FastqBLAST

a tool to quickly assess the organismal and genomic diversity present in a FASTQ file

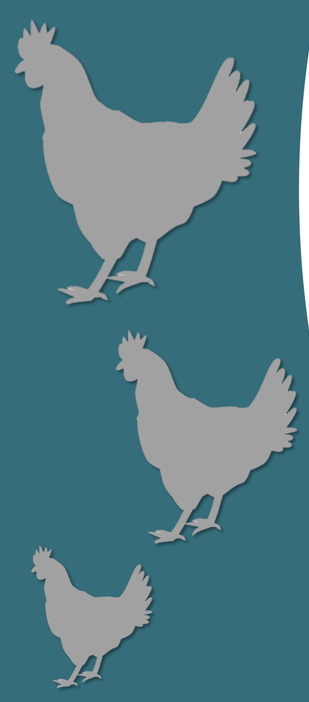
Juniper A. Lake, M. Joseph Tomlinson IV, Behnam Abasht
University of Delaware

background

FastqBLAST was developed to help diagnose issues of low mapping rates of next generation sequencing reads (RNA- and DNA-seq) to a reference genome. Users of alignment tools such as TopHat and STAR may encounter mapping rates for particular samples or tissues that are substantially lower, but have difficulty identifying the specific cause of the discrepancy. Often the first step in addressing the problem is to run a sample of the unmapped reads against NCBI's databases using the well-known BLAST tool. However, there is currently no easy way to BLAST large samples of reads and summarize the results for easy diagnosis of the problem without downloading the BLAST databases onto a local server. FastqBLAST was built to address this issue and can quickly elucidate potential causes of low mapping rates associated with sequence quality, reference genome assembly, and contamination from poor library preparation or an organism's microbiome.

objective

The program is intended to be a user-friendly means of running multiple BLAST searches and making sense of the results. FastqBLAST can be run on a personal computer and can handle very large FASTQ files. Additionally, it does not require users to download and constantly update BLAST databases on a local server because NCBI requests are submitted via the internet.



application

A RNA-Seq study of three tissues – breast muscle, abdominal fat, and liver – in broiler chickens was used to demonstrate the usefulness of FastqBLAST. Alignment of reads to the reference genome, *Gallus gallus* 5.0, resulted in low mapping rates and a lower percentage of uniquely mapping reads specific to the breast muscle. The cause of the lower mapping rate was unclear, so a sample of 1000 unmapped reads from each tissue was submitted to the BLAST database using FastqBLAST. The resulting gene report showed that the most highly represented genes among the breast muscle reads were muscle-specific genes such as enolase 3 (beta, muscle) (ENO3) and myosin light chain, phosphorylatable, fast skeletal (MYLPF) whereas the most highly represented genes in the other two tissues were 28S and 18S ribosomal RNA. An attempt to align these muscle genes to the *Gallus gallus* reference resulted in multiple hits, suggesting copy number variation and/or a poor assembly. This theory was further confirmed by re-running the alignment with chimeric alignment detection turned on. It was found that chimeric alignment was highly correlated ($R^2 = 0.5202$) with the percentage of unmappable reads in a sample (see Fig. 1 and 2).

methods

FastqBLAST is a python-based program that invokes NCBI's BLAST and EFetch services over the internet on a random sample of sequences in a FASTQ file and compiles summary reports that characterize the organismal and genomic diversity of that sample. The program selects a random sample of sequences from the FASTQ file, trims low quality ends from each sequence, BLASTs the trimmed sequences, retrieves additional data on the returned gene list with EFetch, and parses the results to produce tabular reports and key figures. Default parameters of FastqBLAST closely mirror those used on the NCBI website and can be adjusted as needed along with the sample size.

results

Three reports are produced by the program. One report contains all of the reads that were run against the BLAST database(s) and includes the gene ID, sequence, sequence length, gene description, alignment stats, organism, and taxonomy. The second report is a summary of the sequences, with information such as the average sequence length after trimming, the proportion of the reads that returned a hit, the average alignment score, the number of uniquely mapped reads, etc. The third report includes information on the most highly represented genes, including how many hits each gene ID received.

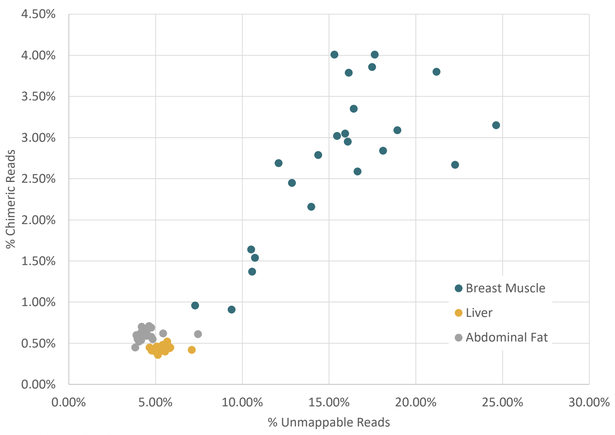


Fig. 1 One potential cause of low mapping rates is the presence of a large number of chimeric reads, which might not be readily apparent during alignment unless the relevant chimera flags are turned on.

FastqBLAST can quickly elucidate potential causes of low mapping rates associated with sequence quality, reference genome assembly, and contamination from poor library preparation or an organism's microbiome.

Tissue	Avg. % Uniquely Mapping Reads	Avg. % Unmappable Reads	Avg. % Chimera Reads
Breast Muscle (n=23)	80.11%	15.39%	2.73%
Abdominal Fat (n=22)	90.78%	4.50%	0.61%
Liver (n=23)	90.98%	5.33%	0.44%

Fig. 2 In FastqBLAST, chimeric reads can show up as a lower percentage of uniquely mapping reads because certain portions of the reference genome are poorly assembled.

The program is user-friendly, can be run on a personal computer, and can handle very large FASTQ files. Additionally, it does not require users to download and constantly update BLAST databases on a local server because NCBI requests are submitted via the internet.

FastqBLAST is available with some demo data at https://github.com/Abasht Laboratory/FastqBLAST

Symptom	Possible Cause
Low mapping rates	Multi-mapping reads associated with specific genes Assembly issues at certain loci
	Low post-trimming sequence lengths
	Low-quality reads
	Large number of reads from other species
	Contamination

interpretation

future work

Keep an eye out for updates to the program that will enhance the summary reports by adding key figures with particular emphasis on taxonomic information and gene IDs. This will allow users to quickly see which large taxonomic groupings are most prevalent in their reads and which genes are most represented in each taxonomic group. The reports will be in a more refined and ready-to-share format instead of their current tabular state.

acknowledgements

The RNA-seq project was jointly funded by Delaware Bioscience Center for Advanced Technology and Heritage Breeders, LLC.