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

24 [13]Current address: Novartis Institutes for Biomedical Research, Oncology Innovative Targets and
25 Technologies, 250 Massachusetts Ave, Cambridge, MA 02139, USA
26
27 Correspondence should be addressed to D.E.R. (droot@broadinstitute.org)
28

29 **Abstract**
30 Saturated variant libraries that encode all single-amino-acid substitutions allow one to test all
31 variants of a gene simultaneously in a single experiment. Because a variant differs from its
32 reference at a single codon-position located anywhere along the coding sequence, variant
33 detection presents a profound technical challenge. A popular but expensive brute-force approach
34 is to divide a library into sub-libraries for direct sequencing. Here we present our approach that
35 features enhanced mutant-detection software that calls variants reliably from massive shotgun
36 sequencing data, and methodologies that enable one to design, construct and screen a single
37 saturated library per gene of any size.
38

39 **Keywords**: saturation mutagenesis, deep mutational scanning (DMS), variants, log-fold change,
40 base miscall, DNA polymerase error rate, next-generation sequencing (NGS).
41

42 **Background**
43 An open reading frame (ORF) variant library of a gene of interest provides a powerful
44 means of systematically characterizing the functions of over thousands of single-amino-acid
45 substitutions in a single experiment. The functional characterization of variants of an ORF has
46 historically been confined to alanine scans[1, 2], random mutagenesis by error-prone PCR[3] or

47 mutator *E. coli* strains[4], site-directed mutagenesis[5] , or transposon insertional
48 mutagenesis[6]. Advances in DNA oligonucleotide and gene synthesis[7] and in next-generation
49 sequencing[8] have accelerated our ability to assess all possible single-amino-acid changes for
50 every amino acid position of a protein. The resulting variant collections are known as saturation
51 mutagenesis libraries, or deep mutational scanning libraries. A well-executed screen of such a
52 library, that is, a pool of ORF variants, produces a comprehensive dictionary that maps variants
53 to phenotype strength.

54 As illustrated in **Figure 1**, a typical pooled screen allows one to test the functions of
55 thousands of genetic perturbation agents in a single experiment. Pooled screens rely on
56 introduction of one genetic perturbation cassette, e.g., an shRNA, a CRISPR sgRNA or an ORF
57 variant, per cell into the genome using delivery by a retrovirus. The perturbed cell population is
58 then subjected to selection pressure, for example, growth-over-time proliferation, treatment with
59 a drug, or flow-sorting for a reporter marker. Eventually, the collected cell population samples
60 are processed to measure the abundance of each library perturbation in each of the samples. The
61 perturbagen counts are the basis to assess the enrichment or depletion of each library
62 perturbation by comparing samples that have undergone selection and samples that have been
63 treated with reference conditions (e.g., pre-selection or vehicle treatment samples). In shRNA or
64 CRISPR screens, the functional elements, hairpins or guides respectively, serve as the perfect
65 barcodes, short and unambiguous, for screening sample deconvolution. In contrast, ORF variant
66 library screens are confronted with challenges in variant detection.

67 The complexity of a saturated ORF variant library is a function of the ORF length and the
68 number of desired variants per codon position. As illustrated in **Supplementary Figure 1**, at a
69 given amino acid position, the template amino acid is typically mutated to 19 other amino acids,

70 a stop codon, and when design rules allow, some synonymous codons (**Methods**). Factoring in
71 the amino acid positions to be mutagenized (N), the library is a collection of approximately 20*N
72 variants, or alleles, each differing from the template ORF at a single amino acid position. The
73 coordinates of all variations in a library are spread out along the full-length ORF. Thus,
74 capturing the variant-defining signature as subtle as a single codon variation located anywhere
75 along the full-length ORF sequence is one of the foremost technical challenges. Over the years,
76 admirable efforts and steady progress have been made in dividing a library into sub-libraries
77 each with mutational regions of initially 25[9, 10], then 47[11], 100[12], and recently 170 [13]
78 amino acids. These sub-libraries were then screened and sequenced separately. In some cases,
79 these sub-libraries were further divided into tiles to allow the direct next-generation sequencing
80 readout of variants[13]. The benefits of these approaches are limited to and deemed practical to
81 small ORFs only. Our methodology, on the other hand, emphasizes robust variant-calling tools
82 that can process massive shotgun NGS data to detect and accurately quantify variants, and
83 consequently allows one to screen a single saturated variant library for an ORF of any size (e.g.,
84 ~600-bp *KRAS[14],* ~1200-bp *TP53[15]* and *SMARCB1,* ~1400-bp region of *PDE3A[16]*,
85 ~1700-bp *SHOC2[17]*, ~1800-bp *PPM1D[18]*, ~2200-bp *EZH2*, ~3600-bp *EGFR*) in a single
86 experiment (**Fig. 1b**). Here, we present our methods and strategies as we draw from our ever-
87 evolving knowledge from designing and constructing over 20 libraries and analyzing the
88 resulting screening data (**Supplementary Table 1**). Furthermore, we have made our library
89 designer, variant-calling software, and data analysis tools freely available (**Supplementary**
90 **Table 2)**. We hope that this study will help lower the access barrier to this technology and pave a
91 path for sharing library resources within the scientific community.
92

Results

Overview of variant-detecting software ASMv1.0.

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. Improving upon our earlier version of variant-calling software, we developed 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 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 PCR primers used to amplify the full-length ORF from genomic DNA. It counts the number of observed molecules that have a 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 the 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 uses an R toolkit to parse ASMv1.0 output files of all screen samples into a single datafile (see **Methods**). Throughout these processes, the toolkit preserves the variant description, which unambiguously defines the called variant and which alone allows the resctruction of mutations in the variant. In

115 our experience, the detail-rich variant descriptions and the annotations derived from them are
116 invaluable datasets that allow us to perform a myriad of in-depth data analyses.
117
118 **The enhanced performance of ASMv1.0.**
119 The early version of variant detection software, ORFCallv1.0, was successfully used in
120 analyzing many saturation mutagenesis screens[15, 19-21]. ORFCallv1.0, however, tallies the
121 intended variants in the space of a single codon, and it ignores the association of an intended
122 codon change and any *in-cis* unintended changes. For a library that has prevalent unintended
123 errors associated with an intended change, ignoring the presence of these errors can mask the
124 true phenotype of the intended variant by the phenotypes caused by the unintended changes. As
125 an improvement, our new software, ASMv1.0, calls variants in the context of full-length read
126 pairs, allowing us to separate the variants that bear only the intended sequence change from those
127 that also harbor extra changes. Furthermore, in ASMv1.0, a pair of reads needs to pass through
128 sequential filters such as minimum base quality, minimum read length, and requirement of
129 consistency in overlapping sequences of a pair of reads, before being admitted for variant calling
130 (**Supplementary Fig. 2)**. **Supplementary Table 3** summarizes the enhanced functionalities
131 from ORFCallv1.0 to ASMv1.0.
132 We used the screens performed in Giacomelli *et al*[15] to compare the two methods.
133 Specifically, the ectopically expressed *TP53* variants were screened in isogenic $p53^{WT}$ and
134 $p53^{NULL}$ A549 lung cancer cells, with selection pressure from either nutlin-3 or etoposide. Under
135 nutlin-3 treatment, the screen of $p53^{NULL}$ cells enriched *TP53* loss-of-function (LOF) variants,
136 and depleted wild-type-like variants (**Fig. 2**). Under etoposide treatment, the $p53^{NULL}$ cells
137 enriched wild-type-like *TP53* variants and depleted the *TP53* LOF variants (**Supplementary Fig.**

138 **3**). Under nutlin-3 treatment, the screen of $p53^{WT}$ cells enriched *TP53* variants exhibiting
139 dominant-negative effects (DNE) (**Supplementary Fig. 4**). Using both the ORFCallv1.0 and
140 ASMv1.0 variant calling methods, we reprocessed the next-generation sequencing data from
141 these three screens. The phenotypic strength of each variant, measured as log-fold change (LFC)
142 in the pre- and post-treatment samples, produced similar heatmaps, indicating that the two
143 variant-calling algorithms are largely consistent with each other (**Fig. 2a, Supplementary Figs.**
144 **3a, 4a**). However, we also noted that several variants exhibited substantial differences in LFC
145 calculated using each algorithm; we define the differences as delta(LFC)=LFC^{ASM}-$LFC^{ORFCall}$ in
146 **Figure 2b, Supplementary Figure 3b** and **4b.**
147 In a perfect library that consists solely of intended variants, ORFCallv1.0 and ASMv1.0
148 should be equal in performance. However, it is inevitable that some planned codon changes are
149 physically associated with one or more unintended nucleotide changes in the same molecule. We
150 define the 'variant purity' of a planned variant as the fraction of pure intended molecules over all
151 molecules that harbor the codon change of interest, whether or not they harbor additional
152 nucleotide changes. When the purity of a planned variant is high, as in a high-quality library,
153 both versions of the software will call variants with similar accuracy. However, when the variant
154 purity is low, the enrichment and depletion scores measured by ORFCallv1.0 can be misleading,
155 as they reflect the sum of effects of all species that happened to have a given codon change. We
156 computed the purity of every planned variant using the reference samples of the Giacomelli
157 screens, and for each purity bin we computed the Pearson correlation coefficient of two sets of
158 LFCs, one using ORFCallv1.0, the other using ASMv1.0 (panel **c** of **Fig. 2, Supplementary**
159 **Figs. 3,4)**. The two methods correlate well when the variant purity is high and diverge as the
160 variant purity decreases.

161 We therefore reasoned that ASMv1.0 is an upgrade over ORFCallv1.0, particularly
162 considering how unintended *in cis* changes are handled. As a way of validating the higher
163 performance of ASMv1.0 over ORFCallv1.0, we compared the Giacomelli screen LFC scores
164 called by both versions of the software with the results from the Kotler *et al*[12] that used a small
165 variant library that focused on the p53 DNA binding domain (DBD) and screened in $p53^{NULL}$
166 H1299 cells (**Fig. 2d**).
167 Kotler *et al* targeted p53 DBD with 4 libraries, each covering the variants of a 47-amino
168 acid region of the ORF. The Kotler screens were processed by direct sequencing readout of the
169 mutational region of each of the 4 libraries. Furthermore, Kotler *et al* processed their screen with
170 stringent analysis thresholds by requiring (1) at least 80bp overlap of the two reads in a read pair,
171 (2) perfect sequence agreement in the overlapping region, and (3) minimum read coverage
172 of >200 reads for each variant. These stringent thresholds and the direct readout of variant counts
173 rendered the Kotler dataset the best available for, independently, validating the ASMv1.0
174 algorithm.
175 Similar to the Giacomelli $p53^{NULL}$ A549/nutlin-3 screen, the Kotler $p53^{NULL}$ H1299 cell
176 screen enriched for *TP53* alleles that exhibit LOF, and depleted wild-type-like variants. From the
177 7890-variant Giacomelli $p53^{NULL}$ A549/nutlin-3 screen data, we extracted a 2964-variant subset
178 that covered the same mutational space as the Kotler H1299 screen. Among these 2964 variants,
179 we identified 76 variants from the Giacomelli $p53^{NULL}$/nutlin-3 screen that were called
180 differently by ORFCallv1.0 and ASMv1.0, with abs(delta(LFC))>1 (**Fig. 2d, right**). Of these
181 differentially scored variants, the LFC scores called by two versions of the software were
182 individually compared with the Kotler screen scores (**Fig.1e).** The ASMv1.0 LFC scores for the

183 Giacomelli p53^{NULL}A549/nutlin-3 screen agree with the Kotler calls better than the ORFCallv1.0
184 LFC scores do (**Fig. 2e**).
185
186 **ASMv1.0-enabled analysis reveals that the secondary mutations (errors) in the library are**
187 **near the targeted mutation and introduced by errors in oligo synthesis.**
188 The ability of ASMv1.0 to detect molecules with intended or unintended changes along
189 the entire read pair allows one to conduct in-depth analyses of each saturated ORF variant
190 library. Thus, in addition to the variant purity discussed earlier, we can also tally and characterize
191 the unintended nucleotide changes to profile the errors introduced during library construction.
192 For example, we were able to determine that in the *TP53* library, 84% of molecules are pure and
193 planned variants (**Fig. 3a left**), and the errors in the library are predominantly single nucleotide
194 changes.
195 Taking one step further with the ASMv1.0 output, we investigated the proximity
196 distribution of errors in relation to the intended codon change. In **Figure 3**, we compared
197 libraries made by two different methods (see **Methods**): the *TP53* library made by MITE
198 (Mutagenesis by Integrated TilEs)[21], and the *PDE3A* library synthesized by Twist Bioscience.
199 The MITE method involved oligo synthesis to cover the length of 'tiles' that are 90 nucleotides
200 long. The Twist Bioscience technology involved the synthesis of oligos that vary, depending on
201 GC content, between 30-50 nucleotides and with the variant-encoding bases in the middle. In
202 **Figure 3b,** we profiled the error rate by distance, in nucleotides, between the locations of the
203 error and the planned change. This distance profile yields an important observation that have
204 shaped our screen deconvolution sequencing strategy, which will be discussed in detail later in
205 this section:

206 The stretches of frequent errors in libraries produced by the two methods, ~90 bases for
207 the MITE method and ~30 bases for the Twist Bioscience method, correspond to the length of
208 chemically synthesized DNA oligonucleotides involved in library production. This indicates that
209 the errors in the libraries are concentrated near the planned codon changes and sequencing reads
210 as short as 100 bases can detect both intended and unintended changes in the library.
211 Furthermore, it reasons that we can focus on the reads that carry expected variant
212 signatures to tally variant counts. Without exception, each full-length ORF variant in the library
213 consists of nucleotides inherited from mutation-introducing DNA strands (e.g., primers) and
214 nucleotides synthesized by high-fidelity DNA polymerase. The errors in the library are dictated
215 predominantly by the former, as it has much higher error rate than DNA polymerase. A high-
216 fidelity DNA polymerase with an error rate of 10^{-6} errors per cycle per amplified nucleotide will
217 produce a small number of erroneous molecules (1.1% for a 1kb ORF; 4.3% for a 4kb ORF),
218 which amount to a small fraction of those introduced from oligo synthesis (10% or 18% between
219 two library production methods) (**Supplementary Table 4**). We therefore argue that by focusing
220 on reads that carry expected mutations, and with NGS reads of 100 bases or longer, we
221 effectively capture the errors in the library.
222
223 **ASMv1.0-enabled analysis allows assessing the effect of miscalls (artifactual errors)**
224 **introduced by processes involving PCR.**
225 While the errors discussed above are true errors that are present in the libraries, it is
226 important to address the artifactual errors in the form of miscalls introduced by the variant
227 quantification process itself. In the screen deconvolution process, PCR was used to replicate the
228 ORF DNA from genomic DNA (up to 30 PCR cycles) and used again in NGS sample

229 preparation (12 cycles). The errors introduced by these PCR steps are artifacts. We define
230 'miscalls' as the artifactual and mispresented base calls relative to what are in the samples. The
231 error rate of DNA polymerases can be as low as one error per one million bases synthesized in a
232 PCR cycle. When we, in parallel with the screening samples, processed and sequenced the
233 reference *TP53* ORFs that are clonal, pure, and without any variants, the sequencing results are
234 striking – at each nucleotide position, we observed miscalls at a rate of ~0.0003-0.0006 miscalls
235 per nucleotide read-out (**Fig. 4a, right**), or 3-6 miscalls per 10,000 sequenced bases.

236 While this miscall rate may appear small, it is a significant factor in variant detection. As
237 an example, illustrated in **Supplementary Figure 5**, these miscalls will lead to (1) inflated
238 counts of variants with codons that are one-nucleotide away from the template codon (we refer to
239 this variant type as '1-nt delta' codon), and (2) the artifactually narrowed log-fold change range
240 of 1-nt delta codons (**Supplementary Fig. 5,** legend**)**. Indeed, in the *TP53* library screen with
241 $p53^{NULL}$ cells/nutlin-3, we observed the inflation of counts (**Fig. 4b**), and the suppression in fold-
242 change (**Fig. 4c**), in the 1-nt delta variant group relative to 2- or 3-nt delta variants. This effect of
243 PCR errors on 1-nt delta codon changes has been observed, without exception, in all libraries we
244 have processed, and the effect magnifies as ORF length increases.

245 A recent report[22] presented software DiMSum that used 1-nt delta codons to model
246 PCR/NGS errors as a method to assess library quality. We, however, took a direct approach and
247 opted to mitigate the effects of PCR/NGS miscall errors by implementing library design rules
248 and adding control experiments. First, we set an important rule for saturated ORF variant library
249 design - minimizing the use of codons that are one nucleotide away from the reference codons.
250 In our experience, to achieve 'saturated' amino acid coverage in a library, we cannot completely
251 avoid variant codons that are one nucleotide away from a reference codon. But we can minimize

252 the use of 1-nt delta codons, reducing them from 15% to 3% of the library members. Secondly,
253 dealing with these 3% 1-nt delta variants, we added a clonal template plasmid sample and
254 processed it as if it were a screening sample. One may opt to use the clonal template data to
255 monitor or even correct the screen sample data (**Fig. 4b**, right).
256
257 **Silent variants are used as a measure of screen baseline.**
258 Another rule we have implemented for library design is the inclusion of silent variants
259 with codons that are 2- or 3-nt different from the reference codon. ORF variant library screens
260 often produce many hits. As a result, one cannot always rely on the large portion of the library
261 producing no phenotype to serve as the baseline. It is therefore essential to include silent codon
262 changes in the library; such nucleotide changes allow us to identify and quantify the silent alleles
263 in parallel with missense and nonsense variants (see rows marked as 'B' in all long-fold-change
264 heatmaps). Often, codons encoding an amino acid differ from one another at the wobble base. As
265 a result, many silent codons are one nucleotide away from the template codon. Requiring that
266 any admitted silent variants have a 2- or 3-nt difference relative to the reference codon results in
267 silent variants comprising about 2% of the library.
268 In the *SMARCB1* library, we included nearly all possible silent variants, including 1-nt
269 delta silent variants. As expected, due to the artifactual miscalls discussed earlier, the abundance
270 of 1-nt delta silent variants is inflated relative to the abundance of 2- and 3-nt delta silent variants
271 (**Supplementary Fig. 6a**). For the same reason, the 1-nt delta silent variants show an
272 artifactually narrowed log-fold change range, compared with 2- and 3-nt delta groups
273 (**Supplementary Fig. 6b**). In our current library designer download (**Supplementary Table 2**),
274 we avoid 1-nt delta silent variants and minimize 1-nt delta missense and nonsense variants.

275

276 **A justification for using next-generation sequencing (NGS) over long-read sequencing**

277 **(LRS).**

278 An often-debated topic is how saturation mutagenesis screen samples should be

279 sequenced. It would seem evident that screen samples should be sequenced with long reads to

280 widen the space of variant characterization and to capture all mutations present on each

281 molecule. However, the technology that can capture the readout of full-length ORF

282 sequences[23, 24] is currently cost-prohibitive for most labs, and more importantly, undesirable

283 for variant detection due to low concordance rate of base calls[25]. For example, in a study by

284 Wenger *et al*.[26], circular consensus sequencing (CCS), the best of its kind, has a concordance

285 rate of 99.772%. This is equivalent to a Q26 Phred score, producing one discordance per 439-

286 base readout. This discordance rate is higher than the error rate in commercial DNA

287 oligonucleotide synthesis, the key contributor of the real errors in the libraries. Since the

288 measurement discordance rate is higher than the underlying error rate which it is expected to

289 measure, the CCS read attrition rate will be high. As a result, one may be forced to either throw

290 out many reads that show secondary mutations or focus locally on the expected codon changes

291 and ignore mutations in the distance.

292

293 It is however encouraging that the long-read sequencing technology is advancing at a

294 steady and rapid pace, and its applications are actively explored. In the recent work by Frank *et*

295 *al*.[27], long read sequencing was used to associate a set of unique external barcodes to each of

296 ~7000 variants in a saturation variant library, so that the screen deconvolution work can be

297 achieved by direct next-generation sequencing of the barcodes. However, for four prevailing

298 reasons, we favor a sequencing platform capable of producing NGS read pairs of 150 bases per
299 read with a Phred score of Q30 or more for each base call, which is comfortably achievable by
300 many NGS platforms: (1) We have demonstrated that the errors in our libraries are near the
301 planned codon changes (**Figure 3, right**), and a 150-base read is adequate to detect both the
302 variant-defining sequences and library errors from DNA oligonucleotide synthesis. (2) We have
303 shown that it is unnecessary to sequence an ORF variant far into its regions synthesized by high-
304 fidelity DNA polymerase during library construction, as the errors introduced by DNA
305 polymerase in regions outside of a 100-base window around the planned codon changes are rare
306 and negligible. Therefore, we rightfully focus on reads that bear a planned codon change, which
307 effectively represents the variant's error-harboring region, that is, the region that was chemically
308 synthesized. (3) We have demonstrated that PCR-introduced miscalls produce predominantly 1-
309 nt delta codon changes. As we minimize the use of 1-nt delta codon variants, the artifactual
310 variants resulting from the miscalls are mostly unintended variants. By ignoring reads carrying
311 only unintended variations, we largely removed the effects of sequencing miscalls. (4) ASMv1.0
312 allows users to set adequate base quality threshold, for example, at Phred score Q37 with
313 Illumina Novaseq NGS. With that, we can comfortably achieve <1 discordance per 5000-base
314 NGS readout. As we focus on the reads that carry an intended-variant signature, at this low
315 discordance rate, >94% (0.9998^300) of 150-base read pairs used to call variants are free of
316 miscalls.
317

318 **Discussion**

319 Variant detection software ASMv1.0 enabled a cascade of in-depth data analyses that
320 allowed us to explore, for instance, variant purity, the rate and nature of library errors, and the

321 artifactual miscalls. The results of these analyses have shaped and reshaped our strategies at
322 many key steps of the projects, on both laboratory and computational fronts. We believe that the
323 analytical and laboratory technical blueprint presented here, along with our library designer,
324 variant-calling software, and data analysis tools will help enable the scientific community to
325 fully utilize this technology in research applications.

326 It should be noted that the strong hits among the 1-nt delta variants may still score well,
327 even with the miscall-related log-fold change suppression. The issue of 1-nt delta codon variants
328 in processing screening samples stems from the fact that we sequence the ORF to quantify the
329 variants. The effect of miscalls on 1-nt delta variant counts becomes less of an issue for short
330 ORFs (e.g., *KRAS*). As illustrated in **Supplementary Figure 5**, the longer the ORF, for each
331 readout of a variant-defining codon, the more copies of the template codon need to undergo
332 PCR-based amplification and NGS, and consequently the more counts of the artifactual 1-nt
333 delta codons will result. Our suggestion of minimizing 1-nt delta codon variants in ORF variant
334 libraries is to insulate the true variant counts in the screening samples from PCR-introduced
335 artifactual miscalls. If one prefers 1-nt delta codons in the variant libraries, we recommend
336 adding the clonal template ORF as a control sample and processing it in parallel with the
337 screening samples. This control sample may be used as a count-correction factor (as shown in
338 **Fig. 4b**, right).

339 The recent establishment of an open-source platform, MaveDB, to enable variant
340 function data sharing, is a significant development[28]. We further advocate that the scientific
341 community find a way to share saturation mutagenesis reagents and resources. In many cases, the
342 content of a saturated ORF variant library is well defined by the gene reference sequence. For a
343 given gene, the library designed by one lab will be highly similar to one done by another lab. As

344 a perfect example of resources sharing, Boettcher *et al*[29], using the *TP53* saturated ORF
345 variant library created by Giacomelli *et al*[15], quickly demonstrated that the dominant-negative
346 activity of *TP53* missense mutations in DNA binding domain was selected for in myeloid
347 malignancies. If the scientific community can come to an agreement for library sharing and
348 repurposing, the available resources could be used to cover more genes, as opposed to building
349 redundant libraries. Imagine what a collection of saturated ORF variant libraries of 1000 disease-
350 related genes could do!
351
352 **Conclusions**
353 Saturated ORF variant libraries are powerful tools that enable one to exhaustively test all
354 possible protein variants simultaneously in a single experiment. Each experiment produces a
355 comprehensive dictionary of variants and functions. Yet, this technology has high access barriers
356 as it demands expertise across many disciplines, and the costs of producing and utilizing the
357 libraries are prohibitive to many academic labs. This report presents the scientific community a
358 full blueprint and a complete set of tools of conducting saturation mutagenesis projects. We are
359 certain that it lowers the technical barrier of this technology.
360
361 **Methods**
362 **ORF expression vector.** The vector we use to clone most of our saturated ORF variant libraries
363 is a lenti-viral vector pMT025 (**Supplementary Fig. 7**). It is suited for restriction/ligation
364 cloning using NheI/BamHI, or NheI/MluI restriction enzyme pairs. The vector has a puromycin
365 resistance selection marker driven by an SV40 promoter. After the cloning, the ORF expression

366 is driven by EF1a. To facilitate cloning, we require that the ORF sequences be properly flanked
367 with constant sequences (in which restriction enzyme sites for cloning are highlighted):
368
369 GGTTCAAAGTTTTTTCTTCCATTTCAGGTGTCGTGAG**GCTAGC**GCCACC [ORF with
370 ATG start and stop]**GGATCC**CGGGACTAGT**ACGCGT**TAAGTCGACAATC
371
372 **Template ORF plasmid**. The starting library we are using is either in a pUC57 plasmid form
373 (libraries made via MITE method), or arrive as linear fragments (libraries purchased from Twist
374 Bioscience). In the cases where the ORF reference sequence contains restriction sites needed for
375 cloning, the template ORF sequences need to be modified to remove these sites. It is important
376 that prior to making the library, screening assay development work is done with control clones in
377 the same vector as the vector for the future library. It is equally important that a clonal plasmid
378 of the template ORF is first cloned into the vector, made into virus, tested for infection, genomic
379 DNA prep, and PCR with the gDNA. This clonal plasmid will be used as the template for
380 mutagenesis.
381
382 **Library design**. A saturation mutagenesis designer written in R is available. The template ORF
383 sequence is first checked for absence of restriction sites that are used in cloning: either
384 NheI/BamHI or NheI/MluI. The ORF template may need to be modified in order to satisfy these
385 requirements. The designer then steps through from N-terminus to C-terminus of the template
386 ORF one amino acid position at a time to select variant codons. For a given template codon there
387 are 63 possible variant codons to choose from at each amino acid position. These 63 codons are
388 divided into two bins: 54 codons that have 2- or 3-nucleotide differences relative to the template

389 codon (2-3-nt delta), and 9 codons with a single-nucleotide difference (1-nt delta). The codons
390 that would violate the restriction sites exclusion rule are removed from both bins. In designing
391 missense or nonsense variants, we first select from the 2-3-nt delta codon bin, and then if needed
392 from the 1-nt delta codon bin. If within a bin there are multiple codon choices available for an
393 amino acid variant, the codon is picked by sampling the available codons with probability
394 weightings according to the codon usage frequency of the organism (*Homo sapiens* in all our
395 experiments to date, http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=9606). For
396 silent codon changes we pick only from 2-3-nt delta bin. In the end, each of the designed variant
397 codons substitutes for its corresponding template codon in the full-length reference ORF as one
398 member of the library.
399
400 **Mutagenized ORF Library cloning**. Saturation mutagenesis library production requires that
401 variant codons be encoded into pools of oligos. These oligos are then either amplified and
402 assembled into cloneable linear fragments using the Mutagenesis by Integrated TilEs (MITE)
403 method or delivered as a linear fragment pool (Twist Biosciences). For MITE method, once the
404 linear fragments bearing the mutagenized codons are produced, they are then cloned into a high
405 copy entry vector for amplification and sequence validation. For Twist Bioscience method, the
406 linear fragment pool is a collection of properly flanked full-length ORFs that are ready for
407 cloning into pMT025 ORF expression vector.
408
409 **MITE method**: The region of the open reading frame to be mutated is partitioned into
410 90-base tiles. Oligos representing all mutations in the space of a tile were synthesized. Each 90
411 base tile is flanked by 30 base sequences that are complementary to the reference ORF. Oligos

412 for adjacent tiles should be synthesized on separate arrays. Oligo tile pools were amplified by
413 emulsion PCR using primers to the 30 base flanking regions and purified on a 2% agarose gel.
414 Entry vector cloning is performed on a per tile basis by linearizing the entry vector at the
415 appropriate tile position using Phusion polymerase (New England Biolabs) and primers to the
416 regions flanking the tiles. The linearized vectors were purified on 1% agarose gel and DpnI
417 treated. The DpnI-treated linear plasmid backbones were then mixed with the relevant PCR
418 amplified tile and assembled by *in vitro* recombination with the NEBuilder HiFi DNA Assembly
419 Master Mix (New England BioLabs). The assembly reactions were purified, electroporated into
420 TG1 *E. coli* cells (Lucigen), and recovered for one hour at 37˚C in Recovery Media (Lucigen).
421 Aliquots from the transformations were used to inoculate overnight cultures of LB containing 25
422 µg/mL of Zeocin (ThermoFisher). Cells were harvested by centrifugation and plasmid DNA was
423 isolated using the QIAGEN Midiprep Plus kit (QIAGEN). Plasmid pools, each corresponding to
424 a tile, were verified by sequencing Nextera XT (Illumina). A pool of all pUC57-tile libraries was
425 made with equal weight per variant. This pool was subjected to restriction digest (e.g., NheI and
426 BamHI). After purification by 1% agarose gel, the excised linear fragment library is ready for
427 cloning ligation with an expression vector that was pre-processed with the compatible restriction
428 enzyme pair.
429
430 **Twist Bioscience method**: The mutagenesis library made by Twist Bioscience was
431 delivered as linear fragments of the full-length ORF flanked with adapters for cloning. The
432 fragment library was first digested with NheI and BamHI overnight, and then underwent ligation
433 with pMT025 vector pre-processed with the same restriction enzyme pair.
434

435 **Cloning ligation:** For either method, the ligation was carried out with 5:1 insert-to-vector
436 molar ratio, using T7 DNA ligase at room temperature for 2 hours. The ligation was cleaned up
437 with isopropanol precipitation. The resulting DNA pellet was used to transform Stbl4 bacterial
438 cells. Plasmid DNA (pDNA) was extracted from the harvested colonies using QIAGEN Maxi
439 Prep Kits. The resulting pDNA library was sequenced via the Illumina Nextera XT platform to
440 determine the distribution of variants within the library.
441
442 **Virus preparation and titer measurement.** Lentivirus was produced and quantified by the
443 Broad Institute Genetic Perturbation Platform (GPP). The detailed protocol is available at
444 http://www.broadinstitute.org/rnai/public/resources/protocols. Briefly, 293T viral packaging
445 cells were transfected using TransIT-LT1 transfection reagent (Mirus Bio) with the pDNA
446 library, a packaging plasmid containing *gag*, *pol* and *rev* genes (psPAX2, Addgene), and an
447 envelope plasmid containing VSV-G (pMD2.G, Addgene). Media was changed 6-8 hours after
448 transfection, and virus was harvested 24 hours thereafter. The virus titer was measured by
449 infecting 3000 A549 cells in 96-w cell culture plate with a series of diluted standard virus of
450 known titer for a standard curve, and a series of diluted library virus of unknown titer to find data
451 points that lie in the linear range of the standard curve. Twenty-four hours post infection, the
452 cells were subjected to puromycin selection for 2 days. The surviving cells were quantified with
453 alamarBlue.
454
455 **Determine cell numbers to start and finish the screen.** In general, infections were performed,
456 per infection replicate, to aim a representation on average of 1000 cells carrying a given variant
457 throughout the entire experiment. At the time of splitting for drug-treatment and reference

458 sample pair, the same rules apply. For positive selection screens, where a majority of the library
459 will be depleted and a small number of cells carrying variants that are a subset of the library will
460 survive, and consequently a lower cell number will be collected. In negative selection screen
461 where the majority of the library will be present, the end-point samples should be of at least 1000
462 cells per variant (i.e., per library member).
463
464 **Genomic DNA preparation.** Upon the completion of the screen, cell pellets were collected.
465 Genomic DNA was extracted from cell pellets using the Machery Nagel NucleoSpin Blood Kit
466 (Midi, Clontech Cat. #740954.20; XL Maxi, Clontech Cat. #740950). The detailed protocol is
467 available at http://www.broadinstitute.org/rnai/public/resources/protocols.
468
469 **Design PCR primers**. The integrated ORF in genomic DNA will be amplified by PCR. The
470 PCR products will be shotgun sheared with transposon, index labeled, and sequenced with next-
471 generation sequencing technology. Each sequenced fragment is generated by 2 transposon
472 insertion events so that both ends of the fragment are tagged with a transposon. Therefore, the
473 ends of the gDNA PCR products will be less frequently represented in the NGS library and will
474 have shallow coverage. For this reason, we advise that the PCR primers be designed in such way
475 that there is a ~100 bp extra sequence at each end leading up to the mutated ORF region. For
476 pMT025 vector, we use these 2 primers:
477
478 Forward: 5'-ATTCTCCTTGGAATTTGCCCTT-3'
479 Reverse: 5'-CATAGCGTAAAAGGAGCAACA-3'
480

481 **Determine the number of PCR reactions per sample.** For positive screen end-point samples,
482 the number of cells that need to go through ORF-extracting PCR is determined in such a way
483 that for each library variant, gDNA from 500 or more cells will be subjected to PCR. For
484 reference samples and negative selection screen end-point samples, the cells/variant ratio should
485 be maintained at 1000 or more. PCR is sensitive to the amount of gDNA input and prior to
486 screening, these conditions must be optimized. For example, in 50ul PCR reactions, we test
487 different amounts of gDNA, using a series of 250ng, 500ng, 1ug, 2ug. This titration will help to
488 determine the highest concentration of DNA that can be used to produce robust PCR products.
489 In general, for human cell lines, we assume a cell produces 6.6 pg DNA. Here is the formula to
490 determine how many PCR reactions are needed for a sample:
491
492 $$\text{PCR reactions per sample} = 6.6 * \frac{library\ size}{ng\ of\ gDNA\ per\ PCR\ well}$$
493
494 **ORF amplification from gDNA by PCR.** PCR reactions were set up in 96-well plates
495 according to the optimized PCR condition and number of PCR reactions per sample calculated
496 above. The volume of each PCR reaction was 50 µL and contained an optimized amount of
497 gDNA. Q5 DNA polymerase (New England Biolabs) was used as the DNA polymerase. All PCR
498 reactions for each gDNA sample were pooled, concentrated with a PCR cleanup kit (QIAGEN),
499 and separated by gel electrophoresis. Bands of the expected size were excised, and DNA was
500 purified, first using a QIAquick kit (QIAGEN), and then an AMPure XP kit (Beckman Coulter).
501
502 **Nextera sequencing.** Nextera reactions were performed according to the Illumina Nextera XT
503 protocol. For each sample, we set up 6 Nextera reactions, each with 1 ng of purified ORF DNA.

504 Each reaction was indexed with unique i7/i5 index pairs. After the limited-cycle PCR step, the
505 Nextera reactions were purified with AMPure XP kit. All samples were then pooled and
506 sequenced using an Illumina sequencing platform chosen among Miseq, Hiseq, or NovaSeq,
507 according to the number of reads needed as calculated below.
508
509 **Determine number of reads needed for the screen.** We need to sequence for every variant
510 500x (for positive selection screen samples) or 1000x (for negative selection screen samples, or
511 reference samples) the full-length PCR product. Thus, the number of total reads depends on ORF
512 size, library size, read length, read quality etc. Below are the formulas:
513
514 For negative selection screen samples, or reference samples:
515

516 $$\text{Reads per sample} = 7000 * \frac{(ORF\ length\ in\ NT)\text{^}2}{read\ length * read\ assembly\ rate}$$

517
518 Where '*read length'* is typically 300, as we use paired-end sequencing reads, 150
519 bases each, and '*read assembly rate'* is in range of 0.7-0.9.
520
521 For a positive selection screen sample:
522

523 $$\text{Reads per sample} = 3500 * \frac{(ORF\ length\ in\ NT)\text{^}2}{read\ length * read\ assembly\ rate}$$

524

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

544 To tune our software parameters, we used the sequencing data of a saturation mutagenesis
545 plasmid DNA library for *PDE3A* gene. This library was built by mutagenizing a region of 1425
546 base pairs using Twist Bioscience services. The errors in the library are concentrated in within 30
547 bases of a planned codon change (**Fig. 2**). These errors can be explained by mutagenesis primer

548 synthesis errors. The codon changes are located in the middle of the mutagenesis primers,
549 sandwiched between two 20-30 base pair flanks. In next-generation sequencing, the transposon
550 sheared fragments are, on average, 140 base pairs, similar to the sequencing read length. The
551 close proximity of variants, planned and unintended, indicates that these variants reside on the
552 same transposon-sheared fragment and on the same sequencing read. Therefore, each library
553 ORF, after transposon shearing, produces one variant-bearing fragment; the rest of the fragments
554 are all wild-type. That is, if we ignore miscalls introduced by PCR, the wild-type/variant read
555 ratio of this library should be: 1425/140 -1 = 9.2. This ratio, 9.2, represents the highest ratio we
556 could observe, as any PCR miscalls would only shift this ratio lower.
557 In determining the PCR miscall rate, we processed a clonal plasmid of pMT025-PDE3A. The
558 sequencing of this clonal plasmid DNA, which had undergone the same PCR process as the
559 saturation mutagenesis plasmid DNA library, showed error rate of 0.00037 miscalls/sequenced
560 nucleotide. Thus, when we factor in PCR miscalls, the read ratio of wild-type/variant becomes
561 5.9:
562
563 Start with miscall-free wild-type/variant read count ratio 9.2:1, the miscalls switch
564 (9.2*140*0.00037)=0.48 from wild-type group to variant group, producing a new wild-
565 type/variant read count ratio:
566 $$(9.2 - 0.48)/(1 + 0.48) = 5.9$$
567
568 By processing the same plasmid DNA library sequencing data set with various parameters, we
569 were able to identify the optimal parameters that produce a wild-type/variant ratio of 5.9, and

570 yield reasonable output of both wild-type and variants: min-q = 30, min-read-length = 40, min-
571 flank-length = 2.
572
573 **NGS Data processing and analysis:** The data from the NGS pipeline are in .fastq format. Each
574 NGS cluster produced 2 long reads (150 bases each), and 2 index reads (each 8 bases long). The
575 2 indexes identify a screen sample, so that a large number of samples can be pooled and
576 sequenced in bulk. Saturated ORF variant library screen data are typically 4-5TB in data size.
577 Generous computer storage is needed to process the data. The reads need to go through the
578 filtering process (**Supplementary Fig. 2**) prior to variant calling.

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

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

587 i) ORFCallv1.0 (an earlier version software; Giacomelli *et al*[15]): Variants are tallied
588 in the context of a single codon space. Without regard of the association of the
589 multiple changes physically in a single molecule, ORFCallv1.0 essentially scans from
590 N- to C-termini, one amino acid position at a time, and sums up the occurrence of
591 codons or amino acids at that position.

592

593 ii) ASMv1.0**:** This is a molecule-centric approach and an enhanced version over
594 ORFCallv1.0. The NGS dataset is analyzed in its entirety, with the following new
595 capabilities: (1) The trimmed reads are evaluated full-length; that is, variants are
596 called in the context of entire read (or read pair). The programmed variant is called
597 when 2 conditions are met: the detection of the programmed codon changes, and the
598 absence of any ***additional*** nucleotide variations throughout the entire read or read
599 pair. (2) Indel detection: This feature was lacking in ORFCallv1.0. This new feature
600 of ASMv1.0 allows us to process indel libraries.

601

602 In ASMv1.0, mutations in a molecule are depicted nucleotide-wise by position along
603 the PCR product, followed by the nature of the nucleotide change. For example, 'a
604 change from G to A at position 1140' is described as '**1140:G>A';** 'a deletion of an
605 A at position 1141' is '**1141:A>-**'; 'an insertion of a T after position 1142' is '**1143:-**
606 **>T**'.

607

608 Also with ASMv1.0, a host of annotations formatted as a datafile can be derived from
609 the variant descriptions. For example, a *TP53* variant with description '**1140:G>A,**
610 **1141:A>C, 1142:G>C**' is annotated as (a pair of brackets marks an annotation):
611 [E343T] [Intended] [one codon change] [343:GAG>ACC] [missense] [substitution].
612 Another *TP53* variant with description '**1140:G>A, 1141:A>C, 1142:G>C,**
613 **1145:G>T**' is: **[**E343T] [Unintended] [two codon changes] [343:GAG>ACC,
614 344:CTG>CTT] [substitution] [missense/nonsense] etc. The first variant is an

615 intended variant that matches a designed variant of the library. The second variant is
616 the same amino acid call as the first variant, but it is annotated, rightfully, as an
617 unintended variant carrying an error 4-nucleotide away from the intended codon
618 change - this error results in a silent codon change.

619 Parsing data into a form ready for hit calling**:** The output files from the above variant-calling
620 software are still large in size (over 100MB), large in numbers of files, and not readily useable.
621 These output files need to be parsed, aggregated and annotated. For this purpose, we built a
622 parser called ASMv1.0_Parser, which was written in R, and is available for download (see the
623 links listed in **Supplementary Table 2**). The Parser Pak consists of a 'readme' file detailing the
624 instructions, and R codes and training data sets. Through the parser, all data files will be
625 organized into a single .csv file, one column per screen sample. These .csv files include a short
626 version that is a subset of the planned variants only, and a complete version that includes all
627 called variants, including planned/unplanned, insertion/deletion/substitution, etc. In addition, we
628 also report the ORFCallv1.0 version of data.
629 Scaling: We use read counts from the NGS data as the quantitative measurements of variants
630 defined by 'changes' relative to the reference sequence. Reads that entirely match the reference
631 sequence are not informative, as they are, as a group, the aggregate of all variants whose changes
632 are located outside of the region covered by the read. At any amino acid position, the total
633 coverage, specifically the sum of counts of variants passing through the position and the sum of
634 counts of the wildtype reads, should be consistent throughout the ORF. But that is not actually
635 the case due to the transposon insertion bias that results in unevenness of total read coverage
636 (**Supplementary Fig. 8**). In cases that address the relative abundance among variants, as in
637 **Supplementary Figure 9**, the raw counts need to be scaled to the fraction of the total read

638 coverage across the codon position. However, in cases that compare the variant abundance
639 change between two samples, such as computing fold-changes in screening samples, scaling the
640 raw counts may not be necessary, as the positional raw count variations caused by transposon
641 insertion bias will cancel out between two samples when computing fold-changes
642 (**Supplementary Fig. 10**).
643

644 **Acknowledgements**

645 We thank our collaborators of saturation mutagenesis projects, Alice Berger, Jesse
646 Boehm, Ben Ebert, Laura Evans, Eejung Kim, Shengwu Liu, Julie-Aurore Losman, Huong Ly,
647 Peter Miller, Utthara Nayar, and Nikhil Wagle for demonstrating the diverse utilities of the
648 technology, Broad Institute Genomics Platform for providing NGS support, Nancy Tran and
649 Cong Zhu for the early work on saturation mutagenesis method, Stephanie Milczarek, Kirsten
650 Milenkowic, Michael Zagieboylo, and Jonah Schwartz for conducting lab work for some
651 projects, Leon Yang for critical reading of the manuscript. This work was supported by NCI U01
652 CA176058 (WCH), the Mark Foundation (WCH), American Cancer Society Professorship
653 (MM), American Cancer Society Mentored Research Scholar Grant MRSG-18-202-01 (ALH),
654 Department of Defense W81XWH-18-PRCRP-CDA (ALH), Team Lick Cancer (ALH), Unravel
655 Foundation (ALH), and Wong Family Award (ALH)
656

657 **Author Contributions**

658 X.Y., D.E.R., A.L.H., A.O.G., and T.S. wrote the manuscript. W.C.H., M.M., H.G.,
659 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,
660 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. A.O.G. is a share and option holder of 10X Genomics.

References

1. Lefèvre F, Rémy MH, Masson JM: **Alanine-stretch scanning mutagenesis: a simple and efficient method to probe protein structure and function.** *Nucleic acids research* 1997, **25:**447-448.
2. Cunningham BC, Wells JA: **High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis.** *Science (New York, NY)* 1989, **244:**1081-1085.
3. McCullum EO, Williams BAR, Zhang J, Chaput JC: **Random mutagenesis by error-prone PCR.** *Methods in molecular biology (Clifton, NJ)* 2010, **634:**103-109.
4. Muteeb G, Sen R: **Random mutagenesis using a mutator strain.** *Methods in molecular biology (Clifton, NJ)* 2010, **634:**411-419.
5. Kato S, Han S-Y, Liu W, Otsuka K, Shibata H, Kanamaru R, Ishioka C: **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* 2003, **100:**8424.

691 6. Chen L, Jenjaroenpun P, Pillai AMC, Ivshina AV, Ow GS, Efthimios M, Zhiqun T, Tan
692 TZ, Lee S-C, Rogers K, et al: **Transposon insertional mutagenesis in mice identifies**
693 **human breast cancer susceptibility genes and signatures for stratification.**
694 *Proceedings of the National Academy of Sciences* 2017, **114:**E2215.
695 7. Hughes RA, Ellington AD: **Synthetic DNA Synthesis and Assembly: Putting the**
696 **Synthetic in Synthetic Biology.** *Cold Spring Harbor perspectives in biology* 2017,
697 **9:**a023812.
698 8. Heather JM, Chain B: **The sequence of sequencers: The history of sequencing DNA.**
699 *Genomics* 2016, **107:**1-8.
700 9. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S:
701 **High-resolution mapping of protein sequence-function relationships.** *Nature Methods*
702 2010, **7:**741-746.
703 10. Fowler DM SJ, Fields S. : **Measuring the activity of protein variants on a large scale**
704 **using deep mutational scanning. Nature Protocols.** 2014.
705 11. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, Hopf TA,
706 Marks DS, Oren M, Segal E: **A Systematic p53 Mutation Library Links Differential**
707 **Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation.** *Mol*
708 *Cell* 2018, **71:**873.
709 12. Mighell TL, Evans-Dutson S, O'Roak BJ: **A Saturation Mutagenesis Approach to**
710 **Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype**
711 **Relationships.** *Am J Hum Genet* 2018, **102:**943-955.
712 13. Weile J, Kishore N, Sun S, Maaieh R, Verby M, Li RJ, Fotiadou I, Kitaygorodsky J, Wu
713 YZ, Holenstein A, et al: **Shifting landscapes of human MTHFR missense-variant**
714 **effects.** *American Journal of Human Genetics* 2021, **108:**1283-1300.
715 14. Awad MM, Liu S, Rybkin, II, Arbour KC, Dilly J, Zhu VW, Johnson ML, Heist RS, Patil
716 T, Riely GJ, et al: **Acquired Resistance to KRAS(G12C) Inhibition in Cancer.** *N Engl*
717 *J Med* 2021, **384:**2382-2393.
718 15. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, Howard TP,
719 Takeda DY, Ly SH, Kim E, et al: **Mutational processes shape the landscape of TP53**
720 **mutations in human cancer.** *Nature Genetics* 2018, **50:**1381-1387.
721 16. Garvie CW, Wu X, Papanastasiou M, Lee S, Fuller J, Schnitzler GR, Horner SW, Baker
722 A, Zhang T, Mullahoo JP, et al: **Structure of PDE3A-SLFN12 complex reveals**
723 **requirements for activation of SLFN12 RNase.** *Nat Commun* 2021, **12:**4375.
724 17. Kwon JJ, Hajian B, Bian Y, Young LC, Amor AJ, Fuller JR, Fraley CV, Sykes AM, So J,
725 Pan J, et al: **Structure-function analysis of the SHOC2-MRAS-PP1C**
726 **holophosphatase complex.** *Nature* 2022, **609:**408-415.
727 18. Miller PG, Sathappa M, Moroco JA, Jiang W, Qian Y, Iqbal S, Guo Q, Giacomelli AO,
728 Shaw S, Vernier C, et al: **Allosteric inhibition of PPM1D serine/threonine**
729 **phosphatase via an altered conformational state.** *Nat Commun* 2022, **13:**3778.
730 19. Persky NS, Hernandez D, Do Carmo M, Brenan L, Cohen O, Kitajima S, Nayar U,
731 Walker A, Pantel S, Lee Y, et al: **Defining the landscape of ATP-competitive inhibitor**
732 **resistance residues in protein kinases.** *Nat Struct Mol Biol* 2020, **27:**92-104.
733 20. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA,
734 Zhang X, Broekema MF, Patterson N, et al: **Prospective functional classification of all**
735 **possible missense variants in PPARG.** *Nat Genet* 2016, **48:**1570-1575.

736 21. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS: **Comprehensive mutational**
737 **scanning of a kinase in vivo reveals substrate-dependent fitness landscapes.** *Nucleic*
738 *Acids Res* 2014, **42:**e112.
739 22. Faure AJ SJ, Baeza-Centurion P, Lehner B.: **DiMSum: an error model and pipeline for**
740 **analyzing deep mutational scanning data and diagnosing common experimental**
741 **pathologies.** *Genome Biol* 2020, **21:**207.
742 23. Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N: **FLAM-seq: full-length mRNA**
743 **sequencing reveals principles of poly(A) tail length control.** *Nature Methods* 2019,
744 **16:**879-886.
745 24. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ,
746 Bowen JE, Tortorici MA, Walls AC, et al: **Deep Mutational Scanning of SARS-CoV-2**
747 **Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.** *Cell*
748 2020, **182:**1295-1310.e1220.
749 25. Marx V: **Long road to long-read assembly.** *Nat Methods* 2021, **18:**125-129.
750 26. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J,
751 Fungtammasan A, Kolesnikov A, Olson ND, et al: **Accurate circular consensus long-**
752 **read sequencing improves variant detection and assembly of a human genome.**
753 *Nature Biotechnology* 2019, **37:**1155-1162.
754 27. Frank F, Keen MM, Rao A, Bassit L, Liu X, Bowers HB, Patel AB, Cato ML, Sullivan
755 JA, Greenleaf M, et al: **Deep mutational scanning identifies SARS-CoV-2**
756 **Nucleocapsid escape mutations of currently available rapid antigen tests.** *Cell* 2022,
757 **185:**3603-3616 e3613.
758 28. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin
759 AF: **MaveDB: an open-source platform to distribute and interpret data from**
760 **multiplexed assays of variant effect.** *Genome biology* 2019, **20:**223-223.
761 29. Boettcher S, Miller PG, Sharma R, McConkey M, Leventhal M, Krivtsov AV,
762 Giacomelli AO, Wong W, Kim J, Chao S, et al: **A dominant-negative effect drives**
763 **selection of TP53 missense mutations in myeloid malignancies.** *Science*
764 2019, **365:**599.
765

766 **Figure legends**

767 **Figure 1. Pooled screens with ORF variant libraries versus shRNA/sgRNA libraries.** A
768 plasmid pool of thousands of genetic perturbation agents is first packaged into lentiviral
769 particles. Cells were then transduced with the pool of virus at an infection rate (the proportion of
770 infected cells over total cells) of 30-50% to ensure that most of the infected cells get a single
771 genetic perturbation. The perturbed cell population is then subjected to either selection pressure,
772 or treatment under a reference condition. Upon the completion of the screen, the remaining cell
773 population is harvested and processed for enrichment or depletion of each library perturbation by

774 comparing a sample that has undergone selection and a sample that has been treated under a
775 reference condition. **(a) A typical pooled screen using shRNA/sgRNA libraries.** For
776 shRNA/sgRNA, the collected cells are first processed for genomic DNA (gDNA), which serve as
777 the PCR template to amplify the perturbagen barcodes. These barcodes are in fact the functional
778 elements themselves, the hairpins of shRNA or guides of sgRNA. These barcodes are short
779 enough to be read straight out by a 50-base NGS read. **(b) ORF variant library screens.** In an
780 ORF variant library, however, the variant signatures are subtle as each member of the library
781 differs from the template by a single codon variation, and the variant signatures of the library are
782 spread out across the entire ORF. Consequently, the detection of variant requires (1) PCR-
783 amplification and then shotgun shearing of the full-length ORF, (2) next-generation sequencing
784 of the resulting fragments, (3) robust variant detecting tools to align the reads to the template,
785 then identify, evaluate, annotate and tally the variant-defining reads. The NGS reads consist of
786 variant signature-carrying reads (shown as red arrows) and wild-type reads (black arrows). The
787 variant-calling software processes the NGS reads to detect variant signature-carrying reads,
788 which is a minority of the total reads, and records the counts of each variant. These counts are
789 the basis for downstream analyses that map the variants to the phenotype strength in the form of
790 log2-fold change (LFC).
791
792 **Figure 2. Reanalysis of *TP53* saturation mutagenesis screen using ASMv1.0.** Screens most
793 often involve the comparison of a sample subjected to selective pressure with a sample processed
794 under a reference condition. This Giacomelli *et al*[15] $p53^{NULL}$ A549 cell/nutlin-3 screen
795 enriches *TP53* variants that exhibit loss-of-function (LOF), and depletes wild-type-like variants.
796 (**a**) Heatmaps: each variant is placed in a 2-dimensional grid, where codon position is presented

797 along the horizontal dimension and the amino acid mutation along the vertical dimension. We
798 assigned the letter 'X' to represent nonsense mutations, and 'B' to represent silent codon
799 changes. The value of the log fold change (LFC) between treatment and reference samples is the
800 heat index. We processed the same original NGS data through the two versions of software and
801 presented the resulting log fold change (LFC). The pair of heatmaps are nearly identical to the
802 naked eye. To help inspect all variants one by one, the third heatmap used the
803 delta(LFC)=(LFC^{ASM} – $LFC^{ORFCall}$) as the heat index. This heatmap revealed variants that were
804 scored differently by the two versions of software. (**b**) Scatter plot comparing all LFC values
805 produced by the two versions of software, each dot representing a planned variant. The colors are
806 scaled by variant purity. (**c**) Demonstrating the impact of variant purity on the improved
807 performance of the new ASMv1.0 software. The new software's massive output allows us to
808 compute 'variant purity' defined by the fraction of pure planned variant over all variants that
809 harbor the planned variant, with or without unintended changes in the same sequencing reads.
810 We binned the variants by variant purity and in each purity bin we assessed the Pearson
811 correlation coefficient between two sets of LFC values produced by the two versions of software.
812 It is clear that high variant purity results in similar LFC values, whereas as the purity decreases,
813 the two sets of fold-changes diverge. The new software's ability to detect and tally variants that
814 carry additional unintended changes is called for when the variant purity is low. Similarly, we
815 analyzed two other Giacomelli *et al*[15] screens of the *TP53* saturated ORF variant library and
816 the results are presented in **Supplementary Figure 3** and **4**. (**d,e**) Validation of ASMv1.0 using
817 the published data by Kotler *et al.*[11] The Kotler dataset includes a screen of $p53^{NULL}$H1299
818 cells with a library covering *TP53* DBD. We first took a subset of Giacomelli library to cover the
819 same *TP53* region as Kotler screens, resulting in a total of 2964 variants (**d**, **left**). We then

820 identified 76 variants that recorded most different LFC scores by our two versions of
821 software **(d, right**). We then compared the Kotler scores against our LFC scores called by
822 ORFCallv1.0 (**e, left**) and ASMv1.0 (**e, middle**). We observed that between two versions of
823 software, ASMv1.0 calls are clearly more in agreement with Kotler calls (**e, right**).
824
825 **Figure 3. The errors in our saturated ORF variant libraries are low in number and local**
826 **relative to the location of the planned codon change.** In screens, the mutational spaces of
827 libraries depend on how many planned variants can be detected in the final sequencing data. A
828 library with a high impurity rate will require a larger experimental effort. (**a**) **Breakdown of the**
829 **planned variants and those with errors.** Going through all called variants that contain a
830 planned codon change, we identified the number of unintended nucleotide changes. This bar-
831 chart shows that in the pMT025-TP53 library, 84% molecules are planned variants without
832 additional mutations. Of 16% imperfect molecules, the majority are those with 1-2 extra
833 nucleotide changes. (**b**) **Major errors are in the vicinity of a planned change.** We first pinned
834 down the planned change in a read and then travelled in both directions and tallied the errors as a
835 function of the distance to the pinned-down codon change. In this figure, we compared two
836 libraries made with different technologies, *TP53* library by MITE and *PDE3A* by Twist
837 Bioscience. The error distribution profile along the distance to a planned variant is strikingly
838 explainable by the oligo synthesis scheme involved in these technologies. MITE technology uses
839 synthesized oligos of 150 bases long, of which 90 bases (called TILE) go into the final products,
840 whereas Twist Bioscience introduced errors through oligo primers with the intended codon
841 changes placed in the middle and flanked with 15-25 bases on either side. The major errors are
842 concentrated within ~100 bases (MITE method) or ~30 bases (Twist Bioscience method) from a

843 planned codon change. This is important because the close proximity of the errors to the
844 intended codon change renders the planned change and errors all detectable in a single read (e.g.,
845 150 bases) or read pair (e.g., 2x150=300 bases).
846
847 **Figure 4. The effect of PCR and sequencing errors on variant quantification. (a)** A clonal
848 *TP53* reference plasmid (i.e. all sequences are wild-type) was processed as if it were a screening
849 sample. The base call at each position is predominantly, but not absolutely, the reference
850 nucleotide. The non-reference calls represent artifactual variants that are not actually in the
851 library. The heatmap on the left shows tallied counts of called nucleotides at each nucleotide
852 position; a blank spot represents zero counts. The heatmap on the right depicts the rate of
853 miscalls. Blank spots in this heatmap are either the wild-type nucleotide or nucleotides of zero
854 counts. (**b**) The miscalls resulting from PCR errors inflate the abundance measurement of
855 variants whose variant-defining codon is 1-nt away from the reference codon. In the *TP53* library
856 screen data, predominately affected are variants of 1-nt (not 2- or 3-nt) difference from the
857 reference codons. As a result, the variants of 1-nt delta codons as a group appear more abundant
858 than groups of 2- or 3-nt delta codons (**left**). The artificial 'variant' calls resulting from PCR
859 errors can be corrected with abundance adjustment by subtracting the miscall rate as determined
860 by the clonal sample experiment (**right**). The variant abundance measurements of the 2-nt and 3-
861 nt delta codons are not affected by the miscalls. (**c**) **Suppression of fold-changes in 1-nt delta**
862 **codon variants.** PCR errors in the workflow inflate the counts of variants that differ from the
863 template codon by one nucleotide. Using the pMT025-TP53 library as an example, 1-nt delta
864 variants showed much narrower fold-change values than 2-nt and 3-nt delta variants. Thus, in the
865 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.

Figure 1

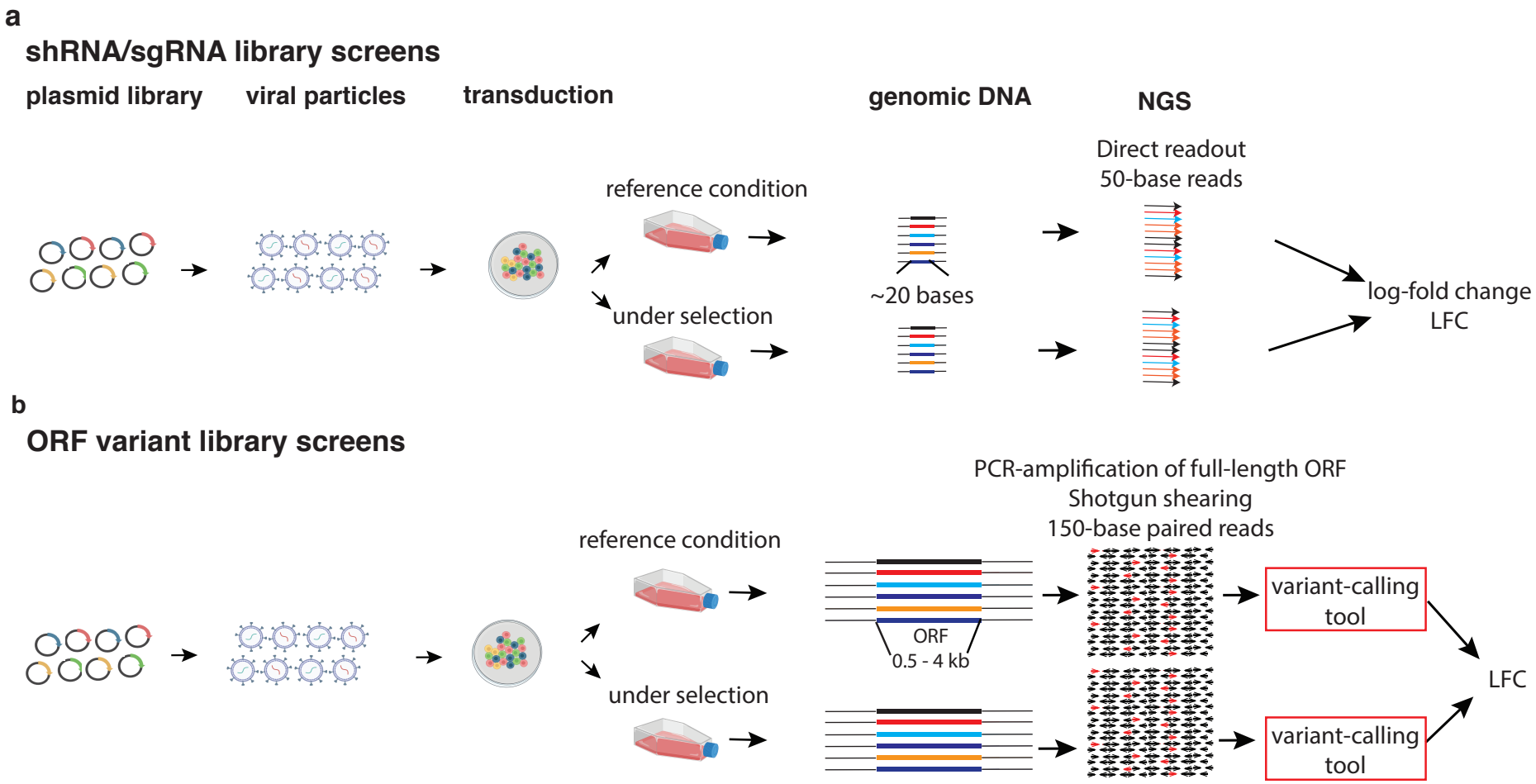

a
shRNA/sgRNA library screens
plasmid library
viral particles
transduction
genomic DNA
NGS
reference condition
under selection
~20 bases
Direct readout
50-base reads
log-fold change
LFC
b
ORF variant library screens
reference condition
under selection
PCR-amplification of full-length ORF
Shotgun shearing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

Figure 2

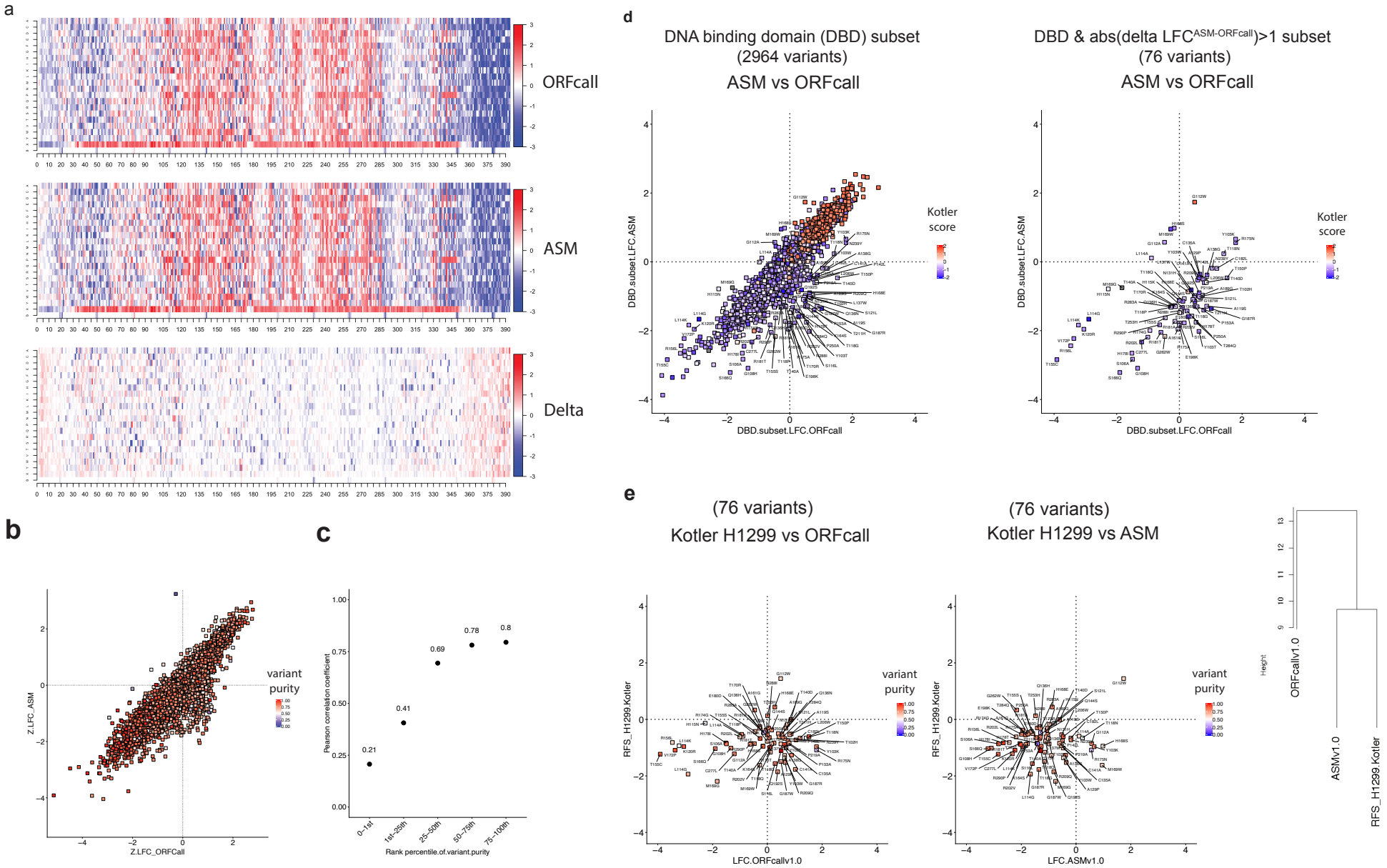

a
ORFcall
ASM
Delta
b
Z.LFC_ASM
Z.LFC_ORFcall
variant purity
c
Pearson correlation coefficient
Rank percentile.of.variant.purity
0.21
0.41
0.69
0.78
0.8
0–1st
1st–25th
25–50th
50–75th
75–100th
d
DNA binding domain (DBD) subset
(2964 variants)
ASM vs ORFcall
DBD & abs(delta LFC^ASM-ORFcall)>1 subset
(76 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 3

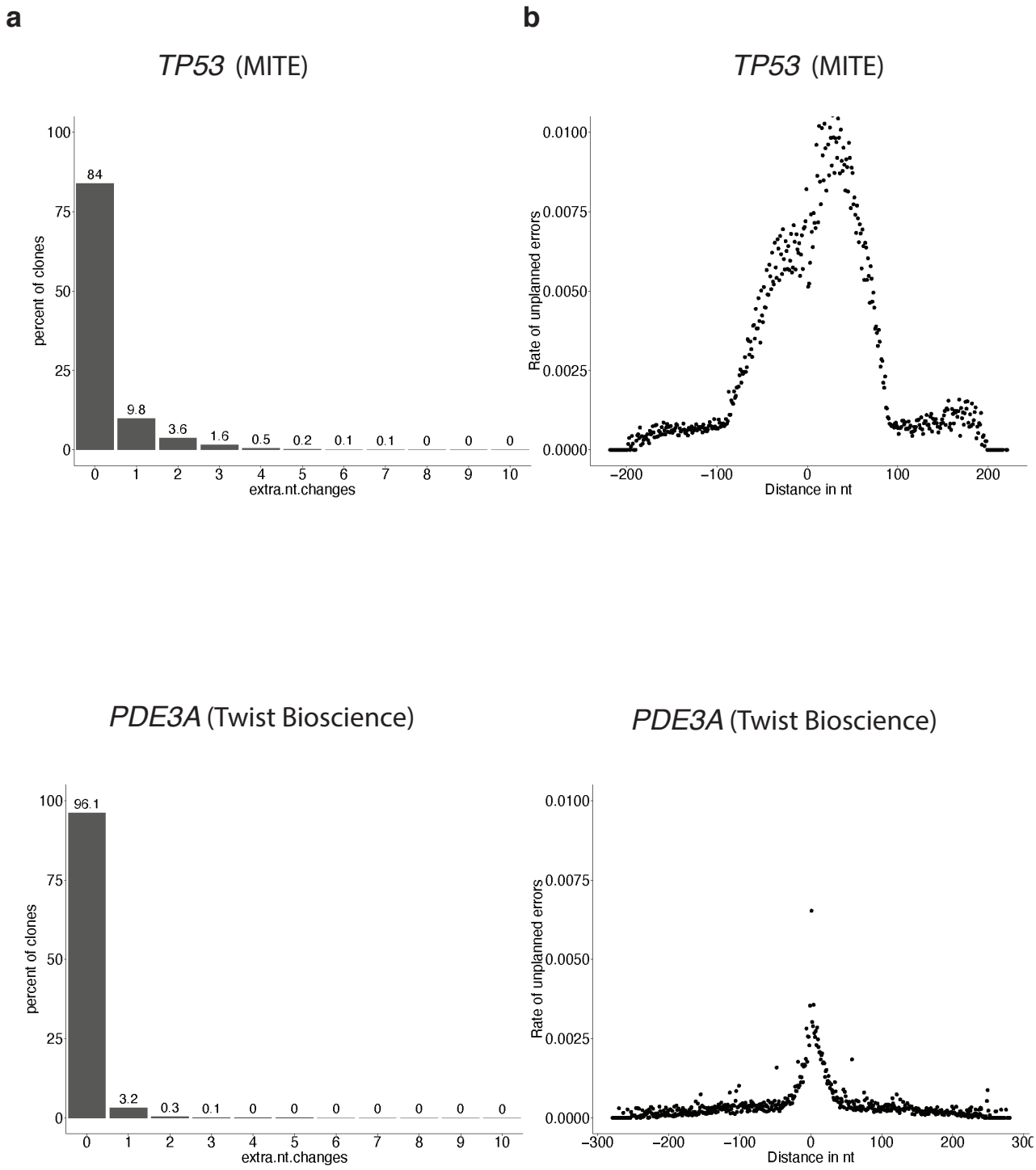
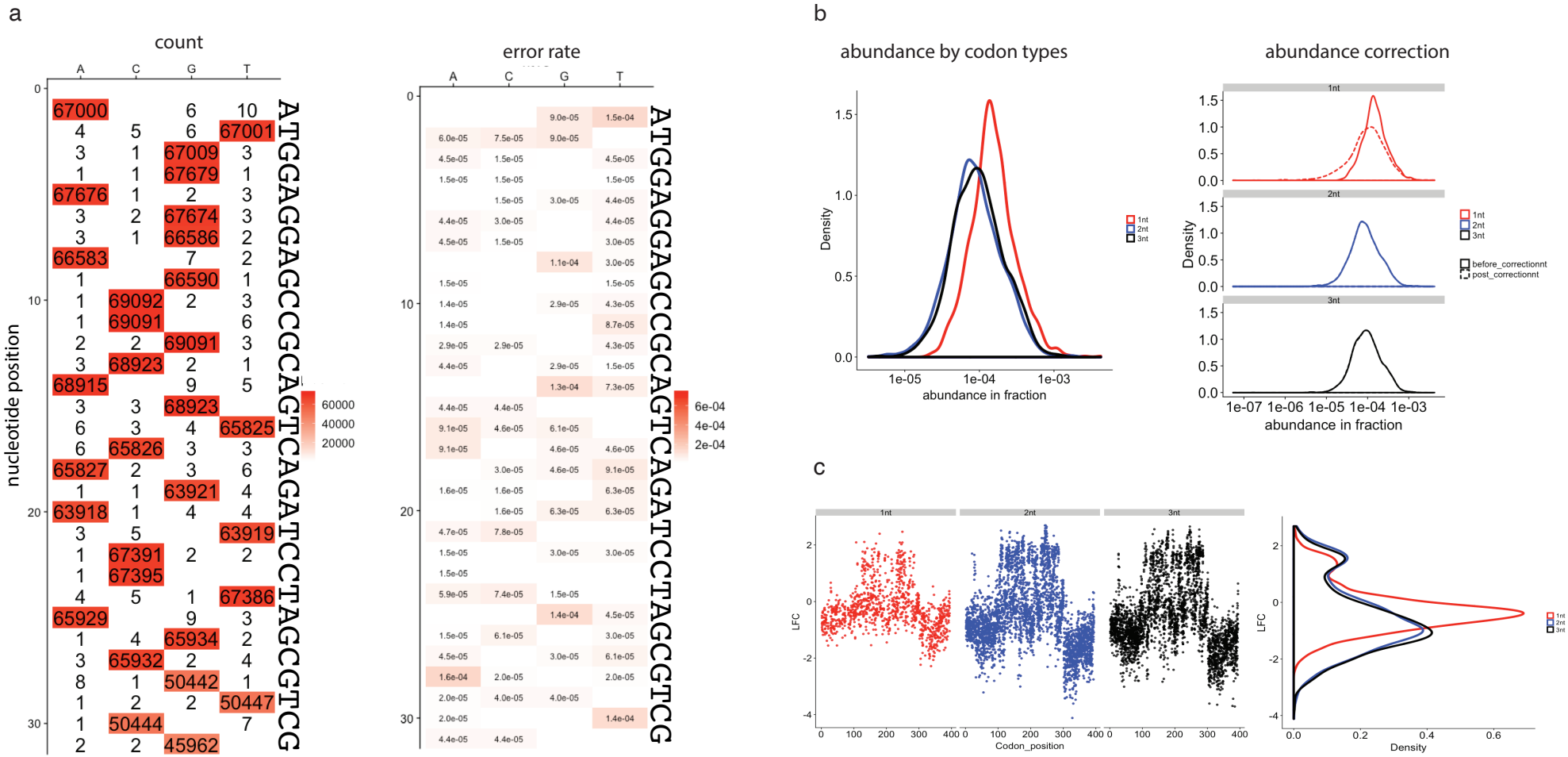

a
b
TP53 (MITE)
percent of clones
extra.nt.changes
84
9.8
3.6
1.6
0.5
0.2
0.1
0.1
0
0
0
TP53 (MITE)
Rate of unplanned errors
Distance in nt
PDE3A (Twist Bioscience)
percent of clones
extra.nt.changes
96.1
3.2
0.3
0.1
PDE3A (Twist Bioscience)
Rate of unplanned errors
Distance in nt

Figure 4


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