Highly accurate variant detection for identification of tumor mutations and mosaic variants

Francesc Muyas Remolar

TESI DOCTORAL UPF / 2019

Thesis supervisor

Dr. Prof. Stephan Ossowski

Dr. Prof. Roderic Guigó

INSTITUTE OF MEDICAL GENETICS AND APPLIED GENOMICS

CENTRE FOR GENOMIC REGULATION

Als que estan i als que han estat al meu costat

"*La perfección es una pulida colección de errores*"

Mario Benedetti

ACKNOWLEDGMENTS

An important part of the success of my PhD thesis depends largely on the guidelines and help of many others. In this first part of the thesis, I would like to express my most sincere gratitude to that people who have given to me support both scientifically and emotionally.

First of all, I would like to thank my supervisor Dr. Stephan Ossowski for his continuous support, patience and guidance. Of course, I would also like to thank him for giving to me the great opportunity to join his lab first in CRG (Center for Genomic Regulation, Barcelona) and afterwards in Tübingen (Germany).

On the other hand, I would like to acknowledge the useful work of my second thesis director Dr. Roderic Guigó from *CRG*, Barcelona. I am grateful to Roderic for the valuable support and suggestions during this period.

Beside my supervisors, I would also like to thank Mattia, Hana and Luis, who also worked in our group in CRG, for the extensive support during the period in Barcelona and for their helpful suggestions. They have been very important to improve my knowledge during my thesis. Moreover, I would like to express my sincere gratitude to other lab mates from Tübingen (Franz, Marc, Axel, Lenni, Jakob, Joo, Sarah…).

I think that my Russian colleague, German, has been much more than a work mate. He has become a very good friend that has supported and partied with me always that I needed. German, thanks a lot for everything and let's continue meeting for a kebab box for lots of years.

I'd also like to remember my friends outside my working place. Facu, Medina, Monfi, Victor, Bertolín, Godman, Edgar, Francis, Nerea, Julia, Diego, Moritz… thanks a lot for being there! Although most probably we will be dispersed around the world in few years, I know we always will feel together.

Of course, I am also grateful to all my family and specially to my parents, Juan Antonio and Paqui. Moltes gràcies pels votres ànims, per preguntar-me com va tot i escoltar-me sempre que ho he necessitat. Tot i que estem lluny, vos sento molt prop. Vos vull molt! Per suposat, també agrair al meu germà Joan, al meus tiets, cosins, "abuelos" i especialment al meus iaios,

que m'haguera agradat que hagueren pogut compartir amb mi aquesta experiència (vos trobo a faltar). Sincerament, moltes gràcies a tots.

Lastly, I would like to thank Carmen. ¿Qué puedo decir de ti? Quizás necesitaría otra tesis para decirlo, pero bueno, lo intento en unas pocas líneas. Supongo que no estaría aquí si no fuera por ti, te debo muchísimo más de lo que nunca podré darte, aunque lo intentaré. Llevamos siendo pareja más de diez años, pero estos dos últimos hemos sido una "pareja a distancia", yo descubriendo el norte y tu Barcelona. Creo que una de las cosas más positivas de esta tesis ha sido pasar esta experiencia contigo, que, aunque a mucha distancia, te he sentido muy pero que muy cerca. No sé a qué país nos llevará el futuro, pero de lo que sí que estoy seguro es que sea donde sea, estaremos juntos. ¡Te quiero mucho!

To sum up, I am very proud to have known all these people during this period. All of them not only have helped me to improve my experience and knowledge in genomics, but also have shown to me to enjoy the science life.

ABSTRACT

The rapid development of high-throughput sequencing technologies pushed forward the fields of medical genomics and precision medicine, creating many new applications for diagnostics and clinical studies that require high quality data and highly accurate analysis methods. Distinguishing errors from real variants in Next Generation Sequencing data is a challenge when systematic errors, random sequencing errors, germline variants or somatic variants at very low allele frequency are present in the same data. In the first part of this thesis, we developed a genotype callability filter (ABB) able to identify systematic variant calling errors that were not found by state-of-the art methods. This tool cleans false positive calls from somatic and germline variant callsets, as well as detects false gene-disease associations in case-control studies. Secondly, we developed a set of novel methods able to distinguish and correct sequencing and PCR errors with the use of molecular barcodes, permitting us to build error rate models for the detection of somatic mutations at extremely low allele frequencies. We demonstrated the applicability of our methods for liquid biopsy and monitoring of cancer treatment response in a longitudinal study of the circulating-tumor DNA (ctDNA) kinetics in 20 head and neck squamous cell carcinoma patients during radiochemotherapy (RCTX). As final part of this thesis, we characterized mosaic mutations in a multi-tissue, multi-individual study using a cohort of thousands of samples from hundreds of healthy individuals. The high number of embryonic mosaic mutations we observed in coding regions implies novel hypotheses and diagnostic procedures for investigating genetic causes of disease and cancer predisposition.

RESUM

El ràpid desenvolupament de les tecnologies de seqüenciació d'alt rendiment ha impulsat els camps de la genòmica mèdica i la medicina d'alta precisió, creant una gran varietat de noves aplicacions, les quals requereixen dades d'una qualitat excel·lent i mètodes d'anàlisi altament precisos. La distinció entre errors i variants reals en dades de seqüenciació de propera generació (NGS) és un repte quan hi ha errors sistemàtics o aleatoris mesclats amb variants germinals o somàtiques a freqüències al·lèliques molt baixes. En la primera part d'aquesta tesi, hem desenvolupat un filtre per al genotipatge de variants (ABB) capaç d'identificar errors sistemàtics durant el procés de detecció de variants que altres mètodes convencionals no poden trobar. Aquesta eina filtra falsos positius del conjunt de variants finals en estudis de variacions somàtiques i germinals, així com també detecta falses associacions de malalties gèniques en estudis de casos-controls. En segon lloc, hem desenvolupat un conjunt de nous mètodes capaços de distingir i corregir els errors de seqüenciació i PCR amb l'ús d'identificadors moleculars. Aquests ens permeten modelar les taxes d'error i conseqüentment detectar mutacions somàtiques a freqüències al·lèliques extremadament baixes. A més, hem demostrat l'aplicabilitat d'aquests mètodes per l'anàlisi de biòpsies líquides i el seguiment de la resposta al tractament contra el càncer en un estudi longitudinal de la cinètica de l'ADN de tumor circulant (ctDNA) en 20 pacients amb carcinoma de cèl·lules escamoses de cap i coll durant radioquimioteràpia (RCTX). Per finalitzar aquesta tesi, hem caracteritzat les mutacions mosaiques en un estudi multi-teixit multi-individu utilitzant una cohort de centenars d'individus sans amb milers de mostres. L'elevat nombre de mutacions mosaiques codificants que ocorren durant el desenvolupament embrionari humà implica noves hipòtesis i procediments diagnòstics per investigar les causes genètiques d'una gran diversitat de malalties i la predisposició al càncer.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTARY FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY TABLES

ABBREVIATIONS

NGS	*Next-Generation Sequencing*
ABB	*Allele Balance Biasds*
AB	*Allele Balance*
AFB1	*Aflatoxin B_1*
AUC	*Area Under the Curve*
BAM	*Sequence Alignment Map (binary)*
BQ	*Base Quality*
BWA	*Burrows-Wheeler Aligner*
cfDNA	*Cell-Free DNA*
CI	*Confidence Interval*
CLL	*Chronic Lymphocytic Leukemia*
CNV	*Copy Number Variant*
COV	*Coverage*
CRG	*Centre for Genomic Regulation*
CT	*Computed Tomography*
ctDNA	*Circulating-Tumor DNA*
cvDNA	*Circulating Viral DNA*
DFS	*Disease-Free Survival*
DNA	*Deoxyribonucleic Acid*
DP	*Depth of Coverage*
EEMM	*Early-Embryonic Mosaic Mutation*
EMM	*Embryonic Mosaic Mutation*
FFPE	*Formalin-Fixed Paraffin-Embedded*
FP	*False Positive*
FPR	*False Positive Rate*
FS	*Fisher Strand bias*
GATK	*Genome Analysis Toolkit*
GIAB	*Genome In A Bottle*
GQ	*Genotype Quality*
GTEx	*Genotype-Tissue Expression*
GTV	*Gross Tumor Volumes*
HC	*High Confident*
HCC	*Hepatocellular Carcinoma*
hGE	*Haploid Genome Equivalent*
HGP	*Human Genome Project*

HNSCC	*Head and Neck Squamous Cell Carcinoma*
HPV	*Human Papiloma Virus*
ICGC	*International Cancer Genome Consortium*
IMRT	*intensity-modulated radiotherapy*
Indel	*Insertion / Deletion*
LCR	*Low-Complexity Region*
LEMM	*Late-Embryonic Mosaic Mutation*
LN	*Lymph Nodes*
LR	*Logistic Regression*
MEMM	*Mid-Embryonic Mosaic Mutation*
MMC	*mitomycin C*
MRD	*Minimal Residual Disease*
OS	*Overall survival*
PCA	*Principal Component Analysis*
PCR	*Polymerase Chain Reaction*
PT	*Primary Tumor*
QC	*Quality Control*
RCTX	*Radiochemotherapy*
RF	*Random Forest*
RIN	*RNA Integrity Number*
RNA	*Ribonucleic Acid*
RPKM	*Reads Per Kilobase per Million*
RVAS	*Rare Variant Association Study*
SAM	*Sequence Alignment Map*
SBS	*Sequencing By Synthesis*
SNP	*Single Nucleotide Polymorphism*
SNV	*Single Nucleotide Variant*
SV	*Structural Variant*
T1	*Time point 1*
T2	*Time point 2*
T3	*Time point 3*
T4	*Time point 4*
T5	*Time point 5*
TCGA	*The Cancer Genome Atlas*
Ti	*Transition*
TMB	*Tumor Mutation Burden*
TP	*True Positive*

TPM	*Transcripts Per Million*
Tv	*Transversion*
UMI	*Unique Molecular Identifier*
UV	*Ultraviolet light*
VAF	*Variant Allele Frequency*
VLC	*Very Low Confident*
VQSR	*Variant Quality Score Recalibration*
WES	*Whole Exome Sequencing*
WGS	*Whole Genome Sequencing*
5-FU	*5-fluorouracil*

INTRODUCTION

1. FROM MENDEL TO NEXT-GENERATION SEQUENCING

The origin of genetics is to be found in Gregor Mendel's experiment on plant hybridization (1865) (Mendel, 1866). In this book, Mendel described and discovered the fundamental laws of inheritance by studying seven different characters of garden pea plant (*Pisum sativum*). However, it was not until the first years of 20th century when his work was re-discovered and the massive impact on biological sciences was appreciated. Thanks to this study and the discovery of chromosomes during the last decades of 19th century, in 1909, Wilhelm Johannsen proposed the new biological concept of 'gene' to describe the functional unit of the heredity and recombination (Gayon, 2016).

It was three decades later (1941) when Beadle and Tatum showed the first proof that a specific gene could control a biochemical reaction (production of vitamin B6) (Beadle and Tatum, 1941). However, although they proved the role of genes on controlling and regulating specific biochemical reactions in the system, the molecular bases remained unknown. Although deoxyribonucleic acid (DNA) was already isolated in 1869 by Friedrich Miescher in Tübingen, it was only in 1953 when Francis Crick and James Watson (thanks to Rosalind Franklyn work) discovered the structure of DNA and its importance in hereditary processes (Watson and Crick, 1953). From this year on, many new findings, such as the discovery of the genetic code or the first model to describe the regulation of the gene expression (Crick et al., 1961), rapidly changed our view and improved our knowledge of molecular genetics.

One of the main technologies that substantially helped to evolve the fields of molecular biology and genetics was DNA sequencing. Although the invention of DNA sequencing was first reported in 1968 (Wu and Kaiser, 1968), other biomolecules were sequenced several years before. The first molecule sequenced was a protein of insulin, in 1953 by Frederick Sanger (Shendure et al., 2017). Sanger's technology fragmented the protein molecule in two chains, deciphered each fragment and afterwards overlapped them to obtain the complete protein sequence (Shendure et al., 2017). Few years later, in 1965, a first sequence of an RNA molecule (alanine tRNA – 76 nucleotides) was obtained using similar processes (Holley et al., 1965): fragmentation of RNA with RNases, separation of pieces by chromatography and electrophoresis, decipheration by

sequential exonuclease digestion, and sequence deduction by overlapping fragments. As previously mentioned, the first DNA sequence was obtained in 1968. It consisted in the sequencing of the cohesive ends of phage lambda DNA using primer extension (Wu and Kaiser, 1968). Almost 10 years later, in 1977, Sanger and Coulson described the chain terminator procedure (Sanger et al., 1977), also known as Sanger sequencing, which changed the way and the speed to obtain DNA sequences, reaching an output of few hundreds bases per day.

The 90s was an important decade for DNA sequencing. In 1995 and 1996, the whole genomes of *Haemophilus influenza* (12 Mb, 1995) and *Saccharomyces cerevisiae* (~ 12 Mb, 1996) were successfully completed and two years later, the *C. elegans* genome (around 100 Mb, 1998) (Fleischmann et al., 1995; Goffeau et al., 1996; The C. elegans Sequencing Consortium, 1998). In 2000, the whole-genome of *Drosophila melanogaster* (around 175 Mb) was obtained using a whole-genome shotgun strategy (Adams et al., 2000), which would represent the pilot project for the Human Genome Project (HGP). HGP released the first draft of the human genome in 2001 (Lander et al., 2001; Venter et al., 2001) and the finished version in 2004 (International Human Genome Sequencing Consortium, 2004). Nevertheless, although sequencing efficiency improved exponentially since 1968, new findings and technologies were needed in order to achieve the goal of sequencing complete eukaryotic genomes for reasonable cost and efforts. The high number of independent steps that Sanger sequencing required, all of them very crucial, and the huge time and financial investments required to generate the human genome (Lander et al., 2001), inspired the science community to investigate alternatives to electrophoretic sequencing technology. The Human Genome Project represented an inflexion point in the DNA sequencing era, as massive parallel sequencing technologies started to replace Sanger sequencing for large-scale sequencing projects.

2. NEXT-GENERATION SEQUENCING ERA

The way and how fast genomes are analyzed has dramatically changed since mid 2000s. The electrophoretic sequencing strategy (Sanger sequencing), based on bacteria cloning and fragment length quantification, was then replaced by the massive parallel sequencing. The strategy of densely multiplexing short DNA fragments on a plate, which are subsequently sequenced in cycles using imaging technologies to

determine the nucleotide sequence (e.g. the two methods pyrosequencing and 'sequencing-by-synthesis' (SBS)) made NGS the optimal technology to perform large-scale genome analysis studies (Margulies et al., 2005; Bentley et al., 2008).

The total amount of nucleotides sequenced per time unit exponentially increased due to massive parallelization. Sanger sequencing achieved a throughput of 0.166 Mb per hour while NGS technologies obtained ~20 Mb per hour in 2008 (Sinville and Soper, 2007; Morozova and Marra, 2008) and ~136 Gb per hour nowadays (in a NovaSeq 6000 S4 dual sequencer - *https://www.illumina.com*). Although read lengths for NGS technology are shorter than in Sanger sequencing, they achieve few hundred bases in length with around 99.9% sequence accuracy in billions of reads per run (Pfeifer, 2017).

In parallel, the sequencing-cost has extremely been reduced during the last 15 years. The total cost of the first draft of the human genome was estimated to be around 2.7 USD billion (Lewin et al., 2018). However, nowadays, the cost for re-sequencing a whole human genome at 30X coverage is around 1,000 USD (Check Hayden, 2014). Thus, both the reduced prize of sequencing and the high throughput achieved by NGS technology permitted to investigate and answer questions that few years ago were almost impossible to be studied.

NGS has enabled the analysis of genomes, transcriptomes, epigenomes, and microbiomes, among others. Since the advent of NGS technologies, many big projects have been successfully performed: many large eukaryotic genomes of a broad range of species have been fully assembled (Pagani et al., 2012); the ENCODE project (ENCODE Project Consortium, 2004; Harrow et al., 2012) characterized the complex structure and regulation of the human genome; the GTEx consortium helped to understand the gene expression complexity across different tissues of the human body (Ardlie et al., 2015; Consortium, 2017); population scale screening projects like the 1000 Genome Project (Consortium, 2010) have provided important findings about the diversity of the human genome; the *International Cancer Genome Consortium* (ICGC, *https://dcc.icgc.org/*) and *The Cancer Genome Atlas* (TCGA, *https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga*) have generated an amazing resource of cancer genomes from different tumor entities, which allowed to investigate the causes, prognosis and treatments of different cancers. These examples, plus many others not listed here, show

that NGS is an effective approach to rapidly generate huge amounts of data and novel insights.

The huge quantity of data generated in NGS studies resulted in many technical challenges that needed to be solved to obtain a complete and accurate record of sequences and to transform them into biologically meaningful results. Therefore, bioinformatics plays a critical role in current *omics* ('genomics, transcriptomics, epigenomics, proteomics etc.') research not only for processing of the massive amounts of genomic data, but also for unlocking the utility of these data for discovering novel knowledge in genomics. Finally, bioinformatics and biostatistics methods are crucial for the creation of computational models that reliably generate or evaluate various hypotheses.

3. NEXT-GENERATION SEQUENCING AND SHORT VARIANT DETECTION

Since the first personal genome was sequenced and more precisely, since sequencing of the first genomes with the Illumina technologies (Bentley et al., 2008), re-sequencing has been widely used for the research of personalized cancer genomics, the discovery of inherited or *de novo* mutations associated with Mendelian diseases, comparative genomics, to reconstruct human population history and to understand mutation processes (Li, 2014; Shendure et al., 2017).

Re-sequencing is a process based on the alignment of short reads against a reference genome in order to detect variants, which mark individual deviations from the reference genome. Variants can be characterized as single nucleotide polymorphisms (SNPs), which are by far the most common type of variants, as small insertions and deletions (indels), or as larger structural variants (SVs) and copy number variants (CNVs). The success of variant identification (often termed 'variant calling') relies on highly precise alignments of reads onto a reference genome and accurate variant calling algorithms that avoid false positive and false negative calls originating from alignment or sequencing issues (Li, 2014).

The typical next-generation re-sequencing workflow is usually split into sequencing, read pre-processing and quality control, read alignment, alignment post-processing, variant calling plus filtering, and variant annotation (Pfeifer, 2017). The NGS workflow starts with the library

preparation and finishes with the interpretation of the high-quality variant calls (Figure 1). Each one of these steps has particular limitations and issues, which require sophisticated algorithms to minimize their impact on the accuracy of the final call list.

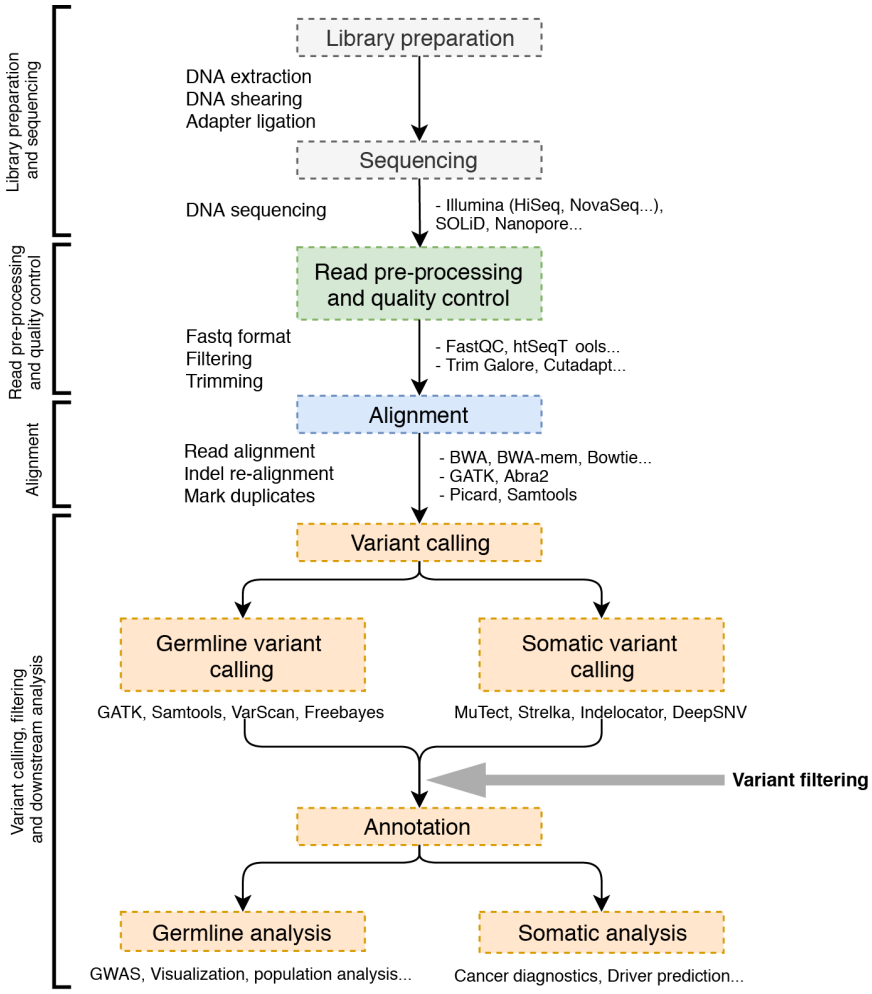


Figure 1 | Typical next-generation re-sequencing workflow.

3.1. Library preparation and sequencing

Before data generation, NGS protocols start with the DNA extraction and library preparation. Extracted DNA is sheared to short DNA fragments of few hundred nucleotides (e.g. 200-500bp) and platform-specific adapters

are ligated. Each of these steps can have important effects on sequence integrity and accuracy of the generated data.

Firstly, Costello et al described how oxidation of DNA during acoustic shearing generated 8-oxoguanine (8-oxoG) lesions, observed afterwards as C to A or G to T changes (Costello et al., 2013). This event is rare and only represents a low percentage of reads generated for a sample, and hence, does not have a huge impact on germline variant detection. However, the introduced errors can have important consequences for the ability to confidently call rare, sub-clonal mutations in e.g. tumor samples or bacterial populations. Secondly, although there are PCR-free protocols, PCR amplification is a step commonly required in many protocols. PCR can incorporate additional errors when generating duplicates of the original DNA fragment due to the *Taq* Polimerase error rate of around 1-20 x 10^{-5} (McInerney et al., 2014).

Finally, each sequencing technology yields erroneous or ambiguous data at particular genomic locations as consequence of systematic sequencing errors, alignment errors, or biases in coverage depending on the GC content (Table 1) (Pfeifer, 2017; Ardui et al., 2018). Nowadays, most genome re-sequencing studies are performed by sequencing-by-synthesis based platforms (Illumina HiSeq, MiSeq, NovaSeq, etc) (https://www.illumina.com). Several causes of sequencing errors of these platforms are well described: (1) crosstalk, occurring if dye frequencies of the laser-excited nucleotides overlap (miscalling A as C, G as T and vice versa) (Ledergerber and Dessimoz, 2011); (2) dephasing, which occurs when the incorporation of a nucleotide is missed in a cycle and the error is propagated to later cycles (leading to errors at the end of the reads); (3) an increase of errors at the end of the read because of reductions in signal intensity due to decreased enzyme activity (Kircher et al., 2009); (4) error in low complexity regions such as homopolymers resulting in false insertion or deletion calls; and/or (5) decreased coverage in regions of very high or low GC content which leads to low-quality base calls (Sleep et al., 2013).

Table 1 | Comparison of several sequencing platforms. Table adapted from Pfeifer, 2017 and Ardui et al. 2018 and complemented with in-house information for Oxford Nanopore sequencing technology.

	Next generation					Third generation*	
	454	Illumina			Ion Torrent	PacBio	Oxford nanopore
Platform	GS FLX+	HiSeq 2500	MiSeq	NextSeq 500	NovaSeq S4	PGM 318	Sequel II System	MinION R9.4
Run time	~ 24 h	~ 6 days	2-3 days	12-30 h	~ 44 h	4-7 h	~ 30 h	24 h**
Output / run	700 Mb	1 Tb	15 Gb	120 Gb	6 Tb	2 Gb	300 Gb	50 Gb
Read length	1 kb	2 x 125 bp	2 x 300 bp	2 x 150 bp	2 x 150 bp	400 bp	10 -16 kb (average)	18 kb**
Error rate	~ 1 %	~ 0.1 %				~ 1 %	~ 14 %	~ 10 % **
Primary errors	Indels	SNVs				Indels	Indels	Indels
Advantages	- Long reads - Relative fast run time	- Highest throughput of all platforms and lowest per-base cost. - Low per-base cost - Low error rate				- Unmodified nucleotides - No optical scanning necessary - Fast run time	- Very long reads - Does not require PCR amplification before sequencing - No bias based on GC content	- Very long reads - Does not require PCR amplification before sequencing - No bias based on GC content -Portable and easy use
Limitations	- High error rate in homopolymers. - Low throughput - High cost - Cumbersome emPCR	-Short reads - Overloading results in overlapping clusters and poor sequence quality - Requirement for sequence complexity. Problems in low-complex regions - Bias in coverage correlated with GC content.				- High error rate in homopolymers - Cumbersome emPCR	- High cost - High error rates	- High error rates, specially in homopolymers and low-complexity regions

* Single molecule real-time (SMRT) sequencing
** In-house calculations by Caspar Gross
References (Pfeifer, 2017; Ardui et al., 2018; Wick et al., 2019)

3.2. Read pre-processing and quality control

After sequencing, platform-specific software is used to obtain the nucleotide sequences in FASTQ format, which provides the read sequences and the corresponding ASCII-encoded PHRED quality scores (Cock et al., 2010). Raw sequencing data often contain complex artifacts and biases produced during the experimental and sequencing steps, which strongly influence the accuracy of the read alignments and consequently, the variant calling and genotyping. Tools like FastQC or htSeqTools (Planet et al., 2012) provide summary statistics of the sequencing performance such as nucleotide and base quality score distributions, as well as characteristics of the sequence (GC-content, k-mer fragments, levels of sequence ambiguity and PCR duplicates). These statistics provide a guidance for the selection of the quality control parameters and thresholds to filter potential problematic reads for later stages of the analysis (Pfeifer, 2017).

Finally, before the alignment step, sequence reads need to be cleaned from undesired sequences at the 3'- or 5'-ends (adapters, primers or barcodes depending on the protocol used to build the library) when reads are longer than the targeted fragments. Tools like *Cutadapt*, *AdapterRemoval* or *SeqPurge*, among many others, are designed for that purpose (Lindgreen, 2012; Sturm et al., 2016).

3.3. Alignment

The alignment is one of the most important steps in genome re-sequencing studies. Its accuracy affects massively the performance of variant calling, and therefore, plenty of algorithms have been developed in order to minimize as much as possible the number of alignment errors. The goal of an alignment algorithm is to map individual reads to the proper position in the reference genome from which they most likely originated (Pfeifer, 2017). Resulting alignments are usually stored in the SAM (sequence alignment/map) format, or its binary and compressed version (BAM format) (Li et al., 2009), which contain information about the location, orientation and alignment quality of each individual read. To this end, alignment algorithms like Bowtie (Langmead, 2010), Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) and BWA-mem (Li, 2013) were developed to identify the correct place of reads in the reference genome. Most alignment algorithms are based on the identification of short *kmers* from the reads in the reference genome ('seed') using e.g. suffix arrays or Burrows-Wheeler transformation, followed by the generation of an accurate pairwise alignment of the read to the most likely position on the reference genome by dynamic programming (Smith-Waterman or Needleman-Wunsch algorithms). During these processes reads can be falsely aligned to the wrong position in the genome, or the pairwise alignment at the correct position could 'misalign' parts of the read.

A particular challenging task during alignment is to place and correctly align short reads originating from repetitive or low-complexity genomic regions (LCRs). LCRs and repetitive regions represent an important fraction of the human genome (around 2 % and 45 %, respectively) and result in multiple possible locations of a read in the genome or multiple equally likely gapped pairwise alignments if indels are involved (Cordaux and Batzer, 2009; Wall et al., 2014). These ambiguous alignments subsequently lead to biases and errors in the variant calling procedure (Cordaux and Batzer, 2009; Wall et al., 2014). Furthermore, systematic alignment errors (non-random, recurrent errors) are found to be highly enriched in low-complexity sequences (homo-, di- and tri-polymers) due to 1) the high difficulty to define the exact position of a read or parts of the read in low-complexity sequences (Li, 2014), and 2) the high error rate of Illumina machines in homo- and di-polymers.

Another important step after the global read alignment is the local re-alignment of reads around indels. Reads spanning insertions or deletions

are often misaligned as most aligners have the tendency to introduce mismatches (SNPs) rather than gaps in the alignments (high penalties for gap opening) (Van der Auwera et al., 2013). Methods like Genome Analysis Tool-Kit (GATK) (Van der Auwera et al., 2013) or Abra2 (Mose et al., 2019) identify these suspicious intervals and locally realign reads in order to obtain a more concise consensus alignment.

As previously mentioned, some protocols require PCR amplification. An important consequence of this step is that DNA molecules might have been sequenced several times due to over-amplification and consequently, it leads to artifacts in the variant analysis (Li, 2014). Hence, it is crucial to identify and mark these 'duplicated reads' to allow the variant caller to ignore them (for instance, with tools like Picard or Samtools (Li et al., 2009)).

3.4. Variant calling and filtering

The process to generate high-quality variant calls remains challenging due to the complexity of errors that arose in any of the previous steps (i.e. systematic errors generated during DNA shearing, amplification, library preparation, sequencing or mapping). To solve this, a plethora of genomic variant prediction tools has been developed available to date, which can be divided in germline and somatic variant callers depending on the type of mutations that one wants to detect. The goal of these tools is to detect mutations such as short variants (also called point mutations), which are further divided into SNVs and short insertions or deletions (indels), as well as other more complex variants such as large-scale rearrangements, copy number alterations, inversion and translocations. However, in this section we only focus on short variant detection methods in both germline and somatic analysis.

A germline variant is defined as a genomic alteration inherited from progenitors and found in all cells of the organism in at least one haploid genome copy of each cell. Some of the most used germline variant callers for diploid genomes (e.g. for human genomes) are *GATK HaplotypeCaller* (McKenna, 2009; Van der Auwera et al., 2013), *Samtools mpileup* (Li, 2011), *Freebayes* and *Varscan* (Koboldt et al., 2009; Koboldt DC, Larson DE, 2013).

GATK HaplotypeCaller (McKenna, 2009; Van der Auwera et al., 2013) uses Bayesian models to genotype the genomic status of each locus based on

machine learning approaches trained on many samples. This can be performed in single- and in multi-sample analysis mode. *Samtools* (Li, 2011) performs the variant calling in two steps, first generating genotype likelihoods, and second, obtaining and filtering the calls. *Freebayes* is a single-sample Bayesian genetic variant detector that performs the variant calling based on the literal sequences of reads aligned to a particular target, and hence manages better the alignment problems than alignment-based variant detectors like *GATK* and *Samtools*. Finally, *VarScan2* (Koboldt et al., 2009; Koboldt DC, Larson DE, 2013) employs a robust heuristic approach to call variants that reach desired thresholds for read depth, base quality, variant allele frequency, strand bias filter, and statistical significance.

Although there are different algorithms for germline variant calling, they show similar performances in sense of quality of the final germline callset. However, in somatic variant detection, the algorithm or tool used influences strongly the accuracy of the final somatic callset (Alioto et al., 2015). A somatic variant is defined as a non-inherited mutation that occurs in any of the cells in a developing somatic tissue and can be transmitted to one of the descent cells (Clancy, 2008). The timing of this mutation in development and the posterior clonal expansion will affect the proportion of cells carrying a specific somatic mutation in a tissue, organ or organism (explained in more detail later).

Somatic variant calling algorithms try to identify mutations, which differ between tumor and normal (healthy) tissues from the same individual (strategy often called 'tumor-normal paired sequencing'). Some of the most widely used somatic variant callers are *MuTect/MuTect2* (Cibulskis et al., 2013; Van der Auwera et al., 2013), *Strelka/Strelka2* (Saunders et al., 2012; Kim et al., 2018), LoFreq (Wilm et al., 2012) and *VarScan2* (Koboldt et al., 2009; Koboldt DC, Larson DE, 2013). These tools use likelihoods of the variant model (Cibulskis et al., 2013), base-call qualities plus other sources of error information (Wilm et al., 2012), or random forest models trained on various call quality features (Kim et al., 2018) to detect somatic mutations. However, recent benchmarking studies reported substantial disagreement between somatic SNV and indel detection methods, especially for indels (Alioto et al., 2015), showing the necessity to develop post-filter algorithms to call variants at low allele frequency with high precision and sensitivity.

Once raw calls have been obtained from a variant caller, it is usually required to perform quality filtering to achieve high quality calls. Variant

callers have complex algorithms to clean false positive calls such as *Variant Quality Score Recalibration (VQSR)* in *GATK* (Van der Auwera et al., 2013), which uses machine learning approaches trained in many samples to filter low quality calls. However, systematic errors due to alignment or sequencing issues, which are not easy to be modeled, might remain and bias the quality of the final callset. Moreover, if somatic mutations at very low variant allele frequency (i.e. in cancer studies) are of interest, systematic errors must be distinguished and filtered in order to achieve good accuracy for real somatic mutations.

Several hard and universal filters have been suggested, although they are only able to capture a fraction of false positives due to systematic errors (Li, 2014) and have the potential to substantially increase the fraction of false negatives. They are based on: (1) filtering variants overlapping low-complexity regions, (2) removing sites with higher than expected depth of coverage (signal of mapping bias due to repetitive regions), (3) filtering sites where the fraction of non-reference reads is too low, (4) for somatic mutation calling only, removing alternative alleles frequently observed in the human population listed in e.g. the *1000 Genomes Project* or the *GnomAD* databases (Auton et al., 2015; Karczewski et al., 2019) or (5) filtering sites where the numbers of reference/non-reference reads are highly correlated with the strands of the reads (Guo et al., 2012).

However, a general issue of many post-filtration strategies is the use of hard thresholds for the various quality metrics, where small changes can dramatically influence false negative and false positive rates, or their dependence on large sample sets to be effective (e.g. VQSR) (Lek et al., 2016; De Summa et al., 2017). Therefore, novel strategies to remove both systematic errors and background noise are needed to get high precision calls, while not reducing the sensitivity.

3.5. Applications

The range of DNA re-sequencing applications has rapidly expanded over the last 13 years and continues to expand to date. The notable decrease of next generation sequencing (NGS) cost during the last decade has significantly changed biomedical and genomics research. Applications in clinical genomics, genome diversity studies with population-scale analysis, transcriptome and expression analysis, metagenome sequencing and developmental biology have become readily available to most researchers at low cost.

Clinical genomics has improved significantly during the last years. Nowadays, both whole genome sequencing (WGS, where no targeted enrichment is used and hence the whole genome is sequenced), whole exome sequencing (WES, which uses oligo-enrichment protocols to target all exons of the human genome) and targeted gene panels (where small regions of genome are targeted) are protocols often used in diagnostics or studies of genetic diseases and cancer, resulting in promising novel diagnostic tools with the potential to transform diagnosis of genetic diseases (Taylor et al., 2015; Schwarze et al., 2018). Early applications of WES rapidly discovered new genes for hundreds of Mendelians disorders and rare diseases, as well as causal coding germline *de novo* mutations in neurodevelopmental disorders (Taylor et al., 2015; Shendure et al., 2017). DNA re-sequencing applications quickly expanded to clinical cancer research discovering novel targets for therapies, cancer predisposition genes and cancer driver genes based on the analysis of mutations in large cohort studies, as widely performed by the ICGC and TCGA consortia (Birkeland et al., 2015; Colli et al., 2017; Bailey et al., 2018; Huang et al., 2018). In addition to WGS and WES of tumor biopsies commonly used for clinical cancer research and diagnostics, DNA sequencing of tumor-released circulating cells or cell-free DNA (cfDNA) has revolutionized the field of non-invasive diagnostics (Wan et al., 2017). Sequencing of cfDNA and the detection of circulating tumor DNA (ctDNA) in plasma or other body liquids enable applications such as early-diagnose of cancer patients, treatment response monitoring during therapy, relapse screening and prognostic prediction of relapse likelihoods (Diehl et al., 2008; Xi et al., 2016; Christensen et al., 2019). In parallel, the use of cfDNA has already been well stablished for non-invasive prenatal testing, where the simple counting of DNA fragments released into the maternal circulation by the fetus during pregnancy can help to detect chromosomal aneuploidies (Norton et al., 2015; Nshimyumukiza et al., 2018; Guy et al., 2019).

DNA re-sequencing is highly important for biomedical genomics research. Population-scale resequencing projects like 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) have helped to understand the diversity of the human genome. However, there are many other applications, which are frequently used and have helped to shape the current knowledge in genetics and cell biology. Firstly, the importance of RNA-sequencing (RNA-seq) protocols to characterize the transcriptome by shotgun sequencing of either full-length or 3' ends of cDNA has already been demonstrated in thousands of publications and by large-scale sequencing projects such as *GTEx* (Lonsdale et al., 2013; Ardlie et al., 2015; Consortium, 2017), ENCODE (Dunham et al., 2012) or GENCODE (Harrow

et al., 2012). Secondly, metagenome sequencing, which uses shotgun sequencing strategies to characterize complex communities of microorganisms, has emerged as a new field in genomics (Shendure et al., 2017; Costea et al., 2018). Moreover, a plethora of epigenome sequencing methods has been introduced over the last decade (Klemm et al., 2019). Furthermore, single cell DNA- and RNA-sequencing is used to understand how a single cell develops into a highly organized mass of cells (tissues, organs or tumors), or how expression differs in various cell types (Potter, 2018). Finally, analysis of multiple tissues from same individuals can help to better understand the differentiation of cells during development.

Although not considered second-generation sequencing (but third or fourth-generation), extremely long reads obtained by real-time, single-molecule sequencing technologies like *PacBio* or *Nanopore* allow not only to obtain better *de novo* assemblies of eukaryotic genomes, but also to investigate structural variants with higher precision and accuracy than achievable with short-reads technologies (Laszlo et al., 2014; van Dijk et al., 2018; Bowden et al., 2019).

As pointed out in previous paragraphs, DNA re-sequencing has many applications. However, in the next sections we will specifically focus on the sequencing of cfDNA to monitor cancer patients and the use of DNA re-sequencing of healthy individuals to find somatic and mosaic mutations.

4. THE USE OF LIQUID BIOPSIES FOR MONITORING CANCER PATIENTS DURING AND AFTER TREATMENT

4.1. Circulating Cell-free DNA

Circulating-cell-free or in short cell-free DNA (cfDNA) are DNA fragments released from cells mostly through apoptosis, necrosis, and (possibly) secretion into various body fluids such as bloodstream, urine, cerebrospinal fluid, pleural fluid and saliva (Botezatu et al., 2000; Jahr et al., 2001; Mithani et al., 2007; Sriram et al., 2012; Wang et al., 2015b) (see Figure 2). The size of these cfDNA fragments in healthy individuals is around 166 bps, which corresponds to the length of DNA wrapped around a nucleosome (around 147 bp) plus linker DNA associated with histone H1 (Wan et al., 2017; Mouliere et al., 2018). Moreover, the half-life of cfDNA is estimated to be between 16 minutes and 2.5 hours, and might be

influenced by the association of these fragments with cell membranes, extracellular vesicles or proteins (Wan et al., 2017).

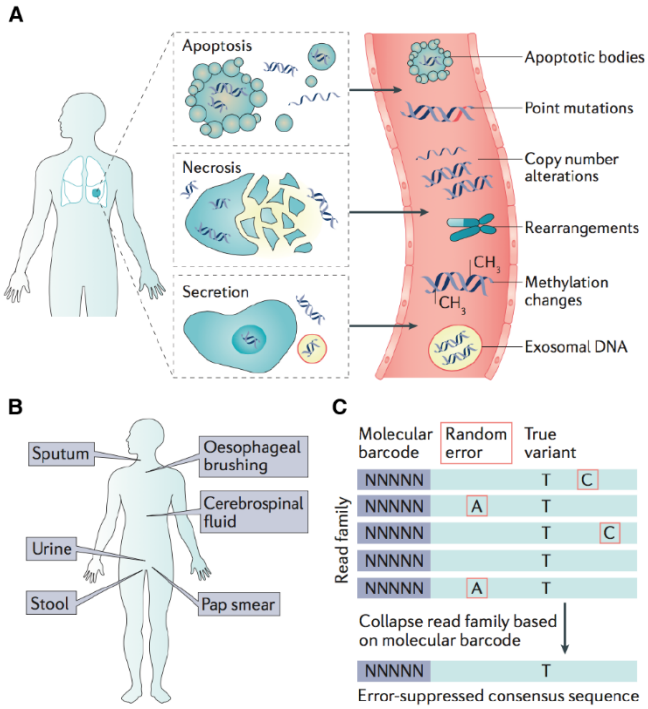


Figure 2 | The origin of the cell-free DNA. (A) Origins and alterations in cfDNA, (B) different body fluids where cfDNA can be released and (C) use of barcodes or unique molecular identifiers (UMIs) to detect sequencing and PCR errors. Figure taken from Wan et al., 2017.

In healthy individuals, the concentration of cfDNA varies between 1 and 10 ng / ml of plasma, representing around 330 to 3,300 haploid human genome equivalents (HGE). Most cfDNA in plasma originates from hematopoietic cells (Lehmann-Werman et al., 2016), nevertheless, small quantities of cfDNA from most organs of the human body are represented in bloodstream. Under specific physiological or clinical conditions like traumas, cerebral infarctions, transplantations, infections or cancer, the concentration of DNA fragments from different tissues of origin can change significantly (Wan et al., 2017; Zwirner et al., 2018a).

4.2. Circulating-tumor DNA and its use in cancer diagnostics

The term 'circulating-tumor DNA' (ctDNA) describes those cfDNA fragments that are released from cancer cells. The facts that ctDNA fragments can carry tumor-specific mutations, can be obtained through a non-invasive biopsy and that the cfDNA half-life is less than 2.5 hours makes ctDNA an interesting biomarker for obtaining a 'real-time' screenshot of the disease burden (Wan et al., 2017).

The high efficiency of PCR and NGS technologies enables novel liquid biopsy protocols to produce a huge clinical benefit through many different applications. First of all, it permits to obtain a more complete characterization of complex tumors compared to conventional sampling methods (invasive biopsies, e.g. punch biopsy), which have difficulties to obtain sufficient material to represent the proper genomic profile due to, for example, intra-tumor heterogeneity (explained later). Secondly, ctDNA sampling is a rapid and non-invasive biopsy method not requiring surgery or other invasive procedures. For these reasons, ctDNA analysis is considered a potential clinical strategy to perform screening for and early-diagnosis of cancer, to characterize the molecular tumor profile, for detection of tumor residual disease after treatment and finally, to perform monitoring of treatment response and tumor clonal evolution (Figure 3).

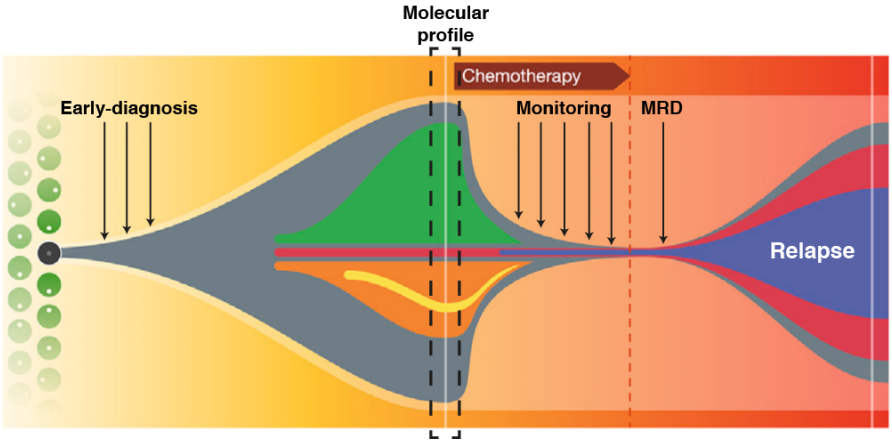


Figure 3 | Potential ctDNA applications in clinical cancer research. The diagram shows how cancer originated from one individual cell acquired cancer driver mutations initiating tumorigenesis and rapid proliferation of cells. Additional mutations at later time points lead to tumor heterogeneity, complexity and clonal expansion of the fittest tumor cells. After treatment (chemotherapy in this case) the tumor mass decreases but some clones resist treatment and form the basis for cancer relapse. Molecular residual disease (MRD) describes the situation were

residual tumor cells remain after treatment, and is a prognostic marker for relapse. Background figure edited and adapted from (Griffith et al., 2015)

4.3. Early-diagnosis using ctDNA

The diagnosis of cancer at early stages could allow earlier intervention and could improve survival. Bettegowda et al found ctDNA fragments in the plasma of around 82 % of cancer patients (solid tumors except brain) with advanced stages, while only 47% of patients at stage I had distinguishable ctDNA fragments (Bettegowda et al., 2014). They described that the capability to detect ctDNA fragments varied strongly among cancer stages, ranging from 10 fragments in 5 ml of plasma in patients at early stages (stage I) up to 100-1000 fragments in patients at advanced stages (stage IV). Additionally, these values also differed depending on the cancer type. Other body fluids than plasma may have a higher tumor DNA content for specific cancer types, as for example, urine for bladder cancers (Birkenkamp-Demtröder et al., 2016), stool for colorectal cancers (Sidransky et al., 1992) or cerebrospinal fluid for various brain cancers.

4.4. Detecting tumor heterogeneity using ctDNA

Cancer is a complex and heterogeneous disease. The intra-tumor heterogeneity occurs when different tumor regions from a single individual present different mutation profiles, which might also differ to the ones found in metastatic sites (Nowell, 1976; Griffith et al., 2015). It has been suggested and later supported that different cancer sub-populations of the same cancer mass have in general a common ancestral origin, but evolve over time by obtaining different sub-clonal mutations, which can also lead to differential speed of proliferation of sub-clones (Nowell, 1976; Gerlinger et al., 2012; Griffith et al., 2015) (see Figure 4). For this reason, an individual biopsy does likely not represent the full complexity and heterogeneity of the whole tumor. Hence, liquid biopsies are preferable, considering that all regions of the tumor release ctDNA fragments to the bloodstream (or other liquids). Therefore, liquid biopsy methods based on ctDNA sequencing have the potential to reflect the complex architecture of a tumor without the need of invasive sampling.

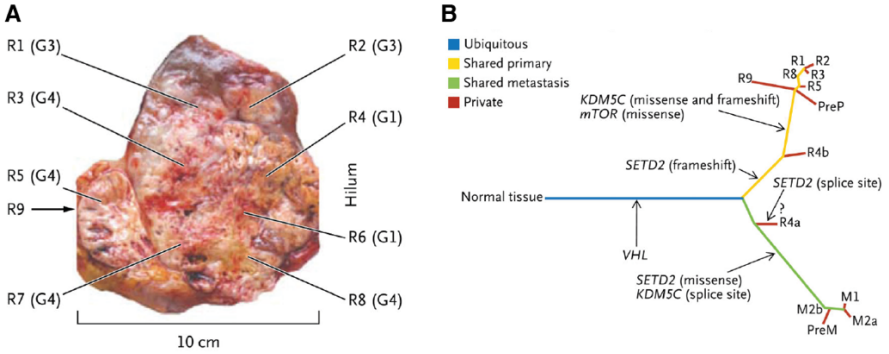


Figure 4 | Genetic Intra-tumor heterogeneity (A) and the correspondent phylogeny (B) in patient with renal carcinoma. Figure taken from Gerlinger et al., 2012.

4.5. Monitoring patients during and after treatment using ctDNA

The short half-life and the easy accessibility to plasma makes liquid biopsies a good choice for longitudinal monitoring of patients and their response to cancer or other types of therapy (Wan et al., 2017). Although it is still a novel approach, there are already some studies that demonstrated that ctDNA dynamics correlates with treatment response and might help to correctly measure treatment response earlier than other clinical detection methods such as tumor imaging.

In 2008, Diehl et al described the importance of ctDNA measurements to monitor tumor dynamics in subjects with cancer who were undergoing surgery or chemotherapy (Diehl et al., 2008). Several years later in 2016, Xi et al suggested that an early spike in ctDNA levels (specifically, an increase in the variant allele fractions of BRAF mutations) in the first week after the initiation of immunotherapy for patients with melanoma could predict response to treatment. The authors speculated that the observed surge in ctDNA frequency might reflect a transient increase in cancer cell death due to therapy (Xi et al., 2016). In 2018, Kurtz et al claimed that pre-treatment ctDNA levels and molecular responses were independently prognostic of outcomes in aggressive lymphomas (Kurtz et al., 2018). Finally, a recent study including 68 patients with localized advanced bladder cancer was able to associate the dynamics of ctDNA during chemotherapy with disease recurrence (Christensen et al., 2019).

These examples show that the longitudinal analysis of ctDNA is becoming an important clinical application enabling personalized medicine and helping oncologists to better diagnose and treat cancer patients.

Following surgery, radiochemotherapy (RCTX) and/or targeted drug treatments, the detection of ctDNA in plasma indicates the presence of minimal residual disease (MRD), i.e. residual tumor cells in the body that survived treatment, even when other clinical features or evidences are absent (Tie et al., 2016). Tests of 230 colorectal cancer patients at first follow-up after surgical resection showed that 90% of the ctDNA-positive (MRD) group suffered from relapse, compared to 0% of the ctDNA-negative group. With these results, Tie *et al* demonstrated that the detection of ctDNA at follow-up after treatment could also indicate poor prognosis, allowing the stratification of patients into high- and low-risk to relapse. (Tie et al., 2016).

4.6. Limitations in ctDNA analysis

All recent advances in ctDNA research and high-throughput sequencing highlight the potential of liquid biopsies for clinical applications. However, there are few limitations that need to be addressed in order to have highly accurate and reproducible results.

The concentration of ctDNA in plasma has been shown to correlate with tumor size and stage (Thierry et al., 2010; Bettegowda et al., 2014). Thus, the ctDNA fragment proportion compared to fragments originated from healthy cells depends on the individual's disease status. Scenarios where ctDNA concentrations are extremely low remain challenging (Wan et al., 2017). Additionally, background noise like oxidation damage during library preparation or systematic mapping and sequencing errors (or any other issues described above) affect sensitivity, specificity and false discovery rates and hence, they must be taken into account. Novel approaches or protocols try to use unique molecular identifiers, i.e. random barcodes attached to DNA fragments, to identify PCR duplicates. Subsequently, information about duplicates is used to reduce the background errors rates through the removal of amplification and sequencing errors (Schmitt et al., 2012; Newman et al., 2016a) (Figure 2C). These strategies permit to detect variants below 0.1 % (Newman et al., 2016a).

A second important limitation appears when variant calling is performed in large target regions, e.g. in panels of hundreds of cancer driver genes. The risk of having false positive calls increases with the number of genomic positions covered by the panel due to multiple hypothesis testing (Wan et al., 2017). For this reason, it is necessary to apply multiple test corrections and filters to increase specificity and precision, which at the same time decrease the sensitivity to discover ultra-low frequency mutations.

In addition to previous limitations, ctDNA analyses are usually performed on a few milliliters of plasma, which contain only around 20ng of DNA representing few thousands of haploid human genome equivalents. For this reason, increasing the theoretical sensitivity and specificity for mutations to below 1 in several thousand DNA fragments may not actually produce any gain in sensitivity due to a lack of unique genome equivalent in the tested DNA sample. It is indeed highly likely that a specific mutation with a variant allele fraction below 1/10,000 (0.01%) is not found due to the insufficient number of haploid genomes present in the sample. Therefore, new strategies and variant calling algorithms need to be developed to solve this problem.

Finally, recent studies have discovered the presence of mutations in genes associated to cancer (*NOTCH1, TP53, KRAS*…) in healthy individuals (Martincorena et al., 2015, 2018). These results show that mutations in specific genes should be considered carefully, as the effect of aging in some tissues of healthy individuals might be misinterpreted as early-stage cancer events.

ctDNA sequencing and the concept of liquid biopsy have the potential to revolutionize biomedical research as well as cancer diagnosis and prognosis. However, there are several important limitations that must be addressed and hence, new algorithms and strategies are needed in order to generate highly accurate results necessary for applications in personalized medicine.

5. SOMATIC AND MOSAIC MUTATIONS IN HEALTHY INDIVIDUALS

The acquisition of DNA mutations during life is unavoidable. Despite the existence of several cell mechanism that preserve genome integrity, cells accumulate mutations during development and life due to aging and many

environmental influences. Accumulation of mutations over time generates populations of cells with different genomic profiles in the same individual, a phenomenon called somatic mosaicism (Acuna-Hidalgo et al., 2016). Moreover, if these mutations are present in gametes, they can be passed to the offspring and might contribute to evolution, i.e they can be positively selected, have neutral effects or lead to sever disorders or abnormalities.

5.1. Importance of timing in somatic mosaic mutations

Postzygotic mosaic mutations, which are those mutations acquired after the fertilization of the egg, lead to the coexistence of distinct cell populations in a single individual (Biesecker and Spinner, 2013; Acuna-Hidalgo et al., 2016).

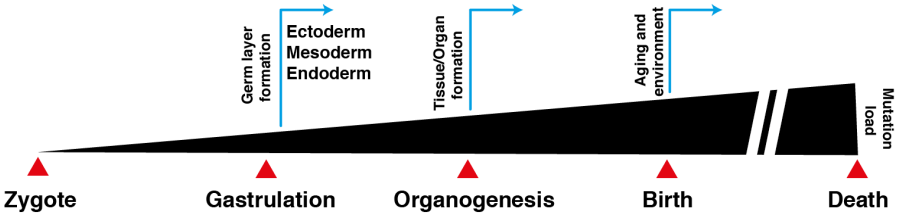


Figure 5 | Different stages of human embryogenesis and life.

Since fertilization, the zygote divides and clonally expands during embryonic development up to around 10^{13} - 10^{14} cells constituting the human body (Bianconi et al., 2013). After fertilization, the one-cell zygote starts a series of cell divisions in which the embryo increases the cell number while maintaining its overall size and a round shape to finally create the blastocyst, a process called cleavage (Rossant and Tam, 2017). Afterwards, the blastocyst is implanted in the uterus and during the stage of gastrulation the three germ layers (ectoderm, mesoderm and endoderm) are formed. These germ layers will further differentiate into tissues and organs through two important processes called histogenesis and organogenesis (see Table 2 and Figure 5) (Yamada et al., 2010; Rossant and Tam, 2018). Finally, the fetal development stage begins around the 9th week and continues until birth (Figure 5).

Table 2 |Tissues derived from the three germ layers (ectoderm, mesoderm and endoderm).

Germ layer	Derived tissues
Ectoderm	Brain, amygdala, pituitary, minor salivary gland, tibial nerve, medulla of adrenal gland, pigment cells…
Mesoderm	Cardiac muscle cells, skeletal muscle cells, adipose cells, kidney, spleen, uterus, gonad organs…
Endoderm	Lung, liver, thyroid, colon, esophagus, stomach, intestine…

The timing of mosaic mutations and the posterior apoptosis and cell migration determine the mosaic pattern of each individual (Biesecker and Spinner, 2013) and the percentage of affected cells in organisms, tissues or groups of tissues (Campbell et al., 2015; Acuna-Hidalgo et al., 2017). For instance, mutations occurring early in embryogenesis (cleavage, blastulation, gastrulation) can be present in a substantial proportion of cells in postnatal humans and therefore, have particularly high likelihood to effect the phenotype or cause disease (Ju et al., 2017) (Figure 6). Moreover, mutations occurring earlier in development should be present in more tissues and in greater proportion of cells, although they don't expand in a symmetric way to adult somatic tissues, as Ju et al claimed (Ju et al., 2017).

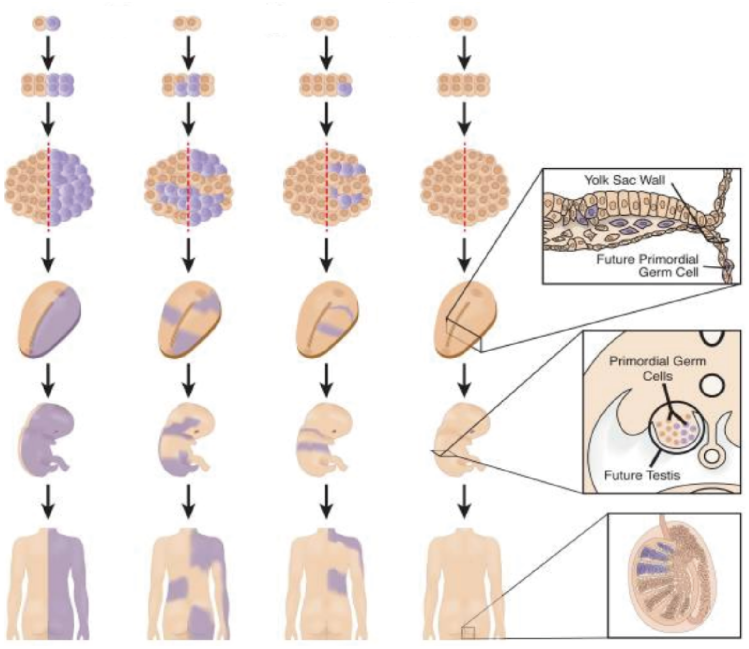


Figure 6 | Importance of timing in mosaic mutations. Figure taken from Campbell et al., 2015).

5.2. Mosaic mutations during human development

Mosaic mutations have been associated with a broad range of genetic diseases (Campbell et al., 2015), including neurological disorders (Poduri et al., 2013; Halvorsen et al., 2016), brain malformation and overgrowth syndromes (Lindhurst et al., 2011; Rivière et al., 2012), autism spectrum disorders (Yurov et al., 2007), and cancer predisposition syndromes (Prochazkova et al., 2009; Ruark et al., 2013). However, although sometimes causing disorders, studies in single tissues demonstrated that mosaic mutations also occur in normal (healthy) tissues (Acuna-Hidalgo et al., 2015; Ju et al., 2017; Wei et al., 2018a). Interestingly, Acuna-Hidalgo and colleagues found that around 7% of presumed germline *de novo* mutations causing rare disease cases treated in their hospital were in fact post-zygotic mosaic mutations (Acuna-Hidalgo et al., 2015). Using whole-genome sequencing of normal blood from 241 adults, Ju et al estimated that approximately three mutations are accumulated per cell division during early embryogenesis (Ju et al., 2017), each of which could hit an essential gene and/or cause severe genetic diseases.

The detection of mosaic variants in tissues is challenging because it requires analysis of many cells within a given tissue and it may be tissue-specific or tissue-limited (Acuna-Hidalgo et al., 2016). Therefore, the detection of the mosaicism in the tissue in which it occurs may require analysis of multiple tissues within an individual. A comprehensive study of all tissues of an individual in a large cohort of individuals has indeed not been performed so far.

5.3. Somatic mutations during life in healthy individuals

Somatic mutations events are not limited to prenatal development and also occur frequently after birth due to environmental effects (sunlight, mutagenic agents etc.) or simply aging (Risques and Kennedy, 2018). Age-related disorders like cancer emerge through the accumulation of somatic mutations during life, creating complex genetic heterogeneity and clonality within tissues or organs. However, as around half of these mutations arise years or even decades before tumor initiation (Tomasetti et al., 2013), it opens the possibility that somatic variants acquired during development and life might be present in non-malignant human tissues (Wei et al., 2018a).

An increase of somatic mutations and clonal expansion events in healthy adults has been reported for peripheral blood, esophagus and skin (Jaiswal et al., 2014; Xie et al., 2014; Martincorena et al., 2015, 2018; Yokoyama et al., 2019). Studies in blood of elder healthy individuals discovered recurrent mutations in genes implicated in myelogenous leukemia (such as *DNMT3A*, *TET2*, *ASXL1* and *JAK2)*, suggesting that these clones might represent early stages of leukemic progression (Jaiswal et al., 2014; Xie et al., 2014).

In the analysis of 74 cancer genes across 234 biopsies of normal skin, Martincorena et al found between 2-6 somatic mutations per megabase, which also exhibited ultraviolet light exposure signatures (Martincorena et al., 2015). Moreover, they also described that some cancer genes were under positive selection and clonal expansion (*NOTCH1*, *TP53* and *FGFR3*) (Martincorena et al., 2015).

Finally, a study of biopsies of normal esophagus (in middle-aged and early donors) has also shown the presence of clones with cancer-associated mutations, caused mainly by intrinsic mutational processes, with *NOTCH1* and *TP53* mutations affecting a high proportion of cells (from 12 to 80 % and 2 to 37 %, respectively) (Martincorena et al., 2018). Additionally, *NOTCH1* mutation frequencies in normal esophagus were found to be several times greater than in esophageal cancer, suggesting a different function (oncogene or tumor suppressor) of *NOTCH1* from the oncogenic function described for instance in leukemia. In parallel, in 2019, both Yizhak *et al* and Yokoyama *et al* found similar results for normal esophagus tissues (Yizhak et al., 2019; Yokoyama et al., 2019).

It is maybe not surprising that mutations conferring proliferative advantages develop into larger and highly-clonal cell populations through the pass of life. However, all these findings have many implications on the way to understand both cancer development and aging.

6. OBJECTIVES

As DNA re-sequencing becomes more and more important for diagnostics of genetic diseases and cancer and can help to make clinical decisions, it also became more and more important to assess the accuracy of variant calls and to understand biases and sources of errors in sequencing and bioinformatics methods. False positive and negative calls in clinical studies

affect dramatically the downstream analysis and bias the interpretation of the results, leading in some cases to a wrong diagnosis or prognosis or suboptimal treatment choices. For this reason, the general goal of this thesis is to develop methods to reduce the amount of false positive calls enriched in both germline and somatic variant analysis, as well as to maximize the sensitivity for the detection of true variants, especially for somatic mutations at very low fraction in cancer patients or normal tissues.

The accuracy of variant detection highly depends on the capacity to distinguish and understand the possible biases that NGS data might have. For this reason, first of all it is crucial to characterize errors that, in general terms, can be summarized as follows:

$$\text{Errors} \sim \text{Systematic errors} + \text{PCR errors} + \text{Sequencing errors} + \text{Others}$$

I. The first part of this thesis (chapter 1) was focused on detecting and characterizing systematic errors. Although random sequencing errors can be modelled statistically and deep sequencing minimizes their impact, systematic errors remain a problem even at high depth of coverage. Therefore, understanding their source is crucial to increase precision of clinical NGS applications. Hence, in the chapter 1 of this thesis, we tried to achieve the following objectives:

- Identifying genomic sites prone to systematic alignment and sequencing errors with the analysis of allele balance bias in a cohort of 987 WES individuals.
- Computing a variant callability score (ABB score) for each position of the human exome, which is able to distinguish systematic and recurrent errors.
- Validating and benchmarking the utility of ABB score to detect false positive calls in somatic and germline variant calling, as well as its utility for finding artifacts and false associations in rare variant association studies.

II. The second part of this thesis (chapter 2 and Appendix) had the main goal of distinguishing mutations at very low allele frequencies for monitoring cancer patients during treatment using cell-free DNA samples from plasma. Once systematic errors are characterized, background noise, mainly caused by sequencing and PCR errors, needs

to be removed to call variants at ultra-low frequency. Hence, the objectives of this chapter 2 were:

- Developing an algorithm to detect and remove sequencing and PCR errors in cfDNA sequences with the use of unique molecular identifiers.
- Creating a variant caller modeling the remaining errors and calling somatic variants at extremely low allele frequency.
- Performing longitudinal analysis of cancer mutation kinetics in 20 head and neck squamous-cell carcinoma (HNSCC) patients across four different time points during treatment (radiochemotherapy), and one extra time point at the first follow-up after treatment in order to detect treatment response.
- Detecting minimal residual disease after treatment and using it to predict poor cancer prognosis.

III. Finally, the third part of this thesis (chapter 3) was focused on the detection of somatic and mosaic mutations in 10,097 RNA-seq samples from up to 49 different tissues of 570 healthy individuals from the GTEx project. Thus, the objectives of this chapter were:

- Developing an algorithm and variant calling method to detect somatic mutations in RNA-seq data.
- Integrating the multi-tissue, multi-individual calls and the human embryonic development tree information to detect and characterize mosaic mutations in healthy individuals during embryogenesis and life.
- Investigating the relation of somatic mutations and age in multiple tissues of healthy individuals.

CHAPTER 1

DETECTION, CHARACTERIZATION AND IMPORTANCE OF SYSTEMATIC ERRORS IN RE-SEQUENCING STUDIES

The enrichment of false positive calls in re-sequencing studies might have important consequences for downstream analysis. Decisions such as the selection of treatment, diagnosis and prognosis of some disease are based on the calls obtained from conventional somatic or germline variant calling pipelines and hence, polishing final callsets to remove remaining errors is crucial to achieve reliable results. Although most of the errors that are random can be modelled statically and high depths of coverage might help to detect them, systematic errors, which are found recurrently across samples, remain even at high depths and are usually not detected by standard aligners and variant callers.

For this reason, in this chapter we have developed an algorithm named ABB ('Allele Balance Bias') able to detect systematic errors in human genome re-sequencing studies. We defined allele balance bias as a recurrent deviation of observed from expected proportion of reads supporting the alternative allele at a genomic position in a large number of samples. The analysis of ABB in a cohort of around 1000 whole exome sequencing (WES) samples has permitted us (1) to build a model able to identify sites in human genome prone to systematic error, (2) to create a genotype callability filter able to remove this type of errors from germline and somatic mutation studies and (3) to detect false gene-disease associations in rare variant association studies.

My contribution in this project has been to design, analyse and build the ABB model, as well as to perform the benchmarking analysis of the tool.

Muyas F, Bosio M, Puig A, Susak H, Domènech L, Escaramis G, et al. Allele balance bias identifies systematic genotyping errors and false disease associations. Human mutation. 2019;40(1):115–26. DOI: 10.1002/humu.23674

CHAPTER 2

IDENTIFYING SOMATIC MUTATIONS IN CELL-FREE DNA FROM BLOOD PLASMA TO MONITOR CANCER PATIENTS PRE-, DURING AND POST-TREATMENT

Characteristics such as short half-life, easy accessibility (non-invasive biopsy) and complete representation of tumor mutation profiles show the potential of cfDNA analysis to become a method for monitoring cancer patients during treatment and different stages of the disease.

In this chapter (adapted version from the manuscript *Dynamics of circulating cell-free tumor DNA in HNSCC patients receiving radiochemotherapy correlates with treatment response* paper – in preparation), we have performed the longitudinal analysis of the ctDNA kinetics in plasma of 20 head and neck squamous carcinoma patients pre-, during and post-treatment to understand how cfDNA behaves compared to treatment response.

My contribution to this project has been the complete computational and statistical analysis related to cfDNA sequencing data. To this end, I have developed different algorithms to use unique molecular identifiers (UMIs) barcodes to correct background noise and obtain highly accurate variant calls at very low allele frequency. Please, find detailed and expanded methods for barcoded deduplication and variant calling in Appendix. Furthermore, I have developed statistical methods to 1) correlate cfDNA levels with response to treatment over time, 2) identify residual tumor cells after treatment by measuring minimal residual disease in plasma, and 3) identify onco-viral DNA in plasma.

F. Hilke*, F. Muyas*, J. Matthes, I. Bonzheim, S. Welz, S. Ossowski, O. Rieß, D. Zips, C. Schroeder, K. Zwirner. Dynamics of circulating cell-free tumor DNA in HNSCC patients receiving radiochemotherapy correlates with treatment response. *In preparation*.

DYNAMICS OF CIRCULATING CELL-FREE TUMOR DNA IN HNSCC PATIENTS RECEIVING RADIOCHEMOTHERAPY CORRELATES WITH TREATMENT RESPONSE

F. Hilke[1, 2*], F. Muyas[1, 3*], J. Matthes[1], I. Bonzheim[4], S. Welz[5, 6], S. Ossowski[1], O. Rieß[1], D. Zips[5, 6], C. Schroeder[1], K. Zwirner[5]

[1] Institute of Medical Genetics and Applied Genomics, Medical Faculty and University Hospital, Eberhard Karls University Tübingen, Tübingen, Germany
[2] Department of Dermatology, Venereology and Allergology, Charité - Universitätsmedizin Berlin, 10117 Berlin, Germany.
[3] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[4] Institute of Pathology and Neuropathology, Comprehensive Cancer Center and University Hospital Tuebingen, Tübingen, Germany.
[5] Department of Radiation Oncology, Medical Faculty and University Hospital, Eberhard Karls University Tübingen, Tübingen, Germany
[6] German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ) partner site Tuebingen, Germany

* These authors contributed equally to this publication and are shared first authors.

Abstract

Purpose: In patients with locally advanced head and neck squamous cell carcinoma (HNSCC), definitive radiochemotherapy (RCTX) is a standard treatment option. However, in spite of intense treatment, two-year overall survival is as low as 60 %. Therefore, novel biomarkers for patient stratification and prediction of treatment success are needed. We tested the prognostic capacity of circulating-tumor DNA (ctDNA) before, during and subsequent to RCTX in patients with HNSCC.

Patients and Methods: We sequenced solid tumors and normal samples of 20 patients with locally advanced HNSCC receiving a primary RCTX to identify driver mutations, and determined the HPV-status of each patient by p16 staining. Subsequently, we performed a longitudinal analysis of circulating tumor DNA dynamics under RCTX by monitoring of driver mutations and HPV levels in the plasma prior, during and after treatment (5 time points, n = 99).

Results: Overall, we detected ctDNA or circulating viral DNA (cvDNA) in 85% of all patients. The pre-therapeutic ctDNA fraction was significantly correlated with the gross tumor volume (p-value 0.032) and ctDNA levels showed a negative correlation between the tumor allele fraction in the

plasma and the course of treatment (p-value < 0.05). Additionally, if ctDNA was still detectable at the first follow-up (molecular residual disease - MRD), the patient presented with a recurring disease later on. Circulating viral DNA (HPV16/18) could be detected in 4 patients, showed a similar dynamic behavior to the ctDNA during treatment, and completely disappeared after treatment in all cases. Hence, circulating HPV DNA mainly originated from tumor cells, which harbor multiple copies of the virus, and therefore can be seen as a promising plasma-based surrogate marker of tumor size.

Conclusion: The detection of ctDNA and cvDNA in plasma of patients with locally advanced HNSCC is feasible, could support the surveillance of treatment response, and the dynamic changes of ctDNA levels throughout the therapy seem to be prognostic for the recurrence of the disease.

Keywords: head and neck cancer, HPV, cfDNA, ctDNA, liquid biopsy, biomarker

1. INTRODUCTION

Head and neck squamous cell carcinomas (HNSCC) represents a relatively high number of the cancers worldwide, with roughly 700,000 newly diagnosed cases in 2018 (including the oral cavity, oropharynx, pharynx and larynx) (Bray et al., 2018). The three major etiologies for the development of HNSCC are tobacco use, heavy alcohol consumption and the infection with the human papillomavirus (HPV), which is especially associated with oropharyngeal squamous cell carcinomas (Leemans et al., 2011, 2018).

In unresectable, advanced tumor stages the primary combination of radiation and chemotherapy (RCTX) with curative intention is a standard treatment option. The treatment regimen includes a cumulative radiation dose of about 70 Gy applied within 6-7 weeks and a concomitant chemotherapy with cisplatin (Pignon et al., 2009) or optionally with 5-fluorouracil (5-FU) and mitomycin C (MMC) (Budach et al., 2015). Despite of this intense treatment, a recent multicenter retrospective study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG) reported a two-year overall survival (OS) rate of only 59.6% (Linge et al., 2016). To date, active treatment monitoring during RCTX is not routinely performed and the follow-up is based on clinical examinations and imaging modalities. This allows an approximation of the therapeutic success over time but gives no insights into the existence of residual disease. Circulating tumor DNA (ctDNA) is a potential biomarker for patient stratification, treatment response assessment and post-treatment tumor surveillance in HNSCC patients (Wang et al., 2015b; Muhanna et al., 2017; Tinhofer and Staudte, 2018). In addition, monitoring of HPV viral particles in plasma was shown to be a surveillance marker for disease recurrence and of prognostic value (Ahn et al., 2014; Wang et al., 2015b; Jeannot et al., 2016).

Thus, in this study we tested the capacity of ctDNA and circulating viral DNA (cvDNA) to monitor treatment response during combined RCTX and to identify molecular residual disease post treatment. The present study provides data of an ultra-sensitive NGS-approach to detect ctDNA and cvDNA dynamics pre-, during- and post RCTX in patients with HNSCC. To our knowledge, this is the first report of ctDNA dynamics during primary RCTX in HNSCC and the according correlation with outcome parameters.

2. PATIENTS AND METHODS

Patients and Clinical Samples

In this prospective pilot study twenty patients with locally advanced HNSCC were enrolled between 2015 and 2016. All participants declared written informed consent, and the study was approved by the local ethics committee (577/2014BO2). The patients received a definitive radiochemotherapy (RCTX) after primary diagnosis by intensity-modulated radiotherapy (IMRT) with a cumulative radiation dose of 70-77Gy and concomitant chemotherapy either with cisplatin weekly or a combination therapy of 5-FU and MMC.

For the analysis of ctDNA, blood samples were collected at 5 times: prior to therapy as baseline (T1), 3 times during therapy to follow the ctDNA kinetic and dynamic (T2-4) and subsequent to chemoradiation (T5) to evaluate the treatment outcome (the time line is shown in the Figure 7). Clinical investigations and computed tomography (CT) scans were terminated 6 respectively 12 weeks after treatment for the first follow-up and consecutively every 3-6 months. Recurrent disease or progression were diagnosed by imaging and endoscopic follow up and - if possible - by histology.

Targeted Panel Sequencing and Bioinformatics Analysis

A HNSCC specific cancer panel containing 327 genes (Eder et al., 2019) was used for the library preparation of DNA from formalin-fixed paraffin-embedded (FFPE) tumor tissues and blood samples were used as normal tissue control. An in-solution capture of the exonic regions was performed using the Agilent HaloplexHS technology (Agilent, Santa Clara, CA) followed by paired-end sequenced using the HiSeq2500 instrument (Illumina, San Diego, CA). Data analysis, quality control and calling of somatic single nucleotide variants, insertions and deletions were performed with an in-house developed pipeline, called “megSAP” as published before (Zwirner et al., 2019). Identification and clinical interpretation of driver mutations was performed using the Cancer Genome Interpreter (Tamborero et al., 2018).

CtDNA and HPV monitoring by ultra-deep sequencing with unique molecular barcodes

We collected peripheral blood samples in Streck (Streck, La Vista, Nebraska) and EDTA tubes (Sarstedt, Nümbrecht, Germany) for the isolation of cell-free DNA (cfDNA), using the QIAmp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany). In each patient, we sequenced the genomic regions of all 135 driver mutations, identified from the total cohort, and the E7 viral DNA sequence from the HPV16 and HPV18 strain at all available times (T1-T5; n = 99). The input amount of cfDNA was limited to either 9 ng, 13 ng or 15 ng per patient and sample. This is equal to 2700– 4500 haploid genome equivalents (hGEs). Therefore, we tried to achieve a minimum fragment recovery of 1000 genome equivalents, which would allow us to detect variants at a minimum allele frequency of 0.1 percent.

To ensure correct variant calling at even low ctDNA proportion we used the unique molecular identifier technology provided by the SureSelectXT-HS kit (Agilent, Santa Clara, CA). For deduplication, fragments were identified and grouped by unique molecular barcodes of 8 bps plus the coordinate information. Then, sequencing and PCR errors were removed with the base-to-base comparison between these PCR duplicates in order to create collapsed fragments (deduplicated reads). Then, resulting reads were processed with the *BamClipOverlap* tool (*https://github.com/imgag/ngs-bits*) to softclip paired-end reads that overlapped. Each one of the final consensus fragments represented a recovered DNA fragment.

For the variant calling step, information of duplicates and error rates were taken into account to calculate error probabilities in a beta-binomial model (method paper in preparation) to obtain the potential calls. Only reads with mapping quality greater than 30 and nucleotides with base quality greater than 20 were considered in the variant calling step. Additionally, only variant sites that passed the beta-binomial model and with at least 2 alternative reads were considered as PASS calls.

A sample was considered ctDNA positive if at least one variant was detected in the plasma that was also observed in the primary tumor of the patient.

Molecular residual disease (MRD)

The *MRD* (Molecular Residual Disease) value summarizes all variant information per sample into one variant calling value (sample-specific value). This value combines and collapses the significance of all variants analyzed as one value. The p-values of all variants (obtained from variant calling step) are combined into one value using Fisher's combined probability test. Afterwards, the resulting p-value is transformed in a 1-100 scale using –log10. Therefore, MRD values, which range from 0 to 100, represent the significance of ctDNA fragments presence in the analyzed sample, giving as maximum significance the value 100. Finally, we considered as significant MRD > 1.3, what is equivalent to p-value < 0.05.

Statistics and Data Correlation

Statistical significance was defined as $p < 0.05$. Events were defined as follows: overall survival (OS), death of any cause; disease-free survival (DFS), loco-regional or distant failure or death of any cause. For longitudinal analysis of ctDNA levels we used a combination of linear regression for removal of confounding effects and Spearman correlation test for correlating ctDNA and treatment dosage. See supplementary materials for extended and detailed methods.

3. RESULTS

The study included 20 patients with locally advanced HNSCC, located in the oropharynx (n=14), hypopharynx (n=4) or in the oral cavity (n=2). In our 20 patients, 99 of 100 planned blood samples could be collected. One patient did not show up at the first follow-up due to inpatient treatment in another clinic. The median follow-up was 823 days (range: 135 – 1168) with 13 patients still alive at the time of analysis. Ten patients developed either a local or distant relapse, whereat the majority within the first year of follow-up. The genetic analysis revealed 667 somatic alteration (synonymous and non-synonymous), of which 127 were annotated as driver mutation, resulting in a median of 4 driver mutations per patient (min = 1, max = 52). The major affected signaling cascades were the TP53 - , NOTCH -, HIPPO -, PI3K – pathways and members of the chromatin modification (Supp. Figure 1). Additionally, five of the patients were screened positive in the pathology for HPV infection by p16ink activation/inactivation analysis (patients 2, 4, 8, 14 and 15).

Detection rate of ctDNA and HPV - E7 genes in the cohort

For monitoring of treatment response, we analyzed the 5 consecutive blood samples taken from each patient (n = 99) (see Figure 7). Overall, we had a positive detection rate of 85%, whereby in 17 out of the 20 patients we detected ctDNA. In three patients (2, 8 and 23) we could not detect ctDNA likely due to variant calls of low quality or low variant allele frequency (VAF) from the solid tumor biopsy, undetectable levels of ctDNA or insufficient fragment recovery and coverage depth. Therefore, we excluded these patients from downstream analysis. In the remaining 17 individuals, we detected circulating viral DNA (HPV) in four patients (patients 4, 5, 14 and 15), of which 3 were validated by p16k analysis in solid tumor. Even though the total amount of cfDNA at each blood collection (median: 9.98 ng/ml, range: 2.89 – 172.9 ng/ml) was limited we achieved an average sequencing depth of 23,206X before and 2049X after deduplication with molecular barcodes (see methods), representing on average 2049 hGEs. This enabled us to detect somatic alterations in the plasma with variant allele fractions (VAF) as low as 0.1 percent, with sensitivity depending on the sequencing depth and error correction efficiency obtained in deduplication with molecular barcodes (see supplemental methods).

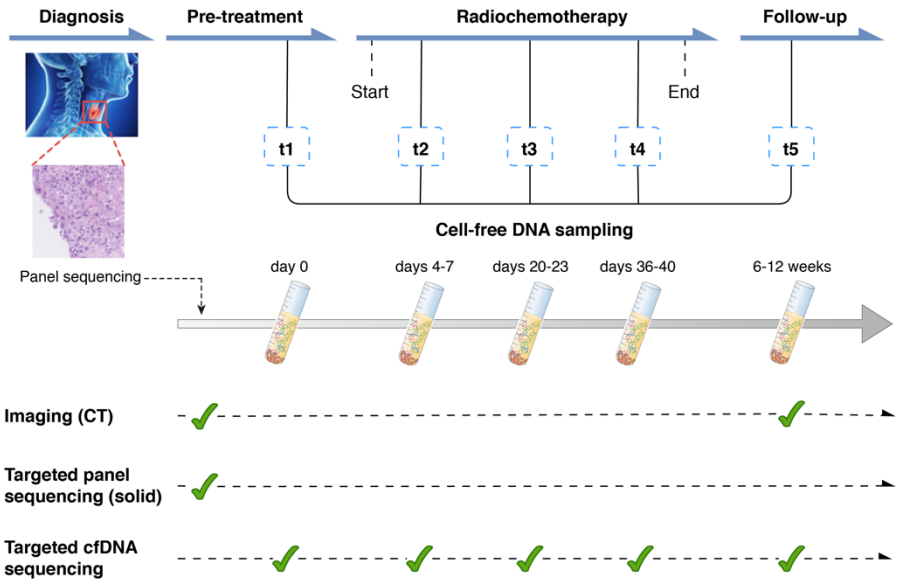


Figure 7 | Timeline for cfDNA sampling, treatment regime and ctDNA analysis. The overview shows the outline of the study design from diagnosis, to the recruitment and treatment until the first follow-up. All solid tumor biopsies were sequenced prior to the analysis of the cfDNA. The lower part shows when the five consecutive blood samples (t1-t5) were taken and how they were analyzed.

Correlation of ctDNA in plasma with the gross tumor volume and fragment proportion

We observed a positive correlation of the macroscopic tumor burden according to the gross tumor volumes (GTVs) in the planning CTs with the allele frequencies of driver mutations observed in the plasma before treatment initiation (T1) (p-value < 0.05 with Pearson correlation test). Bigger volumes of the primary tumors (PT) and their involved lymph nodes (LN) were associated with higher allele frequencies (Supp. Figure 2A).

As already described in literature (Mouliere et al., 2018), ctDNA fragments are substantially smaller (around 90-150 bps) than fragments originated from healthy cells (166 bps). Therefore, we checked the proportion of cfDNA fragments in cfDNA in the range of 90.150 bps. Our analysis revealed that samples with higher ctDNA fractions showed greater proportions of cfDNA fragment in the range size of 90-150 bp (p-value = 0.001 with Pearson correlation test, see Supp. Figure 2B).

Treatment surveillance and ctDNA kinetics

We next investigated the kinetics of ctDNA levels in plasma in response to the combined RCTX. To this end, we surveyed the level of ctDNA in the plasma, as well as the total cfDNA amount and the presence of a manifest infection during radiochemotherapy (Figure 8). The latter two measures are considered confounders that influence the proportion of ctDNA to total cfDNA in the bloodstream, as shown before (Zwirner et al., 2018b). After removal of confounding effects, we observed a dynamic of ctDNA levels, which showed a clear time and dosage dependency. Throughout the treatment, the allele frequency of the tumor- and patient-specific alterations decreased from a median of 1% at T1 to 0.01% at T5 (see Figure 8). Additionally, 7 patients (3, 7, 13, 14, 16, 18 and 20) showed a significant negative correlation (p-value < 0.05) between the tumor allele fraction in the plasma and the course of treatment (Supp. Table 1). Longitudinal analysis of another 8 patients also showed a decrease in the ctDNA levels over time, although non-significant. One patient that did not respond to RCTX indeed showed a positive relation between tumor allele fraction and dosage, i.e. an increase of ctDNA levels over the course of treatment (patient 5). In addition to patient 2, 8 and 23, previously removed from all analysis due to technical issues, patient 11 was excluded from this correlation analysis because had only one variant called, which in some time points did not reach the minimum coverage (see Figure 8).

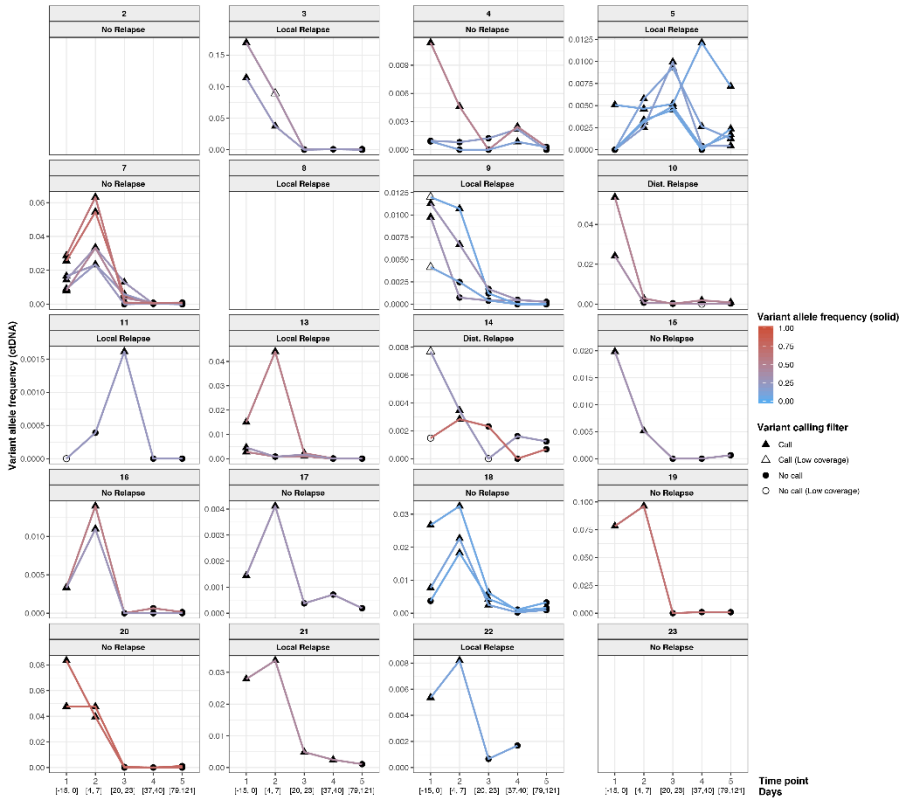


Figure 8 | Variant allele frequencies (VAF) of monitored driver mutations in plasma at different time points during treatment. Each individual picture represents a patient and its relapse status. Lines connect VAF measures of the false same mutation between time points and their colors represent their VAF in the solid tumor (Variant allele frequency (solid) legend). The shapes of the points show the result and significance of variant calling (legend Filter).

Detection of molecular residual disease (MRD)

The second major goal of this study was to detect residual circulating tumor DNA molecules at the first follow-up 6-12 weeks after treatment finished, and to investigate the prognostic value of residual ctDNA molecules. We refer to the significant observation of residual ctDNA molecules after treatment as 'molecular residual disease' (MRD) from here on. We included the 16 patients in the MRD analysis for which we successfully obtained plasma for T5 (post-treatment) and which showed a significant ctDNA level for at least one earlier time point (T1-T4). We compared the rate of observed MRD in the group of patients suffering from a relapse (patients 3, 5, 9, 10, 11, 13, 14, 21) with rate in relapse-free

patients (patients 4, 7, 15 – 20). We detected MRD in 2 out of 8 patients (25%) suffering from a relapse (Figure 9), while none of the 8 relapse-free patients showed MRD.

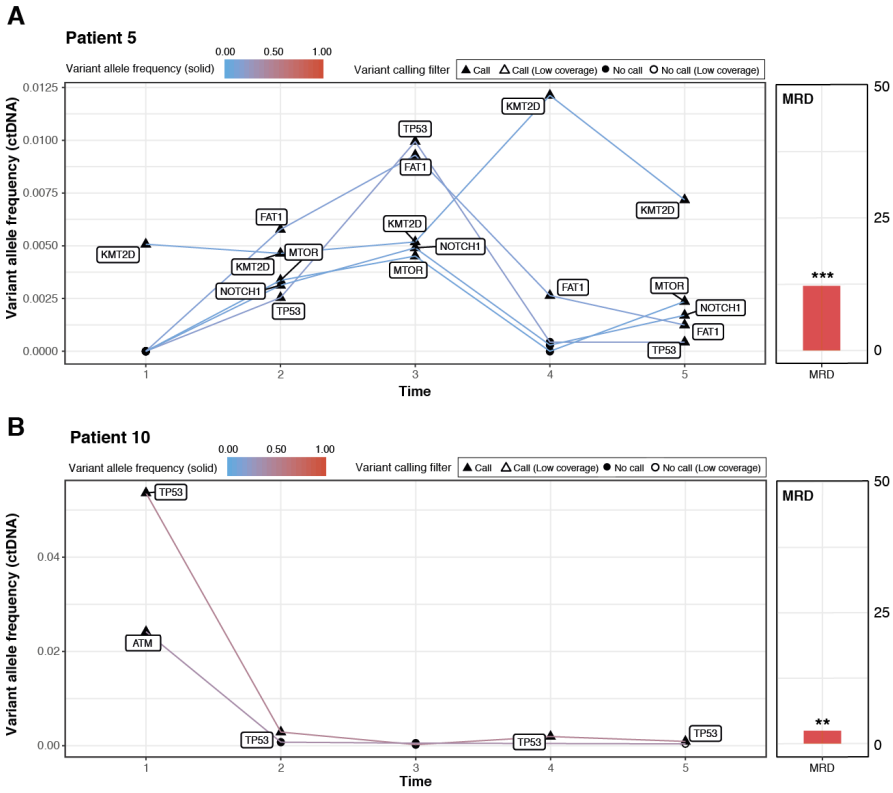


Figure 9 | Longitudinal profiles of ctDNA levels in MRD patients. (A) Patient 5: Longitudinal profile of ctDNA levels for driver mutations (left) and the MRD value at the first follow-up (T5) after treatment. Lines connect identical mutations at different time points and boxes show gene names. The color of the lines represents the VAF in the solid tumor and the shapes of the dots show the variant calling result and significance. (B) Patient 10: Longitudinal profile of ctDNA levels for driver mutations (left) and the MRD value at the first follow-up (T5) after treatment. See (A) for plot description (** for p-value < 0.01 and *** for p-value < 0.001).

Both MRD positive patients had a significant number of tumor fragments with MRD scores of 12.16 (p-value < 10^{-12}) and 2.44 (p-value < 0.004), respectively (see Figure 9). Both patients presented with either a local or distant relapse, within 101 days for patients 5 and 833 days for patients 10, respectively. Further analysis of the residual ctDNA fragments

detected in the two MRD-positive patients revealed that patient 5 still presented ctDNA fragments from five mutated driver genes with a global tumor allele frequency of 0.27 % (Figure 9**Error! Reference source not found.**A). Patient 10 carried only a mutation in *TP53* with a smaller tumor allele frequency (0.09 %), while the mutation in ATM became undetectable (Figure 9B). Interestingly, there is enrichment of significant MRD scores for patients suffering from a relapse in treatment time points 4 and 5 (Supp. Figure 3).

HPV DNA in the plasma as a marker for therapy monitoring and prognosis in advanced HNSCC

HPV-positivity was confirmed on the solid tumor biopsy by p16ink activation/inactivation in 5 patients out of our cohort of 20, and 3 out of 17 after excluding patients 2, 8 and 23. To evaluate the diagnostic potential of liquid biopsy we screened the complete cohort for the existence of cvDNA in the plasma. We could confirm HPV-positive result in 100 % of remaining patients after low-quality sample exclusion (patients 4, 14 and 15) plus one additional patient (patient 5). The number of viral fragments in patients that had a positive p16ink test was very high, showing 2,229 to 11,671 reads per kilobase per million (RPKM) at T1 (see Figure 10), while the additional patient (5) with negative p16ink showed only 352 RPKM in T1. We next assessed the cvDNA kinetics of the three patients with positive p16ink and cvDNA test (4, 14 and 15). The quantification of HPV virus fragments – normalized as RPKM values – showed a steady decrease of the viral load throughout the therapy, mirroring the decrease of ctDNA levels in the respective samples. All three patients showed a significant correlation between the longitudinal change of the normalized cvDNA counts and the administered treatment over time (p-val < 0.05 with Spearman correlation test). Moreover, the longitudinal decrease in normalized cvDNA counts (RPKM) correlated with the decrease of allele fraction of driver mutations in the respective patients (p-value < 0.05 in patients 4 and 14, p-value = 0.0728 in patient 15, with Pearson correlation test, Figure 10).

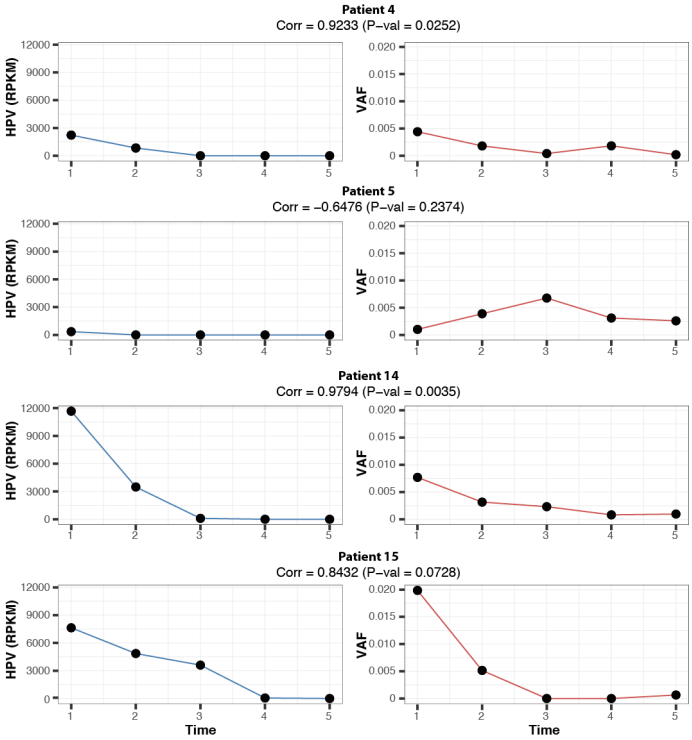


Figure 10 | Longitudinal profiles of cvDNA levels in plasma. Left: normalized counts of HPV fragments observed in cfDNA at five time points for patients with at least one read mapping to HPV. Counts are normalized using the RPKM formula frequently applied for RNA-seq analysis. Right: VAF of driver mutations in cfDNA for the respective patients. Pearson correlation values between VAF and HPV (RPKM) are shown under patient label.

Finally, we estimated the number of HPV copies per cancer cell in patients 4, 14 and 15. Normalizing cvDNA coverage by genomic coverage we estimated 3.86, 2.95 and 11.69 HPV copies per cancer cell for patients 4, 14 and 15, respectively (see Supp. Table 2). HPV copy number estimates of a patient were highly similar between time points 1 and 2, demonstrating the robustness of the approach. Furthermore, this observation indicates that all cells of a given tumor had the same HPV copy number, suggesting that the HPV expansion in the genome predates and potentially caused tumorigenesis.

4. DISCUSSION

In this study, to the best of our knowledge, we document for the first time the kinetics of ctDNA and corresponding outcome parameters in patients with locally advanced HNSCC receiving primary chemoradiation. Unfortunately, the majority of these patients develop either a local or distal relapse within the first 2 years (Linge et al., 2016; Specenier and Vermorken, 2018). Therefore, close monitoring strategies and an early detection of recurrence is needed to initiate salvage strategies. Here, we tested the application of liquid biopsy for monitoring of treatment response during RCTX and detection of molecular residual disease post-treatment, leading to four major findings, all of which could have relevance for the future clinical use of ctDNA as a biomarker in patients with advanced HNSCC.

First, ctDNA can be seen as a surrogate marker of the disease burden, tightly correlating with the gross tumor volume (primary tumor and lymph nodes) prior to the treatment start. Correlation between ctDNA levels and tumor stage has also been reported and observed in a preclinical model of HNSCC (Muhanna et al., 2017). Second, the observed ctDNA kinetics showed a clear time and dosage dependency. This enables a closer monitoring of the dynamic changes of cfDNA as a proxy of tumor size, and hence allowing estimation of the response to the treatment. Indeed, we have shown that the decline of ctDNA levels in plasma observed in most patients corresponded with the primary success of the curative treatment intend. In the only exception, patient 5, the increasing ctDNA fraction in the plasma indicated treatment failure. This patient had a local relapse within 101 days and the patient died after 135 days. Therefore, we believe that surveilling the dynamic changes of ctDNA in the plasma might be a new way to monitor and to adjust the ongoing treatment regime. Interestingly, some patients (5, 7, 13, 16 and 18) showed a peak in the ctDNA levels in in the first cfDNA sampling after treatment (T2). Except the patient 13, all variants analyzed from these five patients showed the same tendency, representing likely and increase of the cell death due to the first round of treatment, as already suggested in BRAF mutations in metastatic melanoma (Xi et al., 2016). However, although this increase of cell death could represent a success of the prognosis outcome, some of these patients suffered of relapse. Moreover, not all individuals, including the relapse-free ones, showed this early spike, meaning that more analysis on that direction should be performed to confirm this hypothesis.

The recurrence rate in locally advanced HNSCC, treated with RCTX, is around 50 percent (Linge et al., 2016). Hence, a dynamic and non-invasive biomarker such as ctDNA is needed, that enables the stratification and or prediction of patients into a high and low risk group for recurrence based on the observed kinetic changes. In many tumor entities such as breast, lung, pancreatic, bladder or colon cancer residual ctDNA in plasma or urine has shown the association of residual disease with poorer overall survival and accelerated disease recurrence rates (Garcia-Murillas et al., 2015; Sausen et al., 2015; Chaudhuri et al., 2017; Christensen et al., 2019; Tie et al., 2019) . In our study, we observed that molecular residual disease (MRD) could be detected after treatment only in patients that had a relapse. However, not all patients suffering from relapse showed MRD, which could either be due to limited sensitivity and low number of targeted variants per individual applied here, or due to the absence of ctDNA despite later relapse. Further studies with increased numbers of monitored SNVs, higher sequencing depth or larger volumes of plasma will be necessary to better understand the sensitivity and specificity of the MRD approach.

The presence of HPV type 16 or 18 was shown to be a promising biomarker for diagnosis of HNSCC, and specifically oropharyngeal SCC (Wang et al., 2015a). In our longitudinal liquid biopsy study, we observed that circulating HPV DNA (cvDNA) detectable in the plasma of patients shows the same dynamic properties during treatment as the ctDNA representing driver genes. Moreover, we observed that cvDNA disappeared post treatment, indicating that only tumor cells harbor (multiple) copies of the virus. Furthermore, pre-therapy and during the first two time points during therapy we were able to detect thousands of unique DNA fragments of the virus in plasma. Therefore, we suggest the use of circulating HPV DNA as a highly sensitive and specific biomarker for diagnosing HNSCC, for monitoring of treatment response and for detection of MRD or relapse. Due to the high sensitivity, cvDNA can furthermore be used as a blood-based screening marker for the early detection of HNSCC (Ahn et al., 2014; Jeannot et al., 2016; Eder et al., 2019). Finally, a sustained detection of HPV following the treatment could be predictive for disease recurrence. On the basis of these results it is mandatory to initiate lager clinical trials to validate our findings.

A major limitation of liquid biopsy approaches is the small fraction of ctDNA present in the total cfDNA convolute (Wan et al., 2017), and the limited amount of cfDNA that can be extracted from a vial of blood. With an average of around 20 ng DNA per vial in cancer patients, representing

roughly 6,000 genome equivalent, the chance of detecting MRD by monitoring a single mutation is limited (Chin et al., 2019). These limitations can be overcome by targeting multiple independent mutations per individual (affecting different genes), which is basically a multiplier of the number of detectable genome equivalents. Hence, the potential of our diagnostic strategy correlates with the total number of targeted mutations per individual (Chaudhuri et al., 2017), and an increase of the number of monitored mutations per individual would significantly increase the power to detect tumor fragments at very low fraction. In this study, we could not detect MRD in all patients that suffered from a recurring disease, most probably due to low number of mutations found in panel sequencing of the biopsy of some patients. In future studies, we will therefore utilize larger gene panels (>700 genes instead of 350 genes) or whole exome sequencing for biopsy analysis to substantially increase the number of monitored mutations in the liquid biopsy.

In conclusion, we have used ctDNA and cvDNA for the first time in patients with locally advanced HNSCC to monitor their treatment response during and post RCTX. We have proven the biological relationship between ctDNA/cvDNA kinetics and the tumor's response to treatment, as well as its prognostic and predictive capability for the detection of disease recurrence. Future clinical trials should include more patients and a more sensitive approach to prove our assumptions and refine our method.

Acknowledgement

We would like to thank the "Stiftung Tumourforschung Kopf-Hals" in Wiesbaden/ Germany for supporting this project by a grant to finance the genetic analysis. Besides, K. Zwirner was supported by the intramural Fortüne / PATE Program of the Medical Faculty, Eberhard Karls University of Tübingen (Funding number: 2447-0-0). Francesc Muyas and Stephan Ossowski received funding from the European Union's H2020 research and innovation programme under grant agreement No 635290 (PanCanRisk).

5. SUPPLEMENTARY MATERIAL

Treatment regime

All participants declared their written and informed consent and received chemoradiation after primary diagnosis with a cumulative radiation dose of 70Gy in areas of the macroscopic tumor, respectively 60Gy were applied in bordering areas and 54Gy were prescribed in adjuvant lymphatic regions. Three patients were treated additionally with an integrated boost on hypoxic tumor volumes with up to 77Gy. For concomitant chemotherapy either cisplatin weekly or a combination therapy of 5-FU and MMC was applied.

Follow-up

Clinical investigations and computed tomography (CT) scans were terminated 6 respectively 12 weeks after treatment for the first follow-up and consecutively every 3-6 months. Recurrent disease or progression were diagnosed by imaging and endoscopic follow up and - if possible - by histology.

Histopathological identification of HPV

Immunohistochemistry was performed on an automated immunostainer according to manufacturer's instruction (Ventana, Tuscon, AZ, USA) and an antibody was used against the p16INK4a protein (mouse monoclonal antibody, clone E6H4®, "ready-to-use" (RTU), Roche mtm laboratories AG, Mannheim, Germany). Strong and homogenous positivity reflects a close link to HPV infection.

Collection of blood samples for the analysis of cfDNA

We collected peripheral blood samples by Streck (Streck, La Vista, Nebraska) and EDTA tubes (Sarstedt, Nümbrecht, Germany) to analyse the ctDNA. Plasma separation was performed within 2 hours of the blood draw. All samples were store at either -20°C or -80°C immediately after plasma separation. Cell free DNA was isolated from 3-5 ml of Plasma with the QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany) according to the manufacturer´s protocol, with the exception that the samples were eluted in 75-150µl of DNAse, RNAse and proteinase free water (AccuGENE, Lonza). The DNA concentration was quantified by Qubit

(Thermo Fischer Scientific, Waltham, USA) fluorescence method following the manufacturer's instructions and normalized by plasma volume.

Targeted panel sequencing of tumour biopsies

Formalin-fixed paraffin-embedded (FFPE) tumour tissues were provided by the pathology department. EDTA blood samples were collected as normal tissue controls. Library preparation and in solution capture of exonic regions were performed using the Agilent HaloplexHS technology (Agilent, Santa Clara, CA). Samples were paired-end sequenced using the HiSeq2500 instrument (Illumina, San Diego, CA). An in-house developed pipeline, called "megSAP" was used for data analysis (*https://github.com/imgag/megSAP*, vers. 0.1-484-g9ad29f4 and 0.1-614-g21d6cfe). In brief, sequencing reads were aligned to the human genome reference sequence (GRCh37) using bwa-mem (vers. 0.7.15) (Li, 2013). Somatic mutations were called using Strelka2 (vers. 2.7.1) and annotated with SNPeff/SnpSift (vers. 4.3i) (Cingolani et al., 2012; Kim et al., 2018). For validity and clinical relevance, an allele fraction of ≥ 5% (i.e. ≥ 10% affected tumor cell fraction) was required for reported mutations. All variants were visually validated with the Integrative Genomics Viewer (version 2.3.97, http:// http://software.broadinstitute.org/software/igv/). Quality control (QC) parameters were collected during all analysis steps (Schroeder et al., 2017). For further interpretation all somatic variants were uploaded to the Cancer Genome Interpreter (CGI) (Tamborero et al., 2018). Somatic nucleotide variants were annotated as driver mutation based on the CGI classifications TIER1, TIER2 or predicted driver mutation.

Targeted cell-free DNA sequencing panel design

We aimed at designing a panel for targeted sequencing of plasma cell-free DNA, which is as small as possible while at the same time covering all relevant patient specific somatic mutations of the whole cohort. Therefore, we limited the panel to the somatic mutations identified as driver mutation by CGI, based on the sequencing results of the solid tumor biopsies. Additionally, we included the sequence of the E7 onco-protein of the HPV16 and HPV18 strains. The panel was designed using the Agilent SureDesign software (*https://earray.chem.agilent.com/suredesign/*) with standard tiling density, most stringent masking and max performance. This resulted in a target size of 26,926kb.

Deep sequencing of plasma samples

The library preparation was done using the Agilent SureSelect^{XT-HS} protocol following the manufacturer's instructions. To ensure uniformity 15 ng of cfDNA for each patient were used. Adapters used in SureSelect^{XT-HS} contain unique molecular identifiers (UMIs), which tag each unique DNA fragment with a unique 8bp barcode prior to the first PCR amplification. These UMI barcodes can subsequently be used to identify any PCR copy of an original DNA fragment. For the pre-hybridization step the samples were amplified with 10 cycles. The entire product was used for the Fast Hybridization protocol (60cycles, taking ~1.5h). Capture was started immediately after the final hybridization cycle and proceeded for 30 minutes at 26°C. For the post-capture procedure, the libraries were amplified with 12 cycles. The post-capture washes were performed at 71°C. The libraries were cleaned with 1x Ampure XP beads, quantified and pooled together. Subsequently, 1.8nM of the library pool of 32 samples was sequenced on NextSeq500 (Illumina, CA).

Identification of unique DNA fragments by UMI-based de-duplication

Sequencing reads originating from the same amplified DNA fragment were identified and grouped by their unique molecular identifier (UMI) of 8 bps and the coordinate of the mapped read on the reference genome. Next, sequencing and PCR errors were identified and removed using multiple alignment and base-to-base comparison between all PCR duplicates. Third, all PCR duplicates with the same UMI were collapsed to one sequence ('de-duplicated'), in which discordant bases between duplicates were masked or marked as low quality. Finally, resulting de-duplicated reads were processed with the *BamClipOverlap* tool (*https://github.com/imgag/ngs-bits*) to soft-clip paired-end reads that overlap.

Ultra-low frequency somatic variant calling using UMI-corrected deep-sequencing data

In order to identify mutations at ultra-low variant allele frequencies (VAF >= 0.1%) in the cfDNA read data we utilize the information from the UMI-based de-duplication and error correction procedure (see above) and developed a statistical model for error probabilities based on the beta-binomial distribution (method paper in preparation). Only reads with mapping quality greater than 30 and nucleotides with UMI-corrected base quality greater than 20 were considered in the variant calling step.

Additionally, only variant sites with at least 2 alternative reads that passed the beta-binomial error model were considered as PASS calls. The lowest AF identified by our model had a VAF of 0.043% (~1 mutated DNA per 2,325 reference-like DNA fragments in plasma). Variant allele frequencies of each mutation per patient and time point were recorded for longitudinal analysis of ctDNA dynamics in plasma in relation to treatment dosage.

Correlation of treatment time point with VAF

In order to obtain the correlation between VAFs of driver mutations in plasma with treatment time points we developed the following longitudinal analysis procedure. First, to reduce the effect of VAF variance and to allow to visualize extremely low VAFs we log-transformed VAF values (log_{10}). As many mutations disappeared in later treatment stages, and VAF = 0 cannot be log-transformed, we summed to a small value to all VAFs, representing the expected background error noise (10^{-5}). Next, as cfDNA concentration and infection status (*Infection* or *Not Infection*) are known confounders that likely affect the concentrations of cfDNA fragments in plasma (Zwirner et al., 2018b), we subtracted ('removed') the effect of these confounders using a linear model. In detail, we applied a linear regression to remove the effect of the variables concentration and infection-status (confounder model):

$$Log_{10}(VAF) \sim [cfDNA] + Infection_{status} + \varepsilon$$

The residuals (ε) of this model represent the variability of the ctDNA fraction in the plasma, which cannot be explained by the two confounders (concentration and infection-status). Once the effects of DNA concentration and infection were subtracted, we considered the residuals (ε) as proxy for the tumor size. Hence, residuals (ε) were next correlated with the variable treatment time-point (T1 to T5) using the Spearman correlation test in order to test if the tumor responded to the RCTX treatment. We defined a positive treatment response if patients showed a decrease of VAF over time and a negative Spearman correlation. The treatment response test was considered significant for $p < 0.05$. A positive Spearman correlation indicates no response to treatment or progress under treatment (only patient 5).

Molecular residual disease (MRD)

The *MRD* value (Molecular Residual Disease, also termed Minimal Residual Disease in other publications) combines information of all monitored mutations per patient and time-point into one measure of presence or absence of residual tumor DNA. We use Fisher's combined probability test to integrate the significance values returned by the variant calling method for each monitored mutation at a time-point *x*. The resulting MRD p-value represents the significance of observing residual tumor DNA in plasma, with p-values < 0.05 indicating that more ctDNA was detected as expected based on the error distribution. Afterwards, the resulting p-value is log-transformed to a 1-100 scale (–log10) forming the MRD score. Higher MRD scores indicate a higher likelihood that a patient has residual tumor DNA in the plasma after treatment, and hence a higher likelihood that the patient has residual tumor cells in the body. Detection of significant MRD score at a later follow-up time-points can also indicate a local or distant relapse. Finally, we considered as significant MRD scores > 1.3, which is equivalent to p-value < 0.05.

Correlation of DNA fragment sizes in plasma and variant allele frequency of driver mutations

Fragment sizes were calculated for all patients and times directly from the insert size of paired reads in bam files. To calculate the proportions of molecules with fragment sizes between 90 and 150 bps, we only took into account fragments with a size less than or equal to 200 bps. Finally, the correlation between the proportion of 90-150 bps fragments to all fragments <= 200 and the $\log_{10}$ of VAF was computed with a *Pearson* correlation test. As ctDNA fragments have been reported to be shorter than other cfDNA fragments (Mouliere et al., 2018) we expect a positive correlation between VAF of driver mutations and the proportion of short fragments.

RPKM for HPV

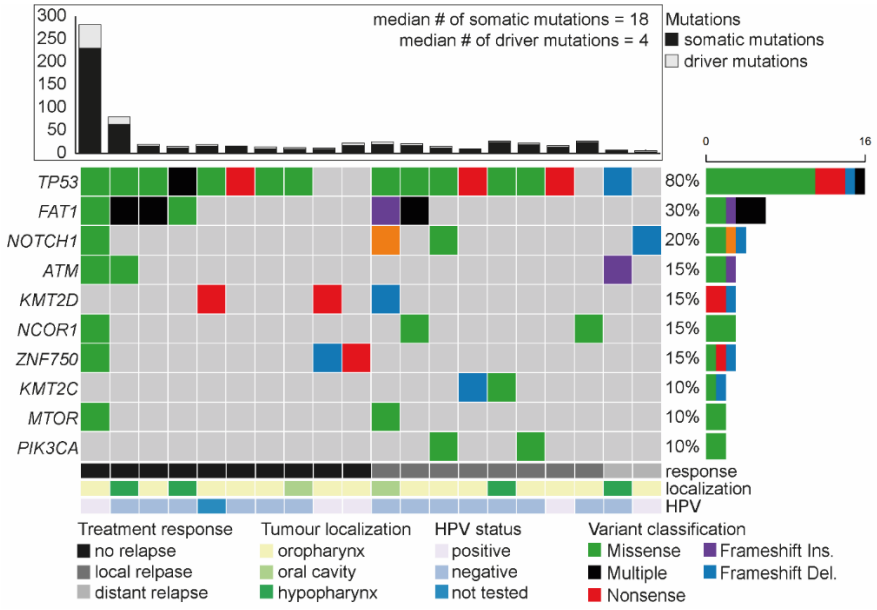
We included the HPV virus strains 16 and 18 in the oligo enrichment panel used for liquid biopsy. After UMI barcode de-duplication, we counted how many reads (i.e. unique DNA fragments) mapped to each of the regions corresponding to the virus strains. In order to normalize these counts, we calculated RPKM values (Reads Per Kilobase per Million), a method well-known from RNA-seq analysis that normalizes by region length and library sequencing depth (Mortazavi et al., 2008). Hence, for RPKM normalization

of viral read counts we used the total on-target read count (viral + genomic regions) and the size of the virus-specific target regions. The correlation between HPV RPKM values and treatment time points was calculated with a Spearman correlation test. Additionally, in order to check the correlation between HPV and VAF we used a Pearson correlation test. Afterwards, the calculation of HPV copies per cell was calculated with the following formula:

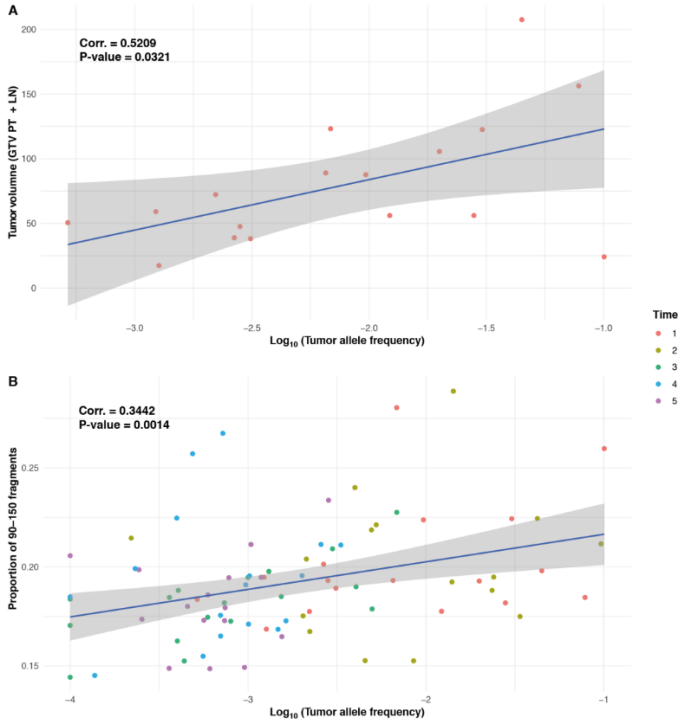
$$\textbf{\textit{HPV copies /cell}} = \frac{RPKM\ HPV}{RPKM\ genome}\ x\ \frac{1}{2*VAF}$$

Where RPKM HPV is the value previously computed, and RPKM per genome is the RPKM value calculated for all targeted regions (except virus regions). ½ x VAF is used as a correction factor that estimates the fraction tumor DNA in the plasma.

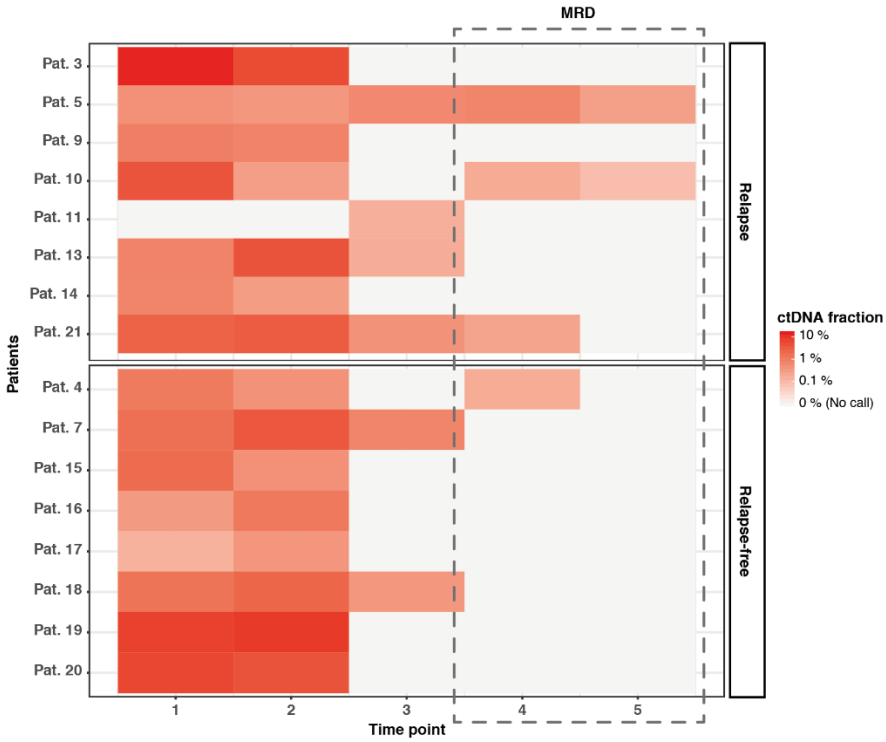
6. SUPPLEMENTARY FIGURES AND TABLES



Supp. Figure 1 | Oncoplot of 10 most frequently mutated genes with driver mutations. The upper part of the panel shows the total number of mutations for each patient as a staged bar chart, where black indicates all somatic alterations and grey highlights the identified driver alterations by the Cancer Genome Interpreter. The oncoplot itself depicts the top most frequently mutated genes based on the selection of all driver mutations, solely. While each row represents the different kinds of driver mutations (color coded) and their alteration frequency as well as the total number of mutations. Each line represents one patient summarizing the identified driver mutations. Additionally, the treatment response is shown in a grey scale, as well as the localization of the primary tumor (yellow-green scale) and the HPV status (purple-blue scale) determined by the pathology.



Supp. Figure 2 | Correlation of ctDNA fraction in the plasma with (A) the fraction of DNA fragments in the size range 90-150 bps, and (B) the tumor volume before treatment. (A) Correlation of the total tumor volume (GTV PT + LN) with variant allele frequencies of driver mutations in plasma before treatment (T1). Variant allele frequencies in plasma correlated positively (p-value < 0.05 with Pearson correlation test) with the total size of the tumor (GTV PT + LN). However, when the two measures of volume were taken separately, none of them showed a significant correlation (p-value > 0.05). (B) The proportion of fragments with size in the range of 90-150 bps correlated significantly and positively with the tumor allele frequencies in plasma of the variants monitored in this study (p-value < 0.01 with Pearson correlation test). This fact supports that, independent of the number of variants we are looking at, high ctDNA fractions in the bloodstream have an impact in the fragment size distribution of a patient.



Supp. Figure 3 | Heatmap with tumor allele frequencies of variants detected in this study (ctDNA fractions) across different patients and time points. Only patients with all time treatment points available and not excluded due to technical issues are included in this plot. There is enrichment of significant MRD scores for patients suffering from a relapse in treatment time points 4 and 5 (dashed line box).

Supp. Table 1 | Correlation analysis between VAF in ctDNA and treatment time points. Firstly, confounder effects were removed from data set, and afterwards, VAF-time correlation was obtained with a Spearman correlation test. Significance was assumed when p-value < 0.05 after FDR correction.

Patient	Num. Mutations	Time-VAF correlation	Rho (Spearman correlation)	Spearman p-val	Confounder Model	Confounder Model P-val
13	3	Significant Decrease	-0,8963	0,00001	Non-significant	0,1907
18	3	Significant Decrease	-0,9069	0,00005	Significant	0,0189
7	6	Significant Decrease	-0,5311	0,00253	Significant	0,0002
3	2	Significant Decrease	-0,8370	0,00252	Significant	0,0441
20	2	Significant Decrease	-0,7878	0,00681	Significant	0,0971
14	2	Significant Decrease	-0,8486	0,00772	Non-significant	0,7704
16	2	Significant Decrease	-0,7139	0,02039	Significant	0,0253
21	1	Not significant change	-0,7000	0,23333	Non-significant	0,2266
10	2	Not significant change	-0,4238	0,25566	Significant	0,087
5	5	Not significant change	0,2563	0,27534	Significant	0,0005
17	1	Not significant change	-0,8000	0,33333	Non-significant	0,8246
15	1	Not significant change	-0,4000	0,75000	Non-significant	0,7596
9	4	Not significant change	-0,0552	0,81713	Significant	0
4	3	Not significant change	-0,0328	0,90773	Non-significant	0,1697
19	1	Not significant change	-0,2000	0,91667	Non-significant	0,4281
22	1	Not significant change	0,2000	0,91667	Non-significant	0,446
2	1	Excluded	NA	NA	Excluded	NA
8	0	Excluded	NA	NA	Excluded	NA
11	1	Excluded	NA	NA	Excluded	NA
23	0	Excluded	NA	NA	Excluded	NA

Supp. Table 2 | HPV-positive individuals' information. Correlations of HPV levels with treatment time points and ctDNA levels, as well as HPV copy number estimates in t1 and t2.

Patient	Rho HPV-Time*[1]	HPV-Time p-val*[1]	Corr HPV-VAF*[2]	HPV-VAF p-val*[2]	HPV Copies/Cancer cell T1	HPV Copies/Cancer cell T2	Solid HPV test
4	-0,8944	0,0405	0,9233	0,0252	3,8583	3,5312	Positive
5	-0,7071	0,1817	-0,6476	0,2374	2,6494	0	Negative
14	-0,9747	0,0048	0,9794	0,0035	11,689	8,4696	Positive
15	-1	0,0167	0,8432	0,0728	2,9525	7,2163	Positive

*[1] = Spearman correlatior

*[2] = Pearson correlatior

APPENDIX

USE OF UNIQUE MOLECULAR IDENTIFIERS TO DETECT ULTRA-RARE SOMATIC VARIANTS IN CELL-FREE DNA

This section of the thesis describes in detail the computational methodology and benchmarking analysis of the tools used in the project *Dynamics of circulating cell-free tumor DNA in HNSCC patients receiving radiochemotherapy correlates with treatment response* (previous section).

The goal of these methods is to detect ultra-rare somatic variants in cfDNA with the use of unique molecular identifiers (UMIs) or barcodes. UMIs are defined as random sequences attached to the original DNA fragment, which afterwards will be present in all duplicates derived from each individual fragment. The use of UMIs has been considered in two main applications:

- Barcode-based consensus sequence correction. Use of barcodes to reduce background noise (PCR and sequencing errors).
- Variant calling and minimal residual disease (MRD) detection.

Both applications will be explained within this appendix and will be part of a method paper (in preparation). Moreover, this method is part of an un-submitted patent application. Currently it has not been decided if the patent will be submitted to the patent office.

1. METHODS

1.1. Processing reads

First of all, the sequences of unique molecular identifiers (UMIs) or barcodes are attached and stored at the end of the read name line of the *fastq* files of read 1 and 2. This step is mandatory to link each read with its own UMI, which afterwards will be used for clustering duplicate reads.

Next, in order to trim possible adapters at the end of the reads, we processed reads with the tool *SeqPurge* (Sturm et al., 2016). This tool is able to distinguish and remove the adapters when, for instance, reads are longer than the insert size.

Following the previous steps, reads were aligned against the human reference genome using *BWA-mem* (default parameters). Afterwards, a QC (quality control) in-house script was applied for removing low quality read pairs. This step removed reads if:

- Reads are not paired or not properly mapped as pairs.
- Mapping quality was lower than 30.
- >= 3 mismatches or >= 1 gap per read alignment.

The filtered bam file was then submitted to the barcode correction tool (deduplication step), where the sequences of duplicate DNA fragments are merged into a single consensus sequence.

1.2. Barcode correction

The goal of this step is to identify PCR duplicates of the same original DNA fragment by their common UMI, and to use the information in duplicate reads to detect and correct errors that occurred during sequencing or PCR amplification steps. This de-duplication step collapses duplicates into error-free consensus fragments.

As previously explained, UMI sequences are available in the bam file within the read name. Therefore, DNA fragments (both paired reads) are grouped by UMI, alignment start, alignment start of the pair, and fragment length (total insert size) in order to identify PCR duplicates originating from the same DNA fragment. These duplicates are grouped and subsequently used to form a consensus sequence, i.e. to correct a nucleotide we compare the sequenced base of each PCR duplicate aligning to the same genomic position. The consensus base at one specific position *x* is the most common base on that site observed across all PCR duplicates of the group (with base quality (BQ) > 20). At that stage, the base quality of the consensus base is chosen based on the maximum base quality observed in the position *x* of all PCR duplicates used to generate the consensus sequence (only base qualities showing the consensus base are considered). Nevertheless, if there is a disagreement between the compared bases or if more than one nucleotide is equally represented, the consensus base is chosen as the most frequent one or as the one with highest BQ in the read data, respectively. In all cases, if the fraction of duplicates supporting the consensus base is less than 75% or there are >= 3 reads showing any other allele, the BQ of the consensus base is set to 0.

Finally, the number of duplicates from which each consensus is created is saved in one extra column in the bam format under the identifier *DP:i:'Number of* duplicates'.

1.3. Error rate calculation

The error rate is calculated dividing the sum of alternative bases over the total number of bases sequenced. Sites with allele balance greater than 10% (reads supporting alternative allele / reads covering the site) were considered true germline or somatic variants and were ignored from the error rate analysis. We computed different error rates for each correction stage and only bases that passed the BQ filter were considered in the calculation. To get the error rate depending on the number of duplicates, we split the bam file based on the DP:i.

1.4. Targeted variant calling

The variant calling is performed using the pileup format generated from BAM files by *Samtools*. To ensure high precision and sensitivity we consider all consensus-reads obtained in the deduplication step, although we take into account the number of duplicates each consensus fragment was created from (i.e. using the information from DP:i).

Using major and minor allele counts per position we model the error rate distribution with a beta-binomial distribution in *n* collapsed and independent samples (n = number of samples analyzed in the same run, or alternatively samples from an in-house database). Errors are distributed as a Binomial distribution with parameter P (error rate), which is a random variable that follows a Beta distribution with parameters α and β.

$$\textit{Error counts} \sim Bin(\textit{Coverage}, \textit{error rate})$$
$$\textit{Error rate} \sim Beta(\alpha, \beta)$$

Our analysis strategy was designed to maximize recall and precision by considering all information obtained during duplicate consensus generation, while also taking into account potential chemical processes leading to a nucleotide-specific error (e.g. oxidative stress or de-amination of methylated cytosines). Hence, we considered two types of error models: a) nucleotide-change specific error signature, and b) duplicate-correction-depth specific error rates.

a. **Nucleotide-change specific error signature**: As the error rate varies across the six different nucleotide-change types, we model the error rate distribution (maximum likelihood estimation – MLE) per each nucleotide change in a non-strand-specific manor (T/A > G/C, T/A > A/T, T/A > C/G, G/C > T/A, G/C > A/T and G/C >C/G) resulting in six different error models.

b. **Duplicate-correction-depth**: The error rate varies significantly depending on the number of duplicates used to create consensus reads. For instance, a consensus generated from more duplicates should have less errors, as more duplicate reads were compared to each other, increasing the chance of distinguishing errors. Considering this criterion, error rates were calculated separately for 4 groups of DP:i (DP = 1x, DP = 2x, DP = 3x and DP >= 4x), which at the same time, every DP group had six nucleotide-change specific error models, resulting in 24 error models in total.

Therefore, for any targeted position in the analysis, we split consensus reads in the 4 different DP:i groups, where each one has different error rates per nucleotide change. Then, using the nucleotide-DP-specific beta-binomial distribution, we obtain a p-value for each site and DP, resulting in 4 p-values for every position (one for every DP:i group). Afterwards, these p-values are combined (Fisher's combined probability test) into a single one (one p-value per targeted site) and normalized (with FDR) to identify likely mutated sites, i.e. sites with an alternative count significantly outside of the respective error rate distribution. Hence, the total number of targeted (genotyped) variant positions affects the p-value correction. Hence, a higher number of monitored variants will slightly decrease sensitivity per variant site. However, as discussed below, a higher number of variants at the same time increases sensitivity of the test based on all variants combined.

Finally, once we got potential calls, other filters were applied:

- **Minimum distance between variants**. Variants clustered in small windows are prone to be false positive calls and can be removed.
- **Strand bias filter**. Variants that are not equally represented on forward and reverse strand are related with false positive calls. To remove these variants, we perform a Fisher exact test comparing reference and alternative counts per strand.
- **Multi-allelic sites**. Positions with more than 1 alternative allele are filtered.

1.5. Minimal Residual Disease (MRD) score

The Minimal Residual Disease (MRD) score is designed for monitoring variants already found in a biopsy of a solid tumor. It is appropriate when very low allele frequency variants (already known) need to be distinguished in cfDNA extracted from liquid biopsies, where the low amount of extracted DNA results in a limited number of distinct (unique) DNA fragments in the sample (only few thousands haploid DNA fragments can be captured per vial of blood).

MRD scores summarize the complete information obtained for all targeted variants per sample/individual into one significance value. This value combines and collapses the significance of each variant analyzed separately (computed as explained in *Targeted variant calling* section), using Fisher's combined probability test. Afterwards, the resulting p-value is transformed using $-\log_{10}$ and floored at maximum value 100. Therefore, MRD values, which range from 0 to 100, represent the significance of ctDNA fragments presence in the analyzed sample, giving as maximum significance the value 100. Finally, we considered as significant sample when MRD > 1.3, which is equivalent to p-value < 0.05.

1.6. Variant calling and MRD performance

Firstly, variant calling limitations were obtained with the theoretical beta-binomial distributions previously computed for 17 cfDNA samples from HNSCC patients. To check the importance of the depth of coverage, we investigated different depths of coverage or haploid genome equivalents (hGEs) (1000, 5000 and 10000).

Secondly, false discovery rate (FDR) of our variant calling was obtained using the real cfDNA data. To avoid the difficult task of background noise simulation, we first randomly selected real reads overlapping with a total of 2,000 random sites where no germline or somatic variants were found, but where real background noise and errors were already in the sequences. Moreover, in order to achieve high hGEs and coverages, we collapsed the bam files of all samples (after deduplication) and extracted reads in different depths of coverage (1000, 5000 and 10000). Then, variant calling with default parameters was run in these 2,000 random sites and FDR was calculated.

Finally, to understand the potential of the MRD calculation we first simulated mutations at different ctDNA fractions. As described in the

literature (Newman et al., 2016b), the number of circulating tumor DNA (ctDNA) fragments in plasma follows a Poisson distribution with $\lambda = n \times d$; where λ represents the expected number of ctDNA fragments, n the number of haplotype genomes and d the fraction of ctDNA molecules in the totality of cfDNA fragments analyzed. Therefore, following this model, we simulated 2,000 variants at different tumor fragment fractions (0.1, 0.05, 0.025, 0.01, 0.0075 and 0.005 %) under different conditions of hGEs (2500, 5000, 7500 and 10000 hGEs, ranging from ~ 10-35 ng of DNA).

In order to check the importance of the amount of variants monitored in different ctDNA fractions, we randomly selected different numbers (N) of variants (from 1 to 150 variants) from the 2,000 simulated mutations and checked the proportion of times that *ctDNA* fragments were significantly detected (MRD > 1.3) (bootstrapping each 'N variant selection and MRD calculation' 1,000 times) for every N at every different ctDNA fraction and hGE content. Additionally, we calculated MRD values assuming three different levels of background error rates (0.01 %, 0.005 % and 0.001 %).

1.7. Samples used

The application, optimization and benchmarking of this method was performed in the samples from 20 HNSCC patients described in chapter 2. Specially, only samples from the last two time points from individuals that not presented relapse were considered in this analysis in order to avoid ctDNA fragment contamination, resulting in 17 samples.

2. RESULTS

2.1. Deduplication and error correction

As described in the Method section, we used UMI and mapping information to collapse and group the reads in family duplicates. The comparison of these fragments allows us to detect errors that occurred during sequencing or PCR amplification, and hence, it allows us to remove them from the background noise amplification, and hence, it allows us to remove them from the background noise (Figure 11).

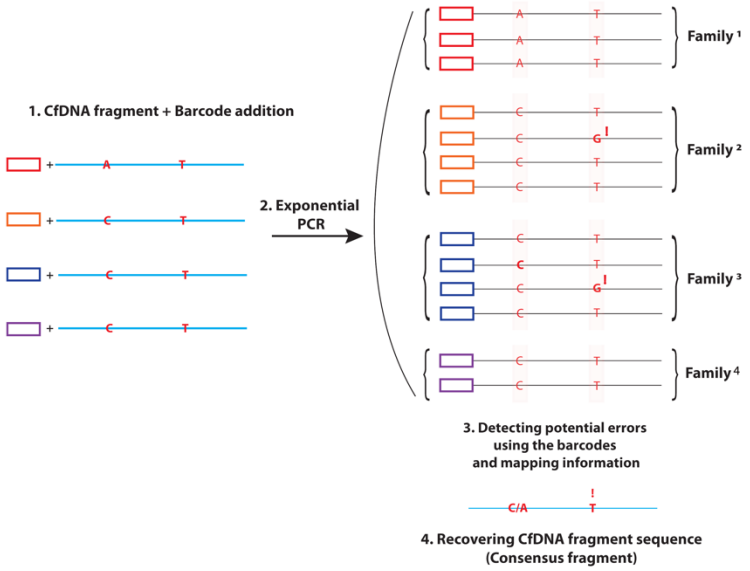


Figure 11 | Barcode correction strategy. (1) Barcodes are attached to each original fragment; (2) exponential PCR creates duplicates, which can be grouped using barcode and mapping information; (3) Comparison of duplicates intra-family-wise permits to detect PCR and sequencing errors; (4) finally, 'error-free' consensus fragments are re-build using duplicate information.

The analysis of 17 cfDNA samples revealed that error rates (background noise) decreased substantially when using barcode correction strategy (Figure 12). Collapsed fragments that originated from more duplicates showed lower error rates, demonstrating that ultra-deep sequencing analysis and therefore, a high saturation of duplicates is useful to reduce the background noise when unique molecular identifiers are used. Nevertheless, some errors remained even after strong deduplication (4x, Figure 12B), showing C to A and its complementary G to T the greatest

errors rates. These errors were most probably caused due to oxidation of DNA during library preparation (before barcode ligation), as already described in Costello et al (Costello et al., 2013). We further observed that the improvements are minor when going from 3 to 4 duplicates, indicating that we likely reach saturation for error correction with less than 10 duplicates. Unfortunately, our data did not provide enough DNA fragments with >=5 duplicates to test this hypothesis.

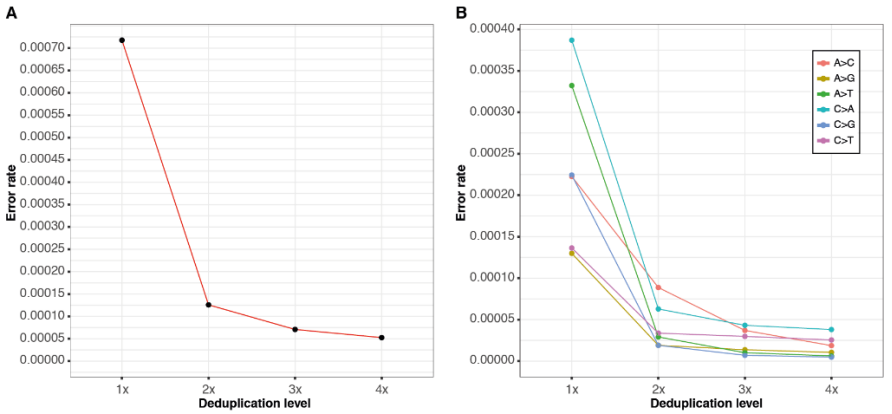


Figure 12 | Error rates based on deduplication level. (A) Global error rates; (B) error rates split by nucleotide change.

2.2. Variant calling

Once we reduced the background noise from data, it was important to take into account the remaining error rates in order to distinguish them from real somatic variants at very low allele frequency. Using beta binomial distributions, we modeled the error rates per each one of nucleotide change and deduplication levels (consensus generated from 1x, 2x, 3x and >=4x duplicates) and use this information to maximize precision and sensitivity. Therefore, the detection limit highly depends on the error rate of the nucleotide change and the deduplication level.

Considering the error rates of the cfDNA samples previously described, we observed that detection limits depended also on the number of fragments covering a specific position, reaching lower detection limits when higher numbers of fragments were analyzed (Figure 13). We noticed that allele frequencies as low as 1-2 in 10,000 fragments were detectable (considering a single variant and p-value < 0.01) when looking at high

fragment depths (5,000 and 10,000) in high deduplicated levels (3x and 4x), and lower values were achieved when only one duplicate was used to create the consensus fragment (1x). As expected, nucleotide changes like C to A and C to T (and their complementaries) had higher detection limits in most of deduplicated levels.

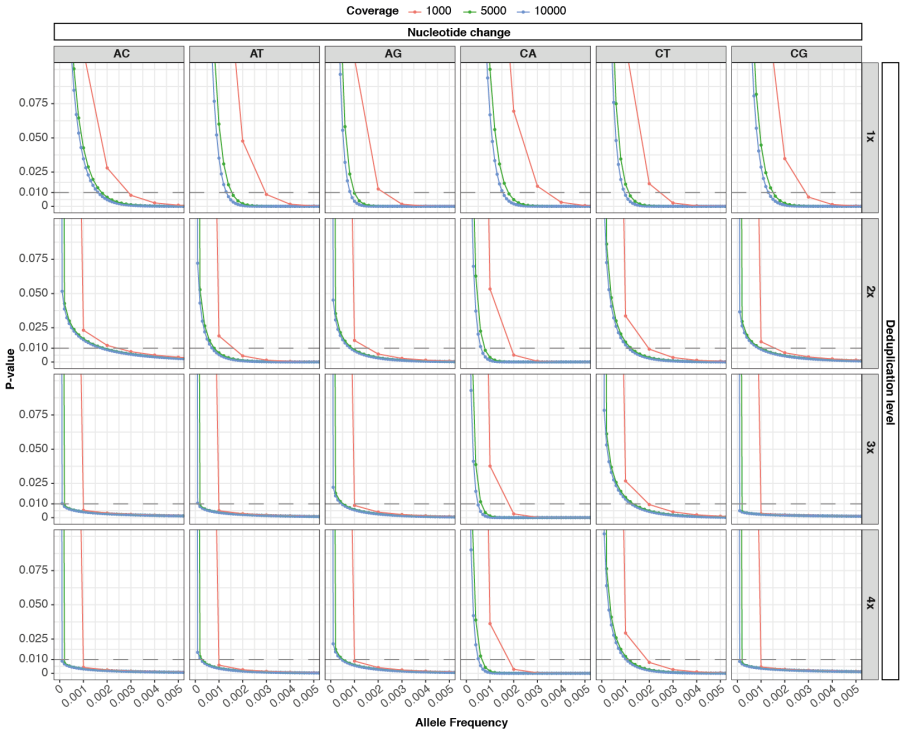


Figure 13 | Detection limits of the variant calling. Detection limits are split based on nucleotide change, deduplication level and depths of coverage simulated (colors of the lines). The detection limit is calculated in a single variant (no FDR correction) with α = 0.01 (grey dashed line).

Additionally, random selection of 2,000 non-variant sites (using collapsed samples in order to achieve high depth of coverage) demonstrated that false discovery rates remained as low as 1.7 %, showing that ultra-rare variants could be detected with a high precision.

2.3. Minimal residual disease

One of the main limitations of cfDNA-based tests is that only few ng of cfDNA per ml of plasma are found per patient, which only represents few

thousands of haploid genome equivalents (hGEs). Hence, detection of a tumor specific mutation present in a frequency below 1 in some thousand fragments is unreliable (Wan et al., 2017). Our MRD strategy tries compensate for this limitation by analyzing many targeted variants per individual in combination. Increasing the number of monitored variants per patient is thought to increase the chances of detecting tumor fragments at very low fraction. Here, we prove this assumption within our simulation based on real HNSCC patients.

Using a Poisson distribution, we simulated variants at different tumor fragment fractions (from 1 % to 0.005 %) under different conditions of hGEs (2500, 5000, 7500 and 10000 hGEs, ranging from ~ 10-35 ng of DNA) in 2000 sites. In order to investigate the effect of the background noise, we called the variants under three different levels of background error rates (0.01, 0.005 and 0.001 %).

As shown in Figure 14, the number of targeted variants, the number of hGEs and the background noise level influenced the sensitivity of the MRD test to detect tumor fragments at very low ctDNA fractions.

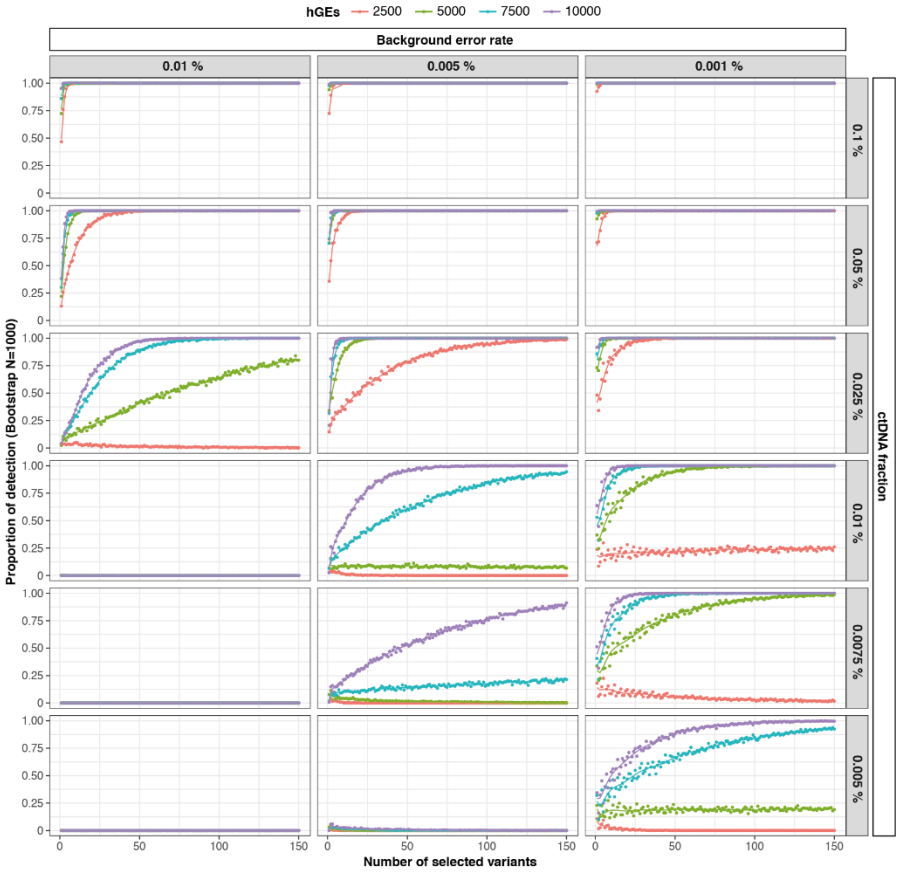


Figure 14 | Minimal residual disease (MRD) detection limit. Y axis represent the proportion of times (boostrap N = 1000) in which the algorithm significantly detected MRD (i.e. detected the mutations in ctDNA) under different conditions of number of targeted variants (lower x-axis), background error rates (upper x-axis), ctDNA fractions (right y-axis) and haploid genome equivalents (hGEs, indicated by curve colors).

As expected, greater error rates increased the fraction limit to detect tumor fragments in plasma, permitting only to detect variants as rare as 0.025 % when a mean error rate of 0.01 % was considered (with a minimum of 25 variants to detect at least 50 % of cases). However, when lower error rates were assumed (0.005 and 0.001 %), the MRD test was able to detect in a high proportion of cases tumor fragments at a frequency as low as 0.005 %. Of course, if very low tumor fractions need to be detected, it requires to increase the number targeted variants or the amount of DNA (hGEs). Indeed, increasing both the number of monitored

variants and the number of samples blood vials would clearly be the ideal solution to increase the sensitivity (Figure 14).

In summary, the MRD method shows that despite the limitations of DNA quantities in liquid biopsies, the increase of variants helps to solve the low input DNA limitation. However, both background noise and input DNA amount influence significantly the capability to detect ultra-low ctDNA fractions, representing a challenge for new barcoding protocols.

CHAPTER 3

DETECTING MOSAIC MUTATIONS IN HEALTHY TISSUES OF THE HUMAN GENOME

The way and rate that mosaic mutations are acquired during human development and life is not fully understood. From the first division of the zygote until death, cells accumulate mutations that in some cases might lead to the development of disorders (Prochazkova et al., 2009; Ruark et al., 2013; Campbell et al., 2015; Halvorsen et al., 2016). However, although the effect of somatic mutations in cancer has been deeply studied, there are very few studies that characterize mosaic mutations acquired during human embryogenesis or adult life in healthy individuals. Studies of mosaic mutations acquired early during embryogenesis have only been performed in single tissues (Ju et al., 2017; Wei et al., 2018a), which means these studies had a high chance of missing mosaic mutations not present in the analysed tissue. For this reason, a multi-tissue view of the mosaic mutations of an individual is lacking to date.

In this chapter, we have used 10,097 RNA-seq samples from up to 49 tissues and 570 individuals from the Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013) consortium cohort to characterize the mosaic mutations acquired during human embryogenesis and adult life. The high number of mosaic mutations in coding regions detected in several normal tissues from the same donor entails novel hypothesis to be considered when searching for genetic causes of diseases, which might impact the development of new diagnostic procedures.

The selection signature analysis of somatic mutations was performed by Dr. Zapata Ortiz. The rest of the analysis of this study were performed by Francesc Muyas. I have also written the paper, with the help of my supervisor Prof. Ossowski.

Muyas F, Zapata L, Guigó R, Ossowski S. The rate and spectrum of mosaic mutations during embryogenesis revealed by RNA sequencing of 49 tissues. Genome medicine. 2020;12(1):49–14. DOI: 10.1186/s13073-020-00746-1

DISCUSSION

The fast development and continuous improvement of high-throughput sequencing technologies pushed forward the field of medical genomics, creating plenty of new applications that require high quality data and methods to obtain relevant results and novel hypotheses. Distinguishing errors from real variants is a challenge when systematic errors, background errors, germline variants or somatic variants at very low frequency are present in the same data (Li, 2014). Therefore, one of the main objectives of this thesis was to develop methods to distinguish the different type of errors from real somatic or germline variants.

For this reason, the first part of the thesis was focused on characterizing errors that, in general terms, can be divided into:

$$Errors \sim Systematic\ errors\ + PCR\ errors + Sequencing\ errors + Others$$

In the chapter 1 of this thesis, we focused on creating a method (called ABB) to detect sites in the human genome that are prone to systematic errors, leading to false calls in germline and somatic variant studies. We described several important implications of these errors in downstream analysis and showed how to reduce their negative impact on rare variant association studies for case-control cohorts.

In the next chapter, chapter 2, we analyzed the kinetics of somatic mutations in cfDNA of 20 HNSCC cancer patients during treatment. For that, we developed methods to remove background noise (PCR and sequencing errors) using unique molecular identifiers (UMIs) and created a somatic variant calling approach able to detect variants at extremely low fraction.

Finally, in chapter 3 we applied previous acquired knowledge to characterize the spectrum and rate of mosaic mutations during early embryogenesis and life from a big cohort of 49 tissues from hundreds of healthy individuals.

Impact and detection of systematic errors

Although the performance of variant callers has been optimized since DNA re-sequencing by sequencing-by-synthesis technologies was introduced,

some systematic errors remain even after strict filtering, which might bias the downstream analysis (Pfeifer, 2017).

In chapter 1, we described a new genotype callability filter (called ABB) able to detect and filter systematic errors from read alignments to the human reference genome. The analysis of allele balance bias recurrence across 987 WES samples permitted us to build a model to recognize and distinguish this type of errors. Sanger validations of randomly selected variants and quality control measures such as transition-transversion ratio (TiTv) (Freudenberg-hua et al., 2003; Pattnaik et al., 2012) confirmed that our method was able to detect false positive calls that standard filters and pipelines were not able to remove (Muyas et al., 2019a).

Around 4% of the genomic positions called as germline variants and 8% of positions called as somatic mutations by conventional methods were labeled as potential systematic errors by ABB, showing that an important fraction of final and "high-quality" variant callsets are enriched by false positive calls. The importance of this finding arises in the downstream analysis of these calls, as they can be used in clinical diagnostics, rare variant association studies and many other applications requiring high precision. The enrichment of false positive calls in this type of analysis might have important consequences such as suboptimal treatment selection, could lead to wrong diagnosis and prognosis of different genetic disorders or could lead to false positive associations of genes with diseases.

Although ABB shows some correlations with other QC measures like Fisher strand bias and repetitive or low complexity regions (Li, 2014), none of these parameters completely overlap with the set of positions flagged by ABB, making ABB a valuable addition to the QC filter setup. Moreover, we found an enrichment of systematic errors in public databases, with dbSNP being by far the database with the highest fraction of ABB low-quality sites, supporting previous observations by other groups (Musumeci, 2011). Our results also demonstrate that variant callsets created consistently by a defined and reproducible pipeline and parameter setting, such as 1000GP, ExAC/GnomAD, EVS, provide higher quality than databases, which are collections provided by many different users. As a consequence databases like GnomAD should be preferred over dbSNP for benchmarking of variant callers (Muyas et al., 2019a).

The high fractions of somatic variant calls that are likely systematic errors revealed that novel algorithms are required to remove false positive calls,

in order to achieve reliable cancer diagnostics. Considering the importance of predicted mutations for cancer diagnostics and treatment selection, filtering of FP calls is essential for the applicability of NGS in precision oncology.

A second important application described in this chapter was the utilization of our ABB tool to identify false phenotype-genotype associations resulting from systematic errors in rare variant association studies (RVAS). A high fraction of significantly associated genes (25 %) we detected for the ICGC Chronic Lymphocytic Leukemia cohort (Quesada et al., 2011; Muyas et al., 2019a) was labeled as FP by ABB, as their association could be better explained by uneven burden of systematic errors between cases and controls. Again, this demonstrates the massive impact of systematic sequencing analysis errors in genetic studies and hence, the necessity of systematic error removal before downstream analysis. However, we also hypothesize that some of these systematic SNV calling errors could be introduced by un-annotated copy number variants (CNVs) in at least a couple of candidate genes, indirectly pointing to the real cause of the genotype-phenotype association, although further analysis in that direction needs to be performed (Abyzov et al., 2013; Muyas et al., 2019a).

In summary, in chapter 1 of this thesis we have presented a novel genotype callability estimator based on allele balance bias (ABB), which can identify systematic variant calling errors not found by other measures. Moreover, ABB can improve the accuracy of germline and somatic variant sets as well as clean disease association studies in large cohorts.

Monitoring cancer patients pre-, during and post-treatment using cell-free DNA

In chapter 2 of this thesis, we performed the longitudinal analysis of ctDNA dynamics in 20 HNSCC patients, pre-, during and post-treatment, as well as developed methods to detect mutations in cfDNA at very low frequency with the use of unique molecular barcodes.

HNSCC is a highly-represented cancer worldwide whose patients frequently develop local or distal relapse in the first two years (Linge et al., 2016; Specenier and Vermorken, 2018). For these reasons, close monitoring strategies are of high importance to anticipate recurrence as early as possible. The analysis of ctDNA kinetics in locally advanced HNSCC patients receiving primary chemoradiation (RCTX) revealed that ctDNA

fractions before treatment correlated positively with the gross tumor volume, a phenomenon already described in a preclinical model of HNSCC (Muhanna et al., 2017).

The ctDNA dynamics during treatment showed a clear dosage dependency. This fact indicates that surveilling the fluctuations of ctDNA in the plasma during treatment could be a new way to monitor patients and might also help to adjust the ongoing treatment regime, as already described for bladder cancer (Christensen et al., 2019). However, the power of this strategy relies on the number of targeted mutations, meaning that greater number of analyzed variants would show more confident results.

The analysis of the ctDNA levels in the first follow-up after treatment seems to be a potential prognostic test for relapse in HNSCC patients, as already described for colorectal cancer (Tie et al., 2016). Using minimal/molecular residual disease (MRD) analysis, we detected significant amount of tumor fragments in two patients of our cohort, which afterwards presented with relapse. Moreover, none of the relapse-free individuals were MRD-positive. However, several patients without detectable MRD also suffered from relapse, indicating that our method and study design had sensitivity problems for detecting very low ctDNA levels. Benchmarking analysis of our MRD algorithm suggested that this low sensitivity could have been caused by insufficient sequencing depth, but more importantly, by the low number of monitored variants per patient in combination with the low input DNA typically obtained from plasma.

Our study also revealed that circulating HPV DNA (circulating virus DNA, cvDNA) is detectable in the patient´s plasma, showing the same dynamic properties as the ctDNA. Hence, we suggest the use of circulating HPV DNA as additional biomarker for the detection of HNSCC with two main applications: (1) as a blood-based marker for early detections similar to the detection of EBV for nasopharyngeal carcinoma (Chan et al., 2013; Wang et al., 2013) and (2) as a post-treatment test to predict disease recurrence.

Finally, all these results were obtained thanks to a previously developed set of tools to detect somatic mutations and ctDNA fragments at very low fraction with the use of barcodes (detailed method described in appendix). This was split in two main parts: (1) barcode correction and (2) variant calling and minimal residual diseases (MRD) detection.

Our barcode correction strategy decreased substantially the error rates (background noise) of our samples, allowing to reduce the detection limits of our somatic variant caller (Newman et al., 2016c). We showed that the use of barcode and mapping information to collapse and group the reads in families of duplicates was extremely useful to distinguish PCR and sequencing errors, allowing variant calling to reduce the false discovery rate at very low allele fractions.

Our somatic variant calling, based on the use of beta-binomial distribution to model errors combined with barcode correction information, permitted us to call ultra-rare somatic mutations (as low as 1-2 in 10,000 fragments). However, although ultra-low frequency variants could be detected with our algorithm, high haploid genome equivalents are hardly recovered in conventional liquid biopsy, where only few milliliters of plasma are isolated (representing few thousands of haploid genome equivalents) (Wan et al., 2017).

Our MRD strategy tries to compensate for the low DNA input by collapsing observations from all targeted variants per individual into a single observations and p-value and hence, increasing the number of monitored variants would increase the chances of detecting tumor fragments at very low fraction. Benchmarking and simulations revealed that if background noise is reduced down to 0.005-0.001 %, ctDNA fractions as low as 0.0075% can be detected with high sensitivity in 5,000 haploid genome equivalents (around 16ng DNA) with only 50 targeted variants.

In summary, the development of this variant calling strategy and the barcode correction permitted us to detect ctDNA fragments present at extremely low fractions, which has important implications for monitoring cancer patients pre-, during and post-treatment as shown in the HNSCC project (chapter 2) and benchmarking analysis (appendix).

Mosaic mutations in healthy individuals

The accumulation of DNA mutations during life is inevitable. Although many cell mechanisms are involved in the preservation of genome integrity, cells still acquire mutations during development and life, generating populations of cells with different genomic profiles in the same individual, a phenomenon named mosaicism (Acuna-Hidalgo et al., 2016; Muyas et al., 2019b).

In the chapter 3, we characterized mosaic mutations across 49 tissues of 570 healthy individuals (a total of 10,097 RNA-seq samples) using the GTEx consortium data (Ardlie et al., 2015; Consortium, 2017; Muyas et al., 2019b). As so far there is not a defined methodology to detect somatic variants in RNA-seq data, we developed an algorithm able to achieve high precision and recall detecting somatic calls from expression data (85 % and 71 % for precision and recall, respectively).

Our multi-tissue, multi-individual approach has allowed us to identify mosaic mutations occurring during various stages of human embryo development and life. We estimated that newborns harbor on average around 0.5 - 1 mosaic mutations in exons affecting multiple tissues/organs, and likely a greater number of organ-specific mutations. These findings suggest that mosaic mutations have similar frequencies to germline *de novo* mutations and could explain a substantial fraction of unresolved cases of sporadic and rare genetic disorders, as well as play a role in cancer predisposition syndromes (Acuna-Hidalgo et al., 2016; Muyas et al., 2019b).

The fact that only 41 % of the early embryonic mosaic mutations are detected in the expressed genes of blood reveals that a high fraction of early mosaic mutations might be missed by blood-based genetic diagnostic tests (Muyas et al., 2019b). The observation could be explained by the asymmetric cell doubling model during early embryogenesis suggested by Ju et al (Ju et al., 2017), which describes an unequal contribution of early embryonic cells to adult somatic tissues. The implications of these findings demonstrate the necessity of developing new diagnostic tests, which should characterize the majority of cell populations present in the human body. Such a test could be liquid biopsy and the analysis of cell-free DNA as described in chapter 2.

Further analysis of mutational signatures showed an association of embryonic mosaic mutations with spontaneous deamination of methylated cytosine (leading to C>T transitions at CpG dinucleotides), which likely reflects a cell-cycle-dependent mutational clock as suggested by Alexandrov et al (Alexandrov et al., 2015).

The investigation of mutations acquired after birth (tissue-specific mutations) revealed that esophagus mucosa and sun-exposed skin accumulated mutations with the pass of years, a fact that correlates with environmental exposure to food or ultra-violate (UV) light during life, respectively (Alexandrov et al., 2013). Moreover, the analysis of mutations

in these two tissues showed an enrichment of non-silent mutations (signature of positive selection) in the genes *NOTCH1* and *TP53*, a phenomenon that was not found in any other tissues analyzed in this study. These two findings (the relation of somatic mutations with age and the positive selection in these two cancer genes) agrees with results recently described in the literature (Martincorena et al., 2015, 2018; Yizhak et al., 2019; Yokoyama et al., 2019), demonstrating the capacity of expression data to be exploited as a source to detect somatic mutations. Moreover, the presence of somatic mutations in cancer genes and non-malignant samples updates the vision of how genomes and cancers behave through life and aging.

The mutational signature analysis of somatic mutations acquired during life across different groups of tissues was able to confirm some expected signatures. For instance, we found the UV signature in sun-exposed skin, which was not present in non-sun-exposed skin (Alexandrov et al., 2013). However, we also discovered a non-expected mutational signature associated with the food-borne carcinogen aflatoxin in tissues of the gastrointestinal tract. Although the aflatoxin signature was previously reported in some liver cancers (Alexandrov et al., 2013; Chawanthayatham et al., 2017), we also found it in the tissues of the gastrointestinal tract of healthy individuals even after exclusion of liver. Therefore, this discovery indicates that aflatoxin-related mutations are spread in many tissues of this tract that are in contact with food, and could play a role in the development of cancer in more organs than previously thought (Muyas et al., 2019b).

In conclusion, in this chapter we have described how RNA-seq data from multiple tissues and individuals have been used to generate a high-resolution landscape of mosaic mutations acquired during different stages of embryogenesis and life, as well as to assign the occurrence of embryonic mutations to specific germ layers or tissues. Our findings have significant implications for clinical diagnostics, as samples from the tissue(s) affected by a mosaic mutation are often unavailable. In summary, our study reveals a surprisingly high number of embryonic mosaic mutations in coding regions, implying novel hypotheses and diagnostic procedures for investigating genetic causes of diseases and cancer predispositions (Muyas et al., 2019b).

Final discussion

In this thesis we have developed methods to distinguish different types of errors from real variants and somatic mutations. The relevance, impact in downstream analysis and the impossibility of filtering some systematic errors with conventional methods make ABB a valuable tool to achieve highly accurate variant calls in somatic and germline studies. Moreover, we have demonstrated that the use of unique molecular identifiers (UMIs) or barcodes is highly effective to remove PCR and sequencing errors, which can impair the detection limit and accuracy of somatic mutations detected at very low variant allele fraction (VAF).

Our variant calling strategy, which uses barcode information to detect ultra-rare somatic variants in cfDNA, is useful to monitor cancer patients pre-, during and post-treatment and correlates with treatment response. This approach also permits to detect minimal residual diseases (MRD) below the DNA input limits classically found in cfDNA analysis (few thousands of haploid genome equivalents), and represents a potential prognosis tool to predict recurrence of disease. Studies in other cancers, with bigger cohorts and higher number of mutations per individual might expand the knowledge of cfDNA applications in clinical research and will bring the personalized medicine to a new era.

Finally, the use of previous knowledge has allowed us to investigate the mutational integrity of human genome of healthy individuals through embryogenesis and life. The high number of embryonic mosaic mutations in coding regions entails novel hypotheses and diagnostic procedures for investigating genetic causes of disorders and cancer predisposition.

CONCLUSIONS

1. Impact and detection of systematic errors

1.1. Our novel genotype callability estimator based on allele balance bias (ABB) identifies systematic variant calling errors not found by other measures and can improve the accuracy of germline and somatic variant sets as well as disease association studies in families or large cohorts.

1.2. Up to 4% of the positions called as germline variants and 8% of positions called as somatic mutations by state-of-the art methods show high ABB scores (indicative of systematic errors).

1.3. Sanger validation of random variants showed that ABB correlates with the likelihood to identify false positive SNVs.

1.4. Systematic errors are highly enriched in low complexity and repetitive regions, although they are also found in other parts of the genome that standard filters cannot distinguish.

1.5. Sites prone to systematic errors are highly enriched in public variant databases, especially in dbSNP, demonstrating that variant callsets created consistently by a defined and reproducible pipelines and parameter setting are preferable for most analysis purposes.

1.6. Systematic errors resulting in false genotype-phenotype associations can be identified by ABB in case-control studies.

2. Monitoring cancer patients pre-, during and post-treatment using cell-free DNA

2.1. Circulating-tumor DNA (ctDNA) fractions correlated positively with the gross tumor volume.

2.2. ctDNA dynamics during treatment showed a clear dosage dependency under radiochemotherapy treatment, suggesting that surveilling the fluctuations of ctDNA in the plasma could be a new way to monitor the patient's response to treatment and might also help to adjust the ongoing treatment regime.

2.3. The presence of minimal/molecular residual disease in the plasma in the first follow-up after treatment predicts cancer recurrence.

2.4. DNA from human papilloma virus (here termed circulating virus DNA, cvDNA) is detectable in the plasma of some HNSCC patients, showing the same dynamic properties as the ctDNA. This suggests the usability of circulating HPV DNA as a post-treatment test to predict disease recurrence.

2.5. Our barcode correction strategy is able to detect and correct PCR and sequencing errors, allowing as to detect somatic variants at ultra-low allele frequency.

2.6. To compensate for the low number of haploid genome equivalents recovered in cfDNA analysis from a few ml of plasma, our MRD strategy collapses the information of all targeted variants per individual to increase the sensitivity to detect tumor fragments at very low fraction.

3. Mosaic mutations in healthy individuals

3.1. We developed an algorithm to detect and call somatic mutations in RNA-seq samples with a precision and recall of 85 % and 71 %, respectively.

3.2. Our multi-tissue, multi-individual approach estimates the embryonic mosaic mutation (EMM) rate around 1.32×10^{-8} per nucleotide per healthy individual, which corresponds to an average of 0.5–1 mutations in the exome of newborns.

3.3. EMMs are as frequent as germline de novo mutations and could explain a substantial fraction of unsolved sporadic diseases and cancer predisposition syndromes.

3.4. Only 41 % of the early EMMs are detected in the expressed genes of blood, revealing that a large fraction of early mosaic mutations could be missed by blood-based genetic diagnostic tests, and implying the necessity to develop novel diagnostic procedures.

3.5. EMMs are dominated by a mutational signature associated with spontaneous deamination of methylated cytosines and the number of cell divisions.

3.6. The single-tissue mutational rates of sun-exposed skin and esophagus mucosa showed significant correlation with age, supporting the idea that tissues heavily exposed to carcinogens (UV and food) accumulate lots of mutations during life even in healthy individuals.

3.7. The cancer genes *TP53* and *NOTCH1* present positive selection in sun-exposed skin and esophagus mucosa of cancer-free individuals, a phenomenon not observed in other tissues.

3.8. Mutations acquired after birth in normal tissues of the gastrointestinal tract seem to be associated with the food-borne carcinogen aflatoxin.

REFERENCES

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526:68–74.

Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L, 1000 Genomes Project Consortium T 1000 GP, Lee C, Gerstein M. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res 23:2042–52.

Acuna-Hidalgo R, Bo T, Kwint MP, Vorst M Van De, Pinelli M, Veltman JA, Hoischen A, Vissers LELM, Gilissen C. 2015. Post-zygotic Point Mutations Are an Underrecognized Source of de Novo Genomic Variation. Am J Hum Genet 97:67–74.

Acuna-Hidalgo R, Sengul H, Steehouwer M, Vorst M van de, Vermeulen SH, Kiemeney LALM, Veltman JA, Gilissen C, Hoischen A. 2017. Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. Am J Hum Genet 101:50–64.

Acuna-Hidalgo R, Veltman JA, Hoischen A. 2016. New insights into the generation and role of de novo mutations in health and disease. Genome Biol 17:241.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, et al. 2000. The Genome Sequence of Drosophila melanogaster. Science (80-) 287:2185–2195.

Ahn SM, Chan JYK, Zhang Z, Wang H, Khan Z, Bishop JA, Westra W, Koch WM, Califano JA. 2014. Saliva and plasma quantitative polymerase chain reaction-based detection and surveillance of human papillomavirus-related head and neck cancer. JAMA Otolaryngol Head Neck Surg 140:846–54.

Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. 2015. Clock-like mutational processes in human somatic cells. Nat Genet 47:1402–7.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, et al. 2013. Signatures of mutational processes in human cancer. Nature 500:415–21.

Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, Heisler LE, Beck T a., Simpson JT, Tonon L, Sertier A-S, Patch A-M, et al. 2015. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun 6:10001.

Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-) 348:648–660.

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 46:2159–2168.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, et al. 2015. A global reference for human genetic variation. Nature 526:68–74.

Auwera G a. Van der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K V., et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. 1–33 p.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK-S, Jeong KJ, et al. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173:371-385.e18.

Beadle GW, Tatum EL. 1941. Genetic Control of Biochemical Reactions in Neurospora. Proc Natl Acad Sci 27:499–506.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, Antonarakis ES, Azad NS, et al. 2014. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. Sci Transl Med 6:224ra24.

Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, Perez-Amodio S, Strippoli P, et al. 2013. An estimation of the number of cells in the human body. Ann Hum Biol 40:463–471.

Biesecker LG, Spinner NB. 2013. A genomic view of mosaicism and human disease. Nat Rev Genet 14:307–320.

Birkeland AC, Ludwig ML, Meraj TS, Brenner JC, Prince ME. 2015. The Tip of the Iceberg: Clinical Implications of Genomic Sequencing Projects in Head and Neck Cancer. Cancers (Basel) 7:2094–109.

Birkenkamp-Demtröder K, Nordentoft I, Christensen E, Høyer S, Reinert T, Vang S, Borre M, Agerbæk M, Jensen JB, Ørntoft TF, Dyrskjøt L. 2016. Genomic Alterations in Liquid Biopsies from Patients with Bladder Cancer. Eur Urol 70:75–82.

Botezatu I, Serdyuk O, Potapova G, Shelepov V, Alechina R, Molyaka Y, Ananév V, Bazin I, Garin A, Narimanov M, Knysh V, Melkonyan H, et al. 2000. Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. Clin Chem 46:1078–84.

Bowden R, Davies RW, Heger A, Pagnamenta AT, Cesare M de, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY, Popitsch N, Ip CLC, et al. 2019. Sequencing of human genomes with nanopore technology. Nat Commun 10:1869.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394–424.

Budach V, Stromberger C, Poettgen C, Baumann M, Budach W, Grabenbauer G, Marnitz S, Olze H, Wernecke K-D, Ghadjar P. 2015. Hyperfractionated accelerated radiation therapy (HART) of 70.6 Gy with

concurrent 5-FU/Mitomycin C is superior to HART of 77.6 Gy alone in locally advanced head and neck cancer: long-term results of the ARO 95-06 randomized phase III trial. Int J Radiat Oncol Biol Phys 91:916–24.

Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. 2015. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet 31:382–392.

Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, Schrock A, Campbell B, Shlien A, Chmielecki J, Huang F, He Y, et al. 2017. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med 9:34.

Chan KCA, Hung ECW, Woo JKS, Chan PKS, Leung S-F, Lai FPT, Cheng ASM, Yeung SW, Chan YW, Tsui TKC, Kwok JSS, King AD, et al. 2013. Early detection of nasopharyngeal carcinoma by plasma Epstein-Barr virus DNA analysis in a surveillance program. Cancer 119:1838–44.

Chaudhuri AA, Chabon JJ, Lovejoy AF, Newman AM, Stehr H, Azad TD, Khodadoust MS, Esfahani MS, Liu CL, Zhou L, Scherer F, Kurtz DM, et al. 2017. Early Detection of Molecular Residual Disease in Localized Lung Cancer by Circulating Tumor DNA Profiling. Cancer Discov 7:1394–1403.

Chawanthayatham S, Valentine CC, Fedeles BI, Fox EJ, Loeb LA, Levine SS, Slocum SL, Wogan GN, Croy RG, Essigmann JM. 2017. Mutational spectra of aflatoxin B 1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma . Proc Natl Acad Sci 114:E3101–E3109.

Check Hayden E. 2014. Is the $1,000 genome for real? Nature.

Chin RI, Chen K, Usmani A, Chua C, Harris PK, Binkley MS, Azad TD, Dudley JC, Chaudhuri AA. 2019.

Christensen E, Birkenkamp-Demtröder K, Sethi H, Shchegrova S, Salari R, Nordentoft I, Wu H-T, Knudsen M, Lamy P, Lindskrog SV, Taber A, Balcioglu M, et al. 2019. Early Detection of Metastatic Relapse and Monitoring of Therapeutic Efficacy by Ultra-Deep Sequencing of Plasma Cell-Free DNA in Patients With Urothelial Bladder Carcinoma. J Clin Oncol 37:1547–1557.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of

somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31:213–9.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. SNPs in the genome of Drosophila melanogaster strain *w1118; iso-2; iso-3* . Fly (Austin) 6:80–92.

Clancy S. 2008. Genetic Mutation. 1(1):187 p.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–71.

Colli LM, Machiela MJ, Zhang H, Myers TA, Jessop L, Delattre O, Yu K, Chanock SJ. 2017. Landscape of Combination Immunotherapy and Targeted Therapy to Improve Cancer Management. Cancer Res 77:3666–3671.

Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, Deardorff MA, Krantz ID, Hakonarson H, Spinner NB. 2010. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum Mol Genet 19:1263–1275.

Consortium Gte. 2017. Genetic effects on gene expression across human tissues. Nature 550:204–213.

Consortium T 1000 GP. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Cordaux R, Batzer M. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet 10:691–703.

Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, de Vos WM, Ehrlich SD, Fraser CM, Hattori M, Huttenhower C, Jeffery IB, et al. 2018. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 3:8–16.

Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, Fostel JL, Friedrich DC, Perrin D, Dionne D, Kim S, Gabriel SB, et al. 2013. Discovery and characterization of artifactual mutations in deep coverage

targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 41:e67–e67.

Crick FHC, Barnett L, Brenner S, Watts-Tobin RJ. 1961. General Nature of the Genetic Code for Proteins. Nature 192:1227–1232.

Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, Thornton K, Agrawal N, Sokoll L, Szabo SA, Kinzler KW, Vogelstein B, et al. 2008. Circulating mutant DNA to assess tumor dynamics. Nat Med 14:985–990.

Dijk EL van, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The Third Revolution in Sequencing Technology. Trends Genet 34:666–681.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a., Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Eder T, Hess AK, Konschak R, Stromberger C, Jöhrens K, Fleischer V, Hummel M, Balermpas P, Grün J von der, Linge A, Lohaus F, Krause M, et al. 2019. Interference of tumour mutational burden with outcome of patients with head and neck cancer treated with definitive chemoradiation: a multicentre retrospective study of the German Cancer Consortium Radiation Oncology Group. Eur J Cancer 116:67–76.

ENCODE Project Consortium TEP. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–40.

Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, Bult C, Tomb J, Dougherty B, Merrick J, al. e. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (80-) 269:496–512.

Freudenberg-hua Y, Freudenberg J, Kluck N, Cichon S, Propping P, Nöthen MM. 2003. Single Nucleotide Variation Analysis in 65 Candidate Genes for CNS Disorders in a Representative Sample of the European Population Single Nucleotide Variation Analysis in 65 Candidate Genes for CNS Disorders in a Representative Sample of the European Popu. Genome Res 2271–2276.

Fuentes Fajardo K V., Adams D, Mason CE, Sincan M, Tifft C, Toro C, Boerkoel CF, Gahl W, Markello T, Markello T. 2012. Detecting false-positive signals in exome sequencing. Hum Mutat 33:609–613.

Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, Cheang M, Osin P, Nerurkar A, Kozarewa I, Garrido JA, Dowsett M, et al. 2015. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci Transl Med 7:302ra133.

Gayon J. 2016. From Mendel to epigenetics: History of genetics. C R Biol 339:225–230.

Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, Varela I, Phillimore B, et al. 2012. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. N Engl J Med 366:883–892.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, et al. 1996. Life with 6000 Genes. Science (80-) 274:546–567.

Griffith M, Miller CA, Griffith OL, Krysiak K, Skidmore ZL, Ramu A, Walker JR, Dang HX, Trani L, Larson DE, Demeter RT, Wendl MC, et al. 2015. Optimizing Cancer Genome Sequencing and Analysis. Cell Syst 1:210–223.

Guo Y, Li J, Li C-I, Long J, Samuels DC, Shyr Y. 2012. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics 13:666.

Guy C, Haji-Sheikhi F, Rowland CM, Anderson B, Owen R, Lacbawan FL, Alagia DP. 2019. Prenatal cell-free DNA screening for fetal aneuploidy in pregnant women at average or high risk: Results from a large US clinical laboratory. Mol Genet Genomic Med 7:e545.

Halvorsen M, Petrovski S, Shellhaas R, Tang Y, Crandall L, Goldstein D, Devinsky O. 2016. Mosaic mutations in early-onset genetic diseases. Genet Med 18:746–749.

Happle R. 1987. Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. J Am Acad Dermatol 16:899–906.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, et al. 2012. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res 22:1760–1774.

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. 1965. Structure of a Ribonucleic Acid. Science (80-) 147:1462–1465.

Huang AY, Xu X, Ye AY, Wu Q, Yan L, Zhao B, Yang X, He Y, Wang S, Zhang Z, Gu B, Zhao HQ, et al. 2014. Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. Cell Res 24:1311–1327.

Huang K-L, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, Scott AD, Krassowski M, et al. 2018. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell 173:355-370.e14.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–45.

Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, Knippers R. 2001. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res 61:1659–65.

Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman P V., Mar BG, Lindsley RC, Mermel CH, Burtt N, Chavez A, Higgins JM, Moltchanov V, et al. 2014. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. N Engl J Med 371:2488–2498.

Jeannot E, Becette V, Campitelli M, Calméjane M-A, Lappartient E, Ruff E, Saada S, Holmes A, Bellet D, Sastre-Garau X. 2016. Circulating human papillomavirus DNA detected using droplet digital PCR in the serum of patients diagnosed with early stage human papillomavirus-associated invasive carcinoma. J Pathol Clin Res 2:201–209.

Ju YS, Martincorena I, Gerstung M, Petljak M, Alexandrov LB, Rahbari R, Wedge DC, Davies HR, Ramakrishna M, Fullam A, Martin S, Alder C, et al.

2017. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature 543:714–718.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, et al. 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210.

Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT. 2018. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 15:591–594.

Kiran AM, O'Mahony JJ, Sanjeev K, Baranov P V. 2012. Darned in 2013: inclusion of model organisms and linking with Wikipedia. Nucleic Acids Res 41:D258–D261.

Kircher M, Stenzel U, Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10:R83.

Klemm SL, Shipony Z, Greenleaf WJ. 2019.

Koboldt DC, Larson DE WR. 2013. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr Protoc Bioinforma 44:15.4.1–15.4.17.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–2285.

Kurtz DM, Scherer F, Jin MC, Soo J, Craig AFM, Esfahani MS, Chabon JJ, Stehr H, Liu CL, Tibshirani R, Maeda LS, Gupta NK, et al. 2018. Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma. J Clin Oncol 36:2845–2853.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Langmead B. 2010. Aligning Short Sequencing Reads with Bowtie. Current Protocols in Bioinformatics, Hoboken, NJ, USA: John Wiley & Sons, Inc., p Unit 11.7.

Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW, Samson JM, Daza R, Doering K, Shendure J, et al. 2014. Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol 32:829–833.

Ledergerber C, Dessimoz C. 2011. Base-calling for next-generation sequencing platforms. Brief Bioinform 12:489–497.

Leemans CR, Braakhuis BJM, Brakenhoff RH. 2011. The molecular biology of head and neck cancer. Nat Rev Cancer 11:9–22.

Leemans CR, Snijders PJF, Brakenhoff RH. 2018. The molecular landscape of head and neck cancer. Nat Rev Cancer 18:269–282.

Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A, Rubertsson S, Nellgård B, Blennow K, Zetterberg H, Spalding K, Haller MJ, et al. 2016. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci 113:E1826–E1834.

Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards S V, Forest F, Gilbert MTP, Goldstein MM, Grigoriev I V, et al. 2018. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A 115:4325–4333.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr arXiv 00:3.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 1–9.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

Lindgreen S. 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes 5:337.

Lindhurst M, Teer JK, Sapp JC, Johnston JJ, Ph D, Finn EM, Peters K, Turner J, Cannons JL, Bick D, Blakemore L, Blumhorst C, et al. 2011. A Mosaic Activating Mutation in. Genome Res 611–619.

Linge A, Lohaus F, Löck S, Nowak A, Gudziol V, Valentini C, Neubeck C von, Jütz M, Tinhofer I, Budach V, Sak A, Stuschke M, et al. 2016. HPV status, cancer stem cell marker expression, hypoxia gene signatures and tumour volume identify good prognosis subgroups in patients with HNSCC after primary radiochemotherapy: A multicentre retrospective study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG). Radiother Oncol 121:364–373.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, et al. 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–585.

Lupski JR. 2013. Genetics. Genome mosaicism--one human, multiple genomes. Science 341:358–9.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, Cagan A, Murai K, Mahbubani K, Stratton MR, Fitzgerald RC, Handford PA,

et al. 2018. Somatic mutant clones colonize the human esophagus with age. Science (80-) eaau3879.

Martincorena I, Roshan A, Gerstung M, Ellis P, Loo P Van, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, Stebbings L, Menzies A, et al. 2015. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348:880–6.

McInerney P, Adams P, Hadi MZ. 2014. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. Mol Biol Int 2014:287430.

McKenna S. 2009. The Genome Analysis Toolkit. Proc Int Conf Intellect Capital, Knowl Manag Organ Learn 254–260.

Mendel G. 1866. Versuche Über Planzen-Hybriden, Verhandlungen Des Naturschenden Vereines in Brünn. 3–47 p.

Mithani SK, Smith IM, Zhou S, Gray A, Koch WM, Maitra A, Califano JA. 2007. Mitochondrial Resequencing Arrays Detect Tumor-Specific Mutations in Salivary Rinses of Patients with Head and Neck Cancer. Clin Cancer Res 13:7335–7340.

Morozova O, Marra MA. 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255–264.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628.

Mose LE, Perou CM, Parker JS. 2019. Improved indel detection in DNA and RNA via realignment with ABRA2. Bioinformatics.

Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, Mair R, Goranova T, Marass F, Heider K, Wan JCM, Supernat A, et al. 2018. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med 10:eaat4921.

Muhanna N, Grappa MA Di, Chan HHL, Khan T, Jin CS, Zheng Y, Irish JC, Bratman S V. 2017. Cell-Free DNA Kinetics in a Pre-Clinical Model of Head and Neck Cancer. Sci Rep 7:16723.

Musumeci L et al. 2011. Single Nucleotide Differences (SNDs) in the dbSNP Database May Lead to Errors in Genotyping and Haplotyping Studies. 20:200–210.

Muyas F, Bosio M, Puig A, Susak H, Domènech L, Escaramis G, Zapata L, Demidov G, Estivill X, Rabionet R, Ossowski S. 2019a. Allele balance bias identifies systematic genotyping errors and false disease associations. Hum Mutat 40:115–126.

Muyas F, Zapata L, Guigó R, Ossowski S. 2019b. The rate and spectrum of mosaic mutations during embryogenesis revealed by RNA sequencing of 49 tissues. bioRxiv 687822.

Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, Stehr H, Liu CL, Bratman S V, Say C, Zhou L, Carter JN, et al. 2016a. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 34:547–555.

Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, Stehr H, Liu CL, Bratman S V, Say C, Zhou L, Carter JN, et al. 2016b. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol Article in press.

Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, Stehr H, Liu CL, Bratman S V, Say C, Zhou L, Carter JN, et al. 2016c. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 34:547–555.

Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Loo P Van, Ju YS, et al. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534:47–54.

Norton ME, Jacobsson B, Swamy GK, Laurent LC, Ranzini AC, Brar H, Tomlinson MW, Pereira L, Spitz JL, Hollemon D, Cuckle H, Musci TJ, et al. 2015. Cell-free DNA Analysis for Noninvasive Examination of Trisomy. N Engl J Med 372:1589–1597.

Nowell P. 1976. The clonal evolution of tumor cell populations. Science (80-) 194:23–28.

Nshimyumukiza L, Menon S, Hina H, Rousseau F, Reinharz D. 2018. Cell-free DNA noninvasive prenatal screening for aneuploidy versus conventional screening: A systematic review of economic evaluations. Clin Genet 94:3–21.

Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40:D571-9.

Pattnaik S, Vaidyanathan S, Pooja DG, Deepak S, Panda B. 2012. Customisation of the exome data analysis pipeline using a combinatorial approach. PLoS One 7:.

Pfeifer SP. 2017. From next-generation resequencing reads to a high-quality variant data set. Heredity (Edinb) 118:111–124.

Pham J, Shaw C, Pursley A, Hixson P, Sampath S, Roney E, Gambin T, Kang SHL, Bi W, Lalani S, Bacino C, Lupski JR, et al. 2014. Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10 362 consecutive cases. Eur J Hum Genet 22:969–978.

Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. 2017. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res 45:D750–D757.

Pignon J-P, Maître A le, Maillard E, Bourhis J, MACH-NC Collaborative Group. 2009. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients. Radiother Oncol 92:4–14.

Planet E, Attolini CS-O, Reina O, Flores O, Rossell D. 2012. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. Bioinformatics 28:589–590.

PLANT KE, Boye E, Green PM, Vetrie D, Flinter FA. 2000. Somatic mosaicism associated with a mild Alport syndrome phenotype. J Med Genet 37:238–239.

Poduri A, Evrony GD, Cai X, Walsh CA. 2013. Somatic mutation, genomic variation, and neurological disease. Science 341:1237758.

Potter SS. 2018. Single-cell RNA sequencing for the study of development, physiology and disease. Nat Rev Nephrol 14:479–492.

Prochazkova K, Pavlikova K, Minarik M, Sumerauer D, Kodet R, Sedlacek Z. 2009. Somatic TP53 mutation mosaicism in a patient with Li-Fraumeni syndrome. Am J Med Genet Part A 149:206–211.

Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, Munar M, Rubio-Pérez C, Jares P, Aymerich M, Baumann T, Beekman R, et al. 2015. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 526:519–524.

Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A, López-Guerra M, Colomer D, et al. 2011. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet 44:47–52.

Risques RA, Kennedy SR. 2018. Aging and the rise of somatic cancer-associated mutations in normal tissues. PLOS Genet 14:e1007108.

Rivière JB, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, Conway RL, St-Onge J, Schwartzentruber JA, Gripp KW, Nikkel SM, Worthylake T, Sullivan CT, et al. 2012. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat Genet 44:934–940.

Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. 2016. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 17:31.

Rossant J, Tam PPL. 2017. New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation. Cell Stem Cell 20:18–28.

Rossant J, Tam PPL. 2018. Exploring early human embryo development. Science (80-) 360:1075–1076.

Ruark E, Snape K, Humburg P, Loveday C, Bajrami I, Brough R, Rodrigues DN, Renwick A, Seal S, Ramsay E, Duarte SDV, Rivas MA, et al. 2013.

Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. Nature 493:406–410.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74:5463–5467.

Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 28:1811–1817.

Sausen M, Phallen J, Adleff V, Jones S, Leary RJ, Barrett MT, Anagnostou V, Parpart-Li S, Murphy D, Kay Li Q, Hruban CA, Scharpf R, et al. 2015. Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. Nat Commun 6:7686.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci 109:14508–14513.

Schroeder CM, Hilke FJ, Löffler MW, Bitzer M, Lenz F, Sturm M. 2017. A comprehensive quality control workflow for paired tumor-normal NGS experiments. Bioinformatics 33:1721–1722.

Schwarze K, Buchanan J, Taylor JC, Wordsworth S. 2018. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet Med 20:1122–1130.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. Nature 550:345–353.

Sidransky D, Tokino T, Hamilton SR, Kinzler K, Levin B, Frost P, Vogelstein B. 1992. Identification of ras oncogene mutations in the stool of patients with curable colorectal tumors. Science (80-) 256:102–105.

Sinville R, Soper SA. 2007. High resolution DNA separations using microchip electrophoresis. J Sep Sci 30:1714–1728.

Sleep J a, Schreiber AW, Baumann U. 2013. Sequencing error correction without a reference genome. BMC Bioinformatics 14:367.

Specenier P, Vermorken JB. 2018. Optimizing treatments for recurrent or metastatic head and neck squamous cell carcinoma. Expert Rev Anticancer Ther 18:901–915.

Sriram KB, Relan V, Clarke BE, Duhig EE, Windsor MN, Matar KS, Naidoo R, Passmore L, McCaul E, Courtney D, Yang IA, Bowman R V, et al. 2012. Pleural fluid cell-free DNA integrity index to identify cytologically negative malignant pleural effusions including mesotheliomas. BMC Cancer 12:428.

Sturm M, Schroeder C, Bauer P. 2016. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. BMC Bioinformatics 17:208.

Summa S De, Malerba G, Pinto R, Mori A, Mijatovic V, Tommasi S. 2017. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics 18:119.

Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, Torres C de, Dienstmann R, et al. 2018. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med 10:25.

Tan MH, Li Q, Shanmugam R, Piskol R, Kohler J, Young AN, Liu KI, Zhang R, Ramaswami G, Ariyoshi K, Gupte A, Keegan LP, et al. 2017. Dynamic landscape and regulation of RNA editing in mammals. Nature 550:249–254.

Taylor JC, Martin HC, Lise S, Broxholme J, Cazier J-B, Rimmer A, Kanapin A, Lunter G, Fiddy S, Allan C, Aricescu AR, Attar M, et al. 2015. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet 47:717–726.

The C. elegans Sequencing Consortium. 1998. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science (80-) 282:2012–2018.

Thierry AR, Mouliere F, Gongora C, Ollier J, Robert B, Ychou M, Rio M Del, Molina F. 2010. Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. Nucleic Acids Res 38:6159–6175.

Tie J, Cohen JD, Wang Y, Li L, Christie M, Simons K, Elsaleh H, Kosmider S, Wong R, Yip D, Lee M, Tran B, et al. 2019. Serial circulating tumour DNA analysis during multimodality treatment of locally advanced rectal cancer: A prospective biomarker study. Gut 68:663–671.

Tie J, Wang Y, Tomasetti C, Li L, Springer S, Kinde I, Silliman N, Tacey M, Wong H-L, Christie M, Kosmider S, Skinner I, et al. 2016. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Sci Transl Med 8:346ra92-346ra92.

Tinhofer I, Staudte S. 2018. Circulating tumor cells as biomarkers in head and neck cancer: recent advances and future outlook. Expert Rev Mol Diagn 18:897–906.

Tomasetti C, Vogelstein B, Parmigiani G. 2013. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci 110:1999–2004.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, et al. 2001. The sequence of the human genome. Science 291:1304–51.

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P, Schaefer C, Risch N. 2014. Estimating genotype error rates from high-coverage next-generation sequence data. 1734–1739.

Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, Pacey S, Baird R, Rosenfeld N. 2017. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer 17:223–238.

Wang W-Y, Twu C-W, Chen H-H, Jiang R-S, Wu C-T, Liang K-L, Shih Y-T, Chen C-C, Lin P-J, Liu Y-C, Lin J-C. 2013. Long-term survival analysis of nasopharyngeal carcinoma by plasma Epstein-Barr virus DNA levels. Cancer 119:963–970.

Wang Y, Springer S, Mulvey CL, Silliman N, Schaefer J, Sausen M, James N, Rettig EM, Guo T, Pickering CR, Bishop JA, Chung CH, et al. 2015a. Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas. Sci Transl Med 7:293ra104.

Wang Y, Springer S, Zhang M, McMahon KW, Kinde I, Dobbyn L, Ptak J, Brem H, Chaichana K, Gallia GL, Gokaslan ZL, Groves ML, et al. 2015b. Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord. Proc Natl Acad Sci 112:9704–9709.

Watson JD, Crick FHC. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171:737–738.

Wei W, Keogh MJ, Aryaman J, Golder Z, Kullar PJ, Wilson I, Talbot K, Turner MR, McKenzie C-A, Troakes C, Attems J, Smith C, et al. 2018a. Frequency and signature of somatic variants in 1461 human brain exomes. Genet Med.

Wei W, Keogh MJ, Aryaman J, Golder Z, Kullar PJ, Wilson I, Talbot K, Turner MR, McKenzie C-A, Troakes C, Attems J, Smith C, et al. 2018b. Frequency and signature of somatic variants in 1461 human brain exomes. Genet Med 1.

Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res 40:11189–201.

Wu R, Kaiser AD. 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol 35:523–537.

Xi L, Pham TH-T, Payabyab EC, Sherry RM, Rosenberg SA, Raffeld M. 2016. Circulating Tumor DNA as an Early Indicator of Response to T-cell Transfer Immunotherapy in Metastatic Melanoma. Clin Cancer Res 22:5480–5486.

Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, McMichael JF, Schmidt HK, Yellapantula V, Miller CA, Ozenberger BA, Welch JS, et al. 2014. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med 20:1472–1478.

Yamada S, Samtani RR, Lee ES, Lockett E, Uwabe C, Shiota K, Anderson SA, Lo CW. 2010. Developmental Atlas of the Early First Trimester Human Embryo. Dev Dyn 239:1585.

Yizhak K, Aguet F, Kim J, Hess JM, Kübler K, Grimsby J, Frazer R, Zhang H, Haradhvala NJ, Rosebrock D, Livitz D, Li X, et al. 2019. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science 364:eaaw0726.

Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, Shiozawa Y, Sato Y, Aoki K, Kim SK, Fujii Y, Yoshida K, et al. 2019. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature 565:312–317.

Youssoufian H, Pyeritz RE. 2002. Mechanisms and consequences of somatic mosaicism in humans. Nat Rev Genet 3:748–758.

Yu Y, Xu T, Yu Y, Hao P, Li X. 2010. Association of tissue lineage and gene expression: conservatively and differentially expressed genes define common and special functions of tissues. BMC Bioinformatics 11 Suppl 1:S1.

Yurov YB, Vorsanova SG, Iourov IY, Demidova IA, Beresheva AK, Kravetz VS, Monakhov V V., Kolotii AD, Voinova-Ulas VY, Gorbachevskaya NL. 2007. Unexplained autism is frequently associated with low-level mosaic aneuploidy. J Med Genet 44:521–525.

Zapata L, Pich O, Serrano L, Kondrashov FA, Ossowski S, Schaefer MH. 2018. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. Genome Biol 19:67.

Zapata L, Susak H, Drechsel O, Friedländer MR, Estivill X, Ossowski S. 2017. Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. Sci Rep 7:13124.

Zhang W, He H, Zang M, Wu Q, Zhao H, Lu L ling, Ma P, Zheng H, Wang N, Zhang Y, He S, Chen X, et al. 2017. Genetic Features of Aflatoxin-Associated Hepatocellular Carcinoma. Gastroenterology 153:249-262.e2.

Zwirner K, Hilke FJ, Demidov G, Ossowski S, Gani C, Rieß O, Zips D, Welz S, Schroeder C. 2018a. Circulating cell-free DNA: A potential biomarker to differentiate inflammation and infection during radiochemotherapy. Radiother Oncol 129:575–581.

Zwirner K, Hilke FJ, Demidov G, Ossowski S, Gani C, Rieß O, Zips D, Welz S, Schroeder C. 2018b. Circulating cell-free DNA: A potential biomarker to differentiate inflammation and infection during radiochemotherapy. Radiother Oncol 129:575–581.

Zwirner K, Hilke FJ, Demidov G, Socarras Fernandez J, Ossowski S, Gani C, Thorwarth D, Riess O, Zips D, Schroeder C, Welz S. 2019. Radiogenomics in head and neck cancer: correlation of radiomic heterogeneity and somatic mutations in TP53, FAT1 and KMT2D. Strahlentherapie und Onkol.

ANNEX

Publications during PhD

Published

Muyas F, Bosio M, Puig A, Susak H, Domènech L, Escaramis G, Zapata L, Demidov G, Estivill X, Rabionet R, Ossowski S. 2019a. Allele balance bias identifies systematic genotyping errors and false disease associations. Hum Mutat 40:115–126.

Bosio M, Drechsel O, Rahman R, Muyas F, Rabionet R, Bezdan D, Domenech Salgado L, Hor H, Schott J, Munell F, Colobran R, Macaya A, et al. 2019. eDiVA—Classification and prioritization of pathogenic variants for clinical diagnostics. Hum Mutat humu.23772.

Preprint and/or under review

Muyas F, Zapata L, Guigó R, Ossowski S. 2019b. The rate and spectrum of mosaic mutations during embryogenesis revealed by RNA sequencing of 49 tissues. bioRxiv 687822.

Waszak SM, Tiao G, Zhu B, Rausch T, Muyas F, Rodríguez-Martín B, Rabionet R, Yakneen S, Escaramis G, Li Y, Saini N, Roberts SA, et al. 2017. Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. bioRxiv 208330.

In preparation

F. Hilke*, F. Muyas*, J. Matthes, I. Bonzheim, S. Welz, S. Ossowski, O. Rieß, D. Zips, C. Schroeder, K. Zwirner. Dynamics of circulating cell-free tumor DNA in HNSCC patients receiving radiochemotherapy correlates with treatment response. *In preparation*.

F. Muyas, J. Admard, F. Hilke, C. Schroeder, S. Ossowski. Use of Unique Molecular Identifiers (UMIs) to detect ultra-rare somatic variants in cell-free DNA. *In preparation.*

H. Susak, L. Serra-Saurina, R. Rabionet, L. Domènech, M. Bosio, F. Muyas, X. Estivill, G. Escaramis, S. Ossowski. Efficient and Flexible Integration of Variant Characteristics in Rare Variant Association Studies Using Integrated Nested Laplace Approximation. *In preparation.*

PANCANCER ANALYSIS OF WHOLE GENOMES (PCAWG) marker paper. PCAWG consortium. *In preparation*.