Spring 5-15-2020

Multi-omics Integration for Gene Fusion Discovery and Somatic Mutation Haplotyping in Cancer

Steven Mason Foltz
Washington University in St. Louis

Recommended Citation
Foltz, Steven Mason, "Multi-omics Integration for Gene Fusion Discovery and Somatic Mutation Haplotyping in Cancer" (2020). *Arts & Sciences Electronic Theses and Dissertations*. 2186.
https://openscholarship.wustl.edu/art_sci_etds/2186

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
Li Ding, Chair
Christopher Maher,
Michael Province
Nancy Saccone
Ravi Vij

Multi-omics Integration for Gene Fusion Discovery and Somatic Mutation Haplotyping in Cancer

by
Steven Mason Foltz

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2020
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

Science is a team sport, and the work documented in this thesis is a reflection of the amazing team of scientists I have been privileged to work with every day. I am thankful to have learned from, listened to, and exchanged ideas with each of you. You have provided daily laughter and heartfelt friendship, and I will carry your memories and perspectives with me wherever I go. Our lab has changed location on campus over the years, but our community has grown around a constant core of hard work, good humor, and shared purpose, the values that best describe our leader, Dr. Li Ding. Her tireless enthusiasm for digging deeper has taught me how to ask better questions and how to be confident in myself as a scientist. Her ability to see our (my) potential and trust us (me) to reach it has illustrated the value of working as a team -- not just that we can cover more angles, but that we draw out the best in each other. Li, thank you for creating this special place where we do good science together, have fun in each other's company, and always want to come back tomorrow.

Steven Mason Foltz

Washington University in St. Louis

May 2020

Dedicated to all patients, families, and medical teams.

ABSTRACT OF THE DISSERTATION

Multi-omics Integration for Gene Fusion Discovery and

Somatic Mutation Haplotyping in Cancer

by

Steven Mason Foltz

Doctor of Philosophy in Biology and Biomedical Sciences

Human and Statistical Genetics

Washington University in St. Louis, 2020

Professor Li Ding, Chair

Cancer is a disease caused by changes to the genome and dysregulation of gene expression. Among many types of mutations, including point mutations, small insertions and deletions, large scale structural variants, and copy number changes, gene fusions are another category of genomic and transcriptomic alteration that can lead to cancer and which can serve as therapeutic targets. We studied gene fusion events using data from The Cancer Genome Atlas, including over 9,000 patients from 33 cancer types, finding patterns of gene fusion events and dysregulation of gene expression within and across cancer types. With data from the CoMMpass study (Multiple Myeloma Research Foundation), we generated the largest gene fusion study in multiple myeloma (742 patients), which is the second most common type of blood cancer, and which is driven by recurrent translocations. We then developed a novel tool for analyzing the haplotype context of somatic mutations. Linked-read whole genome sequencing enables haplotype resolution for analyzing somatic mutation patterns, which is lost during typical short-read sequencing and alignment. We analyzed a cohort of 14 multiple myeloma patients across disease stages, phasing three-quarters of high confidence somatic mutations and enabling us to interpret clonal evolution models at higher resolution. Finally, we also studied the co-evolution of the multiple myeloma tumor and microenvironment using single-cell RNA-sequencing, finding distinct patterns of tumor subclone evolution between disease stages in 14 patients. Our methods and results demonstrate the power of integrating data types to study complex and dynamic evolutionary pressures in cancer and point to future directions of research that aim to bridge gaps in research and clinical applications.

Chapter 1: Cancer is a genetic disease subject to evolutionary dynamics

Cancer is a disease caused by mutations to the genome. When a mutation changes the genome, the highly evolved system of checks and balances in each cell is disrupted, allowing for aberrant and uncontrolled growth. Somatic mutations accumulate randomly over time through mutagenesis, while germline variants are inherited and may confer a predisposition to cancer. An example of an inherited genetic condition is Lynch Syndrome, which is characterized by dysregulation of the mismatch repair (MMR) pathway that normally functions by correcting or eliminating errors in DNA transcription each time the cell divides [1]. Deficiency in this repair mechanism allows somatic mutations to accumulate at an accelerated rate and leads to an increased risk of developing cancer at an early age. Tumors that result from MMR pathway deficiency are frequently characterized by high levels of microsatellite-instability (MSI), most often observed in colorectal, stomach, and uterine cancers [2]. Identifying tumors with high levels of MSI sheds light on how they have evolved and may be treated. In Foltz, *et al.* [3], we approached MSI status prediction by analyzing mutations and methylation levels affecting mismatch repair genes. We built MIRMMR, a user-friendly, computational framework for penalized logistic regression modeling to predict MSI status using alterations in MMR pathway genes and compared its effectiveness to sequence-based methods like MSI-sensor [4]. By integrating multiple data types, leveraging the power of large study cohorts, and developing user-friendly analysis tools, MIRMMR illustrates a common approach to computational cancer genomics research.

Positive selection on mutations leads to cancer

Cancer represents the power of positive selection; it is microevolution at an accelerated rate. Tumors are the result of inherited germline variants and acquired somatic mutations that confer competitive advantage over the surrounding cells. In cancer sequencing studies, mutations may be observed recurrently at the same genomic site across many individuals, may be seen in many patients but dispersed across an entire gene body, or may not occur in genes at all, but in the regulatory regions influencing gene expression. Mutations are not limited to single nucleotide variants (SNV), but may take the form of copy number variation (CNV), or other structural variations (SV). Gene fusions are the result of some previous event of genomic instability, such as a deletion, insertion, inversion, or translocation, that brings two distant parts of the genome into closer proximity. For example, a deletion may eliminate the DNA sequence between the exons of two genes. Transcription of that altered DNA sequence results in a new hybrid mRNA, leading to hybrid fusion protein with an altered function or level of expression. Other mechanisms of fusions relevant to cancer include repositioning the powerful regulatory regions of one gene to be in the neighborhood of a gene whose increased activity may be oncogenic.

Gene fusions have oncogenic potential

Fusions have played an important role in the history of cancer, and present-day studies continue to reveal their functional and clinical relevance [5]. Pioneering discoveries from the 1950s-1970s identified a translocation between chromosomes 9 and 22 associated with chronic myeloid leukemia (CML) [6]. This Philadelphia chromosome encodes a gene fusion of *BCR* and *ABL1*, a tyrosine kinase. Normally, *ABL1* activity is auto-regulated, but the *BCR--ABL1* fusion causes *ABL1* to adopt an “always on” state, leading to increased cell proliferation and genome instability. However, this increased activity can be targeted by tyrosine kinase inhibitors, leading to more effective treatments for patients with CML and other tumors with oncogenic kinase mutations.

Distinguishing important, cancer-causing mutations (driver mutations) from those that have happened but did not confer selective advantage (passengers) has been a major effort in the cancer genomics field [7-9]. Driver mutations can happen in oncogenes (genes that take on a more active or altered role in cancer) or tumor suppressor genes (genes whose lack of functionality releases the brakes that inhibit cancer). Large-scale, multi-platform sequencing studies such as those organized by The Cancer Genome Atlas [10] and The Multiple Myeloma Research Foundation have enabled researchers to paint landscapes of the genetic determinants of cancer in dozens of cancer types. A major motivation for cancer studies is to identify effective therapies for patients with particular mutations and to provide treatment options targeted to an individual patient. With comprehensive overviews of many cancer types already published, doctors and drug developers have a solid baseline when considering effective therapies for individual patients.

Tumors are heterogenous and require multi-omic approaches

Each patient's cancer is a unique and heterogenous entity that can be studied from a variety of angles. Various approaches to data analysis with many data types are carried out, including detection of SNVs, CNVs, and SVs from DNA, and gene expression and fusion detection from RNA. Assessing tumor heterogeneity is an important aspect of cancer genomics since understanding the subclonal structure can reveal important clues for how a tumor has evolved and what targeted therapies may be most effective [11-14]. Further, mutations found in

DNA can be mapped onto protein structures, demonstrating how mutations far apart on the linear DNA sequence can affect the same functional units in three-dimensional protein space [15].

Genomic data from a single individual may occupy dozens of gigabytes of computer storage. With this wealth of data from thousands of samples, multiplied by numerous data types, efficient computational approaches are key to discerning what is important information. Computational tools and databases enable discovery of new trends and allow researchers to generate testable hypotheses that can be rigorously examined under laboratory conditions. Viewing genetic events from different angles with multiple data types colors in the details of each cancer's molecular portrait [16,17]. Further, utilizing new technologies brings more breadth and depth to cancer analysis. One example is single-cell RNA-sequencing, which has allowed unprecedented resolution of gene expression and tumor heterogeneity. Another example is linked-read whole genome sequencing (WGS), which combines the accuracy of Illumina sequencing with the long-range connectivity and improved mapping of haplotypes [18,19].

Case Study 1: Data integration models microsatellite instability

The story each of each individual tumor may not be told by a single data type. We may need to combine information from multiple platforms to gain a more complete picture of what is going on inside a tumor. With that in mind, we approached the problem of predicting microsatellite instability (MSI) by building a logistic regression model based upon two data types: mutations and methylation. Our work, MIRMMR: binary classification of microsatellite instability using methylation and mutations, was published in *Bioinformatics* (2017).[3] Please refer to the publication for any supplementary information. Contributions: As sole first author, SMF developed the modeling concept, wrote the software package, and wrote the manuscript.

Introduction

Microsatellites consist of short DNA sequence repeats and may change in size due to errors in DNA replication, in particular because of strand slippage [20]. Normally, such errors are caught and repaired through mechanisms of the mismatch repair (MMR) pathway. However, changes in the methylation level of gene promoters and deleterious mutations in MMR pathway genes such as MLH1 may be responsible for dysregulation of the MMR pathway and increases in microsatellite instability (MSI) [21]. MSI is strongly associated with inherited cancer syndromes such as Lynch syndrome and is an important diagnostic indicator that may influence treatment options.

Experimental and computational methods exist to detect MSI in patient samples. Experimentally, the length of known microsatellites is measured using gel electrophoresis and compared between normal and tumor samples. Computational methods such as MSIsensor [4] and mSINGS [22] measure the prevalence of unstable microsatellites by examining sequence data from normal and tumor samples. MSIseq [23] and MOSAIC [23] use machine learning classifiers based on microsatellite variants and other microsatellite features.

The experimental measurement process is time consuming and only assays a limited number of markers. Measuring microsatellite length in DNA-seq data requires computational resources to store and process sequencing data. MOSAIC and MSIseq mitigate these issues by working on smaller files but still focus on microsatellite features such as the number of microindels observed in simple repeat regions per mega-base.

Instead of observing microsatellites directly to evaluate MSI status, we created an orthogonal prediction method using methylation levels and mutations in MMR pathway genes. Here we present MIRMMR (pronounced ‘murmur’): Microsatellite Instability Regression using Methylation and Mutations in R. MIRMMR trains logistic regression models using DNA methylation and mutation information from MMR pathway genes to classify MSI status. Once a prediction model has been trained, MIRMMR quickly reports the likely MSI status of new samples.

Methods

MIRMMR consists of several independent modules to build logistic regression models, compare method outcomes, and classify MSI status in new samples. Users may select penalized, stepwise, or univariate modules to perform logistic regression modeling. Given a binary measure of MSI status, MIRMMR trains logistic regression models based on predictors such as MMR pathway gene methylation levels or mutation severity indicators, like Combined Annotation Dependent Depletion (CADD) scores [24].

Penalized regression can perform variable selection by setting the coefficients of unimportant predictors to zero, which is vital to finding an informative and relevant model. MIRMMR’s penalized module performs elastic net regression based on R’s glmnet package [25], which lets users balance the penalty term’s L1 and L2 norms.

A vital task in penalized regression is selecting an appropriate weight (lambda) to give the entire penalty term. Minimizing cross validation (CV) error is one way to find the optimal lambda value. However, due to the randomness of fold selection, the best lambda value may not be consistent between successive CV runs. After many independent CV runs, MIRMMR selects the lambda value with minimal average CV error. It fits a penalized logistic regression model using that lambda value and reports a logistic model based on the automatically selected variables.

See Supplementary Information for a description of all MIRMMR parameters, including options to train and test models on subsets of data.

Results

We used MIRMMR's penalized module to train a model on colorectal (COADREAD), stomach (STAD), and uterine (UCEC) tumor samples from The Cancer Genome Atlas (TCGA) [10]. Of 676 total samples, 123 (123/676, 18.2%) were called MSI-High by TCGA. We trained the model using 10-fold CV with no samples withheld for testing. Model predictors included point mutation rate, methylation beta levels at MMR genes, and CADD scores for mutations found in MMR genes. See Supplementary Information for a full list of MMR pathway genes included and a summary of the final model produced, which highlights predictors important for MSI status prediction. Figure 1 illustrates the distribution of MIRMMR scores and shows a clear separation between TCGA MSI-High and Not-MSI-High groups.

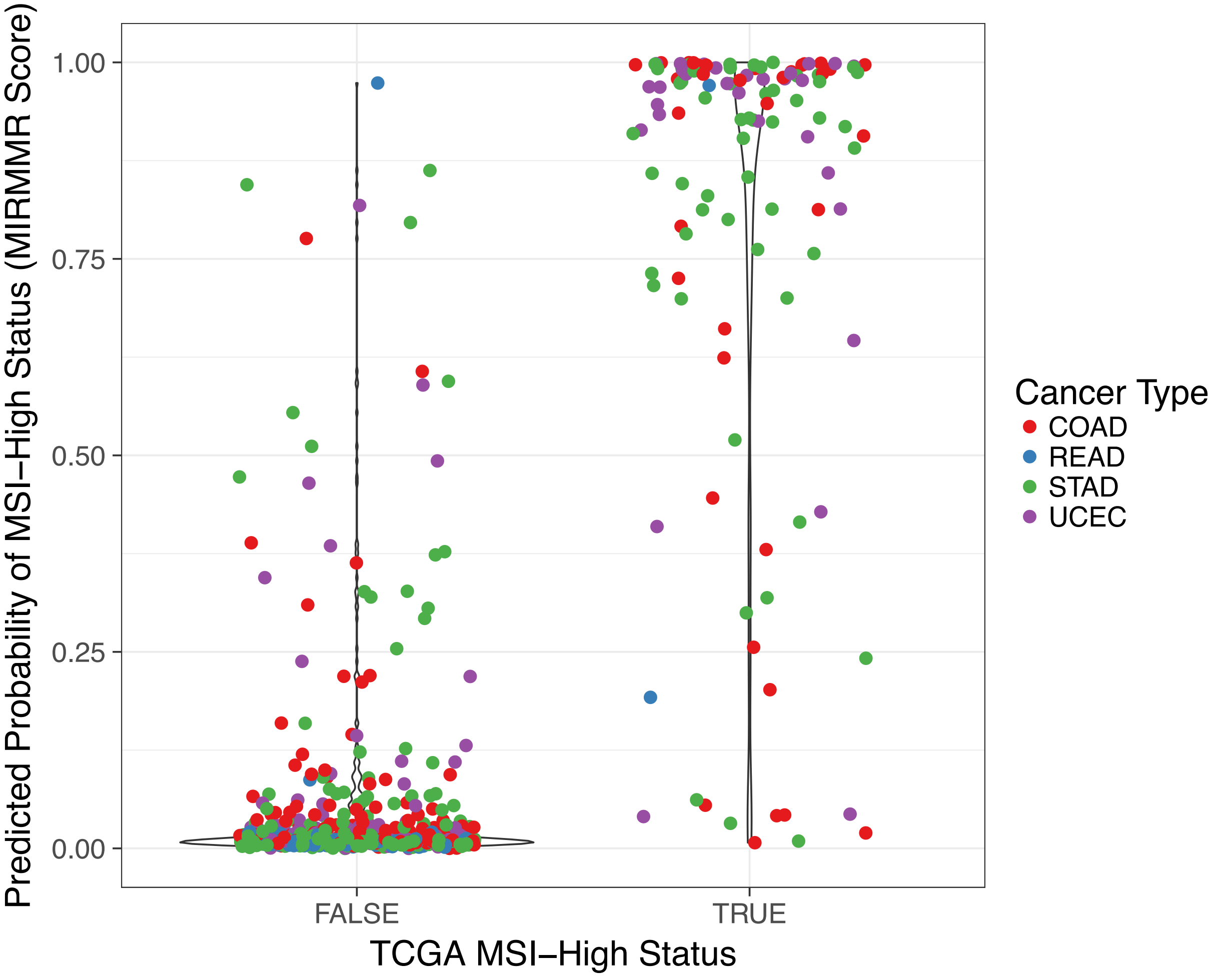


Figure 1. MIRMMR scores. MIRMMR scores (y-axis) indicate a sample's predicted probability of having MSI-High status. Higher scores indicate higher probability of being MSI-High. The x-axis indicates MSI-High status reported by TCGA. The prediction model was built using 676 COADREAD, STAD, and UCEC samples from TCGA.

MIRMMR reports a score between zero and one, so a suitable cutoff to separate MSI-High samples from Not-MSI-High samples is necessary. Individual users may decide on a cutoff to balance their own needs for sensitivity and specificity. We selected a cutoff score of 0.1922 to maximize the sum of sensitivity (0.9187) and specificity (0.9421). With this cutoff, we found 634 samples (634/676, 93.8%) for which the original TCGA experimental MSI status call matched the MIRMMR call. Missed calls could be due to incomplete or inaccurate mutation and methylation reporting. We found similar areas under the curve when comparing the ROC curves of MIRMMR (0.9727), mSINGS (0.9799), and MSIsensor (0.9977), indicating that MIRMMR offers a promising new option for integrated MSI diagnosis that does not rely on measuring microsatellites. Given the high accuracy of existing, sequence-based methods, MIRMMR also offers an orthogonal measurement to reinforce concordant calls and flag potentially misclassified samples for further review.

Conclusion

MIRMMR provides a new dimension in MSI diagnosis and modeling. Although previous studies [23] have used regression to infer relationships between certain gene mutations and MSI, only MIRMMR performs full logistic regression model building for the purpose of MSI status prediction via binary classification. Building a pre-diction model highlights genes contributing to the MSI phenotype, and users can set intuitive classification thresholds based on probabilities.

We trained a logistic regression model to predict MSI status based only on mutation and methylation data using samples from COADREAD, STAD, and UCEC cancer types. MIRMMR's classification performance was on par with methods that rely on measuring microsatellites in BAM files, providing an additional, accurate tool for MSI diagnosis.

Case Study 2: Identifying sample swaps in a large, multi-omics cohort

Beyond data integration for gaining deeper insights into cancer biology, various cancer data types can be compared to ensure data quality and consistency. For example, by matching germline mutations from WGS, WXS, and RNA-seq data samples to ensure they all originated from the same individual. This need for quality control exists for all scales of data analysis, but it is especially visible and necessary to be done robustly in large scale, public consortium data that is shared by multiple institutions and will be the foundation for multiple publications and advancing science for the years following. One such consortium is the Clinical Proteomic Tumor Analysis Consortium (CPTAC), a program of the National Cancer Institute. Throughout various phases of the CPTAC project, genomics, transcriptomics, and proteomic data has been collected and analyzed, leading to several high profile publications.[26-31] Mishaps in sample handling at the data generation stage could have profound downstream effects, potentially causing confusion, inconsistency, mistrust, and misleading results.

Copy number variation can be detected from a group of WXS cancer samples by comparing each cancer sample against a background panel of normals. The assumption is that every sample has been sequenced under the same protocol, so that differences in read depth are directly related to changes in the copy number profile. In theory, normal samples have two copies of each chromosome without any local variation. Cancer samples, however, may have wild fluctuations that result in dysregulation and selective advantage. We developed a somatic copy number profiling pipeline based on the Genome Analysis Toolkit [32] workflow, and we deployed it to analyze over 300 samples from breast, ovarian, and colorectal cancer types. We also intentionally utilized this pipeline to examine germline CNVs.

Our initial visual examination of germline copy number profiles revealed an immediate problem -- one batch of data, nearly 20 patients -- had germline CNV profiles that were not flat and even like we expected (see Figure 2). At the same time, their tumor CNV profiles resembled germline CNV profiles. Therefore, we suspected a sample swap occurred that only affected this batch of data. A simultaneous discovery was that the samples in this batch did not have any *TP53* somatic mutations, which we expected to be above 90% in ovarian cancer. Here we integrated findings from past studies to double check our results against the expectation, and we used two data types to corroborate an error that showed up downstream in both. To fix the error, the simple bioinformatic solution was to swap the files back to match the samples they originated from. In practice, distributing such a fix to widespread collaborators required repeated explanations and careful documentation.

The checking and corroborating process should be built-in to any bioinformatics pipeline, and no results should be taken at face value. They must fit into the context of a dynamically integrated biological system and also fit the paradigm of cancer as an entity responding to evolutionary pressure. The work leading to this thesis has been a struggle against complacent pipelines. Our responsibility in cancer research is to contextualize findings so they are meaningful and useful to others -- first as a resource, then as inspiration for future research, and eventually as a springboard for innovative translation to the clinic.

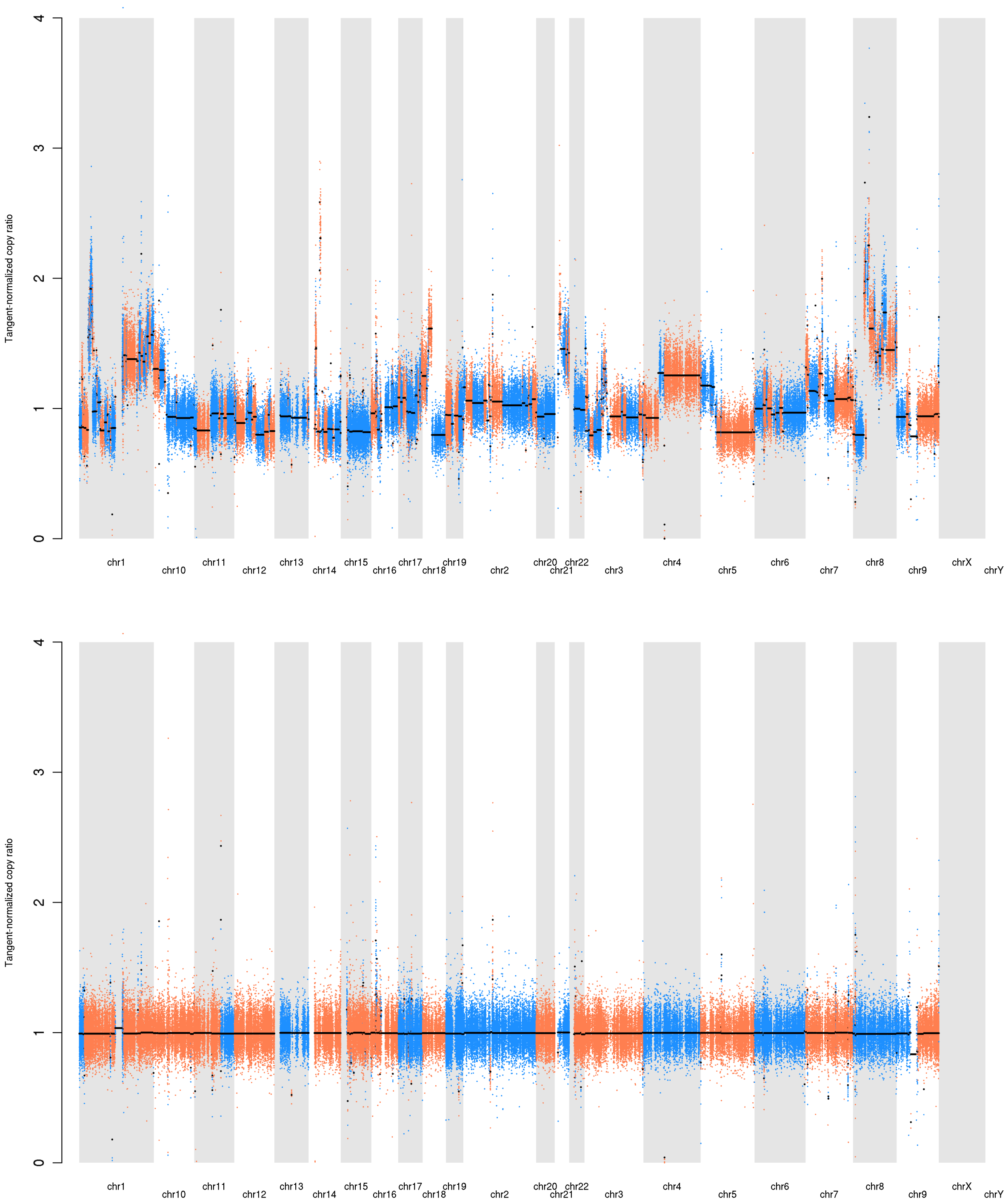


Figure 2. Copy number profiles of swapped samples. Top: a "normal" sample before correcting the swap. Bottom: a swapped "tumor" sample from the same patient.

Chapter 2: Driver Fusions and Their Implications in the Development and Treatment of Human Cancers

Background

Gene fusions are a common cause of cancer that account for 20% of human cancer morbidity [6]. Fusion detection from RNA-sequencing data remains an important challenge in cancer sequencing studies [33]. There are inherent computational difficulties, such as mapping hybrid reads efficiently, which are compounded by biological complexities like tumor heterogeneity. These difficulties are reflected in the poor concordance between fusion detection tools run on the same input data and the large number of false positives events often reported [34]. To overcome these challenges and provide insights into cancer fusions, we developed fusion detection strategies that integrate the results of multiple fusion tools (for higher sensitivity and specificity) and apply multiple layers of filtering (to reduce false positives).

We applied our fusion detection framework to The Cancer Genome Atlas dataset, including 9,624 tumor samples from 33 cancer types [35]. Our comprehensive approach broadened the existing landscape of pan-cancer fusion studies, and we incorporated fusion events with gene expression and mutation data. We found patterns of upregulated gene expression when an oncogene was a fusion partner, and tumor suppressor fusions were often downregulated. We found 6.0% of samples with a fusion that could be a potential drug target. However, a major problem remains:

from the long lists of fusions detected, what fusions are most important and likely to be drivers of disease?

Prior efforts to conduct pan-cancer fusion detection have utilized only a single fusion calling algorithm [36-38]. Since disagreements among different callers are common, a comprehensive approach that combines the strengths of various callers could achieve higher fusion calling accuracy. Further, large-scale analyses are likely to expand the landscape of druggable fusions in cancer, revealing potential treatment options for patients.

We leveraged multiple newly-developed bioinformatics tools to methodically identify fusion transcripts from TCGA using the ISB Cancer Genomics Cloud. These tools included STAR-Fusion [39], Breakfast, and EricScript [40]. Fusion calling across 9,624 TCGA tumor samples from 33 cancer types identified a total of 25,664 fusion transcripts, with a 63.3% validation rate for the samples having available whole genome sequencing data. We investigated the relationship between fusion status and gene expression and analyzed fusions as potential drug targets.

We explored the gene expression of fusions involving oncogenes, protein kinases, and tumor suppressor genes. For example, Figure 1C illustrates the higher expression level of oncogene *RET* in thyroid carcinoma (THCA) samples with a *RET* fusion. Figure 1A shows that samples with fusions in oncogenes are more likely to overexpress that oncogene, while samples with tumor suppressor fusions are more likely to underexpress tumor suppressor genes. Figure 1B shows that the median expression level of oncogenes is higher in samples with fusions than those without fusions, and the median expression of fused tumor suppressors tends to be lower, though the pattern is less consistent.

Overexpressed fusions, especially in-frame kinase fusions, are commonly targeted for therapy due to their susceptibility to kinase inhibitors. We found that 6.0% of samples had a druggable fusion event based on the Database of Evidence for Precision Oncology (DEPO) (http://dinglab.wustl.edu/depo).

Our work, Driver Fusions and Their Implications in the Development and Treatment of Human Cancers, was published in *Cell Reports* (2018) as part of the TCGA Pan-Cancer Atlas.[35] Please refer to the publication for any supplementary information. Contributions: As co-first author with Qingsong Gao and Wen-Wei Liang, SMF developed the fusion calling and filtering pipeline, analyzed gene expression and druggability, produced and edited figures, and wrote and edited the manuscript.

Summary

Gene fusions represent an important class of somatic alterations in cancer. We systematically investigated fusions in 9,624 tumors across 33 cancer types using multiple fusion calling tools. We identified a total of 25,664 fusions, with a 63% validation rate. Integration of gene expression, copy number, and fusion annotation data revealed that fusions involving oncogenes tend to exhibit increased expression, while fusions involving tumor suppressors have the opposite effect. For fusions involving kinases, we found 1,275 with an intact kinase domain, the proportion of which varied significantly across cancer types. Our study suggests that fusions drive the development of 16.5% of cancer cases and function as the sole driver in more than 1% of them. Finally, we identified druggable fusions involving genes such as *TMPRSS2*, *RET*, *FGFR3*, *ALK*, and *ESR1* in 6.0% of cases, and we predicted immunogenic peptides, suggesting that fusions may provide leads for targeted drug and immune therapy.

Significance

The Cancer Genome Atlas project is concluding with a broad finale of analyses on the final data corpus of approximately 11,000 samples across 33 cancer types. Here, we focus on gene fusions, which can arise through various mechanisms, such as translocation and interstitial deletion, and which play crucial roles in cancer diagnosis and prognosis. We conducted a systematic, multi-tool analysis to discover 25,664 fusion events across cancer types. Our integrated analyses, involving gene expression, copy number, and other results shed light on the effects of fusions in oncogenes and tumor suppressors. We also highlighted the cancer types in which fusions play important and even primary driver roles.

Introduction

The ability to determine the full genomic portrait of a patient is a vital prerequisite for making personalized medicine a reality. To date, many studies have focused on determining the landscape of single nucleotide polymorphisms, insertions, deletions, and copy number alterations in cancer genomes [9,41-45]. While such genomic alterations make up a large fraction of the typical tumor mutation burden, gene fusions also play a critical role in oncogenesis. Gene fusions or translocations have the potential to create chimeric proteins with altered function. These events may also rearrange gene promoters to amplify oncogenic function through protein overexpression or to decrease the expression of tumor suppressor genes.

Gene fusions function as diagnostic markers for specific cancer types. For example, a frequent translocation between chromosomes 11 and 22 creates a fusion between *EWSR1* and *FLI1* in Ewing's sarcoma. Also, the Philadelphia chromosome 9-22 translocation is characteristic of chronic myeloid leukemia, resulting in the fusion protein *BCR--ABL1*. This fusion leads to constitutive protein tyrosine kinase activity and downstream signaling of the PI3K and MAPK pathways, which enables cells to evade apoptosis and achieve increased cell proliferation [46-49]. Fibrolamellar carcinoma (FLC) in the liver is characterized by a *DNAJB1--PRKACA* fusion. A recent study of TCGA tumors revealed this fusion transcript is specific to FLC, differentiating it from other liver cancer samples [50]. In contrast, *FGFR3--TACC3* is an inframe activating kinase fusion found in multiple cancer types, including glioblastoma multiforme (GBM) [51,52] and urothelial bladder carcinomas (BLCA) [53]. Other recurrent fusions have also been reported in multiple cancer types [54-56], and functional characterization of a few selected fusion genes in cellular model systems has confirmed their oncogenic nature [57].

Recently, large-scale genomic studies have utilized the TCGA RNA-Seq data corpus to systematically identify and compile fusion candidates across many cancer types. For example, as part of its goal to develop a comprehensive, genome-wide database of fusion genes, ChimerDB [58] has analyzed RNA-Seq data of several thousand TCGA cases. Giacomini et al. performed breakpoint analysis on exon microarrays across 974 cancer samples and identified 198 candidate fusions in annotated cancer genes [59]. A searchable portal of TCGA data includes 20,731 fusions called from 9,966 cancer and 648 normal samples [60]. Some studies focus on important classes of genes, such as kinase fusions [37], which may have particular structural properties that are selected for during oncogenesis and cancer progression. However, most efforts have utilized only a single fusion calling algorithm. Since disagreements among different callers are common, there is a need to develop a comprehensive approach that combines the strengths of various callers to achieve higher fusion calling accuracy. Further, large-scale analyses are likely to expand the targetable landscape of fusions in cancer, revealing potential treatment options for patients.

Here, we leverage multiple newly-developed bioinformatic tools to methodically identify fusion transcripts across the TCGA RNA-Seq data corpus using the ISB Cancer Genomics Cloud. These tools include STAR-Fusion, Breakfast, and EricScript (STAR Methods). Fusion calling across 9,624 TCGA tumor samples from 33 cancer types identified a total of 25,664 fusion transcripts, with 63.3% validation rate for the samples having available whole genome sequencing data. Further, we investigated the relationship between fusion status and gene expression, the spectrum of kinase fusions, mutations and fusions found in driver genes, and fusions as potential drug and immunotherapy targets.

Results

Fusion detection pipeline and WGS-based validation of a subset of fusion predictions

We analyzed RNA-Seq data from 9,624 tumor samples and 713 normal samples from The Cancer Genome Atlas (TCGA) using STAR-Fusion (STAR Methods), EricScript [40], and Breakfast (STAR Methods, Table S1**)**. A total of 25,664 fusions were identified after extensive filtering using several panel-of-normals databases, including fusions reported in TCGA normal samples, GTEx tissues [61] and non-cancer cells [62] (STAR Methods, Fig. 1A, and Table S1). Our pipeline detected 405 out of 424 events curated from individual TCGA marker papers (Table S1) (95.5% sensitivity).

We further cross-confirmed our transcriptome sequencing-based fusion detection pipeline by incorporating whole genome sequencing (WGS) data, where available. WGS paired-end reads aligned to the partner genes of each fusion were used to validate fusions detected using RNA-Seq. Using all available whole-genome sequencing, including both low-pass and high-pass data, from 1,725 of the 9,624 cancer samples across 25 cancer types, we were able to evaluate 18.2% (4,675 fusions) of our entire fusion call set. Of that subset, WGS validated 63.3% of RNA-Seq based fusions by requiring at least three supporting discordant read pairs from the WGS data (Figure S1).

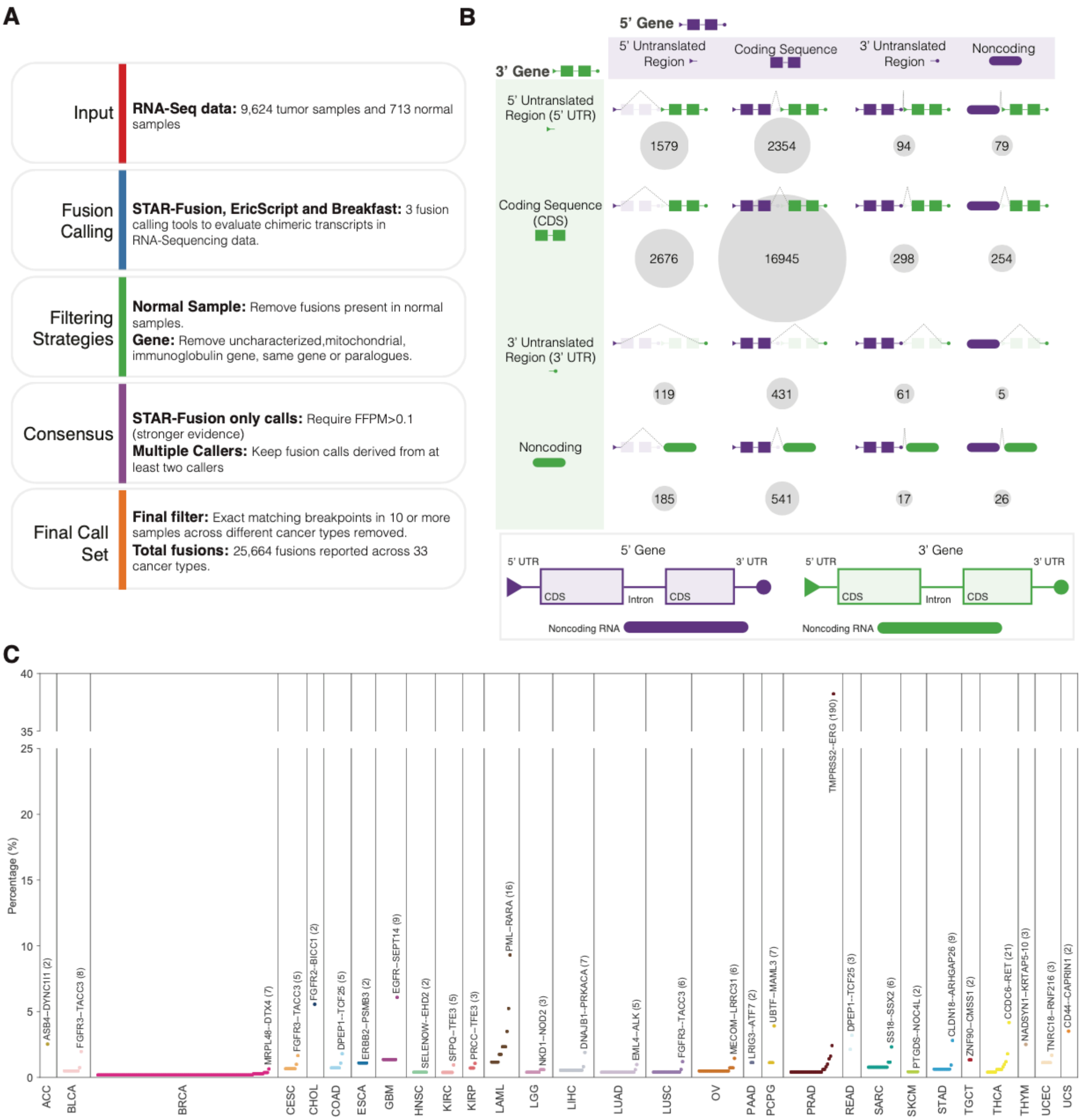


Figure 1 Fusion detection and landscape in cancer. (A) Fusion calling and filtering pipeline. (B) Cartoon overview of fusion gene partner breakpoints. Purple indicates the 5' gene partner and green indicates the 3' gene partner. For both the 5' and 3' gene partner, fusion gene breakpoints can occur in the following genomic regions: 5' untranslated region (5'UTR, triangle), coding sequence (CDS, rectangle), 3'UTR (circle), and noncoding region (rounded rectangle). For each fusion event, a dotted line connects the breakpoints in the 5' and 3' gene partners to create the predicted fusion and the circle size, while number represents the total fusion events classified into the associated fusion category. (C) The dot plot shows the frequency of recurrent fusions found in each cancer type. The most recurrent fusion in each cancer type is labeled. Cancer types without recurrent fusions are not shown.

Fusion landscape across 33 cancer types

Categorizing the 25,664 fusions based on their breakpoints, we found that the majority of breakpoints are in coding regions (CDS) of both partner genes (Fig. 1B). Surprisingly, there are many more fusions in 5’ UTRs compared to 3’ UTRs for both partner genes, given that 3’ UTRs are generally longer (Mann-Whitney U Test, $p<2.2e-16$). This could be explained by having more open chromatin in the 5’ UTR region [63], the larger number of exons in 5’ UTRs than 3’UTRs (Mann-Whitney U Test, $p<2.2e-16$) [64], but could also indicate some regulatory mechanisms, e.g. alternative usage of the promoter region of a partner gene.

For different cancer types, the total number of fusions per sample varies from 0 to 60, with a median value of one (Figure S1). Cancer types having the fewest number of fusions per sample are kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), low grade glioma (LGG), pheochromocytoma and paraganglioma (PCPG), testicular germ cell tumors (TGCT), thyroid carcinoma (THCA), thymoma (THYM), and uveal melanoma (UVM), each with a median of zero. Other cancer types show a range of medians between 0.5 and 5 fusions per sample, although most samples demonstrate zero or only one inframe, disruptive fusion relevant to oncogenesis.

Frequencies of recurrent fusions found in each cancer are illustrated in Figure 1C (Table S1). The most recurrent example within any cancer type was *TMPRSS2--ERG* in prostate adenocarcinoma (PRAD, 38.2%). We found *FGFR3--TACC3* to be the most recurrent fusion in BLCA (2.0%), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, 1.7%), and lung squamous cell carcinoma (LUSC, 1.2%). Other top recurrent fusions include

EML4--ALK in lung adenocarcinoma (LUAD, 1.0%), *CCDC6--RET* in THCA (4.2%), and *FGFR2--BICC1* in cholangiocarcinoma (CHOL, 5.6%).

Fusion gene expression in oncogenes and tumor suppressors

Fusion events may be associated with altered expression of one or both of the fusion gene partners, a well-known example being multiple myeloma tumors in which translocation t(4;14) fuses the highly-expressed IGH locus with the tyrosine protein kinase *FGFR3* [65]. We integrated gene expression, copy number, and fusion annotations to systematically test for associations between gene expression and fusion status.

For each fusion having an oncogene, kinase, or tumor suppressor (Table S2), we determined whether that sample was an expression outlier for that gene and subsequently examined resulting percentages of both under- and overexpressed genes in each cancer type (Table S3). Figure 2A shows that between 6% (mesothelioma, MESO) and 28% (KIRP) of kinase fusions displayed outlier overexpression of the kinase partner. Oncogenes tended to show higher likelihoods of overexpression, while tumor suppressors displayed lower likelihoods. Between 3% (breast invasive carcinoma, BRCA) and 38% (PCPG) of tumor suppressor gene fusions showed outlier under expression, generally higher than both oncogenes and kinases.

Figure 2B illustrates the median percentile expression level of the most highly recurrent oncogenes and tumor suppressors involved in fusions (Table S3). Samples with fusions involving oncogenes, such as *EGFR*, *ERBB2*, and *RET,* showed increased expression of those genes relative to samples without fusions across cancer types. Most tumor suppressor genes (TSGs) showed inconsistent patterns of expression across cancer types. However, the global trend for TSGs is decreased expression compared to non-fusion samples.

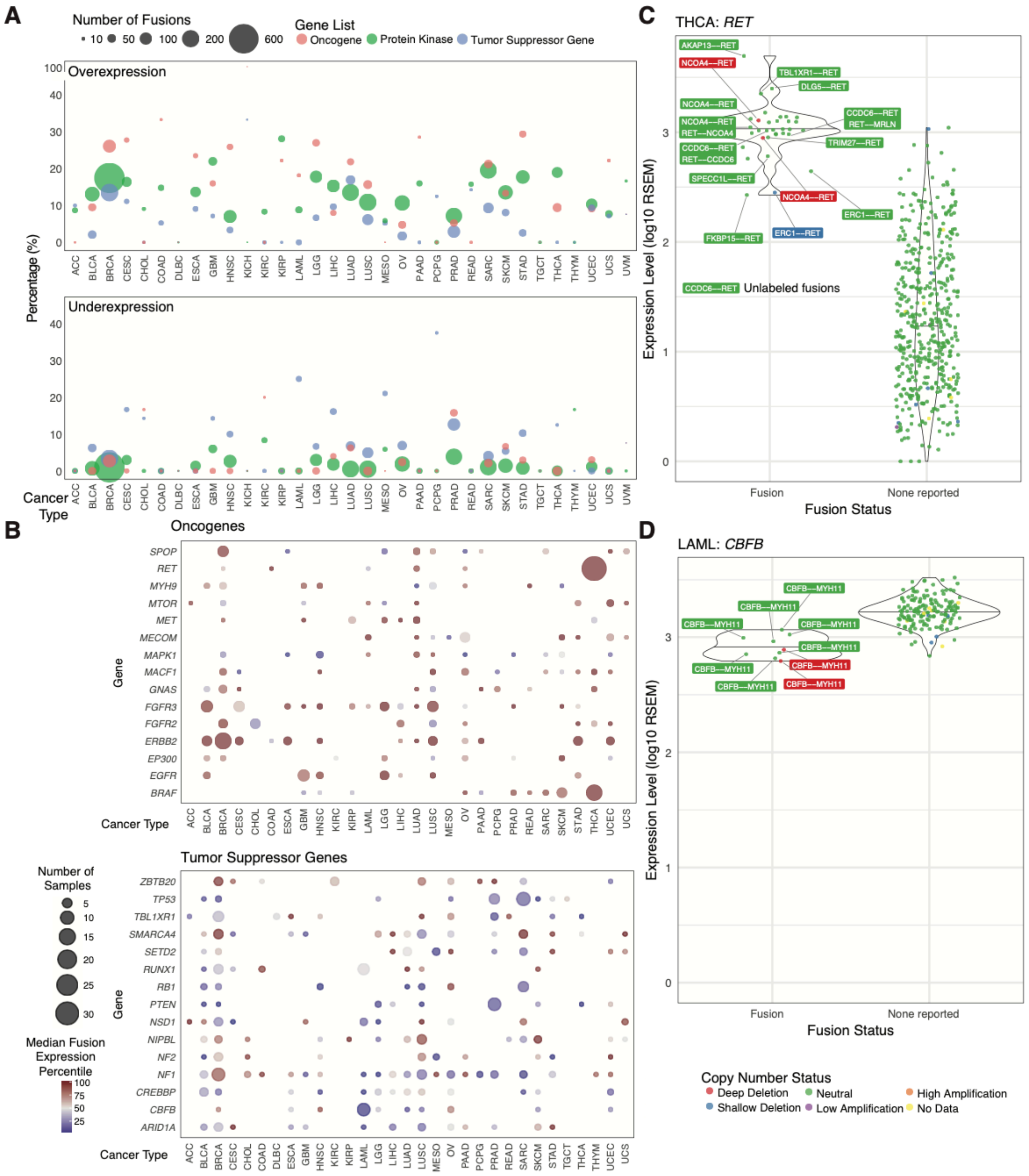


Figure 2 Fusion expression outliers. (A) The dot plot indicates the percentage of fusions called in which one of the partner genes is an expression outlier (overexpression or underexpression). The size of the dot corresponds to the number of fusions called in each cancer type. Color corresponds to genes of interest coming from lists of oncogenes, protein kinases, and tumor suppressor genes. (B) The dot plot shows the relative expression level of samples with fusions compared to those without fusions. Each sample has a particular expression percentile at a given gene, and color indicates the median percentile of samples with a fusion in that gene. Genes are the fifteen most recurrent oncogenes and tumor suppressor genes. Size corresponds to the number of samples in each cancer type with a fusion at that gene. (C)-(D) Expression of samples at *RET* and *CBFB* in thyroid carcinoma (THCA) and acute myeloid leukemia (LAML), respectively. Color indicates a categorical copy number ranging from deep deletion to high amplification.

We also examined the relationship between TSG mutations and fusions to determine whether frequently-fused TSGs were also disrupted by other mutation types. A variety of patterns were noted. For example, *TP53* is affected by mutations rather than fusions in most cancer types. However, in sarcoma (SARC), both fusions and mutations affecting TP53 were detected. In acute myeloid leukemia (LAML), several *CBFB* fusions but no mutations were observed, yet other cancer types also exhibited *CBFB* mutations (Table S3, Figure S2). Our results suggest that alternative mechanisms are utilized by tumor cells in a cancer type-specific manner.

We also observed associations between fusion status and expression level in well-known fusions (Table S3), such as *RET--NTRK1* in thyroid cancer, *EML4--ALK* in lung cancer [37], and *DNAJB1--PRKACA* in the fibrolamellar carcinoma subtype of liver cancer [50]. *RET* fusions in thyroid carcinoma (THCA) and lung adenocarcinoma (LUAD) are inframe protein kinase fusions with overexpression of the 3' *RET* oncogene (Fig. 2C). Recurrent *CBFB--MYH11* fusions in LAML are significantly associated with decreased expression of the tumor suppressor *CBFB*, which functions as a transcriptional regulator [66] (Fig. 2D).

In breast cancer, copy number amplification is a well-known mechanism of *ERBB2* over-expression and treatment of these HER2+ patients with trastuzumab is an established and effective targeted therapy [67]. Interestingly, three out of four samples with *ERBB2* fusions and two samples without a called fusion showed HPV integration within 1Mb of *ERBB2* [68]. *ERBB2* fusion gene partners *PPP1R1B* and *IKZF3* are genomic neighbors of *ERBB2*, suggesting that these fusions could be a by-product of local instability, potentially induced by the viral integration and subsequent breakage fusion events. By careful analysis of the association between fusions and expression, we have identified strategies for improving both sensitivity and specificity of fusion calls.

Structure and spectrum of kinase fusions

Some oncogenic kinase fusions are susceptible to kinase inhibitors [37], suggesting that additional therapeutic candidates might be discovered by examining fusion transcripts involving protein kinase genes. In total, we detected 2,892 such events, comprising 1,172 with kinase at the 3' end (3'-kinase), 1,603 with kinase at the 5' end (5'-kinase), and 117 with both partners being kinases (both-kinase) (Fig. 3A and Table S4). Analysis of the catalytic kinase domains using the UniProt/PFAM domain database (STAR Methods) showed that 1,275 (44.1%) kinase fusions retained an intact kinase domain (Fig. 3A). We further predicted open reading frames for these fusions and separated them into three categories with respect to the frame of the 3' gene: inframe, frameshift, and no frame information (e.g. breakpoint at UTR, intron, or non-coding RNA). In general, there were more inframe fusions than frameshift fusions, especially for 3'-kinase fusions, because preserving the reading frame is required to keep the kinase domain intact. For subsequent kinase analyses, we focused only on those 1,275 fusions having intact domains, further classifying the both-kinase group into 3'-kinase or 5'-kinase based on the position of the intact domain.

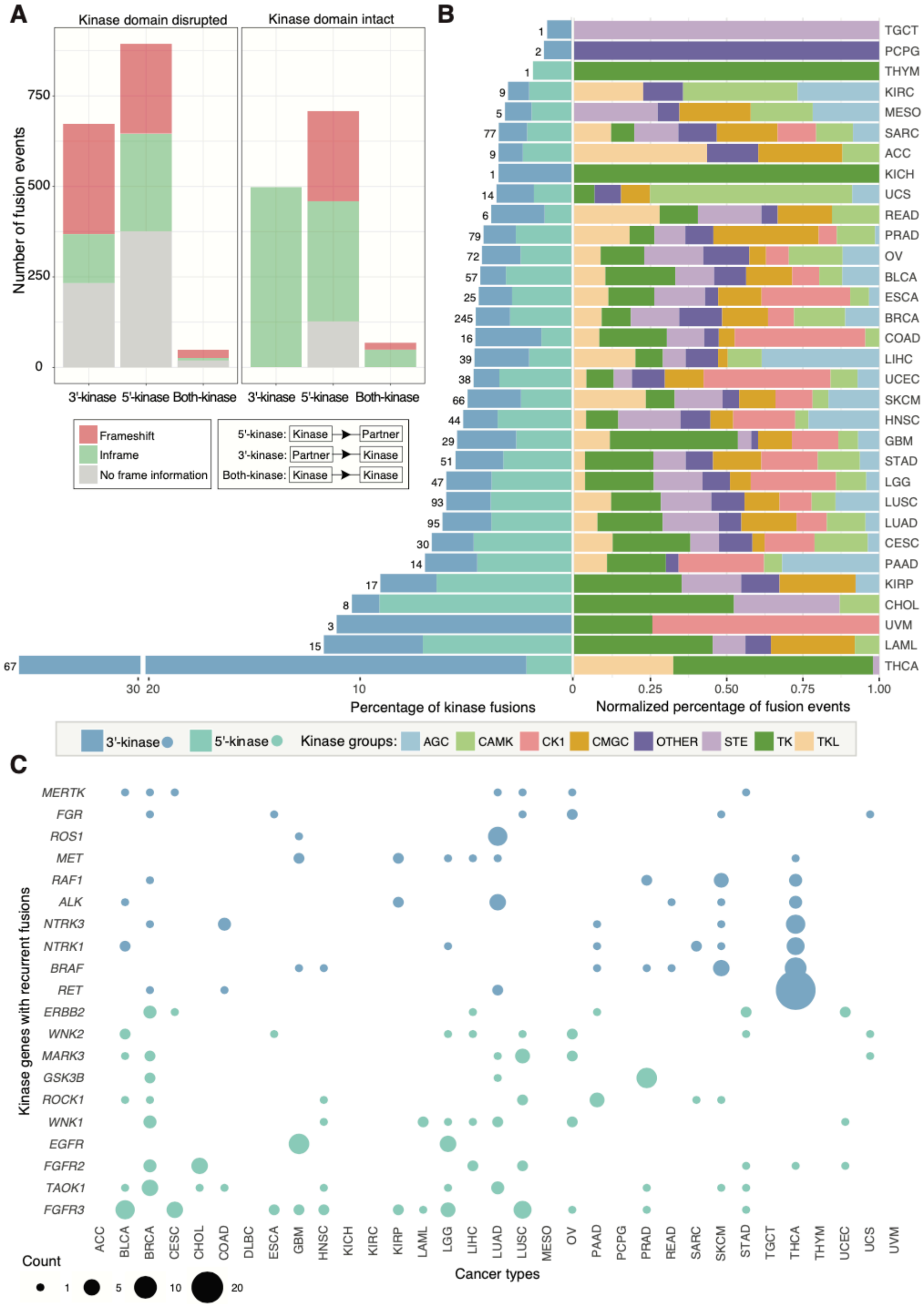


Figure 3 Protein kinase fusions. (A) The bar chart indicates the number of protein kinase fusions with the kinase at the 5’ or 3’ end, inframe or frameshift, and kinase domain intact or disrupted. (B) The left bar plot shows the percentage of samples with kinase fusions across different cancer types. The number of samples with a kinase fusion is also indicated at the end of each bar. 5’ kinase and 3’ kinase fusions are marked in light green and blue, respectively. The right bar plot shows the normalized percentage of kinase fusions broken down by kinase groups. (C) The dot plot shows the numbers of samples for recurrent fusions across different cancer types. 5’ kinase and 3’ kinase fusions are marked in light green and blue, respectively.

Comparison of kinase fusions across different cancer types indicated that kinase fusions are significantly enriched in thyroid carcinoma (THCA, 35.6%, Fisher's Exact Test, $p < 2.2e{-}16$) (Fig. 3B). Moreover, the majority were 3'-kinase fusions (94.0%), a significantly higher percentage than what we observed in other cancer types (Fisher's Exact Test, $p < 2.2e{-}16$). We further divided these fusions into eight categories based on different kinase groups, including AGC, CAMK, CK1, CMGC, STE, TK, TKL. In general, we found that the percentages of different categories vary across cancer types (Fig. 3B). For example, there are more TK fusions in THCA and GBM, more CK1 fusions in uterine corpus endometrial carcinoma (UCEC), colon adenocarcinoma (COAD), and esophageal carcinoma (ESCA), and more AGC fusions in liver hepatocellular carcinoma (LIHC). Across different cancer types, we found an enrichment of TK and TKL kinase fusions for 3'-kinases, but no strong preference for 5'-kinases (Figure S3).

Recurrent kinase fusions are of great interest as potential drug targets. Overall, we detected 744 5'-kinase and 531 3'-kinase fusions. Of these, 147 and 99 were recurrent, respectively, mostly across cancer types rather than within cancer types (Figure S3). As expected, fusions in the *FGFR* kinase family (*FGFR2* and *FGFR3*) are the most frequent 5'-kinase fusions, given their high recurrence in individual cancer types (Fig. 3C). *WNK* kinase family fusions (*WNK1* and *WNK2*) were also detected in multiple cancer types. The *WNK* family is phylogenetically distinct from the major kinase families, and there is emerging evidence of its role in cancer development [69]. Here, we found a total of 23 WNK family fusions, most of which resulted in higher expression of *WNK* mRNA (Figure S4). The increased expression was not generally accompanied by copy number amplification; for example, neither *WNK1* nor *WNK2* were amplified in ESCA or LIHC. Incidentally, *ERC1--WNK1* was also detected recently in an independent Chinese esophageal cancer cohort [70]. For 3'-kinase fusions, all the top 10 kinase genes are tyrosine kinases, most of which are enriched in THCA, including *RET*, *BRAF*, *NTRK1*, *NTRK3*, *ALK,* and *REF1* (Fig 3C). *FGR* fusions were found in 7 samples the same partner gene *WASF2*, 5 of which showed higher expression of *FGR* gene. In these five samples, the breakpoints for the two genes are the same (5'UTR of both genes) resulting in usage of the stronger *WASF2* promoter for the *FGR* gene. Interestingly, recurrent *MERTK* fusions are singletons in each individual cancer type with *TMEM87B* and *PRKACA* fusions are only observed in liver cancer with *DNAJB1* (Figure S3).

To further understand the regulation of kinase fusions, we compared the gene expression patterns between the kinase gene and partner gene. There are in total 1,035 kinase fusions with both gene expression and copy number data available. To control for the effect of copy number amplification on gene expression, we focused on the fusions with copy numbers between 1 and 3, including 439 5'-kinase and 339 3'-kinase fusions (Fig. 4A-B). For 5'-kinase fusions, the kinase gene expression quantiles are uniformly distributed, indicating that the kinase gene expressions in the samples with fusion are not significantly different from the samples without fusion (Fig. 4A). However, 3'-kinase genes tend to show higher expression in samples with a fusion compared to the ones without. To explain this, we classified the fusion events into three categories based on the relative expression pattern between the kinase gene and its partner in samples from the same cancer type. Most (66.7%, 293/439) 5'-kinase fusions showed lower expression in the partner gene compared to the kinase. In contrast, 70.5% (239/339) of 3'-kinase fusions showed higher partner expression (Fig. 4A-B). Moreover, those 3'-kinase fusions involving a more highly expressed 5' partner also show higher kinase expression (Fig. 4C). For example, we found a *TRABD--DDR2* fusion in one head and neck squamous cell carcinoma

(HNSC) sample, which fused the stronger *TRABD* promoter with *DDR2*, resulting in its overexpression (Fig. 4D). This patient could potentially be treated using dasatinib, which targets overexpressed *DDR2* in HNSC [71]. *DDR2* fusions were also detected in another 9 samples from 5 different cancer types, which could be treated similarly given sufficient DDR2 overexpression (Table S1).

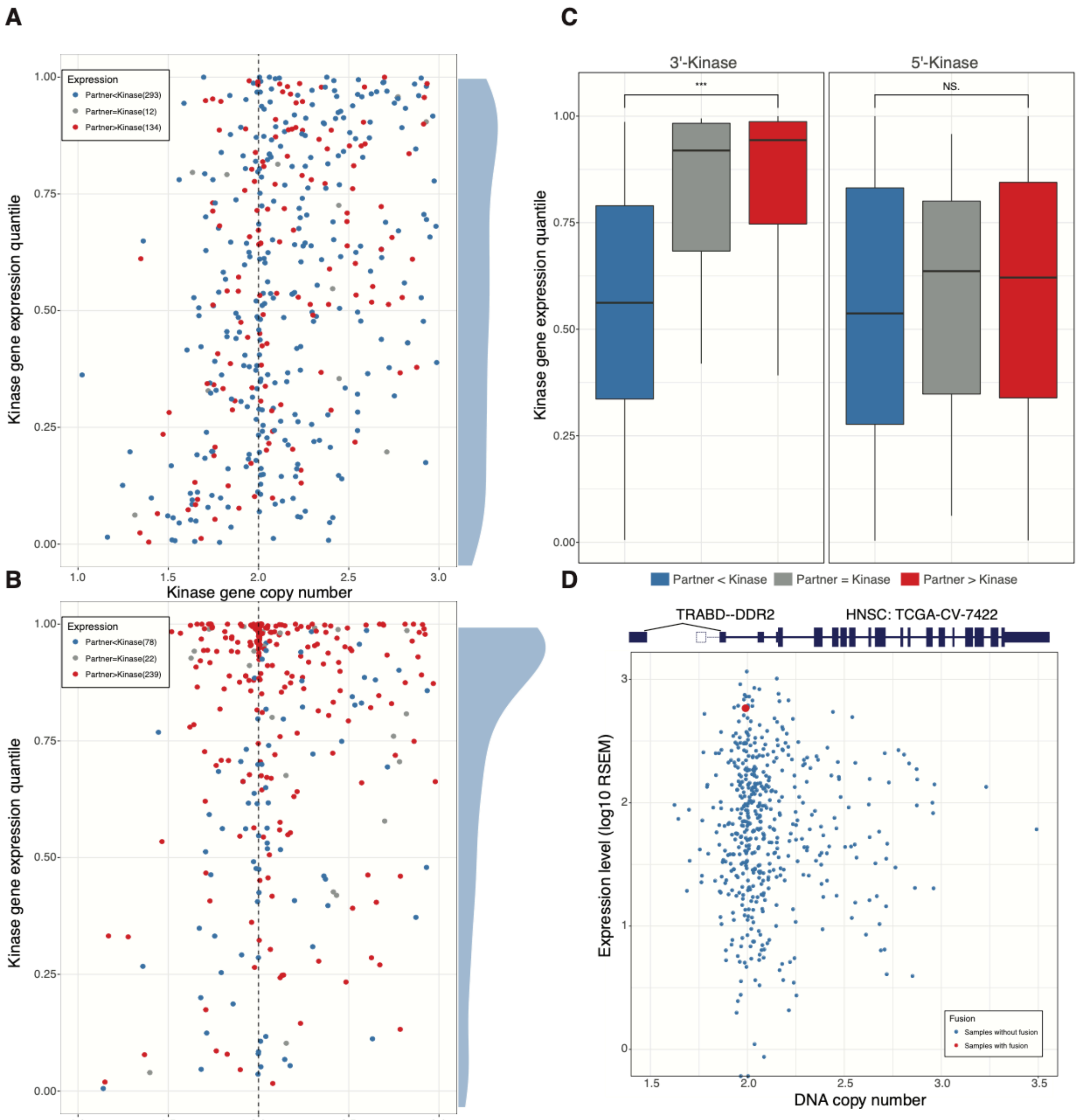


Figure 4 Kinase gene expression regulated by fusion. (A) The scatterplot shows the gene expression quantile (y-axis) for the 5'-kinase without copy number variation (between 1 and 3 copies, x-axis). All genes are classified among three categories: kinase expression higher, equal, and lower, as compared to partner expression, marked in blue, grey, and red, respectively. The density plot for expression quantile is also shown on the right panel. (B) The scatterplot shows the gene expression quantile (y-axis) for the 3'-kinase without copy number variation (between 1 and 3 copies, x-axis). The colors represent the same three categories as (A). The density plot for expression quantile is also shown. (C) Boxplot comparing the distribution of kinase gene expression quantile between the three groups defined in (A) for 5'-kinase and 3'-kinase, respectively. (D) Schematic of *TBABD--DDR2* fusion gene structure in a HNSC sample, and scatter plot of *DDR2* copy number versus mRNA expression in HNSC. The samples with and without this fusion are marked in red and blue, respectively.

Mutual exclusivity between fusions and mutations

While mutations in oncogenes or tumor suppressors may lead to tumorigenesis, fusions involving those genes are also an important class of cancer driver events. We systematically profiled mutations and fusions in 299 cancer driver genes [7] (Table S2) to assess the contributions of fusion genes in carcinogenesis in the 8,963 TCGA patients that overlap between the mutation call set (Public MC3 MAF [72], Key Resources Table) and our fusion call set. We characterized patients as having a driver mutation, a mutation in a driver gene, and/or a driver fusion (fusion involving a driver gene).

Although the majority of cancer cases have a known driver mutation (48.6%, mean 6.8 mutations) or mutations in a driver gene (28.1%, mean 4.2 mutations), we found 8.3% have both a driver mutation and driver fusion event (mean 5.5 mutations and 1.2 fusions), 6.4% have both a mutation and fusion in a driver gene (mean 4.2 mutations and 1.3 fusions), and 1.8% have a driver fusion only (mean 1.1 fusions) (Fig. 5A). This distribution is consistent with the notion that only a few driver events are required for tumor development [9].

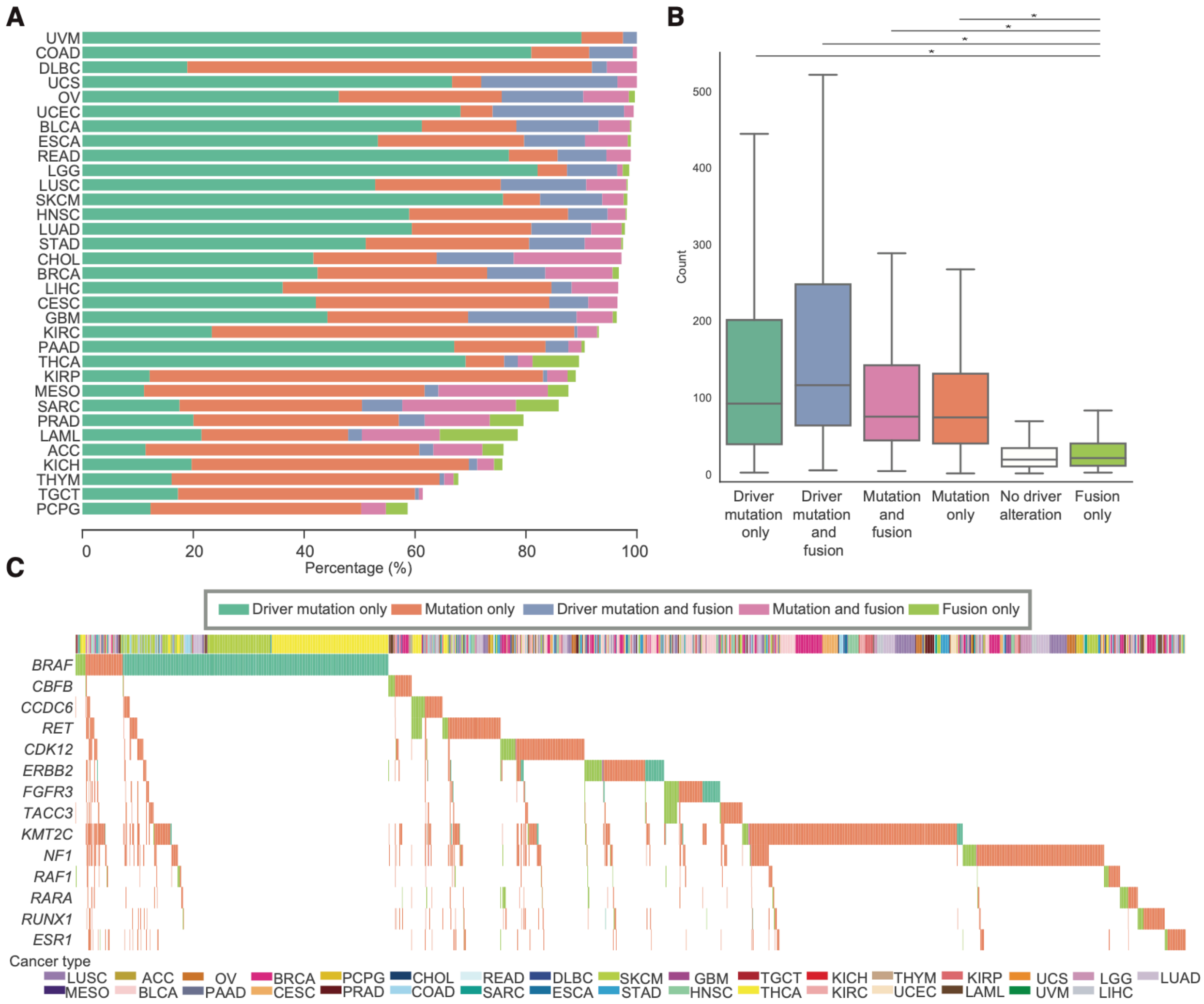


Figure 5 Mutual exclusivity between driver mutations and driver fusions. (A) The bar plot shows the percentages of samples with driver mutations only (green), mutations only (orange), driver mutation and fusion (blue), mutation and fusion (pink), or fusion only (light green) events in 299 cancer driver genes. (B) Distribution of mutation burden across each alteration group designated in all figures. (C) All samples with fusions or mutations in any of the genes indicated on the left are displayed on the x-axis. For each gene, samples are clustered by the alteration group. Bottom bar indicates cancer type.

We further examined the total number of mutations for samples and observed a low mutational burden in the group with driver fusion only, which is comparable with the group with no driver alterations (Fig. 5B). The significant decrease in the numbers of mutations (Mann-Whitney U Test, p<2.2e-16) reflects the functionality of fusions across multiple cancer types. Moreover, within cancer types, we observed a range of 0.2% (HNSC) to 14.0% (LAML) of tumors with fusions but no driver gene mutations. Among those LAML tumors that have fusions and no driver gene mutations, we identified several well-recognized fusions relevant to leukemia, such as *CBFB*--*MYH11* (number of samples=3), *BCR*--*ABL1* (n=2), and *PML*--*RAR* (n=2). We also identified the leukemia-initiating fusion *NUP98*--*NSD1* in two LAML tumors [73].

We then examined the relationship of fusions and mutations in the same driver gene (Fig. 5C). The result shows that when fusion events are present in a gene, mutations in the same gene are rarely found, supporting a pattern of mutual exclusivity of the two types of genomic alteration. This trend was observed across many patients and many cancer types. Our results suggest that a considerable number of tumors are driven primarily or solely by fusion events.

Contributions of fusions to cancer treatment

We investigated potentially druggable fusion events in our call set using our curated Database of Evidence for Precision Oncology (DEPO; Sun, et al. submitted) (Table S5). We defined a fusion as druggable if there is literature supporting the use of a drug against that fusion, regardless of cancer type (allowing for "off-label" drug treatment). We found potentially druggable fusions across 29 cancer types, with major recurrent druggable targets in PRAD (*TMPRSS2*, 205 samples), THCA (*RET*, 33 samples), and LAML (*PML*--*RARA*, 16 samples) (Fig. 6A). *FGFR3* was a potential target (both on-label and off-label) in 15 cancer types. Overall, we found 6.0% of samples (574/9,624 samples) to be potentially druggable by one or more fusion targeted treatments. Further study of fusions in human cancer will facilitate the development of precision cancer treatments.

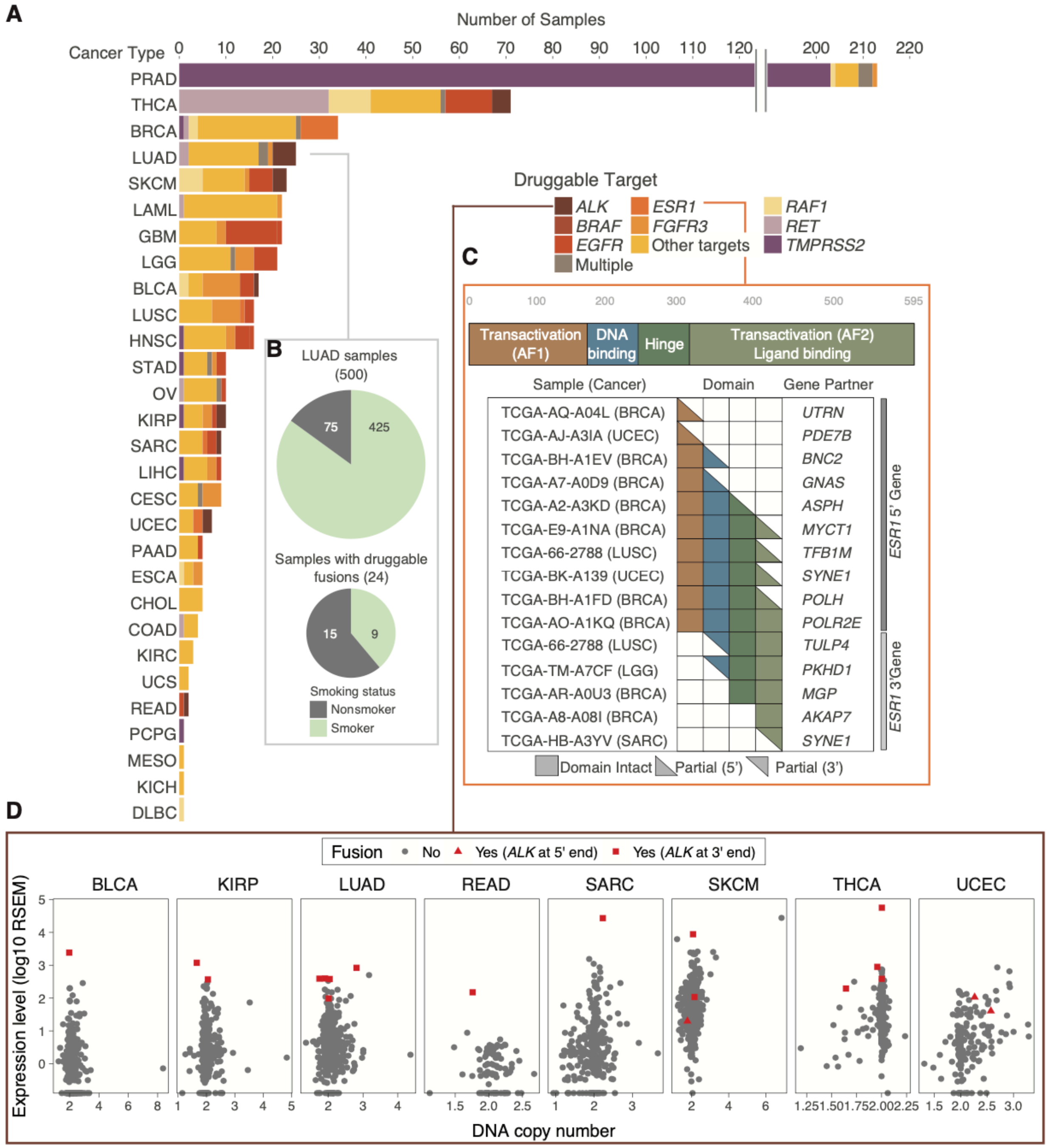


Figure 6 Druggable fusion targets. (A) The bar chart indicates the number of samples potentially treatable based on their fusion status. (B) Percentages of LUAD samples with known smoking status. (C) *ESR1* domains kept in *ESR1* fusions across cancer types. (D) *ALK* expression across cancer types indicating *ALK* fusion status.

We analyzed patterns of fusion druggability in LUAD, stratifying by smoking status. In this data set, 15% of LUAD samples (75 out of 500 samples with known smoking status) were never smokers, while a significantly higher percentage of never smokers (15 out of 75 samples) vs. smokers (9 out of 425 samples) were found to have a druggable fusion (Chi-square test, p<1e-6) (Fig. 6B). Several FDA approved drugs exist to target *ALK* fusions in lung and other cancer types. We observed *ALK* fusions in 20 samples from 8 cancer types (5 samples in LUAD). In most cases, fusion status corresponded to copy number neutral overexpression of *ALK* (Fig. 6D). In 17 out of 20 cases, *ALK* was the 3' partner of the fusion pair, with *EML4* being the most frequent 5' partner (7 out of 17).

ESR1 encodes an estrogen receptor with important and druggable relevance to breast cancer [74]. We detected *ESR1* fusions in 16 samples from 5 different cancer types (9 samples from BRCA). Of the 9 BRCA samples, 8 are known be from the Luminal A or B subtypes. We observed strict mutual exclusivity between *ESR1* mutations and fusions (Fig. 5C). Of the 16 fusions, 11 have *ESR1* at the 5' end, and 5 at the 3' end. When *ESR1* is the 5' gene in the fusion, the transactivation (AF1) domain is always included (Fig. 6D). When *ESR1* is the 3' gene, the transactivation (AF2) domain is always included. Those samples with *ESR1* fusion tend of have higher *ESR1* expression, especially in the 9 BRCA samples (Figure S5). Similarly, *ESR1* expression is higher when *ESR1* is mutated in BRCA, CESC, and UCEC, which are all hormone receptor related cancer types [75-77]. Further functional study to determine the mechanism of *ESR1* fusions could suggest drug development directions.

Immunotherapy based on tumor-specific neoantigens shows promise in treating cancer patients [78]. Gene fusions found in tumor cells can generate peptides, which may serve as neoantigen candidates. However, patients with known driver fusions may be poor candidates for immunotherapy due to their reduced mutational burden, especially without clear evidence of immune cell infiltration and overall immunogenicity. As an exploratory and speculative analysis, we investigated neoantigens produced by gene fusions [79]. On average, there were 1.5 predicted neoantigens per fusion across different cancer types (Figure S6 and Table S5). The mean number of predicted neoantigens per fusion ranged from 0.33 in KICH to 2.88 in THYM. We also compared the number of neoantigens for inframe and frameshift fusions (Figure S6). Results show that frameshift fusions can generate more immunogenic epitopes than inframe fusions (mean value: 2.2 vs 1.0), though nonsense mediated decay might reduce some of this potential difference.

We further investigated seven fusions for which there were at least four samples having one or more neoantigen candidates (Figure S6). In particular, *TMPRSS2--ERG*, *CCDC6--RET,* and *FGFR3--TACC3* have the highest number of samples with predicted neoantigen candidates. Our results show that the fusion product is only immunogenic in a small subset of patients, especially for *TMPRSS2--ERG* fusions. Again, without clear evidence of immune cell infiltration and overall immunogenicity, any fusion neoantigen analysis remains exploratory and speculative.

Discussion

In this study, we applied multiple RNA-Seq fusion callers, namely STAR-Fusion, EricScript, and Breakfast, followed by a stringent filtering strategy, to identify potential driver fusion events across 33 cancer types. We were able to successfully identify 95.5% of fusions reported in TCGA marker papers. While existing studies have published fusion calls across the TCGA cancer cohort [37,60], we have improved on prior analyses by integrating results across multiple fusion callers and by applying stringent filtering to derive a confident dataset of fusion events from 9,624 tumor samples. Importantly, we investigated the biology and evaluated the significance of fusions in the cancer context. Of the 25,664 fusions we detected, 18.2% could be tested for validation using available whole-genome sequencing data, leading to a 63.3% validation rate.

By integrating gene expression, copy number, and fusion annotation data, we evaluated the biological and therapeutic implications of fusion events. Kinase and oncogene related fusions tended to be overexpression outliers, while fusions involving tumor suppressor genes showed the opposite effect overall. When comparing fusion events to the remainder of the cancer cohort, fusions involving oncogenes such as *EGFR*, *ERBB2*, and *RET* had increased expression. Overexpressed fusions, especially inframe kinase fusions, are commonly targeted for therapy due to their susceptibility to kinase inhibitors.

For all 2,892 kinase fusions, we translated the resulting peptide sequence, finding that 1,275 had functional catalytic kinase domains. Comparison of kinase fusions across different cancer types showed that THCA has significantly more kinase fusions, most of which were 3' kinase fusions. In addition to well-known recurrent fusions like *FGFR3--TACC3*, we also detected 245 kinases with recurrent fusions to different partner genes, which may ultimately prove to be successful drug targets.

We showed that a meaningful percentage of patients (16.8%) harbor fusions involving cancer driver genes but have no driver gene mutations. Notably, 6.0% of cancer patients could potentially benefit from existing drugs targeting fusion products. Moreover, our analysis also highlights an important consideration for immunotherapy treatment in patients with fusions. The significant decrease in mutational burden observed in patients with fusions in driver genes points toward a reduced efficacy of immunotherapy in these patients, despite fusion peptides themselves potentially being good immunogenic targets. Many fusions are already known to be drug targets.

Our study demonstrates the necessity of performing fusion analysis across multiple cancer types. Our approach integrated the results of multiple fusion calling algorithms, lending confidence to fusions with lower levels of RNA-seq read support that might otherwise have been discarded. We sought to prioritize fusions relevant to cancer by highlighting their association with gene expression, potential for targeted therapy, and role in cancer hallmark pathways. Fusion allele frequency is an elusive measure from RNA-Seq data and tracking the clonal evolution of fusions within a tumor remains an exciting opportunity for study. Fusions play an increasingly appreciated role in tumorigenesis and progression and represent an important source of improved treatment options. Ultimately, our multi-tool, integrative bioinformatic detection approach helps to define the universe of fusions in cancer. Further, it reminds us that developing robust and widely applicable clinical diagnostic approaches that can document fusions across cancer types is vital. Such approaches are critical to identifying those patients who can benefit from both established treatments and clinical trials.

Methods

Dataset description

Aligned RNA-Seq bam files were analyzed using the ISB Cancer Genomics Cloud (https://isb-cgc.appspot.com/). These 33 cancer types included in this study are adrenocortical carcinoma [ACC], bladder urothelial carcinoma [BLCA], brain lower grade glioma [LGG], breast invasive carcinoma [BRCA], cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC], cholangiocarcinoma [CHOL], colon adenocarcinoma [COAD], esophageal carcinoma [ESCA], glioblastoma multiforme [GBM], head and neck squamous cell carcinoma [HNSC], kidney chromophobe [KICH], kidney renal clear cell carcinoma [KIRC], kidney renal papillary cell carcinoma [KIRP], acute myeloid leukemia [LAML], liver hepatocellular carcinoma [LIHC], lung adenocarcinoma [LUAD], lung squamous cell carcinoma [LUSC], lymphoid neoplasm diffuse large B-cell lymphoma [DLBC], mesothelioma [MESO], ovarian serous cystadenocarcinoma [OV], pancreatic adenocarcinoma [PAAD], pheochromocytoma and paraganglioma [PCPG], prostate adenocarcinoma [PRAD], rectum adenocarcinoma [READ], sarcoma [SARC], skin cutaneous melanoma [SKCM], stomach adenocarcinoma [STAD], testicular germ cell tumors [TGCT], thymoma [THYM], thyroid carcinoma [THCA], uterine carcinosarcoma [UCS], uterine corpus endometrial carcinoma [UCEC], and uveal melanoma [UVM]. The sample set consists of 10,331 total TCGA samples, 9,624 tumor samples, and 713 normal samples.

Level-3 gene expression (RSEM) and segment-based copy number data were downloaded from Broad GDAC firehose (https://gdac.broadinstitute.org) (version: 2016_01_28). Gene-based copy number data were obtained by intersecting with RefSeq gene annotation bed file (version: 2013-07-27). Mutation calls were provided by the Multi-Center Mutation Calling in Multiple Cancers (MC3) working group within TCGA [72] (Key Resources Table).

Fusion detection and filtering

TCGA RNA-Seq data were downloaded from Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu) and analyzed using the ISB Cancer Genomics Cloud (https://isb-cgc.appspot.com/). For each sample, the fastq file was mapped to the human genome (build 38) followed by fusion calling using STAR-Fusion (parameters: --annotation --coding-effect), EricScript (default parameters) (https://sites.google.com/site/bioericscript/) and BREAKFAST (two different minimum distance cut-offs were used: 5 kb and 100 kb) (https://github.com/annalam/breakfast). STAR-Fusion showed higher sensitivity in detecting the fusions reported in previous TCGA studies. Therefore, we focused on the STAR-Fusion output and integrated EricScript and BREAKFAST output in one of the following filtering steps: 1) an exclusion list of genes was curated, including uncharacterized genes, immunoglobulin genes, mitochondrial genes, etc. Fusions involving these genes were filtered; 2) Fusions from the same gene or paralogue genes (downloaded from https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/tree/master/resources) were filtered; 3) Fusions reported in normal samples were filtered, including the ones from TCGA normal samples, GTEx tissues, and non-cancer cell study [62]; 4) For the fusions reported by only STAR-Fusion, a minimum value of FFPM > 0.1 (fusion fragments per million total reads) was required, as suggested by the authors; for the fusions reported by two or more callers, no minimum FFPM was required. 5) Finally, fusions with the same breakpoints in ≥10 samples across different cancer types were removed unless they were reported in previous TCGA studies.

Validation of fusion transcripts

For fusion events where low-pass whole genome sequencing data or whole genome sequencing (WGS) data were available from the ISB Cancer Genomics Cloud (https://isb-cgc.appspot.com/), we obtained high quality (-q 20) reads mapping to each partner gene and the 100kb region up and downstream using SAMtools. At least 3 discordant reads from WGS were required to determine if the fusion prediction was validated.

Gene expression analysis

We collected gene expression, copy number, and fusion annotations to test for associations between gene expression and fusion status. We used Tukey's definition of outliers to determine if the expression level at a given gene was an outlier or not. An overexpression outlier means the sample's expression level at a given gene was greater than (75th percentile) + 1.5*IQR, where IQR is the interquartile range. An underexpression outlier means the sample's expression level at that gene was less than (25th percentile) - 1.5*IQR. To test for a significant association between expression and fusion status, we calculated p-values using both a t-test and Fisher's Exact Test. If either of those results passed stringent FDR multiple test correction, three or more fusions were reported, and if the median expression of the fusions was in the top or bottom decile of the data, we reported those genes for manual review.

Protein kinase fusion analysis

We curated a list of kinase genes from previous publications and public databases (Table S5). Then we compared this list with UniProt/PFAM domain database (http://www.uniprot.org/database/DB-0073) to retain the ones with an annotated kinase domain. For the fusions involving kinase genes, we used AGFusion (https://github.com/murphycj/AGFusion) to check whether the annotated kinase domain was still present in the fusion transcript to separate them into fusions with an intact kinase domain versus those with a disrupted kinase domain. We compared the breakpoint positions in each fusion with the annotation file to check whether the breakpoint was in the 5'UTR, CDS, or 3'UTR region. Kinase genes are classified into eight groups: AGC, CAMK, CK1, CMGC, STE, TK, TKL, and others based on the PhosphoSite Database [80]. The percentage of kinase genes in each group across different cancer types was defined as the number of kinase genes with fusions in each group divided by their sum, denoted as p_g. For each cancer type, the number of kinase genes in each group was first normalized by p_g, denoted as n_g. Then each number was divided by their sum $n_g / \sum n_g$ to calculate a normalized percentage of kinase genes in each group.

Neoantigen prediction

For each predicted fusion, we obtained translated protein sequences for novel transcripts from STAR-Fusion. The wild-type protein sequences are obtained from Ensembl Database. We constructed different epitope lengths (8-11mer) from the translated protein sequence. Each sample's HLA type comes from the TCGA Pan-Cancer Immune Group (Synapse ID: syn5974636). We predicted the binding affinity between epitopes and the major histocompatability complex (MHC) using NetMHC4 [79]. Epitopes with binding affinity ≤ 500nM which are also not present in the wild-type transcript are reported as neoantigens. We required at least 5 splitting reads for supporting junctions to filter fusions with low expression.

Mutual exclusivity analysis

For TCGA tumor samples where both MC3 [72] (Key Resources Table) mutation calls and gene fusion calls were available, we obtained the genetic alteration events, including fusion, inframe deletion, inframe insertion, missense mutation, nonsense mutation, nonstop mutation, splice site mutation, and translation start site mutation in 299 driver genes. We separated all the genomic alterations and events into "driver mutation", "mutation", and "fusion" categories, and compiled a genomic alteration profile for each sample. To test if the total number of mutations are significantly different among groups, we took samples without mutations in the following genes: *POLE, MLH1, MLH3, MGMT, MSH6, MSH3, MSH2, PMS1,* and *PMS2*, to exclude the confounding factor stemming from microsatellite instability. We then calculated p-values by using Mann-Whitney U Test.

DEPO

DEPO is a curated list of druggable variants filtered such that each variant corresponds to one of several categories: single nucleotide polymorphisms or SNPs (missense, frameshift, and nonsense mutations), inframe insertions and deletions (indels), copy number variations (CNVs) or expression changes. Each variant/drug entry in DEPO was paired with several annotations of potential interest to oncologists. DEPO is available as a web portal (http://dinglab.wustl.edu/depo).

Chapter 3: Evolution and structure of clinically relevant gene fusions in multiple myeloma

Our work, Evolution and structure of clinically relevant gene fusions in multiple myeloma, has been accepted for publication in *Nature Communications* (2020). Please refer to the publication for supplementary table information. Contributions: As first author, SMF developed the fusion calling and filtering pipeline, analyzed all data, made all figures, wrote the manuscript, and was responsible for all revisions.

Abstract

Multiple myeloma is a plasma cell blood cancer with frequent chromosomal translocations leading to gene fusions. To determine the clinical relevance of fusion events, we detect gene fusions from a cohort of 742 patients from the Multiple Myeloma Research Foundation CoMMpass Study. Patients with multiple clinic visits enable us to track tumor and fusion evolution, and cases with matching peripheral blood and bone marrow samples allow us to evaluate the concordance of fusion calls in patients with high tumor burden. We examine the joint upregulation of *WHSC1* and *FGFR3* in samples with t(4;14)-related fusions, and we illustrate a method for detecting fusions from single cell RNA-seq. We report fusions at *MYC* and a neighboring gene, *PVT1*, which are related to *MYC* translocations and associated with divergent progression-free survival patterns. Finally, we find that 4% of patients may be eligible for targeted fusion therapies, including three with an *NTRK1* fusion.

Introduction

Fusions are a type of somatic alteration leading to cancer associated with up to 20% of cancer morbidity.[5,6] Translocations, copy number changes, and inversions can lead to fusions, dysregulated gene expression, and novel molecular functions. Fusions occur and have oncogenic roles in hematological, soft tissue, and solid tumors. Fusion rates differ across cancer types, and fusions may define some cancer types, such as *BCR--ABL1* in chronic myeloid leukemia (CML). A balanced translocation t(9;22) leads to *BCR--ABL1*, producing a hybrid protein with constitutive *ABL1* kinase domain activation, signaling cell division and avoiding apoptosis. Imatinib inhibits the *BCR--ABL1* protein hybrid and in 2001 became the first FDA-approved drug to specifically target a fusion protein.[5]

Multiple myeloma (MM) is the second most common blood cancer (10% of blood cancers, 1-2% of all cancers) and involves the clonal proliferation of bone marrow plasma cells, which are fully differentiated B cells. B cells produce a diverse repertoire of antibodies through genomic alterations at immunoglobulin (Ig) loci, including VDJ recombination, somatic hypermutation, and class switch recombination. Aberrant class switch recombination may result in translocations upregulating oncogenes. Ig enhancers get repurposed to drive oncogene expression, myeloma tumorigenesis, and clonal expansion.[81]

Tumor initiating genomic changes may already be present at the pre-malignant stages of MM include monoclonal gammopathy of undetermined significance (MGUS) and smouldering MM (SMM). Primary genomic events in MM distinguish patient groups having hyperdiploidy (HRD, approximately 50%) and non-hyperdiploidy (non-HRD). Non-HRD patients typically have a different primary event, like an Ig translocation. *CCND1* (chr11) and *WHSC1* (chr4) are the two most common translocation partners of IGH (chr14). Patients may have both HRD and translocation events, and secondary events like t(8;14) dysregulating *MYC* are associated with progression.[65,82]

Previous studies used RNA-seq to catalogue fusion events from over 9,000 patients and 33 cancer types from The Cancer Genome Atlas (TCGA).[35,37,38] False positives due to library preparation or bioinformatic errors must be filtered. Overlapping fusion calls from multiple tools can establish concordance. Low expression or low quality RNA may cause false negatives, and translocations may affect expression but not produce detectable fusion transcripts. In myeloma, plasma cell Ig gene expression dominates the transcriptome and masks lower expression fusions. Multi-omic approaches with DNA and RNA resolves some limitations.[5]

Large-scale sequencing efforts to understand multiple myeloma have demonstrated genomic heterogeneity beyond primary copy number and translocation events.[83-86] Several fusion detection studies show complementary results. Cleynen, *et al.* detected gene fusions from 255 newly diagnosed MM patients, finding significant relationships between fusions and gene expression, hyperdiploidy, and survival, and identifying recurrent fusion gene partners.[87] Nasser, *et al.* analyzed MMRF CoMMpass RNA-seq data, reconstructed Tophat-Fusion transcripts, and validated fusions with WGS. [88] Lin, *et al.* used targeted RNA-seq in 21 MM patients, finding several novel fusions with disease relevance.[89] Morgan, *et al.* used targeted sequencing of kinases to understand how translocations dysregulate kinase activity in MM.[90]

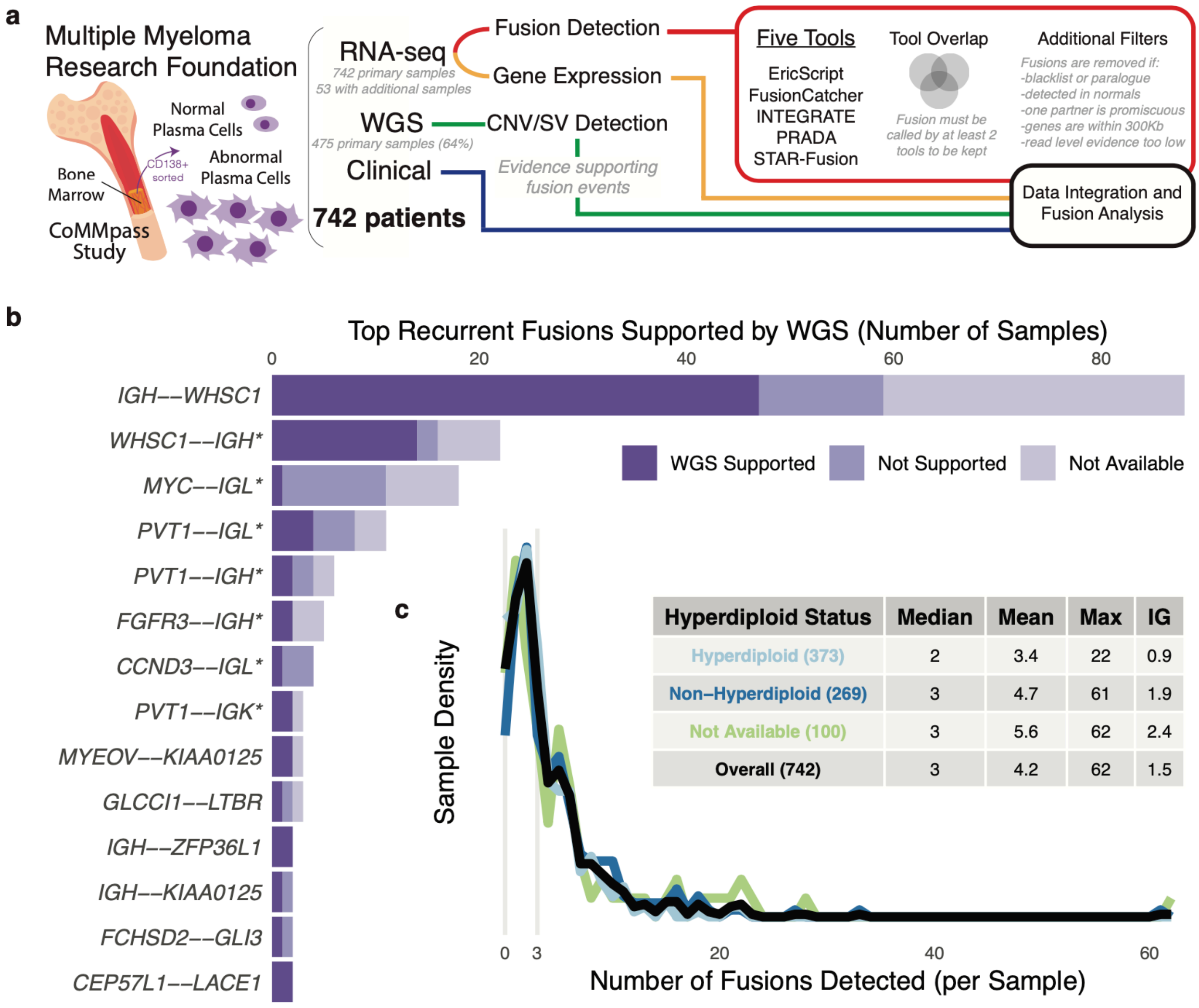
Here, we extend previous efforts by focusing on the clinical implications and evolution of fusions across multiple time points. We leverage RNA and DNA sequencing as well as clinical data types to analyzed fusion genes we detected from the Multiple Myeloma Research

Foundation (MMRF) CoMMpass Study. We analyze fusion genes and gene expression patterns from 742 multiple myeloma patients (806 samples). Patient samples from serial clinic visits enable tumor evolution profiles using fusions and mutations. Further, from patients with both bone marrow and peripheral blood samples collected at the same time, we quantify the concordance of their fusion profiles. We demonstrate fusion event detection at single cell resolution using barcoded scRNA-seq data, pointing to future development of fusion methods. We explore the prognostic relevance of fusions by analyzing progression-free survival and find that those with *IGH--WHSC1* or *PVT1--IGL* fusions have significantly worse outcomes. 4% of patients have a fusion annotated as a drug target in a public database.

Results

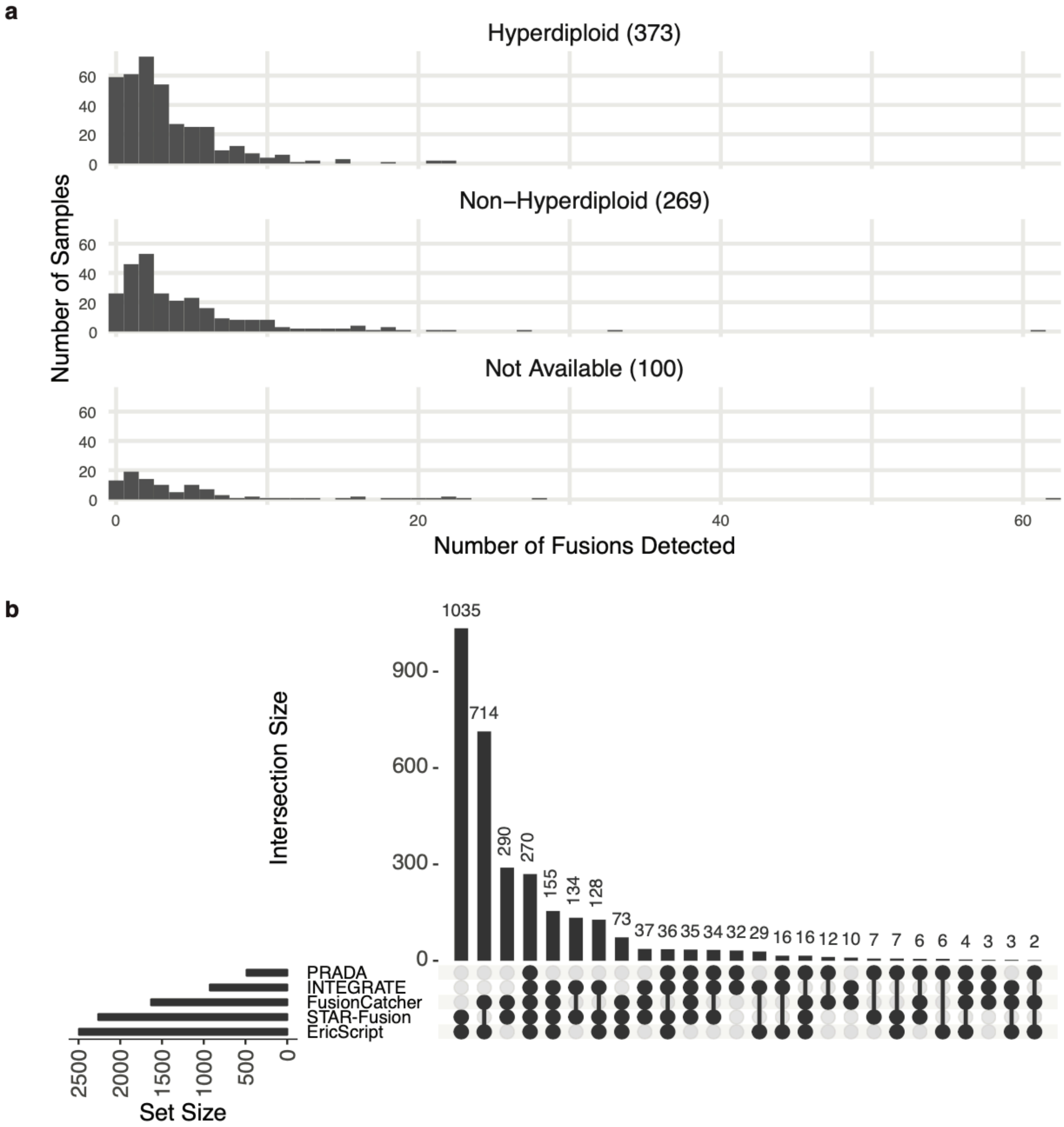
Fusion calling pipeline and clinical characteristics

We detected gene fusions from 742 patients from the Multiple Myeloma Research Foundation (MMRF) CoMMpass Study (see **Data availability**), combining RNA and DNA sequencing data with clinical information to form a landscape of fusion events (Fig. 1, Supplementary Figure 1, Supplementary Tables 1-3). We ran five fusion detection tools, implemented strict filtering criteria, and quantified gene expression to correlate with gene fusions (see **Methods**). We used WGS to detect structural variants and copy number changes potentially related to fusions. Sequencing-based FISH (seq-FISH) results showed major translocations and copy number changes such as hyperdiploidy.[91] We defined a primary sample for each patient as the earliest available sample and favored bone marrow (BM) over peripheral blood (PB) (740 BM, 2 PB). For 97.2% of patients (721/742 patients), the primary sample corresponded with the pre-treatment clinic visit. 53 patients had RNA-seq from multiple samples (BM and PB from the same visit or data from serial visits), for a total of 806 RNA-seq samples. Results come from primary samples only, unless otherwise stated.



Hyperdiploid Status	Median	Mean	Max	IG
Hyperdiploid (373)	2	3.4	22	0.9
Non–Hyperdiploid (269)	3	4.7	61	1.9
Not Available (100)	3	5.6	62	2.4
Overall (742)	3	4.2	62	1.5

Figure 1. Overview of pipeline and fusions reported. a. Project pipeline. **b.** Recurrent fusions with at least one sample's fusion supported by WGS. The asterisk (*) annotation refers to reciprocal fusions with the opposite orientation of the canonical fusion led by an Ig partner gene. **c.** Number of fusions detected per sample, stratified by hyperdiploid status. Source data are provided as a Source Data file.



Supplementary Figure 1. Related to Figure 1. a. Number of fusions detected per sample, stratified by hyperdiploid status. **b.** Fusion caller overlap after filtering. Source data are provided as a Source Data file.

The cohort ranged from 27 to 93 years old (median 63) (Supplementary Table 1). Patients were spread evenly across ISS Stage, with 34.7% of patients from Stage I (247/711 patients with annotated stage), 35.7% Stage II (254/711), and 29.5% Stage III (210/711). Follow-up for progression-free survival ranged from 8 days to 5.7 years (median 2.23 years) with 60.3% of patients progressing (402/667 patients with PFS). Follow-up for overall survival ranged from 8 days to 6.43 years (median 3.19 years) with 27.4% of patients dying (182/665 patients with OS). ISS Stages I, II, and III patients had median PFS of 3.85 years, 2.47 years, and 1.76 years, respectively. 58.1% of patients showed a hyperdiploidy (373/642 patients with annotated HRD status). 77.1% of patients had ancestry reported as White (512/664 patients with annotated ancestry), 15.8% Black (105/664), and 7.1% Other (47/664). Most patients were treated initially with a proteasome inhibitor (bortezomib or carfilzomib) and an immunomodulatory drug (IMID) (68.4%). Others received a proteasome inhibitor-based regimen (25.9%) or an IMID (5.7%). 41.4% of patients received a bone marrow transplant (305/737 with transplant annotated) during first-line therapy. Supplementary Table 1 summarizes clinical information.

Immunoglobulin gene fusions are most frequent

IGH--WHSC1 was the most common fusion reported; it results from t(4;14) typically observed in 15% of patients.[65] *IGH--WHSC1* or *WHSC1--IGH* were found in 12.4% of samples (92/742 samples). 79.7% of *IGH--WHSC1* fusions showed WGS support (47/59 patients with WGS data) (see **Methods**). Figure 1b shows the top recurrent fusions with at least one fusion supported by WGS. Ig fusions (*IGH*, *IGK*, or *IGL*) were reported frequently (35.6%, 1102/3094 fusions) with upregulated partner genes.

Our pipeline reported fusions between Ig loci and *MYC* or its downstream neighbor *PVT1*. *MYC* or *PVT1* was usually the 5' gene and paired with IGH, IGK, or IGL, including 18 samples with *MYC--IGL*, 11 with *PVT1--IGL, 6* with *PVT1--IGH*, and 3 with *PVT1--IGK* (Fig. 1b). One sample had *IGH--MYC* and one had *IGL--PVT1*. Past reports show *MYC* translocations with all three Ig loci.[92] However, previous multiple myeloma fusion studies hypothesized that *MYC* fusions with Ig would not be detected from RNA-seq if there were no hybrid transcript generated after the translocation.[87] Further study is necessary to determine whether these reported fusions are true events, biological by-products, or bioinformatic artifacts, and whether they confer functional or clinical significance. This will complement recent work showing the dysregulation of both *MYC* and *PVT1* in the presence of super-enhancer translocations.[93,94]

The number of fusions reported per sample ranged from 0 to 62 (median 3) (Fig. 1c, Supplementary Figure 1a), similar to breast, glioblastoma, ovarian, and prostate cancers from TCGA.[35] Hyperdiploid samples had significantly fewer fusions reported than non-HRD samples (HRD mean 3.4, non-HRD 4.7, Mann-Whitney U test two-sided p-value 6.71×10^{-3}). There were also significantly fewer Ig fusions between those groups (HRD mean 0.9, non-HRD 1.9, Mann-Whitney U test two-sided p-value 7.88×10^{-8}). We required two or more tools to agree upon a particular fusion call. We removed 18 highly recurrent IGL fusions with low WGS support (see **Methods**). After filtering, the overall WGS support rate was 22.3% (comparable to a previously reported pan-cancer support rate of 32.5% from samples with similar WGS coverage).[35] Most fusions were called by two tools (73.3%, 2269/3094), while 17.9% (555/3094) were called by three or four tools, and 8.7% were called by all five tools (270/3094) (Supplementary Figure 1b).

Fusion gene expression highlights multiple myeloma biology

Fusions may be associated with expression changes of the partner genes. We defined a sample's expression percentile for each gene as their expression level relative to primary samples at that gene (see **Methods**). The median fusion expression percentile of a gene is the median expression percentile of samples with a fusion involving that gene. We identified 51 genes significantly overexpressed in fusion samples (FDR < 0.05 or median fusion expression percentile > 0.9) (Supplementary Table 4). Of those, 9 are cancer-related genes from any cancer type annotated as a driver, drug target, kinase, oncogene, or tumor suppressor (Fig. 2a), including *FGFR3* (12 samples), *MAPKAPK2* (5), *MYC* (19), *NTRK1* (3), *PAX5* (3), *PIM3* (3), *RARA* (3), *TXNIP* (7), and *WHSC1* (97).[7] Expression levels may also identify samples with a false negative fusion call. 12 samples have outlier *WHSC1* overexpression but no *WHSC1* fusion reported, representing false negative *IGH--WHSC1* fusions or indicating samples with t(4;14) but no fusion product formed. Of those 12 samples, 50% (5/10 with seq-FISH) have a *WHSC1* translocation with expression percentile over 0.87. The tumor etiology of samples with high gene expression but no fusion calls may still involve upregulated gene activity. Since gene expression is itself relevant to cancer biology and drug targeting, fusion analysis should always be paired with gene expression.

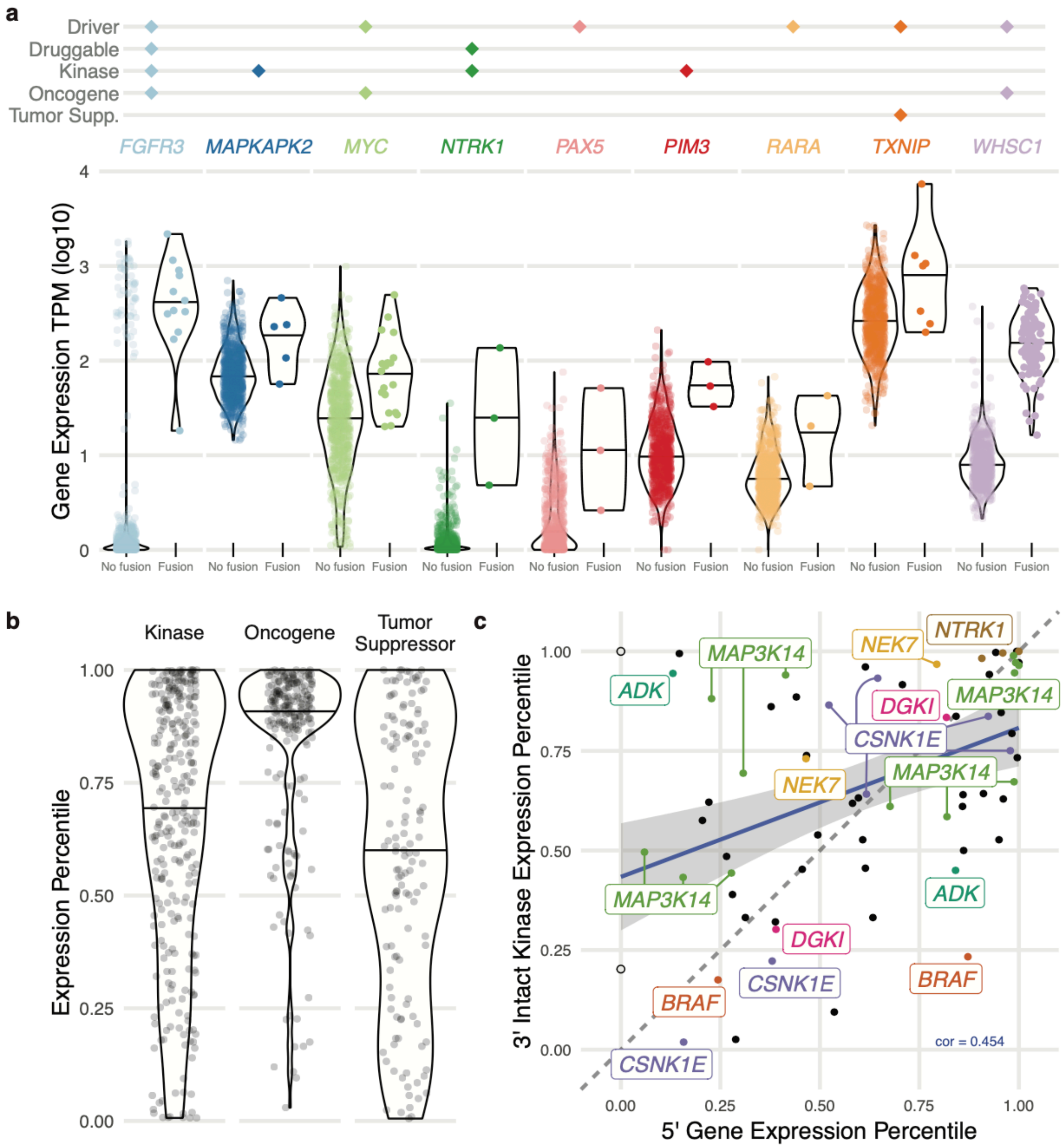
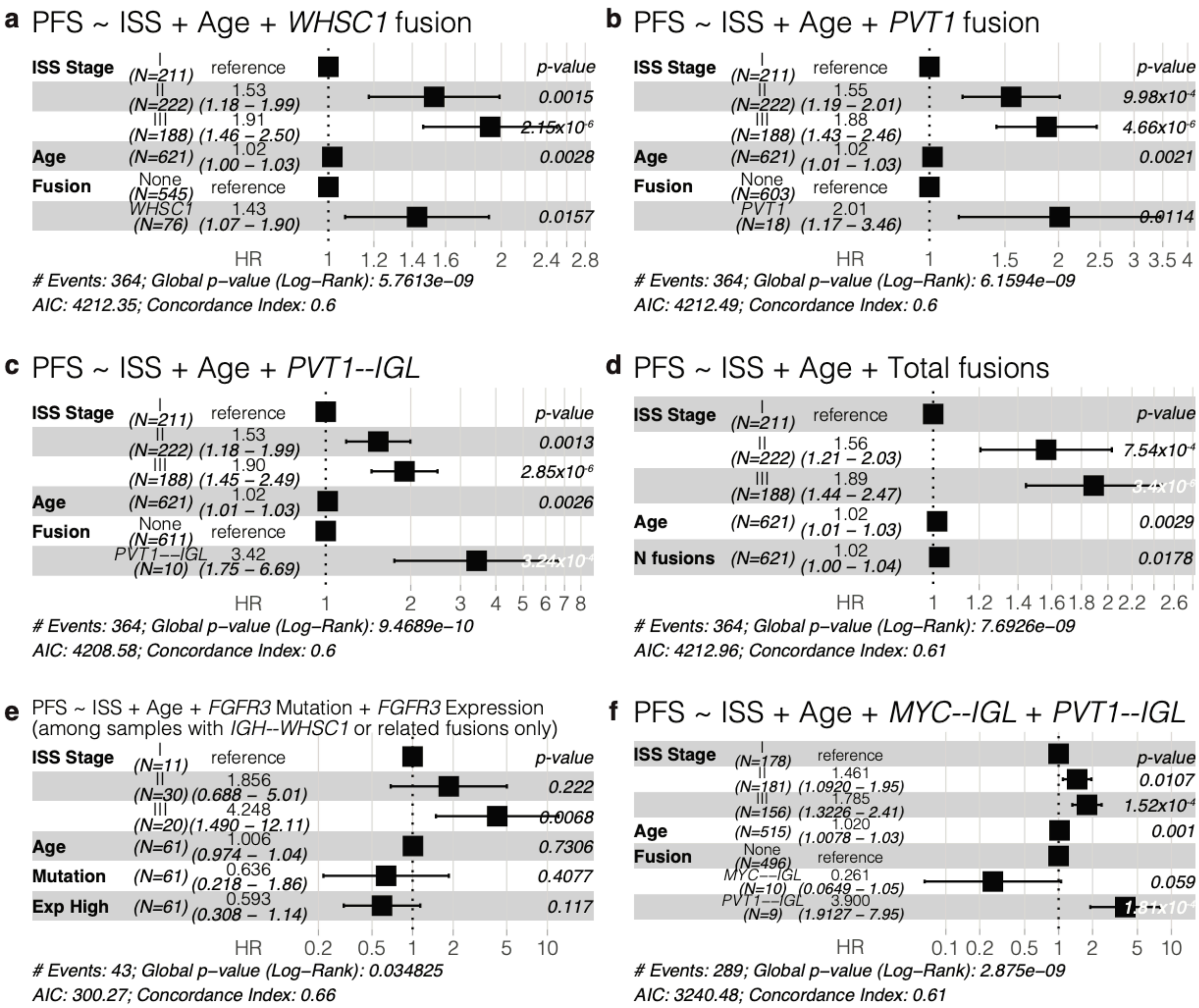


Figure 2. Expression of cancer-related genes. a. Significantly overexpressed fusion genes. **b.** Expression percentile distribution for different gene classes. **c.** 5' and 3' gene expression of fusions with intact 3' kinase genes. The two empty circles have no expression value for the 5' gene (IGH, IGL). Labels refer to genes appearing more than once. Source data are provided as a Source Data file.

Samples with fusions involving kinases, oncogenes, and tumor suppressors show different trends in expression levels of those genes (Fig. 2b). Gene expression of fusion oncogenes tended to be higher relative to other samples, fitting the biological context of oncogenes being deleterious when upregulated. Tumor suppressors, which may be disrupted in cancer in many different ways, displayed no trend of up- or downregulation. Kinases showed a skewed preference toward upregulation and are an important type of gene with implications for cancer development and drug targeting. We investigated the correlation between 5' and 3' partner gene expression when the 3' partner gene is a kinase and contains an intact kinase domain (see **Methods**) (Fig. 2c, Supplementary Table 5). In this subset of fusion partners, the positive correlation between 5' and 3' gene expression is somewhat higher than that of the overall background (0.454 vs. 0.352), indicating a pattern of selection for overexpressed kinase fusion partners. Recurrent 3' kinases with intact domains included: *MAP3K14* (13 patients), *CSNK1E* (7), *NTRK1* (3), *ADK* (2), *BRAF* (2), *DGK1* (2), and *NEK7* (2).

We tested for associations between clinical data (including age, sex, ancestry, ECOG performance, ISS stage, bone lesions, plasmacytoma, bone marrow plasma cell percentage, and LDH) and fusion genes observed in 3 or more samples (see **Methods**). After FDR correction and assessment of model fit, no clinical measures were significantly associated with fusion events. To understand the relationship between fusion events and prognosis in this cohort, we analyzed survival in patients with and without particular fusions or fusion genes. We created baseline PFS multivariate Cox proportional hazards models, including disease stage and patient age as covariates. For each fusion or fusion gene observed in 10 or more samples, we added the event as a covariate and tested for significant improvement in model fit using a chi-squared test. *WHSC1* and *PVT1* fusions were significantly associated with worse prognosis (Supplementary

Figure 2a-c). The estimated hazard ratio (HR) for a *WHSC1* fusion was 1.43 (95% CI 1.07-1.90; two-sided z-score p-value 0.0157). For *PVT1* fusions, the estimated HR was 2.01 (95% CI 1.17-3.46; two-sided z-score p-value 0.0114). For *PVT1--IGL* specifically, the HR estimate was 3.42 (95% CI 1.75-6.69; two-sided z-score p-value 0.000324). After including R-ISS and common translocations as covariates in the model, no fusion events or fusion genes were significantly associated with PFS, likely due to confounding introduced by translocation events directly associated with fusions. Total fusion burden was associated with worse prognosis; each additional fusion was associated with a slight decrease in PFS (HR estimate 1.02; 95% CI 1.00-1.04; two-sided z-score p-value 0.0178), after controlling for disease stage and patient age (Supplementary Figure 2d).



Supplementary Figure 2. Related to Figures 1, 5, 6. a-f. Progression-free survival (PFS) models using multivariate Cox proportional hazards. Error bars indicate a 95% confidence interval on each hazard ratio estimate. Covariate p-values derived from z-scores are two-sided. Source data are provided as a Source Data file.

Patients are stratified into risk groups by genomic events like amp(1q), del(17p), t(4;14), t(14;16), and t(14;20) using mSMART criteria.[95] Patients with multiple high risk events have worse prognosis.[96] Walker *et al.* identified a subgroup of patients with especially poor outcomes having biallelic *TP53* inactivation (for example, del(17p) and inactivating mutation) or Stage III disease and high copy number of *CKS1B* (1q21).[97] In our data, we defined a double hit group of patients with both amp(1q) and del(17p). The median PFS time for this group was 581 days (19 patients, 14 progressed). 5 patients with an additional t(4;14) event and *IGH--WHSC1* fusion had median PFS of 142 days (5 patients, 4 progressed). Ongoing research with larger sample sizes and longer follow-up will enable more robust survival modeling utilizing genomic events to define progression and overall survival risk.[98]

Fusions from multiple time points highlight tumor evolution

53 patients had additional samples allowing for within-patient comparisons across time (serial visits) or from different tissue sources (bone marrow, BM; peripheral blood, PB). 45 patients had BM samples from serial visits, and we compared fusions from the first two visits (Fig. 3a, Supplementary Figure 3). When initiating clonal fusion *IGH--WHSC1* was detected at the earlier visit, it was always detected at the later visit (6/6 patients). In one patient (1/39 patients), *IGH--WHSC1* was observed only at the later visit, but *WHSC1* expression at the earlier visit was above the 98th percentile, indicating a likely t(4;14) and false negative fusion call.

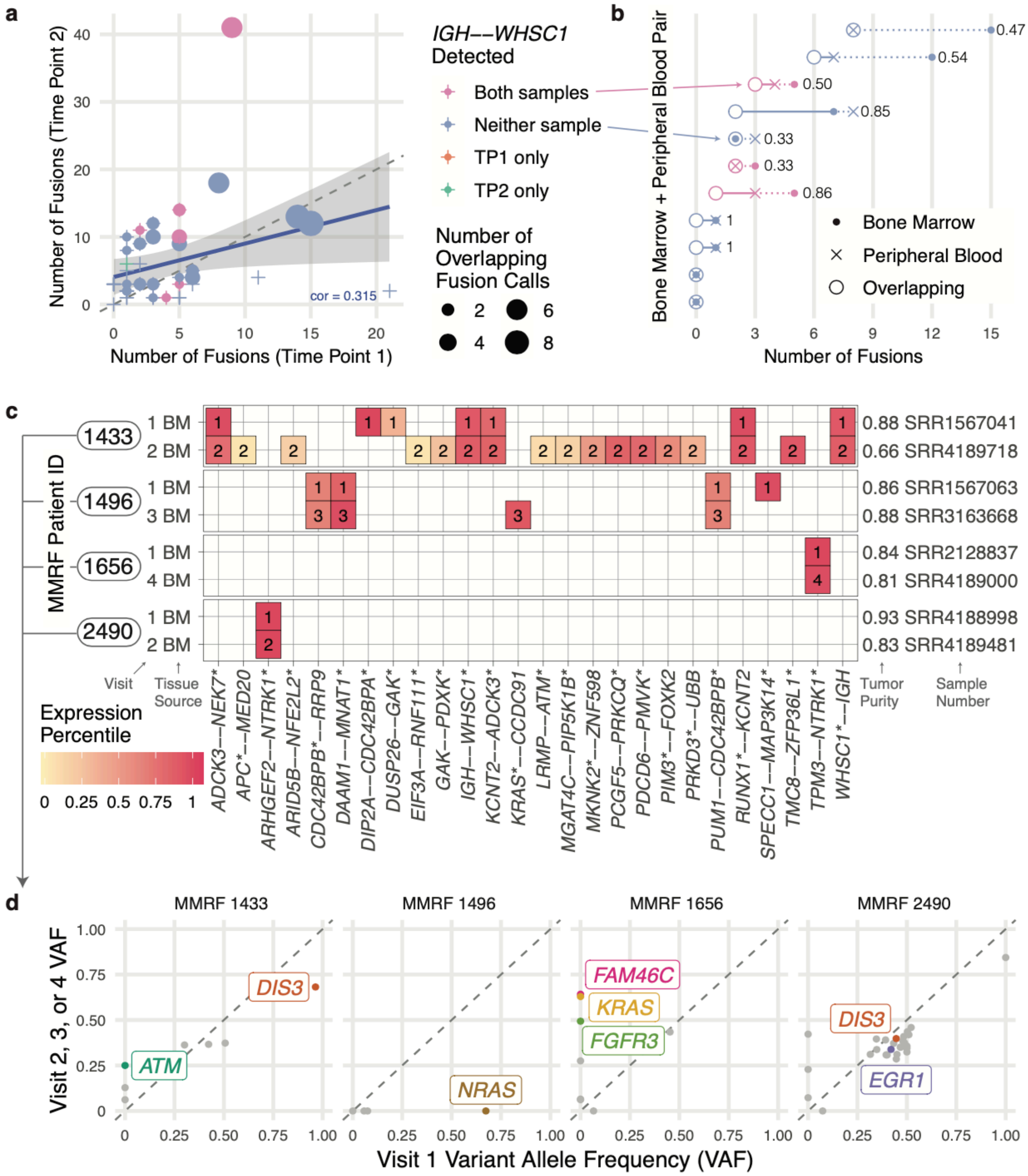
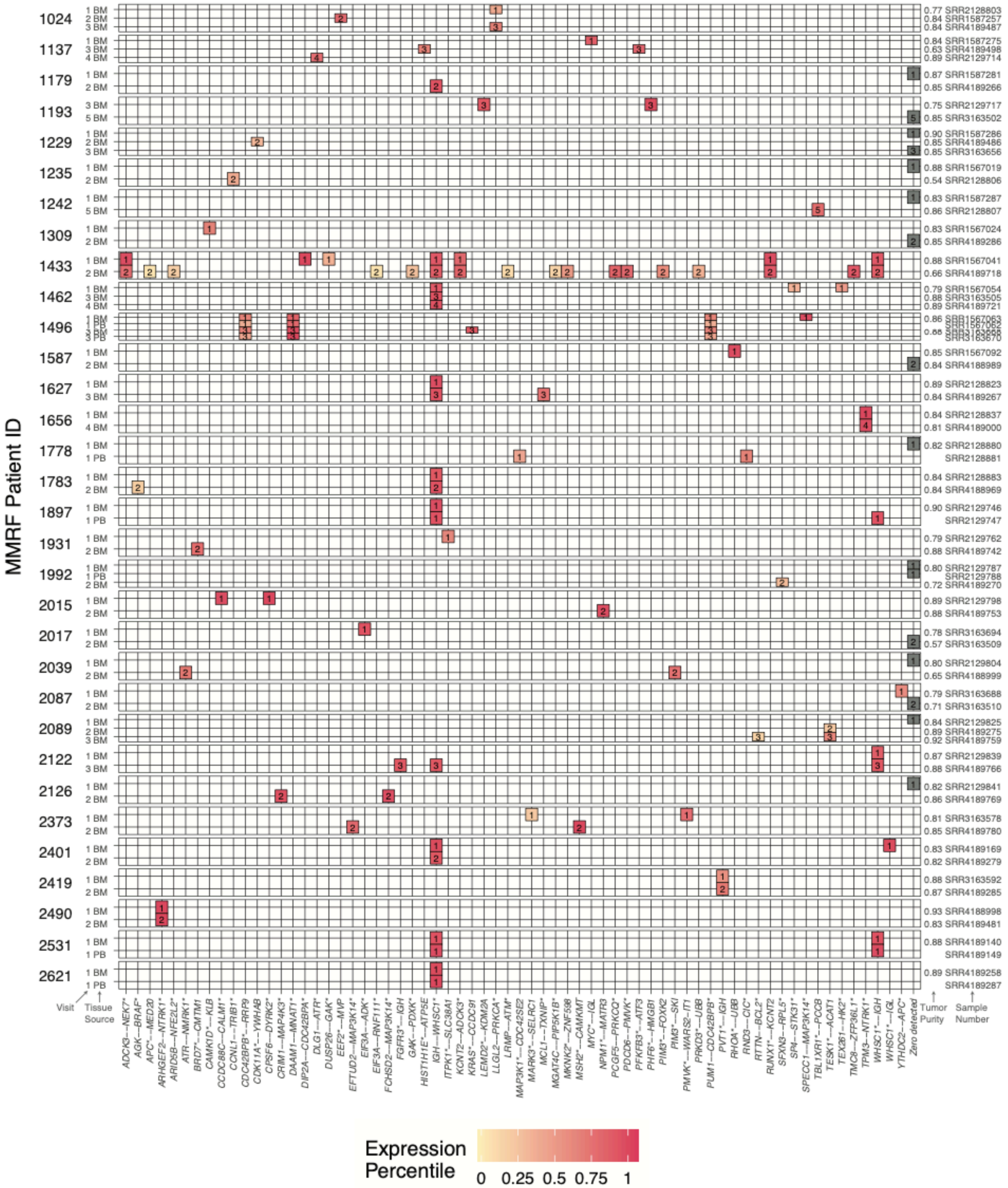


Figure 3. Fusions detected from multiple clinical samples and fusion evolution. a. Number of fusions called at serial clinic visits. Shaded region indicates a 95% confidence interval on the regression line. **b.** Overlap of fusions called from bone marrow (BM) and peripheral blood (PB) from the same clinic visit with normalized Hamming distance (range 0-1, 0 = perfect overlap, 1 = completely discordant). **c.** Fusions from cancer-related genes detected at serial clinic visits. **d.** Somatic mutations from cancer-related genes detected at serial clinic visits. Genes frequently mutated in multiple myeloma are labeled. Source data are provided as a Source Data file.



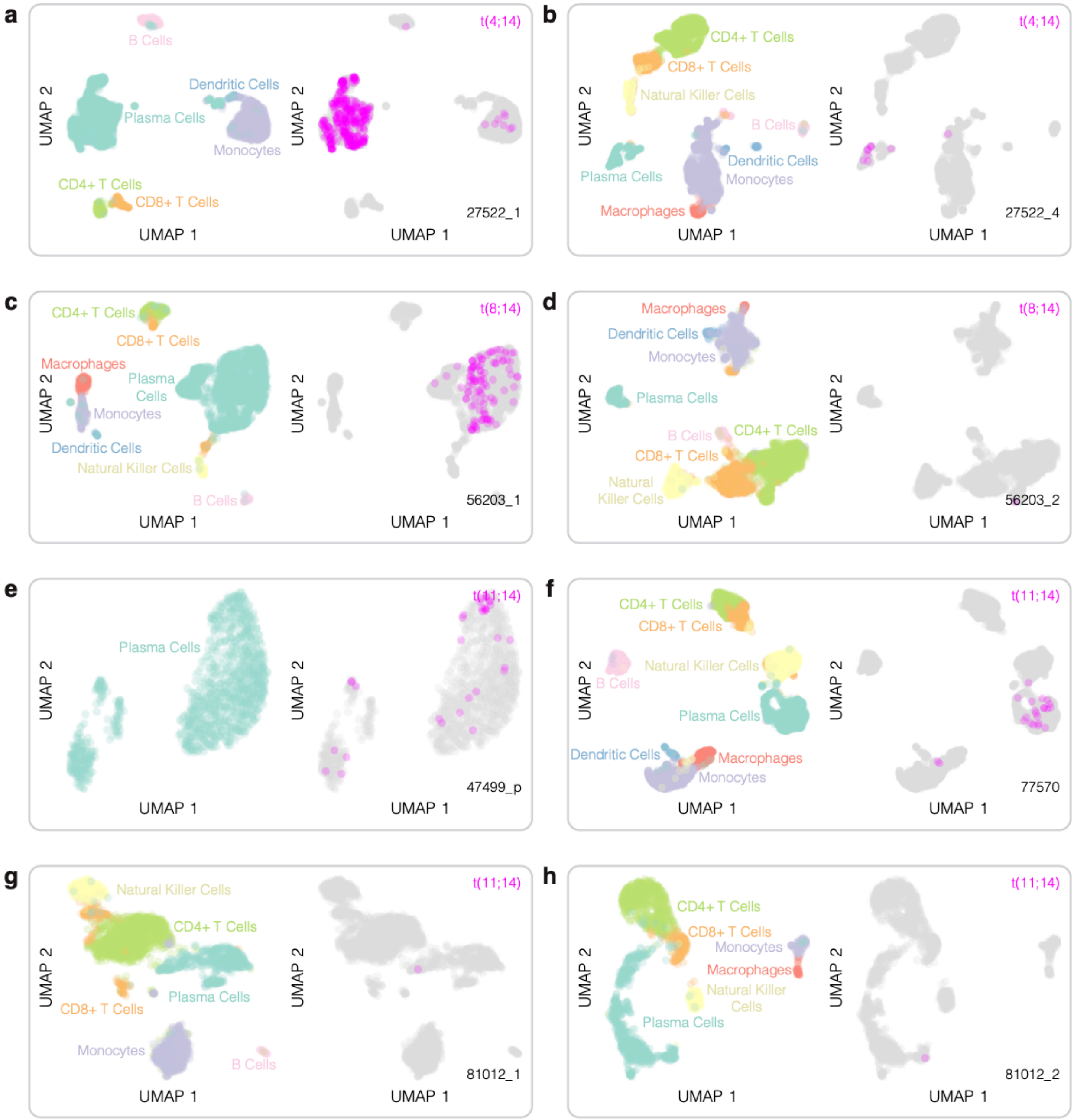
Supplementary Figure 3. Related to Figure 3. Expression percentile of cancer-related genes (marked with *) in patients with samples from multiple clinic visits or tissue sites (bone marrow, BM, or peripheral blood, PB). Tumor purity of PB samples was not quantified. Source data are provided as a Source Data file.

For some samples with sufficient PB tumor burden, such as patients with plasma cell leukemia, both BM and PB samples had RNA-seq. In this subset, we compared fusions detected from both samples from the same visit (Fig. 3b) (10 patients, 11 visits). *IGH--WHSC1* events were always detected in both or neither sample. Overall, more fusions were reported from BM samples than PB samples. We calculated the normalized Hamming distance between each pair of samples to quantify their overlap. Values ranged from 0.33 in pairs sharing 2 out of 3 fusions to 1 in completely discordant pairs. Previous studies have shown that tumor cells derived from peripheral blood have highly similar somatic mutation and copy number profiles.[99] Our comparison, limited to a subset of patients with high tumor burden, quantifies the fusion landscape consistency between BM and PB samples.

Next, we considered the evolution of the fusion and mutation landscape between earlier and later clinic visits, especially in four patients illustrating different patterns of clonal changes (Fig. 3c-d). Analyzing the genetic changes and clonality structures that promote relapse remains important for understanding treatment response.[100] MMRF 1433 had many more fusions reported at Visit 2 compared to Visit 1 (Fig. 3c), and the appearance of *ATM* and other mutations at Visit 2 indicates a shift in clonal architecture (Fig. 3d). Low fusion expression at Visit 2 could indicate tumor heterogeneity or correspond to low tumor purity (66%). In MMRF 1496, the *NRAS* mutation at Visit 1 (VAF 0.673 with copy number loss) was not detected at Visit 3 (no mutation call or read-level evidence), meaning the *NRAS* mutant subclone was lost between visits. The *CDC42BPB* and *MNAT1* fusions remained present, implying the hemizygous *NRAS* mutant subclone arose after or independently of those fusions. In MMRF 1656, there was one clonal missense mutation in kinase *BCR* and one important fusion event, *TPM3--NTRK1*. The absence of a known oncogenic driver mutation at Visit 1 may mean the *NTRK1* fusion played a tumorigenic role and could have been an ideal drug target high on the tumor evolutionary tree. By Visit 4, mutations in *FAM46C*, *FGFR3*, and *KRAS* were detected at or above 50% VAF, indicating a strong clonal expansion of the new mutations after diagnosis. In contrast, another patient with an *NTRK1* fusion, MMRF 2490, had clonal mutations in well-known myeloma tumor suppressors *EGR1* and *DIS3*, meaning that targeting the *NTRK1* fusion alone may not have been sufficient. Those mutations as well as expression levels of the fusion gene indicate tumor stability. Measures of fusion allele frequency useful for tracking clonal dynamics remain complicated by lower detection power and consistency compared to mutations; confident assessment of fusion VAF from expression data is an area of ongoing research and may benefit from cross-platform data integration. Further, the clonal resolution possible from bulk RNA-seq can be improved by methods that detect fusion events from scRNA-seq data.

Chimeric transcripts in scRNA-seq reveal single cell fusions

Fusion detection from bulk RNA-seq returns a fusion list but little further resolution. To detect fusions in single cells or, more broadly, present in tumor subclones, we analyzed barcoded scRNA-seq data from in-house MM patients generated on the 10x Genomics Chromium platform 3' scRNA-seq protocol. Previous MM studies utilized scRNA-seq to investigate variation in heterogenous tumors, and AML mutations in single cells illustrated tumor specificity and subclonality.[101,102] Our method detects chimeric transcripts associated with cell and molecule barcodes and map those to their cell of origin (see **Methods**). We analyzed scRNA-seq data from 5 MM patients (8 samples). Patients had known translocations that guided our discovery, including one t(4;14), one t(8;14), and three t(11;14). The results reflect trends learned from bulk analysis but with additional, informative detail (Supplementary Figure 4). In samples with an initiating t(4;14), fusion events are readily detected and map to specific malignant plasma cell subclones. In the patient with a secondary t(8;14) event, the t(8;14) subclone appears to be lost at relapse, emphasizing patterns of tumor heterogeneity and treatment response. Finally, although evidence of t(11;14) events is often observed in RNA and scRNA-seq due to upregulation of *CCND1*, actual *IGH--CCND1* fusion transcripts may not be present or reported at the RNA level, and we find a similar low detection rate of chimeric transcripts in scRNA.



Supplementary Figure 4. Related to Figure 4. a-b. Cell types and Fuscia scRNA-seq fusion discovery for Patient 27522 (primary, relapse) with t(4;14). Results from overlapping and non-overlapping regions. **c-d.** Patient 56203 (primary, relapse) with t(8;14). **e.** Patient 47499 (CD138+ sorted primary) with t(11;14). **f.** Patient 77570 (primary) with t(11;14). **g-h.** Patient 81012 (primary, relapse) with t(11;14). Source data are provided as a Source Data file.

Quality control steps identified regions with high transcript overlap (see **Methods**). In these regions, true positive chimeric transcripts from real fusions may be detected in addition to chimeric transcripts attributed to high expression of certain genes. We confidently mapped one sample's *IGH--WHSC1* fusion events from non-overlapping genomic regions to single cells. This sample (Patient 27522, primary) comprised plasma cells (54.5%, 2477/4543 cells), monocytes (29.8%), B cells (6.6%), and CD4+ T cells, CD8+ T cells, and dendritic cells, each under 5% (Fig. 4a). We defined a high-confidence subpopulation of tumor cells harboring del(chr13) to evaluate the sensitivity of our approach. In that subpopulation, our non-overlap detection rate was 4.6% (54/1166 tumor cells) (Fig. 4b). Further, no fusions mapped to non-plasma cells. The expression pattern of *WHSC1* and *FGFR3* indicates upregulation across all plasma cells, although there is subregional variation (Supplementary Figure 5a-b). Since t(4;14) and *IGH--WHSC1* are often clonal, our method showed overall low detection power, possibly reflecting the sparsity and positional bias of 3' scRNA-seq sequencing or the stringency of our quality control.

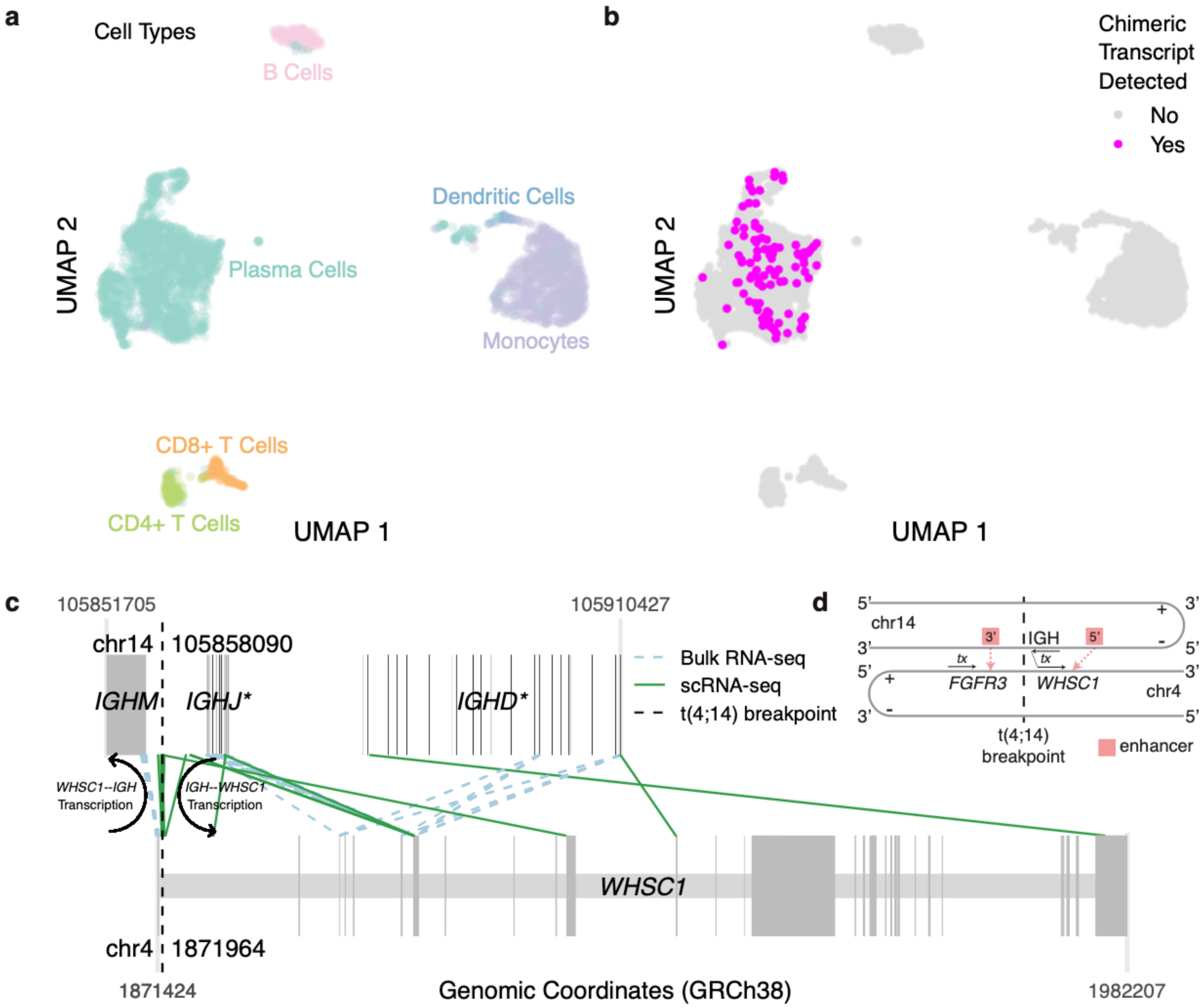
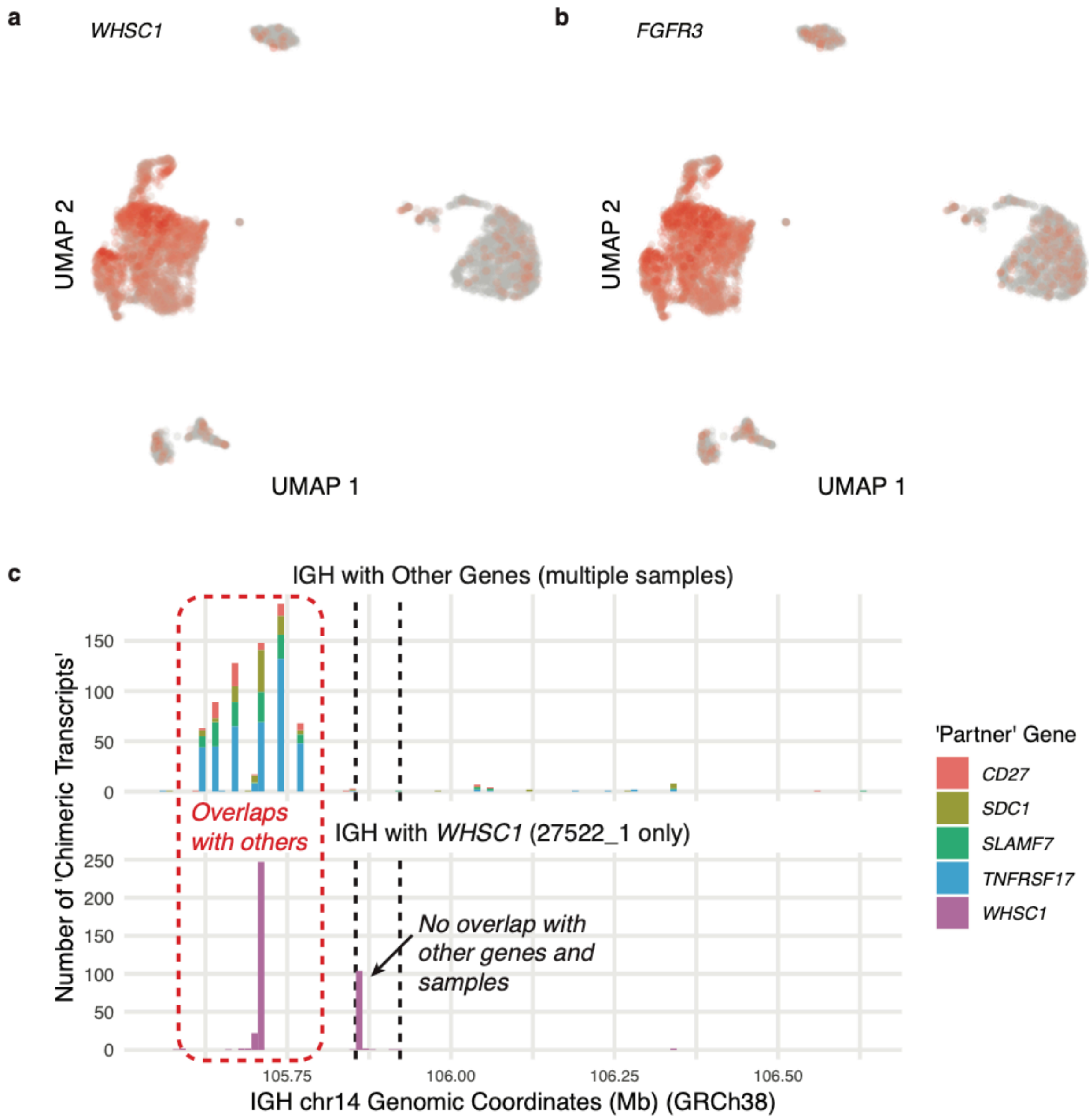


Figure 4. Single cell chimeric transcript detection. a. Cell types present from one patient's scRNA-seq sample (27522 primary disease stage, with t(4;14)). **b.** Cells with chimeric transcripts detected from non-overlap regions. **c.** Mapping location of paired-end (bulk) or same barcode (scRNA-seq) reads. **d.** *IGH--WHSC1* fusion transcription model. Source data are provided as a Source Data file.



Supplementary Figure 5. Related to Figure 4. a. Single cell expression of *WHSC1*. **b.** Single cell expression of *FGFR3*. **c.** Mapping location and number of 'Chimeric Transcripts' linking IGH with various 'Partner' genes which do not form real fusions with IGH except for *WHSC1*. Source data are provided as a Source Data file.

Fusion-support reads from bulk and scRNA-seq reads mapped to similar exonic locations along the IGH region and *WHSC1* gene body (Fig. 4c) and illustrate some transcript heterogeneity. After t(4;14), transcription proceeds from chr14 (negative strand) (IGH region) to chr4 (positive strand) (*WHSC1*) (Fig. 4d). Reads mapping to the right of the t(4;14) breakpoint (vertical dotted black line) on both chromosomes support *IGH--WHSC1*. Reads mapping to the left are transcribed in the opposite direction and support *WHSC1--IGH*. Reads from non-overlapping regions mapped to the *IGHM*, IGHJ, and IGHD regions of the IGH superlocus, precisely where *IGH--WHSC1* and t(4;14) were detected from bulk sequencing. (Supplementary Figure 5c).

Despite the resolution gained from single end scRNA-seq, we lose the benefits of paired reads used for fusion detection from bulk data. Our method demonstrates the potential utility and feasibility of mapping fusions to individual cells. Long-term implications include better understanding of tumor heterogeneity, subclonality, and the relationship of fusion events with gene expression and somatic alterations. Continued methods development, both in sample sequencing and fusion detection, building upon this early work is necessary to improve single cell fusion mapping accuracy and sensitivity. Future methods and data, especially full-length transcript scRNA-seq data, will elucidate complex expression changes due to MM translocations and fusions, which have only been analyzed in bulk RNA-seq.

IGH translocations lead to dysregulated *WHSC1* and *FGFR3*

MM translocations juxtapose highly expressed immunoglobulin loci (IGH, IGK, and IGL) with oncogenes such as *WHSC1* and *MYC*, leading to upregulation and tumor selective advantage. Neighboring genes may also be dysregulated through this process, like when *WHSC1* and *FGFR3* are both dysregulated with t(4;14). Typically, the t(4;14) translocation breakpoint on chr4 occurs between *WHSC1* and its upstream neighbor *FGFR3*. Previous studies showed that *WHSC1* and *FGFR3* are both upregulated in around 70% of patients while the remaining 30% only have high *WHSC1* expression.[103] In our data, 93 patients had a reported *IGH--WHSC1*, *IGH--FGFR3*, or reciprocal fusion; all had high *WHSC1* expression and 72.0% (67/93 patients) had *FGFR3* overexpression (Fig. 5a). No samples had *FGFR3* overexpression without t(4;14). Of patients with high *FGFR3* expression and mutation calls, 15.3% (9/59 patients) had somatic mutations in *FGFR3* (see **Methods**), all of which were copy number neutral at *FGFR3*. Interestingly, when we compared the DNA and RNA VAF of each *FGFR3* mutation, the RNA VAF was always 2-4 times higher than the DNA VAF, indicating a strong pattern of allele specific expression in all 9 cases. We hypothesize that the *FGFR3* mutant allele expression is driven by the 3’ enhancer of IGH located on the same allele as the mutation. In this scenario, expression of the translocation allele dominates the expression landscape, and the RNA VAF reflects the proportion of translocation alleles with the *FGFR3* mutation.

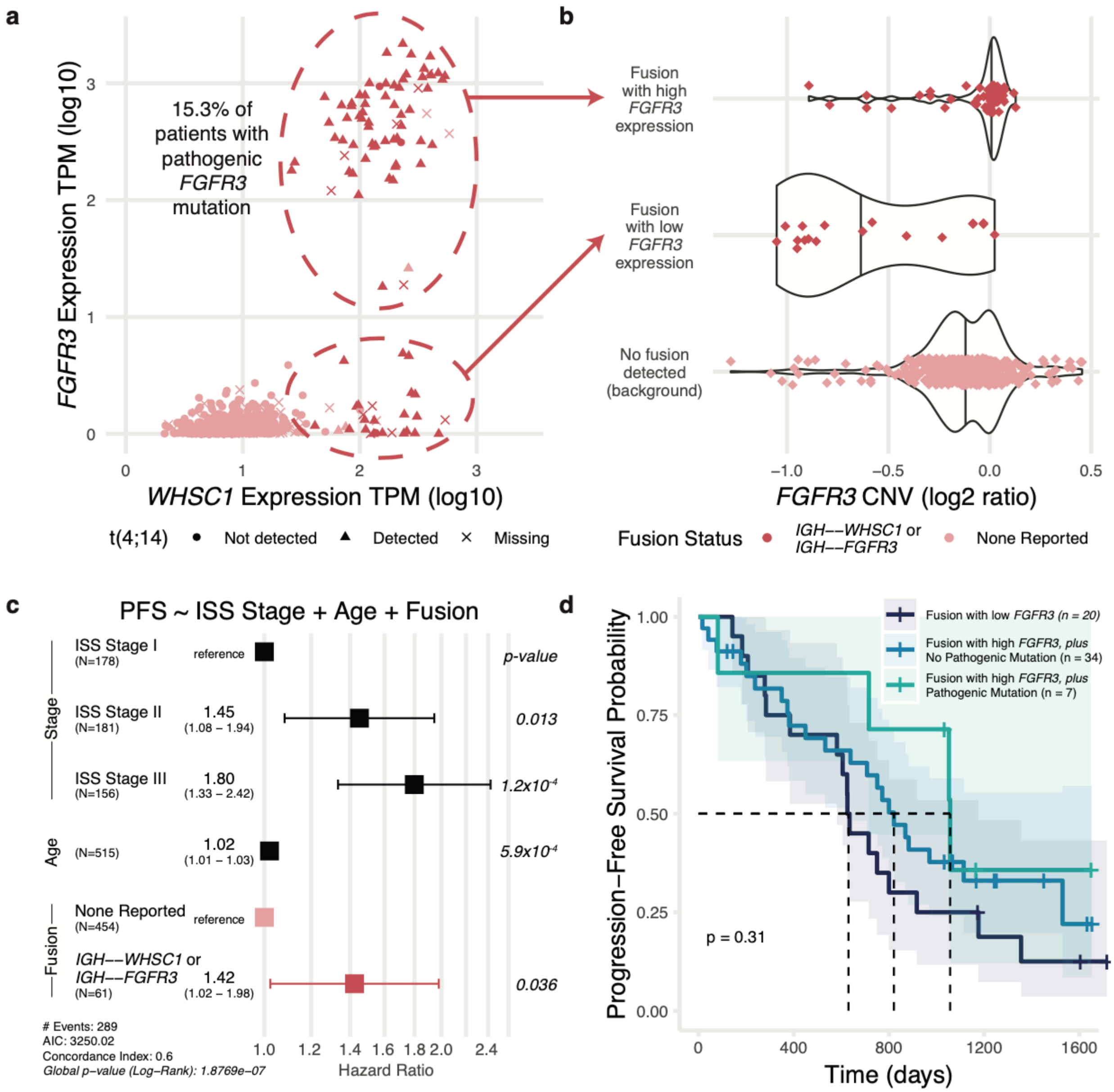
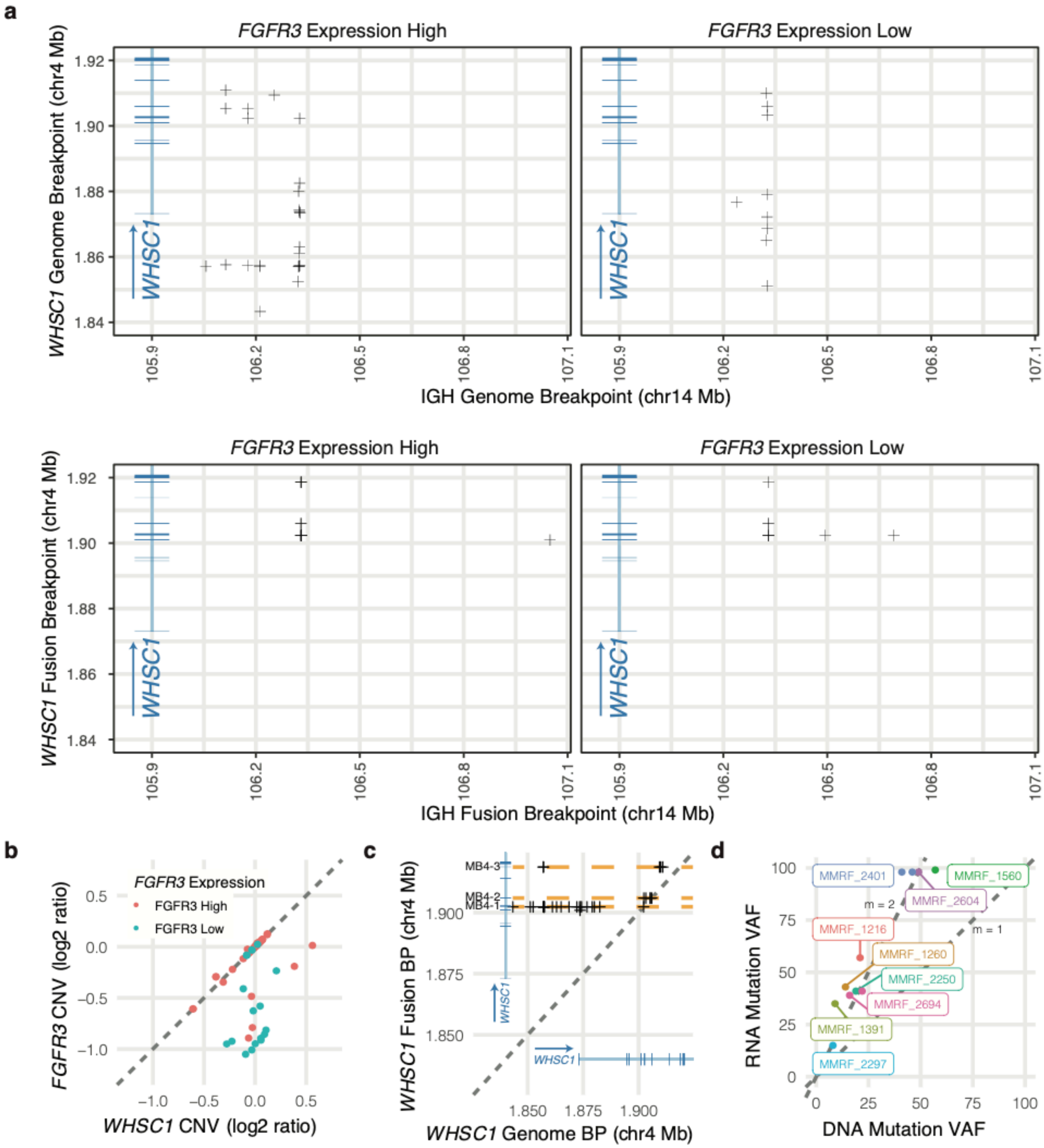


Figure 5. t(4;14) *WHSC1* and *FGFR3* expression and survival patterns. a. Co-expression of *WHSC1* and *FGFR3*, annotated with fusion and translocation status. **b.** *FGFR3* copy number (log2 of the tumor/normal ratio). **c.** Multivariate Cox proportional hazards progression-free survival model including disease stage, age, and fusion status. Error bars indicate a 95% confidence interval on each hazard ratio estimate. Covariate p-values derived from z-scores are two-sided. **d.** Kaplan-Meier curve stratified by *FGFR3* expression among fusion patients. Shaded regions indicate a 95% confidence interval on each survival curve. Significance p-value was calculated by two-sided log-rank test and uncorrected for multiple comparisons. Source data are provided as a Source Data file.

We then used available WGS translocation breakpoint and CNV data available from 34 samples with a reported *IGH--WHSC1* fusion. We observed no relationship between *FGFR3* expression status and the location of genomic or fusion breakpoints (Supplementary Figure 6a). Fusion samples with low *FGFR3* expression had distinctly lower *FGFR3* copy number (Fig. 5b) while corresponding *WHSC1* copy number tended to remain neutral (Supplementary Figure 6b), suggesting a loss of *FGFR3* after t(4;14) translocation.[104] Genomic breakpoints near IGH ranged over 0.27 Mb on chr14, while the chr4 genomic breakpoints ranged over 0.07 Mb, occurring both upstream of and within the gene body of *WHSC1*. As expected, *IGH--WHSC1* fusion breakpoints always occurred downstream of the genomic breakpoints on chr4, with three fusion breakpoint groups coalescing in the documented MB4-1, MB4-2, and MB4-3 regions of *WHSC1* (Supplementary Figure 6c).[105]



Supplementary Figure 6. Related to Figure 5. a. Breakpoint mapping locations at IGH and *WHSC1*, split by *FGFR3* expression. **b.** *WHSC1* and *FGFR3* copy number in samples with IGH fusion. **c.** Breakpoint mapping location at *WHSC1*. **d.** *FGFR3* DNA and RNA variant allele frequency (VAF) comparison. Source data are provided as a Source Data file.

Patients with pre-treatment *IGH--WHSC1* showed poorer PFS in a multivariate Cox proportional hazards model compared to patients with the same ISS stage and age (HR 1.42; HR 95% CI 1.02-1.98; two-sided z-score p-value 0.035880) (Fig. 5c). Among patients with *IGH--WHSC1*, there was no difference in PFS between those with high and low *FGFR3* expression (Fig. 5d, Supplementary Figure 2e). For the few patients with pathogenic *FGFR3* mutation and available survival data (7 patients, 4 events), mutation status was not a significant model predictor, although the small sample size after stratification precludes any robust conclusion.

MYC translocations lead to *MYC* and *PVT1* fusions

Samples with *MYC* mutations or Ig fusions involving *MYC* or its downstream neighbor *PVT1* showed elevated *MYC* expression (Fig. 6a). 10 samples had a *MYC* mutation. *MYC* fusion breakpoints occurred across the *MYC* gene body while *PVT1* fusion breakpoints were located mostly at its 5’ end; Ig breakpoints ranged across each Ig region (Supplementary Figure 7).

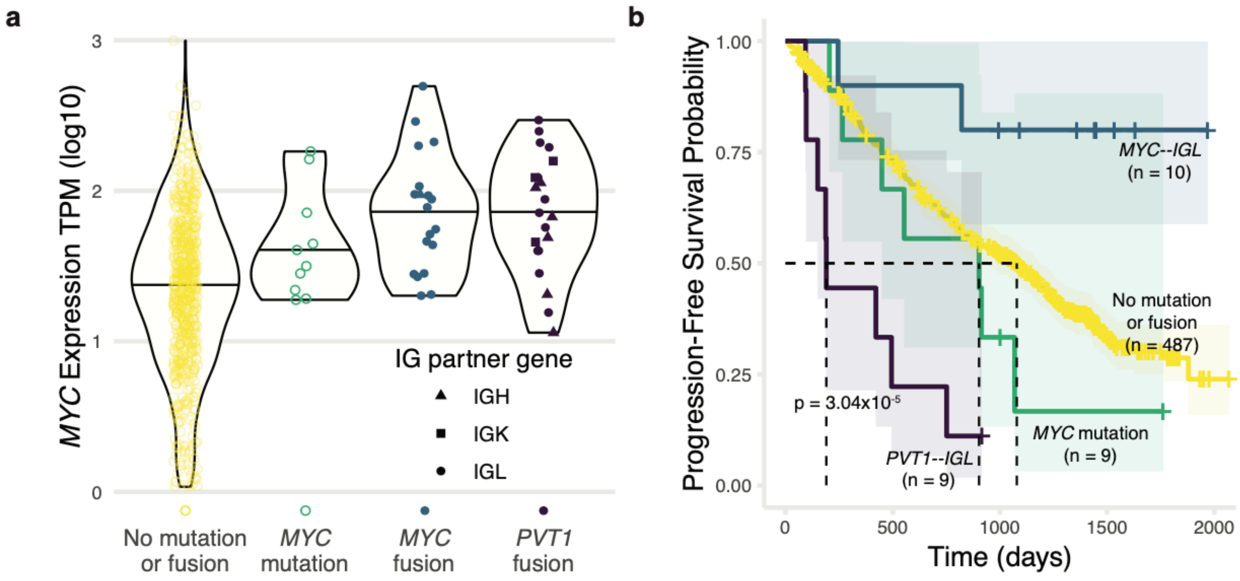
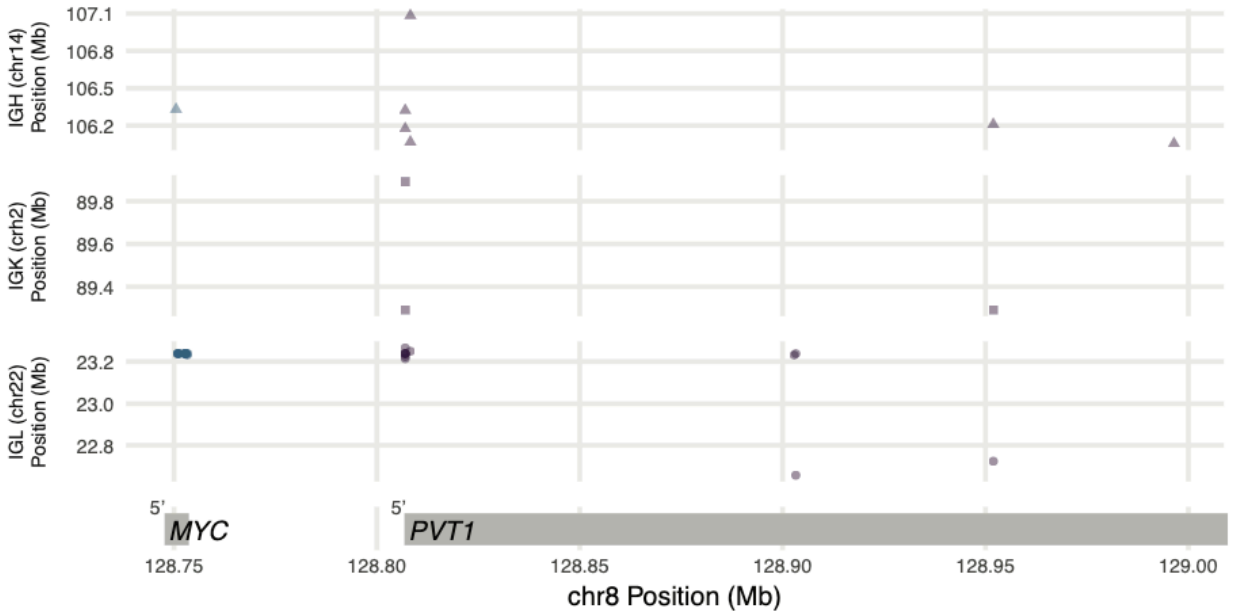


Figure 6. MYC translocation fusion partners and survival differences. a. *MYC* expression by *MYC* mutation or fusion status. **b.** Kaplan-Meier curves stratified by *MYC* mutation or fusion status. Shaded regions indicate a 95% confidence interval on each survival curve. Significance p-value was calculated by two-sided log-rank test and uncorrected for multiple comparisons. Source data are provided as a Source Data file.



Supplementary Figure 7. Related to Figure 6. Fusion breakpoint mapping locations at Ig loci, *MYC*, and *PVT1*. Source data are provided as a Source Data file.

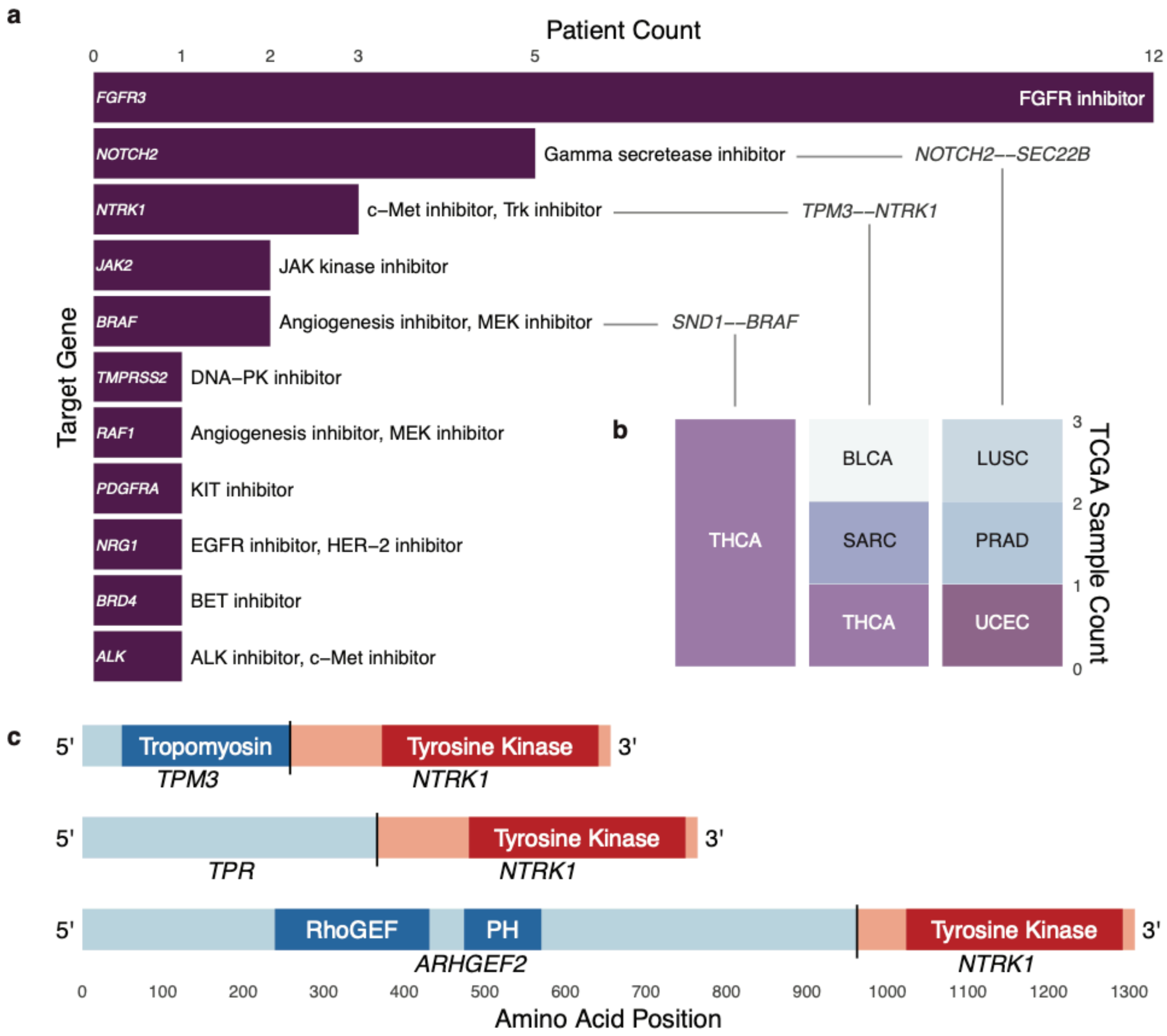
IGL translocations predict decreased survival in MM.[92] Kaplan-Meier curves for *PVT1--IGL* and *MYC--IGL* show that patients with *PVT1--IGL* had worse survival than the background (median PFS 190 days), while patients with *MYC--IGL* showed better survival with more censoring (median PFS not reached) (Fig. 6b). Further, only 18.2% of *PVT1--IGL* patients were ISS Stage I, while 43.8% of *MYC--IGL* patients were ISS Stage I. In a Cox model including ISS Stage and patient age, *PVT1--IGL* status had an estimated HR of 3.90 (95% CI 1.91-7.95; two-sided z-score p-value 0.000181), while the *MYC--IGL* HR estimate was 0.26, (95% CI 0.06-1.05; two-sided z-score p-value = 0.059018) (Supplementary Figure 2f). Of the 15 patients with complete seq-FISH data and *MYC--IGL* or *PVT1--IGL*, 8 had *MYC--IGL* and 7 had *PVT1--IGL*. One of 8 with *MYC--IGL* had t(8;22). Six of 7 with *PVT1--IGL* had t(8;22). Thus, fusions annotated as *PVT1--IGL* may be more closely associated with t(8;22) than fusions annotated as *MYC--IGL*. *PVT1--IGL* has prognostic value to the extent that it is a proxy for t(8;22). Follow-up is needed to evaluate the source and relevance of these reported events. The *MYC/PVT1* relationship and its role in tumorigenesis remains an area of ongoing research.

MYC and *MYC* paralogs can be dysregulated through copy number amplification, viral integration, and translocation.[106] MM Ig translocations dysregulating MYC predict poor survival, and *MYC* can be downregulated by BET domain inhibitors.[92,107] One oncogenic role of lncRNA *PVT1* is to stabilize and upregulate MYC protein, promoting tumorigenesis.[108] In contrast, the *PVT1* promoter may compete with the *MYC* promoter, acting as a tumor suppressor.[109] *PVT1* promoter mutations may disrupt that *MYC* downregulation. Future studies will determine how genomic variation affects *MYC*/*PVT1* interactions. The *MYC* region is a hotbed of genomic rearrangement and instability. The underlying mechanisms contributing to the tumor evolutionary advantage of this complex pattern could be elucidated by ongoing and future studies, especially with haplotype-resolved copy number and translocation calls.[110]

Fusions are potential drug targets with prognostic relevance

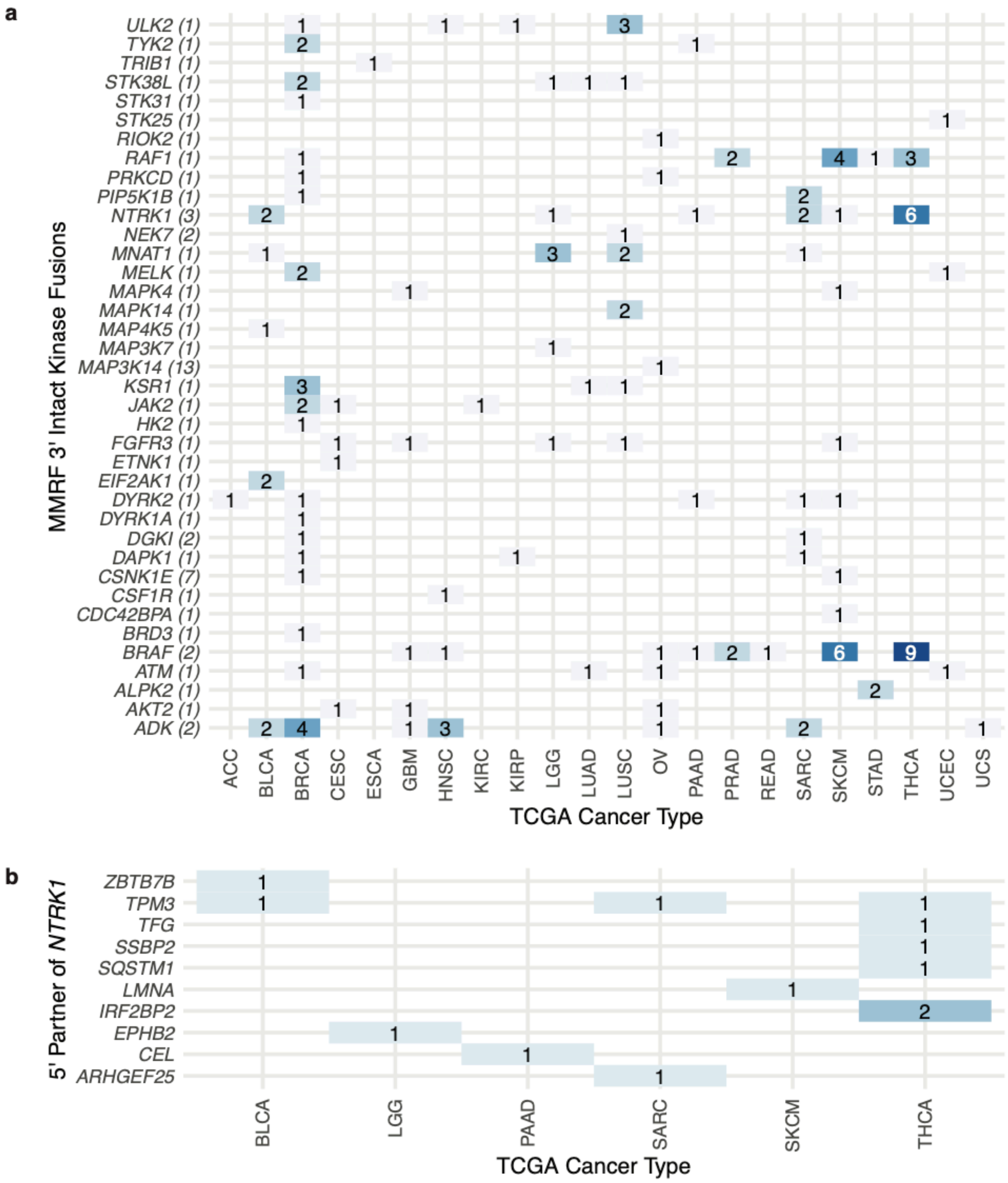
MM treatment often involves combination therapies, including alkylating agents, histone deacetylase inhibitors, immunomodulatory agents, monoclonal antibodies, proteasome inhibitors, and steroids.[82] Patients with actionable mutations in *BRAF*, *KRAS*, *NRAS*, *FGFR3*, or upregulation of *CCND1*, *CCND3*, and *MYC* may be eligible for targeted therapies.[82]

We discovered 11 fusion genes reported in the Database of Evidence for Precision Oncology as potentially sensitive to drug treatment in other cancer types (Supplementary Figure 8a).[111] 4.0% of patients had a fusion annotated as druggable. We observed 2 patients with *BRAF* fusions, and *BRAF* fusions have shown some evidence of sensitivity to MEK pathway inhibitors in the absence of other drivers.[112] We found direct overlap of potentially druggable fusions in six cancer types (Supplementary Figure 8b), pointing toward opportunities for tissue-agnostic clinical trials.



Supplementary Figure 8. a. Overlap of fusion calls with DEPO drug target database. **b.** Cancer types with exact fusion overlaps. **c.** Protein structures of *NTRK1* gene fusions. Source data are provided as a Source Data file.

Kinase fusions are important across cancer types, especially since they may be sensitive to kinase inhibition. In our cohort, common kinase pathways with fusion genes included the NIK, MAPK, and RAS pathways. We compared intact 3’ kinase fusions from our cohort to those reported from a TCGA pan-cancer analysis (Supplementary Figure 9) and found the same 3’ kinase fusions reported across 22 cancer types.[35] Fusions with *ADK*, *BRAF*, and *NTRK1* were reported repeatedly both in our cohort and in multiple cancer types.



Supplementary Figure 9. a. MMRF 3' intact kinase fusion partner genes overlapping with TCGA cancer types (number of samples). **b.** MMRF *NTRK1* partner genes overlapping with TCGA cancer types. Source data are provided as a Source Data file.

NTRK genes, including *NTRK1*, encode cell surface neurotrophin receptor tyrosine kinases. TRK fusions are a drug target in solid cancers, although TRK inhibition may lead to resistance mechanisms.[113] TRK fusions from hematological cancers were responsive to inhibition in cell culture and mouse modeling.[114] We found three patients with 3' *NTRK1* fusions, each with an intact kinase domain (Supplementary Figure 8c), and two had the same fusion detected at a later clinic visit (Supplementary Figure 3). All three primary samples had strong WGS support for their fusion event. *NTRK1* fusion 5' partners all came from the opposite strand of the same chromosome (chr1), indicating that an inversion event may have brought the two genes together. There is also evidence of chromosome 1q copy number amplification in these samples, highlighting overall genomic instability in the region. Each partner gene had expression in the 90th percentile or above, potentially driving *NTRK1* activity higher (Fig. 2c), and *NTRK1* was overexpressed in each case (Fig. 2a), leading to upregulation of downstream pathways.

APOBEC signature is associated with *MAF* and *MAFB* translocations in multiple myeloma, and such translocations are markers of poor prognosis.[115,116] Of three samples with *MAF--IGL*, each had outlier APOBEC signature scores and high *MAF* expression, lending further evidence to the relationship between APOBEC and dysregulated *MAF* (see **Methods**).

Discussion

Our study forms an MM gene fusion landscape and explores clinical relevance. We analyzed the gene expression patterns of fusions, fusions involving kinase genes, druggable targets, evolution of tumor fusion profiles, and translocation and fusion breakpoints of events. We also compared fusions from serial clinic visits and from different tissue sources. We developed methods to map scRNA-seq fusion events to single cells. Our results represent a resource for future studies involving gene fusions in multiple myeloma and other cancer types and highlights several fusion analysis methods. We have built upon prior studies and hope our resource and strategies can be useful for future research and clinical translation.

Targeted sequencing can generate cost-effective reports with clinical utility, including somatic mutations, indels, translocations, and gene expression profiles.[117] Including fusions will require tool development to meet clinical standards, although methodological and study design improvements are being made in this direction.[118] scRNA-seq and long read sequencing will further delineate genomic changes during tumor progression, elucidating subclonal heterogeneity and contextualizing common patterns observed from bulk sequencing.

MM immunotherapies , including checkpoint inhibition, monoclonal antibodies, and chimeric antigen receptor T (CAR-T) cells, represent the forefront of targeted therapy. Pan-cancer studies showed reduced mutational load in patients with driver fusions, meaning they would not be ideal candidates for neoantigen-based immunotherapy.[35,119] However, dramatic responses to immunotherapy have sometimes been observed using gene fusions as neoantigens.[120]

In multiple myeloma, fusions represent an area for continued study, especially as they relate to gene expression, disease progression, tumor evolution, and targeted therapy. Ongoing research to improve fusion detection tools and pipelines that leverage information from multiple data types will enable more complete pictures of patient tumors as bioinformatics analyses become more deeply integrated into clinical decision making.

Methods

Alignment

Paired RNA-seq fastq files were aligned to GRCh37 using STAR version 2.5.3a_modified.[121]

BAM files were sorted and analyzed with flagstat using Samtools version 1.5.[122] Quality control was conducted using FastQC version 0.11.5. (See bioinformatics.babraham.ac.uk/projects/fastqc/.)

Association testing and correlation

Association testing was done using Student's t-test (two-sided) (continuous expression) and Fisher's Exact Test (two-sided) (categorical expression). Clinical associations with fusions and fusion genes were calculated using Fisher's Exact Test (two-sided) for categorical variables and Mann-Whitney U Test for continuous variables. Expression and clinical testing p-values were corrected using the Benjamini and Hochberg false discovery rate (FDR) method.[123] All correlations are calculated as Pearson correlations unless otherwise stated.

Copy number variation detection

We detected copy number variation from WGS data using BIC-seq2[124] (BICseq2-norm version 0.2.4; BICseq2-seg version 0.7.2). In scRNA-seq, we used inferCNV (version 0.8.2) to calculate single cell copy number profiles.[125]

Fusion analysis scripts

Fusion results were analyzed by scripts written in Python (version 3.7.2) and R (version 3.5.3). Python packages included numpy, os, and pysam. R packages included ggrepel, gridExtra, readxl, RColorBrewer, Seurat (version 3.0.0), survival, survminer, tidyverse, and UpSetR. (Please see github.com/ding-lab/griffin-fusion/tree/master/mmrf_fusion for fusion analysis scripts.)

Fusion detection

We used five fusion detection tools including EricScript[40] (version 0.5.5), FusionCatcher[126] (version 1.00), INTEGRATE[127] (version 0.2.6, using RNA-seq samples only, not paired RNA and WGS), PRADA[128] (version 1.2), and STAR-Fusion[39] (version 1.1.0). Gene names from immunoglobulin super-loci were condensed to IGH, IGK, and IGL (including *IGLL5*).

Fusion filtering

Fusions were required to be called by at least two tools. Fusions called by any combination of EricScript, FusionCatcher, or INTEGRATE must also have been called by STAR-Fusion or PRADA in another sample (soft filter tag EFI). Fusions were removed if: partners are the same gene; genes appear on blacklist or are paralogs; fusion comes from list of normal panel fusions (non-cancer cell lines, GTEx, TCGA normal samples)[35,62]; one partner is promiscuous with 25 or more partners (soft filter tag Many Partners); or partner genes are within 300 Kb (soft filter tag Within 300Kb). Additionally, across all samples for a particular fusion pair, we required at least one sample to have 2 or more junction reads or one sample to have 1 or more spanning reads, or that fusion pair was removed from all samples (soft filter tag Low

Count). Finally, fusions with a low WGS support rate compared to the background rate were removed if the binomial test two-sided p-value was less than 0.15 (soft filter tag Undervalidated). See Supplementary Table 6 for a list of all soft filtered fusions and why they were filtered.

Gene expression

Transcripts per million (TPM) was calculated using kallisto[129] (version 0.43.1).

Gene level TPM was calculated as the sum of TPM values from each of that gene's transcripts.

Log transformation of TPM values was calculated as log10(TPM + 1).

Kinase domain analysis

Kinase domain status was determined based on reported gene fusion breakpoints using AGFusion[130] (version 1.231). (See github.com/murphycj/AGFusion.) Following manual review, 15 out of 19 *MAP3K14* fusions were found to possess an intact kinase domain after initially being reported as having disrupted kinase domains due to a lack of annotation.

Mutation signature profiling

We used SignatureAnalyzer[131] to quantify mutation signatures.

Outlier detection

Gene expression outliers were defined as having values greater than 75th + 1.5*IQR or less than 25th - 1.5*IQR, where 75th and 25th represent the 75th and 25th percentile, respectively, and IQR is the interquartile range, defined as the 75th percentile minus the 25th percentile.

Single cell fusion detection -- Fuscia

Given an aligned BAM file, barcode information for each read mapping to fusion gene regions was extracted using the Python module pysam (version 0.15.2), which wraps Samtools[122] (version 1.7). When two reads map to different genes or regions and share the same cell and molecular barcode, we labeled that transcript as a "chimeric transcript". Multiple reads could originate from the same chimeric transcript. We eliminated reads with length > 128 and then selected one representative read from each side of the chimeric transcript by picking the reads mapping closest to the known WGS breakpoint. Transcript overexpression makes false positive detection of chimeric transcripts more likely. We reduced this risk by purposefully looking for chimeric transcripts that may be detected due to overexpression. In plasma cells with IGH translocations, we specifically looked for chimeric transcripts linking IGH and plasma cell markers *SDC1*, *SLAMF7*, and *TNFRSF17*. We called those regions 'overlap' regions because chimeric transcripts from genes not associated with fusions overlap with those from legitimate fusions. (Please see github.com/ding-lab/fuscia.)

We used R (version 3.5.3) and the Seurat[132] package (version 3.0.0) to analyze cell type and gene expression from individual data. Dimensional reduction was performed using UMAP.[133]

Somatic mutation calling

MMRF exome bams were aligned to hg19, and somatic variants were called by our in-house pipeline SomaticWrapper, which includes four established bioinformatic tools (Mutect[134] (version 1.1.7), Pindel[135] (version 0.2.54), Strelka2[136] (version 2.9.2), and VarScan2[137] (version 2.3.83)). (See github.com/ding-lab/somaticwrapper.) We kept SNVs called by at least two out of three tools (Mutect, Strelka, VarScan2). Likewise, we kept INDELs called by at least two out of three tools (Pindel, Strelka, VarScan2). We required 14X coverage for somatic mutation calls and only kept mutations with tumor variant allele frequency (VAF) >= 0.05 and normal VAF <= 0.02.

Structural variant detection

Structural variants were detected from paired normal and tumor WGS samples using Delly[138] (version 0.7.6) and Manta[139] (version 1.1.0). To be analyzed, tumor and normal WGS samples must have had matching sequencing assays and a corresponding RNA-seq sample.

Survival analysis

We performed survival analysis using progression-free survival as the outcome using the survival (version 2.44-1.1) and survminer (version 0.4.6) packages in R. To test for significant improvements in model fit with additional covariates, we implemented a chi-squared test using the anova() function and compared the new model to the baseline model. Only patients whose primary sample corresponded to the pre-treatment clinic visit were included for survival modeling.

Tumor purity

We used the R package estimate[140] (version 2.0) to quantify tumor purity from RNA-seq data. Tumor purity of peripheral blood (PB) samples was not quantified.

WGS support of fusion events

We used WGS data to determine if reported fusions also had genomic support. We defined a breakpoint window centered at each fusion breakpoint. If there were 3 or more discordant read pairs mapping to within 100 Kb of each breakpoint, we determined the fusion to be supported by WGS. Reads were filtered by Samtools[122] (version 1.5) with flags -F 1920 -f 1 -q 20. We removed fusions from all samples if the fusion-specific support rate differed significantly from the background support rate of all fusions.

Data availability

Data was provided by The Multiple Myeloma Research Foundation (MMRF) CoMMpass (Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile) Study (NCT01454297). dbGaP Study Accession: phs000748.

For single cell RNA sequencing of additional patient samples, the Washington University Institutional Review Board approved the study protocol, and we have complied with all relevant ethical regulations, including obtaining informed consent from all participants.

The source data underlying all figures are provided as Source Data files accessible with DOIs 10.6084/m9.figshare.11941494 (for everything except scRNA data) and 10.6084/m9.figshare.11941506 (for scRNA data).

Code availability

Data analysis scripts and single cell fusion detection methods are available under the MIT license at github.com/ding-lab/griffin-fusion/tree/master/mmrf_fusion and github.com/ding-lab/fuscia.

Chapter 4: Somatic mutation phasing and haplotype extension using linked-reads in multiple myeloma

Summary

Somatic mutation phasing informs the relationship of cancer-related events, like copy number loss and inactivating mutations. We analyzed linked-read whole genome sequencing data from 14 multiple myeloma patients across several disease stages (23 total samples). We developed SomaticHaplotype, an open-source tool for analyzing linked-reads and systematically assigning somatic mutations to haplotypes. We report the landscape of phase sets across our samples and show how phase set length can be extended 4.6 fold when pairing samples from the same patient. We also uncover disease-relevant phasing information in cancer genes, phasing 79.4% of high confidence somatic mutations and enabling us to interpret clonal evolution models at higher resolution. For example, our analysis suggested that two *NRAS* hotspot mutations occurred on the same haplotype but were independent events in different subclones. Our framework for haplotype analysis enables phase-aware analysis of somatic events in any cancer type and can be integrated with established methods for structural variation phasing.

Introduction

Humans genomes are diploid, meaning they have two copies of each autosomal chromosome in their normal state. When a zygote forms, the haploid maternal and paternal gametes each contribute one copy of the genome to the diploid zygote. Each copy of each chromosome contains a pattern of inherited germline variation that distinguishes it from the other copy. Genetic relationships between individuals may be established by comparing patterns of germline variation within an individual with parental samples or other ancestral samples from world populations. Whereas a genotype defines a set of alleles at a specific locus and does not include alleles from other sites, a haplotype consists of information across loci and distinguishes between different copies of the chromosome. A variant is phased when it has been assigned to a particular haplotype. Phasing may be achieved through various techniques, including both technological and computational methods.[141] Sequencing of trios (two parents, one child) allows phasing because the child shares stretches of DNA identical-by-descent from each parent. Large-scale SNP databases of world populations are useful resources that enable genotype imputation and computational phasing approaches.[142] 1000 Genomes is one example of a publicly-available, population-scale resource of phased haplotypes.[143,144]

Mutations and structural variation that leads to cancer occur on specific haplotypes, though haplotype information is often lost with next-generation bulk sequencing.[145] Mutations may have cis effects, affecting nearby gene activity, or trans effects, having an impact beyond the immediate neighborhood. For example, a cis-acting mutation at an expression quantitative trait loci (eQTL) may impact expression on the same haplotype (allele specific expression). Knowing that the mutation and the expression change both came from the same haplotype can help determine the impact of the mutation. Knudson’s two-hit hypothesis states that some cancers are driven by two events affecting the same gene or process. For example, biallelic inactivation of TP53 is a marker of poor prognosis in multiple myeloma.[97] Determining the haplotype on which each event occurred informs the oncogenic process. Mutations may also affect the same haplotype, such as double PIK3CA mutations, which have been found to be more oncogenic but also more susceptible to PI3Ka inhibitors.[146]

Technologies that enable determination of the long range information between variants are collectively referred to as Third Generation.[147] The two major categories of Third Generation technologies generate long reads (e.g. PacBio and Oxford Nanopore) or synthetic long reads (e.g. 10X Genomics). PacBio and Oxford Nanopore both offer long, continuous reads and direct observation of epigenetic modifications, with the trade-offs of cost for PacBio, accuracy for Napopore, and large amount of input DNA for both. A benefit of long reads is being able to span repeat regions that may confound short read alignment. With 10X Genomics, short fragments originating from the same haplotype are linked together (linked-reads) with the same barcode. Although this approach has higher sequencing accuracy, requires less input material, and costs less, accurate phasing in regions of low complexity is still a challenge. Zheng et al. established this linked-read approach and describe how this technology enables exon modeling for accurately determining fusion breakpoints and how phasing an inactivating TP53 mutation to one haplotype and a hemizygous deletion to the other proved a two-hit process.[19] Later, Marks et al. established the accuracy and reliability of linked-read whole genome sequencing (lrWGS) and discussed factors that impact phasing performance, such as variant density and heterozygosity.[148]

Linked-reads have impacted cancer study design and enabled novel insights to tumor biology. Greer, et al. compared gastric cancer metastases and delineated a complex structural variant leading to FGFR2 amplification.[18] Viswanathan, et al. utilized linked-reads to determine the order of events in a cohort of prostate cancer patients, showing the ordering of androgen receptor (AR) gene duplications, CDK12 inactivation, phasing somatic variants if the reads supporting it were assigned to a haplotype and phase set, and developing allele-specific copy number detection methods.[149,150] Sereewattanawoot, et al. explore the cis-acting effects of regulatory mutations by using linked reads in lung cancer cell lines to match regulatory variants with allele-specific expression.[151]

In this study, we analyzed a cohort of multiple myeloma patients using linked-read whole genome sequencing (lrWGS) generated by the 10X Genomics Chromium system (see **Methods**).[19] Fig. 1a describes the process of generating lrWGS data. From a bulk sample of cells, long fragments of DNA, also called high-molecular weight (HMW) DNA, is isolated into an individual gel bead in emulsion (GEM). Each GEM contains a gel bead with primers including a 16-bp DNA barcode unique to that GEM. The gel bead dissolves and releases the barcoded primers, which attach to the DNA and undergo isothermal amplification. Now each short fragment of amplified DNA contains a barcode identifying which GEM it originated from. The GEMs break and the barcoded fragments are pooled together and sequenced.

We aligned reads using Long Ranger (v2.2.2, reference GRCh38, see **Methods**). The advantage of lrWGS over traditional WGS is that reads with the same barcode that map to the same region are overwhelmingly likely to have originated from the same haplotype. This gives additional leverage to studies examining the long range information missed by short-read sequencing. Long Ranger aligns reads, calls and phases variants, reports SVs, and produces phasing quality metrics. With enough information (depth and allelic heterogeneity), Long Ranger is able to assign variants and reads to haplotypes. Variants and reads are grouped into phase sets, also called phase blocks, which are genomic ranges in which the haplotype assignments of variants are consistent. Within a particular phase set, all variants assigned to a certain haplotype are thought to have originated from the same biological haplotype. In another phase set, the haplotype order may switch, so haplotype assignments cannot be compared between phase sets.

Long Ranger phasing is designed to work with germline single nucleotide variants. Phasing performance may be suboptimal in regions of copy number variation, for somatic mutations with low variant allele frequency, or in tumor samples with low purity.

Multiple myeloma is the second most common form of blood cancer and, in the United States, has higher incidence among African-Americans.[65] Myeloma is a disease caused by clonal proliferation of plasma cells and is typified by large structural variation or hyperdiploidy. Common primary event translocations join the highly expressed IGH locus (chr14) with cancer genes, including t(11;14) (CCND1), t(4;14) (WHSC1), t(6;14) (CCND3), and t(14;20) (MAFB), and secondary events include MYC translocations. MAPK is the most commonly mutated pathway in MM including somatic mutations in KRAS, NRAS, and BRAF. Better appreciation of the haplotype context of these events, both driver mutations and structural variation, is necessary to improve targeted therapies and understanding of myelomagenesis.

To our knowledge, our cohort is the largest published multiple myeloma linked-reads WGS data set to date and improves our understanding of human haplotype and cancer haplotype analysis. We created novel methods for systematically phasing somatic mutations to haplotypes and inferred haplotype relationships between somatic mutations as well as translocations. We also present methods for extending phase sets using overlapping information from the same individual and ancestral population samples.

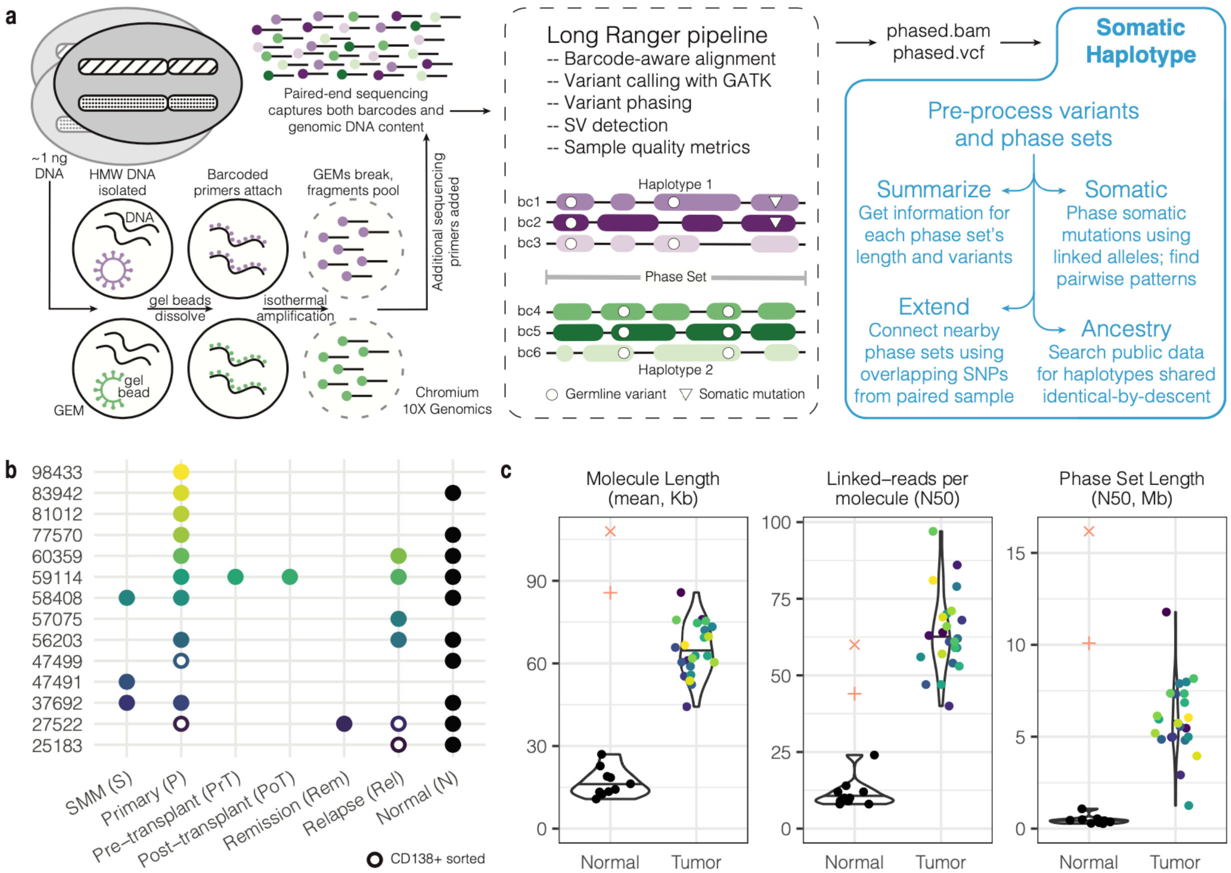

a
~1 ng DNA
HMW DNA isolated
Barcoded primers attach
GEMs break, fragments pool
DNA
gel beads dissolve
isothermal amplification
gel bead
GEM
Paired-end sequencing captures both barcodes and genomic DNA content
Additional sequencing primers added
Chromium 10X Genomics
Long Ranger pipeline
-- Barcode-aware alignment
-- Variant calling with GATK
-- Variant phasing
-- SV detection
-- Sample quality metrics
Haplotype 1
bc1
bc2
bc3
Phase Set
bc4
bc5
bc6
Haplotype 2
○ Germline variant
▽ Somatic mutation
phased.bam
phased.vcf
Somatic Haplotype
Pre-process variants and phase sets
Summarize
Get information for each phase set's length and variants
Somatic
Phase somatic mutations using linked alleles; find pairwise patterns
Extend
Connect nearby phase sets using overlapping SNPs from paired sample
Ancestry
Search public data for haplotypes shared identical-by-descent
b
98433
83942
81012
77570
60359
59114
58408
57075
56203
47499
47491
37692
27522
25183
SMM (S)
Primary (P)
Pre–transplant (PrT)
Post–transplant (PoT)
Remission (Rem)
Relapse (Rel)
Normal (N)
CD138+ sorted
c
Molecule Length (mean, Kb)
0
30
60
90
Normal
Tumor
Linked–reads per molecule (N50)
0
25
50
75
100
Normal
Tumor
Phase Set Length (N50, Mb)
0
5
10
15
Normal
Tumor

Figure 1. Linked-read data generation and analysis pipeline. a. The 10X Genomics Chromium platform tags large DNA molecules with barcodes such that reads originating from the same molecule have the same barcode. The Long Ranger pipeline aligns reads and phases variants. **b.** The SomaticHaplotype tool builds upon Long Ranger output with several modules. **c.** Our cohort comprises 14 multiple myeloma patients across several disease stages for a total of 23 tumor samples. **d.** Quality control measures for tumor and normal samples.

Results

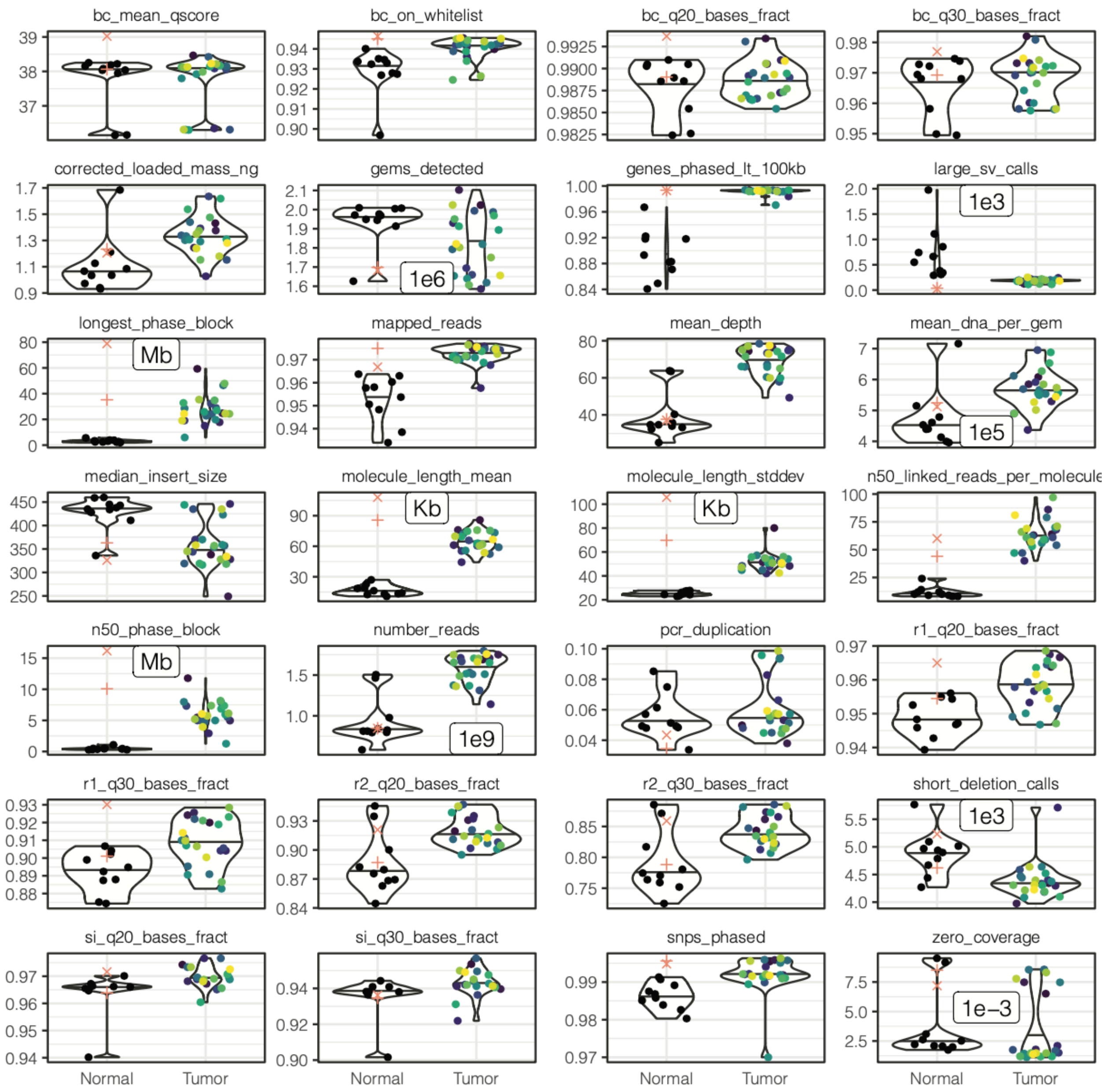
SomaticHaplotype tool builds on phasing information to analyze somatic mutations

To enable further downstream processing of lrWGS data, we developed an open-source software tool called SomaticHaplotype (SH) (Fig. 1b), see **Methods** and **Code Availability**). Given the phased variant call format (VCF) file and a phased bam alignment file produced by Long Ranger, SH pre-processes relevant information with the *phaseset* module by constructing Variant and PhaseSet objects which are used in later modules. The *summarize* module generates a table with information about each phase set, including its genomic range and the number of variants it contains, as well as statistics like phase set length N50. In the *somatic* module, we consider the haplotype origin of high-confidence somatic mutations, especially those which may not have been called or phased by Long Ranger due to low variant allele frequency or tumor purity. We then assess the haplotype relationship between pairs of proximal somatic mutations. With the *extend* module, we utilized phase information from overlapping variants detected in multiple samples from the same individual to bridge the gap between phase sets and recommend how to flip haplotype assignments to make neighboring phase sets have consistent haplotype assignments. Finally, in the *ancestry* module, we developed methods for augmenting phased lrWGS data with large-scale phased resources, like 1000 Genomes.

Our data set comprises lrWGS data from 14 individual patients diagnosed with multiple myeloma (Fig. 1c). Longitudinal samples were taken across disease stages, from the premalignant stage (smoldering multiple myeloma, SMM), to primary diagnosis, pre- and post-transplant, remission, and relapse. In total, 23 tumor samples were collected for lrWGS. In addition, 10 patients had a skin normal sample processed with lrWGS. Four tumor samples were CD138+ sorted to enrich for plasma cells, increasing tumor purity. Other samples were not CD138+ sorted and contain varying compositions of microenvironment cells in addition to tumor plasma cells. Regular paired-end next-generation was also performed on sorted samples, including 13 whole exome (WES) and 6 whole genome (WGS) (see Supplementary Data 1).

To assess sample quality, we considered the summary statistics produced by Long Ranger (as reported in output file summary.csv) (Fig. 1d). We also included publicly available summary information from two 1000 Genomes samples, NA12878 and NA19240 (Long Ranger version 2.2.1, reference hg19) (see **Data Availability**). Those two normal control samples are represented by red symbols (+ for NA12878; x for NA19240). Overall, quality control measures of our tumor samples compared well with data from gold-standard publically available lrWGS. Molecule length refers to the size of the long, HMW DNA fragments isolated into gel beads. In our tumor samples, the mean molecule length per sample ranged from 44.3 Kb to 85.8 Kb with a median of 62.8 Kb, while in our normal skin samples, the median value was 15.3 Kb. Linked-reads per molecule is the number of read pairs originated from each molecule, and the N50 value indicates that half of the molecules have that many reads pairs or more. In our tumor samples, the N50 linked-reads per molecule ranged from 40 to 97 with a median of 62, compared to a median of 10 in our skin samples. Finally, the N50 phase set length in tumor samples ranged from 1.3 Mb to 11.8 Mb with a median of 5.7 Mb, while the median was 0.4 Mb in skin samples. Given the consistent lack of informative linked-read information in our skin samples, we excluded them from downstream analysis. The skin samples were only used as a control for variant calling.

For tumor samples, the median corrected mass of input DNA loaded into the Chromium chip was 1.3 ng, and the median mean sequencing depth was 71.6 reads. The median percentage of SNPs phased by Long Ranger was 99.2%. Please see Supplementary Figure 1 and Supplementary Data 2 for additional summary quality control measures.



Supplementary Figure 1. Phasing performance quality control summary measures for all samples.

Phase set lengths reveal biologically-relevant genomic changes

We examined the distribution of phase set lengths to ascertain and explore patterns in our data. Overall, N50 phase set lengths from each samples were consistent across chromosomes, with the median N50 ranging from 4.42 Mb on chromosome 15 to 7.74 Mb on chromosome 18 (Fig. 2a). Chromosome 1 showed the least variation in N50 phase set length (median 4.52 Mb, standard deviation 1.37 Mb), while chromosome 21 showed the greatest variation (median 5.78 Mb, standard deviation 9.33 Mb) and also had the highest overall values, with 6 samples having N50 phase set lengths above 20 Mb, 4 of which came from Patient 59114. Some samples, such as 25183 (P), had consistently higher N50 values across many chromosomes (Fig. 2b). This may be due to this sample having the highest mean molecule length (85.8 Kb) and percentage of mapped reads (97.7%) of all tumor samples. Another sample, 58408 (P), had consistently shorter phase sets, but quality control measures did not clearly point to a reason for this pattern.

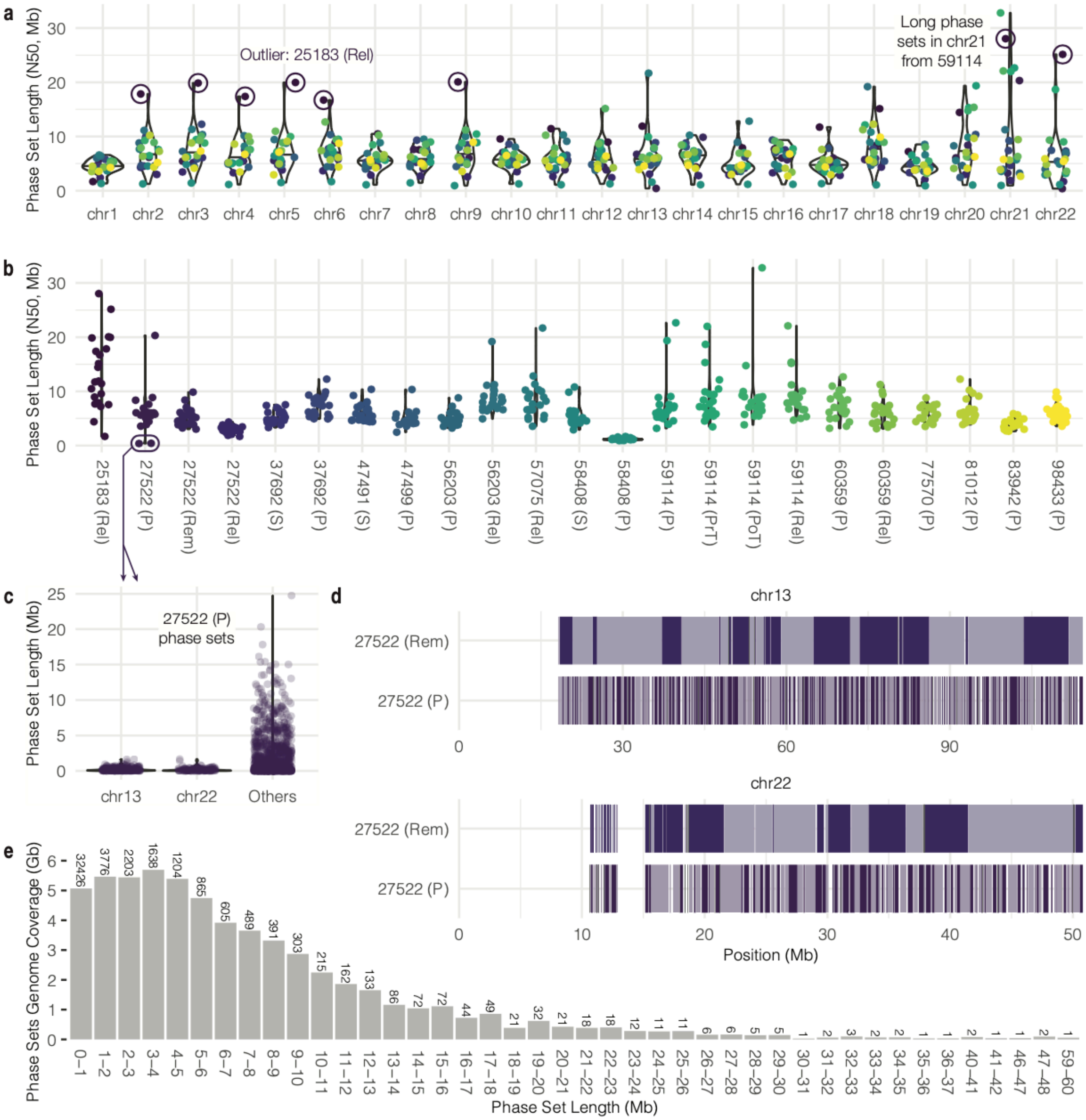
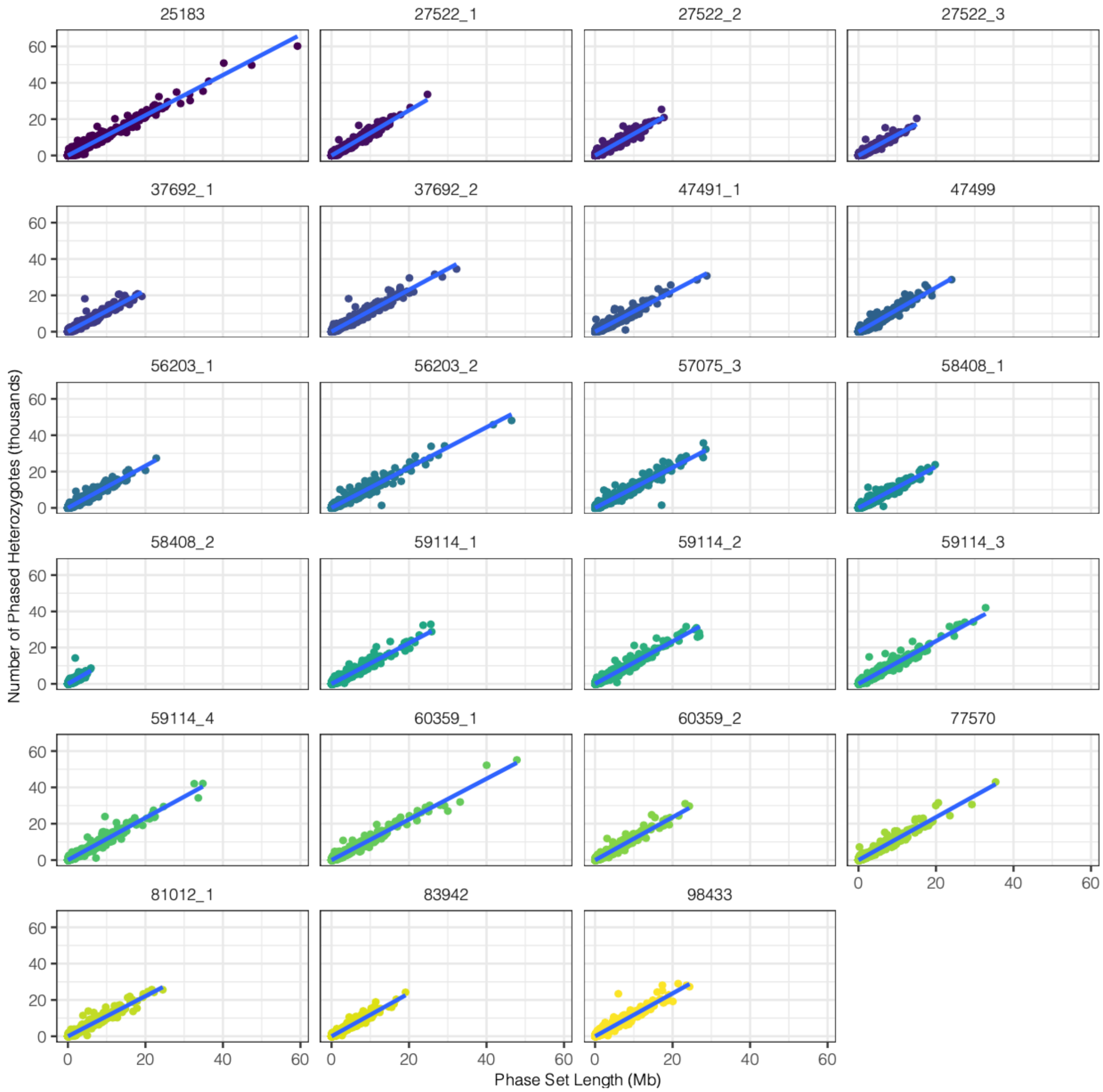


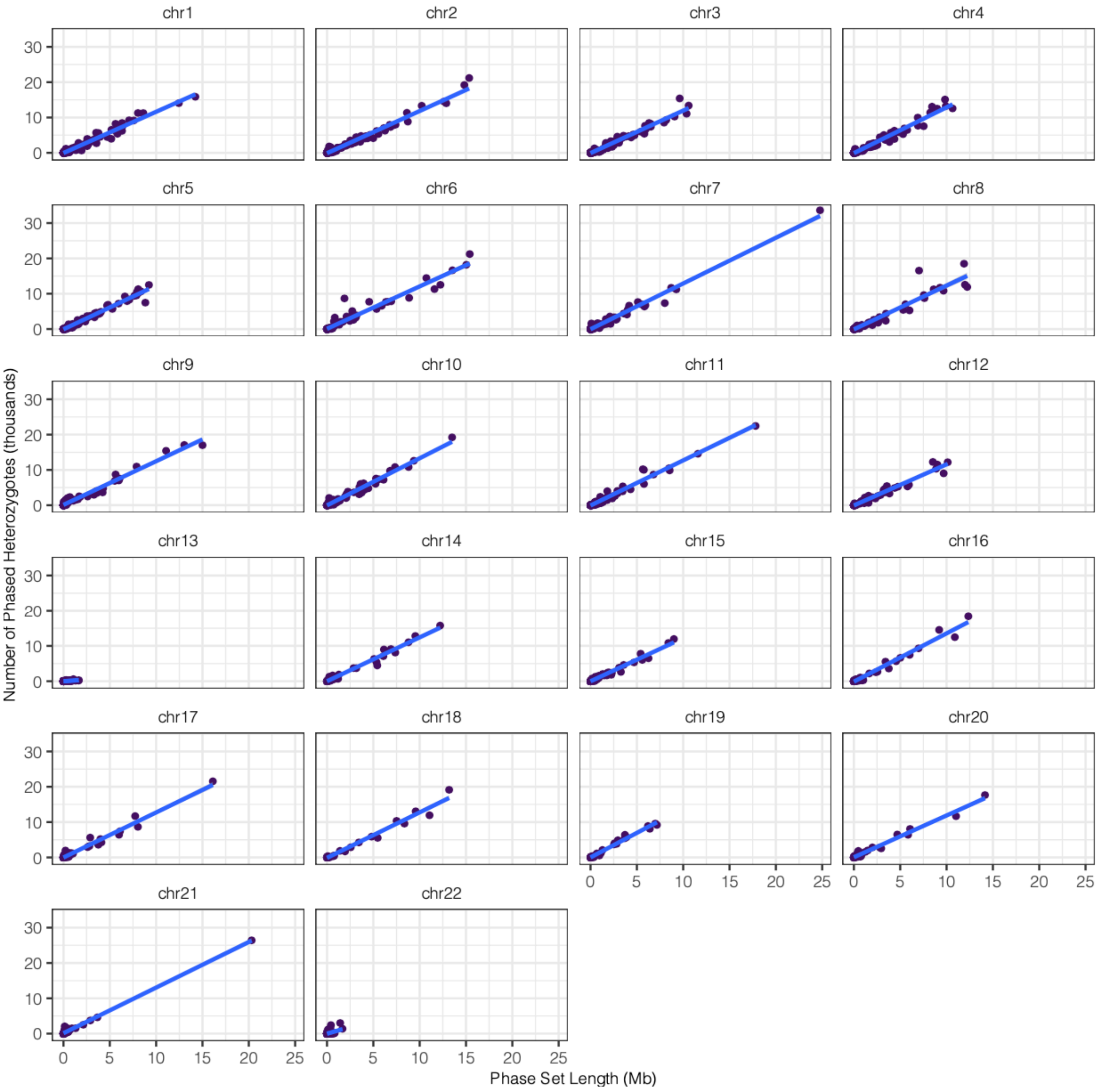
Figure 2. Phase set length distribution. a. Phase set length by chromosome across all samples. **b.** Phase set length per sample across all chromosomes. **c.** Phase set lengths of chr13, chr22, and others from 27522 (P). **d.** Phase set length and locations of chr13 and chr22 from 27522 (P) and 27522 (Rem). **e.** Total phase set genome coverage from all samples combined, grouped by phase set length.

Interestingly, chromosomes 13 and 22 from 27522 (P) showed low N50 phase set lengths, and the distribution of phase set lengths from those two chromosomes is strikingly different than the overall distribution from other chromosomes (Fig. 2c) (phase sets less than 1 kb filtered out for plotting). The N50 phase set lengths for chromosomes 13 and 22 were 0.42 Mb and 0.38 Mb, respectively, compared to that sample's median N50 of 5.9 Mb. Based on copy number variation results from lrWGS and WES data, both chromosomes 13 and 22 had a one copy deletion across the entire chromosome. The homozygosity across these chromosomes due to the deletions explains why the associated phase sets are so short, since the Long Ranger pipeline depends on having heterozygous SNPs to grow phase sets beyond linked-reads from the same barcode. One benefit of this lack of heterozygosity across entire chromosomes is that we can potentially phase the entire length of these chromosomes. In comparison to chromosome 13 and 22 phase sets from the 27522 (Rem) sample, those from 27522 (P) are much shorter, and the short phase sets are distributed across the entire chromosome (Fig. 2d). The power of short phase sets to predict deletion status in this sample may be a reflection of tumor cell content and the proportion of tumor cells affected by copy number loss.

Overall, phase sets cover 60.6 Gb of genome across our 23 tumor samples (Fig. 2e), for an average of 2.6 Gb per sample. 72.2% (32,426/44,918 phase sets) of phase sets are between 0 and 1 Mb, but collectively those short segments account for only 8.4% (5.1/60.6 Gb) of the total amount of genome covered by phase sets in these samples. In comparison, 3,776 phase sets are between 1-2 Mb, but those cover 5.5 Gb (9.0%) collectively. The distribution of genomic coverage by phase sets of increasing length has a right-skewed long tail distribution. There are 19 phase sets longer than 30 Mb, and the longest phase set is 59.2 Mb. As expected, there was a strong linear relationship between phase set length and the number of phased heterozygous variants ($r^2 = 0.96$) and also strong homoscedasticity (Supplementary Figure 2). This linear relationship was not observed for the copy number deleted chromosomes (chr13 and chr22) from 27522 (P) (Supplementary Figure 3).



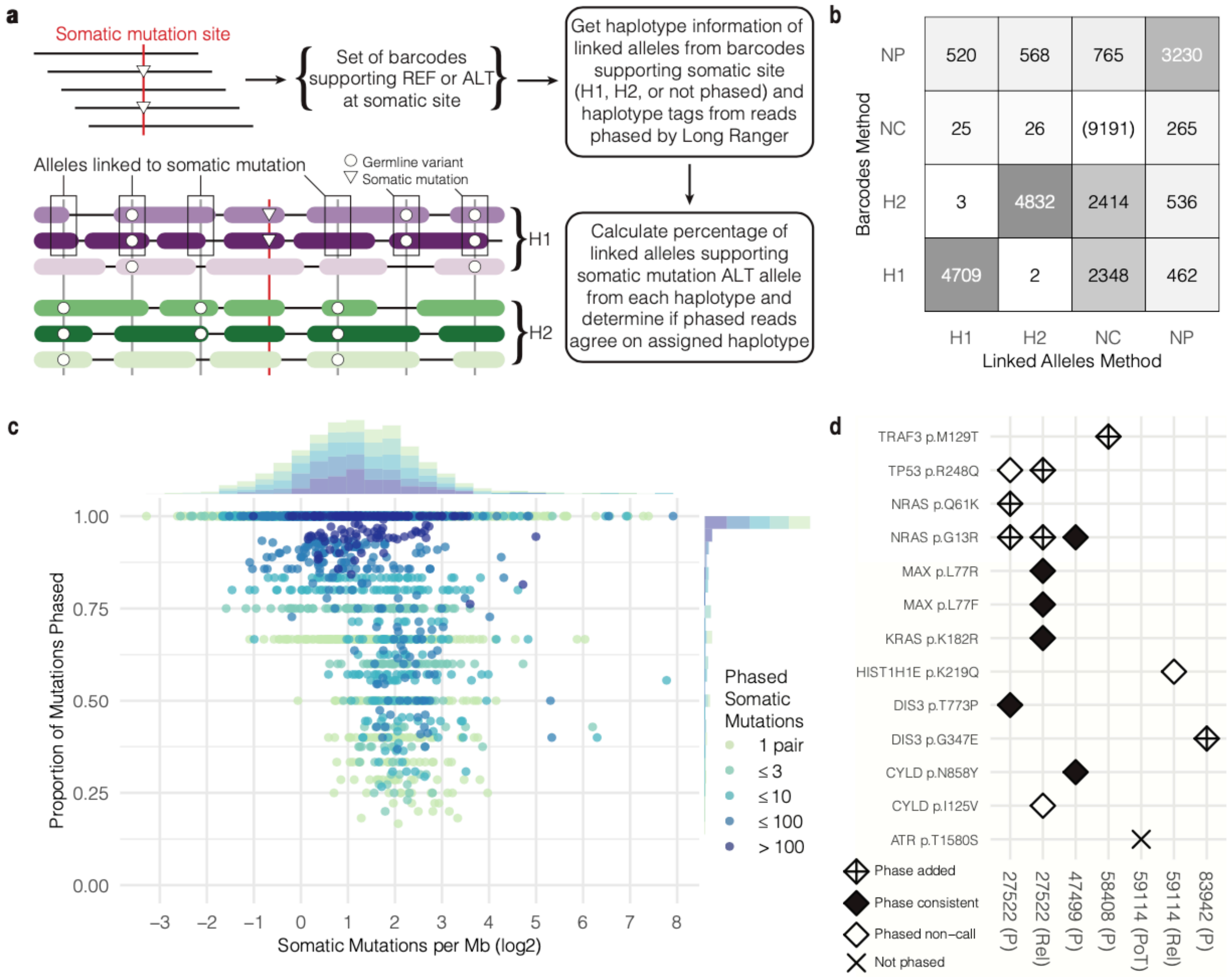
Supplementary Figure 2. Phase set length correlation with the number of phased heterozygotes across all samples.



Supplementary Figure 3. Phase set length correlation with the number of phased heterozygotes across all chromosomes from 27522 (P).

Somatic mutations can be phased to specific haplotypes using linked alleles

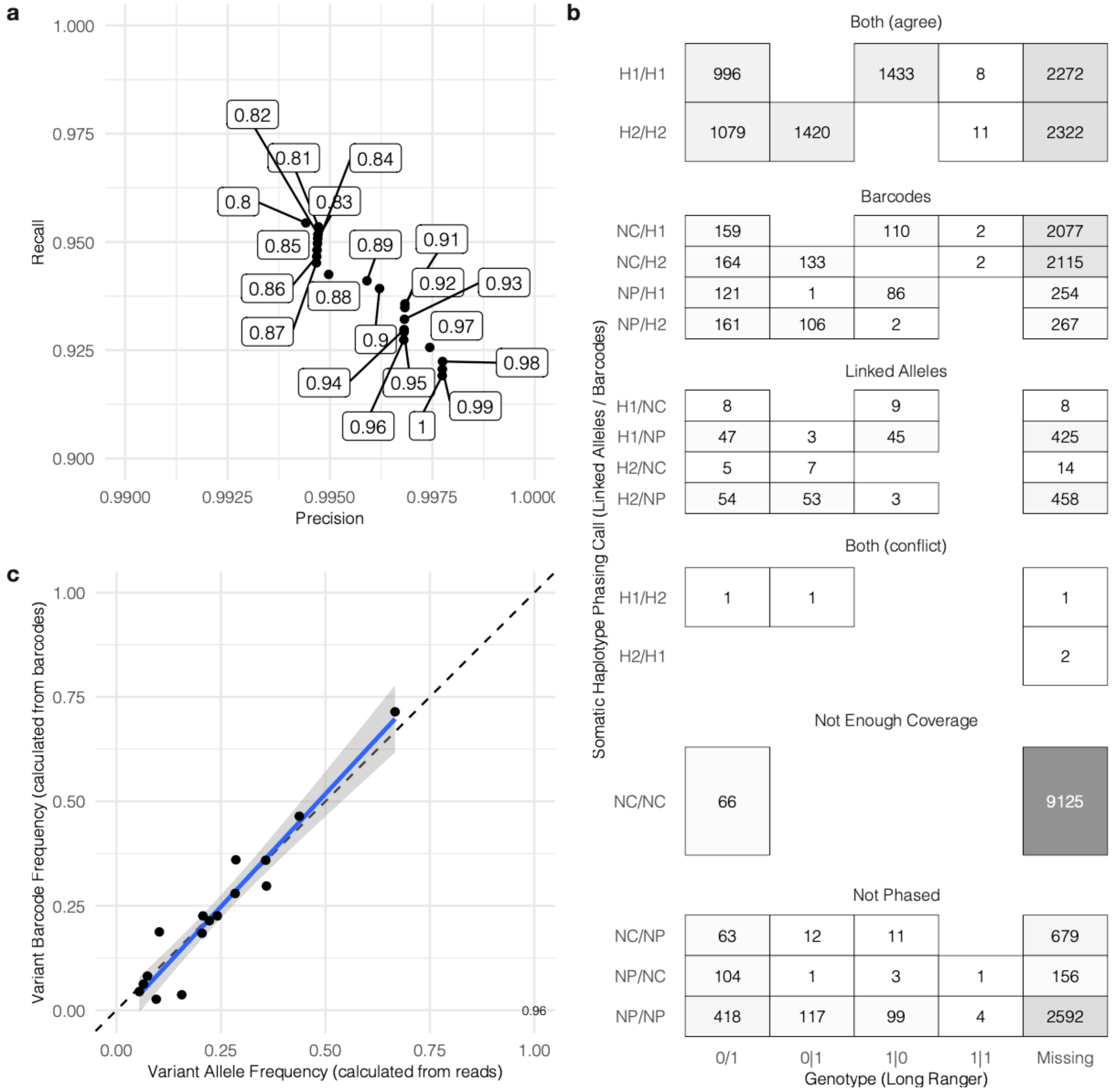
The haplotype context in which somatic mutations occur may be biologically relevant. For example, knowing the phase of two mutations affecting the same gene would indicate whether they occur on opposite haplotypes, possibly leading to biallelic inactivation, or if they are on the same haplotype with only one copy disrupted. However, tumor purity and variant allele frequency make somatic mutations harder to identify and phase using standard approaches. To phase somatic mutations, we built upon the strengths of Long Ranger by examining germline variants that occur on each barcode with reads covering the somatic mutation site (Fig. 3a). We defined linked alleles as alleles co-occurring with either the reference or alternate allele at the somatic mutation site. We know the haplotype assignment of most (~99%) linked alleles, and we know that alleles co-occurring in the same phase set with the same barcode most likely originated from the same molecule of DNA. Thus, if the linked alleles co-occurring with the alternate allele at the somatic mutation site are phased to the same haplotype above a certain threshold, then we can infer the haplotype of the somatic mutation even if it was not initially phased. An alternative method to phasing somatic mutations is to rely on the haplotype assigned to reads supporting the somatic mutation site.[151] This haplotype annotation is reported as a tag in the phased bam output from Long Ranger, but it is not given for all reads. In our tumor sample data, 71.6% of reads overlapping a mutation site were assigned a haplotype. Using this phased barcode approach, we determined that a somatic mutation was phased if at least one barcode supported the mutant allele and that barcodes supporting the mutant allele all agreed on the haplotype assignment. Combining these approaches increases our phasing power when one approach does not have adequate coverage.



Barcodes Method \ Linked Alleles Method	H1	H2	NC	NP
NP	520	568	765	3230
NC	25	26	(9191)	265
H2	3	4832	2414	536
H1	4709	2	2348	462

Figure 3. Phasing somatic mutations to haplotypes. a. Overview of methods used to phase somatic mutations. **b.** Number of somatic mutations phased using two phasing methods (H1 = phased to haplotype 1, H2 = phased to haplotype 2, NC = not enough coverage for phasing, NP = not phased). **c.** Distribution of somatic mutations per phase set and the proportion of mutations phased. **d.** Phasing somatic mutations commonly observed in multiple myeloma.

For six lrWGS samples which also had matched, sorted WGS, we identified high confidence somatic mutations using the sorted WGS sample and normal lrWGS as control (see **Methods**). In total, we detected 33,503 somatic SNVs and small indels from our six sorted WGS sample, or 5,584 somatic mutations per sample. Using those high confidence somatic mutations, we then identified barcodes and linked alleles supporting the somatic site from the lrWGS samples. Of those, 29,896 (4,983 per sample) were SNV positions with coverage in the matched lrWGS samples, and 20,705 (69.2%) of those met our minimum coverage requirement with at least 10 linked alleles on barcodes supporting the somatic allele or at least one phased barcode supporting the alternate allele. We overlapped high confidence somatic mutations with phased Long Ranger calls to create a comparison set. At a linked allele threshold of 0.91, the phasing precision was 0.997 and the recall was 0.936 (Supplementary Figure 4a) (see **Methods**). Overall, 79.4% (16,440/20,705 mutations) of somatic mutations with enough coverage were phased using the established cutoff. Of 3,380 somatic mutations with enough coverage that were called by Long Ranger but not phased, 82.7% were phased by our method. Overall, the linked alleles and barcodes phasing methods were concordant on 99.95% of phasing decisions where both methods made a phasing decision (H1 or H2) (9,546/9,551 calls) (Fig. 3b). The barcodes approach added 5,760 calls where linked alleles did not have enough coverage or did not meet the phasing threshold. The linked alleles approach added 1,139 calls. See Supplementary Figure 4b for an overview of all results by phasing method.



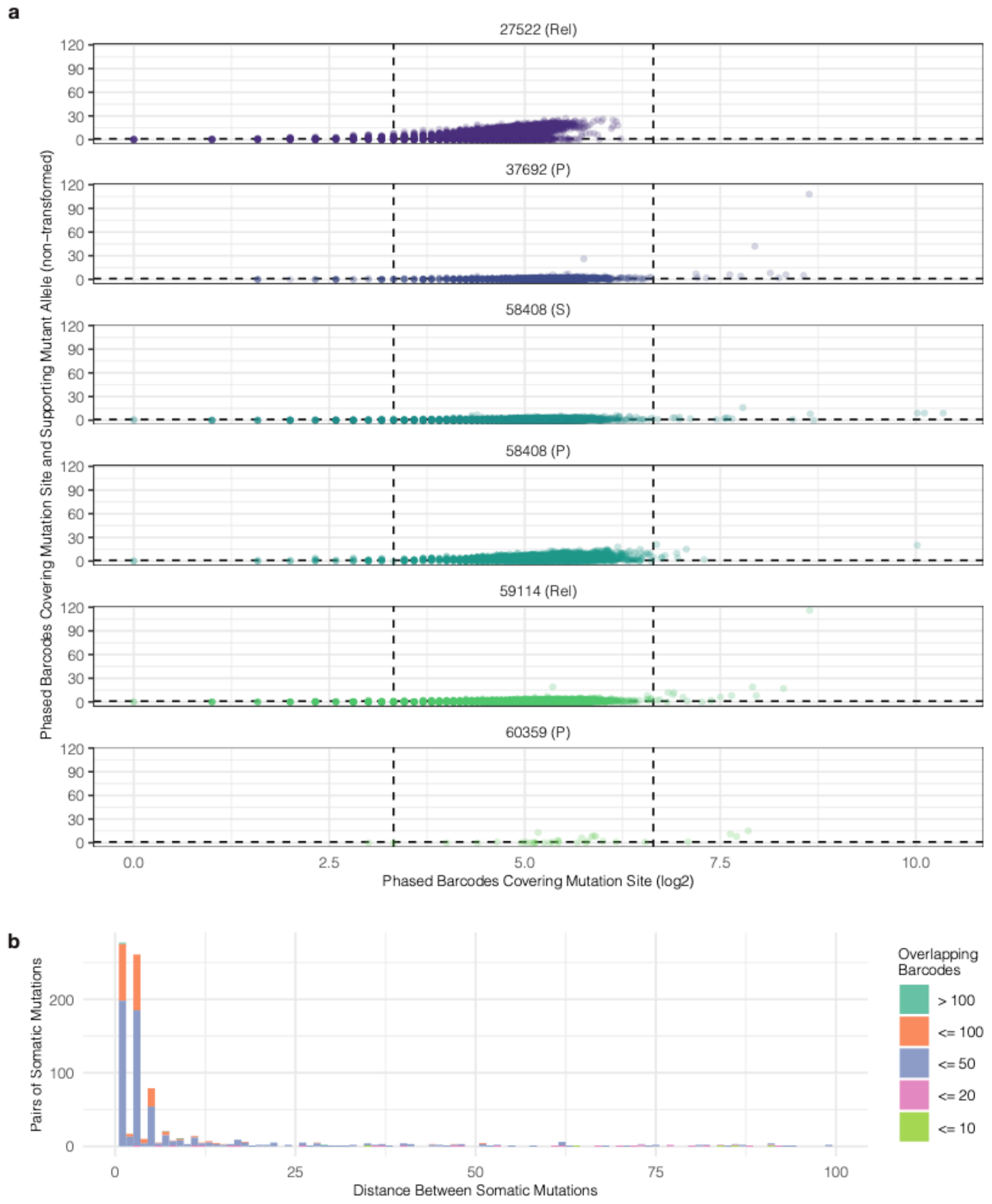
Supplementary Figure 4. Additional information related to somatic mutation phasing. a. Precision/recall rates at various cutoffs for the proportion of linked-alleles assigned to one haplotype. **b.** Comparison of phasing results with Long Ranger genotypes. **c.** Correlation of variant allele frequency derived from reads and barcodes.

Globally, we grouped high confidence somatic mutations from these six samples according to the phase set they map to and found the proportion of somatic mutations phased in our approach (Fig. 3c). Close to half of phase sets longer than 1 kb had zero pairs of somatic mutations (44.8%, 2,212/4,941 phase sets), with 11.1% having zero somatic mutations and 33.6% having only one somatic mutation. But among those 2,729 phase sets longer than 1 kb with at least one pair of somatic mutations, 37.0% had exactly one pair, 20.7% 3 or fewer, 18.6% 10 or fewer, 19.7% 100 or fewer, and the remaining 4.0% had more than 100 pairs. 64.6% had every mutation phased, and 77.5% had at least three-quarters of mutations phased. The number of phased somatic mutations per megabase within each phase set showed a log2-normal distribution ranging from 0.10 to 241.3, with a median of 2.25.

Several samples had somatic mutations commonly associated with multiple myeloma[65], including mutations in CYLD, DIS3, HIST1H1E, KRAS, NRAS, and TP53 (Fig. 3d). In several cases, we were able to confidently phase somatic mutations that were either not called or were not phased by Long Ranger. When Long Ranger did phase the mutation, our results were always consistent. One mutation in ATR was not called by Long Ranger and was not phased by our approach since the linked alleles did not clearly favor one haplotype over the other (60.2% of phased linked alleles supporting the somatic mutation were phased to Haplotype 1, and 39.8% were phased to Haplotype 2). We noted that in 27522 (P), the NRAS G13R mutation was phased by our method to Haplotype 2, but was phased to Haplotype 1 in 27522 (Rel). However, since haplotype numbering is arbitrary, such differences are trivial. Further, we noticed that in 27522 (P), the NRAS G13R and Q61K mutations, two well known hotspot drivers, were both phased to the same haplotype. Later analysis suggested that they arose independently in separate tumor subclones. Comparing the variant allele frequency of non-synonymous mutations in driver genes using two methods, we found a high correlation (r = 0.96) between VAFs calculated using all reads and VAFs calculated using only unique barcodes (Supplementary Figure 4c).

Pairs of somatic mutations illustrate patterns of clonal evolution

In traditional next-generation sequencing, if two mutations are spaced close enough together and occurred together on the same short read pair, then we may directly observe that they originated from the same haplotype and were present in the same cell. With lrWGS, we now have more reads to consider when we look for such co-occurring mutations. Analysis must account for copy number variation, which could present multiple copies of the same haplotype on which somatic mutations could arise. From our samples with both lrWGS and sorted WGS, we focused on mutation sites in copy number neutral regions with between 10 and 100 phased barcodes with linked-reads covering that position and at least one barcode supporting the alternate allele (Supplementary Figure 5a). We defined copy number neutral regions as having a $\log_2$ copy number ratio between -0.25 and 0.2 in the sorted WGS. To assess how likely two mutation sites are to have linked-reads associated with the same barcode covering both sites, we examined 59,063 pairs of mutations from the same phase set, within copy number neutral ranges, and with both mutation sites having adequate coverage. As expected, the probability of one barcode covering both sites decreases at the distance between sites increases, with 98.4% (54,643/55,559 pairs) of mutation pairs greater than 62 kb apart sharing no overlap. (62 kb is the median of the mean molecule lengths described in Fig. 1d). We focused on the 3,504 mutation pairs within relatively close proximity (i.e. less than 62 Mb apart) (Fig. 4a, Supplementary Figure 5b). For the 2,648 mutation pairs within this genomic distance but greater than 100 bp apart, the 13.0% were covered simultaneously by zero barcodes, 77.3% were covered by between 1 and 10 barcodes, 8.3% between 11 and 20 barcodes, and 1.4% greater than 20 barcodes (Fig. 4a). For the 856 mutation pairs less than 100 bp apart, each pair had at least one shared barcode (Supplementary Figure 5b). Overall, 5.9% (3,504/59,063 pairs) of somatic mutation pairs were within 62 Kb (Fig. 4b). Of those, 90.2% (3,159/3,504 pairs) share at least one barcode in common. From that reduced group, 64.6% (2,042/3,159 pairs) have a barcode on which one or both somatic mutations is represented, potentially allowing for direct observations or inference related to mutation patterns in the same cell.



Supplementary Figure 5. Additional information related to the relationship of pairs of somatic mutation. a. Number of barcodes covering each mutation site and those supporting the mutant allele. **b.** Number of overlapping barcodes by distance between somatic mutations less than 100 bp apart.

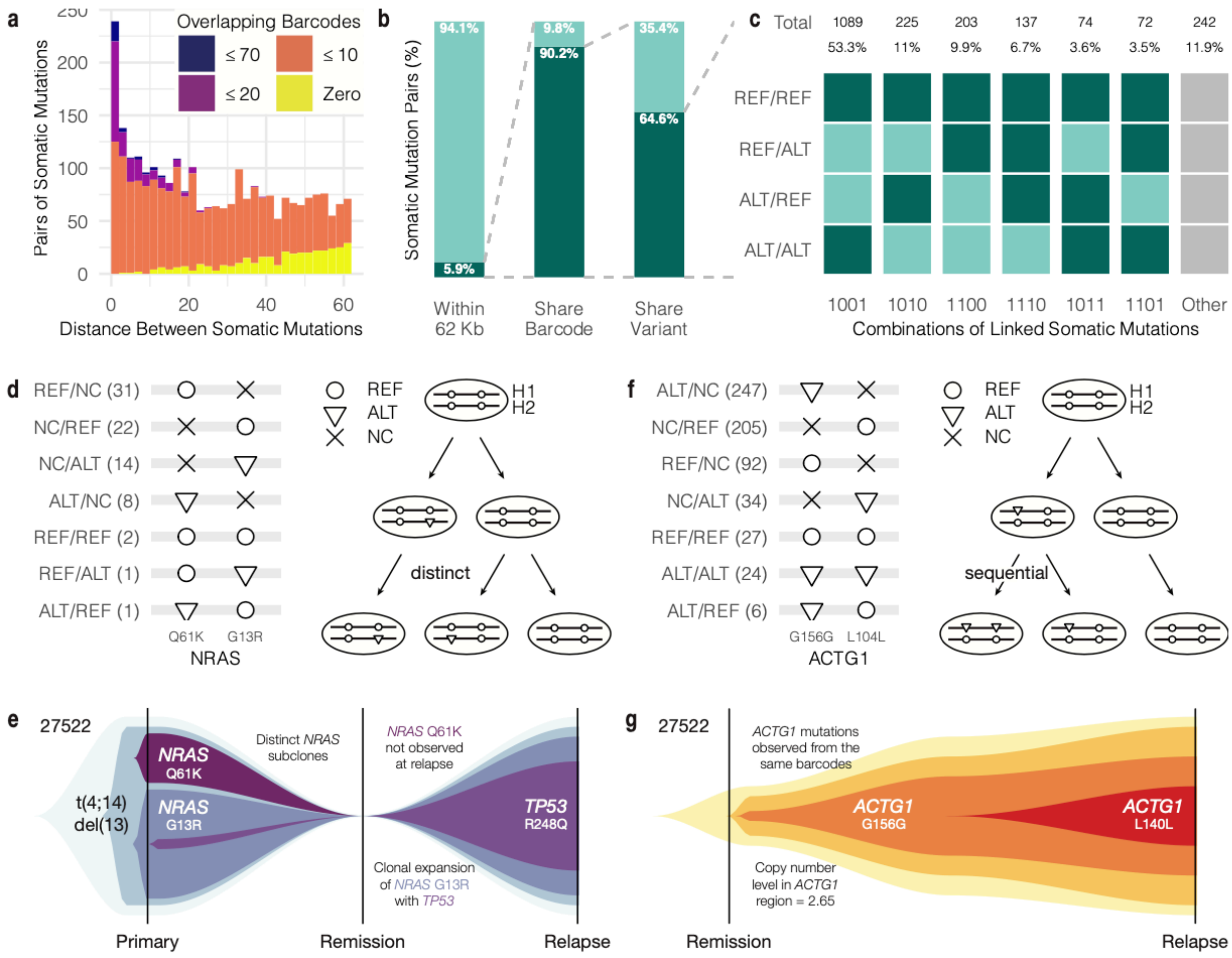
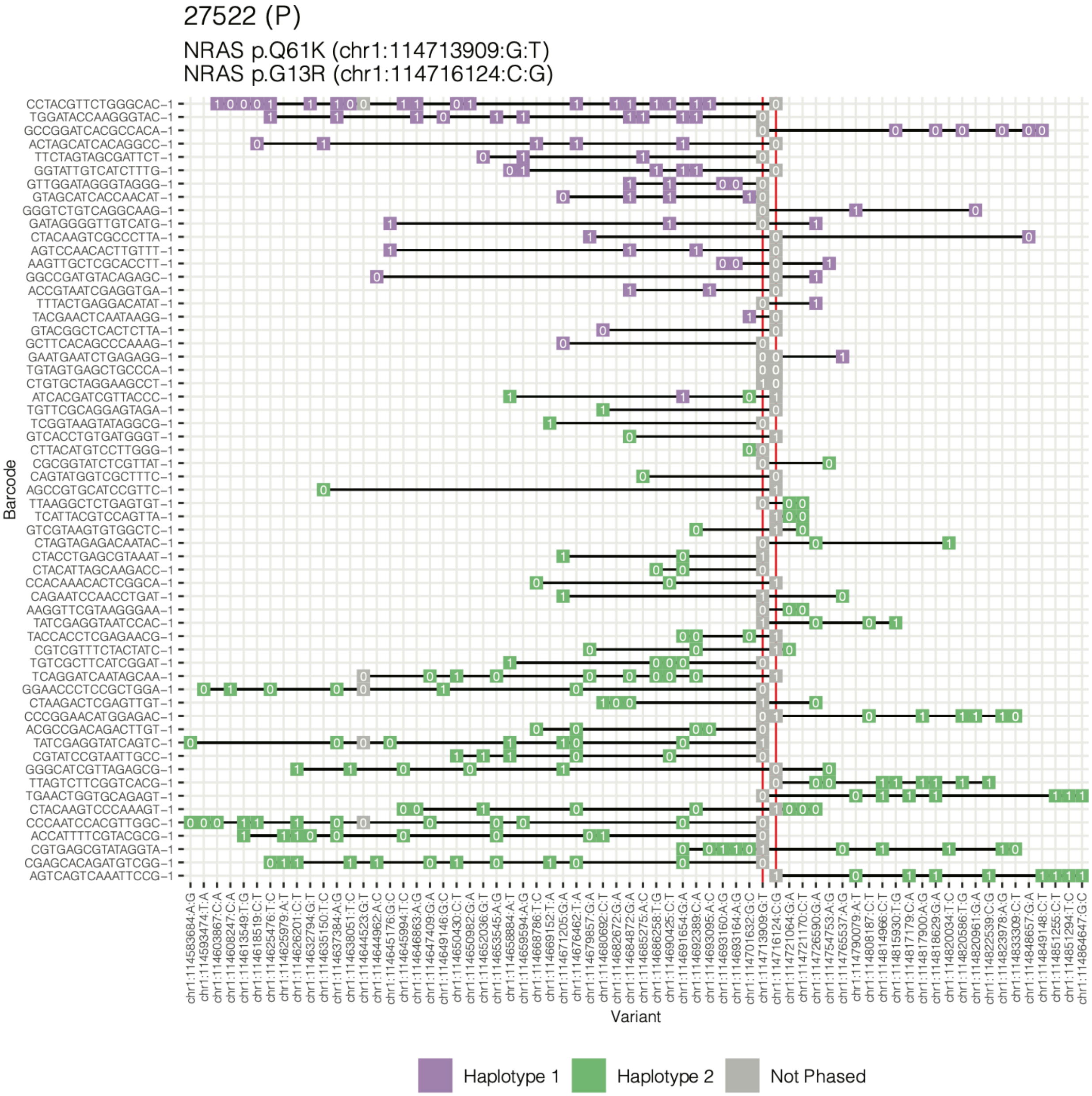


Figure 4. Tumor evolution models derived from observed mutation pairs. a. Number of overlapping barcodes by distance between somatic mutations. **b.** Proportion of somatic mutation pairs in close proximity sharing barcodes. **c.** Patterns of mutation pairs observed on barcodes. **d.** NRAS mutation pair observed in 27522 (P) and evolution model. **e.** Interpretation of evolution model observed from NRAS mutation pair in 27522 (P). **f.** ACTG1 mutation pair observed in 27522 (Rel) and evolution model. **g.** Interpretation of evolution model observed from ACTG1 mutation pair in 27522 (Rel).

Next we considered the pairwise relationship of the reference and alternate alleles observed on the same barcodes (Fig. 4c). Of the 2,042 remaining mutation pairs, most (53.3%) only share barcodes that cover both reference alleles (REF/REF) and both alternate alleles (ALT/ALT). This means they have at least one barcode where both alleles are reference and at least one barcode where both alleles are alternate. Other observed patterns of partnership are less common, but include REF/REF in addition to REF/ALT or ALT/REF, in which there is at least one barcode supporting one of the alternate alleles but not both, despite both sites having some alternate allele present. 6.7% of pairs show reads supporting REF/ALT and ALT/REF, and if the two alternate alleles are phased to the same haplotype, this could indicate that the two mutations do not co-occur on the same molecule. Finally, 7.1% of pairs have a pattern of REF/ALT or ALT/REF along with ALT/ALT. Since the same mutation is not likely to recur independently within the same tumor, it is more parsimonious to conclude that one mutation preceded the other in the clonal evolution tree.

Such patterns of somatic mutations may be informative for refining tumor phylogenies and may have clinical implications. For example, in 27522 (P) we observed two hotspot mutations in NRAS (G13R and Q61K) (Fig. 4d). NRAS is a known cancer driver oncogene and mutations may lead to dysregulation of the Ras pathway. Our phasing analysis placed both mutations on the same haplotype (H2) (Supplementary Figure 6). We observed 2 barcodes supporting REF/REF, 1 barcode supporting REF/ALT, and 1 barcode supporting ALT/REF. Based on sorted lrWGS data, at the primary timepoint, the G13R VAF was 35.7% and the Q61K VAF was 22.2%, while at relapse, the G13R VAF was 20.5% and the Q61K mutation was not detected (VAF 0.0%). Without the benefit of phasing, one possible interpretation could have been that Q61K occurred in the same clone as G13R and then the double mutant subclone was eliminated after therapy. However, with linked reads, we directly observed both mutations occurring without the other, and we do not observe them together, guiding the interpretation that these mutations occurred independently in separate subclones, and then the Q61K subclone was later lost (Fig. 4e).



Supplementary Figure 6. Barcodes supporting 27522 (P) NRAS hotspot mutation pair.

In another instance, we detected a pair of mutations that may have occurred in sequential order on the same haplotype. Two synonymous mutations in ACTG1 showed a pattern of first affecting G156 site and then later L104 in a co-occurring way. Six barcodes demonstrate the ALT/REF pattern, and 24 barcodes had ALT/ALT (Fig. 4f). Under a parsimonious model in which the same mutation occurs only once, the mutation at the first position preceded the mutation at the second position. Since there are barcodes supporting both mutant alleles simultaneously, the mutations must occur within the same cells and we interpret this to mean the cells with both mutations form a later subclone within the subclone of cells with only one ACTG1 mutation (Fig. 4g). We also noted that there is an elevated copy number in this region (estimated to be 2.65). This would often preclude clonality analysis due its effect on the VAF.[152] However, the combination of alleles present on the same barcodes enables us to interpret a sequential order of events.

Common myeloma translocations map to specific haplotypes

Multiple myeloma is characterized by recurrent clonal translocations that take advantage of overexpressed IGH locus by dysregulating oncogene expression. Barwick, et al. [92] analyzed 795 newly-diagnosed multiple myeloma patients from the Multiple Myeloma Research Foundation CoMMpass study (NCT01454297), reporting clonal translocations in across the cohort, including 16% of patients having t(11;14) impacting CCND1, 11% with t(4;14) (WHSC1), 3.3% with t(14;16) (MAF), 1.1% with t(6;14) (CCND3), and 1.0 % with t(14;20) (MAFB). Manier, et al. [65] reported similar translocation recurrence rates. In our cohort of 14 patients, we detected common myeloma translocations from lrWGS using the Long Ranger pipeline as well as gemtools.[153] We detected translocations affecting patients across multiple disease stages, including t(11;14) in 2 patients and t(4;14) in 1 patient and identified the haplotype affected by each translocation.

We focused on events reported in Patients 27522 and 77570. In 27522, 6 out of 7 SVs detected from both Primary and Relapse samples were also detected by Manta from a later sorted WGS sample (Fig. 5a). In 27522 (P), the barcodes supporting t(4;14) originated from two phase sets in the IGH region on chr14, and in both phase sets the assigned Haplotype was H2. In 27522 (Rel), the same t(4;14) event was detected but the two IGH phase sets were in opposite phase. Patient 27522 had a t(4;14) event detected at primary diagnosis and at relapse (Fig. 5b-c). The barcodes supporting the translocation are also associated with a deletion in the IGH region (Supplementary Figure 7a-b). The translocation juxtaposes IGH with WHSC1 and FGFR3, leading to overexpression of both oncogenes. WHSC1 overexpression increases methylation of H3K36 and further dysregulation across the genome. The coverage heat map showing where discordant barcodes map on chr4 and chr14 clearly identified the translocation breakpoint within the first intron of WHSC1 at chr4:1871962 and near IGHM on chr14. The barcodes supporting the translocation also showed a deletion in the IGH region.

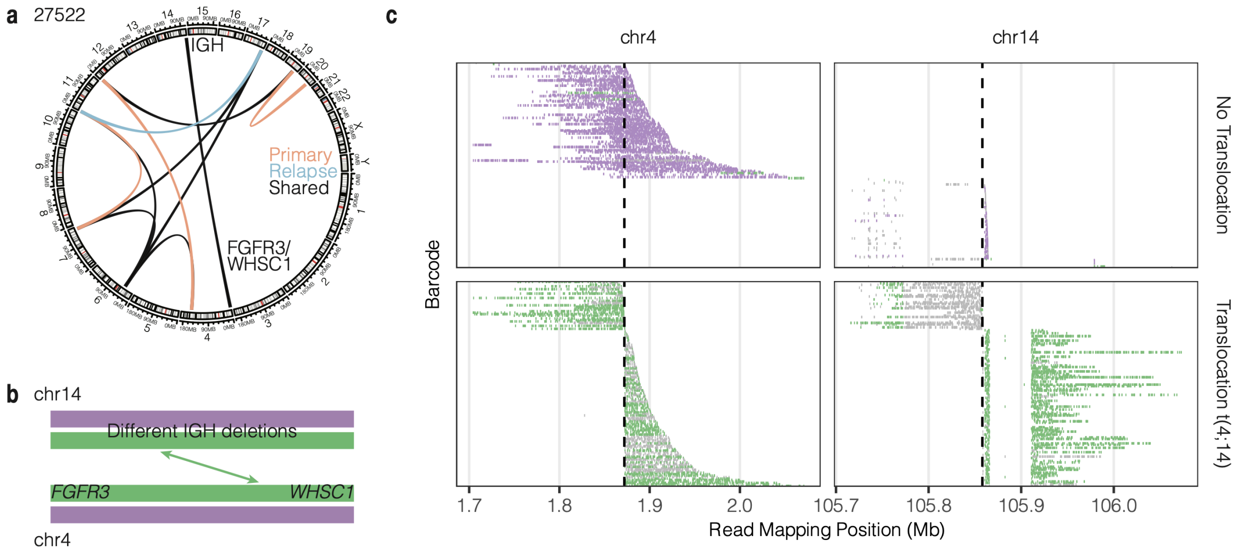
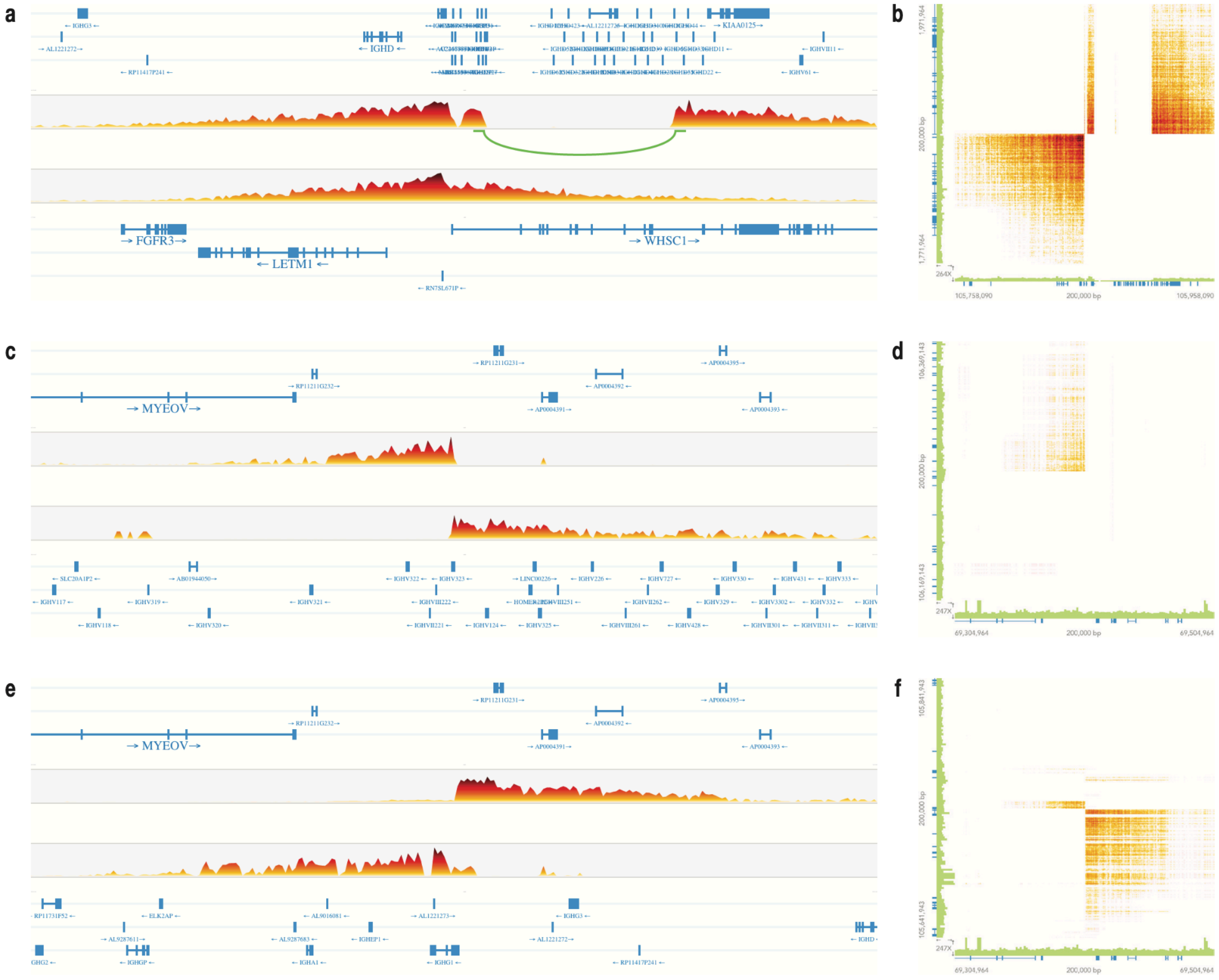


Figure 5. Common myeloma translocations mapped to haplotypes. a. Overlap of translocations observed in 27522 (P) and (Rel). **b.** Model of t(4;14) translocation. **c.** Barcodes supporting translocation indicate a single haplotype origin.



Supplementary Figure 7. Barcode support for common myeloma translocations. a-b. 27522 (P) t(4;14). **c-f.** 77570 (P) t(11;14).

For 77570 (P), Long Ranger reported multiple distinct t(11;14) events with translocations affecting two different regions of IGH to the same breakpoint upstream of CCND1 (Supplementary Figure 7c-f). One linked the IGH variable gene region (chr14:106269142) to the region upstream of CCND1 on chr11. The other at chr14:105741942 linked the coding region of IGHG1 to the same upstream CCND1 location. Barcode analysis suggests these are actually one reciprocal event.

An application of translocation mapping would be to match allele specific expression to translocation events, for example if a germline heterozygous coding variant from the same haplotype of the dysregulating translocation were detected from RNA-seq, then the connection between translocation and expression could made more explicitly.

Overlapping germline variants from paired samples enables phase set extension

Phase set size may be limited by random chance due to the distribution of linked-reads mapping locations, and phase set boundaries differ between samples originating from the same patient, in general. However, samples from the same patient do share germline variants, and those should be phased together in the same groups in both samples. By comparing the pattern of germline variants assigned to each haplotype in each sample, we can determine if the two phase sets are oriented the same way, or if one needs to be flipped to be consistent. We built the *extend* module into SomaticHaplotype to compare germline variants overlapping phase sets found in two samples, the target sample and the reference sample (Fig. 6a). When there is significant evidence of exact matches or exact mismatches to know if two phase sets have the same or opposite haplotype orientation, the module may recommend switching the orientation of the target sample phase set. If two phase sets from the target sample both overlap the same phase set from the reference sample, then we may be able to infer the haplotype orientation relationship of the target phase sets. If both target phase sets need to be switched to be consistent with the reference sample, then they already have the same orientation. If one target phase set needs to be switched and the other not switched, then we know they have opposite orientation. The *extend* module builds a bipartite graph in which nodes are phase sets and edges connect overlapping target and reference sample phase sets. Edge weight is defined as 1 if a switch is necessary between the target and reference phase set or 2 if a switch is not necessary. If two target phase sets overlap the same reference phase set, then there is a connected path between the target phase sets; we find the sum (mod 2) of weighted edges that connect pairs of target phase sets. If the sum is even, then the two target phase sets have the same orientation. If the sum is odd, then they have opposite orientation.

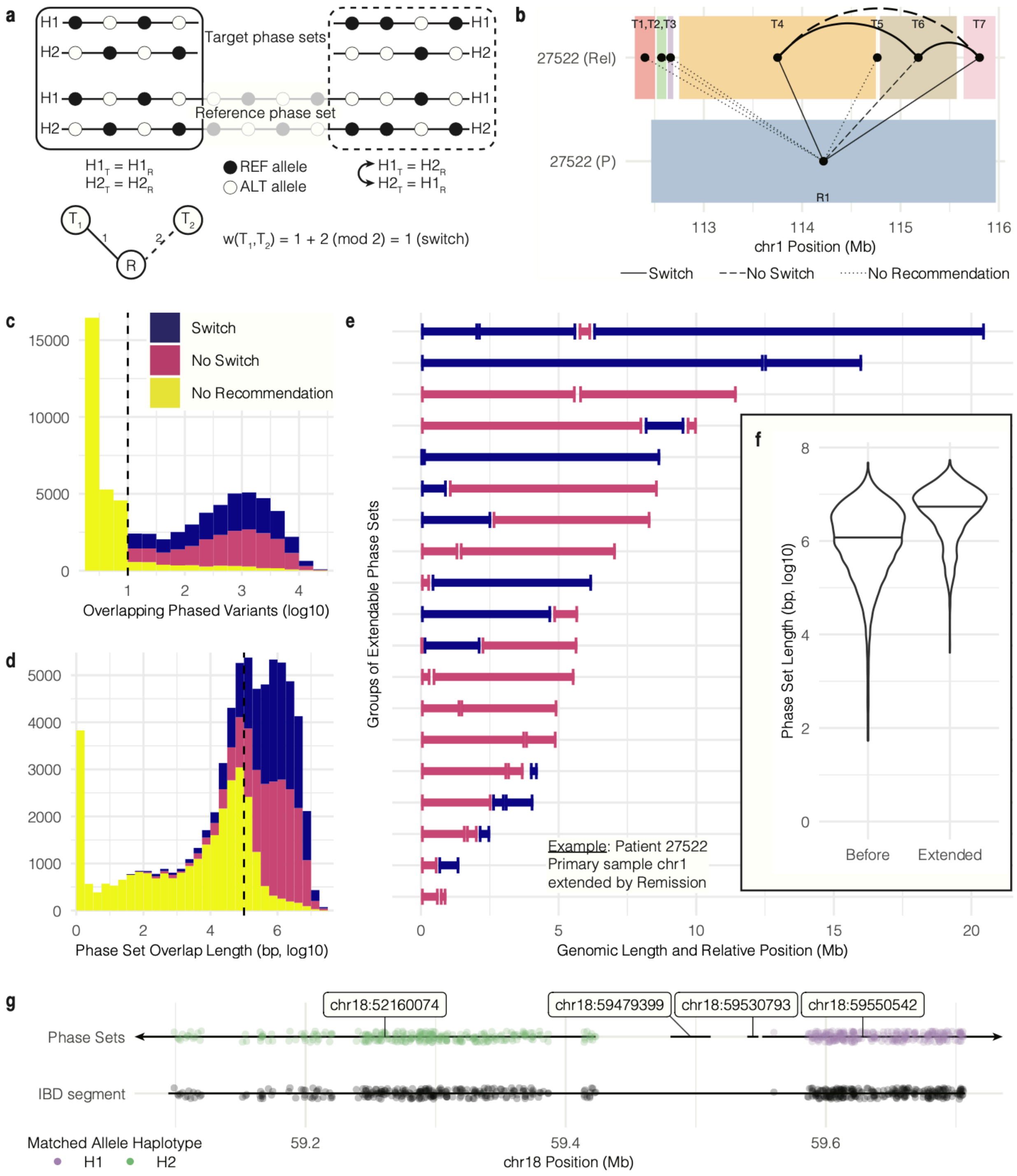


Figure 6. Extension of phase sets using additional sample information. a. Model for phase set extension. **b.** Data-driven example of phase set overlap between samples. **c.** Number of phased variants needed for switch/no switch recommendation. **d.** Length of phase set overlap needed for switch/no switch recommendation. **e.** Phase set groups extended by overlap with another sample. **f.** Distribution of phase set lengths before and after extension. **g.** Use of identity-by-descent segments as overlap between phase sets.

We analyzed data from 6 patients with multiple samples, with a total of 68,374 overlapping phase sets from 26 target and reference sample pairs. As a data-driven example, we examined phase set originating from chromosome 1 of 27522 (P) and 27522 (Rel), using 27522 (P) as the reference sample (bottom) and 27522 (Rel) as the target sample (top) (Fig. 6b). Reference phase set 1 (R1) (colored blue) spans multiple target phase sets (T1-T7). For T1, T2, T3, and T5, there are not enough overlapping variants to draw conclusions about their orientation relative to R1. Phase sets T4 and T7 must be switched in order to be consistent with R1, and T6 is already in the same orientation. Thus, since T4 and T7 have the same orientation relative to R1, T4 and T7 do not need to be switched to be consistent with each other. However, T6 must be switched to be consistent with T4 and T7.

We analyzed how much overlap is required before our testing methods give a solid switch or no switch recommendation. In general, at least 10 overlapping phased variants are required before making a switch or no switch recommendation (Fig. 6c). Since the number of shared variants correlates with the length of the overlap, the length of overlap tends to be greater than 100 kb before a recommendation can be made (Fig. 6d). Since haplotype numbering is random, we were not surprised to find roughly equal proportions of recommendations to switch (28.3%) and not switch (27.6%). The algorithm made no recommendation for the remaining 44.1%. For extendable phase sets from chromosome 1 in target sample 27522 (P) (extended by reference 27522 (Rem)), we found that, before extension, the median phase set length was 1.6 Mb, and after extension, it was 5.7 Mb, a nearly 3.5-fold increase. Similarly, from all samples with extendable phase sets, we found that median phase set length increased from 1.2 Mb (6.1 on log10 bp scale) to 5.5 Mb (6.7 on log10 bp scale), an increase of 4.6 fold increase from before extension to after extension.

We also developed methods (*ancestry* module) to utilize publicly available phased resources to improve our lrWGS data. We used data from 1000 Genomes sample NA12878 to illustrate this. After reporting identical-by-descent segments shared between from 2,504 individuals from 1000 Genomes data (see **Methods**), we identified IBD segments that overlap multiple lrWGS phase sets. Using overlapping, phased heterozygous variants shared between the 1000 Genomes VCF and the VCF output by Long Ranger, we found the proportion of IBD alleles that matches each haplotype in each phase set. IBD alleles matched one haplotype or the other, with the occasional short switch error (calculate error rate). For example, NA12878 shares an IBD segment with NA10851 spanning from position 59,094,547 to 59,706,930 on chromosome 18 (LOD score 15.64, 1.576 cM). That IBD segment bridges multiple lrWGS phase sets. Since the IBD alleles match Haplotype 2 from phase set chr18:52160074 and match Haplotype 1 from chr18:595505042, those two phase sets may be in opposite orientation. The *ancestry* module also reports the population group and subgroup of the individual associated with each IBD segment, linking public database ancestry information to haplotypes from lrWGS data.

Discussion

As sequencing technologies evolve and analysis methods more regularly include haplotype phasing, somatic mutation phasing may become a more common practice. The current methodological approaches to phasing-aware somatic SNP mutation analysis will mature from ad-hoc investigations to standard pipelines. We have developed a systematic approach to somatic mutation analysis in a cohort of multiple myeloma patients over the course of their disease. Our methods build up the backbone of the Long Ranger variant calling and phasing pipeline for linked-read sequencing data. We combine information from additional sequencing data to specifically target somatic mutations and infer their phase based on neighboring linked alleles. We also augment samples with shared data from other samples from the same individual to extend phase sets beyond their original limits. These tools are an open-source opportunity for future methods development in a climate of rapid technological shifts.

In the course of our analysis of this data, we noted several limitations and guidelines for data quality. Our normal controls came from skin samples, and we observed severe limitations in phasing performance potentially due to the DNA molecule input size. We generally expect longer fragments from blood samples used as normal controls. For our somatic analysis, one limitation was the prevalence of copy number changes in our data. Once a copy number alteration occurs, a strict two haplotype paradigm of mutation phasing must adapt. This is especially true for determining the haplotype relationship of pairs of mutations, where we may be confident that that haplotype looks like one of the two inherited copies, but we need additional information to know if two mutations occurred on the same copy or not. Another caveat to our somatic analysis was the tumor purity available in our samples. Four of our samples were CD138+ sorted, and two samples in particular gave us the best results. Higher tumor purity and lower variability in cell type composition are likely important for robust somatic variant analysis. Further, calling somatic mutations with low variant allele frequency, compounded by lower tumor purity, is a challenge for any mutation caller, especially those like Long Ranger built for germline variant detection. In our case, pairing linked-read data with high-confidence somatic mutation calls from a separate sample was necessary to gain sensitivity.

Moving beyond next-generation sequencing to Third Generation and single-cell approaches has the distinct advantage of increasing the resolution of cancer genome analyses.

With long reads and linked-reads, we get haplotype resolution. With single cell RNA-seq, we observe cell-specific patterns of gene expression and can even map coding mutations to specific cells.[102] Single cell DNA offers further resolution of haplotype structures and clonal structure, giving breath to dreams of reconstructing tumor phylogenies, understanding tumor evolution, and identifying optimal treatment targets.[154-156] Methodological integration of single cell data with the resolution gained from haplotype analysis is a direction for continued research. Tools that incorporate single cell copy number may be more robust than single cell RNA-seq mutation mapping and enable phylogenetic inference, especially across longitudinal samples.

Methods

Patient cohort

Fourteen patients with multiple myeloma, 10 male and 4 female Caucasians, were included in the analysis. The median age at diagnosis was 63 (range 46-69). Eight patients had IgG isotype, 4 being kappa light, chain and 4 being lambda light chain, 2 had IgA kappa isotype, 2 had light chain only disease (1 kappa and 1 lambda), and 2 were non-secretory. Five were International Staging System Stage 1, two were Stage 2, 3 were stage 3, and 4 were unreported. The median plasma cell burden by flow cytometry in bone marrow at diagnosis was 24% (range 4-63). By standard fluorescence in situ hybridization (FISH), 1 patient had t(4;14), 3 had t(11;14), and 2 showed del(17p).

The data comprises 14 patients having various combinations of sample types, time-points, data types, and treatment modalities. Most patients have 10x whole genome sequencing (10xWGS) data for skin normal and pre-treatment state, with several having relapse data, as well. Treatment ranges from none for 7 patients (3 of which have an SMM sample) to multi-cycle regimens of several 2-drug and 3-drug cocktails. All WES and WGS data are generated with CD138+ sorted population (tumor cells) within bone marrows. To ensure samples matched across time points, we compared germline variant allele fractions (VAF) at 24 loci (data not shown).

Sequencing data generation

Research bone marrow aspirate samples were collected at the time of the diagnostic procedure. Bone marrow mononuclear cells (BMMCs) were isolated using Ficoll-Paque. BMMCs were cryopreserved in a 1:10 mixture of dimethyl sulfoxide and fetal bovine serum. Upon thawing, whole BMMCs were used for linked-read whole genome sequencing. Plasma cells were separated from a sub-aliquot by positive selection using CD138-coated magnetic beads in an autoMACs system (Miltenyi Biotec, CA) and used for whole genome and exome sequencing. Skin punch biopsies were performed at the time of the diagnostic bone marrow collection to serve as normal controls. Although many studies use peripheral blood mononuclear cells (PBMCs) as a control, abnormal B cells and circulating tumor cells frequently contaminate the peripheral blood of patients with multiple myeloma. Therefore, using PBMCs may lead to omission of genetic events potentially important in disease pathogenesis.

Linked-read whole genome sequencing (lrWGS) -- Normal skin samples were processed with a standard Qiagen DNA isolation kit resulting in 10-50Kb DNA fragments. 250K tumor cells were processed with the MagAttract HMW DNA extraction kit (Qiagen) resulting in 100-150Kb DNA fragments. 600-800ng of normal DNA was size selected on the Blue Pippin utilizing the 0.75% Agarose Dye-Free Cassette to attempt to remove low molecular weight DNA fragments. The size selection parameters were set to capture 30-80 Kb DNA fragments (Sage Science). The resulting size selected DNA from the normal samples and the HMW DNA from the tumor cells were diluted to 1ng/µL prior to the v2 Chromium Genome Library prep (10x Genomics). Approximately 10-15 DNA molecules were encapsulated into nanoliter droplets. DNA molecules within each droplet were tagged with a 16 nucleotide barcode and 6 nucleotide unique molecular identifier during isothermal incubation. The resulting barcoded fragments were converted into a sequence ready Illumina library with an average insert size of 500bp. The concentration of each library was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for sequencing on the HiSeqX/NovaSeq6000 platform (Illumina). 2x150 sequence data were generated targeting 30x (normal) and 60x (tumor) coverage providing linked reads across the length of individual DNA molecules.

Standard whole genome sequencing (WGS) -- Manual libraries were constructed with 50-2000ng of genomic DNA utilizing the Lotus Library Prep Kit (IDT Technologies) targeting 350bp inserts. Strand specific molecular indexing is a feature associated with this library method. The molecular indexes are fixed sequences that make up the first 8 bases of read 1 and read 2 insert reads. The concentration of each library was accurately determined through qPCR (Kapa Biosystems). 2x150 paired end sequence data generated ~200 Gb per tumor sample leading to 60x (tumor) haploid coverage.

Standard whole exome sequencing (WXS) -- A 700ng aliquot of the existing WGS library was used for the exome capture. Five libraries were pooled at an equimolar ratio yielding a ~3.5µg library pool prior to the hybrid capture. The library pools were hybridized with the xGen Exome Research Panel v1.0 reagent (IDT Technologies) that spans a 39 Mb target region (19,396 genes) of the human genome. The concentration of each captured library pool was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for sequencing on the NovaSeq6000 platform (Illumina). 2x150bp sequence data was generated ~50Gb per library targeting a mean depth of coverage of 500x.

Processing linked-read WGS with Long Ranger (alignment, variant calling, phasing) -- We used Long Ranger (v2.2.2) (10x Genomics) for preliminary analysis, including demultiplexing cDNA libraries into FASTQ files and aligning reads to the human genome reference GRCh38 (GRCh38-2.1.0). To call variants using Long Ranger, we used --vcmode with GATK (version 3.7.0-gcfedb67). Long Ranger phasing quality metrics were extracted from the summary output file associated with each sample. Show full code run and output files generated.

Somatic mutation detection

Somatic variants were called by our SomaticWrapper pipeline, which includes four established bioinformatic tools, namely Strelka[136], Mutect[134], VarScan2[137] (2.3.83), and Pindel[135] (0.2.54). We retained SNVs and INDELs using the following strategy: keep SNVs called by any 2 callers among Mutect, VarScan, and Strelka and INDELs called by any 2 callers among VarScan, Strelka, and Pindel. For these merged SNVs and INDELs, we applied coverage cut-offs of 14X and 8X for tumor and normal, respectively. We also filtered SNVs and INDELs with a high-pass variant allele fraction (VAF) of 0.05 in tumor and a low-pass VAF of 0.02 in normal. The SomaticWrapper pipeline is freely available from https://github.com/ding-lab/somaticwrapper.

Copy number profiling

We used BIC-seq2[124], a read-depth-based CNV calling algorithm to detect somatic copy number variations (CNVs) using standard WGS tumor samples and paired skin linked-read WGS data (human genome GRCh38 reference). The procedure involves 1) retrieving all uniquely mapped reads from the tumor and paired skin BAM files, 2) removing biases by normalization

(NBICseq-norm_v0.2.4) 3) detecting CNV based on normalized data (NBICseq-seg_v0.7.2) with BIC-seq2 parameters set as --lambda=90 --detail --noscale --control. In WXS data, we used CNVkit[157] (v0.9.4) to compare our tumor samples to a background panel of normals.

Structural variant detection

Somatic structural variants (SVs) were detected by Manta[139] using tumor/normal sample pairs of standard WGS and paired skin linked-read WGS. To filter false positive SVs, we removed events with somatic score < 30 and junction somatic score < 30.

Extracting lrWGS reads supporting somatic mutations

We used a mapping tool (10Xmapping), to identify reads supporting the reference allele and variant allele covering the variant site for each somatic mutation and gathering molecular barcode and haplotype information from the bam file. The tool is freely available at https://github.com/ding-lab/10Xmapping which is then contained as a submodule in https://github.com/ding-lab/SomaticHaplotype.

Subclonal analysis

The R package SciClone[152] was used to define clonal architecture, and tumor phylogeny was illustrated using the R package Fishplot[158].

Data availability

1000 Genomes samples downloaded from https://support.10xgenomics.com/genome-exome/datasets.

Chapter 5: Co-evolution of tumor and immune cells during progression of multiple myeloma

Our work, Co-evolution of tumor and immune cells during progression of multiple myeloma, was submitted to *Nature Cancer*. Please refer to the eventual publication for any supplementary tables and extended data figures. Contributions: As co-first author with Ruiyang Liu and Qingsong Gao, SMF developed haplotype-based somatic mutation phasing, clonality analysis, plasma cell evolution analysis, single-cell visualization techniques, and helped lead the overall project organization, manuscript writing and revision, and figure design.

Summary

Multiple myeloma (MM) is characterized by the uncontrolled proliferation of plasma cells. To investigate MM and its immune environment, we applied single cell RNA and linked-read whole genome sequencing to profile 29 longitudinal samples at different disease stages from 14 patients. We collected 17,267 plasma cells and 57,719 immune cells, discovering patient-specific plasma profiles and immune cell expression changes. Patients with the same genetic alterations tended to have both plasma cells and immune cells clustered together. We noted distinct T cell clusters in the tumor microenvironment, which may be associated with common translocation events present in the tumor. By integrating genomics and single cell mapping, we tracked plasma cell subpopulations across disease stages and found three patterns: stability (from precancer to diagnosis), and gain or loss (from diagnosis to relapse). In multiple patients, we detected "B cell-featured" plasma cell subpopulations that cluster closely with primitive and mature B cells, implicating their cell of origin. We validated AP-1 complex differential expression (e.g. JUN and FOS) in plasma cell subpopulations using CyTOF-based protein assays, and integrated analysis of single cell RNA and CyTOF data revealed AP-1 downstream targets (e.g. IL6 and IL1B) potentially leading to inflammation regulation.

Introduction

Multiple myeloma (MM) is a disease characterized by clonal proliferation of malignant plasma cells (PCs), sometimes manifesting clinically with anemia, renal impairment, and pathologic bone fractures [159,160]. Over the past three decades, novel therapies, such as autologous hematopoietic cell transplantation, proteasome inhibitors (PIs), immunomodulatory drugs (IMiDs), and targeted monoclonal antibodies have led to dramatic improvements in quality and length of life in patients with multiple myeloma [161-165]. Despite these advances, the disease remains incurable for most patients as it progresses and becomes resistant to these treatments.

Several landmark genomic studies have led to a greater understanding of the molecular pathogenesis of myeloma. These studies have demonstrated recurrent mutations in *KRAS, NRAS*, and *TP53*, as well as a significant percentage of previously unrecognized mutations affecting RNA processing and protein homeostasis [83,166,167]. Other investigations have used bulk sequencing technologies to broadly describe MM clonal heterogeneity and evolution in terms of shifting subclonal dominance and branching evolution, often in response to therapeutic selective pressure [165,168,169]. There is an impetus to translate the growing understanding of the genomic landscape of MM into precision therapies. This is highlighted by the upcoming MyDRUG trial (NCT02884102) being initiated by the Multiple Myeloma Research Foundation (MMRF), which will use genomic and transcriptomic information obtained from the CoMMpass study (relating clinical outcomes to assessment of individual genetic profiles) in order to identify targetable genetic alterations and to evaluate personalized therapies to enrollees.

Single-cell sequencing methods combine novel sequencing technologies with cell-sorting techniques, allowing for a more granular understanding of inter- and intra-tumoral genomics [170]. Early studies used low-throughput systems to analyze the tumor microenvironment in solid tumors, examining the genomes and transcriptomes of malignant cells, as well the immune compartment, confirming the importance of single-cell resolution [125,170,171]. With the advent of high-throughput methods, these technologies are rapidly expanding toward dissecting all malignancies. Ledergor et al. (2018) recently used single cell RNA sequencing (scRNA-seq) to compare plasma cell transcriptomes from patients with newly diagnosed MM (NDMM), precursor states, and healthy controls; they highlighted significant inter-individual heterogeneity and demonstrated variable subclonal divergence leading to new thoughts about the role of intergenic mutations, epigenetics, and environmental transcriptional regulation [172]. Jang et al. (2019) used scRNA-seq to examine 597 CD138+ plasma cells from 15 patients at different stages of MM, associating clusters of gene expression with risk of early disease progression and cytogenetic abnormalities [173].

Multiple myeloma is a dynamic disease characterized by clonal evolution and immune modulation in response to therapeutic pressure. The aforementioned single cell studies did not examine MM patients at multiple points during their disease progressions, nor did they evaluate dynamic alterations in non-malignant components of the tumor microenvironment. Here, we report our analysis of single-cell patterns in 29 longitudinal samples procured at different disease stages from 14 MM patients. We collectively analyzed 74,386 single cells from these patients, including 17,267 plasma cells and 57,719 immune cells. Deeper dissection of plasma cells and B cells identified subpopulations of plasma cells with various genetic changes and marker gene expressions, suggesting cells in transitional states. By single cell sequencing, we discerned co-evolution maps of tumor and immune cells between smoldering multiple myeloma (SMM) and primary stages and between primary and relapse stages after remission. In summary, our study represents the first longitudinal investigation of tumor and immune microenvironment during MM disease development and paves the way for expanding treatment options for this disease.

Results

Patients, treatments, technologies, and landscape of genomic alterations in multiple myeloma

The main data corpus of the study comprises 29 longitudinal samples from 14 individuals with different combinations of disease stages, sequencing data types, and treatments (Fig. 1a and Extended Data Fig. 1a, Supplementary Table 1). All patients have at least one sample with both single-cell RNA sequencing (scRNA-seq) and 10x Genomics linked-read whole genome sequencing (10xWGS), and 9 patients have data from two or more time points, including a mix of CD138+ sorted and unsorted bone marrow aspirate samples. Three patients have data from the SMM and primary stages, and six have both primary and relapse samples. To ensure samples matched across time points, we compared germline variant allele fractions (VAF) at 24 loci (Extended Data Fig. 1b, Supplementary Table 1). In addition, we performed CyTOF based profiling and validation using tumor samples from 4 additional patients.

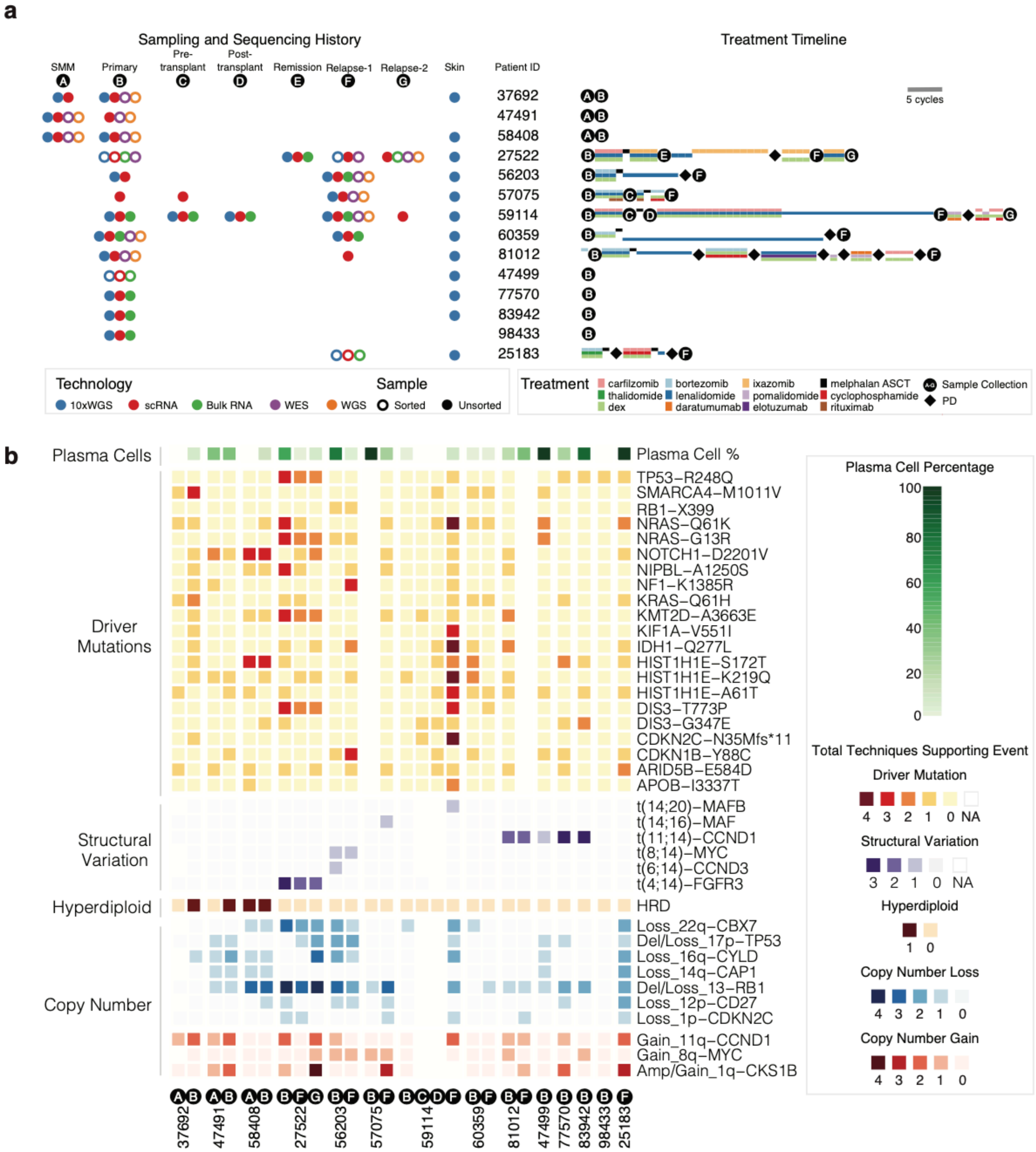


Figure 1. Samples, next generation data set, and genomics landscape. a. Sample type, technology, and treatment timeline broken down by patient. Left portion shows sample technology (10xWGS, scRNA, Bulk RNA, WES, WGS) and sample type (CD138+ sorted vs. unsorted). Right portion shows each patient's treatment timeline. Treatment length corresponds to number of cycles. **b**. Heatmap shows the landscape of Copy Number Variations (CNV), Structural Variants (SV) and driver mutations across 14 patients. Copy number amplification/gain, copy number deletion/loss, SV and driver mutations are shown in red, blue, purple and orange respectively, with colors indicating the number of techniques supporting the event. Techniques for copy number events are FISH, 10xWGS, regular WGS, WES, scRNA-seq. Techniques for SV are FISH, 10xWGS, Bulk RNA-seq, scRNA-seq. Techniques for driver mutations are 10xWGS, WES, WGS and Bulk RNA-seq. Number of techniques supporting an event is 0 if the only technique supporting the event is from scRNA-seq. Plasma cells percentage inferred from scRNA-seq is shown on the top of the heatmap.

Multiple myeloma exhibits a variety of primary and secondary genomic events (Fig.1b, Extended Data Fig. 1c and 2d). We analyzed potential driver events, focusing on known significantly mutated genes and structural and copy number variation (Fig. 1b, Supplementary Table 1). Three patients had hyperdiploid (HRD) copy number profiles with little evidence of translocation events, and in 12 patients, we observed loss of 13q supported by at least one level of evidence [174,175]. Most translocations in MM involve the highly-expressed *IGH* locus on chromosome 14, with t(11;14) being the most frequent [176] and t(4;14) being associated with adverse prognosis [81,177-180]. We have multiple evidence levels of t(11;14) in 3 patients and t(4;14) in 1 patient.

We detected a median of 55 coding mutations from whole exome sequencing (WES) and 6702 total mutations from whole genome sequencing (WGS) (Supplementary Table 1). The variant allele fraction distribution was consistent across sequencing platforms for key driver mutations, including *TP53*, *NRAS*, *KRAS*, and *DIS3* [7,65,181]. We observed VAF changes during disease progression for several mutations in cancer genes, notably *TP53* and *NRAS* in Patient 27522 and *APOB*, *CDKN2C*, *HIST1H1E*, and *IDH1* in Patient 59114. For example, *TP53*-R248Q in 27522 expands from 0.4% to 33.1%, while *NRAS*-Q61K recedes from 17.1% to 0.6% during progression from Primary to Relapse-1 (Extended Data Fig. 1c, Supplementary Table 1).

Tumor and immune populations influenced by genetic alterations and treatments during disease progression

We integrated scRNA-seq data from all 14 patients; after quality control and cell type detection (Methods), we retained 74,986 cells from 11 patients, including 17,267 plasma cells and 57,719 non-plasma cells. The proportions of plasma and immune cell types vary across patients and disease stages (Fig. 2a). Plasma cells in primary tumor samples ranged from 0.9% to 84.1%. Other cell types detected include B cells (3,686), macrophages (16,183), monocytes (4,249), CD4+ T cells (18,250), CD8+ cells (8,334), natural killer (NK) cells (6,282), and dendritic (DC) cells (735) (Figure 2a, Extended Data Fig. 3B). Different patients show a range of cell type compositions, such as complete loss of NK cells in Patient 27522 at the primary stage, but presence of 22% NK cells in Patient 77570 at the primary stage. Different stages from the same patient can also have different compositions as well. For example, in Patient 59114, CD4+ T cells change from 36% at Primary to 9% at both Pre and Post-transplant, and increase back to 35% at Relapse-1.

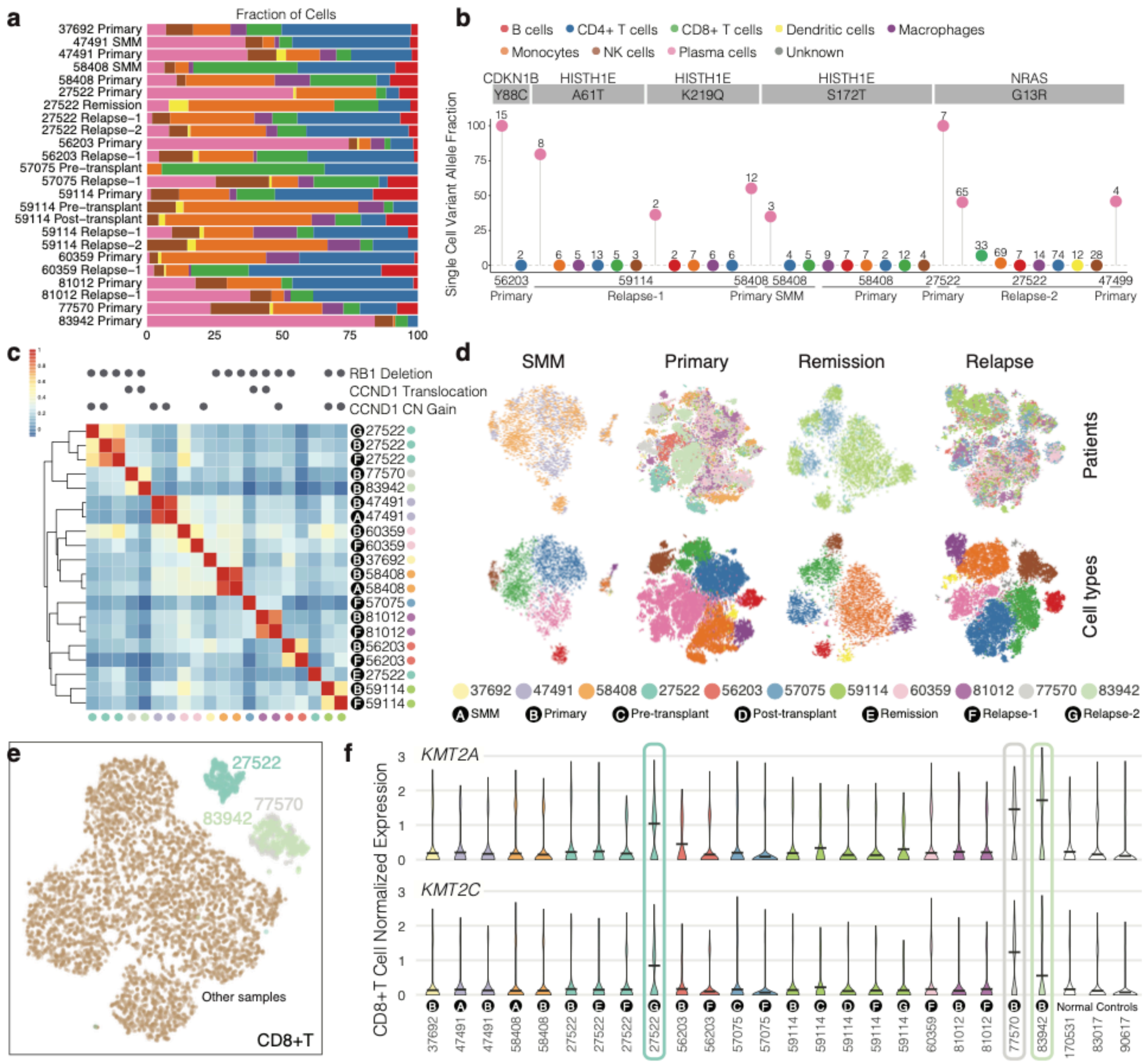


Figure 2. Integration analysis across 14 multiple myeloma patients revealing distinct cancer populations and immune microenvironments during disease progression. a. Bar plots showing cell type fractions for each sample. Colors indicate cell type. **b**. Single cell variant allele fractions (VAF) for driver mutations. Each bubble is colored by the cell type with the associated VAF, and total cells supporting the variant are labeled atop each bubble. **c**. Heatmap showing pairwise correlation of average expression for malignant cells in each sample. Genomic alterations with either FISH evidence or at least another two levels of evidence shown above. **d**. t-SNE plots showing the integration of samples from multiple patients for a given timepoint. Clustering of cells from different timepoints are colored by patient (top) or by cell type (bottom). The remission group includes one remission sample, one pre-transplant, and one post-transplant. **e**. t-SNE plot showing CD8+ T cells from all the patients where CD8+T cells are available. Cells from the primary sample of Patients 77570 and 83942 and Relapse-2 sample of Patient 27522 are colored specifically. **f**. Expression pattern of *KMT2A* and *KMT2C* in CD8+T cells for each sample.

Mapping somatic mutations to individual scRNA cells has the potential to identify tumor cells that cannot be discerned purely by expression data or subclonal populations with different mutational patterns [102]. Overall, we mapped 48 mutations to 198 cells from 14 samples (Extended Data Fig. 2b-c, Supplementary Table 2, Methods). Variants in key driver genes, such as *NRAS* G13R mutation, were primarily detected in plasma cells (158 cells) relative to non-malignant cell types (39 cells), which are much more numerous. The reference allele was detected more readily across cell types (1212 plasma cells and 5278 non-plasma cells) (Fig. 2b). We also examined mutations co-residing in the same cells (Extended Data Fig. 2a), finding that mutations *NRAS*-G13R, *YBX1*-F74L, *ACAT1*-S14N, *CLPTM1L*-T33S, and *DIS3*-T773P serve as hubs for a mutational network in the 27522 Relapse-2 sample.

Single cell expression profiles of plasma cells primarily clustered by each individual patient, with different disease stages of the same patient showing high similarity (Fig. 2c, Extended Data Fig. 4a-b), while expression of non-plasma cells largely cluster by cell types (Extended Data Fig. 3a). Notably, we observed the highest correlation between SMM and primary tumors (0.92 for Patient 47491 and 0.91 for Patient 58408), but lower and more variable correlation between primary and relapse samples in other patients. Expression profiles also partially clustered by genetic alterations; Patients 77570 and 83942 both harbor *CCND1* translocation, and their plasma cell expression profiles are more similar than others (Fig. 2c, Supplementary Table 3, Methods).

We also integrated samples from multiple patients by disease stage (Fig. 2d, top row colored by patient, bottom row colored by cell type). We observed again that plasma cells tended to cluster by patient, and found that non-plasma cells clustered by cell type and included a broader mix of patients. We then identified genes with variable expression across disease stages in multiple patients. For example, we found CD4+ T cells from primary tumors show a higher expression of *NFKBIA* when compared to SMM. In Patient 27522, *NFKBIA* expression was lost during remission, but regained in relapse. *NFKBIA* is a negative regulator of NFkB, meaning cell types with higher *NFKBIA* might implicate altered NFkB activity. In another example, we found higher expression of *CD69* in CD4+ T cells of remission samples, which was subsequently lost during relapse. Higher expression of *IL1R2* was observed in primary sample monocytes but was then lost in remission monocytes. In monocytes found in the Relapse-2 sample of Patient 59114, there was a slight increase in *IL1R2* expression, and a similar trend was observed for *IL1B* expression in the monocytes of Patient 60359 (Extended Data Fig. 3d). Together these suggest a role of *IL1* signaling during myeloma, which should be further explored.

To evaluate differences of the tumor microenvironment across patients, we did another integration including an additional 4 samples from healthy donors. We then extracted cells from each non-tumor population for subclustering analysis. We found cells from different patients generally mixed well, but cells from some samples exhibited a consistent outlier pattern across cell types. This phenomenon is particularly seen for the Relapse-2 sample of Patient 27522 and for the Primary samples of Patients 77570 and 83942. Specifically, NK cells and especially CD4+ and CD8+ T cells from 77570 and 83942 overlapped showed similar overall expression profiles, further suggesting similar genetic alterations could shape similar tumor microenvironment (Fig. 2e, Extended Data Fig. 3b). Further investigation identified a set of genes exhibiting outlier expression pattern in these samples. For example, in CD8+T cells of these three samples, we found higher expression of *KMT2A* and *KMT2C*, two genes belonging to the lysine methyltransferase family, suggesting epigenetic changes in the T cell population (Fig. 2e-f). There is strong evidence of t(11;14) (*CCND1* translocation) in Patients 77570 and 83942, and t(4;14) (*WHSC1/MMSET* translocation) in Patient 27522, suggesting further study into the role these events might play in modifying the tumor microenvironment. We found high expression of *CTSS* in the macrophages and monocytes of Patient 77570 (Extended Data Fig. 3c). *CTSS* encodes Cathepsin S, a major endoprotease processing the MHCII complex prior to antigen presentation. It has been shown in mouse models that CTSS is necessary for the release of IL1B in macrophages[182], and that macrophage-derived cathepsin S induces chemoresistance in breast cancer [183] and invasion in pancreatic cancer [184]. Interestingly, for the general myeloid lineage cell types (macrophages, monocytes, DC), the Relapse-2 sample of Patient 27522 shows outlier expression of *TNFSF13* (APRIL) (Extended Data Fig. 3c). *TNFSF13* engages with the plasma cell-specific receptor *TNFRSF13B* (TACI) [185] and induces secretion of proinflammatory mediators such as IL-8 and MMP-9 [186], which could implicate the complex interaction within the tumor microenvironment.

Delineating B cell lineage by gene signature analysis and genetic alteration mapping

To study B cell lineage and the transition between normal and malignant plasma cells, we integrated B cells and plasma cells from 21 tumor samples with both cell types along with 4 healthy donors (Methods). After integration, we found that clusters separated by cell type (Fig. 3a), with mature B cells from each patient mapping to the same cluster as B cells from normal samples. There are three small B cell clusters predominantly from healthy donors that exhibit high expression of *SOX4, VPREB3*, and *MME*, suggesting a primitive B cell state [187,188]. Interestingly, we found plasma cells from healthy donors mixing with some MM plasma cells, meaning that these particular MM plasma cells exhibit an expression pattern similar to normal plasma cells. The rest of the MM plasma cells largely clustered by patient, as shown previously (Extended Data Fig 4a).

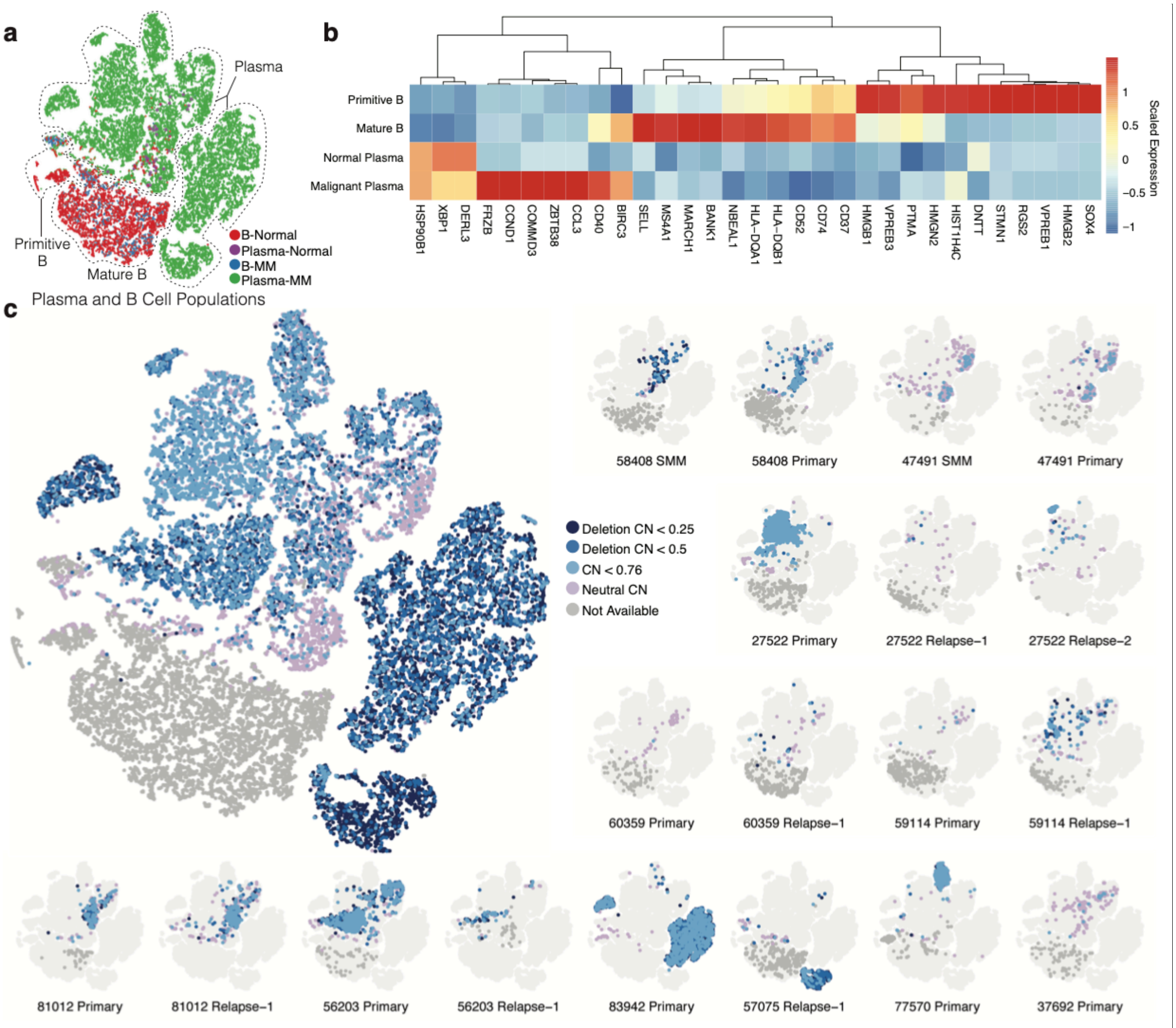


Figure 3. Analysis of B cell lineage markers and landscape for copy number events. a. t-SNE plot showing the distribution of B cells and plasma cells from all patients and four healthy "normal" donors. **b**. Heatmap showing genes specifically expressed at certain stages of B cell development. **c**. Landscape of chromosome 13 deletion status showing all samples (left), with sample-specific maps for samples with at least one cell with chromosome 13 copy number (CN) < 0.76 (right).

To investigate whether the malignancy of plasma cells is implicated from the early B cell stages, we also subset only the B cell populations for analysis. We found cells from some patient samples along with two normal samples (090617 and 170531) to be outliers. We found substantial B cell signatures in the plasma cells -- with high expression of typical plasma cell markers, such as *SDC1* and *TNFRSF17* -- for Patients 56203 and 83942, and to a lesser degree for 77570, as illustrated by expression of B cell marker *MS4A1* (Extended Data Fig. 4c). It has been previously reported that a subset of patients with high *CCND1* expression exhibits a B cell phenotype (CD2 group) [177], consistent with our observation for 77570 and 83942. For Patient 56203, there is also an aberrant *CCND1* regulation according to FISH report (data not shown), although *CCND1* translocation/overexpression is not observed. This suggests that aberrant *CCND1* regulation, not necessarily overexpression, may drive a switch back to B cell phenotype. Patient 81012, harboring a *CCND1* translocation, had elevated expression of *FYN* and *SETD7* (Extended Data Fig. 4c), consistent with the previously reported CD1 group [177].

We then identified genes differentially expressed across the B cell lineage, from primitive B cells to mature B cells and ultimately to normal and malignant plasma cells. We found four groups of overexpressed genes that defined each stage (Fig. 3b, Supplementary Table 4). The Primitive B group included *SOX4* and *DNTT*, along with several less-investigated genes in terms of lineage, such as *HMGB1* and *HMGB2*, both of which are involved in DNA double-strand breakage [189,190] and might be associated with VDJ recombination. The Mature B group was defined by *CD20* (*MS4A1*) and MHC-associated genes. The third group showed increased expression along the B cell lineage, with high expression in both normal and malignant plasma cells. As expected, ER stress response gene *XBP1* was overexpressed since plasma cells produce high levels of secreted proteins [191]. The final group showed high gene expression for malignant plasma cells only. Typical genes for this category include *FRZB*, *CD40*, *BIRC3,* and *ZBTB38.* Our discovery of B cell lineage genes is confirmed by the observation of increased expression of MHC II-related genes from primitive B cell to mature B cell stage prior to differentiation to plasma cells. Further, this observation is validated in our independent CyTOF experiment, where the CD38-low, CD45-high, mature population exhibits higher levels of HLA-DQA1 (Extended Data Fig. 5).

We also analyzed single cell copy number in B and plasma cells and found that 17 out of 21 samples showed chromosome 13 deletion (Fig. 3c, Extended Data Fig. 2d). Complete loss of chromosome 13 is associated with more aggressive malignancy than partial loss, in part because tumor suppressors such as *RB1* reside there. We identified clusters with deeper chromosome 13 deletion in 83942 Primary, 57075 Relapse-1, and 27522 Relapse-2, indicating possible homozygous deletion in their plasma cells. Clusters with deeper deletion tended to be patient and subpopulation-specific, while cells mapping to the same location as normal plasma cells tended to come from multiple patients and showed greater variability, including some with neutral CNV.

Distinct plasma cell subpopulations remain stable during transition from SMM to primary

We investigated how clonal structure evolves from SMM to primary diagnosis in three patients, 37692, 47491, and 58408 (Fig. 1a). Without exception, we found that plasma cells grouped into two geometrically distinct t-SNE subclusters (subpopulations) in both disease stages (Fig. 4a, Extended Data Fig. 6a-b).

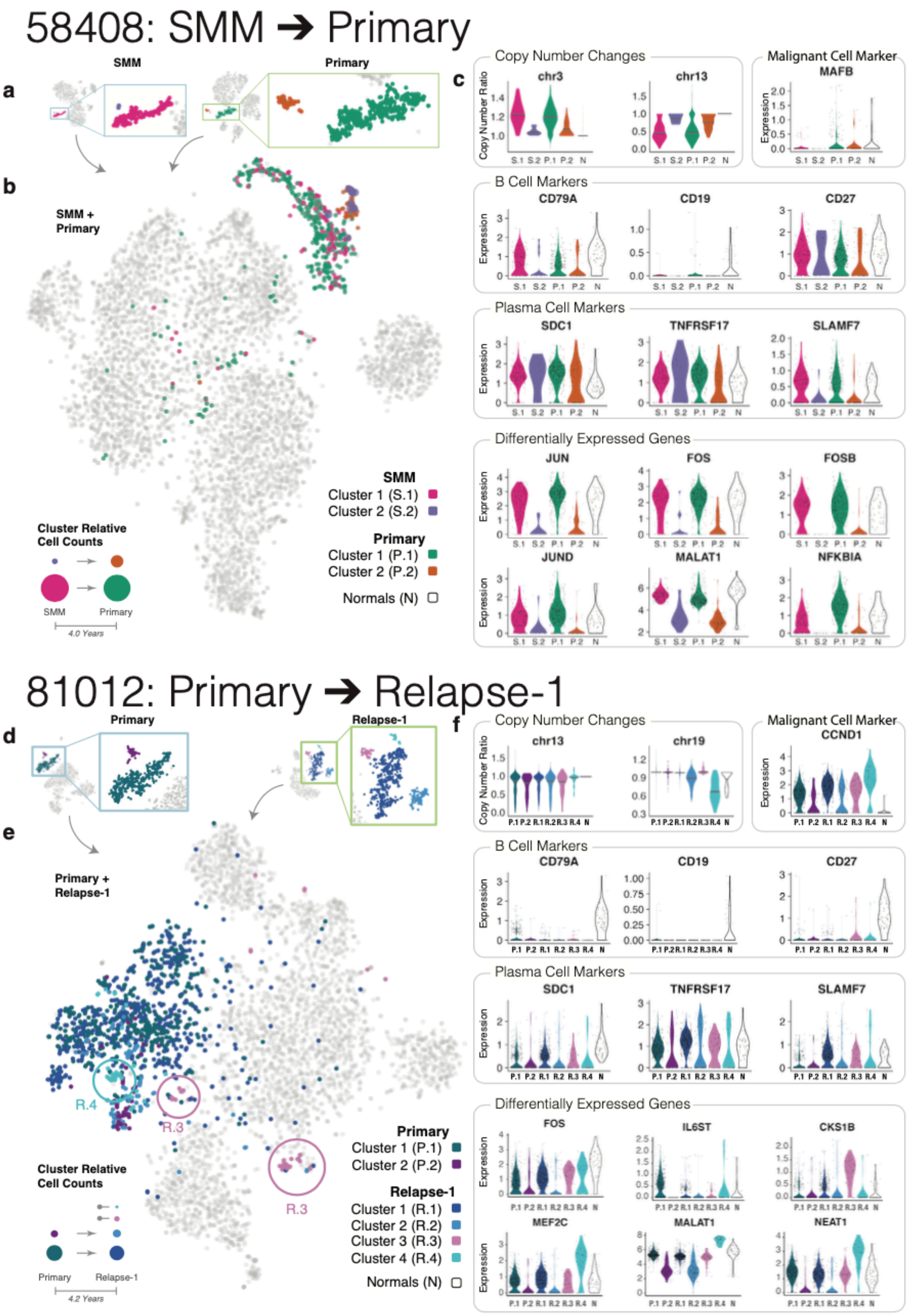


Figure 4. Patterns of plasma cell subpopulation shift from SMM to Primary (58408) and from Primary to Relapse (81012).

Figure 4. Patterns of plasma cell subpopulation shift from SMM to Primary (58408) and from Primary to Relapse (81012).

a. Plasma cell t-SNE subclusters for Patient 58408 at SMM and Primary time points. **b**. Plasma cell subclusters identified in **a** mapped to the integrated t-SNE of all cells from Patient 58048 SMM and Primary time points. Bottom left: possible explanation for plasma cell subpopulation shift from SMM to Primary. **c**. Copy number and expression patterns for plasma cells from different time point subclusters and plasma cells from healthy donors. The first row shows copy number changes and expression of genes associated with genetic alterations detected in Patient 58408. The second and third rows show the expression of B cell markers and plasma cell markers. The last two rows show differentially expressed genes found between the clusters. **d**-**f**. Similar illustrations as **a**-**c** except for Patient 81012, who progressed from Primary to Relapse-1.

To investigate whether primary plasma cell subpopulations descended from subpopulations present at SMM, we integrated data from the two disease stages and examined how the respective cells cluster. In Patient 58408, we found a good mixture for clusters 1 and 2 from the two stages, which occurred 4.0 years apart (Fig. 4b). We then compared genetic alterations and the expression profiles of these clusters (Fig. 4c), finding clear chromosome 13 loss in cluster 1 of both the SMM and primary stage, while cluster 2 of both stages exhibited normal copy number. Gains on chromosomes 5 and 15 show a similar concordance (Fig. 4c, Extended Data Fig. 6c). This evidence collectively suggests that Primary subpopulation 1 probably descended from SMM subpopulation 1, and likewise for subpopulation 2 at the two time points.

We repeated this analysis in the other two patients (47491 and 37692) (Extended Data Fig. 6d-h) and found the same pattern. In Patient 47491, cluster 2 from SMM matches cluster 1 from primary, and the remaining two clusters are associated with each other. This is illustrated by the slight gain of chromosomes 5 and 15, as well as clusters overlapping in the integrated t-SNE plot (Extended Data Fig. 6d-e). For Patient 37692, we also found cluster 1 from SMM and cluster 2 from primary overlapping, while the other two clusters overlapped (Extended Data Fig. 6f).

In Patient 37692, we did not find compelling evidence at the CNV level, possibly due to limited coverage resulting from a low number of plasma cells recovered at the SMM stage. A notable difference regarding Patients 37692 and 58408 is that the dominant subpopulation (the subcluster with more cells) for 58408 at SMM stage remains dominant at primary stage, while the minor subpopulation for 47491 and 37692 at SMM becomes dominant at the primary stage, suggesting differences in the survival/ proliferation of distinct plasma subpopulations. Nevertheless, plasma cell population structures are maintained from SMM to primary diagnosis, suggesting a stable population evolution pattern during this transition.

To further understand subpopulation expression profiles, we investigated expression patterns for Patient 58408. We found slightly higher expression of canonical B cell markers *CD79A* and *CD19* in cluster 1 for both time points (Fig. 4c), while expression of plasma cell markers is similar (Fig. 4c), suggesting plasma cell subpopulation 1 represents a more ancestral "B cell-like" phenotype. Given the presence of chromosome 13 deletion in this cluster, it is possible that malignant transformation of this clone occurs at the B cell rather than plasma cell stage though this could also arise through a reprogramming process. We also conducted an unbiased differential expression analysis and found high expression of *JUN*, *FOS*, *FOSB*, and *JUND* in cluster 1 (Fig. 4c). Notably, differential expression for *FOS* and *JUN* is also found within clusters for the other two patients (Extended Data Fig 6g-h). *JUN* and *FOS* encode proteins JUN and FOS which dimerize to assemble the AP-1 transcription factor. AP-1 has been implicated in a variety of biological processes, including cell proliferation, differentiation, and apoptosis [192].

We found chromosome 13 deletion in cluster 1 in Patient 58408, suggesting a more malignant phenotype. However, we also detected high levels of *JUN* and *FOS* in normal plasma cells, similar to what we found in this cluster. Based on these observations, it is difficult to determine whether high AP-1 activity could be an indicator of malignancy, especially given that the oncogenic role of the AP-1 pathway is very context-dependent [192].

Dynamic gain and loss of plasma cell subpopulations observed from primary to relapse

We followed plasma cell populations from the primary diagnosis to relapse and noticed the emergence of distinct plasma cell subpopulations. In each of six patients with primary and relapse time points (27522, 56203, 57075, 59114, 60359, and 81012), we observed two or more t-SNE subclusters of plasma cells, which arose in the context of treatment-related selective pressure (Fig. 1a). Plasma cell subclusters tended to be more similar to (i.e. clustered more closely to) each other than other cell types. The proportion of plasma cells present at the primary and relapse stages varied across patients, with some tumors exhibiting a higher proportion at the primary stage and vice-versa; this could reflect sampling variability, patient-to-patient differences in disease progression and treatment efficacies, and/or the snapshot nature of data collection (Fig. 1a, Fig. 2a). Next, using single cell gene expression and copy number changes, we determined the relationship between plasma cell subpopulations at primary and relapse stages. Within a particular patient, subclusters with similar expression and copy number patterns at different time points likely represent the same subpopulation of cells observed over the course of tumor progression. Three patients (81012, 56203, and 27522) illustrate this dynamic population shift in detail.

Patient 81012 displayed variable plasma cell subpopulation dynamics over the course of progressive disease (Fig. 4d-f). At the primary stage, we observed two plasma cell subpopulations (named P.1 and P.2). Later, at relapse, we observed four plasma cell subpopulations (R.1-R.4). In this case, two new plasma cell subpopulations emerged at relapse which had not been observed at the primary stage. Integrated t-SNE mapping showed that the overall expression profiles of P.1 and R.1 match, that P.2 and R.2 match, and that R.3 and R.4 are distinct new clusters (Fig. 4e). Looking more closely at expression markers, P.1 and R.1 showed elevated expression levels of B cell marker *CD79A*. P.1, R.1, and R.3 had similar levels of plasma cell markers (*SDC1*, *TNFRSF17*, and *SLAMF7*). For FOS, one component within the AP-1 complex, we found the lowest expression in P.2 and R.2; P.1, R.1 and R.2 exhibit higher expression, while R.4 shows highest expression. We then took a closer look at R.3 and R.4, since the two newly-derived populations have similar expression of FOS as R.1. We found R.3 exhibits the highest *CKS1B* expression, overexpression of which promotes myeloma cell growth and survival [193] and is associated with a poorer prognosis [194]. *CKS1B* overexpression could be caused by gain at chromosome 1q21 region, but this was not observed in our analysis, suggesting the change is independent of chromosome alteration. R.4 has the highest expression for *MEF2C*, a transcriptional factor typically regarded as playing a role in muscle cell differentiation.[195] Recently, ATAC-seq profiling suggested *MEF2* family is preferentially enriched in the open chromatin regions in myeloma cells and MEF2C inhibition resulted in reduced myeloma cell growth and survival [196]. At the copy number level, R.4 exhibits chromosome 19 loss, a feature absent in all the other subpopulations (Fig. 4f). Together, the evidence suggests that the newly arisen R.3 and R.4 both exhibit enhanced growth and survival, though through different mechanisms of regulation.

Patient 56203 progressed from the primary stage, with three plasma cell subpopulations (P.1, P.2, and P.3), to the relapse stage, with two plasma cell subpopulations (R.1 and R.2) (Extended Data Fig. 7a). P.1, R.1, and R.2 showed similar levels of chromosome 13 loss, while R.1 and R.2 demonstrated chromosome 17 loss, which distinguished the relapse clusters from the primary clusters. Following drug therapy and ASCT, primary cluster P.1 showed similarity to the two subpopulations present at relapse, while primary clusters P.2 and P.3 appear to have been lost (Extended Data Fig. 7a).

However, tumor subpopulation relationships during disease progression can be more complex than Patients 81012 and 56203 illustrated, as seen in the four time points of Patient 27522 (Fig. 5). The primary time point plasma cells comprise 4 distinct subpopulations (P.1-P.4) (Fig. 5a). Subpopulations P.1, P.2, and P.3 each show partial loss of chromosome 13, while P4 does not (Fig. 5d). Projection of P.1-P.4 from Patient 27522 onto the integrated cross-sample B cell and plasma cell t-SNE map shows two groupings of P.4, both of which map distantly from P.1-P.3, largely confirming the original sample-level clustering as well as indicating a high level of population complexity (Fig. 3a, Fig 5b).

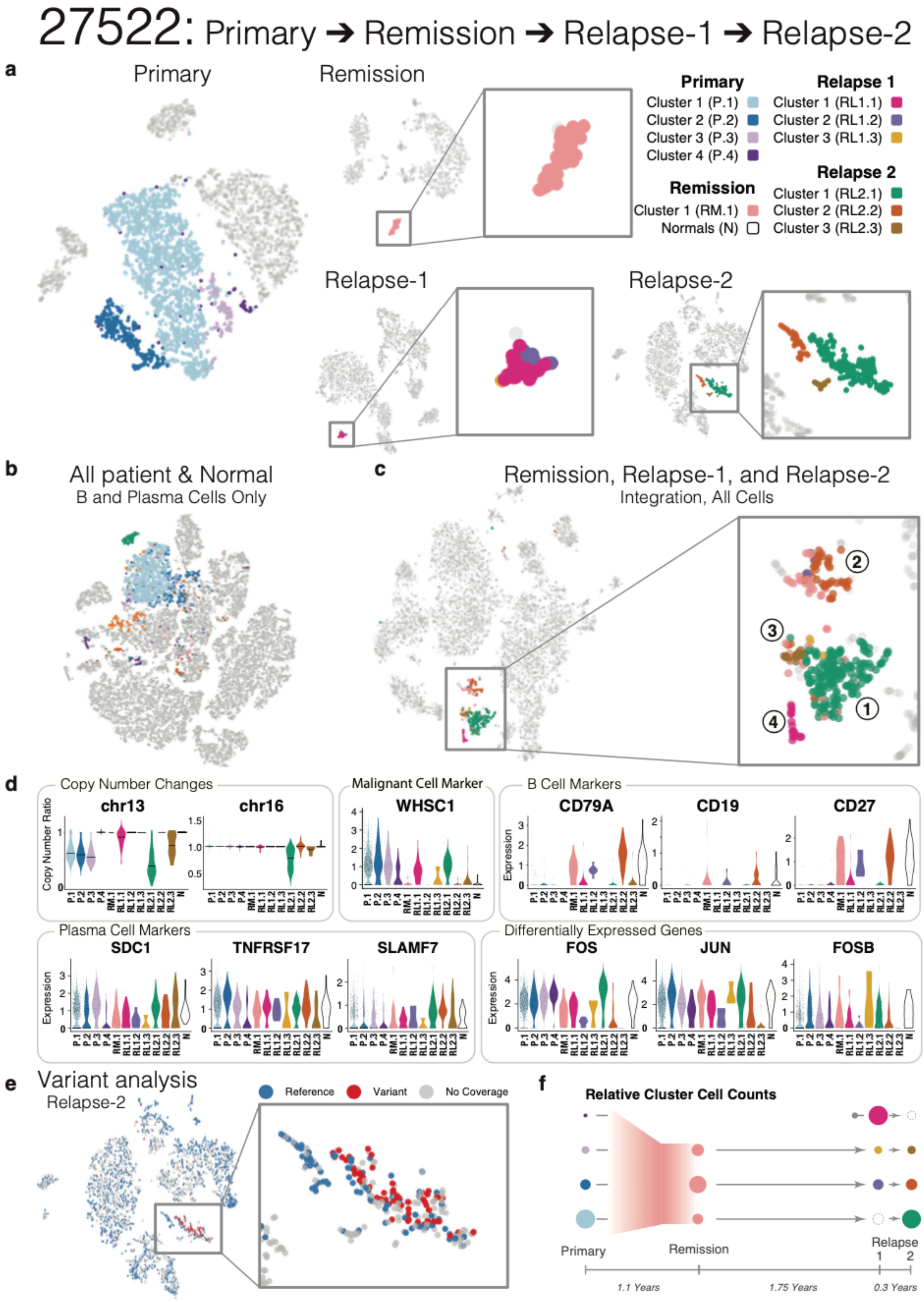


Figure 5. Detailed analysis of plasma cell subpopulation shift for Patient 27522.

Figure 5. Detailed analysis of plasma cell subpopulation shift for Patient 27522.

a. t-SNE mapping of plasma cell subclusters for Patient 27522 at Primary, Remission, Relapse-1, and Relapse-2 disease stages. Colors indicate different subclusters within each time point. **b**. Plasma cell subclusters identified in **a** mapped to the integrated t-SNE of B and plasma cells from all samples plus healthy donors (as in Figure 3a). **c**. Plasma cell subclusters identified in **a** mapped to the integrated t-SNE of all cells from Patient 27522 Remission, Relapse-1, and Relapse-2 disease stages. **d**. Subcluster level copy number changes and expression of malignant cell markers, B cell markers, plasma cell markers, and differentially expressed genes. **e**. Somatic mutations mapped onto Relapse-2 t-SNE (blue, reference allele only; red, variant allele detected; grey, no coverage). **f**. Possible explanation for plasma cell subpopulation shift from Primary to Relapse-2.

We then looked at subpopulations from Remission (RM), Relapse-1 (RL1), and Relapse-2 (RL2) separately from Primary. At Relapse-2, we observed three subpopulations of plasma cells (RL2.1, RL2.2, and RL2.3), with chromosome 13 and chromosome 16 loss in RL2.1, partial loss of chromosome 13 in RL2.3, and t(4;14) translocation in both RL2.1 and RL2.3. RL2.2 remained copy number neutral at chromosome 13 and chromosome 16 (Fig. 5d). Further, we looked for somatic mutations detected from WES data in our scRNA-seq-seq data and noted the occurrence of reference (blue dots) and mutant (red dots) alleles in cells with read coverage. Mutant alleles were detected exclusively in RL2.1, but never in RL2.2 or RL2.3 (Figure 5E). Somatic events observed in these cells included *NRAS* G13R mutation and t(4;14) translocation (inferred from *FGFR3* and *WHSC1* upregulation) (Fig. 5d, Extended Data Fig. 7b). All three clusters expressed high levels of standard plasma cell markers, such as *SDC1*, *SLAMF7 (CS1)*, and *TNFRSF17 (BCMA)*, while *FGFR3* and *WHSC1* were primarily expressed in the malignant (RL2.1) and the "transitional" malignant (RL2.3) populations. *CD27*, a marker associated with normal plasma cells [197], *CD79A*, a member of the B cell antigen receptor complex, and *CD19*, a marker for B cell development, were exclusively detected in RL2.2, supporting the normal "B cell-like" classification (Fig. 5d, Extended Data Fig. 7c). RL2.2 is composed of cells with either high expression of IgA or IgG, while the patient exhibited IgA in isotype identification, which suggests some plasma cells from this subpopulation are normal. These data represent the first confirmed observation that combining mutation and CNV/SV mapping and single cell expression data enables precise identification of cell subpopulations in multiple myeloma that are either rare or undergoing transitional states, with important clinical implications.

In summary, Relapse-2 comprises three distinct subpopulations, one malignant (RL2.1, with somatic mutations and deep chromosome 13 deletion), one “B cell-like” (RL2.2, with strong B cell marker expression), and one “transitional” (RL2.3, without somatic mutations detected but with shallow chromosome 13 deletion). We then traced the origin of these three subpopulations by integrating Relapse-2 with the Remission and Relapse-1 time points.

Based on an integration of Remission (RM), Relapse-1 (RL1), and Relapse-2 (RL2), we found four groups of cells, which are colored by their time point-specific clusters (Group 1: mostly RL2.1; 2: mostly RL2.2; 3: mostly RL2.3; 4: exclusively RL1.1) (Fig. 5c). Some cells from both Remission and Relapse-1 mapped with RL2.2 (Group 2); it is likely that part of these cells are non-malignant plasma cells based on the expression of IgA and IgG. Likewise, other groups of cells from Remission and Relapse-1 mapped with RL2.3 (Group 3). There was one major subpopulation of cells from RL1 that mapped on its own without any clear connection to the previous or later time points (Group 4). Finally, cells present at Remission mapped with the malignant subpopulation RL2.1 (Group 1). This subpopulation was not seen at Relapse-1, potentially due to low cell count or sampling variability. According to B cell marker expression (*CD79A*, *CD19*, *CD27*), the cell population at Remission shows a “B cell-like” pattern, but the co-clustering of Remission cells to multiple relapse populations indicates there is still some malignancy lurking at Remission. Taken together, one interpretation is there were cells present at remission that evaded treatment and survived to seed the relapse. Expression and copy number changes seen in Relapse-1 split according to their grouping with Relapse-2 and Remission on the integrated map, justifying the use of three clusters for downstream analysis although sample-level clustering did not resolve such clusters (Fig. 5f).

Haplotype-based mutation analysis increases resolution of clonal evolution subclustering

We examined how cell type and tumor clonal composition change over time and focus here on Patients 58408 and 27522 to illustrate such evolution. In Patient 58408, the population share of CD4+ and CD8+ T cells dropped from being the two most observed cell types at SMM, with monocytes later emerging as the most prevalent cell type at the primary stage (Figure 2a). Within the plasma cells, we previously described a relatively stable transition of two subpopulations from SMM to Primary (Figures 4a-c), with both hyperdiploidy (HRD) and chromosome 13 deletion detected at the SMM and primary disease stages. Using a mutation VAF-based approach, we observed little genomic change over the 4.0 years separating SMM and Primary (Fig. 6a-b). We detected mutated driver genes (*HIST1H1E*-S172T and *NOTCH1*-D2201V) in the main subclone at both time points.

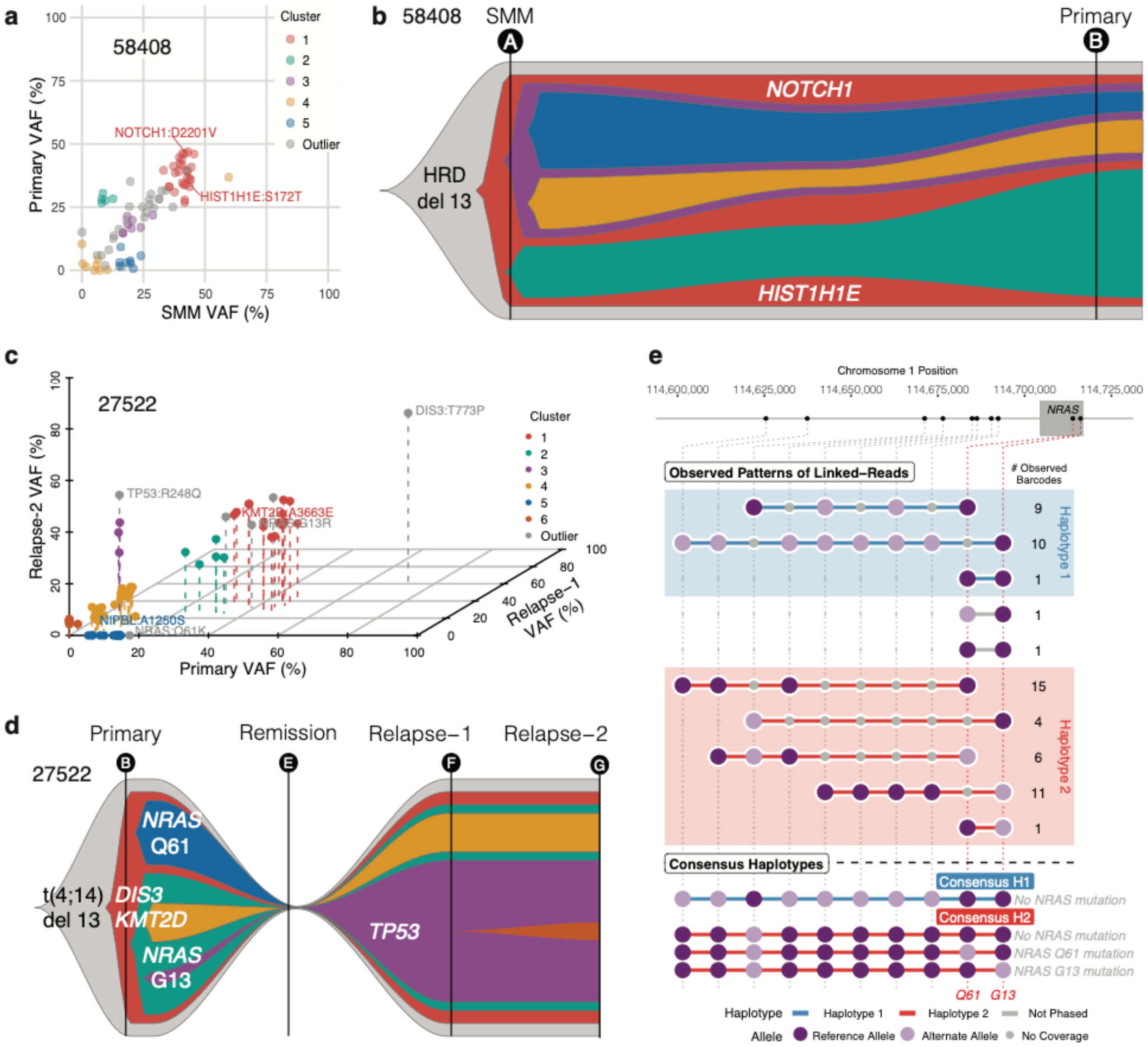


Figure 6. Linked-read DNA sequencing maps somatic mutations to germline haplotypes and clonal evolution maps.

a. Variant allele frequency clustering of subclonal populations from Patient 58408 SMM and Primary samples. **b**. Somatic mutation VAF-based clonality models for Patient 58408. **c**. Variant allele frequency clustering of subclonal populations from Patient 27522 Primary, Relapse-1, and Relapse-2 samples. **d**. Somatic mutation VAF and haplotype-based clonality model for Patient 27522. **e**. Barcode analysis of two *NRAS* somatic mutations showing both mutations occurred on Haplotype 2 did not co-occur, suggesting an independent subclonal relationship. Each set of linked-reads represents a particular pattern of support for the two somatic *NRAS* mutations. The number of observed barcodes refers to total barcodes demonstrating the same pattern of *NRAS* somatic mutations.

In Patient 27522, we observed NK cells only at the relapse stages (Figure 2a), and the population share of CD4+ T cells expanded following remission. The overall proportion of plasma cells observed declined over time with treatment from being a prominent cell type at the Primary stage to later being only a minor cell type. Patient 27522 had one primary and two relapse samples with t(4;14) and del(13q). *NRAS*- G13, *NRAS*-Q61 and *DIS3*-T773 were secondary mutations in the primary sample and *TP53*-R248 was detected at relapse (Figures 6c-d). The *TP53* subclone showed higher VAF at relapse compared to other subclones, implying relevance to subclonal expansion. As previously shown with plasma cell subpopulation analysis, we detected somatic mutations only in Relapse-2 cluster 1 (RL2.1, green) (Fig. 5e).

The primary sample of Patient 27522 displayed two *NRAS* hotspot mutations at G13 (chr1:114716124, C>G) and Q61 (chr1:114713909, A>T). We noted that the Q61 mutation was nearly lost (VAF) at relapse wanted to know if the Q61 mutation occurred in a secondary subclone of the G13 subclone or if G13 and Q61 occurred independently. We utilized 10x Genomics linked-read whole genome sequencing (10xWGS) [19] to address this question. Compared to previous tumor clonality methods which rely mainly on somatic variant allele fractions [152,198-201], linked-reads have the advantage of placing variants in their haplotype context and providing direct observations of the relationship between proximal somatic mutations at distances not captured by short reads alone. Surrounding germline variation showed that these two mutations occurred on the same haplotype, but they did not co-occur in linked-reads covering both positions (n=4), leading us to interpret that they arose independently in distinct subclones, not sequentially in the same subclone (Figure 6e).

Targeted protein assay confirms differential AP-1 expression populations in plasma cells

To better understand how heterogeneity within a single tumor may be reflected in the functional roles of plasma cell subpopulations, we sought to identify common patterns of pathway enrichment across the subpopulations of multiple tumors (Methods). We first divided the plasma cell fractions into a total of 53 discrete subpopulations based on differential gene expression. We then performed pathway enrichment analysis on the differentially expressed genes of each sub-population. Correlation analysis of enrichment results resolved three groups with highly similar enrichment profiles (Extended Data Fig. 8). Group 1 subpopulations share enrichment for pathways related to translation regulation, including nonsense-mediated mRNA decay, as well as PD-1 signaling. These findings are consistent with previous work showing the relevance of active translation[65] and T cell exhaustion[202] to myeloma pathogenesis. Group 2 shares enrichment of cell cycling and proliferation pathways, and may represent highly proliferative subgroups of their respective tumors. Group 3 is enriched in various metabolic pathways as well as in Toll-Like Receptor signaling cascades. These pathways may signify differential interaction with the immune microenvironment.

In addition to database-driven pathway enrichment, we identified pathways in which differentially expressed genes are known key players. Strikingly, out of 13 cases in which enough plasma cells were detected in each sample to perform subpopulation analysis, we observed 7 cases with tumor subpopulations showing differentially expressed members of the heterodimeric AP-1 transcription factor complex, which we call AP-1-high subpopulations. High expression of AP-1 was not solely associated with a specific chromosome alteration event, but the AP-1-high subpopulation was usually enriched for CNV events. There is also a positive correlation between the expression of single cell and bulk RNA-seq expression for *FOS* (r=0.43) and *JUN* (r=0.56) across samples (Extended Data Fig. 9a). We then evaluated the expression of *FOS* and *JUN* across subclusters and across samples, finding at least one plasma cell subpopulation with high expression of *FOS* or *JUN* in all cases, regardless of AP-1 expression differences (Fig. 7a). Interestingly, plasma cells from the multiple sample collections of Patients 58408 and 81012 showed subpopulations exhibiting differential expression of both *FOS* and *JUN,* and we manually defined plasma cell subclusters for each sample based on their t-SNE mapping location. The preservation of the AP-1-high population across samples suggests this population potentially plays a role in the pathogenesis of myeloma.

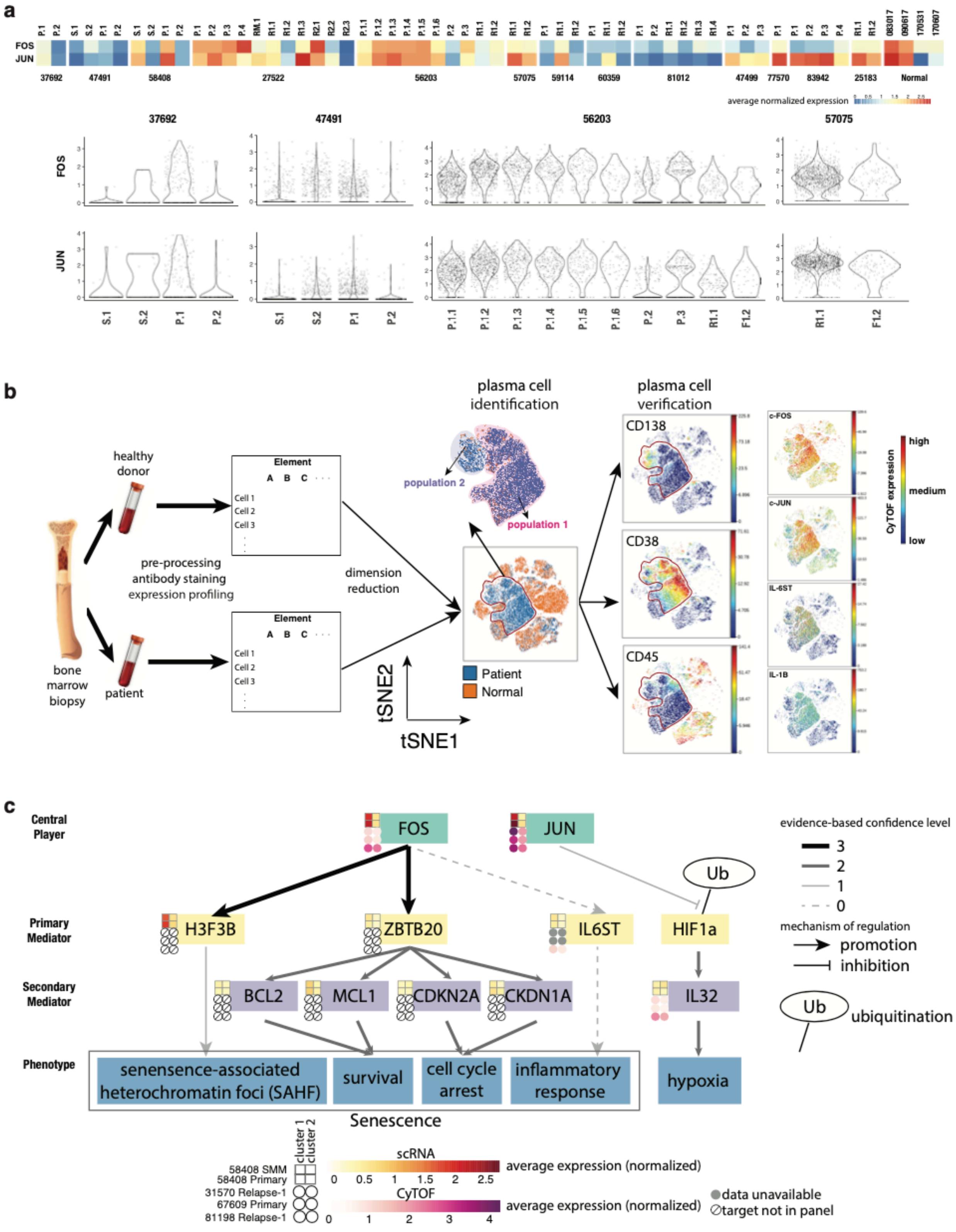


Figure 7. AP-1 expression population in plasma cells confirmed by independent cohort.

Figure 7. AP-1 expression population in plasma cells confirmed by independent cohort.

a. AP-1 components expression across plasma cell subpopulations across samples. Upper: average expression for *FOS* and *JUN* for each sub-population. Lower: violin plot showing the expression patterns of *FOS* and *JUN* for some cases of interest. S, SMM; P, Primary; RM, Remission; R1, Relapse-1; R2, Relapse-2. **b**. CyTOF experiment workflow and data analysis. Bone marrow samples from patient and healthy donors are preprocessed, stained for target antibodies of interest, and expression is profiled in parallel. Samples from patients and healthy donors are merged together and visualized with t-SNE. Regions where only patient samples occupy are further checked for CD138, CD38 and CD45 for verification of their plasma cell identity. Expression profile for FOS, JUN, IL-1B and IL-6 within plasma cells are shown. **c**. Proposed mechanism of how AP-1 complex influences the phenotype of myeloma cells. Heatmap beside each gene indicates normalized expression for different populations of plasma cells in Patients 58408 (SMM and Primary), 31570, 67609, 81198. scRNA-seq expression data, yellow scale; CyTOF expression data, purple scale. Solid arrows, presence of evidence from literature or database. Dashed arrows indicate indirect evidence. Color of solid arrows indicates the confidence level of the evidence of origin. 3, evidence from ChIP-seq database; 2, evidence from myeloma associated literature; 1, evidence from non-myeloma associated literature. Clusters 1 and 2 for each case are manually defined.

Given the frequently observed AP-1 differences within plasma cell populations, we further investigated whether and how differences in the AP-1 pathway could lead to biological differences in plasma cell subpopulations. We performed CyTOF experiments with four additional MM patient samples, three of which had good cell viability. We designed two target panels to separate relevant cell types, quantify signaling pathways (e.g. JAK-STAT, NK-kB), and investigate interleukin activity [203,204]. As expected, we found distinct clusters with differential expression of JUN and FOS (Extended Data Fig. 9b). In fact, a closer look at sample 81198 indicates the two populations with differential AP-1 expression are evident after t-SNE dimension reduction using only cell surface markers (Fig. 7b), consistent with the other two samples (data not shown).

We then combined results from scRNA-seq and CyTOF experiments for a deeper analysis of AP-1 targets (Fig. 7c). We noticed the expression of *H3F3B* and *ZBTB20*, known downstream targets of FOS, are concordant with AP-1 expression within plasma cell populations. *H3F3B* encodes H3.3, a variant of histone H3. Ectopic overexpression of H3.3 is sufficient to induce senescence-associated heterochromatin foci (SAHF), an important marker for cellular senescence [205]. ZBTB20 reportedly plays a role in B cell terminal differentiation; its expression in plasma cell lines induces cell survival and blocks cell cycle progression [206]. Consistent with this, we found slightly upregulated expression of MCL1, a survival marker, and CDKN1A, a cell cycle inhibitor, in the AP-1-high population. Enhanced survival, decreased cell proliferation, as well as the presence of SAHF, all suggests a senescent phenotype for the AP-1 upregulated population. We also found higher expression of IL6ST in the AP-1-high population. IL6ST is a signal transducer shared by IL-6 family cytokine members and is implicated in the progression of a various cancer types [207,208]. IL-6, one of the ligands for IL6ST, and IL1B were upregulated in Patients 81198 and 31570. Given that both samples have undergone prior treatment, it is possible that different populations of plasma cells respond to treatment differently by producing differential amounts of cytokines, especially those involved in senescent-associated-secretory profile (SASP).

It should be noted that, while FOS and JUN are co-dysregulated for a specific cluster in most cases, there are situations where only one of the molecules is dysregulated while the other one is much less obvious. For example, in sample 83942, where no AP-1 differences among clusters are observed, we found all the clusters exhibit low expression of FOS while JUN expression is high. A more interesting case is for sample 81198, where the AP-1-high population exhibits higher upregulation of JUN compared to FOS. In this sample, the AP-1-high population exhibits downregulated CD138 expression and upregulated IL32 expression compared to AP-1-low population. Hypoxia could downregulate CD138 expression in myeloma cells[209] and induce

IL-32 in myeloma cells[210], suggesting AP-1 high population has a more obvious hypoxic signature. Meanwhile, JUN has been shown to stabilize *HIF1A* in a transcriptionally independent manner[211]. It is likely that JUN stabilizes *HIF1A*, promoting the expression of a series of downstream targets, including IL-32, a phenomenon not expected for a cluster where only FOS is high. In summary, different components within AP-1 complex could play different roles in shaping the downstream effector, contributing to diverse phenotypes of plasma cells.

Discussion

In this study, we applied a combination of conventional and single-cell technologies to systematically study multiple myeloma in 14 patients with different treatments at multiple stages of disease progression. We performed scRNA sequencing for ~75K single cells, including both malignant and non-malignant cells, to better understand the transcriptome profiles of these tumors and their interactions with the microenvironment. Varying compositions of cell types over the disease course (e.g. fluctuation of numbers of CD4+ T cell numbers in Patient 59114 discussed above) support the view that the tumor microenvironment is fluid and plays an active role in inter-tumor heterogeneity, as well as disease progression. Patients with the same genetic alterations tended to have both plasma cells and immune cells clustered together. For example, in our two patients with t(11;14), we noted distinct T cell clusters in the tumor microenvironment as well as upregulation of lysine methyltransferase genes *KMT2A* and *KMT2C* in CD8+ T cells. After integrating the data from inferred plasma cells and B cells from healthy donors, we were able to catalog a lineage from primitive B cells to mature B cells and ultimately to normal or malignant plasma cells. Many genes related to this lineage were identified, including known genes like *XBP1,* as well as novel genes requiring further characterization. The overall result indicates that single cell transcriptome profiling of B cells and plasma cells could be used to trace the origin of multiple myeloma, and we identified some patients with plasma cells that exhibit a B cell signature.

We investigated how plasma cell population structure evolves from SMM to primary diagnosis to relapse by integrating somatic alterations mapping, cell lineage marker gene expression, and differential gene expression. Although previous studies have characterized the stability of the SMM to primary transition, we traced specific plasma cell subpopulations across disease stages to illustrate this process and extended the analysis to highlight dynamic changes from diagnosis to relapse. Our analysis is the first to delineate the plasma cell subpopulation structure during multiple myeloma disease progression. By integrating scRNA-seq and genomic alternations, we built plasma cell evolution models representing transitions between disease stages and highlighted co-evolution with the tumor microenvironment. In contrast to malignant cells, non-malignant cells clustered by cell type, independent of their tumor of origin and disease stage. However, detailed characterization of individual immune cell types showed some patients with distinct expression profiles, suggesting a potential interplay between the genomic landscape and an altered microenvironment.

We identified distinct subpopulations of plasma cells in most samples and observed three major patterns of subpopulation shift during disease progression: stable, gain, and loss. Stable pattern is seen in all three patients from SMM to primary, while gain and loss of subpopulations are found from primary to relapse. We extend conventional mutation VAF-based tumor evolution inference models by directly observing subclonal relationships using single cell and single molecule mutation mapping. In the future, mutation mapping should provide more useful information as scRNA-seq technology keeps evolving. We believe mutation and CNV mapping carried out in conjunction with gene expression clustering strategies may be generalizable to other cancer types to trace the origins of malignant cells.

Plasma cells from different populations within the same sample usually exhibit differential expression for components within the AP-1 complex, e.g. JUN and FOS. Tracing the co-differentially expressed genes, together with ChIP-seq data analysis, revealed potential downstream targets which contribute to enhanced survival but decreased proliferation of the AP-1-high population. CyTOF experimentation revealed a similar pattern in FOS and JUN expression. The presence of additional differentially expressed genes from the CyTOF panel, such as IL6 and IL1B, potentially suggests a greater inflammatory response happening in the AP-1 high population.

Future study designs will enable us to compare greater numbers of patients within the same treatment regimen to better understand effects of treatments on tumor and immune cells. In addition to single cell transcriptomics, integrating single cell proteomics will bolster our ability to comprehensively investigate disease progression and treatment response in multiple myeloma.

Methods

Patient Cohort

Fourteen patients with multiple myeloma, 10 male and 4 female Caucasians, were included in the analysis. The median age at diagnosis was 63 (range 46-69). Eight patients had IgG isotype, 4 being kappa light chain and 4 being lambda light chain, 2 had IgA kappa isotype, 2 had light chain only disease (1 kappa and 1 lambda), and 2 were non-secretory. Five were International Staging System Stage 1, two were Stage 2, 3 were stage 3, and 4 were unreported.

The median plasma cell burden by flow cytometry in bone marrow at diagnosis was 24% (range 4-63). By standard fluorescence in situ hybridization (FISH), 1 patient had t(4;14), 3 had t(11;14), and 2 showed del(17p). Four additional patients were included for validation. Two patients have IgG isotype, 1 being kappa light chain and 1 being lambda light chain. One has IgA lambda isotype. One patient has light chain disease (lambda).

Processing

Research bone marrow aspirate samples were collected at the time of the diagnostic procedure. Bone marrow mononuclear cells (BMMCs) were isolated using Ficoll-Paque. BMMCs were cryopreserved in a 1:10 mixture of dimethyl sulfoxide and fetal bovine serum. Upon thawing, whole BMMCs were used for scRNA-seq (unless otherwise specified), 10x WGS, and RNA-seq, as described below. Plasma cells were separated from a sub-aliquot by positive selection using CD138-coated magnetic beads in an autoMACs system (Miltenyi Biotec, CA) and used for WGS, IDT exome, and RNA-seq, as descried below. Skin punch biopsies were performed at the time of the diagnostic bone marrow collection to serve as normal controls for WGS. Although many studies use peripheral blood mononuclear cells (PBMCs) as a control, abnormal B cells and circulating tumor cells frequently contaminate the peripheral blood of patients with MM. Therefore, using PBMCs may lead to omission of genetic events potentially important in disease pathogenesis.

Single cell library prep and sequencing

Utilizing the 10x Genomics Chromium Single Cell 3’ v2 or 5' Library Kit and Chromium instrument, approximately 17,500 cells were partitioned into nanoliter droplets to achieve single cell resolution for a maximum of 10,000 individual cells per sample. The resulting cDNA was tagged with a common 16nt cell barcode and 10nt Unique Molecular Identifier during the RT reaction. Full length cDNA from poly-A mRNA transcripts was enzymatically fragmented and size selected to optimize the cDNA amplicon size (approximately 400 bp) for library construction (10x Genomics). The concentration of the 10x single cell library was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for the HiSeq 4000 or NovaSeq 6000 platform (Illumina). 26x98bp (3' v2 libraries) or 2x150bp (5' libraries) sequence data were generated targeting between 25K-50K read pairs/cell, which provided digital gene expression profiles for each individual cell. For all the samples included in this study, only Patient 27522 Relapse-2 was processed with the 5’ Library Kit.

10x WGS

The normal skin samples were processed with a standard Qiagen DNA isolation kit resulting in 10-50Kb DNA fragments. 250K tumor cells were processed with the MagAttract HMW DNA extraction kit (Qiagen) resulting in 100-150Kb DNA fragments. 600-800ng of normal DNA was size selected on the Blue Pippin utilizing the 0.75% Agarose Dye-Free Cassette to attempt to remove low molecular weight DNA fragments. The size selection parameters were set to capture 30,000 - 80,000bps DNA fragments (Sage Science). The resulting size selected DNA from the normal samples and the HMW DNA from the tumor cells were diluted to 1ng/µL prior to the v2 Chromium Genome Library prep (10x Genomics). Approximately 10-15 DNA molecules were encapsulated into nanoliter droplets. DNA molecules within each droplet were tagged with a 16nt 10x barcode and 6nt unique molecular identifier during an isothermal incubation. The resulting barcoded fragments were converted into a sequence ready Illumina library with an average insert size of 500bp. The concentration of each 10x WGS library was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for the HiSeqX/NovaSeq6000 platform (Illumina). 2x150 sequence data were generated targeting 30x (normal) and 60x (tumor) coverage providing linked reads across the length of individual DNA molecules.

Standard WGS

Manual libraries were constructed with 50-2000ng of genomic DNA utilizing the Lotus Library Prep Kit (IDT Technologies) targeting 350bp inserts. Strand specific molecular indexing is a feature associated with this library method. The molecular indexes are fixed sequences that make up the first 8 bases of read 1 and read 2 insert reads. The concentration of each library was accurately determined through qPCR (Kapa Biosystems). 2x150 paired end sequence data generated ~100Gb per normal and ~200Gb per tumor sample which lead to ~30x (normal) and 60x (tumor) haploid coverage.

IDT Exome

A 700ng aliquot of the existing WGS library was used for the exome capture. Five libraries were pooled at an equimolar ratio yielding a ~3.5µg library pool prior to the hybrid capture. The library pools were hybridized with the xGen Exome Research Panel v1.0 reagent (IDT Technologies) that spans a 39 Mb target region (19,396 genes) of the human genome. The concentration of each captured library pool was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for the NovaSeq6000 platform (Illumina). 2x15bp sequence data was generated ~50Gb per library targeting a mean depth of coverage of 500x.

RNA-seq

Total RNA was isolated from ~700K cells utilizing the AllPrep DNA extraction kit (Qiagen). ERCC RNA Spike-In Mix 1 was added to 100-250ng of total RNA as outlined by the manufacturer (Ambion, Life Technologies). The ERCC control mix is a set of external RNA controls that enable performance assessment for gene expression experiments. The cDNA library was prepared with the TruSeq Stranded Total RNA Sample Prep with Ribo-Zero Gold kit (Illumina). The concentration of each cDNA library was determined through qPCR (Kapa Biosystems). 2x150 reads were generated on the HiSeq4000/NovaSeq6000 instrument (Illumina) generating ~83 million read pairs/sample.

Dataset Description

The data corpus is comprised of 14 patients having various combinations of sample types, time-points, data types, and treatment modalities (Figure 1A). Most patients have 10x whole genome sequencing (10xWGS) data for skin normal and pre-treatment state, with several having relapse data, as well. Patients 59114 and 81012 underwent relatively long treatment periods before relapse (Supplementary Table 1). Treatment ranges from none for 7 patients (3 of which have an SMM sample) to multi-cycle regimens of several 2-drug and 3-drug cocktails, for example in Patient 27522. There are 9 patients having at least one time point with both WES and WGS data. Some patients, such as 27522 also have regular whole exome and whole genome shotgun data at several time points. All WES and WGS data are generated with CD138+ sorted population (tumor cells) within bone marrows. Two patients have data from a first and a second relapse (Relapse-1 and Relapse-2), with Patient 59114 having an additional complement of pre-/post-transplant samplings. To ensure samples matched across time points, we compared germline variant allele fractions (VAF) at 24 loci (Extended Data Figure 1b, Supplementary Table 1).

Somatic mutation detection

Somatic variants were called by our SomaticWrapper pipeline, which includes four established bioinformatic tools, namely Strelka, Mutect, VarScan2 (2.3.83), and Pindel (0.2.54) [134,137,212,213]. We retained SNVs and INDELs using the following strategy: keep SNVs called by any 2 callers among Mutect, VarScan, and Strelka and INDELs called by any 2 callers among VarScan, Strelka, and Pindel. For these merged SNVs and INDELs, we applied coverage cut-offs of 14X and 8X for tumor and normal, respectively. We also filtered SNVs and INDELs with a high-pass variant allele fraction (VAF) of 0.05 in tumor and a low-pass VAF of 0.02 in normal. The SomaticWrapper pipeline is freely available from GitHub at https://github.com/ding-lab/somaticwrapper.

Copy Number and Structural Variation Detection

We used BIC-seq2 [124], a read-depth-based CNV calling algorithm to detect somatic copy number variations (CNVs) using standard WGS tumor samples and paired skin 10xWGS data (human genome GRCh38 reference). The procedure involves 1) retrieving all uniquely mapped reads from the tumor and paired skin BAM files, 2) removing biases by normalization (NBICseq-norm_v0.2.4) 3) detecting CNV based on normalized data (NBICseq-seg_v0.7.2) with BIC-seq2 parameters set as --lambda=90 --detail --noscale --control. In WES data, we used CNVkit (v0.9.4) [157] to compare our tumor samples to a background panel of normals. For scRNA-seq data, we used inferCNV (v0.8.2) [125].

Since we analyzed copy number alteration data from multiple different platforms and varying tumor purity levels, we used five ordered categories to describe copy number changes: deletion < loss < neutral < gain < amplification. The CNV category cutoffs (log2 copy number ratio) were -1, -0.25, 0.2, and 0.7, based on BIC-seq2 and CNVkit documentation. For scRNA based copy number, we transformed the inferCNV results to the log2 scale and set cutoffs at -1, -0.4, 0.3, and 0.7.

Somatic structural variants (SVs) were detected by Manta [139] using tumor/normal sample pairs of standard WGS and paired skin 10xWGS. To filter false positive SVs, we removed events with somatic score < 30 and junction somatic score < 30. We used bulk RNA and single cell RNA data to confirm if translocation events showed overexpression compared with non-translocation samples. We collected translocation and gene expression results relevant to MM based on literature (Supplementary Table 7).

Analysis of 10x Genomics whole genome sequencing data

The proprietary Long Ranger system (v2.2.2) from 10x Genomics was used for preliminary analysis, including demultiplexing cDNA libraries into FASTQ files and aligning reads to the human genome reference GRCh38 (GRCh38-2.1.0). To call variants using Long Ranger, we used --vcmode with GATK (version 3.7.0-gcfedb67) [32]. Long Ranger phasing quality metrics were extracted from the summary output file associated with each sample. For haplotype analysis of somatic variants, we relied on phase information of germline variation from surrounding loci on the same set of linked-reads.

Ancestry analysis

We used a reference panel of genotypes and clustering based on principal components to identify the likely ancestry of our 14 multiple myeloma individuals, with an additional 856 Multiple Myeloma Research Foundation (MMRF) cases (including 31 multiple time point cases). We randomly selected 10,000 coding SNPs from minor allele frequency > 0.02 from the 1000 Genomes Project [143]. From that set of loci, we measured the depth and allele counts of each sample's bam using the tool bam-readcount (version 0.8.0). Genotypes were called using these criteria: 0/0 if reference count ≥ 8 and alternate count < 4; 0/1 if reference count ≥ 4 and alternate count ≥ 4; 1/1 if reference count < 4 and alternate count ≥ 8; and ./. (missing) otherwise. After filtering markers with vacancies > 5% in our multiple myeloma samples, 6,349 markers were left for analysis. We performed principal component analysis (PCA) on the 1000 Genomes samples to identify the top 20 principal components. We then projected our multiple myeloma samples onto the 20-dimensional space representing the 1000 Genomes data. To predict the likely ancestry of our multiple myeloma samples, we built a random forest classifier using these 20 principal components, which has known ancestry information for each sample. Using an 80%/20% split between training and test data, our classifier had 99.6% test accuracy. We then predicted the likely ancestry of our multiple myeloma samples based on this classifier.

Analysis of bulk RNA-seq data

Gene expression was estimated using Kallisto (v0.43.1) [129] and gene fusions were detected using STAR-Fusion (v1.4.0) [39]. We used GRCh38_v27_CTAT_lib_Feb092018 from the STAR-fusion website as the human reference and corresponding GENCODE annotation sets.

Analysis of scRNA-seq data

For single cell RNA-seq analysis, the proprietary software tool Cell Ranger (v2.1.1) from 10x Genomics was used for de-multiplexing sequence data into FASTQ files, aligning reads to the human genome (GRCh38), and generating gene-by-cell UMI count matrix. The R package Seurat (v2.0) was used for all subsequent analysis [214]. First, a series of quality filters were applied to the data to remove those barcodes which fell into any one of these categories: too few genes expressed (possible debris), too many UMIs associated (possible more than one cell), and too high mitochondrial gene expression (possible dead cell). The cut-offs for these filters were as recommended by the Seurat package. Next, the data were normalized and scaled and dimensional reduction was performed using PCA. The cells were then clustered using graph-based clustering (default of Seurat) approach. Cell types were assigned to each cluster by manually reviewing the expression of marker genes. The marker genes used were *CD79A*, *CD79B*, *MS4A1* (B cells); *CD8A*, *CD8B*, *CD7*, *CD3E* (CD8+ T cells); *CD4*, *IL7R*, *CD7*, *CD3E* (CD4+ T cells); *NKG7*, *GNLY* (NK cells); *MZB1*, *SDC1*, *IGHG1* (Plasma cells); *FCGR3A* (Macrophages); *CD14*, *LYZ* (Monocytes); *FCER1A*, *CLEC10A* (Dendritic cells); and *AHSP1*, *HBA*, *HBB* (Erythrocytes). All cells that were labeled as erythrocytes were removed from subsequent analysis.

scRNA-seq data integration

Different scRNA gene expression matrices were integrated using the Seurat R package. We controlled for batch effects using the CCA method and the data were integrated using the top 1000 variable genes from each sample and the first 15 CCs. Cell types were assigned based on manual review of marker gene expression (as described above). Cells with inconsistent cell type assignments between the integrated and individual analyses were filtered out. In some cases, the inconsistencies arose from evident clustering issues (for example, when reviewing marker gene expression, two sub-clusters were obvious within one cluster). Such instances were manually resolved and the cells were rescued. All differential gene expression analyses were carried out using the FindMarkers function of the Seurat package. The default Wilcox test was used and hits with adjusted p-value < 0.05 were deemed significant.

scRNA-seq correlation analysis

After integration, for each cell type, we compared the gene expression to other types to identify the significant highly expressed genes (adjusted p-value < 0.05 and log fold change > 0). Then their average expressions in each sample were calculated. Their pairwise correlations were then estimated.

Clustering of sub-populations of plasma cells based on pathway enrichment

We used differentially expressed genes (DEGs, fold change >1.5 and FDR < 0.1) to detect clusters in plasma cells for each sample. We then used the DEGs for each sub-cluster in samples to do pathway enrichment analysis. For the integration pathway analysis, we used the q-value (FDR) associated with each pathway and only used pathways that had at least one significant (FDR < 0.05) association with a cluster in order to filter non-significant pathways. We then calculated the correlation between sub-clusters from different samples based on the 764 pathway FDR values, to see which sub-clusters shared similar enrichment in pathways.

10Xmapping

scRNA data provide an unprecedented resource for studying tumor heterogeneity and clonal evolution. Connecting somatic mutations to individual cells can help to better understand these aspects and have the potential to identify tumor cells which cannot be unveiled purely based on expression data or is difficult to be separated by expression alone. Here, we developed a mapping tool (10Xmapping), which can identify reads supporting the reference allele and variant allele covering the variant site in each individual cell by tracing cell and molecular barcode information in the bam file. The tool is freely available at https://github.com/ding-lab/10Xmapping. For mapping, we used high-confidence somatic mutations from WES data; mutations were combined if data from multiple time points existed.

Single cell RNA CNV Detection and Clustering

To detect large-scale chromosomal copy number variations using single-cell RNA-seq data, inferCNV (version 0.8.2) [125] was used to obtain relative expression intensity of plasma cells in comparison to a set of reference "normal" cells, including B cells, T cells, Erythrocytes, NK cells, etc. Cutoff=0.1 was used for revealing CNV signals. inferCNV took the raw expression matrix generated from Seurat after several filtering steps, as described above. Subsequently, samples were clustered on inferCNV expression data for 30 genes implicated in MM. Cells for each sample underwent a dimensionality reduction using PCA and t-SNE before clustering. Cells were then clustered with the DBSCAN algorithm. Optimal values for epsilon and minimum points were selected via a grid search. Parameters resulting in the highest Silhouette coefficient were ultimately selected.

CyTOF

Thawed bone marrow suspensions were stained with two panels of metal-conjugated antibodies as listed in (Supplementary Table 6). The concentration of the antibodies were either based on the suggestions from manufacturer (Fluidigm) or based on titration experiments. We used two distinct protocols for cell staining. For panel 1, we included a series of signaling molecules specifically, such as the ones from JAK-STAT pathway and NF-kB pathway [215]. Within this panel, we used three conditions by adding either PBS, PVO4 or TNFa to stimulate samples. Final concentrations for PVO4 and TNFa are 125uM and 20ng/mL, respectively. For panel 2, we included a series of interleukins and interleukin receptors. The inclusion of the aforementioned targets are based on their dysregulation in multiple myeloma [216,217]. We included two components within AP-1 complex, JUN and FOS, in panel 2 as well. To stimulate the production of cytokines, we used three conditions by adding either PBS, R848, or TNFa. Final concentrations for R848 and TNFa are 5ug/mL and 20ng/mL, respectively. Protein transporter inhibitors were added to each condition 2 hours after the beginning of stimulation, and co-incubation lasted for another 2 hours. Gating and data analysis were done using WUSTL Cytobank. Live, single cells are selected by gating out cells/debris with outlier cisplatin and DNA intercalator staining. To perform t-SNE analysis, we used the scaled expression of cell surface marker, including CD34, CD123, CD38, CD3, CD4, CD8, CD19, CD138, CD14, CD16, CD11c, CD56.

AP-1 targets were identified using ChIP-seq data (ENCODE accession number ENCSR000EYZ)[218,219]. We included 4 additional myeloma patient samples for expression profiling via CyTOF experiment. For each CyTOF run, a sample from healthy donor would be included. Expression of cell surface markers are used for t-SNE. Cells from patient samples which does not overlap those from healthy donors on t-SNE plot are further checked for their expression of CD138, CD38 and CD45. Accordingly, the qualified cells are termed as plasma cells.

Subclonal analysis

The R package SciClone [152] algorithm was used to define clonal architecture, and tumor phylogeny was illustrated using Fishplot [158].

Chapter 6: Future Directions

Automating high-resolution multi-omic data integration for cancer story-telling

As scientists, we are called to be responsible story tellers. Responsible because what we say should be accurate and in the public interest. Story tellers because our mission requires others benefitting from our findings, and we have failed if we toil in isolation. Many fields require story telling, but scientific story telling is unique because our role in society is predicated upon being trustworthy, data-driven, and unbiased. Our work is meaningless if we do not communicate, and our work is wasted if we squander public trust through disingenuous behavior. We tell stories by interpreting data and contextualizes our conclusions in a way that is meaningful and beneficial to others.

When we approach a problem, we are blinded to the whole of reality. Our instruments and methods of observation restrict our field of view to specific conditions and outputs, and each data type alone tells only part of the story. When we combine data types, we must do so understanding the limitations of each. But where one data type falls short, another may add value, and so we are compelled to integrate different viewpoints to form a more complete picture. Do so responsibly, repeatably, and transparently requires intentional effort from the beginning of study design to the implementation of analysis code. We have seen this applied successfully to study complex tumor dynamics over time and within samples (Fig. 1). Designing tools to purposefully integrate and magnify the impacts of each data type is the computational complement to ongoing technology development that we rely on to do good genome science.

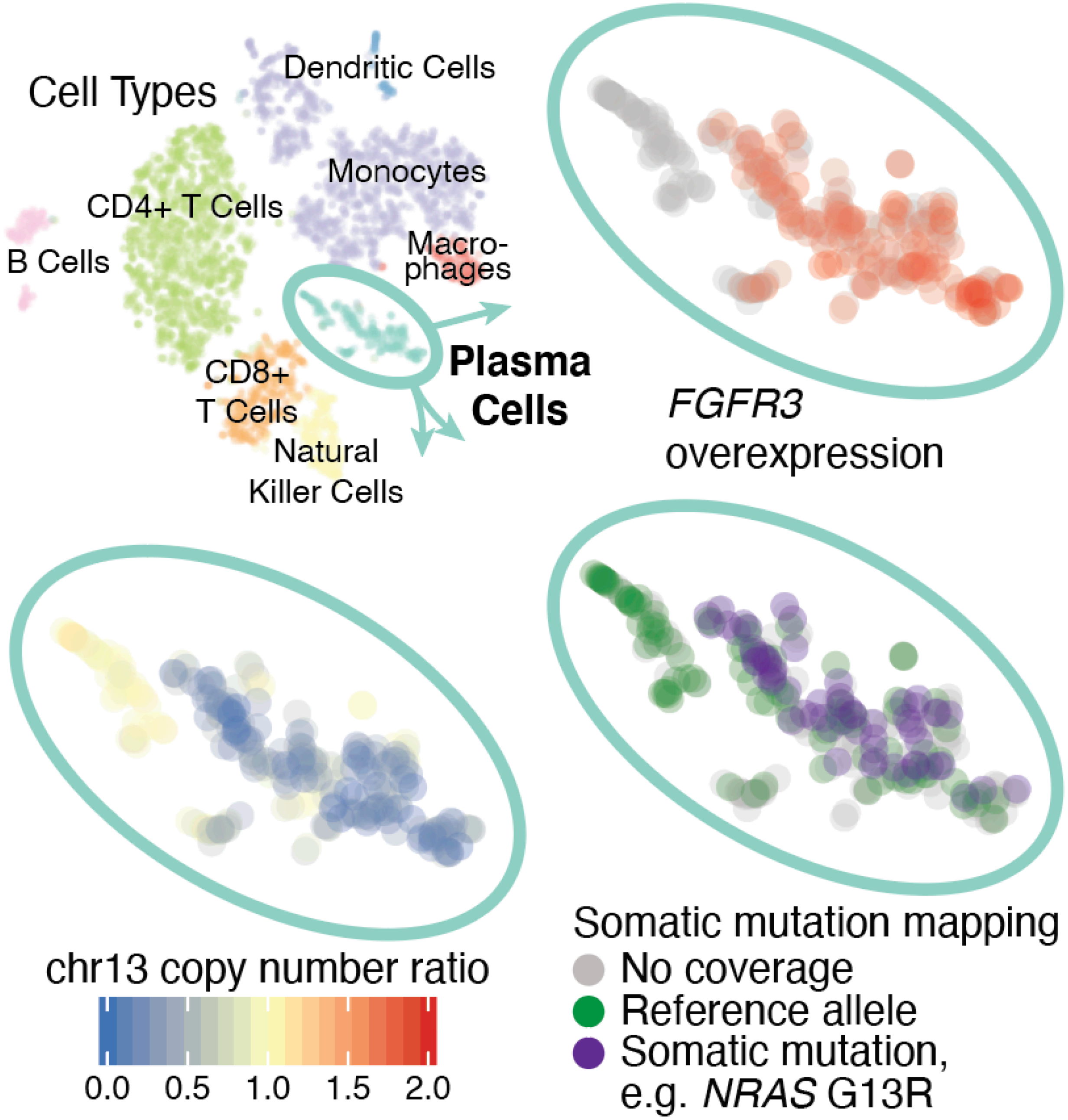


Figure 1. Automating high-resolution multi-omic data integration for cancer story-telling. Integrated single-cell cell type, gene expression, copy number, and somatic mutations from a relapse sample of multiple myeloma patient 27522. This manual curation of data types, carefully scaled and mapped to align each data type, can be automated for deeper understanding. In this case, the relationship of three tumor subclones can be delineated by focusing on various mutation and expression patterns. Overexpression of *FGFR3* is seen in two plasma cell subclusters showing association with t(4;14). The same two subclusters have a later chr13 deletion, followed then by *NRAS* somatic mutations, which appear only in one subcluster.

Cancer disparities research can bridge gaps in health care outcomes

Multiple myeloma (MM) is the second most common hematologic malignancy in the United States, diagnosed in approximately 14,500 Americans each year, and is responsible for 2% of all cancer deaths (SEER.cancer.gov). MM is a malignancy of antibody-secreting plasma B-cells whose etiology is poorly understood and is incurable in the vast majority of patients. MM is always preceded by monoclonal gammopathy of undetermined significance (MGUS) [220-222] a condition that can be detected with a simple blood test. MGUS is an asymptomatic condition for which patients are not routinely screened since there is currently no treatment that has demonstrated efficacy in reducing the risk of progression of MGUS to MM[223]. There has been progress in the treatment of MM, but due to the aging population, the incidence of MM is expected to increase along with the associated costs. Total healthcare costs in the first year after diagnosis of MM is $118,353[224].

There is well-established and long-standing disparity with excess incidence and mortality among African Americans[225]. Incidence of multiple myeloma is approximately two times higher in African-Americans compared to the general population (12 compared to 6 per 100,000) (Figure 3A-C).[226] Our collaborative germline predisposition study with Dr. Lucy Godley from the University of Chicago using >900 spontaneous cases and 57 families has identified *BRCA2*, *ATM*, *CHEK2*, and *KDM1A* as predisposition genes in MM (Figure 3E-G). Previous studies have begun to examine genomic differences present in African-Americans compared to others, includes SNPs and translocations, finding preliminary evidence that translocations associated with lower risk (e.g. t(11;14)) are more prevalent in African-Americans than others.[227,228] Ongoing studies to understand why the incidence rate of multiple myeloma is higher in African-

Americans as well as how to improve health care access and outcomes are important for reducing health care disparities. Equal access to quality care is of major importance to reduce health disparities in all underserved communities.

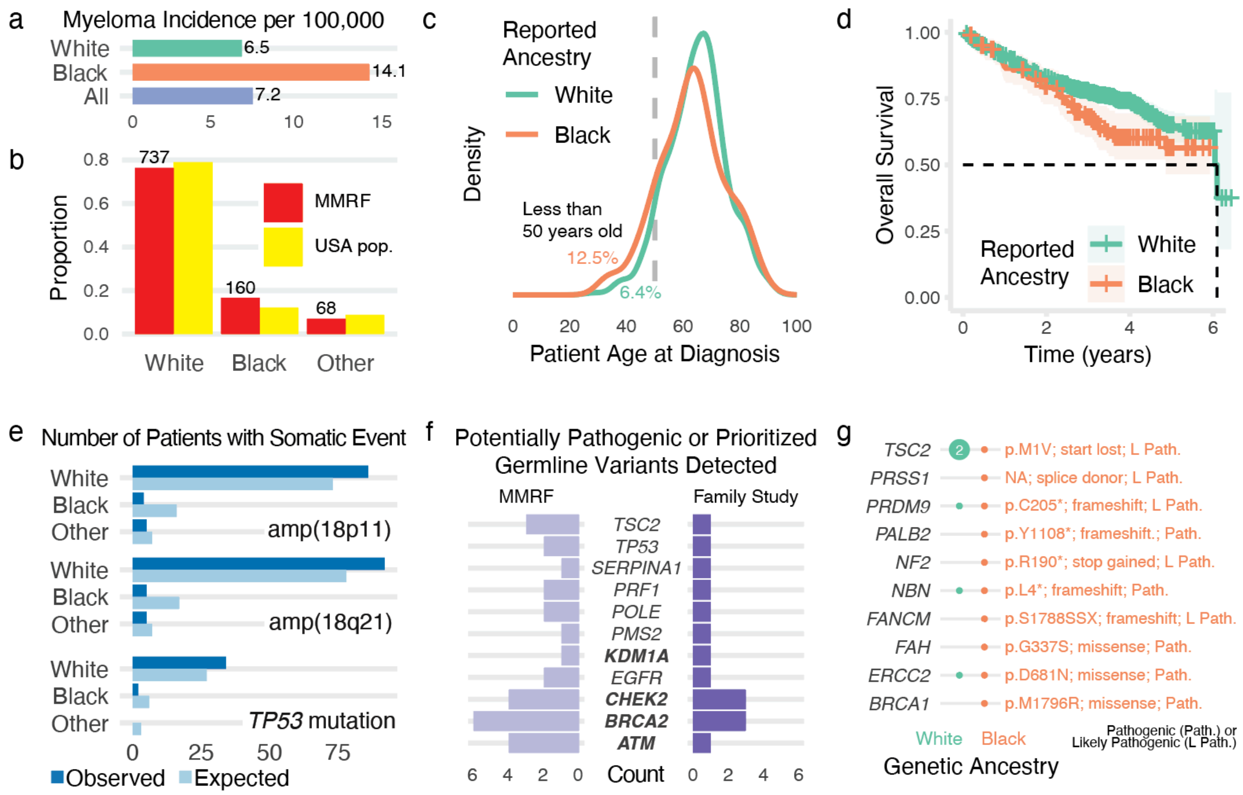


Figure 2. Multiple myeloma incidence and outcome disparity of African American patients. **A.** Incidence of multiple myeloma per 100,000 people in the United States. **B.** Proportional representation of patients in the MMRF cohort compared to the general USA population, stratified by reported ancestry. **C.** Comparison of MMRF patient age at diagnosis, annotated with the proportion of patients under 50 years old. **D.** Kaplan-Meier curves modeling overall survival rates in MMRF. **E.** Somatic events detected in the MMRF cohort with a significantly different number of patients observed to have that event compared to expectation. **F.** Number of pathogenic, likely pathogenic, or prioritized variants of uncertain significance reported from MMRF and family study cohorts. **G.** Pathogenic and likely pathogenic germline variants reported in black patients from MMRF.

References

1 Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D. & Hitchins, M. P. Milestones of Lynch syndrome: 1895-2015. *Nat Rev Cancer* **15**, 181-194, doi:10.1038/nrc3878 (2015).

2 Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med* **22**, 1342-1350, doi:10.1038/nm.4191 (2016).

3 Foltz, S. M., Liang, W. W., Xie, M. & Ding, L. MIRMMR: binary classification of microsatellite instability using methylation and mutations. *Bioinformatics* **33**, 3799-3801, doi:10.1093/bioinformatics/btx507 (2017).

4 Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015-1016, doi:10.1093/bioinformatics/btt755 (2014).

5 Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* **15**, 371-381, doi:10.1038/nrc3947 (2015).

6 Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233-245, doi:10.1038/nrc2091 (2007).

7 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e318, doi:10.1016/j.cell.2018.02.060 (2018).

8 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).

9 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339, doi:10.1038/nature12634 (2013).

10 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

11 Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510, doi:10.1038/nature10738 (2012).

12 Fakhri, B. & Vij, R. Vol. 16 S130-S134 (2016).

13 Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306-313, doi:10.1038/nature10762 (2012).

14 McGranahan, N. & Swanton, C. Vol. 168 613-628 (Elsevier, 2017).

15 Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* **48**, 827-837, doi:10.1038/ng.3586 (2016).

16 Sengupta, S. *et al.* Integrative omics analyses broaden treatment targets in human cancer. *Genome Med* **10**, 60, doi:10.1186/s13073-018-0564-z (2018).

17 Vasaikar, S. V., Straub, P., Wang, J. & Zhang, B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res* **46**, D956-D963, doi:10.1093/nar/gkx1090 (2018).

18 Greer, S. U. *et al.* Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med* **9**, 57, doi:10.1186/s13073-017-0447-8 (2017).

19 Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-311, doi:10.1038/nbt.3432 (2016).

20 Vilar, E. & Gruber, S. B. Microsatellite instability in colorectal cancer-the stable evidence. *Nat Rev Clin Oncol* **7**, 153-162, doi:10.1038/nrclinonc.2009.237 (2010).

21 Kim, T. M., Laird, P. W. & Park, P. J. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155**, 858-868, doi:10.1016/j.cell.2013.10.015 (2013).

22 Salipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H. & Pritchard, C. C. Microsatellite instability detection by next generation sequencing. *Clin Chem* **60**, 1192-1199, doi:10.1373/clinchem.2014.223677 (2014).

23 Huang, M. N. *et al.* MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. *Sci Rep* **5**, 13321, doi:10.1038/srep13321 (2015).

24 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

25 Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).

26 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).

27 Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765, doi:10.1016/j.cell.2016.05.069 (2016).

28 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).

29 Vasaikar, S. *et al.* Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **177**, 1035-1049 e1019, doi:10.1016/j.cell.2019.03.030 (2019).

30 Clark, D. J. *et al.* Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **179**, 964-983 e931, doi:10.1016/j.cell.2019.10.007 (2019).

31 Dou, Y. *et al.* Proteogenomic Characterization of Endometrial Carcinoma. *Cell* **180**, 729-748 e726, doi:10.1016/j.cell.2020.01.026 (2020).

32 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

33 Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res* **44**, 4487-4503, doi:10.1093/nar/gkw282 (2016).

34 Carrara, M. *et al.* State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int* **2013**, 340620, doi:10.1155/2013/340620 (2013).

35 Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep* **23**, 227-238 e223, doi:10.1016/j.celrep.2018.03.050 (2018).

36 Hu, X. *et al.* TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res* **46**, D1144-D1149, doi:10.1093/nar/gkx1018 (2018).

37 Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat Commun* **5**, 4846, doi:10.1038/ncomms5846 (2014).

38 Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845-4854, doi:10.1038/onc.2014.406 (2015).

39 Haas, B. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*, doi:10.1101/120295 (2017).

40 Benelli, M. *et al.* Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* **28**, 3232-3239, doi:10.1093/bioinformatics/bts617 (2012).

41 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

42 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).

43 Kanchi, K. L. *et al.* Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun* **5**, 3156, doi:10.1038/ncomms4156 (2014).

44 Wang, L. *et al.* Novel somatic and germline mutations in intracranial germ cell tumours. *Nature* **511**, 241-245, doi:10.1038/nature13296 (2014).

45 Kumar-Sinha, C., Kalyana-Sundaram, S. & Chinnaiyan, A. M. Landscape of gene fusions in epithelial cancers: seq and ye shall find. *Genome Med* **7**, 129, doi:10.1186/s13073-015-0252-1 (2015).

46 Sinclair, A., Latif, A. L. & Holyoake, T. L. Targeting survival pathways in chronic myeloid leukaemia stem cells. *Br J Pharmacol* **169**, 1693-1707, doi:10.1111/bph.12183 (2013).

47 Hantschel, O. Structure, regulation, signaling, and targeting of abl kinases in cancer. *Genes Cancer* **3**, 436-446, doi:10.1177/1947601912458584 (2012).

48 Ren, R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat Rev Cancer* **5**, 172-183, doi:10.1038/nrc1567 (2005).

49 Cilloni, D. & Saglio, G. Molecular pathways: BCR-ABL. *Clin Cancer Res* **18**, 930-937, doi:10.1158/1078-0432.CCR-10-1613 (2012).

50 Dinh, T. A. *et al.* Comprehensive analysis of The Cancer Genome Atlas reveals a unique gene and non-coding RNA signature of fibrolamellar carcinoma. *Sci Rep* **7**, 44653, doi:10.1038/srep44653 (2017).

51 Singh, D. *et al.* Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science* **337**, 1231-1235, doi:10.1126/science.1220834 (2012).

52 Lasorella, A., Sanson, M. & Iavarone, A. FGFR-TACC gene fusions in human glioma. *Neuro Oncol* **19**, 475-483, doi:10.1093/neuonc/now240 (2017).

53 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322, doi:10.1038/nature12965 (2014).

54 Palanisamy, N. *et al.* Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat Med* **16**, 793-798, doi:10.1038/nm.2166 (2010).

55 Jones, D. T. *et al.* Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res* **68**, 8673-8677, doi:10.1158/0008-5472.CAN-08-2097 (2008).

56 Bass, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-968, doi:10.1038/ng.936 (2011).

57 Lu, H. *et al.* Engineering and Functional Characterization of Fusion Genes Identifies Novel Oncogenic Drivers of Cancer. *Cancer Res* **77**, 3502-3512, doi:10.1158/0008-5472.CAN-16-2745 (2017).

58 Lee, M. *et al.* ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res* **45**, D784-D789, doi:10.1093/nar/gkw1083 (2017).

59 Giacomini, C. P. *et al.* Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. *PLoS Genet* **9**, e1003464, doi:10.1371/journal.pgen.1003464 (2013).

60 Hu, X. *et al.* TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res*, doi:10.1093/nar/gkx1018 (2017).

61 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

62 Babiceanu, M. *et al.* Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res* **44**, 2859-2872, doi:10.1093/nar/gkw032 (2016).

63 Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322, doi:10.1016/j.cell.2007.12.014 (2008).

64 Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol* **3**, REVIEWS0004 (2002).

65 Manier, S. *et al.* Genomic complexity of multiple myeloma and its clinical implications. *Nat Rev Clin Oncol* **14**, 100-113, doi:10.1038/nrclinonc.2016.122 (2017).

66 Haferlach, C. *et al.* AML with CBFB-MYH11 rearrangement demonstrate RAS pathway alterations in 92% of all cases including a high frequency of NF1 deletions. *Leukemia* **24**, 1065-1069, doi:10.1038/leu.2010.22 (2010).

67 Smith, I. *et al.* 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *Lancet* **369**, 29-36, doi:10.1016/S0140-6736(07)60028-2 (2007).

68 Cao, S. *et al.* Divergent viral presentation among human tumors and adjacent normal tissues. *Sci Rep* **6**, 28294, doi:10.1038/srep28294 (2016).

69 Moniz, S. & Jordan, P. Emerging roles for WNK kinases in cancer. *Cell Mol Life Sci* **67**, 1265-1276, doi:10.1007/s00018-010-0261-6 (2010).

70 Chang, J. *et al.* Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nat Commun* **8**, 15290, doi:10.1038/ncomms15290 (2017).

71 von Massenhausen, A. *et al.* Targeting DDR2 in head and neck squamous cell carcinoma with dasatinib. *Int J Cancer* **139**, 2359-2369, doi:10.1002/ijc.30279 (2016).

72 Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281 e277, doi:10.1016/j.cels.2018.03.002 (2018).

73 Cancer Genome Atlas Research, N. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).

74 Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep* **4**, 1116-1130, doi:10.1016/j.celrep.2013.08.022 (2013).

75 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).

76 Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

77 Cancer Genome Atlas Research, N. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378-384, doi:10.1038/nature21386 (2017).

78 Bobisse, S., Foukas, P. G., Coukos, G. & Harari, A. Neoantigen-based cancer immunotherapy. *Ann Transl Med* **4**, 262, doi:10.21037/atm.2016.06.17 (2016).

79 Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511-517, doi:10.1093/bioinformatics/btv639 (2016).

80 Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512-520, doi:10.1093/nar/gku1267 (2015).

81 Gonzalez, D. *et al.* Immunoglobulin gene rearrangements and the pathogenesis of multiple myeloma. *Blood* **110**, 3112-3121, doi:10.1182/blood-2007-02-069625 (2007).

82 Kumar, S. K. & Rajkumar, S. V. The multiple myelomas - current concepts in cytogenetic classification and therapy. *Nat Rev Clin Oncol* **15**, 409-421, doi:10.1038/s41571-018-0018-y (2018).

83 Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472, doi:10.1038/nature09837 (2011).

84 Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell* **25**, 91-101, doi:10.1016/j.ccr.2013.12.015 (2014).

85 Walker, B. A. *et al.* Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. *Leukemia* **28**, 384-390, doi:10.1038/leu.2013.199 (2014).

86 Lagana, A. *et al.* Integrative network analysis identifies novel drivers of pathogenesis and progression in newly diagnosed multiple myeloma. *Leukemia* **32**, 120-130, doi:10.1038/leu.2017.197 (2018).

87 Cleynen, A. *et al.* Expressed fusion gene landscape and its impact in multiple myeloma. *Nat Commun* **8**, 1893, doi:10.1038/s41467-017-00638-w (2017).

88 Nasser, S. *et al.* Comprehensive Identification of Fusion Transcripts in Multiple Myeloma: An Mmrf Commpass Analysis. *Blood* **130**, 61-61 (2017).

89 Lin, M. *et al.* Identification of novel fusion transcripts in multiple myeloma. *J Clin Pathol* **71**, 708-712, doi:10.1136/jclinpath-2017-204961 (2018).

90 Morgan, G. J. *et al.* Kinase domain activation through gene rearrangement in multiple myeloma. *Leukemia* **32**, 2435-2444, doi:10.1038/s41375-018-0108-y (2018).

91 Miller, C. *et al.* A Comparison of Clinical FISH and Sequencing Based FISH Estimates in Multiple Myeloma: An Mmrf Commpass Analysis. *Blood* **128**, 374-374 (2016).

92 Barwick, B. G. *et al.* Multiple myeloma immunoglobulin lambda translocations portend poor prognosis. *Nat Commun* **10**, 1911, doi:10.1038/s41467-019-09555-6 (2019).

93 Mikulasova, A. *et al.* Microhomology-mediated end joining drives complex rearrangements and over expression of MYC and PVT1 in multiple myeloma. *Haematologica*, doi:10.3324/haematol.2019.217927 (2019).

94 Misund, K. *et al.* MYC dysregulation in the progression of multiple myeloma. *Leukemia*, doi:10.1038/s41375-019-0543-4 (2019).

95 Mikhael, J. R. *et al.* Management of newly diagnosed symptomatic multiple myeloma: updated Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART) consensus guidelines 2013. *Mayo Clin Proc* **88**, 360-376, doi:10.1016/j.mayocp.2013.01.019 (2013).

96 Binder, M. *et al.* Prognostic implications of abnormalities of chromosome 13 and the presence of multiple cytogenetic high-risk abnormalities in newly diagnosed multiple myeloma. *Blood Cancer J* **7**, e600, doi:10.1038/bcj.2017.83 (2017).

97 Walker, B. A. *et al.* A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia* **33**, 159-170, doi:10.1038/s41375-018-0196-8 (2019).

98 Bolli, N. *et al.* Analysis of the genomic landscape of multiple myeloma highlights novel prognostic markers and disease subgroups. *Leukemia* **32**, 2604-2616, doi:10.1038/s41375-018-0037-9 (2018).

99 Vij, R. *et al.* Deep sequencing reveals myeloma cells in peripheral blood in majority of multiple myeloma patients. *Clin Lymphoma Myeloma Leuk* **14**, 131-139 e131, doi:10.1016/j.clml.2013.09.013 (2014).

100 Pawlyn, C. & Morgan, G. J. Evolutionary biology of high-risk multiple myeloma. *Nat Rev Cancer* **17**, 543-556, doi:10.1038/nrc.2017.63 (2017).

101 Melchor, L. *et al.* Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. *Leukemia* **28**, 1705-1715, doi:10.1038/leu.2014.13 (2014).

102 Petti, A. A. *et al.* A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun* **10**, 3660, doi:10.1038/s41467-019-11591-1 (2019).

103 Mirabella, F. *et al.* MMSET is the key molecular target in t(4;14) myeloma. *Blood Cancer Journal* **3**, e114-e114, doi:10.1038/bcj.2013.9 (2013).

104 Keats, J. J. *et al.* Overexpression of transcripts originating from the MMSET locus characterizes all t(4;14)(p16;q32)-positive multiple myeloma patients. *Blood* **105**, 4060-4069, doi:10.1182/blood-2004-09-3704 (2005).

105 Walker, B. A. *et al.* Characterization of IGH locus breakpoints in multiple myeloma indicates a subset of translocations appear to occur in pregerminal center B cells. *Blood* **121**, 3413-3419, doi:10.1182/blood-2012-12-471888 (2013).

106 Schaub, F. X. *et al.* Pan-cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas. *Cell Syst* **6**, 282-300 e282, doi:10.1016/j.cels.2018.03.003 (2018).

107 Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).

108 Cui, M. *et al.* Long non-coding RNA PVT1 and cancer. *Biochem Biophys Res Commun* **471**, 10-14, doi:10.1016/j.bbrc.2015.12.101 (2016).

109 Cho, S. W. *et al.* Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell* **173**, 1398-1412 e1322, doi:10.1016/j.cell.2018.03.068 (2018).

110 Ashby, C. *et al.* Chromothripsis and Chromoplexy Are Associated with DNA Instability and Adverse Clinical Outcome in Multiple Myeloma. *Blood* **132**, 408-408, doi:10.1182/blood-2018-99-117359 (2018).

111 Sun, S. Q. *et al.* Database of evidence for precision oncology portal. *Bioinformatics* **34**, 4315-4317, doi:10.1093/bioinformatics/bty531 (2018).

112 Hutchinson, K. E. *et al.* BRAF fusions define a distinct molecular subset of melanomas with potential sensitivity to MEK inhibition. *Clin Cancer Res* **19**, 6696-6702, doi:10.1158/1078-0432.CCR-13-1746 (2013).

113 Cocco, E., Scaltriti, M. & Drilon, A. NTRK fusion-positive cancers and TRK inhibitor therapy. *Nat Rev Clin Oncol* **15**, 731-747, doi:10.1038/s41571-018-0113-0 (2018).

114 Taylor, J. *et al.* Oncogenic TRK fusions are amenable to inhibition in hematologic malignancies. *J Clin Invest* **128**, 3819-3825, doi:10.1172/JCI120787 (2018).

115 Walker, B. A. *et al.* APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat Commun* **6**, 6997, doi:10.1038/ncomms7997 (2015).

116 Maura, F. *et al.* Biological and prognostic impact of APOBEC-induced mutations in the spectrum of plasma cell dyscrasias and multiple myeloma cell lines. *Leukemia* **32**, 1044-1048, doi:10.1038/leu.2017.345 (2018).

117 Bolli, N. *et al.* A DNA target-enrichment approach to detect mutations, copy number changes and immunoglobulin translocations in multiple myeloma. *Blood Cancer J* **6**, e467, doi:10.1038/bcj.2016.72 (2016).

118 Heyer, E. E. *et al.* Diagnosis of fusion genes using targeted RNA sequencing. *Nat Commun* **10**, 1388, doi:10.1038/s41467-019-09374-9 (2019).

119 Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet* **51**, 202-206, doi:10.1038/s41588-018-0312-8 (2019).

120 Yang, W. *et al.* Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat Med* **25**, 767-775, doi:10.1038/s41591-019-0434-2 (2019).

121 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

122 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

123 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

124 Xi, R., Lee, S., Xia, Y., Kim, T. M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res* **44**, 6274-6286, doi:10.1093/nar/gkw491 (2016).

125 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).

126 Nicorici, D. *et al.* FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. Report No. 011650, 011650-011650 (Cold Spring Harbor Laboratory, 2014).

127 Zhang, J. *et al.* INTEGRATE: gene fusion discovery using whole genome and transcriptome data. *Genome Res* **26**, 108-118, doi:10.1101/gr.186114.114 (2016).

128 Torres-Garcia, W. *et al.* PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics* **30**, 2224-2226, doi:10.1093/bioinformatics/btu169 (2014).

129 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

130 Murphy, C. & Elemento, O. AGFusion: annotate and visualize gene fusions. *bioRxiv*, 080903, doi:10.1101/080903 (2016).

131 Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606, doi:10.1038/ng.3557 (2016).

132 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

133 McInnes, L. a. H., John and Saul, Nathaniel and Großberger, Lukas. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861, doi:10.21105/joss.00861 (2018).

134 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

135 Ye, K. *et al.* Systematic discovery of complex insertions and deletions in human cancers. *Nat Med* **22**, 97-104, doi:10.1038/nm.4002 (2016).

136 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594, doi:10.1038/s41592-018-0051-x (2018).

137 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

138 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).

139 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222, doi:10.1093/bioinformatics/btv710 (2016).

140 Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612, doi:10.1038/ncomms3612 (2013).

141 Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet* **16**, 344-358, doi:10.1038/nrg3903 (2015).

142 Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-223, doi:10.1016/j.ajhg.2009.01.005 (2009).

143 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

144 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

145 Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat Rev Genet* **12**, 215-223, doi:10.1038/nrg2950 (2011).

146 Vasan, N. *et al.* Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kalpha inhibitors. *Science* **366**, 714-723, doi:10.1126/science.aaw9032 (2019).

147 van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet* **34**, 666-681, doi:10.1016/j.tig.2018.05.008 (2018).

148 Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29**, 635-645, doi:10.1101/gr.234443.118 (2019).

149 Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174**, 433-447 e419, doi:10.1016/j.cell.2018.05.036 (2018).

150 Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**, 1881-1893, doi:10.1101/gr.180281.114 (2014).

151 Sereewattanawoot, S. *et al.* Identification of potential regulatory mutations using multi-omics analysis and haplotyping of lung adenocarcinoma cell lines. *Sci Rep* **8**, 4926, doi:10.1038/s41598-018-23342-1 (2018).

152 Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol* **10**, e1003665, doi:10.1371/journal.pcbi.1003665 (2014).

153 Greer, S. U. & Ji, H. P. Structural variant analysis for linked-read sequencing data with gemtools. *Bioinformatics* **35**, 4397-4399, doi:10.1093/bioinformatics/btz239 (2019).

154 Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C. & Beerenwinkel, N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun* **10**, 2750, doi:10.1038/s41467-019-10737-5 (2019).

155 Ramazzotti, D., Graudenzi, A., De Sano, L., Antoniotti, M. & Caravagna, G. Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. *BMC Bioinformatics* **20**, 210, doi:10.1186/s12859-019-2795-4 (2019).

156 Bohrson, C. L. *et al.* Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat Genet* **51**, 749-754, doi:10.1038/s41588-019-0366-2 (2019).

157 Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873, doi:10.1371/journal.pcbi.1004873 (2016).

158 Miller, C. A. *et al.* Visualizing tumor evolution with the fishplot package for R. *BMC Genomics* **17**, 880, doi:10.1186/s12864-016-3195-z (2016).

159 Greipp, P. R. *et al.* International staging system for multiple myeloma. *J Clin Oncol* **23**, 3412-3420, doi:10.1200/JCO.2005.04.242 (2005).

160 Richardson, P. *et al.* The treatment of relapsed and refractory multiple myeloma. *Hematology Am Soc Hematol Educ Program*, 317-323, doi:10.1182/asheducation-2007.1.317 (2007).

161 Stewart, A. K. *et al.* Carfilzomib, lenalidomide, and dexamethasone for relapsed multiple myeloma. *N Engl J Med* **372**, 142-152, doi:10.1056/NEJMoa1411321 (2015).

162 Dimopoulos, M. A. *et al.* Carfilzomib or bortezomib in relapsed or refractory multiple myeloma (ENDEAVOR): an interim overall survival analysis of an open-label, randomised, phase 3 trial. *Lancet Oncol* **18**, 1327-1337, doi:10.1016/S1470-2045(17)30578-8 (2017).

163 Lokhorst, H. M. *et al.* Targeting CD38 with Daratumumab Monotherapy in Multiple Myeloma. *N Engl J Med* **373**, 1207-1219, doi:10.1056/NEJMoa1506348 (2015).

164 Durie, B. G. *et al.* Bortezomib with lenalidomide and dexamethasone versus lenalidomide and dexamethasone alone in patients with newly diagnosed myeloma without intent for immediate autologous stem-cell transplant (SWOG S0777): a randomised, open-label, phase 3 trial. *Lancet* **389**, 519-527, doi:10.1016/S0140-6736(16)31594-X (2017).

165 Fakhri, B. & Vij, R. Clonal Evolution in Multiple Myeloma. *Clin Lymphoma Myeloma Leuk* **16 Suppl**, S130-134, doi:10.1016/j.clml.2016.02.025 (2016).

166 Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997, doi:10.1038/ncomms3997 (2014).

167 Walker, B. A. *et al.* Mutational Spectrum, Copy Number Changes, and Outcome: Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma. *J Clin Oncol* **33**, 3911-3920, doi:10.1200/JCO.2014.59.1503 (2015).

168 Keats, J. J. *et al.* Clonal competition with alternating dominance in multiple myeloma. *Blood* **120**, 1067-1076, doi:10.1182/blood-2012-01-405985 (2012).

169 Egan, J. B. *et al.* Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. *Blood* **120**, 1060-1066, doi:10.1182/blood-2012-01-405977 (2012).

170 Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res* **25**, 1499-1507, doi:10.1101/gr.191098.115 (2015).

171 Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94, doi:10.1038/nature09807 (2011).

172 Ledergor, G. *et al.* Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat Med* **24**, 1867-1876, doi:10.1038/s41591-018-0269-2 (2018).

173 Jang, J. S. *et al.* Molecular signatures of multiple myeloma progression through single cell RNA-Seq. *Blood Cancer J* **9**, 2, doi:10.1038/s41408-018-0160-x (2019).

174 Fonseca, R. *et al.* Deletions of chromosome 13 in multiple myeloma identified by interphase FISH usually denote large deletions of the q arm or monosomy. *Leukemia* **15**, 981-986 (2001).

175 Chiecchio, L. *et al.* Deletion of chromosome 13 detected by conventional cytogenetics is a critical prognostic factor in myeloma. *Leukemia* **20**, 1610-1617, doi:10.1038/sj.leu.2404304 (2006).

176 Chesi, M. *et al.* Dysregulation of cyclin D1 by translocation into an IgH gamma switch region in two multiple myeloma cell lines. *Blood* **88**, 674-681 (1996).

177 Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108**, 2020-2028, doi:10.1182/blood-2005-11-013458 (2006).

178 Chang, H. *et al.* The t(4;14) is associated with poor prognosis in myeloma patients undergoing autologous stem cell transplant. *Br J Haematol* **125**, 64-68 (2004).

179 Prideaux, S. M., Conway O'Brien, E. & Chevassut, T. J. The genetic architecture of multiple myeloma. *Adv Hematol* **2014**, 864058, doi:10.1155/2014/864058 (2014).

180 Avet-Loiseau, H. *et al.* Genetic abnormalities and survival in multiple myeloma: the experience of the Intergroupe Francophone du Myelome. *Blood* **109**, 3489-3495, doi:10.1182/blood-2006-08-040410 (2007).

181 Weissbach, S. *et al.* The molecular spectrum and clinical impact of DIS3 mutations in multiple myeloma. *Br J Haematol* **169**, 57-70, doi:10.1111/bjh.13256 (2015).

182 Hughes, C. S. *et al.* Extracellular cathepsin S and intracellular caspase 1 activation are surrogate biomarkers of particulate-induced lysosomal disruption in macrophages. *Part Fibre Toxicol* **13**, 19, doi:10.1186/s12989-016-0129-5 (2016).

183 Shree, T. *et al.* Macrophages and cathepsin proteases blunt chemotherapeutic response in breast cancer. *Genes Dev* **25**, 2465-2479, doi:10.1101/gad.180331.111 (2011).

184 Gocheva, V. *et al.* IL-4 induces cathepsin protease activity in tumor-associated macrophages to promote cancer growth and invasion. *Genes Dev* **24**, 241-255, doi:10.1101/gad.1874010 (2010).

185 Mantchev, G. T., Cortesao, C. S., Rebrovich, M., Cascalho, M. & Bram, R. J. TACI is required for efficient plasma cell differentiation in response to T-independent type 2 antigens. *J Immunol* **179**, 2282-2288, doi:10.4049/jimmunol.179.4.2282 (2007).

186 Lee, S. M., Jeon, S. T., Suk, K. & Lee, W. H. Macrophages express membrane bound form of APRIL that can generate immunomodulatory signals. *Immunology* **131**, 350-356, doi:10.1111/j.1365-2567.2010.03306.x (2010).

187 Sun, B. *et al.* Sox4 is required for the survival of pro-B cells. *J Immunol* **190**, 2080-2089, doi:10.4049/jimmunol.1202736 (2013).

188 Rodig, S. J. *et al.* The pre-B-cell receptor associated protein VpreB3 is a useful diagnostic marker for identifying c-MYC translocated lymphomas. *Haematologica* **95**, 2056-2062, doi:10.3324/haematol.2010.025767 (2010).

189 Stros, M., Bacikova, A., Polanska, E., Stokrova, J. & Strauss, F. HMGB1 interacts with human topoisomerase IIalpha and stimulates its catalytic activity. *Nucleic Acids Res* **35**, 5001-5013, doi:10.1093/nar/gkm525 (2007).

190 Stros, M., Polanska, E., Struncova, S. & Pospisilova, S. HMGB1 and HMGB2 proteins up-regulate cellular expression of human topoisomerase IIalpha. *Nucleic Acids Res* **37**, 2070-2086, doi:10.1093/nar/gkp067 (2009).

191 Yoshida, H., Matsui, T., Yamamoto, A., Okada, T. & Mori, K. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107**, 881-891 (2001).

192 Shaulian, E. AP-1--The Jun proteins: Oncogenes or tumor suppressors in disguise? *Cell Signal* **22**, 894-899, doi:10.1016/j.cellsig.2009.12.008 (2010).

193 Zhan, F. *et al.* CKS1B, overexpressed in aggressive disease, regulates multiple myeloma growth and survival through SKP2- and p27Kip1-dependent and -independent mechanisms. *Blood* **109**, 4995-5001, doi:10.1182/blood-2006-07-038703 (2007).

194 Chang, H. *et al.* Multiple myeloma patients with CKS1B gene amplification have a shorter progression-free survival post-autologous stem cell transplantation. *Br J Haematol* **135**, 486-491, doi:10.1111/j.1365-2141.2006.06325.x (2006).

195 Potthoff, M. J. *et al.* Regulation of skeletal muscle sarcomere integrity and postnatal muscle function by Mef2c. *Mol Cell Biol* **27**, 8143-8151, doi:10.1128/MCB.01187-07 (2007).

196 Ott, C. J. *et al.* Chromatin Accessibility Profiling Reveals Cis-Regulatory Heterogeneity and Novel Transcription Factor Dependencies in Multiple Myeloma. *Blood* **132**, 1313-1313, doi:10.1182/blood-2018-99-119941 (2018).

197 Guikema, J. E. *et al.* CD27 is heterogeneously expressed in multiple myeloma: low CD27 expression in patients with high-risk disease. *Br J Haematol* **121**, 36-43 (2003).

198 Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20**, 1472-1478, doi:10.1038/nm.3733 (2014).

199 Zhang, X., Lv, D., Zhang, Y., Liu, Q. & Li, Z. Clonal evolution of acute myeloid leukemia highlighted by latest genome sequencing studies. *Oncotarget* **7**, 58586-58594, doi:10.18632/oncotarget.10850 (2016).

200 Dang, H. X. *et al.* ClonEvol: clonal ordering and visualization in cancer sequencing. *Ann Oncol* **28**, 3076-3082, doi:10.1093/annonc/mdx517 (2017).

201 Deshwar, A. G. *et al.* PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* **16**, 35, doi:10.1186/s13059-015-0602-8 (2015).

202 Zelle-Rieser, C. *et al.* T cells in multiple myeloma display features of exhaustion and senescence at the tumor site. *J Hematol Oncol* **9**, 116, doi:10.1186/s13045-016-0345-3 (2016).

203 Shen, C. J., Yuan, Z. H., Liu, Y. X. & Hu, G. Y. Increased numbers of T helper 17 cells and the correlation with clinicopathological characteristics in multiple myeloma. *J Int Med Res* **40**, 556-564, doi:10.1177/147323001204000217 (2012).

204 Shain, K. H. *et al.* Beta1 integrin adhesion enhances IL-6-mediated STAT3 signaling in myeloma cells: implications for microenvironment influence on tumor survival and proliferation. *Cancer Res* **69**, 1009-1015, doi:10.1158/0008-5472.CAN-08-2419 (2009).

205 Duarte, L. F. *et al.* Histone H3.3 and its proteolytically processed form drive a cellular senescence programme. *Nat Commun* **5**, 5210, doi:10.1038/ncomms6210 (2014).

206 Chevrier, S. *et al.* The BTB-ZF transcription factor Zbtb20 is driven by Irf4 to promote plasma cell differentiation and longevity. *J Exp Med* **211**, 827-840, doi:10.1084/jem.20131831 (2014).

207 Howlett, M., Menheniott, T. R., Judd, L. M. & Giraud, A. S. Cytokine signalling via gp130 in gastric cancer. *Biochim Biophys Acta* **1793**, 1623-1633, doi:10.1016/j.bbamcr.2009.07.009 (2009).

208 Selander, K. S. *et al.* Inhibition of gp130 signaling in breast cancer blocks constitutive activation of Stat3 and inhibits in vivo malignancy. *Cancer Res* **64**, 6924-6933, doi:10.1158/0008-5472.CAN-03-2516 (2004).

209 Kawano, Y. *et al.* Hypoxia reduces CD138 expression and induces an immature and stem cell-like transcriptional program in myeloma cells. *Int J Oncol* **43**, 1809-1816, doi:10.3892/ijo.2013.2134 (2013).

210 Muz, B. *et al.* A CD138-independent strategy to detect minimal residual disease and circulating tumour cells in multiple myeloma. *Br J Haematol* **173**, 70-81, doi:10.1111/bjh.13927 (2016).

211 Zahoor, M. *et al.* Hypoxia promotes IL-32 expression in myeloma cells, and high expression is associated with poor survival and bone loss. *Blood Adv* **1**, 2656-2666, doi:10.1182/bloodadvances.2017010801 (2017).

212 Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-1817, doi:10.1093/bioinformatics/bts271 (2012).

213 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

214 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420, doi:10.1038/nbt.4096 (2018).

215 Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687-696, doi:10.1126/science.1198704 (2011).

216 O'Gorman, W. E. *et al.* Mass cytometry identifies a distinct monocyte cytokine signature shared by clinically heterogeneous pediatric SLE patients. *J Autoimmun*, doi:10.1016/j.jaut.2017.03.010 (2017).

217 Lin, D., Gupta, S. & Maecker, H. T. Intracellular Cytokine Staining on PBMCs Using CyTOF Mass Cytometry. *Bio Protoc* **5**, doi:10.21769/BioProtoc.1370 (2015).

218 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

219 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801, doi:10.1093/nar/gkx1081 (2018).

220 Kyle, R. A. *et al.* A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *N Engl J Med* **346**, 564-569 (2002).

221 Landgren, O. *et al.* Monoclonal gammopathy of undetermined significance (MGUS) precedes multiple myeloma: a prospective study. *Blood* (2009).

222 Weiss, B. M., Abadie, J., Verma, P., Howard, R. S. & Kuehl, W. M. A monoclonal gammopathy precedes multiple myeloma in most patients. *Blood* **113**, 5418-5422, doi:10.1182/blood-2008-12-195008 (2009).

223 Korde, N., Kristinsson, S. Y. & Landgren, O. Monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM): novel biological insights and development of early treatment strategies. *Blood* **117**, 5573-5581, doi:10.1182/blood-2011-01-270140 (2011).

224 Teitelbaum, A., Ba-Mancini, A., Huang, H. & Henk, H. J. Health care costs and resource utilization, including patient burden, associated with novel-agent-based treatment versus other therapies for multiple myeloma: findings using real-world claims data. *Oncologist* **18**, 37-45, doi:10.1634/theoncologist.2012-0113 (2013).

225 Brown, L. M. *et al.* Multiple myeloma and family history of cancer among blacks and whites in the U.S. *Cancer* **85**, 2385-2390 (1999).

226 Smith, C. J., Ambs, S. & Landgren, O. Biological determinants of health disparities in multiple myeloma. *Blood Cancer J* **8**, 85-85, doi:10.1038/s41408-018-0118-z (2018).

227 Kazandjian, D. *et al.* Molecular underpinnings of clinical disparity patterns in African American vs. Caucasian American multiple myeloma patients. *Blood Cancer J* **9**, 15, doi:10.1038/s41408-019-0177-9 (2019).

228 Manojlovic, Z. *et al.* Comprehensive molecular profiling of 718 Multiple Myelomas reveals significant differences in mutation frequencies between African and European descent cases. *PLoS genetics* **13**, e1007087, doi:10.1371/journal.pgen.1007087 (2017).