Usage Report for Fiscal Year 2025

Project Title: Machine Learning-Based identification of epigenetic regulation of basal Transposable Elements During Viral Infection.

Name:
○ **Hiroshi Masuya (1), Magdaline Otieno (1)**
Laboratory at RIKEN:
(1) RIKEN BioResource Centre, Integrated Bioresource Information Division

1. Background and purpose of the project, relationship of the project with other projects

Transposable elements (TEs) occupy nearly half of the human genome and are increasingly recognized as key contributors of the innate immunity. During viral-induced stress, specific TEs can form double stranded ribonucleic acid (dsRNA) that activate cytosolic sensors such as RIG-I (retinoic acid-inducible gene I) triggering interferon (IFN) responses. Importantly, TE activation typically precedes the induction of robust interferon stimulated genes (ISGs) and failure to activate TEs is associated with weaker antiviral responses Viruses such as severe acute respiratory syndrome coronavirus (SARS-COV-2) actively supress TE activation, underscoring TE importance in antiviral defence. Although TEs are epigenetically repressed to prevent chronic inflammation and genome instability, low (basal) expression levels persist across cells. These basal TE levels contribute to immune readiness and are equally associated with inter-individual differences in IFN responsiveness. Recent findings show that viral load correlates with basal TE levels and basal IFN signatures. However, whether variation in basal TE levels directly affects inter-individual variation in IFN response remains unclear. TEs are repressed by multiple epigenetic mechanisms, including DNA methylation and H3K9 methylation such as the human silencing hub (HUSH) complex and the SET bifurcated histone lysine methyltransferase 1/Tripartite motif-containing protein 28 (SETDB1/TRIM28). Therefore, variations in epigenetic repression could be a plausible contributor to individual differences in basal TEs and consequently IFN responsiveness. This project will apply machine learning and multi-omics integration to determine whether basal TE expression can predict the strength of antiviral immune response using the IFN score calculation and map predictive TEs to epigenetic pathways. Specifically, we will ask: (1) whether supervised models trained on basal TE expression from healthy samples can accurately predict IFN scores in virus-infected samples; (2) whether predictive TEs are enriched for epigenetic marks and regulators known to silence TEs such as the HUSH complex and SETDB1/TRIM28. (3) in which tissues and cell types these predictive TEs show the strongest coupling with IFN responses; and (4) whether baseline TE levels can predict disease severity across individuals. Publicly available RNA-sequencing datasets from healthy and viral-infected human samples will be used to quantify TE and gene expression and to derive IFN scores from interferon-stimulated genes. Supervised models will be trained on basal TE expression profiles and evaluated for their ability to predict IFN scores in infected samples, with feature-selection approaches highlighting the most informative TEs. Predictive TEs will then be annotated using ATAC-seq and

ChIP-seq datasets to assess chromatin accessibility, H3K9me3 enrichment, and occupancy of HUSH-related factors (TASOR, TASOR2, MPP8, Periphilin), SETDB1, and TRIM28. Taken together, this work will define how basal TE regulation shapes antiviral immunity across tissues and cell types and evaluate whether TE-based signatures can serve as predictors of disease severity.

2. Specific usage status of the system and calculation method

The project is currently processing large RNA sequencing data (bulk and single cell) on Hokusai using SLURM workload manager. All analyses were conducted within Conda managed environments to ensure reproducibility. The analysis involved high-throughput transcriptomics data processing and large-memory genome alignment. The intended output are Binary Alignment/Map files (BAM) and gene count matrices to support downstream TE-based machine learning analysis. The computational approach consisted of three main stages:

- **Acquisition of FASTQ files**

Raw FASTQ files were downloaded using parallel processing and stored on home directory. The job could not be submitted to compute nodes due to no internet errors therefore downloads were done on the login node as shown ***cat download_links.txt | xargs -n1 -P2 wget -c***

- **Quality Control (FastQC)**

Raw FASTQ files were assessed using FastQC tool on the mpc partition. Each job utilizes 8 CPU cores and 32 GB memory with a maximum runtime of 8 hours. Quality reports were generated for all sequencing files, compiled using MultiQC tool and stored in home directory.

- **Read Preprocessing (fastp)**

Adapter trimming and read filtering were performed using fastp for bulk RNA- sequencing data only. Jobs were executed with 8 CPU cores and 8 GB memory per task, with a maximum runtime of 2 hours. Output included trimmed FASTQ files and quality reports.

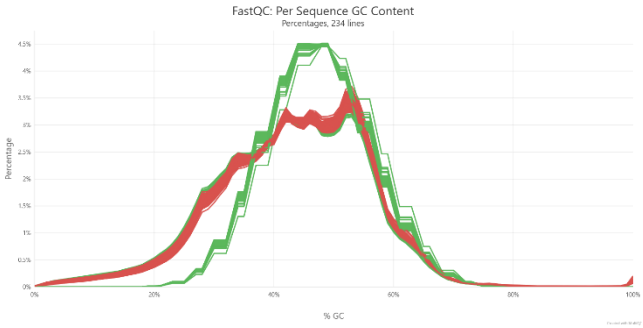
- **scRNA-seq Alignment (STAR / STARsolo)**

Single-cell RNA-seq alignment to the human hg38 reference genome was performed using STAR on the mpc_l partition. A SLURM job array was used to process 29 samples. Each task utilizes 12 CPU cores and 96 GB memory, with a maximum runtime of 72 hours.

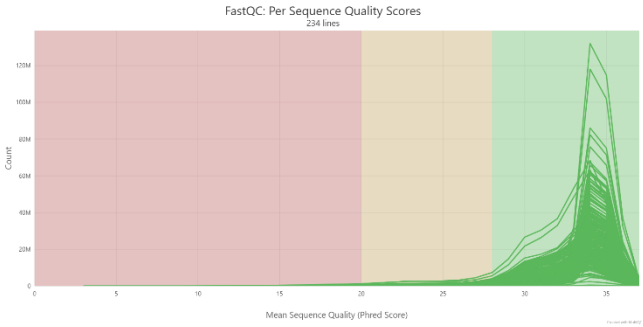
3. Result

Quality control (FASTQC) for single cell reads

The following represents quality reports for single cell-sequencing data as an example. Three plots; the Guanine-Cytosine (GC) content, sequence quality and mean sequence quality are elaborated.

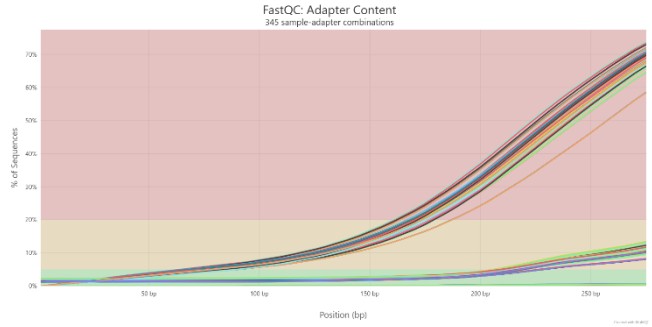


The graph shows the distribution of GC content in the analyzed reads compared to reference genome GC content. The reads originated from paired end sequencing data comprised of Read 1 and Read 2 from samples. The two reads showed distinct GC content distributions due to differences in their composition. Read 1 primarily contains barcode and technical sequences while Read 2 represents biological cDNA fragments, representing transcriptional GC composition.



The quality score graph indicated that most reads

had high quality (phred score>20). Therefore, no reads were discarded from the analysis.



Adapters were observed in the reads consisting mostly of polyA and polyG tails. The alignment tool (STAR) soft clips the reads therefore read trimming was not done on the reads.

Mapping results

The project has been completed 18 of 29 mapping jobs using 12 CPUs and 96 G RAM as indicated.

JOBID	Partition	Nodes	Node list
6934635_ [20-28%1]	mpc_l	1	1 (JobArrayTaskLimit)
6934635_19	mpc_l	1	bwmpc150

Upon analysis of individual task, it was noted that each mapping used at least 40G RAM as shown below in green, therefore, to process the remaining FASTQs, the work will apply parallelization of two FASTQs to cut down on runtime.

JobID	MaxRSS	Elapsed	State
6934635_9		04:46:35	Completed
6934635_9. b+	37847524K	04:46:35	Completed
6934635_9. e+	20K	04:46:35	Completed

4. Conclusion

- The mapping for the reads was successfully done on Hokusai; subsequent analysis will include the quantification of TEs from the BAM files generated.
- Parallelization should be implemented via CPU multi-threading and array-based sample distribution to speed up alignment.

5. Schedule and prospect for the future

Research Project Timeline (Year 1)

Stage	Activity	Estimated duration	Start date	End date	Deliverables	Comments ✓ (completed) - (Incomplete)
Research design	Research problem/questions	3 months	Oct 2025	Dec 2025	Confirmed RQs	✓
	Develop research design				Research design draft	✓
	Prepare research proposal				Proposal draft	✓
Literature review + Data collection	Systematic literature review	4-6 months	Jan 2026	Jun 2026	Draft literature review sections	-
	Begin writing literature review chapter					
	Search and collect data				Raw datasets downloaded	✓
	Organize data on Hokusai					✓
Data preparation & Pilot analysis	Perform QC, trimming, alignment, TE quantification	1 month	Feb 2026	Mar 2026	Quality control reports	✓
					Processed FASTQs/	✓
	Run pilot TE-IFN analysis on transcriptomics data				First TE-IFN results	-
	Generate first TE-IFN correlation plots					-
	TE- IFN coupling in target tissue					-
Internship	Epigenetic data analysis (ChIP-seq, ATAC-seq, methylation)	1 month	April 2026	April 2026	Insights for RQ2	-
Full data analysis	TE-IFN coupling in other virus datasets (Influenza, Rhinovirus, WNV, ZIKV)	6 months	May 2026	Oct 2026	Viral TE-IFN results (RQ1, RQ3)	-
	Epigenetic regulation of predictive TEs				Epigenetic regulation (RQ2)	-
	Build TE-based IFN prediction model				Donor-level prediction model (RQ4)	-
First thesis report	Write preliminary draft	1 month	Sep 2026	Oct 2026	First draft manuscript	-
YEAR 2						
Stage	**Activity**	**Estimated duration**	**Start date**	**End date**	**Deliverables**	**Comments ✓ (completed) - (Incomplete)**
Method validation	Re-analysis/ reproducibility checking	6 months	Oct 2026	Mar 2027	Validated ML model Identified predictive TEs/Identified epigenetic factors	-
Experimental validation	Validation of identified TEs in lab	6 months	Oct 2026	Mar 2027	-	-
Journal submission	Submit paper to journal	-	Mar 2027	Oct 2027	-	-
YEAR 3						
Stage	**Activity**	**Estimated duration**	**Start date**	**End date**	**Deliverables**	**Comments ✓ (completed) - (Incomplete)**
Review manuscript based on journal comments	Rerun additional analysis	8 months	Oct 2027	June 2028	Completed thesis	-
Thesis final draft	Compile all sections of PhD thesis	8 months	Oct 2027	June 2028	Completed thesis	-

Continued access to Hokusai is required throughout the PhD period to ensure reproducibility, scalability of analyses, and potential re-analysis during peer review. Regions indicated by orange represent task which will utilize Hokusai system for analysis.