HEIBRiDS Lecture Series – Wednesday 21st November, 14.00 - 17.00 @ Einstein Center Digital Future

Programme

Location: Room 009 (Only for HEIBRiDS PhD students)

14:00 – 15:15 Personality assessment tests

15:15 – 15:45 TEDx talk by Jorge Cham on public dissemination of science, followed by round table discussion

15:45 – 16:00 Coffee break

Location: Room 104/105/106 (Open to HEIBRiDS supervisors)

16:00 – 16:30 Machine Learning for understanding tumour evolution
Speaker: Roland Schwarz, MDC Berlin (see next page for **Abstract**)

16:30 – 17:00 Binning directories: An efficient data structure to query k-mers in partitioned string sets
Speaker: Knut Reinert, FU-Berlin (see next page for **Abstract**)

Next Lecture Series: Wednesday, December 5th

Abstract 1

Machine Learning for understanding tumour evolution

Cancers are characterised by ubiquitous somatic alterations, including single-nucleotide variants (SNVs), short insertions and deletions, structural variants and somatic copy-number alterations (SCNAs) that contribute to intra-tumour heterogeneity (ITH). ITH has been identified as the major cause for treatment failure in the clinic, yet most cancer sequencing projects have focused on the identification of somatic driver SNVs that are shared between samples and patients.

Our group was one of the first to demonstrate that SCNAs rather than SNVs affect resistance development and patient outcome in high-grade serous ovarian cancer (Schwarz, et al., 2015). This was enabled through dedicated machine learning algorithms for the quantification of ITH from SCNA data (Schwarz, et al., 2014). I will report on these efforts, as well as subsequent developments that confirmed SCNAs as the major driving force behind genome evolution in non-small cell lung cancer (Jamal-Hanjani, et al., 2017; Abbosh, et al., 2017). I will demonstrate how allele-specific SCNA profiling of multiple tumour regions can detect convergent evolution in clinical samples with direct effects on patient survival and will show how these findings extend to a pan-cancer setting. Lastly, I will report on our recent efforts tracking the functional implications of both SCNA and SNV heterogeneity on the cancer transcriptome through whole-genome and RNA-sequencing of 1200 clinical samples in the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (Calabrese, et al., 2017).

References

Abbosh, C., Birkbak, N. J., Wilson, G. A., Jamal-Hanjani, M., Constantin, T., Salari, R., . . . Swanton, C. (2017, 4). Phylogenetic ctDNA analysis depicts early stage lung cancer evolution. *Nature*. doi:10.1038/nature22364

Calabrese, C., Lehmann, K.-V., Urban, L., Liu, F., Erkek, S., Fonseca, N., . . . Stegle, O. (2017). Assessing the Gene Regulatory Landscape in 1,188 Human Tumors. *bioRxiv*. doi:10.1101/225441

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B., Veeriah, S., . . . Consortium, T. (2017, 4). Tracking the Evolution of Non-Small-Cell Lung Cancer. *The New England journal of medicine*. doi:10.1056/NEJMoa1616288

Schwarz, R. F., Ng, C. K., Cooke, S. L., Newman, S., Temple, J., Piskorz, A. M., . . . Brenton, J. D. (2015, 2). Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS medicine, 12*(2), e1001789. doi:10.1371/journal.pmed.1001789

Schwarz, R. F., Trinh, A., Sipos, B., Brenton, J. D., Goldman, N., & Markowetz, F. (2014, 4). Phylogenetic quantification of intra-tumour heterogeneity. *PLoS computational biology, 10*(4), e1003535. doi:10.1371/journal.pcbi.1003535

Abstract 2

Binning directories: An efficient data structure to query k-mers in partitioned string sets

Motivation: Mapping-based approaches have become limited in their application to very large sets of references since computing an FM-index for very large databases (e.g. >10 GB) has become a bottleneck. This affects many analyses that need such index as an essential

step for approximate matching of the NGS reads to reference databases. For instance, in typical metagenomics analysis, the size of the reference sequences has become prohibitive to compute a single full-text index on standard machines. Even on large memory machines, computing such index takes about 1 day of computing time. As a result, updates of indices are rarely performed. Hence, it is desirable to create an alternative way of indexing while preserving fast search times.

Results: To solve the index construction and update problem we propose the DREAM (Dynamic seaRchablE pArallel coMpressed index) framework and provide an implementation. The main contributions are the introduction of an approximate search distributor via a novel use of Bloom filters. We combine several Bloom filters to form an interleaved Bloom filter and use this new data structure. called binning directories, to quickly exclude reads for parts of the databases where they cannot match. This allows us to keep the databases in several indices which can be easily rebuilt if parts are updated while maintaining a fast search time.