Projet de Recherche Doctoral Concours IPV 2021

Intitulé du Projet de Recherche Doctoral :

Directeur de Thèse porteur du projet (titulaire d'une HDR) :

NOM : **Morillon** Prénom : **Antonin**

Titre : **DR1**

e-mail : Antonin.morillon@curie.fr

Adresse professionnelle : **Centre de recherche institut Curie 26 rue d'Ulm 75005 Paris**

Unité de Recherche :

Intitulé : Dynamique de l'information génétique, bases moleculaires et cancer

Code : UMR3244

Equipe de Recherche *(au sein de l'unité)* **:**

Intitulé : **ARN non codant epigentique et fluidité du génome**

Thématique de recherche : Transcription pervasive, lncRNA, épigénétique

Responsable d'équipe :

NOM : **Morillon** Prénom : **Antonin**

Ecole Doctorale de rattachement de l'équipe & d'inscription du doctorant :
CDV

Doctorants actuellement encadrés par le directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) :
Torrossian, 2019, 50%
Cipolla, 2019, 50%
Andus, 2018, 50%

CO-DIRECTION *(obligatoire)*

Co-Directeur de Thèse (titulaire d'une HDR) :

NOM : **Gautheret** Prénom : **Daniel**

Titre : **PU** HDR

e-mail : daniel.gautheret@universite-paris-saclay.fr

Unité de Recherche :

Intitulé : Institut de Biologie Intégrative de la Cellule

Code: I2BC

Equipe de Recherche *(au sein de l'unité)* **:**

Intitulé : **Séquence, Structure et fonction des ARN**

Thématique de recherche : Bioinformatique, ARN

Responsable d'équipe :Daniel Gautheret

NOM : **Gautheret** Prénom : **Daniel**

Ecole Doctorale de rattachement :

Ou si ED non SU : **SDSV, Univ Paris-Saclay**

Doctorants actuellement encadrés par le co-directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) :

Haoliang Xue, 2018, 100% (3e année)
Yunfeng Wang, 2018, 50% (3e année)

Cotutelle internationale : Non Oui, précisez Pays et Université :

Précisez ici les éventuels co-encadrants (non HDR)
(cotutelle industrielle pour doctorant Yunfeng Wang)

Co-encadrant :

NOM : **DU** Prénom : **Yang**

Titre : **PhD** HDR

e-mail : yangdu@annoroad.com

Affiliation: Annoroad Technologies, Inc, Beijing

Résumé (2 000 caractères maximum) :

Exploring tissues and cell populations heterogeneity, to **discover cellular identities, reconstruct cell developmental trajectories and infer regulatory networks** is now possible through the rapid development of single-cell RNA sequencing (scRNA-seq). The description of malignant cell heterogeneity and the identification of clonal cell populations associated with resistance to therapy has become a major challenge in personalized medicine that led to the development of numerous computational tools. Most (sc)RNA-seq gene expression analysis pipelines use an external genome reference to infer the genomic location of the reads. The need for an external reference hinders the detection of events that are not annotated and thus misses key information to discriminate cells. This is particularly true for cancer cells, where many genomic rearrangements/mutations and genetic/ epigenetic rewiring are common. Reference-free approaches were developed to tackle this problem, by decomposing reads in k-mers (sequence of length k), and then directly comparing k-mers across

samples. However, while we were successful in applying k-mer strategies to compare differentiated cells (Audoux et al 2017) or patients (Pinskaya et al, 2019) in bulk RNA samples, application to single-cell analysis require major developments. In this project, we propose to **develop a novel reference-free method for scRNA-seq analysis and to apply it to discover novel predictive signatures of cell types that will be used to explore tumor heterogeneity associated with cancer progression**. This interdisciplinary project is at the interface of 2 teams, the group of **A.Morillon** (I. Curie), expert in assessing functional significance of lncRNA and **D. Gautheret** (I2BC) expert in non coding RNA bioinformatics and predictive signatures. In addition, the project benefits of close collaborations with the single-cell platform (lead by C. Vallot) at Curie who provide scRNA transcriptomes in tumoral and cellular contexts, excellent models for testing and challenging our **experimental machine learning procedures**.

Joindre en annexe un descriptif du PRD avec références au format pdf (« NOM_2_IPV_2021 » / 3 pages maximum, taille police 11)

AVIS et VALIDATION de l'ECOLE DOCTORALE :

à envoyer simultanément par e-mail à l'ED de rattachement et au programme : interfaces_pour_le_vivant@listes.upmc.fr avant le lundi 15 février minuit.

Reference free classification of single-cell RNA-seq data to reveal novel cell identities in tumors

Introduction

Understanding biological systems requires the characterization of their components, ultimately at the cell level. Technological advances of the last 20 years now allow genome-wide molecular profiling at this level, for hundreds or thousands of cells in parallel. The colossal amount of data that has been produced has led molecular biology into the "big data" era, and research may now benefit from more data-driven approaches (Wagner et al, 2016). Single-cell RNA sequencing (scRNA-seq) technologies can "read" RNA content of individual cells, and vary in the type of RNA they target, the sequencing depth and transcript coverage they achieve or the number of cell they can process in parallel (Svensson, 2017).

This new level of resolution has allowed to explore tissues and cell populations heterogeneity, to discover new cell types, to reconstruct cell developmental trajectories and infer regulatory networks (Chen et al, 2019). Applied to tumor, it revealed patterns in malignant cell heterogeneity across patients and led to the description of cell populations that may be associated with tumor development and resistance to therapy (Suvà et al. 2019).

Because of the small amount of material available per cell for sequencing, scRNA-seq data are more affected by technical noise than "bulk" RNA-seq (population level). Also, as cellular heterogeneity comes from multiple process occurring at the same time, it might be useful to distinguish the relative impact of these processes (eg. cell-cycle and cell differentiation). Finally, the high dimensionality of single-cell transcriptomic data generally requires feature (gene) selection and dimension reduction before further analysis. The particular challenges of scRNA-seq data has led to the development of numerous tools to address them (Andrews et al, 2018).

Most (sc)RNA-seq analysis pipelines need an external genome reference to infer the genomic location of the reads - short (or long) nucleotidic sequences – obtained from sequencing technologies, and a map of the genes location (annotation) to quantify gene expression. Since the human genome is widely transcribed – resulting in widespread expression of non-coding RNA (Djebali et al, 2012) - lots of efforts were made to enrich such maps and describe the different classes of transcripts and their role in biological processes and diseases (Hombach and Kretz, 2016). Today, the tools developed allow to efficiently reconstruct transcripts from (sc)RNA-seq experiments, and anyone can analyze data with tailored annotation (Shao and Kingsford, 2017 ; Nip et al, 2020).

The need for an external reference may hinder the detection of events that are not able to map to the reference (mutation, post-transcriptional event, genomic rearrangements, repeat elements, novel genes) and thus miss information that may be valuable to discriminate between cells. This is particularly true for cancer cells, where many genomic rearrangements happen. Reference-free approaches were developed to tackle this problem, first by decomposing sequenced reads in k-mers (nucleotidic sequence of length k), and then directly (or after assembly) comparing them across samples. They were able to detect new transcriptomic events in EMT cell-line or find markers that efficiently discriminate between several cancer sub-types (Audoux et al, 2017 ; Thomas et al., 2019; Lorenzi et al, 2020; Nguyen et al, 2020). Also, our lab has used this approach to detect novel lncRNA markers in prostate cancer (Pinskaya et al, 2019).

However, so far the k-mer methodology enables the comparison between conditions at the population level only and require further development to study biological systems at the cell level, where well-described cell types are mixed with cells in transient states or unknown types, such as heterogeneous tumors.

In this project, we propose to develop a novel pipeline for a reference-free analysis of single-cell RNA-seq data, and with it explore tumor heterogeneity to reveal novel cell sub-populations involved in cancer progression.

Aims

1) The first aim of this project will be to adapt reference-free analysis tools designed for bulk RNA-seq to make them compatible with major single-cell RNA-seq applications (cell-classification, clustering etc.).

2) The second aim of this project will be to assess the performance of reference-free scRNA-seq approaches for cell type clustering and definition of molecular signature.

3) The third aim of this project will be to apply this new approach to an original breast cancer tumor scRNA-seq dataset, to analyze tumor evolution at the cell level upon treatment.

Methodology

1) The first 9 months of the project will be dedicated to the adaptation of the existing scRNA-seq pipelines to k-mer counts.

scRNA-seq pipelines generally begin with a gene expression matrix, with as many rows as genes (about 20,000) and columns as cells (from hundreds to thousands). As k-mer count for a single RNA-seq library can reach 10^9, we will need to integrate the strategies developed for reference-free analysis of bulk RNA-seq into a tools that can efficiently handle this high number of count for hundreds or thousands of cells, and perform feature selection to keep the informative data. We will make use of the method of Lorenzi et al (2020), based on KMC3 (Kokot et al, 2017), that enabled them to work with more than a thousand of libraries, on the work of Nguyen et al (2020) for merging overlapping k-mer, which showed to reduce k-mer count per library of one order of magnitude, along with more classical feature selection used in scRNA-seq pipelines (based on mean/variance relationship for example).

Gene count distribution is generally assumed to follow a negative binomial distribution or be zero-inflated, and dimension reduction and clustering approaches may rely on this assumption. We will then need to assess to what extent k-mer count can be taken as gene count regarding statistical distribution or whether we will have to adapt existing tools.

2) The next 9 month will be used to assess the ability of a reference-free approach to efficiently cluster cell types/states and extract molecular signatures that can identify them.

Many methods for dimension reduction and clustering were developed. As in Duò et al (2018) or Qi et al (2020), we will compare several unsupervised clustering methods to assess which ones are best fit for k-mer counts. We test “traditional” ones such as SC3 (Kiselev et al, 2017), based on consensus clustering or the one implemented in Seurat v3 (Satija et al, 2015), which use K-nearest neighbor graph and modularity optimization (Blondel et al, 2008), along with deep learning approaches (Li et al, 2020).

From the cluster obtained, we will extract k-mer signature that characterize each of these groups, using random forest algorithm or Bayesian method.

Once obtained the k-mer signature, we will test whether we can use them to classify cells from similar tissus. We will use the method developed by Nguyen et al (2020) to retrieve k-mer signature from one data-set to another.

To assess the efficiency of the whole process, we will use 2 data-set for which we know cell type labels of similar tissues. One will be used as training set and the other as test set, for example the human PBMC data that Stuart et al (2019) used for their integration analysis.

3) The last part of the project will be the application of the reference-free approach to analyze breast cancer response to therapy.

In breast cancer patient-derived xenograft (PDX), Grosselin et al (2019) or Marsolier et al (2021) were able to find cells sharing chromatin states - before and after chemotherapy - for genes that play a key role in breast cancer and drug resistance. It reinforces the idea that cell population existing prior to treatment and responsible for drug resistance can be detected and described at the molecular level to help the development

of new treatment strategies and diagnostic/prognostic. We will perform scRNA-seq on breast cancer PDX, with or without treatment, compare cell population molecular signatures in both conditions to find cell populations related to drug resistance, and the associated transcripts, including novel unreferenced RNAs.

For validation, we will computationally extract the best unreferenced RNA marker candidates that identifies resistant cells and then use smFISH to visualize them in untreated and treated samples. This will be done in collaboration with the Morillon and Vallot's teams and the Plateforme d'Imagerie of Institut Curie

Environment

The candidate will be supervised by Antonin Morillon (UMR3244, I. Curie), expert in lncRNA and non coding transcriptomics, at the Institut Curie, which host a sequencing platform and a computational biology unit, and by prof Daniel Gautheret (I2BC), who developed DE-kupl (Audoux et al, 2017) & KaMRaT (Nguyen et al, 2020).

References

1. Andrews TS, Hemberg M. Mol Aspects Med. 2018 Feb;59:114-122. doi: 10.1016/j.mam.2017.07.002. Epub 2017 Jul 25. PMID: 28712804.
2. Audoux J, Philippe N, Chikhi R, Salson M, Gallopin M, Gabriel M, Le Coz J, Drouineau E, Commes T, Gautheret D. Genome Biol. 2017 Dec 28;18(1):243. doi: 10.1186/s13059-017-1372-2. PMID: 29284518; PMCID: PMC5747171.
3. Blondel V.D., Guillaume J.L., Lambiotte R. and Lefebvre, E. Journal of Statistical Mechanics: Theory and Experiment. Doi:10.1088/1742-5468/2008/10/p10008

4. Chen G, Ning B, Shi T. Front Genet. 2019 Apr 5;10:317. doi: 10.3389/fgene.2019.00317. PMID: 31024627; PMCID: PMC6460256.

5. Duò A, Robinson MD, Soneson C. F1000Res. 2018 Jul 26;7:1141. doi: 10.12688/f1000research.15666.3. PMID: 30271584; PMCID: PMC6134335.

6. Djebali S, Davis CA, Merkel A, Dobin A, et al. Nature. 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233. PMID: 22955620; PMCID: PMC3684276.
7. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, Dahmani A, Lameiras S, Reyal F, Frenoy O, Pousse Y, Reichen M, Woolfe A, Brenan C, Griffiths AD, Vallot C, Gérard A. Nat Genet. 2019 Jun;51(6):1060-1066. doi: 10.1038/s41588-019-0424-9. Epub 2019 May 31. PMID: 31152164.
8. Hombach S, Kretz M. Adv Exp Med Biol. 2016;937:3-17. doi: 10.1007/978-3-319-42059-2_1. PMID: 27573892.
9. Kim KT, Lee HW, Lee HO, Kim SC, Seo YJ, Chung W, Eum HH, Nam DH, Kim J, Joo KM, Park WY. Genome Biol. 2015 Jun 19;16(1):127. doi: 10.1186/s13059-015-0692-3. PMID: 26084335; PMCID: PMC4506401.
10. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. Science. 2014 Feb 14;343(6172):776-9. doi: 10.1126/science.1247651. PMID: 24531970;
PMCID: PMC4412462.
10. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. Nat Methods. 2017 May;14(5):483-486. doi: 10.1038/nmeth.4236. Epub 2017 Mar 27. PMID: 28346451; PMCID: PMC5410170.
11. Kokot M, Dlugosz M, Deorowicz S. Bioinformatics. 2017 Sep 1;33(17):2759-2761. doi: 10.1093/bioinformatics/btx304. PMID: 28472236.
12. Lorenzi C, Barriere S, Villemin JP, Dejardin Bretones L, Mancheron A, Ritchie W. Genome Biol. 2020 Oct 13;21(1):261. doi: 10.1186/s13059-020-02165-2. PMID: 33050927; PMCID: PMC7552494.
13.Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, Susztak K, Reilly MP, Hu G, Li M. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020 May 11;11(1):2338. doi: 10.1038/s41467-020-15851-3. PMID: 32393754; PMCID: PMC7214470.
14. Marsolier J., Prompsy P., Durand A., Lyne A.-M., Landragin C. , Trouchet A. , Tenreira Bento S., Eisele A., Foulon S., Baudre L., Grosselin K., Bohec M., Baulande S., Dahmani A., Sourd L., Letouzé E., Marangoni E., Perié L. and Vallot C. Biorxiv. 2021. doi:
https://doi.org/10.1101/2021.01.04.423386
15. Ha Nguyen, Haoliang Xue, Virginie Firlej, Yann Ponty, Mélina Gallopin, et al. 2020. ⟨hal-02948844⟩
16. Nip KM, Chiu R, Yang C, Chu J, Mohamadi H, Warren RL, Birol I. Genome Res. 2020 Aug;30(8):1191-1200. doi: 10.1101/gr.260174.119. Epub 2020 Aug 17. PMID: 32817073; PMCID: PMC7462077.
17. Pinskaya M, Saci Z, Gallopin M, Gabriel M, Nguyen HT, Firlej V, Descrimes M, Rapinat A, Gentien D, Taille A, Londoño-Vallejo A, Allory Y, Gautheret D, Morillon A. Life Sci Alliance. 2019 Nov 15;2(6):e201900449. doi: 10.26508/lsa.201900449. PMID: 31732695; PMCID: PMC6858606.
18. Qi R, Ma A, Ma Q, Zou Q. Brief Bioinform. 2020 Jul 15;21(4):1196-1208. doi: 10.1093/bib/bbz062. PMID: 31271412; PMCID: PMC7444317.
19. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Nat Biotechnol. 2015 May;33(5):495-502. doi: 10.1038/nbt.3192. Epub 2015 Apr 13. PMID: 25867923; PMCID: PMC4430369.
19. Severson DT, Owen RP, White MJ, Lu X, Schuster-Böckler B. Nat Commun. 2018 Mar 22;9(1):1187. doi: 10.1038/s41467-018-03608-y. PMID: 29567991; PMCID: PMC5864873.
20. Shao M, Kingsford C. Nat Biotechnol. 2017 Dec;35(12):1167-1169. doi: 10.1038/nbt.4020. Epub 2017 Nov 13. PMID: 29131147; PMCID: PMC5722698.
21. Suvà ML, Tirosh I. Mol Cell. 2019 Jul 11;75(1):7-12. doi: 10.1016/j.molcel.2019.05.003. PMID: 31299208.
22. Svensson V, Vento-Tormo R, Teichmann SA. Nat Protoc. 2018 Apr;13(4):599-604. doi: 10.1038/nprot.2017.149. Epub 2018 Mar 1. PMID: 29494575.
23; Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Cell. 2019 Jun 13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031. Epub 2019 Jun 6. PMID: 31178118; PMCID: PMC6687398.
23. Thomas A, Barriere S, Broseus L, Brooke J, Lorenzi C, Villemin JP, Beurier G, Sabatier R, Reynes C, Mancheron A, Ritchie W. Commun Biol. 2019 Jun 20;2:222. doi: 10.1038/s42003-019-0456-9. PMID: 31240260; PMCID: PMC6586863.
24. Wagner A, Regev A, Yosef N. Nat Biotechnol. 2016 Nov 8;34(11):1145-1160. doi: 10.1038/nbt.3711. PMID: 27824854; PMCID: PMC5465644.