Extended methods

Enrollment and exclusion criteria

Enrollment criteria included locally advanced or metastatic cancer, predominantly having received one or more lines of therapy in the metastatic setting, and ECOG ≤ 1. Patients with life expectancy <6 months were excluded. Of consented and enrolled patients, 692 (79%) underwent biopsies. Patients who did not receive a biopsy did so for a variety of reasons including high risk to the patient, deterioration of performance status, or loss to follow up. 592 patients (86% of biopsied patients) had samples that passed quality control, with the majority of sample failures due to a lack of sufficient tumor content as identified in pathology review (<40%). An additional 22 patients were excluded from the analysis cohort after sequencing for reasons including poor sample quality at nucleic acid extraction or sequencing library preparation, poor sequencing data obtained, insufficient tumor content based upon sequencing data analysis or incomplete clinical data. 570 patients were included in the final cohort.

Tissue collection and pathology review

Tumor specimens were collected using ultrasound- or CT-guided needle core biopsies, endobronchial ultrasound biopsies, or tissue resection after which samples were immediately embedded in a small amount of optimal cutting temperature (OCT) compound (Tissue-Tek, Sakura) for snap freezing on dry ice. Liquid biopsies or procedures such as pleurocentesis were performed as needed with material spun down into a cell pellet resuspended in PBS and submitted for extraction. Scrolls from the tissue were sectioned at 50 µm, with intervening 10 µm sections for hematoxylin & eosin staining, until sufficient scrolls were harvested or the tissue was exhausted. The stained sections were used for pathology review of tumor content and cellularity, and remaining sections used for DNA and RNA extractions. Four 50 µm sections were added to 2.0 mL tubes containing 420-600 µL RLT Plus lysis buffer (Qiagen) containing the reducing agent tris (2-carboxyethyl) phosphine (TCEP). Co-extraction of DNA and RNA from 3 to 11 tubes, selected for optimal tumor content and cellularity, was performed using an ALine EvoPure kit (Aline Biosciences, R-907-400-C5) automated on MicroLab NIMBUS (Hamilton) liquid handling robot. Constitutional DNA representing normal cells was extracted from peripheral blood using an AutoGen instrument.

Whole genome sequencing library construction

To minimize library bias and coverage gaps associated with PCR amplification of high GC or AT-rich regions, a version of the TruSeq DNA PCR-free kit (E6875-6877B-GSC, New England Biolabs), automated on a Microlab NIMBUS liquid handling robot (Hamilton) was employed. Briefly, 500ng of genomic DNA was arrayed into wells in a 96-well microtitre plate and subjected to shearing by sonication (Covaris LE220). Sheared DNA was end-repaired and size-selected using paramagnetic PCRClean DX beads (C-1003-450, Aline Biosciences), targeting a 350-450 bp size range. After 3' A-tailing, full length TruSeq adapters were ligated to DNA fragments. Libraries were purified using paramagnetic (Aline Biosciences) beads. Prior to sequencing, PCR-free genome library concentrations were quantified using a qPCR Library Quantification kit (KAPA, KK4824).

Strand-specific RNA library construction

Qualities of total RNA samples were determined using an Agilent Bioanalyzer RNA Nanochip or Caliper RNA assay and arrayed into a 96-well plate (Thermo Fisher Scientific). Polyadenylated (poly(A)) RNA was purified using the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490L, NEB) from 500 ng total RNA normalized in 35 µL for DNase I-treatment (1 Unit, Invitrogen). DNase-treated RNA was purified using RNA MagClean DX beads (Aline Biosciences, C-1005-250) on a Microlab NIMBUS liquid handler (Hamilton Robotics, USA). Messenger RNA (mRNA) selection was performed using NEBNext Oligod$(T)_{25}$ beads (NEB) with incubation at 65°C for 5 minutes followed by snap-chilling at 4°C to denature RNA and facilitate binding of poly(A) mRNA to the beads. mRNA was eluted in 36 µL of tris buffer (pH 7.4).

In cases with RNA Integrity Numbers <7.0, ribosomal RNA depletion RNA sequencing was employed. To remove cytoplasmic and mitochondrial ribosomal RNA (rRNA) species from total RNA NEBNext rRNA Depletion Kit for Human/Mouse/Rat was used (NEB, E6310X). Enzymatic reactions were set up in a 96-well plate (Thermo Fisher Scientific) on a Microlab NIMBUS liquid handler (Hamilton Robotics, USA). 120 ng of DNase I treated total RNA in 12 µL was hybridized to rRNA probes in a 15 µL reaction. Heat-sealed plates were incubated at 95°C for 2 minutes followed by incremental reduction in temperature by 0.1°C per second to 22°C (730 cycles). The rRNA in DNA hybrids were digested using RNase H in a 20 µL reaction incubated in a thermocycler at 37°C for 30 minutes. To remove excess rRNA probes (DNA) and residual genomic DNA contamination, DNase I was added in a total reaction volume of 50 µL and incubated at 37°C for 30 minutes. RNA was purified using RNA MagClean DX beads (Aline Biosciences, USA) with 15 minutes of binding time, 7 minutes clearing on a magnet followed by two 70% ethanol washes, 5 minutes to air dry the RNA pellet and elution in 37 µL DEPC water.

First-strand cDNA was synthesized from the purified polyadenylated mRNA or rRNA depleted total RNA using the Maxima H Minus First Strand cDNA Synthesis kit (Thermo-Fisher, FSSP9760210) and random hexamer primers at a concentration of 5µM along with a final concentration of 1 µg/µL Actinomycin D, followed by PCRClean DX bead purification on a Microlab NIMBUS robot (Hamilton Robotics, USA). Second strand cDNA was synthesized following the NEBNext Ultra Directional Second Strand cDNA Synthesis protocol (NEB) which incorporates dUTP in the dNTP mix, allowing the second strand to be digested using USER^{TM} enzyme (NEB) in the post-adapter ligation reaction, thus achieving strand specificity.

cDNA was fragmented using Covaris LE220 sonication for 55 seconds at a "Duty cycle" of 20% and "Intensity" of 5 to achieve 200-250 bp average fragment lengths. The paired-end sequencing library was prepared using a strand-specific, plate-based library construction protocol on a Microlab NIMBUS robot (Hamilton Robotics, USA). Briefly, the sheared cDNA was subject to end-repair and phosphorylation in a single reaction using an enzyme premix (NEB) containing T4 DNA polymerase, Klenow DNA Polymerase and T4 polynucleotide kinase, incubated at 20^{o}C for 30 minutes. End-repaired cDNA was purified in 96-well format using PCRClean DX beads, and 3' A-tailed (adenylation) using Klenow fragment (3' to 5' exo minus) and incubation at 37^{o}C for 30 minutes prior to enzyme heat inactivation. Illumina PE adapters were ligated at 20^{o}C for 15 minutes. The adapter-ligated products were purified using PCR Clean DX beads, then digested with USER^{TM} enzyme (1 U/µL, NEB) at 37^{o}C for 15 minutes followed immediately by 13 cycles of indexed PCR using Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) and Illumina's PE primer set. PCR parameters were: 98˚C for 1 minute, followed by 13 cycles of 98˚C 15 seconds, 65˚C 30 seconds and 72˚C 30 seconds, and then 72˚C for 5 minutes. The PCR products were purified twice and size-selected using a 1:1 PCRClean DX beads-to-sample ratio. The eluted DNA quality was assessed using the Caliper LabChip GX for DNA samples and the High Sensitivity Assay (PerkinElmer, Inc. USA). Quantification was performed using a Quant-iT dsDNA High Sensitivity Assay Kit on a Qubit fluorometer (Invitrogen). Libraries were then pooled and size-selected to adjust the final library molar concentration for sequencing.

Somatic alterations

Sequence reads from normal and tumor whole genome libraries were aligned to the human reference genome (hg19) using the Burrows-Wheeler Alignment tool[1] (v0.5.7 for up to 125 bp reads and v0.7.6a for 150 bp reads). Tumor genome sequences were compared to those from the patient's constitutive (normal) DNA to identify somatic alterations. Regions of copy number variation and losses of heterozygosity were identified using the Hidden Markov model-based approaches CNAseq[2] (v0.0.6) and APOLLOH[3] (v0.1.1), respectively. Regions of amplification were defined as those with total copies greater than twice the estimated tumor ploidy, and deep deletions were defined as those with less than half the estimated tumor ploidy. Somatic single nucleotide variants (SNVs) were identified using two approaches: (1) putative somatic variant calls from SAMtools[4] (v0.1.17) with subsequent scoring by machine-learning based MutationSeq[5] (v1.0.2 and v4.3.5), and (2) identification and scoring with the joint caller Strelka[6] (v1.0.6). Only consensus SNVs with MutationSeq probability >= 0.85 and Strelka QSS >= 15 were used in downstream analyses. Small (<20 bp) insertions and deletions (indels) were identified using Strelka with QSI >= 15. Structural variants (SVs) in RNA-Seq data were identified using the assembly-based tools ABySS[7] v1.3.4 and TransABySS[7,8] (v1.4.10) and alignment-based tools Chimerascan[9] (v0.4.5) and DeFuse[10] (v0.6.2); SVs in the DNA sequence data were identified using assembly-based tools ABySS and Trans ABySS and alignment-based tools Manta v1.0.0[11] and Delly[12] v0.7.3. Putative SV calls identified from the DNA and RNA sequences were merged into a consensus caller MAVIS[13] (v2.1.1), where they were clustered, computationally validated and annotated against constitutional DNA to provide somatic and germline structural variant calls. Both DNA- and RNA-derived structural variant calls were additionally filtered to identify those called by more than one tool, and for which a contig could be assembled that aligned across a candidate genomic breakpoint. DNA SV calls were further filtered to exclude events with identical genomic breakpoints in multiple samples, removing potentially confounding germline variants and technical artifacts. Variants were annotated to genes using SNPEff[14] (v3.2) with the Ensembl database[15] (v69).

Mutation positional clustering and kataegis

To focus on events most likely to have biological significance, any cluster with less than 5 mutations was removed from downstream analysis. To remove artifacts and germline events remaining in the dataset, all mutations in the remaining clusters were filtered against matched normal data from the entire cohort, summarized at equivalent positions using samtools[4] mpileup (v0.1.17); any mutation that occurred in at least 2% of the matched normal data and at an allele fraction equal to or higher than the tumor variant was filtered out. The remaining mutations were then re-clustered and evaluated for patient frequency. Any cluster found in less than 5 patients (equivalent to <1% of the cohort) was removed.

Cluster significance was calculated using a binomial distribution[16], where n is the total number of patients and p_i is the probability of mutation for the patient:

$$P = 1 - \sum_{j=0}^{k-1} \left(\quad n j \quad \right) p_i^j (1 - p_i)^{n-j}$$

We calculated p_i based on the length L_i of the cluster as well as the background mutation rate q_i. The background mutation rate was calculated as the average of all non-clustered mutations within 10kb upstream and downstream of the cluster:

$$p = 1 - (1 - q_i)^{L_i}$$

Clustered mutations were not included in the calculation to avoid an overestimation of the rate of mutation.

Kataegis events[17] were defined as six or more consecutive mutations with average intermutation distances of ≤1 kb. Additionally, only events where ≥ 50% of mutations in the region of interest were comprised of C>T or C>G substitutions were considered to be true kataegis events. Mutations in kataegis regions were then filtered against matched normal data from the entire cohort, as described above for non-coding clustering. Filtered mutations were subsequently regrouped into kataegis events for subsequent analysis.

Tumour heterogeneity

For SNVs and copy number alterations used to predict the presence of subpopulations using EXPANDS[18], one case was excluded from tumour heterogeneity analyses as no SNVs were detected. As EXPANDS has a recommended upper limit of 8,000 mutations per sample, and many POG570 samples had a greater number than this, a subset of mutations was selected to include all coding variants and TERT promoter mutations with the addition of randomly sampled non-coding variants, up to a maximum of 8,000 per case. To evaluate the confidence of heterogeneity detection in our cohort, for each sample with at least 2000 somatic SNVs we additionally performed 100 random resamplings of 1000 mutations and analyzed these with EXPANDS; the average percent standard error of heterogeneity was 0.84, confirming the stability of our results. In addition, we found that while overall heterogeneity per sample was lower in the results using an input of 1000 versus the input of 8000 (Extended Data Fig. 2d), supporting the concept that increasing the amount of input data increases the potential for detection of additional subpopulations, there was a strong correlation in heterogeneity between the two datasets (Extended Data Fig. 2e, R=0.74, $p<2.2e^{-16}$, Spearman correlation).

Comparison to primary tumour datasets

As the distribution of therapies and mutated genes varies substantially by tumor type in our cohort, we performed comparison of alteration frequency in POG570 and TCGA[19] on each tumor type separately. Drug treatments grouped by mechanism of action (Supplementary Table 2) given to at least 10% of the tumor type cohort were examined for associations. Power analysis indicated that a sample size of 31 patients was needed to have 80% power of identifying an effect size of 0.5 at a significance level of 0.05. Tumor type groups of this size were: BRCA, LUNG, COLO, OV, SARC and PANC (see Fig. 1a). Of these, SARC was excluded because the distribution of subtypes in our sarcoma cohort is substantially different to that in TCGA, making the frequency of gene alterations not directly comparable. Protein altering mutations were identified using SNPEff annotations, analyzed as described above, with ‘Moderate’ or ‘High’ impact. Amplifications were defined as regions of copy number 5 or more as defined by CNASeq[2], and deletions were defined as regions of copy number 0.

DNA repair and genotoxic therapy

As the number of gene mutations is related to overall TMB, we sought to confirm that the increase in TMB in cases with DNA repair pathways is more than would be expected due only to this correlation. We performed 1000 iterations of random mutation subsampling; for each iteration 181 genes were randomly selected to represent a ‘pathway’ and the ratio of mutation burden in cases with ‘pathway’ mutations to those without was computed. The distribution of ratios computed based on these random 'pathway' groups were then used to compute a p-value for the observed ratio for DNA-repair mutated cases.

Mutation signatures and timing

SBSs were categorized based on 6 variant types and 16 trinucleotide context subtypes to yield a total of 96 mutation classes. Indels were categorized into five broad classes and sub-categorized based on repeat content to yield 83 mutation classes. Double base substitutions were classified into on the 78 possible strand-agnostic dinucleotide substitutions.

Signature stability estimates were obtained by bootstrap resampling with 1,000 iterations. The solution which best maximizes signature stability and minimizes Frobenius reconstruction error was chosen for each cohort with the formula:

$$\mathrm{argmin}_n \frac{R_n - \min(\mathrm{R})}{\max(\mathrm{R}) - \min(\mathrm{R})} - \frac{S_n - \min(\mathrm{S})}{\max(\mathrm{S}) - \min(\mathrm{S})}$$

Sn and Rn are the signature stability and reconstruction error values for the n-signature solution and S and R are vectors containing stability and reconstruction error values for all values of n. Mutation signature analysis of a total of 23 tumor type cohorts was attempted. Of the 23 cohorts tested, all but 12 cohorts failed SBS mutation signature analysis because of (1) too few samples, (2) too few SBSs, or (3) excessive heterogeneity in mutation signatures (as was observed for the MISC cohort). An analysis was marked failed if every sample had its own private mutation signature (meaning dimensionality reduction did not take place) or if the stability and reconstruction error estimates were poor across all attempted models.

SignIT (https://github.com/eyzhao/SignIT) uses a Bayesian hierarchical model to jointly infer mutation signatures and the prevalence and size of temporally distinct tumor subpopulations. Cases which fit models described by greater than one subpopulation can be subject to mutation signature timing analysis. SignIT requires the annotation of SNV calls with tumor and normal copy number. Prior to annotation, CNVs from CNAseq were first corrected for ploidy using the following formula

$$C^{(T)} = \frac{(\bar{R}+1)\left(TP + C^{(N)}(1-T)\right) - C^{(N)}(1-T)}{T}$$

Where R is the mean tumor-to-normal read depth ratio across the segment, T is the tumor content, and P is the ploidy. CT is the estimated absolute tumor copy number of the segment and was rounded to the nearest whole number, and CN is the normal copy number, assumed to be 2. SNVs in regions with greater than 5 copies were filtered out, as precise copy number estimation becomes difficult.

The fraction of late-arising mutations was computed as

$$\frac{\text{late exposure}}{\text{late exposure} + \text{early exposure}}$$

and could vary from 0 for early mutation signatures to 1 for late mutation signatures.

Germline mutations

Germline SNVs and indels were identified in normal DNA using samtools[4] (v0.1.17), annotated using SNPEff[14] v4.1 , population minor allele frequencies derived from the 1000 genomes[20] v.1000g2015aug, and pathogenicity annotated using ClinVar[21] v.20180905. Copy number variants (CNVs) in the germline genomes were identified using ControlFREEC[22]. An orthogonal approach was used for combined CNV, structure variants (SV) and indel calling and annotation. Germline CNV, SV and indel calls from DELLY[12] v0.7.3, ABySS[7] v1.3.4 and manta[11] v1.0.0 were aggregated and further annotated with Trans-ABySS[7,8] v1.4.10 transcriptome data using MAVIS[13].

Immune signatures and clonotypes

The LM22 cell subtype signature, composed of 547 genes, was used to predict the presence of 22 immune cell subtypes in each RNA-Seq sample. CIBERSORT[23] was run without quantile normalization with 1000 permutations on reads per kilobase per million mapped reads (RPKM) data across all samples to generate absolute scores for each cell type. RPKMs were calculated by aligning RNA-Seq reads against a database of exon junction sequences, processing to reposition all read alignments with gaps onto the same genomic reference using Jaguar[24] (v2.0.3), and calculation of gene-level RPKMs based on Ensembl gene models (v69).

TCR clonotypes were determined from RNA-Seq data using MiXCR[25] (v2.1.2). Raw RNA-Seq reads were aligned to reference V, D, J and C genes of human T cell receptors. Two rounds of partial assembly were followed by an alignment extension step and a final assembly step to determine T cell clonotype sequences. Out-of-frame sequences and sequences containing a stop codon were filtered out. Post-analysis of TCR repertoire data was performed using VDJtools[26] (v1.1.9). Shared clonotypes and distances between repertoires were determined based on shared CDR3 amino acid sequences.

References

1. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **26**, 589–595 (2010).
2. Jones, S. J. *et al.* Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* **11**, R82 (2010).
3. Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* **22**, 1995–2007 (2012).
4. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
5. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).
6. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
7. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).

8. Birol, I. *et al.* De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872–2877 (2009).

9. Iyer, M. K., Chinnaiyan, A. M. & Maher, C. A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27**, 2903–2904 (2011).

10. McPherson, A. *et al.* deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. *PLoS Comput. Biol.* **7**, e1001138 (2011).

11. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

12. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

13. Reisle, C. *et al.* MAVIS: merging, annotation, validation, and illustration of structural variants. *Bioinformatics* **35**, 515–517 (2019).

14. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

15. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).

16. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).

17. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).

18. Andor, N., Harness, J. V., Müller, S., Mewes, H. W. & Petritsch, C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinforma. Oxf. Engl.* **30**, 50–60 (2014).

19. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).

20. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**,

68–74 (2015).

21. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

22. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).

23. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

24. Butterfield, Y. S. *et al.* JAGuaR: Junction Alignments to Genome for RNA-Seq Reads. *PLoS ONE* **9**, e102398 (2014).

25. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).

26. Shugay, M. *et al.* VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Comput. Biol.* **11**, e1004503 (2015).