Bioinformatique et Séquençage Haut Débit

24 mars 2010, ENS Paris

Objectif

Les technologies de séquençage à haut débit (SHD) révolutionnent de nombreuses approches expérimentales en biologie moléculaire et évolutive. Leurs applications couvrent des domaines aussi divers que la génomique, l'épigénomique, la transcriptomique, ou encore la métagénomique. Ces applications génèrent toutes des quantités impressionnantes de séquences courtes (reads en anglais), et exigent des traitements bioinformatiques spécifiques, aussi fiables qu'efficaces.

L'objectif de ce colloque est de présenter l'état de l'art dans divers domaines d'applications, ainsi que les questions ouvertes qui se posent pour le futur.

Lieu

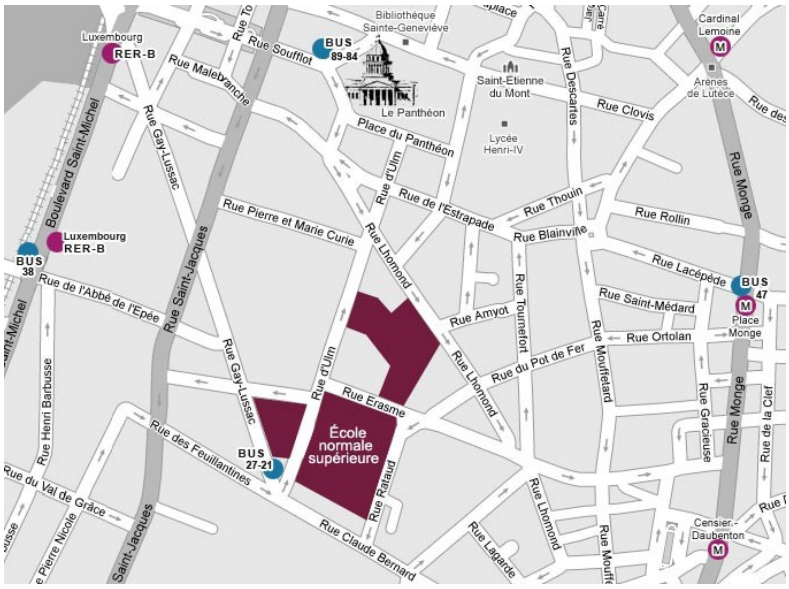
Ecole Nationale Supérieure (ENS)

45 Rue d'Ulm, Paris 5°

Salle Dussane

Site du colloque :

www.lirmm.fr/~rivals/SHD-2010



Comité d'organisation

- Eric Rivals, LIRMM - CNRS, Univ. Montpellier II
- Emmanuel Barillot, Institut Curie, Paris
- Stéphane Robin, INRA - AgroParisTech, Paris
- Hugues Roest Crollius, ENS, Paris

Programme

8h30 : **Accueil des participants**

9h00 : **Les nouvelles technologies de séquençage au Genoscope: assemblage et annotation de génomes**, *Jean Marc Aury*, CNS.

9h50 : **Statistique et métagénome**, *Jean-Jacques Daudin*, AgroParisTech.

10h40 : Pause

11h00: **Analyses de la biodiversité microbienne grâce au séquençage massif des séquences d'ARN ribosomiques. Résultats & problèmes**, *Richard Christen*, Univ. Nice.

11h50: **Discovery and quantification of RNA by RNAseq experiments**, *Roderic Guigo*, CRG, Barcelone.

12h40: Déjeuner

14h30: **Seed design framework for mapping AB SOLiD reads**, *Gregory Kucherov*, INRIA, Lille.

15h20: **High Throughput Transcriptomics**, *Gunnar Raetsch*, Max Planck Institute, Tübingen.

16h10 : Pause

16h30 : **Detection and Annotation of Alternative splicing events with RNA-Seq data**, *Hughes Richard*, Univ. Pierre et Marie Curie, Paris.

17h20 : **Prediction of transcription factor binding sites from ChIP-Seq data through de novo TFBS motif discovery**, *Valentina Boeva*, Curie, Paris.

18h10 : Fin

Résumés

Les nouvelles technologies de séquençage au Genoscope: assemblage et annotation de génomes

Jean Marc Aury, CNS

Les nouvelles technologies de séquençage au Genoscope: assemblage et annotation de génomes

Les nouvelles technologies de séquençage haut débit permettent de générer plusieurs gigabases de séquences par semaine. Le séquençage de 'grands' génomes complets s'accélère et les méthodes utilisées auparavant doivent être adaptées ou revues. Le Genoscope, en tant que centre national de séquençage, est impliqué dans plusieurs projets de séquençage */de novo/* de génomes complets. L'exposé détaillera les méthodes mises en place au genoscope pour l'assemblage et l'annotation de ces grands génomes eucaryotes.

Prediction of transcription factor binding sites from ChIP-Seq data through de novo TFBS motif discovery.

Valentina Boeva, Institut Curie Paris

Next-generation sequencing technologies enabled genome-wide identification of binding sites of DNA-associated proteins. Recently, a number of applications predicting binding sites from ChIP-Seq data have been published [1-4]. One of the major problems in this kind of approaches is the determination of the threshold choice for DNA tag coverage. Generally, the threshold is selected using the False Discovery Rate estimation, which is done either by Monte-Carlo simulation, by using data from a control experiment or by using the Poisson distribution for tag density. Still, our experience showed that regions which have relatively low DNA tag coverage to pass the selection and thus are discarded by most of tools, very often contain binding site motif occurrences just in the area of the peak top coverage. Since the length of the top coverage area is rather small, the probability that a predicted binding site motif is found there by chance is extremely small as well. The observed number of such regions is much greater than expected, so we conclude de that some of these regions should be included in the final peak selection. We proposed an algorithm which serves two ends in ChIP-Seq data analysis: de novo binding site motif identification and binding site extraction without explicit threshold selection. First, one chooses a set of peaks with high DNA tag coverage. Then, one identifies de novo motifs in the top regions of those peaks. Next, using extracted PSSMs one selects peaks with lower DNA tag coverage which contain motifs in their central area. A user defined threshold is set for the total number of expected false positives hits among selected peaks. The algorithm is implemented as a Java package MICSA (Motif Identification for ChIP-Seq Analysis) which is available at our website http://bioinfo-out.curie.fr/projects/micsa/ . The MICSA package was tested on real data for the oncogenic transcription factor EWS-FLI1. Through the comparison with transcriptomic data, dozens of putative direct targets of EWS-Fli1 were discovered.

1. A. Valouev et al. (2008), Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data, Nature Methods, 5(9):829-834.
2. H. Ji et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data, Nature Biotechnology, 26(11):1293-1300.
3. D.A. Nix et al. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks, BMC Bioinformatics, 9(523).

4. A. Fejes et al. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology, Bioinformatics, 24(15):1729-1730.

Analyses de la biodiversité microbienne grâce au séquençage massif des séquences d'ARN ribosomiques. Résultats & problèmes.

Richard Christen, Université de Nice

L'analyse d'amplicons PCR de séquences d'ARN ribosomiques (ARN ou ADN) est la méthode standard dans les analyses de la diversité microbienne (archaea, bacteria et eukaryota). Leur séquençage direct et massif (Roche 454) a remplacé le séquençage Sanger de bibliothèques de clones. Le résultat est l'obtention de millions de séquences en terme de semaines au lieu de l'obtention de centaines de séquences au bout de plusieurs mois. De nombreux projets ont ainsi analysé les compositions microbiennes de l'environnement (eau, sol, tube digestif,...). Les résultats publiés montrent tous une diversité beaucoup plus importante que celle estimée jusque là. Mais de nombreuses incertitudes demeurent.

- Quels sont les biais dus aux erreurs de la PCR et du séquençage, aux amorces de PCR et aux méthodes utilisées pour identifier les espèces à partir des séquences...
- Quel domaine de la molécule doit-on séquencer ?
- L'évolution de la capacité de ces machines est environ un doublement de la longueur séquencée et du nombre de séquences chaque 18 mois ! Comment et sous quelle forme stocker les résultats ? A-t-on les moyens de calcul nécessaires pour analyser les centaines d'échantillons qui vont être produits annuellement ?

La discussion portera sur les réponses actuelles et possibles à ces questions.

Statistical challenges from the analysis of NGS-Metagenomics experiments

Jean-Jacques Daudin, AgroParisTech, Paris, France

All Metagenomics experiments take the same first step: DNA is extracted directly from all the microbes living in a particular environnment (sea, soil, human gut, cheese...). Sequences-based metagenomics captures a massive amount of information on the microbial community under study. Then the tags are filtered for quality, assembled and aligned on a reference gene set containing known genes and unknown ORFs. Then the statistic questions arise: Is the experiment repeatable? What are the sources of variability? How many species or genes were really present? Is there any difference between two conditions ? What about experimental design ? Some of these questions may be solved using standard tools and some of them needs new methods.

Discovery and quantification of RNA by RNAseq experiments

Roderic Guigo, Centre de Regulació Genòmica, Barcelone, Espagne

Transcribed regions have been long been regarded as a distinguishing characteristic of functional portions of the human genome. Massively parallel sequencing of RNAs through next generation sequencing NGS instruments promises, for the first time, sufficient sequencing depth for full transcriptome characterization, that is for the identification of every transcript species in the cell, and their quantification; in particular, for the accurate estimation of the relative abundances of alternative transcript isoforms from the same gene, and of the expression of novel non-coding RNAs. However, the most cost-effective such technologies typically produce very short sequence reads, which compounds transcript reconstruction and quantification. We will discuss computational approaches being developed to address this issue, and produce accurate estimation of transcript quantities in the cell.

Seed design framework for mapping AB SOLiD reads

Gregory Kucherov, INRIA, Lille,

Throughput Transcriptomics

Gunnar Raetsch, Max Planck Institute, Tübingen

Detection and Annotation of Alternative splicing events with RNA-Seq data

Hughes Richard, Univ. Pierre et Marie Curie, Paris

Remerciements

Ce colloque est organisé sous l'égide du GDR Bioinformatique Moléculaire, du réseau ReNaBi (Réseau National de Bioinformatique) et de la Société Française de Génétique. Il est financé et soutenu par :

- le Groupe de Recherche (GDR) CNRS 3003 bioinformatique moléculaire www.gdr-bim.u-psud.fr/ ;
- ReNaBI : Réseau National des plates-formes BioInformatiques www.renabi.fr/ ;
- le LIRMM (Lab. d'Informatique Robotique et Microélectronique de Montpellier) www.lirmm.fr ;
- la Société Francaise de Génétique ;
- l'Ecole Normale Supérieure de Paris www.ens.fr ;
- Institut Curie, Paris www.curie.fr/.