ECCB 2014 Accepted Posters with Abstracts

A: Sequencing and sequence analysis for genomics

A01: Pedro Oliveira, Marie Touchon and Eduardo Rocha. Genetic Mobility and the Distribution of Restriction Modification Systems in Prokaryotes

Abstract: Restriction modification (R-M) systems are nearly ubiquitous among Bacteria and Archaea. They are involved in the immune response against foreign mobile genetic elements (MGEs) and used by MGE to infect prokaryotic cells. RMSs typically consist of an endonuclease that cleaves DNA at a particular sequence context, and a cognate methyltransferase that prevents cleavage within the same sequences by methylation. So far, few studies have precisely addressed their distribution particularly in relation to horizontal transfer. Here we used comparative genomics to quantify R-M abundance, distribution, evolution and association with MGEs.

R-M systems were found to be much more abundant in chromosomes than in MGEs. MGEs over-represent solitary methylases, possibly to protect themselves from the host systems. Fused R-M (Type IIC) systems were unexpectedly abundant presumably compacting genetic information and facilitating rapid changes in target specificity. Small sexually isolated genomes have few R-M systems. The remaining genomes have an average of ~2 systems independently of their size, suggesting weak selection for R-M systems beyond this value. Our data also shows that R-M systems are under strong purifying selection, whose intensity was found to correlate with the structural complexity of the system. Type IV REases locate preferentially in the close vicinity of Type I systems, which may be regarded as a strategy for diversification of the resident methyltransferase (m6A). R-M systems were found to be significantly more abundant in genomes in which MGEs and CRISPR-Cas immune systems were present. Finally we provide for the first time quantitative evidence linking natural competence for transformation and increased R-M abundance.

In conclusion, we present an extensive study on R-M systems. Our data suggests a role of structurally compact systems such as Type IIC and IV on R-M evolvability and on antagonizing other R-M Types. Moreover, given their distribution biases and the associations established among them as well as with MGEs and other defense systems, this work highlights their intricate and complex relations.

A02: Vivek Srinivas, Swati Patankar and Santosh Noronha. Translatability can predict functional small open reading frames in intergenic regions of the Plasmodium falciparum genome.

Abstract: Small ORFs (smORFs) are open reading frames of length less than 300 nucleotides (Basrai et al. 1997). They are found throughout the genome and can be classified as intergenic smORFs and upstream ORFs (uORFs). Intergenic smORFs are present between two genes while uORFs lie within the 5' un-translated regions (UTRs) of messenger RNAs. Recent reports have shown translatability of smORFs in yeast (Ghaemmaghami et al. 2003), Drosophila(Ladoukakis et al. 2011) and other organisms.

Human malaria parasite, P. falciparum, has an AT rich (81%) genome (Gardner et al. 2002). Previous reports (Marin et al. 2003), and our comparison of the frequency of smORFs in different Plasmodium species, suggests that the AT content of the genome is directly proportional to the frequency of smORFs, thereby predicting a large number of smORFs in P. falciparum. Further, uORFs have emerged as one of the key players in mediating translational regulation in parasite (Le Roch et al. 2004).

Available gene prediction methods are inaccurate in predicting translatable smORFs (Wang al. 2003) and mass spectrometric fail to detect proteins synthesised by smORFs due to abundance and technical issues (Zhu et al. 2004) (Merrell et al. 2004). This has resulted in poor annotation of large numbers of smORFs (Basrai et al. 1997). Therefore, a specialized predictor was used for identifying protein coding smORFs (CDS smORFs) from the P. falciparum genome - SORF finder (Hanada et al. 2010). It gave an accuracy of 86.3% in predicting CDS smORFs, but 73.5% in predicting NCDS smORFs. In an attempt to increase the accuracy, “ORFpred”, a Machine learning algorithm was developed.

ORFpred was trained on coding (CDS) and non-coding sequences (NCDS, ORFs from non-coding RNA and randomly generated ORFs). The trained predictors were tested on known smORFs. ORFpred gave an accuracy of 88.0% in predicting CDS smORFs and 99.9% in predicting NCDS. Thus, ORFpred performed better than SORF finder in predicting translatable smORFs from the P. falciparum genome. The algorithm uses multiple parameters in training the predictor: these are compositional features (nucleotide, codon, di-codon frequencies) and positional features (ribosomal binding sites known as Kozak sequences in untranslated regions). Thereby, ORFpred measures the probability of translation initiation and elongation of smORFs, giving a detailed mechanism of translational selection of smORFs and uORFs in genome.

ORFpred identified 257,003 smORFs and 22,670 uORFs to be translatable in P. falciparum. Out of these, 221 are conserved, 207 have homologs in other organisms and 8 have functional domains. Finally, reverse transcriptase PCR (RT-PCR) was employed to check whether the predicted translatable smORF genes are expressed as mRNA. RT-PCR identified 3 out of 4 conserved smORFs to be transcribed, giving experimental proof for transcription of some of the predicted translatable smORFs. Further, the role of uORF in mRNAs is also being examined.

A03: Amin Ardeshirdavani, Erika Souche, Luc Dehasbe, Jeroen Van Houdt, Joris Vermeesch and Yves Moreau. NGS logistics: federated analysis of ngs sequence variants across multiple locations

Abstract: Next-Generation Sequencing (NGS) is a key tool in genomics, in particular in research and diagnostics of human disorders. Multiple projects now aim at mapping the human genetic variation on a large scale, such as the 1,000 Genomes Project, the UK 100k Genome Project. Meanwhile with the dramatic decrease of the price and turnaround time, large amounts of human sequencing data have been generated over the past decade. As a result, about 100,000 human genomes have been sequenced so far. Crucially, the speed at which NGS data is produced greatly surpasses Moore’s law and challenges our ability to conveniently store, exchange, and analyze this data. The collection of files resulting from the analysis of a single whole genome study can take up to 50Gb of disk space. This raises significant issues in terms of computing and data storage and transfer, with off-site data transfer currently being a key bottleneck. Moreover, the analysis of NGS data also raises the major challenge of how to reconcile federated analysis of personal genomic data and confidentiality of data to protect privacy. In many situations, the analysis of data from a single study alone will be much less powerful than if it can be correlated with other studies. In particular, when investigating a mutation of interest, it is extremely useful to obtain data about other patients or controls sharing similar mutations. However, confidentiality of this data must be guaranteed at all times and only duly authorized researchers should access such personal data.

To address all challenges described above, we developed a data structure NGS-Logistics, which fulfills all requirements of a successful application that can process data inclusively and comprehensively from multiple sources while guaranteeing privacy and security. NGS-Logistics is a web-based application providing a data structure to analyze NGS data in a distributed way. The data can be located in any data center, anywhere in the world. NGS-Logistics provides an environment in which researchers do not need to worry about the physical location of the data. With respect to users rights, queries will be sent to each remote server. The host will process the request and return the results back to the main server where all the privacy limitations are controlled for the data.
The pilot version of NGS-Logistics has been installed and is currently being beta-tested by users at the Center for Human Genetics of the University of Leuven. Currently we have two installations of the system, the first one at the Leuven University Hospitals and the second one at the Flemish Supercomputing Center (VSC). The development of NGS-Logistics has significantly reduced the effort and time needed to evaluate the significance of mutations from full genome sequencing and exome sequencing, in a safe and confidential environment. This platform provides more opportunities for operators who are interested in expanding their queries and further analysis.

A04: Susan Jones, Linda Milne and Glenn Bryan. Genetic marker discovery in potato: using RNA-Seq for single nucleotide polymorphism discovery in wild Solanum species

Abstract: Potato (Solanum tuberosum) is essential to global food security, with worldwide cultivation >330 million metric tons. The completion of the potato genome sequence provides a baseline upon which to assess genetic variation in tuber bearing Solanum species, of which there are ~150 distributed across South and Central America. Many species contain traits useful for potato breeding. Single nucleotide polymorphisms (SNPs) can be used to design markers that can facilitate dissection of complex traits. The current work presents SNP marker discovery in potato using wild species. The aim of the work is to identify SNPs linked to traits that confer resistance to abiotic stress, such as extreme temperatures. This could lead to the development of new crop varieties required in the face of increasing climate change.
Paired end RNA reads from Illumina sequencing were obtained for 8 wild potato species, including S.Chacoense, a tuber bearing species native to South America that is adapted to grow in a wide variety of habitats. Trimmed reads were mapped to the genome for the double monoploid (DM) cultivar, from the Potato Genome Sequencing Consortium (PGSC_DM_v4.03) using TopHat2. SNPs were identified from the alignments using FreeBayes. For S.chacoense different alignment parameters were used (including miss-matches (N=1,2,3,5,7) and intron length ranges (70-500K and 10-15K) to assess the influence of mapping parameters on SNP density. Alignments were then made for 7 additional wild species, using a consensus set of parameters. A set of "high quality" SNPs were extracted and initial work conducted on mapping these SNPs to specific genes families.
Alignments for S.Chacoense that gave ≥60% of reads mapped to PGSC_DM_v4.03 genome were used to investigate the effects of parameter variation on SNP density. The number of miss-matches gave the most variation: averaging 24% and 48% for N2 (546K SNPs) compared to N3 (708K) and N7 (>1million). Read trimming using varying Phred scores and minimum length constraints with the same value of N gave on average 2.6% variation, and different intron length ranges gave 0.1% variation. For a consensus set of parameters, 203 SNPs were identified per base pair (bp), which is significantly lower than the 1/20-50 per bp previously observed.
This work clearly shows that mapping parameters significantly influence the density of SNPs reported. The parameters used to define "high quality" SNPs is dependent upon the level of variation being measured, and how it will be applied in further work. Initial work on mapping

SNP subsets to specific genes families within potato shows SNPs mapping to genes with very diverse functions. This work is being further developed to identify SNPs linked to traits that allow adaptation to different environments.

A05: Javier Perez-Florido, F.Javier Lopez-Domingo, Antonio Rueda, Joaquin Dopazo and Javier Santoyo-Lopez. ngsCAT: an easy-to-use tool to evaluate the efficiency of targeted enrichment sequencing

Abstract: Targeted enrichment sequencing by Next Generation Sequencing (NGS) experiments is becoming a common way to interrogate specific loci or the whole exome at a relatively low cost. The efficiency and the lack of bias in the enrichment process need to be assessed as a quality control step before performing downstream analysis of the sequence data. We have developed the next-generation sequencing data Capture Assessment Tool (ngsCAT), an application that takes the information of the mapped reads (BAM file) and the coordinates of the targeted regions (BED file) as input and generates a self-explanatory report with metrics, summary tables, figures and plots that allow a comprehensive assessment of the efficiency of the targeted enrichment process in terms of sensitivity, specificity and uniformity. In the case of sensitivity, ngsCAT assesses the quality of the coverage on target regions through (i) the percentage of target bases covered at different coverage thresholds and (ii) a saturation curve of the coverage as a function of the number of reads, which can serve to estimate whether sequencing a higher number of reads will produce a significant increase of the coverage in the regions of interest (ROIs). When assessing the specificity, the tool measures (iii) the number and the percentage of reads on/off target and (iv) the number of duplicated reads on/off target and (v) calculates bedgraph tracks of off-target regions with high coverage. Regarding the uniformity, ngsCAT evaluates sequencing biases in the targeted regions reporting (vi) the distribution of the coverage in the ROIs, (vii) the variability of the coverage within the ROIs and (viii) the distribution of the coverage as a function of GC content. The tool has been used as a quality control for >600 whole-exome runs in our facility in the context of the Medical Genome Project (http://www.medicalgenomeproject.com), allowing us to detect samples not properly hybridized, optimize targeted enrichment protocols and adjust data analysis pipelines. ngsCAT is a Linux command application written in Python. An efficient multi-threaded implementation enables even a full execution for human exome data in a standard computer (Intel Core 2 Duo, 4GB RAM). The tool can also process two samples at a time, which facilitates the comparison of two samples. Documentation, examples and downloads for ngsCAT can be found at http://www.bioinfomgp.org/ngscat

A06: Arnaud Kerhornou, Dan Bolser and Paul Kersey. The polyploid bread wheat genome in Ensembl Plants

Abstract: Bread wheat is a major global cereal grain essential to human nutrition. Its genome is composed of three closely-related and independently maintained genomes that are the result of a series of naturally occurring hybridization events. Ensembl Plants incorporates the chromosome survey sequence (CSS) and gene annotation for Triticum aestivum cv. Chinese Spring, generated by the International Wheat Genome Sequencing Consortium (IWGSC). The IWGSC gene models have been functionally annotated and run through the Ensembl comparative analysis pipelines, generating "gene trees" that show the inferred evolutionary history of each gene family. Gene trees put the bread wheat genes in the context of homologues from its diploid progenitors (Aegilops tauschii and Triticum urartu) and the wider groups of cereals and plants. To accommodate polyploid genomes, we have introduced in Ensembl another class of homologues, homoeologues, to represent homologous genes between the A, B and D component genomes. In addition, a new polyploid view allows users to visualize the homoeologues aligned to each other in the context of their location on the assembly.

A07: Jasmin Schlotthauer, Agnes Hotz-Wagenblatt, Karl-Heinz Glatting, Sabine Weiss and Annette Altmann. Two pipelines for the identification of tumor specific peptides from Next Generation Sequencing data originating from display analysis

Abstract: Specific peptides can help in directing tumor specific drugs or radioactive isotopes to the tumor. The identification of these peptides by phage and ribosome display is done with large libraries presenting scaffold peptides with a variable region of 6-10 amino acids by exposure of the libraries to tumor cells and/or purified target proteins for several rounds (1). Unbound peptides are washed off and will not be multiplied and sequenced in the next rounds. The inserts with the cloned peptides are extracted, multiplied by PCR (fragment length around 100 bases), and sequenced by Illumina HiSeq2000 sequencer with a read length of 100. We created two pipelines for the analysis of the reads. The first one, PeptideDisplayAnalysis, purifies the reads and extracts and clusters the peptide inserts from the available reads according to the PCR primers given. The output is a sorted fasta file with the number of reads in the description for each of the peptide sequences. In the next pipeline, PepEnrich, we check for the enrichment which should occur with the specific inserts by presenting the peptides in several rounds to the targets. The input is a list of PeptideDisplayAnalysis output files together with the round number. The output shows the best enriched peptides together with the read values of all rounds for further analysis. The pipelines are implemented in the pipeline framework and web interface of W2H(2).

References:

1. Zoller F, Markert A, Barthe P, Zhao W, Askoxylakis V, Altmann A, Mier W, Haberkorn U. Combination of Phage Display and Molecular Grafting Generates Highly Specific Tumor-targeting Miniproteins. Angewandte Chemie Int Edition: 2012;51(52):13136-9. doi: 10.1002/anie.201203857

2. P.Ernst, K.-H. Glatting, S.Suhai, A task framework for the Web interface W2H. Bioinformatics: 2003;19(2):278-282.

A08: Laurent Jourdren, Sandrine Perrin, Sophie Lemoine and Stéphane Le Crom. Aozan: an automated post sequencing data processing pipeline

Abstract: Data management and quality control from Illumina HiSeq sequencers is a disk space and time consuming task. Dealing manually with the various input and output file formats from HiSeq softwares [1] requires many hours to generate FASTQ files and QC reports. Moreover this work requires command line knowledge that many biologists do not master. This is the reason why we develop Aozan [2] in order to automatically handle data transfer, read demultiplexing conversion and quality control once a HiSeq run has been finished.

Aozan can handle the output of several sequencers and each step can be dedicated to a computer. The program workflow runs in 3 main steps: data synchronization, read demultiplexing and result quality control. First, Aozan synchronize sequencer outputs with the data storage. Second, Aozan demultiplex data to produce FASTQ files using the Illumina Bcl2fastq tool [3]. To avoid issues with the strict syntax of the Illumina samplesheet, Aozan can use files in Excel file format and can replace indexes alias by their real sequence. This avoids errors due to manual retranscription during the samplesheet creation. To help users, we provide an online samplesheet validator [4] to verify files before Aozan start. Third, Aozan generates an HTML quality control report. It is composed of tests for each lane and sample of the run. Aozan uses many data sources for these tests: sequencer logs (InterOp files), demultiplexing results, FastQC [5] and Fastq Screen [6] output executions.

In addition, we enhanced the bundled version of FastQC in Aozan. First by adding a module that can detect bad tiles caused by the “Bottom Middle Swath” (BMS) problem. Then by improving the Overrepresented Sequences module when 'no hit' source are detected. For that, we launch a blast command on the ”non-redondant” nucleotide database from NCBI and add the best hit to the FastQC report. As for FastQ Screen, it has been reimplemented in Java to speed-up computation.

To conclude, Aozan provides a great help for genomic labs in managing FASTQ creation and accurate quality control report of sequencing runs. This solution dramatically reduces manpower needs for post-sequencing tasks. Once Aozan is configured, biologists can manage all the sequencing process from library preparation to FASTQ generation and QC report without any bioinformatican help.

Acknowledgements: The IBENS high-throughput sequencing facility is a member of the “France Génomique” consortium (ANR-10-INBS-0009).

[1] http://support.illumina.com/sequencing/sequencing_software.ilmn
[2] http://transcriptome.ens.fr/aozan/
[3] https://support.illumina.com/downloads/bcl2fastq_conversion_software_184.ilmn
[4] http://www.transcriptome.ens.fr/designvalidator/
[5] Andrews S: FastQC 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[6] Andrews S: FastQ Screen 2011.
http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen

A09: Lin Wang, Bingding Huang, Agnes Hotz-Wagenblatt, Karl-Heinz Glatting and Christopher Previti. Theoretical and practical coverage achieved by targeted exome enrichment

Abstract: Exome sequencing is widely used in cancer research area due to its relatively high efficiency and low cost. These target-enrichment procedures capture the regions of interest from samples before sequencing. The selection of target regions is crucial to the performance of exome sequencing. Currently, there are different human exome enrichment platforms available provided by different companies. By comparing the target design and performance of these enrichment strategies, we found that important genes are not covered by a particular platforms, both on a theoretical basis as well as in practice.

We compared the exome coverage of three platforms: Agilent's SureSelect (1), Illumina's Nextera Rapid Capture Exome (2) and NimbleGen's SeqCap EZ Exome Library (3). We assessed the coverage of these different platforms in two ways: one is the coverage of coding mutations of genes in cancer census from COSMIC database (4) (release 67), the other is the coverage of all the protein coding genes from the Ensembl database (5) (release 74). On average the platforms covered about 92.51% of the coding mutations of COSMIC and 66.56% of the coding exons of Ensembl.

In doing this analysis we developed a pipeline, called GenePanelCoverage, which calculates the theoretical read coverage of a certain gene or gene list. The input is a list of gene identifiers from the most common databases. Additionally, the target file of the platform has to be specified from a selection list of commonly used platforms. If the platform is not in the list, the user can choose the “User Customized target file” option and upload the target file in bed format. Finally, the user has to specify whether all exons or only coding regions are to be considered in the coverage analysis.

GenePanelCoverage generates a report with detailed information about the coverage of the queried genes. For a particular gene entry, the report contains total length, covered length, percent of the transcript covered and biotype for each of its transcripts. An average of the all

transcripts' coverage is also given as the gene's coverage. Also, GenePanelCoverage will provide the download link of files in bed format containing the query transcripts' positions and coverage analysis of each region. Users can also download the final result in an XML format file for further analysis. GenePanelCoverage was implemented under the W3H pipeline system (6).

References

1. http://www.genomics.agilent.com/
2. http://www.illumina.com/products/nextera-rapid-capture-exome-kits.ilmn
3. http://www.nimblegen.com/seqcapez/
4. Forbes, S.A., et al., COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res, 2011. 39(Database issue): p. D945-50.
5. Flicek, P., et al., Ensembl 2014. Nucleic Acids Res, 2014. 42(Database issue): p. D749-55.
6. Ernst, P., K.H. Glatting, and S. Suhai, A task framework for the web interface W2H. Bioinformatics, 2003. 19(2): p. 278-82.

A10: Kévin Vervier, Jean-Philippe Vert, Maud Tournoud, Jean-Baptiste Veyrieras and Pierre Mahé. Towards Large-scale Machine Learning for Metagenomics Sequence Classification

Abstract: Assigning taxonomic labels to DNA sequences from high throughput sequencing data is one of the main challenges in metagenomics, often referred to as taxonomic binning [3]. Two main computational strategies have been proposed to perform this task: (i) similarity-based approaches, like BLAST [2] and TMAP [1], and (ii) compositional approaches, where a machine learning

model such as a naive Bayes (NB) classifier [7] or a support vector machine (SVM, [5] [6]) is trained to label the sequence based on the set of k-mers it contains. In practice, compositional approaches offer significant gain in terms of classification time over similarity-based approaches. Nevertheless, it seems that similarity-based and compositional approaches achieve comparable performances in terms of classification accuracy [6] [8]. Compositional approaches must be trained on a set of sequences with known taxonomic labels, typically obtained by sampling fragments from reference genomes. For example, [6] sampled approximately 10,000 fragments from 1768 genomes to train a SVM (based on a k-mer representation with k=4,5,6), and reported an accuracy competitive with similarity-based approaches.

Increasing the number of fragments should better capture the genomic diversity of the reference genomes and improve the ability of the SVM to learn in high dimensional feature spaces, which may therefore improve its accuracy. However it also raises computational challenges, as it involves machine learning problems where a model must be trained from potentially millions or billions of training examples, each represented by a vector in 10^9 dimensions for, e.g., k=15. This is out of reach of most standard implementations of SVM. In this work, we investigated the potential of modern, large-scale SVM implementations ([4]) for taxonomic binning. We considered a reference database with 356 complete genome sequences from 52 bacterial species to mimic the expected flora of a Human respiratory sample and simulated test sets of around 130k Roche 454 and IonTorrent PGM reads. We observed that increasing the number of training fragments (up to five millions) and longer k-mers (up to k=15) improved the accuracy of SVM models to values similar to TMAP. In terms of speed, the best SVM was at least 17 times faster than TMAP and took no more than 3 minutes to classify the 130k test sequences on a single core. These first results demonstrate the potential of SVM-based methods with massive training set for sequence classification in metagenomics.

[1] N. Homer et al. https://github.com/iontorrent/TMAP.
[2] D.H Huson et al. Genome research, 2007.

[3] V. Kunin et al. Microbiology and Molecular Biology Reviews, 2008.
[4] J. Langford et al. Technical Report, Yahoo, 2007.
[5] A.C. McHardy et al. Nature methods, 2006.
[6] K.R. Patil et al. PloS one, 2012.
[7] Q. Wang et al. Applied and environmental microbiology, 2007.
[8] D.H. Parks et al. BMC Bioinformatics, 2011.

A11: Aurelia Caputo, Gregory Dubourg and Olivier Croce. A different approach to assemble a whole genome directly from human stool

Abstract: Alterations of gut microbiota composition under antibiotic pressure have been widely studied, revealing a restricted diversity of gut flora associated with colonization by organisms, such as Enterococci, despite inconsistent effects on bacterial load in stool. We study the gut repertoire of a patient being treated with broad-spectrum antibiotics. The impact of antibiotics on the gut microflora was important and a drastic reduction in bacterial diversity was observed. High-level colonization by Akkermansia muciniphila, 84% of the total bacterial population (by pyrosequencing) was found in the sample. It belongs to the Verrucomicrobia phylum and is a Gram-negative mucin-degrading bacterium. The interest to obtain this genome is important, especially to understand this large proportion of this species. However, several attempts to cultivate this microorganism in our laboratory were unsuccessful.

Whole genomes have previously been sequenced directly from samples, such as Chlamydia trachomatis from the vagina, uncultured Termite Group 1 bacteria from protist cells, and Deltaproteobacteria from ocean samples. Here, we propose a different approach for sequencing the genome of Akkermansia muciniphila directly from metagenomics data from the stool sample. We used different technology as Roche 454 and SOLiD (3,844,884 reads). Reads that were generated from Roche 454 were in shotgun (122,354 reads) and paired-end (268,104 reads). We used a mapping method that consisted of aligning metagenomic sequencing reads with the single available reference genome from the Akkermansia muciniphila MucT strain chromosome. This mapping method was possible because Akkermansia muciniphila was overrepresented in the sample data. The sequence data were mapped, using a range of various parameters (percent identity and fraction length of the reads) against the reference genome using CLC workbench software.

Consensus sequence was extract of the previous mapping and do manual curation. Several gaps were closed by PCR.

Ultimately, we obtained a draft genome with only 56 gaps of the Akkermansia muciniphila strain Urmite (CCDQ000000000). This work highlights the potential of metagenomics to provide assembled genomes directly from human stool.

A12: Michał Kierzynka, Wojciech Frohmberg, Piotr Żurkowski and Jacek Błażewicz. A novel approach to the whole genome assembly problem

Abstract: Sequencing has recently become a primary method used by life scientists to investigate biologically relevant problems related to genomics. As modern sequencers can only read very short fragments of the DNA strands, an algorithm is needed to assemble them into the original sequence. When this sequence is not known beforehand, this process is called DNA de novo assembly. There are a couple of methods available that address this problem. However, one of them, based on the overlap-layout-consensus approach, despite its high accuracy, has recently been nearly supplanted from the market due to its time consumption, especially in the context of constantly increasing number of sequences. In response to this, we propose a new algorithm based on this classical approach, but being able to accurately handle large data sets coming from next generation sequencing machines.

We proposed a unique way to construct the DNA graph model by employing the power of alignment-free sequence comparison. The novelty of our solution lies in a special sorting technique that puts similar sequences close to each other without performing the sequence alignment. This phase is very fast and serves as preselection of similar pairs of sequences. Then, an ultra fast exact sequence comparison on GPUs verifies previously selected candidates, resulting in a very accurate results. Both sensitivity and precision parameters of the algorithm are very high: 99% and 97%, respectively. The high performance computations employing both multiple CPUs and GPUs make the method very efficient even for large data sets.

Having the DNA graph, the algorithm goes on to traverse it in a parallel way to obtain so called contigs and scaffolds, i.e. long fragments of reconstructed genome. Again the approach is novel, as resulting contigs are cut in places where the repetitive fragments are detected. Therefore, the results are more accurate compared to other state-of-the-art algorithms. The information about paired-end reads is used to further increase the accuracy of the method. Moreover, the user may visualize the dependencies and possible connections between resulting contigs and scaffolds in a form of a graph which should greatly facilitate any further genome analysis.

The software was tested on a variety of real data coming from modern Illumina sequencing machines, and was proved to deal with them particularly well. Tests show that the accuracy of the algorithm is higher compared to many well-established assemblers like WGS Celera Assembler, SOAPdenovo, Velvet or AS-ASM. As a result, we think that our method for the DNA de novo assembly may revolutionize the world of DNA assemblers.

The software is an academic and non-commercial tool and will be available publicly soon. The poster is meant to present the main algorithm and its high accuracy measured on real data.

A13: Dominik Forster, Lucie Bittner, Slim Karkar, Micah Dunthorn, Sarah Romac, Stéphane Audic, Philippe Lopez, Eric Bapteste and Thorsten Stoeck. Testing ecological theories with sequence similarity networks

Abstract: Next generation sequencing technologies are lifting major limitations to molecular-based ecological studies of eukaryotic microbial diversity, but in silico analyses of the resulting millions of short amplicons remain a major bottleneck for these approaches. We developed a network protocol exploiting key notions from graph theory such as clustering, community detection, path analysis, assortativity and closeness measures which provides an original solution to test central theories of microbial ecology that have been controversially debated for decades. The analyzed dataset comprised novel V4 SSU-rDNA pyrosequencing data of marine ciliates collected at European coastal sampling sites as part of the BioMarKs (www.biomarks.eu) project. To set our analyses in the context of previous studies, we also included publicly available SSU-rDNA Sanger sequence data from cultured ciliates and from previous environmental diversity inventories into our dataset. Even after two centuries of microscopy studies and ca. 20 years of molecular diversity analyses, we could identify novel hidden diversity in the resulting sequence similarity networks. Additionally, this approach revealed patterns that remained elusive in individual phylogenetic studies demonstrating that marine ciliates —just like bacteria, plants and animals—are under strong environmental and geographical selection at intermediate geographical scale. And finally, our graphs showed an inherent conservative bias of targeted metagenomics and metatranscriptomics caused by the use of ‘universal’ primers, which tend to rediscover more of what was already known in terms of genetic diversity. We argue that this fundamental methodological problem can be overcome by exploiting the structure of sequence similarity networks and by designing

specific targeted PCR-primers based on the identification of clusters of most peripheral sequences in these graphs. Generally, our interdisciplinary approach opens new avenues for the understanding of evolutionary processes and provides new types of evidence to test ecological theories of microbial diversity, in particular regarding the endemicity of microbes. This novel application of network analyses is broadly applicable to other molecular diversity studies which seek to better exploit massive sequence datasets.

A14: Mathieu Labernardiere, Patrice Baa-Puyoulet, Jean-Pierre Gauthier, Gérard Febvay, Federica Calevro, Yvan Rahbé, Hubert Charles, Jean-Christophe Simon and Stefano Colella. SymbAphidBase: a new database dedicated to aphid symbionts to store novel sequenced genomes and standardize their annotations.

Abstract: Complete sequences of bacterial genomes are accumulating with an unprecedented speed due to the democratization of NGS technologies. This is also true for symbiotic bacteria and genomic comparisons are key to understand their contribution to host biology. We developed SymbAphidBase [1]: an ad hoc genome database to store and analyse aphid symbionts' genome sequences. Aphids harbour an obligate primary endosymbiont, Buchnera aphidicola, and several facultative secondary symbionts. SymbAphidBase is designed to integrate data from all these bacteria. At present it includes the sequenced genomes of 17 strains of B. aphidicola from 8 different aphid species available in GenBank. To implement this database we used the GMOD's tools: the chado database to store the genomic data and annotations, coupled with the JBrowse genome browser. SymbAphidBase includes an interface that gives access to data in different formats: a genome browser, a Blast server, comparative genes/proteins statistics and downloadable files. From the beginning of the project, the need to generate a unified gene annotation and identification scheme was apparent. In fact, if we were to use the original gene functional annotations and names, often a small fraction of genes would be found to be common in the different B. aphidicola genomes when performing pairwise comparisons (as low as 10%). In light of these results, we decided to re-annotate the genomes using EuGene-PP, a prokaryotic gene finder tool. The genes are later re-annotated (or annotated for the new genome sequences) using a Blastx analysis against the HAMAP [2] protein database that includes 10 highly curated B. aphidicola genomes. The final assignment of gene names is prioritized in a filtered pipeline to include the SwissProt or TrEmbl IDs when available with variable homology criteria (that are registered in the new gene ID). With this approach we are able to increase the number of common genes when performing pairwise comparisons among B. aphidicola genomes (40-99% with our method depending on the chosen parameters). For genes that do not get a name and functional annotation using this automated method, we are working on other approaches that would use phylogeny and/or expert manual annotation. Beyond this novel unified gene annotation, to facilitate the direct comparison of different genomes, we implemented a double browser interface to facilitate the contemporary visualization of two genomes at the same time. All these database generation steps are automated with specific pipelines developed using mainly Perl, PHP, and jQuery languages. In conclusion, SymbAphidBase is a companion database to AphidBase [3] (the aphid genome database) to facilitate genomic data storage and analysis to study symbiosis in the aphid model.

[1] http://symbaphidbase.cycadsys.org/; [2] http://hamap.expasy.org/; [3] http://www.aphidbase.com/aphidbase/

A15: Nik Shazana Nik Mohd Sanusi, Rozana Rosli, Chan Kuang Lim, Low Eng-Ti Leslie, Meilina Ong Abdullah, Rajinder Singh and Ravigadevi Sambanthamurthi. SNPs Discovery from RNA-seq Data of dura, pisifera and tenera Fruit Forms of Oil Palm

Abstract: Single nucleotide polymorphisms (SNPs) are the most common types of genetic variation. Next-generation sequencing (NGS) technology has made SNP discovery affordable even in complex genomes. With the recent completion of the oil palm genome sequence, the characterization of genetic variation present in the oil palm genome is now possible. Here we describe the identification of good quality SNPs in RNA-seq data of three fruit forms of the African oil palm Elaeis guineensis, namely dura (thick-shelled), pisifera (shell-less) and tenera (thin-shelled), a hybrid between dura and pisifera. The identification of SNPs from 454 raw reads involves many processing steps and the application of a diverse set of tools. The high quality reads were mapped to oil palm reference genome using Bowtie2. The mapping files in SAM format were converted to BAM format and sorted by SAMtools. Picard's MarkDuplicate was used to remove PCR duplicates. SNPs and its' corresponding genotypes were called using SAMtools mpileup. Low quality SNPs (<30) or covered by low depth (<5) were filtered out via SAMtools ('vcfutils.pl'). Comparing the reference genome and the RNA-seq data from three different fruit forms of oil palm yielded 53,041 SNPs . This numbers varied from ~9,000 to ~37,000, with tenera showing the highest occurrence of SNPs. A total of 2,155 common SNPs were also identified across all three samples, where 1,181 are homozygous representatives and 974 are heterozygous representatives. When comparing the two fruit forms most distant from the reference genome , dura and tenera, 11,069 SNPs were common to both fruit forms, three fold higher than the common SNPs found in dura and pisifera, as well as common SNPs found in pisifera and tenera. In addition, a total of 8,033 insertions and deletions (InDels) were identified. The filtered VCF files containing the homozygous and heterozygous SNPs were annotated based on their genomic location with the SnpEff software (version 3.5c). Annotation of the called variants demonstrated that less than 11% of the SNPs occurred in coding regions. The annotation results also showed that a total of 3,606 non-synonymous (N) SNPs and 3,676 synonymous (S) SNPs were identified. The ratio of N/S is ~0.981. The 191 SNPs in significant regions (splice sites, start and stop sites variation) were found in 146 genes . This study has demonstrated that high number of genetic variations exist in the three oil palm fruit forms. The annotated SNPs and associated genes affected by it will facilitate various genetic studies in oil palm.

A16: Rozana Rosli, Chan Kuang Lim, Low Eng Ti Leslie, Meilina Ong-Abdullah, Rajinder Singh and Ravigadevi Sambanthamurthi. Comparative genomics in oil palm

Abstract: Comparative genomics analysis to identify genes and their functional annotations is an invaluable tool to get a better understanding of genomes. The availability of the oil palm genome sequence has made it possible to carry out comparative genomics with the sequenced plants. This can help determine the function and evolution of genes, gain an insight into how the species evolved as well as identify sequences unique to the species. In this study, 17,774 E. guineensis gene model sequences were compared with 4 different plant gene models: namely; Arabidopsis thaliana, Phoenix dactylifera, Musa acuminata and Oryza sativa. Several tools, that are publicly available to identify orthologs genes via sequence comparison such as OrthoMCL, InParanoid and Clusters of Orthologous Groups (COG) were utilized. However, identification of orthologous genes between two genomes was performed using InParanoid. This resulted in a total of 80,739 pairwise orthologs groups. A total of 869 orthologs pairs were identified between oil palm and and date, comprising 217 orthologs groups. The results showed that oil palm and date palm were closely related. Multiparanoid was then used to analyze protein relationship between oil palm and 4 other plant. We obtained 244 clusters in 6,189 orthologs genes (3,204 main orthologs) from 3 difference group of tree conflicts. The individual ortholog clusters were aligned using Multiple Alignment using Fast Fourier Transform (MAFFT) in order to produce an accurate multiple sequence alignment of amino acid sequences to identify the conserved regions. A total of 6,060 orthologs genes were

successfully annotated via BLASTP against RefSeq protein database. To determine the true orthologous relationships, phylogenetics analysis was also carried out for several genes.

A17: Electra Tapanari, Julien Lagarde, Javier Santoyo, Laurens Wilming, Jose Manuel Gonzalez, Barbara Uszczynska, Anne-Maud Ferreira, Alexandre Reymond, Roderic Guigo and Jennifer Harrow. Examining tissue specific RACEseq extension of lncRNAs in the Human GENCODE gene set

Abstract: There is a plethora of lncRNA transcripts in the GENCODE Human annotation gene set (around 15,000 transcripts in Gencode version7). Many of these have presumambly incomplete 5' and 3' ends as they have ~15% less CAGE (5' end support) and PET (3' end support) than the protein coding transcripts.(Derrien et al, Genome Res. 2012). In order to extend them to their full length, we have designed an experimental workflow, called RACE-seq that uses 3' and 5' RACE (Rapid Amplification of cDNA ends) followed by 454 sequencing.

In the pilot experiment we designed primers for 400 manually annotated lncRNA loci from Gencode 7 that were lacking CAGE data and GIS-PET support and also had an RPKM > 5 in at least one Human Bodymap tissue. We then ran standard and nested RACE in seven human tissues (brain, testis, heart, liver, kidney, lung, spleen). The RACE products were sequenced using long read RNA-seq (Roche 454 GS FLX+).

The aligned read sequences were provided to the Havana annotation team to manually edit the targeted models.

Overall, the edits resulted in the emergence of novel transcript objects -that had new intron combinations- and to the extension of existing transcripts. Furthermore, a number of the transcript extensions gave rise to their locus expansion. We then investigated the quality of the extensions using standard 5' end support evidence (Fantom5 CAGE data, Encode CAGE, CPG islands) and 3' end support evidence (PET and PolyA seq data) of the curated isoforms and also the protein coding potential of the edited models. We also assessed the differences in the number and quality of extensions between standard and nested RACE and used different alignment methods of 454 reads to curated models to compare tissue specific differences of lncRNA extensions and splicing.

A18: Jérôme Compain, Renaud Jullien, Sivasangari Nandy, Olivier Collin, Jean-Francois Gibrat, Valentin Loux, Véronique Martin and Sophie Schbath. Mapdecode : inventory and benchmark of read mapping tools

Abstract: New sequencing technologies are able to produce enormous amounts of data, up to a billion reads per run. The first step of many analyses of these data, for instance the study of gene differential expression (RNA-seq), the study of gene regulation (ChIP-seq, Methyl-seq), the search for genomic variants (SNPs, chromosomal rearrangements) to name but a few, starts with mapping the reads on the reference genome. Over the last seven years, many mapping methods have been proposed to efficiently cope with the avalanche of data produced by the new generation sequencing (NGS) technologies.

There exist exact algorithms for mapping reads on reference genomes, with or without indels but these algorithms are far too slow to be used with NGS data and large genomes. Therefore, most mapping methods implement heuristics that provide a trade-off between speed and accuracy. This trade-off often leads to the development of complex software with many ad-hoc options whose effects on the mapping results are usually difficult to predict beforehand.

To help users choosing a particular mapping program that best suits their needs a number of benchmarks have been recently published that differs in the mapping tools they consider and

in their methodology for carrying out the benchmarks (criteria for evaluating the mapping results, data used, etc.). In this work,
we extend the work done in Schbath et al., 2012 by focusing, more specifically, on paired reads and extending the test cases (longer reads, consideration of
indels, etc.)
For the purpose of this study, we first compiled an inventory of 93 published mapping tools. However, only 25 of those tools seem actively maintained and have been updated since March 2013.
From this list, we evaluated the performance of seven mapping tools on simulated datasets. We generated 3 different datasets from the human genome. The first one contains 10 millions 40 bp reads, the second one, contains 10 millions 100 bp reads. The last one contains 5 millions 2x100 bp paired reads. From these 3 datasets we then created further datasets with 1, 2 or 3 mismatches per reads. For the long and paired datasets, we also created a dataset with an insertion
of three consecutive random nucleotides and a dataset with a deletion of three consecutive nucleotides. Finally, we generated a more realistic dataset by
using a software that can simulate reads according to errors profiles observed in real datasets (we used bacterial data sequenced with an Illumina Hiseq2000).
All the mapping tools were evaluated for correctness of mapping against these datasets.
The inventory of the mapping tools and the benchmark results for the one tested are available on the Mapdecode website (http://mapdecode.genouest.org).
This work is supported by the "France Génomique" project (ANR-10-INBS-0009)

A19: Aurélien Bernard, Nicolas Guilhot, Fréderic Choulet, Etienne Paux and Philippe Leroy.
The TriAnnot pipeline and its application to the wheat chromosome 3B annotation

Abstract: Today, a combined approach of genetics, genomics and bioinformatics is recognized as a primary driver to make agriculture more sustainable particularly with the climate change underway. This is particularly true for bread wheat a major crop in the world. Therefore, structural and functional annotation of genes in large plant genomes such as wheat is a daunting task but essential to transform draft data into biological knowledge. To achieve a systematic and comprehensive annotation of the giant bread wheat genome (6x=2n=42, AABBDD, ~17Gb/1C), we have developed the automated structural and functional annotation pipeline TriAnnot in the framework of the International Wheat Genomic Sequencing Consortium (http://www.wheatgenome.org) under the umbrella of the Wheat Initiative.
TriAnnot is hosted at INRA URGI in France and has been parallelized on a cluster of 900 cores. The pipeline relies on the use of 15 bioinformatics tools and more than a hundred databanks to predict transposable elements, protein-coding genes and non-coding RNA into genomic sequences, and to design primers for molecular markers such as microsatellites, ISBPs and SNPs. Specific EMBL/GFF output files are produced to be displayed with GBrowse, Artemis and GenomeView to help further manual expertise.
Recently, the first reference sequence of the wheat chromosome 3B has been produced by our group (Choulet et al. Science submitted). TriAnnot has been used to annotate a pseudomolecule of 1,358 scaffolds representing 774.4 Mb (93% of the complete sequence). Overall, 5,326 protein-coding genes and 1,938 pseudogenes were predicted, as well as ~234,000 transposable element-related sequences representing more than 85% of the whole chromosome.
Beside wheat, TriAnnot has been successfully implemented to other plant genomes such as barley, maize, rice and oak and can be adapted easily for the annotation of other genomes.
The pipeline is already available through a friendly web interface

(http://www.clermont.inra.fr/triannot) and the source code and full documentation will be made publicly available on SourceSup (https://sourcesup.renater.fr/projects/triannot/) in the near future.

The global architecture of the latest stable version of TriAnnot as well as the main results of the wheat chromosome 3B annotation will be presented.

A20: Chloé Cabot, Mélissa Mary, Chadi Saad, Alexandre Renaux, Alexis Bertrand, Amandine Velt, Arnaud Lefebvre, Caroline Bérard, Nicolas Vergne and Hélène Dauchel. GC-VC/DGE: a user-friendly web application for Going over Concordance across results from NGS bioinformatics analytic pipelines

Abstract: Next-generation sequencing (NGS) technologies have definitively revolutionized the experimental approaches in genomics and transcriptomics, allowing to address a wide range of questions whatever the biological area and the organism of interest. DNA-sequencing (DNA-seq) and RNA-sequencing (RNA-seq) became the major platforms for variant discovery and differential gene expression (DGE) studies, respectively. However, the different existing NGS technologies combined to the plethora of new algorithms, new statistical methods and the resulting complex bioinformatics analytic pipelines led to the recurrent question of the concordance across final results obtained from different methods [1-4].

Here, we present GC-VC/DGE, the first dedicated web application (in our knowledge) for Going over Concordance across results from the most popular Variant Calling (VC) pipelines and Differential Gene Expression pipelines (DGE). The GC-VC/DGE friendly interface offers an easy way for a non-programming user, (i) to submit up to five files from pipeline outputs in a standard formats: VCF [5] such as provided by BWA-GATK, BWA-SamTools or Bowtie-VarScan VC pipelines or VCF-like Gold Standard from Sanger sequencing including at least for required fields, chromosome number, genomic position, reference and alternative nucleotides and tab-separated value files from DGE pipelines such as provided by EdgeR, DESeq and Cuffdiff including at least three required fields, Gene ID, Log2FoldChange and p-value, (ii) to define analysis parameters (such as the first n genes most significantly differentially expressed or a type I error setting to retain only records whose adjusted p-value is lower than the risk attached), (iii) to accurately analyze the concordance through comprehensive and interactive graphs such as pie chart, Venn diagram and stacked column chart for intersection analysis, dendrogram and similarity heatmap for pipeline distance evaluation, p-value-based rank correlation and Log2FoldChange distributions for DGE concordance, (iv) to friendly retrieve sublists of the common or specific variants (VC) or genes (DGE) of each pipeline by clicking on a selected graph portion. All ouputs are exportable in standard formats. GC-VC/DGE application uses a MVC (Model-View-Controller) model as a design pattern. This flexible and easy-to-maintain architecture warrants further developments according to Variant Calling and RNA-seq technology improvements and scientists needs.

In conclusion, GC-VC/DGE provides researchers a convivial and efficient web tool for comparing results from DNA-seq variant calling and RNA-seq differential analysis pipelines.

1. Kim SY and Speed TP, BMC Bioinformatics 2013, PMID:23758877
2. O'Rawe et al., Genome Med. 2013, PMID:23537139
3. Reeb et al., Front Genet. 2013, PMID:24062766
4. Kvam et al., Am J Bot. 2012, PMID:22268221
5. VCF (Variant Call Format), 1000 genomes http://www.1000genomes.org/wiki/analysis/variant-call-format/

A21: Samia Benamar, Morgan Gaia and Olivier Croce. Genome analysis of a new virophage to highlight the specificity with its host

Abstract: The giant viruses of the Mimiviridae family, infecting the acanthamoeba genus, include 3 groups: group A, group B and group C. Virophages have a functional analogy with bacteriophages, but they only infect giant viruses. Sputnik Virophages have been isolated with Mimiviridae of group A, although experiments, in vitro, show that they can grow in all three groups. In this study, we describe « Zamilon », the first virophage isolated from Mimiviridae group C. We showed that Zamilon is able to multiply with members of groups B and C but not with the ones from group A.

In order to understand this specificity of Zamilon, we have sequenced its whole genome using a MiSeq sequencer (Illumina). A total of 1,165,648 reads were obtained and assembled into a genome size of 17,276 bp with a 29.7% G+C content. The assembling was performed using the MIRA assembler and CLC Genomics Workbench version 4.9 (CLC BIO Aarhus, Denmark). 20 Open Reading Frames (ORFs) were predicted using Prodigal and GeneMarkS. The genome was manually annotated by searching protein homology using BLASTp searches against the non-redundant protein collection in the NCBI database, UniProt, Pfam, COG, and InterPro databases.

Comparisons against other Sputnik virophages proteins showed 15 homologues proteins with Zamilon (ORF4-ORF7, ORF9-ORF18 and ORF20) that share an identity comprised between 40 and 80%. Among the nonhomologous proteins, 2 ORFs (3 and 19) presented similarities with some genes of Megavirus chiliensis (67 % and 50 %) and one (ORF 8) with Moumouvirus monve’s transpoviron (72 %). Phylogenetic trees were constructed, using the Maximum Likelihood algorithm for each ORF of Zamilon and the closest proteins from the BLASTp results. The tree including ORF19 of Zamilon reveals that this protein clusterize perfectly with proteins from the group B and C and not with those from the group A and their associated Sputnik virophages. This single gene thus highlights the infection specificity of Zamilon to its mimiviridae groups B and C.

These results show that, even within a single lineage, virophages are more complex than initially thought and can target specific genotypes within a same virus family.

We hypothesize that the specificity of Zamilon may be due to the ORF 19 of Zamilon and/or the four other proteins (ORF1-3 and ORF8) which are nonhomologous with those of the Sputnik virophages. An ongoing study may help us to find a potential interaction between these proteins and the mimiviridae genome, which could explain the Zamilon specificity. A better comprehension of this capacity can thus be used, to understand the regulation of some pathogenic viral populations.

A22: Paul Bailey, Sarah Ayling, Cristobal Uauy, Ksenia Krasileva, Hans Vasquez-Gross and Jorge Dubcovsky. Development Of An Exome Capture Resource For Functional Genomics In Bread Wheat

Abstract: Recently chromosome survey sequences (CSS) of the 17 Gb hexaploid genome of bread wheat have been assembled by the IWGSC (International Wheat Genome Sequence Consortium). In this work we are using this assembly to develop an exome capture resource for identifying gene mutations in a TILLING (Targeted Induced Local Lesions In Genomes) population of bread wheat for use in functional genomics. Previously we assembled a set of 82,511 wheat protein coding sequences (Krasileva et al, 2013). These sequences were aligned to the CSS contigs to identify exon-intron gene models and the exon regions plus flanking intronic regions were used as the target for the preparation of exome capture bait sequences (NimbleGen). By taking advantage of the fact that the IWGSC assembly is composed of individual assemblies for each flow-sorted chromosome arm, it has been possible to identify the chromosome arm from which each reference transcript is derived. We are now using this exome capture platform to identify EMS-induced mutations in gene homoeologs belonging to the A, B or D subgenomes for up to 1536 TILLING lines. The resulting database of EMS mutations in bread wheat, and the identification of mutations that abolish or modify protein activity will provide a functional genomics resource for both basic and applied wheat research.

We will present a summary for an initial set of capture experiments that were tested for the effectiveness of read mapping and mutation calling to the CSS reference. On average only 2 out of 3 homoeologue copies exist in the current reference. Therefore, reads from unrepresented homoeologs are likely to map incorrectly to the other homoeologs that exist in the reference, particularly in places where the sequences are highly similar. To reduce this problem, we are attempting to find reads in the CSS chromosome arm data sets that correspond to genes with homoeologs still missing in the reference and reassemble them.

Reference:

Krasileva et al (2013) Separating homeologs by phasing in the tetraploid wheat transcriptome Genome Biology 2013, 14:R66

A23: Vincent Walter, Julie Thompson, Olivier Poch and Hoan Nguyen. NeoPipe : A workflow for protein family analysis

Abstract: New advanced technologies including the next-generation DNA, and information technology have significantly improved our capacities of developing biological knowledge and changing our understanding of diseases, phenotype and genotype. In this post-genomic context, protein sequence analysis is a key issue to better understand the evolutionary, structural and functional aspects. NeoPipe is a tool of analyzing a protein family , which consists of 8 steps concerning the search for homologous sequences in multiple databases (protein, 3D structures,…) and functional and structural annotations of clustered multiple alignment of complete sequences (MACS). Those indicate the relationship between the protein subfamilies. The emphasis is to get a high quality alignment by performing refinement and corrections (evaluated by a quality score at each steps) and giving a clustered and annotated alignment of potential subgroups. NeoPipe's application, APIs and REST Web Services are implemented in Java and supported on Linux. NeoPipe is open-source (under the LGPL license) and the source code is available on SourceForge athttp://sourceforge.net/projects/neopipealign/ using a GIT repository. NeoPipe website is developed in Java, JavaScript (JQuery), AJAX with all major browser supported. The website is available at http://lbgi.fr/neopipe/.

A24: Koen Illeghems, Luc De Vuyst and Stefan Weckx. Unravelling ecosystem composition and functional potential of the microbial metagenome involved in the cocoa bean fermentation process

Abstract: Metagenomic sequencing allows for a more detailed characterisation of microbial ecosystems, including those involved in food fermentation processes, compared to previously used techniques. However, results may depend on the analysis software applied. To assess the performance of several software tools and to unravel its ecosystem composition and functional potential, a 750 Mb metagenomic sequence data set obtained by shotgun 454 pyrosequencing of a sample of a cocoa bean box fermentation process taken after 30 h of fermentation was used. Such fermentation processes last typically four to six days, during which a specific consecution of yeasts, lactic acid bacteria, and acetic acid bacteria takes place that is inextricably linked with the usage of substrates and/or the production of metabolites, encompassing glucose, fructose, citric acid, ethanol, lactic acid, acetic acid, and

mannitol.
The ecosystem composition was analysed relying on several similarity-based as well as composition-based computational methods. The results between the two methods differed, mainly regarding the lower abundant microorganisms, due to biases in the public databases used. Indeed, the overrepresentation of genome sequences of mainly (pathogenic) model bacteria and human-associated bacteria in public databases may influence the results for some microbial ecosystems. By considering only operational taxonomic units that were consistently predicted by the different computational methods, both prevailing ones as well as rare microbial species could be identified, providing a more complete and reliable insight into the microbial ecosystem composition of the cocoa bean fermentation process.
As a first step in functional analysis, the metagenomic sequence data set was assembled into metagenomic contigs. Hereto, different assembly programs were used to assess their performance. Overall, Genovo outperformed as it provided the largest number of contigs and bases retrieved, the largest N50 and average contig length, and the highest number of prokaryotic and eukaryotic genes predicted. Next, annotation of the contigs obtained was performed by means of the GenDB annotation platform and the resulting genes were used to reconstruct meta-pathways, i.e., combinations of specific metabolic pathways of multiple microorganisms in an ecosystem. Besides the known functional role of lactic acid bacteria, a possibly important functional role of Enterobacteriaceae was elucidated. Indeed, evidence was provided that these microorganisms possess appropriate metabolic pathways to convert substrates present in the cocoa pulp-bean mass, including the usage of citrate as energy source and as alternative external electron acceptor, and to contribute to pectinolysis.
The present study provided a more detailed characterisation of the cocoa bean fermentation ecosystem and will lead to a better supported selection of appropriate starter cultures to perform controlled cocoa bean fermentation processes.

A25: Gabriel Renaud, Udo Stenzel, Tomislav Maricic, Victor Wiebe and Janet Kelso. deML: Likelihood-based approach for robust demultiplexing of next-generation sequencing data

Abstract: The cost of sequencing multiple samples using the Illumina technology can often be reduced by pooling the samples and sequencing them on a single lane. This multiplexing is typically achieved by using a short DNA index to uniquely identify each sample. Once sequencing is complete, reads must be assigned in silico to their sample of origin, a process referred to as demultiplexing. The software supplied by the vendor to perform this task uses a fixed number of mismatches to identify the correct index sequence. This may fail if the sequence quality of the indices was poor, or when the initial index list was poorly designed. We introduce deML, a maximum likelihood algorithm aimed at robustly demultiplexing NGS sequencing data. Our algorithm computes the likelihood of each sequence read belonging to a given sample. The resulting assignment includes quality values which reflect the probability of the assignment being correct. This approach allows end-users to demultiplex even very problematic datasets.

A26: Anna Ershova, Ivan Rusinov, Anna Karyagina, Sergei Spirin and Andrei Alexeevski. Underrepresented Words in Prokaryotic Genome Shed Light on Lifespan of Restriction-Modification Systems in Genome

Abstract: Restriction-modification (R-M) systems have three activities: interaction with specific DNA sequences (recognition site), methylation of DNA and cleavage of unmethylated DNA. R-M system genes are often localized on the mobile elements of a genome and can move between genomes through horizontal genes transfer. R-M systems can also lose activities and change their specificity. So, we can speak about a lifespan of an R-M

system in a genome. R-M system is a bacterial defense system against phages, but it also has a certain toxicity for bacteria due to occasional errors with the host DNA methylation. It was shown [1] that the presence of an R-M system leads to the avoidance of its recognition site in the genomes of prokaryotes and their phages. We supposed that recognition site avoidance can relate to the lifespan of the corresponding R-M system in a genome. To prove this assumption, a large-scale analysis of R-M system sites avoidance in the currently available genomes was performed. 2141 prokaryotes and 1409 phages were used.

To estimate site avoidance in a genome, the observed number of recognition sites is compared with their expected number. Two methods are frequently used to estimate the expected number of short sequences in a genome: the first one is based on the maximal order Markov model [2], the second one is S. Karlin's method [3]. The former method takes into account the observed frequencies of subwords in a given word obtained by deleting one letter from either 5’-, 3’- or both word ends, whilst the latter, of all the subwords, including discontinuous ones.

We found that the used method has a significant impact on the obtained results because the lists of the most underrepresented sites differ significantly. Our data demonstrate that Karlin's method is more reliable.

Our analysis showed that the sites of R-M systems, encoded in a genome, are underrepresented in half of the cases. At the same time, the number of underrepresented potential R-M systems sites is greater than that can be explained by the encoded R-M systems. According to our data, the sites of the R-M systems recently acquired by a genome are often not avoided, while at least a part of underrepresented sites are the traces of the lost R-M systems.

Our results link occurrences of recognition sites in the genome and the lifespan of the corresponding R-M systems. We thus can suppose that R-M systems with underrepresented recognition sites persist in a genome for a long time. R-M systems with non-underrepresented sites in a genome can be recently acquired by the genome. If a recognition site is underrepresented but there is no R-M system with such a site in the genome, then it is likely, that the corresponding R-M system was recently lost by the genome.

1. Rocha EP, Danchin A, Viari A. Genome Res 2001, 11:946-958.
2. Schbath S, Prum B, de Turckheim E. J Comput Biol. 1995, 2(3):417-437.
3. Karlin S, Cardon LR. Annu Rev Microbiol 1994, 48:619-654.

A27: Rian Pierneef, Oliver Bezuidt and Oleg Reva. Getting insight into ontological relations between mobile genetic elements in bacterial genomes using SeqWord LingvoCom tools

Abstract: Motivation: Horizontally transferred genomic islands have been referred to as important factors which contribute towards the emergences of pathogens and outbreak instances. The development of tools towards the identification of such elements and retracing their distribution patterns will help to understand how such cases arise. Sequence composition has been used to identify genomic islands, infer their phylogeny; and determine their relative times of insertions. We propose several combinatorial parametric optimizations to further enhance the performances of such approaches.

Results: This paper introduces SeqWord Genomic Islands Sniffer (SWGIS) together with LingvoCom tools which utilize composition based approaches to study distribution patterns of genomic islands in prokaryotes. SWGIS uses a set of revised statistical parameters as those introduced in our previously published SeqWord Genome Browser. It is a standalone program that detects genomic islands using a set of combinatorial optimized parametric measures with estimates of acceptable false positive and false negative rates. This study also illustrates the need for parametric optimization towards the prediction of genomic islands, as such has been shown through the use of combinatorial parametric measures and comparing results with those produced by other available prediction methods. The other tools mentioned in the study illustrate the use of DNA composition based phylogenetic analysis, 2-D and 3-D projections, which could be utilized to infer ontological links between different groups of genomic islands in order to trace their donor organisms and be able to distinguish between recent and ancient acquisitions.

Availability: Freely available on the web at http://www.bi.up.ac.za/SeqWord/sniffer/index.html.

Acknowledgement: This work was supported by the National Research Foundation (NRF, South Africa) grant #86941.

A28: Julien Pelé, Matthieu Moreau, Hervé Abdi, Patrice Rodien, Hélène Castel and Marie Chabbert. Evolutionary hubs in protein families: A comparative analysis of sequence covariation methods

Abstract: A wide variety of methods have been developed to analyze covariation between positions in a multiple sequence alignment, in order to gain information on structural and/or functional constraints. However, the sequences of a protein family are not independent but phylogenetically related. The intrinsic inhomogeneity of a sequence set containing different subfamilies, with subfamily specific covarying positions, leads to a phylogenetic bias. This bias is usually considered deleterious, but may also represent a potential source of information about the evolutionary mechanisms of a protein family. In this study (Pelé et al., 2014, PROTEINS, in press), we exploit the phylogenetic bias to identify the correlated mutations that contribute to the sub-family divergence within a protein family. We compare several widely used covariation methods for their adequacy with this aim. The selected methods represent the main four classes of covariation methods (chi2 test, mutual information, substitution matrices, and perturbation of the alignment). They are tested on a model system composed of several sets of G-protein-coupled receptors (GPCRs). In each set, a divergence event is related to specific sequence patterns. We analyze the dependence of the covariation scores on the characteristics of the multiple sequence alignment (residue conservation and number of sequences) and the networking structure of the top pairs. Out of the seven methods investigated, two methods, OMES (Observed minus Expected Squared) and ELSC (Explicit Likelihood of Subset Covariation), are well adapted to find pairs of covarying residues important for the divergence of a protein family. They favor (1) pairs with intermediate entropy on a wide range of set size and (2) residues with high connectivity. The resulting network has a hub structure with a central residue involved in several high scoring pairs. In each case investigated, the central residue corresponds to a residue known to be crucial for the evolution of the receptor family and the subfamily specificity. These data support an epistasis model of family divergence in which new protein functions may arise from the coevolution of several residues.

A29: Anne-Laure Abraham, Mathieu Almeida, Nicolas Pons, Charlie Pauvert, Sophie Schbath and Pierre Renault. A shotgun metagenomic method to characterise low abundant species and assign precisely taxonomy in complex microbial ecosystems

Abstract: A first step for a better understanding of complex microbial ecosystems, such as cheese or human gut microbiota, is the characterisation and quantification of their species composition. Once DNA is extracted from samples, two main techniques are usually used: the sequencing of evolutionary conserved genes, such as those coding for the 16S or 18S RNA or ITS, or whole genome shotgun sequencing. The first approach is widely used and provides a rapid view of the ecosystem, but often fails to provide taxonomy information more precise than the genus level. Shotgun metagenomic sequencing approaches may circumvent such

issues. They are often based on the use of gene marker sets to discriminate species rapidly and without ambiguity, although the use of a small part of the genomes decreases sensitivity of this approach.

We present a new tool that gives encouraging results in the characterisation of low abundance species and may allow distinguishing strains under the subspecies taxonomic assignations. This approach is based on a mapping of metagenomic reads on the whole genomes of a dataset of reference genomes. We then identify species present in the sample based on the number and distribution of reads along each reference genome. Preliminary analyses indicate that we can determine the taxonomy up to the strain level, and identify low abundant species. Our method is complementary to methods based on a set of marker genes. It is computationally more expensive, but can provide a more precise view of the ecosystem. Indeed, methods based on a set of marker genes usually fail to go up to the strain level in the taxonomy identification, and can miss low abundant species if the marker genes are not sequenced at a sufficient depth. We tested our tool on several datasets, including simulated reads and artificial metagenomic dataset in order to test the power and limits of this method. We will present examples on cheese ecosystems.

A30: Heiner Klingenberg and Peter Meinicke. Tools for fast and accurate metatranscriptome analysis

Abstract: While metagenomics can highlight the metabolic potential of microbial communities metatranscriptomics can provide a snapshot of the actual activities. In particular, this includes the possibility to study gene expression in organisms which cannot be cultivated. In that way, metatranscriptomic studies can directly explore the impact of various external influences such as the available C-sources on microbial life. The experimental data usually consists of large amounts of short reads obtained from next generation sequencing techniques. After quality filtering and trimming of the reads all sequences related to rRNA genes need to be removed.

We apply a machine learning-based method [1] to realize a fast rRNA filtering on metatranscriptomic reads. For further analysis the sequencing reads have to be assigned to taxonomic and functional categories, which is computationally expensive for large datasets. With UProC (http://uproc.gobics.de/) a fast protein domain detection tool is available, which for short reads (100 bp) even shows a higher sensitivity than profile-based methods like HMMER or RPS-BLAST. The database of UProC consists of a large dictionary of labeled protein words, which are used to functionally classify the words in a query sequence. Recently we started to include taxonomic information into the database to integrate phylogenetic and functional classification of metatranscriptomic reads. Each word in the dictionary is assigned to a specific taxonomic rank, which can range from species to superkingdom. A query read is then taxonomically classified based on a combination of all matching word labels, using an algorithm similar to the lowest common ancestor scheme as applied to BLAST hits. In addition to the Pfam domain-based database, we also provide a KEGG-based UProC database that directly supports the analysis of a metatranscriptome in terms of the active metabolic pathways.

[1] Heiner Klingenberg, Robin Martinjak, Frank Oliver Glöckner, Rolf Daniel, Thomas Lingner, and Peter Meinicke. Dinucleotide distance histograms for fast detection of rRNA in metatranscriptomic sequences. German Conference on Bioinformatics 2013, volume 34 of OpenAccess Series in Informatics (OASIcs), pages 80–89

A31: Rita Pancsa, Mauricio Macossay-Castillo, Simone Kosol and Peter Tompa. Structural and functional implications of stop codon readthrough in an evolutionary context

Abstract: Translational readthrough (TR) occurs when the translating ribosome fails to stop at the first in-frame termination codon but continues to decode the mRNA, thereby extending the C-terminus of the nascent protein. Recent high-throughput analyses showed that TR is likely to be functional, it is often subject to regulatory control, and is more abundant in eukaryotes than expected. However, the associated advantages are still unclear, because the functions of TR-derived protein extensions have never been investigated. In this study, based on a variety of computational methods we describe the structure-function properties of Drosophila, yeast and human TR candidate proteins and extensions. We found that the long and highly modular fly TR proteins engage mainly in regulatory roles, and their C-termini tend to be structurally disordered. Fly TR-extensions are structurally disordered and rich in binding motifs, which, together with their cell-type- and developmental stage-dependent occurrence, suggest roles in the time- and space-regulated rewiring of cellular interaction networks. In contrast, yeast TR proteins are mainly involved in basic housekeeping functions, like translation and certain biosynthetic pathways. They are rather short and domain rich, with their extensions lacking structural disorder and short linear motifs. In yeast, besides carrying signals for subcellular localization, the extensions could be inhibitory tails of enzymatic domains or have destabilizing effects influencing complex formation properties. Also, the data showed that the extensions of identified orthologous TR proteins are unlikely to bear conserved functionalities and hence probably stem from convergent evolution. Based on the biological processes involving TR proteins, the predicted properties of extensions and the analysis of orthologs, we propose that readthrough serves different purposes in yeast and fruit fly, and suggest that TR-mediated functions are mainly specific to lower taxonomic levels and contribute to species differentiation.

A32: Valentina Boeva, Tatiana Popova, Maxime Lienard, Sebastien Toffoli, Maud Kamal, Christophe Le Tourneau, David Gentien, Nicolas Servant, Pierre Gestraud, Thomas Rio Frio, Philippe Hupé, Emmanuel Barillot and Jean-François Laes. OncoCNV: a multi-factor data normalization method for the detection of copy number aberrations in amplicon sequencing data

Abstract: High-throughput sequencing (HTS), although widely used in biomedical research, is increasingly making its way into clinics. Here, it is helping to identify and designate personalized treatment for cancer patients based on the information on 'actionable' mutations, i.e. mutations influencing cell sensitivity to a particular targeted therapy. Since only a limited number of genes are informative for treatment, expensive genome-wide profiling would not be required. Thus, clinics often employ a less costly amplicon sequencing method that focuses on a selection of actionable genes. However, although amplicon sequencing allows us to reliably detect point mutations, so far we have been unable to assess DNA copy number aberrations, essential for detecting the involvement of some oncogenes. As a result, HTS is often complemented in clinics by other techniques, in particular microarrays (such as array CGH or SNP).

Here, we show how amplicon sequencing data can be effectively exploited in clinics in a way that would mean the use of microarrays would become unnecessary. We present ONCOCNV (http://oncocnv.curie.fr/), the method which is, to the best of our knowledge, the first to allow for the accurate detection of gene copy number aberrations based solely on amplicon sequencing data. ONCOCNV includes a multi-factor normalization and annotation technique enabling the detection of large copy number changes from amplicon sequencing data. We validated our approach on high and low amplicon density datasets and demonstrated that ONCOCNV can achieve a precision comparable to that of array CGH techniques in detecting copy number aberrations.

A34: Sebastien Tempel, David Servillo and Emmanuel Talla. Survey of insertion sequence domestication in prokaryotic genomes

Abstract: Background: Insertion sequences (IS) are small and simple transposable elements that are widely disseminated in bacteria and archea genomes [1]. IS insertions are mainly responsible of the mutations and recombination of prokaryotic genomes and can participate to the creation/modification of host genes: IS domestication.

Results: For our purpose, 1129 selected prokaryotic genomes (covering the 31 main prokaryotic phylum) were downloaded from NCBI. Then, based on IS reference sequences extracted from ISFinder [2], IS sequences were identified using the BLASTn program (default parameters with alignment hits longer than 50p) [3]. As results, we found 3235 distinct IS families from 817 genomes. Among them, 1247 IS families are found in one given genome and 1800 are widespread in less than 10 genomes, mostly in the same prokaryotic phyla. This result clearly suggests that most IS are subject to vertical inheritance and therefore remains specialized in the solely host genome. One hundred and fourty-two IS families are located in two or more distinct phyla (e.g. ISArsp6 in four phyla). Indeed, most of the IS belongs to Proteobacteria and Firmicutes, indicated multiple horizontal transfer events with the two phyla. Interestingly, ISAzs17, the most widespread IS is founded in 55 proteobacterial species and in two strains of Salinibacter ruber (Bacteroides). Genetic context analysis shows all that ISAzs17 occurrences are inserted in coding sequences, mainly represented by 'Arsenate reductase' genes (10 copies), 'Arsenite efflux pump' genes, and 'Arsenical-resistance' genes (25 insertions). Moreover, comparative analysis of the genetic context of these gene families indicates that they are member of similar functional operons. Therefore, these results suggest the ISAzs17 copy have acquired a functional role (IS domestication) in those operons during the evolution.

References:

[1] Siguier P, Gourbeyre E, Chandler M. (2014) Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. doi: 10.1111/1574-6976.12067.

[2] Kichenaradja P, Siguier P, Pérochon J, Chandler M. (2010) ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. Nucleic Acids Res. 38:D62-8.

[3] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. J (1990) Basic local alignment search tool. Mol Biol. 215(3):403-10.

A35: Anaïs Gouin, Anthony Bretaudeau, Claire Lemaitre and Fabrice Legeai. Identification and correction of genome mis-assemblies due to heterozygosity

Abstract: Assembly tools are more and more efficient to reconstruct a genome from next-generation sequencing data but some problems remain. One of them corresponds to mis-assemblies due to heterozygosity. Indeed, the assembly of an heterozygous region for which there is a significant divergence between the two haplotypes, could lead to the construction of two different contigs, instead of one consensus sequence. This problem causes an assembly of an heterozygous genome larger than expected, and also a loss of information (heterozygous SNPs or indels cannot be found in the erroneous regions). We propose a strategy to detect and correct false duplications in assemblies based on several metrics.

We identified two specific cases highlighting problems of heterozygosity. The first case involves scaffolds that are completely matching on another one. The second case corresponds to scaffolds matching together by their extremities. The two sequences involved in the match may actually correspond to two distinct alleles of a specific locus instead of two different locations in the genome. Ideally, an erroneous duplication would involve two divergent but similar assembly parts, not containing any heterozygous polymorphisms, and for which the

merge of the two would lead to the expected read coverage for the resulting consensus sequence. As a consequence, to distinguish between true genomic duplications and alleles, we used various metrics : sequence similarity, length of the match, average read coverage, presence/absence of SNPs in the two concerned regions, number of mate pairs with expected (or not) insert size…

As a result, selected allelic regions are used to construct a single sequence by removal of one of the two alleles or joining of scaffolds by their extremities. This allows to decrease redundancy in the genome assembly, to improve the scaffolding and then to increase the N50 statistic.

We applied this method to a 526Mb highly heterozygous wild type insect genome assembly for which we expected a genome size around 400Mb only. A set of user-validated false duplications in this assembly enabled us to validate the method and to fit the set of criteria, in order to distinguish between true and artefactual duplications.

We took advantage of this study to compare classical assemblers (Minia, Soap) with more recent tools that handle heterozygosity, such as Platanus. This highlighted the advantages of such new assemblers for diploid genomes. However, for already-built assemblies, we showed that our approach is a fast and easy way to discard as much as possible erroneous duplications, allowing their correction without resorting to a complete new assembly that would be more time-consuming.

A36: Amin Zarif Saffari, Marie-Theres Gansauge, Matthias Meyer, Svante Pääbo, Janet Kelso and Kay Prüfer. DNA damage analysis of ancient DNA

Abstract: The DNA in every living organism is under the constant risk of degeneration through environmental or cell-intrinsic processes. Several pathways exist that repair damaged DNA during the cell’s life. However, after the cell's death, these repair pathways cease to work and DNA damage starts to accumulate.

These damages become visible when sequencing ancient DNA: some types of base damages lead to the incorporation of mismatching bases during amplification and lead to errors that cannot be recognized by the base caller during sequencing. Here, we make use of the high coverage genome sequence from a Neandertal and a Denisovan individual to analyze damage patterns by comparing the archaic reads against the archaic consensus sequences. The read data for both archaic genomes were generated by a single stranded library protocol that preserves the orientation of miscoding lesions (i.e. does not show the reverse complement in addition to the original damage substitution). Since DNA extraction, library preparation and specialized computational procedures for ancient DNA may produce false signals of substitutions, we sequenced a modern human sample after identical treatment as a baseline for comparisons.

Our analysis confirms that the most common miscoding lesion in ancient DNA is cytosine deamination, a hydrolytic damage leading to characteristic C→T exchanges that rise towards the ends of molecules. The increased resolution of the dataset reveals a further type of substitution (A→G) following this pattern, that may be explained by a different type of hydrolytic damage affecting adenine. Interestingly, we find two other substitutions, G→T and G→C, that do not follow the pattern of hydrolitic damage. These substitutions can be explained by oxidative damage affecting guanine. Understanding the biochemical processes of ancient DNA damage can help devise new laboratory procedures, is crucial for the computational processing of DNA and can help to identify endogenous molecules. Our results indicate that several additional types of damage are recognizable in ancient DNA sequences.

However, further study is needed to pinpoint the types of biochemical modifications that underly these substitutions.

A37: Kirsley Chennen, Corinne Stoetzel, Jean Muller, Julie Thompson, Hélène Dollfus and Olivier Poch. VarScrut: a tool for deciphering new genes involved in rare diseases in the post-genomic era

Abstract: The molecular diagnosis and study of rare diseases has been revolutionized by recent high-throughput biotechnologies and notably, by next generation sequencing approaches. Today, although many rare diseases present unexpected clinical and molecular heterogeneity, approximately 50% of the roughly 7000 rare monogenic diseases have an identified causative gene. Exome sequencing, which involves the targeted sequencing of 1% of the human genome including all protein-encoding exons, is currently the favorite approach used to identify causal variants and genes. However, the effectiveness of this approach is still limited and only 25% of exome sequencing experiments succeed in identifying a causative variant/gene. This is partly due to the complexity and amount of information to be considered in exome analysis, including: the quality of the sequencing, the various steps for variant identification, the number of patients available, the mode of genetic inheritance, the annotation, filtering and evaluation of variants and candidate gene roles.

In this context, our preliminary studies have demonstrated that no existing tool integrates all the recommendations made by the American College of Medical Genetics for variant analysis, concerning: i) an annotation module, ii) a filtering system for polymorphism frequencies, iii) simultaneous variant evaluation in the context of the 6 modes of inheritance (autosomal recessive, autosomal recessive with inbreeding scenario, X-linked recessive, autosomal dominant, de novo mosaicism), iv) a prioritization module... Consequently, we have developed VarScrut to combine automatic variant exome analysis and simultaneous exploration of all 6 scenarios. In addition, VarScrut integrates rules and optimized parameters for each scenario as well as an annotation module, a filtering module fined-tuned for each scenario and a prioritization module using multi-level knowledge (evolution, function, interactions and metabolic networks). Finally, a literature-tracking module has been included to allow regular automated re-evaluation of candidate genes resulting from unsolved exome analyses.

The current version of VarScrut has been used to analyse exomes from patients with inherited ocular syndromes, notably Bardet-Bield syndrome (BBS), allowing the identification of the BBS18 gene (Scheidecker et al., 2014) as well as other genes currently in the final wet-lab validation phase.

A38: Jose de Vega-Bartol, Leif Skot, Matthew Hegarty, Wayne Powell, Mario Caccamo and Sarah Ayling. Annotation of the the red clover (Trifolium pratense) genome

Abstract: Red clover (Trifolium pratense L.) is a major forage legume cultivated in temperate regions. It is an outcrossing species with a diploid genome (2n = 2x = 14) of approximately 430 Mb. WGS from Illumina paired end and mate pair libraries was integrated with Sanger-based bacterial artificial chromosome (BAC) end sequences and a genetic map. As result, 83.2 % of the genome (354 Mb) was contained in 94,512 scaffolds longer than 1 Kb, including 12,922 scaffolds longer than 10 Kb that comprised 154 Mb. Repetitive and low complexity regions were annotated with RepeatModeler and RepeatMasker, and constituted 51.7 % of the total assembly. A gene annotation of the assembly resulted in 64,629 genes and 66,276 transcripts. The annotation was based on the integration of the genes predicted by Augustus, GeneID and SNAP modelers, supported by alignments of Fabaceae proteins, transcripts from three related model species, and de novo and genome-guided transcripts assembled from RNAseq reads. Alternative splicing based on these assembled transcripts was incorporated with PASA. The following functional annotation evidenced that 94.3 % of the transcripts had a homologous protein in UniprotKB database, and 3.2 % of the remaining transcripts had a novel ORF. After clustering similar transcripts, the red clover transcriptome contained 47,968 unique transcripts. The future publication of the first draft of the red clover genome and its annotation will facilitate breeding improved varieties for animal feed.

A39: Felicia Ng, David Ruau, Evangelia Diamanti, Rebecca Hannah, Berthold Gottgens and Judith Schütte. Features of motif co-occurrence in blood cell development

Abstract: We have previously reported the development of the HAEMCODE repository for curated public ChIP-seq datasets in mouse blood cells (Ruau et al., Nature Methods 2013). To date, almost 400 ChIP-seq samples covering ~130 transcription factors (TFs) across 30 blood cell types are available on the repository and it is the largest collection of TF binding maps in mouse haematopoiesis. This vast collection provides a rich resource not only for genome-wide analysis of gene regulation in blood cells but also for studying important features of TF binding and co-operation. At the beginning of this study, 289 ChIP-seq datasets covering 75 TFs have been processed and so we focused our analysis on this dataset. We searched for enriched de-novo motifs in each of the 289 samples using HOMER and then matched de-novo motifs to known Jaspar motifs using TOMTOM. By integrating all the motif analysis results, we examined the motif co-occurence patterns and distances between motifs. To do this, we developed a motif-pair discovery pipeline to identify motif co-occurrences (within +/- 100bp) and significant preferential distance(s) between motifs in all TF-bound regions. Taking into account distinct pairs, motif orientation and distance values, 7444 significant results were obtained at a q-value threshold of 1e-4. Overall, the distribution of all significant distance values shows a symmetric pattern with preferential distance values of ~+/-1bp and ~+/-11bp. Furthermore, majority of the motif pairs (69.1%) have 1 preferred offset value and very rarely more than 1 offset value was observed for a given motif pair of a specific orientation. Some of the most common types of motif pairs in haematopoiesis include ‘Ets + Homeobox’, ‘Cebp + Homeobox’, ’CTCF + Homeobox’ and ‘Ets + Zinc-coordinating’. We also found partnering between motifs of the same class. For example, the ‘Ets + Ets’ type of motif pairs constitute ~15.1% of all the motif pairs involving an Ets motif while ‘Homeobox + Homeobox’ make up 42% of all Homeobox motif pairs. It is worth noting that Ets motifs can form pairs with all clusters of motifs while, in constrast, CRE motifs form pairs almost exclusively with Ets motifs. To find out if the motif spacing were functionally important, we examined the genetic variation data from the Wellcome Trust Sanger Mouse Genome Project and found that regions containing 1 or more overlapping motif pairs have fewer SNPs and Indels than regions without motif pairs. Moreover, we found that specific types of motif pairs are more prominent in certain cell types than others.

A40: Aslihan Gerhold-Ay, Johanna Mazur and Harald Binder. Optimal Mapping of Methylation and RNA-Seq Data with Prediction Performance as a Measure for Optimality

Abstract: Next-generation sequencing data are becoming more and more important for medical research. Combining RNA-Seq and methylation data still offers open challenges for best prognostic modelling. However,they enable us to develop gene signatures for prediction of clinical endpoints via the integration of the information present in RNA-Seq data on the gene expression leveland methylation data on CpGs. This still has the challenge which CpGs should be considered as being related to one specific gene.

Our goal is to investigate how the prediction performance measure can be used as a measure for optimality to find the mapping of CpGs to their related genes.

For our analysis we define a length of nucleotides around all genes, a so-called window, to find the optimal mapping for methylation to gene informationIn a two-step approach, we first use a likelihood-based componentwise boosting approach to estimate a gene signature only with RNA-Seq data. In the following step, the methylation data of the CpGs that are falling in this window are used to estimate a new signature.

We use different window sizes for the mapping and show the effect on the prediction performance with respect to the clinical endpoint. For finding prognostic signatures, we use RNA-Seq and methylation data of kidney clear cell carcinoma patients from the TCGA ("The Cancer Genome Atlas") platform.

Prognostic gene signatures can be a powerful tool for the classification of cancer patients. To underpin this tool, we propose the prediction performance measure as a criterion to find the optimal mapping window for RNA-Seq and methylation data and show its usefulness.

A41: Patrick Durand, Erwan Drezen, Sébastien Brillet and Dominique Lavenier. KLAST: a new high-performance sequence similarity search tool

Abstract: KLAST is a fast, accurate and NGS scalable bank-to-bank sequence similarity search tool providing significant accelerations of seeds-based heuristic comparison methods, such as the Blast suite. Relying on unique software architecture, KLAST takes full advantage of recent multi-core personal computers without requiring any additional hardware devices. KLAST achieves two major goals: provide a fast and cost-effective general purpose sequence comparison tool and provide high-quality results.

KLAST is a new optimized implementation of the PLAST algorithm(1), to which several improvements have been made. KLAST is fully designed to compare query and subject comprised of large sets of DNA, RNA and protein sequences using KLASTn, KLASTp, KLASTx, tKLASTx and tKLASTn methods. It is significantly faster than original PLAST, while providing comparable sensitivity to BLAST and SSearch algorithms. KLAST contains a fully integrated engine capable of selecting relevant hits with user-defined criteria (E-Value, identity, coverage, alignment length, etc.).

KLAST has been benchmarked on metagenomic data sets from the Tara Oceans International Research Project(2). The main goal of the test was to evaluate speedup and quality of results obtained by KLAST in comparison with BLAST, which is usually used at Genoscope to run sequence comparisons. Quality was evaluated in two ways. First, crude results from both tools were compared, i.e. how much results from BLAST are also found by KLAST. Second, by using results from both tools to assign each query to a taxonomy entry. KLAST achieved sequence comparisons up to 18x times faster than BLAST, while covering up to 96% of the results produced by BLAST. This benchmark illustrates the benefits of using KLAST both in terms of quality results and speed on the Tara Oceans metagenomic data.

To provide users with an advanced sequence similarity search platform, the KLAST engine has been integrated into several tools, from the command-line up to full-featured graphical data analysis platforms such as ngKLAST, KNIME and CLC bio's Genomics Workbench. In all cases, the KLAST system provides an integrated algorithm analysis workflow that includes similarity searches, hits annotations, and data filtering.

(1) V.H. Nguyen & D. Lavenier. BMC Bioinformatics 2009, 10:329

(2) This benchmark has been conducted by the research team of Jean-Marc Aury at the French National Center for Sequencing / Genoscope / CEA.

A42: Ruslan Soldatov, Svetlana Vinogradova and Andrey Mironov. Translation causes global unfolding of mRNA structures in vivo

Abstract: Monitoring RNA structure in vivo has long been impossible due to technical limitations. Recent application of dimethyl sulfate modifications of unpaired adenine and cytosine coupled with deep sequencing (DMS-seq) by Rouskin et al. provides genome-scale profiling of RNA structure in vivo. Genome-wide survey reveals global mRNA unfolding inside cells compared to the in vitro condition. Rouskin et al. hypothesized that action of RNA helicases and other ATP-dependent processes are causing the unfolding. On the other hand, single-molecule experiments reveal complex interplay between single ribosomes and mRNA secondary structure, where the ribosome unwinds RNA structure and structured RNA slows protein synthesis. The ability of a locally disrupted structure to refold depends on the translation intensity, speed and biochemical conditions. When focusing on translation in general rather than individual ribosomes, the question arises whether translation pervasively destabilizes RNA structure. In this work we study the role of translation in mRNA unfolding by considering separately coding regions, untranslated regions (UTRs) and long-noncoding RNAs. The process of translation explains RNA unfolding in human foreskin fibroblasts and is a major force of RNA unfolding in yeast. We also compare in vivo and in vitro structural patterns of protein-coding and long-noncoding genes and show distinction between them.

A43: Stephen Newhouse, Amos Folarin, Hamel Patel and Richard Dobson. eaNGS (Easy Analysis of Next Generation Sequencing): a flexible, easy to use automated NGS pipeline for research and clinical laboratories

Abstract: We present eaNGS (Easy Analysis of Next Generation Sequencing), a flexible and easy to use NGS pipeline for automated alignment, quality control, variant calling and annotation for research and clinical laboratories.

The software pipeline allows users with minimal computational/bioinformatic skills to be able set up and run a NGS pipeline on their own samples in less than an afternoon, as a virtual machine or container on any operating system (Windows, iOS, Linux). eaNGS can be deployed on any medium to high-end workstation, high performance computer cluster and the cloud (public/private cloud computing) - enabling instant access without investment overheads for additional hardware and providing a solution for rapid and efficient deployment of NGS pipelines for clinical and research laboratories of all sizes.

The pipeline is controlled from a single configuration file, produced in excel, that allows the user to upload sample/run related information and control multiple run parameters eg RAM, CPU usage, sample names, project names etc. eaNGS also provides flexibility to use multiple open-source aligners (stampy, bwa, bowtie2) and provide guidance for use with novoalign (commercial) .

We have adapted the current best practices from the Genome Analysis Toolkit (GATK) for processing raw alignments in SAM/BAM format and variant calling. The current workflow, has been optimised for Illumina Platforms, but can easily be adapted for other sequencing platforms, with minimal effort.

We provide a modular workflow that runs from raw fastq file to final report and allow users to easily select from a range of modules: 1) Full Pipeline (fastq to full reports), 2) Alignment, 3) Indel Realignment and Base Recalibration, 4) SNP and small INDEL Variant Calling (single/multi sample), 5) Structural Variant Calling, 6) Alignment Quality Control Reports, 7) Variant Annotation, 8) Variant Reports and 9) Custom Reports (HTML) . The pipeline can be used for population level, family based, single samples and cancer related studies.

The eaNGS pipeline will be made available as virtual machine, along with all scripts and detailed documentation for installation and configuration for alternative sequencing platforms and operating systems.

A44: Léa Siegwald, David Hot, Yves Lemoine, Hélène Touzet and Ségolène Caboche. How can you trust your metagenomic analysis pipeline ?

Abstract: Next generation targeted metagenomic sequencing allows biologists to reveal the microbial diversity of their samples without the limitations caused by culture and phenotypic identification techniques. Unlike whole genome shotgun sequencing, targeting 16S rDNA hypervariable regions does reduce the complexity of datasets and allows a quicker comparison to databanks for bacterial taxonomic identification. With the revolution of high throughput sequencing, there are now a lot of different bioinformatic methods with few good practice guidelines. Choice between softwares is often empirical and habit-driven : non-experts tend to focus on user experience and result display. GUIs and graphical outputs are usually favoured over command line and flat-text files, sometimes at the expense of results, more adequate methods and statistics.

Beyond usage convenience, users may have trouble evaluating the reliability of the softwares, parameters and thresholds they choose. A priori knowledge of their samples is often the only element used to criticize the results. It is also difficult to measure the adequacy of a sequencing technique, to evaluate analysis biases, or to choose a reference databank.

Therefore we propose an evaluation framework to easily assess targeted metagenomics analysis pipelines and their configuration.

Based on the suggestion of Mavromatis et al. (Nature Methods 2007), we profiled three artificial communities with more than a hundred taxa and different biodiversity complexities. For each one, we used read simulators to generate amplicon sequencing reads outputs fitted to a high-throughput sequencing technology, its inherent error model and read length profile. Each taxa has a known number of reads in each community, thus giving us full control on the analyses input. We also varied sequencing coverage and generated replicates to represent varied sequencing strategies. Those informations will free the user from biological uncertainty and biases, and allow him to precisely evaluate the methodology used on different experimental models.

We provide guidelines regarding how the user should perform tests with our datasets. A particular attention has been paid to the evaluation of results and the adjustment of parameters and thresholds. We established a set of suited metrics taking into consideration exactness and reproducibility of results, running time and computer usage, ease of use and output formatting…

This framework has been used to assess our own targeted metagenomic analysis pipeline. It allowed us to identify the improvable stages and to sharpen its configuration. We used real biological samples of known bacterial composition to validate those adjustments. We hope that our framework will allow people to easily test several targeted metagenomics methods, to choose and validate their pipelines, parameters and thresholds, and to polish their experimental and analysis designs.

A45: Alexandre Loywick, Gael Even, Sophie Merlin, Renaud Blervaque and Christophe Audebert. A targeted metagenomic analysis pipeline dedicated to Ion Torrent PGM Data

Abstract: Targeted metagenomics, such as the 16S rRNA-based surveys for microbe identification, consists in sequencing targeted regions to produce a profile of diversity in a sample. The launch of benchtop sequencers, such as the Ion Torrent PGM (Life Technologies), makes the sequencing accessible to a broad community. However, to our knowledge, no tool was specifically developed, neither tested, to analyze targeted metagenomic data produced by the Ion Torrent sequencer. We present a complete pipeline for this task. A particular attention has been paid to the design of the pipeline making it fast, automatic, reproducible and easy to use by non-experts. It has been validated with public

available datasets and with a house-made biological benchmark.

The pipeline is composed of publicly available tools, databanks and Perl/python scripts. It can be divided into three main modules. The first module corresponds to a pre-processing step producing a curated and filtered collection of reads. In the second step, classification and phylogenetic analysis are performed, returning a list of OTUs for each taxonomy level. In the third module, reads are clustered in OTUs and classified on a reference taxonomy, producing a list of OTUs corresponding to given distances. The two-step approach (phylogenetic analysis and clustering analysis) are complementary and allows the users to check the consistency of the classification as discussed by Zemanick(2013). For each steps, summary statistics are available to check the quality of the analysis (e.g. number of reads discarded, diversity score, rarefaction plots, ...). Graphical outputs are provided (Krona charts) and all results can be exported in BIOM standard format to be used for other post-analyses. The whole pipeline was integrated into our Galaxy server, which provides a simple and intuitive interface that allows the user to easily create, run and share analysis for large datasets. Moreover, in order to execute multiple instances of the pipeline (e.g multiple indexes) in a simple toolbox, we linked the pipeline with the Galaxy API bioblend.

All the tools integrated in the pipeline have been chosen by performing tests with public available data provided by Jünemann (2012), simulated data, and with an house-made biological benchmarks. This benchmark contains five species for the two superkingdoms of bacteria, five eukarya and five fungal species with varying known proportions. This benchmark represents a simplified metagenomic sample. It is a complementary approach to the use of simulated data and an original way to test the pipeline compared to the many existing studies; it allows us to take into account biological biases (e.g. library preparation, PCR bias, host contamination). Our results showed that even if Ion Torrent post-light reads show similar characteristics than the reads produced by 454 pyrosequencing reads (Roche), some steps of the analysis like filtering and clustering parameters have to be adapted.

A46: Martina Visnovska, Petra Hlouskova, Terezie Mandakova and Martin Lysak. Fragility of genomic block I among species of the mustard family (Brassicaceae)

Abstract: The mustard family, containing the well-known model species Arabidopsis thaliana as well as several other species with already sequenced genomes, provides a remarkable source of data for plant comparative genomics. The genome of every Brassicaceae species is built up by 24 conserved genomic blocks (A-X) showing a high level of sequence similarity across species. Here we focused on genomic block I, as several previously published comparative cytogenetic and genomic studies differed in definition of its sequence limits.

We have inspected sequenced genomes of seven Brassicaceae species (Schrenkiella parvula, Thellungiella halophila, Arabidopsis lyrata, Capsella rubella, A. thaliana, Camelina sativa, and Brassica rapa) and focused on two candidate regions for the 3' border of genomic block I. In S. parvula, T. halophila, A. lyrata, and C. rubella, natural 3' border of genomic block I is the centromere on the fourth chromosome. As plant centromeric regions are rich in repetitive sequences, a high number of short contigs in these regions are usually difficult to assemble. Therefore, we decided to identify the less continuous candidate region of the four species with the 3' border of genomic block I. We have created syntenic maps between A. thaliana and every other species and searched for regions homologous to the candidate regions. We have found that the first candidate region divides in S. parvula into two contigs, whereas in T. halophila only a fragment of the candidate region was localized on a relatively short contig. The second candidate region falls into a long contig in both species. In A. lyrata and C. rubella, the first candidate overlaps with a long region of unknown bases representing the centromere on the fourth chromosome and the second candidate falls into a well assembled region. Hence, we used the fragmentation pattern as a criterion to decide, that the first

candidate should be the 3' border of genomic block I. We have performed comparative chromosome painting experiments showing that an area between the two candidate regions is located downstream from the centromere and therefore the first candidate region really should be the border of genomic block I. We have also realized that during evolution of C. sativa and B. rapa genomic block I has been broken in three distinct regions on the 3' end. Two of them correspond to the two analyzed candidate regions and the third one lays between them. This observation suggests that several sites downstream of genomic block I are predisposed to chromosome breakage.

We have studied borders of genomic block I in seven Brassicaceae genomes, and summarized information about fragile sites within block I. A detail analysis of sequence features within and around the fragile sites may help to elucidate principles of large-scale genomic rearrangements in Brassicaceae species.

A47: Boris Nagaev and Andrei Alexeevski. NPG-explorer: a new tool for nucleotide pangenome construction and analysis of closely related prokaryotic genomes

Abstract: Genomes of closely related bacteria have highly similar sequences of orthologous fragments but usually undergo multiple rearrangements, long deletions, insertions of mobile elements and occasionally horizontally transferred regions.

We developed a new tool, Nucleotide PanGenome explorer (NPG-explorer), designed for aligning and analysis of a number of input closely related genomes. NPG-explorer constructs nucleotide pangenome - a set of aligned blocks, each block consisting of orthologous fragments. Fragments having no orthologs are considered dummy blocks of one fragment. Each nucleotide from input genomes belongs to exactly one block of NPG (it is a reason for NPG terminology). Minimum length of block (default 100 bp) and minimum identity (default 90%) are algorithm parameters. NPG-explorer iterates block detection algorithm until the following criterion is satisfied: BLAST search all-against-all block consensuses detects no hits of appropriate size and identity.

In addition NPG-explorer provides: (1) Multiple alignments of input chromosomes represented by a sequence of block identifiers. These alignments allow to detect chromosomal rearrangements. (2) File with consensus sequences of all blocks and file with description of all mutations with respect to consensuses. Thus, all input genome sequences can be completely reconstructed from these two files. (3) Phylogenetic trees of core blocks and of whole genomes. Core blocks are those that contain exactly one fragment of each genome. These trees are computed on the base of diagnostic positions in block alignments. (4) All gene annotations, mapped on blocks. This data are useful for detection and correction mis-annotations, gene corruption etc.

NPG-visualization tool presents interactively a list of blocks, the alignment with mapped genes, alignments of block identifiers.

NPG-explorer is written in C++ and is licensed under the GNU GPL. Simple script language for program modules invocation is introduced.

NPG-explorer was applied to 17 genomes of Brucella genus, each genome consists of about 3 Mb. NPG-explorer worked approximately one hour on Intel Core i5 computer. Among 527 detected blocks with two or more fragments there are 270 core blocks, they cover 95% of nucleotides. Average sequence similarity in core blocks is 99.2%. Phylogenetic tree of genomes computed by NPG-explorer by using diagnostic positions is in agreement with published data for 10 Brucella genomes [1]. Detected 25787 point mutations and 2334 deletions of more than two bp describes the evolution of sequences within blocks. The program found large translocation from first to second chromosome in Brucella suis ATCC 23445 and large inversion in chromosome 2 of Brucella abortus, also described earlier [2].

The work was supported by RFBR grants 14-04-01693, 13-07-00969.

[1] Wattam et al., J.Bacteriology, 191:3569-79 (2009)
[2] Tsoktouridis et al., J.Bacteriology, 185:6130-6 (2003)

A48: Gaetan Benoit, Dominique Lavenier, Claire Lemaitre and Guillaume Rizk. Bloocoo, a memory efficient read corrector

Abstract: Next generation sequencing technologies generate a high amount of short DNA sequences, but may contain imperfections. Some applications, such as assembly, yield better results with high quality data, triggering the need for short-read correction software. Most methods proposed in the past do not scale well when dealing with a large dataset, often requiring a very large amount of memory.

Bloocoo is a k-mer-spectrum based read error corrector. It relies on k-mer frequency to discriminate between correct solid k-mer and k-mer containing sequencing errors. This is a traditional approach used by many read correctors. Bloocoo distinguishes itself by requiring an order of magnitude less memory than other state-of-the art correctors, while still providing equivalent correction. This is achieved thanks to a constant-memory k-mer counting algorithm, and a bloom filter to store solid k-mers.

The read correction workflows consists in 3 main steps as follows:

1) k-mer counting.
2) Insertion of solid k-mer in a bloom filter.
3) Multi-stage read correction.

The first stage is the constant-memory k-mer counting step provided by the DSK counter by Rizk et al. It outputs on disk the list of solid k-mer, i.e. k-mers with a high enough abundance.

The second step is the insertion of all solid k-mers in a bloom filter. This probabilistic data structure allows to know if a given k-mer is solid, using only 11 bits per solid k-mer inserted, at the cost of introducing some false positives.

The correction step is a multi-step approach largely similar to the Musket correction algorithm by Liu et al. All k-mers of a read are queried in the bloom filter and classified as solid or non-solid k-mers. Errors are then corrected through multiple iterations of two-sided conservative, one-sided aggressive and voting-based correction algorithm. To neutralize the effect of false positive solid k-mers due to the bloom filter, errors are corrected only if several solid k-mers cover the corrected nucleotide, greatly reducing the risk that all of them are false-positives and induce false correction.

Bloocoo was also designed for fast correction. First, dispatching blocks of reads to several threads is an easy way to parallelize the correction procedure. Secondly, the bloom filter used is a blocked bloom filter, which greatly increases performance thanks to higher cache coherence.

Tests were conducted on simulated datasets with different error rates and genomes against state-of the art competitors. For example, with reads from human chromosome 1 with 1% error rate and 70x coverage, Bloocoo completed correction in 1390 sec using 740 MB of memory, while Musket required 5330 sec and 12190 MB of memory. Quality of correction was 90.28 % recall / 96.93 % precision for Bloocoo and 90.92 % recall / 97.86 % precision for Musket. Bloocoo correction is roughly the same as Musket, while taking 16x less memory and 3.8x less time.

A49: Leo Colmet Daage, Nadia Bessoltane, Virginie Bernard, Eve Lapouble, Nathalie Clement, Angela Bellini, Gaëlle Pierron, Valérie Combaret, Jean Michon, Isabelle Janoueix-Lerosey, Olivier Delattre and Gudrun Schleiermacher. Whole-Genome Sequencing Analysis of Neuroblastoma's Clonal Evolution

Abstract: Neuroblastoma, a clinically heterogeneous pediatric cancer, is characterized by distinct genomic profiles but few recurrent mutations. As neuroblastoma is expected to have high degree of genetic heterogeneity, study of neuroblastoma's clonal evolution with deep coverage whole-genome sequencing of diagnosis and relapse samples will lead to a better understanding of the molecular events associated with relapse.

Whole genome sequencing was performed on trios (constitutional, diagnose and relapse samples) from seven patients using Illumina Hi-seq2500 leading to paired-ends 90x90 for 6 of them and 100x100 for the last one. Expected coverage was higher than 80X for tumor samples and 50X for constitutional samples. Following alignment with BWA (Li et al., Oxford J, 2009 Jul) allowing up to 4% of mismatches, bam files were cleaned up following the Genome Analysis Toolkit (GATK) recommendations (Van der Auwera et al., Current Protocols in Bioinformatics, 2013). Variant calling was performed using GATK, SAMTOOLS and MUTECT (McKenna et al., Genome Res, 2010; Li et al., Oxford J, 2009 Aug; Cibulskis et al., Nature, 2013). As a first step, we focused on SNVs (single nucleotide variants), within and around coding genes, unknown as polymorphisms in the 1000 genomes and Exome Sequencing Project. We focused on tumor specific SNVs. Then we analyzed CNVs (copy number variants) with HMMtools using a window of 2000bp (Gavin et al., Genome Res, 2012). Finally we explored SVs (Structural variants) including deletions, inversions, tandem duplications and translocations using DELLY (Rausch et al., Oxford J, 2012). At least 10 supporting reads were required to predict SVs in tumors. As for SNVs, tumor SVs were filtered with constitutional following Pitkänen analysis (Pitkänen et al., Oncotarget, 2014).

As results for SNV predictions, a median of 17 tumor specific SNVs per sample was predicted (range: 10-236). While a minority of them were shared between diagnosis and relapse (13.4%), 21.9% were specific to diagnosis and 64.7% to relapse. ALK was the only recurrent mutated gene with four patients having a mutation in both tumors and one having a relapse specific mutation. Copy number analysis highlighted an expected chromosome 1p deletion in five patients (Schleiermacher et al., Br J Cancer, 2012), two on relapse and three in both tumors. A chromosome 17 or 17q gain was found in all patients. DELLY predicted a median of 33 tumor specific SVs per sample (range: 9-65). 24.8% of SVs were specific to diagnosis, 38.2% specific to relapse and 37% shared between diagnosis and relapse.

These results confirm the high level of mutation heterogeneity between diagnosis and relapse tumors in term of SNVs, CNVs and SVs. Tumor events are mostly specific and only a few are shared between diagnosis and relapse. As perspective, we plan to combine all our prediction in a more comprehensive report allowing to link SNVs, CNVs and SVs and then better predict their impact.

A50: Guillaume Rizk, Anaïs Gouin, Rayan Chikhi and Claire Lemaitre. MindTheGap : integrated detection and assembly of short and long insertions

Abstract: Structural variants (SV) are large-scale structural changes in the genome, that have been shown to play an important role in evolution and disease. There are several types of SVs, in this work we focus on insertion variants: sequences that are present at one site (position) in the donor genome but are absent from the reference genome at this site.

Such variants are difficult to detect from short read sequencing data, especially when they exceed the paired-end insert size. Many approaches have been proposed to call short insertion variants based on paired-end mapping, such as SOAPindel. However, there remains a lack of practical methods to detect and assemble long variants.

We propose here an original method, called MindTheGap, for the integrated detection and assembly of insertion variants from re-sequencing data. Importantly, it is designed to call insertions of any size, whether they are novel or duplicated, homozygous or heterozygous in the donor genome. The software performs three steps: (1) construction of the de Bruijn graph of the reads, (2) detection of insertion breakpoints on the reference genome (find module) and (3) local assembly of inserted sequences (fill module). Both the detection step and the assembly step rely solely on the constructed graph.

The graph is constructed using the algorithms implemented in the Minia assembler, a memory-efficient data structure enabling high scalability.

Insertion sites are detected by scanning the reference genome and testing membership of reference k-mers in the de Bruijn graph. Homozygous and heterozygous insertions are handled using two different methods: they generate distinctive patterns, respectively a sequence of at most k-1 missing k-mers spanning the insertion site, or two consecutive k-mers with the first being right-branching, and the second left-branching.

The fill module starts from a given insert site represented by flanking kmers (L, R) and performs de novo assembly to attempt to reconstruct the inserted sequence. In a nutshell, a graph of contigs is constructed by performing breadth-first traversal of k-mers, starting from L. Then all paths between L and contigs containing R are returned as putative insertion sequences.

MindTheGap was first evaluated on simulated read datasets of various genome complexities (E. coli, C. elegans and a full human chromosome). This showed that MindTheGap is able to detect with high recall and precision insertions of any size in simple genomes. On a simulated C. elegans dataset of homozygous insertions, MindTheGap and SoapIndel showed similar recall and precision, while needing much less computational resources. Moreover, SoapIndel appeared to be limited to insertions smaller than 200bp, whereas MindTheGap still got 79.5 % recall and 97.3 % precision on simulated 1KB insertions. When applied to real C. elegans and human NA12878 datasets, MindTheGap detected and correctly assembled insertions longer than 1 kb, using at most 14 GB of memory.

A51: Nadia Bessoltane Bentahar, Virginie Bernard and Olivier Delattre. Pubmed Based Gene Annotation Mining, tool helps clinicians and biologists to highlight candidate genes of interest

Abstract: Accurate annotations of genomic sequence features are required to effectively perform causal gene investigations as for Next Generation Sequencing (NGS) analysis or microarray analysis. However not only expert curators assigned function to genes, which would be the less error-prone approach. Often automatic approaches are used to annotate the exponential flow of genomic data continuously submitted in the public databases. This makes database search for curated annotations challenging and time consuming. Several software analyse the functional annotations of genes, as the well known DAVID website suite (Huang DW et al., 2009). Nevertheless, to our knowledge, none does perform any filtering on annotation quality. Therefore the need to provide a support to clinicians and biologists aiming to provide accurate gene annotations is becoming imperative.

Pubmed Based Gene Annotation Mining (PuBGAM) is a versatile tool designed to explore public databases, regularly updated as KEGG (Tanabe M. et al., 2012), GeneOntology (Gene Ontology Consortium, 2013), DrugBank (Kanehisa M et al., 2008), and PubMed (http://www.ncbi.nlm.nih.gov/pubmed). It selects accurate gene annotations characterized by pubmed publication evidence, and then helps biologists to highlight genes making sens relative to their projects.

PuBGAM is command-line driven and easy to integrate into any pipeline analysis involving genes. A user friendly tabulated output file provides links to database annotations and pubmed references filtered in. Not only designed to annotate human genes, PuBGAM was deployed to other species including mouse and yeast.

At the Curie Institute, PuBGAM has been used in various NGS projects. Exome as RNA sequencing projects lead a large list of genes to investigate aiming to prioritize the one(s)

likely to be of interest. While it takes time digging on databases, with PuBGAM it is achieve in a few minutes, guarantying a reliable and useful gene annotations such as functions, pathways, diseases, drugs and MeSH terms significantly related to genes (http://www.nlm.nih.gov/mesh/meshrels.html).

In order to extend PuBGAM benefits outside of our institute, we plan to make it available. PuBGAM is a promising tool helping scientists to rank genes with respect to their putative causal impact and then prioritize genes of interest.

A52: Sven Schuierer and Guglielmo Roma. The exon quantification pipeline (EQP)

Abstract: In recent years RNA-seq has become a widely used approach for expression profiling studies. Besides determining differentially expressed genes RNA-seq also offers the possibility to investigate differential splicing of genes. One common and robust approach to differential splicing analysis is based on the analysis of exon counts (see, for instance, DEXSeq^1).

Here, we present the exon quantification pipeline (EQP) – a novel pipeline to quantify genes, exons, and exon-exon junctions from RNA-seq reads. EQP starts from the Fastq files of an experiment and provides its own alignment module.

The alignment module is based on computing un-spliced alignments against the genome, transcript, and a custom junction Fasta file from either RefSeq or Ensembl. For quantification the three alignment files are first combined – where an effort is made to select the best alignment for each fragment. Based on the combined alignment file EQP generates counts for genes, exons, and junctions where a fragment is only counted for a feature (gene, exon, or junction) if it is consistent with its exon splicing pattern.

EQP also computes the number of genomic alignments nf of a fragment and uses it to weight the fragment with the factor 1/nf in the generation of the counts. In addition to moderating the effects of multi-reads this approach also allows for a flexible way of filtering out fragments with too many genomic alignments; for instance, if the fragment weight threshold is set to one, then only uniquely mapping fragments are counted. Finally, EQP also allows excluding fragments that map to multiple features (as is the standard in htSeq^2).

EQP can generate a file with genomic alignments which can be loaded into a genome browser from the combined alignment file. On the other hand, the alignment module of EQP can be skipped altogether and an alignment file of a splice-aware aligner can be used as input to EQP to generate the gene, exon, and junction counts.

EQP has been developed for a cluster environment (based on SGE) and provides a mechanism to split the input files into chunks of a specifiable size (by default 25 M reads) on which the alignment and quantification are then performed.

Finally, EQP ensures full reproducibility of all results by creating a complete copy of all reference files as well as scripts and executables in a separate, self-contained directory; all computations only use the scripts and files in this directory.

1- S Anders, A Reyes, W Huber, Detecting differential usage of exons from RNA-seq data. Genome Res 22 (2012)

2- S Anders, P Pyl, W Huber, HTSeq — A Python framework to work with high-throughput sequencing data, bioRxiv preprint (2014), doi: 10.1101/002824

A53: Claire Kuchly, Gerald Salin, Gaelle Vilchez, Jerome Mariette and Frederic Escudie. NGS Goes Automatic : From library preparation to quality control of data

Abstract: With the standardization of Next Generation Sequencers, such as Illumina HiSeq2000 and MiSeq, for genomics, transcriptomics and epigenetics and their growing output, the challenge is to sequence at the same time, in a single run, more and more samples of any kind. If we want to benefit from this very high throughput and the many related applications, we have to achieve a high level of automation concerning library preparation and data quality analysis. To get rid of these time-consuming tasks, we developed, at the INRA GeT-PlaGe facility, an automated process to produce more libraries using 3 TECAN liquid handling robots and worked with Genotoul bioinformatic platform to automate the data quality control step. Currently, several protocols are completely or partially automated allowing library preparation and specific data quality analysis for Whole Genome sequencing, Rad-seq, Amplicon sequencing (e.g. 16S sequencing on MiSeq in metagenomic studies), stranded RNA seq, Mate-Pair and Whole Genome Bisulfite Sequencing.

A54: Kevin Lebrigand. De Novo Genome Assembly of Drechmeria Coniospora using ION Torrent, SOLiD and Optical Mapping data.

Abstract: We present the first de novo genome assembly of Drechmeria coniospora, fungus which is the most common and best documented endoparasitic nematode parasite. In this study we describe a novel de novo assembly strategy using Ion Torrent reads for initial assembly followed by a two-steps scaffolding process using SOLiD Mate-paired reads then Optical Mapping data integration for chromosomal organization. Then we have performed the functional annotations of this genome and compared it with close relative fungus species to determine special Drechmeria coniospora characteristics for nematode parasitism.

A55: Katarina Matthes and Mark D. Robinson. A comparison of count-based and assembly-based methods for differential splice detection

Abstract: Detection of differential isoform usage between experimental conditions, such as control versus treatment or disease versus non-diseased, is of significant biomedical relevance, especially given that splicing patterns are aberrant in many diseases. In particular, knowledge of pathological alternative splicing may allow the development of new treatments or better management of patients. To date, several methods to detect changes in isoform usage from RNA-Seq data have been proposed. Popular methods include Cuffdiff2, which calculates transcript-level differences, MISO, which compares percent-spliced-in values using “event” counts and DEXSeq, which compares exon-level counts. Using simulations for both human and fruit fly to span a range of transcript complexity, we assess the detection performance for these methods as well as some new variations, such as voom-diffSplice and alternative methods for counting. In particular, we randomly introduced isoform preferences between two transcripts in a subset of the genes. We investigated the true positive rate and the ability of methods to control the false discovery rate (FDR). Interestingly, for count-based methods, event counting performed best in terms of their true positive rate while also controlling the FDR. Simulation results are split by expression level, number of isoforms, and complexity of known isoforms highlighting that. Not surprisingly, all methods exhibit better performance for higher expressed genes and there are tradeoffs in performance depending on the complexity of the gene structure. Cuffdiff2 seems to be somewhat conservative and had difficulties to detect differentially spliced genes that contain a large number of isoforms.

A56: Susete Alves Carvalho, Raluca Uricaru, Jorge Duarte, Claire Lemaitre, Nathalie Rivière, Gilles Boutet, Alain Baranger and Pierre Peterlongo. Reference-free high-throughput SNP detection in pea: an example of discoSnp usage for a non-model complex genome.

Abstract: Detecting Single Nucleotide Polymorphisms (SNPs) between genomes is becoming a routine task with Next Generation Sequencing. Generally, SNP detection methods use a reference genome. As non-model organisms are increasingly investigated, the need for reference-free methods has been amplified. Most of the existing reference-free methods have fundamental limitations: they can only call SNPs between exactly two datasets, and/or they require a prohibitive amount of computational resources. In this work we used the discoSnp method (Uricaru et al. in prep.) which detects isolated SNPs from any number of read datasets, without a reference genome or data assembly, and with very low memory and time footprints (billions of reads can be analyzed with a standard desktop computer). The discoSnp method is based on the de Bruijn Graph (without needing its explicit representation) in which motifs witnessing the presence of SNPs are detected and analyzed.

We compared results generated by discoSnp with those obtained with a previous SNP discovery pipeline developed by Biogemma, on the sequence reads of eight pea cDNA normalized libraries (Duarte et al. 2014). We launched discoSnp with a "kmer" size of 27 (which specifies the minimal number of bases without error before and after the SNP), and filtered the results according to various parameters such as coverage rates (min. 2x) and number of missing data (max. 5), etc.

Thus, 40k SNPs were identified with discoSnp compared to 35k with the Biogemma pipeline, 18k SNPs being common to both data sets. From the 35k SNP developed with the Biogemma pipeline, a 1920 SNP subset was genotyped using the Illumina Golden Gate assay to generate a high density composite genetic map including 1340 newly developed SNPs, anchored to the M. truncatula physical map. Close to 74% of the genotyped SNP subset and 77% of the mapped SNP subset were called by discoSnp, which seems to preferentially eliminate "false positives" or "unusable" SNPs in genotyping and mapping. The combination of both methods therefore improved the quality and genotyping ability of selected SNPs. The reason why SNPs specifically generated by either method were lost in the genotyping process may be due to low coverage of the sequencing data or/and to the large "kmer" size, when using discoSnp, which is not able to detect SNPs that are close to other polymorphisms.

The high quality of results generated by discoSnp, together with its simplicity and its low memory and time requirements led us to choose this software for a SNP discovery and direct Genotyping By Sequencing project on a set of 48 pea genomic DNA libraries from a recombinant inbred lines subpopulation sequenced with Illumina HiSeq2000 technology. The analysis enabled to identify 88,851 SNP polymorphs on this population, from which around 60k SNPs will be genetically mapped (Boutet et al. 2014).

Duarte et al. (2014). BMC Genomics 15:126

Boutet et al. (2014). IFLRC VI 2014; Canada; Oral Communication

A57: Bettina Halwachs, Henrik R. Nilsson, Kessy Abarenkov and Gerhard G. Thallinger. Integration and validation of resources for high-throughput classification of fungal communities

Abstract: Fungi represent the second largest kingdom of eukaryotic life and hundreds of fungal species are regularly involved in human and animal mycoses. From the estimated 1.5 million fungal species only about 100,000 have been described worldwide with about 1,200 new fungal species per year over the last decades. One reason for this slow progress is that the large proportion of fungi cannot be kept in culture and thus cannot be examined by traditional culture dependent methods; other fungi do not seem to produce tangible fruiting bodies, such that morphological examination becomes hard or impossible. The advent and rapid development of sequencing-based classification methods give rise to new possibilities in fungal identification and description. Although databases and methods for molecular based classification and analysis of archaea and bacteria, based on their ribosomal RNA (rRNA), are already well established, these resources and methods are to some extent lagging for fungi. Here we introduce an effort to integrate existing sequence collections (the UNITE system for DNA-based fungal species circumscriptions) and classification approaches (the Ribosomal Database Project (RDP) classifier trained on internal transcribed spacer (ITS) sequences) for

fungi, into the Straightforward Novel Webinterface for Microbiome Analysis (SnoWMAn). To validate the performance as well as the scope of these resources, an artificial ITS1 mock community was created by in silico amplification of rRNA sequences manually selected and extracted from GenBank. The mock community was then used to evaluate the newly integrated classification resources for fungal communities. Comparison of the classification results with the true community composition showed that although there are still incomplete or ambiguous linage annotations in the reference databases, relative community composition can be determined properly down to order level. For lower taxonomic levels such as family and genus, the fraction of unclassified sequences reaches nearly 50 % using unsupervised classification approach of RDP at a classification confidence of 80 %.

A58: Sergei Lebedev and Oleg Shpynov. A switching hidden markov model for bisulfite sequencing

Abstract: Bisulfite sequencing is a widely used experimental protocol for obtaining base-resolution genome-wide methylation data. Due to various sources of noise (bisulfite conversion errors, sequencing errors, genetic variation) the reads coming from bisulfite sequencing usually contain errors. This calls for statistical models, which will enable accurate characterization of genome-wide DNA methylation from bisulfite sequencing data.

A common representation for the output of bisulfite sequencing is a vector of (k, n) pairs, where k is the number of reads voting for methylation of some cytosine and n is read coverage in this position. We extend this representation by adding genomic distances between cytosines. This allows us to capture spatial dependence between methylation status of the consequent cytosines.

We propose a two-state binomial switching hidden markov model to perform methylation status calling from bisulfite sequencing data. The two states of the model correspond to cytosine methylation status: methylated or unmethylated.

A switching HMM is a generalization of the classical HMM, which conditions transition probability on the distance between the consequent observations.

The use of genomic distances directly results in an impractically large number of parameters in a model. To reduce the number of parameters we cluster similar distances together with head/tails algorithm [1] and then use a single state transition probability matrix for each distance cluster.

We apply our model to the bisulfite sequencing data of mouse embryonic stem cells [2]. Estimated transition probabilities show that the methylation status of consequent cytosines is affected by the genomic distance between them. Namely, nearby cytosines are more likely to be in the same state than distant cytosines.

To assess the relative performance of our model we compare it with the recently introduced FMSC procedure [3], which is based on a two-state binomial mixture. The procedure permits controlling FDR in the obtained methylation status calls with the null hypothesis being: cytosine is methylated. With FDR controlled at level 0.01 our model discerns all the unmethylated cytosines detected by FMSC as well as a significant proportion of novel unmethylated cytosines.

References

[1]: Jiang, B. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. Prof. Geogr. 1–13 (2013).

[2]: Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490–5 (2011).

[3]: Cheng, L. & Zhu, Y. A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data. Bioinformatics 30, 172–9 (2014).

A59: Pierre-Marie Chiaroni, Denis Thieffry and Morgane Thomas-Chollier. Prediction of transcription factor motifs and binding sites from multiple histone mark ChIP-seq datasets

Abstract: The identification and characterisation of cis-regulatory elements is crucial to understand gene regulation. Thanks to novel sequencing technologies, genome-wide datasets for transcription factor (TF) binding, and epigenomic marks are quickly accumulating, fostering the development of novel computational approaches to extract meaningful regulatory interactions. ChIP-seq datasets targeting specific histone modifications are particularly interesting for the computational identification of TF binding sites (TFBS), as several histone modifications (H3K27ac, H3K4me1) are clearly associated with cis-regulatory elements. However, standard motif discovery or pattern matching approaches applied to chromatin mark ChIP-seq peaks proved unreliable, mostly because these peaks can be very large, thus decreasing the signal-to-noise ratio.

To narrow the search space, we have implemented a pipeline exploiting related ChIP-seq datasets targeting different histone modifications to segment the genome according to epigenomic status, using ChromHMM (http://compbio.mit.edu/ChromHMM/). This approach relies upon a Hidden Markov Model, where hidden states correspond to chromatin states, with emissions denoting the presence or the absence of histone modifications. The genome is consequently segmented into regions annotated with a limited number of different chromatin states. Next, we developed a method to compute the enrichment/depletion of each state in a set of loci of interest (peaks from ChIP-seq targeting transcription factors, or exons). A p-value is associated with each enrichment score based on random state assignment for the same segments. Sequences with enhancer-like state(s) were then analysed with two complementary approaches: (i) discovering de novo motifs in the enhancer-like sequences, (ii) scanning these sequences with known motifs to predict TFBS, using tools included in the RSAT suite (http://rsat.eu).

We have applied this pipeline on ENCODE datasets obtained with two different cell lines: lymphoblastoid cells (ChIP-seq for 8 histone modifications) and normal human osteoblasts (ChIP-seq for 5 histone modifications). In both cases, we have identified an enhancer-like state that was significantly enriched in ChIP-seq peaks for over 30 TFs from ENCODE datasets.

Applying the motif discovery approach allowed us to identify over-represented motifs (position-weight matrices) similar to those of HIF1A:ARNT, GATA1 and FOXO1.

We further evaluated the motif scanning approach to identify putative c-Myc and RUNX2 binding sites. Up to 15% of predicted sites overlap the peaks from ChIP-seq experiments targeting each transcription factor, a 30 times increase compared to control. Up to 60% of the ChIP-seq peaks for a TF were recovered using this method.

In conclusion, this generic approach exploits several histone modification tracks to successfully identify and characterize cis-regulatory elements.

A60: Svetlana Vinogradova and Andrew Mironov. New structure-based RNA alignment method

Abstract: Alignments of RNA have a very wide range of applications but aligning structural or non-coding RNAs (ncRNAs) remains a notoriously difficult task: RNA sequences may evolve by compensatory mutations, which maintain base pairing but destroy sequence homology. When aligning ncRNA, alignment programs would take RNA structure into account. Recent studies, mainly based on approximations of the Sankoff algorithm, have resulted in considerable improvement in the accuracy of pairwise structural alignment. At the same time there are some faster algorithms that work quite well on high identities. However, when sequence identity falls below 60-70%, called the “twilight zone” of RNA alignments,

the accuracies of most fast sequence alignment methods drop considerably. More accurate algorithms perform better, but are extremely slow.

We present a fast method for structural alignment of ncRNA based on scoring function that takes into account sequence similarity and normalized up- and downstream pairing probability. To estimate the algorithm precision and performance, we scored alignments produced by our method against a large set of published reference alignments. We show that in the “twilight zone” our algorithm performance is comparable to the performance of slow, Sankoff-style algorithms and better than the performance of fast algorithms.

A61: Adam Clooney, Marcus Claesson, Roy Sleator and Aisling O'Driscoll. High-resolution of microbiota, inflammation and diet in Inflammatory Bowel Disease using parallelised big data processing and analytics

Abstract: Crohn’s disease and ulcerative colitis are inflammatory bowel diseases characterised by chronic inflammation of the gastrointestinal tract. These diseases cause lifelong suffering and considerable consumption of national and personal health care resources. There are over 1.4 millions sufferers in the U.S. costing an approximate $1.7 billion each year. There are 15,000 in Ireland and there is evidence of a substantial and sustained increase.

Sufferers go through periods of flare and remission without any real indication as to the cause. Symptoms of a flare vary from diarrhea and abdominal pain to more serious events such as toxic mega-colon and rupture of the bowel. Although the aetiology is unclear, accumulating evidence suggests a genetic predisposition and a significant microbial factor.

Initial studies investigating the links between the gut microbiota and Inflammatory Bowel Disease have often been limited by sample size and a lack of metadata. This study explores the microbiota composition of stool samples from 360 Canadian and 80 Irish IBD patients, along with 60 healthy control subjects. The microbial faecal DNA of these patients will be sequenced over 3 time points. Time points will enable the investigation of microbiota stability over time and health states. To date, studies have not investigated the change in the microbiota over time which has been a major downfall. The study will take with inflammatory markers, symptom surveys, dietary and drug information into consideration also. This aims to find microbial biomarkers that could predict relapse and remission of symptoms.

To assess to link between the gut microbiota and Inflammatory Bowel Disease effectively and accurately, it is important to examine external factors which may affect results. Due to remote areas in the west of Ireland, along with a possible lack of patient compliance, it can be difficult to ensure the time-period at which a faecal sample is at room temperature before reaching the laboratory for analysis. It is therefore important to investigate if the length of time faecal samples are exposed to room temperature has an effect on the microbial composition. Faecal samples from 4 healthy subjects were stored at various time points (1-7 days) at room temperature. The DNA was extracted and 16S amplification was performed through PCR. The samples were sequenced via the Roche 454 technology. The resulting reads were quality filtered based on length and chimers were removed via the QIIME suite of tools. Samples were then clustered into Operational Taxonomic Units (OTUs) based on their sequence similarity (97%). Examining the microbial composition of each sample showed that the time at which a faecal sample is at room temperature (up to 7 days) does not have an effect on bacterial abundance. These results allow for an easier approach to sample collection thus enabling the accusation of a larger sample size.

A62: Evgenii Kurbatckii. Detecting differential histone modification sites from ChIP-seq data via ranking

Abstract: Chemical modifications to histone tails is one of the key epigenetic mechanisms involved in transcription regulation. Detecting differential histone modification sites are of great interest for the study.

ChIP-seq is a popular NGS protocol for DNA-protein interaction analysis. The conventional post-processing for ChIP-seq results consists of three steps. First, ChIP-seq reads are mapped to the reference genome sequence. Then the reference genome is partitioned into non-overlapping segments, called bins. And finally each bin is assigned the number of reads starting within its boundaries. The problem of detecting differences in histone tail modification using ChIP-seq thus reduces to finding the differently enriched bins constructed for two ChIP-seq samples. Similar to RNA-seq analysis, ChIP-seq enrichment can be defined in terms of fold change in read counts between the two samples [1]. However, due to the differences in signal / noise ratio in different libraries, can be a source of errors with this approach.

We propose a novel method for finding differences in two ChIP-seq samples. The method assigns an enrichment cluster to each bin, based on the number of reads the bin contains. Each enrichment cluster is characterized by a single parameter -- the probability of observing a read. Enrichment clusters within ChIP-seq sample are ranked w.r.t. the cluster parameter. The total number of enrichment clusters is a parameter of the model.

The key assumption of our method is that most genomic regions should have similar enrichment patterns in the two samples being analyzed. The method seeks to find cluster parameters minimizing the number of differences between the samples. A bin is declared similarly enriched if it is assigned clusters with the same rank in each of the samples. Otherwise, the bin is marked as increased or decreased, based on the relation between the cluster ranks.

The method is implemented as a hidden markov model with three states: unchanged, increased, decreased.

We applied our method to the ChIP-seq data from a study of myogenic differentiation (GSE25308). The data covered a wide range of histone tail modifications: H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K36me3, H3K9ac, H3K18Ac, H4K12Ac. We compared the results of our method to ChIPDiff. Our method detected most of the differential sites found by ChIPDiff and also some sites, which ChIPDiff marked as unchanged. We plan to investigate the nature of the disagreement between the two methods in future work.

1. Han Xu, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data

A63: Imene Boudellioua and Victor Soloviev. Investigation of Prokaryotic Gene Regulation in Distant Phylogenetic Groups

Abstract: Thousands of microbial genomes have been sequenced. However, functional sites have been studied experimentally only for few model organisms, mainly, E.coli and B.subtilis. The aim of this work is to study and analyze the gene regulation functional motifs in prokaryotes and reveal the similarities and differences between distant phylogenetic groups. Such study will lead to a better-informed design of novel biosynthetic pathways and aid in the construction of phyla-specific promoter models. The considered regulatory motifs of gene expression in this study are promoter region short sequences upstream transcription start site; the -10 box and the -35 box.

We classified all completely sequenced prokaryotic data from NCBI database per phyla. Their candidates of promoter region functional sites were predicted using FGENESB suite for automatic annotation of microbial genome. Furthermore, we developed an iterative method for optimal alignment of corresponding functional elements found in different microbes and different genes. Our method uses a scoring scheme for the most similar sequence selection based on position weight matrix with log likelihood that reflects the common structure of the set of the predicted functional motifs. We found that our method yielded more than 44% increase in information content for all prokaryotic promoter regions in comparison to the regions originally suggested by FGENESB annotator. We observed that the GC content of bacteria in the -10 box, -35 box, and in the whole promoter region (150 bp) are, 20.27%, 33.31%, and 36.85% respectively. Such low GC content, especially in the -10 box demonstrated the lowest DNA stability of this site among the other regions. We examined detailed structure of these sites within groups of archaea and bacteria. We found that the average distance between the two boxes in prokaryotes was 24 bp. In addition, we observed that across all phyla, there is a very similar consensus sequence of -10 box that has the consensus sequence TATAAT with three almost fully conserved bases. However, for the –35 box, the reported TTGACA consensus sequence was less conserved across all groups and varied in terms of location, number of bases conserved, and the level of conservation. For example, some bacterial phyla such as Elusimicrobia and Deferribacteres don't exhibit any very strongly conserved bases. Another interesting finding was in a big phylum, Firmcutes, which has no strong consensus sequence visible for the -35 box of all its organisms. However, analyzing the -35 box of one of its representative organisms, Bacillus anthracis, we observed that its -35 box is highly conserved and has higher information content. This may reveal that not all species of Firmcutes share similar promoter structures around the -35 box. Hence, we plan to further study smaller subgroups of prokaryotes to analyze the relationship between their features, their functional site structure and those of other prokaryotic organisms.

A64: Takeru Nakazato, Tazro Ohta and Hidemasa Bono. DBCLS SRA: Functional mining and characterization of public NGS data

Abstract: High-throughput sequencing, also called next-generation sequencing (NGS), is rapidly spread and numerous sequence read data have been archived in public database, the sequence read archive (SRA) by DDBJ, EBI and NCBI. SRA contains approximately 38,000 projects as of June 2014, which is double that of the previous year. The archived SRA data contains not only raw read sequences but also information on the experimental design including project titles, species or cell lines of samples, and sequencing platforms as metadata. However, complicated data structure prevents researchers from utilize these SRA data. Therefore we categorized public NGS data by these study types (e. g. whole genome analysis, transcriptomics, and metagenomics), sequencing platforms, and species of samples. In addition, many results using NGS platform are reported and the NGS data referred in these journal articles should be focused. We thus constructed a publication list that refers NGS data and make hyperlinks to corresponding journal articles and SRA data. We also characterize each SRA entry with diseases by extracting keywords from retrieved journal articles. Recently, DDBJ/EBI/NCBI launched new database for projects and samples called BioProject and BioSample, respectively, because recent projects are sometimes so large that the results are submitted to various databases such as GenBank/EMBL-bank (nucleotide sequence), GEO/ArrayExpress (gene expression) and SRA (high-throughput sequencing). Especially, gene expression data with NGS platforms are archived to both GEO and SRA. Project information is originally archived as "study" in SRA, but these information has moved to BioProject. We therefore indexed each entry in various databases to a corresponding data set in SRA database.
To visualize these information, we developed a web-service called DBCLS SRA (http://sra.dbcls.jp/). This service will drastically improve accessibility to SRA data of interest with hight-quality.
[Reference]
Experimental design-based functional mining and characterization of high-throughput

sequencing data in the Sequence Read Archive.
Nakazato T., Ohta T., Bono H., PLOS One, 8 (10): e77910 (2013)

A65: Chaehwa Seo, Sangok Kim, Jieun Kim, Wan Kyu Kim and Sanghyuk Lee. Exome and Transcriptome Profiling of Lung Adenocarcinoma in Female Never-Smokers

Abstract: Background: Lung cancer is the leading cause of cancer-related mortality in the world. TCGA projects for lung adenocarcinoma and squamous cell carcinoma are the largest datasets publicly available so far. They include diverse subtypes and population groups, and thus small population groups such as female never-smokers in Asian are under-represented. Female never-smokers are an important population cohort where mutations due to genetic factors are enriched rather than environmental factors such as smoking. We applied the high throughput deep sequencing strategy to catalog somatic mutations and to profile gene expression levels for Korean patients.
Results: Deep sequencing data of exome and transcriptome were generated for tumor and matched normal tissues from 100 non-small-cell lung cancer (NSCLC) patients of never-smoker female adenocarcinoma (NSFAd). The mutational catalogue includes numerous somatic mutations, indels, and gene-fusion events. Comparison with TCGA lung cancer data shows that our patients harbor much less and homogeneous somatic mutations as expected. Mutational frequency and driver gene analyses identified many novel candidates of causal genetic variations. For example, our mutation spectrum reveals four key modules of enriched somatic mutations such as EGFR, p53, Wnt signaling pathways, and focal adhesion. Furthermore, network analysis of integrating mutation and expression data illustrates the detailed mechanism of dysregulation in tumorigenesis of lung adenocarcinoma.

A66: Ari Ugarte, Juliana Bernardes and Alessandra Carbone. CASH : a tool to identify domains and functionally annotate metagenomic and metatranscriptomic sequences

Abstract: The improvements of next-generation sequencing have allowed researchers to study the genomic diversity in microbial communities. The increased complexity of metagenomics data poses computational challenges in assembling, annotating, and classifying genomic fragments from multiple organisms. Domain identification provides insights to the biological function of a protein. Hence, domain annotation is a crucial step to identify and quantify the genes in a microbial community that are known and those that are completely new. Traditional protein annotation methods describe known domains with probabilistic models representing the consensus among homologous domain sequences. When relevant signals become too weak to be identified by consensus, attempts for annotation fails. CASH (Bernardes J., Zaverucha G., Vaquero C., Carbone A., submitted, 2014), a new method for protein domain identification which achieves highly accurate predictions for single genomes compared to Pfam methodology, is based on the observation that many structural and functional protein constraints are not globally conserved through all species but might be locally conserved in separate clades. Hence, it uses an extension of the probabilistic model library in order to characterize a large number of local models to improve domain detection. We applied CASH to simulated and real metagenomics and metatranscriptomics data sets, and evaluated its performance. First, we simulated a data set containing 500,000 reads of Roche's 454 FLX titanium sequencer. We built this data set from 40 bacterial and archaeal complete genome sequences assuming equal abundance. Predicted genes in simulated sequences were translated to proteins and annotated with CASH and Pfam. CASH identifies substantially more domains than Pfam in simulated reads (~30% more detected domains) at the same False Discovery Rate threshold. Besides the improvement in domain recognition, CASH agrees with 96,5% of Pfam domain predictions and reinforces the signal of domains that are also detected by Pfam. To prove that this new method is suitable for real data, it was applied to 5 data sets containing unicellular eukaryotic metatranscriptomic sequences. CASH outperforms Pfam methodology in domain recognition, and signal detection in agreed domains for all data sets. For each dataset, we mapped all domains obtained by CASH for functional annotation using Pfam2Go and a list of GO Terms. The same was done for domains obtained with Pfam. CASH allows to greatly extend the list of significant GO Terms compared to Pfam. Moreover, it permits to have a better resolution of significant GO Terms and highlights the functional characteristics of each sample. Therefore, our results show that CASH is suitable not only for domain recognition but also to improve functional annotation in metagenomics or metatranscriptomics studies.

A67: Sabine Van Dillen, Anne-Claire Coûté-Monvoisin and Philippe Horvath. Development of a pan-genome tool and its application to comparative genomics analysis in Lactobacillus plantarum

Abstract: The pan-genome of a bacterial species, defined as the union of the gene sets of all the strains, comprises the core-genome, with genes present in all strains, and the dispensable (or accessory) genome, corresponding to genes that are only found in a subset of strains. Within this latter category is included the fluctuating sub-category of unique genes, whose existence heavily depends on strain sampling bias. The determination of core and accessory genes, although not trivial, provides critical insights into the dynamics and evolution of bacterial genomes, and further facilitates the identification of genetic traits responsible for strain-to-strain variations.

Here we present the development of a comparative genomics pipeline that automatically calculates pan/core/accessory/unique-genome characteristics. This Linux-based tool relies on BioPerl and MySQL, and takes as input a number of annotated sequences from which all coding sequences (CDS, and possibly other features) are extracted and compared against all using BLAST (BLASTn or BLASTp). The results are then filtered according to user-defined parameters, notably BLAST match identity and coverage, to cluster all features into gene families, for which multiple sequence alignments and dendrograms are automatically generated.

This approach was applied to Lactobacillus plantarum, a lactic acid bacterium commonly used as a starter culture in various food (meat, vegetable, juice, and wine) fermentations. Some L. plantarum strains are also used as probiotics and silage inoculants. The analysis was performed with 20 L. plantarum genomes publicly available in GenBank, including 6 complete chromosomes and the occasional plasmid complement.

Overall, 63,203 annotated CDS of L.plantarum were classified into 5,374 orthologous clusters (pan-genome), a number which corresponds to 1.7 fold the average total number of CDS (3,160) in each of the 20 genomes. Among these families, 1,823 clusters were identified as core, representing 34% of the pan-genome, whereas the remaining 3,551 clusters were identified as dispensable, with 2,110 clusters shared by two to 19 strains, and 1,441 strain-specific clusters. This study shows that the gene composition of L. plantarum strains is very variable, explaining its adaptability and ability to colonize various ecological niches.

A68: Ronald Schuyler and Simon Heath. Nucleosome-influenced DNA methylation gain and loss during differentiation

Abstract: Cell identity is dependent on the spacial organization of chromatin. In the presence of the DNA binding protein CTCF, nucleosomes are known to form consistently spaced arrays which are reflected in local patterns of DNA methylation, but the dynamics of this process during cellular differentiation, activation and aging have not previously been

determined.
We integrated whole genome bisulfite sequencing (WGBS) data from ~50 samples from Blueprint and other sources with CTCF occupancy data and nucleosome positioning data from ENCODE to determine global patterns of methylation at occupied and unoccupied CTCF binding motifs.
The nucleosome-associated DNA methylation pattern near constitutively occupied CTCF binding sites becomes more pronounced (deeper and broader) with successive stages of normal lymphocyte differentiation and in lymphoid cancers, but not in cancers derived from the myeloid lineage. The bias for methylation gain at linkers and loss at nucleosomes is cumulative over differentiation and activation, possibly contributing to cell memory and adding weight to cell fate commitment at this major determinant of higher order genome organization.

A69: Francesca Nadalin and Alberto Policriti. A new approach to RNA-Seq data analysis based on local paired reads assembly

Abstract: Motivation: RNA-Seq reads alignment against the reference genome is at the basis of genome annotation and expression levels estimation. Standard approaches for spliced alignment are unable to correctly map a read to an exon when the overlap is very short; moreover, models for expression levels computation often make assumption that are not justified from a biological point of view.
Both these facts negatively affect precision of transcripts quantification.
Our contribution is based on (i) a deep exploiting of paired reads information aimed at increasing mapping specificity, and (ii) the introduction of gene structure and coverage profile information in the model for accurate expression levels prediction.
Methods: We propose a pipeline for spliced alignment that envisages prior in-silico reconstruction of paired read inserts, in order to overcome mapping ambiguities. The alignment pipeline is structured as follows: (1) insert assembly and layout information storage (i.e., information on reads used for assembly); (2) inserts alignment against the reference genome and refinement around gaps; and (3) indirect reads alignment retrieval. The tool GapFiller is used to reconstruct inserts of a paired read library; several experiments demonstrated that GapFiller has low FP rate and thus provides a trustable output. In order to save memory, a compact representation of insert layout is employed that guarantees irredundancy and efficient retrieval of original data.
Transcripts abundancies are found solving the linear system M x = c associated to each gene, where M is a binary matrix, x contains expression levels, and c contains exons coverages (computed from the reads aligned with our pipeline). Both gene structure and coverage profile are taken into account for expression levels computation. Each exon is associated a quality value, depending on exon length, exon coverage uniformity, and coverage consistency with other exons. Then, x is computed on a minimal sub-system of M x = c identified by the set of highest-quality exons. Both exons selection and solution computation are done in ploynomial time.
Results: Evaluation was done on simulated RNA-Seq reads and comparison with state-of-the-art tools was performed as well. Our method is precise in detecting spliced alignments: 99% of reads are split in the correct number of hits, meaning that the pipeline can accurately detect reads spanning splicing sites. Compared to other tools, our pipeline turns out to have the lowest FP rate; in fact, other methods can report more events, but at the expenses of a high amount of incorrect predictions. Results on expression levels estimation are also encouraging. In the vast majority of cases,
RPKM values are predicted with a mean error rate < 10%. Some cases of high error rates are

still present, but they are due to very low RPKM values, meaning that the absolute error rate is low.

A70: Dimitrios Zisis, Iris Hovel, Rurika Oka, Blaise Weber, Maike Stam, Jan-Jaap Wesselink and Pawel Krajewski. 4C-seq data processing, normalization and differential analysis of chromosomal contact profiles in Arabidopsis thaliana

Abstract: Numerous studies indicate that long-range chromosomal interactions contribute to gene and genome regulation. It was however not until the development of the chromosome conformation capture (3C) technology that the widespread role of chromosomal interactions became clear. One of the 3C-based methods is 4C-seq. This method provides information on physical interactions between a known fragment of interest and the rest of the genome. 4Cseqpipe is a computational analysis pipeline providing support for the analysis of 4C-seq experiments (van de Werken et al., 2012). It includes sequence extraction, mapping, normalization and the generation of high resolution contact profiles around the “viewpoint”. To our knowledge, it has been used only for mammalian genomes, which happen to have different characteristics than a plant like A. thaliana. For this reason we propose a 4C-seq data processing schema partially based on tools commonly used for NGS data analysis and we show its application for four data sets obtained using the Arabidopsis FLC locus as a viewpoint. FLC serves as a very useful model system for epigenetic studies, because its expression can be modulated easily by changing growing conditions or by using mutants in various pathways. Concentrating on inter-chromosomal contacts, we present considerations concerning sources of data bias resulting from the distribution of A. thaliana genomic restriction sites and propose a data normalization method. Finally, we use statistical functional data analysis (Ramsay and Silverman, 2012) to find differences between contact profiles obtained for different samples.

A71: Renaud Vanhoutrève, Julie Thompson, Pierre Collet and Olivier Poch. YAMSA (Yet Another Multiple Sequence Alignment) method using Parisian Evolution on GPGPU

Abstract: The incredible increase in the output of new generation sequencing technologies has completely exceeded Moore's Law, making data analysis a major bottleneck for the biologist. New bioinformatics solutions are necessary to allow end-users to fully exploit the progress of these technologies in various applications, including genome annotation, structural modelling of proteins, analysis of genetic mutations, phylogenetic studies, etc. In this context, multiple sequence alignments (MSA) play a central role in processing the huge amount of data. Recently, new methods for the construction of MSA (such as MAFFT, MUSCLE, KALIGN) have been developed that are fast enough to handle ‘big data’ problems, but they generally produce alignments that are of lower quality than slower algorithms (such as ProbCons).

YAMSA is a new MSA approach, which exploits an alternative genetic algorithm, called Parisian Evolution, in order to combine the advantages of the different, complementary methods. The conventional genetic algorithm tries to find an optimal MSA by creating a population of random MSA and allowing them to evolve following the theory of evolution (with crossovers and mutations). In this case, an individual in the population represents one solution to the problem, i.e. an individual is an alignment of all sequences with gaps at given positions. In the Parisian Evolution approach, an individual represents a part of the solution and in the final stage of the algorithm the main solution is reconstructed from the individual parts.

In YAMSA, each individual represents a part of the alignment (subset of sequences), which is aligned with one or more of the standard MSA methods. Although this is more complex to implement, the different approaches form a co-evolving system where the individual parts help to explore the search space and generate a robust composite result.

YAMSA has 2 types of crossover: merge a sequence in both parents or create a mixed alignment based on both parents. The current mutation operator is basic and involves adding or removing a sequence in individuals. One or more of the alignment approaches is then selected, creating one or more new individuals. In the final stage, the individual parts are combined to create the complete MSA by ranking the sub-alignments and progressively incorporating the high scoring individuals.

YAMSA is implemented on the EASEA platform (easea.unistra.fr), which allows us to exploit the massively parallel architectures such as GPGPU. The performance of YAMSA is demonstrated in a large scale evaluation involving more than 200 multiple alignments from the BAliBASE benchmark.

A72: Mi Ni Huang, John R. McPherson, Bin Tean Teh, Patrick Tan and Steven G. Rozen. Assessing Microsatellite Instability in Tumor Exome Sequences

Abstract: Microsatellite instability (MSI) is a form of hypermutation that arises during the development of several types of cancer due to inactivation of the DNA mismatch repair system. MSI is characterized by frequent somatic mutations (i.e., cancer-specific mutations) that change the length of mononucleotides (e.g., AAAAA….) or microsatellites (e.g., GATAGATAGATA….) as well as by frequent single-nucleotide-substitution mutations. Standard clinical-laboratory tests for MSI evaluate the lengths of only a few mononucleotides and/or microsatellites (collectively, simple-sequence repeats) in the human genome. With next-generation sequencing (NGS), we can now assay a far larger number of simple-sequence-repeat sites for somatic mutations. As a result, we can develop a more complete view of the mutations' frequencies and distributions. In this study, we analyzed publicly-available somatic mutation data from the exomes of 612 colorectal, endometrial, and gastric tumors with known laboratory-determined MSI statuses. We then developed and evaluated classifiers to determine MSI status based on this exome mutation data. We used four machine-learning frameworks: logistic regression, decision trees, random forests, and naïve Bayes. All four frameworks performed similarly as assessed by five-fold cross validation in a training set. We chose a decision tree classifier that had > 96% concordance with the laboratory assessments in training-set cross validation and almost 98% concordance in a separate test set. This classifier was robust, in that it retained high concordance even when classifying based on subsets of whole-exome data. The high concordance of this classifier with laboratory-based assessments indicates that whole-exome NGS data contains adequate information for assessing MSI status. We have submitted an R package, MSIseq, based on this classifier, to Bioconductor. MSIseq will be useful for genomic studies in which laboratory-assessed MSI status is unavailable, for detecting MSI in cancers for which laboratory-based MSI testing is not routine, and for detecting possible misclassifications by laboratory assessments.

A73: Jaime Castro-Mondragon and Jacques van Helden. Comparing, clustering and aligning transcription factor binding motifs with RSAT matrix-clustering

Abstract: Transcription factors binding motifs (TFBM) are classically represented either as consensus strings (IUPAC, regular expressions), or as position-specific scoring matrices (PSSM). Thousands of curated TFBM are available in specialized databases (JASPAR, RegulonDB, TRANSFAC, etc), built from collections of transcription factor binding sites (TFBS) obtained from various experimental methods (e.g. EMSA, DNAse footprinting, SELEX). TFBM can also be discovered ab initio from genome-scale data sets: promoters of co-expressed genes, ChIP-seq peaks, phylogenetic footprints, etc.

Motif collections sometimes contain groups of similar motifs, for different reasons: curation of alternative motifs for a same TF; homologous proteins sharing a particular DNA binding domain, motifs discovered with analytic workflows combining several algorithms (e.g. RSAT peak-motifs, or MEME-chip). In order to address the increasing need for efficient tools enabling to discover groups of similarities among motif collections, we developed matrix-clustering, which presents significant advantages over existing solutions.
1) Segmentation of the input set of TFBM into separated clusters, displayed as a motif forest rather than a single motif tree (alternative software tools force all motifs to be aligned).
2) Multiple alignment of all motifs belonging to a same cluster.
3) User-friendly display of motif trees with aligned logos and consensuses.
4) At each level of the hierarchical tree, computation of a merged motif (matrix and consensus) summarizing all the descendant motifs.
5) Support for a large series of alternative metrics (correlation, Euclidian distance, SSD, Sandelin-Wasserman, logo dot product, and length-normalized version of these scores).
6) Possibility to select a custom combination between these scores to compute an integrative score (rank mean).
The potentialities of the tool are illustrated by study cases: (1) clustering of matrices extracted from ChIP-seq peaks using several motif discovery algorithms; (2) Extraction of a motif-to-motif network and clustering of all motifs from the JASPAR taxon-wise collections. (3) The significance of the clustering results is further assessed by analysing collections of randomized matrices (column-permuted). In this negative control, most motifs are correctly assigned to a singleton, except for low complexity motifs (e.g. AAAAAA).
We analysed the effect of hierarchical clustering parameters (hierarchical agglomeration rule, similarity metrics) on the number of clusters and on the relationships between motifs, and identified suitable parameters to obtain relevant results.
Availability: matrix-clustering is available on the Regulatory Sequence Analysis Tools (RSAT) Web site (RSAT; http://www.rsat.eu/). It can also be downloaded with the stand-alone RSAT distribution to be run from the Unix shell.

A74: Ching Chang, Mei-Ju May Chen, Tony Kuo, Jian-Long Huang, David S. Haymer, Ju-Chun Hsu and Chien-Yu Chen. Improving completeness of de novo transcriptome assembly and gene annotation by multi-species transcriptome sequencing in fruit fly genus Bactrocera

Abstract: RNA-seq tool provides considerable information without reference genome, which largely facilitates the studies of non-modeled organisms. Transcripts can be de novo assembled from sequencing reads and then used for subsequent analysis, e.g. gene quantification. Meanwhile, de novo assembled transcripts needs to be annotated to infer their molecular functions. Most studies annotated such transcripts by using an adjacent model organism. However, this approach suffers from an inevitable limitation when the target organism has a great distance from a model organism in the evolutionary tree, such as the relationship between Bactrocera dorsalis and Drosophila melanogaster. Gene annotation would be incomplete because some genes in B. dorsalis have no homologues in D. melanogaster. Sometimes, incorrect annotation may also happen because of the incompletely assembled transcripts. In this study, a method was proposed to improve the completeness of assembly and annotation by utilizing de novo transcripts assembled from the RNA-seq data of two species sharing the same genus, i.e., Bactrocera Cucurbitae and Bactrocera dorsalis. Ideally, these two species can support de novo assembled transcripts from each other since they contain a great amount of homologues owing to a close evolutionary distance. The results show that more annotated transcripts can be achieved when compared to only using D. melanogaster protein sequences as the reference. The traditional approach can obtain about 10,000 annotated transcripts through annotation against D. melanogaster protein sequences by

using BLASTx. After we utilized the linkage of the overlapping transcripts from the two species, we identified about 1,000 more annotated transcripts for future studies, revealing the contribution of the proposed approach.

A75: Jocelyn De Goer De Herve, Myoung-Ah Kang, Xavier Bailly and Engelbert Mephu Nguifo. A perceptual hash algorithm for indexing and similarity search in a database of DNA sequences

Abstract: In the pipeline of data analysis from high-throughput genome sequencing[1], the similarity search between DNA sequences is a common problem for most of genomic project. It is an important part of the work during the sequence assembly tasks, for the identification of consensus sequences, during the annotation process, for detection of mutations or to determine a sample of biological diversity during metagenomics studies.

This work proposes a novel approach to perform the similarity search between a DNA sequence and a reference sequence database. This method is based on a perceptual hashing[2] function based on DCT (Discrete Cosine Transform) Signs[3] and designed for digital image indexation[4], which has been adapted for biological sequences. The indexing process splits the DNA sequence into multiples sub-sequences, and applies the perceptual hashing function on each of them in order to compute the hash keys. Each key has a size of 32 bits (4 characters) and is smaller compared to the complete sequence encoded in FASTA format. The query of the hash table process starting with a sequence, consists in generating all hash keys from this sequence, and comparing them with the whole hash table. The perceptual function is not affected by the avalanche effect[5], thus the similarity indices between two hash keys can be evaluated. To optimize the operations (read-write) related to the storage, the hash table, corresponding to the reference sequences is stored in a NoSQL[6] In-Memory Key-Value database.

A first implementation of the method has been released in C++ language, and the process of evaluation by theoretical simulations (128 billions of comparisons) reports good execution time results, and also allowed to validate the perceptual hash indexation and comparison method.

BIBLIOGRAPHY:

1. Metzker, M.L. (2010) - Sequencing technologies - the next generation. Nature Review Genetics, 11 : 31-46
2. Ton Kalker ; Jaap Haitsma ; Job C. Oostveen - Issues with digital watermarking and perceptual hashing - Proc. SPIE 4518, Multimedia Systems and Applications IV, 189 (November 12, 2001); doi:10.1117/12.448203
3. Hiroshi KONDO, Satoshi KUNIFUSA, Zhimei YANG, Takaharu KODA, and Lifeng ZHANG (2004) - Binary Signal Compression using DCT Signs - 2nd International Conference on Autonomous Robots and Agents December 13-15, 2004 Palmerston North, New Zealand
4. Zauner, Christoph (2010) - Implementation and Benchmarking of Perceptual Image Hash Functions. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg Campus.
5. Horst Feistel (1973) - Cryptography and Computer Privacy - Scientific American, May, 228(5) : 15-23
6. A B M Moniruzzaman, Syed Akhter Hossain (2013) - NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison - arXiv:1307.0191 [cs.DB].

A76: Joanna Sasin-Kurowska, Piotr Borsuk, Jan Gawor, Robert Gromadka, Jakub Grzesiak and Marek Zdanowski. Supraglacial community responces to increasing global temperatures as revealed by comparative metagenomics

Abstract: Increasing global temperatures have a profound impact on the supraglacial ecosystem. The majority of life in the Antarctic is microbial. Recent evidence supports important role of polar microorganisms in the regional and global carbon cycling [Stibal et al. 2012]. Considering that ice currently covers 11% of the terrestrial landmass the potential for these environments to contribute to global biogeochemical cycles on glacial-interglacial timescales is enormous [Boyd et al. 2011]. Understanding the structure of polar microbial communities is essential for predicting the response of these microbiota to climate change. The comparative metagenomic studies of glacial microbiomes in three model polar glaciers (The Ecology- the King George Island, The Werenskiold- Spitsbergen and The Hans Glacier – Spitsbergen) were performed. The analyzed data came from three distinct areas of glacier ablation zone. The structures of bacterial communities were analyzed by pyrosequencing technique. The most numerous bacteria phyla were Acidobacteria, Cyanobacteria, Proteobacteria, Bacteroidetes. The discriminant analysis of physicochemical properties of glacial environments were performed. Finally, we studied the relationship between taxonomic diversity and the environmental factors.

A77: Mathias Vandenbogaert, Anne-Sophie Delannoy-Vieillard, Laure Diancourt, Aurélia Kwasiborski, Jean-Michel Thiberge and Valérie Caro. Alignment-free sequence composition analysis of High-Throughput Sequencing runs, for pre-assembly read clustering

Abstract: High-Throughput Sequencing (HTS) projects that aim to re-sequence known genomes, or to reliably identify bacterial or viral species in a diagnostics context, are increasingly required to be more sensitive than before. Clinical samples often contain either a number of (meta-) genomes that are at best divergently related to known references, or a limited number of genomes with very low coverage. Variable levels of similarity challenges both identification through mapping and “genome finishing” in diagnostics settings, where the distinction between presence and absence of single species or relative abundance levels are of eminent importance. To overcome the limitations of alignment-based comparisons, ultrafast alignment-free methods are exploited, providing a powerful comparison tool to distinguish different organisms present in meta-genomics HTS read datasets.

Meta-genomic samples represent a large number of reads, rendering assembly without pre-condition computationally inefficient, and often proves to result in under-assembly and chimeric contigs. We emphasize on clustering of sequencing reads, based on n-mer counts, as a preliminary step prior to assembly, which can be a valuable brick in a processing pipeline, as reads belonging to possibly different species show different n-mer compositions.

Preliminary results are presented on alignment-free clustering of HTS sequenced read-data based on word counts, in both a memory-less scheme as through higher order Markov chain models, in overlapping mode, and clustered through an Expectation-Maximization-clustering approach, on metagenomics datasets displaying various distributions of count-data, from both bacterial and viral origin. In essence, each sequence is represented by the vector of n-mer counts. EM-clustering is applied, via iterative refinement of cluster centers. The Kullback-Leibler log-likelihood divergence measure is used, to measure the difference between P (the count-data; the "true" probability distribution) and Q (the model or approximation to P; the "target" probability distribution).

Results indicate that, as alignment-free methods relying on clustering of word enumerations are obviously less accurate than direct sequence alignments, they should only be used when direct alignment is either impossible (due the high level of meta-genomic divergence) or computationally too complex. In addition, to more thoroughly distinguish pathogen from host, and “self” from “foreign”, the word-size in the counting procedure and the order of the Markov chain model, are parameters to be optimized as a function of the length of the reads. Indeed, the length of the reads is steadily increasing, with the different sequencing technologies at hand, i.e. the Illumina HiSeq/MiSeq technology generates 100 to 250 bp reads, whereas Life Technology’s Ion Torrent PGM/Proton systems generate reads up to 300, 400 bp and above.

A78: Ivan Kel, Christoph Dieterich, Zisong Chang, Luciano Milanesi and Ivan Merelli. Applying eQTL analysis to RNA-Seq data to study genetic regulation of miRNA expression in C.elegans

Abstract: Expression quantitative trait loci (eQTL) mapping has become a popular and widely used approach to study the impact of polymorphisms within gene regulatory regions on the regulation of expression of genes (Rockman and Kruglyak, 2006). This technique aims to identify loci based on the substitutions in the genome that result in differential measurable transcript levels and thus can take advantage from high-throughput sequencing methods like RNA-Seq. The potent statistical instrument provided by eQTL-mapping can be exploited to analyse a very large number of genes and in particular seems to be suitable for the study of small RNAs such as miRNAs and their regulatory networks.

As suggested by Gilad et al., 2008 a combination of the eQTL results with traditional linkage studies can establish a link between a genomic locus, known to be associated with a disease, and a specific gene-regulatory role. Such disease-linked loci were successfully identified as candidates for involvement in hypertension in rats (Hubner et al., 2005). In human the studies delivered a variety of associations across a diversity of diseases, including HIV and diabetes (Schadt et al., 2008).

In this work we report the results of an eQTL analysis study performed on miRNA RNA-Seq dataset of 30 recombinant inbred lines (RILs) of the crosses between two different strains of C. elegans, laboratory strain N2, and a wild isolate from Hawaii (CB4856). A modular, generic and functionally highly flexible pipeline for eQTL processing of NGS data was developed during the course of the project. The pipeline integrates a number of univariate, as well as multivariate eQTL methods, but most importantly provides a unique framework for the eQTL analysis embedding complex preliminary steps of analysis both for the genotype and the expression part. It also integrates a hotspot identification method on large association intervals, thus enabling the discovery of trans-eQTLs. Further, it is also designed to provide scalable (cluster based) performance, in order to manage large datasets and to allow the correlation of the variants on the whole genome. The pipeline can be downloaded as an open source package.

As the results of this work we predicted a number of eQTLs for C. elegans miRNA, using both various univariate as well as a multivariate method represented by Random Forest approach. Upon others, the genes correlated with the identified eQTLs are involved in numerous events of C. elegans larval development and other essential processes. Furthermore, the eQTL hotspot analysis identified a significant hotspot in the beginning of the chromosome 1 overlapping with a gene encoding a transcription factor. Performing a search for binding sites of transcription factors (TFBS), we found in the regions around the miRNAs located in the hotspot a significant increase of the average frequency of this binding motif, confirming the hypothesis of the trans-regulatory role of the identified gene.

A79: Sangok Kim, Yukyung Jun, Charny Park, Pora Kim, Jieun Kim, Chaehwa Seo, Kyoohyoung Rho, Jong-Eun Lee, Wan Kyu Kim, Harkyun Kim and Sanghyuk Lee. Multi-dimensional Genomic Study of Early-Onset Gastric Cancer

Abstract: Background: Gastric cancer is among the most frequent tumors in Asian countries, and yet the genetic basis of tumorigenesis is largely unknown. Most genomic studies focused on elderly patients whose etiology is complex due to the dependence on patient-specific environmental factors. Cases of early-onset gastric cancer (EOGC) are an important patient cohort with presumably homogeneous background of genetic factors rather than environmental causes.

Results: We have generated the deep sequencing data of whole exome and transcriptome for tumor and matched normal samples from 50 gastric cancer patients of early-onset cases under the age of 45. Mutation analysis revealed a number of somatic mutations, indels, and gene-fusion events. Most commonly mutated genes (TP53, FAT4, ARID1A, CDH1, FMN2, etc.) were in good agreement with the previous reports even though we obtained several EOGC-specific somatic mutations. Comparison with the cancer genome atlas (TCGA) data indicated that the mutation frequency of EOGC was distinct from that of elderly patients. Gene set analysis of somatic mutations showed the Wnt signaling, cell cycle, and p53 pathways as frequently dysregulated pathways. We further identified a number of germline mutations that had been reported to be somatic in other types of tumor. Integrative network analysis of somatic and germline mutations as well as transcriptome expression provided many novel insights on the tumorigenesis mechanism in gastric cancer patients. Detailed analyses covering diverse biological and clinical factors such as ethnicity, mutation status, and disease subtypes are currently in progress.

Conclusion: Our study elucidates many clinically important mutations and regulatory elements in EOGC cases. It is expected that our dataset would serve as an important resource for gastric cancer research especially in Asian populations.

A80: Margus Lukk. Disk-free computing framework for NGS Big Data – how fast can we compute?

Abstract: The speed of Big Data processing is determined by the availability of computing power. The individual units for many types of the Big Data come in reasonable size and can be entirely processed in the memory of the computer. In this respect, the NGS data is different. The short reads data comes typically in very large files. While being processed, a few equally large temporary files may have to be created. This makes NGS data processing not only computation but additionally I/O, and more generally, I/O performance bound.

Faster data turnaround times are obtained by parallel computing. Processing short read data in shared memory systems which are limited by a small number of CPU cores, requires a smaller number of temporary files. Handling short read data in distributed systems, on the other hand, demands more temporary space and in some cases even additional copies of the original data. The time spent on I/O operations together with the high I/O induced decrease of file system performance has a particularly dramatic effect for distributed parallel processing of very large datasets.

In this work we introduce a novel concept of disk-free computing and test it on NGS Big Data. We present a disk-free framework designed for combining non-pipeable software into pipelines, thus eliminating the need to read and write large temporary files. Besides optimising local computing, the framework takes advantage of distributed resources and allows parts of the pipeline to be run in foreign hosts. Most importantly, the framework is built for massive distributed parallelisation of existing software tools without the need to re-write any code.

We apply the framework to NGS sequencing data and demonstrate the efficiency and unprecedented speed-gain a) in disk-free distributed short read mapping compared with traditional mapping strategies; b) in distributed sorting of large bam files; c) and in conversions/manipulations of large NGS data files. Finally, we show that removed I/O burden

allows large NGS datasets, normally processed in days or hours, to be turned around even in minutes.

A81: Andre Kahles, Cheng Soon Ong and Gunnar Rätsch. SplAdder: Integrated Quantification, Visualization and Differential Analysis of Alternative Splicing

Abstract: The alternative choice of exons during the maturation process of mRNA, termed alternative splicing (AS), is one of the key mechanisms that shape the complexity of higher organism's transcriptomes and that ensure the necessary flexibility in expression from a single locus that is vital for development and gene regulation. The analysis of AS-events from RNA-Sequencing (RNA-Seq) data plays a central role in understanding the mechanisms of gene regulation and elucidating the development of diseases. We present SplAdder, the first integrated analysis framework, that can detect AS-events from RNA-Seq alignments, quantify them and differentially test them between two given sample sets. AS-events are detected from a given annotation or can be added to it based on the given RNA-Seq evidence. The detected events can then be quantified on the same or a different set of RNA-Seq alignments. Differential analysis, employing a negative binomial test, is realized with the rDiff package. SplAdder further comes with several visualization routines, producing publication ready plots of the quantified splicing graph, displaying one or many events or showing sashimi plots for different isoforms. SplAdder can easily handle several thousand samples from human and has been developed and tested on data from The Cancer Genome Atlas project. However, SplAdder is not limited to human and we demonstrate applications in the plant A. thaliana as well as the nematode C. elegans. The software is implemented as open source software and is available both as Python code as well as Octave implementation. For more information visit www.bioweb.me/spladder.

A82: Aleksander Jankowski, Jerzy Tiuryn and Shyam Prabhakar. MOCCA: accurate identification of transcription factor footprints by modeling DNase I cut profiles

Abstract: The identification and characterization of active regulatory elements is essential to decipher the regulatory mechanisms in eukaryotic genomes. This involves the identification of transcription factor (TF) binding sites, in different cell types and conditions, within these elements. The traditional method involves the digestion by DNase I and subsequent identification of regions where TFs are bound to DNA and protect the DNA from degradation by the enzyme. These protected sites, or TF footprints, can be identified on a large scale by a more recent protocol, DNase I digestion followed by high-throughput sequencing (DNase-seq).

Here, we propose a computational method, codenamed MOCCA, to accurately identify TF footprints from sequence information and DNase-seq data. For a given TF, we identify candidate binding sites that have reasonable sequence affinity, using a position weight matrix. Then, we employ an Expectation-Maximization-based approach to simultaneously learn the DNase I cut profiles and classify the binding sites as bound or unbound. Our method is unique in allowing for multiple bound states for a single TF, differing in their cut profile and overall number of DNase I cuts. To make the model robust and limit its number of free parameters, we employ a systematic approach to group the DNase I cuts into bins, according to their location and strand.

Our approach outperforms two existing methods, CENTIPEDE and Wellington, when benchmarked on ChIP-seq data for 11 TFs in K562 cells. Moreover, when allowing for more than one bound state, we found that the additional DNase I cut profiles can capture half-site

binding, as well as potential cofactors. We also found some of the additional DNase I cut profiles to be highly asymmetric, and hypothesize about the possible explanations.

A83: Tahila Andrighetti, Gunther Johannes Lewczuk Gerhardt, Agnes Alessandra Sekijima Takeda, Ney Lemke and Jose Luiz Rybarczyk-Filho. Identification of metagenomic data by genome signatures and n-entropy analysis

Abstract: The advances on sequencing technologies enabled the development of metagenomics, helping with one of the greatest challenges in microbiology: the full characterization of microbiomes. It allowed a deeper analysis of microbial environments, such as species quantification and functional characterization. However, researchers have been facing a challenge to analyze the data obtained, since most of the available bioinformatics tools are based on sequence homology; implying that computational cost increases exponentially as the fragments size decreases. An alternative to improve the characterization is the detection of genome signatures, which are sequence organization patterns shaped by the selection pressure of the environment on the microorganisms. Herein we propose a new technique for metagenomics based on genomic signatures and entropy. Sequence organization measures were used in microorganisms DNA fragments, in order to find genome signatures that can be used to help the recognition of the fragments obtained by metagenomics. Regions of 70kbp in the genome of 2164 species available on GenBank were randomly selected, resulting in two groups of different regions from the same species to be analyzed, corresponding to training and test groups. All the regions were subdivided a hundred times in random fragments of 4096, 2048, 1024, 512, 128 and 64 base pairs to evaluate the influence of the sequences length on the methodology. We calculated GC content, 3bp periodicity and N-entropy (where N = 2, 3 or 4) using two ways to determine the entropy. The average and the standard error were calculated for both groups to determine a high confidence interval. Each species were allocated in its family taxon, totalizing 149 families. To evaluate the best combinations of genomic signatures to filter metagenomic data, 49 combinations were analyzed aiming a high sensitivity and specificity for prediction in metagenomic taxonomic data. The preliminary results has shown that 50% of the metagenomic data can be identified by the combinations of n-entropy, and 28% were identified by N-entropies and GC-N-entropy combinations, totalizing the taxonomy identification of 78% of the sequences of a metagenome. These data suggest that the entropy sequence analysis can be applied in studies of microbiomes as a marker for structural organization of organisms.
Supported by: FAPESP (Process number : 2013/15174-4)

A84: Séverine Gagnot, Mireille Ansaldi and Emmanuel Talla. Identification and characterization of small viral proteins in bacteriophage and prokaryotic genomes.

Abstract: Phages are viruses that inject their genetic material (DNA or RNA) into bacteria and replicate themselves inside. Virulent phages can only develop short lytic life cycles to generate new phage particles whereas temperate phages can switch from a lysogenic life cycle to a lytic life cycle under certain conditions. The lysogenic cycle is characterized by the integration of the viral genome into the bacterial chromosome to produce a latent form: a prophage. Although the integrase is the main protein involved in phage DNA excision or integration within a bacterial chromosome, the Recombination Directionality Factor (RDF, also called excisionase or Xis) plays a critical role to direct the site-specific recombination reaction towards excision [1]. Most of the RDFs (e.g. TorI in Escherichia coli K12) are small size proteins (less than 100 aa). Even if their primary sequences are very divergent due to the high rate of viral protein mutation, we notice that they adopt a similar structural model that essentially consists of a winged-helix DNA binding motif composed of a helix-turn-helix

motif and a loop (wing) [1].
The small size of the RDFs, their poor sequence conservation, but strikingly 2D or 3D structure conservation led us to choose a tool of the HH-suite package, HHblits, based on HMM-HMM comparison and taking into account the protein secondary structure prediction, as a very fast and powerful tool to detect RDF homologs [2]. For this purpose we built a customized HMM database. We downloaded 1057 phage and 1382 selected prokaryotic proteomes. Protein sequences were clustered and then the clusters were enriched by searching a HH-suite HMM database (uniprot20) for homologs and by adding of secondary structure predictions. These multiple sequence alignments were transformed into HMMs. The next step consists of the comparison of the HHM profiles of 21 experimentally determined RDFs [3] to our HHM database via the HHblits software to detect RDF homologs. As preliminary results, we obtained a high number of RDF homologs (compared to a simple similarity analysis) for which some of them exhibit less than 10 % sequence identity with the queried RDFs.
Complete analysis of the RDF homologs, including protein characterization, genomic context and classification will be described on our poster.

[1] Elantak L et al. Structural and genetic analyses reveal a key role in prophage excision for the TorI response regulator inhibitor. J Biol Chem. 2005 Nov 4;280(44):36802-8.
[2] Remmert M et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011 Dec 25;9(2):173-5.
[3] Lewis JA et al. Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins. Nucleic Acids Res. 2001 Jun 1;29(11):2205-16.

A85: Thomas Abeel, Bruce Walker, Alex Salazar, Terrance Shea, Chris Desjardins, Jennifer Wortman, Sarah Young and Ashlee Earl. Identifying large sequence variants in collections of bacterial genomes with Pilon and Emu

Abstract: Advances in sequencing technology allow us to generate sufficient data to analyze hundreds of bacterial genomes from a single Illumina machine in a single day. This potential for sequencing massive numbers of genomes calls for an integrated method to identify genetic variants. Currently, the identification of large sequence variants (LSVs) that exceed the sequenced fragment length is particularly challenging. Accurately identifying LSVs is critical to associate this class of mutations to phenotypes that are relevant to human health, such as drug resistance and virulence. We introduce a pair of complementary tools, called Pilon and Emu, which, together, enable consistent LSV identification among large genomic data sets.
Pilon is a computational tool for correcting draft assemblies and calling LSVs of multiple sizes, including very large insertions and deletions. To identify LSVs, Pilon searches for discontinuities in the aligned read-data and then locally reassembles that region, yielding, in many cases, the complete sequence of the inserted bases. However, due to differences in the alignment evidence from one data set to the next, Pilon can represent a single variant in multiple ways resulting in inconsistent reporting of the same LSV across strains. This inconsistency can have confounding effects on downstream analyses that e.g., attempt to associate LSVs to phenotypes of interest.
Emu is an algorithm that normalizes different representations of the same LSV to a canonical form. Emu compares variants that have similar content and are located in the same genomic neighborhood across all included samples. By merging variants that result in the same change with respect to the reference, Emu reduces the total number of unique LSVs by as much as 50% among a collection of 161 genomes. This normalization of LSVs enables us to study the association of LSVs with clinically relevant phenotypes.

A86: Jakob Hull Havgaard, Kortine Kleinheinz, Sachin Pundhir and Jan Gorodkin. Combining read profile alignment with structural alignment to predict structured RNAs in transcriptomic data

Abstract: High-throughput sequencing often reveals completely novel transcripts which lack protein coding potential. It can be hard to distinguish whether these transcript are functional ncRNAs or transcriptional noise.

In this work we combine two types of methods to predict if two transcripts are of the same type of structured ncRNA. The first type of methods is the read profile alignment method where the patterns of reads mapped to the genome are aligned. A well known read profile is that of miRNA where there is a very high read count at the positions corresponding to the mature miRNA, and a second smaller read count peak at the positions corresponding to the miRNA* sequence. It is the processing and structure of the ncRNA that leads to the observed read profiles, and it is therefore often possible to align read profiles from different ncRNAs of the same type. For example two different miRNAs.

The second type of methods is the structural alignment methods where the secondary structure of the RNA sequences are used to align the sequences. Structural alignment is necessary for ncRNA since the structures are usually better conserved than the primary sequences. The main disadvantage of structural alignment is the high computational cost, but by using the read profile alignment to filter which sequences to align this cost can be greatly reduced.

We show that combining read profile alignment, using the deepBlockAlign algorithm, with structural alignment, using the Foldalign algorithm, is an efficient way of conducting a screen for small expressed structural ncRNAs in transcriptomic data. The combination of deepBlockAlign and Foldalign is tested on nine sets of sequencing data from the ENCODE project.

A87: Sepideh Mazrouee and Wei Wang. Single Individual Haplotyping - third generation sequencing

Abstract: Haplotype problems, which aim to reconstruct copies of Diploid organism's chromosome, find applications in various fields of human genetics, from prognosis of organ transplant matching and clinical medicine to understanding origin of specific phenotypes and personalized drug design. Research in this area relies on sequencing datasets, which have undergone significant improvements in terms of quality and length of the sequencing data. Modern sequencing technologies enable us to reconstruct haplotype copies using DNA short fragments for individuals with higher accuracy. Quality of the reconstruction process is significantly affected by the quality of the sequenced data, making reconstruction of the haplotypes highly challenging. This necessitates development of algorithms that not only reconstruct haplotypes accurately but also require low computation time and therefore enable scalability of the reconstruction process. Haplotype research has utilized Next Generation Sequencing (NGS) data in the last decade to construct haplotypes of higher quality and lower error rates. Yet, several limitations of the data generated by NGS, such as the length of fragments, have limited capabilities of the haplotype reconstruction algorithms. Utilizing recent advancements in sequencing technology can be a solution to overcome these limitations. The third generation sequencing fragments have extra-long fragment lengths at the cost of higher sequencing error rate. The objective, therefore, is to design algorithms that mitigate error impact and rebuild the most likely copies of each chromosome accurately and fast while accessing extra-long fragments.

We introduce a framework for accurate and fast haplotype assembly. The presented framework takes the sequencing data as input and produces various performance metrics regarding haplotype reconstruction based on a new fragment partitioning approach. We

present a novel similarity metric for the distance between pairs of fragments which then is utilized to introduce a new graph model (fuzzy conflict graph) that captures the amount of inter-fragment dis/similarity. Using the graph model we developed a fast, yet accurate, fragment partitioning and haplotype reconstruction algorithm which is one order of magnitude faster than the state-of-the-art haplotype assembly algorithms. We will demonstrate that the new generation of sequencing data (Pacific Bio-Sciences and 1000G dataset), which essentially demands highly scalable algorithms, can benefit from our haplotype assembly approach mainly due to the high level of flexibility and scalability. Our analysis demonstrates that the proposed approach is up to 17 times faster than HapCut, a historically known fast and accuracy haplotype assembly algorithm. It is also up to 7 times faster than simple greedy approaches such as FastHare. Our approach also outperforms these algorithms with 3.0% to 7.1% reduction in minimum error correction (MEC) values.

A88: Anaïs Vittu, John Randy Clayton and Stéphanie Blandin. A pipeline for the identification of contaminant microorganisms in high-throughput RNA/DNA sequencing data

Abstract: Context: Malaria is a parasitic infection transmitted by mosquitoes of the genus Anopheles. The etiological agents of malaria are Apicomplexan parasites of the genus Plasmodium. These pathogens alternate between vertebrate and invertebrate hosts during their complex life cycle. The vectorial capacity of Anopheles gambiae varies extensively between individuals, with some mosquitoes resistant to parasites and therefore unable to transmit the disease. The capacity of a mosquito to transmit malaria is determined by mosquito genetic factors, parasite virulence factors, but also environmental factors such as bacteria, viruses or other parasites that may colonize mosquitoes and shape their basal immunity or interfere with Plasmodium development. We have isolated several lines of mosquitoes with different degrees of susceptibility to malaria parasites. In this project, we aim to identify the species of bacteria/viruses/fungi that populate these lines and look for correlations between their susceptibility and the microbiota they carry.

Methods: We have built a pipeline to identify novel organisms present in libraries of mRNAs, small RNA and genomic DNA that were sequenced from mosquito samples using Illumina technology. For this, we first align reads on the mosquito genome or its transcripts using Bowtie and we select reads that do not map to the mosquito reference sequence. This step may be reiterated several times on the remaining reads to remove all reads that map on additional species that we know are present in the sample but not interesting to us for this analysis.

After filtering out all the “expected” reads, we use Velvet to perform a de novo assembly of all remaining reads to obtain contigs. We then retrieve large contigs using a perl script and launch a standard nucleotide BLAST in NCBI to determine whether these contigs are part of an organism or assembly artefacts. Small contigs are mapped on the NCBI Reference Sequence Databases for bacteria, fungi, viruses… using Bowtie. Our pipeline has a specific Perl script that summarizes the results in a new file and highlights significant findings from each output alignment file.

This pipeline is easily to organise, to use and gives results in one day.

Conclusions: Using this pipeline, we have identified several microorganisms present in our mosquito populations. In adult mosquitoes, we found numerous reads mapping to Saccharomyces cerevisiae which we feed young larvae with for a short time after hatching. We also identified water-related viruses, insect viruses and some bacteria. We aim to compare the lists of microorganisms that are present in resistant and susceptible mosquitoes to identify those that may contribute to determine mosquito vectorial capacity.

A89: Aleksandra Pfeifer, Barbara Jarzab and Joanna Polanska. Algorithms for fusion transcripts detection in RNA-seq data - comparison and improvement

Abstract: Introduction: In many carcinomas, fusion transcripts (caused by chromosomal rearrangements) are present. During last few years, RNA-seq have been widely used to successfully detect fusion transcripts in multiple cancers. In the same time, multiple bioinformatic tools have been created for detection of fusion transcripts in RNA-seq data. Those tools differ in performance, some of them generate a lot of false positives or have low sensitivity. It would be useful to compare those tools to help in algorithm selection. It would be also useful to improve the results by careful selection of parameters and applying additional filters.

Material and Methods: We performed comparison of 8 programs for fusion detection: FusionMap, TopHat Fusion, ChimeraScan, BreakFusion, SnowShoes, DeFuse, Trans-ABySS and FusionAnalyser. We used public RNA-seq data from 16 samples (4 breast cancers, 6 melanomas, 3 prostate cancers, 1 leukemia, 2 normal samples) with total number of 79 known fusion transcripts. We performed both single-end and paired-end analysis. We also created additional filters to get rid off false positive fusions and we carefully selected programs' parameters to improve the sensitivity.

Results: The most sensitive program was ChimeraScan. It gave high sensitivity of 82% but also high false detection rate of 96%. The program with lowest false detection rate was SnowShoes. It gave false detection rate of 9% and medium sensitivity of 36%. Another program that gave good results was ChimeraScan which gave medium values of both sensitivity and false detection rate.

Applying additional filters and optimizing the parameter selection caused significant improvement in the performance of fusion detection.

Coclusions: Careful selection of algorithm, selection of parameters and applying additional filters significantly impact the performance of fusion detection in RNA-seq data.

The Project was financed by the National Science Center Poland based on the decision no. DEC-2011/03/N/NZ2/03495. This research was supported in part by PL-Grid Infrastructure.

A90: Gift Nuka, Matthew Fraser and Maxim Scheremetjew. InterProScan 5: Large-scale protein sequence analysis

Abstract: InterProScan 5 is a function prediction tool built to efficiently handle large-scale sequence analysis of both protein and nucleic acid sequences. This poster presents new developments including new analysis types, output formats (GFF, SVG) and techniques to achieve scalable large-scale distributed data analysis.

New analysis include combined transmembrane topology and signal peptide predictor (Phobius) and new features include ability to infer potential membership of proteins in pathways (via InterPro entry mappings to KEGG, MetaCyc, UniPathway, etc.).

InterProScan 5 parallelizes search tasks at three levels using different techniques including message passing to achieve a high level of parallelization of the analysis steps on a cluster or supercomputer.

A91: Konstantin Okonechnikov, Aki Imai-Matsushima, Lukas Paul, Alexander Seitz, Thomas F. Meyer and Fernando Garcia-Alcalde. InFusion: advancing discovery of fusion genes and chimeric transcripts from deep RNA-sequencing data

Abstract: Gene fusions and chimeric transcripts occur frequently in cancers and in some cases drive the development of the disease. An accurate detection of these events is crucial for cancer research and in a long-term perspective could be applied for personalized therapy. RNA-seq technology has been established as an efficient approach to investigate

transcriptomes and search for gene fusions and chimeric transcripts on a genome-wide scale. While a number of computational methods for the detection of gene fusions from RNA-seq data have been developed, recent studies have found differences between commonly used approaches in terms of specificity and sensitivity. Moreover their ability to detect gene fusions on the isoform level has not been studied carefully so far. Here we propose a novel computational approach called InFusion for fusion gene detection from deep RNA sequencing data. InFusion introduces a number of unique features such as discovery of chimeric RNAs involving intergenic regions and detection of anti-sense transcription in fusions based on the strand-specificity of the sequencing library. Our approach demonstrated superior detection accuracy on simulated and on several public RNA-seq datasets. We also performed deep RNA sequencing of two well-established prostate cancer cell lines. From these data InFusion identified 26 novel fusion events that were validated in vitro, including alternatively spliced gene fusion isoforms and chimeric transcripts that include non-exonic regions.

A92: Andigoni Malousi, Justine Guégan, Vincent Guillemot, Vincent Perlbarg, Arthur Tenenhaus and Ivan Moszer. Scaling Sequencing Pipelines for Whole Genomes

Abstract: Running whole-genome sequencing data analyses timely and cost-effectively is an imperative need, especially given the foreseen massive production of such data. Contrary to exome data, whole-genome data of higher eukaryotes can be analysed only with the support of high-throughput technologies (cloud or local cluster infrastructures) that allow effective management and parallelisation of the pipelined tasks. We have implemented a fully automated pipeline for whole-genome sequencing data analyses, running in three stages. The first one includes splitting the original sequencing files into chunks of millions of reads that subsequently feed the sequence aligner in parallel. The resulting sorted alignments are then merged and cleaned from PCR duplicates. The above processes run in parallel with quality control tools that build reports on the quality of the reads and the extracted alignments. The second stage includes the per chromosome minimisation of the alignment artifacts by locally realigning incorrectly mapped reads and subsequently recalibrating base quality scores. The final step refers to variant calling. This includes the detection of both SNPs and short deletions/insertions but also complex structural variants, e.g. inversions, intra-, inter-chromosomal rearrangements, etc. The latter is performed using a combinatorial method, SVDetect [1], which also incorporates the detection of copy number variations. Besides the inherent parallelisation capabilities of each variant calling method, this stage is performed in parallel for each chromosome and, since these methods are independent, they can also be launched in parallel.

The developed pipeline uses bpipe [2] for the definition and execution of each stage together with publicly available resources that address different steps of the analysis. The proposed pipeline is deployed in a local cluster infrastructure and evaluated in terms of time-efficiency and quality of the results using single-sample executions on paired-end reads of human genomes. The novelty of the proposed implementation is two-fold. a) Multi-parallelisation: The pipeline increases the degree of parallelisation by using three levels of parallel execution: i) Within-task, ii) between-tasks, and iii) per chromosome/chunk/sample. Thereby, time-efficiency depends on the available resources but also on how these resources are distributed among tasks running in parallel. b) Adoption/Extensibility: Due to the built-in support of multiple resource managers by bpipe, the proposed pipeline can be easily adopted to different cluster environments. Besides, the pipeline is cross-platform and self-contained: once the included tools are set, no further administrative rights are required to use the pipeline. Finally, due to its modular implementation, the pipeline can be easily enriched with additional study-specific functionalities.

[1] Zeitouni B. et al. Bioinformatics. 2010 26:1895-6
[2] Sadedin SP. et al. Bioinformatics. 2012 28:1525-6

A93: Jacques Lagnel, Khalid Belkhir, Tereza Manousaki, Erick Desmarais, Anastasia Tsagkarakou and Alban Mancheron. New tools to optimise the analysis of a large RNA-Seq dataset from non model species: development of a hybrid assembly strategy and assessment of library complexity from raw sequencing output

Abstract: With the advance of new sequencing technologies major challenges have emerged on different levels, from library preparation up to data analysis. Here, we present one tool that improves the quality and feasibility of the assembly process and another that assesses the complexity of the sequenced library.

Currently, transcriptome assembly strategies are either reference-based or “de novo” depending on the availability of high-quality reference sequences. For non-model species where a high-quality reference transcriptome/genome is lacking, a closely-related species reference sequence can be used as a proxy to improve the quality of the reconstructed transcriptome and decrease the computational requirements. By bringing together the two complementary assembly strategies, we can take advantage of the high sensitivity of reference-based assemblers, while leveraging the ability of de novo assemblers to detect novel transcripts. The procedure includes three steps. First, the RNA-seq reads of the target species are aligned to the reference genome/transcriptome of a closely-related species. Reads that map within the same genomic location are grouped into clusters. Then, each “cluster” of aligned reads combined with the remaining unaligned reads will serve as input to a de novo assembly process running in parallel. Finally, all the resulted de novo assemblies are merged to form the final transcriptome. The de novo assembly requires important computing resources, particularly memory. The proposed strategy solves the problem of intensive memory requirements of a de novo assembly and greatly reduces the computational time by parallelising both the first and second step of the process. Here, we test this strategy on a simulated Drosophila group dataset using various reference model species in a wide range of divergence times to assess the mapping success. Our pipeline led to great improvement when closely related species were used as a reference. This was further tested on an experimental dataset decreasing the required computational resources.

Prior to the assembly process, sequencing experiments can be evaluated based on the success of library construction and sequencing. Often, technical failures can lead to biased RNA representation and over-sequencing of particular molecules that do not represent the starting biological sample and cause reduced complexity. The level of complexity reflects the coverage of the transcriptome. We developed a new metric that infers the diversity of unique sequences and assesses the complexity of a given library using two clustering steps of identical reads. This metric can be used to compare the library construction success and for evaluation of multiple experiments.

The presented tools facilitate the analysis of RNA-Seq for non-model species improving the feasibility of the assembly and allowing for the post-sequencing evaluation of the library construction.

A94: Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants

Abstract: The interpretation of human genetic variation on a genome-wide scale is a crucial challenge in both research and clinical settings. Available annotations tend to exploit a single information type (e.g. conservation) and/or are restricted in scope (e.g. missense changes). A broadly applicable metric is needed that objectively weights and integrates the large, diverse, and otherwise unwieldy collection of annotation data available. We developed Combined Annotation Dependent Depletion (CADD), a framework that integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations. We show that our scores correlate with allelic diversity, pathogenicity of both coding and non-coding variants, and experimentally measured regulatory effects, and also highly rank causal variants within individual genome sequences. We pre-computed SNV scores for the whole human genome and enable scoring of short InDels (http://cadd.gs.washington.edu). We describe our method and discuss the integration of additional annotations as well as methodological improvements.