Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes

Erin Pleasance[1*], Emma Titmuss[1*], Laura Williamson[1*], Harwood Kwan[1], Luka Culibrk[1], Eric Y. Zhao[1], Katherine Dixon[6], Kevin Fan[1], Reanne Bowlby[1], Martin R. Jones[1], Yaoqing Shen[1], Jasleen K. Grewal[1], Jahanshah Ashkani[1], Kathleen Wee[1], Cameron J. Grisdale[1], My Linh Thibodeau[1,3,6], Zoltan Bozoky[1], Hillary Pearson[7], Elisa Majounie[1], Tariq Vira[1], Reva Shenwai[1], Karen L. Mungall[1], Eric Chuah[1], Anna Davies[1], Mya Warren[1], Caralyn Reisle[1], Melika Bonakdar[1], Gregory A. Taylor[1], Veronika Csizmok[1], Simon K. Chan[1], Stuart Zong[1], Steve Bilobram[1], Amir Muhammadzadeh[1], Darryl D'Souza[1], Richard D. Corbett[1], Daniel MacMillan[1], Marcus Carreira[1], Caleb Choo[1], Dustin Bleile[1], Sara Sadeghi[1], Wei Zhang[1], Tina Wong[1], Dean Cheng[1], Scott D. Brown[1], Robert A. Holt[1], Richard A. Moore[1], Andrew J. Mungall[1], Yongjun Zhao[1], Jessica Nelson[1], Alexandra Fok[1], Yussanne Ma[1], Michael K.C. Lee[2], Jean-Michel Lavoie[2], Shehara Mendis[2], Joanna M. Karasinska[8], Balvir Deol[2], Ana Fisic[2], David F. Schaeffer[5,8], Stephen Yip[5], Kasmintan Schrader[3,6], Dean A. Regier[9], Deirdre Weymann[9], Stephen Chia[2], Karen Gelmon[2], Anna Tinker[2], Sophie Sun[2], Howard Lim[2], Daniel J. Renouf[2,8], Janessa Laskin[2], Steven J.M. Jones[1,4,6], Marco A. Marra[1,6]

[1]Canada's Michael Smith Genome Sciences Centre at BC Cancer, Vancouver BC, Canada; [2]Department of Medical Oncology, BC Cancer, Vancouver BC, Canada; [3]Hereditary Cancer Program, BC Cancer, Vancouver BC, Canada; [4]Department of Molecular Biology and Biochemistry, Simon Fraser University, Vancouver BC, Canada; [5]Department of Pathology and Laboratory Medicine, Vancouver General Hospital, Vancouver BC, Canada; [6]Department of Medical Genetics, University of British Columbia, Vancouver BC, Canada; [7]Department of Medicine, University of British Columbia, Vancouver BC, Canada;

[8]Pancreas Centre BC, Vancouver BC, Canada; [9]Canadian Centre for Applied Research in Cancer Control , Cancer Control Research, BC Cancer, Vancouver BC, Canada; [*]These authors contributed equally;

Corresponding author: Marco A. Marra; mmarra@bcgsc.ca

ABSTRACT

Advanced and metastatic tumors with complex treatment histories drive cancer mortality. Here we describe the POG570 cohort, a comprehensive whole genome, transcriptome and clinical dataset, amenable for exploration of the impacts of therapies on genomic landscapes. Prior exposure to DNA-damaging chemotherapies, and mutations affecting DNA repair genes, including *POLQ* and genes encoding Polζ, were associated with genome-wide, therapy-induced mutagenesis. Exposure to platinum therapies coincided with signatures SBS31 and DSB5, and when combined with DNA synthesis inhibitors, signature SBS17b. Alterations in *ESR1, EGFR, CTNNB1, FGFR1, VEGFA*, and *DPYD* were consistent with drug resistance and sensitivity. Recurrent non-coding events were found in regulatory region hotspots of genes including *TERT, PLEKHS1, AP2A1*, and *ADGRG6*. Mutation burden and immune signatures corresponded with overall survival and response to immunotherapy. Our data offer a rich resource for investigation of advanced cancers and interpretation of whole genome and transcriptome sequencing in the context of a cancer clinic.

INTRODUCTION

Application of tumor sequencing in cancer management[1], and large-scale cancer genomic profiling has transformed our understanding of the genomic events that drive cancers[2]. The majority of clinical cancer sequencing has used panel approaches[3,4], and whole exome, genome, and transcriptome profiling studies have often concentrated on primary tumors with limited prior therapy[5,6]. Despite the high mortality rate associated with advanced disease[7], few studies describe comprehensive pan-cancer profiles of advanced, post treatment tumors from mixed histology patient populations[8,9]. Treatment with anti-cancer compounds can impact the tumor genomic landscape through selection for resistant clones and a few resistance mechanisms to targeted therapies have been described[10,11]. Therapies such as platinum-based compounds have been associated with DNA mutagenesis[12–14] and specific mutational signatures[13]. Genomic features beyond gene mutations, including mutation signatures and transcriptome profiles, have clinical implications[15,16].

Here we describe analysis of 570 advanced and metastatic cancers from patients treated at a tertiary care center, referred to as the 'POG570' cohort, profiled using whole genome and transcriptome sequencing as part of the Personalized OncoGenomics (POG) Program at BC Cancer[17–20] (NCT02155621; https://www.personalizedoncogenomics.org/cbioportal/). Examination of the interaction between drug treatment and genomic landscapes in this predominantly post-treatment population identified gene alterations, mutation signatures, increased mutation burden and genome instability in treated tumors. Results from gene expression profiles of the immune microenvironment correlated with overall survival, and with patient outcomes upon subsequent treatment with immune checkpoint inhibitors (ICIs). Our analyses reveal potential mechanisms of resistance and impacts of therapy, and illustrate the value of performing whole genome and transcriptome profiling on advanced cancers in a clinical context.

RESULTS

POG570 cohort characteristics

The cohort is composed of advanced and metastatic tumors from patients treated in a tertiary cancer clinic, and represents 25 histologies with biopsies from 18 organ groups. Most biopsies were taken from metastatic sites (n=438, 77%) while others represented local recurrences or refractory disease (Fig. 1a, Supplementary Table 1). Rare cancers were profiled, including eccrine porocarcinoma[21], carcinoma ex pleomorphic adenoma[22], and ghost cell odontogenic carcinoma[23].

Most patients (n=466, 82%) received systemic therapy prior to biopsy (Fig. 1b). 110 different drugs were received by patients before genomic analysis, with mean treatment duration for each drug ranging from four days to over four years (Extended Data Fig. 1a, Supplementary Table 2). Treatment duration reflected a combination of standard treatment protocol, discontinuation due to toxicity, or discontinuation due to disease progression. Overall, 72% of patients received more than one drug prior to biopsy, and drug combinations used were tumor type dependent as demonstrated by differences in drug co-occurrence (Fig. 1c, Extended Data Fig. 1b).

Recurrent mutations and genomic alterations in advanced cancers

Whole tumor and normal genome sequences were analyzed to identify somatic events. A total of 7,441,311 somatic substitutions and 701,166 small (<20 bp) insertions or deletions were detected, with mutation burdens ranging from <0.1/megabase (Mb) to 159/Mb (Fig. 2a). The highest mutation loads were associated with microsatellite instability (MSI), or with C>T substitution patterns in melanomas and

skin cancers, associated with UV exposure. Deep deletions affected a median of nine genes per genome, and mid- to high-level amplifications affected a median of 709 genes per genome. In 4% of cases, viral or microbial sequences were detected, including *Fusobacterium*, human herpesvirus, and human papillomavirus (HPV).

58,638 structural variants encompassing genomic regions >1 kb consisted of deletions (28%), inversions (24%), duplications (22%), and interchromosomal events (26%). Expressed transcripts for 2,291 events were detected and predicted to encode altered protein products. Ninety-five percent of samples had evidence for potential gene fusions, of which 31% (1,815 out of 5,831 events) were supported by RNA sequencing data. Sarcomas, breast, and ovarian cancers had the highest burden of potential gene fusions (96-100% of samples, averaging 12-17 events per sample), while colorectal and pancreatic cancers had lower rates of gene fusions (90-93% of samples, averaging five to eight per sample), consistent with results from primary cancers[24,25]. The most frequent therapeutically targetable driver fusion was EML4-ALK (Fig. 2a), observed exclusively in lung cancers (nine samples, 13%). NRG1 fusions, an emerging therapeutic target[19], were observed in cholangiocarcinoma, lung, and pancreatic cancers. Targetable RET and ROS1 fusions were detected, including a ROS1-GOPC fusion in a colorectal cancer, an event which has not previously been described in gastrointestinal tumors[26].

The most frequently altered oncogenes and tumor suppressor genes included *TP53*, *NF1*, *RB1*, *KRAS*, *CDKN2A/B*, and *MYC,* which we note are among the most frequently altered genes in primary cancers from The Cancer Genome Atlas (TCGA)[2] (Fig. 2a). Significantly mutated genes included *KEAP1* in lung, *SF3B1, GATA3* and *SOX10* in breast, and *MAPK8/JNK1* and *NCOA4* in a pan-cancer analysis (Supplementary Table 3; see Methods). *SOX10* was altered in 5 samples, two of which were triple negative breast cancers (TNBC) that harbored frameshift mutations with concurrent loss of *SOX10* gene

expression. SOX10 protein expression is a proposed adjunct diagnostic marker to GATA3 for TNBC[27]. SOX10-negative TNBC tumors are associated with elevated androgen receptor (AR) protein expression[28]. Consistent with this, both *SOX10* frameshift-containing TNBC samples demonstrated *AR* gene expression, supporting the relationship between these two clinical markers of TNBC and indicating *SOX10* mutations could account for a proportion of SOX10-negative TNBCs with concurrent AR expression.

Comparison of mutation frequencies of POG570 significantly mutated genes to primary tumors from the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset revealed higher *ESR1* mutation frequencies in advanced and metastatic breast cancers (all: 12.9% POG570 vs 1.2% PCAWG $p = 3.0x10^{-5}$, ductal: 11.5% POG570 vs. 0.9% PCAWG $p = 2.4x10^{-4}$, lobular: 37.5% POG570 vs. 0% PCAWG, $p = 1.5x10^{-3}$, Chi-squared test, FDR adjusted), and was validated in a pan-cancer, propensity-matched cohort (p = 0.017, Chi-squared, FDR adjusted), consistent with recent studies of advanced, pretreated tumors[8,29]. We observed a higher frequency of mutations affecting *PTEN* in ductal breast cancers compared to primary cancers (10.6% POG570 vs. 4.0% PCAWG, p = 0.09, Chi-s quared, FDR adjusted), supporting previous observations[29]. A similar elevation of *PTEN* mutation frequency was observed in metastatic ovarian serous cystadenocarcinomas (15% POG570 vs. 0% PCAWG, p value = 0.011, Chi-squared, FDR adjusted), consistent with results from MSK-IMPACT targeted sequencing in advanced cancers[3] in comparison to TCGA data (MSK-IMPACT 6% vs TCGA 1%).

Evaluation of copy number and single nucleotide variation revealed an increased frequency of *FGFR1* amplification (17% POG570 vs. 11% TCGA, p = 0.038, Chi-squared). Similarly, *NF1* was more frequently altered in breast and ovarian tumors when small mutations and copy number variants were combined (breast: 7% POG570 vs. 3% TCGA, p = 0.013, ovarian: 21% POG570 vs 10% TCGA, p = 0.057, Chi-squared).

Elevated alteration frequencies of *ESR1*, *FGFR1*, and *NF1* in our cohort were confirmed in MSK-IMPACT breast cancers (*ESR1*: 10%, *FGFR1*: 13%, *NF1*: 7%). Such gene alterations were previously associated with resistance to endocrine therapies[11,30,31], which is consistent with the treatment histories of patients in our cohort (Fig. 1b), indicating that these events may have contributed to treatment resistance.

Mutation hotspots include non-coding regulatory events

Almost all known highly recurrent cancer hotspot mutations are in protein coding regions[2–4]. Across POG570, 1.1% of small mutations were observed in coding and non-coding exons, 42% in intronic and 56% in intergenic regions. The mutation frequency in non-coding regions of the genome was, on average, 18% higher than in exonic regions (Extended Data Fig. 2a, $p<2.2x10^{-16}$, Wilcoxon rank sum), although several cases that harbored mutations in transcription-coupled repair genes *ERCC6*, *UVSSA* and *GTF2H1* did not exhibit this trend[32,33]. We identified 2,596 mutation clusters, of which 66.5% were in intergenic regions, 30.5% were in introns, and 3% were in exons (Fig. 2b, Extended Data Fig. 2b, Supplementary Table 4). Of the regulatory region clusters, including promoters, untranslated regions (UTRs) and enhancer regions, three were seen in more than 2% of patients (Fig. 2b): mutations in the *TERT* promoter, *ADGRG6* enhancer, and the *PLEKHS1* promoter. *TERT* promoter mutations are well-studied[34,35], and consistent with their oncogenic role, were associated with increased *TERT* expression (Fig. 2b,c, p=0.038). *PLEKHS1* promoter mutations showed a trend towards an increased expression (p=0.08, Wilcoxon rank sum, FDR correction), in contrast to previous reports from primary tumours[36]. Six patients with basal cell melanoma or head and neck cancer had *AP2A1* promoter mutations situated downstream of the transcription start site and elevated *AP2A1* expression (Fig. 2b, c, p=0.038), a relationship that was not reported previously[36]. We found recurrent hotspot G>A (chr. 6: 142,706,206 bp) and C>T (chr. 6: 142,706,209 bp) changes in *ADGRG6*, which were described in 26% of primary

bladder cancers and are proposed to promote angiogenesis[37]. Our findings extend the tumor types in which such mutations are found to include advanced breast, lung and cervical cancers.

Tumor mutation burden, heterogeneity and survival

Tumor mutation burden (TMB) and intratumor heterogeneity (ITH) were variable across tumor types (Fig. 2a, Extended Data Fig. 2c, p=0.0046). We sought to evaluate the relationship between TMB and ITH, and impact on patient survival. ITH was surveyed using multiple parameters (see Methods), the results of which were strongly concordant (Extended Data Fig. 2e, r=0.74, $p<2.2x10^{-16}$), and revealed elevated ITH in tumors characterized by a high level of subclonality (Extended Data Fig. 2f, p=0.00011). The median size of subpopulations associated with driver mutations was higher than sub-populations associated with only non-driver mutations, indicating driver mutations were preferentially found in dominant subpopulations (Extended Data Fig. 2g, p=0.024, Wilcoxon rank sum).

ITH was highest in breast, and colorectal tumors, and was low in lymphomas and melanomas, despite a high TMB in the latter (Fig. 2d), consistent with a recent PCAWG report (Dentro *et al.*, http://biorxiv.org/lookup/doi/10.1101/312041). ITH was positively correlated with TMB (r= 0.58, $p<2.2x10^{-16}$, Spearman correlation), possibly reflecting an increased potential for ITH facilitated by greater TMB. Increased TMB was associated with poorer overall survival (Fig. 2e, $p=7.03x10^{-6}$), even when accounting for tumor type (Extended Data Fig. 2h, HR=1.52, p=0.000826), while ITH did not independently contribute to prognosis (Extended Data Fig. 2h, HR=1.05, p=0.68).

Mutations and copy number alterations associated with prior therapy

To study alterations potentially associated with drug resistance, we identified coding small mutations and copy number alterations that were more frequent in treated than untreated patients, and in comparison to primary tumors (see Methods). We expected to detect alterations related to therapy

alignment in addition to those that may have contributed to resistance. Of 35 significant coding small mutations, we focused on 13 that were clustered or truncating (Fig. 3a, Supplementary Table 5), as these mutation patterns are often associated with oncogenic effects. This revealed drug-mutation associations for *ESR1*, *EGFR*, *SMAD4*, *TP53*, *ARID1A*, *OR5H2*, *ANKRD12*, and *TCF7L2*. The relevance and validity of our approach was supported by the observation that 3 of the 13 associations detected were well-known resistance mutations in aromatase inhibitor-treated breast cancers (*ESR1*; D538X)[11] and EGFR inhibitor-treated lung cancers (*EGFR;* T790M)[10], as well as treatment sensitivity mutations in *EGFR* (L858R), which are used to select patients for therapy (Fig. 3a, Extended Data Fig. 3a). Interestingly, the presence of *ESR1* and *EGFR* mutations was associated with longer treatment durations prior to biopsy (Fig. 3c and Extended Data Fig. 3a, *ESR1* p=0.0025, *EGFR* p =0.081). Notably, two cases with *EGFR* T790M resistance mutations that were exposed to therapy for the longest duration additionally harbored mutations in the β-catenin encoding gene, *CTNNB1* (Extended Data Fig. 3a) (Fig. 3a). Alterations affecting genes in the WNT pathway have previously been reported in *EGFR*-mutant cancers treated with EGFR inhibitors[10]. Together, these observations suggest WNT signaling may cooperate with *EGFR* resistance mutations, thus contributing to EGFR inhibitor resistance, as previously proposed[38]. We identified 20 loci containing 153 copy number variants associated with treatment, which additionally correlated with gene expression changes (Fig. 3b, Supplementary Table 6). *ERBB2* (HER2) amplifications are used as a clinical biomarker for HER2-inhibitor therapy[39], and consistent with this, *ERBB2* (HER2) amplifications on 17q were associated with HER2-inhibitor therapy (Fig. 3b). *FGFR1* amplifications on 8p exhibited higher *FGFR1* expression ($p=7x10^{-7}$, Wilcoxon rank sum) and were associated with aromatase inhibitor treatment in breast cancer patients (Fig. 3b, p=0.022, Chi-squared). As *FGFR1* amplifications are not used as clinical markers to select patients for therapy, these observations are consistent with a role for *FGFR1* amplification in aromatase inhibitor resistance, as previously proposed[30].

Expression changes associated with therapy

Clustering of POG570 samples based on gene expression data revealed sample relationships compatible with the tissue of origin (Extended Data Fig. 3b). Treatment-associated changes in gene expression may reflect gene alterations used clinically to select patients for treatment, such as elevated *ESR1* expression associated with hormonal therapy (Fig. 3d, $p=8.1x10^{-6}$), or alterations that arose concomitant with treatment. *VEGFA*, a target of bevacizumab, exhibited higher expression in patients who had received bevacizumab than in those who had not (Fig. 3d, p=0.00058), even in patients treated for less than 90 days. As *VEGFA* is not used as a marker to align patients to therapy, the rapid increase in *VEGFA* expression may represent a potential compensatory resistance mechanism, counteracting the inhibitory effect of bevacizumab. Expression of *DPYD* was lower in colorectal cancer patients treated with 5-FU, particularly in patients treated for more than 90 days (Fig. 3d, p=0.0084). The enzyme, DPYD, degrades 5-FU. Reduced activity due to germline variants in *DPYD* predict toxicity to 5-FU[40], whereas somatic loss may contribute to sensitivity to 5-FU, as indicated by a recent case study harboring a *DPYD* structural variant[18]. Reduced *DPYD* gene expression in patients with longer duration of 5-FU is consistent with a role for reduced *DPYD* expression in 5-FU sensitivity.

Novel mutation signatures in advanced, pre-treated tumors

To characterize the broader mutational landscape in POG570, *de novo* single base substitution (SBS), insertion and deletion (ID) and double base substitution (DBS) mutation signature analyses were performed on tumor groups with sufficient patient numbers (see Methods), and the temporal distribution of SBS signatures was inferred[41,42] (see Methods) (Fig. 4 and Extended Data Fig. 4). Fifteen SBS, six DBS, and nine ID signatures matching known patterns were identified (Fig. 4a, Extended Data

Fig. 4)[43]. We identified six additional novel SBS signatures (MSBS1-MSBS6) and seven novel ID signatures (MID1-MID7) (Fig. 4a, Extended Data Fig. 5).

Early arising, aging-related SBS1 and indel signatures ID1 and ID2 were observed in most tumor types (Fig. 4a). Tobacco-associated SBS4 and ID3 were found in lung cancers and correlated with DBS2, a widely observed signature also associated with exogenous mutagens including tobacco smoke (Fig. 4a,b and Extended Data Fig. 6). Early-arising SBS7a, attributed to UV-associated mutations, and late-arising SBS38 were identified in the SKCM group and were anti-correlated (Fig. 4b, Extended Data Fig. 6). SBS38 was elevated in acral melanoma and a mucosal melanoma of the vulva that were not characterized by UV-associated SBS7a, indicating SBS38 may be a UV-independent, melanoma-associated signature. SBS3, SBS8, and ID6, associated with HRD, were observed in breast and ovarian cancers while SBS3 and SBS8 were also found in pancreatic and stomach cancers and sarcomas, as previously described[44,45]. APOBEC-associated SBS2, SBS13, and DBS11 signatures were found in breast cancers. HRD and APOBEC-associated SBS signatures were observed with both early and late timing, suggesting these mutational processes accompany tumor evolution (Fig. 4c). Of the six novel SBS signatures (Fig. 4a and Extended Data Fig. 5), some resembled known signatures. The predominantly early-arising MSBS1 signature matched signature 1B previously found in many primary tumors[46]. MSBS2 had a mutation profile most similar to APOBEC-associated SBS2 and SBS13 (cosine similarities 0.67 and 0.63 respectively, Extended Data Fig. 4a), potentially implying a similar mutational mechanism. The late-arising MSBS6 was similar to SBS7c (cosine similarity 0.66, Extended Data Fig. 4a) and correlated with SBS7a (Spearman correlation 0.59, Fig. 4b and Extended Data Fig. 6). The predominantly late-arising MSBS3 was most similar to SBS9 and SBS17b (cosine similarities 0.72 and 0.69 respectively, Extended Data Fig. 4a) and was observed in pancreatic and stomach cancers (Fig. 4a). Interestingly, samples with MSBS3 also had exposure of SBS36 and SBS30 (Spearman correlation 0.46, 0.32 respectively, Fig. 4b and Extended Data Fig. 6), both of which are associated with deficient base excision repair (BER)[47,48], despite the absence of an obvious

mutation affecting BER in these samples. MSBS5, characterized by T to G transversions and T to C transitions, was detected in two samples and is of unknown origin. Of the novel indel signatures, MID1, identified in multiple tumor groups (Fig. 4a), exhibited large insertions and deletions (Extended Data Fig. 5b), particularly in regions of varying microhomology length, similar to HRD-associated ID6 and radiation-associated ID8. These observations add to the growing literature supporting indel signatures as markers of mutagen exposure and DNA damaging etiologies[49,50].

Tumor alterations increased by DNA repair mutations and long exposure to genotoxic therapy

Increased TMB as a result of mismatch repair (MMR) defects has been well documented[51]. Increased TMB in nucleotide excision repair- and BER-defective preclinical models has also been reported[52]. To examine the impact of DNA repair deficiencies on genomic landscapes, we surveyed 181 DNA repair and damage response genes representing 12 DNA damage response (DDR) pathways (Supplementary Table 7, see Methods) and identified 357 patients with somatic mutations. The most frequently altered genes included *TP53* (36%), *ATM* (2.6%), and *BRCA2* (2.3%) (Extended Data Fig. 7a). Across the DDR pathways analyzed, TMB was consistently higher in cases with DNA repair mutations, even after excluding hypermutated cases (Fig. 5a, $p<2.2x10^{-16}$). This observation held even when accounting for the relationship between overall mutation burden and mutations in a specific gene set (p=0.033, see Methods). Genomes with DDR mutations also exhibited increased structural instability, measured by HRD score, which has been associated with, but not limited, to loss of the DNA repair genes *BRCA1* and *BRCA2*[53] (Extended Data Fig. 7b, $p=7.1x10^{-10}$). Together, our data indicate that mutations affecting DDR pathways impact overall mutation burden and genome stability.

An estimate of the mutational toxicity of chemotherapy by assessing therapy-associated mutational footprints supports the notion that prior therapy significantly contributes to the overall TMB[54]. Excluding patients with DDR mutations, we found that patients treated with genotoxic chemotherapy therapy for more than one year exhibited a significant increase in TMB compared to those treated for less than one year or not treated at all (see Methods) (Fig. 5b, untreated vs >1 year p=0.00018), even after accounting for tumor type (untreated vs >1 year p=0.00012, linear regression). Patients undergoing long-term therapy prior to biopsy showed an average increase of 4,304 (2.0-fold increase) somatic mutations and had, on average, 32 (2.0-fold increase) more mutated genes than non-treated patients. These observations are similar to those reported in a large advanced pan-cancer cohort[54]. Interestingly, we also observed a trend towards increased genomic instability, as measured using the HRD score, in patients with prolonged exposure to genotoxic therapy (Extended Data Fig. 7c, p=0.07). Overall, our analyses indicate that treatment with DNA damaging agents remodels the genomic landscape of advanced cancers and may drive a more mutated and genomically unstable phenotype.

In contrast to global TMB, we did not observe an increase in localized mutation showers, or kataegis[6], with prior genotoxic therapy exposure (p=0.8, linear regression). Tumors with DDR pathway alterations demonstrated a lower proportion of kataegis-associated mutations ($p=2.3x10^{-6}$, Wilcoxon rank sum). In total, 62% of cases had evidence for kataegis, the highest proportion occurring in breast cancer (n=113, 78%) and cholangiocarcinoma (n=11, 79%), which was associated with increased APOBEC3B ($p=5.3x10^{-10}$, Wilcoxon rank sum), as previously reported[55].

Alterations in error-prone polymerases are associated with increased mutation burden in patients exposed to genotoxic therapy

In seeking specific alterations, particularly those affecting DDR genes, that might contribute to elevated TMB in patients treated with genotoxic therapy, we identified mutations in translesion synthesis polymerase genes. Tumors with somatic mutations affecting members of the translesion polymerase ζ (Polζ) complex, including *REV3L* and *POLD3* encoded proteins, which were common in our cohort (Extended Data Fig. 7a), were associated with significantly increased TMB in patients previously exposed to genotoxic agents (Fig. 5b, p=0.0016). The Polζ complex is a low fidelity polymerase involved in error-prone replication, bypass and repair of interstrand crosslinks such as those induced by cisplatin, and has been associated with mutagenesis and chemoresistance[56]. Similarly, mutations in *POLQ*, which encodes another low fidelity polymerase involved in error-prone microhomology-mediated end joining and BER, were only associated with increased TMB in tumors treated with genotoxic therapy (Fig. 5b, p=0.00066). These observations were robust even when accounting for tumor type (POLζ: p=0.026, POLQ: $p<2x10^{-16}$, linear regression). Our results indicate a relationship between DNA damaging therapies and DDR pathways in advanced tumors, and highlight the potential mutagenic effect of genotoxic therapy in specific DDR-altered contexts.

Mutation signatures associated with treatment

Signatures SBS17b, SBS31, and MSBS1-MSBS6, which are late-arising or of unknown etiologies, in addition to ID and DBS signatures were examined for association with prior therapy. SBS31 was elevated in patients exposed to platinum agents (cisplatin $p=3.69x10^{-4}$, carboplatin $p=2.41x10^{-6}$, all platinum $p=3.4x10^{-10}$, Wilcoxon rank sum, Holm-Bonferroni correction), as previously reported[13,29,43,54], and

capecitabine ($p=2.15x10^{-2}$) (Fig. 5c, Extended Data Fig. 7d). Our analysis expanded the number of tumor types associated with SBS31 to include cholangiocarcinoma, sarcoma, and breast, lung and ovarian cancers, with a striking signal in sarcomas, where nearly all cisplatin-treated cases demonstrated increased SBS31 (Extended Data Fig. 7d). Tumor samples from patients treated with platinum-based therapy for longer durations demonstrated increased SBS31 exposure compared to patients with shorter treatment duration (Fig. 5c), and remained significant after accounting for tumor type ($p =1.53x10^{-13}$, linear regression). We also observed an association between prior exposure to platinum agents and DBS5 ($p=5.81x10^{-7}$, Wilcoxon rank sum, Holm-Bonferroni correction). This signature, characterized by CT dinucleotide mutation and associated with platinum exposure[49], was found in several tumor types, expanding previous tumor associations to include colorectal cancers and sarcomas. Similar to SBS31, DBS5-associated mutations were elevated in patients with long exposure to therapy (Fig. 5d, $p=4.85x10^{-4}$, linear regression accounting for tumor type), consistent with the higher TMB we observed in patients that underwent prolonged genotoxic therapy (Fig. 5b).

Given the role for homologous recombination repair in resolving platinum-associated interstrand crosslinks[57], we investigated the relationship between HRD and platinum-associated mutation signatures. HRD tumors were more likely to have experienced longer durations of platinum therapy (HRD vs HR-proficient on platinum for > one year, p=0.033, Chi-squared), which may reflect the clinical benefit of platinum-based therapy for HRD tumors. Interestingly, HRD samples (see Methods) unexposed to platinum therapy had increased SBS31 (Fig. 5e, $p=4x10^{-7}$, Wilcoxon rank sum) and SBS31 was greater in HRD platinum-treated samples compared to platinum-treated HR-proficient samples, even when considering only patients with prior therapy of two months to one year (Fig. 5e, p=0.065, Wilcoxon rank sum, p=0.12 after accounting for tumor type, linear regression), suggesting that inherent DNA repair deficiency can influence therapy-associated mutagenesis. DBS5 was not elevated in HRD samples after accounting for tumor type (Extended Data Fig. 7e, p = 0.87, linear regression) indicating

DBS5 was independent of HRD status. ID6 was observed in breast and ovarian cancers, correlated with SBS3 and was elevated in HRD tumors, consistent with an HRD association (Extended Data Fig. 7f, $p=2.6x10^{-5}$). Samples from patients previously exposed to platinum revealed a slight increase in ID6 after accounting for tumor type, even when HR-proficient (Extended Data Fig. 7f, p=0.053, linear regression). Together, these observations indicate that mutations associated with SBS31 and ID6 are elevated in HRD tumors and that SBS31 in particular is predominant in tumors also exposed to platinum therapy. In contrast, DBS5 appears to be solely associated with platinum therapy, and is consistent with platinum-associated intrastrand, and to a lesser extent interstrand, crosslinks[13].

SBS17b was biased towards late timing (Fig. 4c, $p=4x10^{-17}$, Wilcoxon rank sum), consistent with mutations arising after treatment, and was elevated in cancers exposed to platinum-based therapies (all platinum therapies $p=4.02x10^{-3}$, Wilcoxon rank sum, Holm-Bonferroni correction), DNA synthesis inhibitors including capecitabine, gemcitabine, and 5-FU ($p=1.33x10^{-6}$), and doxorubicin ($p=2.65x10^{-10}$). SBS17b was recently reported to be elevated in metastatic cancer compared to primary lesions and was associated with prior exposure to 5-FU, taxanes, platinum-based chemotherapy, and eribulin[29]. As chemotherapy regimens with combined platinum-based compounds and DNA synthesis inhibitors are common across multiple cancer types (Fig. 1b, 1c and Extended Data Fig. 1), we hypothesized that SBS17b mutations may result from exposure to combined DNA damaging agents. Indeed, SBS17b was not significantly elevated in breast cancer patients treated with either DNA synthesis inhibitors or platinum-based therapies alone, whereas patients treated for long periods of time with both therapies demonstrated increased mutations attributed to SBS17b (Fig. 5f, p=0.0020). Thus, treatment with platinum-based therapy and DNA synthesis inhibitors may both contribute to SBS17b but combined therapy may have an even more significant effect on SBS17b-associated mutations. As signature 17 (COSMIC v2) has been found in a small number of primary cancers including liver[41] and breast[6], our observation that platinum and DNA synthesis inhibitor combination therapy is associated with increased

SBS17b suggests that therapy-associated mutagenesis may share or mimic an underlying etiology in primary cancers characterized by this signature[6,41].

Radiation therapy is another potential source of therapy-associated mutations. Indel signature ID8, identified in breast cancer and sarcoma, was higher in samples from irradiated tumors compared to non-irradiated tumors (Fig. 5g, $p=5x10^{-3}$, linear regression). ID8 is characterized by deletions greater than 5bp and is enriched for microhomology at breakpoints, proposed to arise from non-homologous end joining of radiation-induced double strand breaks[58].

In addition to therapy-directed mutational profiles, we also observed elevated mutation signatures consistent with potential therapeutic resistance. APOBEC-mediated mutagenesis is associated with acquired resistance to tamoxifen[59]. APOBEC-associated SBS2 was correlated with prior tamoxifen exposure, particularly for longer treatment durations, suggesting a potential selection for APOBEC-mediated mutagenesis in response to therapy (Fig. 5h). Estrogen receptor status, as determined by a clinical immunohistochemistry test, was not significantly correlated with SBS2 (Extended Data Fig. 7g, p=0.33), indicating molecular subtype was not responsible for this association. Concomitant elevation of *APOBEC3A* expression was also observed in tumors from patients treated with tamoxifen for a longer duration compared to patients not treated with tamoxifen (Extended Data Fig. 7h, p=0.052), consistent with the hypothesis that intrinsic APOBEC-mediated mutagenesis contributes to acquired resistance to tamoxifen in breast cancer.

Somatic second-hits and mutation signatures in patients with germline alterations

We investigated 98 cancer predisposition genes identified by ACMG/AMP (Supplementary Table 7) and identified 84 pathogenic and likely pathogenic germline variants in 13.5% of cases, spanning 17 cancer

types and 27 cancer predisposition genes (See Methods, Fig. 6a). This frequency is comparable to prevalence estimates of 12.2-17.8% in similar advanced cancer cohorts[9,60]. Although 90% of the variants were SNVs and small indels, we also identified eight large copy number and structural variants (Fig. 6b and Extended Data Fig. 8a), including a complex rearrangement at the *NTHL1* and *TSC2* loci[61].

39% of cases harboring pathogenic or likely pathogenic germline variants had a second event consistent with a classic two-hit model of tumorigenesis in which both copies of a tumor suppressor gene are affected (Fig. 6c). Secondary somatic alterations occurred most often in *BRCA2*, *BRCA1*, *ATM*, and *CHEK2* (Fig. 6c), primarily through loss of heterozygosity (LOH) (Fig. 6d). Many germline variants in tumor suppressor genes were associated with low tumor expression (40% below the 25th percentile across the POG570 cohort), while two of three oncogenic *MITF* variants were associated with high tumor expression (above the 75th percentile) (Extended Data Fig. 8b).

Several somatic mutational processes were associated with the presence of pathogenic and likely pathogenic germline variants. A strong presence of signature SBS30 was observed in one highly mutated case (Extended Data Fig. 8c) with deleterious germline alterations in the *NTHL1* and *TSC2* genes[61]. In addition, SBS18 predominated in one case with bi-allelic germline *MUTYH* variants[62] but not in heterozygous *MUTYH* carriers (Extended Data Fig. 8c). Cases with both germline and somatic defects in *BRCA1/2* had higher SBS3 exposure compared to those with only one or the other (Fig. 6e, $p=1x10^{-5}$). HRD was increased in the presence of either germline or somatic *BRCA1/2* alterations (Fig. 6f, Extended Data Fig. 8c, $p=7x10^{-23}$). Thus, the germline background was relevant to tumor alterations, expression patterns, and mutation signatures.

Immune landscapes of advanced cancers are diverse and predict survival

To profile the immune microenvironment, we determined expression signatures of immune cells (see Methods). Consensus clustering of these signatures identified eight clusters (Fig. 7a), independent of biopsy site, tumor content, and prior treatment. Clusters included those characterized by $CD8^+$ T cells (cluster 3), both B and T cells (cluster 5), and neutrophil signatures (cluster 7). Notably, these clusters were independent of tumor type (Extended Data Fig. 9a), emphasizing the importance of the immune landscape in the understanding of advanced cancers.

Overall survival of patients varied by immune cluster (Fig. 7b, p=0.00011), with cluster 5 exhibiting the highest overall survival, even after accounting for tumor type and tumor content (Extended Data Fig. 9b, HR=0.4, p=0.001). Analysis of immune clusters in TCGA[63] did not reveal a similar lymphocyte-enriched cluster, suggesting that features of this cluster may be enriched in advanced tumors, or at metastatic sites. Presence of B cells in the microenvironment has been associated with improved prognosis in several cancer types[64,65], however we did not observe a difference in survival when stratifying patients by B cell signatures alone. The combination of immune signatures in cluster 5 may indicate the presence of tertiary lymphoid structures, which could contribute to the improved prognosis in these patients[65].

Several clusters exhibiting higher levels of macrophage-associated, monocyte, and neutrophil expression (clusters 1,7,8) were associated with poorer survival. These clusters are similar to the 'lymphocyte-depleted' subgroup in primary tumors, similarly described as having poor long-term survival[63]. Our data thus indicate that the immune composition of metastatic tumors shares some features with primary tumors.

We mined transcriptome data for T cell receptor (TCR) sequences (see Methods) in tumor samples, identifying 10,732 unique TCR β locus (TRB) sequences, and computed dominance and diversity for each

sample's repertoire (see Methods). TCR diversity was positively associated with T cell expression signatures (r = 0.76, $p=2x10^{-16}$, Spearman) and T cell-enriched cluster 3 ($p=2x10^{-5}$, Spearman), presumably because the elevated presence of T cells resulted in increased reads aligning to the TCR region, and was negatively correlated with the presence of a dominant clonotype (Fig. 7c) (r = -0.46, $p=2x10^{-16}$). The most commonly observed CDR3β clonotype sequence, CSARESTSDPKNEQFF, was detected in fourteen samples (3%), and was highly dominant in two (Extended Data Fig. 9c). We did not detect a shared alteration among these samples, suggesting another source of shared tumor-associated antigen may be associated with expansion of this TCR clone.

Immune and genome landscapes interact to predict ICI response

Immune checkpoint inhibitors (ICIs) have become transformative therapies of relevance in tumor types with high mutation loads and MSI. Biomarkers of response in addition to MSI are still being developed[66]. We examined the interaction between tumor genomes and immune microenvironments in predicting response to ICIs. In 76 patients from the POG570 cohort treated with ICIs after biopsy (ICI cohort) (Extended Data Fig. 9d), high exonic TMB was associated with longer treatment duration (Extended Data Fig. 9e, p=0.028) indicative of improved prognosis. Similarly, high T cell signatures were associated with prolonged ICI therapy (Extended Data Fig. 9e, p=0.0094), as was TCR diversity (p=0.008, Log rank test), reflecting the relationship between diversity and T cell signatures. Notably, coding mutation burden and T cell scores were poorly correlated (r=0.07, p=0.103, Pearson), and we hypothesized that combining these markers could more effectively predict response. Indeed, we observed that patients with both high exonic TMB and high T cell signatures had the longest duration of ICI therapy (Fig. 7d, p=0.0055), even after accounting for tumor type (HR=0.18, p=0.037, Cox proportional hazards). This was in contrast to patients with low mutation burden and T cell signatures who had the shortest duration of ICI therapy. The combined predictive value of mutation burden and T cell signatures are consistent with a recent

large study of pembrolizumab in clinical trials[15], demonstrating that these associations also hold in a heavily treated pan-cancer cohort, unselected for PD-L1.

DISCUSSION

We leveraged our uniquely comprehensive pan-cancer whole genome, transcriptome and clinical dataset to explore the impacts of prior therapy on the genomic landscapes of advanced cancer patients. As our capacity to sequence human tumors grows, it will be of increasing importance to also comprehensively capture treatment and clinical data if we are to fully understand the treatment-associated evolution of advanced cancers and the clinical implications thereof.

Advanced tumors showed evidence for variants that may have contributed to therapy resistance, in agreement with another recent advanced pan-cancer analysis[8], and revealed therapy-induced mutagenesis across multiple cancer types. Long-term exposure to therapies such as cisplatin resulted in increased tumor mutation burden (TMB), and signatures SBS31, and DBS5[13,49]. DNA damage response (DDR) deficiency, including homologous recombination deficiency (HRD), was associated with elevated TMB and patients with HRD tumors were more likely to remain on platinum therapy for longer durations[16] suggesting the combined effect of prolonged platinum exposure, in addition to the underlying accumulation of mutations associated with HRD, may particularly elevate TMB in a subset of relapsed patients. The notion of a general role for the interaction of DDR-deficiency and therapy-associated mutagenesis in shaping tumor evolution is supported by observations of a recurrent glioblastoma subtype, characterized by temozolomide-induced hypermutation associated with *MGMT* promoter methylation[67]. Our findings that mutations in low fidelity DNA polymerases were only

associated with increased TMB when present in tumors with prior exposure to genotoxic agents suggest alterations in DDR-proficiency and DNA-damaging treatments can also synergistically enhance mutation burden in advanced, treatment-resistant human tumors.

Evidence in our cohort for increased TMB related to DDR and therapy implies that increased therapy-associated mutagenesis may lead to enhanced sensitivity to immunotherapy. Clinical trials to examine the combination of chemotherapy with immunotherapy are underway[68]. Additionally, immune expression signature clusters have implications for patient survival, independent of tumor type and biopsy site, and combined TMB and T cell expression signatures predict response to ICIs. These findings can be used to inform future trial design and aid therapeutic decision-making in patients.

Our findings demonstrate that specific therapeutic combinations can also induce mutations associated with distinct signatures. SBS17b has recently been proposed to be elevated in response to therapy, but overlap in therapies prohibited elucidation of the causative agent[29]. Dissection of our breast cancer cases to examine specific drug combinations revealed an important role for contribution from both DNA synthesis inhibitors and platinum-based therapies in the elevation of SBS17b. Our results indicate that combination therapy-associated mutagenesis characterized by SBS17b either shares or mimics an underlying etiology found in primary cancers exhibiting this signature[43]. Considering the mechanism of action of these drugs, one hypothesis is that SBS17b results from replication stress that may be inherent to tumors, consistent with its presence in primary untreated samples, and also induced by therapy in advanced cancers. Our evaluation of mutation signatures has provided further insight into the mechanisms that drive mutagenesis in human cancers, including evidence that melanoma-associated SBS38 is not a result of indirect UV damage as previously proposed, and suggests several directions in which longitudinal studies could further examine mechanisms and specific drug-gene interactions.

We identified treatment-associated genomic and gene expression alterations that may play a role in treatment resistance, including in *FGFR1*, *CTNNB1*, *EGFR*, and *VEGFA*. Potential therapy sensitivity associated with reduced *DPYD* expression, combined with recent evidence that somatic *DPYD* loss-of-function can be associated with a prolonged response[18], suggests that somatic mutations or deletions in *DPYD* may be useful biomarkers for selecting patients for treatment with 5-FU. The emergence of resistance alterations, combined with significant changes in mutation burden, highlights the relevance of using advanced or metastatic tumor samples, rather than solely primary samples at diagnosis, for clinical genetic testing and personalized therapy[17].

In addition to contributing to fundamental research insights, the results from this study have informed clinical patient management, including selection of patients for immunotherapy, use of genome signatures for predicting drug sensitivity[16], transcriptome-based changes in diagnosis[69], and identification of drug targets including fusions and overexpressed genes[17–20]. As further large-scale efforts move towards sequencing of pre-treated and metastatic disease[3,8,58], the availability of our rich dataset serves as a foundation for understanding the genomic landscape and treatment impacts in advanced tumors, and brings whole genome and transcriptome sequencing closer to the cancer clinic.

ACKNOWLEDGEMENTS

This work would not be possible without the participation of our patients and families, the POG team, Canada's Michael Smith Genome Sciences Centre technical platforms, the generous support of the BC Cancer Foundation and their donors, and Genome British Columbia (project B20POG). We acknowledge contributions from Genome Canada and Genome BC (projects 202SEQ M.A.M & S.M.J, 212SEQ M.A.M & S.M.J, 12002 GBC M.A.M, S.M.J & J.L), Canada Foundation for Innovation (projects 20070 M.A.M & S.M.J, 30981 M.A.M, S.M.J & J.L, 30198 M.A.M, 33408 M.A.M & S.M.J) including the CGEN platform

(35444 S.M.J) and the BC Knowledge Development Fund. We acknowledge the generous support of the CIHR Foundation Grants program (FDN 143288, M.A.M), University of British Columbia Clinician Investigator Program (M.L.T) and the CIHR Vanier Canada Graduate Scholarship (E.Y.Z). The results published here are in part based upon analyses of data generated by the following projects and obtained from dbGaP (http://www.ncbi.nlm.nih.gov/gap): The Cancer Genome Atlas managed by the NCI and NHGRI (http://cancergenome.nih.gov); Genotype-Tissue Expression (GTEx) Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health (https://commonfund.nih.gov/GTEx).

CONTRIBUTIONS

M.A.M, J.L., M.R.J, Y.S and E.P conceptualized the study. C.R, E.Y.Z, K.L.M, E.C, A.D, M.W, S.K.C, S.Z, S.B, A.M, D.D, R.D.C, D.M, M.C, C.C, D.B, S.Sadeghi, W.Z, T.W., D.C, Y.M and S.D.B contributed to software development and implementation. Analyses were performed by E.T, E.P, L.W, E.Y.Z, H.K, K.F, R.B, K.D, L.C, J.K.G, J.A, K.W, C.J.G, M.L.T, M.R.J, Z.B, H.P, T.V and R.S. Data was collected, and experiments performed by A.J.M, R.A.M, Y.Z, M.R.J, Y.S, M.B, G.A.T, E.M, V.C, K.S, S.Y, D.A.R, D.W and R.A.H. Provision of patient samples and curation of data was conducted by K.G, A.T, S.Sun, H.L, D.J.R, S.C., D.F.S, J.L, M.K.C.L, J.M.L, B.D, A.Fisic, J.N and S.M. The original draft was written by E.P, E.T, L.W, E.Y.Z, H.K, K.D, K.W, M.R.J, Y.S, and M.A.M, J.L and S.J.M.J reviewed and edited the manuscript. Data visualization was conducted by E.T, H.K, R.B, E.Y.Z, L.C, K.D, E.P, K.W, K.F, Z.B and J.K.G. Project management and coordination was performed by J.N, A.Fok and J.M.K. Funding was acquired by M.A.M, J.L and S.J.M.J.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Schwaederle, M. *et al.* Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. *J. Clin. Oncol.* **33**, 3817–3825 (2015).
2. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).
3. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
4. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* **9**, 34 (2017).
5. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
6. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
7. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J Clin* **69**, 7–34 (2019).
8. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
9. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
10. van der Wekken, A. J. *et al.* Overall survival in EGFR mutated non-small-cell lung cancer patients treated with afatinib after EGFR TKI and resistant mechanisms upon disease progression. *PLoS ONE* **12**, e0182885 (2017).

11. Jeselsohn, R., De Angelis, C., Brown, M. & Schiff, R. The Evolving Role of the Estrogen Receptor Mutations in Endocrine Therapy-Resistant Breast Cancer. *Curr Oncol Rep* **19**, 35 (2017).
12. Szikriszt, B. *et al.* A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol.* **17**, 99 (2016).
13. Boot, A. *et al.* In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.* **28**, 654–665 (2018).
14. Murugaesu, N. *et al.* Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov* **5**, 821–831 (2015).
15. Cristescu, R. *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* **362**, (2018).
16. Zhao, E. Y. *et al.* Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. *Clinical Cancer Research* **23**, 7521–7530 (2017).
17. Laskin, J. *et al.* Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. *Cold Spring Harb Mol Case Stud* **1**, a000570 (2015).
18. Majounie, E. *et al.* Fluorouracil sensitivity in a head and neck squamous cell carcinoma with a somatic DPYD structural variant. *Molecular Case Studies* mcs.a004713 (2019) doi:10.1101/mcs.a004713.
19. Jones, M. R. *et al.* NRG1 Gene Fusions Are Recurrent, Clinically Actionable Gene Rearrangements in KRAS Wild-Type Pancreatic Ductal Adenocarcinoma. *Clin. Cancer Res.* **25**, 4674–4681 (2019).
20. Jones, M. R. *et al.* Response to angiotensin blockade with irbesartan in a patient with metastatic colorectal cancer. *Annals of Oncology* **27**, 801–806 (2016).
21. Thibodeau, M. L. *et al.* Whole genome and whole transcriptome genomic profiling of a metastatic eccrine porocarcinoma. *NPJ Precis Oncol* **2**, 8 (2018).

22. Chooback, N. *et al.* Carcinoma ex pleomorphic adenoma: case report and options for systemic therapy. *Curr Oncol* **24**, e251–e254 (2017).
23. Bose, P. *et al.* Integrative genomic analysis of ghost cell odontogenic carcinoma. *Oral Oncol.* **51**, e71-75 (2015).
24. Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* **171**, 950-965.e28 (2017).
25. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
26. Davare, M. A. *et al.* Rare but Recurrent ROS1 Fusions Resulting From Chromosome 6q22 Microdeletions are Targetable Oncogenes in Glioma. *Clin. Cancer Res.* **24**, 6471–6482 (2018).
27. Tozbikian, G. H. & Zynger, D. L. A combination of GATA3 and SOX10 is useful for the diagnosis of metastatic triple-negative breast cancer. *Hum. Pathol.* **85**, 221–227 (2019).
28. Harbhajanka, A. *et al.* Clinicopathological, immunohistochemical and molecular correlation of neural crest transcription factor SOX10 expression in triple-negative breast carcinoma. *Hum. Pathol.* **80**, 163–169 (2018).
29. Angus, L. *et al.* The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nat. Genet.* **51**, 1450–1458 (2019).
30. Drago, J. Z. *et al.* FGFR1 Amplification Mediates Endocrine Resistance but Retains TORC Sensitivity in Metastatic Hormone Receptor-Positive (HR+) Breast Cancer. *Clin. Cancer Res.* **25**, 6443–6451 (2019).
31. Sokol, E. S. *et al.* Loss of function of NF1 is a mechanism of acquired resistance to endocrine therapy in lobular breast cancer. *Ann. Oncol.* **30**, 115–123 (2019).
32. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29**, 948–960 (2015).

33. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).

34. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).

35. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).

36. 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).

37. Wu, S. *et al.* Whole-genome sequencing identifies ADGRG6 enhancer mutations and FRS2 duplications as angiogenesis-related drivers in bladder cancer. *Nat Commun* **10**, 720 (2019).

38. Arasada, R. R. *et al.* Notch3-dependent β-catenin signaling mediates EGFR TKI drug persistence in EGFR mutant NSCLC. *Nat Commun* **9**, 3198 (2018).

39. Howie, L. J. *et al.* FDA Approval Summary: Pertuzumab for Adjuvant Treatment of HER2-Positive Early Breast Cancer. *Clin. Cancer Res.* **25**, 2949–2955 (2019).

40. Matsusaka, S. & Lenz, H.-J. Pharmacogenomics of fluorouracil -based chemotherapy toxicity. *Expert Opin Drug Metab Toxicol* **11**, 811–821 (2015).

41. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun* **8**, 1315 (2017).

42. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med* **7**, 283ra54 (2015).

43. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

44. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).

45. Alexandrov, L. B., Nik-Zainal, S., Siu, H. C., Leung, S. Y. & Stratton, M. R. A mutational signature in gastric cancer suggests therapeutic strategies. *Nat Commun* **6**, 8683 (2015).

46. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

47. Grolleman, J. E. *et al.* Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell* **35**, 256-266.e5 (2019).

48. Viel, A. *et al.* A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **20**, 39–49 (2017).

49. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).

50. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat Commun* **9**, 1744 (2018).

51. Baretti, M. & Le, D. T. DNA mismatch repair in cancer. *Pharmacol. Ther.* **189**, 45–62 (2018).

52. Denver, D. R., Feinberg, S., Steding, C., Durbin, M. D. & Lynch, M. The relative roles of three DNA repair pathways in preventing Caenorhabditis elegans mutation accumulation. *Genetics* **174**, 57–65 (2006).

53. Telli, M. L. *et al.* Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).

54. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).

55. Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* **2**, e00534 (2013).

56. Lee, Y.-S., Gregory, M. T. & Yang, W. Human Pol ζ purified with accessory subunits is active in translesion DNA synthesis and complements Pol η in cisplatin bypass. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2954–2959 (2014).

57. McHugh, P. J., Spanswick, V. J. & Hartley, J. A. Repair of DNA interstrand crosslinks: molecular mechanisms and clinical relevance. *Lancet Oncol.* **2**, 483–490 (2001).

58. Behjati, S. *et al.* Mutational signatures of ionizing radiation in second malignancies. *Nat Commun* **7**, 12605 (2016).

59. Law, E. K. *et al.* The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer. *Sci Adv* **2**, e1601737 (2016).

60. Schrader, K. A. *et al.* Germline Variants in Targeted Tumor Sequencing Using Matched Normal DNA. *JAMA Oncol* **2**, 104–111 (2016).

61. Wong, H.-L. *et al.* Molecular characterization of metastatic pancreatic neuroendocrine tumors (PNETs) using whole-genome and transcriptome sequencing. *Cold Spring Harb Mol Case Stud* **4**, (2018).

62. Thibodeau, M. L. *et al.* Base excision repair deficiency signatures implicate germline and somatic MUTYH aberrations in pancreatic ductal adenocarcinoma and breast cancer oncogenesis. *Cold Spring Harb Mol Case Stud* **5**, (2019).

63. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

64. Dieu-Nosjean, M.-C. *et al.* Long-term survival for patients with non-small-cell lung cancer with intratumoral lymphoid structures. *J. Clin. Oncol.* **26**, 4410–4417 (2008).

65. Petitprez, F. *et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **577**, 556–560 (2020).

66. Topalian, S. L., Taube, J. M., Anders, R. A. & Pardoll, D. M. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nature Reviews Cancer* **16**, 275–287 (2016).

67. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).
68. Zhou, C. & Zhang, J. Immunotherapy-based combination strategies for treatment of gastrointestinal cancers: current status and future prospects. *Front Med* **13**, 12–23 (2019).
69. Grewal, J. K. *et al.* Detection and genomic characterization of a mammary-like adenocarcinoma. *Cold Spring Harb Mol Case Stud* **3**, (2017).

FIGURE LEGENDS

Fig. 1: **POG570 cohort description**.

a, Disease types and associated biopsy sites. Width of joining lines reflects the number of patients with the corresponding tumor type and biopsy site. Tumor-biopsy pairings are only shown if there are three or more patients in the pairing. *n* = number of samples of that tumor type or biopsy site. **b**, Frequency of drug usage for the cohort (*n*=570 patients) by tumor type (top), proportion of patients in each tumor type treated with each therapy (indicated by color on heatmap), number of patients given each therapy (right), and distribution of number of days between first and last dose of treatment (far right). The 20 most frequently administered cancer drugs are displayed. Boxplots represent the median, upper and lower quartiles of the distribution, and whiskers represent the limits of the distribution (1.5 * interquartile range). The central line on the violin plots in **b** represent the median and the tips extend to the minimum and maximum values of the distribution. **c**, Drug co-occurrences by patient in breast (BRCA, top) and colorectal (COLO, bottom) cancer patients; darker circles indicate drugs used more frequently, and darker lines show drugs frequently used in combination. Disease cohorts: BRCA, breast cancer; COLO, colorectal cancer; LUNG, lung cancer; SARC, sarcoma; PANC, pancreatic cancer; OV, ovarian cancer; CNS-PNS, nervous system tumors; CHOL, cholangiocarcinoma; SKCM, melanoma; SECR, tumors of secretory organs; LYMP, blood and lymphoid cancers; STAD, stomach cancer; UCEC, uterine

cancers. Biopsy sites: Liver, liver; Lymph, lymph node or blood samples; Resp, respiratory system; ChWall, chest wall; SoftTis, soft tissue; Abdom, abdominal; Breast, breast; Repro, reproductive system; Brain, brain; H&N, head and neck.

Fig. 2: Tumor genomic landscape and frequent alterations.

a, Genomic alterations across tumor types. Somatic total mutation load is shown on a log scale of small mutations per genome. The mutation substitutions show the proportion of specific base changes in SNV and small indel events. Alterations in the 25 most frequently altered oncogenes (red, left) and tumor suppressors (blue, left) are shown as defined by OncoKB (see Methods); only one representative gene is shown when frequent co-amplification occurs. Bar plots on the right-hand side display the frequency of hits in that gene across the whole cohort. HRD, homologous recombination deficiency[16]; MSI, microsatellite instability as calculated by MSIsensor (see Methods); HHV, human herpesvirus; HPV, human papillomavirus. **b**, Recurrent mutation clusters; those seen in at least five patients are shown (*n*=2596 clusters). The significance of each cluster was calculated using a binomial distribution as described in the Methods, and was multiple test corrected using the false discovery method. **c**, Association of non-coding clusters in regulatory regions of *TERT* and *AP2A1* with gene expression (log TPM). P values shown are calculated by a two-sided Wilcoxon rank sum with multiple test correction using the false discovery rate method. Boxplots in **c** represent the median, upper and lower quartiles of the distribution, and whiskers represent the limits of the distribution (1.5 * interquartile range). **d**, Tumor heterogeneity, measured as the Shannon diversity index of subpopulations determined by EXPANDS (see Methods), and median number of somatic SNVs (Spearman correlation). **e**, Kaplan-Meier survival from advanced disease diagnosis across the cohort split on the median genomic tumor mutation burden (TMB, 6237 mutations). The P value was determined by comparison of samples above (*n*=285 patients) and below (*n*=285 patients) the median using a two-sided log rank test.

Fig. 3: Treatment-associated recurrent alterations.

a, Drug-mutation associations, where the drug association score is the -log(P value) of the co-occurrence of a drug with a mutation, 'clustered' refers to more than one mutation within 9 bp which may indicate activating mutations, and 'truncated' refers to mutations that are predicted to result in a truncated protein and loss of function. No non-coding alterations showed association with therapy. **b**, Drug-copy change associations, where the drug association score is the -log(P value) of the co-occurrence of a drug with a copy change, and expression is compared between amplified and non-amplified samples for the most significant gene in the genomic region; genes are grouped by chromosomal band. P values for drug-alteration associations shown in **a-b** were calculated using Chi square statistics with multiple test correction (see Methods). **c**, Time on treatment for patients with mutations in *ESR1* (BRCA samples) and *EGFR* (LUNG samples). **d**, Expression of drug targets (TPM) and time on treatment for *ESR1* (fulvestrant and tamoxifen, in BRCA samples), *VEGFA* (bevacizumab, in COLO samples) and *DPYD* (5-FU, in COLO samples). P values in **c** and **d** were calculated by two-sided Wilcoxon rank sum tests. Drugs are grouped as in Supplementary Table 2. Boxplots in **c** and **d** represent the median, upper and lower quartiles of the distribution, and whiskers represent the limits of the distribution (1.5 * interquartile range). Sample sizes used for statistical tests in **a** and **b** are as described in Fig.1a for individual tumor types and are as follows: BRCA, *n*=144; COLO, *n*=87; LUNG, *n*=67; PANC, *n*=42; OV, *n*=28; n=number of patients.

Fig. 4: Novel mutation signatures are identified in metastatic tumors.

a, *De novo* mutation signatures (SBS, ID and DBS) deciphered from 482 POG570 advanced cancer samples reflect COSMIC signatures[43] and additional novel signatures with no COSMIC match. **b**, Representative network of selected pairwise Spearman correlations between exposures of signatures, with proposed etiology. Only samples with both signatures detected (as demonstrated in **a**) were included in calculation of each paired correlation (edge). Correlations displayed have a minimum

correlation of 0.3. Negative correlations are indicated by a dashed line. **c**, Mean timing of mutations associated with each signature by tumor type, for samples with at least 10% of mutations that could be timed. SBS, single base signature; ID, indel, insertion and deletion signature; DBS, double base signature. Sample sizes for each edge in **b** are: DBS2-ID3, *n*=26; ID3-SBS4, *n*=64; DBS2-SBS4, *n*=26; SBS13-SBS2, *n*=144; MSBS6-SBS7a, *n*=12; SBS7a-SBS38, *n*=13; SBS30-SBS36, *n*=42; SBS36-MSBS3, *n*=42; MSBS3-SBS30, *n*=42; SBS17b-SBS31, *n*=144; SBS31-SBS3, *n*=233; SBS31-DBS5, *n*=127; DBS5-SBS3, *n*=123; SBS3-ID6, *n*=171; ID6-SBS17b, *n*=143; SBS3-SBS17b, *n*=231; n=number of patients with both signatures.

Fig. 5: Prior therapy shapes the tumor genomic landscape.

a, Tumor mutation burden (TMB) in tumors with somatic mutations in genes in DNA repair pathway mutations. The P value was calculated by an Anova test. **b**, TMB and duration of prior treatment with genotoxic agents in tumors with no DNA repair mutations and in tumors with somatic mutations in the translesion polymerases, *POLQ* and genes encoding Polζ (including *REV3L* and *POLD3*). **c**-**d**, Exposure to signatures SBS31 (**c**) and DBS5 (**d**) in tumors with prior platinum therapy, median days = 111 and 114 respectively. **e**, Exposure of signature SBS31 and prior therapy (treated for two months to one year) and HRD (homologous recombination deficiency) status. Samples were defined as HR deficient if they had a somatic or germline variant in an HR gene (Supplementary Table 7) and exhibited an HRD score > 35, corresponding to the 70^{th} percentile of this cohort. **f**, Exposure of signature SBS17b and status of prior therapy with platinum agents (cis-, carbo- or oxaliplatin) and DNA-synthesis inhibitors (cape-, gemcitabine or 5-FU) in BRCA tumors. Long and short treatments are split on the median time of platinum agents (71 days). **g**, Exposure of signature ID8 and prior radiation status for all samples. **h**, Exposure of SBS2 and prior tamoxifen therapy in BRCA tumors. P values in **b**-**h** are calculated by two-sided Wilcoxon rank sum tests. Boxplots in **a**-**h** represent the median, upper and lower quartiles of the distribution, and whiskers represent the limits of the distribution (1.5 * interquartile range).

Fig. 6: Germline alterations and effects on the genomic landscape.

a, Proportion of cases within each tumor type with high- and moderate-penetrance pathogenic germline variants in 98 cancer predisposition genes (Supplementary Table 7). **b**, Cancer predisposition genes (*n*=27) with pathogenic germline variants detected in this cohort, including small mutations, copy number variants (CNVs), and structural variants (SVs). **c**, Number of carriers and second hits for the 27 genes with pathogenic germline variants detected in this cohort, including bi-allelic germline variants, loss of heterozygosity (LOH, both deletion and copy neutral), somatic mutations, and low tumor expression (see Methods). Larger dots indicate more genes at the same position. **d**, Proportions and types of germline variants (inner circle) and second hits (outer ring) in the four cancer predisposition genes from **c** most frequently altered with second hits. **e-f** Exposure of SBS3 (**e**) and HRD score (**f**) in cases based on BRCA1/2 germline and somatic mutation status. The central line on the violin plots in **e-f** represent the median, the top and bottom of the coloured box represent the upper and lower quartiles and the tips extend to the minimum and maximum values of the distribution. P values for **e-f** were calculated using Dunn's test on Kruskal-Wallis multiple comparison. P values were adjusted using the Benjamini-Hochberg method.

Fig. 7: Immune landscapes of metastatic cancers.

a, Clustering of samples by composition of immune cell expression signatures (see Methods). **b**, Overall patient survival (*n*=568) based on immune clusters described in **a**. *n* values for each cluster are defined in the table below, and the P value was determined using a two-sided log-rank test. **c**, Relationship between T cell receptor diversity and dominance (see Methods, *r* and P determined using a Spearman correlation) in 372 non-lymphoid samples with at least 20 reads aligned to the TRB region; inset circles

indicate examples of V/J gene usage in samples with high (1), moderate (2), and low (3) dominance. The shaded region around the trend line represents the standard error, with a confidence interval of 0.95. **d**, Probability of continued therapy for patients receiving immune checkpoint inhibitors after the biopsy after exclusion of lymphoid-related tumors (*n*=57, see Methods), stratified by T cell signature and exonic TMB. Kaplan-Meier statistical significance was calculated using a two-sided log rank test.

METHODS

Ethical oversight, consent and enrollment

This work was approved by and conducted under the University of British Columbia – BC Cancer Research Ethics Board (H12-00137, H14-00681), and approved by the institutional review board. The POG program is registered under clinical trial number NCT02155621. Patients residing in the province of British Columbia were referred to the POG program by their treating oncologist and were selected for functional status, available treatment options, and ability to undergo biopsy procedures. Selected patients were approached for study participation by a POG trained oncologist or study nurse.

The assembled cohort was comprised of 878 adult patients who gave informed consent and enrolled in the POG trial between July 2012 and August 2017. For enrollment and sample exclusion criteria, see the Supplementary Note. Complete high quality comprehensive clinical tumor profiles were generated for 570 patients, who were included in the POG570 cohort (Details in Supplementary Table 1). The patients were 359 (63%) female, 211 male (37%), with a median age of 59 years (range: 19–86).

Clinical data collection and processing

Treatments related to patients' cancer diagnosis were systematically abstracted from the BC Cancer Pharmacy database. This database captures all approved cancer therapies administered in regional cancer centers, community hospitals, or taken at home in BC and therefore captures the vast majority of

treatments delivered. Additional treatments were identified using patients' charts where available. Information collected included the type of systemic therapy delivered and the duration of treatment, including all lines of therapies. All analyses related to therapy exposure were performed only on therapies received prior to biopsy, unless specifically stated otherwise. Date of advanced disease diagnosis was defined as the date of incurable, advanced or metastatic disease determined by radiology or by overt clinical finding, whichever was earlier, if progression was documented with subsequent imaging. Kaplan-Meier survival analysis was performed from the date of advanced disease diagnosis to the date of death or censoring as of January 2019 using the R packages survival (v2.42.3) and survminer (v0.4.2). Differences in non-parametric survival functions were assessed across subgroups using log-rank tests. Survival analysis was further reviewed using cox proportional hazards models to ensure there were no competing risk factors from observable patient characteristics, including tumor type, which are shown in extended data figures where relevant.

Tissue collection and library construction

Tumor specimens were collected from biopsies or resections, pathology reviewed, and nucleic acids extracted as described in the Supplementary Note. Constitutional DNA representing normal cells was extracted from peripheral blood. PCR-free DNA libraries and either strand-specific or ribodepleted RNA libraries were constructed as described in the Supplementary Note.

Whole genome and transcriptome sequencing

Tumor genomes were sequenced to a target depth of 80X coverage and normal peripheral blood samples to 40X coverage (see Supplementary Table 8 for coverage by sample) on Illumina (San Diego, California) HiSeq 2500 using V3 or V4 chemistry and paired-end 125 base reads, or on HiSeqX using v2.5 chemistry and paired-end 150 base reads. Transcriptomes were sequenced targeting 150-200 million 75-base paired end reads on Illumina HiSeq2500, or on NextSeq500 using v2 chemistry.

Discovery of somatic alterations

Sequence reads from normal and tumor whole genome libraries were analyzed to identify somatic single nucleotide variants (SNVs), insertions and deletions (indels), copy number variants, loss of heterozygosity, and structural variants (SVs), as described in the Supplementary Note. Structural variants in either RNA or DNA considered to represent putative fusion genes were defined as events with breakpoints in two different protein-coding genes which were predicted to affect the sequence of the expected protein product[70]. Events were defined as having evidence in both DNA and RNA if the same gene pair was predicted in both. Tumor suppressors and oncogenes were defined using OncoKB annotations[71]. Total genomic tumor mutation burden (TMB) was the total number of SNVs and indels per sample; exonic TMB was the total number of SNVs and indels annotated as overlapping or affecting protein coding regions.

Gene expression profiling

RNA-Seq reads were aligned using STAR[72] (v2.5.2b) and expression was quantified using RSEM[73] (v1.3.0) as transcripts per million (TPM) to minimize computational batch effects between POG570 RNA-seq samples and over 20,000 recomputed and publicly available RNA-seq samples in Xena Public Data Hubs (https://xena.ucsc.edu/public-hubs/). All required input indexed files for STAR and RSEM were generated from the hg38 reference genome (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/), and gene annotations were based on Ensembl version 85[74].

Microsatellite instability, homologous recombination deficiency (HRD), and microbial detection

MSI scores determined using MSIsensor[75] (v0.2) from genome alignments were computed as the percentage of total sites displaying MSI. HRD scores were computed using the R package HRDtools[16]

(v0.0.0.9) as the arithmetic sum of loss of heterozygosity (LOH), telomeric-allelic imbalance (TAI), and large-scale state transitions (LST) scores, determined based on published guidelines[76]. Microbial detection was performed using BioBloomTools[77] (v2.0.11b; https://github.com/bcgsc/biobloom), which compared tumor and normal sequences to reference viral, bacterial and fungal sequences. Candidate positive microbial matches were evaluated by manual review.

Significantly mutated genes and primary tumor comparison

Significantly mutated genes were identified using MutSig2CV[78] v3.11, using default parameters and a genome-wide coverage file, based on variants annotated using vcf2maf v1.6.6 (https://github.com/mskcc/vcf2maf) with Ensembl version 83. Significantly mutated genes were computed for (1) the entire POG570 cohort and (2) each of the six largest tumor types (BRCA, COLO, LUNG, SARC, PANC, OV). Results were filtered to identify genes with q<=0.1. Significant genes were also computed for the entire cohort excluding hypermutated cases with >10 mutations/Mb, to provide insight into genes which appeared significant due to frequent mutation in hypermutated cases, including genes with multiple polynucleotide repeats.

PanCancer Analysis of Whole Genomes (PCAWG) consensus SNV, MNV and indel calls were obtained through https://dcc.icgc.org/pcawg and annotated with the same approach. For each significantly mutated gene in POG570, we queried whether that gene was mutated more frequently in POG570 than in primary tumors from PCAWG by: (1) grouping PCAWG and POG570 by tumor subtype, considering those subtypes with at least 10 samples in each cohort and (2) propensity matching based on tumor type, age and gender using R packages MatchIt v3.0.2 and cobalt v3.9.0 to create matched PCAWG and POG570 cohorts. All pediatric, recurrent and metastatic samples from PCAWG were excluded in this analysis.

Mutation positional clustering

All SNVs in the POG570 cohort residing within 50 bp of one another were grouped into clusters using the R package *ClusteredMutations* (v1.0.1, https://cran.r-project.org/web/packages/ClusteredMutations/), with subsequent filtering to remove potential germline artefacts and identification of events in >5 patients as described in the Supplementary Note. The significance of each cluster was calculated using a binomial distribution as previously described[36], with multiple test correction using the false discovery rate method[79]. Positional clusters were verified using an alternative sliding window approach. Intermutational distances were calculated for all mutations in each patient and kataegis events were identified using the definition generated through a computational model by Alexandrov *et al*.[80], as described in the Supplementary Note. Kataegis mutational burden for each patient was calculated as:

$$kataegis\ mutation\ burden = \frac{number\ of\ mutations\ in\ kataegis\ events}{total\ number\ of\ mutations}$$

Gene promoter regions were defined as 1500 bp upstream to 500 bp downstream of all transcription start sites using Ensembl (v69) gene models. Enhancers were defined using GeneHancer[81], only considering "double elite" enhancers. 5'UTR and 3'UTR regions were defined by Ensembl (v69). microRNA (miRNA) binding sites were defined as the conserved miRNA families binding site predictions available in the TargetScan database[82] (v7.2). For clusters in regulatory regions of genes with expression data, expression p-values were calculated using an unpaired two-samples Wilcoxon test on TPM values, with multiple test correction using the false discovery rate method.

Tumor subpopulation analysis

SNVs and copy number alterations were used to predict the presence of sub-populations in each sample with EXPANDS[83] (v2.1.1) (Extended Data Figs. 2d-e), as described in the Supplementary Note. The

Shannon index was calculated as a measure of tumor genome heterogeneity using the R vegan package (v2.5.3, https://cran.r-project.org/web/packages/vegan/index.html).

As an alternative method of evaluating heterogeneity, the cancer cell fraction (CCF) of SNVs (Extended Data Fig. 2 c and f) was calculated for samples with at least 2000 somatic SNVs as follows:

$$corr = \frac{cn\ x\ tc}{cn\ x\ tc + \left(2\ x\ (1 - tc)\right)}$$

$$ccf = \frac{alt}{(alt + ref)\ x\ corr}$$

For analysis, subclonal mutations were defined as those with a CCF of 0.2 or less, and samples were defined as having a high proportion of subclonal mutations if 20% or more of all mutations were subclonal.

To examine the subpopulation frequency of driver gene mutations, high or moderate impact variants (SNPeff annotation) in genes in The Cancer Genome Atlas (TCGA) driver gene list[2] which were mutated in at least 3 cases within a POG570 tumor type were considered.

Gene alterations associated with therapy

To identify small mutations (< 20 bp) and copy number changes with increased prevalence in patients that had received treatments, we used an occurrence measurement and filtered by increased prevalence of gene mutations compared to untreated TCGA primary tumors from the TCGA PanCan cohort[2] (https://gdc.cancer.gov/about-data/publications/pancanatlas, version v0.2.8.PUBLIC) as described in the Supplementary Note (see Supplementary Table 9 for barcodes). Associations between genetic alterations and treatments were determined using chi square statistics with multiple test correction.

For mutations, genes with treatment associations and a fold change greater than 1 were then selected for those with mutational hotspots (multiple mutations within 9 bp) providing evidence for gain of function, or a minimum of two truncating mutations providing evidence for loss of function. For copy changes, genes with associations and a fold change greater than 1 were then examined for differential expression. Differential expression was calculated between groups of patients with or without the copy number alteration, overall and specifically within the treated group, using an unpaired two-tailed Wilcoxon test with multiple test correction using the false discovery rate method. Only genes with a difference in means in the anticipated direction (positive for amplifications, negative for deletions) were considered in the analysis. Genes were then grouped by chromosomal band. Regions producing non-coding transcripts were defined using Ensembl 69 biotype[84]. The most frequent non-coding alterations were also included in analysis of treatment associations, with no significant findings.

Gene expression clustering and association with therapy

t-SNE (t- Distributed Stochastic Neighbour Embedding) decomposition plots were generated using TPM values and default parameters from the scikit-learn[85] package in Python, with min-max scaling per sample (rescaling to a range of 0-1). For each pair of cancer types in TCGA, pairwise-ANOVA was used to determine the highest discriminating genes between the two cancer types ($p<0.05$); the aggregated set of 1559 genes were used as input to the decomposition.

We assessed the impact of drugs on gene expression for 45 literature-informed combinations of drug, gene and disease. We considered three groups of patients for each cancer type: (1) those who never received the drug, (2) those that were on treatment for less than 90 days, and (3) those who were on treatment for more than 90 days prior to biopsy. Patients were excluded if days on treatment was <14

days, or if PLACEBO was included in the drug name. If a patient was given multiple doses of the same drug, the most recent duration of therapy was used.

Genomic alterations with DNA repair and genotoxic therapy

To investigate the relationship between somatic DNA repair defects and mutation burden we examined mutations in 181 genes across 12 DNA repair pathways distilled from the literature (Supplementary Table 7), p-values were calculated using an ANOVA method and also a resampling approach as described in the Supplementary Note. We excluded samples exhibiting a hypermutated phenotype, defined as $\geq$ 10 mutations per Mb across the genome. An additional case was excluded as no coding mutations were detected.

To investigate the effects of genotoxic treatments on genomic landscapes, drugs were grouped as genotoxic if they belonged to any of the following drug classes (Supplementary Table 2): anthracyclines, DNA alkylating, DNA synthesis inhibitor, topoisomerase I inhibitor, topoisomerase II inhibitor. Comparisons between treatment groups used a Wilcoxon rank-sum test. Significance was computed both including and excluding samples with *TP53* mutations, which made up half of the DNA repair mutations. To control for bias from covariates for this analysis, we performed linear regression models using tumor type and other DNA repair mutation for the POLQ/POL zeta groups.

Mutation signatures

Somatic mutation signature analysis was performed from 6,181,180 somatic single base substitutions (SBS), 974,629 indels, and 54,042 double base substitutions (DBS) using a published framework[80] for non-negative matrix factorization (NMF) of the mutation catalog matrix into *de novo* mutation signatures and the relative exposure of each signature for each cancer genome. Fractional exposure was

defined as the proportion of a genome's total mutation burden contributed by a particular signature. Signature classes, stability estimates, and cohorts analyzed are as described in the Supplementary Note.

A total of 482 patients remained in the 12 completed SBS cohorts, 415 in the six indel cohorts, and 373 in five double base substitution cohorts. Signatures were compared against their respective COSMIC reference signature (version 3, May 2019, https://cancer.sanger.ac.uk/cosmic/signatures/) using the cosine similarity metric. Signatures with cosine similarity to a COSMIC signature greater than 0.6 were considered for matching, and verified through manual inspection of the similarity matrix and the signatures themselves. Manual review was focused on assessment of cosine similarity with special attention paid to subclass-specific trends in the signature analysis. Signature exposures of less than 0.01 were excluded from downstream analyses.

Temporal analysis of SBS mutation signatures based on mutation types and NGS variant allele counts was performed using SignIT (https://github.com/eyzhao/SignIT), as described in the Supplementary Note.

Within the cohort, 106 cases (22%) best fit a model with multiple temporally distinct subpopulations thus enabling signature timing. Mean early and late mutation signature exposures were computed by fitting a weighted linear model of exposure fraction versus subpopulation prevalence.

Analysis of drug-signature associations

Among the 20 most commonly used chemotherapy agents, 7 with known DNA damaging qualities were chosen for investigation: cyclophosphamide, doxorubicin, fluorouracil, cisplatin, capecitabine, carboplatin, and oxaliplatin. Late-arising mutation signatures, signatures of unknown etiology and all

novel signatures (SBS5, SBS17, SBS31, MSBS1-MSBS6) were each assessed for differences in exposure between therapy-exposed and non-exposed patients by the Wilcoxon signed-rank test. Resulting p-values were adjusted for multiple hypothesis testing using the Bonferroni-Holm method. Median days on treatment for platinum associated signatures are as follows: SBS31, 111 days; DBS5, 114 days; ID6, 117 days; SBS17b, 71 days.

Germline mutation analysis

Germline variants, including SNVs, small indels, CNVs, and SVs in normal blood genomes were identified as described in the Supplementary Note. All coding and splice site germline variants in 98 cancer predisposition genes (Supplementary Table 7) were classified according to the American College of Medical Genetics (ACMG) 2015 guidelines[86] using InterVar (Li & Wang, 2017) for partially automated classification followed by manual review. Structural and copy number variants were validated by Multiplex Ligation-dependent Probe Amplification, Sanger sequencing or long-range PCR where possible.

Tumor genome and transcriptome data for the 98 cancer predisposition genes (Supplementary Table 7) was reviewed to identify potential somatic second hits, which included LOH due to deletion of the wild-type allele or allele-specific imbalance, non-synonymous somatic SNVs or indels and low mRNA expression. Due to the limited size of some tumor type-specific cohorts, expression percentiles were calculated for each gene with respect to the entire POG570 cohort using the R package dplyr v.0.8.1, and low expression was defined as values below the 2.5th percentile.

Immune cell deconvolution and repertoire analysis

Gene expression from RNA-Seq was deconstructed using the CIBERSORT[87] R package (v1.04), as described in the Supplementary Note. In general, samples with higher tumor content had lower

predicted immune scores (Extended Data Fig. 9a).

To identify tumor clusters based on immune cell content, we ran ConsensusClusterPlus (v1.44.0) on the CIBERSORT scores for the 568 cases that had evidence of immune content using all 22 leukocyte cell types. The optimal number of clusters was assessed based on 80% cell type and tumor resampling over 1000 iterations of PAM clustering for k2-k8 using Pearson distance metric for clustering and Ward's method for linkage. To plot the heatmap and annotation tracks, we used the ComplexHeatmap (v.1.18.1) R package. The total CIBERSORT score is the sum of all the cell types in each patient.

The Kaplan-Meier survival analyses shown for the immune clusters were censored at 10 years, at which time 18 patients remained alive.

After exclusion of T cell lymphomas and samples obtained from lymph node, bone marrow and peripheral blood biopsies (Supplementary Table 1), a total of 459 samples were included in the T cell receptor (TCR) repertoire analysis using MiXCR[88] (v2.1.2) (see Supplementary Note).For each sample, dominance (presence of a single very common TRB sequence), and diversity (Shannon diversity index, number and proportion of unique TRB sequences) of the TCR repertoire were calculated using the formulae below.

$$Dominance = \frac{Number\ of\ reads\ supporting\ most\ abundant\ clonotype}{Number\ of\ reads\ supporting\ all\ TRB\ clonotypes}$$

$$Shannon\ Diversity\ Index = -\sum_{i=1}^{R} p_i \ln p_i$$

Where p_i is the proportion of reads supporting the *i*th most abundant clonotype.

Evaluation of treatment and survival in immune checkpoint inhibitor treated cohort

A sub-cohort of 76 patients treated with checkpoint inhibitors (PD-1, PD-L1, CTLA-4, OX40, NKG2A and combinations of these with other therapies, including chemotherapy, IDO-1, LAG-3, and HER2 inhibitors) following a POG biopsy was used for Kaplan-Meier survival analyses of the time to treatment failure. Drug information post-biopsy for these patients was collected by chart review, and data was censored at April 2019. High mutation burden was defined as an exonic mutation burden (non-synonymous SNVs and indels) of 10 mutations per Mb or more. High T cell infiltration for this analysis was defined as higher than the 80th percentile of total T cell scores (sum of all CIBERSORT T cell scores excluding regulatory T cells) of all POG570 samples, excluding LYMP and THYM tumors and biopsies taken from lymphatic sites as these tumors are inherently populated with lymphoid cells. Kaplan-Meier analyses using T cell scores in this cohort also excluded these lymphoid-related tumors, resulting in a cohort of 57 patients.

Statistics and reproducibility

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to groups during analyses. Tumor types with less than 10 samples were collected into the “other” disease group to not present any misleading data. Any sample exclusions for analyses are reported in the relevant section of the Methods or Supplementary Note. Unless otherwise stated all statistical tests were performed in R (https://cran.r-project.org/) and p values stated reflect two-sided tests.

DATA AVAILABILITY

Genomic and transcriptomic sequence datasets including metadata with library construction and sequencing approaches have been deposited at the European Genome–phenome Archive (EGA, http://www.ebi.ac.uk/ega/) as part of the study EGAS00001001159 with accession numbers as listed in

Supplementary Table 1. Data on mutations, copy changes and expression from tumor samples in the POG program organized by OncoTree classification (http://oncotree.mskcc.org) are also accessible from https://www.personalizedoncogenomics.org/cbioportal/. Complete small mutation catalog is available for download from http://bcgsc.ca/downloads/POG570/. Previously published TCGA and PCAWG data that were re-analysed here are available from data portals (https://portal.gdc.cancer.gov/ and https://dcc.icgc.org/) with sample barcodes as listed in Supplementary Table 9. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

CODE AVAILABILITY

The bioinformatics analyses were performed using open-source software, including Burrows-Wheeler alignment tool (v0.5.7 for up to 125bp reads and v0.7.6a for 150bp reads), CNAseq[89] (v.0.0.6), APOLLOH[90] (v0.1.1), SAMtools[91] (v0.1.17), MutationSeq[92] (v1.0.2 and v4.3.5), Strelka[93] (v1.0.6), SNPEff[94] (v3.2 for somatic and v4.1 for germline), ABySS[95] (v1.3.4), TransABySS[95,96] (v1.4.10), Chimerascan[97] (v0.4.5), DeFuse[98] (v0.6.2), Manta[99] (v1.0.0), Delly[100] (v0.7.3), MAVIS[70] (v2.1.1), STAR[72] (v2.5.2b), RSEM[73] (v1.3.0), MSIsensor[75] (v0.2), HRDtools[16] (v0.0.0.9), BioBloomTools[77] (v2.0.11b), EXPANDS[83] (v2.1.1), SignIT (https://github.com/eyzhao/SignIT), samtools[91] (v0.1.17), ClinVar[101] (v20180905), InterVar[102], ControlFREEC[103] (v5), CIBERSORT[87] (v1.04), Jaguar[104] (v2.0.3), MiXCR[88] (java, v2.1.2), and VDJtools[105] (v1.1.9). Additional packages used for meta-analyses include R packages ClusteredMutations (v1.0.1), vegan (v2.5.3), ConsensusClusterPlus (v1.44.0), ComplexHeatmap (v1.18.1), survival (v2.42.3), survminer (v0.4.2), and Python package scikit-learn[85] (python, v0.20). Additional processing involved in-house scripts that are available upon request.

METHODS-ONLY REFERENCES

70. Reisle, C. *et al.* MAVIS: merging, annotation, validation, and illustration of structural variants. *Bioinformatics* **35**, 515–517 (2019).
71. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).
72. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
73. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
74. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754–D761 (2018).
75. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
76. Timms, K. M. *et al.* Association of BRCA1/2defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res* **16**, (2014).
77. Chu, J. *et al.* BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics* **30**, 3402–3404 (2014).
78. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer genes. *Nature* **499**, 214–218 (2013).
79. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).
80. 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).

81. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, (2017).

82. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, (2015).

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

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

85. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* **12**, 2825–2830 (2011).

86. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

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

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

89. Jones, S. J. *et al.* Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biology* **11**, R82 (2010).

90. 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 Research* **22**, 1995–2007 (2012).

91. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

92. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).

93. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

94. 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).

95. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Research* **19**, 1117–1123 (2009).

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

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

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

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

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

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

102. Li, Q. & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *The American Journal of Human Genetics* **100**, 267–280 (2017).

103. 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).

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

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