Novel cell-free genomic sequencing data: Technique, Application and Analysis of Broad Range cell-free DNA sequencing

Neeti Swarup, Irene Choi, Jordan Cheng, Akanksha Arora, Mohammad A Aziz, David T. W. Wong

Cell free genomic analysis of various biofluids including blood, saliva, urine, or CSF is ushering in a new era of non-invasive liquid biopsy. Previously, mutated sequences were the primary feature in cell-free DNA (cfDNA) for determining mutation burden, residual cancer monitoring, or drug section. However, mutated sequences are often rare and difficult to find amongst DNA released from non-tumor cells. More recently, other non-mutational features have been identified within the cfDNA which are promising for cancer detection. Some features include, copy number, human genome mapping origins, and cfDNA fragment length ratio.

To the ever-expanding landscape of cfDNA our group has established a novel workflow: Broad Range cell-free DNA sequencing (BRcfDNA-Seq) which has introduced a new population of cfDNA which is ultrashort (40-70bps) and single-stranded (uscfDNA). The uscfDNA is recovered in addition to conventionally observed mononucleosomal cfDNA (~167 bps) and shorter pieces of cfDNA (71-119bps), short cfDNA (scfDNA). Isolation of these novel populations of cfDNA opens newer avenues for analysis of recently described non-mutational features. Our group has also described a bioinformatic pipeline for processing the fastq.gz files to human-aligned .bam files and then categorizing into different size-based populations for assessment.

At present, we have employed features like fragmentomic ratios, end motif diversity, functional genomic elements, differential exonic counts, differential counts across chromosomal bins, and occurrences of G-Quadruplex complexes of the three cfDNA populations to differentiate cancer from non-cancer samples. These features yield a very large number of quantitative values for each metric. As a proof of concept, we have employed these features individually and collectively to differentiate plasma derived from lung cancer patients from non-cancer controls; oral premalignant lesions which progressed to cancer, and which did not; malignant indeterminate pulmonary nodules (IPN) samples from benign IPN samples; saliva from gastric cancer patients and non-cancer controls.

When utilizing multiple features derived from BRcfDNA-Seq we tested several learning models (random forest and XGBoost) for differentiating the 2 groups. We opted for the investigation of learning techniques that allow for the prevention of overfitting, structured data, and simple learning. Through different learning techniques coupled with the BRcfDNA-Seq datasets across various cancers, we believe this novel dataset will yield a more robust differentiation between malignant and benign disease. Additionally, apart from utilizing quantitative values from described features to detect cancer, the emergence of genomic large language models (LLMs) could allow us to use .bam files from different categories of cfDNA in order to identify undiscovered features such as base pair entropy and develop new classifiers based on these features.

Both the BRcfDNA-Seq (downstream metrics) and LLM approaches output large datasets wherein employing learning techniques like random forest or XGBoost on identified metrics or LLM-engineered features to discover highly differentiating signatures of malignant disease could significantly improve the clinical utility and viability of BRcfDNA-derived cfDNA biomarkers.