Dimension Reduction Approches for the Statistical Analysis of NGS data

Next Generation Sequencing (NGS) technologies are profoundly impacting many domains of biomedical research, ranging from basic biology to translational research and soon to personalized medicine. Rapid technological advances have dramatically increased the throughput of sequencers, while sequencing costs are falling down spectacularly. This already has a strong impact in terms of developing personalized medical treatments based on individual genomic backgrounds. NGS raises huge challenges to store, analyze and exploit the unprecedented wealth of data produced by current and upcoming NGS machines. New methodological and computational developments are needed to elucidate the multiscale structures of genomes, transcriptomes and epigenomes, their relationships, and their variations across individuals, from NGS data. The particular nature of the data and the complexity and variety of problems addressed calls for new statistical and machine learning approaches, while the scale of data produced by NGS experiments, which easily reaches Terabytes and continuously increases, calls for computationally extremely efficient procedures. Although the field of NGS Bioinformatics is also moving quickly, it is still in its infancy with many problems remaining unsolved or calling for better solutions than those currently available.

One major statistical challenge raised by NGS data is the ultra-high dimension which refers to the explosion of the number of recordings to be compared with a moderate/low number of individuals (NGS putatively deals with as many recordings as genomic positions). This curse of dimensionality requires the development of new statistical methods even for standard questions like clustering and classification. Lasso-type methods based on L1 penalization have received enormous success these past years, due to their joint computational and statistical efficiencies. Among different strategies, fused-lasso penalties have been defined to control for sparsity for spatially organized data. **The development of lasso and fused-lasso methods in the context of aligned-based NGS data is the central challenge of this PhD project.** NGS data are counts that can be over-dispersed, which makes Generalized Linear Models an appropriate framework for this purpose. Another possible research direction of the project is to develop penalized versions of Partial Least Square (PLS) methods. PLS is widely used for efficient dimension reduction by compressing variables on the basis of an empirical covariance criterion. PLS-Lasso strategies would be an interesting direction to compress and select relevent biological features based on NGS data.

This project will be part of the ABS4NGS project recently selected by the "investissement d'avenir" call. This project gathers a consortium of Algorithmicians, Bioinformaticians, Statisticians, and Biologists funded for 4 years to develop mathematical approaches for the analysis of NGS data. The student will be based at the LBBE, Lyon, and will work in collaboration with Sophie Lambert-Lacroix (TIMC-UPMF, Grenoble), Vivian Viallon (IFSTAR-ICJ, Lyon) and Franck Picard (LBBE, Lyon).