Interplay between whole genome doubling and the accumulation of deleterious alterations in cancer evolution

Saioa López (1,2), Emilia L. Lim (2,3), Stuart Horswell (4), Kerstin Haase (5), Ariana Huebner (1,2,3), Michelle Dietzen (1,2,3), Thanos P. Mourikis (1,2), Thomas B.K. Watkins (3), Andrew Rowan (3), Sally M. Dewhurst (6), Nicolai J. Birkbak (3,7), Gareth A. Wilson (3), Peter Van Loo (5, 8), Mariam Jamal-Hanjani (2,9), TRACERx Consortium (10), Charles Swanton* (2,3), Nicholas McGranahan* (1,2)

(1) Cancer Genome Evolution Research Group, University College London Cancer Institute, London, UK

(2) Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK

(3) Cancer Evolution and Genome Instability Laboratory, the Francis Crick Institute and University College London Cancer Institute, London, UK

(4) Bioinformatics and Biostatistics Group, The Francis Crick Institute, London, UK

(5) Cancer Genomics Laboratory, The Francis Crick Institute, London, UK

(6) Laboratory for Cell Biology and Genetics. The Rockefeller University. New York, US

(7) Department of Molecular Medicine, Aarhus University, Aarhus, Denmark

(8) Department of Human Genetics, University of Leuven, Leuven 3000, Belgium

(9) Department of Medical Oncology, University College London Hospitals NHS Foundation Trust, London, UK

(10) A full list of authors can be found at the end of the article

*joint corresponding authors, email: charles.swanton@crick.ac.uk or nicholas.mcgranahan.10@ucl.ac.uk

Abstract

Whole genome doubling (WGD) is a prevalent event in cancer, involving a doubling of the entire chromosome complement. However, despite its prevalence and prognostic relevance, the evolutionary selection pressures for WGD have not been investigated. Here, we combine evolutionary simulations with an analysis of cancer sequencing data to explore WGD during cancer evolution. Simulations suggest WGD can be selected to mitigate the irreversible, ratchet-like, accumulation of deleterious somatic alterations, provided they occur at a sufficiently high rate. Consistent with this, we observe an enrichment for WGD in tumor types with extensive loss of heterozygosity (LOH), including lung and triple negative breast cancers, and we find evidence for negative selection against homozygous loss of essential genes prior to, but not after, WGD. Finally, we demonstrate that LOH and temporal dissection of mutations can be exploited to identify novel tumor suppressor genes and to obtain a deeper characterization of known cancer genes.

Introduction

Whole genome doubling (WGD), involving the duplication of a complete set of chromosomes, is a common feature of cancer genomes[1,2]. WGD has been linked to increased tumor cell diversity, accelerated cancer genome evolution and worse prognosis[1,3,4].

Polyploidy is found across several plants and animal species, and many hypotheses have been proposed regarding its selective advantages and/or disadvantages [5,6]. Polyploidy could be a mechanism to mitigate the Muller's ratchet effect[7], originally described for asexual populations, where due to the lack of recombination, deleterious mutations would accumulate in an irreversible manner over long evolutionary times. This phenomenon has been observed in nature in organisms such as the diploid Amazon molly[8], the self-fertile worm *C. elegans*[9], asexual DNA-based microbes such as *Salmonella typhimurium*[10] and amoebae[11]. The phenomenon is also described in the haploid setting in the context of the evolution of the Y-Chromosome[12] and propagation of mitochondrial DNA[13] in humans.

Cancer development can be considered analogous to asexual evolution. Although cancer cells are not long-lived lineages, they may also be subject to the irreversible accumulation of deleterious passenger alterations[14,15], and this effect may be particularly marked in genomic segments exhibiting loss of heterozygosity (LOH).

Here, we investigate whether WGD buffers the deleterious impact of somatic mutations and somatic copy number alterations (SCNA) in regions of LOH and explore scenarios in which WGD may be selected. We focus on non-small cell lung cancer (NSCLC), one of the cancer types with the highest frequencies of WGD[1]. We explore this in the TRACERx (Tracking Non-Small-Cell Lung Cancer Evolution through Therapy) cohort[16], a prospective and longitudinal study with multiregional data, and use the lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) data from The Cancer Genome Atlas (TCGA)[17]. Furthermore, we also investigate how gene duplication can be exploited for identification of novel cancer genes and apply this approach to 33 different cancer types.

Results

LOH and WGD are common events in NSCLC

First, we explored WGD, LOH and mutational burden in NSCLCs from the TRACERx[16] and TCGA datasets[17]. WGD was a frequent event in NSCLC (64.45% in LUAD, 66.12% in LUSC tumors), consistent with previous work[1,16]. Using TRACERx data, where multiregion data were available, we confirmed that WGD is primarily an early event in cancer progression, with only a small subclonal proportion (6.25% in LUSC and 1.63% in LUAD) as previously reported in Jamal-Hanjani et al. (2017) (Figure 1a,b).

We noted a lower proportion of the genome subject to LOH in non-WGD tumors compared to their genome doubled counterparts (Figure 1a-c), but as expected, by virtue of the additional genome copies, on average only 1.32% of the genome was haploid in WGD tumors, compared to 9.55% in nWGD cases (Figure 1a,b). Clonal LOH was significantly more common than subclonal LOH (average of 25.85% vs 5.96%, Figure 1c), suggesting the majority of LOH occurs early, likely before WGD.

Additionally, we observed a significant positive correlation between the amount of haploid LOH in non-WGD tumors and the frequency of WGD events across tumor types (Figure 1d). These observations led us to explore whether WGD could mitigate the potential negative effects of haploidy and the accumulation of deleterious alterations (mutations and/or SCNA) in these regions.

The fitness cost of WGD is offset by its positive impact

To explore the impact of a WGD event on cancer cell viability and fitness and whether a duplication event may be selected to buffer deleterious alterations (Figure 2a,b), we adapted a tumor progression model from [14].

In our model, alterations in cancer genes is associated with a fitness gain (s_d) while passenger alterations in haploid regions are associated with a weak fitness cost (s_p), reducing the birth rate. WGD influences the fitness cost of these passenger alterations, reducing their negative impact, but is itself associated with a fitness cost (s_{WGD}).

To explore WGD we simulated the evolution of cancer populations (starting with a 1,000 cells), and varied the fitness costs of passenger alterations ($s_p \in \{0\text{-}0.01\}$, where 0.01 represents a 1% increase in waiting time to birth) and WGD ($s_{WGD} \in \{0\text{-}1.5\}$) (Supplementary Table 1). Deleterious passenger alterations represent either mutations or SCNA losses affecting genes

in regions of LOH. To explore whether WGD is selected, at the end of each simulation we calculated the proportion of WGD in the population. We initially assume the rate of deleterious alterations, u_p, is 0.5 and that cells adapt to the cost of WGD after ten generations. The deleterious alteration rate reflects the product of number of putatively deleterious sites (i.e. the extent of haploid LOH before doubling) and the alteration rate per-base-per-cell-division. We simulate until the population reaches 20,000 cells or 2,000 generations. Using these parameters, after 1,000 generations we observed an average of 489 [381-514] passenger and 8 [7-9] driver alterations.

Our simulations suggest a relationship between fitness cost of passenger alterations and the likelihood of WGD being selected; higher fitness cost of passenger alterations was associated with increased WGD (Figure 2c). However, this relationship was also dependent on the cost of WGD; a higher fitness cost was associated with reduced WGD, when passenger alterations were associated with very minor fitness costs (s_p <0.00002). Notably, provided the fitness cost of passenger alterations exceeded 0.0002, WGD was selected, even when the fitness cost of WGD was high (s_{WGD}>0.5). Moreover, if the fitness cost of WGD was low (s_{WGD}=0.05) and the cost of passengers relatively high (s_p=0.001), WGD was selected even when the fitness cost was maintained for >100 generations (Supplementary Figure 1a).

However, when the average passenger fitness cost was negligible (s_p <0.00001), the benefit of WGD did not compensate for its cost. Likewise, if the deleterious alteration acquisition rate was low (u_p<0.2 for $s_p = 9.1 \times 10^{-5}$ or u_p<0.05 for $s_p = 4 \times 10^{-4}$), e.g. due to a low proportion of the genome subject to LOH, or if the WGD fitness cost was high (sWGD>0.5) and durable (>25 generations) WGD was rarely selected (Figure 2d).

Notably, a stable genome without any haploid regions, will exhibit a deleterious acquisition rate approaching zero (i.e $u_p \simeq 0$). This may be the case for hyper-mutator tumors (for example *POLE* mutant tumors), which display a lower burden of SCNA and low frequency of WGD [18]. By contrast, a high deleterious rate of deleterious alteration acquisition may occur in with a high degree of haploidy and chromosomal instability, whereby there is a high probability of a SCNA leading to homozygous loss. These results are in agreement with the relationship between LOH and WGD proportions across cancers (Figure 1d).

To evaluate whether the fitness costs explored are compatible with previously documented lack of detectable negative selection[19] we explored the relationship between the cost of passenger alterations and purifying selection. Despite being associated with WGD, weakly deleterious alterations (s_p<0.0005) largely escaped detectable negative selection

(Supplementary Figure 1b). Conversely, alterations with a substantial fitness cost (s_p >0.01) were associated with strong negative selection.

Taken together, this model, despite simplifying cancer evolution, suggest WGD in cancer cells can act as an evolutionary mechanism that might serve to buffer the deleterious effect of somatic alterations. The model suggests that provided there is a sufficient deleterious alteration rate, through SCNA or mutations, WGD can potentially be selected.

Accurate timing of mutations pre and post WGD

To quantify the impact of WGD in cancer evolution using sequencing data and to determine whether there is a shift in detectable selection following a duplication event, it is imperative to be able to time somatic alterations in relation to WGD. Previous work has suggested WGD provides a natural mechanism to temporally dissect mutations[20,21]. In brief, mutations occurring prior to WGD should be present at multiple copies, while those occurring after a doubling event would only be present at one copy (Figure 3a).

However, this assumption has not been subject to experimental validation. Therefore, we utilized an isogenic genome doubling model system involving genome-doubled HTC-116 clones deriving from a non-genome doubled common ancestor[3] and exome-sequenced the ancestor in addition to two diploid and four tetraploid cell lines at two time points (passages 4 and 50). Given that the common ancestor was diploid, all clonal mutations should occur before WGD, while private mutations should almost all occur after WGD.

Reassuringly, 90.19% of pre-genome doubled mutations were correctly timed as such. Conversely, 94.70% tetraploid private mutations were correctly classified as occurring after WGD (Figure 3b). Applying temporal dissection of mutations to tumors exhibiting WGD in TRACERx and TCGA datasets suggests that the majority of detectable clonal mutations in NSCLC accumulate prior to WGD (Figure 3c).

Purifying selection on essential genes prior to duplication

If WGD has a significant impact upon selection and the evolutionary course of the disease, we reasoned one would expect to see a difference in the selection of mutations before and after doubling. While our simulations suggest weakly deleterious alterations may escape detectable selection, purifying selection will likely operate on clones bearing deleterious alterations in housekeeping or essential genes, located in haploid regions of the genome (where s_p > 0.01). We reasoned that if WGD buffers deleterious alterations, purifying selection pressures will be relieved after duplication of essential genes.

To investigate the selective pressures acting before and after WGD, and the effect of LOH, we applied a modified dNdS ratio test [19] to early (pre-WGD) and late (post-WGD) mutations within segments of LOH. Under the assumption that synonymous mutations are neutral, this ratio can be informative about the direction of selection: ratios >1 indicate positive selection, while ratios <1 are consistent with negative selection. We investigated different sets of genes, including essential genes identified by using mutagenesis screens in haploid human cells[22] and LUSC and LUAD-specific cancer genes described in the literature[19,23-26]. We focussed on dNdS values for truncating mutations (including nonsense and splice-site mutations), which would be most closely associated with protein dysfunction and should in theory be subject to stronger purifying selection (Figure 4a). dNdS values for missense mutations are shown in Supplementary Figure 2.

Analyses at the whole exome level (all genes) showed no deviations from neutrality (Figure 4a), as previously reported[19], and consistent with our simulations (Supplementary Figure 1). However, signals of purifying selection were observed for essential genes occurring in early mutations (pre-WGD) encoded within genomic segments of LOH (Figure 4a, Supplementary Figure 3). These data suggest homozygous disruption to essential genes in regions of LOH in diploid genomes resulting in disruption of both alleles, leads to a detectable fitness impairment; indeed, on average, approximately half of clones harboring early truncating mutations in regions of LOH are predicted to be lost due to purifying selection. However, due to the limited number of point mutations in essential genes, we can estimate that on average only one clone harboring a nonsense mutation in an essential gene is eliminated per tumor prior to WGD during its evolution.

We hypothesized that following a WGD event, new mutations in genomic segments of LOH occur less frequently in a haploid context and therefore purifying selection would be weaker. Consistent with this, mutations occurring after WGD in essential genes were not found to be subject to significant purifying selection (Figure 4a). Likewise, in non-LOH regions (which harbor more mutations), dNdS values were not significantly different from 1. These results suggest that genomic segments that accumulate truncating or nonsense mutations but still maintain one or more wild-type copies of the gene are not subject to strong negative selection as they are still viable. Therefore, by reducing the haploid genome fraction of cancer cells, WGD may act to mitigate the detrimental effect of deleterious mutations.

To rule out an excess of SNP contamination in essential genes leading to spurious signatures of negative selection, we compared the proportion of somatic mutations contained in the SNP database (dbSNP) in essential[22] vs non-essential genes and all genes (Supplementary Figure 4). The proportion of mutations classified as SNPs was low in essential, non-essential and all

genes in both LUAD (0.017, 0.022 and 0.022, respectively) and LUSC (0.025, 0.026 and 0.025, respectively), suggesting germline contamination is not responsible for the observed negative selection.

In order to assess the effect of using different essential gene lists, in addition to the essential genes from[22](Figure 4a), we took advantage of CRISPR screens and calculated the dNdS ratio for pre-WGD mutations in LOH in NSCLC specific essential genes[27]. We noted strong purifying selection acting on genes with the highest gene-knockout effects (i.e. genes whose deletion has the most negative impact on cell viability) (Supplementary Figure 5). We also quantified selection in other essential gene lists from additional CRISPR-based model-systems[28,29]. The results were broadly consistent. For TCGA-LUSC, we observed the same trend regardless of the gene-set used (Supplementary Figure 6). For TCGA-LUAD we noted that while negative selection was observed prior to doubling in all cases, for genes derived from [29], a weak signal of negative selection was observed for post WGD mutations, potentially due to some genes acting in a haploinsufficient manner.

Taken together these results suggest that WGD may mitigate against nonsense mutations in essential genes. However, given that these mutations are relatively rare events during cancer evolution (u_p<0.0001), these data, together with the simulations, suggest WGD cannot be selected solely to buffer against nonsense mutations in essential genes.

We therefore next explored whether WGD may also have an impact on the acquisition of SCNAs. Homozygous deletions are rare in NSCLC genomes, consistent with complete loss of gene function generally having a negative effect on cell fitness. Genomic segments which are haploid may be particularly susceptible to copy number loss events, which would result in homozygous deletions. To explore whether the SCNA landscape is influenced by negative selection we considered whether the proportion of haploid LOH in WGD tumors was less than expected by chance. We simulated copy number losses randomly (at the chromosome arm level) in WGD tumors, fixing the number of chromosome arms in pre-WGD LOH based on the observed data. Next, we quantified the number of post-WGD losses resulting in haploid LOH. Through simulations in WGD tumors exploring where copy number losses occur across the genome, we found significantly less haploid LOH than would be expected by chance, consistent with negative selection removing those tumor clones (Figure 4b). Furthermore, we explored the correlation between copy number losses and essentiality of the chromosome arms (as captured by the density of essential genes)[30] and observed a significant negative correlation between the essential score and haploid LOH occurrence for each chromosome arm (Figure 4c), consistent with negative selection resulting in fewer copy number losses where there is an abundance of essential genes. This correlation, however, was not observed

when we focused on copy number losses occurring after the WGD event (Figure 4d), consistent with negative selection sculpting the SCNA landscape, reducing haploidy and homozygous deletions.

Similar selection patterns across other cancer types

Prompted by these findings, we sought to further investigate selection across other cancer types. Consistent patterns were observed among most of those cancer types with sufficient patients/mutations to perform selection analyses (Supplementary Figure 7). Significant signatures of purifying selection in pre-WGD truncating mutations in LOH in essential genes (but not after WGD) were observed in skin cutaneous melanoma (SKCM), which is characterized by high mutation rates and high LOH (Supplementary Figure 6). In other cancer types (liver hepatocellular carcinoma -LIHC-, colon adenocarcinoma -COAD-, uterine corpus endometrial carcinoma -UCEC-, bladder urothelial carcinoma -BLCA-, cervical squamous cell carcinoma and endocervical adenocarcinoma -CESC-, head and neck squamous cell carcinoma -HNSC-, esophageal carcinoma -ESCA-), although not significant, a similar trend was also observed.

Leveraging mutations in LOH to identify novel cancer genes

Positive selection is a feature of cancer, reflected as an enrichment of non-silent mutations in cancer genes[19,31,32], and thus the dNdS ratio can be exploited to identify new cancer genes. Next, we explored dNdS values specifically for known cancer genes. We observed highly significant positive selection of truncating mutations prior to WGD in regions of LOH, followed by reduced positive selection after duplication (Figure 4a). These data may reflect strong selection for inactivation of both copies of tumor suppressor genes through LOH and mutation, prior to WGD.

We hypothesized that by specifically looking into genomic segments of LOH we could identify tumor suppressor genes that would not be otherwise detected. Moreover, we could explore whether selection intensities varied depending on whether the mutation co-occurred with LOH or the wild-type allele remained intact.

We obtained dNdS values for nonsense mutations at the gene level and compared the selection intensities for early mutations in genomic regions of LOH (Figure 5a, x-axis) compared to mutations not occurring in segments of LOH (Figure 5a, y-axis). Notably, two genes, with very high selection coefficients in LOH but lower in non-LOH regions in LUSC were *TP53* and *PTEN* (Figure 5a, left). These are well-established tumor suppressor genes that, concordant with the “two-hit hypothesis”[33], appear to require disruption of both alleles.

ZNF750 (Zinc Finger Protein 750), is subject to high positive selection in regions of LOH in LUSC and would have remained undetected with standard procedures. Notably, *ZNF750* has previously been described as a lineage-specific tumor suppressor gene in squamous cell carcinoma[34], consistent with its occurrence as an early event in 87% (7/8) of LUSC tumors. Other genes that were only identified as significant when looking at mutations in regions of LOH were *NOTCH1* and *SMAD4* in LUSC (Figure 5b and Supplementary Figure 3). On the other hand, *CUL3* (Cullin-3), which plays an important role in the ubiquitin-proteasome system[35], showed high signatures of selection in genomic regions not subject to LOH, possibly reflecting haploinsufficient activity in cancer, and, conceivably, the requirement for an intact wild-type allele.

Finally, we extended the analyses to the remaining cancer subtypes in TCGA. A total of 33 potential tumor suppressor genes were identified by limiting the analyses to segments of LOH (Figure 5b). Some of these are included in the COSMIC and/or Network of Cancer Genes (NCG6.0) databases as well-characterized cancer genes (*RB1*, *PTEN*, *SMAD4*, *BAP1*, *SETD2*, *NOTCH1*, *BRCA1*, *NF2* and *CDK12*). In other cases, we re-discovered genes that have been identified in other cancer types or pan-cancer studies, but not specifically in that cancer type, like *SMAD4* in CESC, *CDH1* in BLCA, *ZNF750* in LUSC, *ARHGAP35*, *TUBA3C* and *PTCH1* in HNSC, *SMAD3* in COAD and *ITGAV* in READ. Finally, 42% of these were additional novel cancer genes may be acting as tumor suppressor genes not included in the current release of COSMIC and NCG6.0: *WWC1*, *FPF12*, *NT5DC3, NCLN*, *KRTAP19-5*, *GRIK2*, *GLRA1*, *FAXDC2*, *FAM19A3*, *CRYGC*, *CLEC4E*, *BC02, ARPP21* and *AC061992.1*.

As an additional approach for the identification of driver genes under different scenarios (mutations in LOH, non-LOH, all) we also implemented MutSigCV[26,36]. Supplementary Figure 8 shows those genes with q-values<0.1. We detected six potential cancer genes when looking at LOH mutations that were not significant when assessing all mutations together. These include the well-established cancer genes *RB1*, *NCOR*, *CDNK2A*, *FBXW7* and *TSC1*, as well as *PCDHGA3* in LUAD.

Discussion

Despite the fact that WGD is associated with poor prognosis and an acceleration of chromosomal instability [1,3], a rational basis for the observation of the recurrence of WGD events in human tumors remains unclear. Cancer progression is an evolutionary process, and as such, the fundamental principles of Darwinian evolution can be applied to study tumor development. In this work, we explore the evolutionary importance of WGD in cancer, and show how this can be exploited to identify novel cancer genes.

While human germline evolution is dominated by a relatively low mutation rate coupled with negative or purifying selection, which removes deleterious or harmful mutations, cancer evolution is characterized by high mutation rates and positive selection[19]. Indeed, the extent of negative selection during somatic evolution has been subject to debate. For example, while negative selection has been reported in transcription factor binding motifs [37], hemizygous regions[19,38] and splicing-associated sequences[39], conflicting reports regarding the extent to which the immune system results in negative selection and mutation loss have been presented[19,32,40,41]. Previous works showing that hemizygous regions may be subject to purifying selection[19,38] only consider genomic segments that are currently haploid, reducing the analysis to a very small proportion of the genome (<5%).

Here we consider regions of LOH which are not haploid but once were, prior to genome doubling. Thus, we are able to explore the presence of purifying selection in a considerably larger proportion of the genome (>25%). And, crucially, we are able to time mutations relative to doubling, permitting an exploration of whether selection pressures change following WGD.

Focussing on genomic segments exhibiting LOH, we demonstrate that truncating mutations in essential genes occurring before WGD in a haploid context are subject to negative selection in lung cancer evolution. Analogous to haploid asexual and non-recombining populations in nature, cancer cells will accumulate these alterations irreversibly, in a ratchet-like process. Thus, in cancers with a high rate of deleterious alteration acquisition (reflecting both point mutations and SCNA losses) and high levels of LOH, and in the absence of other compensating mechanisms, this may lead to the attrition of subclones. We find evidence that a WGD event in cancer progression seemingly relieves the impact of deleterious alterations. Thus, regardless of the underlying reason for why WGD occurs, by duplicating haploid genomic segments (which hence become diploid), WGD may attenuate cancer cell attrition through disruption of the genome.

Although tumors may exhibit an advantage as a consequence of WGD, it cannot evolve to prevent acquisition of deleterious mutations in the future and may itself be associated with a

fitness cost[3]. Our simulations suggest the cost of WGD can be offset by its protective effect against the accumulation of weakly deleterious alterations in haploid genomic regions (Figure 2). However, crucially WGD is only selected when the rate of deleterious alteration acquisition is sufficiently high. This is consistent with the significant relationship between the extent of LOH and WGD across cancer types (Figure 1d). Our simulations and data suggest that for WGD to be selected to buffer deleterious alterations, this cannot solely be due to truncating mutations in essential genes in haploid regions, but likely also reflects ongoing chromosomal instability (CIN) with loss of the remaining haploid genomic regions, sometimes encoding genes essential for survival[3]. These results are also in keeping with a conspicuous absence of WGD in hypermutator tumors, which generally do not harbor extensive LOH or chromosomal instability.

Further experimental work is required to explore the fitness consequences of WGD and to obtain a detailed understanding how different types of alterations, including nonsense and SCNA may negatively impact upon the fitness of cancer cells and how this changes during the disease course. In addition, there are likely many other benefits to WGD which have not been explored here. Indeed, previous work has suggested a triploid karyotype represents an optimal fitness state for cancer cells [42].

An additional contribution from this work is the identification of 33 potential cancer genes, identified by considering mutations specifically occurring in segments of LOH. Temporal dissection of mutations, coupled with a focus on regions of the genome exhibiting LOH, enables elucidation of genes subject to two-hits (mutation and LOH) and strong signals of positive selection. Importantly, this signal may be missed without such dissection. Our framework enabled elucidation of 14 putative cancer genes that are not currently included in the COSMIC/NCG database, and the “re-discovery” of other cancer genes in a different cancer type. In addition, our results confirm that many established tumor suppressor genes, including *PTEN* and *RB1*, likely require both hits to be subject to positive selection, while other cancer genes, including *CUL3*, are subject to strong positive selection without two hits. Conceivably, a similar framework could be applied to identify cancer genes subject to either mutation and methylation.

In conclusion, our study highlights the parallels between species and cancer evolution and emphasizes the importance of punctuated events such as WGD in cancer development and cell survival. Identifying cellular mechanisms that lead to WGD, and cancer cell vulnerabilities that ensue from this event, may provide a unique approach to limit cancer evolution, adaptation and disease progression.

Acknowledgements

S.L. receives funding from Rosetrees. P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. C.S. is Royal Society Napier Research Professor. This work was supported by the Francis Crick Institute that receives its core funding from Cancer Research UK (FC001169,FC001202), the UK Medical Research Council (FC001169, FC001202), and the Wellcome Trust (FC001169, FC001202). C.S. is funded by Cancer Research UK (TRACERx, PEACE and CRUK Cancer Immunotherapy Catalyst Network), the CRUK Lung Cancer Centre of Excellence, the Rosetrees Trust, NovoNordisk Foundation (ID16584) and the Breast Cancer Research Foundation (BCRF). This research is supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (Grant Number: SU2C-AACR-DT23-17). Stand Up To Cancer is a program of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. N.M is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant Number 211179/Z/18/Z), and also receives funding from CRUK Lung Cancer Centre of Excellence, Rosetrees, and the NIHR BRC at University College London Hospitals.

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) Consolidator Grant (FP7-THESEUS-617844), European Commission ITN (FP7-PloidyNet 607722), an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement 835297), and Chromavision from the European Union's Horizon 2020 research and innovation programme (grant agreement 665233).

The results published here are in part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. The data were retrieved through database of Genotypes and Phenotypes (dbGaP) authorization (Accession No. phs000178.v9.p8). Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

We also thank Christopher McFarland for kindly sharing code for simulating deleterious alterations in cancer evolution.

Author contributions

Conceptualization and supervision: N.M and C.S. Manuscript preparation: S.L.,N.M. Manuscript review/editing: S.L., C.S., N.M. Simulations: S.H., S.L. Formal analysis: S.L., E.L. Visualization/data presentation: S.L. Data curation and interpretation of results: S.L., E.L., A.H., M.D., T.M., T.W., N.B., G.W., N.M. Resources: S.W. A.R. K.H., P.V.L, M.J.H, C.S, N.M.

Author Information

The authors declare competing financial interests: C.S. receives grant support from Pfizer, AstraZeneca, BMS, and Ventana. C.S. has consulted for Boehringer Ingelheim, Eli Lily, Servier, Novartis, Roche-Genentech, GlaxoSmithKline, Pfizer, BMS, Celgene, AstraZeneca, Illumina, and Sarah Cannon Research Institute. C.S. is a shareholder of Apogen Biotechnologies, Epic Bioscience, GRAIL, and has stock options and is co-founder of Achilles Therapeutics. N.M. and G.W. has stock options and has consulted for Achilles Therapeutics. Correspondence and material requests should be addressed to C.S. (Charles.Swanton@crick.ac.uk) and N.M. (Nicholas.Mcgranahan.10@ucl.ac.uk).

References

1. Bielski, C.M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* **50**, 1189-1195 (2018).
2. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-40 (2013).
3. Dewhurst, S.M. *et al.* Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov* **4**, 175-185 (2014).
4. Storchova, Z. & Pellman, D. From polyploidy to aneuploidy, genome instability and cancer. *Nat Rev Mol Cell Biol* **5**, 45-54 (2004).
5. Huxley, J. *Evolution. The modern synthesis*, (George Alien & Unwin Ltd., London: , 1942).
6. Madlung, A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity (Edinb)* **110**, 99-104 (2013).
7. Muller, H.J. The relation of recombination to mutational advance. *Mutat Res* **106**, 2-9 (1964).
8. Loewe, L. & Lamatsch, D.K. Quantifying the threat of extinction from Muller's ratchet in the diploid Amazon molly (Poecilia formosa). *BMC Evol Biol* **8**, 88 (2008).
9. Loewe, L. & Cutter, A.D. On the potential for extinction by Muller's ratchet in Caenorhabditis elegans. *BMC Evol Biol* **8**, 125 (2008).
10. Andersson, D.I. & Hughes, D. Muller's ratchet decreases fitness of a DNA-based microbe. *Proc Natl Acad Sci U S A* **93**, 906-7 (1996).
11. Maciver, S.K. Asexual Amoebae Escape Muller's Ratchet through Polyploidy. *Trends Parasitol* **32**, 855-862 (2016).
12. Engelstadter, J. Muller's ratchet and the degeneration of Y chromosomes: a simulation study. *Genetics* **180**, 957-67 (2008).
13. Loewe, L. Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. *Genet Res* **87**, 133-59 (2006).
14. McFarland, C.D., Korolev, K.S., Kryukov, G.V., Sunyaev, S.R. & Mirny, L.A. Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci U S A* **110**, 2910-5 (2013).
15. McFarland, C.D., Mirny, L.A. & Korolev, K.S. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proc Natl Acad Sci U S A* **111**, 15138-43 (2014).
16. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2109-2121 (2017).
17. Campbell, J.D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet* **48**, 607-16 (2016).

18. Shlien, A. *et al.* Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat Genet* **47**, 257-62 (2015).
19. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
20. 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).
21. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
22. Blomen, V.A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092-6 (2015).
23. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
24. Berger, A.H. *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* **30**, 214-228 (2016).
25. Bertrand, D. *et al.* ConsensusDriver Improves upon Individual Algorithms for Predicting Driver Alterations in Different Cancer Types and Individual Patients. *Cancer Res* **78**, 290-301 (2018).
26. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
27. Meyers, R.M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* **49**, 1779-1784 (2017).
28. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515-26 (2015).
29. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-101 (2015).
30. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-62 (2013).
31. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-44 (2013).
32. Zapata, L. *et al.* Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol* **19**, 67 (2018).
33. Knudson, A.G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-3 (1971).
34. Hazawa, M. *et al.* ZNF750 is a lineage-specific tumour suppressor in squamous cell carcinoma. *Oncogene* **36**, 2243-2254 (2017).
35. Chen, H.Y. & Chen, R.H. Cullin 3 Ubiquitin Ligases in Cancer Biology: Functions and Therapeutic Implications. *Front Oncol* **6**, 113 (2016).
36. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
37. Vorontsov, I.E. *et al.* Negative selection maintains transcription factor binding motifs in human cancer. *BMC Genomics* **17 Suppl 2**, 395 (2016).
38. Van den Eynden, J., Basu, S. & Larsson, E. Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution. *PLoS Genet* **12**, e1006506 (2016).
39. Hurst, L.D. & Batada, N.N. Depletion of somatic mutations in splicing-associated sequences in cancer genomes. *Genome Biol* **18**, 213 (2017).
40. Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479-485 (2019).
41. Eynden, J.V.d., Jiménez-Sánchez, A., Miller, M.L. & Larsson, E. Lack of detectable neoantigen depletion in the untreated cancer genome. *bioRxiv*, 478263 (2018).
42. Laughney, A.M., Elizalde, S., Genovese, G. & Bakhoum, S.F. Dynamics of Tumor Heterogeneity Derived from Clonal Karyotypic Evolution. *Cell Rep* **12**, 809-20 (2015).

43. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

Figure legends:

Figure 1. Prevalence of whole genome duplication (WGD) and loss of heterozygosity (LOH) in NSCLC. a-b) Proportion of WGD, subclonal WGD and non-WGD genomes (left), and proportion of the genome subject to LOH and haploid LOH in WGD vs nWGD (right) in LUSC (a) and LUAD (b). For all boxplots, the upper whisker indicates the largest value (no further than 1.5* inter-quartile range (IQR) of the box-edge), and the lower whisker corresponds to the smallest value at most 1.5* IQR of the box-edge; the median is indicated by the thick horizontal line; and the first and third quartiles are indicated by box edges; data beyond whiskers are 'outliers' and plotted individually. c) Differences in the proportion of clonal vs subclonal LOH in TRACERx data. Significant differences between groups was assessed with a t-test. d) Proportion of WGD tumors vs haploid LOH in nWGD tumors across 34 cancer types in the TCGA cohort. LUSC, lung squamous carcinoma; LUAD, lung adenocarcinoma

Figure 2. Whole genome doubling (WGD) buffers the deleterious effect of passenger alterations. a) The principle of Muller's ratchet in asexual and sexual organisms. The red dots represent the mutations acquired over time on chromosome segments. Asexual organisms with no recombination accumulate mutations in an irreversible manner leading towards cell death or extinction, while sexual populations with recombination are viable for longer periods of time. b) WGD buffers the effects of late (post-WGD) deleterious mutations in regions of loss of heterozygosity (LOH) by providing additional mutation-free segments. c) Proportion of WGD cells in the tumor at the end of simulations with varying values of WGD-associated cost, s_{WGD}, and passenger fitness costs, s_p. The cost relates to a proportional reduction in birth rate, such that 0.5 cost represents a 50% increase in the waiting time to the next birth. d) Relationship between the proportion of cells subject to WGD and deleterious alteration rate, u_p, for two different values of s_p.

Figure 3. Timing mutations relative to whole genome doubling (WGD). a) Those mutations occurring before genome duplication (red dots) will be present at multiple copies, whereas those occurring after the duplication event (blue dots) will only be present at only one copy. b) Validation of the mutation timing approach using an isogenic genome doubling system involving genome-doubled HTC-116 clones deriving from a non-WGD common ancestor. Barplots show the proportion of mutations correctly vs incorrectly classified as either pre-WGD or post-WGD by our timing approach for common mutations in the tetraploids (left) and tetraploid private mutations (i.e., not present in the diploid genomes) (right). c) Number of mutations in regions of LOH across LUSC and LUAD datasets, grouped by WGD status and timing. LUSC, lung squamous carcinoma; LUAD, lung adenocarcinoma

Figure 4. Purifying selection before but not after whole genome doubling (WGD). a) dNdS values for truncating mutations in WGD tumors calculated for all genes, essential genes and lung-specific

cancer genes, grouped by loss of heterozygosity (LOH) status and timing of the mutations in lung squamous carcinoma (LUSC) and lung adenocarcinoma (LUAD) from the TRACERx dataset (n=93), LUSC from TCGA (n=325) and, LUAD from TCGA (n=398). A dNdS ratio of 1 (red line) is consistent with neutrality. Values significantly higher than 1 (consistent with positive selection) are shown in dark blue. Values significantly lower than 1 (indicating purifying selection) are shown in red. Each point represents a dNdS estimate, with 95% CI shown. b) In WGD tumors difference in the number of chromosome arms in haploid LOH found through simulations vs the observed data in the TCGA NSCLC cohorts (n=970) and the ploidy of the tumors (below). Tumors with very high ploidy (>4.5) rarely exhibit any haploid LOH in simulations or in observed data. No multiple test correction was performed. c) Pearson Correlation between the essential genes score [30] and the frequency of haploid LOH per chromosome arm (n=39) for TCGA NSCLC (n=970). d) Correlation between the essential genes score[30] and the frequency post-WGD losses for TCGA NSCLC. Shaded region indicates 95% confidence interval, and Pearson's correlation coefficient is indicated.

Figure 5. Exploiting loss of heterozygosity (LOH) to identify cancer genes. a) dNdS selection coefficients for truncating mutations in early mutations in LOH (i.e. mutations present on all copies of remaining allele that is not subject to LOH) (x-axis) vs truncating mutations in genomic regions without evidence of LOH (y-axis). The background color indicates whether the gene was identified as significant using mutations in early LOH (dark blue) using all mutations (grey), or identified as significant in both cases (light blue) (q-value<0.05). The border color represents whether the gene is currently included in the COSMIC/NCG databases as a cancer gene. Data from TCGA and TRACERx LUSC (n=356) and TCGA and TRACERx LUAD (n=460) is shown b) dNdS selection coefficients for truncating mutations in early mutations in LOH ("LOH") vs truncating mutations in regions without evidence of LOH ("noLOH") across cancer types. Only genes that are significant in at least one cancer type in the LOH category are shown. Barplots show the total number of cancer genes that are significantly (q-value<0.05) identified using the "all approach" (grey) and the number of cancer genes that are only identified (q-value<0.05) using the "LOH" approach (dark blue). The latter is also represented with a number above the bars, and the specific genes are marked with dots in the heatmap. Only those cancers where we identify additional cancer genes using the "LOH" approach (this is, only looking at early mutations in LOH) are shown. (BRCA TNBC= triple negative breast cancer, HNSC=head and neck squamous cell carcinoma, BRCA ER+=ER positive breast cancer, COAD=colon adenocarcinoma, LUSC=lung squamous cell carcinoma, CHOL=cholangiocarcinoma, KIRP=kidney renal papillary cell carcinoma, LIHC=liver hepatocellular carcinoma, OV=ovarian serous cystadenocarcinoma, BLCA=bladder urothelial carcinoma, CESC=cervical squamous cell carcinoma and endocervical adenocarcinoma, KIRC=kidney renal clear cell carcinoma, READ=rectum adenocarcinoma, UCS=uterine carcinosarcoma.)

Online Methods

Data processing

We analyze 93 patients from the first cohort of patients of NSCLC (61 LUSC and 32 LUAD) obtained through the lung TRACERx (TRAcking Cancer Evolution through therapy (Rx)) project and thoroughly described in [16].

Additionally, raw .bam files for LUAD (n=398) and LUSC (n=325) samples from the TCGA repository were downloaded and processed through the TRACERx pipeline. Briefly, we used BWA-MEM to align the reads to the reference genome (build hg19). We used Platypus for SNP calling on the germline, and Varscan2 and MuTect for somatic mutation calling. Functional annotation of genomic variants was performed using ANNOVAR [43]. Purity, ploidy and copy number profiles of tumor cells were obtained with ASCAT [44], using the matching germline data. Mutations in regions of LOH were timed as early or late based on the mutation and major allele copy number. Following a conservative approach, we considered early mutations those with mutation copy number >=1.75 and major allele copy number >=1.75. Mutations were classified as late if mutation copy number <=1.25 and major allele copy number >=1.75. Clonal mutations that could not be timed were classified as "unknown". Mutations were defined as mutations in LOH if the minor allele copy number was <0.25. The WGD status for each tumor was obtained using the genome doubling algorithm described in [3].

For the identification of new drivers across other cancer types, we downloaded the MC3 somatic mutation tables [45] from whole-exome sequencing data across the 33 cancer types from TCGA. Further processing of the samples was performed as described above.

Simulations

We performed simulations to model the viability of a cancerous cell over time and illustrate the potential of increasing DNA copies via genome duplication in the mitigation of damaging effects of mutations using an adapted version of the cancer evolution model developed by McFarland et al.[14,15]. Briefly, this model simulates cancer progression as a stochastic system of birth and death events using a standard Gillespie algorithm, and is defined by the mutation rate, the target sizes for drivers and passenger mutations, and the effect on the fitness of drivers and passenger mutations [14]. We have extended this model incorporating WGD events. The viability of the population of cells varies by a function of the mutations that the cell has accumulated over time, which exert deleterious or beneficial effects on cell fitness to varying degrees, depending on the type of genes where they are accumulated. Passenger alterations in regions of haploid LOH decrease the fitness by increasing the time to the next birth. Conversely, driver alterations decrease the time to the next birth.

For all the simulations we used the following fixed parameters based on the simulations by McFarland et al.[14]: starting population size (No) 1000 cells, 3000 driver loci (Td) with a maximum of 10 driver mutations in a cell, selection coefficient of the drivers (sd) 0.15. Different selection coefficients of passenger mutations (s_p) were tested, in the range of 0-0.01 (Table S1). We also explored different passenger alterations rates.

The effects of WGD events were incorporated by reducing the deleterious effects of passenger mutations by half. Note that given we are considering passenger mutations in regions of haploid LOH, we are likely underestimating the buffering effect of WGD.

The probability of WGD in a given cancer cell was defined by a normal distribution: if(rnorm(1, 0) > qnorm(0.001, lower = FALSE)). Every cell (and therefore its subsequent daughter cells) was only allowed to accumulate one WGD event. Additionally, we incorporated a fitness cost (range: 0-1) associated with surviving the WGD event that was maintained during fixed period.

In each simulation, the tumor progressed until the cell population size reached 20,000 cells - happening when a lineage reaches fixation in tumor progression very quickly - or until 2,000 generations in the cases where we don't reach the maximum number of cells but the final population reaches an equilibrium in tumor size and composition. At that point we calculated the proportion of WGD cells in the tumor. An average of 16 iterations were run for each fitness cost.

Negative selection in simulations was evaluated by comparing simulations where there was no fitness cost associated with passenger alterations, i.e. $s_p = 0$, to those where $s_p > 0$. For each s_p value, the median number of mutations at 1,000 generations was divided by the median number of mutations when $s_p = 0$. This ratio is equivalent to a dNdS ratio, such that a value <0 indicates positive selection, while a value <0 indicates a depletion of mutations and negative selection.

Copy number simulations

In order to explore whether there is less haploid LOH than would be expected by chance, we performed simulations based on copy number changes. For each of the WGD tumors in LUSC+LUAD from TCGA we determined the observed LOH and haploid LOH proportion and the total copy number gains and losses at the chromosome arm level. Using the total copy number gains and losses at the chromosome arm level as probabilities for gains and losses we simulated 10,000 random copy number scenarios and compared the simulated haploid LOH proportions with the observed ones. Essential genes scores per Chromosome arm were obtained from [30].

Selection tests

Selection across LUSC and LUAD was quantified using the dNdScv R-package [19], an implementation of the traditional dNdS ratio, adapted and refined for cancer genomes. In all cases we used the trinucleotide substitution model with 192 rate parameters and default parameters, removing ultra-hypermutator samples (3,000 maximum number of coding mutations per sample) and limiting the analyses to 3 mutations per gene per sample. dNdS ratios were quantified for missense, nonsense and essential splice mutations, both globally and at gene level.

Global dNdS values for nonsense mutations were calculated for different lists of genes: a) *all genes*, b) *cancer genes*, which includes a compilation of LUSC and LUAD-specific tumor suppressor genes and oncogenes described in the literature [19,23-26] and c) *essential genes*, which includes the essential genes reported by Blomen et al., [22], identified by using extensive mutagenesis in haploid human cells (1,154 genes). Different lists of essential genes obtained through CRISPR-based systems were also evaluated for comparison [28,29]. Essential genes from Hart et al. [28] included the 1,580 hits observed in three or more of the five cell lines. In the case of Wang et al., (2015), we included the genes with adjusted p-values <0.05 in three or more of the four cell lines used (a total of 1,282 genes). In all the analyses, cancer genes (present in the cancer gene census) were removed from the essential genes list.

Selection tests were performed on different subsets of mutations considering the timing (pre-WGD vs post-WGD mutations) and the LOH presence (LOH vs nonLOH) in WGD genomes.

Additionally, we calculated dNdS values for early mutations in LOH in essential genes inferred by NSCLC CRISPR analyses (Achilles) available at the DepMap portal https://depmap.org/portal/. Genes were ranked by the average gene-knockout effect (a measurement of the consequences of gene deletion on cell viability) across 283 samples annotated as lung cancer for the primary disease and took lists of top genes of different lengths ranging from 500 to 7,000.

Identification of driver genes

Maximum-likelihood dNdS estimates (MLE) at the gene level obtained with the dNdScv package were used to identify potential driver genes under positive selection. We compared the MLE values for nonsense mutations (wnon) in regions of LOH where the mutation was present on all copies of the remaining allele vs non-LOH in WGD and nWGD tumors combined. As an additional method to identify driver genes we used MutSigCV [36], comparing the adjusted p-values for mutations in LOH, non-LOH and all mutations combined. In order to

reduce false positives and multiple testing correction we filtered the genes that are expressed at low levels (below the median expression value across all the genes and patients) in that particular cancer type. RNAseq-based normalized expression data for each cancer type was obtained from the TCGA data portal.

Validation of the timing approach using mutation and copy number data from diploid and tetraploid clones from the HCT-116 cell line.

In order to demonstrate that we can accurately quantify which mutations and copy number losses occur pre and post the WGD event we used mutation and copy number data from 2 diploid and 4 tetraploid subclones derived from diploid human colon carcinoma HCT-116 cells, isolated at different passages (4 and 50 passages) [3].

For each sample exome capture was performed on 1-2 ug DNA isolated from genomic libraries with median insert size of 190 bp, using a customized version of the Agilent Human All Exome V5 kit, according to the manufacturer's protocol (Agilent). Samples were 100 bp paired-end multiplex sequenced on the Illumina HiSeq 2500 at the Advanced Sequencing Facility at the Francis Crick Institute. The data was aligned to the reference human genome (build hg19) and somatic mutations were then obtained with Varscan [46] and MuTect [47] as previously described [16], using the parental diploid clone as the germline. Copy number profiles were inferred with PICNIC using genotype chip data, as previously described [3]. The early (pre-WGD)/late (post-WGD) classification of the mutations was assessed using the same procedure as above.

Statistics

R version 3.3.1 was used to analyze the data. No statistical tests were used to predetermine the same size. Tests involving comparison of distributions were done using 't.test' using the unpaired option, unless otherwise stated. Confidence intervals for dNdS analysis was obtained using the dNdSCV package [19].

Data availability

HTC-116 clones' sequence data used during the study has been deposited at The National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), under the accession code: PRJNA595067.

Code availability

R code to reproduce figures is available at:

https://github.com/ucbtsl1/lopez_etal_2019_wgd-cancer

References

44. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).
45. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281.e7 (2018).
46. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-76 (2012).
47. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).

TRACERx consortium members

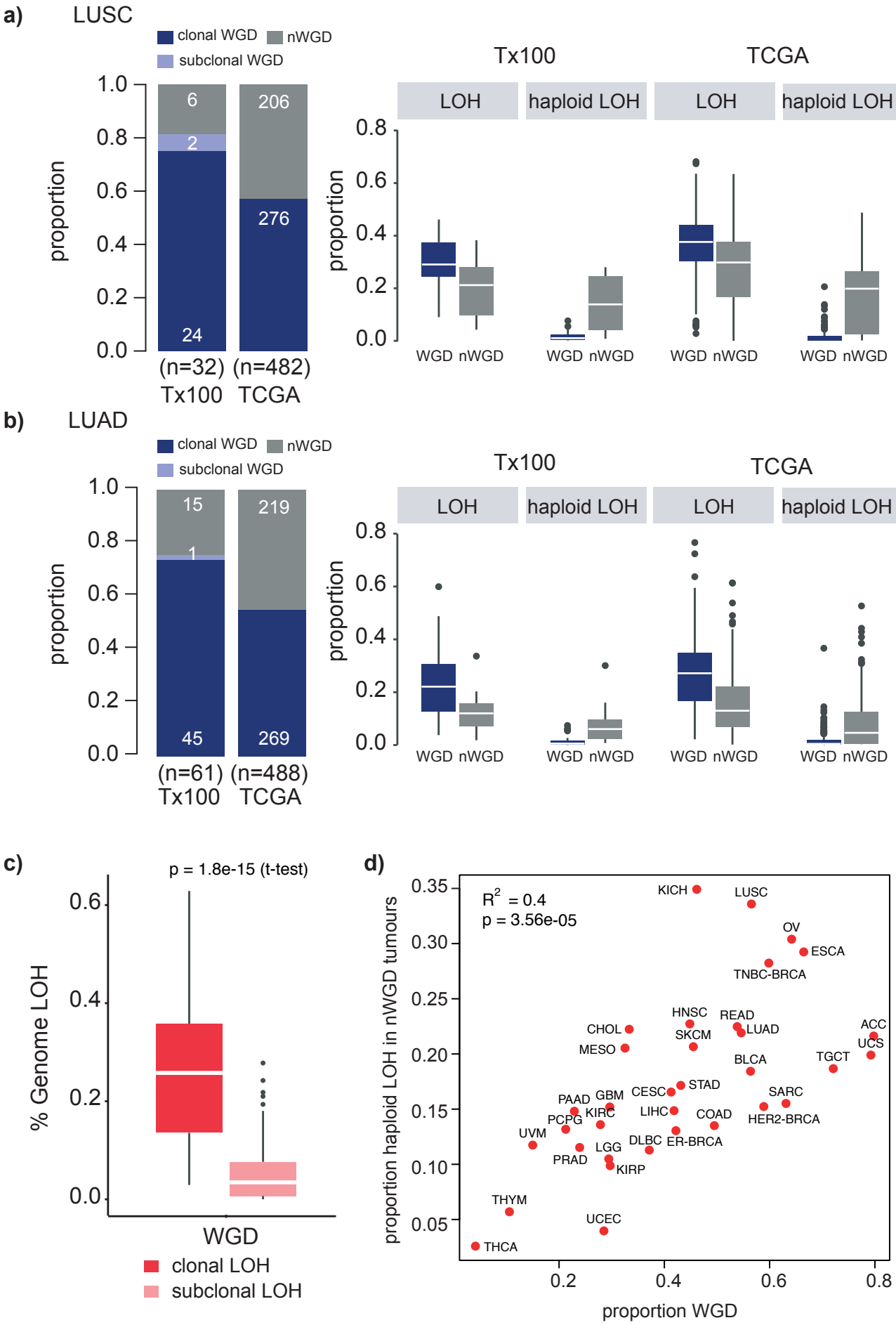
Charles Swanton (2,3,11), Mariam Jamal-Hanjani (2,11), Gareth Wilson (12), Nicolai J Birkbak (12), Thomas B K Watkins (12), Nicholas McGranahan (1,2), Christopher Abbosh (13), Rachel Rosenthal (12), Yin Wu (13), Mickael Escudero (12), Aengus Stewart (12), Andrew Rowan (12), Jacki Goldman (12), Selvaraju Veeriah (13), Marcin Skrzypski (13), Robert E Hynds (13), Andy Georgiou (13), Mariana Werner Sunderland (13), James Reading (13), Sergio Quezada (13), Karl Peggs (13), Teresa Marafioti (13), Peter Van Loo (12), John A Hartley (13), Richard Kevin Stone (12), Tamara Denner (12), Emma Nye (12), Sophia Ward (12), Emilia Lim (12), Stefan Boeing (12), Maria Greco (12), Maise Al-Bakir (12), Kevin Litchfield (12), Pat Gorman (13), Helen L Lowe (13), Leah Ensell (13), Victoria Spanswick (13), Angeliki Karamani (13), David Moore (13), Dhruva Biswas (13), Maryam Razaq (13), Jerome Nicod (12), Stephan Beck (13), Ariana Huebner (13), Michelle Dietzen (13), Cristina Naceur-Lombardelli (13), Mita Afroza Akther (13), Haoran Zhai (13), Nnennaya Kannu (13), Elizabeth Manzano (13), Clare Puttick (12), Katey Enfield (12), Emma Colliver (12), Brittany Campbell (12), Supreet Kaur Bola (13), Ehsan Ghorani (13), Marc Robert De Massy (13), Elena Hoxha (13), Emine Hatipoglu (13), Stephanie Ogwuru (13), Benny Chain (13), Jason Lester (14), Fiona Morgan (15), Malgorzata Kornaszewska (15), Richard Attanoos (15), Haydn Adams (15), Helen Davies (15), Dean Fennell (16), John Le Quesne (17), Apostolos Nakas (18), Sridhar Rathinam (18), William Monteiro (19), Hilary Marshall (19), Louise Nelson (18), Kim Ryanna (18), Alan Dawson (20), Mohamad Tuffail (18), Amrita Bajaj (18), Jan Brozik (18), Mark R Lovett (20), Jacqui A Shaw (21), Joan Riley (17), Lindsay Primrose (17), Luke Martinson (17), Nicolas Carey (17), Girija Anand (22), Sajid Khan (22), Gillian Price (23), Marianne Nicolson (23), Keith Kerr (23), Shirley Palmer (23), Hardy Remmen (23), Joy Miller (23), Keith Buchan (23), Mahendran Chetty (23), Lesley Gomersall (23), Sara Lock (24), Kayleigh Gilbert (24), Babu Naidu (25), Gerald Langman (25), Andrew Robinson (25), Hollie Bancroft (25), Amy Kerr (25), Salma Kadiri (25), Charlotte Ferris (25), Gary Middleton (25), Madava Djearaman (25), Akshay Patel (25), Yvonne Summers (26), Raffaele Califano (26), Paul Taylor (26), Rajesh Shah (26), Piotr Krysiak (26), Kendadai Rammohan (26), Eustace Fontaine (26), Richard Booton (26), Matthew Evison (26), Phil Crosbie (26), Stuart Moss (26), Juliette Novasio (26), Leena Joseph (26), Paul Bishop (26), Anshuman Chaturvedi (26), Helen Doran (26), Felice Granato (26), Vijay Joshi (26), Angela Ieek (27),

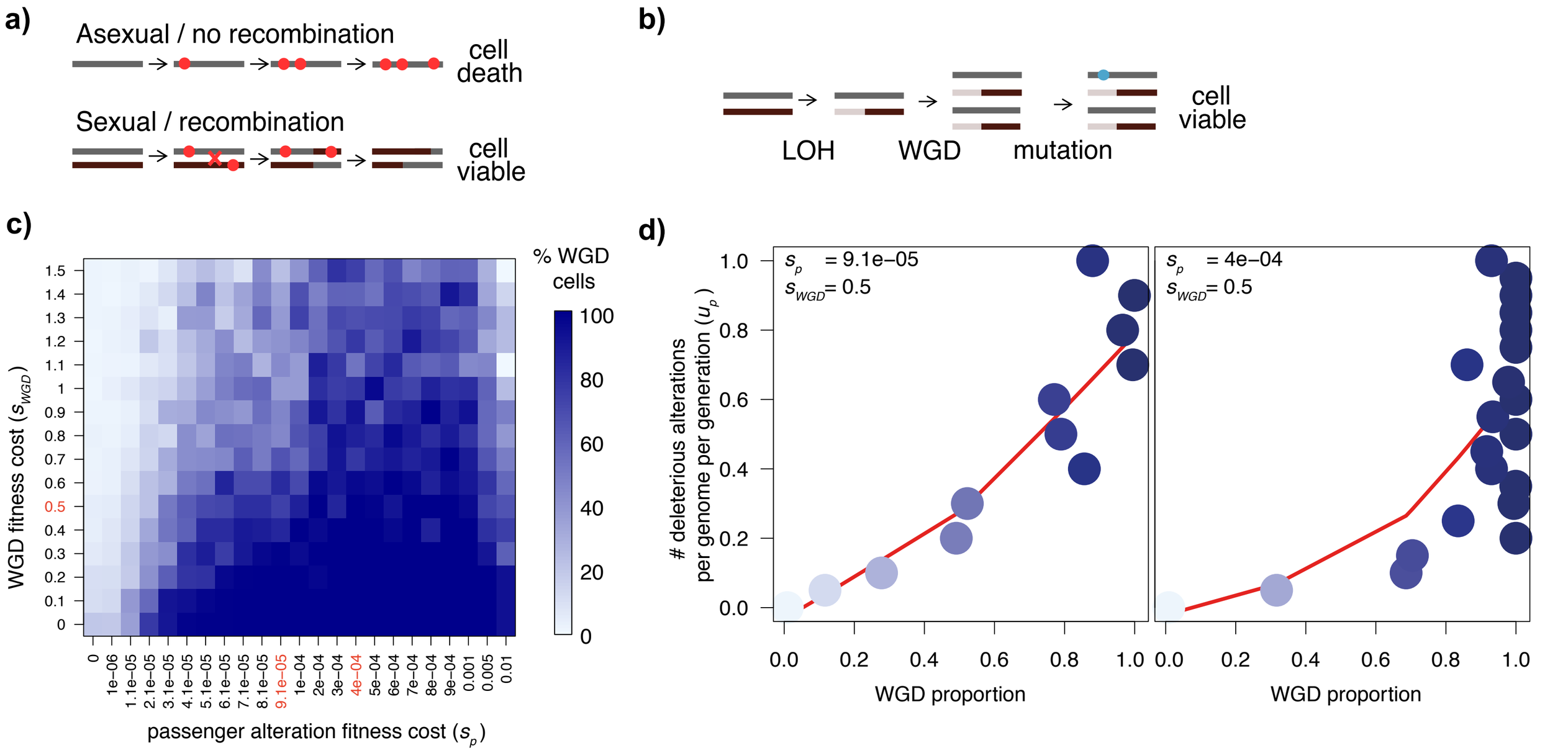
Phil Harrison (27), Katrina Moore (27), Rachael Waddington (27), Fiona Blackhall (28), Jane Rogan (27), Elaine Smith (26), Katie Baker (28), Mathew Carter (28), Lynsey Priest (28), Matthew G Krebs (28), Lindsay CR (28), Fabio Gomes (28), Angeles Montero (26), Caroline Dive (29), Ged Brady (29), Dominic G Rothwell (29), Francesca Chemi (29), Jonathan Tugwood (29), Jackie Pierce (29), David Lawrence (30), Martin Hayward (30), Nikolaos Panagiotopoulos (30), Robert George (30), Davide Patrini (30), Mary Falzon (30), Elaine Borg (30), Reena Khiroya (30), Crispin Hiley (12,31), Asia Ahmed (30), Magali Taylor (30), Junaid Choudhary (30), Penny Shaw (30), Sam M Janes (30), Martin Forster (30), Tanya Ahmad (30), Siow Ming Lee (30), Javier Herrero (30), Dawn Carnell (30), Ruheena Mendes (30), Jeremy George (30), Neal Navani (30), Dionysis Papadatos-Pastos (30), Marco Scarci (30), Roberto Salgado (32), Elisa Bertoja (30), Robert CM Stephens (30), Emilie Martinoni Hoogenboom (30), James W Holding (30), Steve Bandula (30), Hugo Aerts (33,34), Roland Schwarz (35), Zoltan Szallasi (36), Istvan Csabai (37), Miklos Diossy (37), Mairead MacKenzie (38), Maggie Wilcox (38), Allan Hackshaw (39), Yenting Ngai (39), Abigail Sharp (39), Cristina Rodrigues (39), Oliver Pressey (39), Sean Smith (39), Nicole Gower (39), Harjot Dhanda (39), Christian Ottensmeier (40), Serena Chee (40), Benjamin Johnson (40), Aiman Alzetani (40), Emily Shaw (40), Eric Lim (41), Paulo De Sousa (41), Simon Jordan (41), Alexandra Rice (41), Hilgardt Raubenheimer (41), Harshil Bhayani (41), Morag Hamilton (41), Lyn Ambrose (41), Anand Devaraj (41), Hema Chavan (41), Sofina Begum (41), Aleksander Mani (41), Daniel Kaniu (41), Mpho Malima (41), Sarah Booth (41), Andrew G Nicholson (41), Nadia Fernandes (41), Jessica E Wallen (41), Pratibha Shah (41), Kelvin Lau (42), Michael Sheaff (42), Peter Schmid (42), Louise Lim (42), John Conibear (42), Veni Ezhil (43), Vineet Prakash (43), Sarah Danson (44), Jonathan Bury (44), John Edwards (44), Jennifer Hill (44), Sue Matthews (44), Yota Kitsanta (44), Jagan Rao (44), Sara Tenconi (44), Laura Socci (44), Kim Suvarna (44), Faith Kibutu (44), Patricia Fisher (44), Robin Young (44), Joann Barker (44), Michael Shackcloth (45), John Gosney (45), Sarah Feeney (45), Julius Asante-Siaw (45), Teresa Light (46), Tracey Horey (46), Dionysis Papadatos-Pastos (46), Peter Russell (46), Kevin G Blyth (47), Craig Dick (47), Andrew Kidd (47), Alan Kirk (48), Mo Asif (48), John Butler (48), Rocco Bilanca (48), Nikos Kostoulas (48)

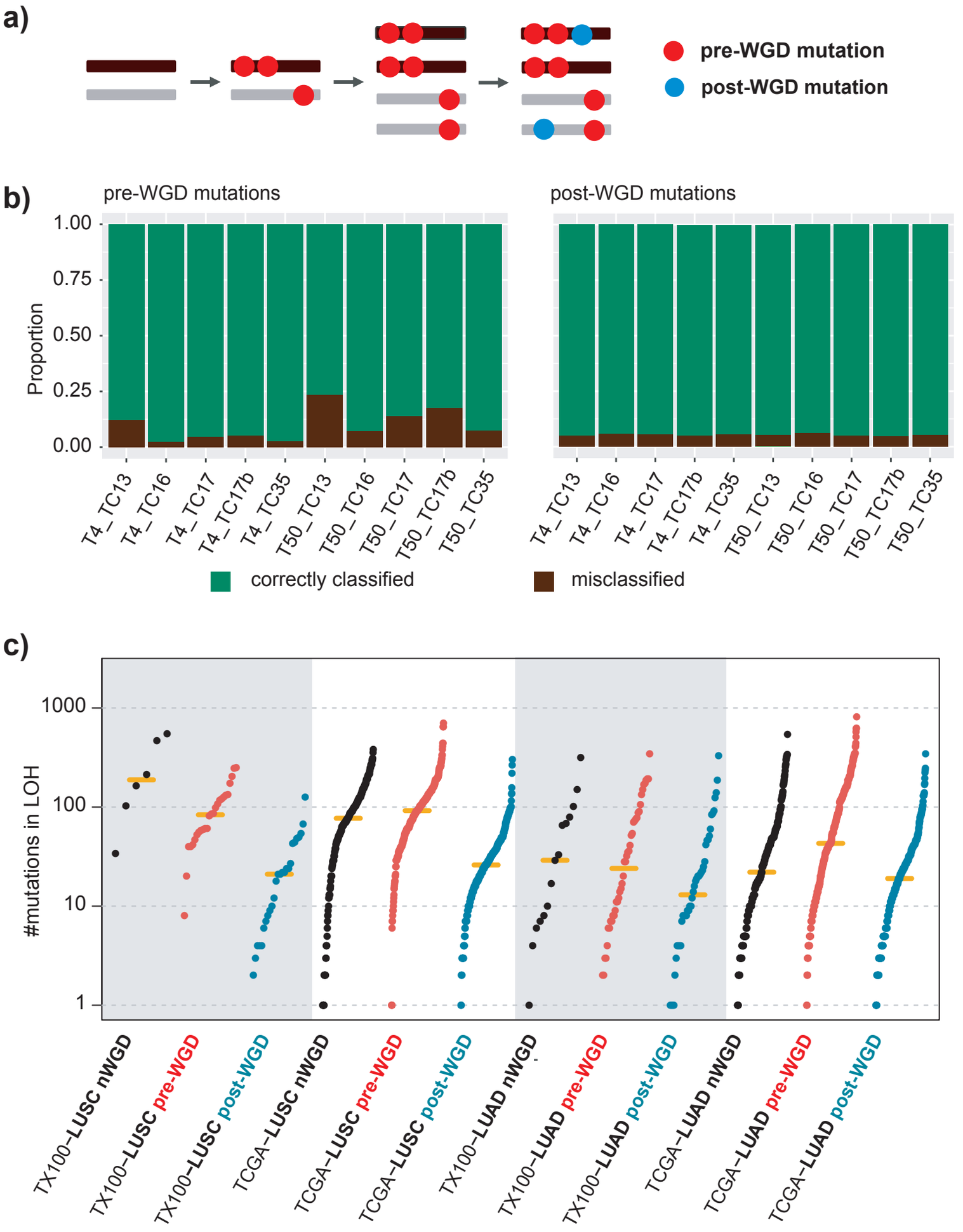
Affiliations:

(11) Department of Medical Oncology, University College London Hospitals, London, United Kingdom
(12) The Francis Crick Institute, London, United Kingdom
(13) University College London Cancer Institute, London, United Kingdom
(14) Velindre Cancer Centre, Cardiff, Wales
(15) Cardiff & Vale University Health Board, Cardiff, Wales
(16) Cancer Studies and Molecular Medicine, University of Leicester, Leicester, United Kingdom & Leicester University Hospitals, Leicester, United Kingdom
(17) Cancer Studies and Molecular Medicine, University of Leicester, Leicester, United Kingdom
(18) Leicester University Hospitals, Leicester, United Kingdom
(19) National Institute for Health Research Leicester Respiratory Biomedical Research Unit, Leicester, United Kingdom
(20) University of Leicester, Leicester, United Kingdom
(21) Cancer Studies and Molecular Medicine, University of Leicester, Leicester, United Kingdom.
(22) Barnet Hospital, Barnet, United Kingdom
(23) Aberdeen Royal Infirmary, Aberdeen, United Kingdom
(24) The Whittington Hospital NHS Trust, London, United Kingdom
(25) University Hospital Birmingham NHS Foundation Trust, Birmingham, United Kingdom

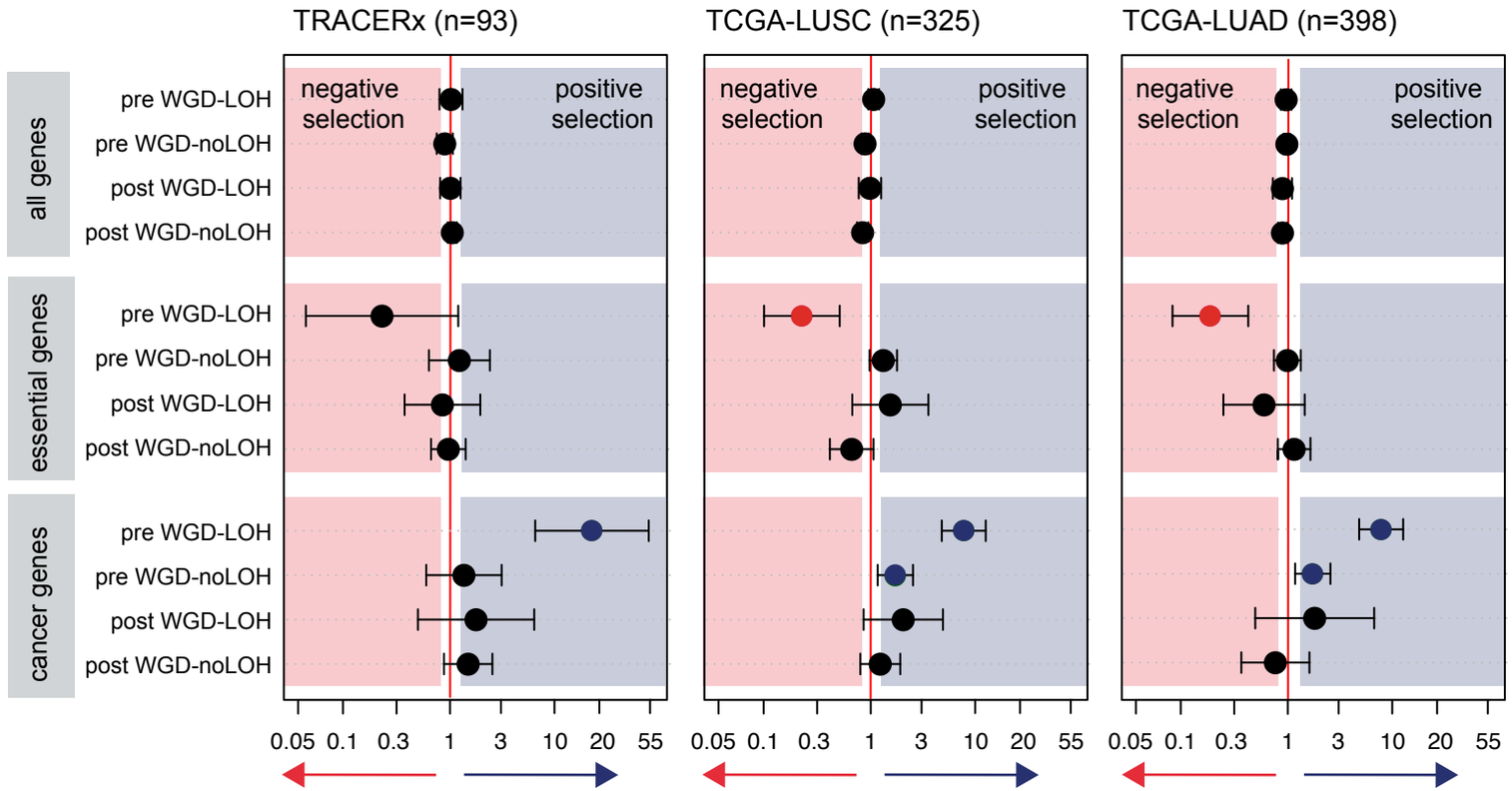
(26) University Hospital of South Manchester, Manchester, United Kingdom
(27) Manchester Cancer Research Centre Biobank, Manchester, United Kingdom
(28) Christie NHS Foundation Trust, Manchester, United Kingdom
(29) Cancer Research UK Manchester Institute
(30) University College London Hospitals, London, United Kingdom
(31) Cancer Research UK Lung Cancer Centre of Excellence, UCL Cancer Institute, London
(32) Department of Pathology, GZA-ZNA Antwerp, Belgium
(33) Departments of Radiation Oncology and Radiology, Dana Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
(34) Department of Radiology, Netherlands Cancer Institute, Amsterdam, The Netherlands.
(35) Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, Berlin, Germany
(36) Danish Cancer Society Research Center, Denmark
(37) Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest 1117, Hungary
(38) Independent Cancer Patients Voice
(39) Cancer Research UK & UCL Cancer Trials Centre, London, United Kingdom
(40) University Hospital Southampton NHS Foundation Trust
(41) Royal Brompton and Harefield NHS Foundation Trust
(42) Barts Health NHS Trust
(43) Ashford and St. Peter's Hospitals NHS Foundation Trust
(44) Sheffield Teaching Hospitals NHS Foundation Trust
(45) Liverpool Heart and Chest Hospital NHS Foundation Trust
(46) The Princess Alexandra Hospital NHS Trust
(47) NHS Greater Glasgow and Clyde
(48) Golden Jubilee National Hospital


a)
LUSC
clonal WGD
nWGD
subclonal WGD
proportion
6
2
24
206
276
(n=32)
(n=482)
Tx100
TCGA
Tx100
TCGA
LOH
haploid LOH
LOH
haploid LOH
WGD nWGD
b)
LUAD
15
1
45
219
269
(n=61)
(n=488)
c)
p = 1.8e-15 (t-test)
% Genome LOH
WGD
clonal LOH
subclonal LOH
d)
R^2 = 0.4
p = 3.56e-05
proportion haploid LOH in nWGD tumours
proportion WGD
KICH
LUSC
OV
ESCA
TNBC-BRCA
HNSC
READ
LUAD
CHOL
SKCM
MESO
ACC
UCS
BLCA
TGCT
STAD
CESC
PAAD
GBM
KIRC
LIHC
COAD
SARC
HER2-BRCA
PCPG
UVM
DLBC
ER-BRCA
LGG
PRAD
KIRP
THYM
UCEC
THCA

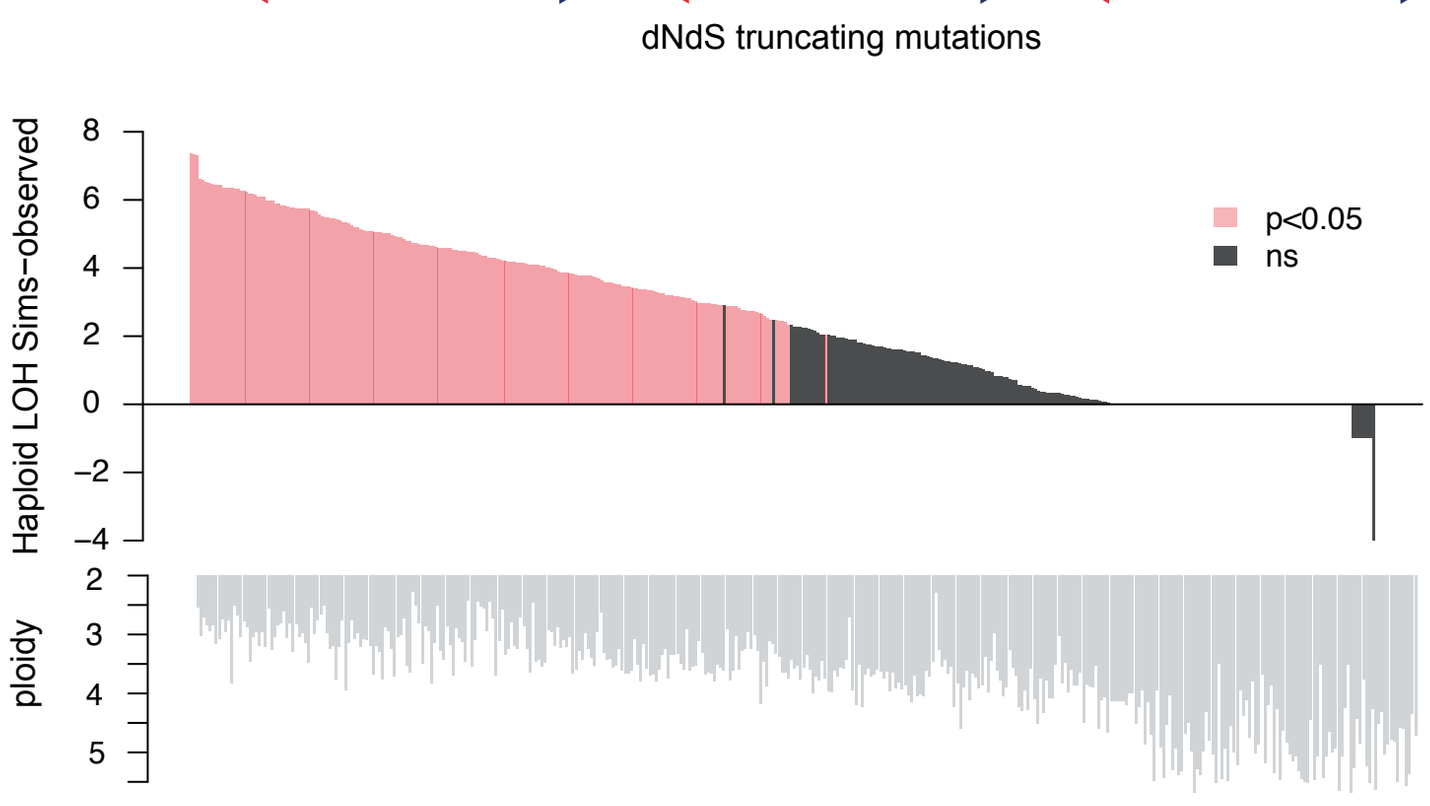

a)
Asexual / no recombination
cell death
Sexual / recombination
cell viable
b)
LOH
WGD
mutation
cell viable
c)
% WGD cells
100
80
60
40
20
0
WGD fitness cost (s_{WGD})
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1e−06
1.1e−05
2.1e−05
3.1e−05
4.1e−05
5.1e−05
6.1e−05
7.1e−05
8.1e−05
9.1e−05
1e−04
2e−04
3e−04
4e−04
5e−04
6e−04
7e−04
8e−04
9e−04
0.001
0.005
0.01
passenger alteration fitness cost (s_p)
d)
s_p = 9.1e−05
s_{WGD}= 0.5
s_p = 4e−04
s_{WGD}= 0.5
deleterious alterations per genome per generation (u_p)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
WGD proportion
WGD proportion


a)
pre-WGD mutation
post-WGD mutation
b)
pre-WGD mutations
post-WGD mutations
Proportion
1.00
0.75
0.50
0.25
0.00
T4_TC13
T4_TC16
T4_TC17
T4_TC17b
T4_TC35
T50_TC13
T50_TC16
T50_TC17
T50_TC17b
T50_TC35
correctly classified
misclassified
c)
#mutations in LOH
1000
100
10
1
TX100–LUSC nWGD
TX100–LUSC pre-WGD
TX100–LUSC post-WGD
TCGA–LUSC nWGD
TCGA–LUSC pre-WGD
TCGA–LUSC post-WGD
TX100–LUAD nWGD
TX100–LUAD pre-WGD
TX100–LUAD post-WGD
TCGA–LUAD nWGD
TCGA–LUAD pre-WGD
TCGA–LUAD post-WGD

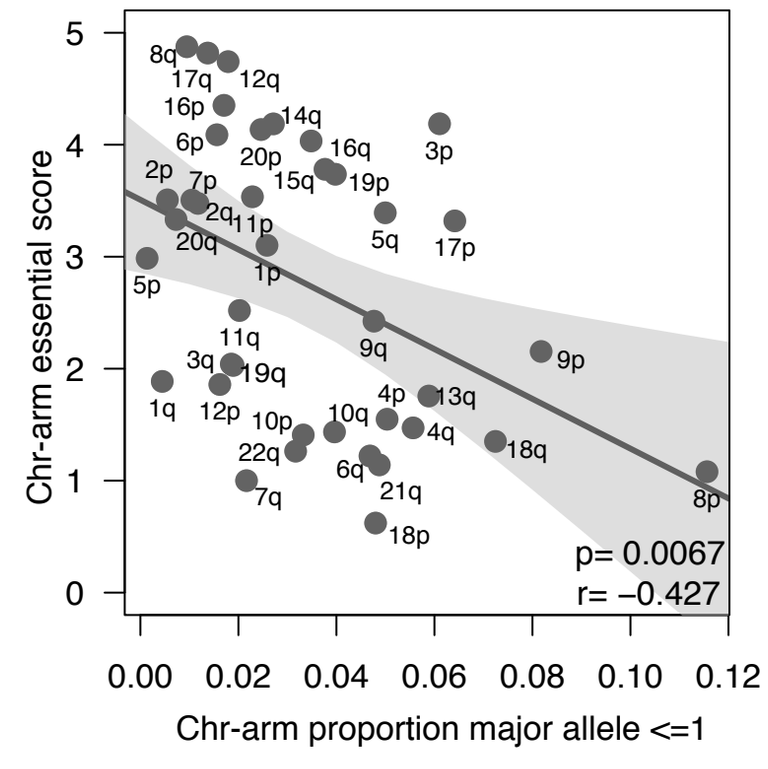
a)


TRACERx (n=93)
TCGA-LUSC (n=325)
TCGA-LUAD (n=398)
all genes
essential genes
cancer genes
pre WGD-LOH
pre WGD-noLOH
post WGD-LOH
post WGD-noLOH
negative selection
positive selection
0.05 0.1 0.3 1 3 10 20 55

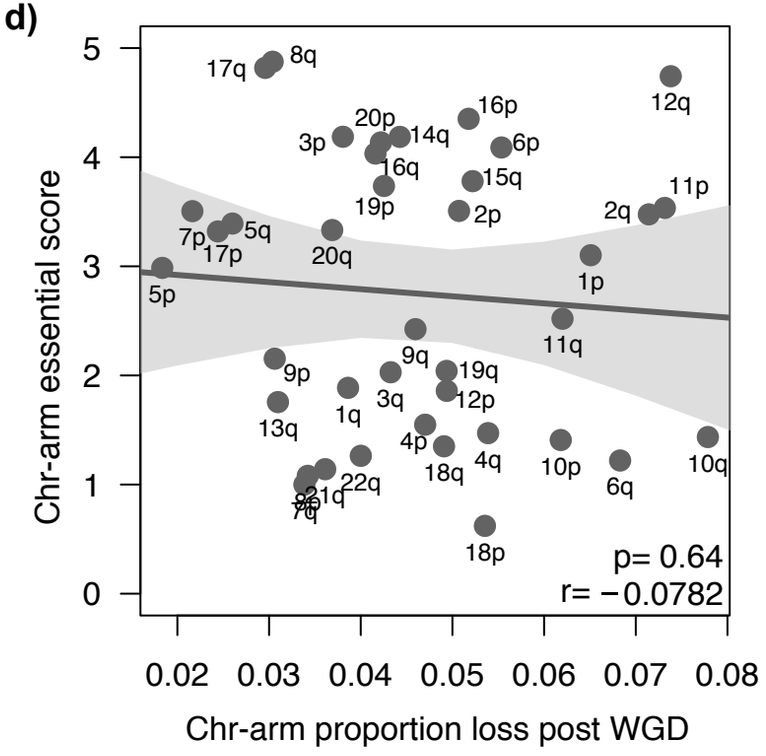
b)

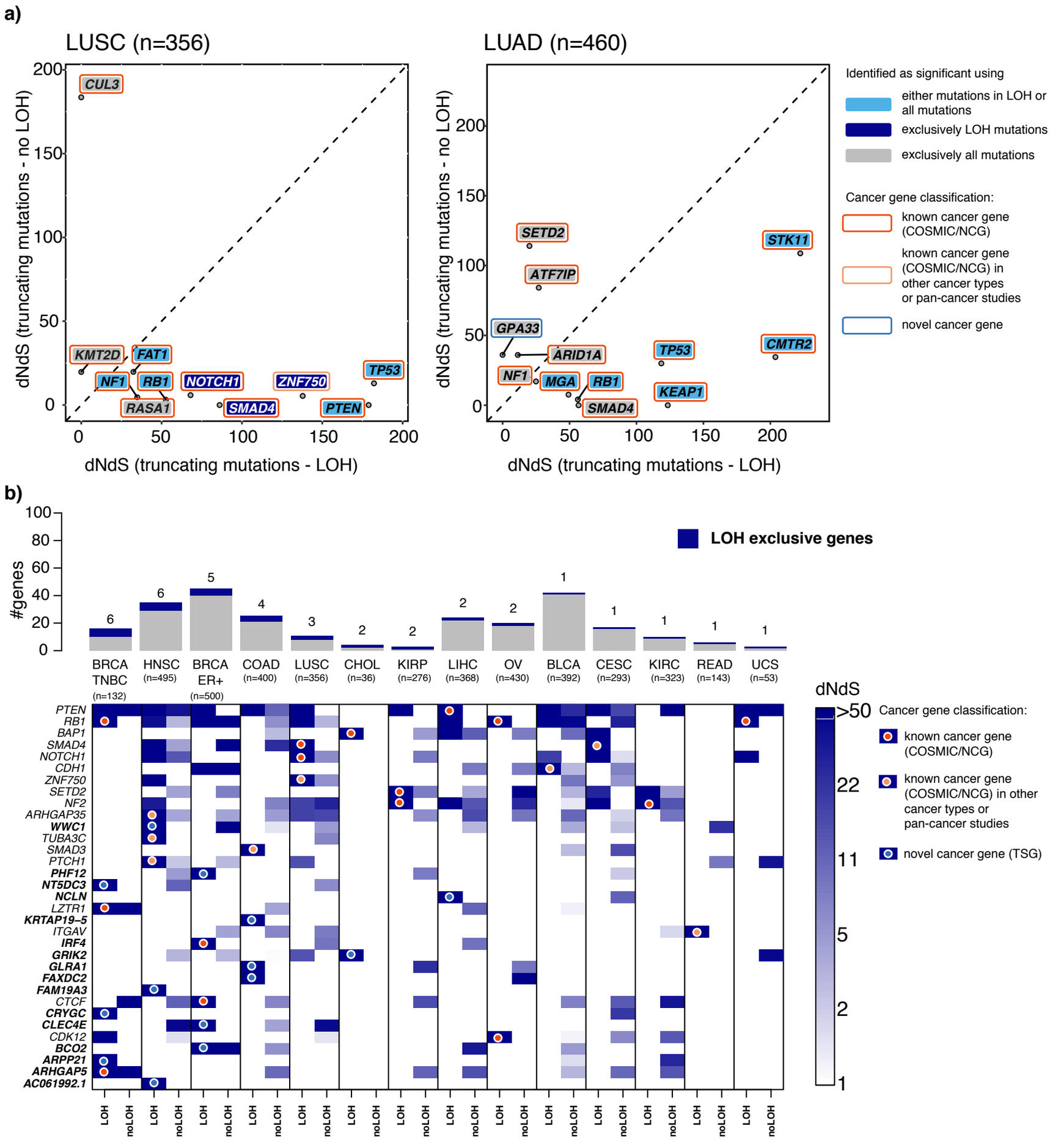

dNdS truncating mutations
Haploid LOH Sims−observed
ploidy
p<0.05
ns

c)


Chr-arm essential score
Chr-arm proportion major allele <=1
p= 0.0067
r= −0.427

d)


Chr-arm essential score
Chr-arm proportion loss post WGD
p= 0.64
r= −0.0782


a)
LUSC (n=356)
LUAD (n=460)
dNdS (truncating mutations - no LOH)
dNdS (truncating mutations - LOH)
CUL3
KMT2D
FAT1
NF1
RB1
NOTCH1
ZNF750
TP53
RASA1
SMAD4
PTEN
SETD2
ATF7IP
STK11
GPA33
ARID1A
TP53
CMTR2
NF1
MGA
RB1
KEAP1
SMAD4
Identified as significant using
either mutations in LOH or all mutations
exclusively LOH mutations
exclusively all mutations
Cancer gene classification:
known cancer gene (COSMIC/NCG)
known cancer gene (COSMIC/NCG) in other cancer types or pan-cancer studies
novel cancer gene
b)
LOH exclusive genes
#genes
BRCA TNBC (n=132)
HNSC (n=495)
BRCA ER+ (n=500)
COAD (n=400)
LUSC (n=356)
CHOL (n=36)
KIRP (n=276)
LIHC (n=368)
OV (n=430)
BLCA (n=392)
CESC (n=293)
KIRC (n=323)
READ (n=143)
UCS (n=53)
PTEN
RB1
BAP1
SMAD4
NOTCH1
CDH1
ZNF750
SETD2
NF2
ARHGAP35
WWC1
TUBA3C
SMAD3
PTCH1
PHF12
NT5DC3
NCLN
LZTR1
KRTAP19–5
ITGAV
IRF4
GRIK2
GLRA1
FAXDC2
FAM19A3
CTCF
CRYGC
CLEC4E
CDK12
BCO2
ARPP21
ARHGAP5
AC061992.1
LOH
noLOH
dNdS
>50
22
11
5
2
1
Cancer gene classification:
known cancer gene (COSMIC/NCG)
known cancer gene (COSMIC/NCG) in other cancer types or pan-cancer studies
novel cancer gene (TSG)