Precision Oncology Opportunities And Disease Insights From Next-Generation-Sequencing Profiling Of Routine Clinical Biospecimens

by

Daniel H. Hovelson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2017

Doctoral Committee:

Associate Professor Scott A. Tomlins, Chair
Professor Arul M. Chinnaiyan
Professor Jun Z. Li
Assistant Professor Ryan E. Mills
Adjunct Professor Daniel R. Rhodes

Daniel H. Hovelson
hovelson@umich.edu
ORCID id: 0000-0003-1881-4451

ACKNOWLEDGEMENTS

This dissertation is the product of hard work, invaluable mentorship from Dr. Scott Tomlins, perpetual guidance and patience from all Tomlins lab members and collaborators, and the love and support of my wife and 3 awesome kids.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AOHC - AcroMetrix Oncology Hotspot Control
AR – Androgen receptor
cfDNA – Cell-free DNA
ctDNA – Cell-free tumor DNA
CNAs - Copy number alterations
CRPC – Castration-resistant prostate cancer
CTC – Circulating tumor cell
ddPCR – Digital droplet polymerase chain reaction
FFPE - Formalin fixed paraffin embedded
GoF - Gain-of-function
H&E - Hematoxlyin and eosin
indels - Insertions/deletions
LoF - Loss-of-function
LU - Lung cohort
MCR - Minimal common region
MO - Molecular cohort
NCCN - National Comprehensive Cancer Network
NGS - Next generation sequencing
OCP - Oncomine Comprehensive Assay
PGM - Personal Genome Machine
PR - Prostate cohort
QMRS - Quantitative Multiplex Reference Standard
SNV – Single nucleotide variant
SCC - Small cell carcinoma
TCGA - The Cancer Genome Atlas
trDNA – trans-renal cell-free DNA
WGS – Whole genome sequencing

ABSTRACT

Rapid technological developments in next-generation sequencing (NGS) and inter-institutional collaborations including The Cancer Genome Atlas (TCGA) have enabled comprehensive characterization of the genomic, transcriptomic, and epigenetic landscapes from bulk tissue specimens in a wide range of cancers. Emerging work has focused on scaling NGS-based profiling strategies to guide precision medicine approaches in clinical oncology using routine clinical biospecimens such as formalin-fixed, paraffin-embedded (FFPE) tissue or less-invasive liquid (e.g., blood or urine) samples. Technical challenges associated with limited tumor lesion size, low nucleic acid quantities, disease-specificity applications, and disease and histological heterogeneity present hurdles to widespread adoption and utility of extant NGS-based precision oncology approaches. Here, several analytical advances are described supporting democratization of precision oncology approaches from clinical tissue and liquid biospecimens, while revealing disease insights and important clinical considerations in the context of both localized and advanced (including multifocal and/or heterogeneous) disease. First, development and validation of a targeted DNA and RNA NGS assay compatible with small quantities of DNA and RNA isolated from routine, archived FFPE tissue specimens is described. This assay, targeting recurrently mutated oncogenic hotspots, tumor suppressors, copy-number-altered genes, and recurrent gene fusions is applied to a cohort of >300 FFPE tissue samples, revealing high sensitivity with orthogonal molecular diagnostic assays for *BRAF, KRAS,* and *EGFR* oncogenic alterations. Second, I describe a rapid, inexpensive, low-pass cell-free DNA (cfDNA) whole-genome sequencing (WGS) copy-number profiling approach, including a novel heuristic

tumor content approximation method, capable of establishing genome-wide copy-number profiles from 0.01-0.1x sequencing coverage. Application of our approach in plasma samples from patients with advanced cancer with matched comprehensive tissue NGS revealed high concordance with tissue-based molecular profiles, while highlighting important areas of potential utility from noninvasive profiling of overall disease burden. Third, I describe the systematic assessment of expression-based molecular subtypes in histologically heterogenous bladder cancers, revealing robust identification of basal/luminal molecular subtypes in a cohort of >100 bladder cancer cell lines and tumor tissue specimens, and recapitulation of basal/luminal subtypes in >400 samples profiled by TCGA using selected marker subsets. Importantly, I describe divergent expression profiles in the context of shared genomic alterations for individual histologically divergent tumor components from the same tumor, confounding proposed clinical utility of expression-based subtypes for disease prediction and prognosis. Fourth, I describe the development of a targeted RNAseq panel capable of assessing major transcriptional programs and disease biomarkers across the full spectrum of prostate cancer disease, while deriving commercially available prognostic scores that show limited robustness to disease multifocality. Lastly, I describe extensions of our cfDNA WGS approach to urine cfDNA samples from patients with advanced cancer, while exploring the potential utility of pairing described analytic tools with existing and emerging molecular profiling strategies to improve our understanding of disease biology and maximize clinical utility.

CHAPTER I: Introduction

Previously published in *The Cancer Journal*, co-authored with Tomlins, S.A.

Comprehensive next-generation sequencing (NGS) of primary prostate cancer and castration-resistant prostate cancer (CRPC) has provided a foundational understanding of the prostate cancer genomic and transcriptomic landscape, elucidating key biological and molecular components of progression and potential therapeutic opportunities[1, 2]. NGS-based profiling of CRPC has identified the most frequent molecular alterations in advanced, treatment refractory disease, as well as demonstrated the unique therapeutic challenges in using molecular information to guide treatment[1]. At present, NGS-based profiling can enable relatively fast, accurate, and comprehensive assessment of driving genomic and transcriptomic alterations in advanced cancer. However, prostate cancer is a dynamic, inherently heterogeneous disease, and within this context, considerable challenges remain around how best to leverage NGS-based screening, prognostic, and disease monitoring strategies in the context of current standards of care [3]. Here we review some of the key NGS-based approaches and findings that are enabling the tracking of the evolution of metastatic CRPC, including applications for informing treatment, and explore challenges for prospective implementation of NGS-based assays aimed at guiding precision medicine approaches for CRPC.

Genomic/Transcriptomic Landscape and NGS Profiling in CRPC

Multiple recent large-scale sequencing studies have helped to characterize the diverse genomic and transcriptomic landscape of both primary prostate cancer and CRPC, as well as small cell/neuroendocrine prostatic carcinoma (NePC)[1, 2, 4-6]. These studies have leveraged comprehensive DNA and RNA sequencing of fresh frozen tissue samples, describing a heterogeneous set of somatic alterations present in CRPC and/or NePC, including those enriched or unique in CRPC or NePC compared to primary disease [1, 2]. Alterations of particular relevance include frequent adaptive *AR* amplifications and mutations often conferring resistance to first and second generation anti-androgen therapies, *TP53* and *RB1* mutations and deletions particularly in NePC, and an increased prevalence of germline or somatic alterations in DNA repair pathway genes in CRPC[1, 2, 4-6]. Comprehensive RNA sequencing of advanced prostate cancers have also been recently reported, building on prior expression profiling studies of CRPC[2, 4-9]. Sequencing based approaches for assessing the CRPC transcriptome may have particular relevance given that the presence of *AR* splice variants in both primary and advanced prostate cancer may lead to increased resistance to second generation anti-androgens[10]. Overall, these sequencing initiatives have helped to outline the feasibility and efficacy of comprehensive (whole genome, whole exome) sequencing-based profiling of CRPC patients in large-scale single- or multi-institutional collaborations[1].

Technical challenges persist, however, for widespread prospective implementation of comprehensive NGS based profiling of patients with advanced prostate cancer. Access to fresh frozen tissue biopsy samples is often limited, leaving formalin-fixed, paraffin-embedded (FFPE) tissue samples as the primary source analyte for many sequencing-based assays[11-13]. Whole genome or transcriptome scale sequencing of routine FFPE clinical core biopsy samples has proven challenging[14]. Further, even when fresh frozen tissue is obtainable, it is still generally

cost-prohibitive for many clinical centers to deploy comprehensive genomic and transcriptomic NGS-based profiling of CRPC samples in a prospective fashion [12, 15]. Additionally, routine biopsy sampling of metastases in patients with advanced disease is not always performed given the utility of serum PSA as a recurrence/response biomarker, limiting tissue availability for widespread understanding of molecular relationships between primary and metastatic lesions and hindering development of personalized treatment approaches for individuals with CRPC[16].

Several groups have shown that targeted DNA and RNA sequencing of FFPE tissue samples may be a feasible strategy for profiling clinically relevant somatic alterations in both primary and advanced prostate cancer [11-13]. Targeted strategies have shown promise in assaying recurrently altered oncogenes and tumor suppressors, genes with frequent copy number alterations, and driver gene fusions such as *TMPRSS2-ERG* in order to identify the salient driving molecular alterations present in an advanced prostate cancer. Both targeted and more comprehensive approaches have also proven effective at identifying alterations that define well-established prostate cancer subtypes, including samples with ETS family gene fusions as well as those with *SPOP* mutations, *SPINK1* overexpression, *CHD1* mutations and deletions, and *IDH1* mutations [1, 2, 6, 17-19]. Given the initial success in tissue-based targeted sequencing of CRPC, some have even proposed strategies for monitoring disease via rebiopsy of lesions profiled pre- and post-treatment paired with NGS profiling[8]. However, these targeted and comprehensive approaches all require repeat invasive procedures for individual patient tracking, presenting limited feasibility for widespread clinical implementation, particularly in an era where biopsy of metastatic lesions to obtain material for molecular testing is not routinely reimbursed.

For this reason, recent efforts reporting efficacy of non-invasive NGS-based approaches to identify and track clinically relevant somatic alterations over time within patients with CRPC

may be particularly relevant in the near term[17, 20-24]. These approaches have shown that by targeted and more comprehensive sequencing of cell-free DNA (cfDNA), somatic point mutations, insertions/deletions, and copy-number alterations can be detected across a broad spectrum of tumor-derived cfDNA fractions. Alterations detected have highlighted or confirmed a number of important resistance mechanisms that emerge over the course of anti-androgen treatment (including both *AR* amplifications and point mutations), as well as suggesting that *AR* amplification alone may be a strong predictor of resistance to second generation anti-androgens abiraterone and enzalutamide[17, 21-23]. Perhaps most importantly, this work has demonstrated dynamic temporal changes in circulating tumor DNA fractions in cfDNA representing different tumor subclones over the course of anti-androgen treatment, hinting at myriad molecular changes in primary and metastatic lesions occurring in response to substantial therapeutic and fitness-related selective pressures[21].

Further work suggesting utility of whole exome and RNA sequencing from circulating tumor cells (CTCs) in patients with advanced prostate cancer has also been reported [25, 26], however the clinical utility of these approaches have not been fully investigated. Overall, these non-invasive approaches represent an important first step in understanding the dynamic nature of tumor clone and subclone representation detectable in the blood, as well as identifying technical hurdles that must be overcome for widespread clinical use of non-invasive NGS-based monitoring of molecular alterations in patients with advanced disease. Substantive work is required to enhance the sensitivity of these non-invasive approaches and validate the prognostic utility of these tools in personalizing patient care for individuals with advanced prostate cancer. Of note, very focused assays (including single gene assays) may be the final clinical assays used

after more discovery based NGS approaches have defined critical alterations, such as RT-PCR based assays for ARv7 expression in CTCs[10].

Intertumoral heterogeneity

Despite the broad characterization of the genomic and transcriptomic landscape of castrate-resistant prostate cancer (CRPC) and efforts to non-invasively characterize molecular alterations during treatment, a complete understanding of the intra-patient progression from localized primary prostate cancer to metastatic castrate-resistant disease remains elusive, limited primarily by the long timeline of typical prostate cancer progression that complicates longitudinal sample collection. Complicating the long arc of disease progression is the relatively recent discovery of substantial intra- and inter-individual heterogeneity for patients with metastatic disease, which may complicate development of personalized approaches to CRPC treatment, particularly for AR based therapies[3, 5, 27]. Prostate cancer is an inherently multifocal disease[28], with recent reports describing multiple clonal expansions even within a single morphologic tumor focus[29, 30]. While multiple reports support the monoclonal origin of lethal metastatic CRPC[3, 5, 6, 31], recent evidence supports the potential of lethal metastases arising from one or several clones or subclones in the primary tumor [5, 27, 31]. Likewise recent work in heavily treated patients suggests there may be a more complex series of metastasis-to-metastasis or metastasis-to-surgical bed seeding events that enable widespread metastatic spread as well as elimination and recurrence of individual clones during treatment [3, 21, 32]. While truncal mutations are typically shared across most lesions, these reports suggest substantial inter-tumoral heterogeneity may be present in patients with CRPC, although the degree of “relevant”

heterogeneity is less well established[3, 5]. Nevertheless, this heterogeneity presents significant challenges for using NGS to inform clinical decision-making in patients with CRPC, primarily owing to the limited molecular resolution available from a single core biopsy sample of a particular lesion in an inherently multi-focal, heterogeneous disease. Likewise, use and interpretation of non-invasive NGS-based disease monitoring approaches must account for the heterogeneous mix of physical locations from which tumor derived cfDNA or CTCs being assayed were originally shed.

Prognostic and screening considerations

While NGS-based profiling has played a critical role in elucidating key components of prostate cancer biology and some aspects of disease progression, NGS-based prognostic assays are still limited. Existing tissue-based prognostic assays, including Oncotype DX, Prolaris, Promark and Decipher, use RT-PCR, protein expression, or genomewide expression arrays to determine gene/protein expression for their component markers[33]. Ultimately, orthogonal NGS-based validation of these assays, incorporation of DNA based alterations, and their robustness to multifocality and intratumoral heterogeneity will likely be necessary to further improve prostate cancer prognosis and prediction.

Meanwhile, prospective sequencing of select genes may be an important consideration for germline screening and advanced disease monitoring in patients at risk for primary or advanced prostate cancer. Germline alterations in *BRCA2* and *BRCA1* have been shown to increase the lifetime risk for prostate cancer [34-37] and germline *BRCA2* carriers show worse prognosis than non-*BRCA2* carriers [38]. DNA damage repair genes are also an important

consideration, particularly in advanced prostate cancer, as approximately 20% of CRPC patients have been shown to harbor germline and/or somatic alterations in DNA damage repair genes such as *BRCA1*, *BRCA2*, or *ATM* [1, 6, 11, 12]. With only ~3% of primary prostate cancer reporting germline or somatic alterations in *BRCA1* or *BRCA2*, there may need to be a particular focus in screening for or monitoring *BRCA1*/*BRCA2* alterations in men with previous primary prostate cancer diagnosis or at higher baseline risk for primary disease, particularly in light of the potential predictive nature of these alterations (see below). Sequencing of additional genes that predispose men to higher risk of prostate cancer (e.g., *HOXB13*) may also be warranted [39].

Neuroendocrine/small cell prostate cancer

NGS profiling has also informed on the subset of patients who develop AR-independent small cell/neuroendocrine prostate cancer (NePC)[4, 40]. The increasing relevance of NePC (whether due to selection by more potent AR signaling therapies or increased survival of patients with CRPC beyond AR driven disease) has led to investigation on both the morphologic and molecular characterization of this disease subtype[4, 12, 40, 41]. Importantly, both single gene and comprehensive NGS approaches support transdifferentiation as the typical mechanism of NePC development, where NePC is clonally related to preceding AR driven disease[4, 42, 43]. NePC, particularly small cell carcinoma, shows a unique transcriptional profile (typically AR signaling low, neuroendocrine gene expression high and proliferation high) as well as characteristic genomic alterations including *RB1* and *TP53* loss and *MYCN* (or *MYCL*) amplification[40, 44-46]. Of particular relevance, comprehensive NGS interrogation has demonstrated that typical adenocarcinoma and small cell carcinoma represent a spectrum, with

the opportunity for molecular assessment to complement clinicopathologic assessment in determining treatment strategies[4, 12, 41, 46, 47].

Clinical trial design

NGS-based molecular stratification strategies have emerged as a way to more intelligently enroll patients most likely to benefit in targeted therapy clinical oncology trials. However, recent reports indicate only 2% of all clinical trials enrolling patients with prostate cancer from September 2011 to September 2014 used biomarkers or molecular alterations to select patients for trial enrollment [48]. Conversely, the 20% of CRPC tumors showing germline or somatic alterations in DNA damage repair genes (most frequently *BRCA2*, *BRCA1*, or *ATM*) carry clear implications for ongoing and prospective clinical trial design, given the success of poly ADP ribose polymerase (PARP) inhibitors in BRCA-deficient advanced breast and ovarian cancers[49]. Of particular note, Mateo *et al*. recently reported a phase II study of PARP inhibition with olaparib in metastatic CRPC, with response rates >80% in cases with germline or somatic alterations in DNA damage repair genes (*BRCA2*, *BRCA1*, *ATM*, *CHEK2*, *FANCA*, and *PALB2*) compared to 6% in patients without DNA damage repair gene alterations[50], leading to breakthrough status. This study underscores the benefit for employing NGS assay guided patient selection for clinical trial design, where even rare potentially targetable alterations (e.g. those in RAF family members and IDH1) can be assessed enabling umbrella or basket trials, similar to the approach taken by the NCI-MATCH trial (NCT02465060), where NGS from metastatic FFPE samples guides enrollment on patient-specific molecular alterations.

Additional NGS-based Applications in Treatment of CRPC

Given the reported inter-tumoral heterogeneity and temporal changes in circulating DNA from tumor subclones in response to therapeutic pressures, utilizing molecular sequencing to improve prostate cancer prognostication and therapeutic prediction may be particularly challenging [51]. Challenges including technical limitations, tissue availability, the inherent biological variability in prostate cancer and the established utility (and known limitations) of serum PSA mean that serial monitoring and disease tracking in patients at risk or with CRPC is still a fledgling enterprise. For patients on anti-androgen therapy, PSA monitoring and imaging are typically used as a primary metric for response to treatment, but we expect prospective NGS-based tracking strategies may improve sensitivity in screening for and detecting genomic & transcriptomic alterations – including *AR* mutations, splice variants, and amplifications – signaling the start of or susceptibility to treatment resistance at earlier time points than existing strategies [52]. It must be stressed however, that the clinical adoption of NGS to detect recurrence or resistance based on ultrasensitive detection of molecular alterations will require proven benefit of initiating/changing therapy at that time vs. waiting for clinical progression.

Although comprehensive NGS is critical to characterize the molecular landscape of CRPC, we anticipate that small, customized targeted sequencing panels compatible with DNA or RNA isolated from tissue, blood, or urine will prove invaluable for the eventual treatment guidance and monitoring of disease- or progression-associated alterations in patients with CRPC, much like those employed in recent reports [17, 21]. Alternatively, some groups have reported utility in using low coverage whole-genome sequencing to screen cfDNA in patients with CRPC for clinically informative copy number alterations (including *AR* amplifications), a strategy

which could help complement a more targeted NGS approach given the high prevalence of driving copy-number alterations in CRPC [1, 24].

Recent discoveries have also characterized a series of long non-coding RNAs (lncRNAs) associated with aggressive prostate cancer, most notably *SChLAP1*, which is prognostic in localized prostate cancer[7, 9, 53], and the landscape of lncRNAs in CRPC remains poorly described. Together with work summarizing the expression of myriad *AR* splice variants (several of which may confer resistance to second-generation anti-androgens) in both primary and advanced prostate cancer[10, 54], these reports highlight a potential key role for serial RNA-based NGS profiling in guiding treatment of patients with CRPC. However, established clinical benefit associated with these newly discovered mechanisms and biomarkers is still being explored in ongoing trials, and systematic validation of the clinical and prognostic utility is warranted prior to widespread implementation.

Epigenomic analyses in localized and advanced prostate cancer have also reported preliminary evidence supporting the role of epigenetic alterations as potential biomarkers for both aggressive and castrate-resistant prostate cancer, however limited work has been carried out to determine whether these markers can be reliably detected non-invasively[55]. These analyses have, however, helped to identify the role that epigenetic AR co-activators such as *TIF2*, *p300*, *CBP*, and *EZH2* play in CRPC, nominating important candidates for NGS-based gene expression profiling over the course of disease [55, 56]. Ultimately, the prognostic ability for proposed epigenetic biomarkers will require more systematic evaluation before being considered for use in guiding treatment decisions in CRPC.

Conclusions

In the near term future, tissue-based NGS profiling coupled with non-invasive (cfDNA- or CTC-based) NGS profiling will likely present a powerful approach for capturing a relatively complete assessment of the intra- and inter-tumoral molecular heterogeneity present in patients with advanced cancers (including prostate) and identifying the most promising treatment hypotheses. Precision oncology (e.g., molecular profiling of disease within an individual to better tailor personalized treatment decisions) is primarily employed in this advanced cancer context, making valid and robust assessment of clinically informative molecular alterations and prognostic biomarkers from relevant clinical biospecimens an extremely important consideration. While standard current practice dictates broad disease inferences are typically made from comprehensive profiling of bulk tissue specimens, scalable strategies for sequencing-based profiling of more routine biospecimens (including formalin-fixed paraffin embedded (FFPE) tissue specimens, blood, and urine) are starting to emerge that may offer several advantages to bulk tissue specimen-based approaches. While challenges remain around isolation of hiqh-quality nucleic acid from these tissue samples yielding quantities amenable to sequencing-based profiling, leveraging both routine tissue and more readily available liquid biospecimens is quickly becoming an attractive approach that may complement comprehensive characterization of driving molecular alterations from bulk tissue.

Accordingly, this dissertation summarizes several advances supporting such scalable precision oncology approaches from routine tissue and liquid biopsy specimens. In Chapter II, I describe the validation and application of a targeted next-generation assay compatible with minute quantities of DNA and RNA isolated from formalin-fixed paraffin-embedded tissue samples in a cohort of >300 clinical tissue specimens. This assay has been extended to a number

of unique and underprofiled cancers in projects I have led through my graduate work[57, 58], and applications of this panel when paired with cell-free DNA whole-genome sequencing (Chapter III) and innovative tissue-based targeted RNA sequencing approaches (Chapters IV and V) are described herein.

In Chapter III, I describe a pan-cancer, rapid, inexpensive low-pass plasma cell-free DNA whole genome sequencing approach capable of establishing genome-wide copy-number profiles, identifying therapeutically-relevant focal copy-number alterations, and facilitating heuristic tumor content estimation to inform next-generation sequencing workflows. Extensions of this approach across multiple advanced cancers are described, and utility as a "screening" tool for precision oncology workflows is explored. Systematic concordance of plasma cfDNA profiles with comprehensive profiles from matched bulk tissue specimens is described, while ongoing work applied to individual and serial urine cell-free DNA samples in patients with advanced cancer, with novel urine tumor-specific cell-free DNA is discussed in Appendix A.

Given the purported clinical utility of expression-based molecular subtypes and prognostic classifiers for making disease predictions and stratifying therapies across a number of cancer types, I also sought to test whether these expression-based tools provide consistent results in the context of heterogeneous or multifocal disease. Chapter IV explores derivation of expression-based molecular subtypes in bladder cancer (a frequently histologically heterogeneous disease) using consensus clustering of expression values assessed via a custom targeted RNA sequencing panel, with a particular focus on pairing genomic profiling and expression-based basal/luminal subtyping in a series of paired histologically divergent components from the same tumor across a series of cases. In Chapter V, I demonstrate derivation of commercially-available prognostic expression-based scores used in patients with prostate

cancer from a targeted RNA sequencing panel, and explore whether these scores are robust to routinely multifocal disease. Chapter VI (and Appendix A) summarize ongoing work, and describe ways the analytic tools presented in this dissertation can be paired with existing and emerging molecular profiling strategies to improve our understanding of disease biology with maximal clinical utility.

Figure 1.1: Next Generation Sequencing in Castration Resistant Prostate Cancer Treatment

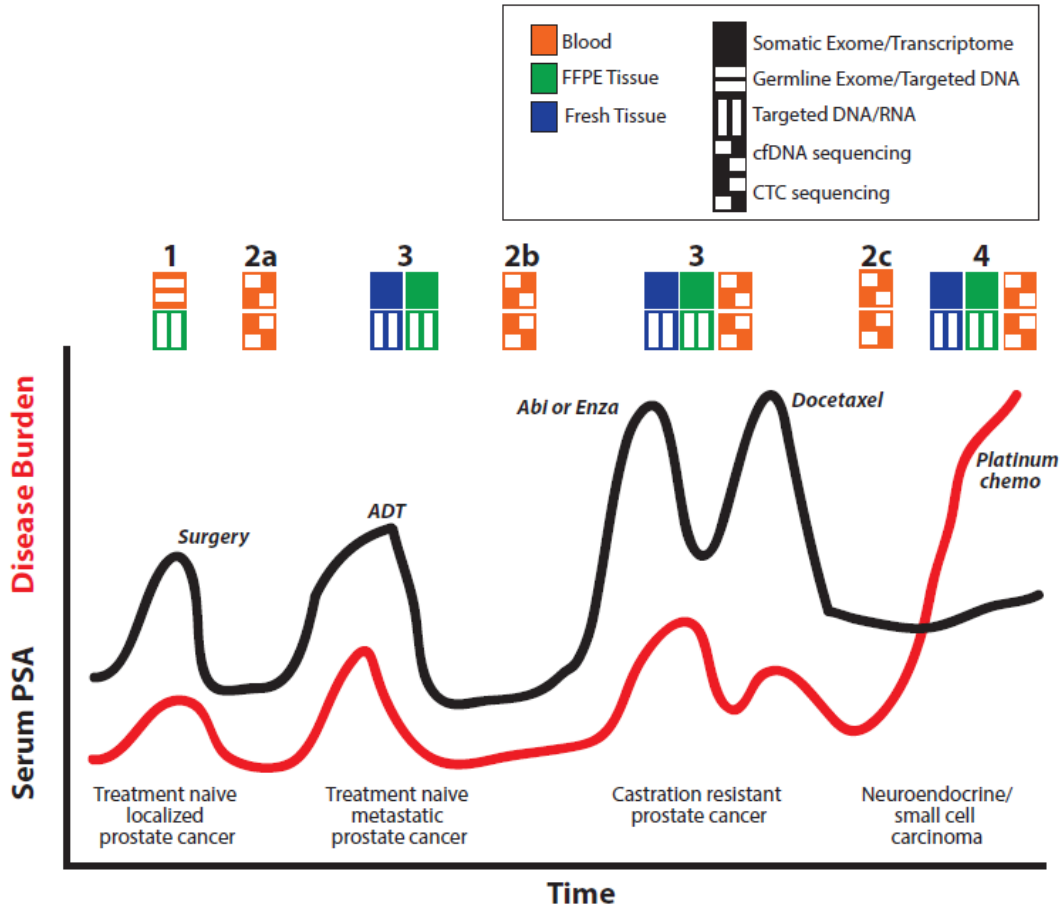


Figure 1.1. Potential clinical utility of next generation sequencing (NGS) during prostate cancer progression. A timeline of serum PSA (black line) and disease burden (red line) along with treatments (italics) are shown for a hypothetical patient who progresses from localized untreated prostate cancer diagnosed and treated by radical prostatectomy to untreated treatment naïve metastatic prostate cancer to castration resistant prostate cancer (CRPC) and eventually neuroendocrine/small cell carcinoma. Opportunities for NGS to guide clinical management are shown above the graph according to the biocompartment assessed (color of the box) and NGS approach (pattern of the box) as indicated in the legend. At diagnosis (1), germline NGS assessment may be utilized to identify predisposing germline variants that may inform on later therapy and identify hereditary predisposition. Likewise, targeted DNA and RNA based assessment of FFPE biopsy and/or prostatectomy tissues may be used for prognosis and assessment of presumed clonal alterations that can be tracked and/or targeted during progression. NGS of cfDNA and/or CTCs isolated from blood may be used for non-invasive assessment of disease recurrence (2a) and assessment of clonal dynamics upon treatment. Diagnosis of metastatic disease by biopsy enables targeted DNA and RNA assessment of FFPE tissue (or comprehensive assessment if fresh tissue is obtained [most likely in the translational research setting]), and may have utility in predicting response to ADT or enrollment on clinical trials in the castration sensitive space (3). In addition to monitoring for development of CRPC after ADT (2a), NGS of cfDNA and/or CTCs may have particular utility for predicting response to second generation anti-androgens (such as abiraterone [abi] or enzalutamide [enza]) based on assessment of AR amplifications, mutations, or splice variant expression. Likewise, targeted or comprehensive NGS of CRPC biopsy tissue may have utility for identifying resistance mechanisms, novel targetable alterations, and identification of alterations enabling enrollment on umbrella and/or basket studies (3). NGS assessment of cfDNA and/or CTCs may be useful as a non-invasive complement to serum PSA to identify the development of AR independent clones (2c) and neuroendocrine/small cell prostate carcinoma when serum PSA may not be an accurate measurement of disease burden. Lastly, NGS of neuroendocrine/small cell prostate carcinoma (4) tissue may identify potential novel targetable alterations that developed during progression.

Chapter I References

1. Robinson, D., et al., *Integrative clinical genomics of advanced prostate cancer.* Cell, 2015. **161**(5): p. 1215-28.
2. Cancer Genome Atlas Research, N., *The Molecular Taxonomy of Primary Prostate Cancer.* Cell, 2015. **163**(4): p. 1011-25.
3. Gundem, G., et al., *The evolutionary history of lethal metastatic prostate cancer.* Nature, 2015. **520**(7547): p. 353-7.
4. Beltran, H., et al., *Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer.* Nat Med, 2016. **22**(3): p. 298-305.
5. Kumar, A., et al., *Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer.* Nat Med, 2016. **22**(4): p. 369-78.
6. Grasso, C.S., et al., *The mutational landscape of lethal castration-resistant prostate cancer.* Nature, 2012. **487**(7406): p. 239-43.
7. Prensner, J.R., et al., *RNA biomarkers associated with metastatic progression in prostate cancer: a multi-institutional high-throughput analysis of SChLAP1.* Lancet Oncol, 2014. **15**(13): p. 1469-80.
8. Rajan, P., et al., *Identification of a candidate prognostic gene signature by transcriptome analysis of matched pre- and post-treatment prostatic biopsies from patients with advanced prostate cancer.* BMC Cancer, 2014. **14**: p. 977.
9. Ylipaa, A., et al., *Transcriptome Sequencing Reveals PCAT5 as a Novel ERG-Regulated Long Noncoding RNA in Prostate Cancer.* Cancer Res, 2015. **75**(19): p. 4026-31.
10. Antonarakis, E.S., et al., *AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer.* N Engl J Med, 2014. **371**(11): p. 1028-38.
11. Beltran, H., et al., *Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity.* Eur Urol, 2013. **63**(5): p. 920-6.
12. Hovelson, D.H., et al., *Development and validation of a scalable next-generation sequencing system for assessing relevant somatic variants in solid tumors.* Neoplasia, 2015. **17**(4): p. 385-99.
13. Cheng, H.H., et al., *A pilot study of clinical targeted next generation sequencing for prostate cancer: Consequences for treatment and genetic counseling.* Prostate, 2016.
14. Cieslik, M., et al., *The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing.* Genome Res, 2015. **25**(9): p. 1372-81.
15. Van Allen, E.M., et al., *Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine.* Nat Med, 2014. **20**(6): p. 682-8.
16. Beltran, H. and M.A. Rubin, *New strategies in prostate cancer: translating genomics into the clinic.* Clin Cancer Res, 2013. **19**(3): p. 517-23.
17. Romanel, A., et al., *Plasma AR and abiraterone-resistant prostate cancer.* Sci Transl Med, 2015. **7**(312): p. 312re10.
18. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.
19. Tomlins, S.A., et al., *The role of SPINK1 in ETS rearrangement-negative prostate cancers.* Cancer Cell, 2008. **13**(6): p. 519-28.

20. Azad, A.A., et al., *Androgen Receptor Gene Aberrations in Circulating Cell-Free DNA: Biomarkers of Therapeutic Resistance in Castration-Resistant Prostate Cancer.* Clin Cancer Res, 2015. **21**(10): p. 2315-24.
21. Carreira, S., et al., *Tumor clone dynamics in lethal prostate cancer.* Sci Transl Med, 2014. **6**(254): p. 254ra125.
22. Wyatt, A.W., et al., *Genomic Alterations in Cell-Free DNA and Enzalutamide Resistance in Castration-Resistant Prostate Cancer.* JAMA Oncol, 2016.
23. Lallous, N., et al., *Functional analysis of androgen receptor mutations that confer anti-androgen resistance identified in circulating cell-free DNA from prostate cancer patients.* Genome Biol, 2016. **17**: p. 10.
24. Ulz, P., et al., *Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer.* Nat Commun, 2016. **7**: p. 12008.
25. Lohr, J.G., et al., *Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer.* Nat Biotechnol, 2014. **32**(5): p. 479-84.
26. Miyamoto, D.T., et al., *RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance.* Science, 2015. **349**(6254): p. 1351-6.
27. Haffner, M.C., et al., *Tracking the clonal origin of lethal prostate cancer.* J Clin Invest, 2013. **123**(11): p. 4918-22.
28. Wise, A.M., et al., *Morphologic and clinical significance of multifocal prostate cancers in radical prostatectomy specimens.* Urology, 2002. **60**(2): p. 264-9.
29. Boutros, P.C., et al., *Spatial genomic heterogeneity within localized, multifocal prostate cancer.* Nat Genet, 2015. **47**(7): p. 736-45.
30. Cooper, C.S., et al., *Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue.* Nat Genet, 2015. **47**(4): p. 367-72.
31. Liu, W., et al., *Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer.* Nat Med, 2009. **15**(5): p. 559-65.
32. Hong, M.K., et al., *Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer.* Nat Commun, 2015. **6**: p. 6605.
33. Ross, A.E., A.V. D'Amico, and S.J. Freedland, *Which, when and why? Rational use of tissue-based molecular testing in localized prostate cancer.* Prostate Cancer Prostatic Dis, 2016. **19**(1): p. 1-6.
34. Thompson, D., D.F. Easton, and C. Breast Cancer Linkage, *Cancer Incidence in BRCA1 mutation carriers.* J Natl Cancer Inst, 2002. **94**(18): p. 1358-65.
35. Liede, A., B.Y. Karlan, and S.A. Narod, *Cancer risks for male carriers of germline mutations in BRCA1 or BRCA2: a review of the literature.* J Clin Oncol, 2004. **22**(4): p. 735-42.
36. Tischkowitz, M., et al., *Mutations in BRCA1 and BRCA2 and predisposition to prostate cancer.* Lancet, 2003. **362**(9377): p. 80; author reply 80.
37. Ostrander, E.A. and M.S. Udler, *The role of the BRCA2 gene in susceptibility to prostate cancer revisited.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(8): p. 1843-8.
38. Attard, G., et al., *Prostate cancer.* Lancet, 2016. **387**(10013): p. 70-82.
39. Ewing, C.M., et al., *Germline mutations in HOXB13 and prostate-cancer risk.* N Engl J Med, 2012. **366**(2): p. 141-9.
40. Beltran, H., et al., *Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets.* Cancer Discov, 2011. **1**(6): p. 487-95.

41. Epstein, J.I., et al., *Proposed morphologic classification of prostate cancer with neuroendocrine differentiation.* Am J Surg Pathol, 2014. **38**(6): p. 756-67.
42. Nadal, R., et al., *Small cell carcinoma of the prostate.* Nat Rev Urol, 2014. **11**(4): p. 213-9.
43. Kadakia, K.C., et al., *Comprehensive serial molecular profiling of an "N of 1" exceptional non-responder with metastatic prostate cancer progressing to small cell carcinoma on treatment.* J Hematol Oncol, 2015. **8**(1): p. 109.
44. Beltran, H., et al., *Molecular Characterization of Neuroendocrine Prostate Cancer and Identification of New Drug Targets.* Cancer Discov, 2011. **1**(6): p. 487-495.
45. Grasso, C.S., et al., *Integrative molecular profiling of routine clinical prostate cancer specimens.* Ann Oncol, 2015. **26**(6): p. 1110-8.
46. Aggarwal, R., et al., *Neuroendocrine prostate cancer: subtypes, biology, and clinical outcomes.* J Natl Compr Canc Netw, 2014. **12**(5): p. 719-26.
47. Aparicio, A.M., et al., *Platinum-Based Chemotherapy for Variant Castrate-Resistant Prostate Cancer.* Clin Cancer Res, 2013.
48. Khemlina, G., S. Ikeda, and R. Kurzrock, *Molecular landscape of prostate cancer: implications for current clinical trials.* Cancer Treat Rev, 2015. **41**(9): p. 761-6.
49. Lord, C.J., A.N. Tutt, and A. Ashworth, *Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors.* Annu Rev Med, 2015. **66**: p. 455-70.
50. Mateo, J., et al., *DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer.* N Engl J Med, 2015. **373**(18): p. 1697-708.
51. Spratt, D.E., et al., *Translational and clinical implications of the genetic landscape of prostate cancer.* Nat Rev Clin Oncol, 2016.
52. Prensner, J.R., et al., *Beyond PSA: the next generation of prostate cancer biomarkers.* Sci Transl Med, 2012. **4**(127): p. 127rv3.
53. Prensner, J.R., et al., *The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex.* Nat Genet, 2013. **45**(11): p. 1392-8.
54. Sprenger, C.C. and S.R. Plymate, *The link between androgen receptor splice variants and castration-resistant prostate cancer.* Horm Cancer, 2014. **5**(4): p. 207-17.
55. Valdes-Mora, F. and S.J. Clark, *Prostate cancer epigenetic biomarkers: next-generation technologies.* Oncogene, 2015. **34**(13): p. 1609-18.
56. Xu, K., et al., *EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent.* Science, 2012. **338**(6113): p. 1465-9.
57. Warrick, J.I., et al., *Tumor evolution and progression in multifocal and paired non-invasive/invasive urothelial carcinoma.* Virchows Arch, 2015. **466**(3): p. 297-311.
58. McDaniel, A.S., et al., *Genomic Profiling of Penile Squamous Cell Carcinoma Reveals New Opportunities for Targeted Therapy.* Cancer Research, 2015. **75**(24): p. 5219-5227.

CHAPTER II: Development and Validation of a Scalable Next-Generation Sequencing System for Assessing Relevant Somatic Variants in Solid Tumors

Previously published in *Neoplasia*, co-authored with McDaniel AS, Cani AK, et al**.** http://dx.doi.org/10.1016/j.neo.2015.03.004

Contributions: Dr. McDaniel procured and reviewed slides, facilitating tissue scraping for DNA and RNA isolation. Mr. Cani isolated all DNA and RNA for profiled tissue specimens, and prepared all sequencing libraries and carried out necessary sequencing. I collated, annotated, and filtered all point mutation and indel calls, copy number alterations, and gene fusion calls, and carried out all analyses reported in the main manuscript and supplementary materials.

INTRODUCTION

Precision medicine approaches, where patients are treated with therapies directed against the specific molecular alterations driving their tumors, have revolutionized oncology [1-4]. Such approaches require identification of driving molecular alterations (which may occur only in a subset of a given histologic cancer type or in cancers arising from diverse organs), development of targeted therapies, and diagnostic tests to identify appropriate patient populations for clinical trials and eventual implementation [5-7]. The early successes of trastuzumab (a monoclonal antibody against ERBB2) in the subset of breast adenocarcinomas with *ERBB2* amplifications[8], and imatinib (an ABL kinase inhibitor) in the subset of leukemia driven by *BCR-ABL* gene fusions (chronic myeloid leukemia)[9], have been replicated in numerous cancers [1-4]. For example, multiplexed assessment of driving somatic alterations in lung cancer has been shown to aid in physician selection of therapy, and patients with drivers receiving a matched therapy lived significantly longer than those not receiving a matched therapy[10].

Recent advances in genome sciences, including next generation sequencing (NGS), have led to the identification of hundreds of recurrent somatically altered genes through the analysis of tens of thousands of cancer samples from individual investigators and large consortia, such as The Cancer Genome Atlas (TCGA)[11-15]. These technological advances are also changing routine molecular pathology practice from single gene based tests (i.e. Sanger sequencing to assess *EGFR* mutations in lung adenocarcinoma) to multiplexed NGS assays. Several NGS approaches have been successfully clinically implemented in oncology, including multiplexed PCR based panels assessing tens to hundreds of genes, hybrid capture based panels targeting hundreds of genes, as well as comprehensive exome/genome/transcriptome sequencing[16-26]. These approaches vary in sample requirements, nucleic acids assessed, cost, throughput, genes and alteration types assessed, and performance. For example, most clinically implemented multiplexed PCR based approaches fail to assess copy number alterations (CNAs) and/or gene fusions [16, 19, 22, 25], which guide current treatment selection for several cancers.

The primary challenge with comprehensive NGS approaches, however, is the specialty infrastructure and expertise needed to interpret the results and convey treatment strategies to clinicians. Several centers using comprehensive NGS-based oncology approaches require NGS-based tumor boards [21] to guide interpretation and inform clinical decision-making. Large companies have also been established with the goal of providing comprehensive NGS-based precision oncology services [18], however interpretation of results and prioritizing treatment strategies may still be outsourced. Scalability limitations hinder widespread adoption of such initiatives in reference laboratories.

In order to enable precision medicine approaches for all patients with cancer, rapid, inexpensive, scalable NGS solutions capable of assessing all classes of current and near term

clinically relevant targets (point mutations, short insertions/deletions [indels], CNAs and gene fusions) from routine formalin fixed paraffin embedded (FFPE) tissues are required. Such a technical solution must be coupled with a dynamic, scalable, analytical approach capable of prioritizing treatment options. To begin to address these challenges, we report the development and validation of the Oncomine Comprehensive Panel (OCP), a multiplexed PCR-based NGS assay and analytical system to identify and prioritize potential treatment strategies from predefined somatic solid tumor genome variants. The OCP is compatible with 20ng of FFPE isolated DNA and 15ng FFPE isolated RNA and bench top Ion Torrent sequencers. Demonstrating the potential for a scalable solution to enable widespread precision medicine oncology applications, the OCP will be utilized in the NCI Match Trial to assess 3,000 cancer samples for trial selection in a multi-arm umbrella study with sequencing conducted at multiple sites.

MATERIALS & METHODS

Analysis of relevant somatic variants in solid tumors

The Oncomine Comprehensive Panel (OCP) was designed to interrogate somatic mutations, CNAs and gene fusions involving oncogenes and tumor suppressors recurrently altered in solid tumors with the potential for near term clinical relevance. To define OCP content, we used evidence-based analysis of genomic alterations present in Oncomine, a resource comprised of mutation, copy number, and gene fusion data from >700,000 cell line, xenograft and clinical cancer samples as of December 2013[27-29]. Candidate genes with somatic driver mutations were derived from gain-of-function (GoF) and loss-of-function (LoF) analyses performed on 686,530 tumor samples with mutation data in Oncomine. Candidate driver CNA events were identified by performing a minimal common region (MCR) assessment on a pan-

cancer subset of 10,249 tumor samples in Oncomine[28]. In addition, single cancer type assessments were performed to identify private candidate copy number drivers. Candidate driver gene fusions were identified from the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman), as well as from analyzing 6,438 primary tumor sample RNA-seq profiles contained within Oncomine. Complete details on somatic variant analysis to define candidates are provided in **Appendix B**. All candidate driver GoF, LoF and CNA genes, as well as gene fusions, were then assessed for evidence of near term potential clinical relevance as defined in the **Appendix B**.

OCP NGS assay design

We developed multiplexed PCR (Ion Ampliseq) NGS panels to characterize DNA (mutations and CNAs) and RNA (gene fusions) based alterations. For GoF alterations, amplicons were designed to assess recurrent hotspots as defined above. For LoF alterations, amplicons were designed to tile the gene's entire coding sequence. For CNAs, sufficient amplicons (n=3 to 38) were designed from coding and noncoding regions to facilitate copy number profiling. For the RNA based panel, primers were included to detect known gene fusion junctions, assess the 5' and 3' regions of *RET*, *ROS1*, and *ALK* (to enable testing for fusions involving novel 5' fusion partners through 3'/5' expression imbalance), and quantify housekeeping/positive expression genes (*HMBS*, *ITGB7*, *LMNA*, *MYC*, and *TBP*); a small subset of the prostate cancer samples (n=12) were sequenced using a version of the RNA panel that did not contain the 3'/5' expression imbalance assays. Ampliseq panels were designed using Ampliseq Designer and multiplexed pools were obtained from Ion Torrent. Two versions of both the DNA and RNA

based panels were assessed herein during iterative optimization, and complete information on the panels is provided in **Table S1**.

Molecular Standards

We utilized two commercially available molecular standards to assess performance of the DNA component of the OCP. The AcroMetrix Oncology Hotspot Control (AOHC; Life Technologies) was designed to assess somatic mutation detection performance by NGS assays. The custom version used herein contained 365 applicable single/ multiple nucleotide variants (SNVs/MNVs) and 33 indels each at an estimated allele frequency of 0.20 on the GM24385 cell line genomic background. AOHC DNA was used directly for library preparation.

The Quantitative Multiplex Reference Standard (QMRS, Horizon Diagnostics, Cambridge, UK) consists of 1-3 FFPE tissue sections from multiplexed FFPE cell lines with a known set of 30 engineered and endogenous mutations present at specific variant allele frequencies quantified by ddPCR. Of the 30 mutations, 16 (all 11 primary engineered mutations, and 5 of 19 secondary endogenous mutations) were targeted by the OCP and were used for evaluation. QMRS tissue was processed as for the remaining tissue cohorts for DNA isolation and library preparation.

Tissue Cohorts

We used three cohorts of routine FFPE tissues for OCP evaluation (molecular [MO], lung [LU] and prostate [PR]). All FFPE specimens were obtained from the University of Michigan (UM) Department of Pathology Tissue Archive with IRB approval. Diagnostic hematoxlyin and

eosin (H&E) stained slides were reviewed by board certified Anatomic Pathologists (A.S.M. and S.A.T.).

The MO cohort consisted of all cancer specimens (including biopsy, resection and cell block specimens) sent during a five month period to the CLIA certified UM Molecular Oncology/Genetics Laboratory for 1) *EGFR*, *BRAF* or *KRAS* mutation testing or 2) *ALK* rearrangement testing. Complete details of the MO cohort and clinicopathologic information for all cases is provided in **Appendix B**. The LU and PR cohorts consisted of 104 and 118 retrospectively identified FFPE tissue specimens, respectively. A subset of the PR samples (*n*=37) have previously been assessed by a combined capture based NGS (Agilent Haloplex followed by Ion Torrent NGS) and Taqman low density array q-RT-PCR panel[30]. Clinicopathologic information for all included cases in the LU and PR cohorts is provided in **Appendix B**. Targeted next generation sequencing of all tumor tissues was performed with IRB approval.

Nucleic acid isolation

For each specimen, 3-10 x 10um FFPE sections were cut from a single representative block per case, using macrodissection with a scalpel as needed to enrich for tumor content. DNA and RNA were isolated using the Qiagen Allprep FFPE DNA/RNA kit (Qiagen, Valencia, CA) as described[31]. DNA and RNA were quantified using the Qubit 2.0 fluorometer (Life Technologies, Foster City, CA).

DNA/RNA libraries

DNA/RNA libraries were generated essentially as described[31, 32]. DNA libraries were generated from 20ng of DNA per sample using the Ion Ampliseq library kit 2.0 (Life Technologies, Foster City, CA) and the OCP Ampliseq panel according to manufacturer's instructions with barcode incorporation. RNA libraries were generated from 15ng of RNA per sample using the Ion Ampliseq RNA Library kit. OCP Ampliseq Libraries were quantified using the Ion Library Quantification Kit according to the manufacturer's instructions.

Template generation and sequencing

Templates for DNA and RNA libraries were prepared using the Ion PGM Template OT2 200 Kit (Life Technologies, Foster City, CA) on the Ion One Touch 2 according to the manufacturer's instructions. Sequencing of multiplexed templates was performed using the Ion Torrent Personal Genome Machine (PGM) on Ion 318 chips using the Ion PGM Sequencing 200 Kit v2 (Life Technologies, Foster City, CA) according to the manufacturer's instructions. For the LU and PR cohorts, a single DNA template and 4-8 RNA templates were assessed separately on a single 318 chip. For the MO cohort, a single DNA template was combined with a single RNA template in a 4 to 1 ratio and assessed on a single 318 chip. For experiments with molecular standards, single DNA templates were assessed on one 318 chip.

Data analysis

Data analysis was performed using Torrent Suite (4.2.0) and the Coverage Analysis (or Coverage Analysis RNA) Plug-ins (both v4.0-r73765), along with the Ion Reporter (4.2.0) Fusion analysis workflow essentially as described [31, 32]. For DNA sequencing, alignment was performed using TMAP with default parameters, and variant calling was performed using the

Torrent Variant Caller plugin (version 4.2-8-r87740) using default low-stringency somatic variant settings. Somatic variant identification was performed essentially as described [31, 33] using read and base level filtering, which we have previously confirmed to identify variants that pass Sanger sequencing validation with >95% accuracy. Copy number analysis from total amplicon read counts provided by the Coverage Analysis Plug-in was performed essentially as described [19, 31, 32]. As an estimate of data quality, we determined the standard deviation of the amplicon-level copy number estimates relative to the gene-level estimate for each gene per sample (**Fig B1**). Gene fusion analysis was performed within the Ion Reporter (4.2.0) Fusion analysis workflow, with reads from the RNA AmpliSeq panel aligned using TMAP to a gene reference of targeted chimeric fusion transcripts as well as reference sequences for expression imbalance and expression control gene targets. Complete description of all data analysis is provided in **Appendix B**.

Alteration prioritization and potential actionability assessment

Somatic SNVs/indels passing filtering in a GoF gene were considered GoF if occurring at the predefined hotspot residue targeted in OCP. Somatic variants in a LoF gene were considered LoF if deleterious (nonsense or frame shifting) or occurring at a predefined hotspot residue. Somatic CNAs were considered for potential actionability analysis if they were concordant with the predicted alteration (amplification or deletion) from Oncomine analysis as described above. Somatic gene fusions were considered for actionability analysis if they represented known gene fusions from the Mitelman database or Oncomine analysis, or involved known 3' or 5' drivers with novel partners (i.e. *ERC1-BRAF* fusion in MO-17, with recurrent fusions involving *BRAF* as a 3' partner reported previously[34]).

These prioritized variants were then associated with potential actionability using the Oncomine database. Briefly, for each patient the "most actionable" alteration was identified by prioritizing 1) variants referenced in FDA drug labels, 2) variants referenced in NCCN treatment guidelines in the patient's cancer type, 3) variants referenced in an NCCN guideline in another cancer type, and 4) variants referenced as inclusion criteria in a clinical trial. Actionable variants were identified by manual curation of FDA labels, NCCN guidelines and by keyword searches and manual curation of clinical trial records in the TrialTrove database. Alterations associated with specific treatments are shown in **Appendix B.**

qRT-PCR and immunohistochemistry (IHC) validation

Details of qRT-PCR validation of *ERC1:BRAF* and *TPR:NTRK1* fusions, as well as ERBB2 IHC to confirm copy number gains are provided in the **Appendix B**.

Statistical tests

All statistical tests were performed in R (3.1.0) using two sided tests. *P*-values < 0.05 were considered statistically significant.

RESULTS

Oncomine Comprehensive Panel (OCP) development

To define relevant somatic cancer genome variants based on near term potential actionability, we first interrogated data from >700,000 tumor samples in the Oncomine database (including >8,000 exomes, >7,000 transcriptomes and >30,000 copy number profiles in addition to tumors studied by single gene/targeted approaches) to identify pan-cancer, recurrently altered

oncogenes (enriched in gain of function [GoF] hotspot mutations), tumor suppressors (enriched in loss of function [LoF] deleterious mutations), genes targeted by high level amplifications or deletions, and driving gene fusions (**Fig 2.1A**). Genes with these variants were then filtered based on near term potential actionability (see **Methods**). The distribution of these variants across >7,000 TCGA samples from 23 cancer types is shown in **Figure 2.1B**.

To translate the relevant somatic cancer genome to a NGS assay capable of detecting mutations, copy number alterations and gene fusions (including multiple splice isoforms) but compatible with limited amounts of routine FFPE tissues, we developed custom Ion Torrent multiplexed PCR based DNA and RNA sequencing (-seq) panels, together comprising the Oncomine Comprehensive Panel (OCP), as shown in **Figure 2.1C**. In total, the final OCP version assessed herein (v0.9b) interrogates 143 unique cancer genes including 73 oncogenes, 49 CNA genes, 26 tumor suppressor genes and 22 fusion driver genes. The targeted DNA-seq panel includes 2,530 amplicons covering 260,717 base pairs in 130 different genes. To minimize panel size and focus on predefined relevant alterations, only GoF mutations were targeted in oncogenes, while high level CNA genes were targeted by 3 to 38 probes to facilitate copy number profiling[19] and the entire coding sequence of tumor suppressors were targeted to identify LoF mutations and GoF mutations. The targeted RNA-seq panel included a total of 154 primer pairs targeting known gene fusion isoforms (n=148) as well as 5' and 3' expression assays for *RET*, *ROS1*, and *ALK* to enable novel fusion discovery through 3'/5' expression imbalance ratios. To enable appropriate normalization in downstream analyses, the targeted RNA-seq panel also includes 5 additional primer pairs targeting a pre-determined set of housekeeping/positive expression genes (*HMBS*, *ITGB7*, *LMNA*, *MYC*, and *TBP*). Details of the

two versions of the OCP panel (v0.9a and v0.9b) validated and applied herein are presented in **Appendix B**.

Molecular standards validation

To validate OCP performance, we first assessed the DNA component using the Acrometrix Oncology Hotspot Control molecular standard, a cell line DNA sample engineered to contain 398 OCP targeted variants at 0.20 expected variant allele frequencies. OCP detected 364 of 365 (99.7%) targeted single/multiple nucleotide variants (SNVs/MNVs), with a median variant allele frequency of 0.24 [interquartile range 0.21-0.28] as shown in **Fig B2**. Of the 33 OCP targeted indels, we detected 25 (75.8%) at a median variant allele frequency of 0.22 (interquartile range 0.18-0.28] (**Fig B2**). Of the 8 indels that were not detected, 3 were over 10 bases in length (12, 30, 41 bases) and 5 were single nucleotide insertions or deletions occurring within 2 to 7 base homopolymer runs. Accurate indel identification in homopolymer runs is a known challenge with current Ion Torrent sequencing technology[16].

We also profiled DNA isolated from commercially available FFPE sections containing a cell line mixture (QMRS cell line) with engineered and endogenous mutations at precise variant frequencies. In total, 16 known mutations in the QMRS cell line (11 primary induced and 5 endogenous mutations; median variant allele frequency 0.10, range 0.01-0.33) are targeted by the OCP. To prioritize high-quality somatic variants, we applied our standard filtering approach (which includes filters at <5% or <10% depending on alteration type, see **Methods**) to default variant calls. Ten of 16 (63%) OCP-targeted mutations (including 8/11 induced and 2/5 endogenous) were called by the Torrent Variant Caller (TVC) using our standard approach at variant allele frequencies highly correlated with those expected (r^2=0.99, **Appendix B**). Five of

the remaining 6 (83%) OCP-targeted mutations were detectable at close-to-expected frequencies—including indels and point mutations at <1-5% variant allele frequencies—via automated variant calling (**Appendix B**). The only variant not detected by OCP was a secondary *NF1* frameshift deletion at the start of a 6bp homopolymer run (expected frequency of 7.5%); hence, in total, 15 of 16 (94%) of OCP-targeted known mutations in the QMRS cell line were detected via our automated variant calling procedures (**Appendix B**). Highly concordant results were observed with both OCP versions, as well as separate Ion Torrent PGM template preparation and sequencing runs performed at two locations (Ann Arbor, MI and Carlsbad, CA) from aliquots of the same DNA library (**Appendix B**).

OCP performance in FFPE tumor tissue cohorts

To validate performance and demonstrate applicability, we applied the OCP to three cohorts of routine FFPE tissue specimens: a cohort comprised of tumor samples sent for routine molecular diagnostics (molecular [MO] cohort, *n*=105 samples), and retrospective lung cancer (LU, *n*=104 samples) and prostate cancer (PR, *n*=118 samples) cohorts. For each cohort, 3-10 x 10um FFPE sections were used for DNA/RNA co-isolation after macrodissection, with an overall average of 52% estimated post-dissection tumor content per sample (range 5%-90%), as assessed by histology (**Appendix B**). Across the MO, LU and PR cohorts, we isolated an average of 1.3/2.8ug, 2.0/3.3ug and 2.2/6.7ug DNA/RNA per sample, respectively. Overall, 32% of the FFPE specimens were at least 3 years old (average 2 years, [range <1 to 10 years]).

Multiplexed PCR based DNA and RNA libraries were generated from each sample for template preparation and NGS on the Ion Torrent PGM using Ion 318 chips. We excluded DNA and RNA libraries from 1/105 MO, 3/104 LU and 2/118 PR samples due to low quality libraries,

resulting in a total of 321/327 (98%) informative samples. An additional two samples (MO-46 and LU-141) were excluded from CNA analysis due to excessively noisy copy number profiles (**Fig B1**). Across the three cohorts, using the DNA panel, we achieved an average of 5,142,690 mapped reads (97% on-target), 1,941x coverage across targeted bases, 93.6% of targeted bases covered by at least 20 reads and 202 called variants per informative sample. Using the RNA panel, we achieved an average of 306,872 total mapped reads (including 210,712 reads mapped to the five housekeeping/positive expression genes) per sample. Complete sequencing statistics and DNA variants are provided in **Appendix B**. All high level OCP prioritized (see **Methods**) copy number alterations and gene fusions across the cohorts are also provided in **Appendix B**.

OCP validation in a clinical molecular diagnostics cohort

To validate the performance of the OCP and identify additional relevant variants beyond current routine practice, we assessed a cohort of 105 FFPE cancer samples sent for molecular testing for *EGFR*, *BRAF*, *KRAS* and *ALK* alterations in a Clinical Laboratories Improvement Ammendments/College of American Pathologists (CLIA/CAP) certified molecular diagnostics laboratory. The 104 informative MO samples from 104 patients were comprised of colorectal adenocarcinomas (*n*=29), lung adenocarcinomas (*n*=23), melanomas (*n*=48) and 4 other cancers (see **Appendix B**). After filtering to the predefined Oncomine variants, we identified an average of 1.7, 0.8 and 1.7 relevant somatic point mutations, indels and high level CNAs, respectively, per sample. Genes most frequently harboring relevant alterations across the MO cohort were *TP53* (33%), *BRAF* (31%) and *APC* (24%). An integrative heatmap of prioritized alterations across the MO cohort is shown in **Fig 2A**, and copy number profiles for all samples are shown in **Fig B3**.

A total of 4 prioritized gene fusions were identified across the cohort: *EML4:ALK* in two lung cancer samples positive for *ALK* rearrangement by molecular testing (MO-100 and MO-106), *ERC1:BRAF* in a melanoma sample (MO-17) and *TPR:NTRK1* in a colon cancer sample (MO-35) (**Fig 2A&B**). Importantly, multiple isoforms of the *ERC1:BRAF* fusion were identified in MO-17 due to combinatorial priming/amplification, including fusions of *ERC1* and *BRAF* exons 17 to 8 (designated E17B8), 12 to 9 (E12B9) and 12 to 10 (E12B10), respectively. qRT-PCR confirmed expression of *ERC1:BRAF* and *TPR:NTRK1* fusions in MO-17 and MO-35, respectively (**Fig 2C**).

The 104 informative MO samples underwent 129 total molecular diagnostic tests for *EGFR*, *BRAF*, *KRAS* and *ALK* alterations. The DNA component of the OCP demonstrated 100% sensitivity (44 of 44, 100%) and specificity (61 of 61, 100%) for detecting the clinically identified *EGFR*, *BRAF* and *KRAS* mutations (**Figure 2.2A** & **Appendix B**). Likewise, as described above, the RNA component of the OCP detected gene fusions involving *ALK* in 2 of the 3 (66%) samples with *ALK* rearrangements by FISH testing in the molecular diagnostics laboratory. In MO-66, a lung adenocarcinoma with an *ALK* rearrangement by molecular testing, OCP profiling identified only 9 *EML4:ALK* fusion reads, which was below our threshold for calling a gene fusion present; however, as described below, we observed 3'/5' *ALK* expression imbalance in this case (see **Fig 3B**). In total, considering MO-66 as failing to detect the *ALK* rearrangement, the 129 molecular tests performed across the MO cohort involving integrative DNA/RNA profiling by OCP showed 99.2% accuracy compared to molecular diagnostic testing. Additional findings from the MO cohort, including identification of relevant alterations not assessed by molecular testing, are described below.

OCP application in a lung cancer cohort

We also applied the OCP to a retrospectively identified cohort of 104 primary lung tumors given the assessment of somatic variants in lung cancer management[10]. The 101 informative samples from 96 individuals, which were chosen to represent the pathologic/histologic spectrum, consisted of 69 adenocarcinomas, 21 squamous cell carcinomas, 5 adenosquamous carcinomas, 2 bronchioalveolar carcinomas [1 adenocarcinoma *in situ* and 1 well differentiated lepidic predominant adenocarcinoma], 2 pulmonary small cell carcinomas (SCCs) and 2 carcinoid tumors (**Fig 3A** & **Appendix B**). After filtering to the predefined Oncomine variants, we identified an average of 1.2, 0.3, and 1.9 relevant somatic point mutations, indels and high level CNAs, respectively, per sample. *TP53* (38%), *KRAS* (28%), and *EGFR* (24%) were the genes most frequently harboring relevant alterations across the LU cohort. Alteration frequencies varied between histologic subtypes as expected. For example, high level CNAs in *NKX2-1*, which represent the most significant focal gain in lung adenocarcinoma[35], were observed in 15 of 69 (22%) adenocarcinomas in our LU cohort, but were not observed in the 21 squamous cell carcinomas (p=.0083, two-sided Fisher's exact test). Of note, both SCCs harbored nonsense *RB1* mutations, while both carcinoid tumors lacked prioritized alterations. An integrative heatmap of prioritized alterations across the LU cohort is shown in **Fig 3A**, and copy number profiles for all LU samples are shown in **Fig B4.**

Fourteen samples in the LU cohort underwent successful diagnostic molecular testing (as in the MO cohort) for *EGFR* and/or *ALK* alterations (27 total tests). OCP demonstrated 100% sensitivity and specificity for *EGFR* alterations in these samples; in LU-49, two somatic *EGFR* hotspot gain-of-function mutations that were not assayed for via molecular testing were identified by OCP (p.S768I, 37% variant allele frequency; p.G719C, 35% variant allele

frequency). All three samples with *EML4:ALK* fusions by OCP (LU-1, LU-30, and LU-150) harbored *ALK* rearrangements by FISH. In addition, in LU-61, an adenocarcinoma lacking other actionable alterations, we identified an *EZR:ROS1* fusion (exon 10 of *EZR* fused to exon 34 of *ROS1*). As shown in **Figure 2.3B**, all LU samples with detectable *ALK* and *ROS1* fusions by targeted RNAseq, as well as MO-100 and MO-106 (*EML4:ALK* fusion positive as described above) showed 3'/5' expression imbalance in the involved 3' partner. In total, across the LU and MO cohorts, 6 of the 8 (75%) samples with the greatest 3'/5' *ALK* expression imbalance by OCP harbored *ALK* rearrangements by FISH (**Fig 2.3B**), supporting the complementary information provided by this approach. Additional assessment of OCP performance in cases with known gene fusions is provided in the PR cohort below.

OCP validation and application in a prostate cancer cohort

Lastly, we applied the OCP to a cohort of 118 retrospectively identified prostate cancers for validation and application. The PR cohort was selected to enrich for samples poorly represented in standard frozen tissue cohorts, with the 116 informative samples (from 114 patients) including 35 diagnostic biopsy samples, 20 samples from individuals ≤55 yrs of age, and 50 previously treated samples (**Fig 2.4A** & **Appendix B**). After filtering to the predefined Oncomine variants, we identified an average of 1.0, 0.2, and 1.2 relevant somatic point mutations, indels and high level CNAs, respectively, per sample. Besides *T2:ERG* gene fusions (see below), the genes most frequently harboring relevant alterations in the PR cohort were *TP53* (27%), *PTEN* (18%), and *ATM* (11%). An integrative heatmap of prioritized alterations across the PR cohort is shown in **Fig 2.4A**, and copy number profiles for all PR samples are shown in **Fig B5**.

Approximately 40-60% of prostate cancers harbor recurrent gene fusions, typically

involving 5’ androgen regulated genes fused to 3’ ETS transcription factor family members, with the most common fusion being *TMPRSS2*(*T2*)*:ERG*[36, 37]. The RNA component of the OCP is designed to detect recurrent gene fusions in prostate cancer through inclusion of forward primers in known 5’ fusion partners (including *TMPRSS2*, *SLC45A3* and *C150RF21*) and reverse primers in known 3’ fusion partners (including *ERG*, *ETV1*, *ETV4* and *BRAF*). Across the PR samples, OCP detected ETS gene fusions in 58 of 100 (58%) samples, as shown in **Figure 2.4B**. Of note, amongst the 54 *T2*:*ERG* fusion positive samples, we identified a median of 3 unique fusion isoforms (range 1 to 9) due to combinatorial priming allowed by targeted RNAseq, consistent with the known expression of multiple *T2:ERG* splice variants in fusion positive tumors, including those reported to drive aggressive disease[38].

Thirty seven informative PR samples were previously assessed using an integrative DNA/RNA molecular profiling assay (MiPC) based on Haloplex target capture and Ion Torrent NGS coupled with qRT-PCR[30], providing an opportunity for additional OCP validation. Using automated variant calling and filtering, OCP profiling demonstrated 97% sensitivity (29 of 30) for detecting commonly targeted somatic variants (from 37 samples assessed by both approaches) with highly concordant observed variant allele frequencies (**Table S13**). High level CNAs in 34 genes targeted by both OCP and MiPC were also strongly correlated (Pearson’s r: 0.95; p <.001, ref [30]).

The thirty seven PR samples assessed by the OCP were also assessed by the RNA component of MiPC, which used a validated TaqMan qRT-PCR assay for *T2:ERG* (exon 1 of *TMPRSS2* fused to exon 4 of *ERG*, designated T1E4) and 3’ TaqMan expression assays for *ERG*, *ETV1*, *ETV4* and *ETV5* expression[30], with outlier expression of these genes indicative of gene fusions. We observed 100% concordance for *T2:ERG* isoform T1E4 expression by OCP and

MiPC (**Fig 2.4B**). Importantly, in the 3 cases identified as *ERG* expression outliers (without *T2:ERG* isoform T1E4 expression) by MiPC, we identified a *SLC45A3:ERG* gene fusion in PR-23 (3 detected fusion isoforms) and expression of non-T1E4 *T2:ERG* isoforms in PR-30 and PR-57 (PR-57 had *T2:ERG* T1E4 fusion reads detectable at <1/10,000th of non-T1E4 reads). Additionally, by OCP, we detected a *TMPRSS2:ETV1* gene fusion (supported by three fusion isoforms) in 1 of the 4 samples with *ETV1* outlier expression by MiPC (PR-7-3). No fusions were detected involving *ETV4* or *ETV5* in the PR cohort, although 2 of the 37 samples profiled previously by MiPC harbored *ETV4* or *ETV5* outlier expression, consistent with fusions involving 5' partners not targeted by the OCP. Taken together, with results from the MO and LU cohorts described above, these results support the ability of targeted RNAseq to identify isoform specific gene fusions through combinatorial priming and suggest that inclusion of 5'/3' expression amplicons (as for lung fusions) may improve detection of fusions involving novel 5' partners.

The inclusion of a large number of treated samples in the PR cohort enabled comparisons related to treatment status and unique histology/immunophenotype post treatment (i.e. prostatic neuroendocrine/SCC and samples with no/low canonical AR signaling by immunohistochemistry [AR$^-$]). For example, although *TP53* was the most frequently altered gene (besides *ERG*) in the PR cohort, *TP53* alteration frequency varied significantly across sample types, from 8.4% (6 of 71) of untreated or single modality treated samples (androgen deprivation [ADT] or radiation therapy [XRT]) to 100% of prostatic SCC (**Appendix B,** $p < .001$). Likewise, *ATM* alteration frequency varied across treatment subtypes, with 7 of 22 (32%) of samples treated with ADT + XRT and/or chemotherapy [ADT+] harboring *ATM* alterations compared to 0 of 8 (0%) of SCCs (**Appendix B**, $p=.14$). Robust prostate cancer molecular subtypes have been identified,

including those defined by ETS gene fusions, *SPOP* hotspot mutations and rare alterations (i.e. FGFR or RAF family fusions)[36]. Of interest, PR-122 harbored an *IDH1* R132 hotspot mutation (at 18% variant allele frequency) but lacked ETS gene fusions, *SPOP* mutations, or other prioritized alterations (**Fig 2.4A** & **B6A**). Assessment of the current PR cohort combined with 353 prostate cancer samples in the cBioPortal database identified *IDH1* R132 mutations in 6/453 (1%) prostate cancers, all of which lacked ETS gene fusions or *SPOP* mutations (p=.004, Fisher's exact test, **Fig B6B** & **Appendix B**), supporting *IDH1* mutations as defining a unique prostate cancer molecular subtype.

Lastly, OCP allowed us to assess paired samples that can inform on molecular correlates of disease progression, which is particularly challenging in prostate cancer given the long follow-up typically required to obtain sequential progressive specimens and the lack of routine biopsy confirmation of metastatic disease. In the PR cohort, PR-77 represents a primary, untreated Gleason score 9, pT3b N0 prostatectomy sample, while PR-88 is a paired urinary bladder tumor resected 4 years later after ADT, XRT and docetaxel chemotherapy with AR^- phenotype. Both samples showed focal prioritized *MCL1* and *MYC* amplifications (and non-prioritized high level *BRCA1* amplification), consistent with clonality, however a *TMPRSS2:ERG* fusion (exons T2E2) was identified by the OCP RNA-seq panel exclusively in PR-77, consistent with the AR^- phenotype in PR-88 (**Fig B7A**). In contrast, PR-88, the AR^- metastasis, uniquely harbored prioritized *AR* amplification and *CDKN2A* deletion, as well as a *CTNNB1* (*beta catenin*) GoF mutation (S37C, variant allele frequency 10%). Of note, no read support for *CTNNB1* S37C was present in PR-77, despite >5,000 covering reads. Likewise, PR-160, a post-therapy (ADT+chemotherapy) epidural metastasis resected after rapid progression in a man who presented with metastatic disease at the age of 49, harbored a focal, prioritized *CTNNB1*

amplification, which was not present in a pre-treatment, diagnostic prostate biopsy specimen that shared other clonal alterations with PR-160 (**Fig B7B**). These results demonstrate utility of OCP for identifying alterations associated with treatment resistance through profiling pre-/post-treatment limiting FFPE specimens.

Actionability assessment

An important component of the OCP is a knowledgebase of therapies and clinical trials associated with the predefined potential actionable variants targeted by the NGS assay. Potential therapeutic strategy prioritization for each OCP assessed sample is based on histologic cancer type and level of evidence associated with the potential actionability of each variant (FDA approved agent, within cancer type National Comprehensive Cancer Network (NCCN) guideline, outside cancer type NCCN guideline and biomarker directed/informed clinical trials; see **Appendix B**). In cases with multiple potential actionable variants, potential treatment strategies are prioritized, including consideration of detected variants that preclude treatment strategies based on other identified variants (i.e. *KRAS* mutations and potential EGFR inhibitor based treatment in colorectal adenocarcinoma),

To assess the potential utility of the OCP in identifying treatment options, we identified the highest priority alteration for each sample assessed herein, as shown in **Figure 2.5A**. These analyses only include positively associated variants (i.e. *KRAS* mutations in colorectal cancer excluding EGFR inhibitors are not prioritized). In the MO cohort, OCP confirmed the presence of *BRAF*, *EGFR* and *ALK* alterations in 29 samples (28%), each associated with FDA approved indications. In an additional 15 MO samples (14%), OCP identified an actionable variant that is not routinely tested for in that cancer type but which is associated with same- (n=2) or other-

cancer type (*n*=13) approved therapies referenced in NCCN clinical guidelines (e.g. *ERBB2*, *BRAF* and *EGFR* alterations). These findings are especially important because emerging evidence supports benefit, in some cases substantial, to an available targeted therapy. For example, responses to the BRAF inhibitor dabrafenib in lung cancer patients with *BRAF* mutations in a phase II trial led to Breakthrough Therapy designation by the FDA, and combination trials with the MEK inhibitor trametinib—which proved superior to single agent BRAF therapy in melanoma [39, 40]—are enrolling. An additional 44 samples (42%) harbored alterations in a gene that is a positive eligibility criteria for a clinical trial involving a targeted therapy (e.g. *PIK3CA, NRAS*, etc*).*

Likewise, in the LU cohort, OCP identified alterations associated with FDA approved therapies, NCCN guidelines and clinical trial eligibility in 21 (21%), 15 (15%; n=11 same-cancer; n=4 other-cancer) and 52 (51%) samples, respectively. Lastly, in the PR cohort, OCP identified alterations associated with FDA approved therapies, NCCN guidelines and clinical trial eligibility in 0 (0%), 7 (6%; all other cancer) and 42 (36%) samples, respectively, demonstrating that "actionable" alterations occur with variable frequency across cancers from different organs. As an example of a highly actionable alteration that is not routinely tested for in the specific cancer type (lung cancer) nor assessed by targeted NGS approaches that do not assess CNAs, OCP prioritized high level gains in *ERBB2* in MO-86 and LU-31 (lung adenocarcinoma and lung SCC, respectively), with over-expression in both cases confirmed by IHC (**Fig 2.5B**).

DISCUSSION

Here we report the development, validation and assessment of a highly scalable, FFPE-compatible, targeted NGS based system to prioritize potential treatment strategies from

predefined relevant somatic variants in solid tumors. To identify candidate driving somatic alterations for inclusion in the OCP, we queried genomic data from over 700,000 tumor samples to define pan-solid tumor, recurrent driving somatic alterations through defining GoF mutations in oncogenes, LoF (and GoF) mutations in tumor suppressors, CNAs through minimal common region analysis, and recurrent gene fusions. These alterations were combined with a comprehensive knowledgebase of currently available oncology therapeutics and clinical trials to define variants with immediate or near-term relevance. We then developed a targeted multiplexed PCR based NGS panel compatible with limited amounts of routine FFPE tissue samples (20ng DNA/15ng RNA) to detect these variants. These nucleic acid requirements are 2-50 fold less than those for comprehensive capture-based precision oncology approaches [18, 21]. To balance OCP panel size and clinical relevance, we excluded genes without near term clinical actionability and only currently identified hotspots are targeted. Hence, additional genes/amplicons may be included in future OCP versions or supplemental panels to target novel relevant alterations, including treatment resistance hotspots poorly represented in most publically available profiling studies.

We validated OCP performance using over 300 FFPE tumor specimens, including a prospective cohort of 104 samples undergoing concurrent molecular diagnostics testing for *BRAF*, *KRAS* or *EGFR* point mutations and indels, achieving a sensitivity and specificity of 100%. We and others have previously validated the utility of multiplexed PCR based Ion Torrent sequencing for CNA assessment[19, 25, 31-33] and herein confirm high level *ERBB2* CNAs identified by OCP using IHC. OCP identified mutations and high level CNAs were also highly concordant with results from Haloplex capture based NGS in a subset of the PR cohort profiled by both technologies. Taken together, these results demonstrate the ability of OCP to identify

these relevant classes of alterations. Frameshifting indels in long homopolymer runs are challenging to detect with current Ion Torrent approaches (and are excluded using our filtering criteria) and multiplexed PCR approaches cannot detect large structural rearrangements, however these alterations predominantly result in LoF alterations in tumor suppressors [41], which represent a minority of current therapeutic targets. We anticipate that our cohort and additional OCP profiled samples will enable the development of panel- and laboratory-specific error models to improve performance in homopolymer regions.

The RNA component of the OCP is designed to identify known recurrent gene fusions (through primers spanning known exon junctions) as well as fusions of *RET, ROS1* and *ALK* with novel 5' partners (or novel fusion isoforms) through 3'/5' expression imbalance. We confirmed 100% concordance for *T2:ERG* gene fusion isoform specific detection between OCP and a validated qPCR assay in a subset of our PR cohort profiled by both methods, with multiple splice variants detected in the majority of fusion positive cases. Likewise, in 7 lung cancers known to harbor *ALK* rearrangements across our cohorts, OCP profiling identified *EML4:ALK* fusions in 5 (71%), with these 5 samples also showing 3'/5' expression imbalance by OCP. MO-66 (the known *ALK* rearranged sample with fusion read support below our threshold criteria) and LU-38 (known *ALK* rearrangement without fusion read support) also showed 3'/5' expression imbalance. Of note, in our MO cohort, we identified two additional relevant fusions (*ERC1:BRAF* in a melanoma sample negative for *BRAF* mutation [MO-17] and *TPR:NTRK1* in a colon cancer [MO-35]), validating both fusions by qPCR. Of note *ERC1:BRAF* was not directly targeted in the OCP RNA panel design, as *ERC1* had previously only been reported as a fusion partner with *RET[42]*, highlighting the utility of the combinatorial nature of targeted multiplexed PCR based RNAseq.

Taken together, our results validate the multiplexed PCR based RNA sequencing approach for detecting targeted gene fusions. Characterization of additional cohorts will be required to determine performance and optimal 3'/5' expression imbalance cutoffs for *ALK*, *RET* and *ROS1* fusions in lung cancer (and other cancer types) involving unknown partners. Likewise, we anticipate that inclusion of additional 3'/5' expression imbalance amplicons will improve fusion detection involving other genes. Lastly, splice variant detection of non-gene fusion events, such as *AR* splice variants in prostate cancer[43, 44] or alternatively spliced tyrosine kinases (e.g. *MET*) in other cancers[45, 46], may also be assessed in OCP through inclusion of additional amplicons. Importantly, although comprehensive capture based NGS approaches assessing only DNA can identify gene fusions through sequencing introns of involved genes [18], such approaches cannot detect or quantify potentially relevant splice variants.

As a demonstration of the utility of OCP for translational research, we applied this approach to a cohort of 116 prostate cancers, including 50 previously treated samples. We recapitulated known molecular subtypes and alterations with specific histology, including the high prevalence of *TP53* alterations in prostatic SCC[47-52]. We also identified a high burden of *ATM* alterations in heavily treated patients, which can be investigated in future efforts characterizing this understudied population. Of note, through integration with previous profiling studies, we identify *IDH1* R132 mutant prostate cancer as a novel molecular subtype that lacks other subtype defining lesions. This finding is especially important as IDH1 inhibitors are now in early phase clinical trials. Lastly, two pairs of pre- and post-treatment samples each demonstrated *AR* amplifications (a known adaptive response to ADT [53]) and *CTNNB1* GoF mutation/amplification exclusively in the post-treatment sample. Although activation of the WNT/CTNNB1 pathway has been identified in ADT treated prostate cancer[29, 54, 55], our

report is the first to demonstrate that ADT and/or subsequent chemotherapy specifically induces (or selects) for *CTNNB1* amplification/activating mutation, supporting a functional role in treatment resistance..

The OCP is compatible with routine Ion Torrent workflows, and the DNA/RNA components of the OCP can be combined for template preparation and concurrent PGM sequencing on a single Ion Torrent 318 chip in a standard ~4hr PGM run, with the potential for higher throughput using the Ion Torrent Proton. Although complete analytic validation will need to be performed in individual laboratories, we demonstrate highly concordant results with typical specimens sent for molecular diagnostic testing as well as molecular standards (performance with down-sampled reads is shown in **Fig B8**). Hence, this approach provides a rapid, highly scalable approach requiring small amounts of routine tissue specimens with performance comparable to previous multiplexed PCR based Ion Torrent panels assessing DNA alterations[16, 19, 22], capture based approaches[18, 56] and anchored multiplexed PCR based NGS[24]. A critical component of the OCP is a highly automated analysis pipeline that links to a knowledgebase of potential treatment options, facilitated by predefining the actionable cancer genome prior to panel development. As shown through our actionability assessment, a significant number of samples currently harbor relevant alterations that are identifiable using our approach. As clinical sequencing efforts and expertise become more prevalent, a key advantage of the OCP is the potential for integration into multiple independent institutions (rather than a single centralized testing center), enabling valuable direct involvement from molecular biologists, pathologists and oncologists. Taken together, the highly scalable assay and framework described herein may have utility in future oncology precision medicine approaches, such as the NCI Match Trial, where multiple sites will sequence 3,000 cancer samples using the OCP.

Figure 2.1. Pan-solid tumor cancer somatic alteration analysis to identify relevant variants

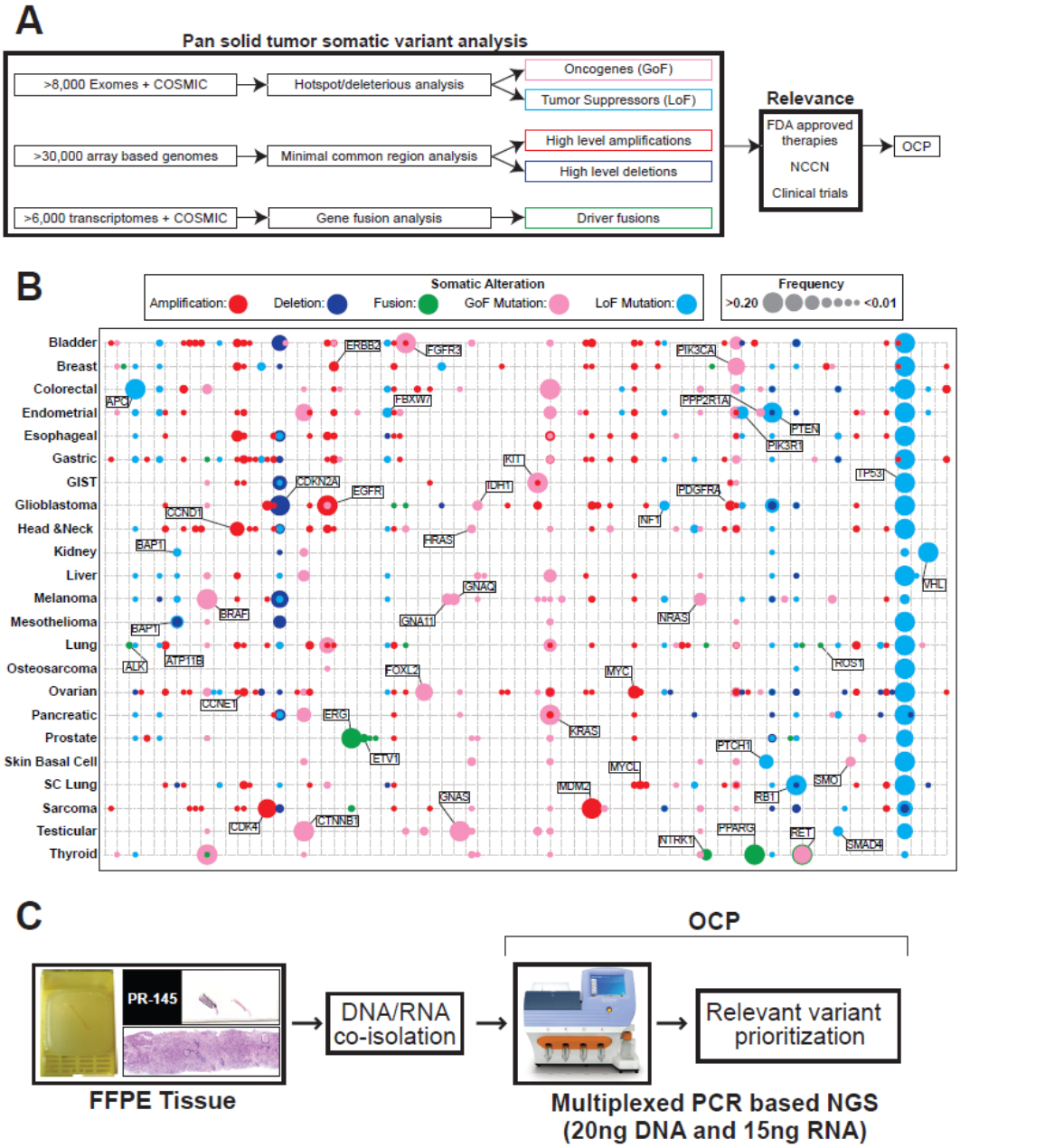


Figure 2.1. **A**. Using the Oncomine database supplemented with data from COSMIC, over 700,000 tumor samples (including >8,000 cancer exomes) were used to assess genes for over-representation of hotspot (gain of function [GoF] and deleterious (loss of function [LoF]) mutations to identify oncogenes and tumor suppressors, respectively. Array based copy number profiles from >30,000 tumors were assessed by minimal common region analysis to identify targets of focal, high level amplifications or deletions. Transcriptomes from >7,000 cancers were similarly assessed for driver gene fusions. Prioritized genes were further filtered to include only near term relevant alterations for inclusion into the Oncomine Comprehensive Panel (OCP). **B**. Frequency of somatic alterations (type according to color in the legend) in OCP included genes across publically available The Cancer Genome Atlas (TCGA) data. For each gene per cancer type, alteration frequency (<0.01 to >0.20) is indicated by the size of the circle according to the legend. Selected genes of interest are highlighted. **C**. The OCP was designed for compatibility with routine formalin fixed paraffin embedded (FFPE) tissues, with co-isolation of DNA/RNA from FFPE tissues used in our validation. The OCP consists of multiplexed PCR (Ampliseq) panels compatible with 20ng DNA and 15ng RNA, which can be combined after library generation for NGS on Ion Torrent benchtop sequencers. By predefining relevant somatic variants, identified variants can be linked to potential treatment strategies.

Figure 2.2. Validation of the Oncomine Comprehensive Panel (OCP) using an oncology cohort undergoing molecular diagnostics testing.

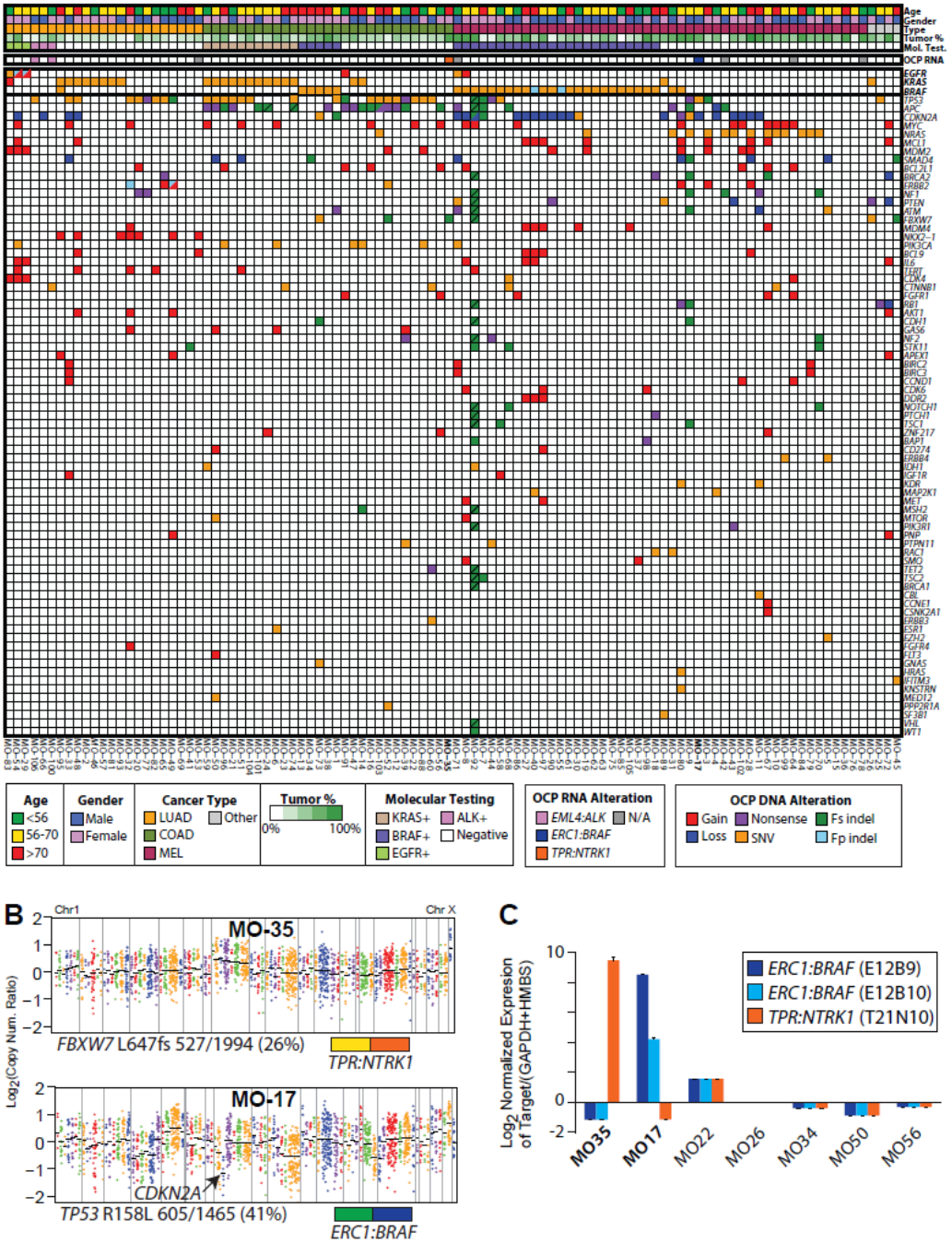


Figure 2.2 A. We applied the OCP to a prospectively identified cohort of formalin fixed paraffin embedded (FFPE) cancer samples undergoing molecular diagnostics testing for somatic mutations in *BRAF*, *KRAS* or *EGFR*, or *ALK* rearrangements (MO cohort). All OCP defined relevant alterations from the RNA (in header) and DNA components of the OCP for the 104 informative samples are shown in the heatmap. Specific alteration types are indicated according to the legend (Nonsyn. SNV = nonsynonymous SNV; Fs. and Fp. indel = frame-shifting and frame-preserving indels, respectively). Slashed boxes indicate two alterations. Samples not sequenced in OCP RNA analysis are indicated as in the legend. Samples excluded from copy number analysis due to noisy profiles are named in italics. Clinicopathological information is given in the header according to the legend (LUAD= lung adenocarcinoma, COAD = colon adenocarcinoma, MEL= melanoma). 100% concordance with molecular testing was observed for mutations (see **Table S10**). Detailed OCP RNAseq results, including 3'/5' expression imbalance, for the *ALK* rearrangement positive lung cancers are shown in **Fig 3B**. **B**. Integrative OCP results from two cases, MO-17 and MO-25 (names

bolded in **A**) harboring relevant gene fusions. Copy number plots show log_2 copy number ratios (compared to a composite normal sample) per amplicon, with each individual amplicon represented by a single dot, and individual genes indicated by different colors. Gene-level copy number estimates are shown as black bars. By OCP, MO-17 (top), a *BRAF* wildtype melanoma by clinical testing, harbored *CDKN2A* high level copy number loss, *TP53* R158L mutation, and a novel *ERC1:BRAF* gene fusion. OCP profiling of MO-35, a *KRAS/BRAF* wildtype colon adenocarcinoma by clinical testing, identified an *FBXW7* L647fs mutation and a *TPR:NTRK1* gene fusion. For mutations, variant allele containing reads/total reads and the variant allele frequency are shown. **C**. Validation of OCP identified gene fusions using quantitative (q) RT-PCR for *ERC1:BRAF* (*ERC1* exon 12 fused to *BRAF* exon 9 [E12B9, blue] or 10 [E12B10, cyan]) and *TPR:NTRK1* (*TPR* exon 21 fused to *NTRK1* exon 10 [T21N10, orange]). qRT-PCR was performed on MO-17, MO-35 and five control MO samples without OCP detected gene fusions. Mean log_2 expression (normalized to the arithmetic mean of *GAPDH+HMBS* calibrated to the mean of the MO control samples) + S.D. of triplicate qPCR reactions are plotted. No detectable expression of *ERC1:BRAF* or *TPR:NTRK1* was present in any sample other than that identified by OCP.

Figure 2.3. OCP identified relevant somatic alterations, including gene fusions, in a lung cancer cohort.

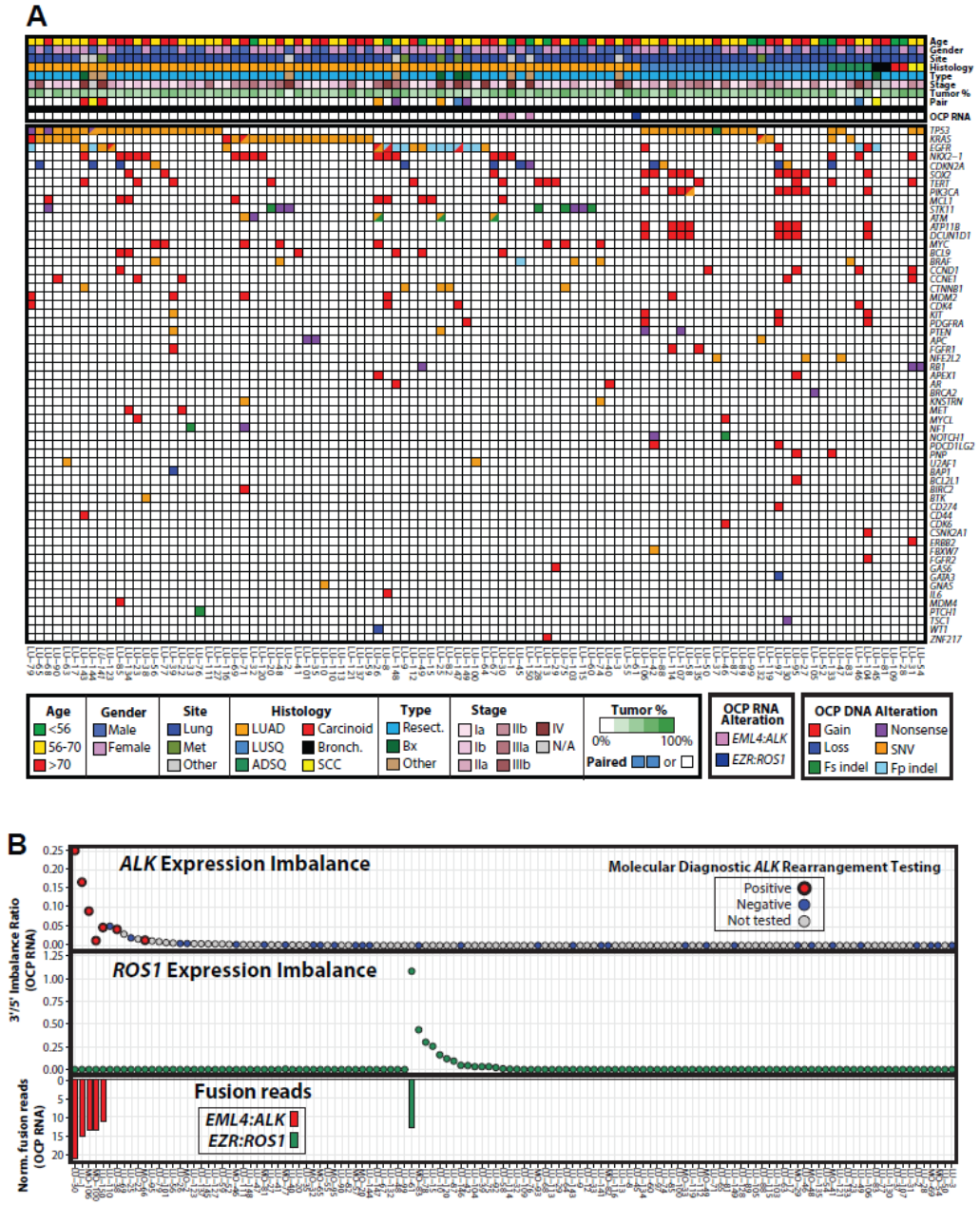


Figure 2.3. A. We applied the OCP to a retrospective cohort of FFPE lung tumors selected to represent diverse pathology (LU cohort). All OCP defined relevant alterations from the RNA (in header) and DNA components of the OCP for the 101 informative samples are shown in the heatmap. Clinicopathological information is given in the header according to the legend (Met = metastasis; LUSQ = squamous cell carcinoma, ADSQ = adenosquamous carcinoma, BAC = bronchioloalveolar carcinoma (adenocarcinoma in situ or well differentiated lepidic predominant adenocarcinoma), SCC = small cell carcinoma; Resect. = resection, Bx = biopsy). All 101 informative lung samples were included in OCP RNA analysis. Samples excluded from copy number analysis due to noisy profiles are named in italics. **B**. In addition to primers for pan-cancer prioritized 5' and 3' gene fusion partners, OCP includes 5' and 3' amplicons for *ALK*, *ROS1* and *RET* to identify 3'/5' expression imbalance indicative of gene fusions. For all lung tumors (including those from MO cohort), normalized OCP RNAseq expression of gene fusions involving *ALK* (red) and *ROS1* (green) are plotted. No fusions involving *RET* were detected. Corresponding normalized 3'/5' expression imbalance for *ALK* (top panel) and *ROS1* (middle panel) for each sample are plotted. *ALK* rearrangement positive (bolded red), negative, (blue) or untested (gray) samples by molecular testing are indicated.

Figure 2.4. Application of OCP to a prostate cancer cohort identifies variable alterations across histologic and treatment subtypes and confirms isoform specific gene fusion detection.

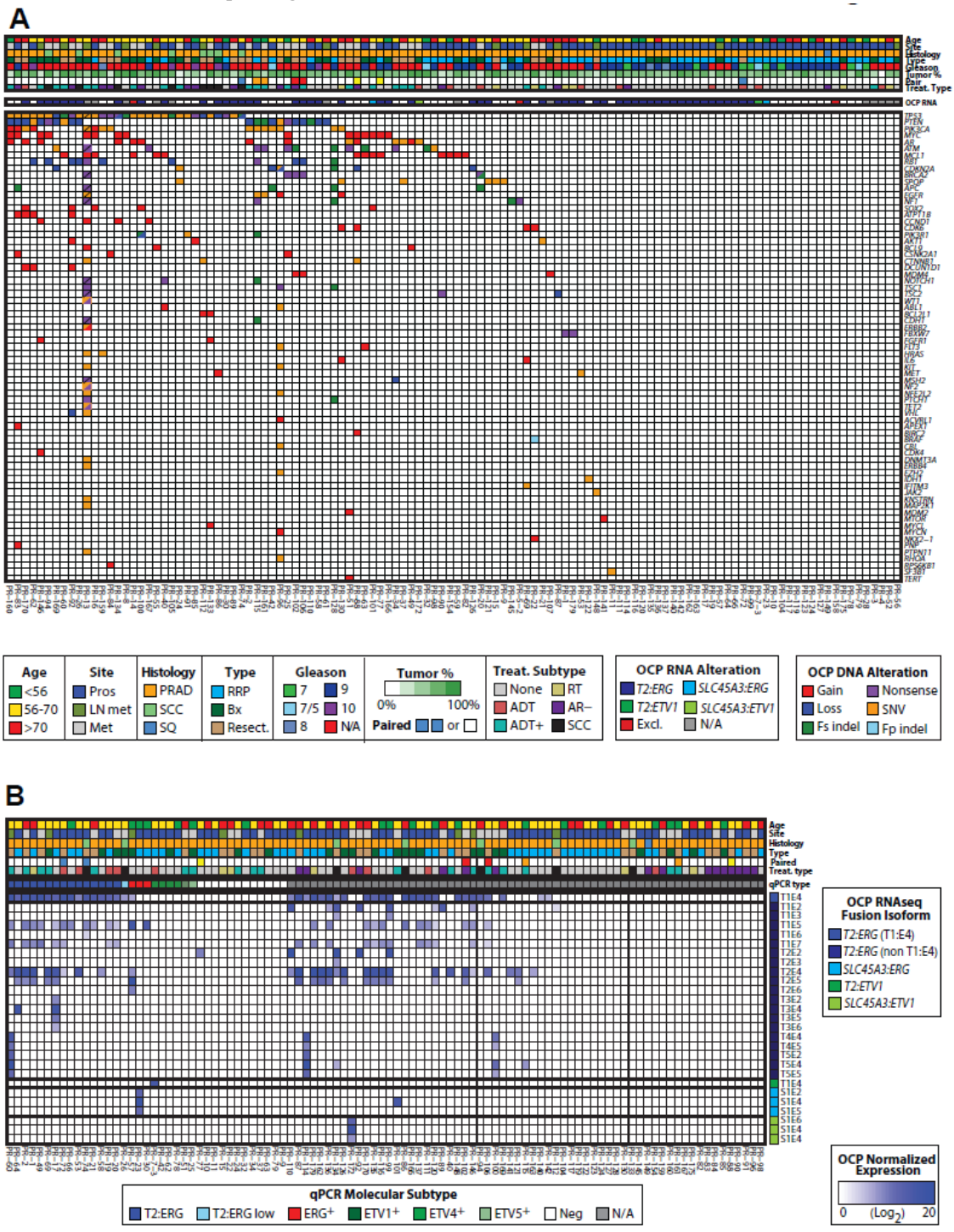


Figure 2.4 A. We applied the OCP to a retrospective cohort of aggressive FFPE prostate cancers. All OCP defined relevant alterations from the RNA (in header) and DNA components of the OCP for the 116 informative samples are shown in the heatmap. Clinicopathological information is given in the header according to the legend (Met = metastasis; Pros.= prostate, LN met= lymph node metastasis; PRAD= prostatic adenocarcinoma, SCC = small cell carcinoma, SQ= squamous differentiation; RRP = radical prostatectomy). For treatment subtype, ADT = prior androgen deprivation therapy, XRT = radiation therapy, ADT+ = ADT plus XRT and/or chemotherapy, AR- = no (or reduced) AR signaling as indicated by no/focal PSA staining.

Samples excluded from or not sequenced in OCP RNA analysis are indicated as in the legend. **B**. The RNA component of the OCP contains forward primers in known 5' fusion partners and reverse primers in known 3' fusion partners for recurrent gene fusions in prostate cancer. Normalized log_2 read counts for indicated gene fusion isoforms are indicated in each cell according to the color scale, with individual fusions indicated by the color blocks (right) and fusion isoforms named by the exon junctions of the involved genes (e.g. *T2:ERG* T1E4 indicates a fusion junction of *TMPRSS2* exon 1 and *ERG* exon 4). qRT-PCR was previously performed on a subset of these cases, as indicated in qPCR type. *T2:ERG* T1E4 status (including low expression), and *ERG* outlier expression without T1E4 isoform detection (ERG^+), *ETV1* ($ETV1^+$), *ETV4* ($ETV4^+$) or *ETV5* ($ETV5^+$) are indicated in the header. Samples without any of these alterations (Neg) or not tested (N/A) by qPCR are indicated.

Figure 2.5. Automated treatment prioritization by OCP identifies relevant alterations beyond routine molecular testing.

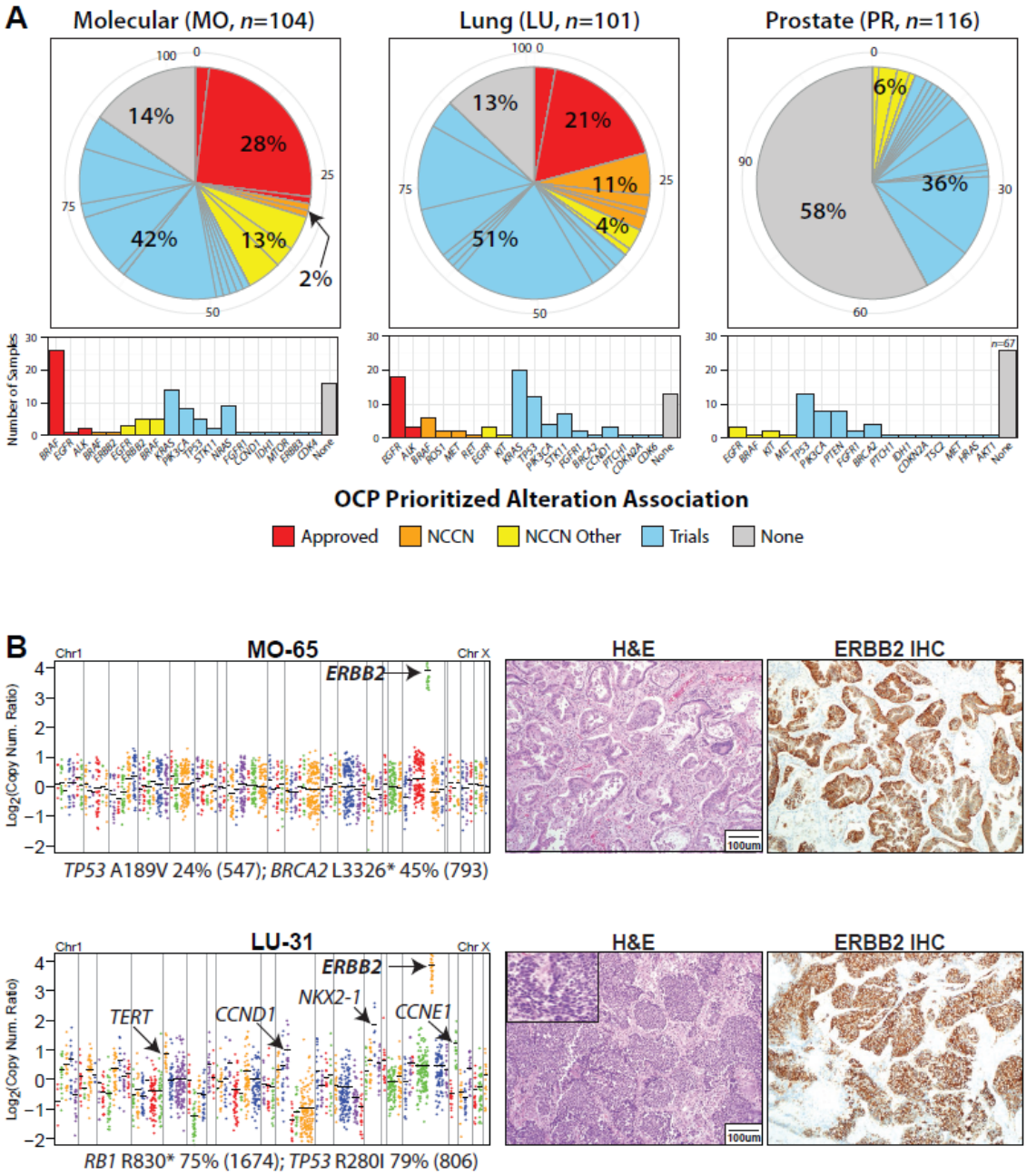


Figure 2.5A. For each OCP assessed cohort, the breakdown of the highest prioritized alteration per sample is shown, according to whether the alteration is associated with: 1) FDA approved therapies (red), 2) therapies within NCCN indications (orange), 3) therapies outside that specific cancer type's NCCN indication (yellow), 4) clinical trial entry requirements (blue). This assessment incorporates variants precluding treatment strategies based on other identified variants, but does not prioritize variants that only exclude approved agents. Individual prioritized alterations are indicated as slices of each pie, and are shown in the histogram. **B.** Integrative OCP profiling prioritized high-level *ERBB2* copy gains in two lung carcinomas. Integrative OCP results are shown as in **Figure 2B** (gene fusions were not identified in either sample). OCP profiling prioritized high level *ERBB2* copy number gains in MO-65 (top), an *EGFR/ALK* wildtype lung adenocarcinoma by diagnostic molecular testing, and LU-31 (bottom), a lung small cell carcinoma with no previous molecular diagnostic testing. Morphology by hematoxylin and eosin (H&E) staining is shown (inset of LU-31 shows typical small cell morphology). Diffuse 3+ ERBB2 protein expression was confirmed by immunohistochemistry (IHC).

Chapter II References

1. Garraway, L.A., J. Verweij, and K.V. Ballman, *Precision oncology: an overview.* J Clin Oncol, 2013. **31**(15): p. 1803-5.
2. Mendelsohn, J., *Personalizing oncology: perspectives and prospects.* J Clin Oncol, 2013. **31**(15): p. 1904-11.
3. Arteaga, C.L. and J. Baselga, *Impact of genomics on personalized cancer medicine.* Clin Cancer Res, 2012. **18**(3): p. 612-8.
4. McDermott, U., J.R. Downing, and M.R. Stratton, *Genomics and the continuum of cancer care.* N Engl J Med, 2011. **364**(4): p. 340-50.
5. Pant, S., R. Weiner, and M.J. Marton, *Navigating the rapids: the development of regulated next-generation sequencing-based clinical trial assays and companion diagnostics.* Front Oncol, 2014. **4**: p. 78.
6. Olsen, D. and J.T. Jorgensen, *Companion diagnostics for targeted cancer drugs - clinical and regulatory aspects.* Front Oncol, 2014. **4**: p. 105.
7. Parkinson, D.R., B.E. Johnson, and G.W. Sledge, *Making personalized cancer medicine a reality: challenges and opportunities in the development of biomarkers and companion diagnostics.* Clin Cancer Res, 2012. **18**(3): p. 619-24.
8. Mass, R.D., et al., *Evaluation of clinical outcomes according to HER2 detection by fluorescence in situ hybridization in women with metastatic breast cancer treated with trastuzumab.* Clin Breast Cancer, 2005. **6**(3): p. 240-6.
9. Druker, B.J., *Translation of the Philadelphia chromosome into therapy for CML.* Blood, 2008. **112**(13): p. 4808-17.
10. Kris, M.G., et al., *Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs.* JAMA, 2014. **311**(19): p. 1998-2006.
11. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.
12. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
13. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
14. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-1140.
15. Tomlins, S.A., et al., *Analysis of 2,700 Cancer Exomes to Identify Novel Cancer Drivers and Therapeutic Opportunities.* European Journal of Cancer, 2012. **48**: p. 134-134.
16. Beadling, C., et al., *Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping.* J Mol Diagn, 2013. **15**(2): p. 171-6.
17. Borad, M.J., et al., *Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma.* PLoS Genet, 2014. **10**(2): p. e1004135.
18. Frampton, G.M., et al., *Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing.* Nat Biotechnol, 2013.
19. Grasso, C., et al., *Assessing Copy Number Alterations in Targeted, Amplicon-Based Next-Generation Sequencing Data.* J Mol Diagn, 2014.

20. Hadd, A.G., et al., *Targeted, high-depth, next-generation sequencing of cancer genes in formalin-fixed, paraffin-embedded and fine-needle aspiration tumor specimens.* J Mol Diagn, 2013. **15**(2): p. 234-47.
21. Roychowdhury, S., et al., *Personalized oncology through integrative high-throughput sequencing: a pilot study.* Sci Transl Med, 2011. **3**(111): p. 111ra121.
22. Singh, R.R., et al., *Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes.* J Mol Diagn, 2013. **15**(5): p. 607-22.
23. Van Allen, E.M., et al., *Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine.* Nat Med, 2014. **20**(6): p. 682-8.
24. Zheng, Z., et al., *Anchored multiplex PCR for targeted next-generation sequencing.* Nat Med, 2014.
25. Hoogstraat, M., et al., *Simultaneous Detection of Clinically Relevant Mutations and Amplifications for Routine Cancer Pathology.* J Mol Diagn, 2015. **17**(1): p. 10-18.
26. Samorodnitsky, E., et al., *Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing.* J Mol Diagn, 2015. **17**(1): p. 64-75.
27. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2014.
28. Rhodes, D.R., et al., *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-80.
29. Grasso, C.S., et al., *The mutational landscape of lethal castration-resistant prostate cancer.* Nature, 2012. **487**(7406): p. 239-43.
30. Grasso, C.S., et al., *Integrative molecular profiling of routine clinical prostate cancer specimens.* Ann Oncol, 2015.
31. Warrick, J.I., et al., *Tumor evolution and progression in multifocal and paired non-invasive/invasive urothelial carcinoma* Virchows Arch, 2014.
32. Cani, A.K., et al., *Next-Gen Sequencing Exposes Frequent MED12 Mutations and Actionable Therapeutic Targets in Phyllodes Tumors.* Mol Cancer Res, In press.
33. McDaniel, A.S., et al., *HRAS mutations are frequent in inverted urothelial neoplasms.* Hum Pathol, 2014. **45**(9): p. 1957-1965.
34. Palanisamy, N., et al., *Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma.* Nat Med, 2010. **16**(7): p. 793-8.
35. Weir, B.A., et al., *Characterizing the cancer genome in lung adenocarcinoma.* Nature, 2007. **450**(7171): p. 893-8.
36. Tomlins, S.A., *Molecular clues assist in the cancer clinic.* Sci Transl Med, 2013. **5**(193): p. 193fs26.
37. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.
38. Wang, J., et al., *Expression of variant TMPRSS2/ERG fusion messenger RNAs is associated with aggressive prostate cancer.* Cancer Res, 2006. **66**(17): p. 8347-51.
39. Robert, C., et al., *Improved Overall Survival in Melanoma with Combined Dabrafenib and Trametinib.* N Engl J Med, 2014.
40. Long, G.V., et al., *Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma.* N Engl J Med, 2014. **371**(20): p. 1877-88.
41. Baca, S.C., et al., *Punctuated evolution of prostate cancer genomes.* Cell, 2013. **153**(3): p. 666-77.

42. Nakata, T., et al., *Fusion of a novel gene, ELKS, to RET due to translocation t(10;12)(q11;p13) in a papillary thyroid carcinoma.* Genes Chromosomes Cancer, 1999. **25**(2): p. 97-103.
43. Sita-Lumsden, A., et al., *Circulating microRNAs as potential new biomarkers for prostate cancer.* Br J Cancer, 2013. **108**(10): p. 1925-30.
44. Antonarakis, E.S., et al., *AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer.* N Engl J Med, 2014. **371**(11): p. 1028-38.
45. Kong-Beltran, M., et al., *Somatic mutations lead to an oncogenic deletion of met in lung cancer.* Cancer Res, 2006. **66**(1): p. 283-9.
46. Druillennec, S., C. Dorard, and A. Eychene, *Alternative splicing in oncogenic kinases: from physiological functions to cancer.* J Nucleic Acids, 2012. **2012**: p. 639062.
47. Barbieri, C.E. and S.A. Tomlins, *The prostate cancer genome: perspectives and potential.* Urol Oncol, 2014. **32**(1): p. 53 e15-22.
48. Beltran, H. and M.A. Rubin, *New strategies in prostate cancer: translating genomics into the clinic.* Clin Cancer Res, 2013. **19**(3): p. 517-23.
49. Brenner, J.C., A.M. Chinnaiyan, and S.A. Tomlins, *ETS Fusion Genes in Prostate Cancer*, in *Prostate Cancer: Biochemistry, Molecular Biology and Genetics*, D.J. Tindall, Editor. 2013, Springer New York: New York. p. 139-183.
50. Rubin, M.A., C.A. Maher, and A.M. Chinnaiyan, *Common gene rearrangements in prostate cancer.* J Clin Oncol, 2011. **29**(27): p. 3659-68.
51. Beltran, H., et al., *Aggressive variants of castration-resistant prostate cancer.* Clin Cancer Res, 2014. **20**(11): p. 2846-50.
52. Chen, H., et al., *Pathogenesis of prostatic small cell carcinoma involves the inactivation of the P53 pathway.* Endocr Relat Cancer, 2012. **19**(3): p. 321-31.
53. Bluemn, E.G. and P.S. Nelson, *The androgen/androgen receptor axis in prostate cancer.* Curr Opin Oncol, 2012. **24**(3): p. 251-7.
54. Kumar, A., et al., *Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers.* Proc Natl Acad Sci U S A, 2011. **108**(41): p. 17087-92.
55. Rajan, P., et al., *Next-generation sequencing of advanced prostate cancer treated with androgen-deprivation therapy.* Eur Urol, 2014. **66**(1): p. 32-9.
56. Beltran, H., et al., *Targeted Next-generation Sequencing of Advanced Prostate Cancer Identifies Potential Therapeutic Targets and Disease Heterogeneity.* Eur Urol, 2012.

CHAPTER III: Rapid, Ultra Low Coverage Copy Number Profiling of Cell-Free DNA as a Precision Oncology Screening Strategy

Previously published in *Oncotarget*.

INTRODUCTION

Clinical and commercial next-generation sequencing (NGS) based precision oncology strategies have expanded rapidly[1, 2]. Both targeted [3-8] and more comprehensive [9, 10] NGS assessment of frozen and archived formalin-fixed paraffin-embedded (FFPE) tissue samples have proven effective in identifying certain categories of clinically informative somatic DNA-based alterations, but tissue and re-biopsy requirements serve as considerable hurdles for widespread clinical implementation for identifying and tracking clinically relevant genomic alterations.

Myriad noninvasive ('liquid biopsy') approaches for identifying and tracking clinically relevant genomic alterations from cell-free DNA (cfDNA) have emerged as viable and potentially more broadly applicable alternatives to tissue-based assays using technologies including quantitative PCR (qPCR), digital droplet PCR (ddPCR), targeted DNA sequencing, and whole exome (WES) or whole genome sequencing (WGS)[2, 10-31]. Identifying a tractable, scalable precision oncology workflow with utility across patients with various advanced cancers, however, is still a substantial challenge given the variability of tumor-derived circulating cfDNA content, relevant genomic alterations, and frequent need for ultra-deep (e.g. >10,000x), high-

sensitivity sequencing in order to ensure detection (or absence) of clinically relevant alterations in pan-cancer cohorts[25, 32].

Genome-wide copy number profiles derived from low-pass cfDNA whole genome sequencing (WGS) are routinely used to detect large-scale aneuploidy events in clinical applications such as screening for fetal anomalies during pregnancy [33-36]. Multiple experiments have leveraged similar principles using low-pass cfDNA WGS to infer somatic whole-genome copy-number profiles in patients with advanced cancer, occasionally deploying higher depth disease-specific strategies for approximating cfDNA tumor content [22, 37-41]. However, these approaches often rely on disease specificity trade-offs that limit widespread prospective implementation[39]. Applicability across cancers, routine identification of actionable CNAs, correlation with comprehensive tissue based NGS profiling, and use as a precision oncology screen strategy have not yet been comprehensively addressed[40, 41]. Initiatives comparing comprehensive tissue-based molecular profiles to those obtained from cfDNA have also thus far been limited in size, particularly in metastatic castration resistant prostate cancer (mCRPC) [30, 31, 40].

Here, as part of an effort to facilitate precision medicine for all patients with advanced cancer, we propose a comprehensive approach deploying rapid, inexpensive, ultra-low pass cfDNA WGS as a broadly applicable potential screening strategy through: 1) directly identifying actionable CNAs, 2) informing needed sequencing depth for additional comprehensive/targeted cfDNA assessment (through cfDNA tumor content approximation) and 3) reserving ultra-deep cfDNA sequencing or tissue-based profiling for patients with low cfDNA tumor content. We show that with effective whole-genome coverage as low as 0.01x (<100,000 single end reads) per sample on a benchtop Ion Torrent sequencer from as little as 10 pg of double-stranded DNA,

we can recapitulate known whole-genome copy number profiles in cell lines and advanced prostate, colon, lung, and breast cancer patient samples, while retaining the ability to identify both focal and broad CNAs with megabase-level resolution. To confirm the utility of this screening approach to guide additional precision oncology assessment, we also paired this ultra-low-pass WGS with targeted multiplexed PCR based NGS of the same cfDNA, validating CNAs and identifying clinically relevant somatic mutation profiles at depth tuned by WGS-informed cfDNA tumor content approximation. Further, we directly compare cfDNA copy-number and mutational profiles with molecular profiles from synchronous or asynchronous tissue samples, highlighting high overall concordance and unique considerations for comprehensive precision oncology workflows, while exploring associations between putative cfDNA biomarkers and therapeutic outcomes in patients with mCRPC.

RESULTS

*Rationale for a **p**an-cancer, **r**apid, **i**nexpensive, ultra-low pass **N**GS **c**fDNA (PRINCe) approach to guide precision oncology*

The major impetus for ultra-deep, high sensitivity cfDNA profiling in precision oncology is the need for robust sensitivity and specificity for somatic alterations detection at extremely low cfDNA tumor content [42]. While many cfDNA-based detection approaches thus rely heavily on targeted, ultra-sensitive methodologies, many patients with elevated tumor burden or metastatic treatment refractory cancer—where precision oncology NGS is most commonly employed—have relatively high cfDNA tumor contents of 5-50% [22, 25, 42] (**Figure 3.1A**). If tumor-derived cfDNA characteristics could be rapidly leveraged to approximate tumor content and

potentially identify clinically relevant alterations across cancer types, unique and potentially more optimized precision medicine strategies may be achievable. Given that somatic copy-number alterations (CNAs) are pervasive in cancer [43] and somatic copy-number burden may be an important marker for aggressive or treatment-resistant disease [44], we first assessed the prevalence of extended copy-number burden in a pan-cancer TCGA cohort using 11,576 copy number profiles from 32 tumor types (**Figure 3.1B**). Overall, 56% of tumors had elevated copy-number burden (defined by having >15% fraction of the genome altered [FGA]), with FGA increasing with pathologic tumor stage, tumor grade and clinical stage (**Figure C1**). Importantly, per-sample FGA was also increased in a cohort of advanced/metastatic tumors (*n*=129) profiled as part of the MI-ONCOSEQ project[45] compared to the TCGA cohort, with 81% of Mi-ONCOSEQ profiled tumors having >15% FGA (**Figure 3.1B**). As CNAs can be robustly detected at substantially lower sequencing coverage (and cost) than typically required for somatic mutation calling in genome-wide or targeted pan-cancer workflows, we sought to exploit genome-wide CNAs as a biomarker through a pan-cancer, rapid, inexpensive, ultra-low pass NGS cfDNA (PRINCe) precision oncology screening approach, which has the potential to directly inform precision oncology workflows through genome-wide CNA detection and tumor content approximation (**Figure 3.1C)**.

Validation of ThruPLEX cfDNA WGS for Ion Torrent Benchtop Sequencers and cfDNA Tumor Content Approximation

Validation of cfDNA WGS using a three hour ThruPLEX RGP-0003 WGA single tube library construction approach (compatible with ≤ 50pg double stranded DNA) for rapid

sequencing on Ion Torrent benchtop sequencers was carried out on 10 normal control cfDNA samples, all of which displayed high sequencing coverage uniformity (>90%) (**Appendix C**). *In vitro* dilution experiments of sheared genomic DNA for VCaP (prostate cancer) and UMUC-5 (bladder cancer) cell lines confirmed our ability to leverage Ion Torrent cfDNA WGS for recapitulation of whole-genome copy number profiles and detection of therapeutically relevant focal amplifications (including *AR* and *EGFR* amplifications), with high observed concordance with orthogonal targeted and genome-wide copy-number profiles at tumor contents as low as 5% (see **Figures C2 and C3, Methods, Appendix C**) [46, 47].

Subsequent *in silico* dilution and downsampling experiments of cell line (sheared gDNA) and patient cfDNA WGS data facilitated development of a heuristic tumor content approximation metric (least squares statistic; LSS), while highlighting our ability to recapitulate both broad and focal copy-number alterations across tumor contents as low as 5% (see **Figures C3-5, Appendix C**). While detection of focal amplifications by low-pass cfDNA WGS is also dependent on absolute copy-number of amplified gene(s) in the tumor, high-level focal amplifications (>4 copies) are frequent across TCGA and advanced cancers [48, 49], and abundant and detectable in our patient cohort (described below). An illustrative example of a genome-wide copy-number profile from cfDNA collected from a patient (TP1337) with mCRPC after progression on second generation anti-androgens abiraterone and enzalutamide is shown in **Figure 3.2A**. TP1337 harbored focal *AR* amplification, chr8q gain, focal 2-copy *PTEN* loss, and one-copy loss on chr13 including *RB1*, representing the majority of the most common CNAs in mCRPC [45]. **Figure 3.2B** further displays the ability of our approach to detect both broad and focal CNAs down to 0.005x (~82,000 reads) in TP1337, with routine robust detection of focal amplifications in cell lines and high tumor content mCRPC samples at 0.01x coverage (**Figures**

C6 and 7). While ultra-low-pass (0.005x) is expected to have greatest clinical utility in high-tumor content cfDNA samples, these results support the fidelity of copy number profiling from cfDNA using our low-pass WGS based PRINCe approach and the capacity to leverage this workflow to both approximate tumor content and identify high level focal amplifications, a key therapeutic class of somatic alterations in cancer.

Application of PRINCe *to patient cfDNA sample cohorts and utility in disease monitoring*

To demonstrate feasibility and utility of PRINCe in representative clinical scenarios, we next assessed cfDNA from two patient cohorts, one comprised of 31 samples from 24 individual patients with metastatic colorectal, breast, or lung cancers, uterine leiomyosarcoma, sarcoma, or leukemia, and another comprised of 93 samples from 75 patients with mCRPC (including patients with both low and high volume disease) (**Appendix C**). Across the 124 total patient samples, 74 (59%) had LSS values ≥ 0.1, and thus an estimated cfDNA tumor content of >8.75% (**Figure 3.2C, Appendix C**). PRINCe enabled routine detection of actionable focal copy-number alterations (including focal *EGFR* and *FGFR1* amplifications) across patient samples in our non-mCRPC cohort (**Figure 3.2D**); combining this approach with targeted cfDNA enabled robust detection of ddPCR validated informative point mutations or indels (including *EGFR* exon 19 deletions) (see **Appendix C**). PRINCe profiling of serial cfDNA samples from several patients highlighted utility in evaluating treatment response, disease monitoring, and identification of candidate biomarkers of treatment response in a patient (PD-L1006_1) with stage IV lung adenocarcinoma who achieved a complete response to PD-L1 checkpoint inhibition immunotherapy (see **Appendix C**). While there remains clear utility in specific contexts for

profiling disease recurrence at extremely low tumor content using high-depth, ultra-sensitive or personalized sequencing/ddPCR methodologies [32, 50, 51], our results suggest substantial potential clinical utility across cancer types from low-cost identification of pre-treatment genome wide CNA profiles and cfDNA tumor content estimates via highly scalable whole-genome and targeted cfDNA NGS-based profiling strategies to monitor disease burden and molecular evidence of response.

PRINCe applied to metastatic castration resistant prostate cancer (mCRPC)

Given the potential impact of CNA detection in cfDNA—particularly *AR* amplification—on therapeutic decision-making in prostate cancer[23, 52, 53], we next focused on the 76 patients with mCRPC. All patients had progressive disease after androgen deprivation therapy, and the clinical characteristics are shown in **Appendix C**. PRINCe was carried out on 5 normal male and 93 mCPRC patient samples (including one technical replicate, TP1052B) to average whole-genome coverage of 0.32x (range: 0.02-1.30x)). Of 93 mCRPC cfDNA samples, 60 (65%) had estimated tumor contents greater than 8.75% by LSS analysis (LSS $\geq$ 0.1), our minimum threshold for accurately estimating tumor content, and were considered as high tumor content. Low-pass WGS of one cfDNA sample (TP1330) identified a single 19Mb deletion on chr20 (20q11.21-20q13.2) leading to elevated LSS, while by targeted NGS this sample also carried a *U2AF1* S34F COSMIC hotspot mutation (variant fraction = 30%, 527 covering reads; **Appendix C**), consistent with contaminating white blood cell cfDNA in the presence of concurrent myelodysplastic syndrome[54], and thus this sample was considered as low tumor content for subsequent analyses (**Figure 3.2B and Figure C8; see Methods**). In total, the 63% (59 of 93) of

mCRPC samples with estimated tumor content >8.75% represent a similar proportion of mCRPC samples to that reported as having sufficient tumor derived cfDNA for array CGH and targeted NGS based assessment described by Wyatt et al.[55].

Unsurprisingly, 68 of 93 mCRPC cfDNA samples (73%) showed evidence of detectable chromosome 8p losses and/or 8q gain (known early alterations in prostate carcinoma progression[56, 57]), including 58 of 59 (98%) high tumor content samples (**Figure C9**). In total, 14 of 93 (15%) mCRPC cfDNA samples also demonstrated detectable segmented 21q22.2 copy-number deletions consistent with deletion leading to *TMPRSS2:ERG* gene fusion, another known early event in prostate oncogenesis[58, 59] (**Figure C10**). Focal copy number alterations were also frequent, including *PTEN* deletion (20 of 59 (28.8%) high tumor content cfDNA samples, 11 (65%) of which are focal deep deletions (**Figure C10**)), and focal *AR* amplification (36 of 93 (39%) cfDNA samples, including 32 of 59 (54%) high tumor content mCRPC samples) (**Figure C10**, **Appendix C**), both of which are biomarkers of poor prognosis and/or resistance to second-line anti-androgens (abiraterone and enzalutamide), particularly when observed in cfDNA[23, 52, 53, 60] (see **Appendix C**). Focal *RB1* deletion, a frequent alteration in neuroendocrine/small-cell prostatic carcinoma[45, 61], was also detectable by our approach, with 4 samples (4.3%) (4 patients) exhibiting focal deep deletions (**Figure C10**), including 1 from a patient (TP1320) with detectable *AR* amplification, who (post-ADT and a single course of docetaxel) progressed rapidly on abiraterone over the course of 3 months on therapy with PCa-related death 4 months after cfDNA profiling (**see Appendix C**).

Notably, PRINCe assessment of cfDNA sample TP1291 paired with targeted NGS of the matched unamplified cfDNA (described below) identified a broad 1-copy copy-number loss affecting *BRCA2* and *RB1* in combination with a Clinvar pathogenic *BRCA2* germline R2494X

stop-gain SNV at a variant fraction (71%, 1,022 variant-containing reads) consistent with copy-number deletion of the non-mutated copy of the gene and biallelic inactivation of *BRCA2* (**Figure C11)**. Prior to cfDNA sample collection, the corresponding patient progressed rapidly through courses of abiraterone, enzalutamide, docetaxel, and cabazitaxel over the 11 months prior to cfDNA sample collection, consistent with known poor prognosis for *BRCA*-mutant men with prostate cancer[62], confirming important utility for cfDNA profiling in guiding PARP inhibitor treatment in patients with advanced prostate cancer[30, 31]. Additional PRINCe assessments detected a putative complex rearrangement affecting *BRCA1* in a patient with mCRPC, along with clinically relevant copy-number alterations in advanced treatment-naïve patients with heavy tumor burden (**Figure C11; Appendix C**). Overall, these results highlight our capacity to detect therapeutically relevant focal copy-number deletions from low-pass cfDNA WGS in patients with mCRPC and support potential clinical utility in informing precision oncology workflows for patients with advanced prostate cancer.

PRINCe to guide additional precision oncology testing

In the absence of immediately actionable copy-number alterations by low-pass WGS, a priori tumor content approximation from low-pass cfDNA WGS can enhance subsequent precision medicine workflows by directly informing requisite strategies or coverages needed for meaningful NGS profiling (**Figure 3.1C)**. For example, we hypothesized that in patients with relatively high cfDNA tumor content (e.g. >10%), routine tumor tissue profiling NGS strategies would be sufficient to detect relevant alterations, rather than ultra-high depth, high fidelity (e.g. single molecule barcoding) sequencing as typically performed for cfDNA NGS. Hence, we

subjected separate 1-20ng aliquots of unamplified cfDNA from 61 of our patient samples (including 46 mCRPC samples, 11 high tumor content non-mCRPC samples, and 4 male control samples with sufficient DNA; see **Appendix C**), as well as the undiluted artificial VCaP and UMUC5 cfDNA samples as positive controls, to targeted multiplexed PCR based NGS using the DNA component of the Oncomine Cancer Assay (OCP)[4], the panel being used in the NCI sponsored MATCH trial performing NGS on tumor tissue.

Sequencing of pooled patient samples resulted in a median average coverage of 1,075x (range: 42-17,944x), with average uniformity of 96.0% (higher than typically observed for FFPE DNA samples[4]). OCP on cfDNA confirmed high level *EGFR* amplification in UMUC-5, and high level *AR* amplifications in VCaP and 23 of 23 (100%) high tumor content mCRPC samples. In TP1337 (see **Fig 3.2A**), OCP on cfDNA validated all key somatic copy-number alterations detected by low-pass cfDNA and detected a 28bp *TP53* frameshift deletion (L264del28bp, variant frequency 20.8% with 504 covering reads) (**Figure 3.2A**). Of note, we observed high correlation between gene-level copy number alterations (absolute_value[targeted NGS log2(CopyNumberRatio)] ≥ 0.5) by targeted sequencing and low-pass WGS calls from PRINCe assessment of patient cfDNA samples (Pearson correlation coefficient: 0.92, $p < 0.001$), and *in silico* down-sampling experiments in patient and cell line cfDNA samples suggest mean coverages as low as 50x enable reliable detection of known putative clonal somatic point mutations, indels, and copy number variants in samples with high tumor content (**Figures C12 and C13**). Taken together, these results underscore the potential for PRINCe followed by targeted sequencing (tuned to cfDNA tumor content) as part of a high-throughput, cost-effective clinical or translational research NGS workflow.

PRINCe concordance with comprehensive tissue-based profiling

To assess the potential utility of PRINCe cfDNA assessment in the context of comprehensive tissue-based precision oncology workflows, we focused on 26 of the 76 men (34%) with mCRPC profiled by cfDNA low-pass WGS (corresponding to 31 of 93 (33%) mCRPC cfDNA samples) where synchronous or asynchronous comprehensive whole exome and whole transcriptome profiling was attempted on fresh frozen or FFPE biopsy tissue specimens (median number days between tissue- and cfDNA specimen collection: 137 (range: 0-682 days)). Of 26 men, 4 (15%) had either insufficient tumor content for comprehensive tissue profiling or incomplete tissue profiling data for analysis. Notably, all 4 men had cfDNA samples that yielded clinically informative results, including 4/4 (100%) with detectable focal *AR* amplification, while 4 of 5 patient-matched cfDNA samples were taken pre-biopsy highlighting important opportunities for optimized resource allocation in precision medicine workflows (see **Appendix C**). Collectively, this supports complementary clinical utility for plasma cfDNA profiling when paired with comprehensive tissue-based NGS workflows as a first-stage "screening" strategy.

Global copy number concordance across tissue and cfDNA profiling has been poorly explored in mCRPC and other cancers. Hence, we next assessed the 22 men with comprehensive tissue-based profiling and at least 1 profiled cfDNA sample (range of cfDNA samples per individual: 1-3), of which 18 (82%) had a cfDNA sample w/high cfDNA tumor content amenable to analysis (**Figure 3.3A**, **Appendix C**). Despite variable specimen tumor content and sample synchronicity, genome-wide segmented tissue-based copy-number profiles were highly correlated (median r = 0.87 [range: 0.54-0.95]; **Figure 3.3B**) with whole genome cfDNA segmented copy-number profiles for the 16 of 18 (89%) individuals with fresh frozen tissue specimens, and this concordance was not significantly associated with time between cfDNA and

tissue specimen collection (p=0.72, two sample t-test) (**Figure C14**, **Appendix C**). For 6 of 18 men (33%) with high tumor content cfDNA samples and tissue-based profiles, clear 21q22.2 copy-number deletions (consistent with TMPRSS2:ERG gene fusion) detected by cfDNA WGS was also detected in tissue-based DNA profiling, with *TMPRSS2:ERG* fusion isoform expression confirmed by tissue-based RNAseq in 5 of 6 men (**Figure C10, Appendix C**). Of 18 men with tissue profiling data, 12 (67%) harbored focal *AR* amplifications and 11 of 12 (92%) patient-matched high tumor content cfDNA samples show concordant detectable *AR* amplifications (**example in Figure 3.3C**; **Appendix C**). By targeted NGS of patient-matched cfDNA samples, 24/28 (86%) somatic point mutations and indels present in tissue specimens at variant fractions ≥10% targeted by our panel were detected in cfDNA samples, including 20/21 (95%) in matched high tumor content cfDNA samples and 15/15 (100%) in high tumor content cfDNA samples collected ≤ 200 days from tissue collection (**Figure C15, Appendix C**). Collectively, these results suggest PRINCe assessment of routine cfDNA samples from men in mCRPC may enable highly scalable, robust identification of putative clonal somatic alterations consistent with comprehensive profiling results from synchronous tissue samples.

Clinically relevant discrepancies between synchronous cfDNA and tissue profiles, however, were also identified. In one patient with a history of both primary prostatic adenocarcinoma and a metastatic lesion with small cell carcinoma/neuroendocrine features (TP1034/MO_1215), PRINCe assessment of synchronous (same-day) specimens detected a clear focal *AR* amplification in the cfDNA that was absent in the tissue based profiling of a prostatic neuroendocrine/small cell carcinoma focus (despite identical prioritized somatic point mutations), consistent with circulating evidence of both *AR*-driven and *AR*-independent clones (**Figure 3.4A**). Further, while previous reports suggest cfDNA clonal representation of known

early copy-number events (including chr8p/8q changes) in men with mCRPC may vary over time and therapy[24, 40], analyses in our cohort reveal stable representation of early genomic events in tissue and serial patient-matched plasma cfDNA samples (**Figure 3.4B**, **Appendix C**). Overall, these results suggest noninvasive profiling may yield high concordance with near-synchronous tissue profiling for clinically relevant molecular alterations, and may provide unique complementary advantages and opportunities for expansion into treatment-naïve patient cohorts.

Evaluating prognostic utility of cfDNA biomarkers

cfDNA detectable *AR* amplification has been reported as a biomarker predicting therapeutic resistance to second generation anti-androgens (abiraterone/enzalutamide) in several studies[23, 52, 53], while circulating tumor cell (CTC) detectable ligand independent *AR* splice variant (*AR-V7*) has been reported as predictive of abiraterone/enzalutamide resistance *and* taxane chemotherapy sensitivity[63, 64]. While our mCRPC cohort was not designed specifically to assess associations between circulating biomarkers and clinical outcome or therapeutic response, our cohort contained a large number of men on—or starting—second generation anti-androgens, as well taxane based chemotherapies. In exploratory analyses in our full cohort, we observed an enrichment of cfDNA detectable *AR* amplification in samples from patients with limited PSA response (**Figure 3.5A**), with both cfDNA detectable *AR* amplification (Kaplan-Meier log-rank test, chi-square=15.3, $p<0.0001$; **Figure 3.5B**) and elevated cfDNA tumor content (Kaplan-Meier log-rank test, chi-square=8.2, $p<0.0042$; **Figure 3.5C**) showing a significant association with reduced time on therapy. Further, stratifying by therapy (starting or

on taxane vs. abiraterone/enzalutamide), we see that both *AR* amplification (yes/no) (Kaplan-Meier log-rank test, chi-square=21.9, $p<0.0001$; **Figure 3.5D**) or cfDNA tumor content (Kaplan-Meier log-rank test, chi-square = 18.9, $p=0.0003$; **Figure 3.5E**) again show significant differences in time on therapy, suggesting cfDNA detectable *AR* amplification (and high cfDNA tumor content) may be a potentially prognostic marker for resistance to both second generation anti-androgen therapy and taxane chemotherapies. These results are consistent with those seen when restricting analyses to samples from patients on or starting therapy separately (**Figure C16**), and together confirm previous reports that cfDNA detectable *AR* amplification predicts resistance to abiraterone or enzalutamide[23, 52, 53], while supporting *AR* amplification (and high tumor content) as a more general poor prognostic factor, similar to circulating tumor cell (CTC) count[65, 66].

DISCUSSION

Many comprehensive precision oncology NGS approaches carry up-front coverage and sequencing requirements (aimed at maximizing sensitivity and specificity) that limit clinical implementation across cancer types in the current era of limited reimbursement, particularly using cfDNA (where estimated tumor content can be <0.01% in early stage disease[27]). Given current precision oncology NGS testing is typically performed in patients with multiple-therapy refractory advanced cancers usually exhibiting significant disease burden[67], here we describe a pan-cancer, rapid, inexpensive, ultra-low pass NGS cfDNA (PRINCe) based precision oncology first stage “screening” approach. Our approach can 1) direct therapy in patients with actionable CNAs, 2) guide precision oncology workflows based on cfDNA tumor content approximation in

the absence of actionable CNAs, and 3) identify genome wide CNA profiles that can be used for treatment monitoring. We show this highly scalable approach cfDNA WGS approach can be deployed at effective whole-genome coverages down to 0.01x from as little as 10pg of DNA, and that it facilitates robust detection of clinically relevant CNAs and tumor content approximation in samples with tumor contents >~10%, suggesting substantial utility as a high-throughput, cost-effective screening tool in research and clinical laboratories (with appropriate validation).

As CNAs may not be informative in all cancers, and many patients may have insufficient tumor content to identify high level CNAs, results from our approach can be used to guide additional precision oncology NGS profiling of the same cfDNA sample or fresh frozen or archived FFPE tissue-based NGS profiling, with sequencing approach and coverage tuned to tumor content. Supporting this tiered approach, we performed targeted multiplexed PCR based NGS on residual unamplified cfDNA from 61 cfDNA samples from patients with advanced cancer, confirming focal amplifications and identifying potentially informative mutations and indels at high concordance with known putative clonal alterations (25/26, 96%) in cfDNA samples with high tumor content. Comparisons between cfDNA and comprehensive tissue-based profiling in a subset of patients highlight substantial concordance for both somatic mutational and copy-number profiles, while elucidating important potentially complementary utility for cfDNA-based profiling strategies.

Limitations of our approach include the need for multiple assays, particularly in tumor types with few CNAs or where chromosomal rearrangements must be assessed. Likewise, in clinical scenarios where cfDNA tumor content is expected to be very low, up front ultra-deep cfDNA sequencing or ddPCR (as currently performed) is more appropriate, though our ability to detect known early broad copy-number events (e.g., 8p loss, 8q gain) in prostate carcinoma

progression at low cfDNA tumor content (see **Fig C9**) suggests potential expanded utility of our approach at lower tumor contents than currently implemented (paired with more comprehensive approaches when necessary). Further refinement of our tumor content approximation approach (see **Appendix C**) through assessment of informative heterozygous SNPs or incorporation of a matched normal genomic DNA would enhance the precision and lower limits of our tumor-derived cfDNA fraction estimates, though costs and feasibility in a clinical sequencing workflow are key considerations. While PRINCe is necessarily limited to megabase resolution for copy number alteration detection at ultra-low-pass (~0.01x) whole-genome coverage, smaller (multi-kilobase) clinically relevant focal alterations (including focal *PTEN* deletion) can clearly be detected at 0.01-0.1x genome-wide coverages with sufficient cfDNA tumor content (**Figure 3.5B**). Importantly, our approach can be routinely completed in 2-3 days and when performed at 50% capacity on an Ion Torrent Proton sequencer (currently limited by Ion Torrent barcodes incorporated in ThruPLEX library construction), 96 samples could be sequenced per single Ion Torrent Proton P1 chip at list reagent costs of ~$70 per sample for library construction and NGS. Taken together, these observations suggest the proposed workflow may be amenable to high volume, cost-effective ultra-low-pass WGS screening protocols.

Applied to a large mCRPC cohort, our approach showed high overall concordance between our cfDNA genome-wide CNA profiles with tissue-based profiles derived from whole exome sequencing in a precision medicine program[45]. In addition, we demonstrated that cfDNA detectable *AR* amplification not only predicts poor response to second generation anti-androgens, consistent with other published reports[23, 53], but it also portends poor prognosis for patients treated with taxane based chemotherapy. Hence, cfDNA detectable *AR* amplification may be a more general poor prognostic factor, unlike *AR-v7*, which has been reported to confer

resistance to anti-androgens and sensitivity to taxanes[63, 64]. An important limitation of these results is that this was not assessed in the context of a clinical trial, and men in our study treated with taxanes were more advanced and had been treated with more lines of therapy post-ADT. Hence, prospective confirmation of our findings will be required.

In summary, we have demonstrated the feasibility and potential utility of PRINCe, a broadly applicable, rapid, inexpensive cfDNA WGS screening assay for precision oncology that can robustly detect clinically informative CNAs from cfDNA at low tumor content using effective whole-genome coverage as low as 0.01x. This screen, while most informative in those patients with actionable CNAs and tumor content >10%, can nevertheless be used to guide additional testing in all patients based on cfDNA tumor content approximation. Our approach highlights important potential clinical utility when paired with targeted cfDNA NGS and/or tissue-based workflows, and demonstrates unique possibilities for inexpensive disease monitoring. More generally, our study supports the potential utility of tiered approaches in precision oncology, rather than using costlier front-line approaches defined by performance necessary in the extremes.

METHODS

TCGA Data Analysis

TCGA pan-cancer copy number analyses were run on somatic segmented Affymetrix SNP6 array-based copy-number calls for 11,576 tumor samples across 32 tumor types contained in the January 28, 2016 TCGA GDAC Firehose standard data run (stddata__2016_01_28)[68] (see **Appendix C**).

Cell-free DNA extraction

Five milliliters of peripheral blood were collected for 93 samples from 76 patients with mCRPC and 10 healthy controls (5 male, 5 female) using K2 EDTA blood collection tubes (Cat: 366643, BD, NJ), and cfDNA was isolated as described (see **Appendix C**). For 31 samples from 24 patients with other advanced cancers, 10 mL peripheral blood was collected using Streck Cell-Free DNA BCT tube (Streck; NE) and cfDNA was isolated as detailed (see **Appendix C**).

VCaP and UMUC-5 In vitro Dilution

We carried out *in vitro* dilution experiments using serially diluted genomic DNA from 1) VCaP cells (metastatic prostate cancer cell line) with normal male human cell-free genomic DNA at 50%, 25%, 10%, 5%, 1% and 0% dilutions, and 2) UMUC-5 cells (urothelial cancer cell line) with normal male human cell-free genomic DNA at 50%, 10%, 5%, 0% dilutions. Cell line DNA was fragmented to approximately 180bp by Covaris AFA (Woburn, MA) focused ultrasonication. Library preparation and sequencing from undiluted and serial dilution samples was performed as for patient samples described below.

ThruPLEX Library preparation

Whole genome amplified (WGA) libraries were prepared from either cell-free DNA (cfDNA) isolated from plasma samples (median of 2.9ng cfDNA, interquartile range [IQR] 1.73-5.79ng, see **Appendix C**) or Covaris-sheared and size selected (~ 180bp size) VCaP (1.9ng) or

UMUC-5 (2.0ng) genomic DNA (gDNA) using the ThruPLEX RGP-0003 prototype (Takara Bio USA; Ann Arbor, MI) according to the manufacturer's protocol. Libraries were quantified using Ion Library Quantification kit by qPCR, and sequenced with 2-16 samples per Proton PI chip on an Ion Proton sequencer (Ion Torrent, Carlsbad, CA) according to the manufacturer's instructions.

Low-pass WGS and copy-number detection

Sequencing alignment and coverage analyses were performed using Torrent Suite version 5.0.2 (Ion Torrent, Carlsbad, CA). Genome-wide copy number alterations were first called from aligned, non-PCR-duplicate reads using the QDNASeq R package (version 1.6.1) [69]. Segmented copy-number events were identified using bin-level corrected, median- and control-normalized read counts using the circular binary segmentation algorithm implemented by the DNACopy (1.44.0) R package, and final segment- and bin-level copy-number values were used for subsequent analyses as described (see **Appendix C**). Focal CNAs were defined as CNAs 1.5-20Mb long with a log2(CopyNumberRatio) $\geq$ 0.2.

Targeted sequencing: Oncomine Comprehensive Assay (OCP)

For 61 patient cfDNA samples (see **Appendix C**) and both sheared UMUC-5 and VCaP gDNA samples, we performed targeted NGS using the DNA component of the OCP, a custom multiplexed PCR-based panel of 2,530 amplicons targeting 126 genes[4]. Library preparation, data analysis, and variant and copy-number annotation and prioritization was carried out

essentially as described for each sample [4, 70-72] using validated in house pipelines (**Appendix C**).

In silico experiments and tumor content approximation

To establish theoretical segment-level copy-number distributions for tumor content approximation and examine efficacy across variable effect whole-genome coverages (0.005-0.01x), we carried out serial *in silico* dilution and downsampling experiments on artificial cfDNA VCaP and UMUC-5 WGS data and patient cfDNA samples (see **Appendix C).** Using computational experiments on *in vitro* and *in silico* VCaP and UMUC-5 cell line dilution data as described in **Appendix C**, a heuristic least squares based distance metric (LSS) was used to approximate tumor content from whole-genome copy-number data, and guide tumor content approximation for patient samples, with low tumor content samples (LSS < 0.1) specifically scanned for focal CNAs as described (see **Appendix C**).

Cell line cfDNA WGS vs COSMIC array-based CN calls

To evaluate the capacity of low-pass cfDNA WGS to detect copy-number alterations across variable tumor content, segmented cfDNA WGS copy-number calls for VCaP and UMUC-5 *in vitro* dilutions were compared to publically available COSMIC and targeted NGS copy-number calls, respectively (see **Appendix C**).

Concordance with tissue-based whole-exome sequencing copy-number profiles

Segmented log2 copy number ratio and point mutation data from whole-exome sequencing of fresh frozen tissue specimens[10, 45] was available for 22 of 26 patients also profiled by cfDNA low-pass WGS and compared to patient-matched cfDNA WGS profiles (see **Appendix C**).

Clinical information

All clinical and outcome information was collected, retrieved, and analyzed from internal patient tracking databases and University of Michigan Health System (UMHS) electronic health records by IRB-approved personnel.

Statistical analyses

All statistical analyses described were carried out in R (3.2.3).

Figure 3.1: **Leveraging tumor-derived cfDNA distribution in advanced cancer to develop a pan-cancer, rapid, inexpensive, ultra-low pass whole genome next generation sequencing (NGS) cfDNA precision oncology workflow (PRINCe).**

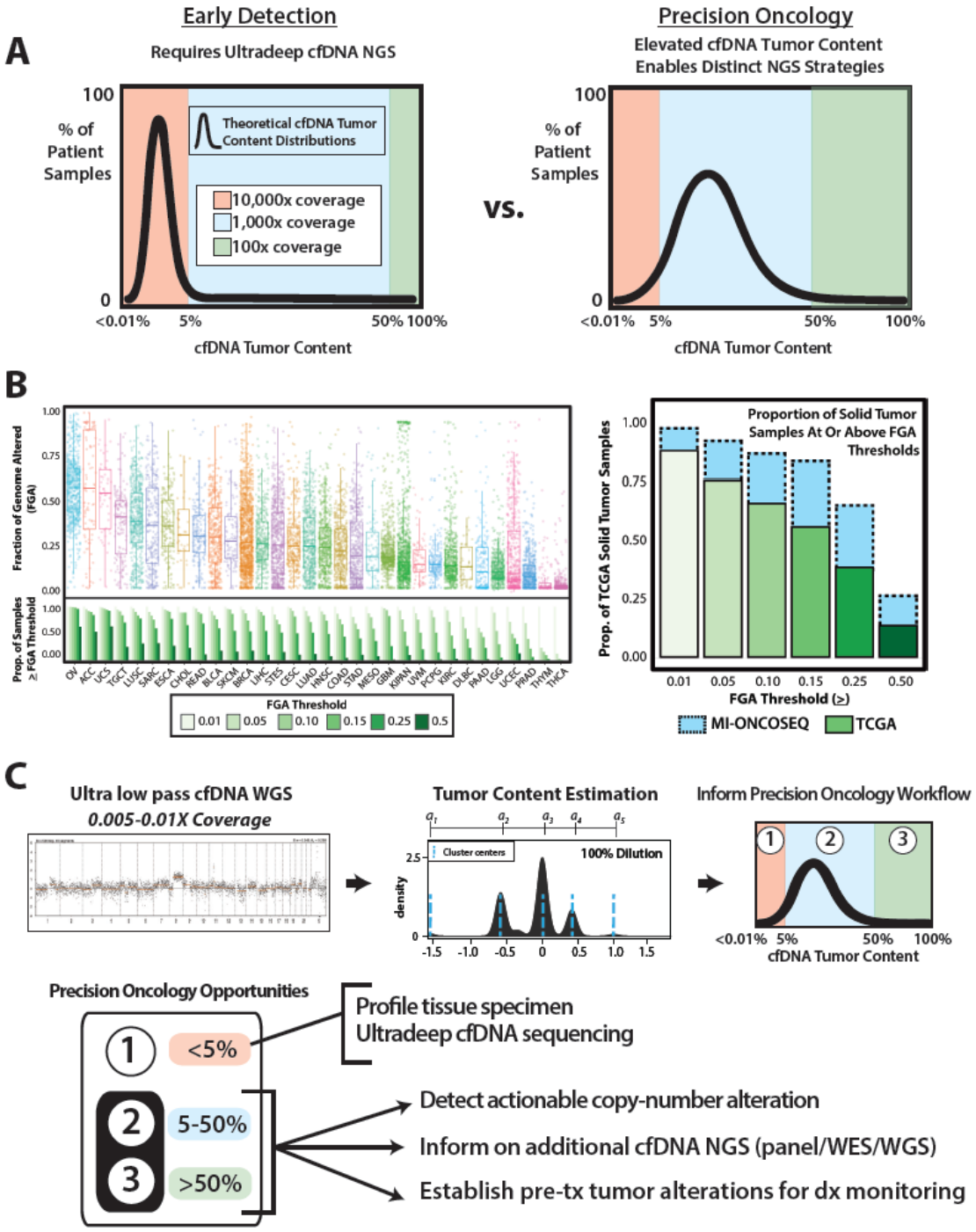


Figure 3.1 A. Theoretical cfDNA tumor content distributions and typical next-generation sequencing (NGS) coverage requirements for mutation profiling are presented for early detection (left) and precision oncology in advanced disease (right) applications. In early detection context, the majority of cfDNA samples are expected to have a low proportion of tumor-derived cfDNA fragments (e.g., <<5%), whereas advanced cancers have an elevated proportion of tumor-derived cfDNA. Tumor content requiring ultra-deep, extreme-fidelity (e.g. 10,000x coverage) targeted sequencing are shaded red, while those amenable to targeted sequencing on larger panels or whole-exome/whole-genome (WES/WGS) are shaded blue and green, respectively. **B.** Copy number alterations (CNAs) are frequent across human cancers. The fraction of the genome altered (FGA, see **Methods**) by CNAs in 11,576 The Cancer Genome Atlas (TCGA) samples from 32 solid tumor types is shown across multiple thresholds

(overall cohort on left, individual tumor types on right). Increased FGA in a cohort of 129 advanced/metastatic cancers (prostate, kidney, lung and breast cancers) subjected to exome sequencing in the MI-ONCOSEQ program (plotted on the right panel) is seen in comparison to the TCGA cohort, consistent with increasing frequency of CNAs in advanced/metastatic cancers. **C.** Schematic for **p**an-cancer, **r**apid, **i**nexpensive, ultra-low pass **N**GS **c**fDNA workflow (PRINCe). Segmented copy-number calls from ultra-low-pass cfDNA whole-genome sequencing are generated, followed by CNA-clustering based tumor content approximation to inform on precision oncology management. In patients with sufficient tumor content by PRINCe (e.g. >5-10%), CNA profiles may directly guide treatment (if focal targetable alterations are identified), enable routine panel, WGS, or WES based cfDNA NGS tuned to tumor content, as well as establish pre-treatment (tx) CNA profiles for disease (dx) monitoring post-therapy. More costly ultra-deep, extreme fidelity cfDNA and tissue based profiling can thus be reserved for patients with low cfDNA tumor content.

Figure 3.2: cfDNA tumor content approximation and disease monitoring applications for targeted and ultra-low-pass whole-genome sequencing (WGS) of cell-free DNA from patients with advanced cancer.

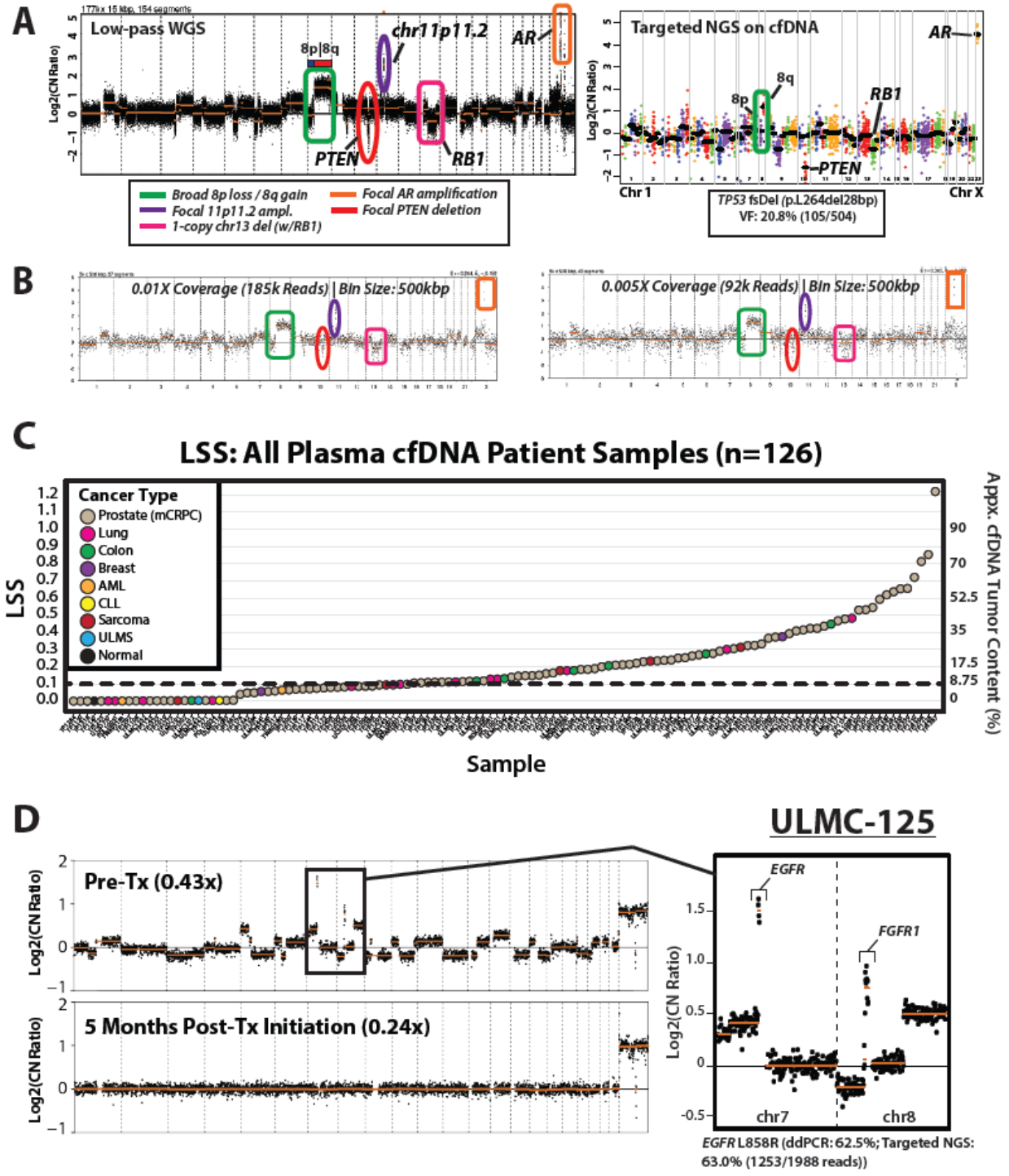


Figure 3.2 A. Genome-wide log2(CopyNumberRatio) (Log2CN) calls for TP1337, a high tumor content cfDNA sample from a patient with mCRPC, are displayed for low-pass WGS data (0.82x whole-genome coverage) and targeted NGS data. Key copy-number alterations (CNA) detected are circled, including broad gain of 8q (green), focal amplification of chr11p11.2 (purple) and *AR* (orange), and focal *RB1* (1-copy; pink) and *PTEN* (2-copy; red) deletions. Copy number and mutation data from deep coverage targeted NGS data is provided at right from unamplified TP1337 cfDNA (1,102x targeted NGS coverage) using the DNA component of the Oncomine Comprehensive Assay (OCP), a pan cancer NGS panel developed for FFPE tissue samples. For genes with sufficient amplicons for CNA calling, amplicon (dots) and gene (black bars)-level log base 2 copy number ratio

(Log2 [CN Ratio]) estimates (compared to a composite reference sample) are plotted. All CNAs seen by low-pass WGS are detected via targeted NGS CNA analysis (chr11p11.2 is not targeted by OCP). A prioritized high confidence somatic 28bp frameshift deletion in *TP53* (p.L264del28bp; variant fraction (VF) = 20.8% (105/504 total sequencing reads)) detected by OCP is shown in the inset box. **B.** *In silico* downsampling experiments highlight the ability to detect both focal and broad copy-number alterations from TP1337 cfDNA WGS data at whole-genome coverages down to 0.005x. Bin size and number of high-quality (MAPQ ≥37) mapped reads used for copy-number analysis are indicated at each coverage, and regions affected by copy-number alterations detected in original low-pass WGS are circled. **C.** Distribution of cfDNA tumor content estimates (right axis) from least-squares distance metric (LSS) values (left axis) for 124 patient cfDNA samples (123 from patients with advanced cancer, including TP1178 (from a patient with untreated advanced prostate cancer), along with 1 normal control sample (TP1147). All patient samples are colored by cancer type (indicated in the legend). **D.** Low-pass WGS copy-number for pre- and post- EGFR inhibitor (erlotonib) treatment plasma cfDNA samples from ULMC-125, a patient with metastatic lung cancer. Multiple whole-chromosome and arm-level gains/losses as well as focal amplifications are present in the pre-treatment cfDNA sample with high tumor content. A zoomed view of chromosomes 7 and 8 show focal *EGFR* and *FGFR1* amplifications in the pre-treatment sample (an activating *EGFR* L858R mutation was previously detected at 62.5% variant fraction by digital droplet PCR [ddPCR]). Low-pass WGS sequencing of a cfDNA sample taken 5 months post-treatment initiation (bottom) showed no detectable copy-number alterations genome-wide, and no detectable L858R mutation by ddPCR analysis (L858R variant fraction: 0.0%). Low-pass WGS copy-number bin size: 500kbp; segmentation p-value threshold: 0.01.

Figure 3.3: Comparison of synchronous and asynchronous tissue and cfDNA biospecimens collected from patients with metastatic castration-resistant prostate cancer (mCRPC) yields highly concordant genome-wide copy number profiles.

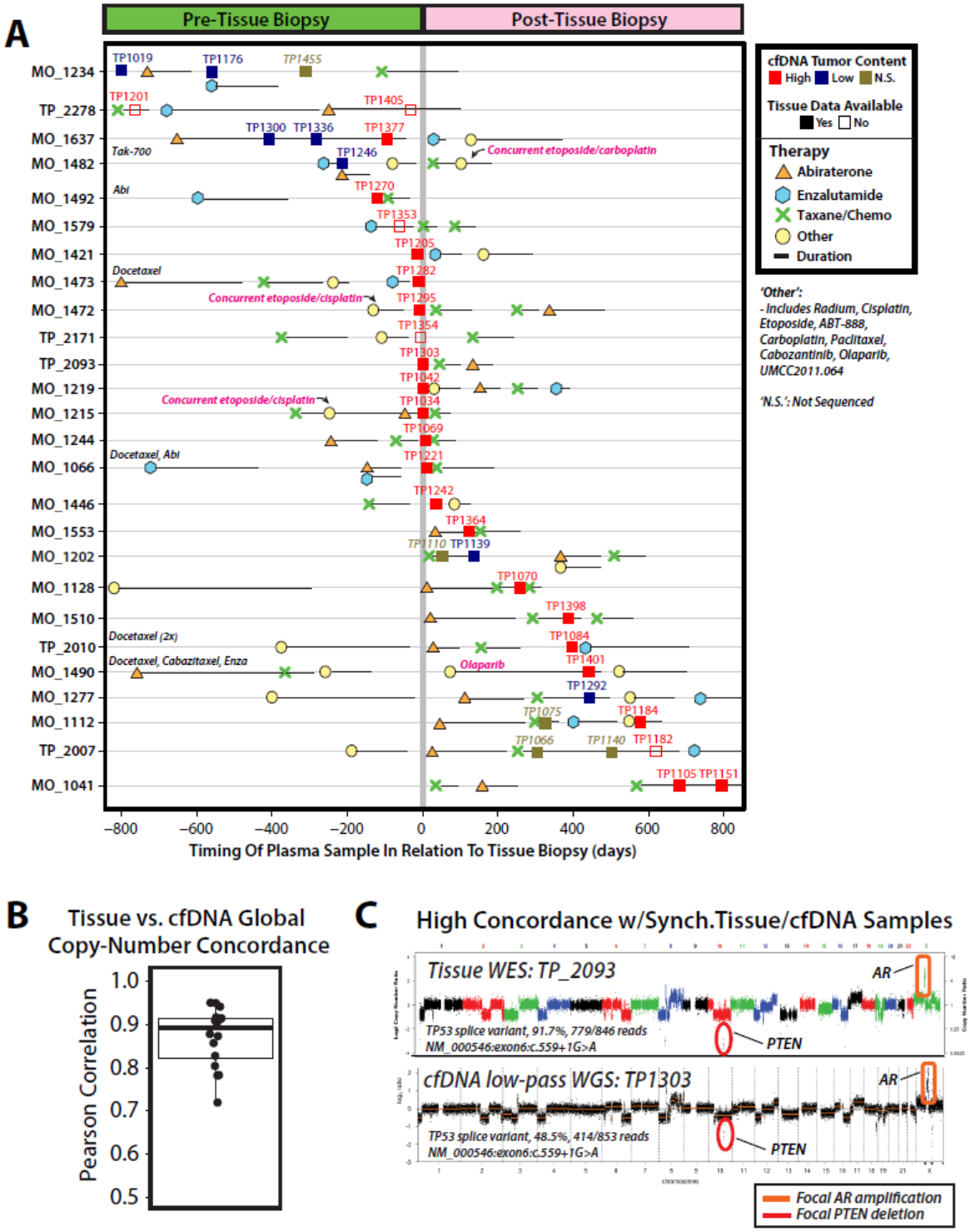


Figure 3.3 A. Treatment and cfDNA sample collection timeline plotted in relation to tissue specimen collection date for 26 men with metastatic castration-resistant prostate cancer (mCRPC) eligible for tissue-based comprehensive whole-exome and whole-transcriptome NGS profiling. Treatment start and cfDNA sample dates are plotted relative to tissue specimen collection date (denoted by solid vertical gray line) for each individual. As indicated in the legend, treatments have been divided into 4 separate categories, including: abiraterone (orange triangle), enzalutamide (blue hexagon), taxane-based chemotherapy (green 'X') and

other (yellow circle), and treatment duration is indicated by solid black horizontal lines extending rightward from treatment start dates. Therapies categorized as 'other' include: radium, cisplatin, etoposide, ABT-888, carboplatin, paclitaxel, cabozantinib, olaparib, and UMCC2011.064. Where appropriate, 'other' treatment including etoposide and cisplatin or carboplatin for individuals with prostate cancer containing small cell/neuroendocrine features are noted. As indicated in the legend, samples are colored by LSS-based tumor content approximation with high (LSS > 0.1, red), low (LSS < 0.1, blue), and not sequenced ('N.S.', brown). For a subset of men, tissue-based molecular data was not available, as indicated by filled (tissue data available) or unfilled (tissue data not available) squares. Displayed sample dates are restricted to +/- 800 days from date of tissue specimen collection, and therapies administered >800 days before tissue specimen collection are written at the left-hand side of corresponding individual timelines. **B.** Correlations between genome-wide tissue and cfDNA segmented copy-number profiles are plotted for 16 patients with available comprehensive tissue NGS profiling data and PRINCe assessment of ≥1 high tumor content cfDNA sample (see **Methods**). Each point represents the correlation of genome-wide copy number profile for a single cfDNA sample as compared to the patient-matched tissue-based copy-number profile. A box-and-whisker plot behind points indicates the interquartile range (IQR), with the top and bottom of box representing 25^{th} and 75^{th} percentile, respectively, while bold horizontal line within the box represents the median correlation value. Whiskers stretch to 1.5 times the IQR for this sample distribution. **C.** Tissue whole exome sequencing (WES) (top; tissue id: TP_2093) and cfDNA low-pass whole genome sequencing (WGS) (bottom; cfDNA id: TP1303) genome-wide copy-number profiles for biospecimens collected on the same day from a patient with mCRPC (TP_2093). Genome-wide copy-number concordance is statistically significant (Pearson correlation coefficient: 0.94, $p < 0.001$), and focal 2-copy deletion of *PTEN* and focal high-level *AR* amplification are cleared detected in both the tissue and cfDNA as indicated. A *TP53* splice variant (NM_000546:exon 6:c.559+1G>A) identified via WES tissue profiling (91.7% variant fraction (VF), 846 covering reads) is also detected by cfDNA targeted NGS (48.5% VF, 853 covering reads).

Figure 3.4: Unique precision oncology considerations identified via serial and synchronous tissue and cfDNA NGS-based profiling in patients with advanced prostate cancer.

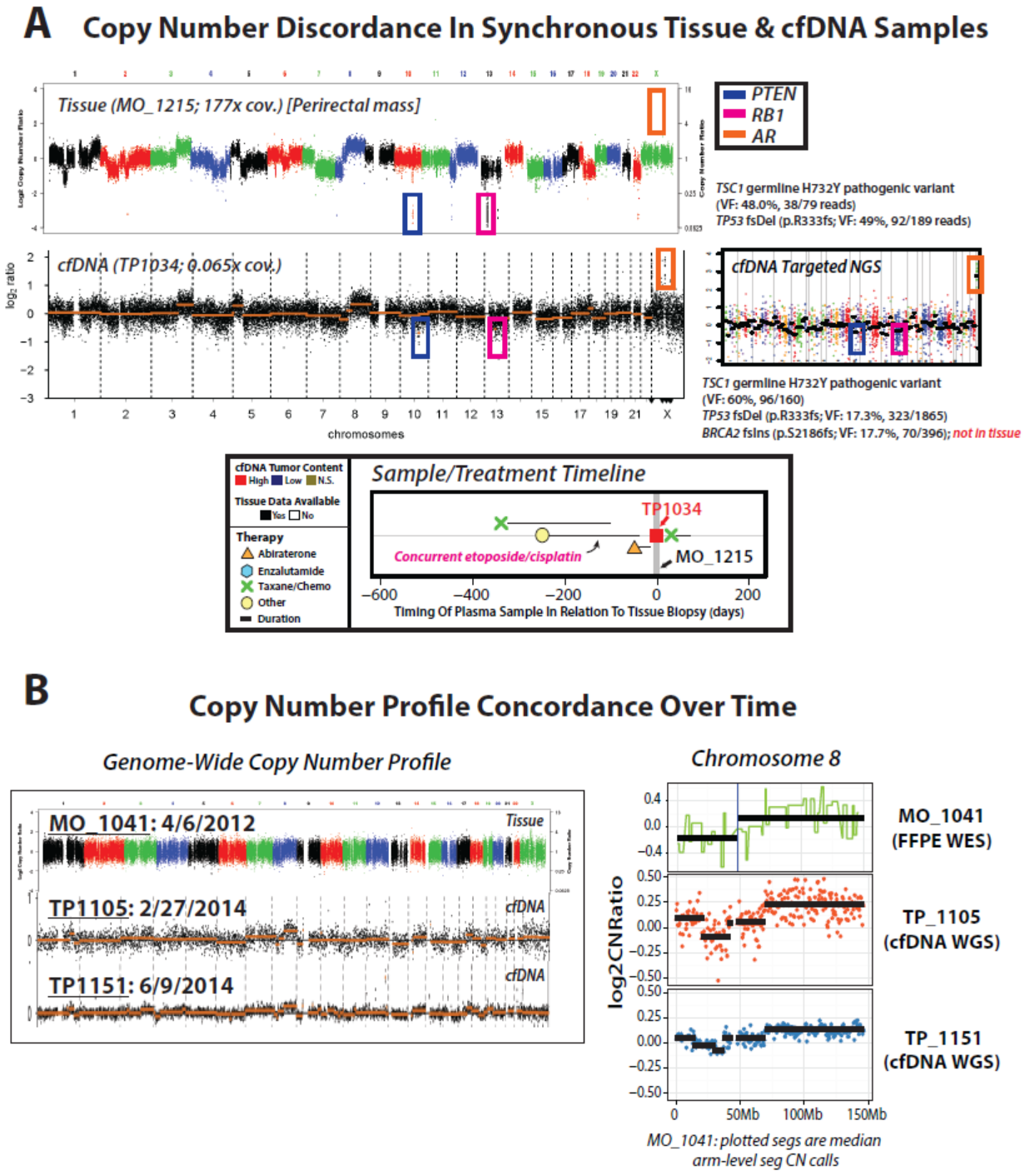


Figure 3.4 A. Genome-wide (tissue and cfDNA) and targeted (cfDNA only) NGS copy number profiles are displayed, along with treatment and sample timeline, for synchronous (same-day) tissue and cfDNA specimens from a patient with metastatic castration-resistant prostate cancer (mCRPC) with a history of both primary prostatic adenocarcinoma and a metastatic lesion with small cell carcinoma/neuroendocrine features. Tissue whole exome sequencing (WES) copy number analysis of a frozen perirectal mass tissue biopsy specimen (top left) revealed focal, deep deletions in both *PTEN* and *RB1*, and no *AR* copy-number alterations, consistent with histological reports of high-grade poorly differentiated carcinoma with neuroendocrine features. Individual dots in tissue WES copy-number profile represent exon-level copy-number estimates displayed in genome order, and

dots are colored by corresponding chromosomes. cfDNA low-pass whole genome sequencing (WGS) copy number profiling identified focal deep deletions in *PTEN* and *RB1*, as well as high level focal *AR* amplification, highlighting circulating evidence of both *AR*-driven and *AR*-independent clones. Individual dots in cfDNA low-pass WGS plot represent bin-level copy-number estimates displayed in genome order (left to right), with segmented copy-number alterations represented by orange horizontal lines. Both tissue (WES) and cfDNA (targeted NGS; copy-number prfile) mutation profiling identified a *TSC1* germline H732Y pathogenic variant (tissue: 48% variant fraction (VF), 79 covering reads; cfDNA: 60% VF, 160 covering reads) and somatic *TP53* frameshift deletion (p.R333fs; tissue: 49% VF, 189 covering reads; cfDNA: 17%, 1865 covering reads), while cfDNA targeted NGS identified a *BRCA2* frameshift insertion (p.S2186fs; 18% VF, 396 covering reads) not present in the tissue sample, further supporting detection of multiple clones via cfDNA PRINCe assessment. The cfDNA targeted NGS copy number profile is presented at right, showing confirmation of focal *PTEN* and *RB1* deletions along with high-level focal *AR* amplification as seen by low-pass WGS. Zoomed view of treatment and sample timeline for this patient is presented at bottom, as previously described (see **Fig 3**). **B.** Genome-wide (left) and chromosome 8 (right) copy number profiles from multiple biospecimens taken over time from a single patient with metastatic castration-resistant prosate cancer (mCRPC) (tissue id: MO_1041; cfDNA ids: TP1105 and TP1151). WES of a formalin fixed paraffin embedded (FFPE) tissue biopsy specimen (top left) revealed low but detectable tumor content, and identified copy-number loss affecting 8p and arm-level gain of 8q (at right). Low-pass WGS of a cfDNA specimen collected almost 2 years after tissue biopsy (TP1105, middle left) revealed elevated cfDNA tumor content with frequent copy-number alterations genome-wide, including copy-number loss affecting chr8p and arm-level gain of chr8q (displayed at right), as detected in initial tissue profiling. A subsequent cfDNA sample (TP1151) again showed detection of elevated cfDNA tumor content and a highly concordant genome-wide copy number profile, with faithful representation of the 8p loss and 8q gain events detected in previous specimens. Overall, these results highlight the consistent representation of early genomic events as inferred from circulating tumor DNA profiled in our cohort.

Figure 3.5: Exploratory analyses of association between circulating biomarkers and outcome in patients with metastatic castration-resistant prostate cancer (mCRPC) supports cfDNA detectable *AR* amplification as a poor overall prognostic factor independent of treatment type

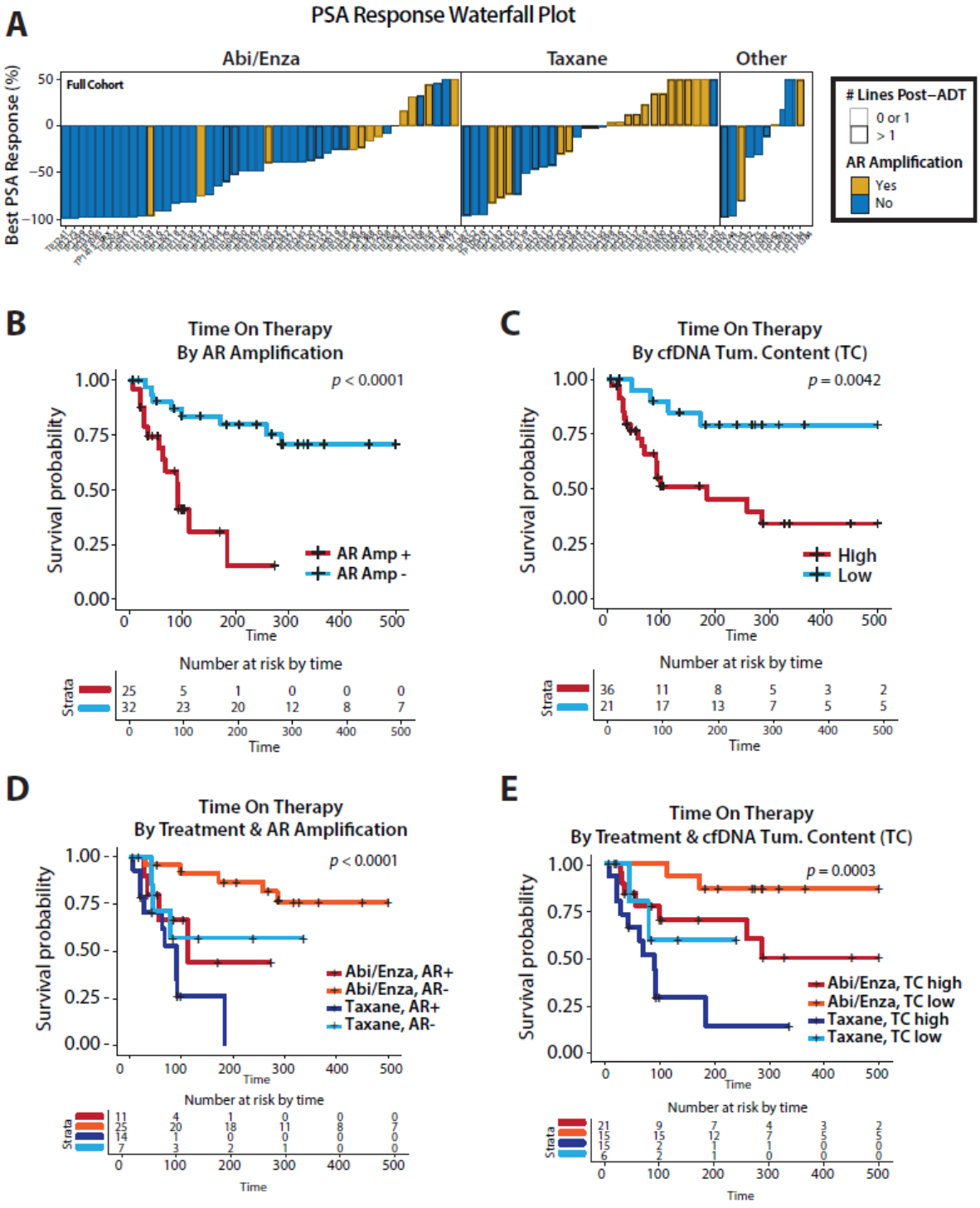


Figure 3.5. A. Waterfall plot summarizing prostate specific antibody (PSA) response for all samples from men with mCRPC with complete PSA data (n=90). Height of bars represent the percentage change in PSA response as calculated by subtracting the

PSA level at sample date from the best PSA observed after sample date while on the current or initiated treatment, and dividing by starting PSA value. Bars are ordered horizontally within treatment category (Abi/Enza, Taxane, or Other) by PSA response. Bars are colored by cfDNA detectable AR amplification status (yellow = cfDNA detectable AR amplification; gray = no cfDNA detectable AR amplification) and bars corresponding to samples taken from men who have received more than one line of therapy post-ADT are outlined in bold. B-E. Kaplan-Meier survival curves are plotted for analyses exploring association between cfDNA detectable AR amplification (B, D) or cfDNA tumor content (C, E) and total time on therapy in both unstratified (B-C) and stratified (D-E; by treatment type) analyses of our mCRPC cohort. Unstratified analysis of single cfDNA samples from men on or starting taxane-based chemotherapy or second-generation anti-androgens abiraterone or enzalutamide (n=57 men) highlight significant differences in time on therapy for both (B) cfDNA detectable AR amplification (Kaplan-Meier log-rank test, chi-square=15.3, $p<0.0001$) and (C) elevated cfDNA tumor content (Kaplan-Meier log-rank test, chi-square=8.2, $p<0.0042$). Analyses stratified by treatment (starting or on taxane vs. abiraterone/enzalutamide) show (D) cfDNA detectable AR amplification (yes/no) (Kaplan-Meier log-rank test, chi-square=21.9, $p<0.0001$) and (E) cfDNA tumor content (Kaplan-Meier log-rank test, chi-square = 18.9, $p=0.0003$) again demonstrate significant differences in time on therapy. Survival curves are colored by corresponding strata, and risk tables at selected timepoints are displayed below each Kaplan-Meier plot.

Chapter III References

1. Roychowdhury, S. and A.M. Chinnaiyan, *Translating cancer genomes and transcriptomes for precision oncology.* CA Cancer J Clin, 2016. **66**(1): p. 75-88.
2. Hyman, D.M., et al., *Precision medicine at Memorial Sloan Kettering Cancer Center: clinical next-generation sequencing enabling next-generation targeted therapy trials.* Drug Discov Today, 2015. **20**(12): p. 1422-8.
3. Cheng, D.T., et al., *Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology.* J Mol Diagn, 2015. **17**(3): p. 251-64.
4. Hovelson, D.H., et al., *Development and validation of a scalable next-generation sequencing system for assessing relevant somatic variants in solid tumors.* Neoplasia, 2015. **17**(4): p. 385-99.
5. Grasso, C., et al., *Assessing copy number alterations in targeted, amplicon-based next-generation sequencing data.* J Mol Diagn, 2015. **17**(1): p. 53-63.
6. Beadling, C., et al., *Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping.* J Mol Diagn, 2013. **15**(2): p. 171-6.
7. Wagle, N., et al., *High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing.* Cancer Discov, 2012. **2**(1): p. 82-93.
8. Zheng, Z., et al., *Anchored multiplex PCR for targeted next-generation sequencing.* Nat Med, 2014. **20**(12): p. 1479-84.
9. Mody, R.J., et al., *Integrative Clinical Sequencing in the Management of Refractory or Relapsed Cancer in Youth.* JAMA, 2015. **314**(9): p. 913-25.
10. Roychowdhury, S., et al., *Personalized oncology through integrative high-throughput sequencing: a pilot study.* Sci Transl Med, 2011. **3**(111): p. 111ra121.
11. Lih, C.J., et al., *Analytical Validation and Application of a Targeted Next-Generation Sequencing Mutation-Detection Assay for Use in Treatment Assignment in the NCI-MPACT Trial.* J Mol Diagn, 2016. **18**(1): p. 51-67.
12. Overman, M.J., et al., *Utility of a molecular prescreening program in advanced colorectal cancer for enrollment on biomarker-selected clinical trials.* Ann Oncol, 2016.
13. Meric-Bernstam, F., et al., *Incidental germline variants in 1000 advanced cancers on a prospective somatic genomic profiling protocol.* Ann Oncol, 2016. **27**(5): p. 795-800.
14. Meric-Bernstam, F., et al., *Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials.* J Clin Oncol, 2015. **33**(25): p. 2753-62.
15. Gray, S.W., et al., *Oncologists' and cancer patients' views on whole-exome sequencing and incidental findings: results from the CanSeq study.* Genet Med, 2016.
16. Van Allen, E.M., et al., *Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine.* Nat Med, 2014. **20**(6): p. 682-8.
17. Janku, F., et al., *Actionable mutations in plasma cell-free DNA in patients with advanced cancers referred for experimental targeted therapies.* Oncotarget, 2015. **6**(14): p. 12809-21.

18. Schwaederle, M., et al., *Detection rate of actionable mutations in diverse cancers using a biopsy-free (blood) circulating tumor cell DNA assay.* Oncotarget, 2016. **7**(9): p. 9707-17.
19. Schwarzenbach, H., D.S. Hoon, and K. Pantel, *Cell-free nucleic acids as biomarkers in cancer patients.* Nat Rev Cancer, 2011. **11**(6): p. 426-37.
20. Thierry, A.R., et al., *Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA.* Nat Med, 2014. **20**(4): p. 430-5.
21. Oxnard, G.R., et al., *Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA.* Clin Cancer Res, 2014. **20**(6): p. 1698-705.
22. Leary, R.J., et al., *Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing.* Sci Transl Med, 2012. **4**(162): p. 162ra154.
23. Romanel, A., et al., *Plasma AR and abiraterone-resistant prostate cancer.* Sci Transl Med, 2015. **7**(312): p. 312re10.
24. Carreira, S., et al., *Tumor clone dynamics in lethal prostate cancer.* Sci Transl Med, 2014. **6**(254): p. 254ra125.
25. Bettegowda, C., et al., *Detection of circulating tumor DNA in early- and late-stage human malignancies.* Sci Transl Med, 2014. **6**(224): p. 224ra24.
26. Forshew, T., et al., *Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA.* Sci Transl Med, 2012. **4**(136): p. 136ra68.
27. Newman, A.M., et al., *An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.* Nat Med, 2014. **20**(5): p. 548-54.
28. Murtaza, M., et al., *Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA.* Nature, 2013. **497**(7447): p. 108-12.
29. Chan, K.C., et al., *Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing.* Clin Chem, 2013. **59**(1): p. 211-24.
30. Goodall, J., et al., *Circulating Free DNA to Guide Prostate Cancer Treatment with PARP Inhibition.* Cancer Discov, 2017.
31. Quigley, D., et al., *Analysis of Circulating Cell-free DNA Identifies Multi-clonal Heterogeneity of BRCA2 Reversion Mutations Associated with Resistance to PARP Inhibitors.* Cancer Discov, 2017.
32. Newman, A.M., et al., *Integrated digital error suppression for improved detection of circulating tumor DNA.* Nat Biotechnol, 2016. **34**(5): p. 547-55.
33. Lo, Y.M., et al., *Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus.* Sci Transl Med, 2010. **2**(61): p. 61ra91.
34. Norton, M.E., et al., *Cell-free DNA analysis for noninvasive examination of trisomy.* N Engl J Med, 2015. **372**(17): p. 1589-97.
35. Amant, F., et al., *Presymptomatic Identification of Cancers in Pregnant Women During Noninvasive Prenatal Testing.* JAMA Oncol, 2015. **1**(6): p. 814-9.
36. Bianchi, D.W., et al., *Noninvasive Prenatal Testing and Incidental Detection of Occult Maternal Malignancies.* JAMA, 2015. **314**(2): p. 162-9.
37. Yu, S.C., et al., *Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing.* Proc Natl Acad Sci U S A, 2014. **111**(23): p. 8583-8.

38. Heitzer, E., et al., *Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing.* Genome Med, 2013. **5**(4): p. 30.
39. Jiang, P., et al., *Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients.* Proc Natl Acad Sci U S A, 2015. **112**(11): p. E1317-25.
40. Ulz, P., et al., *Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer.* Nat Commun, 2016. **7**: p. 12008.
41. Belic, J., et al., *Rapid Identification of Plasma DNA Samples with Increased ctDNA Levels by a Modified FAST-SeqS Approach.* Clin Chem, 2015. **61**(6): p. 838-49.
42. Diaz, L.A., Jr. and A. Bardelli, *Liquid biopsies: genotyping circulating tumor DNA.* J Clin Oncol, 2014. **32**(6): p. 579-86.
43. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-40.
44. Hieronymus, H., et al., *Copy number alteration burden predicts prostate cancer relapse.* Proc Natl Acad Sci U S A, 2014. **111**(30): p. 11139-44.
45. Robinson, D., et al., *Integrative clinical genomics of advanced prostate cancer.* Cell, 2015. **161**(5): p. 1215-28.
46. Teles Alves, I., et al., *Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line.* Hum Genet, 2013. **132**(6): p. 709-13.
47. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(Database issue): p. D805-11.
48. Robinson, D.R., et al., *Integrative clinical genomics of metastatic cancer.* Nature, 2017. **548**(7667): p. 297-303.
49. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-1140.
50. Tie, J., et al., *Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer.* Sci Transl Med, 2016. **8**(346): p. 346ra92.
51. Garcia-Murillas, I., et al., *Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer.* Sci Transl Med, 2015. **7**(302): p. 302ra133.
52. Azad, A.A., et al., *Androgen Receptor Gene Aberrations in Circulating Cell-Free DNA: Biomarkers of Therapeutic Resistance in Castration-Resistant Prostate Cancer.* Clin Cancer Res, 2015. **21**(10): p. 2315-24.
53. Wyatt, A.W., et al., *Genomic Alterations in Cell-Free DNA and Enzalutamide Resistance in Castration-Resistant Prostate Cancer.* JAMA Oncol, 2016.
54. Bacher, U., et al., *Investigation of 305 patients with myelodysplastic syndromes and 20q deletion for associated cytogenetic and molecular genetic lesions and their prognostic impact.* Br J Haematol, 2014. **164**(6): p. 822-33.
55. Wyatt, A.W., et al., *Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer.* Genome Biol, 2014. **15**(8): p. 426.
56. El Gammal, A.T., et al., *Chromosome 8p deletions and 8q gains are associated with tumor progression and poor prognosis in prostate cancer.* Clin Cancer Res, 2010. **16**(1): p. 56-64.
57. Sato, K., et al., *Clinical significance of alterations of chromosome 8 in high-grade, advanced, nonmetastatic prostate carcinoma.* J Natl Cancer Inst, 1999. **91**(18): p. 1574-80.

58. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.
59. Perner, S., et al., *TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion.* Am J Surg Pathol, 2007. **31**(6): p. 882-8.
60. Lotan, T.L., et al., *PTEN loss is associated with upgrading of prostate cancer from biopsy to radical prostatectomy.* Mod Pathol, 2015. **28**(1): p. 128-37.
61. Beltran, H., et al., *Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer.* Nat Med, 2016. **22**(3): p. 298-305.
62. Castro, E., et al., *Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer.* J Clin Oncol, 2013. **31**(14): p. 1748-57.
63. Antonarakis, E.S., et al., *AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer.* N Engl J Med, 2014. **371**(11): p. 1028-38.
64. Scher, H.I., et al., *Association of AR-V7 on Circulating Tumor Cells as a Treatment-Specific Biomarker With Outcomes and Survival in Castration-Resistant Prostate Cancer.* JAMA Oncol, 2016. **2**(11): p. 1441-1449.
65. de Bono, J.S., et al., *Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer.* Clin Cancer Res, 2008. **14**(19): p. 6302-9.
66. Lorente, D., et al., *Decline in Circulating Tumor Cell Count and Treatment Outcome in Advanced Prostate Cancer.* Eur Urol, 2016.
67. Stockley, T.L., et al., *Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial.* Genome Med, 2016. **8**(1): p. 109.
68. *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run.*, B.I.T.G.D.A. Center, Editor. 2016: Broad Institute of MIT and Harvard.
69. Scheinin, I., et al., *DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly.* Genome Res, 2014. **24**(12): p. 2022-32.
70. Cani, A.K., et al., *Next-Gen Sequencing Exposes Frequent MED12 Mutations and Actionable Therapeutic Targets in Phyllodes Tumors.* Mol Cancer Res, 2015. **13**(4): p. 613-9.
71. McDaniel, A.S., et al., *Genomic Profiling of Penile Squamous Cell Carcinoma Reveals New Opportunities for Targeted Therapy.* Cancer Res, 2015. **75**(24): p. 5219-27.
72. McDaniel, A.S., et al., *Next-Generation Sequencing of Tubal Intraepithelial Carcinomas.* JAMA Oncol, 2015. **1**(8): p. 1128-32.

CHAPTER IV: Targeted DNA and RNA Sequencing of Paired Urothelial and Squamous Bladder Cancers

INTRODUCTION

Expression-based molecular subtypes have been widely reported in both muscle-invasive (MIBC) and non-muscle-invasive (NMIBC) bladder cancer[1-5]. Similar to those identified in breast cancer[6], these subtypes have typically been established through gene expression microarray or whole transcriptome next-generation sequencing (NGS) profiling of bulk tissue specimens, and are considered intrinsic to the tumor, demonstrating prognostic and predictive clinical utility largely based on basal vs. luminal differentiation[2, 7-9]. Importantly, gene expression based assessment of intrinsic molecular subtypes compatible with routine formalin fixed paraffin embedded (FFPE) tissue specimens has been reported and offered clinically with advocates for clinical introduction to guide neoadjuvant therapy decision making.

Genomic profiling of fresh frozen tissue specimens has highlighted substantial genomic heterogeneity in bladder cancer[4, 10, 11], with this molecular diversity shaped over time by both disease- and treatment-induced phenomena[12]. Histologically, bladder cancer also shows remarkable diversity with conventional urothelial and divergent components (including 15-25% with squamous differentiation) often co-existing[13, 14]. Technical challenges associated with isolating and profiling individual components of histologically heterogeneous frozen tissue specimens have limited comprehensive concurrent DNA and RNA based profiling of NMIBC and MIBC to fresh frozen bulk tissue specimens, with limited previous reports summarizing

targeted DNA NGS profiling from microdissected tissue specimens in aggressive, high-risk, or invasive urothelial carcinoma[15-18]. To our knowledge, no previous reports have systematically integrated comprehensive DNA and RNA molecular analysis of individual conventional urothelial and divergent components of the same tumor, nor evaluated robustness of expression-based molecular subtypes in such paired conventional and squamous tumor components.

Here, we report multiplexed DNA (mxDNAseq) and RNA sequencing (mxRNAseq)-based analysis of FFPE bladder cancer tissue specimens with diverse histologies, including cases with paired urothelial/squamous components, as well as bladder cancer cell lines. Through in silico assessment and application of our mxRNAseq panel, we validated robustness of observed profiles and the ability to determine basal and luminal gene expression-based subtypes in FFPE tissue specimens. Extending our previous work in detecting both mutations and copy number alterations (CNAs) from mxDNAseq data[19, 20], we report a novel strategy for detecting sub-gene copy-number alterations, with several detectable multi-exon sub-gene deep deletions confirmed using whole transcriptome RNA sequencing, and application to a retrospective compendium of over 1,100 pan-cancer tumor specimens identifying numerous clinically-relevant sub-gene copy-number deletions. Importantly, comprehensive DNA and RNA analysis of paired urothelial and squamous components enabled correlation of gene expression with genetic alterations of interest, and in multiple cases highlighted divergent expression profiles for paired components of the same tumor.

METHODS

Samples

For each tissue specimen, 4-10 x 10um FFPE sections were cut from a single representative block per case, using macrodissection with a scalpel as needed to enrich for tumor content. DNA was isolated using the Qiagen Allprep FFPE DNA/RNA kit (Qiagen, Valencia, CA), according to the manufacturer's instructions except for adding a 2 minute room temperature incubation and extending centrifugation time to 5 minutes during the xylene deparaffinization (step 1) and ethanol washing of xylene (step 2). DNA was quantified using the Qubit 2.0 fluorometer (Life Technologies, Foster City, CA).

Targeted next generation sequencing (NGS) - DNA

Genomic DNA was used for library generation using the Ion Ampliseq library kit 2.0 (Life Technologies, Foster City, CA) according to manufacturer's instructions with barcode incorporation. Templates were prepared using the Ion PGM Template OT2 200 Kit or Ion PI Template OT2 200 kit (Life Technologies, Foster City, CA) on the Ion One Touch 2 according to the manufacturer's instructions. Sequencing of multiplexed templates was performed using the Ion Torrent Personal Genome Machine (PGM) on Ion 318 chips with the Ion PGM Sequencing 200 Kit v2 or on the Ion Torrent Proton machine using Ion PI chips using the Ion PI Sequencing 200 Kit v2 (Life Technologies, Foster City, CA) according to the manufacturer's instructions.

Targeted next generation sequencing (NGS) - RNA

Multiplexed PCR-based RNA sequencing was performed on all samples as indicated in Table S1. For each sample, 20 ng RNA was reverse transcribed, bar-coded, and subjected to

multiplexed PCR to generate libraries using a custom Ampliseq panel and the Ion Ampliseq RNA Library kit. The custom Ampliseq panel contained 8 housekeeping genes and 103 target genes assessing major transcriptional programs in bladder cancer identified from publically available data and the Oncomine database [2, 4, 9]. As described for DNA, template generation and sequencing were performed on the Ion Torrent PGM or Proton according to the manufacturer's instructions. Data analysis was performed using Torrent Suite (5.0.2) and the Coverage Analysis Plugin (v5.0.2.0). For each amplicon, read counts were log2 transformed (read count +1). Then, to determine normalized expression for each target gene, the log2 count was normalized to the median of the log2 counts of the 8 housekeeping genes. Samples with >10,000 mapped end-to-end reads were retained for analyses. Consensus clustering (with number of clusters 'k' evaluated from k=2-6) and unsupervised hierarchical clustering using median-centered gene expression values were performed with R (v3.2.3) using the ConsensusClusterPlus (v1.24.0) and NMF (v0.26.0; aheatmap function) packages. Basal signature scores were calculated for each sample as the average of log2 normalized targeted RNAseq expression values for select basal markers minus the average across select luminal markers.

Whole Transcriptome RNASeq

All cell lines were profiled by whole transcriptome sequencing on Illumina HiSeq 2500 sequencers per manufacturers' instructions. Gene expression analysis was performed on aligned RNAseq reads for each sample using the Cufflinks pipeline[21]. Briefly, cDNA fragment size distribution means and standard deviations were estimated from unspliced alignment to the

genome. Reads were prepared with Tophat (v2.0.4)'s ‘prep_reads’ software tool, using the previously estimated fragment size distributions, prior knowledge of Ensembl gene annotations (hg19, GRCh37.69), and the following parameters: *--min-anchor 8; --splice-mismatches 0; --min-report-intron 50; --max-report-intron 500000; --min-isoform-fraction 0.15; --max-multihits 20; --max-seg-multihits 40; --segment-length 25; --segment-mismatches 2; --min-closure-exon 100; --min-closure-intron 50; --max-closure-intron 5000; --min-coverage-intron 50; --max-coverage-intron 20000; --min-segment-intron 50; --max-segment-intron 500000; --max-mismatches 2; --max-insertion-length 3; --max-deletion-length 3; --no-closure-search; --no-coverage-search; --no-microexon-search; --library-type fr-firststrand*. Alignment was then performed with Tophat (v2.0.4), using Bowtie2, and again using the previously estimated fragment size distributions and Ensembl gene annotations. Expression of each Ensembl gene was estimated using Cufflinks (v2.0.2), using the following parameters: *--library-type fr-firststrand; and --multi-read-correct*.

Variant calling

Sequencing data was analyzed using Torrent Suite 5.0.2 with alignment by TMAP using default parameters, and variants called via the Torrent Variant Caller plugin (version v5.0.2.1) using default low-stringency somatic variant settings. Called variants were filtered to remove low-quality or panel-specific errors as previously described [22], including flow-corrected sequencing depth of <40 reads and variants at variant fractions of <10% in tumor suppressors or <5% in oncogenic hotspots. Synonymous and non-coding variants passing these initial filters

were also removed, as were all variants present in the ExAC database at >0.1% (except those with pathogenic Clinvar annotation; **Appendix D**).

Copy number analysis

Normalized, GC-content corrected read counts per amplicon for each sample were divided by those from a pool of normal male genomic DNA samples (FFPE and frozen tissue, individual and pooled samples), yielding a copy number ratio for each amplicon. Gene-level copy number estimates were determined as described previously[19, 23, 24] by taking the coverage-weighted mean of the per-probe ratios, with expected error determined by the probe-to-probe variance. Genes with a $\log_2$ copy number ratio estimate of <-1 or >0.8 were considered to have high level loss and gain, respectively.

Sub-gene copy number detection

Amplicon-level copy-number estimates for all tumor suppressor genes with >10 targets on their respective targeted panels were considered for sub-gene copy-number detection in samples with >85% sequencing uniformity[19]. Circular binary segmentation was carried out on outlier-smoothed log2 amplicon-level copy-number ratio estimates using the DNAcopy (v1.44.0)[25] package in R (v3.2.3), and a subsequent sliding window function was used to identify maximum differences ('max-diff') in median amplicon-level log2 copy-number estimates for consecutive segments across analyzed genes in samples with segment standard deviations of <0.75. While variable max-diff thresholds were considered, a conservative

threshold of max-diff values ≥ 0.8 was applied to prioritize candidate sub-gene copy-number alterations.

TCGA data

Absolute gene-level RNASeq V2 RSEM expression and z-score quantitative values for whole transcriptome RNA sequencing data from 405 bladder cancer samples profiled in the TCGA project ('TCGA provisional') were downloaded from cBioPortal[26, 27]. Sample molecular subtype assignment for 234 samples was obtained from Aine et al [28] and the 126 samples with TCGA cluster membership assignment were confirmed via TCGA cluster assignments[4]. Consensus clustering was evaluated for numbers of clusters ('k') for k=2-6 using the ConsensusClusterPlus (v1.24.0) package in R (v3.2.3).

RESULTS

Validation of targeted RNAseq panel

To establish the validity of our custom RNAseq panel comprised of 103 targets (normalized to 8 housekeeping genes) capturing major biologically-relevant transcriptional programs in bladder cancer, we first assessed concordance between targeted RNAseq gene expression and whole transcriptome RNAseq from high quality frozen RNA from 21 independent bladder cancer cell lines (**Appendix D**). We previously demonstrated the reproducibility of this approach across both biological and technical replicates[29], and here we confirm our custom targeted RNAseq assay yields normalized target gene expression values

highly correlated with those from conventional whole transcriptome RNAseq (**Figure 4.1A**; median Pearson correlation coefficient = 0.90). Further, unsupervised hierarchical clustering of our urothelial cancer cell lines using a subset of well-characterized basal and luminal markers[1] was essentially identical using both targeted and conventional RNAseq data (**Figure 4.1B**). Together these results highlight the validity of our targeted RNAseq approach, supporting the ability of this assay to faithfully assess individual components of expression-based molecular subtypes with high fidelity.

Targeted RNAseq

To demonstrate the potential utility of our targeted RNAseq panel, we first analyzed in silico gene expression data from 234 bladder cancer samples profiled by whole transcriptome RNAseq through The Cancer Genome Atlas (TCGA) project[4, 26, 27] focusing on the 103 cancer-associated targets on our targeted RNAseq panel. Using TCGA RNAseq data from our targeted genes, which includes subsets of genes from multiple previously reported molecular subtype classifiers[2, 4, 9, 28], we demonstrate the ability to recapitulate known molecular subtypes including basal/luminal expression modules, with high fidelity by unbiased consensus clustering (**Figures D1 and D2**). These results support the ability of the genes in our panel to assess essential transcriptional programs informing expression-based molecular subtyping.

To explore potential clinical utility and applicability of our approach on routine clinical FFPE bladder cancer tissue specimens, we integrated targeted RNAseq on our panel from a total of 110 samples, the above described 21 bladder cancer lines, 16 previously profiled FFPE tissue specimens[29], and 73 unreported FFPE tissue specimens representing a spectrum of bladder

cancer histology, including both pure squamous cell carcinoma and paired urothelial and squamous cell carcinoma components from the same resection (**Appendix D**). Targeted RNAseq yielded an average 1,664,3369 mapped reads per sample, and enabled robust assessment of sets of inter-correlated transcripts from key biologically relevant transcriptional modules (**Figure D3**). Consensus clustering of normalized expression values across all 103 targets for tissue samples passing rigorous QC metrics (*n*=77; see **Methods**) identified 4 sample clusters across our tissue cohort (**Figure 4.2A**) with luminal+/basal- (Cluster 1), basal+/luminal- (Cluster 2), basal-/luminal- (Cluster 3), or basal+/luminal+ (Cluster 4) expression profiles. Clusters 1 and 2 contain the majority of samples and segregate mostly along histological lines, with conventional urothelial lesions (enriched in Cluster 1) generally showing high luminal marker expression (consistent with low basal signature scores; see **Methods**), while squamous lesions (enriched in Cluster 2) show elevated basal marker expression (as evidenced by elevated basal signature scores) (**Figure 4.2A, Appendix D**). Cluster 3 contains all samples with non-squamous divergent differentiation, while Cluster 4 contains a mix of urothelial and squamous lesions (**Figure 4.2A, Table S1**). Examining expression patterns in key basal/luminal markers alone reinforces these observations (**Figure 4.2B**), and unsupervised hierarchical clustering is consistent with consensus clustering results (**Figure D4**). Consensus clustering incorporating profiled cell lines identified 5 clusters that map as expected to Clusters 1-4 derived without cell lines, with an additional cell-line-only cluster (UM CL2) comprised entirely of cell line samples with no epithelial-to-mesenchymal (EMT) marker expression (**Figure D5**). Together these results suggest our approach is capable of profiling individual components of histologically heterogeneous bladder cancer tissue specimens, enabling robust recapitulation of previously reported expression-based basal/luminal subtypes.

Targeted DNAseq

To characterize the underlying genomic context of our targeted RNAseq results, high-quality targeted DNA sequencing data was assessed for 106 of 110 (96%) DNA samples (**Appendix D**). For all high-quality cell line and tissue samples (*n*=90; see **Methods**) profiled by targeted DNA sequencing and not previously reported, prioritized somatic DNA alterations are summarized in **Appendix D** and **Figure 4.3**, and copy-number heatmaps are displayed in **Figure D6**.

Frequent *TP53* (56%, 39 of 69 samples) and activating hotspot *PIK3CA* (30%, 21 of 69) somatic mutations were observed in our tissue cohort (**Figure 4.3A**), with a larger (though non-significant) proportion of squamous lesions carrying detectable *PIK3CA* point mutations than urothelial components (42% vs. 19%, Fisher's exact test; p=0.06, 95% CI = (0.9,12.9)). Activating point mutations in *FGFR3* were seen in 12/69 (17%) samples, with no significant difference in *FGFR3* mutational frequency between UCC and squamous lesions (15% vs. 14%, Fisher's exact test p=1.0). Copy-number gains of *MYC* (26%) and *CCND1* (13%) were common overall in our tissue cohort, as were focal deletions of *CDKN2A* (31%). A significantly larger proportion of squamous samples carried copy-number gains in *MYC* than those with urothelial histological differentiation (Fisher's exact p=0.002, 95% CI (1.9, 96.5)), while no significant differences were observed for frequency of *CCND1* gains (Fisher's exact test, p=0.70) or *CDKN2A* deletions (Fisher's exact test, p=0.42). Focal, high-level amplifications (log2(CopyNumberRatio) > 1.6) of *EGFR* (6%) and *ERBB2* (6%) were also observed in this cohort. These results suggest substantial potential utility for characterization of underlying

driving somatic genomic alterations by targeted DNA sequencing to refine interpretation of expression-based profiles in the context of histologically heterogeneous bladder cancers.

Targeted DNA sequencing of 21 bladder cancer cell lines identified a series of recurrent point mutations, indels, and copy-number alterations across cell lines. Of 21 cell lines, 15 (71%) carried somatic point mutations or indels in *TP53*, with 10/21 (48%) cell lines carrying at least one loss-of-function mutation in other tumor suppressors (**Figure 4.3B**). Activating hotspot mutations in *PIK3CA* (*n*=5), *FGFR3* (*n*=4), *ERBB2/3* (*n*=4), *NFE2L2* (n=3), *RAS* family members (*n*=2), and *AKT1* (n=1) were seen across cell lines (**Figure 4.3B**). Focal copy-number deletions (log2(CopyNumberRatio) < -1.0) of *CDKN2A* were observed in 11 cell lines (52%), while 2 cell lines (UMUC-1 and UMUC-9) showed high-level focal amplifications (log2(CopyNumberRatio) > 1.6) of *CCND1*. Additional high-level focal amplifications of *PPARG* (in UMUC-9; CopyNumberRatio = 44.7) and *EGFR* (in UMUC-5; CopyNumberRatio=16.6) were identified. Together, these results suggest the profiled cell lines contain a majority of the recurrent genomic alterations seen in bladder cancer *in vivo*, and include high-level copy-number amplifications with important therapeutic implications for bladder cancer patient populations[7, 30].

Sub-gene copy-number detection

While targeted DNA sequencing has shown promise in assessing diverse genomic alterations for precision oncology initiatives, few reports have explored whether targeted, amplicon-based DNA sequencing is capable of detecting sub-gene copy-number alterations, a unique but important class of alterations given their expected impact on gene function,

particularly for tumor suppressor genes. Thus, to evaluate whether a systematic approach could facilitate sub-gene copy-number deletion detection from our targeted DNA sequencing data, we developed a novel heuristic strategy for scanning sub-gene (e.g., exon- or multi-exon-level) copy-number deletions in the 24 tumor suppressor genes on our panel with >10 target amplicons (overall range of target amplicons: 6-138); see **Methods**). Initial testing in our highly altered bladder cancer cell line cohort identified putative sub-gene copy-number deletions in J82 and UMUC-14 cell lines in *PTEN* and *RB1* genes, respectively (**Figure 4.4, Figure D7**). For J82, genome-ordered, amplicon-level copy-number ratio profiles and segmented copy-number calls in *PTEN* highlight the structure of these sub-gene copy-number aberrations (**Figure 4.4B**), with similar results seen in *RB1* for UMUC-14 (**Figure D7B**). Importantly, orthogonal whole transcriptome RNAseq of both cell lines identified expression of only the first 6 exons of *PTEN* in J82 (**Figure 4.4C-D**) and first 5 exons of *RB1* in UMUC-14 (**Figure D7C-D**), consistent with the observed genomic sub-gene copy-number deletions.

We subsequently applied our approach to a retrospective cohort of high-quality targeted DNA NGS data from 1,105 internally sequenced FFPE tissue samples, and identified 48 additional samples (4.3%) with candidate sub-gene copy-number alterations in tumor suppressor genes with low gene-level standard deviation (Probe Error < 0.2)[19], including *APC, ATM, MSH2, RB1, PTEN*, and *TP53*. This included three separate prostate cancer tissue samples with putative sub-gene deletion in *MSH2* (two of which are paired primary prostate adenocarcinoma (PRAD) and bladder metastasis/recurrence samples from a single case (PR-115 and PR-161)), with the paired bladder metastasis/recurrence (PR-161) clearly hypermutated by mutation profiling (**Figure D8**). Further, in the third unpaired prostate cancer sample (PR-34), NGS pileup data from multiple sequencing modalities (both amplicon and hybrid capture) consistently

supported the presence of the detected sub-gene *MSH2* copy number deletion affecting exons 12-16 (**Figure D8**). High-quality sub-gene deletion events in additional prostate cancer tissue samples included those in *PTEN* (**Figure D8**), *RB1* (**Figure D8**), and *TP53*. Sub-gene deletions were observed across additional cancer types, including an *RB1* sub-gene deletion in MO-32 (a fine needle aspirate sample from a lung adenocarcinoma lymph node metastasis) affecting exons 19 through 27, and sub-gene deletions in both *PTEN* and *RB1* in MO-72 (a metastatic uterine leiomyosarcoma) (**Figure D9**). Together these results highlight important potential clinical opportunities for leveraging a heuristic sub-gene copy-number deletion detection approach to scan for potentially therapeutically relevant set of molecular alterations in targeted DNA sequencing data.

Paired urothelial and squamous cases

Considerable work has gone into characterizing putative intrinsic expression-based molecular subtypes in bladder cancer from bulk tissue specimens, but limited work has evaluated these molecular subtypes in the context of histologically diverse components of the same tumor, which is encountered frequently in transurethral resection and cystectomy specimens[2, 4, 8, 9]. To determine whether expression-based subtypes for paired urothelial and squamous components of the same tumor appeared “intrinsic” to the tumor, and thus stable between the divergent histologic components, we systematically evaluated targeted DNA and RNA sequencing data for paired urothelial and squamous components from 11 separate bladder cancer cases in our cohort.

Overall, most pairs show a high degree of genomic similarity, with paired samples showing nearly identical prioritized DNA point mutations, indels, and copy-number alterations

(**Figure 4.2, Figure D6**). However, a number of pairs with identical prioritized genomic alterations show discordant gene expression profiling results. In pair 9, DNA sequencing of paired squamous (BL-360) and urothelial (BL-361) components highlighted identical *TP53* and *NFE2L2* nonsynonymous SNVs, along with focal *ERBB2* amplification, yet targeted RNA sequencing results shows differential classification by consensus clustering (**Figure 4.2**) and highly divergent basal/luminal expression profiles, with BL-360 demonstrating markedly higher basal marker expression than BL-361 (see **Figures 4.5A and B**). Notably, samples such as BL-360/BL-361 in our cohort with focal *ERBB2*, *EGFR*, or *PPARG* amplifications show clear outlier expression of the amplified gene product (**Figure 4.5C**), a phenomenon that may complicate proposed use of individual targets as expression subtype proxies for guiding clinical decision-making, such as EGFR expression to identify basal subtype urothelial carcinoma[7]. Analysis of conventional whole transcriptome RNAseq and copy-number data from 405 bladder cancer samples profiled in TCGA reinforces these observations for *ERBB2*, *EGFR*, and *PPARG*, with highest expression in samples with corresponding gene amplifications, independent of basal/luminal expression-based subtype signature (**Figure 4.5D**).

We also observed basal/luminal expression discordance for Pair 1, where targeted RNA sequencing of paired squamous (BL-340) and urothelial (BL-341) components of the same tumor showed substantially elevated basal expression in the squamous lesion, despite identical somatic hotspot mutations in *CTNNB1* (p.S37F) and *PIK3CA* (p.E542K) identified by DNA sequencing (**Figure D10**). Together, our results demonstrate that although histologically divergent components typically have concordant driving genomic alterations (supporting clonality), they may have markedly different expression of canonical basal/luminal genes. Hence, expression-

based subtyping from bulk tissue specimens may be confounded by divergent sub-component expression profiles and challenge the “intrinsic” nature of these subtypes.

DISCUSSION

Histological divergent differentiation is extremely common in bladder cancer, and few studies have explored the extent to which this histological differentiation may confound clinical application of recently reported expression-based subtypes. Here we describe validation of a multiplexed targeted RNA-seq assay capable of assessing major biologically relevant transcriptional programs in bladder cancer from FFPE tissue samples, and demonstrate the ability to identify expression profiles from individual components of bladder cancers with varying histology. We show that while some paired urothelial and squamous components demonstrate concordant expression profiles across key transcriptional modules in the context of identical genomic alterations, several sets of paired samples show markedly different expression of key basal/luminal markers despite shared genomic alterations, potentially complicating proposed clinical utility of expression-based molecular subtyping. Our results support the need for clinical validation of expression based subtyping assays, including outcome and prediction of benefit from chemotherapy, in the context of histologic and transcriptomic heterogeneity, particularly when both squamous and urothelial components are present. For example, if both conventional urothelial and squamous cell carcinoma are present in a transurethral resection specimen, it is unclear from currently available knowledge if both components should be sampled for expression based subtyping given that they may give divergent profiles. Stu

Likewise, although strategies for predicting neoadjuvant chemotherapy response may be feasible by expression based basal/luminal subtyping alone, combined DNA and RNA profiling of individual tumor components may provide a more reliable portrait of the driving (and potentially actionable) molecular alterations. For example, although *ERBB2* (luminal) and *EGFR* (basal) expression have been included in several basal vs. luminal subtyping schemas and proposed therapeutic strategies (e.g. EGFR based therapy in all basal subtype cancers[7]), results from our targeted RNA/DNA sequencing and the TCGA demonstrate a clear difference between low level expression of these markers in basal vs. luminal subtypes, with marked over-expression, regardless of basal vs. luminal subtype, exclusively in cases with high level amplification of the respective gene.

We further describe a heuristic strategy for sub-gene copy-number deletion detection through segmentation of amplicon-level copy-number estimates that enabled detection of multi-exon deletions in tumor suppressors from targeted DNA sequencing data in bladder cancer cell lines which was validated by whole transcriptome RNA-seq analysis. Subsequent application of this heuristic approach identified a subset of samples from a large FFPE tissue compendia with candidate sub-gene deletions, several of which may be therapeutically relevant. Refinements of our sub-gene copy-number detection approach are warranted to expand potential applications, including corrections for tumor content and sequencing uniformity, improved parameterization of the amplicon-level copy-number estimate smoothing and circular binary segmentation process, and applications to updated iterations of the OCP and expanded panels. Overall, conservative thresholds were used for this implementation to reduce the likelihood of false positives, and a more sophisticated appreciation of error/noise modeling could enable detection

of exon-level deletion events in genes such as *BRCA1*, *BRCA2* with more variable amplicon-level performance.

Our current targeted RNA sequencing panel has notable limitations in terms of applicability beyond basal vs. luminal type expression subtyping. Given the long-standing clinical use of Bacillus Calmette-Guerin (BCG) intravesical immunotherapy in high-risk early stage bladder cancer[31], recent approval of multiple checkpoint inhibitors for treatment of locally advanced and/or metastatic bladder cancer[32-35], and oncogenic role of *FGFR3* gene fusions in bladder cancer[36], our panel will need to be refined to assess these important therapeutic targets to support predictive and prognostic biomarker development and monitoring. Likewise, our panel would need to undergo usual analytical and clinical validation before clinical introduction.

Understanding the impact of histological differentiation on expression-based subtyping of biopsy or bulk tissue specimens is essential when considering their proposed use in guiding clinical decision-making. Thus, comprehensive molecular profiling of expanded sample sets with divergent differentiation is warranted. As the prevalence of squamous differentiation (the most common type of divergent differentiation) with increasing stage and grade[14], understanding the impact of histologic heterogeneity on expression profiles and patient outcome in both the presence and absence of neoadjuvant therapy is urgently needed to guide appropriate use of expression based subtyping assays.

In summary, we describe the ability of a targeted RNA-seq platform to assess key biologically-relevant transcriptional programs in bladder cancer, including luminal vs. basal subtyping. Importantly, through pairing with targeted DNA sequencing, we demonstrate

important considerations for proposed clinical use of expression-based subtypes in a histologically heterogeneous disease, as our results show paired urothelial and squamous components of the same tumor may have identical driving genomic alterations but markedly different transcriptional profiles. We anticipate that continued work in profiling individual tumor components with divergent differentiation will help refine our understanding and interpretation of expression-based subtypes in bladder cancer, and may guide identification of disease biomarkers that more precisely stratify patients for optimal treatment prioritization.

Figure 4.1 – Validation of custom bladder cancer targeted RNAseq panel and comparison to conventional whole transcriptome sequencing in 21 profiled cell lines.

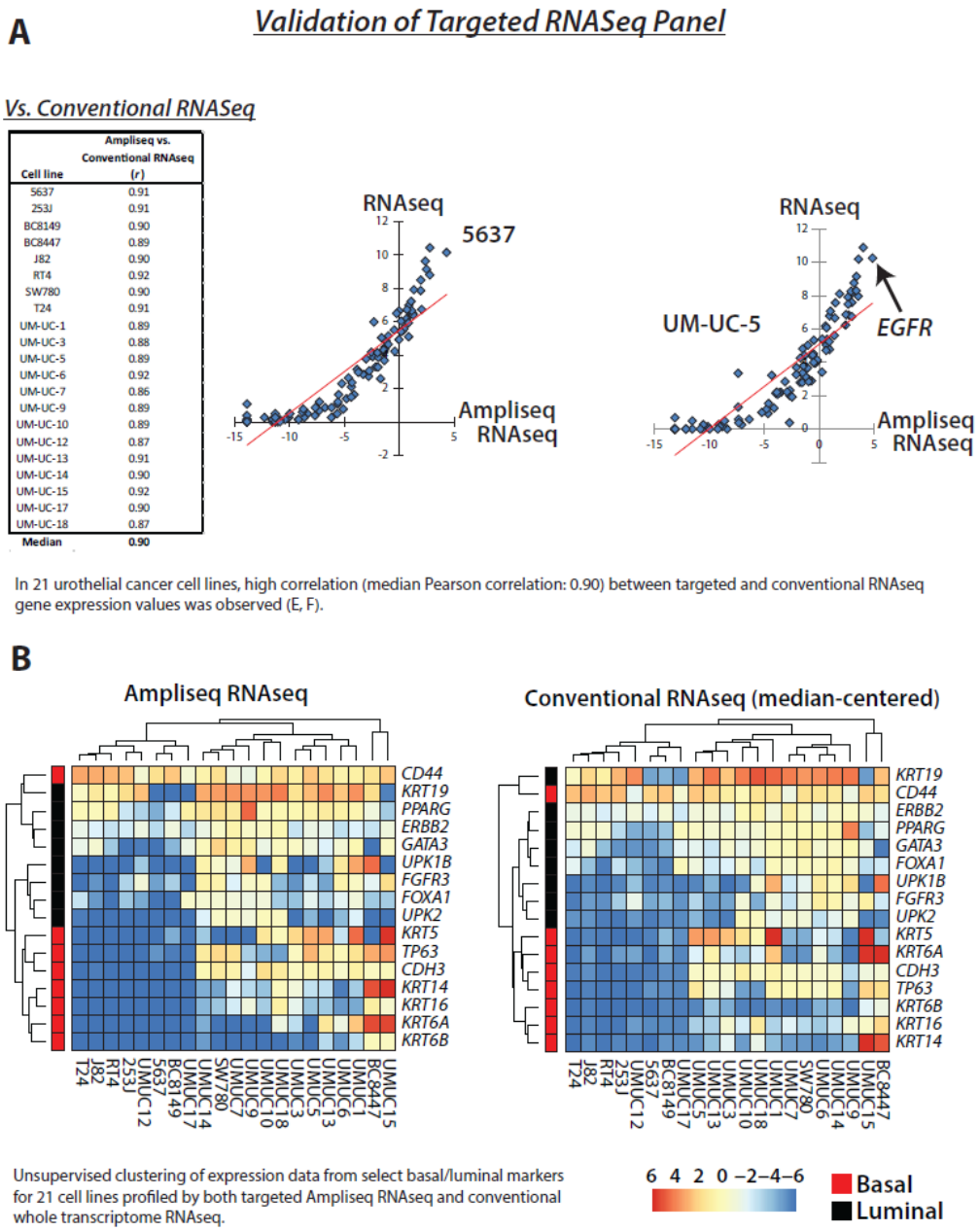


Figure 4.1. A. At left, a table summarizing Pearson correlation values across 21 cell lines between normalized log2-transformed gene-level targeted RNAseq expression values and gene-level expression values from conventional whole transcriptome RNAseq for the 103 non-housekeeping gene targets on our targeted RNAseq panel. At right, expression values from targeted and conventional RNAseq are plotted separately for two different bladder cancer cell lines (5637 and UM-UC-5), showing highlight correlated values across targets. **B.** Unsupervised hierarchical clustering of 16 consensus basal or luminal markers across 21 cell lines using targeted RNAseq (left) and conventional RNAseq (right) expression values yields similar clustering of both gene targets and samples, supporting the ability of our custom targeted RNAseq assay to robustly assess major basal/luminal expression programs previously profiled by whole transcriptome RNA sequencing.

Figure 4.2 – Unsupervised clustering of targeted RNAseq expression data for high-quality tissue specimens profiled on custom targeted RNAseq panel.

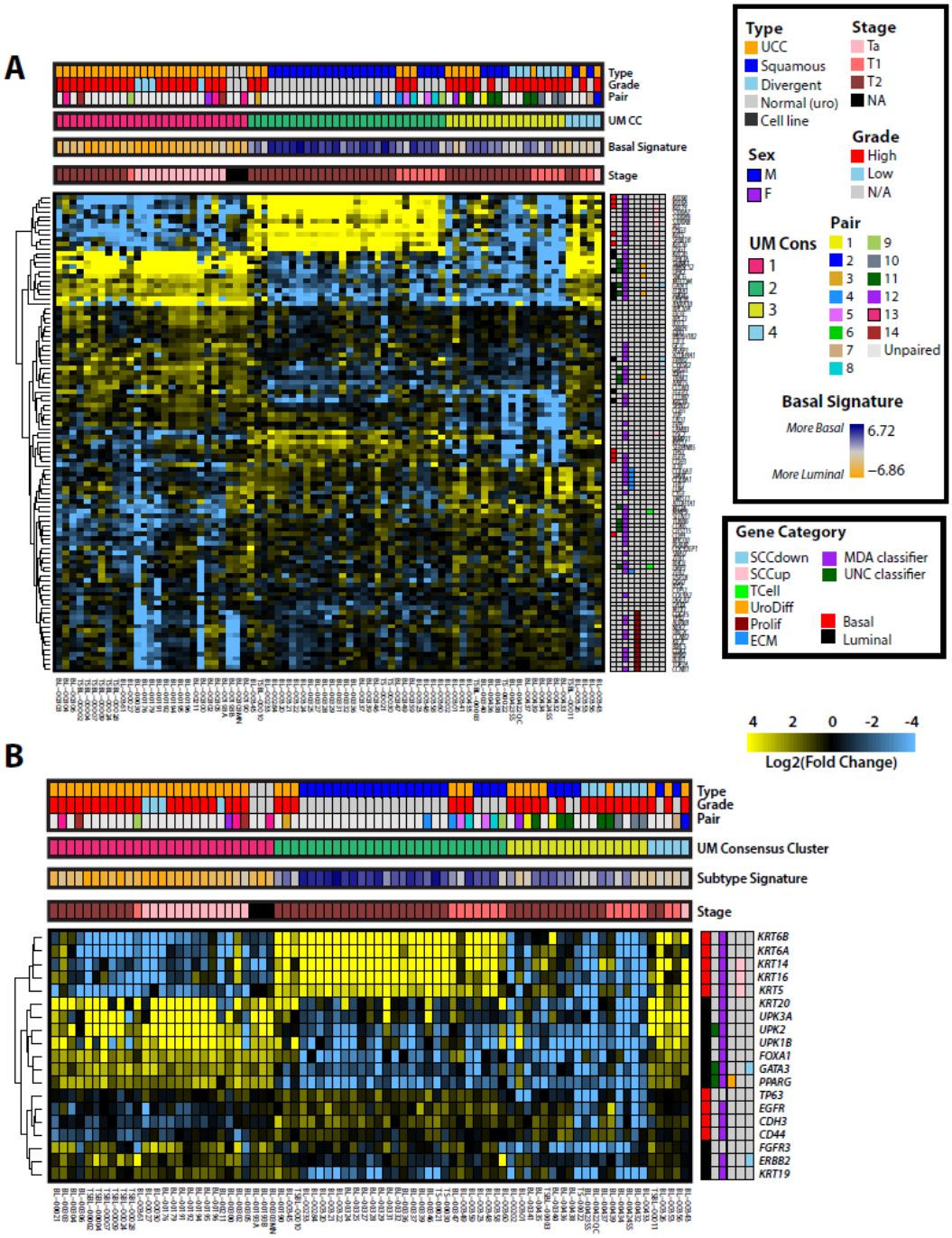


Figure 4.2 A. Unsupervised clustering of normalized log2 expression values from all non-housekeeping gene targets for 77 high-quality tissue specimens profiled on our custom targeted RNAseq panel. Samples are sorted left to right by consensus cluster, then stage, then histological subtype. Sample annotation (header annotation rows at top) is colored corresponding to annotations contained the figure legend, while target annotation (at right) is colored according to gene category annotations provided. **B.** Unsupervised clustering of normalized log2 expression values from select basal/luminal genes for 77 high-quality profiled tissue specimens enables delineation of individual gene target expression, and highlights substantially elevated expression of *ERBB2* and *EGFR* in samples with focal copy-number amplifications.

Figure 4.3 – Integrative table summarizing prioritized somatic point mutations, insertions, and deletions detected from targeted DNA sequencing of high-quality tissue specimens

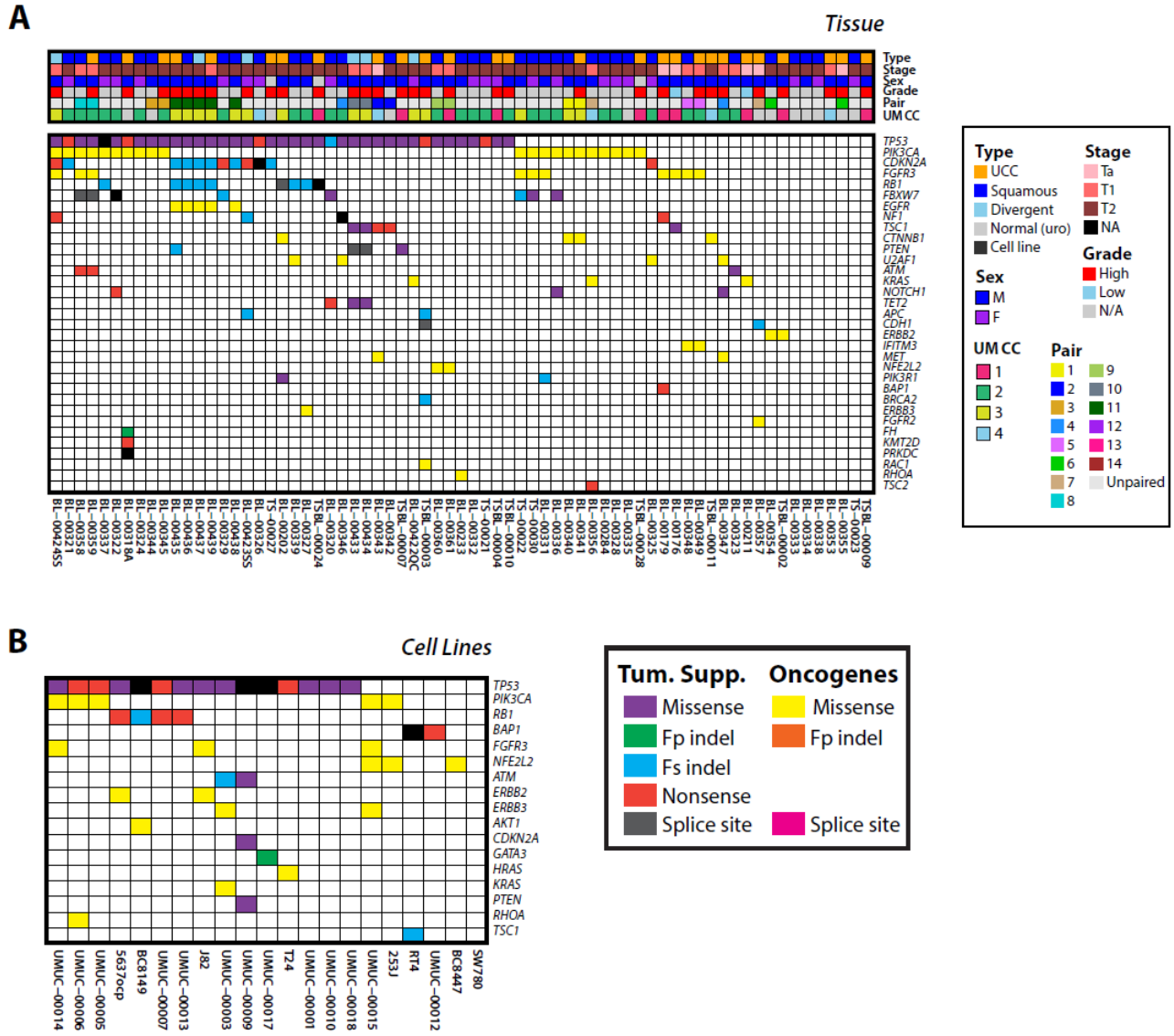


Figure 4.3. Integrative tables summarizing prioritized somatic point mutations, insertions, and deletions identified from targeted DNA sequencing of (**A**) high-quality tissue specimens and (**B**) urothelial carcinoma cell lines. Sample annotation is provided at top with colors corresponding to annotation legend, and samples are sorted from left to right by presence or absence of alteration in the corresponding genes. Genes are sorted from top to bottom by decreasing total number of alterations across the cohort, and cells are colored by alteration types provided in the legend.

Figure 4.4 – Validation of sub-gene copy-number deletion detection from targeted DNA sequencing by conventional whole-transcriptome RNA sequencing.

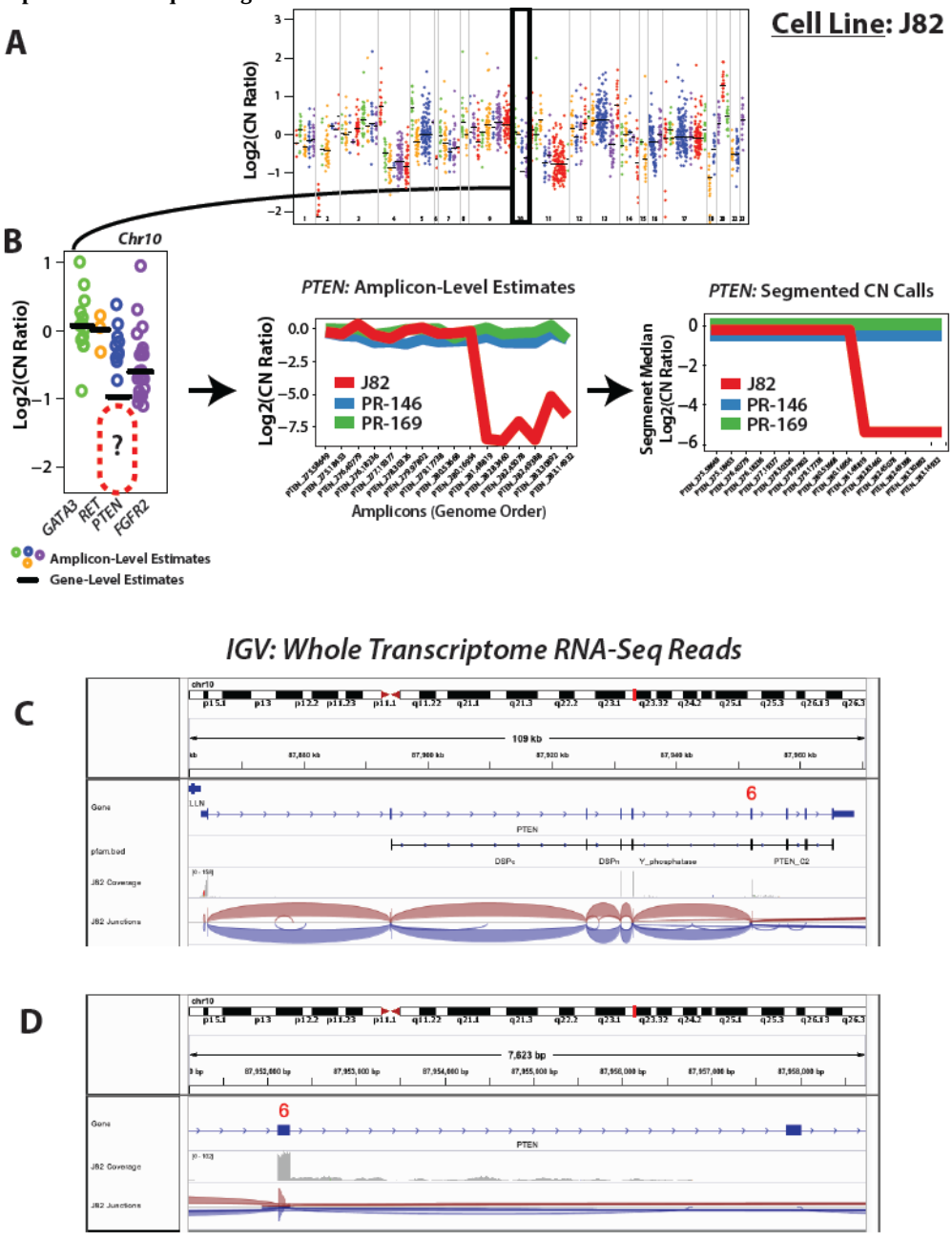


Figure 4.4 A. Genome-wide copy-number plot from targeted DNA sequencing of urothelial cancer cell line J82. Individual dots represent amplicon-level log2 copy-number ratio estimates, with horizontal black lines representing log2 gene-level copy-number ratio estimates. Black rectangle highlights portion of the plot (chr10) presented in panel B. **B.** At left, a zoomed view of amplicon- and gene-level copy number ratios on chr10 for J82 demonstrates the absence of amplicon-level copy-number ratios for a subset of *PTEN* target amplicons. The middle panel highlights amplicon-level copy-number ratios sorted in genome order, suggesting a sub-gene deletion affecting the last several exons of *PTEN*. At right, a sliding-window function applied to segmented copy number values from amplicon-level data, provides a smoothed, segmented sub-gene copy-number call for clinical or research reporting. **C.** Integrated Genome Viewer (IGV) screenshot of spliced read alignment data across *PTEN* coding regions for conventional whole-transcriptome RNAseq data from J82 shows depleted expression of last several exons. **D.** Zoomed view of conventional RNAseq data for exon 6 and 7 shows limited read mapping and depleted expression of exon 7 consistent with the observed sub-gene copy-number deletion affecting the 3' region of *PTEN*.

Figure 4.5 – Divergent expression profiles of histologically diverse components of the same tumor with shared genetic alterations, including focal ERBB2 amplification.

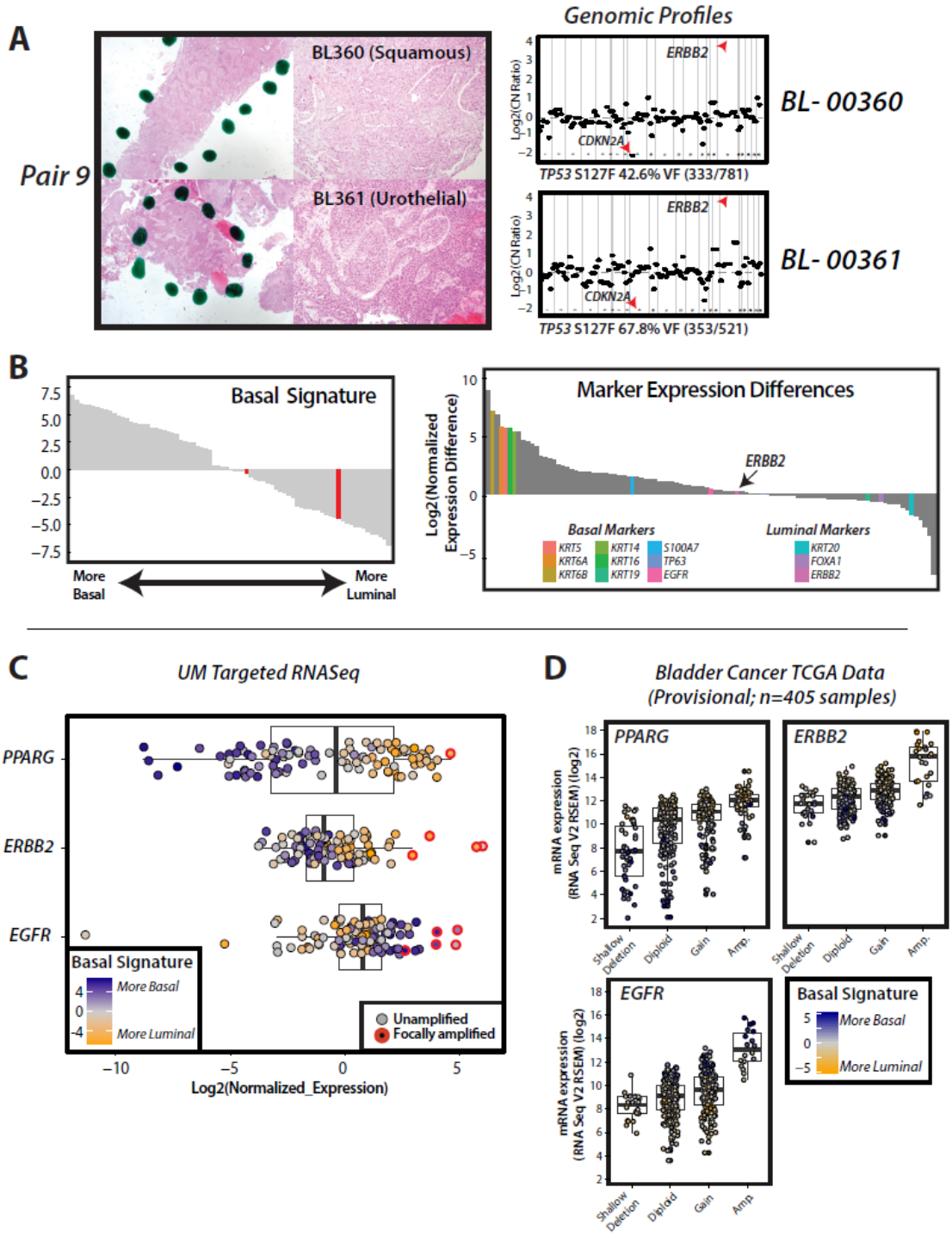


Figure 4.5: **A.** Haematoxylin and eosin staining images of individual squamous and urothelial components profiled for pair 9 are shown. At right, similar genome-wide copy-number profiles derived from targeted DNA sequencing are shown for each sample, and focal 2-copy deletion of *CDKN2A*, focal amplification of *ERBB2*, and a *TP53* S127F somatic point mutation seen in both samples are indicated. **B**. At left, divergent basal signature values for BL-360 and BL-361 are highlighted in red in the context of all basal signatures for profiled tissue specimens in our study. At right, individual expression differences between BL-360 and BL-361 are plotted for 103 non-housekeeping markers, with select basal or luminal markers colored according to the legend.

Similarity in *ERBB2* expression values between the two samples as indicated is consistent with concordant high level focal amplifications identified in both lesions. **C.** Box plots of targeted RNAseq expression values for 3 individual genes shows elevated expression is enriched for samples with focal copy-number amplifications identified by targeted DNA sequencing of the same sample. Points are colored by basal signature score as indicated in legend. **D.** For the 3 genes displayed in panel C, TCGA copy-number and expression data was analyzed in 405 bladder cancer samples, showing similar outlier expression levels for most samples with focal copy-number amplifications.

Chapter IV References

1. Choi, W., et al., *Intrinsic basal and luminal subtypes of muscle-invasive bladder cancer.* Nat Rev Urol, 2014. **11**(7): p. 400-10.
2. Choi, W., et al., *Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy.* Cancer Cell, 2014. **25**(2): p. 152-65.
3. Sjodahl, G., et al., *A molecular taxonomy for urothelial carcinoma.* Clin Cancer Res, 2012. **18**(12): p. 3377-86.
4. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of urothelial bladder carcinoma.* Nature, 2014. **507**(7492): p. 315-22.
5. Hedegaard, J., et al., *Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma.* Cancer Cell, 2016. **30**(1): p. 27-42.
6. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
7. Rebouissou, S., et al., *EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype.* Sci Transl Med, 2014. **6**(244): p. 244ra91.
8. Seiler, R., et al., *Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy.* Eur Urol, 2017.
9. Damrauer, J.S., et al., *Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology.* Proc Natl Acad Sci U S A, 2014. **111**(8): p. 3110-5.
10. Morrison, C.D., et al., *Whole-genome sequencing identifies genomic heterogeneity at a nucleotide and chromosomal level in bladder cancer.* Proc Natl Acad Sci U S A, 2014. **111**(6): p. E672-81.
11. Hurst, C.D., et al., *Novel tumor subgroups of urothelial carcinoma of the bladder defined by integrated genomic analysis.* Clin Cancer Res, 2012. **18**(21): p. 5865-77.
12. Faltas, B.M., et al., *Clonal evolution of chemotherapy-resistant urothelial carcinoma.* Nat Genet, 2016. **48**(12): p. 1490-1499.
13. Amin, M.B., *Histological variants of urothelial carcinoma: diagnostic, therapeutic and prognostic implications.* Mod Pathol, 2009. **22 Suppl 2**: p. S96-S118.
14. Lopez-Beltran, A., et al., *Squamous differentiation in primary urothelial carcinoma of the urinary tract as seen by MAC387 immunohistochemistry.* J Clin Pathol, 2007. **60**(3): p. 332-5.
15. Longo, T., et al., *Targeted Exome Sequencing of the Cancer Genome in Patients with Very High-risk Bladder Cancer.* Eur Urol, 2016. **70**(5): p. 714-717.
16. Ross, J.S., et al., *Advanced urothelial carcinoma: next-generation sequencing reveals diverse genomic alterations and targets of therapy.* Mod Pathol, 2014. **27**(2): p. 271-80.
17. Millis, S.Z., et al., *Molecular profiling of infiltrating urothelial carcinoma of bladder and nonbladder origin.* Clin Genitourin Cancer, 2015. **13**(1): p. e37-49.
18. Bellmunt, J., et al., *HER2 as a target in invasive urothelial carcinoma.* Cancer Med, 2015. **4**(6): p. 844-52.
19. Grasso, C., et al., *Assessing copy number alterations in targeted, amplicon-based next-generation sequencing data.* J Mol Diagn, 2015. **17**(1): p. 53-63.

20. Hovelson, D.H., et al., *Development and validation of a scalable next-generation sequencing system for assessing relevant somatic variants in solid tumors.* Neoplasia, 2015. **17**(4): p. 385-99.
21. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
22. Hovelson, D.H., et al., *Development and Validation of a Scalable Next-Generation Sequencing System for Assessing Relevant Somatic Variants in Solid Tumors.* Neoplasia, 2015. **17**(4): p. 385-399.
23. Cani, A.K., et al., *Next-Gen Sequencing Exposes Frequent MED12 Mutations and Actionable Therapeutic Targets in Phyllodes Tumors.* Mol Cancer Res, 2015. **13**(4): p. 613-9.
24. Warrick, J.I., et al., *Tumor evolution and progression in multifocal and paired non-invasive/invasive urothelial carcinoma* Virchows Arch, 2014.
25. Venkatraman, E.S. and A.B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array CGH data.* Bioinformatics, 2007. **23**(6): p. 657-63.
26. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.
27. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
28. Aine, M., et al., *Biological determinants of bladder cancer gene expression subtypes.* Sci Rep, 2015. **5**: p. 10957.
29. Warrick, J.I., et al., *Tumor evolution and progression in multifocal and paired non-invasive/invasive urothelial carcinoma.* Virchows Arch, 2015. **466**(3): p. 297-311.
30. Korpal, M., et al., *Evasion of immunosurveillance by genomic alterations of PPARgamma/RXRalpha in bladder cancer.* Nat Commun, 2017. **8**(1): p. 103.
31. Redelman-Sidi, G., M.S. Glickman, and B.H. Bochner, *The mechanism of action of BCG therapy for bladder cancer--a current perspective.* Nat Rev Urol, 2014. **11**(3): p. 153-62.
32. Powles, T., et al., *MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer.* Nature, 2014. **515**(7528): p. 558-62.
33. Massard, C., et al., *Safety and Efficacy of Durvalumab (MEDI4736), an Anti-Programmed Cell Death Ligand-1 Immune Checkpoint Inhibitor, in Patients With Advanced Urothelial Bladder Cancer.* J Clin Oncol, 2016. **34**(26): p. 3119-25.
34. Rosenberg, J.E., et al., *Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial.* Lancet, 2016. **387**(10031): p. 1909-20.
35. Bellmunt, J., et al., *Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma.* N Engl J Med, 2017. **376**(11): p. 1015-1026.
36. Williams, S.V., C.D. Hurst, and M.A. Knowles, *Oncogenic FGFR3 gene fusions in bladder cancer.* Hum Mol Genet, 2013. **22**(4): p. 795-803.

CHAPTER V: Comprehensive Molecular Profiling of Multifocal Prostate Cancer

INTRODUCTION

Considerable academic and commercial efforts across cancers endure to both identify and robustly assess prognostic and predictive biomarkers from routine clinical biospecimens including urine, blood, and tissue [1-3]. In prostate cancer, progress to characterize disease biomarkers and (given the widely acknowledged limitations of PSA screening[4, 5]) enhance opportunities for treatment and disease prognosis and risk stratification have been made across multiple biospecimens in both localized and metastatic/advanced disease [6]. Notably, tissue biopsy-based gene expression assays capable of predicting risk of high-grade disease, as well as disease metastasis or recurrence risk, have emerged as tools with substantial clinical utility[7-9]. As an inherently multifocal disease, prostate cancer presents challenges for disease prognostication from single tissue biopsies, yet tissue-based expression assays are used frequently in clinical practice and some claim robustness to disease multifocality[10]. To date, technical challenges have limited systematic exploration of expression-based profiles using these assays to evaluate concordance of expression-based prognostic scores from individual foci in the context of true multifocal disease.

Accordingly, we have designed a custom multiplexed, PCR-based targeted RNA sequencing panel compatible with minute formalin fixed paraffin embedded tissue (FFPE) tissue samples comprised of 306 targets to assess major transcriptional modules and disease biomarkers

relevant for both localized and metastatic prostate cancer. This panel includes transcripts enabling derivation of commercial expression-based Oncotype DX™, Prolaris™ and Decipher™ scores, key prostate cancer specific long noncoding RNA (lncRNA) and fusion (e.g., *TMPRSS2-ERG*) isoforms, expressed somatic mutations (e.g., *BRAF, SPOP, IDH1*), expressed hereditary risk variants (*HOXB13*), and potentially predictive/prognostic biomarkers (*AR-V7, SCHLAP1*). In addition, this panel enables robust assessment of *AR*-driven transcriptional modules for disease subtyping and potential prognostic application. By pairing this assay with targeted DNA sequencing, we comprehensively profiled 195 FFPE tissue specimens from benign prostatic tissue, localized tumors across a wide range of grades, and metastases and castration-resistant prostate cancer (CRPC), validating the performance of our custom RNAseq assay and highlighting our ability to comprehensively assess the diverse set of intended transcriptional modules and biomarkers. We specifically assessed >80 individual disease foci from 14 separate multifocal prostate cancer cases, deriving commercially available prognostic scores for individual foci elucidating challenges of interpreting single biopsy prognostic assays in the context of multifocal disease. Lastly, we profiled a cohort of >30 disease foci from 10 cases in which a subset of lesions was not visible through traditional magnetic resonance imaging (MRI) to assess whether genomic and transcriptomic characteristics varied between visible and invisible groups.

METHODS

Patients

Following institutional review board approval, we assembled consecutive patients who underwent radical prostatectomy at three different centers (Michigan Medicine, Ann Arbor, USA; Medical University Vienna, Vienna, Austria; and Rennes University Hospital, Rennes, France). Inclusion criteria for this study included the presence of multifocal areas of cancer detected in the fresh-frozen paraffin-embedded (FFPE) specimens. Additionally, patients with preoperative mpMRI and/or synchronous matched lymph node metastasis were included. Exclusion criteria included any previous form of prostate ablative treatment or androgen deprivation therapy. We abstracted relevant demographic, clinical and pathologic data from each patient's medical chart and recorded in a secure electronic HIPAA-compliant database.

Tissue Procurement and Nucleic Acid Isolation

Whole mount FFPE prostate and lymph node tissue (where available) were retrieved for each study participant. An anatomic pathologist with genitourinary interest (S.A.T.) reviewed all slides to confirm cancer foci, Gleason score and volume of cancer in the prostate and lymph nodes. Areas for NGS were outlined for each patient as shown in **Figure 5.5**. We obtained punch biopsies (5 punches using a 1-mm biopsy punch) of each outlined focus. In cases with small foci of cancer deemed insufficient for punching, 8 10µm unstained slide sections were obtained for microdissection. We co-isolated DNA and RNA from each primary tumor, corresponding matched lymph node metastatic foci and normal tissue (where available) using the Qiagen Allprep FFPE DNA/RNA kit (Qiagen, Valencia, CA) as described.[4] DNA and RNA were quantified using the Qubit 2.0 fluorometer (Life Technologies, Foster City, CA).

Targeted DNA/RNA sequencing

DNA libraries were generated from 1- 20ng of DNA per sample using the Ion Ampliseq library kit 2.0 (Life Technologies, Foster City, CA) and the OCP Ampliseq panel with barcode incorporation. RNA libraries were generated from 1 - 15ng of RNA per sample using the Ion Ampliseq RNA Library kit. OCP Ampliseq Libraries were quantified using the Ion Library Quantification Kit. We prepared templates for DNA and RNA libraries using the Ion PI Template OT2 200 Kit v3 on the Ion One Touch 2 and sequenced on Ion Proton P1 chips using the Ion PI Sequencing 200 Kit v3 (200 base pair reads) as described.[5]

Variant Calling (DNA)

Raw reads were aligned to the reference genome (hg19) using TMAP on Torrent Suite v. 5.0.4 (Thermo Fisher Scientific, Waltham MA). Somatic variants for DNA samples were called using Torrent Variant Caller v. 5.0.4, and annotated and filtered using previously described internal pipelines.[5-10] For cases with >1 profiled disease foci, detected alterations were evaluated across control (benign) tissue, samples from disease foci and, if applicable, lymph node metastases.

Copy Number Analysis (DNA)

Normalized, GC-content corrected read counts per amplicon for each sample were divided by those from a pool of normal male genomic DNA samples (FFPE and frozen tissue, individual and pooled samples), yielding a copy number ratio for each amplicon. Gene-level copy number estimates were determined as described previously[11-13] by taking the coverage-weighted mean of the per-probe ratios, with expected error determined by the probe-to-probe

variance. Genes with a $\log_2$ copy number ratio estimate of <-1 or >0.8 were considered to have high level loss and gain, respectively.

RNAseq Analysis

To characterize key transcriptional programs in prostate cancer and facilitate detection of alterations associated with known molecular subtypes, we developed a custom target Ion AmpliSeq RNA-sequencing panel to measure 306 amplicons measuring many markers of prostate cancer, including proliferation, stromal activity, androgen signaling, and immuno-oncology. This custom panel assays amplicons from 202 genes, multiple isoforms of 25 unique gene fusions, and 27 long non-coding RNAs. This panel assays all genes included in Myriad's Prolaris Cell Cycle Progression (CCP) score, Oncotype DX's Genome Prostate Score (GPS), and GenomeDX's Decipher Prostate Cancer Test to compare their robustness to tumor multifocality and heterogeneity. This panel also enables detection of expressed genomic variants through targeted RNAseq amplicons located at positions of interest in *NRAS, HOXB13, SPOP, IDH1*, and *HSD3B1*.

End to end read counts for RNA expression runs were calculated using Torrent Suite's Coverage Analysis plugin v5.0.4. All further analyses were conducted using The R Project for Statistical Computing v3.2.3. Housekeeping genes from Oncotype DX panel (n=5) were considered for normalization, and 4 of 5 (*ATP53, AFR1, CLTC1,* and *PGK1*) were used for normalization prior to downstream analyses. Non-fusion amplicons were filtered to ensure that all amplicons retained for analysis had >= 200 reads in at least two samples, or >1000 reads in least one sample. Raw read counts were subsequently log2-transformed, (i.e., log2(read_count +1)) and normalized to the geometric mean of expression values for the 4 retained housekeeping

genes. For heatmap visualization only, the median amplicon-level expression was calculated across all samples, and subtracted from each target amplicon's expression value prior to plotting.

Sample-level inclusion criteria for RNA data included at least 500,000 total mapped sequencing reads, with at least 60% of all sequenced reads mapping end-to-end. Housekeeping gene read mapping and expression variability were also assessed to filter out low quality samples. For each sample, the proportion of mapped reads mapping end-to-end to each housekeeping gene ('hk_prop_filt') was evaluated in a cohort of 255 samples (including a set of 66 blinded tissue specimens from GenomeDx), and the following hard gene-level hk_prop_filt thresholds were applied (based on percentile expression across the full cohort) to exclude low-quality samples: (ATP < 0.000133, ARF < .001266, CLTC < 0.001894, PGK < 0.000352). Samples with < 0.8% of all reads mapping to housekeeping genes or standard deviation of log2-normalized expression values across housekeeping genes < 0.0015 were also excluded from our analyses.

Derivation of Prognostic Scores

For each sample, we derived CCP and GPS scores[7, 9] based on previously published methods integrating expression data from component genes robustly assessed by our assay. Our custom RNAseq assay targeted all 30 transcript components used for CCP score calculation and 16 of 17 transcripts comprising GPS assay, and 12 were retained respectively after amplicon-level filters were applied.

Log2-normalized expression values for the 29 high quality CCP transcripts were floored at -5 prior to score derivation to ensure technical artifacts of RNAseq normalization did not impact score derivation. For derivation of CCP score, the previously published formula for non-

replicate experiments was used, taking the mean of each retained CCP gene's median-centered expression value to the power of 2, then log2 transforming the mean[9].

For GPS, scores were derived by adding or multiplying log2 normalized gene expression values for components of each core module as previously published[7]. The lower bound of log2 normalized expression values for *TPX2* and *SRDA5* were capped at 5 and 5.5 respectively, as described by the original authors[7]. However, we omitted multiplying individual expression values by coefficients in previous publication, as these were tuned for a qPCR methodology. As such, each module score was derived as follows:

Cellular Organization Module = FLNC + GSN + TPM2 + GSTM2

Stromal Module = BGN + COL1A1 + SFRP4

Androgen Module = FAM13C + KLK1 + SRDA5 + AZGP1

Proliferation Module = TPX2

To derive the full unscaled score, the previously published methodology was used, including the coefficients for adding component modules[7]:

*GPSu = .735*Stromal -.368*Cellular Organization -.352*Androgen + 0.95*Proliferation*

After score derivation, CCP and GPS scores were converted to percentile distributions, respectively, for ease in downstream interpretation. A one-way ANOVA was computed for each score type to determine whether there was any difference in mean score among grade groups.

Tukey's Honest Significant Difference test was used to evaluate which groups' means differed for each score type. P-values < 0.05 were deemed statistically significant.

Fusion isoform- and partner-level analyses

For initial validation analyses, fusion isoform-specific amplicons were filtered to those with >1000 reads on at least one sample. Isoform-level (e.g., *TMPRSS2:ERG.E1E4*) log2 normalized read counts were calculated as described above. For fusion partner-level (e.g., *TMPRSS2:ERG*) status, read counts for all retained isoforms were then totaled for each sample, and a normalized fusion partner value was calculated by taking the log2 of the sum of the all reads over the sum of housekeeping reads for each sample. A sample was determined as *TMPRSS2:ERG* fusion positive if it had more than 500 total reads across isoforms, and it's fusion value was greater than log2(.01). Investigation of novel fusion isoforms was carried out by mapping all targeted RNAseq reads to the hg19 reference genome with STAR (v2.3.0) using Gencode v19 splice junction annotation.

RESULTS

Validation of targeted RNAseq assay

To validate the performance of our custom multiplexed PCR-based Ampliseq targeted RNAseq panel in representative clinical biospecimens, we profiled RNA isolated from 195 formalin fixed paraffin embedded (FFPE) tissue tumor and (where available) matched normal samples from primarily untreated, primary localized or metastatic prostate cancer. A total of 167 high-quality samples were retained after sample-level quality control filters were applied. **Figure 5.1** highlights unsupervised hierarchical clustering of all high-quality non-fusion targets across

this cohort (n=167), and demonstrates our ability to broadly assess a number of relevant transcriptional modules with our assay, including proliferation, stromal, prostate cancer-specific, and immune-oncology transcriptional programs. Across ascending grades, we see proliferation marker module expression increase, consistent with increasingly more aggressive disease at higher grades, while observing expected expression patterns for markers associated with prostate adenocarcinoma vs. benign tissue (e.g., *PCA3, DLX1*). Importantly, Grade Group 1 samples taken from tumors with only Gleason Grade 3+3=6 lesions appear identical to Grade Group 1 samples taken from tumors with both high- and low-grade foci, reinforcing our ability to robustly assess expression differences across individual prostate cancer disease foci in most major clinically relevant contexts. Together these results suggest this panel is capable of assessing major transcriptional programs relevant in prostate cancer, and may offer potential as a tool for evaluating important prostate cancer biomarkers from FFPE tissue specimens.

An additional advantage of our panel is the ability to assess the major expression and mutation-based molecular subtypes in prostate cancer, including ETS family gene fusions (present in approximately 40% of prostate cancers), as well as expressed *SPOP* and *IDH1* mutations, and *SPINK1* overexpression. Figure 5.2 summarizes fusion expression, *SPINK1* gene expression, and *SPOP* mutational status supporting subtype characterization in our cohort. Fusion partner and isoform level expression data highlights our ability to capture diverse fusion isoforms across our cohort, including both canonical *TMPRSS2:ERG* isoforms as well as ETS gene family fusions involving alternate 5' (*SLC45A3*) or 3' (*ETV1*) fusion partners. In 6 samples with elevated *ERG* or *ETV1* expression, but no detected fusion isoform expression across predefined isoform targets, unbiased realignment of targeted RNAseq reads to the whole transcriptome identified robust expression of ETS family fusion isoforms not directly targeted on

our panel, likely due to combinatorial priming. We further show that amplicons targeting the activating F133* hotspot in *SPOP* enabled detection of somatic hotspot point mutations by targeted RNAseq in a subset of samples, with expressed variant fractions consistent with those observed by DNA sequencing. A number of samples with activating *SPOP* mutations also show over-expression of *SPINK1*, observations consistent with an overlap between *SPINK1* over-expression and *SPOP* mutation in some prostate cancers previously reported (**Figure 5.2**). Together these results confirm our ability to assess the major molecular subtypes of prostate cancer using a single targeted RNAseq assay, suggesting important potential clinical utility for prospective diagnostic and disease subtyping applications.

Score derivation and validation

Commercially available expression-based assays are currently used to predict risk of high-grade disease upon radical prostatectomy (RP), risk of disease recurrence post-RP, and predictions of prostate cancer specific survival. To confirm our CCP and GPS score derivations demonstrate trends across grade concordant with previous reports, we evaluated CCP and GPS component score distributions across ISUP grades (including benign and lymph node metastatic lesions). **Figure 5.2** highlights expression of all CCP transcripts, with higher-grade and lymph node metastases showing elevated CCP scores compared to benign and lower-grade samples (**Figure 5.2B**). Scores for individual GPS component modules trend in expected directions across increasing grade, and GPS scores overall increase with grade (data not shown). Together these results underscore our ability to derive several clinically-relevant prognostic scores from a

single custom targeted RNAseq assay, suggesting unique opportunities for use in both research and clinical contexts.

Multifocal cohort

Current gene expression-based assays claim utility in predicting risk of high grade disease, disease recurrence and prostate cancer-associated survival from single tissue prostate tissue biopsies, even in the context of multifocal disease, observed in >80% of prostate cancers. Thus, we first sought to evaluate our derived prognostic signatures and other candidate prognostic biomarkers (e.g. *SChLAP-1*) to multifocality using an extreme case design. We paired DNA and RNA sequencing for 84 samples (including 67 primary and 17 lymph node metastatic loci) from spatially independent disease foci in 14 separate cases of multifocal prostate cancer, including several multifocal cases with extremes of tumor foci aggressiveness. For example, case MF1 (Figure 5.4A) harbored a large Gleason score 9 [Grade Group 5] index tumor focus and a positive lymph node (pT3B N1). In addition, a small focus of Gleason score 3+3=6 cancer with PIN-like morphology (separate from the high grade focus on all levels) was present at the extreme periphery of the prostate. DNA and RNA were co-isolated from multiple regions of the large index tumor, the involved lymph node, the low grade focus, and the uninvolved prostate stroma. We first assessed the clonality of these foci using targeted mxDNAseq for 409 cancer related genes, where somatic copy number profiles demonstrated the presence of chromosome 9p somatic copy-number loss in all samples from the high grade tumor focus and the lymph node metastases, but not in the low grade Gleason score 6 focus (Figure 5.4B). Further, a shared *TP53* (chr17:7577568, C>T) somatic mutation was present in all samples from the high grade tumor

focus and the lymph node metastasis, but again not detected in the Gleason score 6 focus (Figure 5.4C). Taken together, these results demonstrate that the low and high grade tumor foci in this case represent true multifocality at the extremes of aggressiveness by both molecular and usual histopathologic criteria, making it ideal to assess the robustness of transcriptomic biomarkers to multifocality.

Co-isolated RNA from all samples from the high and low grade foci, benign stroma and lymph node metastasis in this case were subjected to mxRNAseq with our assay, with 6 samples being informative. All tumor samples had detectable *T2:ERG* expression and over-expression of genes most discriminative of prostate cancer vs. benign prostate tissue (e.g. *AMACR*, *DLX1* and *PCA3*), consistent with the morphologic impression of carcinoma. Likewise, both derived mxGPS and mxCCP scores (Figure 5.4D) were similarly higher in the high grade tumor foci compared to the low grade tumor focus. These results suggest that prognostic scores derived from gene expression profiling of single tissue biopsy samples may not truly be robust to multifocality in the context of multifocal disease with clonally independent disease exhibiting extremely different histopathological grades (e.g., ISUP Grade Group 5 vs Grade Group 1).

Results from a second case with extremes of tumor foci aggressiveness by histopathology show similar results. In MF9, we co-isolated DNA and RNA from 6 areas of a large, high grade pT3a tumor (overall Gleason score 4+5=9, Grade Group 5), including 4 areas of Gleason score 4+5, one area of Gleason score 4+3 in an area of extraprostatic extension in the apex, and one area of Gleason score 3+4. Multiple, histologically separate Gleason score 3+3=6 tumors were present in the base, with two samples taken from the largest Gleason score 3+3=6 focus. Lastly, we also sampled benign prostate tissue (mixed epithelium and stroma) in close proximity to the sampled 3+3=6 focus. By mxDNAseq of 7 high-quality specimens, a somatic *MED12*

(chrX:70349252, G>C) point mutation was detected in a single low-grade tumor foci (PR-406_RNA), Gleason 3+3=6; Grade Group 1), but not present in any of the high grade foci, with a focal 2-copy deletion of PTEN seen in all high grade loci, but not observed in low-grade foci, suggesting independent clonal origins for low-grade and high-grade foci. By mxRNAseq, *T2:ERG* fusions were detected in all high grade tumor focus samples, but not in the low grade tumor focus samples, provide further support for true disease multifocality. Importantly, markedly different mxCCP and mxGPS scores are observed across foci, with high grade foci showing substantially higher CCP scores than the low-grade focus (PR-406_RNA). Together, these results clearly demonstrate multiclonality of these tumor foci, and show that in some multifocal cases with extremes of tumor aggressiveness, potentially prognostic biomarkers and derived signatures appear closely aligned with histologic grade.

To determine the frequency in which an under- or unsampled high grade disease focus may lead to extreme upgrading at radical prostatectomy (and thus was missed by the initial biopsy), we used the Radical Prostatectomy database at the University of Michigan from all cases with complete biopsy and prostatectomy pathology information from 2005-2013. We identified a total of 1,418 men with biopsy Gleason score 6 (710 men) or 3+4 (708 men) who underwent prostatectomy. Of these men, 283 (20.0%) had OncotypeDX defined adverse pathology at radical prostatectomy (>pT2, primary pattern 4 or any pattern 5; 69/710 [9.7%] and 214/708 [30.2%] with Gleason score 6 and 3+4=7, respectively). Though it cannot be determined whether the biopsy undersampled the higher grade focus or simply did not sample the higher grade focus, 21 of the 1,418 men (1.5%; 5/710 [0.07%] and 16/708 [2.3%] of men with biopsy Gleason score 6 and 3+4=7, respectively) showed extreme upgrading on prostatectomy (to only Gleason scores >=4+4=8 [Grade groups 4 and 5]), where the biopsy almost certainly missed the

RP defined index tumor focus. Hence, our results suggest that ~1% of prognostic tests performed on men with Gleason score 6 or 3+4=7 prognostic tests should report expression signatures consistent with extremely aggressive prostate cancer if they are truly robust to multifocality.

DISCUSSION

Herein, we describe the validation of a multiplexed, PCR-based targeted RNAseq assay compatible with routine clinical FFPE tissue specimens that can robustly assess major transcriptional modules and molecular alterations relevant to prostate cancer biology from, enabling characterization of major prostate cancer molecular subtypes, including samples with *ETS* (including *ERG* and *ETV1*) family gene fusions, expressed *SPOP* point mutations, and over-expression of *SPINK1*. We see expected trends in individually prognostic biomarkers (e.g., SChLAP-1) and expression modules (e.g., proliferation) across grades, and demonstrate the ability to robustly derive expression-based prognostic scores routinely used for optimal case management and treatment stratification. Importantly, we show that in the context of true disease multifocality and extremely divergent histopathological grade, prognostic scores from individual disease foci may not yield consistent results, and thus may yield false negative results in the presence of un- or undersampled high grade components or individual foci at diagnostic biopsy.

This study leveraged paired DNA and RNA sequencing on a large cohort of routine clinical tissue specimens to comprehensively characterize molecular profiles at both the genomic and transcriptional level. Importantly, these profiles enabled systematic comparison of genomic and transcriptional alterations across individual foci from the same tumor, supporting evaluation of derived prognostic scores in the context of multifocal disease with very high and low grade

disease. We characterize the diverse set of ETS family fusion isoforms present across samples sequenced on our panel, and show our panel has the capacity to identify novel fusion isoforms through combinatorial priming of 5' and 3' fusion partners targeted on the panel.

Continued iterations to expand the utility of this panel will focus on more robust assessment of potentially predictive biomarkers such as androgen receptor splice variants (e.g., *ARv7*). By removing genes such as *MALAT1* and *NEAT1* (which, in many samples, collectively account for >60% of all mapped reads), we anticipate being better able to characterize the dynamic range of individual target and fusion isoform expression, as well as add expanded immune-oncology and long noncoding RNA targets for exploration of prognostic biomarker potential. Further, while this panel targets genes included in GenomeDx's Decipher test, appropriate gene- and/or sub-component-specific weights used for this proprietary classifier are unknown, preventing recapitulation of Decipher test scores in this cohort. Additional computational work will explore the best way to leverage Decipher markers, in conjunction with targets used in CCP and OncotypeDx assays, for improved clinical prognostication.

Overall, this work summarizes a powerful NGS-based approach compatible with FFPE tissue specimens for characterizing molecular profiles of individual tumor foci, demonstrating robust assessment of biologically relevant genomic and transcriptional alterations in a representative cohort (from benign prostatic tissue through high-grade (or neuroendocrine) and castration-resistant disease) covering the full spectrum of prostate cancer. By systematically assessing prognostic scores in the context of true multifocal prostate cancer using our custom targeted RNAseq assay, we show divergent score predictions in multiple cases with clonally-independent high- and low-grade disease foci. We also note the frequency of extreme disease-upgrading at prostatectomy in a clinical prostate cancer cohort, suggesting high grade

components of the same tumor or independent high-grade disease foci were undersampled or simply missed at biopsy. Together, these results highlight important clinical scenarios in which current prognostic classifiers may have limited utility, and elucidate important opportunities for continued prognostic biomarker classification and development work.

Figure 5.1 – Unsupervised hierarchical clustering of mxRNAseq data enables assessment of major biologically-relevant transcriptional modules in prostate cancer

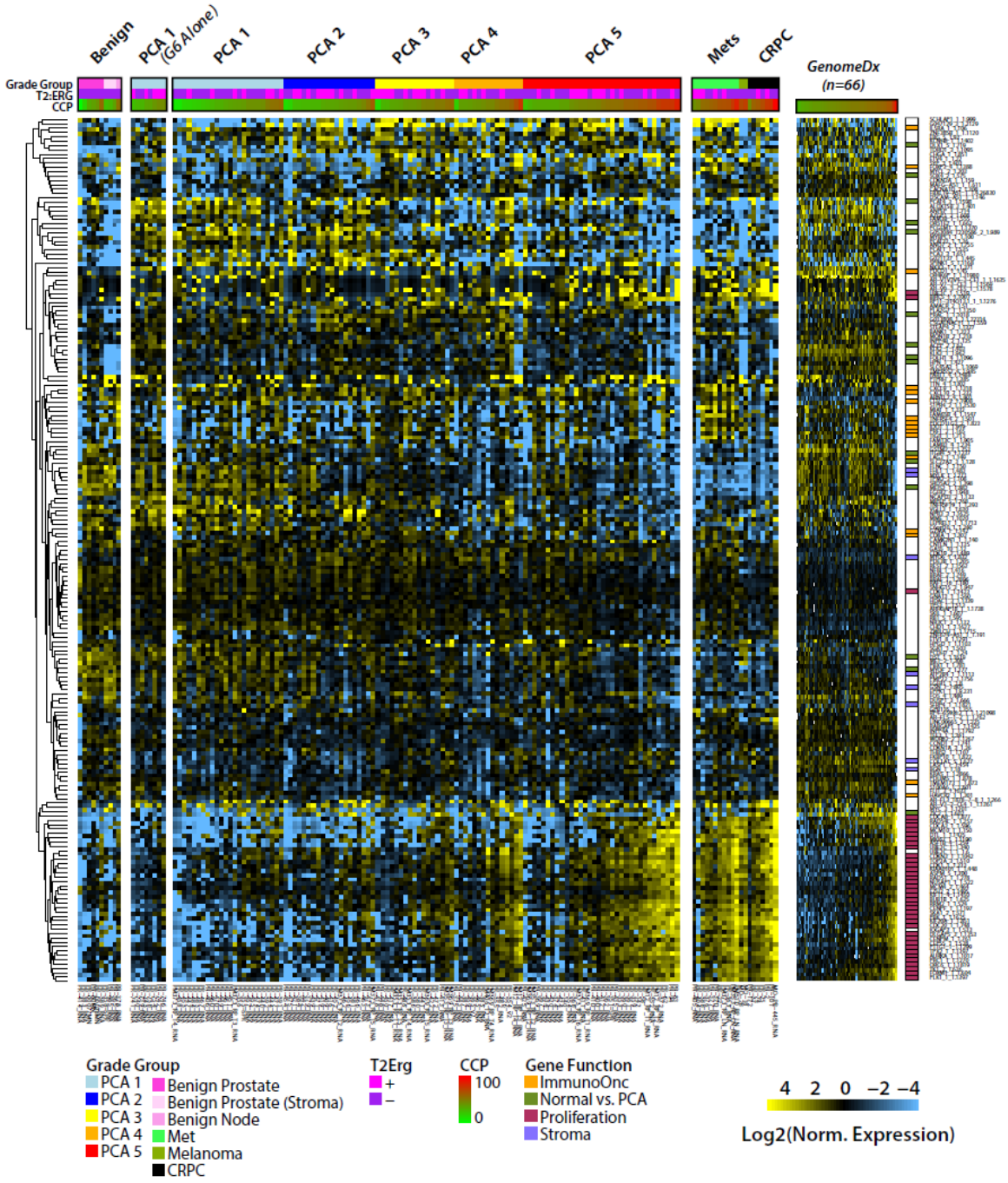


Figure 5.1 Unsupervised hierarchical clustering of log2 normalized expression values for 235 non-fusion gene targets across 215 high-quality tissue samples profiled via mxRNAseq. Samples are ordered from left to right by increasing ISUP grade group, including (in order): benign, PCA1 (from tumors with only gleason 3+3=6 lesions), PCA1, PCA2, PCA3, PCA4, PCA5, metastatic lesions, and lesions from individuals with castration resistant prostate cancer (CRPC). ISUP grade group, TMPRSS2:ERG (T2:ERG) fusion status, and derived CCP score annotation (header rows) are colored according to the legend at bottom. 66 blinded (lesion grade unknown) samples from industry collaborator GenomeDx are displayed at right. On far right, gene annotations are provided vertically, with colors assigned according to legened at bottom. Expression values have been capped at +5 and -5 for clarity in display.

Figure 5.2 – Robust assessment of major prostate cancer molecular subtypes via mxRNAseq

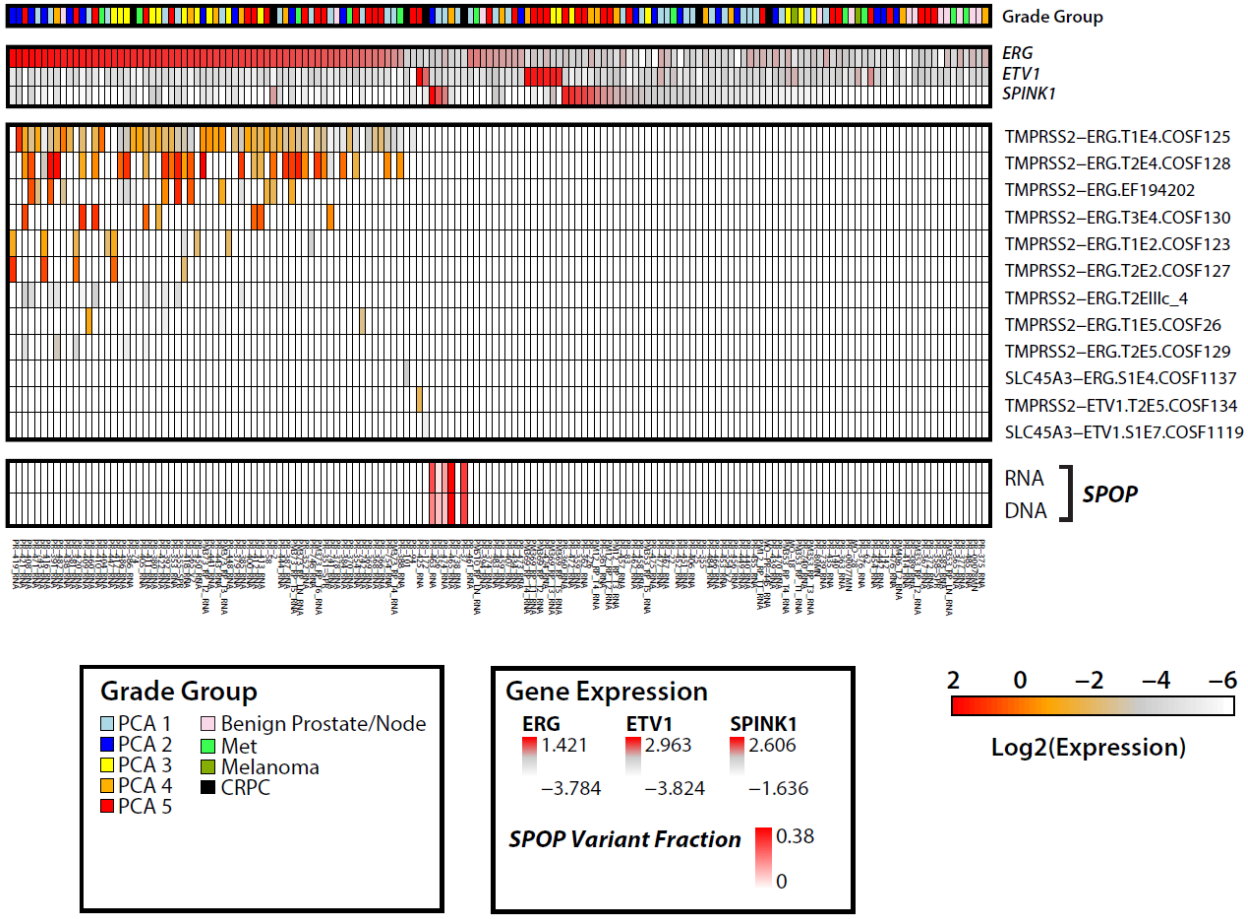


Figure 5.2 Fusion isoform heatmap for 149 high-quality tissue specimens profiled by mxRNAseq. Samples are sorted from left to right by: decreasing total log2 ERG fusion isoform expression values, decreasing total log2 ETV1 fusion isoform expression, decreasing SPINK1 expression for samples with expressed SPOP point mutation, outlier ERG expression (no targeted fusion isoform expression), outlier ETV1 expression (no targeted fusion isoform expression), then decreasing SPINK1 expression. ISUP grade group is identified on top header row, with ERG, ETV1, SPINK1 individual gene expression in subsequent 3 header rows, each of which is colored according to expression value scales in legend at bottom. Fusion isoforms are listed at far right in main heatmap, and sorted from top to bottom by decreasing total expression across samples within fusions involving ERG, then fusions involving ETV1. The bottom two rows depict expressed variant fraction for SPOP hotspot mutation in RNA, and variant fraction detected by paired targeted DNA sequencing of the same sample.

Figure 5.3 – Derived CCP scores increase with grade, demonstrating robust expression across individual gene targets contained by mxRNAseq

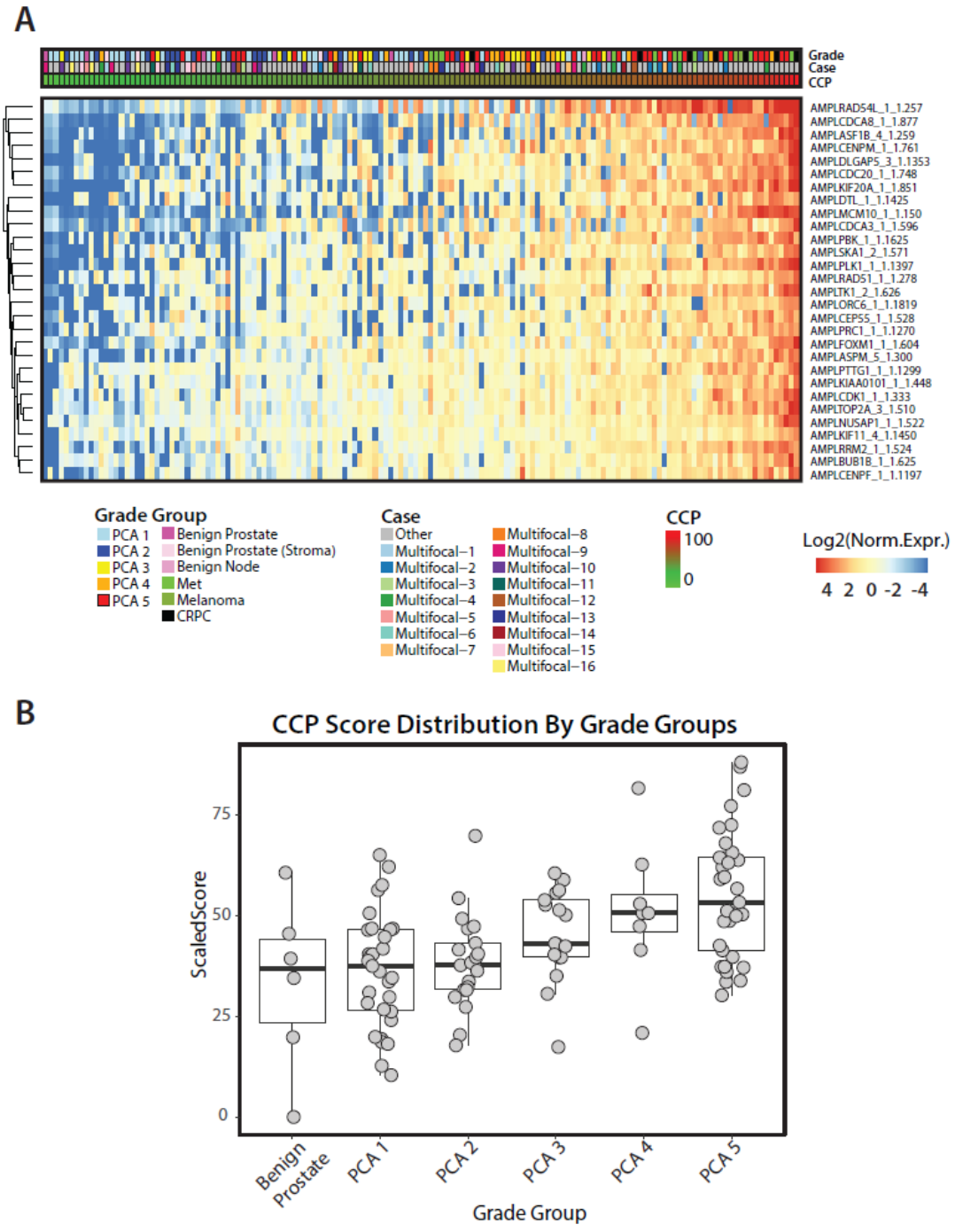


Figure 5.3 A. Unsupervised hierarchical clustering of high-quality CCP gene targets across 149 high-quality tissue samples profiled on mxRNAseq panel. Samples are sorted from left to right by ascending derived CCP score percentile, with grade, case, and CCP score info colored according to legend at bottom. Individual CCP gene targets are labeled at right. Expression values were capped at +5 and -5 to control for individual sample/target outliers. **B.** Box plots of derived CCP score percentile are plotted across grade (including benign lesions), demonstrating increasing CCP score with increasing ISUP grade.

Figure 5.4 - **Divergent prognostic scores in the context of true disease multiclonality**

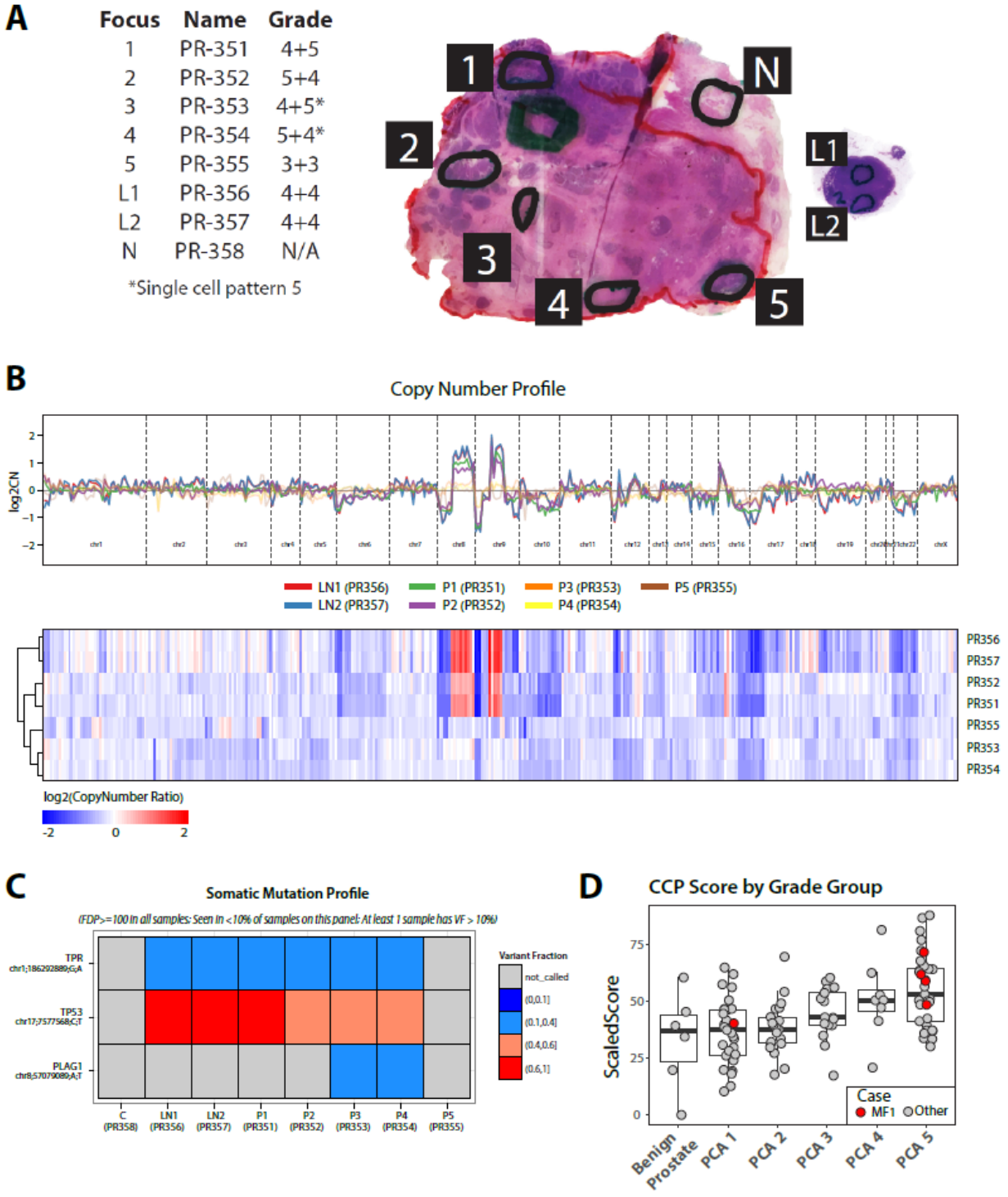


Figure 5.4 A. Image of whole mount prostatectomy specimen with multifocal prostate cancer (Case 1). Each of 8 individual foci profiled by mxDNAseq and mxRNAseq are labeled, including: one primary Gleason 3+3=6 focus [PR-355], 4 primary Gleason 9 (4+5 or 5+4) foci [PR-351, PR-352, PR-353, PR-354], 2 lymph node 4+4 metastases [PR-356, PR-357], and 1 non-tumor (control) prostate tissue specimen [PR-358]. **B.** Genome-wide traces (top) and unsupervised hierarchical clustering of genome-wide copy-number profiles for 7 tumor specimens profiled by mxDNAseq using the 409 gene Oncomine Comprehensive Cancer Panel. For these plots, the control (normal) prostate tissue specimen (PR-358) was used a reference for computing copy-number

ratio estimates. In top panel, lines represent genome-wide trace of log2 gene-level copy-number ratios for each profiled sample, and are colored according the legend between the trace graphic and clustered heatmap. In clustered heatmap, primary Gleason Grade 9 foci PR-351 and PR-352 cluster with lymph node mets PR-356 and PR-357, showing concordant chr8p loss / chr8q gain, as well as highly altered chr9, while Gleason Grade 6 primary foci PR-355 shows limited genome wide copy-number alterations (with Gleason Grade 9 primary foci PR-353 and PR-354 clustering separately). **C.** Variant fractions for prioritized somatic variants detected across lesions profiled for case 1. Somatic TPR and TP53 nonsynonymous SNVs are detected in all foci with Gleason Grade >8, but not the Gleason 3+3=6 (PR-355) or control (PR-358) lesion, and when considered with copy-number data support true multiclonal disease. The two primary foci with Gleason 9 disease also show a PLAG1 nonsynonymous SNV not seen in any other specimen supporting alterations acquired independent of other lesions. **D.** Derived CCP scores for all primary foci are displayed for case 1 in the context of derived values for all benign and primary (PCA1-5) tumor specimens, highlighting discordant results for the low- vs. high-grade disease.

Chapter V References

1. Schwarzenbach, H., D.S. Hoon, and K. Pantel, *Cell-free nucleic acids as biomarkers in cancer patients.* Nat Rev Cancer, 2011. **11**(6): p. 426-37.
2. Ludwig, J.A. and J.N. Weinstein, *Biomarkers in cancer staging, prognosis and treatment selection.* Nat Rev Cancer, 2005. **5**(11): p. 845-56.
3. Tomlins, S.A., et al., *Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA.* Sci Transl Med, 2011. **3**(94): p. 94ra72.
4. Andriole, G.L., et al., *Mortality results from a randomized prostate-cancer screening trial.* N Engl J Med, 2009. **360**(13): p. 1310-9.
5. Schroder, F.H., et al., *Screening and prostate-cancer mortality in a randomized European study.* N Engl J Med, 2009. **360**(13): p. 1320-8.
6. Prensner, J.R., et al., *Beyond PSA: the next generation of prostate cancer biomarkers.* Sci Transl Med, 2012. **4**(127): p. 127rv3.
7. Knezevic, D., et al., *Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies.* BMC Genomics, 2013. **14**: p. 690.
8. Erho, N., et al., *Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy.* PLoS One, 2013. **8**(6): p. e66855.
9. Cuzick, J., et al., *Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study.* Lancet Oncol, 2011. **12**(3): p. 245-55.
10. Klein, E.A., et al., *A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling.* Eur Urol, 2014. **66**(3): p. 550-60.
11. Grasso, C., et al., *Assessing copy number alterations in targeted, amplicon-based next-generation sequencing data.* J Mol Diagn, 2015. **17**(1): p. 53-63.
12. Cani, A.K., et al., *Next-Gen Sequencing Exposes Frequent MED12 Mutations and Actionable Therapeutic Targets in Phyllodes Tumors.* Mol Cancer Res, 2015. **13**(4): p. 613-9.
13. Warrick, J.I., et al., *Tumor evolution and progression in multifocal and paired non-invasive/invasive urothelial carcinoma* Virchows Arch, 2014.

CHAPTER VI: Conclusion

In this work, I have devised and validated multiple next-generation sequencing (NGS) based analytic approaches capable of supporting clinically-feasible molecular profiling from routine clinical biospecimens including blood, urine, and formalin-fixed paraffin embedded tissue samples. These approaches offer unique opportunities for leveraging rapid, scalable and statistically robust analytic workflows to inform on molecular profiles characterizing clinically-relevant somatic alterations.

As outlined in Chapter I, NGS-based profiling work has played an important role in the elucidation of both canonical molecular alterations common in prostate cancer, as well as the evolution of molecular profiles in the context of progressive disease and in response to treatment over time. Chapter II describes the analytic validation of a targeted DNA and RNA sequencing system compatible with formalin-fixed paraffin embedded tissue specimens (and extended successfully to other plasma cell-free DNA profiling), paving the way for scalable sequencing efforts in high-throughput translational research contexts, including initiatives I have led in rare cancer cohorts or unique clinical contexts[1, 2]. Further, this assay was utilized in the initial phase of the NCI-MATCH trial[3], demonstrating clear opportunities for prospective clinical trial and/or CLIA-validated clinical workflows.

As flexible approaches to profiling repeat non-invasive biospecimens become more feasible, diverse NGS-based approaches (including targeted and genome-wide) will likely hold

utility for extracting clinically meaningful intra-individual or cohort-specific trends from rich, high-dimensional NGS-based datasets. Analytic workflows deploying blood and urine cell-free DNA WGS sequencing such as those described in Chapter III and Appendix A) can be refined and paired with orthogonal profiling techniques (RNA- or protein-based) to both generate and digest diverse clinical and translational datasets. Applications of our work to single-cell sequencing from circulating tumor cells, for instance, or individually isolated cells from heterogeneous solid tumor cell populations will be extremely informative, albeit with uncertain potential clinical utility. Ongoing work as a member of the Blood Profiling Atlas Consortium[4] stimulated by 2016's Precision Medicine Initiative launch will provide me with continued opportunities to engage with other experts in the field to pursue innovative strategies for maximizing the potential of liquid biopsy applications in concert with relevant tissue-based profiling workflows.

Extensions of this work to explore the effects of histological divergent differentiation on expression-based subtyping across cancers will be extremely important, particularly in understanding the role DNA alterations play in informing things such as expression-based subtype membership. As pathology workflows become digitized, automated analyses capable of leveraging inferred tissue histological heterogeneity to inform NGS-based molecular profiling results will likely become more plausible, making integrative assessment of driving somatic DNA and RNA alterations from individual tumor components increasingly relevant. Additionally, recently formed initiatives such as the Pre-Cancer Genome Atlas[5] will take wide aim at better understanding what molecular alterations in precursor lesions may confer selective advantages for dysregulated cellular proliferation and tumor formation, making the precise and

robust assessment of DNA and RNA based molecular profiles from minute individual tumor components as described in Chapters IV and V extremely important.

Further, as relevant datasets abound, applications of principles from this work can potentially be expanded to infer and identify potentially predictive and prognostic biomarkers in a wide set of clinical contexts. Chapters IV and V highlight important opportunities for expanding biomarker identification and assessment, characterizing robust derivation of commercially available prognostic scores or biologically-relevant molecular subtypes, suggesting important advantages for leveraging validated customized targeted sequencing assays in translational research or clinical settings. As clinical oncology evolves in response to biological and technological advances, assays and analytic approaches capable of assessing relevant predictive and prognostic biomarkers may become essential for guiding clinical decision-makers. For instance, the emergence and power of immunotherapeutic approaches for inducing dramatic clinical responses in subsets of patients with certain types of cancer will inevitably require continued identification and monitoring of biomarkers of resistance or response, especially in the context of widely-acknowledged heterogeneity in response profiles across cancers. Flexible disease- or system-specific NGS-based approaches for identifying predictive and prognostic biomarker discovery will require innovative and statistically robust analytic platforms that may benefit from principles of the work reported herein.

Continued work enhancing the precision of blood- or urine-based copy-number profiling and tumor content approximation will help to more precisely estimate the relative contribution of tumor-derived circulating DNA in samples from patients with advanced cancer, supporting cost-effective disease monitoring and treatment response evaluations from routine fluid biospecimens. Leveraging both tissue- and liquid biospecimen profiling described herein to complement

existing comprehensive tissue-based molecular profiling strategies in the context of longitudinal translational research or clinical trials may help refine and expand an understanding of relevant disease biology and clinically-useful biomarkers over time. Further, such work should shed light on optimal clinical utility for proposed cfDNA and tissue-based targeted RNAseq profiling strategies. As the goal of any precision oncology efforts are ostensibly to decrease patient mortality and improve quality of life and clinical outcomes for patients with cancer, this dissertation describes a collection of scalable analytic approaches to increase NGS-guided precision oncology opportunities from routine clinical biospecimens, offering important potential for impactful translational research and clinical work.

Chapter VI References

1. Warrick, J.I., et al., *Tumor evolution and progression in multifocal and paired non-invasive/invasive urothelial carcinoma.* Virchows Arch, 2015. **466**(3): p. 297-311.
2. McDaniel, A.S., et al., *Genomic Profiling of Penile Squamous Cell Carcinoma Reveals New Opportunities for Targeted Therapy.* Cancer Research, 2015. **75**(24): p. 5219-5227.
3. Lih, C.J., et al., *Analytical Validation of the Next-Generation Sequencing Assay for a Nationwide Signal-Finding Clinical Trial: Molecular Analysis for Therapy Choice Clinical Trial.* J Mol Diagn, 2017. **19**(2): p. 313-327.
4. Grossman, R.L., et al., *Collaborating to Compete: Blood Profiling Atlas in Cancer (BloodPAC) Consortium.* Clinical Pharmacology & Therapeutics, 2017. **101**(5): p. 589-592.
5. Campbell, J.D., et al., *The Case for a Pre-Cancer Genome Atlas (PCGA).* Cancer Prev Res (Phila), 2016. **9**(2): p. 119-24.

APPENDICES

APPENDIX A: Urine CfDNA Copy-Number Profiling

INTRODUCTION

While we and others have demonstrated potential clinical utility for identifying therapeutically informative alterations and improving precision oncology patient stratification with a plasma-based sequencing workflow, the emerging role of urine-based cellular or cell-free DNA assessment in precision oncology is still an active area of investigation [1]. Given the relative ease and absolute noninvasive nature of urine sample collection (compared to tissue and even blood) and the ability to serially collect reasonably high volumes ostensibly each day, urine-based biomarker assessment and monitoring presents as an attractive potential strategy for informing precision oncology workflows[2]. Despite several reports of point mutation profiling from urine cfDNA across various cancers[3, 4], technical limitations to scalable, genome-wide urine cfDNA profiling for non-urothelial cancers[5] persist, including questions around optimal DNA isolation and purification strategies, fidelity of isolated DNA, variability in non-tumor urine cfDNA sources, fragment lengths of tumor-derived cfDNA, and uncertainty around whether trans-renal tumor DNA can be reliably assessed for prospective treatment monitoring and response.

Here, we leverage single-strand DNA library preparation protocols to establish genome-wide copy-number profiles from whole-genome sequencing of urine cell-free DNA samples in a

set of patients with acute myeloid leukemia (n=14) or solid tumors (n=9), highlighting high concordance with profiles from clinical karyotype and/or synchronous plasma cfDNA WGS. We demonstrate the ability of our approach to identify therapeutically relevant copy-number alterations (including both focal amplifications and broad copy-number gains or losses) from urine cfDNA samples, suggesting important potential clinical utility for monitoring disease burden and treatment response. Importantly, we establish comprehensive fragment length distribution profiles across patient sampels, and identify an enrichment of ultra-short (20-40bp) cell-free tumor DNA in both patients with AML and those with solid tumors. We confirm ultra-short fragments are also present in synchronous plasma samples and demonstrate genome-wide copy-number profiles using only reads from ultra-short cfDNA fragments profiles mirror those generated from unrestricted mapped sequencing read analysis. These results are consistent with trans-renal passage of ultra-short cfDNA fragments from the blood to the urine, and suggest a number of interesting potential biological and translational opportunities.

METHODS

Preparation of sequencing libraries

Plasma cfDNA and fragmented blood genomic DNA sequencing libraries were prepared with the ThruPLEX Plasma-seq kit (Beckman Coulter) at 1-10 ng of DNA input, individually barcoded, and pooled for purification using AMpure beads, according to the manufacturer's instructions. Urine DNA sequencing libraries were prepared at 5-10 ng of cfDNA input using a protocol adapted from the single-stranded DNA library preparation method (3) with the

following modifications: (i) uracil excision and DNA cleavage step was omitted. (ii) single stranded ligations were performed with the addition of 1 µL instead of 4 µL of Circligase II (Epicentre) (4), and incubated overnight instead of 1 hr, which reduced reagent cost while maintained the ligation efficiency. (iii) double stranded ligations were performed for 2 hr at room temperature instead of 1 hr (5). (iv) for hybrid select target capture, double stranded Adapter 2 with16 randomized bases (hand mixed) was used for double stranded ligations and modified indexing primers were used for library amplification. For whole genome sequencing, dual indexed urine DNA libraries were amplified at optimal cycles determined by SYBR qPCR (3), and followed by size selection using 10% TBE gel (Bio-Rad). For hybrid select target capture, libraries were amplified into saturation and purified using AMpure beads according to the manufacturer's instructions. Purified pooled plasma DNA libraries and individual urine DNA libraries were quantified using ddPCR™ Library Quantification Kit for Illumina TruSeq (Bio-Rad), in which the annealing temperature for urine libraries was modified from 60 ℃ to 55 ℃, due to the truncated P5 adapter.

DNA Sequencing and primary data processing

All libraries were sequenced on MiSeq or HiSeq 2500 instruments (Illumina) otherwise stated, with details of sequencing are provided in Table S2. The raw fastq files were run through FastQC v0.10.1 to check quality before trimming. Adapters were removed using cutadapt with the settings '-q 20,15 -m 5 -overlap=4'. After adapter trimming, FastQC v0.10.1 was run again to confirm that the trimming was successful and the quality of the reads was suitable for analysis. The reads were aligned to the hg19 Human reference sequence using BWA v.0.7.15-r1140 with

options ‘mem -t 4 -k 8’. The resulting SAM files were sorted, indexed and compressed to a BAM file using Picard Tools v2.5.0 ‘SortSam’. Various metrics were calculated using the Picard Tools ‘CollectAlignmentSummaryMetrics’, ‘CollectInsertSizeMetrics’ and ‘QualityScoreDistribution’ functions in addition to the Samtools (v0.1.19) ‘idxstats’ function.

Copy-number variant detection: WGS

Non-PCR-duplicate reads (samtools v1.3) were used to identify candidate copy-number alterations using the QDNASeq R package (version 1.6.1) [6]. Briefly, the genome was divided up into variable bin sizes (15, 25, 50, 100, 500, and 1,000 kilobase-pair bins), and bin-level counts of high-quality mapped reads (MAPQ $\geq$ 37) were calculated separately for each sample. Raw bin-level counts were simultaneously corrected for GC content and mappability by fitting a LOESS surface through median read counts for bins with the same combination of GC content and mappability and dividing raw bin-level counts by the corresponding LOESS fitted value. GC- and mappability-corrected bin-level counts were then normalized by median bin-level corrected counts within each sample. Bins previously shown using either ENCODE or 1000G data to yield anomalous copy-number results due to germline copy number variants (CNVs), low mappability, or large stretches of uncharacterized nucleotides were excluded [6]. For each bin in each tumor sample, high-quality, corrected, median-normalized read counts were divided by average corrected, median-normalized read counts from our 5 normal male samples. Segmented copy-number events were called from bin-level corrected, median- and control-normalized read counts using the circular binary segmentation algorithm implemented by the DNACopy (1.44.0) R package, and final segment- and bin-level copy-number values were used for subsequent

analyses as described. Focal CNAs were defined as CNAs 1.5-20Mb long with a log2(CNRatio) ≥ 0.2, thresholds similar to those described elsewhere [7].

RESULTS

To confirm our ability to detect genome-wide copy-number profiles from whole genome sequencing of urine cfDNA, we first profiled urine cfDNA samples and matched blood genomic DNA samples from 14 patients with AML using single-strand DNA library preparation protocols. Initially, using all high-quality (>Q30) sequencing reads from single-strand urine cfDNA sequencing libraries and a pool of normal genomic DNA samples as a reference, only 3 of 14 (21%) AML samples showed urine cfDNA copy-number profiles concordant with those reported in clinical karyotyping and verified via plasma cfDNA and blood genomic DNA WGS sequencing. However, by stratifying fragment length distributions based on whether fragments mapped with high-quality to genomic regions known to be gained, lost, or unaltered by clinical karyotype, we observed a significant enrichment of ultra-short (20-40bp) fragments mapping to regions affected copy-number gain, suggesting a tumor-specific enrichment of ultra-short cfDNA fragments in urine cfDNA samples (data not shown). By restricting our analysis to sequenced fragments <40bp in length in all urine cfDNA samples from patients with AML, we show that 14/14 (100%) samples show detectable copy-number profiles consistent with clinical karyotype and plasma cfDNA profiles (**Figure A1A**). Together these results confirm that tumor-derived DNA detectable in urine can be leveraged to establish genome-wide copy-number profiles consistent with clinical karyotype assays and profiles derived from patient-matched plasma cfDNA samples. They further support that despite low overall tumor content in most samples,

tumor-specific cfDNA fragments in the urine appear to be enriched at ultra-short lengths, carrying important implications for prospective urine cfDNA assay development and implementation.

We next sought to evaluate whether WGS of urine cfDNA samples from patients with solid tumors could yield detectable genome-wide copy-number profiles. Here again, unrestricted analysis of all high-quality sequenced fragments generally indicated low overall tumor content estimates, limiting the ability to detect variable-sized copy number alterations originally identified via plasma cfDNA. However, stratified fragment length distribution analyses based on fragments mapped to regions identified as copy-number altered in plasma cfDNA sequencing again highlighted an enrichment of ultra-short (20-40bp) in regions of copy-number gain, with enhanced signal for reads mapped to regions focally amplified. These observations were reinforced by genome-wide copy-number profiles from urine cfDNA WGS data stratified by fragment length (**Figure A1B**). These results reinforce tumor-specific urine cfDNA fragment length observations from patients with AML, and suggest that detectable trans-renal tumor DNA in urine can facilitate genome-wide copy-number profiles for patients with solid tumor samples.

One hypothesis may be that cfDNA fragments detected in urine samples are originally present in blood, and are detectable in urine after passage through the glomerulus filter (which may filter out longer, higher molecular weight cfDNA). To evaluate whether ultra-short tumor-specific cfDNA fragments exist in the blood, we first analyzed a single cfDNA sample from a patient with squamous cell carcinoma of the lung sequenced at high-depth (>25x) genome wide in an orthogonal plasma cfDNA profiling cohort [8]. Though default analysis parameters in the original study prevented mappability of cfDNA fragments less than ~35bp, we were still able to

see clear tumor-specific copy-number profile signals at similar tumor contents when comparing profiles derived from ultra-short (35-50bp) and known peak plasma circulating tumor DNA fragment length ranges (**Figure A1C**). Analyses on plasma cfDNA samples from patients with AML or solid tumors profiled in our cohort (with libraries prepared using either single-stranded DNA protocols) show similar signal, supporting the possibility that tumor-derived ultra-short cfDNA fragments detected in the urine may, to some extent, be derived from populations of cfDNA molecules circulating in the blood. Together with our fragment length restriction and urine cfDNA whole-genome copy-number analyses, these results identify important utility and analytic considerations for prospective urine-based cfDNA NGS profiling, particularly for the development and application of urine-based cfDNA molecular profiling assays.

DISCUSSION

High-dimensional sampling from patients with both localized and metastatic cancer in the context of disease progression and response to treatment may provide a more complete picture of driving molecular alterations over time, guiding clinical decision-making and enhancing potential for true precision oncology in clinical practice. To this end, we have extended whole-genome sequencing genome-wide copy-number profiling to urine cell-free DNA in patients with both circulating and solid tumor disease, and showed a robust ability to establish genome-wide copy-number profiles, while characterizing the size distribution of urine circulating tumor DNA (ctDNA) with precision across a diverse sample and cancer compendium. Here, we have leveraged single-strand DNA library protocols to sequence the full spectrum of urine cell-free DNA fragments, showing that ultra-short (20-40bp) cfDNA fragments are over-represented in

reads mapping to regions with elevated tumor contributions (e.g., regions gained or amplified), supporting peak urine ctDNA lengths of <40bp. Further we show that while the bulk of plasma cfDNA peak fragments are sized in the 150-200bp range, ultra-short (<40bp) fragments exist and contain similar tumor contents as plasma fragments in the larger size range.

Interesting questions arise around whether ultra-short ctDNA or cfDNA fragments detectable in the blood are the same fragments that make their way into the urine (or rather, fragments in the urine simply represent degraded versions of the longer plasma cfDNA molecules). However, our results clearly demonstrate that copy-number profiles derived from ultra-short trans-renal DNA (trDNA) recapitulate profiles from matched plasma or blood genomic DNA in patients with both solid and hematologic malignancies. Together with the relative ease of repeat and/or large volume urine collection in comparison to blood, these results support important potential utility for pairing urine-based molecular profiling approaches with clinical tissue and blood-based assessment strategies in managing treatment and monitoring disease burden in patients with advanced cancer.

Continued computational experimentation will help to refine our urine cfDNA copy-number profiling approach, informing parameterization best suited for ultra-short fragment analyses. For instance, due to enrichment of ultra-short ctDNA fragment length, precise determination of optimal seed length during the alignment phase (for standard NGS or even plasma cfDNA experiments, read mapping seed lengths of 30-40bp are typically used) will be critical for prioritization of high-confidence alignment of ultra-short urine cfDNA fragments in prospective work. Further, we expect systematic downsampling of high-depth WGS urine cfDNA sequencing datasets to enable a more precise estimate of the relationship between

coverage and assessable tumor content from urine cfDNA samples. Additional analyses exploring copy-number profiling performance using alternative reference cohorts for solid tumor plasma and urine cfDNA samples (e.g., matched blood genomic DNA vs. cohort of normal plasma cfDNA samples) will aid in identifying the most appropriate controls for use in potential prospective clinical implementation. Lastly, strategies for evaluating tumor heterogeneity and clonal dynamic representation from plasma cfDNA sequencing[9-11] may have utility in urine cfDNA NGS profiling, and could provide a relevant framework for extensions of (or synergy with) our work.

Collectively, this work suggests that by pairing expanded profiling of serial urine samples with synchronous plasma cfDNA and longitudinal clinical outcome variables, our approach may support a more robust assessment of disease burden and DNA-based biomarker representation over time from routine clinical liquid biospecimens. Further work integrating assessment of circulating RNA, exosomal or circulating tumor cell (CTC) DNA and RNA, high-depth targeted urine cfDNA NGS mutation profiling, and even patient-matched comprehensive tissue-based profiles will help to shape the potential utility of urine cell-free DNA NGS for noninvasively characterizing relevant disease-specific DNA alterations and informing clinical care. Systematic evaluation of sensitivity and specificity for our approach across cancer types and alteration size will also enable more targeted clinical utility, perhaps supporting candidate predictive or prognostic biomarker assessment (such as *AR* amplification in prostate cancer) in relevant disease cohorts.

Ultimately, we have shown that urine cfDNA WGS can recapitulate genome-wide copy-number profiles assessed by clinical karyotype and/or plasma cfDNA WGS, including both focal

amplifications and broad copy-number gains and losses, and demonstrate these profiles are most enriched for tumor content when prioritizing ultra-short (20-40bp) urine cfDNA fragments. These results support the presence of detectable ultra-short trDNA fragments that may ostensibly be present in blood and transported to the urine through the glomerulus filter (which may filter out higher molecular weight DNA molecules). By urine cfDNA WGS of routine and abundant urine samples, we demonstrate noninvasive characterization of genome-wide copy-number profiles that carries important potential for complementing alternative (more targeted) urine-, blood-, and tissue-based molecular profiling strategies. Continued computational work will enable sensitivity and specificity optimization for our approach, and guide strategies for maximizing the utility of this approach for potential use in clinical oncology workflows. Overall, this work, paired with existing strategies for routine clinical biospecimen molecular profiling, carries significant potential for improving the temporal resolution of disease burden, circulating tumor clone representation, and treatment response in patients with advanced cancer.

Figure A1 – Urine cell-free DNA copy-number profiling recapitulates genome-wide copy-number profiles from patient-matched clinical karyotype and plasma cell-free DNA analyses, with evidence for enrichment of ultra-short tumor-specifc cfDNA fragments in urine

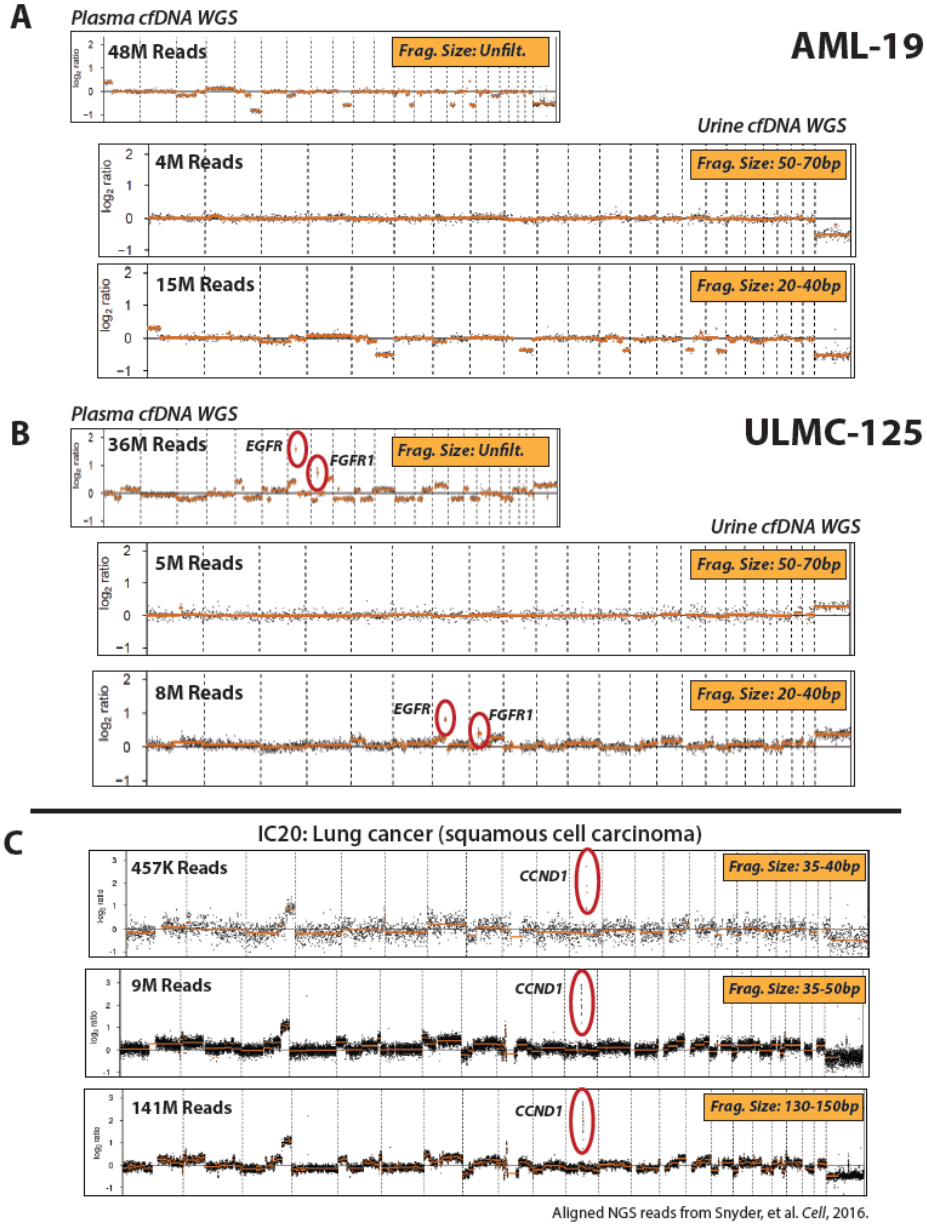


Figure A1 A. Genome-wide copy-number profiles derived from plasma and urine cell-free DNA sample WGS data in a patient with acute myeloid leukemia. Plasma cell-free DNA copy-number profile (top) represents known profile identified by classical karyotype analysis, with robust detection of arm- and chromosome-level copy-number events. Synchronous single-strand urine cfDNA libraries from the same patient show variable signal with restricting to specific urine cfDNA fragment length, suggesting urine tumor-specific cfDNA fragments may be enriched at ultra-short (<40bp) lengths. **B.** Similar, highly concordant results are seen for patient-matched plasma and urine cell-free DNA samples from a patient with lung adenocarcinoma, with similar enrichment of tumor-specific cfDNA fragments at 20-40bp in urine cfDNA WGS data enabling detection of both broad and focal copy-number alterations (here, *EGFR* and *FGFR1*) originally detected by plasma cfDNA WGS. **C.** Analysis of high depth plasma cfDNA WGS sequencing data from Snyder, et al. shows detectable tumor-specific signal at ultra-short fragment lengths in plasma, suggesting such fragments may be detectable in the urine after passage through the glomerulus filter.

Appendix A References

1. Wan, J.C., et al., *Liquid biopsies come of age: towards implementation of circulating tumour DNA.* Nat Rev Cancer, 2017. **17**(4): p. 223-238.
2. Siravegna, G., et al., *Integrating liquid biopsies into the management of cancer.* Nat Rev Clin Oncol, 2017. **14**(9): p. 531-548.
3. Fujii, T., et al., *Mutation-Enrichment Next-Generation Sequencing for Quantitative Detection of KRAS Mutations in Urine Cell-Free DNA from Patients with Advanced Cancers.* Clin Cancer Res, 2017. **23**(14): p. 3657-3666.
4. Hyman, D.M., et al., *Prospective blinded study of BRAFV600E mutation detection in cell-free DNA of patients with systemic histiocytic disorders.* Cancer Discov, 2015. **5**(1): p. 64-71.
5. Togneri, F.S., et al., *Genomic complexity of urothelial bladder cancer revealed in urinary cfDNA.* Eur J Hum Genet, 2016. **24**(8): p. 1167-74.
6. Scheinin, I., et al., *DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly.* Genome Res, 2014. **24**(12): p. 2022-32.
7. Ulz, P., et al., *Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer.* Nat Commun, 2016. **7**: p. 12008.
8. Snyder, M.W., et al., *Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin.* Cell, 2016. **164**(1-2): p. 57-68.
9. Oxnard, G.R., et al., *Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA.* Clin Cancer Res, 2014. **20**(6): p. 1698-705.
10. Mok, T., et al., *Detection and Dynamic Changes of EGFR Mutations from Circulating Tumor DNA as a Predictor of Survival Outcomes in NSCLC Patients Treated with First-line Intercalated Erlotinib and Chemotherapy.* Clin Cancer Res, 2015. **21**(14): p. 3196-203.
11. Carreira, S., et al., *Tumor clone dynamics in lethal prostate cancer.* Sci Transl Med, 2014. **6**(254): p. 254ra125.

APPENDIX B: Supplementary Materials for Chapter II

Supplementary Tables:

http://www.sciencedirect.com/science/article/pii/S1476558615000445?via%3Dihub#s0115

Supplementary Materials and Methods

Candidate genes with somatic driver mutations were derived from gain-of-function (GoF) and loss-of-function (LoF) analyses performed on 686,530 tumor samples with mutation data in Oncomine. GoF genes (oncogenes) were defined as those with a hotspot missense mutation (i.e. recurrent) rate >20% and deleterious mutation (i.e. nonsense and frameshift indels) rate <10%. Additionally, gene-level *p* values were estimated by the likelihood that a hotspot residue will have a given number of mutations by chance given the total number of mutations in that gene, with a false discovery rate (FDR) adjusted *p* value <0.1 required for classification as a GoF gene. LoF genes were defined as those with deleterious mutations in at least three samples and a combined deleterious and hotspot mutation frequency greater than 20%. Additionally, gene-level *p*-values were estimated, representing the significance of the proportion of deleterious mutations observed in each gene compared to all other genes, with a FDR adjusted *p*-value <0.1 required for classification as a LoF gene. Genes failing to meet GoF/LoF criteria were considered passenger genes. This approach was previously validated using a trained classifier in 2,711 TCGA profiled samples from 13 cancer types[1].

Candidate driver CNA events were identified by performing a minimal common region (MCR) analysis that identified regions of recurrent CNA (defined ≥ 3.7 copies) or deletion (≤~1 copy) in ≥ four samples (pan-cancer and in specific cancer type). Candidate regions were further filtered by imposing a requirement that at least one sample must have a copy number ≥ 8 for amplifications or ≤ 1 for deletions, and a further requirement that median event frequency was ≥0.5%. MCRs observed in different cancer types that shared common genes were identified. The most frequently amplified or deleted gene(s) within each set of overlapping MCRs was included in the candidate copy number gene list.

To identify additional fusions or novel 5'/3' fusion partners not present in the Mitelman database, we analyzed 6,438 primary tumor sample RNA-seq profiles contained within Oncomine using publicly available fusion prediction algorithms[2, 3]. This generated a large number of predicted fusions; we filtered out the following fusions to nominate driver candidates: fusions predicted in normal samples, those involving adjacent genes, homologous genes, or repetitive regions, and those involving transcriptional units in opposite orientations.

All candidate driver GoF, LoF and CNA genes, as well as gene fusions, were then assessed for evidence of near term potential clinical relevance. Genes (with or without a candidate variant) were considered for inclusion in OCP if they were 1) a target of FDA approved therapies, 2) associated with treatment recommendations from organizations such as the National Comprehensive Cancer Network (NCCN), 3) used as a biomarker for enrollment into ongoing clinical trials or 4) reported as associated with treatment response in clinical trials (or published case reports). Additional genes were considered for inclusion based on 1) membership in the Sanger Cancer Gene Census, 2) known cancer involvement or 3) were associated with investigational therapies.

Tissue Cohorts

The MO cohort consisted of all cancer specimens (including biopsy, resection and cell block specimens) sent during a five month period to the CLIA certified UM Molecular Oncology/Genetics Laboratory for 1) *EGFR*, *BRAF* or *KRAS* mutation testing or 2) *ALK* rearrangement testing. Testing for *EGFR* (exon 19 indels and residue 858 mutations by PCR based fragment analysis), *KRAS* (codon 12, 13 and 61 mutations by Sanger sequencing) and *BRAF* (codon 600 by allele-specific PCR or Sanger sequencing) was performed as described [4-6]. FISH for *ALK* rearrangement was performed using the FDA approved dual color break apart probe strategy (Abbott Molecular).

Only cases testing UM FFPE tissue blocks were considered for inclusion in the MO cohort. H&E slides and tissue blocks were reviewed after molecular testing to ensure sufficient material remained for OCP evaluation. A single FFPE sample was chosen if multiple blocks were tested. In cases where insufficient tissue remained in the block sent for molecular testing, concurrent blocks or blocks from prior diagnostic procedures were used. From 130 cases assessed during the above time period, 105 cases were from UM samples and had sufficient remaining tissue. Clinicopathologic information for all included cases is provided in **Table S4**.

Somatic variant identification

Variants were annotated using Annovar[7]. VCF-level filtering was applied to annotated variants to remove synonymous or non-coding variants, those with flow corrected read depths (FDP) less than 20, flow corrected variant allele containing reads (FAO) less than 6, variant allele frequencies (FAO/FDP) less than 0.10 in tumor suppressors or less than 0.05 in oncogenes,

extreme skewing of forward/reverse flow corrected reads calling the variant (FSAF/FSAR <0.2 or >5), or indels within homopolymer runs ≥4. Any variants called in >25% of all research samples sequenced herein or in other cohorts using any OCP version (*n*=776 total) were excluded as technical artifacts, unless occurring at known Oncomine-prioritized hotspot variants. Variants with allele frequencies >0.5% in ESP6500 or 1000 Genomes (from Annovar) or those reported in ESP6500 or 1000 Genomes with observed variant allele frequencies between 0.40 and 0.60 or >0.9 were considered germline variants.

Base-level filtering was then applied to candidate somatic variants passing the above criteria to exclude additional technical artifacts or poorly supported variants, including removal of variants located at the last mapped base (or outside) of amplicon target regions, variants with the majority of supporting reads harboring excess additional mismatches or indels (likely sequencing error), those in repeat-rich regions (likely mapping artifacts), and variants occurring exclusively in one amplicon if overlapping amplicons cover the variant. Variants passing these filters were visually confirmed in IGV. We have previously confirmed this filtering criteria identifies variants that pass Sanger sequencing validation with >95% accuracy[8, 9].

Copy number analysis

To identify CNAs, we utilized total amplicon read counts provided by the Coverage Analysis Plug-in. Read counts per amplicon for each sample (normalized to total number of reads for that sample) were divided by normalized counts from a composite normal male genomic DNA sample (comprised of multiple FFPE and frozen tissue, individual and pooled samples run on the same OCP version), yielding a copy number ratio for each amplicon. These

copy number ratios were then corrected for GC content, and gene-level copy number estimates were determined by taking the coverage-weighted mean of the GC-corrected per-probe ratios, with expected error determined by the probe-to-probe variance, as described[8-10]. Genes with a $\log_2$ copy number estimate of <-1 or >0.81 were considered to have high level loss or gain, respectively. As an estimate of data quality, we determined the standard deviation of the amplicon-level copy number estimates relative to the gene-level estimate for each gene per sample (**Fig B2**). Samples with median values >0.75 were deemed low quality and excluded from further analysis.

Gene fusion analysis

Within the Ion Reporter (4.2.0) Fusion analysis workflow, reads from the RNA AmpliSeq panel were aligned using TMAP to a gene reference of targeted chimeric fusion transcripts as well as reference sequences for expression imbalance and expression control gene targets. Read alignment required at least 70% overall homology to each side of the fusion breakpoint. Read counts were determined for expression control gene and expression imbalance targets; the exon imbalance metric for a given gene is calculated as the count of 3' target reads minus the count of 5' target reads divided by the sum of the expression control gene target read counts.

In the MO and LU cohorts, individual absolute fusion isoform read counts <200 and non-prioritized gene fusions were excluded. In the PR cohort, individual absolute fusion isoform read counts <30 were excluded. Individual isoform (i.e. *TMPRSS2:ERG* fusions involving *TMPRSS2* exon 1 fused to *ERG* exon 4 [T1E4]) and gene level (all *TMPRSS2:ERG* isoforms) were summed

and normalized to the summed read count of the five housekeeping genes. For visualization, the $\log_2$ [(normalized read counts)*100,000] was used.

qRT-PCR

qRT-PCR was performed to confirm the expression of *ERC1:BRAF* in MO-17 and *TPR:NTRK1* in MO-35 as detected by OCP. Primers and probes (5' FAM; ZEN/Iowa Black FQ dual quenchers) were designed using PrimerQuest (www.idtdna.com/Primerquest/Home/Index, hg 19 genome assembly) and obtained from IDT. Primer/probes sequences are given in **Table S16**. Reverse transcription (RT) of 1µg RNA was performed using Omniscript RT (Qiagen) in the presence of RNAse Inhibitor (Qiagen) and gene specific priming using a pool of the 5 reverse primers used in qPCR (50nM final concentration of each primer) at 37 C for 1 hour. qPCR reactions (15ul) were performed in triplicate using TaqMan Universal Master Mix II (Applied Biosystems), 50ng cDNA equivalent per reaction and a final concentration of 0.9uM each primer and 0.25uM probe in 384 well plates on the QuantStudio 12K Flex (Applied Biosystems). Baseline and C_t thresholds were set using QuantStudio 12K Flex Real-Time PCR System Software. All C_t threshold values >40 were set to 40. Log_2 expression of *TPR:NTRK1*(T21N10), *ERC1:BRAF*(E12B9) and *ERC1:BRAF*(E12B10) were determined by the $\Delta\Delta^{Ct}$ method using the *GAPDH* and *HMBS* C_t geometric mean as the reference and the average of the 5 assessed MO samples without gene fusion detection by OCP as the calibrator. A no template control (water subjected to RT as above) was processed in parallel.

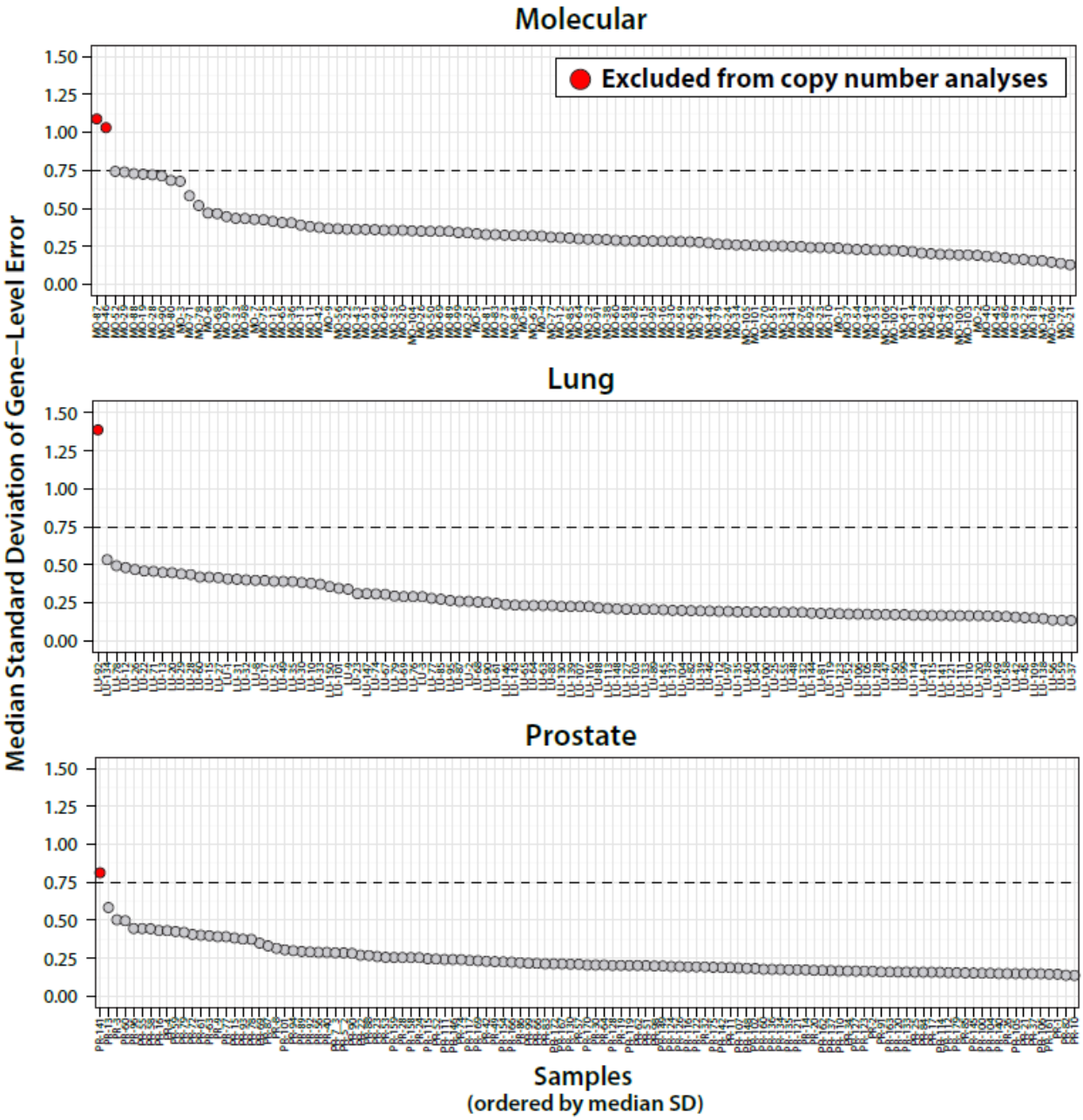
ERBB2 immunohistochemistry (IHC)

IHC for ERBB2 was performed using the Ventana Benchmark System (Ventana Medical Systems; Tucson, Arizona) on 4-5µm thick FFPE tissue sections in the University Of Michigan Department of Pathology Clinical IHC Laboratory using pre-dilute mouse anti-ERBB2 monoclonal antibody (clone 4B5).

Comprehensive Cancer Panel profiling

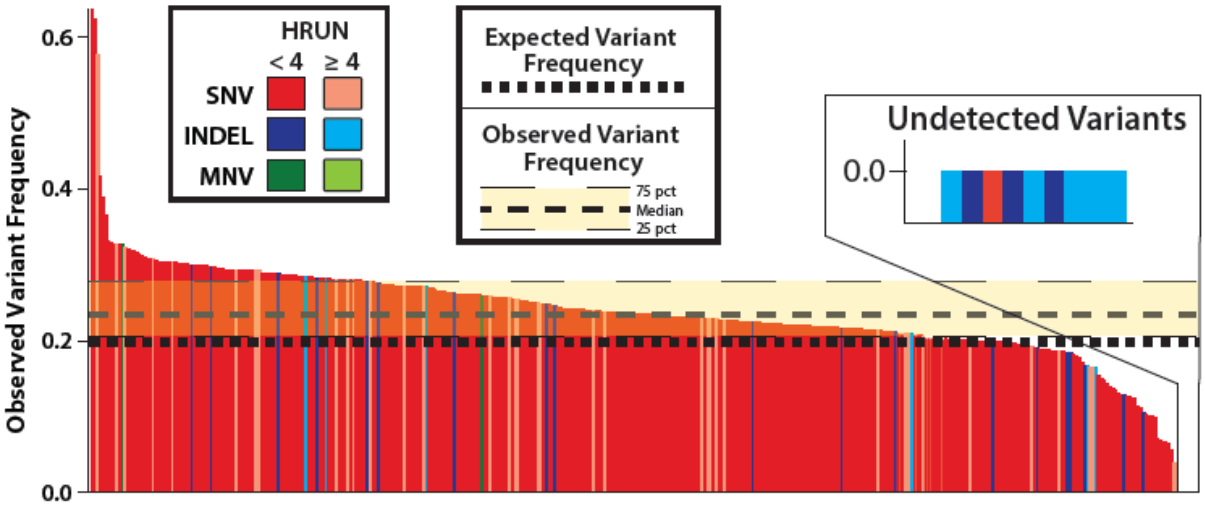
PR-185 and PR-186 (FFPE prostate cancer specimens) were profiled using the Ion Torrent Comprehensive Cancer Panel (CCP) as described[9].

Figure B1. Assessment of OCP copy number alteration (CNA) profiling data noise.



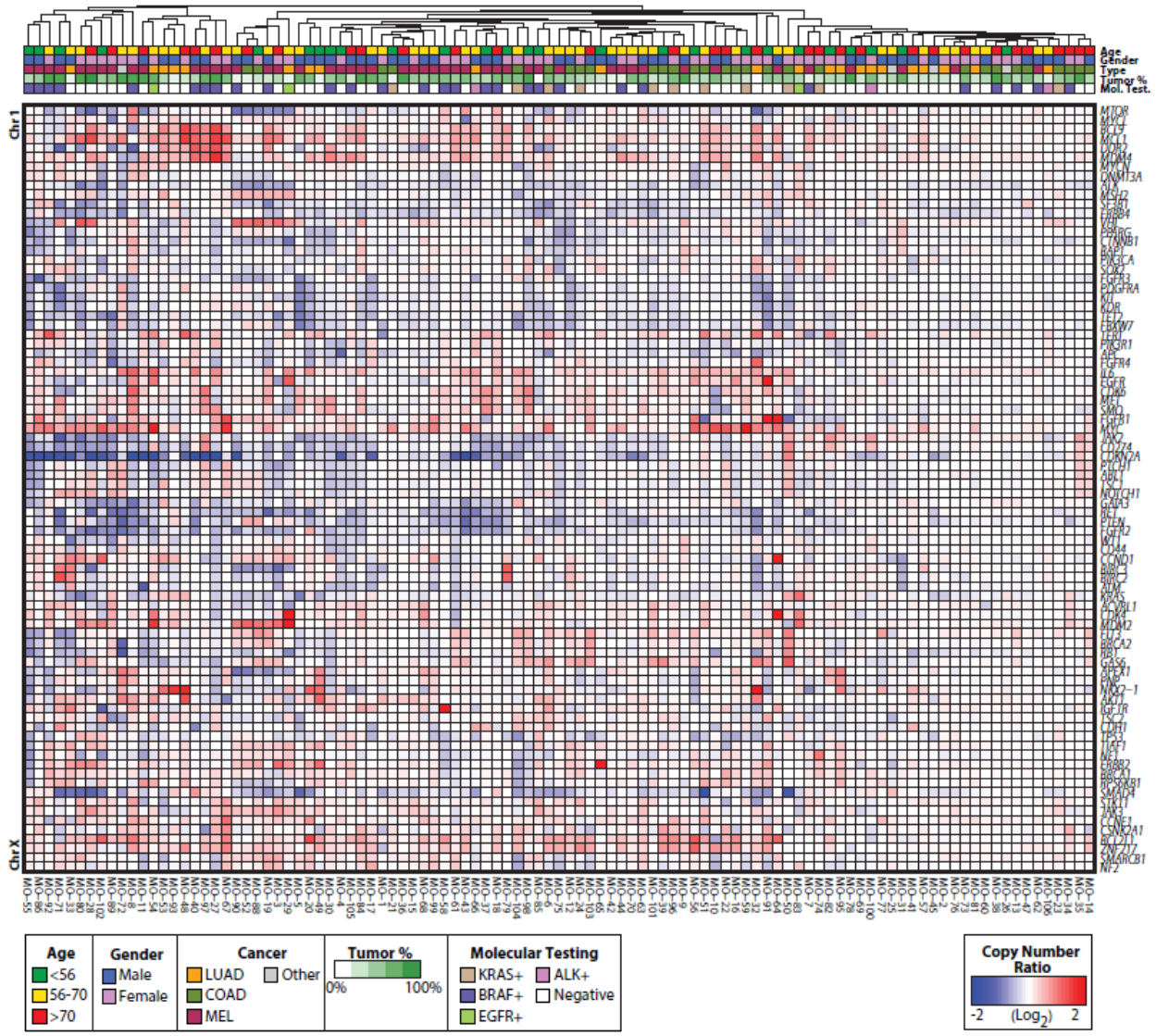
For each sample across the MO, LU and PR cohorts, the amount of noise in the copy number profiling data was assessed by determining the standard deviation of the target-level copy number estimates relative to the gene-level estimate for each gene in the sample. Values for all genes are plotted per sample, and samples with median values >0.75 (shown in red) were deemed low quality and excluded from CNA analysis.

Figure B2. Assessment of AcroMetrix Oncology Hotspot Control (AOHC) panel.



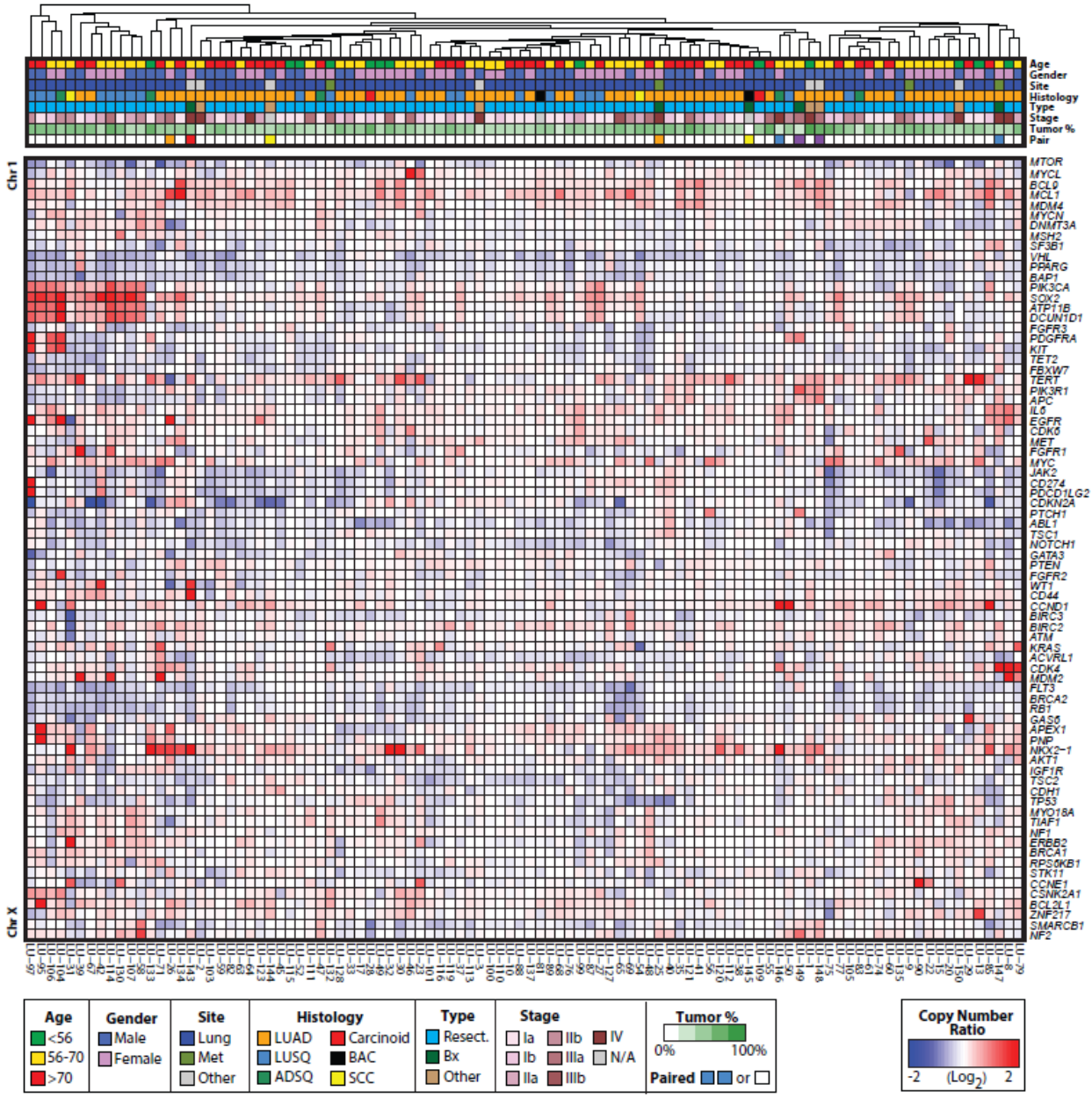
For each single nucleotide variant (SNP), short insertion/deletion (INDEL), and multi-nucleotide polymorphism (MNP) present in AOHC and targeted by OCP (n=398 total variants), the observed variant allele frequency is plotted. Each bar corresponds to a single variant, and variants are sorted in order of descending observed variant frequency, with the expected variant allele frequency (0.20 for all alleles) indicated. Bars are colored by variant type and homopolymer context (< or ≥ 4bp in length). Variants undetected by automated variant calling are indicated in inset. The median and interquartile range for observed variant frequency is indicated as in the legend.

Figure B3. Copy number profiles from the molecular diagnostics (MO) cohort.



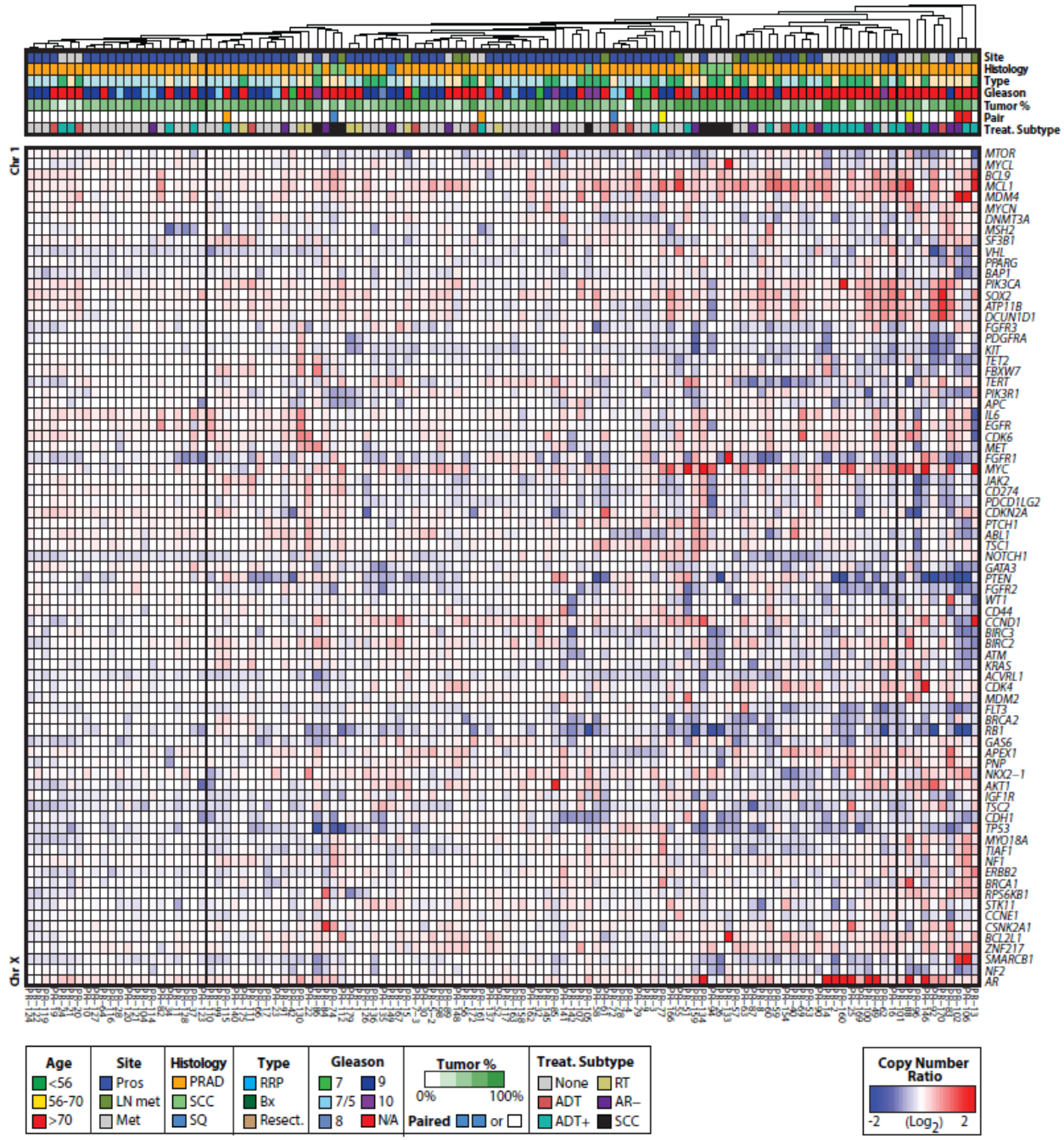
Unsupervised hierarchical clustering of copy number profiles from MO samples. Copy number ratios (log_2) for genes targeted by OCP are shown according to the color scale. Genes are arranged in genome order (from top to bottom). Pathological information is given in the header according to the legend.

Figure B4. Copy number profiles from the lung cancer (LU) cohort.



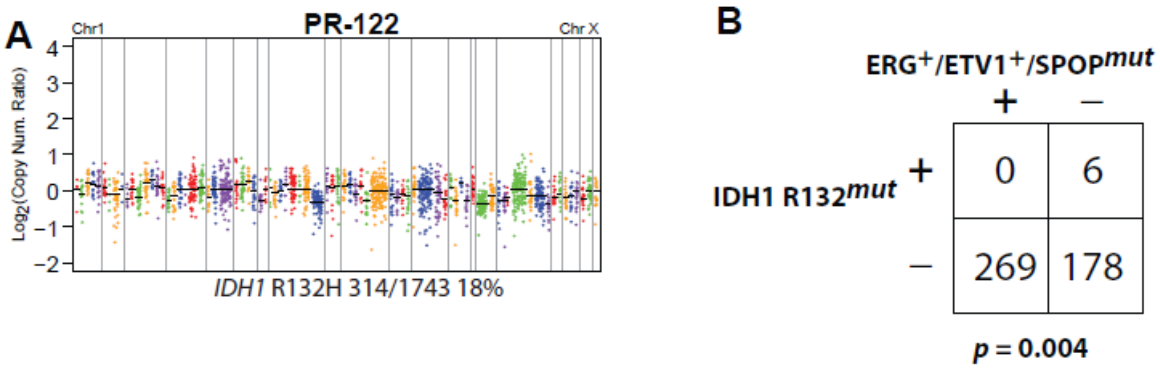
Unsupervised hierarchical clustering of copy number profiles from LU samples. Copy number ratios (log_2) for genes targeted by OCP are shown according to the color scale. Genes are arranged in genome order (from top to bottom). Pathological information is given in the header according to the legend.

Figure B5. Copy number profiles from the prostate cancer (PR) cohort.



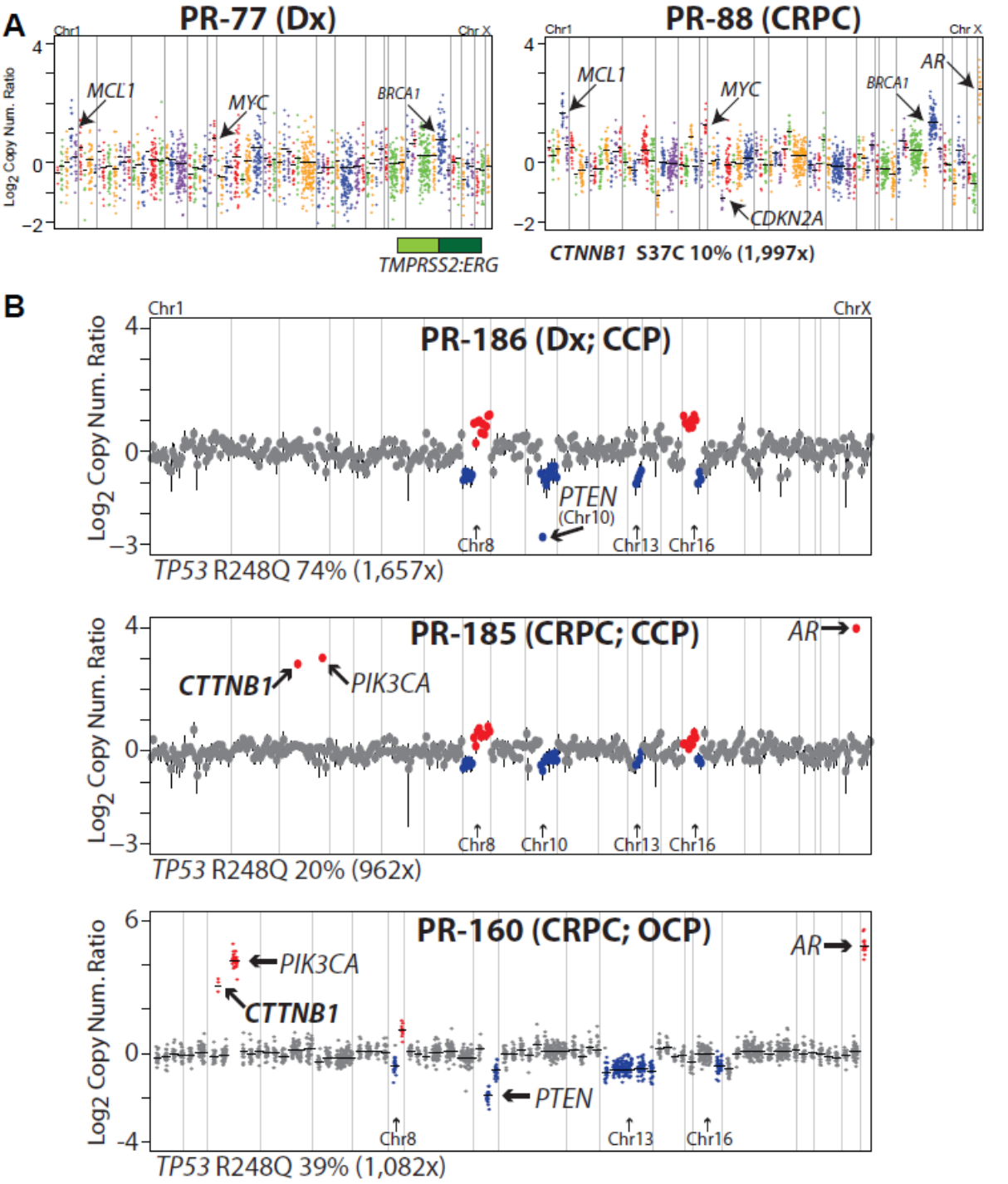
Unsupervised hierarchical clustering of copy number profiles from PR samples. Copy number ratios (log_2) for genes targeted by OCP are shown according to the color scale. Genes are arranged in genome order (from top to bottom). Pathological information is given in the header according to the legend.

Figure B6. OCP as a translational research tool identifies *IDH1* R132 mutations as defining a rare subtype of ETS^{-} prostate cancer.



A. An ETS fusion negative prostate cancer (PR-122) without other OCP defined actionable alterations harbored an *IDH1* R132H mutation. **B**. Distribution of *IDH1* R132 mutations and combined *ERG* fusions, *ETV1* fusions and *SPOP* mutations from 453 publically available sequenced prostate cancers (see **Table S13**). Two sided Fisher's exact test significance is given.

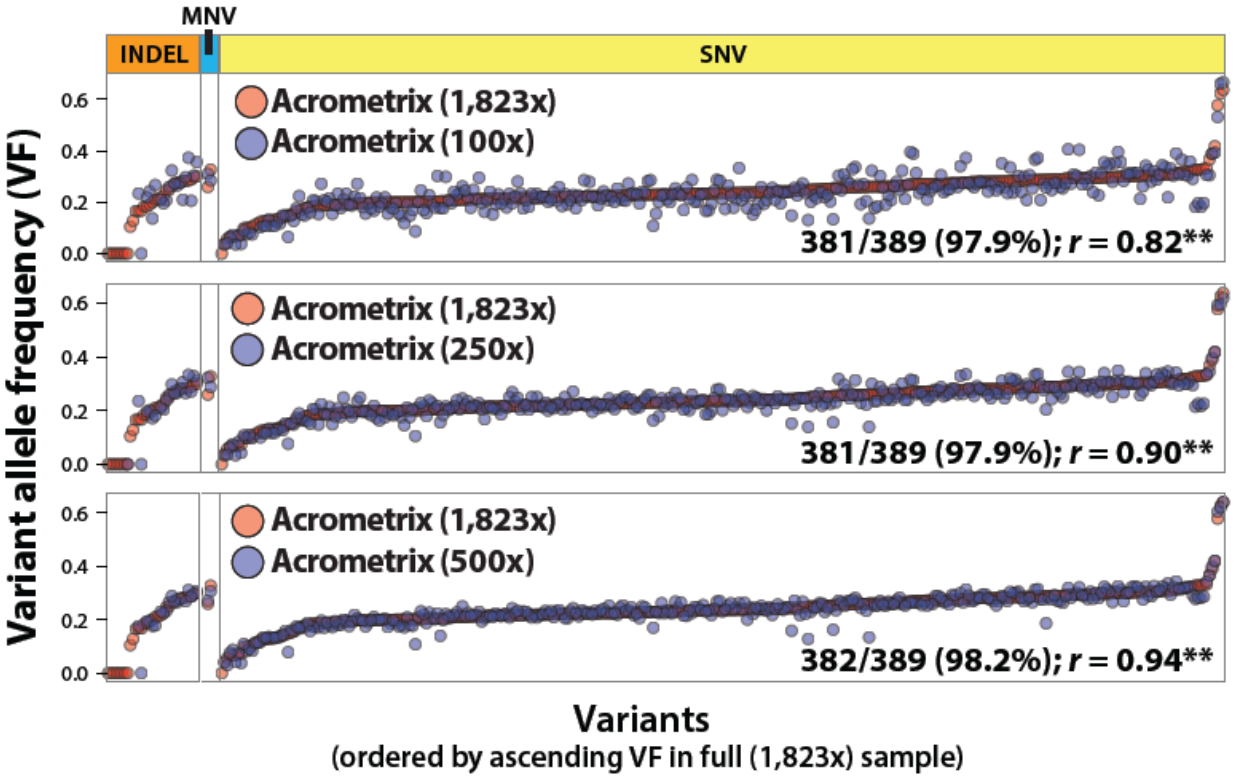
Figure B7. OCP profiling of paired pre-/post-therapy prostate cancer specimens identifies *CTNNB1* amplification/ mutation as an adaptive (or selected) response to ADT and/or chemotherapy.



A. OCP profiling of pre- and post-treatment prostate cancer supports activating *CTNNB1* mutation as an adaptive response. PR-77 is an untreated diagnostic (dx) primary Gleason score 9 prostate cancer and PR-88 is a subsequent castration resistant prostate cancer (CRPC) bladder metastasis obtained after ADT, XRT and chemotherapy that had AR$^-$ phenotype. OCP profiling demonstrates shared high level *MCL1* and *MYC* CNAs (and non-prioritized high level *BRCA1* amplification), consistent with clonality; however a *TMPRSS2:ERG* fusion (exons T2E2) was only identified by the OCP RNA-seq panel in PR-77, consistent with the AR$^-$ phenotype in PR-88. PR-88 uniquely harbored *AR* amplification (a known ADT resistance mechanism) and

CDKN2A deletion, as well as a *CTNNB1* S37C (variant allele frequency 10%). No read support for *CTNNB1* S37C was present in PR-77 (>5,000 reads). **B**. Using the Ion Torrent Comprehensive Cancer Panel (CCP), which targets all coding exons of 409 cancer related genes, we profiled the diagnostic prostate biopsy tissue (PR-185, top) from a 49 year old man presenting with M1 (lymph node and liver metastases) prostate cancer. After rapidly developing CRPC after ADT and chemotherapy, liver biopsy of a metastasis (PR-185, middle) and an epidural metastasis resection specimen (PR-160, bottom) were obtained. PR-185 was profiled on the CCP and PR-185 was profiled using the OCP. All three tumors were gene fusion negative by the RNA component of the OCP. Integrative profiles for each tumor are shown as in **A**, except for CCP copy number plots, gene level copy number ratios are plotted as points with 95% confidence intervals indicated. Shared *TP53* R248 mutations and broad low level CNAs (shown in red and blue points/amplicons, including 1 or 2 copy *PTEN loss*) were present in each sample,, consistent with clonal progression. High level, focal *AR*, *PIK3CA* and *CTNNB1* amplifications were present in both CRPC specimens but not the pretreatment sample, consistent with adaptive (or selected) alterations in response to therapy.

Figure B8. Comparison of variant detection in complete and downsampled sequencing data using the Acrometrix Oncology Hotspot Control (AOHC) molecular standard.



Variant allele frequencies (VFs**)** independently derived from complete and downsampled sequencing data across a set of 389 known indels, multi nucleotide variants (MNV) and single nucleotide variants (SNV) called in the Acrometrix Oncology Hotspot Control (AOHC) sample. Original VFs calculated by TVC (in orange) utilized the complete set of mapped reads for the AOHC sample (average per-base coverage across OCP targeted regions: 1,823x). Random downsampling of original sequencing data enabled concordance analyses at 100x, 250x, and 500x effective average coverage across OCP targeted regions, with VFs for all variants plotted in blue. Percentages indicate proportion of original variant calls that were also made from each donwsampled dataset. Pearson correlation coefficients (*r*) between complete and downsampled VFs are provided; **=p <0.001.

Appendix B References

1. Tomlins, S.A., et al., *Analysis of 2,700 Cancer Exomes to Identify Novel Cancer Drivers and Therapeutic Opportunities.* European Journal of Cancer, 2012. **48**: p. 134-134.
2. Kim, D. and S.L. Salzberg, *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome Biol, 2011. **12**(8): p. R72.
3. McPherson, A., et al., *deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data.* PLoS Comput Biol, 2011. **7**(5): p. e1001138.
4. Betz, B.L., et al., *The application of molecular diagnostic studies interrogating EGFR and KRAS mutations to stained cytologic smears of lung carcinoma.* Am J Clin Pathol, 2011. **136**(4): p. 564-71.
5. Bernacki, K.D., et al., *Molecular diagnostics of melanoma fine-needle aspirates: a cytology-histology correlation study.* Am J Clin Pathol, 2012. **138**(5): p. 670-7.
6. Hookim, K., et al., *Application of immunocytochemistry and BRAF mutational analysis to direct smears of metastatic melanoma.* Cancer Cytopathol, 2012. **120**(1): p. 52-61.
7. Chang, X. and K. Wang, *wANNOVAR: annotating genetic variants for personal genomes via the web.* J Med Genet, 2012. **49**(7): p. 433-6.
8. Cani, A.K., et al., *Next-Gen Sequencing Exposes Frequent MED12 Mutations and Actionable Therapeutic Targets in Phyllodes Tumors.* Mol Cancer Res, 2015.
9. Warrick, J.I., et al., *Tumor evolution and progression in multifocal and paired non-invasive/invasive urothelial carcinoma.* Virchows Arch, 2015. **466**(3): p. 297-311.
10. Grasso, C., et al., *Assessing copy number alterations in targeted, amplicon-based next-generation sequencing data.* J Mol Diagn, 2015. **17**(1): p. 53-63.

APPENDIX C: Supplementary Materials for Chapter III

SUPPLEMENTARY METHODS

TCGA Data Analysis

TCGA pan-cancer copy number analyses were run on somatic (scna_minus_germline_cnv_hg19__seg) segmented Affymetrix SNP6 array-based copy-number calls for 11,576 tumor samples across 32 tumor types contained in the most recent (01/28/2016) TCGA GDAC Firehose standard data run (stddata__2016_01_28)[1]. Data was downloaded from the TCGA GDAC Firehose repository using the firehose_get utility (v0.4.6), and the fraction of genome altered (FGA) was calculated as in cBioPortal (https://groups.google.com/forum/#!topic/cbioportal/HKLa9C9m4y4). Specifically, FGA was calculated for all tumor samples as the total number of bases in regions affected by copy-number alterations with log2(CopyNumberRatio) > 0.2 or < -0.2 divided by 3 billion (the approximate median number of bases in all segments for each sample across all analyzed samples and tumor types).

Cell-free DNA extraction

Five milliliters of peripheral blood were collected for 92 samples from 76 patients with metastatic castration resistant prostate cancer (mCRPC) and 10 healthy controls (5 male, 5

female) using K2 EDTA blood collection tubes (Cat: 366643, BD, NJ) (**Table S1**). Within 4 hr, blood was mixed with equal volume of PBS and Ficoll-Paque Plus (Sigma-Aldrich; MO) was used to separate plasma from red blood cells and peripheral mononuclear cells (PBMC). Plasma was centrifuged twice at 1500 g for 12 min to limit cell contamination and stored in -80° C.

For 11 patients (13 samples) with metastatic lung adenocarcinoma, 4 patients (7 samples) with metastatic colorectal cancer, 3 patients (3 samples) with leukemias, and 2 patients (4 samples) with sarcoma, one patient with both sarcoma and breast cancer, and a patient with uterine leiomyosarcoma, 10 mL peripheral blood was collected using Streck Cell-Free DNA BCT tube (Streck; NE) (**Table S1**). Within 4 hr, blood was centrifuged at 1600 g for 10 min, and then plasma was centrifuged at 1600 g for 10 min to remove cell debris and stored in -80° C.Cell free DNA was extracted from all plasma (2 mL) samples with QIAamp Circulating Nucleic Acid Kit (Qiagen; CA) according to the manufacturer's instructions. Sample collection and NGS was performed with Institutional Review Board approval.

Low-pass whole-genome sequencing and copy-number detection

Sequencing alignment and coverage analyses were performed using Torrent Suite version 5.0.2 (Ion Torrent, Carlsbad, CA). Initially, reads were aligned to the hg19 version of the human reference genome using tmap (v5.0.7) and aligned, non-PCR-duplicate reads (samtools v1.3) were used as input for our copy-number calling workflow. Genome-wide copy number alterations were first called using the QDNASeq R package (version 1.6.1) [2]. Briefly, the genome was divided up into variable bin sizes (15, 25, 50, 100, 500, and 1,000 kilobase-pair

bins), and bin-level counts of high-quality mapped reads (MAPQ ≥ 37) were calculated separately for each sample. Raw bin-level counts were simultaneously corrected for GC content and mappability by fitting a LOESS surface through median read counts for bins with the same combination of GC content and mappability and dividing raw bin-level counts by the corresponding LOESS fitted value. GC- and mappability-corrected bin-level counts were then normalized by median bin-level corrected counts within each sample. Bins previously shown using either ENCODE or 1000G data to yield anomalous copy-number results due to germline copy number variants (CNVs), low mappability, or large stretches of uncharacterized nucleotides were excluded [2]. For each bin in each tumor sample, high-quality, corrected, median-normalized read counts were divided by average corrected, median-normalized read counts from our 5 normal male samples. Segmented copy-number events were called from bin-level corrected, median- and control-normalized read counts using the circular binary segmentation algorithm implemented by the DNACopy (1.44.0) R package, and final segment- and bin-level copy-number values were used for subsequent analyses as described. Focal CNAs were defined as CNAs 1.5-20Mb long with a log2(CNRatio) ≥ 0.2, thresholds similar to those described elsewhere [3].

Targeted sequencing: Oncomine Comprehensive Assay

For 60 patient cfDNA samples (31 high tumor content mCRPC samples, 13 low tumor content mCRPC samples, 11 high tumor content non-mCRPC samples, 1 mCRPC sample with germline chr20 deletion, and 4 male normals; see **Table S1**) and both sheared UMUC-5 and VCaP gDNA samples, we performed targeted NGS using the DNA component of the Oncomine

Comprehensive Assay (OCP), a custom multiplexed PCR-based panel of 2,530 amplicons targeting 126 genes. These genes were selected based on pan-cancer analysis that prioritized somatic, recurrently altered oncogenes, tumors suppressors and genes subject to high level copy alterations, combined with a comprehensive analysis of known/investigational therapeutic targets[4]. Barcoded libraries were generated from 1-20ng of cfDNA per sample and multiplexed sequencing was performed using the Ion Torrent Proton sequencer. Library preparation with barcode incorporation, template preparation on the OneTouch 2 and sequencing using the Ion Torrent Proton sequencer (Ion Torrent, Carlsbad, CA) were performed according to the manufacturer's instructions. Data analysis was performed using Torrent Suite 5.0.2, with alignment by TMAP using default parameters, and variant calling using the Torrent Variant Caller plugin (5.0.2.1) using default low-stringency somatic variant settings. Variant annotation filtering and prioritization, along with gene-level copy number estimation, were performed essentially as described [4-7] using validated in house pipelines, and gene level copy-number calls, and prioritized point mutations, small insertions/deletions (indels), and copy-number variants were reported for each patient sample (**Table S2 & S3**). Copy number alterations called from targeted NGS data with log2(copy number ratio) >= 0.6 or <= -1.0 were prioritized.

VCaP and UMUC-5 In silico Dilution

To establish theoretical segment-level copy-number distributions for tumor content estimation, we carried out serial *in silico* dilution experiment by mixing read proportions derived from undiluted VCaP and UMUC-5 whole-genome sequencing data and our set of normal male patient samples. Briefly, we combined FASTQ files for the whole-genome sequencing

experiments from our 5 normal male samples (n=85,981,532 total unaligned reads). We then shuffled reads (*"awk '{OFS=\"\\t\"; getline seq; getline sep; getline qual; print \$\$0,seq,sep,qual}' <norm_fastq_file> | shuf | awk '{OFS=\"\\n\"; print \$\$1,\$\$2,\$\$3,\$\$4}'*), and randomly sampled an identical number of non-PCR-duplicated reads as was present for the VCaP (n=6,670,015 reads; whole-genome coverage= 0.26x) and UMUC-5 (n=16,570,486 reads; whole-genome coverage = 0.74x) undiluted whole-genome sequencing samples.

In silico dilutions were subsequently carried out on both undiluted whole-genome sequencing cell line samples with our coverage-matched normal male sample (for all integer percent dilutions 0-100%), where for each dilution the following steps were executed:

1) Shuffle undiluted cell line & normal male FASTQ files (using code above)

2) Sample appropriate portion of reads from each file using seqtk NGS toolkit (v1.0-r31) (*seqtk sample –s100 <FASTQ file><proportion_to_sample>*)

3) Concatenate proportional FASTQ files (*cat <vcap_prop_file><normal_prop_file>*)

4) Map mixed read set to the reference genome (hg19) using identical mapping approach to that used for original undiluted cell line and patient whole-genome sequencing samples:

tmap mapall –f hg19.fasta –r input.fastq –s output.bam –v –Y –u –prefix-exclude 5 –o 1 stage1 map 4

5) Sort and index aligned bam files for input to copy-number calling workflow

Genome-wide copy number variation calls were subsequently generated for each *in silico* dilution as described (see Methods).

Clustering

Mean-shift, k-means, and xmeans clustering approaches were assessed and deployed to identify relevant clusters from segment- (whole-genome sequencing) or gene-level (targeted sequencing) copy number ratio data. All clustering analyses were carried out in R (3.2.3) using packages LPCM (v0.45-0), RWeka (0.4-26), or base packages as applicable. For mean-shift clustering, variable bandwidths were evaluated, supporting a static bandwidth value of 0.01 on exome or whole-genome copy-number calls. Mean-shift clustering showed the most consistent expected cluster identification across *in vitro*/*in silico* dilutions, and was used for all analyses described herein.

Tumor Content Estimation

For whole-genome sequencing samples, reference segment-level copy-number ratio distributions were established through serial *in vitro* and *in silico* VCaP and bladder (UMUC-5) cell line dilutions as described. A heuristic least squares based distance metric (LSS) was used to approximate tumor content from whole-genome copy-number data. LSS between cluster centroids was calculated as a proxy for tumor content using the following formula:

$$\sum_{i=2}^{n} \sqrt{a[i]\text{^}2 - a[i-1]\text{^}2}$$

where a is the vector of cluster centroids for clusters identified by the mean-shift algorithm, n is the length of the cluster vector, and i is ith element of this vector. If only one cluster was assigned for a given sample, LSS was calculated as the square root of the cluster center squared (equivalently, the absolute value of the cluster centroid):

$$\sqrt{cluster\_center\text{^}2} = |cluster\_center|$$

Reference LSS distributions were established across serial *in silico* dilution experiments at all integer percent dilutions 0-100% as described, and these distributions were used to guide tumor content estimation for patient samples. While tumor content estimates were not generated for samples with LSS values < 0.1, these samples were specifically scanned for focal CNAs, as described above.

In silico Experiments: Downsampling

For the VCaP and UMUC-5 *in silico* dilutions, as well as 9 patient cfDNA samples (5 w/highest tumor content, 1 germline chr21 deletion, 2 no tumor content), we carried out *in silico* downsampling experiments to evaluate capacity to call copy-number alterations at variable effective whole-genome coverages (range: 0.005–0.1x). After downsampling (using *samtools view –s <proportion of reads to sample> –bh <original.bam.file>*) for each sample, copy-number alterations were called across variable bin sizes as described. Given the effective coverages analyzed, bin sizes were not analyzable across all coverages (e.g., 0.01x whole-genome coverage corresponds to approximately 150k single-end reads, leaving <10 reads per

100kb bin, on average). For this reason, we considered effective coverage & bin combinations ≥30 reads per bin as analyzable for this analysis.

Serial *in silico* downsampling experiments were also carried out on targeted sequencing data from 10 mCRPC patient plasma cfDNA samples (5 high tumor content, 1 germline chr20 deletion, and 3 normals) to 500, 250, 100, 50, and 25x effective target coverage by the same sampling approach taken with whole-genome data.

VCaP cfDNA WGS vs COSMIC array-based CN calls

Of 500 segment-level copy-number calls for chromosomes 1-22 & X reported as present in VCaP by COSMIC, 464 (92.8%) overlapped ≥90% of at least one 15kbp bin from our low pass (0.26x whole-genome coverage) analysis of undiluted VCaP, with 496 (99.2%) showing at least some (≥1 bp) overlap of one bin or more. We calculated median of bin-level integer copy number values for all 15kbp bins overlapped at ≥90% by a COSMIC-reported copy-number segment, and compared these low-pass sequencing derived values to segment-level integer copy-number values reported in COSMIC. Given the known variability in reported copy-number estimates for VCaP focal AR amplification (copy number of 14 reported by COSMIC; at least 3-18 copies by FISH [8]), we explored correlations between COSMIC segmented copy-number and both raw and capped (copy-number = 14) sequencing-derived copy-number values.

UMUC-5 cfDNA WGS vs Targeted NGS CN Calls

Copy number calls from whole genome sequencing of sheared UMUC-5 genomic DNA (gDNA) were compared to calls derived from targeted sequencing (OCP) of sheared DNA in this study. Of 126 genes targeted on the OCP, 90 had more than 3 amplicons and amplicon-level estimate variability sufficient for gene-level copy-number analysis. Coding sequence for 87/90 genes (97%) overlapped at least one 15kbp bin-level call from whole-genome sequencing data of sheared gDNA. Gene-level copy number estimates from whole-genome sequencing data were calculated as mean log2 copy number ratio for 15kb bins overlapping genome space from first to last coding base pair for each of the 87 genes.

Application to exome sequencing segmented copy-number calls

In order to test the efficacy of this particular approach for approximating tumor content from alternate datasets, we tested our LSS approach on segmented-copy-number calls from 129 clinical advanced/treatment refractory cancer tissue samples subjected to exome sequencing as part of the MI-ONCOSEQ project at the University of Michigan [9, 10]. Tumor content for all MI-ONCOSEQ samples is estimated through a model fitting variant allele frequencies of all somatic mutations and a model assessing zygosity shift of heterozygous SNPs and local copy number [9, 10]. As our analysis of TCGA copy-number data, the fraction of genome altered (FGA) was calculated for each MI-ONCOSEQ sample as the total number of bases in regions affected by segmented copy-number alterations with log2(CopyNumberRatio) > 0.2 or < -0.2 divided by 3 billion (the approximate median number of analyzable bases across all analyzed samples).

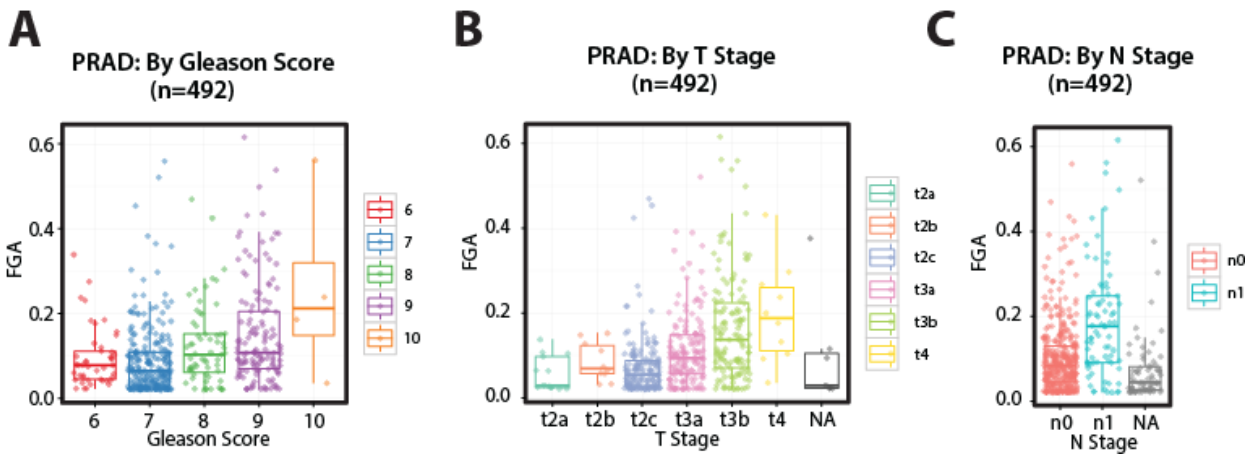
Concordance with tissue-based whole-exome sequencing copy-number profiles

Segmented log2 copy number ratio data from whole-exome sequencing of fresh frozen tissue specimens[9, 11] was available for 23 of 27 patients also profiled by cfDNA low-pass WGS. Each of these 23 patients had at least 1 cfDNA plasma sample (range: 1-3), and 18 of 23 (78.3%) had at least 1 cfDNA sample with elevated tumor content (LSS ≥ 0.1) suitable for concordance analyses. For these 18, the median of cfDNA low-pass WGS bin-level copy number values for all 500kbp bins overlapped at ≥90% by a tissue-based copy-number segment was calculated as a pseudo-cfDNA segment call, and correlations between tissue- and cfDNA-based copy number ratios were evaluated.

Focal AR amplification determination

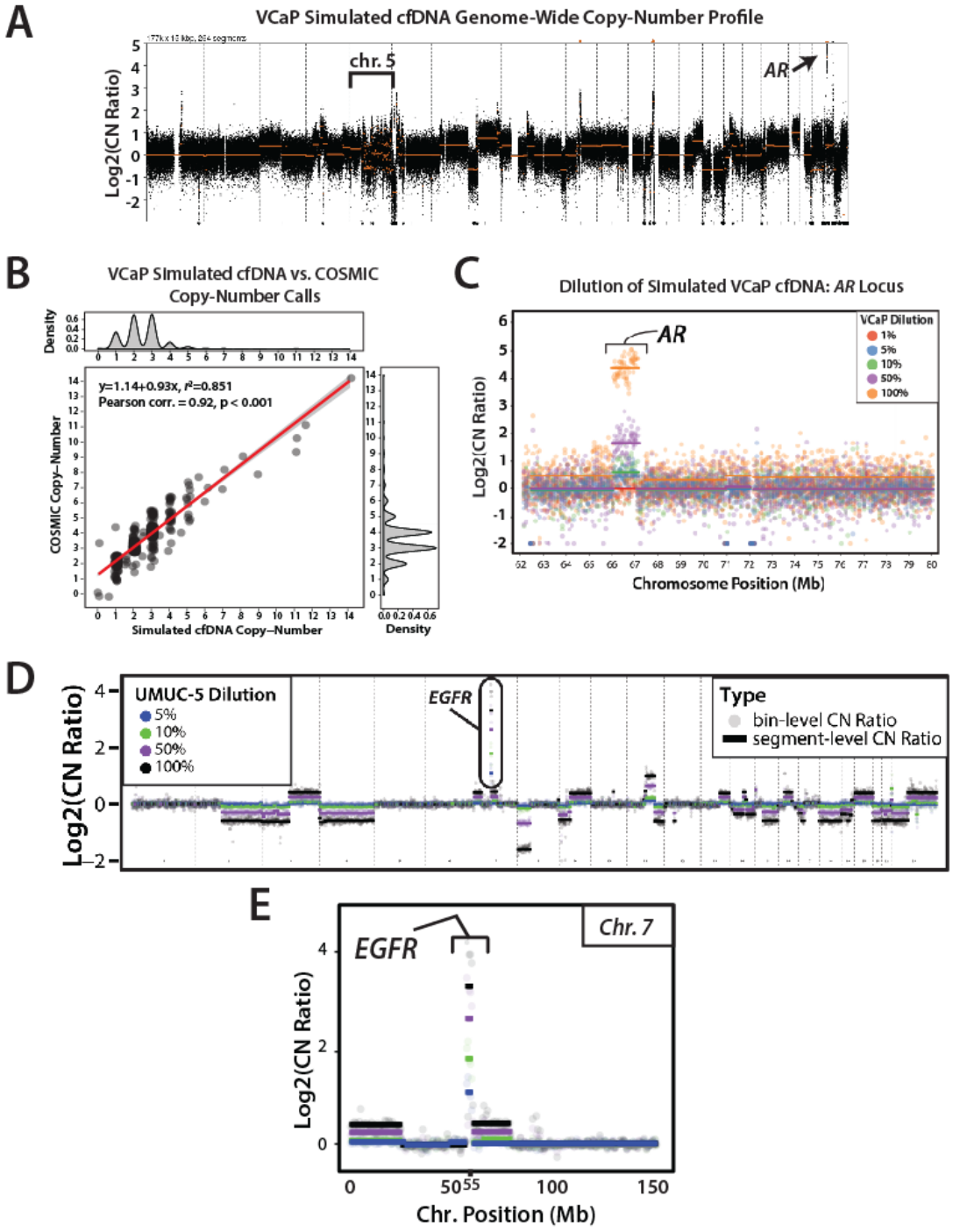
Given the difficulty of appropriate copy number segmentation on chrX, median 100kb bin-level copy number estimates across chrX q-arm were subtracted from mean 100kb bin-level copy-number estimates at *AR* locus (chrX:66.0-67.5Mb), and difference values >= 0.2 were used to call focal *AR* amplifications in our mCRPC cohort. Two cfDNA high tumor content samples (TP1216 and TP1295) met the above criteria, but were excluded as potential false positives due to use of 100kb bin width at low coverage (<300,000 total high-quality (MAPQ >= 37) mapped reads). An additional low tumor content sample (TP1139) met the amplification criteria, but with excessive variability in chrX bin-level copy-number estimates, was considered negative for *AR* amplification for all subsequent analyses.

<u>Figure C1:</u> Fraction of genome altered (FGA) analysis by stage/grade in TCGA prostate adenocarcinoma (PRAD) samples.



Fraction of genome altered (FGA) analysis was carried out on 492 PRAD samples using segmented Affymetrix SNP6 array-based segmented calls extracted from the most recent standard analysis set generated by GDAC Firehose (stddata__2016_01_28). FGA was calculated for all PRAD tumor samples as the total number of bases in regions affected by copy-number alterations with log (base 2) copy number ratio (Log2CN) > 0.2 or < -0.2 divided by 3 billion (the approximate median number of bases in all segments for each sample across all analyzed TCGA samples and tumor types). **A**. PRAD cohort FGA proportions are stratified by Gleason score, showing an increase in FGA as Gleason score increases. **B & C**. PRAD cohort FGA proportions are stratified by tumor stage (**B**; T Stage) and clinical stage (**C**; N Stage), showing increased FGA in high stage and N stage disease as well.

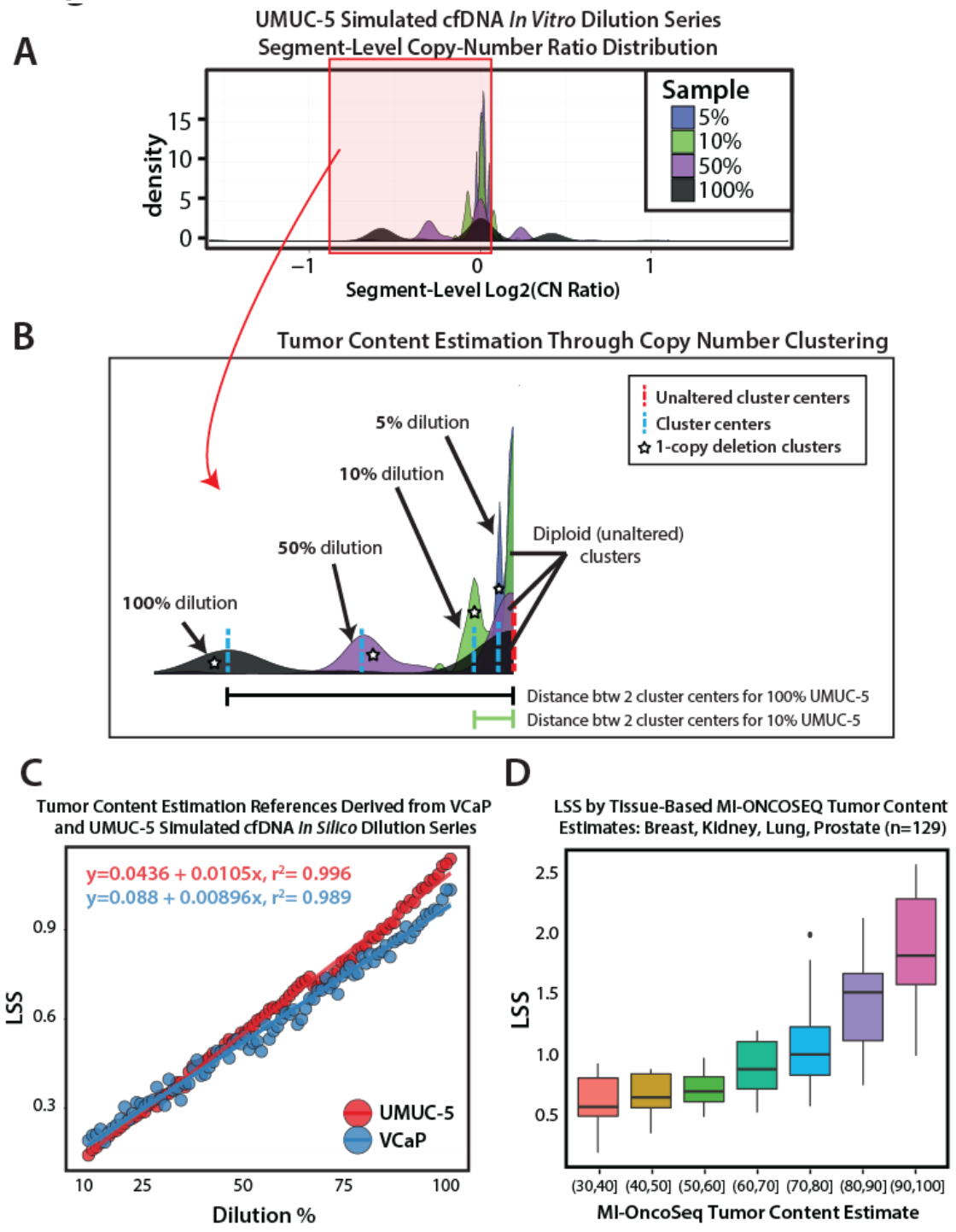
Figure C2: Robust copy number alteration (CNA) detection by low-pass whole genome sequencing (WGS) of artificial cfDNA on bench top sequencers.



A. Low-pass WGS generated genome wide copy number profile of the VCaP prostate cancer cell line using sheared genomic DNA to simulate cfDNA. The known focal *AR* amplification on chr X and the chromothriptic event on chr 5 are indicated. Bin-level estimates are plotted as black dots, and segmented copy-number calls are plotted as orange lines. **B**. Correlation of integer copy number values from low-pass WGS artificial cfDNA to reported VCaP copy number values in the COSMIC database. *AR* copy number for unamplified VCaP was capped at 14 given variability in reported copy number (see **Methods**). Pearson correlation and density plots are shown. **C.** The high-level *AR* amplification in simulated VCaP cfDNA can be detected down to

5% tumor content. *In vitro* dilution of simulated VCaP cfDNA to the indicated tumor content using was performed using a healthy male control cfDNA sample prior to low-pass WGS. Log (base 2) copy number ratios (Log2 CN Ratio) are plotted. **D.** Bin-level and segmented genome-wide copy number calls from a similar *in vitro* dilution series of simulated UMUC-5 (a urothelial cancer cell line) cfDNA subjected to low-pass WGS. Broad whole-chromosome and arm-level copy-number alterations, including both 1- and 2-copy deletions, are called at expected log2CN values across dilutions. The *EGFR* locus is highlighted. **E.** The known focal *EGFR* amplification is clearly detected down to effective tumor content of 5%. Bin sizes: 15Kbp (**A** & **C**) or 1Mbp (**D & E**). Segmentation p-value threshold: 0.01.

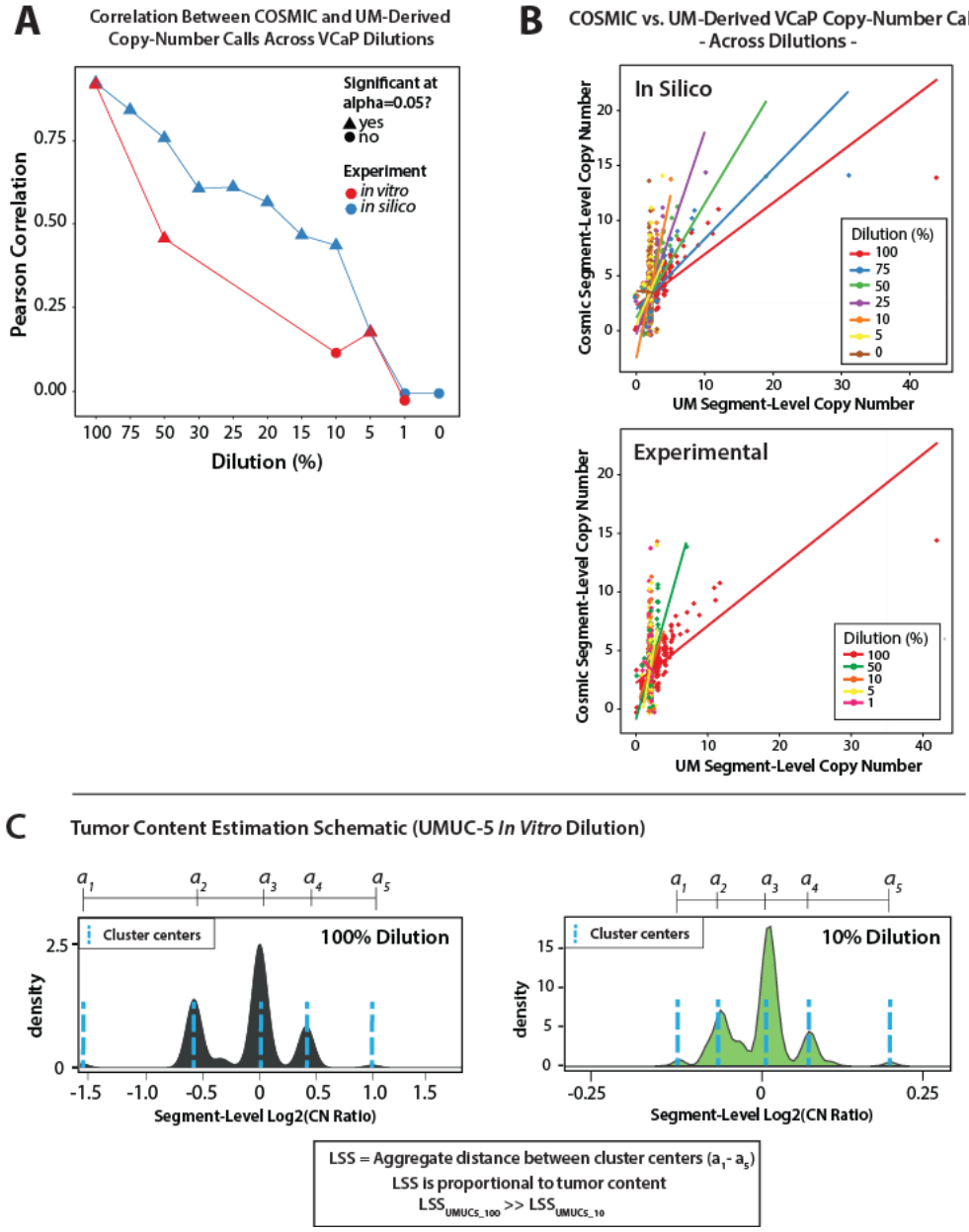
Figure C3: Cell free DNA (cfDNA) tumor content approximation from low-pass whole genome sequencing (WGS) derived copy number profiles.



Unlike in tissue based next generation sequencing (NGS), tumor content cannot be assessed a priori for cfDNA. Such information is critical to guide sequencing depth. Hence, most cfDNA approaches employ ultra-deep, high fidelity sequencing at limited loci to guide therapy with or without direct tumor content approximation. Here we leverage the near ubiquity of copy number alterations (CNAs) across tumors and our ability to rapidly generate whole genome copy number profiles from cfDNA subjected to low-pass WGS to estimate cfDNA tumor content based on the distribution of segment-level copy-number calls as part of our

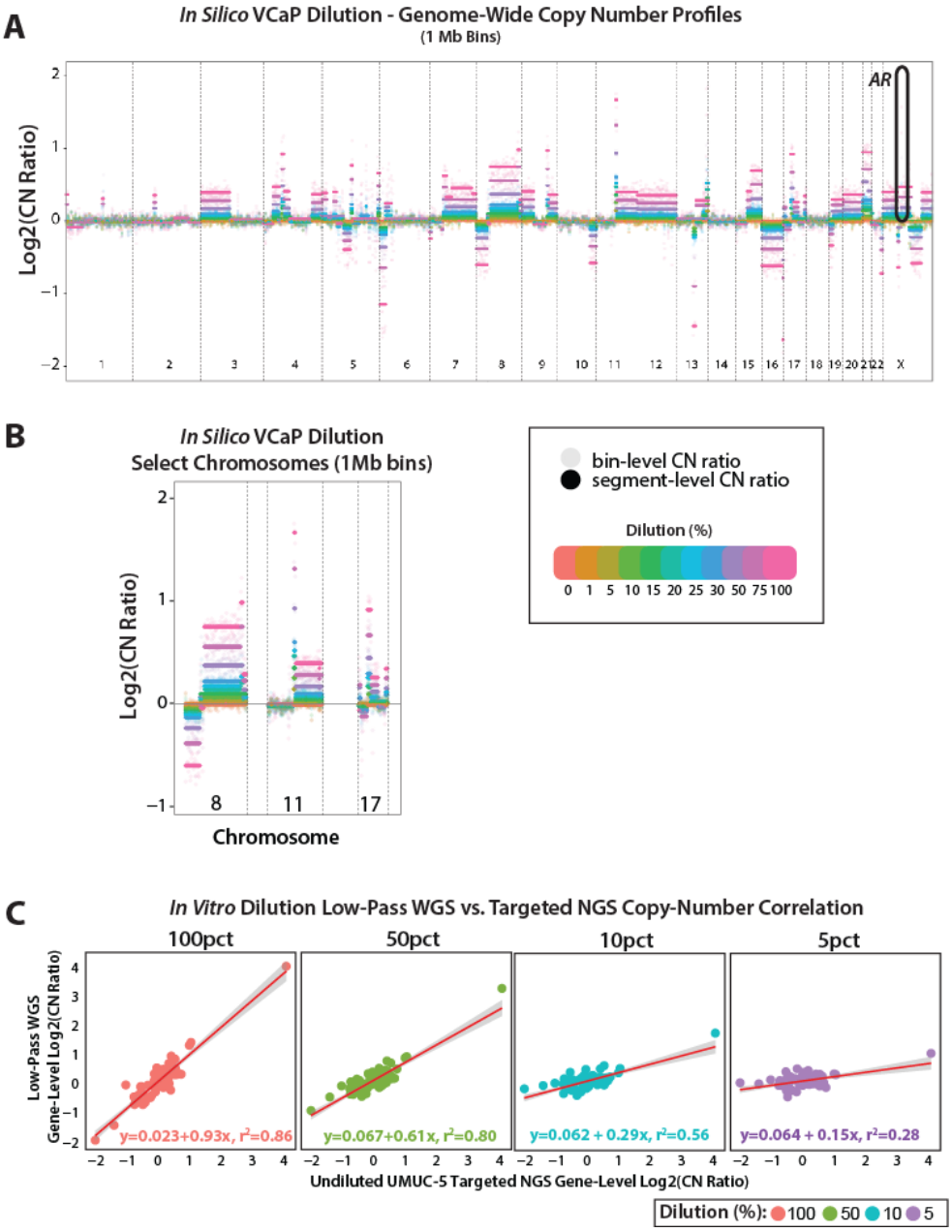
PRINCe workflow. **A**. The relative density of segment-level log (base 2) copy-number ratio (log2CN) values from low-pass WGS of the *in vitro* simulated UMUC-5 cfDNA dilution series (samples according to the legend). **B.** The basic principles of copy-number clustering and tumor content approximation as part of the PRINCe workflow are shown using the density of segment-level log2 copy-number ratio values for the simulated UMUC-5 cfDNA *in vitro* dilution series (from highlighted region of **A**). Clusters are called using a mean-shift clustering algorithm on segmented log2CN values, and cluster centers are used to determine a least-squares distance metric (LSS) for tumor content approximation (see **Methods, Fig B2**). Cluster assignment for presumed 1-copy deletions detected by low-pass WGS in the UMUC-5 simulated cfDNA dilution series are labeled (stars), as are 1-copy deletion (blue dashed vertical line) and diploid/unaltered (red dashed vertical line) cluster centers. As tumor content decreases so does the distance between cluster centers. Aggregate distance between all cluster centers for a given cfDNA sample is calculated (as LSS) and translated to estimate the cfDNA tumor content. **C.** Tumor content approximation from segmented log2CN calls (bin size: 1Mbp; segmentation p-value threshold: 0.01) across *in silico* dilution of simulated VCaP and UMUC-5 cfDNA were used to establish reference distributions for LSS interpretation and tumor content approximation. **D**. Validation of our LSS based tumor content approximation approach on segmented whole exome sequencing based copy number profiles from 129 advanced/metastatic cancer (prostate, kidney, lung and breast cancer) tissue samples sequenced as part of the MI-ONCOSEQ program. Box-plots of our LSS metric stratified by MI-ONCOSEQ estimated tumor contents (through modeling SNVs and heterozygous SNPs, lower estimate is 30%) are shown.

Figure C4: Validation of low-pass WGS copy number estimation for use in cfDNA tumor content estimation.



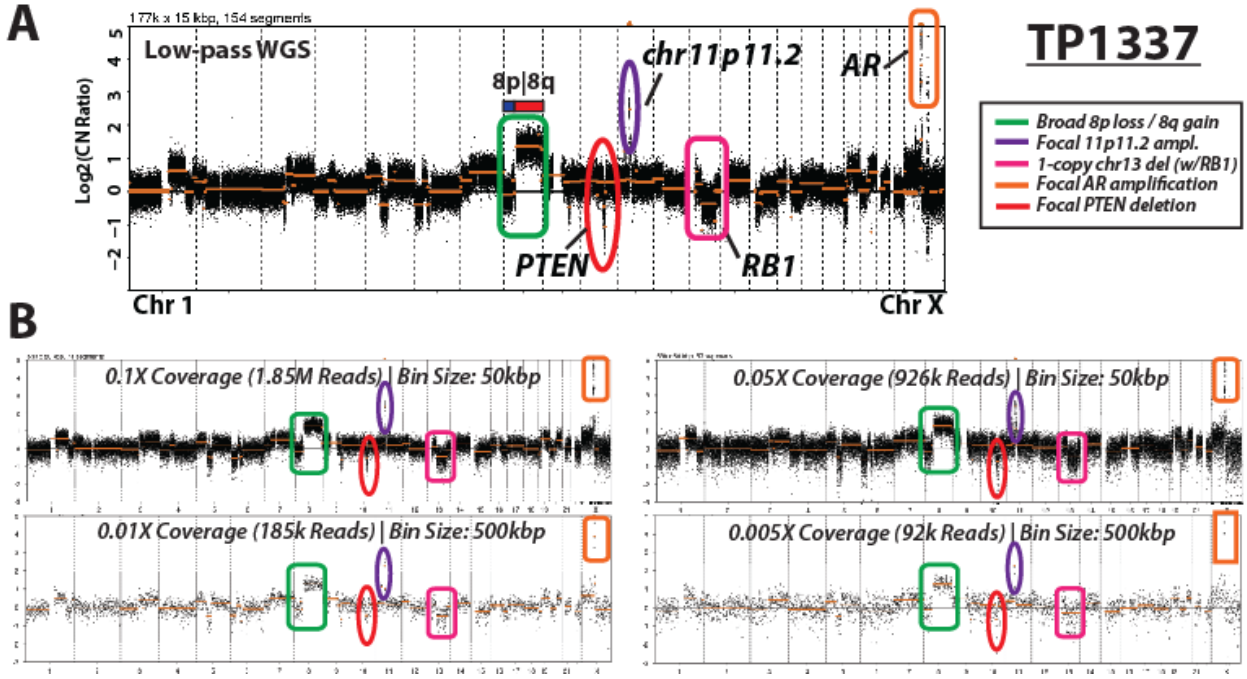
A. Correlation between low-pass WGS and COSMIC copy number calls for *in vitro* and *in silico* dilutions of simulated VCaP cfDNA. Copy number analysis was performed on data from low-pass whole-genome sequencing (WGS) of *in vitro* and *in silico* dilution series for simulated VCaP cfDNA (see **Methods**). Pearson correlations between COSMIC integer-level segmented copy-number and low-pass WGS copy-number (UM-Derived) values are shown across select *in silico* and all *in vitro* dilutions. **B**. Scatterplot of COSMIC integer-level copy number values compared to UM-Derived values for select *in silico* and all *in vitro* dilutions. Points are colored by dilution, and fitted linear regression lines are plotted for each dilution. **C**. Key parameters for cfDNA tumor content estimation based on WGS copy-number profiles. Relative density of segment-level log (base 2) copy number ratio (Log2[CN Ratio]) values from low-pass WGS of UMUC-5 simulated cfDNA is plotted separately for 100% (undiluted) and 10% dilutions. Hypothetical cluster centers are denoted as blue dashed vertical lines, and correspond to elements *a1 - a5* labeled above each plot. A least-squares distance metric (LSS) is calculated (see **Methods**) from cluster centers assigned via a mean-shift clustering algorithm, and LSS is translated to approximate cfDNA tumor content. LSS is proportional to approximate tumor content, with larger LSS values representing higher effective cfDNA tumor content.

Figure C5: Genome-wide low-pass whole genome sequencing (WGS) copy number calls for *in silico* dilution of simulated VCaP and UMUC-5 cfDNA.



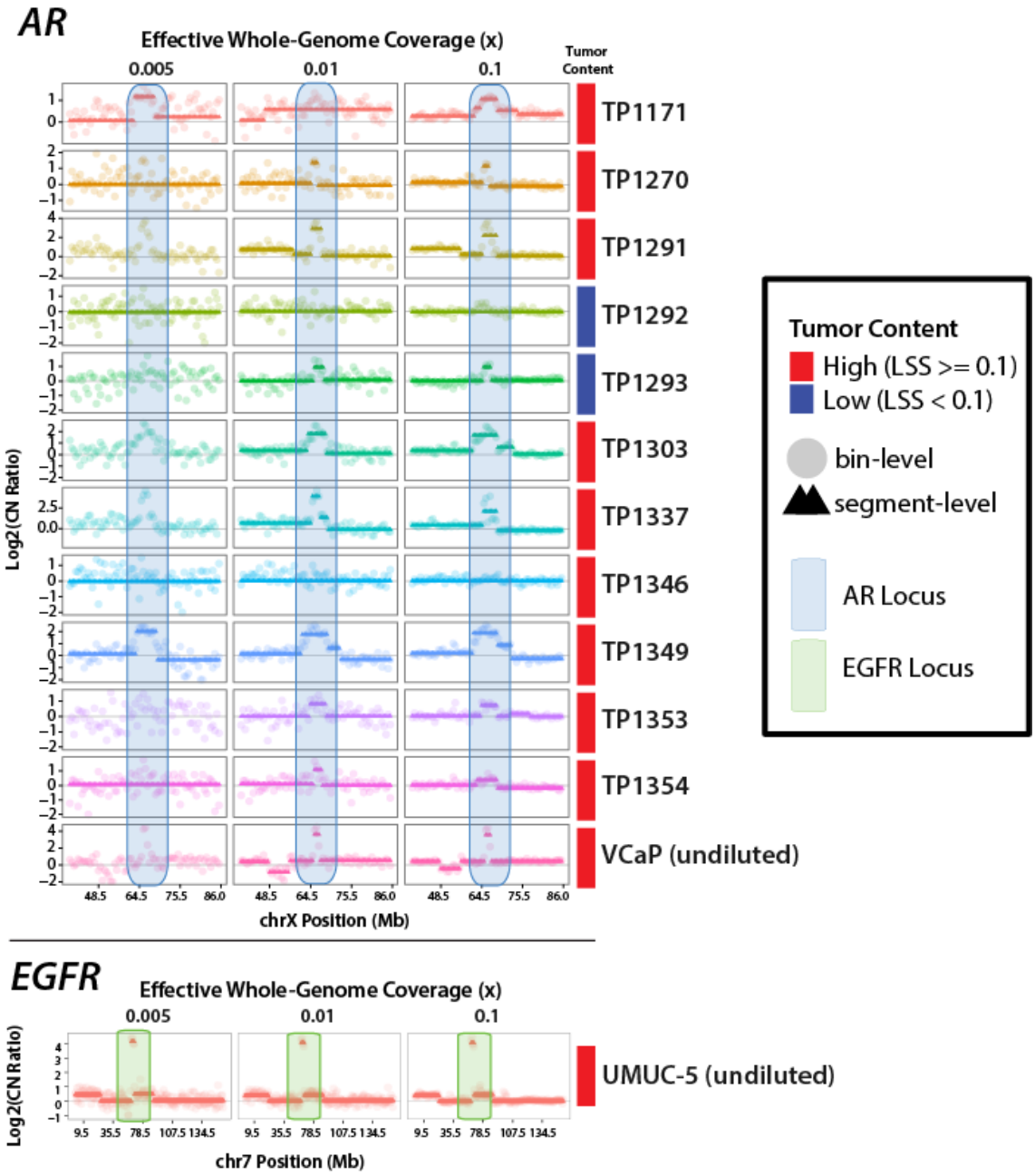
A. Whole-genome bin- (gray points) and segmented (colored bars) copy-number calls (bin size: 1Mb, segmentation p-value threshold: 0.01) at select *in silico* dilutions for VCaP low-pass WGS data highlight log (base 2) copy number ratio values (Log2 CN Ratio) values at expected gradations across dilutions for alterations >2Mbp in length. The known focal *AR* amplification in VCaP is ~1Mbp in length (COSMIC *AR* amplification call: chrX:66031108-67075149) and can be seen via bin-level estimates at *AR* loci as shown. **B.** Zoomed view of bin- and segmented copy-number calls for chromosomes 8, 11, and 17 shows both broad and focal copy number alterations at Log2 CN Ratios consistent with *in silico* dilution. **C.** Comparison of gene-level Log2CN values from targeted NGS of undiluted, unamplified simulated UMUC-5 cfDNA and low-pass WGS gene level calls for *in vitro* UMUC-5 dilution (see **Methods**). Points are colored by *in vitro* dilution, and fitted linear regression lines and 95% confidence intervals are plotted. Linear models and r2 values are provided for each *in vitro* dilution.

Figure C6: Bioinformatic analysis highlighting potential feasibility of ultra-low pass (<0.01x) whole genome sequencing (WGS) of cfDNA as a disease monitoring application from cell-free DNA in patients with advanced cancer.



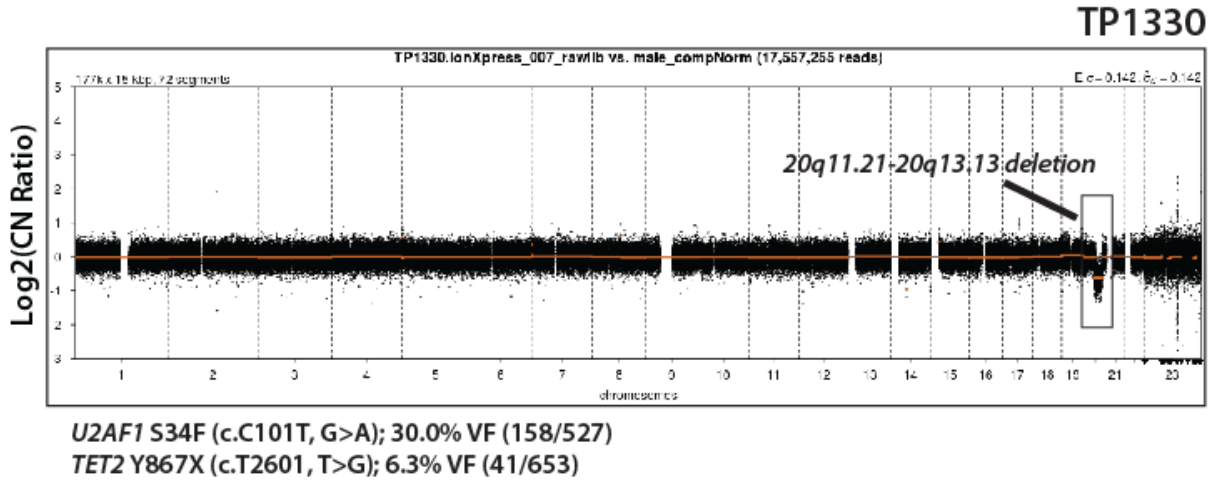
A. Genome-wide log2(CopyNumberRatio) (Log2CN) calls for TP1337, a high tumor content cfDNA sample from a patient with mCRPC, are displayed for low-pass WGS data (0.82x whole-genome coverage). Key copy-number alterations detected are circled, including broad gain of 8q (green), focal amplification of chr11p11.2 (purple) and *AR* (orange), and focal deletions of *RB1* (1-copy; pink) and *PTEN* (2-copy; red). **B.** *In silico* downsampling experiments highlight the ability to detect both focal and broad copy-number alterations from TP1337 cfDNA WGS data at whole-genome coverages down to 0.005x. Bin size and number of high-quality (MAPQ ≥37) mapped reads used for copy-number analysis are indicated at each coverage, and regions affected by copy-number alterations detected in original low-pass WGS are circled.

Figure C7: *AR* and *EGFR* amplifications detected in *in silico* downsampling of simulated cell line cfDNA and patient cfDNA samples.



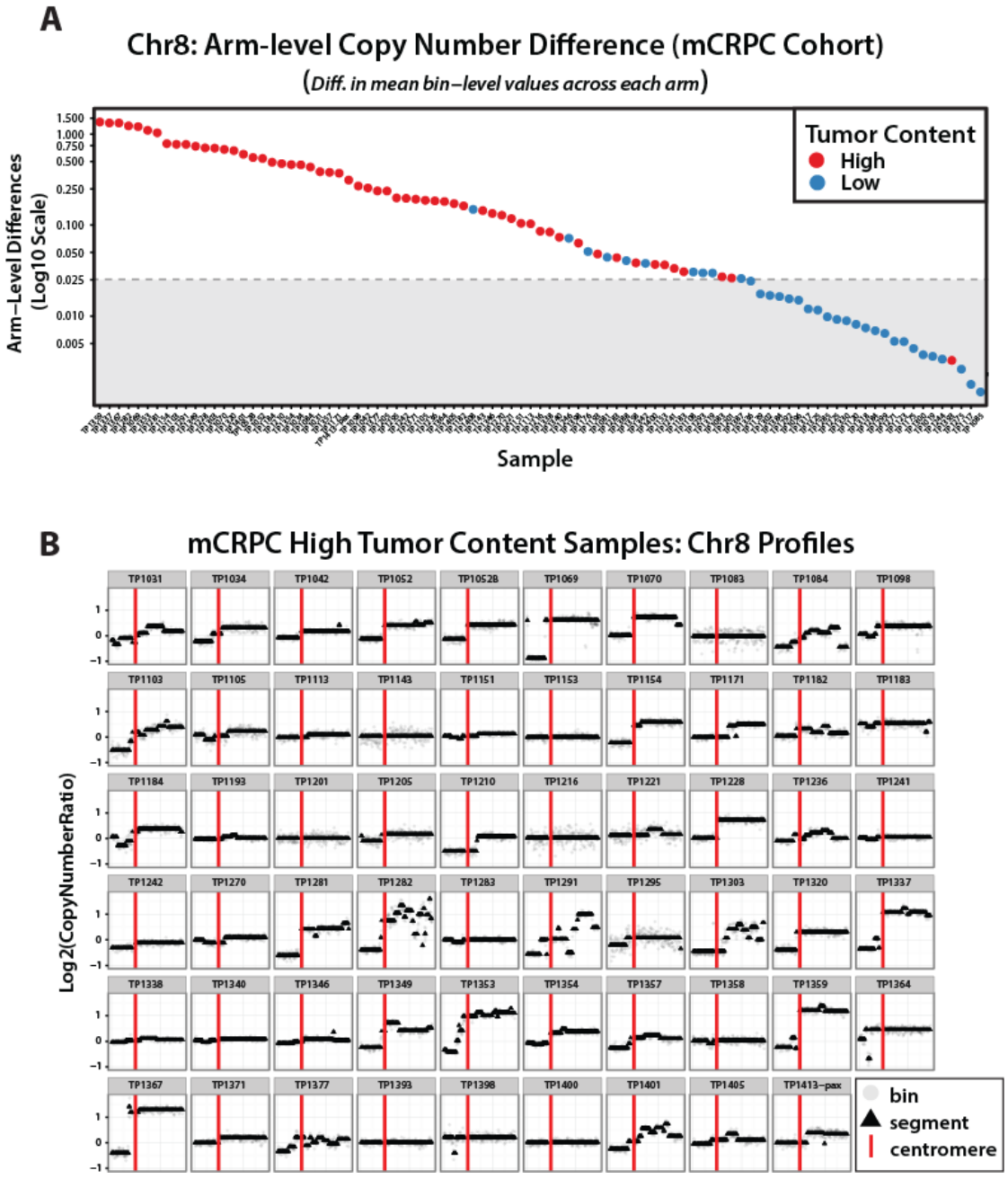
In silico downsampling experiments were carried out on low-pass whole genome sequencing (WGS) data from simulated cell line cfDNA (VCaP and UMUC-5) and 11 cfDNA samples from patients with mCRPC to yield ultra low effective whole genome coverages (0.1x, 0.01x, and 0.005x). **A.** Bin- and segment-level log (base 2) copy number ratio (Log2 [CN Ratio]) calls are presented across effective whole genome coverages in *AR* region on chrX for mCRPC samples with detectable *AR* amplifications by low-pass WGS as well as the undiluted simulated VCaP cfDNA sample. Points are colored by sample, 500kbp bin-level and segment-level Log2 (CN Ratio) estimates are represented by lightly shaded circles and densely colored triangles, respectively. Tumor content estimates are highlighted by red (high) and blue (low) boxes at right. The *AR* locus is highlighted in light blue boxes. **B.** Bin- and segment-level Log2 (CN Ratio) copy number calls for undiluted artificial UMUC-5 cfDNA sample are presented across effective whole-genome coverages for chr7, and the *EGFR* locus is highlighted in light green boxes. Bin- and segment level estimates are indicated as in **A**.

<u>Figure C8</u>: PRINCe assessment of sample from patient with metastatic castration-resistant prostate cancer (mCRPC) identifies unique molecular alterations consistent with contaminating cell-free DNA from white blood cells in the context of concomitant myelodysplastic syndrome.



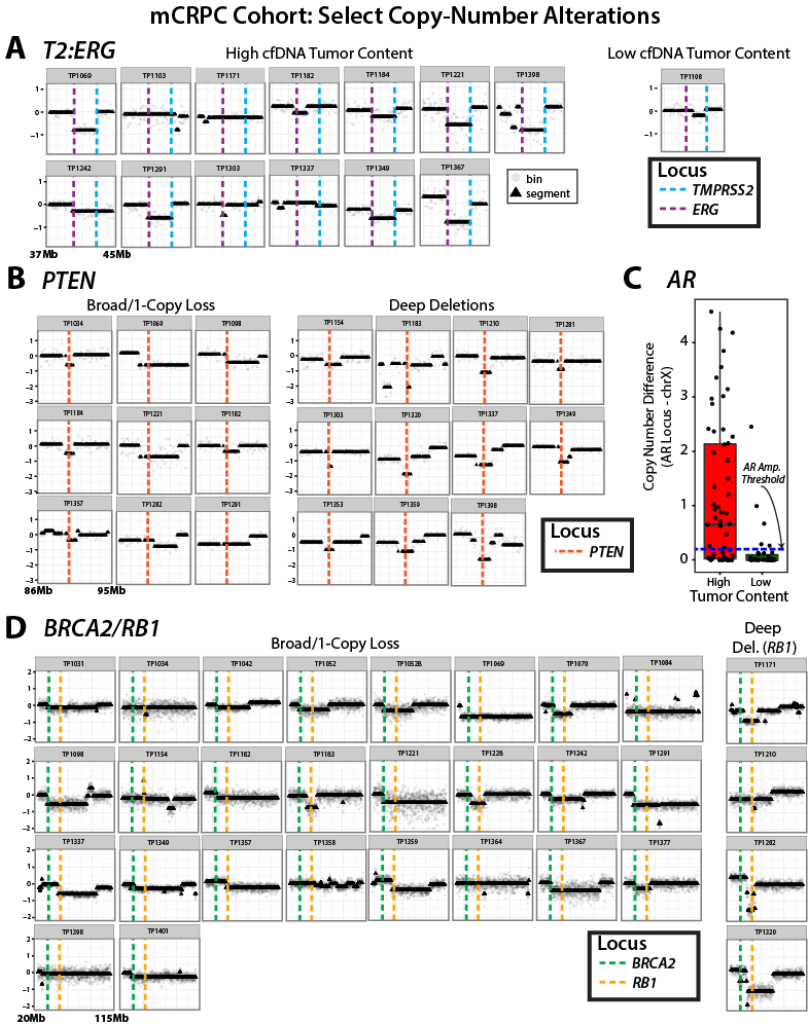
Low-pass whole-genome sequencing (WGS) copy-number calls (bin size: 15kbp, segmentation p-value threshold: 0.01) are plotted for a cfDNA sample from TP1330, a patient with mCRPC. A unique 19Mbp deletion (affecting chr20q11.21-20q13.13) is present on chr20, with no other copy-number alterations detected genome-wide. By targeted NGS of unamplified residual cfDNA for this same sample, we identified a *U2AF1* S34F hotspot mutation (30% variant fraction (VF), 527 covering reads) that in combination with the chr20 deletion is strongly suggestive of contaminating cell-free DNA (likely from white blood cells) in the context of concomitant myelodysplastic syndrome, consistent with clinical reports of anemia, and potentially arising in response to prior therapy.

Figure C9: Low-pass whole genome sequencing (WGS) copy number profiles from cell-free DNA (cfDNA) in patients with metastatic castration-resistant prostate cancer (mCRPC) highlight detection of arm- and sub-arm level copy-number alterations on chromosome 8 (chr8), even at low cfDNA tumor contents.



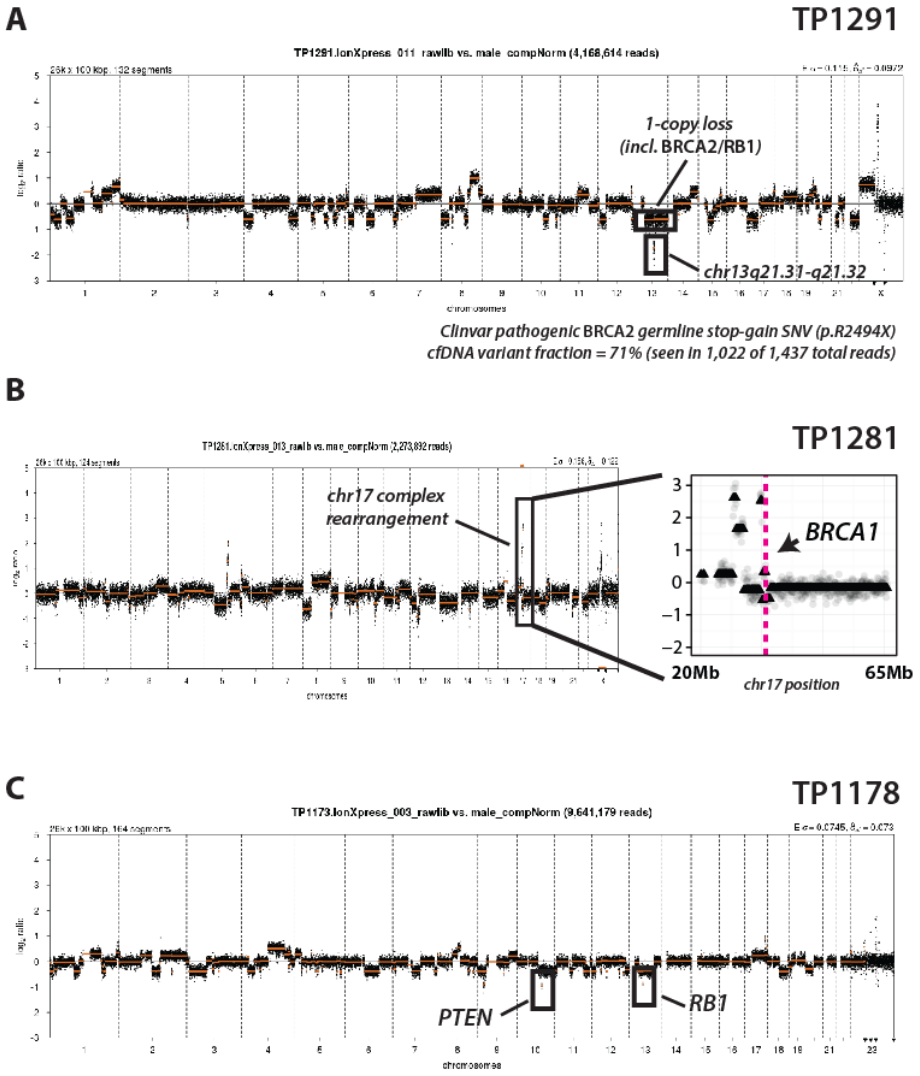
A. Points representing chr8 'difference values' (the absolute value of the difference in mean bin-level log2 copy-number estimates between p and q arm of chr8) for all samples in the mCRPC cohort (*n*=93). Samples are sorted in order of descending difference value, and colored by cfDNA tumor content as assigned by LSS analysis (red=High (LSS ≥ 0.1; blue=Low (LSS < 0.1)). A threshold of 0.025 was applied to difference values to determine whether each cfDNA sample had detectable chr8 copy number alterations (≥ 0.025 = 8p or 8q copy-number alterations) consistent with copy number events known to occur early in prostate cancer progression. **B.** Low-pass WGS chr8 copy-number profiles for all high tumor content cfDNA samples (*n*=59) from men with mCRPC. As indicated, gray dots correspond to bin-level copy-number estimates, while black triangles denote the segment-level copy number value for the corresponding bin. The vertical red line in each plot indicates the centromere region to aid in p- and q-arm determination.

Figure C10: Clinically relevant somatic copy number alterations detected via low-pass whole genome sequencing (WGS) of cell-free DNA (cfDNA) in patients with metastatic castration-resistant prostate cancer (mCRPC).



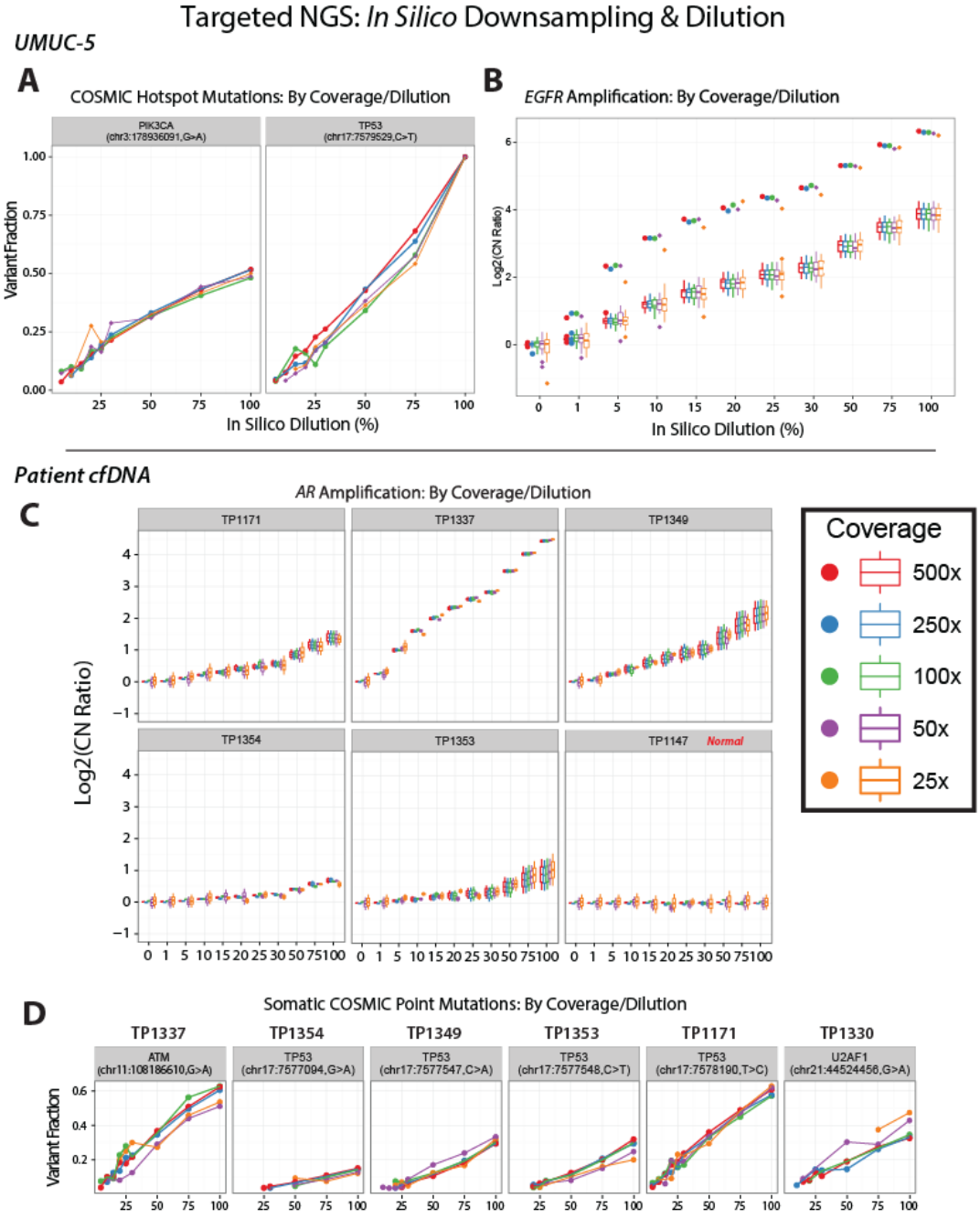
A. Bin- and segment-level copy-number values from low-pass WGS data in 14 cfDNA samples from patients with mCRPC that have copy-number alterations or breakpoints on chr21 consistent with genomic events leading to TMPRSS2:ERG or ETS family gene fusions (displayed region: chr21:37.0-45.0Mb). Tumor content for the corresponding cfDNA sample is listed at top. Dashed vertical lines at 40Mb (purple) and 42.8Mb (cyan) represent loci corresponding to *ERG* and *TMPRSS2* coding sequence (hg19 reference coordinates), respectively. **B.** Bin- and segment-level copy-number values from low-pass WGS data in 20 high tumor content cfDNA samples from patients with mCRPC with chr10 copy-number deletions affecting the *PTEN* locus (displayed region: chr10:86.0-95.0Mb). Samples are grouped by deletion type (broad/1-copy or deep). A dashed vertical line at 90Mb (orange) represents the location of *PTEN* coding sequence (hg19 reference coordinates). **C.** Combined box and scatterplot for *AR* deviance values (mean 100kb bin-level log2 copy-number estimates at *AR* locus [chrX:66.0-67.5Mb] minus median 100kb log2 bin-level copy number estimates across chrX q-arm) used to identify focal *AR* amplifications in our mCRPC cohort (see **Supplementary Methods**). Samples with deviance values ≥ 0.2 were considered positive for *AR* amplification, and this threshold is represented by the blue horizontal dashed line as annotated on the plot. Combined box and scatter plots are plotted separately for high (red) and low (green) cfDNA tumor content. **D.** Bin- and segment-level copy-number values from low-pass WGS data in 30 high tumor content cfDNA samples from patients with mCRPC with chr13 copy-number deletions affecting *BRCA2/RB1* loci (displayed region: chr13:20.0-115.0Mb). Samples are grouped by deletion type (broad/1-copy vs deep deletion), and dashed vertical lines at 33Mb (green) and 49Mb (yellow) indicate *BRCA2* and *RB1* loci, respectively. Reference genome coordinates: hg19. Bin-width: 100kb. Copy number segmentation switch-point threshold (p-value): 0.01.

<u>Figure C11:</u> Low-pass whole genome sequencing (WGS) of cell-free DNA (cfDNA) identifies likely copy-number alteration affecting *BRCA1* and *BRCA2* in patients with mCRPC as well as clinically relevant alterations (including focal PTEN and RB1 deletions) in treatment-naïve patient with aggressive disease.



A. Genome-wide bin- (black dots) and segment-level (orange lines) log2 copy number estimates from low-pass WGS sequencing data for TP1291, a patient with mCRPC who progressed rapidly through treatment with abiraterone, enzalutamide, docetaxel, and cabazitaxel over the course of 11 months preceding cfDNA sample collection. Broad 1-copy loss on chr13 (including *BRCA2* and *RB1*) is indicated, as is the focal 2-copy deletion of a nearby loci absent any coding transcripts (chr13q21.31-q21.32). Targeted NGS of paired unamplified cfDNA for TP1291 identified a germline Clinvar pathogentic *BRCA2* stop-gain SNV (p.R2494X; variant fraction = 71%,with 1,437 total covering reads), suggesting biallelic inactivation of *BRCA2* through copy-number deletion of the non-mutated copy of *BRCA2*. **B.** Bin- (black dots) and segment-level (orange lines) log2 copy number estimates from low-pass WGS sequencing data are presented genome-wide at for a 45Mb section of chr17 for TP1281, a sample from a patient with mCRPC. Region affected by putative complex rearrangement on chr17 is highlighted on the genome-wide plot, and at right, a zoomed version indicates the location of *BRCA1* (dashed vertical pink line; hg19 reference coordinates). **C.** Genome-wide bin- (black dots) and segment-level (orange lines) log2 copy number estimates from low-pass WGS sequencing data (0.52x, 14.2 million reads) for TP1178, a cfDNA sample from a treatment-naïve patient with mCRPC. Focal deep deletions of *PTEN* and *RB1* are identified in addition to multiple arm- and sub-arm level copy number gains or losses genome-wide, suggesting potential clinical utility for PRINCe assessment in treatment-naïve patients with advanced cancer and/or likely to have high disease burden. Bin width: 100kbp.

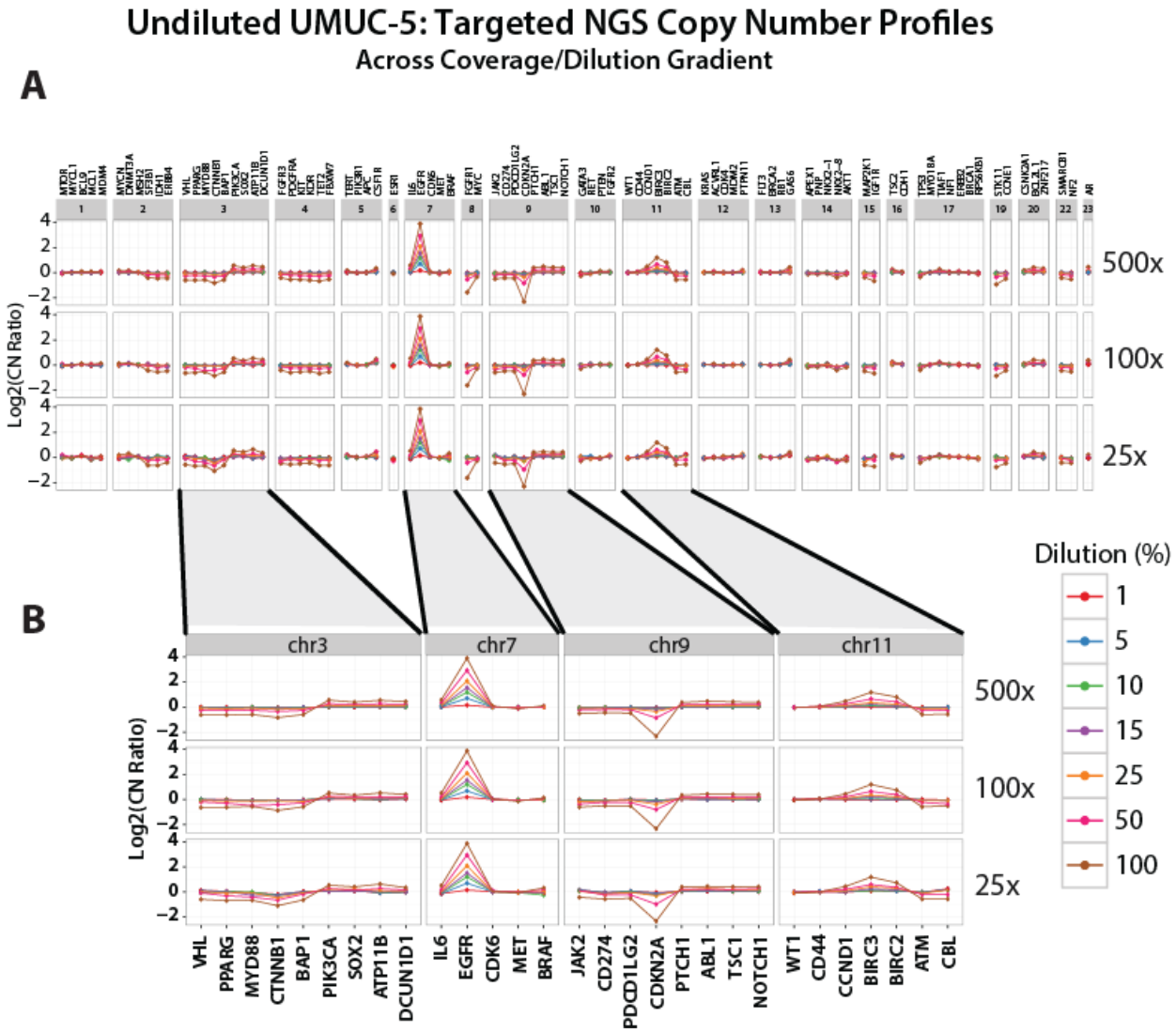
<u>Figure C12:</u> Automated point mutation and copy number alteration calls across *in silico* dilution and downsampling of targeted next generation sequencing (NGS) from simulated cell line cfDNA and patient cfDNA samples.



In silico dilution and downsampling experiments were carried out on targeted NGS data for unamplified, undiluted genomic DNA (gDNA) from the UMUC-5 cell line, as well as unamplified aliquots of 10 patient cfDNA samples (5 high tumor content mCRPC samples, 1 mCRPC sample with germline chr20 deletion, and 4 male normal controls). These samples were sequenced using the DNA component of the Oncomine Comprehensive Assay (OCP), a targeted NGS panel comprised of 2,530 amplicons targeting 126 genes, including oncogenes, tumor suppressors, and copy-number targets recurrently altered across cancers. *In silico* dilutions were carried out at all integer-level dilutions (0-100%) across 5 different effective coverage thresholds (500x, 250x, 100x, 50x, and 25x) (see **Methods**). **A.** Analyses of select *in silico* dilution and downsampling data from two COSMIC hotspot point mutations detected in targeted NGS of undiluted simulated UMUC-5 cfDNA are presented. A heterozygous

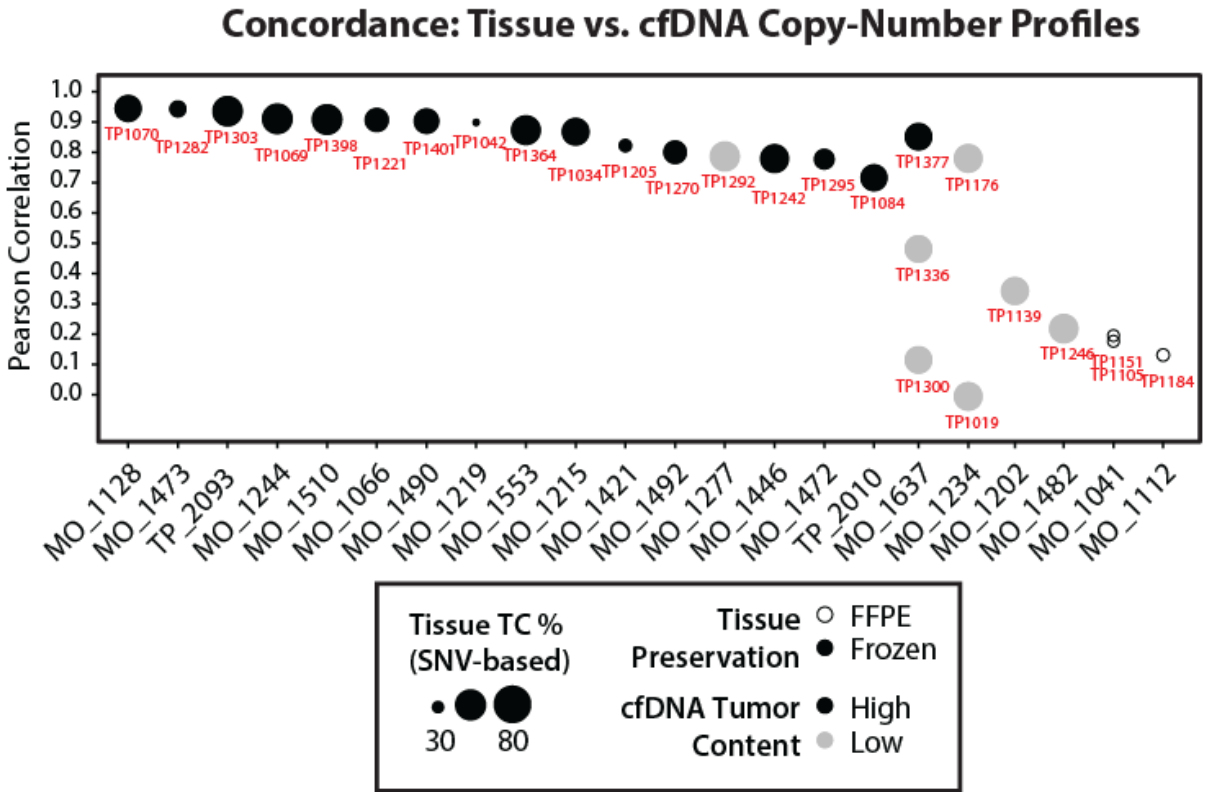
nonsynonymous *PIK3CA* hotspot mutation (p.E545K, detected at 49.6% (FAO/FDP: 916/1848)) and homozygous stop-gain *TP53* hotpot mutation (p.W53X, 100% (668/668)) are reliably detected at expected variant fractions across targeted NGS coverages as low as 50x down to effective tumor contents of 10-15%. **B.** Box-and-whisker plots of amplicon level log base 2 copy number ratio (Log2 [CN Ratio]) estimates from OCP sequencing of undiluted simulated UMUC-5 cfDNA are plotted for all OCP *EGFR* target amplicons (n=33) across select *in silico* dilutions and coverages. Known focal *EGFR* amplification (undiluted UMUC-5 OCP gene-level *EGFR* Log2 (CN Ratio) value = 3.89) in UMUC-5 cell line is reliably detected (median Log2 [CN Ratio] $\geq$ 0.6; see **Methods**) across coverages down to 25x at 5% effective tumor content. **C.** Box-and-whisker plots of Log2 (CN Ratio) values for all OCP *AR* amplicons (n=17) are shown across in silico dilutions and coverages for 6 patient cfDNA samples (5 mCRPC samples with detectable AR amplifications by low-pass whole genome sequencing (WGS) copy-number analysis, and 1 control sample). Depending on starting tumor content of undiluted ('100%' dilution) patient cfDNA samples, *AR* amplifications can be reliably detected in targeted NGS data from CPRC cfDNA samples at coverages down to 25x at 5% dilution. **D.** Putative clonal somatic COSMIC hotspot mutations detected via targeted NGS in 6 mCRPC patient samples are plotted across select *in silico* dilutions and effective coverages. All mutations are detected at heterozygous variant fractions in undiluted ('100%' dilution) OCP targeted NGS data, adjusting for cfDNA tumor content. Depending on starting (undiluted) cfDNA tumor content, hotspot mutations were reliably detected by OCP targeted NGS at coverages as low as 50x with 15% effective tumor content.

Figure C13: **Targeted NGS gene-level copy-number analysis across *in silico* dilution and downsampled coverages for simulated UMUC-5 cfDNA.**



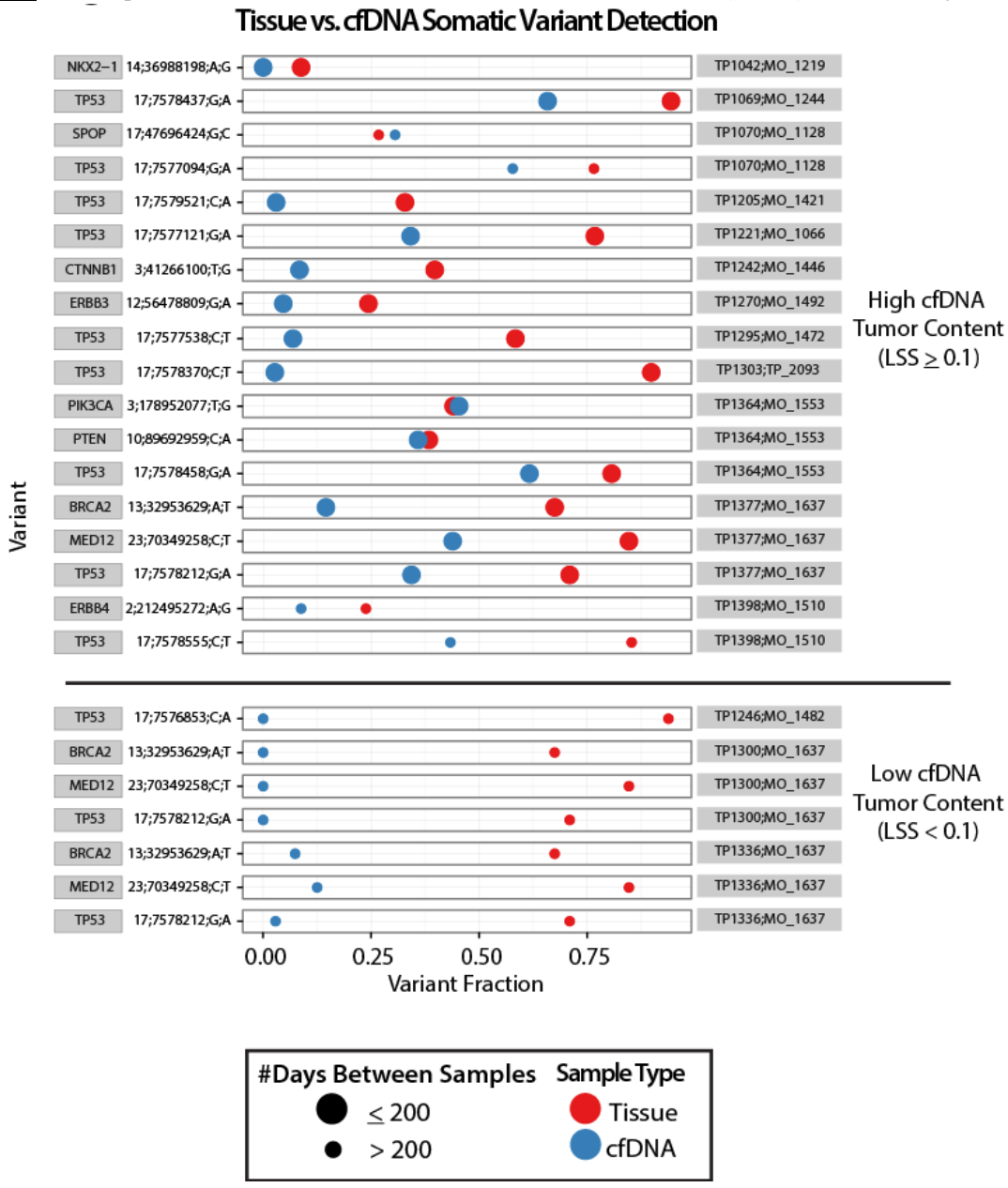
A. Gene-level log (base 2) copy number ratio (Log2 CN Ratio) values derived from Oncomine Comprehensive Assay (OCP) targeted NGS data for simulated UMUC-5 cfDNA are plotted across select *in silico* dilutions at three separate coverages (500x, 100x, and 25x) for all OCP target genes with at least 3 target amplicons (n=90). Points represent gene-level Log2CN Ratio values, with points (and lines connecting points) colored by *in silico* dilution proportion. The known focal *EGFR* amplification can be seen as peak on chromosome 7. **B.** Zoomed view of OCP gene-level Log2CN Ratio values for select chromosomes (chr3, chr7, chr9, and chr11). Focal amplifications or deletions identified by low-pass WGS can be detected at targeted NGS coverages down to 25x fordilutions with as low as 5% effective tumor content.

<u>Figure C14:</u> Genome-wide copy number profile concordance for cfDNA low-pass whole genome sequencing (WGS) as compared to patient-matched tissue whole exome sequencing (WES) copy-number profiles.



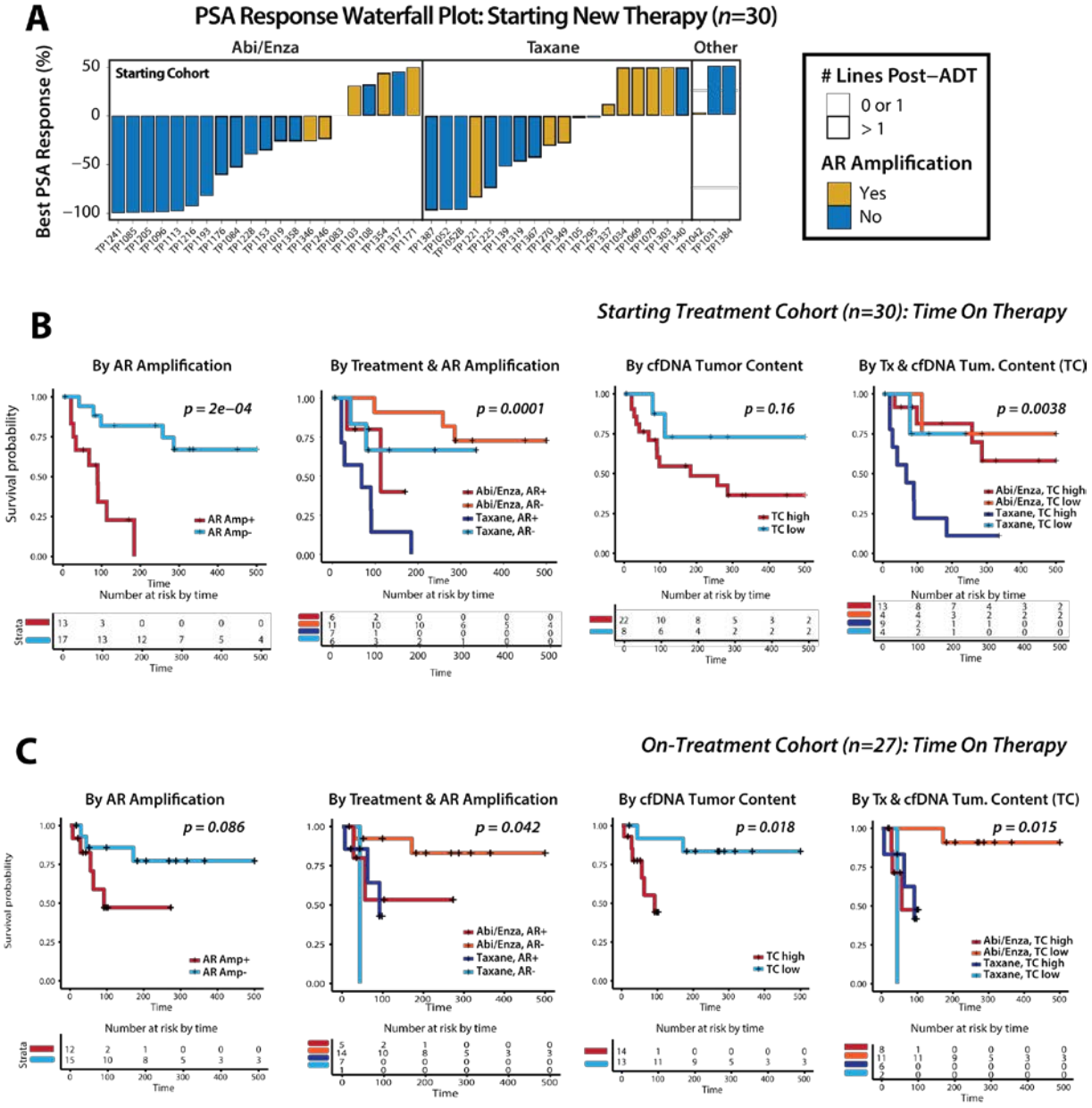
Pearson correlation coefficients are plotted for genome-wide segmented copy-number profiles from 23 patients with comprehensive tissue NGS profiling and PRINCe assessment of at least 1 cfDNA sample (see **Supplementary Methods**). As indicated, each point represents the correlation of a single cfDNA sample's low-pass WGS genome-wide profile as compared to the patient-matched whole exome sequencing tissue copy-number profile (see **Supplementary Methods**). The size of each point corresponds to the SNV-based estimated tissue tumor content (which varies from 30 to 80%), while the color represents cfDNA tumor content (black: high; gray: low). Circle filling represents patient-matched tissue preservation type (filled: frozen; unfilled: FFPE). cfDNA sample identifers are provided in red.

Figure C15: **Somatic point mutation concordance between tissue and cell-free DNA (cfDNA) mutation analyses.**



Tissue and cfDNA variant fractions are plotted for 26 point mutations identified in patient-matched tissue-based whole exome sequencing (WES) that fall in regions targeted by the Oncomine Comprehensive Assay (OCP). Variants are sorted vertically by increasing cfDNA sample identifier, and all cfDNA/tissue id combinations are listed on right hand side. Each row corresponds to a single variant detected in the comprehensive patient-matched tissue profile, and the gene, genomic coordinates, and allelic changes are indicated on left hand side of each row. For each variant, both tissue- (red) and cfDNA-based (blue) variant fractions are plotted (variant fraction of 0% = not detected), and points are sized by whether the corresponding cfDNA sample was taken within 200 days of the patient-matched tissue biopsy as indicated in the legend. Variants are grouped vertically by cfDNA tumor content for the corresponding cfDNA sample (top: high tumor content (LSS ≥ 0.1); bottom: low tumor content (LSS < 0.1)). Overall, 17 of 18 (94.4%) point mutations detected in patient-matched tissue specimens with ≥ 1 high tumor content cfDNA sample were also detected by OCP targeted NGS of the corresponding cfDNA sample.

Figure C16: PSA waterfall and outcome analyses in samples from patients starting and on therapy.



Exploratory analyses of association between circulating biomarkers and outcome in patients with metastatic castration-resistant prostate cancer (mCRPC) supports cfDNA detectable AR amplification as a poor overall prognostic factor independent of treatment type. **A.** Waterfall plot summarizing prostate specific antibody (PSA) response for all samples from men with mCRPC with complete PSA data taken between therapies (n=42). Height of bars represent the percentage change in PSA response as calculated by subtracting the PSA level at sample date from the best PSA observed after sample date while on the current or initiated treatment, and dividing by starting PSA value. Bars are ordered horizontally within treatment category (Abi/Enza, Taxane, or Other) by PSA response. Bars are colored by cfDNA detectable AR amplification status (yellow = cfDNA detectable AR amplification; gray = no cfDNA detectable AR amplification) and bars corresponding to samples taken from men who have received more than one line of therapy post-ADT are outlined in bold. **B-C**. Kaplan-Meier survival curves are plotted for analyses exploring association between cfDNA detectable AR amplification or cfDNA tumor content and total time on therapy in samples taken from men with CRPC (**B**) starting treatment and (**C**) on therapy. For each subset, Kaplan-Meier time on therapy analyses are plotted separately (from left to right) for cfDNA AR amplification, treatment by cfDNA AR amplification, cfDNA tumor content, and treatment by cfDNA tumor content. Survival curves are colored by corresponding strata, and risk tables at selected timepoints are displayed below each Kaplan-Meier plot.

Appendix C References

1. *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run.*, B.I.T.G.D.A. Center, Editor. 2016: Broad Institute of MIT and Harvard.
2. Scheinin, I., et al., *DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly.* Genome Res, 2014. **24**(12): p. 2022-32.
3. Ulz, P., et al., *Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer.* Nat Commun, 2016. **7**: p. 12008.
4. Hovelson, D.H., et al., *Development and validation of a scalable next-generation sequencing system for assessing relevant somatic variants in solid tumors.* Neoplasia, 2015. **17**(4): p. 385-99.
5. Cani, A.K., et al., *Next-Gen Sequencing Exposes Frequent MED12 Mutations and Actionable Therapeutic Targets in Phyllodes Tumors.* Mol Cancer Res, 2015. **13**(4): p. 613-9.
6. McDaniel, A.S., et al., *Genomic Profiling of Penile Squamous Cell Carcinoma Reveals New Opportunities for Targeted Therapy.* Cancer Res, 2015. **75**(24): p. 5219-27.
7. McDaniel, A.S., et al., *Next-Generation Sequencing of Tubal Intraepithelial Carcinomas.* JAMA Oncol, 2015. **1**(8): p. 1128-32.
8. Liu, W., et al., *Homozygous deletions and recurrent amplifications implicate new genes involved in prostate cancer.* Neoplasia, 2008. **10**(8): p. 897-907.
9. Roychowdhury, S., et al., *Personalized oncology through integrative high-throughput sequencing: a pilot study.* Sci Transl Med, 2011. **3**(111): p. 111ra121.
10. Robinson, D.R., et al., *Activating ESR1 mutations in hormone-resistant metastatic breast cancer.* Nat Genet, 2013. **45**(12): p. 1446-51.
11. Robinson, D., et al., *Integrative clinical genomics of advanced prostate cancer.* Cell, 2015. **161**(5): p. 1215-28.

APPENDIX D: Supplementary Materials for Chapter IV

Figure D1 – Assessment of major transcriptional programs for 234 bladder cancer specimens profiled via TCGA using markers targeted on a bladder RNAseq panel

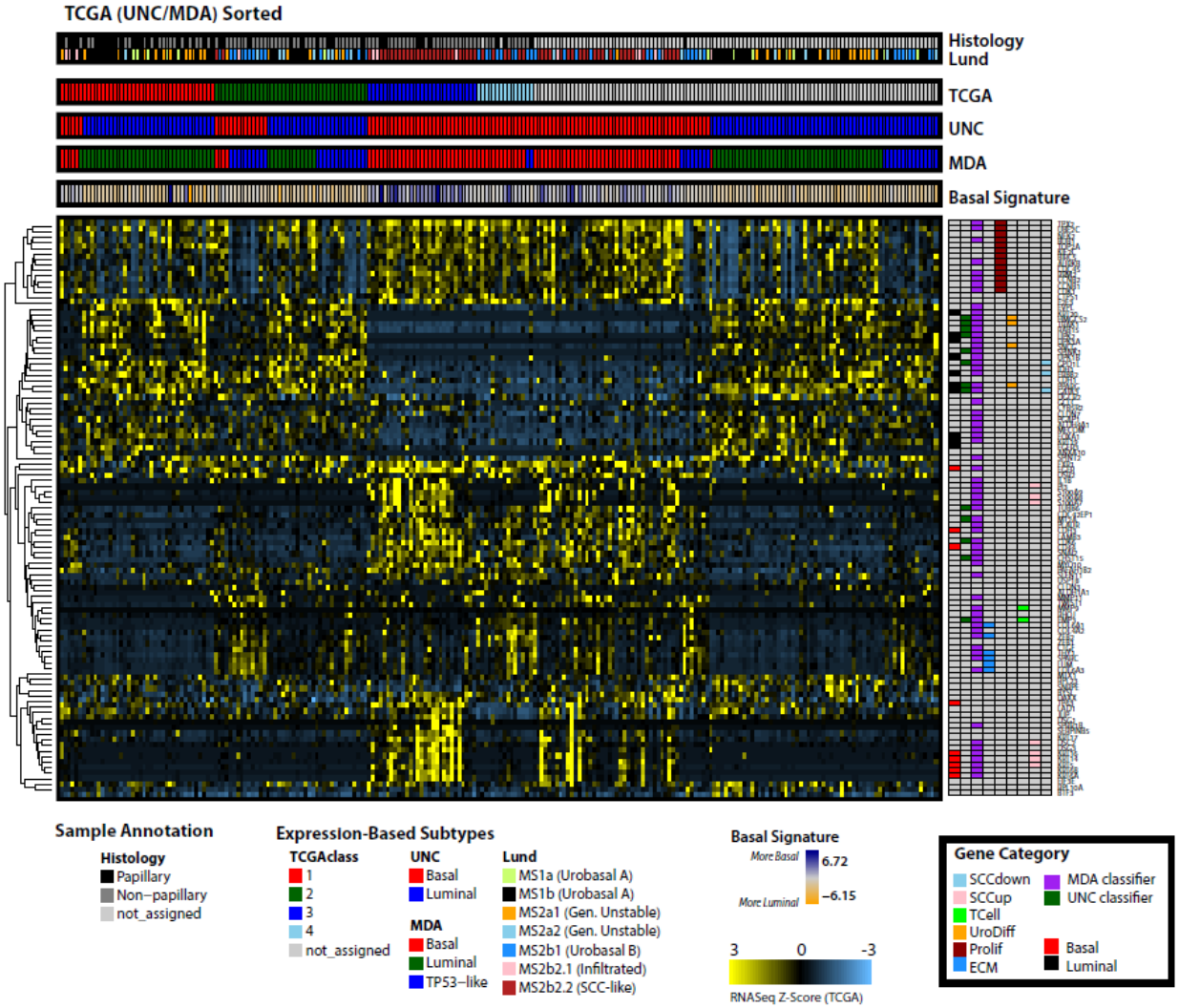


Figure D1 – Unsupervised hierarchical clustering of TCGA gene expression Z-score values for 234 bladder cancer tissue specimens profiled via TCGA. Only TCGA processed data from genes targeted on our bladder targeted RNAseq panel are included. Expression-based subtype annotation corresponding to TCGA (I, II, III, IV), UNC (basal, luminal), MD Anderson (MDA; basal, luminal, TP53-like), and Lund classification approaches for each sample are displayed at the top, and colored according to the legend at the bottom. UM basal signature values are also displayed, and samples with higher basal expression are more blue, while samples with higher luminal expression are more orange. Gene annotation is indicated at right, and major transcriptional programs (including proliferation and EMT markers) are indicated.

<u>Figure D2</u> – Recapitulation of UNC and MDA expression-based subtypes using markers targeted on a custom targeted RNAseq panel

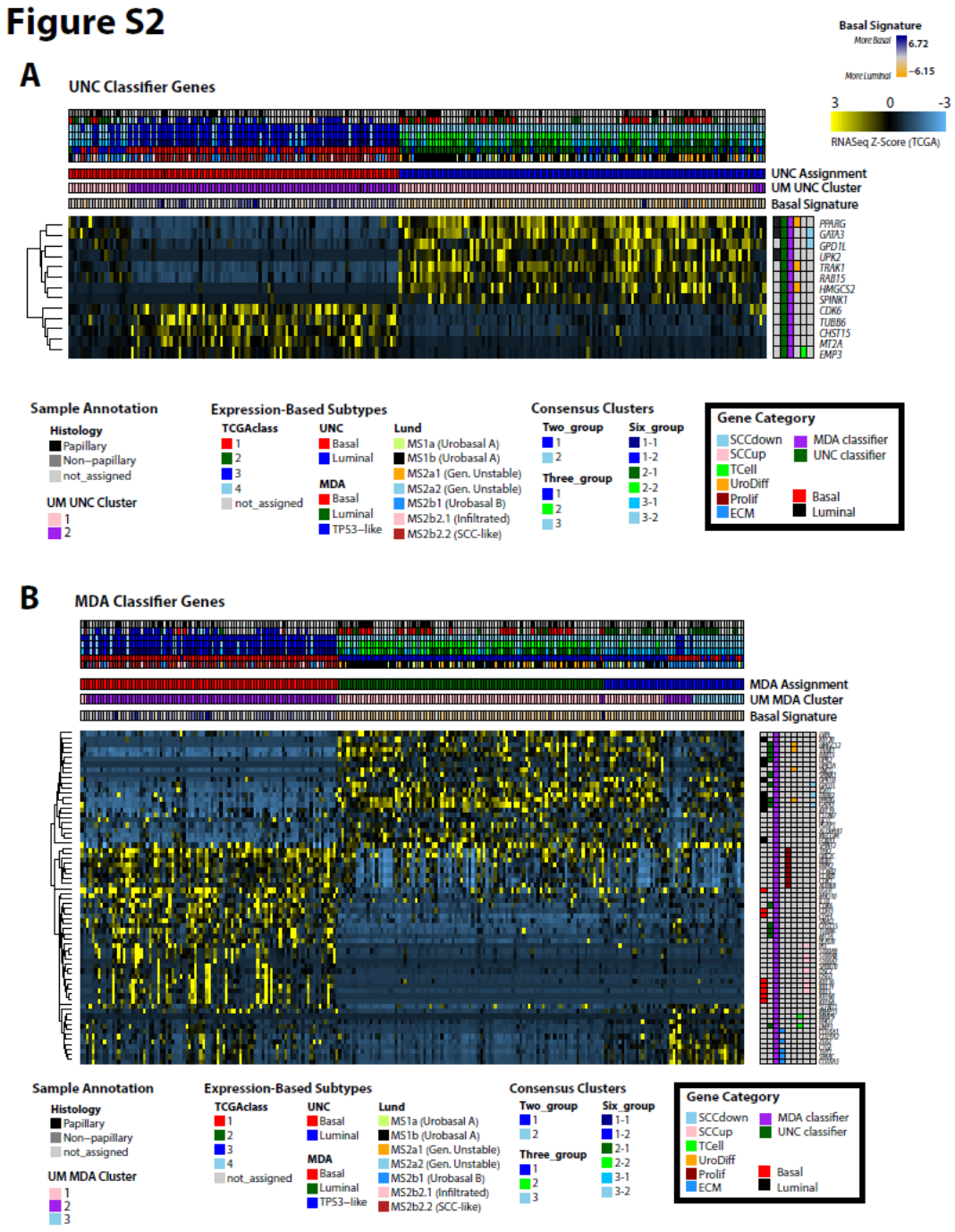


Figure D2 A. Consensus clustering of gene expression Z-score values for 234 TCGA bladder samples using TCGA-processed expression values from genes in UNC classifier that are targeted on our custom bladder RNAseq panel. Number of clusters used to define UM consensus clusters was pre-defined at 2 to evaluate recapitulation of known subtypes using only markers from UNC classifier targeted on our panel. Expression-based subtype annotation corresponding to TCGA (I, II, III, IV), UNC (basal,

luminal), MD Anderson (MDA; basal, luminal, TP53-like), and Lund classification approaches for each sample are displayed at the top, and colored according to the legend at the bottom. UM consensus cluster and basal signature values are also displayed, with basal signature scores indicating whether a sample is more basal (blue) or luminal (orange). Gene annotation is indicated at right and colored according to gene classification/transcriptional group as indicated at the bottom. **B.** Here, consensus cluster number was pre-defined as 3 to evaluate the capacity to re-discover clusters aligned with MDA class annotation. As in A, sample and gene annotation is colored according to the legend at the bottom.

Figure D3 – Correlation matrix for all targets on custom bladder targeted RNAseq panel

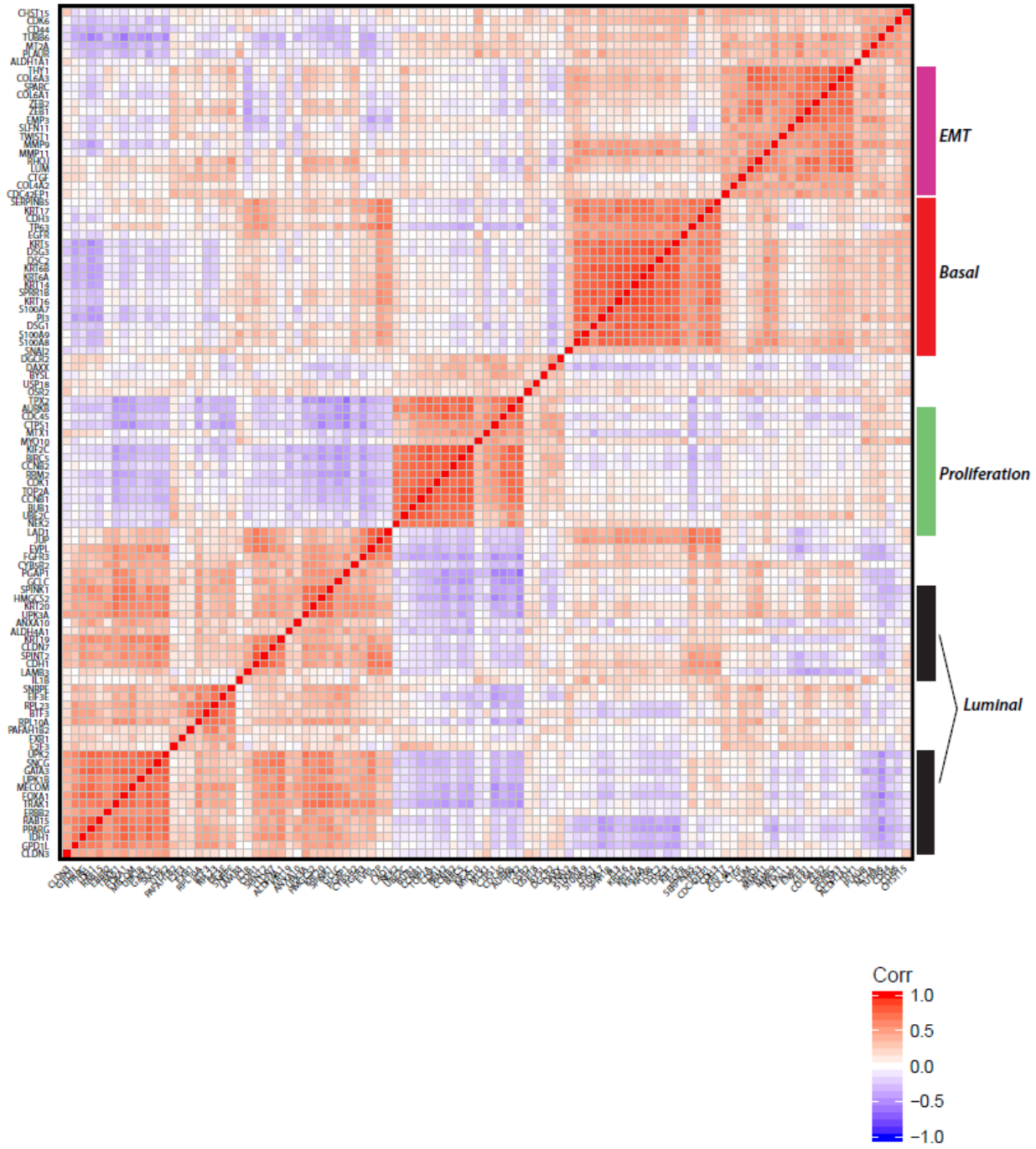


Figure D3 - Pearson correlation matrix for expression values from 77 high-quality tissue specimens across all 103 targets on a custom bladder targeted RNA sequencing panel. Major transcriptional modules assessed on this panel (e.g., proliferation, basal, luminal) are represented by highly inter-correlated markers as annotated at right. Genes are ordered by hierarchical clustering distance.

<u>Figure D4</u> – Unsupervised clustering of normalized log2 expression values from all non-housekeeping gene targets and 77 high-quality tissue specimens profiled on our custom targeted RNAseq panel

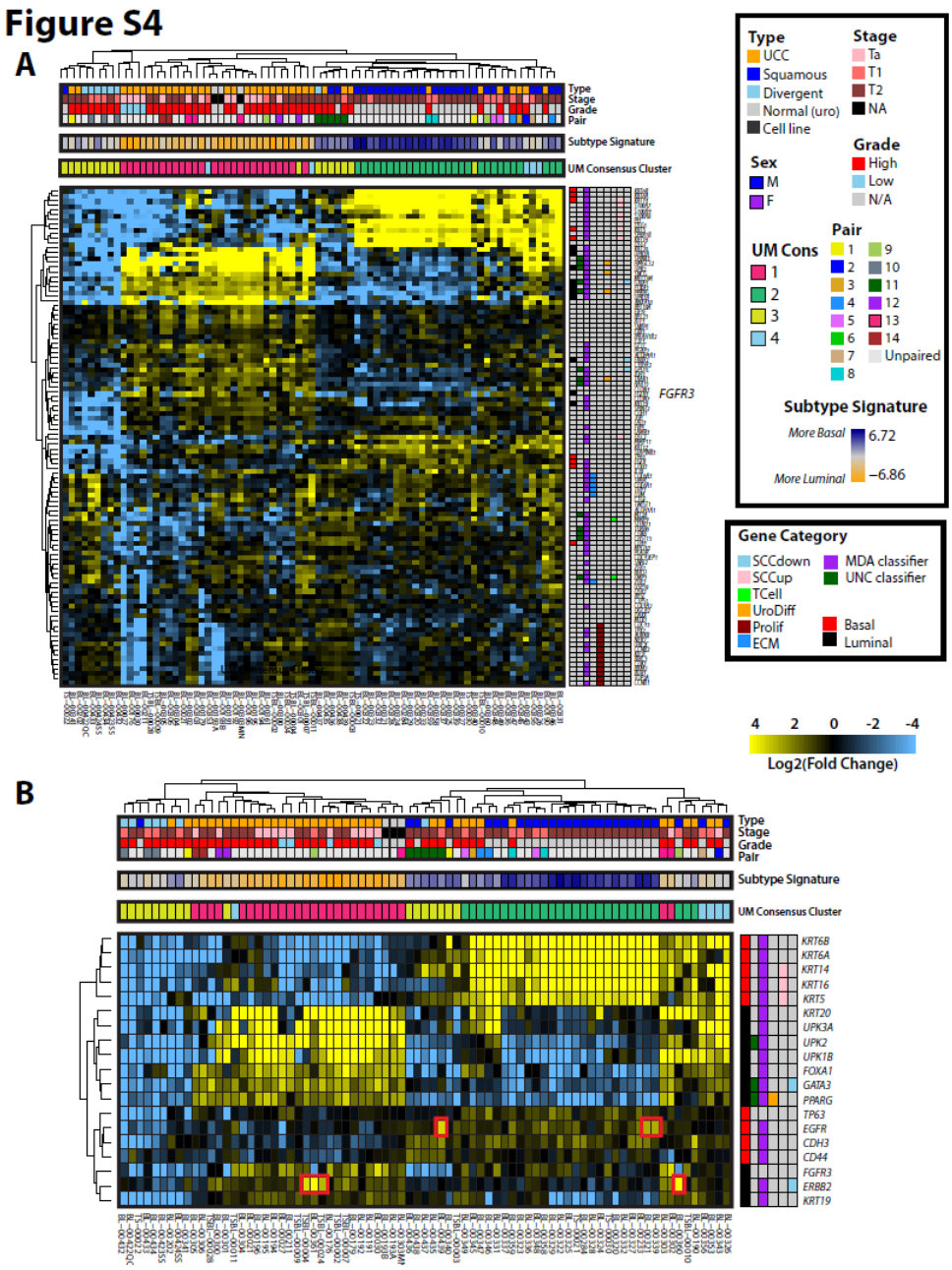


Figure D4 A. Unsupervised clustering of normalized log2 expression values from all non-housekeeping gene targets and 77 high-quality tissue specimens profiled on our custom targeted RNAseq panel. Sample annotation (header annotation rows at top) is colored corresponding to annotations contained the figure legend, while target annotation (at right) is colored according to gene category annotations provided. **B.** Unsupervised clustering of normalized log2 expression values from select basal/luminal genes and 77 high-quality profiled tissue specimens enables delineation of individual gene target expression, and highlights substantially elevated expression of *ERBB2* and *EGFR* in samples with focal copy-number amplifications.

Figure D5 – Unsupervised clustering of normalized log2 expression values from all non-housekeeping gene targets for 98 high-quality tissue specimens and cell lines profiled on a custom targeted RNAseq panel

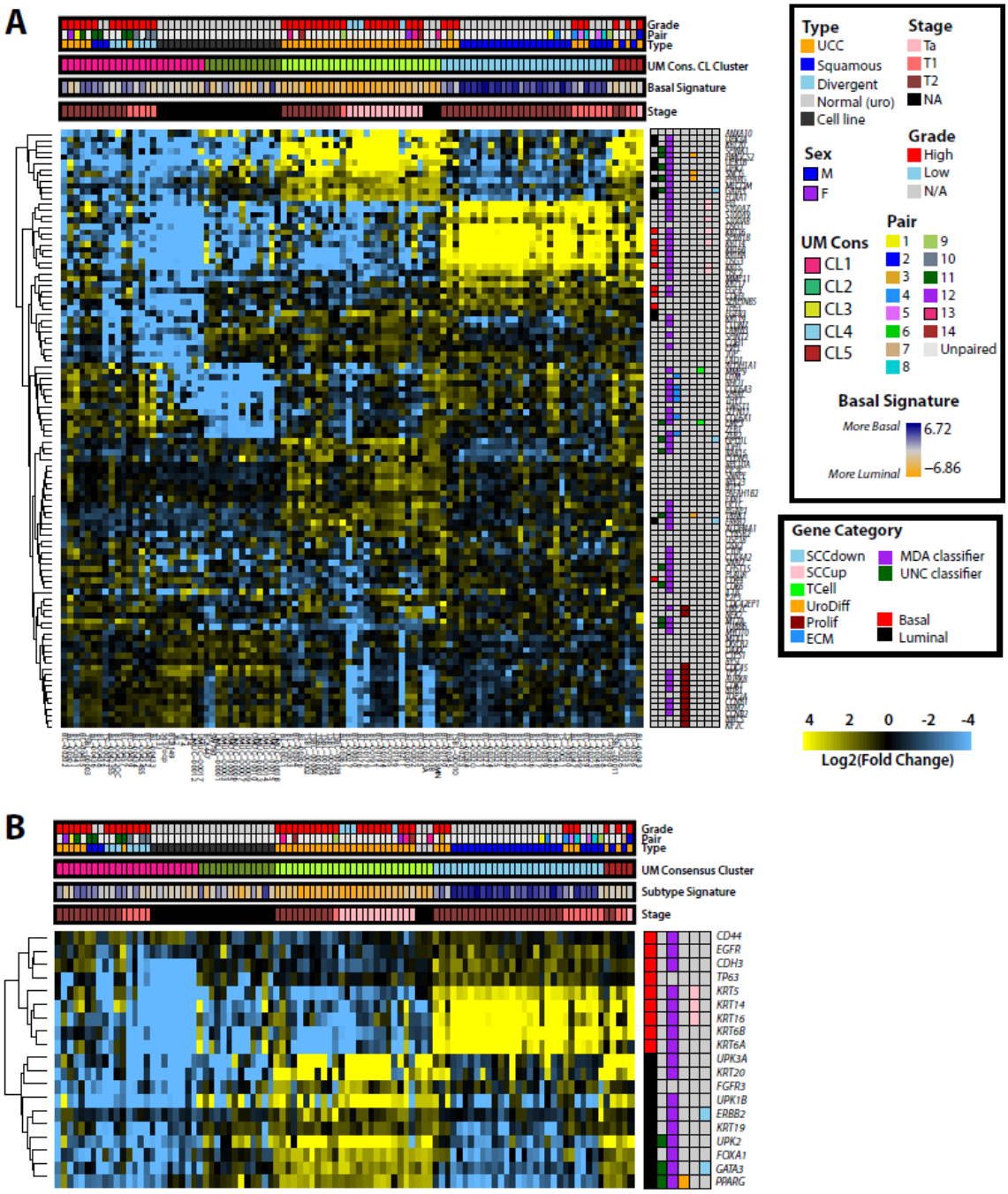


Figure D5 A. Unsupervised clustering of normalized log2 expression values from all non-housekeeping gene targets for 98 high-quality tissue specimens and cell lines profiled on our custom targeted RNAseq panel. Samples are sorted left to right by consensus cluster, then stage, then histological subtype. Sample annotation (header annotation rows at top) is colored corresponding to annotations contained the figure legend, while target annotation (at right) is colored according to gene category annotations provided. **B.** Unsupervised clustering of normalized log2 expression values from select basal/luminal genes for 77 high-quality profiled tissue specimens enables delineation of individual gene target expression, and highlights substantially elevated expression of ERBB2 and EGFR in samples with focal copy-number amplifications.

Figure D6 – Copy-number heatmap for bladder tissue and cell line samples

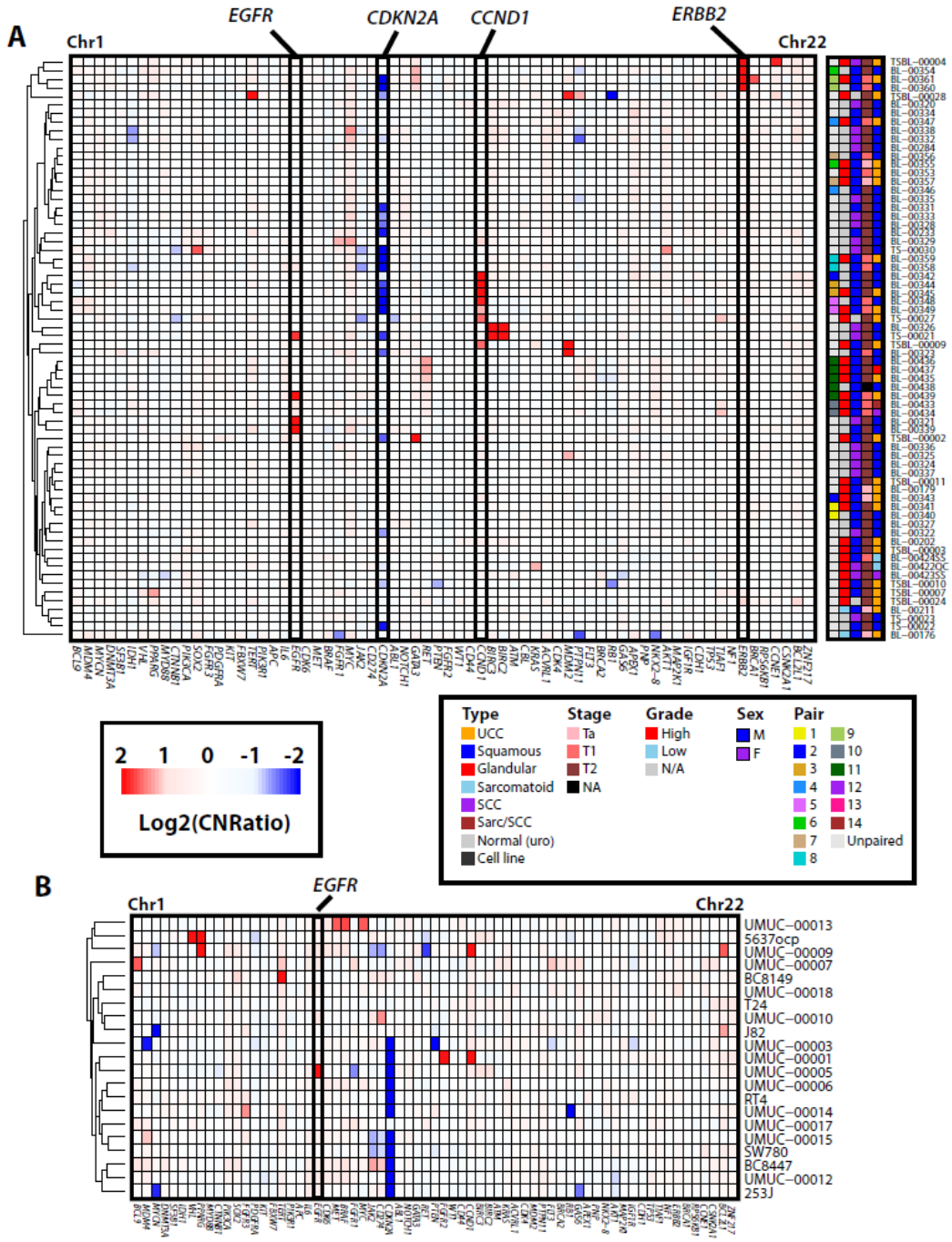


Figure D6 Unsupervised hierarchical clustering of gene-level copy-number ratios for 77 tissue specimens (**A**) and 21 cell lines (**B**) with high-quality DNA profiled by targeted DNA sequencing. Gene-level copy-number ratios are displayed left to right in genome order, and and sample annotation (including cancer type, stage, grade, sex, and case (if relevant)) are colored at right as indicated in the legend. Focal alterations of relevance are highlighted in both plots.

Figure D7 – Validation of sub-gene RB1 copy-number deletion in UMUC-14 bladder cancer cell line

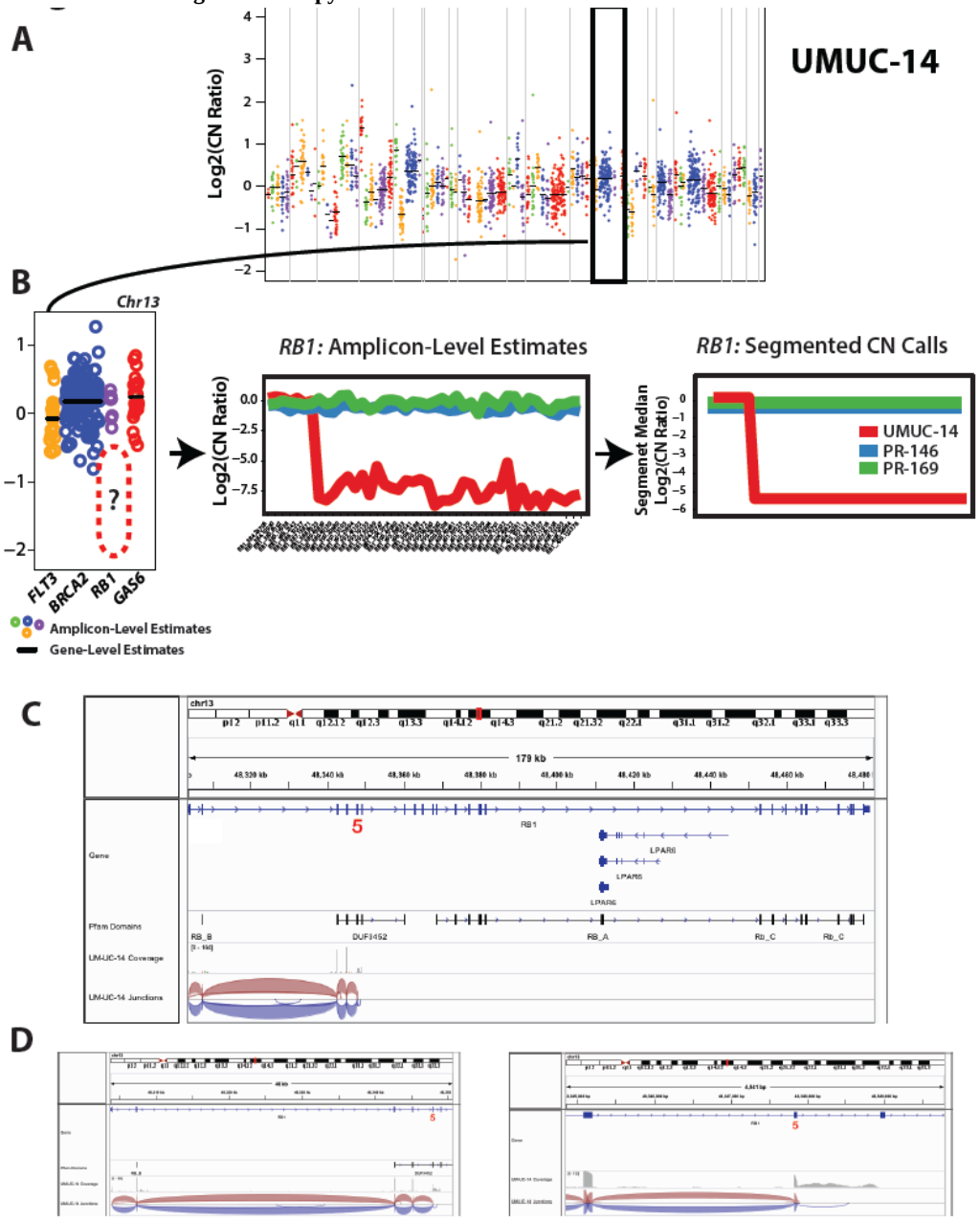


Figure D7 A. Genome-wide copy-number plot from targeted DNA sequencing of urothelial cancer cell line UMUC-14. Individual dots represent amplicon-level log2 copy-number ratio estimates, with horizontal black lines representing log2 gene-level copy-number ratio estimates. Black rectangle highlights portion of the plot (chr10) presented in panel B. **B.** At left, a zoomed view of amplicon- and gene-level copy number ratios on chr13 for UMUC14 demonstrates the absence of amplicon-level copy-number ratios for a subset of *RB1* target amplicons. The middle panel highlights amplicon-level copy-number ratios sorted in genome order, suggesting a sub-gene deletion affecting the majority of exons of *RB1*. At right, a sliding-window function applied to segmented copy number values from amplicon-level data, provides a smoothed, segmented sub-gene copy-number call for clinical or research reporting. **C.** Integrated Genome Viewer (IGV) screenshot of spliced read alignment data across *RB1* coding regions for conventional whole-transcriptome RNAseq data from UMUC14 shows depleted expression of all exons after exon 5. **D.** Zoomed view of conventional RNAseq data for the first five exons of RB1 and exon 5 shows limited read mapping and depleted expression of downstream exons consistent with the observed *RB1* sub-gene copy number deletion.

Figure D8 – Sub-gene copy-number deletions detected in retrospective cohort of samples from patients with prostate cancer

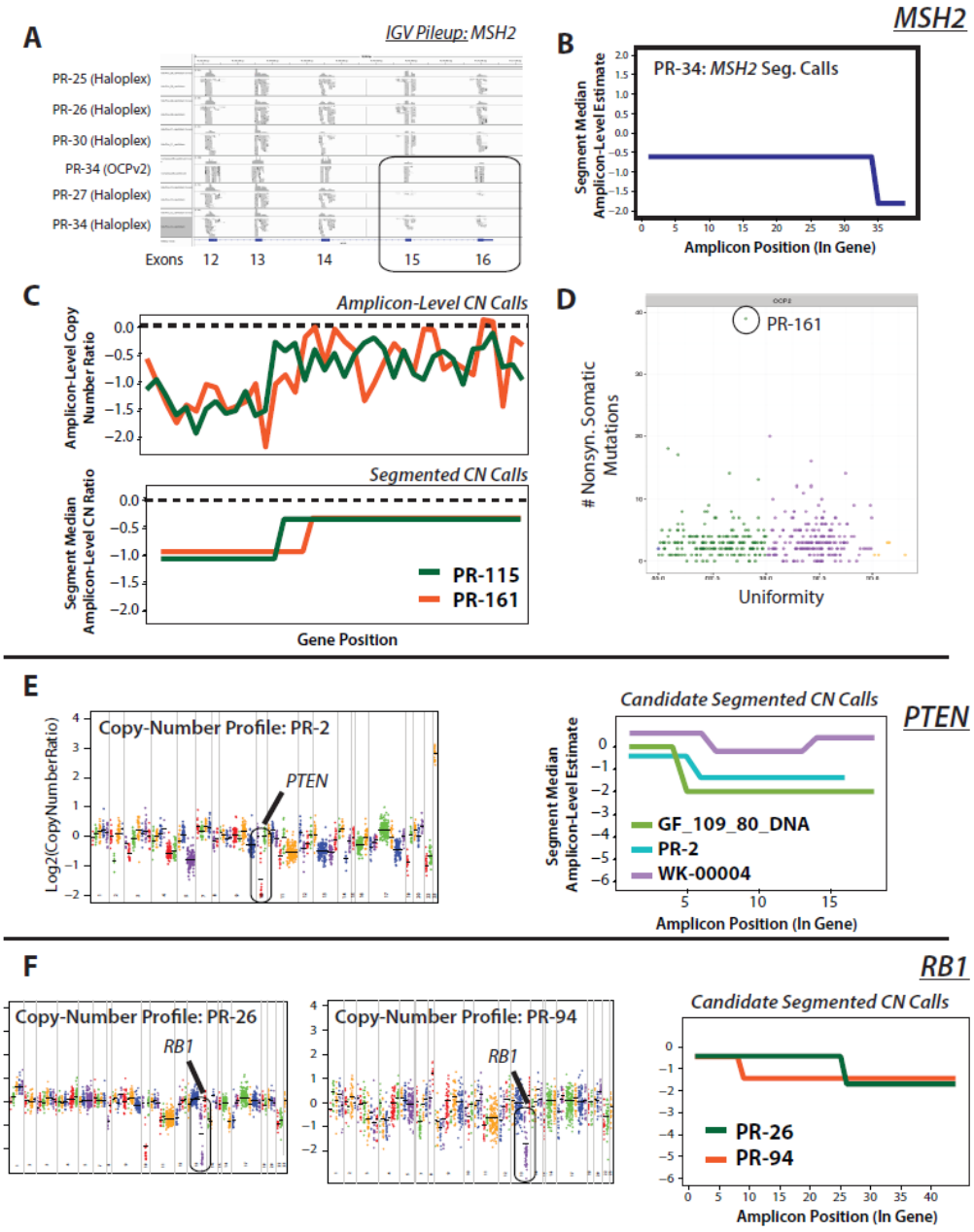


Figure D8 A. IGV pileup of capture- (Haloplex) and amplicon-based (OCPv2) targeted DNA sequencing reads for *MSH2* exons 12-16 across a series of prostate cancer samples. PR-34 (profiled by both targeted NGS approaches) shows concordant depletion of reads mapping to exons 15 and 16. **B.** Segmented sub-gene MSH2 copy-number deletion call for PR-34 highlights sub-gene copy-number deletion call. **C.** *MSH2* amplicon- and segmented CN calls cross *MSH2* coding sequencing for paired PR-115 and PR-161 prostate cancer tissue samples. **D.** Elevated prioritized nonsynonymous mutation load in PR-161 is shown, consistent with impaired mismatch repair function of *MSH2*. **E.** At left, OCP copy-number profile for PR-2, highlighting amplicon- and gene-level ratios for all targeted genes with >= 3 target amplicons (*PTEN* is highlighted). At right, candidate *PTEN* sub-gene deletion calls, including PR-2. **F.** At left, copy-number profiles for both PR-26 and PR-94, with amplicon- and gene-level calls displayed. Calls for *RB1* are circled. At right, segmented sub-gene *RB1* copy-number calls are displayed.

Figure D9 – Sub-gene copy-number deletions detected in retrospective pan-cancer cohort

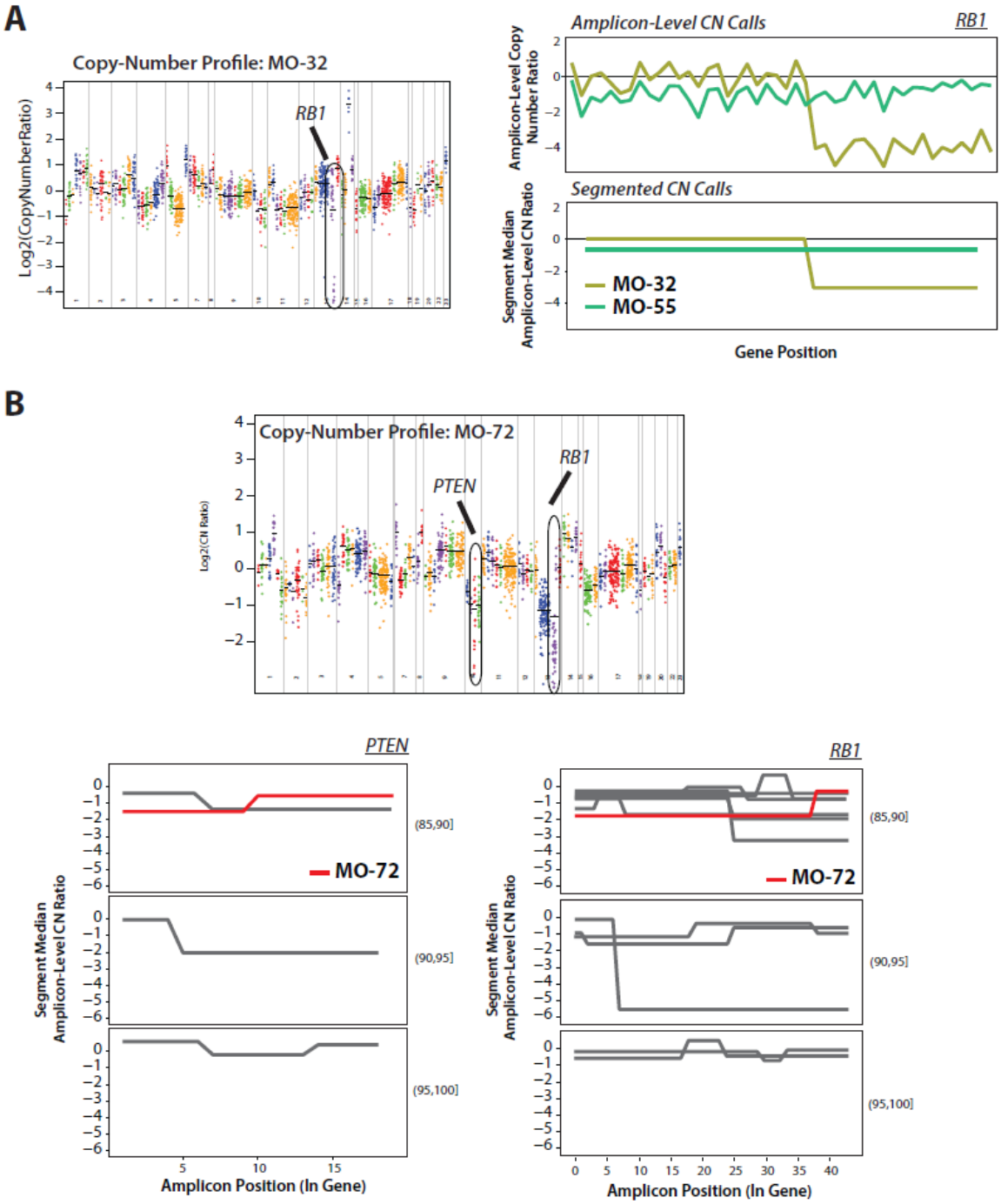


Figure D9 A. At left, OCP copy-number profile for MO-32, highlighting amplicon- and gene-level ratios for all targeted genes with >= 3 target amplicons (*RB1* is highlighted). At right, candidate *RB1* sub-gene deletion calls, including MO-32 and an unaltered sample (MO-55). **B.** At top, copy-number profiles for MO-72, with amplicon- and gene-level calls displayed. Calls for *PTEN* and *RB1* are circled. At bottom, candidate segmented sub-gene calls for *PTEN* (left) and *RB1* (right) are displayed, with calls for MO-72 highlighted. Each gene-specific graphic is separated into 3 separate panels based on sequencing uniformity (top: 85-90% uniformity; middle: 90-95% uniformity; bottom: 95-100% sequencing uniformity). .

Figure D10 – Divergent expression profiles in the context of identical genomic profiles for paired urothelial and squamous differentiation lesions from the same tumor.

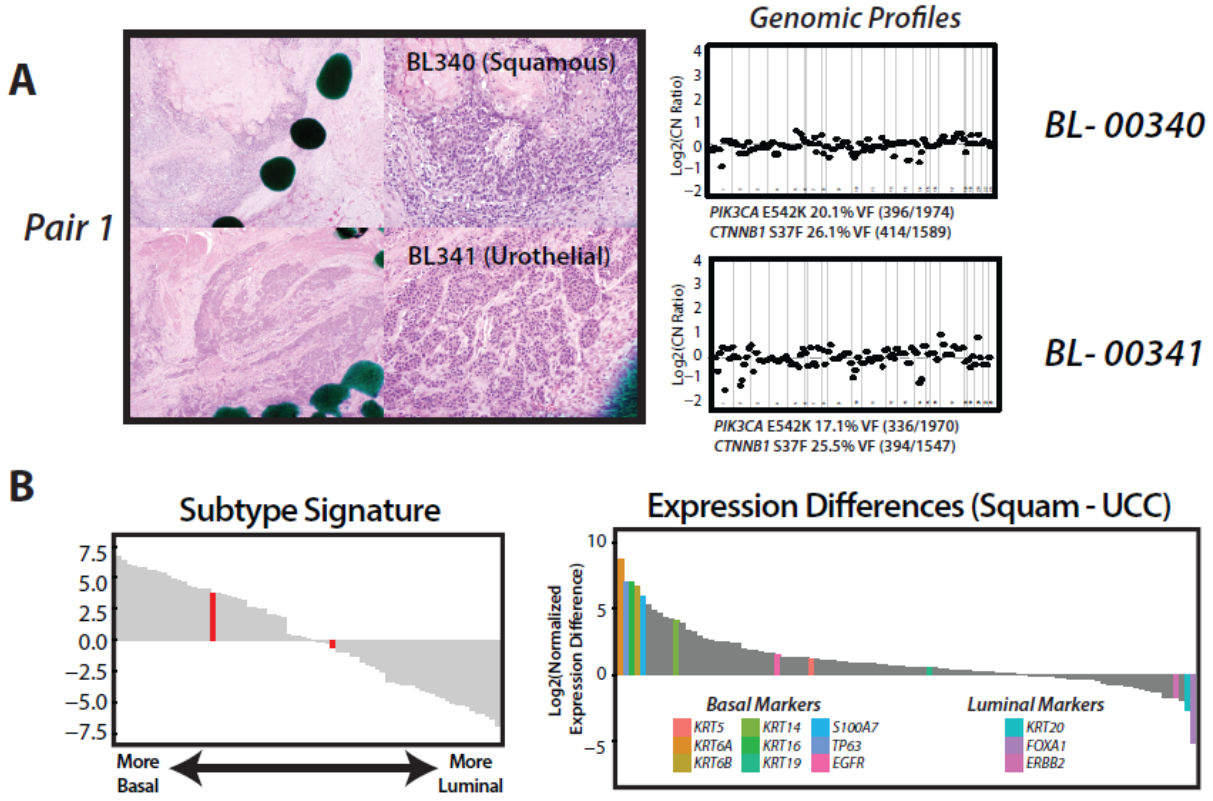


Figure D10 A. Haematoxylin and eosin staining images of individual squamous and urothelial components profiled for pair 1 are shown. At right, similar genome-wide copy-number profiles derived from targeted DNA sequencing are shown for each sample, *PIK3CA* E542K and *CTNNB1* S37F somatic point mutations seen in both samples are indicated. **B**. At left, divergent basal signature values for BL-340 and BL-341 are highlighted in red in the context of all basal signatures for profiled tissue specimens in our study. At right, individual expression differences between BL-340 and BL-341 are plotted for 103 non-housekeeping markers, with select basal or luminal markers colored according to the legend.