Define protein variant functions with high-complexity mutagenesis libraries and enhanced mutation detection software

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 Institute/Princess Margaret Cancer Centre, 610 University Avenue, Toronto, ON, 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

Open reading frame (ORF) variant libraries have advanced our ability to query the functions of a large number of variants of a protein simultaneously in a single experiment. Variant libraries targeting full-length ORFs typically consists of all possible single-amino-acid substitutions and a stop codon at each amino-acid position. Because a variant differs from the template ORF by merely a single codon variation, variant quantification presents the most profound challenge to this technology. Efforts such as dividing a library into sub-libraries for direct sequencing, or tag-directed subassembly are practical only for small ORFs. Our approach, however, features generating and screening libraries for genes sized up to 3600 bases, shotgun sequencing and an enhanced variant-detecting tool. Having processed screens of ~20 ORF variant libraries, our tool calls variants reliably, and also presents variant annotations in datafiles enabling analyses that have reshaped our strategies governing library design, screen deconvolution, sequencing and its analysis.

An open reading frame (ORF) variant library of a gene of interest provides a powerful means of systematically characterizing the functions of thousands of single amino acid substitutions in a single experiment. The functional characterization of variants of an ORF have historically been confined to alanine scans[1,2], random mutagenesis by error-prone PCR[3] or mutator *E. coli* strains[4], site-directed mutagenesis[5] , or transposon insertional mutagenesis[6]. Advances in DNA oligonucleotide and gene synthesis[7] and in next-generation sequencing (NGS)[8] have accelerated our ability to assess all possible single amino acid changes for every amino acid position of a protein. The resulting collections are also known as saturation mutagenesis libraries, or deep mutational scanning (DMS) libraries. A screen of such a library, that is, a pool of all variants, produces a comprehensive dictionary that maps variants to phenotype strength.

As illustrated in **Figure 1**, pooled screens allow testing the functions of thousands of genetic perturbation agents in a single experiment. Pooled screens rely on introduction of one genetic perturbation cassette, e.g. an shRNA, a CRISPR sgRNA or an ORF variant, per cell into the genome using delivery by a retrovirus. The perturbed cell population is then subjected to selection pressure, for example, growth-over-time proliferation, treatment with a drug, or flow-sorting for a reporter marker. Eventually, the collected cell population samples are processed to measure the abundance of each library perturbation in each of the samples. The perturbagen counts are the basis to assess the enrichment or depletion of each library perturbation by comparing samples that have undergone selection and samples that have been treated with reference conditions (e.g. pre-selection or vehicle treatment samples). In shRNA or CRISPR screens, the functional elements, hairpins or guides respectively, serve as the perfect barcodes, short and unambiguous, for screening sample deconvolution. In contrast, ORF variant library screens are confronted with challenges in 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 a given amino acid position, the template amino acid is typically mutated to 19 other amino acids, a stop codon, and when design rules allow, one or two synonymous codons (**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. The coordinates of all variations in a library are spread out along the full-length ORF. Thus, capturing the variant-defining signature as subtle as a single codon variation located anywhere along the full-length ORF sequence is one of the foremost technical challenges. Over the years, efforts that have been made include (1) dividing an ORF variant library into sub-libraries with mutational regions of 25 amino acids initially[9,10], then 47 and 100 amino acids[11,12], to allow the sub-libraries to be screened and sequenced for direct readout of the variants, or (2) subassembly approaches involving tag-directed PCR amplification[13]. The benefits of these efforts are limited to and deemed practical to small ORFs. Our methodology, on the other hand, emphasizes robust variant-calling tools that process massive shotgun NGS data to detect and accurately quantify variants, and consequently allows one to screen a single saturated variant library for an ORF of any size (e.g. ~500-bp *KRAS* or ~3600-bp *EGFR*) in a single experiment (**Fig. 1b**). Here, we present our methods and strategies as we draw from our ever-evolving knowledge from designing and constructing nearly 20 libraries and analyzing the resulting screening data (**Supplementary Table 1**). We hope that this study will help lower the access barrier to this technology and pave a path for sharing library resources within the scientific community.

Results

Despite differences in screening protocols, all ORF variant library screens converge at the quantification of library alleles in each of the screening samples (**Fig. 1b**). Variant quantification requires resolution of subtle differences among variants. We introduce a new variant-calling software, Analyze Saturation Mutagenesis v1.0 (ASMv1.0; see **Methods** and **Supplementary Fig. 2**), to more accurately process the NGS data and measure variant abundance.

ASMv1.0. ASMv1.0 is a part of the Genome Analysis Toolkit (GATK v4.2.0.0). For each screening sample, this program analyzes NGS reads aligned to the reference sequence that includes the ORF template sequence and flank sequences defined by the primers used to amplify the full-length ORF from genomic DNA. It counts the number of observed molecules that have some given set of mutations. ASMv1.0 observes this mutational landscape without regard to the library design. Because the NGS reads are typically far shorter than the full-length ORF and because each planned variant in the library carries a single codon change from the template ORF, most aligned reads will be wild-type, and serve only to indicate depth of read coverage across a codon position. The relative abundance of reads that do contain mutations gives information on the identity and abundance of the variant defined by those mutations.

The ASMv1.0 variant-calling process is followed by a process that parses ASMv1.0 output files of all screen samples into a single datafile (see **Methods**). Throughout these processes, the toolkit preserved the variant description, which unambiguously defines the called variant. In our experience, the detail-rich variant descriptions and the annotations derived from them are invaluable datasets that allow us to perform a myriad of in-depth data analyses.

The enhanced performance of ASMv1.0. The early version of variant detection software, ORFCallv1.0, was successfully used in analyzing many saturation mutagenesis screens[14-17]. ORFCallv1.0, however, tallies the intended variants in the space of a single codon, and it ignores the association of an intended codon change and any *in-cis* unintended changes. For a library that has prevalent unintended errors associated with an intended change, ignoring the presence of these errors can mask the true phenotype of the intended variant by the phenotypes caused by the unintended changes. As an improvement, our new software, ASMv1.0, calls variants in the context of full-length read pairs, allowing us to separate the variants that bear only the intended sequence change from those that also harbor extra changes. Furthermore, in ASMv1.0, a pair of reads needs to pass through sequential filters such as minimum base quality, minimum read length, and requirement of consistency in overlapping sequences of a pair of reads, before being admitted for variant calling (**Supplementary Fig. 2)**. **Supplementary Table 3** summarizes the enhanced functionalities from ORFCallv1.0 to ASMv1.0.

We used the screens performed in Giacomelli *et al*[14] to compare the two methods. Specifically, the ectopically expressed *TP53* variants were screened in the presence or absence of endogenous wild-type *TP53*, with selection pressure from either nutlin-3 or etoposide. Under nutlin-3 treatment, the screen of p53$^{\mathrm{NULL}}$ cells enriched *TP53* loss-of-function (LOF) variants, and depleted wild-type-like variants (**Fig. 2**). Under etoposide treatment, the p53$^{\mathrm{NULL}}$ cells enriched wild-type-like *TP53* variants and depleted the *TP53* LOF variants (**Supplementary Fig. 3**). Under nutlin-3 treatment, the screen of p53$^{\mathrm{WT}}$ 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 largely consistent with each other (**Fig. 2a, Supplementary Figs. 3a, 4a**). However, we also noted that several variants exhibited substantial differences in LFC calculated using each algorithm; 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, it is inevitable that some planned codon changes are physically associated with one or more unintended nucleotide changes in the same molecule. We define the 'variant purity' of a planned variant as the fraction of pure intended molecules over all molecules that harbor the codon change of interest, whether or not they harbor additional nucleotide changes. When the purity of a planned variant is high, as in a high-quality library, 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, as they reflect the sum of effects of all species that happened to have a given codon change. We computed the purity of every planned variant using the reference samples 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 (panel **c** of **Fig. 2, Supplementary Figs. 3,4)**. It is clear that the two methods correlate well when the variant purity is high, and diverge as the variant purity decreases.

We therefore reasoned that ASMv1.0 is an upgrade over ORFCallv1.0, particularly considering how unintended *in cis* changes are handled. As a way of validating the higher performance of ASMv1.0 over ORFCallv1.0, we compared the Giacomelli screen LFC scores called by both versions of the software with the results from the Kotler *et al*[12] screen that used $p53^{NULL}$ H1299 cells and a small variant library that focused on the p53 DNA binding domain (DBD) (**Fig. 2d**).

Kotler *et al* targeted p53 DBD with 4 libraries, each covering the variants of a 141-base region of the ORF. The Kotler screens were processed by direct sequencing of the mutational region of each of the 4 libraries. Furthermore, Kotler *et al* processed their screen with stringent analysis thresholds by requiring (1) at least 80bp overlap of the two reads in a read pair, (2) perfect sequence agreement in the overlapping region, and (3) minimum read coverage of >200 reads for each variant. These stringent thresholds and the direct readout of variant counts rendered the Kotler dataset the best available for validating the ASMv1.0 algorithm.

Similar to 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 mutational space as the Kotler H1299 screen. Among these 2964 variants, we identified 76 variants from the Giacomelli p53^{NULL}/nutlin-3 screen that were called differently by ORFCallv1.0 and ASMv1.0, with abs(delta(LFC))>1 (**Fig. 2d, right**). Of these differentially scored variants, the LFC scores called by two versions of the software were individually compared with the Kotler screen scores (**Fig.1e).** The ASMv1.0 LFC scores for the Giacomelli p53^{NULL}A549/nutlin-3 screen agree with the Kotler calls better than the ORFCallv1.0 LFC scores do (**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 one to conduct in-depth analyses of each saturated ORF variant library. 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,b, left**), and the errors in the library are predominantly single nucleotide changes.

Taking one step further with the ASMv1.0 output, we investigated the physical proximity distribution of errors in relation to the intended codon change. In **Figure 3**, we compared libraries made by two different methods (see **Methods**): the *TP53* library made by MITE (Mutagenesis by Integrated TilEs)[17], and the *PDE3A* library synthesized by Twist Bioscience. The MITE method involved oligo synthesis to cover the length of 'tiles' that are 90 nucleotides long. The Twist Bioscience technology involved the synthesis of oligos that vary, depending on GC content, between 30-50 nucleotides and with the variant-encoding bases in the middle. In **Figure 3,** we profiled the error rate by distance, in nucleotides, between the locations of the error and the planned change. What this distance profile reveals are significant:

First, it indicates that the errors in the libraries are concentrated near the planned codon changes. The lengths of stretches with frequent errors in the library made by each method, ~90 bases for the library made by the MITE method and ~30 bases for the library made by Twist Bioscience, suggest that errors in the libraries are predominantly DNA oligo synthesis errors. In other words, the errors are near a planned codon change. Therefore, both intended and unintended changes can be detected comfortably by a read of 100 bases or more.

Secondly, it suggests that we can focus on the variant-signature-carrying reads to tally variant counts, and assume that the unseen *in cis* portion of the ORF, the ORF sequence outside of the variant-signature-carrying reads, should be otherwise wild-type and free of errors. In fact, the unseen portion of the ORF is subjected to errors introduced by DNA polymerase during the library build procedure (Twist Bioscience method) or during vector preparation (MITE method). However, ignoring such errors is not unreasonable, because a high-fidelity DNA polymerase with an error rate of 10^{-6} errors per cycle per amplified nucleotide will produce a small number of erroneous molecules (1.1% for a 1kb ORF; 4.3% for a 4kb ORF), which amount to a small fraction of those introduced from oligo synthesis (10% or 18% between two methods) (**Supplementary Table 4**).

ASMv1.0-enabled analysis allows assessing the effect of miscalls (artifactual errors) introduced by processes involving PCR. While the errors discussed above are true errors that are present in the libraries, it is important for us to address the artifactual errors in the form of miscalls introduced by the variant quantification process itself. In the screen deconvolution process, PCR was used to replicate the ORF DNA from genomic DNA (up to 30 PCR cycles) and used again in NGS sample preparation (12 cycles). The errors introduced by these PCR steps are artifacts. We define 'miscalls' as the artifactual and mispresented base calls relative to what are actually in the samples. The error rate of DNA polymerases can be as low as one error per one million bases synthesized in a PCR cycle. When we, in parallel with the screening samples, processed and sequenced the reference *TP53* ORFs that are clonal, pure, and without any variants, the sequencing results are striking – at each nucleotide position, we observed miscalls at a rate of ~0.0003-0.0006 miscalls per nucleotide read-out (**Fig. 4a, right**), or 3-6 miscalls per 10,000 sequenced bases.

While this miscall rate may appear small, it is a significant factor in variant detection. As an example, illustrated in **Supplementary Figure 5**, these miscalls will 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) the artifactually narrowed log-fold change range of 1-nt delta codons (**Supplementary Fig. 5,** legend**)**. Indeed, in the *TP53* library screen with p53$^{\text{NULL}}$ 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, without exception, in all libraries we have processed, and the effect magnifies as ORF length increases.

A recent report[18] presented software DiMSum that used 1-nt delta codons to model PCR/NGS errors as a method to assess library quality. We, however, took a direct approach and opted to mitigate the effects of PCR/NGS miscall errors by implementing library design rules and adding control experiments. First, we set an important rule for saturated ORF variant library design - minimizing the use of codons that are one nucleotide away from the reference codons. In our experience, to achieve 'saturated' amino acid coverage in a library, we cannot completely avoid variant codons that are one nucleotide away from a reference codon. But we can minimize the use of 1-nt delta codons, reducing them from 15% to 3% of the library members. Secondly, dealing with these 3% 1-nt delta variants, we added a clonal template plasmid sample and processing it as if it were a screening sample. One may opt to use the clonal template data to monitor or even correct the screen sample data (**Fig. 3b**, right).

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 variants present on each molecule. However, the technology that can capture the read-out of full-length ORF sequences[19,20] is currently cost-prohibitive for most labs, and more importantly, undesirable for variant detection due to low concordance rate of base calls[21]. For example, in a study by Wenger *et al.*[22], 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 bases read out. This discordance rate is higher than the error rate in oligo synthesis, the key contributor of the real errors in the libraries. Consequently, until long-read sequencing (LRS) technology achieves a polymerase-level rate of concordance, for instance, 99.999%, or Q50 on the Phred scale, one should use NGS for variant detection.

For variant quantification, we favor a sequencing platform capable of producing NGS read pairs of 150 bases per read with a Phred score of Q30 or more for each base call, for four main reasons: (1) We have demonstrated that the errors in our libraries are near the planned codon changes (**Figure 3, right**), and a 150-base read is adequate to detect both the variant-defining sequences and library errors from DNA oligo synthesis. (2) We have shown that it is unnecessary to sequence far into the regions synthesized by DNA polymerase during library construction, as the errors introduced by DNA polymerase in regions outside of the planned codon changes are rare and negligible. In principle, our variant-detecting method is targeting the library ORF regions whose nucleotides are originated from oligo synthesis. To achieve this, we focus on reads that bear a planned codon change, with or without additional changes. (3) We have demonstrated that PCR-introduced miscalls produce predominantly 1-nt delta codon changes. As we minimize the use of 1-nt delta codon variants, the artifactual variants resulting from the miscalls are mostly unintended variants. By ignoring reads carrying only unintended variations, we largely removed the effects of sequencing miscalls. (4) With ASMv1.0 base quality threshold set at Phred score Q37 (with Novaseq NGS), 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 discordance rate, >94% (0.9998^300) of 150-base read pairs used to call variants are free of miscalls.

Silent variants in library are used as a measure of screen baseline. Another rule we have implemented for library design is the inclusion of silent variants with codons that are 2- or 3-nt different from the reference codon. ORF variant library screening often produces a large number of hits. As a result, one cannot always rely on the large portion of the library producing no phenotype to serve as the baseline. It is therefore essential to include silent codon changes in the library; such nucleotide changes allow us to identify and quantify the silent alleles in parallel with missense and nonsense variants (see rows marked as ‘B’ in all fold-change heatmaps). Often, codons encoding an amino acid differ from one another at the wobble base. As a result, many silent codons are one nucleotide away from the reference 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.

In the *SMARCB1* library, we included nearly all possible silent variants, including 1-nt delta silent variants. As expected, due to the artifactual miscalls discussed earlier, the abundance of 1-nt delta silent variants is inflated relative to the abundance of 2- and 3-nt delta silent variants (**Supplementary Fig. 6a**). For the same reason, the 1-nt delta silent variants show an artifactually narrowed log-fold change range, compared with 2- and 3-nt delta groups (**Supplementary Fig. 6b**). 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.

Discussion

Variant detection software ASMv1.0 enabled a cascade of in-depth data analyses that allowed us to explore, for instance, variant purity, the rate and nature of library errors, and the artifactual miscalls. The results of these analyses have shaped and reshaped our strategies at many key steps of the projects, on both laboratory and computational fronts. We believe that the analytical and laboratory technical blueprint 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.

It should be noted that the strong hits among the 1-nt delta variants may still score well, even with the miscall-related log-fold change suppression. The issue of 1-nt delta codon variants in processing screening samples stems from the fact that we sequence the ORF to quantify the variants. 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 5**, the longer the ORF, for each readout of a variant-defining codon, the more copies of the template codon need to undergo PCR-based amplification and NGS, and consequently the more counts of the artifactual 1-nt delta codons will result. Our suggestion of minimizing 1-nt delta codon variants in ORF variant libraries is to insulate the true variant counts in the screening samples from PCR-introduced artifactual miscalls. If one prefers 1-nt delta codons in the variant libraries, 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. 4**, right).

The recent establishment of an open-source platform, MaveDB, to enable variant function data sharing, is a significant development[23]. We further advocate that the scientific community find a way to share saturation mutagenesis reagents and resources. In many cases, the content of a saturated ORF variant library is well defined by the gene reference sequence. For a given gene, the library designed by one lab will be highly similar to one done by another lab. As a perfect example of resources sharing, Boettcher *et al*[24], using the *TP53* saturated ORF variant library created by Giacomelli *et al*[14], quickly demonstrated that the dominant-negative activity of *TP53* missense mutations in DNA binding domain was selected for in myeloid malignancies. If the scientific community can come to an agreement for library sharing and repurposing, the available resources could be used to cover more genes, as opposed to building redundant libraries. Imagine what a collection of saturated ORF variant libraries of 1000 disease-related genes could do!

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. 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)

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 saturated ORF 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.

References

1 Lefèvre, F., Rémy, M. H. & Masson, J. M. Alanine-stretch scanning mutagenesis: a simple and efficient method to probe protein structure and function. *Nucleic Acids Res* **25**, 447-448, doi:10.1093/nar/25.2.447 (1997).

2 Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science (New York, N.Y.)* **244**, 1081-1085, doi:10.1126/science.2471267 (1989).

3 McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random mutagenesis by error-prone PCR. *Methods Mol Biol* **634**, 103-109, doi:10.1007/978-1-60761-652-8_7 (2010).

4 Muteeb, G. & Sen, R. Random mutagenesis using a mutator strain. *Methods Mol Biol* **634**, 411-419, doi:10.1007/978-1-60761-652-8_29 (2010).

5 Kato, S. *et al.* Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proceedings of the National Academy of Sciences* **100**, 8424, doi:10.1073/pnas.1431692100 (2003).

6 Chen, L. *et al.* Transposon insertional mutagenesis in mice identifies human breast cancer susceptibility genes and signatures for stratification. *Proceedings of the National Academy of Sciences* **114**, E2215, doi:10.1073/pnas.1701512114 (2017).

7 Hughes, R. A. & Ellington, A. D. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harb Perspect Biol* **9**, a023812, doi:10.1101/cshperspect.a023812 (2017).

8 Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1-8, doi:10.1016/j.ygeno.2015.11.003 (2016).

9 Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* **7**, 741-746, doi:10.1038/nmeth.1492 (2010).

10 Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature protocols* **9**, 2267-2284, doi:10.1038/nprot.2014.153 (2014).

11 Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *American journal of human genetics* **102**, 943-955, doi:10.1016/j.ajhg.2018.03.018 (2018).

12 Kotler, E. *et al.* A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol Cell* **71**, 873, doi:10.1016/j.molcel.2018.08.013 (2018).

13 Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods* **7**, 119-122, doi:10.1038/nmeth.1416 (2010).

14 Giacomelli, A. O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet* **50**, 1381-1387, doi:10.1038/s41588-018-0204-y (2018).

15 Persky, N. S. *et al.* Defining the landscape of ATP-competitive inhibitor resistance residues in protein kinases. *Nat Struct Mol Biol* **27**, 92-104, doi:10.1038/s41594-019-0358-z (2020).

16 Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* **48**, 1570-1575, doi:10.1038/ng.3700 (2016).

17 Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res* **42**, e112-e112, doi:10.1093/nar/gku511 (2014).

18 Faure AJ, S. J., Baeza-Centurion P, Lehner B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol* **21**, 207, doi:10.1186/s13059-020-02091-3 (2020).

19 Legnini, I., Alles, J., Karaiskos, N., Ayoub, S. & Rajewsky, N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat Methods* **16**, 879-886, doi:10.1038/s41592-019-0503-y (2019).

20 Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e1220, doi:10.1016/j.cell.2020.08.012 (2020).

21 Marx, V. Long road to long-read assembly. *Nat Methods* **18**, 125-129, doi:10.1038/s41592-021-01057-y (2021).

22 Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155-1162, doi:10.1038/s41587-019-0217-9 (2019).

23 Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* **20**, 223-223, doi:10.1186/s13059-019-1845-6 (2019).

24 Boettcher, S. *et al.* A dominant-negative effect drives selection of <em>TP53</em> missense mutations in myeloid malignancies. *Science* **365**, 599, doi:10.1126/science.aax3649 (2019).

Figure 1

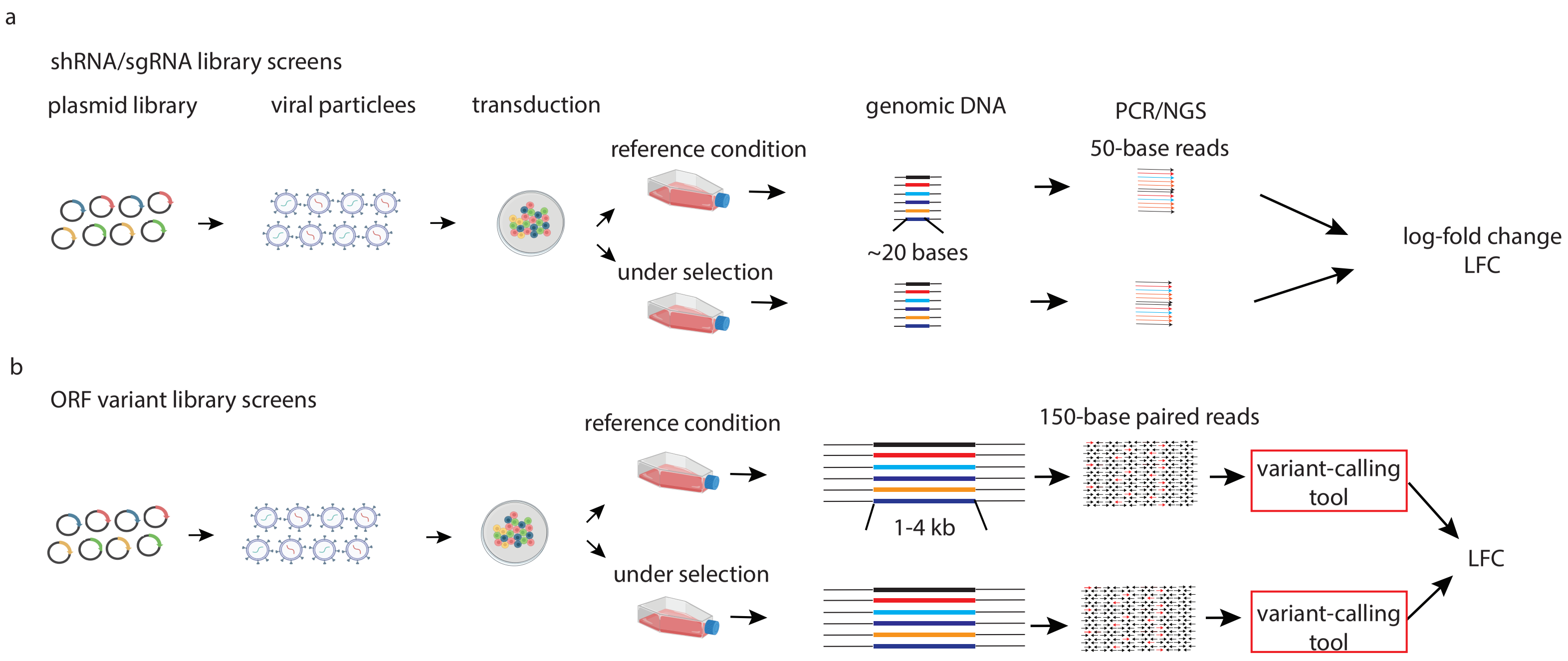
a
shRNA/sgRNA library screens
plasmid library
viral particlees
transduction
reference condition
under selection
genomic DNA
~20 bases
PCR/NGS
50-base reads
log-fold change
LFC
b
ORF variant library screens
reference condition
under selection
1-4 kb
150-base paired reads
variant-calling tool
variant-calling tool
LFC

Figure 1. Pooled screens with ORF variant libraries versus shRNA/sgRNA libraries. A plasmid pool of thousands of genetic perturbation agents is first packaged into lentiviral particles. Cells were then transduced with the pool of virus at an infection rate (the proportion of infected cells over total cells) 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 remaining cell population is harvested and processed for enrichment or depletion of each library perturbation 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) robust variant detecting tools to align the reads to the template, then identify, evaluate, annotate 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 the NGS reads to detect variant signature-carrying reads, which is a minority of the total reads, and records the counts of each variant. These counts are

the basis for downstream analyses that map the variants to the phenotype strength in the form of log2-fold change (LFC).

Figure 2

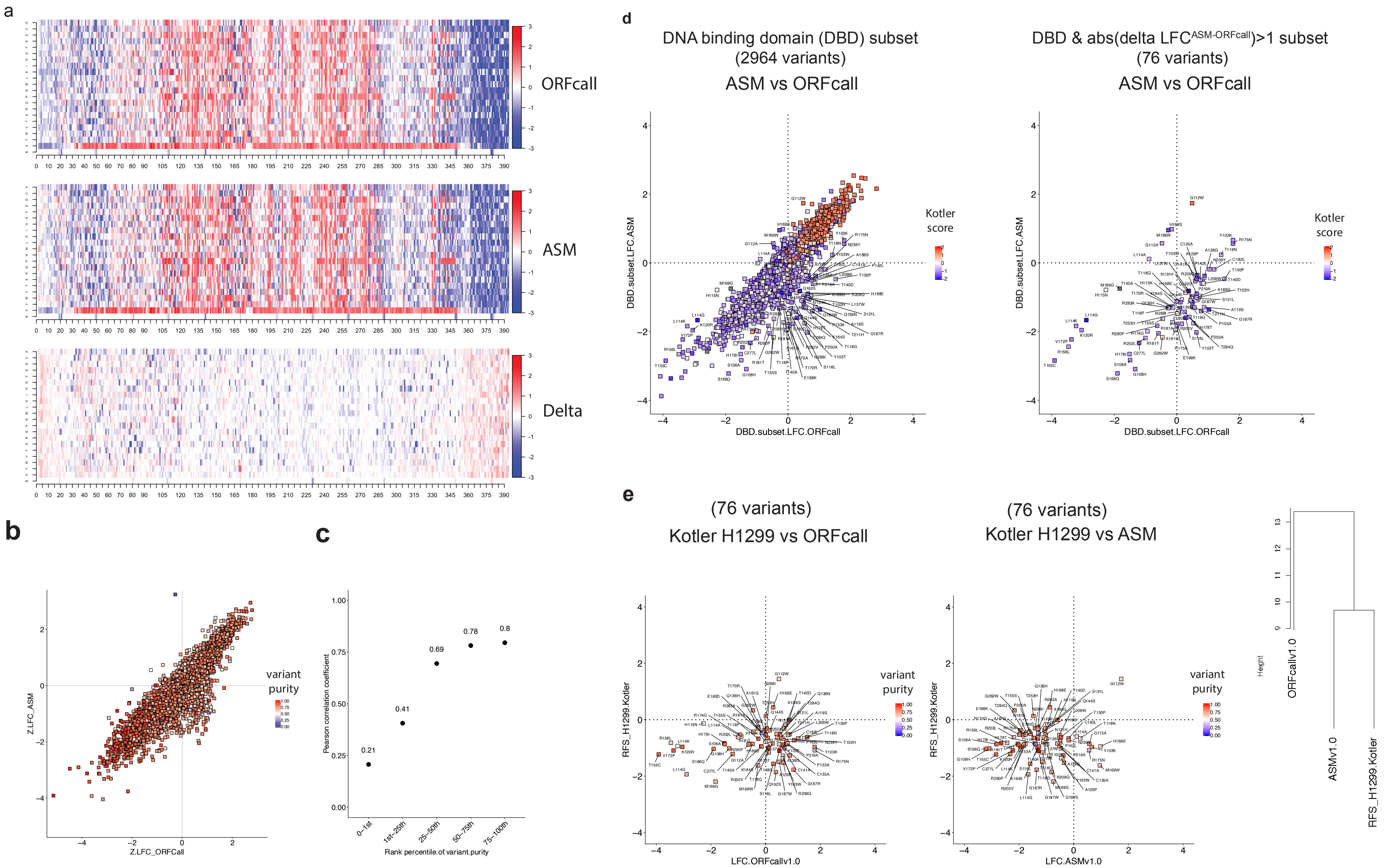
a
ORFcall
ASM
Delta
0 10 20 30 40 50 60 70 80 90 105 120 135 150 165 180 195 210 225 240 255 270 285 300 315 330 345 360 375 390
3
2
1
0
-1
-2
-3
b
Z.LFC_ASM
Z.LFC_ORFCall
variant purity
1.00
0.75
0.50
0.25
0.00
c
Pearson correlation coefficient
0.21
0.41
0.69
0.78
0.8
0–1st
1st–25th
25–50th
50–75th
75–100th
Rank percentile.of.variant.purity
d
DNA binding domain (DBD) subset
(2964 variants)
ASM vs ORFcall
DBD & abs(delta LFC$^{ASM-ORFcall}$)>1 subset
(76 variants)
ASM vs ORFcall
DBD.subset.LFC.ASM
DBD.subset.LFC.ORFcall
Kotler score
e
(76 variants)
Kotler H1299 vs ORFcall
(76 variants)
Kotler H1299 vs ASM
RFS_H1299.Kotler
LFC.ORFcallv1.0
LFC.ASMv1.0
variant purity
Height
ORFcallv1.0
ASMv1.0
RFS_H1299.Kotler

Figure 2. Reanalysis of *TP53* saturation mutagenesis screen using ASMv1.0. Screens most often involve the comparison of a sample subjected to selective pressure with a sample processed under a reference condition. This Giacomelli *et al*[22] 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 log fold change (LFC). The pair of heatmaps are nearly identical to the naked eye. 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. The colors are scaled by variant purity. (**c**) Demonstrating the impact of variant purity on the improved performance of the new ASMv1.0 software. The new software's massive output allows us to compute 'variant purity' defined by the fraction of pure planned variant over all variants that harbor the planned variant, with or without unintended changes in the same sequencing reads. 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. It is clear that high variant purity results in similar LFC values, whereas as the purity decreases, the two sets of fold-changes diverge. The new software's ability to detect and tally variants that carry additional unintended changes is called for when the variant purity is low. Similarly, we

analyzed two other Giacomelli *et al*[22] screens of the *TP53* saturated ORF variant library and the results are presented in **Supplementary Figure 3** and **4**. (**d,e**) Validation of ASMv1.0 using the published data by Kotler *et al.*[25] The Kotler dataset includes a screen of p53$^{\text{NULL}}$H1299 cells with a library covering *TP53* DBD. We first took a subset of Giacomelli library to cover the same *TP53* region as Kotler screens, resulting in a total of 2964 variants (**d**, **left**). We then identified 76 variants that recorded most different LFC scores by our two versions of software **(d, right**). We then compared the Kotler scores against our LFC scores called by ORFCallv1.0 (**e, left**) and ASMv1.0 (**e, middle**). We observed that between two versions of software, ASMv1.0 calls are clearly more in agreement with Kotler calls (**e, right**).

Figure 3

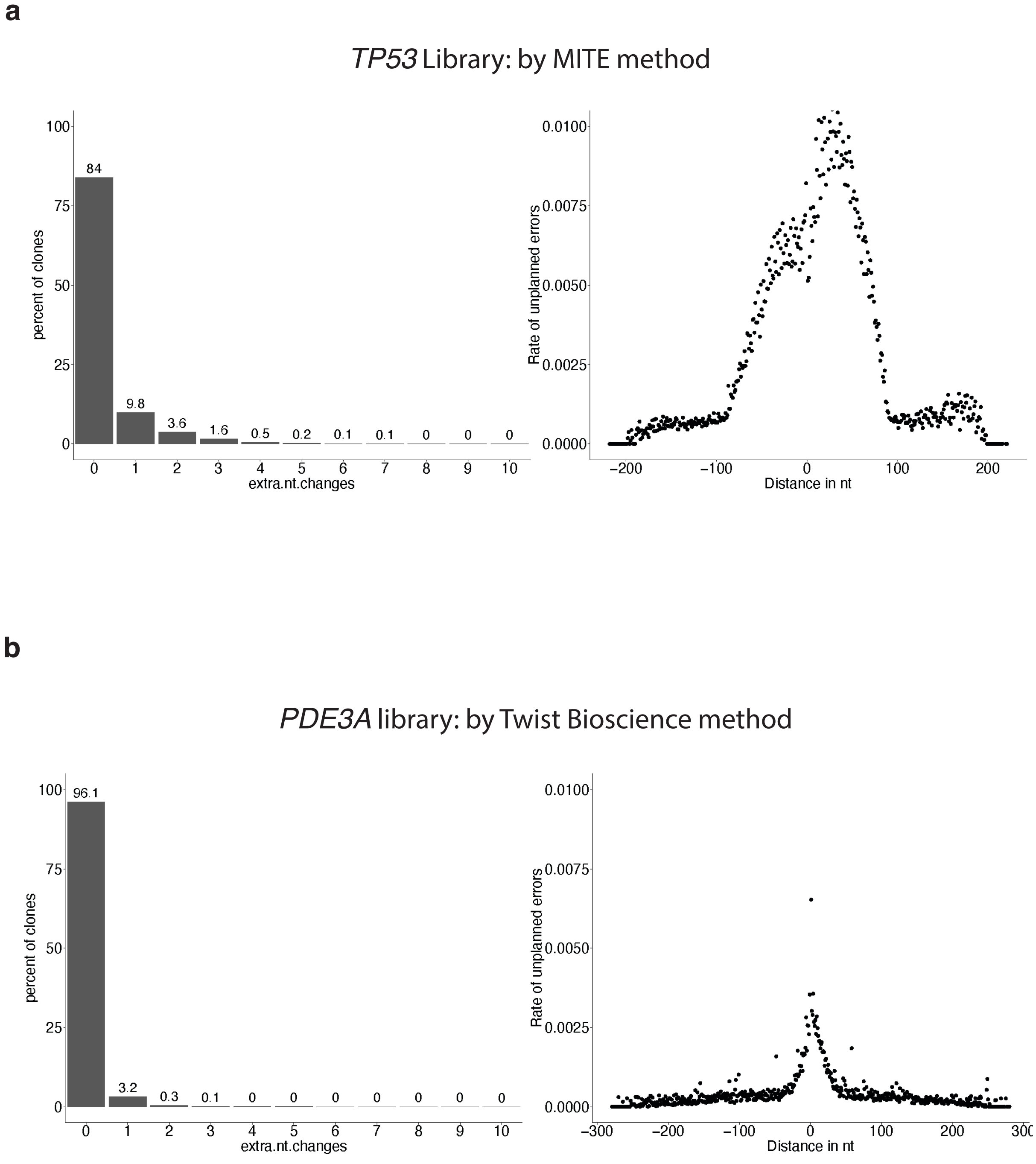

Figure 3. The errors in our saturated ORF variant libraries are low in number and local relative to the location of the planned codon change. In screens, the mutational spaces of libraries depend on how many planned variants can be detected in the final sequencing data. A

library with a high impurity rate will require a larger experimental effort. (**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 strikingly explainable by the oligo synthesis scheme involved in these technologies. MITE technology uses synthesized oligos of 150 bases long, of which 90 bases (called TILE) go into the final products, whereas Twist Bioscience introduced errors through oligo 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. This is important because the close proximity of the errors to the intended codon change renders the planned change and errors all detectable in a single read (e.g. 150 bases) or read pair (e.g. 2x150=300 bases).

Figure 4

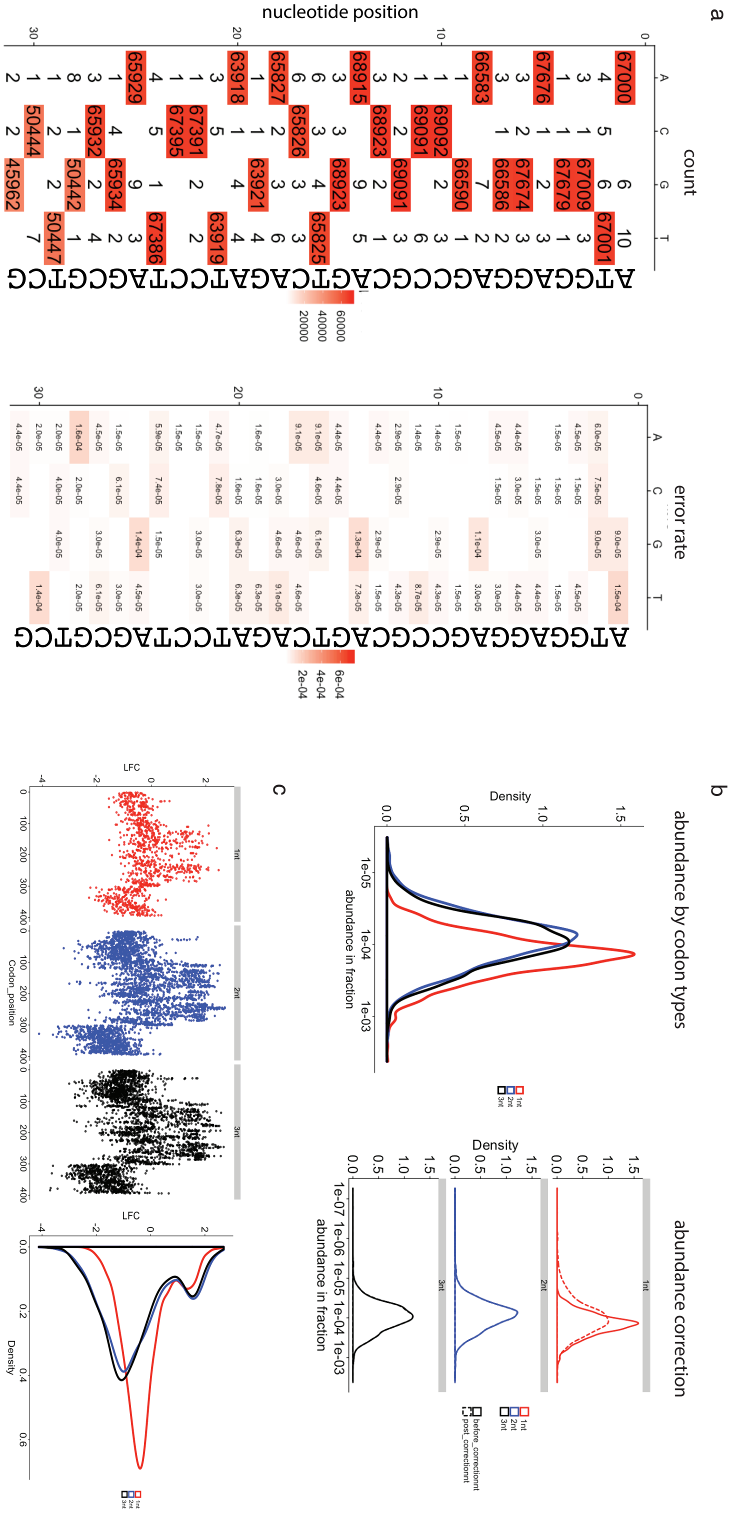
a
count
nucleotide position
A
C
G
T
ATGGAGGAGCCGCAGTCAGATCCTAGCGTCG
60000
40000
20000
error rate
6e-04
4e-04
2e-04
b
abundance by codon types
Density
abundance in fraction
1e-05
1e-04
1e-03
1nt
2nt
3nt
abundance correction
1e-07 1e-06 1e-05 1e-04 1e-03
before_correctionnt
post_correctionnt
c
LFC
Codon_position
Density

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. 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 in the library. The heatmap on the left shows tallied counts of called nucleotides at each nucleotide position; 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**) The miscalls resulting from PCR 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 of 1-nt (not 2- or 3-nt) difference from the reference codons. 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 artificial 'variant' calls resulting from PCR 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. (**c**) **Suppression of fold-changes in 1-nt delta codon variants.** PCR errors in the workflow inflate the counts of variants that differ from the template codon by one nucleotide. Using the pMT025-TP53 library as an example, 1-nt delta variants showed much narrower fold-change values than 2-nt and 3-nt delta variants. Thus, in the library design stage, when we have choices of 2- or 3-nt delta codons, we pick them over 1-nt delta codons. However, we do admit 1-nt delta codons when they are the only choices available; without them, the library would have gaps in coverage. The fold-change suppression in the 1-nt delta group leads to false negative calls. In order to mitigate this error, correction of variant counts with data from a carefully constructed clonal plasmid control experiment may be helpful

for 1-nt delta variants. Altogether, we have arrived at an important rule for saturated ORF variant library design: minimizing the use of codons that are 1-nt away from the reference codons.