Séminaire – Math For Genomics

Séance du mercredi 6 février 2019. 14h.
Evry. IBGBI. LaMME.

Stem cells

Characterization of human hematopoietic stem cells through in vivo tracking of integration sites in gene therapy patients

Emmanuelle Six[1,2], Adeline Denis[1,2], Marina Cavazzana[1,2,3], Agathe Guilloux[4]

1. INSERM UMR 1163, Laboratory of Human Lymphohematopoiesis, Paris, France
2. Paris Descartes–Sorbonne Paris Cité University, Imagine Institute, Paris, France
3. Biotherapy Clinical Investigation Center, Groupe Hospitalier Universitaire Ouest, AP-HP, INSERM, Paris, France
4. LaMME, CNRS, Evry University, Paris-Saclay University, Evry, France

Emmanuelle SIX (Institut Imagine)

Introduction of a therapeutic gene into hematopoietic stem and progenitor cells (HSPC) and transplantation in patients with inherited hematological disorders is a successful strategy to restore functional immune cells. In these gene therapy trials, the lentiviral therapeutic vector (self-inactivated) integrates into the genome at unique positions in each HSPC and is consequently transmitted to its progeny, offering a unique opportunity to understand human hematopoiesis. Through the identification of integration sites (IS) in the various blood cell subsets, it is possible to reconstitute the progeny of the various repopulating HSPCs and hence, to address the question of their potential and heterogeneity. In the context of human gene therapy trials for Wiskott–Aldrich syndrome (WAS) and beta-hemoglobinopathies performed in children and young adults conducted at the Biotherapy department of the Necker Hospital, we have already isolated and identified more than 100 000 retroviral integration sites (IS) in the genome of circulating blood cells. We sorted peripheral blood samples for 5 cell types: myeloid (granulocytes and monocytes) and lymphoid subpopulations (T, B and NK cells), and analysed their IS profile using the INSPIIRED pipeline. Clonal abundance was quantified using the sonicAbundance method which provided an estimate of the numbers of cells contributing to each IS cell clone, and therefore allow robust analysis of HSPC function in term of clonal lineage output. However, the data generation process comprises several steps, which add uncertainty to determining the true population compositions. One challenge comes from sparse sampling, that we take into account by applying stringent abundance filtering (allowing 95% retrieval of IS between replicates). We also corrected IS data for residual contamination and unbalanced sampling, then quantified cell lineage output. Cluster analysis of long-term hematopoietic stem cells (HSC) (focused on later time points) revealed the existence of several human HSC subsets with distinct lineage potential: myeloid-dominant, lymphoid-dominant, and balanced HSC subsets, that are detected up to 4-5 years in the two types of genetic diseases. Our results thus highlight the heterogeneity of human HSCs and introduced a novel, rigorous approach for tackling the technical challenges associated with the use of IS data for human HSC lineage ouput tracking in gene therapy trials.

Agathe GUILLOUX (LaMME, Université d'Evry)

Pour répondre au problème biologique : on assimilera le potentiel d'une cellule souche au vecteur de proportions sur les 5 types cellulaires dans sa descendance. La question biologique peut alors être traduite en un problème de clustering (classification non-supervisée) : si nous parvenons à montrer l'existence de clusters de potentiels dans les données, nous pourrons conclure qu'il existe des sous-types de cellules souches aux potentiels différents. Du point de vue de l'apprentissage statistique, il s'agit donc de développer des algorithmes de clustering pour des données dites compositionnelles (la somme des valeurs sur chaque ligne de données vaut 1). Il faut donc adapter les algorithmes classiques à ces données dans le simplexe. Dans la littérature, on trouve des transformations des données (CLR, ou logCLR) qui permettent d'utiliser l'algorithme du kmeans. Il existe également des modèles de mélange (Dirichlet Multinomiale ou Dirichlet généralisé Multinomiale) que nous présenterons. Nous monterons les limites de ces algorithmes quand les composantes multinomiales ont des paramètres situés sur la frontière du simplexe (présence de proportions nulles). Nous présenterons les résultats obtenus sur des données synthétiques et sur les données réelles.