EMBnet.journal
Volume 19
Supplement A
May 2013

The Next NGS Challenge Conference: Data Processing and Integration

14-16 May 2013, Valencia, Spain
www.thenextngschallenge.org

cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

ESF
EUROPEAN SCIENCE FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

ESF provides the Cost Office through an EC contract

COST is supported by the EU RTD Framework Programme

Editorial

Two million years of stone technology represent the first long era of discovery at the start of human history;the use of fire, more than 500,000 years ago, was also a significant discovery.

Fire has been called the foundation of human civilisation, and there are few things more vital than starting and maintaining one when you're in a survival situation.

The first Next Generation Sequencing technology in use in the realm of Life Sciences was "Pyrosequencing", a technology based on the use of the fire fly enzyme.

Therefore, the Editorial board of *EMBnet .journal* has chosen for this supplement an image of a man and a torch in a balancing act, symbolising what this joint COST, EMBnet and ISCB Conference represents: rapid technological advances at the edge of technology, ushering in a new era of great discoveries.

EMBnet.journal Editorial Board

Man and a torch in balancing act Cape Town, South Africa, 2012
[© Erik Bongcam-Rudloff]

Contents

Editorial 2
The Next NGS Challenge conference: Data processing and integration.....3
Scientific Programme.....5
Keynote Lectures.....7
Oral Communications 13
Posters.....34

The Next NGS Challenge Conference: Data Processing and Integration





Erik Bongcam-Rudloff[1] ✉, Teresa K Attwood[2], Ana Conesa[3], Andreas Gisel[4], Burkhard Rost[5]

[1]Swedish University of Agricultural Sciences, Uppsala, Sweden
[2]University of Manchester, Manchester, United Kingdom
[3]Bioinformatics and Genomics Department of the Centro de Investigación Príncipe Felipe, Valencia, Spain
[4]CNR, Institute for Biomedical Technologies, Bari, Italy
[5]Department for Bioinformatics and Computational Biology, Fakultät für Informatik, Garching, Germany

Next Generation Sequencing (NGS) is a highly parallelised approach for quickly and economically sequencing new genomes, re-sequencing large numbers of known genomes, or for rapidly investigating transcriptomes under different conditions. Producing data on an unprecedented scale, these techniques are now driving the generation of knowledge (especially in biomedicine and molecular life sciences) to new dimensions. The massive data volumes being generated by these new technologies require new data-handling and -storage methods. Hence, the Life Science community urgently needs new and improved approaches to facilitate NGS data management and analysis.

A "moving target", this field requires that bioinformaticians, computer scientists and biomedical scientists join their expertise to bring NGS data management and analysis to new levels of efficiency and integration.

In your hands, you hold the proceedings of the *"Next NGS Challenge, Data Processing and Integration, Conference"* organised during 16-17 May 2013 in Valencia, Spain. This conference – a joint event of the EU COST Action BM1006, SeqAhead; the Global Bioinformatics Network, EMBnet; the FP7 Project, STATegra; and the International Society for Computational Biology, ISCB – aims to become a regular, dedicated meeting on cutting-edge NGS applications. As such, it will continue to bring together biologists, computational biologists and bioinformaticians to discuss new challenges in high-throughput sequencing, and to highlight new trends in NGS-based genome research.

The Next NGS Challenge
data processing and integration

Sponsors and Supporters

Single-Molecule, Real-Time (SMRT™) DNA Sequencing: Technology Overview and Recent Applications

Understanding the dynamics of biological processes is fundamental to understanding life itself. At Pacific Biosciences, we are developing applications to observe individual biomolecules at work in real time in nanostructures. The first is monitoring DNA synthesis by single DNA polymerase molecules, allowing the speed, processivity, and efficiency of the enzyme to be exploited for new capabilities in DNA sequencing. The power of this new sequencing technology - characterized by very long readlengths (up to 30.000 bp) and fast run times - is highlighted through examples from diverse applications, such as finishing and improving genomes (*de novo* assembly), targeted resequencing and haplophasing of difficult to sequence regions like GC- or AT-rich, resolving complex genomic regions (repeats, indels, HLA), characterizing transcript and gene fusion diversity, rapid pathogen sequencing, and the direct detection of epigenetic base modifications by making use of the kinetic information collected during DNA-sequencing.

GI Tech, a subsidiary of BGI - the world's largest genomics organization, is a leading provider of next-generation sequencing and bioinformatics analysisservices for global customers. Equipped with the industry's broadest array ofcutting-edge technologies, BGI Tech delivers rapid, cost-effective, and high-qualityresults that enable researchers to achieve scientific breakthroughs.

Scientific Programme

The Next NGS Challenge Conference: Data Processing and Integration

14-16 May 2013, Valencia, Spain

www.thenextngschallenge.org

Conference Programme

8:30AM	CONFERENCE REGISTRATION	
9:15AM	CONFERENCE WELCOME	
9:30AM	Keynote #1. Janet Kelso: Sequencing ancient genomes	
10:30AM	* Ralph Vogelsang: Automated Finished Microbial Genomes and Epigenomes to Understand Infectious Diseases.	
11:00AM	Break	
11:30AM	* Francisco M. De La Vega: Toward highly accurate and fast variant and de novo mutation identification from high-throughput sequencing data by joint Bayesian family calling.	
11:50AM	* Vladimir B. Teif: Developing a Software Suite to analyze the interplay between Nucleosome arrangement, DNA Methylation and Transcription factor binding.	
12:10AM	* John William Whitaker: Interplay between DNA Sequence motifs and the Human Epigenome.	
12:30	Lunch	
2:30PM	* Morgane Thomas-Chollier: RSAT Peak- Motifs: Efficient prediction of transcription factor motifs and binding sites from Genome- Wide sequencing peaks sets.	
2:50PM	* Babette Regierer: ESFRI - Infrastructure for Systems Biology Europe - (ISBE)	
3:10PM	* Javier Santoyo: Challenges in whole exome sequencing to identify disease-causing variants in human rare diseases	
3:30PM	Break	
4:00PM	Keynote #2. Karla Neugebauer: The Role of Gene Architecture in Gene Expression	
5:00PM 7:00PM	Posters&Beers	MC Meeting SeqAhead

Thursday 16th, 2013

9:00AM	Keynote #3. Leif Andersson: How domestic animal genomics can teach human medicine and medicine and evolutionary biology	
10:00AM	Keynote #4. Michele Morgante: Structural variation and the plant pan genomes	
11:00AM	Break	
11:30AM	* Carine Poussin: Translational Systems Biology Understanding the limits of Animals Models as Predictors of Human Biology.	
11:50AM	* Kay Nieselt: Automated transcription start site prediction for comparative comparative transcriptomics using the Supergenome.	
12:10AM	* Oskar Erik Karlsson: Viral Metagenomics – New applications for the broad-range detection of viromes in veterinary and public health settings.	
12:30AM	Lunch	
2:30PM	* Ignacio Blanquer Blanquer: Supporting NGS pilelines in the cluod.	
2:50PM	* Luca Pireddu: Automated and traceable processing for large-scale high-throughput sequencing facilities.	
3:10PM	* Thomas Svensson: SciLifeLab:New national infrastructures for NGS data production and applied "down stream " bioinformatics analysis in order to meet the demands of the scientific community	
3:30PM	Break	
4:00PM	Keynote #4. Ivo Gut: High-Throughput DNA Analysis	
5:00PM 7:00PM	Posters&Beers	

Keynote Lectures

How domestic animal genomics can teach human medicine and evolutionary biology

Leif Andersson

Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Sweden

Domestic animals provide unique opportunities for exploring genotype-phenotype relationships. Firstly, selective breeding during thousands of years has enriched for mutations that have adapted domestic animals to a new environment, i.e. farming under various environmental conditions. Secondly, the population structure is often favorable for genetic studies, large families and more or less closely related subpopulations (breeds). Thirdly, strong positive selection leaves genomic footprints that facilitate positional cloning. The combined use of whole genome resequencing, linkage mapping and linkage disequilibrium (LD) mapping within and between breeds provides a powerful approach for positional identification of both monogenic and multifactorial trait loci. The successful use of this approach for identifying genes underlying phenotypic traits will be illustrated on the basis of our research program in chickens, pigs, dogs, horses and rabbits. Several emerging features as regards the phenotypic evolution of domestic animals will be illustrated including:

- the importance of tissue-specific regulatory mutations;
- the importance of structural changes (duplications, deletions, inversions);
- evolution of alleles at loci under strong directional selection.

High-throughput sequencing data generation and the translation to medical practice

Ivo Gut

Centro Nacional de Analisis Genomico, Barcelona, Spain

2nd generation sequencing offers exquisite resolution of genetic features at an affordable cost. Whole human genomes can be sequenced for around 5000 Euros, focussing on the exomic regions of the human genome only costs less than 1000 Euros. However, many issues remain to be resolved before this supremely powerful technology can be taken into the clinic. Whole-genome sequencing is still fairly slow and can take even under optimal conditions close to one month. A further issue is that the sequence data interpretation requires dedicated informatics pipelines and a substantial number of cpu hours (computer time). Accelerated computational methods that preferably can run on a small commodity-type computer are required. No best practices for sequence data analysis exist and even systems in the hands of experts can lead to results that have less agreement with each other than one would desire. It is clear that substantial efforts by the community will be needed for standardisation of variant calling. Another problem that we face is that current clinical practice, in the treatment of cancer patients, disposes of roughly 100 molecular tests that look at somatic variants present in a tumour or the expression of specific surface markers by the tumour. Nearly all of these tests are related to a treatment decision. An exome sequence analysis of a tumour can easily show hundreds of somatic variants in a tumour and hardly any of these overlap with the commonly used tests and will have an accepted treatment associated with it. The question is how to deal with information that is clearly disease-specific, but for which we do not yet know a clinical action? Or an alternative is that we observe a clinically actionable variant that has an approved treatment for another form of cancer. Can a drug be repositioned? The next problem is that as treatments become more individualized and adapted to a patient profile, e.g. if a sample shows multiple actionable variants, can different drug-treatment regimes be combined. The corresponding clinical trial would not available and clinicians are moving on uncertain legal ground. These are just a few of the issues that will need to be resolved before high-resolution genome sequencing will be able to be integrated into clinical practice.

Sequencing ancient genomes

Janet Kelso

Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany

The genomes of extinct hominin groups closely related to contemporary humans offer a unique opportunity to identify genetic changes specific to anatomically fully modern humans.

Although it is possible to recover mtDNA and occasionally even nuclear DNA sequences from well-preserved remains of organisms that are less than a few hundred thousand years old, the determination of ancient hominid sequences is particularly challenging due to DNA degradation, chemical modifications in the sequence and contamination.

Recent advances in large- scale sequencing technologies have made it possible to perform direct sequencing of fossil extracts. Using these next-generation sequencing technologies we have generated high quality genome sequences for two archaic hominin groups: Neandertals who were known to have lived in Europe and Western Asia until approximately 30 000 years ago, and Denisovans, a group that was newly described based on the genome sequence generated from a bone found in Southern Siberia. The availability of these high coverage genomes provides multiple insights into the population history of both archaic and modern humans. They also allow us to identify sequence differences that have come to fixation or reached high frequency in modern humans since their divergence from Neandertals and Denisovans, some of which may have important functional effects in modern humans. I will outline some of the challenges in the generation and analysis of ancient genome sequence data, and discuss the evolutionary insights that have resulted from the sequencing of these genomes.

Structural variation and the plant pan genomes

Michele Morgante
Dipartimento di Scienze Agrarie ed Ambientali, Università di Udine, Udine, Italy
Istituto di Genomica Applicata, Parco Scientifico di Udine, Italy

The analysis of variation in plants has revealed that their genomes are characterised by high levels of structural variation, consisting of both smaller insertion/deletions, mostly due to recent insertions of transposable elements, and of larger insertion/deletion similar to those termed in humans Copy Number Variants (CNVs). These observations indicate that a single genome sequence might not reflect the entire genomic complement of a species, and prompted us to introduce the concept of the plant pan-genome, including core genomic features common to all individuals and a Dispensable Genome (DG) composed of partially shared and/or non shared DNA sequence elements. The very active transposable element systems present in many plant genomes may account for a large fraction of the DG. Both the mechanisms by which the CNV-like variants are generated and the direction of the mutational events are still unknown. Uncovering the intriguing nature of the DG, i.e. its composition, origin and function, represents a step forward towards an understanding of the processes generating genetic diversity and phenotypic variation. Additionally, since the DG clearly appears to be for the most part the youngest and most dynamic component of the pan genome, it is of great interest to understand whether it is a major contributor to the creation of new genetic variation in plant evolution as well as in the artificial selection processes of plant breeding. We will discuss the extent and composition of the pan genome in different plant species, the different mechanisms that generate and maintain the dispensable portion, the phenotypic effects of the DG and the rates and modes of creation of new genetic variation due to DG components.

The role of gene architecture in gene expression

Karla Neugebauer
Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

My lab is interested in the interplay between transcription and splicing. Recent global studies in several species have indicated that most intron removal is co-transcriptional, underscoring the potential for coupling between the transcription and splicing machineries as well as chromatin. Indeed, we recently discovered Terminal Exon Pausing, in which transcription elongation slows in short last exons of genes in budding yeast (Carrillo Oesterreich *et al.*, 2010). The effect of TEP is to ensure that pre-mRNA splicing is completed before transcription termination. We have further pursued the role of gene architecture in transcription and processing, showing that short first exons act as transcriptional enhancers (Bieberstein *et al.*, 2012). In addition, we study RNA-protein interactions and their functions (SR proteins and the cap binding complex) as well as the structure and function of RNA rich cellular compartments, such as Cajal bodies. Our work is conducted in mammalian tissue culture cells, zebrafish embryos, and budding and fission yeasts. On this occasion, I will report on a new study investigating the role of gene architecture in the activation of transcriptional programs in zebrafish embryos.

References:

Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM. (2012) First exon length controls active chromatin signatures and transcription. *Cell Rep.* **2**(1), 62-68.

Carrillo Oesterreich F, Preibisch S, Neugebauer KM. (2010) Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol Cell.* **40**(4), 571-81.

Oral Presentations

Supporting NGS Pipelines in the cloud

Ignacio Blanquer Blanquer[1], Goetz Brasche[2] Jacek Cala[3], Fabrizio Gagliardi[2], Dennis Gannon[2], Hugo Hiden[3], Hakan Soncu[2], Kenji Takeda[2], Andrés Tomás[1], Simon Woodman[3] ✉

[1]Insitute of Instrumentation for Molecular Imaging – I3M, Universitat Politècnica de València, València, Spain
[2]Microsoft Research, Aachen, Germany
[3]Newcastle University, Newcastle Upon Tyne, United Kingdom

Motivation and Objectives

The availability of workflow management systems (Romano, 2007, Smedley *et al.*, 2008) and public cloud computing infrastructures (Schatz *et al.*, 2010) have become a major breakthrough in the usage of computing resources for scientists / the science community. However, the combination of both approaches has shortcomings (Magellan Final Report, 2011, Blanquer *et al.*, 2012), such as the need to reduce administration effort to user, or the need for simple programming models for the transition from previous more conventional computing approaches and the support of legacy software. Although projects such as GATK, galaxy, 1000genomes (http://www.1000genomes.org/) use cloud, end users must have sophisticated knowledge of IT to deploy and use such resources. With this in mind, Microsoft Research (Microsoft Research– Cloud Research Engagement, 2013) has started several initiatives to improve the use of clouds in science. The "cloud4science" initiative, see http://www.cloud4science.eu/, considers next generation sequencing (NGS) as an excellence reference use case. This initiative builds on the results of the VENUS-C and e-Science Central projects, see http://www.venus-c.eu/ and http://www.escientificcentral.co.uk/, in which two different scientific workflow engines, namely the *Generic Worker and e-Science Central* were applied to solve specific bioinformatics problems requiring intensive computing. We propose an integration and enhancement of these two workflow engines with a set of selected bioinformatics tools to provide an easy-to-use framework for a cloud-enabled NGS pipeline for mutation analysis.

The resulting framework and components will simplify the deployment of processing services, the access to data and the sharing of the results.

Methods

The development of the framework focuses on three different categories: Computing resources; workload management and orchestration software; and bioinformatics legacy software.

As public cloud computing resources we selec-ted Microsoft's Windows Azure, see http://www.windowsazure.com. Windows Azure provides both users and application developers with different abstraction levels (PaaS and IaaS), including the support of Windows or Linux Virtual Machines (which is a requirement for many Bioinformatics tools). This makes Windows Azure an attractive platform to build upon.

The workload management and orchestration software need to deal with the two main types of NGS workflows: Coarse-grained data flows and fine-grained complex workflows. To tackle the first problem, the Generic Worker has been selected as it has been proven to be efficient in large-scale alignment problems (Carrión *et al.*, 2012). To deal with finer-grain complex workflows, eScience Central is used. E-science Central can scale very well in drug discovery problems (Cala *et al.*, 2012). Reference data for alignment is obtained from the UCSC Genome Bioinformatics repository (http://hgdownload.cse.ucsc.edu), and it is replicated in the Azure storage for convenience.

Finally, for the implementation of the pipeline, several tools have been selected for the initial prototype, covering sequence conversion (seqtk – https://github.com/lh3/seqtk and samtools - http://samtools.sourceforge.net/), Quality Control Analysis (fastqc - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), alignment (bowtie2 - http://bowtie-bio.sourceforge.net/bowtie2 and HPG aligner - http://docs.bioinfo.cipf.es/projects/hpg-aligner), Variant call file generation (GATK - http://www.broadinstitute.org/gatk/) and visualization (GenomeMaps - http://www.genomemaps.org/ and JBrowse - http://jbrowse.org/).

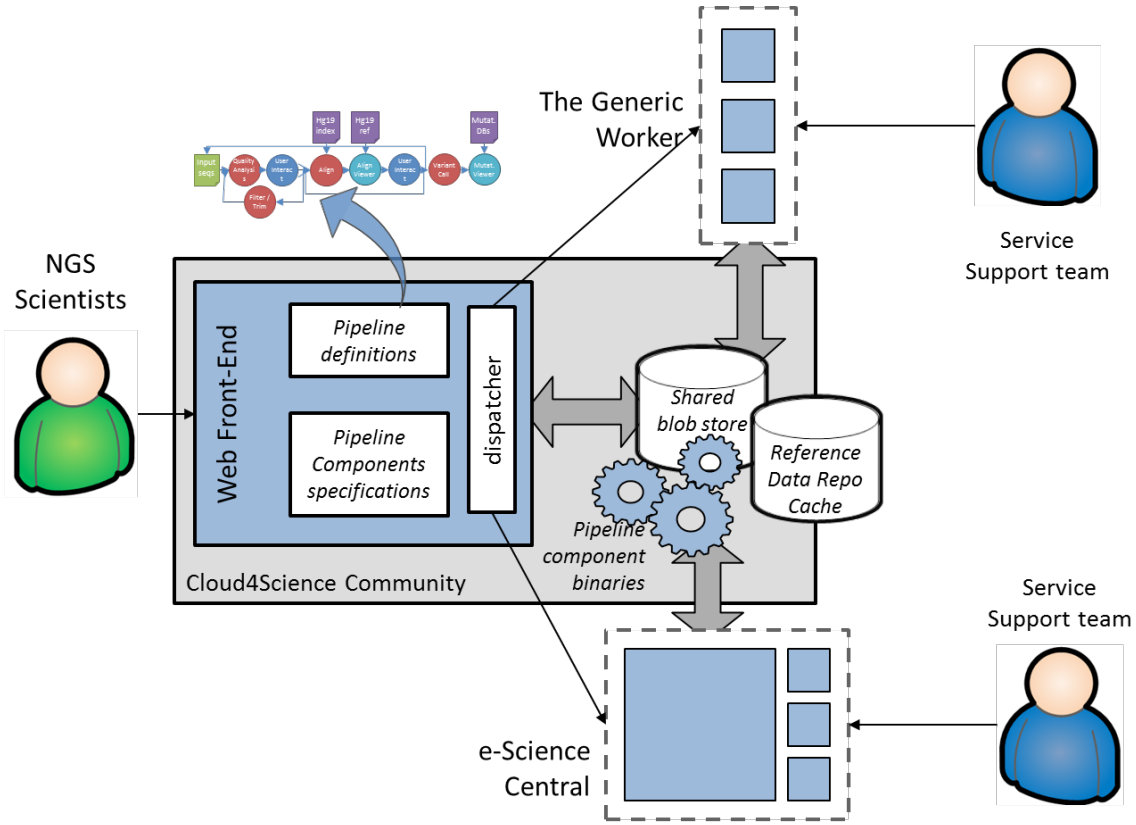


Figure 1. Simplified Architecture of the C4S platform-NGS framework

Results and discussion

The components are integrated through a shared blob store that hosts user-specific and shared data. As shown in Figure 1, the two different deployments of the enactment services are orchestrated through a web front-end, which exposes the interface to the users. The user interface is customized to guide the user through the different steps: uploading, conversion, quality control, trimming and filtering (repeatedly), alignment, visualization, quality control, variant call and final visualization. Interaction is managed through the use of server-based visualization tools, such as GenomeMaps. Early results of the performance achieved by the individual components demonstrated speed-up factors in the order of 63 times faster with 100 Azure cores (efficiency of 63%) using Generic Worker (Carrión *et al.*, 2012) and 176 times faster with 200 Windows Azure cores (efficiency of 88.2%) using e-Science Central (Cala *et al.*, 2012), . Improved results will be obtained by optimizing the data transfer among the processing nodes.

This first prototype will be improved in the first half of 2013 providing the users with more flexibility to alter the workflow, repeating or skipping individual or grouped steps and enhanced management of data to foster sharing and collaboration.

Eventually, it is important to note that the software in development in this project aims to becoming self-supported in the future by the establishment of an open end-user community, by the adoption of an open source software license scheme. Generic Worker and e-Science Central are released through Open Source licenses and the entire Cloud4Science environment will be released in the form of an open and collaborative project.

Acknowledgements

The authors want to thank Microsoft and the cloud4Science project for funding this research activity.

References

Blanquer I, Brasche G, Lezzi D: Requirements of Scientific Applications in Cloud Offerings. Proceedings of IBERGRID 2012, September 2012, in press.

Cala J, Hiden H, Woodman S, Watson P: Fast Exploration of the QSAR Model Space with e-Science Central and Windows Azure. Microsoft Cloud Futures, Berkeley, May 2012.

Carrión A, Blanquer I, Hernández V (2012) "A service-based BLAST command tool supported by cloud infrastructures", *Stud Health Technol Inform.* **175**, 69-77.

Magellan Final Report, December 2011, http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_final_report.pdf, last visited Jan 2013.

Microsoft Research – Cloud Research Engagement, http://research.microsoft.com/en-us/projects/azure/, last visited February 2013.

Romano P (2007) Automation of in-silico data analysis processes through workflow management systems. *Briefings in Bioinformatics* **9**(1), 57–68.

Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race, *Nature Biotechnology* **28**, 691–693. doi:10.1038/nbt0710-691.

Smedley D, Swertz MA, Wolstencroft K, Proctor G, Zouberakis M, *et.* al. (2008) Solutions for data integration in functional genomics: a critical assessment and case study. *Briefings in Bioinformatics* **9**(6), 532–544

Toward highly accurate and fast variant and de novo mutation identification from high-throughput sequencing data by joint Bayesian family calling

Francisco M. De La Vega, Mehul Rathod, Richard Littin, Len Trigg, John G. Cleary ✉
Real Time Genomics Inc., San Bruno, United States

Motivation and Objectives

Whole-genome sequencing (WGS) has become a fundamental tool in human disease research and is being adopted in clinical settings at an unprecedented rate. Whole-genome and exome sequencing has been successful in the elucidation of highly penetrant genes in early childhood diseases and its making inroads in complex trait studies entailing thousands of samples. As WGS becomes faster and moves into the clinic, e.g. into neonatal ICUs (Saunders *et al.*, 2012) and in prenatal screening (Talkowski *et al.*, 2012), there is an unmet need for both speed and accuracy in the analysis workflow. Due to its shotgun nature, mis-mapping of short reads in complex genomic regions and high sequencing error rates, calling variants from human high-throughput sequencing (HTS) data still results in substantial false positives and false negatives (Ajay *et al.*, 2011). The problem is magnified when looking for *de novo* mutations in affected offspring of families, as this enriches for sequencing artifacts (Veltman and Brunner, 2012). This is problematic since *de novo* mutations are thought to be responsible for about half of all early neurodevelopmental childhood disorders (Veltman and Brunner, 2012) and likely a similar fraction of neonatal/prenatal cases (Saunders *et al.*, 2012; Talkowski *et al.*, 2012).

Methods

In order to alleviate these problems, we developed a joint Bayesian calling framework which calls variants simultaneously across a pedigree leveraging shared haplotypes in its members and incorporating a Mendelian segregation model, to produce accurate variant and *de novo* mutation calls from HTS data. We present how our Bayesian framework escapes combinatorial explosion (as compared to more simplistic approaches), is highly scalable to large pedigrees, can deal with low coverage and missing data, and can call *de novo* mutations if desired (Conrad *et al.*, 2011). Coupled with our fast alignment method, a family of three 40X whole genomes can collectively be analyzed from reads to variant calls in ~30 hours on a single commodity server, and is amenable to large-scale parallelization for further speed improvements. To validate our method, we analyzed WGS data from a 3-generation CEPH family of 17 members produced by Illumina Inc. as part of their "Platinum Genomes" resource(). Each genome was sequenced with the HiSeq® 2500 system to 40X average depth using 2x100bp libraries of ~350bp insert size. We aligned reads and performed calls in 3 nuclear family subsets and the entire pedigree for comparison.

Results and Discussion

We focus our analysis on NA12878, a female in the second generation, for which extensive orthogonal validation data exists including fosmid-end Sanger sequence data (Kidd *et al.*, 2008), Complete Genomics WGS data, OMNI SNP-array genotype data (Consortium *et al.*, 2013) and experimentally validated germline and cell-line somatic *de novo* mutation data (Conrad *et al.*, 2011). As compared to naïve singleton calling, our family caller produced more high quality SNV/indel/MNP calls and eliminates low quality calls, as judged by commonly used quality metrics such as Ti/Tv, Het/Hom ratios, and dbSNP/OMNI array concordance. All this with a low 2.5% FP rate as assessed by variants called at monomorphic sites in the OMNI array (Consortium *et al.*, 2013); cf. Table 1, below.

Table 1. Summary statistics and quality metrics comparing singleton and family calling.

Quality metrics	SNVs	Indels/ MNP	Ti/Tv	Het/Hom	% dbSNP (r129)	OMNI TP	OMNI FP
Singleton calls	3573672	775857	2.05	1.66	89.5	98%	2.4%
Family calls	3469745	665964	2.1	1.59	89.2	98%	2.5%
Pedigree analysis	**Mendelian errors**	**de novo candidates**	**de novo segregants**	**de novo germline**	**Germline sensitivity (%)**	**de novo somatic**	**Somatic sensitivity (%)**
Singleton calls	101204	16902	14341	47	96%	878	92%
Family calls	8672	2667	295	47	96%	872	92%

As compared with the Conrad *et al* (Conrad *et al.*, 2011) validated *de novo* mutations set, we observed 96% and 92% sensitivity in detecting reported germline and *de novo* mutations, respectively (note that the cell line batch may be different and thus have different somatic mutations). While high sensitivity can be achieved by simply reporting variants that pass less stringent accuracy thresholds (and in so doing increasing substantially the number of variants that violate Mendelian segregation), our family calling achieves high sensitivity, delivering a 10X reduction in Mendelian errors from 101,204 to 8,672 (cf. Table 1). A further 10X reduction in Mendelian violations can be achieved without using the *de novo* priors, which would be appropriate when assuming inherited disease. Through the analysis of variant segregation to the third generation, we confirmed 99% of the Conrad *et al*(Conrad *et al.*, 2011) germline mutations (somatic variants do not segregate, as expected) and observed about ~250 new *de novo* mutation candidates, which is close to expectation (about 100 from previous studies (Conrad *et al.*, 2011)). Importantly, the high de novo sensitivity of 96% was achieved while reducing the number of candidate *de novo* mutations by greater than 6-fold, from 16,902 candidates to 2,667 *de novo* candidates, without using empirical filters (this is ongoing work we will report at the conference). Our results suggest that joint family calling produces more accurate calls than singleton calling and allows for the assessment of *de novo* mutation candidates with much less noise. We illustrate the impact of an improved call set in the downstream interpretation analysis of a simulated cased from the literature, and a real case from a cardio-pulmonary syndrome. We believe the analytical advances we present are crucial for the clinical adoption of genome and exome sequence data in family disease studies and beyond.

References

Ajay,S.S. *et al.* (2011) Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498–1505.

Conrad,D.F. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, **43**, 712–714.

Consortium,T.1.G.P. *et al.* (2013) An integrated map of genetic variation from 1,092 human genomes. *Nature* **490**, 56–65.

Kidd,J.M. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64.

Platinum Genomes Platinum Genomes Illumina, Inc.

Saunders,C.J. *et al.* (2012) Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. *Sci Transl Med* **4**, 154ra135–154ra135.

Talkowski,M.E. *et al.* (2012) Clinical Diagnosis by Whole-Genome Sequencing of a Prenatal Sample. *N Engl J Med* **367**, 2226–2232.

Veltman,J.A. and Brunner,H.G. (2012) De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565-575

Automated transcription start site prediction for comparative transcriptomics using the SuperGenome

Alexander Herbig[1], **Cynthia Sharma**[2], **Kay Nieselt**[1] ✉

[1]University of Tübingen, Tübingen, Germany
[2]University of Würzburg, Würzburg, Germany

Motivation and objectives

RNA deep-sequencing (RNA-Seq) has been revolutionizing eukaryotic and prokaryotic transcriptome analyses. Next-generation sequencing platforms allow one to sequence all RNA species in an organism within a couple of hours to days, and so keep accumulating massive amounts of transcriptome data. However, the bioinformatics-based analysis of this data is lagging behind. Very often, transcriptome features such as transcription start sites (TSS) and novel ncRNAs are still manually annotated, which is laborious and time-intensive. The problem is compounded for comparative transcriptomics of several species within a genus. A comparative approach would allow for refining the transcriptome annotation of the individual species by integrating the information from multiple species. This would not only lead to much better transcriptome and genome annotations but can also reveal differences in gene expression among species.

However, due to differences between the genomic architectures of the genomes, which are the result of insertions, deletions or genomic rearrangements, a direct comparison of RNA-seq data is infeasible.

Here we present two complementary approaches to solve this problem. Firstly, we developed the SuperGenome algorithm, which computes a common coordinate system for all genomes in a multiple alignment (Herbig *et al.*, 2012). The SuperGenome can be utilized for comparative analyses such as gene expression analysis, promoter sequence comparison or SNP calling. Furthermore it can be used for the comparative detection of TSS for which we - secondly - developed an automated TSS prediction method for dRNA-seq experiments (differential RNA-seq, Sharma *et al.*, 2010).

Methods

For the construction of the SuperGenome first a multiple genome alignment using *Mauve* (Darling *et al.*, 2004) is computed. Based on this alignment, the SuperGenome as a common genomic coordinate system and a mapping of each position of each single genome to a position in the SuperGenome is calculated. Next, all genome-specific data can be mapped to the common coordinate system, which includes genomic annotations or sequence information but also expression values derived from mapped read data, thus making a direct comparison of these data possible.

Our automated TSS prediction approach consists of several steps: first an initial detection of TSS in each genome (species) used in the experiment is conducted. Here, positions are localized, where a significant number of reads start (in comparison to local background). To evaluate if the reads starting at this position originate from primary transcripts, the enrichment factor is calculated by comparing the data from the standard library with a library that has been treated with terminator exonucleases (TEX), which specifically degrades processed RNAs with a 5'-mono-phosphate (Sharma *et al.*, 2010). For all positions where these values exceed the thresholds a TSS candidate is called. In the next step the TSS candidates of each species are mapped to the SuperGenome to assign each TSS to the corresponding TSS in the other species. Finally, all TSS are then characterized on the SuperGenome level with respect to their occurrence in the different species.

Results and discussion

Here we present the application of our SuperGenome approach and TSS prediction method to an RNA-seq experiment using four *Campylobacter jejuni* strains.

Mapping of RNA-seq data into the SuperGenome allows for a direct comparison of expression patterns among the four strains, e.g. for a visual comparison in a genome browser (Figure 1). In combination with our TSS prediction

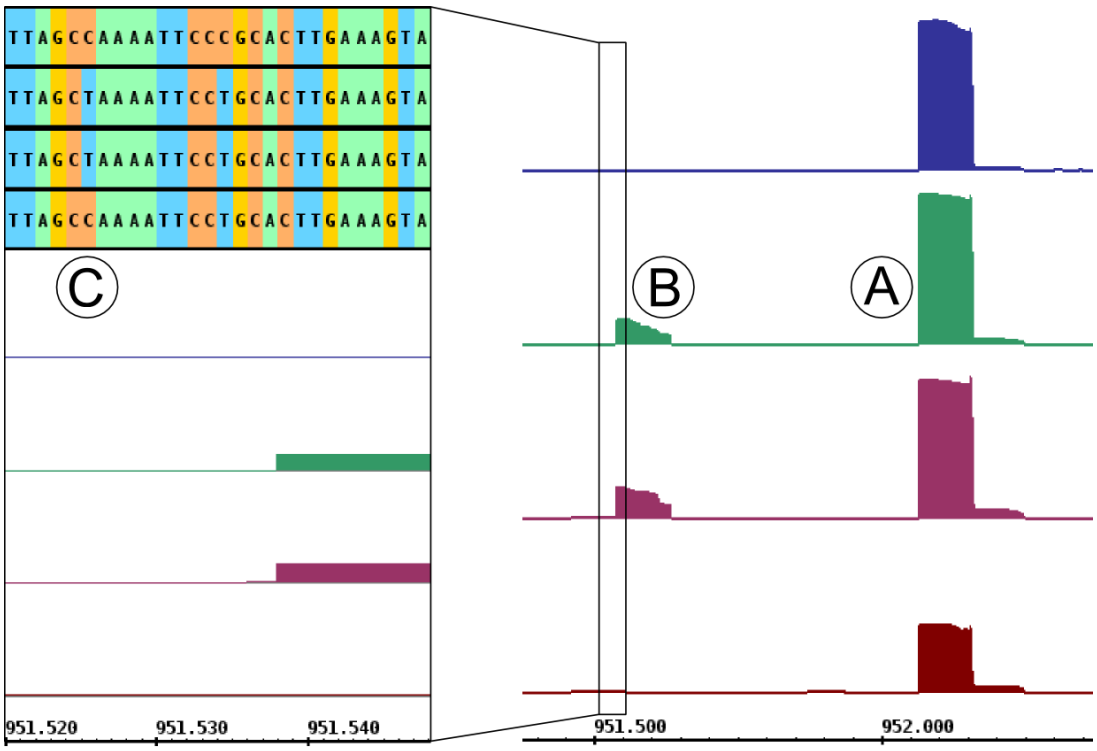


Figure1. Visualization of RNA-seq expression data of four C. jejuni strains. The graphs have been projected into the SuperGenome coordinate system (bottom). A: Expression pattern indicating a TSS that is conserved in all strains. B: Expression pattern showing a TSS that is only conserved in two of the strains. A close-up view of the respective promoter region is shown on the left. C: Alignment information from the SuperGenome reveals a SNP in the promoter region, which possibly induces the strain-specific expression pattern.

method, we are thus able to identify TSS in several species simultaneously and classify on the SuperGenome level whether they are detected in all strains (Figure 1A) or whether they are specific for only a subset of strains (Figure 1B).

Genome-wide application of our automated TSS prediction resulted in the annotation of more than 3000 TSS of which more than 1000 were detected in all four strains.

In addition, the SuperGenome allows for a comparative analysis of promoter regions related to the detected TSS. By this means, variations in the promoter sequence can easily be identified potentially explaining differences in the transcriptomic architectures of the investigated organisms (Figure 1C). Combining these information can help to elucidate novel mechanisms of transcriptional regulation and explain phenotypic diversity, e.g., in the context of pathogenicity. Overall our high-resolution transcriptome map revealed regulatory elements and their conservation in multiple C. *jejuni* strains on a genome-wide scale.

In summary, our TSS prediction procedure in combination with the SuperGenome provides a novel approach to comparative analysis of RNA-seq data, facilitating the cross-genome annotation of transcriptome features such as TSS maps and promoter regions.

References

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**(7): 1394-403. doi:10.1101/gr.2289704

Herbig A, Jäger G, Battke F, Nieselt K (2012) GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*. **28**(12): i7-15. doi:10.1093/bioinformatics/bts217

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*. **464**(7286): 250-5. doi:10.1038/nature08756

Viral Metagenomics – New applications for the broad-range detection of viromes in veterinary and public health settings

Oskar Erik Karlsson[1,2,3], Martin Norling[2], Fredrik Granberg[1,3], Sándor Belák[1,3,4], Erik Bongcam-Rudloff[2] ✉

[1]Department of Biomedical Sciences and Veterinary Public Health (BVF), Swedish University of Agricultural Sciences, Uppsala, Sweden.
[2]SLU Global Bioinformatics Centre, Department of Animal Breeding and Genetics (HGEN), Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[3]The Joint Research and Development Division of SLU and SVA, OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine (OIE CC), Uppsala,Sweden
[4]Department of Virology, Immunobiology and Parasitology, (VIP), National Veterinary Institute (SVA), Uppsala

Motivation and Objectives

Metagenomic methods for detection of viruses provide new diagnostic tools to the veterinary and public health laboratories, with powerful capacities to detect and to monitor the viromes in clinical samples. The Metagenomics methodology is divided into three main activities or steps: (1) wet-lab methodology ;(2) sequencing; and (3) data analysis. Integrating all three parts is of critical importance for the result as well as the interpretation of those. Our groups at the OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine, Uppsala, Sweden and at the SLU Global Bioinformatics Centre, Uppsala, Sweden are working with the development and evaluation of the methodological and technological platforms for viral metagenomics. Together with the National Veterinary Institute (SVA), we develop and test methods for extraction of viromes, feasibility of sequencing platforms to deliver metagenomic data-sets and evaluate bioinformatics tools as well as combine them into software packages for analysis and exploration of metagenomes, for separation, classification, assembly and visualization of genomic data in metagenomic samples. The aim of the work is to provide insight into using the metagenomics approach for detection of emerging viruses, monitoring of wild life for known pathogens as well as providing a tool for rapid characterization of viral pathogens in outbreak situations.

Methods

Clinical samples are collected through the continued monitoring performed by the SVA, as well as through international contacts, both from wild-life and from domestic animals. Viromes are prepared for analysis by homogenization of material, DNase/RNase treatment and nucleic acid purification. The viral nucleic acids in the pre-processed samples are quantified and, depending on chosen sequencing platform, either processed directly or pre-amplified using random amplification methods. Data from sequencing is processed following a general paradigm; read QC, quality based filtering, rough assembly, homology search and visualization. The viromes are then characterized and the results are disseminated to the clinicians and to the health authorities.

Results and Discussion

The development of metagenomics as a tool for exploration of viromes has proven to be an extremely powerful technique. We have previously published several articles and are continuously developing both Wet-lab methodology, evaluating sequencing platforms and developing the bioinformatics analysis (Belák *et. al*. 2013; Granberg *et. al.* 2013).

Our current work focuses at virome isolation and amplification, development of modular bioinformatics tools for use within the diagnostic setting, and integration with the clinical sciences, as well as ongoing evaluation of sequencing technologies together with the Swedish National Infrastructure for Large-scale Sequencing and NGS equipped laboratories within Europe.

With the current development of new technological platforms, the availability of high-throughput sequencing moves from the core facilities out into the medium and small scale diagnostic labs. This provides re-emerging challenges in data analysis and interpretation as well as enormous educational needs. We aim at providing the capability to utilize these methods and

technologies in a meaningful way within the field of veterinary virology and public health.

Acknowledgements

This work was supported by the Award of Excellence (Excellensbidrag) provided to SB by the Swedish University of Agricultural Sciences (SLU).

The authors would also like to acknowledge support of Uppsala Genome Centre and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at Uppsala Genome Centre has been funded by RFI/VR "SNISS" Swedish National Infrastructure for large Scale Sequencing and Science for Life Laboratory, Uppsala.

Writing of this publication has been supported by the framework of the EU-project AniBioThreat (Grant Agreement: Home/2009/ISEC/AG/191) with the financial support from the Prevention of and Fight against Crime Programme of the European Union, European Commission – Directorate General Home Affairs. This publication reflects views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Belák S, Karlsson OE, Blomström AL, Berg M, Granberg F (2013) New viruses in veterinary medicine, detected by metagenomic approaches. Vet Microbiol. Feb 1. doi: 10.1016/j.vetmic.2013.01.022.

Granberg G, Vicente-Rubiano M, Rubio-Guerri C, Karlsson OE, Kukielka D, Belák S, Sánchez-Vizcaíno JM. (2013) Metagenomic Detection of Viral Pathogens in Spanish Honeybees: Co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses. PLOS One. Feb 27. doi:10.1371/journal.pone.0057459

Automated and traceable processing for large-scale high-throughput sequencing facilities

Luca Pireddu, Gianmauro Cuccuru, Luca Lianas, Matteo Vocale, Giorgio Fotia, Gianluigi Zanetti ✉
CRS4, Pula, Italy

Motivation and Objectives

Scaling up production in medium and large high-throughput sequencing facilities presents a number of challenges. As the rate of samples to process increases, manually performing and tracking the center's operations becomes increasingly difficult, costly and error prone, while processing the massive amounts of data poses significant computational challenges. We present our ongoing work to automate and track all data-related procedures at the CRS4 Sequencing and Genotyping Platform, while integrating state-of-the-art processing technologies such as Hadoop, OMERO (Allan, 2012), iRODS (Rajasekar, 2010), and Galaxy (Goecks, 2010) into our automated workflows.

Our main objective is to completely automate workflows from the moment the sequencer is started to the delivery of the processed sequencing data while keeping complete data traceability. The second, future, objective will be to reach full automation of selected pipelines for NGS downstream analysis such as variant calling. Through this effort, the Platform aims to cut the number of operator-hours required to process each sample while lowering its problem rate and costs. Simultaneously, the system will enable immediate access to detailed documentation of the processing undergone by each sample, even for special processing recipes.

Methods

To overcome the challenges related to the processing and tracking of large volumes of data and a large number of samples, we are building an automated and traceable processing platform based on four core services: Hadoop, OME Remote Objects, iRODS and Galaxy.

The computational heavy lifting with the sequencing platform is performed with the Hadoop framework. To enable this technology within the context of a sequencing center, we have adopted the Hadoop-based Seal toolkit (Pireddu, 2011), SeqPig (http://seqpig.sf.net), and Pydoop (Leo, 2010). Though CRS4 has its own computing facility with more then 3200 processors, it is not dedicated exclusively to the sequencing platform; therefore, an "elastic" Hadoop cluster allocation scheme has been developed in-house to allow effective use of shared storage and distributed computational cluster. In short, a Job Tracker (head node) is always up and ready to accept jobs. New worker nodes are allocated on demand through a standard Grid Engine queuing system. For this approach to work, we forego the HDFS file system and instead use a shared GPFS.

As the sequencing platform processes a relatively large number of samples, tracking their progress and tracing the specific processing steps applied to the resulting data is not a simple feat. To address this issue we have extended the OMERO system with models for high-throughput sequencing data; OMERO is a flexible, client-server, model-driven data management platform for experimental biology. Within the architecture presented in this work, it is the element that recalls how a given datum was produced – both in the source data used and the operations applied.

The sequencing operation generates a significant amount of data split over a large number of files and data sets. In addition, frequent collaborations with geographically dispersed entities introduce a requirement for fast and controlled remote access to data. To simplify and manage access to the data we have adopted iRODS, which provides a single point of access for all data sets, which may instead be distributed across a number of disjoint storage systems. In addition, the system allows one to associate meta-data and storage policies to data files and collections – e.g., compress all text files larger than 500 KB – and implements an optimized data transfer protocol.

To glue all these components together our architecture relies on the Galaxy web-based workflow engine. We have extended the Galaxy fork

created by Brad Chapman at the Harvard School of Public Health Bioinformatics to handle native Illumina flow cell descriptors and support the retrieval of all such information via a web service. Thus, the laboratory enters the complete flow cell composition through the modified Galaxy web interface; then, through the web service the sample tracking software can fetch the flow cell information and integrate it with the operations-related meta-data.

In addition, we have integrated with Galaxy the Hadoop-based tools used by our pipeline. Therefore, Galaxy is used to manage all workflow-based operations, while a custom "automator" daemon is used to execute and monitor workflow progress and to link workflows to each other – e.g., execute sample-based workflows after a flow cell-based workflow. One of the main advantages of this approach is that Galaxy natively tracks operations with its histories. In our system, successful completion of a workflow will trigger an action within the automator that will commit the data set to iRODS, extract the associated history from Galaxy and save it to OMERO. Additional integration work exposes these data sets registered within iRODS through the Galaxy libraries feature, from where users can perform their additional analyses on the data.

Results and Discussion

Currently, the core system is in its testing phase and it is on schedule to be in production use at CRS4 by May 2013. The results thus far obtained by combining Hadoop, OMERO, iRODS and Galaxy are encouraging and the authors are confident that the CRS4 Platform will increase its efficiency and capacity thanks to this system. In the near future, we plan to release the integration components as open source software. In addition, we are also working on the extending the framework to integrate different pipelines for downstream analysis of the sequencing data with a focus on microbiology and metagenomics.

Acknowledgements

This work has been funded by the Sardinian (Italy) Regional Grant L7-2010/COBIK.

References

Allan C., Burel J.-M., Moore J., Blackburn C., Linkert M., *et al.* (2012). OMERO: flexible, model-driven data management for experimental biology. *Nature Methods*, **9**(3), 245–253. doi:10.1038/nmeth.1896

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

Leo S, Zanetti G. (2010) Pydoop: a Python MapReduce and HDFS API for Hadoop. *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, 819–825. doi:10.1145/1851476.1851594

Pireddu L, Leo S, Zanetti G. (2011) SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics* **27**(15), 2159-2160. doi:10.1093/bioinformatics/btr325

Rajasekar A, Moore R, Hou CY, Lee CA, Marciano R, *et al.* (2010) iRODS Primer: Integrated Rule-Oriented Data System. *Synthesis Lectures on Information Concept, Retrieval and Services* **2**(1), 1-143.

Challenges in whole exome sequencing to identify disease-causing variants in human rare diseases

Javier Santoyo Lopez ✉
Medical Genome Project (MGP), Genomics & Bioinformatics Platform of Andalusia (GBPA), Seville, Spain

Next Generation Sequencing (NGS) Technologies have greatly improved our ability to mine variants out of the entire genome, and the extensive use of Whole Exome Sequencing (WES) has allowed the discovery of variants responsible of disease especially in monogenic diseases. However, finding disease-causing variants requires a primary and secondary analysis where several factors that can affect a successful outcome need to be taken into consideration.

Using Life Technologies SOLiD sequencers and Nimblegen SeqCapEZ exome capture library, we have systematically analyzed more than 300 exomes accounting for 20 different diseases and more than 200 exomes from control Spanish population within the Andalusian Medical Genome Project (MGP), a public-private partnership that aims to find mutations responsible of hereditary Rare Diseases and to establish first steps towards the implementation of personalized medicine.

Thus, the reliability of calling variants in a sample is highly related to the sequencing instrument used due to the sequencing chemistry and the intrinsic properties of each sequencing technology, so in the first place we have optimized a primary analysis pipeline to obtain high-quality variants from color-space. This pipeline provides high-quality variants with an extremely low rate of false positives as shown by Sanger sequencing validations. Furthermore, the analysis of whole exome data, shows more than 98% correlation of normalized coverage for library pairs and the distribution of coverage is consistent between samples, which shows a good reproducibility of the enrichment and sequencing technology.

After sample variant calling a secondary analysis pipeline filters those variants that are present in variant public databases and have a MAF that indicates that are SNPs. The filtering of SNPs from local population is a critical step, as there are many SNPs in our study population that are not represented in the public databases (e.g. dbSNP or 1000 Genomes), which is done by using MGP database for WES Spanish Healthy Control Individuals. Finally mutations are filtered based in family pedigree keeping only those mutations that segregate with the disease.

All this WES approach and subsequent variant analysis has allowed us to find the gene and the mutations responsible of disease in a great number of retinal dystrophies in particular for Retinitis Pigmentosa Autosomal Recessive (RPAD).

Developing a software suite to analyze the interplay between nucleosome arrangement, DNA methylation and transcription factor binding

Vladimir B. Teif, Daria A. Beshnova, Yevhen Vainshtein, Thomas Höfer, Karsten Rippe ✉
German Cancer Research Center, Heidelberg. Germany

Motivation and Objectives

Many of the underlying cell fate decisions occur via changes of chromatin features that affect gene expression. The specific location of nucleosomes on the DNA has important functions in controlling access to the DNA. Binding of protein factors to the 145-147 bp of DNA wrapped around the histone octamer core is frequently impeded, while the linker DNA between nucleosomes is more easily accessible. Recent advancements in high-throughput sequencing methods allowed genome-wide mapping of individual nucleosomes at single base pair resolution, with yeast serving as a model system for the initial pioneering studies. More recently, tissue- and disease-specific features of nucleosome positions in higher organisms were reported. These include for example studies of human and mouse cells. However, the large amount of data, as well as complex relation between different data types, poses a challenge to the understanding the underlying biology. Here we set to develop computational software capable to tackle this problem and applied it to the analyze the interplay between nucleosome arrangement, DNA methylation and transcription factor binding.

Methods

In a recent work (Teif *et al.*, 2012), we determined genome-wide nucleosome occupancies in mouse embryonic stem cells and their neural progenitor and embryonic fibroblast counterparts to assess features associated with nucleosome positioning during lineage commitment. Cell type and protein specific binding preferences of transcription factors to sites with either low (e.g. Myc, Klf4, Zfx) or high (e.g. Nanog, Oct4 and Sox2) nucleosome occupancy as well as complex patterns for CTCF were identified. Nucleosome depleted regions around transcription start and termination sites were broad and more pronounced for active genes, with distinct patterns for promoters classified according to CpG-content or histone methylation marks. Throughout the genome nucleosome occupancy was correlated with certain histone methylation or acetylation modifications. In addition, the average nucleosome-repeat length increased during differentiation by 5-7 base pairs, with local variations for specific regions. Our results revealed regulatory mechanisms of cell differentiation that involve nucleosome repositioning. Here we have used this dataset to develop and test the computational software for the analysis of this type of data (Teif *et al.*, 2013; Teif and Rippe, 2012). To demonstrate its applicability, we combined these data with DNA methylation and hydroxymethylation data, as well as the maps of histone variants and DNA-modifying enzymes and studied the interplay of nucleosome positioning DNA methylation and TF binding.

Results and Discussion

We have developed a software suite for the biophysical analysis of high-throughput sequencing experiments, and applied it to the interplay between nucleosome positioning, DNA methylation and transcription factor binding during ES cell differentiation. DNA cytosine methylation (5mC) and hydroxymethylation (5hmC) are among the most important epigenetic marks. The interplay of 5mC/5hmC marks, the arrangement of nucleosomes and transcription factors (TFs) links DNA methylation with cellular gene expression programs but the underlying mechanisms are poorly understood. Here, we analyzed nucleosome positioning, DNA methylation and TF binding in conjunction with additional dinucleosome occupancy maps. Our study provides a novel quantitative description for the relations between DNA methylation/demethylation, TF binding and nucleosome occupancy changes.

Acknowledgements

We are grateful to the DKFZ Sequencing Core Facility for conducting the sequencing. This work was funded within project EpiGenSys by the German Federal Ministry of Education and Research (BMBF) as a partner of the ERASysBio+

initiative in the EU FP7 ERA-NET Plus program. Computational resources and data storage were provided via grants from the BMBF (01IG07015G, Services@MediGRID) and the German Research Foundation (DFG INST 295/27-1). V.T. acknowledges the support from the Heidelberg Center for Modeling and Simulation in the Biosciences (BIOMS) and a DKFZ intramural grant.

References

Teif VB, Erdel F, Beshnova DA, Vainshtein Y, Mallm JP, Rippe K (2013) Taking into account nucleosomes for predicting gene expression.

Teif VB and Rippe K (2012) Calculating transcription factor binding maps for chromatin. *Brief Bioinform* **13**, 187-201.

Teif VB, Vainstein E, Marth K, Mallm J-P., Caudron-Herger M, Höfer T, Rippe K (2012). Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol* **19**, 1185-1192.

RSAT peak-motifs: Efficient prediction of transcription factor motifs and binding sites from genome-wide sequencing peak sets

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

[1]Institut de Biologie de l'Ecole Normale Supérieure, Paris, France
[2]Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico
[3]CNRS-UMR8199 Institut de Biologie de Lille, Lille, France
[4]INSERM U928 & Université de la Méditerranée, Marseille, France

Motivation and objectives

ChIP-seq is increasingly used to characterize transcription factor binding and chromatin marks at a genomic scale. Although 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 they typically restrict motif discovery to a few hundred peaks, or to the central-most part of the peaks. To interpret genome-wide location data, there is a crucial need for time- and memory-efficient algorithms, interfaced as user-accessible tools to extract relevant information from high-throughput sequencing data.

For this purpose, we developed the software tool *peak-motifs* (Thomas-Chollier *et al.*, 2012a), which takes as input a set of peak sequences of interest, discovers key motifs, compares them with transcription factor binding motifs from various databases, predicts the location of binding sites within the peaks and exports them in a format suitable for visualization in the UCSC Genome Browser. Notably, all these steps, including motif discovery, are performed on the full-size sets of peak sequences, without restrictions on peak number or width.

Methods

The motif discovery step relies on a combination of algorithms integrated in the software suite regulatory sequence analysis tools (RSAT, http://rsat.ulb.ac.be/rsat/) (Thomas-Chollier *et al.*, 2011), which use complementary criteria to detect exceptional words (oligonucleotides and spaced motifs): global over-representation of oligonucleotides (*oligo-analysis*) or spaced pairs (*dyad-analysis*), heterogeneous positional distribution (*position-analysis*) and local over-representation (*local-word-analysis*).

The motif comparison step is performed by *compare-matrices* (Thomas-Chollier *et al.*, 2011), which supports a wide range of scoring metrics and displays the results as multiple alignments of logos, enabling to grasp the similarities between a discovered motif and several known motifs. This feature is particularly valuable to reveal adjacent fragments of the discovered motif showing similarities with two distinct known motifs, suggesting a bipartite motif for two factors.

Sequences are scanned with the discovered motifs to locate binding sites, and their positioning within peaks is analyzed (coverage, positional distribution along peaks).

Peak-motifs generates an HTML report summarizing the main results and giving access to each separate result file. The report page includes links, allowing users to upload input peaks and predicted sites to the UCSC Genome Browser in order to visualize them in their genomic context.

Results and discussion

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 (Chen *et al.*, 2010; Thomas-Chollier *et al.*, 2012a).

Similarly, for ChIP-seq data targeting the generic transcriptional co-activator p300, peak-motifs identified motifs bound by tissue-specific transcription factors consistent with these two tissues (Visel *et al.*, 2009; Blow *et al.*, 2010; Thomas-Chollier *et al.*, 2012a).

We assessed the time efficiency of *peak-motifs* by analyzing data sets of increasing sizes (from 100 to 1 000 000 peaks of 100 bp each), with total sequence sizes ranging from 10 kb to 100 Mb. The computing time of the motif discovery algorithms integrated in peak-motifs increases linearly with sequence size and outperforms all the other existing motif discovery tools used in our comparison (Thomas-Chollier *et al.*, 2012a). Data sets of several tens of megabytes are processed in a few minutes on a personal computer (the most efficient tool, *oligo-analysis*, treats 100Mb in 3min). This linear time response enables peak-motifs to scale up efficiently with sequence size, and allows us to provide an easy access via a web interface, without any data size restriction. This moreover gives us the possibility to run four distinct algorithms in order to detect motifs of various types (oligonucleotides, spaced pairs) based on complementary criteria (over-representation, positional heterogeneity).

In conclusion, peak-motifs supports time-efficient and statistically reliable analysis of complete ChIP-seq datasets, while offering an online user-friendly and well-documented interface, as well as a detailed protocol (Thomas-Chollier *et al.*, 2012b)

Acknowledgements

M.T-C is supported by the Alexander von Humboldt foundation.

References

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-10.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-17.

Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J (2011). RSAT 2011: Regulatory Sequence Analysis Tools. *Nucleic Acid Research* **39**: W86-91.

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2012a). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acid Research* **40**: e31.

Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012b). From peaks to motifs: a complete workflow for full-sized ChIP-seq (and similar) datasets. *Nature Protocols* **7**: 1551-68.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-8.

Translational systems biology understanding the limits of animal models as predictors of human biology

Carine Poussin[1], Leonidas Alexopoulos[2], Vincenzo Belcastro[1], Erhan Bilal[3], Carole Mathis[1], Pablo Meyer[3], Raquel Norel[3], Jeremy J Rice[3], Gustavo Stolovitzky[3], Julia Hoeng[1], Manuel Peitsch[1] ✉

[1]Philip Morris Research and Development, Neuchatel, Switzerland
[2]Protatonce Ltd, Glyfada, Greece
[3]IBM, New York, United States

Motivation and objectives

Inferring how humans respond to external cues such as drugs, chemicals, viruses or hormones is an essential question in biomedicine. Very often, however, this question cannot be addressed since it is not possible to perform experiments in humans. A reasonable alternative consists of generating responses in animal models and "translating" the results to humans. The limitations of such translation, however, are far from clear, and systematic assessments of its actual potential are urgently needed.

A series of challenges have been designed in the context of the 'sbv IMPROVER' project (Industrial Methodology for Process Verification in Research; http://sbvimprover.com/) to address the issue of translatability between humans and rodents. Our aim is to understand the limits and opportunities of species to species translatability at different levels of biological organization: signalling, transcriptional, and release of secreted factors (such as cytokines, chemokines or growth factors).

Methods

Normal Bronchial Epithelial Cells of both human and rat origin were selected to address this question. The cells were exposed to more than 50 different substances and for each stimulus, samples were collected at different time points to generate phosphoproteomic (After 5 and 25min stimulus exposure), gene expression (after 6 hours) and secreted protein (after 24 hours) data.

Our challenge will provide participants with both training and test data sets which are designed to assess the ability of methods to predict the responses in Normal Human Bronchial Epithelial cells, from the responses observed in Normal Rat Bronchial Epithelial primary cells.

The central questions that we will pose in this challenge are: (1) Can the phosphoproteomic responses induced by stimuli addressing several distinct signalling pathways in human cells be predicted given the responses generated with the same stimuli in rat cells? How does the accuracy of the prediction depend on the nature of the applied perturbation? (2) Which gene expression regulatory processes (biological pathways / functions) are translatable and therefore predictable across species, and which are too divergent?

Results and discussion

We will present the community with questions and large scale omics data aimed at assessing methodologies designed to infer human biology from non-human biology.

The sbv IMPROVER project, the website, and the Symposium are part of a collaborative project designed to enable scientists to learn about and contribute to the development of a new crowd sourcing method for verification of scientific data and results.

Acknowledgements

The project team includes scientists from Philip Morris International's Research and Development department and IBM's Thomas J. Watson Research Center. The project is funded by PMI.

References

Meyer P, Hoeng J, Rice JJ, Norel R, Sprengel J et al. (2012). Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. Bioinformatics **28**(9), 1193-201. doi:10.1093/bioinformatics/bts116

Meyer P, Alexopoulos LG, Bonk T, Califano A, Cho CR, et al. Verification of systems biology research in the age of collaborative competition (2011). Nat Biotechnol **29**(9):811-5. doi:10.1038/nbt.1968

Automated finished microbial genomes and epigenomes to understand infectious diseases

Ralph Vogelsang
Pacific Biosciences, United States

Motivation and objectives

Understanding the genetic basis of infectious diseases is critical to enacting effective treatments, and several large-scale sequencing initiatives are underway to collect this information[1]. Sequencing bacterial samplesis typically performed by mapping sequence reads against genomes of known reference strains. While such resequencing informs on the spectrum of single nucleotide differences relative to the chosen reference, it can miss numerous other forms of variation known to influence pathogenicity: structural variations (duplications, inversions), acquisition of mobile elements (phages, plasmids), homonucleotide length variation causing phase variation, and epigenetic marks (methylation, phosphorothioation) that influence gene expression to switch bacteria from non-pathogenic to pathogenic states2 (Srikhanta *et al.*, 2010). Therefore, sequencing methods which provide complete, *de novo* genome assemblies and epigenomes are necessary to fully characterize infectious disease agents in an unbiased, hypothesis-free manner.

Methods

Hybrid assembly methods have been described that combine long sequence reads from SMRT® DNA sequencing with short reads (SMRT CCS or second-generation reads), wherein the short reads are used to error-correct the long reads which are then used for assembly. We have developed a new paradigm for microbial *de novo* assemblies in which long SMRT sequencing reads (average read lengths>5,000 bases) are used exclusively to close the genome through a hierarchical genome assembly process, thereby obviating the need for a second sample preparation, sequencing run and data set.

Results and discussion

We have applied this method to achieve finished de novo genomes with accuracies exceeding QV50 (>99.999%) to numerous disease outbreak samples, including *E. coli*, *Salmonella*, *Campylobacter*, *Listeria*, *Neisseria*, and *H. pylori*. The kinetic information from the same SMRT sequencing reads is utilized to determine epi-genomes. Approximately 70% of all methyltransferase specificities we have determined to date represent previously unknown bacterial epigenetic signatures.

Conclusions

Our method allows for rapid and comprehensive elucidation of the genetic and epigenetic basis of infectious disease agents. The process has been automated and requiresless than 16 hours from an unknown DNA sample to its complete *de novo* genome and epigenome.

References

Srikhanta YN, Fox KL, Jennings MP (2010) The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat Rev Microbiol.* **8**(3), 196-206. doi: 10.1038/nrmicro2283

1 e.g., the 100K Foodborne Pathogen Genome Project (www.100kgenome.vetmed.ucdavis.edu/)

Interplay between DNE sequence motifs and the human epigenome

John William Whitaker[1], Zhao Chen[2], Wei Wang[2] ✉

[1]UCSD, San Diego, United States
[2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, United States

Motivation and objectives

Different mammalian cell-types display distinct phenotypes but posses the same genome. It is well established that epigenomic modification is important in establishing cell-type specific patterns of gene expression. Epigenomic modifications include the covalent modification of histone tails and the methylation of DNA. Epigenomic modifications function to mark regions of the genome as being active or repressed and their correct establishment is a critical aspect of mammalian development. Furthermore, correct recapitulation of the epigenome is key during cellular reprogramming, such as induced pluripotent stem cells (Lister *et al.* 2011; Won *et al.* 2012). Moreover, alterations in the epigenome are associated with disease such as cancers (Hon *et al.* 2012) and autoimmune diseases (Nakano *et al.* 2012).

The establishment and maintenance of the epigenome is regulated by many factors including: modifying enzymes, DNA binding proteins, non-coding RNAs, signaling molecules and three dimensional genomic organization. Herein, we investigate the involvement of DNA sequence motifs in the regulation of the epigenome. We demonstrate the involvement of DNA sequence motifs by constructing a series of predictive models that can predict the presence of six histone modifications in five different developmental cell-types, including human embryonic stem cells. Epigenomic modifications can span long variably length regions that are associated with GC content biases. Thus, great care was taken to avoid biases from influencing predictive performance.

Methods

We analyzed a comprehensive dataset of six core histone modifications (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3) in five primitive cell-types (human embryonic stem cells (H1), trophoblast-like, neural progenitor cell, mesendoderm and mesenchymal cells). To identify a broad set of DNA motifs that are associated with epigenome we used two *de novo* motif discovery programs: Homer (Heinz *et al.* 2010), and our own, Epigram. Then a LASSO logistic regression was use to identify the subset of motifs that had the greatest prediction performance (Friedman *et al.* 2010). Then a Random forest was trained to distinguish genomic regions that posses a modification from regions that do not.

Results and discussion

An integrative analysis of over 70 separate ChIP-Seq experiments shall be presented. The analysis pipeline first used to distinguish genomic regions that possess a modification from regions that do no possess any modifications. In H1 the average prediction performance across the six modifications was AUC = 0.85 (Figure 1). Further comparisons, identified that prediction performance was constituent in other cell-types and that models could be trained to distinguish a specified modifications from other modifications.

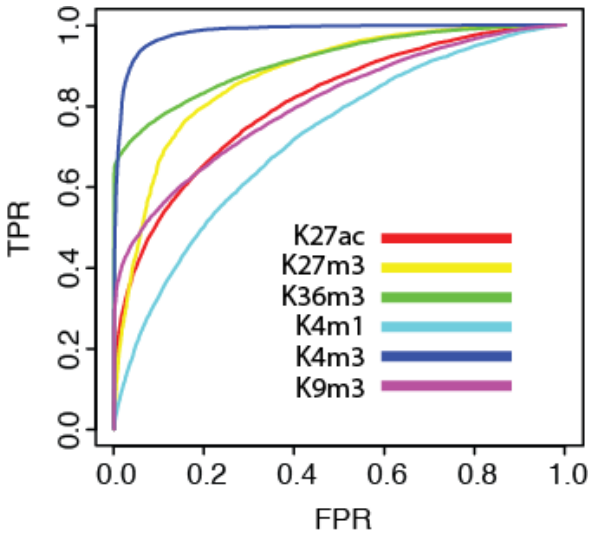


Figure 1. The prediction performance in of six histone modifications in human embryonic stem cells

Comparison of the identified motifs to known motifs identified interplay between known factors and the epigenome. For example, cell-type

specific marker genes are identified as being associated with H3K27ac, which marks active regions of the genome. Motif location preference analysis revealed that motifs occur at the edge of modification peaks and suggests they may function by establishing barriers.

The identified DNA motif and epigenome associations demonstrate interplay between sequence and epigenome. This work demonstrates the importance of large-scale integrative genomic analysis to gain complex biological insight. Identification of factors that interplay with epigenome in stem cells should improve the efficacy of cellular reprogramming strategies.

Acknowledgements

This work was partially supported by NIH.

References

Friedman, J., T. Hastie and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**(1): 1-22.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, *et al.* (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**(4): 576-589.

Hon, G. C., R. D. Hawkins, O. L. Caballero, C. Lo, R. Lister, *et al.* (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome research* **22**(2): 246-258.

Lister, R., M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, *et al.* (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**(7336): 68-73.

Nakano, K., J. W. Whitaker, D. L. Boyle, W. Wang and G. S. Firestein (2012). DNA methylome signature in rheumatoid arthritis. *Ann Rheum Dis.* **72**(1): 110-117

Won, K. J., Z. Xu, X. Zhang, J. W. Whitaker, R. Shoemaker, *et al.* (2012). Global identification of transcriptional regulators of pluripotency and differentiation in embryonic stem cells. Nucleic Acids Res. **40**(17): 8199-8209

Posters

Application of whole genome resequencing in the dissection of QTLs affecting boar taint

Rahul Agarwal[1,2], Maren Van Son[2,3], Matthew Peter Kent[1,2], Sigbjørn Lien[1,2], Eli Grindflek[3] ✉

[1]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, Norway
[2]Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, Ås, Norway
[3]NORSVIN (The Norwegian Pig Breeders Association), PO Box 504, 2304 Hamar, Norway

Motivation and objectives

Male piglets are usually castrated to remove urine like smell from the meat of some entire male pigs. The urine like smell also known as boar taint. However, castration also affects the reproductive and growth traits, which in turn causes the loss in revenue of pig industries. Andostenone and skatole are the major compounds behind the incidence of boar taint in intact boars. Hence, alternative method of boar taint selection in Norwegian pig breeding scheme requires the detection of functional genetic polymorphisms in association with boar taint without simultaneously interfering with the other traits especially male sex steroids responsible for reproduction and growth in male pigs. .

Methods

Fourty-seven Norwegian male pigs were resequenced using Illumina technology with mean coverage >10x. These animals are representing the whole Norwegian Duroc and Landrace population including the previously studied boar taint animals (Grindflek *et al.*, 2011). About 19.19 billion 100-bp paired end reads were obtained, which in turn filtered to remove duplicate reads, adaptor sequence from the reads, bad quality bases, and reads shorter than cutoff length using perl custom scripts. Remaining reads were mapped to pig build 10.2 reference sequence using the Bowtie2 mapping tool (Langmead *et al.*, 2012). Mapped reads were used for calling single nucleotide polymorphisms (SNPs) within the quantitative trait loci (QTL) regions on SSC13 across 23 Duroc samples and on SSC7 across total 47 samples using the Freebayes variant caller (Erik *et al.*,2012). To eliminate the bad quality SNPs, we applied some stringent filtering upon the detected SNPs. Filtered SNPs was classified to the different functional classes on the basis of the current annotation file available in the pig Ensembl. Of filtered SNPs, the subset of the SNPs will be genotyped using MassARRAY assays and then genotyped SNPs will be tested for associations to androstenone and other male sex steroids using AS-Reml. The BEAGLE version 3.3.1 software will be employed to deduce phase and impute missing genotypes. The estimation of linkage disequilibrium (LD) of the genotyped SNPs and then construction of haplotype blocks will be done with the Haploview software (Barrett, 2009).

Results and discussion

Recent mapping study for finding QTLs underlying boar taint compounds in Norwegian male pig breeds (Grindflek *et al.*, 2011) revealed the significant regions of interest on SSC7 which affect major boar taint compounds and on SSC13 which affect both androstenone and other three male specific sex steroids (testosterone, estradiol and estrone sulfate). Thus, fine mapping of these two regions and detection of underlying functional mutations would be attractive aims to pursue in this study. Two QTL regions on SSC7 are most likely located from 60.67 Mb to 68.40 Mb and 74.80 Mb to 80.53 Mb in build 10.2 with 77 and 46 protein coding genes respectively while region on SSC13 probably located from 25.2Mb to 25.9 Mb and having a total of 10 genes. Therefore, whole genome resequencing of 47 Norwegian Duroc pigs was performed and generated paired end reads were filtered and afterward filtered reads were mapped to latest pig reference genome with 77% mapping percentage. We detected a large number of SNPs within these three QTLs from mapped reads of 23 Duroc and 24 Landrace pigs simultaneously. Initially, many thousands of SNPs (with minimum support of 2 reads across samples) were called within QTLs on SSC7 and on SSC13 respectively. These SNPs was sorted out to remove those SNPs located in repeats and SNPs likely to be false. Finally, a sum of 8,591 and 1,231 SNPs within 2 QTLs on SSC7 and single QTL on SSC13 respectively were sorted out and annotated to exon, intron, 1000bp up-

stream and intergenic regions. A total of 59 non-synonymous SNPs (ns-SNP) , 93 synonymous SNPs (s-SNP), 6231 intergenic SNPs, 1995 intronic SNPs and 121 upstream SNPs were detected within two QTLs on SSC7 using SNPEff version 3.1. Similiarily, in case of SSC 13, a thousand of SNPs assigned under intronic and intergenic region, 60 exonic SNPs including both ns-SNP and s-SNP, and 12 SNPs located 1000bp upstream. The minor allele frequency (MAF) of the SNPs within QTLs on SSC7 and SSC13 was varying from 0.07-0.50 and the average values of total read depth comprising the sum of the read depth of reference allele and alternate allele were 250 and more than 400 in case of SSC13 and SSC7 respectively . As expected, the ratio between the number of transistions and transversions was estimated more than 2 within all the QTLs. So far, a potential candidate gene *ACVR2B* was identified within QTL on SSC13 and responsible for inhibitory activity of steroidogenic acute regulatory protein (*StAR*) and 3β-hydroxysteroid dehydrogenase (*3β-HSD*) which encode important enzymes for the biosynthesis of androstenone and other androgens. Seven SNPs were identified in this particular gene which might play important role in selection.

Acknowledgements

The work was financed by the Research Council of Norway. The samples and phenotypes are provided by the Norwegian pig breeders association (NORSVIN). We also highly appreciate the contribution from BioBank AS for collection and handling of samples, CIGENE for data handling, performing the primer design, sample preparations and SNP genotyping. Boar taint group (Sigbjørn Lien, Matthew P Kent, Maren van son, Eli Grindflek and me) for planning and executing this study.

References

Grindflek E, Lien S, Hamland H, Hansen MH, Kent M et al. (2011) Large scale genome-wide association and LDLA mapping study identifies QTLs for boar taint and related sex steroids. *BMC Genomics* **12**: 362. doi: 10.1186/1471-2164-12-362.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359. doi: 10.1038/nmeth.1923.

Erik Garrison, Gabor Marth (2012) Haplotype-based variant detection from short-read sequencing. http://arxiv.org/abs/1207.3907 (Submitted on 17 Jul 2012 (v1), last revised 20 Jul 2012 (this version, v2)).

Barrett JC (2009) Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harbor Protocols* **2009**: pdb.ip71 doi:10.1101/pdb.ip71.

Deep sequencing exposes small RNA transcriptome differences between low- and high-temperature stress responses in Arabidopsis

Vesselin Baev, Ivan Milev, Mladen Naydenov, Tihomir Vachev, Elena Apostolova, Nikolay Mehterov, Mariana Gozmanova, Ivan Minkov, Galina Yahubyan

University of Plovdiv, Plovdiv, Bulgaria

Motivation and Objectives

Plant small RNAs (sRNAs) include two major groups distinguished by their different modes of biogenesis - microRNAs (miRNAs) and small-interfering RNAs (siRNAs). Despite of numerous studies on the involvement of particular sRNA in plant stress response, there are only few reports on the genome-wide sRNA profiles generated under different stress treatments (Borsani *et al.,* 2005; Yan *et al.,* 2011; Katiyar-Agarwal *et al.,* 2011). To understand the involvement of the various sRNA groups in plant response to different stress factors, we characterized sRNA datasets produced by high-throughput sequencing (HTS) from Arabidopsis plants suffering from low-temperature (LT) and high-temperature (HT) stress for 24 hour. Our study for the first time presents the genome-wide sRNA profiles of LT and HT treated plants and reveals that they greatly differed in the first 24 hours of stress onset.

Methods

Using Solexa technology for HTS that produces highly accurate and quantitative readouts of sRNAs, three sRNA libraries were generated and sequenced. The libraries derived from RNA extracted from leaf tissue of A. thaliana plants grown under normal temperature conditions (21°C, NT library), and treated either with low temperature (4°C, LT library) or high temperature (36°C, HT library) for 24 h. We obtained 21,364,883, 20,607,777 and 22,542,401 total reads for NT, LT and HT library resp. After applying a quality-read filter and discarding low-quality reads, adaptor sequences were trimmed. Additionally, reads attributed to ligation contaminants or adaptor self-ligation were also removed, that finally resulted in 19,881,084 (NT), 19,231,094 (LT) and 21,852,807 (HT) clean reads. For the sRNA analysis, we further clustered all clean reads into unique sequences with associate copy numbers generating 2,134,578 (NT), 1,534,898 (LT) and 3,308,837 (HT) unique tags for the three libraries.

Results and Discussion

Under the NT and LT conditions, the sRNA populations were dominated by 21-nt sRNAs. Another important finding is that there are many protein-coding genes that give rise to differentially expressed sRNAs following temperature shifts. The LT treatment caused increased production of sRNAs of sense polarity from numbers of cold-responsive genes such as COR15A, COR47, COR413, KIN2, KIN1. The HT treatment induced production of sRNAs of sense and antisense polarity from many genes encoding functionally diverse proteins. Amongst the temperature induced sRNA-producing loci, several can be selected whose sRNA profiles can be used to distinguish HTS libraries generated under low-temperature stress from those generated under high-temperature treatment.

Acknowledgements

This work was supported by FP7 BIOSUPPORT project.

References

Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK (2005): Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell.* **123**(7):1279-91.

Katiyar-Agarwal S, Morgan R, Dahlbeck D, Borsani O, Villegas A Jr, Zhu JK, Staskawicz BJ, Jin H. (2006): A pathogen-inducible endogenous siRNA in plant immunity., *Proc Natl Acad Sci U S A.* **103**(47):18002-7.

Yan Y, Zhang Y, Yang K, Sun Z, Fu Y, Chen X, Fang (2011): Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice. *R. Plant J.* **65**(5):820-8. doi:10.1111/j.1365-313X.2010.04467.x

A reliable pipeline for a transcriptome reference in non-model species

Hicham Benzekri[1], Rocío Bautista[1], Darío Guerrero-Fernández[1], Noé Fernández-Pozo[2], M. Gonzalo Claros[1] ✉

[1]University of Málaga, Spain
[2]The Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, United States

Motivation and objectives

Next-generation sequencing (NGS) platforms can sequence a particular transcriptome in a fast and cost-effective way. However, most bioinformatics tools are focused in model species where a reference sequence is available. But *de novo* transcriptome assemblies occur commonly when working with non-model species.

Methods

Here it is presented a pipeline for obtaining a reliable transcriptome in a plant non-model species, such as pine and sole, using NGS reads. It should be noted that the non-model species selected do not have a homogeneous genome, since individuals sampled were highly heterozygotic from natural populations. That means that single-nucleotide variations in reads can be due to SNPs or sequencing errors, and there is no way to discern both possibilities. The pipeline is outlined in figure 1.

The pipeline starts with the pre-processing software SeqTrimNext, developed by the authors, that extracts the reliable reads and removes from low-quality ends to contaminant fragments. It can work both on short and long reads. The assembling strategy is based on well-know, dedicated software for transcriptomics. Several assemblers were tested (SOAP, Trinity, ABySS, CABOG, Newbler, etc.) and the ones that better behaved were Oases, MIRA3 and EULER-SR. CD-HIT has used for selection of longest contigs derived from short reads. Since there is no reference to compare the reliability of assemblies, several confirmations were included: (1) original reads were mapped on contigs using Bowtie2 in order to discard artefacts; (2) FullLengtherNext (developed by the authors) was used to discard contigs that do not seem to be coding. Finally, all (long) contigs were reconciled using CAP3 (Minimus can also be used) to obtain the final candidate unigenes. FullLengtherNext can also be used to

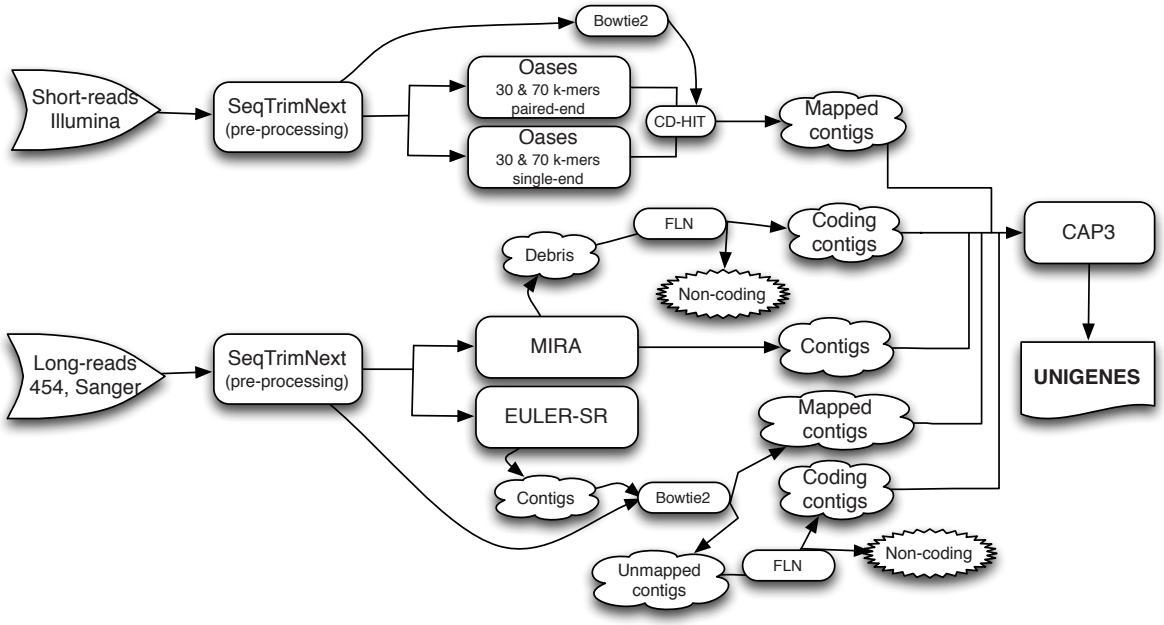


Figure 1. Pipeline for assembling non-model transcriptomes using well-known software as well as new algorithms developed by the authors. It can also handle both long-read and short-reads, including paired-ends.

test which assemblies are the better ones when several strategies are conducted. Unigenes were finally annotated using Sma3Annot (developed in collaboration with O. Trelles) and AutoFact.

Results and discussion

Reference transcriptomes obtained with this approach have been used for printing microarrays (whose hybridisation provided significant and useful results), and perform RNA-Seq analyses that were confirmed by RT-PCR, suggesting that the pipeline is adequate for transcriptome assembling of non-model organisms.

RNA-Seq expression profiling of genes related to neurodegenerative disorders affecting the human retina

Laura Campello, José Martín-Nieto ✉
Universidad de Alicante, Alicante, Spain

Motivation and objectives

Sight is likely the most important human sense. In this context, it is well known that human neurodegenerative diseases, such as Parkinson's disease (PD) and the neuromuscular disorders called dystroglycanopathies (DGPs), cause retinal impairments and consequently vision loss (Muntoni and Voit, 2004; Bodis-Wollner, 2009). We have characterized the expression of PD-related genes *SNCA* (α-synuclein), *PARK 2* (parkin) and *UCHL 1* in the mammalian retina (Martínez-Navarrete *et al.*, 2007; Esteve-Rudd *et al.*, 2010) and have found that a number of DGP-related genes are expressed in this tissue as well (Martín-Nieto *et al.*, 2012). We have also described morphological (Cuenca *et al.*, 2005) and proteomic (Esteve-Rudd *et al.*, 2013) alterations taking place in the primate retina associated with parkinsonism. In this work we have attempted to catalog all known genes linked to PD and DGPs expressed in the human retina and quantify their mRNA levels. We have also focused in identifying transcript variants of these genes, in order to possibly correlate them with propensity to visual impairment.

Methods

Human retina reference RNA extracted from a pool of 29 Caucasian donors (both sexes, ages 20-60) was obtained from Clontech-BD. Total RNA was reverse-transcribed and amplified using the SMART PCR cDNA Synthesis kit (Clontech-BD). The obtained cDNA was mechanically cut into 100 bp fragments by ultrasonication, and a cDNA library was constructed using NEBNext reagents (New England Biolabs). There after, the cDNA was sequenced on an Illumina HiSeq 2000 system by Otogenetics Corp. using a read length of 100 bp, paired-end sequencing and a depth coverage of 100 million reads. Subsequent bioinformatic analyses of the obtained sequenceswere performed by Otogenetics and Genometra companies. The data processing protocol included the following computational tools:

- Sequence data quality control: FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- Sequence data files handling: Samtools (http://samtools.sourceforge.net/)
- Mapping: TopHat software (http://tophat.cbcb.umd.edu/), including the ultra high-throughput short read aligner Bowtie (http://bowtie.cbcb.umd.edu).
- Transcript identification: Cufflinks (http://cufflinks.cbcb.umd.edu/).
- Expression level quantification: Cufflinks software and Qualimap platform (García-Alcalde *et al.*, 2012; http://qualimap.bioinfo.cipf.es/).
- Sequence data alignment visualization: Integrative Genome Viewer (IGV) (www.broadinstitute.org/igv/v1.4).

Results and Discussion

We have evidenced that most of the neurodegenerative disease-related genes assessed are expressed in the human retina, and their mRNA expression levels have been quantitated in terms of fragments per kilobase per million reads (FPKM) through RNA-Seq technology. These include the PD-linked genes *SNCA*, *PARK2*, *UCHL1*, *DJ1* and *PINK1*, and the DGP-linked genes *POMT1*, POMT2, *POMGNT1*, *FKTN* (fukutin), *FKRP* and *LARGE*, among others. Besides, we have characterized the expression profile of such genes in the retina by determining their exonic, intronic and exon-intron junction expression levels. These data have allowed us to examine the alternative splicing pattern of particular genes, and as a result a number of new transcript variants have been identified. We are currently attempting to correlate particular splice variants with loss of gene function. We believe that this research should be of potential usefulness to understand the molecular bases of sight deficiencies associated with neurodegenerative disorders.

Acknowledgements

This research has been supported by the Instituto de Salud Carlos III grant ref. PI09/1623 (to J.M.-N.). L.C. was the recipient of a predoctoral contract from the Universidad de Alicante.

References

Bodis-Wollner I (2009) Retinopathy in Parkinson disease. *J Neural Transm* **116**(11), 1493–1501. doi: 10.1007/s00702-009-0292-z.

Cuenca N, Herrero M-T, Angulo A, De Juan E, Martínez-Navarrete GC et al.(2005) Morphological impairments in retinal neurons of the scotopic visual pathway in a monkey model of Parkinson's disease. *J Comp Neurol* **493**(2), 261–273.doi: 10.1002/cne.20761.

Esteve-Rudd J, Campello L, Herrero M-T, Cuenca N, Martín-Nieto J (2010) Expression in the mammalian retina of parkin and UCH-L1, two components of the ubiquitin-proteasome system. *Brain Res* **1352**,70-82. doi: 10.1016/j.brainres.2010.07.019.

Esteve-Rudd J, Campello L, Bru-Martínez R, Fernández-Villalba E, Herrero MT, Cuenca N, Martín-Nieto J (2013) Alterations in energy metabolism, neuroprotection and visual signal transduction in the retina of parkinsonian, MPTP-treated, monkeys. *PLoS One* (submitted)

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A (2012) Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**(20),2678-2679. doi: 10.1093/bioinformatics/bts503

Martín-Nieto J, Uribe ML, Arenas CM, Rubio M, Aza M, Cruces J, Campello L (2012) All dystroglycanopathies-causing genes are expressed in the retina of adult mammals. *FEBS J* **279**(Suppl. 1), 311

Martínez-Navarrete GC, Martín-Nieto J, Esteve-Rudd J, Angulo A, Cuenca N (2007) α–Synuclein gene expression profile in the retina of vertebrates. *Mol Vis* **13**, 949-961.

Muntoni F, Voit T (2004) The congenital muscular dystrophies in 2004: a century of exciting progress. *Neuromusc Disord* **14**(10), 635-649.

Next-masigpro: dealing with RNA-SEQ time series

Ana Conesa[1], María José Nueda[2] ✉
[1]Prince Felipe Research Centre, Valencia, Spain
[2] University, Alicante, Spain

Motivation and Objectives

During the last decades the development of specific statistics methods to deal with microarray data has been key in transcriptome study. Most of the developed statistics methods have become as reference or classic methods due to their ability to deal with transcriptomics data. However, recent advances in sequencing technologies have created alternatives to microarrays. These new type of data require an appropriate statistical treatment to get good results. New methods are needed but it is also important the study of the adequateness of the existing methods and the adaptation of them to the new type of data.

maSigPro (Conesa *et al.,* 2006) is a method to deal with time course microarray (TCM) data that has been applied in several biological scenarios. maSigPro is in Bionconductor since 2005 and it is implemented in several web-services (Nueda *et al*., 2010 and Medina *et al,.* 2010). However, maSigPro has been designed to deal with normal microarray intensity signals, rather than with count data. In this work, we adapt maSigPro to RNA-Seq time series analysis.

Methods

maSigPro deals with regression linear models where the response is considered as normally distributed data, a continuous variable. Sequencing technologies give us counts data which distribution is discrete. Therefore, applying the original version of maSigPro to discrete data can not be appropriate and results can be wrong.

The statistical model for counts data may be Poisson or Binomial. However, there are studies (Lu *et al.,* 2005) that show overdispersion of the data and suggest the negative binomial (NB) distribution for being more flexible to estimate the variance of the data. Generalized linear models (GLMs) are an extension of linear models to non-normally distributed response data (McCullagh and Nelder 1989, Dobson 2002). We have modified maSigPro package replacing linear models functions by GLMs functions and giving them the appropriate statistical treatment.

To study the need of adapting the maSigPro package to RNA-Seq data several binomial negative time series datasets have been simulated in different scenarios with different number of replicates in each experimental condition (example in table 1). Linear regression models and GLMs have

Table 1: 4 RNA-Seq simulated datasets with 6000 genes, 300 differentially expressed genes, 6 time-points and different number of replicates in each one. FP: false positives. FN: false negatives. R2: model good of fit threshold for gene selection..

		LM MODEL					GLM MODEL				
replicates	**R2**	**selection**	**FP**	**FN**	**Sensitivity**	**Specificity**	**selection**	**FP**	**FN**	**Sensitivity**	**Specificity**
1	0.5	0	0	300	0.000	1.000	663	454	91	0.697	0.920
	0.6	0	0	300	0.000	1.000	657	453	96	0.680	0.921
	0.7	0	0	300	0.000	1.000	601	523	122	0.260	0.926
2	0.5	11	0	289	0.037	1.000	420	144	24	0.920	0.975
	0.6	11	0	289	0.037	1.000	330	75	45	0.850	0.996
	0.7	11	0	289	0.037	1.000	214	20	106	0.647	0.995
3	0.5	217	11	94	0.687	0.998	325	29	4	0.987	0.999
	0.6	171	5	134	0.553	0.999	273	6	33	0.890	1.000
	0.7	81	1	220	0.267	1.000	198	1	103	0.657	1.000
5	0.5	238	0	62	0.793	1.000	299	0	1	0.997	1.000
	0.6	120	0	180	0.400	1.000	280	0	20	0.933	1.000
	0.7	32	0	268	0.107	1.000	174	0	126	0.580	1.000

been applied to the simulated datasets to compare the results.

Results and Discussion

Results show an improved performance of maSigPro to deal with RNA-Seq when using generalized linear models. Therefore the maSigPro package has been updated to include RNA-Seq compatible statistical model. This new version is available in Bioconductor 2.12. The package main structure, analysis steps and visualization options are maintained, hence current maSigPro users can upgrade seamlessly to RNA-Seq time series analysis.

References

Conesa A, Nueda MJ, Ferrer A and Talón M (2006) maSigPro: a Method to Identify Significantly Differential Expression Profiles in Time-Course Microarray Experiments. *Bioinformatics*, **22**(9), 1096-1102. doi: 10.1093/bioinformatics/btl056

Dobson AJ (2002) An introduction to generalized linear models. Chapman & Hall/CRC, Boca Ratón, Florida, 2nd edition.

Lu J, Tomfohr JK and Kepler TB (2005) Identifying diferential expression in multiple SAGE libraries: an overdispersed log-linear model approach. BMC Bioinformatics, 6, 165.

McCullagh P and Nelder JA (1989) Generalized linear models. Chapman & Hall/CRC, Boca Ratón, Florida, 2nd edition.

Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, et al. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling", (2010) Nucleic Acids Research (38) Web Server Issue, W210-213. doi: 10.1093/nar/gkq388

Nueda MJ, Carbonel J, Medina I, Dopazo J and Conesa A (2010) Serial Expression Analysis: a web tool for the analysis of serial gene expression data. Nucleic Acids Research (38) Web Server Issue, 239-245. doi:10.1093/nar/gkq488

Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data

Martin Dahlö ✉
SciLifeLab, Stockholm, Sweden

Motivation and Objectives

As sequencing get cheaper more and more researchers turn to this technology for answers to their questions. The large amounts of generated data will have to be stored somewhere, and the tools to analyse it will have to be updated constantly. UPPNEX tries to solve these problems through high performance computing, large scale and high availability storage, an extensive bioinformatics software suite, up-to-date reference genomes and annotations and a support function with systems and application experts. The software maintained on the computers are mostly the popular open source applications used by the international bioinformatics community, and comprises functions such as alignment, de novo assembly, SNP calling, methylation and RNA-seq analysis.

There are over 300 separate projects at UPPNEX which all belong to different research projects, such as the sequencing of the flycatcher and Norwegian spruce.

The focus of the poster is on the technical and organizational implementations at UPPNEX to handle its task.

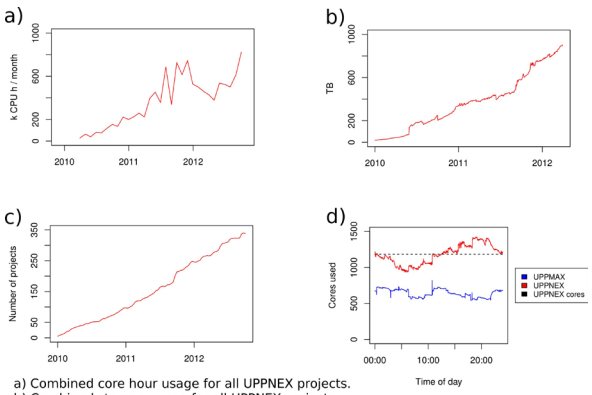
Methods

UPPNEX consists of half a cluster and storage system, and the other half is owned by Uppsala University's high performance computing (HPC) centre, UPPMAX. This sharing of resources has made it possible to compare the typical system usage by bioinformatics projects and the other more general HPC users. The attributes we have been measuring the last 2 years are storage, core hour usage and the size of the booked resources. We have summarized and plotted these attributes and compared the bioinformatics projects to the other general HPC projects (see Figure 1). We have also kept record over how many processor cores each job has used and compared this with non-bioinformatics projects.

Results and Discussion

The number of bioinformatics projects at UPPNEX, and their individual storage usage is increasing everyday and the trend shows no signs of decreasing. The amount of core hours used by UPPNEX is not increasing as much as the storage. Looking at the core hour usage during an average day, it is clear that UPPNEX users are clearly more prone to run interactive sessions and shorter jobs than the other HPC users.

The decision to pool our resources with the existing HPC centre in Uppsala (UPPMAX) gave us a running start since the systems experts already had a lot of the infrastructure in place.



a) Combined core hour usage for all UPPNEX projects.
b) Combined storage usage for all UPPNEX projects.
c) Number of active UPPNEX projects.
d) The number of CPU cores used during an average day. Comparing UPPNEX projects with other general HPC projects.

Figure 1. a) Combined core hour usage for all UPPNEX projects. b) Combined storage usage for all UPPNEX projects. c) The number of active UPPNEX projects. d) The number of cores used during an average day. Comparing UPPNEX projects (red) with other general HPC projects (blue).

Acknowledgements

The UPPNEX project is funded by a KAW grant, and maintenance is carried out by the system experts at UPPMAX at Uppsala University.

Improving automated de-novo transcriptome definition in non-model organisms by integrating manually defined gene information

Ester Feldmesser, Shilo Rosenwasser, Assaf Vardi, Shifra Ben-Dor✉
Weizmann Institute of Science, Rehovot, Israel

Motivation and Objectives

Non-model organisms are of great ecological and economic significance, consequently the understanding of their unique metabolic pathways by investigating their gene expression profiles is essential. The bloom-forming alga *Emiliania huxleyi* is a cosmopolitan unicellular photoautotroph that plays a prominent role in the marine carbon cycle. Its intricate calcite coccoliths account for a third of the total marine CaCO3 production, making it highly susceptible to future ocean acidification.

The advent of next generation sequencing (NGS) technologies and corresponding bioinformatics analysis tools has allowed the definition of transcriptomes in non-model organisms and automated transcriptome assemblies have become common over time. Several methods that integrate de-novo assembly together with genome based assembly have been proposed for non-model organisms (Martin and Wang, 2011). Yet, there are many open challenges in defining genes, where genomes are not available or incomplete. The available genome assembly of *E. huxleyi* is a draft and was constructed from 454 reads in one round of assembly. A large number of available unassembled reads, numerous repeats and duplications, as well as holes in the genome, indicated that the genome alone would not provide a good basis for building transcripts.

In spite of the high numbers of transcriptome assemblies that have been performed, quality control of the transcript building process is rarely performed, if ever. To test and improve the quality of the automated transcriptome definition, we used 63 manually defined and curated genes, several of them experimentally validated. After each step in the automated definition pipeline, the presence of the manually defined genes was checked, allowing troubleshooting of missing genes and improving our pipeline. To the best of our knowledge, this is the first time that an automated transcript definition is subjected to quality control using manually defined and curated genes and thereafter the process is improved.

Methods

Three different approaches were applied in parallel to the automated definition of *E. huxleyi* transcripts, two of them utilizing the read data. The first was de-novo assembly using CLC Assembly Cell (http://www.clcbio.com/products/clc-assembly-cell/) and then CAP3 (Huang and Madan, 1999) to remove redundancy. The second was a genome-based alignment, in which the reads of each sample were aligned separately to the genome using TopHat (Trapnell *et al*, 2009). After the alignment, Cufflinks and Cuffcompare (Trapnell *et al*, 2010) were applied to all the TopHat outputs to define transcripts. In the third approach, available *E. huxleyi* ESTs were clustered using TGICL (http://compbio.dfci.harvard.edu/tgi/software/).

In parallel, genes were manually defined. Protein sequences of the target genes from human, Arabidopsis and yeast, were compared to the *E. huxleyi* genome on the JGI genome website. Hits were inspected to see if any transcript or EST evidence was available. If there were ESTs available, they were assembled into transcripts, and compared to the predictions available. When (NGS) reads became available, they were used to correct and improve the gene definition. If more than one hit were retrieved, each successive hit was also checked to see if it was truly an independent hit, representing a family member, or a duplication, which was then classified as real or artificial. If no ESTs were available as an anchor for a predicted transcript, then a combination of reads (if available), prediction based on blast hits and the JGI predictions were used to construct a transcript. If there was no genomic hit, searches were performed against *E. huxleyi* ESTs, in order to identify sequences that might not have been mapped to the genome. Transcripts were then constructed and extended as far as possible by running successive blasts.

The improvements added to the automatic pipeline after troubleshooting missing manually defined genes were: (1) The Partek (http://www.partek.com) software was applied to find regions to which reads were aligned, but where

Cuffcompare transcripts were not defined, (2) Artificially fused transcripts were split using in-house developed PERL scripts and (3) At the end, two different clustering algorithms were applied to the collection of potential transcripts, TGICL that strongly removed redundancy but loses genes and CAP3 that does not lose genes, but leaves redundancy in the collection. The two approaches were integrated.

Results and Discussion

The final transcriptome collection included 75092 transcripts. The transcript lengths ranged from 301 to 34193 base pairs (bp), 34680 of the transcripts have a length of more than 1000 bp and 23993 are between 500 and 1000 bp. Open reading frames (ORFs) covered the entire transcript in 44% of the transcripts and in approximately 70% of the transcripts, the ORFs covered more than 80%. A high percentage of the reads (80%) were successfully mapped to the transcript collection.

The inter-play between the automated pipeline and the quality control using manually defined genes indicated which additional processes were required to improve the transcriptome definition. In the first assessment of the transcriptome quality, presence of the 63 genes in the three transcript definition approaches was examined. Four of the genes had no coverage in the RNA-Seq, and were not expected to be found in the read-based arms of the assembly. In the genome based transcript collection, of the 59 possible genes, eleven genes were missed. In the de novo assembly, twelve genes were missed.

E. huxleyi has a very high percentage of non-canonical splice junctions, and relatively high rates of intron read-through, which caused unique issues with the currently available tools. While individual tools missed genes and artificially joined overlapping transcripts, combining the results of several tools improved the completeness and quality considerably. The final collection, created from the integration of several quality control and improvement rounds, was compared to the manually defined set both on the DNA and protein level. 61 transcripts and 47 proteins were found, an improvement of 20% versus any of the read-based approaches alone.

References

Huang, X, and Madan A. (1999) CAP3: A DNA sequence assembly program. *Genome Res* **9**(9), 868-877.

Martin, J A and Wang Z. (2011) Next-generation transcriptome assembly. *Nat Rev Genet* **12**(10): 671-682. doi:10.1038/nrg3068

Trapnell, C., L. Pachter, and S. L. Salzberg. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C, Williams BA, Pertea G, Mortazavi A, Kwan G et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5): 511-515.doi: 10.1038/nbt.1621

Understanding biology of potato- virus PVY interaction

Kristina Gruden
National Institute of Biology, Ljubljana, Slovenia

Motivation and Objectives

In nature plants encounter various factors, which influence their growth and development and consequently affect plant product quantity and quality. Potato virus Y (PVY) is a severe plant pathogen responsible for yearly losses in production of *Solanaceae* crops worldwide. Plant responses to viruses and the disease development are different and much less explored in comparison to bacterial or fungal infections. In single component studies the complexity of the plant – pathogen interaction at molecular level can lead to limited conclusions that may fail to notice important changes in physiological processes. Omics approaches, that offer a more holistic view of the processes, are therefore a major step forward in understand these interactions. In addition, we are aiming at identifying novel components of plant defence like noncoding RNAs, potato endo- and epiphytes as well as estimate sequence variability on the pathogen site.

Methods

In our studies, gene expression in the disease response of the susceptible, tolerant and resistant potato (*Solanum tuberosum L.*) cultivars to PVY infection was investigated at different times after infection, using transcriptomics approaches.

Most of our studies were performed combining microarrays and real-time PCR for transcriptomics. Recently, we have complemented our results with the RNA-seq analysis on Solid platform. In addition, we have performed analysis of small RNAs libraries both on Solid as well as on Illumina platform.

Results and Discussion

With 'standard' transcriptomic approaches we have shown that not only the components involved but also the timing and intensity of response are extremely important for the outcome of plant-virus interaction. Small RNA analysis using Solid platform identified significant differences of several miRNAs after viral infection that were so far not implied in plant defence against viruses. Results of other NGS related experiments are still in progress.

Retroviral diversity of laboratory and wild mice M. musculus domesticus

Stefanie Hartmann[1], Jens Mayer[2], Camila Mazzoni[3], Alex D Greenwood[4]
[1]University of Potsdam, Potsdam, Germany
[2]University of Saarland, Saarbrücken, Germany
[3]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany
[4]Leibniz-Institute for Zoo and Wildlife Research, Berlin, Germany

Motivation and Objectives

The sensitivity and cost-effectiveness of Next Generation Sequencing (NGS) technologies has transformed almost every biological discipline, allowing to address questions whose answers seemed out of reach just a few years ago. One such area is the inventory and analysis of endogenous retroviruses of non-human origin. PCR-amplification and low-throughput Sanger-sequencing generally detects only retroviral variants that predominate in a given species or population, and rare sequence variants cannot easily be detected. NGS technologies, in contrast, allow to survey diversity and distribution of retroviruses among individuals, populations and species. Our poster will describe an analysis of targeted murine retroviral sequences from five mouse samples.

Methods

Regions of approximately 400 bp that are conserved in most MLVs and in all known XMRV, PreXMRV-1, and PreXMRV-2 sequences were amplified. Amplicon clones were sequenced using the GS FLX Titanium platform, generating approximately 162,000 reads; thousands for each retroviral region. Although efficient algorithms for computing large multiple sequence alignments and phylogenetic trees exist, the visualization and interpretation of such large trees is technically and conceptually difficult. Furthermore, the fine-scale resolution of relationships within and between clades that phylogenetic trees provide may often not even be necessary. Instead of studying the retroviral diversity using standard phylogenetic analyses, we employed a clustering approach: The sequence reads, together with reference sequences, were clustered using the Markov Clustering algorithm as implemented in Tribe-MCL (Enright, 2002). The resulting clusters were then used to inventory retroviral sequences in the mouse samples. For sequences of selected clusters we also computed Maximum Likelihood phylogenies using RAxML (Stamatakis, 2006).

Results and Discussion

For questions that are generally answered in the context of a phylogeny, massive amounts of sequence data often are a curse as much as a blessing. We will discuss clustering as an alternative to phylogenetic inference for the analysis and visualization of NGS technology-generated sequence data. Specifically, we characterized and will describe the distribution of mouse gamma retroviruses Xmv (xenotropic), Pmv (polytropic), and Mpmv (modified polytropic) in the three inbred laboratory mouse strains and two wild-caught mice. We also examined the distribution of Xenotropic Murine Leukemia Virus-related virus (XMRV), which is a laboratory recombinant of the two precursors PreXMRV-1 and PreXMRV-2 that have been reported to have very different distributions in mice. In addition, phylogenetic trees were computed for clusters containing XMRV and/or PreXMRV sequences.

References

Enright AJ, Van Dongen S, Ouzounis CA. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**(7), 1575-1584.

Stamatakis A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21), 2688-90.

A better sequence-read generator program for metagenomics

Stephen Eric Johnson, Brett Trost, Jeffrey R Long, Anthony Kusalik✉
University of Saskatchewan, Saskatoon, Canada

Motivation and Objectives

There are many programs available for generating simulated metagenomic sequence reads. The data generated by these programs follow rigid models, which limits the use of a given program to the author's original intentions. For example, many popular simulator programs only generate reads that follow uniform or normal distributions. To our knowledge, there are no programs that allow a user to generate simulated data following non-parametric read-length distributions and quality profiles based on empirical next-generation sequencing (NGS) data.

We present BEAR (Better Emulation for Artificial Reads), a program that uses a machine learning approach to generate reads with lengths and quality values mimicking empirically derived distributions. BEAR is able to emulate reads from various NGS platforms, including Illumina, 454 and Ion Torrent. BEAR requires minimal user input, as it automatically determines appropriate internal parameter settings.

Methods

Multiple popular sequence simulator programs were tested to gauge their ability to emulate real data available to our lab. The tested programs were SimSeq (Earl *et al.*, 2011), MetaSim (Richter *et al.,* 2008), Grinder (Angly *et al.,* 2012), and 454sim (Lysholm *et al.*, 2011). Shortcomings were identified and used to guide the development of our improved sequence simulator.

BEAR was written using a combination of Perl and Python scripts. An advantage of BEAR is that it requires only three files as input:

1. the training set, a multi-FASTA file that exhibits the desired read-length and quality distributions;
2. organism database, a multi-FASTA file containing the genomes from which reads will be generated. Each genome is a single sequence;
3. a tab-delimited file containing a sequence identifier for each sequence in the database and the desired relative abundance of that sequence.

BEAR uses a two-step process: in the first step, the organism database and abundance file are used to generate a simulated metagenomic dataset containing reads of uniform quality and length. In the second step, a model of the distribution of read lengths is generated from the training data, and a Markov chain is created based on the quality scores in the training data. The reads from step 1 are then trimmed using a Monte Carlo process based on the read length distribution and have quality scores generated using the Markov chain. The quality score of the current position and the average quality of the five previous positions determine the quality score of the next position.

To evaluate BEAR, actual read data from various sequencer technologies were obtained. Each of the aforementioned programs was used to generate simulated data for these technologies, using appropriate parameter settings. In the case of BEAR, the real data was used as input. (Not all programs could generate all types of simulated reads.) Since metagenomic data was the goal, a list of species and abundances as outlined by Pignatelli and Moya (2011) was used. The characteristics of simulated reads from each program were then determined and compared to the characteristics from the real data.

BEAR is available from the authors upon request. It requires that the user have BioPerl and BioPython installed.

Results and Discussion

As shown in Figure 1a, modern sequencing simulators are limited in their ability to model actual read lengths, being restricted to uniform or normal distributions. In addition, Figure 1b demonstrates the inflexibility of quality score generation within these programs. In contrast, BEAR is better able to mimic the read length and quality distribution characteristics of reads from next-generation sequencing technologies. For example, Figure 1a shows that BEAR more accurately emulates the read distribution of the Ion Torrent data when compared to Grinder, the program that was second closest to matching the read-length

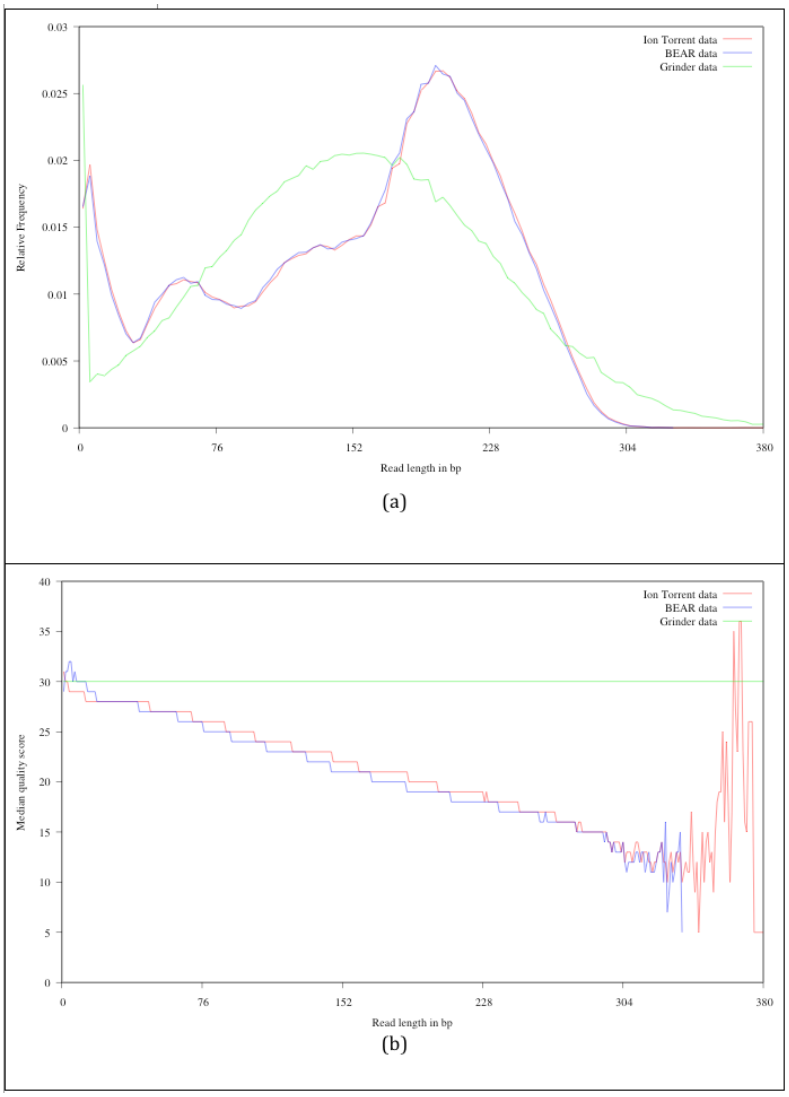


Figure 1. Top: The read length distribution of real Ion Torrent data (red) is compared to the trained data from BEAR (blue) and the second-most accurate program Grinder (green). Bottom: The median quality scores for real Ion Torrent data (red) are compared to the trained data from BEAR (blue) and Grinder (green). As evident in the figure, the longest read generated by BEAR is 338 bp, while the longest read in the Ion Torrent data is 380bp. This is due to the fact that reads longer than 338bp comprise less than 0.0005% of all Ion Torrent reads in our training data.

distribution of the Ion Torrent data. A plot of the median quality scores for the Ion Torrent, BEAR, and Grinder data (Figure 1b) suggests that BEAR generates reads that better emulate the quality profile of real data. Similar results were observed for 454 and Illumina data, suggesting that BEAR is a versatile tool for emulating various NGS platforms.

Acknowledgements

This work was supported by MAVEN, a project funded by Western Economic Diversitification Canada and Enterprise Saskatchewan. Oversight of the MAVEN project is provided by Genome Prairie and Dr. Reno Pontarollo.

References

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson, GW. (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research* **40**(12), e94-e94. doi:10.1093/nar/gks251.

Earl D, Bradnam K, St. John J, Darling A, Lin D, et al. (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* **21**(12), 2224-41. doi:10.1101/gr.126599.111.

Lysholm F, Anderson B, Persson B. (2011) An efficient simulator of 454 data using configurable statistical models. *BMC Research Notes* **4**, 449. doi:10.1186/1765-0500-4-449.

Pignatelli M, Moya A. (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One* **6**(5), e19984. doi:10.1371/journal.pone.0019984.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH. (2008) MetaSim--A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE* **3**(10), e3373.

De novo assembly and annotation of the grey reindeer lichen (Cladonia rangiferina) transcriptome

Sini Junttila, Stephen Rudd ✉
Turku Centre for Biotechnology, University of Turku, Turku, Finland

Motivation and Objectives

Lichens are symbiotic organisms that have a remarkable ability to survive in some of the most extreme terrestrial climates on earth. Lichens can endure frequent desiccation and wetting cycles and are able to survive in a dehydrated state for decades at a time. Genetic resources have been established in lichen species for their taxonomic classification, but no lichen species have been characterised yet using genomics, and the molecular mechanisms underlying the lichen symbiosis and the fundamentals of desiccation tolerance remain undescribed. Research on lichen gene expression is very limited, and as yet there is little in the way of either high-throughput genome sequence or expressed sequence tag (EST) data available for any lichen species. For non-model organisms, de novo genome assembly of short read data is complicated but many transcriptomes have been sequenced from non-model species and published over the last years (Alagna *et al.*, 2009, Novaes *et al.,* 2008, Vera et al. 2008). The annotation process remains challenging, especially for species with no close relatives with a sequenced reference genome.

Our objective was to de novo assemble and annotate a lichen transcriptome using both next-generation sequencing and traditional Sanger sequencing. We produced additional Sanger EST sequences from axenically grown symbiotic partners (*C. rangiferina* and *Asterochloris* sp.) to train classification models for predicting the genome of origin of lichen sequences. We have obtained a basic view of the ongoing molecular processes and have identified the most active biological pathways in *C. rangiferina* (Junttila and Rudd, 2012). Our transcriptome data brings an increase to the amount of publicly available lichen sequences and provides a starting point for further studies into lichen functional transcriptomics.

Methods

cDNA from lichen tissue was sequenced with Roche GS FLX platform, which was chosen for its longer read length compared to other next-generation sequencing platforms, and Sanger sequencing with ABI PRISM 3130xl Genetic Analyzer capillary DNA sequencer was performed to complement the FLX run data with its long sequence reads. EST sequences from cDNA libraries prepared from the axenically grown symbiotic partners, the fungus and the alga, were obtained using Sanger sequencing.

The lichen sequences were de novo assembled with CLC Genomics Workbench software version 4.9 (CLCBio, Denmark). Prior to the assembly the sequences were trimmed in the CLC Genomics Workbench, and sequences shorter than 15bp were removed from the analysis. Eclat (Friedel *et al.,* 2005) was used to identify the genome of origin for the assembled contigs and singletons derived from lichen tissue. The Sanger sequences obtained from the axenically grown algal and fungal symbiont cDNA libraries were used to train Eclat and build a model file for the classification. The minimum sequence length for classification was set at 100 bp.

Blast2GO (Conesa *et al.,* 2005) tool was used for BLASTX, Gene Ontology (GO) term annotation, Interpro scans and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of the contigs and singletons. BLASTX (Altschul *et al.,* 1990) was used to compare the assembled contigs and singletons to a non-redundant (nr) protein sequence database from the NCBI GenBank database (Benson *et al.* 2008). BLASTX matches were filtered using an arbitrary cut-off of 1e-10.

Results and Discussion

Altogether 243,729 high quality reads were de novo assembled into 16,204 contigs and 49,587 singletons. 62.8% of the sequences were classified as being of fungal origin while the remaining 37.2% were predicted as being of algal origin. In

the annotation 34.4% of the sequences had a BLAST match, 29.3% of the sequences had a GO term match and 27.9% of the sequences had a domain or structural match following an InterPro search. 60 KEGG pathways with more than 10 associated sequences were identified.

Our results present a first transcriptome sequencing and de novo assembly for a lichen species and describe the ongoing molecular processes and the most active pathways in *C. rangiferina.* This brings a meaningful contribution to publicly available lichen sequence information. These data provide a first glimpse into the molecular nature of the lichen symbiosis and characterize the transcriptional space of this remarkable organism. These data will also enable further studies aimed at deciphering the genetic mechanisms behind lichen desiccation tolerance.

Acknowledgements

The authors would like to acknowledge Janne Isojärvi, Andras Kiraly, Asta Laiho and Attila Gyenesei. The work was funded by the Academy of Finland grant (project number FI-2960501) to Stephen Rudd.

References

Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, et al. (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* **10**:399. doi: 10.1186/1471-2164-10-399

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *J Mol Biol* **215**(3):403–410.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. *Nucleic Acids Res* **36**:D25–30

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18):3674–3676.

Friedel CC, Jahn KH, Sommer S, Rudd S, Mewes HW, et al. (2005) Support vector machines for separation of mixed plant-pathogen EST collections based on codon usage. *Bioinformatics* **21**(8):1383–1388.

Junttila S and Rudd S (2012) Characterization of a transcriptome from a non-model organism, Cladonia rangiferina, the grey reindeer lichen, using high-throughput next generation sequencing and EST sequence data. *BMC Genomics* **13**:575. doi: 10.1186/1471-2164-13-575.

Novaes E, Drost D, Farmerie W, Pappas GJ, Grattapaglia D et al. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics* **9**:312. doi: 10.1186/1471-2164-9-312

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* **17**(7):1636–1647. doi: 10.1111/j.1365-294X.2008.03666.x

Biologist-friendly analysis software for NGS data

Aleksi Kallio[1], Taavi Hupponen[1], Massimiliano Gentile[2], Jarno Tuimala[3], Kimmo Mattila[1], Ari-Matti Saren[1], Petri Klemelä[1], Ilari Scheinin[4], Eija Korpelainen[1] ✉
[1]CSC – IT Center for Science, Helsinki, Finland
[2]Blueprint Genetics Oy, Helsinki
[3]SPR Veripalvelu, Helsinki
[4]VU University Medical Center, Netherlands

Motivation and Objectives

NGS technology offers unprecedented possibilities for life science, motivating efforts to develop new data analysis tools and techniques. However, available tools are scattered and often require some programming skills, leaving them out of reach of non-computational researchers. Chipster provides a clear and biologist-friendly interface to analysis tools for NGS data. It has a graphical user interface that connects to server environment for heavy data processing. Chipster is a free and open source software, and it is available as a virtual machine for easy server installation.

Methods

Chipster (http://chipster.csc.fi) provides data analysis tools for many NGS applications, including DNA-, RNA-, miRNA-, ChIP-, methyl- and CNA-seq. Users can easily save and share analysis workflows, and built-in genome browser allows seamless viewing of reads and results.

Users can perform their whole data analysis in Chipster from quality control to downstream applications such as pathway enrichment and motif discovery. Popular tools such as FastQC, FASTX, PRINSEQ, SAMtools, BEDTools, Bowtie, BWA, TopHat, HTSeq and Cufflinks are included, and care has been taken to serve them in a biologist-friendly manner. Also several R/Bioconductor packages have been integrated, including edgeR, DESeq and MEDIPS.

Chipster's built-in genome browser allows visualization of reads and results in their genomic context using Ensembl annotations. Users can zoom in to nucleotide level, highlight SNPs and view the automatically calculated coverage. Cross-talk between the genome browser and BED, VCF and GTF 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 (Kallio *et al.*, 2011). Recently system's data handling capabilities have been enhanced to cope with NGS scale data. The system is capable of tracking data copies across the distributed system (both server and client side), minimizing data transfers and always using the closest copy for best performance. We are also working with the Hadoop MapReduce framework so that large jobs can be run in a massively parallel way (Niemenmaa *et al.*, 2012).

To maintain good user experience with large scale NGS datasets, Chipster server allows users to save their analysis sessions on the server side. For the administrators it also provides tools to monitor disk space usage on the server environment.

New analysis tools can easily be added using a simple mark-up language. Chipster places no restrictions on what type tools can be integrated. Currently we are introducing intuitive graphical interfaces also for tool development and server administration.

Results and Discussion

Taken together, Chipster provides an easy way to serve NGS data analysis tools in a biologist-friendly manner. In our national role of providing bioinformatics services for whole country we have noticed that a user-friendly software together with training on data analysis methods is a powerful combination for enabling life scientists to analyze their own data.

The complete Chipster server system is freely available under the open source GPL license at http://chipster.sourceforge.net, and it has been adopted by many institutes worldwide. The recommended option is to download virtual machine images that contain all tools and databases, bundled together with a ready-to-run Chipster installation. Virtual machine images support all major virtualization platforms.

Acknowledgements

This work was supported by the Intelligent Monitoring of Health and Well-being program of SalWe – the Strategic Centre for Science, Technology and Innovation in Health and Well-being, and by the Cloud Software Program of the Finnish Strategic Centre for Science, Technology and Innovation TIVIT.

References

Kallio, Tuimala, Hupponen *et al.* (2011) Chipster: User-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12**, 507.

Niemenmaa, Kallio, Schumacher *et al.* (2012) Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud. *Bioinformatics* **28**(6), 876.

Metagenomics sample preparation and sequencing

Oskar Erik Karlsson[1,2,3], Martin Norling[2], Erik Bongcam-Rudloff[2] ✉
[1]Department of Biomedical Sciences and Veterinary Public Health (BVF), Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[2]SLU Global Bioinformatics Centre, Department of Animal Breeding and Genetics (HGEN), Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[3]The Joint Research and Development Division of SLU and SVA, OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine (OIE CC), Uppsala, Sweden

Motivation and Objectives

Metagenomic methods for detection of viruses provide new diagnostic tools to veterinary and public health laboratories, with powerful capacities to detect and to monitor the viromes in clinical samples. The Metagenomics methodology is divided into three main activities or steps: (1) wet-lab methodology; (2) sequencing; and (3) data analysis. Integrating all three parts is of critical importance to the results as well as their interpretation. Our groups at the OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine, Uppsala, Sweden and at the SLU Global Bioinformatics Centre, Uppsala, Sweden are working with the development and evaluation of the methodological and technological platforms for viral metagenomics. Together with the National Veterinary Institute (SVA), we develop and test methods for extraction of viromes, feasibility of sequencing platforms to deliver metagenomic data-sets and evaluate bioinformatics tools as well as combine them into software packages for analysis and exploration of metagenomes, for separation, classification, assembly and visualization of genomic data in metagenomic samples. The aim of the work is to provide insight into using the metagenomics approach for detection of emerging viruses, monitoring wildlife for known pathogens as well as providing a tool for rapid characterization of viral pathogens in outbreak situations.

The main goal of the work is now to define Standard Operating Procedures for metagenomic investigations of clinical material, integrate current research within environmental ecology for experimental design in a clinical setting, continuously evaluate sequencing technologies (given the rapid turnover of sequencing platforms and chemistry) and integrating the methodology into the diagnostic labs by a risk based model for initiation of metagenomic investigations.

Methods

Samples are processed by whole genome nucleic acid extraction. Targeting the virome fraction of the sample by DNase/RNase treatment, degrading most of the host genome and microbiome genomes present in the original sample. Depending on sequencing technology, see next paragraph, sample might need amplification treatment or nucleic acid concentration. In the case of nucleic acid amplification the included bias must be considered in the final result (Belak *et al.*, 2013).

Sequencing strategies for metagenome retrieval range from clonal amplification combined with Sanger sequencing, to direct sequencing using the Illumina HiSeq system. Depending on chosen strategy, bias might be introduced into the final result. Besides the known bias introduced by the sequencing platforms, the platform can introduce indirect bias due to constraints on input material. By requiring high or low concentrations of input material while handling a mixed or genome depleted sample, bias is introduced by pre-sequencing processing of samples. The most obvious such bias is the whole genome amplification bias, inherent to both Multiple Displacement Amplification (MDA) and Sequence-independent, single-primer amplification (SISPA). Currently the Unknown Virus discovery platform at the OIE CC in Uppsala has successfully deployed our methodology into several sequencing systems; 454 Life Science/Roche, Illumina (HiSeq as well as MiSeq), IonTorrent. As NGS applications for metagenomics is now moving into its first decade the choice of sequencing technology seems to be fairly open, with good experimental design and biologically relevant questions taking precedence for good results.

Results and Discussion

The group has a fairly extensive history of both method adaption/development and applica-

tion in veterinary virology, providing an unbiased toolset for detection of emerging viruses and screening for known families of viruses (Granberg *et al.*, 2013; (Blomstrom *et al.*, 2009; Blomstrom *et al.*, 2010)

Currently large scale screening, prevalence studies as well as studies in viral hotspot zones are performed using the methodology. Integrating the methodology into the framework of bio-preparedness in outbreak situations is also a main goal within the scope of the AniBioThreat program.

With increasing availability of sequencing equipment the field of metagenomisc is only just entering its golden era. Given the possibility of an unbiased methodology for characterization of the microbiome several labs will provide great insight over the coming few years.

Acknowledgements

This work was supported by the Award of Excellence (Excellensbidrag) provided to SB by the Swedish University of Agricultural Sciences (SLU).

The authors would also like to acknowledge support of Uppsala Genome Centre and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at Uppsala Genome Centre has been funded by RFI/VR "SNISS" Swedish National Infrastructure for large Scale Sequencing and Science for Life Laboratory, Uppsala.

Writing of this publication has been supported by the framework of the EU-project AniBioThreat (Grant Agreement: Home/2009/ISEC/AG/191) with the financial support from the Prevention of and Fight against Crime Programme of the European Union, European Commission – Directorate General Home Affairs. This publication reflects views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

Support by BILS (Bioinformatics Infrastructure for Life Sciences) is gratefully acknowledged.

References

Belak, S., O. E. Karlsson, A. L. Blomstrom, M. Berg and F. Granberg (2013). "New viruses in veterinary medicine, detected by metagenomic approaches." *Vet Microbiol*.

Blomstrom, A. L., S. Belak, C. Fossum, J. McKillen, G. Allan, P. Wallgren and M. Berg (2009). "Detection of a novel porcine boca-like virus in the background of porcine circovirus type 2 induced postweaning multisystemic wasting syndrome." *Virus Res* **146**(1-2): 125-129.

Blomstrom, A. L., F. Widen, A. S. Hammer, S. Belak and M. Berg (2010). "Detection of a novel astrovirus in brain tissue of mink suffering from shaking mink syndrome by use of viral metagenomics." *J Clin Microbiol* **48**(12): 4392-4396.

Granberg G, Vicente-Rubiano M, Rubio-Guerri C, Karlsson OE, Kukielka D, Belák S, Sánchez-Vizcaíno JM. (2013) Metagenomic Detection of Viral Pathogens in Spanish Honeybees: Co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses. PLOS One. Feb 27. doi: 10.1371/journal.pone.005759

DGW: an exploratory data analysis tool for clustering and visualisation of epigenomic marks

Saulius Lukauskas[1], Gabriele Schweikert[2], Guido Sanguinetti[3]✉
[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[2]School of Informatics and Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh, United Kingdom
[3]School of Informatics & SynthSys, Synthetic and Systems Biology, University of Edinburgh, Edinburgh, United Kingdom

Motivation and Objectives

Novel sequencing based technologies such as ChIP-Seq and DNAse-Seq (reviewed e.g. in Furey 2012) are revolutionizing our understanding of chromatin structure and function, yielding deep insights in the importance of epigenomic marks in the basic processes of life. The emergent picture is that gene expression is controlled by a complex interplay of protein binding and epigenomic modification, leading to a hypothesis of a major regulatory role for the histone code of each gene (Wang *et al.,* 2008). While histone marks (and other epigenomic marks) can be measured in a high throughput way, exploratory data analysis techniques for these data types are still largely lacking. Epigenomic marks exhibit characteristics that distinguish them fundamentally from e.g. mRNA gene expression measurements: they are spatially extended across regions as wide as several kilobases, and often present interesting local structures, such as the presence of multiple peaks and troughs. These patterns often have a biological origin, such as the displacement of a nucleosome or the length of the first exon of a gene (Bieberstein *et al.,* 2012), so that analysis tools that take into account these spatial features would be desirable. However, each (combination of) epigenomic mark(s) at different locations in principle represents a multivariate data point *of different length* (as peaks for the same mark in different locations can have widely differing lengths): this prevents the straightforward extension of well established data analysis techniques such as hierarchical clustering to these data types. In this work, we present Dynamic Genome Warping (DGW), an open source clustering tool for epigenomic marks which addresses this problem by introducing a *local rescaling* which allows to match (multiple) epigenomic marks based on maximum similarity between shapes. DGW is based on Dynamic Time Warping, a well-established technique in signal processing and speech recognition. Our tool handles simultaneously multiple epigenomic marks and is freely available as a Python stand-alone tool. It consists of a worker module, which distributes the computationally intensive parts across multiple processes automatically (thus using all available CPU cores), and an explorer module, which allows easy and adaptive inspection of the data set.

Methods

The basic algorithm underlying DGW is the classical dynamic time warping algorithm (Sakoe and Chiba, 1978). This is a dynamic programming algorithm closely related to the classical sequence alignment algorithms. Specifically, given two sequences $\mathbf{a}=(a1,\ldots,aN)$ and $\mathbf{b}=(b1,\ldots,bM)$, and a local distance between the elements of each sequence (e.g. Euclidean distance or Cosine distance), it constructs a *warping path*, i.e. a sequence of points in the two sequences that are mapped to each other. The warping path has the property of minimising the sum of the distances between the aligned points; furthermore, it is monotonic (i.e. there are no inversions in each sequence) and maps the first and last point of sequence **a** to the first and last point of sequence **b**. The warping path also computes a warping distance between the two sequences (intuitively, how much one sequence has to be stretched to match the other). In order to avoid large stretches of a sequence being mapped to a single point of the other sequence, we implement the constrained approach suggested in (Sakoe and Chiba, 1978). A modern review of the basic concepts can be found in e.g. (Muller 2007).

DGW takes as input a series of genomic regions (as a bed file outputted by a peak finder, or as a set of predefined regions, e.g. defined windows around transcription start sites) and a number of bam files for different epigenomic marks. Peaks are discretised in bins of 50 bp width. The DGW worker module then computes the warp-

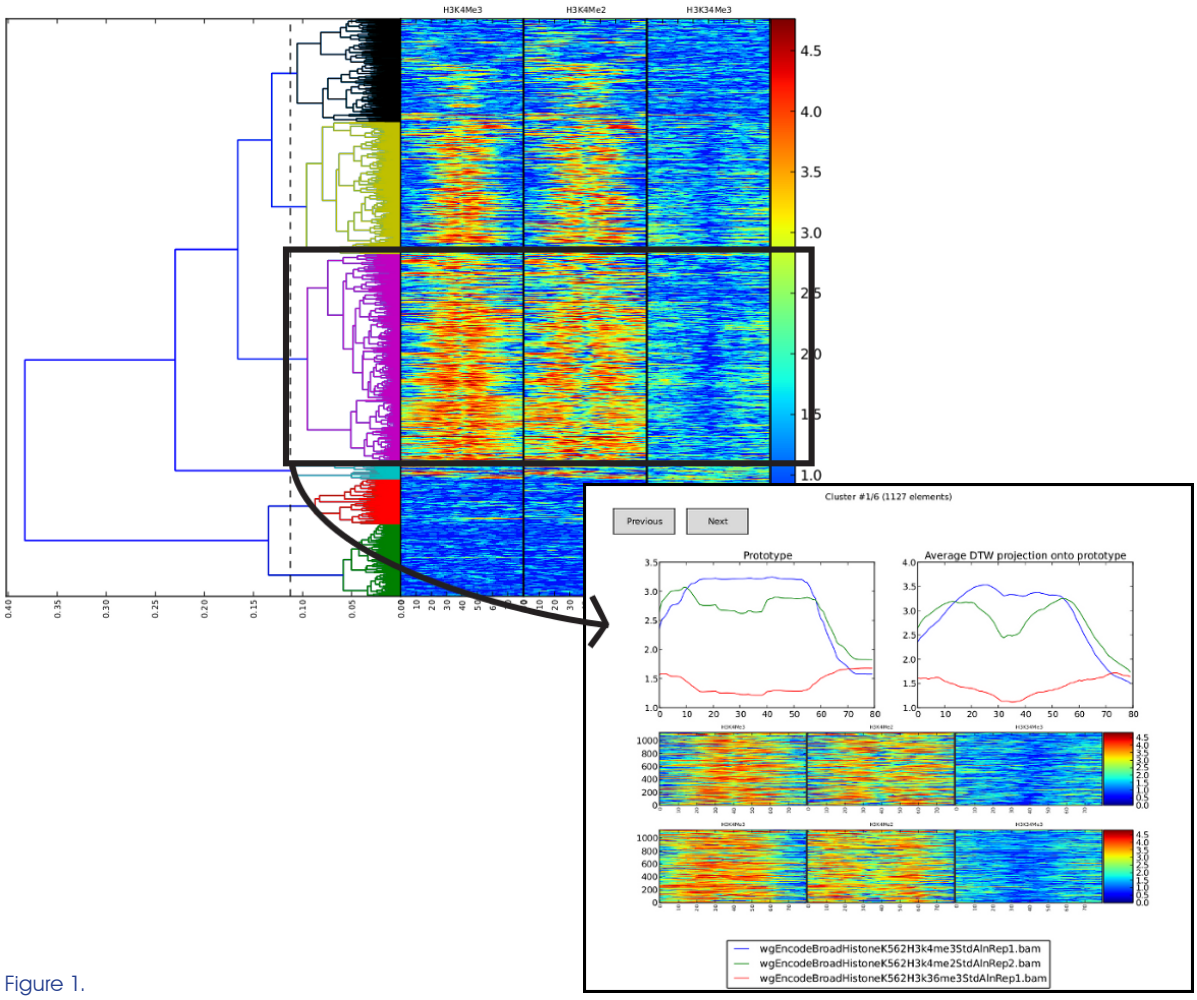


Figure 1.

ing distance between each pair of peaks (to account for antisense promoters, the distance is in fact computed also with peaks with direction reversed). The worker module then computes and outputs a dendrogram, exactly as in hierarchical clustering.

Results are then displayed by the DGW explorer module. This contains several customisable features which facilitate data inspection. The DGW dendrogram is displayed horizontally, with a movable vertical line that allows easy selection of the desired cut distance threshold, so that visually evident clusters can be selected. The module then computes a prototype for each histone mark in each cluster; the prototype can then be displayed in a new window alongside the average of each histone mark across the cluster, and heatmaps of the original and warped data.

Results and Discussion

We stress tested the tool by applying it to randomly simulated sequences generated from five fixed different seed sequences; here, as expected, the dendrogram returned five well defined clusters. We then applied the tool to real data; Figure 1 shows example results obtained applying DGW to a ChIP-Seq data set of histone modifications from the ENCODE project (http://encodeproject.org/ENCODE). The marks selected are H3K4me3, H3K4me2 and H3K36me3 in the leukaemia cell line K562 (accession code wgEncodeBroadHistoneK562). The background panel shows the dendrogram outputted by the DGW worker module; the vertical dashed line can be moved horizontally by the user to select the number of clusters. The foreground panel shows the cluster analysis window, which is opened upon double clicking on a cluster. The output of the program

can then be easily used with downstream tools to perform further biological analysis. The code is in the process of being released to GitHub (www.github.org) as an open source Python package.

The results clearly show that DGW provides a practical and user friendly tool for exploratory data analysis of high throughput epigenomic data sets, much like classical hierarchical clustering is for microarray time series. While evaluation of results is clearly an important step still to be performed, we believe the availability of exploratory data analysis tools will play an important role in generating hypotheses and eventually clarify the role of epigenetics in fundamental biology.

Acknowledgements

G.Schw. acknowledges support from EC through the FP7- Marie Curie project "Epigenome Informatics". G.S. acknowledges support from the European Research Council through grant MLCS306999.

References

Bieberstein N, Carrillo Oesterreich, F, Straube, K and Neugebauer, K (2012). "First exon length controls active chromatin signatures and transcription". Cell, **2** (1), 62-68.

Furey TS (2012) "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions." Nat Rev Genet. 13:840-52.

Muller M, Dynamic Time Warping, Springer, 2007.

Sakoe H and Chiba S (1978). "Dynamic programming algorithm optimisation for spoken word recognition" IEEE Trans. On Speech, Acoustics and Signal Processing 26(1), 43-49.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome.". Nat Genet **40** (7): 897–903.

Ion Torrent sequencing and pipeline assembly of the first genome sequence of a mesophilic syntrophic acetate oxidizing bacterium (SAOB)

Shahid Manzoor[1], Erik Bongcam-Rudloff[1], Anna Schnürer[2], Bettina Müller[2] ✉
[1]Department of Animal Breeding and Genetics Science, Swedish University of Agricultural Science, SLU Global Bioinformatics Centre, Uppsala, Sweden
[2]Department of Microbiology, Swedish University of Agricultural Sciences, Uppsala BioCenter, Uppsala, Sweden

Motivation and Objectives

Syntrophic acetate oxidising bacteria (SAOB) have been identified as key organisms for efficient biogas production from protein rich materials under moderate conditions. *Tepidanaerobacter acetatoxydans* strain Re1 is the first reported mesophilic SAOB which genome has been sequenced (Manzoor *et al.,* 2013). Growth experiments and genetic studies allocated *T. acetatoxydans* to the physiological group of homoacetogens producing acetate through the Wood-Ljungdahl pathway when growing heterotrophically (Manzoor *et al.,* 2013, Westernholm *et al.,* 2011). However, when growing in syntrophy with methanogens, SAOB reverse this pathway and oxidise acetate to hydrogen and carbon dioxide (Müller *et al.,* 2012, Hattori *et al.,* 2005, Schnürer *et al.,* 1997).

Motivation and Objectives

Bioinformatics analysis might aid us to define general features being essential for maintaining a syntrophic lifestyle and to answer question concerning regulation, energy conservation and electron transfer mechanisms. This knowledge will enable us to further understand the mechanisms triggering SAO in different environments, to monitor the activities of known SAOB and to find new isolation strategies.

Methods

The genome was sequenced by Ion Torrent PGM™ Systems and we assembled the genome by using a comparative pipelined approach by using MIRA3 for mapping and Newbler 2.8 for *de novo* assembly. The finished genome was annotated with the annotation tools provided by MaGe (Microbial Genome Annotation & Analysis Platform).

Results and Discussion

Our comparative approach with mapping and *de novo* assembly contributed successfully for finishing the bacterial genomes with high accuracy and less PCR work. After complete genome assembly and annotation (Table 1), a first genome analysis confirmed the two *fhs* clusters identified by Müller *et al.* (Müller *et al.,* 2012) as well as the suggested operon structure including all genes encoding the Wood-Ljungdahl pathway with one exception: No format dehydrogenase has been identified underlying the observed inability of this organism to establish an autotrophic lifestyle (Müller *et al.,* 2012). ATP seems to be generated by substrate level phosphorylation exclusively because no complete F0F1-ATP synthase has been found. Further two V-type ATPase operons, a cluster encoding an Rnf complex and several hydrogenases clusters have been identified and might contribute to the energy conserving systems of the cell.

Table 1. Genomic features of SAOB Genome.

Strain	T. size (bp)	T. Protein coding genes	Avg. CDS Lgth.	Avg. Interg. Lgth.	Total rRNAs	Total tRNAs	Protein Coding Density	G+C %
TepRe1	2,759,867	2,524	917	136	6	52	86.92	38

References

Hattori S, Galushko AS, Kamagata Y, Schink B. (2005) Operation of the CO dehydrogenase/acetyl coenzyme A pathway in both acetate oxidation and acetate formation by the syntrophically acetate-oxidizing bacterium *Thermacetogenium phaeum*. *J. Bacteriol*. **187:** 3471-3476.

Manzoor S, Bongcam-Rudloff E, Schnürer A, Müller B. (2013) *Tepidanaerobacter acetatoxydans* strain Re1: The first genome sequence of a syntrophic acetate oxidizing bacteria. *Genome Announc*. **1:** e00213-12.

Müller B, Sun L, Schnürer A. (2012) First insights into the syntrophic acetate oxidising bacteria (SAOB) – a genetic study. *Microbiology Open* (accepted).

Schnürer A, Svensson BH, Schink B. (1997). Enzyme activities in and energetics of acetate metabolism by the mesophilic syntrophically acetate-oxidizing anaerobe *Clostridium ultunense*. *FEMS Microbiol letters* **154**: 331-336.

Westerholm M, Roos S, Schnürer A. (2011): *Tepidanaero-bacter acetatoxydans* sp. nov., an anae-robic, syntrophic acetate-oxidizing bacterium isolated from two ammonium-enriched mesophilic methanogenic processes. *Syst. Appl. Microbiol.* **34:** 260-266.

Comparison of variant calling methods in exome sequencing of matched tumor-normal sample pairs

Sara Monzon[1,2], Javier Alonso[2], Gonzalo Gómez[3], David Gonzalez-Pisano[3], Isabel Cuesta[1] ✉
[1]Bioinformatic Unit, National Centre of Microbiology, Instituto de Salud Carlos III (ISCIII), Majadahonda, Madrid, Spain
[2]Chilhood Solid Tumor Unit, Instituto de Investigación de Enfermedades Raras, ISCIII, Majadahonda, Madrid, Spain
[3]Bioinformatic Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

Motivation and Objectives

Next Generation Sequencing techniques are allowing the determination of new somatic mutations involved in cancer development. One of the Bioinformatics' challenges is to develop computational tools able to distinguish in a reliable way the germline polymorphisms present in healthy tissue from the somatically acquired mutations in tumor cells. There has been described two families of somatic variant calling approaches, in earlier one somatic variants have been detected by independently genotyping both samples and subtracting the results (i.e Samtools, Unified Genotyper), in contrast to new one which make simultaneous analysis of tumor and normal datasets from the same individual (i.e. Strelka, JointSNVMix). In our knowledge, just a few reliable studies compare their results. The aim of this work is to compare these two different somatic variant calling approaches analyzing sequenced exome of tumor-matched normal sample pairs.

Methods

A new workflow that allows comparison of different variant calling methods (Samtools v. 0,1,16 (Li *et al.*, 2009), Unified Genotyper v. 1,6 (DePristo *et al.,* 2011), Strelka v. 0,4,6 (Saunders *et al.,* 2012) and JointSNVMix v. 0,8 (Roth *et al.,* 2012)) has been developed. This workflow has been tested using two exome paired-end matched tumor-normal data sets obtained from pediatric cancers: dataset A (Illumina GAIIx, two patients – two tumors) and dataset B (Illumina HiSeq, 11 patients – 13 tumors). The threshold used for Samtools was selected by default, for Unified Genotyper was the recommended configuration in the GATK best practice guidelines, for Strelka was the recommended configuration with a quality score ≥15, and for JointSNVMix was P(somatic) ≥ 0,8. Somatic variants reported by all methods were manually curated using IGV (Integrative Genomics Viewer; Robinson *et al.,* 2011).

Results and Discussion

The results obtained analyzing Dataset A are showed in Table 1. The simultaneous analysis of tumor–normal paired sequence used by Strelka or JointSNVMix, gives a lower false positive variant number than independent analysis of the tumor and normal data, approach followed by software like Samtools and Unified Genotyper, both commonly used in variant calling workflows.

Preliminary results prove that the determination of somatic mutations in tumors requires that the specific algorithms are able to analyze, in a combined way, the information provided by tumor DNA and constitutional DNA, and thus enabling better precise distinction between germinal and somatic variants.

Table 1. Number of somatic variants validated by manual curation with IGV. FP: False positive, TP: True Positive.

Methods	Variant number	FP	TP
Samtools - pileup	71	63	8
Unified Genotyper	14	13	1
Strelka	13	0	13
JointSNVMix	7	5	2

Acknowledgements

We acknowledge ASION (Chilhood Cancer Association of The Community of Madrid), special Program "Hucha de Tomás", to support this work. Chilhood Cancer Genome grant TVP 1278/21.

References

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078-9. doi: 10.1093/bioinformatics/btp352

DePristo M, Banks E, Poplin R, Garimella K, Maguire J et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data**.** *Nature Genetics.* **43**: 491-498. doi:10.1038/ng.806.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES et al. Integrative Genomics Viewer (2011). *Nature Biotechnology* **29**, 24–26. doi:10.1093/bib/bbs017

Roth A, Ding J, Morin R, Crisan A, Ha G, et al., (2012). JointSNVMix: A Probabilistic Model For Accurate Detection Of Somatic Mutations In Normal/Tumour Paired Next Generation Sequencing Data. *Bioinformatics* **28** (7), 907–913. doi:10.1093/bioinformatics/bts053

Saunders CT, Wong W, Swam S, Becq J et al. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28** (14), 1811-1817. doi:10.1093/bioinformatics/bts271

PIPELINER: a tool to evaluate NGS pipelines and optimize experimental designs for resequencing studies

Bruno Nevado[1], Miguel Perez-Enciso[1,2,3]✉
[1]Centre for Research in Agricultural Genomics (CRAG), Barcelona, Spain
[2]Universitat Autononoma de Barcelona, Bellaterra, Spain
[3]Institut Catala de Recerca i Estudis Avancats (ICREA), Barcelona, Spain

Motivation and Objectives

The choice of technology and bioinformatics' approach is central in the analysis of Next-Generation-Sequencing (NGS) experiments. The pace with which new software and methodological guidelines are published, together with the fact that many of these choices will depend on the particularities of the study-system, mean researchers are often unable to produce informed decisions regarding these central questions. To address these issues, we introduce Pipeliner, a tool to simulate and validate the performance of NGS analysis pipelines, and optimize experimental designs.

Methods

Pipeliner is written in Object-Oriented Perl and is highly customizable, allowing the user to write and test his own bioinformatics' pipelines. A simulation is then performed for each pipeline defined, and statistics describing their performance in variant calling are calculated and reported.

The first step in the analysis performed with Pipeliner is to specify the experimental design, which includes defining the study system, i.e. number of individuals sequenced, population structure, depth and sequencing technology. Pipeliner uses coalescent simulations (with the software ms; Hudson, 2002) to obtain the "true" Single-Nucleotide Polymorphism (SNP) data for the population under study, allowing for specific conditions to be explored such as the effect of the distance between the sampled individuals and the reference genome available, the effect of population subdivision, or different levels of variability or selection. As for the NGS reads, Pipeliner uses the program ART (Huang *et al.,* 2011) to simulate illumina NGS reads (solid and 454 reads can also be simulated with ART), with the user defining the read length, the average depth per individual, paired or single ends run, etc.

Once the experimental design is defined, the next step is to choose the bioinformatics' pipeline with which to analyze the genetic data obtained. The three crucial decisions in this step are (i) how to align the short reads to the reference genome, (ii) how to call variants from the aligned short reads, and (iii) how to filter the variants obtained. Pipeliner implements a wrapper to the commonly used software bwa (Li and Durbin, 2009), and to samtools (Li *et al.,* 2009), which is the default SNP calling tool in Pipeliner. However, Pipeliner also implements a simple interface that allows using any other software for these tasks.

In the final step, Pipeliner calculates a number of statistics that summarize the performance of the defined bioinformatics' pipeline and provide plots to make interpretation easy. The statistics calculated include Recovery (% of SNPs correctly identified in relation to the original SNP number obtained with the coalescent simulations), Power (% of SNPs correctly identified in relation to all SNPs that pass the quality and depth filters set by the user), False Discovery Rate (FDR, % of SNP calls that are incorrect) as well as the frequency with which different errors occur.

Results and Discussion

As an example of the type of analysis possible in Pipeliner, we investigate the effect of individual SNP calling vs. multiple individual SNP calling. Results (Figure 1) show that, while the overall Power increases when joint SNP calling, singletons for the alternative allele actually become more elusive (lower Power), while other SNP sites become easier to detect. This situation is likely to lead to skewed allele frequency spectrum calculations, however such detailed bias has not, to our knowledge, been reported before.

Qualitatively similar results were obtained with higher coverage per individual (12x).

The choice of experimental design and bioinformatics' pipelines are central issues in

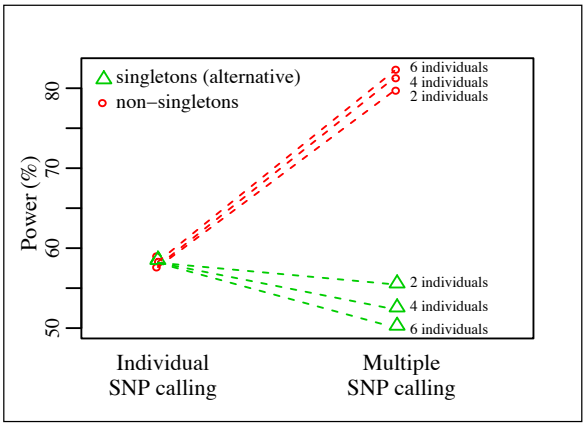


Figure 1. The effect of multiple-sample SNP calling.
Power in identifying heterozygous SNPs when doing individual or multiple SNP calling with samtools, under low coverage (average 6x per diploid individual) and few individuals (2, 4 and 6).

the analysis of NGS datasets. With Pipeliner, we provide the tools that empower researchers to carefully plan their study's sampling design, and compare the suitability of alternative software for their specific study systems. Pipeliner can be obtained from its website: https://github.com/brunonevado/pipeliner.

Acknowledgements

The authors would like to thank S. Ramos-Onsins, W. Sanseverino and R. Tonda for helpful discussions and comments. This work was supported by the Spanish Ministerio de Ciencia e Innovacion [AGL2010-14822 to M.P.E.]; and the Center for Research in Agricultural Genomics [consolider project CSD2007-00036].

References

Huang W, Li L, Myers JR, Marth GT (2011) ART: a next-generation sequencing read simulator. *Bioinformatics* **28** (4): 593-594. doi: 10.1093/bioinformatics/btr708

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18** (2): 337–338. doi: 10.1093/bioinformatics/18.2.337

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25** (14): 1754-1760. doi: 10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al.* (2009): The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25** (16): 2078–2079. doi: 10.1093/bioinformatics/btp352

Genome sequence of plant associated rhizobacterium Bacillus amyloliquefaciens strain UCMB5033

Adnan Niazi, Shahid Manzoor, Sarosh Bejai, Johan Meijer, Erik Bongcam-Rudloff ✉
Swedish University of Agricultural Sciences, Uppsala, Sweden

Motivation and Objectives

Over the past years, different measures have been adopted to increase crop production including chemical fertilizers and pesticides, and more recently genetic manipulation of plants. Such methods are costly tools for increasing the yield. In addition, side effects such as nutrient leakage, development of pesticide resistant strains, and negative effects on the environment call for other means to maintain food safety and security. Bacteria mediated bio-control is an alternative strategy that have great promise to overcome such problems. *Bacillus amyloliquefaciens* strain UCMB5033 is a spore forming Gram-positive rhizobacterium that has shown great potential to serve as biocontrol agent. It has shown growth promotion and disease protection against insects and pathogens after developing symbiotic relationship with host plant; *A. thaliana* and oilseed rape *B. napus* (unpublished). Our ultimate goal is to eliminate the use of agrochemicals for production of crops. To achieve this, the aim of the study is to reveal and explain the genetic architecture that contributes to the plants ability to overcome biotic and abiotic stress based on bacterial bio-control. The availability and analysis of genome sequence will throw light on several biological aspects behind plant-bacterium interaction, such as plant colonization, priming, and stress tolerance in order to support durable plant protection and eliminate chemical pesticides.

Methods

The genome of *B. amyloliquefaciens* UCMB5033 was sequenced with Illumina multiplex technology and Ion Torrent PGM systems. Whole genome assembly was accomplished by comparative assembly approach i.e. integrating both mapping and *de novo* assembly. The paired-end reads from Illumina were provided to MIRA v.3.4 (Chevreux *et al.,* 1999) for mapping assembly and reads data from Ion Torrent was assembled with Newbler v.2.8 by *de novo* assembly method (Zerbino and Birney, 2008). Mapping assembly was performed against the available genome of *B. amyloliquefaciens* UCMB5036 (accession no. HF563562) (Manzoor *et al.,* 2013). The contigs produced by Newbler assembler were moved according to the reference genome and aligned to the sequence obtained through mapping assembly, using Mauve genome alignment software (Darling et al., 2010). The genome sequence was annotated with a collection of annotation tools via Magnifying Genome (MaGe) Annotation Platform (Vallenet *et al.,* 2009).

Results and Discussion

The combination of assemblies from reads data generated by two different NGS platforms contributed in speeding up the assembly process with high accuracy that resulted in *B. amyloliquefaciens* strain UCMB5033 assembled genome sequence. The genome sequence confirmed the presence of NRPS and PKS gene clusters: surfactin (*srf*), fengycin (*fen*), difficidin (*dfn*), bacilysin (*bac*), macrolactin (*mln*), bacillaene (*bae*), bacillomycin D (*bmy*), and bacillibactin (*dhb*) responsible for the synthesis of secondary metabolites, including antifungal and antibacterial compounds (Chen *et al.*, 2007). Other genes involved in metabolism of plant derived compounds, resistance to drugs, root colonization and other functions that presumably give the bacterium an advantage in developing symbiotic relationship with plants were also present.

Acknowledgements

This work was supported by the grants from Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) and the Higher Education Commission (HEC), Pakistan. Sequencing was performed by the SNP&SEQ Technology Platform, Science for Life Laboratory at Uppsala University, a national infrastructure supported by the Swedish Research Council (VR-RFI) and the Knut and Alice Wallenberg Foundation. Bioinformatics analysis

were also supported by the BILS infrastructure at SLU.

References

Chen XH, Koumoutsi A, Scholz R, Eisenreich A, Schneider K et al. 2007. Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nature Biotechnol.* **25**:1007–1014. doi: 10.1038/nbt1325

Chevreux B, Wetter T, Suhai S. 1999. Genomic sequence assembly using trace signals and additional sequence information. *Computer Science and Biology*: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**:e11147. doi: 10.1371/journal.pone.0011147

Manzoor S, Niazi A, Bejai S, Meijer J, Bongcam-Rudloff E. 2013. *Bacillus amyloliquefaciens* strain UCMB5036: The genome sequence of a plant associated bacterium. *Genome Announc.* **1**(2): e00111-13. doi: 10.1128/genomeA.00111-13

Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L et al. 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)* 2009:bap021. doi: 10.1093/database/bap021

Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**:821–829. doi:10.1101/gr.074492.107

The bioinformatics of viral metagenomics

Martin Norling[1], Oskar Karlsson[1,2,3], Erik Bongcam-Rudloff[1] ✉
[1]SLU Global Bioinformatics Centre, Department of Animal Breeding and Genetics (HGEN), Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[2]Department of Biomedical Sciences and Veterinary Public Health (BVF), Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[3]The Joint Research and Development Division of SLU and SVA, OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine (OIE CC), Uppsala, Sweden

Motivation and Objectives

Metagenomic methods provide the veterinary and public health sciences with the promise of new and improved diagnostic tools with unprecedented ability to detect a plethora of known and unknown viromes in clinical samples. Successful metagenomics is based on three main activities where each one must be consistent and reliable for the method to be useful, these are (1) wet-lab preparation, (2) sequencing, and (3) bioinformatics analysis of the results.

We are a collaborative group from the OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine, Uppsala, Sweden and the SLU Global Bioinformatics Centre, Uppsala, Sweden who are working with the development and evaluation of platforms and methods for viral metagenomics. Together with the National Veterinary Institute (SVA), we develop and test methods for extraction of viromes, feasibility of sequencing platforms to deliver metagenomic data sets within constraints of money and time as well as evaluate bioinformatics tools to do the final analysis. We also combine the tools that evaluate well into software packages for separation, classification, assembly and visualization of genomic data in metagenomic samples.

The aim of the work is to provide insight into the feasibility of using the metagenomics approach for detection of emerging viruses, monitoring wildlife for known pathogens as well as providing a tool for rapid characterization of viral pathogens in outbreak situations from a veterinary standpoint.

Methods

The bioinformatical challenge separates itself from the preparation and sequencing steps in that methodologies in bioinformatics evolve comparatively fast. A stable wet-lab and sequencing platform will still require constant updates of bioinformatical databases and tools, requiring that any system is build in a modular way where each part can easily be exchanged for an updated variety.

The current bioinformatical metagenomics pipeline is implemented with two separate front-ends. One is a classic command-line interface suitable for server automation and implementation into further pipelines and the other is a combined HTML5 and jQuery interface intended to give the power of the command-line interface to everyday users. This two-fold approach is to allow as many users as possible to use the same tools, making results easier to reproduce in different settings.

The back-end pipeline is based on a simple plug-and-play configuration where any program or script can easily be replaced to customize or update the system. The current default configuration is based on FastQC (unpublished) for quality control, Prinseq-Lite (Schmieder and Edwards 2011), MetaVelvet (Namiki, *et al.,* 2012) and BLAST (Altschul, *et al.,* 1990), but tests are being run with several other programs as well. The system can be configured to use scheduling systems in the back-end to distribute load and processing. The default scheduler is SLURM, but the system could easily be configured to use a different system. This modular approach is evident throughout the entire project – allowing the system to be used in a wide range of situations.

Results and Discussion

The system is still in BETA, and every part of the pipeline and interface is still being tested and evaluated, but a few common themes are sure to live throughout the project.

First of all, the web version of the system uses the high level of user interaction allowed by HTML5 and jQuery. The system allows for upload of large data files by fragmenting and resuming, allowing arbitrary size data sets to be uploaded

without changing server settings, as well as allowing a broken download to resume from where it was cut-off. The web system is built around responsive, intuitive interfaces giving feedback and process information in real time, sane defaults giving new users an easy start with quick guides for common tasks and clear documentation for server implementations. The entire system will also be released as open source to contribute to the public as much as possible.

Acknowledgements

This work is funded by SIDA grant SWE-2011-154: "Using next generation sequencing to support the development of infection and treatment vaccination methods in East Africa", Sweden. The bioinformatics work at SLU Global Bioinformatics Centre was also supported by grants from SLU and BILS.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403-410.

Namiki T, Hachiya H, Tanaka, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* **40**(20), e155. doi:10.1093/nar/gks678

Schmieder R, Edwards R (2011) Quality cotrol and preprocessing of metagenomic datasets. *Bioinformatics*, **27**(6), 863-864. doi:10.1093/bioinformatics/btr026

Modulation of the host cell RNA splicing program by the gastric pathogen helicobacter Pylori

Frithjof Glowinski, Fernando Garcia-Alcalde, Konstantin Okonechnikov, Thomas F Meyer✉
Department of Molecular Biology, Berlin, Germany

Motivation and Objectives

Helicobacter pylori is a Gram-negative bacterial pathogen colonizing the human stomach. Infection with H. pylori causes chronic inflammation of the gastric mucosa and may lead to peptic ulceration and/or gastric cancer. Using a quantitative phosphoproteomic approach several splicing factors were found to be differentially phosphorylated upon infection (Holland *et al.*, 2011).

Serine arginine rich (SR) proteins, in particular, which are involved in the regulation and control of alternative splicing, are affected by the differential phosphorylation. SR proteins are regulated in their subcellular location dependent on their different phosphorylation states. To investigate the functional consequences of such alterations in phosphorylation we analysed the changes in splicing of a small set of known targets of SR-protein dependent alternative splicing. Within this set, two genes, BRCA1 and BMF, were confirmed to be differentially spliced after infection. For a more comprehensive picture of changes in splicing the host cell mRNA was sequenced using next generation sequencing, RNA-seq.

Methods

The mRNA of infected host cells was analyzed using by RNA-seq and further characterized for changes in gene expression and alternative splicing. Sequencing samples were obtained from *in vitro* infections. Human gastric AGS cells were infected with Helicobacter pylori strain P12. Total RNA was collected and the polyA enriched fraction was used for strand-specific library preparation. Sequencing was performed on the Illumina HiSeq2000 platform with 90 bp strand paired-end reads and a fragment size of 300 bp. To estimate variability 3 biological replicates were performed, each with a non-infected vs infected sample. Reads were mapped with TopHat (Trapnell *et al.*, 2009) and defaults parameters. TopHat mapping was supplemented with splice junctions previously identified by PASSion (Zhang *et al.*, 2012), and mapping quality assessed using QualiMap (García-Alcalde *et al.*, 2012). All further analyses were based on this combined TopHat-PASSion mapping. Differential alternative splicing was analyzed using either Cufflinks (Trapnell et al. 2010) or DEXSeq (Anders *et al.*, 2012). Both pipelines were compared to an experimental in-house pipeline. This in-house tool uses statistical testing on the obtained isoform expression levels for detection of differentially spliced transcripts.

Results and Discussion

All six samples were sequenced with approx. 40 Mio paired-end reads. Reads were mapped and further analysed for differential expression and splicing.

Differential exon usage was analysed using the DEXSeq package. 15 genes were found to contain exons, which are significantly changed in their abundance after infection. Differential expression and alternative splicing was further analysed using the Cufflinks pipeline. Using the workflow described in Trapnell *et al.* (2012) 892 genes were found to be differentially expressed between the non-infected and infected cells. Accordingly, 178 genes were identified to be differentially spliced between the two conditions.

Notably both tools identified a considerably different number of genes to be affected by regulation of cellular splicing with just minimal overlap. To further investigate this contradicting results we developed a new statistical approach to detect genes significantly differing in their isoform distribution between conditions.

The results of this new method and a detailed comparison with the tools mentioned above will be presented.

Acknowledgements

This work was supported by the International-Max-Planck-Research-School IMPRS-IDI (FG).

References

Anders, S., Reyes, A. & Huber, W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome research* **22**(10), 2008–2017.García-Alcalde, F. *et al.* (2012) Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**(20), 2678–2679.

Holland, C. *et al.* (2011) Quantitative phosphoproteomics reveals link between Helicobacter pylori infection and RNA splicing modulation in host cells. *Proteomics* **11**(14), 2798–2811.

Trapnell, C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**(3), 562–578.

Trapnell, C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**(5), 511–515.

Trapnell, C., Pachter, L., Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9), 1105–1111.Zhang, Y. *et al.* (2012) PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. *Bioinformatics* **28**(4), 479–486.

REDITOOLS: efficient RNA editing detection by RNA-SEQ data

Ernesto Picardi[1,2], **Graziano Pesole**[1,2]✉
[1]University of Bari, Bari, Italy
[2]Institute of Biomembrane and Bioenergetics of National Research Council, Bari, Italy

Motivation and Objectives

RNA editing and alternative splicing are post-transcriptional modifications that increase the complexity of eukaryotic transcriptomes and proteomes. Both phenomena have been efficiently investigated by massive sequencing according to recent high-throughput technologies. In particular, the RNA-seq methodology is the de facto technology to investigate entire eukaryotic RNA populations at single nucleotide level providing fruitful snapshots of cell/tissue activities in a variety of normal and non-homeostatic conditions (Wang, *et al.*, 2009).

RNA editing can modify specific RNAs at selected locations (Maas, 2011) and, in human, frequently involves the deamination of adenosines to inosines by the family of ADAR enzymes acting on double RNA strands (Hogg, *et al.*, 2011). Inosine is commonly interpreted as guanosine by splicing and translation machineries other than sequencing enzymes. A-to-I RNA editing has a plethora of biological effects, strictly related to the RNA region involved in the modification. Changes in 5' and 3'UTRs, for example, can lead to altered expression, preventing the efficient ribosome binding at 5'UTR or the recognition by small regulatory RNAs at 3'UTR. In contrast, alterations in coding protein regions can induce amino acid replacements with more or less severe functional consequences (Hood and Emeson, 2011).

RNA editing events can be detected at large scale by adopting the RNA-seq technology and, thus, employing multiple read alignments onto the corresponding reference genome to look at A-to-G changes (Eisenberg, *et al.*, 2010). Recently in human, thousands of candidates have been identified and validated by direct comparison with whole genome sequencing data in order to purge single nucleotide variations (SNPs) (Ramaswami, *et al.*, 2012).

Although a variety of methodologies have been developed to explore the RNA editing impact on eukaryotic transcriptomes, no comprehensive software for this aim has been released to date. The main challenge is to implement effective filters to mitigate the detection of false positives due to sequencing errors, mapping errors and SNPs. Very recently, we released the web service ExpEdit to explore known RNA editing events in RNA-seq experiments and the first computational strategy to predict de novo RNA editing events without any a priori knowledge of the genomic information or the nature of RNA editing process (Picardi, *et al.*, 2012). Here we present REDItools, a suite of python scripts aimed to the study of RNA editing at genomic scale.

The package is freely available at Google Code repository (http://code.google.com/p/reditools/)and released under the MIT license.

Methods

REDItools consist of three main python scripts and several accessory scripts. Main scripts are based on Pysam module (http://code.google.com/p/pysam/) a wrapper of SAMtools (Li, *et al.*, 2009) for easy manipulation of big alignment files. Pysam includes methods and functions to handle read alignments in SAM/BAM format facilitating the browsing of multiple read alignments position by position along a reference genome. REDItools enable the analysis of RNA editing at three levels: 1) REDItoolDnaRna.py identifies RNA editing changes by comparing RNA-seq and DNA-seq reads from the same individual; 2) REDItoolKnown.py explores the RNA editing potential of entire RNA-seq experiments using known sites stored in public databases as DARNED or provided by users; 3) REDItoolDenovo.py implements our methodology to detect RNA editing events using RNA-seq data alone.

Several accessory scripts are also provided in order to facilitate the browsing of results and assisting users through the annotation of predicted positions by using widespread databases from UCSC genome browser. Additional annotation and filtering steps are performed using the tabix program, for which the wrapper is included in the Pysam module.

All results are provided in tabulated tables for easy parsing and filtering.

Results and Discussion

REDItools work on machines running unix/linux operating systems and accept BAM files from whatever sequencing technology or organism.

REDItools have been extensively tested on public human RNA-seq experiments. In particular, we used REDItools to explore the impact of RNA editing on RNA-seq reads from the Illumina Human Body Map 2.0 Project using known events annotated in DARNED database (Kiran, *et al.*, 2013). Fastq files for RNA-seq experiments were downloaded from ArrayExpress database and mapped onto the hg18 human genome by GSNAP (Wu and Nacu, 2011) including known splice sites from UCSC, RefSeq, Ensembl and AspicDB (Martelli, *et al.*, 2011). SAM files were converted to BAM by SAMtools, duplicated reads were marked by Picard MarkDuplicates.jar (http://sourceforge.net/projects/picard/) and quality scores were recalibrated by GATK (McKenna, *et al.*, 2010). We added also further positions not yet present in DARNED and available from Ramaswami *et al.* (2012). The complete set of known RNA editing events comprises 564,135 positions.

Filtering output tables in order to focus only on positions supported by at least 10 reads and RNA editing frequency ^{3} 0.1, yielded the following table ordered by the descending number of RNA editing sites per tissue (Table 1). It is quite interesting to note that the number of RNA editing sites is tissue dependent as expected and the brain is not the human tissue with the predominant number of events. Moreover, this naïve experiment directly demonstrates the power and effectiveness of our REDItools in investigating at genomic level the intriguing phenomenon of RNA editing.

Table 1

Accession	OrganismPart	Editing Sites
ERR030874	ovary	6599
ERR030881	adrenal	5603
ERR030872	thyroid	4573
ERR030873	testes	4122
ERR030882	brain	3992
ERR030880	adipose	3877
ERR030883	breast	3739
ERR030877	prostate	3238
ERR030885	kidney	3146
ERR030886	heart	2393
ERR030887	liver	2376
ERR030884	colon	2320
ERR030875	white blood cells	2288
ERR030878	lymph node	1978
ERR030879	lung	1591
ERR030876	skeletal muscle	536

Acknowledgements

This work was supported by the Italian Ministero dell'Istruzione, Università e Ricerca (MIUR): PRIN 2009 and 2010; Consiglio Nazionale delle Ricerche: Flagship Project Epigen, Aging Program 2012-2014 and by the Italian Ministry for Foreign Affairs (Italy-Israel actions).

References

Eisenberg, E., Li, J.B. and Levanon, E.Y. (2010) Sequence based identification of RNA editing sites, *RNA biology*, **7**, 248-252.

Hogg, M., *et al.* (2011) RNA editing by mammalian ADARs, *Advances in genetics*, **73**, 87-120.

Hood, J.L. and Emeson, R.B. (2011) Editing of Neurotransmitter Receptor and Ion Channel RNAs in the Nervous System, *Current topics in microbiology and immunology*. **353**, 61-90

Kiran, A.M., *et al.* (2013) Darned in 2013: inclusion of model organisms and linking with Wikipedia, *Nucleic acids research*, **41**, D258-261.

Li, H., *et al.* (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics (Oxford, England)*, **25**, 2078-2079.

Maas, S. (2011) Gene regulation through RNA editing, *Discovery medicine*, **10**, 379-386.

Martelli, P.L., *et al.* (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing, *Nucleic acids research*, **39**, D80-85.

McKenna, A., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome research*, **20**, 1297-1303.

Picardi, E., *et al.* (2012) A Novel Computational Strategy to Identify A-to-I RNA Editing Sites by RNA-Seq Data: De Novo Detection in Human Spinal Cord Tissue, *PLoS One*, **7**, e44184.

Ramaswami, G., *et al.* (2012) Accurate identification of human Alu and non-Alu RNA editing sites, *Nature methods*, **9**, 579-581.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature reviews*, **10**, 57-63.

Wu, T.D. and Nacu, S. (2011) Fast and SNP-tolerant detection of complex variants and splicing in short reads, *Bioinformatics (Oxford, England)*, **26**, 873-881.

Semi- supervised ensemble learning to boost miRNA target predictions

Gianvito Pio[1], Domenica D'Elia[2], Donato Malerba[1], Michelangelo Ceci[1] ✉
[1]Department of Computer Science - University of Bari, Bari, Italy
[2]CNR, Institute for Biomedical Technologies, Bari, Italy

Motivation and Objectives

The huge amount of data produced by the advent of Next Generation Sequencing (NGS) technologies is providing scientists with an unprecedented potential to investigate and shed light on remote secrets of genomes. In particular, many interesting insights are coming from the growing number of evidences about the regulatory function of the non-coding RNA (ncRNA) component of the genome. Indeed an increasing number of new ncRNAs have been recently discovered, most of them showing a primary role in the regulation of genome expression at different levels (Rossi Paschoal *et al.,* 2012). Some of these ncRNAs, although their existence is known since many years ago, have been only recently characterised for their functional role in important biological processes, in a wide variety of organisms and in human diseases. These findings represent one of the most important outcomes of the recent NGS revolution and it would not be so surprising if still unsuspected functions of this so called "dark matter" of the genome will be discovered in the near future.

Among functional classes of ncRNAs with a role in the regulation of gene expression, microRNAs (miRNAs) are those for which more functional data are available and on which the interest of scientists has been more focused over the last decades. The growing number of evidences of their key role in cancer and recent evidences about their presence in body fluids, such as serum and plasma, have further sparked the interest of the scientific community, emphasizing the possibility of using them as therapeutic targets and noninvasive biomarkers of disease and of therapy response.

We have developed a new tool based on biclustering techniques, i.e. HOCCLUS2 (Pio *et al.,* 2013) which is able to significantly correlate multiple miRNAs and their targets to detect potential miRNA:mRNA regulatory networks. However, experiments performed on predicted interactions led to observe that the noise (i.e., false positives) introduced by prediction algorithms can substantially affect the significance of the discovered modules. In order to overcome this issue, we have developed a probabilistic method which is able to build a more reliable dataset, combining data produced by several well-known miRNA target site prediction algorithms. This tool could greatly help in the interpretation of NGS miRNA profiles analysis with respect to their effects, by using genome-wide predictions of their targets.

Methods

The main goal of this work is to combine the prediction score of several prediction algorithms in a single stronger classifier, in order to improve the reliability of the obtained predictions. We propose a probabilistic approach to identify a probability function which, on the basis of the score returned by several prediction algorithms, estimates the probability of the presence of actual miRNA:mRNA interactions. In particular, it exploits information conveyed by datasets of validated interactions to learn such probability function. The identified function is then applied to large sets of predicted interactions. In this context, classical supervised learning algorithms cannot be directly applied, since: *i)* datasets of experimentally verified interactions provide only positive instances; *ii)* the number of labeled (positive) instances is imbalanced with respect to the number of unlabeled (unknown) instances. In order to overcome the first issue, the proposed approach works in the semi-supervised learning setting (Chapelle *et al.,* 2006) which exploits both information conveyed by (positively) labeled and unlabeled instances in the learning phase. Furthermore, the proposed method resorts to an ensemble learning solution in order to deal with the imbalancing.

The results obtained with the proposed approach, which "learns to combine" the output of several prediction algorithms, can be used to identify regulatory modules. This last task is performed by HOCCLUS2 which *i)* extracts possibly overlapping biclusters, to catch multiple roles of

both miRNAs and their target genes; *ii)* extracts hierarchically organized biclusters, to facilitate bicluster browsing and to distinguish between universe and pathway-specific miRNAs; *iii)* extracts highly cohesive biclusters, to consider only reliable interactions; *iv)* ranks biclusters according to the functional similarities, computed on the basis of Gene Ontology (GO)(Ashburner *et al.,* 2000), to facilitate bicluster analysis.

Results and Discussion

Experimental results obtained using human data from miRTarBase (Hsu *et al.,* 2011), as the set of labeled (positive) interactions, and mirDIP (Shirdel *et al.,* 2011), as the set of unlabeled (unknown) interactions, show a significant improvement in the quality of biclusters with respect to the baseline approach of averaging the scores obtained by selected prediction algorithms. In particular, the application of the proposed approach leads to identify biclusters containing miRNAs and mRNAs that are more functionally related each other, according to the GO classification. Moreover, a deep evaluation of the functional consistency of identified miRNA:mRNA modules, by investigating the current literature and data extracted from different web resources (e.g., DAVID, Reactome, GeneCards tool suite and STRING), shows performances of HOCCLUS2 which are comparable with those obtained when applied on experimentally validated interactions in miRTarBase.

Acknowledgements

This work is partial fulfillment of the research objective of "DM19410 - Laboratorio di Bioinformatica per la Biodiversità Molecolare" and "PON01 _ 02589 - MicroMap project "Caratterizzazione su larga scala del profilo metatrascrittomico e metagenomico di campioni animali in diverse condizioni fisiopatologiche".

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25-29. doi:10.1038/75556

Chapelle O, Schölkopf B, Zien A (2006) Semi-*Supervised Learning. Adaptive computation and machine learning.* MIT Press, Cambridge, Mass., USA.

Hsu SD, Lin FM, Wu WY, Liang C, Huang WC et al. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, **39**, 163-169. doi: 10.1093/nar/gkq1107

Pio G, Ceci M, D'Elia D, Loglisci C, Malerba D (2013) A novel biclustering algorithm for the discovery of meaningful biological correlations between miRNAs and mRNAs. *BMC Bioinformatics*, **14**(Suppl 7):S8. doi:10.1186/1471-2105-14-S7-S8

Rossi Paschoal AR, Maracaja-Countinho V, Setubal JC, Simoes ZLP, Verjovski-Almeida S, Durham AM (2012) Non-coding transcription characterization and annotation. A guide and web resource for non-coding RNA databases. *RNA Biology*, **9**, 274-282. http://dx.doi.org/10.4161/rna.19352

Shirdel EA, Xie W, Mak TW, Jurisica I (2011) NAViGaTing the Micronome - Using Multiple MicroRNA Prediction Databases to Identify Signalling Pathway-Associated MicroRNAs. *PLoS ONE*, **6**(2), e17429. doi:10.1371/journal.pone.0017429

A comprehensive comparison between reference-based and 'de novo' isoform assembly approaches

Oscar Rodriguez, Juan Carlos Triviño, Rebeca Miñambres, Sheila Zuñiga, Sonia Santillán, Mayte Gil, Reyes Claramunt, Celia Buades ✉
Sistemas Genómicos, Paterna, Spain

Motivation and Objectives

RNA-seq has recently become an attractive method of choice in the studies of transcriptomes, promising several advantages compared to microarrays such as higher sensibility and reproducibility. In addition, RNA-seq offers a broader dynamic range of detection and the capability of identifying novel isoforms as well as non-translated regions that may act in regulating gene expression. The reconstruction of the transcriptome can be performed following two different approaches, a reference-based method in which reads are mapped back to a reference genome, and a 'de novo' assembly strategy where reads are compared to each other to reconstruct expressed isoforms without the need of using a reference genome.

In the present studio we provide a comprehensive comparison between these two transcriptome analysis methodologies for isoforms reconstruction based on genome annotation and isoform expression levels using a Human sample. In addition, our work provides new insights into Human isoform diversity and the composition of non-canonical isoforms.

Methods

Total RNA was extracted from a Hapmap cell line culture. Strand-specific fragment libraries were built for Illumina HiSeq2000 sequencing using a paired-end strategy. A total of 30Gbs of raw data were produced for the sample.

Following standard reference-based approaches for RNA-seq data analysis, high quality reads were mapped with Tophat (Trapnell *et al.,* 2009) against the Human reference genome GRhg37/hg19. Gene expression levels were estimated using FPKM values as given by Cufflinks (Trapnell *et al.,* 2010) and DESeq (Anders and Huber, 2010).

In the 'de novo' transcriptome reconstruction approach, two algorithms, Trinity (Grabherr *et al.,* 2011) and Oases (Schulz *et al.,* 2012), were used. Resulting isoform assemblies were merged with CAP3 (Huang and Madan, 1999) to obtain a final consensus assembly. Isoform annotation and chimera detection were based on the Human annotations available at Ensembl (http://www.ensembl.org/).

Results and Discussion

Our results showed a high correlation between the reference-based approach and the 'de novo' assembly strategy in terms of the number of detected/reconstructed isoforms and their global expression. However, both methodologies showed specific differences suggesting higher susceptibility to different technical parameters and biases depending on sequencing depth, sequencing errors and the presence of complex or large variants.

Acknowledgements

This work has been financed by the 7th Framework Program.

References

Anders, S., Huber W. (2010): Differential expression analysis for sequence count data. *Genome Biology* **11**, R106+. doi:10.1186/gb-2010-11-10-r106.

Grabherr, M. G. et al. (2011): Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnology 29, 644-652. doi:10.1038/nbt.1883.

Huang, X., Madan, A. CAP3 (1999): A DNA sequence assembly program. *Genome Research* **9**, 868-877. doi:10.1101/gr.9.9.868

Trapnell, C., Pachter, L., Salzberg, S. L. TopHat (2009): discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C. *et al.* (2010): Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-515. doi: 10.1038/nbt.1621.

Schulz, M. H., Zerbino, D. R., Vingron, M., Birney, E. Oases (2012): robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092. doi:10.1093/bioinformatics/bts094.

Generation of expression calls for RNA-seq data

Marta Rosikiewicz[1,2], Marc Robinson-Rechavi[1,2]✉
[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Motivation and Objectives

The Bgee database (database for Gene Expression Evolution; Bastian *et al.,* 2008) provides information about genes that are expressed in different organs and tissues. In order to introduce RNA-seq results into Bgee we had to develop methodology for deriving expressed/unexpressed calls for genes. Such detection calls can be used for characterization of the tissue gene expression profile. Additionally detection calls are widely used in transcriptomic studies for filtering the genes used for differential expression analysis, clustering samples or building more reliable classifiers (Archer and Reese, 2010). The goal of our work is to find an automatic way to define the cut-off value on a transcription level that allows discrimination between expressed and non-expressed genomic features for each library individually.

Methods

RNA-seq data preprocessing

Reads from RNA-seq libraries from experiment GSE30352 (Brawand *et al.,* 2011) were mapped to gene models from Ensembl database and to selected intergenic regions of the reference genome. The mapping of the reads was performed using TopHat2 (Trapnell *et al.,* 2009). The maximum number of mapping sites allowed for a read was set to 1. The intergenic regions are chosen in such a way that the distribution of their lengths matches the distribution of lengths of the transcriptome. Reads that map to the features are summed up using the HTSeq-count software (http://www-huber.embl.de/users/anders/HTSeq/). The RPK (read per kilobase) value for every feature is obtained by dividing the number of reads that match a given feature by its length.

The present/absent calls

Our approach to define present/absent calls is based on Hebenstreit *et al.,* 2011. For each RNA-seq library independently, we define a RPK cut-off, k, for determining "present/absent" calls, set to be equal to the minimal value for which the ratio of relative abundance of intergenic regions and genes, with RPK values above k, is equal or lower than α (in Bgee, $\alpha = 0.05$). In other words, a RPK threshold is defined for each sample independently, such that a randomly chosen feature, from the set of genes and intergenic regions, with a RPK value above the threshold, has at least 95% probability of being a gene.

Cut-off determination procedure

1) For every value of x define the ratio r: $r = n_{ix}N_g / n_{gx}N_i$ where:

n_{ix}: number of intergenic regions with RPK values higher than x
n_{gx}: number of genes with RPK values higher than x
N_i: number of all intergenic regions
N_g: number of all genes

2) The cutoff value k is the minimal value of x for which r is equal or lower than α.

Results and Discussion

The procedure described for expression calls generation was applied to all samples from the analyzed dataset. In general we decided to use selected random intergenic fragments to estimate transcription level coming from experimental noise or background activity of the transcription machinery. Despite up to 4 times differences in the number of aligned reads between libraries the proportion of genes called expressed by our algorithm remained consistent among different samples reaching in case of mouse data 39.1% $\pm$ 1.49 SD and human 34.4% $\pm$ 3.9 SD (56.4% $\pm$ 2.22 and 71.59% $\pm$ 5.28 for protein coding genes respectively). In contrast only 3.4% $\pm$ 0.12 in case of mouse intergenic regions and 4% $\pm$ 0.55 of human intergenic regions were above the cut-off (example distributions of expression values for different types of genomic features is shown on Figure 1). Less than 15% and 20% of intergenic regions for mouse (n=17) and human (n=16) data accordingly were ever called "expressed".

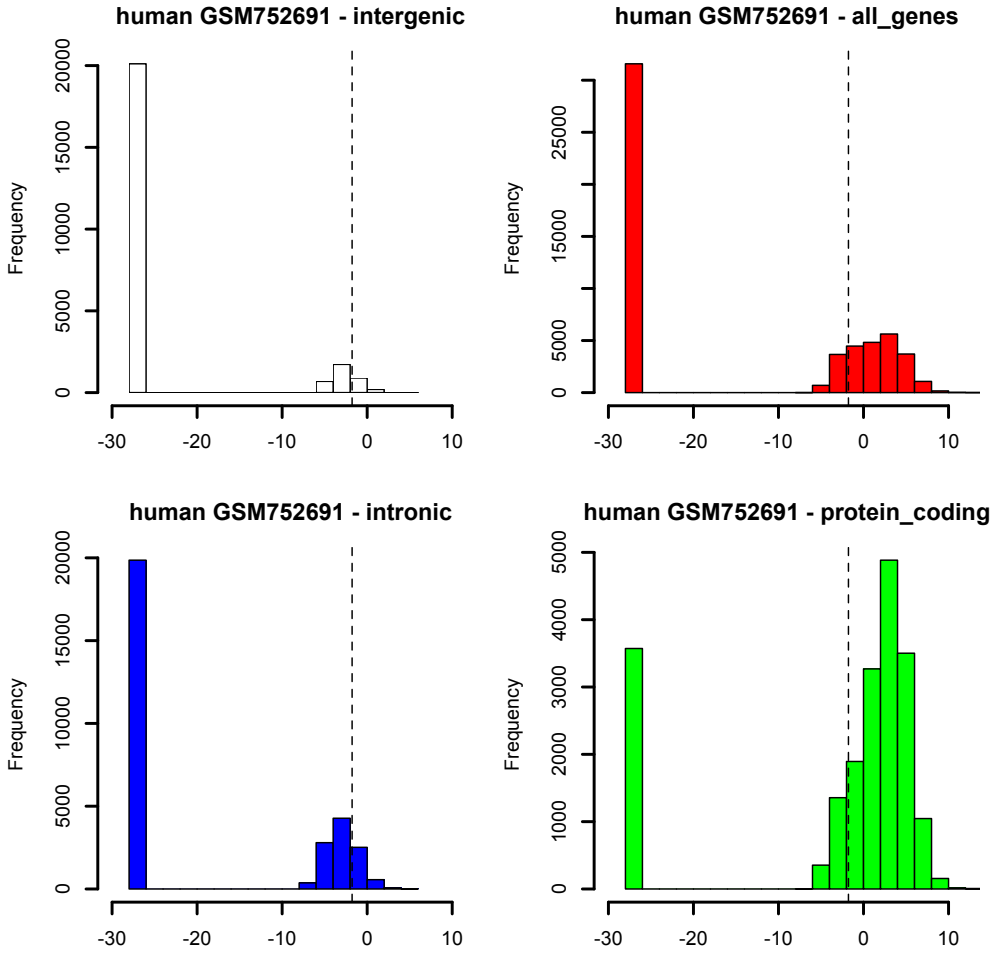


Figure 1. Distribution of log2 (RPK + 1e-08) values for different feature categories, dashed line specify cutoff.

In contrary more than 80% of mouse and 90% of human protein coding genes were at least once called "expressed". Moreover according to our results, among protein coding genes more than 50% in case of human data and 40% in case of mouse data are expressed ubiquitously in all analyzed samples. If we took, as criterion of transcription, the presence of at least one uniquely mapped read then many intergenic regions would have to be classified as expressed, which we believe would be less informative. Moreover, thanks to our methodology it is possible to avoid applying a single arbitrary cut-off for all libraries.

Acknowledgements

This work was supported by the Swiss Institute of Bioinformatics, by the Swiss National Science Foundation [grant number 31003A 133011/1], and by Etat de Vaud.

References

Archer, K.J., and Reese, S.E. (2010). Detection call algorithms for high-throughput gene expression microarray data. *Briefings in bioinformatics* **11**, 244-252.

Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In Data Integration in the Life Sciences, A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux, eds. (Springer Berlin Heidelberg), pp. 124-131.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-348.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., and Teichmann, S.A. (2011). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular systems biology* **7**, 497.

HTSeq: Analysing high-throughput sequencing data with Python. [http://www-huber.embl.de/users/anders/HTSeq/]

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111.

Analysis pipeline for the detection of mutations causative of rare diseases on whole exome sequencing data

Antonio Rueda[1], Francisco Javier López[1], Javier Pérez[1], Pablo Arce[1], Luis Miguel Cruz[1], José Carbonell[2], Jorge Jiménez-Almazán[2], Enrique Vidal[2], Guillermo Antiñolo[1,3], Joaquín Dopazo[1,2], Javier Santoyo[1] ✉

[1]Andalusian Human Genome Sequencing Centre (CASEGH), Medical Genome Project (MGP), Sevillia, Spain
[2]Institute of Computational Genomics, Principe Felipe Research Centre (CIPF), Valencia, Spain
[3]Unidad de gestión clínica de genética, reproducción y medicina fetal. Instituto de Biomedicina de Sevilla (IBIS), Hospital Universitario Virgen del Rocío-CSIC-University of Seville, Sevilla, Spain

Motivation and Objectives

Recent advances in high-throughput sequencing technologies have made exome sequencing to be an outstanding tool for finding disease associated mutations at a relatively low cost. However, it is a non-trivial task to transform the vast amount of sequence data into meaningful variants to improve disease understanding. Several challenges arise when dealing with this approach, being critical checkpoints the raw read preprocessing, mapping procedure, variant calling and posterior variant selection. A number of computational algorithms and pipelines have been reported for variant analysis (Kumar *et al.,* 2009; Lam *et al.*, 2012; Li *et al.*, 2012; San Lucas *et al.,* 2012; Yandell *et al.,* 2011; Wang *et al.*, 2010) although none of them provide a complete strategy from raw data to mendelian analysis results. Here, we present a methodology that spans from SOLiD raw reads processing to mendelian analysis and variant selection, and its application over a set of samples from The Medical Genome Project, which proves the good performance of the applied methodology.

Methods

The input of the pipeline is an xsq file generated by Applied Biosystem SOLiD 5500 XL sequencers, while the output is the result of variant annotation and mendelian analysis, assuming samples to be derived from a group or a family. A brief description of the steps is provided below:

1. Fasta and qual files generation from xsq files.
2. Duplicated reads removal.
3. BLAT-like Fast Accurate Search Tool v0.7.0a (BFAST) (Homer *et al.,* 2009) for read mapping.
4. BAM cleaning: duplicated alignments and mismatched reads removal.
5. BAM realignment and SNV calling using the Genome Analysis Toolkit v1.4.14 (GATK) (DePristo *et al.,* 2011)
6. Variant quality filter based on GATK Best Practices V3 and depth filter.
7. Annotate Variation package (ANNOVAR) for variant annotation (Wang *et al.,* 2010); SIFT (Kumar *et al.,* 2009), Polyphen (Adzhubei *et al.,* 2010), 1000 genomes frequency (The 1000 genomes project consortium, 2010) and dbSNP (Sherry *et al.,* 2001) for assessment of variant frequency.
8. Mendelian filter of deleterious variants.

Results and Discussion

The Medical Genome Project (MGP) aims to characterize a large number of rare genetically-based diseases. As a proof of concept, we selected from the MGP a set of affected individuals by several hereditary rare diseases, their healthy relatives and a set of 50 control healthy individuals from Spanish population. The full methodology was run and the results reveal a number of deleterious haplotypes in several genes which could be directly associated with the diseases.

The validation of some of the predicted variants by the pipeline shows the good performance of our methodology analysis. Critical aspects to achieve such good performance are (i) BAM filtering, since an excessive number of mismatches are allowed by BFAST for short reads; (ii) the selection of variant filters and quality thresholds as recommended by GATK Best Practices V3 in combination with a depth threshold allowing high quality calls and (iii) the inclusion of control individuals in the analysis, which is essential since they remove population variants which can disturb the interpretation of the final variant set.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A et al. (2010) A method and server for predicting damaging missense mutations. *Nature methods* **7**(4), 248-249. doi:10.1038/nmeth0410-248

DePristo M, Banks E, Poplin R, Garimella KV, Maguire JR et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics,* **43**, 491-498. doi: 10.1038/ng.806

Homer N, Merriman B, Nelson SF (2009) BFAST: An Alignment Tool for Large Scale Genome Resequencing. *PLoS ONE* **4**(11). doi:10.1371/journal.pone.0007767

Kumar P, Henikoff S, Ng P.C. (2009): Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073-1081. doi:10.1038/nprot.2009.86

Lam HYK, Cuping P, Clark MJ, Lacroute P, Chen R et al. (2012) Detecting and annotating genetic variations using the HugeSeq pipeline. *Nature Biotechnology* **30,** 226-229. doi:10.1038/nbt.2134

Li MX, Gui HS, Kwan JSH, Bao SY and Sham PC (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucl. Acids Res.* **40**(7). doi: 10.1093/nar/gkr1257

San Lucas FA, Wang G, Scheet P and Peng B (2012) Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* **28**(3), 421-422. doi:10,1093/bioinformatics/btr667

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L et al. (2001) dbSNP: the NCBI database of genetic variation. *Nucl. Acids Res.* **29**, 308-311. doi: 10.1093/nar/29.1.308

The 1000 genomes project consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1051. doi:10.1038/nature09534

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl. Acids Res.* **38**(16). doi: 10.1093/nar/gkq603

Yandell M, Huff C, Hu H, Singleton M, Moore B et al. (2011) A probabilistic disease-gene finder for personal genomes. *Genome Research* **21**(9), 1529-1542. doi: 10.1101/gr.123158.11

Error profiles for next generation sequencing technologies

Melanie Schirmer[1], Linda D'Amore[2], Neil Hall[1], Christopher Quince[2]✉
[1]University of Glasgow, United Kingdom
[2]University of Liverpool, United Kingdom

Motivation and Objectives

Next generation sequencing has revolutionized genome research and marked the start of a new era. These new technologies present us with unprecedented amounts of data - but with this sequencing data come errors that are not only platform specific but also depend on the library preparation method and the type of sequencing (i.e. amplicon or metagenome). Illumina's sequencing platforms are currently among the most utilized platforms as they are able to generate millions of reads at relatively low cost – but Illumina error profiles are still poorly understood. A better knowledge of the error profiles is essential for sequence analysis and vital in order to draw valid conclusions. It has been reported that the major source of errors for Illumina are substitution-type miscalls (Archer *et al.*, 2012). We developed a program that enables us to infer error profiles based on sequencing data from mock communities. This allows us to study and compare different errors and biases introduced by different sequencing machines, different library preparation methods as well as different types of sequencing. Here, we present the metagenome error profiles for a mock community that was sequenced on the Genome Analyzer (GA) II for the standard Illumina library preparation method (TruSeq). Being able to infer error profiles for individual sequencing runs has the potential to greatly improve our ability to correct errors and thus enhance further sequencing analysis.

Methods

For our error profiles we used a diverse mock community that was constructed by combining even amounts of purified genomic DNAs (Shakya *et al.,* 2013). The mock community consists of 49 bacterial genomes and 10 archaeal genomes covering most phyla and the community also contains closely related species and strain pairs. We sequenced a sample of the mock community on the GA II. The libraries for the sample were prepared with the standard Illumina library preparation method (TruSeq) with a starting amount of 500ng of DNA. This yielded about 6 million forward and reverse reads of 101bp.

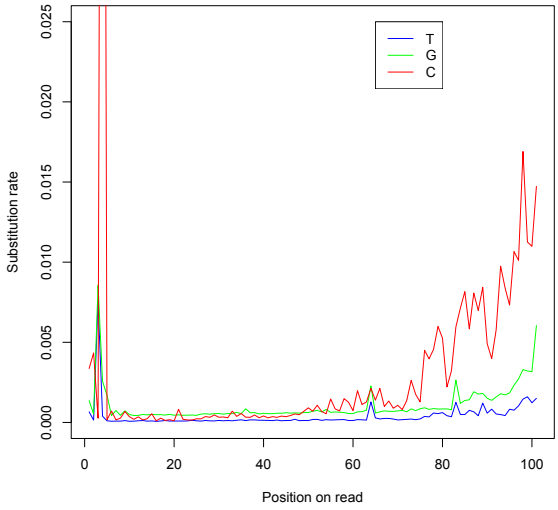
First, we aligned the reads with BWA (Li and Durbin, 2009) against the 59 reference genomes. Then we converted the files to SAM format and generated the MD tag with samtools (Li *et al.*, 2009). Based on the resulting files our program infers position and nucleotide specific substitution rates. Whenever a substitution is encountered, we identify the reference nucleotide based on the MD tag and the substituting nucleotide on the read is determined based on the extended CIGAR string. The output of our program consists of four 4x101 matrices (one for each possible "original" nucleotide) for the set of forward and reverse reads, respectively, in which we store the number of observed substitution types for each position of the read. We then normalize these matrices as follows: We count the number of occurrences of, for example, A on the read for each position, add the number of detected substitutions from A to T, G and C, respectively, and subtract the number of substitutions from T, G and C, respectively, to A at this position. This accounts for errors and reflects the true number of A's. In addition, our program computes the overall insertion and deletion rate.

To verify our algorithm we extended our read simulation program (Schirmer *et al.*, 2012) to generate reads based on error profiles of the above described format. We simulated one million forward and reverse reads based on the error profiles inferred from the GA II run. The error profiles, inferred from the simulated reads, concurred with the original error profiles used to simulate the reads and thus validates the algorithm.

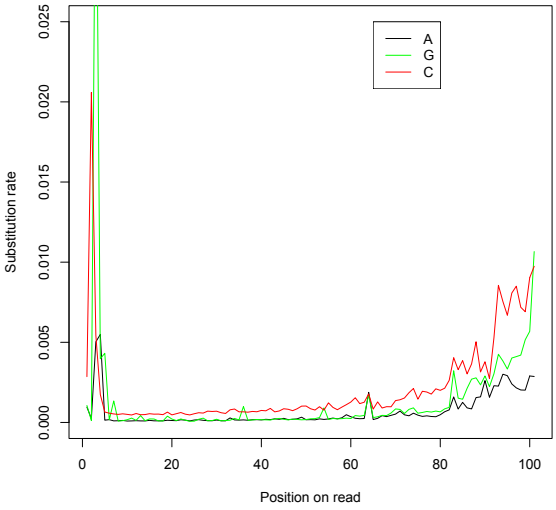
Results and Discussion

The GA II error profiles show a strong increase in the number of substitutions towards the end of the read. The average substitution rate for the forward reads is $\approx$0.004, where several spikes were observed across the first 10bp as well as an increase in substitutions starting from the middle of the read towards the end of the read. For

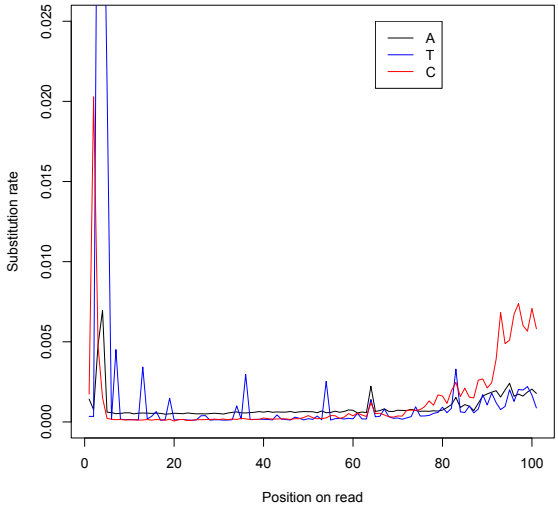
the reverse reads the average substitution rate is ≈0.012. The start of the reverse reads also shows several spikes in the error profile but less compared to the forward reads. Though smaller spikes were observed across the whole read length. The substitution rate starts to increase after the first third of the read and is overall significantly higher towards the end of the read compared to the forward reads. We observed the highest substitution rates for A and the lowest substitution rates for G for both forward and reverse reads (disregarding the first 10bp). Subsequently we examined the frequencies of the nucleotides for each position across the reads to test for possible



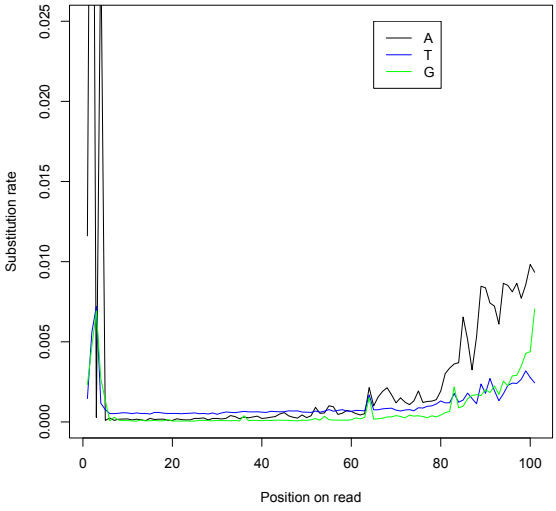
(a) R1 reads: original nucleotide A



(b) R1 reads: original nucleotide T



(c) R1 reads: original nucleotide G



(d) R1 reads: original nucleotide C

Figure 1: Error profile for forward reads: The x-axis indicates the position on the read and the y-axis the substitution rate (# of observed substitution/# of occurrences of the "original" nucleotide). Each subfigure represents one of the four possible original nucleotides for which different types of substitutions are indicated by different colors.

artifacts, as these could explain the spikes in the error profile at the read-start. For metagenomic data sets we expect a uniform frequency distribution across the reads for all nucleotides. Here, we identified fluctuations within the first 10bp that sufficiently account for the increased error rates across these positions. Separating the error profiles according to the different substitution types presented further insights. Figure 1 shows that - independent of the original nucleotide – a substitution with C is the most common error towards the end of the read. If the original nucleotide is a C a substitution with A is the most common error. Inferring error profiles for different sequencing machines, library preparation methods and sequencing types has great potential for error correction. It also enables us to infer error profiles for individual sequencing runs by including a mock community (e.g. instead of PhiX). We will extend our research to different sequencers, more library preparation methods and different types of sequencing to identify differences and similarities in the error profiles and as a possible guideline for experimental design.

Acknowledgements

This research is part of a project funded by the Technology Strategy Board. M.S. is supported by Unilever R&D Port Sunlight, Bebington, United Kingdom.

References

Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A et al. (2012) Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics* **13**(1), 47.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*,**25**(16), 2078–2079

Schirmer M, Sloan WT and Quince C (2012) Benchmarking of viral haplotype reconstruction programmes: An overview of the capacities and limitations of currently available programmes. *Brief. Bioinfo.* (online - ahead of print]). doi: 10.1093/bib/bbs081

Shakya M, Quince C, Campbell J, Yang ZK, Schadt C *et al.* (2013) Comparative meta-genomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology*. doi:10.1111/1462-2920.12086

Scripting for large-scale sequencing based on Hadoop

André Schumacher[1,2,3], Luca Pireddu[4], Aleksi Kallio[5], Matti Niemenmaa[6], Eija Korpelainen[5], Gianluigi Zanetti[4], Keijo Heljanko[2,3]✉
[1]ICSI, Berkeley, USA
[2]Helsinki Institute for Information Technology HIIT, Helsinki, Finland
[3]Aalto University, Espoo, Finland
[4]CRS4, Pula, Italy
[5]CSC-IT Center for Science, Helsinki, Finland
[6]Aalto University, Espoo, Finland

Motivation and Objectives

The large volumes of data generated by modern sequencing experiments present significant challenges in their manipulation and analysis. Traditional approaches, such as scripting and relational database queries, are often found to be inadequate, frustratingly slow, or complicated to scale. These problems have already been faced by the "big data revolution" in data-based activities resulting in novel computational paradigms such as MapReduce and scalable tools such as Hadoop and Pig.

We describe our ongoing work on SeqPig, a tool that facilitates the use of the Pig Latin scripting language to manipulate, analyze and query sequencing data. SeqPig provides access to popular data formats and implements a number of high level functions. Most importantly, it grants users access to the proven to be scalable platform that is Hadoop from a high level scripting language, whether the cluster is run locally or *in the cloud*.

Methods

SeqPig operates on top of *Hadoop* and *Pig* and augments them to facilitate their use to process sequencing data. Hadoop is a distributed computing framework that implements the MapReduce programming model, which expresses computations as sequences of *side-effect* free Map and Reduce functions. Hadoop was initially developed at Yahoo!, but has since been widely adopted, e.g. by Facebook, Twitter and LinkedIn. Pig is a set-based scripting language whose instructions are compiled to a sequence of MapReduce jobs, which are then executed on a Hadoop cluster. It effectively simplifies the use of a Hadoop cluster through its concise SQL-like logic. Both Hadoop and Pig are projects supported by the Apache Software Foundation (http://hadoop.apache.org, http://pig.apache.org).

SeqPig

SeqPig extends Pig with a number of features and functionalities conceived for processing sequencing data. Specifically, it provides: 1) data input and output components, 2) specialized functions to extract fields and to transform data and 3) a collection of scripts for frequent tasks (e.g., pileup, QC statistics).

SeqPig provides import and export functions for file formats commonly used for sequencing data: Fastq, Qseq, SAM and BAM. SeqPig supports ad hoc – scripted or even interactive – distributed manipulation and analysis of large sequencing datasets. Unlike traditional methods, the scalable nature of Pig allows the speed of its operations to scale with the computing resources available. SeqPig includes functions to access SAM flags, split reads by base (for computing base-level statistics), reverse-complement reads, calculate read reference positions in a mapping (for pile-ups, extracting SNP positions), and more. The authors are currently working on expanding the library of functions, and SeqPig is an open source project that welcomes and encourages contributions from the community.

Using cloud-based resources

SeqPig has been tested on Amazon's Elastic MapReduce service. Users may rent computing time on the cloud to run their SeqPig scripts, and even share their S3 storage buckets with other cloud-enabled software.

Dependencies

SeqPig builds on Hadoop-BAM (Niemenmaa *et al.,* 2012), Seal (Pireddu *et al.,* 2011), and Picard (http://picard.sourceforge.net). Hadoop-BAM implements a number of file formats for Hadoop, while Seal and Picard implement some of the

sequence analysis functiona-lity that SeqPig exposes at a higher level.

Results and Discussion

SeqPig enables the manipulation and analysis of sequencing data on the Hadoop big-data computational platform. At CRS4 SeqPig is already used routinely for some steps in the production workflow; in addition, SeqPig scripts have been used for ad hoc investigations into data quality issues, comparison of alignments tools, and reformatting or packaging data. In the future we plan to expand its function library and thoroughly test its scalability and performance characteristics.

Acknowledgements

This work was supported by the Cloud Software and D2I Programs of the Finnish Strategic Centre for Science, Technology and Innovation TIVIT and by the Sardinian (Italy) Regional Grant L7-2010/ COBIK.

References

Niemenmaa M, Kallio A, Schumacher A, Klemelä P, Korpelainen E, and Heljanko K. (2012) Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. *Bioinformatics* **28**(6):876-877. doi:10.1093/bioinformatics/bts054

Pireddu L, Leo S, and Zanetti G. (2011) SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics* **27**(15):2159-2160. doi:10.1093/bioinformatics/btr325

Shape matters: Differential peak detection for Chip-seq data sets

Gabriele Schweikert, Guido Sanguinetti ✉
University of Edinburgh, Edinburgh, United Kingdom

Motivation and Objectives

ChIP-Seq has rapidly become the dominant experimental technique in functional genomic and epigenomic research. Statistical analysis of ChIP-Seq data sets however remains challenging, due to the highly structured nature of the data and the paucity of replicates. Current approaches to detect differentially bound or modified genomic regions are mainly borrowed from RNA-Seq analysis tools, e.g. (Ross-Innes, 2012, Anders and Huber, 2010). With these methods a given peak is represented by a single number: the total number of reads mapping to the peak region. Any information encoded in the structure of the peak is ignored. However, the shape of an enrichment peak at a given genomic location tends to be highly reproducible across biological replicates and increasing evidence hints towards a functional role of these profile structures (The ENCODE Project Consortium, 2012). To complement count-based methods, we present MMDiff, a new non-parametric statistical testing methodology to identify significant shape differences in profile patterns of enrichment peaks between different conditions.

Methods

The underlying idea is to treat each peak as a *distribution* over a finite space given by the starting positions of all reads. The problem of testing for differential binding is then reduced to testing whether two samples are generated by the same probability distribution. As there is a large variability for observed peak profiles at different genomic locations we cannot make any assumption about the type of distribution. We therefore take advantage of recent kernel-based tests developed in the machine learning community (Gretton *et al.*, 2012). Here, the non-linearity of the original data is captured with a positive definite Kernel and the data is mapped into a high-dimensional reproducing Kernel Hilbert space (RKHS). In the RKHS the mean element of a distribution contains the information of all higher-order moments. Furthermore, the distance between the mean elements of two distributions, the maximum mean discrepancy (MMD), can be used as test statistic. Intuitively, the greater the distance, the more different the distributions are. Here we use the 5' position of the mapped reads as features and an RBF Kernel, where the width of the Kernel is heuristically determined. To obtain empirical p-values we compute the probability of observing MMD values between biological replicates which are at least as extreme as those observed between conditions. To correct for multiple testing we compute false discovery rates (FDRs) according to (Benjamini and Hochberg, 1995).

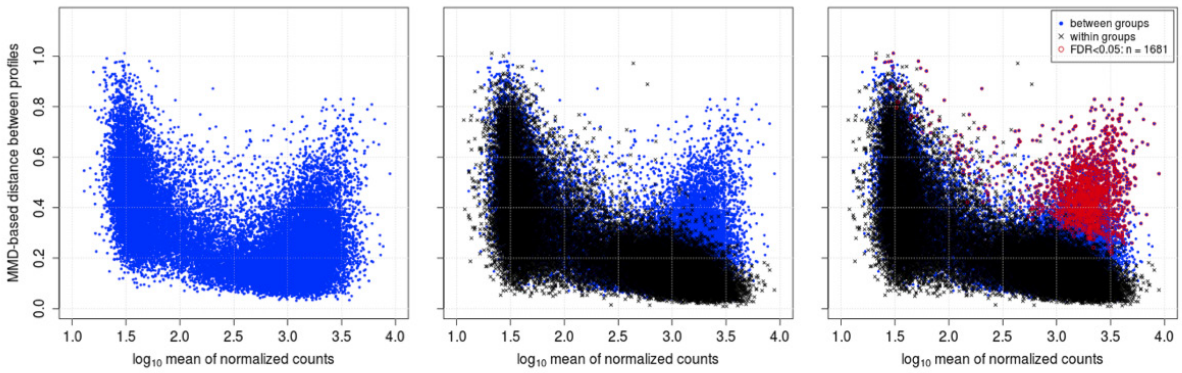


Figure 1. Scatter plots showing MMD as a function of averaged normalized total counts, where each dot represent one examined promoter. Left: MMD determined between WT and Null. Middle: Overlayed are the MMDs for biological replicates (black) Right: Additionally, profiles that are significantly different in WT vs Null ($p < 0.05$) are shown in red.

Results and Discussion

One of the best studied epigenomic marks is trimethylation of Lysine 4 at histone 3 (H3K4me3), which is associated with active gene promoters. However, the mechanisms and dynamics by which this mark is established at its target locations are not well understood. We apply our method to a recently published data set by (Clouaire *et al.*, 2012), which examines the role of one key player, the DNA binding protein Cfp1. To this end, H3K4me3 profiles in wt ES cells and mutant cell lines lacking Cfp1 are compared.

Our empirical analysis shows that MMDiff is complementary to count based methods, that it provides highly reproducible results and that it is able to detect biologically relevant changes in histone modifications.

To make the method available to a wider range of users we have developed an R package, called MMDiff which is released with the la-test Bioconductor version (2.12).

Acknowledgements

We would like to thank Arthur Gretton, Rory Stark and Gunnar Raetsch for helpful discussions. Shaun Webb is thanked for computing support. Thomas Clouaire and Adrian Bird provided the data and helped with discussions.

G.Schw. acknowledges support from EC through the FP7- Marie Curie project "Epigenome Informatics". G.S. acknowledges support from the European Research Council through grant MLCS306999.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol* **11**(10), R106.

Benjamini and Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing; Journal of the royal statistical society. series a, Statistics in Society. JRSS. Series A, Statistics in society, **57**(1), 289.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., et al. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev* **26**(15), 1714– 28.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schoelkopf, B., and Smola, A. J. (2012). A kernel two-sample test. *JMLR* **13**, 723–773.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., et al. (2012). Differential Oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**(7381), 389–93.

Comparison of oligonucleotide microarray and RNA-SEQ technologies in the context of gene expression analysis

Nicolas Sierro, Florian Martin, Carine Poussin, Julia Hoeng, Nikolai V. Ivanov✉
Philip Morris R&D, Neuchatel, Switzerland

Motivation and Objectives

For more than a decade, the microarray technology has been widely and extensively used to profile gene expression in various study types. Transcript abundance measurement is currently revolutionized by RNA-seq technology. Indeed, RNA-seq enables the sequencing of the whole transcriptome (sequence-centric data) while only predefined transcripts/genes can be measured on arrays (gene-centric data). In addition, the nature of RNA-seq data renders the analysis more flexible for addressing biological questions going from transcript/gene (differential) quantification to transcript structure (splice variants) identification (qualitative analysis) to cite a few. This latest application requires the use of additional specific arrays (e.g. exon), which have some limitations. The purpose of our work here was to compare both affymetrix GeneTitan array and Illumina HiSeq-2000 sequencing technologies in the context of gene expression analysis.

Methods

For this study, mRNA samples from lung tissue of ApoE mice exposed to conventional cigarette smoke (CS) or fresh air (Sham) for 3 and 6 months, or from ApoE mice exposed to CS for 3 months and then exposed to fresh air for 3 months (Cessation) were hybridized on Affymetrix MG-HT430PM GenTitan array or sequenced on Illumina HiSeq-2000 (2x100bp paired-end run, ~30 to 100 million paired reads per sample). Quality control analysis was performed for each data type accordingly. RNA-seq data were cleaned using the fastx toolkit. Briefly, 3'-ends of reads were trimmed with a quality threshold of 20, and filtered to retain only reads of at least 50 bases, at least 90% of which with a base-calling quality of 20 or more. Quantification at the gene level was performed using RSEM and VOOM-LIMMA to accommodate for mean-variance trend or Cufflinks followed by Cuffdiff for differential count analysis. Array data were pre-processed using RMA and differential expression was carried out using LIMMA.

Results and Discussion

Comparing gene expression abundances measured by both technologies revealed a significant correlation for highly expressed genes, while this correlation significantly decreased for low expressed abundance genes. The Illumina HiSeq-2000 showed a wider range of expression values than the Affymetrix GeneTitan array for low expressed genes. This result indicates that the Illumina HiSeq-2000 sequencing technology has a higher sensitivity to detect low expressed genes. When comparing differential expression (treatment vs control samples), a higher fold change magnitude was observed in the volcano plot of data generated with the Illumina HiSeq-2000. The magnitude of the significance of the observed fold changes was similar between both platforms. However, in-more-depth analysis varying False Discovery Rate and Fold Change thresholds showed that Illumina HiSeq-2000 is more sensitive to detect differentially expressed genes. Further investigations will be undertaken to understand if these low expressed transcripts/genes detected by RNA-seq reveal new biological functions or not compared to those identified with the array technology. Overall, this study shows that RNA-seq is a very powerful technology since it is more sensitive to detect low and differential expressed transcripts/genes compared to the Affymetrix GeneTitan technology.

Acknowledgements

The project is funded by PMI.

Rapid whole genome sequencing investigation of a familial outbreak of E. coli O121:H19 with a sheep farm as the suspected source

Robert Söderlund[1,2], Cecilia Jernberg[3], Christine Källman[2], Ingela Hedenström[3], Erik Eriksson[2], Erik Bongcam-Rudloff[1], Anna Aspán[2]✉
[1]SLU Global Bioinformatics Centre, Swedish University for Agricultural Sciences, Uppsala, Sweden
[2]National Veterinary Institute, Uppsala, Sweden
[3]Swedish Institute for Communicable Disease Control, Solna, Sweden

Motivation and Objectives

Finding the source of an outbreak of zoonotic bacterial disease requires comparison of patient isolates to isolates found in food, animal or environmental samples. The current gold standard for this is pulsed field gel electrophoresis (PFGE), but multi-locus VNTR analysis (MLVA) has also been gaining popularity in recent years. Methods with too low discriminatory power can lead to false positives which are highly disruptive for food producers and costly for society, while excessively discriminatory markers can be affected by genetic changes which have occurred within an outbreak, causing false negatives. The recent emergence of benchtop high-throughput sequencing instruments has made whole genome sequencing (WGS) of bacteria a viable alternative to the currently available methods in terms of speed and cost, while potentially providing far more reliable genetic comparison data.

In 2012, verotoxin 2 (vtx2)-positive *E. coli* O121:H19 was isolated from a Swedish EHEC patient and two asymptomatic family members. A single isolate of O121:H19 vtx2$^+$ was found in samples taken from sheep at a farm where the patient had patted the animals. PFGE and MLVA analyses were inconclusive with limited variation both within the family and between the family and animal isolates. To resolve this and evaluate WGS as a tool for routine molecular epidemiology, the genomes of the three familial outbreak isolates, the isolate from the farm and two unrelated patient isolates were sequenced and compared.

Methods

MLVA was performed according to a generic *E. coli* protocol targeting a total of 10 loci. PFGE was performed according to PulseNet standard laboratory protocols (pulsenetinternational.org). For sequencing, crude lysates of bacteria harvested from agar plates were treated with Qiagen RNAseA and DNA extracted using a Qiagen EZ1 Biorobot. The DNA was quantified using the Qubit BR kit, and the integrity was checked by slab gel electrophoresis. Libraries for sequencing were prepared using the Nextera XT DNA Sample Preparation Kit with indexing, using 1 ng of extracted DNA as starting material. Libraries were verified with an Agilent Bioanalyzer HS DNA kit and run on a Illumina MiSeq instrument using the MiSeq V2 500 cycle run kit (2*250bp). The generated sequences were assembled using MIRA with contig filtering based on size and coverage. SNP discovery with the publicly available MT#2 draft genome sequence (GenBank AGTJ01000000) as reference was performed using MUMmer with further quality filtering, and SNP states were extracted from all contig sets using BLAST+ with the input and output treated by custom R scripts. Phylogenetic trees were drawn using the neighbor-net algorithm in SplitsTree. Regions of interest, e.g. known virulence factors in O121 and other types of EHEC as well as targets for multi locus sequence typing (MLST) and clustered regularly interspaced short palindromic repeats (CRISPR) typing were extracted from the contig sets. The time required to complete the full process was approximately one working week, with limited hands-on time.

Results and Discussion

The sequenced draft genomes had estimated sizes around 5.5 Mbp with an average coverage of 33x - 44x, and N50 values of 55 kbp - 96 kbp. Comparison of the contig sets identified 369 high quality SNPs in regions conserved in all included isolates. Analysis of the SNP data strongly supported a recent common origin for two of the three outbreak isolates, one of which was from the patient. These differed by a single SNP located in the coding region of an uncharacterized protein. However, the third outbreak isolate as well as the sheep isolate were distinct from each other and

Table 1. Characteristics of the sequenced O121:H19 isolates based on whole genome sequencing. *Symptomatic patient, reference for SNP similarity comparison

Source	wzx/wzy/ fliC	CRISPR	MLST	SNP similarity [%] *	Vero-toxins	Secondary virulence factors						
						eae	tir	hlyA	toxB	espP	efa1 OI-122	terB/ iha OI-48
Outbreak isolate 1*	O121:H19	CB8124	ST655	(100)	vtx2a	ε	+	+	+	+	+	+ / -
Outbreak isolate 2	O121:H19	CB8124	ST655	99.7	vtx2a	ε	+	+	+	+	+	+ / -
Outbreak isolate 3	O121:H19	CB8124	ST655	80.8	vtx2a	ε	+	+	+	+	+	+ / -
Sheep farm isolate	O121:H19	CB8124	ST655	62.4	vtx2a	ε	+	+	+	+	+	+ / -
Unrelated pat. isolate A	O121:H19	CB8124	ST655	94.3	vtx2a	ε	+	+	+	+	+	+ / -
Unrelated pat. isolate B	O121:H19	CB8124	ST655	67.2	vtx1a vtx2a	ε	+	+	+	+	+	+ / -
MT#2 (ref)	O121:H19	CB8124	ST655	45.0	vtx2a	ε	+	+	+	+	+	+ / -

from the isolate from the patient (Table 1). In fact, the outbreak patient isolate was far more similar to one of the unrelated reference isolates. Thus, there was no evidence that the sheep farm was the source of the infection.

Regions of interest were extracted from the generated sequences to produce backward compatible typing data traditionally produced by PCR or Sanger sequencing (Table 1). This analysis supported the PFGE and MLVA typing in indicating that all four isolates in the outbreak investigation belonged to the same clone of O121:H19. The prevalence of this clone in ruminants and asymptomatic humans in Sweden merits further investigation. In retrospect, sufficient typing data to exclude the farm as the source of the outbreak was produced by either PFGE or MLVA alone. However, the low total variation combined with the surprising outcome meant that WGS data was necessary for a definite answer. To see if multiple strains of O121:H19 were present at the farm, sampling of the animals was repeated at a later date, but on this occasion no O121 could be found. The source of the infection remains unknown.

The emergence of quick and affordable lab methodology combined with standardized data analysis workflows will see WGS taking an increasingly important role in the routine work of veterinary and public health authorities in the next few years.

Acknowledgements

This study was financially supported by the Swedish Civil Contingencies Agency. The Bo Segerman Group, SVA is thanked for sharing bioinformatics software and hardware. The authors also thank the staff at the SVA and SMI EHEC laboratories for excellent technical asssistance.

SOFTvenom: an omics drug discovery approach from animal venoms

Juan Carlos Triviño[1], Miñambres Rebeca[1], Raquel Rodríguez-dePablos[1], Mayte Gil[1], Pierre Escoubas[2], Marion Verdenaud[3], Sheila Zuñiga[1], Sheila Zuñiga[1] ✉
[1]Sistemas Genomicos, Paterna, Spain
[2]Venomtech, Valbonne, France
[3]CEA, Gif sur Yvette cede, France

Motivation and Objectives

Animal venoms have been proven to be a rich source for drug development due to their efficiency and target selectivity and the subsequent reduction in side effects in a wide range of therapeutic conditions such as pain or cancer where medical needs are not properly addressed by the existing treatments. Since no reference genome is currently available for most venomous animals, research in this field has been economically restricted to small animal groups and species.

Here we present SOFTVENOM, an efficient strategy to reconstruct and characterize animal venoms. Our approach was applied to the transcriptomes analysis of three animal venoms as a pilot project using RNASeq techniques in two different NGS platforms, Illumina-HiSeq2000 and 454-GSTitanium. We functionally compare the different datasets to provide new insights into a fascinating field where little information is currently known.

Methods

Strand-specific libraries were constructed from the RNA poly(A) fraction of tarantula (*Poecilotheria regalis*), scorpion (*Parabutus transvaalicus*) and viper (*Bitis arietans*). Libraries were then sequenced with Illumina Hiseq2000 and 454-GSTitanium following a paired-end and a single-end strategy respectively. 10-20Gbs were obtained after Illumina sequencing and 0.6Gb after 454-sequencing.

Reads were trimmed, filtered and collapsed to reduce dataset complexity using a combination of Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit), Trimmomatic (Lohse *et al.,* 2012) and 'in house' scripts. Assemblies were performed with Oases (Schulz *et al.,* 2012) and Trinity (Grabherr *et al.,* 2011) and further merge with CAP3 (Huang and Madan, 1999). Functional annotation and comparison between venom assemblies was carried out using the UniProt database (http://www.uniprot.org) and custom scripts.

Results and Discussion

SOFTVENOM is an analysis framework that includes different tools to process RNA-Seq data from scratch to finally obtain a list of well-established isoforms and associated functional information.

Our results provide new insights into compounds, which may play a role in venom function. In addition, our work highlights the limitations of the applied analysis strategy and the differences found after functionally comparing both NGS platforms.

The results of this work are currently being integrated with proteomics data from mass spectrometry to obtain a complete view of the venom composition. The work done so far will serve as the fundamental basis for the study of a total of 200 animal venoms in the following three years as part of the VENOMICS project, an international effort to uncover the secrets behind venom activity and their potential use in the development of drugs to improve Human Health.

Acknowledgements

This work has been financed by the 7th Framework Program.

References

Grabherr, M. G. et al. (2011): Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* **29**, 644-652. doi:10.1038/nbt.1883

Huang, X. & Madan, A. CAP3 (1999): A DNA sequence assembly program. *Genome Research* **9**, 868-877. doi: 10.1101/gr.9.9.868

Lohse, M. et al. (2012): RobiNA: a user-friendly, integrated software solution for RNA-seq-based transcriptomics. *Nucleic Acids Research* **40**, W622-W627. doi:10.1093/nar/gks540

Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases (2012): robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092. doi: 10.1093/bioinformatics/bts094

Integrated analysis of diverse genomic data

Georgia Tsiliki[1], Konstantinos Tsaramirsis[2], Sophia Kossida[1] ✉
[1]Biomedical Research Foundation of the Academy of Athens (BRFAA), Athens, Greece
[2]King's College London, London, United Kingdom

Motivation and Objectives

The increasing growth of high throughput genome-wide assays, such as next generation sequencing (NGS), is enabling the simultaneous measurement of several genomic features in the same biological samples. As a consequence forefront genome consortia have faced the challenge of integrating these diverse data types (Parson *et al.,* 2008; Network TCGAR, 2008), including RNA transcriptional levels, genotype variation, DNA copy number variation, and epigenetic marks. Often such data types produce controversial or partly overlapping results, towards a particular disease of interest, resulting in only a limited number of successful applications to everyday medical practice. Particularly in cancer research, this overall failure to translate modern advances in basic cancer biology is also attributed to the lack of comprehensively organizing and integrating all of the 'omics' features now technically acquirable on virtually any type of cancer (Vaske *et al.,* 2010). Two methodological approaches are mainly presented, namely meta-analysis techniques and integration techniques considering all the data types simultaneously, however until now there has not been developed a general and scalable statistical framework ready to incorporate as many diverse 'omics' features (Tyekucheva *et al.,* 2011). We present a model-based methodology, which considers all data together and aims in estimating important gene-sets for the pathology under study. An important objective is to validate established gene signatures, although emphasis is given on those sets which can only be found through integrated analysis. Special technical issues considered are different data formatting as well as data rescaling. The use case presented here considers breast cancer disease.

Methods

We consider microarray gene expression, RNA-seq and copy number variation measurements taken from the same samples as derived by The Cancer Genome Atlas (TCGA; https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) database. An extra source of variation introduced in the data is due to the different technology platforms they are derived from. Particularly, the microarray gene expression data are derived from Agilent and Affymetrix platforms (AgillentG4502A _ 07 _ 03, U133A, U133Plus2), the expression mRNA sequencing data from Illumina platform (IlluminaGA _ RNASeq, IlluminaHiSeq _ RNASeq) and the copy number variants from Affymetrix platform (Genome _ Wide _ SNP _ 6.0). Nevertheless diverse data are taken from the same breast cancer samples, to avoid further discrepancies.

We present a Bayesian partition model to detect genetic interactions in the data, where a Markov Chain Monte Carlo (MCMC) algorithm is designed to simultaneously search across datasets (Denison *et al.,* 2002). The above methodology is a powerful modelling approach, which can handle large number of data, and also allow for interaction across data samples. Our aim is to present a stand-alone tool able to independently analyse the data using standard clustering methodologies (hierarchical clustering algorithms) and also provide the option of an integrated analysis for all data types simultaneously.

Towards this end, we first establish a common annotation mirror, where all entries are 'translated' to chromosomal locations. Data from only chromosomal regions across data sets are considered. The above procedure results in variable number of entries per chromosomal region, given the data set, for example a chromosomal location could include one gene and two copy number variants. This is addressed by averaging the data entries over the region. Similar averaging techniques have been applied to cross-platform microarray data in the past.

Regression modelling is employed for each chromosomal location and each sample, to test the null hypothesis that the particular chromosomal location is of no interest to breast cancer disease. The above procedure results in a single data set for all data types which is then further

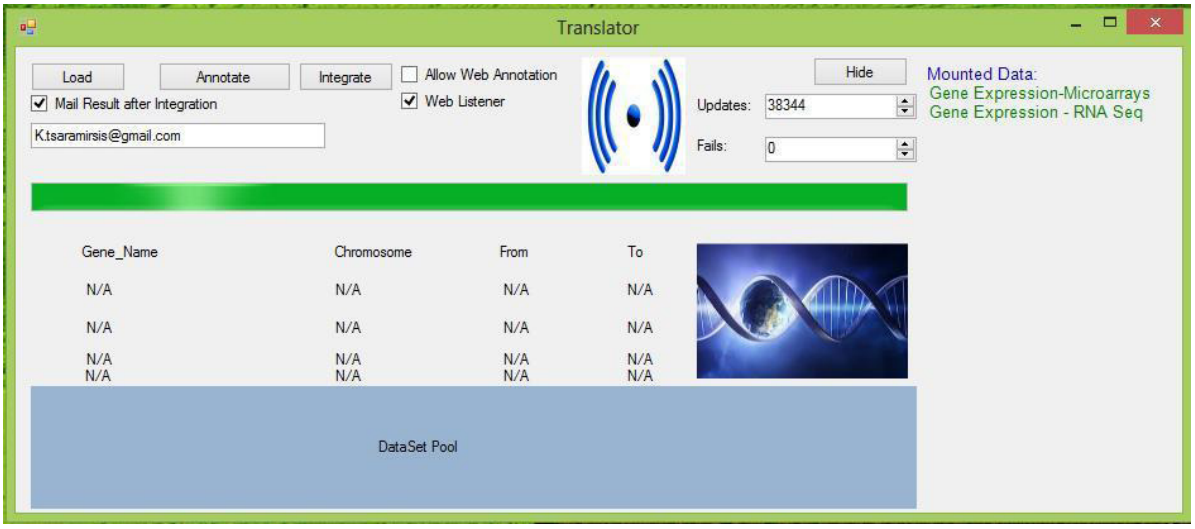


Figure 1. The integrated analysis stand-alone tool. The first form of the tool is shown, where users can upload the data and proceed with annotating or integrating the diverse data types.

analysed for estimating interesting gene signatures. The partition modelling is employed to the derived data set aiming to estimate clusters of homogeneous data which are then interpreted as breast cancer gene-signatures and cross validated with those derived from the individual data analysis and well-established gene-signatures.

In Figure 1, we show an instance of the application tool produced; users are prompted to upload the data sets one at a time, and can either proceed with 'annotating' the data using the chromosomal region scheme introduced above, or directly proceed with the integrated analysis.

Results and Discussion

Both simulated and empirical data examples demonstrated our method's ability to detect highly correlated data groups across platforms and provided key insights into previously defined gene expression subtypes. The accompanied stand-alone tool will be freely available via our website. Future plans include extending the methodology presented to other data types and technological platforms. An interesting extension is also considering other pathologies to identify significant molecular heterogeneity.

Acknowledgements

Research carried out in the context of this study has been funded by the EU DICODE (Mastering Data-Intensive Collaboration and Decision Making) Collaborative Project (FP7, ICT- 2009.4.3, Contract No. 257184) and EU COST Action SeqAhead (BM1006).

References

Denison DGT, Adams NM, Holmes CC, Hand DJ (2002) Bayesian partition modelling, *Comput Stat Data Anal*, **38** (4): 475– 485.

Network TCGAR (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**: 1061-1068 doi: 10.1038/nature07385

Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**: 1807-1812. doi: 10.1126/science.1164382

Tyekucheva S, Marchionni L, Karchin R, and Parmigiani G (2011) Integrating diverse genomic data using gene sets. *Genome Biol*, **12**: R105. doi: 10.1186/gb-2011-12-10-r105

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C *et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**: i237-i245. doi: 10.1093/bioinformatics/btq182

Computational cleaning of noisy 5' end tag sequencing data sets from rare in vivo cells

Johannes Eichler Waage[1], Ilka Hoof[1], Jette Bornholdt[1], Esben Pedersen[2], Mette Jørgesen[1], Kim Theilgaard[3], Cord Brakebusch[1], Bo Porse[4], Albin Sandelin[1] ✉
[1]Bioinformatics Centre, University of Copenhagen, Denmark
[2]Biomedical Institute, BRIC, University of Copenhagen, Copenhagen, Denmark
[3]Biotech Research and Innovation Centre, University of Copenhagen, Denmark
[4]The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Motivation and Objectives

CAGE-seq, cap analysis of gene expression followed by next-generation sequencing, allows for precise profiling of the promoterome (Plessy *et al.,* 2010) Here, we present a data filtration and processing pipeline for analysis of nanoCAGE-seq, a variant of the method allowing for very small amounts of input material (~50 ng per sample), and thus expanding the number of tissue- and cell types available for proteome profiling. We show, however, that low-intensity signal across exons, mRNA degradation and other method-specific noise is common to this technique, obfuscating true promoters in the dataset. Rigorous filter methods, including tag clustering, cluster width and profile filtering, and variance filtering rescue bona fide promoters, allowing for detection of promoter usage, inter-sample promoter switching and detection of new putative promoters. These types of filtering methods could potentially also be used on other noisy next-generation data sets. Here, we present result from nanoCAGE from two different studies; data from a mouse melanoma skin cancer model, as well as data from human acute promyelocytic leukemic blast populations.

Methods

For both studies, samples were sequenced in biological triplicates on the Illumina Genome Analyzer II and the Illumina HiSeq 2000, quality validated by fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed when necessary, and were mapped to the *mus* (mm9) and *homo* (hg19) genomes by Bowtie (Langmead *et al.,* 2009).

For *homo* and *mus* nanoCAGE data from the Rac1 project, single tags were removed, and all sample were merged followed by consensus generation by merging all tags within 20bp. Next, we required a 2/1 cluster height to width ratio. Clusters having a width of 5 tags or less were removed to filter for PCR-amplification artifacts. Tags were counted in consensus clusters per sample, and expression values were quantified as TPM (tags per million mapped). Clusters with <5 TPM in the highest sample were removed to filter out noise in the lowest band. The intra-replicate coefficient of variance (CV) was calculated for all samples, and clusters with a CV higher than 1 were removed.

All statistical analyses were performed in the statistical package R (Ihaka *et al.,* 1996), and the Bioconductor package edgeR (Robinson *et al.,* 2010) was used for differential testing.

Results and Discussion

We present preliminary results from nanoCAGE in two different studies. First, we show data from nanoCAGE of epidermal cells in a model of melonama skin cancer, harvested from Rac1 KO vs. WT mice, treated with or without the proliferative agent tetradecanoylphorbol acetate (TPA). This four-way experiment allows a detailed characterization of promoter usage of treated vs. untreated mice, and how the Rac1 gene contributes to the gene expression in hyperplastic vs. normal skin cells. After the initial stringent filtering, we present a confined set of high confidence promoters (figure 1) and their interactions between the samples and treatments. Secondly, we present preliminary results from nanoCAGE-seq of blast cells of human acute promyelocytic leukemic populations versus the corresponding normal hematopoietic progenitor cell, revealing, among other things, a pattern of promoter switching from full length transcript to shorter transcripts in the cancer cells.

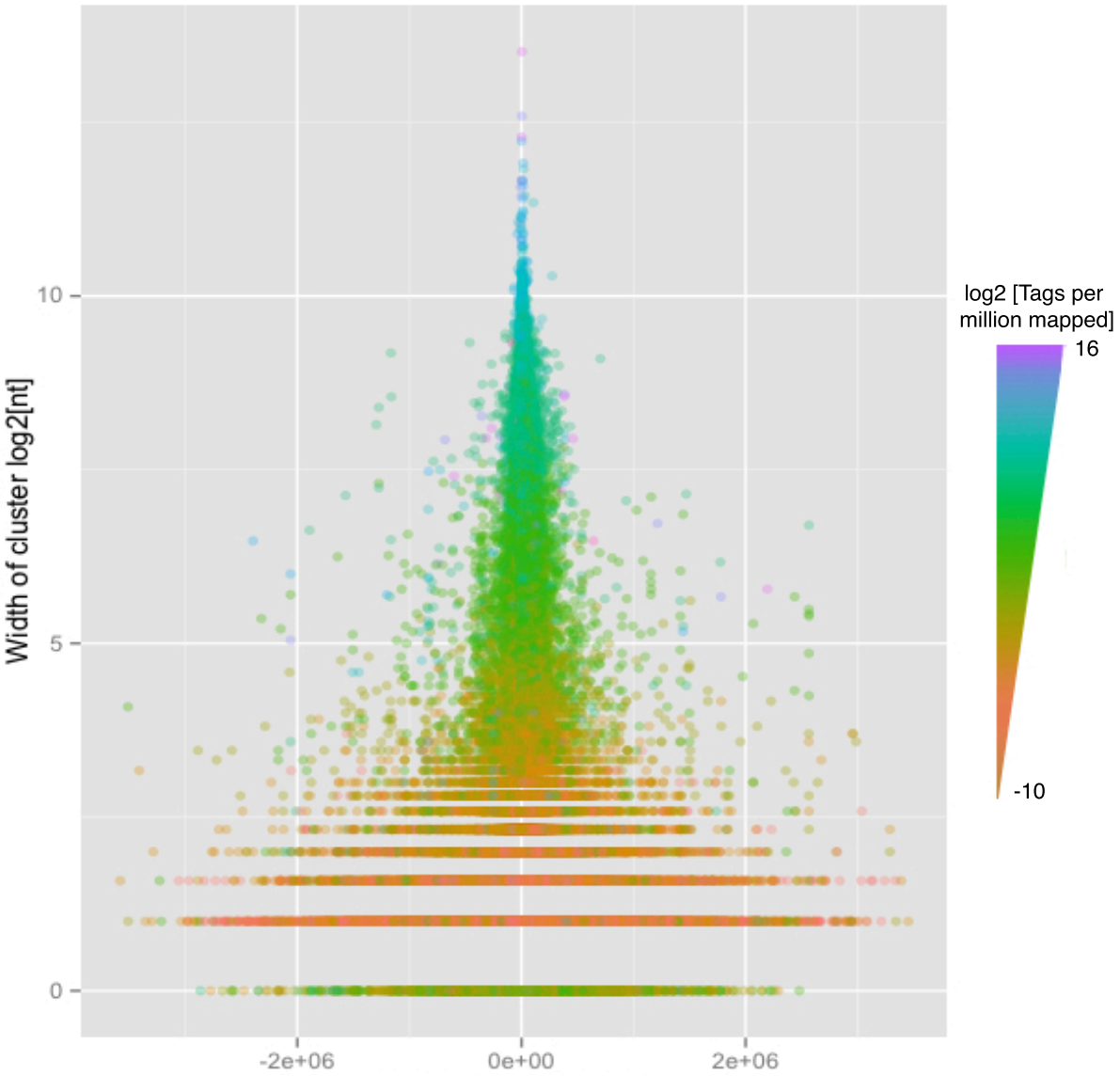


Figure 1. nanoCAGE-seq data requires rigourous filtering. Scatterplot of all clusters before filtering. X-axis: distance to nearest UCSC knownGene transcription start site, y-axis: width of cluster in nt (log2). Clusters are color-coded by expression amount (TPM). As evident, higher expressed clusters are closer to the TSS and wider, while the lowest expressed clusters, much of if noise, are spread across the genome and are slim.

References

Ihaka, Ross, and Robert Gentleman. (1996) R: A language for data analysis and graphics. *Journal of computational and graphical statistics* **5**(3), 299-314.

Langmead B *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3), R25.

Plessy C *et al.* (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature methods* **7**(7), 528-534.

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.

National Nodes

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

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

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

China
Centre of Bioinformatics, Peking University, Beijing

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

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

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network

Greece
Biomedical Research Foundation of the Academy of Athens, Athens

Hungary
Agricultural Biotechnology Center, Godollo

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

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

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan
COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

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

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

Russia
Biocomputing Group, Belozersky Institute, Moscow

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

South Africa
SANBI, University of the Western Cape, Bellville

Spain
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

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

Sweden
Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland
Swiss Institute of Bioinformatics, Lausanne

United Kingdom
The Genome Analysis Centre (TGAC), Norwich

Specialist- and Assoc. Nodes

CASPUR
Rome, Italy

EBI
EBI Embl Outstation, Hinxton, Cambridge, UK

Nile University
Giza, Egypt

ETI
Amsterdam, The Netherlands

IHCP
Institute of Health and Consumer Protection, Ispra. Italy

ILRI/BECA
International Livestock Research Institute, Nairobi, Kenya

MIPS
Muenchen, Germany

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

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

The New South Wales Systems Biology Initiative
Sydney, Australia

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

EMBnet.journal
ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Instructions for authors" at http://journal.EMBnet.org/index.php/EMBnetnews/about and send your manuscript and supplementary files using our on-line submission system at http://journal.EMBnet.org/index.php/EMBnetnews/about/submissions#onlineSubmissions.

Past issues are available as PDF files from the Web site:
http://journal.EMBnet.org/index.php/EMBnetnews/issue/archive

Publisher:
EMBnet Stichting p/a
CMBI Radboud University
Nijmegen Medical Centre
6581 GB Nijmegen
The Netherlands

Email: erik.bongcam@slu.se
Tel: +46-18-67 21 21