D2.1

Targeted ultra-deep sequencing of cancer-gene loci

Project number:	668858
Project acronym:	PrECISE
Project title:	PrECISE: Personalized Engine for Cancer Integrative Study and Evaluation
Start date of the project:	1st January, 2016
Duration:	36 months
Programme:	H2020-PHC-02-2015

Deliverable type:	Report
Deliverable reference number:	PHC-668858 / D2.1 / 3.0
Work package contributing to the deliverable:	WP 2
Due date:	December 2016 – M12
Actual submission date:	1st December, 2017

Responsible organisation:	IBM
Editor:	Marìa Rodrìguez Martìnez
Dissemination level:	PU
Revision:	V3.0

Abstract:	This report provides targeted profiles of selected biopsies and will be used to improve clone inference in WP1, prognostic-biomarker inference in WP3, and tumour classification in WP4.
Keywords:	sequencing, clone inference, mutations, castrate-resistant prostate cancer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858.

This work was supported (in part) by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0324-2. The opinions expressed and arguments employed therein do not necessarily reflect the official views of the Swiss Government.

Editor

Rodrìguez Martìnez, Marìa (IBM)

Contributors (ordered according to beneficiary numbers)

Sumazin, Pavel (BCM)

Wagner, Ulrich (UZH)

Wild, Peter (UZH)

Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The users thereof use the information at their sole risk and liability.

Executive Summary

We set out to test our ability to infer mutations and their clonality using 10 castrate-resistant prostate cancer (CRPC) tumor biopsies. Each of these biopsies was profiled using an Illumina enrichment kit, an Agilent SureSelectXT, and an Ion Xpress Plus Fragment Library Kit that targets coding exons of ERG, TMPRSS2, AR, PTEN, and SPOP. Conclusions from studying these biopsies informed methodology for selecting additional tumors for profiling by exome sequencing, for predicting mutations from Ion Xpress captures and exome sequencing, and allowed us to devise criteria for estimating mutation cellularity (a necessary step for inferring clonality in WP1). Following these efforts we profiled two areas in each of 39 proCOC patients, and 5 areas in each of 10 additional CRPC patients. These, together with profiles of over 40 additional prostate cancer patients will be used to infer phylogenies, test inferences, and produce prognostically predictive biomarkers for prostate cancer. Clonality inference is a building block for prognostic-biomarker inference in WP3, and tumor classification in WP4.

Revision V2.0: Our original submission of D2.1 focused on the profiling and analysis of 10 areas by whole-exome sequencing using two kits, and on an initial effort to study the capture of 5 known driver genes. This was our first effort to (1) test the two kits, (2) reconstruct phylogeny for CRPCs based on whole-exome sequencing, and (3) profile these tumors at ultra-deep coverage. This update includes the description of a new capture, which we decided to use when searching for additional candidate patients for profiling and analysis, and a clinical description about the profiling of 39 proCOC patients, and 5 areas in each of 10 additional CRPC patients, which occurred following our initial efforts. Clinical data is given in Supplementary table 1.

Revision V3.0: In this revision the supplementary tables for the current deliverable are made available at a public Box link. Details are reported in Chapter 6.

List of Figures

List of Tables

Chapter 1 CRPC DNA profiling

Molecularly profiled tumor sections are depicted for Patient 3 in Figure 1. Each section was profiled using Illumina enrichment kit, Agilent SureSelectXT—both exome captures—and Ion Xpress Plus Fragment Library Kit that targets coding exons of ERG, TMPRSS2, AR, PTEN, and SPOP. Exome sequencing data was used to detect mutations, estimate their frequency, and predict copy number changes in exome-wide fashion. Analysis protocols are described below.

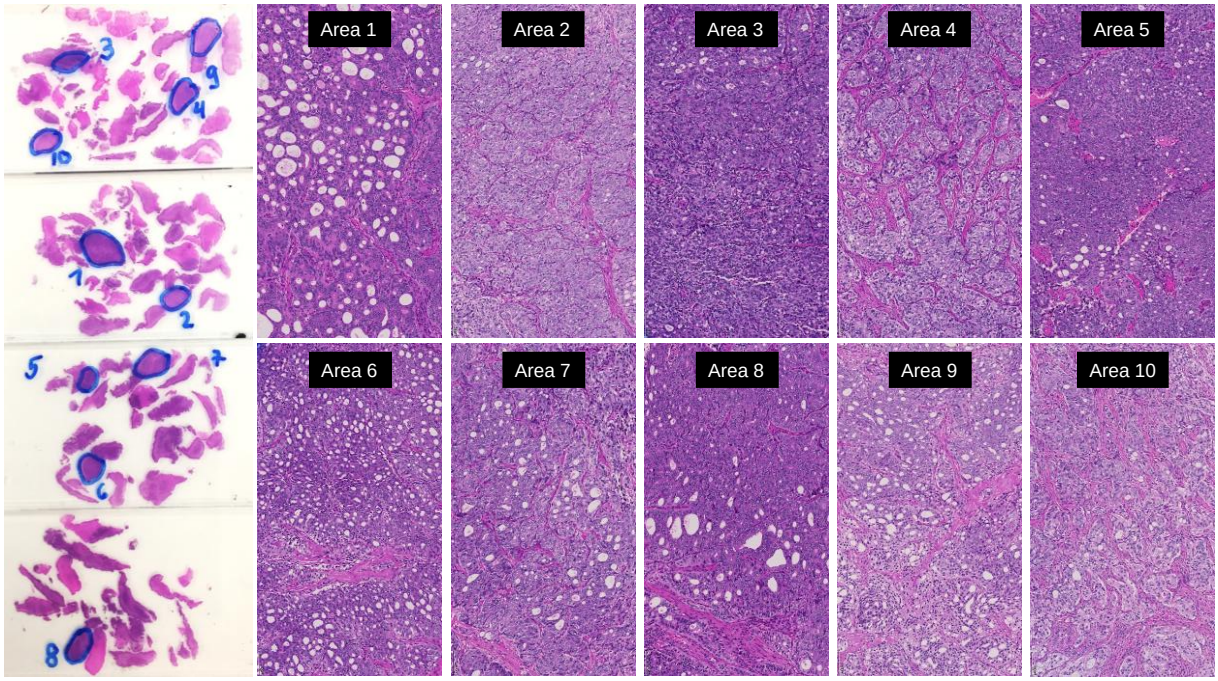

Figure 1: Ten tumor areas were chosen for profiling by exome sequencing and an Ion Xpress Plus Fragment Library Kit. Areas were selected in an effort to maximize diversity, and, subsequently, information content for clonal inference across biopsies.

1.1 Mutation calling

Mutation calling followed protocols established by TCGA and ExAC.1,2 First reads were aligned to hg19 using BWA, then variants were called with GenomeAnalysisTK.

1.1.1 Alignment and quality control

Fastq files from exomes and capture were aligned to the human genome reference (hg19) using bwa on a per lane basis using the following command lines. Picard MarkDuplicates was used to mark likely PCR duplicate reads and indels were accounted for using GATK's RealignerTargetCreator and a list of known indel sites. Using this interval list, local realignment was then performed by GATK IndelRealigner. The base quality scores were then recalibrated using GATK BaseRecalibrator and a list of known variant sites and recalculated using GATK PrintReads.

```
bwa aln hg19.fasta -q 5 -l 32 -k 2 -t $NSLOTS -o 1 -f $output.1.sai $input.1.fastq.gz

bwa aln hg19.fasta -q 5 -l 32 -k 2 -t $NSLOTS -o 1 -f $output.2.sai $input.2.fastq.gz

bwa aln hg19.fasta -q 5 -l 32 -k 2 -t $NSLOTS -o 1 -f $output.unpaired.sai $output.unpaired.fastq.gz

bwa sampe -t $NSLOTS -T -P -f $output.aligned_bwa.sam hg19.fasta $output.1.sai $output.2.sai $input.1.fastq.gz $input.2.fastq.gz
```

1.1.2 Variant calling

GATK HaplotypeCaller algorithm was used to generate gVCFs and known sites were annotated with dbSNP135. Command lines were as follows.

```
java –jar GenomeAnalysisTK.jar -T HaplotypeCaller --
disable_auto_index_creation_and_locking_when_reading_rods -R hg19.fasta -o $output.vcf.gz -
I $input.bam -L $input.intervals --minPruning 3 --maxNumHaplotypesInPopulation 200 -ERC
GVCF   --max_alternate_alleles 3 -variant_index_parameter 128000 -variant_index_type LINEAR
-contamination 0.0

java –jar GenomeAnalysisTK.jar -T CombineGVCFs
-disable_auto_index_creation_and_locking_when_reading_rods -R hg19.fasta -o $output.vcf.gz
-V gvcf.list --sample_rename_mapping_file rename_alias_file.txt


java –jar GenomeAnalysisTK.jar -T GenotypeGVCFs --
disable_auto_index_creation_and_locking_when_reading_rods -R hg19.fasta -o
$output.unfiltered.vcf.gz -D hg19.dbsnp.vcf -L Input.intervals  -V all_combined_gvcfs.list

java –jar GenomeAnalysisTK.jar -T VariantRecalibrator --
disable_auto_index_creation_and_locking_when_reading_rods -R hg19.fasta -input
$input.sites_only.unfiltered.vcf.gz --num_threads 2 -recalFile $output.snps.recal
-tranchesFile $output.snps.tranches -allPoly -tranche 100.0 -tranche 99.95 -tranche 99.9 -
tranche 99.8 -tranche 99.6 -tranche 99.5 -tranche 99.4 -tranche 99.3 -tranche 99.0 -tranche
98.0 -tranche 97.0 -tranche 90.0 -an QD -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an
InbreedingCoeff -resource:hapmap,known=false,training=true,truth=true,prior=15
hapmap_3.3.b37.vcf.gz -resource:omni,known=false,training=true,truth=true,prior=12
1000G_omni2.5.b37.vcf.gz -resource:1000G,known=false,training=true,truth=false,prior=10
1000G_phase1.snps.high_confidence.b37.vcf.gz
-resource:dbsnp137,known=false,training=false,truth=false,prior=7 dbsnp_138.b37.vcf.gz
-resource:dbsnp129,known=true,training=false,truth=false,prior=3
dbsnp_138.b37.excluding_sites_after_129.vcf.gz --maxGaussians 6 -mode SNP -rscriptFile
$output.snps.recalibration_plots.rscript
```

Predicted mutations—after excluding poor-quality mutations and those that appeared in fewer than three biopsies—were crossed with previously identified mutations in the COSMIC database. 119 COSMIC mutations were identified in 118 genes; see Table 1 below.

For every mutation in Table 1, we estimated the frequency—proportion of reads covering the mutated positions—and the copy number using VarScan using default parameters and setting the maximum amplification to 8x.3 The result produced, for each mutation in each biopsy, (1) the number of reads supporting the reference sequence, (2) the number of reads supporting the alternate (somatic) sequence, and (3) the average copy number for the given position across all profiled cells. Comparing our copy number estimates suggested that mutations identified in our biopsies had greater genomic instability than what was observed for genes in TCGA prostate2 and breast4 carcinomas (PRAD and BRCA); see Figure 2.

Symbol	Chromosome	StartPosition	Symbol	Chromosome	StartPosition	Symbol	Chromosome	StartPosition
ABCA13	chr7	48269459	SOGA1	chr20	35441382	NDRG2	chr14	21487315
ABCC5	chr3	183679331	SPCS1	chr3	52741601	NEO1	chr15	73536675
ABLIM2	chr4	8034546	SPEG	chr2	2.2E+08	NKAIN3	chr8	63768953
ADCK4	chr19	41198272	SPO11	chr20	55907112	PILRB	chr7	99957210
ANGPTL6	chr19	10206984	SRP14	chr15	40328471	POTEC	chr18	14534886
ANHX	chr12	133796008	SRRM1	chr1	24998286	POTEM	chr14	19990801
AP1S2	chrX	15870503	SYDE2	chr1	85644102	POTEM	chr14	20019383
AQP6	chr12	50369285	SYNE1	chr6	1.53E+08	PRELP	chr1	203453296
ASH1L	chr1	155447740	TAS2R30	chr12	11285909	PREPL	chr2	44548425
ATG2B	chr14	96769617	TCEA3	chr1	23724019	PRMT8	chr12	3649768
BANF1	chr11	65771358	TCHH	chr1	1.52E+08	PRODH2	chr19	36293637
CACNA1B	chr9	140851169	TGM7	chr15	43585279	PROM1	chr4	16077540
CASKIN2	chr17	73497272	TMBIM6	chr12	50151844	PTGDR	chr14	52735190
CD1E	chr1	158325745	TMC3	chr15	81641692	PXDNL	chr8	52361805
CERK	chr22	47087599	TMEFF2	chr2	1.93E+08	RANBP17	chr5	170626462
COG3	chr13	46057444	TMEM161B	chr5	87517667	RFX7	chr15	56394536
CUBN	chr10	16989293	TP73	chr1	3638549	RPL10L	chr14	47120123
DGKB	chr7	14880879	TRAM1L1	chr4	1.18E+08	RPS6KA1	chr1	26870962
DNMT3A	chr2	25469200	TRAM2	chr6	52441810	RSPH10B2	chr7	6826763
DPY19L4	chr8	95738579	TRAPPC9	chr8	1.41E+08	RYR1	chr19	39075660
DYM	chr18	46904950	TRMT2B	chrX	1E+08	RYR2	chr1	237777756
EDN3	chr20	57897621	TRUB2	chr9	1.31E+08	SAP130	chr2	128735644
ENTHD1	chr22	40271433	TSPAN6	chrX	99887416	SEC22A	chr3	122944283
F5	chr1	169492462	TTBK1	chr6	43222945	SERHL2	chr22	42970288
FBN1	chr15	48722747	TUBGCP3	chr13	1.13E+08	SERPINF2	chr17	1648406
FGFR3	chr4	1806540	UBAC2	chr13	99970267	SH2B3	chr12	111856399
FOXO3	chr6	108984667	UBE4B	chr1	10132505	SH3TC1	chr4	8216010
GOLGB1	chr3	121435808	UBR1	chr15	43282241	SIGLEC10	chr19	51919085
GPT	chr8	145732105	UNC5C	chr4	96143426	SIGLEC5	chr19	52129411
IDH1	chr2	209103715	USP49	chr6	41773507	SIPA1	chr11	65413699
JAKMIP1	chr4	6050742	UTP18	chr17	49362592	SLC16A4	chr1	110932130
KIRREL	chr1	157963400	VIPR2	chr7	1.59E+08	SLC1A2	chr11	35287017
LCMT2	chr15	43622103	WHAMM	chr15	83499273	SLC26A4	chr7	107312534
LOXHD1	chr18	44113308	ZBTB46	chr20	62421700	SLC6A9	chr1	44459219
LRP2	chr2	169997124	ZHX1	chr8	1.24E+08	SLC9A9	chr3	142984890
LRRC6	chr8	133622253	ZNF17	chr19	57929158	SMARCA5	chr4	144451500
MACF1	chr1	39888527	ZNF24	chr18	32917874	SMPD4	chr2	130915668
MON2	chr12	62887931	ZNF341	chr20	32358141	SNRNP200	chr2	96970517
MS4A12	chr11	60269505	ZNF778	chr16	89294574	SNX18	chr5	53814224
MYLK2	chr20	30414340	ZNF790	chr19	37310662			

Table 1: 119 COSMIC mutations were in 118 genes. All mutations were identified in at least 3 of the ten biopsies from Patient 3 using both Agilent and Illumina exome sequencing kits. Coordinates refer to the hg19 assembly.

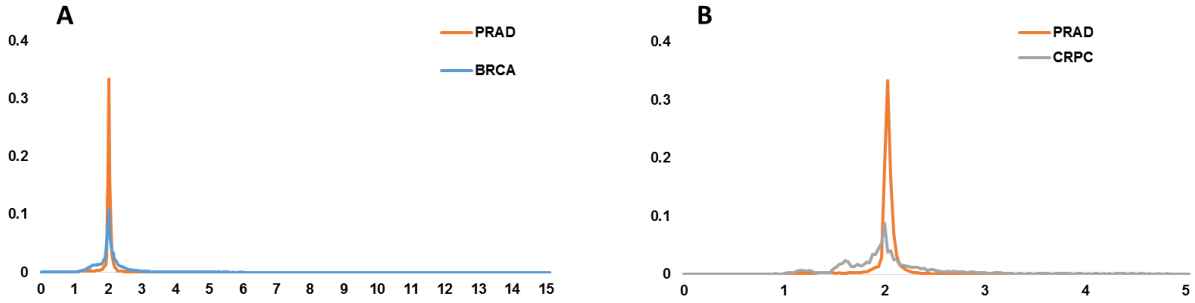

Figure 2: Distribution density plots of copy numbers of (**A**) all genes in profiled PRADs and BRCAs by TCGA and (**B**) all genes in profiled PRADs compared to all Table 1 mutations (CRPC). The results suggest that mutations in our biopsies take on a greater range of copy numbers:

1.1.3 Capture

Our initial capture, focusing on ERG, TMPRSS2, AR, PTEN, and SPOP failed to identify any mutations targeting the covered loci. Consequently, this assay did not inform about our ability to estimate clonality from capture data. Instead, the results suggested that captures are most useful to segregate patients and identify those that have and those that do not have common mutations. However, we needed a method to select patients for analysis. Consequently, we designed a 2nd capture. This is reported within D2.2.

1.1.4 Conclusions

Information about mutation copy numbers across biopsies suggested that copy number information is essential for accurate estimates of mutation cellularity. Mutation cellularity is required for identifying ancestral relations between clones, as depicted in Figure 3 and Figure 4. We therefore elected to include copy number analysis as a part of any future profiling.

Chapter 2 Estimating clonal composition and ancestral relations between clones

Ancestral relations between clones are estimated based on clonal cellularity (Figure 3), and cellularity, in turn, is estimated from observed mutation frequencies in the sequencing experiment. Genomic instability can dramatically alter cellularity estimates, as shown in Figure 3C. Here, we propose an interpretation of observed frequencies derived from 3A,B when copy numbers are neutral (Figure 3C top) and when copy numbers are taken from a distribution matching that of CRPC Figure 2B.

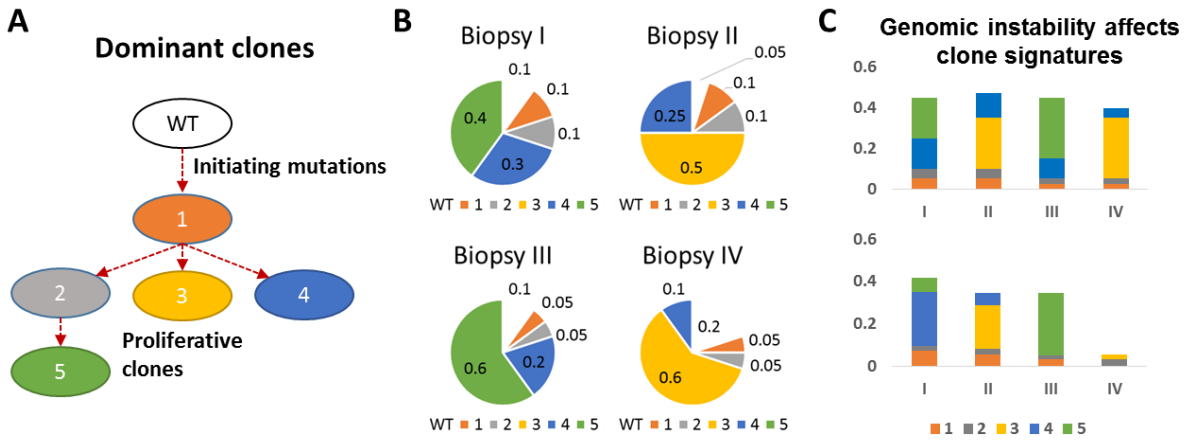

Figure 3: An anecdotal example for clonal phylogeny describing the evolution of a tumor. (**A**) Each node represent a subclone—a dominate cell type within the biopsy—and the edges describe ancestral relations between subclones. Subclones 2, 3 and 4 are decedent from subclone 1 and inherit any genetic alterations present in subclone 1. Mutations in Subclone 1 are likely to be tumor initiating, while mutations in subclones 4 and 5 have proliferative advantage. (**B**) Each biopsy is composed of subclones, and the proportion is termed *cellularity*; cellularity can refer to clones or mutations, where mutation cellularity is the proportion of tumor cells with the mutation. (**C**) Genomic instability, i.e. allele copy number, affects clone signatures and our ability to estimate mutation cellularity and frequency from sequencing data.

The results suggest that failure to account for genomic instability can lead to large deviations in the accuracy of cellularity estimation. Errors in cellularity estimation can have a dramatic impact on our ability to estimate ancestral relations between clones and reverse engineer clonal phylogeny (Figure 4D,E).

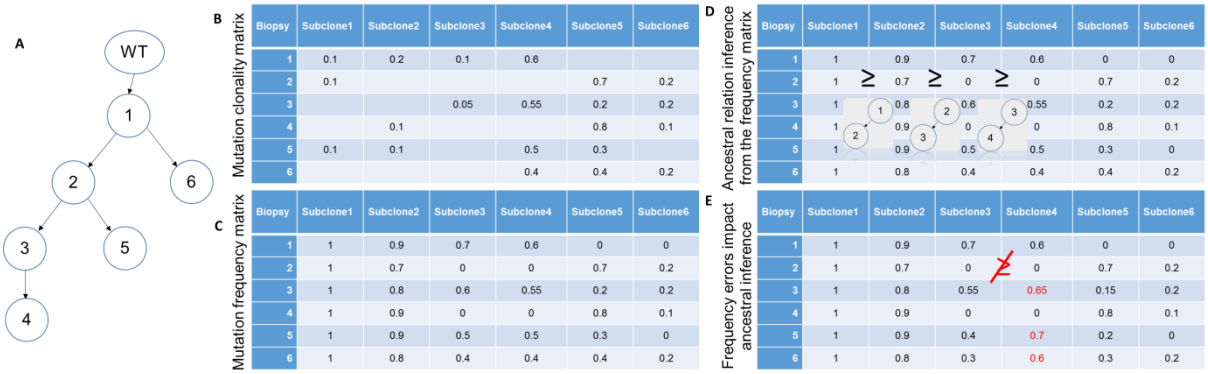

Figure 4: An anecdotal example for clonal phylogeny describing the evolution of a tumor. **(A)** Each node represent a subclone—a dominate cell type within the biopsy—and the edges describe ancestral relations between subclones. Subclones 2 and 6 are decedent from subclone 1 and inherit any genetic alterations present in subclone 1. **(B)** The mutation cellularity matrix informs about the composition of each biopsy, as a sum of its composing subclones. **(C)** The mutation frequency matrix assigns expected observations of mutations associated with each subclone in each biopsy. **(D)** The mutation frequency matrix can be used to reverse engineer phylogeny from the root down, by noticing a conserved relationship between clones across biopsies. **(E)** Frequency estimation errors (in red) can inhibit reverse engineering efforts; in (D) mutation frequencies of subclone 4 were always smaller than those of subclone 3, but here this relation is lost.

Chapter 3 Modeling copy number variations for improved cellularity estimates

Based on observations from our profiled biopsies, we produced the following model (Figure 5). This model accounts for genomic variability at each allele, and makes no assumptions about copy numbers of the allele in mutated or un-mutated cells. Based on this model, we are redesigning methods for cellularity estimation and for reverse engineering both ancestral relations between clones and clonal phylogeny.

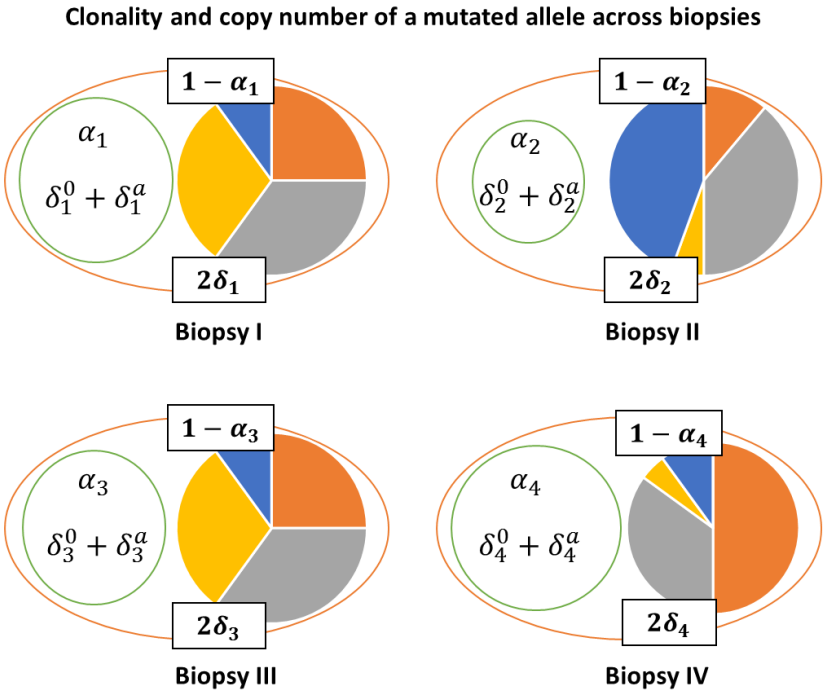

Figure 5: For each mutation, in each biopsy (biopsy I to IV), we model the cellularity of the mutation α, and copy number of the allele in cells (tumor or wildtype) without the mutation ($2\delta_i$), and copy number of the wildtype and mutated allele (δ_i^0 and δ_i^a, respectively) in cells with the mutation.

3.1 Tests using synthetic data

Our observations from TCGA tumor profiles and our CRPC data suggested that copy numbers of mutated alleles can vary widely, and that copy numbers of these alleles vary also in biopsies and tumors where the allele appears to be exclusively wild type. These results suggest that δ_i, δ_i^0 and δ_i^a can vary from one biopsy to another, and copy numbers in profiled tumors can range from 0 to over 200 per gene. To simulate our observations we generated simulated phylogenies, cellularity values, and copy number data. Copy number was taken from truncated normal distributions, with mean ranging from 1 to 4 and standard deviations ranging from ½ to 2. Some simulations produce narrower and some produce wider ranges of copy numbers. All were used to evaluate methods for estimating cellularity. Results from this evaluation will be reported as a part of Work package 1.

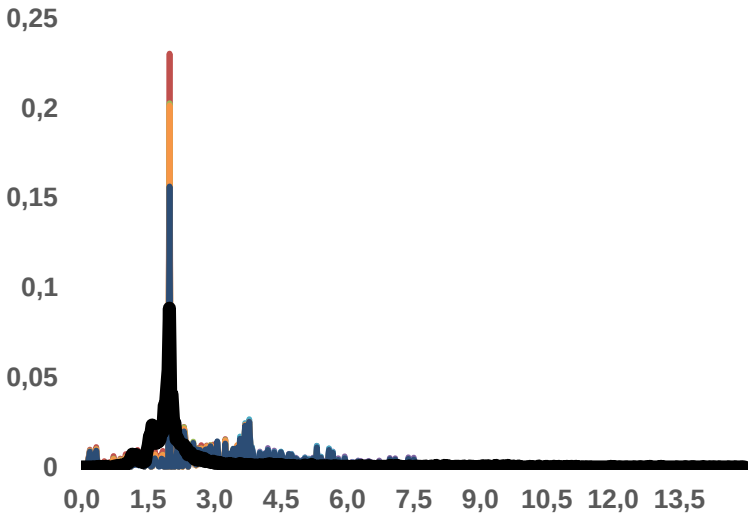

Figure 6: Copy number distributions in our CRPC data (black) and 8 additional synthetically generated that were used for simulations.

Chapter 4 Informing D2.2 and D2.3

Based on our observations from the analysis of the 10 CRPC biopsies from Patient 3 we have devised methodology to identify patients using targeted capture. These patients, over 40 in total were profiled using a targeted capture approach, and are reported on in D2.2. Here, we include profiles of 5 areas from 10 additional CRPC patients. Their identified mutations, including mutations that were found to be of specific interest, are given in Supplementary table 2.

Chapter 5 Summary and Conclusion

We set out to test our ability to infer mutations and their clonality using 10 castrate-resistant prostate cancer (CRPC) tumor biopsies from one CRPC patient. Each of these biopsies was profiled using an Illumina enrichment kit, an Agilent SureSelectXT, and an Ion Xpress Plus Fragment Library Kit that targets coding exons of ERG, TMPRSS2, AR, PTEN, and SPOP. Conclusions from studying these biopsies informed methodology for selecting additional tumors for profiling by exome sequencing, for predicting mutations from captures and exome sequencing, and allowed us to devise criteria for estimating mutation cellularity (a necessary step for inferring clonality in WP1). Clonality inference is a building block for prognostic-biomarker inference in WP3, and tumor classification in WP4. Following these efforts, we profiled a total of 39 patients with RNA-seq and whole-exome sequencing of 2 areas per patient. This data will be used to test predictive biomarkers and inferred phylogeny from a cohort of 25 patients, including 10 CRPC patients that we report on here, and 15 candidate patients that were selected for additional profiling (D2.2).

Chapter 6 Supplementary tables

Supplementary Table 1. (Tab 1, 39 proCOC_patient_info) Clinical data on 39 proCOC patients that where profiled by RNA-seq and whole-exome sequencing of two areas. (Tab 2, 10 CRPC patients) Clinical data on 10 additional CRPC patients that were profiled at 5 areas per patient. (Tab 3, Selected for more profiling) A sample of patients that were selected for additional ultra-deep profiling.

Supplementary Table 2. Mutations identified in the 10 CRPC patients using targeted sequencing (36 genes). Including selected mutations that are candidates as predictive mutations (2nd tab).

The supplementary tables are available and can be downloaded from the following link:

https://bcm.app.box.com/s/oafhiuejoiodayuv74g9653nlkstrh70

Chapter 7 List of Abbreviations

Abbreviation	*Explanation*
TCGA	*The Cancer Genome Atlas*
CRPC	*Castration-resistant Prostate Cancer*
PC	*Prostate Cancer*
DNA	*Deoxyribonucleic acid*

Chapter 8 Bibliography

[1] Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-291.

[2] Network CGAR. The molecular taxonomy of primary prostate cancer. Cell. 2015;163(4):1011-1025.

[3] Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research. 2012;22(3):568-576.

[4] The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.