Quantifying Circulating Tumour DNA from Liquid Biopsies and Application to Lymphoma

by
Kristena Daley

B.Sc., Mount Allison University, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2022

Declaration of Committee

Name:	**Kristena Daley**
Degree:	**Master of Science (Molecular Biology and Biochemistry)**
Title:	**Quantifying Circulating Tumour DNA from Liquid Biopsies and Application to Lymphoma**
Committee:	**Chair:** **Peter Unrau** Professor, Molecular Biology and Biochemistry
	Ryan Morin Supervisor Associate Professor, Molecular Biology and Biochemistry
	Amy Lee Committee Member Assistant Professor, Molecular Biology and Biochemistry
	Christian Steidl Committee Member Professor, Pathology and Laboratory Medicine University of British Columbia
	Ly Vu Committee Member Assistant Professor, Pharmaceutical Sciences University of British Columbia
	William Hsiao Examiner Associate Professor, Health Sciences

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Diffuse large B-cell lymphoma is an aggressive and heterogeneous type of non-Hodgkin lymphoma. Circulating tumour DNA (ctDNA) is composed of highly fragmented tumour-derived cell-free DNA (cfDNA) and can be extracted from a patient's bloodstream. This "liquid biopsy" contains tumour-specific genetic alterations inclusive of simple somatic mutations and copy number variations (CNVs). Quantifying ctDNA is challenging, as existing tools are inconsistent in determining the fraction of ctDNA in a plasma sample (known as the purity) and have variable sensitivity at low levels. Leveraging CAPP-Seq and low-pass WGS (lpWGS), I developed a bioinformatic program called PurEctDNA that estimates cfDNA purity levels with high accuracy across a broad range (5-100%). With this, I modified the CNV caller, WisecondorX, to infer purity and produce improved copy number profiles from lpWGS data. Utilizing these new methods could enable more accurate and sensitive detection of ctDNA from lymphoma patients thereby improving our ability to monitor disease progression non-invasively.

Keywords: Diffuse large B-cell lymphoma; circulating tumour DNA; liquid biopsy; purity estimation; copy number profiling

Dedication

To my mom and sisters, for the overwhelming love and support throughout my degree,
My dad, a cancer survivor who continuously inspires me to pursue my passions,
Friends and family who battled cancer until the end,
And of course, those who survived the disease or are still fighting it.

Acknowledgements

I would like to begin by acknowledging my supervisor, Dr. Ryan Morin. From our first meeting over Skype with an 18-hour time difference (Melbourne, Australia to Vancouver, Canada) to me joining the lab from Nova Scotia at the start of the pandemic, to me finally moving to Vancouver, you continuously supported my work while accommodating every time zone difference that we had. You ensured that I was always included and understood topics at CLC- and lab meetings, and took time out of your busy schedule to make sure that I didn't get lost with all of the new terms and knowledge that I was learning at the beginning of my degree. Your passion and enthusiasm for cancer genomics and bioinformatics is infectious, and I am so thankful that I got to attend two of your courses (even if they were both through Zoom). Thank you for always being there for me and helping me through all of the many challenges that grad school has thrown my way, from working late to help me run my first bioinformatic pipeline (ProDuSe) on the command line to working through all of the bugs in my first snakefile with me (which, let's be honest, I had no idea what was happening at the time so that was all you), and even to celebrating my smaller accomplishments such as changing the colour of those offset WisecondorX segments for the first time. I will always appreciate your random dad jokes and stories from your grad school days, as these made my days a little brighter through having to work from home during the entirety of my degree. I highly valued your attempts to make the labs working environment better throughout the pandemic, including the Zoom socials and lab logo contests (leading to your first ever tattoo, of course). I absolutely cannot thank you enough, Ryan, for everything.

I also of course must thank the Morin lab members, known as the "Morons", for providing a supportive online learning environment, especially for a new grad student who didn't get to meet anyone in person until my second semester. I always valued our Zoom socials (playing a lot of skribbl.io, crosswords, codenames, and among us) and summer pool parties at Ryan's, including the impromptu one that I organized literally the day before (I still can't believe we convinced Ryan to host that). A very special thank you to Laura, Kostia, Annie, and Krysta for the incredible amount of support, guidance and laughs that you all have given me throughout these past two years. I absolutely could not have done this without you.

Thank you to my committee members, of course, Drs. Ly Vu, Christian Steidl, and Amy Lee for your encouragement and advice during my degree. Although my committee meeting presentations were long and every presentation had a slightly different topic, you all followed along and made sure to ask the right questions to keep me on the right path. I'd like to thank Amy for joining during my second year, I really appreciate and value your contributions as part of my committee. You were all very convincing to switch to a PhD, but also very supportive when I decided to stick with a masters, and I'd like to thank you all so much.

Lastly, I would like to say a big thank you to my friends and family, who have all been incredible throughout this degree. From supporting my decision to start a semester early during the pandemic, through the massive ups and downs that grad school had to offer, I could not have completed this degree without you all. Thank you so much to my mom, even though you put up quite the fight to try and keep me in Nova Scotia, you always told me to do what I love, even if it meant moving across the country. I always loved our many facetime calls and rant sessions and will always cherish your pep talks to keep me going. Thank you to my dad, a cancer survivor who inspired me to pursue this field of research in the first place, and who has always supported my dreams (and drove me to and from the airport too many times to count). A massive thank you also goes out to my sister, Breanna, for listening to me practice my presentations (even though I know you had no idea what anything I was saying meant), always being there for me when I was down and making every single day a million times better. I love you (and Kobe, the cat) so much. Another thank you goes to my oldest sister, Erika, her fiancé, Andrew, and Jake the dog for encouraging me to attend grad school. I have received so much life advice from you both, and our incredible trips around LA and Vancouver together definitely helped me through the stresses of grad school life. Of course, I have to thank my friends from Nova Scotia and England, and the lovely people I have met from my time in Vancouver (special shout out to Charlotte and Annie) for understanding the pressures of grad school and helping celebrate the accomplishments that I have experienced from my time at SFU. Thank you all from the bottom of my heart.

Table of Contents

List of Figures

List of Acronyms

ABC	Activated B-cell
BAF	B-allele frequency
bp	Base pair(s)
CAPP-Seq	Cancer personalized profiling by deep sequencing
CBS	Circular binary segmentation
cfDNA	Cell-free DNA
cfRNA	Cell-free RNA
CNV	Copy number variation
COO	Cell-of-origin
CTC	Circulating tumour cell
ctDNA	Circulating tumour DNA
ddPCR	Droplet digital PCR
DLBCL	Diffuse large B-cell lymphoma
EV	Extracellular vesicle
PET/CT	Positron emission tomography/computed tomography
FISH	Fluorescence *in situ* hybridization
FL	Follicular lymphoma. A type of indolent NHL
GAMBL	Genome-wide analysis of mature B-cell lymphomas
GAMBLR	Genome-wide analysis of mature B-cell lymphomas R package
GCB	Germinal centre B-cell
HMM	Hidden Markov model
Ig	Immunoglobulin
IGV	Integrated genomics viewer
IPI	International prognostic index
kb	Kilobase
LCR-modules	Lymphoid cancer research - modules
LDH	Lactate dehydrogenase
LOH	Loss of heterozygosity
lpWGS	Low-pass whole genome sequencing

MCL	Mantle cell lymphoma
MRD	Minimal residual disease
NGS	Next generation sequencing
NHL	Non-Hodgkin lymphoma
PCR	Polymerase chain reaction
PurEctDNA	Purity estimation of circulating tumour DNA
R-CHOP	Rituximab plus cyclophosphamide, doxorubicin, vincristine and prednisone
rrDLBCL	Relapsed and refractory diffuse large B-cell lymphoma
SNV	Single nucleotide variant
SSM	Simple somatic mutation
SV	Structural variation
V(D)J	Variable, diversity, and joining gene segments
VAF	Variant allele frequency
WGS	Whole genome sequencing

Chapter 1.

Introduction to Diffuse Large B-cell Lymphoma and Circulating Tumour DNA

1.1. Non-Hodgkin Lymphoma

Lymphomas are tumours that originate from cells of the lymphatic system. This malignancy is broadly divided into classical Hodgkin and non-Hodgkin lymphomas (NHLs), with the latter encompassing 90% of subtypes[1]. NHL is the 5th most commonly diagnosed cancer in both males and females in Canada as of 2021, accounting for more than 540,000 new cases around the world[2,3]. Diagnosis more often occurs at advanced disease stages resulting in a poor prognosis[4].

NHLs arise from the clonal expansion of B, T, or natural killer lymphocytes during various stages of development and differentiation, with 85% of cases coming from the B cell lineage[1,5,6]. B cell development takes place in the bone marrow where V(D)J recombination occurs to rearrange the immunoglobulin (Ig) heavy chain genes, thereby creating antibody diversity[7,8]. During this process, double-stranded DNA breaks can occur elsewhere in the genome and are resolved through DNA repair processes, leading to chromosomal translocations[7]. Whereas these steps occur in the bone marrow, the terminal stages of B cell differentiation take place in the germinal centre of the lymph node. Here, B cells are activated through antigen binding and T cell signalling, causing B cell expansion and further differentiation into memory B cells or plasma cells[7,8]. Germinal centres are also the main site of antibody generation and Ig alteration via somatic hypermutation and class-switch recombination, respectively[7]. Both processes can cause DNA damage and have the potential to contribute somatic mutations to B cells that may facilitate lymphomagenesis.

B-NHLs are a diverse group of malignancies that can be broadly separated into various types, dependent on the differentiation stage of B cells. Based on the clinical features of the individual cancers, they can be categorized as indolent lymphomas such as follicular lymphoma (FL) and aggressive subtypes such as diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma[9].

1.2. Diffuse Large B-cell Lymphoma

DLBCL is an aggressive type of NHL and the most common NHL diagnosed in adults, accounting for 30-40% of newly diagnosed cases in North America[10]. This malignancy can arise *de novo* or through the histologic transformation of other lymphoma types such as FL[11,12]. DLBCL is a clinically and genetically heterogeneous disease, where patients respond differently to frontline therapy due to inter- and intra-patient genetic variability as well as the clonal diversity of this subtype.

1.2.1. Clinical Features and Diagnosis

DLBCL can manifest in any of the primary or secondary lymphoid organs including lymph nodes, the spleen, as well as among extranodal sites[13,14]. The predominant extranodal site is the gastrointestinal tract, whereas other common areas include the stomach, central nervous system, testis, breast, mediastinum, skin, and bone, yet almost any organ can be affected[13,15–18]. Typical symptoms of DLBCL consist of enlarged lymph nodes and occasionally B symptoms in approximately 30% of patients that present as fever, loss of more than 10% body weight or excessive night sweats[17–19].

The standard diagnostic methods for DLBCL are a positron emission tomography/computed tomography (PET/CT) scan and surgical excision biopsy. Immunohistochemistry and fluorescence *in situ* hybridization (FISH) can also be performed to genetically characterize the tumour[20]. PET/CT imaging is used to determine the affected location(s) and can aid in determining the preferred site to biopsy[21]. After a biopsy is performed, the tissue's morphology is examined. DLBCL tumours have B cells that appear in a diffuse pattern with occasional areas of necrosis, although around 10% of cases display a starry sky pattern representative of high proliferation rates, a feature that is more commonly attributed to Burkitt lymphoma[17]. A large diversity of morphological variations have been reported in DLBCL including the centroblastic, immunoblastic, and anaplastic variants[17]. Morphology alone cannot be used to differentiate the subtypes of NHL and make a diagnosis. To accomplish such, the immunophenotype must be characterized. Immunohistochemistry and flow cytometry can identify NHL subtypes based on the presence of cell surface markers and protein expression levels[22]. This can be variable due to the heterogeneity of the malignancy, although the most common immunophenotypes contain surface markers of CD19,

CD20, CD22, CD79A/B, PAX5, and IgM[17,23]. Other markers can be expressed less frequently or in specific subtypes of DLBCL such as MUM1, CD10, BCL6, Ki67, MYC, TP53, and CD5[17,23]. Lastly, FISH is often used to detect high grade B-cell lymphoma, a subtype that contains double- or triple hit rearrangements involving the MYC and BCL6 or BCL2 oncogenes[24].

1.2.2. Prognostic Methods

After diagnosis, a patient's prognosis can be assessed through several tools: the International Prognostic Index (IPI), cell of origin classification, and genetic classification[25–29].

The IPI is a numerical risk score that is derived from the combination of five clinical variables: age at diagnosis, ECOG performance status, serum lactate dehydrogenase (LDH) concentration, number of extranodal sites, and disease stage[25]. This system assigns one point for each of these negative prognostic factors and categorizes patients into four groups assessing their risk of death: 0 or 1 indicates low risk, 2 equates to low intermediate risk, 3 is a high intermediate risk, and the two highest values indicate high risk[25]. Naturally, a higher IPI score correlates to a worse prognosis in general and aids with outcome prediction, but this is only a course tool. Features such as these, suggesting a patient has higher or lower risk, may inform on treatment options for patients with aggressive types of NHL but in general practice, patients all receive the same initial treatment. Leveraging the IPI to determine the risk of death or potential treatment alternatives of a patient can enhance diagnosis and in turn improve the corresponding clinical management strategy.

Gene expression profiling has allowed DLBCL to be separated into two molecular subtypes, based on gene expression patterns that are characteristic of distinct B cell differentiation stages[26]. These so-called "cell of origin" (COO) groups differ by their clinical presentation and molecular features. The germinal centre B-cell-like (GCB) subtype more closely resembles the expression patterns of normal B cells in the germinal centre. The activated B-cell-like (ABC) subtype, in contrast, contains signatures similar to those of activated plasma B cells[26]. Cases that do not fit into either group are known as unclassifiable. Patients with GCB DLBCL have a significantly longer overall survival in comparison to those with the ABC subtype [26]. Though there is not one characteristic genetic alteration that is pathognomonic in DLBCL, some of the more

common genes found to be mutated in cases regardless of COO subtype encode histone or chromatin modifiers such as ARID1A and TET2 mutations[6,30]. Frequent alterations found in the GCB DLBCL subtype include mutations affecting the histone modifiers EZH2, CREBBP, KMT2D and EP300, as well as the B cell homing regulators GNA12, GNA13, and BCL2 translocations that regulate apoptosis[31–35]. Finally, somatic mutations preferentially associated with the ABC DLBCL subtype are genes that activate the NF-kB and B-cell receptor pathways including IRF4, MYD88, CD79A, CARD11, PIM1 and TNFAIP3[33,34,36,37].

The recently developed LymphGen classifier uses genetic features to distinguish DLBCL into six genetic subgroups, with the intention of ultimately using this information to guide therapy[27,38–40]. These subgroups are defined by different combinations of recurrent genetic aberrations such as mutations, copy number aberrations and gene fusions[41,42]. The LymphGen algorithm classifies tumours based on how well they belong to one of the defined genetic subtypes, or a mixture thereof. DLBCL tumours that are classified into multiple subgroups reflects the genetic heterogeneity of the disease along with the imperfection of existing classification tools.

1.2.3. Tumour Heterogeneity and Resistance to Treatment

Tumorigenesis occurs via the acquisition of one or more somatic mutations within a non-malignant cell. These alterations provide a selective advantage typically for abnormal cell growth[43]. In aggressive NHLs, this results in uncontrolled cellular proliferation and the emergence of additional somatic driver and passenger mutations, allowing further progression of the neoplasm[44]. Sustaining malignancy requires sequential somatic mutations, resulting in clonal and sub-clonal selection among the tumour cells. These clones are subject to evolutionary selection pressure where only subclones that contain advantageous mutational profiles for tumour survival proliferate and maintain neoplastic growth[44,45]. As development progresses, the heterogeneity of the tumour increases, making extranodal involvement and treatment resistance more likely to occur[45]. DLBCL is a phenotypically, clinically, and genetically heterogeneous malignancy.

Approximately 60% of patients experience long-term benefit from the combination chemotherapy R-CHOP (rituximab plus cyclophosphamide, doxorubicin, vincristine and prednisone)[46]. Patients with DLBCL that do not respond to frontline

therapy or relapse after treatment (termed rrDLBCL) have a 2-year complete response rate of ~20% and a 5-year overall survival rate of ~50%, indicating dismal survival outcomes[47,48]. Minimal residual disease (MRD) describes when a limited number of tumour cells remain after treatment and can lead to patient relapse. DLBCL displays a high level of inter- and intrapatient heterogeneity where clonal diversity is largely associated with MRD and relapsed or refractory cases (Fig 1.1). The specific acquired mutations that contribute to treatment resistance in rrDLBCL are not well known, with recent work from our group and others indicating some examples[49–53]. Genetic alterations that may associate with rrDLBCL cases are largely in line with pathways and cellular processes that are perturbed more generally in DLBCL. These include genes involved in epigenetic regulation, immune surveillance, cell-cycle regulation, and the JAK-STAT and NF-kB signalling pathways[49,50,54]. Mutations in more limited number of specific genes have been reported to be relevant to the acquisition of treatment resistance, such as TP53, KMT2D and MS4A1[50,55]. Although the genetic landscape of DLBCL has been extensively studied and characterized, it is uncommon for patients with rrDLBCL to undergo a tissue biopsy at the time of progression, making the identification and characterization of chemotherapy-resistant subclones extremely difficult[56]. Understanding the genetic composition of a tumour that has survived frontline therapy is crucial, as the knowledge of mutations that affect treatment response may ultimately lead to changes in up-front or subsequent clinical management.

1.3. Circulating Tumour DNA is a Non-Invasive Biomarker

The current gold standard diagnostic method for lymphoma involves a tissue biopsy, where a piece (or core) of tumour tissue, often an entire lymph node, is removed from the patient. This procedure allows for an accurate histological assessment of the malignancy yet contains some major drawbacks as it cannot reflect real-time spatial and temporal heterogeneity during disease progression[57,58]. Spatial heterogeneity describes genetic diversity among different regions of the same tumour or between extranodal sites, and temporal heterogeneity reflects the genetic diversity of an individual tumour over time (Fig 1.1)[45]. Serial tissue biopsies are not typically performed in a clinical setting due to patient discomfort, high costs, inaccessible tumour location (potentially leading to tumour cell seeding and further complications), and the inability to view macroscopic tumours after treatment response[58,59].

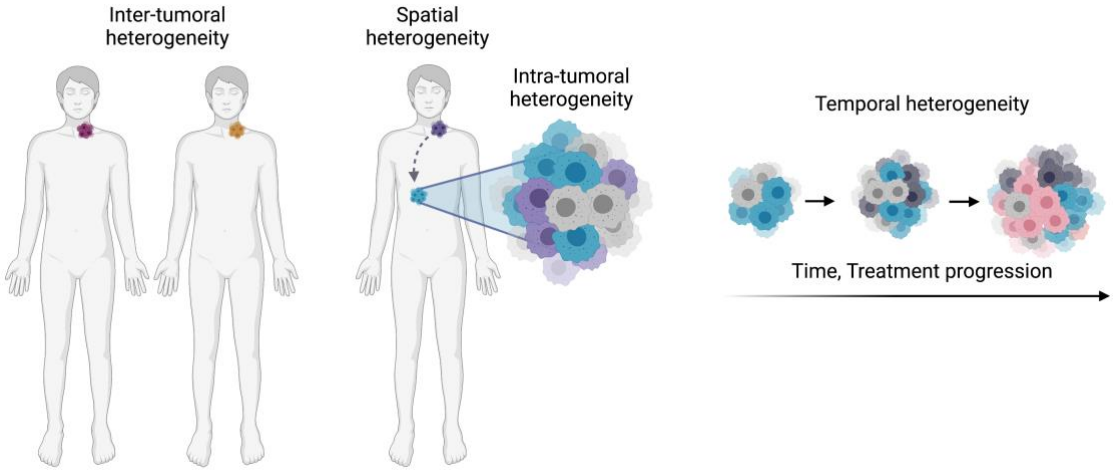

Figure 1.1. Summary depicting lymphoma tumour heterogeneity
Differently coloured cells within the tumours represent individual clones with distinct mutational profiles. This figure was created with biorender.com

A cancer biomarker is a biological molecule found in a patient that can be either tumour-derived or produced by the body when a tumour is present[60]. Many biomarkers that have been discovered and studied to date are present in various bodily fluids of a diseased patient, and can be obtained through non-invasive methods such as a blood draw or urine sample[61–64]. During tumour development, apoptotic and necrotic cells release circulating tumour cells (CTCs), tumour-derived extracellular vesicles (EVs), along with cell-free DNA (cfDNA) and RNA (cfRNA) into the bloodstream[65,66]. The detection and quantification of these components in blood or other non-tissue samples is termed as a liquid biopsy. These can broadly be utilized as a non-invasive approach for studying tumour genetics and have been shown to be useful to study clonal evolution and real-time disease burden through serial monitoring[57,67,68].

Exosomes are EVs released from the tumour via fusion in response to stress factors such as chemotherapeutics[69–71]. These can be extracted from various bodily fluids including the blood, saliva, CSF and urine[67,71–73]. Exosomes can contain tumour-derived DNA, RNA, protein and metabolites, making them a substrate for real-time early cancer detection, prognosis, disease monitoring and resistance prediction in a variety of cancer types[72,74,75]. Exosomes have been found to not only contribute to metastasis but can also aid in the prediction of metastatic location due to their integrin and oncogenic gene expression profile[72,76,77]. In DLBCL, exosome-derived nucleic acid and protein content can be leveraged for molecular analyses. These include the use of exosomal

miRNA to determine chemoresistant patients and the monitoring of relapse post-treatment[78,79]. While utilizing exosomes as a DLBCL biomarker holds potential, current studies are limited with cohort sizes and differing protocols require standardization[79,80].

In addition to using membrane-bound molecules as a biomarker for DLBCL, fragments of circulating cfDNA and cfRNA have also demonstrated clinical value. Different forms of cfRNAs are present in the blood such as messenger RNA (mRNA), micro RNA (miRNA) and long non-coding RNA (lncRNA)[81]. Increased mRNA expression of common DLBCL oncogenes and tumour suppressors has been found to positively correlate with a poor overall survival[82,83]. The majority of studies focusing on miRNA (most commonly the upregulation of miR-21 and miR-155) have consistently shown a significant association with disease survival, prognosis, and response prediction in rrDLBCL patients treated with R-CHOP[84–86]. lncRNAs are infrequently investigated for biomarker discovery in DLBCL, though have demonstrated potential as a future diagnostic and chemotherapeutic resistance evaluation tool[87]. In comparison with mRNA, encapsulated cell-free miRNA and lncRNA are known to be more resistant to ribonuclease degradation, facilitating detection and isolation[81]. Studies are limited within the cfRNA field for DLBCL and require additional validation[88].

While biomarkers that can be extracted from liquid biopsies such as exosomes and cfRNA have shown many advancements and are potential candidates for early detection, disease monitoring and resistance prediction in DLBCL, in depth biological and technological knowledge is still lacking to provide better patient outcome in a clinical setting. Alternatively, tumour-derived cfDNA, known as circulating tumour DNA (ctDNA), particularly shows great promise and rapid development as a tool to assess tumour burden, clonal evolution, response to therapies, MRD, and relapse in DLBCL[67,89].

CtDNA is released into the peripheral bloodstream from tumour cells undergoing apoptosis, necrosis and secretion[90]. This cell-free double-stranded DNA is approximately 147-167bp in length due to its correspondence to nucleosomes and chromatosomes (Fig. 1.2)[91]. Furthermore, the size of mononucleosomal ctDNA is only about 10-20bp smaller than the germline cfDNA continuously present in the bloodstream of a healthy individual[92,93]. CtDNA has been shown to reflect genetic and epigenetic alterations inclusive of simple somatic mutations (SSMs), copy number variations (CNVs), IgH gene rearrangements and methylation patterns in a multitude of cancers, making it the ideal biomarker to leverage for my project[67,94–96].

1.3.1. ctDNA Quantification for Monitoring and Prognostication

The small size difference between the cfDNA molecules from tumour cells and healthy cells along with the often exceedingly low levels of ctDNA (0.01% of the total cfDNA) in the plasma provide significant limitations for robust and accurate quantification[97,98]. The amount of ctDNA is elevated in a cancer patient and directly correlates to tumour burden[99]. With this, many studies have demonstrated that higher levels of ctDNA in the bloodstream correlate to an overall worse survival outcome, establishing its value as a prognostic biomarker in DLBCL[100–102].

To assess the value of ctDNA for rrDLBCL prediction during interim and post-treatment, ctDNA quantification is commonly examined alongside standard PET/CT imaging parameters. The metabolic tumour volume (a measure of tumour burden) is particularly valuable and highly correlates to increased ctDNA levels[103,104]. Quantifying ctDNA alongside the use of PET/CT imaging to assess disease outcome and predict relapse in DLBCL has rapidly gained popularity, as its use and improvement from the limitations of PET/CT have been validated[100,105]. Such limitations include subjecting patients to radiation, false positive or negative results, as well as the discouragement of routine surveillance imaging after a patient achieves complete response[21,106,107]. With the advancement of sensitive ctDNA quantification methods (such as NGS) in DLBCL, it is possible to predict relapse between 3 and 6 months prior to clinical presentation[108,109]. Overall, liquid biopsies have shown to be highly accurate for real-time prognosis, monitoring and relapse detection in DLBCL, and in addition to the evaluation of ctDNA through quantification and mutational profiling, studies have recently started focusing on the epigenetic landscapes and fragmentomics of ctDNA[110–112]. Liquid biopsies have begun the process of adoption into clinical practices as complimentary tool alongside the existing gold standard procedures[113]. With this, it is difficult to quantify ctDNA in a sensitive and standardized manner to broadly incorporate into the clinical setting, presently making most detection techniques suitable only as research-based tools. Additional

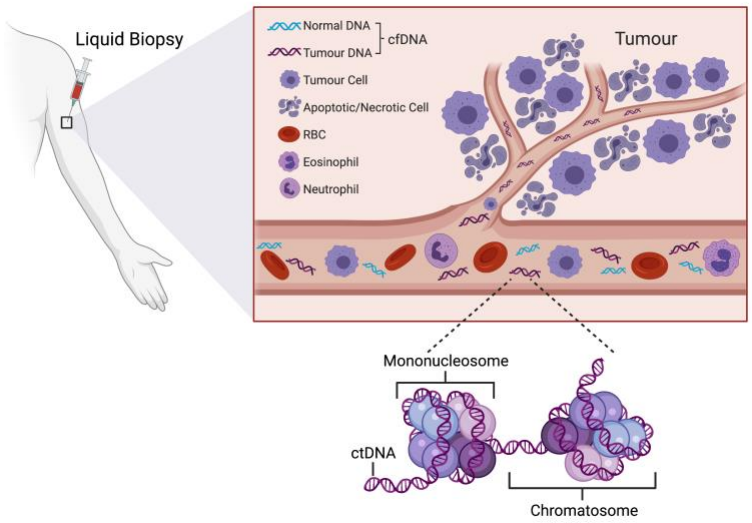

Figure 1.2. Contents of the peripheral blood near a malignant tumour
CTCs and ctDNA are shed from the tumour and circulate along with normal cfDNA and red blood cells in the bloodstream. CtDNA is wrapped around nucleosomes or chromatosomes, defining their fragmentation length. This figure was created with biorender.com.

1.4. Methods of ctDNA Quantification

Cell-free DNA extracted from a liquid biopsy contains a mixture of DNA originating from both tumour and normal cells. SSMs and CNVs can be present in both DNA derivatives, making the determination of which genomic aberrations are solely from the tumour important, as this can significantly alter how clinical decisions are made. This is called tumour purity, where the fraction of cancer cells in a liquid biopsy sample are quantified. Other terms for tumour purity when used in the context of ctDNA are tumour fraction, cellularity or ctDNA level. For example, IchorCNA and MRDetect utilize tumour fraction, whereas PurBayes utilizes tumour purity and Sequenza describes the term as tumour cellularity[114,59,115,116]. For consistency, in this thesis I use tumour purity both in the context of ctDNA and tissue biopsies.

Sensitive and accurate ctDNA quantification is important for the detection of changes in tumour burden, including treatment response and relapse. This becomes increasingly challenging with patients that have a low tumour burden (e.g. after treatment), as ctDNA is present in low abundances. Quantification and analysis approaches for ctDNA are broadly categorized into targeted and untargeted (or unselected) approaches[117,118]. Targeted ctDNA detection techniques leverage the use of gene panels or “selectors” to monitor a subset of tumour-specific mutations or structural

rearrangements and require prior knowledge of the genetic landscape for the cancer of interest[117,118]. Detecting SSMs via targeted sequencing enables ctDNA level estimation using the variant allele frequency (VAF). The VAF describes how often a specific mutation is found in a sample and is calculated by dividing the number of reads that match the variant allele by the overall coverage on that locus. Alternatively, untargeted screening methods allow the inference of CNVs and the discovery of novel genomic alterations with no prior knowledge of specific somatic mutations necessary. These assays can be expensive resulting in limited sensitivity[117,119]. Another method to estimate purity from a liquid biopsy, in addition to VAF, involves the deviation of non-diploid CNVs from the expected log2 ratio of a pure tumour.

Most studies utilize one sole method for ctDNA quantification, such as ultra-deep targeted sequencing or WGS to infer VAFs that reflect tumour burden and aid in the detection of MRD or rrDLBCL[120–123]. Various bioinformatic programs are available for the molecular analysis and estimation of purity from plasma samples. Though a seemingly simple concept, the accurate estimation of tumour purity is extremely difficult, with various inconsistencies shown between the current bioinformatic techniques[124]. Programs that estimate purity utilize either a copy number-based approach (IchorCNA, Sequenza, Battenberg, ABSOLUTE and THetA) or a variant-/B-allele frequency (VAF/BAF) -based approach (PurityEst, PurBayes, or VAF-inferral via multiplex droplet digital PCR (ddPCR) or CAPP-Seq)[114,116,125–128,115,129]. To date, three programs exist to calculate purity based on both CNV and single nucleotide variant (SNV) integration of WGS data; PyLOH, Accurity, and MRDetect[130–132]. PyLOH and Accurity both estimate tumour purity and ploidy from modelling somatic CNVs and heterozygous germline SNVs (using BAF). MRDetect on the other hand, is specifically designed for ctDNA and estimates purity using genome-wide SNVs and CNVs to detect extremely low tumour fractions in MRD[59].

1.4.1. Leveraging Targeted Sequencing to Infer Somatic Mutations

Utilizing conventional PCR-based methods to quantify ctDNA such as quantitative PCR or FAST-SeqS was initially popular due to their low costs and straightforward procedures, though their inadequate sensitivity when detecting ctDNA at low levels (for applications such as the prediction of MRD) has led to further advancements in the field[133]. Digital PCR methods such as ddPCR have an improved

ability to detect clinically relevant mutations with low VAFs (~0.1%), as this technique is highly precise and has a low false-positive rate when counting individual ctDNA molecules[118,134–137,89,138,139]. The amplification and high throughput sequencing of Ig gene rearrangements (known as IgHTS) is also utilized to quantify ctDNA in NHL[140]. This method has been shown to be highly specific and sensitive, and is frequently compared to PET/CT for the detection, prognosis, and surveillance of several NHL subtypes including DLBCL[141,142,108,140]. Other ctDNA quantification techniques utilize targeted gene panels include tagged amplicon sequencing, the safe sequencing system, and cancer personalized profiling by deep sequencing (CAPP-Seq)[120,138,143–146]. The CAPP-Seq technique has been gaining adoption due to its low error rate, high sensitivity and specificity and low input ctDNA requirements, making it the most broadly applied ctDNA quantification technique used for lymphomas[109,120,138,147–149]. CAPP-Seq leverages the use of a gene panel or “selector” to detect SNVs, insertions and deletions (indels), structural rearrangements, and somatic CNVs for individual cancer subtypes[120,149]. This method uses a pool of biotinylated DNA oligonucleotide probes to target the frequently mutated regions in the cancer of interest through hybridization-based library enrichment or “capture”. The selector probes anneal preferentially to the desired regions of the DNA library, resulting in enrichment followed by sequencing and analysis[150]. With the many benefits of this technique, it nevertheless has limitations including the inefficient capture of breakpoints that underlie structural rearrangements, the requirement for a lower detection threshold for early stage tumours, and the inability to comprehensively identify CNVs[150].

For ctDNA quantification methods that utilize SSMs, the first step after alignment of reads is the inference of somatic variants from the alignments or “variant calling”. This can be accomplished from a matched tumour sample or directly from the sequencing data from the ctDNA. Commonly used tools for this include Strelka, LoFreq, MuTect2, and SAGE[151–154]. These programs have a high sensitivity for calling low-VAF variants which is necessary during ctDNA analysis, as liquid biopsies generally have lower VAFs compared to tissue biopsy samples[155–158].

1.4.2. Detecting Copy Number Variants from Untargeted Methods

While targeted approaches are highly sensitive, specific and flexible, comprehensive/untargeted methods offer the opportunity for unbiased genetic profiling

of ctDNA on a genome-wide scale[61,138]. Multiple techniques are available to detect CNVs and genomic rearrangements without the requirement of any prior knowledge of the tumour. These include whole genome- and whole exome sequencing (WGS and WES, respectively), where the sensitivity to detect low-VAF variants is dependent on coverage depth which increases with cost[159,160,138]. Low-pass WGS (lpWGS, coverage ~0.1-1x) has emerged as a rapid, high-throughput and inexpensive alternative for ctDNA quantification in cancers such as DLBCL. This method allows the estimation of ploidy, CNVs, and ctDNA levels from liquid biopsies[114,161–163].

Somatic CNVs are large regions of the genome that are amplified or deleted during tumour development[164,165]. These genotypic alterations are present in a different number of copies from the germline or healthy state, therefore determining the number of CNVs at a locus is called “copy number calling”. Due to the expression level of a gene being correlated to its copy number, many cancer-related genes can be affected by CNVs, thereby deregulating their expression in cancer cells[164,166]. Identifying somatic copy number events have been valuable in the prognosis and treatment decision of cancers including DLBCL[164,167]. Various programs are available for this purpose, though the majority utilize data from standard WGS. Calling CNVs from lpWGS data specifically can be performed using a limited number of bioinformatic tools, with IchorCNA and WisecondorX being the most prevalent due to their explicit intended use of lpWGS data and direct applicability to liquid biopsies[168,169]. IchorCNA was developed to quantify tumour purity while accounting for ploidy and subclonality through the use of a hidden Markov model (HMM) segmentation algorithm[168]. WisecondorX infers CNVs utilizing an optimized data normalization process, followed by circular binary segmentation (CBS), and finally the assignment of aberrant segments to their respective copy number state[169]. Although both programs can adequately detect CNVs, each have unique limitations that hinder the accurate quantification of ctDNA from lpWGS input. Within our group, IchorCNA has been found to have limited resolution for focal events and consistently overfits CNV calls by wrongly assigning aberrant segments to an overly high or low copy state (e.g. if a segment is neutral or gained it is called incorrectly as amplified, or vice versa for deletions) particularly in samples containing low ctDNA levels (<10%). Despite WisecondorX having an easily customizable set of input arguments, the software package lacks the ability to calculate tumour purity, ploidy, subclonality, and often outputs biased copy number calls in samples with a considerate number of CNVs when many more amplifications are present than deletions or the opposite.

1.5. Gap in Knowledge and Project Novelty

In general, there remains a paucity of open source bioinformatic tools for analyzing ctDNA for the purpose of purity estimation. Moreover, to our knowledge, there are no existing programs that utilize both copy number information from lpWGS and the VAF of simple somatic mutations from targeted capture sequencing to estimate tumour purity from liquid biopsies. To address this unmet need, I have developed a new tool for this application, which I named PurEctDNA (Purity Estimation of circulating tumour DNA). The goal of PurEctDNA is to enable reliable tumour purity estimates for non-invasive treatment monitoring and relapse assessment of lymphoma patients with improved accuracy. To date, there is no established method of leveraging both data types to calculate tumour purity from serial sampling of liquid biopsies.

In preparation for the development of PurEctDNA, I utilized both IchorCNA and WisecondorX to estimate the copy number profiles of samples based on lpWGS data. To address the limitations of IchorCNA and WisecondorX detailed above, I improved the performance of WisecondorX through the extensive modification of the source code. Furthermore, I altered WisecondorX to generate enhanced copy number plots, as well as a new file type and tumour purity estimate that was not previously produced by the program. Improving this pipeline, in addition to developing PurEctDNA, resulted in two new methods of reliably estimating tumour purity from liquid biopsies.

Chapter 2.

Improving Genomic Aberration Detection and Tumour Purity Estimation for Liquid Biopsies

2.1. Introduction

In general, circulating cfDNA from both healthy and tumour cells are present in a liquid biopsy, although the relative representation from each source can vary through a patient's clinical course. There is a broad division of approaches for analyzing liquid biopsies to quantify the relative contribution of tumour cells (tumour purity) that rely either on 1) tracking the presence of tumour-derived simple somatic mutations (SSMs) and structural variants (SVs) or 2) estimating the levels of circulating tumour DNA (ctDNA) in the plasma (known as the tumour purity). The first approach is more amenable for the general study of tumour genetics including changes to the mutational landscape of subclonal populations changes over time. This has been accomplished through serial sampling of liquid biopsies. In theory, the eventual clinical application of such approaches may afford the opportunity to detect mutations that predict early relapse, disease progression, and even the acquisition of mutations that contribute to treatment resistance. A more common clinical application of ctDNA is to employ its quantification as a proxy to measure tumour burden, or the total volume of tumour cells in the body. At the most extreme end, the use of high-sensitivity assays allows for the detection of minimal residual disease (MRD) that have been shown capable of predicting an upcoming relapse weeks before the current gold standard tissue biopsy or PET/CT imaging techniques[108,109]. To assess the somatic single nucleotide variants (SNVs) and CNVs present in liquid biopsies, the majority of programs leverage only one of two sequencing data types for tumour purity estimation. The most common programs leverage copy number calling to determine the purity and ploidy from matched tumour and normal samples using whole genome sequencing (WGS)[114,116,125–127,123].

Inferring copy number alterations requires the normalization, segmentation, and interpretation of aberrant (non-diploid) segments into estimates of discrete copy number states, which refers to the number of copies per tumour cell. The discrete aberrant states are generally referred to as single-copy gain (3), multi-copy gain (>=4; amplification),

and either heterozygous (1) or homozygous deletion (0). This process begins with normalization, a data pre-processing step that accounts for known causes of systematic variation in the data. Common normalization processes include correction for GC content and mappability biases[169,170]. After the data are normalized, segmentation is performed, where the genome is divided into bins that are then grouped into sets of continuous bins estimated to have equal copy number. The log2 ratio is calculated for each bin, and these values are included when segmentation and plotting take place. Due to noise, the individual bins eventually assigned to a segment can have a broad range of values. Two common methods used for segmentation to infer tumour-derived copy number profiles are circular binary segmentation (CBS) and the hidden Markov model (HMM). CBS identifies regions of equal copy number using change-point detection where segments are conceptualized to be spliced at either end to form a circle[171]. Considering this, the arc spanning an individual segment is compared to the following segment, and for different copy numbers to be assigned the neighbouring segments must have statistically significantly different means from each other[171]. HMM characterizes “hidden” states through multiple probability parameters and estimates the maximum likelihood of each copy number state through the expectation-maximization algorithm[114]. Because this method includes a model that estimates the copy number for each segment directly, they have the benefit of not requiring the application of a threshold to the segmented data. After segmentation is complete, copy number profiles are analyzed. Different programs incorporate various parameters to interpret features of the malignant genome such as tumour purity, ploidy, and subclonality estimates. In this work, I use the copy number calling programs IchorCNA and WisecondorX as they both leverage lpWGS data from liquid biopsies[114,169].

Primarily, this work focuses on the application of WisecondorX as it is a relatively new tool for ctDNA application with the potential to overcome some of the limitations of IchorCNA described in Chapter 1. These limitations include IchorCNA’s requirement for manual curation of copy number profiles. IchorCNA leverages HMM for segmentation and produces a set of solutions with different log likelihoods, tumour fraction estimates, and copy number calls for each sample. The solution with the highest log likelihood appears first and the corresponding segmental information is based off this primary solution. However, this copy number profile may be suboptimal and so manual inspection of every sample is necessary, especially those with a non-diploid tumour ploidy. Usually, the optimal solution is clear to identify, as incorrect profiles will contain

many subclonal events, whole genome amplification or deletion events, or segments that are wrongly being called as neutral. This process is not necessary for WisecondorX, as the program performs aberration calling in three steps: BAM to NPZ file format conversion, reference creation, and CNV prediction[169]. Segmentation occurs via the CBS algorithm and CNV determination can be accomplished through two routes; either a default Z-score calculation or a user-definable cut-off associated with log2 ratios to separate aberrant and normal segments. When applied to the same data, WisecondorX often produces copy number profiles comparable to IchorCNA, with exceptions more common when applied to samples with a less balanced suite of alterations (e.g. many more gains than losses). With these samples, WisecondorX assigns the mean log2 ratio as zero resulting in the inaccurate assignment of copy number states. I noted that the copy neutral state is defined as the median value of all bins, which is a fundamental limitation of the software. In the most common example, this shifts the estimated copy number downward and can lead to the incorrect assignment of diploid regions as deletions. To address this issue, the log2 ratios and cut-off values require re-calculation to adjust final outputs and increase accuracy of aberration calls to ultimately improve ctDNA quantification.

Estimating tumour purity from targeted sequencing methods such as CAPP-Seq or ddPCR and the corresponding variant calling pipelines is also commonly performed, as the variant allele frequency (VAF) of somatic SNVs is directly calculated[120,135,151–154]. The VAF represents the fraction of sequencing reads that match a variant from the overall coverage at that genetic locus. The main limitation when leveraging only VAF is that tumour purity has several confounding factors including ploidy, subclonality and the effect of CNVs on VAF, which are not taken into consideration by existing tools. These factors are important to consider, as a variant's allele frequency that is calculated from the same subclonal population changes depending on the copy number and timing of the alteration.

In this chapter, I implemented a strategy to leverage both somatic variant calls from CAPP-Seq data and copy number profiles from lpWGS data as a means to estimate tumour purity with high accuracy. The purity estimation tool that I developed, called PurEctDNA (Purity Estimation of circulating tumour DNA; pronounced "pure ctDNA"), annotates cancer-specific variants to their respective copy number state and calculates the mutational VAF, ploidy and purity for each sample. It also produces the

final purity values along with a summary table for each copy-annotated variant and their relevant parameters. This program is deployed as a component of the open source R package “GAMBLR”. This is the codebase for the genomic analysis of mature B-cell lymphomas (GAMBL) project, which is introduced briefly in a subsequent section.

Here, I demonstrate the validation of PurEctDNA through two different analyses involving ground truth genome comparisons and an *in silico* dilutions. No other bioinformatic programs exist to date that estimate tumour purity of cfDNA through the same parameters as PurEctDNA, making this tool a novel method to assess the mutational and overall ctDNA level of liquid biopsies. PurEctDNA is easily customizable, as users can specify their preferred CNV caller for the input copy number profiles, an optional gene panel and other options for subsetting the data in meaningful ways. These parameters allow for a facilitated and reliable analysis of the tumour purity and mutational profile from plasma-based cfDNA samples.

2.2. Methodology

2.2.1. Cohorts

A total of 897 plasma samples from 5 clinical trials were used in this study: LY17, Obinituzumab-GDP (OZM073), Epizyme, Montreal (a cohort of rrDLBCL patients), and the BC ctDNA cohort (a group of DLBCL patients from British Columbia). lpWGS was performed on 610 samples from 594 patients and CAPP-Seq was performed on 376 samples from 249 patients (Table A.1). Plasma samples were obtained from collaborators at the Jewish General Hospital (JGH) in Montreal, Quebec and the British Columbia Cancer Agency (BCCA) in Vancouver, British Columbia.

The GAMBL project is a meta-analysis of all available genomes from mature B-cell neoplasms, mostly non-Hodgkin lymphomas with over 1300 tumours having fully analyzed genomes available including copy number and mutation profiles. For validation, I selected 495 genomes (491 from solid tissue biopsies, and 4 from liquid biopsies) from GAMBL for two separate analyses. Using the estimated purity from the Battenberg pipeline, all 495 of these were utilized to evaluate the accuracy of PurEctDNA. The subset of 4 genomes derived from ctDNA were also used for *in silico* dilution experiments. These ctDNA genomes were from the OZM073 and LY17 clinical trials.

2.2.2. Plasma Sample Processing, Sequencing and Read Alignment

All patient samples that underwent lpWGS were processed and sequenced by collaborators at the JGH (Montreal, Quebec) or the BCCA (Vancouver, BC), depending on the cohort, whereas those that underwent CAPP-Seq were processed and sequenced by a former member of the Morin lab.

cfDNA Isolation

Blood samples were either processed to separate plasma from the cellular fraction promptly after collection (<4 hours) or stored in blood collection tubes, which prevent cell lysis, (Streck, La Vista, NE, USA) and processed within 2 weeks. All plasma was separated into aliquots (typically 1-2 mL) and stored at -80°C for future extraction. Cell-free DNA was isolated using the QIAamp® circulating nucleic acid kit (Qiagen, Hilden, Germany) or the MagMAX Cell-Free DNA isolation kit (ThermoFisher Scientific, Waltham MA, USA), where samples were lysed with the Proteinase K treatment prior to magnetic bead binding with the latter kit.

Library preparation and sequencing

Library preparation for the 376 plasma samples that were sequenced via CAPP-Seq (~1000x coverage), was performed in concordance with the methodologies described in Rushton et al., 2020[172]. In short, ctDNA libraries were pooled with xGen lockdown oligonucleotide probes (Integrated DNA Technologies, Coralville, IA, USA) and custom gene capture pools, and then enriched via hybridization capture targeting a panel of 63 lymphoma-specific genes (Table A.2). Libraries were multiplexed and sequenced on either the Illumina MiSeq or NextSeq instruments, depending on the cohort.

For the 610 plasma samples with lpWGS (~0.4x coverage) performed, libraries were constructed using the xGen cfDNA & FFPE DNA Library Prep MC kit (Integrated DNA Technologies, Coralville, IA, USA). Sequencing was performed on the Illumina HiSeq and MiSeq systems.

Genomes derived from the GAMBL project were sequenced to between 40x and 80x coverage on a variety of Illumina sequencers, dependent on their age and the source of each genome. The details for the cohorts used here are provided in the papers

detailing each individual study[173,174]. All cases utilized from GAMBL had a matched normal sample (a sample of healthy tissue from the same individuals) sequenced to an average depth of 40x.

Sequencing read alignment

For the sample data obtained from targeted sequencing, all read alignment methodology described in this section was completed by a member of the Morin lab. Raw reads were aligned to the hg38 using bwa-mem[175]. Reads from lpWGS as well as the ctDNA genomes in GAMBL were aligned to grch37 using bwa-mem[175]. Picard MarkDuplicates was used to flag duplicate reads, and quality control was performed using Picard CollectWGSMetrics (http://broadinstitute.github.io/picard/). ReadCounter by HMMcopy (https://github.com/shahcompbio/hmmcopy_utils) was used to count the reads in individual 500 kb genomic windows, utilizing only reads with a mapping quality of at least 20. Copy number profiles were separately determined using both IchorCNA and WisecondorX [114,169]. Samples processed through IchorCNA were manually inspected, and the optimal fit for each sample was manually selected from the full set of solutions.

2.2.3. Determination and Improvement of Tumour-Specific CNVs and SNVs from ctDNA

Modification of the WisecondorX source code to improve aberration calls

To test the capabilities and parameter requirements for WisecondorX, I processed 14 samples from the BC ctDNA cohort. Of these samples, 10 were DLBCL, 3 were FL, and 1 was mantle cell lymphoma (MCL). As per the published WisecondorX pipeline, samples were converted from the BAM to NPZ file type, a reference was created using a set of 86 plasma samples that were identified as having minimal (<1%) ctDNA levels. I adjusted the bin size to detect CNVs using 500kb bins, as the default (5kb) produced very noisy plots, making CNVs difficult to accurately identify.

The first step to improve the biased copy number calls generated from this program (as seen in Fig. 2.1) was determining an approach to center the mean log2 ratio to a baseline that more accurately represents the segments in the diploid state such that new cut-off values for gains and losses can be determined. To accomplish this, I visualized the distribution of the log2 ratio per segment as a mixed histogram and

density plot. This allowed the visualization of each assigned copy number state, with individual density curves representing each ploidy (Fig. 2.2).

These visualizations revealed that the mean was commonly not aligned with the mode of the highest peak. To infer the individual distributions contributing to this, I used the R package Mclust (https://cran.r-project.org/web/packages/mclust/) to apply a Gaussian mixture model to the bin-level reads counts per sample. Within this model, I was able to specify the number of mixture components ("clusters") that ultimately correlated to the number of peaks on the distribution curve. Setting this cluster number to 1 and taking the mean of that cluster allowed me to determine how offset the genomic alterations were from the copy neutral state. This mean was positive if there were more gains and CNVs were shifted downwards, and negative if there were more losses and the variants were all shifted upwards in the initial plot (Figs 2.3 and A.2 show segments being moved upwards due to a positive mean being calculated). With this, I incorporated this "offset value" as a new argument into the WisecondorX source code within the main python script, where this value was added (or subtracted) to the pre-existing calculations for log2 ratios, Z-scores, and normalization weight values. These three variables were all used in the downstream code for segmentation, aberration calling and plot generation, so modifying them based on the offset value resulted in all steps within the program to be adjusted accordingly. This new argument can be specified when running my improved version of WisecondorX to fix biased copy number profiles.

After leveraging the offset value to shift segments to their correct states, there was a need to similarly re-define the cut-off values to assign segments as gains and losses. This was a pre-existing argument available in the original WisecondorX code defined as "beta", that should optimally be close to the purity. I similarly leveraged the mixture model approach to calculate beta. Here, I explicitly specified 3 clusters for the model to fit bins in each segment, given the common pattern of three main clusters representing diploid, gain and heterozygous loss (Fig. 2.2A shows the clusters and their correspondence to each copy state). Using the clusters, I determined the deviation of each from the copy neutral state. I calculated the tumour content and read depth ratio for bins that are in the gained or lost states and used these ratios to calculate the purity of the sample. This estimate was reiterated back into WisecondorX to accurately define aberration cut-off values as a loss, neutral, or gain for resulting tables and scatter plots (Fig A.2).

In addition to correcting CNV calls, I modified other features from the WisecondorX source code to improve copy number profiles and tables, allowing for more detailed analyses. Initially, WisecondorX only assigned copy number states 1, 2, and 3 to aberrant segments (i.e. discrete states of high-level gains and homozygous losses were not considered). Because the CNV state is used in the inference of purity, I enhanced the downstream analyses by adding an additional cut-off value. This was based on the standard read depth ratio calculation that assigns the appropriate segments as amplifications (copy state 4 or higher) and lists them within the resulting tables and plots. The addition of this state is utilized further in the next section.

With this modified code, I created a new plot that incorporates the density curve generated during mixture modelling to aid in the visualization of the biased and shifted segments. A key output file for representing CNV is known as a SEG file. Unfortunately, this file type was not initially produced by WisecondorX, therefore I modified the program to generate this SEG file and confirmed that the corresponding calls were assigned properly in IGV.

I applied all modifications described above to increase the accuracy of CNV detection in a standardizable manner through the development of a snakemake workflow[176]. This workflow automatically performs an initial run of the unmodified WisecondorX program defining 500 kb bins, calculates the offset mean and purity, inputs those values as arguments back into the modified version of WisecondorX, re-running the program and re-calculating the purity using the values of the adjusted segments.

Detecting Simple Somatic Mutations (SSMs)

To determine lymphoma-specific simple somatic mutations (SSMs) from liquid biopsy samples, I created a pipeline that consists of three main steps: variant calling, post-filtering, and MAF file manipulation. Many variant callers are available to accomplish the first step, whereas the latter two steps require specific criteria that is highly dependent on the application and data quality.

I implemented a snakemake workflow that runs two modules (the SAGE variant caller and vcf2maf) that have been implemented as part of the LCR-modules (https://github.com/LCR-BCCRC/lcr-modules) analytical pipeline. This pipeline required a table as input that specified the identification name for each sample along with the sequencing type, patient identifier, tissue status, and the genome build. This table was

used within the LCR-modules snakefile to run the proper samples through SAGE, and the resulting VCF files produced by SAGE were converted to MAF using the vcf2maf module. The only parameter that differed from the default whilst running SAGE was setting the validation stringency to lenient to allow variants with low read support to be detected. The output of vcf2maf is a tab-delimited table from each sample containing one line per variant using the standard mutation annotation file (MAF) format implemented for The Cancer Genome Atlas (TCGA). Importantly, these files contain the read counts for reference and alternate allele.

After variants were identified and VCFs were converted to MAFs, I performed post-filtering on these MAFs to remove artifacts using a custom python script previously written by a member of the Morin lab. This accounted for strand bias, mapping quality, mismatch bias, and depth filtering, while also removing common sequencing errors and germline variants that might have been present after somatic variant calling was performed. I subset the filtered outputs to the capture space, meaning the mutations called were only kept within the MAF file if they were within the region of interest as defined by the experiment. Next, I identified the subset of samples that contained data from both lpWGS and CAPP-Seq to view their mutational profiles.

2.2.4. PurEctDNA

Goals and implementation

To improve tumour purity estimation for liquid biopsies by leveraging information from both lpWGS and CAPP-Seq data, I developed a novel bioinformatic program called PurEctDNA. The input for this tool is a SEG and a MAF file as input, respectively from any common CNV caller and SSM calling/annotation pipeline. Notably, this tool can also be applied to low-pass WGS data or deeply sequenced genomes but was specifically designed to estimate tumour purity from cfDNA samples that have undergone targeted sequencing and lpWGS. To develop PurEctDNA, I utilized the segment information from IchorCNA and my modified version of WisecondorX along with the post-filtered MAF files curated from SAGE.

PurEctDNA is implemented in R and utilizes dplyr, data.table and custom functions to assign the absolute copy number to each mutation, ultimately performing mutation-level corrections to allow estimation of individual cancer cell fraction and overall estimation of tumour purity. This program utilizes these variants and their annotated

copy number, along with the tumour variant reference- and alternate allele counts, to calculate the purity (Eq. 1) and VAF (Eq. 2). Within PurEctDNA, mutations assigned as copy number state 1 or 2 and those assigned as copy number state 3 and above are analyzed separately. The mean of these grouped mutational purities is calculated, and the final purity is determined by taking the mean of all annotated variants from that sample. If the final purity is over 1 (100%), PurEctDNA estimates that value as 1, as no sample can contain more ctDNA than cfDNA.

Equation 1:

$$\text{Purity} = \frac{\text{Copy Number} * \text{VAF}}{\text{Ploidy}}$$

Equation 2:

$$\text{VAF} = \frac{\text{t\_alt\_count}}{\text{t\_ref\_count} + \text{t\_alt\_count}}$$

PurEctDNA produces a purity estimate and summary table displaying all copy number annotated variants along with their respective VAF, ploidy, and mutational purity estimates. Two plots are also optionally generated from the program. These are both histograms, showing the distribution of VAFs and purity estimates present in each copy number state within the selected sample.

I incorporated the R script where these calculations take place into GAMBLR and developed a command line workflow that runs multiple samples through the function efficiently.

Validation analyses

I performed two different methods of validation for PurEctDNA, where I primarily tested genomes from tissue biopsies that had accurate purity estimates determined with high concordance between at least two separate pipelines. I compared these estimates to PurEctDNA by evaluating correlation between methods across all samples. I further explored the performance across a range of purities using an *in silico* dilution approach.

The first validation incorporated samples containing sequenced genomes from solid tissue biopsies that were previously analyzed in GAMBL. I selected samples containing WGS data where purity estimates contained a 5% concordance between two bioinformatic tools that perform copy number calling and tumour purity estimation

specifically for WGS data (Battenberg and Sequenza[116,125]). These 495 samples were considered to have highly accurate purity values and were therefore used a ground truth to compare against PurEctDNA estimates. Copy number profiles (SEG files) used in this analysis were produced by Battenberg. Different parameters available for PurEctDNA were leveraged in this validation to test how the tumour purity estimates change when they are utilized. A Pearson correlation value was calculated for each comparison.

The second validation entailed the generation of *in silico* dilutions of tumour samples from liquid biopsies that have undergone WGS. Four samples were used in this analysis from patients with DLBCL, as these were the only genomes that were accessible from GAMBL with deeper sequencing done. To generate these dilutions, I determined the fraction of reads from the tumour (F_t) and matched normal (F_n) samples, taking coverage and purity into account using the following equation:

Equation 3:

$$F_t = \frac{P_f * C_n * F_n}{C_t\,(P_i - P_f)}$$

Where F_t is the fraction of the tumour sample required, F_n is the fraction of the normal sample required, C_t is the coverage of the tumour, C_n denotes the coverage of the normal, P_i is the initial purity of the tumour sample, and P_f is the final tumour purity. For each of the four ctDNA genomes utilized in this analysis, purity estimates were down sampled in increments of 5%, starting from 5% and leading up to 5% below the initial purity value (Table A.3). For example, if a sample has an initial purity estimate of 40%, I diluted the sample to values of 5%, 10%, 15%, 20%, 25%, 30% and 35% using reads from the matched normal genome.

After the necessary amount of tumour and normal sample required for the dilution was determined, I used samtools to combine the tumour and normal reads, creating one indexed cram file for each dilution. Next, I used a pipeline within GAMBL to dilute the cram files and return these in MAF format. To test PurEctDNA, I used the diluted MAFs and corresponding SEG files generated from Battenberg as inputs to determine if the resulting tumour purity was accurate. As a last validation measure, I ran

the diluted samples through Battenberg using the LCR module's pipeline to determine how well this program estimated the down-sampled purity compared to PurEctDNA.

2.3. Results

2.3.1. Modified WisecondorX Improves ctDNA Copy Number Inference

WisecondorX occasionally generates inaccurate CNV calls

As described in Chapter 1, liquid biopsy samples that are processed through WisecondorX are known to have biased results when a significant amount of either gained or lost aberrations are called (Fig. 2.1). This is due to the underlying calculation that WisecondorX performs, where the mean of the log2 ratio is set as zero (copy neutral) and all copy number calls are shifted in the opposite direction, leading to an inaccurate output. When examining the results from this program on their own, it is not always clear that the copy number profiles are wrong. To overcome this, I compared each of the original ctDNA samples' CNV calls to IchorCNA, allowing for the direct detection of any genome-wide or focal errors.

In addition to analyzing copy number profiles, I plotted the log2 ratio values of the individual bins from the entire sample as a density plot to assess the distribution of bins across each copy number state. Ideally, three curves should be seen. The largest curve peaking at zero (corresponding to copy-neutral segments in diploid tumours) and deleted or gained regions should be less common (showing smaller density curves on either side of the largest neutral curve) however this was not always the case (Fig. 2.2). Of the 14 ctDNA samples initially analyzed through this pipeline, only 3 required a substantial shift while the remaining 11 required little to no shift as their neutral peak was centered very close to zero.

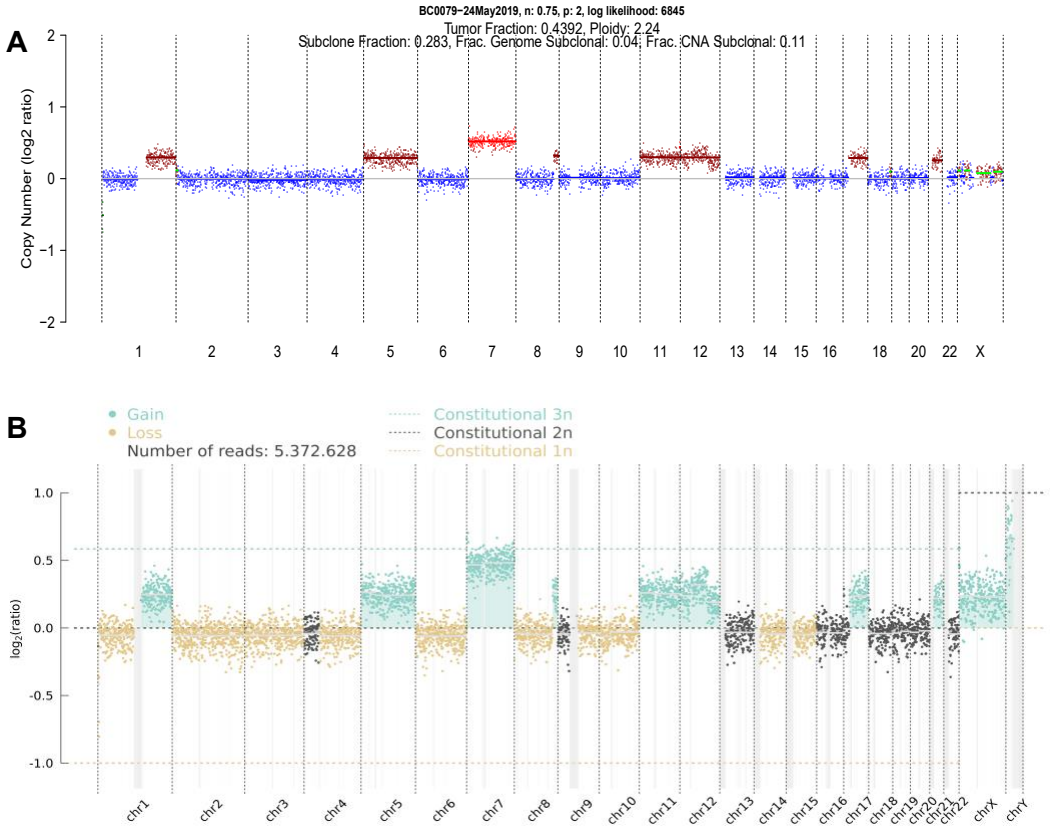

Figure 2.1. Initial copy number profiles from IchorCNA and WisecondorX
A higher number of gained and amplified regions are present in this sample, causing WisecondorX to offset genome-wide CNVs downward. Most segments that are copy neutral from IchorCNA (A) are identified as deletions from WisecondorX (B). This representative sample illustrates the effect of unbalanced amounts of gains and losses can lead to an inaccurate analysis of the tumour molecular profile.

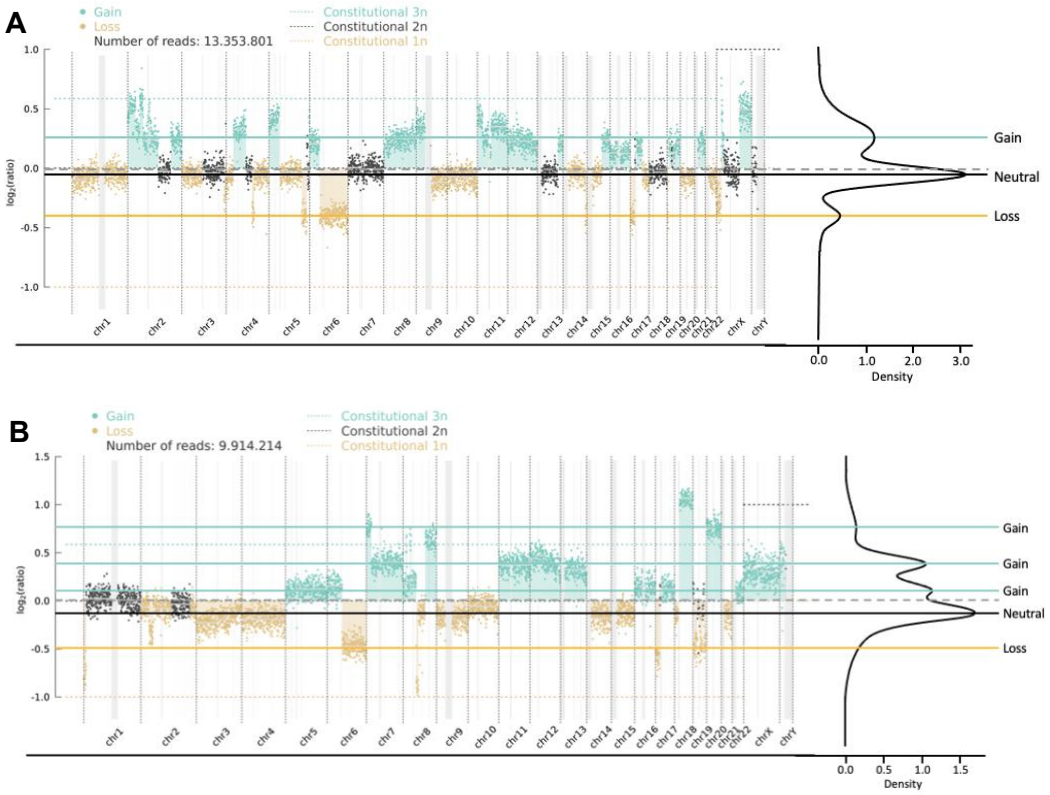

Figure 2.2. Inferring copy number state from local density of log2 ratio values
(A) The ideal three peaks are present at the expected density levels, as the tumour is diploid with the majority of bins assigned as neutral, and less common deleted and gained regions below and above the neutral curve, respectively. (B) More than three peaks are present due to a surplus of gained and amplified regions in the tumour. Both plots contain more gained CNVs than deletions, resulting in biased copy number profiles.

Correcting biased aberration calls and estimating tumour purity

To overcome some of the inaccuracies in CNV profiles inferred by WisecondorX, I applied a Gaussian mixture model and calculated the offset of the mean of all alterations relative to the copy neutral state such that segments could be shifted to the appropriate position (see Methods). This resulted in segments being at an accurate log2 ratio, yet the aberration calls were still incorrect. When bins are shifted, the beta argument must be set, otherwise the appropriate copy number state will not be assigned to the new segment location due to WisecondorX using Z-scores as a default instead of log2 ratios. Therefore, I consistently ran clustering to force the mixture model to infer 1 or 3 clusters from the data respectively to calculate the proper offset and beta values. These parameters were chosen because they were observed to produce the most consistent fit and best calculations across all samples. I found that higher positive or

negative offset values correlated to a larger deviation away from the neutral state and thus represent how much the log2 ratio must be shifted. For example, a sample with an over-representation of gains and amplifications would automatically be shifted downwards (segments will contain lower log2 ratio values) by the original WisecondorX program due to an over-estimation of the baseline for the neutral/diploid state. This is caused by the use of all bins to calculate the median, including those representing regions affected by gains and losses (Fig. 2.3).

To identify copy number profiles in need of improvement from WisecondorX, I used IchorCNA profiles from the same samples with low ctDNA levels. Calling aberrations in samples with low purity is important because an accurate copy number profile is required to accurately quantify ctDNA levels. This applies to clinically relevant scenarios where patients commonly have low ctDNA, for example the detection of MRD or assessing a patient's response during and after treatment. Owing to this, bioinformatic programs must provide this high sensitivity for an accurate estimate that can then be interpreted clinically. IchorCNA claims to reliably estimate purity in liquid biopsies with values as low as 3%[168], yet members of our group have found this to be untrue. Aberrations called at a purity lower than 10% are not as reliable, due to significant overfitting that has been seen in these sample types. This is partially explained by IchorCNA's method of segmentation and how it requires extensive manual curation of the resulting plots to ascertain the appropriate CNV profiles. WisecondorX, on the other hand, does not have this issue and performs aberration calling very well with samples below 10% tumour purity (Fig. 2.4). Adding a prior purity estimate to WisecondorX allows for improved analysis of the copy number profiles and largely correlates to the estimate produced from IchorCNA.

Feature extension and standardization of the new WisecondorX pipeline

Alongside the modification of WisecondorX to improve ctDNA copy number inference, I updated the plots and tables that are produced and standardized the new snakemake pipeline to make the program more user-friendly. To produce outputs more consistent and directly comparable with the outputs of IchorCNA, the program was modified to consider a new CNV gain class (amplifications) and a colour encoding consistent with IchorCNA (Fig. 2.5).

SEG files list the loci of each segment with its corresponding copy number event and absolute copy number, flags any loss of heterozygosity (LOH) events, and lists the log2 ratio values. Changing the source code of WisecondorX to produce a SEG file has many benefits, as this file type is useful for analyzing aberration calls in the integrative genomics viewer (IGV) platform between serial plasma samples, tissue biopsy samples, and for comparing CNVs from multiple bioinformatic programs simultaneously (Fig. 2.6). Leveraging SEG files in IGV from diagnostic tumour and serial plasma samples allows for the analysis of tumour progression and treatment response. This type of analysis is ideal for tumour biopsy samples containing a high sequencing coverage and plasma samples with a high level of ctDNA, as CNVs are called more reliably in these cases and a high correlation between CNVs present in both sample types can be seen (Fig. 2.6).

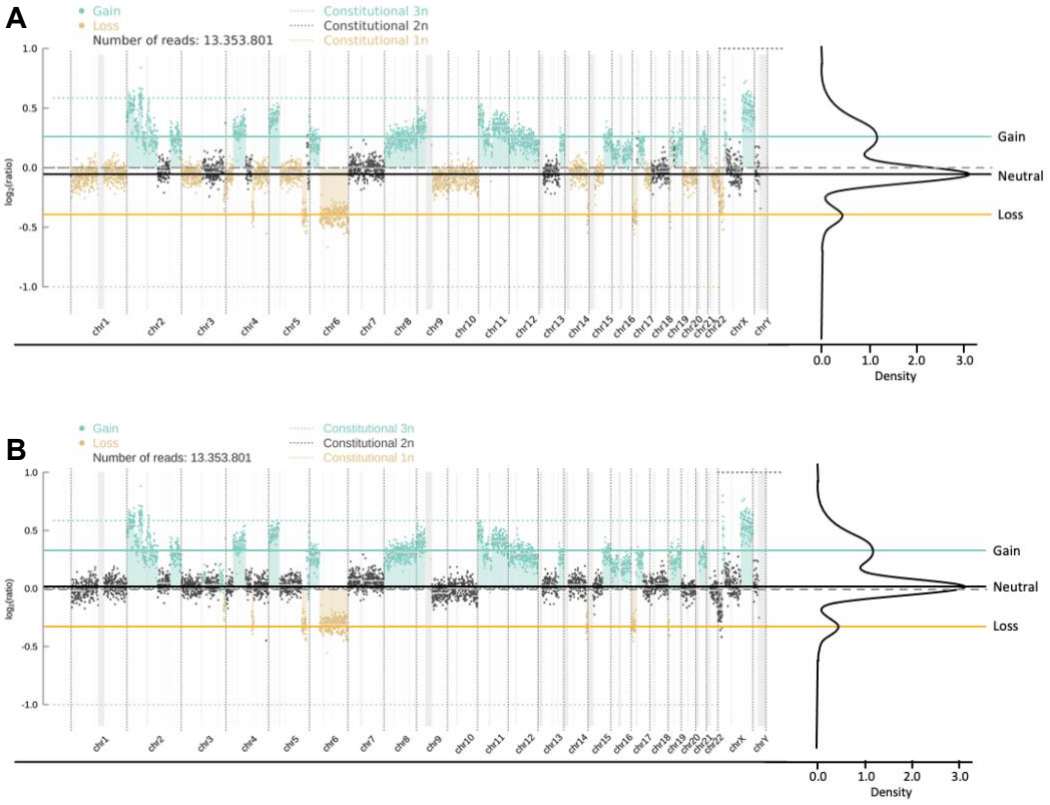

Figure 2.3. Implementing the offset value and purity back into WisecondorX sucessfully adjusts segments

(A) A copy number profile from the original WisecondorX program, showing offset segments. This results in the neutral peak of the density plot on the right to be a negative log2 ratio value and aberrations are called inaccurately for this cfDNA sample. (B) After segments are shifted, the neutral peak of the density plot on the right is centered at zero and CNVs are in turn appropriately assigned.

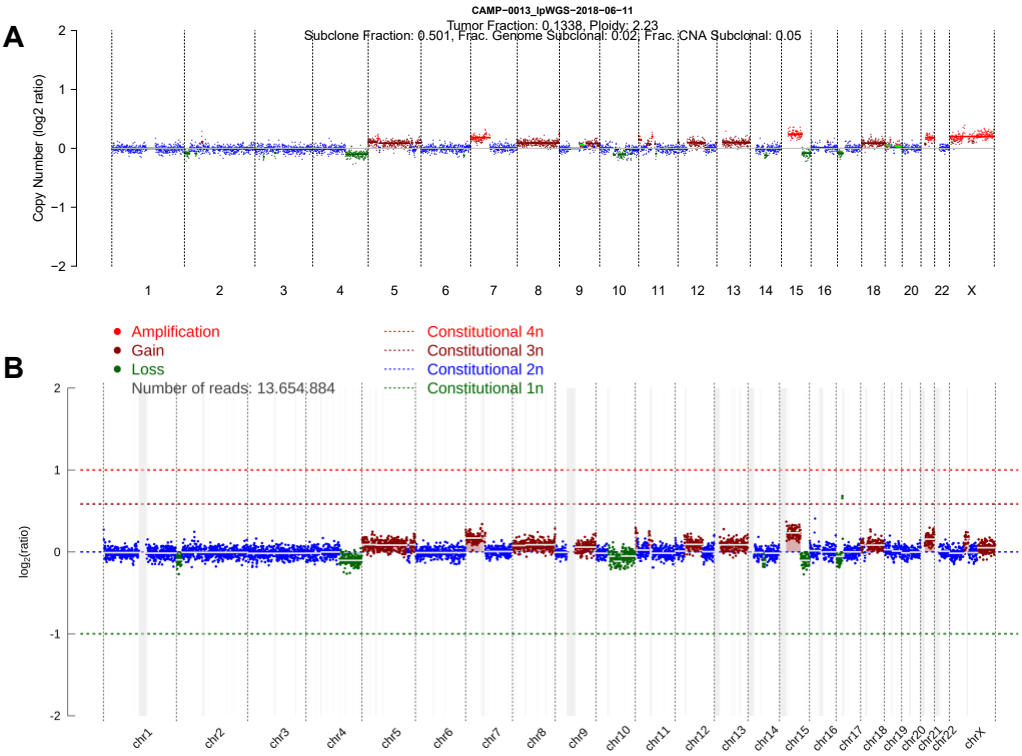

Figure 2.4. IchorCNA overfits CNV calls for samples with low ctDNA levels
(A) Copy number profile from IchorCNA, a tumour purity of 13% was estimated for this sample. Many segments are assigned as amplifications and horizontal lines where bins assigned to segments through HMM can be clearly viewed, specifically on chromosomes 7p and X. (B) Copy number profile from the improved WisecondorX program. No amplifications are called, and beta was specified as the same 13% to match the IchorCNA tumour purity. This direct comparison shows the difference in segmentation and aberration cutoff values between IchorCNA and WisecondorX, as well as a slight difference in focal aberration calls between WisecondorX and IchorCNA at low ctDNA levels.

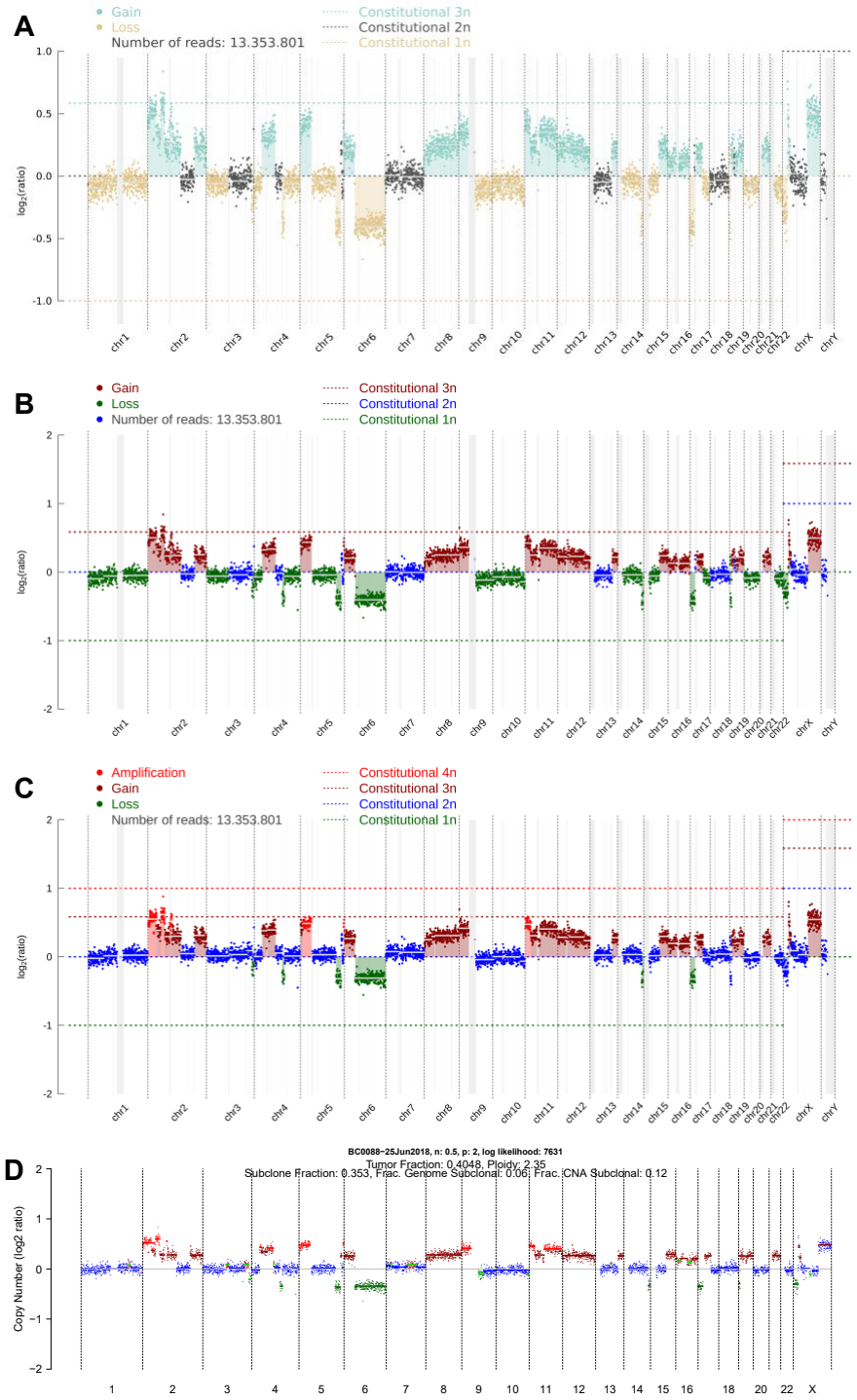

Figure 2.5. Representative copy number profile from original and modified versions of WisecondorX

(A) The original copy number profile produced by WisecondorX prior to any modification. (B) The same offset genome after adjusting the source code, prior to the use of beta. Tumour purity estimate = 42.5%. (C) After incorporating beta and adjusting aberration calls, these now correspond to the profiles determined by IchorCNA. Amplifications are also displayed. Tumour purity estimate = 40.3%. (D) The corresponding IchorCNA copy number profile, with an estimated purity of 40.5%.

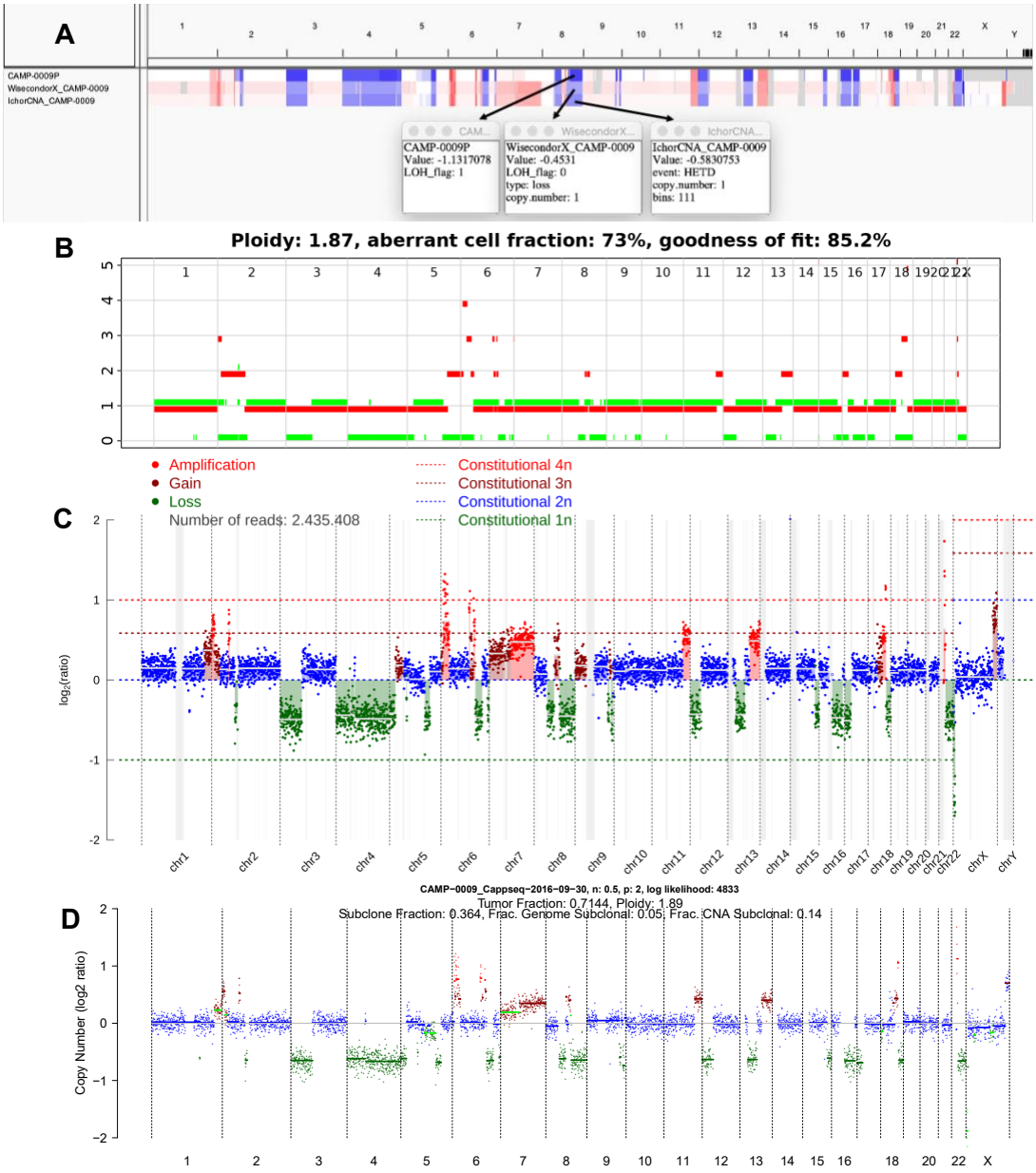

Figure 2.6. Examining CNVs in IGV facilitates aberration comparison between tissue biopsies and plasma samples

(A) An IGV representation of genome-wide CNV calls from Battenberg-derived SEG files for a tissue biopsy sequenced at high depth (top row) along with the corresponding liquid biopsy samples produced by WisecondorX (middle) and IchorCNA (bottom row). CNVs determined between these programs can differ between the tumour and plasma-derived genomes, with the largest discrepancy in aberration calls for this sample located on chromosome 7. Copy number profiles are shown of the tumour genome from Battenberg (B) and of the liquid biopsy samples taken on the same date from WisecondorX (C) and IchorCNA (D).

2.3.2. SNV analysis to infer tumour genetics and improve ctDNA quantification

In this analysis, the prevalence of common lymphoma-specific variants was determined, allowing the VAF to be calculated within PurEctDNA to ultimately contribute to cfDNA purity estimation. Detecting SSMs and their corresponding VAF from deep targeted sequencing techniques such as CAPP-Seq enables ctDNA quantification and is especially helpful in the identification of low VAF variants that are common to liquid biopsies. Here, SSMs were identified from 376 samples with CAPP-Seq data using the SAGE program and a custom post-filtering script (Table A.1). Of these, 89 ctDNA samples had sequencing data from both CAPP-Seq and lpWGS and thus were used in the implementation and evaluation of PurEctDNA. In addition to utilizing the VAF in the development of PurEctDNA, detecting the prevalence of SSMs aids in the inferral of tumour specific molecular profiles such as the heterogeneity and clonal evolution of a tumour, and can therefore help assess a patient's response to therapy and chemoresistance.

Variant calls were subset to the set of 63 lymphoma-related that were specifically targeted in the hybridization capture panel. These genes were selected for this panel based on their recurrent mutation in relapsed and refractory cases of DLBCL and include genetic and epigenetic modifiers of common pathways that lead to lymphoid tumorigenesis (Table A.2). The top 3 most commonly mutated genes among all samples were TP53 (39%), KMT2D (39%), and CREBBP (22%), which is consistent with the known frequency of mutations in rrDLBCL and other B-cell lymphomas (Fig. 2.7). These tumour suppressor genes are typically found mutated in approximately 20-25%, 20-30%, and 5-25% of aggressive B cell lymphoma subtypes, respectively[177–179]. With this, I also found the expected occurrence of commonly mutated genes in rrDLBCL including B2M, TNFRSF14, EZH2 and ARID1A from the samples analyzed among all 5 clinical cohorts. Here, the top three variants remained the same yet were present in elevated numbers, potentially due to these samples containing a higher level of ctDNA and are thereby more likely to have recurrently mutated genes identified than the samples where capture was not performed.

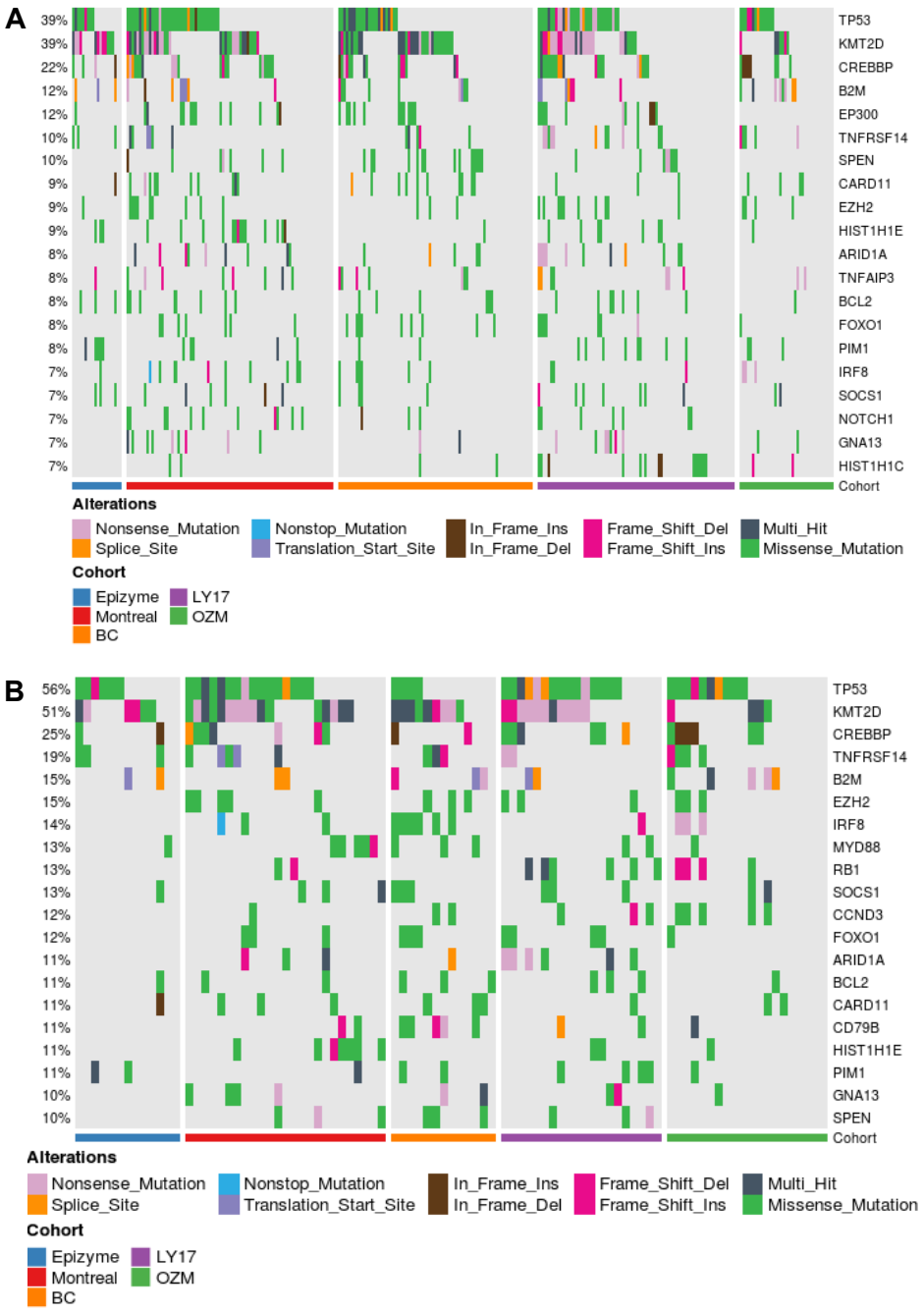

Figure 2.7. Mutations in rrDLBCL genes observed at the expected frequency
The above oncoplots show somatic variants identified from 376 samples from 5 cohorts (A), and 89 samples (a subset of the 376) that contain data from both lpWGS and CAPP-Seq techniques (B). The subset of cases with matched lpWGS, which were used in the development of PurEctDNA, have a similar frequency of mutations in these genes relative to the full cohort.

2.3.3. Analysis and Validation of PurEctDNA Purity Estimates

PurEctDNA parameter description and initial analysis using cfDNA samples

PurEctDNA estimates the level of ctDNA in a plasma sample (or tumour purity) based on the relationship between absolute copy number, VAF and ploidy of the individual mutation called from CAPP-Seq and lpWGS. In implementing PurEctDNA, I provided the user with options to provide their own MAF and SEG file or, if the user has access to the Morin labs GAMBL repository, directly refer to samples within this dataset using the sample identifier. Providing custom copy number information is optional, as these calls may not always be available or reliable depending on the tumour ploidy and the protocol used for sample collection and sequencing. I therefore customized PurEctDNA to estimate purity in either scenario using all the data provided. If a SEG file is not provided, the somatic variants must still be assigned a copy number state for purity to be calculated, which is assumed to be diploid at every position otherwise. Users also have the option to only include coding genes from the input MAF file, along with subsetting the somatic variant calls to a cancer-specific gene panel. This panel can (but is not limited to) include the same genes used in the panel from when targeted sequencing was performed and is input as a BED file into PurEctDNA. Lastly, as described in Methods, two histograms are optionally produced to visualize the calculated VAF and purity for the distribution of mutations within the sample. These are faceted based on copy number state.

As an initial comparison of purity estimates between two copy number tools, I analyzed 63 samples with somatic variant calls from CAPP-Seq and corresponding SEG files from my modified version of WisecondorX and IchorCNA. These estimates showed a high positive Pearson correlation ($r = 0.965$, slope = 0.96), indicating that PurEctDNA can robustly estimate tumour purity using copy number information from different tools (Fig. 2.8).

Comparisons were not only performed between programs, but also between changes in purity estimates when a SEG file is given and when it is not (Fig. 2.9). A high positive Pearson correlation ($r > 0.9$, slope = 1) was found for both CNV callers when comparing estimates from when a SEG is utilized or not. This shows that copy number alterations affect how purity is calculated, as estimates differ from 0-20% when using copy number information from IchorCNA compared to when variants are assigned to the copy neutral state. WisecondorX displays a much lower range here, where the purity

estimates differ less than 10% from when a SEG is used and when all variants are assumed diploid.

To assess the inferred CNVs and purity estimates produced from PurEctDNA when the copy number profile was provided from WisecondorX or IchorCNA, I calculated the percent of the genome that was altered ("PGA") between the two programs and correlated this PGA value to purity from PurEctDNA. Here, many similarities were seen between WisecondorX and IchorCNA for the fraction of CNVs called (Fig. 2.10). Various outliers were also present, where more CNVs were detected by one of the algorithms compared to the other. With this analysis, neither PGA values from WisecondorX nor from IchorCNA represent "true purity", therefore samples were grouped into thirds depending on their purity estimate given from PurEctDNA when all variants were assumed diploid. The values produced by PurEctDNA when copy number profiles were utilized from WisecondorX and IchorCNA were also defined, allowing for the comparison between purity estimates from these copy number callers and their correspondence to the grouped "truth" estimates (Fig. 2.10A & D, B & E, C & F). This showed frequent similarities between estimates when CNV profiles from IchorCNA and WisecondorX were used, though samples with purity estimates that differ from the "truth" values were also observed.

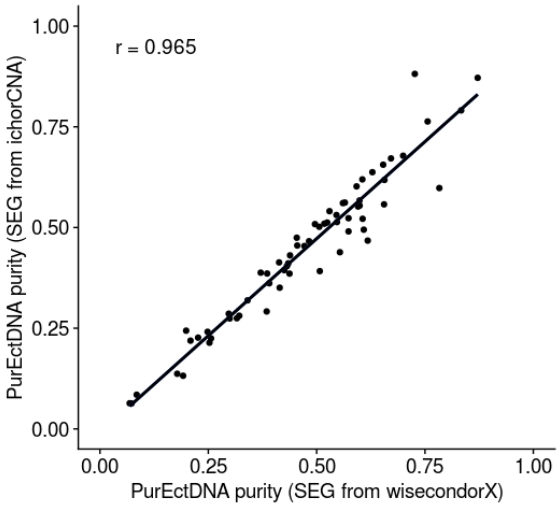

Figure 2.8. Purity estimates when leveraging the custom copy number profiles from IchorCNA and WisecondorX are highly correlated
Scatterplot showing purity estimates from the same cfDNA samples, comparing copy number information from IchorCNA and WisecondorX. The slope of the linear relationship is 0.96, indicating a strong positive correlation.

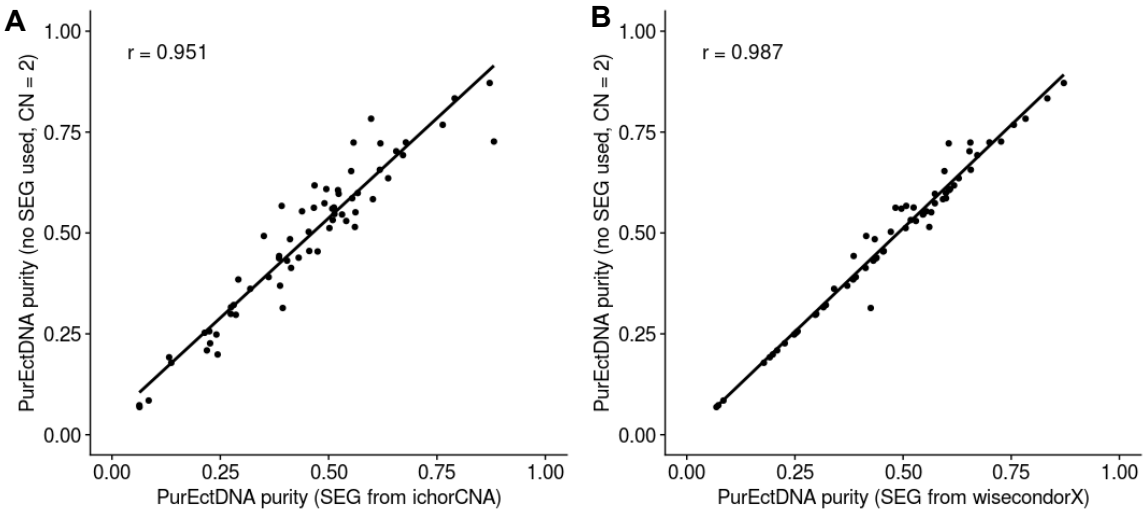

Figure 2.9. Purity estimates differ between IchorCNA and WisecondorX when aberrations are derived from the copy number callers or all are assigned as diploid by PurEctDNA

A higher Pearson correlation (r = 0.987, slope = 1) is found for WisecondorX (compared to 0.951 and a slope of 0.99 for IchorCNA), showing that the tumour purity estimated with specific copy number profiles from IchorCNA's SEG file differs from when all copy states are assigned as neutral more than the estimates from WisecondorX.

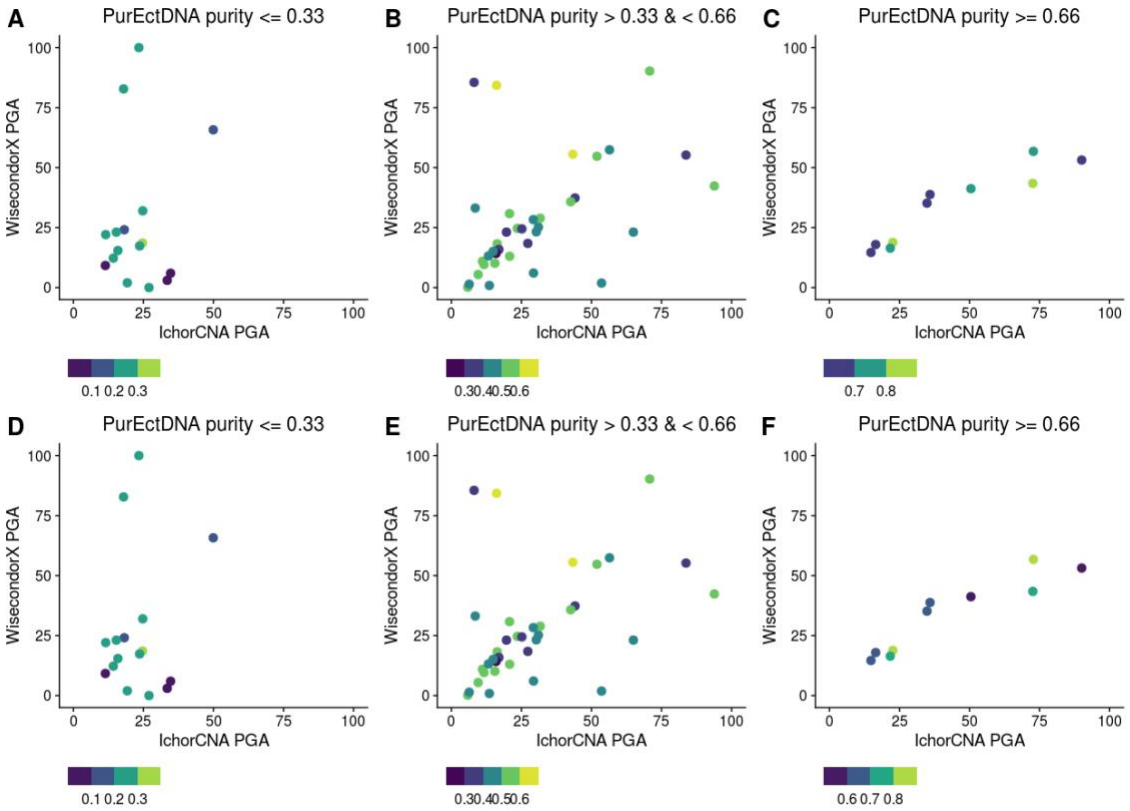

Figure 2.10. The percent of the genome altered (PGA) between WisecondorX and IchorCNA is representative of differences between the programs and how well they infer CNVs in each range of ctDNA levels

Scatterplots showing the distribution of 63 samples containing different levels of CNVs called (the PGA) by WisecondorX and IchorCNA. When PGA values on the x and y axes correspond, the amount that the genome is altered from the diploid state is equivalent between both programs. When they do not correspond, one algorithm is assigning more or less CNVs than the other. Samples are grouped into three sections, by their purity estimate given by PurEctDNA when variants are assumed diploid; A & C, B & E, and C & F. Samples are coloured by their purity estimates given when the copy number profiles from WisecondorX (A-C) or IchorCNA (D-F) are used in PurEctDNA.

GAMBL genome validation

The previous analysis has various limitations that precluded the ability to draw firm conclusions about the performance of PurEctDNA. Primarily, this related to the lack of a high-confidence gold standard measurement of ctDNA level for each sample. Using the data analyzed in the GAMBL project, I identified genomes with robust measurements of purity by selecting cases with values having nearly identical tumour purity estimates between the Battenberg and Sequenza programs. These were considered as “ground truth” to use as a comparison against PurEctDNA (Fig. 2.11). I estimated purity in each of these using PurEctDNA and found a high Pearson correlation between the ground truth values and those from PurEctDNA (r = 0.88, Fig. 2.12). A small number of outliers

are present when examining the purity values from PurEctDNA in this validation analysis. To determine the potential factors influencing these discrepancies, I separately explored the estimates for samples within four ranges of ploidy estimates. This was done separately using the ploidy estimation inferred by Sequenza and the estimate given by Battenberg. These values represent the average copy number of the genome. These comparisons show that many purity estimates that are discordant between PurEctDNA and one of these algorithms are in tumours with more extreme ploidy estimates (Figs 2.12-2.14, C, D).

These analyses provided PurEctDNA with all somatic variants identified genome-wide, which is more than is typically available in a CAPP-Seq experiment. To simulate this scenario, I restricted the mutations to the regions covered by our sequencing panel, which represents 63 lymphoma-related genes. I then compared the estimates to ground truth tumour purity. Here, a lower Pearson correlation was observed, showing that the accuracy of this approach is diminished when fewer mutations are available ($r = 0.71$ and $r = 0.88$, respectively, Fig. 2.13). An increased number of outliers were present below the line of best fit, where PurEctDNA underestimated the tumour purity for several samples. Faceted plots to examine ploidy states were created, again explaining some of the outliers.

I performed a similar analysis to determine how purity estimates differ when PurEctDNA assumes all regions are diploid. Interestingly, this resulted in a very similar Pearson correlation, presumably due to the reduction in the number of extreme outliers observed among genomes with more extreme ploidy estimates (Fig. 2.14). Generally, values were calculated at a higher mutational- and overall purity estimate when all copy number states were set to 2, due to deleted variants shifting the mean VAF upwards.

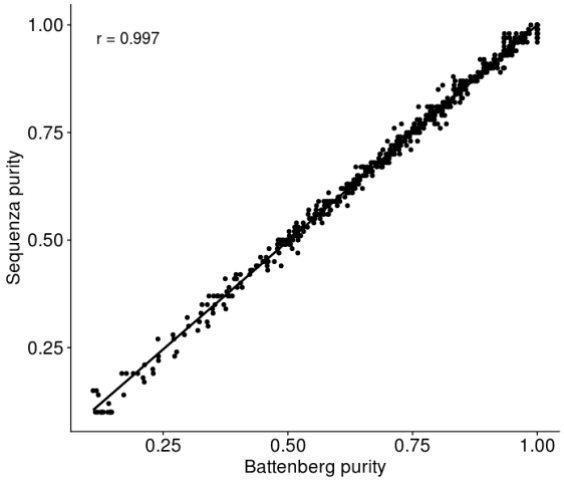

Figure 2.11. Samples obtained from the GAMBL project with a 5% concordance between the Battenberg and Sequenza

A total of 495 samples are included in this analysis. Purity estimates range from 0.10 to 1.00 (10-100%), with a smaller fraction of samples that contain a purity estimate of 0.50 (50%). These samples are highly positively correlated (Pearson r = 0.997, slope = 1), enabling them to be labelled as ground truth in this validation.

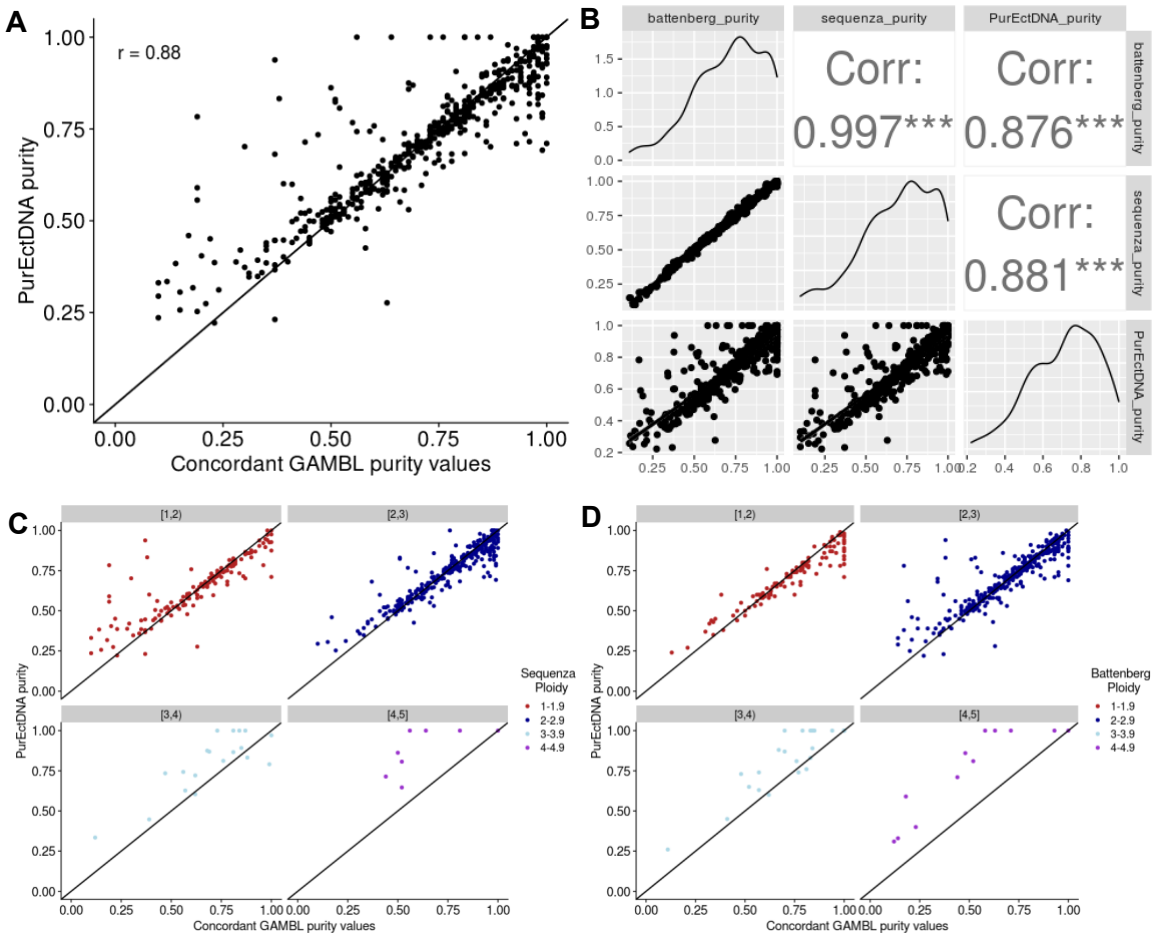

Figure 2.12. PurEctDNA accurately estimates purity from tumour tissue samples
(A) Scatterplot showing a high linear correlation between PurEctDNA purity and ground truth estimates (Pearson r = 0.88). Here, concordant purity values are from Sequenza. (B) Correlation between PurEctDNA, Battenberg, and Sequenza shows high Pearson correlation coefficients, with PurEctDNA and Battenberg displaying the lowest correlation (r = 0.876). (C) Dividing samples by their ploidy as determined by Sequenza showing that cases having more extreme ploidy estimates (1 and 4) exhibit lower correlation. (D) In contrast, separating samples by their ploidies assigned by Battenberg shows more outliers in the diploid state. Tumour samples with a genome-wide ploidy of 3 or above show overestimated tumour purity values by PurEctDNA, according to both Battenberg and Sequenza.

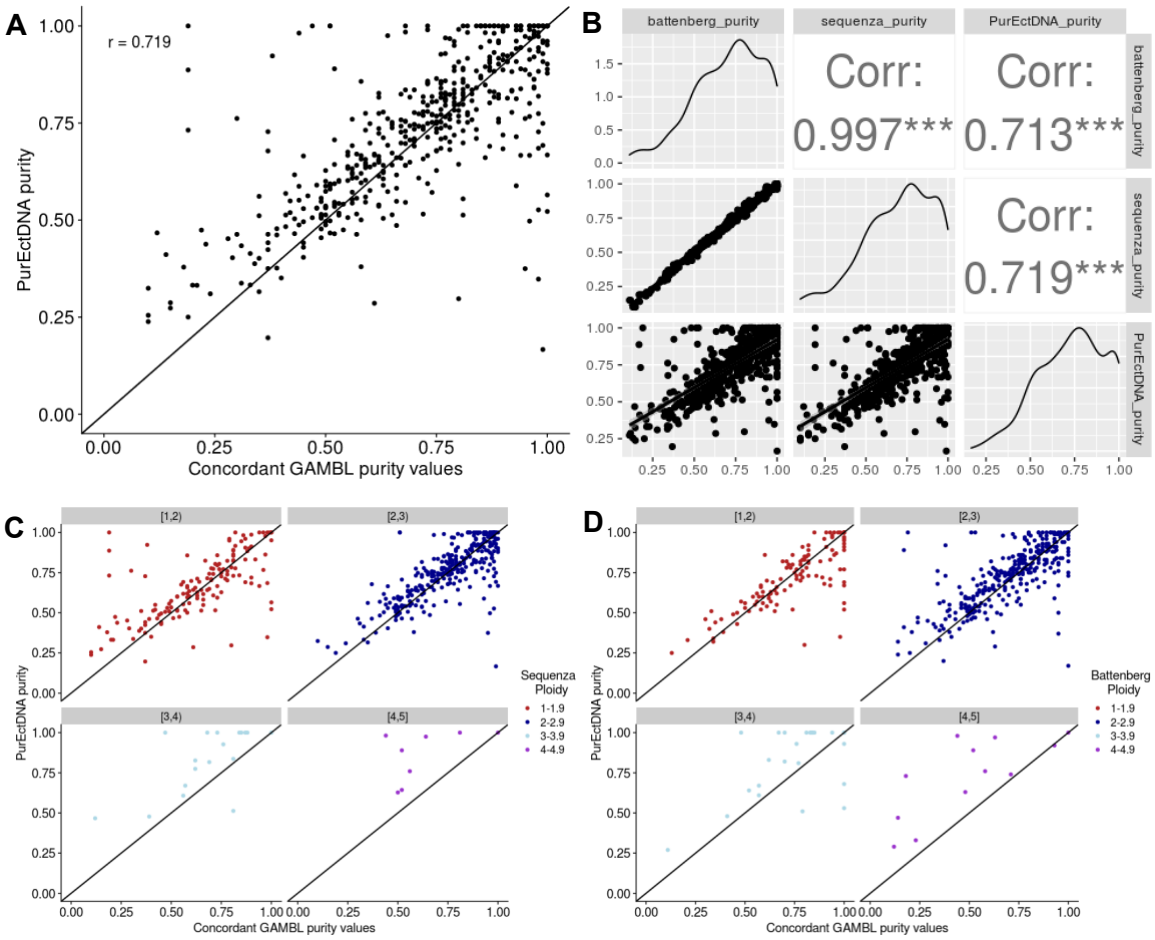

Figure 2.13. Tumour purity estimates are more dispersed when a limited set of somatic variants are used

(A) The correlation between ground truth estimates from Sequenza and PurEctDNA purity estimates derived from only somatic variants in the region targeted by a lymphoma gene panel is shown. (B) Three correlations amongst PurEctDNA, Battenberg, and Sequenza. Pearson correlation coefficients are displayed, with PurEctDNA and Battenberg containing the lowest value (r = 0.713). (C) Determining the dispersion of samples per ploidy state estimated by Sequenza shows outliers in samples with more extreme ploidies. (D) Separating based on Battenberg ploidy estimates shows a larger dispersion of outliers at extreme ploidy states compared to those seen in Figure 2.12.

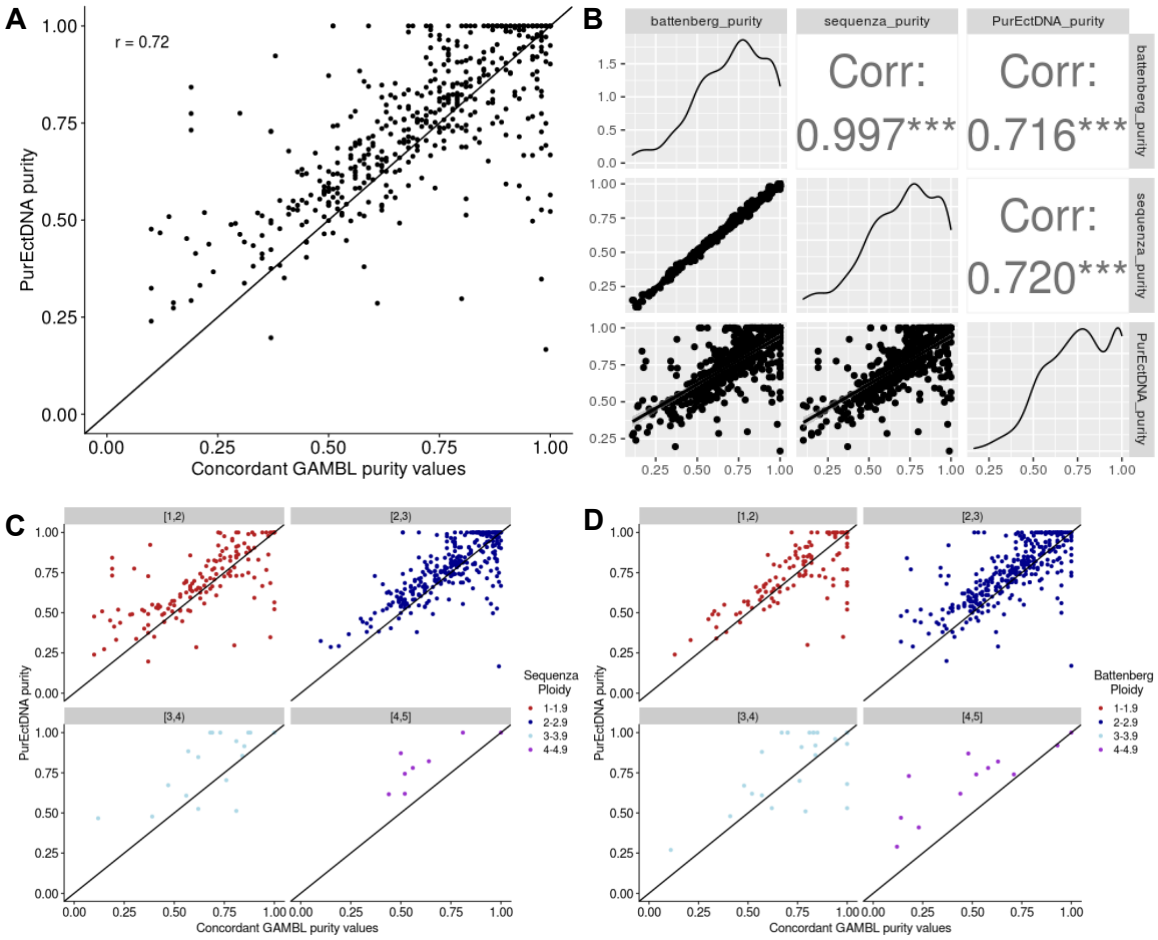

Figure 2.14. PurEctDNA purity values are not altered when variants are subset to genes of interest and assigned as copy neutral

(A) A moderate linear correlation between PurEctDNA purity and ground truth estimates from Sequenza is shown. A slope of 1 was manually added to better view how points differ from this ideal linear fit. (B) The correlation between PurEctDNA, Battenberg, and Sequenza show the lowest correlation (r = 0.716) between PurEctDNA and Battenberg, as seen in figures 2.12 and 2.13 as well. (C) Samples are grouped by their ploidy as assigned by Sequenza, where most outliers are contained in samples with more extreme ploidies. (D) Samples are grouped by their respective ploidies assigned by Battenberg with a lower number of outliers found at extreme ploidy states compared to when CNV profiles are utilized.

In silico dilution validation

Four ctDNA genomes with tumour and normal WGS data from the GAMBL project were down sampled to test the validity of PurEctDNA, as well as the programs accuracy at pre-determined high and low ctDNA levels. A total of 43 dilutions were generated *in silico* and purity estimates were produced by PurEctDNA and Battenberg. PurEctDNA estimates tumour purity at a very high accuracy for all ctDNA levels, as demonstrated by a Pearson correlation to the dilutions of 0.968 (Fig. 2.15A). Battenberg also estimates purity extremely well on samples equivalent to or above 15% ctDNA,

although fails to accurately estimate values below this as 6/8 (75%) of the dilutions with generated purity values of 5% and 10% were calculated to be above 95% (Fig. 2.15B). Due to the dilution calculation not utilizing purity given from the pre-determined ground truth estimates (from Sequenza and Battenberg), an overestimation of PurEctDNA's purity estimates was seen for two of the four samples (Fig. 2.15A). Owing to this, I utilized purity values given directly from Battenberg as a second measure of PurEctDNA's accuracy (Fig. 2.15C). This use of diluted ground truth purity estimates further validated PurEctDNA and increased the Pearson correlation (r = 0.986) of the program as an accurate purity estimation tool.

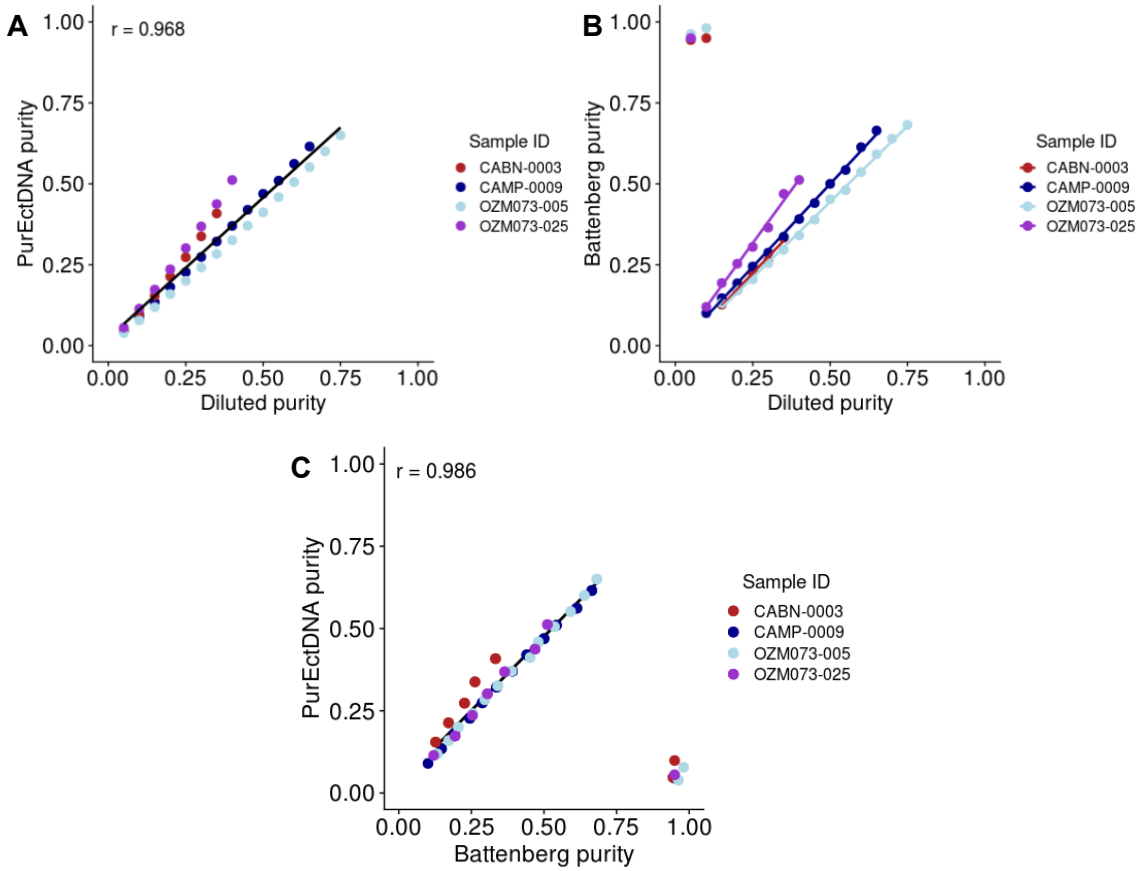

Figure 2.15. PurEctDNA estimates tumour purity from *in silico* dilutions with high accuracy

(A) A linear correlation for the *in silico* diluted ctDNA genomes with coverage ranging from 27-37x. A high Pearson correlation value (r = 0.968) demonstrates that PurEctDNA estimates tumour purity with high accuracy, down to 5% purity. (B) A linear correlation comparing purity calculated from the dilutions to Battenberg values. A high correlation is seen when examining sample purity estimates above 15%, yet the program is limited with purity values below this. (C) A high positive correlation between PurEctDNA purity estimates and ground truth estimates from Battenberg. Outliers (5% and 10%) from Battenberg (B) are included.

Chapter 3.

Discussion, Limitations and Future Directions for the Purity Estimation of cfDNA

3.1. Discussion and Conclusions

Determining the proportion of cfDNA in a plasma sample that originates from tumour cells is crucial for assessing tumour dynamics, heterogeneity, response to treatment, and the detection of relapsed or refractory cases of various cancers. Tumour purity from both tissue and liquid biopsy sample is rarely confidently known and difficult to accurately determine. This is readily apparent given that many of the available tools do not give concordant estimates. Available programs for inferring tumour purity from sequencing data commonly rely on the variant allele frequency (VAF), or the B-allele frequency (BAF) and the log ratio of somatic copy number variations (CNVs). The available bioinformatic tools for estimating the purity from a liquid biopsy are scarce. At the outset of this project there were no programs available that estimate the level of ctDNA in a plasma sample using data from both targeted sequencing and lpWGS. Therefore, I developed PurEctDNA, a novel program that utilizes somatic variants and the copy number profile of plasma samples and estimates the mutational and overall tumour purity. In addition to the creation of this tool, I modified the copy number caller WisecondorX to improve the detection of accurate CNV profiles and incorporated a purity calculation that did not previously exist.

The improvements to WisecondorX utilized the results from IchorCNA as ground truth because that tool was specifically developed to identify CNVs on ctDNA from lpWGS data. This program is used often in our group but has known limitations when applied to samples with a tumour purity below approximately 10-15%. In samples with lower purity, we observed that IchorCNA overfits CNV calls to a model with an implausible combination of ploidy and subclonal CNVs (Fig. 2.4). From direct comparisons, WisecondorX produced superior aberration calls in these conditions, qualifying itself as a suitable program to use alongside or as a replacement for IchorCNA to analyze CNV events from liquid biopsies. Despite this benefit of WisecondorX, the software originally lacked the ability to calculate tumour fraction, ploidy, subclonality, and

often output biased copy number calls when samples contain an imbalance of copy number gains and losses. I modified the source code to address some of these limitations and implement new features into the program for improved ctDNA quantification.

Initially, a Gaussian mixture model was used to manually determine cut-off values and allow the inference of the segments most likely representing regions that were diploid, gained and lost. I determined a suitable number of mixture model clusters that were necessary to center the data allowing more accurate genome-wide detection of diploid and non-diploid regions of the genome (Fig. A.1). These new ctDNA level estimates are frequently similar to IchorCNA, although have been found to differ in cases where many CNVs are present, as ploidy is not taken into consideration (Fig. 2.10). Ultimately, given the shortcomings of each tool that were identified in these analyses, it is preferrable that another program be used in conjunction with WisecondorX, particularly when purity estimation is a priority and when a cfDNA sample is not diploid. The modified WisecondorX software now infers CNV profiles more reliably than IchorCNA in samples with low ctDNA levels. Also, due to differences in segmentation methods (HMM *vs* CBS), WisecondorX occasionally assigns more focal events that may not be detected by IchorCNA. Such events may be of utility if the user is seeking CNVs of diagnostic or prognostic relevance. Moreover, WisecondorX appears more robust than IchorCNA in the samples that are overfit by the IchorCNA model (Fig. 2.4). Ultimately, adjusting how WisecondorX calls CNVs, along with the modification to the visual outputs and the addition of amplifications has significantly improved the copy number caller's functionality and ability to accurately analyze ctDNA from lpWGS data.

In its current implementation, PurEctDNA has eight customizable options for the user to apply this tool to different combinations of inputs and to perform different computations. The only required input is a MAF file containing somatic variants and reference/alternate allele read counts from either genome-wide or targeted sequencing. One option is to allow the variants to be subset to a user-defined set of regions, in my case this was genes of relevance to DLBCL. This work shows that use of this feature can lead to occasional underestimation or overestimation of the final tumour purity, resulting in a lower concordance with the ground truth estimate. Although this is a limitation, the use of gene panels is a typical use case in ctDNA sequencing. In many cases, the number of variants is significantly reduced, causing purity estimates to further

deviate from their accurate value. This is also seen when only coding mutations (rather than lymphoma-associated mutations) are specified, due to the reduction of data. These experiments highlight a caveat for users of CAPP-Seq, namely that samples with a low number of somatic variants will, in general, have less accurate tumour purity/ctDNA level estimates.

When a copy number profile is not provided to PurEctDNA, the estimated tumour purity is derived solely from VAF information. In cases where tumour purity or ctDNA levels are high and a significant number of genomic alterations are present in the ctDNA, copy number profiles will differ significantly from the neutral state. As such, if copy number is assumed to be diploid through PurEctDNA, the VAF of variants found in deleted regions would be proportionally higher as a function of tumour purity, which would systematically and erroneously increase tumour purity estimation for the sample. Similarly, mutations on chromosomes with increased copy number will have higher VAF (if on the gained copy) or lower VAF otherwise. Samples with genomes that appear to have minimal copy number alterations can often be explained by low purity, which can limit our sensitivity to detect CNVs accurately. This seems to be more pronounced with WisecondorX. For example, in Figure 2.9 the samples with lower purity are more strongly correlated between purity estimates when the use of a copy number profile from IchorCNA (A) or WisecondorX (B) is toggled.

Leveraging the percentage of the genome that is altered from the normal diploid state (called the PGA) showed a frequent correlation to the fraction of CNVs observed between WisecondorX and IchorCNA (Fig. 2.10). Here, a number of outliers were present, with the most extreme cases found in the lower (PurEctDNA purity below 33%) and upper (PurEctDNA purity above 66%) groupings of purity. I further examined some of the largest outliers, including one sample with a purity estimate of 31% where the PGA of WisecondorX was 100% and that of IchorCNA was 23% (Fig. A.3). This specific sample contained a copy number profile where the full genome was amplified, which is clearly incorrect as seen when compared to the IchorCNA profiles and that from the unadjusted plot. Examination of those case led to the discovery of a bug in the WisecondorX snakefile, thus fixing this error. Another outlier that is present from Figure 2.10C and 2.10F represents a similar error, yet for the opposite case (Fig. A.3). This sample with a 72% estimated purity contained a PGA from WisecondorX of 53% and a PGA from IchorCNA of 90%. Here, WisecondorX assigned the genome as diploid, where

IchorCNA inferred it to be a triploid tumour. In cases such as this, it is difficult to determine which algorithms' CNV calls are correct, as the BAF information or a tissue biopsy sample from a similar timeframe as the plasma sample would be necessary to discern the ploidy.

While originally obtaining genomes that have undergone WGS from the GAMBL project with purity values within a 5% concordance, samples were chosen from Sequenza, Battenberg and IchorCNA. Only 33 genomes contained this concordance, whereas 495 were found when the IchorCNA estimate was not considered. To explore this further, I compared the IchorCNA purity estimate among samples with a high concordance between Battenberg and Sequenza. This demonstrated that IchorCNA had a tendency to underestimate purity in cases with purity above 50% according to Sequenza and Battenberg (Fig. A.4). In contrast, when these same comparisons are made against PurEctDNA there is a high concordance. Also, comparison of the PurEctDNA tumour purity estimates to IchorCNA estimates revealed the same pattern of purity underestimation by IchorCNA among high-purity samples (Fig. A.4). The number of cases with highly discrepant underestimates of purity by IchorCNA is striking. It should be noted that IchorCNA generates multiple copy number profiles with different fits of its model and these need to be manually curated to identify the best fit. This curation was not done for these samples. Also, IchorCNA was developed for low-pass genome data from ctDNA and the samples used in this validation had high sequencing depth. Because of this unexplained discrepancy I excluded the IchorCNA purity estimates from this GAMBL genome validation analysis and instead relied on concordant purity estimations from Sequenza and Battenberg as ground truth to compare against PurEctDNA validates the tool's accuracy to estimate purity from genomes with tumour fractions mainly above 50%. I had technical constraint preventing the ability to adequately measure the performance of purity estimates below 50% due to the bias towards high purity tumours with WGS data. This is due to the genomes being derived from solid tissue biopsies and not liquid biopsies and the preferential use of high purity tumours for most WGS experiments. Also, Sequenza and Battenberg are specifically designed to call CNVs and estimate tumour purity values from genomes that have undergone deeper WGS methods, unlike genomes from lpWGS that have a very low coverage. Therefore, estimates below 50% from Sequenza and Battenberg are less reliable for this analysis (Fig. 2.12A), hence why more scattered purity values are seen from PurEctDNA.

Despite a high correlation between PurEctDNA purity estimates and the ground truth estimates, there are still various outliers. Separation of cases into ranges of ploidy estimates from either Battenberg or Sequenza allows some of these discrepancies to be explained. PurEctDNA assigns ploidy to the copy-annotated mutations in each sample depending on two factors: the copy number state of that region and the prior estimate of local purity. When a variant is assigned as deleted or neutral, PurEctDNA automatically sets the mutational ploidy to 1, which is the only possible ploidy in a heterozygous deletion, as copy neutral LOH events are not incorporated into the PurEctDNA calculation. A prior purity estimate is calculated for these lost or neutral copy number cases by calculating the mean of all mutational purities with these classifications. For variants with a copy number of 3 or above (representing gained and amplified structural variants), PurEctDNA considers all possible ploidy states and calculates the individual purity for each of these, taking the ploidy producing the value closest to the prior estimate of purity (based on diploid regions). Owing to how PurEctDNA assigns ploidy and omits copy neutral LOH events, this can explain the limited number of higher purity estimates seen from haploid or diploid tumour genomes. Within the GAMBL genome validation analysis, it was found that PurEctDNA occasionally overestimates tumour purity when the sample ploidy is higher. This could be due to the outliers containing insufficient read depth therefore affecting the VAF, or when samples contain a low number of somatic variants, resulting in an overly high or low purity estimation.

After performing the *in silico* dilution analysis on four ctDNA genomes, I found that PurEctDNA estimates purity with high accuracy across a lower range. The true purity of these was eventually found to be incorrectly calculated and this was not corrected. To account for this, I determined the correlation between purity estimated from PurEctDNA and Battenberg as values from Battenberg are the best estimate of purity available at this time due to their concordance to ground truth (Fig. 2.15C). Running Battenberg on these samples provides a good estimate of purity across most of the range with the exception of the sample with the highest dilution. Interestingly, this confirmed that Battenberg can also over-fit in some situations, leading to striking over-overestimates of purity in samples with low purity (here, ~15%). In contrast, PurEctDNA performed well down to 5% when correlated to the *in silico* dilutions and 15% with Battenberg dilution purity values, demonstrating its capabilities to correctly estimate purity from samples at all ctDNA levels.

In conclusion, I have developed two improved methods for tumour CNV detection and purity calculation to address an unmet need for ctDNA quantification. Both utilize lpWGS data as input, which is beneficial as it is a cost effective and high-throughput method for the analysis of structural variants in liquid biopsies. I modified WisecondorX to more accurately infer aberrant regions of the genome and improved its visualization and outputs, thus allowing more accurate calculation of tumour purity by using these outputs in PurEctDNA. I created a snakemake workflow that automatically incorporates these adjustments as a standardization measure for the future use of this improved program. This aspect of my project is beneficial to researchers who analyze ctDNA data using lpWGS and are looking for an improved copy number calling method from IchorCNA, which obviates the required manual curation of IchorCNA results. Tumour purity is extremely difficult to correctly estimate and does not often correlate between programs. PurEctDNA is a tool that leverages CAPP-Seq and lpWGS data to accurately estimate purity from liquid biopsies. This program has the potential to benefit researchers in the assessment of ctDNA to track tumour burden, treatment response, and patient relapse using plasma-derived sequencing data.

3.2. Limitations

While WisecondorX has been modified for improved ctDNA quantification, the program still has notable limitations and missing features. Every copy number calling tool that estimates purity utilizes different parameters and produces copy number profiles with a different extent of information. Most of these programs incorporate ploidy, purity and maximum likelihood estimates, along with clonal and subclonal copy number inferral within their algorithms and resulting files for tumour samples that contain over 30x coverage and a matched normal[116,125,127]. As described in Chapter 1, a limited number of tools currently exist to infer copy number and purity from ctDNA from lpWGS data. These consist mainly of the IchorCNA and WisecondorX pipelines. IchorCNA integrates 33 optional parameters and produces estimates for tumour purity, ploidy, and subclonal status, data before and after the application of normalization metrics, and a set of copy number profile solutions from different fits of a modal, each with an individual log-likelihood that does not always relate to the most biologically sensible fit of the data[168]. IchorCNA also calls homozygous deletions, losses, copy neutral events, gains, amplifications, and subclonal events. Alternately, the original WisecondorX contains 19

optional parameters and generates estimates for a copy-profile abnormality score as well as segmental, bin-wise, and aberration-based log2 ratios and Z-scores for each sample[169]. WisecondorX only assigns losses, copy neutral events, and gains during copy number profiling. I added support for amplifications while modifying the source code. These events can only be inferred when the argument beta is used after shifting. I could not include amplifications without beta since Z-scores are used in WisecondorX as the default for aberration calling. Due to Z-scores leveraging the standard deviation within their calculations and cut-off values, there was no way to accurately implement this calculation to infer amplifications. This can only be accomplished when beta is used to leverage the log2 ratio where a standardized calculation involving the read depth ratio can be applied to add such events. Homozygous deletions or subclonal events were also not included in the modified WisecondorX. Ideally, subclonal events are important to consider when performing CNV calling and purity estimation, as ctDNA has been shown to describe both inter- and intra-tumoral heterogeneity compared to the gold standard tissue biopsy. Having the ability to detect these events is extremely beneficial for research purposes as well as clinical interpretations of a patient's tumour progression, relapse assessment, and treatment decisions. Considering this, the lack of subclonality inferral and ploidy estimation in WisecondorX is a significant limitation.

PurEctDNA appears to be an accurate tumour purity estimation method that leverages both SNV and CNV data from cfDNA samples. This program incorporates all confounding factors when estimating tumour purity apart from subclonality. When reliable somatic variant calls and copy number profiles are used as input for PurEctDNA, the resulting mutational and overall purity estimate is highly accurate on a large range of ctDNA levels (as low as 5%). Owing to this, PurEctDNA is dependent on these inputs and estimates purity. If the somatic mutations are not subset to the genes of interest after variant calling and post-filtering is performed, PurEctDNA contains two parameters that can aid with this: allowing the sample to be subset to a cancer-specific gene panel, and filtering to coding-only mutations. Otherwise, the quality of ctDNA variants within the program are dependent on the variant calling pipeline used. The same situation is relevant with copy number calls, where the reliability of the absolute copy number for each locus are dependent on the bioinformatic program used. However, if these are incorrect, PurEctDNA can label all variants as copy neutral. This is occasionally a larger issue for programs such as IchorCNA, where we have relied on the copy number profile with the highest log likelihood estimate, which is not always the most accurate. For rarer

cases where the solution is not correct (e.g. a non-diploid tumour with extreme CNV events), this causes downstream issues with purity estimates from PurEctDNA.

The *in silico* dilution validation analysis demonstrated that PurEctDNA can estimate purity with a high level of accuracy, yet was limited in its sample size and number of dilutions performed. Only five ctDNA genomes were available in GAMBL for this analysis, with one excluded from the analysis as it did not contain a 5% concordance between Battenberg and Sequenza. Because of this, no ground truth purity estimate was available, making the resulting purity estimation unreliable. The remaining four genomes were diluted by only 5% increments as storage space quickly became limited due to the large size of these genomes. Samples were only diluted down to 5% purity, as the ~30x coverage restricted the number of reads that supported the variants. Owing to this, testing the limits of sensitivity that PurEctDNA can attain was not possible at this time. This can be addressed as a future experiment, as CAPP-Seq data is available on these samples, and this higher coverage can be utilized for further dilutions *in silico* to determine the sensitivity of PurEctDNA below 5%.

3.3. Future Directions

A limited number of samples were available for the initial ctDNA analysis and *in silico* dilution validation of PurEctDNA. Using larger cohorts would be ideal for examining the lowest tumour purity value that PurEctDNA can accurately estimate and determining the sensitivity of the program. Discovering the lowest detection limit can aid in the extension of PurEctDNA to estimate purity after treatment or during low tumour burden to detect relapse or MRD. As described in Chapter 1, the only other existing tool that is able to detect extremely low levels of ctDNA is called MRDetect. This program also leverages SNVs and CNVs yet does not utilize targeted sequencing data or lpWGS, and has been developed for MRD detection from WGS data (~35x coverage)[132]. A future study that can be performed is comparing PurEctDNA to MRDetect using WGS and lpWGS data, as well as samples at a larger range of purity values (0.01-100%).

As a gold standard ctDNA quantification method does not currently exist, no ground truth estimates were available for this project. However, the use of IgHTS to track clonal Ig V(D)J gene rearrangements from ctDNA has been shown as a highly sensitive and specific technique for disease detection and surveillance in cancers including NHL and could therefore be utilized as a method to compare purity estimates

to PurEctDNA. Due to the unresolved differences seen between the current mutational and copy number approaches for ctDNA quantification, IgHTS affords the opportunity to contribute to the determination and resolution of such discrepancies and ultimately advance the development of a gold standard measurement of ctDNA in DLBCL. Owing to this, performing IgHTS on the samples utilized in this project and comparing them to values produced from PurEctDNA is a future study that could be done to assess the accuracy of the estimated ctDNA levels.

Further testing on how well PurEctDNA estimates purity on cfDNA samples from serial time points with the corresponding clinical data would be desirable, as this would allow the program to have a more clinical application rather than more research based as it currently stands. With this, I would incorporate a patient-level summary with new plots and tables describing how ctDNA-derived tumour purity changes over the course of treatment, response, and post-therapy monitoring. Ideally, the ctDNA levels would be compared with clinical variables measured by imaging techniques or other complementary approaches for quantifying or detecting MRD.

As PurEctDNA is a newly developed program, there are a variety of extensions that can be added to increase performance. Although the existing parameters allow for adjustments regarding somatic SNV and CNV calls, they can be expanded to improve precision when serial samples are being analyzed. Currently, PurEctDNA estimates purity for one sample at a time and does not recognize the relationship between samples from the same patient. PurEctDNA can be modified to leverage information from samples from same patient, identify matching mutations between them, and assign the same ploidy states for those variants with preference from samples with higher purity. The copy number profile from the higher purity sample can also be utilized, as this profile would be the most reliable with more detectable CNVs present. This can be a parameter that is specified by the user only when samples are obtained in a short timeframe, as liquid biopsies taken months or years apart are more likely to contain mutations at different copy number states due to tumour clones evolving in response to treatment or relapse.

Another feature that exists in PurEctDNA but was not directly evaluated is the individual estimation of cancer cell fraction (CCF) for all mutations. The CCF is the proportion of cancer cells that each variant is present in which aids in describing a tumours clonal composition[180]. Two CCF values are currently calculated in PurEctDNA

and listed in the resulting summary table, though testing is still required to establish their validity. The first is a mutational CCF where the maximum value is 1, representing a clonal variant. This is calculated where the purity estimate of the individual variant is divided by the highest mutational purity among the listed variants in the sample (Eq. 4). The second value is the CCF of the sample that incorporates the individual purity of each variant and the final purity (Eq. 5). This CCF of the sample is flawed, though, as it produces values above 1. An analysis was not completed to assess this measure due to time limitations to attain lpWGS and CAPP-Seq data on a suitable cohort with numerous serial samples available. Ultimately, this variable would be beneficial as it allows the visualization of how mutational clonality changes over time and if it correlates to treatment or relapse.

Equation 4:

$$\mathrm{CCF}_{\mathrm{mut}} = \frac{\mathrm{Purity}_{\mathrm{individual\ mutation}}}{\mathrm{Purity}_{\mathrm{max\ mutation}}}$$

Equation 5:

$$\mathrm{CCF}_{\mathrm{sample}} = \frac{\mathrm{Purity}_{\mathrm{individual\ mutation}}}{\mathrm{Purity}_{\mathrm{final}}}$$

In addition to continuous testing, optimization, and feature extension to PurEctDNA, an interesting future experiment could examine the effect of fragment enrichment on CNV calling and cfDNA purity estimation. Here, sequencing reads could be selected *in silico* to fragment lengths below 167bp, as this is the peak size of ctDNA that corresponds to chromatosomes after being released from the tumour via apoptosis and necrosis. This might aid in determining if fragment enrichment enhances ctDNA detection, accuracy of somatic CNV calls, and if the sensitivity of PurEctDNA is affected. Analyzing a subset of fragment lengths to select for tumour-derived fragments has been shown to enhance detection of ctDNA as well as identify clinically actionable mutations and CNVs that were not previously found[181]. This is advantageous as fragments can be enriched after lpWGS rather than more expensive sequencing techniques where a larger depth is necessary.

In summary, many modifications and extensions can be performed on PurEctDNA, as it has only been tested on a limited subset of liquid biopsy samples from patients with NHL. Many additional analyses and extended projects can be performed

with PurEctDNA to further improve ctDNA quantification through genetic and epigenetic analyses, on patients with a variety of cancer types, and even to carry over into a clinical setting. Considering this, the present version of PurEctDNA is a highly accurate tool that can be used in the non-invasive monitoring of tumour burden, treatment response and relapse assessment in lymphoma patients.

References

1. Armitage, J. O., Gascoyne, R. D., Lunning, M. A. & Cavalli, F. Non-Hodgkin lymphoma. *The Lancet* **390**, 298–310 (2017).
2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).
3. Canadian Cancer Statistics 2021.
4. Sehn, L. H. & Gascoyne, R. D. Diffuse large B-cell lymphoma: optimizing outcome in the context of clinical and biologic heterogeneity. *Blood* **125**, 22–32 (2015).
5. Hachem, A. & Gartenhaus, R. B. Oncogenes as molecular targets in lymphoma. *Blood* **106**, 1911–1923 (2005).
6. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298–303 (2011).
7. Lenz, G. & Staudt, L. M. Aggressive Lymphomas. *New England Journal of Medicine* **362**, 1417–1429 (2010).
8. Hamel, K. M., Liarski, V. M. & Clark, M. R. Germinal Center B-cells. *Autoimmunity* **45**, 333–347 (2012).
9. Shankland, K. R., Armitage, J. O. & Hancock, B. W. Non-Hodgkin lymphoma. *The Lancet* **380**, 848–857 (2012).
10. The 2016 revision of the World Health Organization classification of lymphoid neoplasms | Blood | American Society of Hematology. https://ashpublications.org/blood/article/127/20/2375/35286/The-2016-revision-of-the-World-Health-Organization.

11. Agbay, R. L. M. C., Jain, N., Loghavi, S., Medeiros, L. J. & Khoury, J. D. Histologic transformation of chronic lymphocytic leukemia/small lymphocytic lymphoma. *American Journal of Hematology* **91**, 1036–1043 (2016).
12. Montoto, S. *et al.* Risk and clinical implications of transformation of follicular lymphoma to diffuse large B-cell lymphoma. *J Clin Oncol* **25**, 2426–2433 (2007).
13. Boussios, S. *et al.* Extranodal diffuse large B-cell lymphomas: A retrospective case series and review of the literature. *Hematol Rep* **10**, 7070 (2018).
14. Møller, M. B., Pedersen, N. T. & Christensen, B. E. Diffuse large B-cell lymphoma: clinical implications of extranodal versus nodal presentation – a population-based study of 1575 cases. *British Journal of Haematology* **124**, 151–159 (2004).
15. Krol, A. D. G. *et al.* Primary extranodal non-Hodgkin's lymphoma (NHL): the impact of alternative definitions tested in the Comprehensive Cancer Centre West population-based NHL registry. *Annals of Oncology* **14**, 131–139 (2003).
16. d'Amore, F. *et al.* Clinicopathological features and prognostic factors in extranodal non-Hodgkin lymphomas. *European Journal of Cancer and Clinical Oncology* **27**, 1201–1208 (1991).
17. Li, S., Young, K. H. & Medeiros, L. J. Diffuse large B-cell lymphoma. *Pathology* **50**, 74–87 (2018).
18. Lee, S. Diffuse large B-cell lymphoma. *Canadian Cancer Society* https://cancer.ca/en/cancer-information/cancer-types/non-hodgkin-lymphoma/what-is-non-hodgkin-lymphoma/diffuse-large-b-cell-lymphoma.

19. Thieblemont, C. *et al.* Non-Hodgkin's lymphoma in very elderly patients over 80 years. A descriptive analysis of clinical presentation and outcome. *Annals of Oncology* **19**, 774–779 (2008).

20. Liu, Y. & Barta, S. K. Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment. *American Journal of Hematology* **94**, 604–616 (2019).

21. El-Galaly, T. C. *et al.* FDG-PET/CT in the management of lymphomas: current status and future directions. *Journal of Internal Medicine* **284**, 358–376 (2018).

22. De Leval, L. & Harris, N. L. Variability in immunophenotype in diffuse large B-cell lymphoma and its clinical relevance. *Histopathology* **43**, 509–528 (2003).

23. Colomo, L. *et al.* Clinical impact of the differentiation profile assessed by immunophenotyping in patients with diffuse large B-cell lymphoma. *Blood* **101**, 78–84 (2003).

24. Scott, D. W. *et al.* High-grade B-cell lymphoma with MYC and BCL2 and/or BCL6 rearrangements with diffuse large B-cell lymphoma morphology. *Blood* **131**, 2060–2064 (2018).

25. A Predictive Model for Aggressive Non-Hodgkin's Lymphoma. *New England Journal of Medicine* **329**, 987–994 (1993).

26. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).

27. Wright, G. W. *et al.* A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551-568.e14 (2020).

28. Ennishi, D. *et al.* Double-Hit Gene Expression Signature Defines a Distinct Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. *J Clin Oncol* **37**, 190–201 (2019).

29. Yoon, S. O. *et al.* MYC translocation and an increased copy number predict poor prognosis in adult diffuse large B-cell lymphoma (DLBCL), especially in germinal centre-like B cell (GCB) type. *Histopathology* **53**, 205–217 (2008).

30. Pasqualucci, L. & Dalla-Favera, R. Genetics of diffuse large B-cell lymphoma. *Blood* **131**, 2307–2319 (2018).

31. Pasqualucci, L. *et al.* Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* **471**, 189–195 (2011).

32. Morin, R. D. *et al.* Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265 (2013).

33. Karube, K. *et al.* Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia* **32**, 675–684 (2018).

34. Chapuy, B. *et al.* Diffuse large B-cell lymphoma patient-derived xenograft models capture the molecular and biological heterogeneity of the disease. *Blood* **127**, 2203–2213 (2016).

35. Schuetz, J. M. *et al.* BCL2 mutations in diffuse large B-cell lymphoma. *Leukemia* **26**, 1383–1390 (2012).

36. Davis, R. E. *et al.* Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature* **463**, 88–92 (2010).

37. Yang, Y. *et al.* Exploiting Synthetic Lethality for the Therapy of ABC Diffuse Large B Cell Lymphoma. *Cancer Cell* **21**, 723–737 (2012).

38. Schmitz, R. *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine* **378**, 1396–1407 (2018).

39. Chapuy, B. *et al.* Molecular Subtypes of Diffuse Large B-cell Lymphoma are Associated with Distinct Pathogenic Mechanisms and Outcomes. *Nat Med* **24**, 679–690 (2018).

40. Morin, R. D., Arthur, S. E. & Hodson, D. J. Molecular profiling in diffuse large B-cell lymphoma: why so many types of subtypes? *British Journal of Haematology* **n/a**,.

41. Chapuy, B. *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med* **24**, 679–690 (2018).

42. Schmitz, R. *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine* **378**, 1396–1407 (2018).

43. Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *Science* **194**, 23–28 (1976).

44. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).

45. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology* **15**, 81–94 (2018).

46. Coiffier, B. *et al.* CHOP Chemotherapy plus Rituximab Compared with CHOP Alone in Elderly Patients with Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine* **346**, 235–242 (2002).

47. Rovira, J. *et al.* Prognosis of patients with diffuse large B cell lymphoma not reaching complete response or relapsing after frontline chemotherapy or immunochemotherapy. *Ann Hematol* **94**, 803–812 (2015).

48. Crump, M. *et al.* Outcomes in refractory diffuse large B-cell lymphoma: results from the international SCHOLAR-1 study. *Blood* **130**, 1800–1808 (2017).

49. Morin, R. D. *et al.* Genetic Landscapes of Relapsed and Refractory Diffuse Large B-Cell Lymphomas. *Clin Cancer Res* **22**, 2290–2300 (2016).

50. Rushton, C. K. *et al.* Genetic and evolutionary patterns of treatment resistance in relapsed B-cell lymphoma. *Blood Advances* **4**, 2886–2898 (2020).

51. Mareschal, S. *et al.* Whole exome sequencing of relapsed/refractory patients expands the repertoire of somatic mutations in diffuse large B-cell lymphoma. *Genes Chromosomes Cancer* **55**, 251–267 (2016).

52. Nijland, M. *et al.* Mutational Evolution in Relapsed Diffuse Large B-Cell Lymphoma. *Cancers (Basel)* **10**, E459 (2018).

53. Nesic, M. *et al.* The mutational profile of immune surveillance genes in diagnostic and refractory/relapsed DLBCLs. *BMC Cancer* **21**, 829 (2021).

54. Melchardt, T. *et al.* Clonal evolution in relapsed and refractory diffuse large B-cell lymphoma is characterized by high dynamics of subclones. *Oncotarget* **7**, 51494–51502 (2016).

55. Juskevicius, D. *et al.* Distinct genetic evolution patterns of relapsing diffuse large B-cell lymphoma revealed by genome-wide copy number aberration and targeted sequencing analysis. *Leukemia* **30**, 2385–2395 (2016).

56. Morin, R. D., Arthur, S. E. & Hodson, D. J. Molecular profiling in diffuse large B-cell lymphoma: why so many types of subtypes? *British Journal of Haematology* **n/a**,.

57. Ma, M. *et al.* "Liquid biopsy"—ctDNA detection with great potential and challenges. *Ann Transl Med* **3**, (2015).

58. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol* **10**, 472–484 (2013).

59. Zviran, A. *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med* **26**, 1114–1124 (2020).

60. Henry, N. L. & Hayes, D. F. Cancer biomarkers. *Mol Oncol* **6**, 140–146 (2012).

61. Wu, F.-T., Lu, L., Xu, W. & Li, J.-Y. Circulating tumor DNA: clinical roles in diffuse large B cell lymphoma. *Ann Hematol* **98**, 255–269 (2019).

62. De Mattos-Arruda, L. *et al.* Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nat Commun* **6**, 8839 (2015).

63. Birkenkamp-Demtröder, K. *et al.* Genomic Alterations in Liquid Biopsies from Patients with Bladder Cancer. *European Urology* **70**, 75–82 (2016).

64. Sriram, K. B. *et al.* Pleural fluid cell-free DNA integrity index to identify cytologically negative malignant pleural effusions including mesotheliomas. *BMC Cancer* **12**, 428 (2012).

65. Jahr, S. *et al.* DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells. *Cancer Res* **61**, 1659–1665 (2001).

66. Perakis, S. & Speicher, M. R. Emerging concepts in liquid biopsies. *BMC Medicine* **15**, 75 (2017).

67. Wu, F.-T., Lu, L., Xu, W. & Li, J.-Y. Circulating tumor DNA: clinical roles in diffuse large B cell lymphoma. *Ann Hematol* **98**, 255–269 (2019).

68. Perdigones, N. & Murtaza, M. Capturing tumor heterogeneity and clonal evolution in solid cancers using circulating tumor DNA analysis. *Pharmacology & Therapeutics* **174**, 22–26 (2017).

69. Ofori, K., Bhagat, G. & Rai, A. J. Exosomes and extracellular vesicles as liquid biopsy biomarkers in diffuse large B-cell lymphoma: Current state of the art and unmet clinical needs. *British Journal of Clinical Pharmacology* **87**, 284–294 (2021).

70. Stoorvogel, W., Kleijmeer, M. J., Geuze, H. J. & Raposo, G. The Biogenesis and Functions of Exosomes. *Traffic* **3**, 321–330 (2002).

71. Ha, D., Yang, N. & Nadithe, V. Exosomes as therapeutic drug carriers and delivery vehicles across biological membranes: current perspectives and future challenges. *Acta Pharm Sin B* **6**, 287–296 (2016).

72. Tai, Y.-L., Chen, K.-C., Hsieh, J.-T. & Shen, T.-L. Exosomes in cancer development and clinical applications. *Cancer Science* **109**, 2364–2374 (2018).

73. Rajagopal, C. & Harikumar, K. B. The Origin and Functions of Exosomes in Cancer. *Frontiers in Oncology* **8**, (2018).

74. Chaput, N. & Théry, C. Exosomes: immune properties and potential clinical implementations. *Semin Immunopathol* **33**, 419–440 (2011).

75. Melo, S. A. *et al.* Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature* **523**, 177–182 (2015).

76. Peinado, H. *et al.* Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. *Nat Med* **18**, 883–891 (2012).

77. Costa-Silva, B. *et al.* Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nat Cell Biol* **17**, 816–826 (2015).

78. Feng, Y. *et al.* Exosome-derived miRNAs as predictive biomarkers for diffuse large B-cell lymphoma chemotherapy resistance. *Epigenomics* **11**, (2018).

79. Zare, N., Haghjooy Javanmard, S., Mehrzad, V., Eskandari, N. & Kefayat, A. Evaluation of exosomal miR-155, let-7g and let-7i levels as a potential noninvasive biomarker among refractory/relapsed patients, responsive patients and patients receiving R-CHOP. *Leukemia & Lymphoma* **60**, 1877–1889 (2019).

80. Zare, N., Eskandari, N., Mehrzad, V. & Javanmard, S. H. The expression level of hsa-miR-146a-5p in plasma-derived exosomes of patients with diffuse large B-cell lymphoma. *J Res Med Sci* **24**, 10 (2019).

81. Decruyenaere, P., Offner, F. & Vandesompele, J. Circulating RNA biomarkers in diffuse large B-cell lymphoma: a systematic review. *Experimental Hematology & Oncology* **10**, 13 (2021).

82. Garcia, V. *et al.* Extracellular Tumor-Related mRNA in Plasma of Lymphoma Patients and Survival Implications. *PLOS ONE* **4**, e8173 (2009).

83. Attia, F., Moustafa, A., El-Maraghy, N. & Ibrahim, G. Clinical significance of suppressor of cytokines signaling-3 mRNA expression from patients with non-Hodgkin lymphoma under chemotherapy. *Cancer biomarkers : section A of Disease markers* **11**, 41–7 (2011).

84. Yuan, W. X., Gui, Y. X., Na, W. N., Chao, J. & Yang, X. Circulating microRNA-125b and microRNA-130a expression profiles predict chemoresistance to R-CHOP in diffuse large B-cell lymphoma patients. *Oncology Letters* **11**, 423–432 (2016).

85. Song, G. *et al.* Serum microRNA expression profiling predict response to R-CHOP treatment in diffuse large B cell lymphoma patients. *Ann Hematol* **93**, 1735–1743 (2014).

86. Li, J., Fu, R., Yang, L. & Tu, W. miR-21 expression predicts prognosis in diffuse large B-cell lymphoma. *Int J Clin Exp Pathol* **8**, 15019–15024 (2015).

87. Senousy, M. A., El-Abd, A. M., Abdel-Malek, R. R. & Rizk, S. M. Circulating long non-coding RNAs HOTAIR, Linc-p21, GAS5 and XIST expression profiles in diffuse large B-cell lymphoma: association with R-CHOP responsiveness. *Sci Rep* **11**, 2095 (2021).

88. Lopez-Santillan, M., Larrabeiti-Etxebarria, A., Arzuaga-Mendez, J., Lopez-Lopez, E. & Garcia-Orad, A. Circulating miRNAs as biomarkers in diffuse large B-cell lymphoma: a systematic review. *Oncotarget* **9**, 22850–22861 (2018).

89. Suehara, Y. *et al.* Mutations found in cell-free DNAs of patients with malignant lymphoma at remission can derive from clonal hematopoiesis. *Cancer Science* **110**, 3375–3381 (2019).

90. Jahr, S. *et al.* DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells. *Cancer Res* **61**, 1659–1665 (2001).

91. Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from

maternal blood. *Proceedings of the National Academy of Sciences* **105**, 16266–16271 (2008).

92. Jiang, P. *et al.* Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci USA* **112**, E1317–E1325 (2015).
93. Underhill, H. R. *et al.* Fragment Length of Circulating Tumor DNA. *PLOS Genetics* **12**, e1006162 (2016).
94. Alix-Panabières, C., Schwarzenbach, H. & Pantel, K. Circulating tumor cells and circulating tumor DNA. *Annual review of medicine* **63**, 199–215.
95. Arzuaga-Mendez, J. *et al.* Cell-free DNA as a biomarker in diffuse large B-cell lymphoma: A systematic review. *Critical Reviews in Oncology/Hematology* **139**, 7–15 (2019).
96. Thierry, A. R. *et al.* Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. *Nat Med* **20**, 430–435 (2014).
97. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A* **102**, 16368–16373 (2005).
98. Roschewski, M., Staudt, L. M. & Wilson, W. H. Dynamic monitoring of circulating tumor DNA in non-Hodgkin lymphoma. *Blood* **127**, 3127–3132 (2016).
99. Hohaus, S. *et al.* Cell-free circulating DNA in Hodgkin's and non-Hodgkin's lymphomas. *Annals of Oncology* **20**, 1408–1413 (2009).
100. Macaulay, C. *et al.* Interim Circulating Tumor DNA As a Prognostic Biomarker in the Setting of Interim PET-Based Adaptive Therapy for DLBCL. *Blood* **134**, 1600 (2019).

101. Kristensen, L. S. *et al.* Aberrant methylation of cell-free circulating DNA in plasma predicts poor outcome in diffuse large B cell lymphoma. *Clinical Epigenetics* **8**, 95 (2016).

102. Kurtz, D. M. *et al.* Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma. *J Clin Oncol* **36**, 2845–2853 (2018).

103. McEvoy, A. C. *et al.* Correlation between circulating tumour DNA and metabolic tumour burden in metastatic melanoma patients. *BMC Cancer* **18**, 726 (2018).

104. Decazes, P. *et al.* Correlations between baseline 18F-FDG PET tumour parameters and circulating DNA in diffuse large B cell lymphoma and Hodgkin lymphoma. *EJNMMI Res* **10**, 120 (2020).

105. Roschewski, M. *et al.* Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B-cell lymphoma: a correlative biomarker study. *The Lancet Oncology* **16**, 541–549 (2015).

106. Cheson, B. D. *et al.* Recommendations for Initial Evaluation, Staging, and Response Assessment of Hodgkin and Non-Hodgkin Lymphoma: The Lugano Classification. *J Clin Oncol* **32**, 3059–3067 (2014).

107. Schoder, H., Gonen, M. & Franklin, B. Screening for Cancer with PET and PET/CT: Potential and Limitations. 15.

108. Kurtz, D. M. *et al.* Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood* **125**, 3679–3687 (2015).

109. Scherer, F. *et al.* Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. *Sci Transl Med* **8**, 364ra155 (2016).

110. Esfahani, M. S. *et al.* Inferring gene expression from cell-free DNA fragmentation profiles. *Nat Biotechnol* **40**, 585–597 (2022).

111. Mehrmohamadi, M. *et al.* Distinct Chromatin Accessibility Profiles of Lymphoma Subtypes Revealed By Targeted Cell Free DNA Profiling. *Blood* **132**, 672 (2018).

112. Lianidou, E. Detection and relevance of epigenetic markers on ctDNA: recent advances and future outlook. *Molecular Oncology* **15**, 1683–1700 (2021).

113. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* **17**, 223–238 (2017).

114. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017).

115. Larson, N. B. & Fridley, B. L. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29**, 1888–1889 (2013).

116. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* **26**, 64–70 (2015).

117. Gorgannezhad, L., Umer, M., Nazmul Islam, M., Nguyen, N.-T. & A. Shiddiky, M. J. Circulating tumor DNA and liquid biopsy: opportunities, challenges, and recent advances in detection technologies. *Lab on a Chip* **18**, 1174–1196 (2018).

118. Elazezy, M. & Joosse, S. A. Techniques of using circulating tumor DNA as a liquid biopsy component in cancer management. *Comput Struct Biotechnol J* **16**, 370–378 (2018).

119. Chin, R.-I. *et al.* Detection of Solid Tumor Molecular Residual Disease (MRD) Using Circulating Tumor DNA (ctDNA). *Mol Diagn Ther* **23**, 311–331 (2019).

120. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**, 548–554 (2014).

121. Forshew, T. *et al.* Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA. *Sci. Transl. Med.* **4**, (2012).

122. Frenel, J. S. *et al.* Serial Next-Generation Sequencing of Circulating Cell-Free DNA Evaluating Tumor Clone Response To Molecularly Targeted Drug Administration. *Clin Cancer Res* **21**, 4586–4596 (2015).

123. Lightbody, E. D., Dutta, A. K. & Ghobrial, I. M. MRDetect: An Ultrasensitive Solution to Monitor Low Tumor Burden in Liquid Biopsies. *The Hematologist* **17**, (2020).

124. Diaz, L. A. & Bardelli, A. Liquid Biopsies: Genotyping Circulating Tumor DNA. *J Clin Oncol* **32**, 579–586 (2014).

125. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

126. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413–421 (2012).

127. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology* **14**, R80 (2013).

128. Su, X., Zhang, L., Zhang, J., Meric-Bernstam, F. & Weinstein, J. N. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* **28**, 2265–2266 (2012).

129. A Novel Multiplex Droplet Digital PCR Assay to Identify and Quantify KRAS Mutations in Clinical Specimens - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1525157818301144.

130. Li, Y. & Xie, X. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics* **30**, 2121–2129 (2014).

131. Luo, Z., Fan, X., Su, Y. & Huang, Y. S. Accurity: accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants. *Bioinformatics* **34**, 2004–2011 (2018).

132. Zviran, A. *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med* **26**, 1114–1124 (2020).

133. Volik, S., Alcaide, M., Morin, R. D. & Collins, C. Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies. *Mol Cancer Res* **14**, 898–908 (2016).

134. Vogelstein, B. & Kinzler, K. W. Digital PCR. *PNAS* **96**, 9236–9241 (1999).

135. Alcaide, M. *et al.* Multiplex Droplet Digital PCR Quantification of Recurrent Somatic Mutations in Diffuse Large B-Cell and Follicular Lymphoma. *Clinical Chemistry* **62**, 1238–1247 (2016).

136. Sanmamed, M. F. *et al.* Quantitative Cell-Free Circulating BRAFV600E Mutation Analysis by Use of Droplet Digital PCR in the Follow-up of Patients with Melanoma Being Treated with BRAF Inhibitors. *Clinical Chemistry* **61**, 297–304 (2015).

137. Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine* (2015) doi:10.1126/scitranslmed.aab0021.

138. Gorgannezhad, L., Umer, M., Islam, M. N., Nguyen, N.-T. & Shiddiky, M. J. A. Circulating tumor DNA and liquid biopsy: opportunities, challenges, and recent advances in detection technologies. *Lab Chip* **18**, 1174–1196 (2018).

139. van Ginkel, J. H., Huibers, M. M. H., van Es, R. J. J., de Bree, R. & Willems, S. M. Droplet digital PCR for detection and quantification of circulating tumor DNA in plasma of head and neck cancer patients. *BMC Cancer* **17**, 428 (2017).

140. Faham, M. *et al.* Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* **120**, 5173–5180 (2012).

141. Scherer, F., Kurtz, D. M., Diehn, M. & Alizadeh, A. A. High-throughput sequencing for noninvasive disease detection in hematologic malignancies. *Blood* **130**, 440–452 (2017).

142. Roschewski, M. *et al.* Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B-cell lymphoma: a correlative biomarker study. *The Lancet Oncology* **16**, 541–549 (2015).

143. Gale, D. *et al.* Development of a highly sensitive liquid biopsy platform to detect clinically-relevant cancer mutations at low allele fractions in cell-free DNA. *PLOS ONE* **13**, e0194630 (2018).

144. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences* **108**, 9530–9535 (2011).

145. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med* **4**, 136ra68 (2012).

146. Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* **34**, 547–555 (2016).

147. Iwahashi, N. *et al.* Liquid biopsy-based comprehensive gene mutation profiling for gynecological cancer using CAncer Personalized Profiling by deep Sequencing. *Sci Rep* **9**, 10426 (2019).

148. Bratman, S. V., Newman, A. M., Alizadeh, A. A. & Diehn, M. Potential clinical utility of ultrasensitive circulating tumor DNA detection with CAPP-Seq. *Expert Review of Molecular Diagnostics* **15**, 715–719 (2015).

149. Chin, R.-I. *et al.* Detection of Solid Tumor Molecular Residual Disease (MRD) Using Circulating Tumor DNA (ctDNA). *Mol Diagn Ther* **23**, 311–331 (2019).

150. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**, 548–554 (2014).

151. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

152. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research* **40**, 11189–11201 (2012).

153. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213–219 (2013).

154. hmftools/sage at master · hartwigmedical/hmftools. *GitHub* https://github.com/hartwigmedical/hmftools.

155. Chen, Z. *et al.* Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep* **10**, 3501 (2020).

156. Wang, M. *et al.* SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Sci Rep* **10**, 12898 (2020).

157. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal* **16**, 15–24 (2018).

158. Odegaard, J. I. *et al.* Validation of a Plasma-Based Comprehensive Cancer Genotyping Assay Utilizing Orthogonal Tissue- and Plasma-Based Methodologies. *Clin Cancer Res* **24**, 3539–3549 (2018).

159. Leary, R. J. *et al.* Detection of Chromosomal Alterations in the Circulation of Cancer Patients with Whole-Genome Sequencing. *Sci Transl Med* **4**, 162ra154 (2012).

160. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).

161. Chen, X. *et al.* Low-pass Whole-genome Sequencing of Circulating Cell-free DNA Demonstrates Dynamic Changes in Genomic Copy Number in a Squamous Lung Cancer Clinical Cohort. *Clin Cancer Res* **25**, 2254–2263 (2019).

162. Sumanasuriya, S. *et al.* Elucidating Prostate Cancer Behaviour During Treatment via Low-pass Whole-genome Sequencing of Circulating Tumour DNA. *European Urology* **80**, 243–253 (2021).

163. Bouzidi, A. *et al.* Low-Coverage Whole Genome Sequencing of Cell-Free DNA From Immunosuppressed Cancer Patients Enables Tumor Fraction Determination and Reveals Relevant Copy Number Alterations. *Front Cell Dev Biol* **9**, 661272 (2021).

164. Liu, B. *et al.* Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget* **4**, 1868–1881 (2013).

165. Wei, T. *et al.* Genome-wide profiling of circulating tumor DNA depicts landscape of copy number alterations in pancreatic cancer with liver metastasis. *Molecular Oncology* **14**, 1966–1977 (2020).

166. Shao, X. *et al.* Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical Genetics* **20**, 175 (2019).

167. Smith, J. C. & Sheltzer, J. M. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. *eLife* **7**, e39217 (2018).

168. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017).

169. Raman, L., Dheedene, A., De Smet, M., Van Dorpe, J. & Menten, B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Research* **47**, 1605–1614 (2019).

170. Janevski, A., Varadan, V., Kamalakaran, S., Banerjee, N. & Dimitrova, N. Effective normalization for copy number variation detection from whole genome sequencing. *BMC Genomics* **13**, S16 (2012).

171. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

172. Rushton, C. K. *et al.* Genetic and evolutionary patterns of treatment resistance in relapsed B-cell lymphoma. *Blood Adv* **4**, 2886–2898 (2020).

173. Grande, B. M. *et al.* Burkitt Lymphoma Genome Sequencing Project (BLGSP): Integrative Genomic and Transcriptomic Characterization of Burkitt Lymphoma. *Blood* **130**, 39 (2017).

174. Arthur, S. E. *et al.* Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat Commun* **9**, 4001 (2018).

175. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

176. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

177. Zenz, T. *et al.* TP53 mutation and survival in aggressive B cell lymphoma. *International Journal of Cancer* **141**, 1381–1388 (2017).

178. Iqbal, J. *et al.* Genomic signatures in B-cell lymphoma: How can these improve precision in diagnosis and inform prognosis? *Blood Reviews* **30**, 73–88 (2016).

179. Huang, Y.-H. *et al.* CREBBP/EP300 mutations promoted tumor progression in diffuse large B-cell lymphoma through altering tumor-associated macrophage polarization via FBXW7-NOTCH-CCL2/CSF1 axis. *Sig Transduct Target Ther* **6**, 1–14 (2021).

180. Cmero, M. *et al.* Inferring structural variant cancer cell fraction. *Nat Commun* **11**, 730 (2020).

181. Mouliere, F. *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Science Translational Medicine* **10**, eaat4921 (2018).

Appendix

Table A.1. Summary of plasma sample cohorts used

	# Samples with CAPP-Seq	# Samples with lpWGS
LY17	92	39
OZM073	59	21
Epizyme	59	24
Montreal	88	511
BC ctDNA	78	15
Total	376	610

Table A.2. List of lymphoma-specific genes used in the CAPP-Seq panel and to subset genes during PurEctDNA purity estimation

Chromosome	Hugo_Symbol
chr1	TNFRSF14
chr1	SPEN
chr1	ID3
chr1	ARID1A
chr1	ZC3H12A
chr1	BCL10
chr1	CD58
chr1	BTG2
chr2	BIRC6
chr2	FBXO11
chr3	MYD88
chr3	RHOA
chr3	GNAI2
chr3	NFKBIZ
chr3	TBL1XR1
chr3	KLHL6
chr4	NFKB1
chr5	EBF1
chr6	IRF4
chr6	HIST1H1C
chr6	HIST1H1E
chr6	PIM1
chr6	CCND3
chr6	NFKBIE
chr6	TMEM30A
chr6	SGK1

chr6	TNFAIP3
chr7	GNA12
chr7	CARD11
chr7	BRAF
chr7	EZH2
chr8	MYC
chr9	NOTCH1
chr10	FAS
chr11	MPEG1
chr11	MS4A1
chr11	CCND1
chr11	ETS1
chr12	KMT2D
chr12	STAT6
chr12	HVCN1
chr13	FOXO1
chr13	RB1
chr14	ZFP36L1
chr15	B2M
chr15	RFX7
chr15	SIN3A
chr16	CREBBP
chr16	SOCS1
chr16	IL4R
chr16	IRF8
chr17	TP53
chr17	CD79B
chr17	GNA13
chr18	BCL2
chr19	TCF3
chr19	S1PR2
chr19	MEF2B
chr19	POU2F2
chr22	EP300
chrX	P2RY8
chrX	TMSB4X
chrX	DDX3X

Table A.3. Variables and resulting purity estimates used in the *in silico* dilution of four ctDNA genomes from PurEctDNA and Battenberg

Sample ID	Normal fraction	Final purity	Initial purity	Coverage normal	Coverage tumour	Tumour fraction	PurEctDNA purity	Battenberg purity
CAMP-0009	1	0.05	0.74	27.907	36.658	0.058	0.045	0.95
CAMP-0009	1	0.1	0.74	27.907	36.658	0.125	0.090	0.101
CAMP-0009	1	0.15	0.74	27.907	36.658	0.204	0.135	0.147
CAMP-0009	1	0.2	0.74	27.907	36.658	0.299	0.181	0.193
CAMP-0009	1	0.25	0.74	27.907	36.658	0.414	0.227	0.245
CAMP-0009	1	0.3	0.74	27.907	36.658	0.557	0.274	0.287
CAMP-0009	1	0.35	0.74	27.907	36.658	0.740	0.322	0.337
CAMP-0009	0.5	0.4	0.74	27.907	36.658	0.491	0.370	0.391
CAMP-0009	0.5	0.45	0.74	27.907	36.658	0.659	0.420	0.441
CAMP-0009	0.5	0.5	0.74	27.907	36.658	0.906	0.469	0.500
CAMP-0009	0.25	0.55	0.74	27.907	36.658	0.654	0.510	0.543
CAMP-0009	0.2	0.6	0.74	27.907	36.658	0.830	0.562	0.613
CAMP-0009	0.1	0.65	0.74	27.907	36.658	0.825	0.615	0.665
CABN-0003	1	0.05	0.4	20.304	27.866	0.104	0.054	0.944
CABN-0003	1	0.1	0.4	20.304	27.866	0.243	0.099	0.950
CABN-0003	1	0.15	0.4	20.304	27.866	0.437	0.155	0.127
CABN-0003	1	0.2	0.4	20.304	27.866	0.729	0.213	0.171
CABN-0003	0.5	0.25	0.4	20.304	27.866	0.607	0.273	0.226
CABN-0003	0.25	0.3	0.4	20.304	27.866	0.546	0.338	0.262
CABN-0003	0.1	0.35	0.4	20.304	27.866	0.510	0.408	0.333
OZM073-005	1	0.05	0.74	20.774	34.136	0.041	0.039	0.963

OZM073-005	1	0.1	0.74	20.774	34.136	0.087	0.078	0.981
OZM073-005	1	0.15	0.74	20.774	34.136	0.140	0.119	0.132
OZM073-005	1	0.2	0.74	20.774	34.136	0.203	0.159	0.171
OZM073-005	1	0.25	0.74	20.774	34.136	0.277	0.200	0.205
OZM073-005	1	0.3	0.74	20.774	34.136	0.365	0.241	0.254
OZM073-005	1	0.35	0.74	20.774	34.136	0.473	0.283	0.296
OZM073-005	1	0.4	0.74	20.774	34.136	0.609	0.326	0.341
OZM073-005	1	0.45	0.74	20.774	34.136	0.782	0.371	0.389
OZM073-005	0.5	0.5	0.74	20.774	34.136	0.507	0.412	0.452
OZM073-005	0.5	0.55	0.74	20.774	34.136	0.669	0.459	0.481
OZM073-005	0.5	0.6	0.74	20.774	34.136	0.913	0.506	0.536
OZM073-005	0.25	0.65	0.74	20.774	34.136	0.659	0.552	0.592
OZM073-005	0.2	0.7	0.74	20.774	34.136	0.852	0.601	0.639
OZM073-005	0.1	0.75	0.74	20.774	34.136	0.913	0.650	0.682
OZM073-025	1	0.05	0.61	22.681	37.213	0.076	0.056	0.095
OZM073-025	1	0.1	0.61	22.681	37.213	0.174	0.114	0.120
OZM073-025	1	0.15	0.61	22.681	37.213	0.305	0.173	0.194
OZM073-025	1	0.2	0.61	22.681	37.213	0.488	0.236	0.253
OZM073-025	1	0.25	0.61	22.681	37.213	0.762	0.301	0.305
OZM073-025	0.5	0.3	0.61	22.681	37.213	0.609	0.368	0.365
OZM073-025	0.25	0.35	0.61	22.681	37.213	0.533	0.437	0.469
OZM073-025	0.1	0.4	0.61	22.681	37.213	0.488	0.512	0.512

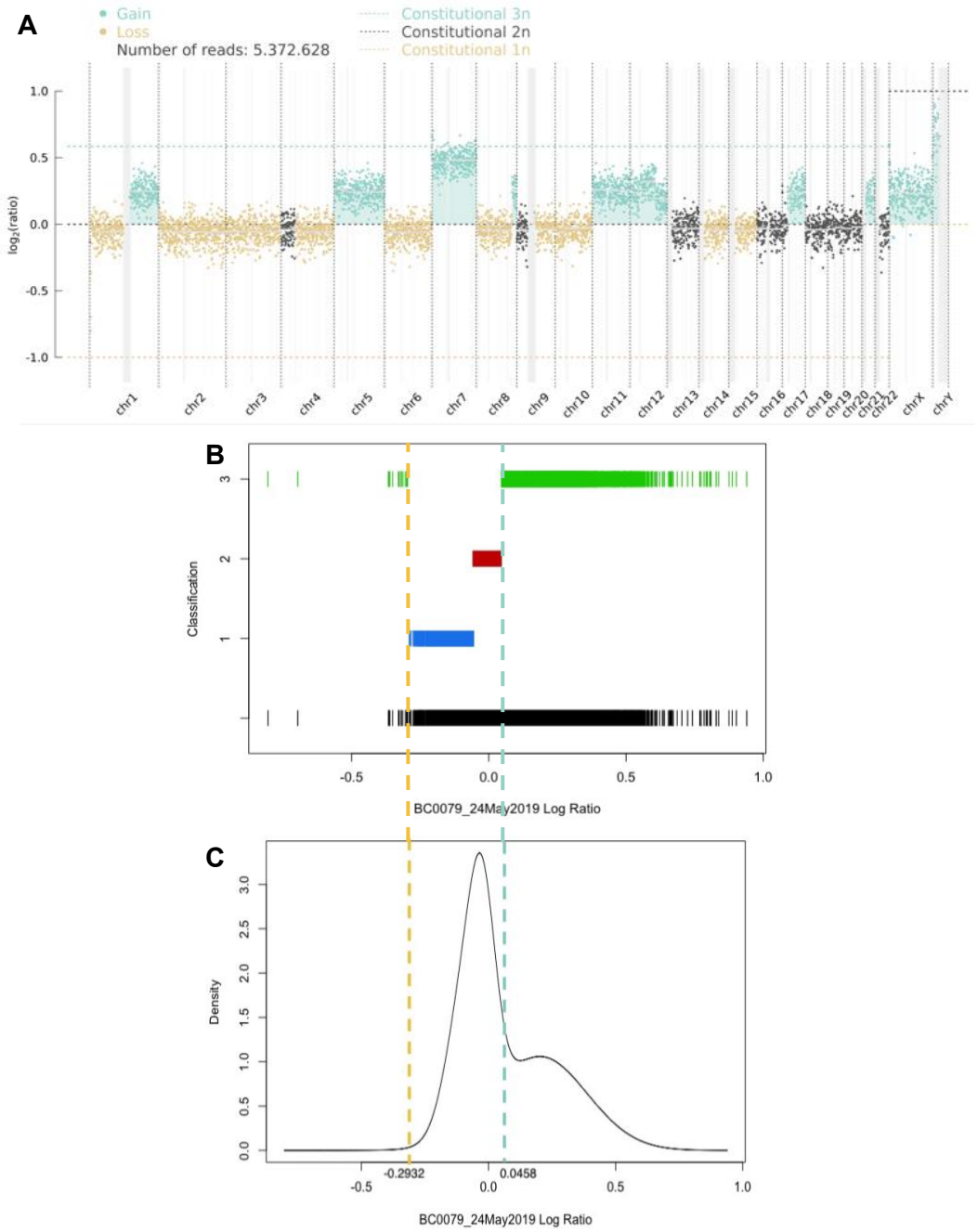

Figure A.1. Manual discovery of the mixture model-based cutoff values and best cluster values to reiterate back into WisecondorX

Using a mixture model approach, I manually determined the appropriate cluster number for each of the 14 samples originally run through WisecondorX. I calculated the cut-off values between copy lost and gained states utilizing the samples classification plot (B) and aligned it to the corresponding density plot (C). With this manual calculation, the offset value to shift bins by was 0.0458 and the beta was 0.2932. After incorporating the standardized cluster numbers of 1 and 3 for the offset and beta arguments respectively, WisecondorX calculated the offset number to shift bins as 0.0692 and the beta (also the tumour purity estimation) as 0.3728. This purity estimate of 37% is similar to the estimate of 43% from IchorCNA, and properly shifts and calls structural aberrations for this sample.

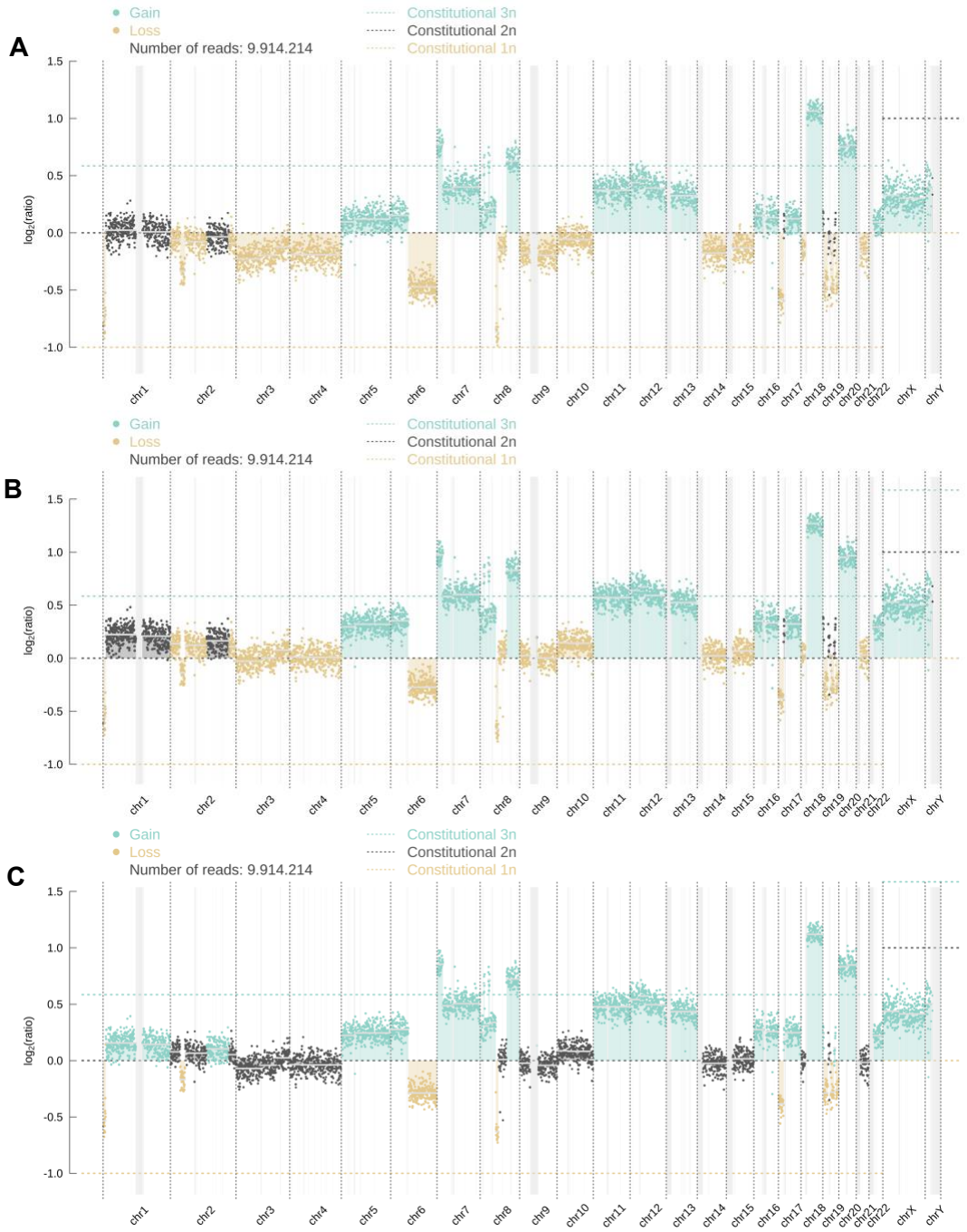

Figure A.2. Modifying segment positions and aberration calls from WisecondorX using the offset mean and beta

(A) Original copy number profile produced by WisecondorX before any modifications were performed on the program. Here, the mean log2 ratio of segments is centered at zero though many gains and amplifications are present resulting in offset CNV calls. (B) After all bins are clustered into one mixture component and the mean is utilized to adjust the segments to their appropriate locations. (C) After the offset mean is incorporated to move segments and the purity estimate is used as the argument “beta” to adjust cut-off values for copy number aberrations to be called according to their new log2 ratio values.

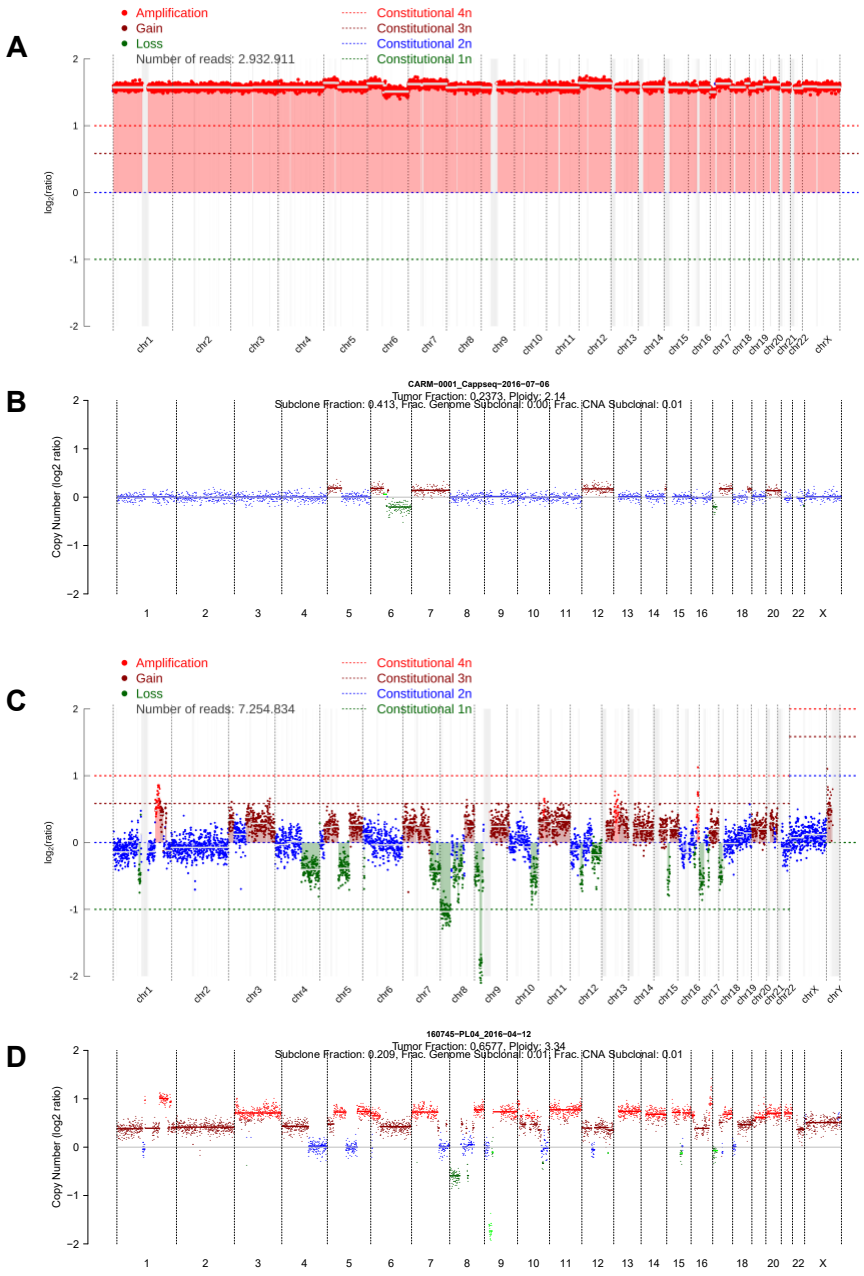

Figure A.3. Copy number profiles from WisecondorX and IchorCNA that correspond to extreme outliers during the PGA analysis

(A) This shows a copy number profile from WisecondorX that is one of the largest outliers from the PGA analysis (Fig 2.10, Chapter 2). Here, a limitation of WisecondorX is shown, where all segments are wrongly assigned as amplifications, with 100% of the genome shown as altered, compared to IchorCNA with a PGA of 23% (B). This much of an over-inferral of CNVs is extremely rare for WisecondorX and has since been found to be a bug in the snakefile when biased segments are adjusted due to the unadjusted profile showing a diploid tumour (not shown). A similar phenomenon is seen with the CNV profile from a sample containing a high PGA from IchorCNA and a low PGA from WisecondorX (C, D). This highlights IchorCNA's ability to either assign tumours as triploid or could be a case of overfitting. This cannot be confirmed whether the IchorCNA or WisecondorX copy number profile is more accurate to the tumour without the BAF or tissue biopsy sample from a similar timepoint.

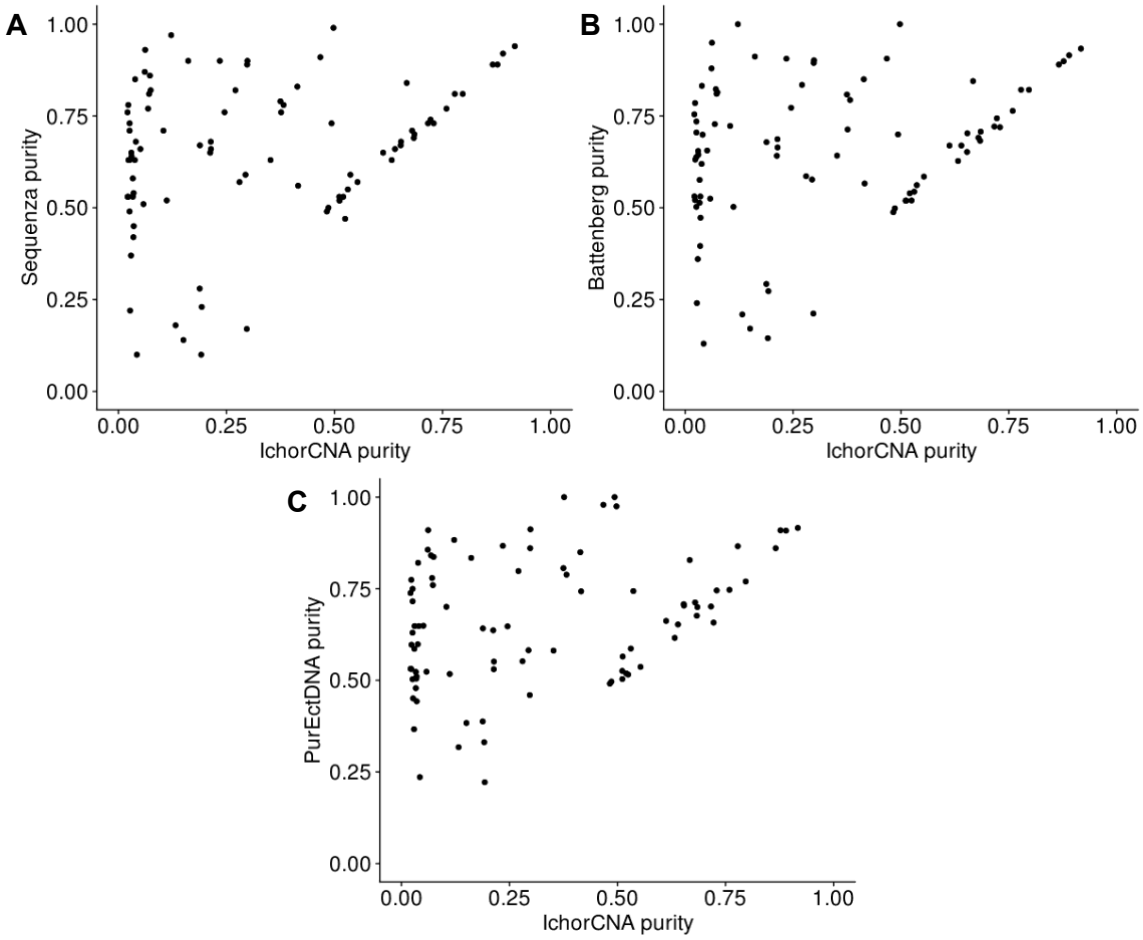

Figure A.4. IchorCNA underestimates purity values for tissue tumour genomes
This underestimation results in a poor correlation to Sequenza (A), Battenberg (B) and PurEctDNA (C) estimates.