EMBnet.journal

Volume 17
Supplement B
March 2012

Proceedings SeqAhead Scientific Meeting (COST Action BM1006)

7-9 November 2011
Brussels, Belgium

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

The completion of more than 30 mammalian genome sequences has triggered world-wide efforts to unravel their information content. Many genomics and proteomics projects have been successfully completed in areas such as gene hunting, functional annotation, prediction of post-translational modifications, identification of protein-protein interactions, and so on; others have been stimulated across the fields of transcriptomics and systems biology. These projects have resulted in the design and deployment of large numbers of novel analytic and predictive computer programs throughout the global life science community.

The majority of these developments have necessarily focused on harvesting the fruits of the human genome, transcriptome, proteome, epigenome, etc., given the universal pursuit of pinnacles in human health, wealth and happiness. Today, however, we live in interesting times (as the saying goes), experiencing a revolution in which thousands of other genomes relating to a host of unicellular organisms, plants and non-human animals are being sequenced in diverse research fields across the life sciences. This has been made possible by the advent of Next Generation Sequencing (NGS) (and, more recently, 3Gen) technologies, allowing vast amounts of genomic information to be generated at drastically reduced cost. Analysis of the rapidly accumulating sequence data brings new challenges: above all, it requires integration of bioinformatics and statistical approaches, throwing computational biology, computer science and mathematics into the melting pot in order to extract the wealth of information sequestered in the ever-growing pool of sequenced genomes.

It is in this context that the COST Action "SeqAhead" is working to help address the urgent community need for new and improved approaches to facilitate NGS data management and analysis. EMBnet.journal is honored to publish this Supplement, which has been generated by the SeqAhead consortium following its first combined Management Committee Business Meeting and Scientific Meeting, held in Brussels from November 7th-9th, 2011.

EMBnet.journal Editorial Board

This publication is supported by COST

Ana Conesa explaining their approach using RNA-seq for calculating differential gene expression

Contents

Editorial...2
COST Action BM1006 (SeqAhead): MC Business Meeting and Scientific Meeting ...4
Programme of the SeqAhead MC Business Meeting and Scientific Meeting...8
Oral Presentations ...10
Demos ...25
Posters ...27

Editorial Board:

Cover picture: Uppmax HPC Facilty, Uppsala Sweden, 2010.
[© Erik Bongcam-Rudloff]

COST is an intergovernmental framework for European Cooperation in Science and Technology, allowing the coordination of nationally-funded research on a European level. Established by the Ministerial Conference in November 1971, COST is presently used by the scientific communities of 36 European countries.

COST enables break-through scientific developments leading to new concepts and products and thereby contributes to strengthen Europe's research and innovation capacities.

COST is a unique means for European researchers to jointly develop their own ideas and new initiatives across all scientific disciplines through trans-European networking of nationally funded research activities. COST key features are:

- building capacity by connecting high-quality scientific communities throughout Europe and worldwide;
- providing networking opportunities for early career investigators;
- increasing the impact of research on policy makers, regulatory bodies and national decision makers as well as the private sector.

Through its inclusiveness COST supports integration of research communities, leverages national research investments and addresses issues of global relevance.

COST is a building block of the European Research Area[1], instrumental for successful innovation strategies and global cooperation.

Website: www.cost.eu

[1] http://ec.europa.eu/research/era/index_en.htm

COST Action BM1006 (SeqAhead): MC Business Meeting and Scientific Meeting

Andreas Gisel[1], Teresa Attwood[2], Jacques van Helden[3], Josè R. Valverde[4], Ana Conesa[5], Ralf Herwig[6], Eija Korpelainen[7], Veli Mäkinen[8], Steve Pettifer[9], Alberto Policriti[10], Thomas Svensson[11], Gert Vriend[12], Erik Bongcam-Rudloff[13]

[1]CNR, Institute for Biomedical Technologies, Bari, Italy
[2]Faculty of Life Sciences and School of Computer Science, University of Manchester, Machester, United Kingdom
[3]Lab. Technological Advances for Genomics and Clinics, Université d'Aix-Marseille , Marseille, France
[4]Centro Nacional de Biotecnología, CSIC, Madrid, Spain
[5]Centro de Investigaciones Príncipe Felipe, Valencia, Spain
[6]Max Planck Institute for Molecular Genetics, Berlin, Germany
[7]IT Center for Science, CSC, Helsinki, Finland
[8]Department of Computer Science, University of Helsinki, Helsinki, Finland
[9]School of Computer Science, The University of Manchester, United Kingdom
[10]University of Udine, Italy
[11]Karolinska Institut, Stockholm, Sweden
[12]Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands
[13]Swedish University of Agricultural Sciences, Uppsala, Sweden

SeqAhead is a COST Action created by a group of researchers involved in the development of Next Generation Sequencing (NGS) data-analysis software and pipelines. The primary objective of SeqAhead is to develop a coordinated action plan to help the scientific community to handle the flood of NGS data in an efficient and coherent manner, using state-of-the-art bioinformatics. Establishment of a strong European network of NGS, data-analysis and informatics centres will facilitate and stimulate the exchange of data, protocols, software, experiences and ideas ([1,2]).

Following its kick-off meeting on 13 March 2011, SeqAhead organised its first Management Committee (MC) meeting, combining the event with a scientific meeting and Working Group (WG) discussions, from 7-9 November 2011, in Brussels, Belgium.

The first part of the 3-day event was the official MC meeting. The local organiser, Jacques van Helden, and the Action Chair, Erik Bongcam-Rudloff, opened the event. Erik Bongcam-Rudloff gave an overview of the Action, pointing out that it had grown from the 18 member countries represented at the kick-off meeting to 22 signatories, highlighting both the relevance of, and the need for, such an Action within the biological community.

The budget for the first year of the Action was also presented. This included 3 meetings, one training school, 2 Short-Term Scientific Missions (STSMs), and a variety of other ad hoc dissemi-

All participants of the COST scientific meeting

[1] www.sequahead.eu
[2] http://journal.embnet.org/index.php/embnetjournal/article/view/218

nation activities. Preliminary ideas were outlined for the training school, planned for the end of May 2012, in Uppsala, in conjunction with a COST Action workshop; Erik confirmed that he had already booked the training facilities, with access to computer resources at the UPPMAX computer centre, to allow hands-on classes with 'real' NGS data.

The second part of the MC meeting was dedicated to the Working Groups. Each WG Chair briefly presented the Group's principal tasks and how these might be achieved.

***WG1, Technology Watch**, Chaired by Ralf Herwig and Thomas Svensson:* Ralf Herwig summarised the role of this group in providing timely alerts on new technology developments in topics such as sequence technology, analysis tools, applications and projects. The major activity of this WG will be scanning, reading and summarising scientific and technological articles. He proposed to organise a 'journal club', which will meet on a monthly basis to exchange the latest news on NGS technology developments.

***WG2, Action Plan for NGS Bioinformatics**, Chaired by Andreas Gisel and Ana Conesa:* Andreas Gisel described the role of this group first, in reviewing the challenges and gaps in analysis pipelines in parallel with WG1, and then in formulating actions that would be tackled in collaboration with WG3 and WG4. The creation of sub-committees on specific topics was discussed in order to galvanise as many Action and non-Action members as possible to work on these topics.

COST scientific session: Robert Lyle presenting the Norwegian Sequencing Center

Laurent Falquet presenting the SIB NGS infrastructure

Day 3: Joint work group 2 and work group 3 meeting

WG3, Software*, Chaired by Eija Korpelainen and Steve Pettifer:* Eija Korpelainen described the role of WG3 in gathering information on current data-analysis tools, including those under development, aiming to collaborate with WG4 to provide solutions in cases where these tools need to be customised to handle vast amounts of NGS data. A list of NGS tools developed by Action members has already been seeded on the WG4 page of the Action website.

WG4, Generic Informatics Topics*, Chaired by Veli Makinen and Alberto Policriti:* Veli Makinen described how this WG will focus on computer technology problems, such as data storage, interoperability, Grid and Cloud computing, and semantic applications.

WG5, Dissemination, Education and Training*, Chaired by Gert Vriend and Jacques van Helden:* Jacques van Helden explained that this WG will use several different media, including the portal and printed matter, to distribute information about NGS; it will also implement courses and teaching materials. An important role for this WG will be to propose standards for publishing NGS tools.

Before lunch, the COST Action BM0902, *"Network of experts in the diagnosis of myeloproliferative disorders (MPD)"* was presented by Sylvie Hermouet and Robert Kralowicz; their aim was to establish a close collaboration with SeqAhead, as they will be involved in extensive data-analysis scenarios using NGS technology in future.

During the afternoon, there was an open session on common aims and planned activities. In particular, a joint training school with TD0801 (Statseq, *"Statistical challenges on the €1000 Euro genome sequences in plants"*) and FA0806 (Plantivax, *"Plant virus control employing RNA-based vaccines: a novel non-transgenic strategy"*) was discussed, as were the form and location of the next MC meeting, together with a summary of the remaining activities and proposals for year 2 of the Action.

The second day of this COST event was organised as a scientific session, in which selected Action members and non-members (according to their submitted abstracts) were invited to present their work on NGS data-analysis platforms, tools and applications. There were presentations on 4 different platforms, given by Robert Lyle from The Norwegian Sequencing Centre (NSC), Oslo, Norway; Kjell Petersen, representing NGS research and services at the Computational Biology Unit (CBU), Bergen, Norway; Laurent Falquet from the Swiss Institute for Bioinformatics, Lausanne, Switzerland, representing the Vital-IT HPC and Swiss-Prot groups; and Ning Li, from Beijing, China, presenting the Beijing Genomics Institute (BGI) sequencing and bioinformatics strategy.

A range of applications, providing broad coverage across the NGS data-analysis debate, were also presented: Frank Picard outlined bioinformatics developments for NGS data analysis at PRABI; Ana Conesa reviewed NOIseq, an RNA-seq differential-expression method robust for sequencing-depth biases; Eric Rivals presented a combinatorial and integrated method to analyse RNA-seq reads; Jacques van Helden introduced RSAT peak-motifs, a pipeline for discovering motifs in massive ChIP-seq peak sequences; Luca Pireddu spoke about the Seal suite of distributed software for high-throughput sequencing; Keijo Heljanko presented scalable Cloud computing solutions for NGS data; Andreas Gisel discussed smallRNA data analysis; Eija Korpelainen presented Chipster 2.0, a user-friendly NGS data-analysis suite with built-in genome browser and workflow functionality; and Petr Baldrian reviewed the current possibilities and limitations in data analysis of environmental metagenomes and metatranscriptomes.

There was also a 'Miscellaneous' session, in which Jean Imbert presented HTS Science and gave a technology-watch tour; Matthias Steinbrecher spoke about innovation and trends with In-Memory technology; and José Ramón Valverde talked of NGS data-analysis from the user perspective. All presentations are available on the SeqAhead[3] site.

The final day of the event involved a series of parallel meetings in which each WG met to discuss its activities for the forthcoming Action year.

WG1 agreed to meet frequently and exchange information from publications they planned to analyse, and to provide frequent updates on the WG1 page of the Action website.

WG2, WG3 and WG4 agreed both to a number of important actions that SeqAhead should initiate, and to establish a primary repository of NGS data-analysis tools. The latter will be made available via the WG3 page of the Action website, and on the SeqAnswers[4] software hub. The agreed actions were to formulate, develop and publish on the website:

- a list of existing tools and platforms for NGS data analysis
- parallelisation (distributed computing) approaches for NGS data analysis
- protocols (descriptions on how to analyse NGS data), focusing initially around:
  - oncogenomics (e.g., how to align ChIP-seq data to a normal reference)
  - metagenomics

Protocols for other topics will be proposed and formulated in future, including, for example:

- variant annotation, especially for non-coding variants, association with phenotype
- genome annotation quality
- ncRNA analysis and annotation

A small group, mainly members of WG4, broke out from the WG2, WG3 and WG5 discussions to hold a first action meeting focusing on parallelisation approaches in NGS data analysis. The group agreed to organise a second action meeting on 'Hadoop technology', in February or March 2012, to develop solutions for the parallelisation of NGS data analysis.

WG5 discussed the practicalities and processes both for accepting applicants on Training Schools, and for awarding STSMs. The group discussed sets of criteria to facilitate these processes, and agreed to post further information, template application forms, etc., on the WG5 page of the Action website. An important role for this WG will be to propose standards for documenting NGS tools in order to make them usable by external users (user manual, demos, annotated study cases, utilization protocols).

The next SeqAhead, COST Action BM1006, events will be its inaugural training school and workshop, in Uppsala, Sweden, at the end of May and beginning of June 2012. Follow the Action on www.seqahead.eu and become active as an external expert in NGS data analysis.

[3] http://seqahead.cs.tu-dortmund.de/meetings:slides-2011-11

[4] http://seqanswers.com/wiki/SEQanswers

Programme of the SeqAhead MC Business Meeting and Scientific Meeting

Monday		**November 7 - Management Committee (MC) business meeting - By invitation only**
12:00	12:15	Welcome and lunch at COST office
		• *Welcome by local organizer (Jacques van Helden)*
		• *Welcome by the chair (Erik Bongcam)*
12:15	14:00	Lunch
14:00	18:00	MC (Management Committee) business meeting
		Auditorium room (COST office, 15th floor)
		• *Introduction by the chair Erik Bongcam-Rudloff*
		• *WG presentations by the WG chairs*
		WG1: Technology watch for new developments - Ralf Herwig
		WG2: Development of an Action Plan for NGS bioinformatics to cope with challenges for ERA - Andreas Gisel
		WG3: Design, implementation, and incorporation of software solutions - Eija Korpelainen
		WG4: Generic informatics topics - Veli Makinen
		WG5: Development of a strategic dissemination and education program for NGS bioinformatics - Jacques van Helden
		• *Discussion of common aims*
		• *Discussion programme COST Action year 2*
		• *Discussion training school and workshop in Uppsala*
		• *Alliances:*
		Sylvie Hermouet: BM0902: Network of experts in the diagnosis of myeloproliferative disorders (MPD)
Tuesday		**November 8 – Scientific program meeting**
		Scientific program meeting
		Auditorium room (COST office, 15th floor)
09:00	10:20	Session 1 - Facilities
09:00	09:20	Robert Lyle. The Norwegian Sequencing Centre (NSC)
09:20	09:40	Kjell Petersen. NGS research and service at the CBU
09:40	10:00	Laurent Falquet. The Vital-IT HPC and the Swiss-Prot group
10:00	10:20	Ning Li. BGI: combination of sequencing and bioinformatics strategy
10:20	11:00	Coffee break
11:00	12:00	Session 2 – Various

11:00	11:20	Matthias Steinbrecher. Innovation and Trends with In-Memory Technology
11:20	11:40	Jean Imbert. HTS Science and Technology Watch Tour
11:40	12:00	José Ramón Valverde. NGS data analysis: the user POV
12:00	14:00	Lunch + posters
14:00	16:00	Session 3 - Tools and applications
14:00	14:20	Frank Picard. Bioinformatics developments for NGS data analysis at PRABI
14:20	14:40	Ana Conesa. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases
14:40	15:00	Eric Rivals. A combinatorial and integrated method to analyse RNA-seq reads
15:00	15:20	Jacques van Helden. RSAT peak-motifs: fast extraction of transcription factor binding motifs from full-size ChIP-seq datasets
15:20	15:40	Luca Pireddu. The Seal suite of distributed software for high-throughput sequencing
15:40	16:00	Keijo Heljanko. Scalable Cloud Computing Solutions for Next Generation Sequencing Data
16:00	16:30	Coffee break
16:30	17:30	Session 3 - Tools and applications (continued)
16:30	16:50	Andreas Gisel. smallRNA data analysis.
16:50	17:10	Eija Korpelainen. Chipster 2.0: User-friendly NGS data analysis software with built-in genome browser and workflow functionality
17:10	17:30	Petr Baldrian. Exploration of environmental metagenomes and metatranscriptomes: current possibilities and limitations in data analysis
17:30	18:15	Discussions: Challenges and perspectives
20:00	22:00	Dinner
Wednesday		**November 9 - Work group meetings**
09:00	09:20	Introduction by Erik Bongcam
09:20	11:30	Split meeting (parallel sessions)
		• *WG1: Technology watch for new developments - Ralf Herwig* • *WG2: Development of an Action Plan for NGS bioinformatics to cope with challenges for ERA - Andreas Gisel* • *WG3: Design, implementation, and incorporation of software solutions – Eija Korpelainen* • *WG4: Generic informatics topics - Veli Makinen* • *WG5: Development of a strategic dissemination and education program for NGS bioinformatics - Gert Vriend*
11:30	12:30	Conclusion
12:30		End of the meeting

Oral Presentations

A combinatorial and integrated method to analyse RNA-seq reads

Nicolas Philippe, Mikael Salson, Therese Commes, Eric Rivals
Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS,
Université de Montpellier II, Montpellier, France;
LIFL, CNRS, INRIA Lille, Univ. Lille I, Villeneuve d'Ascq, France;
CRBM, UMR 5237 CNRS, Montpellier, France
http://www.lirmm.fr/~rivals

RNA sequencing enables a complete investigation covering the full dynamic spectrum of a transcriptome. It thus paves the way to a better understanding of the function of gene expression in different tissues, during development or pathological states. However, the splicing process, which generates both co-linear and non co-linear RNAs, the inclusion of sequencing errors, somatic mutations, polymorphisms, and rearrangements make the reads differ from the reference genome in a variety of ways. This complicates the task of comparing reads with a genome. Currently, the analysis paradigm consists in:

1. mapping the reads to a reference genome contiguously allowing as many differences as one expects to be necessary to accommodate sequence errors and small polymorphisms;
2. using uniquely mapped reads to determine covered genomic regions, either for computing a local coverage to predict mutations and filter out sequence errors (cf. program ERANGE), or for delimiting expressed exons approximately (cf. program TopHat);
3. re-aligning unmapped reads, which were not mapped contiguously at step one, to reveal splicing junctions.

Limitations of this approach include lack of precision, redundant computations due to multi-mapping steps, error propagation due to heuristics and the absence of back-tracking. We propose a novel, integrated approach to analyze nowadays longer reads (> 50 bp). The idea is to adopt a k-mer approach that combines the genomic positions and local coverage to perform a complex analysis of each read and detect in a single step, mutations, indels, errors, as well as both normal and chimeric splice junctions. Comparisons with other tools demonstrate the feasibility of this approach, which yields both sensitive and highly specific inferences.

References

1. N. Philippe, M. Salson, T. Lecroq, M. Leonard, T. Commes and E. Rivals; Querying large read collections in main memory: a versatile data structure. BMC Bioinformatics, Vol. 12, p. 42, doi:10.1186/1471-2105-12-242, 2011.

Relevant Web sites

2. http://crac.gforge.inria.fr/gkarrays/
3. http://www.atgc-montpellier.fr/ngs/

Bioinformatics developments for NGS data analysis at PRABI

Franck Picard, Guy Perrière

Pôle Rhône-Alpes de Bioinformatique, Bât. Gregor Mendel, Université Claude Bernard Lyon 1, Villeurbanne, France

http://www.prabi.fr/

The recent developments performed at PRABI for NGS data analysis are led in three main directions: i) short reads clustering for metagenomic data; ii) Open Reading Frames (ORFs) detection in metagenomes; and iii) statistical detection of peaks applied to the identification of replication origins on the human genome and to chIP-Seq data.

One of the problems frequently encountered with present day metagenomic data is the large amount of reads that have no significant homologs in the repository sequence data banks. In order to see if, at least, those "orphans" share some similarities among themselves, a lot of different clustering strategies have been developed. The strategy we have chosen to explore at PRABI is a distance-based one, as opposed to the model-based ones. More precisely, we have focused on the use of Correspondence Analysis (CA) and derived methods [1]. Due to its simplicity, this method is easy to use, very fast and efficient with large data sets containing hundreds of thousands of reads. On the other hand, its efficiency rapidly decreases when the number of different taxa present in the samples is high.

The approach chosen for ORFs detection is also based on CA. In this case, the analysis is computed on the codon composition of the six possible reading frames of a sequence [2]. The main advantage of this method is that it does not require a training step (like in Glimmer), therefore it can be used on metagenomic data, even if the biodiversity expected in the samples is very high. Tests on simulated metagenomic data sets show that the sensitivity of the program is 59% while specificity is 89%. The low sensitivity is due to fact that the efficiency of the method is highly dependant on the intensity of the codon bias in the coding sequence. Therefore, weakly biased genes (such as lowly expressed genes when there is translational selection in the species considered) are often missed by the method.

Lastly, for the detection of peaks in NGS data, the novelty is to develop a rigorous statistical framework to detect exceptional enrichment of reads using Poisson processes and scan statistics. It is a powerful framework that allows to define a proper P-value and FDR for the peaks, and our project is now to focus on the realistic modeling of the coverage function along the genome in order to adapt the significance of the peaks to a background noise that is highly dependent on the genomic context. As an extension and perspective, we plan to develop a statistical methodology to compare chIP-Seq data between conditions, and to assess the significance of differential peaks. This strategy will be applied also to the detection of differentially expressed small RNAs.

References

1. Perrière, G. and Thioulouse, J. (2002) Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res., **30**, 4548-4555.
2. Fichant, G. and Gautier, C. (1987) Statistical method for predicting protein coding regions in nucleic acid sequences. Comput. Appl. Biosci., **3**, 287-295.

Relevant Web sites

3. http://metasoil.univ-lyon1.fr/

Exploration of environmental metagenomes and metatranscriptomes: current possibilities and limitations in data analysis

Petr Baldrian
Laboratory of Environmental Microbiology, Institute of Microbiology of the ASCR, Prague, Czech Republic
http://www.biomed.cas.cz/mbu/lbwrf

Environmental metagenomes and metatranscriptomes are extremely complex, considering that one gram of soil may harbor tens of thousand species of bacteria and thousands of species of eukaryotic microorganisms. Their exploration thus currently relies on methods delivering relatively long sequence reads, i.e. these obtained with the Roche or Pacific Biosciences instruments. Shotgun approaches are combined with sequencing of PCR amplicons of genes with sufficient taxonomic resolution (rDNA) or, less frequently, functional genes. Our recent experience shows, that a description of total (DNA sequencing) and active (sequencing of cDNA derived from environmental RNA) soil microbial communities or the pools of functional genes and their transcripts (mRNAs) can be sufficiently well characterised using amplicon sequencing (1). The analysis of metagenomes is much more challenging since the sequence identity has to be determined and the assignment of functions and microbial producers to such sequences is not trivial. Current possibilities of metagenomic data analysis would benefit mainly from the tools allowing to search not only in GenBank (as most of the current tools do) but also in the full genomes of individual microorganisms, or, as a best option, in a database covering all these genomes. Furthermore, amplicon sequencing, that now relies on the construction of consensus sequences representing putative microbial species (OTUs, operational taxonomic units) would greatly advance if an automatic tool of consensus construction of all identified similarity clusters is developed. As our first results in the field of environmental metaproteomics show, even more sophisticated tools would be needed if metaproteomic data, typically short sequences of amino acids, need to be compared with nucleotide sequences obtained using DNA or cDNA sequencing.

References

1. Baldrian, P., Kolarik, M., Stursova, M., Kopecky, J., Valaskova, V., Vetrovsky, T., Zifcakova, L., Snajdr, J., Ridl, J., Vlcek, C., Voriskova, J. (2011) Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME Journal in press, doi:10.1038/ismej.2011.95.

HTS Science and Technology Watch Tour

Jean Imbert
TAGC UMR _ S 928, Inserm, Université de la Méditerranée, Marseille, France
http://www.yourwebsite.org/

I have recently performed on behalf of Inserm a Science and Technology Watch Tour on HTS in USA from March 25 to April 12, 2011 as the chairman of the Scientific Board of Inserm Workshops. These workshops are dedicated to high level and innovative training. Inserm workshops were created in 1987 with a triple objective: (i) to investigate emerging or rapidly evolving questions ; (ii) to diffuse quickly information ; (iii) to promote the rapid efficient acquisition of news techniques for a direct and immediate impact on the development of ongoing research programs in biomedical research in France. They are organized under the direction of leading international experts and with the participation of researchers, engineers, technicians and M.DS working in Academic institutes, universities, hospitals and industries. They are divided in 2 phases. Phase I presents a critical assessment, initiation and information for the best choice of research strategies. Phase II is a training session to acquire a particular technique to be used in a well-defined research project.

The tour has involved the visit of the major companies on the San Francisco Bay area in California (Applied Biosystems, Illumina, Ion Torrent, Pacific Biosciences, Complete Genomics) as well as 3 major academic genome centers (Human Genome Sequencing Center, Baylor College of Medicine, Houston TX; The Genome Center, WUSL, St Louis, MI; NIH Intramural Sequencing Center, Rockville, MD).

I will present a synthesis of my visit oriented toward the real performances of the present machines as well as on what we can expect in few months.

Relevant Web sites

1. http://www.rh.inserm.fr/INSERM/IntraRh/RHPublication.nsf/mDisplayMotsClefsWeb?OpenForm&arg1=19&arg2=#/
2. http://rechercher.rh.inserm.fr/cgi-bin/findall?C=193&X=2&KEYWORDS=atelier&SORT_ORDER=afs:relevance|DESC-DATE|DESC&CAT=DOCUMENTATION&UNIQUE=user2
3. http://www.rh.inserm.fr/INSERM/IntraRH/RHPublication.nsf/vPubRH/641A388D42288202C125785B004919CB?OpenDocument

NGS data analysis: the user POV

Jose R. Valverde, Jose M. Rodriguez, Alexandro Rodriguez-Rojas, Alejandro Couce, Jesus Blazquez
CNB/CSIC, C/Darwin 3, Madrid, Spain
http://www.es.embnet.org/

Bioinformaticians working in NGS are used to in-depth involvement in difficult problems and developing ingenious solutions to solve each and every specific user need. The users' point of view (POV) however tends to drift from their initially specific plans into fuzzier forays.

When used in the wet lab, NGS data opens a hoard of potential studies to carry out, empowering users to address several complex problems at once. This broad potential compels users to aim towards exhaustive mining of their NGS data in a multidimensional approach in an attempt to extract maximum information from their experimental results (e. g. deep sequencing for theoretical model validation may help characterise novel strains, identify mutations, understand evolutive events and do genome reconstruction as well). However, data analysis is still a difficult task requiring strong bioinformatics support, and while attractive, post hoc multidirectional analysis entails major challenges that may some times be better served by careful planning in close collaboration with a bioinformatician or a bioinformatics community.

Deeper understanding of users' initial expectations and how they evolve after data has been collected, their demands, analysis patterns, and requirements provides useful insight on the major problems faced and to be addressed by bioinformaticians and software developers involved in SEQAHEAD.

In this talk we draw on our experience working in close collaboration with users and applications at CNB to present the users' point of view on NGS data analysis, its inherently polifacetic approach to laboratory problems and raise some concerns with the way NGS is currently being considered by users vs. developers, suggesting possible approaches to deal with this post hoc complexity by exploiting SEQAHEAD collaborative infrastructure.

Relevant Web sites

1. http://www.es.embnet.org/
2. http://www.cnb.csic.es/
3. http://www.cnb.uam.es/content/research/microbial/stress/

NGS research and service at the CBU

Kjell Petersen, Inge Jonassen

Computational Biology Unit, Uni Computing, Uni Research AS, Bergen, Norway

http://www.bioinfo.no/

http://www.uni.no/computing/units/cbu

CBU consists of seven research groups and one service group, specialising in different aspects of computational biology. A common denominator in many of our projects are high throughput data sets, with Next Generation Sequencing as a prominent data providing technology. CBU also co-ordinates the Norwegian Bioinformatics platform that offers both helpdesk support and training to scientists in the field of functional genomics.

A natural research focus in Bergen is marine genomics. The recently published genome of cod [1] were accomplished with CBU as an active partner in the bioinformatics work, in particular the assembly of the 454 reads. Through this and other projects special competence on analysis of high-throughput sequencing (in particular, 454) data has been built, as documented in the work to realise the FlowSim tool [2].

Metagenomics on samples from extreme environments along the mid-Atlantic ridge is another field of high interest in Bergen, due to the Centre of Geo-biology (a National Centre of Excellence) situated next to CBU. Through this collaboration, new approaches to handle amplicon sequencing datasets from 454 have been developed, and implemented in the AmpliconNoise software tool [3].

Through our role in the National bioinformatics helpdesk and our close collaboration with the Norwegian Microarray Consortium, we have extensive experience in designing experiments and analysing gene expression data from high throughput datasets. Both analysis of data in research projects and training that we provide through the Bioinformatics platform and NMC have successfully been based on the J-Express analysis software suite [4].

In addition to algorithms and tools, a suitable infrastructure for step-by-step analysis of your work-flow, as well as sharing of data, results and methods across disciplines in a project group, is vital for proper utilization of your data. This is the aim of the eSysbio project, and components of the system are currently in use to implement the StoreBioinfo portal for providing high capacity storage for Life Science data sets in NorStore storage resource (national e-infrastructure).

Based on the total research experience and expertise of CBU and on the analysis and e-infrastructure competence built in the national network operated over the 9 previous years, we have coordinated an application for establishing a Norwegian node of the ELIXIR pan-European infrastructure network for bioinformatics.

References

1. Star et al. The genome sequence of Atlantic cod reveals a unique immune system, Nature, **2011**, 477, 207-210.
2. Balzer et al. Characteristics of 454 pyrosequencing data, Bioinformatics, **2010**, 26, i420-i425.
3. Quince et al. Removing noise from pyrosequenced amplicons, BMC Bioinformatics, **2011**, 12, 38.
4. Stavrum et al. Analysis of gene-expression data using J-Express, Curr Protoc Bioinfo, **2008**, Chapter 7, Unit 7.3.

Relevant Web sites

5. http://www.bioinfo.no/
6. http://www.microarry.no/
7. http://jexpress.bioinfo.no/
8. http://www.esysbio.org/
9. http://storebioinfo.norstore.no/

Innovation and Trends with In-Memory Technology

Matthias Steinbrecher
Innovation Center Potsdam, TIP HPI Strategic Projects SAP AG, Potsdam, Germany
http://www.sap.com

High performance in-memory computing will change how enterprises work. Currently, enterprise data is split into two databases for performance reasons. Usually, disk-based row-oriented database systems are used for operational data and column-oriented databases are used for analytics. Since hardware architectures have evolved dramatically during the past decade, this scenario has now changed. Multi-core architectures and the availability of large amounts of main memory at low costs are about to set new breakthroughs in the software industry. Traditional disks are one of the last remaining mechanical devices in a world of silicon and are about to become what tape drives are today: a device only necessary for backup. With in-memory computing and hybrid databases using both row and column-oriented storage where appropriate, transactional and analytical processing can be unified, allowing data analysis algorithms to run inside the database.

Relevant Web sites

1. http://www.sap.com/

NOIseq: a RNA-seq differential expression method robust for sequencing depth biases

Sonia Tarazona, Fernando García, Alberto Ferrer, Joaquín Dopazo, Ana Conesa
Genomics of Gene Expression Lab, Centro de Investigaciones Príncipe Felipe, Valencia, Spain
http://bioinfo.cipf.es/aconesa

Introduction

Next Generation Sequencing (NGS) technologies have brought a revolution to research in genome and genome regulation. One of the most breaking applications of NGS is in transcriptome analysis. RNA-seq has revealed exciting new data on gene models, alternative splicing and extra-genic expression. Also RNA-seq permits the quantification of gene expression across a large dynamic range and with more reproducibility than microarrays. Several methods for the assessment of differential expression from count data have been proposed but biases associated to transcript length and transcript frequency distributions have been reported. It is still not clear how much sequencing reads should be generated in a RNA-seq experiment to obtain reliable results and what's exactly being detected. In general we observed that many RNA-seq datasets have not reached saturation for detection of expressed genes and that the relative proportion of different transcript biotypes changes with increasing sequencing depth. In this work we investigate the effect that library size has on the assessment of differential expression on different aspects of the selected genes. We show that current statistical methods suffer from a strong dependency of their significant calls on the number of mapped reads considered and proposed a novel differential expression methodology – **NOISeq**[1]- that is robust to the amount of reads.

Results

NOISeq is a non-parametric approach for the differential expression analysis of RNseq-data. NOISeq creates a null or noise distribution of count changes by comparing the number of reads of each gene in samples within the same condition. This reference distribution is then used to assess whether the change in count number between two conditions for a given gene is likely to be part of the noise or represents a true differential expression. Two variants of the method are implemented: NOISeq-real uses replicates, when available, to compute the noise distribution and, NOISeq-sim simulates them in absence of replication. We compared our method with edgeR2, DESeq3, baySeq4 and Fisher Exact Test (FET) using three different experimental datasets. Results are presented for MAQC experiment where the transcriptome of brain and Universal Human Reference (HUR) samples were sequenced at about 45 million Solexa reads each.

We first determined that although protein-coding gene is the most abundant transcript type within differential expression calls for all methodologies, other RNA types, such as processed-transcript, pseudogenes and lincRNAs are readily detected. NOISeq dected comparatively more protein-coding genes than other methods that called significant a considerable number of non-coding and small RNA transcripts. Additionally, all comparing methods except FET greatly increased the number of detected (non-coding) genes as sequencing depth raised while NOISeq showed a constant pattern. Also these other methods tend to select shorter genes and smaller fold change differences with the increasing amounts of reads. In general, parametric approaches selected much more genes than NOISeq, specially at high sequencing depth rates. When analyzing the functional content of these genes by functional enrichment analysis, we observed that the pool of genes detected both by NOISeq and the parametric methods where highly enriched in functional categories, while genes selected only by parametric methods did not. To check whether this differences were indicative of different false calls between methods, we used the RT-PCR data available at the MAQC project that contains 330 true positive and 83 true negative differentially expressed genes. Performance plots indicate that edgeR, DESeq, baySeq strongly increased the number of false calls with sequencing depth, while NOISeq was constant and low. On the contrary true discoveries were slightly better for these methods, presumably consequence of their large number of selected genes. FET showed in low false and true discovery rates, due to its general lower detection power.

Conclusions

We showed that most current RNA-seq statistical analysis methods fail to control the number of false discoveries as the size of the sequenced library increases. These false positive are mainly short, non-coding genes and/or genes with small fold changes. NOISeq, but adopting an empirical approach to model the null distribution of differential expression captures better the shape of noise in RNA-seq data, resulting in a sequencing-depth robust method for differential expression analysis.

References

1. Tarazona S., Garcia-Alcalde F., Ferrer A., Dopazo J., Conesa, A. Differential expression in RNA-seq:a matter of depth. Genome Research, Sep 2011, doi:10.1101/gr.124321.11.
2. Robinson, MD, McCarthy, DJ, and Smyth, GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139,140.
3. Anders, S and Huber, W. 2010. Differential expression analysis for sequence count data. Genome Biology 11(10):R106.
4. Hardcastle, T and Kelly, K. 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11(1):422+.

Relevant Web sites

5. http://bioinfo.cipf.es/noiseq

RSAT peak-motifs: fast extraction of transcription factor binding motifs from full-size ChIP-seq datasets

Morgane Thomas-Chollier[1], Matthieu Defrance[2], Olivier Sand[3], Carl Herrmann[4], Denis Thieffry[4] and Jacques van Helden[,4,5]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

[2]Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México. Av. Universidad, Cuernavaca, Morelos, Mexico

[3]CNRS-UMR8199 Institut de Biologie de Lille. Génomique et maladies métaboliques, Lille, France

[4]Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée. Campus de Luminy, Marseille, France

[5]Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, Bruxelles, Belgium

http://rsat.ulb.ac.be/rsat/

ChIP-seq has become a method of choice to study binding preferences of transcription factors, and localization of epigenetic regulatory marks at a genomic scale. There is a crucial need for specialized software tools to make sense of these data. While various programs have been developed to perform read mapping and peak calling, the subsequent steps have not yet reached proper maturation: identifying relevant transcription factor binding motifs and the precise location of their binding sites remains a bottleneck. Most existing tools present limitations on sequence size, and typically restrict motif discovery to a few hundreds peaks.

We present a pipeline called peak-motifs, integrated in the Regulatory Sequence Analysis Tools[1], which takes as input a set of peak sequences, discovers exceptional motifs, compares them with motif databases, predicts binding site positions, and offers different visualization interfaces. The pipeline relies on tried-and-tested algorithms whose computing time increases linearly with sequence size, ensuring scalability to massive datasets of several tens of Mb. In addition to the website, peak-motifs can be used as stand-alone application, as well as SOAP/WSDL web services.

We assessed *peak-motifs* performances on several published datasets. In all cases, relevant motifs are disclosed. For example, we discovered individual Oct and Sox motifs in Sox2 and Oct4 peak collections, whereas the original study only found the composite Sox/Oct motif. For the generic transcriptional co-activator p300 examined in heart and midbrain, *peak-motifs* identified motifs bound by tissue-specific transcription factors consistent with these two tissues.

In summary, *peak-motifs* supports time-efficient and statistically reliable analysis of complete ChIP-seq datasets, while offering an online user-friendly and well-documented interface.

References

1. Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. and van Helden, J. (2011). RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Res 39, W86-91.
2. Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D. and van Helden, J. (2011). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets Nucleic Acids Res accepted.

Relevant Web sites

3. http://rsat.ulb.ac.be/rsat/

[1] http://rsat.ulb.ac.be/rsat/

smallRNA data analysis

Angelica Tulipano, Andreas Gisel
Institute for Biomedical Technologies – CNR, Bari, Italy
http://www.ba.itb.cnr.it/

The discovery of small RNA, such as miRNA and siRNA, opened a new dimension in gene regulation on the level of transcriptional and post-transcriptional regulation (1). To understand the distribution and expression levels of small RNAs is therefore crucial to understand tissue development (2), diseases (3), therapies with xenobiotic medicaments (4) or with small RNAs (5). Furthermore, each cell type, each tissue has a different onset of small RNAs and their expression. Only a large amount of samples from all these tissues will reveal the whole "small RNA-om". Technologies such as NGS heavily supports the investigations of these small RNA such as that a deep sequencing approach gives a holistic view of a snapshot of the small RNA regulatory activity in a biological sample. With the increasing number of sequence output offered by the different NGS technologies, the analysis of these large numbers of sequences especially for small RNA data analysis become time consuming and prone of errors in respect of the prediction of new small RNAs.

Because NGS produces in one experiment such a large number of sequences the technologies offer to run in parallel several samples labelled with a short barcode sequences. Therefore a typical workflow to analyse such a deep sequencing small RNA data set starts with the identification of these barcodes at the 5′ end of the reads from up to 100 million sequences, remove the barcode sequence and search at the 3′ end for the adaptor sequence and remove also these sequence fragments; logistically not too complex but computational very intensive. An intelligent distribution on different threads per CPU, on a GPU, in a cloud or over the GRID would dramatically reduce this process. The next step is the mapping of these cleaned reads onto the reference genome and find potential precursor sequences from known or new miRNA genes which would fold in a proper stem-loop secondary structure. This second more complex step is also computational demanding but more important includes a process for the selection of the proper folding for the cutting site to produce the mature miRNA and the miRNA. Since the feature of such a folding is quite broad the workflow needs to be flexible and user controllable to adjust a range of parameter to extract a list of significant potential miRNA genes and the corresponding mature miRNA.

We are developing a workflow, which starts with the read processing from multiplexed sequencing data (Illumina) and offers a mapping procedure and a corresponding miRNA recognizing procedure with a range of parameters to adjust the output.

References

1. Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., & Mattick, J. S. (2010). Non-coding RNAs: regulators of disease. The Journal of pathology, 220(2), 126–139. doi:10.1002/path.2638.
2. Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research, 18(4), 610–621. doi:10.1101/gr.7179508.
3. Joyce, C. E., Zhou, X., Xia, J., Ryan, C., Thrash, B., Menter, A., Zhang, W., et al. (2011). Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome Human molecular genetics, 20(20), 4025–4040. doi:10.1093/hmg/ddr331.
4. Rodrigues, A. C., Li, X., Radecki, L., Pan, Y.-Z., Winter, J. C., Huang, M., & Yu, A.-M. (2011). MicroRNA expression is differentially altered by xenobiotic drugs in different human cell lines Biopharmaceutics & drug disposition, 32(6), 355–367. doi:10.1002/bdd.764.
5. Gandellini, P., Profumo, V., Folini, M., & Zaffaroni, N. (2011). MicroRNAs as new therapeutic targets and tools in cancer. Expert opinion on therapeutic targets, 15(3), 265–279. doi:10.1517/14728222.2011.550878.

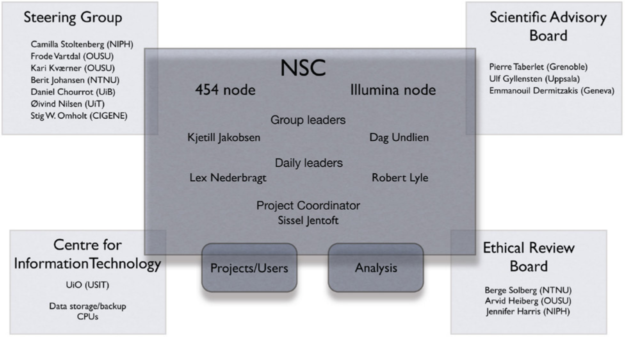
The Norwegian Sequencing Centre (NSC)

Robert Lyle, Tim Hughes, NSC, Dag Undlien, Kjetill Jakobsen

Department of Medical Genetics and Norwegian Sequencing Centre Oslo University Hospital, Kirkeveien 166, 0407, Oslo, Norway

http://www.sequencing.uio.no/

The Norwegian Sequencing Centre (NSC) was established in 2009 to provide the Norwegian research community with access to high-throughput sequencing services. Currently, 16 people work at the NSC. Funding has been provided from a number of sources, including Health South-East, the University of Oslo, and the Norwegian Research Council.



Two main activities of the NSC will contribute to the aims of the SEQAHEAD project. 1. Experience providing a very broad range of sequencing applications based on a range of technology platforms. 2. The development of a national storage and analysis platform for human genetic data.

1. The NSC has 454 (GS FLX+), Illumina (GAIIx, HiSeq) and Pacific Biosciences (RS) platforms. In addition Illumina MiSeq and Ion Torrent (PGM) machines have been ordered. This enables us to support a broad range of projects and sequencing applications. This includes large scale de novo projects, such as the cod genome (doi:10.1038/nature10342), transcriptomics (mRNA, miRNA), epigenetics (RRBS, ChIP), and resequencing (exomeSeq).

2. The medical genetics department at the Oslo University Hospital has initiated and received funding for the development of a national storage and analysis platform for DNA sequence data to be used by the Norwegian health service. Partners in the project are the University High Performance Computing unit, the Informatics Department, and the hospital IT and Data Protection units. The system should enable the secure transfer of sequence data and meta-data from production sites to the system, strict access control functionality, secure communication between users of the system, and interfaces for power users (e.g. bioinformaticians and medical genetics clinicians) and expert computer systems (e.g. pharmaco-genetics expert system). In addition, the system needs to be highly scalable to accomodate what is anticipated to be the explosive use of genetic information in the treatment of a broad range of pathologies. The above requirements will require the design and development of a secure high performance computing infrastructure that not only satisfies the technical requirements, but also complies with the strict data security laws that apply to sequence data in Norway. In addition, secure data software services will need to be developed and run on top of this infrastructure. The goal is to have a working pilot of the system installed by the spring of 2015.

Relevant Web sites

1. http://www.sequencing.uio.no/
2. http://codgenome.no/
3. https://wiki.uio.no/usit/suf/vd/hpc/index.php/Tsd

The Seal suite of distributed software for high-throughput sequencing

Luca Pireddu, Simone Leo, Gianluigi Zanetti

CRS4, Polaris, Ed. 1, Pula, Italy

http://www.bioinfo.no/

http://www.uni.no/computing/units/cbu

Modern DNA sequencing machines have opened the flood gates of whole genome data; and the current processing techniques are being washed away. Medium- sized sequencing laboratories can produce Terabytes of data per week that need processing. Unfortunately, most programs available for sequence processing are not designed to scale easily to such high data rates, nor are the typical bioinformatics workflow designs. As a consequence, many sequencing operations are left struggling to cope with the high data loads, often hoping that acquiring additional hardware will solve their problems. In contrast, we believe that a change in paradigm is required to solve this problem: a shift to highly parallelized software is required the handle the parallelization that has taken place in sequencing.

In response to the growing processing requirements of the CRS4 Sequencing and Genotyping Platform (CSGP), which now houses 4 Illumina HiSeq 2000 sequencers for a total capacity of about 7000 Gbases/month, we began the development of Seal [3], a new suite of sequence processing tools based on the MapReduce [1] programming model that run on the Hadoop framework. Seal aims to replace many of the tools that are customarily used in sequencing workflows with Hadoop-based, scalable alternatives. Currently, Seal provides distributed MapReduce tools for: demultiplexing tagged reads, mapping reads to a reference (it includes a distributed version of the BWA aligner [2]), and sorting reads by alignment position. In the near future we will also be adding tools for read quality recalibration.

Seal tools have been shown to scale well in the amount of input data and the amount of computational nodes available [4]; therefore, with Seal one can increase processing throughput by simply adding more computing nodes. Moreover, thanks to the robust platform provided by Hadoop, the effort required by operators to run the analyses on a large cluster is generally reduced, since Hadoop transparently handles most hardware and transient network problems, and provides a friendly web interface to monitor job progress and logs. Finally, the Hadoop Distributed File System (HDFS) provides a scalable storage system that scales its total throughput hand in hand with the number of processing nodes. Thus, it avoids creating a bottleneck at the shared storage volume and avoids the need for an expensive high-performance parallel storage device.

Seal is currently in production use at the CRS4 Sequencing and Genotyping Platform and is being evaluated at other various sequencing centers.

References

1. J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In OSDI '04: 6th Symposium on Operating Systems Design and Implementation, 2004.
2. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25(14):1754—1760, 2009.
3. Luca Pireddu, Simone Leo, and Gianluigi Zanetti. Mapreducing a genomic sequencing workflow. In Proceedings of the second international workshop on MapReduce and its applications, MapReduce '11, pages 67–74, New York, NY, USA, 2011. ACM.
4. Luca Pireddu, Simone Leo, and Gianluigi Zanetti. Seal: a distributed short read mapping and duplicate removal tool. Bioinformatics, 27(15):2159–2160, 2011.

Relevant Web sites

5. http://biodoop-seal.sourceforge.net/
6. http://hadoop.apache.org/

The Vital-IT HPC and the Swiss-Prot group

Laurent Falquet
Vital-IT, Swiss Institute of Bioinformatics, Genopode-UNIL, Lausanne, Switzerland
http://www.vital-it.ch/

Biomedical research requires increasing computing power to analyse the huge amounts of data researchers accumulate using high-throughput technologies. However computing power itself is not sufficient, the joint knowledge and expertise of qualified bioinformaticians, statisticians, and IT specialists is essential to provide an efficient support to large-scale projects in biology. Vital-IT is a High Performance Center dedicated to support biological projects within Switzerland. In conjunction with the Swiss-Prot group in Geneva, it forms a unique entity providing both infrastructure and a set of experts in all fields required by modern biology projects. A few examples of genome assembly projects are presented.

References

1. Wurm et al., The genome of the fire ant Solenopsis invicta. PNAS 2011 Apr 5;108(14):5679-84. PMID: 21282665.
2. Andres-Barrao et al., Genome sequences of the high-acetic acid-resistant bacteria Gluconacetobacter europaeus LMG 18890T and G. europaeus LMG 18494 (reference strains), G. europaeus 5P3, and Gluconacetobacter oboediens 174Bp2 (isolated from vinegar). J Bacteriol. 2011 May;193(10):2670-1. PMID: 21441523.
3. Calderon et al., The Mycoplasma conjunctivae genome sequencing, annotation and analysis. BMC Bioinformatics. 2009 Jun 16;10 Suppl 6:S7. PMID: 19534756.

Relevant Web sites

4. http://www.vital-it.ch/
5. http://www.isb-sib.ch/

Demos

Linking research data with scholarly publications

Teresa K Attwood [1,2], Philip McDermott [1,2], James Marsh[3], Steve R Pettifer[2], David Thorne[3]

[1]School of Computer Science, University of Manchester, Manchester, UK

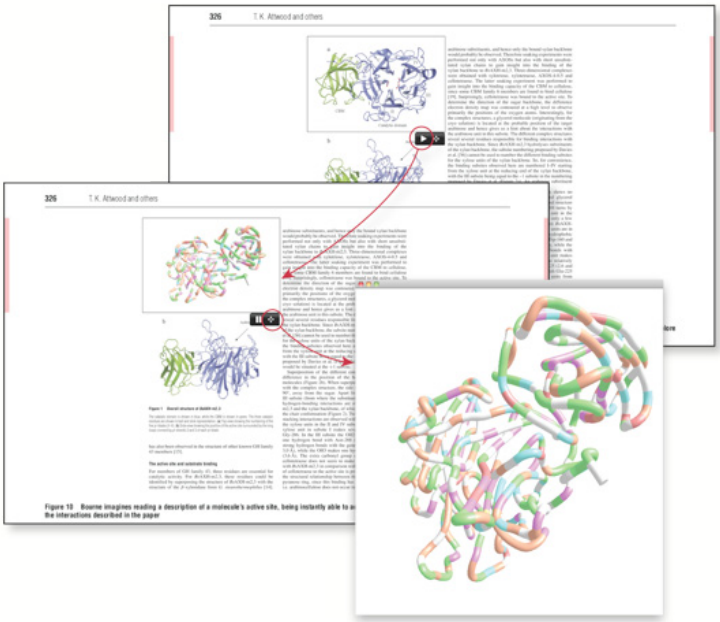
[2]Faculty of Life Sciences, University of Manchester, Manchester, UK

[3]School of Chemistry, University of Manchester, Manchester, UK

http://utopia.cs.man.ac.uk/

Motivation: in recent decades, a vast gulf has opened between the mass of accumulating research data and the massively expanding literature describing and analysing those data. The problem is not so much data generation per se, but rather, the way in which we've buried the knowledge embodied in those data: there is now so much information available that we simply no longer know what we know, and finding what we want is hard, because he knowledge we seek is often spread across thousands of databases and millions of articles in thousands of journals. The intellectual energy required to search this array of archives, and the time and money this wastes, has prompted the development of new software tools to help link these resources, and ultimately liberate the knowledge that's been systematically trapped within them.

Results: to address some of these issues, we have developed Utopia Documents. Building on Utopia, a suite of semantically integrated protein sequence/structure visualisation and analysis tools (1,2), Utopia Documents is a PDF reader that integrates Utopia's functionality with research articles. The system was piloted in a project with Portland Press to create the Semantic Biochemical Journal (BJ) (3) – in this project, Utopia Documents was used to transform static document features into objects that can be linked, annotated, visualised and analysed interactively, thereby transforming the reading experience and making further analysis from within a PDF file possible for the first time. The Semantic BJ was launched in December 2009 (see www.biochemj.org/bj/424/3/), and Utopia Documents is now being used by BJ editors within their routine publication pipelines. With support from other publishers, and groups like SeqAhead, this new software could also make significant advances towards tighter coupling of NGS literature and data in future.


326
T.K. Attwood and others
326
T.K. Attwood and others
Figure 10 Bourne imagines reading a description of a molecule's active site, being instantly able to a
the interactions described in the paper

References

1. Attwood, T.K. et al. (2010) Utopia Documents: linking scholarly literature with research data. Bioinformatics, 26, i568-i574.
2. Pettifer, S.R. et al. (2004) UTOPIA - User-friendly Tools for OPerating Informatics Applications. Comparative and Functional Genomics, 5, CFG359.
3. Attwood, T.K. et al. (2009) Calling International Rescue - knowledge lost in literature and data landslide! Biochemical Journal, 424(3), 317-333.

Relevant Web sites

4. http://getutopia.com/
5. http://utopia.cs.man.ac.uk/utopia/

Posters

Algorithm for error detection in metagonomics NGS data

Dimitar Vassilev[1,] Milko Krachunov[2], Ivan Popov[1], Elena Todorovska[1], Valeria Simeonova[2], Pawel Szczesny[3,4], Pawel Siedlecki[3,4], Urszula Zelenkiewicz[3], Piotr Zelenkiewicz[3]

[1]Bioinformatics group, AgroBioInstitute, Sofia, Bulgaria

[2]Faculty of Mathematics and Informatics, Sofia University "St.Kliment Ohridski", Bulgaria

[3]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland

[4]Institute of Experimental Plant Biology and Biotechnology, University of Warsaw, Poland

http://www.abi.bg/

http://www.ibb.waw.pl/

Because of the nature of metagenomics data, it is neither possible to resample the data to account for the sequencing errors that inevitably occur, nor it is possible to clearly differentiate between an error and a biological variation [6]. Small errors in the sampled data often lead to significant changes in the results of any further analyses and studies based on the data, for example during the construction of phylogenetic trees or during the evaluation of the biological diversity in the sampled environment [2]. For improving the quality of such studies, it is essential that an approach for detecting probable errors is devised.

There are numerous published methods for error detection and correction in NGS data, however none of them are designed to work with metagenomics data, but instead focus on applications such as de novo sequencing of genomes where the appearance of biological variations that are undistinguishable from the errors is not an issue [1,2,3,4]. An example of such software is SHREC (used as a point of reference in this study), which corrects errors in short-read data using a generalized suffix tree [5].

The input data for the initial tests consists of tens of thousands of 16S RNA short-reads with lengths between 300 and 500 bases. For the proposed method to be applied, the read sets need to be filtered of obvious noise and then aligned to each other.

The basic idea behind error correction is that if a given a bit of data, such as a single base, appears too rare in the dataset it is more likely for it to be an error than a biological variation (SNP). A threshold defining "too rare" can be established using the error rate of the sequencing equipment. Higher weights assigned to reads that are locally more similar to the read in question can improve the error recognition by excluding irrelevant data from species that have diverged. . The outline of our evaluation algorithm is as follows:

1. we go over the reads evaluating each base individually;
2. for each base in question, we create a window containing the base at its centre;
3. we calculate a similarity score between the read in question and every other read in the dataset within that particular window. The score excludes the evaluated base, while the bases closest to it are assigned the highest weights;
4. we calculate an evaluation score for the base by calculating a frequency weighted with the similarity score. The result is the ratio of the sum of the similarity scores for the reads that contain the base and the sum of the similarity scores for all the reads;
5. we compare the score of the base to a threshold that has been calculated in advance and experimentally verified. Any scores below the threshold are considered errors and the bases are replaced with the base candidate that would score most using the outlined algorithm.

The biggest challenge in the implementation of this approach is the pre-processing of the data, i.e. the sequence alignment. It is both a difficult and resource intensive task. Trading alignment accuracy for speed is not desirable as alignment errors affect both the evaluation and any further studies.

References

1. Chaisson MJ, Pevzner PA. (2007) Short read fragment assembly of bacterial genomes. Genome Research 18:324-330.
2. Flicek P., Brudno M. (2009) Sense from sequence reads: methods for alignment and assembly. Nature Methods Supplement 6(11) S6-S11.

An Integrated RNA-seq Atlas of the Murine T-Helper Cell Transcriptome

Andrew Deonarine
MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK
http://www.mrc-lmb.cam.ac.uk/tcb/

T-helper cells play an important role in mediating the immune response, and with the advent of next generation sequencing, significant insights can be gained into the T-helper cell transcriptome. One of the barriers to analyzing next-generation sequencing data, such as that generated by RNA-seq analyses, is that many of the statistical properties concerning quantification (ie. RPKM [1] vs. FPKM [2]), normalization [3], and differential expression (using methods such as edger [4], DESeq [5], and Cuffdiff [6]) of the data are still not clearly understood. Building on previous investigations into the bimodality of transcript expression [7], a computational pipeline was created to integrate various methods of expression quantification, cell type clustering, differential expression analyses, gene annotation methods, and novel transcript identification into a murine T-helper cell expression atlas. By integrating these various analyses, it was possible to identify key signature genes (transcription factors, cytokines, receptors, and other molecules) that distinguish the various T-helper cell types from each other, in addition to novel transcripts. This expression atlas, which is easily accessible as a user-friendly online resource at http://www.thelpercell.com, will form the basis for future investigations into immune regulation and function using network-based analyses.

This work is relevant to the goals of SEQAHEAD because it represents a major step forward in the integration and comparison of various methods of expression quantification, differential expression analysis, and annotation of RNA-seq data. The computational principles presented here could potentially be applicable to many other fields of molecular biology and medicine.

References

1. Mortazavi, A., Williams, BA., McCue K., Schaeffer, L., Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* (2008) 5: 621-8.
2. Roberts, A., Trapnell, C., Donaghey, J., Rinn, JL., Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* (2011) 12: R22.
3. Robinson, MD., Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* (2010) 11: R25.
4. Robinson, MD., McCarthy, DJ., Smyth, GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (2010) 26: 139-40.
5. Anders, S., Huber, W. Differential expression analysis for sequence count data. *Genome Biol* (2010) 11: R106.
6. Trapnell, C. Cufflinks Manual. Downloaded from http://cufflinks.cbcb.umd.edu/manual.html on Sept. 12th, 2011.
7. Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenarrden, A., Teichmann, SA. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol* (2011) 7: 497.

Relevant Web sites

8. http://www.thelpercell.com

bcbio-nextgen: Automated, distributed next-gen sequencing pipeline

Roman Valls Guimera, Science for life genomics staff, Brad Chapman

Harvard School of Public Health, Bioinformatics Core; Cambridge, MA USA Science for Life Laboratory, Stockholm, Sweden

http://www.hsph.harvard.edu/research/bioinfocore/

http://www.scilifelab.se/

bcbio-nextgen is an Python framework for next generation sequencing analysis. The fully automated pipeline interacts with the sequencing machine, runs sequences through configurable processing pipelines, and uploads the data into Galaxy for visualization and additional processing. The variant calling analysis pipeline handles alignment to a reference genome, variant identification with GATK and preparation of summary PDF files for assessing run quality.

The pipeline is fully distributed and will run on single multicore machines or in compute clusters managed by LSF, SGE or SLURM. The CloudBioLinux and CloudMan projects utilize this pipeline for distributed analysis on Amazon cloud infrastructure.

The Galaxy web-based analysis tool can be optionally integrated with the analysis scripts. Tracking of samples occurs via a web based LIMS system, and processed results are uploading into Galaxy Data Libraries for researcher access and additional analysis.

Relevant Web sites

1. http://bcbio.wordpress.com/2011/01/11/next-generation-sequencing-information-management-and-analysis-system-for-galaxy/
2. http://bcbio.wordpress.com/2011/09/10/parallel-approaches-in-next-generation-sequencing-analysis-pipelines/
3. https://github.com/brainstorm/bcbb/tree/master/nextgen
4. http://www.slideshare.net/chapmanb/developing-distributed-analysis-pipelines-with-shared-community-resources-using-cloudbiolinux-and-cloudman

BioinformaticsTools@bioacademy.gr

Athanasia Pavlopoulou, Sophia Kossida
Biomedical Research Foundation, Academy of Athens, Athens, Greece
http://www.bioacademy.gr/bioinformatics/

We are presenting some selected computational platforms developed in Dr Kossida's laboratory in the Biomedical Research Foundation, Academy of Athens. These in-house developed software tools include: SAFE1 for the analysis of gene fusion events; GIBA2 for detecting protein complexes; Brukin2D3 for 2D visualization and comparison of LC-MS data; GOmir4 for microRNA target analysis and gene ontology clustering.

References

1. Tsagrasoulis D, Danos V, Kissa M, Trimpalis P, Koumandou VL, Karagouni AD, Tsakalidis A, Kossida S. In press. SAFE Software and FED database to uncover protein-protein interactions using gene fusion analysis.
2. Moschopoulos CN, Pavlopoulos GA, Schneider R, Likothanassis SD, Kossida S. 2009. GIBA: a clustering tool for detecting protein complexes. BMC Bioinformatics 10 Suppl 6:S11.
3. Tsagkrasoulis D, Zerefos P, Loudos G, Vlahou A, Baumann M, Kossida S. 2009. 'Brukin2D': a 2D visualization and comparison tool for LC-MS data. BMC Bioinformatics 10 Suppl 6:S12.
4. Roubelakis MG, Zotos P, Papachristoudis G, Michalopoulos I, Pappa KI, Anagnou NP, Kossida S. 2009. Human microRNA target analysis and gene ontology clustering by GOmir, a novel stand-alone application. BMC Bioinformatics 10 Suppl 6:S20

Relevant Web sites

5. http://www.bioacademy.gr/bioinformatics/projects/ProteinFusion/index.htm
6. http://www.bioacademy.gr/bioinformatics/projects/GIBA/
7. http://www.bioacademy.gr/bioinformatics/Brukin2d/index.html
8. http://www.bioacademy.gr/bioinformatics/projects/GOmir/bioinformatics _ home.htm

Digital gene expression data, cross-species conservation and noncoding RNA

Nicolas Philippe, Florence Ruffle, Elias Bou-Samra, Anthony Boureux, Thérèse Commes, Eric Rivals

Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS, équipe MAB, Université de Montpellier II, Montpellier, France,

CRBM, UMR 5237 CNRS, Montpellier, France

http://www.lirmm.fr/~rival

Recently developed sequencing technologies offer massively parallel production of short reads and become the technology of choice for a variety of sequencing-based assays, including gene expression. Among them, digital gene expression analysis (DGE), which combines generation of short tag signatures for cellular transcripts with massively parallel sequencing, offers a large dynamic range to detect transcripts and is limited only by sequencing depth. As recently described (Philippe et al, 2009), tag signatures can easily be mapped to a reference genome and used to perform gene discovery. This procedure distinguishes between transcripts originating from both DNA strands and categorizes tags corresponding to protein coding gene (CDS and 3'UTR), antisense, intronic or intergenic transcribed regions. Here, we have applied an integrated bioinformatics approach to investigate tags' properties, including cross-species conservation, and the ability to reveal novel transcripts located outside the boundaries of known protein or RNA coding genes. We mapped the tags from a human DGE library obtained with Solexa sequencing, against the human, chimpanzee, and mouse genomic sequences. We considered the subset of uniquely mapped tags in the human genome, and given their genomic location, determined according to Ensembl if they fall within a region annotated by a gene (CDS, UTR and intron) or an intergenic region. We found that 76.4 % of the tags located in human also matched to the chimpanzee genome. The level of conservation between human and chimpanzee varied among annotation categories: 85 % of conserved tags in the CDS, 81 % in the UTR, 76% and 73% respectively in intron and intergenic regions. With the same procedure applied to human and mouse, we obtained 11% of conserved tags in the CDS, 7% in UTR, 1% and 3% respectively in intronic and intergenic regions. We analysed in depth the common CDS and UTR tags in human and mouse for their functional relevance: 90% of them correspond to orthologous genes with a common HUGO. We used DAVID database to extract biological features, the gene clustering revealed specific molecular functions belonging to transcription cofactor and regulator activity, nucleotide binding, ligase and protein kinase, hormone receptor, histone methyltransferase or GTPase activity, and also important signaling pathways like WNT pathway. Indeed, intergenic transcription includes mainly new, non protein-coding RNAs (npcRNAs), which could represent an important class of regulatory molecules. By integrating also SAGE genie and RNA-seq expression data, we selected intergenic tags conserved across species and assayed experimentally the npcRNA transcriptome with Q-PCR. We validated 80% of the 32 tested biological cases. These results demonstrate that considering tag conservation helps to identify conserved genes and functions, which is of great relevance when investigating expressed tags located in intergenic regions.

References

1. N. Philippe, A. Boureux, L. Bréhèlin, J. Tarhio, T. Commes, E. Rivals (2009). Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. Nucleic Acids Research (NAR) Vol. 37, No. 15 e104, doi:10.1093/nar/gkp492; 2009.

Relevant Web sites

2. http://www.atgc-montpellier.fr/ngs/

Epigenomic and transcriptional effects of Dnmt3b mutations in human ICF syndrome-derived B cell lines

Sole Gatto[1,2], Claudia Angelini[2], Sylwia Leppert[1], Valentina Proserpio[3], Sarah Teichmann[3], Maurizio D'Esposito[1], Maria R. Matarazzo[1]

[1]Institute of Genetics and Biophysics "ABT", CNR, Napoli, Italy

[2]Istituto per le applicazioni del calcolo "M. Picone", CNR, Napoli, Italy

[3]MRC Laboratory of Molecular Biology, Cambridge, UK

http://www.igb.cnr.it/

Immunodeficiency, Centromeric region instability, Facial anomalies (ICF) syndrome (OMIM 242860), is a human autosomic recessive disease due to mutations in the Dnmt3b gene, characterized by inheritance of aberrant patterns of DNA methylation and heterochromatin defects (1). How mutations in Dnmt3B and the resulting deficiency in DNA methyltransferase activity result mainly in immunodeficiency has not been clarified yet. Patients show variable agammaglobulinemia and a reduced number of T cells, making them prone to infections and death before adulthood. It is already known that the expression of several genes and microRNAs is deregulated in ICF lymphoblastoid cell lines (LCLs), being both up- and down-regulated (2,3). Surprisingly, subtle but significant reduction of promoter methylation was seen in only few analyzed upregulated genes and approximately half of them were marked with loss of repressive histone modifications, particularly H3K27 trimethylation, and gain in transcriptionally active H3K9 acetylation and H3K4 trimethylation marks, while an extensive change of histone modifications of upregulated miRNAs was always observed.

It is clear that Dnmt3B mutations affect not only DNA methylation, but also several other expression regulators. In order to assess to what extent these mutations affect the epigenetic landscape of the whole genome we examined the global DNA methylation profile using the Infinium assay from Illumina, the genome-wide mapping of 3meK4H3, 3meK27H3 and RNA Polymerase II (Pol II) by chromatin immunoprecipitation-sequencing (ChIP-seq) and correlated those to mRNA transcriptome (obtained by RNA-seq) and to microRNA expression (by previous microarray results) in ICF and control LCLs. We found a positive correlation between active genes, binding of Pol II and 3meK4H3 binding and an opposite correlation with 3meK27H3 binding and DNA methylation as expected. Moreover, we identified several regions of interest, which are differentially enriched between the patient and the controls. The complete results will be shown in the poster.

Beyond its relevance to ICF syndrome, by addressing the impaired DNMT3B functions in abnormal epigenome cases and how these reflect to the transcriptomes of the affected cells, these data will provide new insights in the field, unravelling the physiological contribution of DNMT3B to the epigenetic network.

References

1. Matarazzo MR, De Bonis ML, Vacca M, Della Ragione F, D'Esposito M (2009) Int J Biochem Cell Biol 41 (1):117-126.
2. Jin B, Tao Q, Peng J, Soo HM, Wu W, Ying J, Fields CR, Delmas AL, Liu X, Qiu J, Robertson KD (2008). Hum Mol Genet 17 (5):690-709.
3. Gatto S, Della Ragione F, Cimmino A, Strazzullo M, Fabbri M, Mutarelli M, Ferraro L, Weisz A, D'Esposito M, Matarazzo MR (2010). Epigenetics 5 (5):427-443.

EU COST Action TD0801: Statistical Challenges On The 1000 Euro Genome Sequences In Plants

Marco C.A.M. Bink[1], **Thomas Schiex**[2]

[1]Biometris Wageningen UR, Wageningen, The Netherlands [2]MIA - INRA Chemin de Borde Rouge, Castanet-Tolosan, France

http://www.statseq.eu/

New DNA sequencing technologies either currently available or under development will eventually enable eukaryotic genomes to be sequenced for less than a thousand euros. This technology-push will have a major impact on plant genomics and biological research and lead to a dramatic expansion in both the availability of sequence data and the range of sequence based applications. New innovative techniques are required to unlock the information contained in the sequence data and to apply the acquired knowledge for plant science and crop improvement. The wide variety and often unique characteristics of plant genomes pose additional challenges and opportunities. The need for and the dissemination of efficient strategies for handling and analysing high throughput sequence data in plants requires cooperation at the international level to develop new approaches analytical tools and share best practice. This COST Action will establish a network of researchers that coordinate, focus and strengthen national and pan-European statistical genomics and bioinformatics. It will be built on close interactions with other disciplines such as genetics, genomics and breeding. The Working Groups will arrange workshops, Short Term Scientific Missions, a website and Wiki, training courses, and publications to disseminate aims and achievements.

Relevant Web sites

1. www.statseq.eu (COST Action TD0801)
2. https://colloque.inra.fr/statseq _ 2011/ (3rd StatSeq workshop, Toulouse 2011)
3. www.bioinf.boku.ac.at/statseq (WG1 meeting on RNA seq, Vienna 2011)

From cutadapt to sequencetools (sqt): a versatile toolset for sequencing projects

Marcel Martin, Sven Rahmann

Bioinformatics, Computer Science XI, TU Dortmund, 44221 Dortmund, Germany

Genome Informatics, Institute of Human Genetics, Faculty of Medicine, University of Duisburg-Essen, Essen, Germany

http://www.rahmannlab.de/

We are developing a suite of scriptable tools for both small and large typical tasks arising in high-throughput sequencing projects. Following the nix philosophy, each tool has a specific task, and power and flexibility come from the ability to combine these tools in various ways.

As an example, we present cutadapt in details: When small RNA is sequenced on current sequencing machines, the resulting reads are usually longer than the RNA and therefore contain parts of the 3' adapter. That adapter must be found and removed error-tolerantly from each read before read mapping. Previous solutions are either hard to use or do not offer required features, in particular support for color space data. As an easy to use alternative, we developed the command-line tool cutadapt, which supports 454, Illumina and SOLiD (color space) data, offers two adapter trimming algorithms, and has other useful features.

This, and other tools, are presently organized in a toolset that will be available under the name sqt. We will briefly outline the design idea of this set of tools and report on the current state of development.

References

1. Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads EMBnet. journal 17(1), May 2011.

Relevant Web sites

2. Cutadapt, including its MIT-licensed source code, is available at http://code.google.com/p/cutadapt/
3. Sqt website: http://code.google.com/p/sqt/

Improved analysis of fungal communities using the next-generation-sequencing analysis of rpb2 genes

Tomas Vetrovsky, Jana Voriskova, Lucia Zifcakova, Michaela Urbanova, Petr Baldrian

Laboratory of Environmental Microbiology, Institute of Microbiology of the ASCR, Prague, Czech Republic.

http://www.biomed.cas.cz/mbu/lbwrf

Current exploration of the ecology of soil fungal and bacterial communities and microbe-catalyzed processes in soils largely rely on community composition analysis using next-generation-sequencing of PCR amplicons (1). Typically, the relative abundance of individual members of microbial communities are derived from the analyses of 16S rRNA region of prokaryotic microorganisms and 18S rRNA or internal transcribed spacer (ITS) region of the rDNA for fungi and other microeukaryots. The analysis of fungal ITS sequences is helpful tool for molecular systematics at the species level, and even within species, but the quantitative information on the relative abundance of individual taxa is skewed due to the presence of multiple rDNA gene copies per genome, ranging from 10 to 200 (2). On the other hand, it was demonstrated that there is a group of genes like the elongation factor-1 alpha (*tef1*) or RNA polymerase II second largest subunit (*rpb2*) that are consistently present in one copy per fungal genome and exhibit sufficient variation to be used for phylogenetic analysis and taxonomic assignment (3). The use of such genes offers the possibility to directly count fungal genomes and improve the knowledge on the relative importance of individual taxa of fungi in the environmental processes. Here we demonstrate that the amount of ITS copies per nanogram DNA shows high variation among soil basidiomycetes and even closely related species largely differ in this respect. We also demonstrate that the use of the *rpb2* gene is applicable for analysis of soil fungal communities and that the data derived using this molecular marker are largely different from those based on the amplicon sequencing of the ITS. Although the phylogeneti discriminative power of the *rpb2* gene is limited, it still offers a suitable tool to infer fungal taxonomy at least on the level of families.

References

1. Baldrian et al. (2011) ISME Journal in press, doi:10.1038/ismej.2011.95.
2. Corradi, et al. (2007) Applied and Environmental Microbiology 73, 366-369.
3. Matheny et al. (2007), Molecular Phylogenetics and Evolution 43, 430–451.

In the Shadow of the Genome: A Challenging Journey to Diversity in Leishmania donovani

Hideo Imamura[1], An Mannaert[1], Tim Downing[2], Matthew Berriman[2], Jean-Claude Dujardin[1]

[1]Institute of Tropical Medicine, Antwerp, Belgium;

[2]Wellcome Trust Sanger Institute, Hinxton, UK

http://www.itg.be/

Leishmaniasis is a disease complex caused by protozoan parasites of the genus *Leishmania*, which are transmitted by sandflies. It affects 350 million people worldwide, but the most severe form, visceral leishmaniasis (VL), is most prevalent in India, Nepal, Bangladesh, Sudan and Brazil. In the Indian subcontinent, VL is caused by *Leishmania donovani*, and efficient treatment is highly challenged by the emergence of drug resistance. We are running two projects – Kaladrug (2) and GeMInI (3) – to characterize the genetic and phenotypic diversity of *L. donovani* in India and Nepal using comparative genomics and metabolomics in order to identify genetic and metabolic signatures associated with drug resistance. Both omics represent two extremes, from genotype to phenotype. The ultimate goal is to identify the factors that lead to the different clinical phenotypes (cure versus treatment failure). Therefore, different strains have been and are being subjected to whole-genome sequencing and metabolic profiling. Sequencing, genome assembly and comparative analyses are performed in collaboration with the Parasite Genomics group of the Wellcome Trust Sanger Institute. One phenotypically well-characterized *L. donovani* strain was chosen as a reference and a *de novo* draft genome sequence was generated with 454 and Illumina sequencing technologies, resulting in a 35 Mbp genome distributed over 36 chromosomes. Approximately 50 additional strains with known in vitro drug susceptibility from VL patients with differential response to treatment are sequenced and analyzed to identify natural variation. Despite high sequence conservation and thus a limited number of single nucleotide polymorphisms (SNPs), a substantial amount of structural variation has been observed among the different strains (1). The chromosome ploidy is highly variable between different strains, such that all strains examined so far have a different chromosome content, and contractions and expansions of tandem gene repeats appear to occur frequently. In addition, we observed extra-chromosomal amplification of three gene loci. These structural polymorphisms can result in a change in the gene dosage and can have an effect on the metabolome and thus the phenotype of the parasite. Preliminary genome analyses identified a number of SNPs and structural changes that may contribute to the resistant phenotype, and the first metabolome analyses of the same samples revealed a significant difference in metabolites between drug susceptible and drug resistant strains. The amount of information generated by next-generation sequencing and other technologies is growing, as is the challenge to process and interpret the increasing amount of data. The major task here is the integration of the information coming from both ends of the omics chain in order to understand how complex biological traits, such as drug resistance, are acquired.

References

1. Downing, Imamura et al. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Research, in press.

Relevant Web sites

2. http://www.leishrisk.net/kaladrug
3. http://www.leishrisk.net/gemini

IT Future of Medicine: Next Generation Sequencing is the Key to Future Personalized Medicine

Hans Lehrach and Babette Regierer for the IT Future of Medicine Consortium
Max Planck Institute for Molecular Genetics,
Berlin, Germany
http://www.itfom.eu/

The IT Future of Medicine (ITFoM) initiative will produce computational models of individuals to enable the prediction of their future health risks, progression of diseases and selection and efficacy of treatments while minimizing side effects. As one of six Future and Emerging Technologies (FET) Flagship Pilot Projects funded by the European Commission, ITFoM will foster the integration of technology development in functional genomics and computer technologies to enable the generation of patient models to make them available for clinical application. The realization of the patient model is based on the recent breakthroughs in sequencing technology that enables the high-throughput analysis of a large number of individual genomes and transcriptomes. The genome profile will be integrated with proteome and metabolome information generated *via* new powerful chromatography, mass spectrometry and nuclear magnetic resonance techniques. Computational and mathematical tools enable the development of systems approaches for deciphering the functional and regulatory networks underlying the complex biological systems and form the basis for the future patient model.

The recent increases in the capacity of next-generation sequencing systems will provide huge amounts of genome, epigenome and transcriptome data, making it feasible to apply deep sequencing in the clinic to characterize not only the patient's genome, but also individual samples e.g. from tumors. The –omics information will provide the basis to establish integrated molecular, physiological and anatomical models of every individual in the health care system. The first approach to the "Virtual Patient" modeling system that has been generated at the Max Planck Institute for Molecular Genetics combines general information available about cancer relevant pathways with the individual tumor/patient information (genome, transcriptome). This individualized model will not only be able to analyze the current situation, but will allow the prediction of the response of the patient to different therapy options or intolerance for certain drugs.

IT Future of Medicine will have long lasting beneficial effects for medicine of the future offering new routes to improve clinical practice, reduce health care costs but also to accelerate the development and the approval process for new drugs.

IT Future of Medicine is an initiative of more than 50 academic and industrial partners from over 15 countries to set up a research concept for the development of the "virtual patient".

References

1. Manolopoulos VG, Dechairo B, Huriez A, Kühn A, LLerena A, van Schaik RH, Yeo K-TJ, Ragia G, and Siest G (2011): *Pharmacogenomics* and personalized medicine in clinical practice Pharmacogenomics 12(5):597-610. doi:10.2217/pgs.11.14.
2. Daskalaki A, Wierling C, Herwig R (2009): Computational tools and resources for systems biology approaches in cancer. In Computational Biology - Issues and Applications in Oncology, Series: Applied Bioinformatics and Biostatistics in Cancer Research, Pham, Tuan (Ed.), Springer, New York Dordrecht Heidelberg London. 2009:227-242.

Relevant Web sites

4. http://www.itfom.eu/
5. http://www.fet-f.eu/
6. http://www.molgen.mpg.de/research/lehrach/

Massive-scale RNA-Seq experiments in human genetic diseases

Valerio Costa[1], Marianna Aprile[1], Roberta Esposito[1], Maria Rosaria Ambrosio[1], Margherita Scarpato[1], Carmela Ziviello[1], Italia De Feis[2], Claudia Angelini[2] and Alfredo Ciccodicola,

[1]CNR, Institute of Genetics and Biophysics "A. Buzzati-Traverso" (IGB), Naples, Italy

[2]CNR, Istituto per le Applicazioni del Calcolo (IAC),
Naples, Italy

http://www.igb.cnr.it

Since 2008, our research group is actively working in the field of NGS, with particular attention to RNA-Seq as innovative approach to understand cells' transcriptome in disease states (Costa et al., 2010a). In particular, combining molecular biology and computational expertise, we have recently analysed (Costa et al., 2011) by RNA-Seq - for the first time in Down syndrome (DS) - the global transcriptome of endothelial progenitor cells (EPCs), morphologically and functionally impaired in DS (Costa et al., 2010b). After rRNA depletion - followed by strand specific sequencing - we measured expression from (even) low expressed genes, we identified new regions of active transcription outside annotated loci, novel splice isoforms and extended untranslated regions for known genes, potentially new microRNA targets or regulatory sites. However, although RNA-Seq provided a huge amount of useful data for DS, showing a genome-wide alteration of gene expression (not limited to HSA21 genes), the experiment revealed only a fraction of the underlying complexity, giving no information about the reasons of such global deregulation. Therefore, in this ongoing project we aim to study: 1) by ChIP-Seq, the binding maps of some (preliminarily selected) transcription factors (TFs), key players in gene expression modulation, and 2) by RNA-Seq, the related gene expression changes in the same cells. ChIP-Seq, combining standard chromatin immunoprecipitation and massively parallel sequencing, allows to identify DNA sequences bound by TFs in vivo, helping to decipher gene regulatory networks (Park 2009). We believe that integrating RNA- and ChIP-Seq data would provide much more biological insights into gene expression regulation in DS cells, helping us to better understand some blood-related pathological aspects of the syndrome.

Our group is also participating to a large-scale collaborative industrial project aimed to develop a diagnostic kit for personalized therapeutic strategies in type 2 diabetic (T2D) patients resistant to conventional drug therapies. In particular, to elucidate some mechanisms of drug resistance, our group will perform massive-scale transcriptome analysis by RNA-Seq in a well-selected subset of individuals (~50), also collaborating with bioinformaticians to further data analysis.

In the light of these considerations, and given the objectives of the COST Action BM1006, our group will contribute to the goals of the SEQAHEAD project by actively integrating in the newborn European network of NGS, providing its expertise in sequencing technologies with a particular contribution (protocols, experimental data and pipelines for data analysis) to the RNA-Seq.

References

1. Costa V, Angelini C, De Feis I, Ciccodicola A. (2010) "Uncovering the complexity of transcriptomes with RNA-Seq." J Biomed Biotechnol. 2010:853916.
2. Costa V, Angelini C, D'Apice L, Mutarelli M, Casamassimi A, Sommese L, Gallo MA, Aprile M, Esposito R, Leone L, Donizetti A, Crispi S, Rienzo M, Sarubbi B, Calabrò R, Picardi M, Salvatore P, Infante T, De Berardinis P, Napoli C, Ciccodicola A. (2011) "Massive-scale RNA-Seq analysis of non ribosomal transcriptome in human trisomy 21." PLoS One. 6(4):e18493.
3. Costa V, Sommese L, Casamassimi A, Colicchio R, Angelini C, Marchesano V, Milone L, Farzati B, Giovane A, Fiorito C, Rienzo M, Picardi M, Avallone B, Marco Corsi M, Sarubbi B, Calabrò R, Salvatore P, Ciccodicola A, Napoli C. (2010) "Impairment of circulating endothelial progenitors in Down syndrome." BMC Med Genomics. 3:40.
4. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009 Oct;10(10):669-80. Epub 2009 Sep 8.

Oncogenomics of the Hormone-responsive Breast Cancer Phenotype by NGS

Alessandro Weisz
Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, Molecular Pathology and Medical Genomics Unit, S. Giovanni di Dio e Ruggi d'Aragona University Hospital, University of Salerno, Salerno, Italy
http://www.labmedmolge.unisa.it/

Breast cancer (BC) comprises an heterogeneous group of diseases characterized by different biological history, clinical phenotype and responsiveness to therapy. Among the factors that contribute substantially to breast carcinogenesis and tumor progression, ovarian hormones – in particular estrogen – have long been known to play a pivotal role. In the mammary gland these steroid hormones control differentiation and growth via two intracellular receptors, ERα and β, which are ligand-dependent transcription factors belonging to the nuclear receptor family of transcriptional regulators. Upon hormone binding, ERs bind to multiple sites in chromatin of BC cells and thereby act at gene and epigene level to exert a direct control on specific genetic networks driving proliferation, survival and differentiation status of the cell. In breast tumors, elevated levels of ERα and reduced levels of ERβ are observed from the early stages of the disease, suggesting a dual role for these regulatory factors in breast cancer initiation and progression, with ERα exerting an oncogenic role and ERβ oncossuppressive functions. Specific gene expression patterns mark the clinical and pathological status of BC lesions and its responsiveness to pharmacological treatments, including hormone therapy with anti-estrogens. The functional relationships between the two ER subtypes and genes characterizing the different clinical phenotypes of breast cancer are not known and are the main focus of our research. To this end, our laboratory is implementing several genome-wide analytical approaches based on the use of NGS and microarrays to investigate ER actions and estrogen signalling in cell models and primary BCs. These studies include: (a) identification of ER and their coregulatory factors interaction with chromatin and mapping of histone marks and other epigenetics codes by ChIP-Seq; (b) quantitative analyses of mono- and multi-allelic CpG island methylation by high-throughput DNA methylation microarrays and NGS (methylated DNA IPP sequencing MeDIP-Seq and MBD-Seq, bisulfite-based MethylC-Seq and RRB-Seq); (c) miRNA and other small RNA profiling by miRNA-Seq; (d) ribonucleoprotein-associated RNA identification by RIP-Seq; (e) genome wide search for gene mutations (transitions/transversions, indels, etc) in primary BCs by exome sequencing.

Interpretation and application to the clinical setting of the results obtained with these global analytical approaches require robust statistical tools and innovative bionformatics analyses, that we are interested to implement also in scientific collaborations to be established within SEQAHEAD.

Power and limits of capture-based, targeted DNA resequencing for mutation detection

Fabrice Lopez, Hélène Holota, François-Xavier Théodule and Jean Imbert
TAGC UMR _ S 928, Inserm, Université de la Méditerranée, Marseille, France
http://www.yourwebsite.org/

The IBiSA TGML platform (Sci.Dir.: Dr. J. Imbert) is integrated to TAGC (Inserm U928, Dir.: Dr C. Nguyen) located on the Science Park of Luminy (Université de la Méditerranée). It offers access for academics and companies to transcriptomic and functional genomic studies. The TGML platform and TAGC provide expertise in the analysis of various types of DNA microarrays and sequencing dataset. Our researchers and engineers actively contribute to the development of new computational tools and data processing pipelines. The TGML platform is member of the France-Génomique network. The high throughput sequencing service (TGML DeepSeq) is equipped with a LifeTech SOLiD4 sequencer that can produce up to 100 Gb (fragments 50 nt). Barcoding allows sequencing up to 256 samples in one run. Upgrade g to SOLiD 5500XL (300 Gb, 75 nt) is scheduled Fall 2011 as well as the purchase of an Ion Torrent PGM machine for a fast and cost-effective sequencing alternative for smaller sized projects. Applications performed as a service for external users or in collaboration include: full exome and targeted DNA resequencing (Homo sapiens, Mus musculus, etc.) with customized capture design on microarrays, ChIP-seq, FAIRE-seq, Mnase-seq (Epigenomics), and some projects of full resequencing for small genomes. Collaborators and clients include teams from CIML, IBDML, CRCM, IAB in Grenoble, a partnership with the GIS Institut GIS Maladies Rares (Paris, Marseille, Dijon, Montpellier, etc.), etc. We are planning to implement shortly: whole transcriptome analysis (lncRNA), SAGE-seq, DNAseI-HS-seq, de novo bacteria sequencing, FAIRE-seq, etc. During the last 2 years we have acquired a robust experience in targeted genomic resequencing and we are currently developing of a new bioinformatics pipeline for the characterization of genomic variants (SNPs and small InDels) and a new Java-based graphic software named GeVarA for Genomic Variant Analyzer.

I will present the power and the limit of these approaches with an emphasis on the challenges faced by the bioinformatician and by our computing and data storage resources, as well as our ongoing solutions.

References

1. Lopez,F., Textoris,J., Bergon,A., Didier,G., Remy,E., Granjeaud,S., Imbert,J., Nguyen,C., and Puthier,D. (2008). TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database. PLoS ONE 3, e4001.
2. Benoukraf,T., Cauchy,P., Fenouil,R., Jeanniard,A., Koch,F., Jaeger,S., Thieffry,D., Imbert,J., Andrau,J.C., Spicuglia,S., and Ferrier,P. (2009). CoCAS: a ChIP-on-chip analysis suite. Bioinformatics 25, 954-955.
3. Pekowska,A., Benoukraf,T., Zacarias-Cabeza,J., Belhocine,M., Koch,F., Holota,H., Imbert,J., Andrau,J.C., Ferrier,P. and Spicuglia,S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J. doi:10.1038/emboj.2011.295.

Relevant Web sites

4. http://tagc.univ-mrs.fr/welcome/spip.php?rubrique1
5. http://tagc.univ-mrs.fr/welcome/spip.php?rubrique2

Statistical approaches for the analysis of RNA-Seq and ChIP-seq data and their integration

Claudia Angelini and Italia De Feis
Istituto per le Applicazioni del Calcolo "Mauro Picone", Naples, Italy
http://www.iac.cnr.it/

The recent introduction of Next-Generation Sequencing (NGS) platforms, able to simultaneously sequence hundreds of millions of DNA fragments, has dramatically changed the landscape of genetics and genomic studies. However, to benefit of this novel sequencing technology, advanced laboratory and molecular biology expertise must be combined with a strong multidisciplinary background in data analysis. In addition, since the output of an experiment consists of a huge amount of data, terabytes of storage and clusters of computers are required to manage the computational bottleneck.

Recently, the Institute of Genetics and Biophysics (IGB) and the Istituto per le Applicazioni del Calcolo (IAC) have started a close collaboration aimed to set up a novel NGS facility in Naples that integrates both the wet laboratory and the bioinformatics core. Therefore, the IGB acquired a SOLiD system (now version 4) and, nowadays it provides all the wet laboratory capabilities and its experience in molecular biology for a wide range of experiments. Our team at IAC provides the experience in the usage and the development of computational methods for their analysis and it is also equipped with a powerful cluster of workstations (http://lilligridbio.na.iac.cnr.it/wordpress/) capable of handling massive computational tasks.

The research activities are directed toward two directions: from one side the effort of our group is devoted to the use of efficient software, the maintenance and development of bioinformatics pipeline for specific applications required by the sequencing facility, on the other hand the scientific interest is also devoted to the development of innovative statistical techniques for the NGS data analysis and to the implementation of novel algorithms using both CPU and GPU systems.

Till now our group has been involved the analysis of a series of independent studies on both RNA-seq and ChIP-seq. The experiments were conducted on the local sequencing facility by dr. Ciccodicola (for the RNA-seq data) and dr. Matarazzo (for the ChIP-seq data) groups at IGB-CNR, which are also members of the SEQAHEAD Cost Action. In this context our ongoing activities are devoted to the implementation of specific pipeline on our local cluster and to the definition of a probabilistic approach to model in terms of "signal plus noise" both transcriptional profiles and chromatin profiles. However, since we believe that integrating ChIP-seq and RNA-seq data is expected to provide much more biological insights for a better understanding of the mechanisms involved in gene expression regulation, rather than using one dataset only, we will focus our attention on the integration of these types of data in a unified statistical framework.

In the light of these considerations our group is aimed to contribute to the goals of the SEQAHEAD project by actively participating to the discussion concerning the development of novel statistical and computational methods for the analysis of RNA-Seq and ChIP-seq data and their integration, and to the development of educational programs on the statistical analysis of NGS data.

References

1. V. Costa, C. Angelini, et al., *Massive-scale RNA-Seq analysis of non ribosomal transcriptome in human trisomy 21*, PLoS ONE 2011.
2. V. Costa, C. Angelini, I. De Feis, A. Ciccodicola. *Uncovering the complexity of transcriptomes with RNA-Seq*. Journal of Biomedicine and Biotechnology vol. 2010, Article ID 853916, 19 pages, (2010).
3. C. Angelini, A. Ciccodicola, V. Costa and I. De Feis. *Analyzing the Whole Transcriptome by RNA-Seq data: the Tip of the Iceberg*, ERCIM NEWS July 2010, Special Theme Computational Biology, pp.16-17. 2010.

TAPYR: An efficient high-throughput sequence aligner for re-sequencing applications

Francisco Fernandes[1,2], Paulo G.S. da Fonseca[2], Luis M.S. Russo[1,2], Arlindo L. Oliveira[1,2], Ana T. Freitas[1,2]

[1]Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento, Lisboa, Portugal

[2]Instituto Superior Tecnico, Universidade Tecnica de Lisboa (IST/UTL), Lisboa, Portugal

During the last two decades most laboratories used Sanger's "shotgun" method in many significant large-scale sequencing projects, being this method considered the 'gold standard' in terms of both read length and sequencing accuracy. Recently, several next generation sequencing (NGS) technologies have emerged, including the GS FLX (454) Genome Analyzer, the Illumina's Solexa 1G Sequencer, the SOLiDTM and the Ion Torrent Systems, which are able to generate three to four orders of magnitude more sequences and are considerably less expensive than the Sanger method. However, the read lengths of NGS technologies create important algorithmic challenges. While the 454 platform (using Titanium technology) is able to obtain reads in the 400-600 base pairs (bp), the Illumina's Solexa 1G Sequencer and the Ion Torrent Systems present reads with an average length of 100 bp and the SOLiD platform is currently limited to 25-50 bp.

Several assembly tools have recently been developed for generating assemblies from short, unpaired sequencing reads. However, the sheer volume of data generated by these technologies (0.4 Gbp/run for the 454 and 16 Gbp/run for the SOLiD), and the need to align reads to increasing large reference genomes limits the applicability of standard methods.

One way to speed up the read alignment task is to resort to software based on approximate indexing technologies. This means that the whole reference genome is scanned while applying a dynamic programming algorithm. Indexed alignment algorithms, which preprocess the reference genome into an index data structure that can then be searched, correspond to more efficient approaches. On one hand it can discard irrelevant portions of the reference genome much more efficiently. On the other hand the computation on relevant regions can be factored out. However, building indexes is time and space consuming. State of the art algorithms are using techniques from a new class of indexes, compressed indexes, which have smaller space requirements by using data compression techniques to eliminate regularities in the indexes.

In this work we present TAPyR (http://www.tapyr.net) a new method for the alignment of NGS reads that uses compressed indexing build an index of the reference genome sequence to accelerate the alignment. Being firstly proposed to handle the 454 GS FLX data, it can also be used with Illumina and Ion Torrent data. Like other algorithms, TAPyR uses in a second stage a multiple seed heuristic to anchor the best candidate alignments. This heuristics has the advantage that it dispenses the need of determining the number and length of the seeds beforehand, relying on the assumption that the optimal alignments are mostly composed of relatively large chunks of exact matches interspersed by small, possibly gapped, divergent regions. At the ultimate stage banded dynamic programming is used to finish up the candidate multiple seed alignments considering user-specified error constraints.

TAPyR was evaluated against other mainstream mapping tools namely BWA-SW, SSAHA2, Segemehl, GASSST, and Newbler. The analyses were performed with real and simulated data sets, with the objective of assessing the efficiency and accuracy of the aforementioned tools in the context of re-sequencing projects. As the results show the new method manages to achieve convincing performance in terms of speed and in terms of the number and precision of aligned reads. In fact, TAPyR has displayed class-leading CPU-time performance and excellent use of input reads in comparison to other mainstream tools.

UPPNEX - A solution for Next Generation Sequencing data management and analysis

Samuel Lampa[1,2], Jonas Hagberg[1], and Ola Spjuth[1,3]

[1]Uppsala Multidisciplinary Center for Advanced Computational Science (SNIC-UPPMAX) , Uppsala, Sweden

[2]Science for Life Laboratory, Uppsala University,Uppsala, Sweden

[3]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

https://www.uppnex.uu.se/

We present a solution for Next Generation Sequencing (NGS) data management and analysis using a cluster-based approach with a shared parallel file system, together with a graphical client and a web-based knowledge base. The initiative is named UPPNEX, and has emerged as the leading platform for the vibrant NGS community in Sweden.

For analysis, 900 000 computing hours per month are available via a cluster of 2784 cores through the SLURM queuing system. For primary storage, more than 420TB of parallel storage are attached locally to the computing resources. For archiving, more than 1 PB of storage is available via the Swedish national long time storage system SweStore. To protect project data, UPPNEX is equipped with snapshots, disaster backup on tape, optional data encryption, and a tight security policy permitting only SSH connections.

To simplify for novice users of HPC systems, we have developed a graphical client for accessing UPPNEX resources based on the Bioclipse workbench1. Bioclipse leverages on the plugin-architecture of Eclipse, which allows for easy extensions and reuse of plugins from a large user community. A proxy component translates information from the local queuing system and exposes a transparent API, which is accessed via a persistent SSH connection provided by the Eclipse Parallel Tools Platform. Via Bioclipse, users can access their files via a graphical file browser, they are able to monitor jobs, inspect file and project quotas, and start new analyses. The Bioclipse-plugin takes advantage of the tool configuration files from the Galaxy platform to provide wizard-based configuration of cluster jobs for common bioinformatics tools, but users can also interact with UPPNEX via a regular terminal. A history view allows for inspecting the commands sent to UPPNEX and enables reuse and sharing of analysis scripts.

Apart from hardware and software, the UPPNEX project has several associated human resources ("system experts" and "application experts") serving the national NGS community with experience and know-how in both HPC and bioinformatics analysis via the UPPNEX Knowledgebase web portal3. The distinct focus on end-users has attracted over 130 projects in only 2 years at an increasing rate, and UPPNEX is currently serving over 400 TB of NGS data with the sequencing of 'Norwegian spruce' as one of its largest projects.

UPPNEX was originally funded by the Knut and Alice Wallenberg foundation and the Swedish National Infrastructure for Computing (SNIC) and is formally part of Uppsala Multidisciplinary Center for Advanced Computational Science4 (SNIC-UPPMAX).

References

1. Spjuth et al. *Bioclipse: an open source workbench for chemo- and bioinformatics*, BMC Bioinformatics 2007, 8:59.
2. Giardine et al. *Galaxy: a platform for interactive large-scale genome analysis*. Genome Res. 2005 Oct;15(10):1451-5.

Relevant Web sites

3. https://www.uppnex.uu.se/
4. http://www.uppmax.uu.se/

Read indexing

Nicolas Philippe, Mikael Salson, Thierry Lecroq, Martine Leonard, Therese Commes, Eric Rivals

Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS, équipe MAB, Université de Montpellier II, Montpellier, France,

LITIS, Univ. Rouen, Mont Saint Aignen, France

CRBM, UMR 5237 CNRS, Montpellier, France

http://www.lirmm.fr/∼rivals

The question of read indexing remains broadly unexplored. However, the increase in sequence throughput urges for new algorithmic solutions to query large read collections efficiently. We propose a solution, named Gk arrays, to index large collections of reads, an algorithm to build the structure, and procedures to query it. Once constructed, the index structure is kept in main memory and is repeatedly accessed to answer various types of queries. We compare our data structure to other possible solutions to investigate its scalability and computational efficiency. Gk arrays are implemented in a general purpose library, which may prove useful for assembly purposes, for evaluating the expression level in RNA-seq, and others high throughput sequencing applications.

References

1. Querying large read collections in main memory: a versatile data structure. N. Philippe, M. Salson, T. Lecroq, M. Leonard, T. Commes and E. Rivals. BMC Bioinformatics, Vol. 12, p. 42, doi:10.1186/1471-2105-12-242, 2011.

Relevant Web sites

2. http://crac.gforge.inria.fr/gkarrays/
3. http://www.atgc-montpellier.fr/ngs/

National Nodes

Argentina
IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata

Brazil
Lab. Nacional de Computação Científica, Lab. de Bioinformática, Petrópolis, Rio de Janeiro

Chile
Centre for Biochemical Engineering and Biotechnology (CIByB). University of Chile, Santiago

China
Centre of Bioinformatics, Peking University, Beijing

Colombia
Instituto de Biotecnología, Universidad Nacional de Colombia, Edificio Manuel Ancizar, Bogota

Costa Rica
University of Costa Rica (UCR), School of Medicine, Department of Pharmacology and ClinicToxicology, San Jose

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network

Greece
Biomedical Research Foundation of the Academy of Athens, Athens

Hungary
Agricultural Biotechnology Center, Godollo

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

Mexico
Nodo Nacional de Bioinformática, EMBnet México, Centro de Ciencias Genómicas, UNAM, Cuernavaca, Morelos

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan
COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

Poland
Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warszawa

Portugal
Instituto Gulbenkian de Ciencia, Centro Portugues de Bioinformatica, Oeiras

Russia
Biocomputing Group, Belozersky Institute, Moscow

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

South Africa
SANBI, University of the Western Cape, Bellville

Spain
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

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

Sweden
Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland
Swiss Institute of Bioinformatics, Lausanne

Specialist- and Assoc. Nodes

CASPUR
Rome, Italy

EBI
EBI Embl Outstation, Hinxton, Cambridge, UK

Nile University
Giza, Egypt

ETI
Amsterdam, The Netherlands

IHCP
Institute of Health and Consumer Protection, Ispra. Italy

ILRI/BECA
International Livestock Research Institute, Nairobi, Kenya

MIPS
Muenchen, Germany

UMBER
Faculty of Life Sciences, The University of Manchester, UK

CPGR
Centre for Proteomic and Genomic Research, Cape Town, South Africa

The New South Wales Systems Biology Initiative
Sydney, Australia

for more information visit our Web site
www.EMBnet.org

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: +46-18-4716696