Somatic copy number alterations reveal insight in drivers of resistance as a response to specific treatments

Nathalie A.M. Nijs*, Chris Williams, Arne van Hoeck[1] and Francisco Martinez Jimenez[1]

[1] Center for Molecular Medicine and Oncode Institue, University Medical Centre Utrecht, Heidelberglaan 100, 3584CX Utrecht, The Netherlands

Abstract

There is no cancer treatment that is 100% effective; during treatment of the tumor, drug resistance can arise, which is a big obstacle for treating patient with cancer successfully. The identification of drivers of resistance may help to improve future treatment decision making. Here, we aim to unbiasedly detect resistance mechanisms using tumor whole genome sequencing data from human cancer patients previously treated with standard of care anti-cancer drugs. For this, we analyzed somatic copy number alteration (SCNAs) from 2250 unique samples from Hartwig Medical Foundation (HMF), classified into 41 custom treatment groups. The control dataset contains of samples from HMF and the Pan-Cancer Analysis of Whole Genomes (PCAWG) study. We searched for genes that underwent positive selection after they were treated. By looking into the positive selected genes for different treatment groups, we reveal insight into the potentially drivers of resistance against cancer treatments. Known drivers of resistance were found, like androgen receptor (AR) for anti-androgen, EGFR for anti-EGFR and TYMS for pyrimidine antagonist. EDA2R was found amplified in taxanes and anti-androgen, this is a large candidate to be a driver of resistance against these treatments. MSL3, CASC9, CSPP1, PNMT and HLA-A were also found amplified in specific treatments, further research is needed to find out if they are drivers of resistance. Overall, our findings show some known resistance mechanisms as well as novel gene resistance candidates which require further validation. Nevertheless, this compendium is the first step to better understand resistance mechanisms that are activated by copy number alterations.

Layman's Summary

When you have cancer, you get treated with a drug, but this treatment is not always fully effective. There is a chance that you get resistant against this drug, with other words the drug won't work anymore. This resistance can occur when you get mutations in your DNA, here we focus on the copy number alterations (CNAs). CNAs are mutations that result in the gain (amplification) of loss (deletion) in copies of sections of DNA. When we know which genes play a role in these resistance mechanisms, the future treatment decision making, which drug the patient gets, may be improved. We used HMF and PCAWG datasets, which contains samples from treated and untreated patients. With this approach, we were able to detect already known CNAs that are linked with resistance, but we also find some potential resistance candidates that require further validation analysis.

Introduction

Although the survival rate has increased slightly due to the early detection, cancer is still the second leading cause of death (Ritchie & Roser, 2018). One of the reasons for that is that there is no cancer treatment that is 100% efficient since most cancer develop resistance mechanisms against the treatment and thus remains one of the biggest obstacles for treating patient with cancer successfully (Sun & Hu, 2018). The development of cancer drug resistance is based on treatment specific positive selection on resistant cells with a genetic mutant. It can also be the case that a treatment has positive selection on non-genetically resistant cells. When these cells multiply the chance of generating genetically mutant increases, which can lead to resistant cells [Fig. 1] (Melguizo et al., 2011).
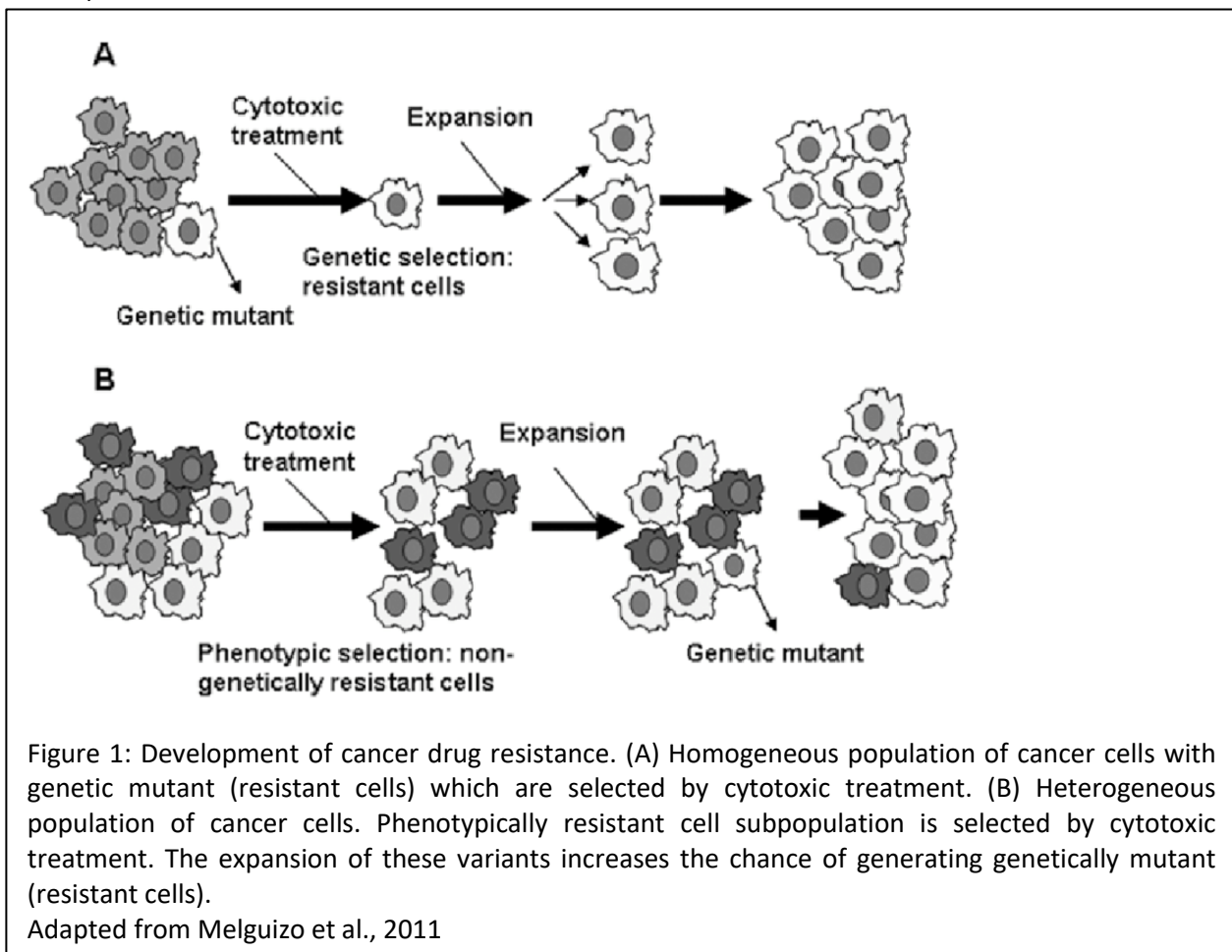

Figure 1: Development of cancer drug resistance. (A) Homogeneous population of cancer cells with genetic mutant (resistant cells) which are selected by cytotoxic treatment. (B) Heterogeneous population of cancer cells. Phenotypically resistant cell subpopulation is selected by cytotoxic treatment. The expansion of these variants increases the chance of generating genetically mutant (resistant cells).
Adapted from Melguizo et al., 2011

Within cancer development there are a lot of different genetic alterations (Futreal et al., 2004). These alterations ranges from nucleotide-level changes, such as single nucleotide variants (SNVs) and short insertions or deletions (indels), to large chromosomal events, like structural variations (SVs) (Greenman et al., 2007) (Mitelman et al., 2007) (Kim et al., 2013). Somatic single nucleotide variants (SNVs), also known as point mutations, are the simplest class of mutation and occur when a single nucleotide (A, T, C or G) is altered in the DNA sequence (Spencer et al., 2015). Somatic copy number alterations (SCNAs) are a large subtype of SVs and are changes to the chromosome structure, which results in gains or losses in the copy number of a segment of DNA (Beroukhim et al., 2010). The recurrence of several SCNAs is associated with particular cancer types, which makes it very interesting and important to study (Baudis, 2007) (Alkodsi et al., 2015). Parallel on this study, another

study focusing on the point mutations was done. However, in this study we will focus on somatic copy number alterations (SCNAs).
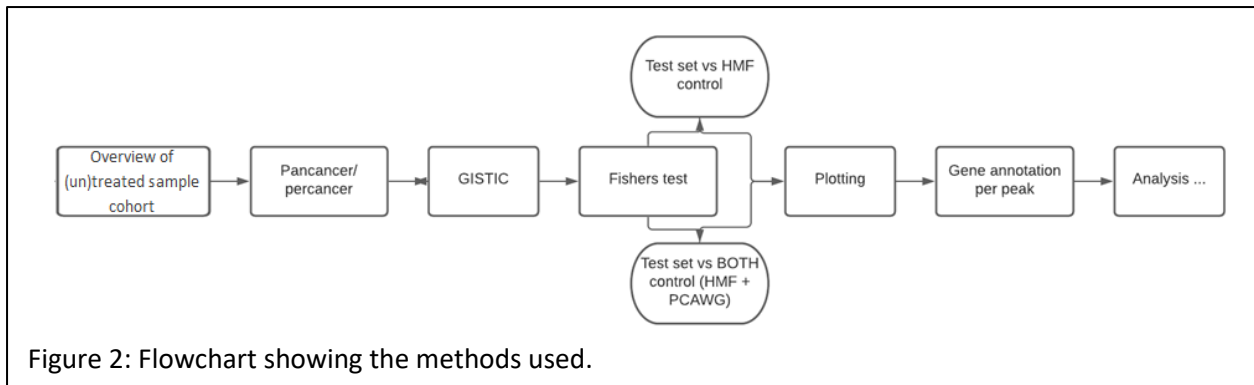
The epidermal growth factor receptor (EGFR) is a protein encoding gene and member of the HER family, which consists of four members; EGFR, HER2 (ErbB2), HER3 (ErbB3) and HER4 (ErbB4). EGFR is a transmembrane receptor tyrosine kinase (RTK) (Morgillo et al., 2016) (Zhao et al., 2016). A known cancer treatment is based on the inhibition of EGFR (Anti-EGFR), these drugs bind to the extracellular domain of EGFR. The binding of these drugs do not only prevent the activation of the RTK, but also the activation of various downstream signal transduction cascades, which are linked to proliferation, cell survival and metastasis (Li et al., 2009) (Wieduwilt & Moasser, 2008). Numerous studies have reported that overexpression of EGFR is linked to the resistance against anti-EGFR cancer treatment (Chong & Jänne, 2013) (Sequist et al., 2011) (Sigismund et al., 2018) (Moroni et al., 2005). Another known driver of resistance is androgen receptor (AR) for the cancer treatment anti-Androgen (Lallous et al., 2016) (Boudadi & Antonarakis, 2016). Androgen receptor is a ligand-dependent transcription factor which gets mostly activated by androgens and can regulate downstream signaling pathways that are dependent of androgen (Sakkiah et al., 2016). The anti-androgen therapies against prostate cancer initially form a positive response which leads to massive prostate cell death. However, later on the cancer cells appear as androgen-independent, i.e. resistant (Yang et al., 2005).

Cancer driver genes are genes that have mutations which play a significant role in cancer development and progression (Pham et al., 2021). There are genes that are mutated in most of the cancer types, while others are cancer type specific. (Tamborero et al., 2013) (Iranzo et al., 2018) (Bailey et al., 2018). For instance the gene CCND1, the amplification of this gene is associated with breast cancer, non-small cell lung cancer and head and neck squamous cell carcinomas initiation and progression (Zhang & Zhang, 2017) (Elsheikh et al., 2008) (Ortiz et al., 2017) (Shan et al., 2017). This gene encodes the protein cyclin D1, which forms a complex and regulate the activation of CDK4 or CDK6. This activity is required for the G1/S transition in the cell cycle. The amplification of CCND1 correlates with tumor differentiation, poor survival and increased metastasis (van Diest et al., 1997) (Shan et al., 2017). In order to find the drivers of resistance, the cancer driver genes need to be filtered out. Because there are also cancer type specific driver genes, the analysis in this study will be done on cancer type level (per cancer).

Thus far, an unbiased characterization of SCNA induced resistance mechanisms from human whole genome cancer data has never been conducted. By analyzing SCNA that underwent positive selection after treatment, we aim to identify genes that play a role in cancer treatment resistance. For this, we used HMF and PCAWG datasets that contains WGS data of treated and untreated cancer patients. With this approach, we were able to detect known alterations, but also some potential interesting resistance candidates that require further follow up validation analysis.

Results

To identify positive selected genes of resistance by copy number alterations we developed this approach shown in Figure 2.

Figure 2: Flowchart showing the methods used.

Overview of (un)treated sample cohort

To identify drivers of drug specific resistance alterations, the different drugs were classified based on their mechanism of action, based on literature and by their FDA labels. [Fig. 3] The drug-grouping used in this study has two layers, while the FDA labelling only has one. The extra layer consists of drugs that have a sufficient number of samples (>300) to do the analysis on drug level. Besides that this drug is also in a multiple drug treatment group based on their mechanism of action.

A dataset (test set) of samples from pretreated cancer patients was created using samples from the Hartwig Medical Foundation (HMF) and the drug classification [Fig. 4A] (Priestley et al., 2019). There are drug-groups that have only samples from one or two cancer types and drug-groups with a variety of cancer types. The test set contains 2240 unique samples and 41 unique custom treatment groups. The actual number of used samples is a lot higher because some samples are used more than once. For instance with the drugs that have both their own drug-group and a multiple-drug-group. [Table 1] Another dataset with samples from untreated cancer patients was created and used as the control set. This dataset contains samples from HMF and the Pan-Cancer Analysis of Whole Genomes (PCAWG) study [Fig. 4B] (Giunta, 2021) (Campbell et al., 2020). This control dataset contains approximately 2600 samples.

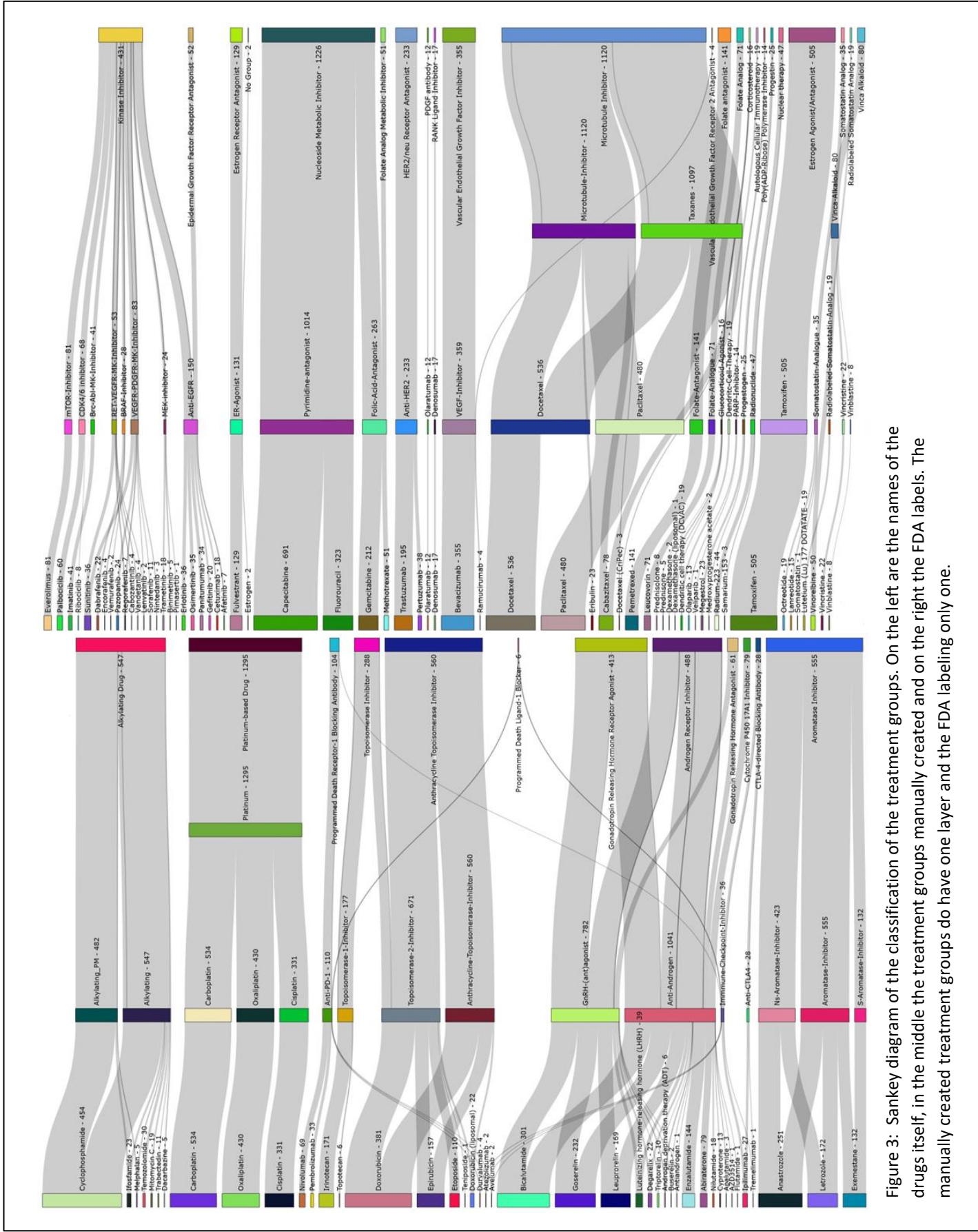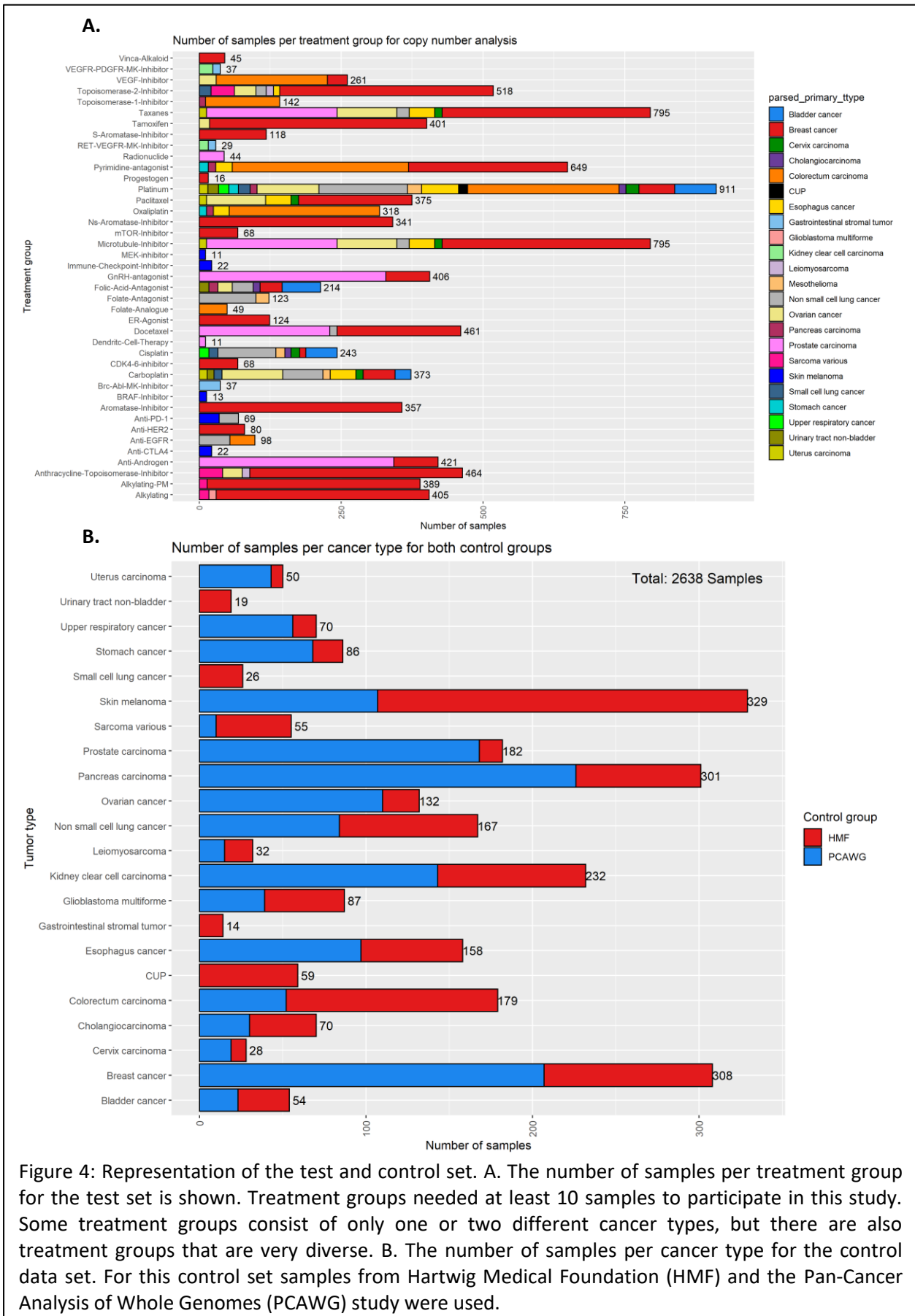|	
Number of unique samples	2240
Number of drugs	106
Number of custom groups	41
Number of FDA groups	36
Number of targeted based groups	6
Number of mechanism based groups	35
Table 1. Overview of the data used for the test set	

Figure 3: Sankey diagram of the classification of the treatment groups. On the left are the names of the drugs itself, in the middle the treatment groups manually created and on the right the FDA labels. The manually created treatment groups do have one layer and the FDA labeling only one.

Figure 4: Representation of the test and control set. A. The number of samples per treatment group for the test set is shown. Treatment groups needed at least 10 samples to participate in this study. Some treatment groups consist of only one or two different cancer types, but there are also treatment groups that are very diverse. B. The number of samples per cancer type for the control data set. For this control set samples from Hartwig Medical Foundation (HMF) and the Pan-Cancer Analysis of Whole Genomes (PCAWG) study were used.

GISTIC

GISTIC identifies positive selected amplified and deleted regions per cohort for every treatment group, this region will show up as a peak. Figure 5 shows the number of amplification and deletion peaks per treatment group on per cancer level. In general there are more deletion peaks compared to the amplification peaks per treatment group and some groups don't even have significant amplified regions. Every sample per amplified region gets annotated in a group; amplified or neutral, this was done using a threshold. If the difference in copies from the genome-wide median is above 0.9, the sample is amplified in that region, if it's below that it's marked as neutral. A lot of the same peaks were found across the results, this was the result of the use of the samples in multiple treatment groups. Besides that we also saw the same peaks for different treatment groups, but with the same cancer type, indicating cancer drivers of specific cancer types. As expected in breast cancer we do observe that basically every treatment group show a amplification peak in chromosome 11 which probably relates to CCND1, which is a known driver gene in breast cancer. We would expect that every peak only represents one gene, but in reality this wasn't true, there were on average 23.28 genes per amplification peak, with a range from 1 to 1554 genes. For instance, in the treatment group anti-HER2 for breast cancer, CCND1 is one of the 7 genes in this peak. The reason that we observe this peak is mostly because of CCND1.
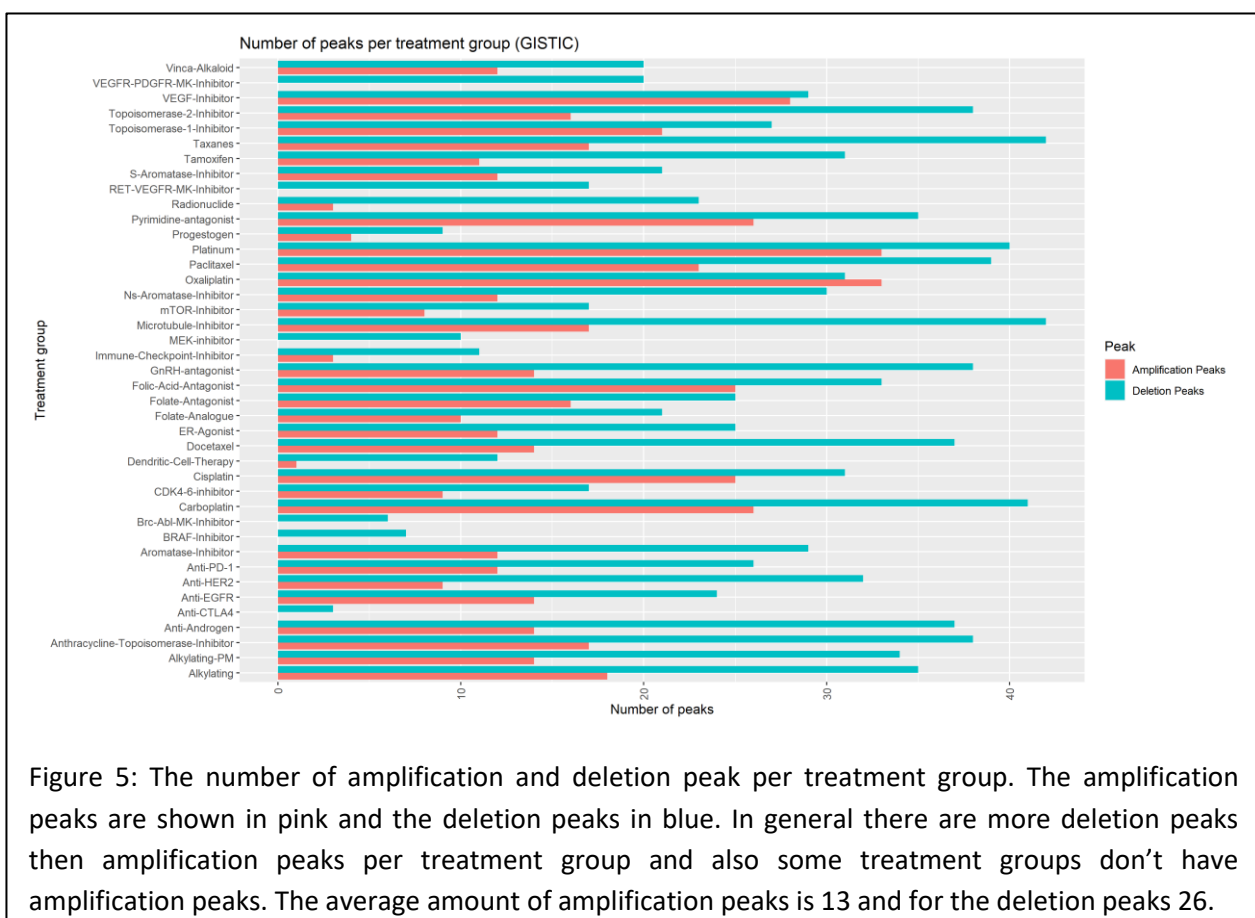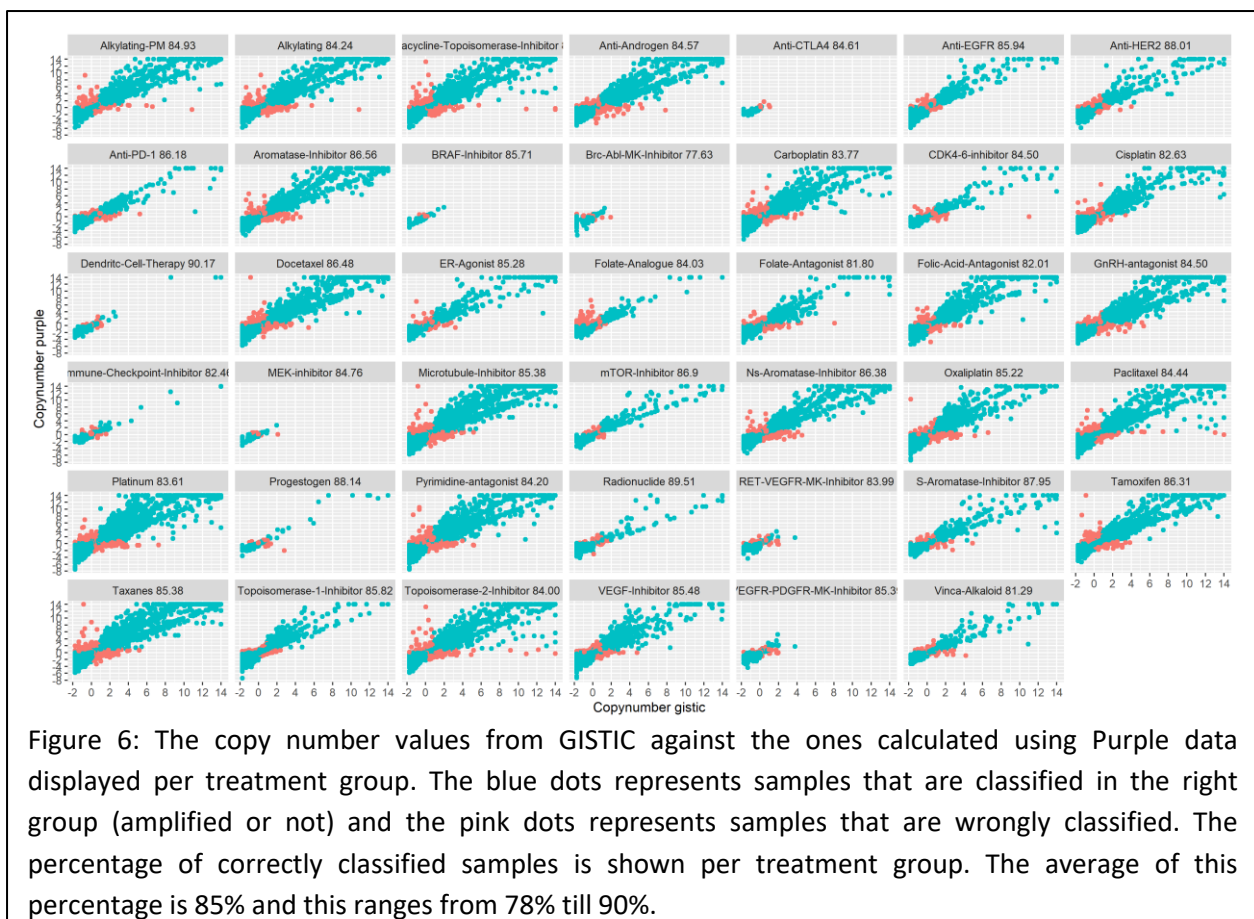

Figure 5: The number of amplification and deletion peak per treatment group. The amplification peaks are shown in pink and the deletion peaks in blue. In general there are more deletion peaks then amplification peaks per treatment group and also some treatment groups don't have amplification peaks. The average amount of amplification peaks is 13 and for the deletion peaks 26.

Fishers' exact test

GISTIC gives the amplified and deleted regions per sample for every treatment group, but now drivers of tumorigenesis need to be separated from the (possible) drivers of resistance. To identify the genes that are related to resistance mechanisms and are not cancer driver genes, we search for genes that show more amplification events in the treated group then in the control group.

As the GISTIC group annotations for each gene of the control set are not given, because you would expect that most of the regions don't show the same positive selected amplified and deleted regions compared to the treated group, the copy number values needed to be computed from raw copy number profiles similar as GISTIC approach. To assess whether our calculation is in line with GISTIC approach, we plotted the correlation between the copy numbers from GISTIC against the ones calculated with the use of purple files [Fig. 6]. The blue dots represent the samples that are right classified (amplified or not) compared to the GISTIC output and the pink dots the ones that are wrongly classified. On average 85% of the samples are classified in the same group annotation, with as lowest percentage of 78 and highest with 90. [Fig. 6], showing we can reproduce most of the GISTIC annotations.

Figure 6: The copy number values from GISTIC against the ones calculated using Purple data displayed per treatment group. The blue dots represents samples that are classified in the right group (amplified or not) and the pink dots represents samples that are wrongly classified. The percentage of correctly classified samples is shown per treatment group. The average of this percentage is 85% and this ranges from 78% till 90%.

The Fishers' exact test was performed using the (calculated) group annotations per sample for the test and control set. The amount of genes was reduced to 11.62 significant genes per amplification peak, with a range from 1 to 348 genes.

Gene annotation per peak

Since we still obtained a high number of genes for every peak, we assessed two methods used to narrow the amount of genes to one gene per peak; based on the p-value and based on the peak level. There were genes that popped up no matter which method was used, but also genes that only showed up for the p-value or peak level method alone. Due to this a combination of these two methods were used, to minimize the chance of removing genes that might be drivers of resistance. The average amount of significant genes is reduced till 2.36, with a range from 1 to 13. Figure 7 shows the most significant genes for all of the drug-groups.
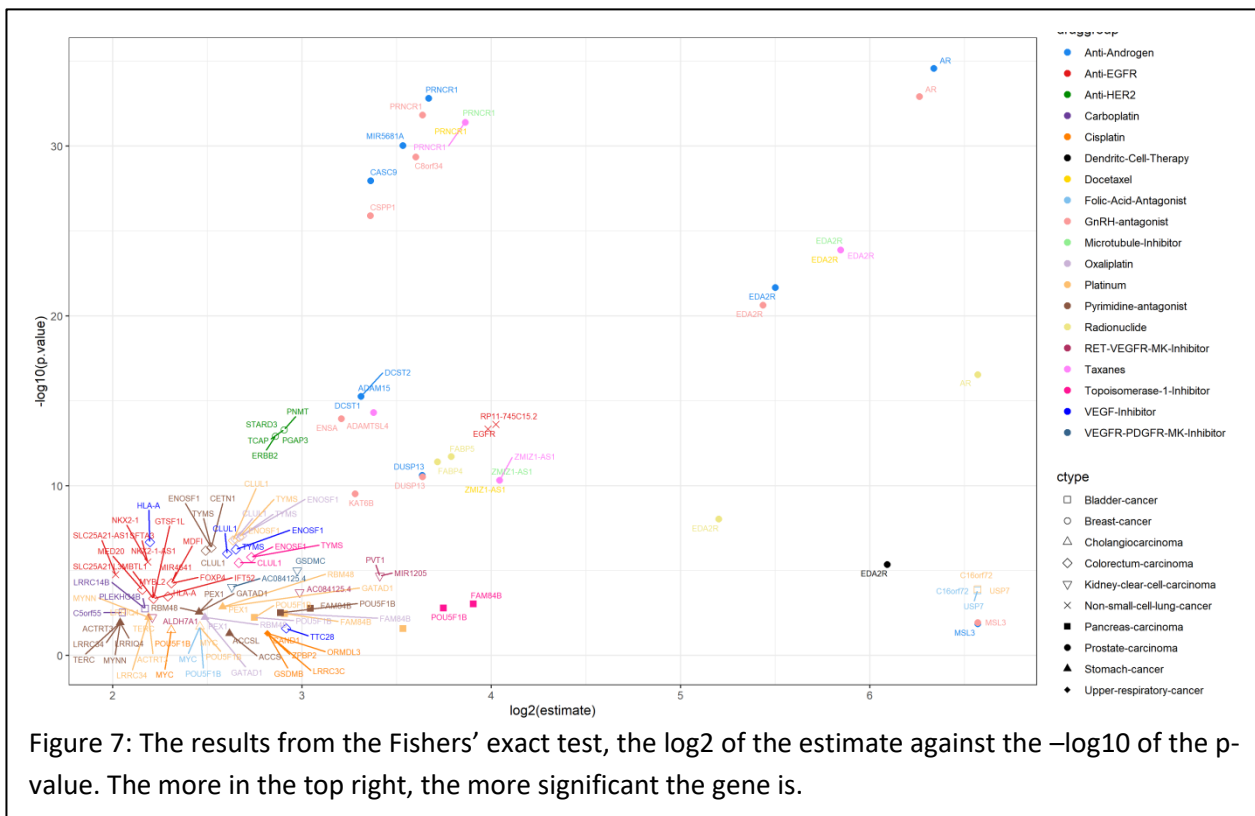

Figure 7: The results from the Fishers' exact test, the log2 of the estimate against the –log10 of the p-value. The more in the top right, the more significant the gene is.

Top hits

After narrowing down the genes identified by GISTIC, the top hits are reported. [Table 2]

The elevated expression of thymidylate synthase (TYMS) has been linked before with the resistance to 5-Fluoruracil (5-FU) (Varghese et al., 2019) (Christensen et al., 2019). 5-FU is a drug that has been classified in the treatment group Pyrimidine-antagonist. That we found TYMS as significant amplified with the treatment pyrimidine-antagonist is thus a positive control. There are more genes that are known as driver genes for cancer; Androgen receptor (AR) is a known driver for the treatment group Anti-androgen (Lallous et al., 2016) (Suzuki et al., 2003). That we found these known drivers shows that our approach works.

Gene	Treatment group	Cancer type	Estimate	P.value	Analysis
AR	Anti-androgen GnRH-antagonist	Prostate-carcinoma	Different	Different	Both
PRNCR1	Anti-androgen GnRH-antagonist	Prostate-carcinoma	Different	Different	Both
PRNCR1	Docetaxel Taxanes Microtubule-Inhibitor	Prostate-carcinoma	14.5359	3.90E-32	Both
MIR5681A	Anti-androgen GnRH-antagonist	Prostate-carcinoma	11.56572	9.12E-31	P-Value
C8orf34	GnRH-antagonist	Prostate-carcinoma	12.13019	4.36E-30	P-Value
CASC9	Anti-androgen GnRH-antagonist	Prostate-carcinoma	10.28247	1.07E-28	Top
CSPP1	GnRH-antagonist	Prostate-carcinoma	10.27215	1.24E-26	Top
EDA2R	Docetaxel Taxanes Microtubule-Inhibitor	Prostate-carcinoma	57.47194	1.31E-24	Both
EDA2R	Anti-androgen GnRH-antagonist	Prostate-carcinoma	Different	Different	Both
ADAM15	Anti-androgen	Prostate-carcinoma	9.917184	5.45E-16	Both
ADAM15	Docetaxel Taxanes Microtubule-Inhibitor	Prostate-carcinoma	10.39571	4.77E-15	Both
EGFR	Anti-EGFR	Non-small-cell-lung-cancer	15.80596	4.59E-14	Top
PNMT	Anti-HER2	Breast-cancer	7.489237	5.00E-14	P-Value
MSL3	Anti-androgen GnRH-antagonist	Prostate-carcinoma	94.96668	Different	Both
TYMS	Pyrimidine-antagonist	Colorectum-carcinoma	5.736882	4.41E-07	P-Value
HLA-A	VEGF-Inhibitor	Colorectum-carcinoma	4.579559	2.10E-07	Both

Table 2. Table representing the top hits. Per gene the treatment group, cancer type, estimate, p-value and analysis is given.

PRNCR1 is a long noncoding RNA (lncRNA) and was identified as a possible component in disease progression (Prensner et al., 2014). However, it's still not clear which roll it plays in prostate cancer. What is known is that PRNCR1 the proliferation of colorectal cancer promotes and a potential oncogene of this cancer is (Yang et al., 2016). Another lncRNA is CASC9, this gene is highly expressed in various cancers (Ma et al., 2017) (Liang et al., 2018) (Yang et al., 2019). Nevertheless there is not yet any explanation for the amplification in prostate cancer, and also not for the treatment groups anti-androgen or GnRH-ant(agonist).

The protein that is encoding by the gene EDA2R is a protein from het Tumor Necrosis Factor Receptor (TNFR) superfamily (Ihara et al., 2021). Other studies already indicate that there is a correlation between the amplification of EDA2R and the development of resistance for the treatments irinotecan and oxaliplatin (Sun et al., 2017).

It is found that the gene ADAM15 is highly upregulated in Prostate cancer (Najy et al., 2008) (Kuefer et al., 2006). It seems that it is cancer type specific and not treatment group specific. The gene MSL3 is significant upregulated with Anti-androgen in breast cancer. [Table 2] Surprisingly low expression of this gene is associated with ovarian cancer. This gene inhibits the malignant phenotype of ovarian cancer cells (Zheng et al., 2021) (Yamanoi et al., 2019).

Most of the top hits are from the cancer type *Prostate-carcinoma.* The treatment groups '*Anti-androgen* and *GnRH-ant(agonist)*' and '*Docetaxel*, *Taxanes* and *Microtubule-inhibitor'* are connected to each other, because they both share a lot of the same samples. [Fig. 3] Because of this it wasn't a surprise that the significant genes that showed up for these treatment groups are overlapping.

Discussion and conclusion

Here, we used the tool GISTIC that can identify regions of the genome that are significantly amplified or deleted across a set of samples. By using this tool on different treatment groups, we reveal insight into the potentially drivers of resistance against cancer treatments.

We found genes that are known driver genes for specific cancer treatments, like androgen receptor (AR) for anti-androgen, EGFR for anti-EGFR and TYMS for pyrimidine antagonist. These finding validate that the method we used was successful for finding drivers for specific cancer treatments. There are also genes that are not known specific as driver genes for cancer, these genes are possible candidates to be drivers of resistance. One of those genes is EDA2R, there is already a correlation between the amplification of this gene and the development of resistance against the treatments irinotecan and oxaliplatin (Sun et al., 2017). We found this gene amplified in Taxanes and anti-androgen, there is a big chance that this gene also has resistance against these treatments.

The low expression of MSL3 is associated with ovarian cancer, while we found it amplified with anti-androgen in breast cancer (Zheng et al., 2021) (Yamanoi et al., 2019). Further research is needed to determine what the role of this gene is within breast cancer and especially for the anti-androgen treatment. There are more genes found that are highly significant in specific treatment groups compared to the control; CASC9, CSPP1, PNMT and HLA-A. These genes are possible candidates for resistance and more research is needed to find out if this is the reason why they are amplified for these specific treatment groups.

The biggest difference when we look at our classification of the drugs into treatment groups and the FDA labeling is that our classification has two layers and the FDA only one. [Fig. 4] This will only give us a chance to also look at drug level with drugs that have a big test set and don't have any disadvantages. The rest was quite the same, so we can conclude that our classification of drugs into treatment groups was successful.

A potential artifact may be hidden in the clinical dataset as there is a chance that there are still samples in the control group that have been pretreated before, but are marked as untreated due to time. This result in that there might be drivers of resistance that are unwanted filtered out by the control set. To minimize the chance of this effect to happen, we wanted to add samples from the The Cancer Genome Atlas (TCGA) program. This dataset contains data from primary tumors from different cancer types before treatment (Neary et al., 2021). To reproduce the GISTIC scores and use them in the control group, the raw copy number data per sample per region was needed. The data

that we used was collected from FireBrowse, a service form the Broad Institute, that contains the TCGA data (Deng et al., 2016) (Deng et al., 2017). This data only contains the segment mean per sample per region. These values are already normalized using a log2 ratio. Multiple attempts were done to get to the raw copy number values, so we could use the same pipeline for the TCGA data as we did before. Every time the values seemed way off and within the time span of this study we weren't able to use the TCGA data as a control set. For further research within this scope it would be useful to find a way to use the TCGA data in the same way as we handled the HMF and PCAWG data.

We reduced the average amount of genes per peak from 11.62 to 2.36. This was done by selecting the gene with the lowest p-value (most significant) and the gene that was closest to the peak level. We would expect that these genes would be the same for both the methods, but this wasn't the case for all the peaks. Other extra post filtering steps can be the integration of the raw copy number levels as high ploidy genes (double minutes amplify genes with ploidy levels up to more than 50) are currently scored equally with genes with one extra ploidy level. Also, the integration of (matching) rnaseq data may be helpful to select for genes that are transcribed.

Nevertheless, it is still doubtable if you would like to select only one gene per peak. There might be a little chance that in that same peak region, two or more genes are co-amplified or co-deleted. For this analysis we chose to select only one gene per peak, because we needed to narrow down the number of genes to do a proper analysis on and the ones we selected are the most significant ones. For further research it would be interesting to not only look at the most significant genes but also genes that are not in the top hits but still significant. With that analysis the average amount of genes would be higher than the one used for our analysis, this may result in the need of using another method to do a proper analysis.

Methods section
A flowchart of the used methods is shown in Figure 2.

Create datasets
We performed the analyses using two datasets; the test set and the control set. For the test set, samples from Hartwig Medical Foundation (HMF) were used and one sample per patient was selected using the biopsy date. In total 2240 unique samples were used.
Samples that weren't pretreated (pretreatment='No') from HMF and the Pan-Cancer Analysis of Whole Genomes (PCAWG) study were used for the control set. The control set also contains only one sample per patient based on the biopsy data and consist of approximately 2600 samples.
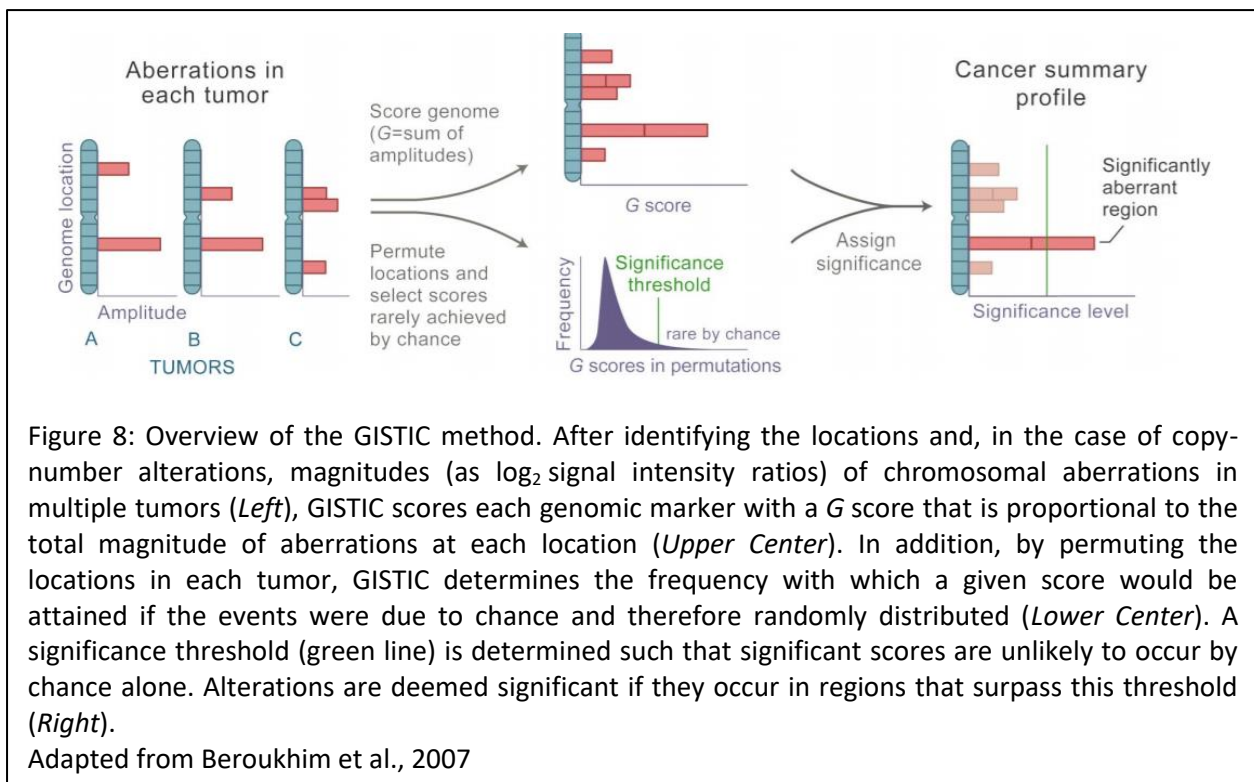
Classify treatment groups
We created treatments groups manually based on the mechanism of action of the drugs. This grouping was validated using known literature and their FDA labeling. Our treatment grouping has an extra layer. There are drugs that fall in the same treatment group when we look at their mechanism of action, but these drugs also have a sufficient number of samples (>300) to do the analysis also on drug level. Because of this, the actual amount of used (not unique) samples is much higher. Besides the extra layer, there are also samples that underwent multiple treatments and are therefore in multiple treatment groups.

GISTIC
The next step was to run the tool GISTIC with the copy number data from these samples, this can be done on pan-cancer and per-cancer level. Genomic Identification of Significant Targets in Cancer (GISTIC) is a great tool for analyzing chromosomal aberrations in cancer that underwent positive

selection. With this tool it is possible to identify copy number alterations, which might have a roll in resistance. This method will first identify the location and magnitudes of chromosomal aberrations in multiple tumor samples. [Fig. 8] Secondly the genome is scored based on the frequency and amplitude of the aberration. At the same time the frequency with which a given score would be obtained by chance is determined and is set as a threshold for the significance. This threshold is used to determine the significant aberration regions and these aberration regions are unlikely to occur by chance alone (Beroukhim et al., 2007). We ran GISTIC on the test set. GISTIC gives the amplified and deleted regions per sample for every treatment group and for every sample the difference in copies from the genome-wide median in that region. Every sample per region was annotated in a group, amplified, deleted or neutral. This annotation was done using a threshold, difference above 0.9 is amplified, below -1.3 is deleted and in between is neutral. The next step was to separate drivers of tumorigenesis from the (possible) drivers of resistance.

Figure 8: Overview of the GISTIC method. After identifying the locations and, in the case of copy-number alterations, magnitudes (as $\log_2$ signal intensity ratios) of chromosomal aberrations in multiple tumors (*Left*), GISTIC scores each genomic marker with a *G* score that is proportional to the total magnitude of aberrations at each location (*Upper Center*). In addition, by permuting the locations in each tumor, GISTIC determines the frequency with which a given score would be attained if the events were due to chance and therefore randomly distributed (*Lower Center*). A significance threshold (green line) is determined such that significant scores are unlikely to occur by chance alone. Alterations are deemed significant if they occur in regions that surpass this threshold (*Right*).
Adapted from Beroukhim et al., 2007

Fishers' exact test
We needed to distinguish the cancer driver genes from (possible) drivers of resistance. The cancer driver genes should also pop up in the control dataset, so we compared the test set with the control set using the Fishers' exact test. Four values are needed to perform the Fishers' exact test per treatment group; 1. The number of samples that were amplified in that peak for the test set. 2. The number of samples that were not amplified in that peak for the test set. 3. The number of samples that were amplified in that peak for the control set and 4. The number of samples that were not amplified in that peak for the control set. As the GISTIC group annotations for each gene of the control set are not given, we needed to compute these from raw copy number profiles similar as the GISTIC approach. The copy numbers given by GISTIC are the difference in copies from the genome-wide median. So to determine this value we took the raw copy number of the region of the gene and subtracted the mean ploidy level of that sample (CNpurple = (CN of that region in purple) - mean ploidy level). Now that we had all four values needed for the Fishers' exact test, we performed the test for every gene.

Gene annotation per peak

After performing the fishers' exact test, there were still too many genes to make conclusions. The solution to that was selecting only one gene per peak in GISTIC. Two methods were used to narrow down the amount of genes per peak. The first approach was based on the p-value, the gene with the lowest (most-significant) p-value was selected. The second approach was based on selecting the gene that was closest to the peak level (top). The peak level is the region with the most amplified samples. We used a combination of both methods, so we are sure we don't exclude (possible) resistance mechanisms. The top hit genes were reported based on their p-value and estimate and analyzed using literature.

References

Alkodsi, A., Louhimo, R., & Hautaniemi, S. (2015). Comparative analysis of methods for identifying somatic

copy number alterations from deep sequencing data. *Briefings in Bioinformatics*, *16*(2), 242–254.

https://doi.org/10.1093/bib/bbu004

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A.,

Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang,

F., Liu, E. M., Mularoni, L., … Ding, L. (2018). Comprehensive Characterization of Cancer Driver Genes

and Mutations. *Cell*, *173*(2), 371-385.e18. https://doi.org/10.1016/j.cell.2018.02.060

Baudis, M. (2007). Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis

of chromosomal CGH data. *BMC Cancer*, *7*(1), 226. https://doi.org/10.1186/1471-2407-7-226

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J.

H., Alexander, S., Du, J., Kau, T., Thomas, R. K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R.

M., Demichelis, F., … Sellers, W. R. (2007). Assessing the significance of chromosomal aberrations in

cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*,

104(50), 20007–20012. https://doi.org/10.1073/pnas.0710052104

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S.,

Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y.-J., Haery, L.,

Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., … Meyerson, M. (2010). The landscape of

somatic copy-number alteration across human cancers. *Nature*, *463*(7283), 899–905.

https://doi.org/10.1038/nature08822

Boudadi, K., & Antonarakis, E. S. (2016). Resistance to Novel Antiandrogen Therapies in Metastatic

Castration-Resistant Prostate Cancer. *Clinical Medicine Insights: Oncology*, *10s1*, CMO.Ss34534.

https://doi.org/10.4137/CMO.Ss34534

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H.

K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C.-L., Faquin, W. C.,

Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … The ICGC/TCGA Pan-Cancer Analysis of

Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–

93. https://doi.org/10.1038/s41586-020-1969-6

Chong, C. R., & Jänne, P. A. (2013). The quest to overcome resistance to EGFR-targeted therapies in cancer.

Nature Medicine, *19*(11), 1389–1400. https://doi.org/10.1038/nm.3388

Deng, M., Brägelmann, J., Kryukov, I., Saraiva-Agostinho, N., & Perner, S. (2017). FirebrowseR: An R client

to the Broad Institute's Firehose Pipeline. *Database: The Journal of Biological Databases and*

Curation, *2017*, baw160. https://doi.org/10.1093/database/baw160

Deng, M., Brägelmann, J., Schultze, J. L., & Perner, S. (2016). Web-TCGA: An online platform for integrated

analysis of molecular cancer data sets. *BMC Bioinformatics*, *17*(1), 72.

https://doi.org/10.1186/s12859-016-0917-9

Elsheikh, S., Green, A. R., Aleskandarany, M. A., Grainge, M., Paish, C. E., Lambros, M. B. K., Reis-Filho, J. S.,

& Ellis, I. O. (2008). CCND1 amplification and cyclin D1 expression in breast cancer and their relation

with proteomic subgroups and patient outcome. *Breast Cancer Research and Treatment*, *109*(2),

325–335. https://doi.org/10.1007/s10549-007-9659-8

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., & Stratton, M. R.

(2004). A census of human cancer genes. *Nature Reviews Cancer*, *4*(3), 177–183.

https://doi.org/10.1038/nrc1299

Giunta, S. (2021). Decoding human cancer with whole genome sequencing: A review of PCAWG Project

studies published in February 2020. *Cancer and Metastasis Reviews*. https://doi.org/10.1007/s10555-

021-09969-z

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler,

A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G.,

Buck, G., Choudhury, B., … Stratton, M. R. (2007). Patterns of somatic mutation in human cancer

genomes. *Nature*, *446*(7132), 153–158. https://doi.org/10.1038/nature05610

Ihara, K., Skupien, J., Krolewski, B., Md Dom, Z. I., O'Neil, K., Satake, E., Kobayashi, H., Rashidi, N. M.,

Niewczas, M. A., & Krolewski, A. S. (2021). A profile of multiple circulating tumor necrosis factor

receptors associated with early progressive kidney decline in Type 1 Diabetes is similar to profiles in autoimmune disorders. *Kidney International*, *99*(3), 725–736.

https://doi.org/10.1016/j.kint.2020.07.007

Iranzo, J., Martincorena, I., & Koonin, E. V. (2018). Cancer-mutation network and the number and specificity of driver mutations. *Proceedings of the National Academy of Sciences*, *115*(26), E6010–E6019. https://doi.org/10.1073/pnas.1803155115

Kim, T.-M., Xi, R., Luquette, L. J., Park, R. W., Johnson, M. D., & Park, P. J. (2013). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Research*, *23*(2), 217–227. https://doi.org/10.1101/gr.140301.112

Kobayashi, S., Boggon, T. J., Dayaram, T., Jänne, P. A., Kocher, O., Meyerson, M., Johnson, B. E., Eck, M. J., Tenen, D. G., & Halmos, B. (2005). EGFR Mutation and Resistance of Non–Small-Cell Lung Cancer to Gefitinib. *New England Journal of Medicine*, *352*(8), 786–792.

https://doi.org/10.1056/NEJMoa044238

Kuefer, R., Day, K. C., Kleer, C. G., Sabel, M. S., Hofert, M. D., Varambally, S., Zorn, C. S., Chinnaiyan, A. M., Rubin, M. A., & Day, M. L. (2006). ADAM15 Disintegrin Is Associated with Aggressive Prostate and Breast Cancer Disease. *Neoplasia*, *8*(4), 319–329. https://doi.org/10.1593/neo.05682

Lallous, N., Volik, S. V., Awrey, S., Leblanc, E., Tse, R., Murillo, J., Singh, K., Azad, A. A., Wyatt, A. W., LeBihan, S., Chi, K. N., Gleave, M. E., Rennie, P. S., Collins, C. C., & Cherkasov, A. (2016). Functional analysis of androgen receptor mutations that confer anti-androgen resistance identified in circulating cell-free DNA from prostate cancer patients. *Genome Biology*, *17*(1), 10.

https://doi.org/10.1186/s13059-015-0864-1

Li, C., Iida, M., Dunn, E. F., Ghia, A. J., & Wheeler, D. L. (2009). Nuclear EGFR contributes to acquired resistance to cetuximab. *Oncogene*, *28*(43), 3801–3813. https://doi.org/10.1038/onc.2009.234

Liang, Y., Chen, X., Wu, Y., Li, J., Zhang, S., Wang, K., Guan, X., Yang, K., & Bai, Y. (2018). LncRNA CASC9 promotes esophageal squamous cell carcinoma metastasis through upregulating LAMC2 expression

by interacting with the CREB-binding protein. *Cell Death & Differentiation*, *25*(11), 1980–1995.

https://doi.org/10.1038/s41418-018-0084-9

Ma, P., Zhang, M., Nie, F., Huang, Z., He, J., Li, W., & Han, L. (2017). Transcriptome analysis of EGFR

tyrosine kinase inhibitors resistance associated long noncoding RNA in non-small cell lung cancer.

Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie, *87*, 20–26.

https://doi.org/10.1016/j.biopha.2016.12.079

Melguizo, C., Prados, J., Rama, A., Quesada, R., Alvarez, P., Fernández, J., & Aránega, A. (2011). Multidrug

resistance and rhabdomyosarcoma (Review). *Oncology reports*, *26*, 755–761.

https://doi.org/10.3892/or.2011.1347

Mitelman, F., Johansson, B., & Mertens, F. (2007). The impact of translocations and gene fusions on cancer

causation. *Nature Reviews Cancer*, *7*(4), 233–245. https://doi.org/10.1038/nrc2091

Morgillo, F., Della Corte, C. M., Fasano, M., & Ciardiello, F. (2016). Mechanisms of resistance to EGFR-

targeted drugs: Lung cancer. *ESMO Open*, *1*(3), e000060. https://doi.org/10.1136/esmoopen-2016-

000060

Moroni, M., Veronese, S., Benvenuti, S., Marrapese, G., Sartore-Bianchi, A., Di Nicolantonio, F.,

Gambacorta, M., Siena, S., & Bardelli, A. (2005). Gene copy number for epidermal growth factor

receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: A cohort study. *The

Lancet Oncology*, *6*(5), 279–286. https://doi.org/10.1016/S1470-2045(05)70102-9

Najy, A. J., Day, K. C., & Day, M. L. (2008). ADAM15 Supports Prostate Cancer Metastasis by Modulating

Tumor Cell–Endothelial Cell Interaction. *Cancer Research*, *68*(4), 1092–1099.

https://doi.org/10.1158/0008-5472.CAN-07-2432

Neary, B., Zhou, J., & Qiu, P. (2021). Identifying gene expression patterns associated with drug-specific

survival in cancer patients. *Scientific Reports*, *11*(1), 5004. https://doi.org/10.1038/s41598-021-

84211-y

Ortiz, A. B., Garcia, D., Vicente, Y., Palka, M., Bellas, C., & Martin, P. (2017). Prognostic significance of cyclin D1 protein expression and gene amplification in invasive breast carcinoma. *PLOS ONE*, *12*(11), e0188068. https://doi.org/10.1371/journal.pone.0188068

Pham, V. V. H., Liu, L., Bracken, C., Goodall, G., Li, J., & Le, T. D. (2021). Computational methods for cancer driver discovery: A survey. *Theranostics*, *11*(11), 5553–5568. https://doi.org/10.7150/thno.52670

Prensner, J. R., Sahu, A., Iyer, M. K., Malik, R., Chandler, B., Asangani, I. A., Poliakov, A., Vergara, I. A., Alshalalfa, M., Jenkins, R. B., Davicioni, E., Feng, F. Y., & Chinnaiyan, A. M. (2014). The lncRNAs PCGEM1 and PRNCR1 are not implicated in castration resistant prostate cancer. *Oncotarget*, *5*(6), 1434–1438.

Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., Roepman, P., Voda, M., Bloemendal, H. J., Tjan-Heijnen, V. C. G., van Herpen, C. M. L., Labots, M., Witteveen, P. O., Smit, E. F., Sleijfer, S., … Cuppen, E. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, *575*(7781), 210–216. https://doi.org/10.1038/s41586-019-1689-y

Ritchie, H., & Roser, M. (2018). Causes of Death. *Our World in Data*. https://ourworldindata.org/causes-of-death

Sakkiah, S., Ng, H. W., Tong, W., & Hong, H. (2016). Structures of androgen receptor bound with ligands: Advancing understanding of biological functions and drug discovery. *Expert Opinion on Therapeutic Targets*, *20*(10), 1267–1282. https://doi.org/10.1080/14728222.2016.1192131

Sequist, L. V., Waltman, B. A., Dias-Santagata, D., Digumarthy, S., Turke, A. B., Fidias, P., Bergethon, K., Shaw, A. T., Gettinger, S., Cosper, A. K., Akhavanfard, S., Heist, R. S., Temel, J., Christensen, J. G., Wain, J. C., Lynch, T. J., Vernovsky, K., Mark, E. J., Lanuti, M., … Engelman, J. A. (2011). Genotypic and Histological Evolution of Lung Cancers Acquiring Resistance to EGFR Inhibitors. *Science Translational Medicine*, *3*(75), 75ra26-75ra26. https://doi.org/10.1126/scitranslmed.3002003

Shan, Y.-S., Hsu, H.-P., Lai, M.-D., Hung, Y.-H., Wang, C.-Y., Yen, M.-C., & Chen, Y.-L. (2017). Cyclin D1

overexpression correlates with poor tumor differentiation and prognosis in gastric cancer. *Oncology*

Letters, *14*(4), 4517–4526. https://doi.org/10.3892/ol.2017.6736

Sigismund, S., Avanzato, D., & Lanzetti, L. (2018). Emerging functions of the EGFR in cancer. *Molecular*

Oncology, *12*(1), 3–20. https://doi.org/10.1002/1878-0261.12155

Spencer, D. H., Zhang, B., & Pfeifer, J. (2015). Chapter 8—Single Nucleotide Variant Detection Using Next

Generation Sequencing. In S. Kulkarni & J. Pfeifer (Red.), *Clinical Genomics* (pp. 109–127). Academic

Press. https://doi.org/10.1016/B978-0-12-404748-8.00008-3

Sun, X., & Hu, B. (2018). Mathematical modeling and computational prediction of cancer drug resistance.

Briefings in Bioinformatics, *19*(6), 1382–1399. https://doi.org/10.1093/bib/bbx065

Sun, X., Wang, X., Feng, W., Guo, H., Tang, C., Lu, Y., Xiang, X., & Bao, Y. (2017). Gene signatures associated

with drug resistance to irinotecan and oxaliplatin predict a poor prognosis in patients with colorectal

cancer. *Oncology Letters*, *13*(4), 2089–2096. https://doi.org/10.3892/ol.2017.5691

Suzuki, H., Ueda, T., Ichikawa, T., & Ito, H. (2003). Androgen receptor involvement in the progression of

prostate cancer. *Endocrine-Related Cancer*, *10*(2), 209–216.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.

S., Getz, G., Bader, G. D., Ding, L., & Lopez-Bigas, N. (2013). Comprehensive identification of

mutational cancer driver genes across 12 tumor types. *Scientific Reports*, *3*(1), 2650.

https://doi.org/10.1038/srep02650

van Diest, P. J., Michalides, R. J., Jannink, L., van der Valk, P., Peterse, H. L., de Jong, J. S., Meijer, C. J., &

Baak, J. P. (1997). Cyclin D1 expression in invasive breast cancer. Correlations and prognostic value.

The American Journal of Pathology, *150*(2), 705–711.

Varghese, V., Magnani, L., Harada-Shoji, N., Mauri, F., Szydlo, R. M., Yao, S., Lam, E. W.-F., & Kenny, L. M.

(2019). FOXM1 modulates 5-FU resistance in colorectal cancer through regulating TYMS expression.

Scientific Reports, *9*(1), 1505. https://doi.org/10.1038/s41598-018-38017-0

Wieduwilt, M. J., & Moasser, M. M. (2008). The epidermal growth factor receptor family: Biology driving

   targeted therapeutics. *Cellular and molecular life sciences : CMLS*, *65*(10), 1566–1584.

   https://doi.org/10.1007/s00018-008-7440-8

Yamanoi, K., Baba, T., Abiko, K., Hamanishi, J., Yamaguchi, K., Murakami, R., Taki, M., Hosoe, Y., Murphy, S.

   K., Konishi, I., Mandai, M., & Matsumura, N. (2019). Acquisition of a side population fraction

   augments malignant phenotype in ovarian cancer. *Scientific Reports*, *9*(1), 14215.

   https://doi.org/10.1038/s41598-019-50794-w

Yang, L., Qiu, M., Xu, Y., Wang, J., Zheng, Y., Li, M., Xu, L., & Yin, R. (2016). Upregulation of long non-coding

   RNA PRNCR1 in colorectal cancer promotes cell proliferation and cell cycle progression. *Oncology*

   Reports, *35*(1), 318–324. https://doi.org/10.3892/or.2015.4364

Yang, Q., Fung, K.-M., Day, W. V., Kropp, B. P., & Lin, H.-K. (2005). Androgen receptor signaling is required

   for androgen-sensitive human prostate cancer cell proliferation and survival. *Cancer Cell*

   International, *5*(1), 8. https://doi.org/10.1186/1475-2867-5-8

Yang, Y., Chen, D., Liu, H., & Yang, K. (2019). Increased expression of lncRNA CASC9 promotes tumor

   progression by suppressing autophagy-mediated cell apoptosis via the AKT/mTOR pathway in oral

   squamous cell carcinoma. *Cell Death & Disease*, *10*(2), 1–16. https://doi.org/10.1038/s41419-018-

   1280-8

Zhang, J., & Zhang, S. (2017). Discovery of cancer common and specific driver gene sets. *Nucleic Acids*

   Research, *45*(10), e86. https://doi.org/10.1093/nar/gkx089

Zhao, B., Wang, L., Qiu, H., Zhang, M., Sun, L., Peng, P., Yu, Q., & Yuan, X. (2016). Mechanisms of resistance

   to anti-EGFR therapy in colorectal cancer. *Oncotarget*, *8*(3), 3980–4000.

   https://doi.org/10.18632/oncotarget.14012

Zheng, J., Zhou, C., Zhang, T., Guo, W., Cui, X., Han, C., Xu, Y., & Cai, C. (2021). *Biomarkers Associated with*

   CD8+ T Cell Infiltration in Childhood AML [Preprint]. In Review. https://doi.org/10.21203/rs.3.rs-

   659529/v1