sequence that can act as a true taxon barcode. DNA barcode represents a valuable type of information that can be integrated with ecological, genetic, and morphological data in order to obtain a more consistent taxonomy. Recent studies have shown that, for the animal kingdom, the mitochondrial gene cytochrome c oxidase I (COI), about 650 bp long, can be used as a barcode sequence for identification and taxonomic purposes of animals. The analysis of DNA barcode sequences is carried out with well known bioinformatics techniques: for example the most common approach to create a phylogenetic tree for a group of species uses the multi-alignment of genetic sequences, the computation of a dissimilarity matrix, using one of the current available evolutionary distance model, and finally the building of a tree by means of hierarchical algorithms such as Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbor Joining (NJ). In the present work we aim at introducing the use of an alignment-free approach in order to make taxonomic analysis of barcode sequences. Our approach is based on the use of two compression-based versions of non-computable Universal Similarity Metric (USM) class of distances. This way we try to overcome some flaws of classic techniques, such as the time-consuming and parameter-dependent alignment procedure and the use of stochastic evolutionary distance models, that do not represent a distance metric.

### Methods

Universal Similarity Metric represents a class of distance measures based on the non-computable Kolmogorov complexity. That means it needs some approximation in order to be used. USM is said to be "universal" because it can be applied

for computing a distance matrix among input data belonging to very different application domains. In fact it has been used for the analysis of text, images, music. In bioinformatics, it has been applied for obtaining phylogenetic trees from complete mitochondrial genome of mammalian species. Our purpose is to justify the employ of USM also for the analysis of short DNA barcode sequences, showing USM is able to correctly extract taxonomic information among those kind of sequences. We, then, downloaded from Barcode of Life data System database (BOLD) 20 datasets of barcode sequences belonging to different animal species. For each dataset we computed dissimilarity matrices by means of two compression-based approximation of USM, namely Normalized Compression Distance (NCD) and its conditional compression version. In both cases we used GenCompress compressor, that is a dictionary-based compressor suited to work with DNA sequences. From those matrices we built, using UPGMA and NJ algorithms, phylogenetic trees of every dataset and compared them, in terms of topology preservation, with the trees obtained through Kimura 2-parameter evolutionary distance model.

### Results

Experimental tests aim to evaluate the quality of phylogenetic reconstruction in terms of both topological similarity and differences in the relative branch length. As regard the tree similarity, we obtain good results with a percentage of similarity between evolutionary and compression-based tree greater than 82% and, for the most of datasets, between 90% and 100%. Lower results are for datasets having an high percentage of sequences with ambiguous bases. We detect the same trend for differences in the relative branch of trees, except that poorer results are reached by those datasets containing some COI-5P gene sequences longer than the other ones.

## NGS TREX: next generation sequences transcriptome profile explorer

I. Boria<sup>1</sup>, G. Pesole<sup>2</sup>, F. Mignone<sup>3</sup>✉

<sup>1</sup>Department of General and Environmental Physiology, University of Bari, Bari, Italy

<sup>2</sup>Department of Biochemistry and Molecular Biology, University of Bari, Bari, Italy

<sup>3</sup>Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale, Alessandria, Italy

### Motivations

Next-generation sequencing (NGS) technology has exceptionally increased the ability to sequence DNA in a massively parallel and cost-effective manner. Nevertheless, the management and the analysis of NGS data requires significant expertise in bioinformatics and hardware infrastructure still beyond the possibilities of many laboratories focused on "wet biology". Moreover some projects only need few deep sequencing cycles and standard tools or workflows to carry out suitable analyses for the identification and annotation of genes, transcripts and splice variants found in the biological samples under investigation. The development of easy to use systems to automatically analyze and annotate NGS data is needed to allow researchers from different backgrounds to take full benefit of NGS technologies.

### Results

We developed an automatic system targeted to the analysis of Next Generation Sequencing

data obtained from large-scale transcriptome studies. This system, we named NGS-Trex (NGS Transcriptome profile Explorer) is available through a simple web interface and allows the user to upload raw sequences and easily obtain an accurate characterization of the transcriptome profile after the setting of few parameters required to tune the analysis procedure. The system is also able to assess differential expression at both gene and transcript level (i.e. splicing isoforms) by comparing the expression profile of different samples. By using simple query forms the user can obtain list of genes, transcripts, splice sites ranked and filtered according several criteria. Data can be viewed as tables and downloaded as text files to allow further analysis. Moreover a simple genome browser helps the visual inspection.

### Availability

<http://www.ngs-trex.org/>

## Integrated analysis of epigenetic and transcriptional circuits in gliomagenesis

G. Bucci<sup>1</sup>, E. Signaroldi<sup>2</sup>, P. Laise<sup>2</sup>, P.-L. Germain<sup>2</sup>, L. Zammataro<sup>1</sup>, H. Muller<sup>1</sup>, G. Testa<sup>2</sup>✉

<sup>1</sup>Center For Genomics Science IIT@SEMM, Istituto Italiano di Tecnologia, Italy

<sup>2</sup>European Institute of Oncology at the, IFOM-IEO Campus, Milan, Italy

### Motivations

Next Generation Sequencing (NGS) technologies have conceptually changed the planning of a molecular biology experiment, modifying the balance between the 'wet-lab' and the 'dry' aspects of the research in favor of the latter. Better and reliable results could be gained with the new technologies when adequate bioinformatics resources are allocated. In the last couple of years new and refined methods and algorithms have been developed to fully exploit this new data generation. Good results have been achieved especially for mRNA-Seq and ChIP-Seq experiments, taking advantage of the publicly available bioinformatics pipelines and of the robust analytical tools [1, 2]. In particular, combined analysis at both the transcriptomic and epigenomic level affords the opportunity of a much deeper characterization of biological samples, enabling the elucidation of the interlaced connections between genotypes and phenotypes. The development of ever more effective methods to integrate high-throughput data is thus a major challenge in the life sciences, as captured by Venkatesh and Harlow in 2002 [3] in their reframing of the concept of "integration" in molecular biology, also referred to as "Integromics", as the bioinformatic integration of high throughput 'omics' data [4, 5].

### Methods

Here we advance in this integration effort and present new results obtained with a defined protocol of analysis, which combines the most recent NGS bioinformatic tools into an integrated pipeline aimed at investigating the transcriptional and epigenetic deregulation that underlies gliomagenesis. Malignant gliomas represent the most common form of primary brain tumor, comprising a pathologically and genetically heterogeneous set of tumor types, whose extremely poor prognosis has not significantly improved in the last decades, with Glioblastoma multiforme (GBM), the most malignant glioma subtype, char-

acterized by an average prognosis of less than one year [6, 7]. While genetic lesions and transcriptional profiles have been widely investigated in GBM [8], the chromatin-wide deregulation that mediates transcriptional changes, and whose effectors may constitute rational therapeutic targets, has not been uncovered. Specifically, despite convergent lines of evidence pointing to profound epigenome aberrations at the level of DNA methylation [9-11] and histone modifications [12], we still do not understand the role that epigenetic alterations play in the initiation, progression and the recurrence of disease, which in turn prevents the definition of epigenetic pathways suitable as prognostic signatures or rational interventions.

### Results

Here we employed a murine model that faithfully recapitulates the most common genetic lesions (Ink4a/Arf inactivation; constitutive EGFR signaling) and pathological hallmarks of human GBM [13]. We used ChIPseq for Histone 3 Lysine 27 trimethylation (K27Me3) and mRNA-seq in order to integrate the analysis of the transcriptomic and epigenomic aberrations in tumorigenic astrocytes and glioma initiating cells (GICs) at, respectively, the onset and end of gliomagenesis.

### References

1. McCarthy DJ, Chen Y, Smyth GK: Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research* 2012:gks042-.
2. Micsinai M, Parisi F, Strino F, Asp P, Dynlacht BD, Kluger Y: Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic acids research* 2012:1-16.
3. Venkatesh T, Harlow H: Integromics: challenges in data integration. *Genome Biology* 2002, 3:reports4027.1-reports4027.3.
4. Hawkins RD, Hon GC, Ren B: Next-generation genomics: an integrative approach. *Nature reviews. Genetics* 2010, 11:476-86.

5. L  Cao K-A, Gonz lez I, D jean S: integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics (Oxford, England)* 2009, 25:2855-6.
6. Jiang Y, Uhrbom L: On the origin of glioma. *Upsala journal of medical sciences* 2012:1-9.
7. van den Bent MJ, Kros JM: Predictive and prognostic markers in neuro-oncology. *Journal of neuropathology and experimental neurology* 2007, 66:1074-81.
8. Eoli M, Silvani A, Pollo B, Bianchessi D, Menghi F, Valletta L, Broggi G, Boiardi A, Bruzzone MG, Finocchiaro G: Molecular markers of gliomas: a clinical approach. *Neurological research* 2006, 28:538-41.
9. Martinez R, Esteller M: The DNA methylome of glioblastoma multiforme. *Neurobiology of disease* 2010, 39:40-6.
10. Martinez R, Martin-Subero JI, Rohde V, Kirsch M, Alaminos M, Fernandez AF, Ropero S, Schackert G, Esteller M: A microarray-based DNA methylation study of glioblastoma multiforme. *Epigenetics*: official journal of the DNA Methylation Society 2009, 4:255-64.
11. Weller M, Stupp R, Reifenberger G, Brandes AA, van den Bent MJ, Wick W, Hegi ME: MGMT promoter methylation in malignant gliomas: ready for personalized medicine? *Nature reviews. Neurology* 2010, 6:39-51.
12. Lee J, Son MJ, Woolard K, Donin NM, Li A, Cheng CH, Kotliarova S, Kotliarov Y, Walling J, Ahn S, Kim M, Totonchy M, Cusack T, Ene C, Ma H, Su Q, Zenklusen JC, Zhang W, Maric D, Fine HA: Epigenetic-mediated dysfunction of the bone morphogenetic protein pathway inhibits differentiation of glioblastoma-initiating cells. *Cancer cell* 2008, 13:69-80.
13. Bachoo RM, Maher EA, Ligon KL, Sharpless NE, Chan SS, You MJ, Tang Y, DeFrances J, Stover E, Weissleder R, Rowitch DH, Louis DN, DePinho RA: Epidermal growth factor receptor and Ink4a/Arf: convergent mechanisms governing terminal differentiation and transformation along the neural stem cell to astrocyte axis. *Cancer cell* 2002, 1:269-77.

## A next generation sequencing-based approach to identify piRNAs in breast cancer cells

C. Cantarella<sup>1</sup>✉, C. Stellato<sup>1</sup>, G. Giurato<sup>1</sup>, M.R. De Filippo<sup>2</sup>, R. Tarallo<sup>1</sup>, A. Weisz<sup>3</sup>

<sup>1</sup>Laboratory of Molecular Medicine and Genomics, University of Salerno, Baronissi, Italy

<sup>2</sup>Fondazione IRCCS SDN, Naples, Italy

<sup>3</sup>Division of Molecular Pathology and Medical Genomics, 'SS. Giovanni di Dio e Ruggi d'Aragona' Hospital, University of Salerno, Salerno, Italy

### Motivations

There are RNA species produced by eukaryotes, classified as non-coding RNAs, which are not involved in translation, but play a major role in regulation of gene expression at the transcriptional, post-transcriptional and translational level. The small non-coding RNA (sncRNA) molecules, along with the Argonaute family of proteins, have been identified as key players in various forms of sequence-specific gene silencing, including RNA interference (RNAi), translational repression and heterochromatin formation. Next-generation sequencing methods have allowed researchers to quickly sequence and profile sncRNA populations. Recently, a new class of non-coding small RNAs has been found, named Piwi-interacting RNAs (piRNAs), which interact with Piwi and are produced by a Dicer-independent mechanism. These non-coding small RNAs regulate a series of small RNA-mediated mechanisms that modulate a large variety of biological processes, such as silencing of selfish DNA elements, development, genome stability, and DNA integrity maintenance. It was reported that piRNA-pathway disorders increase the repeats of retrotransposon and cause DNA damage, both common occurrence in tumorigenesis of germline and somatic cells. Currently, piRNAs have been identified in human cancer cells, regulated by the Hili protein and involved in carcinogenesis. Another report describes how piRNAs are aberrantly expressed in human cancer cells. Other small molecules playing a central role in many RNAi mechanism are: small-interfering RNAs (siRNA) and transcription initiation RNAs (tiRNAs). While it has not been definitively demonstrated that they play a causal role in human disease, this does not necessarily preclude the existence of disease-mediating siRNAs or tiRNAs. Recent studies showed as endogenous siRNAs are derived from naturally occurring double-stranded RNAs (dsRNAs) and have roles in the regulation of gene expression in mouse oocytes. On the other hand, tiRNAs

are nuclear localized 18nt RNAs derived from sequences immediately downstream of RNA polymerase II (RNAPII) transcription start sites. Several reports have shown that tiRNAs are intimately correlated with gene expression, RNA polymerase II binding and behaviours, and epigenetic marks associated with transcription initiation, but not elongation.

### Methods

Small non coding RNAs sequences were obtained by sequencing human cancer cells RNA with an Illumina Genome Analyzer. In order to identify piRNAs expressed in the samples, we used the miRanalyzer tool with the reference databanks RNAdb and Rfam. Two databanks were filtered to reduce redundancies and integrated with data available in public repositories. Programs developed in house (using Perl and PHP languages) have been applied to identify piRNAs not yet annotated, called putative "ping-pong piRNAs". Software is based on the following experimental evidences: there are primary (p-piRNAs) and secondary (s-piRNAs) piRNAs; the initial cleavage site is located within the base-paired region 10nt downstream of the 5' terminal U of the p-piRNA and the resulting s-piRNAs are therefore distinguished by an A at position 10. Some s-piRNAs are expected to be reverse complementary to the original p-piRNA precursor transcript and may themselves be able to direct cleavage of these to recreate the original p-piRNA. During piRNA-guided cleavage of target transcripts a 19-mer product arises. Although 19-mers do appear to associate with Piwi proteins, the fact that they are not stabilized by 3' end methylation argues against their function as piRNAs. Their size and the apparent lack of the 3' end modification would allow them to load into Argonaute proteins and function in RNAi-like silencing as short interfering RNAs. Results have been compared using the recently developed program piRNAPredictor, which uses a k-mer scheme to identify piRNA se-

quences, relying on the training sets of non-piRNA and piRNA sequences of five model species.

### Results

Starting from available tracks of human piRNAs, a non-redundant databank has been assembled removing sequences with ambiguous nucleotides and sequences mapping within the same locus of the human genome but annotated with different accession number. The databank has been used to identify piRNAs molecules in two distinct sequencing experiments, with different coverages. Some piRNAs are differentially expressed in samples respect to controls, revealing their possible involvement in breast cancer.

The programs developed to identify piRNAs not yet annotated, produced by “ping-pong model”, have been tested both with reference piRNA sequences and experimental data. Analysis of sequences reads highlights the presence of the three molecules produced during piRNA biogenesis: primary, secondary and 19-mer piRNA.

### Acknowledgements

Research supported by: Fondazione con il Sud; Italian Association for Cancer Research; Italian Ministry for Education, University and Research; Regione Campania; University of Salerno; Fondazione Umberto Veronesi.

## Development of pipeline for exome sequencing data analysis

M.R. De Filippo<sup>1</sup> ✉, G. Giurato<sup>1</sup>, C. Cantarella<sup>1</sup>, F. Rizzo<sup>1</sup>, F. Cirillo<sup>1</sup>, A. Weisz<sup>2</sup>

<sup>1</sup>Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, University of Salerno, IT, Naples, Italy

<sup>2</sup>Division of Molecular Pathology and Medical Genomics, 'SS Giovanni di Dio e Ruggi d'Aragona' Hospital, University of Salerno, Italy

### Motivations

Exome sequencing the targeted sequencing of the subset of the protein coding human genome is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Until now many algorithms have been produced, each of them addressing a different task in the downstream analysis of next-generation sequencing (NGS) data. The aim of this work is to combine these algorithms into an analysis pipeline for the detection of SNP and deletion/insertion polymorphisms within DNA sequences obtained by whole exome sequencing. The pipeline tested with data obtained from SRA (<http://www.ncbi.nlm.nih.gov/sra>), will then be applied to studies undergoing in our laboratory.

### Methods

Starting from raw sequence data, we first performed quality statistics and filtering of sequence reads and then aligned them to a reference genome. To this end, BWA was used to align both single- and paired-end reads for its computa-

tional efficiency and multi-platform compatibility. Post-alignment analysis, including removal of duplicate reads and quality score recalibration, was carried out using GATK, which takes into account several covariates such as machine cycle and dinucleotide context. Next, SNP calling was done using GATK UnifiedGenotyper, that uses a Bayesian model to estimate the most likely genotypes and allele frequency in a population of N samples, giving an annotated VCF file as output. Subsequently, variant quality score was recalibrated to estimate the probability of each variant being a true polymorphism, rather than a sequencer, alignment or data processing artifact, and finally filtered to improve the accuracy of genotype and SNP calling.

### Results

The results obtained support the accuracy of our pipeline to identify SNP and short indels, to provide a global and quantitative catalog of nucleotide variants in the exome. The next step will be to apply this pipeline to samples sequenced in our laboratory.

## RNA sequencing data: biases and normalization

F. Finotello<sup>1</sup>✉, E. Lavezzo<sup>2</sup>, L. Barzon<sup>2</sup>, P. Fontana<sup>3</sup>, A. Si-Ammour<sup>3</sup>, S. Toppo<sup>2</sup>, B. Di Camillo<sup>1</sup>

<sup>1</sup>Department of Information Engineering, University of Padova, Italy

<sup>2</sup>Department of Molecular Medicine, University of Padova, Italy

<sup>3</sup>Edmund Mach Foundation, San Michele all'Adige, Trento, Italy

### Motivations

In recent years, RNA sequencing (RNA-seq) has rapidly become the method of choice for measuring and comparing gene transcription levels. Despite its wide application, it is now clear that this methodology is not free from biases and that a careful normalization procedure is the basis for a correct data interpretation. The most common normalization techniques account for: library size, gene or transcript length and sequence-specific biases such as GC-content effects. The aim of the present work is to investigate biases affecting RNA seq data and their effect on differential expression analysis. In order to reduce biases due to over-simplification of gene transcription models, we consider exon-based counts.

### Methods

We two used publicly available RNA-seq data sets from two-group comparison studies which are characterized by multiple technical replicates. We summarized read counts at exon level and investigated their dependence on sequence-specific covariates: GC-content and exon length. In addition, we considered the effect of library size correction on between-groups comparison and the impact of the above mentioned biases on the detection of differentially expressed exons. The assessment is performed on raw data, as well as on data normalized with different approaches: RPKM [1], library size scaling, based on Trimmed Mean of M-values (TMM) [2] and on Poisson goodness-of-fit statistic applied to non differentially expressed genes [3], and within-lane normalization based on loess regression of log-counts on GC-content and exon length [4]. We selected differentially expressed exons using the GLM-based version of edgeR [5] as it can consider an "offset" matrix which codi-

fies counts normalization, that can be computed with the desired approach, and library size scaling factors specified by the user.

### Results

In our study, read counts show a significant dependence on exon length and a moderate dependence on GC-content. Exon length bias also affects differential expression analysis: longer exons tend to have lower P-values and to be selected as differentially expressed more frequently than shorter exons. The tested normalization techniques do not completely remove biases and, in particular, RPKM approach over-corrects for exon length bias. Moreover, the choice of the strategy for library size adjustment has a great impact on the direction of the detected differential expression. The results obtained on these data sets demonstrate that RNA-seq data normalization is still an open issue. Further efforts should be directed towards the clarification of the relationship between read counts and sequence-specific biases, which are, in turn, correlated to each other, and the definition of new models for their correction.

### References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature methods*. 2008;5(7):621-8.
2. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25.
3. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*. 2011.
4. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-seq data. *BMC Bioinformatics*. 2011;12(1):480.
5. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012.

## An accurate pipeline for analysis of NGS data of small non-coding RNA

G. Giurato<sup>1</sup>✉, M.R. De Filippo<sup>2</sup>, C. Cantarella<sup>1</sup>, G. Nassa<sup>1</sup>, M. Ravo<sup>1</sup>, E. Nola<sup>3</sup>, A. Weisz<sup>4</sup>

<sup>1</sup>Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, University of Salerno, Italy

<sup>2</sup>Fondazione IRCCS SDN, Naples, Italy

<sup>3</sup>Department of General Pathology, Second University of Naples, Naples, Italy

<sup>4</sup>Division of Molecular Pathology and Medical Genomics, SS Giovanni di Dio e Ruggi d'Aragona Hospital, University of Salerno, Italy

### Motivations

The discovery of various families of small non-coding RNAs (sncRNAs) in recent years revealed the complexity of the regulation of gene expression at both transcriptional and post-transcriptional level. Of the numerous sncRNAs, microRNAs (miRNAs) constitute a large family of 19-23 nucleotides long RNAs that participate in a variety of processes, such as cell development and differentiation, apoptosis and stress responses to carcinogenesis. Computational analysis indicates that a unique miRNA can regulate hundreds of genes, underlining the potential influence of miRNAs in almost every cellular pathway. Deep sequencing technologies provides a powerful strategy to explore miRNA populations (miRNA-Seq) with high sensitivity and specificity. Different computational approaches have been developed to analyze miRNA-Seq data, allowing to identify known and novel miRNAs, perform differential expression analysis and predict putative miRNAs targets. We combined these algorithms into an analysis pipeline and tested it on data obtained from our experiments in cancer cell lines.

### Methods

The data obtained from the sequencer were filtered following several criteria. Since the sequence of the adapter is known, a Perl script was used to trim, from the raw data, the adaptors. The sequence reads were then filtered for quality and clustered to unique sequences to remove redundancy, retaining their individual read count information. Unique sequences 18 nucleotides or more in length were mapped, allowing up to one mismatch, on miRNA annotation according to miRBase version 18 using miRanalyzer. This detects the reads which correspond to known miRNAs, giving an estimation of expression level. miRBase repository is used because it offers information about mature (the mature sequence of known miRNAs), mature-star (the sequence

which pairs with the mature miRNA in the miRNA secondary structure) and precursor miRNA sequences (sequence of the hairpin). miRNAs have been considered as expressed if they are detected at least 5 reads/sample. After detecting those that correspond to known miRNAs, the remaining reads were mapped to databases of transcribed sequences as mRNA and non-coding RNA (RFam). This step has two goals: (i) the number of matches can be viewed as a sample quality parameter (contamination of the RNA sample with degradation products and poly A tails) and (ii) it might be interesting to see which other known sncRNAs are in the sample. To predict novel miRNAs we used a probabilistic algorithm, miRDeep2, based on miRNA biogenesis model, to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. This tool aligns sequencing reads to potential hairpin structures in a manner consistent with Dicer processing and assigns log-odds scores to measure the probability that hairpins are true miRNA precursors. To detect novel miRNAs by miRDeep2, a score cutoff corresponding to a prediction signal-to-noise ratio >10 was used. Identification of differentially expressed miRNAs was performed with the Bioconductor DESeq package. Starting from the expression values, the first step was to minimize the effect of the systematic technical variations, and then a negative binomial distribution model was used to test differential expression in deep sequencing datasets. Only miRNAs with a p-values less or equal to 0.05 and fold-change less or equal to -1.5 and greater or equal to 1.5 were considered as differentially expressed. Given the critical roles of miRNAs in regulating gene expression and cellular functions, we predicted their putative targets, intersecting results obtained from two resources, TargetScan and microRNA.org. TargetScan provide computationally predicted miRNA gene targets by searching for the presence of 8 and 7 mer sites that match

the seed region of each miRNA, while microRNA.org target prediction incorporates current knowledge on target rules and on the use of a compendium of mammalian miRNAs. A further step of the analysis was to investigate nucleotide variations relative to the reference genome. To this purpose, preliminary steps were to reduce alignment artifacts and compute a more accurate quality estimation, removing biases due to sequencing cycle and preceding nucleotide. Further evidences were used to identify new miRNA variation sites: (i) Sequencing depth of variation sites should be equal to or larger than 5 reads per site, (ii) frequency of Single Nucleotide Variant occurrence >5% and (iii) variants not found in previous SNP annotation databases, like dbSNP.

### Results

We developed an accurate pipeline for integral analysis of next generation sequencing data of

small RNA molecules. Based on solid statistical methods, this allows both detection of known miRNAs and prediction of new miRNAs, integrating steps for differential analysis, sequence analysis and target prediction.

### Acknowledgements

Research support by: Fondazione con il Sud; Italian Association for Cancer Research; Italian Ministry for Education, University and Research; Regione Campania; University of Salerno; Fondazione Veronesi. Giorgio Giurato is a student of PhD School in Experimental and Clinic Medicine / Doctorate in Experimental Physiopathology and Neuroscience, Second University of Naples. Maria Ravo is supported by a 'Vladimir Ashkenazy' fellowship of Italian Association for Cancer Research. Concita Cantarella and Giovanni Nassa are fellows of Fondazione con il Sud.

## Analysis workflow for the identification of allelic variant associated with a complex disease using NGS approach

V. Maselli✉, D. Cittaro, E. Stupka

Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Milan, Italy

### Motivations

Recent advances in sequencing technologies allowed for unprecedented possibilities and applications for clinical data analysis. In particular, Whole Genome sequencing at decent coverage has turned into cost-effective technology to characterize the genetic framework of rare diseases. We here present an example of a complex disease. Our purpose is to identify allelic variant associated with the described syndrome using a Next-Generation Sequencing approach.

### Methods

We performed a 100 bp paired-end sequencing of a single human genomic sample using a Illumina HiSeq 2000 sequencer. Read tags were aligned to the hg19 reference genome using BWA [1]. The Genome Analysis ToolKit was used to pipeline the downstream analysis: local realignment around indels, quality score recalibration, SNP and indel calling and, most important, Variant Quality Score Recalibration. We filtered SNV with a confidence lower than 0.1%. We used the Seattle SNP Annotation tools [2] in order to annotate the SNP on the reference genome. We performed some preliminary statistics in order to identify a threshold that allowed us to identify a reliable subset of SNPs to use for our purpose. Using the sub set of known SNPs as guide we identify the value of the SNP quality, for which we are confident regarding our data. In a first step we used a single lane data to validate the homozygosity analysis procedure. Using a so defined quality threshold of 50 we identi-

fied a subset of SNPs that we analysed with the HomozygosityMapper web tools [3]. Analysis on a double lane is work in progress.

### Results

We sequenced almost 400 million of read pairs achieving a 20x average coverage. We filtered out 2 million of SNPs with a read depth of at least 5 and max 250. The average quality above 400k SNPs is 313 (max 9371, min 30). We found about 100k novel SNPs (10% homozygous). We had a prior knowledge about the pathology: in particular we are interested in large stretches of homozygous SNV calls (Genome-wide linkage analysis performed under the assumption of recessive inheritance identified a common homozygous haplotype for this condition.). Accordingly, we found 6 stretches of homozygosity, two of which on the same chromosomes (6 and 16) described in a previous study. We think that this workflow could be easily automatized and used for different type of re-sequencing projects and that would lead to a strong interaction between clinical and molecular data, which is the purpose of the translational genomics.

### References

1. Li and Durbin (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26 (5), 589-95
2. Seattle SNP Annotation tools: <http://snp.gs.washington.edu/SeattleSegAnnotation134/>
3. Seelow et al. (2009) HomozygosityMapper an interactive approach to homozygosity mapping. *Nucleic Acids Res.* 37 (suppl 2), W593-W599

## A software pipeline for the discovery of variations in exome sequencing projects

E. Mattei, P.F. Gherardini, G. Ausiello, M. Helmer-Citterich 

Department of Biology, University of Tor Vergata, Roma, Italy

### Motivations

The recent advances in the technologies and strategies for DNA sequencing have dramatically facilitated the identification of novel human genes associated with rare and common diseases [1]. However novel methods are needed to identify high-quality variations among all the ones identified in a single experiment. The most successful approach to identify disease-causing mutations consists in using exome arrays [2] that allow the sequencing of only the coding regions in a genome. We developed a novel pipeline to identify high quality variations in the data produced by an exome sequencing experiment using the new 454 Roche sequencer [3].

### Methods

The input data of our pipeline are the sequencer reads mapped on the reference genome and the variations already identified by the sequencer software along with their confidence score. The first step of the procedure consists in associating the confidence score to each variant nucleotide and then filtering out variations with a low score. Variations in the length of the reads lead to misalignments between the reads and the reference genome. The second step of the pipeline produces a more accurate local alignment that can be searched for variations. Since the aim of the procedure is to validate the original variations produced by the sequencer, and not to identify new ones, original variations are compared with the newly identified ones for confirmation. As expected the sequencer alignment error rate increases as the length of the reads decreases, causing nucleotide mismatching. Reads with a perfect alignment with the reference, are marked as FULL by the sequencer. However the sequencer software also reports variations identified on chimeric reads, i.e. reads for which different portions align to separate regions of the genome. The pipeline reports, for each variation, how many FULL reads support the variation and how many do not. Information about CHIMERIC reads are included as well and

variations supported only by CHIMERIC reads are more likely to be incorrect. The next step is to flag single nucleotide polymorphisms as missense or synonymous and to use dbSNP [4] to discard the ones which are already known, and therefore unlikely to be associated with a disease. Subsequently, the propensity of each missense mutation to be deleterious for the function of a protein as opposed to neutral is calculated using CONDEL [5], a software that computes a weighted average of the scores of the SIFT [6] and POLYPHEN [7] methods. Moreover variations in dbSNP which are known to be associated with a disease are flagged. As a last step we prioritize mutations occurring in genes belonging to the same family of other genes known to be implicated in the pathology, if any. When SNP array [8] data of a patient genome are also available the procedure includes an additional test to identify which variations are more likely to be in a heterozygous site.

### Results

We tested our pipeline on the sequenced exomes of two patients suffering from Noonan Syndrome [9] and having no mutations in any of the genes already known to be implicated in the disease. SNP array data was also available for one of the patients. The original number of variations identified by the Roche software was about 105,000. After filtering the original set using the Phred score, the number decreased to about 102,000. Using a statistical test based on comparing the sequencing and SNP array results we reduced the number to 22,000. The removal of known SNPs further reduced the number of newly identified variations to 15,000, only 1,400 of which were missense. CONDEL predicted 800 of them as deleterious and only 60 were found in genes likely implicated in the disease. Our filtering pipeline therefore reduced the initial number of variations by four orders of magnitude, resulting in a very limited number of variations that can be tested in follow-up experiments.

## References

1. Roukos, D.H. (2010). Next-generation sequencing and epigenome technologies: potential medical applications. *Expert Rev. Med. Devices* 7, 723-726.
2. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. (2007). Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39, 1522-1527.
3. Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135-1145.
4. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308-311.
5. Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, *Condel*. *Am. J. Hum. Genet.* 88, 440-449.
6. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812-3814.
7. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894-3900.
8. Mei, R., Galipeau, P.C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R.K., Chee, M.S., Reid, B.J., and Lockhart, D.J. (2000). Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* 10, 1126-1137.
9. Allanson, J.E., and Roberts, A.E. (1993). Noonan Syndrome. In *GeneReviews*, R.A. Pagon, T.D. Bird, C.R. Dolan, and K. Stephens, eds. (Seattle, WA).

## De novo detection of A-to-I RNA editing sites in human mRNAs by massive transcriptome sequencing

E. Picardi<sup>1,2</sup>, A. Gallo<sup>3</sup>, S. Raho<sup>3</sup>, F. Galeano<sup>3</sup>, G. Pesole<sup>1</sup>

<sup>1</sup>Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università di Bari A.Morô Bari, Italy

<sup>2</sup>Istituto Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche, Bari, Italy

<sup>3</sup>Laboratorio di ricerca RNA editing, Ospedale Pediatrico Bambino Gesù, IRCCS, Rome, Italy

### Motivations

RNA editing is a widespread molecular phenomenon which modifies primary transcripts at specific positions [1]. It occurs in a variety of organisms including human and cooperates with alternative splicing in increasing both proteomic and transcriptomic complexity. RNA Editing can modulate gene expression and affect protein functionality. In human, such phenomenon is highly frequent in brain and its deregulation has been linked to a variety of neurological and neurodegenerative diseases [2]. Many editing events have been identified by next generation sequencing technologies employing massive transcriptome sequencing [3] together with whole genome or exome sequencing. Nowadays numerous RNA-Seq experiments are available through public databases and represent a relevant source of yet unexplored RNA editing sites. Hereafter we propose a simple computational strategy to identify de novo genomic positions enriched in novel potential RNA editing events through a new, two-step mapping procedure.

### Methods

To accurately predict RNA editing sites we developed a double mapping procedure in which millions of Illumina short reads were independently mapped onto the human transcriptome and the reference human genome tolerating at maximum two mismatches for each unique alignment. Only concordant alignments from the double mapping procedure were used for downstream analyses. All reads supporting each reference position were explored to calculate the empirical probability to observe a substitution. Such probabilities were then used to detect statistically significant base conversions by applying the Fisher's exact test by comparing the observed and expected occurrences in the aligned reads. Benjamini-Hochberg correction was finally employed to reduce the false discovery rate. A-to-I RNA editing candidates may be then selected according to P-value, coverage

and editing extent for the experimental validation using the classical Sanger sequencing from RNA/DNA extracted from the same individual.

### Results

We initially tested our computational method on the SRA study SRA012427 involving high-throughput transcriptome sequencing of human brain tissues by Illumina technology. Over 22 millions 50 nt long paired-end reads were aligned onto the human reference genome (assembly hg18). Applying the above-described double mapping methodology and stringent filters we found 19 highly significant A-to-I conversions in known human coding regions. Interestingly, 11 of such changes have been already described in literature and 6 were experimentally confirmed. To further corroborate our strategy we carried out an RNA-Seq experiment on total RNA extracted from human spinal cord. More than 20 million of directional paired-end reads were analysed using the above mentioned procedure. At 0.05 significance level (FDR corrected) we obtained 15 RNA editing candidates covered by at least 30 independent reads and showing only A-to-G changes. In this case potential editing candidates were confirmed by whole exome sequencing performed on the individual and tissue in order to optimally exclude SNPs and somatic mutations. Notably, we were able to confirm 12 predicted candidates. Our results, therefore, indicate the feasibility and effectiveness of the above-described strategy to detect de novo A-to-I RNA editing events in human.

### References

1. Gott, J.M. and Emeson, R. (2000) Functions and mechanisms of RNA editing. *Annual review of genetics*, 34, 499-531.
2. Maas, S., Kawahara, Y., Tamburro, K.M. and Nishikura, K. (2006) A-to-I RNA editing and human disease. *RNA biology*, 1-9.
3. Picardi, E., Horner, D.S., Chiara, M., Schiavon, R., Valle, G. and Pesole, G. (2010) Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. *Nucleic acids research*, 38, 4755-4767.

## AnnotateGenomicRegions: a Web application

L. Zammataro, G. Bucci, H. Muller✉

Computational Research, Center for Genomic Science of IIT@SEMM c/o IFOM-IEO-CAMPUS (Italian Institute of Technology), Milano, Italy

### Motivations

Next-generation sequencing (NGS) is producing large data volumes at reasonable cost and new applications are being developed at increasing speed. A common denominator for all applications of NGS technology is the need to annotate genomic regions of interest. Tools such as Galaxy [1], CisGenome [2], or the Bioconductor ChIPpeakAnno package [3] have been published to perform this task. However, using these tools often requires a significant amount of bioinformatics skills and/or downloading and installing dedicated software. A widely accepted, web-based annotation tool available to bioinformaticians and biologists with widely varying skill levels is not available. Indeed, many skilled bioinformaticians rely on self-made scripts to process the data to be annotated in the desired input/output format and in the necessary detail. For many biologists working with new generation sequencing data, annotating a set of genomic regions represents a complicated task that necessarily involves the help of a skilled bioinformatician.

### Methods

Here we present AnnotateGenomicRegions, a web application that accepts genomic regions as input and outputs overlapping and/or neighboring genome annotations chosen on a simple web-form. The application is based on Java Enterprise technology and runs on a Glassfish server. The necessary speed of annotating hundreds of thousands of genomic regions with tens of different annotations within seconds is achieved using a proprietary hash-based data structure.

### Results

We developed an annotation tool that fulfills five basic design criteria:

1. genomic regions shall be used as input query;
2. the output shall be pastable into an Excel table;
3. the application shall be web-based;
4. no programming skills required to use the application;

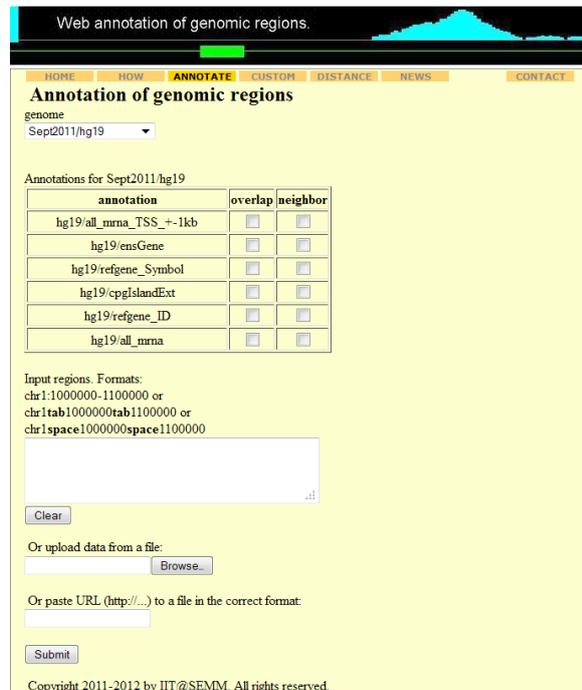


Figure 1.

5. it must be fast enough to annotate hundreds of thousands of genomic regions within seconds.

The tool can be installed on any computer capable of running Java and Glassfish on a Windows or Unix/Linux operating system, which is from a laptop to a mainframe computer.

### Availability

<http://sourceforge.net/projects/annotatelocus/?source=directory>

### References

1. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005;15:1451-5.
2. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol.* 2008;26:1293-300.
3. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics.* 2010;11:237.

## Molecular dynamics simulations reveal the role of five BCR-ABL kinase domain critical residues in TKIS binding

P. Buffa<sup>1</sup>✉, C. Romano<sup>2</sup>, A. Pandini<sup>1</sup>, P. Vigneri<sup>2</sup>, F. Fraternali<sup>1</sup>

<sup>1</sup>Randall Division of Cell and Molecular Biophysics, King's College of London, United Kingdom

<sup>2</sup>Clinical and Molecular Biomedicine Department, University of Catania, Italy

### Motivations

Chronic Myeloid Leukemia (CML) is a myeloproliferative disorder characterized by a well known molecular hallmark: the BCR-ABL chimeric oncoprotein. BCR-ABL displays constitutive tyrosine kinase activity, which favours the expansion of the leukemic clone by increasing proliferation and reducing cell death. Suppression of BCR-ABL catalytic activity by Tyrosine Kinase Inhibitors (TKIs) has dramatically improved the natural history of the disease achieving unprecedented results. However, an increasing number of point mutations inside the BCR-ABL kinase domain has been associated with different degrees of resistance. Insights into the critical residues involved in the interaction between the BCR-ABL tyrosine kinase domain (TKD) and different TKIs could explain the mechanisms allowing BCR-ABL mutants to avoid kinase inhibition.

### Methods

We have determined that five amino acids (E286, T315, M318, I360, D381) are critical for Imatinib (IM) binding. Two of them (T315 and M318) are also necessary for the binding of the second generation (2G) TKI Dasatinib (DAS), while, E286, M318, I360 and D381 are required for the interaction with the third generation (3G) inhibitor Ponatinib (PON). Only one (T315) of these five positions shows amino acidic substitutions in TKI-resistant patients. We used Modeller-v9.8 to generate *in silico* mutants that were predicted to maintain catalytic activity. We subsequently performed Molecular Dynamics (MD) simulations on these mutants in a simulated aqueous environment using the GROMACS package with AMBER force

field (ffamber99sb), obtaining 50ns of data collection for each system. We also carried out MD simulations of these mutants in complex with IM, DAS, and PON. The simulation trajectories were analyzed with both GROMACS analysis packages and Principal Component Analysis (PCA) in order to identify large-scale motions defining the functional dynamics of both wild-type and mutated BCR-ABL.

### Results

We demonstrated that the conservative I360T mutation displaces the helix- $\alpha$ C located in the N-lobe of the BCR-ABL tyrosine kinase domain, moving away E286 from the catalytic pocket. This residue is one of the critical amino acids for IM binding. When we docked IM inside the pocket of the BCR-ABL TKD displaying the helix displaced and performed further 50ns of MD simulation we found that the C-helix moves back towards the catalytic pocket, possibly under the attractive electrostatic field generated by the drug. We also found that IM restores three of the five original h-bond interactions warranting IM binding. This result is in agreement with experimental data showing that BCR-ABL I360T is inhibited by IM. Our data also confirm the efficacy of the 3G inhibitor PON on CML patients failing IM and 2G TKIs because of its ability to maintain the same interactions showed when the drug inhibits BCR-ABL wild-type. These findings could potentially be extended to further protein kinases, thus contributing to the design of other TKIs targeting different protein kinases contributing to the pathogenesis of additional neoplastic diseases.

## Identification and analysis of conserved pockets on protein surfaces

M. Cammisa<sup>1</sup>✉, A. Correr<sup>1</sup>, T. Fioriello<sup>1</sup>, G. Andreotti<sup>1</sup>, M.V. Cubellis<sup>2</sup>

<sup>1</sup>Istituto di Chimica Biomolecolare - CNR, Pozzuoli, Italy

<sup>2</sup>Istituto di Biostrutture e Bioimmagini - CNR, Napoli, Italy

### Motivations

The interaction between proteins and ligands occurs at pockets which are often lined by conserved aminoacids. In order to make the research of new drugs as economic as possible, it is necessary to exploit "in silico" techniques, high throughput and fragment-based screenings, which require the identification of pockets on the surface of proteins, active sites or not, which might be the targets of low molecular weight drugs.

### Methods

We developed a tool to evaluate the conservation of each pocket detected on the protein surface by CastP. This tool was named DrosteP because it recursively searches for optimal input sequences to be used to calculate conservation. DrosteP uses a descriptor of statistical significance, Poisson p-value, as a target to optimize the choice of input sequences. To benchmark DrosteP we used monomeric or homodimer human proteins with known 3D-structure whose active site had been annotated in UNIPROT. DrosteP is able to detect the active site with high accu-

racy because it nearly always coincides with the most conserved pocket. We extended our analysis to all the pockets found on the surface of human proteins.

### Results

Several methods for predicting ligand binding sites on protein surfaces have been proposed which combine 3D-structure and evolutionary sequence conservation, but any method relying on conservation depends critically on the choice of the input sequences. DrosteP chooses how deeply distant homologs must be collected to evaluate conservation and thus optimize the identification of active site pockets. Moreover it recognizes conserved pockets other than those coinciding with the sites annotated in UNIPROT which might represent useful druggable sites. Amino acid composition of conserved pockets differs significantly from that of non conserved pockets. This finding provides useful hints on the fundamental principles underlying protein-ligand interaction.

### Availability

<http://www.sbcentrostorico.unina.it/cammisa>

## A structure based pattern recognition on antibodies

D. Corrada , G. Morra, G. Colombo

Institute of Chemistry of Molecular Recognition, National Research Council, Milano, Italy

### Motivations

Protein-protein interactions are deeply involved in the antibody::antigen recognition process. Crystallographic data offer evidence of conformational changes between apo and holo forms of antibodies complexed with antigens. Nevertheless, the dynamical aspects of intermolecular relationships still remain a challenging issue. Extensive molecular dynamics (MD) simulations offer the suitable tool for generating statistical ensembles of conformations from which various energetic, structural and dynamic properties can be collected. The interaction energy correlations between all residue pairs can be investigated in order to find relevant regions involved in the fold stability; furthermore, a global overview of these sites can highlight preferential signaling pathways along the protein structures. In the present work, we will describe those conformational change events which derive from the formation of antibody::antigen complex. In particular, we will define those pathways that start from the paratope region and propagate through the immunoglobulin domains.

### Methods

We have taken into account a dataset of 24 Fab::antigen complexes, whose structures have been deposited at the Protein Data Bank. For most of them, 50ns unrestrained MD simulations were computed, with a 2fs time-step. Three structures targeting the same epitope and sharing similar paratopes were submitted to 200ns MD simulations. Each case of the dataset was duplicated, considering two systems: the isolated Fab structure (apo form) and the complex (holo form). An amount of 3,300ns of MD simulations was performed. The time-curves of RMSD on backbone atoms were used to monitor system equilibration. For every simulation the first 10ns were discarded for the further analyses. The structures sampled from the trajectories were clustered with the purpose of defining the conformational space explored. From the most representative structures (medoids) of the most populated clusters energy interaction matrices were calculated, consider-

ing all the non-bonded interactions between each residue pairs. Principal component analyses were performed over such matrices (Energy Decomposition Analysis), and only the relevant eigenvectors were considered. Since every component renders the relative contribution of each residue to the overall stabilization energy, we selected those components whose values are higher than a threshold value which depends on the number of the residues in the protein. All the residues were mapped to a common reference, obtained from joining the Chothia numbering schema with structural multiple alignment based schema, for variable and constant domains respectively. Finally, we mapped the energetic relevant residues of each case to the numbering schema depicted above, in order to summarize a common residue pattern referred to a generic Fab structure. We then collected the occurrences for each position, and we termed them as Interaction Energy Recurrent Positions (IERPs).

### Results

The results obtained herein are intended at identifying those interactions that define the formation of antibody::antigen complexes. The correlated motions are also investigated, reducing the complex protein dynamics to its essential degrees of freedom. The analyses of global distance fluctuation matrices show that the motions of residues belonging to the same domain appear more correlated with each other. The constant domains of the heavy chain (CH) show a higher mobility, with respect to the remaining protein. The CH domains are close to the boundary between Fab and Fc fragments; hence, CH domains may benefit from a greater degree of unrestrained motion. From the interaction energy analyses, we collected three subset of IERPs: the first ones are extracted only from apo forms simulations, the second ones are typical of holo forms, the third ones appear as relevant in both forms. The distribution of these subsets along the polypeptide chains of the Fab structures indicates that heavy chains contain most of the shared IERPs; in contrast, it seem that apo and holo forms show

differential IERPs profiles along the sequence of light chains. Then, we compared the distribution of IERPs, typical of apo and holo forms, along the four Ig domains VH, VL, CH and CL. In holo forms CL domains show lesser amount of IERPs, while CH domains appear enriched; variable domains (VH and VL) do not show significant differences between apo and holo. In holo form, CL domains still maintain few sparse IERPs that could be considered as local stabilizing hotspots. The

visual inspection of the structures of the dataset illuminates the spatial distribution of IERPs. In the antigen bound Fabs (holo forms) the selected residues starts from the framework regions of variable domains and propagate downstream, preferentially along the CH domain. Finally, the signal arrive at the base of constant domains, focusing around the C-terminal regions. On the opposite side unbound Fabs (apo forms) do not show a preferential pattern.

## Structural studies on the CXCR7 decoy receptor

T. De Vero<sup>1</sup>, S. Costantini<sup>2</sup>✉, G. Colonna<sup>3</sup>

<sup>1</sup>Dipartimento di Farmacologia Sperimentale, Università degli Studi di Napoli "Federico II", Italy

<sup>2</sup>INT Pascale - Centro Ricerche Oncologiche Mercogliano, Mercogliano, Italy

<sup>3</sup>Dipartimento di Biochimica e Biofisica & Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, Italy

### Motivations

Recent studies have revealed that the chemokine receptor CXCR7 plays an important role in cancer development. However, little is known about the effect of CXCR7 on the process of hepatocellular carcinoma (HCC) cell invasion and angiogenesis. In particular, CXCR7 was overexpressed in HCC tissues. The CXCR7-receptor is a trans-membrane protein that binds two chemokines, CXCL11 and CXCL12. It has the same basic structure of the other chemokine receptors composed by 7 helices and some Cys residues within the primary sequence to stabilize the tertiary structure through intermolecular disulfide bonds. The signal transduction of chemokines occurs, usually via the G proteins whereas CXCR7 is a "decoy receptor" and behaves as the "receptor Duffy". In particular, it does not bind to G protein, and then does not activate its metabolic pathway. When CXCR7 binds to CXCL12, it can form homodimers or heterodimers from the moment that tends to bind also to CXCR4. In the case of heterodimer (CXCL12/CXCR4/CXCR7), CXCR7 changes the conformation of the CXCR4/G-protein complex and repelling signaling. The activation of CXCR4 by CXCL12 leads to the activation of signaling through PI3K/AKT, IP3, and MAPK pathways, that promote cell survival, proliferation and chemotaxis. In addition, the pathway of the  $\beta$ -arrestin may be activated by GRK that leads to the internalisation of CXCR4. When CXCR7 alloy CXCL12, the mobilization of intracellular  $\text{Ca}^{2+}$  + classical does not occur, and the activation of  $\beta$ -arrestin may lead to the scavenging CXCL12. In tumor cells CXCR7 can also activate the signaling via PLC / MAPK that increases the survival of the cells. Therefore, to study the structure-function relationships in CXCR7, we have modeled its three dimensional structure as well as studied the energetic stability of the protein by molecular dynamics simulations. Finally, its complex with CXCL11 and CXCL12 has been simulated by docking methods.

### Methods

The three-dimensional model of human CXCR7 was performed by a comparative modeling strategy using as template the structure of the human CXCR4, recently published, and that of bovine rhodopsin using MODELLER9v5 program. Models have been evaluated in terms of stereochemical, structural packing and energetic quality. PatchDock web server was used to model the complexes between CXCR7 and the two ligands, CXCL11 and CXCL12, using for CXCR7 our model obtained by comparative modeling, for CXCL11 and CXCL12 their experimental structures. The complexes were analyzed by "Cocomaps Server" to identify the amino acids at the interface and to evaluate their solvent accessibility. The presence of putative H-bonds was calculated with Hbplus program.

### Results

Our studies show that CXCR7 model presents seven trans-membrane helices, a long N-terminal segment and three loops (i.e. loops 1, 2 and 3) in the extracellular region and, moreover, other three loops and a C-terminal loop in the cytoplasmic region can also be observed. This model had a Prosa Z-score of -2.19 and 86.54% of residues in most favored regions. In particular, the extracellular loop 2 and the cytoplasmatic C-terminal region comprise two short  $\beta$ -strands and a short helix, respectively, in agreement with the experimental structure of CXCR4. A comparison between the secondary structures of CXCR7 and CXCR4 models shows seven well conserved trans-membrane helices and only few changes in helix length. Furthermore, we have modeled two complexes between CXCR7 and two chemokines, CXCL11 and CXCL12. For these complexes, we evaluated the interaction residues, the related contact maps and the number of inter-chain H-bonds. In general, the analysis of the two complexes shows that the interacting regions are located in the N-terminal region as well as in the loops 2 and 3 of CXCR7. We have

also analyzed the physico-chemical properties of interaction residues of CXCR7 in the two complexes: i) some positively and negatively charged residues are present in the N-terminal while in the CXCR7/CXCL11 complex we also found also an aromatic residue; ii) the loop 2 presents a negatively charged residue (Glu) and an aromatic residue (Tyr); iii) the loop 3 shows a single aromatic residue (Phe). These observations suggest that the predominant interactions found in the models between CXCR7 and the two chemokines

are on hydrophobic and electrostatic basis. In particular, CXCR7 presents the highest affinity for CXCL12 in terms of binding energy, and of number of H-bonds, of charged and hydrophobic interaction residues, and of salt bridges. Moreover, the presence of stacking interactions is also in good agreement with experimental studies which indicated that the affinity of CXCL12 for CXCR7 is approximately ten times higher than the affinity of CXCL12 for CXCR4.

## Modeling and protein function prediction of truncated form of *geobacillus thermocatenulatus* lipase (BTL2)

M.A. Ghafouri<sup>1</sup>✉, A.A. Karkhane<sup>2</sup>, M.R. Azimi<sup>1</sup>, B. Yakhchali<sup>2</sup>, N. Goudarzi<sup>2</sup>

<sup>1</sup>Department of Plant Breeding, University of Zanjan, Zanjan, Iran

<sup>2</sup>Department of Industry and Environment Biotechnology, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran

### Motivations

Lipases, mainly of microbial origin, represent the most widely used class of enzymes in biotechnological applications and organic chemistry. Bacterial lipases are members of the structural superfamily of  $\alpha/\beta$  hydrolases that catalyze the hydrolysis and synthesis of a variety of acylglycerols at the lipid-water interface. Optimization of features and catalytic activity of lipases as the third industrial enzyme are significant for industrial applications. The structural comparison between  $\alpha/\beta$  hydrolases and thermoalkalophilic lipases represented that the insertion of Zn<sup>2+</sup> binding site ( $\alpha$ 3, b1 and b2) has not, until now, been seen within the  $\alpha/\beta$  hydrolase canonical fold that making a tight intramolecular interaction with the main catalytic domain. In this study, we investigated the role of Zn<sup>2+</sup> binding site on lipase activity by using the computational methods.

### Methods

To survey the effect of Zn<sup>2+</sup> binding site on structure and function of lipase, the nucleotide sequence of  $\alpha$ 3 domain was deleted in the *Geobacillus thermocatenulatus* lipase gene. Homology modeling for the native and the mutated lipases was performed using MODELLER v9.10, based on crystal structure of *Geobacillus thermocatenulatus* (PDB: 2w22 (opened form)) as template. Analysis of the 3D-structure of the native and the mutated lipases generated by MODELLER were done using QMEAN and ProSA servers. The molecular docking of native and mutated lipases

with ligands (Dibutyryl, tributyrin, tricaproin, and tricaprylin) was performed using the trial Molegro Virtual Docker 5.0. Before docking, the structure of receptor and ligands was prepared, flexible torsions in ligands detected, explicit hydrogens created and possible missing bonds assigned.

### Results

Z-scores of the two lipases were calculated from PROSA-web. It was found to be -8.32 and -8 for the native and mutated lipases respectively. This result confirmed that there is high similarity between native and mutated lipases in opened conformation. Also the QMEAN Server was used for model quality estimation of the lipases. It was found to be 0.833 and 0.749 for the native and chimera lipases respectively. Results obtained by QMEAN Server specify the reliability of the models and also, confirmed that there is high similarity between two lipases. Furthermore, the root mean-squared deviation (RMSD) value of the native and mutated lipases was calculated by superimposing two lipases (388 residues). The RMSD was found to be 12.62Å° for the opened form of the lipases. We carried out molecular docking with a series of substrates. In comparison of mutated lipase interaction energy with BTL2 lipase, the mutated lipase showed lower energy to break down bonds between ligand and lipase. The function prediction of mutated lipase with deletion on  $\alpha$ 3 domain, represent lipase activity enhancement with C4 to C10 substrates. Therefore function of mutated lipase in hydrolyzing of triacylglycerol is better than BTL2.

## Analysis of molecular recognition features in membrane proteins

I. Kotta-Loizou, G.N. Tsaousis , S.J. Hamodrakas

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens, Greece

### Motivations

During the past few years there has been a growing interest in the field of intrinsically disordered proteins-related research. Intrinsically disordered proteins (IDPs) possess no rigid 3D structure under physiological conditions, yet they are functionally active. A protein may be fully or partly disordered, containing long or short intrinsically disordered regions (IDRs). Molecular Recognition Features (MoRFs), known also as Molecular Recognition Elements (MoREs), are defined as short regions that undergo disorder-to-order transition upon binding to their partners. As their name suggests, they are considered to be implicated in molecular recognition, which serves as the initial step for protein-protein interactions. Membrane proteins comprise approximately 30% of fully sequenced proteomes and they are responsible for a wide variety of cellular functions, including cell signalling. The aim of the current study was to identify and analyse MoRFs in membrane proteins.

### Methods

A dataset of putative MoRFs was constructed from the Protein Data Bank, selecting membrane protein fragments between 10 and 70 residues, which interact with proteins longer than 100 residues. The assumption was made that such short aminoacid sequences would be less likely to form a rigid 3D structure prior to interaction. The initial dataset was further filtered for ambiguous information and a non-redundant dataset was created applying length-dependent thresholds. Subsequently, sequence, structural and functional analysis of the membrane MoRF non-redundant dataset was conducted.

### Results

Initially, we sought to characterize our dataset and assess membrane MoRF associations with intrinsic disorder. Approximately half of the membrane MoRFs are short, between 10 and 20 residues. In addition, membrane MoRFs' aminoacid composition was found to differ not only from that of a typical IDR, as expected, but from globular MoRFs' aminoacid preferences as well. A structure-based criterion, evaluating per-residue surface and interface areas, supported the idea that membrane MoRFs are intrinsically disordered when isolated, and undergo a disorder-to-order transition upon binding. Missing density residues, often associated with disorder, were also assessed and were found to comprise, along with irregular residues, more than half of membrane MoRFs' secondary structure. Membrane MoRFs were categorized as alpha, beta, irregular and complex, depending on their secondary structure after the interaction with their partners. Further studies were focused on MoRF-containing proteins and revealed that the vast majority are eukaryotic single-spanning transmembrane proteins. Moreover, the position of MoRFs in relation to the protein's topology was determined. Finally, functional analyses of MoRF-containing proteins and MoRF-binding partners are currently under way. In conclusion, our goal is to provide insight into potential disorder-based protein-protein interactions involving membrane proteins. A comparison between membrane and globular MoRFs will allow us to determine if they are likewise implicated in molecular recognition procedures and whether a similar mechanism is involved. In the long term, the above information will facilitate identification, and possibly prediction, of membrane MoRFs.

## Assessment of structure-based functional annotation methods on protein 3D models

I. Mangone, M. Helmer-Citterich✉, G. Ausiello

Department of Biology, University of Tor Vergata, Roma, Italy

### Motivations

The three-dimensional structure is more informative of the sole aminoacidic sequence to assign a molecular function to a new protein. For this reason many automated methods have been developed to infer the function of a protein structure using comparison approaches or analyzing its physicochemical characteristics. Unluckily while the genome sequencing projects of organisms have considerably increased the number of available protein sequences, protein structure determination with X-ray crystallography and NMR is still a complex and rather costly procedure. There are indeed more than 20 thousand entries in the database of protein sequences (UniProt) and only 79,600 entries in the database of protein structures (PDB). The big gap separating the number of known sequences from the number of solved structures is increasing every year and is strengthening the need for structure-based functional annotation methods capable to work on homology models instead of crystal structures. The applicability of the existing functional prediction methods to protein models has never been explored so far, even if most of the structural information now available is stored in 3D models. The aim of this work is to study the reliability of different structure-based functional annotation methods when used on protein models and to analyze how the prediction methods performance is correlated with the overall quality of the available homology model.

### Methods

We used an automated procedure to compare the performances of many structure-based functional prediction methods when they work on a set of homology models of different quality or on a crystallographic solved structure. Each different method is tested on the same dataset proposed by the authors in the original publication and on a set of homology models built for each structure in the dataset. All models were generated using MODELLER (v9.9) and evaluated using the GDT<sub>TS</sub> score. To obtain models of different quality only templates are used having a sequence similarity with the solved structures under a set of fixed thresholds.

### Results

We have evaluated five methods: PDBinder, Concavity and Fpocket for the prediction of protein binding pockets and Pfinder and FINDSITE<sub>metal</sub> for the prediction of phosphate and metals binding sites. The performances have been measured using the F-score or the MCC where applicable. Preliminary results show that when using models with a GDT higher than 99% on average the performances drop by about the 22%. When models quality decreases we have a significant decrease of prediction method performances up to 50% (with a GDT of 50%), with some methods that have shown a greater resilience to the decrease of the model quality. These are only to be considered as preliminary results since a number of other methods are being added to the analysis.

## Application of computational methods for structural and functional characterization of mutants of GALT enzyme

A. Marabotti<sup>1</sup>, A. Facchiano<sup>2✉</sup>, L. Milanesi<sup>1</sup>, M. Tang<sup>3</sup>, K. Lai<sup>3</sup>

<sup>1</sup>Istituto di Tecnologie Biomediche, CNR, Segrate, Italy

<sup>2</sup>Istituto di Scienze dell'Alimentazione, CNR, Avellino, Italy

<sup>3</sup>Division of Medical Genetics, Department of Pediatrics, University of Utah, Salt Lake City, Utah, United States

### Motivations

Galactose-1-phosphate uridylyltransferase (GALT) catalyses the conversion of galactose-1-phosphate to UDP-galactose, a key step in the galactose metabolism. Deficiency of GALT activity in humans caused by mutations in the GALT gene are associated to a rare genetic disease called Classic Galactosemia. To date, more than 180 mutations are known on GALT gene, most of which are missense mutations. A previous study applied computational methods to predict the effect of known mutations [1] and results were stored in a publicly available database: <http://bioinformatica.isa.cnr.it/GALT> [2]. In this work, we have characterized the effects on structure and function of GALT enzyme of 14 novel missense mutations, applying not only static modelling, but also molecular dynamics simulations. Results were compared to the biochemical characterization of these mutants, expressed and purified from bacteria.

### Methods

Starting from the 3D model of human GALT enzyme [3] (the crystallographic structure of the human enzyme is not yet available), homodimeric mutants were created using Modeller 9v8 [4]. Each resulting mutant was analysed for variations in structural feature such as intersubunit interactions, secondary structures, solvent accessibility, H-bond and salt bridge patterns, and for predicted stability of the protein. Molecular dynamics simulations were applied to those selected mutants, for which the static modelling of mutations did not allow to highlight any variation in these properties, using GROMACS program [5]. Results were compared with those obtained on wild type molecule and on the most characterized mutant, Q188R. The human homodimeric mutant proteins were expressed in *E. coli*, purified and assayed for their activity and kinetic properties ( $V_{max}$  and  $K_M$ ).

### Results

Our static modelling simulations predicted for most of the mutant GALT enzyme alterations at structural level: in two cases the alteration of intersubunit interactions, in other two cases the involvement of residues indirectly affecting the active site, and in most cases the alteration of the network of H-bonds and salt bridges, with either local or global effects. In addition, in most cases these mutants have predicted stability problems. In two cases, however, static modelling was not able to predict any structural effect. In these cases, the results of molecular dynamics simulations were able to suggest possible effects that impair the correct activity and the stability of the enzyme, too. The comparison with simulations made on mutant Q188R allowed also to confirm the deleterious effects of this mutation at structural level. The biochemical characterization of these mutants showed results that are well explained by the simulations. In particular, those mutations that were predicted to have a large impact on protein structure are also those with minimal or no activity, whereas those mutants showing a residual activity, also show localized effects on protein structure.

### Acknowledgments

This work was supported by a Parents of Galactosemic Children Research Grant (to A.M.), and by project FIRB MIUR ITALBIONET (RBPR05ZK2Z and RBIN064YAT\_003) (to A.M., L.M). This work has been made in the frame of the Flagship Project InterOmics and of the project CNR-Bioinformatics. Research grant support to K.L. includes U.S. NIH grants 5R01 HD054744-04 and 3R01 HD054744-04S1.

### References

1. Facchiano A, Marabotti A. (2010) *Protein Eng Des Sel* 23, 103-113.
2. d'Acerno A, Facchiano AM, Marabotti A. (2009) *Genomics Proteomics Bioinformatics* 7, 71-76.

3. Marabotti A., Facchiano AM (2005) J Med Chem 48, 773-339.
4. Sali A, Blundell T L (1993) J Mol Biol 234, 779-815.
5. Hess B, Kutzner C, Van der Spoel D, Lindahl E (2008) J Chem Theory Comput 4, 435-447.

## Identification of molecular targets for mycotoxins related to autism development

A. Marabotti<sup>1</sup> ✉, M. Landini<sup>1</sup>, A. Mezzelani<sup>1</sup>, L. Milanesi<sup>1</sup>, M.E. Raggi<sup>2</sup>

<sup>1</sup>Istituto di Tecnologie Biomediche, CNR, Segrate, Italy

<sup>2</sup>IRCCS "E.Medea" Associazione "La Nostra Famiglia", Bosisio Parini, Italy

### Motivations

Autistic spectrum disorder (ASD) is an ensemble of developmental disorders impacting on the social relationships, the development of the language and of interpersonal communications and the behaviour of affected people. A number of factors have been supposed to cause ASD, including a direct genetic component (accounting for less than 10% of cases), epigenetics and environmental factors such as pollutants, heavy metals, viruses and vaccines. To date, however, none of these factors have been definitely associated to ASD. Further studies hypothesized the presence of "vulnerability genes" whose variants are not causative of autism per se, but can increase the susceptibility of the individuals towards several factors, ultimately leading to autism development. In people with ASD, gastrointestinal (GI) disorders are commonly reported. The "leaky gut hypothesis" theory speculates that impaired gut permeability permits the entry of molecules such as toxins in the bloodstream, both affecting directly the central nervous system (CNS), and causing sensitization of the gut mucosal immune system. Mycotoxins are food contaminants present ubiquitously and derived from the secondary metabolism of several moulds and fungi. The consumption of several foods, especially those derived from cereals and milk, exposes people to the intake of high level of mycotoxins. Children can be more exposed than adults to mycotoxins, for their particular diet, for a lower diet variability with respect to adults, and for the fact that a newborn assumes more food than an adult, in proportion to his/her body weight. Once introduced in the body, mycotoxins could promote sneaky negative effects in the GI tract and in other tissues, playing a role in the development of syndromes of unknown etiology. In our work, we used bioinformatics approaches to find potential targets for the binding of mycotoxins, in order to verify the existence of a possible role in ASD onset, and we tested a small sample of ASD

patients and controls to verify the alteration of genes identified as potential targets.

### Methods

The Web server TarFisDock was used to identify proteins able to bind four selected mycotoxins: ochratoxin, gliotoxin and fumonisins B1 and B2, with a reverse docking approach. The 3D structures of these molecules were obtained in .sdf format from PubChem. Both neutral and charged formes were taken into account. Targets were searched among the 3D structures of proteins deposited in PDTD database associated to TarFisDock server. Analyses of the involvement of these targets in different pathway networks were made using KEGG database. The further docking studies were made using AutoDock v. 1.5.2. and ADT. A sample of 52 ASD patients and 40 healthy parental and uncorrelated controls were analysed to search for the presence of mutations in selected genes among the targets identified by reverse docking approaches. For each patient, DNA was isolated from peripheral blood and NLGN3 and NLGN4X exons were amplified and sequenced by PCR and Sanger method, respectively.

### Results

The reverse docking approach identified several proteins that are involved in pathways related to CNS development and/or pathologies, as possible targets for the binding of each mycotoxin. Moreover, the relative abundance of proteins related to CNS pathways is increased with respect to the statistical distribution of targets among all the classes identified by the tool. Several proteins in common to all these mycotoxins were identified among the top 10% of all possible targets. In particular, the structural class of cholinesterase was found to be a common target for mycotoxins. This is very interesting, since neuroligins, proteins of the cholinesterase structural class, are directly associated in literature with autism. We performed traditional docking simulations to verify if mycotoxins can bind to these proteins. Results

confirmed the possibility of some mycotoxins to bind to neuroligins, in particular neuroligin 3 and 4, associated to chromosome X and therefore very interesting, given the ratio of 4:1 between males and females for ASD onset. Moreover, with docking studies, we were able to find that some mycotoxins, in neuroligin 4X, are predicted to interact directly with residues whose mutation is related to autism. This suggests that indeed the mutation of this protein can modulate its ability to interact with exogenous toxins, and therefore people carrying these mutations could be more susceptible to their negative effects. The genetic

analysis conducted on selected autistic patients and controls confirmed the presence of SNP and mutations on the exons of NLGN3 and NLGN4X. Further studies will be made to establish a statistical relationship among the presence of mutations and the distribution of mycotoxins into the bodily fluids of these people.

### Acknowledgements

Ricerca Finalizzata e Giovani Ricercatori 2009 (GR-2009-1570296), FIRB-MIUR Italbionet (RBPR05ZK2Z and RBIN064YAT\_003), Flagship Project InterOmics.

## Domain-context information for the improvement of phosphorylation site prediction

A. Palmeri✉, M. Helmer-Citterich, P.F. Gherardini

Centre of for Molecular Bioinformatics, Department of Biology, University of Tor Vergata, Roma, Italy

### Motivations

Our understanding of the determinants of protein phosphorylation is far from being complete. In many predictive systems, linear-motifs represent the main features that have a high power of discrimination between the phosphosites and the non-phosphorylated sites. The majority of the tools consider structural features like the secondary structure, the disorder or the solvent accessibility of the phosphopeptide. However the protein context in which the phosphopeptide is found is in many cases ignored. The aim of this work is to study the distribution of phosphorylation sites with respect to protein domains. Moreover we want to investigate whether including domain information improves the prediction of phosphorylation sites.

### Methods

We collected Human Phosphorylation data from the PhosphositePlus [1], phospho.ELM [2], PHOSIDA [3] and Swissprot databases and mapped these phosphosites on the human proteome downloaded from the Swissprot. We identified the domain boundaries on each protein, using the Pfam scanner [4]. We then performed a statistical test of significance for each domain type to identify the domain types that are enriched or depleted in phosphorylation. The significance test outputs the probability that the relative abundance of each domain type is different between the phosphoproteome and the overall proteome. As the majority of phosphosites are located outside protein domains, we performed the same analysis for Inter Domain Regions. We identify a IDR, as the protein region that is enclosed by two specific domains or by one specific domain and the C-terminal or N-terminal of a protein. A phosphosite predictor has been developed using human phosphorylation data. The training and testing procedures were written in R, using the package LiblineaR. The features that we used in the predictor are: the -5/+5 residues around the phosphosite, a Local Domain Feature and a Best Domain Feature. We encoded in the

predictor the sequence features in standard orthogonal encoding. The Local Domain Feature represents the propensity of a specific domain type to be phosphorylated. It is calculated from the training data as the proportion of phosphorylated domains of a specific domain type on all the occurrences of that domain type. The Best Domain Feature depends on the whole domain-composition of the protein and it is defined as the number of phosphoproteins in which the domain-type is found divided by the total of the proteins that contain that domain-type. When we predict a site in a protein the Best Domain Feature is represented by the maximum among all the propensities of the domains contained in the protein.

### Results

We performed a statistical test to identify domains that are significantly enriched or depleted in phosphorylation. Two tests were performed: one for the Tyr and the other for the Ser/Thr. For the Tyr phosphorylation, we obtained 26 domain types enriched and 19 depleted. For the Ser/Thr phosphorylation we found that there are 22 enriched and 26 depleted domains. The domain types that are enriched and depleted in phosphorylation represent almost the 0.5% of all the existing domain types. But the occurrences of the domains enriched in phosphorylation account for almost the 3% of all the domains in the Human Proteome. Moreover the occurrences of the depleted domains represent almost the 30% of the Human Proteome. However the majority of phosphosites are in protein regions outside domains. Thus, the same analysis was performed on Inter Domain Regions obtaining very similar results. Having observed that domain-context information influences phosphorylation, we wanted to test if it could be useful for phosphosite prediction. Therefore we trained two phosphosite predictors, one using only the sequence and the other with all the domain-contextual features, encoded in the Local Domain Feature and the Best Domain Feature. We compared the per-

formances between these two predictors, using the AUC, in a test dataset independent of the training. The sequence-only predictor obtained an AUC of 0.71 for Ser/Thr prediction and of 0.63 for Tyr prediction. The predictor with all the features reached an AUC of 0.78 for Ser/Thr prediction and of 0.72 for Tyr prediction. In both Ser/Thr and Tyr predictions we observed a 10% improvement in performance, due to the inclusion of the domain-context features.

## References

1. Dinkel et al., Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res.* 2011 Jan;39(Database issue):D261-7. Epub 2010 Nov 9.
2. Hornbeck et al., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research*, 2011, 1-10 doi:10.1093/nar/gkr1122
3. Gnad et al., *Nucleic Acids Res.* 2011 Jan;39(Database issue):D253-60. Epub 2010 Nov 16.
4. Finn et al., The Pfam protein families database. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D211-22. Epub 2009 Nov 17.

## Computer modeling of human delta opioid receptor

F. Sapundzhi<sup>1</sup>, T. Dzimbova<sup>2</sup>✉, N. Pencheva<sup>1</sup>, P. Milanov<sup>3</sup>

<sup>1</sup>South-West University "Neofit Rilski", Blagoevgrad, Bulgaria

<sup>2</sup>Institute of Molecular Biology, Bulgarian Academy of Sciences, Sofia, Bulgaria

<sup>3</sup>Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

### Motivations

The development of strong analgesics without potential for abuse and adverse side effects is connected to understanding of the differences in opioid receptor subtypes as well as the model of interaction of ligands with these receptors. In the absence of crystal structures of opioid receptors, 3D homology models with different templates have been reported in the literature.

### Methods

The aim of our study is to choose within recently published crystallographic structures templates for homology modeling of the human delta-opioid receptor. We generate several models using

different templates and all they were evaluated by docking procedure. Ligands used in this study were already synthesized by our group and their biological activity was evaluated. They are analogues of the endogenous opioid peptides - enkephalins with substitutions in second position.

### Results

The best model of the human delta-opioid receptor was chosen according to data obtained from docking and in vitro biological activity.

### Acknowledgments

This work was supported by NFSR of Bulgaria project DVU 01/197.

## Cooperative fluctuations of PTP1B by an elastic network model analysis

AC Serdaroglu<sup>✉</sup>, AN Ozer

Department of Bioengineering, Marmara University, Istanbul

### Motivations

Protein tyrosine phosphorylation is essential in controlling many vital activities of the cell such as growth, differentiation, metabolism and immune response. Abnormal tyrosine phosphorylation leads to various human diseases including cancers, diabetes, rheumatoid arthritis and hypertension. PTP1B, which is a major negative regulator of insulin signaling, is one of the important forms of tyrosine specific phosphatases that hydrolyze phosphotyrosine containing proteins. As the loss of PTP1B activity leads to enhanced insulin sensitivity and resistance to weight gain, inhibiting PTP1B activity represents a novel approach for the treatment of diabetes and obesity. In this regard, understanding the molecular recognition mechanism in binding processes of PTP1B may provide guidelines for the development of potent PTP1B inhibitors.

### Methods

Here, the structural dynamics of both substrate- and inhibitor-bound PTP1B are studied comparatively using the Anisotropic Network Model (ANM) which performs harmonic vibrational analysis around the equilibrium states and predicts the directionalities of the collective motions as well as their magnitudes.

### Results

The mean-square fluctuations in the most cooperative ANM modes are similar in different PTP1B complex structures and the minimum fluctuating residues correspond to the dynamically correlated hinge regions. Further, the variation in the orientation of the fluctuations is caused mainly by the residues lying along the rotational axes that are responsible for the functional motions. Overall, the elaborated analysis of the structural fluctuations of PTP1B in interaction with its ligands helps to gain insight into the dynamics of the phosphatase in relation to its function.

## Use of BioBlender for all atom morphing of protein structures

M.F. Zini, Y. Porozov, T. Loni, R. Andrei, M. Zoppè✉

Scientific Visualization Unit, Institute of Clinical Physiology - CNR, Pisa, Italy

### Motivations

The vast majority of proteins and other biological macromolecules act in life processes through some form of motion. While this concept has been recognized and is ever more considered in the structural biology field, it is still difficult to handle by the majority of experimental scientists. We reasoned that a simple system that enables biologists to elaborate protein motion would help experimental scientists to better understand the complex spacial behavior of proteins.

### Methods

Using Blender, a complete package for Computer Graphics, Gaming and Visual effects, we have developed BioBlender, which uses the game engine for interpolation between different conformations of a proteins. The conformations can be derived from studies of NMR, or can be calculated according to Normal Mode Analysis. The system can be downloaded as a stand alone

and we are also preparing a server for the elaboration of complex that contain large N of atoms (> 10.000). Elaboration of motion is performed using the Game Engine embedded in Blender, equipped with a set of rules that mimic the main features of chemical motion, i.e. that atoms can move by rotation of the previous bond (implemented by allowing rotation on torsion angles only), and that atoms cannot occupy the same space at the same time (collision detector).

### Results

The motion was recorded and exported in .pdb format as a series of conformations trajecting the molecule between the two given models. Evaluation using the GROMOS force filed in Swiss PDB viewer revealed a very good agreement with data derived experimentally, and /or calculated using classical Molecular Dynamics simulations.

### Availability

<http://www.bioblender.net/>

## Geena, a tool for MS spectra filtering, averaging and aligning

P. Romano<sup>1</sup>✉, A. Profumo<sup>2</sup>, R. Mangerini<sup>3</sup>, F. Ferri<sup>4</sup>, M. Rocco<sup>2</sup>, F. Boccardo<sup>3</sup>, A. Facchiano<sup>5</sup>

<sup>1</sup>Bioinformatics, IRCCS San Martino University Hospital - IST National Cancer Research Institute, Genova, Italy

<sup>2</sup>Biopolymers and Proteomics, IRCCS San Martino University Hospital - IST National Cancer Research Institute, Genova, Italy

<sup>3</sup>Medical Oncology B, IRCCS San Martino University Hospital - IST National Cancer Research Institute, Genova, Italy

<sup>4</sup>Department of Physics and Mathematics, Insubria University, Como, Italy

<sup>5</sup>Bioinformatics, Institute of Food Sciences, National Research Council, Avellino, Italy

### Motivations

Mass spectrometry (MS), one of the most recent high-throughput technologies, produces a high volume of data. Many tools exist for MS data management, but little is available for the automation of related procedures. Geena is a new tool that aims at automating some of the fundamental steps involved in the analysis of  $m/z$  and abundance data from MALDI/TOF MS experiments. Geena was developed by taking into account the following assumptions:

a. in each spectrum, molecules are present in the form of different isotopic abundances

that can be summed together to give a total abundance value;

b. often, experimental data have to be normalized against an internal standard in order to obtain (semi) quantitative results;

c. experimental data can be affected by background noise. The selection of signals above a modulated threshold built on the spectra profile may be useful;

d. the analysis of sample replicates yields multiple spectra which are different because of marginal errors/changes in the experimental phase only.

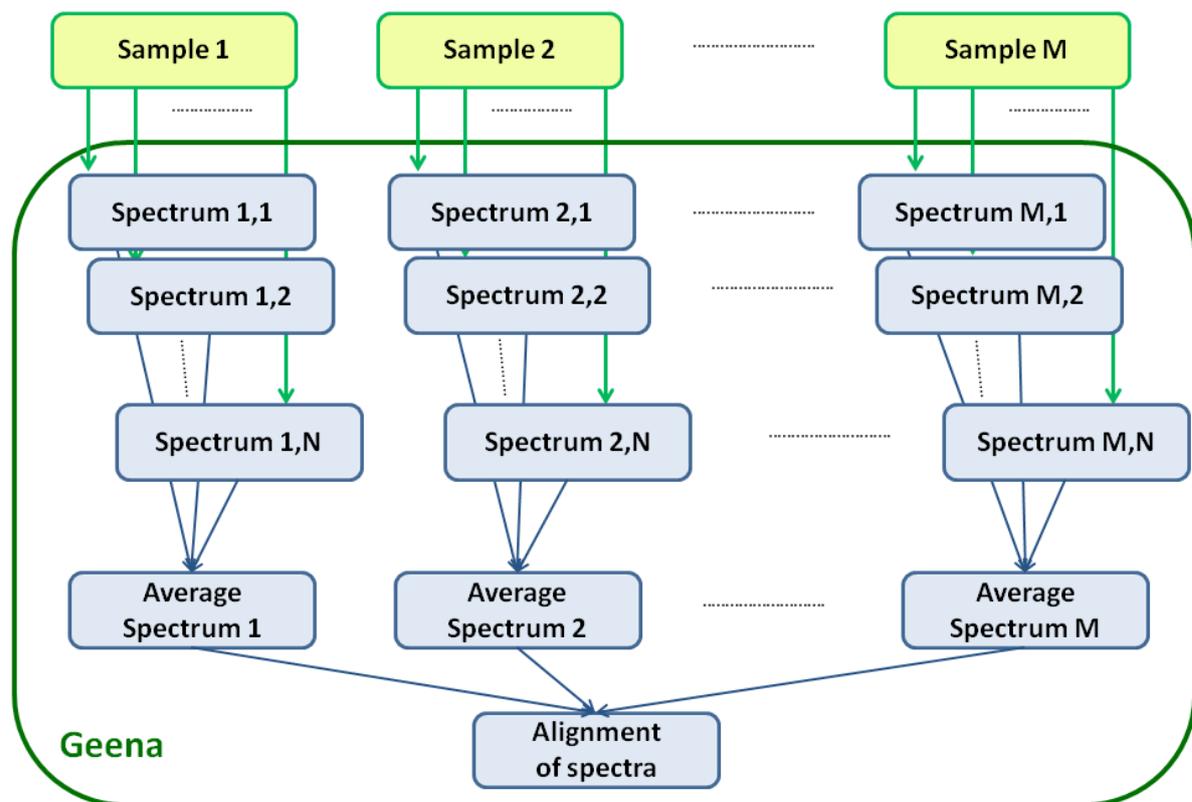


Figure 1. The figure reports an outline of the processing method in Geena.

An average spectrum representative of the sample may be defined by aligning these spectra along the  $m/z$  axis and computing mean intensity values from the corresponding abundances. In order to compare single or average spectra obtained from different samples the alignment along the  $m/z$  axis is required.

### Methods

Geena was written in PHP and partially, for spectra alignment, in perl. Both input and output are managed in simple text files, usually having tab or comma delimited values. Such formats can easily be consumed by MS Excel or any other data management system. Data may be stored on the server in a MySQL database. The processing method is mainly heuristic and it is based on original algorithms. It includes the following steps: a) preprocessing of spectra replicates, which consists in isotopic peaks joining, normalization, and peak selection, b) computing average spectra for replicated analysis of samples, and c) alignment of average spectra.

### Results

Geena is a public web server. The input consists in MALDI/TOF MS spectra. The data file is usually uploaded to the server and removed as soon as it has been used. It may include data from many samples, for each of which more spectra replicates can be provided. The output consists both in the averaged spectra from replicates and in the alignment of averaged spectra. The alignment is shown in the results page, while all results are available for downloading from the same page and sent by email, if a valid address is provided. Many parameters are defined. The analysis range is specified by indicating the lower and upper  $m/z$  values. The presence of a

normalization peak and its corresponding  $m/z$  value must be specified in order to normalize data. The threshold for peak selection is specified by providing abundance threshold values for the upper and lower limits of the analysis range. Further Intermediate thresholds may be specified, in which case a broken line is built by linear interpolation. An alternative method, that is based on data background estimation, is being added to Geena. Isotopic peaks of the same molecule are identified on the basis of the maximum allowed delta between them, i.e. the maximum deviation from expected values to consider a signal as an isotopic abundance of a given peak, and of the maximum number of isotopic replicates. In order to compute average spectra for a given sample, the maximum delta for aligning replicates, i.e. the maximum allowed deviation along the  $m/z$  axis between two signals belonging to replicates of the same sample to align them, and the minimum number of signals in replicates, that defines the minimum number of replicates that should contain a signal to include it in the average spectrum, can be specified. Similarly, the maximum delta for aligning average spectra and the minimum number of signals in average spectra can be specified to support alignment of average spectra. The method has been used for the analysis of long-term cryopreserved sera [1].

### Availability

<http://bioinformatics.istge.it/geena/>

### References

1. Mangerini R, Romano P et al. (2011) The application of atmospheric pressure MALDI to the analysis of long-term cryopreserved serum peptidome. *Analytical Biochemistry* 417, 174-181.

## Topological analysis of co-expression networks in neoplastic tissues

R. Anglani<sup>1</sup>✉, T.M. Creanza<sup>2</sup>, V.C. Liuzzi<sup>1</sup>, P.F. Stifanelli<sup>1</sup>, R. Maglietta<sup>1</sup>, A. Piepoli<sup>3</sup>, S. Mukherjee<sup>4</sup>, F.P. Schena<sup>2</sup>, N. Ancona<sup>1</sup>

<sup>1</sup>Bioinformatics and Systems Biology Lab, Institute of Intelligent Systems for Automation, CNR, Bari, Italy

<sup>2</sup>Department of Emergency and Organ Transplantation, DETO, University of Bari, Italy

<sup>3</sup>Unità Operativa di Gastroenterologia, IRCCS, Casa Sollievo della Sofferenza, San Giovanni Rotondo,

<sup>4</sup>Italy Institute for Genome Science and Policy, Duke University, Durham, United States

### Motivations

Gene expression data carry important information for the study of the complex response patterns of a biological system to cell state modifications. Consequently, gene co-expression networks are useful models to enlighten the coordinated expression of groups of genes that are functionally co-regulated in order to provide the adaptive response to the system modification. In this framework, topology-based approaches to network analysis have yielded unexpected insights of the global properties of biological systems that could not be unveiled with one-gene approaches. The key idea is that topological differences can critically emerge from the comparison between normal and cancer networks and

can identify those non-differentially expressed genes which are involved in the onset and progression of the specific disease.

### Results

In this work, we introduce a novel method for the characterization of disease genes, based on the study of topological differences of co-expression networks inferred from microarray expression profiling of neoplastic and normal tissues. Moreover, we assess the statistical significance associated with the variation of topological observables in the two phenotype conditions. The analysis, that has been focused on different human solid tumors, provides crucial evidences of common characteristics in the response patterns of gene co-expression networks in neoplastic tissues.

## mentha: the interactome browser

A. Calderone , G. Cesareni

Department of Biology, University of Rome "Tor Vergata", Via della Ricerca Scientifica, Rome, Italy

### Motivations

Protein interaction databases archive protein-protein interaction (PPI) information from published articles. However, no database alone has sufficient literature coverage to offer a complete resource to investigate "the interactome". We have developed mentha, a new resource that addresses this problem. mentha archives evidence about PPI in the human proteome – evidence collected from different PPI sources - and offers a series of tools to analyse these data. Users can search the database using a web search engine that returns interaction information for any protein of interest displaying it in the context of the "global interactome". A graphical application to represent this information helps scientists browse the collected data in order to ease and to inspire new experiments and outlooks.

### Methods

All the remote protein-protein interaction (PPI) databases that are relevant to this project, namely Mint, IntAct, BioGRID, DIP and MatrixDB, are members of the IMEx consortium. IMEx databases adopt a common format and use controlled vocabularies that facilitates data integration. In addition they implement PSICQUIC, a project whose aim is to standardise the access to molecular interaction databases. PSICQUIC defines a list of common fields for each database, and specifies a standard web service with a well-defined list of methods. All the PPI databases are queried using the PSICQUIC protocol. mentha is assembled by a merging procedure that runs weekly and that creates non-redundant data. The information archived in the queried PPI databases is manually curated and annotated

with controlled vocabularies. PPI evidence is represented with fixed sets of identifiers such as UniProt IDs and PMIDs. Most of the time proteins are represented with a UniProt identifiers. For any identifier different from UniProt, a sub-procedure gathers all the identifiers and tries to map them using a service offered by UniProt. The merging procedure behind mentha builds a non-redundant database that collects protein interactions from the aforementioned databases, together with their detailed annotation: interaction type, experimental methods and literature references.

### Results

This approach has generated a consistent interactome (graph) that can be used in various analyses as demonstrated by its use in other projects carried out in our group. Most importantly, the procedure assigns to each interaction a reliability score - the MINT score - that takes into account all the supporting evidence. The user can decide to explore a high confidence interactome or a larger one, accepting more false positives. The application and the web site that have been developed are designed to make the data stored in mentha accessible to all users. All the information contained in the local database is accessible through a web server. The Graphical Application embodies a wide selection of functionalities that help the user navigate and interact with a network of interest. The graphical application, formally a Java web applet, can be embedded in a web page such as iGoogle or even a scientific paper just by feeding it with an SDA, a DOI, or with UniProt IDs.

### Availability

<http://mentha.uniroma2.it/>

## miRNA-mRNA integrated pathway analysis: an application to colorectal cancer

T.M. Creanza<sup>1</sup>✉, R. Maglietta<sup>2</sup>, V.C. Liuzzi<sup>2</sup>, R. Anglani<sup>2</sup>, P.F. Stifanelli<sup>2</sup>, A. Piepoli<sup>3</sup>, S. Mukherjee<sup>4</sup>, F.P. Schena<sup>1</sup>, N. Ancona<sup>2</sup>

<sup>1</sup>Department of Emergency and Organ Transplantation - DETO, University of Bari, Italy

<sup>2</sup>Bioinformatics and Systems Biology Lab, Institute of Intelligent Systems for Automation, National Research Council - CNR, Bari, Italy

<sup>3</sup>Unità Operativa di Gastroenterologia, IRCCS, Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy

<sup>4</sup>Institute for Genome Science and Policy, Duke University, Durham, United States

### Motivations

In recent studies, it has been argued that small noncoding microRNAs (miRNAs) can contribute to development and progression of cancer and show a differential expression between normal and neoplastic tissues. To date, it is still not completely clear the functional role of miRNAs in human solid tumors both in terms of changes of expression profiles and in terms of changes of their regulatory activities. Our analysis aims to enlighten the specific contributes of miRNAs in cancer by identifying miRNA-driven pathways deregulated in tumor onset and progression.

### Methods

We present a strategy to infer biological processes potentially altered in tumor development by analyzing the expression of multiple microRNAs as well as by evaluating the cooperative miRNA regulatory activities on them. Our scoring algo-

rithm is based on the inference of pathways in terms of miRNAs determined by tissue-specific and condition-specific correlations evaluated on paired expression levels of human miRNAs and mRNAs. An enrichment analysis of the resulting miRNA pathways allows to identify pathways associated with differentially expressed miRNAs and pathways enriched for different miRNA regulatory activities by using tissue-specific and condition-specific measures, respectively.

### Results

Our novel approach integrates miRNA and mRNA expression data for the understanding of complex pathologies. The application of this approach to colorectal cancer highlights many biological pathways enriched of oncogenes and tumor suppressors that traditional mRNA pathway enrichment analyses were not able to reveal.

## Discovery of conserved long non-coding RNAs in vertebrates

S. Basu , R. Sanges

Bioinformatics - Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Napoli, Italy

### Motivations

Long non-coding RNAs (lncRNA) have been reported as a major class of novel transcripts related to organism development and early neural expression pattern [1-4]. They are reported to be expressed in large numbers in the mammalian transcriptomes [5,6] and recently reported to be expressed in the teleost fishes [7,8]. Computational identification and characterization of lncRNA from public sequence resources have been performed by different groups [9-11]. The focus of attention has been on the mammalian genomes starting by the assumption that they are not well conserved in term of sequence. However, systematic studies measuring their levels of conservation among vertebrates are lacking. Hence we want to computationally evaluate the existence of vertebrate conserved lncRNAs through systematic conservation analyses of both sequence as well as genomic architecture.

### Methods

Mouse lncRNAs reported in an earlier study [2] and predicted by the Ensembl pipeline were considered as a reference dataset. Homology search of the lncRNAs against the zebrafish conserved phastcons elements was performed with the BLAST program. The phastcons elements are regions of conservation in the zebrafish genome with human, mouse, western clawed frog and two teleost fishes, tetraodon and stickleback. The lack of selection pressure in lncRNAs as compared to the protein-coding genes required a calibration of BLAST parameters to define a cut-off score indicative of significant conservation. Using ROC analyses we calculated the best BLAST parameters able to select regions of lncRNA conserved in vertebrates. The predicted conserved candidates were also evaluated in terms of their RNA secondary structure using the RNAfold software. Gene ontology and expression pattern enrichment of flanking protein-coding genes was performed with DAVID software.

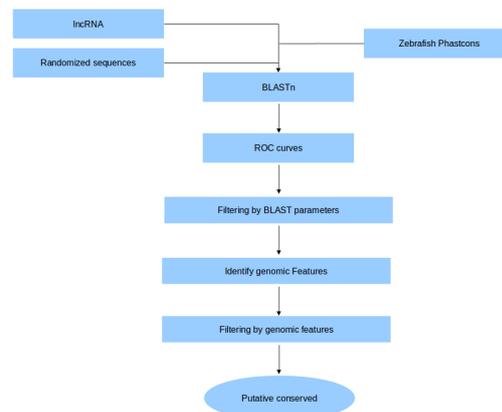


Figure 1. A schematic representation of the pipeline followed to identify putative conserved mouse long non-coding RNAs in the zebrafish phastcons elements.

### Results

Our results show that the usage of the alignment length as cut-off is sufficient to distinguish the conservation of mouse lncRNAs in zebrafish as compared to conservation of random genomic regions. The RNA secondary structure prediction was not able to define any threshold for conservation. From an initial dataset of ~2,800 lncRNAs we could predict that 235 are conserved using the defined cut-off on the alignment length. Gene ontology enrichment analyses, related to the protein-coding genes proximal to the region of conservation in mouse and zebrafish, highlighted corresponding GO classes such as regulation of transcription and central nervous system development. The proximal coding genes exhibited a similar enrichment for their tissue of expression where brain was highly enriched in both mouse as well as zebrafish. Two interesting candidate regions of conservation were chosen for future experimental validation based upon the presence of ESTs overlap and the function of the proximal proteins (in this case the interest being development and functioning of the nervous system). The analysis is poised as an initial pipeline to select interesting candidate lncRNAs conserved among vertebrates.

## References

1. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458: 223-227. doi:10.1038/nature07672.
2. Ponjavic J, Oliver PL, Lunter G, Ponting CP (2009) Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* 5: e1000617. doi:10.1371/journal.pgen.1000617.
3. Qureshi IA, Mattick JS, Mehler MF (2010) Long non-coding RNAs in nervous system function and disease. *Brain Res* 1338: 20-35. doi:10.1016/j.brainres.2010.03.110.
4. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci USA* 105: 716-721. doi:10.1073/pnas.0706729105.
5. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* 36: 40-45. doi:10.1038/ng1285.
6. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420: 563-573. doi:10.1038/nature01266.
7. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, et al. (2011) Systematic identification of long non-coding RNAs expressed during zebrafish embryogenesis. *Genome Research*. Available: <http://genome.cshlp.org/content/early/2011/11/22/gr.133009.111.abstract>. Accessed 23 November 2011.
8. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* 25: 1915-1927. doi:10.1101/gad.174466.11.
9. Khachane AN, Harrison PM (2010) Mining mammalian transcript data for functional long non-coding RNAs. *PLoS ONE* 5: e10316. doi:10.1371/journal.pone.0010316.
10. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, et al. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* 16: 1478-1487. doi:10.1261/rna.1951310.
11. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotech* 28: 503-510. doi:10.1038/nbt.1633.

## Non-coding RNA bioinformatics platform for full backing of the high-throughput sequencing experiments generated by Next-Generation Sequencing technologies

F. Licciulli<sup>✉</sup>, A. Consiglio, G. De Caro, A. Gisel, G. Grillo, A. Tulipano, S. Liuni

Istituto di Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, Bari, Italy

### Motivations

Short non-coding RNA molecules (20-30 nucleotides long) play an important role in the regulation of gene expression by interacting with their target RNAs. This interaction generally downregulates gene expression either affecting RNA stability or repressing translation. Different classes of small regulatory non coding RNAs (sncRNAs) have been discovered and studied so far, and new families continue to be described, which differ in the proteins required for their biogenesis, the mechanism of target recognition and regulation, and the biological pathways they control

[1,2]. In particular, three major classes of sncRNAs have been mostly investigated: small interfering RNAs (siRNAs), micro-RNAs (miRNAs) and PIWI-interacting RNAs (piRNAs) [1,2,3]. siRNAs direct the endonucleolytic cleavage of their target RNAs through a mechanism known as RNA interference (RNAi), miRNAs can repress translation or direct degradation of their target mRNA generally through imperfect complementary pairing on their 3'UTRs, whereas the major role of piRNAs is to ensure germline stability by repressing transposable elements (TEs). Recently, the advent of new Next-Generation Sequencing (NGS) tech-

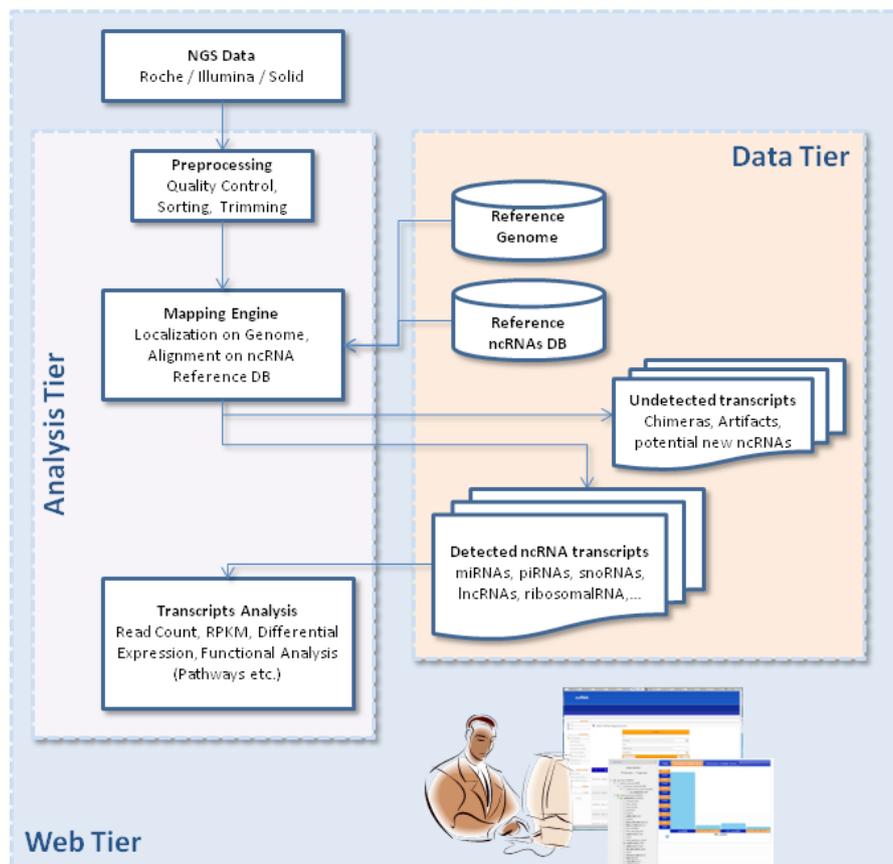


Figure 1: ncRNAs Platform Architectures

nologies has awfully increased the throughput of transcriptome studies, thus allowing an unprecedented investigation of non-coding RNAs. Regulatory pathways involving ncRNAs, such as miRNAs, are now being elucidated in detail and functions for long non-coding RNAs are also emerging. The huge amount of transcript data produced by high-throughput sequencing requires the development and implementation of suitable bioinformatics workflows for their analysis and interpretation. Here we describe here a bioinformatics resource to classify and analyze the non-coding RNA component of human transcriptome sequence data obtained by different NGS platforms (Roche 454, Illumina and Solid).

### Methods

The ncRNAs bioinformatics platform is organized according to a typical three tier architecture: an analysis tier for ncRNAs detection, classification and functional analyses; a data tier made up of a data-warehouse used to store the analysis results, the ncRNAs reference database (a non-redundant collection of ncRNAs sequence retrieved from fRNAdb, RNAdb, mirBASE, NONCODE and others), the reference genome and other useful annotation database like HGNC nomenclature [4], Sequence Ontology (SO) [5] and Entrez Gene; a web tier module for querying the analysis results and the annotation stored in the ncRNAs reference database. The core of the platform is the analysis workflow. In figure 1 we show the pipeline for classification and functional annotations of non-coding RNAs (ncRNAs) fraction obtained through high-throughput sequencing (HTS) experiments using different NGS technologies. The input data for the bioinformatics platform can be either the reads data obtained by different NGS platforms (Roche 454, Illumina and Solid) or previously mapped reads stored in users' SAM/BAM files.

### Results

The ncRNA bioinformatics platform - through a combination of an analyses pipeline, a data-warehouse and a user-friendly web interface - is able to:

- i. detect and classify reads in known functional ncRNA categories using Sequence Ontology classification, HGNC nomenclature, gene names and miRNA accessions;
- ii. extract reads collections belonging to a given category for further analysis;
- iii. quantify ncRNA expression based on annotations derived from different reference ncRNA databases;
- iv. generate some statistics of expressed ncRNAs, indicating the RPKM (reads per kilobase of RNA model per million mapped reads) value for each Sequence Ontology class;
- v. detect differential expression of ncRNAs between two conditions (i.e. normal/pathological);
- vi. create a collections of interesting clusters of reads mapped on the genome but not detected as known ncRNA;
- vii. filter out reads mapping to ribosomal RNAs and mtDNA transcripts;
- viii. create a collection of unmapped residual reads (chimeras, artifacts, and contaminations).

### References

9. Ghildiyal, M. and Zamore, P. D. Small (2009) "Silencing RNAs: an expanding universe". *Nature Rev. Genet.* 10, 94-108.
10. Malone, C. D. and Hannon, G. J. (2009), "Small RNAs as guardians of the genome". *Cell* 136, 656-668
11. Kim, N. V., Han, J. ' & Siomi M. C. (2009), " Biogenesis of small RNAs in animals". *Nature Rev. Mol. Cell Biol.* 10, 126-139
12. Wright M ' & Bruford E (2011), "Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature". *Human Genomics*, VOL 5. NO 2. 90-98.
13. Eilbeck et al.(2005), "The Sequence Ontology: A tool for the unification of genome annotations". *Genome Biology* 6:R44

## An improved procedure for clustering and assembly of large transcriptome data

E. Picardi<sup>1</sup>✉, V. Bevilacqua<sup>2</sup>, F. Stoppa<sup>2</sup>, G. Pesole<sup>1</sup>

<sup>1</sup>Istituto di Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche, Bari, Italy

<sup>2</sup>Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, Bari, Italy

### Motivations

Expressed sequence tags and full-length cDNAs represent an invaluable source of evidence for inferring reliable gene structures and discovering potential alternative splicing events [1]. However, to fully exploit their biological potential, correct and reliable EST clusters are required. To fill this gap we developed the program EasyCluster that resulted the most accurate when compared to software at the state of the art in this field [2]. Recent technological advances are dramatically increasing the number of available transcriptome reads. EST-like sequences can now be generated by pyrosequencing using Roche 454 platform which generates approximately one million reads per run. Handling such huge amount of EST-like data is basic to detect alternative splicing events, improve gene annotations or simply create gene-oriented clusters for expression studies. Sometimes EST-like data provide a fragmented overview of their genomic loci of origin and, thus, transcript assembly may be an optimal solution to annotate user-produced sequences. For these reasons we propose here a new implementation of EasyCluster able to manage genome scale transcriptome data and generate reliable gene-oriented clusters from 454 reads. The new version of EasyCluster software can facilitate downstream analyses because it enables the assembly of full-length transcripts per cluster, improves the clustering procedure using available annotations and embeds a graphical browser to provide an overview of results at genome level.

### Methods

EasyCluster is based on the well-known EST-to-genome mapping program GMAP [3] since it can perform a very quick mapping of whatever expressed sequence onto a genomic sequence and can detect splicing sites according to a so defined "sandwich" dynamic programming that is organism independent. Providing EST-like data from Roche 454 sequencer, EasyCluster initially runs GMAP program and parses results in order

to create an initial collection of pseudo-clusters by grouping EST-like reads according to the overlap of their genomic coordinates on the same strand. Then EasyCluster refines the EST grouping by including in each cluster only expressed sequences sharing at least one splice site. An ad hoc procedure is used to correct potential GMAP errors near splice sites and unspliced ESTs are added to each refined cluster. Finally, full-length transcripts are assembled for each cluster in order to evaluate the alternative splicing extent and provide gene expression levels according to user supplied annotations.

### Results

The new implementation of EasyCluster is written in Java programming language and provides improved clusters of EST sequences. It has been conceived to handle huge amount of EST-like reads produced by Roche 454 machines and supply a unique tool to cluster and assembly such transcriptome reads. Moreover, EasyCluster can now include unspliced reads and take benefit from available annotations. Alternative splicing is also inferred from each cluster after a refining procedure near exon-intron boundaries to reduce mapping errors due to GMAP. Accuracy and performances have been tested on simulated 454 reads by MetaSim software. Preliminary results indicate that the new EasyCluster implementation is highly efficient to manage and analyze deep transcriptome data from Roche 454 technology.

### References

1. Nagaraj, S.H., Gasser, R.B. and Ranganathan, S. (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in bioinformatics*, 8, 6-21.
2. Picardi, E., Mignone, F. and Pesole, G. (2009) EasyCluster: a fast and efficient gene-oriented clustering tool for large-scale transcriptome data. *BMC bioinformatics*, 10 Suppl 6, S10.
3. Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)*, 21, 1859-1875.

## miRandola: extracellular circulating microRNAs database

F. Russo<sup>1</sup>, S. Di Bella<sup>1</sup>, G. Nigita<sup>2</sup>, V. Macca<sup>2</sup>, A. Laganà<sup>3</sup>, R. Giugno<sup>1</sup>, A. Pulvirenti<sup>1</sup>, A. Ferro<sup>1</sup>✉

<sup>1</sup>Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy

<sup>2</sup>Department of Mathematics and Computer Science, University of Catania, Catania, Italy

<sup>3</sup>Department of Molecular Virology, Immunology and Human Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus OH, United States

### Motivations

MicroRNAs (miRNAs) are small (approximately 22 nt) noncoding RNAs that play an important role in the regulation of various biological processes through their interaction with cellular messenger RNAs. They are frequently dysregulated in cancer and have shown promise as tissue-based markers for cancer classification and prognostication. Extracellular miRNAs in serum, plasma, saliva, urine and other body fluids have recently been shown to be associated with various pathological conditions including cancer. miRNAs circulate in the bloodstream in a highly stable, extracellular form, thus they may be used as blood-based biomarkers for cancer and other diseases. Circulating miRNAs are protected by encapsulation in membrane-bound vesicles such as exosomes, but the majority of circulating miRNAs in human plasma and serum cofractionate with Argonaute2 (Ago2) protein, rather than with vesicles. In the present work, we performed a comprehensive classification of different extracellular circulating miRNA types. A direct link to the knowledge base miRò together with the inclusion of datamining facilities allow users to infer possible biological functions of the circulating miRNAs and their connection with the phenotype. To our knowledge miRandola is the first database that provides information about all kind of extracellular miRNAs and we believe that it will constitute a very important resource for researchers.

### Methods

miRandola is a manually curated database of circulating miRNAs. The database catalogs infor-

mation from both published (from literature and public resources) and unpublished studies (from direct researchers submission). The database include information about miRNAs and their extracellular form, samples, fluids and data sources. It has been created using MySQL, Apache and PHP. For better retrieval and analysis of the miRNA data we have integrated various tools such as search, advanced search and browsing. The search results provide links to miR, our miRNA knowledge base to help users to find useful information about miRNAs and their targets. Finally, an export function allows the download of the search results in various formats (CSV, XLS, TXT).

### Results

miRandola contains around 1700 entries. miRNAs are classified in three categories, based on their extracellular form: miRNA-Ago2, miRNA-exosome and miRNA-circulating. The latter is used when authors don't distinguish between Ago2 and exosome and constitutes the largest group. Simple detection and amplification methods, tissue-restricted expression profiles, and sequence conservation between human and model organisms make extracellular miRNAs ideal candidates for noninvasive biomarkers for the diagnosis and the study of various physiopathological conditions. A comprehensive classification of different extracellular circulating miRNA types is needed to help researchers to find miRNA signatures in human cancer and other diseases.

### Availability

<http://ferrolab.dmi.unict.it/>

## National Nodes

### Argentina

IBBM, Facultad de Cs.  
Exactas, Universidad  
Nacional de La Plata

### Brazil

Lab. Nacional de  
Computação Científica,  
Lab. de Bioinformática,  
Petrópolis, Rio de Janeiro

### Chile

Centre for Biochemical  
Engineering and  
Biotechnology (CIByB).  
University of Chile, Santiago

### China

Centre of Bioinformatics,  
Peking University, Beijing

### Colombia

Instituto de Biotecnología,  
Universidad Nacional de  
Colombia, Edificio Manuel  
Ancizar, Bogota

### Costa Rica

University of Costa  
Rica (UCR), School of  
Medicine, Department  
of Pharmacology and  
ClinicToxicology, San Jose

### Finland

CSC, Espoo

### France

ReNaBi, French  
bioinformatics platforms  
network

### Greece

Biomedical Research  
Foundation of the Academy  
of Athens, Athens

### Hungary

Agricultural Biotechnology  
Center, Godollo

### Italy

CNR - Institute for Biomedical  
Technologies, Bioinformatics  
and Genomic Group, Bari

### Mexico

Nodo Nacional de  
Bioinformática, EMBnet  
México, Centro de Ciencias  
Genómicas, UNAM,  
Cuernavaca, Morelos

### Norway

The Norwegian EMBnet  
Node, The Biotechnology  
Centre of Oslo

### Pakistan

COMSATS Institute of  
Information Technology,  
Chak Shahzaad, Islamabad

### Poland

Institute of Biochemistry and  
Biophysics, Polish Academy  
of Sciences, Warszawa

### Portugal

Instituto Gulbenkian de  
Ciencia, Centro Portugues  
de Bioinformatica, Oeiras

### Russia

Biocomputing Group,  
Belozersky Institute, Moscow

### Slovakia

Institute of Molecular Biology,  
Slovak Academy of Science,  
Bratislava

### South Africa

SANBI, University of the  
Western Cape, Bellville

### Spain

EMBnet/CNB, Centro  
Nacional de Biotecnología,  
Madrid

### Sri Lanka

Institute of Biochemistry,  
Molecular Biology and  
Biotechnology, University of  
Colombo, Colombo

### Sweden

Uppsala Biomedical Centre,  
Computing Department,  
Uppsala

### Switzerland

Swiss Institute of  
Bioinformatics, Lausanne

## Specialist- and Assoc. Nodes

### CASPUR

Rome, Italy

### EBI

EBI Embl Outstation, Hinxton,  
Cambridge, UK

### Nile University

Giza, Egypt

### ETI

Amsterdam, The Netherlands

### IHCP

Institute of Health and  
Consumer Protection, Ispra.  
Italy

### ILRI/BECA

International Livestock  
Research Institute, Nairobi,  
Kenya

### MIPS

Muenchen, Germany

### UMBER

Faculty of Life Sciences, The  
University of Manchester, UK

### CPGR

Centre for Proteomic and  
Genomic Research, Cape  
Town, South Africa

### The New South Wales Systems

Biology Initiative  
Sydney, Australia

for more information visit our Web site

[www.EMBnet.org](http://www.EMBnet.org)

# EMBnet.journal

## ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Instructions for authors" at <http://journal.EMBnet.org/index.php/EMBnetnews/about> and send your manuscript and supplementary files using our on-line submission system at <http://journal.EMBnet.org/index.php/EMBnetnews/about/submissions#onlineSubmissions>.

Past issues are available as PDF files from the Web site:

<http://journal.EMBnet.org/index.php/EMBnetnews/issue/archive>

### Publisher:

EMBnet Stichting p/a  
CMBI Radboud University  
Nijmegen Medical Centre  
6581 GB Nijmegen  
The Netherlands

Email: [erik.bongcam@slu.se](mailto:erik.bongcam@slu.se)

Tel: +46-18-4716696