EMBnet.journal
Volume 20
Supplement A
December 2014

NGS Data after the Gold Rush

COST Action BM1006 Scientific Meeting 2014
6-8 May 2014
Norwich, United Kingdom

cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

EUROPEAN SCIENCE FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

ESF provides the Cost Office through an EC contract

COST is supported by the EU RTD Framework Programme

Editorial

Next Generation Sequencing (NGS) is today driving the generation of knowledge (especially in biomedicine and molecular life sciences) to new dimensions. The massive data volumes being generated by these new technologies require new data handling and storage methods. Hence, the life science community needs not only new and improved approaches to facilitate NGS data management and analysis but also hundreds if not thousands of scientists trained in the usage and analysis of these new technologies. In this EMBnet.journal supplement the COST Action Seqahed presents some insights in the work that unites bioinformaticians, computer scientists and biomedical scientists allowing them all together to bring NGS data management and analysis to new levels of efficiency and integration.

EMBnet.journal Editorial Board

Contents

Cover picture: Moon and clouds, Uppsala, 2014
[© Erik Bongcam-Rudloff]

Scientific Meeting 2014 "NGS Data after the Gold Rush" & "Management Committee" Meeting, Norwich, UK


Vicky Schneider[1], Erik Bongcam-Rudloff[2]✉

[1]The Genome Analysis Centre (TGAC), Norwich, United Kingdom
[2]Swedish University of Agricultural Sciences, Uppsala, Sweden

Schneider V and Bongcam-Rudloff E (2014) *EMBnet.journal* **20**(Suppl A), e794. http://dx.doi.org/10.14806/ej.20.A.794

With 70 stakeholders from 28 European countries, the three-day meeting at the The Genome Analysis Centre (TGAC) facilities in Norwich (UK) explored the state-of-the-art in Next Generation Sequencing (NGS) data analysis, its current challenges and applications.

The event was opened by organiser, Dr. Vicky Schneider, who highlighted the importance of being able to host and make the most of this networking opportunity at TGAC. The first session was chaired by Dr. Ana Conesa, from the Prince Felipe Research Centre (ES), and Aleksandra Pawlik, from the Software Sustainability Institute (UK).

Figure 1. Audience.

During the first day of the event, participants heard presentations from leading international scientists, covering data analysis and management, bioinformatics training, functional and pathogen genomics.

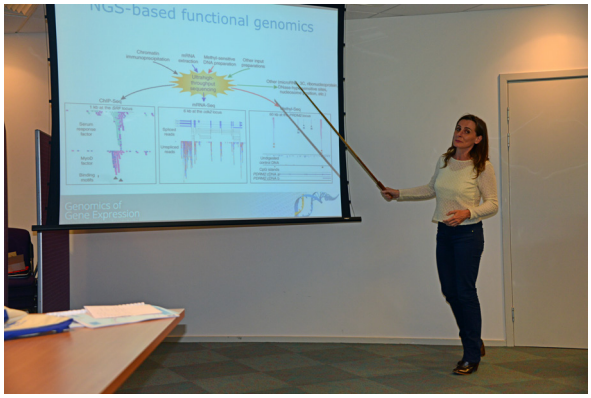

Figure 2. Ana Conesa.

The participants and speakers discussed the importance and value of the SeqAhead-COST Action[1] in enabling the development and coordination of a common action plan for the benefit of the scientific community, in assisting with the rise of NGS data, and in the development and optimal use of state-of-the-art bioinformatics.

The establishment of a strong European network of NGS data-analysis and informatics scientists was underlined as a key facilitator and stimulator towards the exchange of data, protocols, software, experiences and ideas.

The scientific meeting also discussed current and future directions in terms of the use of NGS technologies, further developments of high-throughput approaches, with a particular focus on parallelisation of the sequencing process, producing thousands or millions of concurrent DNA sequences. The large data volumes being generated by these new technologies require

1 seqahead.eu

Figure 3. Ola Spjuth.

new data-handling and storage methods: solutions created and implemented by the speakers were shared and discussed, with emphasis on the critical role and need for training life scientists to make the most of existing solutions, and ensuring best practice in experimental design, in data quality control, analysis and storage, and in facilitating interoperability of data-sets and workflows.

The discussions during the scientific meeting highlighted the importance and need for opportunities to distribute knowledge and expertise via combined education and publication programs, such as the activities supported by this COST action.


Figure 4. Dave Clements, Galaxy.

All of these activities help to unite bioinformaticians, computer scientists and biomedical scientists, harnessing their expertise to bring NGS data management and analysis to new levels of efficiency and integration.

The SeqAhead Management Committee meeting concluded the event, led by Chairman, Erik Bongcam-Rudloff, and Vice-Chair, Terri Attwood


Figure 5. Group picture.

Speakers were happy to share their presentations, which can be found at: http://www.tgac.ac.uk/SeqAhead_Scientific_Meeting_and_Management_Committee_Meeting/

Some voices from the event:

Matt Clark, Plant and Microbial Genomics Group Leader, said, "NGS is a large, growing and rapidly moving field, it's critical to keep your eye on the latest technical breakthroughs and best practices both in the lab and in bioinformatics to do the best possible research. Meetings such as this are critical to sharing knowledge across the EU and to ensure we learn from each other rapidly."

Terri Atwood, Professor of Bioinformatics at the University of Manchester and Vice-Chair of SeqAhead, said: "Networks like SeqAhead are vital both for allowing scientists to share best practice in the development and use of cutting-edge research tools, and for helping to train the 'next generation' of next generation sequencing researchers!"

Chairs and Conference Committees

Scientific Organisers

Vicky Schneider, The Genome Analysis Centre (TGAC), UK
Erik Bongcam-Rudloff, Swedish University of Agricultural Sciences in Uppsala, Sweden
Ralf Herwig, Max Planck Institute for Molecular Genetics, Germany
Thomas Svensson, Karolinska Institute, Sweden
Andreas Gisel, Institute for Biomedical Tecnologies, Italy
Ana Conesa, Prince Felipe Research Centre, Spain
Eija Korpelainen, CSC, Finland
Steve Pettiffer, University of Manchester, UK
Veli Makinen, University of Helsinki, Finland
Alberto Policriti, University of Udine, Italy
Gert Vriend, Netherlands Bioinformatics Centre, Netherlands
Jacques van Helden, University of Brussels, Belgium

Event Organiser

Matt Drew, The Genome Analysis Centre (TGAC), UK

Scientific Committee

Vicky Schneider, The Genome Analysis Centre (TGAC), UK
Jean Imbert, Technological Advances for Genomics and Clinics, France
Pall Melsted, University of Iceland
Pasha Baranov, University College Cork, Ireland
Ana Conesa, Prince Felipe Research Centre, Spain
Pavlos Antoniou, Cyprus Insistute of Neurology and Genetics, Cyprus
Sophia Kossida, Biomedical Research Foundation of the Academy of Athens, Greece
Jose Valverde, CNB/CSIC, Spain
Endre Barta, Agricultural Biotechnology Centre, Hungary
Antonio Marco, University of Essex, UK
Rute Fonseca, University of Copenhagen, Denmark
Claude Muller, Institute of Immunology, Luxembourg
Manolis Christodoulakis, University of Cyprus

Scientific Programme

NGS Data after the Gold Rush

6-8 May 2014, Norwich, United Kingdom

Conference Programme

Day 1	NGS: State of the Art, Current Challenges & Applications Chair: Ana Conesa
08:30-09:00	Registration and Coffee
09:00-09:30	Welcome from **Vicky Schneider & Dr Erik Bongcam-Rudloff**
09:30-10:00	Plenary Lecture: "NGS, then, now and in the future", **Matt Clark**
10:00-10:30	Coffee Break & poster set-up
10:30-11:00	SeqAhead: NGS data analysis network across countries and projects, **Erik Bongcam-Rudloff**
11:00-11:30	Using NGS to answer biological questions, **Björn Usadel**
11:30-12:00	NGS scaling up: A retrospective from the Norwegian Sequencing Centre, **Robert Lyle**
12:00-13:00	Lunch
13:00-13:30	NGS for studying viruses "beyond the consensus", **Jan Kim**
13:30-14:00	Towards sustainability in Bioinformatics training, **Judit Kumuthini**
14:00-14:15	Selected oral presentation: Added value of whole-genome sequence data to genomic predictions, **Marco Bink**
14:15-14:30	Selected oral presentation: Molecular Signalling in interaction between Potato, Potato virus Y and colorado Potato beetle, **Kristina Gruden**
14:45-15:30	Coffee Break and Poster Presentations
15:30-16:00	Training in computational skills, **Aleksandra Pawlik**
16:00-16:30	NGS in Functional Genomics: Where are we now and its impacts, **Ana Conesa**

Day 2	NGS: State of the Art, Current Challenges & Applications Chair: Aleksandra Pawlik
09:00-09:30	Morning Coffee
09:30-10:00	Coding & Best Practice in Programming: Why it matters so much in the NGS era, **Lex Nederbragt**
10:00-10:30	GWAS, where are we now?, **Ashley Farlow**
10:30-11:00	Scaling Galaxy for Big Data, **Dave Clements**
11:00-11:30	Coffee Break & poster set-up
11:30-12:00	Sex, deep sequencing and microRNAs, **Antonio Marco**
12:00-13:00	Lunch
13:00-13:15	Selected Oral Presentation: Transcription microvariability detection by NGS, **Fleur Leenen**
13:15-13:30	Selected oral presentation: From BIG data to RELEVANT data: Application of ribo-seq technology for understanding gene expression, **Pavel Baranov**
13:30-13:45	Selected oral presentation: KmerStream: a Streaming Algorithms for k-mer Abundance Estimation, **Pall Melsted**
13:45-14:15	NGS data management and analysis for hundreds of projects: Experiences from Sweden, **Ola Spujuth**
14:15-15:00	Coffee Break
15:00-15:15	Selected oral presentation: High-throughput Sequencing of the Immunoglobulin Heavy Chain Repertoire of Transgenic Humanized Rats Reveals Convergent Antibody Signatures, **Claude Muller**
15:15-15:30	Selected oral presentation: Analyzing RNA-seq data with RNASeqHUI, **Claudia Angelini**
15:30-16:00	NGS in Pathogen Genomics: It's impact and applications, **Claudio Donati**
16:00-16:30	Impact and sustainability of hands-on training in the analysis of NGS data, **Gabriella Rustici**
16:30-17:15	Closing remarks & What's Next, **Erik Bongcam-Rudloff**

Day 3	SeqAhead Management Committee Meeting
08:30-09:00	Morning Coffee
09:00-09:15	Welcome, Introduction and overview of the day, **Vicky Schneider**
09:15-09:30	Approval of minutes and matters arising from last meeting
09:30-10:00	Update from the Action Chair Status of Action, including participating countries
10:00-10:20	Promotion of gender balance and of Early Stage Researchers (ESR) STSM status and new applications
10:20-10:30	Annual Progress Conference in Malta
10:30-11:00	Coffee Break
11:00-11:40	Follow up of MoU objectives a. WG reports (5 mins each) b. Progress report on the third years workshops (5 mins) c. Publication of State-of-the-art Reports (5 mins) d. Maintaining lists of publications and projects on the website (5 mins)
11:40-12:50	Scientific Planning for last year of the action a. Scientific strategy: WG activity during the last year b. Action Budget Planning c. Long-term planning (including anticipated locations and dates of future activities) d. Dissemination planning (publications and outreach activities)
12:50-13:00	Summary of Management Committee Decisions, close meeting and confirm location and date of next meeting
13:00-14:00	Lunch

Keynote Lectures

Scaling Galaxy for Big Data

Dave Clements, Galaxy Team
Galaxy Project, Johns Hopkins University, Baltimore, USA
Clements D (2014) *EMBnet.journal* **20**(Suppl A), e764. http://dx.doi.org/10.14806/ej.20.A.764

Galaxy[1] is a widely-used, web-based platform for data integration and analysis in the life sciences (Goecks *et al.*, 2010; Blenkenberg *et al.*, 2010; Giardine *et al.*, 2005). It is available as a free public server[2] on the web, and as open-source software that can be installed locally and on the cloud[3]. Galaxy enables life scientists to perform bioinformatics analysis using the large and varied datasets now being generated in biomedical research. It does this without requiring researchers to learn Linux system management, scripting, or command line interfaces.

In addition to making these methods accessible to bench researchers, Galaxy also enables sharing, reproducibility and transparency in research. Galaxy features a robust history mechanism that automatically and unobtrusively records all data, metadata, and analysis steps, allowing the analysis to be shared and published, and run again with the same or different data. The platform also supports creation of reusable pipelines, either *de novo*, or by extracting them from existing analyses.

This talk will introduce Galaxy and then focus on what the project is doing to scale to support complex analysis in experiments with hundreds or even thousands of samples and datasets. It also includes a discussion on the challenges faced, and how they are being addressed

Acknowledgements

The Galaxy Project is an open source project with core team members on three continents, and across the United States. It has a very active community, contributing support, code, documentations, tools, and training back to the project. Principal Investigators are located at Penn State University, Johns Hopkins University, and George Washington University. The Galaxy Project is primarily funded through NIH NHGRI grant 5U41HG006620.

Dave Clements would like to thank The Genome Analysis Center for making it possible for him to present at this meeting.

References

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **89**: 19.10:19.10.1–19.10.21. http://dx.doi.org/10.1002/0471142727.mb1910s89

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**(10): 1451-1455. http://dx.doi.org/10.1101/gr.4086505

Goecks, J, Nekrutenko A, Taylor, J, and The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**(8): R86. http://dx.doi.org/10.1186/gb-2010-11-8-r86

1 galaxyproject.org
2 usegalaxy.org
3 getgalaxy.org

The STATegra project: new statistical tools for analysis and integration of diverse omics data

Ana Conesa (on behalf of the STATegra consortium)

Scientific coordinator of the FP7 STATegra project, Genomics of Gene expression Lab, Prince Felipe Research Center, Valencia, Spain

Conesa A (2014) *EMBnet.journal* **20**(Suppl A), e768. http://dx.doi.org/10.14806/ej.20.A.768

Next generation sequencing has speed up genome analysis and brought omics research closer to many organisms and biological scenarios. Today an increasing number of research projects propose the combined use of different omics platforms to investigate diverse aspects of genome functioning. These proposals ideally seek to provide complementary sources of molecular information that eventually can be put together to obtain systems biology models of biological processes. Hence, it is not rare anymore to find experimental designs involving the collection of genome, transcriptome, epigenome and even metabolome data on a particular system. However, standard methodologies for the integration of diverse omics data types are not yet ready and researchers frequently face post-experiment question on how to combine data of different nature, variability, and significance into an analysis routine that sheds more light than the analysis of individual datasets separately. The STATegra project[1] has been conceived to address these problems and provide the genomics community with user-friendly tools for the integration of different omics data types. STATegra targets several sequencing based functional genomics methods, proteomics and metabolomics. In this presentation I will report about current results of the project that include the STATegraEMS, an experiment management system for storage and annotation of complex omics experiments, novel data integration visualisation tools, statistical approaches to integrate RNA-seq data with different regulators of gene expression, transcriptomics measurements combined with downstream features such as proteomics and metabolomics, and data mining strategies to leverage public domain datasets in the integrative effort. I will also present the STATegRa, a new Bioconductor R package for integrative omics data analysis.

1 stategra.eu

GWAS - where are we now?

Ashley Farlow
Gregor Mendel Institute of Molecular Plant Biology, Vienna, Austria
Farlow A (2014) *EMBnet.journal* **20**(Suppl A), e782. http://dx.doi.org/10.14806/ej.20.A.782

Genome-Wide Association Studies (GWAS) are a powerful tool for establishing correlation between phenotype and genotype. For the self-fertilising plant *Arabidopsis thaliana*, more than 1000 inbred lines have been 'fully sequenced', removing the cost of genotyping for a set of lines that can be phenotyped over and over. Does having full sequence make a difference? How important is sample size and line selection? The answers from *Arabidopsis* are that it is highly dependent on trait architecture and population structure. This offers an important insight into the fundamental advantages and limitation of GWAS.

I will also discuss how Next Generation Sequencing (NGS) data allows one to explore a number of 'genomic traits' such as genome size and centromere length, and how GWAS can be used to follow the fate of new centromeric repeats in the population.

NGS for Studying Viruses "Beyond the Consensus"

Jan T. Kim
The Pirbright Institute, Pirbright, United Kingdom
Kim JT (2014) *EMBnet.journal* **20**(Suppl A), e774. http://dx.doi.org/10.14806/ej.20.A.774

High mutation rates in viruses (especially RNA viruses) have profound consequences on viral evolution, including the formation of quasispecies. These have long been studied in theory (Eigen, 1971) and in silico (Wilke *et al.*, 2001). NGS technologies provide new opportunities to directly observe sequence diversity and its evolution in viruses (Wright *et al.*, 2011) and other systems (Schütze *et al.*, 2011).

Profiles of base frequencies can be constructed from virus NGS data. They are used in a number of well established bioinformatics contexts, including DNA binding site representation (Stormo, 2000) and progressive multiple alignment (Larkin *et al.*, 2007). Profiles can be formalised as elements of a continuous sequence space (Vingron and Sibbald, 1993), and they serve as a basis for information theoretic analyses (Schneider *et al.*, 1986; Kim *et al.*, 2003) and statistical learning approaches (Kim *et al.*, 2004).

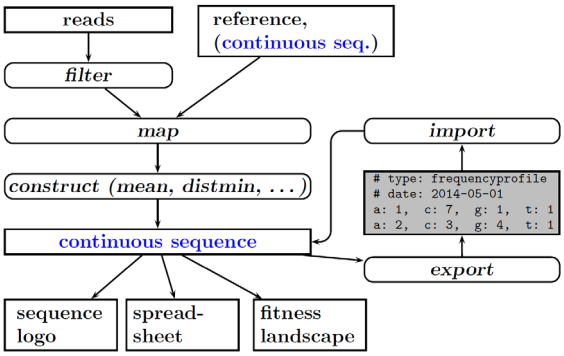

Figure 1. Continuous sequences as a point of departure for many bioinformatic analyses.

Given this basis, profiles provide a point of departure for many types of analyses of NGS data comprising diverse populations, illustrated in Figure 1. I will demonstrate their construction, outline some opportunities for their future use in studying viral diversity and quasispecies, and discuss technical requirements for their appropriate and efficient use.

References

Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523. http://dx.doi.org/10.1007/BF00623322

Kim JT, Martinetz T, and Polani D (2003) Bioinformatic principles underlying the information content of transcription factor binding sites. *J Theor Biol* **220**, 529–544. http://dx.doi.org/10.1006/jtbi.2003.3153

Kim JT, Gewehr JE, and Martinetz T (2004) Binding matrix: A novel approach for binding site recognition. *J Bioinform Comput Biol* **2**, 289–307. http://dx.doi.org/10.1142/S0219720004000569

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948. http://dx.doi.org/10.1093/bioinformatics/btm404

Schneider TD, Stormo GD, Gold L, and Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**, 415–431. http://dx.doi.org/10.1016/0022-2836(86)90165-8

Schütze T, Wilhelm B, Greiner N, Braun H, Peter F, et al. (2011) Probing the SELEX process with next-generation sequencing. *PLoS One* **6**, e29604. http://dx.doi.org/10.1371/journal.pone.0029604

Stormo GD (2000) DNA binding sites: Representation and discovery. *Bioinformatics* **16**, 16–23. http://dx.doi.org/10.1093/bioinformatics/16.1.16

Vingron M and Sibbald PR (1993) Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci U S A* **90**, 8777–8781. http://dx.doi.org/10.1073/pnas.90.19.8777

Wilke CO, Wang JL, Ofria C, Lenski RE, and Adami C (2001) Evolution of digital organisms at high mutation rate leads to survival of the flattest. *Nature* **412**, 331–333. http://dx.doi.org/10.1073/pnas.90.19.8777

Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Merzyk P, et al. (2011) Beyond the consensus: Dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J Virol* **85**, 2266–2275. http://dx.doi.org/10.1128/JVI.01396-10

Bioinformatics Knowledge Transfer Programme (KTP) for Human and Capital Development in South Africa

Judit Kumuthini✉, **Emile Chimusa, Velaphi Masilela**
Centre for Proteomics and Genomic Research (CPGR), South Africa
Kumuthini J (2014) *EMBnet.journal* **20**(Suppl A), e791. http://dx.doi.org/10.14806/ej.20.A.791

A shortage of practical bioinformatics skills and expertise hinders scientific research and sustainable growth in Africa. This largely owes to underdeveloped infrastructures coupled with rare and depleted expertise that can facilitate skills transfer, including slow internet services, limited access and availability of life science research apparatus, and few organised activities that support knowledge expansion. Further compounding the problem is that individuals with some bioinformatics knowledge rarely collaborate with other experts on the African continent and rather look to international collaborative opportunities. This in turn decreases knowledge sharing and capacity development in Africa. The KTP aims to address this problem to develop and advance bioinformatics skill levels in Africa. KTP functions by acquiring the services of global experts to provide expertise and guidance to local research projects. Through this framework, KTP aims to promote knowledge exchange, grow local capacity, and strengthen the quality of local research data with the help of experienced and accredited bioinformaticians. KTP participants will work with an expert using a 'hands-on' approach to ensure that knowledge transfer occurs in a relevant environment using applicable examples. Trainees benefit from reduced training costs, exposure to a professional biotech environment and access to bespoke bioinformatics support. Experts gain exposure and experience in an African context, are afforded an opportunity to publish or patent research findings, and establish long-term collaborative relationships. The KTP recognises that input from global experts will help to ensure sustainable bioinformatics capacity development in Africa and contribute towards the identification of solutions to African problems.

Sex, deep sequencing and microRNAs

Antonio Marco
University of Essex, Essex, United Kingdom
Marco A (2014) *EMBnet.journal* **20**(Suppl A), e765. http://dx.doi.org/10.14806/ej.20.A.765

Expression analyses often quantify gene product levels in samples under diverse conditions, from multiple tissues or from different developmental times. These analyses frequently miss an important aspect in many species: sex. Most animals and some plants have separate sexes, whose morphological differences are the outcome of the biased expression of some genes.

MicroRNAs are important regulators that repress translation by targeting gene transcripts. Recently, we have investigated microRNAs with sex-biased expression pattern from deep sequencing data. We found that sex-biased microRNAs (SBMiR) are, as expected, expressed in the gonads. Strikingly, the evolutionary dynamics of SBMiR is different to that of protein coding genes (SBPG). For instance, SBMiR are frequently located in the sex chromosomes, whilst SBPG tend to be in the autosomes. Interestingly, SBPG often have duplicated copies with no expression bias, and SBMiR usually arise de novo during evolution (no duplication).

In this talk I will explore these and other findings on sex-biased microRNAs, and why we must account for sex differences when planning expression profile experiments.

References

Marco A (2014) Sex-biased expression of microRNAs in Drosophila melanogaster. *Open Biol* **4**, 140024. http://dx.doi.org/10.1098/rsob.140024

Marco A, Kozomara A, Hui JHL, Emery AM, Rollinson D et al (2013) Sex-Biased Expression of MicroRNAs in Schistosoma mansoni. *PLoS Negl Trop Dis* **7**(9), e2402. http://dx.doi.org/10.1371/journal.pntd.0002402

Coding & Best Practice in Programming: why it matters so much in the NGS era

Lex Nederbragt

Norwegian High-Throughput Sequencing Centre (NSC),Centre for Ecological and Evolutionary Synthesis (CEES), Dept. of Biosciences, University of Oslo, Oslo, Norway

Nederbragt L (2014) *EMBnet.journal* **20**(Suppl A), e769. http://dx.doi.org/10.14806/ej.20.A.769

Next generation sequencing has democratised large scale sequencing projects. No longer is generating a large sequencing dataset limited to genome centres. Neither is the analysis of large sequencing datasets limited to dedicated bioinformatics teams at genome centres or large institutions. Many more researchers now regularly obtain substantial NGS data for their projects. Wet lab biologists at all levels find themselves in need of acquiring computational skills to interpret their data. Coupled with a fast growing list of computational tools developed for the analysis of NGS data, this poses several challenges. Researchers inexperienced in judging computational tools need to choose the best/optimal one for the analysis of their data. Self-taught bioinformaticians are not familiar with best practices in computational science, leading to them make beginner's mistakes when they perform their computational analyses. Ultimately, this threatens correctness, reproducibility and reusability of the results obtained.

Luckily, all is not lost. There is an increasing awareness about the above issues, with reproducibility in computational science a topic receiving increased attention. Papers on best practices are published regularly. The non-profit organisation 'Software Carpentry' is helping out by organising two-day bootcamps where volunteers teach computational best practices to researchers.

In this talk, I will discuss what best practices in computational science are important, and why adhering to them, and teaching them, is crucial for our trust in the results obtained through NGS.

Training in computational skills

Aleksandra Pawlik
University of Manchester, Manchester, United Kingdom
Pawlik A (2014) *EMBnet.journal* **20**(Suppl A), e775. http://dx.doi.org/10.14806/ej.20.A.775

This talk will outline the essentials of training in computational skills in bioinformatics and beyond. The need for such training is clear: researchers need to be able to work independently and efficiently using a variety of computational tools. The skills which they require include the ability to automate tasks and build reproducible research pipelines, understand and be able to apply good programming practices in a programming language of choice, as well as being familiar with those software engineering tools that provide relevant support for computational research, such as version control. The skills should scale up enabling researchers to use large computational resources and cloud infrastructure such as Amazon EC2 or Microsoft Azure. The talk will also discuss how much of the training in computational skills is common across biosciences, and how and when it needs to be adjusted for the particular purposes of different disciplines. The example of the successful model of Software Carpentry[1] training shows that building on a common curriculum base makes possible to develop and deliver useful training packages.

1 software-carpentry.org

NGS data management and analysis for hundreds of projects: Experiences from Sweden

Ola Spjuth
Uppsala University, Uppsala, Sweden
Spjuth O (2014) *EMBnet.journal* **20**(Suppl A), e761. http://dx.doi.org/10.14806/ej.20.A.761

UPPNEX is a national e-infrastructure for next-generation sequencing data storage and analysis in Sweden. This presentation features strategic decisions made regarding hardware, software, maintenance and support, resource allocation, and illustrate challenges such as managing data growth in a shared system with over 400 research projects of varying types. Insights into bioinformatics usage patterns are also presented, together with the ongoing development to extend the e-infrastructure with redundant resources, a secure system for analyzing sensitive data, and a private cloud.

References

Lampa S, Dahlö M, Olason PI, Hagberg J, Spjuth O (2013) Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. *Gigascience*, **2**:9. http://dx.doi.org/10.1186/2047-217X-2-9

Using NGS to answer biological questions

Björn Usadel
RWTH Aachen University and Forschungszentrum Jülich, Germany
Usedal B (2014) *EMBnet.journal* **20**(Suppl A), e771. http://dx.doi.org/10.14806/ej.20.A.771

As generating next generation sequencing (NGS) data has become much cheaper, it is becoming more and more commonly used to address questions hitherto tackled by microarrays or by costly large scale EST sequencing. However, the computational challenge to analyse and interpret these data still remains.

Here we present some tools for the processing of NGS data. Focusing on RNA-seq data we show that using these tools, it is possible to get a first idea about major biological stories and to get a first overview which can then be used to develop biological hypotheses which can then be tested in more detail.

One example entails the often performed NGS analysis of non-model plant species and the exploration of metabolic pathways within these species (Lohse *et al.*, 2014). Having established these annotations, NGS data can then be analysed for statistical changes and explored for differences in expression which is demonstrated here (see Figure 1).

A main take home message however is that, even though these tools will help the experimenter in data analysis and interpretation, knowledge of the underlying biological system is of course required.

Acknowledgements

The authors would like to thank the BMBF for funding of the plant primary database 0315961.

References

Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR, Stitt M, Usadel B (2014) Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.* **37**(5):1250-8. http://dx.doi.org/10.1111/pce.12231

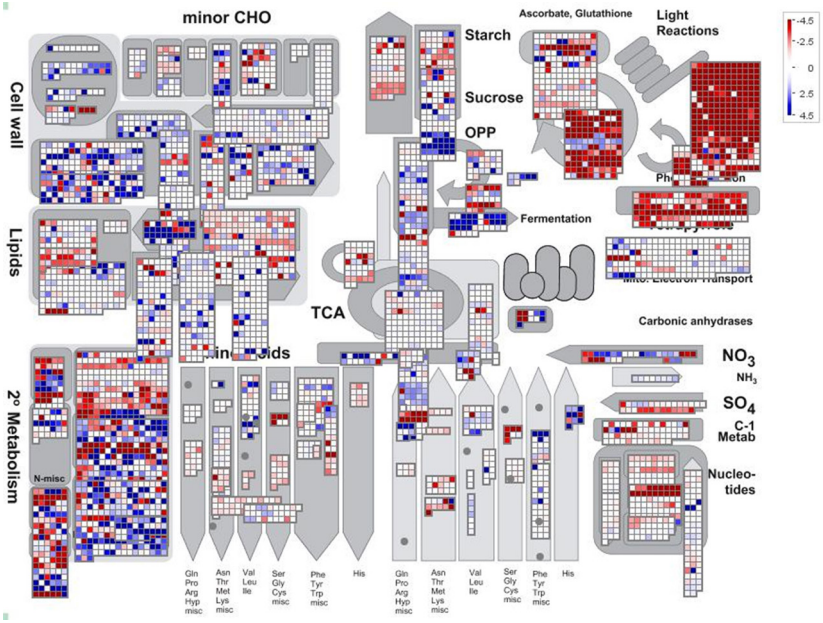

Figure 1.

Oral Presentations

Analysing RNA-Seq data with RNASeqGUI

Claudia Angelini✉, **Francesco Russo**

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

Angelini C and Russo F (2014) *EMBnet.journal* **20**(Suppl A), e783. http://dx.doi.org/10.14806/ej.20.A.783

RNA-seq has quickly become one of the most widely used approaches for identifying differences in the gene expression across multiple biological conditions. Several tools have been already developed to analyse such data. In general, a complete analysis requires carrying out several steps, using different methods and comparing their outputs to obtain more reliable and less biased results.

In this work, we present RNASeqGUI (Russo and Angelini, 2014), a novel R-package that implements a graphical platform devoted to analyse RNA-Seq data, and illustrate both the most recently added features and those ones that will been soon available. RNASeqGUI is not just a collection of some known methods and functions, but it is designed to guide the user during the entire analysis process. In particular, RNASeqGUI allows the identification of differentially expressed (DE) genes and corresponding pathways in RNA-Seq experiments by clicking buttons (see Figure 1).

RNASeqGUI incorporates several other R packages for DE analysis. It includes multiple normalisation procedures, tools for pathway analysis and a large number of functions to explore the data before the analysis and to compare results that have been obtained by different methods. Results are saved both in term of tab-delimited/html files and customizable graphical plots. Finally, in the spirit of "Reproducible Research", a human readable report is automatically generated to keep trace of all steps that have been performed during the analysis. Such report is generated by using R markdown language. Therefore, it incorporates both the documentation and the R code used to generate the results.

Availability: RNAseqGui is freely available at http://bioinfo.na.iac.cnr.it/RNASeqGUI/.

References

Russo F and Angelini C (2014) RNASeqGUI: A GUI for analyzing RNA-seq data. *Bioinformatics*. Advance Access Publication. http://dx.doi.org/10.1093/bioinformatics/btu308

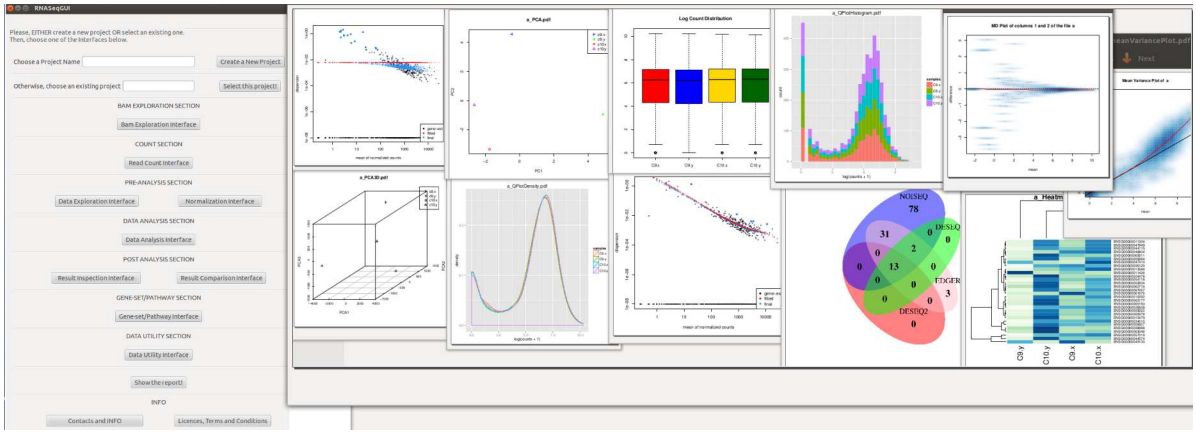

Figure 1. Screenshot of RNASeqGUI main interface and few examples of data analysis output files.

Posters

Contagious Bovine Pleuropneumonia Causative Agent: its Sequencing, Assembly and Annotation

Hadrien Gourlé

Swedish University of Agricultural Science, Uppsala, Sweden

Gourlé H (2014) *EMBnet.journal* **20**(Suppl A), e776. http://dx.doi.org/10.14806/ej.20.A.776

Mycoplasma mycoides mycoides subsp. Small Colony (MmmSC) is a small bacterium affecting cattle and causing the contagious bovine pleuropneumonia (CBPP), the deadliest cattle disease in Africa. This project aims at sequencing and annotating the genome of MmmSC strain Afade, one of the first *Mycoplasma mycoides* that has been directly extracted from African cattle.

The sequencing was realised with the Illumina MiSEQ 300PE technology, and a first quality control (Schmieder and Edwards, 2011) showed that although the quality of the reads and the sequence length distribution were perfect, the number of reads, as well as the duplication level, was way too high. Those being obvious signs of over sequencing, the genome assembler (Chevreux *et al.*, 1999) cannot produce any decent assembly with so many reads and an estimated coverage of 1500x. Several methods for downsampling were tried and the assembly was done with different coverage. The best one was obtained with a coverage assessment of 100x and the downsampling was finally done with a simple shell script that discarded 80% of the reads. 106 contigs and a N50 of 22,554 were obtained. Those contigs were mapped against a reference genome (MmmSC strain Gladysdale, an Australian strain of the bacterium) and merged with a python script. The annotation work is still ongoing.

References

Schmieder R and Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863-864. http://dx.doi.org/10.1093/bioinformatics/btr026

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) **99**, 45-56.

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 Sciences, 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

United Kingdom
The Genome Analysis Centre (TGAC), Norwich

Specialist- and Assoc. Nodes

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

KEMRI
Wellcome Trust Research Programme, Kilifi, 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

SBI
The New South Wales Systems Biology Initiative, Sydney, Australia

for more information visit our Web site
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
Tel: +46-18-67 21 21