BROWN
UNIVERSITY

Explaining Patterns of Widespread Loss of Heterozygosity in Cancer

Ji Yun (Estelle) Han

Senior Honors Thesis in Computational Biology

Thesis Advisor: Dr. Ed Reznik

Second Reader: Dr. Sorin Istrail

ABSTRACT

Widespread genomic instability is a common feature in cancer, yet it is not fully understood what causes it and how cancer cells survive enormous damage to their genomes. Large scale loss of heterozygosity (LOH) across the genome in particular should in theory be deleterious to cells, yet we observe it to be a common feature in many cancer types. We sought to characterize the patterns of LOH across the genome in tumors with a high fraction of homozygous genome (FHG) as well as identify potential causes of extensive LOH. We found that even in high FHG tumors, certain regions of the genome, such as chromosome 7, tend to be exempt from LOH. We also found that somatic mutations in certain genes seem to play a role in driving the indiscriminate accretion of LOH across cancer types, in particular TP53, while others, such as mutations in KRAS, are negatively associated with widespread LOH in some major cancer types. We also found instances of disease-specific associations between mutation and high FHG, specifically MEN1 in pancreatic neuroendocrine tumor and PTEN in thyroid cancer. Finally, we found high FHG to be positively associated with clinical cancer stage in most major cancer types. This is similar to other forms of large scale copy number instability, such as whole genome doubling, and suggests large scale LOH may be a clinically relevant feature in many cancers.

Introduction

While somatic cells originally have two copies of every gene, one maternal and one paternal copy, cancer cells often lose one or more copies of segments of (or entire) chromosomes. Loss of heterozygosity (LOH) is when there is a loss of one of the two copies of a gene, and often occurs in conjunction with duplication of the remaining copy, called copy-neutral LOH (cnLOH). This is a very common genetic alteration in human cancers, yet there is little understanding as to why it occurs (Ryland, et al. 2015). In such cases, heterozygous somatic cells become homozygous and the cancer cell relies on the gene products encoded by a single allele. Widespread homozygosity is typically disadvantageous (Sellis, et al., 2011), and one would expect to have deleterious effects for cancer cells. This is because diploidy typically serves as a safeguard against deleterious mutations. Since many genes are haplosufficient, one copy can afford to be lost. However, when LOH occurs over a region, the cell is now susceptible to total loss of functionality (Szpiech, et al., 2013). For many genes - notably tumor suppressor genes (TSGs) - loss of function is beneficial for the cancer (Chial et al., 2008). However, LOH-affected regions in cancer often cover broad swaths of the genome, sometimes whole chromosomes, affecting useful or even essential genes as well (Nichols, et al., 2020). What causes or allows such large scale LOH to occur in some cancer cells, and what vulnerabilities it may expose in cancer genomes, are thus worth exploring. Prior research suggests that targeting LOH is a viable treatment strategy, so understanding patterns of LOH in cancer is potentially of clinical as well as scientific significance (Zhang, et al., 2021).

There are several motivating questions for this project. Does the fraction of homozygous genome (FHG) vary significantly across cancer types? Do high FHG tumors tend to constitute a distinct subtype within their respective cancer types? What (if any) regions are consistently 'exempt' from high FHG? What genes are mutated in conjunction specifically with high FHG? We further aim to further characterize LOH in high FHG patients - where it is enriched or depleted - across cancer types. Additionally, we explore associations between specific genes/mutations with patterns of LOH.

For this project, we used data from MSK-IMPACT, which has 47,759 patients and information on cancer type/subtype for each patient. We used mutation data on genes in the IMPACT panel to explore associations of certains mutations in cancer-associated genes with LOH. Copy number information was derived for each patient from the algorithm FACETS (Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing), which is an allele-specific copy number analysis (ASCN) (Shen and Seshan, 2016). We calculated the FHG, which is calculated by taking the fraction of the genome with LOH and weighting by the corresponding length of the arm. We first characterized the distribution of FHG across cancer types, and sought to assess whether FHG tended to be bimodally distributed (and thus whether there tended to be a distinct 'high-FHG' subtype). Overall, there was usually no distinct separation between low and high FHG tumors.

We next looked at whether there was depletion or enrichment of LOH on each chromosome arm at the pan-cancer level, before assessing enrichment/depletion for major cancer types. We found that several regions of the genome were significantly depleted of LOH not only at the pan-cancer level, but consistently across cancer types. Specifically, chromosomes 7, 12, 20 and arm 5p are largely 'exempt' from LOH events. We then investigated whether mutations in specific genes were associated with high FHG; we found that mutations in TP53 are positively associated with high FHG across various cancer types while mutations in other genes are associated with high FHG in specific cancer types, notably MEN1 in pancreatic neuroendocrine tumor - MEN1 being a key driver of pancreatic neuroendocrine tumor - while EGFR, and KRAS are associated with LOH for specific cancer types, or associated with LOH on specific chromosome arms (He, et al., 2022).

Finally, motivated by the relationship found between copy number instability and how advanced a tumor is, we hypothesized that higher (clinical) stage tumors would tend to have significantly higher FHG (Bielski, et al., Nature Genetics, 2018). After assessing this question for major cancer types, we found that for most types, higher FHG did indeed correspond to more advanced cancer.

Methods

Patient Samples

Our data set consists of IMPACT targeted sequencing data of 47,759 samples across many different cancer types treated at Memorial Sloan Kettering Cancer Center. The panel consists of 468 genes considered to be cancer-associated genes.

Allele-Specific Copy Number Inference

We inferred total and allele-specific copy number using FACETS algorithm (v0.5.14) in a two-step approach (Bielski et al., Cancer Cell, 2018). In the first step, we ran FACETS in low-sensitivity mode (cval = 100) to estimate purity and identify the log-ratio associated with the diploid copy number state of the tumor. We applied quality control criteria to identify poor or suboptimal fits. These include fits with: (i) extensive homozygous deletions, (ii) inferred logR ratio for diploid state being too high or too low, (iii) extensive hypersegmentation, (iv) too high ploidy, (v) invalid purity estimates such as NA or FACETS default of 0.3, (vi) high fraction of the genome is estimated to be subclonal, (vii) high fraction of the genome where the integer copy number estimate is discordant with the allelic imbalance observed with variant allele log odds ratio. Samples failing these quality control criteria were manually reviewed and re-fit.

Cancer cell fraction (CCF) of somatic mutations was calculated as described previously (McGranahan et al., Sci. Transl. Med. 2015) from the observed allele frequency (obs_{AF}), expected number of mutant copies (exp_{mt_cn}) and the tumor purity (Φ) is calculated as:

$$\frac{obs_{AF} * (tcn * \Phi + (1 - \Phi) * 2)}{\Phi}$$

Mutations were called clonal if the CCF is greater than 0.8 or if the CCF is greater than 0.7 and the upper bound of the 95% confidence interval of the CCF estimate is greater than 0.9. Mutations on subclonal copy number segments were deemed as indeterminate. All other mutations were considered as subclonal.

Assessing FHG Across Cancer Types

To capture the amount of LOH for each sample, we calculated the fraction of the homozygous genome, or FHG. This was calculated by taking the fraction of the genome with LOH and weighting by the corresponding length of the chromosome arm. Then, in order to classify samples as high or low FHG, we used a weighted FHG cutoff of 0.6 such that samples with a weighted FHG greater than 0.6 (that is >60% of the genome is subject to LOH) were classified as high FHG and those with a weighted FHG lower than 0.6 were classified as low FHG. The cutoff of 0.6 was chosen based on prior observations about the typical distribution of FHG in cancer types with apparent bimodal distributions, such as thyroid cancer. It is worth noting that our findings based on comparison of low vs. high FHG tumors were robust to variations in the cutoff.

Widespread loss of heterozygosity is analogous to the phenomenon of whole genome doubling (WGD), which involves the duplication of a complete set of chromosomes (Quinton, et al., 2021). WGD typically occurs in late stage or recurrent tumors (Newcomb, et al., 2021; Lopez, et al., 2020). WGD-affected tumors also had distinct characteristics, being significantly associated both with important genomic lesions and with poor clinical outcome (Bielski, et al., Nature Genetics, 2018). Similarly, we wanted to determine if tumors with LOH across the majority of the genome appeared to constitute a distinct class of tumor. To do this, we assessed the distribution of LOH for all common cancer types (n¿200) and used the dip test to determine multimodality (Hartigan, et al., 1985).

Computing Arm-Level Enrichment/Depletion of LOH

Among patients with LOH occurring throughout the genome, we were interested in the distribution of LOH events across the chromosomes; that is, if it occurred at a similar rate all across the genome or if there were certain areas that were enriched or depleted of LOH. We compared the rate of LOH on each chromosome arm for pan-cancer for high FHG samples and low FHG samples; the significance of the difference in LOH on each arm was determined via the binomial test where p, the hypothesized probability of success, was taken to be the average weighted FHG value of the samples. We also created a heat map showing the LOH enrichment/depletion score for various cancer types and chromosome arms to help assess the distribution of LOH over chromosome arms. The LOH enrichment/depletion score was calculated by taking the fraction of samples with LOH on that arm in a given cancer type and dividing this by the average weighted FHG of that cancer type, akin to an odds ratio. Thus, a score greater than 1 indicates enrichment of LOH, which corresponds to a higher rate of LOH, while a score less than 1 indicates depletion of LOH.

Determining Associations between Somatic Mutations and FHG

In order to assess whether mutations in specific genes are associated with high FHG, we compared FHG between mutated and unmutated samples across cancer-associated genes, for each cancer type with sufficient samples. To determine if differences between the FHG distributions were statistically significant, we used the Wilcoxon test. To correct for multiple hypothesis testing, we used FDR correction.

Assessing Association between Somatic Mutations and Trans-LOH Events

We also sought to determine whether mutations in certain genes were associated with LOH events in other parts of the genome (which we call trans-LOH associations, since the LOH is not in the region of the genome where the mutated gene is located). To do this, we computed the rate of LOH for each chromosome arm and compared the frequency of LOH on each arm between samples that are mutated vs. unmutated for each of a set of commonly mutated TSGs and oncogenes. We used a Chi-squared test to determine if mutation status in each gene was predictive of whether a sample had LOH over each chromosome arm. We then corrected for the family-wise error rate using FDR correction.

Assessing the Relationship between FHG and Cancer Stage

Motivated by research showing that WGD is closely associated with advanced cancer stage, we sought to determine if this was also true of widespread LOH. For each cancer type with at least 500 patients (21 cancer types in total; we used a higher threshold for this analysis because many patients did not have staging information), we determined if there was a significant relationship between patients' clinical cancer stage and FHG using Spearman's Rho test.

Results

FHG Distribution Across Cancer Types

The distribution of FHG varied greatly across cancer types (figure 1a). We looked at cancer types with at least 200 samples to check for bimodality and found that the distribution of FHG typically wasn't bimodal (figure 1b). Ultimately, for most cancer types, there was no distinct separation between high and low FHG groups. We found that, according to the dip test, several cancer types had multimodal FHG distributions, such as thyroid cancer. In some cases, including thyroid cancer, this multimodality seems to be attributable to different subtypes having different distributions (in this case, Hurthle Cell Carcinoma tends to have comparatively high FHG).

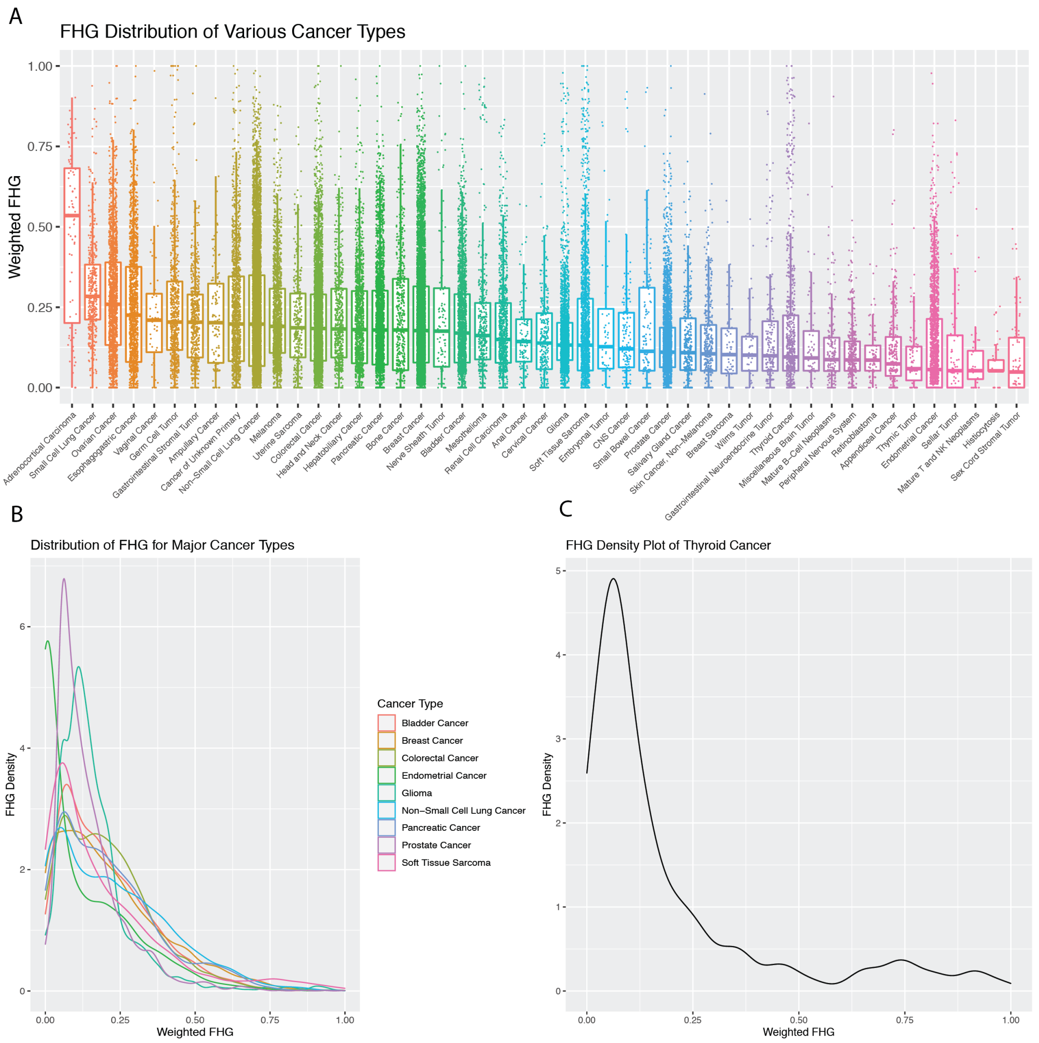


Figure 1: FHG distribution across cancer types. A) Box plots of FHG across cancer types. B) Density plots of FHG by cancer type. C) Example of bimodally distributed FHG: Thyroid cancer.

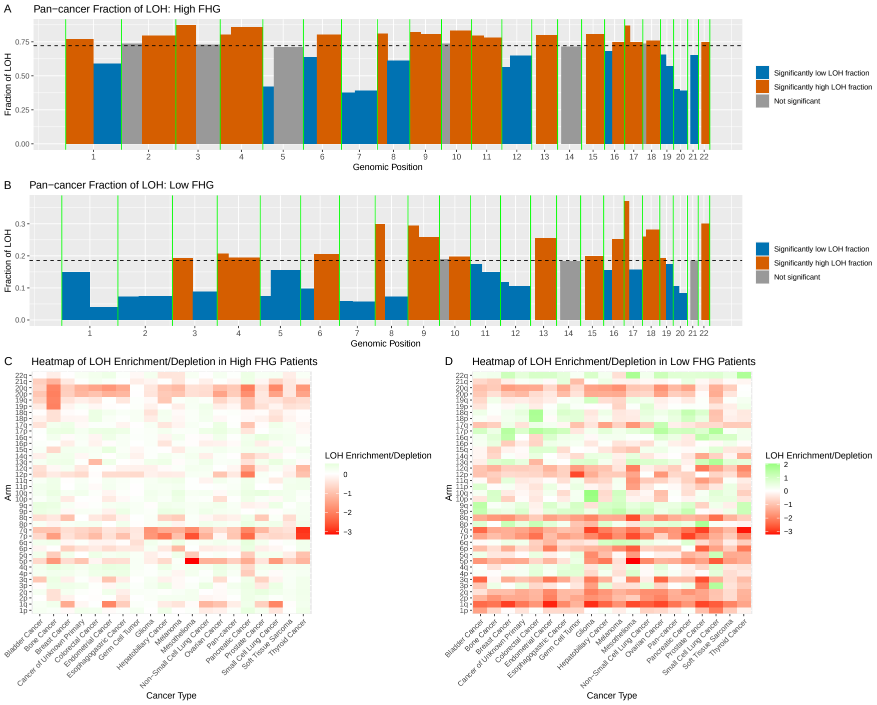


Figure 2: Depletion/enrichment for LOH across the genome in high vs. loh FHG tumors A) LOH fraction in each chromosome arm for high FHG tumors (pan-cancer) B) and for low FHG tumors. C) Heatmap showing LOH enrichment/depletion across chromosome arm and cancer type for high FHG D) and low FHG tumors.

Depletion of LOH at the Chromosome Level

We found that several chromosomes (especially 7, 12, 20 and 5p) show significant depletion of LOH (figure 2A and 2B) compared with the rest of the genome. High FHG tumors were particularly illuminating, as they possess large scale LOH events across the genome, allowing us to discern what regions of the genome appear to be ‘exempt’ from LOH. We also found that some regions are enriched for LOH, such as 17p, but the most pronounced effects observed were the region-specific depletion of LOH, with chromosome 7 being the most LOH-depleted region (the pan-cancer log LOH enrichment score for 7p was -0.94 and for 7q was -0.88).

We looked at the LOH depletion of chromosomes 7, 12, 20, and 5p across various cancer types to see if the effect might be due to some disease-specific phenomenon. We found that the pattern of LOH depletion (that is, less LOH occurring than across the genome on average) in these chromosomes was consistent across cancer types, indicating this depletion is lineage-agnostic (figure 2C and 2D).

Mutations Associated with Genome-Wide LOH

We found mutations in certain genes to be associated with higher FHG in specific diseases. Genes and diseases for which mutation status is associated with FHG at a statistically significant level (using FDR correction for family-wise error) are shown in figure 3a. TP53 mutation was associated with FHG across multiple cancer types, but otherwise, the effects tended to be highly disease specific. Mutation in MEN1, for example, was associated with significantly higher FHG in pancreatic cancer (figure 3b; p = 5.91e-69; log odds ratio: 13.00). We specifically found that the subtype in which MEN1 was associated with FHG was pancreatic neuroendocrine tumor. However, in other cancer types, MEN1 mutation was not associated with FHG. We also found that thyroid cancer patients with mutations in PTEN had significantly higher FHG (figure 3c; p = 1.65e-34, log odds ratio: 19.70).

We also found that mutations in certain genes were significantly associated with LOH in specific regions. In some cases, this is caused by allelic imbalance targeted to the region where the mutated gene is located (e.g. KRAS mutations are often accompanied by LOH on chromosome 12) (Bielski, et al., Cancer Cell, 2018). However, we also found that mutations in some genes tend to accompany LOH in other parts of the genome (figure 4).

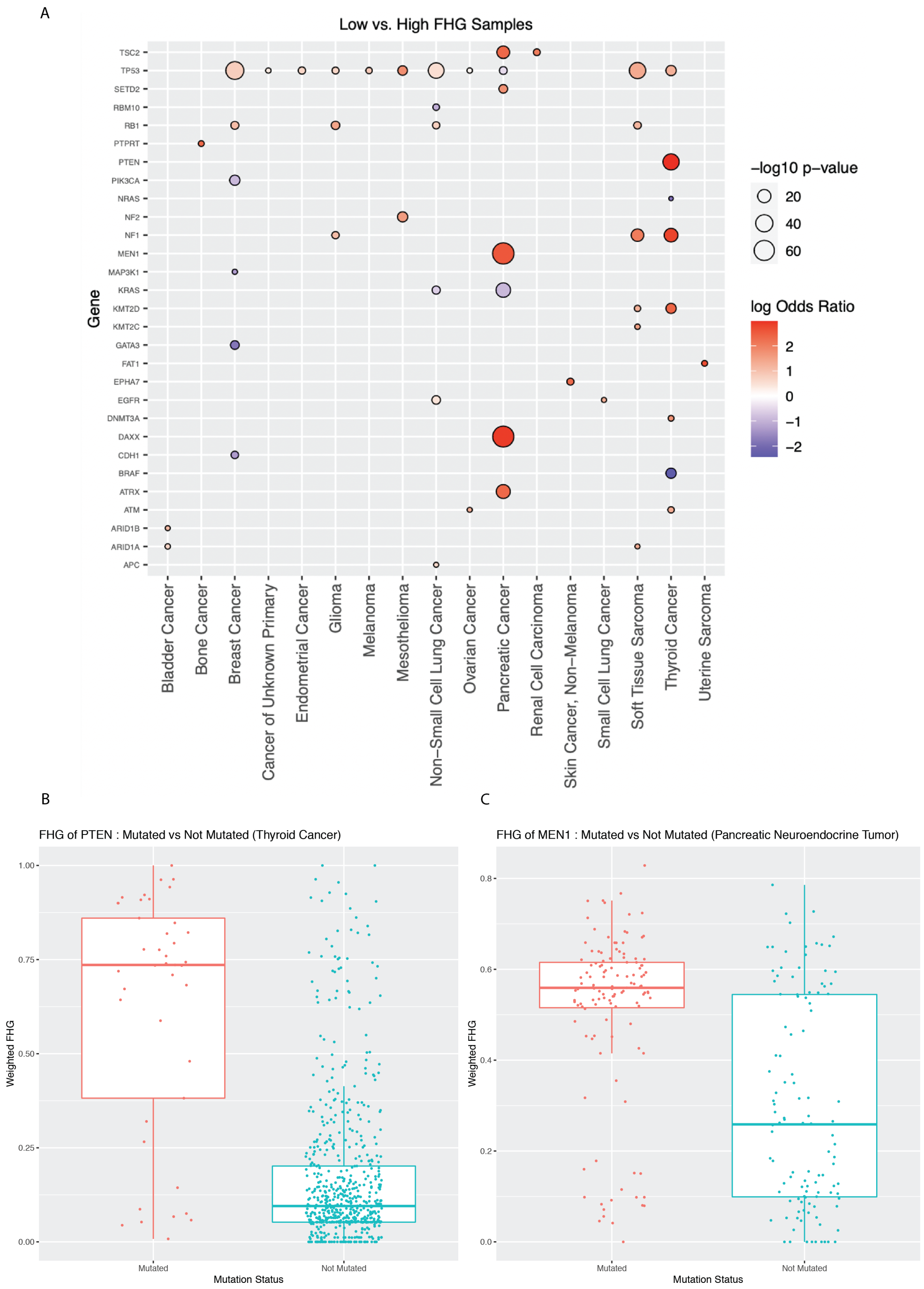
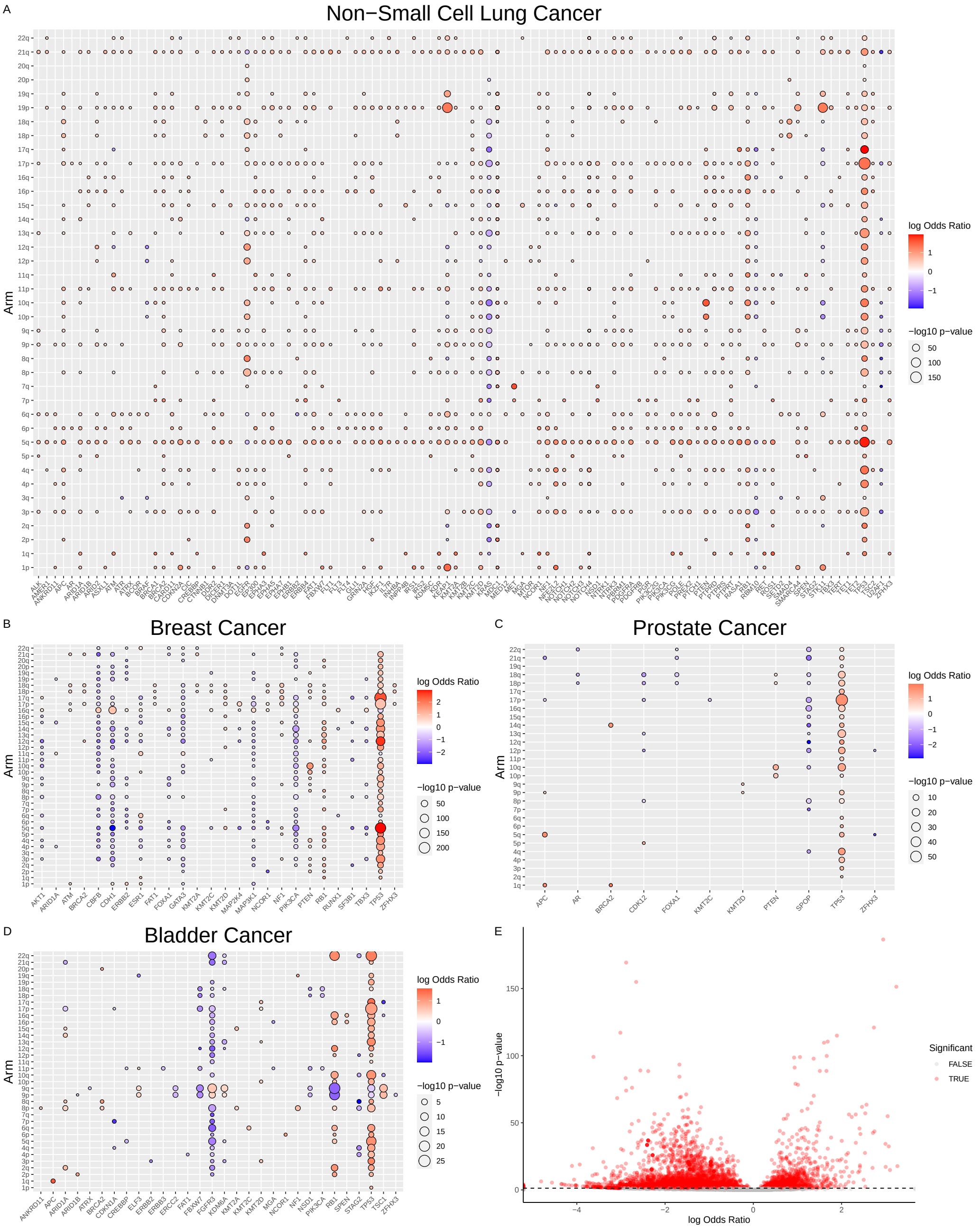


Figure 3: Genes associated with LOH in specific cancer types. A) Heatmap showing magnitude (log odds ratio of mutation in a given gene and cancer type in high vs. low FHG tumors) and significance of association (using Chi-squared test) between mutation in specific genes and high FHG status across major cancer types. B) Relationship between FHG and MEN1 mutation status in pancreatic neuroendocrine tumor. C) PTEN mutation status and FHG in thyroid cancer.

Mutations Associated with Trans-LOH Events

We then assessed whether mutations in specific genes were associated with LOH events in other parts of the genome (which we call trans-LOH associations). We found that mutations in key TSGs are strongly associated with LOH in other parts of the genome, usually across the entire genome. Mutation of TP53, for example, was consistently positively associated with higher LOH across the entire genome and in all major cancer types (figure 4). We also found that mutation of KRAS was negatively associated with LOH across the genome in non-small cell lung cancer, in which KRAS is a key driver. We also found that FGFR3 is strongly negatively associated with LOH across most of the genome in bladder cancer, despite being positively associated with LOH on chromosome 9. Unsurprisingly, we found that many TSGs are positively associated with LOH on their own chromosomes (e.g., TP53, RB1, PTEN).

After assessing associations between specific genes' mutation status and LOH over each chromosome arm, we found that most genes' mutation status are significantly associated with LOH in other parts of the genome. Moreover, the preponderance of significant hits are negative associations between mutation status and LOH.



A

Non−Small Cell Lung Cancer

B

Breast Cancer

C

Prostate Cancer

D

Bladder Cancer

E

Figure 4: Genes with significant association (in terms of mutation status) with LOH in specific regions of the genome, by cancer type. In each heatmap, the size of the dot indicates the negative log10 of the p-value (the larger the dot, the more significant the association) and the color indicates the magnitude and direction of the association (the log odds ratio of LOH frequency on the arm in samples mutated vs. unmutated for associated gene). A) Non-Small Cell Lung Cancer. B) Breast Cancer. C) Prostate Cancer. D) Bladder Cancer. E) A volcano plot of associations between cancer gene mutation status and LOH at other chromosome arms, across genes/cancer types (note: all cancer types with 200 or more samples were included, not just the four shown here).

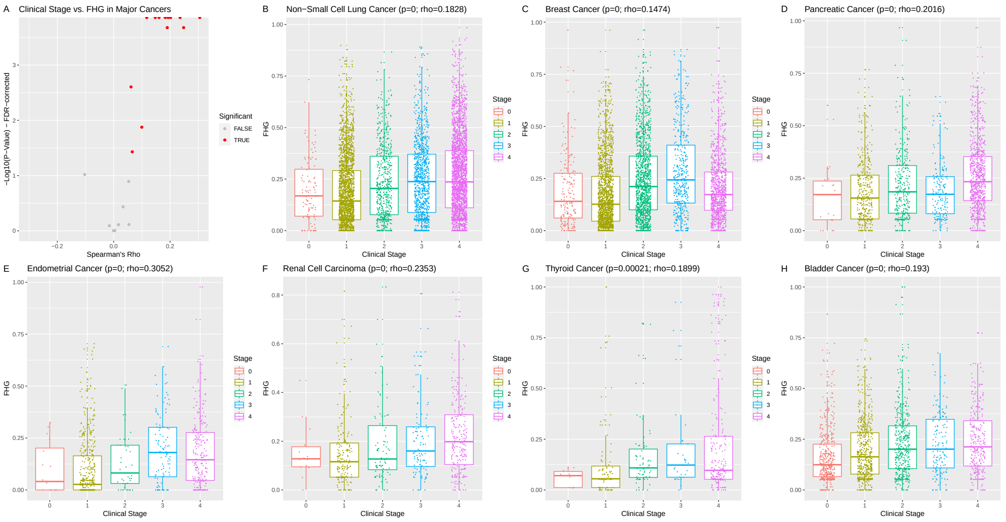


Figure 5: Relationship between cancer stage and FHG. A) A volcano plot showing the negative log10 p-value (FDR corrected) from Spearman's Rho test plotted against the correlation coefficient for each major cancer type. Box plots showing FHG vs. clinical stage are shown for several cancer types with a significant positive association between FHG and stage: B) non-small cell lung cancer; C) breast cancer; D) pancreatic cancer; E) endometrial cancer; F) renal cell carcinoma; G) thyroid cancer; H) bladder cancer.

Clinical Stage and FHG

Finally, we determined if FHG was associated with clinical cancer stage in all cancer types with at least 500 samples. We found that in the majority of these cancer types (12) there was a positive statistically significant relationship between stage and FHG (figure 5A). The remaining 9 cancer types had no significant relationship at all. The relationship between stage and FHG is illustrated for major cancer types in figure 5.

Discussion

We found that highly LOH-ridden tumors usually don't appear to constitute a distinct subtype, but rather FHG usually varies continuously without clear clusters, though some exceptional subtypes do show clear predisposition to acquire many LOH events, such as Hurthle cell carcinoma.

Focusing on high FHG tumors allowed us to find regions of the genome that seem exempt from LOH, which may reflect potential vulnerabilities in cancer that could be targeted in treatment. We found that chromosomes 7, 12, 20, and the p-arm of chromosome 5 are relatively highly 'exempt' from LOH. This proves to be interesting because areas exempt from LOH can tell us what genes are necessary for cancer or maintained by cancer. These areas are indicative of what genes the tumor has to keep intact in order to survive; losing these areas may actually be lethal to the tumors and thus suggestive that those regions could be targeted in therapy. There are several important cancer genes on chromosomes 7 and 12 that could potentially be relevant to this phenomenon, such as BRAF and EGFR on 7 and KRAS and POLE on 12 (POLE, notably, is considered to be an essential gene that cannot be homozygously lost) (Kesti, et al., 1999).

We next showed that mutations in specific genes are often associated with LOH in other parts of the genome. We found that TP53's positive effects on LOH are consistent across the genome and across cancer types. We also found that other genes, such as KRAS and FGFR4, are strong negative predictors of LOH across the genome. Overall, trans-LOH effects - LOH events associated with mutations in another region of the genome - appear to be a significant determinant of LOH, perhaps suggesting that the phenomenon of high-FHG tumors is related to genomic instability caused by somatic mutations.

Mutations in some genes seem associated with LOH over all as well in specific regions, and further research into why those genes predispose cancer cells to LOH might elucidate their function in cancer, e.g., why under some circumstances cancer cells can afford to lose much of their genome. LOH plays an important role in driving certain types of cancers such as Adrenocortical Carcinoma (Greenhill, 2016). Looking into what genes tend to drive this property/characteristic can help explain the genomics of that cancer. Mutation of TP53 is positively associated with high FHG across various cancer types. Although TP53 may have some broad effect across cancer types, there does not necessarily seem to be a singular cause of high FHG across all cancer types. Therefore, there may be idiosyncratic causes for specific cancer types. Additionally, mutations in other genes are associated with high FHG in specific cancer types: most notably, MEN1 in pancreatic neuroendocrine tumors, as well as PTEN in thyroid cancer. Widespread haploidy may be a consequence of lineage-specific relationships between specific driver mutations and genomic instability. Notably, the association between PTEN mutation and MEN1 mutation and FHG in thyroid and pancreatic neuroendocrine cancer, respectively, were not found in other cancer types, indicating specificity to those cancer types.

Finally, we found that cancer stage was positively associated with FHG in most major cancer types, as is the case with WGD. As high FHG tumors tend to be more advanced, the same continuing escalation of genomic instability reflected in WGD also appears to be reflected in large scale loss of heterozygosity. Since WGD has been found to be associated with clinical outcome, this finding suggests future research on the relationship between FHG and clinical outcome may be fruitful (Bielski, et al., Nature Genetics, 2018).

References

1. Bielski CM, Zehir A, Penson AV, Donoghue MTA, Chatila W, Armenia J, Chang MT, Schram AM, Jonsson P, Bandlamudi C, Razavi P, Iyer G, Robson ME, Stadler ZK, Schultz N, Baselga J, Solit DB, Hyman DM, Berger MF, Taylor BS. Genome doubling shapes the evolution and prognosis of advanced cancers. Nat Genet. 2018 Aug;50(8):1189-1195. doi: 10.1038/s41588-018-0165-1. Epub 2018 Jul 16. PMID: 30013179; PMCID: PMC6072608.

2. Bielski CM, Donoghue MTA, Gadiya M, Hanrahan AJ, Won HH, Chang MT, Jonsson P, Penson AV, Gorelick A, Harris C, Schram AM, Syed A, Zehir A, Chapman PB, Hyman DM, Solit DB, Shannon K, Chandarlapaty S, Berger MF, Taylor BS. Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. Cancer Cell. 2018 Nov 12;34(5):852-862.e4. doi: 10.1016/j.ccell.2018.10.003. Epub 2018 Nov 1. PMID: 30393068; PMCID: PMC6234065.

3. Chial H. Tumor suppressor (TS) genes and the two-hit hypothesis. Nature Education 2008, 1(1):177

4. Greenhill C. The genetics of adrenocortical carcinoma revealed. Nat Rev Endocrinol 12, 433 (2016). https://doi.org/10.1038/nrendo.2016.89

5. Hartigan JA and Hartigan PM. (1985). The Dip Test of Unimodality. The Annals of Statistics, 13(1), 70–84. http://www.jstor.org/stable/2241144

6. He L, Boulant S, Stanifer M, Guo C, Nießen A, Chen M, Felix K, Bergmann F, Strobel O, Schimmack S. The link between menin and

pleiotrophin in the tumor biology of pancreatic neuroendocrine neoplasms. Cancer Sci. 2022 Feb 18. doi: 10.1111/cas.15301. Epub ahead of print. PMID: 35179814.

7. Kesti T, Flick K, Keränen S, Syväoja J, Wittenberg C. DNA Polymerase Catalytic Domains Are Dispensable for DNA . Molecular Cell 199 May: 3(5), 679-685.

8. López S, Lim EL, Horswell S, et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nat Genet 52, 283–293 (2020). https://doi.org/10.1038/ s41588-020-0584-7

9. McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci Transl Med. 2015 Apr 15;7(283):283ra54. doi: 10.1126/scitranslmed.aaa1408. PMID: 25877892; PMCID: PMC4636056.

10. Newcomb R, Dean E, McKinney BJ, et al. Context-dependent effects of whole-genome duplication during mammary tumor recurrence. Sci Rep 11, 14932 (2021). https://doi.org/10.1038/s41598-021-94332-z

11. Nichols CA, Gibson WJ, Brown MS, et al. Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. Nat Commun 11, 2517 (2020). https://doi.org/10.1038/ s41467-020-16399-y

12. Quinton RJ, DiDomizio A, Vittoria MA, et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. Nature 590, 492–497 (2021). https://doi.org/10.1038/s41586-020-03133-3.

13. Ryland GL, Doyle MA, Goode D., et al. Loss of heterozygosity: what is it good for?.BMC Med Genomics 8, 45 (2015). https://doi.org/10.1186/s12920-015-0123-z

14. Sellis D, Callahan BJ, Petrov DA, Messer PW. Heterozygote advantage as a natural consequence of adaptation in diploids. Proc Natl Acad Sci U S A. 2011 Dec 20;108(51):20666-71. doi: 10.1073/pnas.1114573108. Epub 2011 Dec 5. PMID: 22143780; PMCID: PMC3251125.

15. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44(16):e131. doi:10.1093/nar/gkw520

16. Szpiech ZA, Xu J, Pemberton TJ, et al. Long runs of homozygosity are enriched for deleterious variation. Am J Hum Genet. 2013;93(1):90-102. doi:10.1016/j.ajhg.2013.05.003

17. Zhang X, Sjöblom T. Targeting Loss of Heterozygosity: A Novel Paradigm for Cancer Therapy. Pharmaceuticals (Basel). 2021;14(1):57. Published 2021 Jan 13. doi:10.3390/ph14010057.