GigaAssay – a high-throughput assay system for molecular functions and cell processes

Martin Schiller (martin.schiller@unlv.edu)
University of Nevada Las Vegas https://orcid.org/0000-0003-1671-6823
Ronald Benjamin
University of Nevada Las Vegas
Christopher Giacoletto
Heligenics
Zachary FitzHugh
University of Nevada Las Vegas
Daniel Eames
University of Nevada Las Vegas
Lindsay Buczek
University of Nevada Las Vegas
Xiaogang Wu
University of Nevada Las Vegas
Jacklyn Newsome
University of Nevada Las Vegas
Mira Han
University of Nevada Las Vegas
Tony Pearson
University of Nevada Las Vegas
Zhi Wei
New Jersey Institute of Technology https://orcid.org/0000-0001-6059-4267
Atoshi Banjeree
University of Nevada Las Vegas
Shirley Shen
University of Nevada Las Vegas
Hong-Wen Deng
Tulane University https://orcid.org/0000-0002-0387-8818

Article

Keywords: HIV, Tat, transcription, High-throughput assay, Next generation sequencing, Mutation, Variant, Protein structure, Intragenic epistasis

Posted Date: July 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-708936/v1

GigaAssay – a high-throughput assay system for molecular functions in cells

Ronald Benjamin[1], Christopher J. Giacoletto[1,2,3], Zachary T. FitzHugh[1], Daniel Eames[1], Lindsay Buczek[1], Xiaogang Wu[1], Jacklyn Newsome[1], Mira V. Han[1,2], Tony Pearson[2,3], Zhi Wei[4], Atoshi, Banjeree[1], Shirley Shen[1], Hong Wen Deng[5], and Martin R. Schiller[1,2,3]

[1]Nevada Institute of Personalized Medicine, and [2]School of Life Sciences, University of Nevada, 4505 S. Maryland Parkway, Las Vegas, Nevada, 89154 United States of America
[3]Heligenics Inc., 833 Las Vegas Blvd. North, Suite B, Las Vegas, NV 89101, United States of America
[4]Department of Computer Science, New Jersey Institute of Technology, GITC 4214C, University Heights, Newark, NJ 07102
[5]Center for Biomedical Informatics & Genomics Tulane University, 1440 Canal Street, Suite 1621, New Orleans, LA 70112

ABSTRACT

High-throughput assay systems have had a disproportionally large impact on uncovering how cells function, as well as how misregulation can lead to disease. However, no high-throughput assay systems have been developed to systematically address how mutations impact molecular functions or cell processes in human cells. This is arguably one of the most critical assays because human pathology and treatment are largely based upon molecular functions. To address this challenge, herein we engineered, developed, and tested the first modular high-throughput molecular function assay system. Note that this is **not** a selection lethality screen! This "GigaAssay" single cell / one-pot assay system was adapted to study how variants impact HIV Tat-driven transactivation of a green fluorescent protein (GFP) reporter. We assayed all 1,615 Tat single and 3,429 double amino acid substitutions with no single mutant dropout. Each mutant was assayed with replicate observations in LentiX293T and Jurkat cells with an average of 100s of separately barcoded cDNA molecules and cell groups for each mutant. Each mutant had ~2,000X-90,000X sequencing coverage to measure its transcriptional activity and had p value ranging as low as 10^{-271}. Five independent assay performance assessments with benchmark data, individually tested clones, and replicate comparisons all indicate exceptional reproducibility, accuracy, and robustness. The shortcomings of alanine scanning mutagenesis and protein truncation studies are revealed by including exhaustive substitution tolerance and intragenic epistasis to the typical structure/function analysis(structure/function/tolerance/epistasis). This flexible and extensible technology enables a far more comprehensive holistic view of protein molecular function and yet with a highly simplified single-pot assay.

Keywords: HIV, Tat, transcription, High-throughput assay, Next generation sequencing**,** Mutation, Variant, Protein structure, Intragenic epistasis

INTRODUCTION

High-throughput screening (HTS) technologies (**Supplementary Table 1**) have transformed biomedical sciences and many of these technologies have been effectively commercialized and impact clinical care. Most of these technologies identify cell components such as DNA, RNA, or protein species, and some assess intermolecular interactions. CRISPR/Cas9 and RNAi genome wide screens can identify a gene as necessary for cellular or organismal process. Pathways and networks are often predicted from the resulting data, but these experiments only indicate a role for a gene, and do not conclusively assess mechanism.

There is no high-throughput assay to broadly assess molecular functions in the context of human or mammalian cells.[1,2] The molecular function are the key to understanding mechanism, disease etiology, and development of therapeutic drugs. Phage or yeast display, yeast 1- or 2-hybrid, DNA encoded libraries (DEL)s, and affinity mass spectrometry assess one general type of function, molecular interactions, although these interactions are not assessed in living mammalian cells.[3–6] More recently, lethality selection screens of mutant libraries in mammalian cells (called deep mutational scanning), identify loss or gain of function mutants.[7–10] For example cells with many different p53 mutants will not survive after two weeks of culture.[9] The presence of variants sequenced before compared to after culturing identify survival mutants with some activity and infer negatives for those mutants that do not survive the screening procedure. However, screens assaying survival are far downstream of many molecular functions.

Herein, we demonstrate proof-of-principle for a new modular high-throughput assay system we call the GigaAssay. The GigaAssay system is a one-pot single cell assay in living mammalian cells where each variant DNA molecule is individually barcoded, assayed by a fluorescent readout, flow sorted into pools, deep sequenced, and then the impact of each variant on a specific activity is bioinformatically deconvolved. The GigaAssay is **not** a survival screen where the negatives are not directly measured. The GigaAssay measures both positives and negatives for millions of individually barcoded DNA molecules producing a highly accurate and reproducible assay. The GigaAssay has the several other advantages over previously developed high throughput assays and screens (**Supplementary Table S1**). It is flexible, readily adapted to many cell process and molecular function assays in the context of living mammalian cells (**Fig. 1A**). It is a high-throughput assay capable of measuring tens of thousands of reads for each of millions of individually barcoded variant DNAs, where different genotypes are pooled for each amino acid substitution, thus reliable statistical probabilities and metrics can be calculated to determine the reliability of each measurement. Because of the high throughput, high reproducibility among samples, and reproducibility in different cell lines, the results are highly accurate with reliable statistics. In this one-pot experiment, each DNA variant molecule is tracked through the plasmid, viral, cell libraries, and then through cell pool separation by flow cytometry sorting.

RESULTS

To develop the GigaAssay, we assayed the HIV Tat transactivation of long terminal repeat-driven green fluorescent protein (GFP) expression in LentiX293T/LTR-GFP reporter cells as a model system. LTR is the long terminal repeat in the HIV genome. This system has the advantages of an established robust reproducible assay and abundant benchmark data is published for performance assessment. Furthermore, *Tat* is a small gene that is suitable for assay development and is of pathological significance for HIV infection and exit from latency.

Development of the GigaAssay system

After multiple rounds of testing and optimization, the current GigaAssay approach for Tat transactivation has several steps. Induction of the reporter by the *Tat* transgene was compared to empty vector and an inactivating mutation as controls for basal reporter expression (**Fig. 1A**). Once the reporter system and cassette were verified, a barcoded plasmid library was generated from a synthetic saturating mutagenesis dsDNA library. Each molecule in this library was randomly barcoded and used to prepare a lentiviral variant library. A human cell line was transduced with the lentiviral library at low MOI (0.1) to minimize double infections. A polyclonal cell library is selected for stable viral DNA integration into each cell with puromycin. Fluorescent and nonfluorescent cells are isolated into GFP^+ and GFP^- pools by flow cytometry sorting. gDNA is purified from each pool and a targeted barcoded *Tat* amplicon was cloned to make a Next Generation Sequencing (NGS) library for sequencing. The resulting paired-end read sequences are analyzed with a bioinformatics pipeline including several custom scripts to group barcodes, interpret variants, and calculate Tat transactivation activities for each mutant.

In the test system, a GigaAssay cassette encoded constitutively expressed Tat translated from a barcoded mRNA (**Fig. 1B**). Tat binds to the LTR of LTR-GFP in the LentiX293T/ LTR-GFP reporter cell line where the HIV LTR drives GFP expression (**Fig. 1B**). The Tat transactivation system was tested by transient transfection of individually prepared clones transfected into separate LentiX293T/LTR-GFP cultures and visualized by epifluorescence microscopy. Cells transfected with empty vector had little detectable GFP fluorescence, while those containing wild type (wt) Tat fluoresced as expected for Tat-driven GFP expression (**Fig. 1C, D**). Cells transfected with a C27S mutant that inactivates Tat transactivation had florescence as expected (**Fig. 1E**).[11] Similar results were obtained when cells with the same control viruses and were confirmed in Jurkat cells, a cell line derived from T cells that is a more suitable model for HIV protein studies (**Supplementary Fig. S1**).

The same control cells were sorted by flow cytometry, and profiles were used to set thresholds. Cells with empty vector or the C27S mutant had low GFP expression that was not different from a control cell with empty vector lacking a *Tat* cDNA. Cells expressing the reporter system with wtTat showed a high number of cells with GFP expression (**Fig. 1F**). These microscopy and flow cytometry sorting experiments verify the assay reporter system, reproducing previous observations.[11–13]

A saturating mutagenesis synthetic mutations ds-DNA *Tat* library (*Tat* accession number: AAK08486.1) was extended with synthetic 32 bp random barcodes into the 3' UTR by polymerase chain reaction (PCR). The library was then subcloned into a lentiviral vector. NGS and bioinformatic analysis of the plasmid library showed no dropout with measurements for all 1,615 possible single amino acid substitutions. A lentiviral library was prepared by co-transfection of lentiviral vectors encoding the library of mutant *Tat* cDNAs into LentiX293T cells. LentiX293T/LTR-GFP cells were transduced with the lentiviral library, and after poison selection for stable cells, GFP^-, mid-GFP, and GFP^+ cells were each sorted by flow cytometry; the mid-GFP cells were not analyzed (**Fig. 1G**). Pools of cells were gated based on GFP fluorescent intensities determined from negative and positive control cell samples. gDNA was isolated from each cell pool, targeted NGS libraries were constructed for the barcoded *Tat* cDNA, and samples were sequenced by NGS on an Illumina platform producing 2 x 250 nt paired-end reads. Samples were then analyzed with a custom NGS analysis pipeline (**Supplementary Fig. S2**). Summary statistics for different stages of the pipeline are shown in **Table 1.**

Impact of Tat single mutants on transcription

The Tat protein sequence and heatmap (**Fig. 2, Supplementary File S1**) show the variable impact of amino acid substitutions on Tat transactivation activity. The synthetic oligonucleotide and optimized cloning approach generated all possible amino acid substitutions for all positions excluding the start Met. Most substitutions (64%) had activities similar to wild type levels (meta $p < 0.05$ under Fisher's method), demonstrating a robustness for mutation tolerance in transcriptional transactivation (**Fig. 2, Supplementary Fig. S3A**). Approximately 36% of mutants had activities levels matching a set of known Tat LOF mutations (meta $p< 0.05$ under Fisher's method), indicating inactivation of transcriptional activity. Of those with reduced activity, 18% has less than 10% activity of wild type indicating that these mutants were largely inactive and likely loss of function (LOF). A bin plot in **Supplementary Fig. S3B** shows the distribution of Tat activities, relative to wtTat. Similar results were obtained among replicate samples and in Jurkat cells (**Supplementary Fig. S3C, D**). The tolerance data for each position is generally consistent with the Shannon entropy score for amino acid variability for Tat sequences in the Los Alamos HIV sequence Database.[14] However, since this is a new assays system we developed and tested a rigorous verification.

GigaAssay performance verification

Five independent verification tests demonstrate that the GigaAssay has very high reproducibility and accuracy (**Fig. 2B**). The first verification approach compared GigaAssay results to benchmark data from previous reports for Tat mutant transcriptional transactivation. 442 mutants for Tat were annotated from 43 papers. We removed mutants that had ambiguous activity reports or had multiple missense mutations or INDELs yielding a final list of 107 mutants from 28 papers (**Supplementary File S2**). The GigaAssay results were compared to these benchmark mutants using an activity threshold of 50%, with those > 50% classified as wild type activity and those <50% classified as reduced activity. Data in different samples were normalized to reads per million (rpm), and variants with a threshold below 2.5 rpm were

discarded. This threshold was determined by comparing performance metrics for different thresholds. Considering these mutant activities, we estimate GigaAssay performance statistics: accuracy = 0.93; sensitivity = 0.94; specificity = 0.89; PPV = 0.95; and NPV = 0.89 when results are compared to true positives and negatives from independently published benchmark data (**Table 2, Supplementary File S2**).[15]

A second verification was based upon an independent source of true negatives and positives measured in the GigaAssay. *Tat* exon 1, encoding the first 58 amino acids of Tat is the minimal region required for Tat transactivation activity.[16–18] Truncations mutants of less than 58 amino acids from the start Met were considered true negatives, while those 58 amino acids or longer were considered true positives.[16–18] Although not designed, due to errors in synthetic oligonucleotide synthesis we observed barcodes for Tat truncation mutants less than 58 amino acids long (n = 70), which were expected to be negatives, and truncation mutants that were longer than 58 amino acids (n = 8) which were expected to be positives with wild type activity (**Supplementary File S3**). Comparing the Tat mutant activities in LentiX293T to the true controls produced perfect GigaAssay performance statistics for: accuracy = 1.0; sensitivity = 1.0, specificity = 1.0; PPV = 1.0; and NPV = 1.0 (**Fig. 2C. Table 2**). This analysis of intrinsic GigaAssay data further supports its high accuracy and verification.

The third verification was comparison to independent testing of a set of separate Tat mutant clones. Prior to the experiment we randomly selected 18 Tat mutants, made stable LentiX293T/LTR-GFP cell lines expressing these mutants, and measured transcription activation of LTR-GFP reporter by flow cytometry sorting (**Supplementary Fig. S4**). These true positive and negative clone results were blinded until the GigaAssay was complete and then compared to the GigaAssay results. The performance statistics were: accuracy = 0.94; sensitivity = 0.75; specificity = 1.00; PPV = 1.00; and NPV = 0.92 (**Table 2**).

The fourth verification approach assessed the reproducibility of GigaAssay results among different samples. The LentiX293/LTR-GFP and Jurkat/LTR-GFP cells were transduced, selected, flow sorted, sequenced, and analyzed separately in duplicate. The global standard deviation for Tat mutant activities between technical duplicates was very low (SD = 0.02) with 1.0 being wild type activity. Mutant activities for replicate samples had a very high correlation (R^2 = 0.99) indicating high reproducibility which is further supported by comparing heatmaps for the two cells lines (**Fig. 2A**, **Supplementary Figs. S5, S6A**).

The fifth verification compared biological replicates in two different cell lines (LentiX293/LTR-GFP and Jurkat/LTR-GFP cells). Similar results for the performance statistics, reproducibility, and mutant activities were observed (**Fig. 2C**, **Table 2, Supplementary Figs. 3, S5B, S6**). There were only minor differences transcriptional activities for each mutant among the LentiX293/LTR-GFP and Jurkat/LTR-GFP cell lines indicated by comparison of activity heatmaps (**Fig. 2A** and **Supplementary Fig. S6A)** and high correlation of results for the two cells lines for matched mutants shown in a scatter plot in **Supplementary Fig. S7**; R^2 = 0.93). The major differences were for Tat

mutants that had intermediate activities. A comparison of Tat activities among the cell lines and samples shown in **Supplementary File S1.** Collectively, the correlation and variance analyses demonstrate reliable high data reproducibility and mutant behavior is nearly identical among replicates and between different cell lines.

We suspect that the high reproducibility is achieved from the experimental design where each individual variant cDNA has a separate random barcode that is tracked through the experiment. During the GigaAssay, each cDNA is individual barcoded and after transduction of recombinant viruses, each cell is uniquely barcoded. During selection, these cells divide forming clonal barcoded cell groups. For the different samples and cell lines there were ~561,000 barcoded cell groups after filtering, 179,763 of which were unique in LentiX293T cells. Each mutant in each replicate sample had an average of 102 independent barcodes with scatter plots showing a high correlation among replicate samples for each cell line (**Supplementary Fig. S8**). This supports saturated testing for all clones in the library. The percentage of GFP^+ reads for each group is calculated from the GFP^- and GFP^+ reads. Each barcoded cell group has an average of 273 filtered reads, while each mutant with multiple barcodes has an average of 25,662 reads for each replicate; the number of reads ranges from approximately 2,000-90,000 (**Supplementary Fig. S9).** Variant are called and Tat transactivation activity is calculated from these reads.

The transcriptional activities for each barcoded cell group with the same mutation are averaged and used to calculate statistics. The global standard deviation for the barcoded cell groups is 0.25 (**Fig. 2B**). While this standard deviation is considerably larger than that for replicate samples, we expect this variance and must consider that there are multiple contributing factors such as stochastic variation in growth of each barcoded cell group, random chromosomal lentiviral insertion sites that impact expression, and cells in different phases of the cell cycle, when larger datasets are analyzed at the level of individually barcoded molecules and cells.

Given the breadth of data produced in this GigaAssay experiment, we can reliably report metrics of confidence for the activity of each mutant. We first tested the hypothesis that the percentage of GFP^+ reads was different from 0.5 for each mutant (null model percentage $GFP^+ = 0.5$). The p value for each mutant in each cell with the distributions p values frequencies are shown in **Supplementary Figs. S10, S11,** and **Supplementary File S4**. These heatmaps and bin plots show that most p values (95%) for mutants in both cell lines reach statistical significance. We then tested whether having a mutant or the WT amino acid at the position influenced the percentage of GFP^+ reads. p and q values for mutants in LentiX293/LTR-GFP and Jurkat/LTR-GFP cells are in **Supplementary Figs. S12, S13,** and **Supplementary File S4**. Many p values are very low due to the large number of barcode replicates with a high average (n=102) for sample replicates. Some mutant p values ranged as low as 10^{-271}.

We compared each mutant's activity to that reported for sets of true positive mutants with established wild type activity and sets of true negative mutants with greatly reduced activity (**Supplementary Figs. S14-S16** and **Supplementary File S4**). In both cell lines,

most substitutions have either wild type (58-60%, $p < 0.05$) or reduced activity (20-23 %, $p < 0.05$) (**Supplementary Fig. S16**). Approximately 22-26% of mutants had neither WT or reduced activity and were moderately inhibited.

Intragenic epistasis of Tat double mutants.
We examined the intragenic epistatic double mutant interactions that arose from random errors in oligonucleotide synthesis and were tracked through the experiment. There was a total of 3,429 double mutants that had at least two barcodes. The activities of double mutants were compared to corresponding single mutants and assigned as positive, negative, or no intragenic epistasis. For the two cell lines 85-90% of double mutants had no epistasis,1-2% had positive epistasis and 9-13% had negative epistasis (**Fig. 2C, D**). Considering that the double mutants were sampled from two cell lines, each with replicate samples, only 51 ambiguous epistatic types were observed yielding a lower bound of and error rate of 1.5%. The observed rate of intragenic epistasis (10-15%) is less than other recent estimates (32%-74%), but our results are for a direct molecular function in living human cells, whereas previous studies of HSP90, TEM-1 β-lactamase, and ΦX174 were of virus, bacteria or yeast fitness.[19–21]

Structure/function/tolerance/epistatic impact of Tat mutants.
The saturation mutagenesis profile enables an improved interpretation of mutation tolerance of secondary structure, post-translational modifications (PTM)s and protein-protein interactions (PPI)s on Tat activity. Based on the saturating mutagenesis experiment and its interpretations, we suggest that the typical structure/function analyses be expanded to structure/function/tolerance/epistasis. The latter terms are added to reflect which amino acid substitutions can be tolerated to preserve structure/function and the interdependence of multiple substitutions. Since our experiments are relevant to the interpretation and/or confirmation of most, if not all of the hundreds of previously published reports on Tat mutants, we limit the discussion to a few examples herein.

Several secondary structure positions are sensitive to mutations. Tat is mostly random coil with one helix and three turns. Mutations were well tolerated in the first turn, but not in the second and third turns (**Fig. 2A**). The only mutations in the first turn (^{7}R-L^{8}) with low activity were R7P, L8P, and L8G The. second turn starting at K28 has the sequence ^{28}KKCCF^{32} (**Fig. 2A**). No mutations at C30 were tolerated and only C31A and C31S with small volume amino acids substitutions retained activity, supporting steric hindrance constraints of ϕ and ψ angles for amino acids in turns. Only conservative large hydrophobic and some small aliphatic substitutions of F21 retained activity. Mutations in the third turn (K41, A42) were generally not tolerated with only reduced activity for A42G and A42C. Scattered mutations in the helices and the random coil regions had reduced activities, and were more tolerant of mutations, especially after position 46 in the C-terminus. Notably, no substitutions were tolerated at K41 or in six C residues in the Cys-rich domain as previously reported.[22] C31 did tolerate substitutions of S, T, or small aliphatic amino acids and C31S was previously known to be active and is a natural variant in clade C Tat proteins.[23]

We examined if mutation of any of the residues with PTMs impacted Tat activity (**Fig. 2A**). Tat has 18 reported PTMs of five different types.[24,25] Mutations of most covalently modified residues at a single PTM positions did not impact transcriptional activity. We show a couple of examples that highlight the benefits of GigaAssay results in interpretation. Tat is ADP ribosylated at E2 and E9 and several mutations at these positions are not known to be ADP ribosylated, but retained activity, suggesting that ADP ribosylation is not necessary for Tat activity[26]. This conclusion cannot be conclusively resolved from previous published work without the new tolerance data. K28 is acetylated and is required for Tat activation, which is reported to be based upon higher affinity and stability of Tat–CycT1–TAR complexes.[27] The tolerance of other amino acids (K28P, K28C, K28R, K28V, and K28A) preserves transcriptional activity and indicates that acetylation is not an absolute requirement, thus there may be other possible mechanisms that increased affinity for the transcriptional complex. For example, K28 is in turn 2, a secondary structure element that is prone to loss of activity when mutated. However, in this report K28R was inactive, thus different genetic background may also explain the difference. Nevertheless, the example shows how GigaAssay results aid in the rigor of interpreting a proposed mechanism

Tat has known binding sites for about 20 other proteins (**Fig. 2A**) of which nine have substitutions that inhibit transcriptional activity. Most interactions of these PPI sites are in a hotspot from residues 29-60.[25,28] The most sensitive interactions are that of Cyclin T1 and Importinβ. CyclinT1 makes contacts with 15 amino acids in a Tat crystal structure, mostly in the Core region; 13 positions have at least one mutant that blocks Tat transactivation, and CyclinT1 is a key protein for recruiting RNA polymerase.[29,30] The Importinβ interaction site (^{50}KKRRQRRRAHQ^{60}) was not very sensitive to single substitutions, although acidic substitutions from 50-56 mildly impaired activity, although this also could be through the overlapping RNApol2 binding site as well. The robustness of this site (^{52}RRQRRRA^{57}) to tolerate all substitutions, likely reflects robustness for key Importinβ and CyclinT1/CDK9 complex recruitment.[27,30–33] The RNApol2 site also overlaps with EGR, P53, CA150 and 11S proteosome binding sites. The Gigassay data aids with interpretation of PPIs to select those that have the largest impact on function.

The Importinβ binding site is also of interest based on Tat truncation mutants. The activities of 78 truncation mutants were measured from nonsense mutants across both cell lines. The mutants were introduced as errors during oligonucleotide synthesis, and were tracked through the experiment (**Fig. 2A**). The 70 mutants with a truncation before amino acid 58 had no or little detectable activity, whereas the 8 mutants after amino acid 58 had nearly full activity. The near-perfect accuracy, PPV and NPV for this analysis is consistent with a protease cleavage site between residues 57-58, which is also an exon boundary, and previous observations support similar truncation tolerance for longer trucations.[16–18,34,35]

However, almost any missense mutation from S46-E86 was tolerated, which is not consistent with the truncation mutants tolerated up to R57. The region of difference (S46-R57) contains a nuclear localization motif, and binding sites for Importinβ, P53, and EGR. The most likely explanation is truncations that remove the nuclear localization

signal block localization of Tat to the nucleus and its transcriptional activity.[36–39] In the presence of any single point mutant in this region, Tat is still localized to the nucleus. This is supported by GigaAssay results showing that mutation of all positively charged residues in the nuclear localization motif to negatively charged residues, reduced, but did not block the transcriptional activity of Tat. Consistent with this hypothesis, others have examined the NLS and only double mutants in this region block nuclear localization, and a peptide containing the NLS is sufficient to localize other proteins to the nucleus.[36–39] We identified 16 binary negative epistatic interactions with mutation of two positions in the NLS and no positive epistasis in the NLS, strongly supporting this hypothesis: R49M/K50H, R49M/K50Q, R49S/K50H, R49E/R52P, R49Y/R55Q, Q50/54E, R49W/R55L, K50Q/K51N, K50N/K51T, K51N/R52I, K51F/R56L, K51W/R56Q, R52L/R53V, R53T/R55L, R53S/R55L, R53W/R55L. The intragenic epistatic interactions cover the entire NLS (**Supplementary Fig. S17K**).

We further examined the impact of mutations on transcriptional activity by coloring the surface of the 3D structure of Tat, which can then be compared to other spatial positions, regions, and secondary structure of Tat (**Fig 2A, 3A-C**).[40] Ala Scanning mutagenesis is an accepted approach to identify positions important for different functions.[41,42] Alanine scanning identifies 18 LOF mutations scattered across the N-terminal half of the protein (**Figs. 2A, 3D)**. Scanning with other amino acids such as Pro or Asp are more sensitive with 23-24 LOF mutations, Cys scanning is less sensitive with only 9 LOF positions that are a subset of Ala scanning, and Gly scanning has less constraints with more flexibility in dihedral angles and can identify additional amino acids (**Figs. 2A, 3E, F, Supplementary Fig. S17E,** respectively).[43] When these approaches are compared to the minimum, average, or maximum activity for all of the 19 amino acid substitutions at each position or the positions that do not tolerate substitutions (**Figs. 2A, 3G-I, Supplementary Fig. S17F**), information from the residue specificity in saturation mutagenesis is not completely captured in these types of 3D surface maps.

A novel approach that better summarizes the additional information gained from saturation mutagenesis, is to score positions for physiochemical groups with similar side chain properties (small aliphatic, large hydrophobic, polar noncharged or charged, negatively charged, positively charged); here positions were scored with the Mathews Correlation Coefficient (MCC). This approach better conceptually segregates substitution tolerance for each position, whereas alanine or other amino acid scanning does not capture this specificity. The heatmap with MCC for groups of substitutions (**Supplementary Fig. S18**) identifies positional tolerance: F32 only tolerates a large hydrophobic residue, G15 only tolerates a polar-noncharged, and C31 only tolerates amino acids with smaller volumes.

Surface plots of MCC heatmaps for different physiochemical properties reveal spatial relationship of tolerance not captured in the heatmaps. There is little specificity for amino acid tolerance over most of the protein including residues S46-E86, a random coiled region that tolerates nearly all substitutions (**Fig. 3J-I**, **Supplementary Figs. S17G-L, S18)**. However, the specificity for different groups of substitutions is clustered

in the Cys-Rich and Core regions (**Figs. 2,3B**). These regions have seven residues that do not tolerate any substitution, have reduced activity in Ala, Pro, and Gly scanning mutagenesis, do not contained buried residues, and include key binding sites for CyclinT1, Importinβ and several other PPIs (**Figs. 2, 3B,D,E, Supplementary Figs. S17E-L, S18G-J, S19**). The new Physiochemical Tolerance Surface plots based upon MCCs better identify residue tolerance of each position and their relative spatial locations, as well as surface accessibility (**Fig. 2, Supplementary Fig. S18J)**.

Likewise, reorganizing the activity heatmap by side chain volume reveals tolerance of amino acid substitutions for sidechain with a constrained volume (**Fig. 2A, Supplementary Fig. S20).** Positions Y26, F32, and F38 prefer large amino acids, positions E9, L10, G15, S16, T23, C31, M39, and A42 prefer small amino acids, and positions D5, C25, L43, and I45 favor medium sized amino acids. Overall, some positions do not tolerate substitutions, some positions tolerate substitutions driven by side chain volume, whereas a well-defined spatial region tolerance is driven by a combination of secondary structure and/or physiochemical properties of sidechains.

DISCUSSION/CONCLUSION

Previous advances in assay or screening throughput such as DNA and RNA sequencing, Y2H, phage display, microarrays, have had a vast impact on biomedical science and we expect that the GigaAssay will be no different. Herein, we demonstrate proof-of-principle of a GigaAssay prototype test a HIV *Tat* variant with an LTR-GFP reporter assay producing one of the most detailed functional maps of a protein.

The GigaAssay has several advantages over existing methods. The method directly assays the function of the protein. Unlike deep mutation scanning (DMS) screens, e.g.[8], all mutations are directly measured in the GigaAssay and not inferred from differences between pre-screen and screened samples. Our assay of >561,000 individually barcoded mutants in living cells is at least a four order of magnitude enhancement over routine low-throughput cell based assays. For example, six Pro residues in a PxxP motifs were used to measure the impact of these SH3 binding motifs on guanine nucleotide exchange factor activity[44]. Because individual molecule are barcoded, we are able to signal average Tat activities from large numbers of independent cells, thereby yielding robust reproducibility, high accuracy, and a statistic for reliability of each mutant. Furthermore, all mutants are assayed under standardized conditions in the same cells with the same genetic background producing a consistency that is not seen in the literature where mutants are often studied by multiple labs with separate assay systems, genetic backgrounds, and conditions.

The saturating mutagenesis map of Tat not only reveals shortcomings in routine interpretation of mutagenesis data, but GigaAssay results create a context to improve interpretation with a new structure/function/tolerance approach. Herein, alanine scanning (**Fig. 2, Supplementary Fig. S3D**) identifies key positions, but lacks sensitivity and misses the importance of some positions. New visualization methods of plotting saturation mutagenesis heatmap data with other structure function data, as well

as physiochemical MCC surface plots enable new opportunities for visual interpretation, enabling a holistic investigation of molecular functions. For example, the missense mutation analysis compared to truncation mutants shows differences that can be explained by an NLS signals (**Fig. 2** and **Supplementary Fig. S18E**). This reveals previously unrecognized limitations of truncation mutant studies. The context also produces a framework to identify ambiguities and potential misinterpretation of mutant data regarding structure, PTMs, and PPIs, especially where sites overlap with each other, as is the case for many regions of Tat.

If impact of the observed percentage of double mutants with of intragenic epistasis 10-15% of double mutants (n = 3,429) is observed in clinically relevant genes and implemented for genetic tests and companion diagnostics, the impact on patient care cannot be overstated. This inclusion of tolerance and epistasis should significantly improve upon routine structure-activity relationship (SAR) studies, extended to structure-activity-tolerance-epistasis relationship (SATER).

We identify a pattern in Tat where PPIs and PTMs are selected for robustness, while key structural elements or substitutions (e.g. proline) that impact structure are not as well-tolerated. The opportunities to interpret this data set are too vast for one paper and will transform interpretation of future experiments on *Tat* that will lead to a better understanding of HIV latency. We see no reason why the GigaAssay cannot be applied to other genes and assays to address other important questions in biomedical sciences.

METHODS

Cloning

All primers and synthetic oligonucleotides used for cloning and PCR are in **Supplementary File S5**.

The plasmid pLjm1_mcs was made by introducing compatible EcoRI, SalI, and AsiSI restriction enzyme sites in the pLjm1-Empty (Addgene) vector for cloning of the *Tat* variant library. *Tat* or mutant *Tat* encoding a C27S mutation was PCR amplified from pNL4-3 as a template with Q5® High-Fidelity DNA Polymerase (New England Biolabs) and cloned into EcoRI/SalI digested pLjm1_mcs1. For generating a LentiX293T/LTR-GFP reporter cell line, a plasmid harboring LTR-GFP and blasticidin S resistance was constructed. The LTR-GFP cassette and Blasticidin S resistance (*bsr*) gene were amplified by PCR with pNL4-3 (NIH AIDS reagent program), pEGFP and LentiCRISPR-v2 Blast (Addgene) as templates. LTR, GFP, and *bsr* amplicons were fused by inverse PCR using Q5® High-Fidelity DNA Polymerase. The fused amplicons were cloned into pAAVS1-Puro-DNR (Origene) previously digested with SpeI and EcoRI.

Generation of barcoded variant plasmid library

A double stranded (ds) DNA library containing HIV-1 *Tat* cDNAs with sequences for all the possible single amino acid mutant mutants (n =1,615 *Tat* mutants) was synthesized by Twist Bioscience. The dsDNA from each well of 96-well plates were pooled and a

single round of overlap PCR extension appended random 32mers oligonucleotides to the 3' untranslated region. The synthesized dsDNA library has a 3'-overhang sequence after the stop codon that overlaps with the 5' overhang sequence upstream of the 32mers random oligonucleotide sequence. The pooled ds DNA library and the random oligomer were mixed in 1:10 molar ratio, denatured, and annealed. Hybridized DNA was extended with the Q5® High-Fidelity DNA Polymerase (New England Biolabs) for one cycle of PCR. The 50 µl of PCR reaction mix was then treated with 2 µl of Exonuclease I (New England Biolabs), incubated at 37°C for 15 min, and DNA was purified by PCR cleanup kit (Macherey-Nagel).

The purified DNA was digested with EcoRI-HF (New England Biolabs) and AsiSI (New England Biolabs) for 3 h at 37°C and ligated into EcoRI-HF/AsiSI digested pLjm1_mcs plasmid (molar ratio vector: insert = 1:3) with electroligase (New England Biolabs). Ligation reactions (12) were pooled, purified with a PCR cleanup kit, and drop dialyzed on MF-Millipore® Membrane Filter, 0.025 µm pore size (Millipore Sigma). The purified ligation reaction mixture was electroporated into E. cloni 10G ELITE electrocompetent cells (Lucigen), plated on prewarmed LB ampicillin plates, and incubated for 18 h at 37°C. Transformants were scrapped and plasmid library from the pooled cell suspension was isolated using EndoFree Plasmid Mega kit (Qiagen).

Production and titering of lentiviral libraries

Lentiviral libraries were produced in LentiX293T cells (Takara). Approximately 3 million LentiX293T cells were seeded in 100 mm petri dish and grown in 10 ml complete DMEM media [(DMEM+10% Fetal Calf serum), Gibco] for 24 h. Plasmids pLjm1_Twist *Tat* Library (8.5 µg); pMDLG/pRRE (Addgene, 7.6 µg); pRSV/pRev (Addgene, 4.0 µg); pMD2.G (Addgene, 4.0 µg) were diluted to a final volume to 613 µl in a 15 ml conical tube. $CaCl_2$ (87 µl of 2M) was added to plasmid mixture. 2XHBS (700 µl) was added dropwise to the above transfection mix with gentle stirring in a circular motion. The transfection mix was incubated for 15 min and added dropwise to the cells in a 100 mm petri dish. The cells were incubated at 37°C for 12 h in a CO_2 incubator at 37°C with a 5% CO2 atmosphere. Post-transfection (12 h), the calcium phosphate-containing medium was replaced with 7 ml complete media (DMEM+10%FBS) and incubated for 48 h in CO_2 incubator at 37°C with a 5% CO_2 atmosphere. Spent media from confluent transfected LentiX293T cells was filtered through a 0.45 µm Uniflow syringe filter (Cytiva Whatman). Aliquots of the filtered spent media with the lentivirus (100 µl to 5 ml) were stored in at -80°C.

Lentiviral vectors for specific clones were produced in LentiX293T cells. Briefly, the 0.6 million LentiX293T cells was seeded in a well of a 6-well plate. After 24 hours, cells were co-transfected with pLjm1-mcs, pLjm1-Tat, or pLjm1-TatC27S (1 µg); pMDLG/pRRE (1.0 µg), and pRsv-Rev (0.5 µg) and pMD2.G (0.5 µg) transfecting with Lipofectamine LTX (Invitrogen) at a 1:3 ratio [DNA(µg): Transfection reagent(µl)]. After 6 h of incubation, media was replaced, and cells were cultured in complete media for an additional 48 h. Cell supernatants containing lentivirus were collected, filtered through a 0.45 µm syringe filter (Millipore), and stored at -80°C.

Lentiviruses were titered by seeding10,000 cells/well in 96 well plate and cultured in 200 µl of complete DMEM media (DMEM+10% FBS). After 24 h, 100 µl of serial dilutions of lentivirus were added after removing majority of the spent media from the wells and incubated 4 h. Complete DMEM media (100 µl) was added and incubated 24 h. Spent media (100 µl) was removed, replaced with DMEM media containing puromycin (Invitrogen, 1.5 µg/ml final concentration), and incubated for 96-120 h. The cells were inspected for viability under the microscope and colonies were counted to calculate the infectious unit/ml.

Generation of LentiX293T/LTR-GFP and Jurkat/LTR-GFP reporter cell lines
LentiX293T cells (0.6 million) were seeded in the well of a 6-well plate and grown in 3 ml of complete DMEM media. After 24 h, a GFP reporter plasmid (1.5 µg) carrying LTR-GFP and the blasticidin S-resistance (*BSR*) gene was transfected in LentiX293T cells and incubated for 48 h. Transfected cells were selected for blasticidin S [(5 µg/ml), Invitrogen] resistance for 14 days, exchanging DMEM media with the poison every 3 days. Cells were trypsinized and 100,000 cells were serially diluted in 96-well plates. After 14 days of incubation, single colonies were screened after expansion.

For confirming lentiviral integration, gDNA was isolated, *Tat* was amplified with GFP-FP and GFP-RP primers, amplicons were subcloned, and sequenced. Tat transcriptional activity was measured in a subculture of each clonal cell line. Cells culture in 96-well plate were transfected with 50 ng of wt*Tat* expression vector and cultured for 48 h. Transactivation-induced GFP expression was evaluated by Nikon TE2000E epifluorescence microscopy. The clonal reporter cell lines were propagated and stored at -80^{o}C.

Stable cell libraries and cell lines
LentiX293T/LTR-GFP cells (33 million) were transduced with the *Tat* variant lentiviral library at a multiplicity of infection (MOI) of 0.1. After 24 h of infection, cells were cultured and maintained in complete DMEM media supplemented with puromycin (1.5 µg/ml). After 5 days, confluent cells were harvested, counted, and washed once with 1X PBS before fixing and isolating gDNA for NGS of the *Tat* amplicon.

Jurkat/LTR-GFP cells (90 million cells) were seeded and transduced with 0.1 MOI of lentiviral library for 4 hours. One day after transduction, the cells were selected for viral survival in RPMI 1640 media [(RPMI 1640+10% FBS), Gibco] supplemented with puromycin (1 µg/ml). After 5 days, the cells were counted, washed with 1X PBS, and fixed for flow sorting and subsequent isolation of gDNA.

For performance evaluation of the GigaAssay, 18 random mutants of Tat, as well as empty vector and wtTat were stably expressed in LentiX293T/LTR-GFP cells**.** Approximately 0.15 million cells were seeded in a well of a 24 well plate and incubated for 24 h. Cells were transduced with lentivirus, selected, and maintained in complete DMEM media with puromycin (1.5 µg/ml) for 96 h. Cells were harvested and sorted by flow cytometry to assess for LTR transactivated GFP expression. The same stable cell

lines were created in Jurkat/LTR-GFP cells. Selected clones for empty vector, wt*Tat*, and *Tat* C27S were stored at -80°C.

Flow sorting of cells and deep sequencing

One fourth of the LentiX293T/ LTR-GFP and one tenth of the Jurkat/LTR-GFP cells were harvested, gDNA isolated using Qiagen DNeasy Blood & Tissue Kit, and sequenced to evaluate library representation before Flow Sorting. The remaining cells were fixed in 2% paraformaldehyde/PBS for 10 minutes, washed twice with 1X PBS and resuspended in 1X PBS for analysis by flow sorting (Sony 800S Cell sorter). Cells were sorting into three bins of GFP signal intensity (low-GFP, mid-GFP and high-GFP) gated with threshold determined for cells stably expressing wtTat for maximal transactivation of LTR-GFP, and cells stable expressing a Tat C27S mutant or empty vector for low background of basal transactivation of LTR-GFP.

For deep sequencing, primers were designed to flank the *Tat* targeted region from gDNA and incorporate the NGS sequencing adaptors. gDNA was amplified by PCR with NEBNext Q5 Hot Start HiFi PCR Master Mix. The PCR protocol denatured strands at 98 °C for 30 sec only in the first cycle followed by: denaturation at 98 °C for 10 s, annealing at 58 °C for 15 s, elongation at 72 °C for 30 s, and a final elongation for 2 min. NGS libraries for each sample category used 10 NGS library forward primers and 1 NGS library reverse primer. The forward primers were common for all the sample categories and the reverse primer being unique for each sample. The *Tat* amplicons were pooled and 20 µl of the sample was purified by gel extraction with Ampure-XP beads (Beckman Coulter). All the samples were pooled and sequenced with a Novaseq 6000 sequencing platform at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign. This SP flow cell produces approximately 2 x 250 nt paired-end reads. 18 samples were sequenced (synthetic dsDNA Tat variant library, plasmid library, selected cell libraries in LentiX293T and Jurkat cells (in duplicate), Flow sorted low-GFP, mid-GFP, and high-GFP cells for each cell line (in duplicate).

Processing NGS data with a bioinformatics pipeline

Paired-end reads were processed with a multistep bioinformatic pipeline BaseSpace and resulting reads in bcl files were converted into FASTQ files with BCL2FASTQ, read quality is assessed with FASTQC (**Supplementary Fig. S2**).[45] Paired end reads for all samples are merged with FLASH to build complete *Tat* contigs.[46] Contigs were quality trimmed with Trimmomatic.[47] Adapters are trimmed, and 32 nt barcodes are isolated with CutAdapt and Starcode is used to group bacodes.[48,49] The sequence reads are demultiplexed into subsets of read sequences for each cell clone based on unique barcodes with a custom Python script that processes the output of Starcode. Resulting reads are then aligned to the *Tat* cDNA with BWA MEM.[50] The BAM file with nucleotide variants are called for each subset of *Tat* contigs (cell clones) and output as a VCF file with BCFtools (mpileup).[51] Custom Python scripts are used to identify the amino acid substitution for the VCFs, the number for reads for each barcode in each sample, and the barcodes groups for cells with the same amino acid substitutions. The PyVCF library was used in scripts that gathered the information for each variant from the VCF files.[52] Read counts and read depths for each barcode and each amino acid substitution in

each sample are normalized to the number of reads/million and activity is measured by the percentage of GFP^+ reads for each barcode and each mutant.

Data analysis, statistics, and figure preparation

Statistics are calculated for each mutation. We assume there are n cell lines (biological replicates) and each cell line has m technical replicates. For each barcode (group) in a sample, we calculate the percentage of the number of reads in the GFP^+ group vs the total number of reads in both GFP^+ and GFP^- groups, denoted as h ratio ($h \in [0,1]$). We expect a high h percentage for wild type, while a low h percentage suggests a mutant. Then 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. We use a one sample t-test to evaluate 1) whether the mutant has a significantly different number of reads in the GFP^+ group compared with the GFP^- group within a technical replicate, and 2) whether the mutant has a significantly different number of reads in the GFP^+ group compared with the GFP^- group among different cell lines based on biological replicates (null hypothesis H_0: h =0.5).

In addition to the t-test comparing the GFP^+ ratio among the mutants, we also devised an association test between the genotype (Variant/WT) and GFP expression (binary variable GFP^+ or GFP^-. We used a mixed effect logistic regression, with random intercepts for barcodes and replicates to model the nested structure in our experimental design. For the WT control populations, we used the cells with no variant calls (sequences identical to the reference). Each variant was compared against the common WT control population. The model M1 with genotype included as fixed effects was compared to a null model M0 without genotype in a likelihood ratio test (LRT). Similar to Genome-Wide Association Studies (GWAS), a significant result indicates that the variant/WT is associated with the percentage of GFP^+ cells. For variants where the model fit was singular, we simplified the model by dropping the random effects. p-values were false discovery rate (FDR)-adjusted 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)

We classify mutants with high h percentage as wild type and a low h percentage as a LOF mutant. To estimate type I error for the classification, we compiled a list of true mutants with wild type transcriptional activity and true LOF mutants with low activity (**Supplementary File S2**). Then we fit their h percentages with a beta distribution as the null distribution. Specifically, for the wild type detection, we use the true mutant as the null, and vice versus, for the mutant detection, we use the wild type as the null. Moment estimators are used for estimating the model parameters. The p values for different cell lines are combined using Fisher's method into a global test p value.

Performance metrics of accuracy, sensitivity, specificity, positive predictive value and negative value are based upon standard formulas.[15]

Figures were prepared with PowerPoint, Excel, FlowJo, R, and Pymol. Bin, Bar, and Pie plots, as well as saturating mutagenesis heatmaps were generated with Excel. Values for saturating mutagenesis heatmaps and 3D surfaces plots were generated with custom python scripts. 3D physiochemical tolerance surface plots for the amino acid tolerance at each position are based upon MCCs for physiochemical properties and colored with gradients from blue to white to magenta. Magenta is the highest MCC and blue is the lowest MCC. MCC is calculated for groups of amino acids with similar physiochemical properties.[15] Solvent accessible surface area (SASA) was calculated for the Tat structure (1TIV) with the Accessible Surface Area and Accessibility Tool.[53] Residues are considered buried if less than 10% of surface area is exposed to solvent (**Supplementary Fig. S18J**).

The MCC formula is calculated with the following data definitions for large hydrophobic amino acids, at a position in Tat as an example: If either Phe, Tyr, or Trp have > 50% activity they are true positives and if the other amino acids have <50% activity they are true negatives. If either Phe, Tyr, or Trp have <50% activity they are false positives and if the other amino acids have >50% activity they are false negatives. We also considered the wild type amino acid to be a true positive when it was in the physiochemical group, and as a true negative when it was not. The MCC captures the tolerance for types of amino acids at each position and when mapped the surface of the 3D structure, is a new visual mining approach to reveal the spatial relationships of amino acids tolerances and their relevance to other Tat functions.

Funding
MRS Grants: NIH: R21AI116411, R15GM107983, R21AI078708, R56AI109156, P20GM121325, COBRE and 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.

Acknowledgements
We thank Dr Edwin Oh, Richard Tillet, and Shang Shen 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 Thanks 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 and for allowing us to use her electroporator. We appreciate discussions with Drs. Qing Wu and Michael F. Lin about statistical assessment of the GigaAssay results.

Conflict of Interest
The GigaAssay 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 and CJG are employees of Heligenics which has licensed the GigaAssay technology 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.

Author Contributions
MRS conceived the general invention of the GigaAssay, obtained funding, and wrote the paper. MRS and RB designed and interpreted experiments. RB performed and supervised all molecular biology, NGS sequencing, and cell biology experiments with technical assistance from AB, LB, and DE. CJG, JN, ZF XW, and MRS were responsible for the bioinformatics work. MRS, ZTF, RB, AB, and CJG prepared figure and tables. DE, LB, and MRS annotated the Tat benchmark data. SS conducted NGS sequencing in developmental experiments. ZW, MH, CJG, and MRS designed and executed the statistical analysis.

Figure Legends

Figure 1. Design and implementation of the GigaAssay for Tat transcriptional activation. A. Design of GigaAssay system. Propagation of the recombined cells under poison selection. Cell sorting based on GFP reporter expression. gDNA is isolated, and a targeted Tat amplicon library is prepared and sequenced by NGS. **B.** Schematic representation of Tat dependent LTR transactivation inducing GFP expression. **C-F.** Epifluorescence microscopic images of LentiX293T/LTR-GFP cells transfected with GigaAssay plasmids: null/LTR-GFP (**C.** - control); wtTat/LTR-GFP (**D,** + control); and an inhibitory mutant[11], C27S-Tat/LTR-GFP (**E,** - control). **F.** Flow cytometry of GigaAssay controls in LentiX293T/LTR-GFP cell to define gates. **G.** Flow cytometry sorting of GigaAssay LentiX293T/LTR-GFP cell library cells with gates defined by - and + controls.

Figure 2. Heatmap showing Tat transactivation activity for a saturating mutagenesis GigaAssay. A. Heatmap for mutated amino acid for each position in Tat. The color gradient represents the level of Tat transactivation activity score measured by GFP^+ / (GFP^+ + GFP^-) reads for each barcode averaged for each mutant. Black boxes are the wild type amino acids and grey boxes are null values. A high ratio (blue) indicates a strong enhancer and a low ratio (red) indicates a strong inhibitor. A color key is shown. **B.** Assay reproducibility and verification. **C, D.** Quantitation of types of intrageneic epistasis for double mutants in LentiX293T/LTR-GFP (n =2,465) and Jurkat/LTR-GFP (n=1,633) cells, respectively

Figure 3. Tat mutant impact on structure/function. All surface maps are on the wtTat 3D structure (PDB: 1TEV) with one member of each pair rotated 180^o about the Z axis: **A.** Amino acid positions on Tat backbone**. B.** Regions of Tat[18]**. C.** Secondary structures. **D.** Ala scanning substitutions. **E.** Pro scanning substitutions. **F.** Cys scanning substitutions. **D-F.** Residues colored black are for reference amino acids that match the type of scanning. A gradient of yellow with no activity to green with full activity is shown. Gly scanning is shown in **Supplementary Fig. S17E.** Minimum **(G)**, average (**H)**, and maximum (**I**) transactivation activity heatmap for all substitutions. A gradient of red with wild type activity to yellow with no activity is shown. Positions that do not tolerate any substitution are shown in **Supplementary Fig. S17F, J-L.** Physiochemical tolerance surface plots for small aliphatic, polar uncharged, and large hydrophobic amino acids, respectively (see Methods). A gradient of blue to white to magenta ranging from lower to higher MCC scores for each position for the class of amino acids indicated is shown. Residues where all substitutions are inactive (C25, C27, C30, C33, C34, C37, and K41) are colored blue purple. MCCs for polar charged amino acids, and those separated by positively and negatively charged amino acids (**Supplementary Fig. S17G-I).** The color key is as in the **Fig. 2A**. For comparison, truncation mutants, PTMs, and PPIs are shown in **Supplementary Fig. S18E-I**. Abbreviations are: Single letter amino acid code, AA = amino acid, SS = secondary structure, SASA = solvent accessible surface areas, PTM = post translational modification, NLS = nuclear localization signal, PPI = protein-protein interaction.

TABLES

Table 1. Summary pipeline statistics for next generation sequencing of GigaAssay libraries.

Step	Reads (start)	% yield (by step)	Barcode	Mutants	Runtime
Fastqc	368,408,071	92.2	-	-	226 min
Flash	327,536,831	88.9	-	-	210 min
Trimmomatic	323,651,918	98.8	-	-	175 min
Adapter Trimming	323,163,108	99.9	-	-	86 min
Barcode Extraction	323,163,108	100	-	-	60 min
Barcode Grouping	294,692,904	91.2	-	-	30 min
Demultiplexing	294,692,904	100	-	-	90 min
Variant Calling	164,852,217	55.9	Total: 561,000 Unique:180,091	1,774	9.5 days
Filtering – 2.5 rpm	-	-	Unique:179,763	1,685	1 sec

Table 2. Summary performance statistics for next generation sequencing of GigaAssay libraries.

Cells	Verification Method	Accuracy	Sensitivity	Specificity	PPV	NPV
LentiX293T	Benchmark data	0.93	0.95	0.89	0.95	0.89
LentiX293T	Nonsense mutations	1.0	1.0	1.0	1.0	1.0
LentiX293T	Verified clones	0.94	0.75	1.0	1.0	0.92
Jurkat	Benchmark data	0.94	0.94	0.92	0.96	0.89
Jurkat	Nonsense mutations	0.91	0.90	0.93	0.96	0.82
Jurkat	Verified clones	0.87	0.50	1.0	1.0	0.86

REFERENCES

1. Picot, J., Guerin, C. L., Le Van Kim, C. & Boulanger, C. M. Flow cytometry: retrospective, fundamentals and recent instrumentation. *Cytotechnology* **64**, 109–130 (2012).
2. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
3. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).
4. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).
5. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032 (1999).
6. Wilson, T. E., Fahrner, T. J., Johnston, M. & Milbrandt, J. Identification of the DNA binding site for NGFI-B by genetic selection in yeast. *Science* **252**, 1296–1300 (1991).
7. Starita, L. M. *et al.* A Multiplex Homology-Directed DNA Repair Assay Reveals the Impact of More Than 1,000 BRCA1 Missense Substitution Variants on Protein Function. *Am. J. Hum. Genet.* **103**, 498–508 (2018).
8. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
9. Kotler, E. *et al.* A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol. Cell* **71**, 178-190.e8 (2018).
10. Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Hum. Genet.* **137**, 665–678 (2018).
11. Ulich, C. *et al.* Functional domains of Tat required for efficient human immunodeficiency virus type 1 reverse transcription. *J. Virol.* **73**, 2499–2508 (1999).
12. Dorsky, D. I., Wells, M. & Harrington, R. D. Detection of HIV-1 infection with a green fluorescent protein reporter system. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirology Off. Publ. Int. Retrovirology Assoc.* **13**, 308–313 (1996).
13. Siekevitz, M., Feinberg, M. B., Holbrook, N., Wong-Staal, F. & Greene, W. C. Activation of interleukin 2 and interleukin 2 receptor (Tac) promoter expression by the trans-activator (tat) gene product of human T-cell leukemia virus, type I. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 5389–5393 (1987).
14. Kamori, D. & Ueno, T. HIV-1 Tat and Viral Latency: What We Can Learn from Naturally Occurring Sequence Variations. *Front. Microbiol.* **8**, (2017).
15. Baratloo, A., Hosseini, M., Negida, A. & El Ashal, G. Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emerg. Tehran Iran* **3**, 48–49 (2015).
16. Link, R. W. *et al.* Investigating the distribution of HIV-1 Tat lengths present in the Drexel Medicine CARES cohort. *Virus Res.* **272**, 197727 (2019).
17. Mele, A. R., Marino, J., Dampier, W., Wigdahl, B. & Nonnemacher, M. R. HIV-1 Tat Length: Comparative and Functional Considerations. *Front. Microbiol.* **11**, 444 (2020).
18. Kuppuswamy, M., Subramanian, T., Srinivasan, A. & Chinnadurai, G. Multiple functional domains of Tat, the trans-activator of HIV-1, defined by mutational analysis. *Nucleic Acids Res.* **17**, 3551–3561 (1989).
19. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. A systematic survey of an intragenic epistatic landscape. *Mol. Biol. Evol.* **32**, 229–238 (2015).

20. Gonzalez, C. E. & Ostermeier, M. Pervasive Pairwise Intragenic Epistasis among Sequential Mutations in TEM-1 β-Lactamase. *J. Mol. Biol.* **431**, 1981–1992 (2019).
21. Poon, A. & Chao, L. The rate of compensatory mutation in the DNA bacteriophage phiX174. *Genetics* **170**, 989–999 (2005).
22. Sadaie, M. R., Mukhopadhyaya, R., Benaissa, Z. N., Pavlakis, G. N. & Wong-Staal, F. Conservative mutations in the putative metal-binding region of human immunodeficiency virus tat disrupt virus replication. *AIDS Res. Hum. Retroviruses* **6**, 1257–1263 (1990).
23. Chopard, C. *et al.* Cyclophilin A enables specific HIV-1 Tat palmitoylation and accumulation in uninfected cells. *Nat. Commun.* **9**, 2251 (2018).
24. Sargeant, D. *et al.* HIVToolbox, an integrated web application for investigating HIV. *PloS One* **6**, e20122 (2011).
25. Sargeant, D. P. *et al.* The HIVToolbox 2 web system integrates sequence, structure, function and mutation analysis. *PloS One* **9**, e98810 (2014).
26. Kameoka, M. *et al.* HIV-1 Tat Protein Is Poly(ADP-ribosyl)ated in Vitro. *Biochem. Biophys. Res. Commun.* **261**, 90–94 (1999).
27. D'Orso, I. & Frankel, A. D. Tat acetylation modulates assembly of a viral-host RNA-protein transcription complex. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 3101–3106 (2009).
28. Sarmady, M., Dampier, W. & Tozeren, A. HIV protein sequence hotspots for crosstalk with host hub proteins. *PloS One* **6**, e23293 (2011).
29. Schulze-Gahmen, U. & Hurley, J. H. Structural mechanism for HIV-1 TAR loop recognition by Tat and the super elongation complex. *Proc. Natl. Acad. Sci.* **115**, 12973–12978 (2018).
30. Schulze-Gahmen, U. *et al.* Insights into HIV-1 proviral transcription from integrative structure and dynamics of the Tat:AFF4:P-TEFb:TAR complex. *eLife* **5**, pii: e15910 (2016).
31. Tahirov, T. H. *et al.* Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature* **465**, 747–751 (2010).
32. Gu, W.-G. Genome editing-based HIV therapies. *Trends Biotechnol.* **33**, 172–179 (2015).
33. D'Orso, I. *et al.* Transition step during assembly of HIV Tat:P-TEFb transcription complexes and transfer to TAR RNA. *Mol. Cell. Biol.* **32**, 4780–4793 (2012).
34. Bertrand, S. J., Aksenova, M. V., Mactutus, C. F. & Booze, R. M. HIV-1 Tat protein variants: Critical role for the cysteine region in synaptodendritic injury. *Exp. Neurol.* **248**, 228–235 (2013).
35. Frankel, A. D., Biancalana, S. & Hudson, D. Activity of synthetic peptides from the Tat protein of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 7397–7401 (1989).
36. Meredith, L. W., Sivakumaran, H., Major, L., Suhrbier, A. & Harrich, D. Potent Inhibition of HIV-1 Replication by a Tat Mutant. *PLoS ONE* **4**, e7769 (2009).
37. Ruben, S. *et al.* Structural and functional characterization of human immunodeficiency virus tat protein. *J. Virol.* **63**, 1–8 (1989).
38. Truant, R. & Cullen, B. R. The Arginine-Rich Domains Present in Human Immunodeficiency Virus Type 1 Tat and Rev Function as Direct Importin β-Dependent Nuclear Localization Signals. *Mol. Cell. Biol.* **19**, 1210–1217 (1999).
39. Hauber, J., Malim, M. H. & Cullen, B. R. Mutational analysis of the conserved basic domain of human immunodeficiency virus tat protein. *J. Virol.* **63**, 1181–1187 (1989).
40. Kuppuswamy, M., Subramanian, T., Srinivasan, A. & Chinnadurai, G. Multiple functional domains of Tat, the trans-activator of HIV-1, defined by mutational analysis. *Nucleic Acids Res.* **17**, 3551–3561 (1989).

41. Schullek, J. R., Ruf, W. & Edgington, T. S. Key ligand interface residues in tissue factor contribute independently to factor VIIa binding. *J. Biol. Chem.* **269**, 19399–19403 (1994).
42. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* **6**, 116-124.e3 (2018).
43. Genevaux, P., Schwager, F., Georgopoulos, C. & Kelley, W. L. Scanning mutagenesis identifies amino acid residues essential for the in vivo activity of the Escherichia coli DnaJ (Hsp40) J-domain. *Genetics* **162**, 1045–1053 (2002).
44. Schiller, M. R. *et al.* Regulation of RhoGEF activity by intramolecular and intermolecular-SH3 interactions. *J.Biol.Chem* **281**, 17774–17786 (2006).
45. *FASTX-toolkit*. (Cold Spring Harbour Laboritories).
46. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
48. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
49. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinforma. Oxf. Engl.* **31**, 1913–1919 (2015).
50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Prepr. ArXiv13033997* (2013).
51. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).
52. *pyVCF library*.
53. *Accessible Surface Area and Accessibility Tool*.

Figures

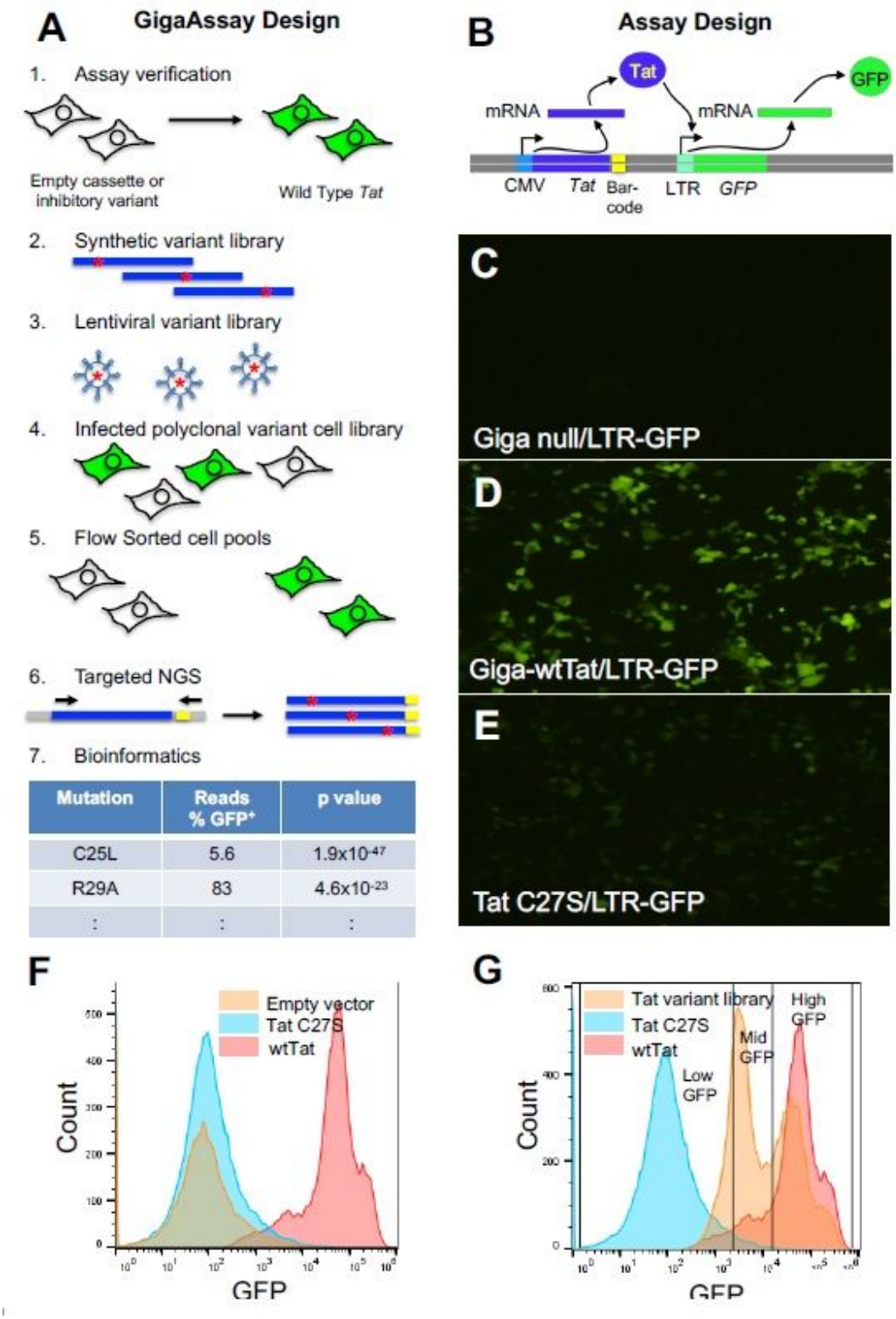

Figure 1

Design and implementation of the GigaAssay for Tat transcriptional activation. A. Design of GigaAssay system. Propagation of the recombined cells under poison selection. Cell sorting based on GFP reporter expression. gDNA is isolated, and a targeted Tat amplicon library is prepared and sequenced by NGS. B.

Schematic representation of Tat dependent LTR transactivation inducing GFP expression. C-F. Epifluorescence microscopic images of LentiX293T/LTR-GFP cells transfected with GigaAssay plasmids: null/LTR-GFP (C. - control); wtTat/LTR-GFP (D, + control); and an inhibitory mutant11, C27S-Tat/LTR-GFP (E, - control). F. Flow cytometry of GigaAssay controls in LentiX293T/LTR-GFP cell to define gates. G. Flow cytometry sorting of GigaAssay LentiX293T/LTR-GFP cell library cells with gates defined by - and + controls.

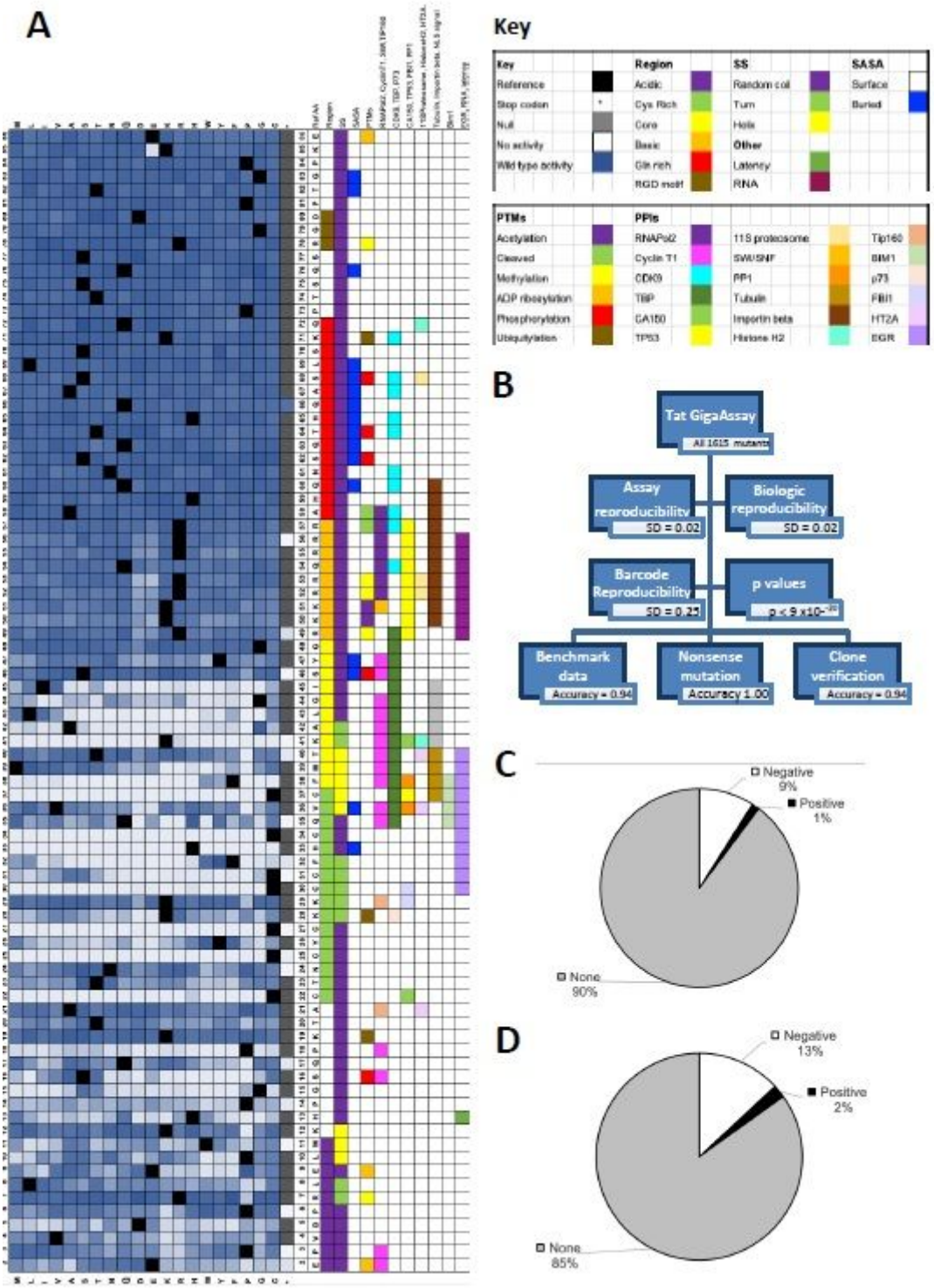

Figure 2

Heatmap showing Tat transactivation activity for a saturating mutagenesis GigaAssay. A. Heatmap for mutated amino acid for each position in Tat. The color gradient represents the level of Tat transactivation activity score measured by GFP+ / (GFP+ + GFP-) reads for each barcode averaged for each mutant. Black boxes are the wild type amino acids and grey boxes are null values. A high ratio (blue) indicates a strong enhancer and a low ratio (red) indicates a strong inhibitor. A color key is shown. B. Assay reproducibility and verification. C, D. Quantitation of types of intrageneic epistasis for double mutants in LentiX293T/LTR-GFP (n =2,465) and Jurkat/LTR-GFP (n=1,633) cells, respectively

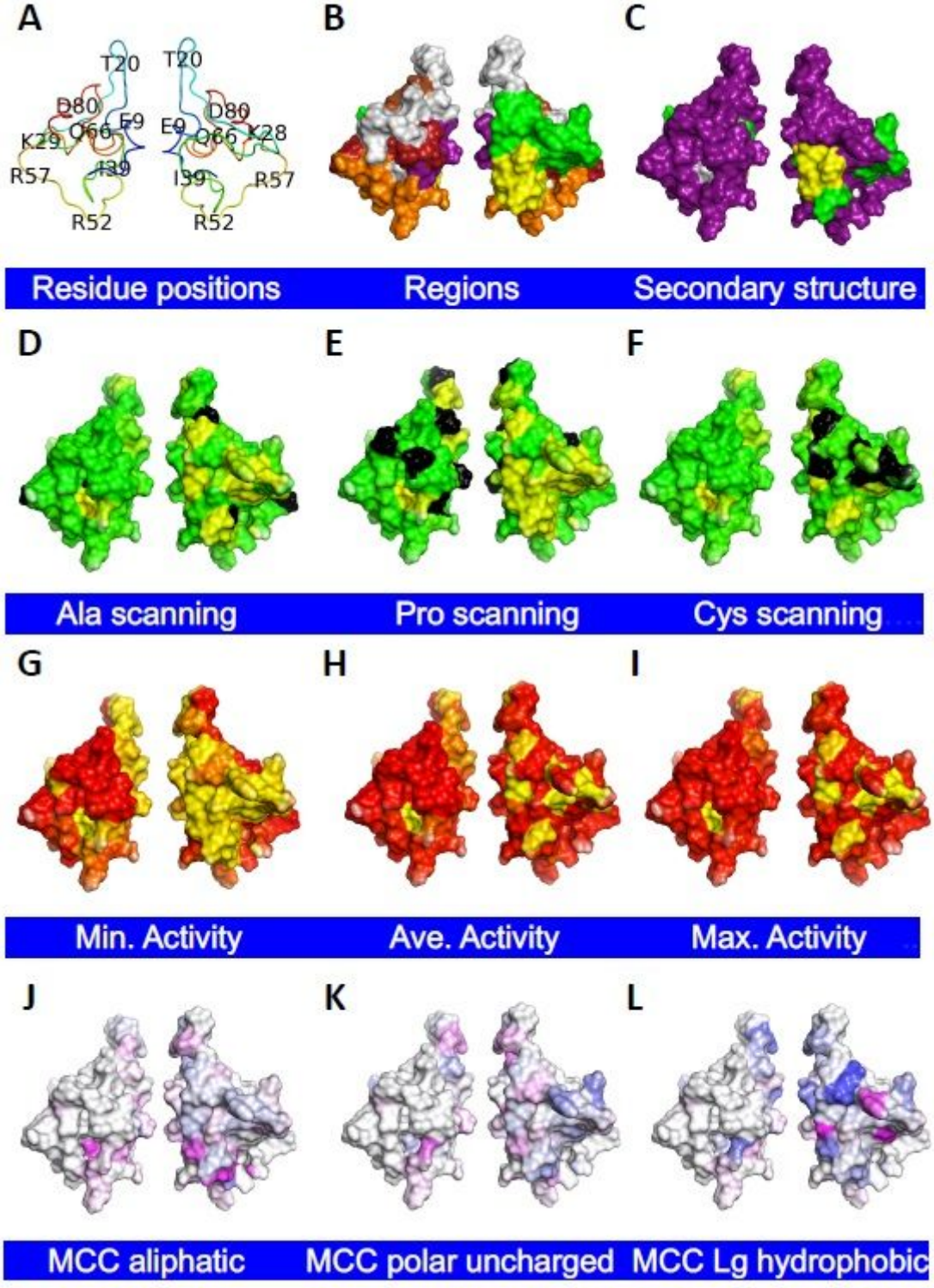

Figure 3

Tat mutant impact on structure/function. All surface maps are on the wtTat 3D structure (PDB: 1TEV) with one member of each pair rotated 180o about the Z axis: A. Amino acid positions on Tat backbone. B. Regions of Tat18. C. Secondary structures. D. Ala scanning substitutions. E. Pro scanning substitutions. F. Cys scanning substitutions. D-F. Residues colored black are for reference amino acids that match the type of scanning. A gradient of yellow with no activity to green with full activity is shown. Gly scanning is shown in Supplementary Fig. S17E. Minimum (G), average (H), and maximum (I) transactivation activity heatmap for all substitutions. A gradient of red with wild type activity to yellow with no activity is shown. Positions that do not tolerate any substitution are shown in Supplementary Fig. S17F, J-L. Physiochemical tolerance surface plots for small aliphatic, polar uncharged, and large hydrophobic amino acids, respectively (see Methods). A gradient of blue to white to magenta ranging from lower to higher MCC scores for each position for the class of amino acids indicated is shown. Residues where all substitutions are inactive (C25, C27, C30, C33, C34, C37, and K41) are colored blue purple. MCCs for polar charged amino acids, and those separated by positively and negatively charged amino acids (Supplementary Fig. S17G-I). The color key is as in the Fig. 2A. For comparison, truncation mutants, PTMs, and PPIs are shown in Supplementary Fig. S18E-I. Abbreviations are: Single letter amino acid code, AA = amino acid, SS = secondary structure, SASA = solvent accessible surface areas, PTM = post translational modification, NLS = nuclear localization signal, PPI = proteinprotein interaction.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFileS1TatMutantActivity.xlsx
- SupplementaryFileS2BenchmarkData.xlsx
- SupplementaryFileS3TatNonsenseMutants.xlsx
- SupplementaryFileS4MutantStatistics.xlsx
- SupplementaryFileS5PrimerSequence.xlsx
- SupplementaryTableS1HighThroughputScreens.pdf
- SupplementaryFiguresFinal.pdf