Joint copy-number segmentation and phasing from multi-region cancer sequencing data

RF Schwarz, MDC Berlin, K Reinert, FU Berlin

Somatic copy-number alterations (SCNAs) are a hallmark of most aggressive tumours and a main contributor to intra-tumour heterogeneity (ITH). The resulting ITH is the main driving force behind tumour evolution and progression, including resistance development to chemotherapy and targeted treatments.

To detect SCNAs, traditional cancer genomics studies follow a one-sample-per-patient strategy in combination with short-read sequencing. After read mapping, allele-specific SCNAs can be detected per sample separately by accumulating read counts at heterozygous germline variants. Computing the total number of reads at a variant site compared to a matched normal tissue (log-ratio, logR) as well as the ratio of alternative over reference allele read counts (B-allele frequency, BAF) allows quantification of allele-specific integer copy-numbers (Figure 1).

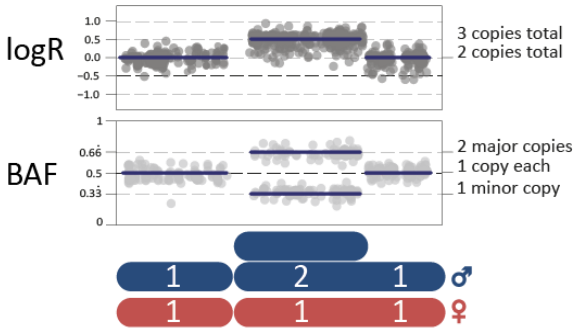

Figure 1: In regions of SCNAs, the BAF deviates from 0.5 as the ratio of available alleles changes. Gains and losses are also reflected in the total read depth (logR).

We[1,2,3] and others have recently demonstrated the power of multi-region sequencing approaches to infer tumour evolution and detect selection acting on cancer genomes. There, multiple samples are acquired and sequenced from the same patient at spatially or temporally separate sites. As the germline background of the patient is constant, multi-region sequencing increases the power to detect SCNAs over single-sample studies by allowing us to determine the phasing of the underlying germline variants. This phasing information in turn increases detection power for SCNAs by enabling the aggregation of read counts over the parental haplotypes within a copy-number altered genomic window. So far, no algorithm exists for joint inference of the most likely haplotype structure (phasing) and most likely SCNA segmentation.

We propose development of such an algorithm based on efficient read mapping and multiple linked Hidden-Markov-Models (HMMs) in the SeqAn library of efficient data types and algorithms.

The Schwarz lab has clinical multi-region data available which is segmented using state-of-the-art single sample methods for reference and benchmarking whereas the Reinert lab has expertise in developing and implementing efficient methods for NGS analysis. Depending on prior knowledge, the candidate will familiarise him-/herself with genomics algorithms including read mapping, genotyping and SCNA calling. He/She will then extend existing prototype implementations in the SeqAn library to achieve the multi-region segmentation tool, followed by data analysis and reconstruction of the evolutionary relationship of tumour clones within a patient.

Prerequisites: Thorough knowledge of HMMs and/or probabilistic graphical models in general; fluent knowledge of C++ and Python; strong interest in cancer biology; good team and communication skills for interactions within the lab and with clinical collaboration partners.

---

[1] *Tracking the Evolution of Non-Small-Cell Lung Cancer.* **Jamal-Hanjani M, et al.** 2017, NEJM.

[2] *Spatial and temporal heterogeneity in high-grade serous ovarian cancer […].* **Schwarz RF, et al.** 2015, PLoS Medicine

[3] *Phylogenetic quantification of intra-tumour heterogeneity.* **Schwarz RF, et al.** 2014, PLoS Comp. Biol., Vol. 10