

# EMBnet journal

Volume 21  
Supplement A  
December 2015

301

**Next Generation Sequencing: a look into the future**

**Final Conference & MC Meeting of COST Action BM1006**

**16-17 March 2015**

**Bratislava, Slovakia**

[http://seqahead.eu/bratislava\\_2015](http://seqahead.eu/bratislava_2015)



ESF provides  
the Cost Office  
through an EC  
contract



COST is supported  
by the EU RTD  
Framework  
Programme



# Editorial

The key task of COST Action BM1006, SeqAhead, *Next Generation Sequencing (NGS) Data Analysis Network*, was, as its name suggests, networking; but SeqAhead also emphasised the dissemination of knowledge. During the four years of the Action, SeqAhead surpassed every expectation: with members participating from 29 European countries, plus one international partner from South Africa, the Management Committee membership reads like a “who’s-who” of European NGS research.

This *EMBnet.journal* Conference Supplement clearly shows that during the four years of SeqAhead’s existence, the Action members actively shared software and experiences, and collaborated in numerous projects spanning diverse topics – from those at the biological end of the spectrum to those at the informatics end.

SeqAhead was acknowledged in more than 100 peer-reviewed articles, demonstrating that the Action’s activities benefited a large and wide audience.

*EMBnet.journal* Editorial Board

# Contents

Editorial .....	2
COST Action BM1006 (SeqAhead) closing conference .....	3
Scientific Programme.....	5
Keynote Lectures.....	9
Oral Presentations .....	13
Posters.....	25

## 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 Science, 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**, University of Fribourg, Fribourg, 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)

**Vicky Schneider-Gricar**, The Genome Analysis Centre (TGAC) Norwich, UK  
[vicky.sg@tgac.ac.uk](mailto:vicky.sg@tgac.ac.uk)

## COST Action BM1006 (SeqAhead) closing conference “Next Generation Sequencing: a look into the future”



Lubos Klucar<sup>1</sup>, Teresa K. Attwood<sup>2</sup>, Erik Bongcam-Rudloff<sup>3</sup>✉

<sup>1</sup>Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia

<sup>2</sup>University of Manchester, Manchester, United Kingdom

<sup>3</sup>Swedish University of Agricultural Sciences, Uppsala, Sweden

Klucar L *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e860. <http://dx.doi.org/10.14806/ej.21.A.860>

The closing conference and final Management Committee (MC) meeting of COST Action BM1006 (SeqAhead) took place over two days, 16-17 March 2015, in Bratislava (SK). Fifty participants from 20 European countries gathered at the event to discuss and share experiences, know-how related to new technologies, and key developments in the field of Next Generation Sequencing (NGS) research.

During the conference, keynote and selected speakers presented the latest NGS technologies and their applications, discussed storage and analysis systems, and presented the training available and still required to enable researchers to utilise these technologies in new fields of research. The event opened with a short welcome from the local organiser, Dr. Lubos Klucar (IMB SAS, Bratislava (SK)), followed by an introduction by the SeqAhead Chair, Prof. Erik Bongcam-Rudloff (SLU, Uppsala (SE)). Fourteen speakers presented their work and discussed new technology highlights in the field, including the latest minION device presented by Ola Wallerman (SLU, Uppsala (SE)).

The final SeqAhead MC meeting took place during the second day, where the principal achievements from the four years of the Action were presented and discussed. The impact of the project has been multi-faceted, ranging from the very abstract ‘improved healthcare’, owing to better trained NGS researchers in medical diagnostic facilities, to the very concrete, in terms of the research described in the nearly 100 peer-reviewed journal articles that acknowledge SeqAhead.

Doubtless, the most tangible result of SeqAhead is the large number of students/scientists who gained access to high-quality training in the emerging field of NGS, training that is not yet part of most formal university curricula, and was



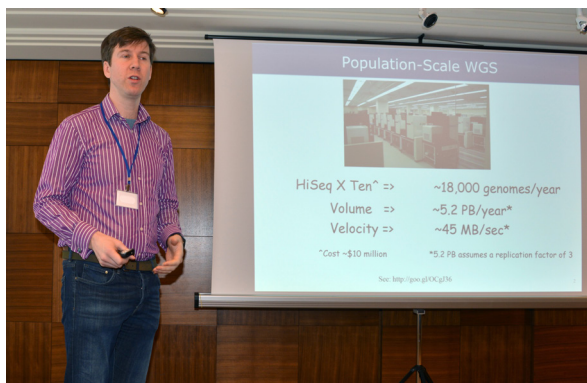
SeqAhead COST conference participants.

certainly not available to scientists who completed their studies a few years ago.

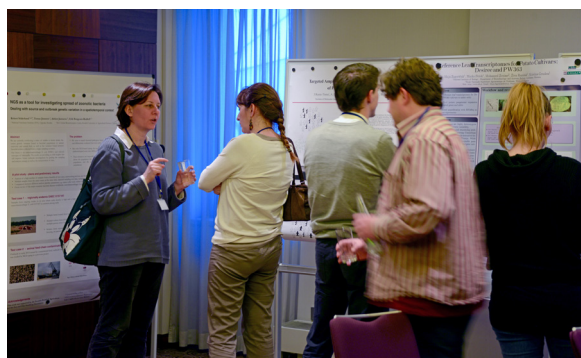
SeqAhead organised more than 30 workshops, four conferences with more than 150 participants each, and several schools, each with 25-60 students. In total, almost 2,000 scientists participated in these activities, the impact of which has been immediate for NGS research in Europe, and will continue to be felt for many years.

As already mentioned, SeqAhead partners have published about 100 articles on NGS and related topics, and many more publications are expected beyond the end of the project. Such articles help disseminate the scientific activities catalysed by SeqAhead and, we hope, will have a lasting impact: healthcare is a major financial burden in most European countries, so better NGS science in hospitals is likely to have both societal and economic benefits.

SeqAhead brought together researchers from across Europe, allowing them to discuss their sci-



Jim Dowling (KTH - Royal Institute of Technology) having talk on the biobankcloud project ([www.biobankcloud.com](http://www.biobankcloud.com)).



Poster session.

entific interests in formal and informal settings, and in physical and virtual (video) meetings. These meetings led to the formation of several large European collaborations, such as the (now complete) [AlIBio project](#)<sup>1</sup> and the newly awarded [COST Action CHARME](#)<sup>2</sup>, which is to kick-off in 2016. AlIBio, in turn, triggered the establishment of the recently initiated EU project [PROLIFIC](#)<sup>3</sup>. Members of SeqAhead also won an IRSES Marie Curie project aiming to create [an NGS network between the Americas and Europe](#)<sup>4</sup>.

We expect these new projects to have a lasting impact in many areas of life-science research, we are proud to have been part of the successful network that seeded them, and we wish them success in the years ahead. Finally, our thanks go out to all who made SeqAhead such a warmly collaborative and productive project!

Lubos Klucar (Local Organiser), Teresa K. Attwood (Action Vice Chair) and Erik Bongcam-Rudloff (Action Chair)



- 1 [www.allbioinformatics.eu/doku.php](http://www.allbioinformatics.eu/doku.php)
- 2 [www.cost.eu/COST\\_Actions/ca/CA15110](http://www.cost.eu/COST_Actions/ca/CA15110)
- 3 [euprolific.eu/index.php/the-project](http://euprolific.eu/index.php/the-project)
- 4 [www.deann.eu](http://www.deann.eu)





301

## Scientific Programme

---

**COST Action BM1006 - Final Conference**  
16-17 March 2015, Bratislava, Slovakia

## Scientific Programme

<b>16 March</b>	<b>COST Conference “NGS: a look into the future”</b>
08:30-09:00	<i>Registration and Coffee</i>
09:00-09:20	Welcome from organisers
	<b>NGS Technologies and their Applications</b>
09:20-10:00	<b>Keynote Lecture</b> <span style="float: right;"><i>Session Chair: Erik Bongcam-Rudloff</i></span> <b>Power and Limitations of RNA-Seq: Findings from the SEQC (MAQC-III) consortium</b> <u>Paweł Łabaj</u> <i>Chair of Bioinformatics Research Group, Boku University Vienna, Austria</i>
10:00-10:30	<i>Coffee Break &amp; posters set-up</i>
10:30-12:00	Oral Presentations
	<b>Reproducible Research in the era of Next Generation Sequencing: current approaches, examples and future perspectives</b> <u>Claudia Angelini</u> , Dario Righelli, Francesco Russo  <b>Current status of nanopore sequencing using the minION device – from full length cDNA sequencing to genome assembly improvements</b> <u>Ola Wallerman</u>  <b>Deep insights into Mecp2-driven transcriptional (de)regulation at embryonic developmental stage through RNA-Seq data analysis</b> <u>Kumar Parijat Tripathi</u> , Maurizio D'Esposito, Mario R Guarracino, Marcella Vacca  <b>Exploring the activity of microorganisms in the forest soil using metatranscriptomics</b> <u>Petr Baldrian</u>
12:00-13:00	<i>Lunch</i>
	<b>Storage and Analysis Systems</b>
13:00-13:40	<b>Keynote Lecture</b> <span style="float: right;"><i>Session Chair: Ralf Herwig</i></span> <b>Population-Scale Genomics with BiobankCloud and Hops/Hadoop</b> <u>Jim Dowling</u> <i>KTH - Royal Institute of Technology, Stockholm, Sweden</i>
13:40- 15:00	Oral Presentations
	<b>Creating a successful facility for large-scale extraction of DNA, an example from the Swedish biobank initiative BBMRI.se</b> <u>James Thompson</u>  <b>Using neural networks to filter predicted errors in NGS Data</b> <u>Dimitar Vassilev</u> , Milko Krachunov  <b>DIANA-TarBase v7: indexing thousands experimentally supported miRNA:mRNA interactions</b> <u>Dimitra Karagkouni</u>  <b>Biobanking for the future, how to prepare for the next generation of next generation sequencing</b> <u>Tomas Klingström</u>
15:00-15:30	<i>Coffee Break</i>
15:30-16:10	<b>Keynote Lecture</b> <span style="float: right;"><i>Session Chair: Claudia Angelini</i></span> <b>Data deluge demands training tsunami</b> <u>Eija Korpelainen</u> <i>CSC, Espo, Finland</i>



16:10-17:10	Oral Presentations
	<p><b>Introducing Meta<sup>2</sup>genomics: the search for the “micro-bee”</b>  <u>Jose R Valverde</u></p> <p><b>Information-theoretic approach for detection of differential splicing from RNA-seq data</b>  <u>Ralf Herwig</u></p> <p><b>iMir: An innovative and complete pipeline for smallRNA-Seq data analysis</b>  <u>Giorgio Giurato</u>, Antonio Rinaldi, Adnan Hashim, Giovanni Nassa, Maria Ravo, Francesca Rizzo, Roberta Tarallo, Angela Cordella, Giovanna Marchese, Domenico Memoli, Alessandro Weisz</p>
17:10-17:20	Closing remarks
17:20-18:00	Poster session
19:30	<i>Dinner</i>

<b>17 March</b>	<b>SeqAhead Management Committee Meeting</b>
08:30-09:00	<i>Morning Coffee</i>
09:00-09:15	Welcome and introduction
09:15-10:00	Business meeting
10:00-10:30	<i>Coffee Break</i>
10:30-12:00	Business meeting
12:00	<i>Lunch</i>

## Chairs and Conference Committees

### Scientific Committee

Prof. Teresa Attwood (University of Manchester, Manchester, United Kingdom)  
 Dr. Claudia Angelini (CNR, Naples, Italy)  
 Dr. Erik Bongcam-rudloff (SLU, Uppsala, Sweden)  
 Dr. Laurent Falquet (University of Fribourg, Fribourg, Switzerland)  
 Prof. Jacques van Helden (AMU, Marseille, France)  
 Mr. Oliver Hunewald (CRP-Sante, Luxembourg, Luxembourg)  
 Dr. Lubos Klucar (IMB SAS, Bratislava, Slovakia)  
 Prof. Claude Muller (CRP-Sante, Luxembourg, Luxembourg)  
 Dr. Guy Perriere (Université Claude Bernard, Lyon, France)  
 Dr. Jose Ramon Valverde (CNB/CSIC, Madrid, Spain)  
 Dr. Dimitar Vassilev (Agro Bio Institute, Sofia, Bulgaria)  
 Dr. Vicky Schneider (TGAC, Norwich, United Kingdom)

### Scientific Organisers

Prof. Teresa Attwood (University of Manchester, Manchester, United Kingdom)  
 Dr. Erik Bongcam-Rudloff (SLU, Uppsala, Sweden)  
 Dr. Lubos Klucar (IMB SAS, Bratislava, Slovakia)

### Session Chairs

Dr. Erik Bongcam-Rudloff (SLU, Uppsala, Sweden)  
 Dr. Ralf Herwig (Max-Planck-Institute for Molecular Genetics, Berlin, Germany)  
 Dr. Claudia Angelini (CNR, Naples, Italy)

# List of presentations

(presenting authors)

## Keynote Lectures

- Power and limitations of RNA-Seq; findings from the SEQC (MAQC-III) Consortium  
*Paweł Piotr Łabaj* ..... 10
- Population-scale genomics with BiobankCloud and Hops/Hadoop  
*Jim Dowling* ..... 11
- Data deluge demands a training tsunami  
*Eija Korpelainen* ..... 12

## Oral Presentations

- Reproducible Research in the era of Next Generation Sequencing: current approaches, examples and future perspectives  
*Claudia Angelini* ..... 14
- Current status of nanopore sequencing using the MinION device – from full length cDNA sequencing to genome assembly improvements  
*Ola Wallerman* ..... 15
- Deep insights into Mecp2-driven transcriptional (de) regulation at embryonic developmental stage through RNA-Seq data analysis  
*Kumar Parijat Tripathi* ..... 16
- Exploring the activity of microorganisms in the forest soil using metatranscriptomics  
*Petr Baldrian* ..... 17
- Creating a successful facility for large-scale extraction of DNA, an example from the Swedish Biobank initiative BBMRI.se  
*James Thompson* ..... 18
- Using neural networks to filter predicted errors in NGS data  
*Dimitar Vassilev* ..... 19
- DIANA-TarBase v7: indexing hundreds of thousands experimentally supported miRNA:mRNA interactions  
*Dimitra Karagkouni* ..... 20
- Biobanking for the future, how to prepare for the next generation of Next Generation Sequencing  
*Tomas Klingström* ..... 21
- Introducing Meta<sup>2</sup>genomics: the search for the “micro-bee”  
*José R. Valverde* ..... 22

- Information-theoretic approach for detection of differential splicing from RNA-seq data  
*Ralf Herwig* ..... 23

- iMir: an innovative and complete pipeline for smallRNA-Seq data analysis  
*Giorgio Giurato* ..... 24

## Posters

- Microbial analysis of ovine cheese by next generation sequencing  
*Vladimir Kmet* ..... 26
- PACMAN: PacBio Methylation Analyzer  
*Laurent Falquet* ..... 27
- NORM-SYS - harmonizing standardization processes for model and data exchange in systems biology  
*Babette Regierer* ..... 28
- Targeted amplicon sequencing in genetic diagnostics of patients with cystic fibrosis  
*Jelena Kusic-Tisma* ..... 29
- Reference leaf transcriptomes for potato cultivars: Desiree and PW363  
*Maja Zagorščak* ..... 30
- NGS as a tool for investigating spread of zoonotic bacteria: dealing with source and outbreak genetic variation in a spatio-temporal context  
*Robert Söderlund* ..... 31
- Biobanks and future emerging technologies: new approaches, new pre-analytical challenges  
*Eva Ortega-Paino* ..... 32
- MetLab: an In silico experimental design, simulation and validation tool for viral metagenomics studies  
*Hadrien Gourelé* ..... 33
- Towards SNP calling in polyploid genomes  
*Anatoliy Dimitrov* ..... 34
- Gene set enrichment analysis of neuroendocrine system of the silkworm *Bombyx mori*  
*Gabor Beke* ..... 35
- The NonCode aReNA DB: a non-redundant and integrated collection of non-coding RNAs  
*Flavio Licciulli* ..... 36



301

## Keynote Lectures

---



## Power and limitations of RNA-Seq: findings from the SEQC (MAQC-III) Consortium



### Paweł Piotr Łabaj

Chair of Bioinformatics Research Group, Boku University Vienna, Vienna, Austria

Łabaj PP (2015) *EMBnet.journal* 21(Suppl A), e831. <http://dx.doi.org/10.14806/ej.21.A.831>.

We present primary results from the Sequencing Quality Control (SEQC) project by US-FDA MAQC consortium. Here we present a multi-centre cross-platform study introducing a landmark RNA-Seq reference dataset comprising 30 billion reads. In addition to NGS also several microarray and qPCR platforms were examined. The study design supports large variety of complementary benchmark metrics by featuring known mixtures, high-dynamic range ERCC-spikes, as well as nested replication structure. With no independent 'gold standard' feasible, these built-in truths support an objective assessment of performance and are critical for the development and validation of novel or improved algorithms and data processing pipelines. We find that measurements of relative expression are accurate and reproducible across sites and platforms

if specific filters are used. Comparisons with microarrays identified complementary strengths, with RNA-Seq at sufficient read-depth detecting differential expression more sensitively, and microarrays achieving higher rank-reproducibility. Measurement performance depends on the platform and data analysis pipeline, and variation is large for transcript-level profiling. On the other hand, even at read-depths >100 million, we find thousands of novel junctions, with good agreement between platforms, and with qPCR validation-rates >80%. We also have shown that the modelling approaches for inferring alternative transcripts expression-levels from read counts along a gene can be applied to probes along a gene in high-density next-generation microarrays. This has advantages in quantitative transcript-resolved expression profiling.



## Population-scale genomics with BiobankCloud and Hops/Hadoop



### Jim Dowling

KTH - Royal Institute of Technology, Stockholm, Sweden

Dowling J (2015) *EMBnet.journal* 21(Suppl A), e825. <http://dx.doi.org/10.14806/ej.21.A.825>

Recent advances in the cost, throughput, and speed of Next-Generation Sequencing (NGS) technology have resulted in huge growth in both the volume and velocity of genomic data available for processing.

Large NGS projects are now starting with the goal of population-based genomics, that is, the analysis of genomes of tens or even hundreds of thousands of individuals. Population-based genomics needs petabyte-scale data analytics platforms. BiobankCloud is an open-source framework, based on [Hadoop](#)<sup>1</sup>, that supports the scalable and secure storage and analysis of NGS data, alongside traditional Biobank meta-data. The platform scales to manage petabytes

of data and it supports a number of Hadoop-based data analytics platforms for scale-out processing of NGS data, such as Cuneiform/HiWAY.

BiobankCloud supports multi-tenancy for studies with sensitive data. A project-based model for authentication has been incorporated into the Hadoop ecosystem, and studies can co-exist on the same platform without users accidentally or maliciously accessing each others study data. The multi-tenancy support is based around a user interface that brings together, in a single platform, the owners of NGS data with bioinformaticians, the researchers who typically analyze NGS data.

<sup>1</sup> [hadoop.apache.org/](http://hadoop.apache.org/)

## Data deluge demands a training tsunami



### Eija Korpelainen

CSC - IT Center for Science, Espoo, Finland

Korpelainen E (2015) *EMBnet.journal* **21**(Suppl A), e833. <http://dx.doi.org/10.14806/ej.21.A.833>

Modern sequencing technologies open unprecedented possibilities for life science research. As new sequencing applications or “seqs” become available and the cost is going down, sequencing has become the measurement technology of choice for more and more researchers. However, in order to exploit sequencing to its full extent, substantial data analysis skills are required. This is a bottleneck for several reasons: each “seq” requires its own analysis methods, methods are changing rapidly as the field hasn’t matured yet, and the sheer volume of the data poses technical challenges. The situation is ag-

gravated by the fact that life science researchers typically have a bio/medical background and very little experience in programming and statistics. Taken together, the need for training in up-to-date data analysis skills is massive, but luckily international efforts are already underway to improve the situation. The truly global effort in this regard is [GOBLET](http://mygoblet.org)<sup>1</sup>, the Global Organisation for Bioinformatics Learning, Education & Training. This talk discusses several aspect of training, such as who should be trained, what should be taught and how, who should provide the training, and who will train the trainers.

---

<sup>1</sup> [mygoblet.org](http://mygoblet.org)



301

## Oral Presentations

---





## Reproducible Research in the era of Next Generation Sequencing: current approaches, examples and future perspectives

Claudia Angelini<sup>✉</sup>, Dario Righelli, Francesco Russo

Istituto per le Applicazioni del Calcolo "M. Picone", Napoli, Italy

Angelini C *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e808. <http://dx.doi.org/10.14806/ej.21.A.808>

Next Generation Sequencing (NGS) has revolutionised the way of thinking and of performing biomedical research. It is now possible to investigate biological aspects of cell functionalities and to understand previously unexplored disease etiologies by analyzing multi-omic data. Several computational (open-source) tools have been developed to analyze NGS data. Despite the technological advances, the way in which data analysis is performed and described in most of scientific papers does not facilitate the reproducibility of scientific results. Complex analyses are often poorly described and the lack of the technical information impedes the possibility for a researcher to reproduce the results available in literature (Nekrutenko and Taylor, 2012). Therefore, the problem of the "Reproducible Research", here-denoted RR, (Stodden *et al.*, 2014) is emerging as a serious issue for all Life Sciences.

In this work, we first introduce the concept of RR, its main benefits and challenges. Then, we discuss a simple way to develop novel computational tools that implement RR by using R and

Bioconductor packages. In particular, we show how RR can be incorporated within a graphical user-friendly interface and how such tools can automatically generate executable analysis reports. Then, we describe how it is possible to speed up repetitive and computational expensive function calls by using results stored in a cache memory (the latter point is crucial for the analysis of "Big Data" as those collected with NGS experiments). As a concrete working example of our approach, we illustrate the advances we have introduced in RNASeqGUI (Russo and Angelini, 2014).

### References

- Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* **13**, 667-672. <http://dx.doi.org/10.1038/nrg3305>
- Stodden V, Leisch F, Peng RD (2014) Implementing Reproducible Research. In: Chapman & Hall/CRC The R Series.
- Russo F, Angelini C (2014) RNASeqGUI: A GUI for analyzing RNA-seq data. *Bioinformatics* **30**(17), 2514-2516. <http://dx.doi.org/10.1093/bioinformatics/btu308>



## Current status of nanopore sequencing using the MinION device – from full length cDNA sequencing to genome assembly improvements

Ola Wallerman

Uppsala University and Swedish University of Agricultural Sciences, Uppsala, Sweden

Wallerman O (2015) *EMBnet.journal* 21(Suppl A), e819. <http://dx.doi.org/10.14806/ej.21.A.819>

The MinION is a small device that may transform the way sequencing is done by taking the sequencer out of the core labs. It detects the current through nanopores as single DNA strands passes through in order to decode the sequence and is capable of generating sequences longer than any other platform currently on the market. The MinION is under active development in an early-access program (MAP). Like other single molecule sequencers, the accuracy of one-pass reads is lower than that of the amplification-based sequencers but through recent improvements to chemistry and basecalling it is now possible to generate long reads with over 90% alignment accuracy.

We are studying the transcription factor ZBED6, which is a repressor of IGF2 in muscle tissue. It is co-expressed with ZC3H11A through intron reten-

tion in a 13 kb transcript. From Illumina RNA-seq after knockout of ZBED6 we observe a drastic change of IGF2 expression, and also a potential shift in promoter usage. As a member of MAP I plan use the MinION to analyse isoforms of these transcripts from full-length cDNA sequencing, but given the nature of the reads it has a great potential also to be used for improvement of genome assemblies. As a first experiment to get a better understanding of the system and its reads I generated long genomic reads from the honeybee *Apis Mellifera*, which is used as a model organism in the lab and here I will describe these results and discuss the current status and future perspectives of nanopore sequencing in terms of data quality and throughput.

Identities = 8228/8961 (92%), Gaps = 560/8961 (6%)  
Strand=Plus/Minus

```

Query 7706      CAAA-TTCGAAAGTC-TTTTGATCACGGCTTTCTAATGGATCACGAATACTTTGGAGATT 7763
          |||| |
Sbjct 293126     CAAAATTCGAAAGTC-TTTTGATCACGGCTTTCTAATGGATCACGAATACTTTGGAGATT 293067

Query 7764      ATCTTCAAACGATTATACTAAGTCAAATCTAAATAGCAAGATTAATTATCAAATCTAAT 7823
          |||| |
Sbjct 293066     ATCTTCAAACGATTATACTAAGTCAAATCTAAATAGCAAGATTAATTATCAAATCTAAT 293007

Query 7824      TTTT-AATCATCAATTCTTAATTCAATAATTTCAAGTAGAAATACAATATGTTT--AAA 7880
          |||| |
Sbjct 293006     TTTTAAATCATCAATTCTTAATTCAATAATTTCAAGTAGAAATACAATATGTTTTTAAA 292947

```

Figure 1. The MinION device is capable of sequencing long unamplified genomic samples with high accuracy. BLAST alignment of an 8.2 kb read shows that most errors are due to gaps from skipped bases.

## Deep insights into Mecp2-driven transcriptional (de)regulation at embryonic developmental stage through RNA-Seq data analysis

Kumar Parijat Tripathi<sup>1</sup>✉, Maurizio D'Esposito<sup>2</sup>, Mario R. Guarracino<sup>1</sup>, Marcella Vacca<sup>2</sup>

<sup>1</sup>Genomic, Proteomic and Transcriptomic Laboratory, National Research Council of Italy (CNR), Institute for High-Performance Computing and Networking (ICAR), Napoli, Italy

<sup>2</sup>Institute of Genetics and Biophysics - ABT, Napoli, Italy

Tripathi KP *et al.* (2015) *EMBnet.Journal* **21**(Suppl A), e816. <http://dx.doi.org/10.14806/ej.21.A.816>

Rett Syndrome (RTT, MIM 312750) is a progressive X-linked neurodevelopmental disorder due to mutation of the Mecp2 gene (encoding the transcription regulator methyl-CpG binding protein 2) (Amir *et al.*, 1999). How mutations in Mecp2 lead to the neuropathological signs of RTT is still unknown and there is no cure for this devastating disorder. RTT typically manifests months after birth (by 5 weeks in null male mice), arguing that key embryonic and perinatal developmental steps take place normally in affected individuals. Despite this, we find an altered transcriptome and a significantly lower primary neurotrophic branching in Mecp2 null cortical neurons dissociated from single embryos (embryonic day 15) compared to wild type cultures. The pharmacological stimulation of a serotonin receptor normalizes defective neurite branching and, partially, transcriptional deregulation due to MeCP2 loss. To understand the biological mechanism behind transcriptional remodeling under the influence of Mecp2 knocking out, we need large-scale study of the transcriptional response of null cortical neurons before and after treatment with serotonin receptor stimulator. RNA-Seq is a new tool, which utilises high-throughput sequencing to measure RNA transcript counts at an extraordinary accuracy. It provides quantitative means of exploring the transcriptome of an organism of interest. Total RNA was extracted from wild type and Mecp2 null cortical neurons, sequenced and analysed with a gene specific approach using Tuxedo pipeline, i.e., Tophat2 (Kim *et al.*, 2013) and cufflinks package (Trapnell *et al.*, 2012). The intersection of differentially expressed genes dataset from untreated Mecp2 null neu-

rons with the dataset from treated versus untreated WT permits to distinguish transcripts influenced only by the specific genotype from those responding to the compound. A subset of latter category rescues a level of physiological expression as consequence of chemical treatment. Using in-house built computational pipeline "Transcriptator" (automated computation pipeline to annotate assembled reads) (Tripathi *et al.*, 2014) we annotated the functional as well gene ontological terms associated with this different set of transcripts. Specific functional terms related to SMART, PIR superfamily, InterPro domains and KEGG pathways are enriched in each set of transcripts. To understand more clearly the precise mechanism of the used drug, we checked for both specific and common functional annotation terms for the different categories of transcripts. We also show distribution of Panther pathways and Biological process GO terms between transcripts responding to the treatment (rescued and newly recruited genes).

### References

- Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U *et al.* (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-188.
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**, 562-578. <http://dx.doi.org/10.1038/nprot.2012.016>
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**:R36. <http://dx.doi.org/10.1186/gb-2013-14-4-r36>
- Tripathi KP *et al.* (2014) Transcriptator: computational pipeline to annotate transcripts and assembled reads from RNA-seq data. *Lecture Notes in Bioinformatics*, In Press.



## Exploring the activity of microorganisms in the forest soil using metatranscriptomics

**Petr Baldrian**

Institute of Microbiology of the ASCR, Prague, Czech Republic

Baldrian P (2015) *EMBnet.journal* **21**(Suppl A), e811. <http://dx.doi.org/10.14806/ej.21.A.811>.

Understanding the ecology of forest soils is very important because they belong to the largest carbon sinks globally. Metatranscriptomics seems to be perfectly suitable for the exploration of microbial involvement in soil processes but until recently it was technically difficult to use it successfully for these highly complex environments. Here we used metatranscriptomics combined with microbial community analysis to describe the roles of individual microbial taxa in the coniferous forest soil in two contrasting seasons. Both the microbial community composition and the pool of microbial transcripts were found to be highly diverse and characterised by the high abundance and activity of fungi. The differences in seasonal functioning of the ecosystem consisted of a combination of moderate changes in microbial community composition and profound

changes in taxon-specific microbial transcription profiles. These differences were more significant in soil than in litter. Most importantly, fungal contribution to total microbial transcription in soil decreased from 33% in summer to 16% in winter. In particular, the activity of the ectomycorrhizal fungi that quantitatively dominate this environment was reduced in winter. The results indicate that plant photosynthetic production was likely the major driver of changes in the functioning of microbial communities in the studied ecosystem across seasons. Technically, the annotation of functions and taxonomic affiliations is relatively precise for bacteria but less reliable for the fungi and archaea, due to the lower number of fully sequenced and annotated genomes. These limitations will hopefully decrease in the future along with the advances in microbial genomics.

## Creating a successful facility for large-scale extraction of DNA, an example from the Swedish Biobank initiative BBMRI.se

**James Thompson**

Karolinska Institutet Biobank, Stockholm, Sweden

Thompson J (2015) *EMBnet.journal* **21**(Suppl A), e830. <http://dx.doi.org/10.14806/ej.21.A.830>

Driven by the needs of several large population cohorts, the Biobanking and BioMolecular Resource Infrastructure of Sweden (BBMRI.se) has established, and now operates, a large-scale facility for the rapid purification of high quality DNA from human whole blood at the KI Biobank. The facility has a throughput of up to 1000 whole blood samples per day on two parallel automated systems. Automation and modern extraction chemistry has given many benefits, and since starting the operation in May 2011, we

have extracted DNA from over 100 000 individuals. Our early experience clearly demonstrated the potential of the new systems in speed and cost. But we also experienced several difficult challenges with the approach. Genotyping and NGS platforms are very sensitive to DNA quality parameters. After gaining significant process improvements we now see study designs that use genotyping and NGS delivering results that promise to transform healthcare.



## Using neural networks to filter predicted errors in NGS data

Milko Krachunov<sup>1</sup>, Ognyan Kulev<sup>1</sup>, Maria Nisheva<sup>1</sup>, Valeria Simeonova<sup>1</sup>, Deyan Peychev<sup>2</sup>,  
Dimitar Vassilev<sup>2</sup>✉

<sup>1</sup>Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

<sup>2</sup>AgroBioInstitute, Bioinformatics group, Sofia, Bulgaria

Krachunov M *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e827. <http://dx.doi.org/10.14806/ej.21.A.827>.

The amount of sequencing errors produced by NGS technologies is low, but not negligible. Some studies, such as single nucleotide polymorphism (SNP) calling in metagenomics, are very sensitive to any noise present in the sequencing data, and would greatly benefit from precise error detection techniques to discover incorrect bases, without flagging the real variation in the data as erroneous. There are still no general solutions to the error detection problem that can be applied to studies like these.

As part of this work, a neural network is trained to recognise errors with a high accuracy in a dataset containing high natural variation. The neural network is then applied to a set of possible errors that were predicted with a sub-optimal detection algorithm that produces a high amount

of false positives. Due to the classification accuracy of the network, there is a net decrease in the number of false positives without a significant increase in the number of false negatives. The combination significantly increases the accuracy of the sub-optimal approach. A 46-61% decrease in the number of predicted errors, all from incorrectly identified errors, is observed when the neural network is applied over a set predicted with frequencies and thresholds.

### Acknowledgements

The study is supported by the National Science Fund of Bulgaria within the "Methods for Data Analysis and Knowledge Discovery in Big Sequencing Datasets" project under contract DFNI-I02/7 of 12.12.2014.

## DIANA-TarBase v7: indexing hundreds of thousands experimentally supported miRNA:mRNA interactions

Dimitra Karagkouni<sup>1</sup>✉, Ioannis S. Vlachos<sup>1</sup>, Maria D. Paraskevopoulou<sup>1</sup>, Georgios Georgakilas<sup>1</sup>, Thanasis Vergoulis<sup>2</sup>, Theodore Dalamagas<sup>2</sup>, Artemis G. Hatzigeorgiou<sup>1</sup>

<sup>1</sup>DIANA-Lab, Department of Electrical & Computer Engineering, University of Thessaly, Thessaly, Greece

<sup>2</sup>'Athena' Research and Innovation Center, Athens, Greece

Karagkouni D *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e824. <http://dx.doi.org/10.14806/ej.21.A.824>

microRNAs are short non-coding RNAs which act as potent post-transcriptional regulators. Accurate identification and cataloging of miRNA targets is crucial to understanding their function. Currently, hundreds of thousands of miRNA:gene interactions have been experimentally identified. Numerous wet lab methodologies have been developed, enabling the validation of predicted miRNA interactions or the high-throughput screening and identification of novel miRNA targets. However, this wealth of information is fragmented and hidden in thousands of manuscripts and raw next generation sequencing data sets.

[DIANA-TarBase v7.0](#)<sup>1</sup> aims to provide for the first time hundreds of thousands of high-quality manually curated experimentally validated miRNA:gene interactions. A text-mining pipeline has been implemented for the identification of all the advanced in experimental methodologies articles which have been subjected to manual curation.

DIANA-TarBase v7.0 has been significantly extended with richer meta-data and detailed information for each interaction, while the interface now supports advanced real-time queering and result filtering. The database enables users to easily identify positive or negative experimental results, the utilised experimental methodology,

experimental conditions including cell/tissue type and treatment. The new interface provides also advanced information ranging from the binding site location, as identified experimentally as well as in silico, to the primer sequences used for cloning experiments.

DIANA-TarBase v7.0 is the first relevant database which breaks the barrier of hundreds of thousands entries by indexing more than half a million interactions in 24 species, 9–250 times more than any other manually curated database. This wealth of information can be utilised for exploratory studies and can significantly boost the understanding of miRNA:mRNA collaboration.

### Acknowledgements

This research has been co-financed by the European Social Fund – ESF and National Resources, Greek national funds through the “Operational Program Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales.

### References

Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T *et al.* (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.* **43** (D1), D153-D159. <http://dx.doi.org/10.1093/nar/gku1215>

<sup>1</sup> [www.microrna.gr/tarbase](http://www.microrna.gr/tarbase)

## Biobanking for the future, how to prepare for the next generation of Next Generation Sequencing

Tomas Klingström

Swedish University of Agricultural Sciences, Uppsala, Sweden

Klingström T (2015) *EMBnet.journal* 21(Suppl A), e815. <http://dx.doi.org/10.14806/ej.21.A.815>

In biobanking the quality of samples is defined depending on their usage. A piece of formalin-fixed, paraffin-embedded tissue may therefore be considered of very high quality when used for immune histochemistry, despite heavy crosslinking and fragmentation of nucleic acids. This, combined with ethical and legal questions on informed consent, makes it important for biobanks to prepare their collections for future technology.

The technology watch of the BBMRI-LPC 2015 will focus on the pre-analytical, ethical and data management issues that may prevent biobanks from providing samples that fully take advantage of the next generation of sequencing technology. Biobanks are a long term commitment, but by preparing for the next generation of technology, biobanks provide researchers with immediate access to material for testing scientific theories that would otherwise take years or decades to evaluate.

To properly support next generation sequencing it is therefore especially important that biobanks consider the following challenges:

- *technical challenges* - DNA is often regarded as stable and easy to work with. But NGS with long read lengths, epigenetic studies and single cell sequencing are all areas where current or upcoming technologies will greatly challenge the quality management of biobanks;
- *ethical challenges* - every sample may be a genetic sample if the research is innovative and meticulous enough. This greatly increases the requirements on how biobanks handle donor consent and privacy when collecting biospecimen;
- *data management* - despite huge cost reduction, NGS remains the costliest step in the process. Data access to ensure reuse and repeatability is therefore important.



## Introducing Meta<sup>2</sup>genomics: the search for the “micro-bee”

**José R. Valverde**

CSIC, Centro Nacional de Biotecnología, Madrid, Spain

Valverde JR (2015) *EMBnet.journal* **21**(Suppl A), e835. <http://dx.doi.org/10.14806/ej.21.A.835>

We will describe what may be called *meta-metagenomics* (or *Meta<sup>2</sup>genomics*), a novel approach to mine metagenomic data and how we apply it to search for microbial sentinels of contamination, the microbial equivalent of bees (or “micro-bee”).

One of the main challenges of Next Generation Sequencing (NGS) is its pervasive use for afterthought research: mining of existing experiments developed with a specific aim for the pursuit of novel, unrelated quests that were not foreseen in advance.

The last years have seen a tremendous increase in the availability of metagenomic data collected from soil samples under a variety of soil textures, locations, circumstances and treatments. All these experiments were carried out with specific aims in mind: understand the variability of the rhizobacterial community under specific environmental conditions.

In parallel with these advances, and partly thanks to some of these studies, society has become increasingly aware of the impact that human activities have on the environment, and is trying to adapt policies and operation procedures to better exploit this newly acquired knowledge. As an example, studies of the impact of herbicide treatments on soil rhizobacteria has been one of the factors considered in the ap-

proval or renewal (or its deny thereof) of the use of herbicidal preparations in the EU and elsewhere.

An orthornormal approach is used in the biological security field. Here, the problem is not knowing and preventing the effect of specific substances, but preventing and detecting the release of biological or toxic agents, and reacting accordingly to minimize their impact on humans and the environment at large. We are interested in the development of cost-effective advisory policies that may help professionals in the detection of potentially harmful spills, specially in developing countries.

Here we describe our quest for a sensitive test that may help detect contamination changes in the soil through the identification of the “micro-bee”, a soil microorganism taxon that may act as a sentinel, developing in the process what can be called *meta-metagenomics* (or *Meta<sup>2</sup>genomics*): mining many unrelated metagenomics experiments to identify a common trend of interest. In effect, this requires to select multiple soil metagenomics experiments from the Sequence Read Archive, re-analyse each of them towards our ends, and compare them for common trends that meet our end criteria, developing new planning, analysis, comparison and statistical technologies in the process.

## Information-theoretic approach for detection of differential splicing from RNA-seq data

Axel Rasche, Ralf Herwig 

Max-Planck-Institute for Molecular Genetics, Dep. Computational Molecular Biology, Berlin, Germany

Rasche A and Herwig R (2015) *EMBNet.journal* **21**(Suppl A), e828. <http://dx.doi.org/10.14806/ej.21.A.828>

The computational prediction of alternative splicing from high-throughput sequencing data is inherently difficult and necessitates robust statistical measures because the differential splicing signal is overlaid by influencing factors such as gene expression differences and simultaneous expression of multiple isoforms, among others. In this work we describe ARH-seq (Rasche *et al.*, 2014), a discovery tool for differential splicing in case-control studies, that is based on the information-theoretic concept of entropy. ARH-seq works on high-throughput sequencing data and is an extension of the ARH method that was originally developed for exon microarrays (Rasche and Herwig, 2010). We show that the method has inherent features, such as independence of transcript exon number and independence of differential expression, what makes it particularly suited for detecting alternative splicing events from sequencing data. In order to test and vali-

date our workflow we challenged it with publicly available sequencing data derived from human tissues, and conducted a comparison with eight alternative computational methods. In order to judge the performance of the different methods we constructed a benchmark data set of true positive splicing events across different tissues, agglomerated from public databases, and show that ARH-seq is an accurate, computationally fast and high-performing method for detecting differential splicing events.

### References

- Rasche A, Herwig R (2010) ARH: predicting splice variants from genome-wide data with modified entropy. *Bioinformatics* **26**(1), 84-90. <http://dx.doi.org/10.1093/bioinformatics/btp626>
- Rasche A, Lienhard M, Yaspo ML, Lehrach H, Herwig R (2014) ARH-seq: identification of differential splicing in RNA-seq data. *Nucleic Acids Res* **42**(14), e110. <http://dx.doi.org/10.1093/nar/gku495>

## iMir: an innovative and complete pipeline for smallRNA-Seq data analysis

Giorgio Giurato<sup>1</sup>, Antonio Rinaldi<sup>1</sup>, Adnan Hashim<sup>1</sup>, Giovanni Nassa<sup>1</sup>, Maria Ravo<sup>1</sup>, Francesca Rizzo<sup>1</sup>, Roberta Tarallo<sup>1</sup>, Angela Cordella<sup>2</sup>, Giovanna Marchese<sup>2</sup>, Domenico Memoli<sup>1</sup>, Alessandro Weisz<sup>1</sup>✉

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

<sup>2</sup>Genomix4Life Srl, Spin-Off of the University of Salerno, Baronissi, Italy

Giurato G *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e810. <http://dx.doi.org/10.14806/ej.21.A.810>

Next-generation sequencing allows researchers to gauge the depth and variation of small non-coding RNA populations, comprising miRNAs, piRNAs, tRNAs and other regulatory small transcripts. The accurate analysis of smallRNA-Seq data remain a non-trivial computational problem, requiring implementation of multiple statistical and bioinformatics tools. Here we present iMir (Giurato *et al.*, 2013), a modular pipeline for comprehensive analysis of smallRNA-Seq data, comprising specific tools for adapter trimming, quality filtering, differential expression analysis, biological target prediction and other useful options by integrating multiple open source modules and resources in an automated workflow (Figure 1).

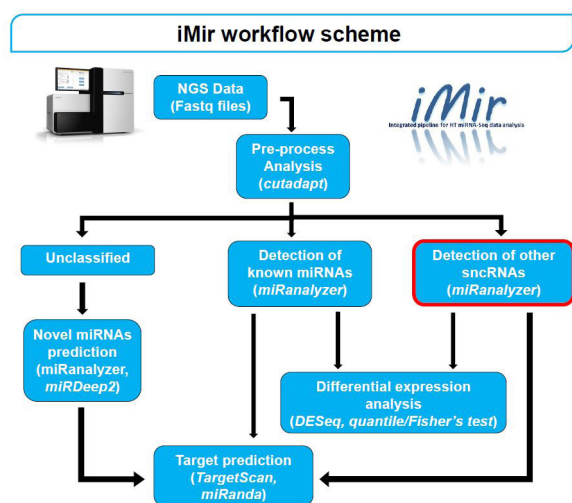


Figure 1. The pipeline accepts NGS data as input and then proceeds automatically to perform several independent analysis, most of them can be selected or excluded according to the user's needs.

iMir is based on reliable, flexible and fully automated workflow, allowing to rapidly and efficiently analyze high-throughput smallRNA-Seq data, such as those produced by the most

recent high-performance next generation sequencers. This pipeline allowed us to investigate piRNA expression patterns in rat liver and their modulation during regenerative proliferation (Rizzo *et al.*, 2014) and to identify >100 human piRNAs in breast cancer, some of which showing significant differences in expression in mammary epithelial compared to cancer cells or in normal respect to cancerous mammary tissues (Hashim *et al.*, 2014), and in endometrial hyperplasia and cancer (Ravo *et al.*, 2015).

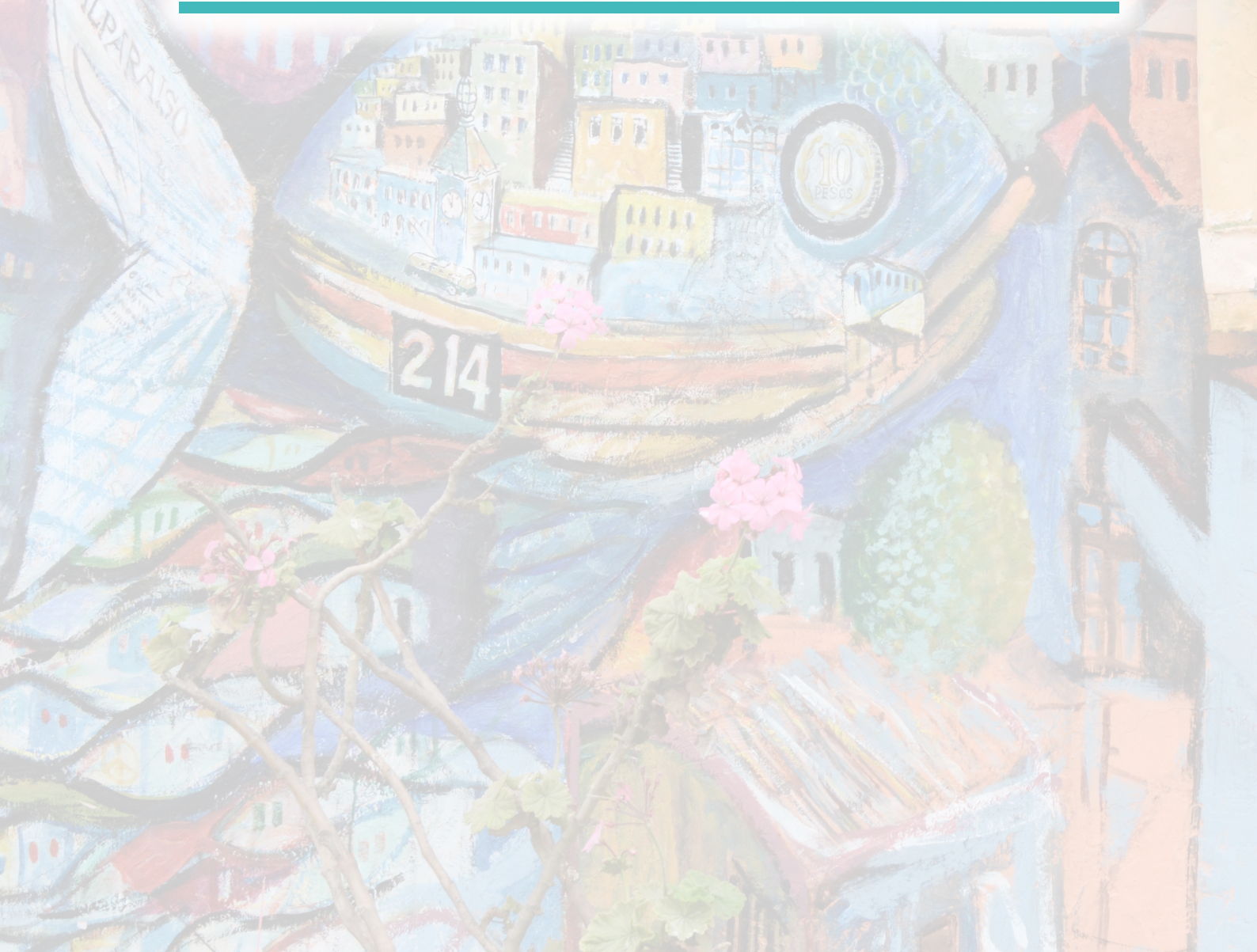
### Acknowledgements

Work supported by: Italian Ministry of Health (Grant Young Researcher GR-2011-02350476 to M.R.), Italian Ministry for Education, University and Research (Grants PRIN 2010LC747T to A.W. and FIRB RBF12W5V5\_003 to R.T.), Italian Association for Cancer Research (Grants IG 13176 to A.W.), National Research Council Flagship Project Interomics. G.N. is supported by a 'Mario e Valeria Rindi' fellowship of the Italian Foundation for Cancer Research, A.R. is a PhD student of the Research Doctorate 'Molecular and Translational Oncology and Innovative Medical-Surgical Technologies', University of Catanzaro 'Magna Graecia'.

### References

- Giurato G, De Filippo MR, Rinaldi A, Hashim A, Nassa G, *et al.* (2013) iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinformatics* **14**:362. <http://dx.doi.org/10.1186/1471-2105-14-362>
- Hashim A, Rizzo F, Marchese G, Ravo M, Tarallo R, *et al.* (2014) RNA sequencing identifies specific PIWI-interacting small non-coding RNA expression patterns in breast cancer. *Oncotarget* **5**(20), 9901-9910.
- Ravo M, Cordella A, Rinaldi A, Bruno G, Alexandrova E *et al.* (2015) Small non coding RNA deregulation in endometrial carcinogenesis. *Oncotarget* (Epub ahead of print).
- Rizzo F, Hashim A, Marchese G, Ravo M, Tarallo R *et al.* (2014) Timed regulation of P-element-induced wimpy testis-interacting RNA expression during rat liver regeneration. *Hepatology* **60**(3), 798-806. <http://dx.doi.org/10.1002/hep.27267>







## Microbial analysis of ovine cheese by next generation sequencing

Vladimir Kmet<sup>✉</sup>, Dobroslava Bujnakova

Institute of Animal Physiology, Slovak Academy of Sciences, Košice, Slovakia

Kmet V and Bujnakova D (2015) *EMBnet.journal* **21**(Suppl A), e806. <http://dx.doi.org/10.14806/ej.21.A.806>

Ovine cheese is a good source for isolation of wild lactobacilli-potential starter cultures. This study aimed to analyse microbiota of the ovine cheese (Slovak Bryndza) and to investigate the presence of *Lactobacillus* antibiotic resistance, virulence or probiotic genes by pyrosequencing.

The V1-V3 regions of the 16S ribosomal DNA were amplified from different ovine cheeses using PCR. In all samples, the microbial populations consisted of *Lactobacillus helveticus*, *Lb. acidophilus*, *Lb. plantarum*, *Lb. rhamnosus*, *Lb. brevis*; *Lactococcus lactis*, *L. raffinolactis*, *L. garviae*; *Enterococcus italicus* and *E. cameliae*; *Streptococcus salivarius*, *St. thermophilus*, *St. caballi* and *St. ferus*.

Furthermore, the genomes of selected *Lb. plantarum*, *Lb. brevis*, *Lb. paracasei* were pyrosequenced. The assembly of *L. plantarum* resulted in 203 contigs longer than 1,000 bp (D'Auria et

al., 2014). There were identified probiotic proteins as an alpha amylase (PF00128), peptidase (PF01433), catalase (PF00199), heat shock protein 33 (PF01430). Nevertheless, there was discrepancy between *Lb. plantarum* ampicillin sensitivity and the presence of serine beta-lactamase like superfamily (PF00144). No virulence factors were detected. Results indicated new properties of lactobacilli, which were not occurred by phenotyping testing.

### Acknowledgements

This work was supported by Slovak grant VEGA No. 2/0014/13.

### References

D'Auria G, Džunkova M, Moya A, Tomaška M, Kološta M, Kmet V (2014). Genome Sequence of *Lactobacillus plantarum* 19L3, a Strain Proposed as a Starter Culture for Slovenska Bryndza Ovine Cheese. *Genome Announc.* **2**(2):e00292-14. <http://dx.doi.org/10.1128/genomeA.00292-14>

## PACMAN: PacBio Methylation Analyzer

Laurent Falquet<sup>1,2</sup>✉, Alexis Loetscher<sup>1</sup>

<sup>1</sup>University of Fribourg, Fribourg, Switzerland

<sup>2</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland

Falquet L and Loetscher A (2015) *EMBnet.journal* **21**(Suppl A), e807. <http://dx.doi.org/10.14806/ej.21.A.807>.

Pacific Biosciences sequencing allows for the simultaneous detection of DNA methylation in particular m6A and m4C (Flusberg et al., 2010). The motifs around the methylation sites are provided by the [SMRT pipeline](#)<sup>1</sup> in GFF format. However the visualization of these motifs methylated or non-methylated, for example on a bacterial genome, is not part of the pipeline. Circos is a software package for producing publication quality images of large scale data (Krzywinski et al., 2009), however mastering the numerous configuration files needed, requires extra skills not easy to acquire for biologists. We developed a tool written in Perl and an associated web site called PACMAN (PacBio Methylation Analyzer) allowing users to easily create images with Circos in a user-friendly interface. The required files are 1) the genome or draft in FASTA format and 2) the

motifs file in GFF format (from the SMRT pipeline). The counts of each motif are calculated according to a customisable sliding window and normalised by their expected frequency. PACMAN generates a publication quality image of the selected methylated motifs counts, locations and non-methylated locations, on one or both strands of the DNA. PACMAN is available on this web site: <http://www.unifr.ch/bugfri/pacman>.

### References

- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465. <http://dx.doi.org/10.1038/nmeth.1459>.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R *et al.* (2009) Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645. <http://dx.doi.org/10.1101/gr.092759.109>.

<sup>1</sup> [www.pacb.com/devnet/](http://www.pacb.com/devnet/)



## NORM-SYS - harmonizing standardization processes for model and data exchange in systems biology

Babette Regierer<sup>1</sup>, Susanne Hollmann<sup>2</sup>, Martin Golebiewski<sup>3</sup>✉

<sup>1</sup>LifeGlimmer GmbH, Berlin, Germany

<sup>2</sup>University of Potsdam, Potsdam, Germany

<sup>3</sup>Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany

Regierer B *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e813. <http://dx.doi.org/10.14806/ej.21.A.813>

The rapid development of modern life science technologies, such as Next Generation Sequencing (NGS) techniques, allows the generation of biological data with increasing speed and precision. All these data have to be accessed, processed, integrated, shared, analysed and compared. Hence, standards for data, resulting computer models and applied workflows have become a critical issue specifically in distributed and interdisciplinary approaches like systems biology. Different stakeholder groups need to be engaged in the standardisation activities to allow an efficient and fast process of standardisation and adoption of the developed standards: researchers from both, academia and industries, with their existing grass-root standardisation communities; as well as the official standardisation bodies (e.g. DIN in Germany, CEN/CENELEC at European level, or ISO at international level) with


their long-standing professional experience in the formal standardization process; or representatives of scientific journals and research funding agencies.

[NORM-SYS](#)<sup>1</sup> (Normalization and standardization for the exchange of models and data in systems biology research) is a new project that started in October 2014 aiming at enhancing and promoting the formal normalisation of existing community standards for computational modelling in systems biology in close collaboration with relevant stakeholder groups and grass-root standardisation initiatives. One major goal of NORM-SYS is to develop a concept for the transformation of existing standards into certified specifications or norms to achieve a more effective transfer of systems biology research results into applications.

---

1 [normsys.de/](http://normsys.de/)

## Targeted amplicon sequencing in genetic diagnostics of patients with cystic fibrosis

Jelena Kusic-Tisma , Aleksandra Divac Rankov, Mila Ljubic, Nadja Pejanovic, Bojana Stanic, Dragica Radojkovic

Institute of Molecular Genetics and Genetic Engineering, Belgrade, Serbia

Kusic-Tisma J *et al* (2015) *EMBnet.journal* **21**(Suppl A), e814. <http://dx.doi.org/10.14806/ej.21.A.814>

Cystic fibrosis (CF), one of the most common autosomal recessive genetic disorders in Caucasians, is caused by mutations in the gene encoding the CF transmembrane conductance regulator (CFTR). Almost 2000 variants in CFTR gene with variable clinical significance have been identified so far.

In this study we performed targeted re-sequencing of CFTR gene in 24 CF patients in which one or no mutations were identified after analysis of seven most common variants (c.1521\_1523delCTT, c.489+1G>T, c.1624G>T, c.1652G>A, c.1657C>T, c.1585-1G>A, c.3909C>G) or after sequencing of the whole coding sequence of CFTR gene. Library pool was generated using multiplex PCR amplification (MASTR, Multiplicom) followed by sequencing on

a MiSeq instrument (Illumina). Sequencing data were evaluated using the software Sequence Pilot (JSI Medical Systems). Identification of disease-relevant CFTR variants was assessed based on the [CFTR database](#)<sup>1</sup> and [CFTR2 website](#)<sup>2</sup>.

The NGS sequencing data correctly confirmed 18 germline variants previously detected in our laboratory. Four out of 16 additionally identified variants were classified as disease-causing mutations according to the literature data, thus enabling us confirmation of clinical CF diagnosis in three patients. The NGS technology in combination with a well-characterised clinically relevant genomic variation database is a good alternative for a time consuming stepwise testing of genes with large allelic heterogeneity such as CFTR.

1 [www.genet.sickkids.on.ca/app](http://www.genet.sickkids.on.ca/app)

2 [www.cfr2.org/](http://www.cfr2.org/)

## Reference leaf transcriptomes for potato cultivars: Desiree and PW363

Maja Zagorščak<sup>1</sup>✉, Marko Petek<sup>1</sup>, Mohamed Zouine<sup>2</sup>, Kristina Gruden<sup>1</sup>

<sup>1</sup>National Institute of Biology - Department of Biotechnology and Systems Biology, Ljubljana, Slovenia

<sup>2</sup>Ecole Nationale Supérieure Agronomique de Toulouse, Toulouse, France

Zagorščak M *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e817. <http://dx.doi.org/10.14806/ej.21.A.817>.

The rapid development of modern life science technologies, such as Next Generation Sequencing (NGS) techniques, allows the generation of biological data with increasing speed and precision. Most potato cultivars are highly heterozygous tetraploids with high genetic variability while being susceptible to pathogens, pests and inbreeding depression. To bypass polyploidy related sequencing problems, *Potato Genome Sequencing Consortium* (PGSC, 2011) sequenced a double monoploid derived from *S. tuberosum* group Phureja.

In order to avoid problems, such as discriminating between paralogous genes, divergence and expression bias between the reference genome and potato cultivars, and to identify traits that are not present in initially sequenced genotype, RNA-sequencing for cv. Desiree and cv. PW363 leaves was conducted on Illumina NGS platform. In house generated raw reads were complemented with data already deposited in the NCBI database and stNIB-v1 *S. tuberosum* gene groups, which included two genome assemblies and two EST datasets (Ramšak *et al.*, 2014). The preliminary transcriptomes were assembled using both hybrid and de novo assembly approaches. The hybrid approach, combining genome-guided and de novo RNA-Seq assembly, was implemented using the pipeline

available from [CLC Genomics Workbench 7.0.3](#)<sup>1</sup>. De novo assembly was performed using Trinity (Grabherr *et al.*, 2011).

The resulting two sets of preliminary transcriptomes were then merged using the CD-HIT clustering algorithm (Limin *et al.*, 2012) and merged with existing gene models with BLAST against stNIB-v1. The presumed novel clusters were annotated using SwissProt, PGSC DM v3.4 super-scaffolds and NCBI-nt databases. Initial potato pangenome containing 35609 genes was expanded with 24999 potential new transcripts, and will serve to further expand knowledge on the potato pathogen interactions.

### References

- Limin F, Beifang N, Zhengwei Z, Sitao W, Weizhong L (2012) CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* **28**(23), 3150-3152. <http://dx.doi.org/10.1093/bioinformatics/bts565>.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* **29**(7), 644-652. <http://dx.doi.org/10.1038/nbt.1883>.
- Ramšak Ž, Baebler Š, Rotter A, Korbar M, Mozetič I, *et al.* (2014) GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology. *Nucleic Acids Research* **42**(D1), D1167-D1175. <http://dx.doi.org/10.1093/nar/gkt1056>.
- PGSC: The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195. <http://dx.doi.org/10.1038/nature10158>.

<sup>1</sup> [www.clcbio.com](http://www.clcbio.com)



## NGS as a tool for investigating spread of zoonotic bacteria: dealing with source and outbreak genetic variation in a spatio-temporal context

Robert Söderlund<sup>1,2✉</sup>, Tomas Jinnerof<sup>2</sup>, Adrien Janssens<sup>1</sup>, Erik Bongcam-Rudloff<sup>1</sup>

<sup>1</sup>Swedish University of Agricultural Sciences, Uppsala, Sweden

<sup>2</sup>National Veterinary Institute, Sweden

Söderlund R *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e818. <http://dx.doi.org/10.14806/ej.21.A.818>

Next generation sequencing (NGS) is increasingly the method of choice for tracing transmission of zoonotic bacterial disease agents; infections that are spread to humans from animals. Establishing the link to a source of infection is dependent on an implicit or explicit definition of a source population from which the outbreak strain originates. The link is also based on assumptions about the relevance of a sampling occasion for the time the infection event took place.

While these problems are the same regardless of the method used, they are magnified by the high resolution of NGS-based typing. In particular, exact genotypic matches are rare leading to a much greater challenge in comparing genetic distances, drawing conclusions from them and communicating the quantitative results to a non-specialist audience.

We are currently conducting a series of studies to better define the source of genetic variation found in bacterial populations in animal reservoirs and animal feed, as well as the variation found between bacterial isolates collected from a single sampling occasion. We are also studying the spatial distribution of endemic clones in regions with dense livestock populations, and the clonal dynamics in such regions. Our focus is on enterohaemorrhagic *E. coli* (EHEC) and selected serotypes of *Salmonella*. The results are providing us with insights that will improve future outbreak investigations by guiding the sampling and analysis strategy as well as the interpretation of data.

### Acknowledgements

This work was financed by the Swedish Board of Agriculture and the Elsa and Ivar Sandberg Foundation.

## Biobanks and future emerging technologies: new approaches, new pre-analytical challenges

Eva Ortega-Paino<sup>1</sup>✉, Tomas Klingström<sup>2</sup>, Johanna Ekström<sup>1</sup>

<sup>1</sup>BBMRI.se Service Center for Southern Sweden, Medicion Village (406), Lund University, Lund, Sweden

<sup>2</sup>SLU Global Bioinformatics Centre, Uppsala University, Uppsala, Sweden

Ortega-Paino E *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e822. <http://dx.doi.org/10.14806/ej.21.A.822>

Establishing a Biobank is a major long-term commitment and samples collected today must be relevant for next generation techniques far into the future.

It is therefore crucial for biobanks to be “future compatible” and rely on sampling and storage methods to maximize the future value of the collections. In its most basic form such future compatibility may be limited to careful sample management such as a strict adherence to quality management and best practices as laid down by authoritative sources such as ISBER (International Society for Biological and Environmental Repositories, 2012) guidelines and BRISQ (Moore *et al.*, 2011; 2012; 2013), among others.

This approach will be sufficient, if adhered to, for most current generation sequencing techniques where read-lengths are limited to 150-500 bp and if biobank samples intended for sequencing carry an abundance of RNA or DNA. But single cell sequencing of circulating tumor cells, DNA methylation analysis and proteogenomic studies, where DNA, RNA and protein from the same sample is analysed (Nesvizhskii, 2014), require new standards for sample collection and storage.

Therefore, and looking at the future, it would be very desirable to run studies similar to SPIDIA-RNA (Malentacchi *et al.*, 2014) and SPIDIA-DNA (Malentacchi, 2013) to ensure that samples can be used for proteogenomics and other applica-

tions where proteomics and nucleic acid based assays are combined. But to achieve this it is necessary to create biobank cohorts with sufficient collection, quality and storage conditions.

### Acknowledgements

Funding for this work has been provided by BBMRI.se.

### References

- International Society for Biological and Environmental Repositories (2012) 2012 Best Practices for Repositories: Collection, Storage, Retrieval, and Distribution of Biological Materials for Research. *Biopreservation and Biobanking* **10**(2), 81-161. <http://dx.doi.org/10.1089/bio.2012.1022>
- Malentacchi F, Pazzagli M, Simi L, Orlando C. *et al.* (2014) SPIDIA-RNA: second external quality assessment for the pre-analytical phase of blood samples used for RNA based analyses. *PLoS One* **9**(11):e112293. <http://dx.doi.org/10.1371/journal.pone.0112293>
- Malentacchi F, Pazzagli M, Simi L, Orlando C, Wyrich R *et al.* (2013) SPIDIA-DNA: an External Quality Assessment for the pre-analytical phase of blood samples used for DNA-based analyses. *Clin Chim Acta*. **424**, 274-286. <http://dx.doi.org/10.1016/j.cca.2013.05.012>
- Moore HM, Kelly A, Jewel SD, McShane LM *et al.* (2011) Biospecimen reporting for improved study quality (BRISQ). *J Proteome Res*. **10**(8), 3429-38. <http://dx.doi.org/10.1021/pr200021n>
- Moore HM, Kelly A, McShane LM, Vaught J (2012) Biospecimen reporting for improved study quality (BRISQ). *Clin Chim Acta*. **413**(15-16), 1305. <http://dx.doi.org/10.1016/j.cca.2012.04.013>
- Moore HM, Kelly A, McShane LM, Vaught J (2013) Biospecimen reporting for improved study quality (BRISQ). *Transfusion* **53**(7):e1. <http://dx.doi.org/10.1111/trf.12281>
- Nesvizhskii AI (2014) Proteogenomics: concepts, applications and computational strategies. *Nature Methods* **11**(11), 1114-1125. <http://dx.doi.org/10.1038/nmeth.3144>

## MetLab: an in silico experimental design, simulation and validation tool for viral metagenomics studies

Martin Norling<sup>1</sup>, Oskar E. Karlsson<sup>2</sup>, Hadrien Gourlé<sup>2</sup>, Erik Bongcam-Rudloff<sup>2</sup>, Juliette Hayer<sup>2</sup>✉

<sup>1</sup>Bioinformatics Infrastructure for Life Sciences, Uppsala, Sweden

<sup>2</sup>SLU Global Bioinformatics Center, Department of Animal Breeding and Genetics (HGEN), SLU, Uppsala, Sweden

Norling M *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e823. <http://dx.doi.org/10.14806/ej.21.A.823>.

The use of metagenomics for detection and characterisation of viruses has several difficult steps, including high throughput sequencing and bioinformatics analysis. We present MetLab, a new program aimed at providing scientists with a simple baseline for experimental design and analysis of viral metagenomics.

MetLab aims to provide support in planning the sequencing, by estimating coverage by implementing an adaptation of Stevens' theorem (Wendl *et al.* 2013). It also provides scientists with several pipelines aimed at simplifying the analysis of viral metagenomes, including quality control, assembly and taxonomic binning. We also implement a tool for simulating metagenomics datasets from a number of sequencing platforms. The overall aim is to provide viro-

gists within veterinary medicine with an easy to use tool for designing, simulating and analyzing viral metagenomes. The results presented here include a comprehensive benchmark towards other suitable software, with emphasis on detection of viruses as well as speed of applications. MetLab is packaged as one comprehensive software package, readily available for Linux and OSX users.

### Acknowledgements

Funding: Formas project number 2012-586.

### References

Wendl MC, Kota K, Weinstock GM and Mitreva M (2013) Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. *J. Math. Biol.* **67**(5), 1141–1161. <http://dx.doi.org/10.1007/s00285-012-0586-x>



## Towards SNP calling in polyploid genomes

Anatoliy Dimitrov<sup>1</sup>✉, Milko Krachunov<sup>1</sup>, Ognyan Kulev<sup>1</sup>, Jérôme Salse<sup>2</sup>, Irena Avdjieva<sup>3</sup>, Dimitar Vassilev<sup>3</sup>

<sup>1</sup>Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

<sup>2</sup>INRA-UBP GDEC, Clermont-Ferrand, France

<sup>3</sup>AgroBioInstitute, Bioinformatics group, Sofia, Bulgaria

Dimitrov *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e826. <http://dx.doi.org/10.14806/ej.21.A.826>

Polyploid genomes such as hexaploid wheat pose significant study challenges. The detection of SNPs is one of the problems that are difficult to solve. The existence of multiple sub-genomes introduces a significant amount of natural noise for the SNP calling algorithms, and the expected occurrence rates of such SNPs is significantly lower than in diploid genomes. This makes them more difficult to distinguish from errors.

Using the assumption that, unlike SNPs, errors are randomly distributed across sub-genomes, we try to reduce the set of predicted errors by excluding rare bases that are clustered together in reads with high probability of coming from a distinct sub-genome. This is accomplished through the application of fuzzy clustering that is based

on the amount of observed variation between the pairs of aligned reads. A secondary filter, relying on machine learning, is also proposed. This is expected to lead to an increase in the number of correctly predicted SNPs that would have been otherwise incorrectly identified as errors. A 56% reduction in the predicted errors, and a 66% increase in the number of predicted SNPs are observed.

### Acknowledgements

The study is supported by the National Science Fund of Bulgaria within the "Methods for Data Analysis and Knowledge Discovery in Big Sequencing Datasets" project under contract DFNI-I02/7 of 12.12.2014.

## Gene set enrichment analysis of neuroendocrine system of the silkworm *Bombyx mori*

Gabor Beke<sup>1</sup>✉, Matej Stano<sup>1</sup>, Ivana Daubnerova<sup>2</sup>, Dusan Zitnan<sup>2</sup>, Lubos Klucar<sup>1</sup>

<sup>1</sup>Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia

<sup>2</sup>Institute of Zoology, Slovak Academy of Sciences, Bratislava, Slovakia

Beke G *et al.* (2015) *EMBnet.journal* **21**(Suppl A), e829. <http://dx.doi.org/10.14806/ej.21.A.829>

Silkworm *Bombyx mori* represents a model organism for studying the neuroendocrine system in invertebrates. This system of nervous organs and endocrine glands regulates large number of life functions including movement, ecdysis, courting and mating. We have analysed strand-specific Illumina RNA-seq data from *B. mori* samples originated from different endocrine glands of both sexes and several developmental stages. Reads were mapped to the *B. mori* transcriptome using the *Bowtie 2* aligner. Since the silkworm genome is abundant in repetitive sequences (Mita *et al.*, 2004), the mapping allowing unlimited multi-mappings (required by *eXpress*) was extremely computing demanding. Transcript level RNA-seq quantification was performed using the *eXpress*

tool based on an online expectation-maximization algorithm. Obtained data were consequently analysed in several gene set enrichment packages, including *topGO* and *gplots* (both *R* packages), where highly expressed genes were clustered and visualised, according to the functional GO enrichment analysis, and clustered by gene expression level.

### Acknowledgements

The study is supported by the APVV-0827-11, VEGA 2/0164/15, IMTS 26230120002 and IMTS 26210120002 grants.

### References

Mita K *et al.* (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res.* **11**(1), 27-35. <http://dx.doi.org/10.1093/dnares/11.1.27>

## The NonCode aReNA DB: a non-redundant and integrated collection of non-coding RNAs

Giorgio De Caro<sup>1</sup>, Arianna Consiglio<sup>1</sup>, Domenica D'Elia<sup>1✉</sup>, Andreas Gisel<sup>1,2</sup>, Giorgio Grillo<sup>1</sup>, Sabino Liuni<sup>1</sup>, Angelica Tulipano<sup>1</sup>, Flavio Licciulli<sup>1</sup>

<sup>1</sup>CNR, Institute for Biomedical Technologies, Bari, Italy

<sup>2</sup>International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria

De Caro G *et al.* (2015) *EMBnet:journal* **21**(Suppl A), e834. <http://dx.doi.org/10.14806/ej.21.A.834>

The recent availability of high throughput technologies, like next generation sequencing (NGS) platforms, has provided the scientific community with an unprecedented opportunity for large-scale analysis of genome in a large number of organisms. However, among others, one of the most challenging task for bioinformaticians is to develop tools that provide biologists with an easy access to curated and non-redundant collections of sequence data.

Non-coding RNAs, for a long time believed to be not-functional, are emerging as the most large and important family of gene regulators. NonCode aReNA Database is a comprehensive and non-redundant source of manually curated and automatically annotated ncRNA transcripts. Originally developed as a component of a bigger project, composed by a datawarehouse for the functional annotation of ncRNAs from NGS data, NonCode aReNA DB is currently available

as a web-resource at <http://ncrnadb.ba.itb.cnr.it/>. Sequences have been classified in diverse biotypes and associated to Sequence Ontology terms. The database can be queried by using multi-criteria and ontological search through an easy-to-use web interface, and data exported as non-redundant collections of transcripts annotated in VEGA, ENSEMBL, RefSeq, miRBase, GtRNAdb and piRNABank. The database is updated through an automatic pipeline and last update was on January 2015. Presently NonCode aReNA DB contains 134,908 human ncRNAs classified in 24 biotypes, and next update will include transcripts of *Mus musculus* and *Arabidopsis thaliana*.

### Acknowledgements

This work was supported by the Italian MIUR Flagship Project "Epigen".



## Organisational Members of EMBnet

### Biocomputing Group

Belozersky Institute, Moscow, Russia

### BMC

Uppsala Biomedical Centre, Computing Department, Uppsala, Sweden

### Centre of Bioinformatics

Peking University, Beijing, China

### CMBI

Radboud University, Nijmegen, The Netherlands

### CNR

Institute for Biomedical Technologies, Bioinformatics and Genomic Group, Bari, Italy

### CSC

Espoo, Finland

### EMBL-EBI

Hinxton, Cambridge, United Kingdom

### Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires, Buenos Aires, Argentina

### Institute of Biochemistry

Molecular Biology and Biotechnology, University of Colombo, Colombo, Sri Lanka

### Instituto de Biotecnología

Universidad Nacional de Colombia, Edificio Manuel Ancizar, Bogotá, Colombia

### Instituto Gulbenkian de Ciencia

Centro Portugues de Bioinformatica, Oeiras, Portugal

### ITICO

Information Technology Infrastructure for Collaborative Organizations, United Kingdom

### KEMRI

Wellcome Trust Research Programme, Kilifi, Kenya

### Lab. Nacional de Computação Científica

Lab. de Bioinformática, Petrópolis, Rio de Janeiro, Brazil

### LCSB

University of Luxembourg, Luxembourg, Luxembourg

### ReNaBi

French bioinformatics platforms network, France

### SIB

Swiss Institute of Bioinformatics, Lausanne, Switzerland

### TGAC

The Genome Analysis Centre, Norwich, United Kingdom

### UMB SAV

Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia

### UMBER

Faculty of Life Sciences, The University of Manchester, Manchester, United Kingdom

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: +46-18-67 21 21