nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0165-1

Genome doubling shapes the evolution and prognosis of advanced cancers

Craig M. Bielski[1], Ahmet Zehir[2], Alexander V. Penson[3,4], Mark T. A. Donoghue[1], Walid Chatila[3], Joshua Armenia[3], Matthew T. Chang[3,4,7], Alison M. Schram[5], Philip Jonsson[3,4], Chaitanya Bandlamudi[1], Pedram Razavi[5], Gopa Iyer[5], Mark E. Robson[5], Zsofia K. Stadler[5], Nikolaus Schultz[1,3], Jose Baselga[4,5], David B. Solit[1,4,5,6], David M. Hyman[5,6], Michael F. Berger[1,2] and Barry S. Taylor[1,3,4*]

Ploidy abnormalities are a hallmark of cancer, but their impact on the evolution and outcomes of cancers is unknown. Here, we identified whole-genome doubling (WGD) in the tumors of nearly 30% of 9,692 prospectively sequenced advanced cancer patients. WGD varied by tumor lineage and molecular subtype, and arose early in carcinogenesis after an antecedent transforming driver mutation. While associated with *TP53* mutations, 46% of all WGD arose in *TP53*-wild-type tumors and in such cases was associated with an E2F-mediated G1 arrest defect, although neither aberration was obligate in WGD tumors. The variability of WGD across cancer types can be explained in part by cancer cell proliferation rates. WGD predicted for increased morbidity across cancer types, including *KRAS*-mutant colorectal cancers and estrogen receptor-positive breast cancers, independently of established clinical prognostic factors. We conclude that WGD is highly common in cancer and is a macro-evolutionary event associated with poor prognosis across cancer types.

Ploidy changes in tumor genomes are a hallmark of human cancer. Tetraploidization—the doubling of a complete set of diploid chromosomes—is one class of ploidy abnormality that results from a whole-genome doubling (WGD). WGD has been studied in both prokaryotic and eukaryotic species, where it has been viewed through an evolutionary lens whereby organisms that have undergone WGD have an advantage that allows them to outcompete their diploid progenitors[1]. In normal human development, tetraploidization of the genome is rare, except in germ cells during meiosis[2].

WGD has been identified in previous studies of cancer in model systems. It is thought to arise from underlying errors in cell division[2], propagate due to a defective G1 checkpoint[3], and contribute to a multitude of malignant phenotypes[4]. In human tumors, WGD has been identified incidentally as part of previous large-scale studies of DNA copy-number alterations (CNAs)[5,6] or analyses defining the phylogenetics of disease evolution[7]. One challenge in studying WGD in solid tumors has been distinguishing a singular WGD event from what may be multiple successive and independent CNAs. This is compounded by the fact that WGD may be permissive of subsequent chromosomal aberrations and genomic instability[8]. Another challenge has been delineating the mutational correlates of WGD in a cohort of diverse cancer types of sufficient population size to draw robust inferences. Finally, due to the limited clinical outcome data available for most large-scale genomic cohorts, little is known about the broader clinical significance of WGD beyond targeted cancer-type-specific studies[5,8,9]. WGD is therefore a common but still cryptic event in human cancers, the evolution and clinical impact of which has not yet been broadly defined in both common and rare cancers.

We inferred WGD status from targeted clinical sequencing of several hundred cancer-associated genes in matched tumor and normal blood specimens acquired as part of a large prospective genomic profiling initiative, the primary goal of which was to inform the care of active cancer patients. Utilizing a computational framework, we developed a simulation-based metric to identify tumors of high ploidy due to a likely singular WGD event as distinct from those with a similar burden of genomic alterations acquired from independent and successive CNAs, all estimated from purity-corrected genome-wide integer copy number. Upon determining the likely presence or absence of WGD in the cancer genomes of 9,692 prospectively sequenced patients[10], we sought to systematically assess its evolutionary impact, genomic associations, and prognostic significance in both common and rare cancers, and to evaluate whether the availability of such information could ultimately impact clinical management.

Results

Genome doubling is among the most common events in cancer. We identified the presence of genome doubling in the tumors of prospectively characterized advanced cancer patients using an analysis of allele-specific DNA copy number, which counts maternal and parental alleles based on the sequencing coverage and genotypes of germline SNPs (Supplementary Fig. 1). In heterozygous regions of a diploid cancer genome, there is one copy of each maternal and paternal allele. However, in a genome-doubled tumor, the number of copies of the more frequent allele (major copy number (MCN)) should be elevated across a substantial fraction of the

[1]Marie-Josée and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [4]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [5]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Department of Medicine, Weill Cornell Medical College, Cornell University, New York, NY, USA. [7]Present address: Genentech, San Francisco, CA, USA. *e-mail: taylorb@mskcc.org

cancer genome. We thus quantified the fraction of the autosomal tumor genome with an MCN of two or greater for all patients in the cohort and found that the distribution of this metric was bimodal, indicating that two distinct groups of cancers exist irrespective of cancer type (Fig. 1a). A considerable number of tumors had 50% or more of their autosomal tumor genome with a somatic MCN of two or more, and were therefore classified as having undergone WGD. While we could not exclude the possibility that individual tumors had acquired this copy-number genotype via successive and independently arising genomic gains of equal copy number, we modeled this scenario by simulating thousands of cancer genomes by randomly selecting 22 autosomes from WGD-negative and -positive tumors alike and were unable to reconstruct a tumor genome with an equal or greater copy-number genotype in WGD-negative cases as determined by this threshold (Supplementary Fig. 2). To confirm that WGD inferred from hybrid-capture, targeted clinical sequencing was representative of what can be achieved with broader-scale sequencing, we generated matched whole-exome sequencing data on the tumor and normal specimens of 149 patients in this cohort. WGD was concordant in 147 cases (99%), confirming the robustness of our analytical inference of WGD in targeting sequencing data. Finally, because WGD is called from allele-specific copy-number inference independent of the somatic mutational data[11], we used mutant allele fractions to assess the accuracy of our calls in WGD-positive genomes where alternative WGD-negative solutions existed. We examined how well these two opposing solutions explained the observed mutant allele fractions of somatic mutations in balanced tetraploid regions of the affected genomes, and confirmed that the WGD-positive solutions were consistent with mutant allele fractions corresponding to one and two copies of four total (after and before WGD, respectively; Supplementary Fig. 3).

In total, 28.2% of cancer patients had tumors that underwent WGD (Supplementary Tables 1 and 2). Notably, this rate of WGD was similar to that of a second orthogonal cohort of whole-exome sequencing data from 6,184 primary untreated tumors generated by The Cancer Genome Atlas (TCGA; see URLs) (31%; see Methods). In our cohort, WGD was one of the most common molecular abnormalities in human cancers, second only to *TP53* mutations (39% of patients affected) in its prevalence. WGD was more than twice as common as oncogenic *KRAS* mutations and *TERT* promoter mutations (~13% each), the next most common molecular aberrations. The median ploidy of tumors having undergone WGD was 3.3 (interquartile range (IQR): 2.9–3.8), compared with 2.1 in tumors lacking WGD (IQR: 1.9–2.4; $P<10^{-16}$, Mann-Whitney *U*-test) (Fig. 1b). As most WGD-positive tumors had sub-tetraploid genomes, we sought to time the emergence of broad single-copy losses relative to WGD in the pathogenesis of these tumors. Of 73,545 total arm- and chromosome-length heterozygous losses in WGD-positive tumors, ~70% arose after the WGD event ($P=2.7\times10^{-68}$, chi-squared test after adjusting for doubled-genome content), reflecting how such tumor cells tolerate a multitude of large-scale losses after WGD to evolve more stable sub-tetraploid tumor genomes[8] (Fig. 1b).

The rate of WGD varied markedly by cancer type, affecting 58% of germ cell tumors versus only 5% or fewer non-Hodgkin lymphomas and gastrointestinal neuroendocrine tumors (Fig. 1c). WGD was also associated with histologically distinct subtypes of disease. For instance, papillary thyroid tumors had little evidence of WGD, consistent with their oncogene-driven but otherwise quiet genomes[12]. Conversely, 46% of all Hürthle-cell thyroid cancers underwent WGD (Supplementary Fig. 4). WGD rates also varied in molecularly distinct subtypes of disease. For example, while 36% of all colorectal cancers underwent WGD, WGD arose exclusively in microsatellite stable tumors ($P=1.8\times10^{-11}$, chi-squared test; Fig. 1d). This pattern was also evident in other cancer types with frequent microsatellite instability (MSI), including endometrial cancers and stomach adenocarcinomas. In total, 0 of 110 tumors with MSI confirmed by conventional immunohistochemistry and orthogonally verified from sequencing data (Supplementary Fig. 5) underwent WGD ($P=4.2\times10^{-13}$, chi-squared test). Given the remodeling of cancer genomes after WGD via large-scale heterozygous losses (Fig. 1b), the absence of WGD in MSI tumors may be due to negative selection in tumor cells against acquiring the likely deleterious presence of both events.

Genomic correlates of genome doubling. Given the rate and variability of WGD across cancer types, we sought to determine whether an association existed between specific genetic lesions and WGD. First, we assessed whether tetraploidization was associated with an increased accrual of somatic mutations, either through having more DNA content to mutate, or because high ploidy buffers tumors against the possible deleterious effect of higher mutational burden[13,14]. Whereas the ploidy-corrected mutational rate of WGD-positive and -negative tumors within individual cancer types was approximately constant, the total mutational load of WGD-positive tumors was significantly higher than WGD-negative tumors (Supplementary Fig. 6). Next, we explored the association between WGD and *TP53*, as intact p53 is thought to prevent genome-doubled cells from re-entering the cell cycle and proliferating[15]. Consistent with these data, we found that WGD was nearly twice as common in *TP53*-mutant tumors (1.8-fold; $P=7.2\times10^{-77}$, chi-squared test)—an association that varied by lineage (Fig. 2a). Nevertheless, 21% of all *TP53*-wild-type tumors still underwent WGD, which represents nearly half (46%) of all the WGD observed here.

To understand the temporal relationship between *TP53* mutations and WGD, we timed the emergence of these events in the molecular pathogenesis of affected tumors using sequencing data (see Methods and Supplementary Fig. 7). Chronologically, WGD arose after functional *TP53* mutations in 97.3% of the patients in whom these two events could be unambiguously timed (1,142 of 1,174 in total; Fig. 2b)—a result consistent with previous estimates[16]. To test this association in tumors where the *TP53* mutation was unequivocally the first molecular event, we examined the tumor genomes of Li–Fraumeni syndrome patients harboring pathogenic germline mutations in *TP53* among 3,136 patients in this cohort who consented for germline analysis of cancer predisposition genes as part of their somatic mutational profiling. Notably, all such patients had tumors with large-scale ploidy defects, with 75% (6 of 8) having undergone WGD. In patients with WGD-positive tumors, other known oncogenic driver mutations[17] similarly preceded WGD in 81.1% of such cases, but this was only true 57.8% of the time for non-hotspot mutations of unknown significance ($P=4\times10^{-39}$, chi-squared test; Fig. 2b). Moreover, while WGD was associated with and followed *TP53* mutations, the incidence of WGD did not vary as a function of the type of *TP53* dysfunction. WGD arose at a similar frequency in tumors with *TP53* mutations that were missense variants of unknown significance, missense likely dominant-negative, or truncating loss-of-function (Fig. 2c). These findings indicate that while WGD likely arises early in the pathogenesis of many cancers, it typically follows earlier-arising transforming mutational events in *TP53* and other cancer genes, although *TP53* dysfunction is not an obligate event for WGD.

To explore additional potential genotypic associations with WGD, particularly in tumors that lack a *TP53* alteration, we constructed a multivariable logistic regression model adjusted for cancer type and the presence of other mutations and CNAs. WGD was significantly associated with multiple histologies of germ cell tumors (mixed histology, yolk sac tumors, seminomas, and embryonal carcinomas; *P* values of 10^{-5} to 0.002, Wald test; Fig. 2d)—a result consistent with germ cells doubling their genomes before meiotic divisions. As expected, this model predicts that *TP53* hotspot mutations increase the odds of a tumor undergoing WGD by a factor of 1.75 (Fig. 2d and Supplementary Table 3). Curiously, focal

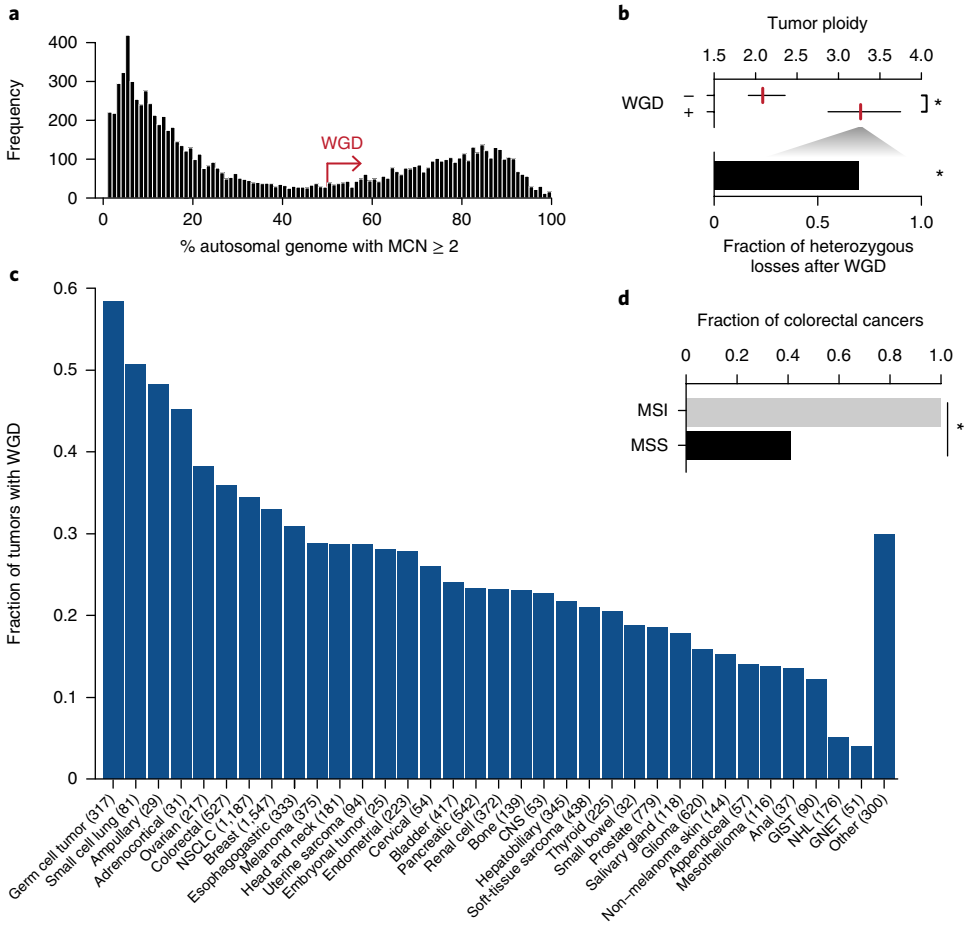

Fig. 1 | Prevalence of genome doubling in advanced cancers. a, Bimodal distribution of the fraction of the autosomal genome with an MCN of two or greater in the prospectively characterized cohort (specimens of largely copy-neutral genomes with <2% MCN of two or greater are not shown). **b**, Top: median (red) and IQR of ploidy among cases with and without WGD ($n = 2,833$ and 7,511, respectively; $^*P < 10^{-16}$, two-sided Mann-Whitney *U*-test). Bottom: fraction of large-scale heterozygous losses that molecular timing analysis indicates arose after WGD. $^*P = 2.7 \times 10^{-68}$, two-sided chi-squared test after adjusting for doubled genome content. **c**, Prevalence of WGD by cancer type. The number of samples per class is indicated in parentheses. CNS, central nervous system; GIST, gastrointestinal stromal tumor; GNET, gastrointestinal neuroendocrine tumor; NHL, non-Hodgkin lymphoma; NSCLC, non-small-cell lung cancer. **d**, Prevalence of WGD in colorectal cancers as a function of their microsatellite status ($n = 430$ for microsatellite stable (MSS) and 72 for MSI; $P = 1.8 \times 10^{-11}$, two-sided chi-squared test).

amplifications of *MDM2*, which inhibits wild-type p53, were detected in 3.5% of tumors and were mutually exclusive with *TP53* mutations ($P = 1.4 \times 10^{-38}$, Fisher's exact test), but were not associated with WGD in the multivariable model ($P = 0.65$, Wald test). Moreover, after adjusting for *TP53* status and cancer type, we found no association between WGD and somatic mutations in *APC*, *LATS1*, and *AURKA*—genes previously speculated to be associated with tetraploidization within and across cancer types[2]. Whereas telomere dysfunction-dependent tetraploidization has been studied extensively[2,18], there was also no association between WGD and *TERT* promoter mutations in this prospective cohort, or with telomere length in retrospectively characterized tumors of the TCGA[19] (see Methods and Supplementary Fig. 8).

Several other recurrent alterations were independently associated with WGD (nominal $P < 0.001$, Wald test) after adjusting for cancer type and other alterations. Among these were amplifications of *CCNE1*, and loss-of-function mutations in *RB1* and *BAP1*. *CCNE1* amplifications have been previously associated with WGD[6], and were associated with WGD here independent of cancer type and *TP53* status. *RB1* loss was also strongly associated with WGD after adjusting for *TP53* status and cancer type (Fig. 2d). Although it has previously been associated with chromosomal aberrations, a role for *RB1* loss in genome doubling has only been speculated[2]. As these findings imply that multiple aberrations converge on a defect in G1 arrest of the cell cycle, it was notable that focal *CCND1* amplifications were also modestly associated with WGD (Supplementary Table 3), a result consistent with experimental data showing that cyclin D1 over-expression in *TP53*-wild-type cancer cells renders them permissive for WGD[20]. Interestingly, *CDK4* amplifications (2.4% of all cases) that likewise inhibit RB1 and therefore E2F-mediated G1 were not associated with WGD ($P = 0.66$, Wald test). Therefore, a limited number of functionally non-redundant genomic aberrations are associated with WGD and converge on the E2F-mediated G1 arrest in both *TP53*-mutant and *TP53*-wild-type tumors. This conclusion is further supported by the association between WGD and *BAP1* mutations ($P = 0.0002$, Wald test), the loss of which has been linked to mitotic progression and chromosome instability[21], as well as G1 arrest via E2F target gene regulation[22]. Overall, 31.8% of *TP53*-wild-type WGD-positive tumors harbored a defect in an effector of E2F-mediated G1 arrest (Supplementary Fig. 9). To verify that cancer type was not a major driver of these associations, we repeated the model after having left out individual cancer types in which key lesions are common. In this subsequent analysis, there was no change in their association or

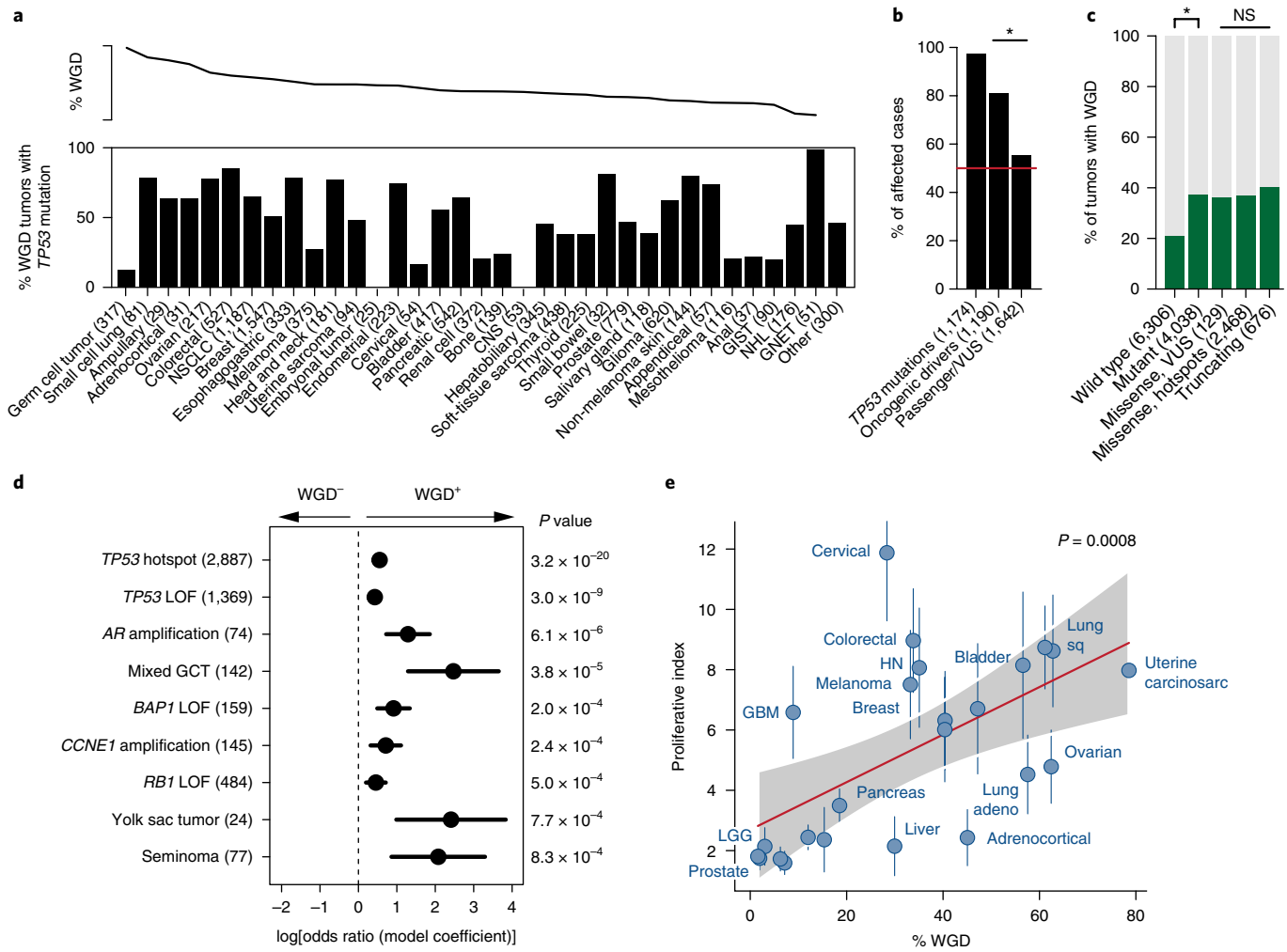

Fig. 2 | Genome correlates of genome doubling. a, Top: percentage of cases with WGD by cancer type, as sorted in Fig. 1c. Bottom: percentage of WGD-positive tumors in each cancer type that also possess a *TP53* mutation. The number of samples per class is indicated in parentheses. **b**, Percentage of WGD-positive cases in which *TP53* mutations, other oncogenic driver mutations, or presumed passenger mutations or variants of unknown significance (VUS) preceded the WGD event. Horizontal red line indicates 50%. The number of samples per class is indicated in parentheses. $^*P = 4 \times 10^{-39}$, two-sided chi-squared test. **c**, Rate of WGD in cases with different *TP53* genotypes, from wild type to mutant, and among different classes of mutations. The number of samples per class is indicated in parentheses. $^*P = 7.2 \times 10^{-77}$, two-sided chi-squared test. NS, not significant ($^*P$ values ranging from 0.10 to 0.98). **d**, Statistically significant associations (nominal $P < 0.001$) with WGD across the cohort as assessed by a multivariable regression model. Error bars on the model coefficients (log odds ratio) are plus or minus two times the standard error. The number of samples per variable is indicated in parentheses. GCT, germ cell tumor, LOF, loss of function. **e**, Correlation between the rate of WGD and the median proliferative index inferred from DNA and RNA sequencing of the same specimens in 24 cancer types from TCGA. Vertical lines represent the median absolute deviation of the proliferative index. The red line is the Spearman's rank correlation ($P = 0.0008$) and the shaded area is the 95% prediction interval. For clarity, cancer types shown but not labeled include (from left to right) thyroid, renal papillary, renal cell, endometrial, stomach, sarcoma, esophageal. adeno, adenocarcinoma; carcinosarc, carcinosarcoma; GBM, glioblastoma multiforme; HN, head and neck; LGG, low-grade glioma; sq, squamous cell.

lack thereof with WGD, indicating that our results reflect fundamental genotypic associations with WGD independent of cancer type.

Taken together, these results indicate that WGD does not result from a clear antecedent aberrant genetic alteration, but instead results from errors in cell division, and that WGD-positive tumor cells with a defect in G1 arrest more readily propagate. This model predicts that cancer types with greater rates of cell turnover would have greater rates of WGD. To test this hypothesis, we used RNA sequencing to calculate a proliferative index[23] for each tumor of 24 cancer types (TCGA) for which we had already inferred the presence or absence of WGD (see Methods). We found that while the rate of WGD in these cancer types was not correlated with the total number of divisions of normal stem cells in these tissues[24] (rho = 0, $P = 1$, Spearman's rank correlation), WGD was strongly correlated with the median proliferative index (rho = 0.65, $P = 0.0008$, Spearman's rank correlation; Fig. 2e). In fact, the variable rate of proliferation in different tumor lineages can explain 42% of the variability we observed in WGD rates across cancer types (Fig. 1c).

Genome doubling predicts worse overall survival pan-cancer. This cohort comprised cancer patients for whom prospective clinical sequencing was performed to guide treatment decisions for the management of advanced and metastatic disease. Detailed characteristics of this cohort have previously been described[10]. The characteristics of this cohort afforded the opportunity to assess the clinical implications of WGD in the setting of advanced disease. First, we explored the effect of WGD on prognosis across the entire cohort and found that it predicted for worse overall survival

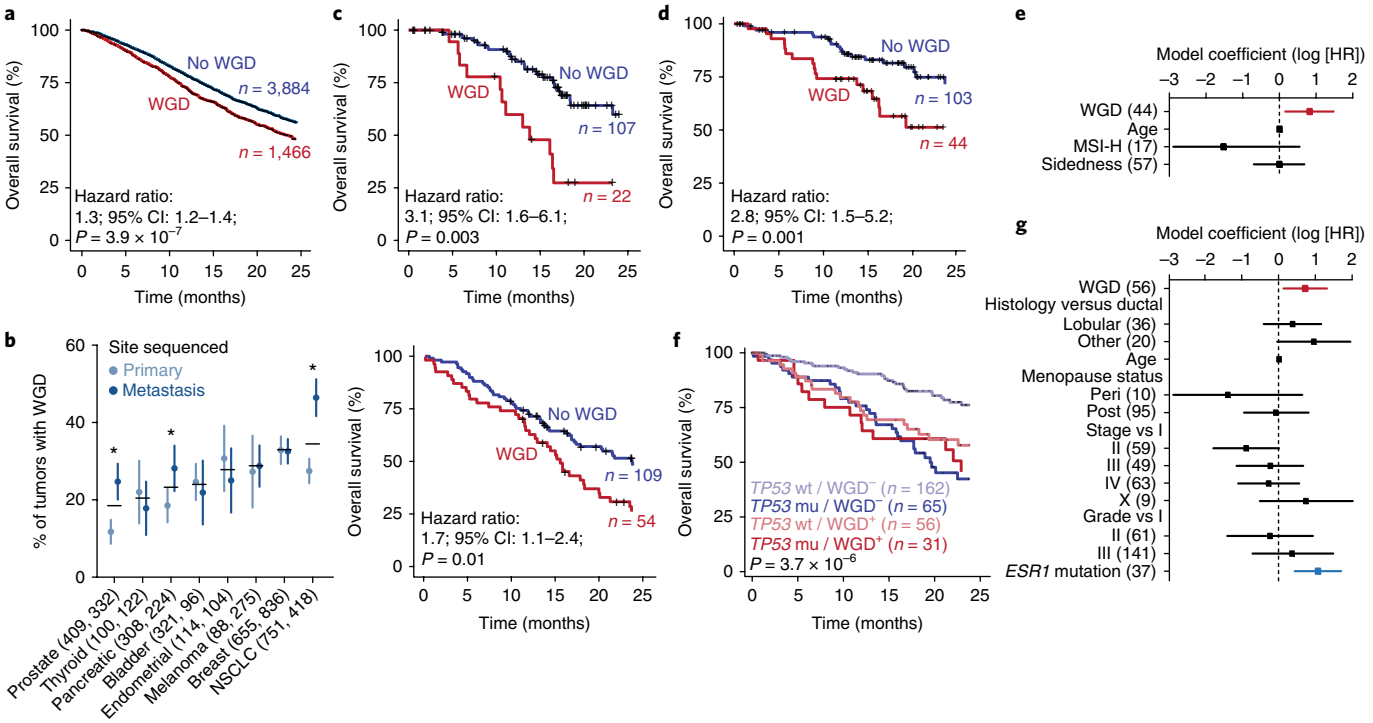

Fig. 3 | Genome doubling and outcome. a, The presence of WGD in the genomes of advanced cancers is associated with worse overall survival. **b**, Prevalence of WGD in primary and metastatic tumors of multiple cancer types. The numbers of primary and metastatic samples, respectively, are indicated in parentheses. Horizontal black bars indicate combined WGD rates for each cancer type. Error bars represent binomial CIs. *$P = 1.3 \times 10^{-4}$ for prostate cancer, 0.042 for pancreatic cancer, and 8.1×10^{-6} for non-small-cell lung cancer (NSCLC), two-sided Fisher's exact test. **c**, WGD in primary pancreatic cancers in the current study cohort (top) and an independent cohort of surgically resected primary pancreatic cancers from the International Cancer Genome Consortium (bottom). Although WGD is significantly more common in metastatic pancreatic cancers (see **b**), here it was associated with a worse prognosis in both our study cohort (even after adjusting for age, resection, and *TP53* mutational status) and the independent cohort (statistical significance determined by LRT for both). **d**, The presence of WGD in the tumor genomes of patients with *KRAS*-mutant colorectal cancers is associated with worse overall survival (statistical significance determined by LRT). **e**, Model showing how the presence of WGD in the tumor genomes of patients with *KRAS*-mutant colorectal cancers is influenced by other variables. Known prognostic variables (age at diagnosis, microsatellite status, and right- versus left-sided disease) are included. The number of samples per variable is indicated in parentheses. HR, hazard ratio; MSI-H, high microsatellite instability. **f**, Tumor-specific WGD in patients with hormone-receptor-positive/HER2-negative *TP53*-wild-type breast cancers is associated with worse overall survival (statistical significance determined by LRT). mu, mutant; wt, wild type. **g**, Multivariate model showing the influence of various prognostic variables at breast cancer diagnosis, as well as the effect of *ESR1* mutations, on overall survival in patients with hormone-receptor-positive/HER2-negative *TP53*-wild-type breast cancers. The number of samples per variable is indicated in parentheses. HR, hazard ratio.

pan-cancer (hazard ratio: 1.3; 95% confidence interval (CI): 1.2 to 1.4; $P = 3.9 \times 10^{-7}$, likelihood ratio test (LRT); Fig. 3a). After adjusting for cancer type, age, and *TP53* mutational status, WGD remained significantly associated with decreased overall survival pan-cancer (hazard ratio: 1.18; 95% CI: 1.08 to 1.32; $P = 0.0005$, Wald test).

Another unique characteristic of this cohort was the inclusion of not only primary tumors from patients with advanced disease, but also metastatic samples in a subset of cases. In total, 42% of samples analyzed here were obtained from metastatic tumors. To control for potential confounding of overall survival based on whether the sample sequenced was a primary tumor versus metastasis, we sought to establish whether WGD was observed more commonly in metastatic compared with primary tumor samples. Adjusting for *TP53* mutation status, WGD was no more common in metastatic than primary tumors in the majority of cancer types, demonstrating that the negative prognostic effect of WGD could not be explained solely by its enrichment in metastatic samples (Supplementary Fig. 10). WGD was, however, significantly more common in non-small-cell lung, pancreatic, and prostate cancer metastases (Fig. 3b). In prostate cancers, we validated that WGD was far more prevalent in prostate cancer metastases than primary tumors in an independent cohort of ~1,000 prostate cancers for which both whole-exome sequencing and detailed clinical data were available (46 and 6%, respectively; $P = 3 \times 10^{-47}$, chi-squared test)[25]. When present in the primary prostate cancers of our prospectively sequenced cohort (14% of 797 patients), WGD was associated with high- rather than low- or intermediate-risk Gleason grade ($P = 7.3 \times 10^{-7}$, chi-squared test; Supplementary Fig. 11). As WGD is associated with the subsequent acquisition of large-scale CNAs (Fig. 1b), this result may explain, in part, the association between an increasing burden of CNAs with biochemical recurrence and metastasis in patients with prostate cancer[26,27]. Similarly, in pancreatic cancers where WGD is significantly more common in metastatic tumors, WGD in primary adenocarcinomas was associated with a worse prognosis (hazard ratio: 3.1; 95% CI: 1.6 to 6.1; $P = 0.003$, LRT) (Fig. 3c)—an association with higher-risk disease that we replicated in an independent cohort of surgically resected primary pancreatic adenocarcinomas of the International Cancer Genome Consortium (Fig. 3c).

As WGD was prognostic for overall survival—even in patients with incurable cancer—we hypothesized that WGD would have clinical significance independent of established prognostic factors in cancer types for which patients have heterogeneous clinical outcomes even in the setting of established metastatic disease. We therefore curated detailed clinical data for two of the most prevalent

cancer types with such clinical heterogeneity: *KRAS*-mutant colorectal cancers[28] and estrogen receptor (ER)-positive/human epidermal growth factor receptor 2 (HER2)-negative breast cancers[29,30]. We found that *KRAS*-mutant colorectal cancers that underwent WGD had a significantly worse prognosis than *KRAS*-mutant cancers that lacked this event (hazard ratio: 2.8; 95% CI: 1.5 to 5.2; $P=0.001$, LRT), even after adjusting for other variables prognostic at metastasis, including age at diagnosis, microsatellite status, and right- versus left-sided disease (hazard ratio: 2.3; 95% CI: 1.2 to 4.4; $P=0.015$, Wald test; Fig. 3d,e). Another common cancer type with substantial clinical heterogeneity in advanced-stage patients is the 70% of breast cancers that are ER-positive and HER2-negative. While WGD was not associated with outcome in *TP53*-mutant ER-positive, HER2-negative breast cancers, it was significantly associated with worse prognosis in *TP53*-wild-type patients (hazard ratio: 2.0; 95% CI: 1.2 to 3.3; $P=0.01$, LRT; Fig. 3f), even after adjusting for clinical features prognostic at breast cancer diagnosis (hazard ratio: 2.1; 95% CI: 1.1 to 3.7; $P=0.016$, Wald test). Notably, WGD had an effect size similar to *ESR1* mutations, which emerge in patients previously treated with antihormonal therapy[31,32] (Fig. 3g). In both the ER-positive, HER2-negative breast and the *KRAS*-mutant colorectal cancers, a quantitative measure of the overall burden of genomic alteration in these tumors (the fraction of the autosomal tumor genome bearing CNAs of any kind) was not significantly associated with survival. This finding suggests that WGD, rather than the chromosomal aberrations that follow, is the basis for these prognostic differences.

Discussion

Here, we establish WGD as one of the most prevalent singular genomic aberrations in human cancer. WGD does not appear to have an obligate antecedent genetic basis. Our analysis instead supports an evolutionary model whereby WGD emerges early in the pathogenesis of affected cancers, but after a preceding oncogenic driver mutation. Such lesions include those that lead to either *TP53* dysfunction or, in *TP53*-wild-type tumors, an E2F-mediated G1 arrest defect. These lesions increase the likelihood of a tumor undergoing WGD, but are not required for it. Nevertheless, a model in which WGD arises early after a preceding oncogenic event that initially transforms the cell is consistent with data indicating that spontaneous tetraploidization of non-transformed human cells is rare. Overall, the data are consistent with an earlier transforming lesion establishing a permissive environment for the proliferation of cancer cells that subsequently undergo a genome doubling after stochastic errors in cell division.

Our findings may have important implications for understanding the molecular pathogenesis and therapeutic management of human cancers. WGD is a macro-evolutionary step in affected cancers, and tumors having undergone WGD evolve sub-tetraploid genomes via an increased burden of subsequent large-scale single-copy losses. These CNAs arise later in the molecular pathogenesis of affected cases, implying that WGD may serve as a precursor of the subclonal diversification of CNAs that has recently been shown to be associated with poor outcomes in lung adenocarcinoma patients[7]. Indeed, the increased prevalence of WGD we observed in metastatic specimens of some cancer types, rather than arising late in the evolution of these tumors and contributing to the transition to metastatic disease, may reflect an early event that, when present, indicates a more aggressive subset of primary disease with a worse prognosis (as in prostate and pancreatic cancers).

Clinically, WGD is associated with adverse survival pan-cancer in patients with advanced disease and in cancers with heterogeneous clinical outcomes, even following the development of metastasis. The ability of WGD to identify poor-prognosis primary tumors, as in the case of the pancreatic cancers profiled here (Fig. 3c), could inform the design of new adjuvant trials in specific populations of high-risk patients. Key questions about how (1) previous therapy impacts the prognostic impact of WGD, (2) WGD contributes to better or worse[9] response to targeted, systemic, or immunotherapies, and (3) WGD may lead to unique therapeutic vulnerabilities, and whether this is due to the WGD event itself or the subsequent evolution of genomic aberrations, will require further clinical and functional investigation. At present, even within the context of the prospective sequencing of cancer patients from which our cohort was drawn, the presence of WGD is not being reported to clinicians.

Overall, prognostics in advanced disease is an understudied area, despite considerable clinical variability among late-stage patients of multiple cancer types. In some instances, prognostic biomarkers may mature into valuable predictive biomarkers. However, for these to inform clinical management at the point of care, they must be captured from current clinical molecular testing methodologies. In the case of WGD, concurrent sequencing of matched normal specimens from cancer patients is essential for its robust detection. This underscores the need for simultaneous sequencing of tumor and matched normal specimens from patients to not only facilitate integrated reporting of germline and somatic findings that simplifies the clinical workflow and hastens the speed of molecular testing, but also inform clinical care beyond the presence of sensitizing therapeutic biomarkers. Indeed, our analysis of WGD was performed in prospectively characterized cancer patients using clinical sequencing data, the results of which could be practice-changing if evidence-based guidelines can be established for the use of this information to inform clinical decision-making.

URLs. cBioPortal for Cancer Genomics, http://cbioportal.org/; The Cancer Genome Atlas, http://cancergenome.nih.gov/; OncoKB Knowledge Base, http://www.oncokb.org/;

UniProt annotations, https://github.com/mskcc/vcf2maf/blob/v1.6.13/data/isoform_overrides_uniprot.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0165-1.

Received: 6 August 2017; Accepted: 22 May 2018;
Published online: 16 July 2018

References

1. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
2. Davoli, T. & de Lange, T. The causes and consequences of polyploidy in normal development and cancer. *Annu. Rev. Cell Dev. Biol.* **27**, 585–610 (2011).
3. Storchova, Z. & Pellman, D. From polyploidy to aneuploidy, genome instability and cancer. *Nat. Rev. Mol. Cell Biol.* **5**, 45–54 (2004).
4. Fujiwara, T. et al. Cytokinesis failure generating tetraploids promotes tumorigenesis in *p53*-null cells. *Nature* **437**, 1043–1047 (2005).
5. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
6. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
7. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
8. Dewhurst, S. M. et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
9. Kuznetsova, A. Y. et al. Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. *Cell Cycle* **14**, 2810–2820 (2015).
10. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
11. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).

12. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).
13. Semon, M. & Wolfe, K. H. Consequences of genome duplication. *Curr. Opin. Genet. Dev.* **17**, 505–512 (2007).
14. Thompson, D. A., Desai, M. M. & Murray, A. W. Ploidy controls the success of mutators and nature of mutations during budding yeast evolution. *Curr. Biol.* **16**, 1581–1590 (2006).
15. Aylon, Y. & Oren, M. p53: guardian of ploidy. *Mol. Oncol.* **5**, 315–323 (2011).
16. 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).
17. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
18. Davoli, T. & de Lange, T. Telomere-driven tetraploidization occurs in human cells undergoing crisis and promotes transformation of mouse cells. *Cancer Cell* **21**, 765–776 (2012).
19. Barthel, F. P. et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).
20. Crockford, A. et al. Cyclin D mediates tolerance of genome-doubling in cancers with functional p53. *Ann. Oncol.* **28**, 149–156 (2017).
21. Peng, J. et al. Stabilization of MCRS1 by BAP1 prevents chromosome instability in renal cell carcinoma. *Cancer Lett.* **369**, 167–174 (2015).
22. Pan, H. et al. BAP1 regulates cell cycle progression through E2F1 target genes and mediates transcriptional silencing via H2A monoubiquitination in uveal melanoma cells. *Int. J. Biochem. Cell Biol.* **60**, 176–184 (2015).
23. Ramaker, R. C. et al. RNA sequencing-based cell proliferation analysis across 19 cancers identifies a subset of proliferation-informative cancers with a common survival signature. *Oncotarget* **8**, 38668–38681 (2017).
24. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
25. Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
26. Hieronymus, H. et al. Copy number alteration burden predicts prostate cancer relapse. *Proc. Natl Acad. Sci. USA* **111**, 11139–11144 (2014).
27. Taylor, B. S. et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
28. Punt, C. J., Koopman, M. & Vermeulen, L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nat. Rev. Clin. Oncol.* **14**, 235–246 (2017).
29. Hart, C. D. et al. Challenges in the management of advanced, ER-positive, HER2-negative breast cancer. *Nat. Rev. Clin. Oncol.* **12**, 541–552 (2015).
30. Zardavas, D., Irrthum, A., Swanton, C. & Piccart, M. Clinical management of breast cancer heterogeneity. *Nat. Rev. Clin. Oncol.* **12**, 381–394 (2015).
31. Chandarlapaty, S. et al. Prevalence of ESR1 mutations in cell-free DNA and outcomes in metastatic breast cancer: a secondary analysis of the BOLERO-2 clinical trial. *JAMA Oncol.* **2**, 1310–1315 (2016).
32. Toy, W. et al. ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat. Genet.* **45**, 1439–1445 (2013).

Acknowledgements

We thank our patients and their families for participating in this study, and members of the Taylor Laboratory and Marie-Josée and Henry R. Kravis Center for Molecular Oncology for discussions and support. We also thank V. Balachandran and the Avner Pancreatic Cancer Foundation for assistance with ICGC pancreas sequencing and clinical data. This work was supported by National Institutes of Health awards P30 CA008748, U54 OD020355 (to D.B.S. and B.S.T.), R01 CA207244 (to D.M.H. and B.S.T.), and R01 CA204749 (to B.S.T.), and the American Cancer Society, Cycle for Survival, Sontag Foundation, Prostate Cancer Foundation, and Josie Robertson Foundation (to B.S.T.).

Author contributions

C.M.B. and B.S.T. conceived the study. C.M.B., A.V.P., M.T.A.D., M.T.C., P.J., C.B., M.F.B., and B.S.T. designed and performed the data analysis. A.Z., W.C., J.A., A.M.S., P.R., G.I., M.E.R., Z.K.S., N.S., J.B., D.B.S., and D.M.H. assisted with the prospective genomic and clinical data collection and sample annotation. C.M.B. and B.S.T. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0165-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.S.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study cohort and prospective sequencing. The study cohort comprised 9,692 advanced cancer patients diagnosed with 1 of 55 principle tumor types who were enrolled onto an Institutional Review Board-approved research protocol (NCT01775072) at the Memorial Sloan Kettering Cancer Center (MSKCC) between January 2014 and November 2016 (Supplementary Tables 1 and 2). In compliance with ethical regulations, all patients provided written informed consent, and this study was conducted with the approval of the Memorial Sloan Kettering Cancer Center Institutional Review Board. Details regarding patient consent, sample acquisition, sequencing, mutational analysis, and clinical reporting have previously been described[10]. Briefly, prospective sequencing of matched tumor and blood specimens was performed using Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT)—a custom hybridization capture-based next-generation sequencing assay approved for clinical use in New York state[10,33]. This study cohort included patients whose tumors were sequenced with 1 of 3 incrementally larger versions of the MSK-IMPACT assay (containing 341, 410, and 468 genes respectively).

Allele-specific DNA copy-number analysis. Estimates of tumor purity and ploidy, as well as genome-wide total, allele-specific, and integer DNA copy number, were inferred from sequencing data using the FACETS algorithm (version 0.3.9)[11]. We utilized a two-pass implementation whereby a low-sensitivity run (cval = 100) first determined the purity and tumor-normal log-ratio corresponding to the diploid state. Gene-level segmentation and integer copy-number calls were inferred from a subsequent run with higher sensitivity for focal events (cval = 50). These calls were used to time mutations and CNAs relative to WGD, while homozygous deletion and focal amplification calls obtained using the MSK-IMPACT analytical protocol[10] were used to model the probability of WGD arising in a given sample. Tumors were considered to have undergone WGD if greater than 50% of their autosomal genome had an MCN (the more frequent allele in a given segment) greater than or equal to two. To evaluate the robustness of this metric, we simulated 1,000 pseudo-cancer genomes constructed from randomly sampling 22 autosomes from subsets of WGD-negative and WGD-positive samples (see Supplementary Fig. 2). Tumor specimens with less than 2% of their autosomal genome having an MCN greater than or equal to 2 were excluded from this simulation as copy-neutral, as were tumor samples with tumor-normal log-ratio values (FACETS dipLogR) falling in the outermost deciles of the WGD-negative and WGD-positive subsets.

We performed FACETS and WGD analysis of an independent cohort of 6,184 primary untreated tumors from 26 tumor types in the TCGA (Supplementary Table 1) using the procedure described above to ensure cross-comparability. The overall rate of WGD in this cohort was 31%, which was similar to the rate measured in our prospective cohort. This estimate is slightly lower than previous analyses of TCGA data[5,6], due primarily to the different composition of cancer types in our cohort, followed by the more conservative threshold we implement here to call WGD. Considering only the same distribution of cancer types of previous large-scale copy-number analyses[6], our estimate of the rate of WGD in our prospective cohort rises to 31%. Similarly, if we relax our threshold for WGD to 40% of the genome having an MCN greater than or equal to 2, our estimate of WGD pan-cancer rises to 33% in the prospective cohort. Telomere length was utilized as previously determined[19]. The TCGA research network data were retrieved through database of Genotypes and Phenotypes authorization accession number phs000178.v9.p8.

Mutation timing analysis. Somatic mutations were timed relative to WGD using a methodology adapted from previous work[16]. Specifically, we inferred a cancer cell fraction (CCF) for all somatic mutations in all tumor samples from variant allele fractions using a binomial distribution and maximum likelihood estimation, normalized to produce posterior probabilities. Mutations were classified as clonal if the upper bound of the 95% CI for the CCF was greater than or equal to 0.95. All other mutations were classified as arising subclonally. The expected number of copies for a given mutation is a function of the variant allele fraction (VAF), total copy number (TCN), and tumor purity (Φ), and is given by:

$$\frac{\mathrm{VAF}}{\Phi} * (\mathrm{TCN} > * \Phi + 2 * (1 - \Phi))$$

The relative timing of mutations was determined using the most parsimonious explanation of an observed copy-number state. Rather than utilizing discretized allelic copy number, we instead tested whether the point estimate of mutant copies was greater than one. For example, a mutation in a region with a TCN of four and an MCN of two was regarded as a single mutation arising before WGD, as opposed to multiple independent but identical mutations affecting different alleles at the same locus arising after WGD. Therefore, clonal mutations in which the TCN and MCN were both two were classified as arising before WGD. In regions with a TCN of greater than or equal to three, clonal mutations with an expected copy number of greater than one were classified as arising before WGD. Clonal mutations in regions with a TCN equal to three and an expected copy number of less than or equal to one were classified as ambiguous and excluded from timing analyses because we could not differentiate between (1) a mutation arising before WGD followed by a single-copy loss of the mutant copy after WGD and (2) a single-copy loss after WGD followed by a mutation. Finally, all other clonal mutations were classified as having arisen before WGD, and all subclonal mutations were classified as having arisen after WGD. Our analysis of single-copy losses relative to WGD compared regions with a TCN and MCN of three and two, respectively (that is, a loss after WGD) versus regions with a TCN and MCN of two and two, respectively (that is, a loss before WGD). Regions affected by multiple copy-number losses after WGD were not considered in our analysis. Known and likely driver mutations were those mutational hotspots identified by previous methods[17] or those alleles whose functional and clinical significance has been curated by the OncoKB Knowledge Base (see URLs). All non-hotspot missense mutations were classified as putative passenger mutations or variants of unknown significance. To compare rates of WGD between classes of *TP53* mutations (that is, hotspot versus truncating), we excluded the subset of samples that harbored variants from both classes. The timing analysis of *TP53* mutations considered only hotspots and loss-of-function mutations (nonsense mutations, splice site mutations, and frameshift insertions and deletions). Overall, 64.4% of all hotspot mutations in oncogenes and 71.2% of all hotspot or loss-of-function mutations in tumor suppressor genes qualified as unambiguously timed and were utilized for the timing analysis.

Multivariable regression model associations with WGD. To explore the genomic correlates of WGD, we modeled the probability of WGD using multivariable logistic regression. Somatic mutations and focal CNAs observed 20 or more times in the prospective cohort in 1 of the 341 genes sequenced in all patients were included in our final model and coded as binary predictor variables. Overall, we considered hotspot mutations, MCN amplifications, and loss-of-function events combining nonsense and splice site mutations, frameshift insertions and deletions, and homozygous deletions. Cancer subtypes were also included in the final model. Variance inflation factors were used to detect multicollinearity arising from correlated predictor variables. To avoid testing mutually dependent observations, amplifications targeting *FGF19*, *FGF4*, *HIST1H3B*, and *IKBKE* were removed due to their proximities to other commonly co-amplified genes in affected cases.

MSI. MSI was determined in colorectal, endometrial, and stomach adenocarcinomas using MSIsensor[37]—an orthogonal bioinformatics approach to identifying MSI based on the percentage of microsatellite loci that are unstable in a tumor genome compared with its matched normal specimen. Tumors with an MSIsensor score greater than or equal to ten were classified as MSI-positive. This MSIsensor score threshold had a validation rate of 99.4% compared with conventional immunohistochemistry testing in a cohort of 180 tumors for which both measures were available (only a single discordant case called MSI by MSIsensor was equivocal by immunohistochemistry; A. Zehir, personal communication).

Correlation of WGD with the proliferative index. Gene expression was quantified from RNA sequencing of 10,535 tumor specimens from TCGA using Kallisto v0.42.4 (ref. [34]) and canonical isoforms per gene based on UniProt annotations (see URLs). After filtering to the subset of these specimens for which we had performed WGD inference from exome sequencing data, we derived the proliferative index scores for all samples from the median expression of the top 1% of genes correlated with the proliferating cell nuclear antigen proliferation marker in a cohort of normal tissues as previously described[23,35].

Statistical analysis. Associations between WGD and both clinico-pathological and genomic features were assessed using Pearson's chi-squared, Fisher's exact, and Wilcoxon tests, as well as multivariable logistic regression. To compare the rates of WGD between primary and metastatic samples, we restricted our analysis to include only the first metastatic sample per patient. To ensure sufficient statistical power for detecting true associations in the context of our multivariable logistic regression model, our analysis satisfied the established minimum number of events per variable criteria[36]. Only those covariates with a minimum of $n = 10 \times k/p$ affected samples were included in the analysis, where k is the number of covariates and p is the proportion of cases in the population being analyzed (~0.3 in this study). This corresponds to a minimum of 30 mutational events present to be included as a covariate in our model for a total of 268 covariates and a suggested n of approximately 9,000 cases, which is fewer than the 9,692 cases analyzed here. Key negative associations were all present in a number of tumors far greater than the events per variable in this study cohort and were present in sufficient numbers to have from 80 to >99% power to detect small effect sizes among individual associations (Cohen's $h = 0.2$ to 0.35).

Univariate and multivariate survival analyses were performed using Cox proportional hazards regression and displayed using Kaplan–Meier methods. Overall survival in days was the difference between the date of the procedure from which prospective sequencing was performed to the date of last follow-up. Only patients whose date of sequencing was less than one year from the date of their procedure were included in the outcome analyses (Supplementary Table 2).

P values for survival analyses were obtained using the LRT or Wald test for the multivariable analyses. All analyses were performed using R, and all figures were generated using the ggplot2 package in R.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The source code for all analyses in this study and the associated allelic data can be found at https://github.com/taylor-lab/GD.

Data availability. All primary genomic results and associated specimen annotations for all patients in this study are accessible as described for the original cohort[10] and have been deposited into the cBioPortal for Cancer Genomics for analysis and visualization at http://cbioportal.org/msk-impact.

References

33. Cheng, D. T. et al. Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
34. Vivian, J. et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).
35. Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7**, e1002240 (2011).
36. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**, 1373–1379 (1996).
37. Niu, B. et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).

nature research

Corresponding author(s): Taylor, Barry S.

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined. — No sample size was pre-determined; all available specimens were utilized.

2. Data exclusions

   Describe any data exclusions. — No exclusion criteria were specified for the study population.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings. — No experimental replication indicated

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups. — Not relevant, no experimental group allocation

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. — No relevant blinding, no group allocation

   Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	FACETS v0.3.9 was used to perform allele-specific DNA copy number analysis. Kallisto v0.42.4 was used to quantify gene expression from TCGA RNA sequencing data. All other statistical analyses were performed using the R environment for statistical computing.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	No restrictions on availability

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used in this study.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used in this study.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used in this study.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used in this study.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	All population characteristics (ages, cancer types, genotypes, etc.) are specified in the main and supplementary text or in an accompanying manuscript (Zehir A, et al. Nat Med. 2017).