Published in final edited form as:

Nat Protoc. 2015 August ; 10(8): 1212–1233. doi:10.1038/nprot.2015.074.

Quantitative Assessment of RNA-Protein Interactions with High Throughput Sequencing - RNA Affinity Profiling (HiTS-RAP)

Abdullah Ozer^{1,†}, Jacob M. Tome^{1,†}, Robin C. Friedman², Dan Gheba³, Gary P. Schroth³, and John T. Lis¹

Abdullah Ozer: ao223@cornell.edu; Jacob M. Tome: jmt343@cornell.edu; Robin C. Friedman: robin-carl.friedman@pasteur.fr; Dan Gheba: dgheba@illumina.com; Gary P. Schroth: gschroth@illumina.com; John T. Lis: jtl10@cornell.edu

¹Molecular Biology and Genetics Department, Cornell University, Ithaca, NY 14853, USA. Phone +1 (607) 255-2441, fax +1 (607) 255-6249

²Molecular Microbial Pathogenesis Unit, Institut Pasteur, 75724 Paris Cedex 15, FRANCE. +33 (0) 1-4438-9437

³Illumina Inc., San Diego, CA 92121, USA. +1 (267) 251-4547, +1 (510) 670-9310

Abstract

Because RNA-protein interactions play a central role in a wide-array of biological processes, methods that enable a quantitative assessment of these interactions in a high-throughput manner are in great demand. Recently, we developed the High Throughput Sequencing-RNA Affinity Profiling (HiTS-RAP) assay, which couples sequencing on an Illumina GAIIX with the quantitative assessment of one or several proteins' interactions with millions of different RNAs in a single experiment. We have successfully used HiTS-RAP to analyze interactions of EGFP and NELF-E proteins with their corresponding canonical and mutant RNA aptamers. Here, we provide a detailed protocol for HiTS-RAP, which can be completed in about a month (8 days hands-on time) including the preparation and testing of recombinant proteins and DNA templates, clustering DNA templates on a flowcell, high-throughput sequencing and protein binding with GAIIX, and finally data analysis. We also highlight aspects of HiTS-RAP that can be further improved and

All correspondence should be addressed to ao223@cornell.edu, or jtl10@cornell.edu.

[†]These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

A.O. and J.T.L. conceived of HiTS-RAP. A.O., J.M.T., D.G., G.P.S., and J.T.L. designed the HiTS-RAP protocol. D.G., and G.P.S. supplied sequencing reagents and equipment, and technical information. A.O. and J.M.T. performed the experiments. J.M.T. wrote the .xml recipe. J.M.T. and R.C.F. wrote the analysis pipeline. A.O., J.M.T., and J.T.L. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests. D. Gheba and G.P. Schroth are employees of Illumina, Inc.

SUPPLEMENTARY INFORMATION

Supplementary Tutorial 1. (How to write an xml recipe.docx) Brief description of the xml code used to operate the GAIIX instrument. Supplementary Software 1. (Recipe_GA2_82Cycle_SR_Transcription_7PointCurve.xml) The .xml recipe used for HiTS-RAP analysis of GFP aptamer-EGFP interaction.

Supplementary Software 2. (extract_intensity_cif.pl) A Perl script used to extract protein binding intensity from .cif files.

Supplementary Software 3. (ExtractIntensitiesFromCIF_max.py) A helper Python script used to extract protein binding intensity from .cif files.

Supplementary Software 4. (ExtractIntensitiesFromCIF_T.py) A helper Python script used to extract protein binding intensity from .cif files.

Supplementary Software 5. (HiTS_RAP_Pipeline_Final.py) A Python script used to match sequence and protein binding information, and calculate the binding affinity of each RNA sequence.

points of comparison between HiTS-RAP and two other recently developed methods, RNA-MaP and RBNS. A successful HiTS-RAP experiment provides the sequence and binding curves for approximately 200 million RNAs in a single experiment.

INTRODUCTION

In addition to the primary role of RNAs as messengers that transmit genetic information from DNA into proteins, some are also dedicated to structural, catalytic, and regulatory roles. Non-coding RNAs play fundamental roles in the structure and catalytic activity of ribosomes and RNA processing machinery along with processes such as chromatin structure and modification, transcriptional regulation, and mRNA translation and stability control¹. Though some non-coding RNAs can function on their own (e.g. self-splicing RNA and ribozymes^{2, 3}), most associate with protein(s) and form ribonucleoprotein complexes (e.g. ribosome and spliceosome complexes⁴). Also, protein coding RNAs interact with proteins that modulate the RNA's splicing efficiency, stability, cellular localization, and translation efficiency. These RNA-protein complexes not only act on the interacting RNA, but also act on other biological molecules including proteins (e.g. ribosome), DNA (e.g. RNA-induced transcriptional silencing (RITS) complex), and other RNAs (e.g. spliceosome and RNA-induced silencing complex (RISC))¹. In HeLa and HEK293 human cells, as many as 860 proteins were found to associate with polyA-tailed RNA^{5, 6}. Given that not every RNA or RNA binding protein (RBP) is expressed in any one cell type under a given condition, and that a single RBP can associate with thousands of RNAs, the total number of potential RNA-RBP interactions in a multicellular organism could be in the millions.

High throughput sequencing (HiTS) technologies have led to a paradigm shift in how we sequence genomes, identify genomic mutations, detect pathogens, and analyze gene expression profiles. Likewise, analysis of how nucleic acids interact with other biomolecules such as proteins has changed dramatically. For example, chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-Seq) is heavily used for identification of DNA regions that are bound by a specific transcription factor or other DNA binding proteins⁷. Similarly, RNA immunoprecipitation-sequencing (RIP-Seq), crosslinking immunoprecipitation (CLIP), and Photoactivatable Ribonucleoside-enhanced CLIP (PAR-CLIP) have grown popular in recent years for the discovery of RNAs that associate with an RBP of interest⁸. SELEX methods utilizing genomic or random libraries have also been used to identify high affinity RNA targets of RBPs^{9, 10}. While the aforementioned methods are capable of identifying potential RNA-RBP interactions and the underlying sequence motifs; these methods are not ideal for quantitative characterizations of these interactions in a high throughput manner. In addition, these methods are dependent on antibodies, unable to discriminate between direct vs. indirect interactions, and/or potentially biased (due to differences in cross-linking efficiency and RNA abundance).

Truly high-throughput, quantitative assays for measuring protein-nucleic acid affinities were limited to DNA, until recently. Microarrays have been used successfully to study interactions of a protein with a large number of double-stranded DNAs (dsDNAs)¹¹ and single-stranded DNA (ssDNA) aptamers¹² (44K and 15K, respectively) in a quantitative

manner, but are costly, possibly requiring a custom microarray for each target protein and/or experiment, and limited to DNA, as RNA microarrays are not commercially available. An extremely high throughput, quantitative method called High Throughput Sequencing-Fluorescent Ligand Interaction Profiling (HiTS-FLIP) was recently developed by C. Burge's group for analysis of DNA-protein interactions¹³. Using Illumina Genome Analyzer IIX (GAIIX) instrument, HiTS-FLIP performs high-throughput sequencing of DNA templates and sequential protein binding at multiple concentrations to measure binding affinity of a target protein to all sequences present on the flowcell.

Overview of the Procedure

Inspired by the seminal HiTS-FLIP method, we developed High Throughput Sequencing-RNA Affinity Profilng (HiTS-RAP) to enable an unbiased, high throughput, and quantitative analysis of RNA-RBP interactions¹⁴. To adapt HiTS-FLIP to measure RNA-protein affinity, a way to present RNA transcripts encoded by the DNA templates in each cluster on the flowcell was lacking. To this end, we decided to use T7 RNA polymerase for transcription because it uses a short, specific promoter sequence, and is widely available commercially or easily prepared in-house¹⁵, and the *E. coli* replication terminator protein, Tus, for stably halting the transcription by T7 RNA polymerase^{16, 17}. The combined action of these two proteins generates RNA transcripts that are stably tethered to the DNA template and thus available for interaction with a target protein, at each DNA cluster on the flowcell.

Like HiTS-FLIP¹³, for HiTS-RAP we utilized an unmodified Illumina Genome Analyzer IIX (GAIIX) sequencer equipped with a Paired-End Module (PEM). At its core, the GAIIX is both an automated microfluidic device, which can precisely control amounts, incubation times, and temperature of reagents delivered to the flowcell, and a sensitive Total Internal Reflection Fluorescence microscope which can quantitatively detect multiple fluorescent entities on the surface of the flowcell. The GAIIX is easily programmable (Supplementary Tutorial 1 and Supplementary Software 1). These features have made the GAIIX an attractive instrument to repurpose for HiTS-RAP¹⁴. HiTS-RAP is fully automated; a GAIIX programmed with a single user-modified .xml recipe (Supplementary Software 1) performs every step of HiTS-RAP, and requires minimal user input once the instrument is set up. The overview of HiTS-RAP RNA-RBP binding affinity determination and the instrument used for this assay, the Illumina GAIIX, are described in Figure 1.

Briefly, HiTS-RAP involves the following biochemical steps that are carried out after a normal sequencing run on a GAIIX. Following standard Illumina sequencing, the DNA strand synthesized during sequencing is stripped away and clean dsDNA template is regenerated using Klenow enzyme. Next, Tus protein is bound to the dsDNA template at the Ter consensus element (5'-AATTAGTATGTTGTAAGTAAAGTCACGTCATG-3'). Then transcription is carried out with T7 RNA polymerase, which halts when it encounters the Tus protein bound to the Ter site. Washing away the transcription reaction then leaves a single, stably halted transcript linked to each DNA molecule. Increasing concentrations of the protein of interest, fluorescently labeled as a fusion with mOrange fluorescent protein, are then equilibrated in the flowcell, and imaged by GAIIX. The standard Illumina software is used to obtain the RNA sequence and to collect protein binding data, which are extracted

from the run's output using a set of software included with this manuscript (Supplementary Software 2–4). Protein binding data is then analyzed, also using software provided here (Supplementary Software 5), to solve the K_d of the interaction between the RNAs sequenced and the protein of interest.

Here, we present a detailed protocol for HiTS-RAP¹⁴ that can quantitatively measure the binding affinities of millions of RNAs in a single experiment. With HiTS-RAP, we measured binding affinities of Green Fluorescent Protein (GFP)- and Negative Elongation Factor subunit E (NELF-E)-binding RNA aptamers. HiTS-RAP measured K_d s (4.3 and 5.2 nM, respectively) are in good agreement with Electrophoretic Mobility Shift Assay (EMSA), Fluorescence Polarization (FP), or Isothermal Titration Calorimetry (ITC) measured affinities (5–15 nM for GFP- and 8.5–19 nM for NELF E-binding aptamers, respectively^{18, 19}). Furthermore, we determined the affinities of over a thousand point mutants of these aptamers, providing structural insights into each aptamer's interaction with the corresponding protein target.

Applications of the method

In addition to performing a mutational analysis of a known single RNA aptamer-target protein interaction, HiTS-RAP can be used for analysis for SELEX enriched libraries to identify the highest affinity aptamers among the enriched ones –SELEX enrichments are not necessarily well-correlated with the binding affinity¹². HiTS-RAP could also be used to analyze interactions of modified nucleic acid aptamers (i.e. 2'-fluoropyrimidine containing RNAs), if the mutant polymerases that are capable of incorporating modified nucleotides into RNA (and DNA) can also be halted by the *E. coli* replication terminator protein Tus. Perhaps most biologically significant, HiTS-RAP can be used to obtain binding affinities of millions of RNAs in genomic or transcriptomic libraries (possibly enriched by RIP, CLIP, or PAR-CLIP methods) for a specific target. HiTS-RAP could be performed with a random library for doing k-mer analysis, to determine *de novo* the binding specificity of an RBP.

Advantages and Limitations of HiTS-RAP

HiTS-RAP uses an unmodified Illumina GAIIx sequencer, which is equipped with a Paired-End Module (PEM) and controlled by a Dell Precision T7500 Tower Workstation. HiTS-RAP is fully automated; a GAIIx programmed with a single user-modified .xml recipe performs every step of HiTS-RAP sequentially, and requires minimal user input once the instrument is set up. Instructions for how to modify the .xml recipe is provided in Supplementary Tutorial 1. Users can modify the provided recipe (Supplementary Software 1) to accommodate their specific needs for HiTS-RAP application such as changing the length of sequencing run and number of protein concentrations used for protein binding. Anyone familiar with operating the GAIIx sequencer should be able to carry out the HiTS-RAP protocol as described below. Loading the flowcell onto the instrument is the step which is likely to require the most practice.

Analysis of sequencing and protein binding data can be done with standard Illumina software and the custom software provided (Supplementary Software 2–5). To this end,

familiarity with basic command line operation of a Linux workstation and some knowledge of the Python programming language will be helpful.

HiTS-RAP measures binding affinities of millions of RNAs for a given target protein. HiTS-RAP can easily be adapted to other biomolecules other than proteins, given that they are labeled with fluorescent dyes that are similar to mOrange fluorescent protein (excitation wavelength: 548 nm and emission wavelength: 562 nm²⁰) and thus compatible with Illumina GAIIx optics.

Arguably the biggest limitation of HiTS-RAP is the length of RNA that can be analyzed. Tus protein can efficiently halt a transcribing T7 RNA polymerase regardless of the length of transcription unit up to 1 kilobase (Kb) (Tome, JM., Ozer, A., and Lis, JT. unpublished data). Rather, the limitation is a direct consequence of restrictions imposed by the cluster generation and sequencing on Illumina flowcell. DNA fragments that are longer than 500 base pairs are poor substrates for cBot and GAIIx instruments. Moreover, as presented, HiTS-RAP performs a single read sequencing, which is limited to 150 nucleotides in length on GAIIx. Therefore, in its current form HiTS-RAP is limited to analysis of 150 nucleotides or shorter RNAs. However, by incorporating a paired-end sequencing protocol into HiTS-RAP both ends of a longer genomic fragment could be mapped, increasing the length limit to the 500 nucleotide limit of cluster generation.

Application of HiTS-RAP to proteins that interact with DNA, T7 RNA polymerase, or Tus protein would be complicated, as these interactions will contribute to the observed protein binding. Such interactions will be evident in the recommended preliminary testing of the materials and would not be pursued by HiTS-RAP. However, contributions of these interactions can be measured apart from RNA interaction if two consecutive protein bindings are carried out; (i) under single-round transcription (i.e. CTP depleted transcription of GFP aptamer template) and (ii) after full-length transcription. Proteins exhibiting these complicating interactions might still be useable if excess carrier is added (e.g. excess salmon sperm DNA, T7 RNA polymerase, or Tus).

As presented, HiTS-RAP does not measure binding kinetics (on- and off-rates), but measures binding affinity (K_d). However, as discussed below, HiTS-RAP can measure off-rates (on-rates can be calculated from measured K_d s and off-rates) with minor modifications to the provided HiTS-RAP recipe.

Despite the good correlation that we observe between binding affinities measured by HiTS-RAP and other methods, HiTS-RAP measured affinities may not capture the true in solution binding affinity. HiTS-RAP is susceptible to steric hindrance effects especially for large proteins due to surface immobilization of RNAs. Most likely as a direct result of these, variations in protein binding at different clusters with the same RNA sequence are observed. Averaging protein-binding intensities at each protein concentration for all clusters with identical sequence and fitting a single K_d for a sequence overcomes this issue. In our experience, sequences that have 10 or more replicates give a reasonably good estimate of the binding affinity, limiting the throughput of HiTS-RAP (20 million sequences per lane of GAIIx flowcell can give a maximum of 2 million K_d s).

Alternatives to HiTS-RAP

Recently, two other methods have been developed for large-scale quantitative analysis of RNA-RBP interactions^{21, 22}. RNA Bind-n-Seq (RBNS) relies on enrichment of RNAs following a single binding reaction performed at different RBP concentrations and analysis of enriched RNAs by high throughput sequencing –much like a single-round SELEX experiment. K-mer analysis of enriched sequences can reveal sequence specificity of RBP binding (RNA motifs). A near perfect correlation between RBNS and SPR measured affinities has been reported ($r = 0.933$). However, in light of recent observations from the SELEX field¹², it is unclear how often and how well the binding affinities would correlate with SELEX-based enrichments. Results reported by Cho *et al.* suggest that the average affinity of aptamer pools increase with additional rounds of SELEX; however, there is no correlation between abundance of a given aptamer sequence in a given SELEX round and its target binding affinity. In other words, an aptamer with higher multiplicity does not necessarily have higher affinity than a lower abundance aptamer and vice versa. In addition, K-mer analysis may fall short in deciphering more complex RNA-RBP interactions that involve longer fragments of RNA or depend on RNA structure more than the sequence.

Quantitative analysis of RNA on a massively parallel array (RNA-MaP) is conceptually identical to HiTS-RAP, where sequencing, transcription of RNA encoding templates, and binding of fluorescently labeled protein to RNA transcripts retained on the DNA template are done on an Illumina GAIIX instrument²¹. Protein binding measured at multiple concentrations is then used to calculate binding affinities, similar to HiTS-RAP. However, RNA-MaP and HiTS-RAP differ in various aspects (Table 1). Some of these differences are major distinguishing features of each method, and others are minor technical differences that can be interchangeably used in both methods.

In our experience, T7 RNA polymerase cannot be efficiently halted by Streptavidin bound to the DNA template strand, as also observed by others²³. We have not yet tested how effective Streptavidin is in halting E.coli RNA polymerase even under single-round transcription conditions when bound to the biotinylated template DNA.

In general, RNA-MaP uses a host of elegant modifications to the Illumina GAIIX to produce a smaller amount of very high-resolution data, while HiTS-RAP utilizes an unmodified instrument to make a larger number of measurements, albeit with fewer protein concentrations and noisier intensity measurements. The modifications to the regular sequencing protocol that RNA-MaP employs, including a separate manual run for protein binding, custom analysis of saved images for protein binding measurements, and necessary modifications to the Illumina GAIIX instrument will be daunting for most researchers and limit its wide-use. However, a few concepts implemented in RNA-MaP, which can be adapted to HiTS-RAP, are desirable, including (i) random barcoding of templates, (ii) direct RNA transcript quantification via hybridization of a fluorescent probe (HiTS-RAP uses DNA sequencing signal as a proxy for RNA transcript amount), (iii) dye labeling of target protein, and (iv) measurement of off-rates (k_{off}). The SNAP-tag is smaller (20 kDa vs. 30 kDa mOrange used in HiTS-RAP), and in general small-molecule dyes can be brighter and more photostable than fluorescent proteins. Multiple dyes that are compatible with GAIIX

instrument settings can be purchased or synthesized permitting multiplex RNA-binding assays to be performed without any modification to the instrument.

Determining binding kinetics (on- and off-rates) in addition to affinities as done with RNA-MaP is particularly useful for certain RNA-RBP interactions. This was achieved by measuring the off-rate (k_{off}) and affinity (K_d), and solving the equation $K_d = k_{\text{off}}/k_{\text{on}}$ for on-rate (k_{on}). However, we suspect this might only be possible, if either the off-rate is relatively slow (as in MS2 RNA-MS2 coat protein interaction $k_{\text{off}} \sim 0.1 \text{ min}^{-1}$ ²⁴) or a small size RNA library is being analyzed (thus many copies of a single RNA are present in each tile of a single lane). On a GAIIX, imaging of a single lane (containing 120 tiles) in one channel takes about 20 minutes, whereas RNA-MaP uses a two channel scan (one for RNA quantification and one for protein binding), thus taking ~40 minutes between consecutive scans (estimated based on the standard Illumina sequencing recipe and an unmodified GAIIX instrument). Regardless of the circumstances, off-rate measurements require separate analysis of saved images to correlate time and protein dissociation, as otherwise the time information gets lost in standard analysis with Illumina's software. We expect that the larger-scale analysis of RNA-RBP interactions, the relative ease of HiTS-RAP run and analysis, and the lack of required modification to the GAIIX instrument are likely to make HiTS-RAP more adaptable to a broader segment of the research community.

Experimental Design

Fusion proteins for HiTS-RAP

The quality of recombinant proteins, especially the protein of interest-mOrange fusion and the Tus protein, have a big impact on the quality of HiTS-RAP data. As outlined in PROCEDURE section, we performed single step affinity purification of mOrange-EGFP and GST-Tus fusion proteins using Ni-NTA- and Glutathione-agarose resins, respectively. Typically, we get >90% full-length protein at >10 μM concentration in these preps. However, some proteins may require additional steps of chromatographic purification to obtain sufficiently pure and concentrated proteins suitable for HiTS-RAP.

Accurate quantification of fluorescently labeled target protein concentration is important for HiTS-RAP, as these values are directly used for the calculation of the binding affinity (K_d). Controlling the amount and activity other proteins used (i.e. T7 RNAP, GST-Tus, Klenow exo-) is also important for efficient producing halted RNA. We typically use Bradford Assay (BioRad) with BSA standards for protein quantification. When possible, prior to HiTS-RAP, we like to test the activity of the recombinant proteins. For example, we perform gel shifts with known protein-nucleic acid interactions to verify that mOrange fusions are active in binding RNA targets, and that GST-Tus binds the Ter sequence element. We also carry out test reactions for enzymes, such as T7 RNAP and Taq.

mOrange fluorescent protein fusions are used for detection of target protein in HiTS-RAP, since its excitation and emission properties are compatible with an unmodified GAIIX. Slow maturation, large size, and weaker fluorescent properties of mOrange are not ideal. Replacement of original filter sets installed on GAIIX to enable use of other fluorescent proteins may not be ideal for general users because it is technically challenging, the

instrument could potentially be damaged in the process, and necessitates separate sequencing and protein binding runs on different instruments. Nevertheless, modification of the GAIIX instrument, as has been done in RNA-MaP¹³ and elsewhere²⁵, is possible and may be beneficial for certain applications. In general, compared to fluorescent proteins, fluorescent dyes have more desirable properties; smaller size, brighter, more photostable, provide a larger variety of choices that are compatible with the GAIIX instrument (i.e. Alexa 532, Atto 532, and Cy3 have mOrange-like excitation and emission spectra), and are likely to improve the detection of protein binding in HiTS-RAP. Target proteins could be labeled with small molecule fluorescent dyes by various labeling strategies including CLIP-, SNAP-, aldehyde-tags, and amine- or thiol-coupling^{26, 27}. Labeling proteins with spectrally distinguishable fluorescent proteins, dyes, or a combination may be used for a multiplex HiTS-RAP assay where binding of multiple proteins can be analyzed simultaneously in one HiTS-RAP run.

RNA encoding DNA Library

As for proteins (see above), the quality of the DNA templates used for HiTS-RAP is very important. Since multiple PCR steps are used for creation of the DNA template (Fig. 2), we optimized the number of PCR cycles for each step (optimized conditions are reported in the PROCEDURE section below) and monitored the quality of the PCR product by polyacrylamide gel electrophoresis at each step. PCR products are purified by PCR purification kit or gel extraction (especially for the final template) to ensure the best quality and sufficient removal of primer dimers and other PCR byproducts. Shorter fragments can easily dominate the flowcell due to better amplification efficiency during cluster generation, and as a result diminish the number of full-length templates that can be analyzed. We use home-made Taq or Pfu-Sso7d (also known as Phusion) polymerases^{28, 29} for PCR amplification. Depending on the purpose (i.e. mutational analysis of a single RNA-single protein or a library of RNAs with a single protein) the choice of DNA polymerase and PCR conditions (e.g. buffer composition, number of PCR cycles) in each PCR step can be modified to increase or to minimize the rate of PCR introduced mutations using Taq polymerase under error-prone PCR conditions³⁰ or high fidelity Phusion polymerase, respectively. Quantification of DNA template concentration, for which we use Qubit dsDNA HS Assay, is also critical to get the optimum number of clusters on the flowcell. For HiTS-RAP, we typically use DNA templates that have no detectable contaminating bands on a native polyacrylamide gel stained with ethidium bromide and have a concentration > 50 ng/μl.

The DNA templates that we have used in HiTS-RAP so far are designed for single-read sequencing of a single template and mutants derived from it (Fig. 2); however, conceptually it is possible to do paired-end sequencing by incorporating TruSeq or Paired-end sequencing oligos into the template design. This would allow analysis of protein interactions of longer RNAs that are derived from the genome or transcriptome. In such a template, the RNA encoding DNA would be flanked by the two Illumina paired-end sequencing oligos on either end, then by a T7 RNA polymerase promoter or Tus binding Ter site, and finally the Illumina flowcell adaptors. The portion of the Illumina sequencing oligo that becomes part of the transcript can be used to quantitate the amount of RNA transcript in each cluster using

a fluorescently labeled complementary oligo as has been done in RNA-Map²¹. A barcode sequence can be introduced between the T7 RNA polymerase promoter and Illumina sequencing oligo and can be read by an index read as part of multiplex sequencing protocol, and allow simultaneous processing and analysis of multiple RNA libraries in a single-lane of GAIIX flowcell, thus minimizing experimental differences between the two libraries. This might be particularly useful for analyzing interactions of a protein with RNA libraries prepared from different sources or under different conditions, and SELEX enriched aptamer libraries from various rounds or performed under different conditions.

If possible, we recommend a preliminary testing of the materials prepared for a HiTS-RAP run in vitro before an actual run. This includes testing transcription, with and without halting, of the template DNA, and analysis by denaturing PAGE³¹ and EMSA³². We also recommend testing of a known interaction of the labeled RBP by EMSA. The cost and labor associated with the HiTS-RAP assay is large, so it is worth spending some time on these tests since they may avoid a failed experiment. Assessing a rough estimate of K_d is also important for deciding what protein concentrations to use for HiTS-RAP protein binding steps.

Modifying the .xml recipe for HiTS-RAP run

The .xml recipe used to operate the Illumina GAIIX instrument, provided as Supplementary Software 1, can be modified using a text editor like Notepad++. A brief description of the .xml code is provided in the Supplementary Tutorial 1. Depending on the user specific application, two main aspects of the HiTS-RAP recipe may need to be modified. First is the length of the sequencing run, which can be easily shortened or lengthened by removing or adding sequencing cycles (as shown below) in the protocol section of the recipe, respectively. The provided recipe contains 82 sequencing cycles:

```
<Incorporation ChemistryName="CompleteCycle" ExposureA="500"
ExposureC="350" ExposureG="200" ExposureT="175" />
```

The second is the number of different protein concentrations used for protein binding. If desired, a larger number of different protein concentrations can be used. This will require a new Chemistry definition similar to that of "TargetBinding16" with the PEM port number adjusted for each additional protein concentration in the ChemistryDefinitions section, and calls for the priming of new port(s) and actual protein binding(s) dictated by ChemistryRef and Incorporation commands, respectively, in the protocol section. Alternatively, new UserWait steps could be added to the recipe, and the five refrigerated ports on PEM are reused for more protein binding steps. If DNA binding of the target protein is suspected, this can be tested by performing the protein binding steps once more just before the transcription. This can be achieved by copy-pasting the cycles 83–90 of the provided recipe in front of the transcription step in the protocol section. It is important that any modified recipe to be placed into the custom recipes folder and tested for any errors by opening with the SCS2.9 software as described in Supplementary Tutorial 1. Once opened, it is important

to read through the steps listed in the SCS software to ensure that the recipe will be implemented by the instrument as intended.

Cluster generation and sequencing

Immobilization of DNA templates on a flowcell and amplification to generate clusters using a cBot instrument is a prerequisite step for the application of HiTS-RAP (Fig. 3). The older Illumina Cluster Station system could also be used to cluster the flowcell; it should be equally effective but requires much more hands-on time and expertise, and these instruments are not as widely available as the cBot instruments. HiTS-RAP starts with sequencing of DNA cluster on the flowcell with GAIIx. A maximum of about 80 million clusters can be sequenced per lane of Illumina GAIIx flowcell; however, for HiTS-RAP assays, we aim for ~20 million clusters per lane to get good separation between neighboring clusters. This generally gives better quality sequencing and protein binding data.

An Illumina GAIIx flowcell has 8 lanes, each of which can be used for a different DNA template. We recommend that one of these lanes be used for PhiX control DNA to ensure high quality sequencing in all lanes. This should be specified as the control lane when starting the Illumina SCS before a run. Inclusion of a control lane is especially important; if the first few bases read in a library are not very diverse, the analysis software is unable to generate a suitable basecalling matrix. If possible and available, we recommend including in each HiTS-RAP experiment a template that has no affinity for the protein of interest (negative control), and one representing a well-established target protein binding RNA (positive control). These can be included in a single control lane or added to other libraries and distinguished based on the sequence of the template or the barcode. Although all 8 lanes of the flowcell can be addressed individually on the cBot instrument during cluster generation, the Illumina GAIIx instrument processes all of them simultaneously with the same solutions at any given time. Recently, modifications to the GAIIx instrument that enable individual processing of 8 lanes have been published²⁵. This could be implemented to perform up to 8 HiTS-RAP runs with a single flowcell.

We perform sequencing and protein binding as part of a single run, unlike RNA-MaP. A single .xml recipe (Supplementary Software 1) is used to operate the sequencer through the entire protocol. Illumina's standard software is used for imaging and image analysis for protein binding. Therefore, anyone who knows how to operate a GAIIx can perform HiTS-RAP.

dsDNA regeneration and transcription halting

Regeneration of a clean full-length dsDNA template following sequencing is necessary for three main reasons; (i) to generate a double stranded Ter site for Tus binding, (ii) to generate a double stranded T7 promoter for transcription by T7 RNA polymerase, and (iii) to generate a template strand for transcription that is full-length and free of any artifacts introduced during sequencing (Fig. 3).

A Tus-bound Ter site has been shown to halt progression of many DNA and RNA polymerases in one orientation but not in the other^{16, 33}. Also, it is well-documented in the literature that an RNA polymerase that has synthesized at least a 9 nt long transcript is

extremely stable when stalled by a physical barrier or the absence of a ribonucleotide^{34, 35}. HiTS-RAP combines these two phenomena to tether RNA transcripts to their DNA templates in each cluster, enabling observation of their interaction with the target protein. Tus's ability to impede progression of T7 RNA polymerase is likely due to its contra-helicase activity³⁶, rather than it acting as a simple roadblock. However, E.coli RNA polymerase can be halted by streptavidin bound to a biotin moiety on template strand of DNA, under single-round transcription conditions, as demonstrated with RNA-MaP²¹.

Protein binding

To get an accurate measure of binding affinity (K_d), concentrations of the fluorescently-labeled protein (used in protein binding steps of HiTS-RAP) need to be centered around the expected K_d . At minimum, 3 concentrations below and 3 concentrations above the K_d (total of 7 points), each separated in 5-fold increments, should be used. For proteins that are known or suspected to bind DNA, contribution of protein binding at each DNA cluster needs to be measured by performing protein binding at the same concentrations without transcription halting.

Our estimates of the halted transcription complex lifetime is ~72 hours, which is sufficient to perform 48 protein binding steps based on a cycle time of 1.5 hours¹⁴. Depending on the experiment, the number of binding steps that could be carried out within this time could change. For example, if a slow on rate were expected, more time would need to be allowed for equilibration before imaging, allowing for fewer protein binding cycles. Similarly, if only a subset of lanes were utilized, imaging times would be reduced, allowing for more potential binding steps. Based on this slow dissociation of the complex, a correction is applied to account for the loss in protein binding signal, and the corrected values are used for calculation of the binding affinities. We have only carried out runs with at most seven binding steps, meaning that most of the original RNA should still be present at each time point we image protein binding.

Data Analysis

After a HiTS-RAP run, the run folder contains a number of files, which were created by Illumina's Real-Time Analysis (RTA) software, that are then analyzed to obtain the binding affinity of each RNA sequenced on the flowcell (Fig. 4). Most notably for the purpose of HiTS-RAP, the RTA analyzes the raw .tif images collected during the run to measure the fluorescence intensity of each cluster. These measurements are stored in binary .cif files. RTA uses this intensity information from all four channels to determine the base added at each cluster, at each cycle, by comparing to a basecalling matrix generated in the first cycles. This dictates which intensity measurements across the four channels correspond to which base. While the intensity measurements recorded within .cif files are generated for the purpose of comparing relative signals to determine the most suitable basecall, they are also sufficient for comparing intensities within a single channel across cycles to make affinity measurements for HiTS-RAP. In this way, the RTA is carrying out the laborious task of identifying clusters and measuring fluorescence intensities from raw images that is necessary for making binding measurements. A set of Perl and Python scripts originally developed for HiTS-FLIP¹³ (Supplementary Software 2, 3, and 4) are used to extract binary

intensity data from .cif formatted files into a plain text format. These text files contain the intensity measurements for every cluster in a lane, for the specified channel and cycles of the run. Each cluster is identified by its tile number, and x and y coordinates within that tile. Intensities measured in the T channel were used for protein binding, as this channel most closely fits mOrange fluorescence with the least amount of noise.

Basecalling is carried out by the Illumina RTA during the run. Thus, after a run is finished, the run folder contains this information stored in binary .bcl files. Each individual .bcl file contains basecalls for clusters within a single tile, at a single cycle. To get the full sequence of each read, the information in these files must be assembled through all cycles. This is done by the .bcl to .qseq function of Illumina's OffLineBaseCaller (OLB) software, which uses .bcl, .stats, .filter, .control, .pos.txt, and the config.xml file created during the run to make .qseq.txt files. Each .qseq.txt file contains all of the sequence information, divided by tile. As in the intensity file created by reading .cif files, clusters are identified by lane, tile number, and x and y coordinates within the tile. In addition to sequence, the entry for each cluster contains a quality string, and a binary pass filter metric, which are important for determining which clusters to include in downstream analyses.

The remainder of the data analysis is done with a single Python script provided as Supplementary Software 5. The first step is to match cluster sequences to their corresponding intensities. We do this using the coordinates of clusters contained in both files. Because this information is recorded in .bcl and .cif files differently, the x and y coordinate numbers in intensity files must be converted to match the sequence file coordinates by multiplying by 10, adding 1,000, and rounding to an integer. During this matching stage, we only include clusters for which we have both sequence and intensities, and which pass Illumina's quality filter as recorded in the last field of every entry in .qseq.txt files. At this stage, the quality string can also be used to implement a more stringent quality filter. For example, when we were performing HiTS-RAP against a library of mutagenized versions of a single sequence, we controlled for sequencing errors resulting in unmutated clusters being called as mutants by requiring that any bases different than the original sequence have a Phred quality score of at least 25. Next, we prepare for determining K_d s by averaging all clusters corresponding to each sequence to generate a single binding curve that is representative of that sequence. We then do a nonlinear, least squares fit to the Hill equation to determine the K_d . We have found that this works best when the four parameters for K_d , hill coefficient, base intensity, and maximum intensity are allowed to vary. We use the covariance matrix returned by SciPy's fitting algorithm to identify good fits; a covariance below ~1,000,000 for K_d generally indicates a good measurement. It is also helpful to look at the other parameters as well when assessing an individual K_d measurement. Our hill coefficients are generally less than one, indicating negative cooperativity in binding, which we interpret as steric hindrance as saturation of the cluster is reached. In addition, when the fitted base intensity is close to the first few measured intensities, and the fitted maximum is close to the measured values of the last intensities, this is indicative of a good fit. Fitted K_d s that are either above or below the range of concentrations assayed generally fail the covariance filter. However, if they do not, they are

not reliable. We expect the dynamic range of the assay to be only within the concentrations of target protein imaged to generate binding curves.

Alternatively, we have considered fitting binding curves from individual clusters and averaging the results to give a representative K_d , rather than averaging intensities and doing a single fit per sequence. These two approaches give similar results, so we recommend the simpler procedure of averaging intensities.

Follow-up verification of HiTS-RAP measured binding affinities

In our hands, HiTS-RAP measured binding affinities correlate well with EMSA measured binding affinities ($r^2 = 0.64$)¹⁴. However, verification of HiTS-RAP measured binding affinities (for a few RNAs) with a complementary method such as EMSA³², SPR³⁷, or nitrocellulose filter binding^{38, 39} may be desirable. These low throughput assays can accommodate many more protein concentrations in binding reactions than HiTS-RAP, and thus may yield a more precise estimate of the K_d . In addition, methods like SPR³⁷ and ITC⁴⁰ can also reveal on- and off-rate of binding which might be of interest for certain applications than K_d .

Potential cost-saving measures

Currently, Illumina reagents including the flowcell, cluster generation reagents, and sequencing reagents, cost around \$7,000 to carry out a HiTS-RAP run with 100 nucleotide sequencing using a full flowcell. However, these costs could be spread over several experiments with some modifications to the method. Gravina et.al. have published a protocol for utilizing only a subset of lanes on the GAIIX by disconnecting the syringe pump from the other lanes²⁵. With this modification experiments could be carried out one lane at a time, in cases where only the amount of data generated from a subset of lanes is sufficient for the desired analyses. In this case, it would not be possible to use a separate PhiX control lane, so the PhiX genome would need to be spiked into each lane if low diversity is likely to affect basecalling. DNA clusters are very stable (Illumina states that a flowcell can be sequenced a month after clustering without degrading), so in cases where the same RNA library is being probed for interactions with multiple proteins, the same flowcell could potentially be reused. In order to use the Illumina image analysis software, as we have presented here, the software must first be allowed to identify clusters during sequencing to be able to measure fluorescence intensities during binding. Thus, sequencing must always be carried out before protein binding and imaging. If images are being analyzed directly, it should be possible to sequence the flowcell once, and then reuse it for protein binding multiple times, using the arrangement of clusters to correlate sequences with binding curves. This would avoid the extra denaturing steps required for resequencing. HiTS-RAP adapted to a newer HiSeq instrument would cost roughly the same (~\$7,000), but would yield 5–10 times more sequence reads and binding affinities (K_d s). With the current pricing of the reagents, the cost of measuring binding affinity of an individual RNA is about 3.5 cents (\$7,000 / 2,000,000 RNAs), whereas a similar measurement can only be made with an alternative method such as EMSA for no less than \$10, making HiTS-RAP ~300 times cheaper.

MATERIALS

Reagents

DNA template encoding the RNA(s) of interest

mOrange pHis-// plasmid (Addgene, cat. no. 53340)

mOrange pHis-// plasmid containing open-reading frame (ORF) of protein of interest

Tus pGST-// plasmid (Addgene, cat. no. 53305)

BL21-CodonPlus(DE3)-RIPL Competent cells (Agilent Technologies, cat. no. 230280)

dNTPs (Life Technologies, cat. no. 10297-117)

rNTPs (Sigma-Aldrich, cat. nos. A7699, G8877, U6625, and C1506)

ColorpHast pH Strips (EMD Millipore, cat. no. M95883)

Diethyl pyrocarbonate (DEPC) (Sigma-Aldrich, cat. no. 40718)

CAUTION: DEPC is a suspected carcinogen. Avoid skin contact and inhalation. DEPC should be handled at all times with gloves and in a fume hood.

Phenol:Chloroform:Isoamyl Alcohol mix, pH 8.0 (EMD Millipore, cat. no. 6810)

CAUTION: Phenol and chloroform are toxic. Avoid skin contact and inhalation. Phenol and chloroform should be handled at all times with gloves and in a fume hood.

Chloroform (EMD Millipore, cat. no. 3150)

CAUTION: Chloroform is toxic. Avoid skin contact and inhalation. Chloroform should be handled at all times with gloves and in a fume hood.

Ethanol (EMD Millipore, cat. no. 100986)

GlycoBlue Coprecipitant (Life Technologies, cat. no. AM9516)

Yeast Inorganic Pyrophosphatase (YIPP) (NEB, cat. no. M2403)

SUPERase• In RNase Inhibitor (Ambion, cat. no. AM2694)

10X NEBuffer 2 (NEB, cat. no. B7002S)

Klenow Fragment (3'→5' exo-) (NEB, cat. no. M0212)

Oligos, 25–100 nmole synthesis scale, standard desalting (IDT)

MassRuler DNA Ladder Mix (Thermo Scientific, cat. no. SM0403)

30% ProtoGel, 37.5:1 Acrylamide:Bisacrylamide Stabilized Solution (National Diagnostics, cat. no. EC-890)

CAUTION: Acrylamide is a potent neurotoxin and can be absorbed through the skin.

Wear gloves and protective clothing when handling solutions containing acrylamide.

E.Z.N.A. Cycle Pure Kit (Omega Biotek, cat. no. D6492)

TruSeq PE Cluster Kit v2-cBot-GA (Illumina, cat. no. PE-300-2001) – includes a flowcell

TruSeq SBS Kit v5-GA (Illumina, cat. no. FC-104-5001)

Genome Analyzer cBot Manifold (Illumina, cat. no. FC-901-1003)

PhiX Control Kit v3 (Illumina, cat. no. FC-110-3001)

All other chemicals common in molecular biology laboratories are purchased from Sigma-Aldrich at ACS reagent grade or equivalents from other reputable companies.

*Taq DNA Polymerase (NEB, cat. no. M0273) [or see Ref²⁸]

*Phusion High-Fidelity DNA Polymerase (NEB, cat. no. M0530) [or see Ref²⁹]

*T7 RNA Polymerase (NEB, cat. no. M0251) [or see Ref¹⁵]

CRITICAL: Although not verified, we suspect enzymes obtained from commercial sources (indicated by *) are likely to work in HiTS-RAP. Homemade Taq and Phusion (Pfu-Sso7d fusion) DNA, and T7 RNA polymerases were used in HiTS-RAP and references for their preparation are given^{15, 28, 29}.

Equipment

PCR Thermal Cycler (MJ Research, PTC-200)

PowerPac Basic Power Supply (BioRad, cat. no. 164-5050)

Mini-PROTEAN Tetra Cell (BioRad, cat.no. 165-8000)

Wide Mini-Sub Cell GT Cell (BioRad, cat. no. 170-4468)

Gel Doc EZ System (BioRad, cat. no. 170-8270)

CAUTION: The UV irradiation, used in Gel Doc EZ system, is a potential mutagen.

Avoid exposure to eyes and bare skin. Wear suitable protective shields while working with UV.

Lens Cleaning Tissues (VWR, cat. no. 52846-001)

Immersion oil, 1.473 refractive index (Cargille, cat. no. 19570)

CRITICAL: Immersion oil with 1.473 refractive index is necessary for proper imaging in Illumina GAIx system. Cargille is the Illumina recommended supplier.

Qubit 2.0 Quantitation Lab Starter Kit (Life Technologies, cat. no. Q32872)

Non-stick, RNase-free Microcentrifuge Tubes (Life Technologies, cat. nos. AM12350, AM12450, and AM12475)

175 ml Falcon conical bottles (BD, cat. no. 352076)

125 ml Nalgene bottles (Thermo Fisher Scientific, cat. no. 342040-0125)

150 ml Storage bottles (Corning, cat. no. 431175)

50 ml Conical tubes (BD, cat. no. 352070)

15 ml Sarstedt conical tubes (VWR, cat. no. 101093-696)

CRITICAL: The bottles and tubes listed above, from the specified companies, are compatible with the GAIIX system and recommended by Illumina. The threading on similar products from other sources may not fit the threads of the holders installed on Illumina GAIIX instrument.

Costar Spin-X Centrifuge Tube Filters (Corning, cat. no. 8162)

Tabletop centrifuge 5430 and 5430R (Eppendorf, cat. nos. 022620511 and 022620623)

cBot (Illumina, cat. no. SY-301-2002)

CRITICAL: For cluster generation on flowcell follow Illumina cBot™ User Guide (Part # 15006165 Rev. J July 2012)

Genome Analyzer IIX with Paired-End Module (Illumina)

CRITICAL: For setting up the GAIIX instrument follow Illumina Genome Analyzer IIX User Guide (Using SCS v2.9) (Part # 15018814 Rev. B November 2011). The provided .xml recipe for HiTS-RAP (Supplementary Software 1), or a modified version, need to be placed in the C:\Illumina\SCS2.9\DataCollection\bin\Recipes\Custom folder.

CRITICAL: Regular maintenance washes are important for ensuring proper reagent delivery on the GAIIX. A maintenance wash flushes water and 1 N NaOH through each port on the instrument (see GAIIX User Guide). We typically do a maintenance wash before and after each run. All reagent ports which are used during runs must be washed. Modification of an existing wash .xml recipe (as briefly described Supplementary Tutorial 1) to include additional ports is necessary; as Illumina's standard wash recipes do not include some of the ports that are used. The wash volumes need to be measured carefully. If the volume is less than expected, measure output from each lane individually. Manual pumping of 1N NaOH through the system at a slow rate (i.e. 10 ml at a flow rate of 25 µl/min) followed by thorough flush of the system with warm water generally restores proper reagent delivery.

Dell Precision T7500 Tower Workstation (Dell): necessary for operating Illumina GAIIX and saving raw images from GAIIX

CRITICAL: SCS 2.9 (Sequencing Control Software) needs to be installed on Dell T7500 workstation to operate Illumina GAIIX. SCS 2.9 can be freely downloaded from "<http://support.illumina.com/sequencing/downloads.html>".

CRITICAL: The ImageCyclePump config file located under C:\Illumina\SCS2.9\DataCollection\bin\Config needs to be set to ImageCyclePump On=true AutoDispense = false before starting a HiTS-RAP run. However, the same file needs to be modified to ImageCyclePump On=false AutoDispense = false after the sequencing just before the protein binding steps of the HiTS-RAP run.

Linux Workstation. We recommend an Ubuntu system with at least 64 GB of memory for analysis. CentOS or RedHat Linux are necessary for installing the OLB and extracting sequences from the raw basecalling data.

CRITICAL: OLB 1.9.4 (OffLineBaseCaller), used to extract sequence information from .bcl files, needs to be installed on a Linux computer. OLB 1.9.4 can be freely downloaded from "<http://support.illumina.com/sequencing/downloads.html>". The custom scripts used for data analysis, provided as Supplementary Software 2–5, can be installed and executed on the Linux Workstation.

Reagent Setup

- Oligos: Resuspend oligos with MilliQ H₂O at 100 µM concentration (e.g. 230 µl for 23.0 nmole oligo. Vortex for 30 seconds, centrifuge at 13,000g for 30 seconds at room-temperature (22 °C), and store up to a year at –20 °C.
- DEPC-H₂O: Transfer 1L of MilliQ H₂O into a clean 1L glass bottle with a clean Teflon-coated magnetic stir bar. Add 1 ml of DEPC and stir the solution at room temperature 300 rpm for ~15 hours. Autoclave in liquid cycle at 121°C for 1 hour. Once cooled down, store up to a year at room-temperature.
- 10X T7 transcription buffer: Combine 3ml of 1M HEPES-KOH pH 7.8, 2 ml of 4M Potassium Glutamate, 1.5 ml of 1M MgAc, 50µl of 0.5M EDTA, 0.5ml of 1M DTT, 250µl of 20% Tween 20 (v/v), 200µl of 1M Spermidine, and 2.5ml DEPC-treated MilliQ H₂O. Mix well by inverting and lightly vortexing the solution. Make aliquots of 0.5ml and store it at –20 °C for up to a year.
- 1X T7 transcription buffer: Dilute 0.5ml of 10X T7 transcription buffer with 4.5 ml of DEPC-H₂O.
- 70 mM rNTP solutions: Dissolve each NTP (ribonucleoside triphosphates) in 13 ml of DEPC-H₂O (~150 mM NTP). While periodically checking the pH of a small aliquot (~100 µl) on a pH dipstick, adjust each solution to pH 7.0 by adding 1M NaOH solution (~3ml for ATP and CTP, and ~1.2 ml for GTP and UTP). Prepare four 3-fold serial dilutions (1ml each) of each neutralized NTP solution. Measure the absorbance of each dilution and calculate the nucleotide concentration using Beer-Lambert's law (Extinction coefficients of each nucleotide are; 1.54×10^4 at 259 nm for ATP, 1.37×10^4 at 253 nm for GTP, 9.10×10^3 at 271 nm for CTP, and 9.60×10^3 at 267 nm for UTP). Adjust the concentration of each nucleotide to 70 mM with DEPC-H₂O. Make 0.5 ml aliquots and store at –20 °C for up to a year.
- 17.5 mM rNTP solutions: Prepare 17.5 mM rNTP mix by combining 1ml of each 70 mM rNTP solutions (ATP, CTP, GTP, and UTP). Store 0.5 ml aliquots at –20 °C for up to a year.

- 2.5 mM rNTP solutions: Prepare 2.5 mM rNTP mix by diluting 250 µl of 17.5 mM rNTP mix with 1.5 ml DEPC-H₂O. Store 0.5 ml aliquots at –20°C for up to a month.
- 1 µM Tus Binding solution: Combine 225 µl of 10X T7 Transcription Buffer, 69.44 µl of 32.4 µM GST-Tus, and 1955.56 µl of DEPC-H₂O. Adjust the protein and water amounts depending on the concentration of the GST-Tus protein prep. Prepare freshly, mix well by gently pipetting up and down, and keep on ice until ready to load to the instrument.
- Multiple Round Transcription Mix: Combine 225 µl of 10X T7 Transcription Buffer, 22.5 µl of 0.1 U/µl YIPP, 22.5 µl of Suprase Inhibitor, 450 µl of 2.5 mM rNTP mix, 6.75 µl of 4 mg/ml home-made T7 RNA Polymerase, and 1523.25 µl of DEPC-H₂O. Prepare freshly, mix well by gently pipetting up and down, and keep on ice until ready to load to the instrument.
- Klenow Enzyme Mix: Combine 225 µl of 10X NEB Buffer 2, 56.25 µl of 10 mM dNTP mix (each), 45 µl Klenow exo- enzyme, 1.13 µl of 20% (v/v) Tween 20, and 1922.63 µl of DEPC-H₂O. Prepare freshly, mix well by gently pipetting up and down, and keep on ice until ready to load to the instrument.
- GFP Aptamer Binding Buffer: Combine 225 µl of 10X PBS, 11.25 µl of 1M MgCl₂, 1.13 µl of 20% (v/v) Tween 20, 11.25 µl of Suprase Inhibitor, and 2001.38 µl of DEPC-H₂O. Prepare freshly, mix well by gently pipetting up and down, and keep on ice until ready to load to the instrument.

CRITICAL: The protein binding buffer will depend upon the application; a buffer that mimicks the physiological conditions under which the assayed RBP-RNA interaction occurs should be used as the HiTS-RAP Protein Binding Buffer.

- Primer Rehybridization Buffer: Dilute 18 µl of 100 µM IllumFORAdapt_T1_IllumFORSeq oligo in 2232 µl of Illumina Hyb1 Buffer. Prepare freshly, mix well by gently pipetting up and down, and keep on ice until ready to load to the instrument.
- 4N NaOH solution: Dissolve 160 g of NaOH pellet gradually in 500 ml MilliQ H₂O while stirring on a magnetic stirrer. Once all the pellets are dissolved, adjust the volume to 1L with MilliQ H₂O, and store at room-temperature.
- 0.1N NaOH solution: Dilute 300 µl of 4N NaOH solution with 11.7 ml of MilliQ-H₂O. Prepare it fresh and discard the remaining unused solution.

CAUTION: NaOH is basic and generates significant heat when dissolving, therefore it can cause skin and eye damage, and can lead to burns. Avoid contact with skin and eyes. Wear protective equipment when handling NaOH.

- 1X NEB Buffer 2: Combine 225 µl of 10X NEB Buffer 2, 1.13 µl of 20% (v/v) Tween 20, and 2023.88 µl of DEPC-H₂O. Prepare freshly, mix well by gently pipetting up and down, and keep on ice until ready to load to the instrument.

- General solutions such as 1M MgCl₂, 1M HEPES-KOH pH7.8, are prepared according to common practices used in molecular biology laboratories^{41, 42}.
- mOrange-EGFP protein solutions: To prepare 3.75 ml of 625 nM mOrange-EGFP protein solution, combine 375 µl of 10X PBS, 18.75 µl of 1M MgCl₂, 1.88 µl of 20% (v/v) Tween 20, 18.75 µl of Suprase Inhibitor, 73.24 µl of 32µM mOrange-EGFP protein, and 3262.4 µl of DEPC-H₂O. Prepare freshly, mix well by gently pipetting up and down, keep on ice, and protect from light until ready to load to the instrument. Prepare 125, 25, 5, 1, 0.2, and 0.04 nM mOrange-EGFP protein solutions by doing a 5-fold serial dilution with GFP Aptamer Binding Buffer. For example, for 125 nM solution, mix 0.75 ml of 625 nM solution with 3 ml of GFP Aptamer Binding Buffer. Mix each well by gently pipetting up and down, keep on ice, and protect from light until ready to load to the instrument.

PROCEDURE

Preparation of mOrange-RNA binding protein and GST-Tus fusion proteins. TIMING 10+ days

1. Transform the two protein expression constructs; i) mOrange pHis-// plasmid containing ORF of the protein of interest (an RNA Binding Protein) and ii) Tus pGST-// plasmid, separately to an E.coli line suitable for recombinant protein expression (e.g. BL21-CodonPlus(DE3)-RIPL). Growth of bacterial cultures and protein purification can be carried out in parallel for GST-Tus and His₆-mOrange-protein of interest fusion.
2. Plate transformants on LB-Agar plate containing 100 µg/ml Ampicillin and incubate at 37°C overnight.
3. On the next day, inoculate 20 ml LB media containing 100 µg/ml Ampicillin and 34 µg/ml Chloramphenicol with a single colony from each transformation.
4. Incubate the cultures at 37°C shaker at 250 rpm for ~16 hours.
5. On the third day, inoculate 1 L LB media containing 100 µg/ml Ampicillin and 34 µg/ml Chloramphenicol with 10 ml of each overnight culture.
6. Incubate the 1 L culture in a 37 °C shaker until OD₆₀₀ reaches 0.6 (about ~3 hours).
7. Induce recombinant protein expression by supplementing the cultures with IPTG to a final concentration of 200 µM.
8. Incubate induced cultures at 18–25°C shaker for 12–16 hours.
9. Collect the bacteria by centrifuging at 4,500g 4°C for 30 min.
10. Discard the media, flash-freeze the bacterial pellet in liquid nitrogen, and store at –80 °C.

PAUSE POINT: Bacterial pellets can be stored at –80 °C until ready to proceed with purification of recombinant proteins (up to a month).

11. Purify GST-Tus and His₆-mOrange-protein of interest fusions using Glutathione-agarose and Ni-NTA Superflow resins, respectively, according to the manufacturers' protocols.
12. Dialyze the purified protein into 10 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, and 5 mM β -mercaptoethanol using a 20K MWCO Slide-A-Lyzer dialysis cassette.
13. Incubate the dialyzed His₆-mOrange-protein of interest fusion protein at 4 °C for at least a week to allow for maturation of the mOrange fluorescent protein. GST-Tus protein does not require any maturation, and can be stored immediately after dialysis
14. Mix both protein preps with an equal volume of 80% glycerol, and store them at –80 °C up to a year in 0.5 ml aliquots.

CRITICAL STEP: mOrange is light sensitive, and should be protected from light.
15. Check the integrity and purity of the protein preps by SDS-PAGE analysis⁴³ and Coomassie Blue staining⁴⁴.
16. Measure the protein concentration using Bradford Assay⁴⁵.

PAUSE POINT: Once the quality and concentration of protein preps are determined, the purified recombinant proteins can be stored at –80 °C until ready to proceed with HiTS-RAP (up to a year).

Preparation of HiTS-RAP DNA templates. TIMING 2 days

17. Set up the PCR reaction for amplification of the DNA template encoding the RNA library of interest. This first PCR reaction will introduce a T7 promoter (on the forward oligo) at the 5'-end of the template DNA (e.g. GFP_Temp_DNA_FOR for GFP Aptamer, Table 2) and a library specific reverse oligo containing the Illumina sequencing oligo and a barcode at its 5'-end (e.g. Temp_DNA_REV for GFP Aptamer, Table 2).

Component	Amount (μ l)	Final concentration
RNA encoding DNA template (100 pg/ μ l)	1	1 pg/ μ l
PCR buffer, 10X	10	1X
dNTP mix, 10 mM (each)	2	200 μ M
Forward oligo, 100 μ M	0.5	0.5 μ M
Reverse oligo, 100 μ M	0.5	0.5 μ M
Taq DNA polymerase	2	
DNase-free water	84	
Total	100	

18. Perform PCR amplification using the following cycling conditions.

Cycle number	Denature	Anneal	Extend
1	95 °C, 3 min		
2–21	95 °C, 30 s	55 °C, 30 s	72 °C, 45 s
22			72 °C, 5 min

19. Verify the size and proper amplification of the template by running 3 µl of the PCR product and MassRuler DNA Ladder on an 8% native polyacrylamide gel, stain the gel with 1 µg/ml Ethidium Bromide (EtBr) solution for 10 min at room-temperature with shaking, and visualize the DNA bands with UV illumination using a GelDoc system.

CAUTION: EtBr is mutagenic and carcinogenic. UV light can cause eye damage, burns on skin, and skin cancer. Wear protective gear when working with EtBr or UV light.

20. Purify the PCR amplified DNA using a PCR purification kit.
21. Quantify the DNA using Qubit dsDNA HS Assay.
22. Repeat steps 17–21 using the purified DNA as the template for this second PCR to introduce Illumina forward flowcell adaptor on the 5'-end and Ter site and Illumina reverse flowcell adaptor on the 3'-end of the template DNA (e.g. Illum_IllumAdaptor_T7 and IllumFORAdapt_T1_IllumFORSeq oligos for GFP aptamer, Table 1).
23. Purify the final PCR amplified template from an 8% native polyacrylamide gel. Run the remaining PCR reaction on a gel with a large (for PCR product) and a small well (for DNA Ladder) (e.g. Prep+1 well, BioRad).
24. Stain the gel with 1 µg/ml EtBr solution for 10 min at room-temperature with shaking.
25. Quickly excise the DNA bands while visualizing on a long-wavelength UV transilluminator box (365 nm UV). Transfer the gel piece into a 0.6 ml centrifuge tube.
26. Poke a hole at the bottom of the tube with a 21G needle and insert the gel containing 0.6 ml tube into a clean 1.7 ml centrifuge tube.
27. Crush the gel into fine pieces by centrifuging the two-tube assembly at 13,000g for 1 minute, which forces the gel piece to go through the needle hole. If there is any remaining gel piece in the 0.6 ml tube, centrifuge it into another clean 1.7 ml tube.
28. Remove the 0.6 ml tube, and soak the gel pieces in 1 ml 10 mM Tris.Cl pH 7.5 overnight at 37°C rotator.
29. Remove gel pieces using Costar SpinX centrifuge tube filters (Corning).
30. Clean up DNA by performing Phenol/Chloroform and Chloroform extractions. Mix 1 ml pH 8.0-buffered Phenol/Chloroform mix (equal volume to DNA

sample) with DNA sample. Vortex for 15 seconds, centrifuge for 5 minutes at room-temperature at 13,000g. Transfer aqueous upper layer to a clean tube. Repeat extraction with Chloroform.

CAUTION: Phenol is highly toxic. Wear protective gear when handling.

CRITICAL STEP: Carryover of Phenol/Chloroform could potentially inhibit downstream enzymatic manipulation of the DNA template. Leave a small fraction of the aqueous layer, to prevent Phenol/Chloroform carryover.

31. Precipitate DNA with EtOH. Split DNA sample (~900 µl left after extractions) into three tubes (300 µl each). Add 1 µl of 15 µg/µl GlycoBlue (Ambion) to aid better precipitation of DNA and visualization of DNA pellet, 1/10th volume of 3M NaOAc, and 3 volumes of cold (-20°C) EtOH. Incubate at -20°C for 30 minutes. Centrifuge at 4°C 20,000g for 25 minutes. Discard supernatant, wash the pellet with 0.6 ml 70% EtOH. Centrifuge at 4°C 20,000g for 5 minutes. Pipet off the 70% EtOH wash and air dry the DNA pellet for ~5 minutes on bench. Resuspend the DNA pellet in 50 µl of 10 mM Tris.Cl pH 8.5, 0.1% Tween-20 solution.
32. Verify the integrity and the size of the DNA template by running 3 µl of it on an 8% native polyacrylamide gel.
33. Measure the DNA concentration using Qubit dsDNA HS Assay (Invitrogen).
34. Store the template DNA in -20 °C.

PAUSE POINT: Once the quality and concentration of DNA template(s) are determined, they can be stored at -20 °C until ready to proceed with HiTS-RAP (up to a year).

Cluster generation on Illumina GAIIX flowcell using cBot. TIMING ~6 hours

CRITICAL: Preparation of Illumina cluster generation is described in the section below; however, the readers are strongly urged to consult the Illumina cBot™ User Guide (Part # 15006165 Rev. J July 2012) for more detailed protocols and critical tips.

35. Thaw reagents for cluster generation either in a room-temperature waterbath for ~1 hour or at 4°C overnight for a maximum of 16 hours.
36. Prepare 20 pM NaOH denatured DNA template. Dilute DNA templates and the PhiX (control) DNA to 2 nM with 10 mM Tris.Cl pH 8.5, 0.1% Tween-20 (v/v) solution. Mix 10 µl of 2 nM DNA template with 10 µl of freshly prepared 0.1 N NaOH. Incubate 5 minutes at room-temperature to denature the dsDNA to single strands. Transfer 20 µl denatured DNA into 980 µl of pre-chilled HT1 Hybridization Buffer (Illumina) to obtain 20 pM denatured DNA. Keep 20 pM denatured DNA on ice until ready to load the cBot.

CRITICAL STEP: The Illumina GAIIX user guide calls for a 20 pM denatured library to be used for cluster generation. In our experience, 20 pM libraries give rise to very few clusters on the flowcell, whereas 200 pM libraries (as quantified by Qubit DNA HS assay) yield more or less the optimal cluster

density (~20 million per lane). If needed, 200 pM library can be further diluted with Buffer HT1 to a lower concentration.

37. Transfer 120 μ l of each of the 200 pM denatured DNA samples to 8-well strip tube. If custom primers (other than the standard Illumina sequencing primer) are to be used, load 120 μ l of 1 μ M primer in buffer HT1 in each well of a second 8-well strip tube.
38. Mix the thawed reagents by inverting a few times and collect reagents by brief centrifugation. Carefully remove non-piercable red foil from row 10 of the reagent tray. This tube strip contains the NaOH solution used for denaturing DNA between amplification cycles, so avoiding spills is critical.
39. Set up the cluster generation run as described in detail in the cBot user guide. Follow cBot instructions displayed on-screen with images. Select a protocol. For a paired end flowcell where custom sequencing primers are to be used (and hybridized on the cBot), the following protocol would be used: PE_Amp_Lin_Block_TubeStripHyb_v7.0. Importantly, PE indicates a paired-end flowcell (SR, single read), TubeStripHyb indicates that custom primers are being used (Hyb would use the standard sequencing primers included in the cluster generation kit), and v7.0 indicates a GA flowcell (v8.0 is for the HiSeq flowcell). Perform a pre-run wash on cBot.
40. Load the reagent plate into the cBot instrument after scanning the reagents barcode. Load and position the GAIIX flowcell after scanning the barcode. Load the manifold matching the GAIIX flowcell and secure the outlet end and the sipper comb of the manifold. Load the DNA templates (and the custom primers if being used) in 8-well strip tubes. When using custom primers for sequencing, load the 8-well tube strip containing 120 μ l of 1 μ M primer in Illumina hybridization buffer (HT1) in each well into the PRIMER position on the cBot. Close the lid of the cBot.
41. Perform a pre-run check.
42. If pre-run check completes successfully start the run to generate clusters of DNA templates on the flowcell.
43. When the run is complete, remove the manifold, then remove the processed flowcell and confirm the successful and even delivery of reagents to individual lanes.
44. Return the clustered flowcell to the storage buffer within the container that it was shipped in.

PAUSE POINT: Clustered flowcells can be stored in the storage buffer at 4 °C for up to a month before starting the HiTS-RAP run. If the primer hybridization was already done on the cBot, primer hybridized flowcells can only be stored for up to a week at 4 °C.
45. Perform a post-run wash of the cBot instrument.

Performing HiTS-RAP on Illumina GAIIX with Paired-End Module. TIMING ~10 days

CRITICAL: See Illumina Genome Analyzer IIX User Guide (Using SCS v2.9) (Part # 15018814 Rev. B November 2011) for more detailed protocol for how to set up the instrument and the sequencing part of HiTS-RAP.

Setting up the Illumina GAIIX Instrument for a HiTS-RAP Run

46. Start the SCS2.9 software. In the Run Parameter dialogue box, make the following changes (See Supplementary Tutorial 1 for images):

Tab	Change
SCS and RTA Setting	<ul style="list-style-type: none"> • Check the following boxes: Base calls per cycle (.BCL), Images (.TIF), Intensities (.CIF), Run Logs and Files, Enable Paired-End Module, Enable base calling and quality scoring (BCL), Use control lane, Enable auto calibration. • Set the values for following fields: Network run folder path: D:\DATA, Control lane: Lane # (the lane containing the PhiX sample), Focus laser exposure time: 20, Find Focus, Channel: [A], Laser: [R], Exposure: 100, Find Edge, Channel: [A], Laser: [R], Exposure: 400, Focus Tile, Lane: 4, Column: 1, Row: 25.
Image Save Setting	Save 10% of Images (starting at tile 5).

CRITICAL STEP: Make sure there is enough disk space for data storage under D:\DATA folder. A HiTS-RAP run with 82 nt sequencing and 7 concentration protein binding can generate upwards of 600 GB data (.bcl, .cif, and .tif files combined). Additional free disk space is required for uninterrupted real-time data analysis by the RTA software. Also, make sure the ImageCyclePump config file is set to ImageCyclePump On=true
AutoDispense = false.

47. Perform a pre-run wash using a previously used flowcell. Load 10 ml PW1 solution (Illumina) on positions 1,3, and 6, 40 ml MilliQ H₂O or PW1 solution (Illumina) on positions 2,4,5, and 7 of GAIIX instrument, and 5 ml MilliQ H₂O on positions 9–22 of PEM. Run the GA2-PEM_PreWash_v8.xml recipe to perform pre-run wash. Make sure 18 ml of solution is collected in the waste container.
48. Thaw sequencing reagents. Determine the number of sequencing kits necessary for sequencing; each 36 bp kit provides enough reagents to perform 42

nucleotide sequencing. Thaw IMR, LFN, and SMX in a room-temperature waterbath for an hour and keep the rest of the reagents at their long-term storage temperature. Thaw CLM in a separate waterbath and keep it separated from other reagents at all times. Once thawed, keep the reagents on ice.

49. Mix PR1, PR2, and PR3 solutions by inverting a few times.
50. Combine the identical reagents from multiple kits and transfer them into 125 ml or 150 ml bottles.
51. Prepare IMR solution. For each 36 bp kit, transfer IMR into 175 ml Falcon bottle, add 3.52 ml of LFN and 330 μ l of HDP into IMR solution, mix by inverting a few times, and collect by centrifugation at 1,000g for 1 minute at 4°C. Keep the IMR/LFN/HDP mix on ice until ready to load onto the GAIIX instrument.
52. Prepare SMX solution. Mix SMX solution by inverting a few times, then transfer into 175 ml Falcon bottle. Keep on ice until ready to load onto the GAIIX instrument.
53. Prepare CLM solution. Mix CLM solution by inverting a few times, collect by centrifugation at 1,000g for 1 minute at 4°C, and then transfer into 175 ml Falcon bottle. Keep on ice until ready to load onto the GAIIX instrument.

CRITICAL STEP: Keep CLM solution away from other reagents at all times and after handling CLM reagent always change gloves. CLM solution is used to chemically cleave the dyes from dNTPs incorporated during sequencing, and thus could inactivate nucleotides even before incorporation during sequencing if mixing occurs.

CRITICAL STEP: When using multiple kits, the containers (125, 150, or 175 ml bottles) may not have enough room to contain the entire reagent necessary for sequencing at once. Therefore, it may be necessary to replenish these reagents at various points during the sequencing (see Illumina GAIIX user guide). UserWait steps can be incorporated into the recipe, to accommodate refilling of reagents before they run out.

54. Load the reagents onto the GAIIX instrument. Centrifuge the reagent containing bottles at 1,000g for 1 minute at 4°C right before loading. Connect the reagent bottles at appropriate positions:

Position	Reagent	Bottle (ml)
1	IMR/LFN/HDP mix	175
2	PW1	125
3	SMX	175
4	PR1	125 or 150
5	PR2	125 or 150

Position	Reagent	Bottle (ml)
7	PR3	125 or 150
6	CLM	175

55. Prime the reagent delivery by running GA2_Prime_v10.xml recipe. Collect the waste and make sure the waste volume is 6.4 ml (+/- 10%).
56. Unload the old flowcell from the instrument.
57. Clean the prism using 100% MetOH and lint-free lens cleaning tissue, and install the cleaned prism into the GAIIX instrument.
58. Clean both the bottom and top faces of the processed flowcell, harboring the DNA clusters on its surface (from step 45), with 100% EtOH using lint-free lens cleaning tissue.

CRITICAL STEP: Do not touch the reagent ports on the top face of the flowcell with the tissue, as this could introduce fibers to the flowcell, and ethanol could disrupt the primers hybridized to clusters. Illumina GAIIX instrument is a TIRF microscope at its core. DNA clusters on the flowcell are illuminated from bottom through the prism and imaged from the top of the flowcell, therefore it is absolutely critical that both surfaces of the flowcell as well as the angled and flat faces of the prism are pristine (free of any dust, dirt, or oil).

59. Load the cleaned flowcell. Scan the flowcell ID. Place the flowcell on top of the prism and secure it. Ensure the flowcell is appropriately seated.
60. Check reagent delivery. Using the Manual Control/Setup tab of the control software, pump 100 µl of PR2 incorporation buffer (Position 5) to the flowcell using a 250 µl/min aspiration rate and 2500 µl/min dispense rate. Visually confirm the flow of liquid in each lane and make sure that air bubbles are cleared. Check the joint between flowcell and the front manifold for leaks using a clean dry lens cleaning tissue. Measure the volume of liquid dispensed into the waste container, expected volume is 800 µl (+/- 10%). Repeat this step twice more.
61. Apply ~150 µl of immersion oil with a refractive index 1.473 (Cargille, catalog # 19570) between the flowcell and the prism from the left side. Make sure that there are no air bubbles between the flowcell and the prism, and no excess oil on the opposite (right) side of the flowcell. Close the instrument door.

Sequencing, dsDNA regeneration, transcription halting, and protein binding

62. Using the SCS2.9 software select the HiTS-RAP recipe (Supplementary Software 1) and start the run.

CRITICAL STEP: The recipe can be modified using an appropriate text editing software such as Notepad++ for Windows (it is helpful to have this program installed this on GAIIX computer as well), or Gedit for Linux. The provided recipe (SI 1) that was used for HiTS-RAP analysis of GFP aptamer –

EGFP-mOrange fusion protein interaction performs an 82 nt sequencing run and protein binding reactions at 7 different concentrations. Many of these steps can be changed to meet desired length of sequencing and number of protein concentrations used for protein binding.

63. After incorporation of the first base, check the quality metrics X obtained after auto-calibration (under Calibration Results). If satisfactory (Goodness of fit 0.9900 and sensitivity between 350–400), resume the recipe to continue with the remaining cycles of sequencing.

64. Replenish the sequencing reagents as needed.

CRITICAL STEP: The provided recipe has UserWait pause after cycle 62 at which point the quality of the sequencing run can be assessed. If the reagents have not run out before and individual DNA clusters are visible in previous sequencing cycles, reagents can be replenished and sequencing can be resumed.

65. Once the sequencing part of HiTS-RAP finishes, the recipe calls for a pause, at which time the reagents for the rest of the HiTS-RAP (see below) can be loaded onto PEM and the recipe can be resumed afterwards.

TROUBLESHOOTING

66. Modify the ImageCyclePump config file by changing ImageCyclePump On=true AutoDispense = false to ImageCyclePump On=false AutoDispense = false.

CRITICAL STEP: This modification will prevent the delivery of scan mix (SMX), used for imaging during sequencing, to the flowcell. SMX will dissociate the halted transcription complex.

67. Load the indicated amounts of following reagents at specified positions on the GAIIX PEM.

Position	Reagent	Amount (ml)
9	1X T7 Transcription buffer	2.25
10	1 μ M Tus binding solution	2.25
11	Multiple Round Transcription Mix	2.25
13	Klenow Enzyme Mix	2.25
14	GFP Aptamer Binding Buffer	2.25
16	Primer Rehybridization Buffer	2.25
19	0.1 N NaOH	11.25
20	1X NEB Buffer 2	2.25
21	Illumina Hyb1 Buffer	4.5

CRITICAL STEP: The recommended amount of each solution, loaded onto the PEM, is larger than what is delivered to the flowcell (i.e. ~1.32 ml of a

2.25 ml solution). The difference accounts for priming (~0.5 ml) and ensures that air bubbles are not introduced to the tubing of PEM and the flowcell.

CRITICAL STEP: Positions 9 to 13 on the PEM are chilled (at 4 °C) and thus used for enzyme and protein solutions, the rest are at room-temperature and used primarily for buffers. If protein solution must be pulled from a room temperature port, we use it before the refrigerated positions and for the lowest protein concentrations. If refrigeration is absolutely essential, positions 1, 3, and 6 on GAIIX can be used with appropriate modifications to HiTS-RAP recipe.

68. Resume the HiTS-RAP run. The GAIIX will automatically perform the following reactions in sequence: (i) regenerate dsDNA template, (ii) bind Tus protein to dsDNA template, and (iii) transcribe the template DNA with T7 RNA polymerase. This will generate a halted transcription complex with the RNA transcript being retained on the DNA template for interaction with protein target.
69. Once the dsDNA regeneration, Tus binding, and transcription steps are finished, the recipe calls for a pause and asks user to load target protein solutions. Load 2.25 ml of the following target protein solutions (at 7 different concentrations with 5-fold increments in GFP Aptamer Binding Buffer) at the indicated positions of PEM.

Position	Target Protein Solutions	Temperature
14	GFP Aptamer Binding Buffer	Room temperature
16	0.04 nM mOrange-EGFP	Room temperature
20	0.2 nM mOrange-EGFP	Room temperature
9	1 nM mOrange-EGFP	chilled (4 °C)
10	5 nM mOrange-EGFP	chilled (4 °C)
11	25 nM mOrange-EGFP	chilled (4 °C)
12	125 nM mOrange-EGFP	chilled (4 °C)
13	625 nM mOrange-EGFP	chilled (4 °C)

CRITICAL STEP: The indicated range of protein concentrations (0.04–625 nM), designed for HiTS-RAP analysis of EGFP-GFP aptamer interaction with 5 nM K_d , may need to be adjusted higher or lower depending on the known (or expected) affinity of target protein for the RNAs present on the flowcell.

70. Resume the HiTS-RAP run. The GAIIX will image protein binding at each of these concentrations sequentially starting from no protein to 625 nM mOrange-EGFP protein. The data will be recorded automatically.

TROUBLESHOOTING

71. Once the protein binding steps are finished, the recipe calls for another pause and asks the user to unload reagents from GAIIX and PEM instruments, to revert

the ImageCyclePump config file to its original form, and to perform maintenance washes to clean up the instruments.

72. Modify the ImageCyclePump config file by changing ImageCyclePump On=false AutoDispense = false to ImageCyclePump On=true AutoDispense = false.

CRITICAL STEP: If not done before the next sequencing run, the modified version of ImageCyclePump config file will prevent the delivery of SMX to the flowcell, which is necessary for proper sequencing.

73. Click OK to complete and terminate the HiTS-RAP run.

Post-HiTS-RAP instrument clean-up

74. Perform a post-run wash using a previously used flowcell. Load 10 ml PW1 solution (Illumina) on positions 1,3, and 6, 40 ml PW1 solution (Illumina) on positions 2,4,5, and 7 of GAIIX instrument, and 5 ml MilliQ H₂O on positions 9–22 of PEM. Run GA2-PEM_PostWash_v7.xml recipe to perform pre-run wash. If additional ports were used, add them to a new custom PostWash recipe.
75. Perform a maintenance wash, as described in the Illumina GAIIX User Guide. If reagent positions used for HiTS-RAP are not washed in the standard protocol, add them in a custom recipe.

PAUSE POINT: Once the HiTS-RAP run is complete and the GAIIX instrument is cleaned up, the sequencing and protein binding data can be transferred to the Linux workstation at any time when the user is ready to proceed with the subsequent data analysis.

Data retrieval and analysis. TIMING ~4 days

Convert .bcl files to .qseq.txt files as described in steps 76–79: This is done as described in the Off-Line Basecaller v1.9.4 User Guide. To install the OLB, we use CentOS Linux. The following steps are command line processes.

76. Move to the basecalls directory within the run folder. `cd path/to/140721_HWUSI-EAS690_00067_FC/Data/Intensities/BaseCalls/` where 140721... is the name of the run folder (named by date_instrument_run number).
77. Run `setupBclToQseq.py` to generate the make file. `path/to/OLB-1.9.4/bin/setupBclToQseq.py -b path/to/140721_HWUSI-EAS690_00067_FC/Data/Intensities/BaseCalls -o path/to/output/directory` The output directory is a new directory created when running the converter.
78. Run the make file from the basecalls directory. When the converter is finished, it will have created the directory indicated with the `-o` argument, and filled it with the output. The directory will contain a `qseq.txt` file for each tile, for each lane. They are named as follows:
`s_<lanenumber>_<readnumber>_<tilenumber>_qseq.txt`

As in the following example for tile number 36 from lane 5:
s_5_1_0036_qseq.txt.

The qseq.txt files are tab delineated text files, where each row is a read. A full description of these files can be found in the Illumina CASAVA v1.8 user Guide. Briefly, each entry consists of: <Machine name> <Run number> <Lane number> <Tile number> <X coordinate> <Y coordinate> <Index (zero for runs with no index read)> <Read number (1 for single read runs)> <Sequence> <Quality String> <Pass filter (0=no, 1=yes)>

The location information (lane, tile, x, and y) is used to match each cluster's sequence, contained within these files, with its intensity stored in .cif files. The quality string and pass filter are used to determine whether a read is suitable for inclusion in affinity analyses.

Extract intensity information from .cif files: A Perl script originally written for analysis of HiTS-FLIP data is used to read the binary .cif files, and output intensities with coordinates as a tab delineated text files¹³.

79. Run the Perl script `extract_intensity_cif.pl` to convert the intensity information from the binary .cif files into plain text files. For a complete list of options, run the script with no flags using “`perl extract_intensity_cif.pl`”. The most important variables to set will be `-intensities_dir`, set to the path/to//140721_HWUSI-EAS690_00067_FC/Data/Intensities/ path, and `-first_protein_cycle`, which provides the cycle number at which to stop recording sequencing intensities and start recording protein-binding intensities. When run, the script will output one text file per lane of the form `L00X_Intensities.txt`. We typically create a directory for the RNA binding data within the run folder, and carry out all analyses there.

Example: `perl extract_intensity_cif.pl -intensities_dir path/to//140721_HWUSI-EAS690_00067_FC/Data/Intensities/ -lanes 1-7 -script_path /usr/local/ -first_protein_cycle 51`

CRITICAL STEP: If the helper scripts “`ExtractIntensitiesFromCIF_T.py`” and “`ExtractIntensitiesFromCif_max.py`” are not found in the same directory that the “`extract_intensity_cif.pl`” script is run in, specify their location using the `-script_path` option. If analyzing a subset of lanes, use the flag `-lanes`, providing a range of the form `-lanes 1-7`.

Solve K_d values (Steps 80 & 81): At this point, two types of files have been produced, containing sequence and intensity information. These data could be used to solve K_d values in any way that the user sees fit. To analyze data using our python pipeline;

80. Edit the input variables at the top of `HiTS_RAP_Analysis_Pipeline.py` according to the run carried out following the comments in the script.

CRITICAL STEP: The following considerations should be taken into account: Firstly, the pipeline generates several intermediate files. These are helpful for

troubleshooting, as the files are large and may take too long to reproduce many times over. This gives the option of restarting the pipeline in several places, with different functions. Secondly, we correct for the slow RNA complex decay rate as previously described¹⁴.

81. Run the pipeline from the terminal:

python HiTS_RAP_Analysis_Pipeline.py. This will generate a tab-delineated text file in which each line contains the following information in the given order: the number of reads representing the sequence, the sequence, the average intensity during sequencing, average normalized protein binding intensities (7 values, one for each protein concentration used), standard deviations of the protein binding intensities (7), fitted parameters (K_d , Hill coefficient, base intensity, maximum intensity), and covariance of each of the fitted parameters (4).

TIMING

Preparation of mOrange-RNA binding protein and GST-Tus fusion proteins

Steps 1–18 10+ days (~3 days hands-on)

Preparation of HiTS-RAP DNA templates

Steps 19–36 2 days

Cluster generation on Illumina GAIIX flowcell using cBot

Steps 37–46 ~6 hours (~30 minutes hands-on)

Performing HiTS-RAP on Illumina GAIIX with Paired-End Module

Steps 47–79 ~10 days (~1 day hands-on)

Data retrieval and analysis

Steps 80–82 ~4 day

ANTICIPATED RESULTS

The number of K_d measurements that can be made in a single HiTS-RAP experiment depends entirely on the library and the application. We have solved as many as 10,000 dissociation constants from a single lane in experiments with SELEX libraries and mutagenized aptamers. In our experience, HiTS-RAP gives reliable results for sequences represented by as few as 5 clusters in a lane, so with an ideal library (20 million reads per lane) up to ~4 million affinity measurements per lane should be possible. When analyzing targeted libraries with long sequences, approaching this maximum might be difficult. However, for other applications, such as k-mer analysis with a longer random library, the number of motifs that can be assessed with a single lane could exceed the number of clusters.

The first step in assessing the success of a HiTS-RAP run is to make sure that the sequencing portion of the run was successful. This can be determined from the run metrics in the Sequencing Analysis Viewer (SAV) software as the run progresses (though some metrics are not available until cycle 25). Favorable run metrics indicate that the instrument is functioning properly. We generally try to reach a cluster density of 200,000–400,000 clusters/mm². Underclustering means that the potential of the assay is underutilized, and overclustering results in lower quality data, and lower pass filter rates. The percent Q30 serves as a measure of overall confidence in basecalling in a lane. In successful runs, at least 85 to 90% of reads are above Q30. A lower Q30 rate indicates that the instrument is not functioning properly: we have observed lower rates when lanes are overclustered and when the optics of the instrument were not functioning properly. The SAV will also report error rates for lanes containing the PhiX genome control. Error rates are typically low (below 1%). It is important to only check the Q30 and error rate values for sequencing cycles, and to keep in mind that they begin to drop precipitously with longer reads (past ~75 cycles). Once the protein binding portion of the HiTS-RAP starts, the aggregate values for these metrics will decrease drastically, as the SCS includes the measurements from these cycles in calculation of the run metrics.

Assessing HiTS-RAP results will depend on the application. When we have applied it to aptamer libraries, we expect most sequences assessed to bind to the target protein, while in other applications binding could be a rare event. We first look at the intensity measurements reported in the Illumina SAV during the protein binding to assess whether protein binding is taking place. With libraries in which the majority of sequences bind the protein, the binding curve can be seen in the T and G channels. If a negative control (or PhiX) lane was included, lack of signal in that lane indicates low background and low nonspecific binding of the labeled protein to DNA and other components of a halted transcription complex (i.e. T7 RNA polymerase and Tus protein).

The results of K_d curve fitting, as described in Data Retrieval and Analysis section, will ultimately determine whether the run was successful or not. Low covariances of fitted parameters, Hill coefficients that are close to one, base and maximum values close to the first and last protein binding intensity measurements, and K_d s within the range of protein concentrations assayed are all indicators of good fits and thus a successful HiTS-RAP run.

The protein binding curves used to determine K_d s should also be inspected to determine whether a run has worked well and the results obtained from data analysis are consistent. In the lowest concentrations of protein, fluorescence intensities of protein binding should be near background, similar to that of the no protein control. As the protein concentration increases, intensities should also increase. We have found that the higher intensity measurements generally appear to be noisier than lower intensity measurements. Systematic shifts in intensity across the lane are also not uncommon. Transition to protein binding is often easily identifiable in raw HiTS-RAP data, but the measurements after that point tend to be noisy. In our most successful runs, we are able to solve reliable K_d s through a range of affinities. Figure 5a shows a series from one such run, in which different sequences, all from the same lane, show low fluorescence in the first few protein concentrations, and then saturate binding at different points. Seeing a range of affinities in one lane is the ultimate

indicator of a successful HiTS-RAP run, and that the binding measurements observed are due to real binding, rather than a systematic bias.

Binding curves that do not match the calculated affinities are indicators of a failed measurement. There could be many reasons for a measurement to be unreliable. Sometimes, the fitting algorithm simply fails to find parameters that are a good match for the data (Fig. 5b): such a fit will usually have a high covariance value. Other times, only one binding measurement rises above background (Fig. 5c). This occurs in cases where the range of protein concentrations assayed were not within the range of the affinities of the RNAs probed. We have also observed instances where the intensity data show no binding at all (Fig. 5d). These data are from a run where several different protein binding were carried out on the same flowcell (sequenced once) after denaturing and remaking double stranded DNA, transcribing and halting, and introducing each labeled protein successively. In that run, we suspect that at least one of the biochemical steps leading up to the halted RNA complex was unsuccessful.

TROUBLESHOOTING

Step	Problem	Possible reason	Possible solution
65	No or very few high quality sequencing reads.	Under- or over-clustering of flowcell	Proper cluster generation is extremely important for extracting the maximum possible amount of high-quality data from a HiTS-RAP run. Both under- and over-clustering will result in reduced yields, and can ruin runs almost entirely. Unfortunately, this takes experience. Thus, we recommend erring on the side of under clustering the first time a run is carried out: we started by loading 10 pM library in our earliest runs. We have found that our libraries actually produce the optimal cluster density (~20 million per lane), when loaded with 200 pM denatured library (quantified by Qubit HS dsDNA Assay, Invitrogen), about ten times the maximum concentration recommended by Illumina. Quantitative PCR (see Illumina's Sequencing Library qPCR Quantification Guide (cat. no. SY-930-1010) for a detailed protocol) is the most precise way to measure library concentration, as it measures not just DNA concentration, but cluster forming units, when primers identical to those on the flowcell are used. Therefore, it likely provides the most reliable measurement of library concentration for new users.
		Inappropriate DNA template	Make sure DNA template is prepared properly, that is, it has all the necessary fragments and in the right orientation. If necessary remake the DNA template.
		Low or high DNA template concentration	Make sure the DNA template was properly quantified and diluted, as both higher and lower amounts leading to under- or over-clustering on the flowcell, respectively, can yield lower than ideal sequences in GAIIX. A 20 pM denatured DNA solution, as recommended by Illumina, can be used for cluster generation on flowcell using cBot.
		Inappropriate GAIIX set up	Make sure the ImageCyclePump config file is set up properly. For proper sequencing, this file should be set to ImageCyclePump On=true AutoDispense = false before starting the sequencing (before step 66). Other potential problems include a clogged port, precipitated Scan Mix or any other solution, expired or poorly stored reagents, a broken syringe, a poor seal of the manifold with the flowcell, bubbles in the oil under the flowcell, oil on top of the flowcell (excess oil can

Step	Problem	Possible reason	Possible solution
			get onto the Peltier cooler and spread to the top of the flowcell), and dust on the flowcell or prism. All of these possibilities should be inspected and if necessary, HiTS-RAP run should be repeated with new reagents and extra care.
		Unknown	Refer to Illumina GAIIX User Guide for potential solutions.
		Unknown	If problem persists, contact Illumina for further help.
70	No protein binding	Inappropriate GAIIX set up	Make sure the ImageCyclePump config file is set up properly. For protein binding, this file should be set to ImageCyclePump On=false AutoDispense = false before the start of protein binding (step 70). Since HiTS-RAP starts with sequencing and ends with protein binding, this modification is done while HiTS-RAP run is in progress.
		Inactive or non-functional reagent	Check reagents in vitro for activity (i) binding of mOrange-target protein fusion to RNA, (ii) binding of GST-Tus to DNA template, (iii) transcription of DNA template by T7 RNA polymerase, (iv) halting of T7 RNA polymerase and retention of RNA transcript by Tus bound to DNA template. If necessary, prepare buffers, solutions or protein preps again.
		RNA degradation	Check reagents for RNase contamination (e.g. RNaseAlert, Life Technologies (Cat. No. AM1964) or pH. RNases and alkaline conditions (> pH 8.0) can hydrolyze RNA. If necessary, prepare buffers, solutions or protein preps again. The GAIIX instrument can be the source of RNase contamination or alkaline conditions. Perform clean-up wash procedures described in Illumina GAIIX User Guide, being sure to wash thoroughly with water after a NaOH wash. We always include SupraseIn in every solution introduced into the flowcell starting with transcription step.
		Low target protein concentration	Check concentrations of mOrange fusion protein used for the assay. Concentrations that are significantly lower than the K_d of the RNA-RBP interaction will yield very low levels of protein binding, which might be difficult to detect by GAIIX. Adjust protein concentrations accordingly.
		Non- or weakly-fluorescent target protein	Verify the mOrange is matured and brightly fluorescent. If not, longer incubations for maturation maybe necessary. A 625 nM solution of the mOrange labeled protein of interest should be visibly orange when held up to a blank sheet of white paper.
		RNA contamination	Check the mOrange fusion protein for RNA contamination. Significant levels of co-purified RNA could hinder binding to RNA transcripts presented on the flowcell. If necessary, remake fusion protein preps after treating bacterial cell lysates with RNases.
70	Protein binding at very few DNA clusters.	Inappropriate DNA template	Check the DNA template. The library may contain very few protein-binding RNAs or is mutated excessively to the extent the protein binding is lost. Enrich the library for target protein-binding (e.g. by SELEX) or reduce the number of mutations in the DNA template.
70	Protein binding at every DNA cluster	High target protein concentration	Check the mOrange protein concentration. Higher concentrations of protein ($>>K_d$) may cause mOrange-target protein fusion to bind even the low affinity RNAs. Adjust protein concentrations accordingly.
		Non-specific binding of target protein	Check the interactions of target protein. If the target protein binds DNA, T7 RNA polymerase, or the Tus protein, this will cause high background in HiTS-

Step	Problem	Possible reason	Possible solution
			RAP. Consider adding excess carrier (e.g. salmon sperm DNA or T7 RNA polymerase or Tus) to target protein binding solutions. Use of an alternative method may be necessary.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J.M. Pagano (Cornell University) for providing NELF-E aptamer and protein constructs, and performing NELF-E EMSA experiments, C.B. Burge (Massachusetts Institute of Technology) for the development of the seminal HiTS-FLIP assay and the scripts used for extraction of protein binding data, B. Mohanty (Medical University of South Carolina) for providing vectors containing the *tus* gene, K. Szeto and D. Shalloway (Cornell University) for advice on data analysis, W. Zipfel and A. Singh (Cornell University) for help in understanding the optics of the GAIIX, A. Rizzi, C.T. Waters and H. Kwak (Cornell University) for bioinformatics advice, the Cornell Sequencing Core Facility for help in running the GAIIX, and H. Craighead (Cornell University) and the members of the Lis lab for helpful discussions on experimental design and the manuscript. This work was supported by US National Institutes of Health grants GM090320 and DA030329 to J.T.L.

REFERENCES

1. Amaral PP, Dinger ME, Mercer TR, Mattick JS. The eukaryotic genome as an RNA machine. *Science*. 2008; 319:1787–1789. [PubMed: 18369136]
2. Kruger K, et al. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*. 1982; 31:147–157. [PubMed: 6297745]
3. Prody GA, Bakos JT, Buzayan JM, Schneider IR, Bruening G. Autolytic processing of dimeric plant virus satellite RNA. *Science*. 1986; 231:1577–1580. [PubMed: 17833317]
4. Staley JP, Woolford JL Jr. Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines. *Current opinion in cell biology*. 2009; 21:109–118. [PubMed: 19167202]
5. Baltz AG, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell*. 2012; 46:674–690. [PubMed: 22681889]
6. Castello A, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*. 2012; 149:1393–1406. [PubMed: 22658674]
7. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics*. 2012; 13:840–852.
8. König J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. *Nature reviews. Genetics*. 2011; 13:77–83.
9. Ray D, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature biotechnology*. 2009; 27:667–670.
10. Ray D, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*. 2013; 499:172–177. [PubMed: 23846655]
11. Berger MF, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*. 2006; 24:1429–1435.
12. Cho M, et al. Quantitative selection and parallel characterization of aptamers. *Proceedings of the National Academy of Sciences of the United States of America*. 2013; 110:18460–18465. [PubMed: 24167271]
13. Nutiu R, et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology*. 2011; 29:659–664.
14. Tome JM, et al. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nature methods*. 2014; 11:683–688. [PubMed: 24809628]

15. Ellinger T, Ehricht R. Single-step purification of T7 RNA polymerase with a 6-histidine tag. *BioTechniques*. 1998; 24:718–720. [PubMed: 9591113]
16. Guajardo R, Sousa R. Characterization of the effects of *Escherichia coli* replication terminator protein (Tus) on transcription reveals dynamic nature of the tus block to transcription complex progression. *Nucleic acids research*. 1999; 27:2814–2824. [PubMed: 10373601]
17. Mohanty BK, Sahoo T, Bastia D. Mechanistic studies on the impact of transcription on sequence-specific termination of DNA replication and vice versa. *The Journal of biological chemistry*. 1998; 273:3051–3059. [PubMed: 9446621]
18. Pagano JM, et al. Defining NELF-E RNA binding in HIV-1 and promoter-proximal pause regions. *PLoS genetics*. 2014; 10:e1004090. [PubMed: 24453987]
19. Shui B, et al. RNA aptamers that functionally interact with green fluorescent protein and its derivatives. *Nucleic acids research*. 2012; 40:e39. [PubMed: 22189104]
20. Shaner NC, et al. Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature biotechnology*. 2004; 22:1567–1572.
21. Buenrostro JD, et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature biotechnology*. 2014; 32:562–568.
22. Lambert N, et al. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular cell*. 2014; 54:887–900. [PubMed: 24837674]
23. Fujita K, Silver J. Surprising lability of biotin-streptavidin bond during transcription of biotinylated DNA bound to paramagnetic streptavidin beads. *BioTechniques*. 1993; 14:608–617. [PubMed: 7682819]
24. Johansson HE, et al. A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein. *Proceedings of the National Academy of Sciences of the United States of America*. 1998; 95:9244–9249. [PubMed: 9689065]
25. Gravina MT, Lin JH, Levine SS. Lane-by-lane sequencing using Illumina's Genome Analyzer II. *BioTechniques*. 2013; 54:265–269. [PubMed: 23662897]
26. Dean KM, Palmer AE. Advances in fluorescence labeling strategies for dynamic cellular imaging. *Nature chemical biology*. 2014; 10:512–523. [PubMed: 24937069]
27. Shi X, et al. Quantitative fluorescence labeling of aldehyde-tagged proteins for single-molecule imaging. *Nature methods*. 2012; 9:499–503. [PubMed: 22466795]
28. Pluthero FG. Rapid purification of high-activity Taq DNA polymerase. *Nucleic acids research*. 1993; 21:4850–4851. [PubMed: 8233838]
29. Wang Y, et al. A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic acids research*. 2004; 32:1197–1207. [PubMed: 14973201]
30. McCullum EO, Williams BA, Zhang J, Chaput JC. Random mutagenesis by error-prone PCR. *Methods in molecular biology*. 2010; 634:103–109. [PubMed: 20676978]
31. Sambrook J, Russell DW. Preparation of denaturing polyacrylamide gels. *CSH protocols*. 2006; 2006
32. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature protocols*. 2007; 2:1849–1861. [PubMed: 17703195]
33. Lee EH, Kornberg A, Hidaka M, Kobayashi T, Horiuchi T. *Escherichia coli* replication termination protein impedes the action of helicases. *Proceedings of the National Academy of Sciences of the United States of America*. 1989; 86:9104–9108. [PubMed: 2556700]
34. Kornberg RD. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104:12955–12961. [PubMed: 17670940]
35. Steitz TA. The structural changes of T7 RNA polymerase from transcription initiation to elongation. *Current opinion in structural biology*. 2009; 19:683–690. [PubMed: 19811903]
36. Lee EH, Kornberg A. Features of replication fork blockage by the *Escherichia coli* terminus-binding protein. *The Journal of biological chemistry*. 1992; 267:8778–8784. [PubMed: 1533620]
37. Katsamba PS, Park S, Laird-Offringa IA. Kinetic studies of RNA-protein interactions using surface plasmon resonance. *Methods*. 2002; 26:95–104. [PubMed: 12054886]

38. Hall KB, Kranz JK. Nitrocellulose filter binding for determination of dissociation constants. *Methods in molecular biology*. 1999; 118:105–114. [PubMed: 10549518]
39. Rio DC. Filter-binding assay for analysis of RNA-protein interactions. *Cold Spring Harbor protocols*. 2012; 2012:1078–1081. [PubMed: 23028069]
40. Salim NN, Feig AL. Isothermal titration calorimetry of RNA. *Methods*. 2009; 47:198–205. [PubMed: 18835447]
41. Moore DD. Commonly used reagents and equipment. *Current protocols in molecular biology* / edited by Frederick M. Ausubel ... [et al.]. 2001 Appendix 2, Appendix 2.
42. Sambrook, J.; Russell, DW. *Molecular cloning : a laboratory manual*. 3rd. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2001.
43. Sambrook J, Russell DW. SDS-Polyacrylamide Gel Electrophoresis of Proteins. *CSH protocols*. 2006; 2006
44. Fairbanks G, Steck TL, Wallach DF. Electrophoretic analysis of the major polypeptides of the human erythrocyte membrane. *Biochemistry*. 1971; 10:2606–2617. [PubMed: 4326772]
45. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry*. 1976; 72:248–254. [PubMed: 942051]

Figure 1.

Overview of High Throughput Sequencing-RNA Affinity Profiling (HiTS-RAP). (A) First, recombinant proteins and DNA template need to be generated. A DNA template suitable for Illumina sequencing and presentation of RNA transcripts on the flowcell is required (see Fig. 2). Recombinant Tus protein and fluorescently labeled protein of interest (either as an mOrange fusion or a comparable dye-labeled form) are required for transcription halting and detection of bound protein, respectively. As an optional step, testing of the recombinant proteins and the DNA template for proper function, and detectable interaction between the

protein of interest and the RNA transcript of the template DNA is recommended before performing HiTS-RAP. The multi-step process of HiTS-RAP (see Fig. 3) is performed on an Illumina GAIIx instrument yielding sequence and protein binding information for about 200 million RNAs on a single flowcell. Finally, HiTS-RAP generated sequencing and binding data are analyzed to determine the protein binding affinity (K_d) of each RNA presented on the flowcell (see Fig.4). A timeline for HiTS-RAP is provided, where generous estimates of time that each step would take for a new user are indicated on the left side. The whole process can be completed in 2–3 weeks. (B) An Illumina Genome Analyzer IIX (GAIIx) equipped with Paired-End Module (PEM) and controlled by a Dell T7500 workstation, is used for HiTS-RAP. The main function of each component is indicated below the schematics with green text. At its core, the GAIIx is both an automated microfluidic device, which can precisely control amounts, incubation times, and temperature of reagents delivered to the flowcell (red parallelogram), and a sensitive Total Internal Reflection Fluorescence (TIRF) microscope, which can quantitatively detect multiple fluorescent entities on the surface of the flowcell. The GAIIx is easily programmable. These features make the GAIIx an attractive instrument to repurpose for applications other than sequencing, as has been done in HiTS-FLIP, RNA-MaP, and HiTS-RAP. Delivery of reagents to the flowcell inside the instrument is indicated by a cyan arrow. HiTS-RAP is fully automated; an unmodified GAIIx programmed with a single user-edited recipe performs every step of HiTS-RAP sequentially, and requires minimal user input once the instrument is set up (see Fig. 3)

Figure 2.

DNA template and RNA transcript of HiTS-RAP. (A) Schematics of the DNA template used for HiTS-RAP and the resulting halted RNA transcript. DNA template encoding the RNA of interest (green) is flanked by the Illumina flowcell adaptor 1 (gray) and T7 RNA polymerase promoter (orange) upstream, and by the Illumina sequencing primer annealing site (purple), Tus-binding Ter site (red), and Illumina flowcell adaptor 2 downstream. Illumina flowcell adaptors 1 and 2 are required for cluster generation on Illumina GA flowcell. T7 RNA polymerase promoter is required for transcription of the RNA of interest. The Illumina sequencing primer is used for sequencing the DNA template of the RNA of interest and serves as a docking site for the T7 RNA polymerase when it is halted. Tus protein binds to Ter site and halts the transcribing RNA polymerase. Direction of transcription and sequencing are indicated by orange and purple arrows, respectively. Tus-bound Ter site that is non-permissive (halting) and permissive (read-through) to RNA polymerase are indicated by solid and open red triangles, respectively. The halted RNA transcript includes a triplet G

derived from the T7 promoter, followed by the RNA of interest and some of the Illumina sequencing primer. The 3'-end of RNA transcript, indicated by a dashed line, is inaccessible. (B) Construction of DNA templates for HiTS-RAP. DNA template is constructed by PCR in two steps using 2 sets of nested oligos. Forward oligos introduce T7 promoter (step 1) and Illumina flowcell adaptor (step 2), whereas reverse oligos introduce Illumina sequencing primer (step 1), and Ter site and Illumina flowcell adaptor 2 (step 2). (C) Sequence of the HiTS-RAP DNA template for GFP aptamer. GFP aptamer encoding sequence is in green, and the rest of the sequences are colored as in (A). The transcription start site is indicated by +1 and a broken arrow. (D) Sequence of the halted RNA transcript from GFP aptamer template. Sequences are colored as in (C). Uppercase indicates the region of the halted RNA transcript that is accessible, and the lowercase indicates a region that is likely to be buried in T7 RNA polymerase and thus inaccessible by other proteins.

Figure 3.

Schematics of HiTS-RAP. (i) The HiTS-RAP DNA template, colored as in Fig. 2, is amplified on the cBot instrument to generate DNA clusters on Illumina flowcell. (ii) HiTS-RAP starts with sequencing of DNA clusters on the flowcell by the GAIIX. (iii) After sequencing, clean, full-length DNA template is regenerated, and (iv) bound by Tus protein at the Ter site. (v) DNA is then transcribed by T7 RNA polymerase, which becomes halted at the Tus-bound Ter site, retaining the RNA transcript on the DNA template. (vi) Finally, increasing concentrations of the target protein are incubated with the halted RNA transcripts,

and the bound protein is monitored via the fluorescence of mOrange fusion protein. Once the assay is finished, the protein binding data is used to calculate the binding affinities that are matched with the sequence of each cluster, yielding K_d measurements for millions of RNAs per lane of the Illumina flowcell. Steps that constitute HiTS-RAP (iii-vii) are indicated with a gray colored box. Critical actions that require user input such as loading reagents and modifying file settings are indicated with a blue open triangle and italicized text.

Figure 4.

Workflow of HiTS-RAP data analysis. First, binary sequence information stored in tile specific .bcl files are read and converted to .qseq file by Illumina's OffLineBaseCaller (OLB). Sequence information in .qseq files are combined by .qseq combiner to generate a single .txt file containing all sequences for a single lane on flowcell (120 tiles). Fluorescence intensity information for sequencing and protein binding steps stored in .cif files are read and converted to a single .txt file by Robin Freedman's .cif reader Perl wrapper script. Sequence and intensity data matched by coordinates are combined into a single .txt file by coordinate matcher. Later, sequence matcher combines reads with identical sequence and calculate average sequence and protein binding intensities. Finally, curve fitting is performed to obtain a .txt file containing all unique sequences with the calculated binding affinity and the fitting parameters. Name of software/scripts used for each step are indicated in gray italicized text.

Figure 5.

Examples of protein binding curves from HiTS-RAP. (A) A set of protein binding curves from a successful HiTS-RAP run. These RNAs represent a range of K_d s (8–160 nM) and copy numbers (N , 13–46,757). In all panels (a–d), protein binding intensities are the average of all clusters with a given sequence in one lane, and are normalized by the average sequencing intensity. Error bars represent s.e.m. Error of the K_d is the square root of the variance returned by the analysis pipeline. (B) Protein binding curves which only have one data point above background. A fit for the K_d can be found, but are likely unreliable. (C) Binding curves with bad fits. In both of these cases, the fitted base is above the maximum, so these measurements should be disregarded. (D) Binding curves from an experiment where no protein binding is observed. These data are from an experiment where one or more biochemical steps leading up to the protein binding have failed.

Table 1

Comparison of HiTS-RAP and RNA-MaP. Major distinguishing (highlighted with gray background) and minor technical differences between these two methods are listed. Many of the differences can be readily implemented in either method.

	HiTS-RAP	RNA-MaP
Instrument	Unmodified Illumina GAIIx	Modified Illumina GAIIx
RNA polymerase	T7 RNA polymerase	E.coli RNA polymerase
RNA tethering strategy (transcription reaction)	Halting by Ter-bound Tus protein (multi-round transcription)	Halting by Biotin-bound Streptavidin protein (single-round transcription)
Control of sequencing and protein binding	Single automated run	Separate sequencing and manual protein binding runs
Protein binding imaging and data extraction	Illumina software for fluorescence intensity measurement and data extraction with custom software	Custom software for independent analysis of saved images with temporal resolution
Protein binding	Binding affinity (K_d)	Binding affinity (K_d) and binding kinetics (on- and off-rates)
Protein labeling	mOrange fluorescent protein fusion	Surface 549 dye labeling via SNAP-tag fusion
Template barcoding	None	Random barcoding with bottleneck PCR
RNA quantification	Indirect, estimation from DNA template	Direct, via hybridization of fluorescent probe

Table 2

Oligos used for HiTS-RAP analysis of GFP Aptamer. Name, sequence, and the usage of each oligo are listed. See Figure 1 for more detail.

Name	Sequence (5' – 3')	Use (Step #)
GFP_Temp_DNA_FOR	GATAATACGACTCACTATAGGGAATGGATCCACA TCTACGAATTCAGCTTCTGGACTGCGATGGGAG	PCR (Step 17)
Temp_DNA_REV	TACACTCTTTCCTACACGACGCTCTTCCGATCTG CTTCTGTGGTGGCCCTCTTTAAGG	PCR (Step 17)
Illum_IllumAdaptor_T7	CAAGCAGAAGACGGCATACGAGATCGGTGATAA TACGACTCACTATAGGGAATGGATCC	PCR (Step 22)
IllumFORAdapt_T1_Illum FORSeq	AATGATACGGCGACCACCGAGATCTCATGACGTG ACTTTAGTTACAACATACTAATTTACACTCTTTCC CTACACGACGCTCTTCCGAT	PCR (Step 22) & dsDNA regeneration (Step 67)