Sequencing technologies have revolutionized the study of genomes by generating high throughput data for various studies which are not cost-efficient when done with Sanger sequencing. The first step in analyzing these high throughput data is often to find the original location from which the data reads are sequenced from a reference genome. Moreover, references genomes can be very large (human genome ~3.2GB). This calls for better methodologies in aligning reads onto a reference genome.

In this presentation, we present three methodologies in producing accurate alignments of DNA-sequencing reads with bisulfite-induced nucleotide conversion, DNA-sequencing reads with mismatches and gaps, and RNA-sequencing reads with intronic spliced junctions. 

Our first contribution is BatMeth; a fast, sensitive and accurate aligner for DNA-sequencing reads derived from sodium bisulfite treatment. BatMeth can handle both base-space and color-space bisulfite-treated reads. Our method was able to avoid examining spurious hits and improve the efficiency and specificity of our alignments. Our experiments also showed that BatMeth can produce better methylation callings across samples of different bisulfite conversion rates.

BatAlign is our next contribution which can align DNA-sequencing reads in the presence of both mismatches and insert-delete (indel) accurately. Two novel strategies called Reverse-Alignment and Deep-Scan are developed to enable the efficient reporting of accurate alignments for these reads. Reverse-Alignment starts the alignment of a read by looking for the most probable preliminary alignments incrementally. Deep-Scan refines the preliminary alignments by searching for a targeted subset of less probable alignments to better distinguish the best alignment from the rest. BatAlign was able to achieve competitive runtime efficiency with SIMD-SSE2 implementations of the Smith-Waterman algorithm for the extension of seeds from a long read in our seed-and-extend strategy.
Our last contribution is BatRNA which is designed to do spliced alignment of RNA-sequencing reads accurately. As RNA-sequencing datasets can have very varying mixture of exonic and spliced reads in them, BatAlign was introduced in BatRNA as a pre-mapping tool to draft up the possible spliced sites of the genome. After which, we filtrate the reads from the mappings of BatAlign to be mapped by BatRNA for possible spliced alignments of the reads. The resultant mappings from both BatAlign and BatRNA are considered for the final alignment of a read. Compared with other popular RNA-sequencing aligners, BatRNA was able to produce very sensitive and accurate alignments in a simulated and real RNA-seq dataset, while maintaining competitive runtimes.

In summary, we have developed various improved methodologies to align reads on to a reference genome, sequenced from various genomic origins.

Jing-Quan first authored BatMeth and BatAlign papers. They are published in Genome Biology (Impact: 10.5) and NAR journals (Impact: 8.8).