SEQAHEAD

EU-funded COST action BM1006

Next Generation Sequencing Data Analysis Network

http://www.nextgensequencing.org

MC Business Meeting and Scientific Meeting

Brussels, 7-9 November 2011

A. Table of contents

A. TABLE OF CONTENTS **2**

B. SCHEDULE **5**

Monday, November 7 - Management Committee (MC) business meeting - By invitation only *5*

Tuesday, November 8 – Scientific program meeting *6*

Wednesday, November 9 - Work group meetings *7*

C. PRACTICAL INFORMATION **8**

Expenses *8*

Arrival *8*

Hotel (arrow A on map below) *8*

Meeting place (arrow B on map below) *8*

Conference dinner (arrow C on map below) *8*

Contact *8*

Detailed information *8*

Map *9*

D. ABSTRACTS – ORAL PRESENTATIONS **10**

1. The Norwegian Sequencing Centre (NSC) 11
   Robert Lyle, Tim Hughes, NSC, Dag Undlien, Kjetill Jakobsen* *11*
2. NGS research and service at the CBU 12
   Kjell Petersen, Inge Jonassen* *12*
3. The Vital-IT HPC and the Swiss-Prot group 13
   Laurent Falquet *13*
4. BGI: combination of sequencing and bioinformatics strategy 14
   Ning LI, BGI-Europe* *14*
5. Innovation and Trends with In-Memory Technology 15
   Matthias Steinbrecher *15*
6. HTS Science and Technology Watch Tour 16
   Jean Imbert *16*
7. NGS data analysis: the user POV 17
   J. R. Valverde, J. M. Rodríguez, A. R. Rojas, A. Couce, J. Blázquez* *17*
8. Bioinformatics developments for NGS data analysis at PRABI 18
   Franck Picard, Guy Perrière* *18*
9. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases 19
   *Sonia Tarazona, Fernando García, Alberto Ferrer, Joaquín Dopazo, Ana Conesa** *19*
10. A combinatorial and integrated method to analyse RNA-seq reads 21
    *N. Philippe, M. Salson, T. Commes, E. Rivals** *21*
11. RSAT peak-motifs: fast extraction of transcription factor binding motifs from full-size ChIP-seq datasets 22
    Morgane Thomas-Chollier[1], Matthieu Defrance[2], Olivier Sand[3], Carl Herrmann[4], Denis Thieffry[4] and Jacques van Helden[,4,5]* *22*
12. The Seal suite of distributed software for high-throughput sequencing 23
    Luca Pireddu, Simone Leo, Gianluigi Zanetti* *23*
13. Scalable Cloud Computing Solutions for Next Generation Sequencing Data 24
    Matti Niemenmaa [1], Aleksi Kallio [2], Andre Schumacher [1], Petri Klemela [2], Eija Korpelainen [2], and Keijo Heljanko[,1]* *24*
14. smallRNA data analysis 25
    *Angelica Tulipano, Andreas Gisel** *25*

15. Chipster 2.0: User-friendly NGS data analysis software with built-in genome browser and workflow functionality 26
*Aleksi Kallio, Taavi Hupponen, Massimiliano Gentile, Jarno Tuimala, Rimvydas Naktinis, Kimmo Mattila, Ari-Matti Saren, Petri Klemelä, Eija Korpelainen** 26
16. Exploration of environmental metagenomes and metatranscriptomes: current possibilities and limitations in data analysis 27
Petr Baldrian 27

E. ABSTRACTS - DEMOS 28

17. Linking research data with scholarly publications 29
Attwood, T.K.[1,2], McDermott, P.[1,2], Marsh, J.[3], Pettifer, S.R.[2] and Thorne, D.[3]* 29

F. ABSTRACTS - POSTERS 30

18. In the Shadow of the Genome: A Challenging Journey to Diversity in Leishmania donovani 31
Imamura H[1], Mannaert A[1], Downing T[2], Berriman M[2], Dujardin JC[*,1]* 31
19. From cutadapt to sequencetools (sqt): a versatile toolset for sequencing projects 32
*Marcel Martin, Sven Rahmann** 32
20. Oncogenomics of the Hormone-responsive Breast Cancer Phenotype by NGS 33
Alessandro Weisz 33
21. An Integrated RNA-seq Atlas of the Murine T-Helper Cell Transcriptome 34
Andrew Deonarine 34
22. MEDIPS - computational analysis of genome-wide methylation using high-throughput sequencing 35
*Lukas Chavez, Jörn Dietrich, Mireia Vilardell-Nogales, Ralf Herwig** 35
23. UPPNEX - A solution for Next Generation Sequencing data management and analysis 36
Samuel Lampa[1,2], Jonas Hagberg[1], and Ola Spjuth[*,1,3]* 36
24. BioinformaticsTools@bioacademy.gr 37
Athanasia Pavlopoulou, Sophia Kossida* 37
25. Read indexing 38
*N. Philippe, M. Salson, T. Lecroq, M. Léonard, T. Commes, E. Rivals** 38
26. Digital gene expression data, cross-species conservation and noncoding RNA 39
*Nicolas Philippe, Florence Ruffle, Elias Bou-Samra, Anthony Boureux, Thérèse Commes, Eric Rivals** 39
27. Power and limits of capture - based, targeted DNA resequencing for mutation detection 40
*Fabrice Lopez, Hélène Holota, François-Xavier Théodule and Jean Imbert** 40
28. bcbio-nextgen: Automated, distributed next-gen sequencing pipeline 41
*Roman Valls Guimera, Science for life genomics staff, Brad Chapman** 41
29. Algorithm for error detection in metagonomics NGS data 42
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]* 42
30. Statistical approaches for the analysis of RNA-Seq and ChIP-seq data and their integration 44
*Claudia Angelini** and Italia De Feis* 44
31. Massive-scale RNA-Seq experiments in human genetic diseases 45
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]* 45
32. EU COST Action TD0801: Statistical Challenges On The 1000 Euro Genome Sequences In Plants 46
Marco C.A.M. Bink[*1], , Thomas Schiex[2]* 46
33. Epigenomic and transcriptional effects of Dnmt3b mutations in human ICF syndrome-derived B cell lines. 47
Sole Gatto[1,2], Claudia Angelini[2], Sylwia Leppert[1], Valentina Proserpio[3], Sarah Teichmann[3], Maurizio D'Esposito[1], Maria R. Matarazzo[*,1]* 47

34. Improved analysis of fungal communities using the next-generation-sequencing analysis of *rpb2* genes 48
Větrovský T., Voříšková J., Žifčáková L., Urbanová M., Baldrian P.* *48*
35. IT Future of Medicine: Next Generation Sequencing is the Key to Future Personalized Medicine 49
Hans Lehrach and Babette Regierer for the IT Future of Medicine Consortium *49*
36. TAPYR: An efficient high-throughput sequence aligner for re-sequencing applications 50
Francisco Fernandes, Paulo G.S. da Fonseca, Luis M.S. Russo, Arlindo L. Oliveira, Ana T. Freitas *50*

G. LIST OF PARTICIPANTS 51

B. Schedule

Meeting location: COST office, Avenue Louise 149, 1050 Bruxelles, Belgium.
Access information: http://www.cost.esf.org/service/contact

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 - TBD
		  ○ 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 – Talks**
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	*Standing 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**
19:30	20:00	*Moving to dinner place*
20:00	22:00	*Dinner at "Rouge tomate"*

Wednesday, November 9 - Work group meetings

09:00-09:20	**Introduction by Erik Bongcam**
09:20-11:30	**Split meeting (parallel sessions)** Albert Einstein room (21 floor – Boardroom /18 people) Competence room (ground floor / 30 people) Excellence room (ground floor / 25 people)
	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 - TBD 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*

C. Practical information

Expenses

All participants entitled to reimbursement have received an official invitation from COST. Reimbursement then will be made using the official COST system and therefore you should finish the registration Erik started by processing all steps on the COST system. At the end you will have to download the personalized reimbursement form.

Reimbursable:

- Hotel nights.
- Air fare.
- Public transportation without receipt up to 25 Euro otherwise receipts are required.
- Taxi up to 40 Euro for whole meeting with receipt.

The following other expenses are already pre-paid: lunches and conference dinner. Beware, there was an error in previous version of the booklet: **hotel expenses are NOT pre-paid**. We apologize for the inconvenience.

Arrival

- By train: take a taxi at Midi station or the Metro direction Simonis, step down at Porte Louise to the hotel Bristol Stéphanie (see map below).
- By plane to Zaventem airport: take the train from the airport to Midi Station then follow the instructions for train.

Hotel (arrow A on map below)

Thon Hotel Bristol Stéphanie

(http://www.thonhotels.be/hotels/countrys/belgique/bruxelles/thon-hotel-bristol-stephanie/)

Avenue Louise 91, 1050 Bruxelles
Tel: +32 2 700 78 78
(5 minutes walk from COST Office, between rue de Florence and rue Defacz)

Meeting place (arrow B on map below)

We draw your attention to the fact that the SEQAHEAD meeting **will take place**

in the COST Office premises located

Avenue Louise 149 - 1050 Ixelles (Bruxelles)
Phone : +32 2 533 38 00
http://www.cost.esf.org/service/contact/

Conference dinner (arrow C on map below)

November 8, 20:00, Restaurant "Rouge Tomate"
190, avenue Louise 1050 Bruxelles (http://www.rougetomate.be/)

Contact

- Myriam Loubriat (mloubria@ulb.ac.be). Mobile phone: +32 478 59 02 38
- Erik Bongcam-Rudloff (erikbong@mac.com)
- Jacques van Helden (jvhelden@ulb.ac.be)

Detailed information

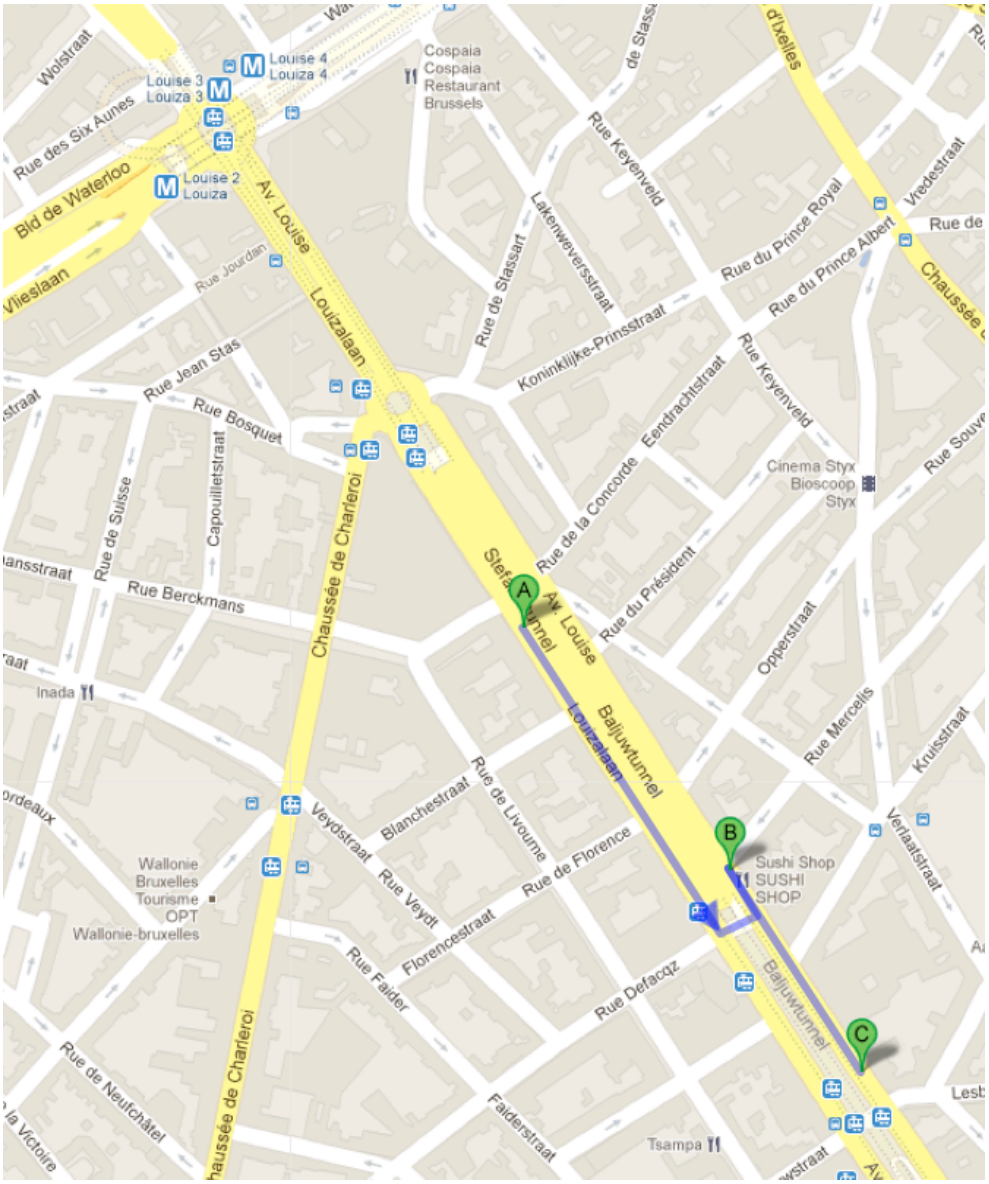
http://seqahead.cs.tu-dortmund.de/meetings:2011-11

Map

A = Hôtel Bristol Stéphanie

B = Meeting place COST Office

C = Restaurant “Rouge Tomate”



D. Abstracts – Oral presentations

1. The Norwegian Sequencing Centre (NSC)

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

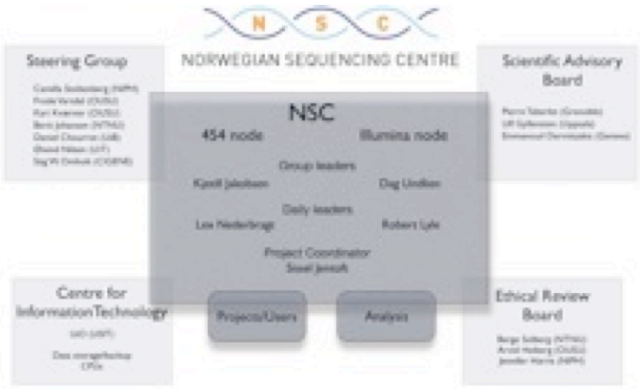
Department of Medical Genetics and Norwegian Sequencing Centre Oslo University Hospital, Kirkeveien 166, 0407 Oslo, Norway

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

Email: Robert.Lyle@medisin.uio.no

Abstract

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

2. NGS research and service at the CBU

Kjell Petersen*, Inge Jonassen

Computational Biology Unit, Uni Computing, Uni Research AS, Thormøhlengst 55, N-5008 Bergen, Norway.
Web: http://www.bioinfo.no/ http://www.uni.no/computing/units/cbu
Email: Kjell.Petersen@uni.no

Abstract

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 coordinates 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 (Star et al) 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 (Balzer et al).

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 (Quince et al).

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 (Stavrum et al).

In addition to algorithms and tools, a suitable infrastructure for step-by-step analysis of your workflow, 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

1. http://www.bioinfo.no/
2. http://www.microarry.no/
3. http://jexpress.bioinfo.no/
4. http://www.esysbio.org/
5. http://storebioinfo.norstore.no/

3. The Vital-IT HPC and the Swiss-Prot group

Laurent Falquet

Vital-IT, Swiss Institute of Bioinformatics, Genopode-UNIL, CH-1015 Lausanne, Switzerland
Web: http://www.vital-it.ch/
Email: laurent.falquet@isb-sib.ch

Abstract

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

1. http://www.vital-it.ch/
2. http://www.isb-sib.ch/

4. BGI: combination of sequencing and bioinformatics strategy

Ning LI*, BGI-Europe

5. Innovation and Trends with In-Memory Technology

Matthias Steinbrecher

Innovation Center Potsdam, TIP HPI Strategic Projects SAP AG
August-Bebel-Str. 88, 14482 Potsdam, Germany
Web: http://www.sap.com
Email: matthias.steinbrecher@sap.com

Abstract

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/

6. HTS Science and Technology Watch Tour

Jean Imbert

TAGC UMR_S 928 - Inserm - Université de la Méditerranée - case 928 -163 Avenue de Luminy
13288 Marseille Cedex 09 - France
Web: http://www.yourwebsite.org/
Email: jean.imbert@inserm.fr

Abstract

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

7. NGS data analysis: the user POV

J. R. Valverde*, J. M. Rodríguez, A. R. Rojas, A. Couce, J. Blázquez

CNB/CSIC, C/Darwin 3. 28049. Spain.
Web: http://www.es.embnet.org/

Abstract

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

- http://www.es.embnet.org/
- http://www.cnb.csic.es/
- http://www.cnb.uam.es/content/research/microbial/stress/

8. 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, 16 rue Raphaël Dubois, 69622 Villeurbanne Cedex, France

Web: http://www.prabi.fr/

Email: franck.picard@univ-lyon1.fr

Abstract

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

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

9. 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.
Web: http://bioinfo.cipf.es/aconesa Email: aconesa@cipf.es

Abstract

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 edgeR[2], DESeq[3], baySeq[4] 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

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

10. A combinatorial and integrated method to analyse RNA-seq reads

N. Philippe, M. Salson, T. Commes, E. Rivals*

Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS,
Université de Montpellier II, 161 rue Ada, 34392 Montpellier, France,
LIFL, CNRS, INRIA Lille, Univ. Lille I, Villeneuve d'Ascq, France
CRBM, UMR 5237 CNRS, Montpellier, France
Web: http://www.lirmm.fr/~rivals
Email: rivals@lirmm.fr

Abstract

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. 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

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

11. 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, Ihnestrasse 73, 14195 Berlin, Germany. Email: morgane@bigre.ulb.ac.be
2. Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México. Av. Universidad, Cuernavaca, Morelos 62210, Mexico. Email: defrance@ccg.unam.mx
3. CNRS-UMR8199 Institut de Biologie de Lille. Génomique et maladies métaboliques. 1, rue du Pr Calmette, 59000 Lille, France. Email: sand@good.ibl.fr
4. Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée. Campus de Luminy, F - 13288 Marseille, France. Email: {thieffry,herrmann}@tagc.univ-mrs.fr
5. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be

Abstract

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 (http://rsat.ulb.ac.be/rsat/), 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

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

12. The Seal suite of distributed software for high-throughput sequencing

Luca Pireddu*, Simone Leo, Gianluigi Zanetti

CRS4, Polaris, Ed. 1, I-09010 Pula, Italy
Email: luca.pireddu@crs4.it

Abstract

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

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

13. Scalable Cloud Computing Solutions for Next Generation Sequencing Data

Matti Niemenmaa [1], Aleksi Kallio [2], Andre Schumacher [1], Petri Klemela [2], Eija Korpelainen [2], and Keijo Heljanko [*,1]

1 = Aalto University, School of Science, Department of Information and Computer Science

2 = CSC — IT Center for Science

Web: http://users.ics.tkk.fi/kepa/

Email: keijo.heljanko@aalto.fi

Abstract

This talk will discuss the use of scalable cloud computing technologies, in particular the use of the massively parallel MapReduce framework for processing next generation sequencing data. We have created an open source library called Hadoop-BAM that uses the open source Apache Hadoop system to parallelize the processing of binary alignment/map (BAM) files.

Hadoop-BAM is a novel library for the scalable manipulation of aligned next generation sequencing data in the Hadoop distributed computing framework. It acts as an integration layer between analysis applications and BAM files that are processed using Hadoop. Hadoop-BAM solves the issues related to BAM data access by presenting a convenient API for implementing map and reduce functions in the Hadoop map-reduce framework. It builds on top of the Picard SAM JDK, so tools that rely on the Picard API are expected to be easily convertible to support large-scale distribution.

We will demonstrate the use of Hadoop-BAM by building a coverage summarizing tool for the Chipster genome browser. Our results show that Hadoop offers good scalability, and the best performance is achieved by minimizing data import and export in analysis pipelines based on the Hadoop framework. We will discuss the challenges faced and outline potential future research directions of using distributed computing in NGS data processing.

This is joint work with Aalto University Department of Information and Computer Science and CSC Chipster team.

References

1. Matti Niemenmaa, André Schumacher, Keijo Heljanko, Aleksi Kallio, Petri Klemelä, Taavi Hupponen, and Eija Korpelainen: *Hadoop-BAM: A Library for Genomic Data Processing*. Poster presented at the Bioinformatics Open Source Conference (BOSC 2011).
2. Matti Niemenmaa, Aleksi Kallio, André Schumacher, Petri Klemelä, Eija Korpelainen, and Keijo Heljanko: *Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud*. Submitted manuscript.

Relevant Web sites

1. http://sourceforge.net/projects/hadoop-bam/
2. http://chipster.csc.fi/
3. http://www.slideshare.net/bosc2011/kallio-bosc2011-hadoopbam/

14. smallRNA data analysis

Angelica Tulipano, Andreas Gisel*

Institute for Biomedical Technologies - CNR
Via Amendola 122/D
70126 Bari, Italy
Web: http://www.ba.itb.cnr.it/
Email: andreas.gisel@ba.itb.cnr.it

Abstract

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

15. Chipster 2.0: User-friendly NGS data analysis software with built-in genome browser and workflow functionality

Aleksi Kallio, Taavi Hupponen, Massimiliano Gentile, Jarno Tuimala, Rimvydas Naktinis, Kimmo Mattila, Ari-Matti Saren, Petri Klemelä, Eija Korpelainen*

CSC – IT Center for Science, Espoo, Finland

Web: http://chipster.csc.fi

Email: eija.korpelainen@csc.fi

Abstract

The open source Chipster software provides an intuitive graphical user interface to NGS and microarray data analysis tools, interactive visualizations and workflow functionality. This presentation covers version 2.0, with the emphasis on the analysis and visualization tools for ChIP-seq, RNA-seq, miRNA-seq and methyl-seq data.

Users can perform their whole data analysis in Chipster from quality control, preprocessing, and alignment to normalization, statistical analysis and downstream applications such as pathway enrichment and motif discovery. Popular packages such as FastQC, FASTX, SAMtools, BEDTools, Bowtie, BWA and TopHat are included, and care has been taken to serve them in a bioscientist-friendly manner. Analysis pipelines can easily be saved as automatic, reusable workflows, which can be shared with other users.

The ChIP-seq analysis tools enable users to detect peaks with MACS, filter them based on p-value, number of reads etc., and scan them for common sequence motifs to be matched against the JASPAR database. Users can also retrieve the nearby genes, filter them based on several criteria, and perform pathway analysis.

Data from miRNA-seq experiments can be normalized and analyzed for differential expression through integration of the edgeR Bioconductor package. Target genes for miRNA:s can be retrieved from several databases and analyzed for pathway enrichment. For RNA-seq data users can also opt for the integrated Cufflinks package.

Chipster's built-in genome browser allows seamless visualization of reads and results in their genomic context using Ensembl annotations. Users can zoom in to nucleotide level, highlight SNP:s and view the automatically calculated coverage. Cross-talk between the genome browser and BED files allows users to quickly inspect genomic regions by simply clicking on the data row of interest.

Technically Chipster is a Java-based client-server system. It is open source, and new tools can easily be added using a simple mark-up language. We are currently working with the Hadoop MapReduce framework so that large jobs can be run in the cloud. Also, a virtual machine distribution of Chipster is being set up to make the server installation even easier. Taken together, Chipster provides an easy way to serve NGS data analysis tools in a biologist-friendly manner.

References

1. Kallio, Tuimala, Hupponen et al.: *Chipster: User-friendly analysis software for microarray and other high-throughput data* (2011) BMC Genomics, submitted.
2. Niemenmaa, Kallio, Schumacher et al: *Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud* (2011) Bioinformatics, submitted.

Relevant Web sites

1. http://chipster.csc.fi
2. http://chipster.sourceforge.net/

16. 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.
Web: http://www.biomed.cas.cz/mbu/lbwrf
Email: baldrian@biomed.cas.cz

Abstract

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., Kolařík, M., Štursová, M., Kopecký, J., Valášková, V., Větrovský, T., Žifčáková, L., Šnajdr, J., Rídl, J., Vlček, Č., Voříšková, 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.

E. Abstracts - Demos

17. Linking research data with scholarly publications

Attwood, T.K.*,[1,2], McDermott, P. [1,2], Marsh, J.[3], Pettifer, S.R.[2] and Thorne, D.[3]

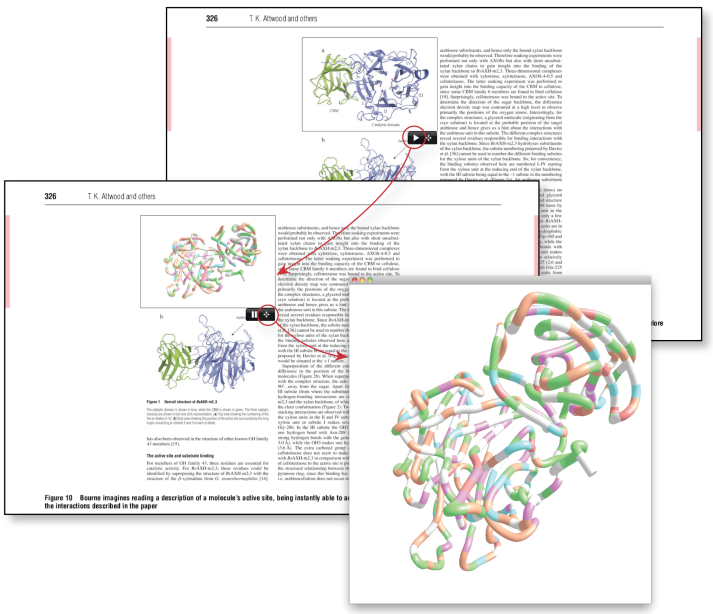
[1]School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK.
[2]Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK.
[3]School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, UK.

Web: http://utopia.cs.man.ac.uk/
Email: Teresa.k.attwood@manchester.ac.uk

Abstract

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.



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

- http://getutopia.com/
- http://utopia.cs.man.ac.uk/utopia/

F. Abstracts - Posters

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

Imamura H[1], Mannaert A[1], Downing T[2], Berriman M[2], Dujardin JC[*,1]

[1]Institute of Tropical Medicine, Antwerp, Belgium; [2]Wellcome Trust Sanger Institute, Hinxton, UK

Web: http://www.itg.be/

Email: jcdujardin@itg.be

Abstract

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. 2011. 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

- www.leishrisk.net/kaladrug
- www.leishrisk.net/gemini

19. 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

Web: http://www.rahmannlab.de/
Email: Sven.Rahmann@tu-dortmund.de

Abstract

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

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

20. 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
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.

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

Andrew Deonarine

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

Abstract

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

- http://www.thelpercell.com

22. MEDIPS - computational analysis of genome-wide methylation using high-throughput sequencing

Lukas Chavez, Jörn Dietrich, Mireia Vilardell-Nogales, Ralf Herwig*

Max Planck Institute for molecular Genetics, Department Vertebrate Genomics, Ihnestr. 73, 14195 Berlin, Germany
Web: http://www.molgen.mpg.de
Email: herwig@molgen.mpg.de

Abstract

The generation of genome-wide data derived from methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq) has become a major tool for epigenetic studies in health and disease. The computational analysis of such data, however, still falls short on accuracy, sensitivity, and speed. We have developed a software package (MEDIPS) that is able to cope with the inherent complexity of MeDIP-seq data. Core of MEDIPS is a normalization procedure that is based on a linear regression model for quantification of the influence of local CpG density. The approach is faster than existing approaches with similar performance compared to public benchmark data. We show improved correlation of normalized MeDIP-seq data in comparison to available whole-genome bisulfite sequencing data, and investigated the effect of differential methylation on gene expression with several applications.

References

1. Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, Herwig R*, Adjaye J*. Genome Res. 2010 Oct;20(10):1441-50. Epub 2010 Aug 27. *equal contribution

Relevant Web sites

1. http://medips.molgen.mpg.de/
2. http://www.bioconductor.org/packages/2.8/bioc/html/MEDIPS.html

23. 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)

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

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

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

Email: ola.spjuth@farmbio.uu.se

Abstract

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 workbench[1]. 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 portal[3]. 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 Science[4] (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/

24. BioinformaticsTools@bioacademy.gr

Athanasia Pavlopoulou*, Sophia Kossida

Biomedical Research Foundation, Academy of Athens
Soranou tou Efessiou 4, 115 27 Athens, Greece
Web: http:// www.bioacademy.gr/bioinformatics/
Email: apavlopoulou@bioacademy.gr

Abstract

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: SAFE[1] for the analysis of gene fusion events; GIBA[2] for detecting protein complexes; Brukin2D[3] for 2D visualization and comparison of LC-MS data; GOmir[4] 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

1 http://www.bioacademy.gr/bioinformatics/projects/ProteinFusion/index.htm
2. http://www.bioacademy.gr/bioinformatics/projects/GIBA/
3. http://www.bioacademy.gr/bioinformatics/Brukin2d/index.html
4. http://www.bioacademy.gr/bioinformatics/projects/GOmir/bioinformatics_home.htm

25. Read indexing

N. Philippe, M. Salson, T. Lecroq, M. Léonard, T. Commes, E. Rivals*

Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS, équipe MAB, Université de Montpellier II, 161 rue Ada, 34392 Montpellier, France,
LITIS, Univ. Rouen, Mont Saint Aignen, France
CRBM, UMR 5237 CNRS, Montpellier, France
Web: http://www.lirmm.fr/~rivals
Email: rivals@lirmm.fr

Abstract

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

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

26. 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, 161 rue Ada, 34392 Montpellier, France,
CRBM, UMR 5237 CNRS, Montpellier, France
Web: http://www.lirmm.fr/~rivals
Email: rivals@lirmm.fr

Abstract

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

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

27. 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 - case 928 -163 Avenue de Luminy
13288 Marseille Cedex 09 - France
Web: http://www.yourwebsite.org/
Email: jean.imbert@inserm.fr

Abstract

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 **Ge**nomic **Var**iant **A**nalyzer.

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

2. **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.
3. 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.
4. 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:emboj2011295 [pii];10.1038/emboj.2011.295 [doi].

Relevant Web sites

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

28. 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
Web: http://www.hsph.harvard.edu/research/bioinfocore/, http://www.scilifelab.se/
Email: bchapman@hsph.harvard.edu , roman.valls.guimera@scilifelab.se

Abstract

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

29. 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*

Web: http://www.abi.bg/ and http://www.ibb.waw.pl/

* Corresponding author: Dimitar Vassilev – jim6329@gmail.com

Abstract

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 a s 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.

3. Kelly D., Schatz M., Salzberg S. (2010) Quake: a quality-aware detection and correction of sequencing errors. *Genome Biology* 11:R116.
4. Salmela L., Schroder J. (2011) Correcting errors in short read my multiple alignments. *Bioinformatics* 27(11): 1455-1461.
5. Schroder J., Schroder H., Puglisi S., Sinha R., Schmidt B. (2009) SHREC: a short read error correction method. *Bioinformatics* 25(17):2157-2163.
6. Simon C., Daniel R. (2011) Metagenomics analyses: past and future trends. *Applied and Env. Microbiology* 77(4):1153-1560.

Relevant Web sites

1. http://redmine.3mhz.net/

30. 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
Web: http://www.iac.cnr.it/
Email: c.angelini@iac.cnr.it

Abstract

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.

31. 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]

[1]CNR, Institute of Genetics and Biophysics "A. Buzzati-Traverso" (IGB), Via P. Castellino 111 Naples, Italy;
[2] CNR, Istituto per le Applicazioni del Calcolo (IAC), Via P. Castellino 111 Naples, Italy
http://www.igb.cnr.it
Email: alfredo.ciccodicola@igb.cnr.it

Abstract

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.

32. 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 Droevendaalsesteeg 1 6708 PB Wageningen The Netherlands

[2] MIA - INRA Chemin de Borde Rouge, BP 52627, Castanet-Tolosan 31326 Cedex France

Web: http://www.statseq.eu/
Email: marco.bink@wur.nl

Abstract

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)

33. 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

.

Web: http://www.igb.cnr.it/
Email: maria.matarazzo@igb.cnr.it

Abstract

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.

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

Větrovský T.*, Voříšková J., Žifčáková L., Urbanová M., Baldrian P.

Laboratory of Environmental Microbiology, Institute of Microbiology of the ASCR, Prague, Czech Republic.
Web: http://www.biomed.cas.cz/mbu/lbwrf
Email: vetrovsky@biomed.cas.cz

Abstract

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.

35. 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
Ihnestr. 63-73, 14195 Berlin, Germany
Web: http://www.itfom.eu/
Email: regerer@molgen.mpg.de

Abstract

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

1. http://www.itfom.eu/
2. http://www.fet-f.eu/
3. http://www.molgen.mpg.de/research/lehrach/

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

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

1 Instituto de Engenharia de Sistemas e Computadores: Investiga_c~ao e Desenvolvimento (INESC-ID), R. Alves Redol 9, 1000-029 Lisboa, Portugal

2 Instituto Superior T_ecnico{Universidade T_ecnica de Lisboa (IST/UTL), Av. Rovisco Pais, 1049-001 Lisboa, Portugal

Email: fjdf@kdbio.inesc-id.pt; pgsf@kdbio.inesc-id.pt ; lsr@kdbio.inesc-id.pt ; aml@kdbio.inesc-id.pt; atf@kdbio.inesc-id.pt;

Abstract

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.

G. List of participants

Name	Address	Country	Email
Lucia ALTUCCI	SUN. Via L. De Crecchio 7, 80138, NA	Italy	Lucia.altucci@unina2.it
Claudia ANGELINI	Istituto per le Applicazioni del Calcolo, CNR. Via Pietro Castellino 111, 80131, Napoli	Italy	claudia.angelini@cnr.it
Teresa ATTWOOD	University of Manchester. Faculty of Life Sciences, Michael Smith Building, Oxford Road, Manchester, M13 9PT	United Kingdom	teresa.k.attwood@manchester.ac.uk
Petr BALDRIAN	Institute of Microbiology of the ASCR. Videnska 1083, 14220, Praha 4	Czech Republic	baldrian@biomed.cas.cz
Endre BARTA	University of Debrecen, Institute of Biochemistry and Molecular Biology. 4032 Debrecen, Nagyerdei krt. 98. POB. 6.	Hungary	barta.endre@unideb.hu
Marco BINK	Wageningen University and Research centre – Biometris. Droevendaalsesteeg 1, 6708 PB Wageningen	Netherlands	Marco.bink@wur.nl
Erik BONGCAM-RUDLOFF	SLU. Undervisningsplan 4 A, Box 7023, 750 07 Uppsala	Sweden	Erik.bongcam@slu.se
Gianluca BONTEMPI	ULB. Bld de Triomphe CP 212	Belgium	gbonte@ulb.ac.be
João CARRIÇO	Faculdade de Medicina, Universidade de Lisboa. Instituto de Microbiologia, FM,UL, Av. Professor Egas Moniz, 1649-028 Lisboa	Portugal	jcarrico@fm.ul.pt
Alfredo CICCODICOLA	CNR, Institute of Genetics and Biophysics "A. Buzzati-Traverso". Via P. Castellino, 111	Italy	alfredo.ciccodicola@igb.cnr.it
Ana CONESA	Centro de Investigacion Principe Felipe. Avda. Autopista Saler s/n, 46012 Valencia	Spain	aconesa@cipf.es
Valerio COSTA	CNR, Institute of Genetics and Biophysics "A. Buzzati-Traverso". Via P. Castellino, 111	Italy	costav@igb.cnr.it
Italia DE FEIS	Istituto per le Applicazioni del Calcolo "M. Picone" - CNR. Via Pietro Castellino 111, 80131 Napoli	Italy	i.defeis@iac.cnr.it
Andrew DEONARINE	Laboratory of Molecular Biology, University of Cambridge. Hills Road, Cambridge	United Kingdom	Andrew.deonarine@mrc-lmb.cam.ac.uk
Jean-Claude DUJARDIN	Instituut voor Tropische Geneeskunde. Nationalestraat, 155 B-2000 Antwerpen	Belgium	jcdujardin@itg.be
Laurent FALQUET	SIB. Vital-IT, Genopode-UNIL, CH-1015 Lausanne	Switzerland	Laurent.falquet@unil.ch
Ana Teresa FREITAS	INESC-ID, Lisbon. Rua Alves Redol 9 1000-029 Lisbon	Portugal	atf@inesc-id.pt

Name	Address	Country	Email
Sole GATTO	Institute of genetics and Biophysics "ABT" - CNR. Via P Castellino 111, 80131 Naples	Italy	gatto@igb.cnr.it
Andreas GISEL	ITB - CNR. Via Amendola 122/D, 70126 Bari	Italy	andreas.gisel@ba.itb.cnr.it
Simon HEATH	Centro Nacional de Análisis Genómico (CNAG). 4 Baldiri i Reixad, PCB-Torre I, 2º Floor, 08028 Barcelona	Spain	simon.heath@gmail.com
Keijo HELJANKO	Aalto University. PO Box 15400, FI-00076 Aalto, Finland	Finland	keijo.heljanko@aalto.fi
Sylvie HERMOUET	INSERM U892. 8 quai Moncousu	France	sylvie.hermouet@univ-nantes.fr
Ralf HERWIG	Max Planck Institute for molecular Genetics . Ihnestr. 73, 14195 Berlin	Germany	herwig@molgen.mpg.de
Hideo IMAMURA	Institute of tropical medicine. Nationalstraat 155, Antwerp, 2000	Belgium	hi1@sanger.ac.uk
Jean IMBERT	Inserm. TAGC UMR928 - Inserm - Université de la Méditerranée - case 928 -163 Avenue de Luminy - 13288 Marseille Cedex 09 - France	France	jean.imbert@inserm.fr
Aleksi KALLIO	CSC – IT Center for Science Ltd.. P.O. Box 405, FI-02101 Espoo, Finland	Finland	aleksi.kallio@csc.fi
Lubos KLUCAR	Institute of Molecular Biology SAS. Dubravska cesta 21, 845 51 Bratislava	Slovakia	klucar@embnet.sk
Eija KORPELAINEN	CSC. PO Box 405, Keilaranta 14, 02101 Espoo	Finland	Eija.korpelainen@csc.fi
Robert KRALOVICS	Center for Molecular Medicine of the Austrian Academy of Sciences. Lazarettgasse 14, BT25.3	Austria	robert.kralovics@cemm.oeaw.ac.at
Ning LI	BGI Europe. Ole Maaløes Vej 3 | DK-2200 Copenhagen N | Denmark	Denmark	lining@genomics.cn
Robert LYLE	Oslo university Hopsital. Kirkeveien 166, 0407 Oslo	Norway	Robert.lyle@medisin.uio.no
Maja MALKOWSKA	Laboratory of Bioinformatics and Biostatistics. Maria Skłodowska-Curie Memorial Cancer Center . Warsaw. 5 Roentgena Street, 02-781 Warsaw, Poland.	Poland	m.malkowska@gmail.com
An MANNAERT	Institute of Tropical Medicine. Nationalestraat 155, 2000 Antwerpen	Belgium	amannaert@itg.be
Maria R MATARAZZO	Institute of genetics and Biophysics "ABT" - CNR. Via P Castellino 111, 80131 Naples	Italy	maria.matarazzo@igb.cnr.it
Ivan MINKOV	University of Plovdiv. 24 Tsar Assen Str, 4000 Plovdiv	Bulgaria	minkov@uni-plovdiv.bg
Athanasia PAVLOPOULOU	Biomedical Research Foundation, Academy of Athens. Soranou tou Efessiou 4, 115 27 Athens	Greece	apavlopoulou@bioacademy.gr
Kjell PETERSEN	Uni Research AS, Uni Computing, Computational Biology Unit. Thormøhlensgt 55, N-5008 Bergen	Norway	Kjell.Petersen@uni.no

Name	Address	Country	Email
Franck PICARD	CNRS. Laboratoire Biometrie et Biologie Evolutive, UCB Lyon 143 Bd du 11 novembre 69622 Villeurbanne cedex	France	Franck.picard@univ-lyon1.fr
Luca PIREDDU	CRS4. Polaris, Ed. 1, I-09010 Pula	Italy	luca.pireddu@crs4.it
Noemi POLGAR	University of Pécs Medical School. 12 Szigeti Road, H-7624 Pecs, Hungary	Hungary	noemi.polgar@aok.pte.hu
Sven RAHMANN	University of Duisburg-Essen. University of Duisburg-Essen. Genome Informatics. Institute of Human Genetics. Faculty of Medicine / University Hospital. Hufelandstr. 55. 45122 Essen, GERMANY	Germany	Sven.Rahmann@tu-dortmund.de
Babette REGIERER	Max Planck Institute for Molecular Genetics. Ihnestr. 63-73, 14195 Berlin	Germany	regierer@molgen.mpg.de
Eric RIVALS	LIRMM – CNRS & Univ. Montpellier 2. CC 477; 161, rue Ada 34095 Montpellier Cedex 5	France	rivals@lirmm.fr
Ola SPJUTH	Uppsala University. Box 591	Sweden	Ola.spjuth@farmbio.uu.se
Matthias STEINBRECHER	Innovation Center Potsdam TIP HPI Strategic Projects SAP AG . August-Bebel-Str. 88, 14482 Potsdam, Germany	Germany	matthias.steinbrecher@sap.com
Thomas SVENSSON	Science for Life laboratory/karolingska Institutet. Tomtebodavägen 23A	Sweden	Thomas.svensson@scilifelab.se
Gerhard THALLINGER	Institute for Genomics and Bioinformatics. Graz University of Technology, Petergasse 14, 8010 Graz	Austria	Gerhard.Thallinger@tugraz.at
Jose R VALVERDE	CSIC. CNB/CSIC. C/Darwin, 3. 28049 Madrid	Spain	jrvalverde@cnb.csic.es
Jacques VAN HELDEN	Université Libre de Bruxelles Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Boulevard du Triomphe, 2 Campus Plaine, 1050 Brussels	Belgium	Jacques.van.Helden@ulb.ac.be
Dimitar VASSILEV	Bioinformatics group, AgroBioInstitute. 8 Dragan Tsankov, Str.	Bulgaria	jim6329@gmail.com
Tomáš VĚTROVSKÝ	Institute of Microbiology ASCR. Vídeňská 1083, 142 20, Praha 4-Krč	Czech Republic	kostelecke.uzeniny@seznam.cz
Gert VRIENDT	CMBI NCMLS UMC Nijmegen. Geert Grooteplein 28 (route 260) 6525 GA Nijmegen	Netherlands	vriend@cmbi.ru.nl
Alessandro WEISZ	Molecular Medicine and Genomics Laboratory, University of Salerno (formerly: Second University of Napoli). Lab. Medicina Molecolare e genomica, Campus di Medicine, Università degli Studi di Salerno, via S. Allende 1, 84081 Baronissi (SA)	Italy	aweisz@unisa.it