Sensitive and specific post-call filtering of genetic variants in xenograft and primary tumors

Brian K Mannakee[1,2], Uthra Balaji[3], Agnieszka K Witkiewicz[2,4,5], Ryan N Gutenkunst[6], and Erik S Knudsen[2,4]

[1]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, Arizona, USA
[2]University of Arizona Cancer Center, University of Arizona, Tucson, Arizona, USA
[3]McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[4]Department of Medicine, University of Arizona, Tucson, Arizona, USA
[5]Department of Pathology, University of Arizona, Tucson, Arizona, USA
[6]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, USA

September 11, 2017

Abstract

Motivation: Tumor genome sequencing offers great promise for guiding research and therapy, but spurious variant calls can arise from multiple sources. Mouse contamination can generate many spurious calls when sequencing patient-derived xenografts (PDXs). Paralogous genome sequences can also generate spurious calls when sequencing any tumor. We developed a BLAST-based algorithm, MAPEX, to identify and filter out spurious calls from both these sources.

Results: When calling variants from xenografts, MAPEX has similar sensitivity and specificity to more complex algorithms. When applied to any tumor, MAPEX also automatically flags calls that potentially arise from paralogous sequences. Our implementation, `mapexr`, runs quickly and easily on a desktop computer. MAPEX is thus a useful addition to almost any pipeline for calling genetic variants in tumors.

Availability: The `mapexr` package for R is available at `https://bitbucket.org/bmannakee/mapexr` under the MIT license.

Contact: mannakee@email.arizona.edu, rgutenk@email.arizona.edu, eknudsen@email.arizona.edu

1 Introduction

Molecular characterization of tumors is an important tool in cancer research, and the large-scale sequencing of cancer genomes has led to a deeper understanding of many aspects of the biology of cancer [Stratton MR, 2011]. It is now common to sequence tumors from large cohorts of patients, as well as patient-derived xenograft (PDX) models from individual patients. Such sequencing enables identification of mutational signatures [Alexandrov *et al.*, 2013], functionally important variants [Ding *et al.*, 2012] and evolutionary history of the tumor [Carter *et al.*, 2012, Nik-Zainal *et al.*, 2012]. These genetic features are relevant in evaluating etiological mechanisms [Yachida *et al.*, 2010], prognostic subtypes [Park *et al.*, 2010, Shah *et al.*, 2009], and acquired therapeutic resistance [Witkiewicz *et al.*, 2015]. All these applications of tumor sequencing depend on sensitive and specific characterization of low-frequency mutations, and as a result may be biased by spurious variant calls. Here we focus on two specific sources of spurious calls, mouse cell contamination in PDX tumors and mis-alignment of paralogous sequences.

PDX models serve as avatars for individual patient tumors when studying intra-tumor heterogeneity and metastasis and when screening anti-cancer compounds [Allaway *et al.*, 2016, Bruna *et al.*, 2016,

Dawson *et al.*, 2012, Day *et al.*, 2015, Knudsen *et al.*, 2017]. The primary difficulty in sequencing these models is that mouse stroma is present in all PDX tumors. The high genetic similarity between mouse and human then causes bias when variants are called using bioinformatic pipelines originally developed for primary tumors [Rossello *et al.*, 2013, Tso *et al.*, 2014]. Several methods have been developed to facilitate the accurate calling of variants in PDX models. Experimentally, human-specific fluorescence tags can be used to label and isolate human cells prior to DNA extraction [Schneeberger *et al.*, 2016]. Bioinformatically, sequence reads can be aligned to both human and mouse reference genomes, either separately [Conway *et al.*, 2012, Khandelwal *et al.*, 2017] or simultaneously [Bruna *et al.*, 2016], to filter out mouse reads prior to variant calling. Although these approaches greatly improve the reliability of variant calls from PDX models, they entail substantial experimental or bioinformatic burdens. Here we describe a lightweight filtering algorithm that achieves equivalent reliability and can be easily added to standard bioinformatic pipelines.

Many human genes have highly similar paralogous sequences in the genome. Spurious variant calls arising from such paralogs have been recognized as an important source of false positives in the study of rare disease-associated germline variants [Jia *et al.*, 2012, Mandelker *et al.*, 2016, Ng *et al.*, 2010, Zhou *et al.*, 2015]. Similarly, paralogs have led to false positives in the study of cancer, including TUBB in non-small cell lung cancer [Kelley *et al.*, 2001], PIK3CA in hepatocellular carcinoma [Müller *et al.*, 2007, Tanaka *et al.*, 2006], and MLL3 in myelodysplastic syndrome [Bowler *et al.*, 2014]. To address the paralog problem, some variant callers, such as MuTect2 (currently in beta but included in the Genome Analysis Toolkit (GATK; McKenna *et al.* [2010])), filter clustered variants, which often result from misalignment of paralogous sequences. Many labs also keep lists of suspect genes that tend to suffer from paralog problems and simply ignore any variants called in these genes. These approaches introduce their own biases. Our approach automatically identifies potential spurious calls from paralogs and enables flexible evidence-based filtering.

Here we fully describe and characterize MAPEX (the Mouse And Paralog EXterminator), a BLASTN-based algorithm for filtering variants that was previously introduced by Knudsen *et al.* [2017]. We also present `mapexr`, a fast and lightweight implementation in R. We show that, when applied to PDX samples, MAPEX generates calls that are highly similar to other methods, but with less bioinformatic and computational overhead. We also show that, when applied to primary samples, MAPEX effectively filters paralogs while avoiding biases of existing heuristics. MAPEX is thus a useful addition to different tumor variant calling pipelines.

2 Approach

2.1 Workflow

The MAPEX algorithm is a post-variant-calling filter designed to fit into a standard tumor variant calling pipeline and flag variants which may arise from mis-alignment of mouse reads or from paralogous sequences. The input for MAPEX is a BAM file containing tumor reads aligned to the human reference genome and a variant callset generated from that alignment. MAPEX scores variants by the fraction of variant-supporting reads that align best to the site of the variant when BLASTed against a combined human/mouse reference genome (Figure 1).

2.2 Algorithm

Each read supporting a variant is BLASTed against the appropriate reference genome for the application. For PDX applications, this is the combined human/mouse reference, and for primary tumor applications, this is just the human reference. The best hit for each read is determined by bit score. Reads for which the best hit overlaps the called variant location are classified as "on target" and assigned a score of 1. Reads for which the best hit is a different region of the human genome or a region of the mouse genome are classified as "off target" or "mouse", respectively, and assigned a score of 0. Reads from genes with close paralogs in the human genome may generate multiple best hits (ties). In this case, the read score is averaged over all best hits, and the read is classified based on the most common result from the best hits. Each variant is then assigned a score that is the average score of all reads supporting that variant and is classified based on the most common classification of the supporting reads.

2.3 Implementation

We have implemented the MAPEX algorithm as an R package (mapexr). The package leverages the Bioconductor packages `Rsamtools`, `GenomicAlignments`, and `GenomicRanges` for fast and memory-efficient BAM file handling and read sequence extraction [Lawrence *et al.*, 2013, Morgan *et al.*, 2017]. The package requires a local BLASTN

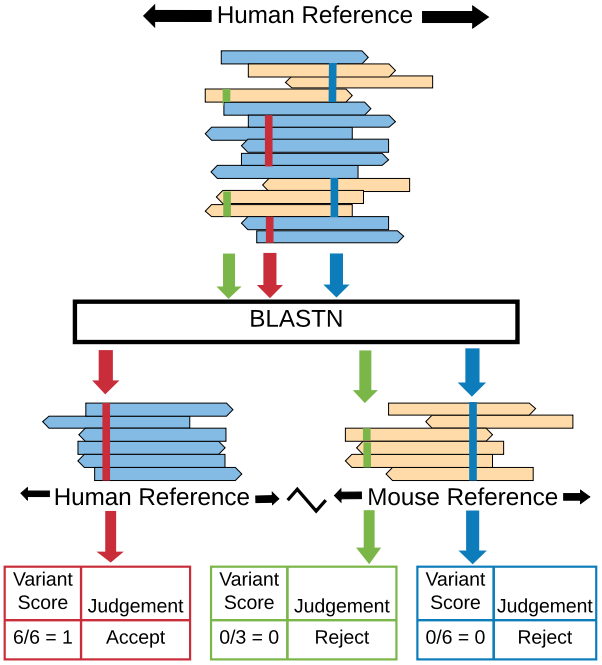

Figure 1: **Illustration of MAPEX applied to a PDX sample.** MAPEX begins with variants called from tumor reads aligned to the human genome. For each variant (red, blue, and green lines), the supporting reads are BLASTed against the combined human and mouse reference genomes. Variants are then scored by the fraction of supporting reads that align to the called site of the variant in the human genome.

installation and a BLAST database constructed from either a combined human/mouse reference genome or a human reference genome, depending on the application.

3 Methods

3.1 Samples

To characterize the performance of MAPEX, we used Whole Exome Sequence trimmed fastq reads obtained from pancreatic ductal adenocarcinoma (PDAC) samples described previously by Knudsen *et al.* [2017] (PDX) and Witkiewicz *et al.* [2015] (primary). For the PDX analysis, we analyzed a total of 34 PDXs derived from 9 primary tumors, sequenced to mean coverage depth of 124x. For the paralog analysis, we analyzed 93 primary tumors sequenced to a mean coverage depth of 40x.

3.2 Alignments and variant callers

All alignments were done using `bwa-mem` with default parameter settings [Li and Durbin, 2009]. For initial variant calling, we aligned all reads in the samples to the human reference genome `GRCh37`. We then called variants using MuTect version 1.1.1 [Cibulskis *et al.*, 2013], MuTect2 (as part of the GATK version 3.6, McKenna *et al.* [2010]), and Varscan 2 [Koboldt *et al.*, 2012], all with default parameter settings. Variants were annotated with Oncotator [Ramos *et al.*, 2015] and the annotation database `oncotator_v1_ds_April052016`. For Varscan 2, two PDX samples that yielded millions of variant calls were not processed by MAPEX, to conserve computational time. We considered only nonsynonymous single nucleotide variants when comparing between methods. For paralog filtering, we used a conservative variant score cutoff of 0.8.

For comparison with Bruna *et al.* [2016], we aligned reads to a combined human/mouse reference genome `GRCh37/mm9` and called variants using MuTect 1.1.1. We calculated the fraction of mouse contamination using the method described in Bruna *et al.* [2016]. Briefly, they generated data comparing the fraction of mouse cells in a sample with the fraction of total reads aligned to the mouse portion of a combined reference genome. We used this data to fit a LOESS regression model for contamination fraction vs fraction aligned, and used this to predict mouse contamination based on the fraction of reads aligned to the mouse genome in our samples.

For comparison with `bamcmp` [Khandelwal *et al.*, 2017], we aligned reads separately to the human and

mouse reference genomes and ran `bamcmp` with default parameters. The output of `bamcmp` includes alignment files for reads that aligned to only the human reference and that aligned to both references but with a higher human alignment score. We merged these two alignments, performed indel realignment and base score recalibration using the GATK, and used the merged alignment to call variants with Mutect version 1.1.1.

4 Results & Discussion

4.1 Methodological

MAPEX is a lightweight filtering algorithm that adds little overhead or complexity to existing variant-calling pipelines. The runtime for MAPEX is linear in the number of variants to be filtered. On a 4-core machine, our implementation `mapexr`, processes roughly 250 variants per minute (Figure S1).

MAPEX has only one tunable parameter, the minimum mapping quality score required for a variant read. The default minimum score is 1, which includes all reads with an unambiguous best mapping. In pipelines in which a minimum mapping quality score is used for variant calling, that score should also be supplied to `mapexr`, to prevent evaluating reads that were not used by the variant caller. The output from `mapexr` is an R data frame with four columns – chromosome, start location, variant score, and variant classification – and one row for each variant evaluated. Users may also optionally provide a file path to `mapexr` which will generate a tab-delimited file with blast results and scores at the read level. The user can choose the variant score threshold used to classify variants as human- or mouse-derived. Here we use a threshold of 0.5, so that a variant is flagged as spurious if less than half of the supporting reads BLAST as "on target". In practice, the distribution of variant scores is bimodal and highly concentrated at 0 and 1, so results are insensitive to the exact threshold (Figure S2).

4.2 Filtering mouse calls from PDX samples

One important use case for MAPEX is as a post-variant-calling filter for PDX samples that have been aligned to a human reference genome. To test the precision of MAPEX, we compared variant calls from aligning reads to the human reference and filtering with MAPEX to calls from two other methods. The first alternate method is to align reads to a combined human and mouse reference and then call variants [Bruna *et al.*, 2016], which we refer to as the "combined reference" method. The second method is to align reads separately to human and mouse references and call variants using only those reads that align better to the human reference, which is the method implemented in `bamcmp` [Khandelwal *et al.*, 2017]. For three representative PDX tumors, all three methods yield similar callsets (Figure 2A). The differences are primarily confined to low-frequency variants, and almost all high-frequency variants are called by all three methods (Figure 2B). Across 34 PDX tumors, all three methods yield a similar dramatic reduction in called variants (Figure 2 C).

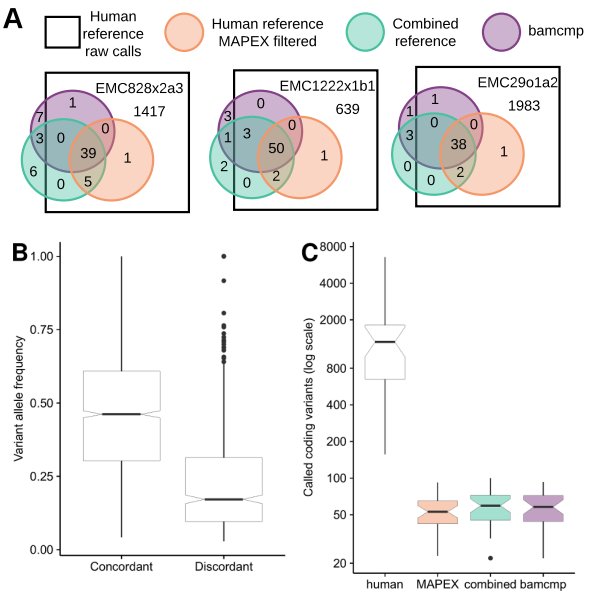

Figure 2: Comparison of MuTect 1.1.1 variants calls between MAPEX, combined reference, and `bamcmp` methods. **A:** Detailed breakdown of variant call overlap between the unfiltered human alignment (white square), MAPEX filtered human alignment (red circle), bamcmp filtered human alignment (purple circle) and unfiltered combined alignment (green circle) for representative PDXs created from three different primary tumors. **B:** Variant allele frequencies for calls that are concordant (n=1663 variants) and discordant (n=552 variants) between the methods. **C:** Comparison of total calls between the methods, n=34 PDX samples. Boxplots depict 25th and 75th percentile with 1.5×IQR whiskers. Notches are Median ± 1.58×IQR/sqrt(n), and represent rough estimates of 95% confidence interval around the median.

To further validate MAPEX, we compared PDX variant calls before and after filtering to the primary tumor from which the PDX was derived, where mouse

contamination is not an issue. Across 34 PDX tumors derived from 9 primaries, MAPEX dramatically enriches PDX calls for variants that were also found in the primary tumor and removes few PDX calls that were found in the primary tumor. Among variants in the PDXs, only 0.3% to 10% called before MAPEX filtering were also found in the primary tumor, but 23% to 90% of variants called after MAPEX filtering were found in the primary tumor (Table S1). This suggests that MAPEX enriches strongly for true variants. Among variants found both in the primary and the PDX before MAPEX filtering, 92% to 100% were retained after filtering (Table S1). This suggests that MAPEX removes few true variants.

To validate the usefulness of MAPEX in practice, we focused on calls within known cancer-associated genes, using the COSMIC database. Among the pancreatic ductal adenocarcinoma (PDAC) samples in COSMIC, 34 genes are mutated in more than 3% of samples. Before filtering with MAPEX, 910 variants were found in these genes among the 34 PDXs we studied. After filtering with MAPEX, only 70 variants were retained. Together, these results suggest that MAPEX removes many false positives, dramatically simplifying variant interpretation. Of particular interest are KRAS, TP53, and SMAD4, which are the most commonly mutated genes in PDAC (Table 1). All of the KRAS mutations filtered by MAPEX are I187V mutants, which result from aligning wild-type mouse KRAS reads to human KRAS, and all 34 PDXs retained the KRAS mutation found in their primary tumor. All of the SMAD4 mutations that were retained by MAPEX in the PDXs also appeared in the corresponding primary tumors. Also of interest is ARID1A, for which the single variant retained by MAPEX was confirmed to appear in the corresponding primary tumor, and none of the filtered variants were present in a corresponding primary tumor.

4.3 Effects of variant call filters on PDXs

We carried out our primary analyses with the variant caller MuTect 1.1.1, but to test the performance of MAPEX with other variants callers, we also considered MuTect2 and Varscan 2.

If mouse contamination were perfectly filtered, the number of called variants should not depend on the level of mouse contamination. For all three variant callers the number of raw calls was strongly correlated with estimated mouse contamination (Fig. 3A,B,C), although MuTect2 did produce substantially fewer calls overall. After filtering with MAPEX, the numbers of variants called with MuTect 1.1.1 and MuTect2 were not significantly correlated with the level of mouse contamination (Fig.. 3D&E). On the other hand, the number of variants called with Varscan 2 was correlated with mouse contamination (Fig. 3F), suggesting that MAPEX is not eliminating all spurious calls.

Table 1: Variants detected in PDX samples for important PDAC genes.

	before MAPEX		after MAPEX	
Gene	Total variants	Samples with a variant	Total variants	COSMIC prevalence
KRAS	56	34	34	0.64
TP53	9	9	7	0.39
SMAD4	5	5	5	0.14
SYNE1	3	3	0	0.05
CSMD3	96	25	0	0.05
GNAS	6	6	6	0.05
HMCN1	10	5	0	0.04
APC	12	11	0	0.04
NEB	31	17	0	0.04
WDFY4	6	4	1	0.04
LRP1B	32	18	1	0.04
ARID1A	131	33	1	0.04

Importantly, as a post-variant-calling filter, MAPEX can not evaluate variants that were not initially called. Filters implemented with a variant caller, generally designed to improve results from primary tumors, can cause problems when using MAPEX. For example, MuTect2 applies a clustered event filter designed to reduce the number of false-positive variant calls due to mis-alignment of highly paralogous sequences. In regions of high similarity between mouse and human, this filter can remove true variants. For instance, Figure 4 shows the result of

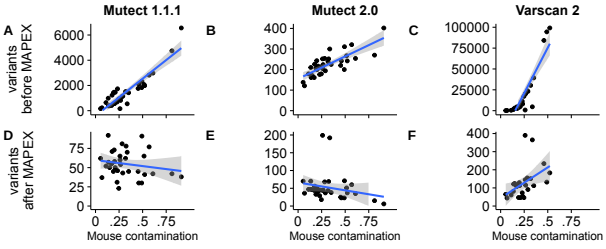

Figure 3: Effects of variant caller on analyzing xenograft samples with MAPEX. A,B,C: For all three calling algorithms and 34 xenograft samples (black dots) the number of raw variants called was strongly dependent on estimated mouse contamination. D,E,F: After filtering with MAPEX, the number of calls was independent of mouse contamination for MuTect 1.1.1 and MuTect2, but not for Varscan 2. Blue lines show linear regressions and shading denotes 95% confidence intervals.

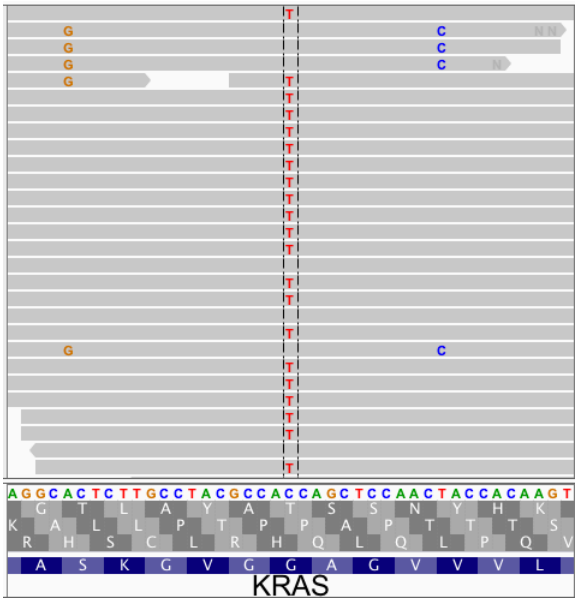

Figure 4: This Integrative Genomics Viewer [Thorvaldsdottir *et al.*, 2013] window covers a portion of the human KRAS gene. The C>T variant is the classic KRAS G12D mutation that appears in many PDAC tumors. The A>G and T>C variants both result from aligning wild-type mouse reads to the human sequence. When used with MuTect 1.1.1 or Varscan 2, MAPEX correctly retains only the G12D variant. MuTect2, however, filters all three variants, so the G12D variant cannot be retained.

aligning a PDX with modest mouse contamination to the human reference for a small portion of the KRAS oncogene. MuTect 1.1.1 and Varscan 2 both called three variants at this locus, and MAPEX correctly rejected the two spurious variants arising from mouse contamination and retained the true G12D variant. MuTect2 fails to call any of these variants, because they are filtered as likely homologous mapping events, so MAPEX does not see and cannot retain the true G12D variant. In our PDX samples, we found instances of the clustered event filter removing true variants from other PDAC oncogenes, including SMAD4 and TP53.

Overall, the performance of MAPEX does not depend sensitively on the variant caller used, but callers can introduce specific biases. In particular, the default parameters for Varscan 2 yield high sensitivity but low specificity. When Varscan 2 is applied to PDX samples with mouse contamination, MAPEX thus does not filter out all spurious calls. As such, we recommend that users of Varscan 2 be cautious when calling PDX samples and perhaps apply additional post-calling filters. By contrast, the default parameters for MuTect2 yield much higher specificity, but at the cost of sensitivity in the PDX context. Currently, the clustered event filter cannot be disabled in MuTect2. We thus advise that users pairing MAPEX with MuTect2 be cautious when interpreting callsets from PDX samples in genes with high similarity between human and mouse.

4.4 Flagging potential false positives resulting from paralogous sequences

In addition to removing mouse contamination from PDX samples, MAPEX can also filter potential paralogs in primary samples. Across 93 PDAC primary tumors, a mean of 11% of total variant calls were flagged by MAPEX as potential paralogs, with a range of 2-33%. The genes in which variants were most frequently flagged as potentially arising from paralogous sequences include members of large gene families, such as mucins, zinc-finger nucleases, and the PRAME family (Table 2). Variants in citrate synthase (CS) were also frequently flagged (Table 2). Citrate synthase has a known pseudogene `NCBI: LOC440514`, which was responsible for all of the spurious calls. We called variants with MuTect 1.1.1 and filtered with MAPEX, but MuTect2 includes new clustered event and read-mapping quality filters to prevent calling variants caused by paralogs. Using MAPEX yielded call sets that were identical with MuTect2 for all the genes in Table 2, with the exception of MUC12 and MUC5B, which differed by 3 variants. MAPEX can thus be efficiently and confidently used to remove variants that likely arise from paralogous sequences.

5 Conclusion

Genome sequencing is an increasingly important tool in cancer research, but spurious variant calls remain a challenge. MAPEX is an algorithm designed to filter spurious variants caused by mouse reads in patient-derived xenografts (PDXs) and caused by paralogous sequences in primary tumors. We showed that MAPEX is as sensitive and specific as more computationally intensive methods for calling variants from PDX tumors. We also showed that MAPEX successfully flags variant calls in potentially problematic gene families in primary tumors. Our implementation, `mapexr`, fits cleanly into standard tumor variant-calling pipelines and runs quickly on modern desktop computers. MAPEX is thus a potentially

Table 2: Top genes for which MAPEX flagged variants as potentially arising from paralogs.

Gene	Variants flagged	Samples with a flagged variant
ZNF814	15	15
CS	12	7
IGFN1	8	6
KMT2C	7	7
FRG1	6	6
LILRB3	6	6
MUC12	6	6
RGPD3	6	6
USP6	6	3
FCGBP	5	4
MUC5B	5	5
NBPF1	5	3
PRAMEF11	5	4
PRB4	5	3
RGPD8	5	4

useful new component for many tumor variant-calling pipelines.

Funding

This work wa supported by the National Science Foundation via Graduate Research Fellowship DGE-1143953 to BKM and by the National Institutes of Health via grants R01CA211878-01 and P30CA023074-36S2 to AKW and ESK.

References

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, **500**(7463), 415–421.

Allaway, R. J., Fischer, D. A., de Abreu, F. B., Gardner, T. B., Gordon, S. R., Barth, R. J., Colacchio, T. A., Wood, M., Kacsoh, B. Z., Bouley, S. J., Cui, J., Hamilton, J., Choi, J. A., Lange, J. T., Peterson, J. D., Padmanabhan, V., Tomlinson, C. R., Tsongalis, G. J., Suriawinata, A. A., Greene, C. S., Sanchez, Y., and Smith, K. D. (2016). Genomic characterization of patient-derived xenograft models established from fine needle aspirate biopsies of a primary pancreatic ductal adenocarcinoma and from patient-matched metastatic sites. *Oncotarget*, **7**(13), 17087–102.

Bowler, T. G., Bartenstein, M., Morrone, K. A., Rohanizadegan, M., Kessel, R. M., Hooda, L., Datt, I., Giricz, O., Bhagat, T. D., Przychodzen, B. P., Parmar, S., Gill, J. B., Yu, Y., Maciejewski, J. P., Steidl, U., and Verma, A. (2014). Exome Sequencing of Familial MDS Reveals Novel Mutations and High Rates of False Positive Mutations in MLL3 Due to Pseudogene Effects. *Blood*, **124**(21), 4591.

Bruna, A., Rueda, O. M., Greenwood, W., Batra, A. S., Callari, M., Batra, R. N., Pogrebniak, K., Sandoval, J., Cassidy, J. W., Tufegdzic-Vidakovic, A., Sammut, S.-J., Jones, L., Provenzano, E., Baird, R., Eirew, P., Hadfield, J., Eldridge, M., Mclaren-Douglas, A., Barthorpe, A., Lightfoot, H., O 'Connor, M. J., Gray, J., Cortes, J., Baselga, J., Marangoni, E., Welm, A. L., Aparicio, S., Serra, V., Garnett, M. J., and Caldas, C. (2016). A Biobank of Breast Cancer Explants with Preserved Intratumor Heterogeneity to Screen Anticancer Compounds. *Cell*, **167**, 260–274.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., and Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology*, **30**(5), 413–21.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, **31**(3), 213–219.

Conway, T., Wazny, J., Bromage, A., Tymms, M., Sooraj, D., Williams, E. D., and Beresford-Smith, B. (2012). Xenome–a tool for classifying reads from xenograft samples. *Bioinformatics*, **28**(12), i172–i178.

Dawson, C. W., Port, R. J., Young, L. S., Yip, K. Y., Ko, C., Tsang, Y., Wong, N., Whitney, B., Lee, J., Tursz, T., Abbott, R., Hoog, J., Dooling, D., Koboldt, D., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wendl, M., McMichael, J., Magrini, V., Cook, L., McGrath, S., Vickery, T., Appelbaum, E., DeSchryver, K., Davies, S., Guintoli, T., and Lin, L. (2012). The role of the EBV-encoded latent membrane proteins LMP1 and LMP2 in the pathogenesis of nasopharyngeal carcinoma (NPC). *Seminars in Cancer Biology*, **22**(2), 144–153.

Day, C.-P., Merlino, G., and VanDyke, T. (2015). Preclinical Mouse Cancer Models: A Maze of Opportunities and Challenges. *Cell*, **163**(1), 39–53.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., McMichael, J. F., Wallis, J. W., Lu, C., Shen, D., Harris, C. C., Dooling, D. J., Fulton, R. S., Fulton, L. L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Wendl, M. C., Heath, S., Watson, M. A., Link, D. C., Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., Graubert, T. A., Mardis, E. R., Wilson, R. K., and DiPersio, J. F. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**(7382), 506–10.

Jia, P., Li, F., Xia, J., Chen, H., Ji, H., Pao, W., and Zhao, Z. (2012). Consensus rules in variant detection from next-generation sequencing data. *PLoS ONE*, **7**(6).

Kelley, M. J., Li, S., and Harpole, D. H. (2001). Genetic Analysis of the Beta-Tubulin Gene, TUBB, in Non-Small-Cell Lung Cancer. *Journal of the National Cancer Institute*, **93**(24), 1886–1888.

Khandelwal, G., Girotti, M. R., Smowton, C., Taylor, S., Wirth, C., Dynowski, M., Frese, K. K., Brady, G., Dive, C., Marais, R., and Miller, C. (2017). Next-Gen Sequencing Analysis and Algorithms for PDX and CDX Models. *Molecular Cancer Research*.

Knudsen, E. S., Balaji, U., Mannakee, B., Vail, P., Eslinger, C., Moxom, C., Mansour, J., and Witkiewicz, A. K. (2017). Pancreatic cancer cell lines as patient-derived avatars: genetic characterisation and functional utility. *Gut*, pages gutjnl–2016–313133.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**(3), 568–76.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, **9**(8), 1–10.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**(14), 1754–60.

Mandelker, D., Schmidt, R. J., Ankala, A., McDonald Gibson, K., Bowser, M., Sharma, H., Duffy, E., Hegde, M., Santani, A., Lebo, M., and Funke, B. (2016). Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genetics in Medicine*, **18**(February), 1–8.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**(9), 1297–303.

Morgan, M., Pagès, H., Obenchain, V., and Hayden, N. (2017). Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import.

Müller, C. I., Miller, C. W., Hofmann, W. K., Gross, M. E., Walsh, C. S., Kawamata, N., Luong, Q. T., and Koeffler, H. P. (2007). Rare mutations of the PIK3CA gene in malignancies of the hematopoietic system as well as endometrium, ovary, prostate and osteosarcomas, and discovery of a PIK3CA pseudogene. *Leukemia Research*, **31**(1), 27–32.

Ng, S. B., Nickerson, D. A., Bamshad, M. J., and Shendure, J. (2010). Massively parallel sequencing and rare disease. *Human Molecular Genetics*, **19**(R2), 119–124.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. S., Papaemmanuil, E., Davies, H. R., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jönsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerød, A., Aparicio, S. A. J. R., Tutt, A., Sieuwerts, A. M., Borg, Å., Thomas, G., Salomon, A. V., Richardson, A. L., Børresen-Dale, A.-L., Futreal, P. A., Stratton, M. R., Campbell, P. J., and Breast Cancer Working Group of the International Cancer Genome Consortium (2012). The life history of 21 breast cancers. *Cell*, **149**(5), 994–1007.

Park, S. Y., Gönen, M., Kim, H. J., Michor, F., and Polyak, K. (2010). Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *Journal of Clinical Investigation*, **120**(2), 636–644.

Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., Meyerson, M., and Getz, G. (2015). Oncotator: Cancer variant annotation tool. *Human Mutation*, **36**(4), E2423–E2429.

Rossello, F. J., Tothill, R. W., Britt, K., Marini, K. D., Falzon, J., Thomas, D. M., Peacock, C. D., Marchionni, L., Li, J., Bennett, S., Tantoso, E., Brown, T., Chan, P., Martelotto, L. G., Watkins, D. N., and Coleman, W. B. (2013). Next-Generation Sequence Analysis of Cancer Xenograft Models. *PLoS ONE*, **8**(9).

Schneeberger, V. E., Allaj, V., Gardner, E. E., Poirier, J. T., Rudin, C. M., and Rots, M. (2016). Quantitation of Murine Stroma and Selective Purification of the Human Tumor Component of Patient-Derived Xenografts for Genomic Analysis. *PLOS ONE*, **11**(9), e0160587.

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., Steidl, C., Holt, R. A., Jones, S., Sun, M., Leung, G., Moore, R., Severson, T., Taylor, G. A., Teschendorff, A. E., Tse, K., Turashvili, G., Varhol, R., Warren, R. L., Watson, P., Zhao, Y., Caldas, C., Huntsman, D., Hirst, M., Marra, M. A., and Aparicio, S. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**(7265), 809–13.

Stratton MR (2011). Exploring the Genomes of Cancer Cells: Progress and Promise. *Science*, **331**(March), 1553–8.

Tanaka, Y., Kanai, F., Tada, M., Asaoka, Y., Guleng, B., Jazag, A., Ohta, M., Ikenoue, T., Tateishi, K., Obi, S., Kawabe, T., Yokosuka, O., and Omata, M. (2006). Absence of PIK3CA hotspot mutations in hepatocellular carcinoma in Japanese patients. *Oncogene*, **25**(20), 2950–2952.

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**(2), 178–192.

Tso, K.-Y., Lee, S., Lo, K.-W., and Yip, K. Y. (2014). Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts? *BMC Genomics*, **15**(1), 1172.

Witkiewicz, A. K., McMillan, E. A., Balaji, U., Baek, G., Lin, W.-C., Mansour, J., Mollaee, M., Wagner, K.-U., Koduru, P., Yopp, A., Choti, M. A., Yeo, C. J., McCue, P., White, M. A., and Knudsen, E. S. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nature communications*, **6**, 6744.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2010). Distant metas.tasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, **467**(7319), 1114–7.

Zhou, W., Zhao, H., Chong, Z., Mark, R. J., Eterovic, A. K., Meric-Bernstam, F., and Chen, K. (2015). ClinSeK: a targeted variant characterization framework for clinical sequencing. *Genome Medicine*, **7**(1), 34.

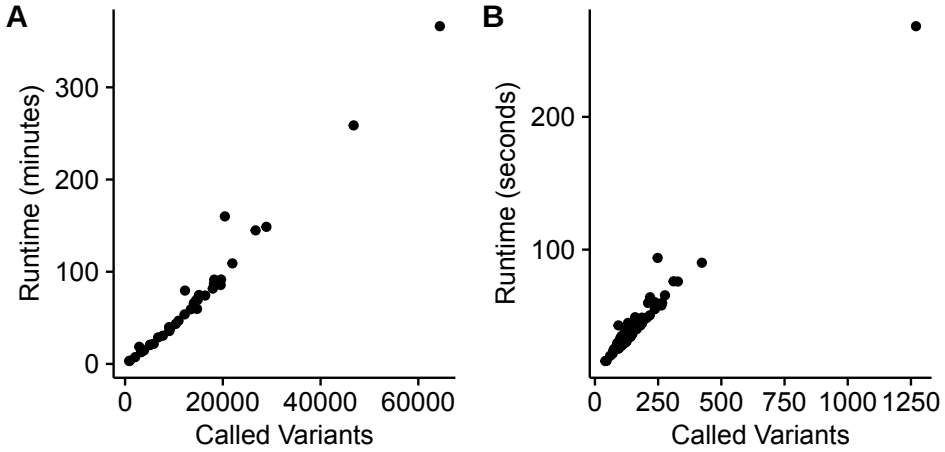

Figure S1: `mapexr` timing for A: xenografts and B: primary tumors. Shown are results from running on 4 cores and filtering all MuTect 1.1.1 calls for each sample. Run time is linear in the number of input variants, roughly one minute per 250 variants. One strategy for reducing run time is to first filter to keep only variants of interest, such as non-synonymous coding variants.

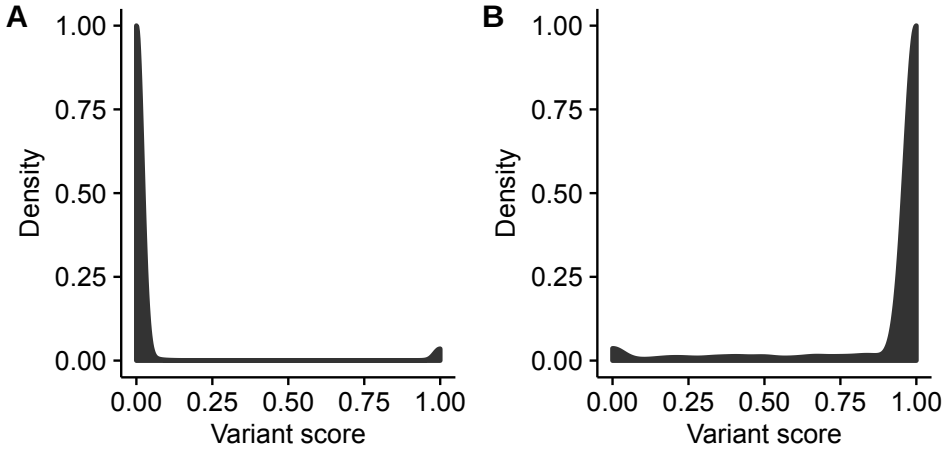

Figure S2: Distribution of variant scores from MuTect 1.1.1 over all A: PDX samples and B: Primary samples.

Table S1: MAPEX removes many potentially spurious PDX variants and retains almost all likely real variants that are also found in the primary.

	Total variants in PDX		Primary variants in PDX		
PDX	Before MAPEX	After MAPEX	Before MAPEX	After MAPEX	Fraction retained by MAPEX
EMC1229x1a1	2086	30	7	7	1.00
EMC828o3a5	1734	46	27	27	1.00
EMC828x2a3	1417	45	28	28	1.00
EMC828x2b2	1461	50	29	29	1.00
EMC828x3b1	228	37	29	29	1.00
EMC129x2b1	688	31	28	28	1.00
EMC1222o2a1	162	62	40	38	0.95
EMC1222o2a2	415	55	40	37	0.93
EMC1222o2a3	556	58	41	38	0.93
EMC1222o2a3duodenalMet	814	54	38	36	0.95
EMC1222o2a3omentalMet	390	50	39	37	0.95
EMC1222o2a3peritonealMet	379	57	39	37	0.95
EMC1222o2a3skinMet	951	59	39	37	0.95
EMC1222o2a3spleenMet	1508	62	40	37	0.93
EMC1222x1b1	639	52	41	39	0.95
EMC1222x3a1	545	69	40	37	0.93
EMC1222x3c2	157	64	39	38	0.98
EMC226o1a5	1495	78	64	64	1.00
EMC226o1a5met	2820	77	64	62	0.97
EMC226x1a1	1229	71	65	64	0.98
EMC226x1a2	1468	72	69	66	0.96
EMC26o1a2	1828	30	17	17	1.00
EMC29o1a1	6559	38	24	24	1.00
EMC29o1a1liverMet	4744	42	28	28	1.00
EMC29o1a1liverMet_1	2981	42	28	28	1.00
EMC29o1a1peritonealMet	1421	45	29	29	1.00
EMC29o1a1spleenMet	1154	43	28	28	1.00
EMC29o1a2	1983	41	27	27	1.00
EMC519x1a1	926	23	14	14	1.00
EMC93o2a3	1122	81	46	45	0.98
EMC93o2a3periMet	2074	65	46	45	0.98
EMC93o2a3spleenMet	2330	91	46	45	0.98
EMC93x1a1	988	92	47	45	0.98