DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Novel computational frameworks for driver gene identification and evolutionary informed genomics analysis in melanoma and prostate cancer

Citation

Conway, Jake. 2021. Novel computational frameworks for driver gene identification and evolutionary informed genomics analysis in melanoma and prostate cancer. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Permanent link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37371185

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division of Medical Sciences

Biomedical Informatics

have examined a dissertation entitled

Novel computational frameworks for driver gene identification and evolutionary informed genomics analysis in melanoma and prostate cancer

presented by Jake Ryan Conway

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: Rameen Beroukhim

Typed Name: Dr. Rameen Beroukhim

Signature: Matthew Meyerson
Matthew Meyerson (Nov 16, 2021 11:37 EST)

Typed Name: Dr. Matthew Meyerson

Signature:

Typed Name: Dr. Nikolaus Schultz

Signature: Joshua Campbell

Typed Name: Dr. Joshua Campbell

Date: November 15, 2021

Novel computational frameworks for driver gene identification and evolutionary informed genomics analysis in melanoma and prostate cancer

A dissertation presented

by

Jake Ryan Conway

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

November 2021

© 2021 Jake Ryan Conway
All rights reserved.

Dissertation Advisor: Eliezer Van Allen Jake Ryan Conway

Novel computational frameworks for driver gene identification and evolutionary informed genomics analysis in melanoma and prostate cancer

Abstract

We performed harmonized molecular and clinical analysis on 1,048 melanoma whole-exomes and discovered markedly different global genomic properties among genomic subtypes (*BRAF*, *(N)RAS*, *NF1*, Triple Wild-Type), subtype-specific preferences for secondary driver genes, and active mutational processes previously unreported in melanoma. Secondary driver genes significantly enriched in specific subtypes reflected preferential dysregulation of additional pathways beyond MAPK, such as induction of TGF-β signaling in *BRAF* melanomas and inactivation of the SWI/SNF complex in *(N)RAS* melanomas. Additionally, select co-mutation patterns coordinated selective response to immune checkpoint blockade. We also defined the mutational landscape of Triple Wild-Type melanomas and revealed enrichment of DNA repair defect signatures in this subtype, which were associated with transcriptional downregulation of key DNA repair genes and may revive previously discarded or currently unconsidered therapeutic modalities for genomically stratified melanoma patient subsets. Broadly, harmonized meta-analysis of melanoma whole-exomes revealed distinct molecular drivers that may point to multiple opportunities for biological and therapeutic investigation. Extension of this analysis to structural variants in 355 melanoma whole-genomes revealed similar secondary driver genes among the genomic subtypes, as well as novel subtype specific drivers specifically affected by structural variants. Integration of Hi-C data also identified histology (cutaneous, acral, mucosal) specific, recurrently altered, topologically associated domain (TAD) boundaries, some of which are adjacent to TADs containing known cancer genes. Finally, the extent to which clinical and genomic characteristics relate to prostate cancer clonal architecture, tumor evolution, and

therapeutic response remains unclear. Here, we also reconstructed the clonal architecture and evolutionary trajectories of 845 prostate cancer tumors with harmonized clinical and molecular data. We show that the clonal architecture of prostate cancer tumors are associated with various clinical risk factors, and demonstrate that a novel approach to evolutionarily-informed mutational signature analysis that leverages clonal architecture can uncover additional cases of homologous recombination deficient and mismatch repair deficient tumors, link the origin of mutational signatures to the specific subclones, and has immediate therapeutic implications. Broadly, clonal architecture and evolutionary informed analysis reveal novel biological insights that are clinically actionable, and more generally may provide multiple opportunities for biological and clinical investigation.

Table of Contents

Chapter 1: Introduction

A Brief Introduction to NGS and Integrated Omics Analysis in Cancer Research

The application of next-generation sequencing (NGS) technologies towards cancer research has enabled integrative analyses of genomic, transcriptomic, and epigenomic data to comprehensively characterize tumors at both the cohort (e.g. cancer type, subtype) and individual patient levels. In 2006, The Cancer Genome Atlas (TCGA) project set out to catalogue the set of genomic alterations that defined several cancer types. Today, TCGA is comprised of over 10,000 tumor-normal exome pairs from 33 different cancer types, totaling over 400 TB of raw omics data[1]. Each cancer type in TCGA features tumor samples that have been profiled with whole-exome sequencing, bulk RNA-sequencing, microRNA sequencing, and bisulfite sequencing, and to a lesser extent whole-genome sequencing (WGS). Integration of these omics data has led to cancer type-specific genomic classifications and subtypes, as well as pan-cancer classifications of immune subtypes that have implications for immunotherapy. Although TCGA has led to many novel biological and clinical findings, the dataset has yet to be completely mined, and is still the focus of many research projects.

The Pan-Cancer Analysis of Whole Genomes (PCAWG) project has aggregated over 2,600 cancer whole-genomes, and expands the body of work published through the TCGA project[2]. Although WGS has been performed on a subset of tumor samples in TCGA, nearly all analysis performed and published through TCGA has been restricted to the exonic, or “coding”, portion of the genome. By sequencing the entire genome, PCAWG offers the advantage of exploring and characterizing the noncoding regions of the genome, in addition to coding regions. Somatic mutation calling in WGS data allows for the inspection of alterations at regulatory elements, such as promoters, enhancers and transcription factor binding sites (TFBSs)[3].

WGS also enables the identification of structural variants (SVs)[4], such as duplications and deletions not detectable through WES or copy number detection algorithms, as well as translocations and transversions. SVs can be analyzed together with copy number alterations (CNAs) to identify the presence of chromothripsis and chromoplexy[5,6]. Chromothripsis is a single complex genomic event characterized by several SVs clustered within genomic regions of oscillating copy number states across one or more chromosomes, while chromoplexy is defined as a chain of balanced (rather than oscillating) translocations. Further, SVs and CNAs identified through WGS can be superimposed with topologically associated domains (TADs) to infer their effect on the 3-dimensional organization of the genome[7]. TADs are regions of the genome where the DNA sequence within a TAD interacts with each other more often than DNA sequences outside of the TAD. The boundaries of each TAD are defined by CTCF followed by cohesin binding to DNA, and can prevent the interaction of regulatory elements, such as promoters and enhancers, with genes in adjacent TADs[8]. SVs overlapping boundaries between TADs can lead to the dysregulation of genes, even without overlapping the genes themselves, providing another molecular mechanism for cancer undetectable with WES data[9]. One example of this is enhancer hijacking, where a SV overlapping a TAD boundary allows an enhancer and gene from adjacent TADs to interact where they normally wouldn't, resulting in aberrant expression of the gene[10–12]. SVs overlapping TAD boundaries can also lead to the formation of new chromatin domains, or “neo-TADs”, which also lead to misregulation of gene expression[13].

Analyses and Methods Central to the Work of this Thesis

Mutational Significance Analysis

The rapid advancement of statistical methods and tools for analyzing NGS data has provided novel insights into the biological underpinnings of various cancer types, and the identification of therapeutic applications. Perhaps the most widely performed analysis of NGS

data is the identification of significantly mutated genes (SMGs) through mutational significance analysis. SMGs are genes that are mutated more than we would expect given some statistical test or framework. Early tests aimed at determining the importance of a gene in a cohort of cancer patients included the ratio of nonsynonymous to synonymous mutations[14], which operates under the assumption that genes important to cancer would be more likely to experience function altering mutations. As the size of cancer cohorts continued to grow, these early tests quickly became inadequate, often leading long lists of putative SMGs with several genes that have no functional role in cancer and potentially dilute the signal of true drivers. Common false positive SMGs include *TTN*, the largest gene in the human genome, and several olfactory receptor genes. To address this issue, several mutational significance algorithms have been developed to correct for additional covariates that may contribute to a gene's functional relevance in cancer. These algorithms can broadly be classified in 3 categories based on whether or not they emphasize mutational recurrence, functional impact, or sequence context.

MutSigCV2 is a popular mutational significance algorithm that emphasizes mutational recurrence, or in layman's terms, seeing the same gene mutated over and over again across samples in a cohort[15]. In addition to mutational recurrence, MutSigCV2 also corrects for the background mutation rate of the cancer type, patient specific tumor mutational burden (TMB), patient specific mutational spectrum, gene expression, and gene replication timing. Gene expression and gene replication timing were used as covariates because of their relationship to gene-specific mutation frequencies. Genes that are more highly expressed generally experience fewer mutations than lowly expressed genes, and genes in late replicating regions generally experience a higher mutation rate than genes in early replicating regions.

OncodriveFML is a mutational significance algorithm that emphasizes the functional impact of alterations on various genetic elements, including protein function, RNA structure, TFBS affinity, and microRNA targets.[16] The functional impact of alterations on each of these features is assessed using frameworks such as the combined annotation dependent depletion

(CADD) framework for genes and promoters[17], and RNAsnp for RNA secondary structure[18]. Simulations of possible sets of gene mutations across the cohort, correcting for sequence context, are used to determine if a cohort's observed alterations in a particular gene have higher functional impacts than expected. By assessing significance via functional scoring frameworks, OncodriveFML also possesses the capability to include noncoding mutations and identify putative driver noncoding regions, such as promoters, enhancers, splice intronic regions, and untranslated regions.

While MutSigCV2 and OncodriveFML correct for sequence context, the scope of this correction is limited. MutSigCV2 compartmentalizes the mutational spectrum of tumors into 1 of 6 predetermined classes (C > T, C > A, TpA > T, CpG > T, TpC > any base, miscellaneous), and OncodriveFML leverages either cohort level or sample level mutation sequence context for simulation of mutation probabilities. MutPanning is a mutational significance algorithm that corrects for many of the same covariates as MutSigCV2, but places greater emphasis on the sequence context of mutations and the mutational processes that generated them[19]. Mutpanning takes advantage of passenger mutations preferentially occuring in nucleotide contexts characteristic of the mutational processes active in the tumor, while driver mutations tend to deviate from characteristic sequence contexts and provide a signal for driver gene detection.

Mutational Signature Analysis and DNA Repair Mechanisms

Statistical advancements have also led to the concept of mutational signatures, which is based on the foundation that all somatic mutations observed in cancer are the result of some exogenous or endogenous mutational exposure. Each of these mutational exposures preferentially results in mutations in particular sequence contexts at specific frequencies, which are called “signatures”[20]. Some of the endogenous mutational processes characterized via mutational signatures include homologous recombination deficiency (HRD; associated with

mutational signature 3), APOBEC activity (associated with mutational signatures 2 and 13), and mismatch repair deficiency (MMRd; associated with mutational signatures 6, 15, 20 and 26). Exogenous exposures characterized by mutational signatures include ultraviolet (UV) mutagenesis (associated with mutational signature 7), and smoking tobacco (associated with mutational signature 4). There also exists a subset of signatures with unknown etiologies, most of which are only observed in a small subset of cancer types.

The first approach used to identify the mutational signatures present in a cohort of tumors was non-negative matrix factorization (NMF) applied to whole-exome sequencing (WES) data in 2013[21,22]. NMF allows for the decomposition of single nucleotide variants (SNVs) into a chosen number of mutational signatures based on NMF metrics and heuristics. Two common approaches for choosing the optimal number of signatures include using the cophenetic correlation coefficient, or choosing the number of signatures such that increasing the number of signatures by 1 doesn't yield a significantly better residual sum of squares error between the original matrix and product of the two decomposed matrices. Cosine similarity between each of the decomposed signatures and a set of reference signatures can be used to determine the identity of the decomposed signatures.

While NMF was the first method used to identify the presence of mutational signatures in tumor samples, there are several drawbacks to its application, specifically when WGS data is not available. WES samples generally produce too few mutations per tumor for accurate and robust mutational signature determination, requiring the mutations of several samples to be pooled together. This leads to generalizing the mutational signatures identified to all of the samples in the cohort, regardless if a subset of those signatures aren't present in certain samples. In 2016, non-negative least squares (NNLS) regression approaches for determining active mutational signatures were popularized through the development of deconstructSigs[23]. NNLS allows mutational signatures to be determined on individual samples without the need to pool mutations across a cohort, which enables the identification of mutational signatures present

at lower frequencies that may not be detectable by NMF-based methods. Additionally, NNLS circumvents the need to generalize each mutational signature identified to every sample in the cohort.

The field of mutational signature analysis is still expanding beyond the application of NNLS, particularly with a focus on clinical utility. In 2019, a novel computational method called Signature Multivariate Analysis (SigMA) was developed specifically to identify the presence of mutational signature 3 (associated with HRD) from clinical panel sequencing data, which generally results in too few mutations for accurate mutational signature detection via other existing methods[24]. SigMA deploys a cancer type specific Bayesian likelihood-based approach, as well as NMF and NNLS, followed by gradient boosting for classification. Briefly, SigMA leverages the prevalence of each mutational signature, and sample specific mutational spectrums, from a cohort of WGS samples of the same cancer type to calculate the likelihood.

Poly (ADP-ribose) polymerase inhibition (PARPi) is a common treatment for advanced breast, ovarian, and prostate cancers that exhibit signs of HRD[25]. Currently, the only FDA approved biomarker for treatment with PARPi is the presence of *BRCA1/2* inactivating alterations, and specifically in ovarian cancer the number of HRD-associated copy number events[26]. These HRD-associated copy number events are loss of heterozygosity (LoH), telomeric allelic imbalance (TAI), and large scale transition (LST) events. The association between each of these types of copy number events and HRD were found independently and in breast cancer cohorts, however their association with mutational signature 3 has held across a variety of cancer types, such as breast cancer, ovarian cancer, prostate cancer and melanoma. For each sample, SigMA provides a strict and “loose” threshold for the identification of signature 3, based on the frequency of signature 3 in the cancer type and the false positive rate (FPR) of the classification. In breast, ovarian, and prostate cancer the sensitivity of signature 3 under the strict threshold is 58.4%, 41.7%, and 64.3% with FPRs of 1.8%, 1.56%, and 4.4%, respectively.

The low FPR of signature 3 in these cancer types demonstrates the potential impact of SigMA in clinical settings.

HRD isn't the only DNA repair-associated mutational process with clinical applicability that can be detected through somatic mutation data and mutational signatures. MMRd is associated with 4 different mutational signatures (6, 15, 20 and 26), and can result in a hypermutator phenotype called microsatellite instability (MSI)[27]. MSI is characterized by DNA polymerase slippage events in microsatellite regions during replication that are usually repaired by MMR machinery. MMRd/MSI-high tumors are associated with improved response to immune checkpoint blockade (ICB) therapy across several solid tumor cancer types, and MMRd/MSI-high status is an FDA approved biomarker for treatment with the PD-L1 checkpoint blockade drug pembrolizumab[28]. The gold standard for determining MSI status in the clinic is through the use of PCR or MSI immunohistochemistry (IHC), however, these procedures can be expensive and time consuming[29]. Further, it has been postulated that the use of NGS to identify MSI status may yield additional MSI patients not identified by these methods. Two computational methods for identifying MSI status from NGS data are MSIsensor-pro[30] and MANTIS[31], both of which still perform well on samples with low sequencing coverage or tumor purity (the proportion of the sample that is tumor cells). Further, MSIsensor has been suggested as a potential diagnostic tool for screening potential MSI-high prostate cancer patients that may benefit from ICB therapy.

Phylogenetic Reconstruction

The order in which somatic mutations occur in the evolution of a tumor can be inferred through calculating their cancer cell fraction (CCF), or the fraction of cancer cells that harbor the mutation[32]. A mutation's CCF is a function of its variant allele frequency (VAF), allelic copy number at the locus, and the purity of the tumor sample (i.e. the proportion of the sample that is tumor cells). The CCF for a mutation can be assigned through maximum likelihood estimation

via a binomial distribution, choosing the CCF that results in the highest probability of observing the actual number of alternate reads (successes) out of the total number of reads sequenced at that position (trials).

Phylogenetic reconstruction algorithms allow for the investigation into the clonal architecture and evolutionary trajectories of tumors by clustering mutations into cell subpopulations, also referred to as clones, based on CCF information[33]. While it is possible for these algorithms to infer cell subpopulations from single samples, more than one temporal or spatial sample is generally recommended for accurate results. Multiple biopsies provide 2 distinct advantages for phylogenetic reconstruction, which are the reduction of noise due to repeated measurements, and the ability to identify sets of mutations whose CCFs shift together due to being in the same clone[34]. Most phylogenetic reconstruction algorithms also output phylogenetic trees based on the cellular prevalence of the inferred clones followed by the application of the pigeon-hole principle. The pigeon-hole principle simply states that the sum of the cellular prevalence of all a parent clone's subclones cannot be greater than the parent clone's cellular prevalence. One novel phylogenetic reconstruction algorithm is PhylogicNDT[35], which was developed for analysis of PCAWG consortium data, and is composed of several modules geared towards tumor evolution analysis. These modules include a clustering algorithm which utilizes a Bayesian dirichlet process with Gibbs sampling to identify cell subpopulations present in a patient's tumor, and a phylogenetic tree reconstruction algorithm which constructs of an ensemble of probable phylogenetic trees while accounting for uncertainties in cluster (clone) identity and mutation memberships.

Melanoma Clinical and Genomic Background

The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program estimated that over 90,000 new cases of melanoma were diagnosed in 2018, and despite the development of novel therapies and approaches for melanoma in the past decade,

an estimated 9,300 patients died from melanoma. Early genetic studies led to the identification of *BRAF* V600E hotpot mutations, and the development of dabrafenib and vemurafenib, which are targeted inhibitors of these mutations[36]. Due to the high mutational burden of UV mutagenesis, cutaneous melanomas are one of the most highly mutated cancer types at approximately 13 mutations per megabase (Mb)[37]. The high TMB of melanoma also makes it one of the most frequent cancer types treated with ICB therapy, where TMB > 10 mutations/Mb is a Food and Drug Administration (FDA) approved biomarker for ICB therapy. Despite biomarker status, TMB is a fairly weak predictor of response to ICB, and has been shown to be associated with varying effect sizes and statistical significance depending on cohort and cancer type[38]. Thus, the identification of novel biomarkers and associations with ICB response remains an important and open area of research[39]. The sheer number of melanoma patients receiving ICB therapy makes melanoma a rich candidate for determining mechanisms of response and resistance to ICB via NGS data, which may lead to the identification of biomarkers and patient population subsets with high likelihoods of responding to therapy.

The TCGA skin cutaneous melanoma (SKCM) study led to the classification of four melanoma genomic subtypes based on the presence of mutations in the most frequently mutated, mutually exclusive, driver genes: *BRAF*, (*N)RAS*, *NF1* and Triple Wild-Type (TWT). Approximately 50%, 25%, 15% and 10% of cutaneous melanomas are classified as *BRAF*, *(N)RAS*, TWT, and *NF1* subtypes, respectively. *BRAF* and *(N)RAS* are oncogenes that activate the MAP kinase pathway, whereas *NF1* is a negative regulator of the MAP kinase pathway and promotes the development and growth of melanoma through inactivating mutations[37,40]. TWT cutaneous melanomas, which lack somatic mutations in *BRAF*, (*N)RAS*, and *NF1*, possess very distinct genomic characteristics relative to the other genomic subtypes. Compared to other subtypes, TWT melanomas experience an enrichment of CNAs, low TMB (2.5 mutations/Mb), and have no known SMGs. Several studies have reported recurrent mutations and CNAs in KIT, but not to a level of statistical significance. The low TMB of TWT melanomas can be partially

explained by the lower contribution of UV mutagenesis (mutational signature 7) to the mutational spectrum of these tumors. Roughly 50-60% of TWT tumors present evidence of mutations due to UV mutagenesis compared to upwards of 90% in the other genomic subtypes. The lack of SMGs identified in TWT melanomas can also be explained by the low TMB of these tumors, in addition to the largest cohort of TWT melanomas analyzed only containing 46 samples[37].

Other than cutaneous melanomas, two other common histological melanoma subtypes are acral and mucosal melanomas. Acral melanomas occur on non-hair bearing skin such as the palms of the hands and soles of the feet[41], and mucosal melanomas occur in mucosal membranes such as the inside of the mouth[41,42]. While the genomic subtypes identified in cutaneous melanoma have been applied to acral and mucosal melanomas, the majority of these tumors are TWT and lack genomic alterations affecting the MAP kinase pathway. Both acral and mucosal melanomas have lower mutational burden, and higher copy number and SV burden compared to cutaneous melanomas. Furthermore, acral and mucosal melanomas frequently lack the presence of UV mutagenesis, while clock-like mutational processes, such as spontaneous deamination of cytosines, are responsible for generating the majority of mutations in these tumors. WGS of acral and mucosal melanomas has also revealed different driver alterations, such as recurrent mutations in *KIT* and *SF3B1*, and recurrent SVs affecting *TERT*, *CDK4*, *MDM2*[41,42].

Prostate Cancer Clinical and Genomic Background

Prostate cancer is the most common non-cutaneous cancer in men with over 1.4 million cases diagnosed and 381,000 deaths annually worldwide. Despite the overall high burden of this disease, the majority of men with prostate cancer die of non-related causes. Early studies aimed at the genomic classification of prostate cancer identified recurrent alterations in androgen receptor (AR) signaling, DNA repair, and phosphoinositide 3-kinases **(**PI3K)-AKT

signaling pathways, all of which are associated with more aggressive phenotypes and advanced disease[43]. Unlike melanoma, prostate cancer has one of the lowest TMBs of all cancer types at between 1-2 mutations per Mb[44]. Due to this TMB, power analysis via mutational significance algorithms suggest that only 700 prostate cancer whole-exomes are required to identify all SMGs mutated at a prevalence of at least 2% across the cancer type[45]. In 2017, Armenia *et al*. aggregated and uniformly analyzed 1,013 prostate cancer whole-exomes, over twice the size of the TCGA-PRAD cohort, with the purpose of saturating the SMG landscape and identifying novel drivers of prostate cancer[43]. This study identified 97 SMGs, 70 of which were novel and included the transcription factor *SPEN*, which functions in the AR signaling pathway.

To further elucidate the role of genomic alterations and features with prostate cancer risk, progression, and recurrence, Boutros *et al.* reconstructed the phylogenies of 293 primary prostate cancer tumor whole-genomes to determine the association between clonal architecture and tumor evolution with clinical characteristics and outcomes[46]. This study identified that SNVs in *FOXA1* and *ATM* preferentially occur as truncal alterations, while SNVs in *TP53* and *SPOP* occur clonally and subclonally. Further, Boutros *et al*. showed that about half of prostate cancer patients experience a shift in the active mutational processes generating mutations over the course of their tumor evolution. Clock-like mutational processes (signature 1) were typically observed early in tumor evolution and were associated with higher mutation CCFs, while DNA repair associated signatures (signature 3 and signature 16; associated with HRD and polymerase eta) were typically observed subclonally. Lastly, Boutros *et al*. demonstrated that polyclonality of prostate cancer tumors was associated with higher rates of biochemical recurrence, regardless of prostate cancer risk determined by gleason score, tumor stage, and prostate-specific antigen (PSA) levels.

While Boutros *et al*. provided seminal work for understanding prostate cancer tumor evolution, both patients with metastatic tumors and patients of African ancestry were not included in the analysis. Metastatic prostate cancers are associated with significantly higher

TMB and copy number burden compared to primary prostate cancers, which yields increased tumor heterogeneity[47]. Metastatic tumors also possess recurrent alterations in cancer genes at higher prevalence, which provides more power to detect clonal and subclonal driver genes than in localized disease. Additionally, metastatic prostate cancer tumors present HRD[48] and MSI[49] at higher frequencies than primary tumors, which would allow for more comprehensive analysis of these DNA repair deficiencies in this cancer type. The use of PARPi for HRD metastatic patients is currently being tested in clinical trials[50], and the prevalence and use of immunotherapy for MSI patients is being explored in metastatic castration resistant prostate cancer patients[49].

Prostate cancer patients of African ancestry are typically associated with a worse prognosis and higher rates of biochemical recurrence compared to European ancestry patients, however it is difficult to correct for the role of socioeconomic status[51–53]. It has been suggested that, in addition to socioeconomic and healthcare disparity factors, the higher tumor volume, metastasis rate and mortality following radical prostatectomy in African ancestry patients may be the result of earlier transformation to a clinically significant cancer[54]. While distinct driver genes have been identified in patients of African ancestry, whether genetic underpinnings relate to these clinical observations is poorly understood, particularly with respect to the treatment of these cancers[55]. Evolutionary informed genomic analysis may provide insight into the initiation, growth, and clinical observations of prostate cancer tumors in men of African ancestry.

Bibliography

1. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271–281.e7 (2018).

2. Consortium, T. I. P.-C. A. of W. G. & The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* vol. 578 82–93 (2020).

3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

4. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

5. Cortés-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* **52**, 331–341 (2020).

6. Voronina, N. *et al.* The landscape of chromothripsis across adult cancer types. *Nat. Commun.* **11**, 2320 (2020).

7. Dekker, J. & Heard, E. Structural and functional diversity of Topologically Associating Domains. *FEBS Lett.* **589**, 2877–2884 (2015).

8. Barrington, C. *et al.* Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nat. Commun.* **10**, 2908 (2019).

9. Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* **52**, 294–305 (2020).

10. Haller, F. *et al.* Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. *Nat. Commun.* **10**, 368 (2019).

11. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).

12. Montefiori, L. E. *et al.* Enhancer Hijacking Drives Oncogenic Expression in Lineage-Ambiguous Stem Cell Leukemia. *Cancer Discov.* (2021) doi:10.1158/2159-8290.CD-21-0145.

13. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

14. Jeffares, D. C., Tomiczek, B., Sojo, V. & dos Reis, M. A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. *Methods in Molecular Biology* 65–90 (2015) doi:10.1007/978-1-4939-1438-8_4.

15. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

16. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).

17. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

18. Sabarinathan, R. *et al.* RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum. Mutat.* **34**, 546–556 (2013).

19. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat.*

Genet. **52**, 208–218 (2020).

20. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

21. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).

22. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).

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

24. Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).

25. Keung, M. Y. T., Wu, Y. & Vadgama, J. V. PARP Inhibitors as a Therapeutic Agent for Homologous Recombination Deficiency in Breast Cancers. *J. Clin. Med. Res.* **8**, (2019).

26. Sztupinszki, Z. *et al.* Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ Breast Cancer* **4**, 16 (2018).

27. Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073–2087.e3 (2010).

28. Chang, L., Chang, M., Chang, H. M. & Chang, F. Microsatellite Instability: A Predictive Biomarker for Cancer Immunotherapy. *Appl. Immunohistochem. Mol. Morphol.* **26**, e15–e21 (2018).

29. Samowitz, W. S., Broaddus, R., Iacopetta, B. & Goldblatt, J. PCR versus immunohistochemistry for microsatellite instability. *The Journal of molecular diagnostics: JMD* vol. 10 181–2; author reply 181 (2008).

30. Jia, P. *et al.* MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability. *Genomics Proteomics Bioinformatics* **18**, 65–71 (2020).

31. Kautto, E. A. *et al.* Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* **8**, 7452–7463 (2017).

32. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).

33. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

34. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat.*

Methods **11**, 396–398 (2014).

35. Leshchiner, I. *et al.* Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. doi:10.1101/508127.

36. Eberlein, T. J. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *Yearbook of Surgery* vol. 2012 353–356 (2012).

37. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696 (2015).

38. Litchfield, K. *et al.* Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614.e14 (2021).

39. Conway, J. R., Kofman, E., Mo, S. S., Elmarakeby, H. & Van Allen, E. Genomics of response to immune checkpoint therapies for cancer: implications for precision medicine. *Genome Med.* **10**, 93 (2018).

40. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).

41. Newell, F. *et al.* Whole-genome sequencing of acral melanoma reveals genomic complexity and diversity. *Nat. Commun.* **11**, 5259 (2020).

42. Newell, F. *et al.* Whole-genome landscape of mucosal melanoma reveals diverse drivers and therapeutic targets. *Nat. Commun.* **10**, 3163 (2019).

43. Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).

44. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).

45. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

46. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* **173**, 1003–1013.e15 (2018).

47. Brady, L. *et al.* Inter- and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. *Nat. Commun.* **12**, 1426 (2021).

48. Lotan, T. L. *et al.* Homologous recombination deficiency (HRD) score in germline BRCA2-versus ATM-altered prostate cancer. *Mod. Pathol.* **34**, 1185–1193 (2021).

49. Abida, W. *et al.* Analysis of the Prevalence of Microsatellite Instability in Prostate Cancer and Response to Immune Checkpoint Blockade. *JAMA Oncol* **5**, 471–478 (2019).

50. Teyssonneau, D. *et al.* Prostate cancer and PARP inhibitors: progress and challenges. *J. Hematol. Oncol.* **14**, 51 (2021).

51. Deka, R. *et al.* Association Between African American Race and Clinical Outcomes in Men Treated for Low-Risk Prostate Cancer With Active Surveillance. *JAMA* **324**, 1747–1754 (2020).

52. Powell, I. J. *et al.* Prostate cancer biochemical recurrence stage for stage is more frequent among African-American than white men with locally advanced but not organ-confined disease. *Urology* vol. 55 246–251 (2000).

53. Freedland, S. J., Jalkut, M., Dorey, F., Sutter, M. E. & Aronson, W. J. Race is not an independent predictor of biochemical recurrence after radical prostatectomy in an equal access medical center. *Urology* vol. 56 87–91 (2000).

54. Powell, I. J., Bock, C. H., Ruterbusch, J. J. & Sakr, W. Evidence supports a faster growth rate and/or earlier transformation to clinically significant prostate cancer in black than in white American men, and influences racial progression and mortality disparity. *J. Urol.* **183**, 1792–1796 (2010).

55. Rayford, W. *et al.* Comparative analysis of 1152 African-American and European-American men with prostate cancer identifies distinct genomic and immunological differences. *Commun Biol* **4**, 670 (2021).

Chapter 2: Melanoma Exome Meta-Analysis

Abstract

We performed harmonized molecular and clinical analysis on 1,048 melanomas and discovered markedly different global genomic properties among subtypes (*BRAF*, *(N)RAS*, *NF1*, Triple Wild-Type), subtype-specific preferences for secondary driver genes, and active mutational processes previously unreported in melanoma. Secondary driver genes significantly enriched in specific subtypes reflected preferential dysregulation of additional pathways, such as induction of TGF-β signaling in *BRAF* melanomas and inactivation of the SWI/SNF complex in *(N)RAS* melanomas, and select co-mutation patterns coordinated selective response to immune checkpoint blockade. We also defined the mutational landscape of Triple Wild-Type melanomas and revealed enrichment of DNA repair defect signatures in this subtype, which were associated with transcriptional downregulation of key DNA repair genes and may revive previously discarded or currently unconsidered therapeutic modalities for genomically stratified melanoma patient subsets. Broadly, harmonized meta-analysis of melanoma whole-exomes revealed

distinct molecular drivers that may point to multiple opportunities for biological and therapeutic investigation.

Introduction

Genomic characterization of melanoma led to the classification of four subtypes based on mutations in the most frequently mutated, mutually exclusive, driver genes: *BRAF*, *(N)RAS*, *NF1* and Triple Wild-Type (TWT)[1-2]. These studies, the largest of which included 333 melanomas[2], have augmented our understanding of the melanoma genomic landscape, informed development of effective therapies with targeted agents[3-4] and enabled molecular stratification strategies for immune checkpoint blockade[5–7]. Still, only a subset of patients exhibit durable responses to therapies targeting these known genetic vulnerabilities. Furthermore, while cancer immunotherapy has revolutionized clinical management of advanced melanoma, only a subset of patients respond to these agents and new molecular targets remain a great clinical need[8].

Identification of new molecular targets is challenging in melanoma due to the extremely high mutational load compared to most solid tumors, which is largely attributed to UV mutagenesis. As a result, power analysis has estimated that thousands of samples are required to saturate the landscape of significantly mutated genes (SMGs) in melanoma[9]. Additionally, while *BRAF*, *(N)RAS*, and *NF1* mutants all converge on MAP kinase signaling, each of these melanoma subtypes is associated with distinctive clinical characteristics, outcomes and immune profiles suggesting molecular differences that could be informed by systematic characterization in sufficiently large patient cohorts[1-2,10-12]. Further, there has been no definitive molecular dissection of TWT melanomas for unbiased gene discovery.

We hypothesized that expanded and harmonized molecular analysis of a larger cohort of melanomas would reveal new genetic drivers within and among these canonical genomic subtypes, and therapeutic vulnerabilities in genomically stratified patient subsets. Thus, we

harmonized 1,048 melanoma tumor and matched germline whole-exome sequencing (WES) samples[1-2,6-7,13-18] and performed uniform molecular analyses across and within established genomic subtypes to redefine the molecular properties that coordinate in heterogeneous melanoma patient populations.

Results

Significantly Mutated Genes in Melanoma

In total, we aggregated and uniformly analyzed WES data from 1,048 melanomas with matched germline samples that passed joint quality control parameters (Methods, Supplementary Figure 2.1, Supplementary Tables 2.1-2.3, Supplementary Data 2.1). Our cohort was comprised of 494 *BRAF*, 290 *(N)RAS*, 102 *NF1* and 162 TWT melanomas, with 5% of all melanomas having acral or mucosal origin. Additional histology information and raw sequencing metrics can be found in Supplementary Tables 2.2-2.3. The median nonsynonymous mutational load for the entire cohort was 7.94 mutations/Mb, and was significantly higher in the cutaneous melanomas compared to acral and mucosal melanomas (8.23 mut/Mb vs 1.87 mut/Mb; Mann-Whitney U, $p = 1.01 \times 10^{-15}$; Figure 2.1A).

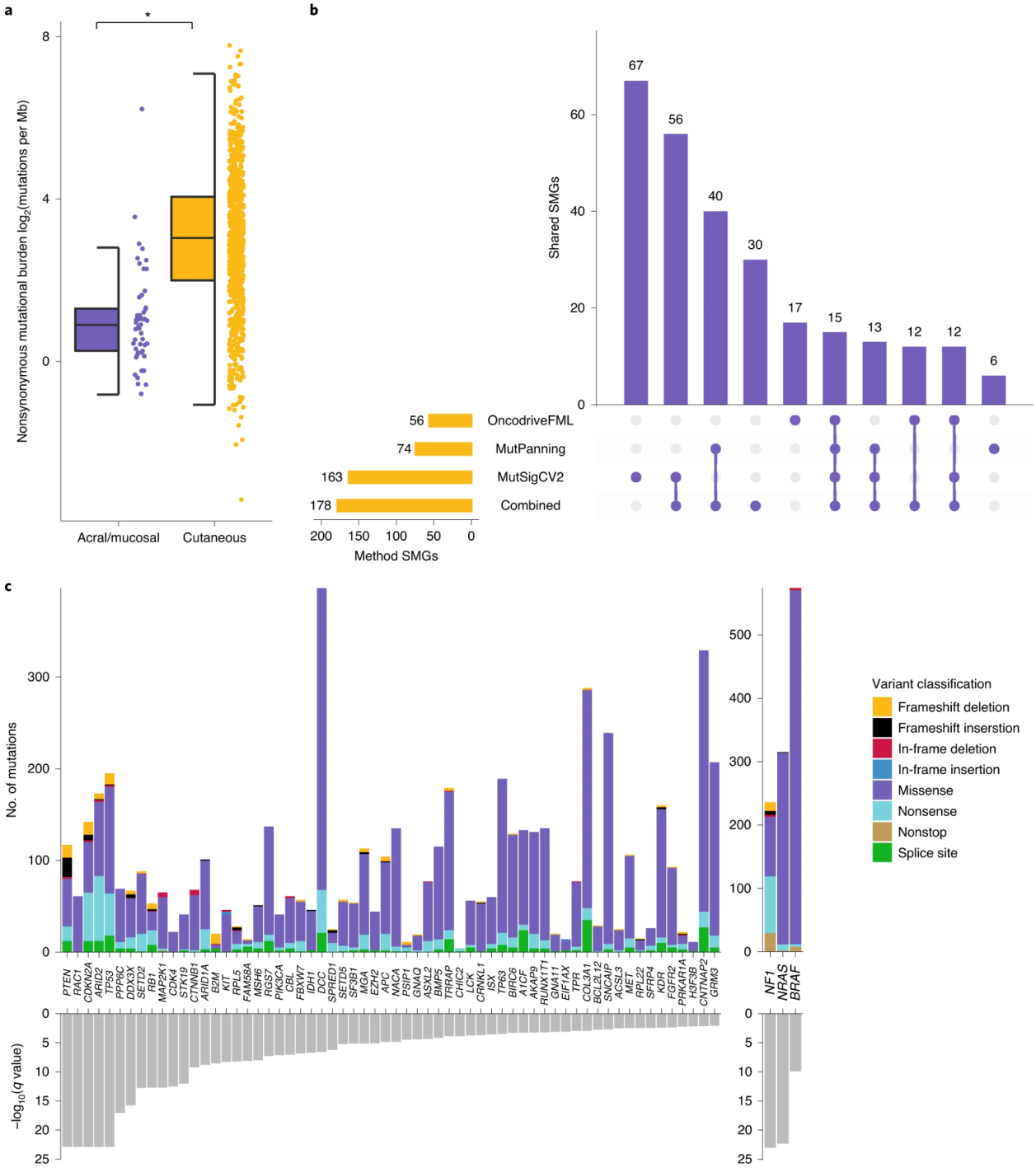


Figure 2.1: Identification of consensus driver genes in melanoma.

a) Nonsynonymous mutational load is significantly elevated in cutaneous ($n = 871$) melanomas compared with acral ($n = 34$) and mucosal ($n = 17$) melanomas (Mann–Whitney U-test, $P = 9.79 \times 10-16$, two sided). The data are represented as a boxplot where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) ×1.5 and beyond the whiskers are outlier points. An asterisk denotes a Mann–Whitney U-test $P < 0.05$. **b)** The overlap between SMGs identified by

Figure 2.1 (continued): each mutational significance algorithm (Benjamini–Hochberg, *q*-value cutoff <0.05), and when combining the *P* values via Brown's method (Benjamini–Hochberg, *q*-value cutoff <0.05). **c)** The distribution of mutation types in melanoma SMGs that are known cancer genes (CGC and OncoKB genes), ordered by statistical significance from left to right.

The statistical challenge of identifying cancer driver genes becomes increasingly difficult in cancers with high background mutation rates like melanoma[9,19]. To identify high-confidence melanoma driver genes, we utilized three orthogonal mutational significance algorithms that emphasize mutational recurrence, sequence context, and accumulated functional impact (MutSig2CV, MutPanning, and OncodriveFML respectively)[9,19-21]. We next applied Brown's method to combine the p-values from each mutational significance algorithm, followed by a strict FDR cutoff ($q < 0.01$) and consideration of transcriptional activity in bulk and single cell melanoma transcriptomes to evaluate SMGs by lineage and potential function (Supplementary Figure 2.1-2.3, Methods), to reduce the number of false positive findings. This process yielded a set of 178 genes (excluding *BRAF*, *(N)RAS*, and *NF1*; Figure 2.1B, Supplementary Figures 2.2-2.5). When restricting to known cancer genes, 46 genes were present in both the COSMIC Cancer Gene Census (CGC v86) and OncoKB, while 10 and 6 genes were only present in the CGC and OncoKB, respectively (Figure 2.1C)[22]. A total of 157 novel candidate melanoma SMGs were identified through this set of high-confidence driver genes (Supplementary Data 2.2), 41 (26%) of which are known cancer genes. These novel SMGs have been experimentally implicated in MAPK signaling and therapeutic response (e.g. *FGFR2* and *LCK*)[23], tumor-intrinsic mediators of cancer immunotherapy (e.g. *ARID1A*, *ASXL2*, *B2M*, *BRD7* and *SETD2*)[24], and oncogenesis in other cancer types (e.g. *CDK4* and *MSH6*)[25-26] (Supplementary Table 2.4). Only 32 of the 83 SMGs previously identified in large melanoma studies (cohort size > 100), were classified as SMGs by any algorithm in our cohort (Benjamini-Hochberg q-value cutoff < 0.1, Supplementary Table 2.5)[1-2,10,17].

Significantly Mutated Genes in Melanoma Genomic Subtypes

The median nonsynonymous mutational load varied widely between genomic subtypes (Mann-Whitney U, $p < 3.82 \times 10^{-8}$ for all pairwise), ranging from 2.06 mutations/Mb in TWT melanomas to 32.29 mutations/Mb in *NF1* melanomas (Figure 2.2A), which is consistent with previous findings[1-2,10,17]. *(N)RAS* melanomas experienced a higher ratio of clonal mutations relative to the other genomic subtypes (Mann-Whitney U, $p < 1.44 \times 10^{-4}$ for all pairwise; Methods), whereas TWT melanomas experienced an elevated ratio of subclonal mutations (Mann-Whitney U, $p < 0.014$ for all pairwise). Further, *BRAF* and *(N)RAS* melanomas frequently had more clonal mutations than subclonal mutations, while *NF1* and TWT melanomas had more subclonal mutations than clonal mutations (Supplementary Figure 2.6). These findings could not be explained by the differences in tumor purity between the genomic subtypes (Kruskal-Wallis, $p = 0.23$).

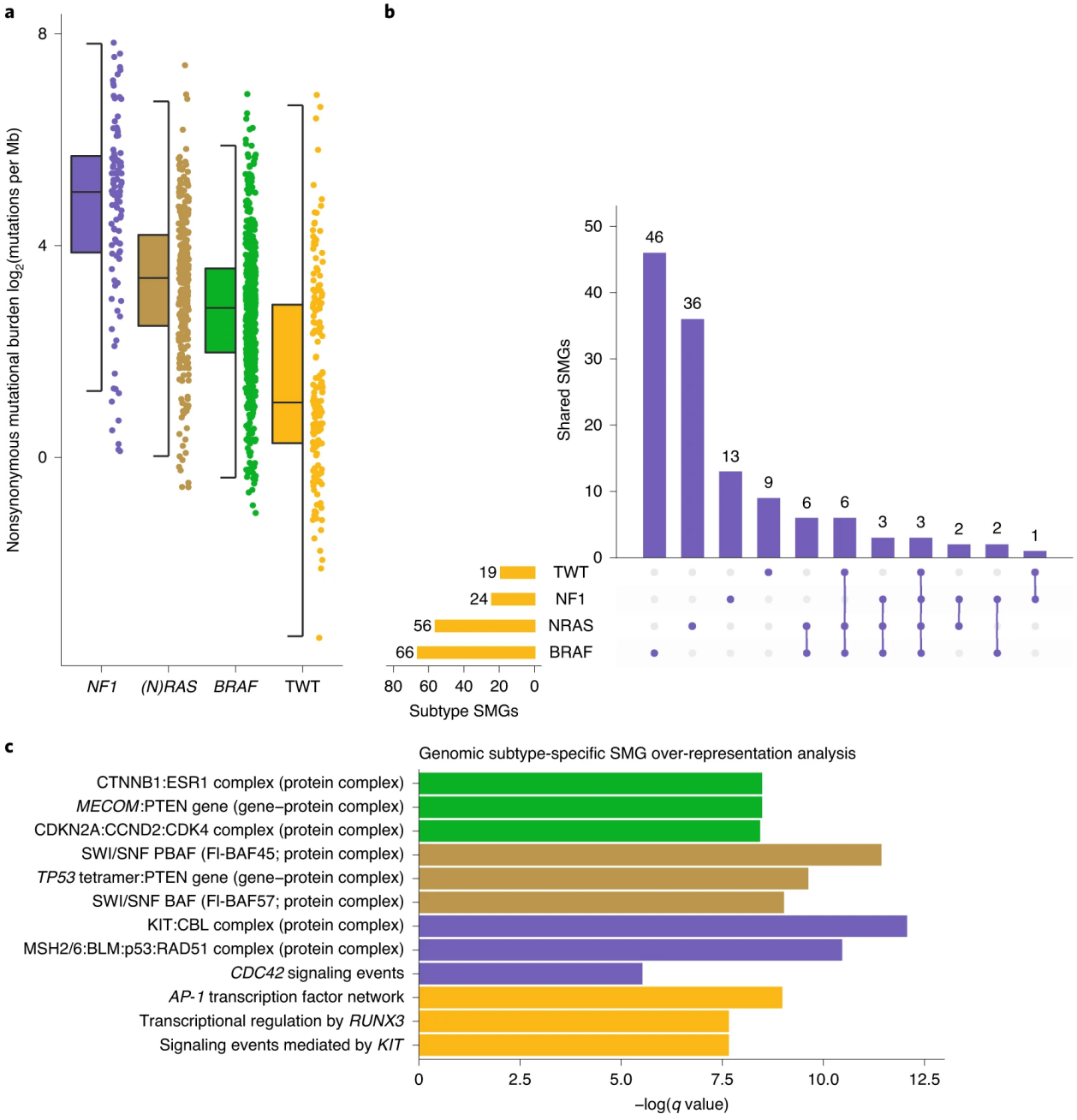


Figure 2.2: Melanoma genomic subtypes have distinct global properties and secondary driver genes.

a) The nonsynonymous mutational load was significantly different between the genomic subtypes (Mann–Whitney *U*-test, $P < 3.82 \times 10{-}8$ for all pairwise, two sided). *NF1* melanomas experienced the largest mutational load, whereas TWT melanomas experienced the lowest mutational load. The data are represented as a boxplot where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the IQR ×1.5 and beyond the whiskers are outlier points. **b)** We identified 66, 56, 24 and 19 SMGs in *BRAF*, *(N)RAS*, *NF1* and TWT melanomas, respectively. Overlapping of the genomic subtype SMGs revealed that genomic subtypes seldom share the same SMGs despite *BRAF*, *(N)RAS* and *NF1* all converging on the MAPK pathway. Specifically, 70% (46/66), 64% (36/56), 54%

Figure 2.2. (continued): (13/24) and 47% (9/19) of the SMGs identified in *BRAF*, *(N)RAS*, *NF1* and TWT melanomas were exclusive to their respective subtypes. In aggregate, only 18% (23/127) of the SMGs identified through the genomic subtype mutational significance analysis were found in more than one genomic subtype. **c)** The top three nongeneric hits (via the *q* value, Benjamini–Hochberg) from pathway and protein–complex over-representation analysis of SMGs specific to each genomic subtype. This analysis revealed several recurring patterns in *BRAF* (for example, cell cycle), *(N)RAS* (for example, chromatin remodeling), *NF1* (for example, DNA damage) and TWT (for example, *RUNX3* and *KIT* signaling) melanomas.

Due to the high mutational load of melanoma, it is unlikely that mutations in specific genes and pathways are restricted to genomic subtypes. However, we hypothesized that specific genes and pathways may be mutated more than expected or preferentially overrepresented in a subtype specific context, despite *BRAF*, *NRAS* and *NF1* converging on the MAP kinase pathway. Indeed, mutational significance analysis within each of the genomic subtypes revealed that candidate SMGs were seldom shared between the subtypes (Figure 2.2B, Extended Data 2.1), and putative function altering mutations in related pathways and protein complexes were significantly associated within those same genomic subtypes (Figure 2.2C). This suggests dysregulation of these pathways may act as co-drivers at different frequencies dependent on the genomic subtype necessitating subtype-specific significance analyses, and that subtype-specific analyses may provide further insights into additional SMGs irrespective of overall co-mutation patterns. Thus, we performed sub-type specific significance analyses to dissect the co-driver dysregulation patterns.

BRAF-mutant

A total of 66 SMGs were identified in *BRAF* melanomas, which included previously established co-mutations (i.e. *PTEN*, Figure 2.2B-C, Supplementary Figures 2.7-2.8). Two of the more frequent *BRAF* co-mutators*, MECOM* and *BMP5* (24.7%, Figure 2.2C, Supplementary Figure 2.9), have been associated with several immune related pathways (e.g. TGF-β, IFN-α, epigenetic modification)[27-30]. Notably, within the subset of melanomas that received

immunotherapy (n = 297), *BRAF* and *MECOM/BMP5* co-mutated melanomas demonstrated improved clinical benefit compared to *MECOM/BMP5* wild-type *BRAF* melanomas (Methods; 77.3% vs. 35.5%, Fisher's exact test, p = 6.4 x 10^{-4}, Figure 2.3A), even when correcting for mutational load, tumor purity, and treatment (logistic regression, p = 0.018). When restricting to *MECOM/BMP5* mutated melanoma, *MECOM/BMP5*-mutant *BRAF* melanomas were significantly associated with improved clinical benefit compared to *MECOM/BMP5*-mutant non-*BRAF* melanomas (77.3% vs. 46%, Fisher's exact test, p = 0.02, Figure 2.3A), further nominating *MECOM* and *BMP5* as subtype-specific mediators of immunotherapy response in melanoma.

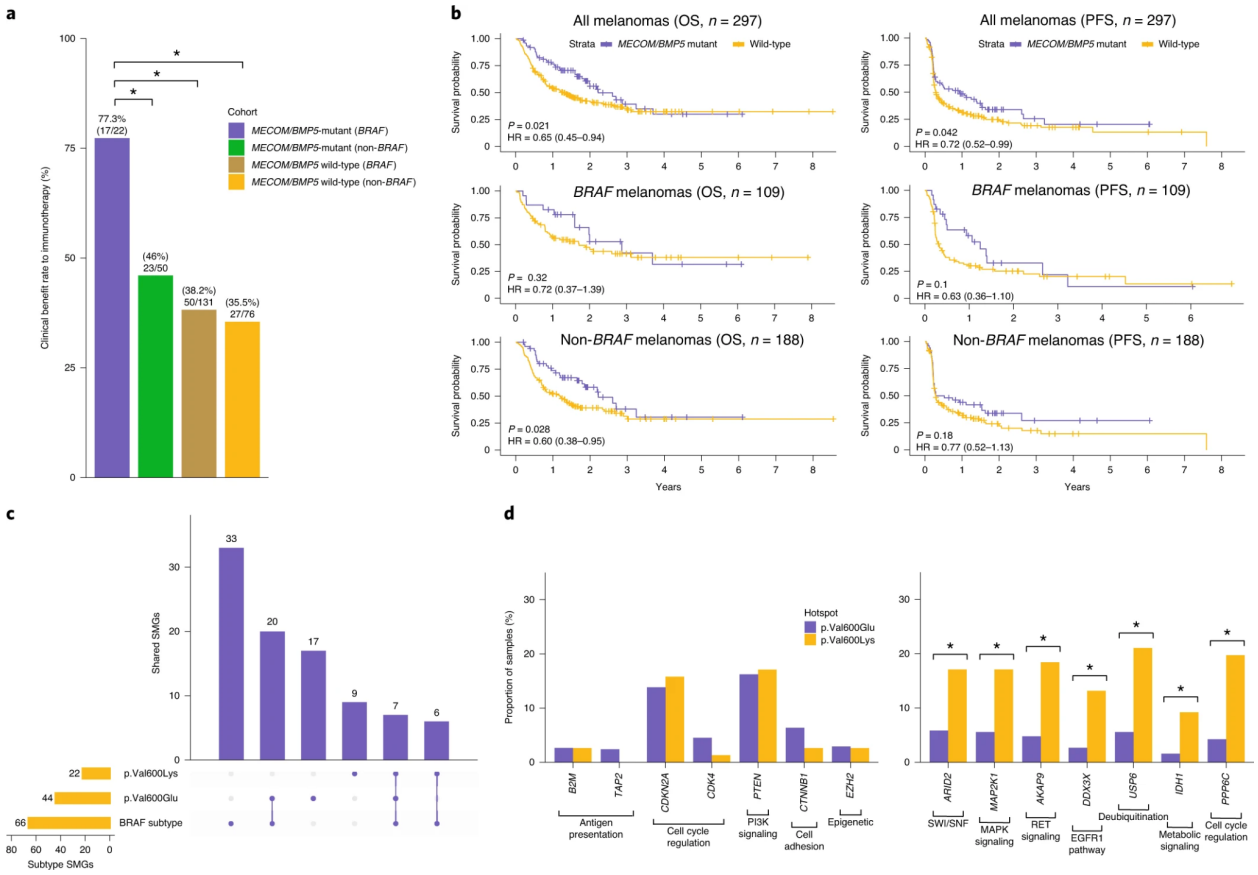


Figure 2.3: SMGs exclusive to *BRAF* melanomas have implications for immunotherapy, and secondary drivers further segregate with p.Val600Glu– and p.Val600Lys–encoding hotspot mutations.

a) In *BRAF* melanomas, but not non-*BRAF* melanomas, mutations in *MECOM* and/or *BMP5* were associated with clinical benefit to immunotherapy, as assessed by RECIST criteria. In

Figure 2.3. (continued): addition, when restricting to *MECOM*/*BMP5*-mutated melanomas, *BRAF* melanomas are associated with significantly better clinical benefit compared with non-*BRAF* melanomas (Fisher's exact test, $P = 0.02$, two sided). **b)** Survival curves between *MECOM*/*BMP5*-mutant and wild-type tumors in all immunotherapy-treated tumors ($n = 297$; top), *BRAF* immunotherapy-treated tumors ($n = 109$; middle) and non-*BRAF* immunotherapy-treated tumors ($n = 188$; bottom). **c)** Overlap of *BRAF*-mutant, BRAF p.Val600Glu– and BRAF p.Val600Lys–encoding SMGs revealed that roughly two-thirds of both the p.Val600Glu– and the p.Val600Lys–encoding SMGs were also identified through the *BRAF*-mutant mutational significance analysis. However, only 16% (7/44) and 32% (7/22) of the p.Val600Glu– and p.Val600Lys–encoding SMGs overlapped with each other, respectively. **d)** Despite p.Val600Lys tumors experiencing a twofold enrichment of nonsynonymous mutational load, some BRAF p.Val600Glu–encoding cancer gene (CGC, OncoKB) SMGs are altered in a similar or greater proportion of samples (left; $P > 0.05$ adjusted for mutational load between subtypes, $\chi 2$, two sided). Conversely, some BRAF p.Val600Lys–encoding cancer gene SMGs are altered in more than double the proportion of samples (right; $P < 1.85 \times 10{-}3$ adjusted for mutational load between subtypes, $\chi 2$, two-sided). EGFR, epidermal growth factor receptor; PI3K, phosphoinositide 3-kinase. **a,d)** An asterisk denotes a Mann–Whitney *U*-test $P < 0.05$.

When considering the entire cohort, patients with *MECOM/BMP5* mutations (including *BRAF* melanomas) demonstrated improved clinical benefit (55.6% vs. 37.2%, Fisher's exact test, p = 0.008), PFS (log-rank, p = 0.042, Figure 2.3B, Supplementary Table 2.6), and OS (log-rank, p = 0.021). Similarly, clinical benefit remained significantly associated with *MECOM/BMP5* mutations after correcting for mutational load, tumor purity, and treatment (logistic regression, p = 0.034). Although the results were concordant in this cohort and a limited external validation cohort (Extended Data 2.2), *MECOM/BMP5* mutations were not statistically associated with improved PFS and OS after correcting for these same covariates (PFS: p = 0.053; OS: p = 0.058; Supplementary Table 2.6).

V600E and V600K mutant melanomas

We then examined *BRAF* V600E (NC_000007.13:g.140453136A>T) and V600K (NC_000007.13:g.140453136_140453137delinsTT) tumors given their diverse clinical and genomic features [31-33]. *BRAF* V600E (n = 376) and V600K (n = 76) hotspot mutations comprised 92% of the *BRAF* melanomas in our cohort. Consistent with prior reports, we observed significant differences in median age of diagnosis, mutational load, and copy number burden (Supplementary Figures 2.10-2.12). Given these global differences, we aimed to determine if secondary drivers were unique to the *BRAF* V600E or V600K subtypes (Figure 2.3C, Methods).

In the V600K and V600E cohorts, 13 (59%) and 19 (61.4%) SMGs were also identified as *BRAF* subtype SMGs. *ARID2*, *CDKN2A*, *MAP2K1*, *PPP6C*, *PTEN, RAC1*, and *TP53* were identified as SMGs in the V600E, V600K and overall *BRAF* subtype cohorts. Despite the elevated mutational load in V600K tumors, *CDKN2A*, *PTEN* and *TP53* were mutated in a similar proportion of V600E tumors (χ^2, $p > 0.05$ adjusted for mutational load between subtypes), among others (Figure 2.3D). However, established cancer genes *AKAP9*, *COL3A1*, *DDX3X*, *FAM131B*, *IDH1*, and *USP6* were identified as SMGs unique to V600K melanomas. Conversely, canonical cancer genes that were SMGs exclusive to the V600E cohort included *B2M*, *CDK4*, *CTNNB1*, *EZH2*, *JAK1*, *PRKAR1A*, *TAP2*, and *TRRAP*, several of which are involved in immune response (Supplementary Table 2.7). Thus, *BRAF*-mutant melanomas have distinct genomic substructures overall, and within specific mutant alleles.

(N)RAS-mutant subtype

We identified 56 SMGs in *(N)RAS* melanomas, excluding *NRAS*, *KRAS* and *HRAS* (Supplementary Figures 2.13-2.14). The chromatin remodeler SWI/SNF complex genes *ARID1A*, *ARID1B*, *ARID2* and *BRD7* were all classified as SMGs in *(N)RAS* melanomas (31% of *(N)RAS*-mutant melanomas and 22.5% in non-*(N)RAS*-mutant melanomas; Fisher's exact test, $p = 8.64 \times 10^{-6}$, Figure 2.4A, Supplementary Figure 2.15). Both *ARID2* and *BRD7* are unique to the SWI/SNF PBAF complex, and mutations in these genes were mutually exclusive in *(N)RAS* melanomas (Figure 2.4A)[34]. *(N)RAS* melanomas were significantly associated with putative inactivating mutations in PBAF complex genes in the multivariate analysis correcting for mutation rate and tumor purity (logistic regression, $p = 0.049$ for all PBAF genes, $p = 0.036$ for unique PBAF genes, Supplementary Table 2.8), but not BAF complex genes ($p = 0.1$ for all BAF genes, $p = 0.338$ for unique BAF genes). Further, nonsynonymous BAF/PBAF complex mutations were disproportionately clonal (Methods) in *(N)RAS* melanomas relative to other genomic subtypes (χ^2, $p < 4.08 \times 10^{-5}$ pairwise adjusted for background subtype proportions,

Figure 2.4B), indicating that BAF/PBAF mutations may be tumor initiating events particularly when paired with activating *(N)RAS* mutations (Supplementary Table 2.9).

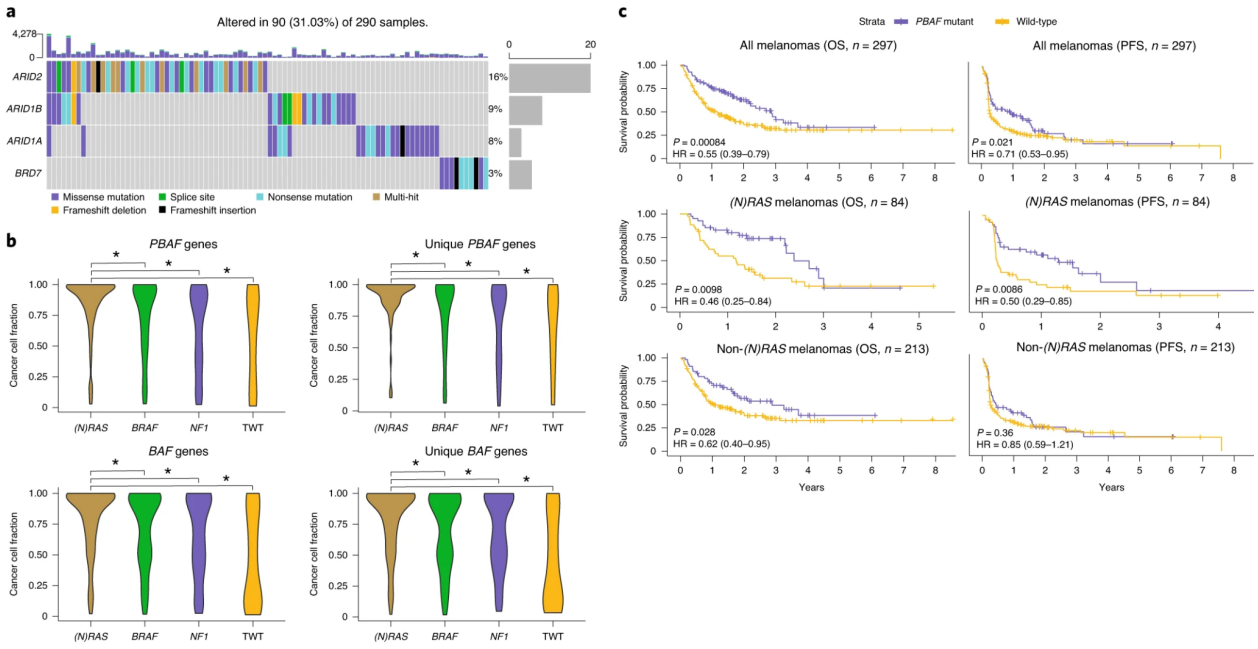


Figure 2.4: *(N)RAS* melanomas frequently experience clonal mutations in the *PBAF* complex, and *PBAF* complex mutations are associated with improved OS and PFS when treated with immunotherapy.

a) The co-mutation plot of *BAF*/*PBAF* complex SMGs identified in *(N)RAS* melanomas. Putative loss-of-function mutations (nonsense, splice site, indels) are almost entirely mutually exclusive to each other. Furthermore, mutations in *ARID2* and *BRD7* (specific to the *PBAF* version of the SWI/SNF complex) were never observed in the same tumor. **b)** The distributions of cancer cell fractions (Methods) for all *PBAF* and *BAF* complex genes among the genomic subtypes. Mutations in *BAF*/*PBAF* complex genes were enriched for being clonal (Methods) in *(N)RAS* melanomas compared with other genomic subtypes ($\chi 2$ pairwise adjusted for subtype proportions, $P < 2.24 \times 10{-}4$, two sided), and *PBAF* gene mutations were clonal more frequently than *BAF* gene mutations in *(N)RAS* melanomas ($P = 0.003$, two-sided Kolmogorov–Smirnov test). An asterisk denotes a Kolmogorov–Smirnov test $P < 0.05$. **c)** Mutations in *PBAF* genes are associated with significantly improved OS and PFS to immunotherapy. Although *PBAF*-mutant *(N)RAS* and non-*(N)RAS* melanomas have significantly better OS compared with their *PBAF* wild-type counterparts, the improvement in OS is much more pronounced in *(N)RAS* melanomas. *PBAF*-mutant non-*(N)RAS* melanomas do not experience significantly better PFS compared with *PBAF* wild-type non-*(N)RAS* melanomas. However, *PBAF*-mutant *(N)RAS* melanomas have significantly improved PFS compared with *PBAF* wild-type *(N)RAS* melanomas, and the PFS signal from *PBAF*-mutant *(N)RAS* melanomas is driving the significant improvement in PFS at the entire cohort level.

Inactivation of the PBAF complex has been associated with improved response to immunotherapy in renal cell carcinoma patients[35], and increased T cell cytotoxicity in melanoma

models (Supplementary Table 2.10)[24]. Within the subset of our cohort that received immunotherapy, *(N)RAS* melanomas co-mutated with PBAF complex mutations were significantly associated with improved PFS (log-rank, $p = 8.6 \times 10^{-3}$, Figure 2.4C, Supplementary Table 2.6) and OS (log-rank, $p = 9.8 \times 10^{-3}$, Figure 2.4C), as well as concordant (but not statistically significant) associations with clinical benefit (56.4% vs. 35.7%, Fisher's exact test, p = 0.076) and OS in a limited external validation cohort (Extended Data 2.3). Non-*(N)RAS* melanomas co-mutated with PBAF complex mutations were also associated to a lesser degree with improved OS (p = 0.028, Figure 2.4C), but not PFS. PBAF complex mutations in the overall cohort were still significantly associated with improved PFS and OS after correcting for mutational load, tumor purity, and treatment (logistic-regression, PFS: p = 0.027; OS: p = 0.007; Supplementary Table 2.6), though this was largely driven by co-mutation with *(N)RAS* melanomas (Supplementary Table 2.6).

NF1-mutant subtype

Consistent with prior studies, we observed that *NF1* melanomas occurred in older patients (Supplementary Figure 2.16A) and harbored higher mutational load than the other genomic subtypes[10-11]. Through our approach, we identified 24 SMGs in *NF1* melanomas (FDR q < 0.01, Supplementary Figures 2.16-2.18). Of the RASopathy genes previously implicated in *NF1* melanomas [10,36], *RASA2* and *SPRED1* were the only ones classified as SMGs. However, *NF1* melanomas were significantly associated with putative inactivating mutations in known RASopathy genes (logistic regression corrected for mutation rate and purity, $p = 1.25 \times 10^{-4}$, Supplementary Table 2.8). One additional RAS-associated gene not previously implicated in melanoma, *RASSF2*, was also identified as a SMG specifically in the *NF1*-mutant subtype[11]. *RASSF2* is a tumor suppressor that regulates the MAP kinase pathway through interactions with *KRAS*, and hypermethylation of its promoter has been observed in several cancer types[37-39].

Triple Wild-Type (TWT) Subtype

Unlike the other subtypes, genomic driver analyses in TWT melanomas have been limited due to insufficient cohort size for unbiased driver discovery. Here, we identified 19 SMGs (FDR q < 0.01, Figure 2.5A, Supplementary Figures 2.19-2.21). Consistent with prior reports, *KIT* was the most frequently mutated SMG[2,17,40]. Three additional SMGs, *GNA11*, *GNAQ*, and *SF3B1*, are known driver genes in uveal melanoma (not included in this study)[41] and predominantly consisted of established hotspot mutations. Though *SF3B1* has been identified as a driver in mucosal melanomas[42], *SF3B1*, *GNA11*, and *GNAQ* were identified as SMGs when considering only cutaneous TWT melanomas (Supplementary Figure 2.19, Supplementary Data 2.2). Through this analysis we identified putative driver events in 91 of 162 (56.17%) tumors, leaving many tumors without a SMG. This large fraction of tumors without known driver mutations may be partially due to the low background mutation rate in this subtype limiting the power to identify more drivers[19-21].

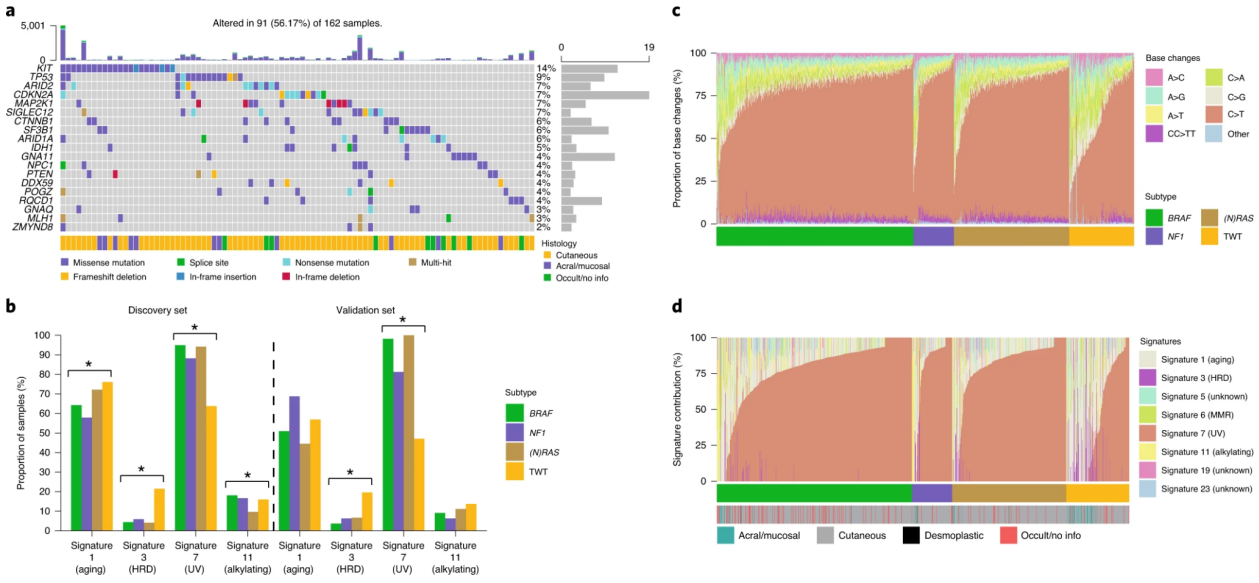


Figure 2.5: Identification of new drivers and enrichment of mutational signature 3 in TWT melanomas.

a) The co-mutation plot of TWT SMGs, including the annotation of tumor histology. These SMGs include canonical melanoma cancer genes (for example, *CTNNB1*, *CDKN2A* and *TP53*) and known uveal melanoma driver genes (for example, *GNA11*, *GNAQ* and *SF3B1*). However, these 19 SMGs explain only the presence of drivers in just over 50% of TWT melanomas. **b)** The

Figure 2.5 (continued): proportion of samples in each genomic subtype exhibiting mutational signatures (Methods) in our discovery and validation ($n = 159$) cohorts. Signature 3 was present in 21.5% of TWT melanomas compared with 4.5% of non-TWT melanomas ($P = 2.20 \times 10-11$, two-sided Fisher's exact test), and was the third most active mutational signature in TWT melanomas. In our validation cohort (Methods), signature 3 was identified in 19.6% of TWT melanomas and 5.6% of non-TWT melanomas ($P = 0.001$, two-sided Fisher's exact test), and was again the third most active signature in TWT melanomas. In both the discovery and the validation cohorts, the proportion of tumors with signatures 3 and 7 was significantly different across the genomic subtypes ($P < 0.05$, $\chi 2$, two sided). An asterisk denotes a $\chi 2$ $P < 0.05$. HRD, homologous recombinant deficiency. **c)** The proportion of base changes ordered by genomic subtype and the proportion of C>T transitions (increasing, left to right). **d)** The relative contribution of each mutational signature ordered by genomic subtype and the relative signature 7 contribution (increasing, left to right).

Similarly, while we observed that TWT melanomas exclusively experienced enrichment of focal amplifications of the *KIT*/*KDR* locus, this subset only represented 18% (29/162) of TWT melanomas (Supplementary Figure 2.22-2.23, Supplementary Table 2.11, Supplementary Data 2.3, Methods)[43]. Half the tumors with *KIT* mutations also had an amplification of *KIT*, including all three tumors with in-frame insertions in *KIT* (Figure 2.5A). SMGs from other genomic subtypes also jointly experienced an enrichment of amplifications or deletions, such as *CDK4* mutations and amplifications in *BRAF* melanomas (Supplementary Table 2.11, Supplementary Data 2.3). Global analysis of structural variants (SVs) revealed that TWT melanomas were enriched in both copy number alterations (CNAs) (Kruskal-Wallis, $p = 9.28 \times 10^{-7}$, Supplementary Figure 2.24A)[44] and fusion events (Kruskal-Wallis, $p = 0.006$, Supplementary Figure 2.24B)[45] compared to other subtypes. *BRAF* was the most common fusion partner in TWT melanomas (n = 3 samples; Supplementary Table 2.7), although there were no recurrent fusion pairs. In aggregate, TWT melanomas demonstrate increased genomic instability relative to the other genomic subtypes, although specific driver events were not highly recurrent in gene-level somatic analysis.

Mutational signatures and DNA repair defects in melanoma

To augment SMG analysis and identify additional TWT melanoma drivers, we next characterized the active mutational processes in this cohort (Methods)[46-47]. Consistent with prior studies, the three most active signatures were signature 1 (aging), signature 7 (UV) and signature 11 (alkylating agents)[48-49]. As expected, performing mutational signature analysis within the *BRAF*, *(N)RAS* and *NF1* subtypes independently revealed these same 3 signatures (Figure 2.5B-D). However, in TWT melanomas, signature 11 was replaced by signature 3, previously associated with homologous recombination (HR) deficiency when observed with *BRCA1/2* mutations in other tumor types, as the third most active signature (Figure 2.5B, Supplementary Data 2.4). Signature 3 was identified in 35 of 162 (21.6%, Methods) TWT tumors and 40 of 886 (4.5%) non-TWT tumors (Fisher's exact test, $p = 5.7 \times 10^{-12}$). Additionally, the average relative contribution of signature 3, when present, was significantly higher in TWT tumors than non-TWT tumors (24.6% vs. 16.7%, t-test, p = 0.015).

Given the flat and ambiguous nature of signature 3[49], we next examined potential confounders to this observation in TWT melanomas. The difference in signature 3 prevalence between TWT and non-TWT tumors was not confounded by histopathological subtype (Fisher's exact test, cutaneous: 19.7% vs. 4.5%, $p = 8.72 \times 10^{-8}$; acral/mucosal: 45.8% vs. 11.1%, p = 0.011), age (logistic regression, 1.72×10^{-10}), or mutational load (logistic regression, $p = 2.6 \times 10^{-3}$). We also replicated our findings using an orthogonal NMF-based method, including enrichment of signature 3 in TWT melanomas (Supplementary Figure 2.25, Extended Data 2.4)[47]. We further confirmed this finding with NMF through downsampling analysis; removing 35 signature 3 tumors vs. 35 non-signature 3 tumors resulted in signature 3 being called in 0% and 92.7% of 1000 simulations, respectively (Extended Data 2.5).

To further evaluate this TWT DNA repair signature finding, we examined its association with copy number loss of heterozygosity (LoH) events[50-51], telomeric allelic imbalance (TAI)[52],

and large scale transitions (LST)[53], which were previously associated with double strand break (DSB) repair and HR deficiency in breast and ovarian cancer[54]. Tumors with signature 3 had significantly greater numbers of LoH regions (Kolmogorov-Smirnov, $p = 0.005$; univariate logistic regression, $p = 5.34 \times 10^{-5}$, Supplementary Figure 2.26, Methods), TAI (Mann-Whitney U, $p = 4.4 \times 10^{-5}$, Supplementary Figure 2.27, Methods), and LST (Mann-Whitney U, $p = 0.007$, Supplementary Figure 2.28, Methods) compared to non-signature 3 tumors (Figure 2.6A). Further, the unweighted sum of these HR deficiency associated CNA events was significantly enriched in tumors with signature 3 (Mann-Whitney U, $p = 6.21 \times 10^{-5}$, Extended Data 2.6, Methods)[54]. To confirm that the association between signature 3 and these DNA repair signatures were not spurious, we performed these same tests for all signatures, but no other signature was significantly associated with all of these associated events (Supplementary Figure 2.29).

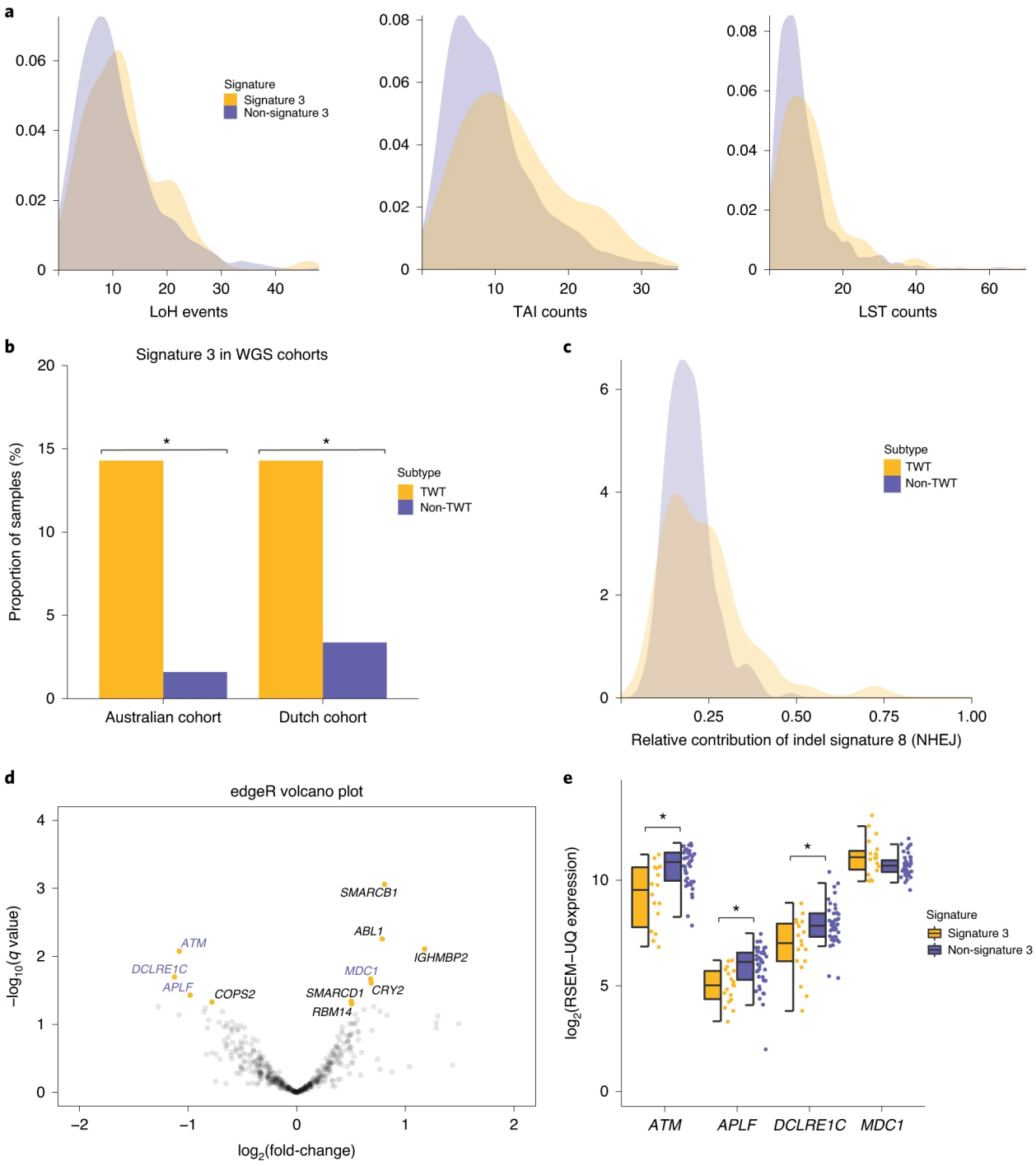


Figure 2.6: Identification of new drivers and enrichment of mutational signature 3 in TWT melanomas.

a) Signature 3 tumors had significantly elevated numbers of LoH ($P = 0.005$; two-sided Kolmogorov–Smirnov test; $P = 5.34 \times 10{-}5$, two-sided univariate logistic regression), TAI ($P = 4.4 \times 10{-}5$, two-sided Mann–Whitney *U*-test) and LST ($P = 0.007$, two-sided Mann–Whitney *U*-test) events. **b)** Signature 3 was also enriched in TWT tumors of two independent melanoma WGS cohorts, suggesting that the assignment of signature 3 was not ambiguous or the result of

Figure 2.6 (continued): noise from lower mutational load in WES data. Asterisks denote a Mann–Whitney *U*-test $P < 0.05$. **c)** Although indel signature ID8 was found in most melanoma WGS samples, the relative contribution of indel signature ID8 was significantly higher in TWT tumors ($P = 3.3 \times 10-3$; two-sided Kolmogorov–Smirnov test). **d)** Significance versus effect size (fold-change) of significantly differentially expressed (Benjamini–Hochberg, *q*-value cutoff <0.05; signature 3 versus non-signature 3) DNA-repair genes via edgeR. Gene names highlighted in purple function in DSB repair pathways including HR. **e)** The distribution of expression between putatively DSB repair-deficient and non-DSB repair-deficient TWT tumors for DSB repair genes that were significantly differentially expressed (Mann–Whitney *U*-test; ordered by two-sided *P* value, increasing, left to right; $*P < 7.6 \times 10-3$). The data are represented as a boxplot where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the IQR ×1.5 and beyond the whiskers are outlier points.

To externally evaluate these mutational signature patterns, we next performed signature analysis in a separate set of melanoma WES tumors that were not included in our original cohort (Methods). Consistent with our findings, signature 3 was observed in 19.6% of TWT tumors and 5.6% of non-TWT tumors (Mann-Whitney U, $p = 9.79 \times 10^{-3}$, Figure 2.5B). To further evaluate whether signature 3 was assigned as a result of ambiguity with lower total called mutations, we analyzed melanoma whole-genome sequenced (WGS) cohorts (n = 390, Hayward *et* al.[17] and Priestley *et al*[55]; Methods). In the Hayward *et al*. cohort, signature 3 was identified in 2 of 14 (14.8%) cutaneous TWT melanomas and 2 of 126 (1.6%) cutaneous non-TWT melanomas (Fisher's exact test, p = 0.0498, Figure 2.6B, Supplementary Data 2.4). In the Priestley *et al*. cohort, signature 3 was identified in 6 of 42 (14.3%) TWT melanomas compared to 7 of 208 (3.8%) non-TWT melanomas (Fisher's exact test, p = 0.011, Figure 2.6B, Supplementary Data 2.4). Signature 3 was still enriched in TWT melanomas when combining the two WGS cohorts (Fisher's exact test, $p = 9.6 \times 10^{-4}$).

Finally, NMF-based indel mutational signature analysis in all 390 WGS tumors revealed that signature 3 was the sole mutational signature associated with indel signature 8 (ID8; Methods), whose proposed etiology is the non-homologous end joining activity component of DSB repair. *BRAF*, (*N)RAS*, and *NF1* melanomas were associated with indel mutational signatures ID1, ID2, and ID13 (associated with UV), while TWT melanomas were associated with indel mutational signatures ID1, ID8, and ID13, even when removing the tumors with

signature 3 (Extended Data 2.7). Although ID8 has been identified in the majority of melanoma tumors[56], as was also the case in the WGS validation cohorts, ID8 was more pronounced in TWT tumors. Single-sample level decomposition (Methods) revealed that although there was no difference in the proportion TWT tumors with ID8 compared to non-TWT tumors (Fisher's exact test, $p > 0.05$), when present, indel signature ID8 contribution was significantly higher in TWT tumors (Kolmogorov-Smirnov, $p = 3.3 \times 10^{-3}$, Figure 2.6C, Supplementary Data 2.4). This may explain why ID8 was only identified in the NMF-based indel signatures for the TWT cohort. Thus, the increased genomic instability of TWT melanomas in general, as evidenced by elevated SV burden, may also manifest in elevated contribution of indel signature ID8 representing double strand DNA repair dysfunction.

Double-strand break repair deficiency in TWT Melanomas

To examine potential sources for this DNA repair dysfunction in TWT melanoma, we surveyed whether alterations in genes previously implicated in DSB repair (N=190, Supplementary Table 2.8, Methods) were enriched in the putative DSB repair defective melanomas, with emphasis on the TWT subtype. While rare deleterious somatic mutations (e.g. *ATM*, *BRCA2*), CNAs (e.g. *CHEK1* and *RNF8* deletions), or germline pathogenic mutations (e.g. *WRN*, *XRCC2*) were observed in DNA repair genes, there was no association with these events and signature 3 (Fisher's exact test, $q > 0.05$, Supplementary Figure 2.30, Supplementary Table 2.12).

Given the lack of genomic features previously correlated with signature 3 in this setting, we then examined whether transcriptional states in DNA repair processes (N=496 genes, Supplementary Table 2.8, Methods), including HR and other DSB repair pathways, could inform the relationship between signature 3 contribution and TWT melanoma (Extended Data 2.8A). Signature 3 contribution was significantly correlated with 19 DNA repair genes (Pearson's, p-value cutoff < 0.05; 9 positive, 10 negative), of which 9 function in DSB repair pathways

(Extended Data 2.8B), suggesting a potential dosage relationship between the degree of signature 3 activity and the expression of these genes. However, none of these genes passed FDR correction when including the full set of genes. To determine if the effect size of expression differences in these genes were also significantly associated with the presence of signature 3, we performed differential expression analysis[57-59]. A total of 11 DNA repair genes were significantly differentially expressed by two methods (Benjamini-Hochberg, q-value cutoff < 0.05, Figure 2.6D, Extended Data 2.9, Supplementary Data 2.5, Methods), four of which are involved in DSB repair (*ATM*, *APLF*, *DCLRE1C*, *MDC1*, Figure 2.6E).

Promoter methylation of *RAD51C* has also been shown to cause HR deficiency in breast cancer[60], although no DNA repair genes were in regions differentially methylated[61-62] between signature 3 and non-signature 3 melanomas (Methods). Analysis of signature 3 contribution correlations with methylation β-values for DNA repair genes (Pearson's, p-value cutoff < 0.05) and anti-correlations with expression identified 6 positions across 6 genes (Extended Data 2.10). *INO80*, which functions in the initial stages of HR[63-64], was the only DSB repair associated gene implicated in this analysis and had significantly higher methylation in signature 3 tumors (Mann-Whitney U, p < 0.015, Extended Data 2.10A, Supplementary Table 2.13). Thus, non-genetic events of the DSB repair genes *ATM*, *APLF* and *INO80*, all of which function early in DSB repair[65-66], were significantly associated with signature 3 contributions in melanoma.

Discussion

Through harmonized and uniform genomic analyses on expanded melanoma WES, we revealed a complex secondary genomic architecture of melanoma that includes multiple oncogenic drivers not previously implicated in this disease. Mutational significance analysis within the genomic subtypes revealed novel, secondary drivers that are rarely shared between the subtypes. Further, several pathways and mechanisms potentially driving tumors in each of the subtypes have remained unappreciated. Over 35% of *BRAF* melanomas had mutations in

the TGF-β pathway genes (*BMP5*, *MECOM*), and roughly 30% of *(N)RAS* melanomas had mutations in SMGs that are core components of the BAF/PBAF complex (*ARID1A*, *ARID1B*, *ARID2*, *BRD7*). Further, nonsynonymous mutations in BAF/PBAF genes were enriched for being clonal in *(N)RAS* melanomas compared to the other genomic subtypes, indicating that aberrant chromatin remodeling and histone modifications may differentially drive tumor progression in a subset of *(N)RAS* melanomas. Each of these observations were linked to associations with selective immune checkpoint blockade response that warrant biological evaluation. Critically, we do not claim prognostic or predictive biomarkers status for these findings, which require randomized prospective analyses.

Prior to this study, TWT melanomas lacked known SMGs, although several studies had proposed *KIT* may be driving a subset of these tumors. Here we've identified 19 SMGs, including *KIT*, in TWT melanomas, and fully characterized their distinct mutational landscape. TWT melanomas have significantly lower mutational load but increased genomic instability (SVs and CNVs). Perhaps most surprisingly, between 14-20% of TWT melanomas display mutational signature 3, which has ambiguous etiology, but in certain histologies has been associated with HR deficiency when co-occurring with *BRCA1/2* mutations.

No one etiology of signature 3 was identified, however transcriptional data suggested a role for downregulation of *ATM* and NHEJ dysregulation. ATM functions in both the initial stages of HR and NHEJ repair[67-68], where it is recruited to DSBs by the MRN complex and subsequently activated, resulting in the phosphorylation of several key HR proteins (e.g. BRCA1, H2AX and MDC1)[69-71], as well as later stages of HR repair after RAD51 filament formation[65]. Further, since APLF is directly phosphorylated by ATM, and accumulates at H2AX foci[66,72-73], this suggests the observed *ATM* down-regulation occurs during the early stages of DSB response. Prior studies in *ATM* deficient cell lines have shown that when BRCA1/2 remain intact, the HR repair pathway is not entirely deficient but rather repairs DSBs at a slower rate, which in turn promotes a more active NHEJ pathway that results in higher rates of DSBs[68]. This

may be similar to the mechanism by which we observe signature 3 in TWT melanoma, and explain why there is an absence of some canonical features associated signature 3 (e.g. *BRCA1/2* alterations)[71], as well as why NHEJ indel signature ID8 is detected but not ID6[56]. Future studies to evaluate the functional consequences of these candidates in melanoma cells, and clinical studies incorporating long read sequencing, may further inform the genetic etiologies underlying these events.

While clinical trials of melanoma patients treated with platinum-based chemotherapy were negative, a subset of patients approximating the frequency of signature 3 positive melanomas in this cohort had definitive clinical responses[74-75]. Prospective assessment of genomically stratified melanomas that consider mutational signatures may enable recovery of a rarely utilized therapeutic modality (platinum-based chemotherapy) and others not widely considered (e.g. ATR inhibitors[68,76]) for melanoma. Moreover, prospective generation of new melanoma models that captures the genomic and phenotypic diversity of the disease (e.g. pre-clinical TWT models with signature 3) will aid identification of novel therapies. Broadly, deep and harmonized exome and genome-wide molecular analysis of increasingly large and histologically uniform tumor types will continue to reveal new biology with immediate translational potential, especially as clinical sequencing programs can directly connect these genomic observations with robust phenotypic measures.

Methods

DNA-seq dataset description

We downloaded publicly available aligned whole-exome sequencing BAM files from 10 previously published studies[1-2,6-7,13-18]. Prior to filtering out samples that failed joint quality control metrics, this sample set consisted of 1,307 tumor with matched normal pairs. Information on pair counts per cohort and dpGaP/ICGC accession numbers are included in Supplementary Table 2.1.

The expression data used in this study is from the TCGA-SKCM cohort, which is publically available from the TCGA-SKCM workspace on FireCloud (TCGA_SKCM_ControlledAccess_V1-0_DATA). The normalized RNA expression data (RSEM-Upper quartile normalized) were used for all expression analysis, except differential expression analysis. The differential expression analysis R packages, edgeR and DESeq2[57-59] both require raw RNA counts since they apply their own normalization methods to the data.

The methylation data used in this study were also downloaded from the TCGA-SKCM workspace on FireCloud. The calculated beta values were used for all methylation analysis.

Genomic data processing

Aligned whole-exome sequencing BAM files were obtained for all samples in the studies mentioned (see DNA-seq dataset description). BAM files aligned to GRCh37 were realigned to Hg19 using the Picard realignment pipeline.

Removal of duplicate samples

To remove duplicate samples from the same patient we calculated the pairwise relationship between all matched normal BAM files in our cohort using Somalier (https://github.com/brentp/somalier). The relatedness between all of the samples used in the analyses of this study can be seen in Supplementary Figure 2.31.

Joint quality control metrics

To pass quality control, we required samples to pass four separate criteria. GATK3.7 (https://hub.docker.com/r/broadinstitute/gatk3/tags) DepthOfCoverage was used to determine the mean target coverage for tumor and normal samples[77]. To pass this metric we required a mean target coverage of at least 50X in the tumor sample and at least 20X in the corresponding normal sample. ContEst (https://software.broadinstitute.org/cancer/cga/contest_download) was

used to determine the extent of cross sample contamination[78]. All samples that had cross sample contamination less than 5% were considered. FACETS (https://github.com/mskcc/facets), an allelic copy number caller that also determines purity and ploidy of tumors, was used to obtain both allelic CNAs (see Copy number analysis) and purity estimates[44]. For tumor samples, we required a tumor purity of 20%. The average purity of the tumor samples that passed this filter was 65% (median: 69%). The last filter we applied was percentage of tumor-in-normal, which was determined by deTiN[79]. All tumor samples with corresponding normal samples that had less than 30% tumor-in-normal passed this filter.

Clinical data

All clinicopathological data was downloaded from the published studies from which we obtained whole-exome sequencing data. The only clinical features used in this study were age at diagnosis, the location of the primary tumor, whether the sample was primary or metastatic, and the histology of the tumor (e.g. cutaneous, acral, mucosal, desmoplastic, occult).

Somatic variant calling

Single-nucleotide variants (SNVs) and other substitutions were called with MuTect (v1.1.6)[80] (https://github.com/broadinstitute/mutect). MuTect (v1.1.6) was used to call SNVs instead of MuTect2 because, at the time of analyses performed herein, the MuTect2 method has not been published and is still actively being developed and compared with other approaches. MuTect mutation calls were filtered for 8-OxoG artifacts[81], and artifacts introduced through the formalin fixation process (FFPE) of tumor tissues[77]. 8-OxoG and FFPE sequencing artifacts were filtered out in a three step process. First, sequence metrics are obtained from running Picard's (https://broadinstitute.github.io/picard/) CollectSequencingArtifactMetrics, which categorizes sequence context artifacts as occurring before hybrid selection (preadapter) or during hybrid selection (bait bias). For 8-OxoG artifacts, Picard's CollectOxoGMetrics was run to obtain

Phred-scaled scores for the 16 trinucleotide sequence contexts implicated in oxidation of 8-oxoguanine. Lastly, orientation bias filtering (C>T transition for FFPE, G>T transversion for 8-OxoG) was applied to these metrics using the GATK tool FilterByOrientationBias. Indels were called with Strelka (v1.0.11). MuTect calls and Strelka[82] calls were further filtered through a panel of normal samples (PoN) to remove artifacts generated by rare error modes and miscalled germline alterations[80]. The cancer cell fraction (CCF) of mutations, defined as the fraction of tumor cells inferred to contain the mutation, were annotated using a modified version of the mafAnno.R script from https://github.com/tischfis/facets-suite, which calculates the CCF likelihoods using the method described in McGranahan *et al*. 2015[83] from FACETS outputs. Clonal mutations were defined as having a CCF of over 80% with a probability of greater than 50% (Prob(CCF > 0.8) > 0.5).

Mutational significance analysis

To identify significantly mutated genes (SMGs) in melanoma, we applied three different algorithms that emphasized mutational recurrence (MutSig2CV; https://github.com/getzlab/MutSig2CV)[9,19], sequence context (MutPanning; https://www.genepattern.org/modules/docs/MutPanning)[20], and accumulated functional impact (OncodriveFML; http://bbglab.irbbarcelona.org/oncodrivefml/home)[21]. Due to the large number of samples, high mutational burden, and wide range of mutational burden, we combined the results (p-values) from each algorithm using Brown's method, and classified SMGs using a strict FDR corrected p-value cutoff (q < 0.01). An expression filter was applied to the list of SMGs, such that only genes that are expressed in melanocytes were considered. All genes that had a RSEM-UQ count of at least 10 passed this expression filter. Additionally, in the event that the mutations were gain of function, genes that failed the initial filter but had a median normalized expression (RSEM-UQ) of at least 10 in the mutated samples were also kept. This is slightly more strict than the expression filter applied in the TCGA-SKCM study, which remains the

largest published study of melanoma exomes to date [2]. To make certain that the gene expression in the bulk transcriptome data was in part due to malignant cells, we leveraged single cell data from Tirosh *et al*.[84], and also required that SMGs had observable expression in malignant melanoma cells. Lastly, hotspot mutations in SMGs were manually run through the UCSC BLAT filter to remove genes that were classified as false positives from mismapped reads. We used UpSetR 1.4.0 to plot the intersection of SMGs between genomic subtypes[85].

Copy number analysis

Allelic copy number alterations (CNAs) were determined using FACETS, which provides information on copy number loss of heterozygosity events[44]. These CNAs were used in all copy number analysis besides identifying regions significantly enriched in focal amplifications/deletions using GISTIC2.0[43], which requires that adjacent segments with the same overall copy number change have not been smoothed into one large segment (See Copy number significance). GATK 3.7 was used to generate segmentation files for all tumor and normal samples that passed quality control, and used as input for GISTIC2.0.

Copy number significance

Focal regions with significant enrichment of amplifications/deletions were identified from a merged segmentation file using GISTIC2.0 (https://github.com/broadinstitute/gistic2). To identify regions harboring germline CNAs to be excluded from the analysis, we ran GISTIC2.0 on the normal samples with amplification and deletion thresholds of 0.1. Any region with a q-value < 0.25 was excluded from the somatic analysis. To identify focal regions with significant enrichment of somatic amplifications/deletions, we ran GISTIC2.0 with amplification and deletion thresholds of 0.3. Any region with a q-value less than 0.1 was considered a peak. Additionally, we examined copy number calls from FACETS for genes associated with DSB repair to determine if they were enriched in tumors with signature 3 via Fisher's exact test.

Immunotherapy survival analysis

To determine if there are significant differences between the survival curves of 2 or more groups of samples we used the log-rank test from the survival R package. We performed this test for both overall survival (OS) and progression free survival (PFS). To evaluate whether tumor mutational burden was a confounding factor in the survival analysis, we also performed cox proportional hazards models adjusting for tumor mutational burden (Supplementary Table 2.6).

Immunotherapy RECIST response analysis

We defined clinical benefit as having complete response (CR), partial response (PR) or stable disease (SD) with overall survival of more than 1 year, per RECIST criteria. Patients classified as having SD with overall survival of 1 year or less were, or progressive disease (PD) were classified as non-responders. To determine if genomic characteristics were associated with clinical benefit to immunotherapy we performed Fisher's exact test.

Whole-exome mutational signatures

Active mutational processes were determined using the deconstructSigs R package (https://github.com/raerose01/deconstructSigs), with a signature contribution cutoff of 6%. This cutoff was chosen because it was the minimum contribution value required to obtain a false-positive rate of 0.1% and false-negative rate of 1.4% via the authors in-silico analysis, and is the recommended cutoff[46]. To confirm the presence of signature 3 as the third most active signature in TWT melanomas, and that the identification of signature 3 was not simply due to the deconstructSigs model, we used the SomaticSignatures R package[47] (https://www.bioconductor.org/packages/release/bioc/html/SomaticSignatures.html), which employs an NMF-based model, rather than the linear-based model used in deconstructSigs. To determine that signature 3 was enriched in TWT melanomas we used Fisher's exact test.

Downsampling of TWT melanomas to determine the robustness of signature 3

To further confirm that signature 3 was indeed the third most dominant signature in TWT melanomas and not being called as a result of the low mutation rate, we first ran 1000 NMF-based simulations (via SomaticSignatures) without the 35 signature 3 TWT samples identified via deconstructSigs to confirm the absence of signature 3. We then ran 1000 NMF-based simulations removing 35 random non-signature 3 TWT samples each run to confirm the existence of signature 3 (Extended Data 5).

Validation of signature 3 and immunotherapy response in independent datasets

To validate the presence of signature 3 in both TWT and non-TWT melanomas, we obtained mutation calls from the supplement of three independent studies: (1) Riaz *et al*. 2017, which included 68 patients with melanoma that were either treated with ipilimumab or ipilimumab-naive[86], (2) Roh *et al.* 2017, which studied 56 melanoma patients of which 53 had mutation calls from pretreatment whole-exome sequencing[87], and (3) Hugo *et al*. 2016, which evaluated 38 pretreatment melanoma patients[88]. The mutation calls for each of these cohorts were obtained from the supplemental information of the original papers, and subsequently run through *deconstructSigs*[41]. To determine that signature 3 was enriched in TWT melanomas we used Fisher's exact test. The 3 cohorts mentioned above, as well as the CheckMate 064 cohort from Rodig *et al*. 2018[89] were used to validate the association *MECOM/BMP5* or PBAF complex mutations with OS and RECIST response.

Calculation of HR deficiency associated copy number events (scores)

To calculate the number of LoH events, TAI events and LST events we used the FACETS copy number calls as input to the scarHRD R package (https://github.com/sztup/scarHRD), which implements the methods used in [50], [52], and [53], respectively. To determine p-values for the association between loss of heterozygosity events and the presence of signature 3, we used a

Kolmogorov-Smirnov test and univariate logistic regression. To determine p-values for the association between the presence of signature 3 with telomeric allelic imbalance (TAI), large scale transitions (LSTs), and the unweighted sum of these homologous recombination associated copy number scores we used a Mann-Whitney U test. We highlighted these specific statistical tests for each score because they were used to find the associations in the original papers, however, Kolmogorov-Smirnov, Mann-Whitney U, and univariate logistic regression were significant for each of the four scores (Supplementary Data 2.4).

Whole-Genome sequenced data analysis

To evaluate whether signature 3 was not being called in the WES data purely because of ambiguity challenges, we performed mutational signature analysis on two melanoma WGS cohorts: (1) the ICGC Hayward *et al*. cohort[17] and (2) the Priestly *et al*. HMF cohort[55]. We received the mutation calls for these cohorts directly from the authors, however, a version of the mutation calls from the Hayward *et al*. cohort is available to download from ICGC. The VCF files from the Priestley *et al*. cohort were annotated using VEP (release 99) to determine the genomic subtype (*BRAF*, *(N)RAS*, *NF1*, TWT) of each sample. To conform with the characterization used in this study, *BRAF* and *(N)RAS* non-hotspot samples were categorized as *BRAF*-mutant or *(N)RAS*-mutant melanomas, respectively.

Indel Mutational Signatures

To call NMF-based indel mutational signatures in the WGS samples we used SigProfiler (v1.0.5)[56] (https://github.com/AlexandrovLab/SigProfilerExtractor), and performed cosine similarity between the global NMF suggested solutions and the known COSMIC signatures. To confirm that the association between signature 3 and ID8 was not random or artefactual, we tested the association between ID8 and all SNV signatures. We did this by running SigProfilier on all tumors with mutational contribution of each signature independently (e.g. running

SigProfiler on all tumors with signature 1, then on all tumors with signature 2, and so on). Besides signature 3, signature 6 was the only other signature to yield ID8. To prove that ID8 was only associated with signature 3 tumors, we reran SigProfiler on the subset of signature 6 tumors that lacked contribution of signature 3, and vice versa. To call single-sample indel signatures we used the deconstructSigs R package[46] and limited the search space to known indel signatures in melanoma[56]. The reference file used for calling indel signatures via deconstructSigs was downloaded from the supplement of Alexandrov *et al*.[56] (https://www.synapse.org/#!Synapse:syn11738318.4).

Germline variant discovery

Germline whole-exome sequencing data were used to perform germline variant calling of single nucleotide variants (SNVs) and small deletions/duplications (indels) across all samples. Genome Analysis Toolkit (GATK) HaplotypeCaller pipeline (version 3.7) was used to call germline variants according to the GATK best practices[77]. GATK Variant Quality Score Recalibration (VQSR) method was used to filter germline variants. The SNP VQSR model was trained using HapMap3.3 and 1KG Omni 2.5 SNP sites, and a 99.6% sensitivity threshold was applied to filter variants. In addition, Mills *et. al*. 1KG gold standard and Axiom Exome Plus sites were used for indel recalibration using a 99% sensitivity threshold[90].

Germline variant pathogenicity evaluation

Pathogenicity of the germline variants that passed filtering were classified according to the American College of Medical Genetics and Genomics and the Association of Molecular Pathology clinical-oriented guidelines[91]. The germline variants were evaluated for pathogenicity using publicly-available databases such as ClinVar and gene-specific databases. Population minor allele frequencies of these variants were obtained from the publicly-available Exome Aggregation Consortium (ExAC) database and Genome Aggregation Database (gnomAD).

Based on the evidence extracted from these resources, germline variants were classified into 5 categories: benign, likely benign, variants of unknown significance, likely pathogenic and pathogenic[91]. Truncating germline variants in genes that have not so far been associated with a clinical phenotype, but are expected to disrupt the protein function, were classified as likely disruptive. Only germline variants classified as pathogenic, likely pathogenic, or likely disruptive were considered in the analysis.

Differential expression analysis

Differential expression analysis was performed using the edgeR[57] (https://bioconductor.org/packages/release/bioc/html/edgeR.html) and DESeq2[58-59] (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) R packages between TWT samples with and without signature 3. Tumor purity was included as a covariate in the models. To classify a gene as significantly differentially expressed we applied a Benjamini-Hochberg corrected p-value threshold of 0.05 (Supplementary Data 2.5). As recommended by the DESeq2 documentation, the output of this method is compatible for input to the Independent Hypothesis Weighting (IHW) R package, and the IHW R package (https://bioconductor.org/packages/release/bioc/html/IHW.html) was used to perform FDR correction for DESeq2 results[92]. Although we ran differential expression analysis on all genes, our downstream analysis focused on DNA repair genes (n = 496, see Gene sets).

Immune Cell Composition

To determine the composition of immune cells in the tumor microenvironment of each tumor we used CIBERSORT[93] (https://cibersort.stanford.edu/). The LM22 immune cell signature matrix was used for deconvolution on the raw TCGA SKCM RNA-seq data. CIBERSORT was run for 1000 permutations and quantile normalization was applied. To determine if there was a

significant shift in the proportion of immune cell types between genomically stratified groups of melanomas we used a Mann-Whitney U test.

Expression correlation analysis

To identify genes whose expression was linearly associated with relative contribution of signature 3, we performed Pearson's correlation between relative signature 3 contribution and normalized RNA expression data for all TCGA-SKCM samples that passed our joint quality control parameters. We performed this analysis on all DNA repair genes (n = 496, see Gene sets), which includes HR genes and other DSB repair pathway genes.

Gene sets

All gene sets and their corresponding genes used for analysis in this study can be found in Supplementary Table 2.8. DNA repair gene sets from the KEGG, GO and REACTOME databases were downloaded from the molecular signatures database (MSigDB v6.2)[94-95] (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). Gene sets from GO, KEGG and REACTOME that specifically contained genes involved in mitotic recombination were considered HR genes (n = 54). HR genes and genes in the GO DSB repair gene set were considered DSB genes (n = 190). The BAF and PBAF gene sets were downloaded from genenames.org. The RASopathy gene set was derived from[96].

Gene fusions

Fusions calls from [45] were leveraged to determine global differences in fusion events between the genomic subtypes and to identify recurrent fusion events. A Kruskal-Wallis test was used to determine if there was a significant difference in the number of fusions events per tumor between the genomic subtypes.

Methylation analysis

Differential methylation analysis was performed between TWT samples with and without signature 3 using bumphunter via the minfi R package[61-62] (http://bioconductor.org/packages/release/bioc/html/minfi.html). We also identified potential sites of methylation associated with signature 3 by applying several joint heuristics and statistical tests. To identify candidate sites we required that there be (1) a significant (p-value cutoff < 0.05) positive Pearson correlation between signature 3 contribution and methylation β-values, (2) a significant median difference in β-values of at least 2% between signature 3 and non-signature 3 tumors, and (3) a significant anticorrelation between methylation β-values and gene expression. The joint heuristic and statistical analysis was restricted to DNA repair genes (n = 496, see Gene sets), while the differential methylation analysis extended to the entire exome.

Pathway over-representation analysis

We performed pathway over-representation analysis on the genomic subtype specific SMGs (including *BRAF* V600E/K) using ConsensusPathDB (v34)[97] (http://cpdb.molgen.mpg.de/). We ran ConsensusPathDB (on March 15, 2020) with default parameters for pathway-based sets, and protein complex-based gene sets (Supplementary Table 2.7).

Statistics and Reproducibility

Statistical analyses were performed using the stats R package for R version 3.6.1. Reported q-values represent Benjamini-Hochberg corrected p-values, and reported p-values represent nominal p-values. All statistical tests performed (e.g. Mann-Whitney U, Kolmogorov-Smirnov, t-test, Fisher's exact test, χ^2) were two-sided.

Declarations

Data Availability

All of the datasets used in this study are publicly available. The raw sequence data can be obtained through dbGaP (https://www.ncbi.nlm.nih.gov/gap) and the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (https://icgc.org/daco), or as described in their original papers (Supplementary Table 2.1). The accession codes can also be found in Supplementary Table 2.1. Publicly available databases used in this study include MSigDB v6.2 (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), ExAC (http://exac.broadinstitute.org/), gnomAD (https://gnomad.broadinstitute.org/), the Broad Institute Single Cell portal (https://singlecell.broadinstitute.org/single_cell), and ConsensusPathDB v34 (http://cpdb.molgen.mpg.de/).

Code Availability

All software and bioinformatic tools used in this study are publicly available.

Acknowledgements

This work was supported by NCI F31CA239347 (J.R.C.), NIH 5T32HG002295-15 (J.R.C.), NIH R01CA227388-02 (E.M.V.A.), NIH R21CA242861 (E.M.V.A.) and the Damon Runyon Clinical Investigator Award (E.M.V.A.). F.D. was supported by the Claudia Adams Barr Program for Innovative Cancer Research and the AWS Cloud Credits for Research Program. The results presented in this study are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This publication and the underlying research are partly facilitated by Hartwig Medical Foundation and the Center for Personalized Cancer Treatment (CPCT) which have generated, analyzed and made available data for this research. We thank Petra

Ross-Macdonald for providing clinical outcomes data used for the immunotherapy response validation analyses.

Author Contributions

J.R.C., A.T.W., S.A., F.D., B.R. and M.X.H. contributed to the analysis of genomic data. J.R.C., C.A.M., B.S., D.S., D.L., E.M.V.A. contributed to aggregation of raw sequence data. N.V., T.K., D.L. and E.M.V.A. contributed to analysis and data interpretation of immunotherapy response. J.R.C., A.T.W., S.A., N.V., F.D., B.R., J.W., R.H., F.S.H., B.S., D.S., D.L., and E.M.V.A. contributed to interpretation of results, and manuscript preparation.

Competing Interests

E.M.V.A. is a consultant for Tango Therapeutics, Genome Medical, Invitae, Enara Bio, Monte Rosa Therapeutics, Manifold Bio, and Janssen. E.M.V.A. provides research support to Novartis and Bristol-Myers Squibb. E.M.V.A. has equity in Tango Therapeutics, Genome Medical, Syapse, Ervaxx, and Microsoft. E.M.V.A. receives travel reimbursement from Roche/Genentech. E.M.V.A. has institutional patents filed on methods for clinical interpretation.

Bibliography

1. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).

2. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696 (2015).

3. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363**, 809–819 (2010).

4. Flaherty, K. T. *et al.* Improved survival with MEK inhibition in BRAF-mutated melanoma. *N. Engl. J. Med.* **367**, 107–114 (2012).

5. Zaretsky, J. M. *et al.* Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).

6. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).

7. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).

8. Wolchok, J. D. *et al.* Overall Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N. Engl. J. Med.* **377**, 1345–1356 (2017).

9. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

10. Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat. Genet.* **44**, 1006–1014 (2012).

11. Cirenajwis, H. *et al.* NF1-mutated melanoma tumors harbor distinct clinical and biological characteristics. *Mol. Oncol.* **11**, 438–451 (2017).

12. Nsengimana, J. *et al.* β-Catenin–mediated immune evasion pathway frequently operates in primary cutaneous melanomas. *Journal of Clinical Investigation* **128**, 2048–2063 (2018).

13. Krauthammer, M. *et al.* Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat. Genet.* **47**, 996–1002 (2015).

14. Van Allen, E. M. *et al.* The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discov.* **4**, 94–109 (2014).

15. Wagle, N. *et al.* MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. *Cancer Discov.* **4**, 61–68 (2014).

16. Miao, D. *et al.* Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nat. Genet.* **50**, 1271–1281 (2018).

17. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).

18. Liu, D. *et al*. Integrative molecular modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat*. *Medicine*. **25**, 1916-1927 (2019).

19. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

20. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genetics.* **52** , 208–218 (2020).

21. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).

22. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).

23. Johannessen, C. M. *et al.* A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. *Nature* **504**, 138–142 (2013).

24. Pan, D. *et al.* A major chromatin regulator determines resistance of tumor cells to T cell–mediated killing. *Science* **359**, 770–775 (2018).

25. Yu, Q. *et al.* Requirement for CDK4 kinase function in breast cancer. *Cancer Cell* **9**, 23–32 (2006).

26. Boland, C. R., Richard Boland, C. & Goel, A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology* **138**, 2073–2087.e3 (2010).

27. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).

28. Alliston, T. *et al.* Repression of bone morphogenetic protein and activin-inducible transcription by Evi-1. *J. Biol. Chem.* **280**, 24227–24237 (2005).

29. Buonamici, S. *et al*. EVI1 abrogates interferon-alpha response by selectively blocking PML induction. *J. Biol. Chem*. **280**, 428-436 (2005).

30. Lee, J. J. *et al*. Targeted next-generation sequencing reveals high frequency of mutations in epigenetic regulators across treatment-naïve patient melanomas. *Clin*. *Epigenetics* **7**, 59 (2015).

31. Menzies, A. M. *et al.* Distinguishing clinicopathologic features of patients with V600E and V600K BRAF-mutant metastatic melanoma. *Clin. Cancer Res.* **18**, 3242–3249 (2012).

32. Flaherty, K. *et al.* Genomic analysis and 3-y efficacy and safety update of COMBI-d: A phase 3 study of dabrafenib (D) + trametinib (T) vs D monotherapy in patients (pts) with unresectable or metastatic BRAF V600E/K-mutant cutaneous melanoma. *J. Clin. Orthod.* **34**, 9502–9502 (2016).

33. Long, G. V. *et al.* Dabrafenib plus trametinib versus dabrafenib monotherapy in patients with metastatic BRAF V600E/K-mutant melanoma: long-term survival and safety analysis of a phase 3 study. *Ann. Oncol.* **28**, 1631–1639 (2017).

34. Kadoch, C. & Crabtree, G. R. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Sci Adv* **1**, e1500447 (2015).

35. Miao, D. *et al.* Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359**, 801–806 (2018).

36. Arafeh, R. *et al.* Recurrent inactivating RASA2 mutations in melanoma. *Nat. Genet.* **47**, 1408–1410 (2015).

37. van der Weyden, L. & Adams, D. J. The Ras-association domain family (RASSF) members and their role in human tumourigenesis. *Biochim. Biophys. Acta* **1776**, 58–85 (2007).

38. Akino, K. *et al.* The Ras effector RASSF2 is a novel tumor-suppressor gene in human colorectal cancer. *Gastroenterology* **129**, 156–169 (2005).

39. Endoh, M. *et al.* RASSF2, a potential tumour suppressor, is silenced by CpG island hypermethylation in gastric cancer. *Br. J. Cancer* **93**, 1395–1399 (2005).

40. Curtin, J. A., Busam, K., Pinkel, D. & Bastian, B. C. Somatic activation of KIT in distinct subtypes of melanoma. *J. Clin. Oncol.* **24**, 4340–4346 (2006).

41. Robertson, A. G. *et al.* Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. *Cancer Cell* **32**, 204–220.e15 (2017).

42. Newell, F. *et al*. Whole-genome landscape of mucosal melanomas reveals diverse drivers and therapeutic targets. *Nat. Commun.* **10**, 3163 (2019).

43. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

44. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).

45. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23**, 227–238.e3 (2018).

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

47. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).

48. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).

49. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

50. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).

51. Timms, K. M. *et al.* Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res.* **16**, 475 (2014).

52. Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).

53. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).

54. Marquard, A. M. *et al.* Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomark Res* **3**, 9 (2015).

55. Priestley, P. *et al*. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210-216 (2019).

56. Alexandrov, A. *et al*. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).

57. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

58. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

60. Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).

61. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200–209 (2012).

62. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

63. Tsukuda, T. *et al.* INO80-dependent chromatin remodeling regulates early and late stages of mitotic homologous recombination. *DNA Repair* **8**, 360–369 (2009).

64. Lademann, C. A., Renkawitz, J., Pfander, B. & Jentsch, S. The INO80 Complex Removes H2A.Z to Promote Presynaptic Filament Formation during Homologous Recombination. *Cell Rep.* **19**, 1294–1303 (2017).

65. Bakr, A. *et al.* Involvement of ATM in homologous recombination after end resection and RAD51 nucleofilament formation. *Nucleic Acids Res.* **43**, 3154–3166 (2015).

66. Fenton, A. L., Shirodkar, P., Macrae, C. J., Meng, L. & Koch, C. A. The PARP3- and ATM-dependent phosphorylation of APLF facilitates DNA double-strand break repair. *Nucleic Acids Res.* **41**, 4080–4092 (2013).

67. Ceccaldi, R., Rondinelli, B. & D'Andrea, A. D. Repair Pathway Choices and Consequences at the Double-Stranded Break. *Trends Cell Biol.* **26**, 52-64 (2016).

68. Balmus, G. *et al*. ATM orchestrates the DNA-damage response to counter toxic non-homologous end-joining at broken replication forks. *Nat. Commun.* **10**, 87 (2019).

69. Zhang, J., Ma, Z., Treszezamsky, A. & Powell, S. N. MDC1 interacts with Rad51 and facilitates homologous recombination. *Nat. Struct. Mol. Biol.* **12**, 902–909 (2005).

70. Scully, R. & Xie, A. Double strand break repair functions of histone H2AX. *Mutat. Res.* **750**, 5–14 (2013).

71. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

72. Iles, N., Rulten, S., El-Khamisy, S. F. & Caldecott, K. W. APLF (C2orf13) is a novel human protein involved in the cellular response to chromosomal DNA strand breaks. *Mol. Cell. Biol.* **27**, 3793–3803 (2007).

73. Macrae, C. J., McCulloch, R. D., Ylanko, J., Durocher, D. & Koch, C. A. APLF (C2orf13) facilitates nonhomologous end-joining and undergoes ATM-dependent hyperphosphorylation following ionizing radiation. *DNA Repair* **7**, 292–302 (2008).

74. Flaherty, K. T. *et al.* Phase III trial of carboplatin and paclitaxel with or without sorafenib in metastatic melanoma. *J. Clin. Oncol.* **31**, 373–379 (2013).

75. Wilson, M. A. *et al.* Correlation of somatic mutations and clinical outcome in melanoma patients treated with Carboplatin, Paclitaxel, and sorafenib. *Clin. Cancer Res.* **20**, 3328–3337 (2014)

76. Rafiei, S. *et al*. ATM Loss Confers Greater Sensitivity to ATR Inhibition than PARP Inhibition in Prostate Cancer. *Clin. Cancer Res.* (2020).

77. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).

78. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).

79. Taylor-Weiner, A. *et al.* DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).

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

81. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).

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

83. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).

84. Tirosh, I. *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).

85. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

86. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934–949.e16 (2017).

87. Roh, W. *et al.* Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci. Transl. Med.* **9**, (2017).

88. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165**, 35–44 (2016).

89. Rodig, S. J. *et al*. MHC proteins confer differential sensitivity to CTLA-4 and PD-1 blockade in untreated metastatic melanoma. *Sci. Transl. Med.* **10**, eaar3342 (2018).

90. Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* **21**, 830–839 (2011).

91. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

92. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580 (2016).

93. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

94. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

95. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

96. Rauen, K. A. The RASopathies. *Annu. Rev. Genomics Hum. Genet.* **14**, 355–369 (2013).

97. Kamburov, A. *et al*. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.* **39**, 712-717 (2011).

Chapter 3: Melanoma SV Analysis

Abstract

We performed harmonized structural variant (SV) and molecular analysis on 355 melanomas and discovered markedly different global genomic properties among melanoma histological subtypes, histology-specific cancer genes recurrently affected by SVs overlapping adjacent topologically associated domain (TAD) boundaries, and a potential mechanism for the downregulation of *ATM* observed in double-stranded break (DSB) repair deficient triple wild-type cutaneous melanomas. Acral melanomas were associated with increased numbers of SVs and prevalence of chromothripsis compared to cutaneous and mucosal melanomas. Additionally, while mucosal melanomas were enriched for SVs relative to cutaneous melanomas, there was no difference in the prevalence of chromothripsis. Irrespective of histological subtype, SVs affecting TAD boundaries were enriched for affecting actively expressed TADs rather than lowly expressed or repressed TADs. The most recurrently altered TAD boundaries in acral and mucosal melanomas was chr11:77750000-77825000, which is adjacent TADs containing the cancer genes *GAB2* and *PAK1*. *GAB2* and *PAK1* both play a role in the MAPK signaling pathway. The most recurrently altered boundary in cutaneous melanomas was chr9:21700000-21775000, which is adjacent to the TADs containing the tumor suppressors *CDKN2A*, *CDKN2B*, and *MTAP*. Lastly, in the subset of cutaneous melanomas in our cohort, we identified that SVs affecting the MRN complex were associated with the presence of DSB repair deficiency, as well as the expression of *ATM*. Broadly, integration of SV analysis with regulatory

element and molecular data revealed subtype-specific modes of melanoma oncogenesis, and a potential causal mechanism for DSB repair deficiency in cutaneous melanoma, both of which provide new opportunities for biological and therapeutic investigation.

Introduction

Cutaneous melanoma is among the most highly mutated cancers due to the impact of UV mutagenesis leading to many C>T transitions across the genome[1,2]. For this reason, molecular analyses of melanoma is often focused on somatic mutations, and increasingly so due to the association of tumor mutational burden (TMB) with response to immunotherapy[3–6]. Somatic structural variant (SV) analyses of cutaneous melanoma whole genomes have been performed[2,7], with emphasis on the counts and frequency of SVs. In contrast to cutaneous melanoma, acral and mucosal melanomas are associated with lower TMBs compared to cutaneous melanomas, with the majority of tumors showing no detectable effect of UV mutagenesis on their mutational spectrums (e.g. *TERT*, *CDK4*, *MDM2*). Instead, in these subtypes, comprehensive SV analysis identified higher SV burden than cutaneous melanomas and the presence of focal SVs targeting known cancer genes[8,9]. Conversely, features relating to the landscape of SVs across histologic or molecular (*BRAF*-mutant, *RAS*-mutant, *NF1*-mutant, and triple wild-type (TWT)) subtypes of melanoma remain incompletely characterized.

Chromothripsis, a single complex genomic event characterized by several SVs clustered in genomic regions of oscillating copy number states across one or more chromosomes, has been systematically characterized in acral melanomas[9], and a subset of cutaneous melanomas available through the PCAWG consortium (n=106)[9,10]. In contrast to chromothripsis characterization, the frequency and effect of SV events on topologically associated domains (TADs), which preserve the regulatory landscape of genes[11], remains unexplored across all melanoma histological subtypes. Disruption of boundaries between TADs has been shown to result in dysregulation of neighboring gene expression through a variety of mechanisms,

including overexpression of oncogenes through enhancer hijacking[12] or inversions overlapping TAD boundaries placing genes near atypical regulatory elements [13,14].

Finally, a subset of cutaneous melanomas exhibit SNV mutational signature 3 (associated with BRCA1/2 mutation and double-stranded break (DSB) repair deficiency in certain cancer types) associated with downregulation of *ATM*[15], among dysregulation of other genes that function early in the DSB repair pathway. However, no DSB repair-associated genomic features identified in cutaneous melanoma whole-exomes were statistically associated with signature 3, and the mechanism leading to the downregulation of *ATM* in the majority of these tumors remains unclear. SV analysis also enables the identification and characterization of various DSB repair mechanisms activity between signature 3 positive and negative tumors[7,15], beyond homologous recombination deficiency (HRD)-associated events that can be obtained through allelic copy number analysis[16,17]. Taken together, we hypothesized that somatic SVs may inform (i) molecularly defined subtype-specific modes of melanoma oncogenesis, (ii) regulatory disruption, and (iii) DNA repair defects that were not identifiable via somatic mutation analysis. Thus, we harmonized WGS from 355 melanomas to investigate the role of SVs in melanoma oncogenesis across these different axes.

Results

We assembled and uniformly analyzed SVs in 355 melanoma WGS (116 acral, 175 cutaneous, and 64 mucosal melanoma; Methods)[1,2,8,9]. Of the cutaneous melanoma samples, 81, 55, 19, and 20 samples were *BRAF*-mutant, *NRAS*-mutant, *NF1*-mutant, and triple wild-type (TWT), respectively. The median sequencing coverage was 57X and 37X in tumor and matched normal samples, respectively, with no statistical difference in tumor sample coverage between the histologies (Wilcoxon-Mann-Whitney, p = 0.08; Supp. Figure 3.1). Additionally, there was no difference in the median tumor purity between the histologies, ranging from 61% in mucosal melanomas to 66% in acral melanomas (Wilcoxon-Mann-Whitney, p = 0.37), while background

ploidy in acral (3.3) and mucosal (2.9) melanomas were significantly higher than in cutaneous (2.1) melanomas (Wilcoxon-Mann-Whitney, $p < 3.8 \times 10^{-5}$). In total, our framework identified a total of 106,032 (median events per tumor; acral: 81, mucosal: 64, cutaneous: 23) somatic genomic rearrangements (Methods; Figure 3.1A), consisting of 25,401 deletions (DEL), 16,297 duplications (DUP), 17,935 inversions (INV), and 46,399 translocations (TRA). Of the 46,399 TRA events, 13,075 (28%) were intrachromosomal while 33,324 (72%) were interchromosomal. Across acral, mucosal, and cutaneous melanomas approximately 72.4%, 71.4%, and 70.7% of TRA events were interchromosomal, respectively.

Global properties of SVs across histological subtypes

The number and features of SVs varied widely across the melanoma histologies. Both acral and mucosal melanomas had significantly higher numbers of TRA, DEL, INV, and DUP events per tumor compared to cutaneous melanomas (Wilcoxon-Mann-Whitney, $p < 2.89 \times 10^{-9}$; Figure 3.1B). However, when compared to mucosal melanomas, acral melanomas had significantly higher numbers of TRA (Wilcoxon-Mann-Whitney, $p = 2.9 \times 10^{-4}$) and INV ($p = 0.01$) events per tumor, but not DEL or DUP events. Acral melanomas were also significantly associated with larger (measured by distance between breakpoints) SV events across all SV categories compared to cutaneous melanomas (Wilcoxon-Mann-Whitney, $p < 0.026$), but not mucosal melanomas. Furthermore, the distributions of DEL and INV sizes in cutaneous melanomas possess distinctive modes surrounding smaller SV events (Kolmogorov-Smirnov, $p < 2.2 \times 10^{-16}$; Figure 3.1C), which may suggest a distinct mechanism of generation. Indeed, pan-cancer analysis of SVs identified small deletions (< 10kb) to be enriched in early replicating regions near TAD boundaries, and inversions enriched in late replicating regions[7].

Within cutaneous melanomas, there was no difference in the number of TRA, INV and DUP events per tumor between the genomic subtypes (Wilcoxon-Mann-Whitney, $p > 0.05$). However, *NF1*-mutant melanomas had significantly higher numbers of DEL events per tumor

compared to the other genomic subtypes (Wilcoxon-Mann-Whitney, $p < 0.022$; Figure 3.1D). Examining the distribution of DEL and INV sizes within cutaneous melanomas revealed that the majority of smaller SV events in this histology were in *NF1* and *NRAS*-mutant tumors (Figure 3.1E). Thus, between histologic and molecular subtypes, the quantity and characteristics of SVs varies widely.

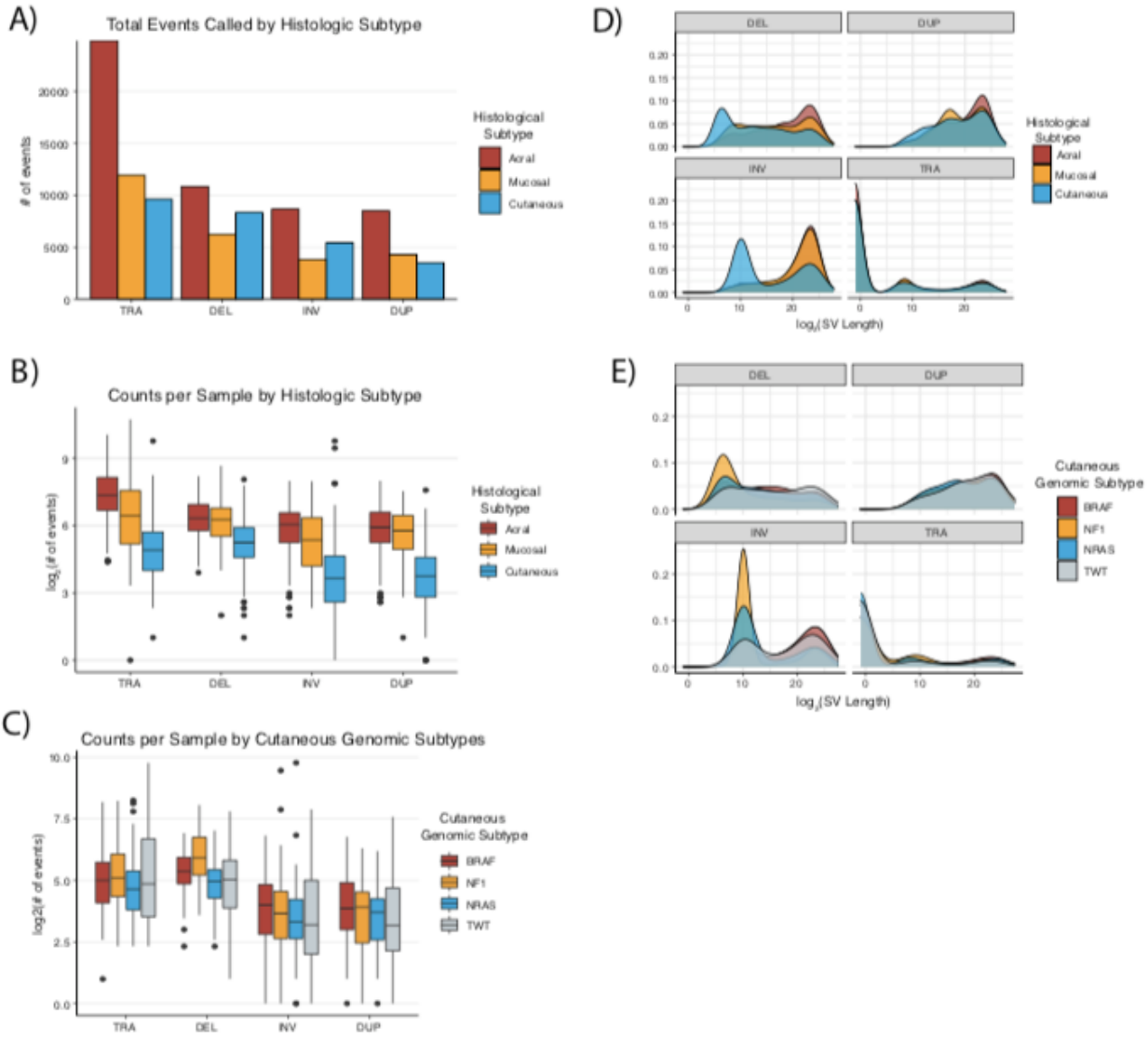


Figure 3.1: Characteristics of histological and cutaneous genomic subtypes in melanoma **(a)** The total number of TRA, DEL, INV, and DUP events in acral, cutaneous, and mucosal melanomas. **(b)** The distribution of the number of TRA, DEL, INV and DUP events in acral, cutaneous, and mucosal melanomas. **(c)** The distribution of the number of TRA, DEL, INV and DUP events across the cutaneous melanoma genomic subtypes. **(d)** The distribution of the sizes of TRA, DEL, INV and DUP events in acral, cutaneous, and mucosal melanomas. **(e)** The

Figure 3.1 (continued): distribution of the sizes of TRA, DEL, INV and DUP events across the cutaneous melanoma genomic subtypes. **(b-c)** Asterisks denote a p-value < 0.05.

Characteristics of chromothripsis

While chromothripsis has been identified in each of the melanoma histological subtypes, prior studies were either unable to differentiate chromothripsis from other complex events[8,9], or used methods for identification with low sensitivity[10,18,19]. Additionally, the relevance of chromothripsis between melanoma histological subtypes, or between genomic subtypes within cutaneous melanomas which have been shown to harbor distinct genomic features, is unknown [1,2,15]. In this cohort, acral melanomas were significantly enriched for chromothripsis events (Methods)[10] compared to both mucosal (Fisher's exact, OR = 5.04, 95% CI = 2.51 - 10.44, $p = 8.23 \times 10^{-7}$) and cutaneous melanomas (Fisher's exact, OR = 6.84, 95% CI = 3.96 - 12.04, $p = 5.01 \times 10^{-14}$; Figure 3.2A), while there was no significant difference in the rate of chromothripsis between cutaneous and mucosal melanomas (Fisher's exact, p = 0.41). Further, 85% of chromothripsis events in acral melanomas involved interchromosomal SVs compared to 65% of mucosal (Fisher's exact, OR = 3.05, 95% CI = 0.85 - 10.49, p = 0.055) and 52% of cutaneous (Fisher's exact, OR = 5.17, 95% CI = 2.07 - 13.53, $p = 1.1 \times 10^{-4}$) melanomas. Of the interchromosomal chromothripsis events, the majority involve more than 1 additional chromosome (> 2 in total; 67% acral, 62% mucosal, and 57% cutaneous). In one extreme case, an Acral melanoma tumor had a single chromothripsis event affecting 18 chromosomes (Figure 3.2B), whereas the most chromosomes involved in a single chromothripsis event in mucosal and cutaneous melanomas were 8 (Figure 3.2C) and 6, respectively. Thus, chromothripsis is the source of genomic instability in the majority of acral melanomas, and may present many opportunities to identify clinically relevant druggable fusions per tumor. Conversely, cutaneous and mucosal melanomas experience chromothripsis at less than half the rate of acral melanomas, and have similar chromothripsis landscapes despite significantly different global SV properties.

Within cutaneous melanomas, 42% (8/19) of *NF1*-mutant melanomas harbored chromothripsis events compared to 20-25% in the other genomic subtypes, although this did not reach statistical significance (Fisher's exact, p = 0.09; Figure 3.2D). All but one (88%) *NF1*-mutant melanomas that harbored chromothripsis involved interchromosomal SVs, compared to just 38% of *BRAF*-mutant melanomas with chromothripsis. Roughly 55% and 50% of *NRAS*-mutant and TWT tumors with chromothripsis involved interchromosomal SVs, respectively. 2 of 3 (67%) *NF1*-mutant melanomas with missense mutations harbored chromothripsis, compared to 6 of 16 (37.5%) *NF1*-mutant melanomas with inactivating mutations, although this difference was not statistically significant (Fisher's exact test, p > 0.05). There was no difference in the proportion of V600E and V600K tumors with chromothripsis within *BRAF*-mutant melanomas.

A subset of samples in each genomic subtype had chromothripsis events that spanned other driver genes that define the subtypes. For example, one *BRAF* melanoma harbored an intra-chromosome chromothripsis event that affected the *BRAF* locus (Figure 3.2E), while 4 other *BRAF* melanomas harbored chromothripsis events that spanned *NRAS*. One tumor with an *NRAS* G12R mutation had an intra-chromosome chromothripsis event spanning *KRAS* (Figure 3.2F), while *BRAF* and *NF1* were involved in chromothripsis events in 1 *NRAS* melanoma each. Additionally, 2, 4, and 1 *NF1* melanomas harbored chromothripsis events spanning *BRAF*, *NRAS*, and *NF1* respectively, with 2 of the tumors harboring chromothripsis at the *NRAS* locus also harboring chromothripsis at the *KRAS* locus. Furthermore, 1 of the 2 *NF1* melanomas with chromothripsis events spanning *NRAS* and *KRAS* was the tumor with the chromothripsis event spanning the *NF1* locus. In TWT tumors, *BRAF* and *NRAS* were involved in chromothripsis events in 1 sample each. Thus, SVs generated via chromothripsis may provide secondary mechanisms of MAPK pathway dysregulation through genes that define the genomic subtypes, which was even more rare outside of chromothripsis events. Furthermore, in

the case of *BRAF* melanomas, these events may result in resistance mechanisms to targeted therapy[20]. Thus, chromothripsis events in cutaneous melanoma are capable of generating alterations that drive tumor initiation and development.

We lastly determined whether the distribution of short INV and DELs observed in *NF1* and *NRAS*-mutant melanomas were the result of chromothripsis. The distribution of small INVs observed in *NF1*-mutant melanomas were largely driven by 2 samples, both of which had chromothripsis. However, only 34.6% and 7.3% of small INVs in these samples were located in chromothripsis regions. Similarly, the distribution of small INVs observed in *NRAS*-mutant melanomas was largely driven by a single sample that harbored chromothripsis, with only 8.5% of these small INVs were located in chromothripsis regions. While the distribution of short DELs observed in *NF1* and *NRAS*-mutant melanomas were not driven by a few outlier samples, there again was no association with the numbers of these events and chromothripsis (Wilcoxon-Mann-Whitney, $p > 0.05$). These results suggest that despite the frequency of SV events differing between the histological subtypes, the differences in the sizes of these SV events are driven by outlier samples.

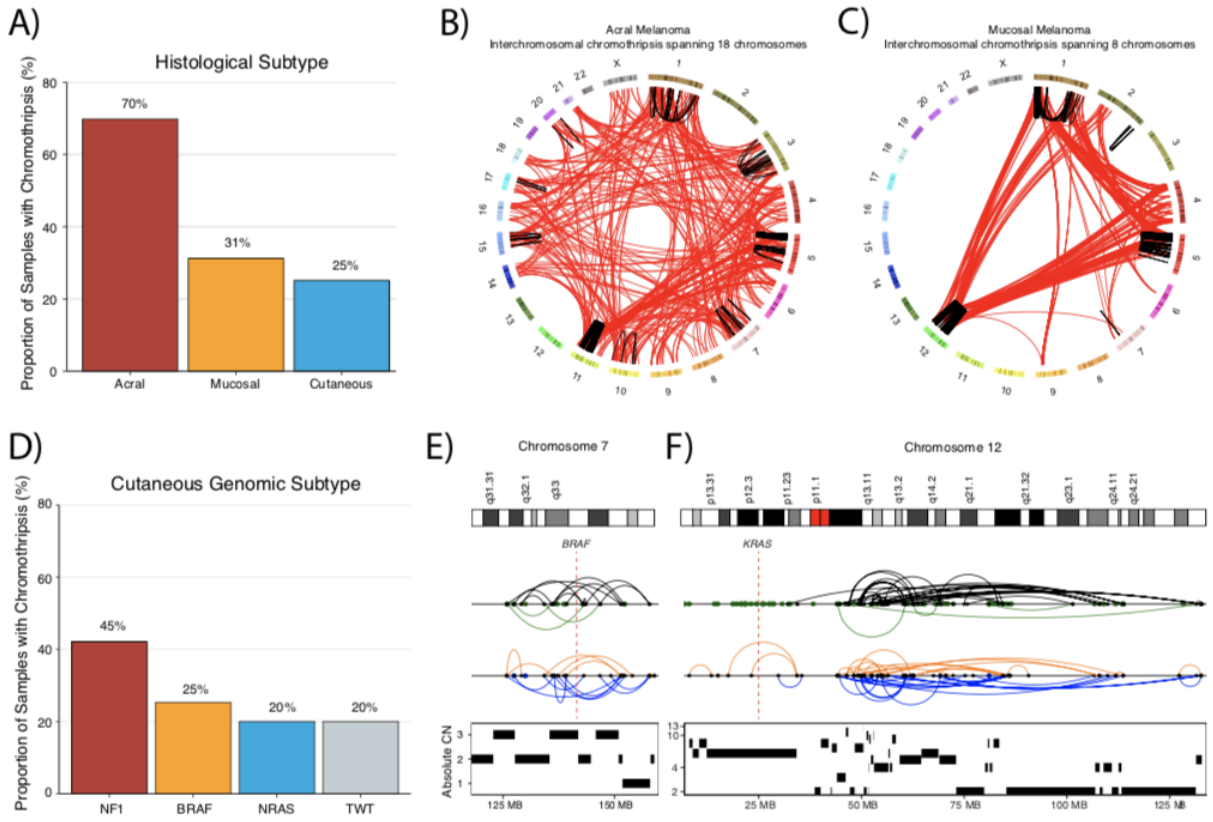


Figure 3.2: The rate and characteristics of chromothripsis events vary by melanoma histologies and cutaneous genomic subtypes.
(a) The frequency of chromothripsis across acral, cutaneous, and mucosal melanomas. **(b)** The most extreme chromothripsis event observed in an acral melanoma tumor, which consisted of SVs spanning a total of 18 chromosomes. **(c)** The most extreme chromothripsis event observed in a mucosal melanoma tumor, which consisted of SVs spanning a total of 8 chromosomes. **(d)** The frequency of chromothripsis across the cutaneous melanoma genomic subtypes. **(e)** An example of an intra-chromosomal chromothripsis event spanning the *BRAF* locus in a *BRAF*-mutant cutaneous melanoma. **(f)** An example of an intra-chromosomal chromothripsis event spanning the *KRAS* locus in a *(N)RAS*-mutant cutaneous melanoma. **(a, d)** Asterisks denote a p-value < 0.05.

Effect of SVs on topologically associated domains (TADs)

Disruption of TAD boundaries through chromothripsis or other SV events can lead to the formation of neo-TADs and dysregulation of gene expression, whereby transcription factors, enhancers[12,21], and silencers[22] that are typically absent from a gene's native TAD may act on the gene as a result of SVs[23]. To investigate the effect of SVs on TADs in melanoma, we focused on SVs unlikely to span multiple TAD boundaries using an established cutoff defined by the

PCAWG consortium (< 2Mb; Methods)[11]. To infer the putative impact of boundary affecting SVs (BA-SVs), we leveraged the 5 TAD type annotations from that same study[11], which were determined using the 15 chromatin state model from the Roadmap Epigenomics Project[24]. These 5 TAD types are Heterochromatin, Low, Repressed, Low-Active, and Active, which are associated with increased expression in the order specified for genes contained within the TADs. We observed that 17.2%, 13.6%, and 7.2% of acral, mucosal, and cutaneous melanoma SVs (< 2Mb) spanned TAD boundaries, respectively. All acral melanoma tumors harbored at least one SV that spanned a TAD boundary, compared to 97% and 86.3% of mucosal and cutaneous melanomas, respectively (Figure 3.3A). Further, when assessing the putative functional impact of BA-SVs across histological subtypes, roughly 97% acral melanomas harbored a TAD boundary spanning SV adjacent to an Active TAD, compared to 83% of mucosal and less than 50% of cutaneous melanomas (Figure 3.3B-C). While there was no significant association between chromothripsis and the presence BA-SVs in a tumor in any histological subtype (Fisher's, $p > 0.05$), tumors with chromothripsis events were associated with higher numbers of BA-SVs per tumor in acral (Wilcoxon-Mann-Whitney, $p = 2.7 \times 10^{-5}$) and cutaneous (Wilcoxon-Mann-Whitney, $p = 0.026$) melanomas, but not mucosal melanomas (Wilcoxon-Mann-Whitney, $p = 0.09$).

Of the total 2477 TAD boundaries, 399 (16.1%), 159 (6.4%), and 105 (4.2%) boundaries were affected by SVs in more than one tumor in the acral, mucosal, and cutaneous cohorts, respectively (Figure 3.3D). Further, SVs affecting the recurrently altered boundaries comprised 56.6%, 35.7%, and 28.4% of all boundary spanning SVs in acral, mucosal, and cutaneous melanomas, respectively. There was no enrichment in the types of TADs adjacent to recurrently altered boundaries (altered in > 1 sample) compared to boundaries only altered in a single tumor across the histological subtypes (Fisher's exact, $p > 0.05$). In general BA-SVs adjacent TADs containing tumor suppressors (Supp.Table 1) were enriched for deletion events (Fisher's exact; OR = 2.34; 95% CI = 1.63 - 3.35; $p = 2.21 \times 10^{-6}$; Methods; Figure 3.3E), whereas

BA-SVs adjacent TADs containing oncogenes (Supp.Table 1) were enriched for complex events (Fisher's exact; OR = 2.62; 95% CI = 1.69 - 4.18; $p = 2.71 \times 10^{-6}$; Methods; Figure 3.3E).

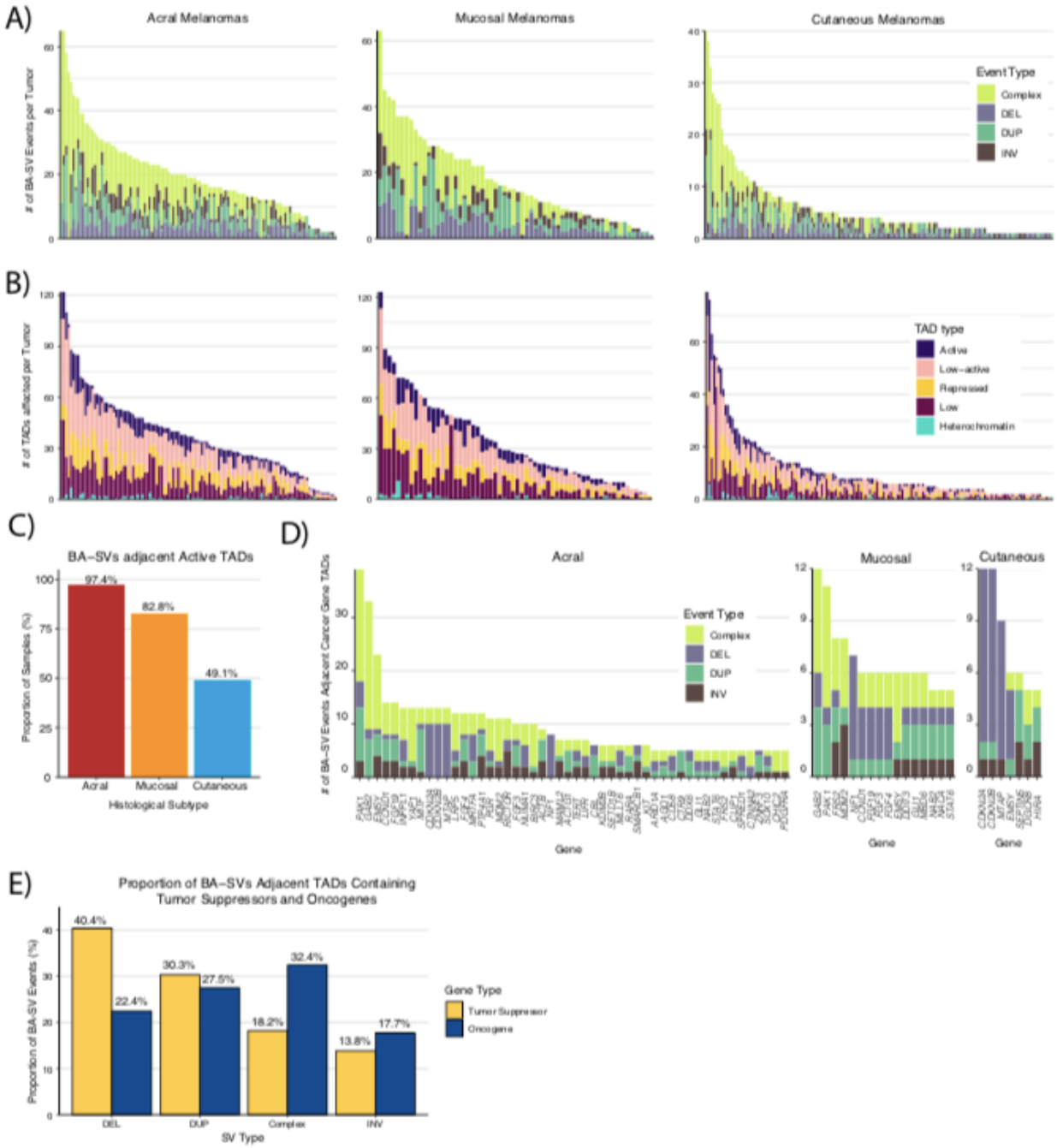


Figure 3.3: Melanomas frequently harbor SVs affecting boundaries adjacent to active TADs and TADs containing cancer genes

(a) The number of BA-SV spanning events per tumor across acral, mucosal and cutaneous melanomas categorized by the type of SV event. Complex SV events are defined as overlapping concomitant DEL, DUP, INV, or TRA events. **(b)** The number of affected TADs per

Figure 3.3 (continued): tumor across acral, mucosal and cutaneous melanomas categorized by functional TAD type (Methods). **(c)** The proportion of acral, mucosal, and cutaneous melanomas with BA-SVs adjacent active TADs. **(d)** Cancer genes that are putatively affected by BA-SVs in at least 5 tumors per histological subtype, characterized by the type of SV event. **(e)** The proportion of event types resulting in BA-SVs that putatively affect tumor suppressors and oncogenes (Methods). **(c, e)** Asterisks denote a p-value < 0.05.

The most recurrently affected TAD boundary in both Acral (n=27, 23%) and Mucosal (n=7, 11%) melanomas was chr11:77750000-77825000, which is adjacent TADs containing the cancer genes *GAB2* and *PAK1* (Figure 3.4A-B). *PAK1* is an oncogene that is involved in activation of the MAPK pathway[25], and has been suggested as a potential target in *BRAF* wild-type melanomas[26]. Further, *PAK1* has been identified as the most recurrently altered kinase gene via fusion events in acral melanomas[2], suggesting *PAK1* may also frequently activate the MAPK pathway outside of boundary affecting events. Similarly, *GAB2* is involved in the activation of the MAPK and PI3K/AKT pathways, and has been proposed to play a role in angiogenesis in melanomas[27]. This TAD boundary was altered in 4% of cutaneous melanomas and was 650kb away from a fragile site (FRA11H)[28]. The most recurrently altered boundary in cutaneous melanomas (all DEL events, n=7) was chr9:21700000-21775000, which is flanked by a Repressed TAD and a Low-active TAD (Figure 3.4C). This boundary is adjacent to the TADs containing the cancer genes *CDKN2A*, *CDKN2B*, and *MTAP*, all of which are tumor suppressors, and this boundary is located within a fragile site region (FRA9C)[29]. One potential mechanism of these BA-SVs is a long range silencer interaction between regulatory elements of the adjacent repressed TAD and these tumor suppressors[30]. The second most recurrently altered TAD boundary (chr22:19600000-19675000) was flanked by Active and Low-Active TADs (Figure 3.4D), and is adjacent to TADs containing the cancer genes *SEPTIN5*, *DGCR8*, and *HIRA*. Unlike the other highly recurrently altered TAD boundaries, this TAD boundary was located several megabases away from the nearest fragile site (8Mb, FRA22B). Both *DGCR8* and *HIRA* are involved in UV-induced DNA damage repair, where *DGCR8* is required for transcription-coupled nucleotide excision repair (NER) at UV-induced lesions[31], and *HIRA* is a

histone regulator required for efficiently priming chromatin for transcriptional reactivation following DNA repair at UV-induced lesions[32,33]. These results suggest an unappreciated role of BA-SVs in tumor development and progression across melanoma histological subtypes, and that BA-SVs can generate histology specific driver events in melanoma. Further, a subset of cutaneous melanomas experience BA-SVs affecting NER genes that may exacerbate the effect of UV mutagenesis on the mutational spectrum of tumors, resulting in an increase of TMB, which may also have implications for immunotherapy.

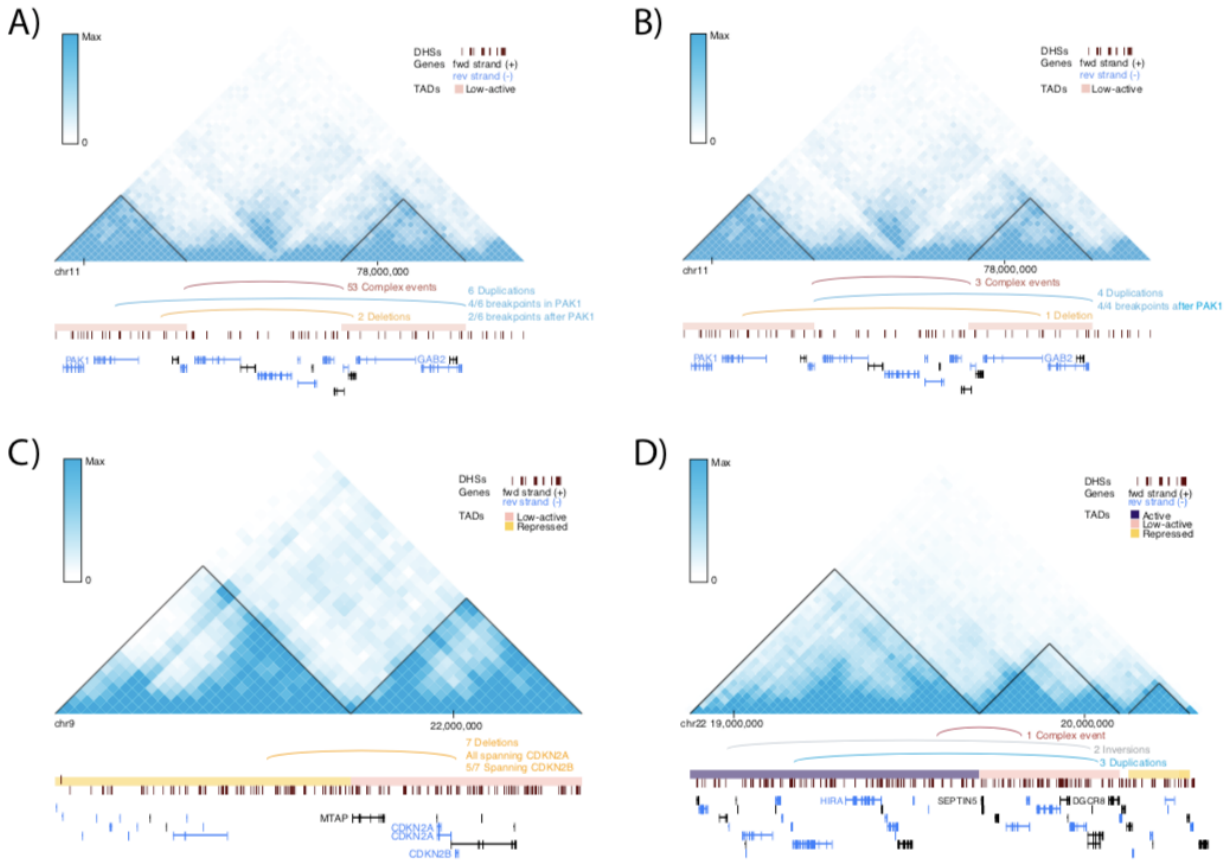


Figure 3.4: Recurrently affected boundaries adjacent cancer gene containing TADs
(a) The contact frequency map and annotations of SV events for the most recurrently altered TAD boundary in acral melanomas. **(b)** The contact frequency map and annotations of SV events for the most recurrently altered TAD boundary in mucosal melanomas. Cancer genes of interest in the adjacent TADs for both acral and mucosal melanomas include *PAK1* and *GAB2*. Both of the adjacent TADs for this boundary were low-active TADs. **(c)** The contact frequency map and annotations of SV events for the most recurrently altered TAD boundary in cutaneous melanomas. Cancer genes of interest in the adjacent TADs include *CDKN2A*, *CDKN2B*, and *MTAP*. The TAD containing these genes is a low-active TAD, and the other adjacent TAD is a

Figure 3.4 (continued): repressed TAD. **(d)** The contact frequency map and annotations of SV events for the second most recurrently altered TAD boundary in cutaneous melanomas. Cancer genes of interest in the adjacent TADs include *HIRA*, *SEPTIN5*, and *DGCR8*. *HIRA* is present in an active TAD, and *SEPTIN5* and *DGCR8* are present in a low-active TAD. The contact frequencies shown here are from the IMR90 cell line, one of the 5 cell lines used to determine the functional TAD classifications by the PCAWG consortium.

The relationship between mutational signatures and SVs in cutaneous melanoma

To further assess the potential functional impact of SVs in melanoma, we next assessed SV pattern relationships with mutational signatures. The predominant mutational signatures in cutaneous melanoma are signature 1 (aging), signature 7 (UV mutagenesis), signature 11 (alkylating), and signature 3 (DSB repair), which is enriched in TWT melanomas[15]. We previously reported an association between signature 3 and indel signature 8 (ID8; NHEJ), as well as homologous recombination deficiency associated copy number events; however, the relationship between mutational signatures and SVs in cutaneous melanoma has remained unexplored[15]. Consistent with prior analyses, mutational signature 3 was enriched in TWT cutaneous tumors in our cohort (Fisher's exact; 5/20 vs. 5/155; OR = 9.75; 95% CI = 2.00 - 47.89; $p = 1.1 \times 10^{-3}$; Figure 3.5A; Methods), and it was the only SNV signature that was associated with increased numbers of SVs per tumor, after correcting for disease stage, genomic subtype, coverage, and tumor purity (multivariate regression, $p = 3.2 \times 10^{-3}$). Specifically, this association was due to increased numbers of DUP and TRA events (multivariate regression, $p = 3.2 \times 10^{-4}$; Figure 3.5B), but not DEL or INV events (multivariate regression, $p > 0.17$). Further, when characterizing SVs as being generated by either NHEJ, MMEJ, or SSA, which are DSB repair mechanisms frequently involved in the repair of SV events and associated with distinct microhomology patterns at SV breakpoint junctions (Methods), signature 3 tumors were significantly associated with increased numbers of SVs arising from NHEJ (multivariate regression, $p = 6.7 \times 10^{-3}$), and decreased numbers of SVs arising from SSA (multivariate regression, $p = 2.8 \times 10^{-4}$). The ratio of NHEJ associated SVs to SSA associated

SVs is also significantly higher in signature 3 tumors (Wilcoxon-Mann-Whitney, $p = 1.95 \times 10^{-3}$; Figure 3.5C). Although with a smaller effect size, higher relative contribution of UV mutagenesis to the mutational spectrum of cutaneous melanomas was associated with lower numbers of SVs (multivariate regression, $p < 2.7 \times 10^{-3}$), particularly TRA and DUP events (multivariate regression, $p < 4.09 \times 10^{-5}$). There was no association between SNV mutational signatures and chromothripsis (multivariate regression, $p > 0.07$).

We then evaluated whether specific SVs affected canonical cancer genes and may directly relate to the mutational processes in cutaneous melanoma. Similar to our finding that cutaneous melanoma genomics were associated with somatic mutations in distinct secondary driver genes[15], several canonical cancer genes were also enriched for SVs on a genomic subtype basis. The most significantly enriched alteration in *BRAF* melanomas were non-duplication SV events in *CDKN2A* (39/81, 48%; Fisher's exact, OR = 2.41, 95% CI = 1.24 - 4.78, $p = 7.8 \times 10^{-3}$). Only 1 *BRAF* and 1 non-*BRAF* melanoma had duplication events overlapping *CDKN2A*. *NF1* melanomas were significantly associated with non-duplication SV events in two RASopathy genes, *RAF1* and *SPRED1* (Fisher's exact, OR = 5.03, 95% CI = 1.46 - 16.42, $p = 4.8 \times 10^{-3}$), the latter of which has also been identified as a significantly mutated gene exclusive to *NF1* melanomas[15,34]. The most statistically significant finding in TWT melanomas was *CBFA2T3*, which was not altered in any of the other genomic subtypes (Fisher's exact, OR = Inf, 95% CI = 5.69 - Inf, $p = 1.3 \times 10^{-4}$), and is a putative tumor suppressor in breast cancer[35,36]. *CBFA2T3* exclusively harbored TRA and INV events in these tumors (n=4).

MRE11A was among the cancer genes significantly enriched for SVs in TWT tumors (Fisher's exact, OR = 5.36, 95% CI = 1.01 25.18, p = 0.024) and is one of the core genes of the MRN complex, which is involved in the initial processes of double-stranded break repair prior to homologous recombination and non-homologous end joining, and is responsible for activating *ATM*[37,38]. We previously found that signature 3 in TWT tumors was associated with downregulation of *ATM*, although we were unable to identify recurrent alterations in somatic

coding regions that might explain the downregulation of *ATM* in a subset of samples[15]. Three of the 5 TWT tumors with SVs affecting *MRE11A* had detectable signature 3. Expanding the analysis to all signature 3 vs. non-signature 3 tumors also revealed the enrichment of *NBN* (Fisher's exact, OR = 7.26, 95% CI = 1.04 - 39.44, p = 0.023), another core gene of the MRN complex. All SVs affecting *MRE11A* and *NBN* were complex events (Methods), compared to less than half (43%) of non-signature 3 tumors (Fisher's exact, OR = 6.74, 95% CI = 1.42 - 32.04, $p = 7.4 \times 10^{-3}$; Figure 3.5D). Pathway overrepresentation analysis (Methods) on the set of cancer genes significantly enriched for SVs in signature 3 tumors identified the MRN complex as the top enriched protein complex ($q = 1.79 \times 10^{-3}$).

Although SVs affecting *RAD50* were not associated with signature 3 tumors, there was no difference in the association between *MRE11A* or *NBN* expression and *ATM* expression compared to the association between *RAD50* and *ATM* expression in TWT tumors (Supp. Figure 3.2). However, the correlation between MRN complex expression and *ATM* expression was significantly stronger in TWT tumors than in non-TWT tumors (r=0.82 vs r=0.69; Fisher's *Z*-transformation, p = 0.03; Figure 3.5E). To assess whether the correlation observed in TWT tumors was not spurious due to having 7-fold less samples, we performed downsampling analysis for 10,000 simulations (Methods). Only 2.57% of these simulations yielded a correlation coefficient higher than that observed for TWT tumors (p = 0.0257, Figure 3.5F). These results suggest that MRN-dependent *ATM* activation may be more frequent in TWT tumors or that *ATM* activation is more tightly regulated by the MRN complex in TWT tumors, potentially explaining why the association between signature 3 and *ATM* downregulation was restricted to TWT tumors. Additionally, these results are consistent with our previous finding that signature 3 in TWT cutaneous melanomas are associated with dysregulation of *ATM*, and affects genes that function early during the initiation process of double-stranded break repair.

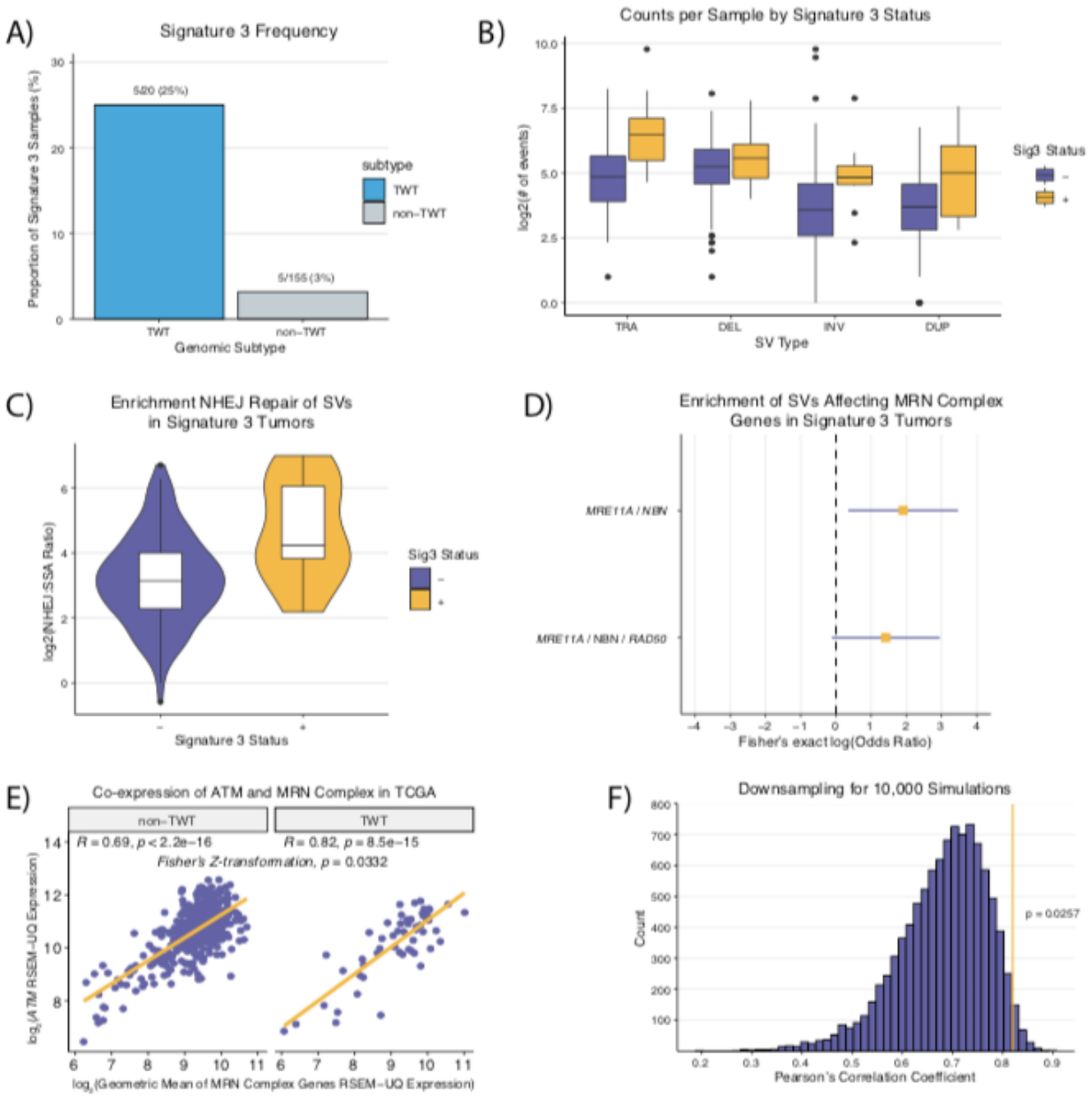


Figure 3.5: Cutaneous signature 3 tumors are enriched for SVs frequently caused by NHEJ, and are associated with SVs affecting the MRN-complex.
(a) The frequency of mutational signature 3 in TWT and non-TWT cutaneous melanomas. **(b)** The distribution of the number of events per tumor between signature 3 and non-signature 3 cutaneous melanomas, characterized by SV type. **(c)** The distribution of the ratio of putative NHEJ to SSA generated events between per tumor by signature 3 status in cutaneous melanomas. **(d)** The odds ratio (yellow square) and 95% confidence interval of the odds ratio (purple line) via Fisher's exact test for SVs overlapping MRN complex genes in signature 3 cutaneous tumors compared to non-signature 3 cutaneous tumors. **(e)** The correlation between *ATM* expression and MRN complex expression (methods) in non-TWT and TWT cutaneous melanoma tumors. **(f)** The distribution of Pearson's correlation coefficients from 10,000 **Figure**

3.5 (continued): randomly sampled simulations where non-TWT cutaneous tumors are downsampled to the number of TWT cutaneous tumors in the cohort.

Discussion

Through uniform analysis of SVs on the largest melanoma WGS cohort to date, we revealed distinct frequencies and drivers of melanoma histological and cutaneous genomic subtypes. Acral and mucosal melanomas were associated with more SVs per tumor relative to cutaneous melanomas regardless of the SV type, and acral melanomas were enriched for chromothripsis events relative to cutaneous and mucosal melanomas. Additionally, in tumors that had chromothripsis events, acral melanomas were associated with higher rates of interchromosomal chromothripsis events compared to cutaneous and mucosal melanomas. While the frequencies of SV events differed between acral and mucosal melanomas, the functional impact and driver gene alterations observed in these histological subtypes were similar. Roughly 97% and 83% of acral and mucosal melanomas, respectively, had BA-SVs that affected functionally active TADs compared to less than half of cutaneous melanomas. In cutaneous melanomas, *NF1*-mutant tumors were enriched for deletion SVs, and had chromothripsis events at nearly twice the rate compared to the other genomic subtypes. Thus, in addition to having the highest TMB of the genomic subtypes, *NF1* also has the highest SV burden[1,15].

Of the 16 genes recurrently affected (observed in at least 5 tumors) by BA-SVs in mucosal melanomas, 13 were shared with acral melanoma. One of these genes was *NF1*, one of the MAPK pathway genes used to define the cutaneous melanoma genomic subtypes. All but one BA-SV affecting *NF1* in acral and mucosal melanomas were deletion events. In addition to sharing similar recurrently affected genes, acral and mucosal melanomas also shared the the most recurrently altered TAD boundary (chr11:77750000-77825000), which is adjacent to the TADs containing *PAK1* and *GAB2*. Only *EMSY* was recurrently affected by BA-SVs in both cutaneous and mucosal melanomas, and only *CDKN2A* and *CDKN2B* were recurrently affected

by BA-SVs in both acral and cutaneous melanomas. Thus, despite having drastically different SV landscapes, acral and mucosal melanomas share many of the same driver SVs. This is akin to the shared somatic mutation derived driver genes from these subtypes[2,39].

The most recurrently affected genes by BA-SVs in cutaneous melanomas were *CDKN2A* and *CDKN2B*, with the majority being deletion events. *CDKN2A* was also enriched for non-duplication events (39/81; 48%) in *BRAF*-mutant cutaneous melanomas compared to the other genomic subtypes. Notably, *CDKN2A* has also been identified as a canonical driver in cutaneous melanoma via analysis of somatic mutations[1,15]. The second most recurrently altered TAD boundary in our cutaneous melanoma cohort affected the NER genes *HIRA* and *DGCR8*, which are involved in the repair of mutations caused by UV mutagenesis. Increased activity of UV mutagenesis is associated with higher TMB[40], and therefore may have implications for immunotherapy treatment decisions or response. Other than *MTAP* and *SEPTIN5*, these were the only recurrently altered cutaneous melanoma genes that were not recurrently altered in acral or mucosal melanomas, which frequently lack the presence of UV-induced mutations.

Mutational significance analysis in cutaneous melanoma has revealed that the genomic subtypes preferentially experience mutations that affect distinct pathways. *NF1* melanomas preferentially experienced alterations in RASopathy genes, with *SPRED1*, *RASA2*, and *RASSF2* being identified as significantly mutated genes within the subtype[15]. *NF1* melanomas in our cohort were enriched for SV events affecting *SPRED1* and *RAF1*, the latter of which has been implicated in activating fusion events in cutaneous melanoma, and enriched in TWT tumors[41].

A subset of cutaneous melanoma tumors have been characterized as having mutational signature 3 (associated with DSB repair deficiency), enriched in TWT tumors[15]. While the prevalence of signature 3 has been characterized in cutaneous melanoma WGS samples, its association with SVs has remained unexplored. Here we show that signature 3 is associated with increased DUP and TRA events in melanoma, and is associated with a higher rate of the

error-prone NHEJ repair. However, we did not find a significant association between signature 3 and chromothripsis, despite prior studies linking HRD to increased prevalence of chromothripsis[42]. Signature 3 in TWT tumors is associated with downregulation of *ATM* and methylation of *INO80*, however the source of *ATM* downregulation is unknown. Here we identified the enrichment of SVs affecting MRN complex genes in signature 3 tumors, which directly interacts with *ATM*. Like *ATM* and *INO80*, the MRN complex functions early in the DSB repair pathway, providing further evidence for the source of signature 3.

Overall we demonstrated that SV analysis of melanoma whole-genomes can identify additional driver mechanisms unique to histological and cutaneous genomic subtypes, some of which may present as clinically relevant druggable events. Still, further experimental work in preclinical models will be required to determine the therapeutic relevance of MRN complex alterations and *ATM* downregulation in melanoma, as well as the functional consequences of SVs at recurrently altered TAD boundaries. Furthermore, the number of whole exome samples far exceeds the number of WGS samples in melanoma. Continued harmonized molecular analysis of a larger melanoma WGS cohorts will help determine the robustness and true prevalence of driver alterations identified in this study.

Methods

Whole-genome sequencing dataset description

We downloaded publicly available aligned WGS BAM files from 4 previously published studies. For SV analysis, we required both tumor and normal samples to have a sequence coverage of at least 20X, and a tumor purity of at least 20%. The median sequencing coverage was 57X in the tumor samples and 37X in the normal samples. The median tumor purity ranged from 61% in mucosal melanomas to 66% in acral melanomas.

The cutaneous melanoma mutation data, which was used to determine genomic subtype and identify mutational signatures (see Mutational signatures), was downloaded from the

supplement of Hayward *et al*. 2017[2], and the ICGC Data Portal (https://dcc.icgc.org/)[43] for TCGA-SKCM WGS samples.

The cutaneous melanoma expression data used in this study is from the TCGA-SKCM cohort, which is publicly available from the TCGA-SKCM workspace on FireCloud (TCGA_SKCM_ControlledAccess_V1-0_DATA) via dbGaP access. The RSEM upper quartile normalized expression data was used for all expression analysis in this study.

SV calling

We called SVs with three different SV calling methods: Manta (https://github.com/Illumina/manta)[44], DELLY2 (https://github.com/dellytools/delly)[45], and SvABA (https://github.com/walaj/svaba)[46]. To identify a set of high confidence SVs per tumor we filtered the calls to only keep SVs identified by 2 or more methods, allowing for a maximum distance of 1kb pairwise between breakpoints and requiring that the calls agree on type, strand, and are at least 30bp long. This filtering was performed using the SURVIVOR R package (https://github.com/fritzsedlazeck/SURVIVOR)[47].

Copy number calling

Allelic copy number calls were determined using FACETS (https://github.com/mskcc/facets)[16], which also provides tumor purity and ploidy information. These copy number calls were used as input to ShatterSeek[10] (see Identification of chromothripsis events) for identifying the oscillating copy number criteria of chromothripsis events.

Identification and visualization of chromothripsis events

To identify chromothripsis events in melanoma cancer genomes we ran ShatterSeek (https://github.com/parklab/ShatterSeek)[10] using the high confidence SVs and allelic copy number data as input. ShatterSeek was also used to visualize chromothripsis events on single

chromosomes, such as in Figures 2E-F. To visualize interchromosomal chromothripsis events we used the circos tool on Galaxy (https://usegalaxy.org/)[48]. Non-chromothripsis complex events were defined as overlapping concomitant DEL, DUP, INV, or TRA events.

SV annotations

To add gene level annotations to our high confidence SV set, we ran AnnotSV v3.0(https://lbgi.fr/AnnotSV/)[49] using the default set of hyperparameters. The SV annotations were run on December 29th, 2020.

TAD and TAD boundary assignments and TAD annotations

TAD and TAD boundary assignments, as well as TAD type annotations were downloaded from Akdemir *et al*. 2020[11]. Here, TAD and TAD boundary coordinate assignments were determined by identifying TAD boundaries that were within 50kb of each other across Hi-C data from 5 different cell types (GM12878, HUVEC, IMR90, HMEC and NHEK). TAD type annotations (Heterochromatin, Low, Repressed, Low-Active, and Active) were determined by k-means clustering to the 15 state ChromHMM model from the Roadmap Epigenomics Project[24], and associating the clusters with gene expression data from GTEx[50] and ICGC[43].

Short range SVs likely to only affect a single TAD boundary were classified as < 2Mb in length, and were the only types of SVs used in the boundary affecting analysis. The cutoff of < 2Mb was defined by the PCAWG consortium (< 2Mb)[11]. For a short range SV to be considered boundary affecting, the entire TAD boundary had to be overlapped by the SV.

Fragile site annotations

Fragile site annotations were obtained from https://webs.iiitd.edu.in/raghava/humcfs/[51]. Specifically, we used the “Fragile site bed files” reference, which provides a directory of bed files containing fragile site regions on a per chromosome basis.

Classification of double-stranded break repair mechanisms

To classify SVs as being repaired by NHEJ, MMEJ, or SSA, we applied the breakpoint microhomology cutoffs identified in Li *et al*. 2020[7], which were determined by fitting linear functions to breakpoint microhomology data across PCAWG. This resulted in the identification of 3 sets of structural variants defined by microhomologies of 1 bp, 2-9 bp, and 10 or more bp, which were classified as NHEJ, MMEJ, and SSA, respectively.

Mutational signatures

To identify mutational signatures present in tumor samples we ran deconstructSigs (https://github.com/raerose01/deconstructSigs)[52] using the COSMIC v2 signatures reference[53,54] and a signature contribution cutoff of 0.06. This contribution cutoff provides a false-positive rate of 0.1% and false-negative rate of 1.4%, and is the recommended cutoff.

Pathway over-representation analysis

We performed pathway over-representation analysis on the set of cancer genes enriched in signature 3 tumors via Fisher's exact method using ConsensusPathDB (v. 34) (http://cpdb.molgen.mpg.de)[55]. We ran ConsensusPathDB (on May 18th, 2021) using the default parameters for both pathway-based gene sets and protein complex-based gene sets.

Expression correlation analysis

We performed correlation between *ATM* expression and MRN complex gene using the TCGA-SKCM RSEM upper quartile normalized RNA-seq data. To calculate one single expression for the entire MRN complex we calculated the geometric mean of *MRE11*, *NBN*, and *RAD50*. Correlation was calculated using the stats R package, and the geometric mean was calculated using the psych R package.

Gene sets

The oncogene and tumor suppressor gene sets used in the BA-SV analysis were downloaded from MSigDB[56,57] on May 27th, 2021 under the curated Gene Families (https://www.gsea-msigdb.org/gsea/msigdb/gene_families.jsp). The set of cancer genes were determined by taking the union of Cancer Gene Census (v. 86) genes and OncoKB[58] cancer genes.

Statistics and reproducibility

Statistical analyses were performed using the stats R package for R v.3.6.1. Reported q-values represent FDR-corrected p-values and reported p-values represent nominal p-values. All statistical tests performed (for example, Wilcoxon-Mann-Whitney, Kolmogorov–Smirnov, Fisher's exact test) were two sided.

Declarations

Data Availability

All of the datasets used in this study are publicly available. The raw sequence data can be obtained through dbGaP (https://www.ncbi.nlm.nih.gov/gap) and the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (https://icgc.org/daco), or as described in their original papers. Publicly available databases used in this study include MSigDB v6.2, AnnotSV v3.0, and ConsensusPathDB v34 (http://cpdb.molgen.mpg.de/).

Code Availability

All software and bioinformatic tools used in this study are publicly available.

Acknowledgements

This work was supported by NCI F31CA239347 (J.R.C.), NIH 5T32HG002295-15 (J.R.C.), NIH R01CA227388-02 (E.M.V.A.), NIH R21CA242861 (E.M.V.A.) and the Damon Runyon Clinical Investigator Award (E.M.V.A.). The results presented in this study are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author Contributions

J.R.C. and J.C. contributed to the analysis of genomic data. J.R.C. and E.M.V.A. contributed to aggregation of processed and raw sequence data. J.R.C., J.C., R.G., B.R., and E.M.V.A. contributed to interpretation of results, and manuscript preparation.

Competing Interests

E.M.V.A. is a consultant for Tango Therapeutics, Genome Medical, Invitae, Enara Bio, Monte Rosa Therapeutics, Manifold Bio, and Janssen. E.M.V.A. provides research support to Novartis and Bristol-Myers Squibb. E.M.V.A. has equity in Tango Therapeutics, Genome Medical, Syapse, Ervaxx, and Microsoft. E.M.V.A. receives travel reimbursement from Roche/Genentech. E.M.V.A. has institutional patents filed on methods for clinical interpretation. J.R.C. is a consultant for Tango Therapeutics.

Bibliography

1. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696 (2015).

2. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).

3. Zaretsky, J. M. *et al.* Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).

4. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma.

New England Journal of Medicine vol. 371 2189–2199 (2014).

5. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).

6. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **168**, 542 (2017).

7. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

8. Newell, F. *et al.* Whole-genome landscape of mucosal melanoma reveals diverse drivers and therapeutic targets. *Nat. Commun.* **10**, 3163 (2019).

9. Newell, F. *et al.* Whole-genome sequencing of acral melanoma reveals genomic complexity and diversity. *Nat. Commun.* **11**, 5259 (2020).

10. Cortés-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* **52**, 331–341 (2020).

11. Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* **52**, 294–305 (2020).

12. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).

13. Gröschel, S. *et al.* A Single Oncogenic Enhancer Rearrangement Causes Concomitant EVI1 and GATA2 Deregulation in Leukemia. *Cell* vol. 157 369–381 (2014).

14. Gillani, R. *et al.* Gene Fusions Create Partner and Collateral Dependencies Essential to Cancer Cell Survival. *Cancer Res.* **81**, 3971–3984 (2021).

15. Conway, J. R. *et al.* Integrated molecular drivers coordinate biological and clinical states in melanoma. *Nat. Genet.* **52**, 1373–1383 (2020).

16. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).

17. Sztupinszki, Z. *et al.* Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ Breast Cancer* **4**, 16 (2018).

18. Zhou, R. *et al.* Analysis of Mucosal Melanoma Whole-Genome Landscapes Reveals Clinically Relevant Genomic Aberrations. *Clin. Cancer Res.* **25**, 3548–3560 (2019).

19. Govind, S. K. *et al.* ShatterProof: operational detection and quantification of chromothripsis. *BMC Bioinformatics* **15**, 78 (2014).

20. Van Allen, E. M. *et al.* The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discov.* **4**, 94–109 (2014).

21. Haller, F. *et al.* Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. *Nat. Commun.* **10**, 368 (2019).

22. Taberlay, P. C. *et al.* Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res.* **26**, 719–731 (2016).

23. Valton, A.-L. & Dekker, J. TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36**, 34–40 (2016).

24. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

25. Shrestha, Y. *et al.* PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling. *Oncogene* **31**, 3397–3408 (2012).

26. Ong, C. C. *et al.* P21-activated kinase 1 (PAK1) as a therapeutic target in BRAF wild-type melanoma. *J. Natl. Cancer Inst.* **105**, 606–607 (2013).

27. Yang, Y. *et al.* GAB2 induces tumor angiogenesis in NRAS-driven melanoma. *Oncogene* **32**, 3627–3637 (2013).

28. Bester, A. C., Kafri, M., Maoz, K. & Kerem, B. Infection with retroviral vectors leads to perturbed DNA replication increasing vector integrations into fragile sites. *Sci. Rep.* **3**, 2189 (2013).

29. Sutherland, G. R., Parslow, M. I. & Baker, E. New classes of common fragile sites induced by 5-azacytidine and bromodeoxyuridine. *Hum. Genet.* **69**, 233–237 (1985).

30. Segert, J. A., Gisselbrecht, S. S. & Bulyk, M. L. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends Genet.* **37**, 514–527 (2021).

31. Calses, P. C. *et al.* DGCR8 Mediates Repair of UV-Induced DNA Damage Independently of RNA Processing. *Cell Rep.* **19**, 162–174 (2017).

32. Adam, S., Polo, S. E. & Almouzni, G. Transcription Recovery after DNA Damage Requires Chromatin Priming by the H3.3 Histone Chaperone HIRA. *Cell* vol. 155 963 (2013).

33. Ray-Gallet, D. *et al.* Functional activity of the H3.3 histone chaperone complex HIRA requires trimerization of the HIRA subunit. *Nat. Commun.* **9**, 3103 (2018).

34. Krauthammer, M. *et al.* Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat. Genet.* **47**, 996–1002 (2015).

35. Kochetkova, M. *et al.* CBFA2T3 (MTG16) is a putative breast tumor suppressor gene from the breast cancer loss of heterozygosity region at 16q24.3. *Cancer Res.* **62**, 4599–4604 (2002).

36. Smith, J. L. *et al.* Comprehensive Transcriptome Profiling of Cryptic Fusion-Positive AML Defines Novel Therapeutic Options: A COG and TARGET Pediatric AML Study. *Clin. Cancer Res.* **26**, 726–737 (2020).

37. Bian, L., Meng, Y., Zhang, M. & Li, D. MRE11-RAD50-NBS1 complex alterations and DNA damage response: implications for cancer treatment. *Mol. Cancer* **18**, 169 (2019).

38. Uziel, T. *et al.* Requirement of the MRN complex for ATM activation by DNA damage. *EMBO J.* **22**, 5612–5621 (2003).

39. Rabbie, R., Ferguson, P., Molina-Aguilar, C., Adams, D. J. & Robles-Espinoza, C. D. Melanoma subtypes: genomic profiles, prognostic molecular markers and therapeutic possibilities. *J. Pathol.* **247**, 539–551 (2019).

40. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

41. Williams, E. A. *et al.* Melanomas with activating RAF1 fusions: clinical, histopathologic, and molecular profiles. *Mod. Pathol.* **33**, 1466–1474 (2020).

42. Simovic, M. & Ernst, A. Chromothripsis, DNA repair and checkpoints defects. *Semin. Cell Dev. Biol.* (2021) doi:10.1016/j.semcdb.2021.02.001.

43. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).

44. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

45. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

46. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).

47. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).

48. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).

49. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).

50. Consortium, T. G. & The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* vol. 369 1318–1330 (2020).

51. Kumar, R. *et al.* HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics* **19**, 985 (2019).

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

53. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

54. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).

55. Herwig, R., Hardt, C., Lienhard, M. & Kamburov, A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature Protocols* vol. 11 1889–1907 (2016).

56. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

57. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

58. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).

Chapter 4: Clonal Architecture

Abstract

The extent to which clinical and genomic characteristics relate to prostate cancer clonal architecture, tumor evolution, and therapeutic response remains unclear. Here, we reconstructed the clonal architecture and evolutionary trajectories of 845 prostate cancer tumors with harmonized clinical and molecular data. We observe that primary tumors in Black patients are associated with higher biochemical recurrence rates, yet in contrast to prior observations relating polyclonal architecture and adverse clinical outcomes, demonstrate these tumors to be more linear and monoclonal. Additionally, we demonstrate that a novel approach to evolutionarily informed mutational signature analysis that leverages clonal architecture can uncover additional cases of homologous recombination deficient and mismatch repair deficient tumors in primary or metastatic tumors, link the origin of mutational signatures to specific subclones, and may have immediate therapeutic implications. Broadly, clonal architecture and

evolutionary informed analysis reveal novel biological insights that are clinically actionable and provide multiple opportunities for subsequent investigation.

Introduction

Tumors consist of cell subpopulations that are characterized by a variety of genomic and epigenomic features including single nucleotide variants (SNVs), insertions and deletions (indels), copy number alterations (CNAs), structural variants (SVs), and methylation profiles[1]. The aggregate of these subpopulations defines the clonal architecture of a tumor, and inherently delineates intra-tumoral heterogeneity. Whole-exome sequencing can provide a snapshot of these cell subpopulations at a point in time and space, and computational methods aimed at determining the number of tumor cell subpopulations and their relative cellular frequencies ("clonal architecture"), as well as inferring linear or branched phylogenetic evolutionary trajectories[2–4], can improve our understanding of tumor evolution. Understanding the relationship between tumor clonal architecture and patient clinical characteristics may provide insight into disease trajectory, outcomes, and inform treatment decision-making[5].

Prostate cancer (PC) is the second leading cause of cancer mortality in men[6]. Genomically-informed therapeutic options in advanced PC have been limited until recently with the approval of immune checkpoint blockade with pembrolizumab for tumors with deficient DNA mismatch repair/microsatellite instability (MSI)[7], or high tumor mutational burden, and PARP inhibition for prostate tumors harboring certain deleterious alterations in homologous recombination (HR) genes[8,9]. Despite their overall benefit, responses to these therapies are heterogeneous[8–10], and little is known to date regarding how these clinical phenotypes relate to the clonal architecture of the tumors.

A prior study leveraged a large cohort of localized PC tumors (n = 293) to analyze associations between clonal architecture and clinical (e.g., Gleason score) or genomic (e.g. mutational signatures) covariates. This work observed that polyclonality is associated with

increased risk of biochemical recurrence after definitive therapy, and that the mutational signature spectrum of a subset of primary tumors shifted over time from clock-like and APOBEC signatures to that of homologous recombination deficiency (HRD)[11]. However, clonal architecture analyses to date have been limited to primary tumors, did not consider the potential effect of ancestry, and were not guided by cell subpopulation inference classifications. We hypothesized that clonal architecture analysis could improve the understanding of race or ancestrally distinct routes to oncogenesis as well as clinical responses to emerging treatment paradigms in metastatic PC. Though spatio-genomic data consisting of multi-focal biopsies from a single tumor are preferred for evolutionary analysis due to the increased power to resolve the true clonality of genomic alterations[12,13], the number of PC patients with this data remains limited. Therefore, we paired harmonized molecular and clinical data from 845 primary and metastatic PC tumors[14] with novel computational methodologies to determine how clinical and genomic components relate to PC clonal architecture and evolutionary dynamics.

Results

Localized Prostate Cancer Clinical Risk Groups and Clonal Architecture

We first evaluated previously identified[11] associations between clonal architecture (Figure 4.1A) and evolutionary trajectories (Figure 4.1B) with clinical characteristics in localized (primary) PCs. All primary PCs in this cohort were treatment naive radical prostatectomy specimens. Based on phylogenetic reconstruction (Methods), tumors were defined as (1) monoclonal vs polyclonal; the former if only a single cell cluster was identified, and (2) having evidence of linear vs branched evolution; the former if each cell subpopulation had a maximum of 1 child subpopulation. We did not detect a statistically significant association between clonal architecture and either age of onset (early: 55 or younger; late: older than 55; univariate linear regression, $p = 0.07$)[15] or ETS fusion status in our cohort ($n = 521$; univariate linear regression, $p = 0.1$). However, we did identify significant associations between clonal architecture and

NCCN clinical risk categories (n = 468 primary tumors with risk information, Supplemental Tables 4.1-2). Approximately 17%, 10%, and 5% of low, intermediate, and high-risk primary tumors were classified as monoclonal, respectively. NCCN low-risk tumors were significantly associated with being monoclonal relative to high-risk tumors in univariate analysis (Figure 4.1C; Fisher's exact test; 95% CI = 1.26-11.88, OR = 3.93; $p = 8.6 \times 10^{-3}$) and were significantly associated with fewer cell subpopulations (i.e., subclones), after adjusting for mutational burden, tumor purity, and coverage (Figure 4.1D; linear regression, $p = 4.6 \times 10^{-4}$). In contrast, intermediate-risk tumors were not associated with monoclonal architecture (Fisher's; 95% CI = 0.68-4.92, OR = 1.92; p = 0.19). Despite differences in the clonal architecture of tumors by risk group, there was no significant association between risk groups and evolutionary trajectories (linear vs. branched evolution; Fisher's, $p > 0.28$ pairwise for all).

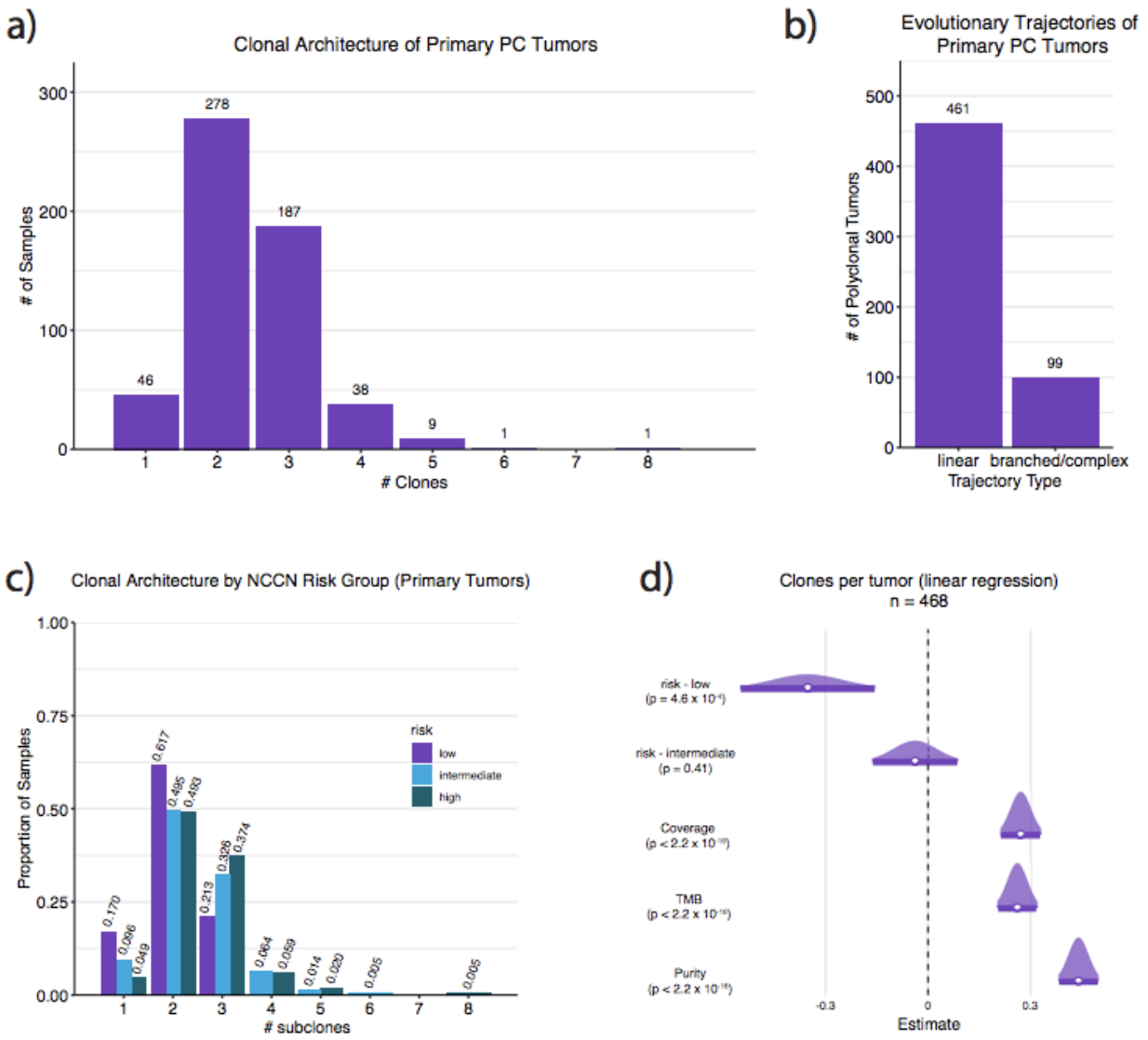


Figure 4.1: The Clonal Architecture and Evolutionary Trajectories of Prostate Cancer, and associations with clinical risk groups

(a) The distribution of the number of cell subpopulations (clonal architecture) per tumor across the cohort of 845 tumors. Of the 845 tumors, 50 (6%) were monoclonal, 303 (36%) were biclonal, and 492 (58%) were polyclonal. **(b)** Roughly three-quarters of PC tumors exhibited linear evolutionary trajectories rather than branched/complex evolutionary trajectories. **(c) T**he distribution of the number of cell subpopulations (subclones) per tumor by risk group for all primary tumors with risk information in our cohort (n = 468). Low-risk tumors were significantly associated with being monoclonal relative to high-risk (Fisher's; 95% CI = 1.26-11.88, OR = 3.93; $p = 8.6 \times 10^{-3}$), but not intermediate-risk (Fisher's; 95% CI = 0.68-4.92, OR = 1.92; p = 0.19) tumors. **(d)** Low-risk tumors were also significantly associated with lower numbers of cell subpopulations per tumor after adjusting for confounding covariates, such as mutational burden, tumor purity, and coverage (linear regression, $p = 4.6 \times 10^{-4}$).

Self-Reported Race and Ancestry

The association between self-reported race, genetic ancestry, and outcomes in men with prostate cancer is complex. Black men have substantially higher rates of prostate cancer incidence and mortality compared to other racial groups, which is considerably influenced by disparities in access to and utilization of equitable healthcare as well as other socioeconomic and health-related factors that influence environmental exposures. Due to the paucity of AA tumors in published genomic studies of PC, whether genetic underpinnings relate to these clinical outcome disparities is poorly understood[16]. We therefore evaluated whether there were associations between self-reported race, genetic ancestry and clonal architecture that may contribute to prostate cancer outcome disparities by race. Importantly, we note that our data set lacked information on the numerous non-genomic factors described above in these analyses, and therefore these results are limited to the available covariates of NCCN risk category and sequencing characteristics such as tumor coverage and purity.

Of the 560 primary PCs in our cohort, 112 were from men with self-reported Black race (Supp. Table 4.1). Black patients were approximately twice as likely to experience biochemical recurrence (Fisher's, 24.4% vs. 14.3%, CI = 1.01-3.61; OR = 1.93, p = 0.03; Supp. Figure 4.1A), even after adjusting for clinical risk groups (logistic regression, $p = 1.3 \times 10^{-4}$), had a significantly higher rate of early onset PC development (Fishers, 36.6% vs. 27%, CI = 0.98 - 2.46, OR = 1.56, p = 0.048, Supp. Figure 4.1B), and had slightly lower mutational burden (Mann-Whitney U, 1.04 mut/Mb vs. 1.27 mut/Mb, $p = 1.13 \times 10^{-7}$; Supp. Figure 4.1C) There was no difference in genomic instability, as measured by proportion genome altered, between Black and non-Black patient tumors (Mann-Whitney U, 13.5% vs. 14.9%, p = 0.14), consistent with previous reports[17].

With respect to clonal architecture, tumors from Black patients were borderline significantly associated with lower numbers of clones after adjusting for clinical risk group, mutational burden, tumor purity, and sequencing coverage (linear regression, p = 0.06, Figure

4.2A). Additionally, mutations in Black patient tumors were found at higher cancer cell fractions (CCFs) (Kolmogorov-Smirnov, $p < 2.2 \times 10^{-16}$), and this relationship remained significant after adjusting for confounding covariate (linear regression, p = 0.025, Figure 4.2B; Supp. Table 4.3). The overall CCFs of tumor cell subpopulations were also significantly higher in tumors from Black patients (Kolmogorov-Smirnov, p = 0.013, Figure 4.2C). Furthermore, the fraction of mutations classified as clonal per tumor was significantly higher in Black patients (Mann-Whitney U, $p = 2.77 \times 10^{-6}$, Figure 4.2D), even when removing monoclonal tumors from the analysis (Kolmogorov-Smirnov, p = 0.002), and this relationship too remained significant when adjusting for confounding covariates (linear regression, p = 0.034, Figure 4.2E; Supp. Table 4.3).

When comparing tumor phylogenetics, Black patient tumors were significantly associated with linear evolutionary trajectories (Fisher's; 95% CI = 1.29-6.24; OR = 2.67; $p = 4.8 \times 10^{-3}$, Methods), and this association still held after correcting for confounding covariates (logistic regression, p = 0.032; Figure 4.2F). Continental ancestry assignments via ancestry inference analysis (Supp. Figure 4.2) revealed these same statistical associations with PC clonal architecture and evolutionary trajectory (Methods, Supp. Table 4.3).

Race is a social construct, and the racial group a patient identifies with is influenced by a multitude of non-biological factors, thus confounding the association between race and clinical outcomes. Therefore, we also tested the association between genetic ancestry and PC clonal architecture by performing local ancestry admixture estimation on our primary PC patients (Methods). Higher AFR ancestry admixture proportions were significantly associated with lower numbers of cell subpopulations per tumor (linear regression, p = 0.02, Figure 4.2G, Supp. Table 4.3), higher proportions of clonal mutations per tumor (linear regression, p = 0.022, Figure 4.2H), and linear evolutionary trajectories as opposed to branched/complex trajectories (logistic regression, p = 0.014, Figure 4.2I; Supp. Table 4.3) These associations were strictly due to AFR admixture proportions, not other ancestry admixture proportions (Methods), although this

analysis also revealed significant inverse associations with these features (less subclones and lower proportions of clonal mutations per tumor) for the EUR admixture proportions (Supp. Table 4.3).

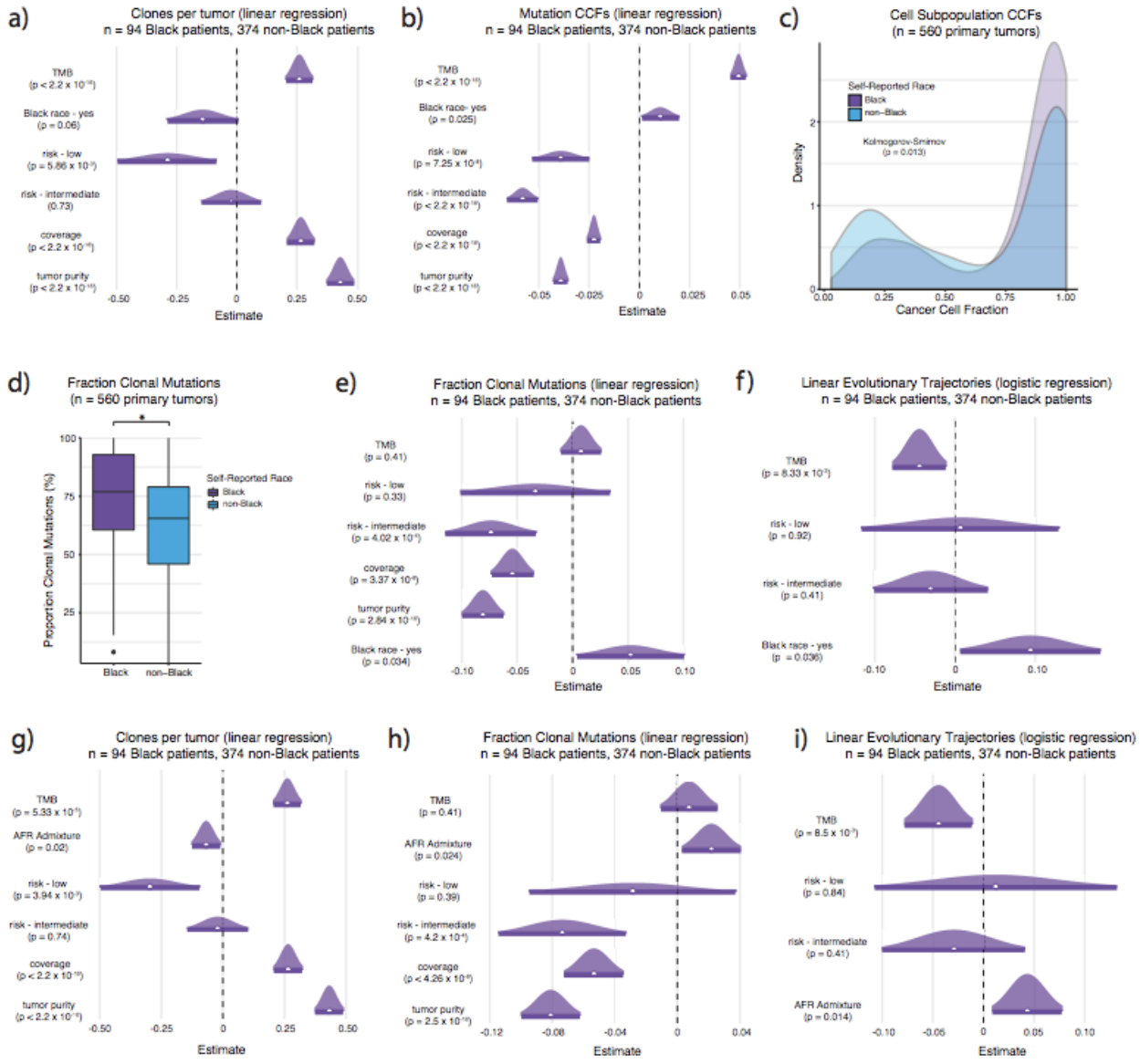


Figure 4.2: Ancestry influences the clonal architecture of primary PCs and tumors in Black men preferentially exhibit evidence of punctuated evolutionary trajectories marked by rapid selective sweeps
(a) Self-report Black race was borderline significantly associated with less cell subpopulations (clones) per tumor (linear regression, p = 0.06), and **(b)** significantly associated with higher mutation CCFs compared to non-Black race (linear regression, p = 0.025) after correcting for confounding covariates. **(c)** The CCFs of cell subpopulations, not just individual mutations, was also significantly higher tumors in Black patients compared to non-Black patients (Kolmogorov-Smirnov, p = 0.013). **(d)** The fraction of clonal mutations per tumor was

Figure 4.2 (continued): significantly higher in Black patient samples compared to non-Black tumor samples (Mann-Whitney U, $p = 2.77 \times 10^{-6}$), **(e)** even after correcting for confounding covariates (linear regression, p = 0.034). **(f)** Reported Black race was also associated with a higher frequency of linear evolutionary trajectories, rather than branched/complex evolutionary trajectories, compared to non-black patient tumors (logistic regression, p = 0.036). **(g)** Higher AFR admixture proportions were significantly associated with less subclones per tumor (linear regression, p = 0.02), **(h)** a higher fraction of clonal mutations per tumor (linear regression, p = 0.024), and **(i)** a higher frequency of linear evolutionary trajectories after correcting for confounding covariates (logistic regression, p = 0.014).

A 3.8 Mb region on chromosome 8q24 has been associated with prostate cancer risk in AA men, most notably the SNP allele A at rs1447295 [18]. Thus, we imputed the genotype at this locus in primary PC samples that had sufficient off target coverage (n = 85, Methods). AFR tumor samples in our cohort were enriched for the A allele (21/33, 63.6%) at rs1447295 compared to EUR patient (15/52, 28.8%) samples (Fisher's exact, 95% CI = 1.55 - 12.16, OR = 4.24, $p = 3.1 \times 10^{-3}$**)** Further, of the 4 samples that were homozygous for the A allele at rs1447295, 3 were of AFR ancestry. The A allele at rs1447295 was associated with fewer subclones per tumor (univariate linear regression, $p = 7.3 \times 10^{-3}$), as well as both the A/A and A/C genotypes relative to the C/C genotype (univariate linear regression, A/A: p = 0.011, A/C: p = 0.026). However, when correcting for confounding covariates (excluding risk, which rs1447295 is strongly associated with[19–22]), only the A allele in general (linear regression, p = 0.039), but not the A/A or A/C genotypes (linear regression, A/A: p = 0.063, A/C: p = 0.081) were associated with fewer subclones per tumor. This difference is likely the result of our imputation cohort being over 5-fold smaller than our original discovery cohort (n = 468 vs. n = 85, 5.5-fold smaller). Proportion of clonal mutations per tumor showed a similar effect, where both the A allele (univariate linear regression, $p = 5.4 \times 10^{-3}$) and genotype (univariate linear regression, A/A: p = 0.017, A/C: p = 0.019) were significantly associated with higher proportions of clonal mutations per tumor in the univariate setting, while only the A allele (linear regression, p = 0.033) but not the genotypes (linear regression, A/A: p = 0.064, A/C: p = 0.067) was significantly associated with higher proportions of clonal mutations after correcting for confounding

covariates. We did not observe a significant association between the risk A allele and evolutionary trajectories (branched/complex vs. linear). All A/A genotype samples had linear trajectories, and 89% of A allele containing samples had linear evolutionary trajectories compared to 83% of non-A allele samples (Fisher's exact, 95% CI = 0.37 - 7.7, OR = 1.55, p = 0.55).

To validate the associations between clonal architecture and ancestry, and demonstrate that our findings were not a consequence of only sequencing coding regions, we performed the same analysis on a small cohort of AFR (n=7) and EUR (n=7) primary prostate cancer whole-genomes[23]. While mutation CCFs and the CCFs of tumor cell subpopulations were slightly higher in EUR tumors, AFR patient tumors had fewer subclones per tumor, and a higher fraction of mutations classified as clonal per tumor (Supp. Figure 4.3A). These associations were borderline significant after correcting for tumor purity, coverage, and mutational burden (linear regression; fraction of clonal mutations: p = 0.08, Supp. Figure 4.3B; number of clones: p = 0.11). Although there were no monoclonal tumors in this cohort, 5 of 7 (71%) AA tumors were bi-clonal compared to 3 of 7 (43%) European ancestry tumors. Further, all AA tumors (100%) had linear evolutionary trajectories compared to 3 of 7 (43%) European ancestry tumors (Fisher's; 95% CI = 0.85 - Inf; OR = Inf; p = 0.07; Figure 4.3). Together, these findings suggest an association between tumors from AFR and/or Black patients and linear as well as monoclonal genomic architecture. This finding contrasts with the positive correlation previously observed between polyclonality and more aggressive disease as measured by biochemical recurrence rate among White patients[11].

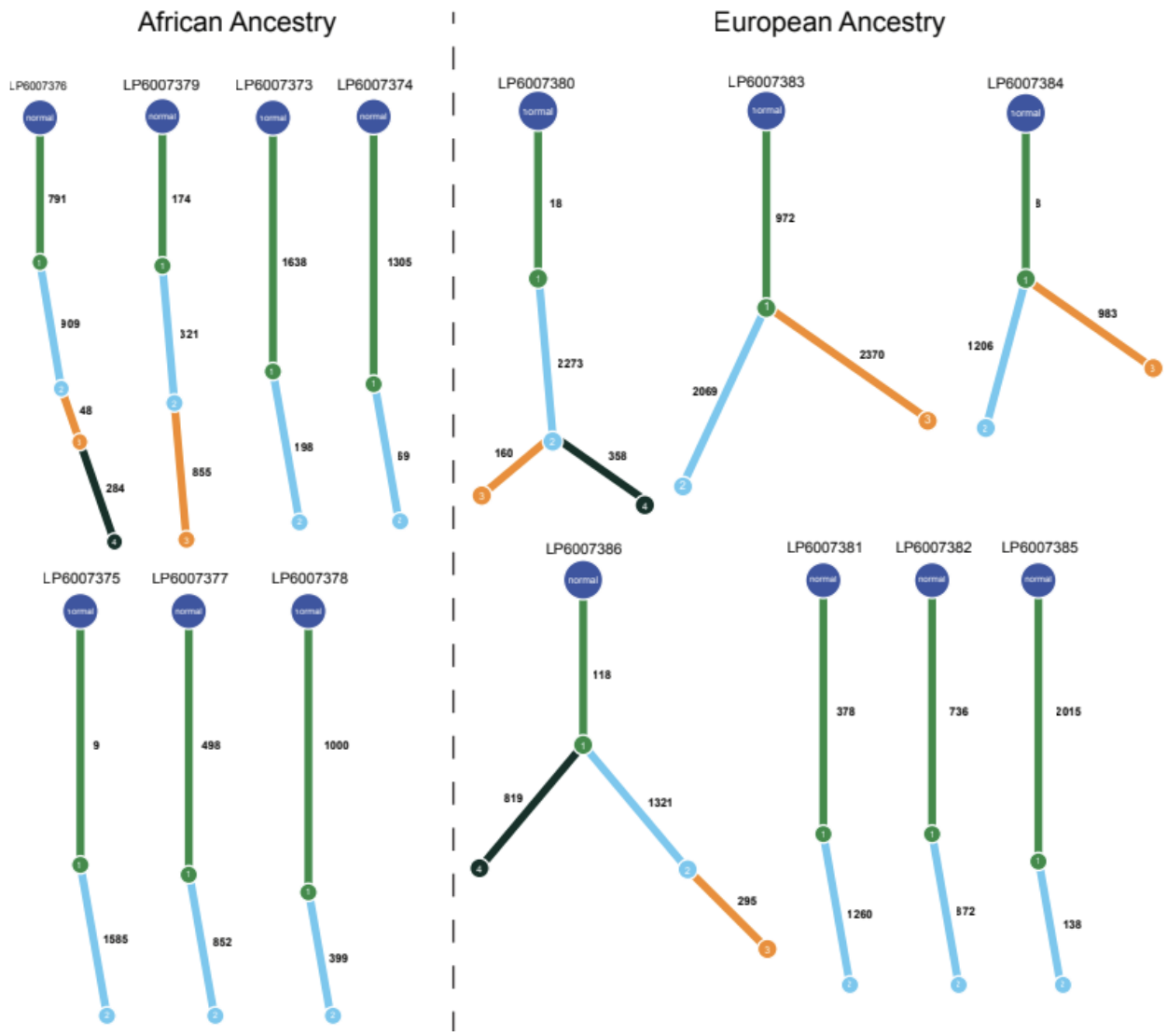


Figure 4.3: Evolutionary trajectories in African and European ancestry whole genome sequencing samples from primary tumors
A total of 2 out of 7 African ancestry samples had more than 2 cell subpopulations, compared to 4 out of 7 European ancestry samples. In both African ancestry cases with more than 2 cell subpopulations the tumor had a linear evolutionary trajectory, whereas all 4 European ancestry cases with more than 2 cell subpopulations had branched evolutionary trajectories (Fisher's, p = 0.07).

Primary vs. Metastatic

We next aimed to compare clonal architecture and evolutionary trajectories between primary (N = 560) and metastatic (N = 285) PC tumors (Supp. Table 4.1). Primary tumors had significantly higher rates of monoclonal tumors (Fisher's; 10% vs. 2.5%; 95% CI = 1.93-11.40,

OR = 4.32; p = 3.98 x 10^{-5}), and less tumor cell subpopulations in general (Kolmogorov-Smirnov, p < 2.2 x 10^{-16}; Supp. Figure 4.4B). When restricting to polyclonal tumors (> 2 subclones), to prevent the increased frequency of monoclonal and bi-clonal tumors in primary PCs from biasing the analysis, metastatic tumors were still significantly associated with branching patterns of evolution compared to primary tumors (Fisher's; 68% vs. 52%; 95% CI = 1.22-3.05, OR = 1.92; p = 0.003), whereas primary tumors were significantly associated with linear evolution. When including monoclonal and bi-clonal tumors, over 80% of primary tumors displayed patterns of linear evolution, consistent with findings in whole-genomes (Supp. Figure 4.4C; Fisher's; 82.4% vs. 67.4%; 95% CI = 1.6 - 3.18, OR = 2.25; p = 1.58 x 10^{-6})[11].

We next aimed to identify clonal and subclonal significantly mutated genes (SMGs) in primary and metastatic PC tumors (q < 0.1). Six genes were identified exclusively as clonal SMGs in both primary and metastatic tumors (e.g. *APC*, *CDK12*, *ERF*, *FOXA1*, *SPOP*, *ZFHX3*), while 13 and 30 genes were identified exclusively as clonal SMGs in primary (e.g. *ATM*, *IDH1*, *SMARCA1*) or metastatic tumors (e.g. *AR*, *BRCA2*, *CUL3*), respectively. There was no overlap between subclonal primary and subclonal metastatic SMGs, however, the primary subclonal SMGs *PIK3CA*, *PTEN*, and *KDM6A* were identified as clonal SMGs in metastatic PCs (Supp. Figure 4.4D), supporting the hypothesis that alterations in the PI3K/AKT/mTOR pathway may increase the metastatic potential of primary PC tumors[24–26]. Indeed, the PI3K-AKT-mTOR signaling pathway was only identified via pathway overrepresentation analysis for primary subclonal SMGs and metastatic clonal SMGs (Methods; q < 0.05). Additional pathways overrepresented by clonal SMGs in metastatic samples included *AR* signaling, canonical *WNT* signaling, and the *RAC1-PAK1-p38-MMP2* pathway (q < 0.05). *ANKT1 and NCOR1* were subclonal SMGs exclusively in primary tumors, and *AFF1* was a subclonal SMG exclusive to metastatic samples.

Tumor-level Mutational Signatures

Previous studies have shown that mutational signatures[27,28] appear and shift over the evolutionary trajectory of localized PC tumors[11]. Thus, we determined the mutational signatures present in each tumor and their association with clonal architecture. Using SigMa[29], signature 3 was identified in 23.6% of metastatic PCs compared to 2% of primary PCs (Methods; Fisher's; 95% CI = 7.79-32.59; OR = 15.21; $p < 2.2 \times 10^{-16}$, Figure 4.4A), and MSI associated signatures (signatures 6, 15, 20, and 26) were identified in 6.7% of metastatic PCs compared to 1.4% of primary PCs (Fisher's; 95% CI = 2.01-13.10; OR = 4.89; $p = 1 \times 10^{-4}$), consistent with prior reports[303132]. Conversely, clock-like signatures (signatures 1 and 5) were enriched as the dominant mutational process in primary PCs (Fisher's; 96.6% vs. 70%; 95% CI = 7.14-21.84; OR = 12.21; $p < 2.2 \times 10^{-16}$). We orthogonally validated these classifications using scarHRD[33] to identify HRD-associated copy number events, and MSIsensor[34] to detect somatic microsatellite changes (Supp. Figure 4.5A-C).

Cell Subpopulation-Level Mutational Signatures

We next performed mutational signature analysis at the level of tumor cell subpopulations after accounting for clonal architecture[4]. A total of 1439 cell subpopulations across 829 tumors were powered for the analysis (Methods). Only 51% (40/78) of signature 3 tumors exhibited evidence of signature 3 at the cell subpopulation level, whereas 96% (26/27) of MSI tumors showed evidence of MSI at the cell subpopulation level (Figure 4.4B). The discrepancy observed in the signature 3 tumors may be due to the reduction in power to call signature 3 when separating mutations into their respective cell subpopulations. To determine the robustness of these signature calls and confirm their clonality, we validated the results using a second phylogenetic reconstruction method (Methods).

Of the 40 tumors with evidence of signature 3 at both the overall tumor and cell subpopulation level, signature 3 was a clonal mutational process in 3 of 5 primary tumors (60%),

and 31 of 35 metastatic tumors (89%). Of the 26 tumors with tumor and cell subpopulation level evidence of MSI, MSI was identified clonally in 6 of 8 primary tumors (75%), and 9 of 18 metastatic tumors (50%). Interestingly, 12 tumors categorized by clock-like mutational signatures at the tumor level showed evidence of signature 3 at the cell subpopulation level (Figure 4B), with the majority of these being the clonal cell subpopulation (83%; 1/2 primary, 9/10 metastatic). While tumors identified with signature 3 exclusively at the cell subpopulation level had higher numbers of HRD-associated CNA events than tumors identified with signature 3 exclusively at the whole tumor level, these differences were non-significant (Supp. Figure 4.6). MSI was also detected in cell subpopulations of tumors categorized by clock-like mutational processes (n = 31, Figure 4.4B). In each of these tumors, MSI was the result of a subclonal mutational process (1 primary, 30 metastatic). This high rate of subclonal MSI is consistent with results from the PCAWG cohort showing mismatch repair (MMR)-associated signatures preferentially result in subclonal mutations[35]. These results suggest that signature 3 and MSI classifications may be masked by more active or dominant mutational signatures, such as clock-like signatures in PC, when performing mutational signature analysis at the whole tumor level.

PC tumors with signature 3 present in the clonal (truncal) subpopulation had significantly elevated numbers of LOH (Kolmogorov-Smirnov, $p = 1.12 \times 10^{-10}$; univariate logistic regression, $p < 2.2 \times 10^{-16}$), TAI (Mann-Whitney U, 23 vs. 4 events, $p < 2.2 \times 10^{-16}$), LST (Mann-Whitney U, 23.5 vs. 8 events, $p < 2.2 \times 10^{-16}$), and total number of HRD-associated CNA (Mann-Whitney U, 61 vs. 18 events, $p < 2.2 \times 10^{-16}$) events compared to PC tumors with no signature 3 (Supp. Figure 4.7A). However, PC tumors with clonal evidence of signature 3 had significantly elevated numbers of TAI (Mann-Whitney U, 23 vs.10 events, $p = 8.4 \times 10^{-4}$) events, but not LOH, LST, and total number of HRD-associated CNA events compared to PC tumors with subclonal signature 3 (Supp. Figure 4.7A). Tumors with subclonal signature 3 had significantly more LST (Mann-Whitney U, 17 vs. 8 events, $p = 5.6 \times 10^{-3}$) and total HRD-associated (Mann-Whitney U,

39 vs. 18 events, p = 0.01) CNA events compared to tumors with no signature 3, as well as borderline significant enrichment of LOH (Kolmogorov-Smirnov, p = 0.20; univariate logistic regression, $p = 2.3 \times 10^{-9}$) and TAI (Mann-Whitney U, 10 vs. 4 events, p = 0.058) events (Supp. Figure 4.7A).

Metastatic tumors with clonal evidence of MSI had significantly higher mutational burden and MSIsensor scores than metastatic tumors with subclonal MSI (TMB: Mann-Whitney U, 59.35 mut/Mb vs. 4.95 mut/Mb, $p = 1.19 \times 10^{-9}$; MSIsensor: Mann-Whitney U, 14.93 vs. 0.12, $p = 3.63 \times 10^{-6}$) or no MSI (TMB: Mann-Whitney U, 59.35 mut/Mb vs. 2.45 mut/Mb, $p = 4.01 \times 10^{-7}$; MSIsensor: Mann-Whitney U, 14.93 vs. 0.12, $p = 2.83 \times 10^{-7}$; Supp. Figure 4.7B-C). Interestingly, metastatic tumors with evidence of subclonal MSI had elevated mutational burden compared to metastatic tumors with no MSI signatures (Mann-Whitney U, 4.95 mut/Mb vs. 2.45 mut/Mb, $p = 1.17 \times 10^{-9}$, Supp. Figure 4.7C), but not higher MSIsensor scores (Mann-Whitney U, 0.12 vs. 0.12, p = 0.84; Supp. Figure 4.7B). These results suggest that the endogenous mutational process driving signature 3 may preferentially occur clonally, while the endogenous mutational process driving MSI may preferentially occur subclonally, and that cell subpopulation specific analyses can uncover additional samples with DNA repair defect-associated mutational signatures[36].

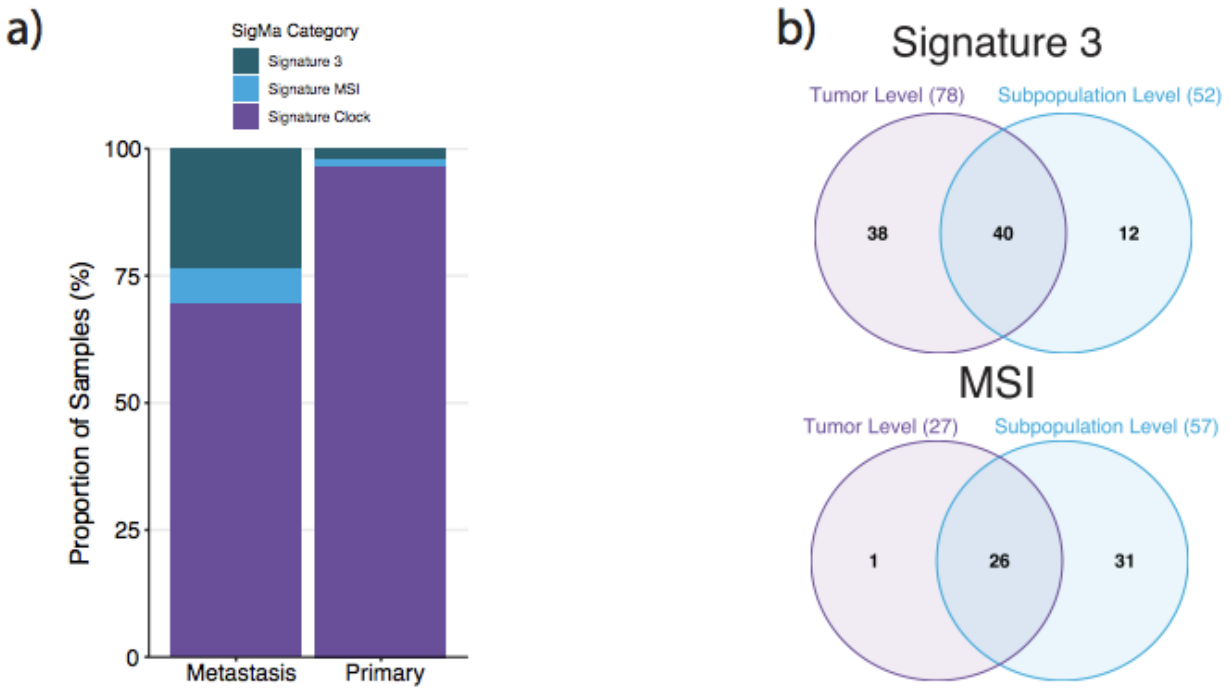


Figure 4.4: Mutational signature analysis at the cell subpopulation level can identify additional PC patients with HRD or MSI
(a) The proportion of metastatic and primary samples that have evidence of mutational signature 3 (associated with HRD) and MSI-associated mutational signatures. Activity of HRD and MSI-associated mutational signatures were more frequent in metastatic samples compared to primary samples. **(b)** The overlap between samples with evidence of mutational signature 3 and MSI-associated mutational signatures at the tumor and cell subpopulation levels. Out of 78 samples identified with signature 3 from running SigMA on all mutations in the tumor, 40 (51%) of those samples were identified as having signature 3 when running SigMA on the cell subpopulations. Conversely, 12 samples that were not classified as having signature 3 at the tumor level were identified as having signature 3 when running SigMA on the cell subpopulations. Similarly, while all but 1 sample identified with the MSI-associated signature at the tumor level were identified as having the MSI-associated signature at the cell subpopulation level, and 31 additional samples were identified as having MSI at the cell subpopulation level.

Mutational mechanisms of clonal and subclonal signature 3 and MSI in PC

We next aimed to leverage cell subpopulation level mutational signatures to identify genomic alterations associated with signature 3 and MSI (Figure 4.5A). Thus, we performed germline mutation calling on a set of double-stranded break repair and MMR genes (Methods), and hypothesized that any germline alteration solely casual of signature 3 or MSI would manifest clonally. Of the 16 samples (all metastatic) with pathogenic germline *BRCA2*

alterations, 10 (62.5%) showed clonal activity of signature 3 (Figure 4.5A-B, Supp. Figure 4.8A-B). Three of those 10 samples also had clonal somatic *BRCA2* mutations (1 missense, 2 frameshift deletions), and therefore signature 3 in these samples may be the result of biallelic loss. Conversely, none of the tumors with germline *ATM* alterations (0 of 10), and only 1 tumor with a *BRCA1* germline alteration (14%, 1/7) had clonal activity of signature 3. One of 2 samples with *PALB2* germline alterations, in this case a primary sample, also had clonal activity of signature 3. Additionally, 1 of 2 samples with *MSH6* germline alterations had clonal activity of MSI (Supp. Figure 4.8C-D), although this sample also had a clonal somatic *MSH6* mutation and therefore may be the result of a biallelic loss (Figure 4.5B).

We next identified cases where signature 3 and MSI may have been caused by somatic alterations. Five of 15 tumors with only somatic putative loss of function (LOF) *BRCA2* mutations (i.e. no germline alterations) had signature 3. Three of these tumors had clonal *BRCA2* mutations, and 2 tumors had subclonal *BRCA2* mutations, with the onset of signature 3 being observed within the same cell subpopulation as the mutations (Figure 4.5B). Like the germline results, none of the tumors with putative LOF somatic *ATM* mutations had signature 3. One tumor with a clonal *BRCA1* mutation had clonal activity of signature 3, and 1 of 2 tumors with putative LOF somatic *PALB2* mutations (Figure 4.5B), each of which were clonal, had clonal evidence of signature 3.

Other than the MSI tumor with the *MSH6* biallelic loss, one other MSI tumor harbored a putative LOF *MSH6* somatic mutation (Figure 4.5B). The tumor with the *MSH6* double hit also had putative LOF somatic mutations in *MSH3* and *PMS1*, as well as missense mutations in 15 other MMR associated genes. Interestingly, this tumor had an intermediate MSIsensor score (1.81), despite having the highest mutational burden of any tumor in the cohort. All 3 samples with putative LOF somatic mutations in *MSH2* (2 clonal, 1 subclonal) experienced the onset of MSI in the corresponding cell subpopulation (Figure 4.5A-B).

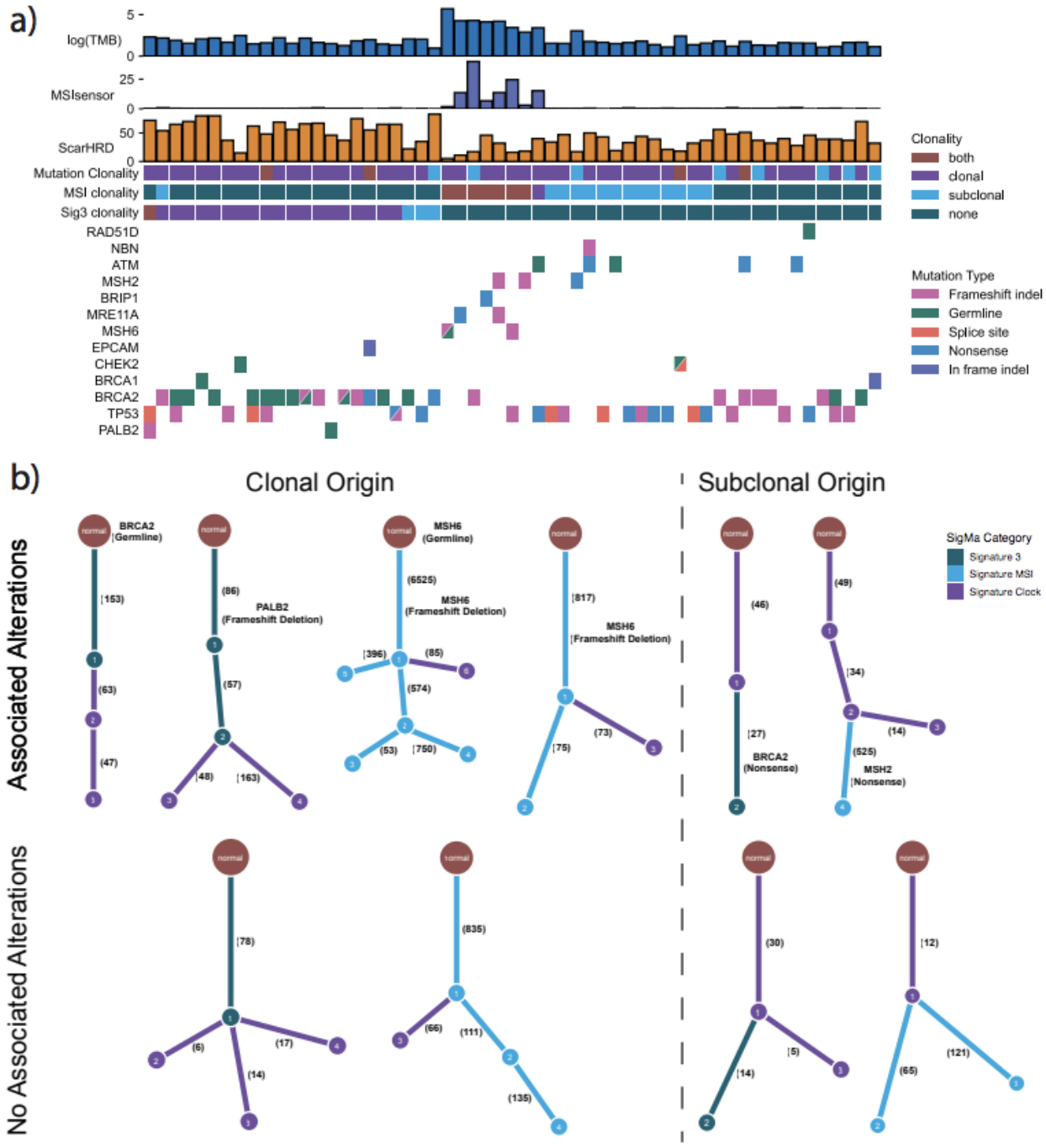


Figure 4.5: Linking the origin of mutational signatures to germline and somatic alterations and tracking how mutational signatures influence tumor evolution
(a) Co-mutation plot of putative LOF somatic and germline alterations in a curated set of DSB and MMR-associated genes. The co-mutation plot is annotated with the clonality of the alterations, clonality of signature 3 (associated with HRD), and clonality of MSI associated signatures. The co-mutation plot is also annotated with the number of HRD-associated CNA events (ScarHRD) per sample, as well as MSIsensor scores and mutational burden which are associated with MSI. **(b)** Phylogenetic trees from a subset of tumors included in the co-mutation plot showing how our novel approach can **(top)** link the origin of mutational signatures to

Figure 4.5 (continued): somatic or germline alterations at the cell subpopulation level, and how (top and bottom) the mutational signatures influence the subclonal diversification over the tumors evolution. The subclone numbers indicate the inferred order in which those subclones occured. For example, subclone 2 is present at a higher CCF than subclone 3.

Therapeutic implications of cell subpopulation mutational signature analysis

To determine the therapeutic implications of integrating cell subpopulation level derived mutational signatures with tumor level derived mutational signatures, we leveraged progression free survival (PFS) data from patients in our cohort that were treated with PARP inhibitors (PARPi) (n=37; Methods)[37]. Although nonsignificant, univariate Cox PH analysis revealed that patients with tumor level (HR = 0.60, p = 0.33; Supp. Figure 4.9A) or cell subpopulation level (HR = 0.45, p = 0.11; Supp. Figure 4.9B) signature 3 had a lower hazard for progression when treated with PARPi, and this effect held when integrating the both levels of signature calls (HR = 0.59, p = 0.20; Supp. Figure 4.9C). When correcting for prior treatment with radiation or systemic therapies, both cell subpopulation level signature 3 calls and the combination of tumor level and cell subpopulation level signature 3 calls were significantly associated with longer PFS (cell subpopulation level: HR = 0.34, p = 0.036; combined: HR = 0.39, p = 0.037; Supp. Figure 4.9D-F; Supp. Table 4.4), while tumor level signature 3 was borderline associated with improved PFS (HR = 0.33, p = 0.09). Thus, these results suggest that integrating cell subpopulation level mutational signatures with traditional approaches of identifying HRD patients can identify more patients that may benefit from PARPi.

To further validate the identification of cell subpopulation signature analysis and its therapeutic implications outside of the PC context, we applied this novel integrative approach to a cohort of ovarian cancer tumors treated with cisplatin chemotherapy[38,39]. We were able to recapitulate the association between signature 3 clonality and HRD-associated CNA events observed in PC (Supp. Figure 4.10A), and our approach identified additional patient tumors with signature 3 present compared to current approaches[40,41]. Furthermore, patients with signature 3

identified via our novel integrative approach had improved PFS on cisplatin therapy (Supp. Figure 4.10B-E), further highlighting the potential expanded therapeutic impact of a clonal architecture-based approach to signature analysis for clinical use.

Discussion

In this study, we revealed novel associations of genomic and clinical characteristics with clonal architecture and evolutionary dynamics in PC tumors through CCF inference and phylogenetic reconstruction of 845 single biopsy samples. Clinical covariates that were associated with the clonal architecture of PC tumors included clinical risk groups, self-reported race, ancestry, and primary versus metastatic status. While the association between clinical risk groups and clonal architecture has already been established, we demonstrated that tumors in patients with germline AFR ancestry or Black race, which have substantial disparities in clinical outcome, paradoxically exhibit clonal architecture and evolutionary features previously reported in lower risk tumors from White patients[11].

Our observations that primary tumors in Black and/or AFR patients are more monoclonal and have linear genomic architecture, but have higher rates of biochemical recurrence, could indicated that these tumors were screened for and detected at a time point after they had underwent rapid selective sweeps or punctuated evolution. Rapid selective sweeps in tumor evolution occur when all cell subpopulations collapse into one dominant clone[5,42,43]. The combination of mutations within this dominant clone may cause or collaborate with tumor transcriptomic, proteomic, and microenvironmental phenotypes that together are associated with higher rates of biochemical recurrence. Through admixture analysis, we provide evidence that the enrichment of linear and monoclonal phylogenies in Black and/or AFR patient tumors may partially be explained by inherited germline SNPs associated with increased risk of PC. The lower abundance of driver intratumoral heterogeneity in AFR and Black patient samples coupled with evidence of rapid selective sweeps and higher clonality genomic alteration events are akin

to observations in clear cell renal cell carcinoma[12], where these features were associated with worse outcomes compared to tumors with increased clonal diversity (i.e. more subclones) and lower clonality of alterations.

Together with prior publications suggesting that Black patients who undergo prostatectomy for PC have higher tumor volume relative to White patients[44], our data also reinforce the critical importance of more nuanced screening approaches and better healthcare resources and access for this patient population. In the context of data illustrating equivalent if not better outcomes for Black patients with advanced PC when treated appropriately with standard of care therapies in relatively equal-access settings[45,46], it is possible that earlier detection of PC in Black patients and its prompt treatment may improve patient outcomes.

Our observations on self-reported race, ancestry, genomic architecture, and outcome in localized PC have several caveats. Within our large cohort, Black and AFR patients are under-represented, limiting the power to detect accurate genomic differences. Our data was aggregated from patients who donated tissue samples primarily at certain academic medical centers through efforts such as The Cancer Genome Atlas, which likely do not fully represent the spectrum of Black and AFR PC. Furthermore, our dataset lacks information on key associations with PC outcome such as health insurance access, zip code, comorbidities such as obesity, tobacco exposure, and others. Recruitment of geographically and racially diverse patient populations and inclusion of information on known risk factors for prostate cancer incidence and mortality are vital for subsequent validation of our findings and other efforts to elucidate the association between tumor biology and outcome. Even more importantly, equal access to resources that may minimize these other risk factors for adverse prostate cancer outcomes are essential.

Using a novel approach that integrates phylogenetic reconstruction[4] with mutational signature analysis[29], we demonstrate that performing mutational signature analysis at the cell subpopulation level can uncover both clonal and subclonal drivers as well as additional cases of

HRD and MSI, which may have therapeutic implications. Mutational significance analysis of clonal and subclonal mutations in primary and metastatic tumors revealed shared clonal drivers (e.g. *ERF*, *FOXA1* and *SPOP*) that may be important for tumor initiation and progression, as well as subclonal drivers in primary tumors that may confer metastatic potential and a selective advantage in metastatic sites (e.g. *KDM6A*, *PTEN* and *PIK3CA*)[24]. We also find that associated genomic features, such as HRD-associated CNA events[33] and mutational burden[47], show activity levels consistent with the clonality of the corresponding mutational signature. Specifically, there is a stepwise increase in the number of HRD-associated CNA events from tumors with no signature 3 to subclonal signature 3 to clonal signature 3, and a stepwise increase in mutational burden from tumors with no MSI signature to subclonal MSI signature to clonal MSI signature. Immunotherapy selection for PC patients based on MSIsensor scores[48] may miss subclonal MSI cases that can be captured through this analytical framework. For a subset of tumors, we further demonstrate the ability to link putative causal germline and somatic alterations to the origin of HRD or MSI[30,31,49,50] in the corresponding cell subpopulation. Additionally, the application of our novel integrative approach to PC tumors treated with PARPi and ovarian cancers treated with cisplatin chemotherapy identified more patients that appeared to benefit from these therapies, demonstrating the potential expanded therapeutic impact of our approach.

While the size of our cohort offers the statistical power necessary to find associations with clonal architecture and evolutionary trajectories, this study is limited by only having a single biopsy with exome sequencing per sample. Single biopsy samples are more prone to sampling bias compared to multifocal biopsies, whereby mutation CCFs may deviate from their true value based on biopsy location (i.e. subclonal mutations presenting more clonally or subclonally)[2,13,51]. To address this issue and increase the confidence of the timing (e.g. CCF or clonality status) of these events in PC, we leveraged the timing of each event across all samples and performed all analyses at the cohort level. Further, while PC whole-genome sequencing datasets do exist,

they lack the breadth of clinical characteristics to fully interrogate ancestry, HRD and MSI but may augment these investigations prospectively. Future studies involving spatio-genomic profiling[12,13], single-cell sequencing[52–54], or long read sequencing[55] may further validate and inform the biology underlying these findings. Nevertheless, clonal architecture and evolutionary informed analysis of increasingly large cohorts of tumors will continue to reveal novel biological insights with immediate clinical potential, especially as clinical sequencing programs integrate such analyses into their workflows.

Methods

Cohort collection, quality control, and somatic variant calling

The somatic variants utilized in this study were taken from supplementary table 2 of Armenia *et. al*[14]. The cohort collection, quality control metrics, and somatic variant calling information can be found in the Methods of that study. However, in Armenia *et al*[14], allelic copy number, purity, and ploidy information were determined using ABSOLUTE[56] in tumors where the purity was too low for FACETS[57] to output a solution. To remain consistent in this study, we re-ran FACETS and only kept tumors where FACETS produced a solution. Out of the original 1,013 PC tumors in the cohort, 845 produced a FACETS output.

Clinical data

All clinical data was downloaded from the original published studies[37,58–64] Clinical data for the TCGA samples were taken from the original study[59], and updated using the clinical data from the MC3 study[39] where applicable. Clinical risk groups were defined using the NCCN guidelines. Specifically, low risk: tumor stages T1-T2, grade group 1, and PSA < 10 ng/mL; intermediate risk: tumor stages T2b-T2c, grade group 2 or 3, or PSA 10-20 ng/mL, and no high risk features; high risk: tumor stage T3a or higher, grade group 4 or higher, or PSA > 20 ng/mL.

Allelic copy number calling

Allelic CNAs were determined using FACETS[57], which provides major and minor allele integer copy number values, tumor purity, tumor ploidy, and the cellular fraction of each copy number segment. The CCF of each CNA was calculated by adjusting the cellular fraction of the CNA by tumor purity.

Calculation of mutation CCFs

Somatic mutation CCFs were calculated using the maximum likelihood method described in McGranahan *et al*.[65] using allelic copy number and tumor purity information from FACETS[57]. Here the maximum likelihood estimation of mutation CCFs are determined using a binomial distribution, taking into account tumor copy number, tumor purity, and variant allele frequency. This process is performed for the scenarios where the mutation occurs on either 1) the major allele, 2) the minor allele, or 3) a single allele copy, and the most likely CCF is chosen. For mutations that fall on normal ploidy segments, there is no difference in the CCF calculations for the mutation occurring on the major allele, minor allele, or a single copy of an allele. For mutations occurring on segments where both the major and minor allele copy numbers do not equal 1, the mutations with the highest likelihood of occurring on a single copy of an allele (rather than the major or minor allele) indicates that the mutation likely occurred after the CNA.

Calculation of copy number alteration CCFs

The cellular frequency estimation column (“cf.em”) output by FACETS[57] is the fraction of all cells with a particular allelic copy number, including normal cells. To get the CCF of any non-diploid copy number segment the “cf.em” column was divided by the FACETS derived tumor purity.

Phylogenetic reconstruction of tumor architecture

To reconstruct the clonal architecture of prostate cancer tumors we used the PhylogicNDT[4] Cluster module, which determines the number of tumor cell subpopulations and the respective assignment of each mutation to a cell subpopulation. The CCF annotated MAF file and FACETS-derived tumor purity were used as inputs to the clustering method. The outputs from the PhylogicNDT Cluster module were then used as inputs to the PhylogicNDT BuildTree module, which produces a series of phylogenetic trees ordered by likelihood. The phylogenetic trees with the highest likelihood were used in the analyses of this study, and were used to determine whether a tumor exhibited linear or branched evolution. Linear evolution is defined as phylogenetic reconstruction where each cell subpopulation in the tumor has a maximum of 1 child node. The number of clones per tumor were defined as the number of clusters identified by PhylogicNDT, and monoclonal tumors were defined as tumors where PhylogicNDT identified only a single cluster.

Although PyClone was not designed for use in whole-exome data[2] and tends to over cluster whole-exome mutation data[66], it is one of the most highly used phylogenetic reconstruction methods. To validate the subclonal and clonal origin of mutational signatures (see Mutational signature analysis) we also ran PyClone with the following hyperparameters: burnin: 1000, density: pyclone_beta_binomial, init_method: connected, mesh_size: 101, num_iters: 10000, prior: major_copy_number, thin: 1. In cases where both methods were powered enough to call signatures at the cell subpopulation level, Pyclone identified MSI in all the same samples as PhylogicNDT (n=46; 100%), whereas Pyclone only identified signature 3 in 41 of 44 (93.2%) samples determined to have signature 3 via PhylogicNDT. Pyclone did not result in any additional MSI or signature 3 cases that weren't detected by PhylogicNDT. Further, when both methods identified MSI and signature 3 in the same samples, MSI was identified at

the same clonality (e.g. clonally or subclonally) in 43 of 46 (93.5%) samples, and signature 3 was identified at the same clonality in 40 of 41 (97.6%) samples.

Mutational significance analysis

To perform mutation significance analysis we used MutSigCV2[67], and classified SMGs as genes with an FDR corrected p-value < 0.1.

Mutational signature analysis

Active mutational processes were determined using both the deconstructSigs[40] and SigMa[29] R packages. To run deconstructSigs we used the recommended, default parameters with the COSMIC (v2) signatures as the signatures reference. To run SigMa we set the data parameter equal to “seqcap” for whole-exome sequencing, the tumor_type parameter equal to “prost” for prostate adenocarcinoma, and check_msi parameter equal to TRUE to identify tumors with MMRd associated signatures. Default values were used for all other parameters. To run SigMa at the cell subpopulation level, we ran SigMa on the clusters of mutations output by PhylogicNDT[4]. In certain cases where there were too few mutations in a cluster (< 10), SigMa failed to produce an output for the cluster. Additionally, since running SigMa at the cell subpopulation level reduces the number of mutations input into SigMa, it may reduce the power to detect signature 3 and MSI in certain cases. Conversely, running at the cell subpopulation level enables the identification of signature 3 and MSI that may be confounded or overpowered by other mutational signatures at the tumor level.

While non-negative least squares (NNLS) methods (e.g. deconstructSigs)[40] are popular for identifying mutational signatures in individual samples, these methods are susceptible to increased rates of false positives in tumors with low mutational burden[29], which is the case with many PC tumors. For instance, the observed prevalence of signature 3, associated with HRD, in primary PC whole-genomes was 5.8%, whereas deconstructSigs called signature 3 in 14.5% of

primary PC tumors. A worse discrepancy was observed for mismatch repair deficient (MMRd) associated signatures. For this reason we reported mutational signature analysis using SigMa[29], which utilizes non-negative matrix factorization (NMF), NNLS, and likelihood-based statistics combined with machine learning to classify known signatures across cancer types.

Classification of clonal vs. subclonal mutations and mutational signatures

Mutations and mutational signatures were classified as clonal if they were identified in the truncal cluster of the tumor. Conversely, mutations and mutational signatures were classified as subclonal if they were not identified in the truncal cluster of the tumor. While mutations identified in the truncal cluster of the tumor are presumed to be present in every other cluster throughout the tumor, it is much more difficult to determine whether this is the case for mutational signatures. For instance, the identification of a mutational signature present in only the truncal cluster may still be active subclonally, but not identified due to power issues, the presence of a more active mutational signature, or that mutations generated by another signature were selected for. Conversely, mutations, CNAs or epigenetic alterations may cause a reversion of the mutational signature, specifically those caused by endogenous mutational processes such as HRD and MMRd.

Calculation of HRD-associated CNA events (scores)

To calculate the number of LOH, TAI, and LST events in each tumor, we used the FACETS[57] allelic copy number calls as input to the scarHRD R package[33]. To determine the enrichment of these events in tumors classified with signature 3, we used the same statistical tests as the original papers the associations were discovered in. That is, Kolmogorov-Smirnov and univariate logistic regression for LOH events[68], and Mann-Whitney U for both TAI and LST events[69,70]. Mann-Whitney U was also used to determine if there was a significant enrichment in

the unweighted sum of events, and denote the significance values of all four scores (LOH, TAI, LST, unweighted sum) in the figures.

Identification of mutations at microsatelites

To identify replication slippage variants at microsatellite regions and quantify the proportion that are somatic (also called MSIsensor score) we used MSIsensor[34]. PC tumors were characterized as having high, intermediate, and low MSIsensor scores using the same criteria as Abida *et. al*[48].

HRD and MMRd gene sets

The list of HRD and MMRd genes were curated from three prostate cancer specific studies: 1) Matteo *et al*. 2015[30], 2) Pritchard *et al*. 2016[31], and 3) de Bono *et al*. 2020[49], as well as Polak *et al*. 2017[50].

Pathway overrepresentation analysis

We performed pathway overrepresentation analysis on genes identified as SMGs, and significantly amplified or deleted, using ConsensusPathDB (CPDB)[71,72]. We ran CPDB on December 3rd, 2019 with default parameters for pathway-based sets.

Germline variant discovery

DeepVariant (v0.8.0)[73] was used to call SNVs and small deletions/duplications (indels) from whole-exome sequencing matched normal samples. Only high quality variants that were classified as “PASS” in the “FILTER” column were kept, and the CombineVariants module from GATK 3.7[74] was used to merge all of the high quality variants into a single Variant Call Format (VCF) file. The vt (v3.13)[75] tool was then used to decompose multiallelic variants, followed by

normalization of variants. The high quality germline variants were annotated using VEP (v2)[76] with the publicly available GRCh37 cache file.

Germline variant pathogenicity evaluation

High quality germline variants were evaluated for pathogenicity using publicly-available databases such as ClinVar and gene-specific databases, and classified according to the American College of Medical Genetics and Genomics and the Association of Molecular Pathology clinical-oriented guidelines[77]. Based on the evidence extracted from these resources, germline variants were classified into 5 categories: benign, likely benign, variants of unknown significance, likely pathogenic and pathogenic[77]. Additionally, truncating germline variants in genes that have yet been associated with a clinical phenotype, but are expected to disrupt the protein function, were classified as likely disruptive. Only germline variants classified as pathogenic, likely pathogenic, or likely disruptive were considered in the analysis.

Ancestry inference

Hail (v0.2.39-ef87446bd1c7) was used to perform ancestry inference for each sample in our cohort. The “variant_qc” method was used on the combined cohort germline VCF to compute common variant statistics. This was followed by filtering out rare variants with an allele frequency less than 0.01, and variants that had a Hardy-Weinberg equilibrium p-value greater than 1×10^{-6}. Additionally, we used the “ld_prune” method to filter out variants with a Spearman correlation threshold less than 0.1. The “hwe_normalized_pca” method was used to obtain the principal component analysis (PCA) eigenvalues and scores. To infer the ancestry of our samples, we also performed PCA on 1000 Genomes reference samples[78,79], and trained a random forest classifier on the first 10 principal components to assign one of the five 1000 Genomes super populations (European, African, Admixed American, East Asian, and South Asian) to each of our samples.

Population admixture estimation

Bcftools *mpileup* was used to generate a VCF of genotype likelihoods with the following options: 1) specifying not to skip anomalous read pairs (-A), 2) recalculating base alignment quality on the fly (-E), and 3) skipping indels (-I). We then called genotypes at local population ancestry reference sites from the 1000 Genomes Project (as opposed to super populations, see "Ancestry inference"; https://www.internationalgenome.org/category/population/) using Bcftools *call* with the option to call genotypes given alleles (-C), followed by removing duplicate genotype calls using Bcftools *norm*. The resulting VCF was used to as input to PLINK (v1.9)[80] *--make-bed*, which was subsequently used as input to fastNGSadmix[81] to determine the admixture proportions for the 1) Utah Residents (CEPH) with Northern and Western European Ancestry (CEU), 2) Han Chinese in Beijing, China (CHB), 3) Yoruba in Ibadan, Nigeria (YRI), and 4) Peruvians from Lima, Peru (PEL) populations.

Imputation of rs1447295

To impute the genotypes of samples at rs1447295, we first calculated off-target coverage of our samples using bedtools genomecov[82] followed by GATK depth of coverage[74]. Only samples with > 0.1X off-target coverage were considered for subsequent analysis. We then calculated the relatedness between samples using somalier (https://github.com/brentp/somalier)[83], and removed samples that were related by the 2nd degree or closer (kinship value > 0.125). We then split the samples by their super population ancestry assignments (European or African), and required that reported race matched the ancestry assignments. The GLIMPSE pipeline[84] was then run on both the European (n=52) and African (n=33) ancestry cohorts independently to infer the genotypes at rs1447295. The info score for rs1447295 was 0.90 in the European ancestry cohort, and 0.96 in the African ancestry cohort.

Survival analysis

To demonstrate the improved clinical utility of using SigMa[29], and determine the therapeutic implication of cell subpopulation level mutational signatures, we performed survival analysis on a subset of patients in our cohort that received PARPi (Olaparib or Veliparib)[37]. In the multivariate model we corrected for whether or not the patient also received prior radiation, hormone therapy, chemotherapy, and immune-related therapy (i.e. yes or no). To validate the presence of signature 3 at both the tumor and cell subpopulation levels, and determine if the associations apply in other cancer contexts, we leveraged data from the TCGA ovarian cancer cohort[38,39], and restricted our analysis to ovarian tumors treated with platinum-based chemotherapy. The drug information, number of platinum-based chemotherapy cycles, patient age, and tumor stage information were downloaded from FireCloud (https://portal.firecloud.org/#workspaces/broad-firecloud-tcga/TCGA_OV_ControlledAccess_V1-0_DATA). The ovarian cancer mutation calls and survival data were downloaded from the MC3 study[39]. To determine if there are significant differences between the survival curves of two or more groups we used the log-rank test from the survival R package. To evaluate whether any covariates were confounding the associations identified in the Kaplan-Meier analyses, we also performed Cox proportional hazards analysis (using the survival R package) correcting for these covariates.

Clonal architecture and evolutionary dynamics in WGS cohort

Somatic mutation VCFs for African ancestry (n=7) and European (n=7) ancestry WGS samples from Petrovics *et al*[23] were obtained directly from the authors, and we restricted the somatic mutation call set to those classified as high-confidence in the original study. Mutation CCFs were calculated as described in “Calculation of mutation CCFs" using FACETS-derived allelic copy number calls and the histologically defined tumor purities from the original study. The

clonal architecture and evolutionary trajectories of these tumors were determined using PhylogicNDT[4] as described in “Phylogenetic reconstruction of tumor architecture”.

Declarations

Data Availability

All of the datasets used in this study are publicly available, and can be downloaded from Armenia *et al*., 2018[14]. Publicly available databases used in this study include MSigDB v6.2 (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), ExAC (http://exac.broadinstitute.org/), gnomAD (https://gnomad.broadinstitute.org/), and ConsensusPathDB v34 (http://cpdb.molgen.mpg.de/).

Code Availability

All software and bioinformatic tools used in this study are publicly available.

Acknowledgements

This work was supported by NCI F31CA239347 (J.R.C.), NIH 5T32HG002295-15 (J.R.C.), NIH R01CA227388-02 (E.M.V.A.), NIH R21CA242861 (E.M.V.A.) and the Damon Runyon Clinical Investigator Award (E.M.V.A.). The results presented in this study are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author Contributions

J.R.C. and A.K.T. contributed to the analysis of genomic and clinical data. J.R.C., A.K.T., S.Y.C., K.S., S.H., J.C., M.X.H., S.H.A., Z.S., M.M.P., M.L.F., P.S.N., M.B., and E.M.V.A. contributed to interpretation of results, and manuscript preparation.

Competing Interests

E.M.V.A. is a consultant for Tango Therapeutics, Genome Medical, Invitae, Enara Bio, Monte Rosa Therapeutics, Manifold Bio, and Janssen. E.M.V.A. provides research support to Novartis and Bristol-Myers Squibb. E.M.V.A. has equity in Tango Therapeutics, Genome Medical, Syapse, Ervaxx, and Microsoft. E.M.V.A. receives travel reimbursement from Roche/Genentech. E.M.V.A. has institutional patents filed on methods for clinical interpretation.

Bibliography

1. Liu, J., Dang, H. & Wang, X. W. The significance of intertumor and intratumor heterogeneity in liver cancer. *Exp. Mol. Med.* **50**, e416 (2018).

2. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).

3. Andor, N., Harness, J. V., Müller, S., Mewes, H. W. & Petritsch, C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30**, 50–60 (2014).

4. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

5. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).

6. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 7–30 (2020).

7. Marcus, L., Lemery, S. J., Keegan, P. & Pazdur, R. FDA Approval Summary: Pembrolizumab for the Treatment of Microsatellite Instability-High Solid Tumors. *Clin. Cancer Res.* **25**, 3753–3758 (2019).

8. Hussain, M. *et al.* Survival with Olaparib in Metastatic Castration-Resistant Prostate Cancer. *N. Engl. J. Med.* (2020) doi:10.1056/NEJMoa2022485.

9. Abida, W. *et al.* Rucaparib in Men With Metastatic Castration-Resistant Prostate Cancer Harboring a BRCA1 or BRCA2 Gene Alteration. *Journal of Clinical Oncology* vol. 38 3763–3772 (2020).

10. Mateo, J. *et al.* Olaparib in patients with metastatic castration-resistant prostate cancer with DNA repair gene aberrations (TOPARP-B): a multicentre, open-label, randomised, phase 2 trial. *Lancet Oncol.* **21**, 162–174 (2020).

11. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* **173**, 1003–1013.e15 (2018).

12. Turajlic, S. *et al.* Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* **173**, 595–610.e11 (2018).

13. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).

14. Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).

15. Gerhauser, C. *et al.* Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell* **34**, 996–1011.e8 (2018).

16. Koga, Y. *et al.* Genomic Profiling of Prostate Cancers from Men with African and European Ancestry. *Clin. Cancer Res.* **26**, 4651–4660 (2020).

17. Huang, F. W. *et al.* Exome Sequencing of African-American Prostate Cancer Reveals Loss-of-Function ERF Mutations. *Cancer Discovery* vol. 7 973–983 (2017).

18. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14068–14073 (2006).

19. Severi, G. *et al.* The common variant rs1447295 on chromosome 8q24 and prostate cancer risk: results from an Australian population-based case-control study. *Cancer Epidemiol. Biomarkers Prev.* **16**, 610–612 (2007).

20. Chen, M. *et al.* The rs1447295 at 8q24 is a risk variant for prostate cancer in Taiwanese men. *Urology* **74**, 698–701 (2009).

21. Robbins, C. *et al.* Confirmation study of prostate cancer risk variants at 8q24 in African Americans identifies a novel risk locus. *Genome Res.* **17**, 1717–1722 (2007).

22. Okobia, M. N., Zmuda, J. M., Ferrell, R. E., Patrick, A. L. & Bunker, C. H. Chromosome 8q24 variants are associated with prostate cancer risk in a high risk population of African ancestry. *The Prostate* vol. 71 1054–1063 (2011).

23. Petrovics, G. *et al.* A novel genomic alteration of LSAMP associates with aggressive prostate cancer in African American men. *EBioMedicine* **2**, 1957–1964 (2015).

24. Wedge, D. C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat. Genet.* **50**, 682–692 (2018).

25. Crumbaker, M., Khoja, L. & Joshua, A. M. AR Signaling and the PI3K Pathway in Prostate Cancer. *Cancers* **9**, (2017).

26. Bitting, R. L. & Armstrong, A. J. Targeting the PI3K/Akt/mTOR pathway in castration-resistant prostate cancer. *Endocr. Relat. Cancer* **20**, R83–99 (2013).

27. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

28. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,

415–421 (2013).

29. Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).

30. Mateo, J. *et al.* DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *N. Engl. J. Med.* **373**, 1697–1708 (2015).

31. Pritchard, C. C. *et al.* Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. *N. Engl. J. Med.* **375**, 443–453 (2016).

32. de Bono, J., Kang, J. & Hussain, M. Olaparib for Metastatic Castration-Resistant Prostate Cancer. Reply. *The New England journal of medicine* vol. 383 891 (2020).

33. Sztupinszki, Z. *et al.* Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ Breast Cancer* **4**, 16 (2018).

34. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).

35. Akdemir, K. C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.* **52**, 1178–1188 (2020).

36. Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat. Commun.* **9**, 3292 (2018).

37. Abida, W. *et al.* Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 11428–11436 (2019).

38. Network, T. C. G. A. R. & The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* vol. 474 609–615 (2011).

39. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271–281.e7 (2018).

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

41. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).

42. Gopal, P. *et al.* Clonal selection confers distinct evolutionary trajectories in BRAF-driven cancers. *Nat. Commun.* **10**, 5143 (2019).

43. Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).

44. Powell, I. J., Bock, C. H., Ruterbusch, J. J. & Sakr, W. Evidence supports a faster growth

rate and/or earlier transformation to clinically significant prostate cancer in black than in white American men, and influences racial progression and mortality disparity. *J. Urol.* **183**, 1792–1796 (2010).

45. Dess, R. T. *et al.* Association of Black Race With Prostate Cancer-Specific and Other-Cause Mortality. *JAMA Oncol* **5**, 975–983 (2019).

46. Spratt, D. E. *et al.* Individual Patient Data Analysis of Randomized Clinical Trials: Impact of Black Race on Castration-resistant Prostate Cancer Outcomes. *Eur Urol Focus* **2**, 532–539 (2016).

47. Haraldsdottir, S. *et al.* Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. *Gastroenterology* **147**, 1308–1316.e1 (2014).

48. Abida, W. *et al.* Analysis of the Prevalence of Microsatellite Instability in Prostate Cancer and Response to Immune Checkpoint Blockade. *JAMA Oncol* **5**, 471–478 (2019).

49. de Bono, J. *et al.* Olaparib for Metastatic Castration-Resistant Prostate Cancer. *N. Engl. J. Med.* **382**, 2091–2102 (2020).

50. Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).

51. Turajlic, S. *et al.* Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell* **173**, 581–594.e12 (2018).

52. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879–893.e13 (2018).

53. Casasent, A. K. *et al.* Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172**, 205–217.e12 (2018).

54. Zhang, K. Stratifying tissue heterogeneity with scalable single-cell assays. *Nature methods* vol. 14 238–239 (2017).

55. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).

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

57. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).

58. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).

59. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).

60. Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat. Med.* **22**, 298–305 (2016).

61. Kumar, A. *et al.* Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.* **22**, 369–378 (2016).

62. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).

63. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).

64. Huang, F. W. *et al.* Exome Sequencing of African-American Prostate Cancer Reveals Loss-of-Function Mutations. *Cancer Discov.* **7**, 973–983 (2017).

65. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).

66. Sun, Y. *et al.* Characterization of genomic clones using circulating tumor DNA in patients with hepatocarcinoma. *Translational Cancer Research* vol. 7 321–329 (2018).

67. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

68. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).

69. Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).

70. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).

71. Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Research* vol. 37 D623–D628 (2009).

72. Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research* vol. 39 D712–D717 (2011).

73. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* vol. 36 983–987 (2018).

74. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

75. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants.

Bioinformatics **31**, 2202–2204 (2015).

76. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

77. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

78. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).

79. Clarke, L. *et al.* The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res.* **45**, D854–D859 (2017).

80. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* vol. 4 (2015).

81. Jørsboe, E., Hanghøj, K. & Albrechtsen, A. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics* **33**, 3148–3150 (2017).

82. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

83. Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).

84. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).

Chapter 5: Conclusions

General Overview

While the work in this dissertation was primarily focused on the development and application of novel computational frameworks to melanoma and prostate cancer, these frameworks can be applied to any cancer type. These include frameworks for 1) identifying driver genes via aggregating information from mutational recurrence, sequence context, and accumulated functional impact-based algorithms followed by pruning of significant results through the integration of bulk and single-cell RNA sequencing, as well as 2) evolutionary

informed driver gene and mutational signature analysis by leveraging bayesian methods for cancer subclone identification and determining the presence of mutational signatures from datasets with few (at least 10) mutations.

Novel driver gene identification framework applied to melanoma

The motivation for developing the driver gene identification framework was largely due to the combined factors of cohort size (> 1,000) and the tumor mutational burden of melanoma (10-13 mutations per megabase). The application of just one, or even multiple algorithms independently, can result in several false positives that frequently include large genes such as *TTN* and olfactory genes that are unrelated to cancer. As the number of samples that undergo whole exome sequencing continues to increase for each cancer type, this will continue to become an issue in driver gene identification analysis, regardless of how low the mutational burden of the cohort may be. Through our novel approach we identified melanoma genomic subtype-specific driver genes that reflected preferential dysregulation of additional pathways outside of MAPK, some of which had implications for immunotherapy response. Accurate identification of driver genes as we continue to saturate the mutational landscape of cancer is imperative for creating driver gene catalogues, understanding biological underpinnings of disease, and identifying potential therapeutic targets and biomarkers. An additional layer of driver gene identification analysis is understanding when and how these genes play a role in a tumor's evolution. Drivers that are more often identified as clonal may be required for tumor initiation, whereas drivers that are more often observed subclonally may be required for advancement to a more advanced disease, or even therapeutic escape mechanisms. Furthermore, how the clonality of driver alterations affects the efficacy and response to various therapies remains an open area of investigation.

Molecular drivers of TWT melanomas

Of the 1,048 melanoma whole-exomes aggregated, 162 were triple wild-type (TWT) melanomas. Unlike the other subtypes, genomic driver analyses in TWT melanomas have been limited due to insufficient cohort size for unbiased driver discovery. Application of our novel driver gene framework to TWT patients identified 19 drivers, including drivers frequently observed in uveal melanomas. Furthermore, while a lack of UV mutagenesis has been reported in TWT tumors, we identified the enrichment of double-stranded break (DSB) repair deficiency signatures in this subtype, which was associated with transcriptional downregulation *ATM* and increased methylation of *INO80*. Follow up analysis in TWT whole-genomes revealed the association between the DSB repair deficiency signature and structural variants affecting the MRN complex. Melanoma patients that exhibit this DSB repair deficiency signature may respond to previously discarded therapeutic modalities such as platinum-based chemotherapy, or those currently unconsidered such as PARP inhibitors (PARPi) or ATR inhibitors (ATRi).

Evolutionary informed genomic analysis in prostate cancer

Mutational signature analysis has traditionally been performed on all mutations from a tumor at once, where accurate inference of the clonality of a mutational signature was rarely possible. Here we demonstrated that integration of phylogenetic reconstruction methods with a likelihood-based mutational signature algorithm designed for clinical panel sequencing data is able to determine the presence of mutational signatures at a cell subpopulation resolution. Through this framework we can link the presence of certain mutational signatures such as homologous recombination deficiency (HRD) and mismatch repair deficiency (MMRd) to specific subclones, and in certain cases identify putative causal somatic or germline alterations within that subclone. This approach also identifies additional patients with HRD and MMRd undetectable by traditional mutational signature analysis, potentially increasing in number of patients amenable to PARPi or platinum-based chemotherapy. Specifically, we showed that the

application of our novel integrative approach to a subset of prostate cancer tumors treated with PARPi, and ovarian cancers treated with cisplatin chemotherapy, identified more patients that benefited from these therapies, demonstrating the potential expanded therapeutic impact of our approach.

Finally, we observed that despite primary tumors in Black patients being associated with higher biochemical recurrence rates, these tumors tend to be more linear and monoclonal, which is contrary to observations between adverse outcomes and polyclonicity in White patient cohorts. One possible explanation is that these tumors had undergone rapid selective sweeps, or “punctuated evolution”, prior to being screened and undergoing whole-exome sequencing. Rapid selective sweeps are one growth model for tumor evolution, where all cell subpopulations in a tumor collapse into one dominant clone, manifesting in a monoclonal tumor. While we do provide evidence that these apparent rapid selective sweeps may partially be explained by inherited germline SNPs associated with increased risk of prostate cancer, we do not possess the relevant socioeconomic or standard of care information necessary to determine the true extent that genetics plays in this observation.

Future Directions

Driver gene identification and saturation in prostate cancer and renal clear cell carcinoma

Since the conclusion of the melanoma whole-exome study we’ve continued to aggregate prostate cancer and renal clear cell carcinoma whole-exomes. To date we’ve aggregated over 2,000 prostate cancer whole-exomes and over 1,000 renal cell carcinoma whole-exomes, both of which are the largest cohorts in their respective cancer type to date. Based on cancer type specific mutational burden, power analysis suggests that roughly 2,000 prostate cancer and 1,000 renal clear cell carcinoma whole-exomes will allow the identification of all driver genes mutated at a frequency of at least 1% and 3%, respectively. Application of our novel driver

identification framework will yield a set of high confidence driver genes for each cohort, and in the case of prostate cancer saturate the landscape of driver genes. The set of drivers identified from these analyses will have the potential to further our understanding of prostate and renal cell cancer biology, and enable the identification of therapeutic targets or biomarkers in genomically stratified patient subsets.

Sensitivity of DSB repair deficient melanomas to relevant therapeutic compounds

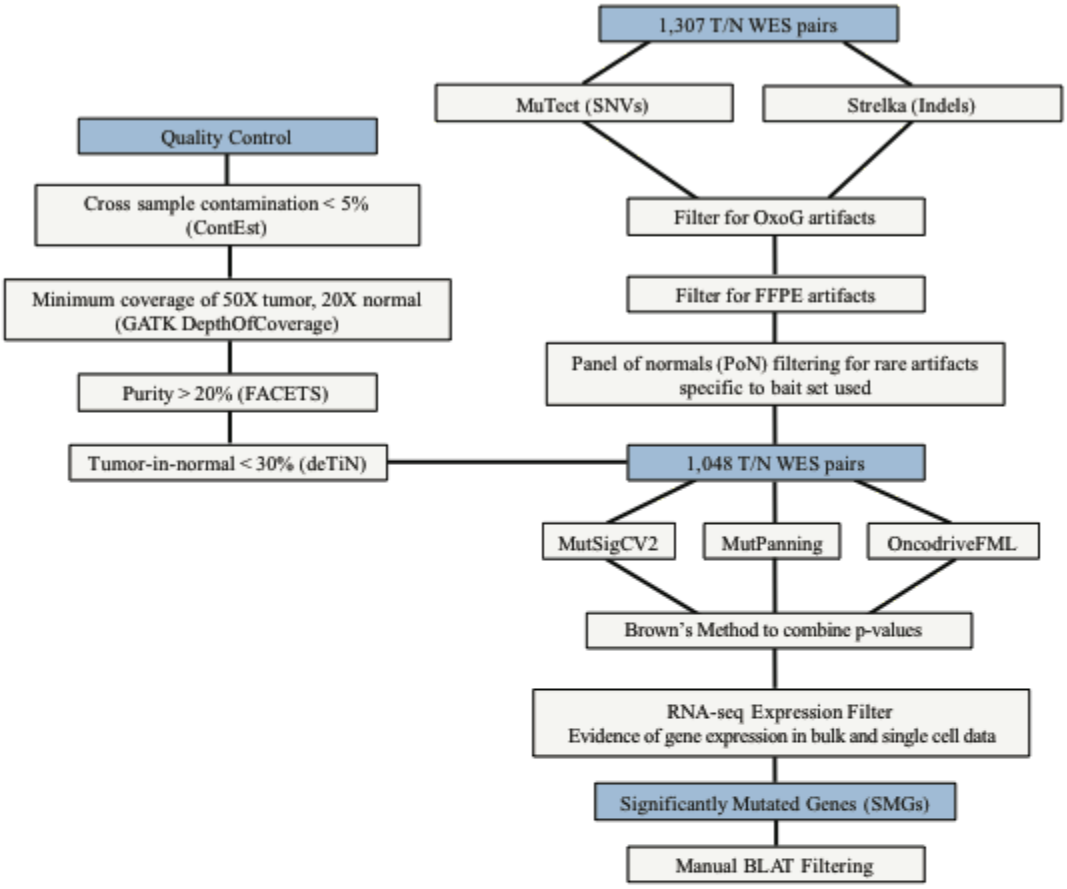
We identified the enrichment of DSB repair deficiency signatures in TWT cutaneous melanomas that is associated with the downregulation of *ATM*. However, we were unable to identify causal mechanisms for the downregulation of *ATM* in our subset of TWT whole-exomes, although follow analysis in whole-genomes suggested that alterations affecting the MRN complex may be a potential mechanism. Additionally, understanding whether tumors with the DSB repair deficiency signature may respond to various therapies previously unconsidered in melanoma, such as PARPi and ATRi, is a necessary first step for translating these findings to the clinic. To address these, we plan to perform CRISPR knockout screens in TWT and non-TWT melanoma cell lines, followed by assessing the sensitivity of these knockout cell lines to PARPi and ATRi.

Effect of subclonal HRD and MSI on response to PARPi and immunotherapy

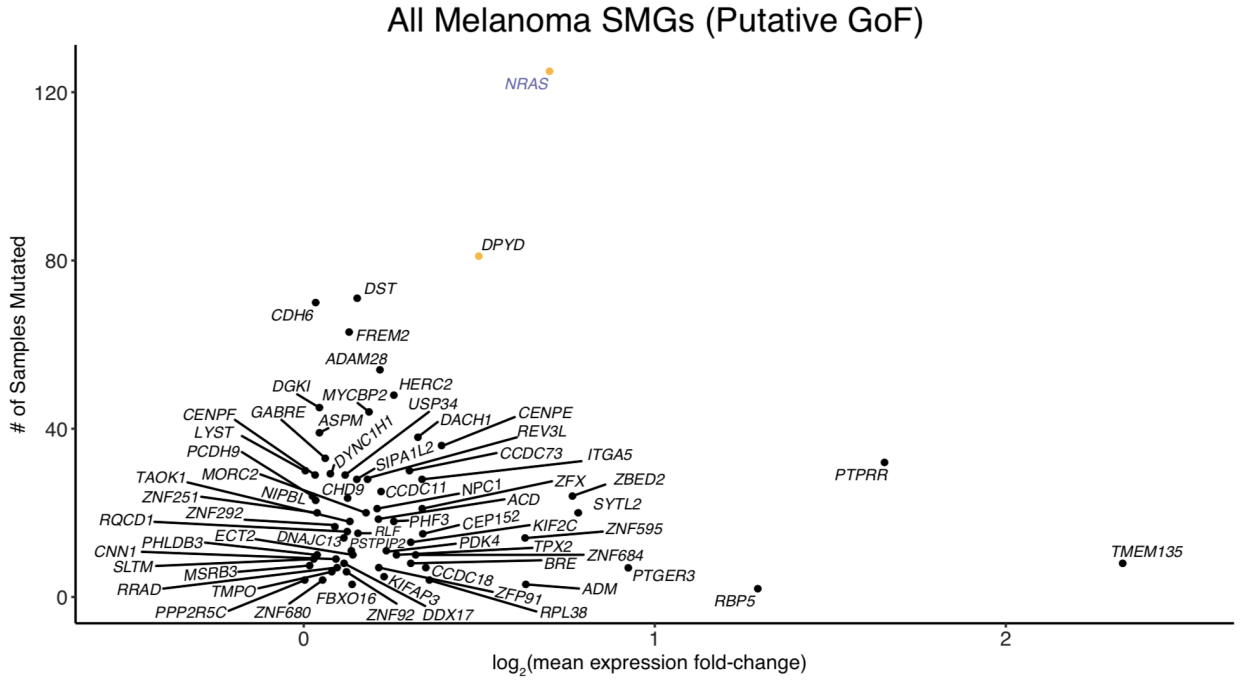
We demonstrated that our novel computational framework for evolutionary informed signature analysis at the cell subpopulation resolution was able to identify additional cases of HRD and MMRd. In a limited cohort of prostate cancer patients treated with PARPi, we showed that HRD patients only identified at the cell subpopulation level had better response than patients without HRD, and similar response to HRD patients identified with more traditional approaches. We also demonstrated similar results in the subset of TCGA ovarian cancer patients treated with platinum-based chemotherapy. To determine the robustness of these

results, as well as determine if the clonality of HRD matters for response to relevant therapies, we would like to apply our framework to larger cohorts of PARPi or platinum-based chemotherapy treated prostate, breast and ovarian cancer patients. Furthermore, we would also like to address these same questions in immunotherapy treated cohorts for patients with MMRd identified via our framework.

Appendix

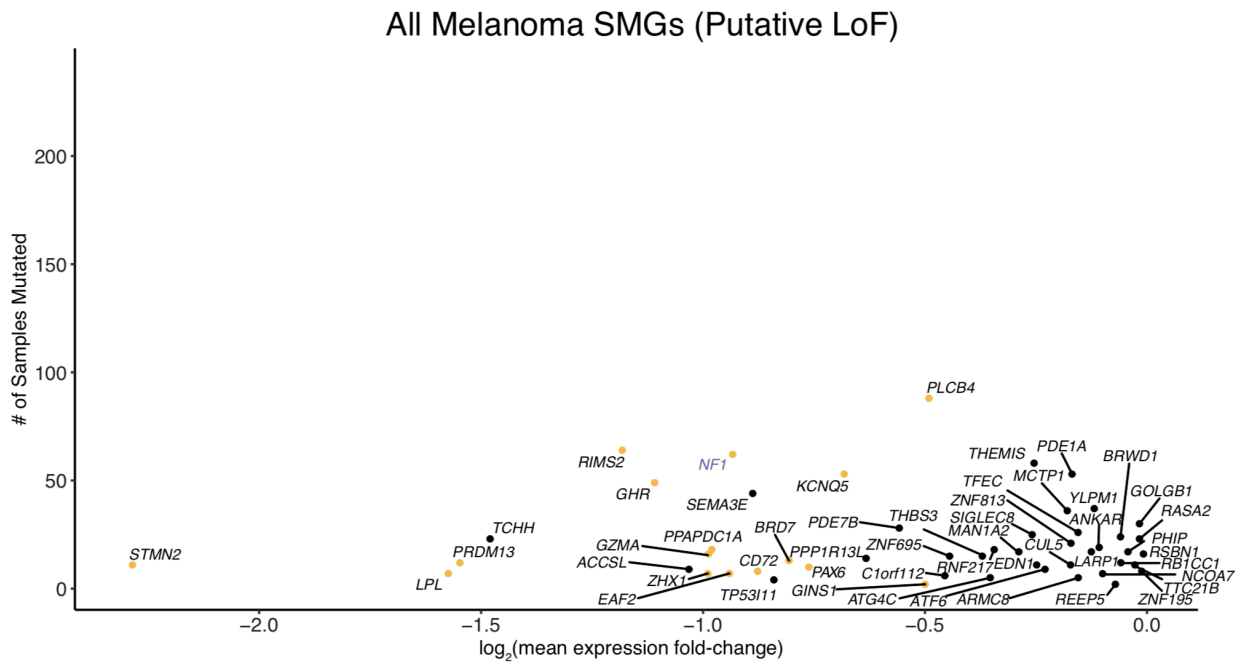


Supplementary Figure 2.1: Quality control, mutation calling and mutational significance workflow



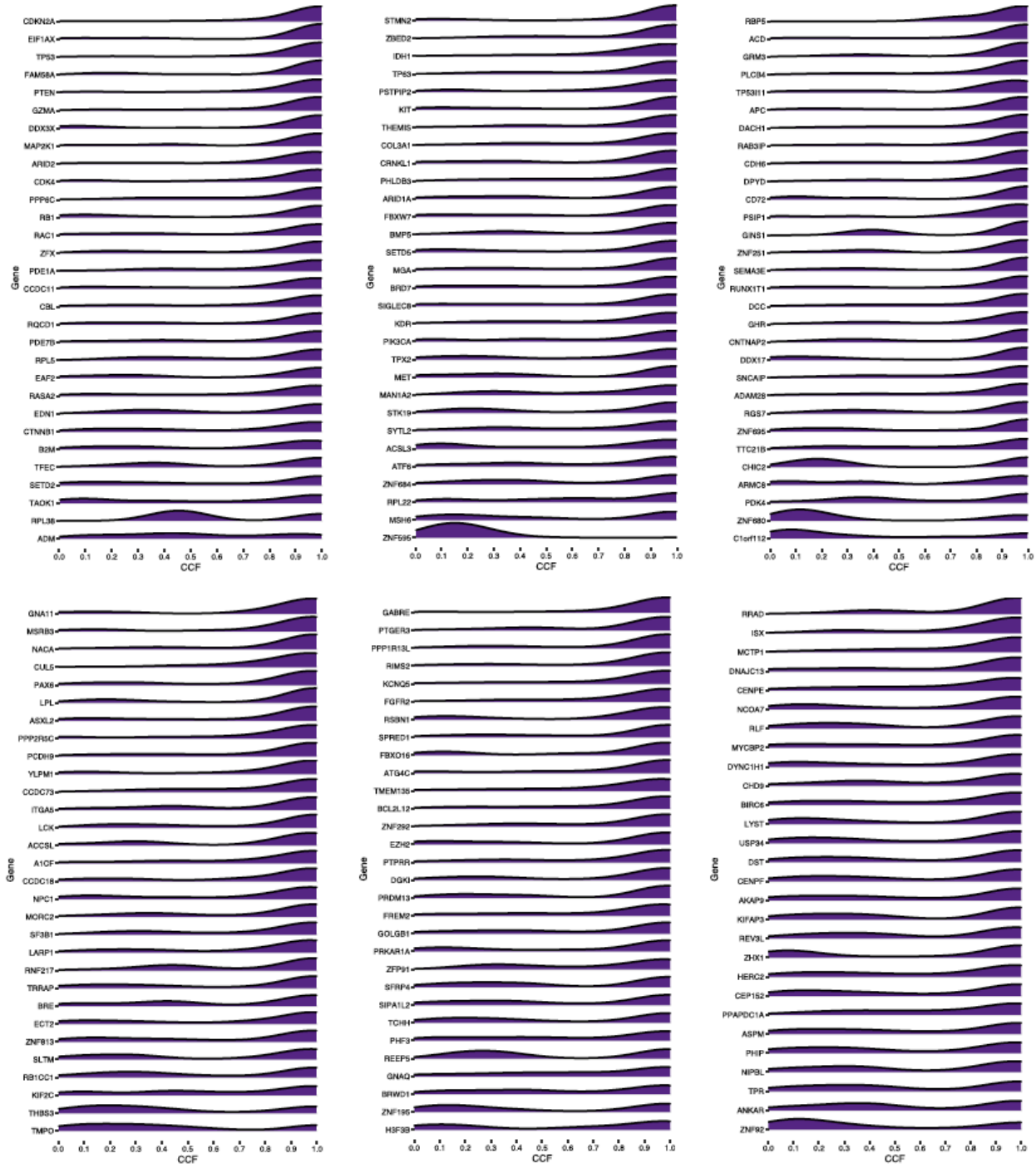
Supplementary Figure 2.2: Expression differences between mutant vs. wild-type putative gain of function, previously unknown cancer gene, melanoma SMGs

Mean expression fold-change differences (in TCGA samples) between mutant vs. wild-type melanoma putative gain of function (GoF) SMGs that are not in the COSMIC Cancer Gene Census or OncoKB databases. *NRAS* is included as a reference for GoF mutations (purple name). Genes highlighted by a yellow point have a statistically significant difference in expression between mutant vs. wild-type tumors.

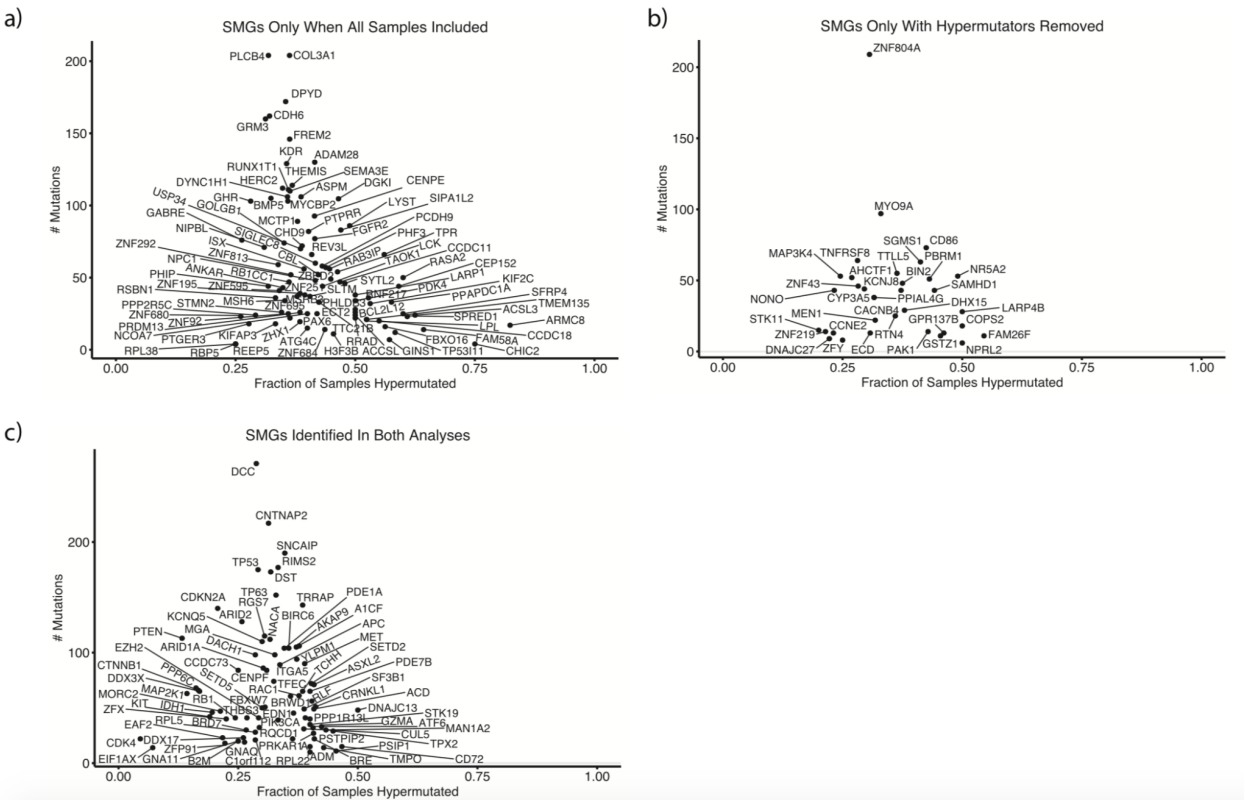


Supplementary Figure 2.3: Expression differences between mutant vs. wild-type putative loss of function, previously unknown cancer gene, melanoma SMGs

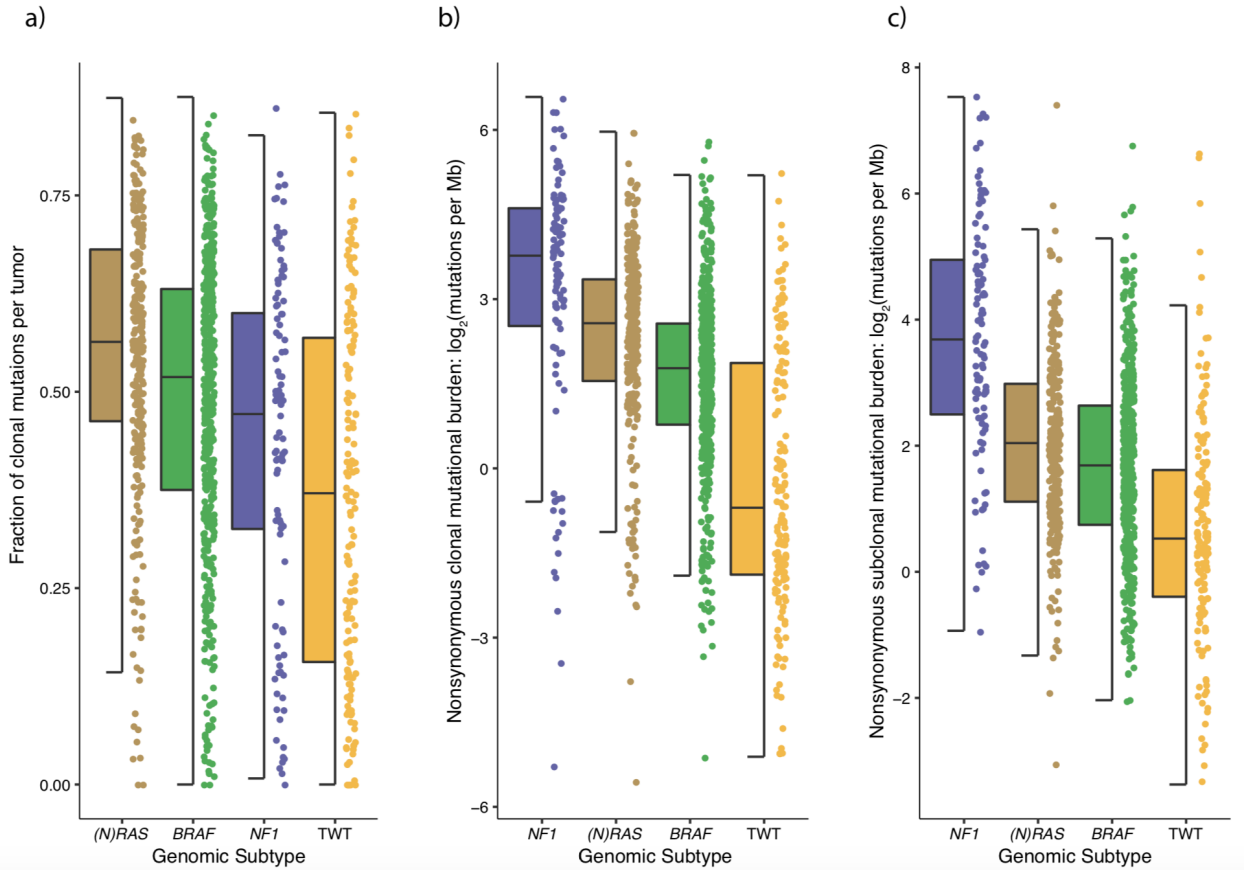
Mean expression fold-change differences (in TCGA samples) between mutant vs. wild-type melanoma putative loss of function (LoF) SMGs that are not in the COSMIC Cancer Gene Census or OncoKB databases. *NF1* is included as a reference for LoF mutations (purple name). Genes highlighted by a yellow point have a statistically significant difference in expression between mutant vs. wild-type tumors.



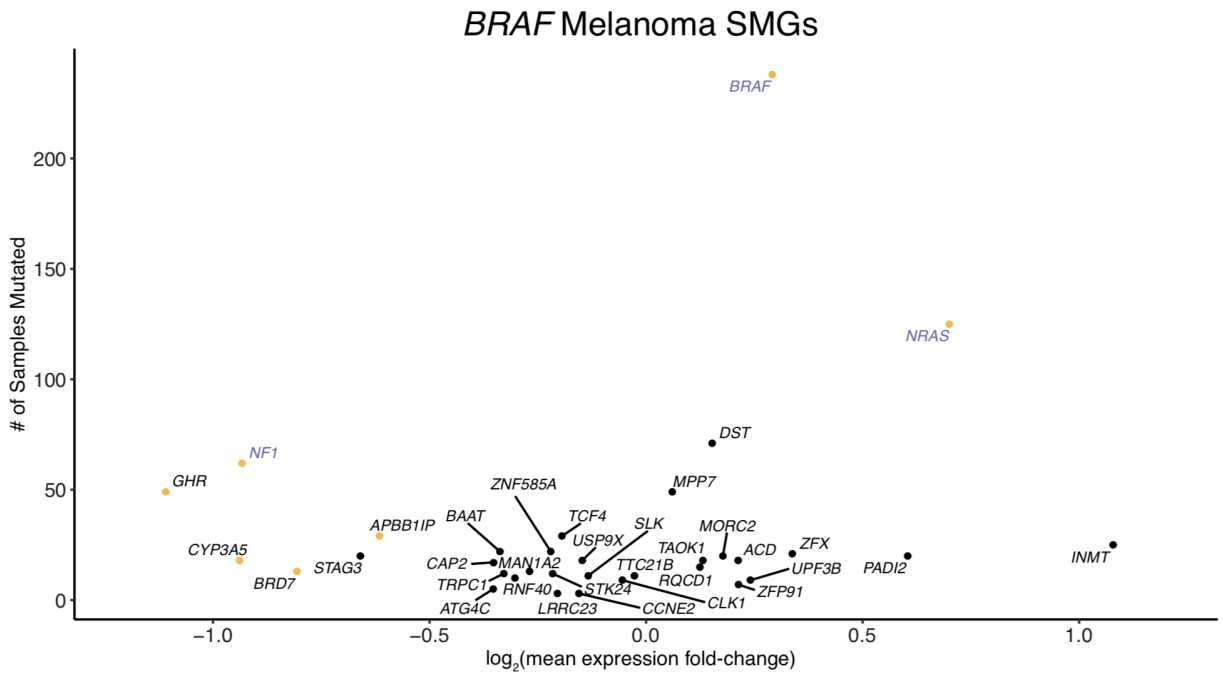
Supplementary Figure 2.4: CCFs of SMGs identified in the full cohort of 1,048 melanomas
Density plots showing the distribution of CCFs for mutations in melanoma SMGs. Some genes are almost always clonal (e.g. *CDKN2A*, *EIF1AX*), while others are bimodal (e.g. *GINS1*, *EZH2*) indicating those genes may be both clonal and subclonal drivers.



Supplementary Figure 2.5: Impact of hypermutated tumors on SMG analysis
To determine the effect hypermutated tumors may have on false positives in our cohort, we classified tumors in the top 10% of mutational burden as hypermutated tumors. Here we show SMGs identified only in the **(a)** entire cohort (n = 1048), **(b)** only when hypermutators are removed (n = 943), and **(c)** both. **(a)** For some genes only called SMGs when including all samples, over half the mutations are from hypermutator tumors, including known cancer genes (e.g. *CHIC2*, *FAM58A*). Further, a nontrivial amount of hypermutator tumors are NF1 melanomas (49%), and several SMGs identified only when including all tumors are driven by NF1 melanoma (e.g. *SPRED1*, *RASA2*). **(c)** For all genes identified in both analyses, the fraction of mutations belonging to hypermutated tumors never exceeded 50%. However, this phenomenon was also observed in genes only identified when **(a)** including all samples, and **(b)** when removing hypermutators. Thus, the covariates included in mutational significance algorithms likely contribute more to statistical significance than the fraction of mutations contributed to hypermutated tumors. Indeed, the Brown's p-values of SMGs (Benjamini-Hochberg, q-value cutoff < 0.1) was not associated with the fraction of mutations contributed by hypermutated tumors (linear regression, $p > 0.05$, two-sided). Expanding to all genes, the fraction of mutations contributed by hypermutated tumors slowly becomes more significantly associated with higher p-values (linear regression, $p < 0.05$, two-sided). This is likely because hypermutated tumors comprise a large percentage of mutations for infrequently mutated genes.

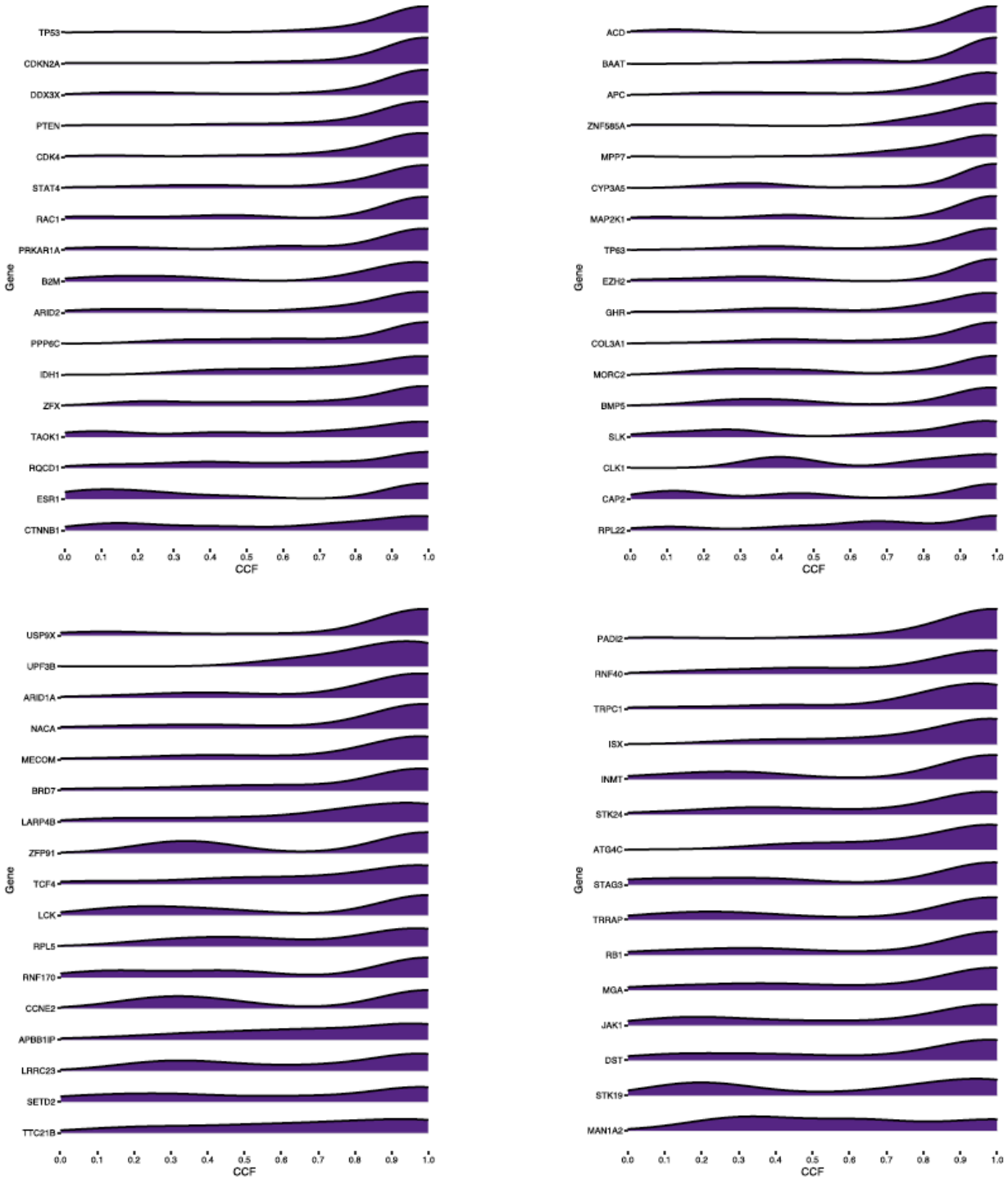


Supplementary Figure 2.6: Clonal and subclonal mutations per genomic subtype
(a) Distribution of clonal:subclonal mutation ratios per genomic subtype (Mann-Whitney U, $p < 1.27 \times 10^{-4}$ for all pairwise). **(b)** Distribution of clonal (Mann-Whitney U, $p < 5.6 \times 10^{-8}$ for all pairwise) and **(c)** subclonal nonsynonymous mutational burdens per genomic subtype (Mann-Whitney U, $p < 6.0 \times 10^{-4}$ for all pairwise). The data are represented as boxplots where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points. The p-values derived from the Mann-Whitney U tests are two-sided.



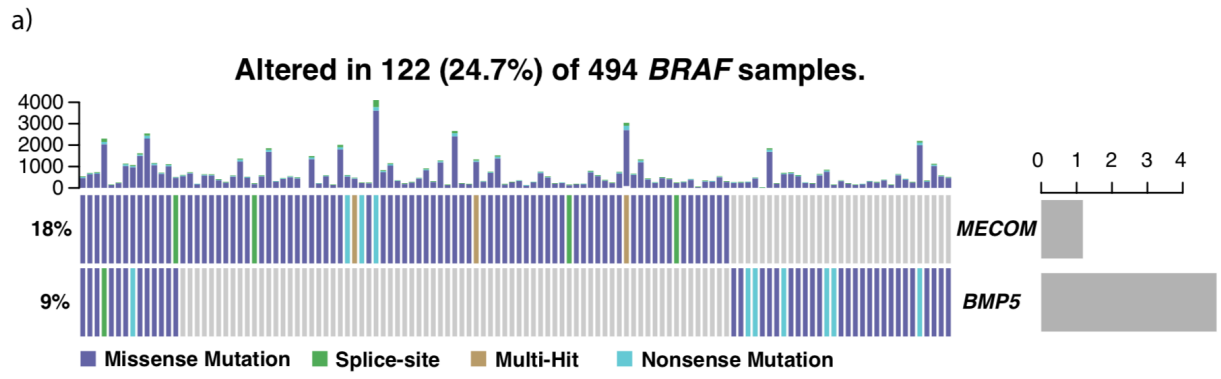
Supplementary Figure 2.7: Expression differences between mutant vs. wild-type, previously unknown cancer gene, *BRAF* melanoma SMGs

Mean expression fold-change differences (in TCGA samples) between mutant vs. wild-type *BRAF* melanoma SMGs that are not in the COSMIC Cancer Gene Census or OncoKB databases. *BRAF*, *NRAS*, and *NF1* are included as references for gain of function (GoF) and loss of function (LoF) mutations (purple names). Genes with a mean fold-change difference above 0 indicate higher expression in mutant samples compared to wild-type (i.e. GoF mutations). Genes with a mean fold-change difference below 0 indicate lower expression in mutant samples compared to wild-type (i.e. LoF mutations). Genes highlighted by a yellow point have a statistically significant difference in expression between mutant vs. wild-type tumors.

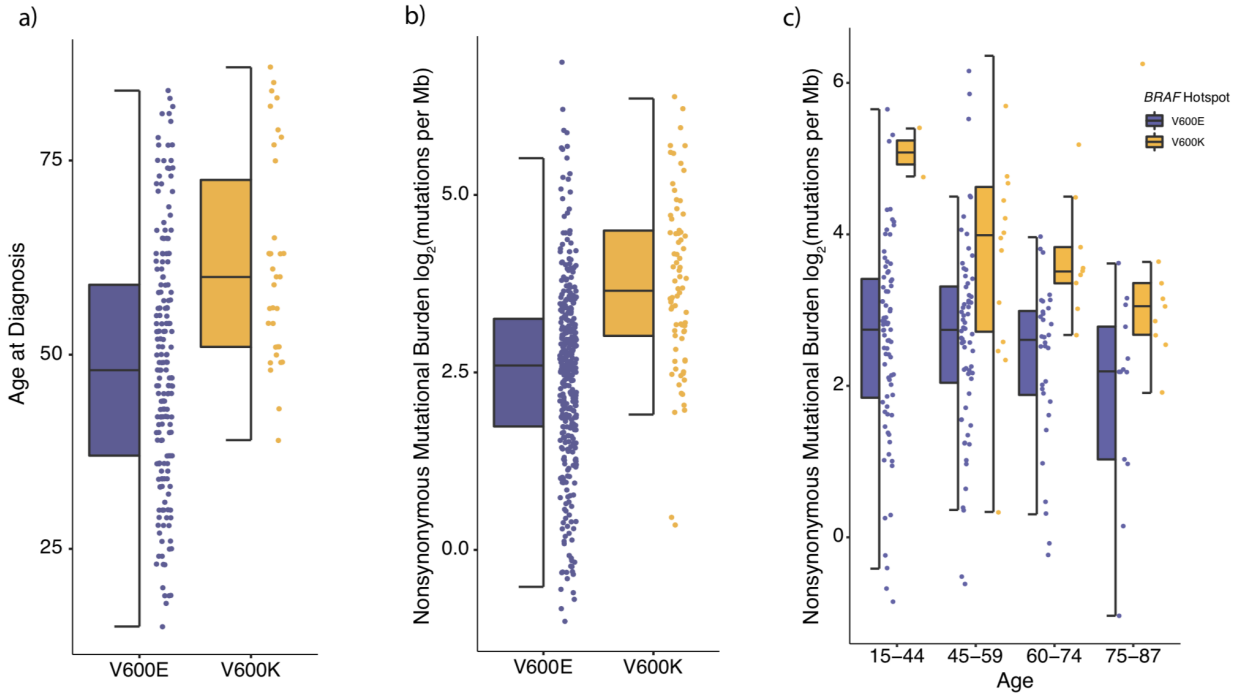


Supplementary Figure 2.8: CCFs of *BRAF* SMGs

Density plots showing the distribution of CCFs for mutations in *BRAF* SMGs. Some genes are almost always clonal (e.g. *STAT4*, *DDX3X*), while others are bimodal (e.g. *STK19*, *ZFP91*) indicating those genes may be both clonal and subclonal drivers.

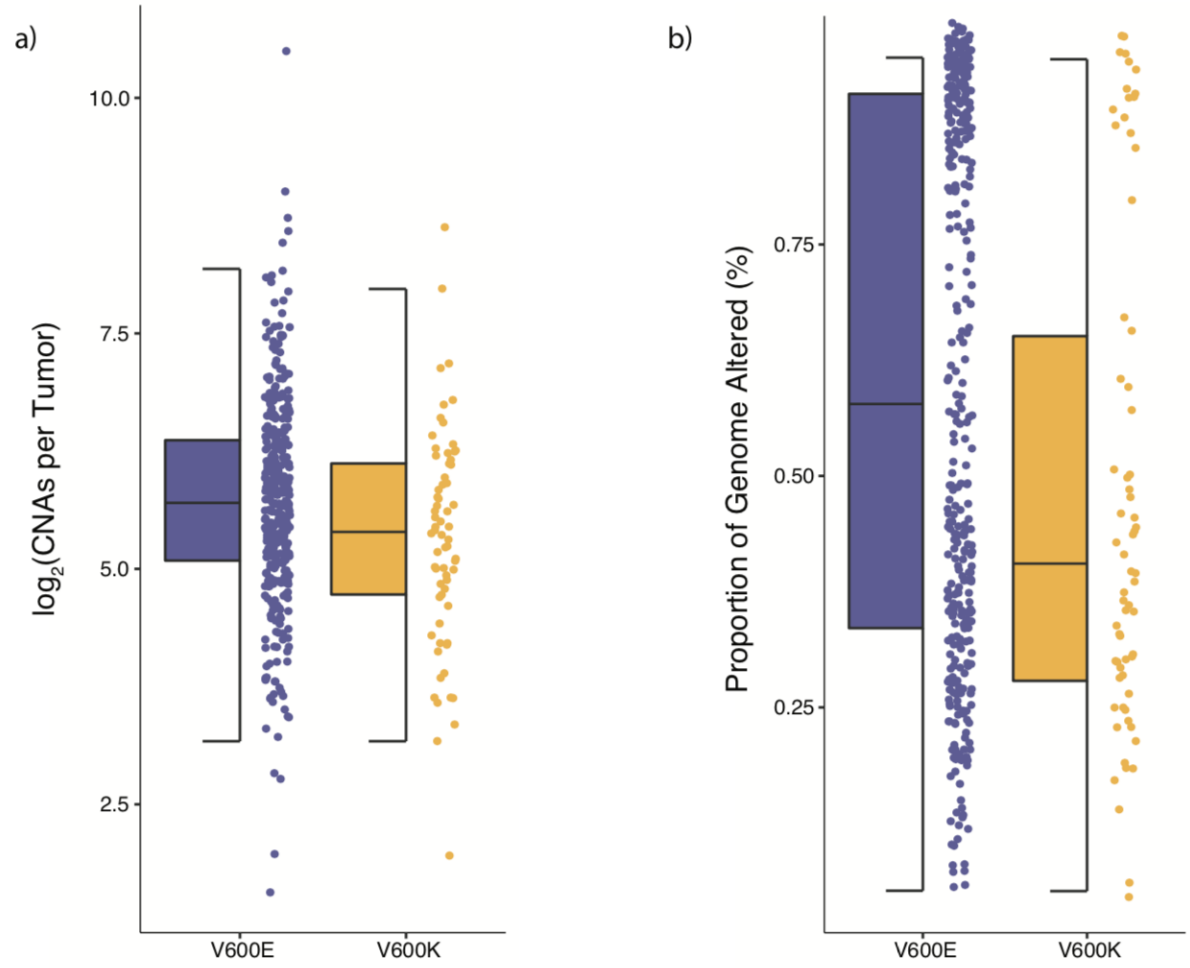


Supplementary Figure 2.9: SMGs in the *BRAF* melanoma subtype
(a) CoMut plot focusing on the TGF-β pathway associated SMGs (*MECOM*, *BMP5*) identified exclusively in *BRAF* melanomas. *MECOM* is an antagonist of the TGF-β pathway (specifically with the SMAD genes), as is *BMP5* (Alliston et al., 2005; Bramlage et al., 2011).

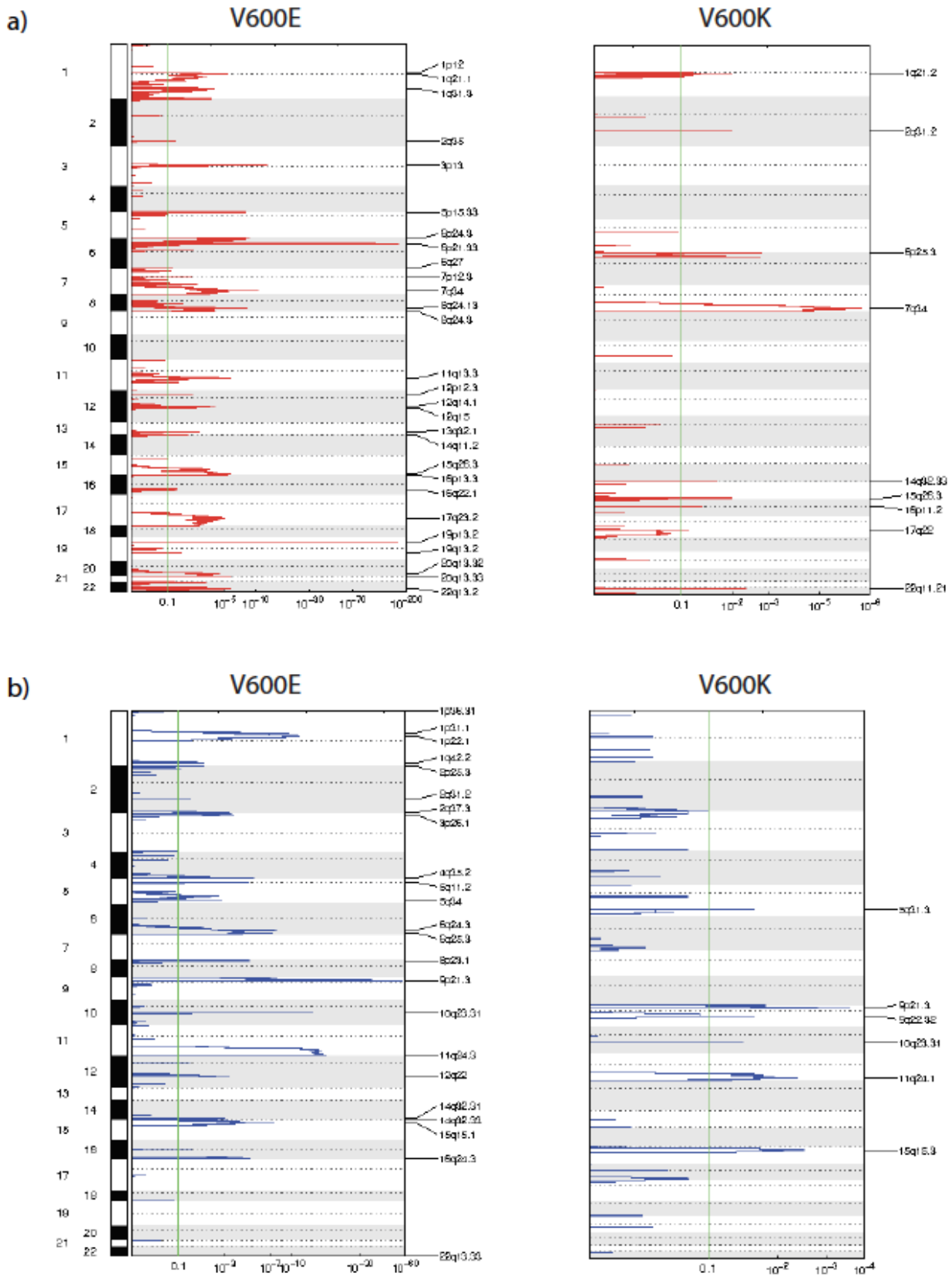


Supplementary Figure 2.10: *BRAF* V600E vs. V600K mutational burden when stratifying by age
(a) Age at diagnosis was significantly older in *BRAF* V600K patients compared to V600E patients (Mann-Whitney U, $p = 1.07 \times 10^{-5}$). **(b)** Nonsynonymous mutational load was significantly elevated in *BRAF* V600K melanomas compared to V600E melanomas (Mann-Whitney U, 1.2×10^{-13}). (c) Even when stratifying by age there is still a significant increase in mutations in V600K tumors compared to V600E tumors (15-44 yrs old: 34.68 mut/Mb vs. 6.69 mut/Mb, Mann-Whitney U, p = 0.024; 45-59 yrs old: 15.88 mut/Mb vs. 6.68 mut/Mb, Mann-Whitney U, p = 0.007; 60-74 yrs old: 11.39 mut/Mb vs. 6.1 mut/Mb,

Supplementary Figure 2.10 (continued): Mann-Whitney U, p = 0.0002; 75-87 yrs old: 8.29 mut/Mb vs. 4.57 mut/Mb, Mann-Whitney U, p = 0.03). The data are represented as boxplots where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points. The p-values derived from the Mann-Whitney U tests are two-sided.

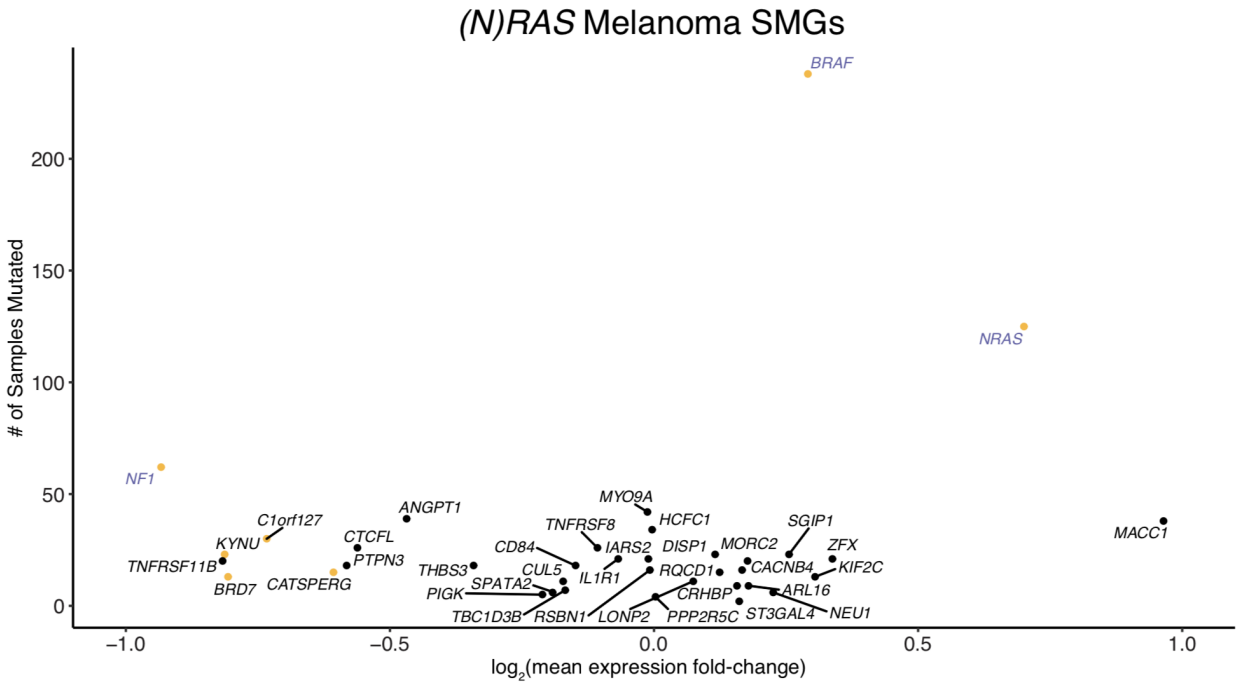


Supplementary Figure 2.11: Global CNA properties of *BRAF* V600E and V600K samples **(a)** *BRAF* V600E tumors experience significantly more copy number events than V600K tumors (49.5 vs. 42, Mann-Whitney U, p = 0.027,). **(b)** Likewise, V600E tumors also have a significantly higher proportion of the genome altered compared to V600K tumors (54.2% vs. 42.5%, Mann-Whitney U, p = 0.028). The data are represented as boxplots where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points. The p-values derived from the Mann-Whitney U tests are two-sided.


a)
V600E
V600K
b)
V600E
V600K

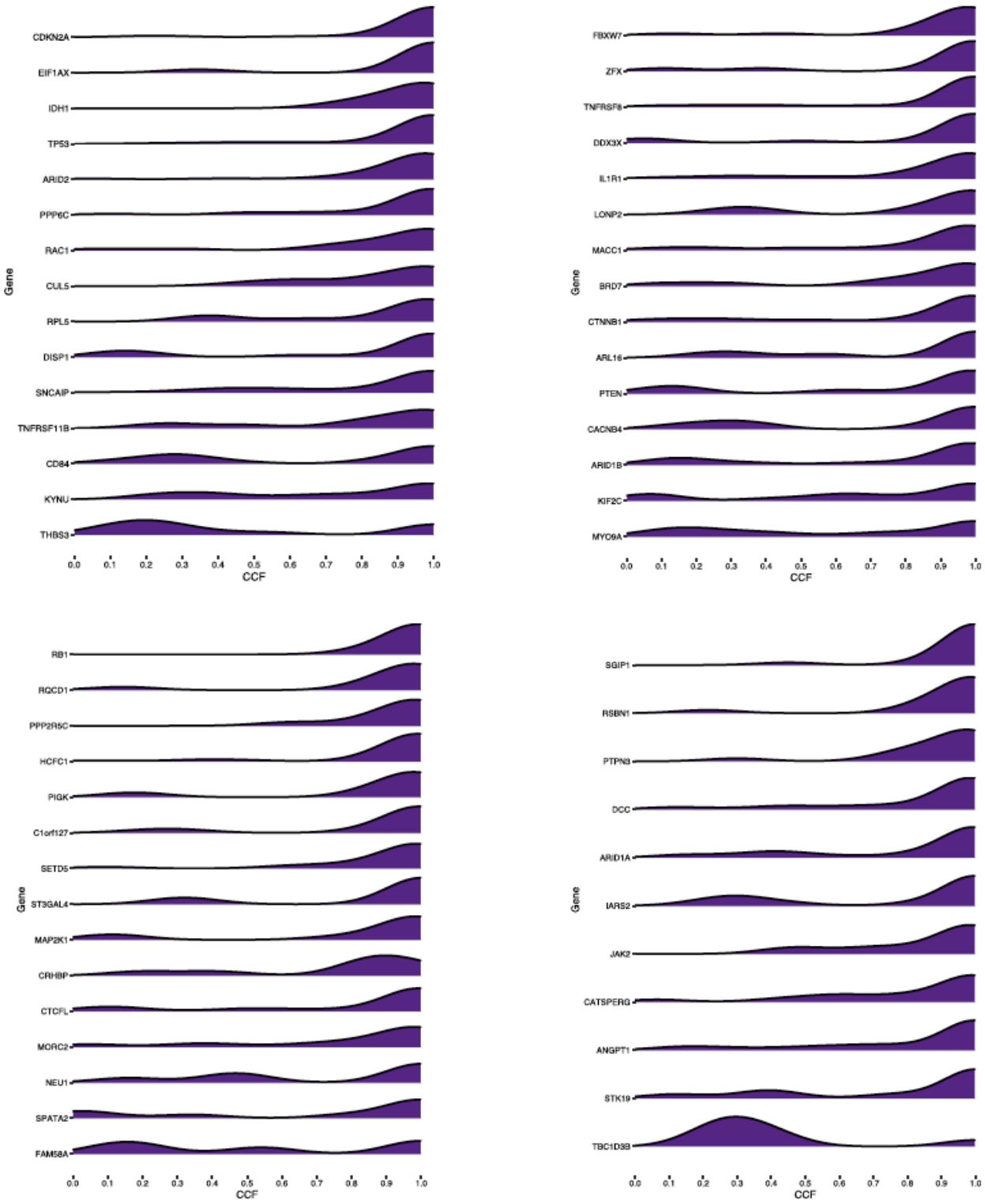
Supplementary Figure 2.12: GISTIC2.0 amplification and deletion peaks for *BRAF* V600E and V600K melanomas

(a) Significant amplification regions in *BRAF* V600E and *BRAF* V600K melanomas (Benjamini-Hochberg, q-value cutoff < 0.1). **(b)** Significant deletion regions in *BRAF* V600E and *BRAF* V600K melanomas (Benjamini-Hochberg, q-value cutoff < 0.1). The complete list of peaks and the genes they contain can be found in SupplementaryTable 2.12. Supplementary Table 2.12 also contains annotations on what genes contained in the peaks are in the CGC and OncoKB, and what genes were called a SMG in the same genomic subtype.



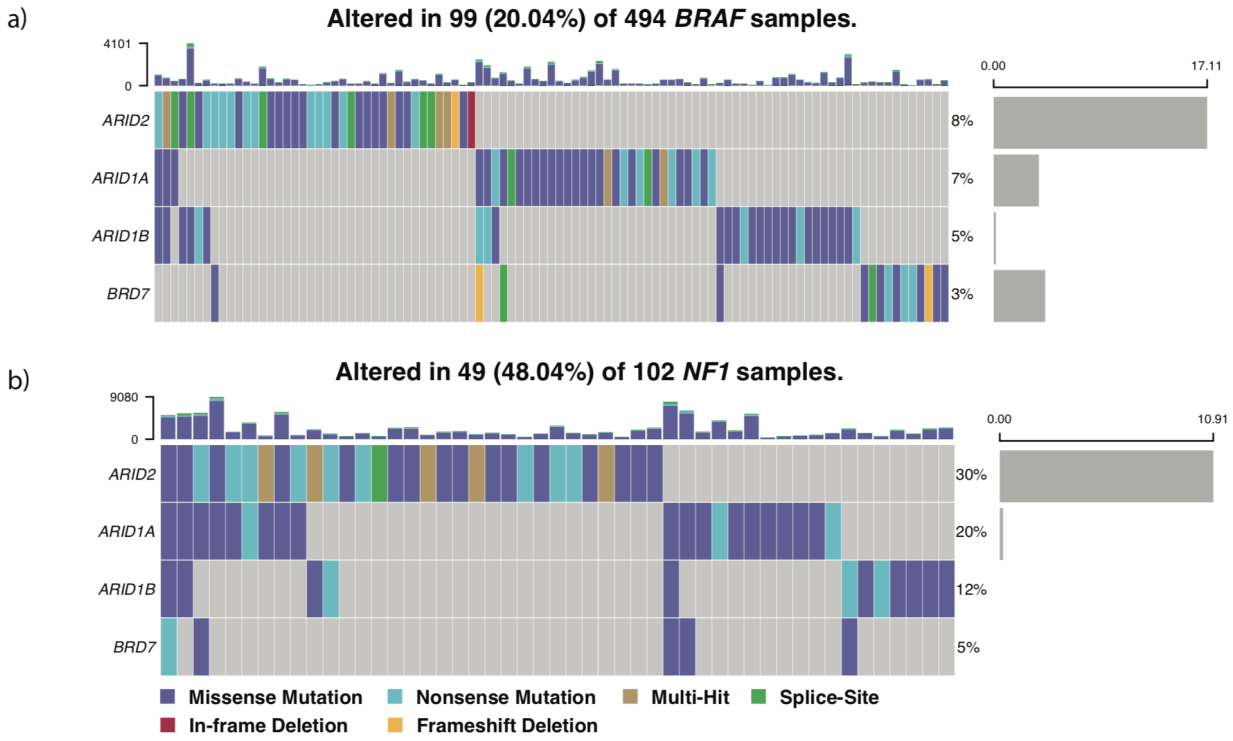
Supplementary Figure 2.13: Expression differences between mutant vs. wild-type, previously unknown cancer gene, *(N)RAS* melanoma SMGs

Mean expression fold-change differences (in TCGA samples) between mutant vs. wild-type *(N)RAS* melanoma SMGs that are not in the COSMIC Cancer Gene Census or OncoKB databases. *BRAF*, *(N)RAS*, and *NF1* are included as references for gain of function (GoF) and loss of function (LoF) mutations (purple names). Genes with a mean fold-change difference above 0 indicate higher expression in mutant samples compared to wild-type (i.e. GoF mutations). Genes with a mean fold-change difference below 0 indicate lower expression in mutant samples compared to wild-type (i.e. LoF mutations). Genes highlighted by a yellow point have a statistically significant difference in expression between mutant vs. wild-type tumors.



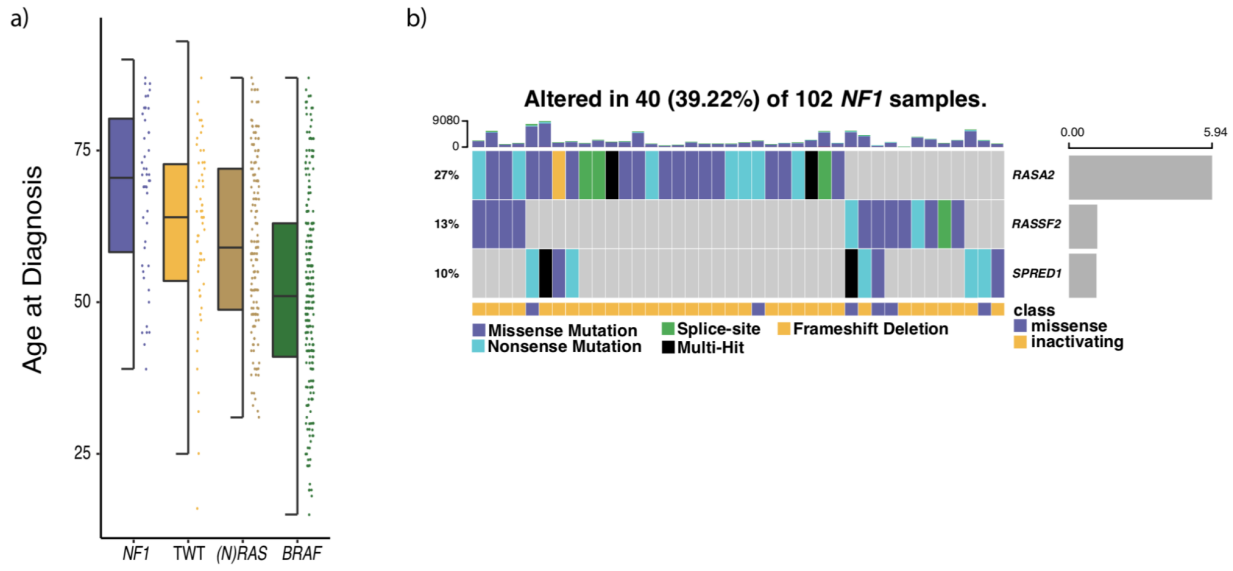
Supplementary Figure 2.14: CCFs of *(N)RAS* SMGs

Density plots showing the distribution of CCFs for mutations in *(N)RAS* SMGs. Some genes are almost always clonal (e.g. *CDKN2A*, *RB1*), while others are bimodal (e.g. *IARS2*, *LONP2*) indicating those genes may be both clonal and subclonal drivers.

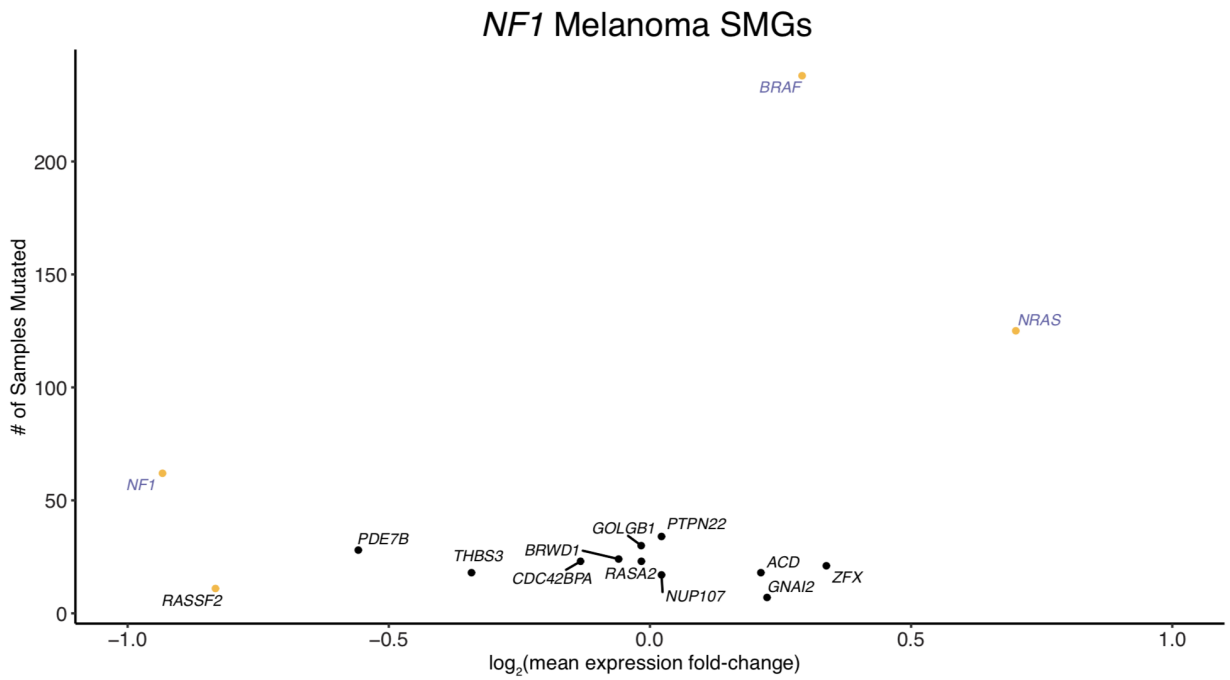


Supplementary Figure 2.15: Mutations in BAF/PBAF complex genes identified as SMGs in *(N)RAS* melanomas

(a) CoMut plot of *(N)RAS* subtype BAF/PBAF complex SMGs in *BRAF* melanomas. *ARID2*, *ARID1A*, and *BRD7* were identified as SMGs in the *BRAF* subtype, although at lower frequencies and statistical significance. **(b)** CoMut plot of *(N)RAS* subtype BAF/PBAF complex SMGs in *NF1* melanomas. Although a higher proportion of *NF1* melanomas harbored mutations in these genes compared to *(N)RAS* melanomas, only *ARID2* was identified as significantly mutated. Further, the majority of mutations in *NF1* melanomas are not putative loss of function (nonsense mutations, splice-site variants and indels).

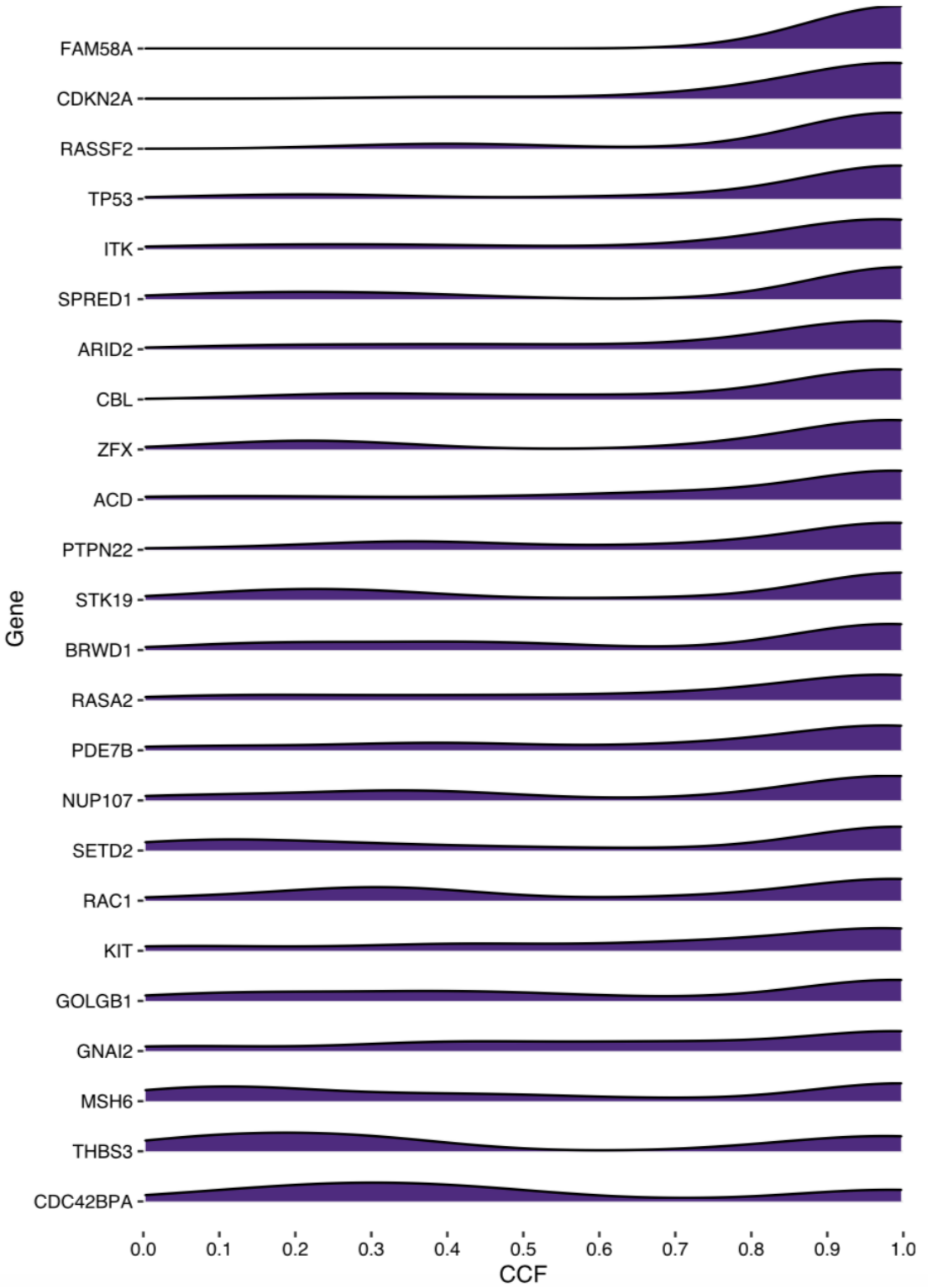


Supplementary Figure 2.16: Clinical characteristics and SMGs in *NF1* melanomas
(a) Distribution of age at diagnosis between the genomic subtypes. *NF1* melanomas are associated with significantly older age at diagnosis compared to the other genomic subtypes (Mann-Whitney U, $p < 0.028$ pairwise for all, two-sided). The data is represented as a boxplot where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points. **(b)** The co-mutation plot of *NF1* RASopathy SMGs and the novel RAS-associated SMG *RASSF2*, including the annotation of missense and inactivating *NF1* mutations. Loss of function mutations in the RASopathy genes (*RASA2* and *SPRED1*) were never observed in the same tumor, as were loss of function mutations between *SPRED1* and *RASSF2*. One tumor harbored loss of function mutations in both *RASA2* and *RASSF2*.

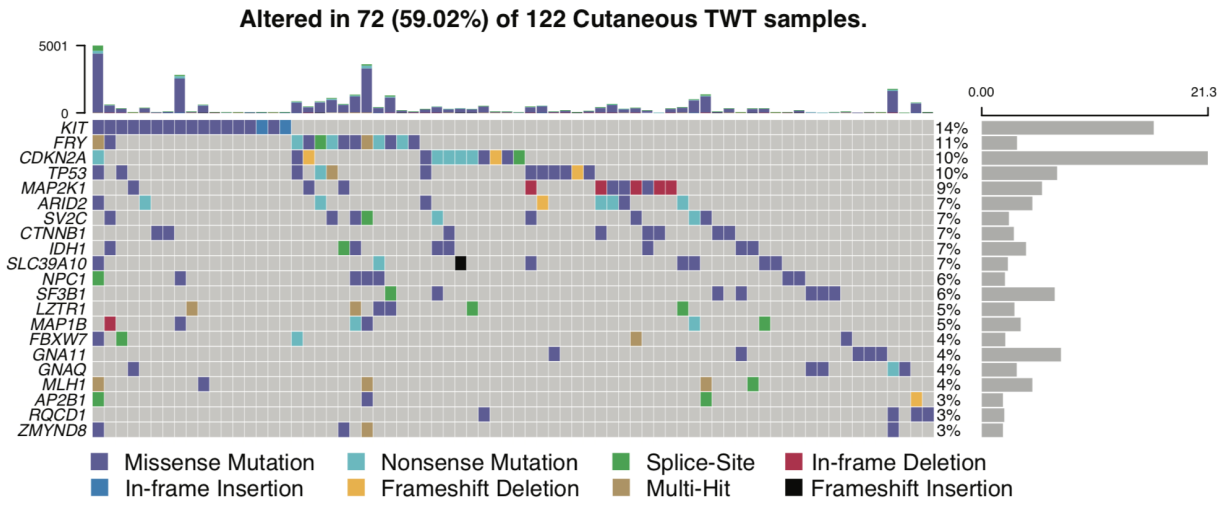


Supplementary Figure 2.17: Expression differences between mutant vs. wild-type, previously unknown cancer gene, *NF1* melanoma SMGs

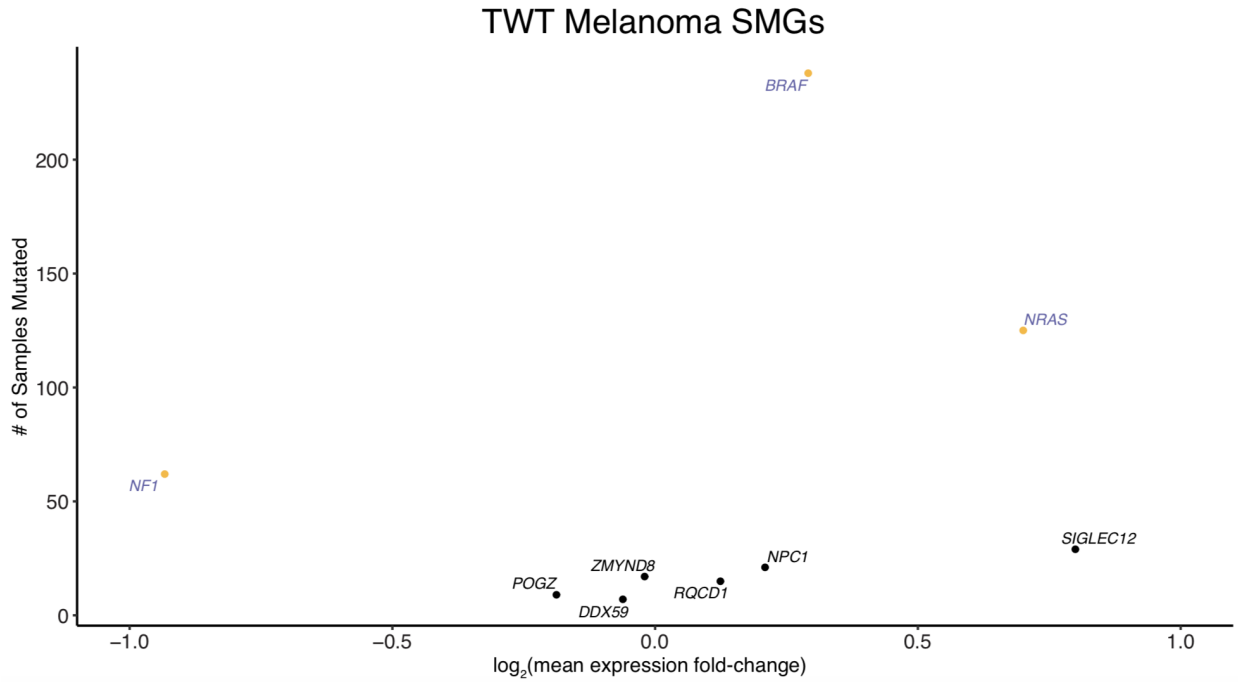
Mean expression fold-change differences (in TCGA samples) between mutant vs. wild-type *NF1* melanoma SMGs that are not in the COSMIC Cancer Gene Census or OncoKB databases. *BRAF*, *NRAS*, and *NF1* are included as references for gain of function (GoF) and loss of function (LoF) mutations (purple names). Genes with a mean fold-change difference above 0 indicate higher expression in mutant samples compared to wild-type (i.e. GoF mutations). Genes with a mean fold-change difference below 0 indicate lower expression in mutant samples compared to wild-type (i.e. LoF mutations). Genes highlighted by a yellow point have a statistically significant difference in expression between mutant vs. wild-type tumors.



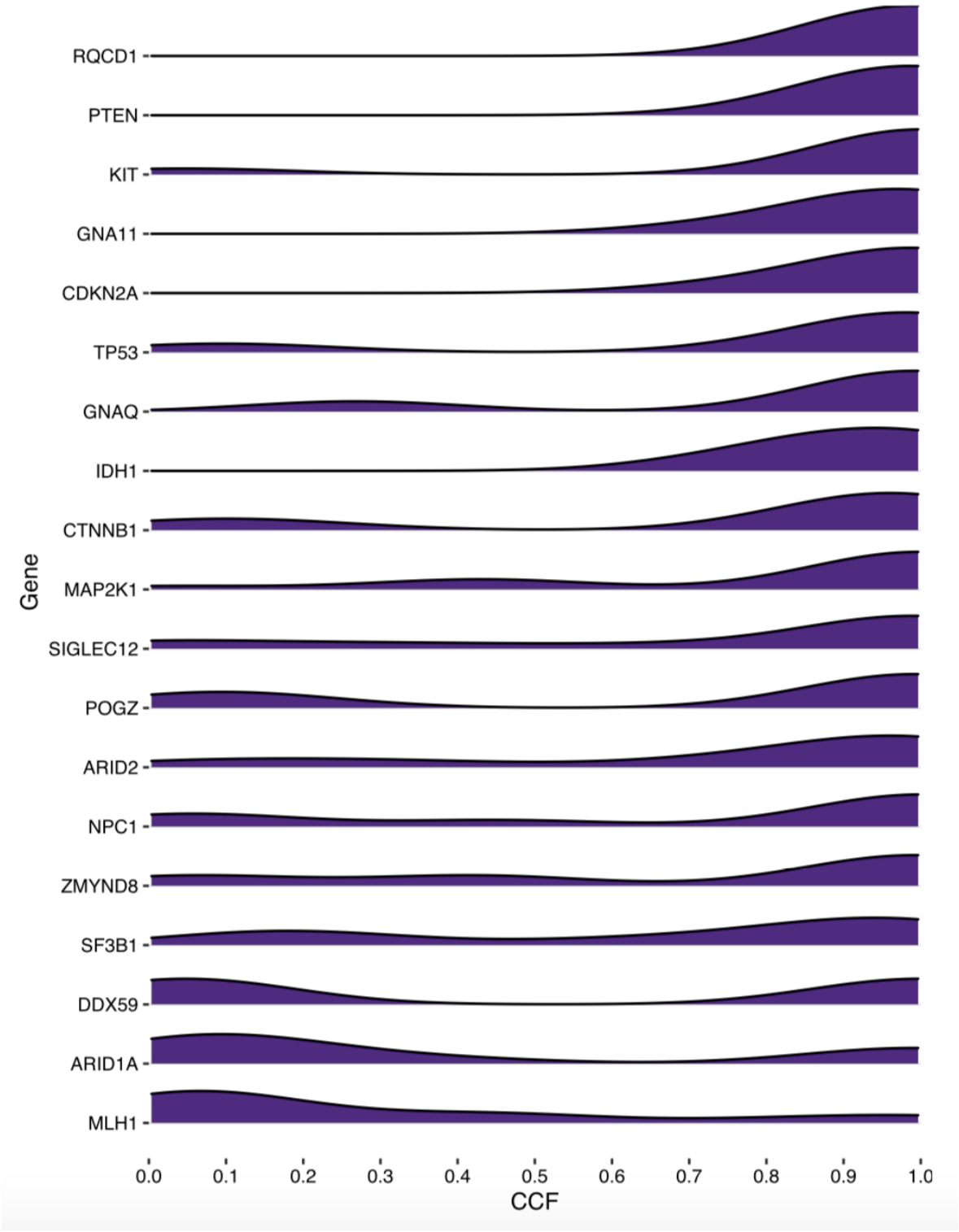
Supplementary Figure 2.18: CCFs of *NF1* SMGs
Density plots showing the distribution of CCFs for mutations in *NF1* SMGs. Some genes are almost always clonal (e.g. *FAM58A*, *RASSF2*), while others are bimodal (e.g. *RAC1*, *MSH6*) indicating those genes may be both clonal and subclonal drivers.



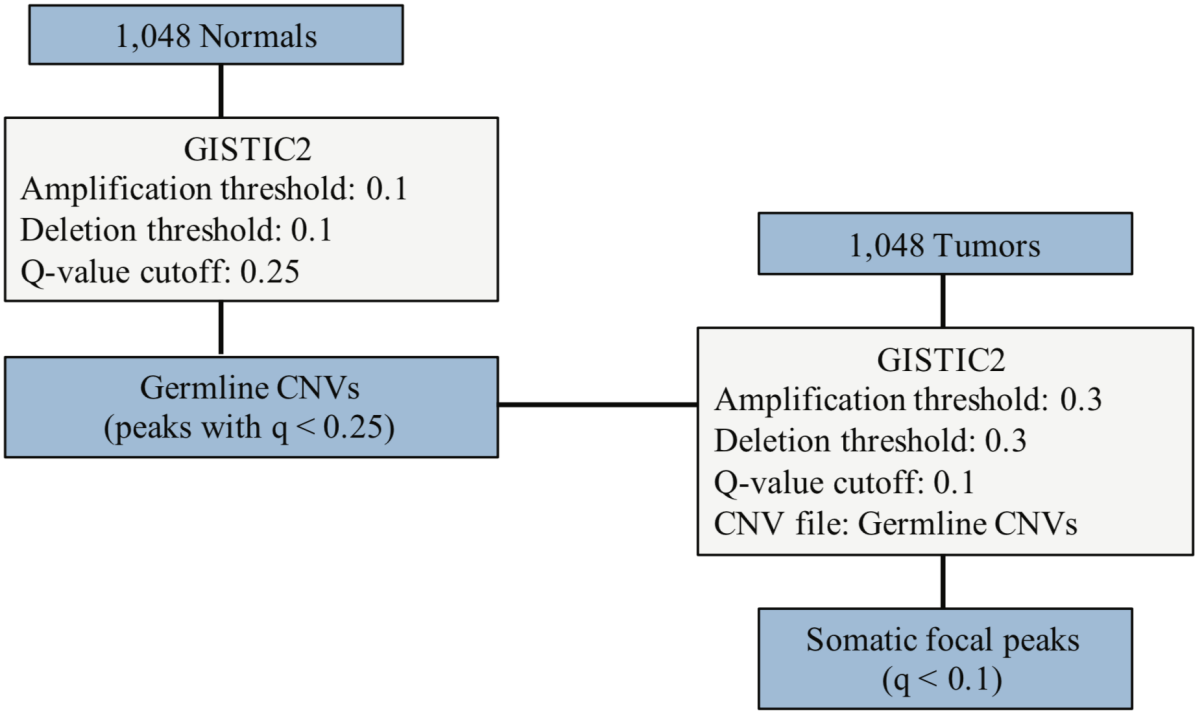
Supplementary Figure 2.19: SMGs in TWT Melanomas
CoMut plot of SMGs in the cohort of only cutaneous TWT melanomas.



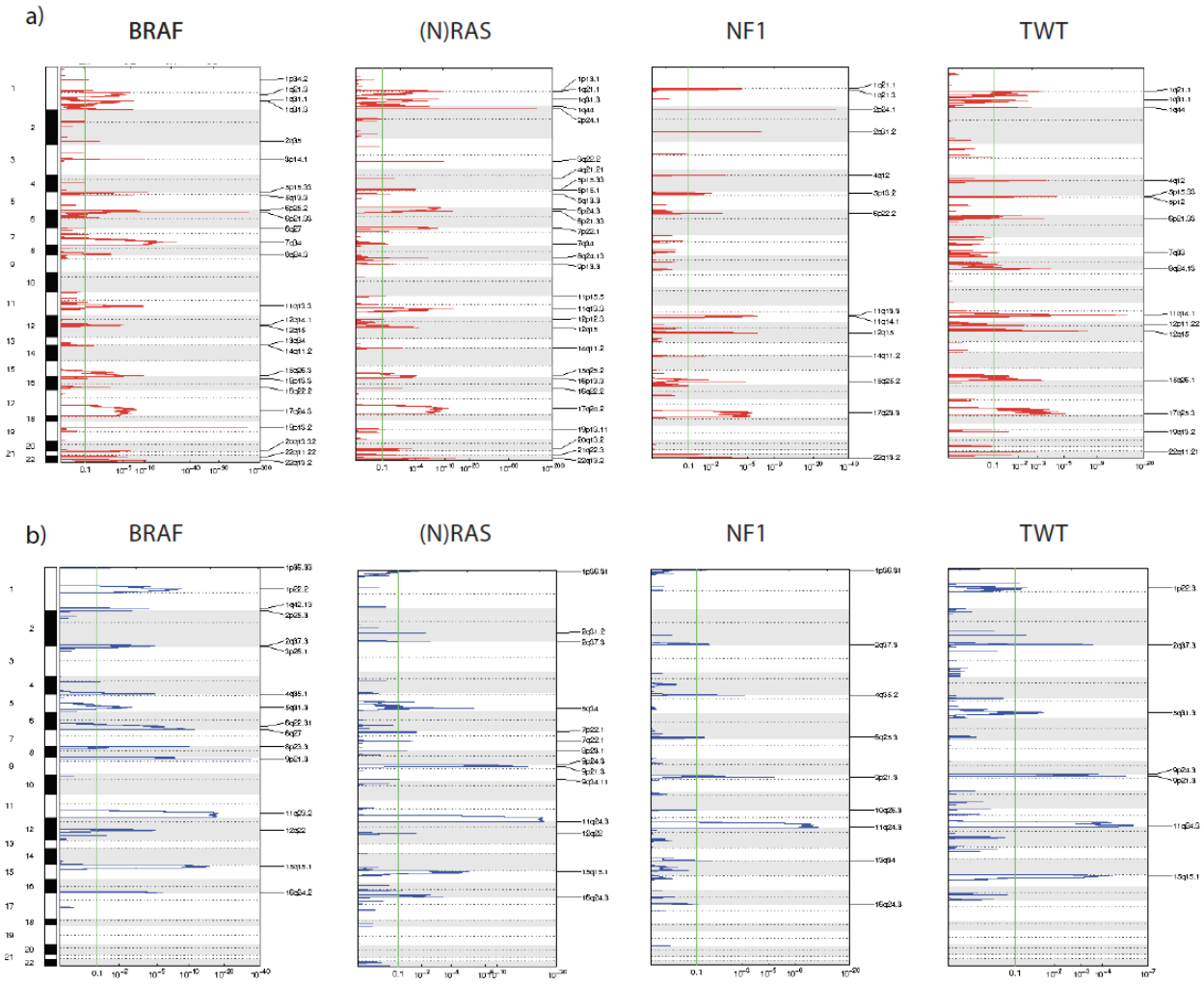
Supplementary Figure 2.20: Expression differences between mutant vs. wild-type, previously unknown cancer gene, TWT melanoma SMGs
Mean expression fold-change differences (in TCGA samples) between mutant vs. wild-type TWT melanoma SMGs that are not in the COSMIC Cancer Gene Census or OncoKB databases. *BRAF*, *NRAS*, and *NF1* are included as references for gain of function (GoF) and loss of function (LoF) mutations (purple names). Genes with a mean fold-change difference above 0 indicate higher expression in mutant samples compared to wild-type (i.e. GoF mutations). Genes with a mean fold-change difference below 0 indicate lower expression in mutant samples compared to wild-type (i.e. LoF mutations). Genes highlighted by a yellow point have a statistically significant difference in expression between mutant vs. wild-type tumors.



Supplementary Figure 2.21: CCFs of TWT SMGs
Density plots showing the distribution of CCFs for mutations in TWT SMGs. Some genes are almost always clonal (e.g. *RQCD1*, *GNA11*), while others are bimodal (e.g. *SF3B1*, *DDX59*) indicating those genes may be both clonal and subclonal drivers.

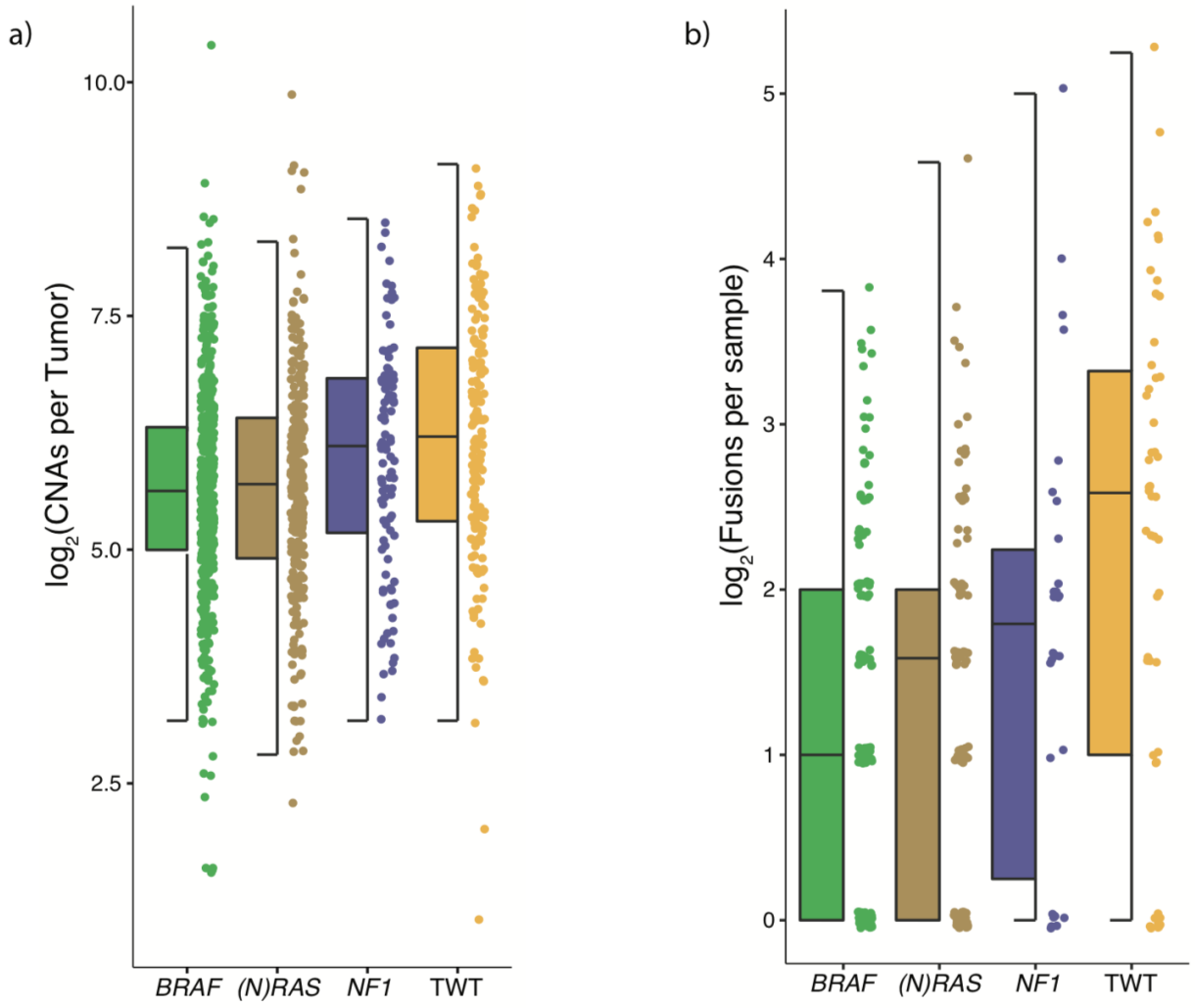


Supplementary Figure 2.22: GISTIC2 workflow for calling focal regions enriched in amplifications and deletions



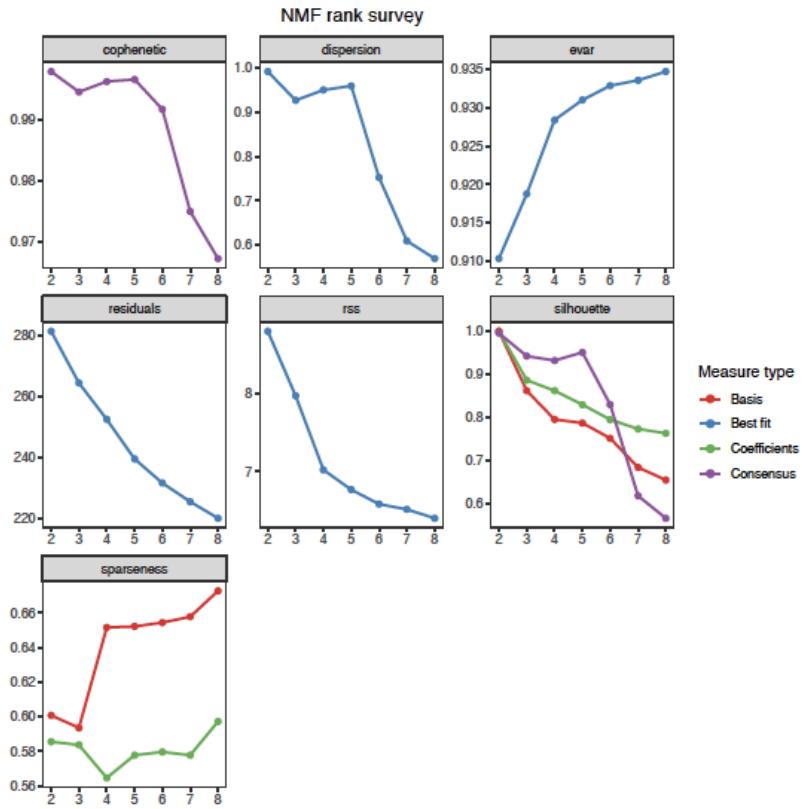
Supplementary Figure 2.23: GISTIC2.0 amplification and deletion peaks for each of the genomic subtypes

We used GISTIC2.0 to identify regions selectively targeted by somatic CNAs in each of the genomic subtypes. (a) A total of 26, 29, 14 and 16 significant focal amplification peaks, and (b) 16, 14, 9, and 7 significant focal deletion peaks, were identified in *BRAF*, *(N)RAS*, *NF1* and TWT melanomas, respectively (Benjamini-Hochberg, q-value cutoff < 0.1). Several of these peaks were in regions containing CGC and OncoKB genes (Supplementary Table 12).

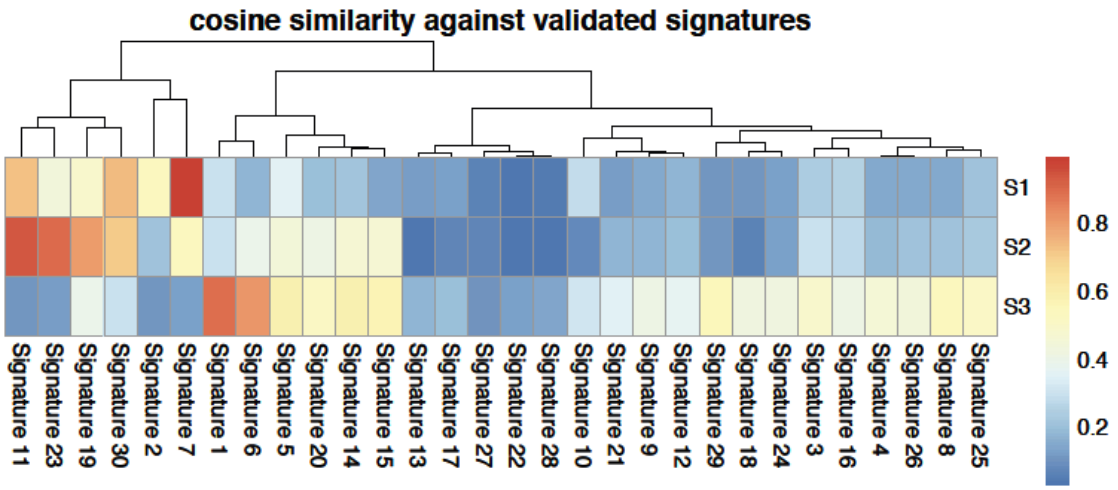


Supplementary Figure 2.24: CNA and fusion events per genomic subtype
(a) There was significant heterogeneity in the number of CNAs between the genomic subtypes (Kruskal-Wallis, $p = 7.78 \times 10^{-8}$, two-sided), ranging from 75 events in TWT melanomas to 47 events in *BRAF* melanomas. However, there was no difference in the proportion of the genome altered by CNA events between the subtypes (Kolmogorov-Smirnov, $p > 0.05$, two-sided). **(b)** The occurrence of gene fusions also differed significantly between the subtypes (Kruskal-Wallis, $p = 0.006$, two-sided), ranging from 6 fusion events in TWT melanomas to 2 fusion events in *BRAF* melanomas. The data are represented as boxplots where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points.

a)

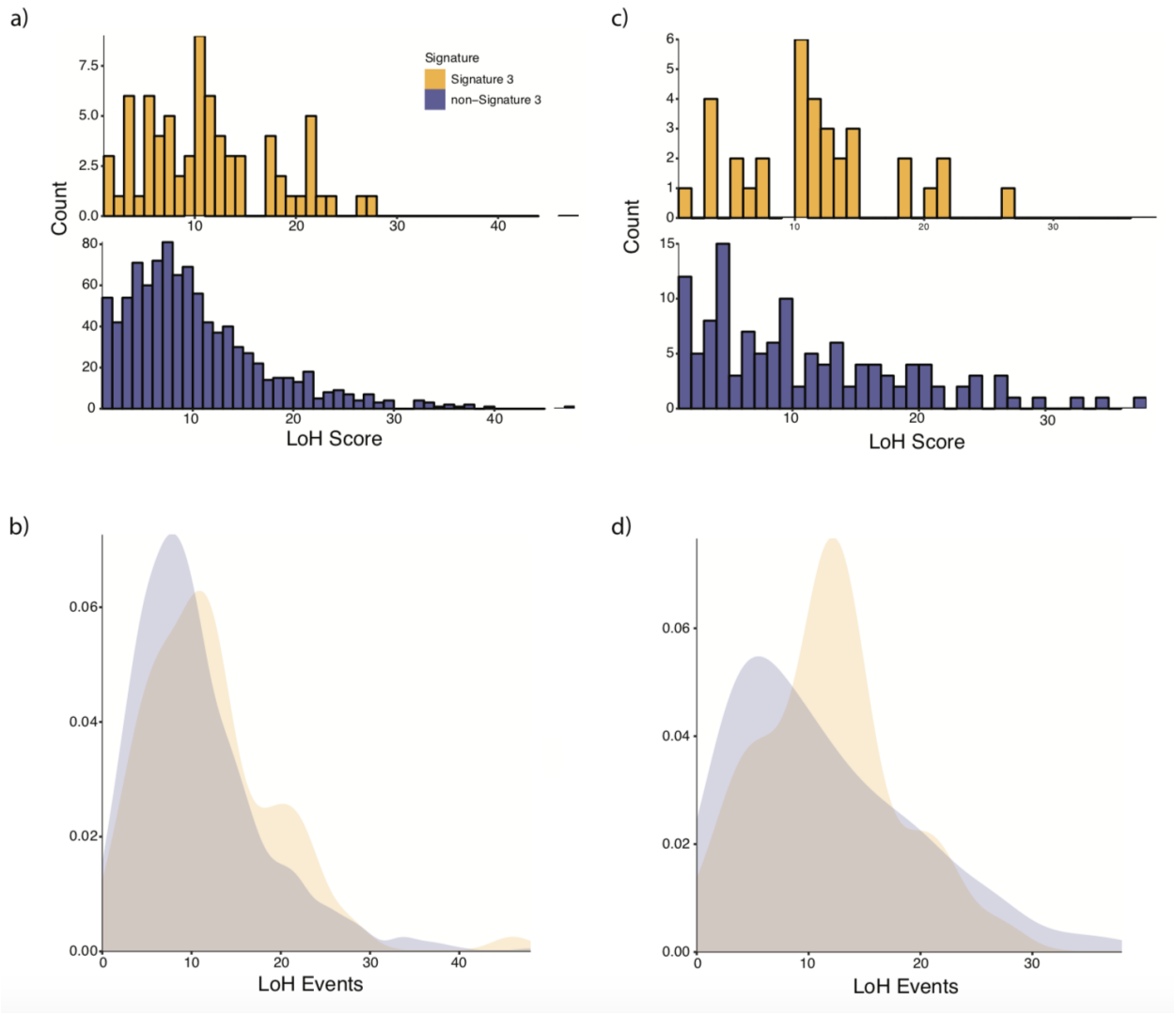


b)

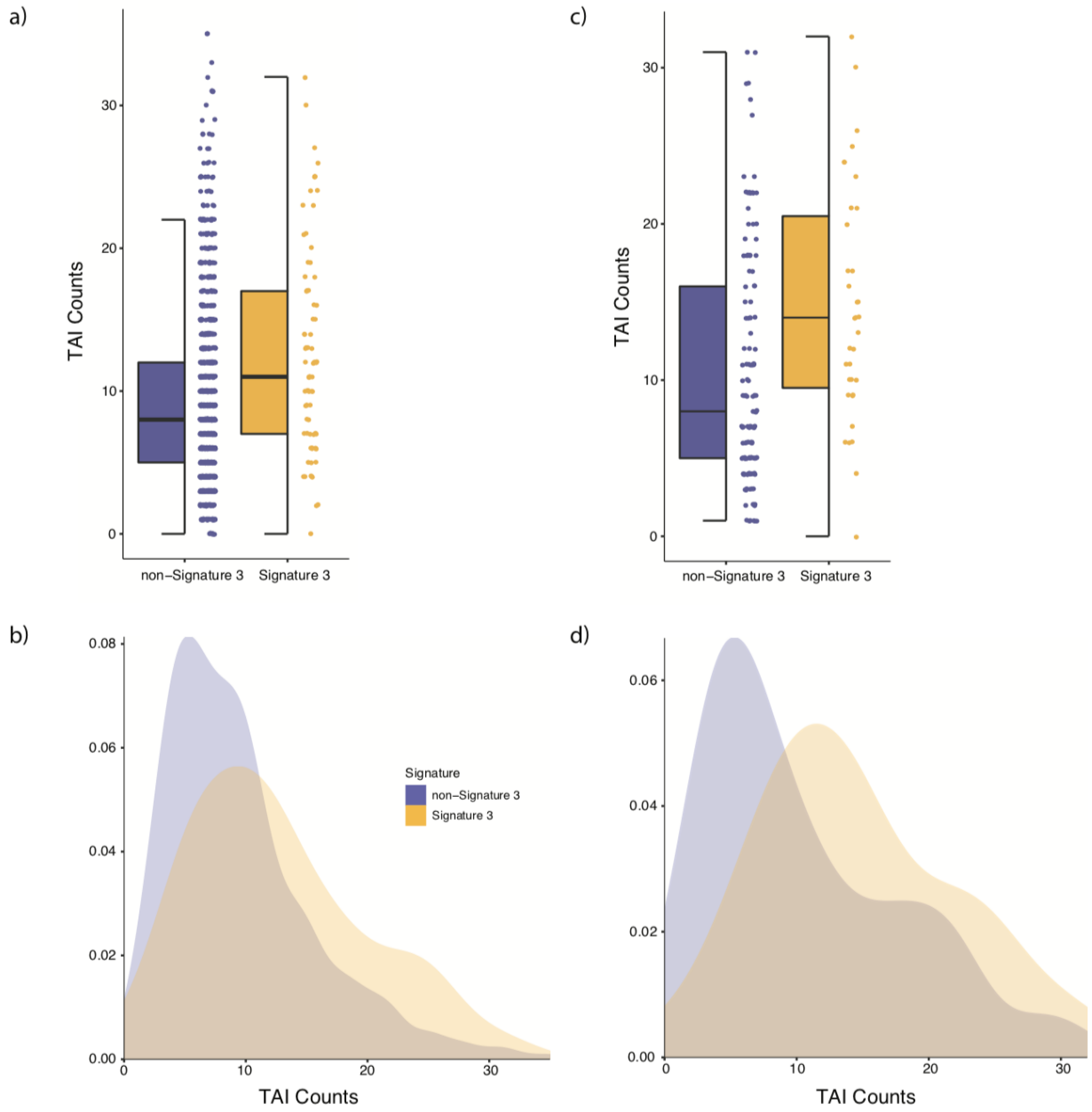


Supplementary Figure 2.25: NMF validation of deconstructSigs results on the entire cohort via SomaticSignatures

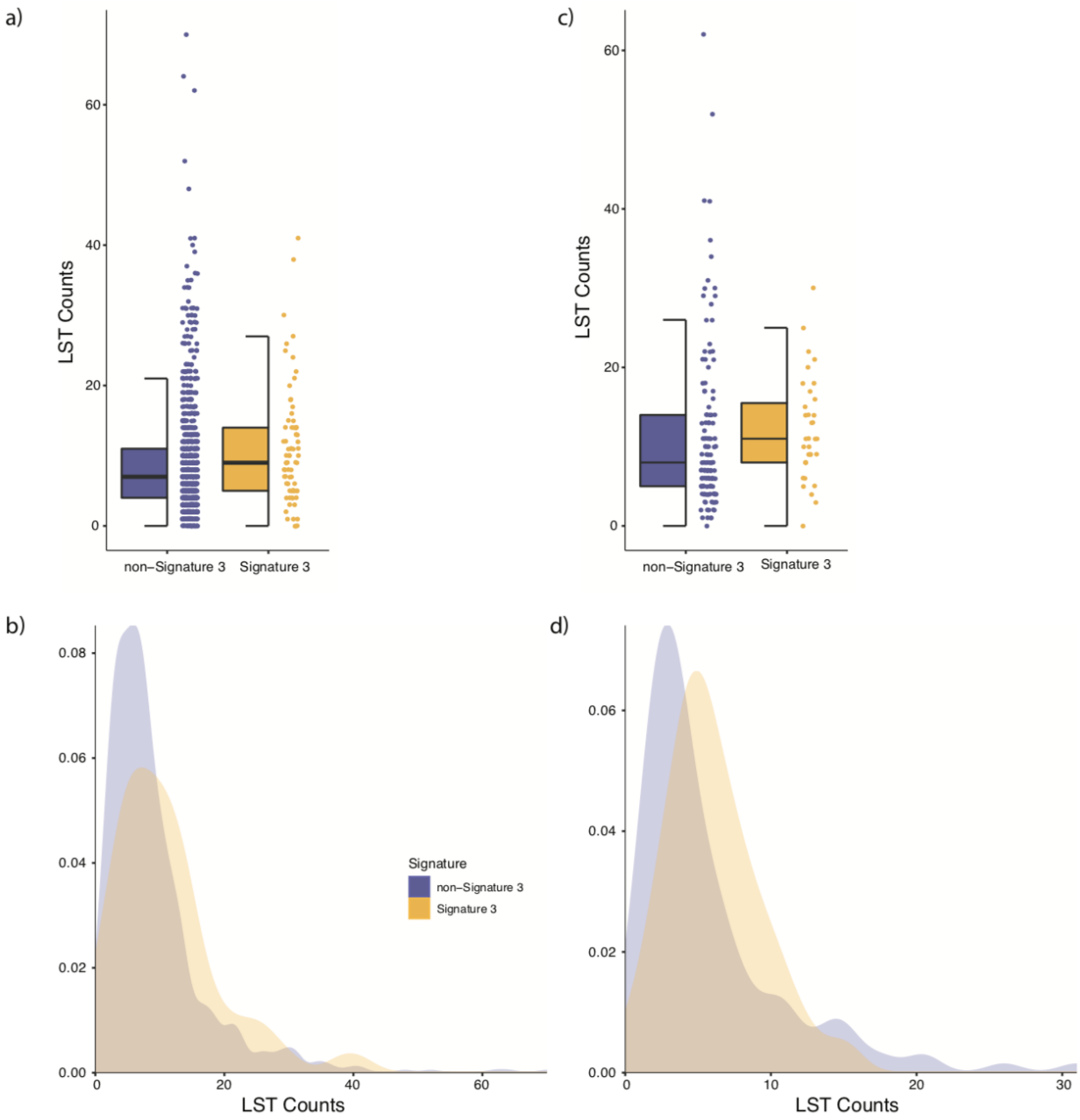
(a) NMF statistics for the entire cohort of melanomas. **(b)** Cosine similarity between COSMIC signatures and signatures decomposed via NMF for the entire cohort of melanomas. Signature S1 had the highest cosine similarity with signature 7 (UV exposure), signature S2 had the highest cosine similarity with signature 11 (exposure to alkylating agents), and signature S3 had the highest cosine similarity with signature 1 (spontaneous deamination of 5-methylcytosine).



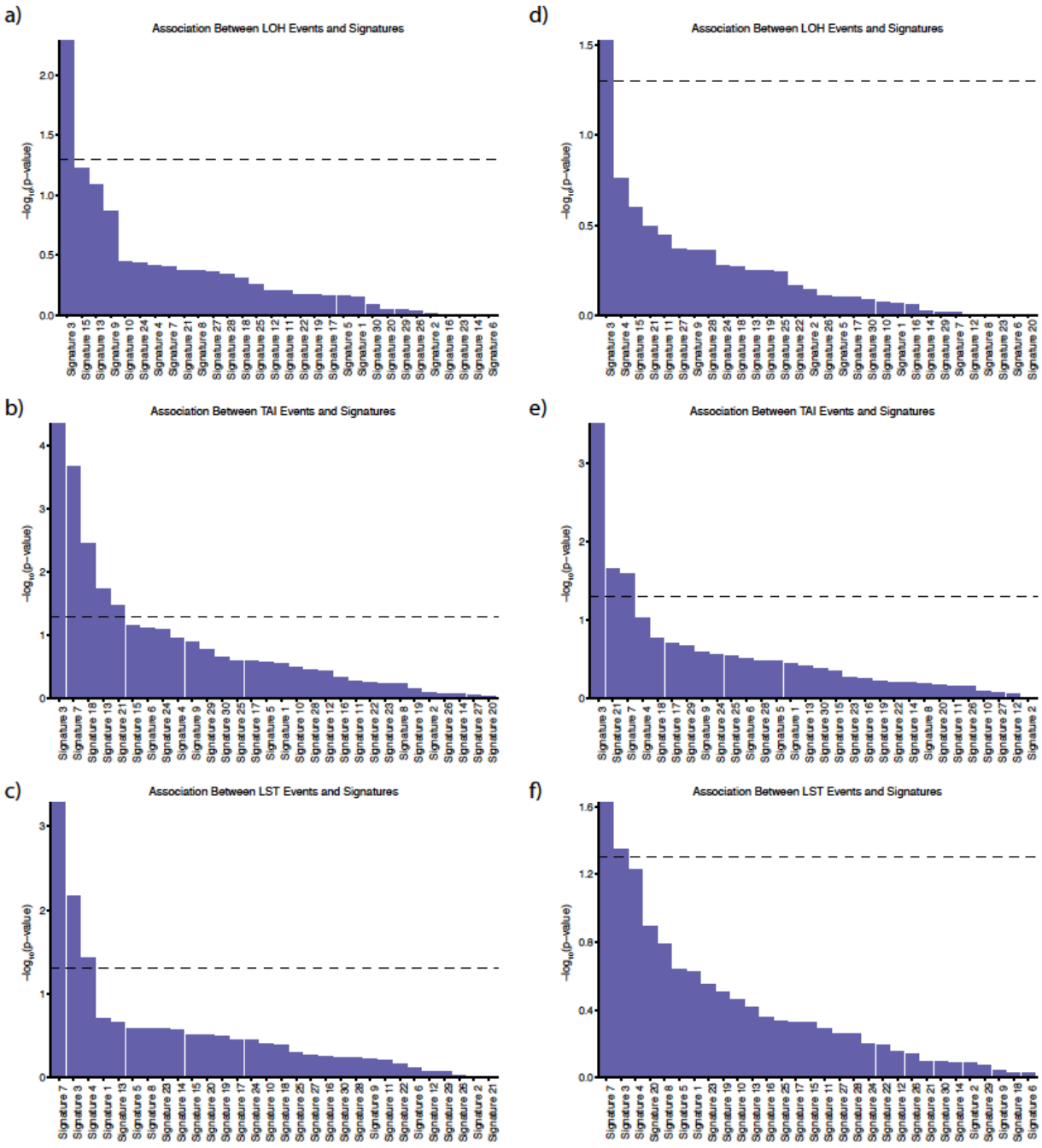
Supplementary Figure 2.26: DSB repair deficiency - loss of heterozygosity (LoH) score **(a)** Distribution of copy number LoH events in signature 3 (yellow) and non-signature 3 (purple) melanomas in the entire cohort. This satisfies the test used in Abkevich et al., 2012 and Timms et al., 2014, as the distribution was significantly different via the Kolmogorov-Smirnov test (p = 0.005, two-sided) and univariate logistic regression ($p = 5.34 \times 10^{-5}$). **(b)** Density plot of copy number LoH events in the entire cohort. **(c)** Distribution of copy number LoH events in signature 3 and non-signature 3 melanomas in TWT melanomas (p = 0.015, Kolmogorov-Smirnov, two-sided; p = 0.077, univariate logistic regression, two-sided). **(d)** Density plot of copy number LoH events in the TWT melanomas.



Supplementary Figure 2.27: DSB repair deficiency - allelic telomeric imbalance score
(a) Distribution of copy number TAI events in signature 3 (yellow) and non-signature 3 (purple) melanomas in the entire cohort. This satisfies the test used in Birkbak et al., as the distribution was significantly different via a Mann-Whitney U test ($p = 4.40 \times 10^{-5}$, two-sided). **(b)** Density plot of copy number TAI events in the entire cohort. **(c)** Distribution of copy number TAI events in signature 3 and non-signature 3 melanomas in TWT melanomas (Mann-Whitney U, $p = 1.8 \times 10^{-3}$, two-sided). **(d)** Density plot of copy number TAI events in the TWT melanomas. In **(a)** and **(c)** the data is represented as a boxplot where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points.

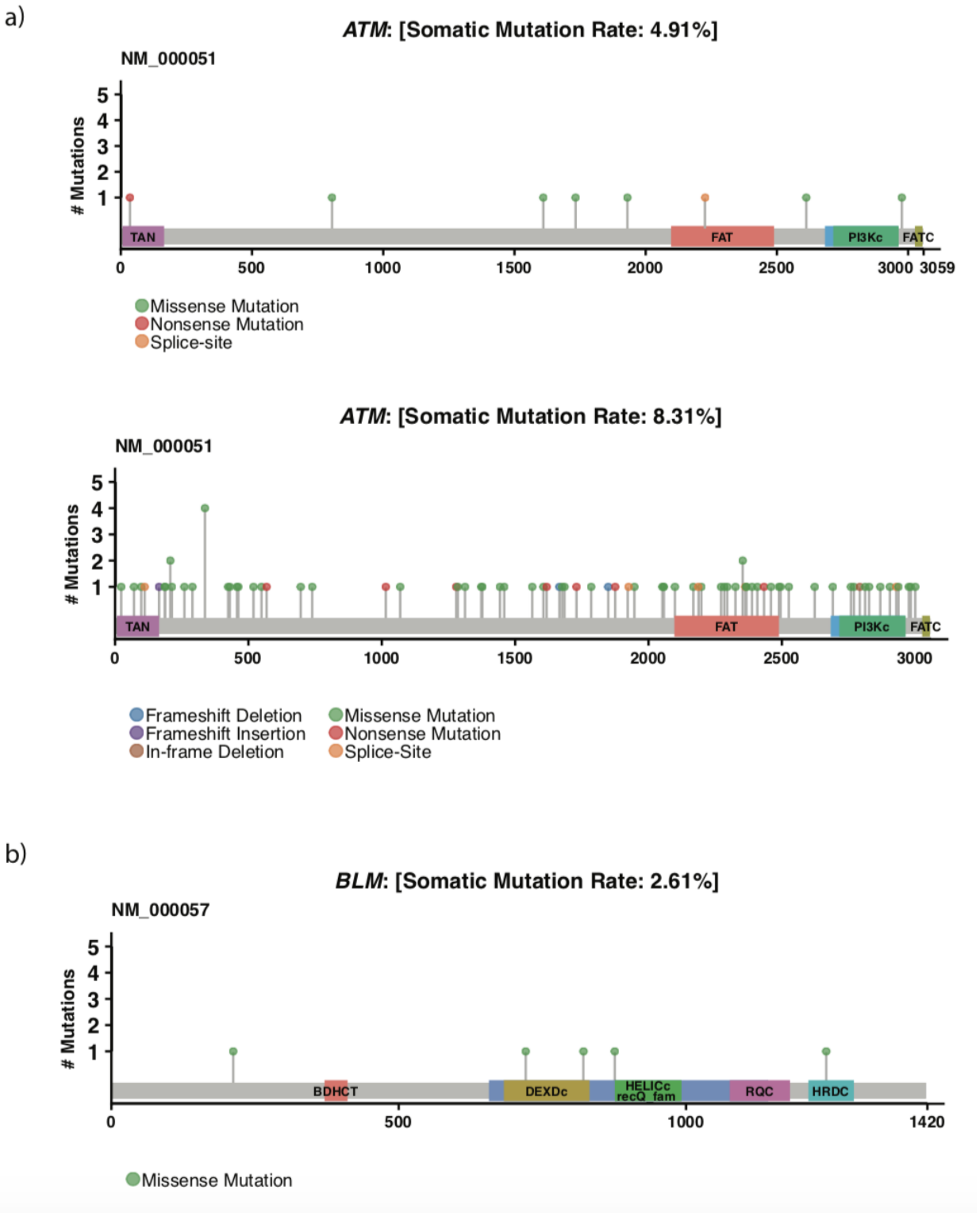


Supplementary Figure 2.28: DSB repair deficiency - large scale transitions
(a) Distribution of copy number LST events in signature 3 (yellow) and non-signature 3 (purple) melanomas in the entire cohort. This satisfies the test used in Popova et al., as the distribution was significantly different via a Mann-Whitney U test ($p = 6.82 \times 10^{-3}$, two-sided). **(b)** Density plot of copy number LST events in the entire cohort. **(c)** Distribution of copy number LST events in signature 3 and non-signature 3 melanomas in TWT melanomas (Mann-Whitney U, $p = 0.056$, two-sided). **(d)** Density plot of copy number LST events in the TWT melanomas. In **(a)** and **(c)** the data is represented as a boxplot where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points.



Supplementary Figure 2.29: Associations between mutational signatures and scarHRD scores

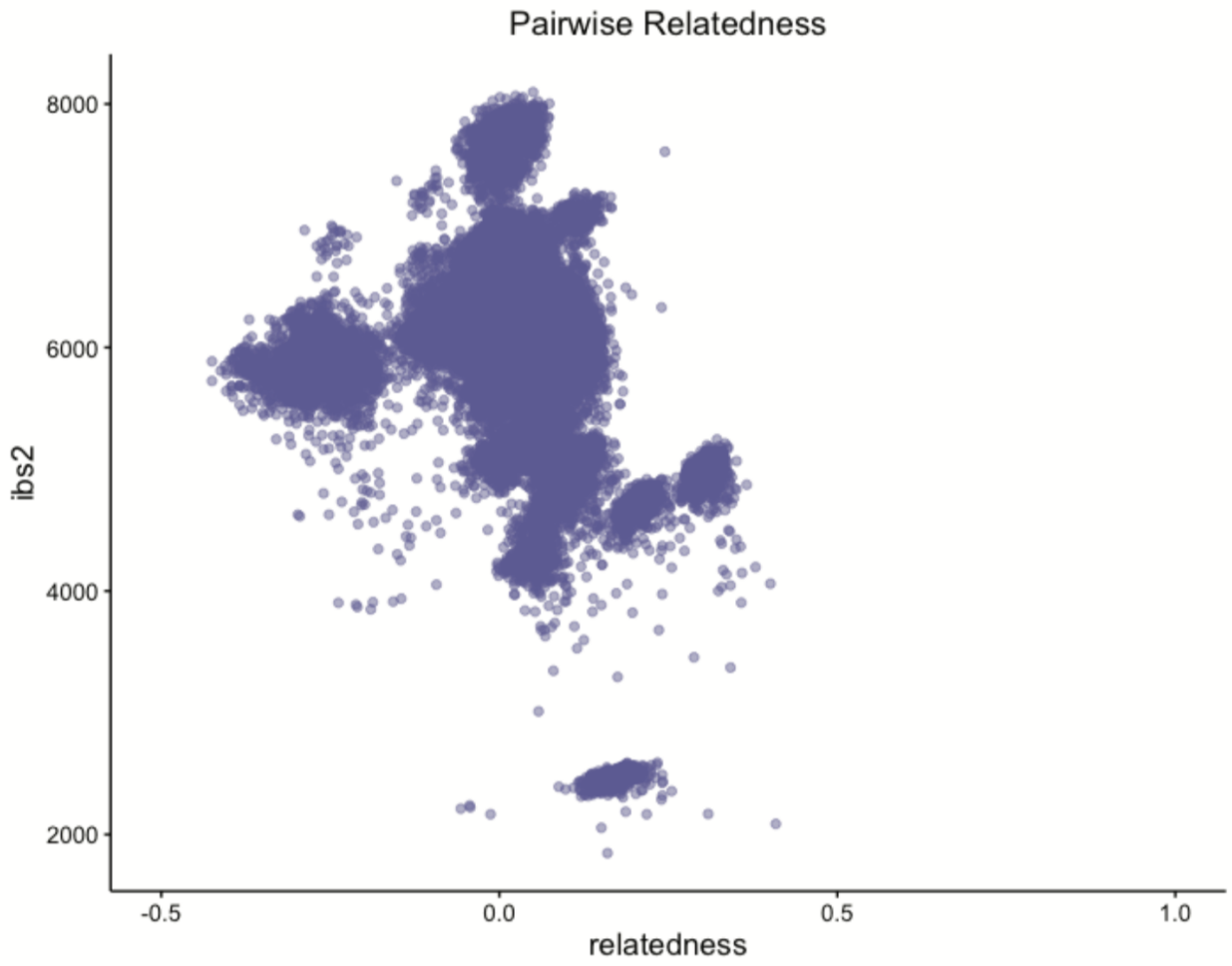
(a-c) Signature 3 was the only mutational signature to be associated with all three scarHRD copy number event scores (loss of heterozygosity, allelic telomeric imbalance, large scale transitions), and **(d-f)** this relationship still held when excluding acral and mucosal melanomas, which are enriched in copy number alterations compared to cutaneous melanomas. The dashed lines represent p-value cutoffs of 0.05.



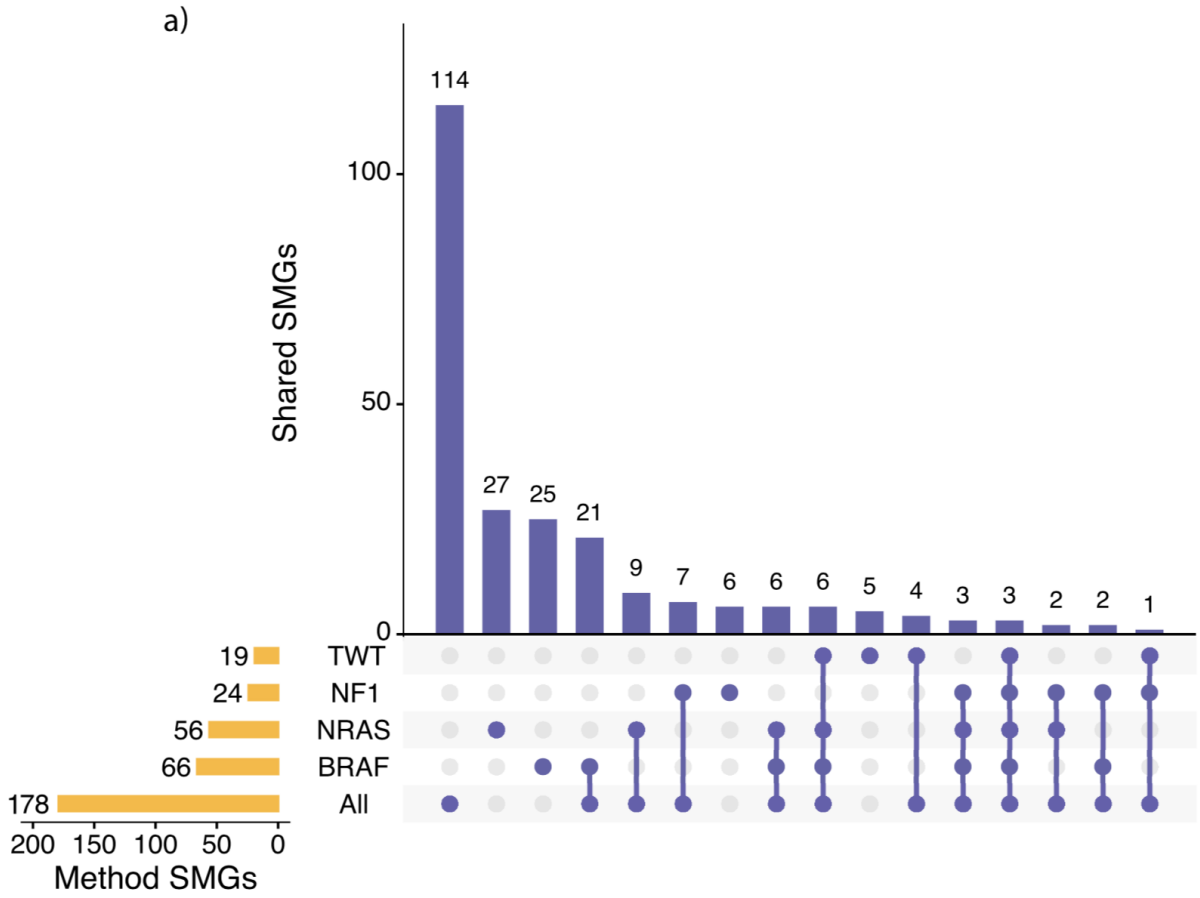
Supplementary Figure 2.30: Somatic alterations of interest in signature 3 TWT melanomas

(a) Top: Lollipop plot of somatic mutations in *ATM* for TWT melanomas. The splice-site variant in the FAT domain of *ATM* was exclusive to a TWT melanoma with signature 3. Bottom: Lollipop

Supplementary Figure 2.30 (continued): plot of somatic mutation in *ATM* for non-TWT melanomas. A signature 3 non-TWT melanoma also had a splice-site variant in the FAT domain of *ATM*. (b) Lollipop plot of somatic mutations in *BLM* for TWT melanomas. The missense mutation in the HRDC domain of *BLM* was exclusive to a TWT tumor with signature 3.

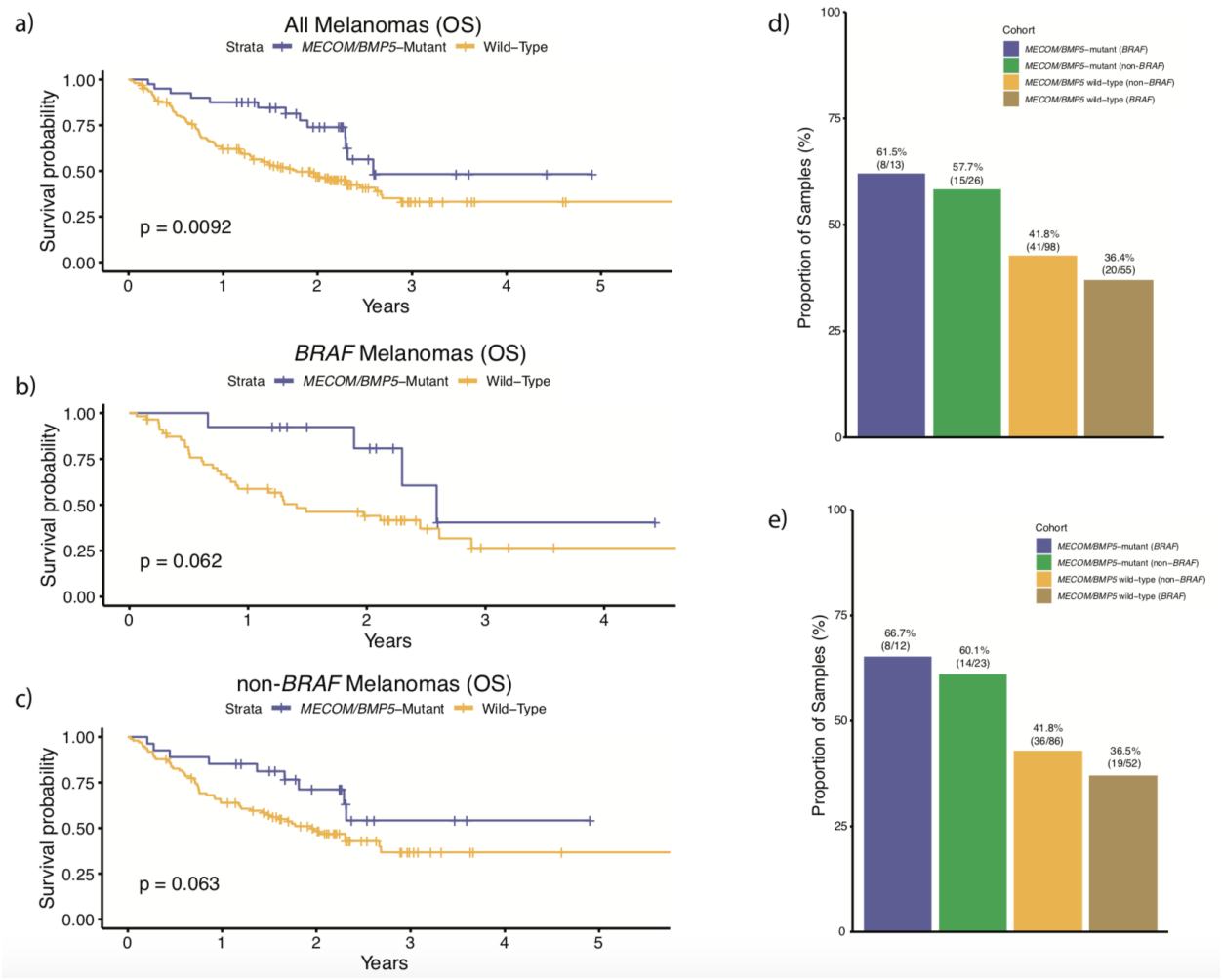


Supplementary Figure 2.31: Relatedness between normal samples
To prevent duplicate mutation calls from the same patient influencing our analyses, we used Somalier to determine the relatedness between normal samples in our cohort. Samples from the same patient would have a relatedness value very close to 1. Opacity was used to show the density of points.



Extended Data Figure 2.1: Overlap between SMGs from the entire cohort and subtype analyses

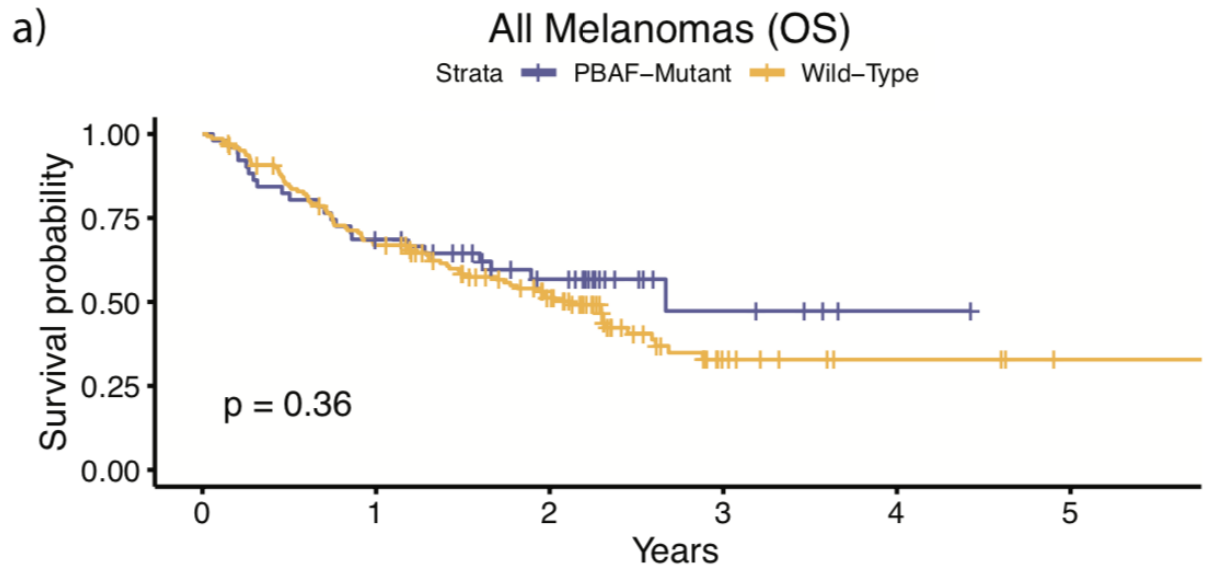
(a) Overlap between the subtype-specific SMGs and the SMGs that were identified via the entire cohort. Most of the SMGs identified in the entire cohort analysis were not identified through the subtype specific analysis (115 of 178, 64.6%).

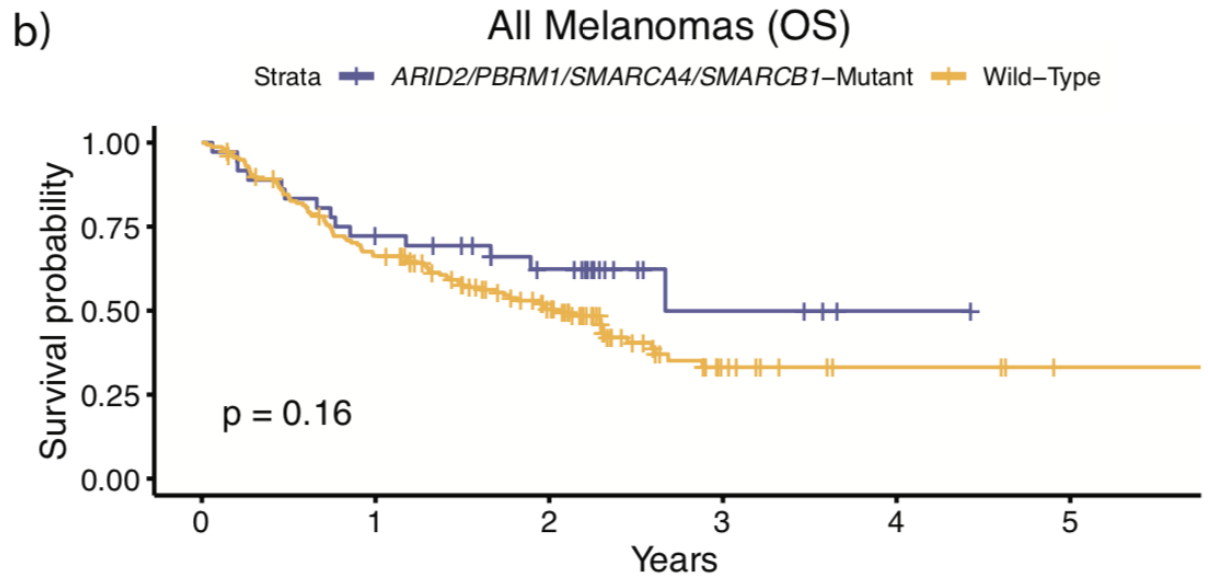


Extended Data Figure 2.2: *MECOM*/*BMP5* immunotherapy validation (overall survival and RECIST response)

External validation analysis of overall survival for *MECOM/BMP5* mutations using the Roh, Riaz, Hugo, and Rodig whole-exome cohorts (n = 194 total) for **(a)** all melanomas, **(b)** *BRAF* melanomas, and **(c)** non-*BRAF* melanomas, excluding post treatment biopsies. These cohorts were chosen because they were immunotherapy treated, whole-exome sequenced, cohorts not included in our discovery cohort. Due to the diverse treatment regimens in each of these trials and cohorts, we were unable to correct for drug. Further, since we did not have access to raw sequencing data from all these studies, we could not calculate and correct for tumor purity and utilized published variant calls. The hazard rate ratios of *MECOM/BMP5* mutations when correcting for only mutational load was **(a)** 0.59 (multivariate Cox proportional-hazards, p = 0.09) for all melanomas, **(b)** 0.46 (multivariate Cox proportional-hazards, p = 0.16) for *BRAF* melanomas, and **(c)** 0.68 (multivariate Cox proportional-hazards, p = 0.31) for non-*BRAF* melanomas. These results are similar to what was observed in the discovery cohort (Supplementary Table 2.8), although this validation cohort size was not powered to achieve statistical significance. **(d)** The association between the *BRAF* subtype and *MECOM/BMP5* mutations for clinical benefit to immunotherapy (via RECIST) in our limited validation cohort was

Extended Data Figure 2.2 (continued): similar to our discovery cohort findings, but not statistically significant. The p-values shown in a-c) are derived from the log-rank test.

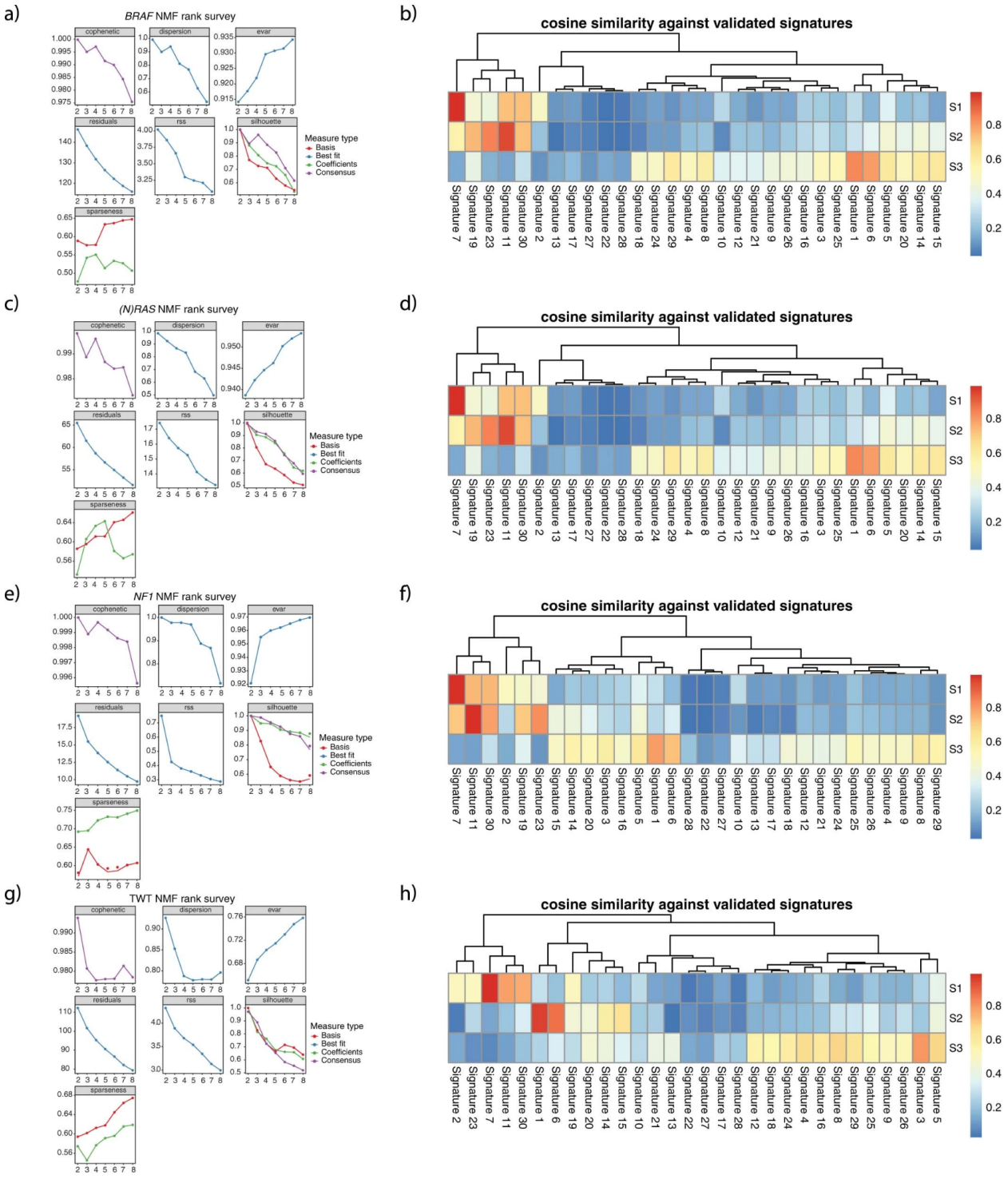




Extended Data Figure 2.3: PBAF complex immunotherapy validation (overall survival and RECIST response)

External validation analysis of overall survival for PBAF mutations using the Roh, Riaz, Hugo, and Rodig cohorts (n = 194), which are immunotherapy treated, whole-exome sequenced, cohorts not included in our discovery cohort. **(a)** Survival curves between PBAF-mutants and

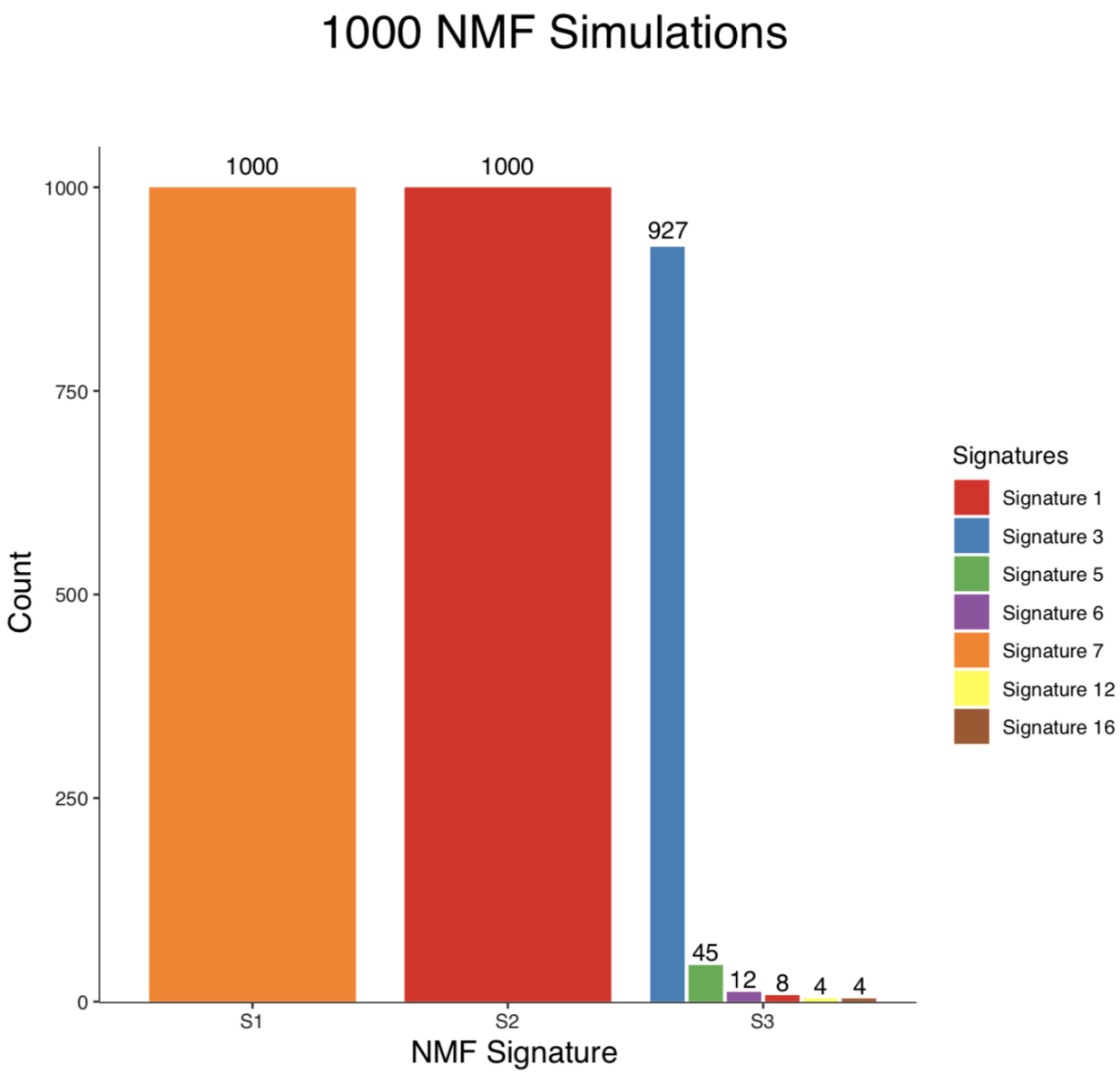
Extended Data Figure 2.3 (continued): non-PBAF mutants. **(b)** Survival curves between PBAF-mutants and non-PBAF mutants where PBAF mutants are classified by having mutations in *ARID2*, *PBRM1*, *SMARCA4*, and *SMARCB1*, which are the 4 PBAF complex genes commonly used in clinical sequencing panels. This limited validation cohort lacked sufficient samples with co-mutation of *(N)RAS* and PBAF complex genes (n = 9), and thus validation analysis was only performed on all tumors. Due to the unique treatment regimens in each of these cohorts, we were unable to correct for drug. Further, because we did not have access to raw sequencing data from these studies, we could not calculate and correct for tumor purity. When correcting only for mutational load the hazard ratio of PBAF mutations in the whole-exome cohorts, **(a)** when considering all genes in the PBAF complex, was 1.07 (multivariate Cox proportional-hazards, p = 0.80). The differences in these findings relative to the primary larger cohort may indicate differences in patient population and study size relative to our discovery cohort. **(b)** When considering only mutations in *ARID2*, *PBRM1*, *SMARCA4*, and *SMARCB1* as PBAF-mutant, the HRR was 0.86 (multivariate Cox proportional-hazards, p = 0.61). The p-values for **(a-b)** are derived from the log-rank test.



Extended Data Figure 2.4: NMF validation of deconstructSigs results on genomic subtypes via SomaticSignatures

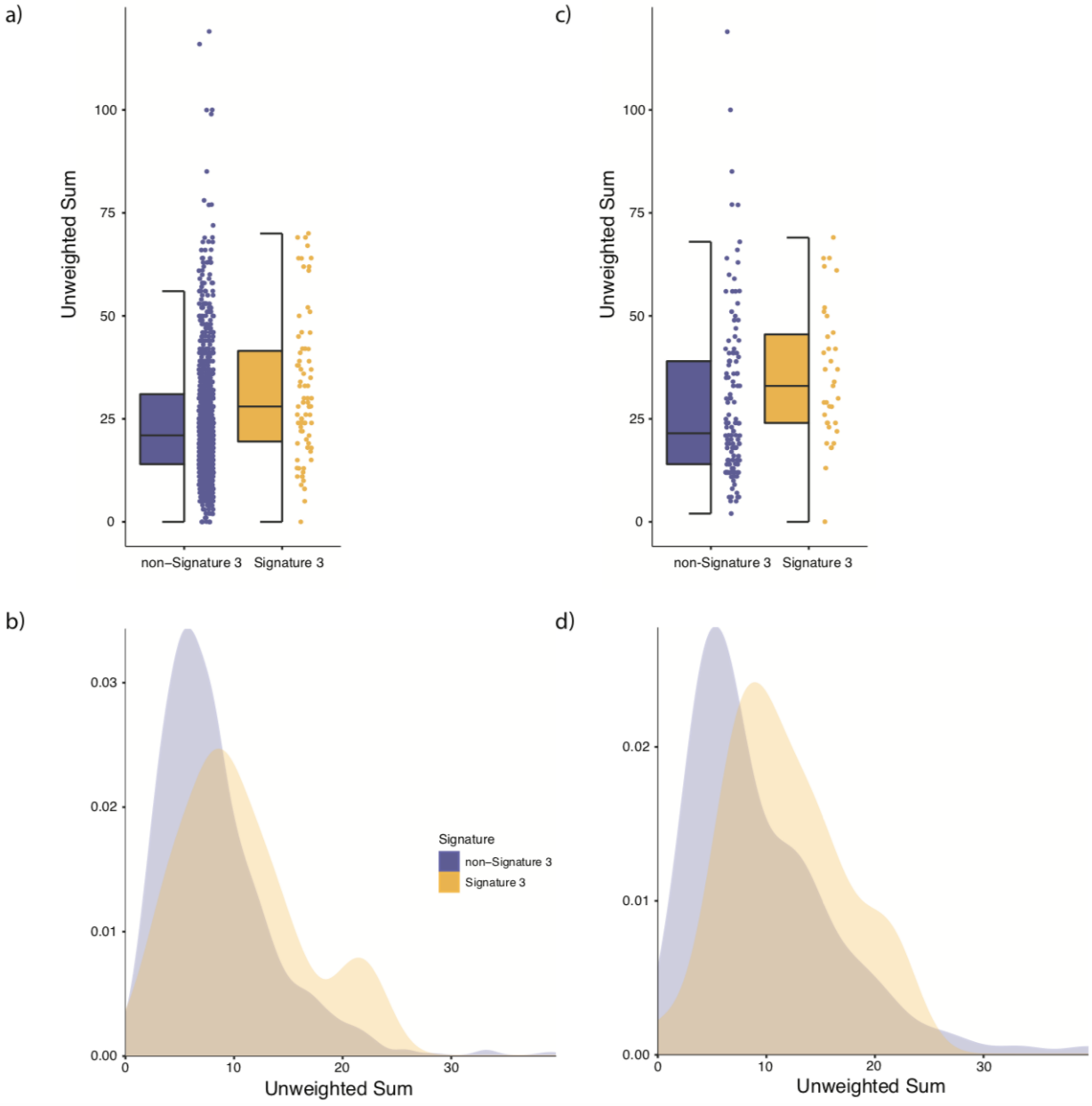
(a) NMF statistics for *BRAF* melanomas. **(b)** Cosine similarity between COSMIC signatures and signatures decomposed via NMF for *BRAF* melanomas. **(c)** NMF statistics for *(N)RAS* melanomas. **(d)** Cosine similarity between COSMIC signatures and signatures decomposed via NMF for *(N)RAS* melanomas. **(e)** NMF statistics for *NF1* melanomas. **(f)** Cosine similarity

Extended Data Figure 2.4 (continued): between COSMIC signatures and signatures decomposed via NMF for *NF1* melanomas. **(g)** NMF statistics for TWT melanomas. **(h)** Cosine similarity between COSMIC signatures and signatures decomposed via NMF for TWT melanomas. The cophenetic correlation coefficient and residual sum of squares (RSS) suggests 3 is the optimal number of signatures for each genomic subtype.



Extended Data Figure 2.5: NMF simulations via SomaticSignature on TWT melanomas removing 35 random non-signature 3 samples each simulation
A total of 35 signature 3 samples were identified via deconstructSigs in our signature analysis. To ensure that our NMF validation in TWT melanomas (Supplementary Fig. 17) is actually identifying signature 3 because it is indeed present, and not because it's a flat signature, we performed 1000 simulations removing 35 random non-signature 3 samples each time. Signature

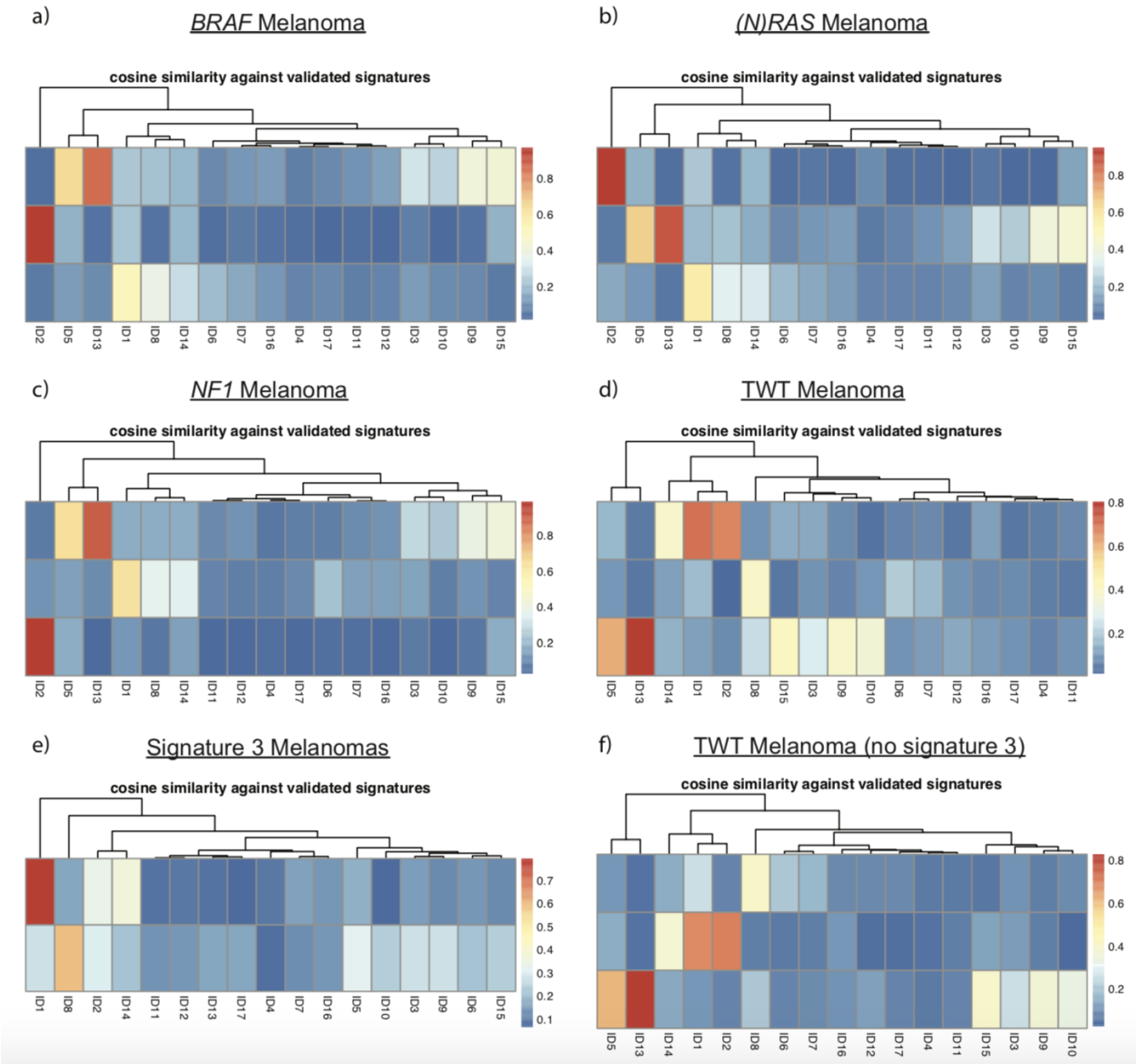
Extended Data Figure 2.5 (continued): 3 was identified 927 times (92.7%), which corroborates the deconstructSigs results and suggests signature 3 is the third most dominant signature in TWT melanomas. Performing 1000 simulations when removing the 35 signature 3 samples each time never yielded the identification of signature 3 via NMF.



Extended Data Figure 2.6: DSB repair deficiency - unweighted sum of HRD associated events

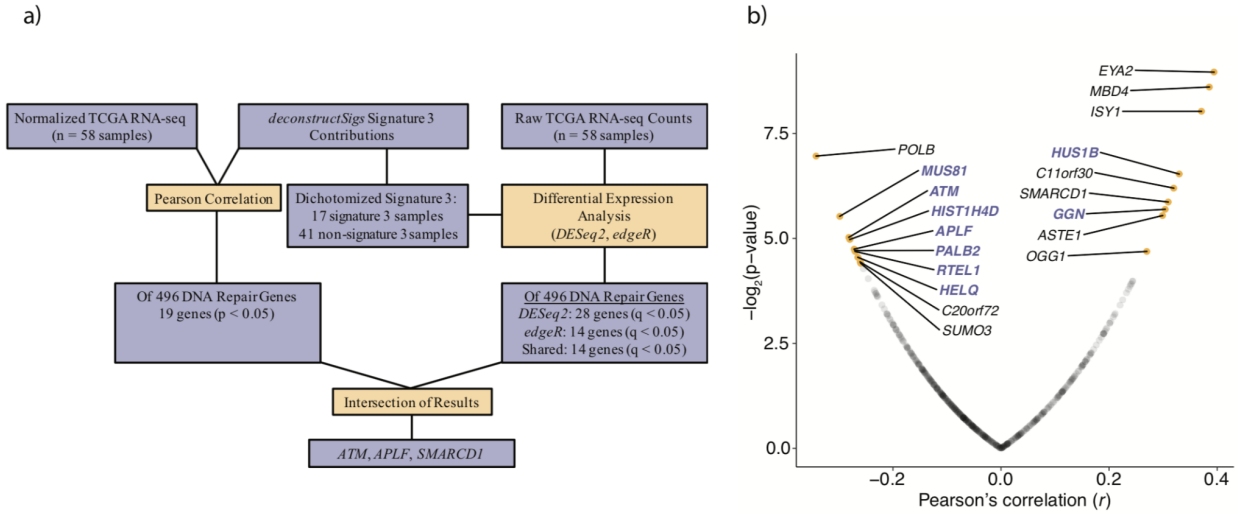
(a) Distribution of the unweighted sum of HRD associated CNA events (loss of heterozygosity, telomeric allelic imbalance, large scale transitions) in signature 3 (yellow) and non-signature 3 (purple) melanomas in the entire cohort. Signature 3 tumors were significantly enriched in HRD

Extended Data Figure 2.6 (continued): associated copy number events via a Mann-Whitney U test ($p = 6.21 \times 10^{-5,}$ two-sided). **(b)** Density plot of HRD associated copy number events in the entire cohort. **(c)** Distribution of HRD associated copy number events in signature 3 and non-signature 3 melanomas in TWT melanomas (Mann-Whitney U, $p = 5.49 \times 10^{-3}$, two-sided). **(d)** Density plot of HRD associated copy number events in the TWT melanomas. In **(a)** and **(c)** the data is represented as a boxplot where the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent the interquartile range (IQR) multiplied by 1.5, and beyond the whiskers are outlier points.



Extended Data Figure 2.7: Indel mutational signatures on the 390 WGS tumors
Cosine similarity between COSMIC indel mutational signatures and the suggested solution NMF results from SigProfileExtractor. Indel mutational signatures revealed that **(a)** *BRAF*, **(b)** *(N)RAS*, and **(c)** *NF1* melanomas were associated with indel signatures ID1, ID2 and ID13 **Extended**

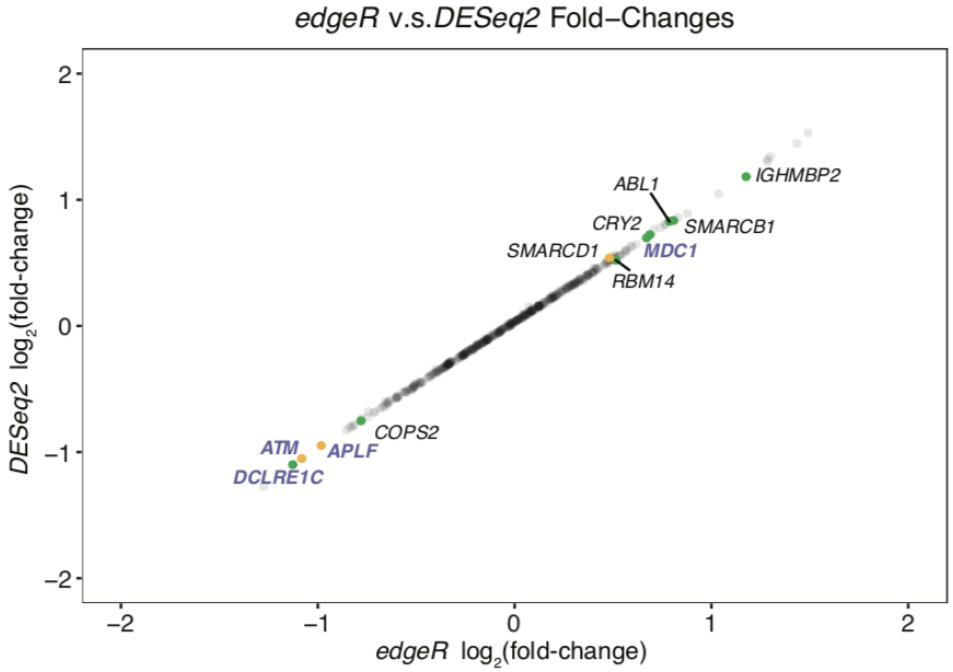
Data Figure 2.7 (continued): (associated with UV), while **(d)** TWT melanomas were associated with indel signatures ID1, ID8 (associated with NHEJ), and ID13. **(e)** Mutational signature 3 was associated with indel signatures ID1 and ID8, and was the sole mutational signature associated with ID8. **(f)** Interestingly, when removing signature 3 tumors from the TWT melanoma cohort, TWT melanomas were still associated with indel signature ID8. Thus, the increased genomic instability of TWT melanomas in general is enough to result in ID8.



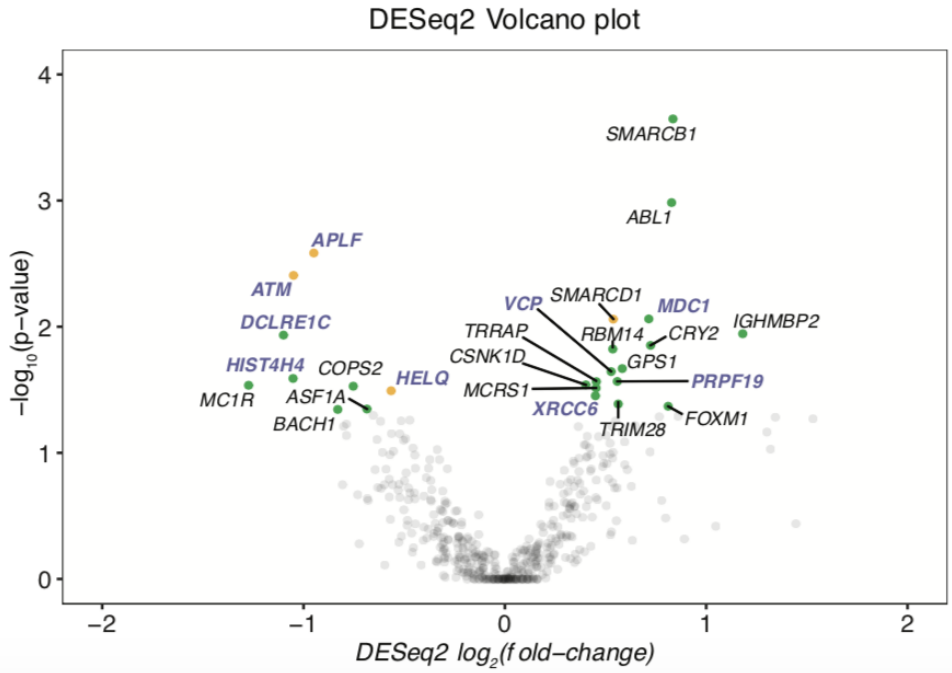
Extended Data Figure 2.8: Comparison of transcriptional profiles between DSB repair deficient and DSB repair intact TWT melanomas

(a) The workflow used to identify transcriptional differences between putative DSB repair deficient (presence of signature 3) and non-DSB repair deficient (no contribution of signature 3) TWT tumors. **(b)** Pearson correlation between signature 3 contribution and normalized gene expression in TWT melanomas (Methods) identified 9 positive and 10 negative signifi- cant correlations for DNA-repair genes (Pearson's, p-value cutoff < 0.05; Methods). Genes highlighted in purple function in DSB repair pathways, including HR. Opacity was used to show the density of non-significant points along both axes.

a)



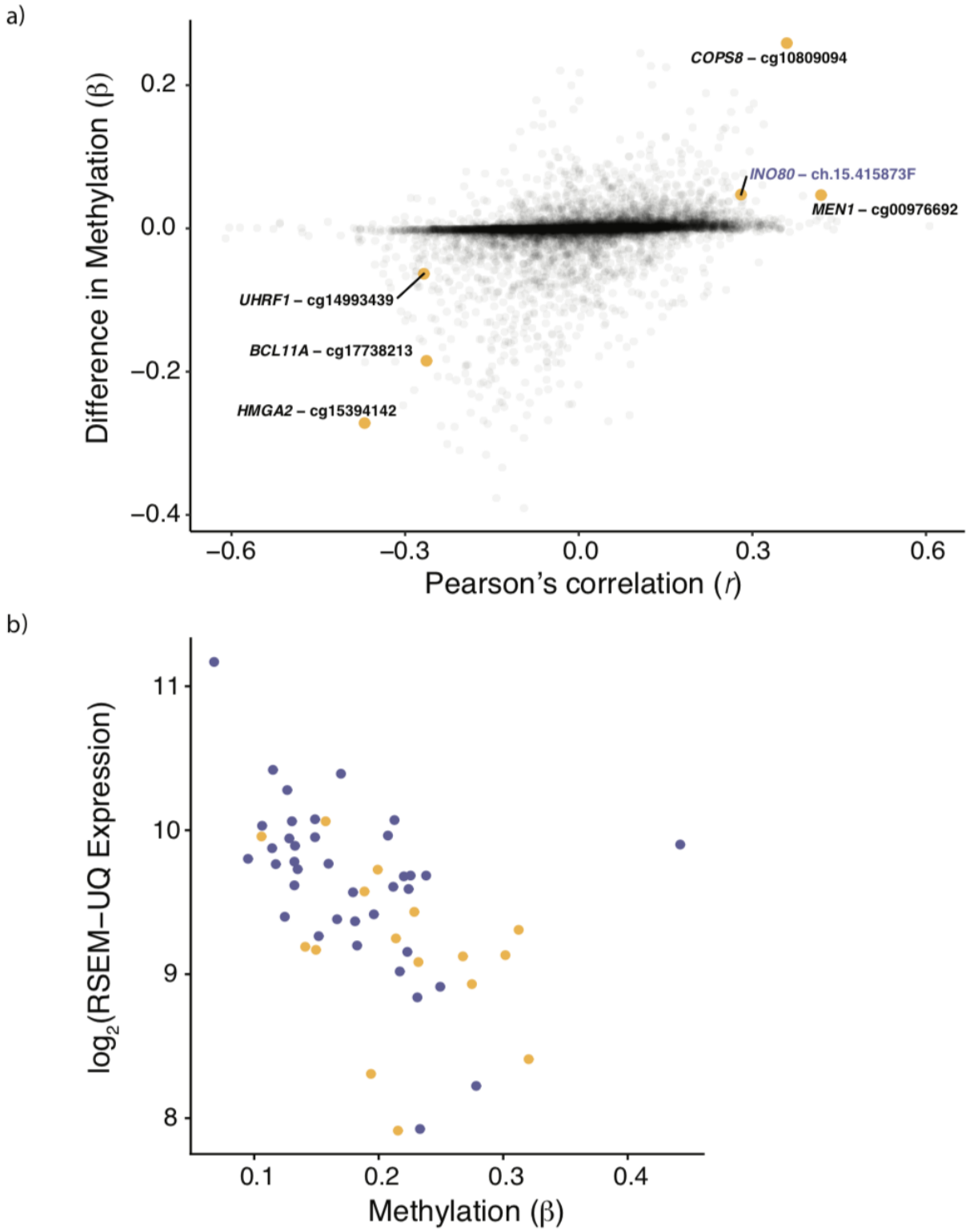
b)



Extended Data Figure 2.9: Differential expression analysis between signature 3 and non-signature 3 TWT melanomas

(a) DESeq2 log2 fold-change vs edgeR log2 fold-change for cumulative set of DNA-repair genes. **(b)** Significance vs log2 fold-change of significantly differentially expressed DNA repair genes as determined by *DESeq2*. Yellow points indicate genes whose expression was significantly correlated with signature 3 contribution and significantly differentially expressed.

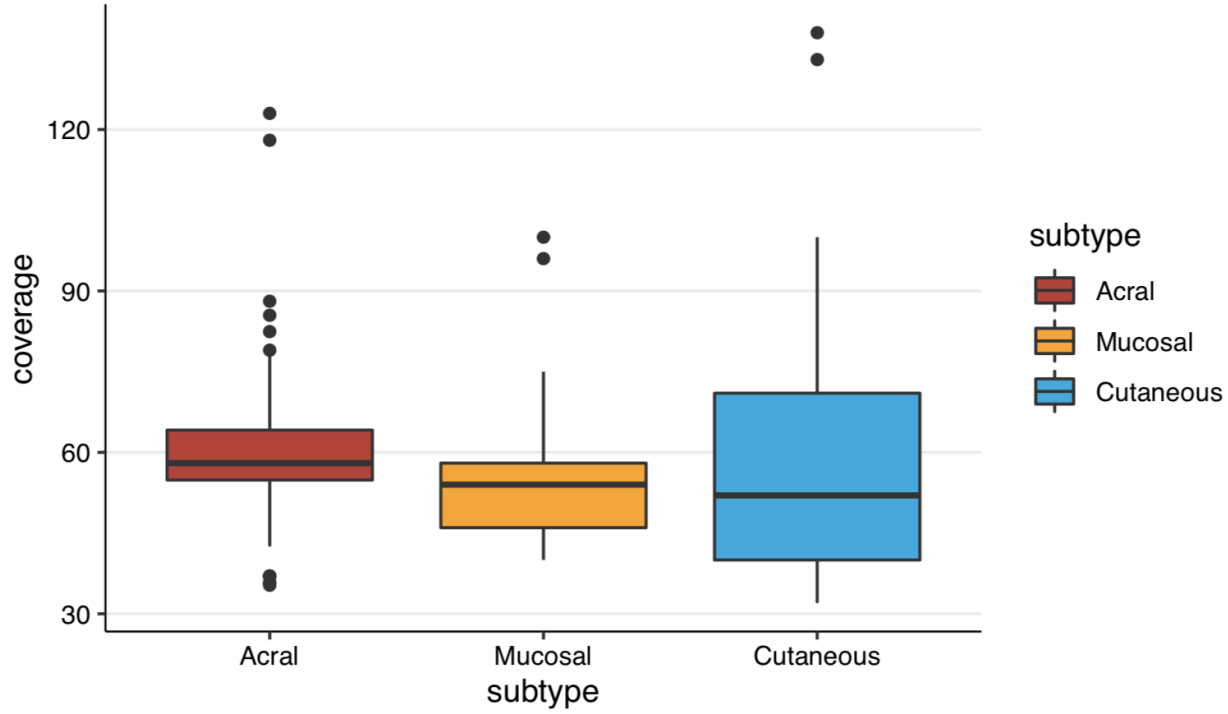
Extended Data Figure 2.9 (continued): Green points indicate genes that were only significantly differentially expressed. Genes highlighted in purple function in DSB repair. Opacity was used to show the density of non-significant points along both axes.



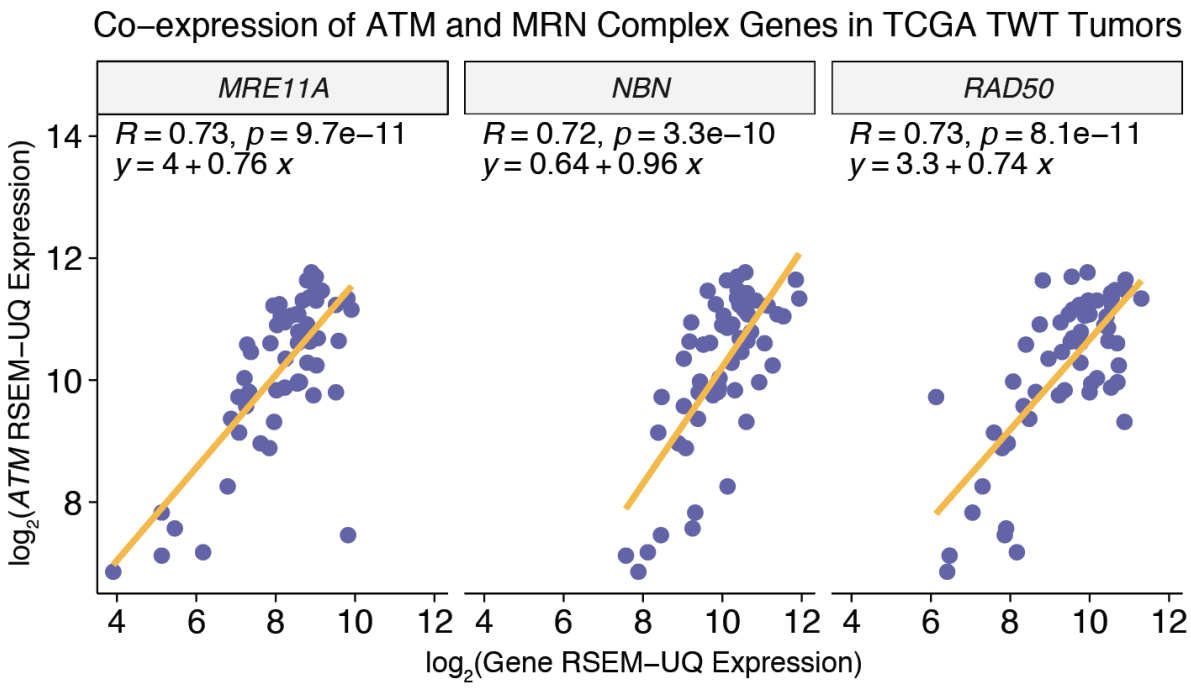
Extended Data Figure 2.10: Methylation and signature 3 contribution

(a) Pearson correlation between signature 3 contribution and methylation β-values plotted on the x-axis vs. difference in median methylation between signature 3 and non-signature 3 TWT samples on the y-axis. Six probe sites were significantly correlated with signature 3 contribution,

Extended Data Figure 2.10 (continued): had a significant difference in median β-values (via Mann-Whitney U), and had methylation β-values significantly associated with gene expression. Of the six probe sites, *INO80* was the only gene involved in HR repair. Opacity was used to show the density of non-significant points along both axes. **(b)** Expression of *INO80* was significantly correlated with methylation β-values at *INO80*-ch.15.415873F (Pearson's, $r = -0.51$, $p = 8.516 \times 10^{-5}$). Points in yellow are from signature 3 TWT samples and points in purple are from non-signature 3 TWT samples.

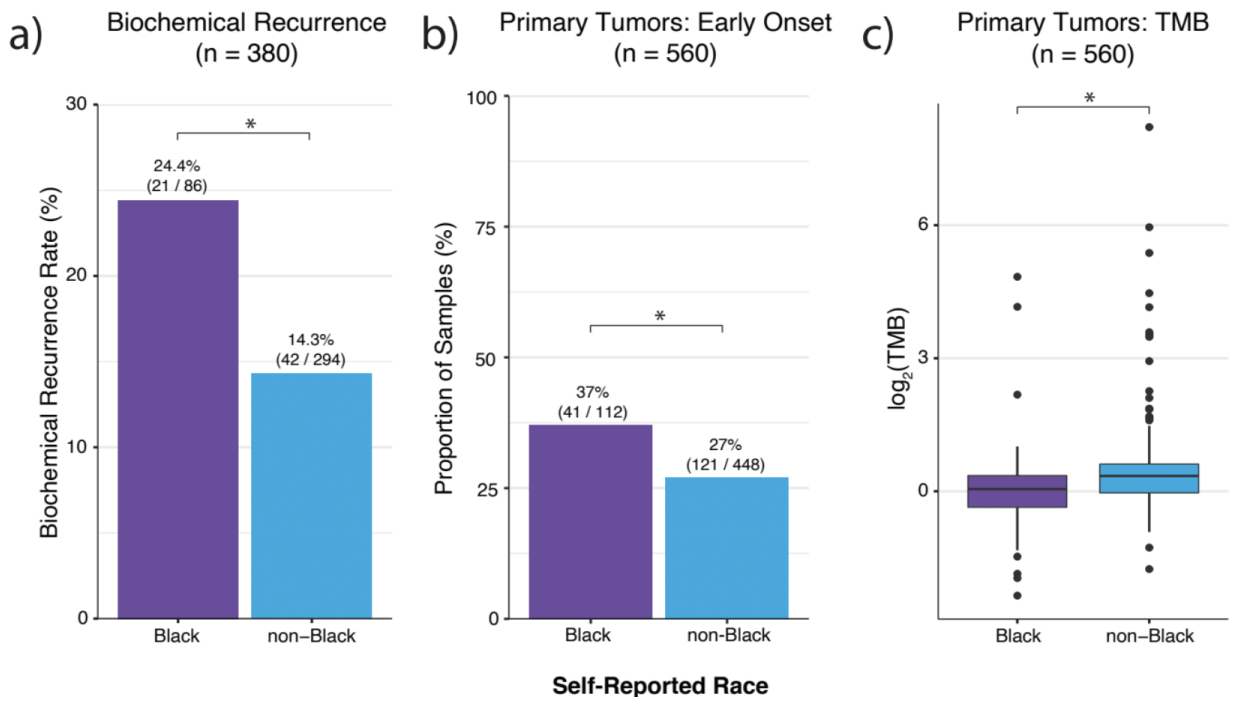


Supplementary Figure 3.1: Tumor sample coverage by melanoma histological subtype The median sequencing coverage of tumor samples in general is 57X, and there is no statistical difference in tumor sample coverage between the histologies (Wilcoxon-Mann-Whitney; pairwise; p = 0.08).



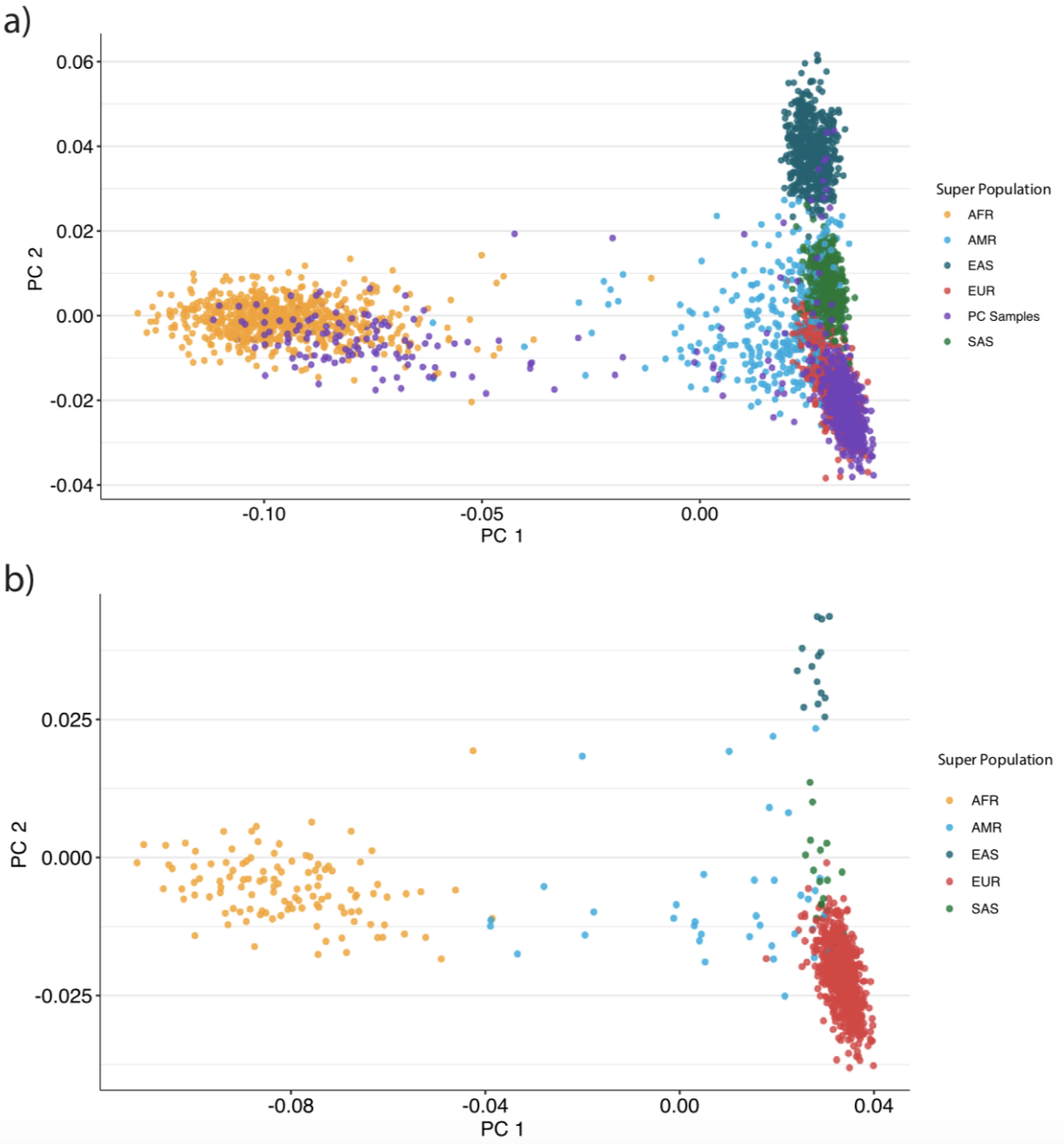
Supplementary Figure 3.2: Association between MRN complex gene expression and *ATM* expression

Although SVs affecting *RAD50* were not associated with signature 3 tumors (Fisher's exact, p > 0.05), there was no difference in the association between *MRE11A* or *NBN* expression and *ATM* expression compared to the association between *RAD50* and *ATM* expression in TWT tumors (Fisher's Z-transformation; pairwise; p > 0.05).

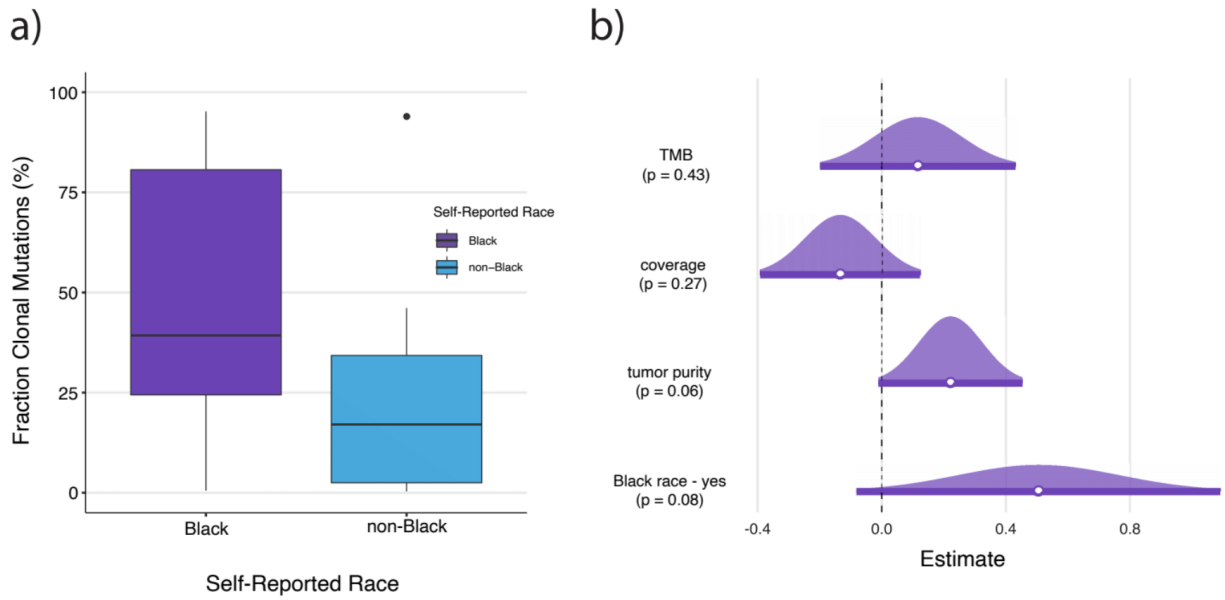


Supplementary Figure 4.1: Clinical and Genomic Characteristics associated with self-reported race in PC

a) The rate of biochemical recurrence between tumors from Black and non-Black patients with information on biochemical recurrence status (n=380). Consistent with previous studies, the rate of biochemical recurrence was significantly higher in Black compared to non-Black patient samples (Fisher's exact test, p = 0.03). **b)** The proportion of early onset (<= 55 years old) tumors in our cohort between Black and non-Black patient tumors. Consistent with previous studies, Black patient samples were associated with a higher rate of early onset tumors (Fisher's exact test, p = 0.048). **c)** The distribution of TMB between Black and non-Black localized tumor samples. Non-Black patient samples were associated with a slightly higher TMB in our cohort (Mann-Whitney U, $p = 1.13 \times 10^{-7}$).

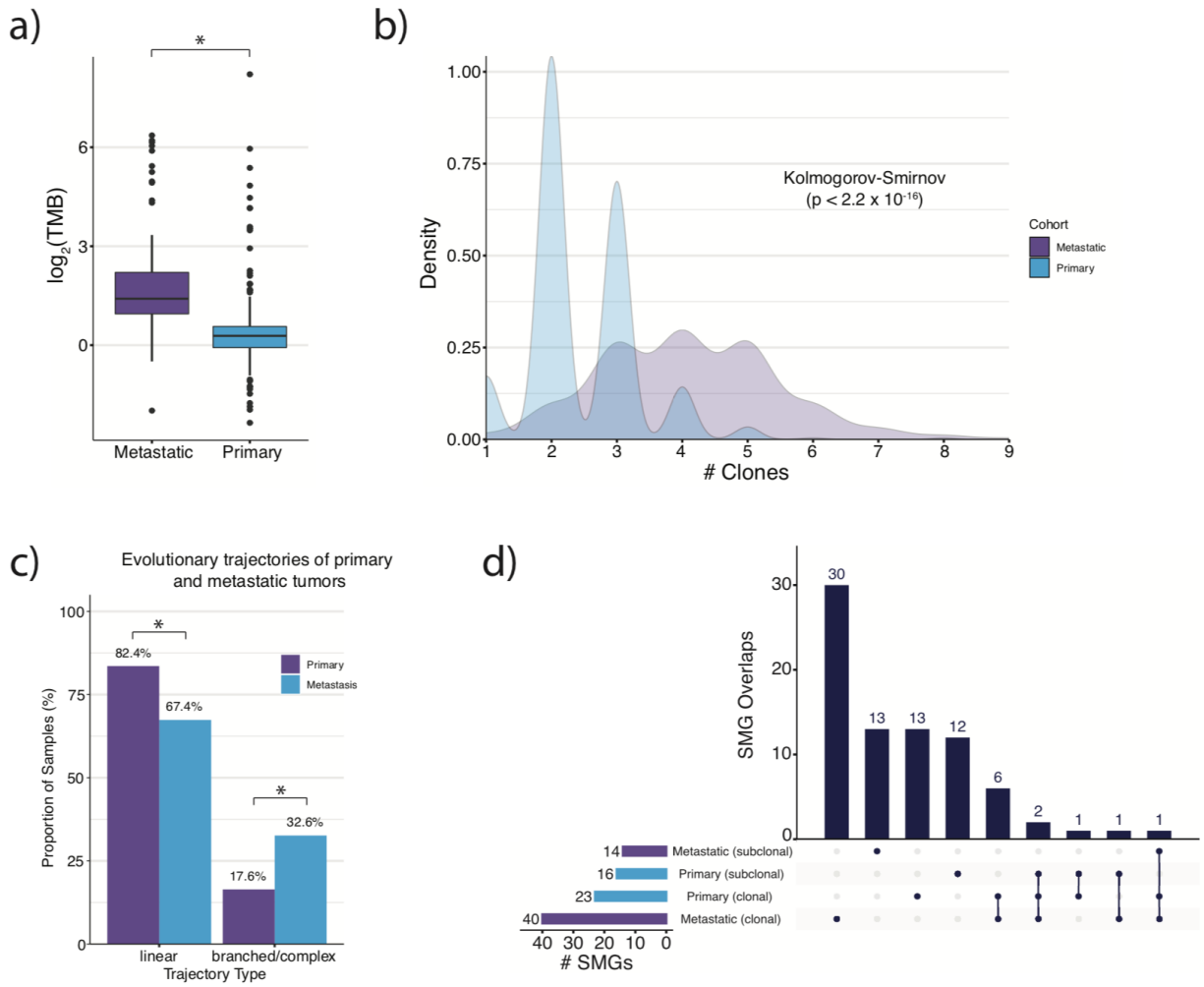


Supplementary Figure 4.2: Continental Ancestry Inference
(a) The first 2 principal components output by Hail "hwe_normalized_pca". Samples from our cohort (n=845) are represented by purple points. 1000 Genomes superpopulation references samples are also included: African (yellow), admixed race (light blue), East Asian (teal), European (red), South Asian (green). **(b)** Ancestry inference assignment of 1000 Genomes super populations for each sample in our cohort using a random forest classifier trained on the first 10 principal components output by Hail "hwe_normalized_pca". The first principal component (PC1) primarily separates samples according to their similarity to the African (yellow) reference population.



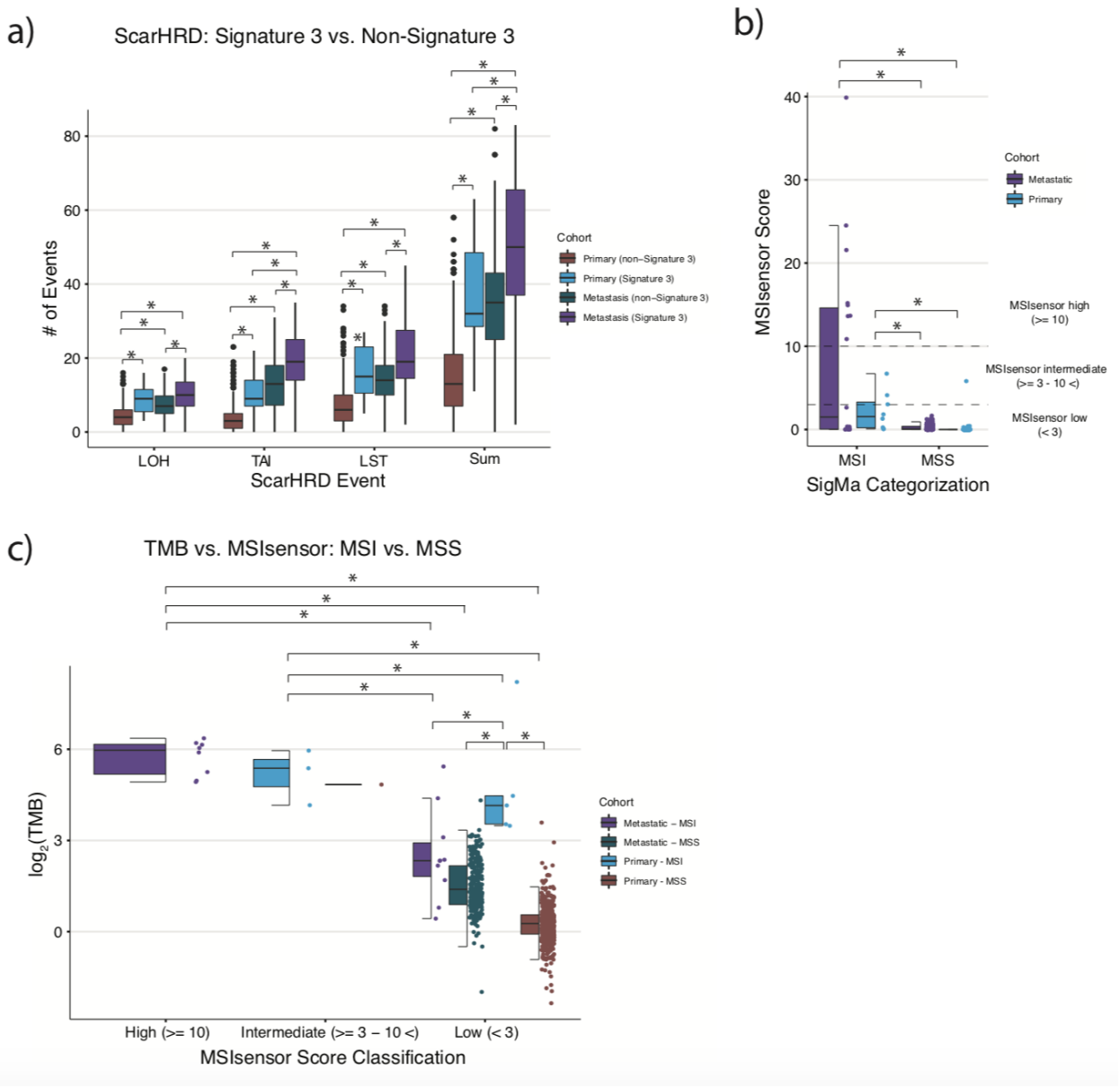
Supplementary Figure 3: WGS validation of preferential punctuated evolutionary trajectories in tumors from Black patients

(a) The distribution of the fraction of clonal mutations in each whole genome sequenced tumor based on self-reported race (n = 7 AA, n = 7 non-AA). Although nonsignificant (Mann-Whitney U, 39.3% vs. 17.1%, p = 0.32), Black patient tumors exhibited higher fractions of mutations in the clonal cluster. **(b)** After correcting for confounding covariates such as TMB, sequencing coverage, and tumor purity, self-reported Black race was borderline significantly associated with a higher proportion of mutations in the clonal cluster (multivariate linear regression, p = 0.08).



Supplementary Figure 4.4: Primary and metastatic PC tumors are associated with distinct genomic and clonal architecture properties

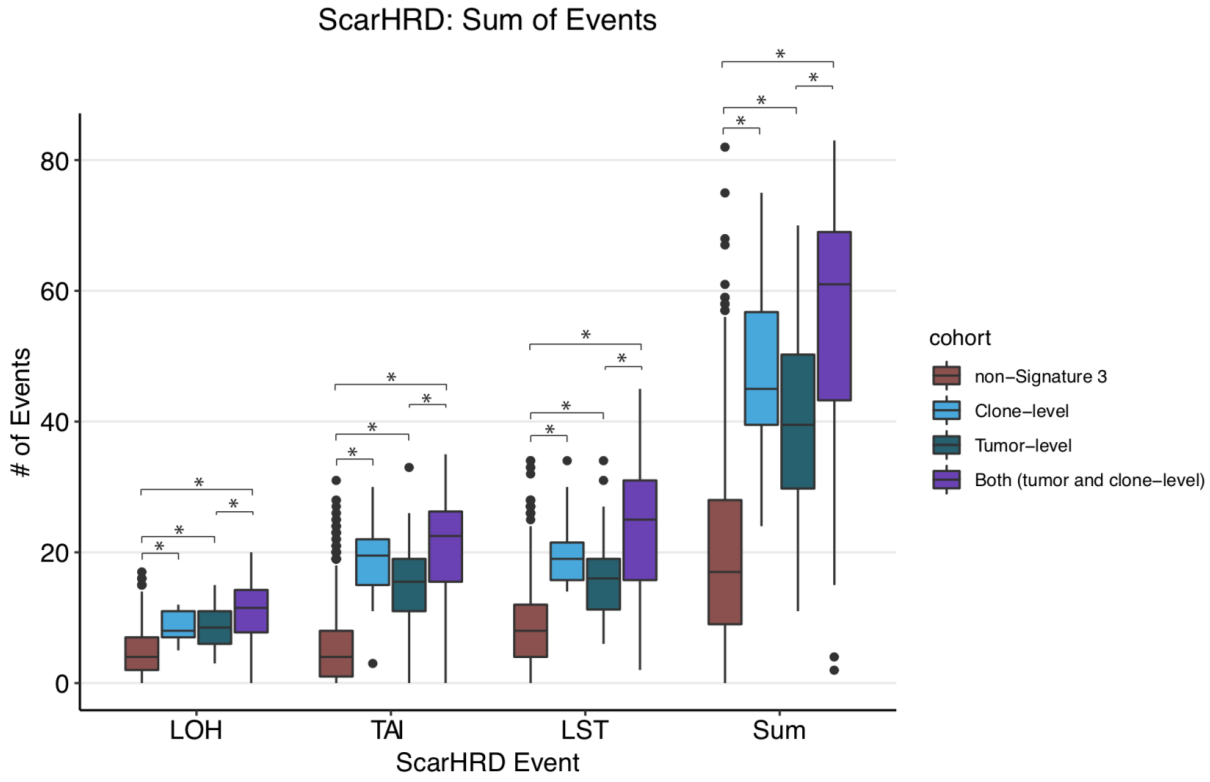
(a) The distribution of tumor mutational burden (TMB) between primary and metastatic samples in our cohort (n = 845). The median TMB observed in the metastatic samples was more than two-fold higher than the TMB observed in primary samples (Mann-Whitney U, 2.66 mut/Mb vs. 1.22 mut/Mb, $p < 2.2 \times 10^{-16}$). **(b)** The distribution of the number of cell subpopulations (clones) between primary and metastatic samples. Metastatic samples were associated with having significantly more cell subpopulations than primary samples (Kolmogorov-Smirnov, p $p < 2.2 \times 10^{-16}$). Similarly, the rate of monoclonality in metastatic samples was 10% compared to 2.5% in primary samples (Fisher's, $p = 3.98 \times 10^{-5}$). **(c)** Primary tumors are associated with a higher rate of linear evolutionary trajectories compared to metastatic tumors (Fisher's; 95% CI = 1.6 - 3.18, OR = 2.25; $p = 1.58 \times 10^{-6}$). **(d)** The overlap between clonal and subclonal primary and metastatic driver genes identified via mutational significance analysis with MutSigCV2.



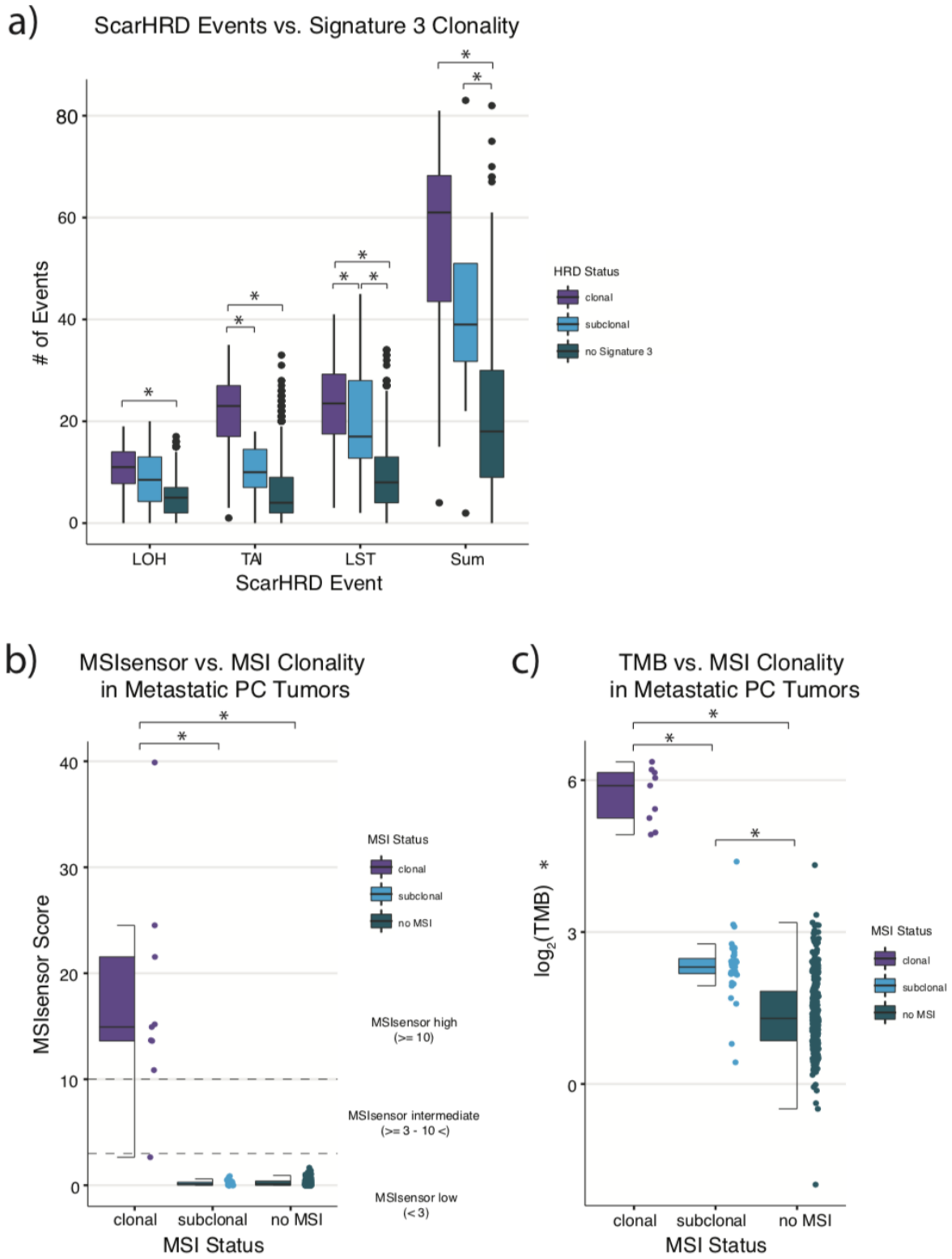
Supplementary Figure 4.5: **Mutational signatures and their associations with genomic events in PC tumors.**

(**a)** The distribution of homologous recombination deficiency (HRD)-associated copy number events between primary and metastatic samples with and without evidence of mutational signature 3. Metastatic samples with signature 3 were associated with higher numbers of HRD-associated copy number events compared to metastatic samples without signature 3, and primary samples with signature 3 were associated with higher numbers of HRD-associated copy number events compared to primary samples without signature 3. Additionally, metastatic samples with signa- ture 3 had higher numbers of HRD-associated copy number events compared to primary samples with signature 3. **(b)** The distribution of MSIsensor scores between primary and metastatic samples with and without MSI-associated mutational signatures. Metastatic samples with MSI-associated mutational signatures had higher MSIsensor scores than metastatic samples without MSI-associated mutational signatures, and

Supplementary Figure 4.5 (continued): primary samples with MSI-associated mutational signatures had higher MSIsensor scores than primary samples without MSI-associated signatures. Additionally, metastatic samples with MSI-associated mutational signatures had higher MSIsensor scores compared to primary tumors with MSI-associated mutational signatures. **(c)** The distribution of mutational burden between primary and metastatic samples with and without MSI-associated mutational signatures. Asterisks denote statistical significance via Mann-Whitney U tests.



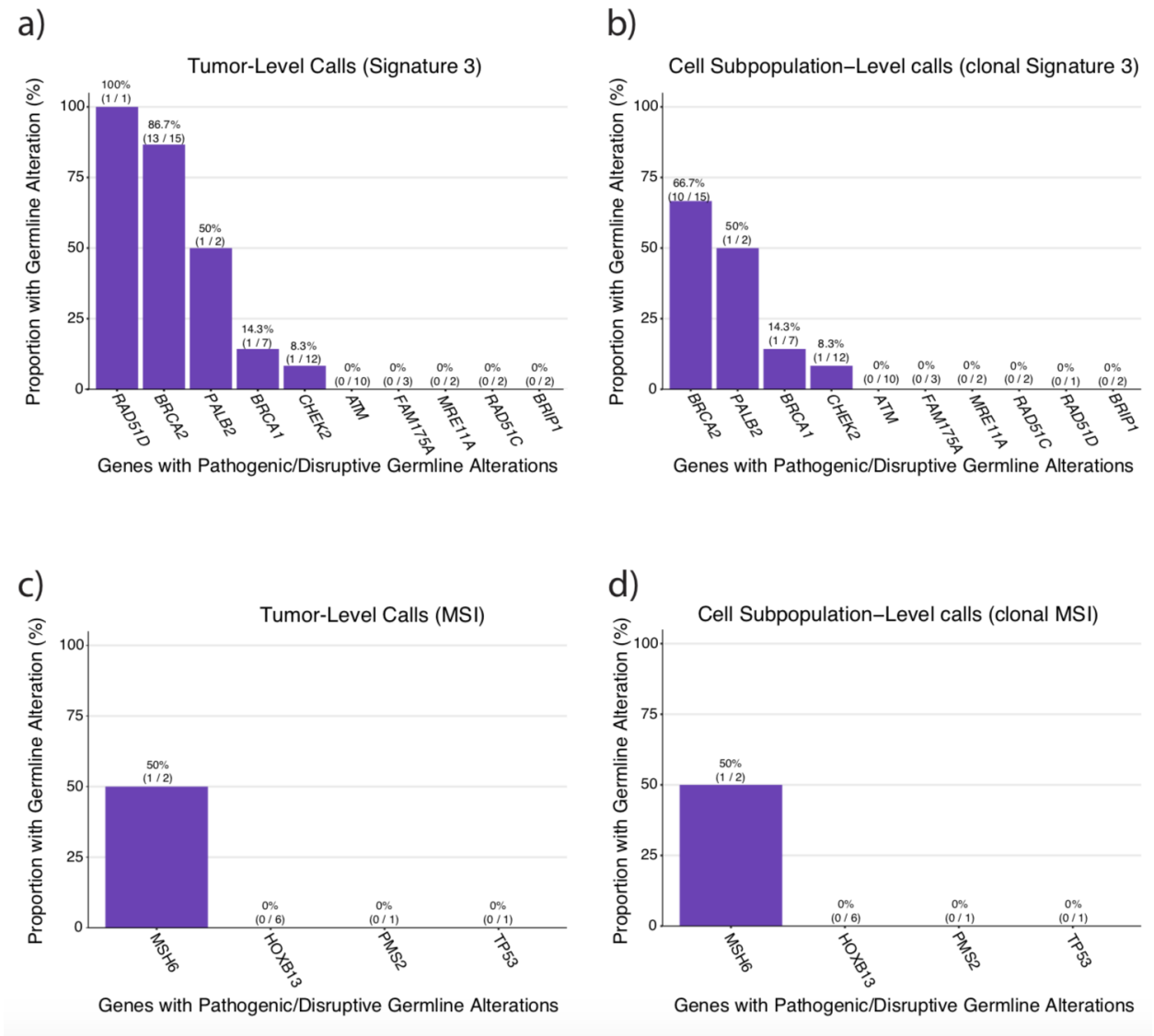
Supplementary Figure 4.6: HRD-associated copy number events are associated with the clonality of mutational signature 3
The distribution of homologous recombination deficiency (HRD)-associated events between tumors where mutational signature 3 was identified at both the tumor and cell subpopulation levels, just at the tumor level, just at the cell subpopulation level, and not at all. Asterisks indicate statistical significance via Mann-Whitney U tests.



Supplementary Figure 4.7: Clonality of mutational signatures and their associations with genomic events in PC tumors.

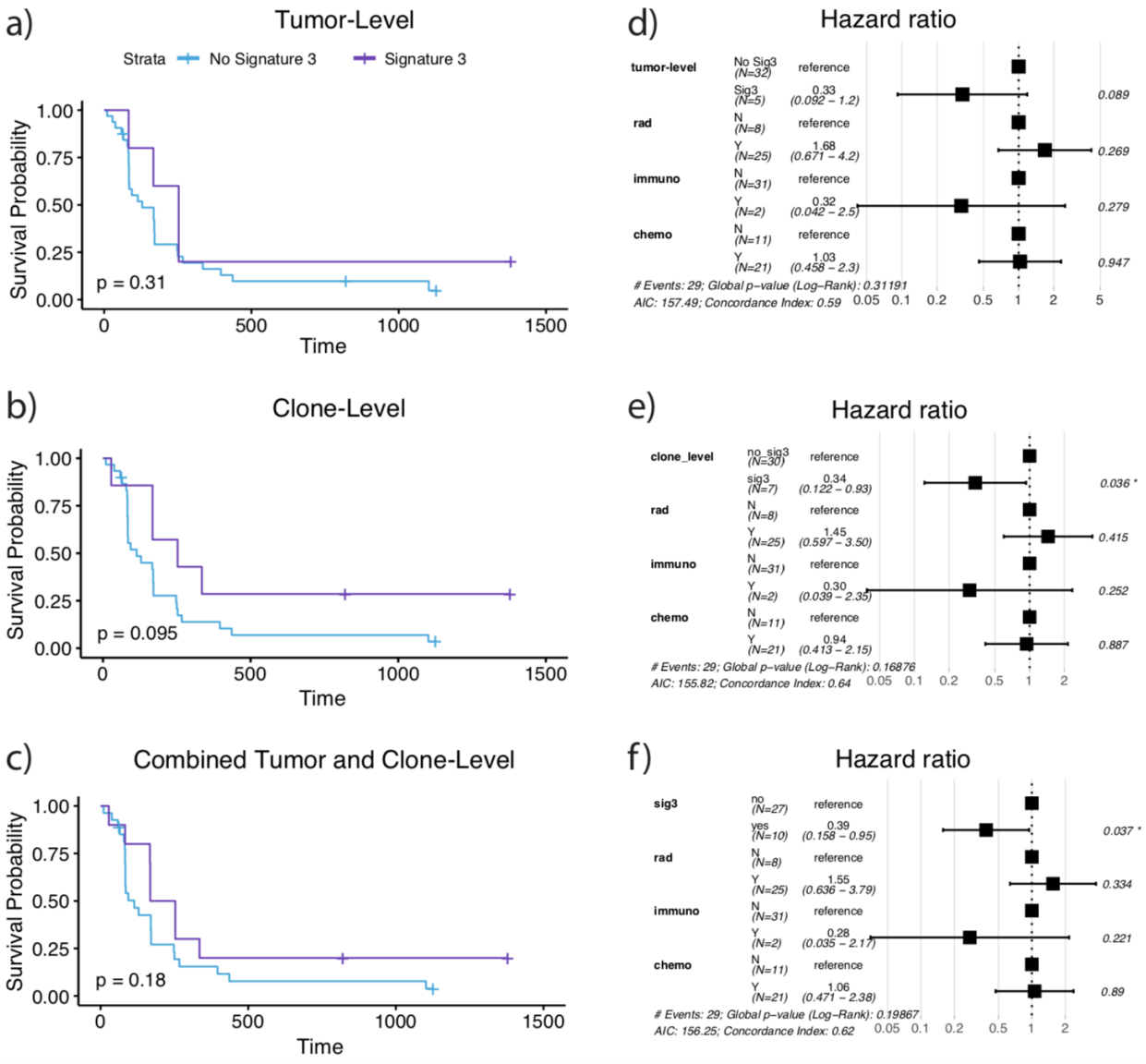
(a) The distribution HRD-associated copy number events based on the clonality of mutational signature 3 from running SigMA at the cell subpopulation level. There is a stepwise increase in

Supplementary Figure 4.7 (continued): the number of HRD-associated copy number events from tumors with no signature 3, to subclonal signature 3, to clonal signature 3. **(b)** The distribution of MSIsensor scores based on the clonality of MSI-associated mutational signatures from running SigMA at the cell subpopulation level. Samples with clonal activity of MSI-associated mutational signatures had significantly higher MSIsensor scores than samples without activity of MSI-associated mutational signatures, however, this was not the case for samples with subclonal activity of MSI-associated mutational signatures. **(c)** The distribution of tumor mutational burden based on the clonality of MSI-associated mutational signatures from running SigMA at the cell subpopulation level. There is a stepwise increase in the TMB from tumors with no MSI-associated signature, to subclonal MSI-associated signature, to clonal MSI associated signature.



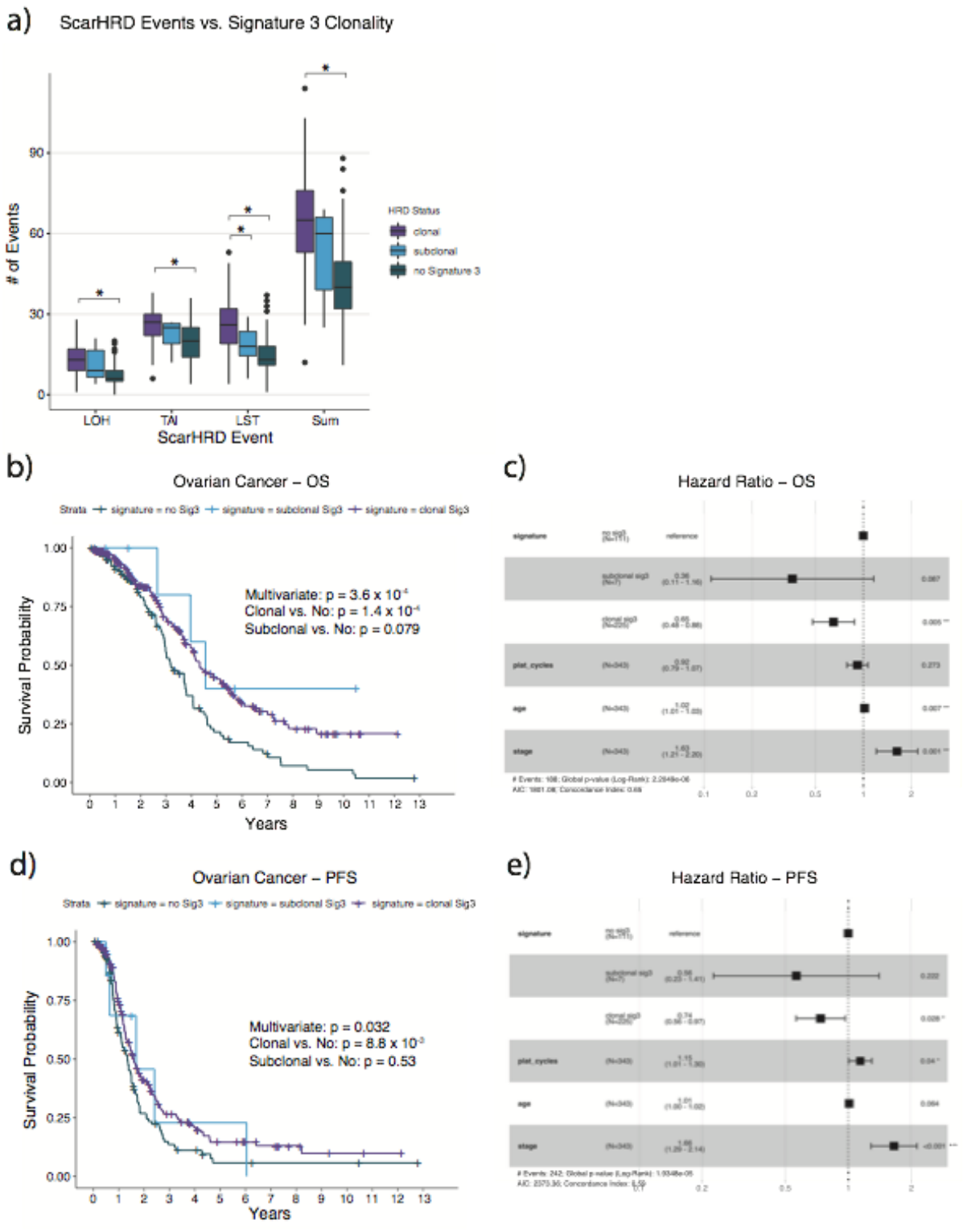
Supplementary Figure 4.8: Germline and putative loss-of-function somatic alterations in DNA repair genes in samples with signature 3 and MSI

(a) The proportion of samples with germline alterations in genes associated with homologous recombination that exhibited mutational signature 3 when running SigMA on all mutations in the tumor and on **(b)** mutations from each cell subpopulation. Samples with germline alterations in *ATM*, *FAM175A*, *MRE11A*, *RAD51C*, and *BRIP1* never exhibited mutational signature 3. **(c)** The proportion of samples with germline alterations in genes associated with mismatch repair deficiency that exhibited microsatellite instability associated mutational signatures when running SigMA on all mutations in the tumor and on **(d)** mutations from each cell subpopulation. Samples with germline alterations in *HOXB13*, *PMS2*, and *TP53* never exhibited microsatellite instability-associated mutational signatures. Some of the genes included in our list of homologous recombination associated and mismatch repair associated genes did not have germline alterations in our cohort, and therefore are not represented in these figures.



Supplementary Figure 4.9: Effect of signature 3 identification framework on PARPi treated prostate cancers

(a) Kaplan-Meier survival curves between signature 3 and non-signature 3 tumors identified by running SigMa at the tumor-level. **(b)** Kaplan-Meier survival curves between signature 3 and non-signature 3 tumors identified by running SigMa at the cell subpopulation-level. **(c)** Kaplan-Meier survival curves between signature 3 and non-signature 3 tumors identified by combining the tumor-level and cell subpopulation-level SigMa calls. **(d)** Multivariate Cox proportional-hazards analysis for tumor-level signature 3, correcting for whether or not the patient also received radiation, immune-related therapy, and chemotherapy. **(e)** Multivariate Cox proportional-hazards analysis for cell subpopulation-level signature 3, correcting for whether or not the patient also received radiation, immune-related therapy, and chemotherapy. **(f)** Multivariate Cox proportional-hazards analysis for the combined tumor-level and cell subpopulation-level signature 3 calls, correcting for whether or not the patient also received radiation, immune-related therapy, and chemotherapy.


a)
ScarHRD Events vs. Signature 3 Clonality
# of Events
ScarHRD Event
LOH
TAI
LST
Sum
HRD Status
clonal
subclonal
no Signature 3
b)
Ovarian Cancer – OS
Strata signature = no Sig3 signature = subclonal Sig3 signature = clonal Sig3
Survival Probability
Years
Multivariate: p = 3.6 x 10-4
Clonal vs. No: p = 1.4 x 10-4
Subclonal vs. No: p = 0.079
c)
Hazard Ratio – OS
d)
Ovarian Cancer – PFS
Strata signature = no Sig3 signature = subclonal Sig3 signature = clonal Sig3
Survival Probability
Years
Multivariate: p = 0.032
Clonal vs. No: p = 8.8 x 10-3
Subclonal vs. No: p = 0.53
e)
Hazard Ratio – PFS

Supplementary Figure 4.10: Effect of signature 3 clonality on survival in cisplatin treated ovarian cancers

(a) The associations between the clonality of mutational signature 3 and the number of homologous recombination deficiency (HRD)-associated copy number events in the ovarian cancer cohort were highly concordant with the observed associations in our prostate cancer cohort. That is, there is a stepwise increase in the number of HRD-associated copy number events from tumors with no signature 3 (n = 111), to subclonal signature 3 (n = 7), to clonal signature 3 (n = 225). Asterisks denote statistical significance via Mann-Whitney U tests. To determine if the clonality of mutational signature 3 affects how tumors respond to therapy, we performed Kaplan-Meier and Cox proportional hazard (PH) analysis on **(b-c)** overall survival (OS) and **(d-e)** progression free survival (PFS). Tumors with clonal activity of mutational signature 3 were associated with significantly improved **(b-c)** OS and **(d-e)** PFS. **(c)** Tumors with only subclonal activity of mutational signature were borderline significantly associated with improved OS (Cox PH, p = 0.087). **(e)** Although nonsignificant, subclonal only activity of mutational signature 3 trended in the direction of improved PFS as well (Cox PH, HR = 0.56, p = 0.22)