Table of contents

DeepBlue: Diving into Epigenomic Data

Felipe Albrecht * [1]

[1] Max Planck Institute for Informatics (MPI-INF) – Germany

DeepBlue: Diving into Epigenomic Data

Felipe Albrecht, Markus List, Christoph Bock, Thomas Lengauer

Large volumes of data are generated by several epigenomic consortia, including ENCODE, Roadmap Epigenomics, Blueprint Epigenetics, and DEEP. To enable users to utilize these data effectively in the studies of epigenetic regulation,we have developed the DeepBlue Epigenomic Data Server [1]. While DeepBlue allows storing, organizing, searching, and retrieving epigenetic data, it does currently not provide the means to conduct epigenomic data analysis. To address this issue for diverse users, we follow several strategies: (i) an R/Bioconductor package (http://deepblue.mpi-inf.mpg.de/R) integrates DeepBlue into The R analysis work-flow. The extracted data is automatically converted to GenomicRanges [2], which are supported by many related packages for analysis and visualization; (ii) a web interface (http://deepblue.mpi-inf.mpg.de) that allows users to find and download epigenomic data of interest from more than 33,000 available experiments; (iii) DeepBlue Dive (http://dive.mpi-inf.mpg.de), which is inspired by EpiExplorer [3] and helps researchers to visually compare their own epigenomic data to data already available in DeepBlue; (iv) a complementary web tool to DeepBlue Dive, which is inspired by EpiGraph [4] and uses LOLA [5], and reports on the enrichment of epigenomic regions provided by the user among the experiments available in DeepBlue..

With the DeepBlue Epigenomic Data Server, we provide programmatic access to vast amounts of epigenomic data. Here, we present a series of tools that build upon the DeepBlue API and enable users not proficient in scripting or programming languages to benefit from our efforts and to analyze epigenomic data in a user-friendly way. DeepBlue and related tools are available at http://deepblue.mpi-inf.mpg.de/. This work has been supported by German Science Ministry Grant No. 01KU1216A (DEEP project) and has been performed in the context of EU grant no. HEALTH-F5-2011-282510 (BLUEPRINT project)

Albrecht,F., List,M., Bock,C. and Lengauer,T. (2016) DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic Acids Research*,doi:10.1093/nar/gkw211

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M and Carey V (2013). "Software for Computing and Annotating Genomic Ranges." *PLoS Computational Biology*,9.

Halachev, K., Bast, H., Albrecht, F., Lengauer, T. & Bock, C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome Biology 13, R96 (2012).

*Speaker

Bock, C., Halachev, K., B́uch, J. & Lengauer, T. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. Genome Biology 10, R14 (2009).

Sheffield N.C., Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics 2016;32:587-589.

Standardizing chromatin research: a simple and universal method for ChIP-seq

Laura Arrigoni *† [1], Andreas Richter [1], Emily Betancourt [1], Kerstin Bruder [1], Sarah Diehl [2], Thomas Manke [1], Ulrike Boenisch * ‡ [1]

[1] Max Planck institute of immunobiology and epigenetics (MPI-IE) – Stübeweg 51, 79108, Freiburg, Germany

[2] Luxembourg Centre for Systems Biomedicine, Université du Luxembourg – avenue du Swing 6, Belvaux, 4366, Luxembourg

Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) is a powerful technique for the genome-wide mapping of DNA-binding proteins and histone modifications.
As a result of our extensive contributions to the German part of the International Human Epigenome Consortium (IHEC/DEEP), we developed a new method to standardize ChIP-seq across cell types and for low input material. While many experimental techniques such as high-throughput sequencing have already been standardized and largely automated, the extraction of chromatin is still a major challenge. Custom-made adjustments and laborious optimizations are often necessary to obtain chromatin of sufficient quality for epitope detection and deep sequencing. Such sample-specific solutions are not transferable, error-prone, and they often result in irreproducible chromatin quality or failures of ChIP-seq.

We demonstrate that harmonization of chromatin extraction for ChIP-seq across cell types is possible when obtaining chromatin from properly isolated nuclei. We identified the common source of inappropriate chromatin quality is the insufficient nuclei extraction. Commonly used approaches, based on chemical treatment and mechanical homogenization, have extremely low nuclei recovery when applied to formaldehyde-fixed cells. Therefore we established a novel procedure which uses ultrasound to break the cell membrane, while leaving the nucleus intact (NEXSON: Nuclei EXtraction by SONication). NEXSON is extremely effective and reliable in many different systems and biological samples, ranging from blood cells to hepatocytes, adipocytes, fibroblasts, cultured embryonic stem cells and even for whole organisms such as *Drosophila* embryos. By including NEXSON in ChIP-seq workflows, we completely eliminated the need for extensive optimizations and sample-dependent adjustments.

Such approach streamline and simplify state-of-the-art ChIP-seq protocols, making possible to apply the same procedure across cell types and saving a substantial amount of time and test material to obtain a working protocol. This is particularly important when dealing with patient-derived unique samples, where there is no possibility of re-iterations. Simple quality checks can be performed in real-time and directly on the sample of interest.

As another attractive side-effect, we found that NEXSON can be used for ChIP-seq on a very low number of cells, without applying any protocol adjustments or extensions. In particular, we successfully generated high-quality chromatins and histone ChIP-seq tracks from as little as

*Speaker
†Corresponding author: arrigoni@ie-freiburg.mpg.de
‡Corresponding author: boenisch@ie-freiburg.mpg.de

10,000 cells.
Our method can replace cell type-specific protocols and will significantly improve the comparability of chromatin maps from different research groups and consortia.

Efficiency and the quality of methylation calls of RRBS and WGBS pipelines

Yassen Assenov [*†] [1], Reka Toth [2]

[1] Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ) – Heidelberg, Germany
[2] Division of Epigenetics and Cancer Risk Factors German Cancer Research Center (DKFZ) – Heidleberg, Germany

We benchmarked multiple publicly and in-house available tools and pipelines dedicated to the processing of RRBS and WGBS raw data. We compared the pipelines in terms of technical requirements such as processing cores, memory usage, possibility of parallelization and speed. Also, the results of their methylation calling in raw and smoothed form were compared.
For the last criterion, we investigated the degree of their concordance using Pearson correlation coefficient and mean methylation difference. To further improve the comparisons we have implemented an improved metric that takes into account the coverage as well. In the presence of a gold standard, we used a coverage corrected version of the mean methylation difference as a measure of the alteration between two pipelines.

As expected, most of the calls were very similar in the different pipelines, with some sites showing larger differences. The raw differences were lowered with increasing coverage. The largest disagreements in the methylation calls between pipelines were mainly in repetitive regions.
We also developed a R based visualizing tool – MeViz – that beautifully shows the degree of local agreements of different pipelines, technologies and/or cohorts. While the tool was designed to visualize different techniques for measuring methylation, it can be easily used to overlay methylation with other sequencing-based data, such as ChIP-seq or RNA-seq.

[*]Speaker
[†]Corresponding author: y.assenov@dkfz-heidelberg.de

Investigating transcription factor binding using DNase-seq datasets

Anaïs Bardet * [1]

[1] Friedrich Miescher Institute for Biomedical Research (FMI) – Basel, Switzerland

Transcription factors (TFs) recognize specific DNA sequence motifs within regulatory regions to control the expression of their target genes. Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) is the method of choice to identify TF binding sites throughout a genome. DNase I digestion followed by high throughput sequencing (DNase-seq) can be used to identify regions of open chromatin typically bound by TFs without prior knowledge of TF identity. Here, I present an analysis and integration of DNase-seq datasets together with TF motifs and other sequencing data types used to study TF binding and its sensitivity to DNA methylation.

*Speaker

Long-range chromatin interaction mapping: From quality assessment to the implementation solutions for the analysis of temporal 3D-chromatin reorganization

Matthias Blum * [1], Valeryia Malysheva [1], Marco Antonio Mendoza Parra [1], Hinrich Gronemeyer [1]

[1] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) – CNRS : UMR7104, Inserm : U964, université de Strasbourg – Parc D'Innovation 1 Rue Laurent Fries - BP 10142 67404 ILLKIRCH CEDEX, France

Today massive parallel DNA sequencing is not only used to decrypt the digital nature of genomes but, in combination with a variety of molecular biology techniques, it provides functional insights into a plethora of regulatory levels, including epigenomics and protein-genome interactions (e.g., ChIP-seq, MeDIP-seq), global transcriptional activity (e.g., RNA-seq, GRO-seq, Ribo-seq), as well as chromatin accessibility (e.g., DNase-seq, FAIRE-seq, ATAC-seq, MNase-seq). More recently, the 3-dimensional chromatin organization, assessed by proximity-mediated ligation strategies combined with massive parallel DNA sequencing (Hi-C, ChIA-PET), is recognized as a further level of complexity interconnecting chromatin architecture with its functional regulatory mechanisms.
During this presentation we will discuss our current efforts for (i) evaluating the quality of Hi-C and related assays; (ii) improving current approaches for correcting systematic biases in raw Hi-C contact maps; (iii) implementing computational solutions for normalizing multiple Hi-C contact maps in order to evaluate 3D-chromatin reorganization in cell fate transition events like cell differentiation or tumorigenesis transformation.

*Speaker

WBS: a computational pipeline for the treatment of whole-genome high-throughput bisulfite sequencing data

Eric Bonnet * [1], Yimin Shen [1], Xavier Benigni [1], Nizar Touleimat [1], Jorg Tost [1], Jean-François Deuleuze [1], François Artiguenave [1]

[1] Centre National de Génotypage (CNG) – CEA – Centre National de Génotypage 2 rue Gaston Crémieux CP5721 91057 EVRY Cedex, France

DNA methylation is an important epigenetic mechanism used by higher eukaryotes and is involved in several key physiological processes, including regulation of gene expression, X-chromosome inactivation, imprinting and silencing of germline-specific genes and repetitive elements. Patterns of methylation are maintained through somatic cell divisions and may be inherited across generations. These patterns are altered in many complex human diseases, such as imprinting disorders and cancer. Understanding methylation patterns is therefore of great importance for many biomedical questions.
Bisulfite treatment of DNA is method of choice to analyse these patterns. Bisulfite treatment leaves methylated cytosines unaffected. Thus, bisulfite treatment introduces specific changes in the DNA sequence that depend on the methylation status of individual cytosine residues, yielding single-nucleotide resolution information about the methylation status of a segment of DNA. Various analyses can be performed on the altered sequence to retrieve this information. Especially, rapidly falling costs of high-throughput sequencing have made the global analysis of DNA methylation at the whole genome level a viable option.

However, there are significant computational challenges associated with the computational treatment of bisulfite generated reads. Here we describe WBS (Workflow Bisulfite), a computational pipeline set-up at the Centre National de Génotypage (CNG) for the analysis of bisulfite whole genome sequencing data. The pipeline is built around standard state-of-the-art tools and workflows for the treatment of bisulfite reads with the aims of being efficient, standardized and easy to run and maintain. We describe the organization of the pipeline, performance and possible evolutions in the framework of the analysis of mammalian (mostly human) whole genome DNA methylation patterns.

*Speaker

On the implications of cohort-based variation analysis as a new paradigm in cancer genomics

Lars Feuerbach * [1]

[1] Deutsches Krebsforschungszentrum (DKFZ) – Im Neuenheimer Feld 280 69129 Heidelberg, Germany

Identifying somatic mutations in tumor samples is a cornerstone of molecular cancer diagnostics and precision oncology. The current state-of-the-art for whole-genome-sequencing data is a sequential comparison of each genomic position in the sequenced tumor sample to a reference and a control genome. Hereby, complex, but yet incomplete, scoring models are applied to discriminate genomic mutations from false positive signals. Especially, for mutations in the non-coding genome this computational problem remains challenging. The recent aggregation of large databases of cancer genomes allows the direct incorporation of large cancer genome cohorts into the mutation calling process. Sharing and efficiently integrating this 'big data' touches aspects such as data privacy, interpretability, maintainability, efficient data structures and compression techniques. Therefore, a new generation of software solutions is required that will allow the distributed, rapid and secure use of large cancer genome databases for cohort-based mutation calling. Here, DeepPileup as a software tool that enables a compressed representation and visualization of individual nucleotide positions of several thousand tumor and control samples is used as an example to discuss challenges and solutions.

*Speaker

The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update

Björn Grüning * [1]

[1] Uni-Freiburg (ALU) – Bioinformatics Group Department of Computer Science
Albert-Ludwigs-University Freiburg Georges-Köhler-Allee 106 79110 Freiburg Germany, Germany

Started in 2005, Galaxy enables biologists without programming and systems administration expertise to perform computational analysis through the web. Existing analysis tools are defined for Galaxy and made available with a consistent web interface and via the API. By bridging the gap between tool developers and researchers, Galaxy helps both constituencies accelerate their research. In this talk we will present the advances over the last years and how these improve the developer and user experience. The interaction of Galaxy with new technologies like Docker and Conda packaging as well as Galaxy's Interactive Environment will be demonstrated. It will be discussed how we can solve the deployment problem in bioinformatics and finally deliver accessible and reproducible pipelines to the end-user without the need to transfer data.

*Speaker

HPC development in the context of large NGS analysis

Frederic Jarlier * [1]

[1] Institut Curie section Recherche (U900) – Institut Curie, Institut National de la Santé et de la Recherche Médicale - INSERM – Institut Curie, 26 rue d'Ulm, 75005, Paris, France

After an intensive study of existing tools we have started computing development to reduce the time of analysis. We opted for MPI (Message Passing Interface) for the parallelisation. MPI is a standard in High Performance Computing, it takes advantage of supercomputer systems equiped with fast network links and distributed file system. Moreover its implementation Open-MPI is free of charge. We propose a framework to optimize the two first steps the alignment and the sorting of NGS whole genome samples.

*Speaker

The 1002 yeast genomes project: Exploring the genotype-phenotype relationship of Saccharomyces cerevisiae

Jackson Peter * [1], Anne Friedrich [1], Agnès Llored , Anders Bergstrom , Anastasie Sigwalt , Kelle Freel , Gianni Liti , Joseph Schacherer *

1

[1] Génétique moléculaire, génomique, microbiologie (GMGM) – CNRS : UMR7156, université de Strasbourg – 28 Rue Goethe 67083 STRASBOURG CEDEX, France

Genome-wide survey of polymorphism patterns in a large sample of individuals is the entry point to a better understanding of the genotype-phenotype relationship. To date, yeast population genomics only focused on a limited number of isolates. Here, we sequenced 1,011 genomes of *Saccharomyces cerevisiae* isolates with high coverage (226x on average). Due to the broad diversity of the selected strains, this dataset reveals an accurate picture of the genomic variation. Our dataset is well suited to reveal the entire repertoire of SNPs (58,934,477 SNPs distributed across 1,544,490 polymorphic sites) and the degree of copy number variation for 7,849 open reading frames of the entire *S. cerevisiae* pangenome. Extensive phenotyping experiments by measuring growth under 36 stress conditions, *i.e.* more than 36,000 trait measurements, allows us to map quantitative trait nucleotide or quantitative trait gene by performing genome-wide association studies (GWAS). The results showed significant associations for 21 out of 35 growth conditions, with approximately of them were explained by a copy number variant and the rest by a single nucleotide variant. This study leads to the identification of a large set of functional polymorphisms that underlie phenotypic variations and constitutes a very rich dataset for the community.

*Speaker

Challenges of analyzing Hi-C data

Fidel Ramirez * [1], Vivek Bhardwaj , Thomas Manke

[1] The Max Planck Institute of Immunobiology and Epigenetics (MPI-IE) – Stübeweg 51 D-79108 Freiburg, Germany

Hi-C is one of the available techniques that prove the 3D conformation of chromatin by using high-throughput sequencing. In Hi-C protocol, chromatin is enzymatically cut and re-ligated, allowing the formation of hybrid DNA segments containing spatially proximal chromatin. After paired-end sequencing the reads need to be processed to infer the conformation. During my talk, I will focus on three particular challenges that we encountered when processing Hi-C data: i) correction of data, ii) identification of long-range contacts and iii) determination of TADs.

*Speaker

A new tool for non hybrid correction of long noisy reads

Eric Rivals * [2,1]

[2] Institut de Biologie Computationnelle (IBC) – Univrsité de Montpellier – Montpellier, France
[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université Montpellier II - Sciences et Techniques du Languedoc – Montpellier, France

Nowadays, several long read sequencing technologies are available and long reads are theoretically advantageous not only to assemble a genome, but also to investigate the linkage of genetic variations or the diversity of transcriptomes. However, current levels of sequencing errors hamper the use of long reads. For instance, even the simple task of read alignment on a reference sequence becomes less reliable, more complex and time consuming than with short reads. Several hybrid error correction methods, such as LoRDEC were recently proposed: they require and take advantage of short reads to correct long reads. Here, we present a non hybrid error correction method, which only uses long reads. This method, embodied in a software called LoRMA, relies on LoRDEC to itera- tively correct long reads using several De Bruijn graphs of increasing order. Then it performs multiple alignment to compute a long read consensus. LoRMA was tested on bacterial and yeast datasets and provides reliable correction in reasonable computing times.

LoRMA is available at: https://www.cs.helsinki.fi/u/lmsalmel/LoRMA/
LoRDEC is available at: http://www.atgc-montpellier.fr/lordec/

Work in collaboration with L. Salmela, R. Wake, E. Ukkonen de l'Université d'Helsinki, Finlande. Early access publication in Bioinformatics: doi: 10.1093/bioinformatics/btw321

*Speaker

Comparative genomics and transcriptomics of several gene families in Vitis genus

Camille Rustenholz *† [1], Gautier Arista [1], Guillaume Barnabé [1], Sophie Blanc [1], Didier Merdinoglu [1], Philippe Hugueney [1]

[1] Santé de la Vigne et Qualité du Vin (SVQV - INRA-UDS) – Institut national de la recherche agronomique (INRA) : UMR1131 – 28 rue de Herrlisheim 68021 Colmar Cedex, France

Grapevine (*Vitis vinifera*) is a fruit crop of great economic importance worldwide and especially in France. The availability of a reference genome sequence since 2007 helped to speed up analyses of agronomic traits like quality and resistance to biotic stresses. At INRA Colmar, we are focusing on gene families especially involved in aroma biosynthesis, like the gene family coding for terpene synthases, or involved in the defence against pathogens, like the gene families coding for NBS-LRR proteins or stilbene synthases (STS).
The STS gene family is composed of 48 members organized into two clusters in the grapevine reference genome. RNA-Seq analyses are performed to establish their expression profile across many conditions in order to understand the regulation of the family. CNV analyses based on resequencing data from 57 *Vitis* genotypes are also carried out on this particular gene family in order to assess its evolutionary dynamic at the family level. Allele assembly would be of great interest to study the evolutionary dynamic at the sequence level. However, the high percentage of identity between the genes (70% to 99.9%) and heterozygosity are our main challenges.
Finally, this increasing knowledge about the gene families in the grapevine genome will help to understand the regulation and the dynamic of genomic regions involved in agronomic traits.

*Speaker
†Corresponding author: camille.rustenholz@colmar.inra.fr

DeepTools2: A flexible platform for deep-sequencing data analysis and exploration.

Fidel Ramirez [1], Devon Ryan * [1], Björn Grüning [2], Vivek Bhardwaj [1,3], Fabian Kilpert [1], Andreas Richter [1], Steffen Heyne [1], Friederike Dündar [4], Thomas Manke [1]

[1] Max Planck Institute of Immunobiology and Epigenetics (MPI-IE) – Germany
[2] University of Freiburg, Department of Computer Science – Germany
[3] Faculty of Biology, University of Freiburg, – Germany
[4] Weill Cornell Medical College – United States

The analysis of data from high-throughput DNA sequencing experiments continues to present a large number of challenges for researchers. The ever expanding repetoire of experimental techniques has led to a concomitant explosion of analysis packages, many requiring a large degree of expertise and quality controlled and normalized input. With this in mind, we developed deepTools, a modular suite of fast and user-freindly tools available both at the command line and within both our public Galaxy instance (http://deeptools.ie-freiburg.mpg.de) and the main Galaxy site (http://usegalaxy.org). deepTools support a wide range of functions, such as various quality controls, different normalization schemes and genome-wide visualizations. Here, we present our latest release, which has greatly expanded the scope of deepTools beyond ChIPseq to now encompass RNAseq and many other *-seq techniques.

*Speaker

How can you trust your metagenetics analysis pipeline ?

Léa Siegwald *† [1,2,3,4], Hélène Touzet [3], Yves Lemoine [2,4], David Hot [2,4], Christophe Audebert [1,4], Ségolène Caboche [2,4]

[1] Gènes Diffusion – Gènes Diffusion – Douai, France
[2] Centre d'infection et d'immunité de Lille (CIIL) – CNRS : UMR8204, Institut Pasteur de Lille, Inserm : U1019, Université Lille Nord (France), CHU Lille – France
[3] Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) – Centre de Recherche en Informatique, Signal et Automatique de Lille, CNRS : UMR9189, Université Lille Nord (France), INRIA – France
[4] PEGASE-Biosciences – Institut Pasteur de Lille – France

Targeted metagenomics, also known as metagenetics, is a high-throughput sequencing application focusing on a nucleotide target in a microbiome to describe its taxonomic content (e.g. 16S rDNA for bacteria). A wide range of bioinformatics pipelines are available to analyze metagenetics outputs, and the choice of an appropriate tool is crucial and not trivial. No standard evaluation method exists for estimating the accuracy of a pipeline for metagenetics analyses, and evaluating its impact on biological interpretations.
In our recent study (paper under review), we designed an evaluation protocol allowing to study the impact of different variables on the biological interpretation of results: sequencing errors, an important feature in error-prone sequencing technology like Ion Torrent, choice of the amplified region, microbiota complexity, sequencing depth, ... To this purpose, we generated simulated metagenetics datasets (using the FaMeS artificial metagenome descriptions, Grinder for *in silico* amplification and CuReSim for sequencing errors simulation), as well as a real metagenetic dataset. We also used adequate metrics (F-measure including precision and recall, richness, diversity and clustering indices) to precisely evaluate the results. This evaluation protocol was used to compare six bioinformatics pipelines in the basic user context: Three pipelines commonly used in metagenetics studies (mothur, QIIME and BMP) based on a clustering-first approach and three emerging ones (Kraken, CLARK and One Codex) using an assignment-first approach and developed for WGS metagenetics studies.

It has been shown that targeting different genomic regions can give different pictures of the same microbiota. Surprisingly, our comparison study proved otherwise when using an error-prone sequencing technology, for which the effect of sequencing errors is predominant over the amplified region. Moreover, we showed that counter-intuitively, increasing sequencing throughput does not improve the results, but increases richness overestimation, even more so for microbiota of high complexity. We also validated that emerging pipelines, usually used in a WGS metagenomic context, can be used for metagenetics analyses. They reach a quality of results comparable to popular clustering-first pipelines on well described microbiomes. However, we demonstrated that analytical biases are not the same for both categories of pipelines: for example, the choice of the reference database has a bigger impact on richness estimation for clustering-first pipelines, and on correct taxa identification for assignment-first pipelines.
To go deeper in our understanding of the impact of analytical steps on biological results, we are

*Speaker
†Corresponding author: lea.siegwald@gmail.com

currently evaluating these pipelines on real clinical study datasets (96 samples, 2 conditions), to demonstrate how using different analytical pipelines can affect statistical analyses and biological findings.

How accurate is the Fluidigm C1 Technology for single-cell transcriptome analysis?

Constance Vagne [*†] [1], Marie Ennen [1], Irwin Davidson [1], Céline Keime [1]

[1] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) – CNRS : UMR7104, Inserm : U964, université de Strasbourg – Parc D'Innovation 1 Rue Laurent Fries - BP 10142 67404 ILLKIRCH CEDEX, France

Single-cell RNASeq enables the exploration of transcriptome heterogeneity at an unprecedented resolution by providing the expression profile of individual cells. However single-cell RNASeq techniques are based on amplification of a low amount of starting material, leading to substantial technical noise. That is why quality filtering of samples is crucial and statistical methods dedicated to bulk RNASeq data analysis are not always fully adapted to single-cell data.
Over the last five years, several single-cell RNASeq methods have been introduced. We selected the Fluidigm C1 system that uses the Clontech SMARTer technology. We used this system for several single-cell RNASeq projects on different cell types and developed a pipeline dedicated to the analysis of these data. This pipeline allows to (i) filter samples based on their quality, (ii) compute and normalize expression levels and (iii) identify and characterize cellular populations. For each of these steps, we will describe the statistical methods we tested and the corresponding results. In particular, we used a known mixture of human melanoma cell lines to evaluate the C1 technology and to compare these statistical methods. This analysis allowed us to validate the technology and the pipeline, but also to highlight their limitations.

[*]Speaker
[†]Corresponding author: keime@igbmc.fr

STARK: A Next-Generation Sequencing data analysis pipeline for clinical diagnosis.

Amandine Velt * [1], Jean Muller [1], Antony Lebéchec *

[2]

[1] Laboratoires de diagnostic génétique, Unité de Génétique Moléculaire, IGMA – Les Hôpitaux Universitaires de Strasbourg (HUS) – France
[2] Institut Régional du Cancer - Alsace (IRC) – Les Hôpitaux Universitaires de Strasbourg (HUS) – France

The great success of Next Generation Sequencing technologies (NGS) due to their capacity to generate millions of base pairs of DNA for an individual patient, has led to the extremely quick transition from research into clinical practice. This large amount of data requires powerful tools for a rapid and robust data analysis and clinical interpretation, with strict guidelines to allow a careful clinical use.
Currently, a plethora of NGS data analysis tools have been made available by research teams and selection of the most reliable and robust tools in automated data analysis in clinical laboratories is a critical task. Here we will discuss STARK (Stellar Tools from raw sequencing data Analysis to variant RanKing) a bioinformatics environment dedicated to NGS data analysis, which aims to provide an interpretative support for clinical diagnosis, with great reproducible results. In this way, STARK is the first analysis pipeline, in France, which has been accredited by COFRAC (COmité FRançais d'ACcréditation). Moreover, STARK is currently being diffused by the INCa (Institut National du Cancer).

Overall, STARK adopts the GATK recommendations, which has rapidly become the standard for identifying SNPs and indels in disease research and clinical applications. STARK performs the demultiplexing, alignment, indels realignment, bam recalibration, variants detection (calling) and variants annotation steps. It takes as input the raw sequencing data (BCL, FASTQ or unaligned BAM) and generates annotated results (VCF) as well as some reports for analysis traceability, notably. It has been developed using shell scripts and uses the Makefiles principle to execute organized rules on the IT resources. Thanks to this, STARK is greatly customizable depending on the IT resources, highly scalable to easily add new data analysis pipelines (with different aligners, callers and annotators), parallelizable and is fully automated.
STARK currently runs, in routine, within four hospital laboratories to identify disease causing variations within patients affected with genetic disorders including rare diseases and cancer (germline and somatic). Due to this specific environment, we will discuss some constraints and issues as well as solutions (IT, clinical, legal) that we have to implement in STARK.

*Speaker

Chromosome conformation dynamics during T cell development

Yousra Ben Zouari * [1], Tom Sexton

[1] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) – CNRS : UMR7104, Inserm : U964, université de Strasbourg – Parc D'Innovation 1 Rue Laurent Fries - BP 10142 67404 ILLKIRCH CEDEX, France

Chromatin conformation capture with high-throughput sequencing (Hi-C) is a widely used technique to study global chromatin organization in vivo (Liberman-aiden et al, 2009) .The analyses of Hi-C data are challenging because of the high complexity of the material which is sequenced, comprising non-specific ligations which can arise from a variety of technical causes. Most Hi-C analyses have focused on identifying biologically significant interactions from a single sample. These analyses require a precise quantitative understanding of the different technical artifacts of the Hi-C experiment. Recently, Hi-C also has been used to study the genome organization changes during cellular development, which makes the analysis even more challenging. Although several studies have performed custom analyses to detect differential interactions from Hi-C data, detecting differential interactions across two or more biological conditions remains a bioinformatic and statistical problem. The resolution of Hi-C has been improved by modifications to reduce the complexity of the material, such as using oligonucleotide pools to enrich for genomic loci of interest (capture-HiC, Schoenfelder et al,2015). However, these approaches require careful analysis to account for further technical biases which are introduced.

Here I will present the different statistical and bioinformatic approaches that we have applied on our capture-HiC datasets in order to detect the changes in genome organization during mouse thymocyte development. In our study, we focus on the potential developmental dynamics of domain borders and enhancer-promoter chromatin loops.

Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009).

Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 25, 582– 597 (2015).

*Speaker

List of participants

- Albrecht Felipe
- Assenov Yassen
- Aubry Marc
- Bardet Anaïs
- Ben Zouari Yousra
- Bhardwaj Vivek
- Blum Matthias
- Bonnet Eric
- Busato Florence
- Carl Sarah
- Chantalat Sophie
- Christiane Bouchier
- Damara Manohar
- Daric Vladimir
- Delepine Marc
- Despons Laurence
- Dormishian Mojdeh
- Dumas Michael
- Etcheverry Amandine
- Feuerbach Lars
- Friedrich Anne
- Geier Florian
- Geoffroy Véronique
- Gounot Jean-Sébastien
- Gourain Victor

- Grabowska Ewa
- Gronemeyer Hinrich
- Grüning Björn
- Jarlier Frédéric
- Jossinet Fabrice
- Jung Matthieu
- Kchouk Mehdi
- Kchouk Mehdi
- Keime Celine
- Kumar Akinchan
- Le Béchec Antony
- Le Gras Stephanie
- Lechner Doris
- Lescure Alain
- Levon Stéphanie
- Malysheva Valeriya
- Marchand Anthony
- Mendoza Parra Marco Antonio
- Meunier Aline
- Misra Nisha
- Molina Nacho
- Mosser Jean
- Naquin Delphine
- Nicolas Kaspric
- Pachchek Sinthuja
- Palomares Marie-Ange
- Peter Jackson
- Pflieger David
- Ramirez Fidel
- Rerra Anna-Isavella
- Rhinn Muriel
- Richer Delphine

- Rivals Eric
- Rustenholz Camille
- Ryan Devon
- Sexton Thomas
- Sikora Katarzyna
- Thomès Luc
- Vagne Constance
- Velt Amandine
- Ye Tao

Author Index