universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Genetic characterization of patient derived bladder cancer cell models“

verfasst von / submitted by

Emily Aline Grubb

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Acknowledgements

First of all, I would like to thank my thesis supervisor at the Department of Urology and Center for Cancer Research, Medical University of Vienna, Bernhard Englinger for providing me with such an interesting project and for welcoming me so warmly into his research group. I really appreciate all the time you took to introduce me to the project and your endless scientific guidance throughout the past months. I would also like to thank fellow Englinger lab members Paula and Neha for their continued scientific input on my project and always being up for an after-work drink. Additionally, I would like to acknowledge Christine for always being available to share her scientific wisdom, helping me with my data interpretation and always making me laugh. On the computational side, I would like to thank Iros Barozzi and Thomas Mohr for their guidance. I would also like to thank all members of the Englinger, Berger and Heffeter research groups, I have really enjoyed working alongside you this year.

I would also like to acknowledge the work of all members of the Department of Urology at Vienna General Hospital, both in the clinical and research teams, and Walter Berger and his team at the Center for Cancer Research, who were instrumental in obtaining the tumor samples and subsequent generation of the primary patient derived cell models. Computational resources were provided by the Vienna Scientific Cluster (VSC). This project would not have been possible without Thomas Rattei, who kindly agreed to act as my internal supervisor at the University of Vienna.

On a personal note, I am hugely grateful for the friendship of my fellow ESB course mates Arne, Flo and Stefan, who helped me stay motivated during the height of the Covid pandemic when all of our teaching was done online.

Table of Contents

Abstract

The heterogeneity of bladder cancer presents a huge challenge to develop effective treatments. Historically, there has been a lack of models that faithfully recapitulate the disease biology *in situ.* To improve treatment discovery, as well as general understanding of tumor biology, primary patient derived cell models had been generated. 42 of these models were subjected to whole exome sequencing with the aim to genetically characterize each model by identifying mutations and copy number alterations. To achieve this aim, bioinformatic tools and parameters within these tools were tested to optimize the workflow for processing this data. The results confirmed the heterogeneity between the different bladder cancer models and will no doubt be helpful in future projects assessing novel therapeutic avenues and resistance mechanisms of the disease.

Die Heterogenität von Blasenkrebses stellt eine bedeutende Herausforderung für die Entwicklung wirksamer Behandlungen dar. In der Vergangenheit fehlte es an Modellen, die die Biologie dieser Krankheit in situ präzise nachbilden konnten. Um die Entwicklung von Therapien zu fördern und das allgemeine Verständnis der Tumorbiologie zu vertiefen, wurden primäre, von Patienten stammende Zellmodelle entwickelt. Von diesen Modellen wurden 42 einer vollständigen Exom-Sequenzierung unterzogen, um jedes Modell auf genetischer Ebene durch die Identifizierung von Mutationen und Kopiezahlvariation zu charakterisieren. Zu diesem Zweck Hierfür, wurden bioinformatische Werkzeuge mit verschiedenem Parameter sorgfältig getestet, um die Verarbeitung der Daten zu optimieren. Die Ergebnisse bestätigten die Heterogenität der verschiedenen Blasenkrebsmodellen und werden zweifellos von großem Nutzen bei künftigen Projekten zur Bewertung neuer therapeutischer Ansätze und der Erforschung von Resistenzmechanismen dieser Krankheit sein.

List of Tables

List of Figures

Introduction

Bladder cancer is the 9th most common malignant disease worldwide. The disease more commonly affects men and can be present as non-muscle-invasive (NMIBC), muscle-invasive bladder cancer (MIBC), or in a metastatic form (Y. Li et al., 2021). Once bladder cancer is diagnosed, patient prognosis strongly varies depending on clinical staging and grading, currently based predominantly on histopathological evaluation. While NIMBC is clinically well manageable in the majority of cases, MIBC and especially metastatic disease has a poor prognosis. Besides surgical tumor resection, the existing solution is chemotherapeutic approaches rooted in (predominately neoadjuvant, i.e., pre-surgical intervention) first-line cisplatin-based single agent or combination treatment, which was first introduced over 30 years ago (F. Wang et al., 2022). Unfortunately, this approach is not effective for all patients and new treatments for the disease are desperately needed. Furthermore, patients' tumors differ in their chemosensitivity due to the heterogeneous nature of tumor cell composition and often develop chemoresistance, resulting in relapse and subsequent tumor progression (Martin-Doyle & Kwiatkowski, 2015).

Technological advances in next-generation sequencing (NGS) have made it possible to improve our understanding of the molecular and genetic processes of cancer biology, including tumor progression. Human genes are made up of exons which are protein coding regions, and non-protein coding regions, i.e., introns and 3' and 5' untranslated regions. NGS techniques can be used to examine the whole genome, so called whole genome sequencing, or parts of the genome, for example whole exome sequencing (WES) (Seaby et al., 2016). Despite only representing 2% of the genome, the exome has been shown to contain ~85% of disease causing variants (EL et al., 2014). When examining WES data, one can detect several types of genetic variation (polymorphism) in the healthy population including single nucleotide variants, Indels (insertion or deletion of a series of nucleotide) or microsatellite variants (structural variants). For instance, short tandem repeats (STRs) are a type of microsatellite frequently used in molecular biology to analyse the degree of similarity between samples, such as cell lines, or individuals in forensics. In addition, WES is a powerful tool to identify genetic alterations of a tumor (Bartha & Győrffy, 2019). This can improve understanding of cancer development and progression, as well as inform targeted therapies to improve subsequent patient outcome.

The genomic defects observed in cancer are complex and wide ranging, with single nucleotide alterations to whole deletions or amplifications of chromosomes. It is well established that activation of oncogenes and suppression of tumor suppressor genes occurs during tumor formation, progression and metastasis (Y. Li et al., 2021). Bladder cancer exhibits very high tumor mutational burden, with only lung cancer and melanoma exhibiting more mutations. This is likely due to the strong association of this disease with smoking and other mutagenic environmental toxins (Maxson & Mitchell et al., 2014). Illustratively, previous WES data from an extensive panel of bladder tumors and matched normal samples identified over 130,000 somatic mutations (Robertson et al., 2017). However, none of these (with the exception of FGFR3 mutation) have so far translated into targeted therapy approaches in the clinic (Kacew & Sweis, 2020). This demonstrates the need for improved understanding of the genetic landscape, also in context with transcriptional composition, of bladder cancer in order to ensure better treatment outcomes for patients.

This project aims at the examination of WES data from primary patient derived bladder cancer models, alongside healthy matched sample data, to identify genetic alterations. These patient-derived, so-called Vienna Urothelial Carcinoma (VUCs) cells have been established directly from bladder cancer patients in the Department of Urology and Center for Cancer Research at Medical University of Vienna. The cell panel (n=42) exhibits full clinical annotations including, e.g., tumor stage, grade, age, sex and the type of treatment the patient has received (if treated). Samples have been obtained during cystectomies or transurethral resections performed at Vienna General Hospital. An advantage of the VUC models in particular is they have a low passage number and are therefore believed to recapitulate the native tumor situation more faithfully than established commercial cell lines. These primary cell models will form the basis for a number of projects investigating factors of cancer aggressiveness, therapy resistance, as well as novel genetic/transcriptional dependencies to identify novel therapeutic vulnerabilities. An imperative prerequisite for the successful achievement of this will be the careful selection of cell models most faithfully representing native tumors.

This work primarily aims to genetically characterise the patient derived bladder cancer cell models individually according to their short nucleotide variants (SNVs and Indels) and copy number alterations (CNA). Additionally, this work aims to integrate these results with clinical data that exists on these models and then to contextualise the results by integrating data with publicly available multi-omic datasets available for bladder cancer.

Materials and methods

Data acquisition

A summary of the project set up can be seen in Fig 1 below. This thesis will describe the bioinformatic pipeline and results for variant calling and detecting copy number alterations (CNA).

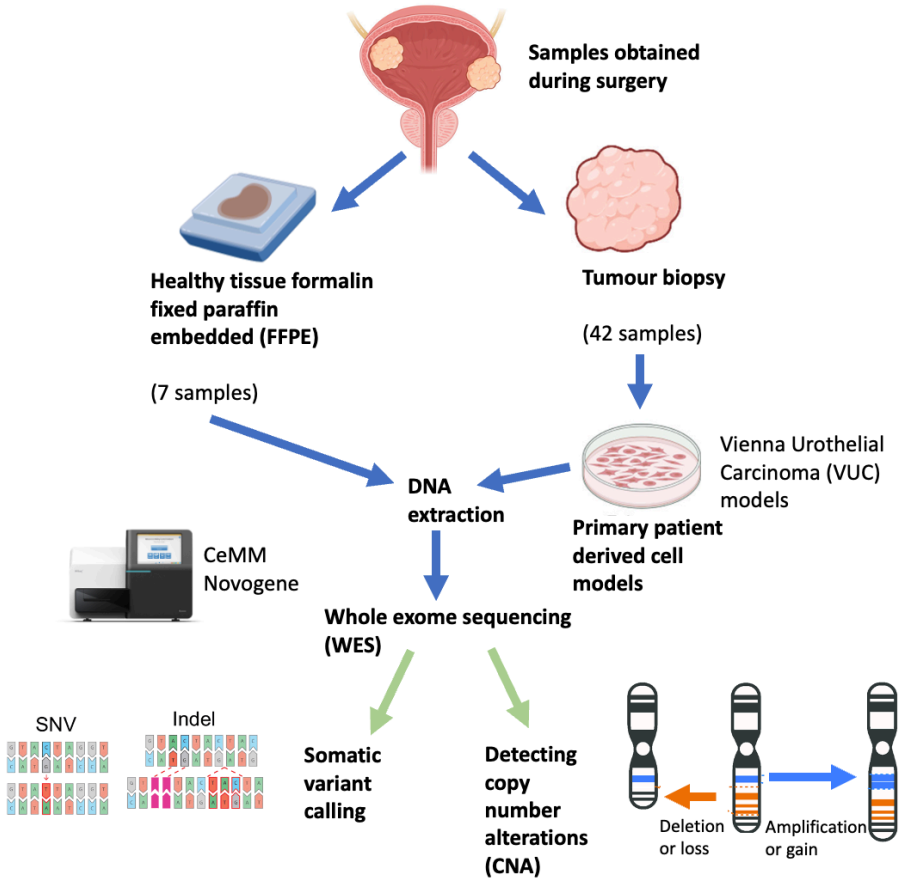


Figure 1. Project set up showing processing of tumor and normal samples from surgery to final bioinformatic outputs. The Department of Urology at Vienna General Hospital obtained samples during surgery. Subsequently, primary patient derived cell models were generated, and DNA extracted at the Center for Cancer Research, Medical Univeristy. The Department of Pathology at Vienna General Hospital identified the healthy tissue and embedded it. Whole exome sequencing (WES) of the models was carried out across two sequencing centers, CeMM and Novogene. SNV = somatic nucleotide variation, Indel = insertion-deletion. Created with BioRender and adapted from Gandawijaya et al., 2021 and Nesta et al., 2021.

Samples were obtained from bladder cancer patients during cystectomies or transurethral resections. From these samples patient derived cell lines were cultured (VUC models) according to standard laboratory protocols (Ertl et al., 2022).

42 VUC models had been subjected to WES at two different sequencing facilities. 9 of the VUC models and 7 normal tissue matches had been sequenced at Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM), Vienna, Austria and the other 33 had been sequenced at Novogene, UK. 7 of the VUC models had normal tissue matches, where DNA was extracted from formalin fixed paraffin embedded (FFPE). The normal samples were tissue sections removed from the bladder during surgery, and were determined as healthy, then formalin fixed paraffin embedded by the Department of Pathology at Vienna General Hospital.

Library preparation and sequencing

For the 16 samples processed at CeMM (9 VUC models and 7 normal tissue matches), DNA was fragmented using enzymatic digestion, before being size selected and amplified using a PCR step. Libraries were hybridised with biotinylated baits and captured with streptavidin-conjugated magnetic beads. Specifically, Twist Bioscience LP Kit Human Core Exome + RefEsq were used. After enrichment, regions were amplified and size selected. Quality control of the libraries were quality controlled using Qubit Fluorometric Quantitation system (Life Technologies) and automated electrophoresis Agilent Bioanalyzer system. Sequencing was carried out on an Illumina Novaseq 6000 using paired-end strategy PE100.

For the 33 samples processed at Novogene the library preparation and sequencing were as follows. DNA was randomly fragmented by sonication to the size of 180 – 280 bp fragments. DNA fragments with ligated adapter molecules on both ends were selectively enriched in a PCR reaction. Then libraries were hybridized in buffer with biotin-labelled probes. To capture the exons magnetic beads with streptomycin were used, after washing beads and digesting probes the captured libraries were enriched in a PCR step to add index tags. Specifically, Agilent SureSelectXT Reagent Kit + Agilent SureSelect Human All ExonV6 were used. Products were purified with AMPure XP system (Beckman Coulter, Beverley, USA) and quantified by using the

Agilent high sensitivity DNA assay on the Agilent Bioanalyzer system. Sequencing was carried out on an Illumina platform Novaseq 6000 V1.5 reagent, S4 flowcell with paired-end strategy PE150.

Data pre-processing

Data was processed from the raw sequencing files (.fastq) for all 42 VUC models and the 7 normal tissue matches. Unless specified otherwise all data analysis was carried out on the Vienna Scientific Cluster (VSC).The initial quality control of the .fastq samples was assessed using FastQC (Andrews.S, 2010) on individual files. To compare the quality between samples from the same batch and compile the results into a simple report, MultiQC, was used (Ewels et al., 2016). After assessing the MultiQC outputs, the samples that had been processed at Novogene were subject to trimming using Trim Galore (Krueger Felix.,), in order to remove adapters still present from sequencing. Default settings were used with the minimum required overlap (stringency) of 10 bp. To check the trimming had been successful MultiQC was used for quality control. The reads were aligned to the reference human genome 38 (hg38) provided by GATK (Genome Analysis Toolkit) (Van der Auwera et al., 2013) using BWA-MEM aligner 0.7.17 (H. Li, 2013). For alignment post processing, SAMtools (H. Li et al., 2009) was used to quantify the alignment output and to sort .bam files. To mark PCR duplicates, the tool MarkDuplicates (Picard) provided by GATK (4.2.2.0) was used (Depristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). Finally, quality score recalibration was carried out using BaseRecalibrator with known SNPs and Indels files from the GATK resource bundle followed by Base Quality Score Recalibration (BQSR) provided by GATK (4.2.2.0). To check the coverage and performance of the WES experiments the tool CollectWgsMetrics provided by GATK (4.2.2.0) was used with the interval list provided by the respective sequencing centres.

Short variant discovery

All tools mentioned below are part of GATK (version 4.2.2.0) and all steps were run on individual samples, unless specified otherwise, on the Vienna Scientific Cluster (VSC). The following input files were always the same for each step that required them and were as follows: human genome 38 (hg38) provided by GATK, interval list (padded by 100 bp) as provided by the sequencing centre with the interval padding (-ip) of 450 (450 bp either side of the padded given intervals). To create

the padded interval list the PreprocessIntervals command was used with the --padding 100 specified.

Variant discovery followed GATK's Best Practices Workflow - Short Variant Discovery pipeline recommendations which is designed to detect somatic SNVs and Indels (Van der Auwera et al., 2013). Mutect2 was run with the following parameters separately on each individual sample and input files; as a reference human genome 38 (hg38), germline resource and panel of normals provided by GATK, af-of-alles-not-in-resource 0.001, -L specified as the padded (by 100bp) interval list provided by the sequencing centre, -ip 450 (interval padding by 450 bp) and –f1r2-tar-gz was specified. Padding the target intervals extended the regions investigated to ensure not to miss variants in coding regions, 100 bp was decided based on previous studies (Corominas et al., 2022).The output from –f1r2-tar-gz was used to learn an orientation biased model using the tool LearnReadOrientationModel, to estimate substitution errors occurring as a result of FFPE damage or during library preparation. All steps described above were used on all samples.

Mutect2 was run in tumor-only mode as well as tumor-normal pairs (where a normal was available). To examine how normal the samples were and to detect potential tumor content, normal samples were also examined running the tumor-only pipeline with the normal samples only. Initially different variations of the pipeline were tested where different panel of normal (PONs) were used. On one hand, the GATK provided PON was used and on the other hand, one PON was created from the normal samples provided (n=7) using the Mutect2 tool in tumor-only mode for each normal sample and then CreateSomaticPanelOfNormals. Additionally, the steps of GetPileUpSummaries and CalculateContamination were tested and run according to "Best Practices Workflow" recommendations after Mutect2 calling in tumor only mode. The output from CalculateContamination was used as a further input for the FilterMutectCalls tool.

To filter the variants called by short variant discovery pipeline the tool FilterMutectCalls was run with the parameters -R hg38, --obs-prior , -L padded interval list and -ip 450.

Short variant annotation and filtering

To just select the variants that were labelled as a 'PASS', the tool SelectVariants using the –exclude-filtered option in addition to the standard parameters. To identify and annotate the variants, the tool Funcotator was used with the data sources path downloaded from GATK and giving an output file in .vcf and .maf format for each sample.

A table for each sample was created showing the short variant detected and the annotations using the VariantsToTable tool on the .vcf output using the -F option to specify the columns required in the table such as chromosome, position, reference allele, alternate allele, quality, funcotation, allelic depth, allelic frequency.

Clinical oncopanel

In the clinic, a next-generation sequencing assay called an oncopanel TruSight™ Oncology 500 is carried out to aid prediction, prognosis and diagnosis by healthcare professionals. Here, depending on the assay kit provider, between 50 and 100 so-called "mutation hotspot" genes are sequenced to a very high coverage (Gong et al., 2022). Where an oncopanel was available and had been carried out in the clinic (n=4), the results were compared against the output from the short variant discovery to validate the results.

Short variant data visualisation

All the .maf files outputted were concatenated together using cat VUC*.maf to create one .maf file for each sequencing run. To visualise the short variants present, the data was imported as an .maf file into R locally (version 4.2.2) (R Core Team, 2022) using the package maftools (version 2.14.0) (Mayakonda et al., 2018). To compare the data to the existing bladder cancer datasets, The Cancer Genome Atlas Program (TCGA) (TCGA Research Network, https://www.cancer.gov/tcga) bladder cancer dataset (n=412) (Robertson et al., 2017; Weinstein et al., 2014) was also imported into R. Plots were generated according to the manual. The top 20 genes most commonly observed with variants was defined as what is listed on the TCGA database website for bladder cancer.

Copy number calling pipeline

CNVkit (version 0.9.10) was run in a conda environment according to the copy number calling pipeline manual (Talevich et al., 2016). For running the pipeline, the refFlat.txt.gz (genome annotation file) was downloaded from https://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/ to enable genes to be correctly annotated in the analysis. To identify sequencing-inaccessible regions and remove them from the analysis the umap_k100_mappability.bed (bed file with hard to map regions) was downloaded from https://github.com/AstraZeneca-NGS/reference_data/blob/master/hg38/tricky_regions/. The same genome sequence hg38 from GATK was used as in the short variant discovery pipeline above. The standard commands were run in addition to cnvkit.py access -x umap_k100_mappability.bed. A reference was created by combining the tumor samples together. The created .cns (copy number segment) file for each model was used to plot the copy number alterations for each in a scatter plot. The diagram command used the created .cns and .cnr (copy number ratio) files to generate an ideogram of the chromosomes. A log2 ratio above 0.2 was classified as an amplification/gain or a deletion/loss. The output scatter and diagrams for each sample was compared to existing array comparative genomic hybridization (aCGH) profiles that had been generated previously in the research group. To summarise the large-scale copy number alterations (CNAs) across all the VUC models the heatmap command was used on the .cns file generated in the previous steps.

Results

Quality control outputs from pre-processing of data

When examining the initial quality of the data, samples processed at Novogene still had adapters present so were subjected to trimming to remove the adapters. The initial quality control outputs on the raw reads for the samples processed at Novogene showed a right shift in the distribution of the percentage of GC content was observed (Fig. 2). For the samples processed at CeMM, there was a bimodal peak observed, as expected for whole exome sequencing (WES) (Fig. 2). For full quality control output graphs see Fig S1 and Fig S2.

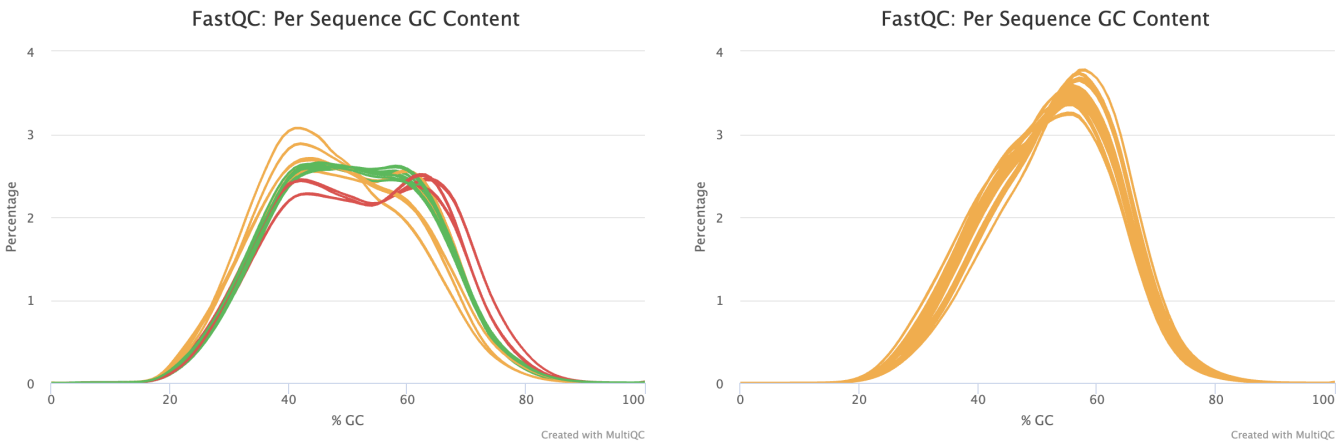


Figure 2. Initial quality control plot showing Per sequence GC Content summary for CeMM processed samples (left) and Novogene processed samples (right). Colors are automatically generated by the programs FastQC and MultiQC, which expect a normal distribution peak of GC content. According to the program manual green indicates the samples have passed, orange indicates the samples have warnings and red indicates the samples have failed. It must be noted for WES a bimodal distribution peak is expected, as observed for the CeMM samples.

All samples showed a successful rate of alignment against the reference genome, where the percentage of reads that were mapped was over 99 %. Both datasets showed differences in a number of quality control parameters including coverage, duplication rate and insert size (Table 1). CeMM processed samples had more total mapped reads, higher median coverage and less duplicated reads. Where normal samples were available, they generally had a reduced number of total mapped reads and lower median coverage as well as a higher percentage of duplicate reads when compared to the tumor VUC models.

Parameter		
Sequencing center	CeMM	Novogene
Number of samples	16 (9 tumor and 7 matched normal)	33 (all tumor)
Exome enrichment kit used	Twist Human Core Exome + Refseq Panel	Agilent SureSelect Human All Exon V6
Total mapped reads (millions)	Tumor 99.7 - 179.4 Normal 79.1 - 157.7	30.1 - 55.5
Median insert size distribution (bp)	Tumor 104 - 145 Normal 126 - 164	214 - 268
Median coverage (over enriched areas)	Tumor 71 - 129 x Normal 51 - 155 x	21- 44 x
Percentage bases covered > 30 x (%)	Tumor 90 - 94 Normal 81 - 94	33 - 68

Percentage of reads that are duplicates (%)	Tumor 10.6 - 14.4 Normal 13.9 - 21.5	20.2 - 41.9

Table 1. Quality control summary metrics after pre-processing of data.

Pipeline establishment for short variant discovery

When examining the number of variants called in tumor-normal and tumor-only mode, less variants were called in tumor-normal mode for all samples except one (VUC99). When classifying the normal samples by the genes with variants most commonly observed in bladder cancer, we observed that the samples also showed variants in these genes (Fig 3).

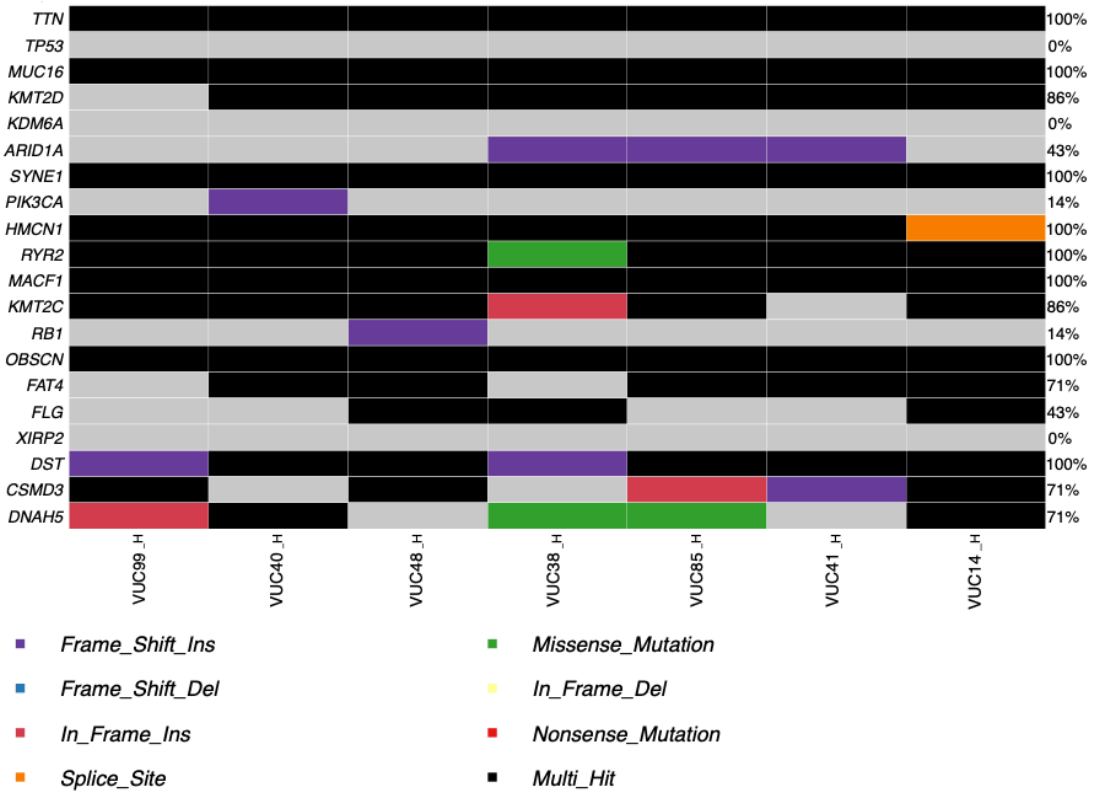


Figure 3. Oncoplot showing many common bladder cancer variants were present in the normal samples. Genes listed are the top 20 most commonly mutated genes in bladder cancer according to the TCGA database. Normal samples are listed along the bottom of the plot. Color corresponds to the variant type present in each of the normal sample. Percentages refer to the percentage of VUC models that have a particular variant present in that gene.

The results from GetPileupSummaries and CalculateContamination showed very low estimated contamination across all samples. When examining the number of variants classified as a 'PASS' after testing the 4 different pipeline variations, the most variants were called with the pipeline using the GATK provided panel of normals (PON) without the GetPileupSummaries and CalculateContamination steps. This was true for all of the samples tested. Using the created PON

and hard filtered resulted in less variants being called (Fig 4). When examining the output from CalculateContamination the estimated contamination across the samples was close to 0.

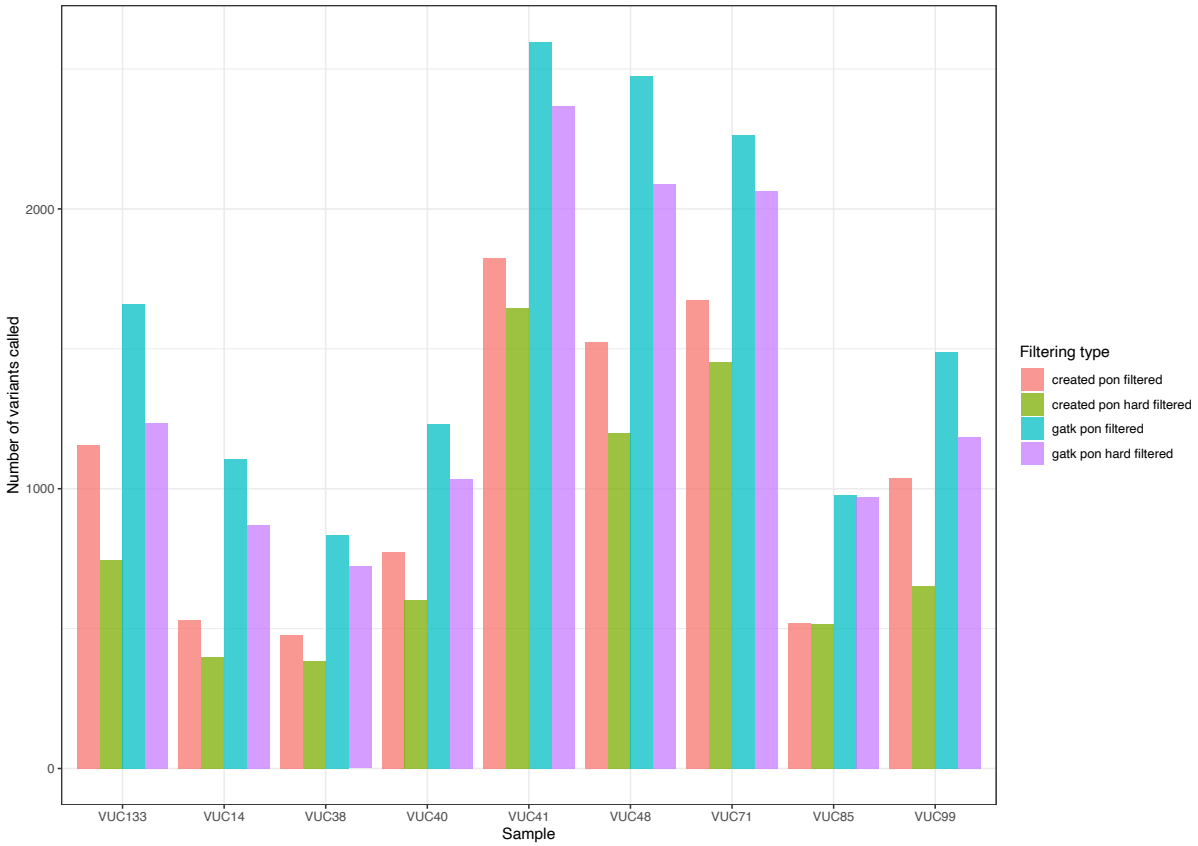


Figure 4. Comparing the number of variants called using different processing steps and panel of normals (PON). Created pon = PON created from the 7 healthy samples. GATK pon = PON available from GATK. Hard filtered = processed with the addition of GetPileupSummaries and CalculateContamination steps. Filtered = processed without the addition of GetPileUpSummaries and CalculateContamination steps.

Short variant calling

The initial output for both batches which had been run in tumor-only mode demonstrated that most variants were classified as missense mutations and just affected a single nucleotide polymorphism (SNP). When examining the single nucleotide polymorphism, the most common nucleotide change was from a C to a T for both batches (Fig 5 and Fig 6). Although there were many similarities in the initial output graphs between the batches there was some differences observed. The results for the CeMM samples showed the second most common variant classification was a frame shift insertion, a variant classification that was barely observed in the Novogene processed samples. Similarly, more diverse variant types were observed the CeMM when compared to the Novogene batch where practically all variant types recorded were SNPs.

The median number of variants per sample was a lot higher at 940 for the CeMM samples and 68 variants per sample for Novogene samples. Similarly, the top 10 most frequently mutated genes were different between the sample batches, with the most frequently mutated gene being *TTN* for the CeMM batch (Fig 5) and *MUC12* for the Novogene batch (Fig 6).

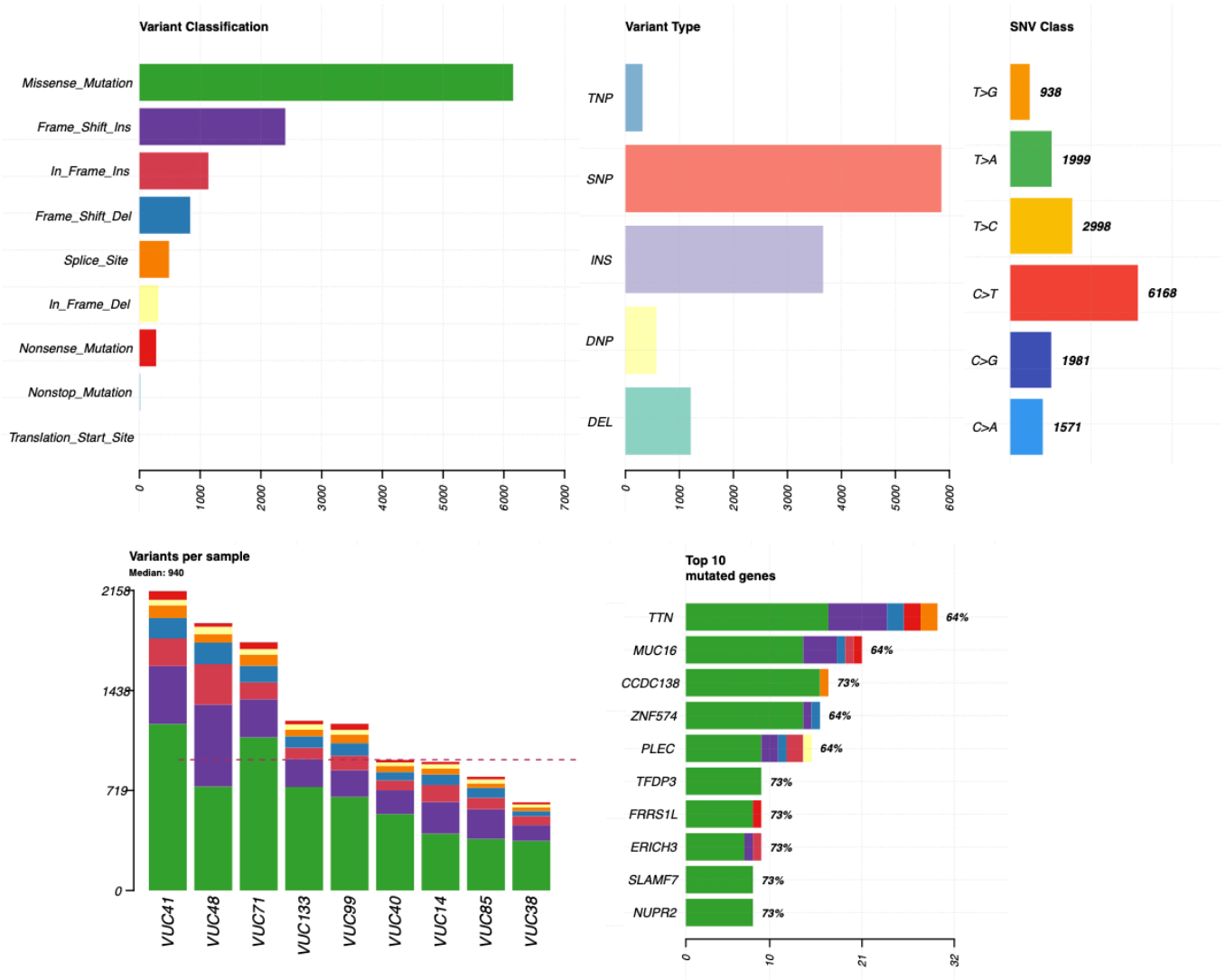


Figure 5. Initial output summaries for the variants detected in the CeMM processed models. Information about the variants is described by their classification, type, the SNV class and the number of variants per sample. The top 10 commonly mutated genes across all the samples, where the x axis refers to the number of mutations in the gene and the percentage refers to the percentage of samples in which this gene is mutated. TNP = tripe nucleotide polymorphism, SNP = single nucleotide polymorphism, INS = insertion, DNP = double nucleotide polymorphism, DEL = deletion and SNV = single nucleotide variant.

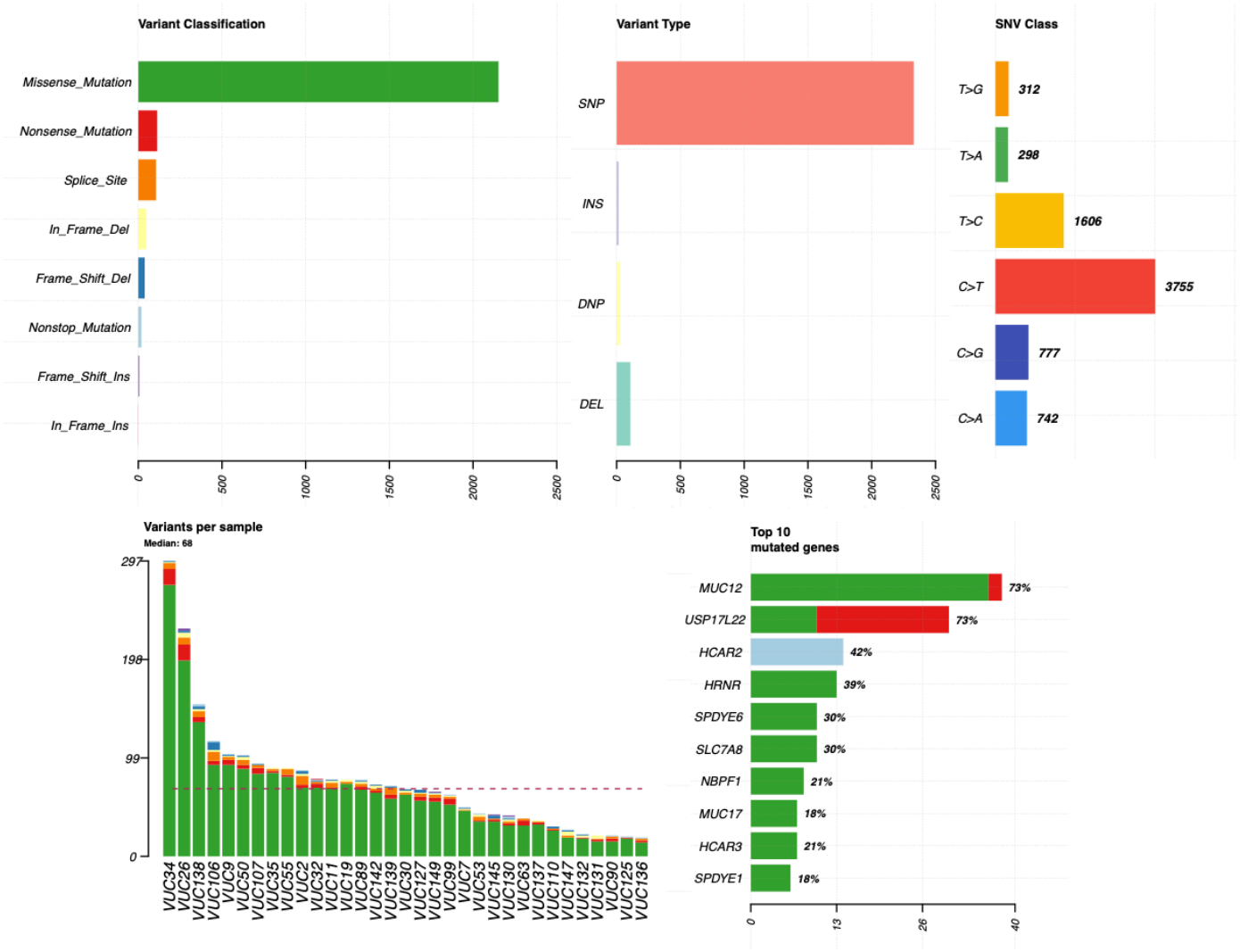


Figure 6. Initial output summaries for the variants detected in the Novogene processed models. Information about the variants is described by their classification, type, the SNV class and the number of variants per sample. The top 10 commonly mutated genes across all the samples, where the x axis refers to the number of mutations in the gene and the percentage refers to the percentage of samples in which this gene is mutated. TNP = tripe nucleotide polymorphism, SNP = single nucleotide polymorphism, INS = insertion, DNP = double nucleotide polymorphism, DEL = deletion and SNV = single nucleotide variant.

Summary of short variant discovery results

When both batches were classified by the most common top 20 variants for bladder cancer as defined by TCGA database (TCGA Research Network, https://www.cancer.gov/tcga), we observed more of these variants present in the CeMM processed samples. All 9 of the VUC models processed at CeMM had variants commonly seen in bladder cancer (Fig 7), whereas only 17 of the 33 VUC models had variants in genes commonly observed (Fig 8). Interestingly, no variant in *FLG* was observed in the VUC models processed at CeMM but was observed in 4 models

processed at Novogene. When examining all of the VUC models together no variant was observed in the gene *PIK3CA* (Fig 7 and Fig 8).

Encouragingly, there were some similarities between models processed at the different sequencing centers, for example for variants observed in the gene *TTN,* where samples have a multi-hit (meaning that there were numerous variants observed in that gene). When examining the VUC99 model, which had been sequenced at both sequencing centres, the same variants in the genes *RYR2* and *KMT2C* were observed in both batches (Fig 7 and Fig 8). The variants were identified as missense variants and were from the same position in the genome. For the CeMM processed sample the variant allele fraction detected was 0.18 for *RYR2* and 0.278 for *KMT2C*. For the Novogene processed samples the variant allele fraction was 0.117 *RYR2* and 0.271 for *KMT2C*.

For all the VUC models, the clinical Tumor-Node-Metastasis (TNM) staging system was also added where the information was available (Fig 7 and Fig 8). T refers to the size and extent of the primary tumor, N refers to the regional lymph node involvement and M if the cancer has metastasized (Kandori et al., 2019). With each of the classifications, a number is assigned to give more information about the cancer, the higher the number the more advanced the cancer is. The T categories are Tx (no information about the primary tumor or it cannot be measured), T0 (no evidence of a primary tumor), T1, T2, T3 and T4 (all describes the tumor size which increases with the number). The N categories are Nx (no information on the nearby lymph nodes or it cannot be measured), N0 (nearby lymph nodes do not contain cancer), N1, N2, N3 (all describe the size and number of lymph nodes affected). Finally, the M category is divided into two M0 (no metastasis has been found) or M1 (metastasis has occurred).

For the models that had been sequenced at CeMM (n=9), 7 models were classified as T2 and 2 models as T4. Only 1 model had affected the lymph nodes classified as N2 and none of the models had metastasized (Fig 7). For the models that had been sequenced at Novogene (n=33), the 17 that had variants in commonly mutated bladder cancer genes and had TNM annotation, 4 of the 12 were described as T3 and above. 4 of the VUC models had an advanced N staging classification

demonstrating the lymph nodes had been affected and only 1 of the models had metastasized (Fig 8).

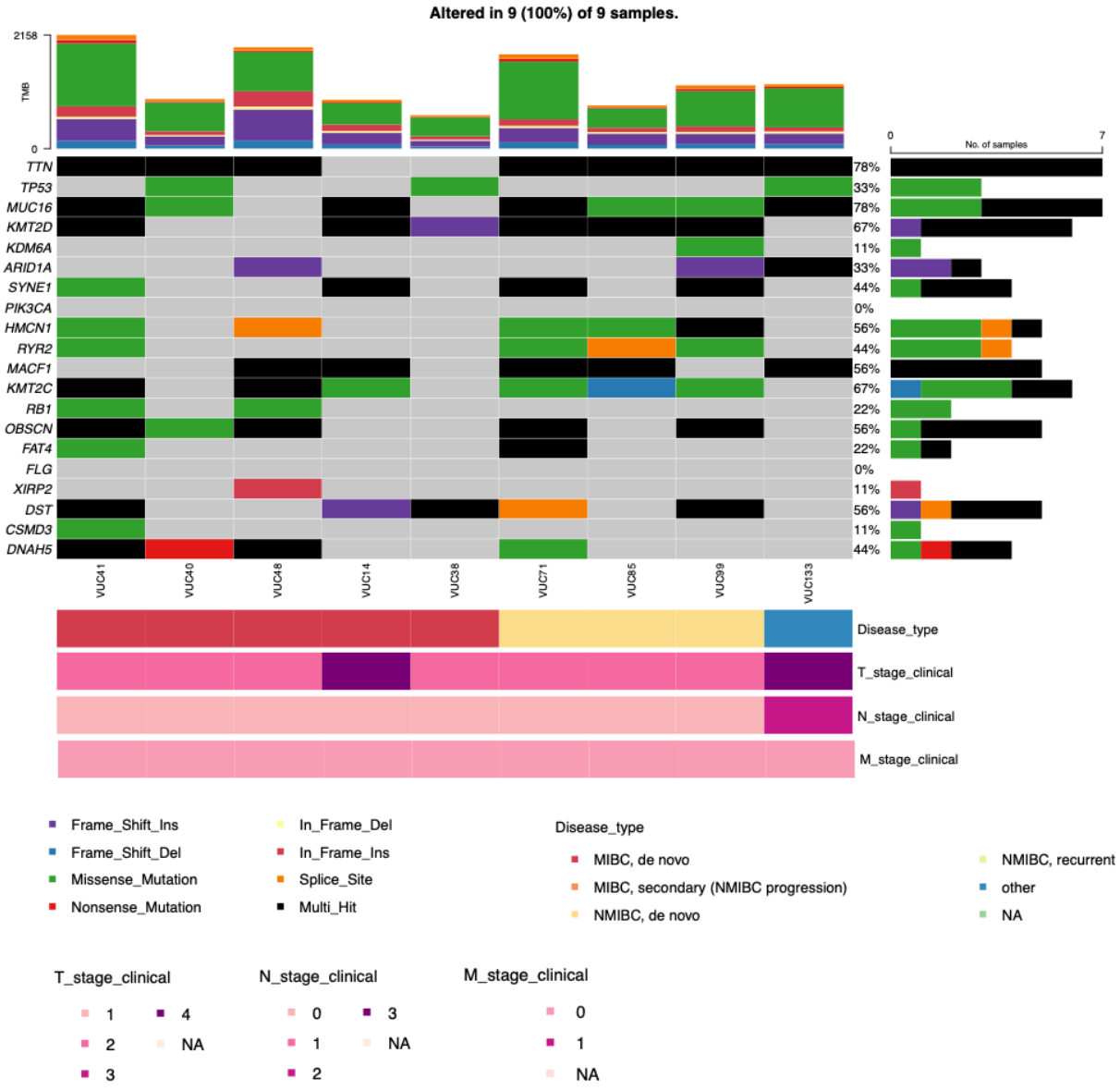


Figure 7. Oncoplot for all the CeMM processed models classified by the top 20 most commonly mutated genes in bladder cancer as defined by the The Cancer Genome Atlas Program (TCGA) dataset. Genes are ordered by how common they are as defined in the TCGA dataset, models are grouped by disease type. TMB = tumor mutation burden, MIBC = muscle-invasive bladder cancer, NMIBC = non-muscle-invasive bladder cancer, de novo = disease has arisen independently in the individual, secondary = disease has arisen from another tumor type, other = disease is neither NMIBC or MIBC, NA = information not available. T stage clinical = tumor clinical stage, 1 indicates the smallest tumor size and 4 indicates a larger tumor size. N stage clinical = regional lymph node involvement, 0 = nearby nodes do not contain cancer, 1 indicates the least number of lymph nodes have been affected and 3 indicates more lymph nodes have been affected. M stage clinical = metastasis clinical stage, 0 = no metastasis detected, 1 = metastasis detected.

Figure 8. Oncoplot for all the Novogene processed models classified by the top 20 most commonly mutated genes in bladder cancer as defined by the TCGA dataset. Genes are ordered by how common they are as defined in the TCGA dataset, models are grouped by disease type. TMB = tumor mutation burden, MIBC = muscle-invasive bladder cancer, NMIBC = non-muscle-invasive bladder cancer, de novo = disease has arisen independently in the individual, secondary = disease has arisen from another tumor type, other = disease is neither NMIBC or MIBC, NA = information not available. T stage clinical = tumor clinical stage, 1 indicates the smallest tumor size and 4 indicates a larger tumor size. N stage clinical = regional lymph node involvement, 0 = nearby nodes do not contain cancer, 1 indicates the least number of lymph nodes have been affected and 3 indicates more lymph nodes have been affected. M stage clinical = metastasis clinical stage, 0 = no metastasis detected, 1 = metastasis detected.

Variant allele fractions of models

For all the VUC models the variant allele fraction (VAF) for the most commonly mutated genes in bladder cancer was defined. Generally, the median VAF observed was below 0.2 with a few above

that value. When comparing the two batches a high penetrance was observed for variants seen in the gene *TP53* (Fig 9).

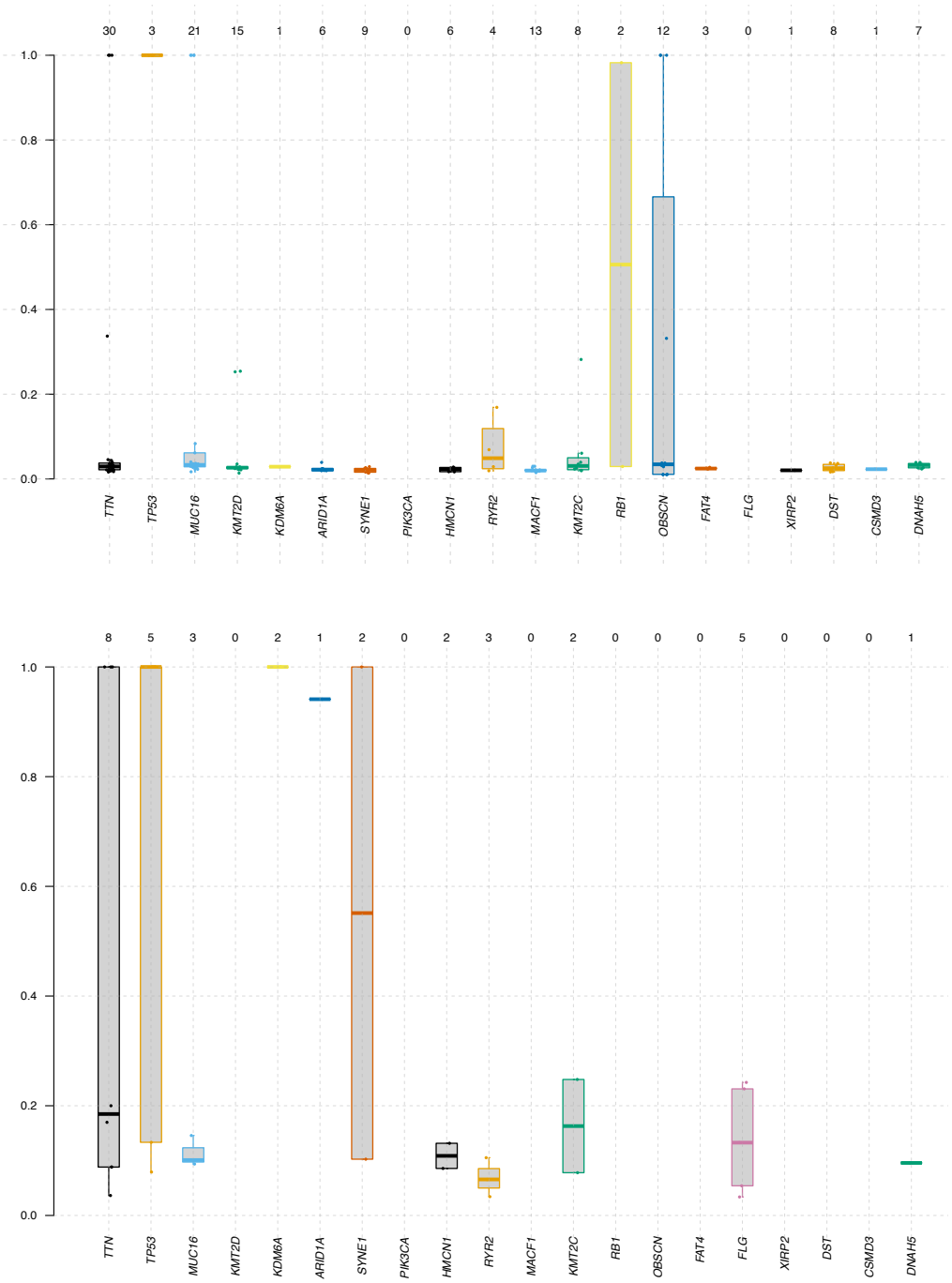


Figure 9. Variant allele fraction (VAF) boxplots, CeMM processed models (top) and Novogene processed models (bottom), for the top 20 most common gene variants in bladder cancer as defined by the TCGA database. Number above denotes how many variants are identified in that gene across all the models in that batch.

Comparing variant calling results and sequencing oncopanel results

Previously in the clinic, 4 of the patients had undergone an sequencing oncopanel for the purposes of genetic counselling to aid diagnosis and treatment. Here, data is expressed as the variant allele frequency, which is the same as the variant allele fraction, but expressed in percentages instead of a fraction. The outcome of the identified variants is summarised for 4 VUC models that were available alongside if the variant was identified by the variant caller and what it was identified as (Table 2). The oncopanel results were entirely in agreement with the outputs from the variant caller for the models VUC107 and VUC142. For the model VUC107, the variant allele frequency reported by the variant caller was higher for all gene variants than reported in the oncopanel. Interestingly, for VUC107 the variants identified as somatic by Mutect2 were the two variants with the highest reported variant allele frequency and the other three variants were determined as germline. For VUC50, the variant caller Mutect2 only suggested the germline variant in the *MUTYH* but detected no variants in the other genes reported in the oncopanel. None of the genes reported in the oncopanel for the model VUC136 could be identified by the variant caller. For VUC142, no genes were reported in the oncopanel, which agreed with the results by Mutect2.

Model	Genes reported in the oncopanel	Variant allele frequency (%) reported in the oncopanel	Identified by Mutect2	Variant allele frequency (%) reported by Mutect2
VUC50	*TP53* *MUTYH* (germline) *CTCF* *RET* *SETD*	59.6 58.8 37.1 19.2 25.9	No Germline No No No	NA 55 NA NA NA
VUC107	*PIK3CA* *CDKN2A* *KRAS* *TP53* *MYCN*	52.2 68.9 53.4 73.4 34.2	Germline Somatic Germline Somatic Germline	72 95.5 94 96.8 40
VUC136	*ATM* *ERBB2* *MDM2* *MEN1* *TOP1*	9.9 12.1 13.5 27.3 6.5	No No No No No	NA NA NA NA NA
VUC142	None	NA	NA	NA

Table 2. Summarising reported gene variants and their corresponding variant allele frequency as identified in a clinical oncopanel and by Mutect2 variant caller. *Variant allele frequency is described as a percentage.*

Copy number alterations (CNA) calling

CNA profiles were generated from the WES data for all VUC models (n=42). When examining the Novogene samples (n=33), a number of samples showed a flat copy number alteration plot. Additionally, the 16 VUC models (38%) that did not contain any variants in genes most commonly found in bladder cancer showed a flat copy number alteration profile. 3 of those models also had existing array-based comparative genomic hybridisation (aCGH) which also showed a similar flat CNA profile (for example see, Fig S6). To validate these findings, the overall CNA profiles were compared to existing aCGH (n=9) (for example see, Fig S5). For all models where aCGH data existed the CNA profiles generated were in general agreement. For VUC99, which had been sequenced at both sequencing centres and had an aCGH profile, all CNA profiles were similar and in agreement with each other (Fig 10). Despite creating the reference for the analysis from combining the tumor samples, the results were comparable to results generated by aCGH which uses diploid reference DNA material from a healthy male.

After examining the overall CNA profiles, the CNA of individual genes were investigated. Firstly, if any of the common top 20 bladder cancer TCGA variants were reflected in the CNA results, to examine if the variants observed had an effect on the detected CNA of genes. The only one of these 20 genes that showed an alteration in the copy number was *FLG.* Where an amplification was observed for *FLG* and the nearby gene *FLG-AS1* (Fig S3).

When examining the generated heatmaps showing large scale CNA in gene segments across the VUC models, many models showed whole chromosome or chromosome arm with extreme amplifications and losses including VUC133, VUC38, VUC40, VUC48, VUC71, VUC99, VUC107, VUC138, VUC26 and VUC34 (Fig 11 and Fig S4). When examining chromosome regions and individual genes, one CNA seen across all the VUC models was a loss of one of the *HLA* Class I genes, including VUC48 (Fig 12). Other common CNA seen in many of the models were *ASNS* amplification, *HRNR* amplification, loss of a member of the *GUSBP* gene family and loss of various members of the *MUC* gene family (Fig 12 and Fig S4).

When examining the CNA profile of VUC48, which showed large sections of amplification and deletion, a particularly interesting area of high gene amplification is on chromosome 17 where a number of genes including *ERBB2* are amplified, threshold log2 ratio > 3 (Fig 12).

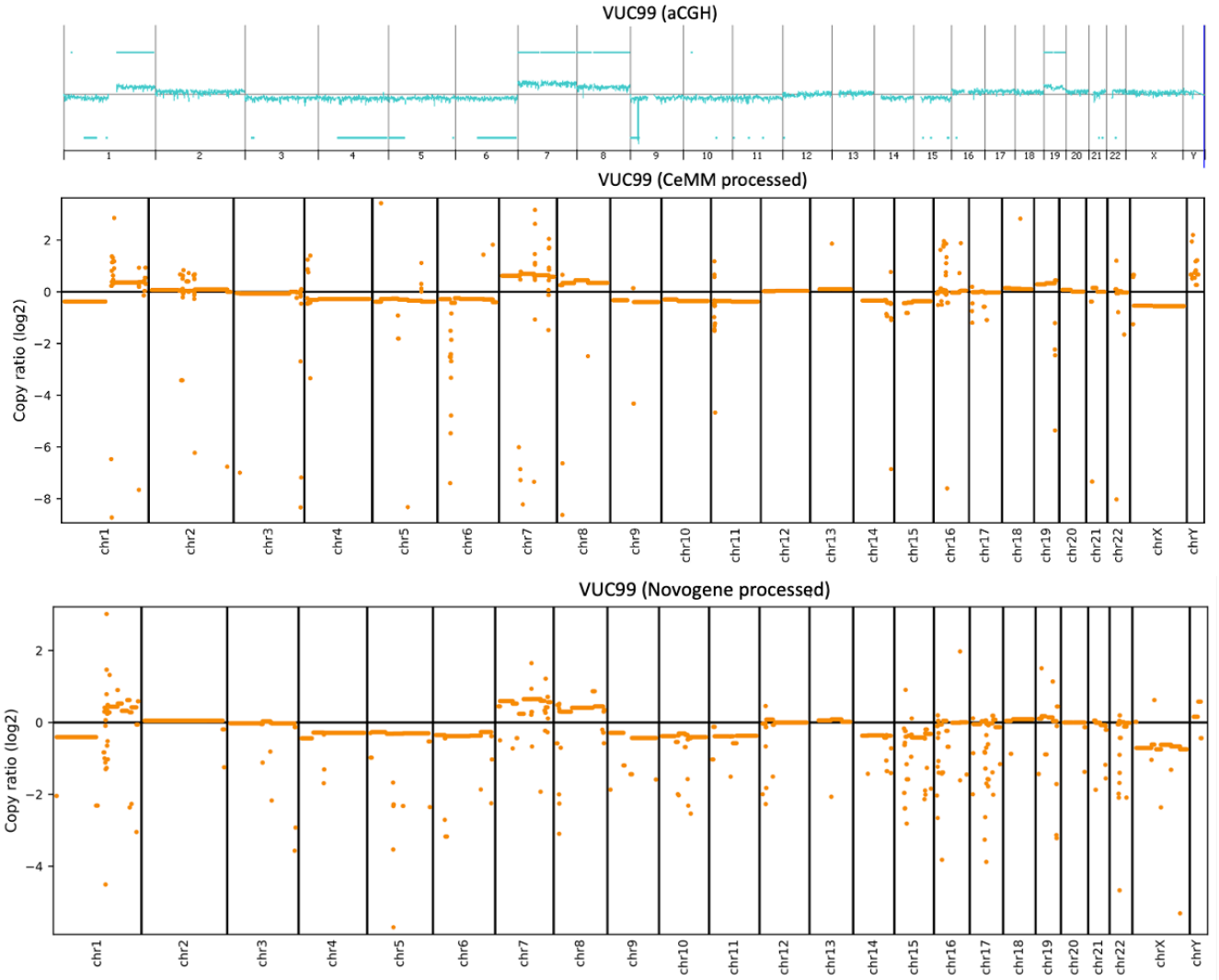


Figure 10. Array comparative genomic hybridisation (aCGH) and copy number alteration (CNA) profiles for VUC99 CeMM and Novogene processed. *CNA profiles showing copy number segments across the chromosomes.*

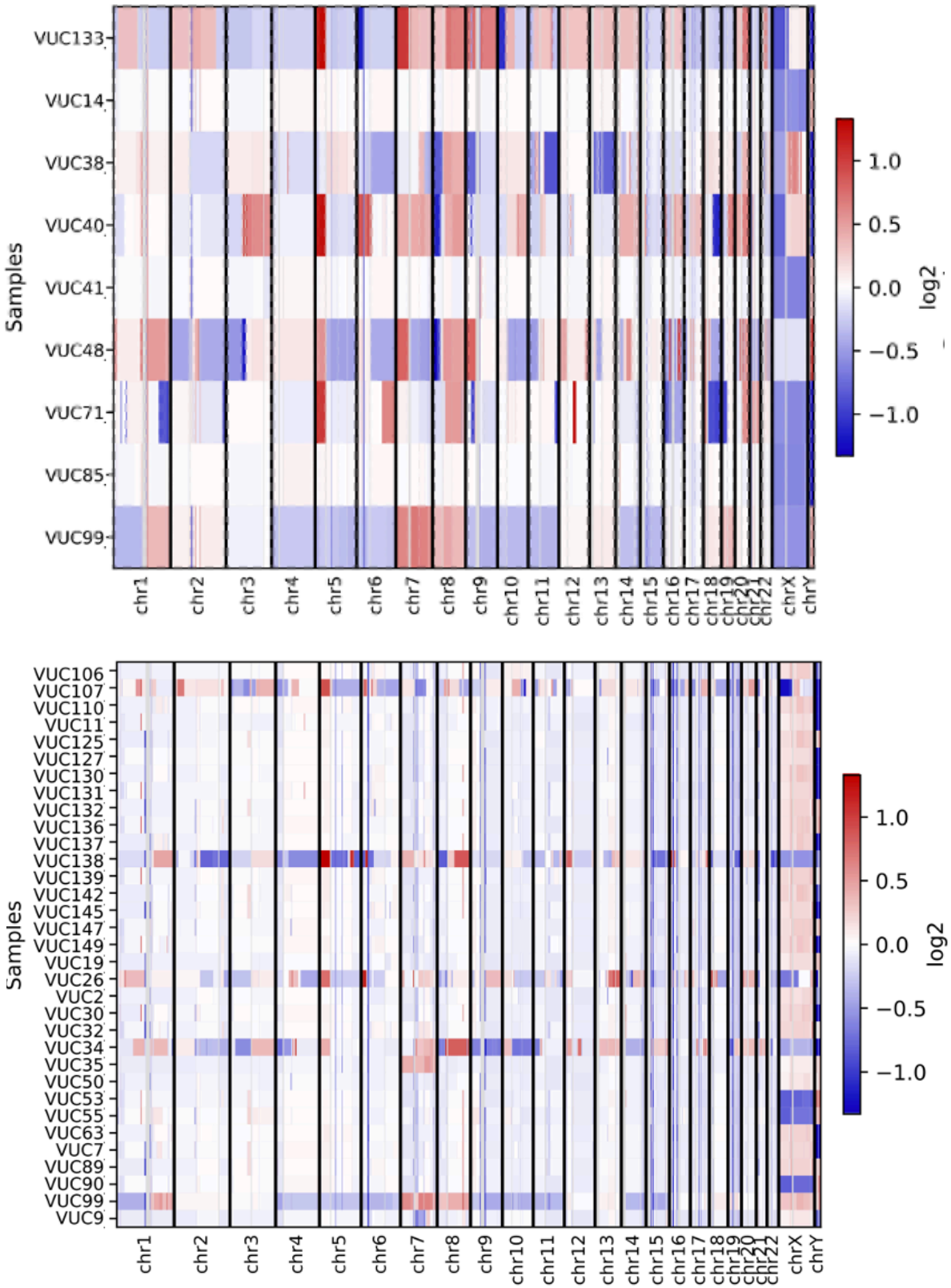


Figure 11. Heatmap summaries for all VUC models showing copy number alterations (CNA) across the chromosomes. *CeMM processed samples (top) and Novogene processed samples (bottom). Red indicates amplification or gain, and blue indicates deletion or loss.*

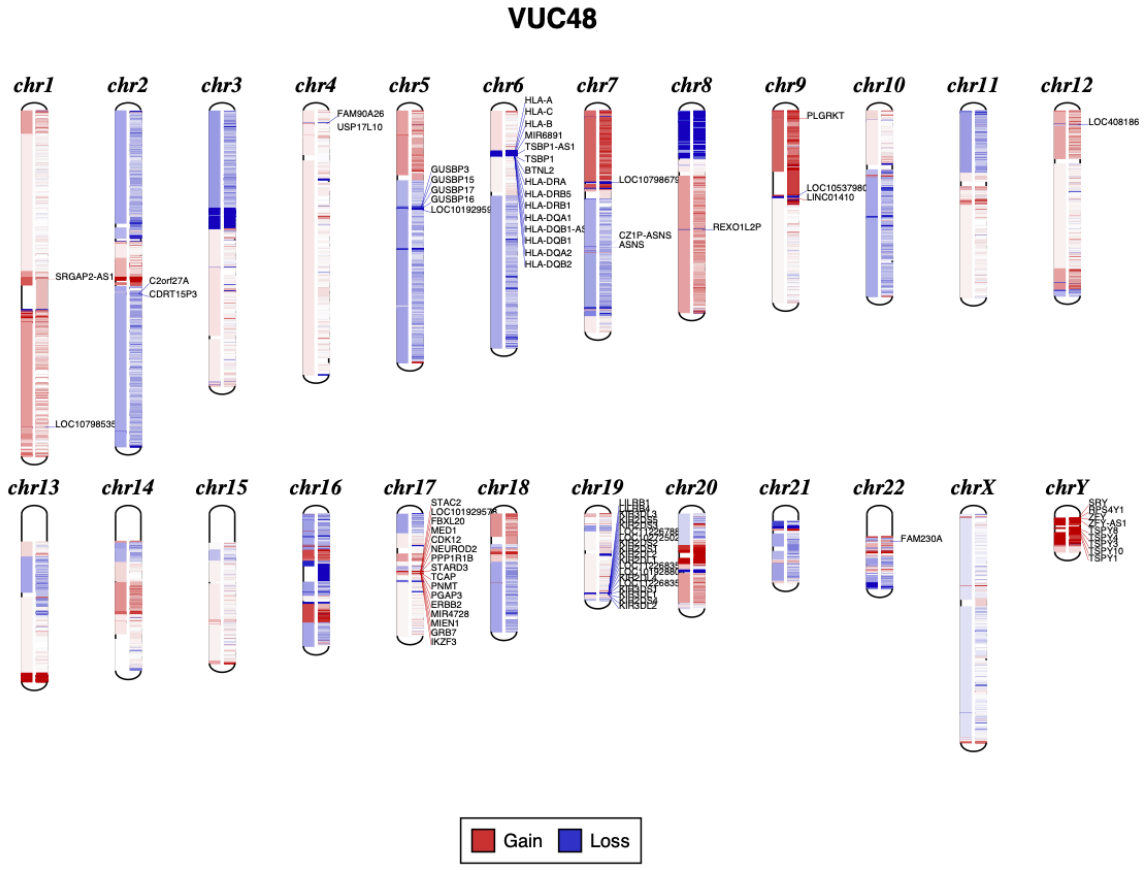


Figure 12. Ideogram of VUC48 chromosomes showing extreme amplification/gain and deletion/loss. Red indicates amplification/gain and blue indicates deletion/loss and threshold log2 ratio >3. Only genes above this threshold have been labelled. Segmentation calls are shown on the left side of the chromosome and bin-level log2 ratios are shown on the right side of the chromosome.

Discussion

Quality of sequencing data

The high percentage of reads that align to the reference genome observed was expected due to the fact that exome sequencing only focuses on 2 % of the human genome, therefore only a very small portion of the genome has been sequenced (EL et al., 2014). This result also demonstrated the short read aligner BWA-MEM is a robust appropriate tool for this purpose. A benchmarking paper comparing mapping efficiencies of BWA-mem, Bowtie2 and Stampy aligners found BWA-mem to have the highest mapping efficiency and a lower number of misaligned reads (when compared to Bowtie2) (Zanti et al., 2021).

There were clear differences in the sequencing methodology used to process the samples at the two different sequencing centers. The right shift in GC content observed in the Novogene

processed samples was abnormal and not the expected distribution typically seen for exome sequencing. This results in a bias as there was a limited sensitivity to AT rich regions and more sensitivity to GC rich regions potentially affecting downstream variant calling. Typically, whole exome sequencing (WES) libraries have a higher GC content (44-54%) than WGS (40-43%) (Xiao et al., 2021). This is due to exomes generally having a higher GC content than the whole genome due to the functional restrictions and selective pressures on protein-coding regions. A higher GC content can stabilise regions of DNA and reduces the likelihood of mutations in these regions, resulting in exonic regions being conserved over periods of evolutionary time. Furthermore, regulatory regions of DNA found outside of exonic regions, which play a crucial role in regulating gene expression, called promoter regions are AT-rich (Tome et al., 2018)

Two different exome capture kits were used to process the samples. The Twist Human Core Exome + Refseq Panel is a double stranded DNA based bait kit, whereas the Agilent SureSelect Human All Exon V6 is a single stranded RNA based bait kit. It has been previously demonstrated that RNA baits, when compared to DNA baits have different binding characteristics and perform better in areas with a higher AT content region (Zhou et al., 2021).

When calling variants, the analysis was restricted to the padded intervals given by the kit manufacturers, plus 450 bp either side to ensure as many variants could be captured as possible in our regions of interest with enough coverage to confidently make the calls. Calling variants on a genome-wide scale without using the intervals would increase the number of variants called but the computational time needed as well as the number of false positives would also increase (Corominas et al., 2022).

The length of genomic DNA sequenced as paired-end reads and flanked by adapters is defined as the insert size. The insert sizes recorded were larger for the Novogene processed samples. A previous methodological study compared different insert sizes to see if it influenced the uniformity of read coverage within targeted interval regions for WES. The study found that increasing the insert size improved the evenness of read coverage (Pommerenke et al., 2016).

There were clear differences between the total number of mapped reads recorded for the samples between the two batches. A previous study examining best practices in cancer mutation detection for WES, demonstrated a higher total number of mapped reads resulted in a higher read coverage. This finding was also supported by the WES data in this study, where a difference in the median coverage between the two sample batches was observed. The recommended coverage for exome sequencing to be able to confidently call variants is 90 x – 100 x, to account for any uneven coverage across the exome (Zverinova & Guryev, 2022). Unfortunately, this meant that the coverage for the Novogene processed samples fell short of what is typically required to confidently call variants.

There was a higher percentage of duplicate reads observed in samples processed at Novogene. Duplicates are a result of PCR amplification during sample preparation for sequencing and were identified as such as not to introduce false positives during variant calling (Corominas et al., 2022; Ebbert et al., 2016). Taking into consideration all the above discussed methodological differences, it was difficult to directly compare the two batches to one another. Furthermore, it could not be ruled out that the subsequent variant calling and copy number alterations results were impacted by methodological differences between the sequencing centres.

Selection of Mutect2 as a variant caller

Mutect2 was the tool selected to do the short variant calling for a number of reasons. First of all, the algorithm was specifically designed by the Broad Institute for the purpose of detecting somatic mutations from sequencing data making it suitable for the data type concerned in this project (Lewis & Moffett, 2021). Secondly, there are numerous published cancer studies that have used the tool successfully to detect somatic mutations (Chen et al., 2020; D'Agostino et al., 2022; Goel et al., 2022; Wilkinson et al., 2021). Thirdly, when examining existing literature that has benchmarked different variant callers against one another, Mutect2 has been reported to have high specificity and sensitivity (Callari et al., 2017). Due to the popularity of the tool, there is extensive literature, manuals and resources available online to be able to easily run the tool. Finally, Mutect2 is an open-source tool and the source code is publicly available and therefore can be readily integrated into a number of different workflows including other tools in the GATK pipeline (Benjamin et al., 2019). It was important that all of the samples were processed along

the same bioinformatic pipeline to ensure consistency in the way they were processed. As most of the samples (33/42) did not have a matched normal, other variant callers such as Strelka2 (Kim et al., 2018) and VarScan2 (Koboldt et al., 2012) could not be used as they do not have a tumor-only variant calling pipeline.

Briefly, Mutect2 variant calling pipeline works by assembling haplotypes, somatic genotyping followed by filtering. The assembling haplotypes step is based on the local assembly code used by the GATK HaplotypeCaller (Poplin et al., 2018) which creates a de Bruijn-like graph from kmerized reads but Mutect2 then carries out adaptive pruning of the possible paths, which is particularly useful for samples with a variable read coverage across the genome as seen in WES. Somatic genotyping works by implementing the Pair Hidden Markov Model (Pair-HMM) probabilistic model, assigning for each read a likelihood for the tumor sample (and normal sample if present), converting this matrix to a fragment-haplotype matrix. Finally, filtering by the GATK FilterMutectCalls tool estimates the probability the variant is somatic and not a germline or sequencing artefact (present in the panel of normals). Additionally, the tool filters out variants with low-quality supporting bases, poor mapping quality and a large mismatch between variant and reference supporting fragments (Benjamin et al., 2019).

Pipeline testing for variant calling

It is recommended to use normal samples, where available in a short variant discovery bioinformatic pipeline, to identify germline variants and to improve the precision of Mutect2's variant calling (Teer et al., 2017). Inconsistency in the total number of variants called in the tumor-normal pipeline, alongside the fact that the normal samples showed evidence of variants commonly found in bladder cancer, indicated the samples could not be considered as completely healthy. Despite normal samples appearing healthy pathologically and histologically, genetically they showed variants present in tumor DNA. This could be due to tumor cells attaching themselves to other areas in the bladder (Khong & Restifo, 2002). Considering these results and that there were only normal samples for only 7 of the 42, it was decided to call variants in tumor-only mode, to ensure the analysis would be comparable between the samples. Studies that have compared results from tumor-normal and tumor-only variant calling pipelines having demonstrated that without a normal there is a higher number of false positives (Teer et al., 2017).

Whether the pipeline is run in tumor-normal or tumor-only mode, a panel of normals (PONs) is always required. A PONs was used to attempt to remove any recurrent technical sequencing artifacts as recommended by the GATK best practices guidelines. The PON should be made up of normal i.e. healthy tissue or blood which is free of somatic alterations, the general guidance is that at least 40 samples should make up the PON. The public GATK PON is generated from exome and whole genomes generated from 1000 Genomes Project samples (Lewis & Moffett, 2021). Although a PONs from the 7 normal samples was generated and tested, the total number of variants called was always less when compared to the GATK PON. One possible reason for this is the sample size of 7 was not large enough to compile a suitable reference panel, one study previously reported 50 normal samples is optimal to identify sequencing artifacts (Koboldt, 2020). Another possible reason could have been the above-mentioned doubts on how healthy the normal samples were, with many of the normal samples contained variants in common bladder cancer genes, resulting in many of these variants being wrongly identified as technical sequencing artifacts.

Using the additional filtering steps of GetPileupSummaries and CalculateContamination, gave a reduced number of total variants detected. Furthermore, the estimated contamination calculated by these steps was very low. Considering all the above information, the pipeline using the GATK provided PON without using the additional GetPileupSummaries and CalculateContamination steps was used for further analysis, however one must bare in mind that with a less stringent pipeline the number of false positives called as a variant also increased.

Examining short variant calling results

The vast majority of variants identified were classified as missense mutations affecting a single nucleotide base, which most often resulted in a change from a C > T. These results have been observed previously in other bladder cancer genetic studies (Hurst et al., 2017; F. Wang et al., 2022). Historically, the most common single nucleotide change observed in human cancer cells has been a cytosine to thymine transition. This is due to the hydrolytic deamination of cytosine and cytosine analogs, resulting in a mispaired intermediate of guanine and later on then a repair of the original lesion as a C >T (Harris, 2013; Hsu et al., 2022). The total number of variants

detected in the different VUC models varied greatly, which is in agreement with the known complex heterogeneity of bladder cancer (Robertson et al., 2017).

Some variants initially identified in the top 10 mutated genes such as *TTN, MUC16*, *PLEC* and *MUC17* had clear links to bladder cancer and had been reported in past studies, however, many of the other genes reported had no obvious link to bladder cancer. Some genes reported such as *NBPF*, *HCAR2* and *TFDP3* have been annotated and associated with different cancer types but not bladder cancer in the GeneCards database (Stelzer et al., 2016). Other genes such as *HRNR* and *SLAMF7* have been associated with immune response pathways (Safran et al., 2021). The presence of variants in genes involved in these pathways is not surprising, given the tumor microenvironment the VUC models are generated from contains malignant and non-malignant cells including immune and stromal cells (Tran et al., 2021). However, a few genes such as *CCDC138* and *USP17L22* have little annotation on their function in the GeneCards database, suggesting they could have been artifacts.

When examining some of these variants more closely, for example the hydroxycarboxylic acid receptor 2 *HCAR2* deletion, found in 42 % of the Novogene processed samples, the variant was exactly the same for each one. Resulting in a removal of the stop codon at the protein level p.P363fs. *HCAR2,* has previously been described as a tumor suppressor gene in breast cancer. Interestingly, one study examining the role of HCARs in breast cancer identified a frame shift at the same amino acid position (McGuire Sams et al., 2021). When examining the other VUC models processed at CeMM, only VUC48 had a variant in *HCAR2* at a different position. Whilst it is not impossible that the *HCAR2* variant observed in the Novogene processed models was linked to the tumor, the fact that it was seen in 42 % of the samples seems unlikely and suggests it is perhaps a sequencing artifact.

Since the variant caller was ran in tumor-only mode and therefore the results were assumed to contain many false positives and sequencing artefacts, it was decided to classify the VUC models by known bladder cancer gene variants as described by The Cancer Genome Atlas (TCGA) (TCGA Research Network, https://www.cancer.gov/tcga). TCGA is a database designed to molecularly characterize 33 cancer types using genomic, epigenomic, transcriptomic, proteomic and imaging

data. Although some variants in these genes were observed in our datasets, some variants in key expected genes were not detected or they were detected but not in as many VUC models as expected. Some possible reasons for this could be the difference in the size of the datasets, our dataset had 42 VUC models in total, whereas the TCGA bladder cancer dataset had 412 samples. Additionally, there were key differences on the data sources and how results were generated. The TCGA bladder cancer dataset, was from primary, untreated tumors with a matched normal and/or blood samples (Weinstein et al., 2014). The WES datasets presented here were from primary patient derived cell models and although a few of the models had a matched normal sequenced, the variant calling was carried out using the tumor-only pipeline for calling. Primary patient derived cell models are commonly monocultures with one cell type population, resulting in less diversity in the genetic composition present (Richter et al., 2021). Furthermore, the TCGA bladder cancer dataset samples were composed of at least 60 % tumor nuclei with equal or less than 20 % necrosis (Robertson et al., 2017). For the primary patient derived VUC models, there was no information on the composition regarding cell type and composition once the models had been generated, which of course means the tumor nuclei composition could differ from the TCGA dataset standards. Together, these factors could account for the differences in the observed between the TCGA dataset and our WES datasets.

When examining the results of the classification of the VUC models by the TCGA bladder cancer genes, less variants were detected in the Novogene processed VUC models when compared to the CeMM processed samples. This can be explained by the higher sequencing coverage of the CeMM processed samples, as well as the GC content bias observed in the Novogene sequences. It has been previously demonstrated, that coverage in regions with extreme (< 30 % or > 70 %) GC content is worse (Q. Wang et al., 2017). This could result in less reads aligning to these regions in the genome and as a result less variants detected in these areas.

Previous work has demonstrated *TTN* and *OBSCN* are frequently mutated genes in cancer studies. This is mainly thought to be due to the large gene size, therefore mutations are more likely to occur as passenger mutations. However there is also some arguments that a fraction of these mutations are driver mutations (Greenman et al., 2007; Mendiratta et al., 2021). A previous study on bladder cancer has demonstrated chromatin remodelling genes (*KMT2D*, *KDM6A* and *KMT2C*)

are present in higher frequencies in non-muscle invasive bladder cancer (NMIBC) than muscle invasive bladder cancer (MIBC), something which was no observed in our datasets. However, this could be due to the cohort size used in our study. Furthermore, when examining TNM clinical staging, previous work has reported high-grade NMIBC and MIBC have a higher number of somatic variants (Goel et al., 2022), an observation we did not clearly observe. For the CeMM dataset this could be due to the small cohort size used and for the Novogene dataset this is potentially caused by the low sequencing coverage.

Examining variant allele fraction results

Lower variant allele fraction (VAF) was observed in the two batches than expected or seen in the TCGA bladder cancer dataset. This is possibly due to the aforementioned differences in cohort size, sample type and cell type composition. It also appeared that the Novogene processed samples had a higher VAF, however this is because less overall variants were detected so the boxplot upper and lower quartiles are stretched out.

Oncopanel and variant caller outputs differ

Only two of the VUC models (VUC107 and VUC142) agree with the oncopanel results. The results for VUC136 disagree completely and the results for VUC50 only show the germline variant was reported. When examining the variant allele frequency reported in the oncopanel, only the variants with a higher frequency were reported by the variant caller Mutect2. This could be due to the much higher sequencing coverage performed for the oncopanel compared to the whole exome sequencing (WES).

For VUC107, three variants were identified as germline in genes *KRAS, PIK3CA* and *MYCN*. To understand why, one must understand how the variant caller assigns a variant as germline or somatic. Mutect2 models the probability that a variant is a germline event by subjecting reads to a diploid genotyping model (Benjamin et al., 2019), which essentially means any variant with a variant allele frequency close to 50 % is identified germline. When examining the likelihood that these identified germline variants were true germline variants, we discovered through literature research this was unlikely to be the case. A germline *KRAS* mutation has been associated with severe developmental disorders that cause phenotypic abnormalities including postnatal growth

deficiency, cardiac defects and cognitive impairment (Gremer et al., 2011). Additionally, the presence of a germline *PIK3CA* mutation in the general population is rare and when present results in childhood onset of tissue overgrowth, brain and vascular malformations and skeletal abnormalities (Cooley Coleman et al., 2023). A Germline *MYCN* mutation is also rare and is characterized by microcephaly, learning disabilities and limb malformations (Kato et al., 2019). Therefore, it seemed unlikely that these identified variants were indeed germline and had only been identified as such due to their variant allele frequency.

A further reason to account for the differences could be that the samples are from a different source, the oncopanel was carried out on a tumor biopsy, whereas the WES was carried out on primary patient derived cell models. It is well known that the composition of cells and subsequently the genetic profile detected varies as primary patient derived cell models are established (Kodack et al., 2017).

Selection of CNVkit tool to examine copy number alterations (CNA)

Copy number changes are structural changes in the DNA that result in gain/amplification or loss/deletion of sections of DNA (Bartha & Győrffy, 2019). CNVkit was selected as the bioinformatic tool to examine copy number alterations across the different VUC models. The tool is designed to examine genome wide copy number detection and visualisation from target and off-target DNA sequences. Although the off-target read coverage is not high enough to do variant calling effectively, these reads result in a very low coverage of the genome and provide effective information on copy number. One advantage over other similar tools is that it allows the user to create a reference from tumor samples. One disadvantage of this created reference is that in our case we were restricted by the small cohort we had as well as the nature of the samples being from tumor tissues, as the CeMM sample reference was created from 9 tumor samples and the Novogene sample reference was created from 33 tumor samples. The recommendation from the authors of CNVkit is to create the pooled reference from a large cohort of healthy samples for accurate copy number alteration detection (Talevich et al., 2016).

Examining copy number alterations (CNA) results

The observation of whole chromosome or chromosome arm amplification or losses is consistent with previous observations in other studies. We observed some chromosomes where this was more likely than others indicating there is a common mechanism across different VUC models, which was in agreement with past studies (Harbers et al., 2021). Equally, we observed VUC models with very few areas of the genome that showed CNA and/or CNA affecting only a few genes, a result also observed in the past (Hurst et al., 2017). One point to bare in mind is that using CNVkit on WES data results in many copy number alterations and some false positives might be present (Talevich, 2016).

Interestingly, the only TCGA common bladder cancer variant that resulted in an observable CNA was *FLG* amplification. One possible reason for this could be that the *FLG* gene contains tandem repeats within the coding regions of the genome and then as a result there is a greater chance of mutation occurring. A previous study demonstrated that *FLG* CNA changes the gene structure and function of tumor cells in gastric cancer (Yicheng et al., 2022). A common CNA observed was the deletion of members of the *HLA* gene family. The deletion of a member or members of this gene family could result in a reduction of the resulting protein or proteins expressed by the tumor cells. Across many different cancer types, it has been demonstrated thar tumor cells commonly lose normal HLA-I expression and subsequently there is a reduction in the ability to stimulate the T-cell response, which results in tumor cells evading immune attack (Gil-Julio et al., 2021). Quite a few VUC models had an amplification in the *ASNS*, which could result in an overexpression in ASNS protein. Several reports over the past years have indicated ASNS promotes cell proliferation, chemoresistance and metastatic behaviour (Chiu et al., 2020). Two further CNA commonly observed in the dataset was *HRNR* amplification and *GUSBP* gene family member or members deletion, which all had no obvious link to tumor progression in bladder cancer. However, these genes are located close to the centromeres, areas where the structure of centromere contains many tandem-repeats and therefore is more likely to be prone to CNA (de Lima et al., 2021).

Within the same tumor type, there is huge variation in the percentage of the genome that is altered, with some tumors having more than 30 % of their genome showing CNA (Harbers et al., 2021). A previous study observed CNA were found at more locations across chromosomes in

NMIBC than MIBC (F. Wang et al., 2022), our results partially agree with this observation with VUC71, VUC99 and VUC138 following this trend. However, for other NMIBC models such as VUC106 and VUC9 this pattern was not observed. One possible reason for this could be due to the aforementioned low coverage and GC content bias of the samples processed by Novogene or due to the limited size of our cohort.

Conclusions

Taking into account the oncopanel results, variants identified, aCGH data and CNA results, it was concluded 16 samples, all Novogene processed, comprised of a majority of non-malignant cells. Generating and establishing primary cancer cells from patient tissues, can be challenging and the success rate depends on a number of different factors. One research group compared the success of cancer cell line generation by tumor type, demonstrating on average only 26 % of attempts were successful (Kodack et al., 2017). One possible explanation for this is that after the tumor cell dissociation step a lot of non-malignant cells remain still, for example haemopoietic and stromal cells. Overgrowth of the cancer cells by fibroblast cells is often observed when establishing patient derived cell models because fibroblasts adapt extremely well to *in vitro* conditions (Miserocchi et al., 2017). In the case of the 16 VUC models, we cannot be certain from the WES which non-malignant cell type or types are present. However, fellow master student in the Englinger research group Neha Singh who analysed RNA-seq data from the same VUC models, found the same 16 VUC models showed RNA expression of gene signatures associated with fibroblast cell types.

In future, more work is needed to confirm the variants present across the various VUC models. Another variant caller or multiple variant callers could be used on the same datasets. One possible solution could be to use SeqMule an automated pipeline that integrates 5 alignment tools (BWA, Bowtie, Bowtie2, SOAP2 and SNAP) and 5 variant calling algorithms (GATK, SAMtools, VarScan 2, Freebayes and SOAPsnp). Additionally, the pipeline also has in built accessory programs to allow quality control. Discrepancies between the output of various programs are to be expected, it was reported the single nucleotide variant agreement across the 5 variant calling algorithms was just 57.4 % (Guo et al., 2015). Inevitably, to install and configure the pipeline to the point where it is

fully up and running would require a significant investment of time. However, the pipeline would help to confirm presence or absence of variants.

In terms of further experiments that could be carried out it would be advantageous to have healthy matched tissue or blood for all VUC models. Additionally, validation studies would need to be performed to help understand the impact of the variants that occur. For example, Sanger sequencing of interesting variants would be need to be carried out to confirm the presence of the variants. Furthermore, integrating the results and data from RNA-seq data from the same VUC models will undoubtably provide further insights into the genetic characteristics as well as multi-omic datasets existing from other studies.

References

Andrews.S. (2010). *FastQC: a quality control tool for high throughput sequence data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bartha, Á., & Győrffy, B. (2019). Comprehensive outline of whole exome sequencing data analysis tools available in clinical oncology. *Cancers*, *11*(11), 1–20. https://doi.org/10.3390/cancers11111725

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. *BioRxiv*, 861054. https://doi.org/10.1101/861054

Callari, M., Sammut, S. J., De Mattos-Arruda, L., Bruna, A., Rueda, O. M., Chin, S. F., & Caldas, C. (2017). Intersect-then-combine approach: Improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Medicine*, *9*(1), 1–11. https://doi.org/10.1186/s13073-017-0425-1

Chen, Z., Yuan, Y., Chen, X., Chen, J., Lin, S., Li, X., & Du, H. (2020). Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports*, *10*(1), 1–9. https://doi.org/10.1038/s41598-020-60559-5

Chiu, M., Taurino, G., Bianchi, M. G., Kilberg, M. S., & Bussolati, O. (2020). Asparagine Synthetase in Cancer: Beyond Acute Lymphoblastic Leukemia. *Frontiers in Oncology*, *9*(January), 1–10. https://doi.org/10.3389/fonc.2019.01480

Cooley Coleman, J. A., Gass, J. M., Srikanth, S., Pauly, R., Ziats, C. A., Everman, D. B., Skinner, S. A., Bell, S., Louie, R. J., Cascio, L., Patterson, W. G., Jones, J. R., Di Donato, N., Stevenson, R. E., & Boccuto, L. (2023). Clinical and functional characterization of germline PIK3CA variants in patients with PIK3CA-related overgrowth spectrum disorders. *Human Molecular Genetics*, *32*(9), 1457–1465. https://doi.org/10.1093/hmg/ddac296

Corominas, J., Smeekens, S. P., Nelen, M. R., Yntema, H. G., Kamsteeg, E. J., Pfundt, R., & Gilissen, C. (2022). Clinical exome sequencing—Mistakes and caveats. *Human Mutation*, *43*(8), 1041–1055. https://doi.org/10.1002/humu.24360

D'Agostino, Y., Palumbo, D., Rusciano, M. R., Strianese, O., Amabile, S., Di Rosa, D., De Angelis, E., Visco, V., Russo, F., Alexandrova, E., Salvati, A., Giurato, G., Nassa, G., Tarallo, R., Galasso, G., Ciccarelli, M., Weisz, A., & Rizzo, F. (2022). Whole-Exome Sequencing Revealed New Candidate Genes for Human Dilated Cardiomyopathy. *Diagnostics*, *12*(10), 1–14. https://doi.org/10.3390/diagnostics12102411

de Lima, L. G., Howe, E., Singh, V. P., Potapova, T., Li, H., Xu, B., Castle, J., Crozier, S., Harrison, C. J., Clifford, S. C., Miga, K. H., Ryan, S. L., & Gerton, J. L. (2021). PCR amplicons identify widespread copy number variation in human centromeric arrays and instability in cancer. *Cell Genomics*, *1*(3), 100064. https://doi.org/10.1016/j.xgen.2021.100064

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501. https://doi.org/10.1038/ng.806

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Kauwe, J. S. K., & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, *17*(Suppl 7). https://doi.org/10.1186/s12859-016-1097-3

EL, van D., H, A., Y, J., & C, T. (2014). Ten years of next-generation sequencing technology.

Trends in Genetics : TIG, *30*(9), 418–426. https://doi.org/10.1016/J.TIG.2014.07.001

Ertl, I. E., Lemberger, U., Ilijazi, D., Hassler, M. R., Bruchbacher, A., Brettner, R., Kronabitter, H., Gutmann, M., Vician, P., Zeitler, G., Koren, A., Lardeau, C. H., Mohr, T., Haitel, A., Compérat, E., Oszwald, A., Wasinger, G., Clozel, T., Elemento, O., … Shariat, S. F. (2022). Molecular and Pharmacological Bladder Cancer Therapy Screening: Discovery of Clofarabine as a Highly Active Compound. *European Urology*, *82*(3), 261–270. https://doi.org/10.1016/j.eururo.2022.03.009

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Gandawijaya, J., Bamford, R. A., Burbach, J. P. H., & Oguro-Ando, A. (2021). Cell Adhesion Molecules Involved in Neurodevelopmental Pathways Implicated in 3p-Deletion Syndrome and Autism Spectrum Disorder. *Frontiers in Cellular Neuroscience*, *14*(January). https://doi.org/10.3389/fncel.2020.611379

Gil-Julio, H., Perea, F., Rodriguez-Nicolas, A., Cozar, J. M., González-Ramirez, A. R., Concha, A., Garrido, F., Aptsiauri, N., & Ruiz-Cabello, F. (2021). Tumor escape phenotype in bladder cancer is associated with loss of HLA class I expression, T-cell exclusion and stromal changes. *International Journal of Molecular Sciences*, *22*(14). https://doi.org/10.3390/ijms22147248

Goel, A., Ward, D. G., Noyvert, B., Yu, M., Gordon, N. S., Abbotts, B., Colbourne, J. K., Kissane, S., James, N. D., Zeegers, M. P., Cheng, K. K., Cazier, J. B., Whalley, C. M., Beggs, A. D., Palles, C., Arnold, R., & Bryan, R. T. (2022). Combined exome and transcriptome sequencing of non-muscle-invasive bladder cancer: associations between genomic changes, expression subtypes, and clinical outcomes. *Genome Medicine*, *14*(1), 1–16. https://doi.org/10.1186/s13073-022-01056-4

Gong, B., Kusko, R., Jones, W., Tong, W., & Xu, J. (2022). Ultra-deep multi-oncopanel sequencing of benchmarking samples with a wide range of variant allele frequencies. *Scientific Data*, *9*(1), 3–11. https://doi.org/10.1038/s41597-022-01359-6

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

Gremer, L., Merbitz-Zahradnik, T., Dvorsky, R., Cirstea, I. C., Kratz, C. P., Zenker, M., Wittinghofer, A., & Ahmadian, M. R. (2011). Germline KRAS mutations cause aberrant biochemical and physical properties leading to developmental disorders. *Human Mutation*, *32*(1), 33–43. https://doi.org/10.1002/humu.21377

Guo, Y., Ding, X., Shen, Y., Lyon, G. J., & Wang, K. (2015). SeqMule: Automated pipeline for analysis of human exome/genome sequencing data. *Scientific Reports*, *5*, 1–10. https://doi.org/10.1038/srep14283

Harbers, L., Agostini, F., Nicos, M., Poddighe, D., Bienko, M., & Crosetto, N. (2021). Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. *Frontiers in Oncology*, *11*(July), 1–11. https://doi.org/10.3389/fonc.2021.700568

Harris, R. S. (2013). Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications. *Genome Medicine*, *5*(9), 1. https://doi.org/10.1186/gm490

Hsu, C. W., Sowers, M. L., Baljinnyam, T., Herring, J. L., Hackfeld, L. C., Tang, H., Zhang, K., & Sowers, L. C. (2022). Measurement of deaminated cytosine adducts in DNA using a novel hybrid thymine DNA glycosylase. *Journal of Biological Chemistry*, *298*(3), 101638. https://doi.org/10.1016/j.jbc.2022.101638

Hurst, C. D., Alder, O., Platt, F. M., Droop, A., Stead, L. F., Burns, J. E., Burghel, G. J., Jain, S., Klimczak, L. J., Lindsay, H., Roulson, J. A., Taylor, C. F., Thygesen, H., Cameron, A. J., Ridley, A. J., Mott, H. R., Gordenin, D. A., & Knowles, M. A. (2017). Genomic Subtypes of Non-invasive Bladder Cancer with Distinct Metabolic Profile and Female Gender Bias in KDM6A Mutation Frequency. *Cancer Cell*, *32*(5), 701-715.e7. https://doi.org/10.1016/j.ccell.2017.08.005

Kacew, A., & Sweis, R. F. (2020). FGFR3 Alterations in the Era of Immunotherapy for Urothelial Bladder Cancer. *Frontiers in Immunology*, *11*(November), 1–7. https://doi.org/10.3389/fimmu.2020.575258

Kandori, S., Kojima, T., & Nishiyama, H. (2019). The updated points of TNM classification of urological cancers in the 8th edition of AJCC and UICC. *Japanese Journal of Clinical Oncology*, *49*(5), 421–425. https://doi.org/10.1093/jjco/hyz017

Kato, K., Miya, F., Hamada, N., Negishi, Y., Narumi-Kishimoto, Y., Ozawa, H., Ito, H., Hori, I., Hattori, A., Okamoto, N., Kato, M., Tsunoda, T., Kanemura, Y., Kosaki, K., Takahashi, Y., Nagata, K. I., & Saitoh, S. (2019). MYCN de novo gain-of-function mutation in a patient with a novel megalencephaly syndrome. *Journal of Medical Genetics*, *56*(6), 388–395. https://doi.org/10.1136/jmedgenet-2018-105487

Khong, H. T., & Restifo, N. P. (2002). Natural selection of tumor variants in the generation of "tumor escape" phenotypes. *Nature Immunology*, *3*(11), 999–1005. https://doi.org/10.1038/ni1102-999

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., & Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594. https://doi.org/10.1038/s41592-018-0051-x

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*(1), 1–13. https://doi.org/10.1186/s13073-020-00791-w

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

Kodack, D. P., Farago, A. F., Dastur, A., Held, M. A., Dardaei, L., Friboulet, L., von Flotow, F., Damon, L. J., Lee, D., Parks, M., Dicecca, R., Greenberg, M., Kattermann, K. E., Riley, A. K., Fintelmann, F. J., Rizzo, C., Piotrowska, Z., Shaw, A. T., Gainor, J. F., … Benes, C. H. (2017). Primary Patient-Derived Cancer Cells and Their Potential for Personalized Cancer Patient Care. *Cell Reports*, *21*(11), 3298–3309. https://doi.org/10.1016/j.celrep.2017.11.051

Krueger Felix. (n.d.). *Trim Galore!* Https://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Trim_galore/.

Lewis, T. E., & Moffett, C. (2021). Notes on Mutect2. *Educational Philosophy and Theory*, *53*(13), 1359–1387.

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *00*(00), 1–3. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., &

Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y., Sun, L., Guo, X., Mo, N., Zhang, J., & Li, C. (2021). Frontiers in Bladder Cancer Genomic Research. *Frontiers in Oncology*, *11*(May), 1–10. https://doi.org/10.3389/fonc.2021.670729

Martin-Doyle, W., & Kwiatkowski, D. J. (2015). Molecular Biology of Bladder Cancer. *Hematology/Oncology Clinics of North America*, *29*(2), 191–203. https://doi.org/10.1016/j.hoc.2014.10.002

Maxson & Mitchell, Daniel Harris, BA, Lynn McNicoll, MD, Gary Epstein-Lubow, MD, and Kali S. Thomas, P., Ann L Coker, Chirag M Lakhani1 2, Arjun K Manrai1, 3, Jian Yang4, 5, Peter M Visscher#4, 5,*, and Chirag J Patel#1, 1Department, B. T. T., Das C, Lucia MS, H. K. and T. J., Levine, R. L., Wadleigh, M., Cools, J., Ebert, B. L., Wernig, G., Huntly, B. J. P., Boggon, T. J., Wlodarska, I., Clark, J. J., Moore, S., Adelsperger, J., Koo, S., Lee, J. C., Gabriel, S., … Heidel, F. H. (2014). 乳鼠心肌提取 HHS Public Access. *Leukemia*, *12*(1), 387–397. https://doi.org/10.1038/nature12213.Mutational

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

McGuire Sams, C., Shepp, K., Pugh, J., Bishop, M. R., & Merner, N. D. (2021). Rare and potentially pathogenic variants in hydroxycarboxylic acid receptor genes identified in breast cancer cases. *BMC Medical Genomics*, *14*(1), 1–13. https://doi.org/10.1186/s12920-021-01126-3

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mendiratta, G., Ke, E., Aziz, M., Liarakos, D., Tong, M., & Stites, E. C. (2021). Cancer gene mutation frequencies for the U.S. population. *Nature Communications*, *12*(1), 1–11. https://doi.org/10.1038/s41467-021-26213-y

Miserocchi, G., Mercatali, L., Liverani, C., De Vita, A., Spadazzi, C., Pieri, F., Bongiovanni, A., Recine, F., Amadori, D., & Ibrahim, T. (2017). Management and potentialities of primary cancer cultures in preclinical and translational studies. *Journal of Translational Medicine*, *15*(1), 1–16. https://doi.org/10.1186/s12967-017-1328-z

Nesta, A. V., Tafur, D., & Beck, C. R. (2021). Hotspots of Human Mutation. *Trends in Genetics*, *37*(8), 717–729. https://doi.org/10.1016/j.tig.2020.10.003

Pommerenke, C., Geffers, R., Bunk, B., Bhuju, S., Eberth, S., Drexler, H. G., & Quentmeier, H. (2016). Enhanced whole exome sequencing by higher DNA insert lengths. *BMC Genomics*, *17*(1), 1–8. https://doi.org/10.1186/s12864-016-2698-y

Poplin, R., Ruano-Rubio, Valentin DePristo, M. A., Fennell, T. J., O Carneiro, M., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Joel, T., Sheila, C., Chris, W., Monkol, L., Stacey, G., Mark, J. D., Ben, N., Daniel, G. M., & Eric, B. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxIV*, 1–22.

R Core Team. (2022). *R Core Team (2022). R: A language and environment for statistical computing*.

Richter, M., Piwocka, O., Musielak, M., Piotrowski, I., Suchorska, W. M., & Trzeciak, T. (2021). From Donor to the Lab: A Fascinating Journey of Primary Cell Lines. *Frontiers in Cell and*

Developmental Biology, *9*(July), 1–11. https://doi.org/10.3389/fcell.2021.711381
Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., Hinoue, T., Laird, P. W., Hoadley, K. A., Akbani, R., Castro, M. A. A., Gibb, E. A., Kanchi, R. S., Gordenin, D. A., Shukla, S. A., Sanchez-Vega, F., Hansel, D. E., Czerniak, B. A., Reuter, V. E., … Zwarthoff, E. C. (2017). Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell*, *171*(3), 540-556.e25. https://doi.org/10.1016/j.cell.2017.09.007
Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T. I., Dahary, D., Fishilevich, S., & Lancet, D. (2021). The GeneCards Suite. *Practical Guide to Life Science Databases*, 27–56. https://doi.org/10.1007/978-981-16-5812-9_2
Seaby, E. G., Pengelly, R. J., & Ennis, S. (2016). Exome sequencing explained: A practical guide to its clinical application. *Briefings in Functional Genomics*, *15*(5), 374–384. https://doi.org/10.1093/bfgp/elv054
Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, *54*(1), 1.30.1-1.30.33. https://doi.org/https://doi.org/10.1002/cpbi.5
Talevich, E. (2016). *CNVkit Documentation*.
Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Computational Biology*, *12*(4), 1–18. https://doi.org/10.1371/journal.pcbi.1004873
TCGA Research Network. (n.d.).
Teer, J. K., Zhang, Y., Chen, L., Welsh, E. A., Cress, W. D., Eschrich, S. A., & Berglund, A. E. (2017). Evaluating somatic tumor mutation detection without matched normal samples. *Human Genomics*, *11*(1), 1–13. https://doi.org/10.1186/s40246-017-0118-2
Tome, J. M., Tippens, N. D., & Lis, J. T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nature Genetics*, *50*(11), 1533–1541. https://doi.org/10.1038/s41588-018-0234-5
Tran, L., Xiao, J. F., Agarwal, N., Duex, J. E., & Theodorescu, D. (2021). Advances in bladder cancer biology and therapy. *Nature Reviews Cancer*, *21*(2), 104–121. https://doi.org/10.1038/s41568-020-00313-1
Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V, Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1110), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43
Wang, F., Dong, X., Yang, F., & Xing, N. (2022). Comparative Analysis of Differentially Mutated Genes in Non-Muscle and Muscle-Invasive Bladder Cancer in the Chinese Population by Whole Exome Sequencing. *Frontiers in Genetics*, *13*(March), 1–14. https://doi.org/10.3389/fgene.2022.831146
Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S., & Girirajan, S. (2017). Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-01005-x
Weinstein, J. N., Akbani, R., Broom, B. M., Wang, W., Verhaak, R. G. W., McConkey, D., Lerner, S., Morgan, M., Creighton, C. J., Smith, C., Cherniack, A. D., Kim, J., Pedamallu, C. S., Noble, M. S., Al-Ahmadie, H. A., Reuter, V. E., Rosenberg, J. E., F.Bajorin, D., Bochner, B. H., … Eley,

G. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, *507*(7492), 315–322. https://doi.org/10.1038/nature12965

Wilkinson, S., Ye, H., Karzai, F., Harmon, S. A., Terrigino, N. T., VanderWeele, D. J., Bright, J. R., Atway, R., Trostel, S. Y., Carrabba, N. V., Whitlock, N. C., Walker, S. M., Lis, R. T., Abdul Sater, H., Capaldo, B. J., Madan, R. A., Gulley, J. L., Chun, G., Merino, M. J., … Sowalsky, A. G. (2021). Nascent Prostate Cancer Heterogeneity Drives Evolution and Resistance to Intense Hormonal Therapy. *European Urology*, *80*(6), 746–757. https://doi.org/10.1016/j.eururo.2021.03.009

Xiao, W., Ren, L., Chen, Z., Fang, L. T., Zhao, Y., Lack, J., Guan, M., Zhu, B., Jaeger, E., Kerrigan, L., Blomquist, T. M., Hung, T., Sultan, M., Idler, K., Lu, C., Scherer, A., Kusko, R., Moos, M., Xiao, C., … Shi, L. (2021). Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nature Biotechnology*, *39*(9), 1141–1150. https://doi.org/10.1038/s41587-021-00994-5

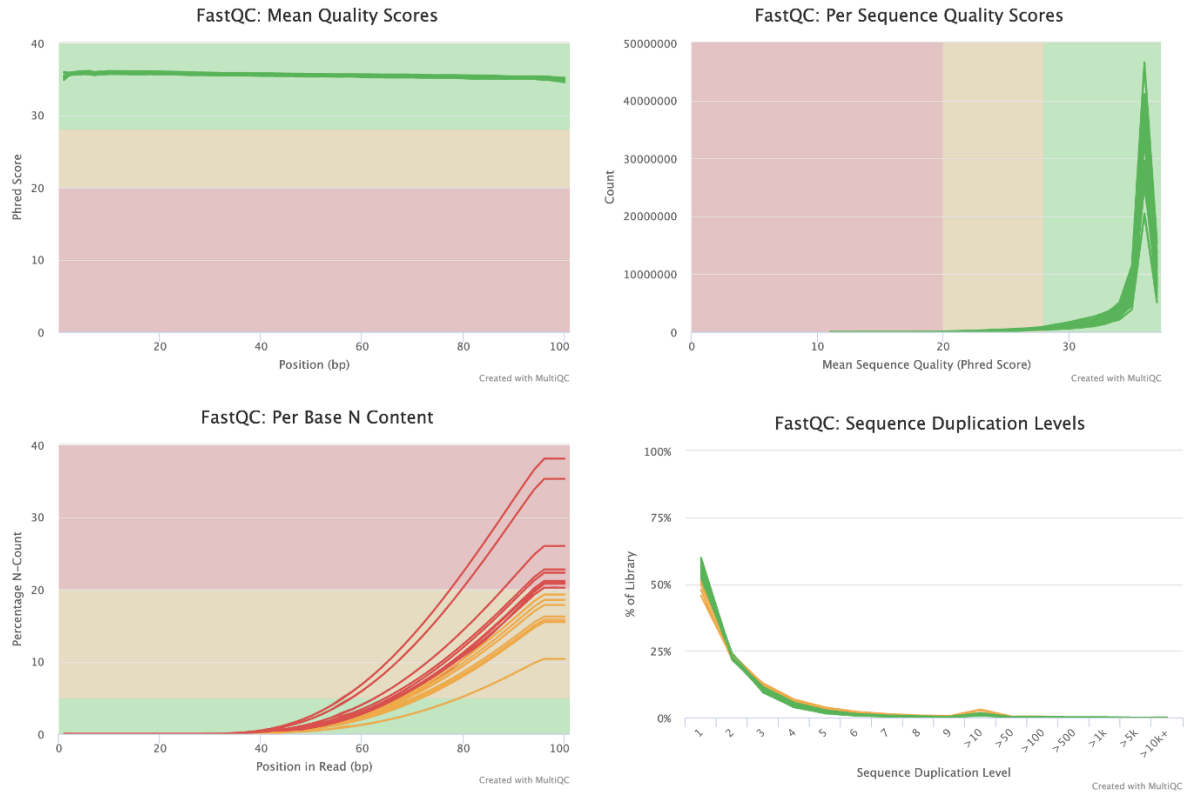
Yicheng, F., Xin, L., Tian, Y., & Huilin, L. (2022). Association of FLG mutation with tumor mutation load and clinical outcomes in patients with gastric cancer. *Frontiers in Genetics*, *13*(August), 1–18. https://doi.org/10.3389/fgene.2022.808542

Zanti, M., Michailidou, K., Loizidou, M. A., Machattou, C., Pirpa, P., Christodoulou, K., Spyrou, G. M., Kyriacou, K., & Hadjisavvas, A. (2021). Performance evaluation of pipelines for mapping, variant calling and interval padding, for the analysis of NGS germline panels. *BMC Bioinformatics*, *22*(1), 1–21. https://doi.org/10.1186/s12859-021-04144-1

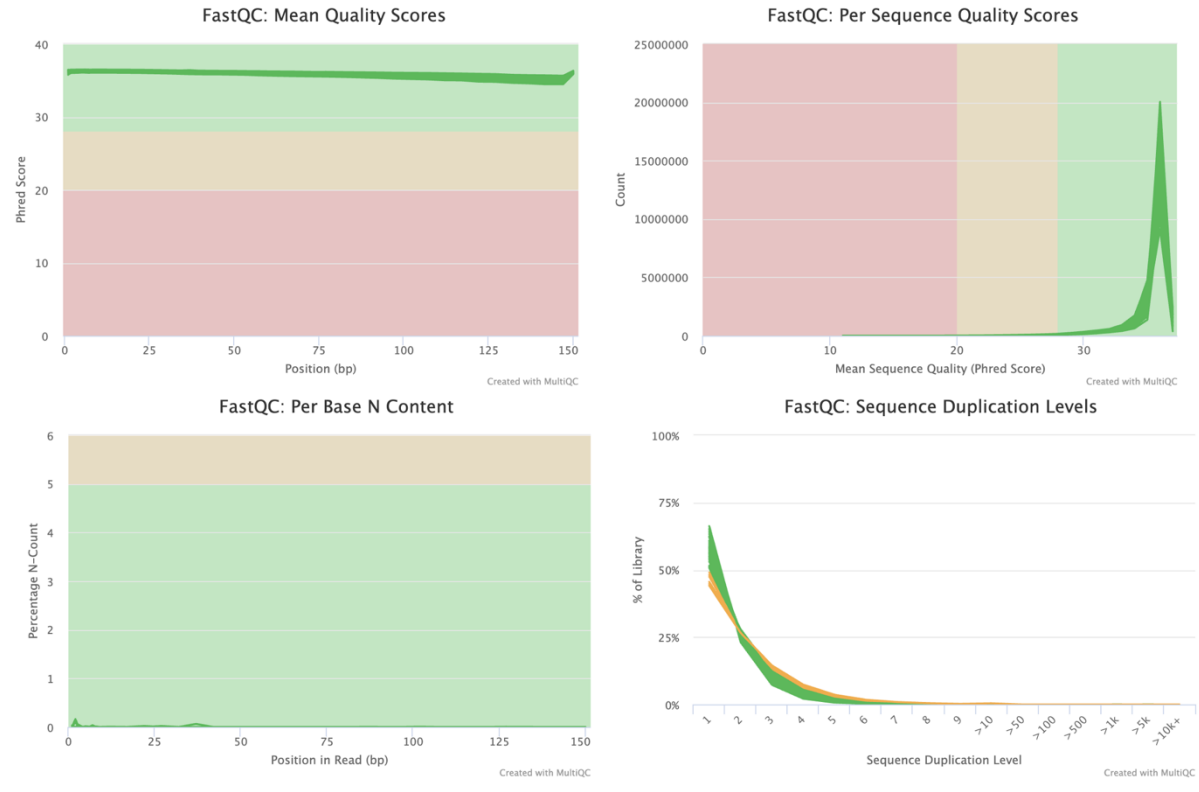
Zhou, J., Zhang, M., Li, X., Wang, Z., Pan, D., & Shi, Y. (2021). Performance comparison of four types of target enrichment baits for exome DNA sequencing. *Hereditas*, *158*(1), 1–12. https://doi.org/10.1186/s41065-021-00171-3

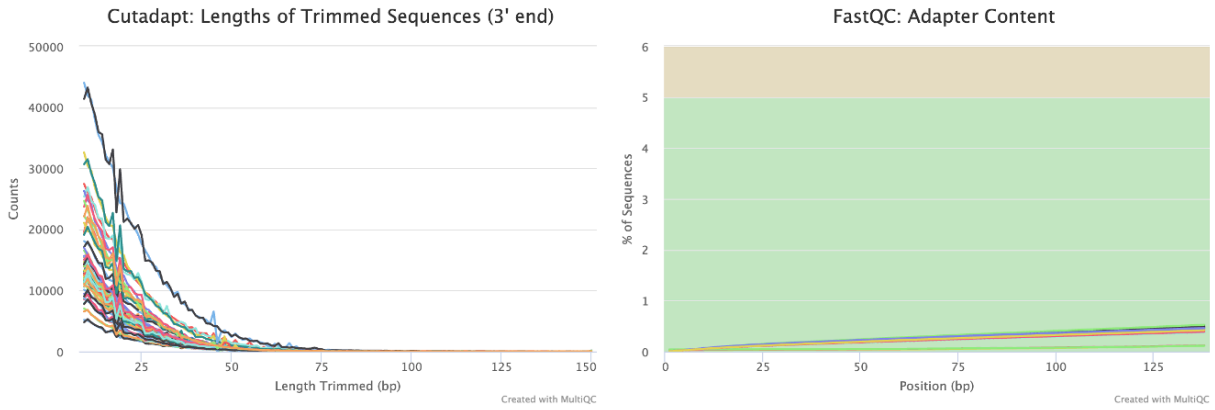
Zverinova, S., & Guryev, V. (2022). Variant calling: Considerations, practices, and developments. *Human Mutation*, *43*(8), 976–985. https://doi.org/10.1002/humu.24311

Supplementary

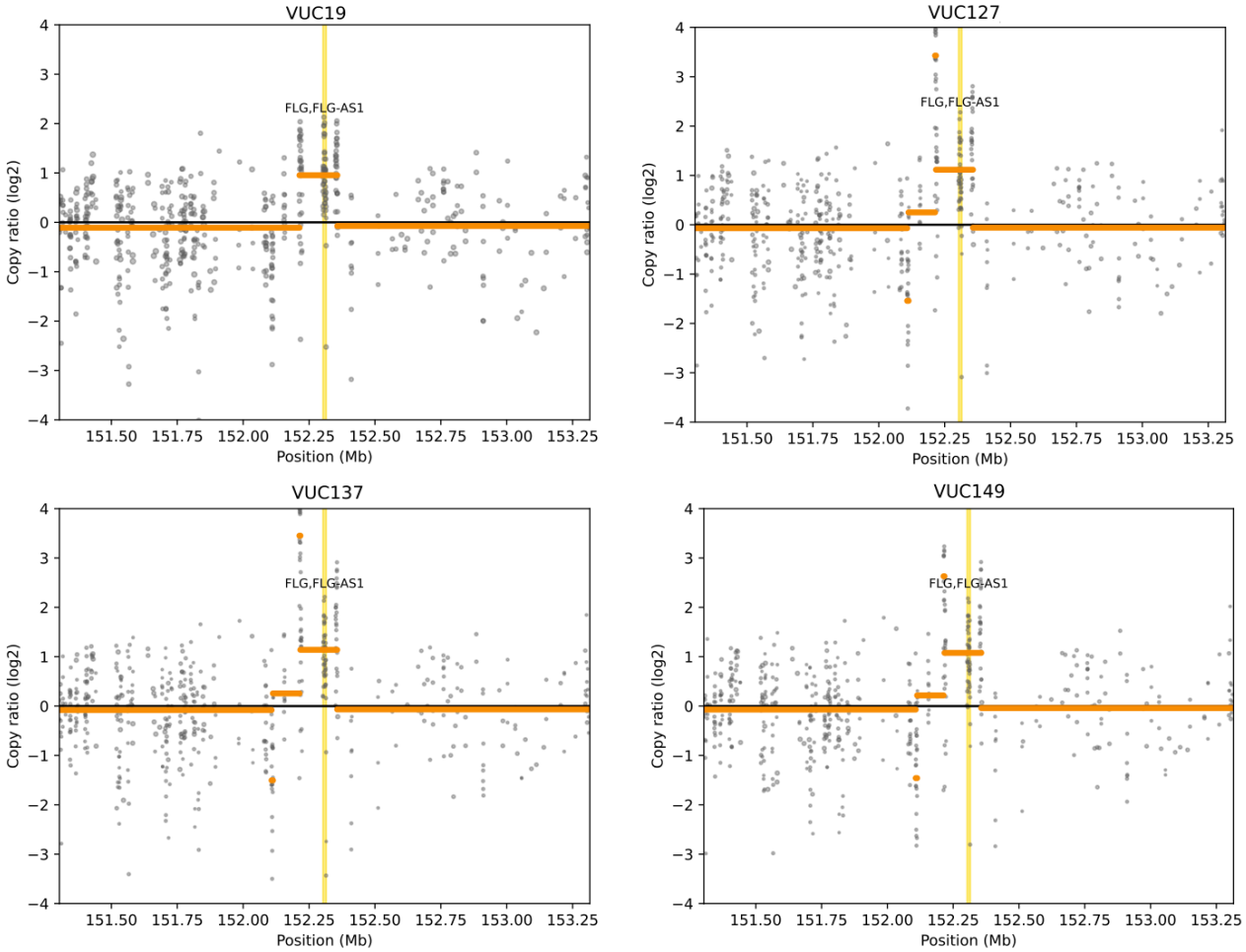


Supplementary Figure 1. Initial quality control outputs on CeMM fastq files.



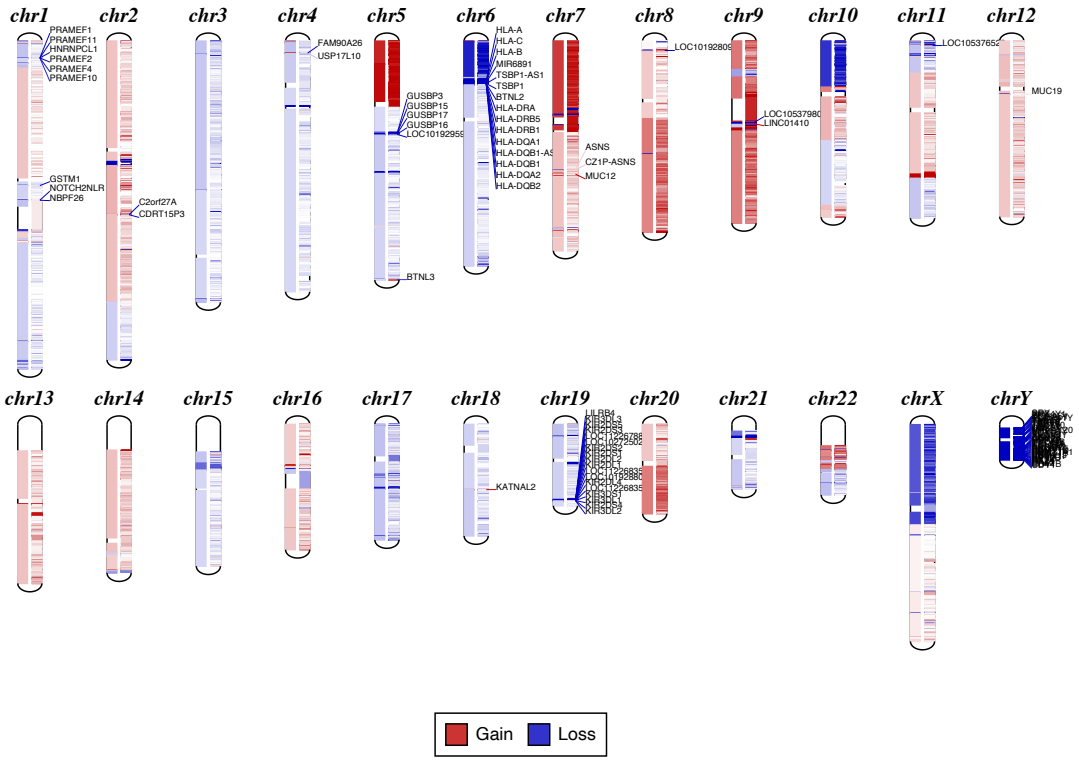


Supplementary Figure 2. Initial quality control outputs on Novogene fastq files (after trimming).

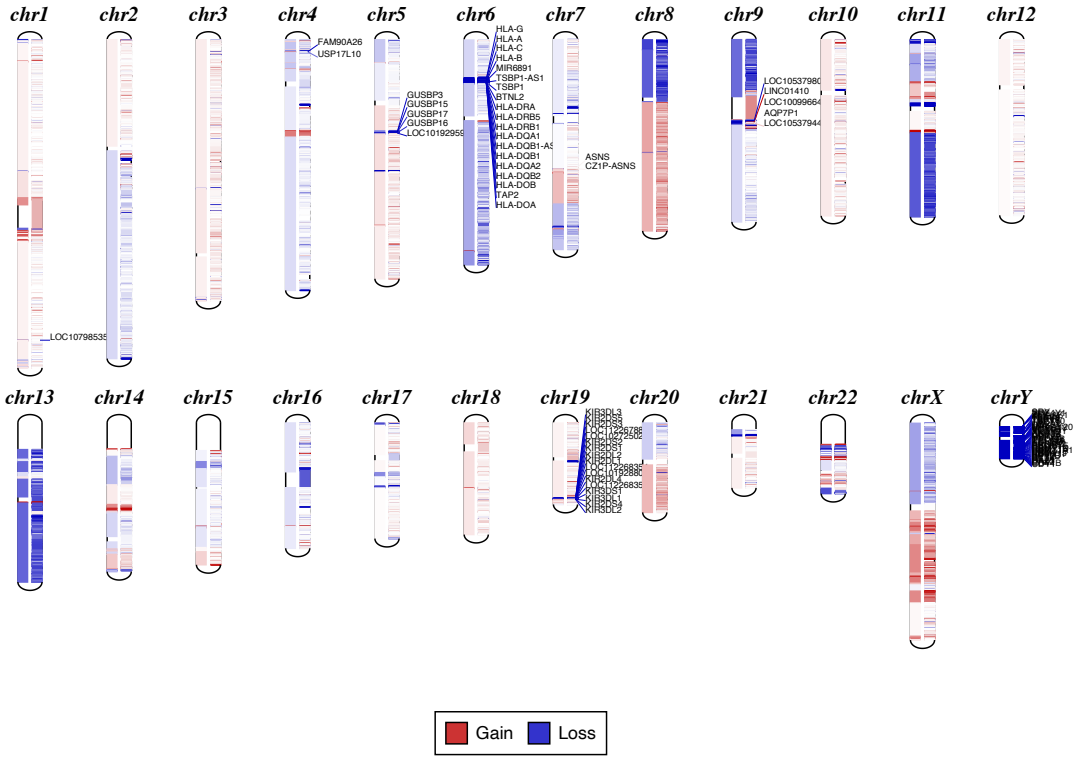


Supplementary Figure 3. Showing amplification/gain of FLG, FLG-AS1 genes observed in 4 VUC models.

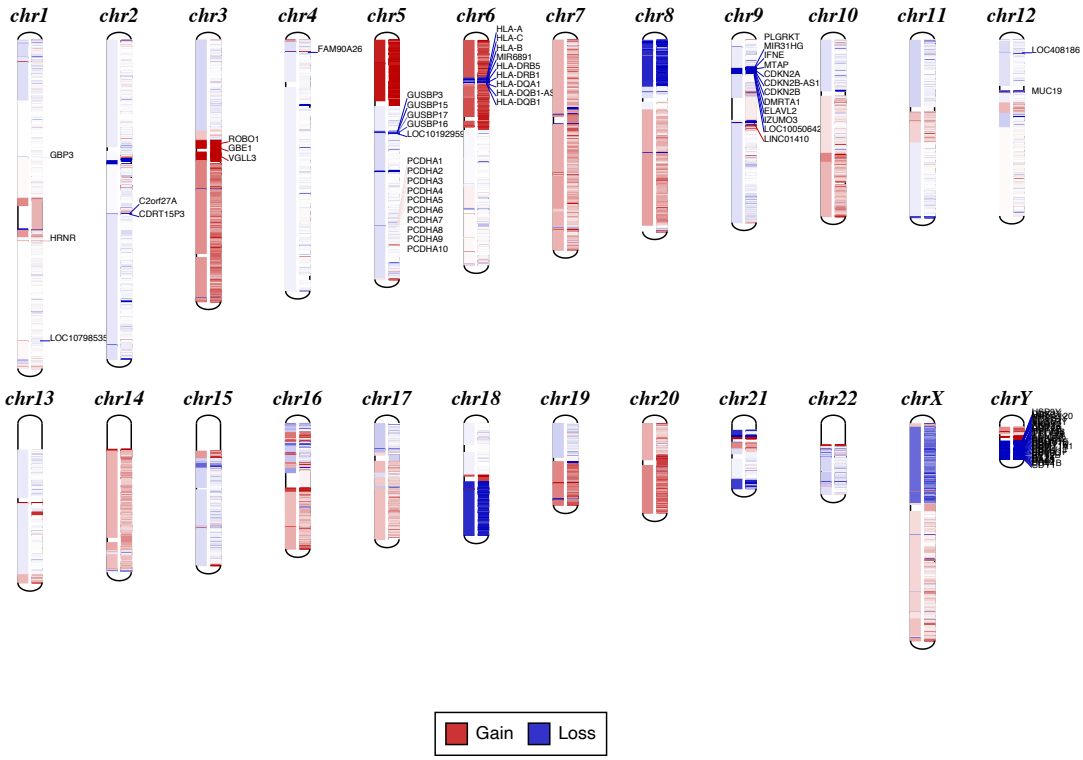
Sample VUC133



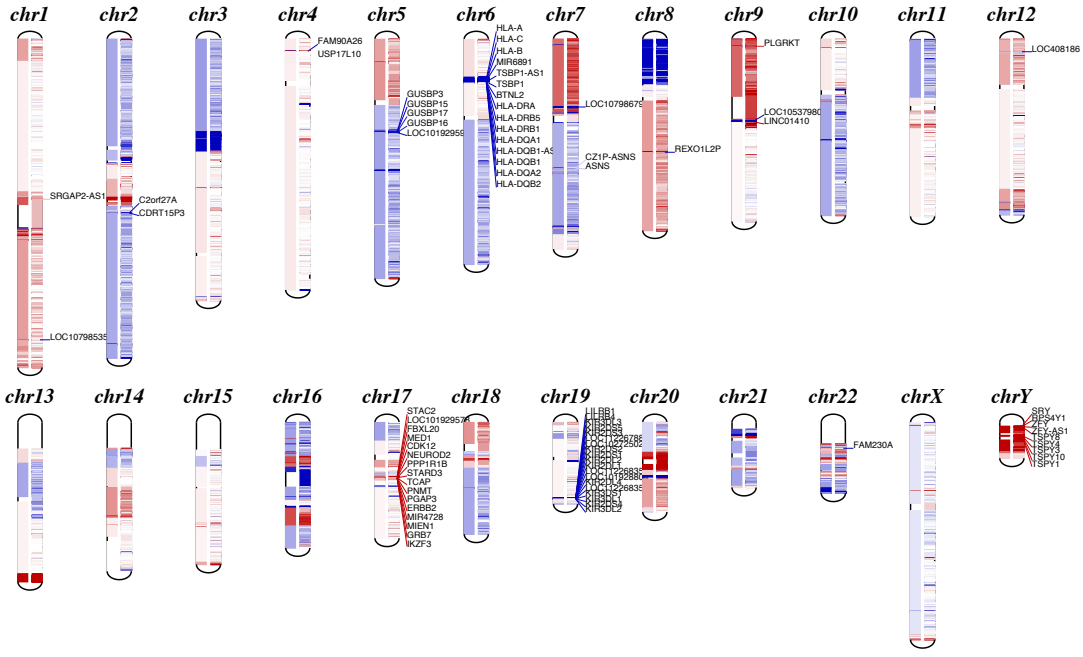
Sample VUC38

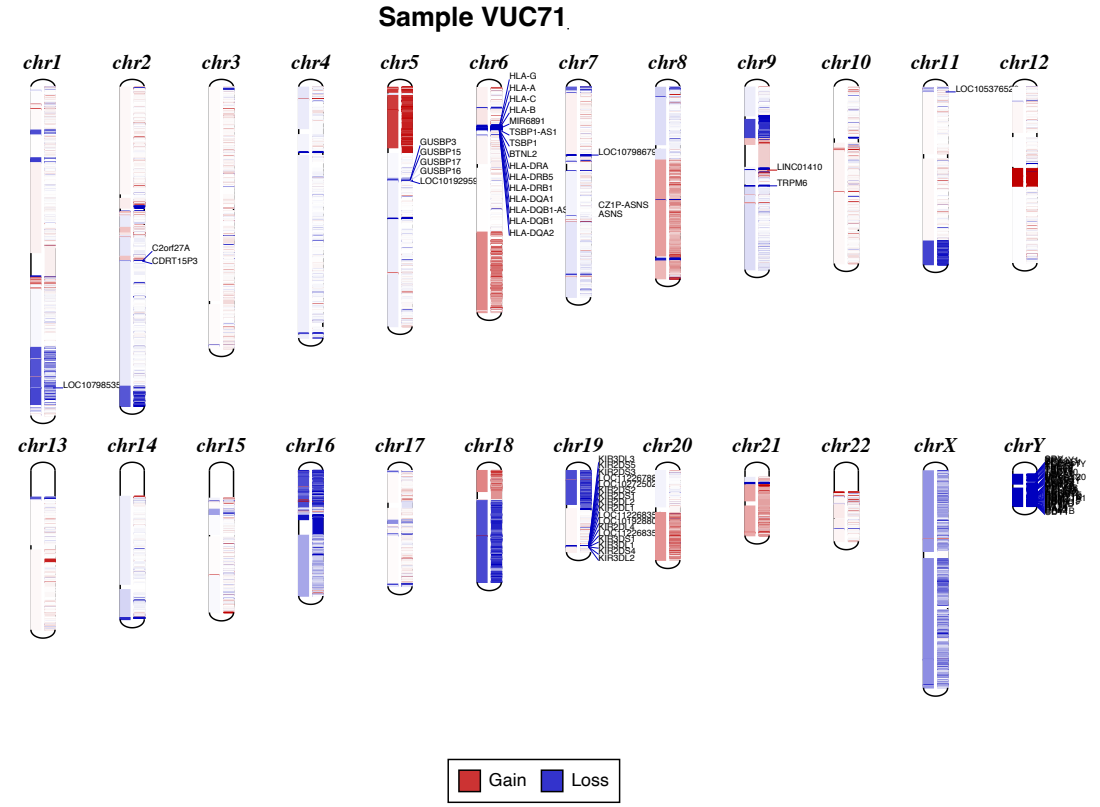


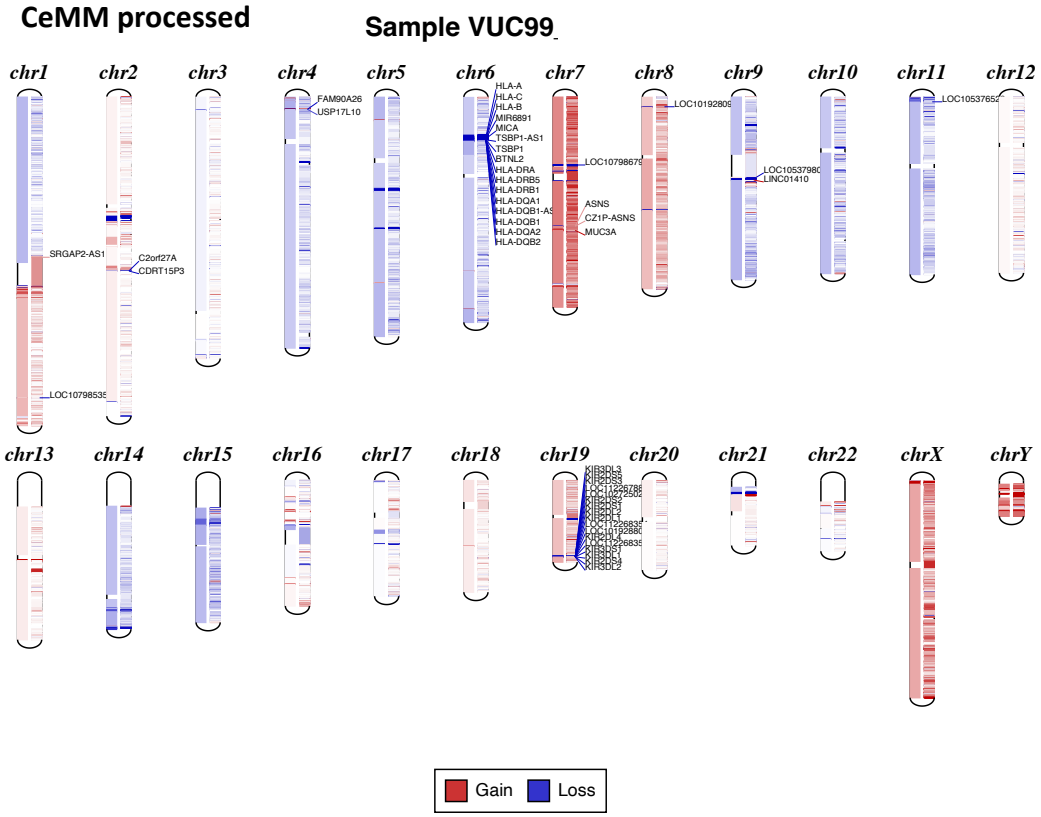
Sample VUC40



Sample VUC48

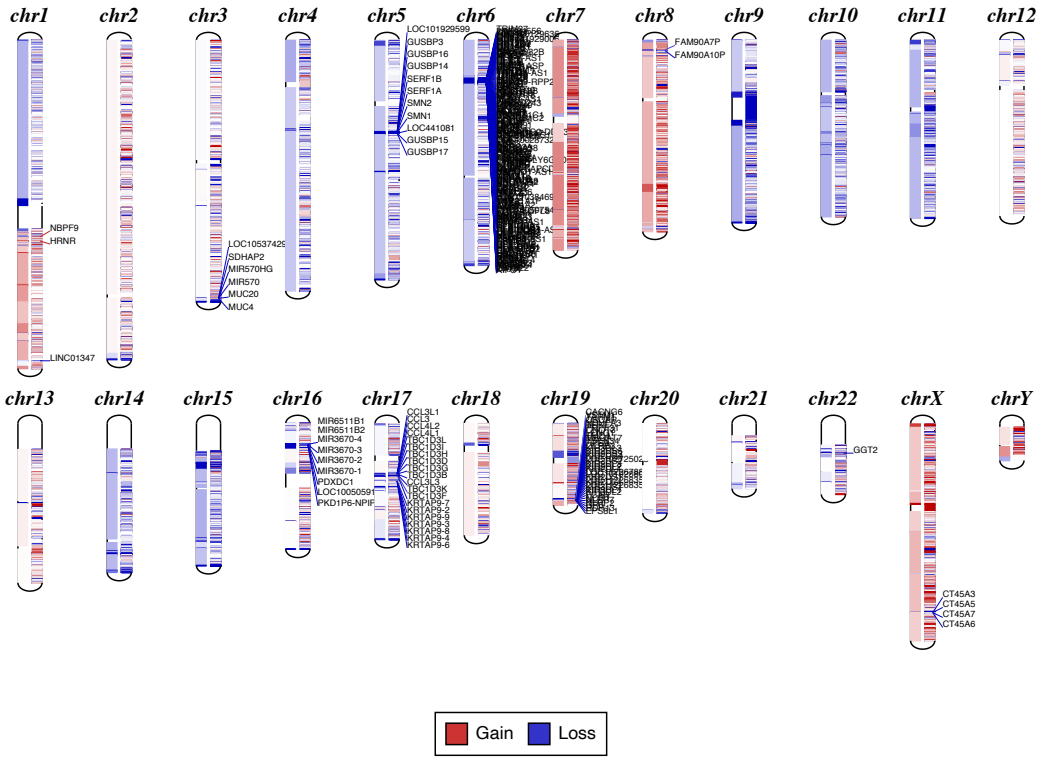



Sample VUC71
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
C2orf27A
CDRT15P3
GUSBP3
GUSBP15
GUSBP17
GUSBP16
HLA-G
HLA-A
HLA-C
HLA-B
MIR6891
TSBP1-AS1
TSBP1
BTNL2
HLA-DRA
HLA-DRB5
HLA-DRB1
HLA-DQA1
HLA-DQB1-AS
HLA-DQB1
HLA-DQA2
CZ1P-ASNS
ASNS
LINC01410
TRPM6
LOC10537652
Gain
Loss

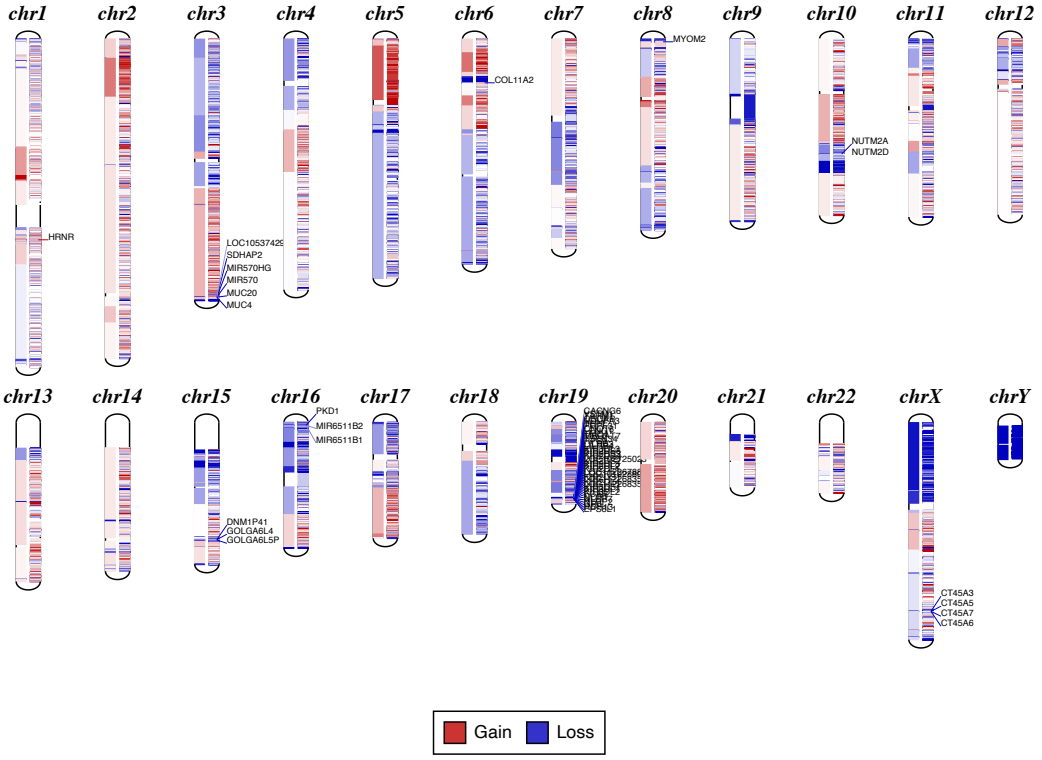

CeMM processed
Sample VUC99
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
SRGAP2-AS1
C2orf27A
CDRT15P3
FAM90A26
USP17L10
HLA-A
HLA-C
HLA-B
MIR6891
MICA
TSBP1-AS1
TSBP1
BTNL2
HLA-DRA
HLA-DRB5
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DQA2
HLA-DQB2
ASNS
CZ1P-ASNS
MUC3A
LINC01410
LOC10537652
Gain
Loss

Novogene processed

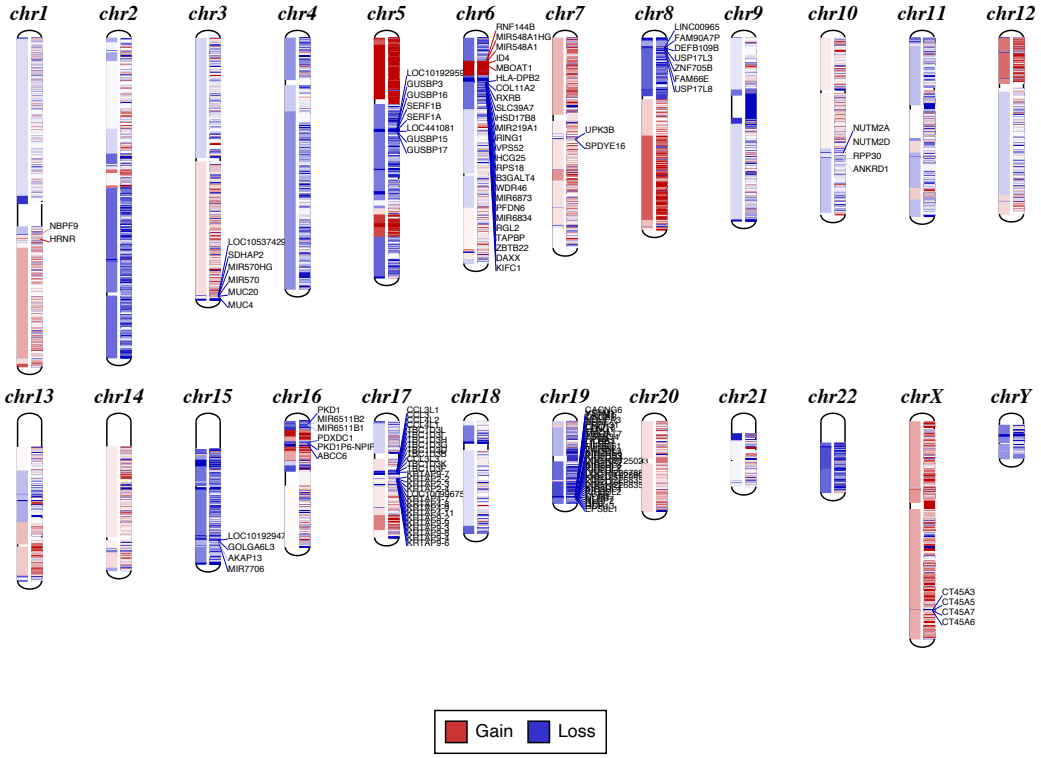
Sample VUC99



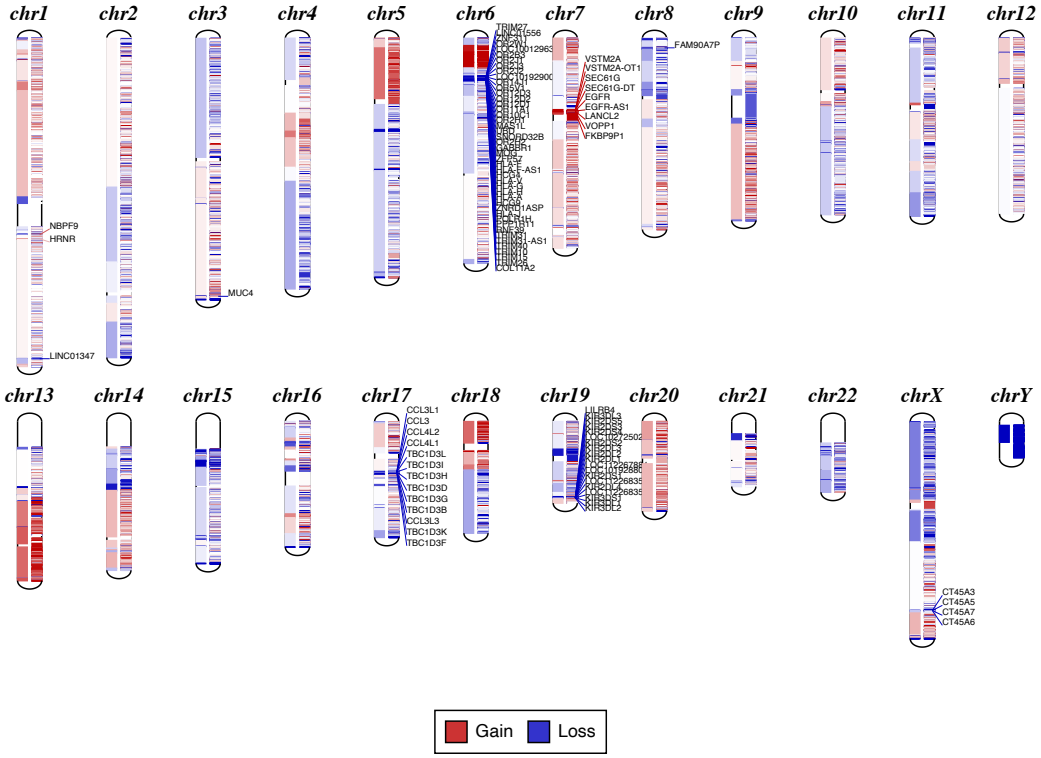
Sample VUC107

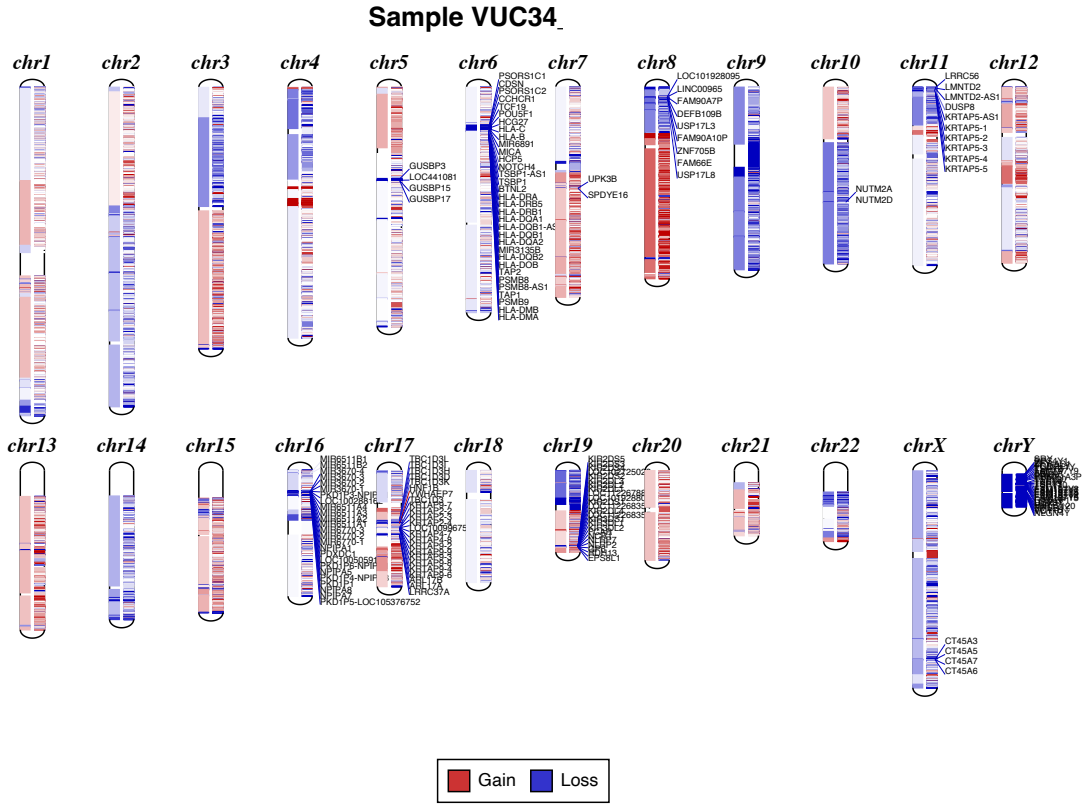


Sample VUC138_

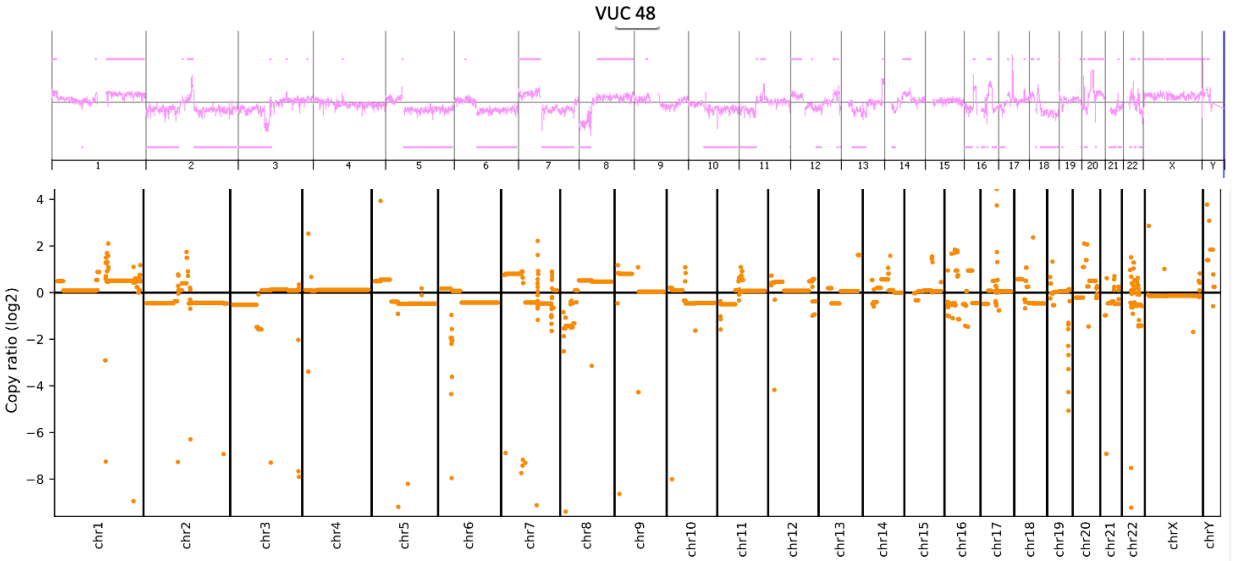


Sample VUC26_

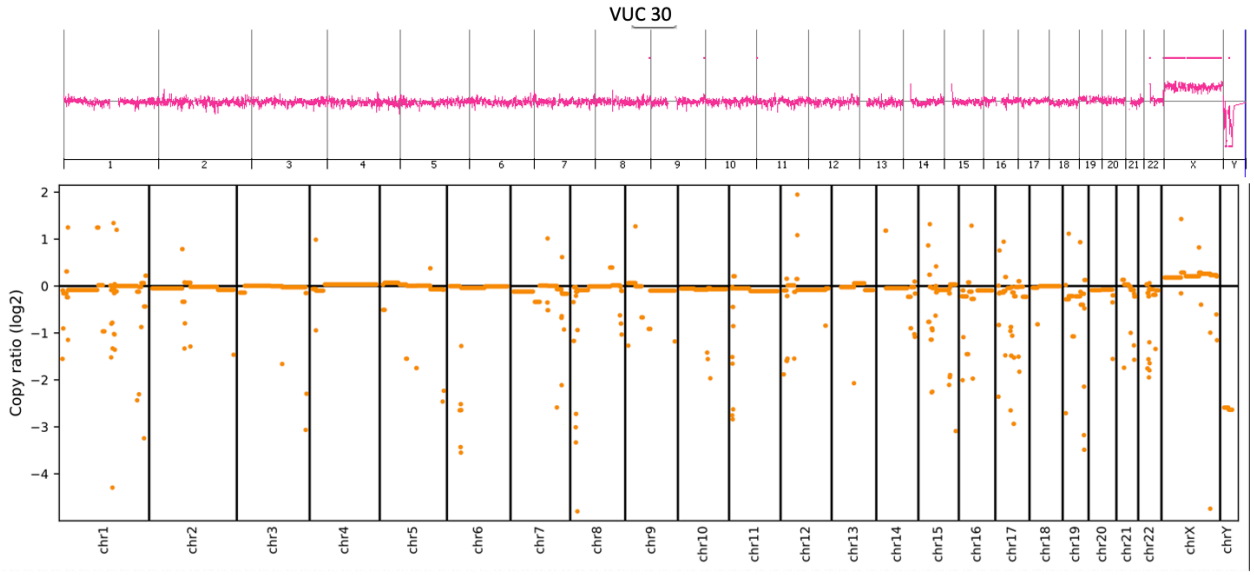




Supplementary Figure 4. Ideogram of VUC models that show the most extreme amplification/gain and deletion/loss of copy number alterations. *Threshold log2 ratio >3 has been specified so only genes that are highly amplified or deleted are labelled.*



Supplementary Figure 5. aCGH and copy number alteration profiles for VUC48.



Supplementary Figure 6. aCGH and copy number alteration profiles for VUC30. *Example of a flatten CNA profile.*