Data Article

Data supporting a saturation mutagenesis assay for Tat-driven transcription with the GigaAssay

Ronald Benjamin\textsuperscript{a}, Christopher J. Giacoletto\textsuperscript{a,b,c}, Zachary T. FitzHugh\textsuperscript{a}, Danielle Eames\textsuperscript{a}, Lindsay Buczek\textsuperscript{a}, Xiaogang Wu\textsuperscript{a}, Jacklyn Newsome\textsuperscript{a}, Mira V. Han\textsuperscript{a,b}, Tony Pearson\textsuperscript{b,c}, Zhi Wei\textsuperscript{d}, Atoshi Banerjee\textsuperscript{a}, Lancer Brown\textsuperscript{c}, Liz J. Valente\textsuperscript{c}, Shirley Shen\textsuperscript{a}, Hong-Wen Deng\textsuperscript{e}, Martin R. Schiller\textsuperscript{a,b,c,*}

\textsuperscript{a} Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, 4505 S. Maryland Parkway, Las Vegas, NV 89154-4004 USA
\textsuperscript{b} School of Life Sciences, University of Nevada Las Vegas, 4505 S. Maryland Parkway, Las Vegas, NV 89154-4004 USA
\textsuperscript{c} Heligenics Inc.\textsuperscript{1}, 833 Las Vegas Blvd. North, Suite B, Las Vegas, NV 89101, USA
\textsuperscript{d} Department of Computer Science, New Jersey Institute of Technology, GITC 4214C, University Heights, Newark, NJ 07102 USA
\textsuperscript{e} Center for Biomedical Informatics & Genomics Tulane University\textsuperscript{2}, 1440 Canal Street, Suite 1621, New Orleans, LA 70112 USA

\begin{tabular}{l}
\hline
\textbf{A R T I C L E I N F O} & \textbf{A B S T R A C T} \\
\hline
\end{tabular}

Article history:
Received 1 August 2022
Revised 21 September 2022
Accepted 22 September 2022
Available online 28 September 2022

Dataset link: GigaAssay, an adaptable high throughput saturation mutagenesis assay for Tat driven transcription (Original data)

The data in this article are associated with the research paper “GigaAssay – an adaptable high-throughput saturation mutagenesis assay” [1]. The raw data are sequence reads of HIV-1 Tat cDNA amplified from cellular genomic DNA in a new single-pot saturation mutagenesis assay designated the “GigaAssay”. A bioinformatic pipeline and parameters used to analyze the data. Raw, processed, analyzed, and filtered data are reported. The data is processed to calculate the Tat-driven transcription activity for cells with each possible

* Corresponding author at: Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, 4505 S. Maryland Parkway, Las Vegas, NV 89154-4004 USA.

\textsuperscript{1} E-mail address: Martin.schiller@unlv.edu (M.R. Schiller).

https://doi.org/10.1016/j.dib.2022.108641

2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Specifications Table

| Subject                           | Biotechnology |
|-----------------------------------|---------------|
| Specific subject area             | Biotechnological research that discovers the sequence-function landscape of Tat. |
| Type of data                      | Tables, images, graphs, figures, files, next generation sequencing (NGS) data (raw and processed). |
| How the data were acquired        | Targeted paired-end sequencing of genomic DNA extracted from flow sorted cell bins. |
| Data format                       | Raw, Analyzed, Filtered, NGS, Fastq, Spreadsheet, Images |
| Description of data collection    | Data were produced from a new type of saturation mutagenesis assay called the GigaAssay. First, a reporter assay system was created to measure Tat-driven transcriptional activity by binding to the long terminal repeat (LTR) promoter driving Green Fluorescent Protein (GFP) expression. Cells with active Tat have high fluorescence, whereas cells lacking functional Tat have basal fluorescence. The assay system was used to measure activities of a saturating single substitution mutagenesis set of Tat mutants. Oligonucleotides were synthesized that carried the Tat sequence with one codon substitution that results in an AA change and a unique molecular identifier (UMI) barcode was ligated into the 3’ untranslated region. There was an average of approximately 100 UMIs for each mutant, which results in 100s of independent measurements of Tat mutant function. Lentiviruses encoding Tat cDNAs were transduced into LentiX293T/LTR-GFP or Jurkat/LTR-GFP cells, and the cells with active Tat mutants became fluorescent. The cells were sorted into three bins based on fluorescence intensity (low, mid, high) with a flow cytometer and gating was set by a loss-of-function (LOF) Tat mutant and wild type (WT) Tat controls. Genomic DNA was extracted from cells in each bin and Tat cDNAs were sequenced by targeted NGS producing paired-ended reads. The resulting dataset has raw sequence reads that were processed into transcription activities for each Tat mutant with a confidence statistic. |
| Data source location              | Nevada Institute of Personalized Medicine (NIPM), University of Nevada, Las Vegas (UNLV) USA; Heligenics Inc., Las Vegas, Nevada USA |
| Data accessibility                | Repository name: Sequence Read Archive (SRA) Data identification number: PRJNA857699 Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA857699/ Instructions for accessing these data: Use the provided link to access the project page. Each sequencing run can be retrieved by following the link at the bottom of the page under “Project Data”. Click each number to show all runs, then select the run to download. Click the run ID on the bottom left of the table. Finally, select the “FASTA/FASTQ Download” header, and select the preferred type to download (fasta or fastq) and download will begin. |

(continued on next page)
Value of the Data

The data can be reused to further additional scientific investigation. Some potential uses are:

- Prediction of mutant activity for multiple mutants and other genes.
- Determining sources of errors in NGS data with 100s of measurements for each variant and each DNA molecule coded with a random UMI barcode.
- Advancing structure/function analyses by including the spatial and physiochemical tolerances for substitutions at each position.
- Comparing functional results between technical and biological replicates for cell lines.
- Interpreting mutagenesis experiments that examine Tat post-translational modifications (PTM)s and protein-protein interactions (PPI)s.

1. Experimental Design

This demonstration of the GigaAssay, measures Tat-driven transcriptional activity with an LTR-GFP reporter. We adapted a previous low-throughput assay where exogenous Tat expression binds to the LTR element and drives transcription of an integrated LTR-GFP reporter [2]. A cDNA library for saturating mutagenesis of Tat was constructed. Oligonucleotides were synthesized that encode the complete Tat cDNA sequence. Each molecule in the library has one codon substitution that results in an amino acid substitution. The cDNAs are subcloned into a plasmid library and a UMI barcode is ligated into the 3’ untranslated region. Lentiviruses encoding these Tat cDNAs were transduced into LentiX293T/LTR-GFP or Jurkat/LTR-GFP cells. Cells with active Tat mutants activate the reporter and fluoresce. The cells were sorted into three bins of graded by fluorescence intensity (low-GFP, mid-GFP, high-GFP) with a flow cytometer. Bins were gated based upon measurements of a known LOF mutant and WT Tat. Genomic DNA was extracted from each bin and Tat cDNAs were sequenced by targeted NGS of Tat producing paired-ended reads. The calculated GFP-high / GFP-low sequence read ratios measure the relative transcriptional activities of each mutant. Mutants that had low activity (<50%) were classified as LOF, and mutants with >50% activity were classified as WT. The assay was repeated for both cell lines. Since each mutant in each run of the assay has, on average, 100 independent measures of its activity due to the incorporation of the UMI, technical replicates are not needed, but we include one for rigor.

2. Data Description

The transcriptional activity associated with every possible single amino acid substitution of HIV-1 Tat drove GFP fluorescence from an LTR-GFP reporter. Tat amplicons were sequenced with
an Illumina MySeq producing paired-ended reads. Contigs of 465 bp were built from forward and reverse read (250 base pairs (bp)) fusing the overlapping reverse complement regions of the forward and reverse reads. Each clone in the library has the full Tat coding region with one codon substitution, as well as a random 32 bp UMI barcode.

Transcriptional activity levels for each mutant were calculated by computing the ratio of normalized reads in the high-GFP and low-GFP flow sorted bins. Mutants where this ratio was \( \geq 50\% \) were considered active for experimental validation. Barcodes passed filtration if there was at least 2.5 reads per million (RPM) for a given barcode. Each mutant’s activity classification and the distribution for assays in Jurkat/LTR-GFP cells are shown in Fig. 1.

Activities for Jurkat/LTR-GFP cells were plotted in heat maps for each particular amino acid substitution at each position organized by physicochemical properties (Fig. 2) or by amino acid side chain volume (Fig. 3). Similar figures are shown for LentiX293T-GFP cells (Figs. 4 and 5). The activities are compared to sites for regions, secondary structures (SS), solvent accessible surface area (SASA), posttranslational modification (PTM) sites, and protein-protein interaction (PPI) sites of Tat. The substitution tolerance for each amino acid’s physicochemical properties was calculated with a Matthew’s correlation coefficient (MCC) score where positive numbers indicate a required chemistry, and negative numbers indicate that the chemistry is not tolerated.

To evaluate the coverage depth of the sequenced library, the number of barcodes, and the reads associated with each mutant are shown in Figs. 6 and 7, respectively.

GigaAssay results were tested for accuracy with three independent approaches. Tat mutant activities in the GigaAssay (see data file 1) were compared to 164 known mutants reported in the literature (see data file 2). We also compared the activities of GigaAssay results to the activity of Tat nonsense mutants measured in the GigaAssay and the known activities of these truncation (see data file 3) [1]. The GigaAssay was further validated by comparing the measured activities against mutants that were independently measured and blinded prior to the GigaAssay experiment. Eighteen stable LentiX293T/LTR-GFP cell lines, each expressing the Tat cDNA with a random Tat point mutation were created and analyzed. The LTR-GFP reporter activities for these mutants were measured by flow cytometry and compared to WT Tat (Fig. 8). These positive controls were then compared to GigaAssay results.

All mutant activities were compared between LentiX293T/LTR-GFP and Jurkat/LTR-GFP to test if cell lines effect the activities of Tat mutants (Fig. 9). Fitting to a line indicates a correlation \( (R^2 = 0.93) \).

A statistical model to classify Tat mutant activities was designed to capture mutants whose activity was like WT or LOF mutants. For each UMI barcode in a sample, the percentage of reads in the high-GFP bin was calculated, and is denoted as the \( h \) ratio \( (h \epsilon [0,1]) \). A high \( h \) percent-
Fig. 2. Heatmap of Tat mutant transcriptional activities in Jurkat/LTR-GFP cells. Amino acids are organized by physiochemical properties. Black squares = reference; Gray square = null. The activity minimum, average, and maximum are shown in yellow – red heatmap. The MCC scores range from -1 (blue) to 0 (white) to 1 (magenta). White indicates no specificity, magenta indicates high specificity for the physiochemical group, and blue indicates high specificity for negative preference against the physiochemical group.

age resembles WT, while a low \( h \) percentage suggests a mutant. For each mutant, we calculate the averaged \( h \) ratio for all the barcodes assigned to the same mutant, denoted as a mutant level summary score. A one sample t-test was used to evaluate (1) whether the mutant has a significantly different number of reads in the high-GFP bin compared to the low-GFP bin for technical replicates, and (2) whether the mutant has a significantly different number of reads in the high-GFP bin compared with the low-GFP bin among biological replicates for different cell
Fig. 3. Heatmap of Tat mutant transcriptional activities in Jurkat/LTR-GFP cells. Amino acids are organized by side chain volumes. Black squares = reference; Gray square = null. The key is as is in Fig. 2. The activity minimum, average, and maximum are shown in a yellow – red heatmap. The MCC scores range from -1 (blue) to 0 (white) to 1 (magenta). White indicates no specificity, magenta indicates high specificity for the physiochemical group, and blue indicates high specificity for negative preference against the physiochemical group.

lines (null hypothesis $H_0: \ h =0.5$). Results of this test in LentiX293T/LTR-GFP cells (Fig. 10) and Jurkat/LTR-GFP cells (Fig. 11) are shown.

In addition, another statistical test was designed to assess the association between the mutant genotype classification (LOF/WT) and Tat-driven transcriptional activity levels (binary variable; high-GFP bin or low-GFP bin). A mixed effect logistic regression was used, with random intercepts for barcodes and replicates to model the nested structure in our experimental design. For the WT control populations, we used Tat cDNAs from cells with no mutant calls (WT sequences identical to the reference). Each mutant was compared against this common WT control population. The model M1 with the genotype included as fixed effects was compared to a null model M0 without genotype in a likelihood ratio test (LRT). Similar to a Genome-Wide
Fig. 4. Heatmap of Tat mutant transcriptional activities in LentiX293T/LTR-GFP cells. Amino acids are organized by physiochemical properties. Black squares = reference; Gray square = null. The key is as is in Fig. 2. The activity minimum, average, and maximum are shown in a yellow – red heatmap. The MCC scores range from -1 (blue) to 0 (white) to 1 (magenta). White indicates no specificity, magenta indicates high specificity for the physiochemical group, and blue indicates high specificity for negative preference against the physiochemical group.

Association Studies (GWAS), a significant result indicates that the variant/WT is associated with the percentage of cells in the high-GFP bins. For mutants where the model fit was singular, we simplified the model by dropping the random effects. p-values were adjusted for false discovery rate (FDR) using Storey’s q-values.

Tests were done at the replicate level with models:

M1 : GFP ~ genotype + (1|barcode)M0 : GFP ~ (1|barcode)

Tests were done at the cell type level with models:

M1 : GFP ~ genotype + (1|barcode/replicate)M0 : GFP ~ (1|barcode/replicate)
Fig. 5. Heatmap of Tat mutant transcriptional activities in LentiX293T/LTR-GFP cells. Amino acids are organized by side chain volumes. Black squares = reference; Gray square = null. The key is as is in Fig. 2. The activity minimum, average, and maximum are shown in a yellow – red heatmap. The MCC scores range from -1 (blue) to 0 (white) to 1 (magenta). White indicates no specificity, magenta indicates high specificity for the physiochemical group, and blue indicates high specificity for negative preference against the physiochemical group.

Results of this test for each mutant in LentiX293T/LTR-GFP cells and Jurkat/LTR-GFP cells are shown in Figs. 12 and 13, respectively. WT and LOF association results from LentiX293T/LTR-GFP cells and Jurkat/LTR-GFP cells are shown in Figs. 14 and 15, respectively. P values for both cell lines testing for LOF and WT activity levels were quantified (Fig. 16). Results of all statistical analyses are reported in data file 4.
Fig. 6. Quantitation of barcodes for each Tat mutant. A, B. Heatmap of barcodes for Tat mutants in LentiX293T/LTR-GFP and Jurkat/LTR-GFP cells, respectively. A key for the heatmap colors is shown. Black cells indicate the reference sequence. C, D. Barcode correlation for technical replicate samples for the mutants in A and B. Each point on the scatter plot is for matched for the same mutant. Barcodes for replicates samples in LentiX293T/LTR-GFP (C, $R^2 = 0.95$) and Jurkat/LTR-GFP (D, $R^2 = 0.96$) cells are fit to a line (red).
Fig. 7. The number of sequence reads for each Tat mutant. Sequence reads for Tat mutants in LentiX293T/LTR-GFP (A) and Jurkat/LTR-GFP (B) cells are shown. A color key is included. Black squares indicate the reference sequence amino acid at each position.
Fig. 8. The flow cytometry profiles of stable LentiX293T/LTR-GFP cell lines engineered to express different Tat mutants. Mutation and the percentage of GFP+ cells are indicated. Tat mutant cells are colored red and WT Tat is colored cyan.
Fig. 9. Tat-driven transcription activities of Tat mutants comparing LentiX293T/LTR-GFP cells to Jurkat/LTR-GFP cells. Matched mutants in LentiX293T/LTR-GFP (open circles) and Jurkat/LTR-GFP cells (blue filled circles) are shown in a scatter plot; Data are fit to a line (red; $R^2 = 0.93$).
Fig. 10. Statistical significance of activities of Tat mutants in LentiX293T/LTR-GFP cells. The hypothesis tested is whether the GFP$^+$ sequence read ratio observed for that mutant is equal to 0.5. Black squares indicate the reference sequence amino acid at each position. A. A heatmap of $-\log(p$ values) for Tat mutant transcriptional activities in LentiX293T/LTR-GFP cells is shown. B. A bin plot showing the distribution of $-\log(p$ values); ($n = 1,615$).
Fig. 11. Statistical significance of activities of Tat mutants in Jurkat/LTR-GFP cells. The hypothesis tested is whether the GFP+ sequence read ratio observed for that mutant is equal to 0.5. Black squares indicate the reference sequence amino acid at each position. A. A heatmap of $-\log(p$ values) for Tat mutant transcriptional activities in Jurkat/LTR-GFP cells is shown. B. A bin plot showing the distribution of $-\log(p$ values); $(n = 1,615)$. 
Fig. 12. Statistical significance of the effect of Tat mutants on Tat activity in LentiX293T/LTR-GFP cells. The hypothesis tested is whether the genotype (Variant/WT) has an effect on the percentage of GFP+ cells. Black squares indicate the reference sequence amino acid at each position. Heatmaps of A. –log(p values) and B. –log(q values) for the LRT test on the significance of the genotype variable in LentiX293T/LTR-GFP cells.
Fig. 13. Statistical significance of the effect of Tat mutants on Tat activity in Jurkat/LTR-GFP cells. The hypothesis tested is whether the genotype (Variant/WT) has an effect on the percentage of GFP+ cells. Black squares indicate the reference sequence amino acid at each position. Heatmaps of A. \(-\log(p \text{ values})\) and B. \(-\log(q \text{ values})\) for the LRT test on the significance of the genotype variable in Jurkat/LTR-GFP cells.
Fig. 14. Heatmaps of q values Tat mutant transcriptional activities in LentiX293T/LTR-GFP cells. Black squares indicate the reference sequence amino acid at each position. q values for comparison of Tat mutant activity to sets of mutants with WT (A) and LOF activity (B). Keys for q value colors are shown.
Fig. 15. Heatmaps of p values Tat mutant transcriptional activities in Jurkat/LTR-GFP cells. Black squares indicate the reference sequence amino acid at each position. q values for comparison of Tat mutant activity to sets mutants with WT (A) and LOF activity (B). Keys for q value colors are shown.
Fig. 16. Bar charts of p values for Tat mutant transcriptional activities compared to wild type Tat and LOF Tat mutants. p values for comparison of Tat mutant activity to sets of mutants with WT activity (A, B) and LOF activity (C, D) for LentiX293T/LTR-GFP (A, C) and Jurkat/LTR-GFP (B, D) cells. WT and LOF percentages are statistically significant (p < 0.05).

Data Files (Files S1–S4):

- SupplementaryFileS1_TatMutantActivity.xlsx:
  - This file shows the transcriptional activity measured for each sample replicate in each cell line for each mutant of Tat. Tabs 1 and 2 contains the results for LentiX293T/LTR-GFP and Jurkat/LTR-GFP, respectively. Column labels are defined in tab 3.

- SupplementaryFileS2_BenchmarkData.xlsx:
  - This file contains published Tat transcriptional activities for 164 previously published mutant activities used for verification of the GigaAssay results. Column labels are defined in tab 2.

- SupplementaryFileS3_TatNonsenseMutants.xlsx:
  - This file contains the transcriptional activity measured for each Tat nonsense mutation. Tabs 1 and 2 contains the results for LentiX293T/LTR-GFP and Jurkat/LTR-GFP cells, respectively. Column labels are defined in tab 3.

- SupplementaryFileS4_MutantStatistics.xlsx:
This file contains the results of a statistical tests each mutant with a mixed effect logistic regression model to test if read distributions amongst the sorted bins are associated with WT or mutant activity, respectively. Tab 1 contains the results of the missense mutants and tab 2 shows nonsense mutants. Tabs 3 and 4 contain the results of a t-test comparing our mutant’s read ratios against 50% activity. Mutants closer to 0% activity are classified as LOF and mutants closer to 100% activity are classified as WT. Tab 3 contains the results of the missense mutants and tab 4 shows nonsense mutants. Column labels are defined in tab 5.

Raw NGS Data (Files 5-38):

- **1_Tatlib_293T_GFP_high_ACTTCTTC_L001_R1_001.fastq.gz**
  o Forward sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_293T_GFP_high_ACTTCTTC_L001_R2_001.fastq.gz**
  o Reverse sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_293T_GFP_mid_GACGCTAT_L001_R1_001.fastq.gz**
  o Forward sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_293T_GFP_mid_GACGCTAT_L001_R2_001.fastq.gz**
  o Reverse sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_293T_GFP_low_CAAGGCGA_L001_R1_001.fastq.gz**
  o Forward sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-low bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_293T_GFP_low_CAAGGCGA_L001_R2_001.fastq.gz**
  o Reverse sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-low bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_293T_rep_GACCGAGA_L001_R1_001.fastq.gz**
  o Forward sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The pre-sort bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_293T_rep_GACCGAGA_L001_R2_001.fastq.gz**
  o Reverse sequence reads from LentiX293T/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The pre-sort bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
- **1_Tatlib_JLTRG_GFP_high_AATGCGTT_L001_R1_001.fastq.gz**
  o Forward sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by
targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **1_Tatlib_JLTRG_GFP_low_AATGGTT_L001_R2_001.fastq.gz:**
  - Reverse sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **1_Tatlib_JLTRG_GFP_low_GTTGGTA_T001_R1_001.fastq.gz:**
  - Forward sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-low bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **1_Tatlib_JLTRG_GFP_high_ATCCGACA_L001_R001.fastq.gz:**
  - Reverse sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **1_Tatlib_JLTRG_GFP_mid_CTATTCAA_L001_R1_001.fastq.gz:**
  - Forward sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **1_Tatlib_JLTRG_GFP_high_CTATTCAA_L001_R2_001.fastq.gz:**
  - Reverse sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **1_Tatlib_JLTRG_rep_ATCCGACA_L001_R1_001.fastq.gz:**
  - Forward sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The pre-sort bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **1_Tatlib_JLTRG_rep_ATCCGACA_L001_R2_001.fastq.gz:**
  - Reverse sequence reads from Jurkat/LTR-GFP cells (replicate 1) were sorted by flow cytometry into bins of graded fluorescence. The pre-sort bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **2_Tatlib_293T_GFP_high_CATCAGAC_L001_R1_001.fastq.gz:**
  - Forward sequence reads from LentiX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **2_Tatlib_293T_GFP_high_CATCAGAC_L001_R2_001.fastq.gz:**
  - Reverse sequence reads from LentiX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-high bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- **2_Tatlib_293T_GFP_low_CATCAGAT_L001_R1_001.fastq.gz:**
  - Forward sequence reads from LentiX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-low bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
• 2_Tatlib_293T_GFP_mid_GATTACAG_L001_R1_001.fastq.gz:
  o Reverse sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-low bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_JLTRG_GFP_high_TGTGCTTA_L001_R2_001.fastq.gz:
  o Reverse sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_293T_GFP_mid_TGTTGTTA_L001_R1_001.fastq.gz:
  o Forward sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The pre-sort bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_JLTRG_GFP_mid_TGTTGTTA_L001_R1_001.fastq.gz:
  o Forward sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_293T_GFP_mid_TGTTGTTA_L001_R2_001.fastq.gz:
  o Reverse sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_JLTRG_GFP_mid_TGTTGTTA_L001_R2_001.fastq.gz:
  o Reverse sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_JLTRG_GFP_mid_TGTTGTTA_L001_R2_001.fastq.gz:
  o Reverse sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_JLTRG_GFP_mid_TGTTGTTA_L001_R2_001.fastq.gz:
  o Reverse sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

• 2_Tatlib_JLTRG_GFP_mid_TGTTGTTA_L001_R1_001.fastq.gz:
  o Reverse sequence reads from LentivX293T/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The GFP-mid bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.
targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- 2_Tatlib_JLTRG_rep_TGTTTAC_L001_R1_001.fastq.gz:
  - Forward sequence reads from Jurkat/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The pre-sort bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- 2_Tatlib_JLTRG_rep_TGTTTAC_L001_R2_001.fastq.gz:
  - Reverse sequence reads from Jurkat/LTR-GFP cells (replicate 2) were sorted by flow cytometry into bins of graded fluorescence. The pre-sort bin in this file was analyzed by targeted NGS, producing Tat cDNA sequence reads. The data in this file is in fastq format and is compressed with gzip.

- Tatlib_plasmid_rep_GGCATAGG_L001_R1_001.fastq.gz:
  - The plasmid library was analyzed by targeted NGS for Tat cDNA producing the forward reads enclosed in this file. In gzipped, fastq format.

- Tatlib_plasmid_rep_GGCATAGG_L001_R2_001.fastq.gz:
  - The plasmid library was analyzed by targeted NGS for Tat cDNA producing the reverse reads enclosed in this file. In gzipped, fastq format.

3. Experimental Design, Materials and Methods

All methods to produce the data are published in Benjamin et al. [1]. Additional details for the bioinformatics methods are provided.

3.1. Bioinformatic Methods for Processing Data

Raw sequence read data (see data files 5–38) were analyzed with a custom bioinformatic pipeline to extract the Tat encoding region and barcode. The read quality for each file was evaluated with Fastqc [3] using the program’s default parameters. Forward and reverse reads were fused in the overlapping reverse complement region to create a complete coding region contig with FLASH [4]. Parameters that were input into FLASH for read fusion were a minimum and maximum reverse complement overlap length of 20 bp and 45 bp, respectively, and a maximum mismatch ratio of 0.15. After fusion, reads were filtered for quality and required to have a minimum size of 365 bp with Trimmomatic [5]. Reads were removed if the average PHRED score dropped below 16 for any 4–bp window. Reads were trimmed to just the Tat encoding region, a small 3’ extension, and the 32 bp barcode by removing adapter sequences with Cutadapt [6]. The parameter “--a GAAATTC...GCGATCGC” was input into Cutadapt’s linked adapter trim function.

After adapter trimming, the 32 bp UMI barcode was extracted with Cutadapt by specifying a 5’ adapter. Extracted barcode UMIs, some contain sequencing errors were then grouped. All UMIs from the 17 sequencing samples were combined and grouped using Starcode, set with a Levenshtein distance of 2 [7]. Reads were demultiplexed with Starcode, first by barcode group and then by sample. A barcode group consensus sequence was created from each barcode group. Each barcode group may have reads in any of the sample files. For example, barcode group AGACGTACCAACAAAGACAATGACAAAAAG was associated with 1,447 reads across the flow sorted files: 34 and 25 reads corresponded to the replicate 1, GFP-high and GFP-low samples, respectively for LentiX293T cells.

Variants were called for each barcode group in each sample with a variant calling pipeline. Fused reads for each barcode group’s samples were aligned to a Tat reference indexed with Burrow’s Wheeler Aligner’s (BWA) BWA mem function [8]. The resulting SAM files were digitized into BAM files, sorted, indexed, and analyzed with Samtools using the view, sort, index, and pileup functions, respectively [9]. VCF files were generated with BCFtools’ call function [10]. To verify
variant calls for a specific codon, we compared each barcode group among all sample VCFs. The minor fraction of variant calls in a particular barcode group that did not agree with a designed codon substitution was filtered and discarded.

Frequencies of reads were calculated for each barcode group in each sample. Read counts supporting the codon substitution were tallied from the flow sorted sequencing files. Read counts were normalized to RPM. Transcriptional activities were calculated by comparing the distributions of read counts across flow sorted samples across all barcodes for a particular variant (see data file 1). Activities were calculated for each mutant using formula 1 below.

\[
A = \left( \frac{\sum_{i=1}^{b} {\text{GFP}^+}}{\left(\sum_{i=1}^{b} {\text{GFP}^+}\right) + \left(\sum_{i=1}^{b} {\text{GFP}^-}\right)} \right)
\]

For each Tat mutant, an activity level \(A\) was calculated from sequence reads. For all barcodes \(b\) associated with the mutant, normalized high-GFP (GFP\(^+\)) bin reads were summed, and normalized low-GFP (GFP\(^-\)) bin reads were summed. The percentage of normalized reads corresponding to the GFP\(^+\) divided by the total reads in both bins equals the activity measurement, \(A\).

**Ethics Statements**

All authors adhered to the Data in Brief ethics guidelines.

**Declaration ofCompeting Interest**

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

The technology is owned by the University of Nevada Las Vegas and is part of a pending patent application with the United States Patent and Trademark Office [Patent No: PCT/US2017/042179 Canadian PCT-CA (0445-02)]. MRS, LV, LB, and CJG are employees of Heligenics, which has licensed the technology from UNLV and is pursuing commercial interest. UNLV manages a conflict-of-interest management plan for Principal Investigator, MRS. ZW is contracted by Heligenics to build and implement a part of a statistical model for the GigaAssay.

**Data Availability**

GigaAssay, an adaptable high throughput saturation mutagenesis assay for Tat driven transcription (Original data) (Mendeley Data).

GigaAssay, an adaptable high throughput saturation mutagenesis assay for Tat driven transcription (Original data) (Sequence Read Archive).

**CRediT Author Statement**

Ronald Benjamin: Conceptualization, Formal analysis, Methodology, Supervision, Validation, Visualization; Christopher J. Giacoletto: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; Zachary T. FitzHugh: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Software, Visualization; Danielle Eames: Data curation; Lindsay Buczek: Data curation; Xiaogang Wu: Conceptualization, Formal analysis, Methodology, Software; Jacklyn
Newsome: Formal analysis, Methodology, Software, Visualization; Mira V. Han: Conceptualization, Formal analysis, Supervision; Tony Pearson: Formal analysis, Methodology; Zhi Wei: Formal analysis; Atoshi Banerjee: Methodology, Visualization; Lancer Brown: Methodology, Validation, Visualization, Writing – original draft; Liz J. Valente: Methodology, Validation, Visualization, Writing – original draft; Shirley Shen: Methodology; Hong-Wen Deng: Writing – original draft; Martin R. Schiller: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgments

Funding: This work was supported by the National Institutes of Health (grant numbers R21AI116411, R15GM107983, R21AI078708, R56AI109156, P20GM121325); the Governor’s Office of Economic Development (Grant Number: 1547526); and the Prabhu endowed professorship. We also acknowledge the UNLV College of Science for a grant to develop the GigaAssay.

We thank Drs. Edwin Oh, and Richard Tillet from the UNLV Nevada Institute of Personalized Medicine Genome Acquisition and Analysis Core for access to a flow cytometer sorter and help with some NGS sequencing and interpretation for GigaAssay development. We Thank Drs. Jefferson Kinney (University of Nevada, Las Vegas) and Tom Metzger (Roseman University) for use of their flow cytometer. We wish to acknowledge the help of Dr. Nora Caberoy with electroporation experiments. We appreciate the discussions we had with Drs. Qing Wu and Michael F. Lin about statistical assessment of the GigaAssay results.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2022.108641.

References

[1] R. Benjamin, C.J. Giaconeotto, Z.T. FitzHugh, D. Eames, L. Buczek, X. Wu, J. Newsome, M.V. Han, T. Pearson, Z. Wei, A. Banerjee, L. Brown, L.J. Valente, S. Shen, H.W. Deng, M.R. Schiller, GigaAssay – an adaptable high-throughput saturation mutagenesis assay platform, Genomics (2022) 110439, doi:10.1016/j.ygeno.2022.110439.
[2] D.I. Dorsky, M. Wells, R.D. Harrington, Detection of HIV-1 infection with a green fluorescent protein reporter system, J. Acquir. Immune Defic. Syndr. Hum. Retrovirol. Off. Publ. Int. Retrovirol. Assoc. 13 (1996) 308–313, doi:10.1097/00042560-199612010-00002.
[3] Fastqc, (n.d.). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
[4] T. Magoc, S.L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies, Bioinformatics 27 (2011) 2957–2963, doi:10.1093/bioinformatics/btr507.
[5] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.
[6] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (2011) 10, doi:10.14806/ej.17.1.200.
[7] E. Zorita, P. Cusco, G.J. Filion, Starcode: sequence clustering based on all-pairs search, Bioinform. Oxf. Engl. 31 (2015) 1913–1919, doi:10.1093/bioinformatics/btv53.
[8] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–Wheeler transform, Bioinform. Oxf. Engl. 26 (2010) 589–595, doi:10.1093/bioinformatics/btp698.
[9] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 genome project data processing subgroup, the sequence alignment/map format and SAMtools, Bioinform. Oxf. Engl. 25 (2009) 2078–2079, doi:10.1093/bioinformatics/btp352.
[10] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, R. Durbin, BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data, Bioinformatics 32 (2016) 1749–1751, doi:10.1093/bioinformatics/btw044.