Unifying genomics and transcriptomics in single cells with ResolveOME amplification chemistry to illuminate oncogenic and drug resistance mechanisms

Jon S. Zawistowski[1*], Isai Salas-González[1], Tatiana V. Morozova[1], Jeff G. Blackinton[1], Tia Tate[1], Durga Arvapalli[1], Swetha Velivela[1], Gary L. Harton[1], Jeffrey R. Marks[2], E. Shelley Hwang[2], Victor J. Weigman[1], and Jay A.A. West[1]

[1]BioSkryb Genomics, Inc., Durham, NC

[2]Department of Surgery, Duke University Medical Center, Durham, NC

* = corresponding author

ABSTRACT

Discovering genomic variation in the absence of information about transcriptional consequence of that variation or, conversely, a transcriptional signature without understanding underlying genomic contributions, hinders understanding of molecular mechanisms of disease. To assess this genomic and transcriptomic coordination, we developed a new chemistry and method, ResolveOME, to extract this information out of the individual cell. The workflow unifies template-switching full-transcript RNA-Seq chemistry and whole genome amplification (WGA), followed by affinity purification of first-strand cDNA and subsequent separation of the RNA/DNA fractions for sequencing library preparation. In the ResolveOME methodology we leverage the attributes of primary template-directed amplification (PTA)[1] to enable accurate assessment of single-nucleotide variation as a DNA feature—not achieved with existing workflows to assess DNA + RNA information in the same cell.

We demonstrated the validity of the technique in the context of two major phenomena in oncology: tumor heterogeneity (leading to cancer progression) and treatment resistance. Material from a primary patient breast cancer and an acute myeloid leukemia (AML) cell line, MOLM-13, was used to highlight multiomic biomarker paradigms enabled by this chemistry. Performance of the PTA-enabled genome amplification was largely unaffected by addition of RNA enrichment, with control WGS results showing > 95% genome coverage, precision > 0.99 and allele drop out < 15%. In the RNA fraction of the chemistry, we were able to routinely retrieve full-length transcripts that demonstrate a ratio of 1 for 5'/3' bias, with increased coverage of intronic regions and 5' regions that are indicative of novel transcripts, showing strength of the template switching mechanism to capture isoform information with sparsity rates < 75%. We find remarkable cellular variability of revealed biomarkers at both in the genome and transcriptome despite employing a relatively small number of individual cells. In our primary patient sample of ductal carcinoma in situ (DCIS)/invasive ductal carcinoma (IDC) we observed oncogenic *PIK3CA* driver mutations and prototypical DCIS copy number alterations binned into heterogenous single-cell classes of genomic lesions. Within our quizartinib-treated MOLM-13 cells, we identified multiple potential mechanisms of resistance within seemingly sporadic changes and were able to associate specific mutation, copy number and expression significantly correlated to treatment. In this latter scenario, the DNA arm of our combined workflow uncovered a *secondary* FLT3 (non-internal tandem duplication (ITD)) mutation as a candidate primary driver of resistance to drug while the RNA arm showed matched transcript upregulation of AXL signal transduction as well as enhancer factor modulation. Importantly, proximal candidate regulatory SNVs, outside of the

CDS, were identified and associated to upregulated transcripts *in cis*. The study highlights that both the genome and transcriptome are dynamic, leading to a set of combinatorial alterations that affect cellular evolution and that fate can be identified through ResolveOME application to individual cells.

INTRODUCTION

Cancer is a disease of remarkable variation and heterogeneity between the individual cells comprising the bulk tumor tissue. While a multitude of studies have described these changes across the evolution of cancer, etiology is still driven by speculation in most cancers. This is borne out in the molecular complexity underlying the resiliency of cancer cells in drug resistance, whereby single nucleotide variation (SNV) and copy number variation (CNV) at the genomic level contributes to resistance in concert with transcriptional adaptation[2]. While one of these modes can be a dominant driver, there is increasing evidence that the modes are not mutually exclusive and instead can synergize to change cell state leading to resistance[3]. It will therefore become paramount to assay these multiple "-omic" tiers (genomic and transcriptomic) in single cells, as bulk sequencing provides an incomplete view of the inherent heterogeneity in each of these tiers. Cancer's evolution is driven through a complex molecular orchestration, where the interdependence of genomic and transcriptomic changes occurring in each cell convey some of the major fitness advantages that drive expansion and drug resistance. The nature of current genomic and transcriptomic assays muddle the underlying clonal structure by reducing genomic data to tissue-based averages. Recent methods aimed at simultaneously monitoring both RNA and DNA in single cells have made this linking possible, but contain uneven genome coverage and low allelic balance, limiting the ability to assess single nucleotide variation genome-wide with accuracy[4,5].

To overcome this challenge, we enhanced our previously-characterized PTA[1] workflow and extended a second modality of transcriptome enrichment. The method is differentiated through enhanced genome coverage and uniformity, along with allelic balance, wherein both copies of the genome are equivalently and uniformly amplified. This is an underlying attribute that allows both CNV and SNV detection from an amplified genome of a sample as finite as a single cell with high accuracy[1]. The ability of PTA to provide this degree of uniformity and accuracy stems from the unfavored recopying of synthesized strands, driven by nucleotide terminators that limit the size of the amplicons, and coincidentally this amplicon-size distribution (500-1500bp) is suitable for the natural distribution of transcript lengths.

We present a single-well integration of single-cell transcriptome and genome amplification where a standard PTA reaction was modified to include a reverse transcription (RT) step prior to single-cell genome amplification and we designate this multiomic enrichment ResolveOME. In this workflow, PTA amplifies the genomes of single cells immediately after the RT reaction is concluded in a single-well reaction. Using template switch-based reverse transcription, we created first-strand cDNA molecules that could be affinity purified and pre-amplified prior to RNA-Seq sequencing library creation. The net result from the combined amplification reaction is a biotin labeled cDNA pool derived primarily from the cytosolic transcripts, available for streptavidin purification, and a pool of amplified genomic material from the single cell. At the conclusion of the genome amplification reaction the cDNA fraction is separated from the amplified genome material whereby libraries from each pool are created. The resulting sequencing data offers the ability to define both genomic and transcriptomic plasticity at single-cell resolution. Specifically, the delineation of isoform expression, combined with ability to annotate the underlying structural variation and single nucleotide changes from the genome of the same cell (**Figure 1a**), allows the assessment of genomic "penetrance", and the definition of mechanisms that drive single-cell fate.

Prior multi-omic efforts have pioneered the pairing of genomic and transcriptomic information from the same single cell but have the primary shortcoming of incomplete genome coverage and associated non-uniformity of coverage—leaving uncovered genomic valleys that may harbor deleterious single nucleotide variants that would remain undetected. Indeed, multiple displacement amplification (MDA)[6] drives the genomic amplification of G&T-seq and DR-Seq has genomic amplification uniformity comparable to that of MALBAC[7], both of which are outperformed by PTA[1] in terms of genomic coverage, allelic balance and SNV calling metrics. Definition of clonal evolution at the SNV/CNV level in a primary patient sample has been accomplished utilizing G&T-seq, yet was limited to a candidate gene survey of exome-level data whereby clusters where defined by 59 oncogenes[8] and another studying employing G&T-seq limited their analysis to the RNA workflow of the method to take advantage of the low input requirement, without assessment of genomic level data[9] . Thus, we address here an unmet need to add genome-wide, high sensitivity and high precision SNV calling capability to a joint DNA/RNA single-cell methodology. Further, we demonstrate the criticality of these measurements, whereby single nucleotide variation fundamentally affects cell state[10] and tumor progression[11,12].

We describe here the utility of these unified "-omic" layers, highlighting heterogenous genomic variation and consequential phenotypic alterations in single cells that both are correlated with the development of resistance to a targeted therapeutic in a cell line model of acute myeloid leukemia, and in oncogenic mechanisms in primary breast cancer cells whereby the insights gained could not be inferred by a single dataset (genome or transcriptome) alone.

RESULTS

Amplification product yield of ResolveOME workflow

Prior to demonstrating biological utility of ResolveOME in a cell line drug resistance model and in a primary patient sample, we sought to demonstrate technical performance of the methodology using a benchmark cell line 1000 Genomes cell line, NA12878[13]. The RNA and DNA arms of the protocol were first assessed using metrics from the template-switching RNA-Seq chemistry or PTA chemistry in isolation to compare to the metrics when the chemistries were unified in the combined ResolveOME protocol.

We first generated ResolveOME data with FACS-sorted NA12878 single cells and with purified total NA12878 RNA or genomic DNA as amplification controls using the workflow shown in **Figure 1a.** Efficiency of the yield of the PTA product and cDNA products from the unified protocol are shown in **Figure 1b**. We obtained approximately 1-1.5 µg of DNA amplification product from single cell genomes and approximately 100-200 ng of cDNA product representing the single cell transcriptome. Importantly, no-template control (NTC) reactions showed lack of detectable product and additionally there was negligible (<50 ng) yield in the DNA fraction from control RNA input using Qubit fluorometer (ThermoFisher). We did detect low-level background amplification of the genomic DNA control input in the cDNA fraction, due to known promiscuity of reverse transcriptase in the absence of mRNA template[14]. By contrast, this background amplification does not occur in reactions with single cells as the genome material is sequestered in the non-lysed nucleus during the reverse transcription workflow of ResolveOME.

Comparative genomic performance of ResolveOME

As default practice prior to passing single cell samples to deep sequencing for SNV analysis we performed low-pass QC sequencing, and as part of the analysis pipeline, determined an estimation of library complexity with the PreSeq count algorithm[15]. QC standards set for ResolveDNA (product solution for PTA) are >3.0E9 PreSeq count value upon low-pass sequencing, an empirically-defined proxy for genomic coverage and uniformity that predicts high-depth sequencing will yield strong allelic balance and high sensitivity and precision of single nucleotide variant calling. The average PreSeq count of ResolveOME single cells from **Supplementary Table 1a** was 3.76E9 with a standard deviation of +/- 2.27E8. The overall robust performance of single cells and genomic DNA controls warranted subsequent deep sequencing for metric comparison of classical PTA (ResolveDNA) to PTA from the ResolveOME multi-omic workflow.

Upon high-depth sequencing (2X150bp, down-sampling to 4.5E8 total reads, ~20x genome depth) and processing through our pipeline, we initially reviewed allelic balance, (ability to represent both alleles through enrichment and a strength of our ResolveDNA methodology[1,16]). The inverse of allelic drop out (ADO) is allelic balance, which is the proportion of known heterozygous loci that are called heterozygous following sequencing. Variants within these loci have allele frequencies between 10% and 90% at each locus. A review of allelic balance of ResolveOME showed 85.5% (+/-3.4%), which is closely comparable to the 88.2% (+/-4%) for ResolveDNA, across 10 replicates each (**Figure 2a**). We then confirmed that genomic coverage at a range of depths did not significantly differ (**Figure 2b**) between the workflows. Lastly, it was critical to demonstrate that the allelic balance and coverage obtained from the ResolveOME workflow culminated in the ability to call SNVs with confidence. **Figure 2c** highlights individual ResolveOME NA12878 cells with a SNV calling sensitivity range of 0.90-0.95 and with precision >0.99, akin to ResolveDNA data[1]. Collectively, these data suggest that, despite the upstream reverse transcription chemistry modifications to generate transcriptome data, amplification performance of single-cell genomes by PTA persists in performance.

Comparative transcriptomic performance of ResolveOME

In choosing a transcriptomic scheme to unite with PTA our goal was to be as comprehensive as possible in capturing the diversity of RNA-based modes of oncogenic and drug resistance mechanisms, and, equally as importantly, to enable the ascertainment of genomic lesions manifesting at the RNA level. We therefore designed a template-switching reverse transcription scheme for ResolveOME that captured full-transcript information as opposed to either 5' or 3' end counting to enhance ability to detect isoforms and identify fusions. This chemistry enables even coverage across transcripts and as shown in **Figure 3a**, where increased coverage of the 5' region (top) which typically is affected by degradation (or reverse transcriptase performance) proportional to the distance from 3'-polyA, is shown. This confirms expected behavior of the template-switching chemistry in the RNA arm workflow. The distribution of read depth across gene bodies of a set of housekeeping genes is presented in **Figure 3a (bottom)**, with all exons equally represented. Feature quantification in the across our defined transcriptome is shown in **Figure 3b**, highlighting the ability to identify a variety of transcript types and components. Progression of the performance is shown in this figure from what is observed in a bulk dataset (bar 1, aggregated datasets listed in **Supplementary Data 1**) vs. features such as bulk RNA isolation (bars 2 and 4) or single cell (bars 3 and 5) against standalone BioSkryb mRNA-seq (bars 2 and 3) or the ResolveOME combined workflow (bars 4 and 5). Notably, intergenic background was routinely below 5% of aligned reads, providing a broader space for isoform detection.

As further performance benchmarking of cell quality post mapping to reference transcriptome, we established performance patterns of common metrics with well characterized Human Brain Reference RNA (HBRR) and Universal Human Reference RNA (UHRR) as additions to the NA12878 cell line and displayed composite features in **Figure 3c**. We identified read and genomic feature mapping percentages, as well as total genes discovered, as criteria for evaluating sequencing quality. We also examined the dynamic range of expression and expression patterns in well-known housekeeping genes. We computed various markers of DNA contamination, sample degradation, and/or bias as a percentage of exonic (more than 55%), and intergenic mapping (less than 5 %) as characteristics of the ResolveOME RNA fraction. Another important metric for measuring the quality of single cell experiments is the number of genes found (>0 counts) per cell. For NA12878 cells there was an average of over 3000, whereas the average number of HBRR and UHRR genes discovered was around 6 and 7 thousand, respectively. Lastly, median absolute deviation (MAD) and percent coefficient of variation (CV) scores were calculated on normalized CPM values for general use housekeeping genes for cross-tissue studies[17]. These metrics measure reproducibility and are robust approaches to measuring sample variability. Overall, we observe comparable monotonous expression metrics across our housekeeping genes of choice, as well as MAD values ranging from 0.25 to 1 for our HBRR and UHRR benchmarks, suggesting that these genes exhibit little variability in expression across cells. With NA12878, we saw a bit more irregularity, which might imply higher variability or unsuitable housekeeping genes. Correspondingly, CV rates varied from 14 to 30 percent, despite NA12878 exhibiting more variation. For each cell, the dynamic range of expressed genes was around 1300 (HBRR), 1400 (UHRR), and 1900 (NA12878) CPM. In all single cells analyzed, mitochondrial read percentage was <10%, with most cells averaging less than 5%, indicating that single-cell lysis was optimal for capturing mRNA and other polyadenylated transcripts and that the amplified cells were healthy[18].

In addition to NA12878 cells, which are relatively transcriptionally quiescent, we also assessed uniquely expressed protein coding genes in single cells from our DCIS and MOLM-13 material, **Figure 3d.** MOLM-13 AML cells averaged ~5000-5500, while FACS-enriched single cells from a primary DCIS/IDC tumor specimen yielded less expressed genes than the cell line models, averaging ~3500, potentially owing to sample integrity of the primary singulated cells and the increased number of workflow steps from surgical resection to FACS.

Generation of a drug resistance model in MOLM-13 acute myeloid leukemia cells

We then expanded from the DNA and RNA performance metrics of ResolveOME on control cells, and moved to generate unified genomic and transcriptomic information from a model of drug resistance. Prior to looking at heterogenous effects of drug resistance, we wanted to make sure the chemistry could regenerate MOLM-13's known genomic features. We started by karyotypically assessing the cell to match published reports and provide context for interpreting CNV analysis. The combined copy number analysis of all MOLM-13 cells used in this study can be found in **Figure 4a.** Prior to drug resistance modeling, our MOLM-13 line exhibited hallmarks of the initial cell line establishment including trisomies of Chr.6 and Chr. 13 (49,,2n.,XY,+6,+8,+13, 49,,2n., XY, +6,+8, ins(11;9)(q23;p22p23), ins(11;9)(q23;p22p23),del(14)(q23.3;q31.3)[19]. Our MOLM-13 line exhibited (**Figure 4b**) additional gains including the presentation of trisomy 5 and pentasomy 8 concomitant with other translocations (52,XY,+5,+6,+8,+8,+del(8p),add(11q),+13,add(17p)).

To demonstrate the utility of concurrent genomic and transcriptomic information in single cells in the context of drug resistance, we created a model by exploiting the presence of an internal tandem duplication (ITD) mutation in MOLM-13 cells [19]. Since the ITD mutation, found in ~20% of AML patients, hyperactivates FLT3 signaling and results in poor prognosis and relapse[20], we treated non-resistant, drug-sensitive cells with a continual dose of 2 nM quizartinib. This drug is a selective type II kinase inhibitor targeting *FLT3*. We found resistance emerged following initial marked growth inhibition/apoptosis (See Methods, **Supplementary Figure 1**).

Distinction in single-cell CNV profiles among parental and quizartinib-resistant MOLM-13 cells

As an initial assessment of single-cell genomic variation in the MOLM-13 quizartinib resistance model we performed CNV analysis following the ResolveOME workflow on 9 parental "P" and 10 quizartinib-resistant "R" cells. Utilizing sequencing data to yield ~25x coverage and a 500 kb window size, copy number gain was evident for chromosomes 5,6,8, and 13 (**Figure 4a**) and concordant with our karyotypic data for the parental cells (**Figure 4b**). Single-cell CNV heterogeneity immediately emerged from the data. Within the "P" cohort, gain to 3N was observed for 9/9 cells for Chr. 5, yet 5/9 cells showed additional 5p gains. Most relevant, we observed heterogenous copy number variation between "P" and "R" single cells. No resistant cells exhibited the additional 5p gain found in the parental cohort, and furthermore, 7/10 resistant cells did not have any amplification of Chr. 5 as a diploid 2n state, suggesting that this was selected for to mediate drug resistance in part by expression consequences on multiple Chr.5-resident genes. In addition to this general implication of Chr. 5 as a candidate contributor to quizartinib resistance, we observed 19q gain uniquely in 4/10 resistance cells. Taken together, we defined a CNV paradigm for the MOLM-13 resistance model that could now be used as context for the SNV and transcriptional layers to be subsequently defined by ResolveOME.

Acquisition of a secondary FLT3 mutation as a key driver of drug resistance

We next sought to determine candidate key drivers of quizartinib resistance beyond gross CNV at the increased level of genomic resolution of the SNV. Expectedly, all parental and resistant single cells harbored *FLT3* ITD (**Figure 5a**). In contrast, a missense mutation N841K was detected in all quizartinib resistant cells (**Figure 5b**). *FLT3* N841K has previously been detected in AML patients[21], resides in the activation loop of FLT3[22], and furthermore, mutation of the residue corresponding to N841 in the closely-related receptor tyrosine kinase KIT is activating[23]. This strongly suggests that N841K is a chief secondary mutation to ITD and is plausibly contributing to quizartinib resistance in this model by preventing efficiency of drug binding.

To assess whether the N841K *FLT3* secondary mutation may have arisen *de novo* or was an existing genetic variant clone in the parental population we employed a custom quantitative PCR-based genotyping assay to distinguish between the two scenarios. This probe set, emitting fluorescence of differing wavelengths for allelic discrimination between N841 and K841 upon probe binding and dequenching, was employed in qPCR assays of genomic DNA isolated from either parental or quizartinib-resistant MOLM-13 cells. In parental cells, while amplification of N841 dominated, a low but detectable level of K841 presented (**Figure 5c**). Resistant cells displayed a contrasting scenario, whereby there was equal signal from N841 and K841. These data suggest that *FLT3* K841 existed as an extremely rare clone in the original MOLM-13 cell line which upon the selective pressure of quizartinib was enriched to domination of the resistant cell line likely due to its ability to affect drug binding—thus highlighting our cell line model's emulation of clonal selection in patient tumors. While this variation independently makes a compelling case, with the increased biomarker resolution, we have very well-defined groups, identified by the heatmap in **Figure 6** that showcases differential genotypes across the 2 groups.

Heterogenous SNV in MOLM-13 quizartinib resistance

We also interrogated a candidate list of genes representing multiple functional classes—signaling, epigenetic, tumor suppressor, spliceosome, cohesion complex genes—previously implicated in AML pathogenesis[24] for SNV. With no resistant-specific coding sequence changes in single cells identified with this candidate approach other than the *FLT3* secondary mutation, we began an unbiased search for mutations that may be contributing to quizartinib resistance and for those mutations representing subclones and not found in all resistant cells (**Supplementary Tables 2 and 3**). We first sought to stratify the variant call file by rarer functional class of mutation, stop codon gain and frameshift mutation, due to the increased likelihood of deleterious functional consequences. We identified a heterozygous nonsense mutation in the splicing and mRNA stability factor *CELF4* in 7/10 quizartinib-resistant cells where the change was not identified in any single cells of the parental cohort. Frameshift mutations were identified in the metabolic enzyme *ADSS1* at K291 (c.870dupC) in 8/10 quizartinib resistant and 0/9 parental cells and in the GTP-binding protein *RRAGC* at A57 (c.167dupG) in 5/10 resistant cells and in 0/9 parental cells. While we initially prioritized these variants, we were unable to detect expression of their cognate transcripts (**Figure 7b**). This suggested that either these genes were lowly expressed in MOLM-13 cells, unexpressed at the time of cell capture and extraction, and/or beyond our limit of detection with ResolveOME. These findings motivated us to more comprehensively quantify the single nucleotide variation in our model, as well as to prioritize genomic variants associated with gene expression, which ResolveOME uniquely enables for single cells.

We thus subsequently employed a variant filtering/prioritization strategy to identify single nucleotide variation present in quizartinib-resistant single cells but not in parental single cells. From this analysis (see Methods), we used multinomial logistic regression analysis and a Wald test to yield 6444 SNVs that were differentially prevalent between parental and resistant single cells ($p < 0.05$). **Figure 6** presents this statistically significant genotypic variation in a heat map and allows visualization of conversion of homozygous reference (0/0) to heterozygous (1/0, 0/1) or homozygous alternate (1/1) alleles in the resistant cells, and, conversely, loss of heterozygous genotypes in the resistant cells to homozygous reference (**Supplementary Table 2,** coding-related **and Supplementary Table 3** intergenic-related). Additional filtration by allowed us to focus on missense variations differing in parental vs resistant line in **Supplementary Figure 2**. As a prioritized missense mutation of biological interest with validated mRNA expression, we report A109V in the E3 ubiquitin ligase gene *RNF167*, found in all 10 quizartinib-resistant cells but not present in cells of the parental cohort.

In addition to prioritizing coding sequence variation above, variant filtration (See details in Methods) allowed us to discern a remarkable degree of single nucleotide variation in intergenic space occurring in our quizartinib resistance model. We catalogued 8601 intergenic SNVs in our parental cells vs 2167 in our quizartinib resistant cell cohort present in at least 25% of all cells within the group. This group-specific variation shows context of both selection of existing genomic variation in response to drug treatment and in de novo mutation and an exemplification of the high degree of plasticity in the genome (**Figure 6, Supplementary Table 3**).

MOLM-13 quizartinib-resistant cells exhibit a distinct transcriptional signature including adaptive bypass

At the SNV level, there was distinction between parental and resistant MOLM-13 single cells in principal coordinate analysis (p<0.05, **Figure 7a**). The same trend was seen in the ResolveOME transcriptomes of the two MOLM-13 single cell cohorts (data not shown). We present in **Figure 7b** a dendrogram highlighting differentially expressed transcripts between the P and R single cells and labeled by biotype indicating the categorical nature of the upregulated or downregulated transcript. We highlight two specific examples here that highlight both DNA and RNA-level contributions to drug resistance in this model.

Firstly, from our differentially expressed gene set we noted marked upregulation of *GAS6*, a ligand for the receptor tyrosine kinase AXL. The AXL pathway, specifically through downstream STAT3 cell proliferation and PI3K/AKT survival signaling, has been shown to be a bypass pathway for FLT3 inhibition[25,26] (**Supplementary Figure 3**). We also observed concurrent transcriptional upregulation of the small GTPase *RAC1*, which may be synergistic with upregulation of the AXL-STAT3 and AXL-PI3K/AKT signaling axes[27,28]. Collectively, these transcriptional responses indicate a mode of adaptive transcriptional bypass that is occurring in the same cell harboring a DNA-level, secondary *FLT3* mutation driving drug resistance.

Intriguingly, we also noted the pioneer transcription factor CEBPA CCAAT/enhancer-binding protein alpha (C/EBPα) transcriptional upregulation in quizartinib-resistant cells (**Figure 7b**). Truncating mutations in *CEBPA* are found in ~10-15% of AML patients[29,30], leading to expression of an N terminal fragment of CEBPA, p30, with potential dominant negative[31] activity. As *CEBPA* resides on Chr. 19q13.11, concomitant with the transcriptional upregulation of *CEBPA*, we observed Chr.19q gain in a subset of quizartinib-resistant cells (**Figure 7c**) suggesting a potential genomic mechanism of *CEBPA* expression upregulation and exemplifying the power of the unification of single-cell genomic and transcriptomic data.

While plausible, we did not observe a positive correlation between copy number gain at *CEBPA* upregulation in individual cells, suggesting that the mode of transcript upregulation is epigenetic in nature. We therefore ascertained the relationship of ploidy to gene expression genome-wide using a zero-inflated linear model. Ploidy and gene expression were not direct correlates using a 500kb window size, except for a set of genes whereby statistically meaningful associations were identified ($p<0.05$) with this model (**Figure 7d**). **Supplementary Table 4** shows each gene identified and summarizes copy number and expression correlates. This highlights the importance of concurrent transcriptomic assessment when interpreting copy number alterations in single cells, as well as highlights the significant single cell heterogeneity that occurs in terms of ploidy across sub-megabase chromosomal intervals.

In addition to these examples of transcriptional drug resistance mechanistic hypotheses informed by combined single-cell genomic and transcriptomic data, we performed differential transcript usage (DTU) analysis (**Figure 7e**) as we were empowered by full-length (vs. 3' end counting) data to make transcript isoform insights. We identified an isoform of *HADHA*, whereby its expression was unique to the quizartinib-resistant population and absent in all but one parental cell—whereby the isoform with biased expression in the resistant cells was shorter (~2688 bp) than the parental isoform (2943 bp). Similarly, 7/10 quizartinib-resistant single cells exclusively expressed an isoform of *PPP1R14B* containing an additional 5' exon while the majority of parental cells expressed none of the isoform. In total, we identified 6 instances of isoform specificity between parental and quizartinib-resistant populations for additional genes RPS3, HSPA4, SUGT1, CAPNS1.

Identification of candidate regulatory SNVs modulating transcript levels in resistant cells

As we identified occurrences of genomic lesions of interest that did not associate with the predicted transcriptional output, we sought to identify single nucleotide variation that would influence the expression of a proximal gene as a candidate regulatory variant in **Figure 8a**. While we earlier failed to identify a correlation between Chr. 19q gain and *CEBPA* mRNA upregulation in resistant cells (**Figure 7c**), we identified a candidate distal promoter/enhancer SNV ~20kb 5' of the *CEBPA* transcriptional start site with a genotypic bias between parental and resistant cells (**Figure 8b**) in the variant call file defining SNVs. We then moved to an unbiased approach, whereby we performed ZLM (zero-inflated linear model) modelling of transcriptional abundance of a gene across the genotypes of the cohorts. For initial analysis we limited our SNV detection to intragenic or promoter (0 to -5000 relative to the transcriptional start site). Upregulation of *MYC* expression was observed in resistant vs parental cells, and we identified a candidate intronic regulatory variant with a genotypic bias to the reference 0/0 allele in resistant cells while all but one of the parental single cells harbored the 0/1 genotype for the candidate regulatory variant (**Figures 8a,c**). An additional example of a candidate proximal regulatory SNVs with a parental/resistant genotypic bias and concomitant expression dichotomy between the parental and resistant cells included a candidate promoter mutation in the *PABPC4* gene, encoding a poly(A) binding protein, within 5' kb upstream of the transcriptional start site (**Figures 8a,d**). All variants identified with this analysis of course warrant functional investigation for validity but emphasize the ability of ResolveOME to generate candidate regulatory SNVs through the pairwise analysis of genotype shifting and transcriptional modulation in individual cells. Extending this analysis to all of intergenic space and associating the SNVs with ENCODE ChIP-Seq data will be a powerful tool to generate larger numbers of candidates influencing drug resistance and oncogenesis.

Primary DCIS/IDC single cells exhibit heterogeneous classes of chromosomal loss

After demonstrating the utility of ResolveOME's unification of genomic and transcriptomic data to elucidate single-cell drug resistance mechanisms in a cell line model, we importantly sought to demonstrate analogous multi-omic utility in elucidating single-cell oncogenic mechanisms in primary human cancer. To this end, we initiated a collaboration with Duke University Medical Center to elucidate genomic and transcriptomic contributions to the transition of premalignant ductal carcinoma in situ (DCIS) to invasive ductal carcinoma (IDC). We first enriched dissociated single cells from tumor tissue from a mastectomy by FACS. The tumor pathology for this patient indicated

ER/PR (estrogen receptor/progesterone receptor) positivity but lack of HER2 expression precluded the use of a HER2 antibody for FACS enrichment. As such, we proceeded with a FACS strategy to enrich for ductal epithelial cells by epithelial cell adhesion molecule (EpCAM) epitope enrichment, and simultaneously to capture "EpCAM low" cells as enrichment controls.

As with our MOLM-13 resistance model, we first assessed CNV in primary DCIS/IDC single cells. We performed the ResolveOME workflow on 16 single cells with pronounced EpCAM expression and 4 single cells with negligible EpCAM expression. Using the same genome coverage (25x) as the MOLMs, and 500 kb windows we first assessed CNV in the "EpCAM high" cohort of single cells. Distinct classes of CNV emerged, whereby single cells exhibited discrete chromosomal losses. As one class, 5/20 cells harbored near complete loss of Chr. 13 with concurrent loss of 16q/17p, **Figure 9**. The most abundant class (12/20 cells) harbored these copy number alterations plus a third discrete loss of Chr. 11q. Two EpCAM high cells lacked any apparent copy number alteration, and one EpCAM high cell had a more aberrant series of genome-wide chromosomal losses. The observed Chr.13 and 16q/17p loss is consistent with reported copy number alteration in multiple stages of DCIS advancement[32] and coincides with the loss of the prototypical tumor suppressor genes *BRCA2, RB1* and *TP53*. Interestingly, we observed gain of Chr. 13p, a heterochromatic "stalk" devoid of genes in 10/20 EpCAM high cells, and Chr. X gain of unknown significance in 2 EpCAM high cells and 1 EpCAM low cell encompassing the centromere and flanking p and q arm segments. Even with this relatively small cohort of single cells, these data highlight copy number heterogeneity of the primary sample.

Identification of an oncogenic PIK3CA mutation

Prior to genome-wide unbiased assessment of SNV, we assessed exons of the *PIK3CA* gene, one of the most frequently mutated genes across diverse molecular subtypes of breast cancer. We identified the missense mutation N345K in 14/18 EpCAM high cells (**Figure 10c**). N345K is second only to H1047R amongst *PIK3CA* hotspot mutations catalogued by TCGA[33] and is known to influence the interaction of the p85 (*PIK3R1*) regulatory/p110 (*PIK3CA*) catalytic subunits by disruption of the C2/iSH2 domain interface[34–36]. The oncogenic N345K mutation was detected only in the single cells where CNV was observed; initially suggesting that we stratified the relevant ductal epithelial cells with our FACS strategy and the two cells lacking CNV + *PIK3CA* N345K either harbored other genomic variation or were a different cell type—requiring the RNA arm of the ResolveOME protocol to further distinguish between the possibilities.

Single nucleotide variation in DCIS/IDC

We then performed variant filtering to identify novel candidate oncogenic SNVs. As validation of our filtering strategy, *PIK3CA* N345K was identified in the 14/16 cells harboring 11q, 13, 16q/17p copy number loss. We did not detect coding sequence mutation in additional candidate genes known to be influential in ER+ breast cancer[37] (**Supplementary Figure 4**). We thus subsequently cataloged variation that existed in the EpCAM high cells but that was not present in the EpCAM low cells (**Supplementary Table 5**). Analogous to our MOLM-13 model of quizartinib resistance, we noted extensive intergenic genomic SNV in EpCAM high vs. EpCAM low cells.

Cell identity and transcriptional state of DCIS/IDC singulated cells

Of noteworthy utility in a combined genomic/transcriptomic single-cell assay is the capability to link genotype to identity of cell type and to inference of cell state. This was critical in the interpretation of the observed CNV and *PIK3CA* single-cell DCIS/IDC genotypes due to the difficulty in designing a FACS marker schema that unambiguously identifies the ductal epithelial cells of interest from surrounding stromal cells and infiltrating immune cells. Gene expression profiles of EpCAM high and EpCAM low cells separated by principal component analysis (**Figure 10a**) using the PAM50 gene set of genes influential in diverse subtypes of breast cancer[38] (**Figure 10b**). Differential gene expression analysis highlighted gene signature blocks between two primary clades: a cluster of exclusively EpCAM high cells, and a cluster comprised of all EpCAM low cells intermixed with 4 EpCAM high cells (**Figure 10c**). Initial ascertainment of transcripts defining the EpCAM low cells revealed enrichment of in IL-2 and CD4 T cell-defining gene sets, suggesting that these cells may be tumor infiltrating lymphocytes present in this patient's singulated tumor sample. However, further rigor into transcriptome-based cellular annotation with Human Cell Atlas data (See Methods) parsed the EpCAM low cells into stem-cell like, endothelial, fibroblastic and monocyte identities/states (**Figures 10b-e**) which was independent of transcript count (Figure 10a). Four outlier EpCAM high cells exhibited a gene expression signature such that they were placed in the same root clade of the dendrogram as the EpCAM low cells. We identified these cells as having two distinct identities/states: epithelial and monocytic. Intriguingly, while all EpCAM low cells lacked *PIK3CA* N345K or characteristic DCIS copy number loss, the EpCAM high cell in the EpCAM low gene expression signature clade with epithelial identity harbored both of these genomic alterations. This is suggestive of a plasticity of cell state of a ductal epithelial cell and the acquisition of phenotype with stemness attributes as suggested by cell annotation profiles more closely matching tissue stem cell or fibroblast identities (**Figure 10d**). One outlier EpCAM high cell in the EpCAM low clade lacked oncogenic *PIK3CA* mutations and the prototypical DCIS chromosomal losses and displayed a monocytic gene expression profile. For this instance, it is suggestive of infiltration of monocytes in the sample, although we cannot formally exclude the possibility of cell state change of a malignant or benign ductal epithelial cell or infiltration of monocytes in the sample. Furthermore, one putative epithelial cell in this outlier EpCAM high class, although differing from the prototypical DCIS chromosome losses observed in the main EpCAM high clade, harbored a grossly aberrant CNV profile and may represent a malignant cell. Our examples of putative plasticity of phenotypic cell state with regard to oncogenicity warrant ResolveOME analysis of additional cells to determine the frequency of this cell state in the sample or whether it represents stochastic genomic variation that did not persist or was not selected for in the population. Collectively, these data suggest profiling a cell at the transcriptome level only could lead to an incorrect cell classification and underscores that understanding both RNA and DNA -omic tiers is critical to provide proper classification.

Holistic view of MOLM-13 and DCIS/IDC single-cell molecular signatures

Having in succession determined CNV, SNV and transcriptional insights in both the MOLM-13 model of drug resistance and in primary DCIS/IDC it was critical to begin to amass and graphically present interrelationships between the "-omic" layers of data. For MOLM-13, we identified a secondary driver mutation likely affecting drug binding in all single cells yet provided evidence for concurrent transcriptional bypass of *FLT3* signaling, highlighting the importance of ascertaining both DNA and RNA-driven mechanisms of resistance in the same cells.

For primary DCIS/IDC, unification of DNA-level and RNA-level data allows the interpretation of genotypes in the context of expression signatures defining cell type and cell state. Harnessing these layers of molecular information in a heat map/dendrogram quickly conveys the finding that EpCAM expressing ductal epithelial cells harbor both prototypical copy number losses and an oncogenic *PIK3CA* mutation while EpCAM low cells with alternative identities by transcriptomic profile from the same singulated cell sample lack chromosomal loss and this mutation (**Figure 10d**). Yet, cell identification cannot be unambiguously assessed solely by EpCAM FACS protein levels but in leveraging more contemporary cellular annotation methods, we see that IDs can be objectively identified that match the cell's known biological origin or reflect a cell state transition.

DISCUSSION

Each “-omic” tier of molecular information allows a greater ability to comprehensively define the molecular mechanisms of oncogenesis and drug resistance in a tumor. In the single cell tumor biology arena, most work to date has been performed at the transcriptome level, owing to the large-scale adoption of droplet-based methodology facilitating workflow ease and single-cell throughput. While there has been unquestionable advance from droplet-based RNA-Seq studies defining diversity and heterogeneity in transcriptional states including those states defined longitudinally, a gap remains in that there have been few studies providing concurrent genomic data with the gene expression data. This is critical for multiple reasons. Firstly, in the absence of DNA-level information, genomic contributions to the transcriptional or phenotypic state cannot be discerned, such as genomic mutation or variation in regulatory elements, in transcription factors, or in chromosomal copy number, each of which has the potential to define transcriptional state. Thus, prior studies have had obvious limitations in resolving the critical link between DNA and transcriptional changes. Secondly, while transcript-level information is frequently employed for molecular subtyping of a tumor[38,39], pharmacological decisions are primarily driven by genomic variation, due to technical and informatics challenges with ascertainment by transcriptional status[40]. This may, in part, explain why tumor DNA molecular data provides imperfect prediction of treatment sensitivity.

Coupling single-cell genomic and transcriptomic information has been hitherto limited due to technical challenges of integrating the RNA and DNA amplification steps. Additionally, in instances where this incompatibility has been overcome, existing methodologies for the amplification of single cell genomes have been employed and thus the shortcomings of incomplete genome coverage, poor coverage uniformity, and less optimal allelic balance have accompanied these joint RNA/DNA protocols. G&Tseq, for example, empowered researchers with transcriptional data of single cells paired with multiple displacement amplification for DNA level information[5]. This has facilitated multi-omic insights at primarily the transcriptome + copy number alteration level due to the incomplete genome amplification inherent with MDA or PicoPLEX, precluding SNV analysis. We designed the ResolveOME chemistry to overcome this limitation by unifying primary template-directed amplification with RNA sequencing in single cells and show its utility by cataloging putative regulatory SNVs affecting gene expression.

The ability to define cell identity and cell state at the single cell level is one chief strength of ResolveOME. While some FACS strategies may sufficiently stratify cell types within a heterogenous sample, one does not always a priori have this biomarker knowledge, and even in the presence of this knowledge we have observed outlier sorted cells where we fail to detect concordant mRNA levels despite the cells being gated on high levels of the corresponding protein biomarker. Thus, joint RNA/DNA single-cell profiling has enabled us here to spotlight instances of diverse, non-epithelial cell types in our primary breast cancer sample, preventing the false interpretation of a ductal epithelial cell lacking prototypical copy number alteration or key oncogenic missense mutations when in fact the lack of genomic variation is due to the cell type being assayed. When armed with joint genomic and transcriptomic information, cell type tumor heterogeneity manifesting in FACS can now be exploited, for example, to understand the contributions of the genome variation of a monocyte to the interaction of the malignant epithelial cell in the given microenvironment, as opposed to considering the monocytes as contaminating the epithelial population of interest in this instance.

Beyond characterizing cell identity with ResolveOME, we identified a continuum and heterogeneity of cell state within a breast tumor specimen at unprecedented resolution. An intermediate transcriptional profile emerged between that of the EpCAM low single cell cohort and that of the core cohort of EpCAM high epithelial cells. This profile was intriguingly observed in an EpCAM high cell that harbored *PIK3CA* N345K and DCIS-characteristic chromosomal losses, thus having the core genomic changes of the main epithelial cell cohort. Nevertheless, it manifested with a different transcriptional stem-like state—indicating a potential state conversion[41] as well as highlighting inherent transcriptional single-cell heterogeneity even within a relatively small sampling of a singulated tumor sample. It will be crucial to determine the prevalence of this cell state as more cells of this sample are sequenced, as well as to define the diversity of additional novel transcriptional states that may be contributing to the advancement of DCIS to invasive cancer. ResolveOME importantly provides the ability to link these diverse transcriptional cell states to genotype (**Figure 8a**).

A second chief strength of ResolveOME is to provide the attributes of primary template-directed amplification to allow comprehensive genomic assessment vs. the sole ascertainment of a small number of candidate loci or copy number alterations of a broad level of resolution. This enablement of SNV detection with high sensitivity and precision over >95%[1] of the genome opens a new realm of discovery. PTA in the ResolveOME workflow opens up a new source of pharmacological targets with genome-wide data and non-exonic space not possible with existing WGA methodologies with low genomic coverage and uniformity. We were struck by the single nucleotide variation present in our parental vs. quizartinib resistant MOLM-13 cells (6444 differentially prevalent SNVs, **Figure 6, Supplementary Tables 2-3**), which further underscores that, while transcriptional plasticity is dogmatic, it is equally as important to recognize *genome* plasticity observed in this model. Furthermore, while there will be a background of passenger mutation or mutation currently not pharmacologically targetable; we put forth that this diversity must be ultimately ascertained and represent a co-evolution of variants for a functional, biologically relevant phenotypic output. Efforts to estimate intergenic variation at putative functional elements—promoters, enhancers, splicing enhancers—is a frontier and an underappreciated aspect of drug resistance studies. The candidate regulatory single nucleotide variation we identified proximal to differentially expressed genes of interest in our parental vs. resistant cells will require obligate functional characterization, but as the cost of genome sequencing begins to plummet, these data and their associated biological insights will necessarily begin to accumulate. For discovery, dual genome/transcriptome ascertainment from single cells not only expedites the generation of candidate regulatory SNV links to transcript modulation but unveils connections obscured by bulk sequencing data.

Both our engineered model of drug resistance in AML and analysis of a primary DCIS/IDC sample have yielded single nucleotide variation that would be predicted, at the outset, to have a deleterious effect on protein function. Frameshift and stop codon gain mutations observed in the single cell genomes of our samples represented an unbiased starting point for the discovery of novel oncogenic and drug resistance loci beyond ascertainment of known candidate genes. Yet, coupling transcriptional information from the same cell revealed that, for some of these novel genomic variants of purported deleterious effect, the single cells did not express the corresponding transcript—indicating the genomic change was passenger or stochastic in nature and not functional. Understanding this genomic variant "penetrance" in terms of manifesting at the transcriptional level is a fundamental capability of ResolveOME, and in our initial sample sets redirected or nullified multiple hypotheses.

In addition to binary "expressed or not expressed" decisions, dual DNA/RNA information assisted in directing hypotheses of molecular mechanism. *CEBPA*, an enhancer factor[42] significantly upregulated in our quizartinib-resistant single MOLM-13 cohort, resides on Chr. 19q, where four resistant cells harbored 2n to 3n genomic gain of 19q. A parsimonious initial hypothesis is that genomic amplification of 19q contributed to the observed transcript upregulation, however the *CEBPA* transcript upregulation was observed in all resistant cells, and did not show a correlation with the single cells that harbored genomic amplification of 19q (**Figure 7c**). This suggests that an alternative mechanism of epigenetic control was at play for this upregulated gene, perhaps via modulation of a transcription factor or an enhancer-level phenomenon that was purported by the SNV between parental and resistant cells proximal to the *CEBPA* gene. More broadly, while we identified statistically significant associations between ploidy and expression of a specific cohort of genes (**Figure 7d**), we found that there was no such association for most loci. Collectively, these examples illustrate the criticality of paired RNA information when positing mechanisms based on genomic data alone and caution that the "penetrance" of the change needs to be ascertained. Conversely, we have identified important correlations between SNV and the expression of a proximal gene, as with the oncogenic driver MYC (**Figures 8a,c**), highlighting instances whereby DNA and RNA information are likely to be functionally linked.

The enablement of simultaneous genomic and transcriptomic data from the same individual cell vastly increases the complexity of putative mechanisms of drug resistance and oncogenesis. This will only increase as additional "-omic" tiers of layers are added, including ascertainment of extracellular protein expression as the nature of ResolveOME template-switching cDNA chemistry allows for the incorporation of CITE-seq-like[43] oligo-tagged antibodies. These data will be complex, requiring development of novel sophisticated bioinformatics tools. However, we envision mechanistic insights analogous to those presented here to accumulate from the research community having the newfound ability to accurately assess single nucleotide genomic variation in conjunction with transcriptional profiles—aiding discovery efforts to generate a new wealth and generation of pharmacological targets.

METHODS

Cell Culture

NA12878 cells (CEPH/Utah Pedigree 1463) were obtained from the Coriell Institute for Medical Research (Camden, NJ). Cells were maintained in RPMI 1640 (Gibco 11875-093) supplemented with 15% FBS and penicillin/streptomycin, and sub-cultured every 2-3 days while maintaining a density range of 1.0-3.0 E6/ml.

MOLM-13 acute myeloid leukemia cells harboring heterozygous FLT3 internal tandem duplication (ITD) were obtained from the DSMZ-German Collection of Microorganisms and Cell Cultures (ACC 554). Cells were maintained in RPMI 1640 (Gibco 11875-093) supplemented with 10% FBS and penicillin/streptomycin, and sub-cultured every 2-3 days while maintaining a density range of 2.5 E5 – 1.5 E6 cells/ml. For generation of the quizartinib-resistant MOLM-13 line, cells were continually treated with 2 nM quizartinib (Selleckchem AC220) or DMSO vehicle control for matched parental control line and drug replenished at each subculturing until emergence of resistant clones at 5 weeks duration in culture. Genomic DNA (Zymo Research Quick-DNA Microprep w\Plus Kit, D3020) or total RNA (Qiagen RNeasy Plus Kit, 74034) was isolated from quizartinib-resistant and matched parental MOLM-13 cells at time of FACS sorting to generate bulk sequencing control libraries for comparison to single cell datasets and for quantitative PCR template.

ResolveOME Workflow

ResolveOME begins with template-switching-based RNA-Seq chemistry to generate biotin-dT-primed, first strand cDNA followed by termination of the reaction and nuclear lysis, at which point primary template-directed amplification proceeds. The mRNA-derived cDNA is affinity purified with streptavidin beads from the combined pool of cDNA and amplified genome. cDNAs are then further purified with subsequent streptavidin bead washes of two stringencies and on-bead pre-amplification of the first-strand cDNA to yield double-stranded cDNA. In parallel, the PTA fraction from the same cell containing genome amplification products, separated from the cDNA, is purified. The separate and distinct fractions of pre-amplified mRNA cDNA and genome-derived DNA amplification fractions undergo SPRI cleanup prior to NGS library are generation.

Karyotyping

MOLM-13 cells were analyzed within 2 weeks of thaw (KaryoLogic, Inc, Durham, NC) with a workflow for complex hyperdiploid karyotypes using 25 metaphase spreads. Live cultures were delivered to the service provider on-site and cultures recovered in 5% CO2 37C incubators on-site for one week prior to metaphase spread creation.

FACS

Prior to FACS, cell lines were first counted and assessed for overall viability by trypan blue staining using a Countess II FL instrument (ThermoFisher Scientific) or by acridine orange + propidium iodide with a Luna FL instrument (Logos Biosystems). Cell line cultures put forth to the FACS protocol exhibited >90% viability.

MOLM-13

For single cell analysis, ~2.0E6 MOLM-13 quizartinib-resistant or matched parental cells were rinsed twice in staining buffer (0.2 µm filtered Dulbecco's Phosphate Buffered Saline lacking calcium and magnesium (Gibco 14190) supplemented with 2% FBS) and kept on ice until BD FACSAria III sorting at the UNC School of Medicine Flow Cytometry

Core Facility. Following Calcein AM (BioLegend 425201), propidium iodide (Millipore Sigma P4864) and DAPI staining, singlet (FSC-A / FSH-H, SSC-A / SSC-W) and live cell (DAPI/PI negative, top 70% Calcein-AM positive) gating was established and single cells were sorted (130 micron nozzle assembly) into low-bind 96 well PCR plates (Eppendorf twin.tec LoBind, semi-skirted, 0030129504) containing ResolveOME Cell Buffer and immediately frozen on dry ice following brief mixing (1400 rpm, 10 sec) and centrifugation.

NA12878

~2.5E6 NA12878 (NA12878/HG001) cells were prepared as above and subjected to Sony SH800 sorting using a 130 micron chip. Singlet (FSC-A / FSC-H, BSC-A / BSC-W) and live-cell (PI negative, top 70% Calcein-AM positive) gating was employed for single cell sorting into low-bind 96 well PCR plates pre-loaded with ResolveOME Cell Buffer as described above.

Primary DCIS/IDC

Tissue for single-cell DCIS/IDC studies was obtained in accordance with the Duke University Medical Center IRB for the clinical trial PRO00034242 “Biologic Characterization of the Breast Cancer Tumor Microenvironment.” Cryo-preserved, singulated cells (~4.2E5) derived from mastectomy tissue were thawed at 37C and centrifuged at 350 x g for 5 min to separate cryo-preservation media. Cells were rinsed once in staining buffer and incubated with 2 µg/ml anti-human CD326 conjugated with AlexaFluor 700 (ThermoFisher 56-9326-42) at 4C in the dark for 1h. Following this, ~8.4E4 cells were reserved for a parallel negative control mock stain lacking any antibody for assessment of background fluorescence levels for viability and EpCAM staining. Then cells were washed 3X with staining buffer with 350 x g 5 min centrifugations in between washes and passed through a 35 micron filter prior to loading for FACS. Singlet (FSC-A / FSC-H, BSC-A / BSC-W) and live-cell (Calcein AM) gating was defined followed by daughter EpCAM high and EpCAM low gates. EpCAM High and Low cells were sorted into the same 96 well plates as described above for to minimize potential batch effects of downstream genomic/transcriptomic amplification.

Quantitative RT-PCR

10 ng of genomic DNA was isolated from a cell collection of quizartinib-resistant or matched parental cells as described above and subjected to a custom Taqman™ genotyping assay, #ANMF9C4 (Invitrogen-Applied Biosystems) using the manufacturer’s suggested conditions for reaction assembly and cycling on a QuantStudio6 instrument. The assay was

designed to distinguish between human N841 and K841 with the C/A nucleotide polymorphism, respectively at the GRCh38 / hg38 coordinate Chr13:28,018,485.

Combined genomic/transcriptomic analysis

Firstly, biotin-conjugated oligo dT primer (Integrated DNA Technologies) was utilized in a template-switching reverse transcription reaction to generate first-strand cDNA from single cells. Primary Template-directed Amplification (PTA) with ResolveDNA reagents was performed in succession following reverse transcription. First-strand cDNA was then affinity-purified using ResolveOME streptavidin beads and subjected to two high-salt washes followed by one low-salt wash. 24-cycles of pre-amplification was performed to generate 2nd strand cDNA and RNA sequencing libraries were prepared using the ResolveOME RNA library preparation module. For preparation of PTA libraries, PTA product not bound to streptavidin beads was purified using BioSkryb ResolveDNA beads and ligated to full-length IDT for Illumina TruSeq adapters using the ResolveOME DNA library preparation module. Sizing for both RNA and DNA amplification products was determined by D5000 TapeStation electrophoresis (Agilent Technologies) while library preparation sizing was determined by HS D1000 electrophoresis. Amplification and library yield was assessed by Qubit 3 or Qubit Flex instrumentation (ThermoFisher Scientific).

Sequencing

Low-pass sequencing was first performed on ResolveOME DNA fraction libraries using an Illumina MiniSeq (2.3 pM library flow cell loading concentration) or NextSeq1000 (640pM library flow cell loading concentration), 2X75 targeting >2.0E6 total reads per library. For RNA fraction libraries, 2X75 MiniSeq or NextSeq1000 sequencing targeting on average >1.0E6 reads per library was employed for flexibility for data down-sampling. For joint clustering of ResolveOME DNA and RNA fraction libraries, a 10:1 molar ratio of [DNA arm]:[RNA arm] libraries was employed. Following low-pass sequencing, ResolveOME DNA arm libraries were 2X150 sequenced on an Illumina NovaSeq6000 S4 flow cell targeting 5.5 E8 total reads to provide down-sampling flexibility at either the Vanderbilt Technologies for Advanced Genomics (VANTAGE) core facility or the Duke University Genomics and Computational Biology (GCB) core facility.

Bioinformatics Approaches

Pre-sequencing Quality Control

Single cell libraries were evaluated utilizing an internal pre- sequencing pipeline that leverages low-pass sequencing data to create multiple quality control metrics to assist in evaluating the single-cell libraries readiness for high-throughput sequencing. Notably we retrieved the PreSeq count to estimate library complexity. This pipeline features additional QC metrics for genomic coverage, percent of reads mapping to chimeras, percent of reads aligned to the reference genome, and percent of nucleotides mismatched to the reference genome. Additionally, the pipeline implements MultiQC for supplementary QC metrics including read length, percent of duplicate reads, number of mapped reads, and total number of mapped reads.

Benchmarking RNA-Seq results

To establish overall benchmarking scores of ResolveOME multi-omic amplification approach, quality control was performed pre- and post-sequencing on Human Brain Reference RNA (HBRR), Universal Human Reference RNA (UHRR), and NA12878 B-lymphocyte cells. We considered several metrics: percent mapping, gene detection, dynamic range of expression, and coefficient of variation for measuring DNA leakage, accuracy, and robustness of this methodology. For each cell the total alignments, reads aligned, and genomic feature alignments were quantified using the Qualimap[44] (v2.2.2) platform for reporting QC metrics and bias estimations of whole transcriptome sequencing data. Furthermore, the platform enables detection of outlier cells, relative consistent performance patterns among these cells, and potential batch or other systematic artifacts that are not apparent when evaluating individual cells in isolation. Using metrics produced from Qualimap findings, we computed the percent mapping of total alignments as well as the percent exonic and intergenic of genomic alignments. Thereafter, we defined the number of genes identified, dynamic range, housekeeping gene variability metrics, and observations of expression patterns in housekeeping genes for each reference cell line, using counts per million (CPM) normalized gene expression counts. Gene detected is defined at the number of genes with non-zero counts in each cell. The dynamic range of all expressed genes was then estimated at 10-90 percent. As an estimate of sample dispersions and reproducibility, the percent coefficient of variation (CV) was calculated as a ratio of standard deviation to mean: CV = . We calculated the median absolute deviation (MAD) as a robust measure of variability between housekeeping genes. This is defined as the median of the absolute deviations from the median (m): MAD= median($|x_i$-m$|$).

Secondary Analysis Pipelines

For the DNA-based analyses coming from the genomic fraction of ResolveOME, we leveraged an internal analytics pipeline modified from Sentieon driver-based tools. Initial FASTQ pairs were trimmed against low quality and library artifacts using fastp[45] (v0.20.1) Alignment was performed using BWA (Sentieon-202112), followed by deduplication (locus_collector v202112 / dedup v202112) of identically-aligned reads. Alignment-based QC and coverage determination was (driver_metrics v202010). Copy number calling was performed using ginko[46] (GitHub commit: 892b2e9f851f71a491cade6297f74f09f17acf4c), with a window size of 500kb. Variant calling at the cell level was performed with haplotyper (v202010). Characteristics for all variants was provided for variant quality score recalibration to VARcall, GVCFtyper (v202010). All variant identification and annotations for gene/coding effect were performed using snpEFF/SnpSIFT[47] (5.0e). Further variant-based tertiary analysis used filtered genomic loci with sequencing depths >4 and >1 variant read candidate SNVs. All candidate SNVs were classified according to allele frequencies.

The RNA-Seq pipeline implemented here was used to generate metrics of feature quantification at the transcript and gene-level. Details about the number and length of reads generated is found in Supplemental Table 1 for the DNA arm (a) and RNA arm (b). Unless specified to be down-sampled (using seqtk[48] v1.3), all reads were leveraged for each analysis. To remove low quality sections and sequencing artifacts, fastp was used for all cells' analysis prior to alignment. Alignment of reads was performed with STAR [49](v 2.7.6a) and were compared against transcript reference made from combining Ensembl[50] (release 104) known transcripts and noncoding. Region assignment and counting of aligned reads was performed with HTSeq4949 (v 0.13.5) and Salmon5050 (v1.6.0) for gene-level metrics. Further, we used the pseudo-alignment algorithm implemented in Salmon to perform both transcript-level and gene-level quantification. Matrices of feature expression were constructed using the Bioconductor package tximport.

Tertiary Analysis

Bulk dataset identification

We identified several datasets in the Short Read Archive (SRA) that had bulk NA12878 in mRNA-stranded RNA library preparation methods that most closely resembled our own ResolveOME approach. To handle variation of an individual dataset, we wanted to capture at least 10 datasets that could represent transcriptome coverage of NA12878 and the full list is provided in **Supplementary Data 1.**

Variant evaluation in NA12878 cells

For the NA12878 cells, first we performed joint genotyping across them utilizing the GVCFTyper, VarCal and ApplyVarCal modules from Sentieon. Then, inputting the re-calibrated variants and evaluating the variant quality score log-odds (VQSLOD), we determined the precision and sensitivity of called SNPs by employing the vcfeval module from the RTG tools using as reference the NA12878/HG001 genome v.3.3.2[51] from the Genome in a bottle (GIAB) consortium[52].

Allelic balance in NA12878 cells

Allelic balance for NA12878 cells was calculated using an *ad hoc* developed module based on a series of bcftools commands that extract the *a priori* defined high confident heterozygous sites, reported in GIAB NA12878/HG001 genome v.3.3.2, from all sequenced NA12878 cells. Then, for each cell and for each heterozygous site, variant allele depth is extracted and converted into proportion. For final reporting, heterozygous sites with at least a total depth >1 are used.

RNA arm: Matrix normalization

For MOLM-13 and DCIS cells, their corresponding Salmon-based transcript and gene matrices were normalized across features utilizing the log norm method. Briefly, feature counts for each cell are divided by the total counts for that cell, multiplied by the scale factor (10^4) whose products is finally log2 transformed. These normalized matrices served as input for downstream analysis including, principal component analysis (PCA), differential transcript expression (DTE), differential gene expression (DGE), differential transcript usage (DTU), heatmap reconstruction including unsupervised clustering of cells and transcripts/genes and zero inflated linear models linking transcript expression to CNV and SNVs.

Principal Component Analysis

MOLM-13 and DCIS normalized transcript level and gene level matrices we centered across samples within a feature using the R function scale. Further, principal component analysis was computed using the oh.pca function from the ohchibi R package taking as input the centered normalized matrices.

Differential Expression

We estimated differential transcript expression and differential gene expression leveraging the zero-inflated linear model (ZLM) implemented in the MAST[53] R package taking as input the log normalized feature matrices described above. For the MOLM-13 dataset, we fitted the following model to identify transcripts/genes that had signatures of differential expression across parental and resistant cells: *Transcript/Gene expression ~ Cell Type (Parental/Resistant) + Number of detected features (transcripts/genes) per cell*

For the DCIS dataset, first we performed principal component analysis using the top 500 most highly variable genes across the dataset and then split the cells into three groups using the PCA projection as guidance. We used this three group scheme to discretize, in an unbiased way the cellular heterogeneity within EpCAM High and EpCAM Low treatment. After dividing the cells into three groups we fitted the following ZLM to identify transcripts/genes that had signatures of differential expression across the aforementioned groups: *Transcript/Gene expression ~ Cell Group + Number of detected features (transcripts/genes) per cell*

Cellular typing

Transcriptome-based cellular typing was performed for the DCIS dataset using the R package SingleR[54] utilizing the Human Primary Cell Atlas expression reference dataset deposited in the celldex[54] R package and taking as input the gene level normalized expression salmon-based matrix.

Differential transcript usage

For the MOLM-13 dataset, we performed differential transcript usage as previously described[55] . Briefly, we took the scaledTPM metric output from tximport and reconstructed a matrix of transcript abundances across cells. Next, we modeled the transcript expression using the Dirichlet-multinomial distribution model implemented in the DRIMSeq R package[56].

Linking transcript expression to CNV

For the MOLM-13 dataset, we linked transcript-level variation in expression with changes in locus ploidy utilizing a zero-inflated linear model framework. Briefly, for each quantified transcript, we extracted its ploidy across cells from the Ginkgo-based estimation by employing genomic-coordinate intersection utilizing the GenomicRanges R package[57]. Next, we fitted the following ZLM design utilizing the MAST R package: *Transcript expression ~ Estimated ploidy at a given locus*

Linking transcript expression to genomic polymorphisms

For the MOLM-13 dataset, we linked transcript-level variation in expression with single nucleotide variations across the genome utilizing a zero inflated linear model framework. Briefly, first we paired the genomic coordinates of SNVs with transcripts utilizing genomic-coordinate intersection via the GenomicRanges R package. With respect to the transcript-coordinates, we used the Ensembl reported transcript start and transcript end to define the gene-body of a transcript, in addition we took the 5000 bps upstream of the Ensembl reported transcription start site (TSS) to define potential cis-regulatory regions affecting the that transcript. After defining the corresponding SNV-Transcripts pairs, we constructed a matrix of expression and genotype locus (SNV) across all cells. Finally, utilizing this matrix, we fitted a zero-inflated linear model using the MAST R package with the following design: *Transcript expression ~ Genotype*

We utilized the GSEA-R tool in conjunction with the molecular signatures database (MSigDB) to conduct a systematic examination of enriched gene sets connected to differentially expressed genes across Molm-13 parental and resistant cells as well as significant SNVs. In addition, we utilized the Reactome Pathways database to find relevant pathways among these genes using a default adjusted p-value of 0.10.

Significant Variant Testing

For identification of differential SNV's between MOLM-13 P and R cells, we generated categorical variables for diploidy status and compared with chi-square test. Two-sided p-values less than 0.05 were considered significant. In addition, we fitted a multinomial logistic regression to identify differences in SNV prevalence across the parental and resistant MOLM-13 types. Specifically, for each SNP, we encoded the three states genotype (0/0, 0/1, 1/1) as dependent variable and the MOLM-13 type (parental, resistant) as independent variable. Significance of the model was tested using a Wald Test.

ACKNOWLEDGEMENTS

We thank the patients involved in this study for their gracious gift of tissue samples that made the scientific insights in this manuscript possible.

FIGURE LEGENDS

Figure 1. ResolveOME workflow.

a.) The high-level workflow of enrichment and preparation of simultaneous RNA and DNA from a single cell and **b.**) the amplification yields of ResolveOME**.** The yields of RNA and DNA isolated (in ng) for each cell used in this study. Samples where purification by streptavidin beads was omitted are highlighted in orange.

Figure 2. Performance characteristics of ResolveOME.

a.) ResolveOME (green) and ResolveDNA (orange) allelic balance in control (NA12878) is shown in deciles of observed allele frequency (AF) across known heterozygous positions. Each dot represents the proportion of variants that showed an AF within the bin frequency for a given cell. Barplots with error bars describe general trend for all cell-replicates for each AF bin. Allelic dropouts are called when AF is < 0.1 or > 0.9. **b.**) A cumulative genomic coverage plot for each sample type performed on ResolveOME, showing the proportion of the entire genome covered (y-axis) at a given depth (x-axis). Each dot represents a cell replicate within a dataset and error plots denote the variability of coverage at a given depth. **c**) SNV calling sensitivity (y-axis) and precision (x-axis), with respect to GIAB NA12878 reference dataset, across the two chemistries

(ResolveDNA, salmon, ResolveOME, turquoise) are shown with both axes having a minimum range of 0.9 and 0.99, respectively.

Figure 3. ResolveOME transcriptome performance.

a.) Summarized coverage plots for all detected transcripts across the full-length chemistry (top). X axis is a normalized fraction of a transcript from 5' to 3', breaking regions into mean depth per percentile of transcript and y-axis are counts. Distribution of counts across coding sequence of two known housekeeping genes: *GAPDH* and *ACTB* (bottom). **b.**) The proportion (averaged across all biosamples of a group) of aligned reads that matches a specific transcript feature or RNA species is reported for each dataset. Features and proportions were derived from Qualimap summarizations of our transcriptome definition file. NA12878 cells were leveraged except for the MOLM-13/DCIS plots. Bulk data was pulled from online repository to serve as reference from typical RNA-Seq. **c.**) Various RNA quality control metrics are displayed for the UHRR and HBRR RNA controls alongside the NA12878 controls used in this study. Clockwise from the top left, the distribution of reads assigned to transcriptome, coding region features, unique genes detected, ranges of counts per million (CPM) and the median absolute deviation (MAD) of common housekeeping genes. **d.**) Expressed protein-coding genes detected with ResolveOME chemistry compared to bulk preparation with the same workflow. Number of uniquely expressed genes across a diversity of cell line models and a primary DCIS patient sample. All sample sets were down-sampled to 75,000 reads.

Figure 4. Copy number profiles of MOLM-13 cells.

a.) Copy number alterations of individual MOLM-13 cells (rows) from parental (turquoise) and resistant (salmon) cells using a bin size of 500kb with Ginkgo. Dendrogram was generated based on distance of each bin's average fold change from 2N. **b.**): Representative metaphase spread of 25 total karyotypic spreads. Red circles denote abnormally amplified chromosomes.

Figure 5. FLT3 mutational characterization of MOLM-13 cells

a.) Genome views showing detection of mutual *FLT3* ITD mutation in parental and quizartinib-resistant single cells. **b.**) Genome views of *FLT3* secondary mutation N841K exclusively in quizartinib-resistant cells. **c.**): qRT-PCR detection of mutant *FLT3* K841 in treatment-naïve parental cells. Cycling traces of *FLT3* N841 (blue) and K841 (red) in MOLM-13 parental and quizartinib-resistant cells are shown.

Figure 6. Differential parental and resistant MOLM-13 genotypes.

Heatmap of SNVs showing statistically significant ($p < 0.05$ by multinomial logistic regression) genotype prevalence across the MOLM-13 parental and resistant cells. Columns represent cells and rows SNV ids. Color within the tiles represent the called genotypes. Both rows and columns were subjected to unsupervised hierarchical clustering.

Figure 7. Distinguishing genotypic and transcriptomic characteristics between MOLM-13 parental and resistant cells.

a.) Scatterplot showing the principal coordinate projection (PCA) of 28,134 SNVs that exhibited statistically significant (chi-square test, $p < 0.05$) differential prevalence across the two MOLM-13 cohorts, parental (turquoise) and resistant (salmon). b.) Clustering of differentially-expressed genes in MOLM-13 model of drug resistance. Parental single cells (turquoise) and quizartinib-resistant (salmon) single cells comprise columns; Gene Symbol/Ensembl transcript ID comprise rows. Biotype and FDR is presented to the right of the heat map; red line indicates $q < 0.1$. c.) CEBPA/B transcript upregulation in single quizartinib-resistant MOLM-13 cells. Each row corresponds to a separate MOLM-13 cell. Resistant cells that also harbor 19q gains are also shown. d.): Heatmap with transcripts in the y-axis that show a statistical (ZLM $p < 0.01$) association with ploidy level across all cells in the MOLM-13 dataset. Color of the tiles represents the average standardized expression value at a given ploidy level. The right panel shown the output of the ZLM model testing the expression given the ploidy. Red line indicates the $p < 0.05$ cutoff of the model. Bars are colored based on the -log10 p-value of the ZLM model testing transcriptional differences between parental and resistant cells. e.) Example of differential transcript utilization (DTU) between MOLM-13 parental and drug-resistant single cells.

Figure 8. Correlation of genotype and expression in parental and resistant MOLM-13 cells.

a.) Bubble plot showing SNV-transcript expression associations ($p < 0.05$). Top: SNVs within 5000 bases of transcriptional start site. Candidate SNVs are shown in the y-axis and genotypes in the x-axis. Size of the circle denotes the genotype prevalence of the variant in the MOLM-13 cell type set (parental or resistant). Colors of points denotes the standardized mean expression level of the transcript in the set. Lateral bars represent significance of the model testing the association between transcript expression and genotype. Red line indicates the $p < 0.1$ cutoff of the model. Bars are colored based on the -log10 p-value of the ZLM model testing transcriptional differences between parental and resistant cells. *PABPC4* and *MYC* are highlighted in yellow. *CEBPA* SNVs were too distal (>5 kb) from transcriptional start site for significance in this plotting. b.) Parental/quizartinib-resistant SNVs proximal to *CEBPA* genomic locus. Stars denote mutation locations. Resistant cells show variant in 60% of cells compared to 11% in the parental line variant 'chr19:33,333,734 – delA' (middle star). For 'chr19:33,361,973 – insA' we observed no mutations in the parental cells and in 50% in quizartinib-resistant cells. **c.**) Intronic SNV of *MYC* gene 'chr8:127,739,932 G>A' correlated with increased expression in drug-resistant MOLM-13 cells. **d**.) Putative promoter variants in *PABPC4* 'chr1:39,579,411 T>G'

& 'chr1:39,579,413 T>G' were found in half of the resistant cells only and also associated with differential expression between MOLM-13 parental and resistant cells.

Figure 9. Copy number analysis of DCIS cells.
Single-cell copy number alterations in primary DCIS/IDC EpCAM cohorts. Status of EpCAM presented for EpCAM High (yellow) and Low (turquoise). Two distinct classes of chromosomal loss are observed in EpCAM high (yellow) cells: 1) combined 11q, 13q, 16q/17p loss and 2) combined 13q and 16q/17p loss. Additionally, 13p gain was identified in 10/20 EpCAM high cells, while Chr. X gain encompassing the centromere and flanking P & Q segments was noted in 3 single cells.

Figure 10. Expression analysis of primary breast cancer single cells.

a.) Principal component analysis of EpCAM high (circles) and EpCAM low (diamonds) primary DCIS/IDC transcriptomes where cells are colored based on the number of detected transcripts **b.**) PAM50 gene expression stratification of EpCAM high and EpCAM low DCIS/IDC transcriptomes. **c.**) Unsupervised clustering yields six primary blocks of differential gene expression between EpCAM high and EpCAM low clades. Average ploidy, *PIK3CA* genotypic status (green=N345 wildtype, pink = K345 heterozygous mutant), and cellular identity call are shown for each single cell (column). Gene biotype and FDR is presented for each transcript (row). **d.**) Prediction of DCIS cell identity/state using Human Cell Atlas data. Heat map showing identity score of diverse cell types (rows) for EpCAM High and EpCAM Low single cells (columns) that were used to identify cell annotations. **e.)** Overlay of cellular annotation for principal component analysis of DCIS cells**.** EpCAM high (circles) and EpCAM low (diamonds) single cell transcriptomes, leveraging isoform counts with overlay of cell identity/state (colors).

SUPPLEMENTARY DATA

Table S1: Sequencing metrics for ResolveOME genomic and transcriptomic libraries.
a.) PreSeq estimation of library complexity is shown along with mitochondrial read percentage as a proxy for ascertainment of efficient cell lysis and cell health of ResolveOME DNA material. **b.**) RNA QC values for matched cells from the RNA feature of ResolveOME.

Table S2: Significant coding variants in quizartinib-resistant vs. parental MOLM-13 cells.
Positions that showed genotype association (Chi-Sq p-value) to parental vs. drug resistant are shown, but filtered for positions within or adjacent to coding sequence. Adjusted p-values are provided but were not filtered. File sorted for p-value and then by chromosome position. Cells were grouped according to genotype along with gene and transcript-level annotation is shown.

Table S3: Significant intergenic variants in quizartinib-resistant vs. parental MOLM-13 cells.

Positions that showed genotype association (Chi-Sq p-value) to parental vs. drug resistant are shown, but filtered for positions outside of coding sequence. Adjusted p-values are provided but were not filtered. File sorted for p-value and then by chromosome position. Cells were grouped according to genotype along with gene and transcript-level annotation is shown.

Table S4: MOLM-13 genes with expression levels correlated with ploidy.

A genome-ordered list of transcripts with shared expression magnitudes within known copy numbers from Figure 4a. Associated Gene IDs and Gene Symbols are provided.

Table S5: Variants associated with DCIS EpCAM High vs Low.

Positions that showed genotype association (Chi-Sq p-value) of EpCAM high vs low are shown, but filtered for positions outside of coding sequence. Adjusted p-values are provided but were not filtered. File sorted for p-value and then by chromosome position. Cells were grouped according to genotype along with gene and transcript-level annotation is shown.

Figure S1. Relative growth rates of parental and quizartinib-resistant MOLM-13 cells.

Counts of cells over culture days after introduction of varying concentrations of quizartinib.

Figure S2. Missense variants in parental vs. resistant MOLM-13 cells.

Variants (rows) identified as significantly associated through logistic regression with drug resistance are displayed, along with individual genotypes (0/0=homozygous reference, 0/1=heterozygous, 1/1=homozygous alternate, NA=not determined). Single cells (columns) are presented for parental (left) or resistant (right) cohorts. P value is shown along the right-hand side.

Figure S3. Model of transcriptional bypass signaling through AXL upon FLT3 inhibition.

Schematic illustrating that upon FLT3 inhibition by quizartinib, GAS6, the ligand for the receptor tyrosine kinase AXL, is upregulated in resistant MOLM-13 cells to drive growth and survival through PI3 kinase and AKT signaling, respectively.

Figure S4. Variants associated with DCIS expression groups.

Variants (rows) identified as significantly associated through logistic regression with expression groups within EpCAM-H DCIS cells are shown, along with individual genotypes are shown (0/1=heterozygous, 1/1=homozygous alternate, NA=not determined). P value is shown along the right-hand side.

REFERENCES

1. Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc Natl Acad Sci U S A* **118**, e2024176118 (2021).

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

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

4. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33**, 285–289 (2015).

5. Macaulay, I. C. *et al.* Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat Protoc* **11**, 2081–2103 (2016).

6. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261–5266 (2002).

7. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).

8. Zhu, Z. *et al.* Genome profiles of pathologist-defined cell clusters by multiregional LCM and G&T-seq in one triple-negative breast cancer patient. *Cell Rep Med* **2**, 100404 (2021).

9. Backhaus, A. E. *et al.* High expression of the MADS-box gene VRT2 increases the number of rudimentary basal spikelets in wheat. *Plant Physiol* kiac156 (2022) doi:10.1093/plphys/kiac156.

10. Ding, J. *et al.* Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat Commun* **6**, 8554 (2015).

11. Griffith, O. L. *et al.* The prognostic effects of somatic mutations in ER-positive breast cancer. *Nat Commun* **9**, 3476 (2018).

12. DiNardo, C. D. & Cortes, J. E. Mutations in AML: prognostic and therapeutic implications. *Hematology Am Soc Hematol Educ Program* **2016**, 348–355 (2016).

13. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**, 561–566 (2019).

14. Verwilt, J. *et al.* When DNA gets in the way: A cautionary note for DNA contamination in extracellular RNA-seq studies. *Proc Natl Acad Sci U S A* **117**, 18934–18936 (2020).

15. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nat Methods* **10**, 325–327 (2013).

16. Luquette LJ, Miller MB, Zhou Z, Bohrson CL, Galor A, Lodato MA, Gawad C, West J, Walsh CA, Park PJ. Ultraspecific somatic SNV and indel detection in single neurons using primary template-directed amplification.

17. Caracausi, M. *et al.* Systematic identification of human housekeeping genes possibly useful as references in gene expression studies. *Mol Med Rep* **16**, 2397–2410 (2017).

18. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).

19. Matsuo, Y. *et al.* Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins(11;9)(q23;p22p23). *Leukemia* **11**, 1469–1477 (1997).

20. Yohe, S. Molecular Genetic Markers in Acute Myeloid Leukemia. *J Clin Med* **4**, 460–478 (2015).

21. Peretz, C. A. C. *et al.* Single-cell DNA sequencing reveals complex mechanisms of resistance to quizartinib. *Blood Adv* **5**, 1437–1441 (2021).

22. Matsuno, N., Nanri, T., Kawakita, T., Mitsuya, H. & Asou, N. A novel FLT3 activation loop mutation N841K in acute myeloblastic leukemia. *Leukemia* **19**, 480–481 (2005).

23. Beghini, A., Magnani, I., Ripamonti, C. B. & Larizza, L. Amplification of a novel c-Kit activating mutation Asn(822)-Lys in the Kasumi-1 cell line: a t(8;21)-Kit mutant model for acute myeloid leukemia. *Hematol J* **3**, 157–163 (2002).

24. DiNardo, C. D. & Cortes, J. E. Mutations in AML: prognostic and therapeutic implications. *Hematology Am Soc Hematol Educ Program* **2016**, 348–355 (2016).

25. Park, I.-K. *et al.* Receptor tyrosine kinase Axl is required for resistance of leukemic cells to FLT3-targeted therapy in acute myeloid leukemia. *Leukemia* **29**, 2382–2389 (2015).

26. Park, I.-K. *et al.* Inhibition of the receptor tyrosine kinase Axl impedes activation of the FLT3 internal tandem duplication in human acute myeloid leukemia: implications for Axl as a potential therapeutic target. *Blood* **121**, 2064–2073 (2013).

27. Zdżalik-Bielecka, D. *et al.* The GAS6-AXL signaling pathway triggers actin remodeling that drives membrane ruffling, macropinocytosis, and cancer-cell invasion. *Proc Natl Acad Sci U S A* **118**, e2024596118 (2021).

28. Abu-Thuraia, A. *et al.* Axl phosphorylates Elmo scaffold proteins to promote Rac activation and cell invasion. *Mol Cell Biol* **35**, 76–87 (2015).

29. Fasan, A. *et al.* The role of different genetic subtypes of CEBPA mutated AML. *Leukemia* **28**, 794–803 (2014).

30. Preudhomme, C. *et al.* Favorable prognostic significance of CEBPA mutations in patients with de novo acute myeloid leukemia: a study from the Acute Leukemia French Association (ALFA). *Blood* **100**, 2717–2723 (2002).

31. Pabst, T. *et al.* Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia. *Nat Genet* **27**, 263–270 (2001).

32. Gorringe, K. L. *et al.* Copy number analysis of ductal carcinoma in situ with and without recurrence. *Mod Pathol* **28**, 1174–1184 (2015).

33. Martínez-Sáez, O. *et al.* Frequency and spectrum of PIK3CA somatic mutations in breast cancer. *Breast Cancer Res* **22**, 45 (2020).

34. Gymnopoulos, M., Elsliger, M.-A. & Vogt, P. K. Rare cancer-specific mutations in PIK3CA show gain of function. *Proc Natl Acad Sci U S A* **104**, 5569–5574 (2007).

35. Huang, C.-H. *et al.* The structure of a human p110alpha/p85alpha complex elucidates the effects of oncogenic PI3Kalpha mutations. *Science* **318**, 1744–1748 (2007).

36. Gkeka, P. *et al.* Investigating the structure and dynamics of the PIK3CA wild-type and H1047R oncogenic mutant. *PLoS Comput Biol* **10**, e1003895 (2014).

37. Griffith, O. L. *et al.* The prognostic effects of somatic mutations in ER-positive breast cancer. *Nat Commun* **9**, 3476 (2018).

38. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167 (2009).

39. Sparano, J. A. *et al.* Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N Engl J Med* **379**, 111–121 (2018).

40. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* **17**, 257–271 (2016).

41. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet* **17**, 693–703 (2016).

42. Repele, A., Krueger, S., Bhattacharyya, T., Tuineau, M. Y. & Manu, null. The regulatory control of Cebpa enhancers and silencers in the myeloid and red-blood cell lineages. *PLoS One* **14**, e0217580 (2019).

43. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865–868 (2017).

44. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

45. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

46. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* **12**, 1058–1060 (2015).

47. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

48. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **11**, e0163962 (2016).

49. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

50. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Research* **49**, D884–D891 (2021).

51. Cleary, J. G. *et al.* Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. 023754 (2015) doi:10.1101/023754.

52. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**, 160025 (2016).

53. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 278 (2015).

54. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).

55. Love, M. I., Soneson, C. & Patro, R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. (2018) doi:10.12688/f1000research.15398.2.

56. Nowicka, M. & Robinson, M. D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res* **5**, 1356 (2016).

57. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9**, e1003118 (2013).

Figure 1

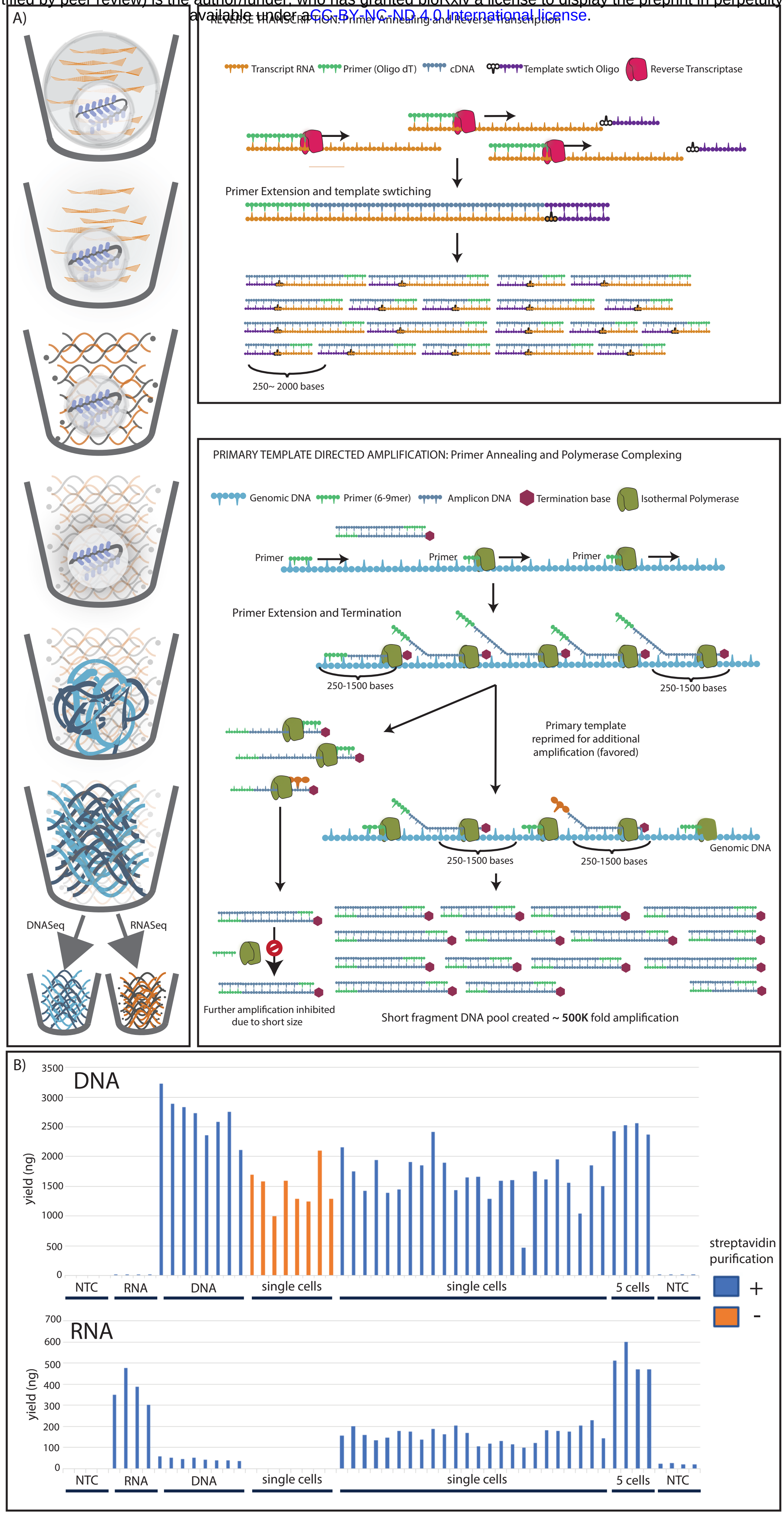

A)
REVERSE TRANSCRIPTION: Primer Annealing and Reverse Transcription
Transcript RNA
Primer (Oligo dT)
cDNA
Template swtich Oligo
Reverse Transcriptase
Primer Extension and template swtiching
250~ 2000 bases
PRIMARY TEMPLATE DIRECTED AMPLIFICATION: Primer Annealing and Polymerase Complexing
Genomic DNA
Primer (6-9mer)
Amplicon DNA
Termination base
Isothermal Polymerase
Primer
Primer
Primer
Primer Extension and Termination
250-1500 bases
250-1500 bases
Primary template reprimed for additional amplification (favored)
250-1500 bases
250-1500 bases
Genomic DNA
Further amplification inhibited due to short size
Short fragment DNA pool created ~ 500K fold amplification
DNASeq
RNASeq
B)
DNA
yield (ng)
3500
3000
2500
2000
1500
1000
500
0
NTC
RNA
DNA
single cells
single cells
5 cells
NTC
streptavidin purification
+
-
RNA
yield (ng)
700
600
500
400
300
200
100
0
NTC
RNA
DNA
single cells
single cells
5 cells
NTC

Figure 2

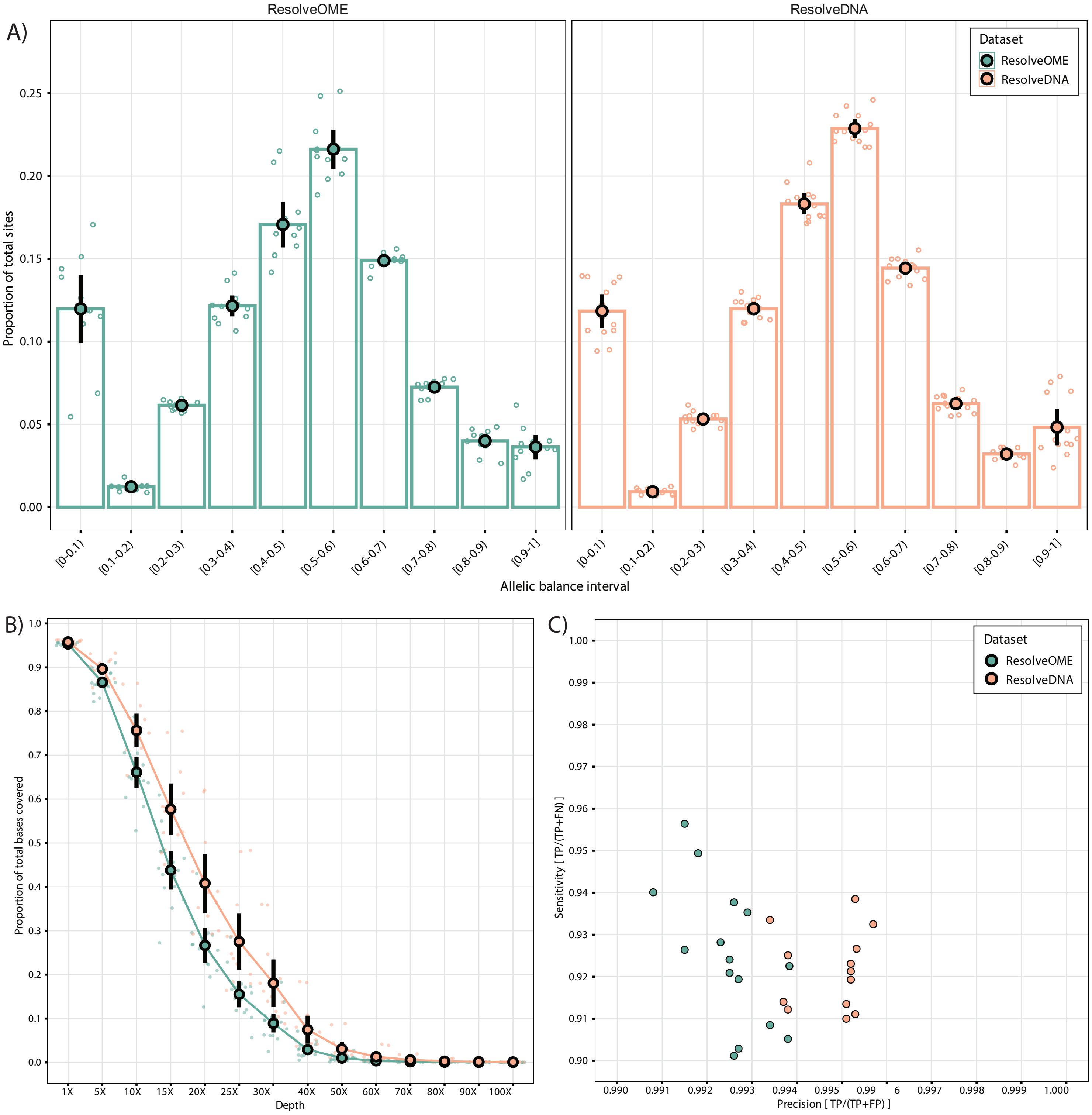
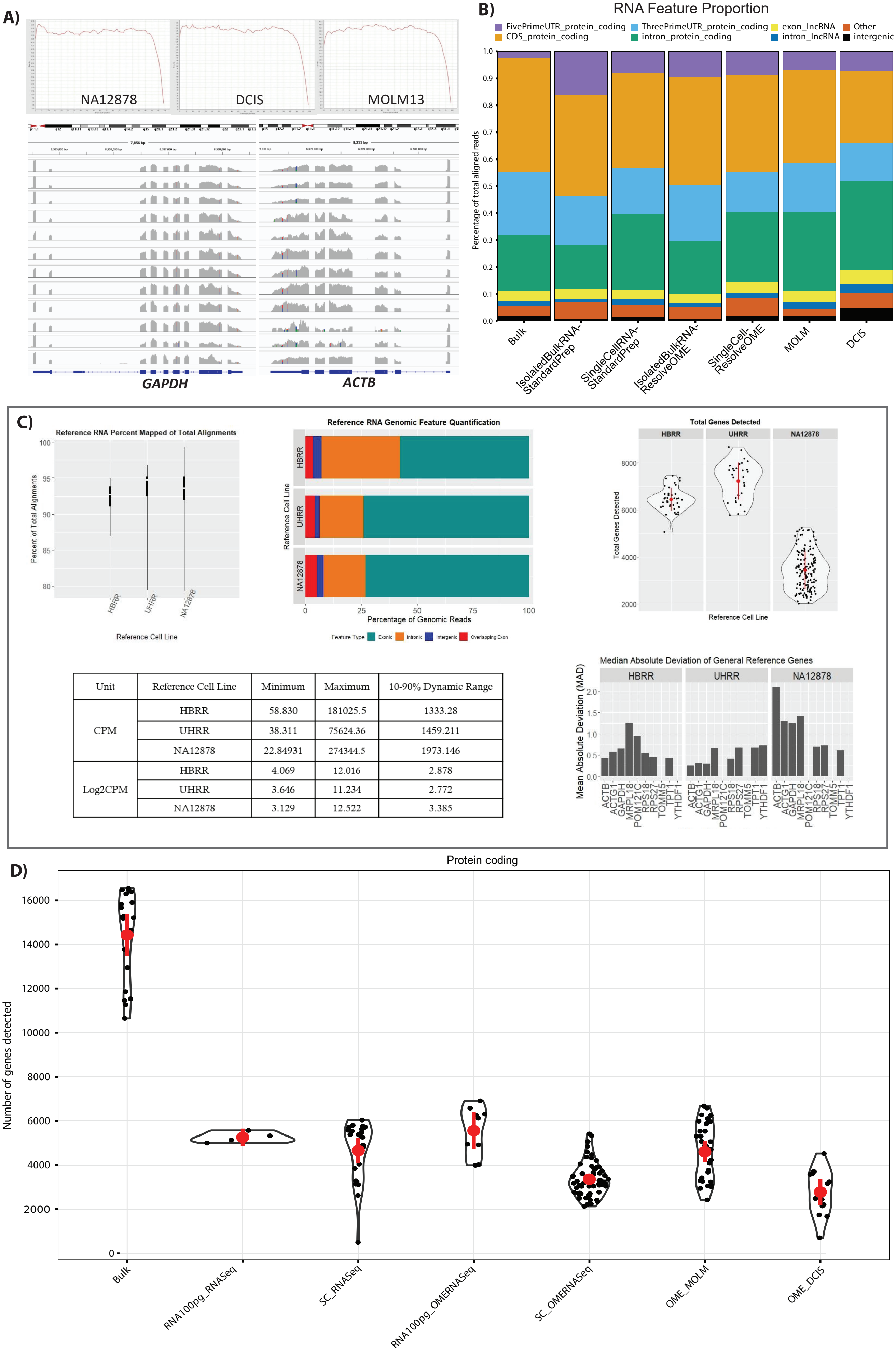

A)
ResolveOME
ResolveDNA
Dataset
ResolveOME
ResolveDNA
Proportion of total sites
0.00
0.05
0.10
0.15
0.20
0.25
[0-0.1)
[0.1-0.2)
[0.2-0.3)
[0.3-0.4)
[0.4-0.5)
[0.5-0.6)
[0.6-0.7)
[0.7-0.8)
[0.8-0.9)
[0.9-1]
Allelic balance interval
B)
Proportion of total bases covered
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1X
5X
10X
15X
20X
25X
30X
40X
50X
60X
70X
80X
90X
100X
Depth
C)
Dataset
ResolveOME
ResolveDNA
Sensitivity [TP/(TP+FN)]
0.90
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1.00
0.990
0.991
0.992
0.993
0.994
0.995
0.99
6
0.997
0.998
0.999
1.000
Precision [TP/(TP+FP)]

Figure 3



Unit	Reference Cell Line	Minimum	Maximum	10-90% Dynamic Range
CPM	HBRR	58.830	181025.5	1333.28
	UHRR	38.311	75624.36	1459.211
	NA12878	22.84931	274344.5	1973.146
Log2CPM	HBRR	4.069	12.016	2.878
	UHRR	3.646	11.234	2.772
	NA12878	3.129	12.522	3.385

Figure 4

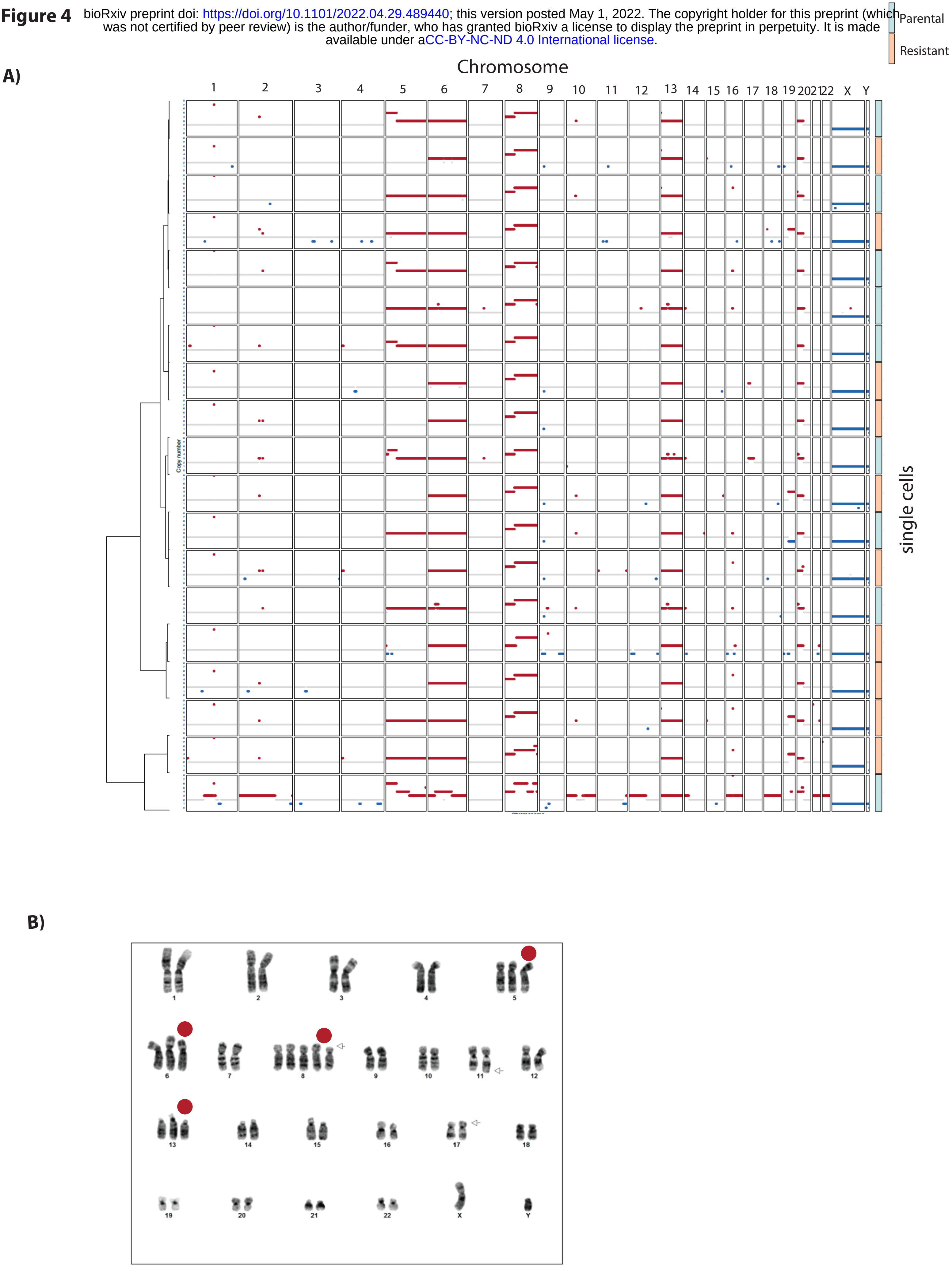

Parental
Resistant
A)
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Copy number
single cells
B)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

Figure 5

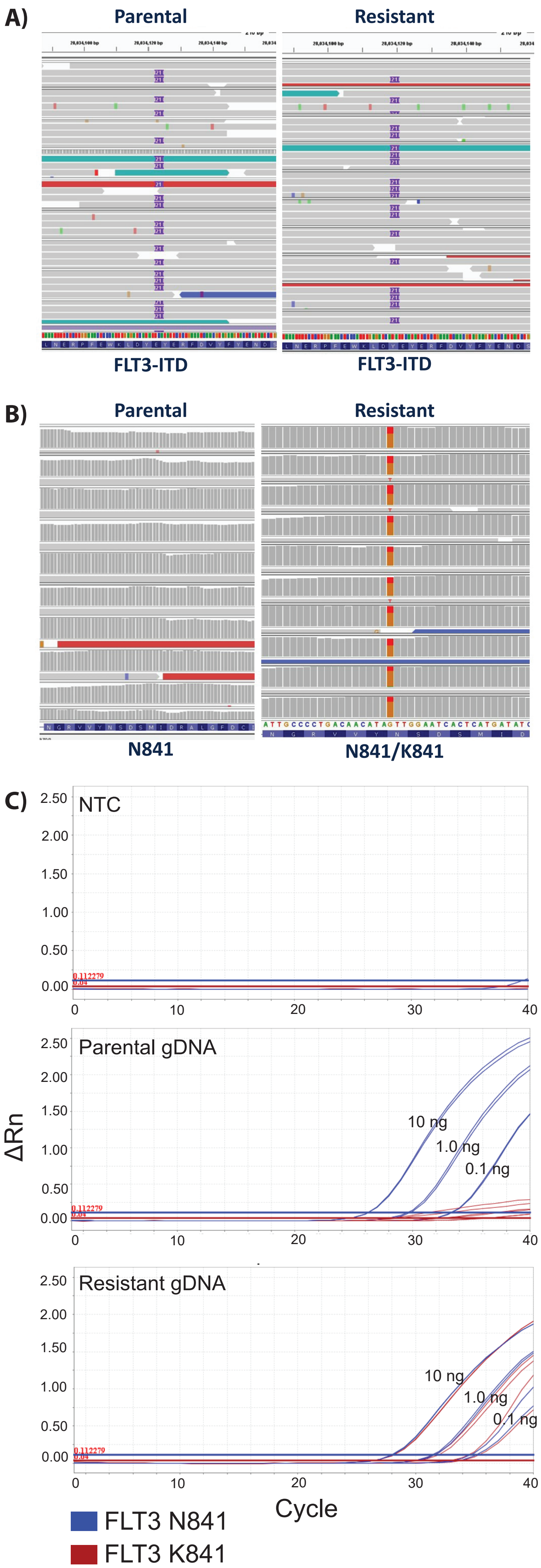

A)
Parental
Resistant
FLT3-ITD
FLT3-ITD
B)
Parental
Resistant
N841
N841/K841
C)
NTC
Parental gDNA
Resistant gDNA
2.50
2.00
1.50
1.00
0.50
0.00
0.112279
0.04
0
10
20
30
40
10 ng
1.0 ng
0.1 ng
ΔRn
Cycle
FLT3 N841
FLT3 K841

Figure 6

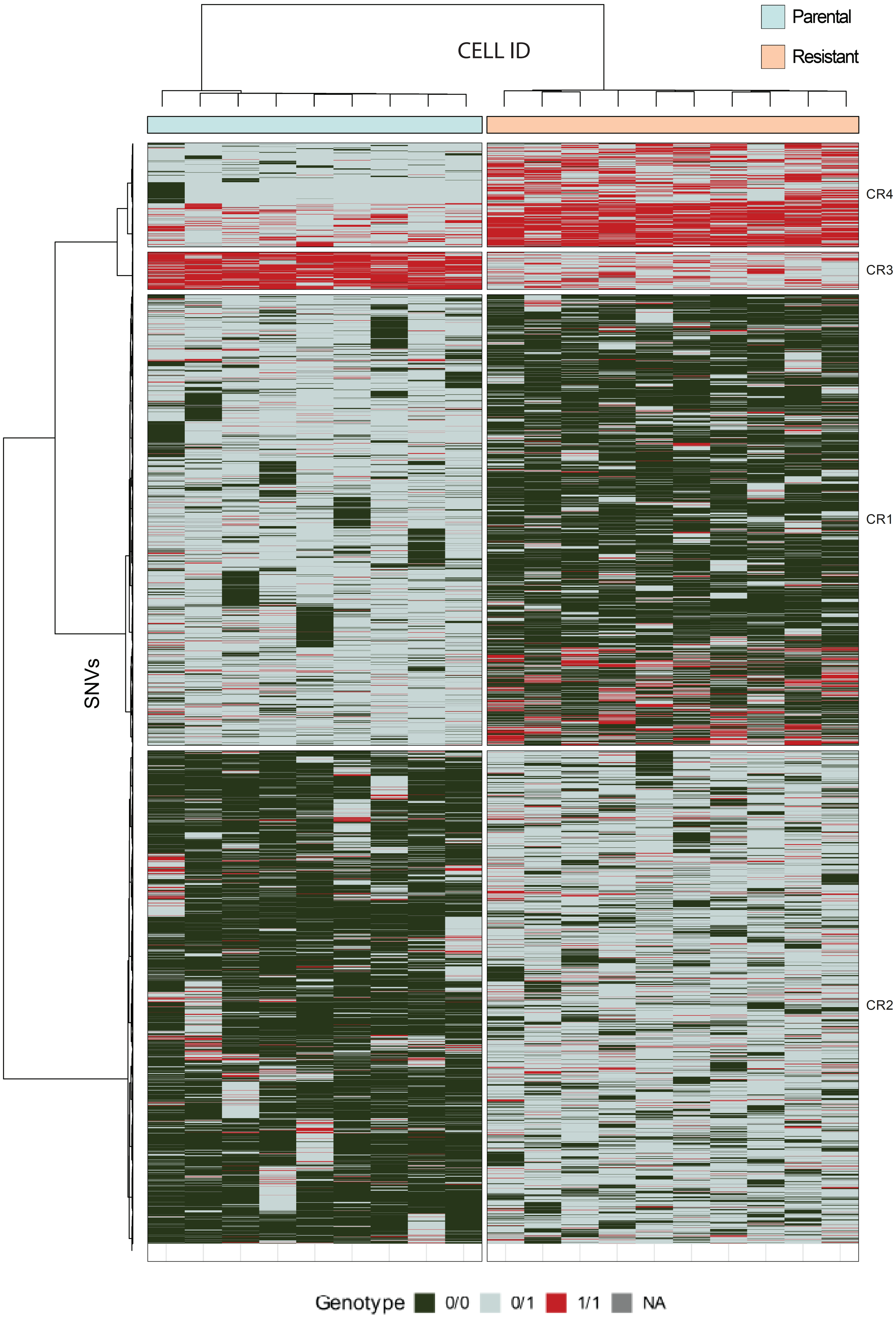

CELL ID
Parental
Resistant
SNVs
CR4
CR3
CR1
CR2
Genotype
0/0
0/1
1/1
NA

Figure 7

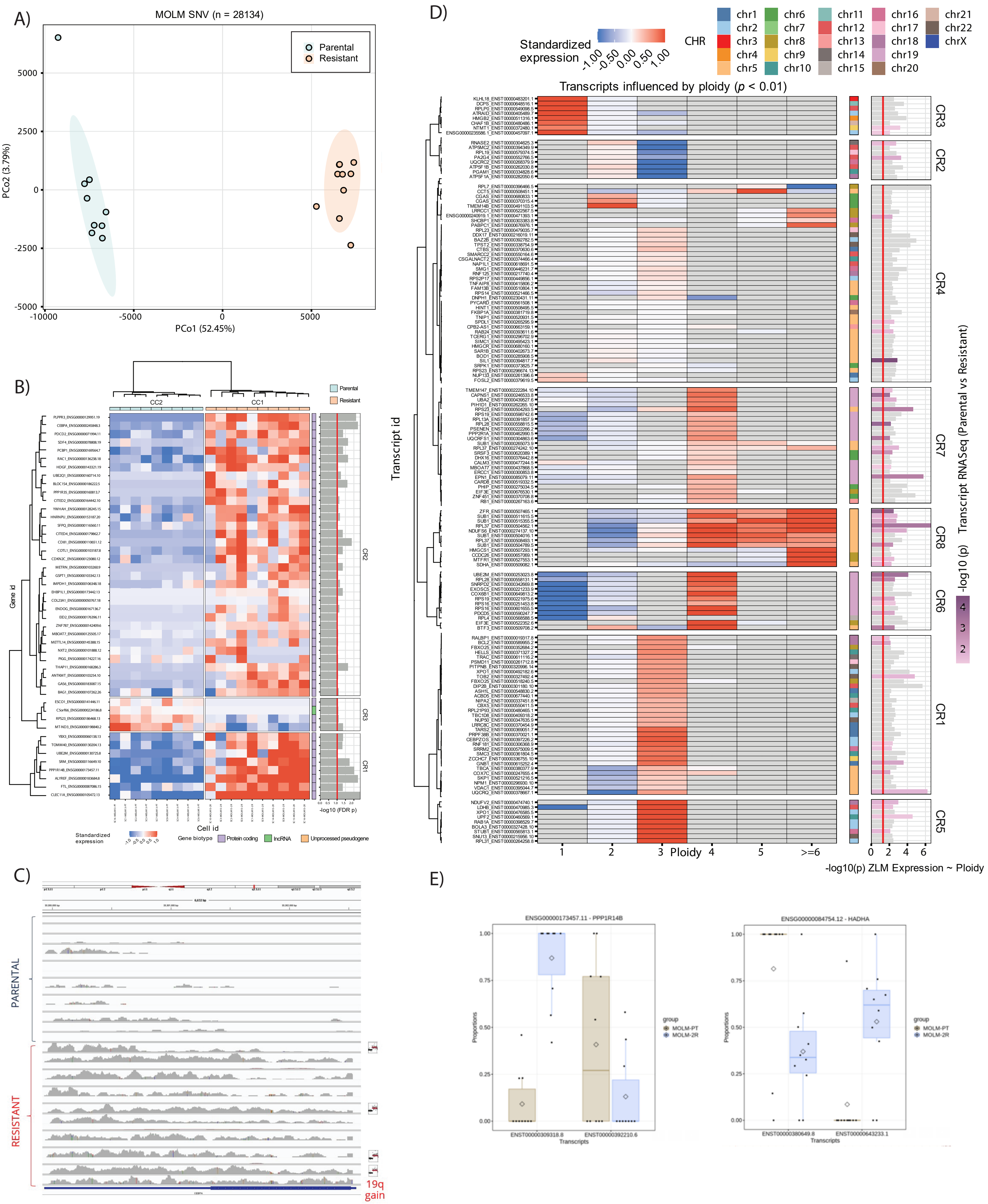

A)
MOLM SNV (n = 28134)
Parental
Resistant
PCo2 (3.79%)
PCo1 (52.45%)
B)
CC2
CC1
Parental
Resistant
Gene id
Cell id
CR2
CR3
CR1
-log10 (FDR p)
Standardized expression
Gene biotype
Protein coding
lncRNA
Unprocessed pseudogene
C)
PARENTAL
RESISTANT
19q gain
D)
Standardized expression
CHR
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX
Transcripts influenced by ploidy ($p < 0.01$)
Transcript id
CR3
CR2
CR4
CR7
CR8
CR6
CR1
CR5
Ploidy
1 2 3 4 5 >=6
Transcript RNASeq (Parental vs Resistant)
-log10 (p)
-log10(p) ZLM Expression ~ Ploidy
E)
ENSG00000173457.11 - PPP1R14B
ENSG00000084754.12 - HADHA
Proportions
Transcripts
ENST00000309318.8
ENST00000392210.6
ENST00000380649.8
ENST00000643233.1
group
MOLM-PT
MOLM-2R

Figure 8

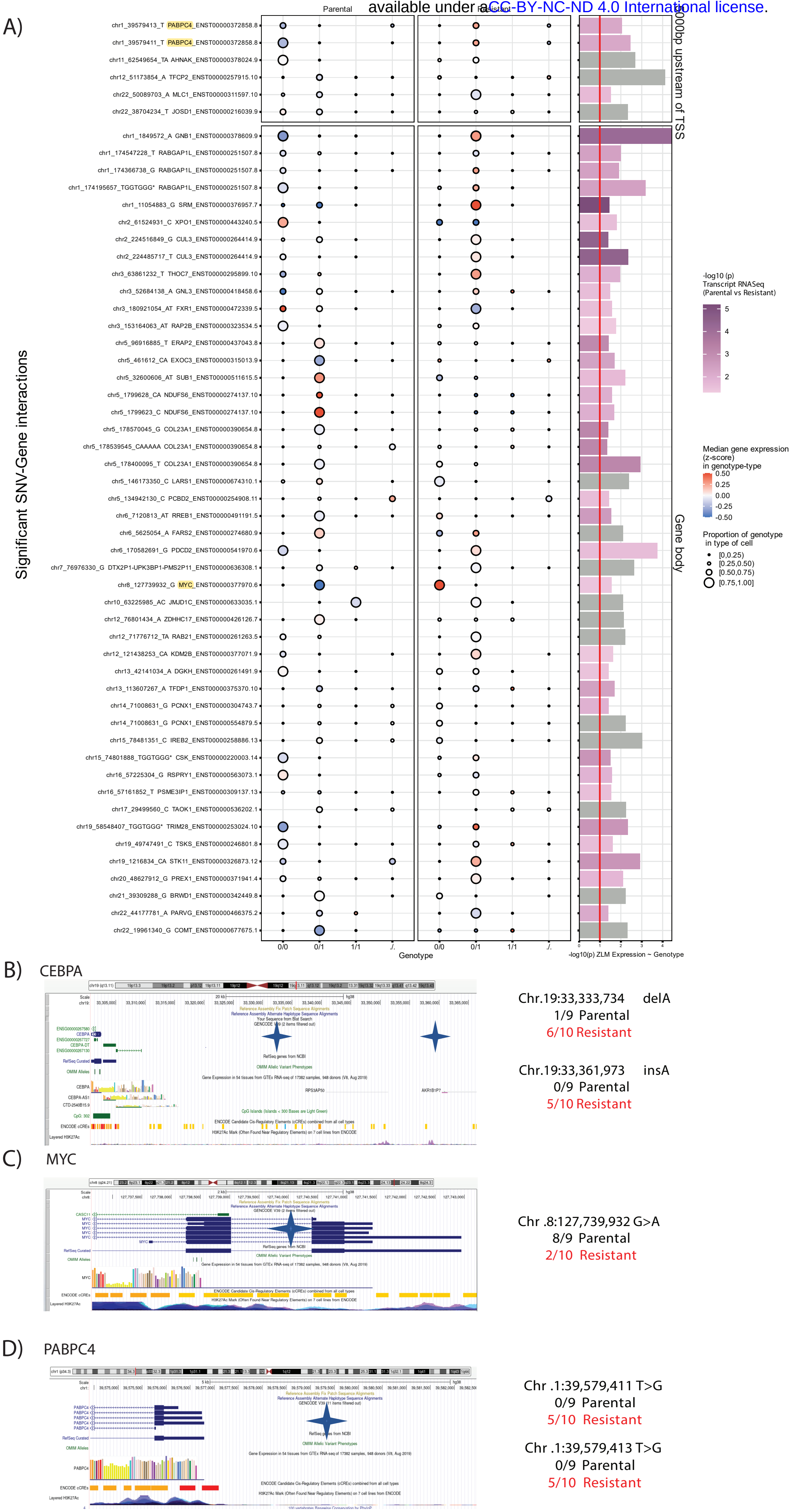

A)
Parental
Resistant
Significant SNV-Gene interactions
chr1_39579413_T PABPC4_ENST00000372858.8
chr1_39579411_T PABPC4_ENST00000372858.8
chr11_62549654_TA AHNAK_ENST00000378024.9
chr12_51173854_A TFCP2_ENST00000257915.10
chr22_50089703_A MLC1_ENST00000311597.10
chr22_38704234_T JOSD1_ENST00000216039.9
chr1_1849572_A GNB1_ENST00000378609.9
chr1_174547228_T RABGAP1L_ENST00000251507.8
chr1_174366738_G RABGAP1L_ENST00000251507.8
chr1_174195657_TGGTGGG* RABGAP1L_ENST00000251507.8
chr1_11054883_G SRM_ENST00000376957.7
chr2_61524931_C XPO1_ENST00000443240.5
chr2_224516849_G CUL3_ENST00000264414.9
chr2_224485717_T CUL3_ENST00000264414.9
chr3_63861232_T THOC7_ENST00000295899.10
chr3_52684138_A GNL3_ENST00000418458.6
chr3_180921054_AT FXR1_ENST00000472339.5
chr3_153164063_AT RAP2B_ENST00000323534.5
chr5_96916885_T ERAP2_ENST00000437043.8
chr5_461612_CA EXOC3_ENST00000315013.9
chr5_32600606_AT SUB1_ENST00000511615.5
chr5_1799628_CA NDUFS6_ENST00000274137.10
chr5_1799623_C NDUFS6_ENST00000274137.10
chr5_178570045_G COL23A1_ENST00000390654.8
chr5_178539545_CAAAAA COL23A1_ENST00000390654.8
chr5_178400095_T COL23A1_ENST00000390654.8
chr5_146173350_C LARS1_ENST00000674310.1
chr5_134942130_C PCBD2_ENST00000254908.11
chr6_7120813_AT RREB1_ENST00000491191.5
chr6_5625054_A FARS2_ENST00000274680.9
chr6_170582691_G PDCD2_ENST00000541970.6
chr7_76976330_G DTX2P1-UPK3BP1-PMS2P11_ENST00000636308.1
chr8_127739932_G MYC_ENST00000377970.6
chr10_63225985_AC JMJD1C_ENST00000633035.1
chr12_76801434_A ZDHHC17_ENST00000426126.7
chr12_71776712_TA RAB21_ENST00000261263.5
chr12_121438253_CA KDM2B_ENST00000377071.9
chr13_42141034_A DGKH_ENST00000261491.9
chr13_113607267_A TFDP1_ENST00000375370.10
chr14_71008631_G PCNX1_ENST00000304743.7
chr14_71008631_G PCNX1_ENST00000554879.5
chr15_78481351_C IREB2_ENST00000258886.13
chr15_74801888_TGGTGGG* CSK_ENST00000220003.14
chr16_57225304_G RSPRY1_ENST00000563073.1
chr16_57161852_T PSME3IP1_ENST00000309137.13
chr17_29499560_C TAOK1_ENST00000536202.1
chr19_58548407_TGGTGGG* TRIM28_ENST00000253024.10
chr19_49747491_C TSKS_ENST00000246801.8
chr19_1216834_CA STK11_ENST00000326873.12
chr20_48627912_G PREX1_ENST00000371941.4
chr21_39309288_G BRWD1_ENST00000342449.8
chr22_44177781_A PARVG_ENST00000466375.2
chr22_19961340_G COMT_ENST00000677675.1
0/0 0/1 1/1 ./.
0/0 0/1 1/1 ./.
Genotype
0 1 2 3 4
-log10(p) ZLM Expression ~ Genotype
1000bp upstream of TSS
Gene body
-log10 (p)
Transcript RNASeq
(Parental vs Resistant)
5
4
3
2
Median gene expression
(z-score)
in genotype-type
0.50
0.25
0.00
-0.25
-0.50
Proportion of genotype
in type of cell
[0,0.25)
[0.25,0.50)
[0.50,0.75)
[0.75,1.00]
B) CEBPA
CEBPA
CEBPA-DT
RefSeq Curated
OMIM Alleles
CEBPA-AS1
CTD-2540B15.9
CpG: 302
ENCODE cCREs
Layered H3K27Ac
RPS13AP50
AKR1B1P7
Chr.19:33,333,734 delA
1/9 Parental
6/10 Resistant
Chr.19:33,361,973 insA
0/9 Parental
5/10 Resistant
C) MYC
CASC11
MYC
RefSeq Curated
OMIM Alleles
ENCODE cCREs
Layered H3K27Ac
Chr .8:127,739,932 G>A
8/9 Parental
2/10 Resistant
D) PABPC4
PABPC4
RefSeq Curated
OMIM Alleles
ENCODE cCREs
Layered H3K27Ac
Chr .1:39,579,411 T>G
0/9 Parental
5/10 Resistant
Chr .1:39,579,413 T>G
0/9 Parental
5/10 Resistant

Figure 9

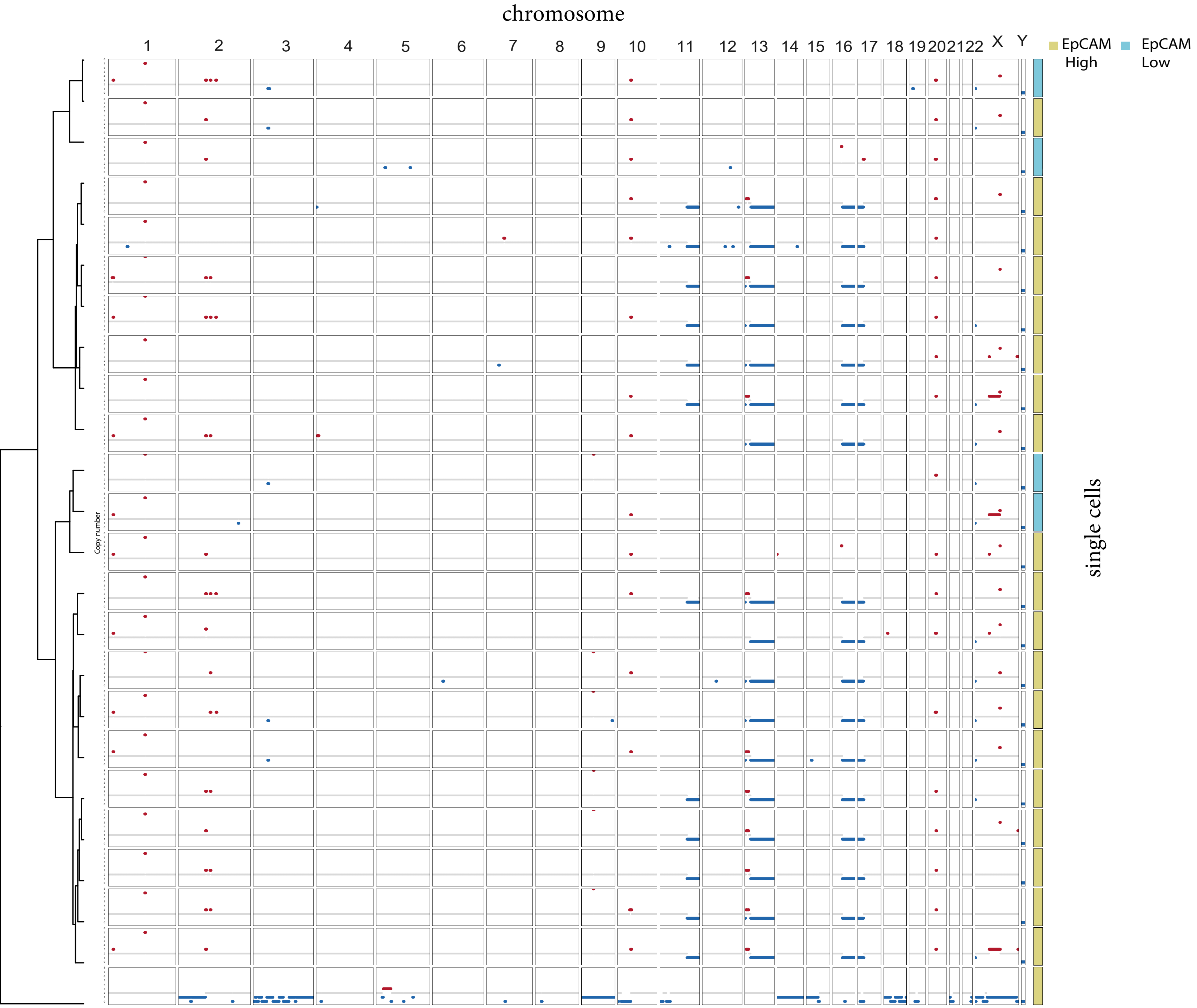
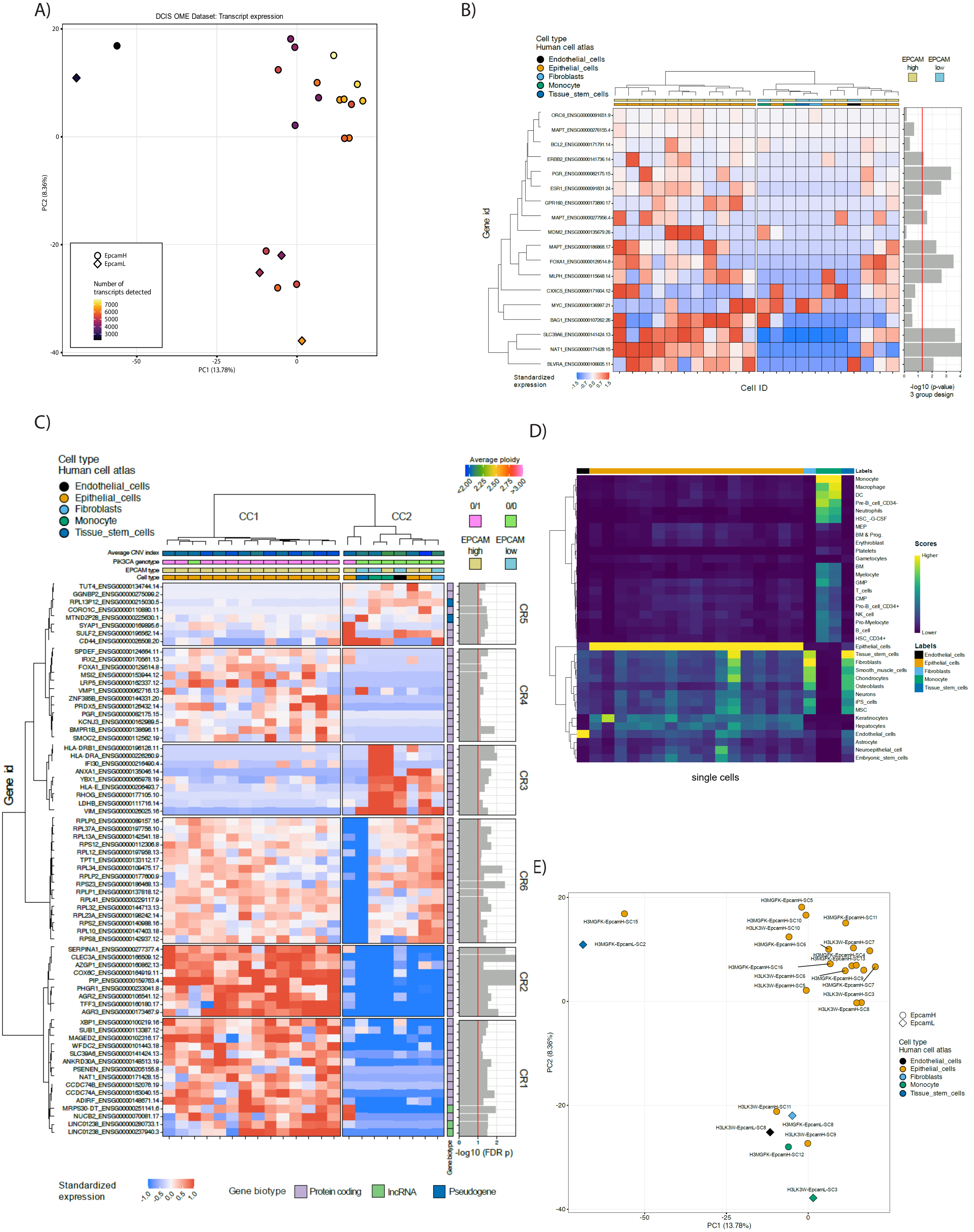

chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
EpCAM High
EpCAM Low
single cells
Copy number

Figure 10


A)
DCIS OME Dataset: Transcript expression
PC1 (13.78%)
PC2 (8.36%)
EpcamH
EpcamL
Number of transcripts detected
7000
6000
5000
4000
3000
B)
Cell type
Human cell atlas
Endothelial_cells
Epithelial_cells
Fibroblasts
Monocyte
Tissue_stem_cells
EPCAM high
EPCAM low
Gene id
ORC6_ENSG00000091651.9
MAPT_ENSG00000276155.4
BCL2_ENSG00000171791.14
ERBB2_ENSG00000141736.14
PGR_ENSG00000082175.15
ESR1_ENSG00000091831.24
GPR160_ENSG00000173890.17
MAPT_ENSG00000277956.4
MDM2_ENSG00000135679.26
MAPT_ENSG00000186868.17
FOXA1_ENSG00000129514.8
MLPH_ENSG00000115648.14
CXXC5_ENSG00000171604.12
MYC_ENSG00000136997.21
BAG1_ENSG00000107262.26
SLC39A6_ENSG00000141424.13
NAT1_ENSG00000171428.15
BLVRA_ENSG00000106605.11
Standardized expression
Cell ID
-log10 (p-value)
3 group design
C)
Cell type
Human cell atlas
Endothelial_cells
Epithelial_cells
Fibroblasts
Monocyte
Tissue_stem_cells
CC1
CC2
Average ploidy
<2.00
2.25
2.50
2.75
>3.00
0/1
0/0
EPCAM high
EPCAM low
Average CNV index
PIK3CA genotype
EPCAM type
Cell type
TUT4_ENSG00000134744.14
GGNBP2_ENSG00000275099.2
RPL13P12_ENSG00000215030.5
CORO1C_ENSG00000110880.11
MTND2P28_ENSG00000225630.1
SYAP1_ENSG00000169895.6
SULF2_ENSG00000196562.14
CD44_ENSG00000026508.20
SPDEF_ENSG00000124664.11
IRX2_ENSG00000170561.13
FOXA1_ENSG00000129514.8
MSI2_ENSG00000153944.12
LRP5_ENSG00000162337.12
VMP1_ENSG00000062716.13
ZNF385B_ENSG00000144331.20
PRDX5_ENSG00000126432.14
PGR_ENSG00000082175.15
KCNJ3_ENSG00000162989.5
BMPR1B_ENSG00000138696.11
SMOC2_ENSG00000112562.19
HLA-DRB1_ENSG00000196126.11
HLA-DRA_ENSG00000204287.14
IFI30_ENSG00000216490.4
ANXA1_ENSG00000135046.14
YBX1_ENSG00000065978.19
HLA-E_ENSG00000204592.7
RHOG_ENSG00000177105.10
LDHB_ENSG00000111716.14
VIM_ENSG00000026025.16
RPLP0_ENSG00000089157.16
RPL37A_ENSG00000197756.10
RPL13A_ENSG00000142541.18
RPS12_ENSG00000112306.8
RPL12_ENSG00000197958.13
TPT1_ENSG00000133112.17
RPL34_ENSG00000109475.17
RPLP2_ENSG00000177600.9
RPS23_ENSG00000186468.13
RPLP1_ENSG00000137818.12
RPL41_ENSG00000229117.9
RPL32_ENSG00000144713.13
RPL23A_ENSG00000198242.14
RPS2_ENSG00000140988.16
RPL10_ENSG00000147403.18
RPS8_ENSG00000142937.12
SERPINA1_ENSG00000277377.4
CLEC3A_ENSG00000166509.12
AZGP1_ENSG00000160862.13
COX6C_ENSG00000164919.11
PIP_ENSG00000159763.4
PHGR1_ENSG00000233041.8
AGR2_ENSG00000106541.12
TFF3_ENSG00000160180.17
AGR3_ENSG00000173467.9
XBP1_ENSG00000100219.16
SUB1_ENSG00000113387.12
MAGED2_ENSG00000102316.17
WFDC2_ENSG00000101443.18
SLC39A6_ENSG00000141424.13
ANKRD30A_ENSG00000148513.19
PSENEN_ENSG00000205155.8
NAT1_ENSG00000171428.15
CCDC74B_ENSG00000152076.19
CCDC74A_ENSG00000163040.15
ADIRF_ENSG00000148671.14
MRPS30-DT_ENSG00000251141.6
NUCB2_ENSG00000070081.17
LINC01238_ENSG00000280733.1
LINC01238_ENSG00000237940.3
CR5
CR4
CR3
CR6
CR2
CR1
Gene id
Gene biotype
-log10 (FDR p)
Standardized expression
Gene biotype
Protein coding
lncRNA
Pseudogene
D)
Labels
Monocyte
Macrophage
DC
Pre-B_cell_CD34-
Neutrophils
HSC_-G-CSF
MEP
BM & Prog.
Erythroblast
Platelets
Gametocytes
BM
Myelocyte
GMP
T_cells
CMP
Pro-B_cell_CD34+
NK_cell
Pro-Myelocyte
B_cell
HSC_CD34+
Epithelial_cells
Tissue_stem_cells
Fibroblasts
Smooth_muscle_cells
Chondrocytes
Osteoblasts
Neurons
iPS_cells
MSC
Keratinocytes
Hepatocytes
Endothelial_cells
Astrocyte
Neuroepithelial_cell
Embryonic_stem_cells
Scores
Higher
Lower
Labels
Endothelial_cells
Epithelial_cells
Fibroblasts
Monocyte
Tissue_stem_cells
single cells
E)
H3MGFK-EpcamH-SC5
H3MGFK-EpcamH-SC15
H3MGFK-EpcamL-SC2
H3MGFK-EpcamH-SC10
H3LK3W-EpcamH-SC10
H3MGFK-EpcamH-SC11
H3LK3W-EpcamH-SC7
H3MGFK-EpcamH-SC6
H3MGFK-EpcamH-SC4
H3MGFK-EpcamH-SC13
H3MGFK-EpcamH-SC16
H3LK3W-EpcamH-SC6
H3LK3W-EpcamH-SC9
H3LK3W-EpcamH-SC5
H3MGFK-EpcamH-SC7
H3LK3W-EpcamH-SC3
H3LK3W-EpcamH-SC8
H3LK3W-EpcamH-SC11
H3MGFK-EpcamL-SC8
H3LK3W-EpcamL-SC8
H3LK3W-EpcamH-SC9
H3MGFK-EpcamH-SC12
H3LK3W-EpcamL-SC3
PC1 (13.78%)
PC2 (8.36%)
EpcamH
EpcamL
Cell type
Human cell atlas
Endothelial_cells
Epithelial_cells
Fibroblasts
Monocyte
Tissue_stem_cells

Supplementary Data 1: Bulk External Datasets

SRR15909920: A public-private-academic consortium, Genome-in-a-Bottle (GIAB), hosted by NIST to develop reference materials and standards for clinical sequencing. https://www.nature.com/articles/s41587-019-0054-x

ERR356372: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. https://www.science.org/doi/10.1126/science.1242463?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

SRR1258218: Transcriptome sequencing of a large human family identifies the impact of rare non-coding variants. https://www.sciencedirect.com/science/article/pii/S0002929714003486#:~:text=Article-,Transcriptome%20Sequencing%20of%20a%20Large%20Human%20Family,Impact%20of%20Rare%20Noncoding%20Variants&text=Recent%20and%20rapid%20human%20population,genetic%20burden%20of%20disease%20risk.

SRR307005: RNA-Seq in GM12878 (ENCODE Project).

SRR307006: RNA-Seq in GM12878 (ENCODE Project).

SRR307007: RNA-Seq in GM12878 (ENCODE Project).

SRR307008: RNA-Seq in GM12878 (ENCODE Project).

SRR065532: ENCODE Caltech RNA-seq

SRR5117664: Analysis of parent-of-origin bias in gene expression levels. https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-019-0674-0

SRR5117665: Analysis of parent-of-origin bias in gene expression levels. https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-019-0674-0

SRR653897: Homo sapiens Transcriptome or Gene expression.

SRR002055: Pilot ENCODE transcriptome data.

SRR002063: Pilot ENCODE transcriptome data.

SRR005091: Pilot ENCODE transcriptome data.

SRR002052: Pilot ENCODE transcriptome data.

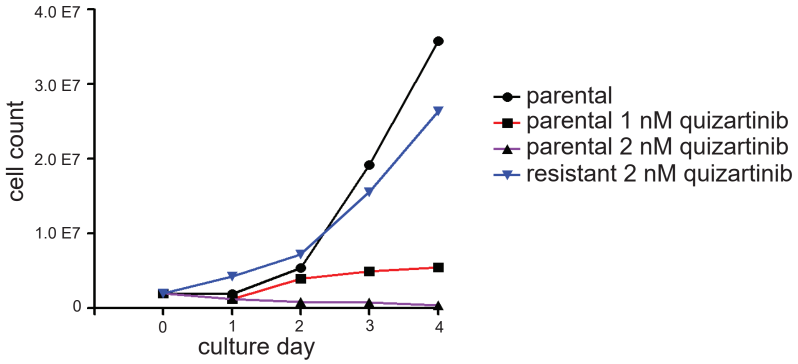
SRR002054: Pilot ENCODE transcriptome data.

SRR002060: Pilot ENCODE transcriptome data.

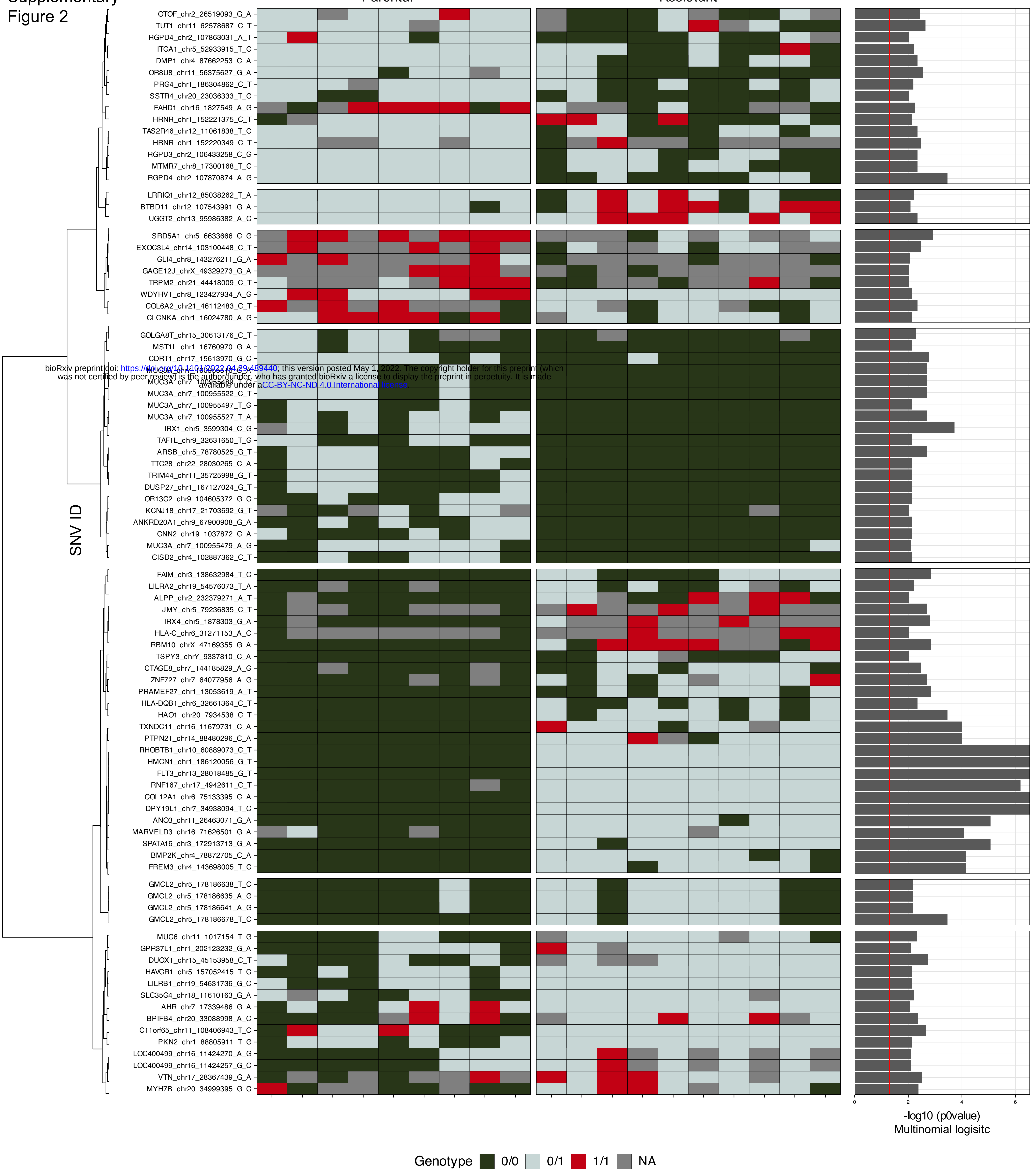
SRR065514: ENCODE Caltech RNA-seq
SRR065515: ENCODE Caltech RNA-seq

SRR065510: ENCODE Caltech RNA-seq

Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3

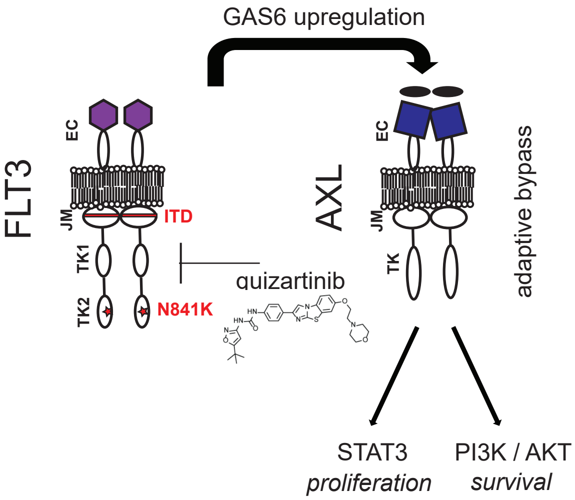
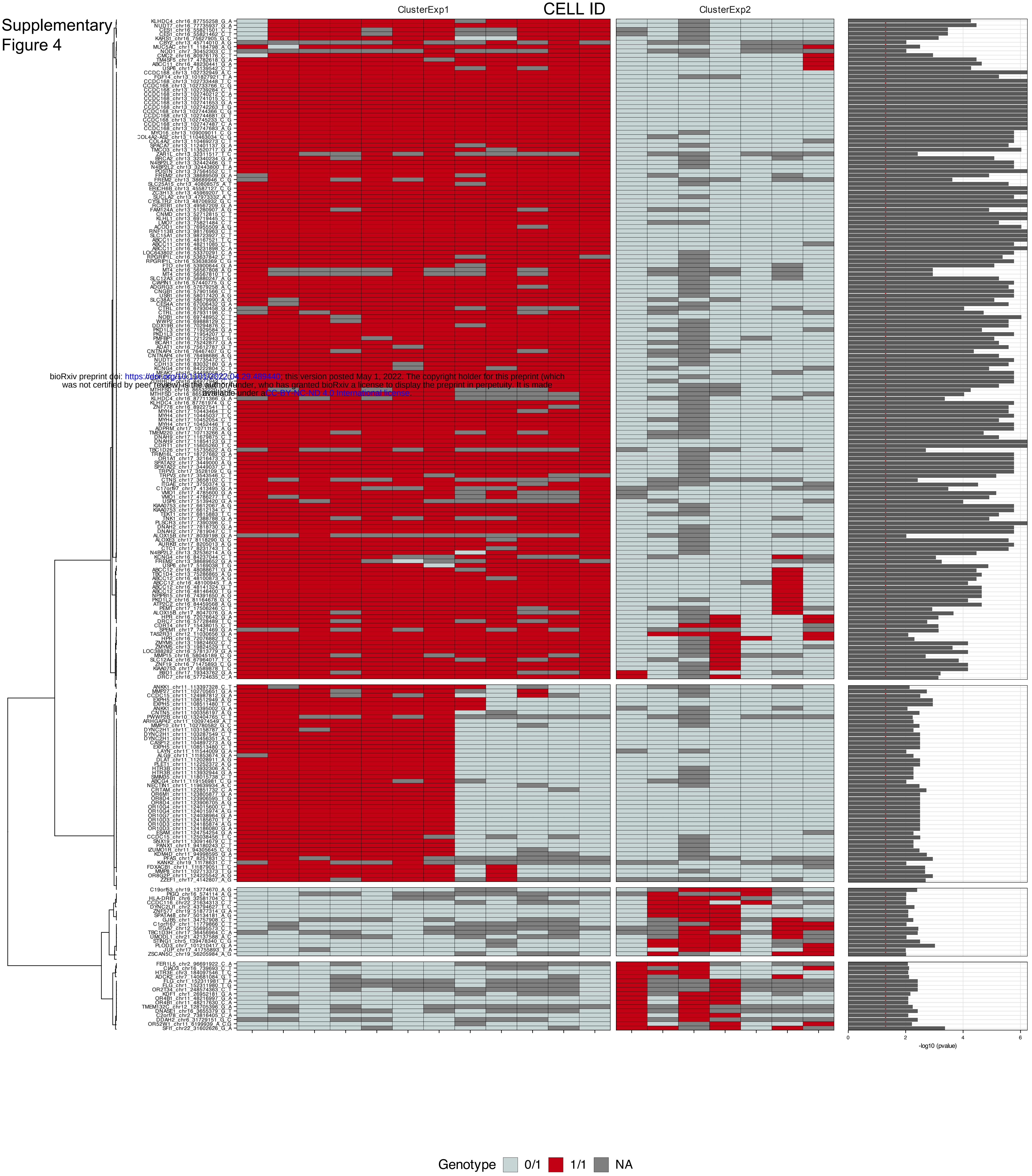


FIGURE 4

Supplementary Figure 4


CELL ID
ClusterExp1
ClusterExp2
KLHDC4_chr16_87755258_G_A
NUDT7_chr16_77735937_G_A
CES1_chr16_55821501_C_T
CES1_chr16_55821462_C_T
KARS1_chr16_75627905_G_C
CBY2_chr13_45714010_A_G
MUC5AC_chr11_1184798_A_G
NOD1_chr7_30452303_C_T
CMC2_chr16_80976176_C_T
TM4SF5_chr17_4782618_G_A
ABCC11_chr16_48230441_G_A
USP6_chr17_5139542_C_T
CCDC168_chr13_102732949_A_C
FGF14_chr13_101827921_T_A
CCDC168_chr13_102733448_T_C
CCDC168_chr13_102733766_C_G
CCDC168_chr13_102739284_C_T
CCDC168_chr13_102740212_C_A
CCDC168_chr13_102741015_C_T
CCDC168_chr13_102741653_G_A
CCDC168_chr13_102742263_T_G
CCDC168_chr13_102744366_C_G
CCDC168_chr13_102744681_G_T
CCDC168_chr13_102745233_C_G
CCDC168_chr13_102747487_C_A
CCDC168_chr13_102747683_A_G
MYO16_chr13_109009011_C_G
COL4A2-AS2_chr13_110463034_C_G
COL4A2_chr13_110469273_C_T
SPACA7_chr13_112401137_G_A
TMCO3_chr13_113520717_G_A
ZAR1L_chr13_32311517_T_C
BRCA2_chr13_32340234_G_A
N4BP2L2_chr13_32442466_G_T
N4BP2L2_chr13_32443800_T_A
POSTN_chr13_37564552_C_T
FREM2_chr13_38689509_G_A
FREM2_chr13_38689946_C_G
SLC25A15_chr13_40808575_A_T
ERICH6B_chr13_45587127_C_G
ZC3H13_chr13_45969207_T_C
SUCLA2_chr13_47973332_A_T
CYSLTR2_chr13_48706932_G_C
RCBTB1_chr13_49567209_G_A
FAM124A_chr13_51280907_A_G
CNMD_chr13_52712815_C_T
KLHL1_chr13_69719445_C_T
LMO7_chr13_75821484_C_T
ACOD1_chr13_76955509_A_G
RNF113B_chr13_98176963_C_T
SLC15A1_chr13_98723927_C_T
ABCC11_chr16_48167521_T_C
ABCC11_chr16_48211085_C_T
ABCC11_chr16_48231898_C_A
LOC643802_chr16_53370291_C_A
RPGRIP1L_chr16_53637842_C_T
RPGRIP1L_chr16_53638369_C_G
FTO_chr16_53900644_G_A
MT4_chr16_56567808_A_G
MT4_chr16_56567810_T_C
SLC12A3_chr16_56880247_A_G
CIAPIN1_chr16_57440775_G_C
ADGRG3_chr16_57679258_A_C
CNGB1_chr16_57901566_C_T
USB1_chr16_58017420_A_G
SLC38A7_chr16_58679990_A_G
CES4A_chr16_67006432_G_A
CTRL_chr16_67930458_G_A
CTRL_chr16_67931196_C_T
NOB1_chr16_69748952_C_T
WWP2_chr16_69888129_C_T
DDX19B_chr16_70294876_C_T
PKD1L3_chr16_71929584_G_A
PKD1L3_chr16_71954207_C_T
PMFBP1_chr16_72122943_T_G
BCAR1_chr16_75242877_G_A
ADAT1_chr16_75612787_G_T
CNTNAP4_chr16_76467407_G_C
CNTNAP4_chr16_76498686_A_G
NUDT7_chr16_77735472_C_T
CDH13_chr16_83032180_G_A
KCNG4_chr16_84222804_C_T
MEAK7_chr16_84498058_C_A
ZDHHC7_chr16_84977942_C_T
MTHFSD_chr16_86532220_C_G
KLHDC4_chr16_87711366_G_A
KLHDC4_chr16_87761974_G_C
ZNF778_chr16_89227541_T_C
MYH4_chr17_10443464_T_C
MYH4_chr17_10445037_T_C
MYH4_chr17_10452054_C_T
MYH4_chr17_10452446_T_C
ADPRM_chr17_10711125_A_G
TMEM220_chr17_10713266_A_G
DNAH9_chr17_11679875_C_T
DNAH9_chr17_11854123_G_T
CDRT1_chr17_15605260_T_C
TBC1D26_chr17_15735622_A_G
TRIM16L_chr17_18727682_G_A
OR1A1_chr17_3216473_C_T
SPATA22_chr17_3449000_A_G
SPATA22_chr17_3449037_C_T
TRPV3_chr17_3528109_C_G
TRPV3_chr17_3543546_C_T
CTNS_chr17_3658102_C_T
ITGAE_chr17_3750374_G_T
C17orf97_chr17_413495_G_A
VMO1_chr17_4785600_G_A
VMO1_chr17_4786277_T_C
USP6_chr17_5139420_G_A
KIAA0753_chr17_6612067_A_G
KIAA0753_chr17_6612134_C_T
TEKT1_chr17_6815883_T_C
TNK1_chr17_7388788_G_A
PLSCR3_chr17_7390396_C_T
DNAH2_chr17_7818730_G_A
DNAH2_chr17_7819047_C_T
ALOX15B_chr17_8039198_G_A
ALOXE3_chr17_8118290_G_C
AURKB_chr17_8205013_A_G
CTC1_chr17_8231743_T_C
N4BP2L2_chr13_32536214_A_G
KCNG4_chr16_84237044_C_T
FREM2_chr13_38689652_G_A
USP6_chr17_5169038_T_G
ABCC12_chr16_48088671_G_A
TBC1D4_chr13_75286865_A_G
ABCC12_chr16_48100873_A_G
ABCC12_chr16_48100945_T_A
ABCC12_chr16_48141324_G_T
ABCC12_chr16_48146400_T_G
NPIPB15_chr16_74391650_A_G
PKD1L2_chr16_81164678_G_C
ATP2C2_chr16_84459568_A_G
PEMT_chr17_17506246_C_T
ALOX15B_chr17_8047076_G_A
HPR_chr16_72076642_G_A
DRC7_chr16_57728489_T_C
CDRT4_chr17_15438015_C_T
SPEM1_chr17_7421469_G_A
TAS2R31_chr12_11030656_G_A
HPR_chr16_72076882_T_C
ZMYM5_chr13_19824602_C_T
ZMYM5_chr13_19824529_T_C
LOC388282_chr16_57813779_G_A
MMP15_chr16_58045189_C_G
SLC12A4_chr16_67964017_T_C
ZNF19_chr16_71475893_C_G
KIAA0753_chr17_6589878_T_C
B9D1_chr17_19343762_G_A
DRC7_chr16_57724635_C_A
ANKK1_chr11_113397328_C_T
MMP27_chr11_102705651_G_A
CCDC15_chr11_124987812_G_A
EXPH5_chr11_108512949_A_G
EXPH5_chr11_108511480_T_C
ANKK1_chr11_113395002_G_A
CNTN5_chr11_100356197_A_G
PWWP2B_chr10_132404765_C_T
ARHGAP42_chr11_100974549_A_T
MMP10_chr11_102780582_G_C
DYNC2H1_chr11_103158787_A_G
DYNC2H1_chr11_103287549_C_T
DYNC2H1_chr11_103456351_A_C
CASP12_chr11_104897273_A_G
EXPH5_chr11_108513480_C_T
LAYN_chr11_111544009_G_A
ALG9_chr11_111853674_G_A
DLAT_chr11_112028911_A_G
PLET1_chr11_112252372_A_G
HTR3B_chr11_113932306_A_C
HTR3B_chr11_113932944_G_A
SMIM35_chr11_118015738_C_T
ABCG4_chr11_119156981_C_G
NECTIN1_chr11_119639934_A_C
CRTAM_chr11_122851732_C_A
OR6M1_chr11_123805877_G_A
OR8D4_chr11_123906595_T_G
OR8D4_chr11_123906705_A_G
OR10G4_chr11_124015600_C_T
OR10G4_chr11_124015974_A_G
OR10G7_chr11_124038964_G_A
OR10D3_chr11_124185670_T_C
OR10D3_chr11_124185874_A_G
OR10D3_chr11_124186080_G_A
ESAM_chr11_124754254_G_A
CCDC15_chr11_125038456_T_C
SNX19_chr11_130914679_C_T
PANX1_chr11_94180243_C_T
IZUMO1R_chr11_94305645_C_G
KDM4D_chr11_94998595_G_A
PFAS_chr17_8257831_C_T
KANK2_chr19_11178631_C_T
FDXACB1_chr11_111879051_T_C
MMP8_chr11_102713373_T_G
OR8G2P_chr11_124225542_A_G
ZZEF1_chr17_4142807_A_G
C19orf53_chr19_13774670_A_G
PIGQ_chr16_574114_A_G
HLA-DRB1_chr6_32581704_C_T
CCDC116_chr22_21634313_C_T
DYNC2LI1_chr2_43794627_T_C
ZNF577_chr19_51877314_G_A
SPATA48_chr7_50134181_A_G
GJB5_chr1_34757908_C_T
C1orf167_chr1_11779866_C_T
ITGA7_chr12_55695573_C_T
TBC1D3H_chr17_36456964_C_A
UMODL1_chr21_42137588_A_C
STING1_chr5_139478340_C_G
PLOD3_chr7_101210417_G_A
JUP_chr17_41755893_T_A
ZSCAN5C_chr19_56205984_A_G
FER1L5_chr2_96691922_C_A
CIAO3_chr16_739693_C_T
HTR3E_chr3_184097546_T_C
ADCK2_chr7_140681084_G_T
FLG_chr1_152311981_T_A
FLG_chr1_152311980_T_G
OR2T34_chr1_248574363_C_T
KDF1_chr1_26952181_G_A
OR4B1_chr11_48216997_G_A
OR4B1_chr11_48217630_C_A
TMEM132C_chr12_128705396_G_A
DNASE1_chr16_3655379_G_T
C2orf78_chr2_73816405_C_A
DDAH2_chr6_31729151_G_C
OR52W1_chr11_6199939_A_C,G
SFI1_chr22_31602626_G_A
-log10 (pvalue)
Genotype
0/1
1/1
NA