Defining protein variant functions using high-complexity mutagenesis libraries and enhanced mutant detection software ASMv1.0

Xiaoping Yang[1,9], Andrew L. Hong[2,5,6,9], Ted Sharpe[3,9], Andrew O. Giacomelli[2,7,9,10], Robert E. Lintner[4,11], Douglas Alan[1], Thomas Green[1], Tikvah K. Hayes[2,7], Federica Piccioni[1,12], Briana Fritchman[1], Hinako Kawabe[1], Edith Sawyer[1], Luke Sprenkle[1], Benjamin P. Lee[5,6], Nicole S. Persky[1], Adam Brown[1], Heidi Greulich[2,7], Andrew J. Aguirre[2,7], Matthew Meyerson[2,7,8], William C. Hahn[2,7], Cory M. Johannessen[2,13], and David E. Root[1,*]

Affiliations:

[1]Genetic Perturbation Platform, [2]Cancer Program, [3]Data Science Platform, and [4]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[5]Department of Pediatrics, Emory University, Atlanta, GA 30322, USA

[6]Aflac Center for Cancer and Blood Disorders, Children's Healthcare of Atlanta, Atlanta, GA 30322, USA

[7]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[8]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[9]These authors contributed equally

[10]Current Address: Ontario Cancer Princess Margaret Cancer Centre, 610 University Avenue, Toronto, ON, M5G 2M9 Canada

[11]Current Address: RAN Biotechnologies, 100 Cummings Center, Ste 434J, Beverly, MA 01915, USA

[12]Current address: Merck Research Laboratories, 33 Avenue Louis Pasteur, Boston, MA 02115, USA

[13]Current address: Novartis Institutes for Biomedical Research, Oncology Innovative Targets and Technologies, 250 Massachusetts Ave, Cambridge, MA 02139, USA

Correspondence should be addressed to D.E.R. (droot@broadinstitute.org)

Abstract

Pooled variant expression libraries can test thousands of variants of a gene in a single multiplexed experiment. In a library encoding all single-amino-acid substitutions of a protein, each variant differs from its reference only at a single codon-position located anywhere along the coding sequence, such that identifying these variants by sequencing to read out the outcomes of the screen is a major technical challenge. A popular but expensive brute-force approach is to divide the pool of variants into much smaller sub-libraries in each of which only a small region is mutated to allow direct sequencing readout of variant abundance. Here we present an approach to screen very large variant libraries with mutations spanning a wide region in a single pool, including library design criteria and mutant-detection algorithms that permits reliable calling and counting of variants from large-scale sequencing data.

Keywords: saturation mutagenesis, deep mutational scanning (DMS), variants, open reading frame (ORF), log-fold change (LFC), base miscall, DNA oligonucleotide synthesis error, DNA polymerase error, next-generation sequencing (NGS).

Background

A pooled open reading frame (ORF) variant expression library for a gene of interest can be screened for cellular phenotypes, providing a powerful means to systematically characterize the functions of thousands of single-amino-acid substitutions in a single experiment. Advances in DNA oligonucleotide and gene synthesis [1] and in next-generation sequencing [2] have enabled production and screening of large variant pools. Pools encompassing all possible single-amino-acid changes for every amino acid position of a protein are known as saturation mutagenesis or deep mutational scanning (DMS) libraries. Screening a DMS library in cells can map each variant to phenotype in a single experiment.

As illustrated in **Figure 1**, a typical pooled screen allows one to test the functions of thousands of genetic perturbation agents in a single experiment. In general, cell-based pooled screens rely on introduction of one genetic perturbation cassette, e.g., a CRISPR sgRNA, an shRNA, or an ORF variant, per cell into the genome using delivery by a retrovirus, so that a single perturbation cassette is inserted into genomic DNA of each cell. The perturbed cell population is then subjected to selection pressure, for example, treatment with a drug, flow-sorting for reporter expression, or simply growth in culture. Post-selection cell samples are collected and processed to measure the abundance of library members in each sample by sequencing. Enrichment or depletion of each library perturbation in samples that have undergone selection versus samples under reference conditions (e.g., pre-selection or vehicle treatment samples) reveals the impact of the perturbations. For CRISPR screens, the functional elements themselves, the sgRNAs, serve as ideal short unambiguous barcodes to use for screen "deconvolution", the measurement of enriched and depleted members of the library. The deconvolution of CRISPR libraries is a direct-readout and software-free process of counting

sequencing reads. In contrast, ORF variant library screens are confronted with greater challenges for variant detection.

The complexity of a saturated ORF variant library is a function of the ORF length and the number of desired variants per codon position. As illustrated in **Supplementary Figure 1**, at each amino acid position, the template amino acid is typically mutated to 19 other amino acids. A stop codon, and when possible synonymous codons that can serve as LOF and negative controls respectively (**Methods**). Factoring in the amino acid positions to be mutagenized (N), the library is a collection of approximately 20*N variants, or alleles, each differing from the template ORF at a single amino acid position. Capturing differences among the variant-defining signatures as subtle as a single codon variation located anywhere along the full-length ORF sequence is a significant technical challenge. A popular approach has been to employ multiple sub-libraries, each with short mutated regions with lengths of 25 [3, 4], then 47 [5], 100 [6], and recently as many as 170 [7] amino acids. These sub-libraries were screened and sequenced separately. Each of these sub-libraries covers a mutated region short enough to be covered by a single next-generation sequencing read for direct readout of variants. While practical for small ORFs, for longer genes this approach requires an unwieldy number of libraries and screens. Our methodology, on the other hand, permits screening of very large pools in a single experiment using robust variant-calling tools that can accurately quantify variants from sequencing data generated by amplifying and randomly shearing the entire ORF. A single saturated variant library is thus adequate a wide range of ORF lengths, e.g., ~600-bp *KRAS* [8], ~1200-bp *TP53* [9] and *SMARCB1*, ~1400-bp catalytic region of *PDE3A* [10], ~1700-bp *SHOC2* [11], ~1800-bp *PPM1D* [12], ~2200-bp *EZH2*, ~3600-bp *EGFR*) (**Fig. 1b**). Here, we present methods and strategies developed in the process designing, constructing and screening more than 20 variant

libraries (**Supplementary Table 1**). We have made available our library designer, variant-calling software, and data analysis tools (**Supplementary Table 2)**. We anticipate that these tools and methods will facilitate variant screening for longer ORFs.

Results

Variant calling with ASMv1.0

The challenge of calling variants correctly from PCR-amplified, fragmented, and sequenced genomic DNA is due to the multiple factors that lead to unexpected or misattributed variant sequences. In addition to the intended, library-designed codon sequence modifications versus the reference sequence, each ORF molecule in our library is subject to two kinds of unintended changes: oligonucleotide chemical synthesis errors and DNA polymerase errors. Although the detailed synthesis strategies can vary, ORF variant libraries are generally all produced using chemically synthesized oligonucleotide strands that encode the variant-defining codon modifications and that constitute part of the final full ORF library sequences together with unmodified reference (often “wild-type”) portions of the ORF that are enzymatically synthesized by high-fidelity DNA polymerases in a PCR process. To fully characterize the composition of the initial library, and of the variant distributions from the genomic DNA of screened cells, it is desirable to first detect empirically all variation of sequence from the reference sequence, without consideration the intended library design. In general, these variations will include nucleotide substitutions, insertions, deletions and combinations thereof. Some variations exactly match design-intended variants, while others carry unexpected nucleotide changes in addition to, or instead of, the expected variation. Aside from the initial library production process, the process of extracting and sequencing ORF molecules from the library pDNA or the genomic

DNA of screened cells, often called “library deconvolution”, is itself subject to errors from imperfect PCR amplification, and from sequencing miscalls (e.g., a Phred Q20 base-call has a 99% chance of being correct). Our enhanced variant-detection software aimed to distinguish the intended variants from those arising from errors of synthesis of the initial library or from the library “deconvolution” process, and to assess library quality and deconvolution miscall rates.

Improving upon our earlier version of variant-calling strategy and the associated software (ORFCallv1.0) cite ref, we developed Analyze Saturation Mutagenesis v1.0 (ASMv1.0; see **Methods** and **Supplementary Fig. 2**) to more accurately interpret pooled variant NGS data and measure variant abundance. ASMv1.0 is now included in the Genome Analysis Toolkit (GATK v4.2.0.0). NGS reads are each first aligned to the reference ORF to produce bam files which are the input to ASMv1.0. ASMv1.0 then performs all the steps required from per-sample read/reference alignments to produce a list of observed nucleotide sequence variations in relation to the reference ORF and the abundance of each of these nucleotide sequence variations for each sample. Notably, the list of observed nucleotide sequence variations is independent of any prior expectations of the library’s variant composition. As shown in **Supplementary Figure 2** , the key steps in this process are: (1) trimming reads of flanking sequences introduced during fragmentation and barcoding of the gDNA, (2) combining overlapping paired-end reads when possible to obtain a “per-molecule” single sequence result per read pair, (3) applying base quality, minimum read length and other filters prior to variant calling, (4) identifying bona fide sequence variation, if any, for each read or read pair versus the reference (substitutions, insertions, deletions), and (5) counting examples of each set of observed nucleotide sequence variations to quantify the abundance of each sequence variation set in each sample. Because the NGS reads are far shorter than the full-length ORF and because each planned variant in the library typically

carries only a single codon change from the template ORF, most observed aligned reads will fully match the reference sequence. These perfectly reference-matching reads are used only to assess the depth of read coverage across the reference ORF. The abundance of reads that contain mutations gives information on the identity and abundance of each variant defined by those mutations.

As noted, the output from ASMv1.0 identifies a list of empirically observed variant-defining sequences and their abundance in each sample. Because this algorithm identifies variants in an entirely data-driven manner that does not rely on expectations from the library design, the output can be used not only to filter for the results for designed ORF variants (i.e. the intended variants), but also to assess the library quality and characterize its actual composition, which will always differ to some extent from the intended design.

The enhanced performance of ASMv1.0

ASMv1.0 incorporates several enhancements versus an earlier variant-detection process, implemented in the software package ORFCallv1.0. ORFCallv1.0 is a simple and effective variant detection tool designed strictly for single amino acid substitution variant libraries, and was employed to interpret a number of saturation mutagenesis screen results [9, 13-15]. ORFCallv1.0 quantifies variant abundance by first aligning reads to the reference sequence, then recording, at each codon position, the count of each of 64 possible codons. ORFCallv1.0 tallies the codon read counts at each amino acid position by staying within the 3-nucleotide codon window and ignoring any additional changes located in the same read. If a library has high rate of synthesis errors (e.g., due to oligonucleotide chemical synthesis errors), these errors will lead to high frequency of unintended changes that are combined, in the same molecule, with an

intended codon change. Ignoring the presence of these unintended changes in the manner of the ORFCallv1.0 algorithm can be costly when these additional nucleotide changes produce phenotypes that differ from the phenotype of the neighboring intended variant. As an improvement, our new software, ASMv1.0, calls variants in the context of full-length read pairs. A called variant is defined by incorporate all nucleotide variation observed between the read pair and the reference sequence. These nucleotide variations include substitutions, insertions, deletions, or the combinations and multiples of them. ASMv1.0 allows one to separate pure variants that bear only the intended sequence changes from those that also harbor extra changes. **Supplementary Table 3** summarizes the enhanced functionalities from ORFCallv1.0 to ASMv1.0.

We used the screens performed in Giacomelli *et al* [9] to compare the two variant-detecting methods. Ectopically expressed *TP53* variants were screened in isogenic $p53^{WT}$ and $p53^{NULL}$ A549 lung cancer cells, with selection pressure from either nutlin-3 or etoposide. Under nutlin-3 treatment, the screen of $p53^{NULL}$ cells enriched *TP53* loss-of-function (LOF) variants and depleted wild-type-like variants (**Fig. 2**). Under etoposide treatment, the $p53^{NULL}$ cells enriched wild-type-like *TP53* variants and depleted the *TP53* LOF variants (**Supplementary Fig. 3**). Under nutlin-3 treatment, the screen of $p53^{WT}$ cells enriched *TP53* variants exhibiting dominant-negative effects (DNE) (**Supplementary Fig. 4**). Using both the ORFCallv1.0 and ASMv1.0 variant-calling methods, we reprocessed the next-generation sequencing data from these three screens. The phenotypic strength of each variant, measured as log-fold change (LFC) in the pre- and post-treatment samples, produced similar heatmaps, indicating that the two variant-calling algorithms are mostly consistent with each other (**Fig. 2a, Supplementary Figs. 3a, 4a**). However, we also noted that a small subset of variants (~3% of the library) exhibited

noticeable differences in LFC calculated using these two algorithms; we define the differences as delta(LFC)=LFC^{ASM}-$LFC^{ORFCall}$ in **Figure 2b, Supplementary Figure 3b** and **4b.**

In a perfect library that consists solely of intended variants, ORFCallv1.0 and ASMv1.0 should be equal in performance. However, owing to errors introduced during library synthesis, such as oligonucleotide synthesis errors and DNA polymerase errors incurred in the production of the ORF variant materials, it is inevitable that planned codon changes are sometimes physically associated with one or more unintended accidental nucleotide changes in the same molecule. We define the 'variant purity' of a planned variant as the ratio of molecules that exhibit only the intended mutation to all molecules that harbor the codon change of interest, regardless of whether they also harbor additional nucleotide changes. When the purity of a planned variant is high, both versions of the software will call variants with similar accuracy. However, when the variant purity is low, the enrichment and depletion scores measured by ORFCallv1.0 can be misleading. We computed the purity of every planned variant using the reference samples (samples without selection pressure) of the Giacomelli screens, and for each purity bin we computed the Pearson correlation coefficient of two sets of LFCs, one using ORFCallv1.0, the other using ASMv1.0 (**Fig. 2c, Supplementary Figs. 3, 4)**. The two methods correlate well when the variant purity is high and diverge as the variant purity decreases.

Work by Kotler *et al* [5] offered an additional way of testing the performance of ASMv1.0 and ORFCallv1.0. We compared the Giacomelli screen LFC scores called by both versions of the software with the results from the Kotler *et al* [5], which used a smaller variant library that was focused on the p53 DNA binding domain (DBD) and screened in $p53^{NULL}$ H1299 cells. Kotler *et al* targeted p53 DBD with four sub-libraries, each covering a 47-amino acid region of the ORF. The four sub-libraries were screened separately, and the screen samples

were processed by direct sequencing readout of the mutational region of each of the 4 libraries. Like the Giacomelli p53^{NULL} A549/nutlin-3 screen, the Kotler p53^{NULL} H1299 cell screen enriched for *TP53* alleles that exhibit LOF and depleted wild-type-like variants. From the 7890-variant Giacomelli p53^{NULL} A549/nutlin-3 screen data, we extracted a 2964-variant subset that covered the same amino acid mutational space as the Kotler H1299 screen. These 2964 Kotler scores and Giacomelli scores, using either ASMv1.0 or ORFCallv1.0, are exhibited good correlation (Pearson coefficient 0.65 (p-value=0), and 0.62 (p-value=2.4e-317) respectively), despite the fact that the screens were performed using different variant detection methods, different libraries, different cell lines, and different conditions consisting in one case of simple growth in culture and in the other nutlin-3 treatment (**Supplementary Fig. 5**). Interestingly, when focusing on p53 DNA binding domain, both Kotler and Giacomelli data exhibit a bimodal pattern representing the enriched p53 loss-of-function variants (**Supplementary Fig. 5a,** upper right quadrant) and the depleted p53 wild-type-like variants (**Supplementary Fig. 5a,** lower left quadrant). There is therefore a natural cutoff to classify each of the 2964 variants as having either LOF or WT-like activity. Consistent with the high similarity noted above between the ASMv1.0 and ORFCallv1.0 scoring for most variants, the agreement between Kotler and Giacomelli classifications is similar whether using the ASMv1.0 method (2366 out of 2964 variants, 80% agreement) or the ORFCallv1.0 method (2314 out of 2964 variants, 78% agreement). This is to be expected for a library that has mostly high purity variants for which ASMv1.0 and ORFCallv1.0 show high agreement. However, focusing on the small subset of 76 variants that ASMv1.0 and ORFCallv1.0 scored very differently in the Giacomelli data (abs(delta(LFC))>1, **Fig. 2d, right**), the ASMv1.0 classification agrees with the Kotler

classification much better (65 out of 76 variants, 86%) than the ORFCallv1.0 classification does (36 out of 76 variants, 47%, **Fig. 2e**).

ASMv1.0-enabled analysis reveals that the secondary mutations (errors) in the library are near the targeted mutation and introduced by errors in oligo synthesis.

The ability of ASMv1.0 to detect molecules with intended or unintended changes along the entire read pair allows in-depth analyses of ORF variant library composition. Thus, in addition to the variant purity discussed earlier, we can also tally and characterize the unintended nucleotide changes to profile the errors introduced during library construction. For example, we were able to determine that in the *TP53* library, 84% of molecules are pure and planned variants (**Fig. 3a, left**), and the errors in the library are predominantly single nucleotide changes.

Using ASMv1.0 output, we investigated the proximity distribution of errors in relation to the intended codon change. In **Figure 3**, we compared libraries made by two different methods: a *TP53* library made by MITE (Mutagenesis by Integrated TilEs) [15], and a *PDE3A* library synthesized by Twist Bioscience [10]. The MITE method used chemically synthesized oligonucleotide strands to encode 90-nucleotide variant 'tiles' that were subsequently incorporated with the rest ORF segments to form the full-length ORF sequence; these ORF segments were prepared by high-fidelity DNA polymerase used in PCR. The Twist Bioscience technology involved the chemical synthesis of oligonucleotides varying in length between 30-50 nucleotides depending on GC content with the variant-encoding bases in the middle of the oligonucleotide; these oligonucleotides were used as the PCR primers and became part of the ORF sequences of the library. In **Figure 3b,** we profiled the error rate by the number of bases between the locations of the error and the intended codon change. This distance profile yields an

important observation that have shaped our screen deconvolution sequencing strategy, which will be discussed in detail later in this section:

We observed that the stretches of more-frequent errors in libraries produced by the two methods, ~90 bases for the MITE method and ~30 bases for the Twist Bioscience method, correspond to the length of chemically synthesized DNA oligonucleotides involved in library production. Since most of the errors in an ORF variant are concentrated near the intended codon change, sequencing read as short as 100 bases can detect most of the unintended changes that accompany each intended change.

It is important to note that all nucleotides in variant ORF libraries come from two sources: (1) chemically synthesized oligonucleotides that encode and introduce the mutations (e.g. PCR primers or oligonucleotide tiles that encode variants), or (2) nucleotides copied from the reference template by high-fidelity DNA polymerase. The errors in these libraries are therefore expected to arise predominantly from oligonucleotide chemical synthesis errors (~2 errors per kb), since DNA polymerase (Q5, Fushion, or Ex Taq) nucleotide misincorporation rates are far lower [16]. A high-fidelity DNA polymerase with an error rate of 10^{-6} errors per cycle per amplified nucleotide will produce only a small number of erroneous molecules. For example, for a 1kb-ORF, after amplification with 12 PCR cycles, 1.1% of ORF molecules may contain an error due to DNA polymerase misincorporation. In contrast, oligonucleotide chemical synthesis errors contribute to far more ORF molecules with errors, 10% by Twist Bioscience method and 18% by MITE method (**Supplementary Table 4**). In this case, NGS reads that is of 100 bases or longer and carry expected mutations, which mark those reads as those spanning chemically synthesized nucleotides, will effectively capture the most of the unintended nucleotide changes in the library.

ASMv1.0-enabled analysis allows assessing the effect of miscalls (artifactual errors) introduced by processes involving PCR.

The library synthesis errors discussed above are actually present in the libraries and screened in cells to which the libraries are applied. It is important also to consider library assessment errors in the form of miscalls introduced by the variant quantification process itself. To deconvolute the variant abundances from a pooled library, PCR was used to replicate the ORF DNA from the genomic DNA of the library-transduced cells (up to 30 PCR cycles) and was used again in to barcode samples in preparation for NGS (12 cycles). We define 'miscalls' as the artifactual base calls that do not correctly represent the sequence that was actually present in the samples. The miscall rates are contributed by the error rate of DNA polymerases used in PCR processes involved in sequencing sample preparation, plus the error rate in sequencing base calls as defined by Phred Q scores. To directly assess the rate of miscalls in the deconvolution of a variant library, we processed and sequenced a clonal pure reference *TP53* ORF sample that should therefore produce only reads that perfectly match the reference sequence. We observed mutations – miscalls - at a rate of ~0.0003-0.0006 per nucleotide (**Fig. 4a, right**), 3-6 miscalls per 10,000 sequenced bases.

This miscall rate is large enough to be a significant factor in variant detection for large complex libraries. As an example, illustrated in **Supplementary Figure 6**, miscalls lead to (1) inflated counts of variants with codons that are one-nucleotide away from the template codon (we refer to this variant type as '1-nt delta' codon), and (2) as a result of these inflated counts, artifactually narrowed log-fold change range between samples of 1-nt delta codons (**Supplementary Fig. 6,** legend**)**. Indeed, in the *TP53* library screen with $p53^{NULL}$ cells/nutlin-3,

we observed the inflation of counts (**Fig. 4b**), and the suppression in fold-change (**Fig. 4c**), in the 1-nt delta variant group relative to 2- or 3-nt delta variants. This effect of PCR errors on 1-nt delta codon changes has been observed in every library we have processed (~20 to date), and the effect increases as ORF length increases (data not shown).

The varying effects of PCR/NGS miscalls on the quantification of variants reshaped our strategy in library design rules and adding control experiments. First, we avoided the use of codons that are one nucleotide away from the reference codons whenever possible. For a typical ORF sequence, to achieve all possible single site amino acid substitutions it is sometimes necessary to use codons deviate from the reference codon at that position by only one nucleotide, but the use of 1-nt delta codons, can be reduced to approximately 3% of the library members. Secondly, to better interpret the results of the 3% 1-nt delta variants, we added a clonal plasmid sample that therefore contains only one uniform sequence and processed it as if it were a screening sample. These clonal template data may be used to identify the 1-nt delta variants that are most affected by miscalls, or even correct the screen sample data based on the background miscall rates (**Fig. 4b**, **right**).

Silent variants are used as a measure of screen baseline.

The inclusion of silent variants provides a set of controls that should serve as independently measured replicates of the reference ORF sequence. Like the missense- and nonsense-variant sequences, in order to reduce contamination by miscalls, these silent mutations should employ codons that are 2- or 3-nt different from the reference codon. Often, codons encoding an amino acid differ from one another at the wobble base. As a result, many silent variant codons are one nucleotide away from the template codon. Requiring that any admitted

silent variants have a 2- or 3-nt difference relative to the reference codon results in silent variants comprising about 2% of the library.

As demonstrated by the bimodal split of phenotypes in *TP53* DBD library screens discussed earlier, ORF variant library screens often produce many "hit" variants that deviate from the reference phenotype. As a result, in contrast to many other common pooled screening situations one cannot always rely on the large portion of the library producing no phenotype to serve as the baseline, making the silent codon change controls a particularly important reference-point for interpreting the screen results. The silent changes LFC results are displayed in the row marked by '**B**' in the heatmaps in **Figure 2a**.

In the *SMARCB1* library, we included nearly all possible silent variants, including 1-nt delta silent variants. This case provides an example of the consequences of miscalls of the reference sequence that masquerade as library variants. As expected, due to the miscalls discussed earlier, the apparent abundance of 1-nt delta silent variants is inflated relative to the abundance of 2- and 3-nt delta silent variants (**Supplementary Fig. 7a**). As a consequence, relative to 2- and 3-nt delta groups, the 1-nt delta silent variants show an artificially narrowed log-fold change range when comparing among samples, because every sample has the same contribution of counts from miscalls of sequence that is actually the reference sequence (**Supplementary Fig. 7b**). In our current library designer download (**Supplementary Table 2**), we avoid 1-nt delta silent variants and minimize 1-nt delta missense and nonsense variants.

A justification for using next-generation sequencing (NGS) over long-read sequencing (LRS).

An often-debated topic is how saturation mutagenesis screen samples should be sequenced. It would seem evident that screen samples should be sequenced with long reads to widen the space of variant characterization and to capture all mutations present on each molecule. However, the technology that can capture the readout of full-length ORF sequences [17, 18] is currently cost-prohibitive for most labs, and more importantly, undesirable for variant detection due to low concordance rate of base calls [19]. For example, in a study by Wenger *et al* [20], circular consensus sequencing (CCS), the best of its kind, has a concordance rate of 99.772%. This is equivalent to a Q26 Phred score, producing one discordance per 439-base readout. This discordance rate is higher than the error rate in commercial DNA oligonucleotide synthesis, the key contributor of the real errors in the libraries as discussed earlier. Since the discordance rate of the measurement method is higher than the underlying error rate which it is expected to measure, the CCS read attrition rate will be high. As a result, one may be forced to either throw out many reads that show secondary mutations or focus locally on the expected codon changes and ignore mutations in the distance.

Long-read sequencing technology is advancing rapidly and may change the ability to get long reads of adequate fidelity. In the recent work by Frank *et al* [21], long read sequencing was used to associate a set of unique external barcodes to each of ~7000 variants in a saturation variant library, so that the screen deconvolution work can be achieved by direct next-generation sequencing of the barcodes. However, there are several factors create a high bar for achieving long-read sequencing that can match the results obtained with NGS paired-end reads of 150 bases each with a Phred score of Q30 or more for each base call, which is comfortably achievable by many NGS platforms: (1) We have demonstrated that the errors in our libraries are

near the intended codon changes (**Figure 3, right**), and a 150-base read is adequate to detect both the variant-defining sequences and library errors from DNA oligonucleotide synthesis. We have shown that it is unnecessary to sequence an ORF variant far into its regions synthesized by high-fidelity DNA polymerase during library construction, because the errors introduced by DNA polymerase in regions outside of a 100-base window around the intended codon changes are rare enough to be inconsequential. ASMv1.0 output offers many choices for one to filter the results in various ways. A good example is that, here, we may filter and focus on only the reads that bear a library-designed codon change, which effectively represents the variant's error-bearing region, i.e., the region that was chemically synthesized. (2) We have demonstrated that PCR-introduced miscalls produce predominantly 1-nt delta codon changes. Because we minimize the use of 1-nt delta codon variants, the artifactual variants resulting from miscalls are mostly unexpected variants that were not part of the library design. In this case, ignoring reads carrying only unexpected variations largely removes the effects of sequencing miscalls. (3) Even with high base-quality threshold, NGS currently produces adequate read yield for sequencing screen samples of variant libraries. ASMv1.0 allows users to set adequate base quality threshold, for example, at Phred score Q37 with Illumina Novaseq NGS. With that filter, we can comfortably achieve <1 discordance per 5000-base NGS readout. As we focus on the reads that carry an intended-variant signature, at this low discordance rate, >94% (0.9998^300) of 150-base read pairs used to call variants are free of miscalls.

Discussion

Variant detection software ASMv1.0 implements in-depth data analyses that allowed us to explore, for instance, variant purity, the characterization of the types and rates of occurrence

of unintended species that were detected in the sequencing data by ASMv1.0; these unintended species include oligonucleotide chemical synthesis errors, which are real and part of the library, and sequencing miscalls, which are artifactual. Furthermore, ASMv1.0 offers choices for one to filter the data for variants that are expected by designs, or filter for high abundance variants independent of the library designs. The results of these analyses have informed the choice of strategies at multiple steps including library design, screen setup, and data analysis. We believe that the analytical principle and screen strategy presented here, along with our library designer, variant-calling software, and data analysis tools will help enable the scientific community to fully utilize this technology in research applications.

Although, as noted, producing the desired amino acid changes using 2 or 3-nt changes versus the reference sequence is always far preferable to using only 1-nt changes versus the reference sequence, it should be noted that the strong hits among the 1-nt delta variants may still score well, even with some suppression of the apparent log-fold change results due to miscalls of reference sequence that match the 1-nt variant. The effect of miscalls on 1-nt delta variant counts becomes less of an issue for short ORFs (e.g., *KRAS*). As illustrated in **Supplementary Figure 6**, the longer the ORF, the more copies of the reference codon will undergo PCR-based amplification and NGS, the more copies of each actual designed 1-nt variant will be also read out, and consequently the more counts of the 1-nt delta miscalls will occur per correct call of the 1-nt variant molecules truly present. To better interpret any 1-nt delta codons that in a variant library, we recommend adding the clonal template ORF as a control sample and processing it in parallel with the screening samples. This control sample may be used as a count-correction factor (as shown in **Fig. 4b**, right) or simply to flag those 1-nt delta variant sequence that occur at high

rates due to miscalls of the reference sequence, contaminating and therefore suppressing any fold-changes in that variant between screen arms.

The recent establishment of an open-source platform, MaveDB, to enable variant function data sharing, is a significant development [22] and is a good step towards more widespread sharing of saturation mutagenesis data, reagents and resources. A well-designed variant library should be useful for many different screens of that same genes, provided that the choice of reference sequence upon which the design is based is satisfactory for each application. For example, Boettcher *et al* [23], using the *TP53* site-saturation variant library created by Giacomelli *et al* [9], quickly demonstrated that the dominant-negative activity of *TP53* missense mutations in DNA binding domain was selected-for in myeloid malignancies. A shared inventory of pooled gene variant libraries would be a powerful tool for variant-to-function mapping and common practices for the design composition and readout of these libraries will enhance their value.

Conclusions

Complex ORF variant libraries with all possible single-amino-acid substitutions are powerful tools that enable one to exhaustively test all possible protein variants simultaneously in a single experiment. Each experiment produces extensive dictionary of variants and functions. Yet, this technology has high access barriers as it demands expertise across many disciplines, and the costs of producing and utilizing the libraries are prohibitive to many academic labs. This report explores some factors that are important for screening very large variant pools and presents a set of well-tested practices and tools to conduct pooled saturation mutagenesis screens.

Methods

Experimental methods for the published data employed in this study are provided in the original publications. However, methods for experimental steps for large-pool variant screening that incorporate our current best practices, based on experience from over 20 variant screens (**Supplementary Table 1**) are provided here along with the analysis and computational methods and the data-driven evaluations of library compositions and variant screen error modes that are presented in this study.

ORF expression vector. A lentiviral delivery expression vector that we have frequently used to clone site-saturation variant libraries is pMT025 (**Supplementary Fig. 7**). This vector was used for the libraries and screens that are re-analyzed in this report and numerous other ORF variant libraries (**Supplementary Table 1)**, It is suited for restriction/ligation cloning using NheI/BamHI, or NheI/MluI restriction enzyme pairs. The vector has a puromycin resistance selection marker driven by an SV40 promoter. After the cloning, the ORF expression is driven by EF1α. To facilitate cloning, we require that the ORF sequences be properly flanked with constant sequences, in which restriction enzyme sites for cloning are highlighted:

GGTTCAAAGTTTTTTTCTTCCATTTCAGGTGTCGTGAG**GCTAGC**GCCACC [ORF with ATG start and stop]**GGATCC**CGGGACTAGT**ACGCGT**TAAGTCGACAATC

Template ORF plasmid. ~~The starting library we are using is either in a pUC57 plasmid form (libraries made via MITE method) or arrive as linear fragments (libraries purchased from Twist Bioscience).~~ An ORF variant library is built from a common reference sequence to which mutations are introduced. This reference sequence may be some version of the “wild type”

coding sequence of a gene, or it may itself already possess known functional mutations to which additional mutations are added to form the library. In this paper, we call this template sequence the “reference sequence”. It is important at the outset of library design to ensure that the chosen reference sequence will be fully compatible with all the requirements of library construction and screening. For example, in the cases where a preliminary choice of ORF reference sequence contains restriction sites needed for the library cloning, the reference sequence need to be modified to remove these sites. If the screening assay has been developed using expression of the the initial reference sequence, then any changes to this sequence necessitate ensuring that these changes have no impact on the reference sequence phenotype and assay. In general, we recommend that, prior to library construction, the screening assay be fully vetted using the *exact* reference sequence and expression vector that the library will employ, even if that requires repeating assay development experiments that have been performed using a very similar vector or ORF sequence. It is equally important that a clonal plasmid of the template ORF is cloned into the vector, made into virus, tested for transduction, selection condition and genomic DNA prep. Finally, the genomic DNA is tested for PCR condition to be used in future screen deconvolution. Twist Bioscience uses this clonal plasmid as the PCR template to produce the mutagenesis library.

Two library construction techniques are described here, one in which the ORF variant sequences originate as pooled linear oligonucleotides encompassing the entire ORF, purchased from Twist Biosciences, and another method in which the variants are assembled from 90-nt tiles in a method described below to create a pool of the variants in a pUC57 plasmid.

Library design. A saturation mutagenesis designer written in R is available (see the links listed in **Supplementary Table 2**). The template ORF sequence is first checked for absence of restriction sites to be used in cloning: either NheI/BamHI or NheI/MluI. As noted, the ORF template may need to be modified in order to satisfy these requirements. The designer then steps through from N-terminus to C-terminus of the template ORF one amino acid position at a time to select variant codons. For a given template codon there are 63 possible variant codons to choose from at each amino acid position. These 63 codons are divided into two bins: 54 codons that have 2- or 3-nucleotide differences relative to the template codon (2-3-nt delta), and 9 codons with a single-nucleotide difference (1-nt delta). The codons that would violate the restriction sites exclusion rule are removed from both bins. In designing missense or nonsense variants, we first select from the 2-3-nt delta codon bin, and then if needed from the 1-nt delta codon bin. If within a bin there are multiple codon choices available for an amino acid variant, the codon is picked by sampling the available codons with probability weightings according to the codon usage frequency of the organism (*Homo sapiens* in all our experiments to date, http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=9606). For silent codon changes we pick only from 2-3-nt delta bin. In the end, each of the designed variant codons substitutes for its corresponding template codon in the full-length reference ORF as one member of the library.

Mutagenized ORF Library cloning. Saturation mutagenesis library production requires that variant-defining codons be encoded in DNA oligo strands, which will be chemically synthesized. For MITE method, once the DNA oligo strands bearing the mutagenized codons are produced, they are first amplified by PCR and then cloned into a high copy number entry vector, pUC57. For Twist Bioscience method, the DNA oligo strands serve as PCR primers to introduce codon

changes into full-length ORF template to produce a pool of properly flanked and variant-carrying full-length ORFs that are ready for cloning into pMT025 ORF expression vector.

MITE method: The region of the open reading frame to be mutated is partitioned into 90-base tiles. Oligos representing all mutations in the space of a tile were synthesized. Each 90 base tile is flanked by 30 base sequences that are complementary to the reference ORF. Oligos for adjacent tiles should be synthesized on separate arrays. Oligo tile pools were amplified by emulsion PCR using primers to the 30 base flanking regions and purified on a 2% agarose gel. Entry vector cloning is performed on a per tile basis by linearizing the entry vector at the appropriate tile position using Phusion polymerase (New England Biolabs) and primers to the regions flanking the tiles. The linearized vectors were purified on 1% agarose gel and DpnI treated. The DpnI-treated linear plasmid backbones were then mixed with the relevant PCR amplified tile and assembled by *in vitro* recombination with the NEBuilder HiFi DNA Assembly Master Mix (New England BioLabs). The assembly reactions were purified, electroporated into TG1 *E. coli* cells (Lucigen), and recovered for one hour at 37˚C in Recovery Media (Lucigen). Aliquots from the transformations were used to inoculate overnight cultures of LB containing 25 µg/mL of Zeocin (ThermoFisher). Cells were harvested by centrifugation and plasmid DNA was isolated using the QIAGEN Midiprep Plus kit (QIAGEN). Plasmid pools, each corresponding to a tile, were verified by sequencing Nextera XT (Illumina). A pool of all pUC57-tile libraries was made with equal weight per variant. This pool was subjected to restriction digest (e.g., NheI and BamHI). After purification by 1% agarose gel, the excised linear fragment library is ready for cloning ligation with an expression vector that was pre-processed with the compatible restriction enzyme pair.

Twist Bioscience method: The mutagenesis library made by Twist Bioscience was delivered as linear fragments of the full-length ORF flanked with adapters for cloning. The fragment library was first digested with NheI and BamHI overnight, and then underwent ligation with pMT025 vector pre-processed with the same restriction enzyme pair.

Cloning ligation: For either cloning method, the ligation was carried out with 5:1 insert-to-vector molar ratio, using T7 DNA ligase at room temperature for 2 hours. The ligation was cleaned up with isopropanol precipitation. The resulting DNA pellet was used to transform Stbl4 bacterial cells. Plasmid DNA (pDNA) was extracted from the harvested colonies using QIAGEN Maxi Prep Kits. The resulting pDNA library was sequenced via the Illumina Nextera XT platform to determine the distribution of variants within the library.

Virus preparation and titer measurement. Lentivirus was produced and quantified by the Broad Institute Genetic Perturbation Platform (GPP). The detailed protocol is available at http://www.broadinstitute.org/rnai/public/resources/protocols. Briefly, 293T viral packaging cells were transfected using TransIT-LT1 transfection reagent (Mirus Bio) with the pDNA library, a packaging plasmid containing *gag*, *pol* and *rev* genes (psPAX2, Addgene), and an envelope plasmid containing VSV-G (pMD2.G, Addgene). Media was changed 6-8 hours after transfection, and virus was harvested 24 hours thereafter. The virus titer was measured by infecting 3000 A549 cells in 96-w cell culture plate with a series of diluted standard virus of known titer for a standard curve, and a series of diluted library virus of unknown titer to find data points that lie in the linear range of the standard curve. Twenty-four hours post infection, the

cells were subjected to puromycin selection for 2 days. The surviving cells were quantified with alamarBlue.

Determine cell numbers to start and finish the screen. In general, infections were performed, per infection replicate, to aim a representation on average of 1000 cells carrying a given variant throughout the entire experiment. At the time of splitting for drug-treatment and reference sample pair, the same rules apply. For positive selection screens, where a majority of the library will be depleted and a small number of cells carrying variants that are a subset of the library will survive, and consequently a lower total cell number can be collected. In negative selection screen where the majority of the library will be present, the end-point samples should be of at least 1000 cells per variant (i.e., per library member).

Genomic DNA preparation. Upon the completion of the screen, cell pellets were collected. Genomic DNA was extracted from cell pellets using the Machery Nagel NucleoSpin Blood Kit (Midi, Clontech Cat. #740954.20; XL Maxi, Clontech Cat. #740950). The detailed protocol is available at http://www.broadinstitute.org/rnai/public/resources/protocols.

Design PCR primers. The integrated ORF in genomic DNA will be amplified by PCR. The PCR products will be shotgun sheared with transposon, index labeled, and sequenced with next-generation sequencing technology. Each sequenced fragment is generated by 2 transposon insertion events so that both ends of the fragment are tagged with a transposon. Therefore, the ends of the gDNA PCR products will be less frequently represented in the NGS library and will have shallow coverage. For this reason, we advise that the PCR primers be designed in such way

that there is a ~100 bp extra sequence at each end leading up to the mutated ORF region. For pMT025 vector, we use these 2 primers:

Forward: 5'-ATTCTCCTTGGAATTTGCCCTT-3'

Reverse: 5'-CATAGCGTAAAAGGAGCAACA-3'

Determine the number of PCR reactions per sample. For positive screen end-point samples, the number of cells that need to go through ORF-extracting PCR is determined in such a way that for each library variant, gDNA from 500 or more cells will be subjected to PCR. For reference samples and negative selection screen end-point samples, the cells/variant ratio should be maintained at 1000 or more. PCR is sensitive to the amount of gDNA input and prior to screening, these conditions must be optimized. For example, in 50ul PCR reactions, we test different amounts of gDNA, using, for example, a series of 250 ng, 500 ng, 1 µg, 2 µg. This titration will determine the optimal concentration of gDNA that can be used to produce robust PCR products. In general, for human cell lines, we assume a cell produces 6.6 pg DNA. Here is the formula to determine how many PCR reactions are needed for a sample:

$$\text{PCR reactions per sample} = 6.6 * \frac{total\ number\ of\ variants\ in\ the\ library}{ng\ of\ gDNA\ per\ PCR\ well}$$

ORF amplification from gDNA by PCR. PCR reactions were set up in 96-well plates according to the optimized PCR condition and number of PCR reactions per sample calculated above. The volume of each PCR reaction was 50 µL and contained an optimized amount of gDNA. Q5 DNA polymerase (New England Biolabs) was used as the DNA polymerase. All PCR

reactions for each gDNA sample were pooled, concentrated with a PCR cleanup kit (QIAGEN), and separated by gel electrophoresis. Bands of the expected size were excised, and DNA was purified, first using a QIAquick kit (QIAGEN), and then an AMPure XP kit (Beckman Coulter).

Nextera sequencing. Nextera reactions were performed according to the Illumina Nextera XT protocol. For each sample, we set up 4 Nextera reactions, each with 1 ng of purified ORF PCR products. Each reaction was indexed with unique i7/i5 index pairs. After the limited-cycle PCR step, the Nextera reactions were purified with AMPure XP kit. All samples were then pooled and sequenced using an Illumina sequencing platform chosen among Miseq, Hiseq, Nextseq, or NovaSeq, according to the number of reads needed as calculated below.

Determine number of reads needed for the screen. We need to sequence for every variant 500x (for positive selection screen samples) or 1000x (for negative selection screen samples, or reference samples) the full-length PCR product. Thus, the number of total reads depends on ORF length, total number of variants in the library, read length, read quality etc. Below are the formulas:

For negative selection screen samples, or reference samples:

$$\text{Reads per sample} = 7000 * \frac{(ORF\ length\ in\ NT)\text{^}2}{read\ length * read\ assembly\ rate}$$

Where '*read length'* is typically 300, as we use paired-end sequencing reads, 150 bases each, and '*read assembly rate'* is in range of 0.7-0.9.

For a positive selection screen sample:

$$\text{Reads per sample} = 3500 * \frac{(ORF\ length\ in\ NT)\text{^}2}{read\ length * read\ assembly\ rate}$$

Determine the optimal parameters for variant-calling software 'ASMv1.0' The next-generation sequences are from transposon-sheared fragments, which are not uniform in size (average fragment size: 140 base pairs, data not shown). For fragments longer than 2x read length (e.g., 2 x 150 = 300 bases), the pair of reads are disjoint reads; for fragments shorter than the read length (e.g., 150 bases), the reads are fully overlapping and each extends into the transposon adaptor located in the opposite end; for fragments sized between 1x and 2x read length (e.g., 150-300 bases), the pair of reads overlap but do not have transposon sequences. The variant calling software first aligns reads to the reference sequence. The transposon tails are then clipped off at the coordinate determined by the start of its paired read. It then applies the ***min-q*** (minimum base quality) filter to trim the end of the aligned reads to ensure there are at least 15 bases from the trimmed end inwards with base quality above the min-q parameter. The ***min-read-length*** filter admits the trimmed reads that meet the threshold. All failed reads go into the low-q bin. Next, the filter of ***min-flank-length*** is applied that requires reference bases at end of the trimmed reads. A min-flank-length of 2 or more bases prevents cases in which a changed codon is split between 2 Nextera fragments. It also prevents reads ending with transposon sequences from being called as variants in cases for which transposon clipping fails to work due to a lack of paired reads. Finally, a variant is called only when the underlying base changes are

above the min-q threshold. In the end, what matters are reads defined as ‘wild-type’ and reads defined as ‘called variants.’

To tune our software parameters, we used the sequencing data of a saturation mutagenesis plasmid DNA library for *PDE3A* gene. This library was built by mutagenizing a region of 1425 base pairs using Twist Bioscience services. The errors in the library are concentrated in within 30 bases of a planned codon change (**Fig. 3**). These errors can be explained by mutagenesis primer synthesis errors. The codon changes are located in the middle of the mutagenesis primers, sandwiched between two 20-30 base pair flanks. In next-generation sequencing, the transposon sheared fragments are, on average, 140 base pairs, similar to the sequencing read length. The close proximity of variants, planned and unintended, indicates that these variants reside on the same transposon-sheared fragment and on the same sequencing read. Therefore, each library ORF, after transposon shearing, produces one variant-bearing fragment; the rest of the fragments are all wild-type. That is, if we ignore miscalls introduced by PCR, the wild-type/variant read ratio of this library should be: 1425/140 -1 = 9.2. This ratio, 9.2, represents the highest ratio we could observe, as any PCR miscalls would only shift this ratio lower.

In determining the PCR miscall rate, we processed a clonal plasmid of pMT025-PDE3A. The sequencing of this clonal plasmid DNA, which had undergone the same PCR process as the saturation mutagenesis plasmid DNA library, showed error rate of 0.00037 miscalls/sequenced nucleotide. Thus, when we factor in PCR miscalls, the read ratio of wild-type/variant becomes 5.9:

Start with miscall-free wild-type/variant read count ratio 9.2:1, the miscalls switch (9.2*140*0.00037)=0.48 from wild-type group to variant group, producing a new wild-type/variant read count ratio:

$$(9.2 - 0.48)/(1 + 0.48) = 5.9$$

By processing the same plasmid DNA library sequencing data set with various parameters, we were able to identify the optimal parameters that produce a wild-type/variant ratio of 5.9, and yield reasonable output of both wild-type and variants: min-q = 30, min-read-length = 40, min-flank-length = 2.

NGS Data processing and analysis: The data from the NGS pipeline are in .fastq format. Each NGS cluster produced 2 long reads (150 bases each), and 2 index reads (each 8 bases long). The 2 indexes identify a screen sample, so that a large number of samples can be pooled and sequenced in bulk. Site-saturation variant library screen data are typically 4-5TB in data size. Generous computer storage is needed to process the data. The reads need to go through the filtering process (**Supplementary Fig. 2**) prior to variant calling.

<u>Read filtering:</u> The reads from an NGS fragment are trimmed (1) by base quality, (2) by removing non-ORF transposon sequences, (3) by requiring that the outermost nucleotide variations in the reads have, at minimum, a number of reference bases as specified by the min-flank-length parameter, and (4) by requiring the trimmed reads be no shorter than the min-read-length (see **Supplementary Fig. 2**)

Variant calling**:** There are two versions of variant calling software, ORFCallv1.0, the earlier version; and ASMv1.0, the current version. The reads that have passed the filtering step are used differently by ORFCallv1.0 and ASMv1.0:

i) ORFCallv1.0 (an earlier version software; Giacomelli *et al*[9]): Variants are tallied in the context of a single codon space. Without regard of the association of the multiple changes physically in a single molecule, ORFCallv1.0 essentially scans from N- to C-termini, one amino acid position at a time, and sums up the occurrence of codons or amino acids at that position.

ii) ASMv1.0**:** This is a molecule-centric approach and an enhanced version over ORFCallv1.0. The NGS dataset is analyzed in its entirety, with the following new capabilities: (1) The trimmed reads are evaluated full-length; that is, variants are called in the context of entire read (or read pair) and by nucleotides, not amino acids. The programmed variant is called when 2 conditions are met: the detection of the programmed nucleotide changes, and the absence of any ***additional*** nucleotide variations throughout the entire read or read pair. (2) Indel detection: This feature was lacking in ORFCallv1.0. This new feature of ASMv1.0 allows us to process indel libraries.

In ASMv1.0, nucleotide mutations in a molecule are depicted nucleotide-wise by position along the PCR product, followed by the nature of the nucleotide change. For example, ‘a change from G to A at position 1140’ is described as ‘**1140:G>A’;** ‘a

deletion of an A at position 1141' is '**1141:A>-**'; 'an insertion of a T after position 1142' is '**1143:->T**'.

Also, with ASMv1.0, a host of annotations formatted as a datafile can be derived from the variant descriptions. For example, a *TP53* variant with description '**1140:G>A, 1141:A>C, 1142:G>C**' is annotated as (a pair of brackets marks an annotation): [E343T] [Intended] [one codon change] [343:GAG>ACC] [missense] [substitution]. Another *TP53* variant with description '**1140:G>A, 1141:A>C, 1142:G>C, 1145:G>T**' is: **[**E343T] [Unintended] [two codon changes] [343:GAG>ACC, 344:CTG>CTT] [substitution] [missense/nonsense] etc. These two variants encode the same amino-acid variant. The first variant is an intended variant that matches a designed nucleotide-variant of the library. The second variant is the same amino acid call as the first variant, but it is annotated, rightfully, as an unintended variant carrying an extra nucleotide change 4-nucleotide away from the intended codon change - this extra nucleotide change results in a silent codon change.

<u>Parsing data into a form ready for hit calling</u>**:** The output files from the above variant-calling software are still large in size (over 100MB), large in numbers of files, and not readily useable. These output files need to be parsed, aggregated and annotated. For this purpose, we built a parser called ASMv1.0_Parser, which was written in R, and is available for download (see the links listed in **Supplementary Table 2**). The Parser Pak consists of a 'readme' file detailing the instructions, and R codes and training data sets. Through the parser, all data files will be organized into a single .csv file, one column per screen sample. These .csv files include a short version that is a subset of the planned variants only, and a complete version that includes all

called variants, including planned/unplanned, insertion/deletion/substitution, etc. In addition, we also report the ORFCallv1.0 version of data.

Scaling: We use read counts from the NGS data as the quantitative measurements of variants defined by 'changes' relative to the reference sequence. Reads that entirely match the reference sequence are not informative, as they are, as a group, the aggregate of all variants whose changes are located outside of the region covered by the read. At any amino acid position, the total coverage, specifically the sum of counts of variants passing through the position and the sum of counts of the wildtype reads, should be consistent throughout the ORF. But that is not actually the case due to the transposon insertion bias that results in unevenness of total read coverage (**Supplementary Fig. 9**). In cases that address the relative abundance among variants, as in **Supplementary Figure 10**, the raw counts need to be scaled to the fraction of the total read coverage across the codon position. However, in cases that compare the variant abundance change between two samples, such as computing fold-changes in screening samples, scaling the raw counts may not be necessary, as the positional raw count variations caused by transposon insertion bias will cancel out between two samples when computing fold-changes (**Supplementary Fig. 11**).

Acknowledgements

We thank our collaborators of saturation mutagenesis projects, Alice Berger, Jesse Boehm, Ben Ebert, Laura Evans, Eejung Kim, Shengwu Liu, Julie-Aurore Losman, Huong Ly, Peter Miller, Utthara Nayar, and Nikhil Wagle for demonstrating the diverse utilities of the technology, Broad Institute Genomics Platform for providing NGS support, Nancy Tran and Cong Zhu for the early work on saturation mutagenesis method, Stephanie Milczarek, Kirsten

Milenkowic, Michael Zagieboylo, and Jonah Schwartz for conducting lab work for some projects, Leon Yang for critical reading of the manuscript, and to the members of the Genetic Perturbation Platform for their help and support. This work was supported by NCI U01 CA176058 (WCH), the Mark Foundation (WCH), American Cancer Society Professorship (MM), American Cancer Society Mentored Research Scholar Grant MRSG-18-202-01 (ALH), Department of Defense W81XWH-18-PRCRP-CDA (ALH), Team Lick Cancer (ALH), Unravel Foundation (ALH), and Wong Family Award (ALH), and the Functional Genomics Consortium (D.E.R.).

Author Contributions

X.Y., D.E.R., A.L.H., A.O.G., and T.S. wrote the manuscript. W.C.H., M.M., H.G., A.J.A., R.E.L., N.P, M.M, A.B., and F.P. reviewed and edited the manuscript. A.L.H., X.Y., T.S, and D.E.R. designed the new variant-calling software ASMv1.0. T.S. wrote the variant-calling software ORFCallv1.0 and ASMv1.0. X.Y., A.L.H., D.E.R., and T.H. optimized the run parameters of ASMv1.0. X.Y. wrote R codes to design site-saturation variant libraries and to parse the output of the variant-calling software. X.Y. conducted data analysis with input from D.E.R., A.L.H., A.O.G., and assistance from B.P.L. D.A and T.G. adapted the variant-calling software into a saturation mutagenesis data pipeline. X.Y., A.L.H., A.O.G, R.E.L., B.P.L., B.F, H.K., E.S., F.P., and L.S. designed and conducted wet-lab experiments. All authors contributed to and approved the manuscript.

Competing Interests

W.C.H. is a consultant for ThermoFisher, Solasta, MPM Capital, iTeos, Jubilant Therapeutics, Tyra Therapeutics, RAPPTA Therapeutics, Frontier Medicines, KSQ Therapeutics and Paraxel. A.J.A has consulted for Oncorus, Inc., Arrakis Therapeutics, and Merck & Co., Inc, and has research funding from Mirati Therapeutics, Syros, Deerfield, Inc., and Novo Ventures that is unrelated to this work. A.O.G. is a share and option holder of 10X Genomics. D.E.R. receives research funding from members of the Functional Genomics Consortium (Abbvie, BMS, Jannsen, Merck, Vir), and is a director of Addgene, Inc.

References

1. Hughes RA, Ellington AD: **Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology.** *Cold Spring Harbor perspectives in biology* 2017, **9:**a023812.
2. Heather JM, Chain B: **The sequence of sequencers: The history of sequencing DNA.** *Genomics* 2016, **107:**1-8.
3. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S: **High-resolution mapping of protein sequence-function relationships.** *Nature Methods* 2010, **7:**741-746.
4. Fowler DM SJ, Fields S. : **Measuring the activity of protein variants on a large scale using deep mutational scanning. Nature Protocols.** 2014.
5. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, Hopf TA, Marks DS, Oren M, Segal E: **A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation.** *Mol Cell* 2018, **71:**873.
6. Mighell TL, Evans-Dutson S, O'Roak BJ: **A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships.** *Am J Hum Genet* 2018, **102:**943-955.
7. Weile J, Kishore N, Sun S, Maaieh R, Verby M, Li RJ, Fotiadou I, Kitaygorodsky J, Wu YZ, Holenstein A, et al: **Shifting landscapes of human MTHFR missense-variant effects.** *American Journal of Human Genetics* 2021, **108:**1283-1300.
8. Awad MM, Liu S, Rybkin, II, Arbour KC, Dilly J, Zhu VW, Johnson ML, Heist RS, Patil T, Riely GJ, et al: **Acquired Resistance to KRAS(G12C) Inhibition in Cancer.** *N Engl J Med* 2021, **384:**2382-2393.
9. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, Howard TP, Takeda DY, Ly SH, Kim E, et al: **Mutational processes shape the landscape of TP53 mutations in human cancer.** *Nature Genetics* 2018, **50:**1381-1387.

10. Garvie CW, Wu X, Papanastasiou M, Lee S, Fuller J, Schnitzler GR, Horner SW, Baker A, Zhang T, Mullahoo JP, et al: **Structure of PDE3A-SLFN12 complex reveals requirements for activation of SLFN12 RNase.** *Nat Commun* 2021, **12:**4375.
11. Kwon JJ, Hajian B, Bian Y, Young LC, Amor AJ, Fuller JR, Fraley CV, Sykes AM, So J, Pan J, et al: **Structure-function analysis of the SHOC2-MRAS-PP1C holophosphatase complex.** *Nature* 2022, **609:**408-415.
12. Miller PG, Sathappa M, Moroco JA, Jiang W, Qian Y, Iqbal S, Guo Q, Giacomelli AO, Shaw S, Vernier C, et al: **Allosteric inhibition of PPM1D serine/threonine phosphatase via an altered conformational state.** *Nat Commun* 2022, **13:**3778.
13. Persky NS, Hernandez D, Do Carmo M, Brenan L, Cohen O, Kitajima S, Nayar U, Walker A, Pantel S, Lee Y, et al: **Defining the landscape of ATP-competitive inhibitor resistance residues in protein kinases.** *Nat Struct Mol Biol* 2020, **27:**92-104.
14. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, et al: **Prospective functional classification of all possible missense variants in PPARG.** *Nat Genet* 2016, **48:**1570-1575.
15. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS: **Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes.** *Nucleic Acids Res* 2014, **42:**e112.
16. Masaki Y, Onishi Y, Seio K: **Quantification of synthetic errors during chemical synthesis of DNA and its suppression by non-canonical nucleosides.** *Sci Rep* 2022, **12:**12095.
17. Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N: **FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control.** *Nature Methods* 2019, **16:**879-886.
18. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, et al: **Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.** *Cell* 2020, **182:**1295-1310.e1220.
19. Marx V: **Long road to long-read assembly.** *Nat Methods* 2021, **18:**125-129.
20. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al: **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.** *Nature Biotechnology* 2019, **37:**1155-1162.
21. Frank F, Keen MM, Rao A, Bassit L, Liu X, Bowers HB, Patel AB, Cato ML, Sullivan JA, Greenleaf M, et al: **Deep mutational scanning identifies SARS-CoV-2 Nucleocapsid escape mutations of currently available rapid antigen tests.** *Cell* 2022, **185:**3603-3616 e3613.
22. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF: **MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect.** *Genome biology* 2019, **20:**223-223.
23. Boettcher S, Miller PG, Sharma R, McConkey M, Leventhal M, Krivtsov AV, Giacomelli AO, Wong W, Kim J, Chao S, et al: **A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies.** *Science* 2019, **365:**599.

Figure legends

Figure 1. Pooled screens with ORF variant libraries versus shRNA/sgRNA libraries. A plasmid pool of genetic perturbation agents is first packaged into lentiviral particles. Cells were then transduced with the pool of virus at an infection rate of 30-50% to ensure that most of the infected cells get a single genetic perturbation. The perturbed cell population is then subjected to either selection pressure, or treatment under a reference condition. Upon the completion of the screen, the cell population is harvested, and the collected screening samples are processed in ways that each library member in each sample is quantified. The resulting abundance scores are used to compute for enrichment or depletion by comparing a sample that has undergone selection and a sample that has been treated under a reference condition. **(a) A typical pooled screen using shRNA/sgRNA libraries.** For shRNA/sgRNA, the collected cells are first processed for genomic DNA (gDNA), which serve as the PCR template to amplify the perturbagen barcodes. These barcodes are in fact the functional elements themselves, the hairpins of shRNA or guides of sgRNA. These barcodes are short enough to be read straight out by a 50-base NGS read. **(b) ORF variant library screens.** In an ORF variant library, however, the variant signatures are subtle as each member of the library differs from the template by a single codon variation, and the variant signatures of the library are spread out across the entire ORF. Consequently, the detection of variant requires (1) PCR-amplification and then shotgun shearing of the full-length ORF, (2) next-generation sequencing of the resulting fragments, (3) variant detecting tools to align the reads to the template, then identify and tally the variant-defining reads. The NGS reads consist of variant signature-carrying reads (shown as red arrows) and wild-type reads (black arrows). The variant-calling software processes these NGS reads for the abundance of each

variant and enables downstream analyses that map the variants to the phenotype strength in the form of log-fold change (LFC).

Figure 2. Reanalysis of *TP53* saturation mutagenesis screen using ASMv1.0. This Giacomelli *et al* [9] $p53^{NULL}$ A549 cell/nutlin-3 screen enriches *TP53* variants that exhibit loss-of-function (LOF), and depletes wild-type-like variants. (**a**) Heatmaps: each variant is placed in a 2-dimensional grid, where codon position is presented along the horizontal dimension and the amino acid mutation along the vertical dimension. We assigned the letter 'X' to represent nonsense mutations, and 'B' to represent silent codon changes. The value of the log-fold change (LFC) between treatment and reference samples is the heat index. We processed the same original NGS data through the two versions of software and presented the resulting LFC. The pair of heatmaps are nearly identical. To help inspect all variants one by one, the third heatmap used the delta(LFC)=(LFC^{ASM} – $LFC^{ORFCall}$) as the heat index. This heatmap revealed variants that were scored differently by the two versions of software. (**b**) Scatter plot comparing all LFC values produced by the two versions of software, each dot representing a planned variant. (**c**) Demonstrating the impact of variant purity on the improved performance of the new ASMv1.0 software. The new software's output allows us to compute 'variant purity.' We binned the variants by variant purity and in each purity bin we assessed the Pearson correlation coefficient between two sets of LFC values produced by the two versions of software. High variant purity results in similar LFC values, whereas as the purity decreases, the two sets of fold-changes diverge. (**d,e**) Testing ASMv1.0 using the published data by Kotler *et al* [5]. The Kotler dataset includes a screen of $p53^{NULL}$H1299 cells with a library covering *TP53* DNA binding domain. We first took a 2964-variant subset of Giacomelli library to match the *TP53* amino acid variants of

the Kotler library (**d**, **left**). We then identified 76 variants that were scored most differently in the Giacomelli screen processed by the two versions of software **(d, right**). We finally compared the 76 Kotler scores against the corresponding Giacomelli LFC scores called by ORFCallv1.0 (**e, left**) and ASMv1.0 (**e, right**).

Figure 3. The errors in our site-saturation variant libraries are low in number and local relative to the location of the planned codon change. (**a**) **Breakdown of the planned variants and those with errors.** Going through all called variants that contain a planned codon change, we identified the number of unintended nucleotide changes. This bar-chart shows that in the pMT025-TP53 library, 84% molecules are planned variants without additional mutations. Of 16% imperfect molecules, the majority are those with 1-2 extra nucleotide changes. (**b**) **Major errors are in the vicinity of a planned change.** We first pinned down the planned change in a read and then travelled in both directions and tallied the errors as a function of the distance to the pinned-down codon change. In this figure, we compared two libraries made with different technologies, *TP53* library by MITE and *PDE3A* by Twist Bioscience. The error distribution profile along the distance to a planned variant is explainable by the oligonucleotide synthesis scheme involved in these technologies. MITE technology uses synthesized DNA strands of 150 bases long, of which 90 bases (called TILE) go into the final products, while Twist Bioscience introduced desired codon changes through oligonucleotide primers with the intended codon changes placed in the middle and flanked with 15-25 bases on either side. The major errors are concentrated within ~100 bases (MITE method) or ~30 bases (Twist Bioscience method) from a planned codon change.

Figure 4. The effect of PCR and sequencing errors on variant quantification. (a) A clonal *TP53* reference plasmid (i.e. all sequences are wild-type) was processed as if it were a screening sample, subjected to PCR amplification, NGS sample preparation and NGS sequencing read out. The base call at each position is predominantly, but not absolutely, the reference nucleotide. The non-reference calls represent artifactual variants that are not actually existing. The heatmap on the left shows tallied counts of called nucleotides at each of the first 30 nucleotide positions; a blank spot represents zero counts. The heatmap on the right depicts the rate of miscalls. Blank spots in this heatmap are either the wild-type nucleotide or nucleotides of zero counts. (**b**) **Abundance distributions of 3 codon types.** *TP53* variants were grouped by the number of differing nucleotides between the variant signature codon and the reference codon; the 3 groups are called 1-nt-delta, 2-nt-delta and 3-nt-delta variants. The miscalls resulting from PCR/NGS errors inflate the abundance measurement of variants whose variant-defining codon is 1-nt away from the reference codon. In the *TP53* library screen data, predominately affected are variants whose signature codon is 1-nt (not 2- or 3-nt) away from the reference codon. As a result, the variants of 1-nt delta codons as a group appear more abundant than groups of 2- or 3-nt delta codons (**left**). The artifactual 'variant' abundance resulting from PCR/NGS errors can be corrected with abundance adjustment by subtracting the miscall rate as determined by the clonal sample experiment (**right**). The variant abundance measurements of the 2-nt and 3-nt delta codons are not affected by the miscalls and the corrections do not alter the abundance distributions of these 2 codon types. (**c**) **Suppression of fold-changes in 1-nt delta codon variants.** Three scatter plots, each for one of the three codon types, depict LFC versus variants ordered by codon positions. A density plot on the right depicts the LFC score distribution within each of the three codon types. Using the pMT025-*TP53* library as an example, 1-nt delta variants

(data-points or curves labeled in red) showed much narrower fold-change values than 2-nt and 3-nt delta variants.

Figure 1

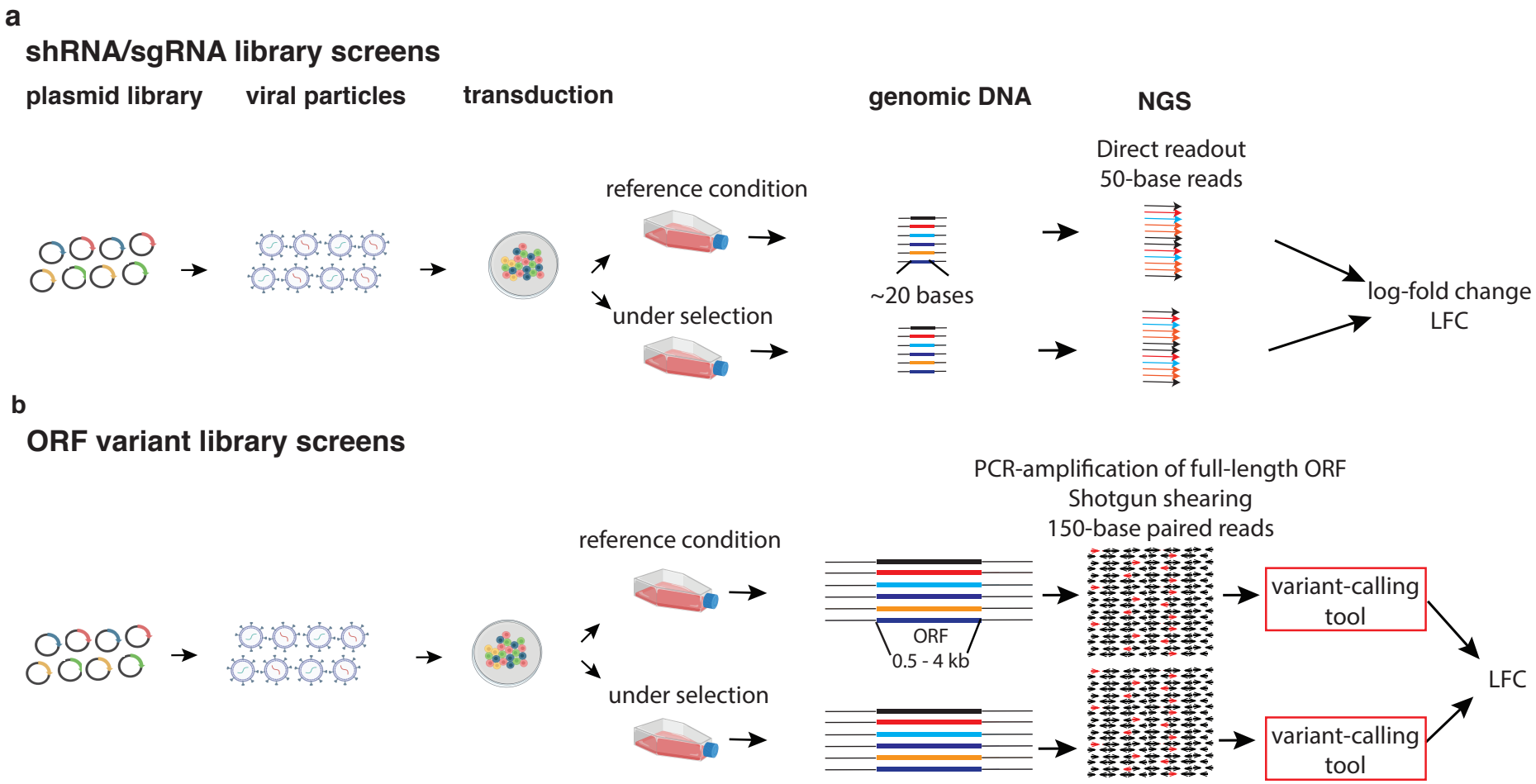

a
shRNA/sgRNA library screens
plasmid library
viral particles
transduction
genomic DNA
NGS
reference condition
under selection
~20 bases
Direct readout
50-base reads
log-fold change
LFC
b
ORF variant library screens
reference condition
under selection
PCR-amplification of full-length ORF
Shotgun shearing
150-base paired reads
ORF
0.5 - 4 kb
variant-calling tool
variant-calling tool
LFC

Figure 2

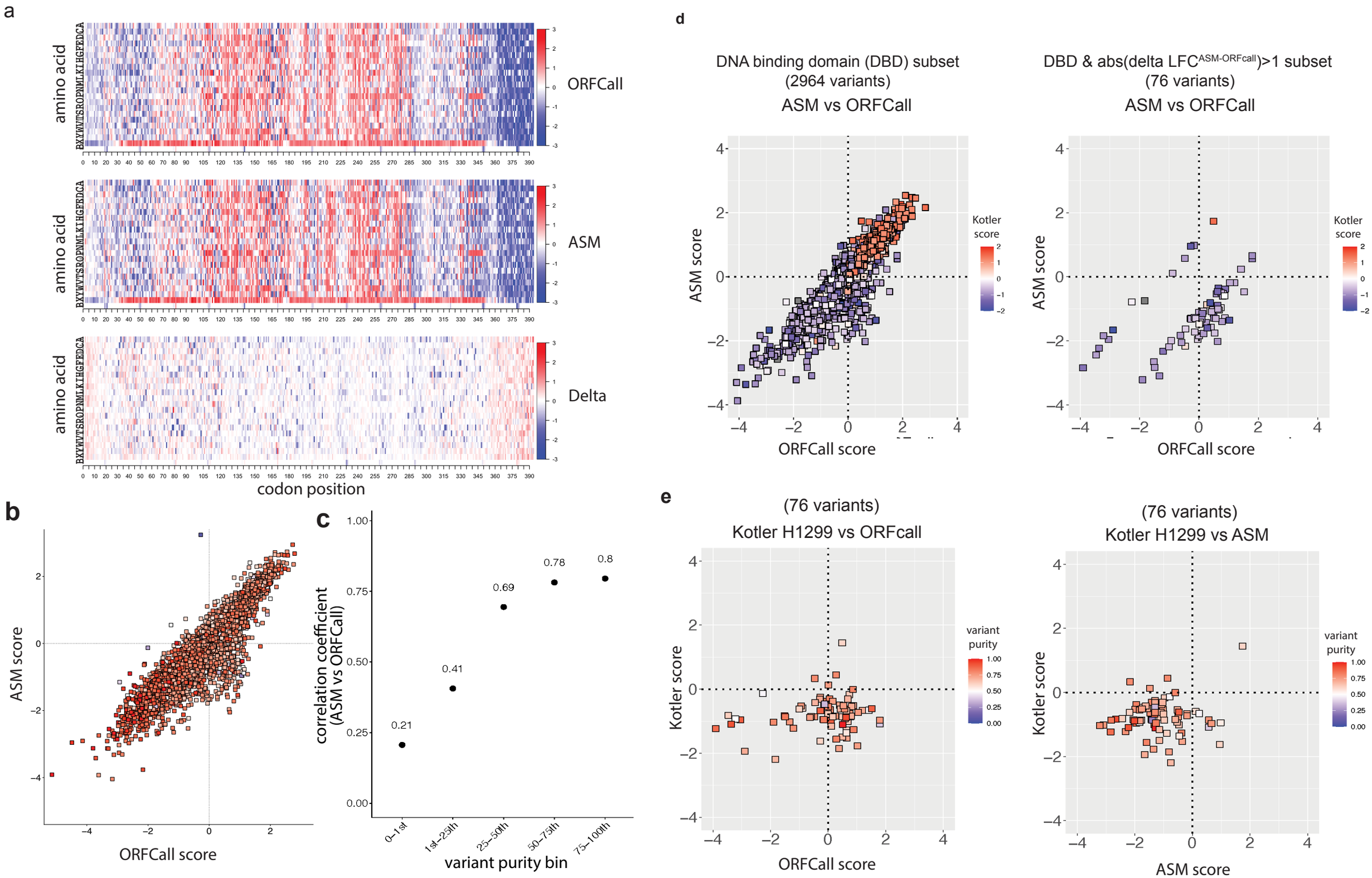

a
amino acid
ORFCall
ASM
Delta
codon position
b
ASM score
ORFCall score
c
correlation coefficient (ASM vs ORFCall)
0.21
0.41
0.69
0.78
0.8
0–1st
1st–25th
25–50th
50–75th
75–100th
variant purity bin
d
DNA binding domain (DBD) subset (2964 variants)
ASM vs ORFCall
DBD & abs(delta LFC^ASM-ORFcall)>1 subset (76 variants)
ASM vs ORFCall
Kotler score
e
(76 variants)
Kotler H1299 vs ORFcall
(76 variants)
Kotler H1299 vs ASM
variant purity
Kotler score
ASM score

Figure 3

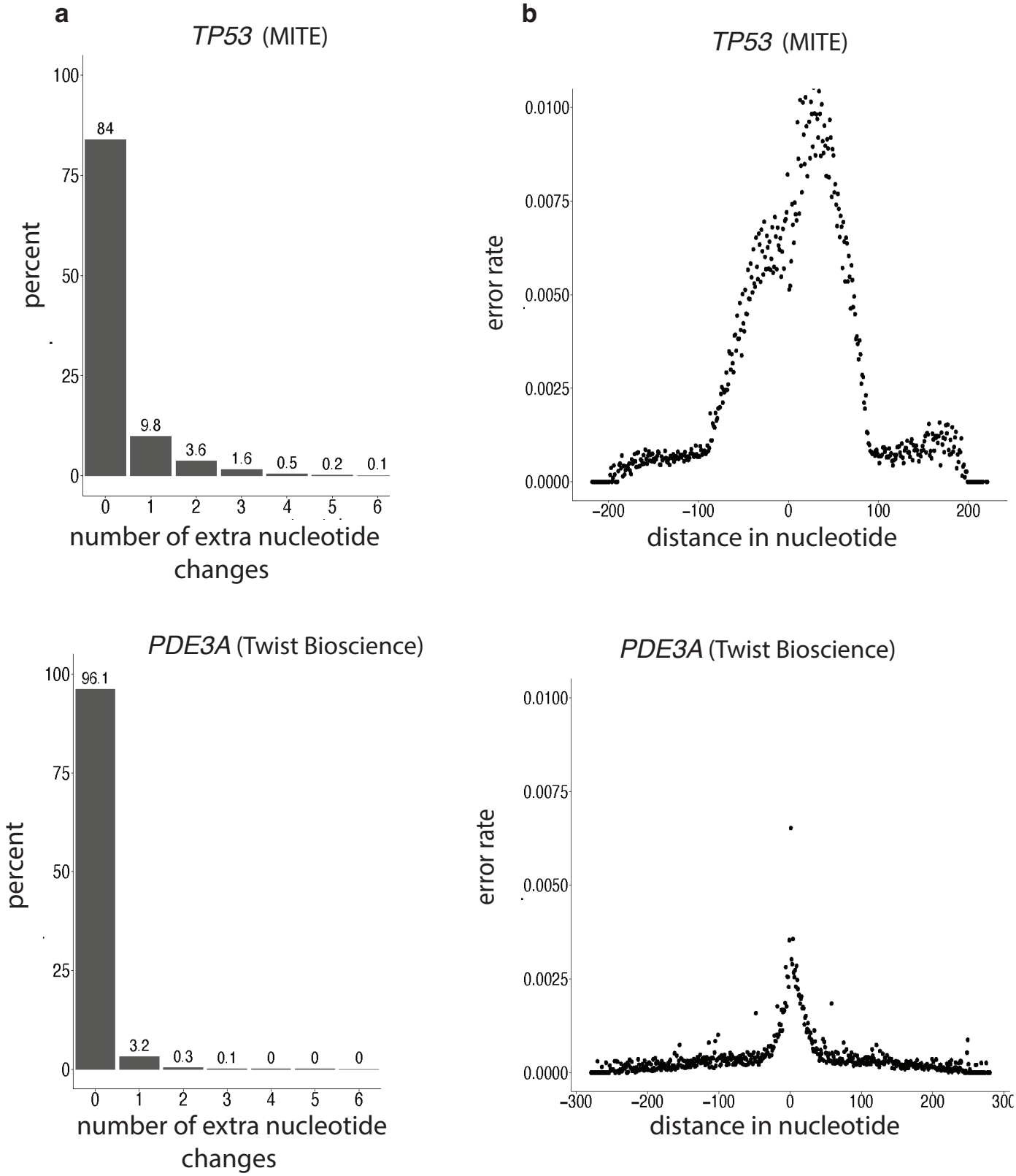
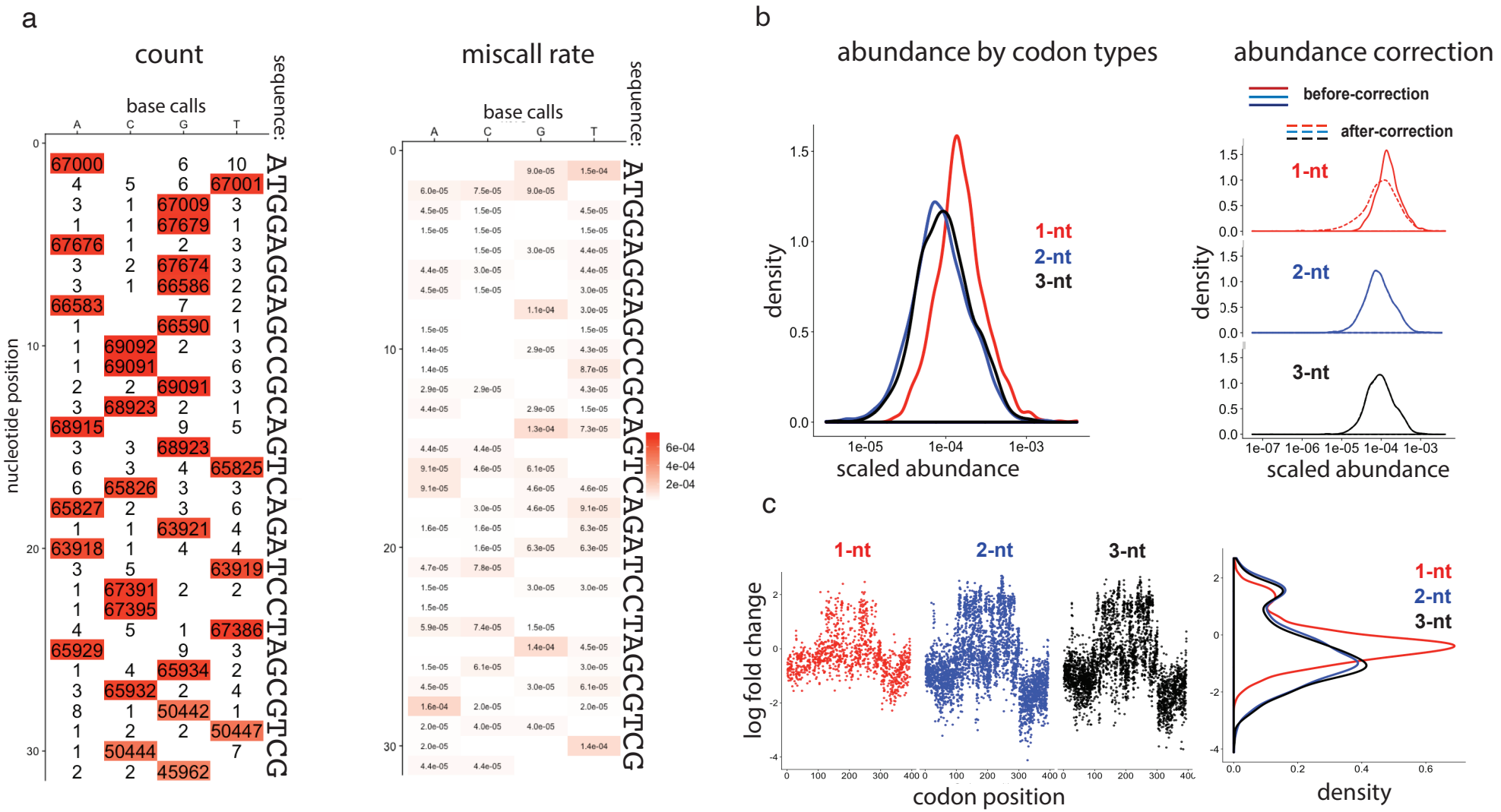

a
TP53 (MITE)
percent
100
75
50
25
0
84
9.8
3.6
1.6
0.5
0.2
0.1
0 1 2 3 4 5 6
number of extra nucleotide changes
b
TP53 (MITE)
error rate
0.0100
0.0075
0.0050
0.0025
0.0000
−200 −100 0 100 200
distance in nucleotide
PDE3A (Twist Bioscience)
percent
100
75
50
25
0
96.1
3.2
0.3
0.1
0
0
0
0 1 2 3 4 5 6
number of extra nucleotide changes
PDE3A (Twist Bioscience)
error rate
0.0100
0.0075
0.0050
0.0025
0.0000
−300 −200 −100 0 100 200 300
distance in nucleotide

Figure 4


a
count
miscall rate
base calls
A C G T
sequence:
nucleotide position
0
10
20
30
6e-04
4e-04
2e-04
b
abundance by codon types
abundance correction
before-correction
after-correction
1-nt
2-nt
3-nt
density
scaled abundance
1e-05 1e-04 1e-03
1e-07 1e-06 1e-05 1e-04 1e-03
c
log fold change
codon position
density