nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-022-01468-y

Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes

Teng Gao[1], Ruslan Soldatov[1], Hirak Sarkar[1], Adam Kurkiewicz[1], Evan Biederstedt[1], Po-Ru Loh[2,3] and Peter V. Kharchenko[1,3,4,5] ✉

Genome instability and aberrant alterations of transcriptional programs both play important roles in cancer. Single-cell RNA sequencing (scRNA-seq) has the potential to investigate both genetic and nongenetic sources of tumor heterogeneity in a single assay. Here we present a computational method, Numbat, that integrates haplotype information obtained from population-based phasing with allele and expression signals to enhance detection of copy number variations from scRNA-seq. Numbat exploits the evolutionary relationships between subclones to iteratively infer single-cell copy number profiles and tumor clonal phylogeny. Analysis of 22 tumor samples, including multiple myeloma, gastric, breast and thyroid cancers, shows that Numbat can reconstruct the tumor copy number profile and precisely identify malignant cells in the tumor microenvironment. We identify genetic subpopulations with transcriptional signatures relevant to tumor progression and therapy resistance. Numbat requires neither sample-matched DNA data nor a priori genotyping, and is applicable to a wide range of experimental settings and cancer types.

Copy number variations (CNVs) and loss of heterozygosity (LoH) events are major genome aberrations found in nearly all cancer cells. Characterization of CNVs in healthy and malignant tissues has informed the early detection, modes of progression and resistance mechanisms of cancer. However, the functional impacts of CNVs on overall cellular activity and how they drive malignant transformation remain largely unclear. Genome instability is also a key contributor to intratumoral heterogeneity. Therapy-resistant subclones frequently arising in the course of treatment pose a major challenge to cancer therapies. In addition to genetic heterogeneity, resistance may also stem from changes in the epigenetic or regulatory state, although the relative importance of different mechanisms has been difficult to establish[1]. All such changes, however, including genomic alterations, are probably reflected in the transcriptional state of the cell.

Single-cell RNA sequencing methods provide an excellent opportunity to bridge genetic heterogeneity with the overall cellular state. It has been demonstrated that CNVs can be inferred from both transcript abundance and allelic imbalance in heterozygous single-nucleotide polymorphisms (SNPs)[2–5]. Reliable inference of copy number states, however, remains challenging using either approach due to the sparse and noisy nature of single-cell measurements. Expression-based methods infer the presence of CNVs based on a general expectation that amplifications (AMPs) or deletions (DELs) will, on average, result in up- or downregulation, respectively, of genes within the affected region of the genome. Such approach can produce false-positive results due to local variations in expression unrelated to genomic copy numbers[6]. Allele-based approaches examine deviations of heterozygous allele frequency ('B-allele frequency', or BAF) caused by CNVs and are less affected by sample or cell type variations[2,5]. They are hindered, however, by data sparsity and allele-specific transcriptional stochasticity in single cells[7].

Existing approaches for CNV detection from scRNA-seq do not use the previous knowledge of haplotypes, or the individual-specific configuration of variant alleles on the two homologous chromosomes, which can enable more sensitive detection of allelic imbalance. Although current sequencing technologies are generally not haplotype resolved, population-based phasing provides a means to computationally phase variants of an individual using population haplotype frequencies[8,9]. The estimated haplotypes are highly accurate within adjacent genomic regions, with a typical span of 50 kb to 1 Mb, but are subject to phase switch errors that accumulate with longer genomic distance[10]. Nonetheless, population-based phasing has been successfully applied to characterize chromosomal aberrations in the context of both germline polymorphisms and cancer evolution, mainly using DNA sequencing/array genotyping data[11–14]. The utility of phasing in detection of CNV signals from scRNA-based assays, however, has not been explored. We hypothesized that previous phasing information would be particularly impactful in the context of the sparse coverage provided by scRNA-seq.

Finally, single-cell sequencing provides a unique opportunity to dissect genetically heterogeneous subpopulations, which are masked in bulk measurements. Because scRNA-seq yields limited coverage per cell, methods that use allele information typically rely on aggregation of information across cells (forming in silico 'pseudobulk' profiles) to confidently define aberrations[2,5]. This approach, however, will increase statistical power only if the aberrations are shared between the cells included in the pseudobulk, and could lead to a dilution of signal with the inclusion of genetically distinct cells. Therefore, reliable identification of subclonal CNV events depends on the accurate inference of clonal cell populations.

We therefore developed a computational method, Numbat, that integrates expression, allele and haplotype information derived from population-based phasing to comprehensively characterize the CNV landscape in single-cell transcriptomes. Numbat employs an iterative approach to jointly reconstruct the subclonal phylogeny and single-cell copy number profile of the tumor sample. Applying our method to 22 tumor samples (59,878 single cells) representing

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Harvard Stem Cell Institute, Cambridge, MA, USA. [5]Present address: Altos Labs, San Diego, CA, USA. ✉e-mail: peter.kharchenko@post.harvard.edu

a variety of cancer types and genomic complexity, we show that Numbat reconstructs high-fidelity copy number profiles from scRNA data alone and accurately distinguishes cancer cells from normal cells in the tumor microenvironment. Within heterogeneous tumors, Numbat readily identifies distinct subclonal lineages that harbor allele-specific alterations. Numbat requires neither sample-matched DNA data nor a priori genotyping, and is applicable to a wide range of experimental settings and cancer types.

Results

Sensitive CNV detection using haplotype information. Previous phasing information can effectively amplify weak allelic imbalance signals of individual SNPs induced by the CNV, by exposing joint behavior of entire haplotype sequences and thereby increasing the statistical power[11–14]. To examine the extent to which expressed heterozygous SNPs can be detected from scRNA-seq data, we genotyped common germline SNPs (>5% population frequency) in 22 tumor samples sequenced by high-throughput, droplet-based protocols (Supplementary Table 1). The density of the detected heterozygous SNPs along the genome and per-cell SNP coverage vary by sample and dataset (16–68 SNPs per megabase and 159–1,045 counts per cell; Supplementary Fig. 1a,b). A large proportion of the SNPs is detected within intronic regions, although with lower coverage than for SNPs within untranslated and exonic regions (Supplementary Fig. 1c,d). To demonstrate the feasibility of population-based phasing in such coverage setting, we first analyzed a triple-negative breast cancer sample (TNBC4) containing widespread loss of heterozygosity. The observed allele counts in chromosome arms with complete LoH allowed us to confidently phase alleles ($P < 0.05$, two-sided binomial test) into their respective haplotypes. We performed population-based haplotype phasing using a reference-based phasing algorithm, Eagle2 (ref. [8]), with respect to two different population genome reference panels: TOPMed and 1000 Genomes (1000G)[15,16].

We found that population-based phasing was effective at inferring the haplotype of long stretches of expressed SNPs (mean: 11.6 SNPs, interquartile range (IQR) 2–15, by TOPMed). SNPs within the same gene were phased with especially high accuracy (96.8%) as compared with coexpression-based phasing (83.7%)[17]. Furthermore, population-based phasing was also able to infer the haplotype across genes, producing perfectly phased blocks containing on average 3.8 genes (IQR 1–5) and achieving a between-gene phasing accuracy of 79.8%. In contrast, coexpression-based phasing relies on haplotype-specific expression of alleles within the same gene and cannot phase across genes. The ability to infer phasing between genes is particularly useful for CNV inference because it provides a potential means to overcome stochastic allele-specific expression (ASE) effects that give rise to bursts of gene-specific allelic imbalances in individual cells. ASE is prevalent in normal diploid cells due to a combination of amplification bias, transcriptional bursting[18] or cis-regulatory effects[19]. However, in diploid regions the direction of ASE is independent between genes—that is, given a transcriptional burst of a gene from a maternal chromosome, the neighboring genes would be on average equally likely to show bursts from either maternal or paternal chromosomes. In contrast, the presence of a CNV would result in a consistent allelic bias among a stretch of neighboring genes towards a particular chromosome. The knowledge of haplotypes provided by population-based phasing enables aggregation of allelic bias signals across SNPs in consecutive genes, thus overcoming noise resulting from ASE (Supplementary Fig. 2).

Hidden Markov models (HMMs) have been used effectively to detect allelic imbalances from noisy signals[2,5,11,14,20,21]. The conventional allele-focused approach (haplotype-naive HMM, such as that used by HoneyBADGER) infers the presence of events by increased variance of allele frequencies in the affected regions (Fig. 1a, left)[2,5,20,21]. On the other hand, a haplotype-aware HMM exploits signed deviations of phased haplotype frequencies to gain additional statistical power (Fig. 1a, right)[11,14]. The aberrant genome state is represented by a pair of mirrored states with reciprocal transitions to account for phase switch errors in the population-derived haplotypes, which can shift between the more abundant (major) haplotype and the less abundant (minor) haplotype (Extended Data Fig. 1b). To reflect the decay in phasing strength over longer genetic distances, we introduced site-specific phase switch probabilities between haplotype states (Methods). This gives rise to an inhomogeneous Markov chain where the haplotype transition probabilities are an exponential function of inter-SNP distance (Extended Data Fig. 1a,b).

To benchmark the extent to which phasing helps with inferring CNVs and single-cell genotypes from scRNA-seq based on allele data, we used the existing cell annotation of TNBC4 and five multiple myeloma (MM) samples with matched whole-genome sequencing (WGS) to create tumor/normal mixture pseudobulk profiles for a range of tumor cell fractions (clonality 0–100%; Supplementary Figs. 3 and 4). Compared with the naive model, the haplotype-aware allele HMM readily identified subtle allelic imbalances that would otherwise be invisible (Fig. 1b) and achieved a higher area under the curve (AUC) at low tumor fractions (Fig. 1c). Phasing also improved CNV detection sensitivity at low coverage settings and for amplification events (Supplementary Fig. 5). We then asked whether we could confidently test for the presence of individual CNVs in single cells using the event characteristics obtained from the pseudobulk profile. An accurately phased haplotype is crucial for identification of genotypes of individual cells, because it helps overcome sparse SNP coverage by aggregating allele counts over affected regions[2]. In a naive HMM, the assignment of alleles to either haplotype is based solely on the observed allele frequencies (an allele is classified as major if its BAF is >0.5), whereas a haplotype-aware HMM combines evidence from previous phasing information and observed allele data to reconstruct haplotypes a posteriori. Using BAF-based allele classification in the all-tumor pseudobulk as ground truth, we found that our haplotype-aware HMM achieved higher allele classification accuracy in aberrant regions, especially at low tumor cell fractions (Fig. 1d,e). As a result, allelic imbalances were more readily discernible in individual tumor cells using posterior allele assignments from the haplotype-aware HMM (Fig. 1f,g and Supplementary Fig. 5). Therefore, incorporation of population phasing signal enables more sensitive characterization of allelic imbalances, and hence CNVs and LoH events, from scRNA-seq data.

Allele-specific copy number inference from transcriptomes. Both allelic imbalance, which reflects the relative copy number of two homologous chromosomes, and changes in expression magnitudes, which reflect total chromosomal dosage, provide signals for characterization of genome aberrations[2,5]. To integrate these two types of signal we designed a joint HMM based on a generative statistical framework (Methods and Extended Data Fig. 2). We expanded the state space of the haplotype-aware allele HMM by combining the expected expression shifts and allele frequencies corresponding to each copy number configuration (Methods, Extended Data Fig. 1c and Supplementary Fig. 6). To increase robustness, Numbat models gene expression as integer read counts using a Poisson log-normal mixture distribution and accounts for overdispersion in allele counts (for example, due to allele-specific detection or transcriptional bursts) using a beta-binomial distribution. The resulting HMM simultaneously calls significantly altered regions and determines their allele-specific copy number states (Fig. 2a). The expression and allele signal in single cells can similarly be integrated to produce probabilistic estimates of event presence in single cells (Fig. 2b).

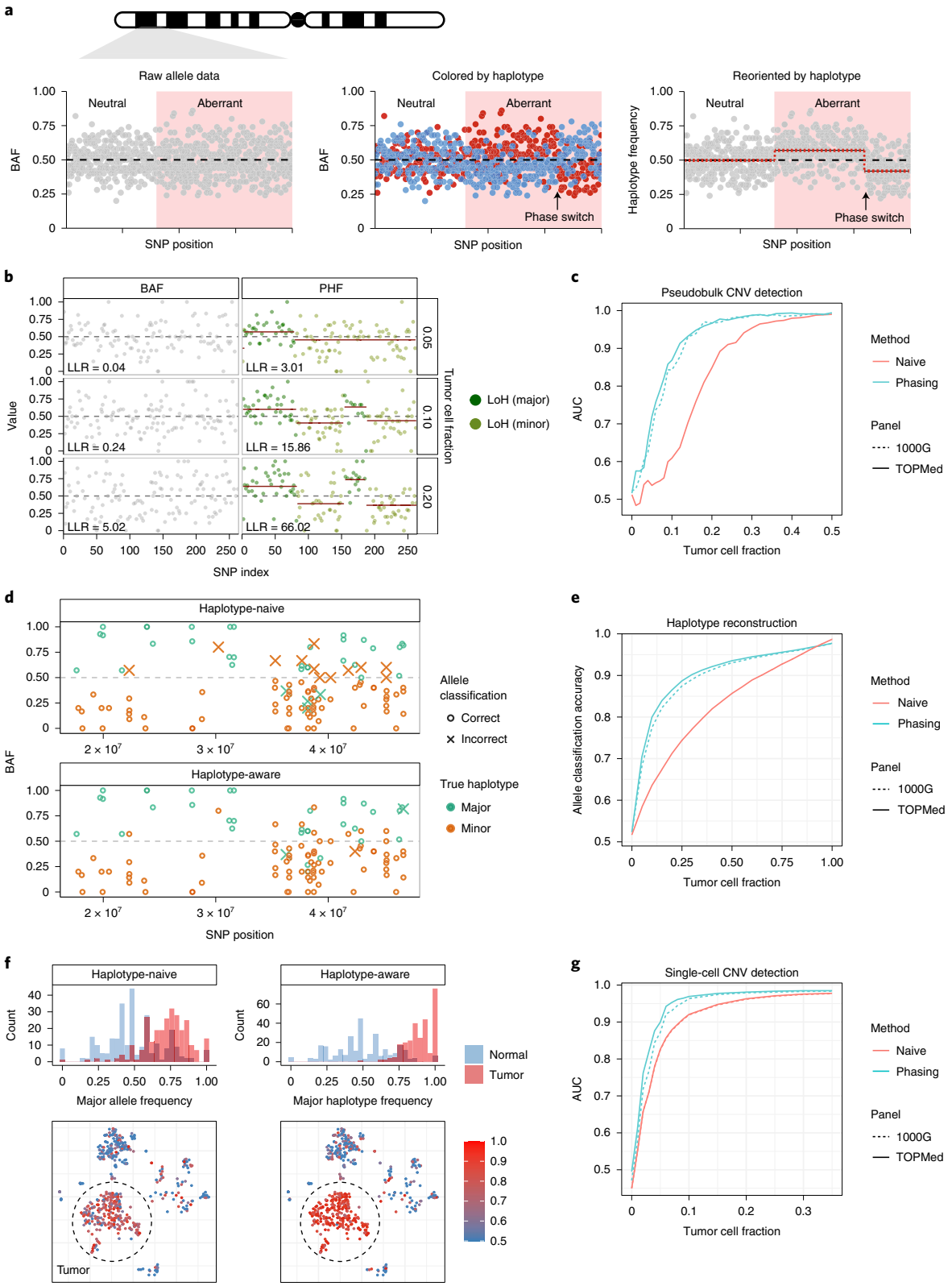


Fig. 1 | Population-based haplotype phasing enables sensitive detection of subclonal allelic imbalances in single-cell transcriptomes. a, Schematic of the use of haplotype information to detect allelic imbalance. Simulated BAF signals are shown for a neutral and aberrant region harboring subclonal CNV. After BAF is transformed into haplotype frequency based on phase information, CNV signals become apparent and can be segmented. **b**, Example of statistical phasing signal uncovering subclonal LoH in TNBC4 tumor/normal cell mixtures that are undetectable using BAF deviation. LLR, log-likelihood ratio. LoH, loss of heterozygosity. PHF, paternal haplotype frequency. **c**, Performance of LoH detection in tumor/normal mixtures with and without haplotype phasing ('phasing' and 'naive'). AUC, area under the ROC curve. **d**, Example of population-based phasing informing allele classification into major and minor haplotypes. **e**, Performance of allele classification accuracy in tumor/normal mixtures. **f**, Example of population-based phasing improving detection of LoH in single cells. **g**, Performance of LoH detection in single cells.

Existing methods infer copy number variations relative to the median ploidy, which can dilute signals of aberrant regions or mistake neutral regions for aberrant due to baseline shifts caused by hyperdiploidy or hypodiploidy[22]. To identify the diploid baseline, Numbat adopts a two-step approach: first, allelically balanced regions are identified through an allele-only HMM. The balanced regions are then clustered based on the expression shifts, and the cluster with the lowest average fold-change is designated as diploid regions (Supplementary Methods).

To validate the performance of copy number inference using the Numbat joint HMM, we turn to scRNA-seq data of the five MM samples with sample-matched, flow-sorted WGS. We detected CNV events from malignant plasma cells using the Numbat joint HMM, Numbat expression-only HMM and three other methods (HoneyBADGER, InferCNV and CopyKat). We found that the copy number events identified by Numbat are highly concordant with the corresponding DNA profiles (Fig. 2c and Extended Data Fig. 3), achieving higher overall accuracy (precision: 99.2%, recall: 95.4%) than other methods (Fig. 2d). Although the number of expressed SNPs varies by event size, incorporation of allele information significantly improved the overall event calling performance (Fig. 2d and Supplementary Fig. 7a). The results are generally not sensitive to specific choices of hyperparameters used to configure the HMM (Supplementary Fig. 7b). In addition, Numbat correctly identified copy-neutral loss of heterozygosity (CNLoH) events in two samples (chr1p of 47491-primary and chr5 of 59114-relapse-1), which are invisible to approaches that consider only expression magnitude, including InferCNV and CopyKAT (Fig. 2c). When tested on nonmalignant cell populations, Numbat made the fewest number of false-positive calls, demonstrating its specificity (Supplementary Fig. 8). Numbat also outperformed other methods on CNV testing on a single-cell level (Fig. 2e).

Numbat correctly identified the diploid baseline in all five cases, whereas the copy number estimates produced by the other three methods are often confounded by baseline shifts caused by hyperdiploidy (for example, 37692-primary and 47491-primary; Fig. 2c). This issue is particularly pronounced in a premalignant breast cancer sample (DCIS1), where CopyKAT denoted chromosomes 3, 9, 10 and 15 as deleted and chromosomes 1, 7q and 14 as copy-neutral (Supplementary Fig. 9a). In contrast, Numbat analysis using both allele and expression data revealed that chromosomes 3, 9, 10 and 15 are largely allelically balanced and therefore probably remain in diploid state, whereas chromosomes 1, 7q and 14 carry widespread allelic imbalance around the two-thirds fraction and are probably in triploid state (Supplementary Fig. 9b).

Inferring tumor clonal architecture and evolutionary history. Single-cell RNA-seq is commonly used to examine a full spectrum of cell states within the tumor microenvironment, including different malignant, immune and stromal subpopulations whose classification is often unknown in advance. Therefore, reconstruction of single-cell copy number aberrations in heterogenous cell populations requires the simultaneous inference of clonal populations and genomic aberrations. In heterogenous tumors, cells with distinct genotypes can generally be assumed to have originated from a common cell of origin and are thus related to each other via a phylogeny. Their evolutionary relationships, if known, can be exploited to improve CNV detection by sharing information across cells in the same lineage[23]. On the other hand, given an estimated single-cell copy number profile, a CNV-based tumor phylogeny can be inferred[24,25].

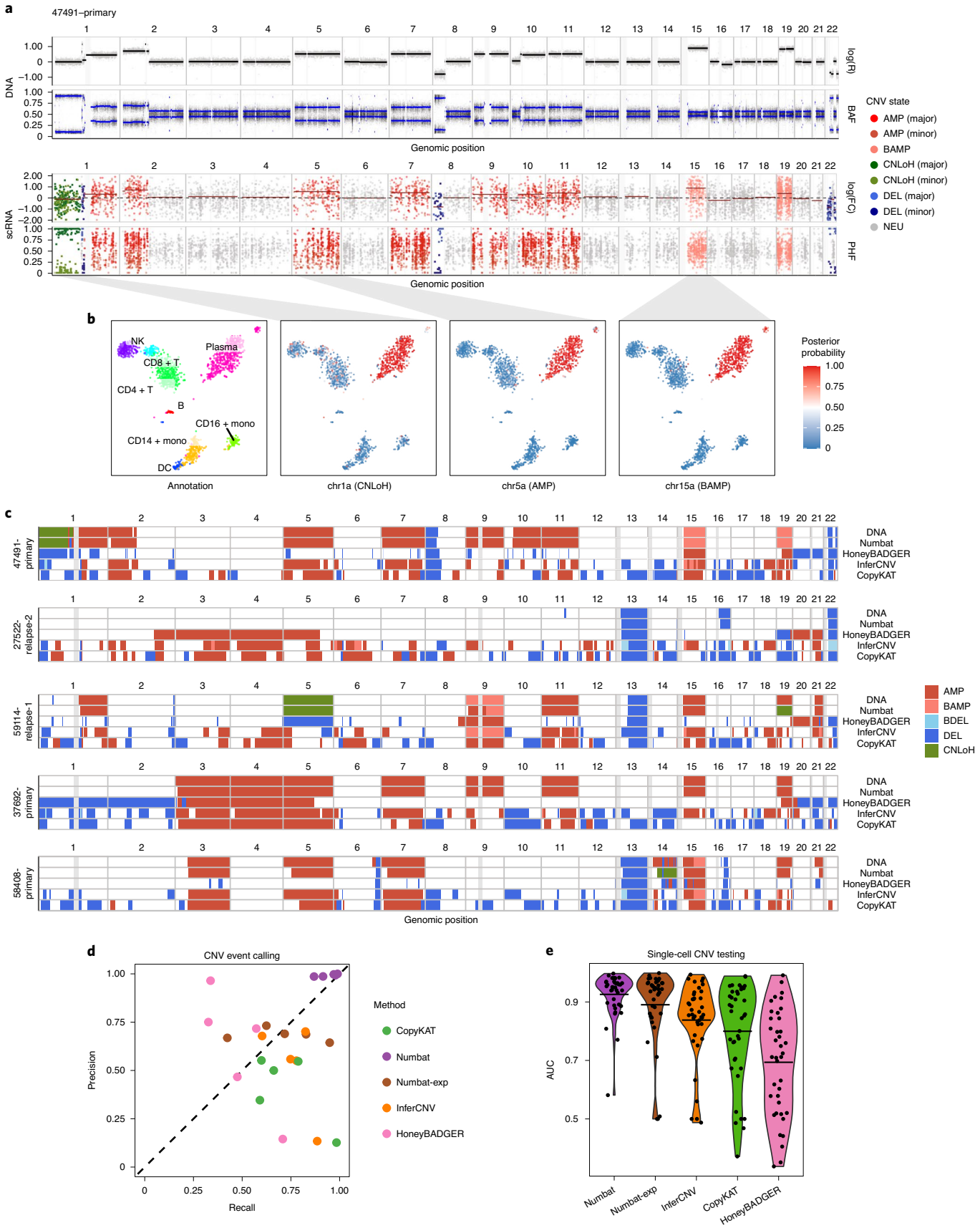
To perform joint inference of single-cell CNV profiles and the associated subclonal phylogeny, Numbat adopts an alternating optimization procedure. In each iteration, it first identifies CNVs in each branch of the clonal phylogeny using the joint HMM on pseudobulk expression and allele profiles (Fig. 3a). Cells are aggregated into pseudobulks by subtrees defined by lineage hierarchy, enabling detection of shared CNV events. CNV calls are then resolved into consensus segments based on the overlap and likelihood evidence (Supplementary Fig. 10). Numbat then evaluates the likelihood evidence for each unique event in individual cells using a Bayesian hierarchical model, producing a matrix of posterior probabilities of CNVs by cell (Fig. 3b). Next, to recover tumor clonal architecture, Numbat infers a single-cell lineage tree using a maximum-likelihood perfect phylogeny approach[26] (ScisTree), fully propagating the uncertainty in single-cell CNV calls (Fig. 3c). Genotype probabilities are used to search for an optimal tree topology using nearest-neighbor interchange (NNI), and mutations are placed on the tree based on maximum likelihood. Clonal populations with distinct genotypes can then be determined from the simplified mutational history (Supplementary Methods). Finally, Numbat uses the inferred single-cell phylogeny to form more precise lineage-specific pseudobulks, iteratively optimizing single-cell copy number profiles and tumor phylogeny. By default, Numbat initializes the phylogeny by hierarchical clustering of window-smoothed expression signals.

Reliable classification of tumor and normal cells. Precisely distinguishing malignant cells within heterogeneous cell mixtures is a well-established problem[3,6]. Because nonmalignant cells do not share aberrations with the tumor, the tumor population should be isolated as a distinct clade in the reconstructed phylogeny (Fig. 3c). To systematically benchmark Numbat's ability to recover this simplest clonal architecture and hence distinguish tumor from nonmalignant cells in the tumor microenvironment, we analyzed five triple-negative breast cancer (TNBC) samples and five anaplastic thyroid cancer (ATC) samples in addition to eight MM samples (Supplementary Table 1). We defined true tumor cell clusters based on the expression of well-established cell type or tumor-specific markers (*EPCAM* for TNBC[27], *KRT8* for ATC[28], *MZB1* for MM), as well as aneuploidy status (Methods). The tumor-versus-normal cell classification performance of Numbat was similar to that of CopyKAT in the two solid tumor panels, and substantially higher in the MM panel (Extended Data Fig. 4). The average classification accuracy for Numbat was 98.4% on TNBC and 98.5% on the ATC series, whereas CopyKAT produced an average accuracy of 98.1% and 98.5%, respectively (Extended Data Figs. 5 and 6). In the MM panel we found that Numbat maintained a stable performance (98.7%) whereas CopyKAT misclassified clusters of cells in five out of eight samples (Extended Data Fig. 7), resulting in lower accuracy (74.7%). The reduced performance of CopyKAT in the MM series is probably due to the lower sequencing coverage per cell and the less pronounced chromosomal aberrations in those samples. Numbat integrates two orthogonal lines of evidence (expression and

Fig. 2 | Numbat achieves accurate copy number inference via joint evaluation of gene expression, allele fraction and previous haplotype phasing information. a, DNA copy number profile of a MM sample juxtaposed with that inferred by the Numbat joint HMM. log(FC), log expression fold-change; NEU, neutral. Gray vertical bars represent centromeres and gap regions. **b**, Cell type annotation and posterior probability of CNV events in single cells visualized on *t*-distributed stochastic neighbor-embedding (t-SNE) of gene expression profiles. **c**, Copy number events detected by WGS, Numbat and other methods. Gray vertical bars represent gap regions. BAMP, balanced amplification; BDEL, balanced deletion. **d**, Performance of CNV event detection by different methods. Each dot represents a distinct sample. **e**, Performance of single-cell CNV testing by different methods. Each dot represents a distinct CNV event ($n = 39$). Shaded areas represent the estimated probability density. Center line, mean.

allele) for aneuploidy status in each cell, thereby enhancing signal and reducing the possibility of deriving erroneous conclusions from either source of information alone (Extended Data Figs. 5–7).

Haplotype-aware CNV analysis reveals subclonal complexity. Accurate detection of subclonal CNVs is a key challenge in characterization of tumor heterogeneity, because both allelic and



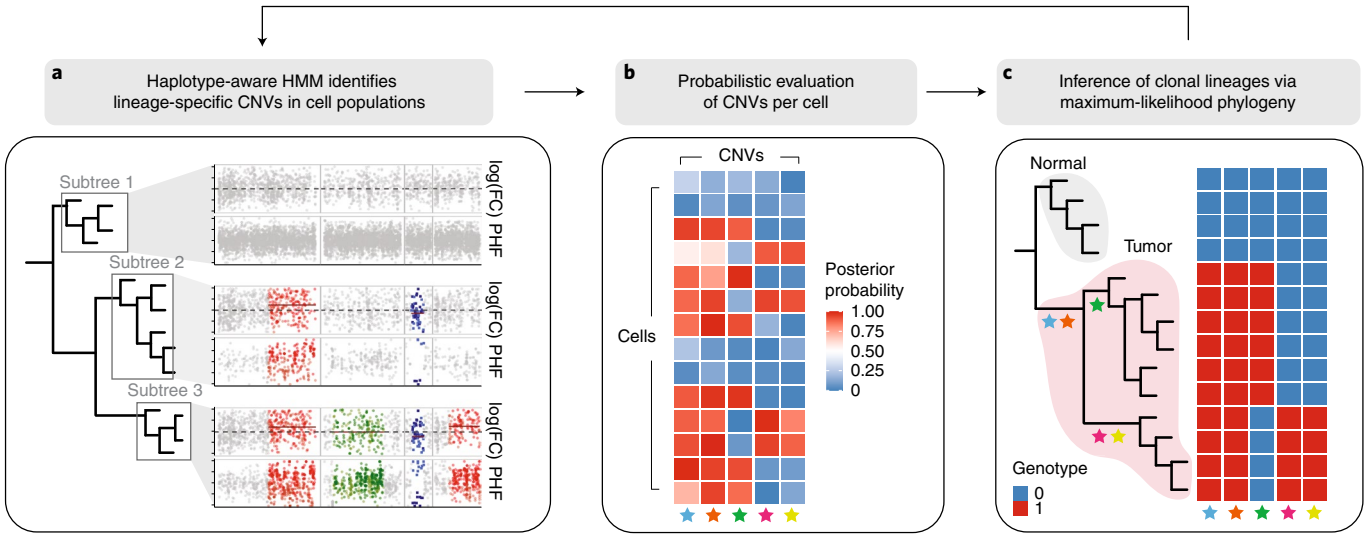


Fig. 3 | Iterative strategy to identify tumor subclones. a, Numbat aggregates data from single cells into pseudobulk profiles by major clades in the single-cell phylogeny, and runs a haplotype-aware HMM on each pseudobulk profile to identify lineage-specific CNVs. **b**, Numbat evaluates the presence of each CNV in each cell probabilistically using a Bayesian hierarchical model. **c**, Numbat then infers a maximum-likelihood phylogeny that captures the evolutionary relationships between single cells.

expression signals diminish with decreasing cellular fraction. Numbat's iterative inference of clonal populations and genomic aberrations should improve subclonal CNV estimation in genetically heterogenous cell populations. To systematically evaluate the extent to which the Numbat iterative strategy provides an advantage in the detection of subclonal CNVs, we applied it to tumor/normal mixtures at various proportions (10–90%) from the five MM samples with matched WGS. We found that the Numbat iterative approach outperformed pseudobulk HMM, as well as other methods, across different tumor cell fractions for both amplifications and deletions (Extended Data Fig. 8). To test Numbat's ability to resolve tumor subclonal structures, we analyzed a gastric cell line sample (NCI-N87) profiled by paired scRNA-seq and scDNA-seq[29]. From the scRNA-seq data, Numbat closely recapitulated the single-cell CNV landscape and subclonal architecture reconstructed by scDNA-seq (Extended Data Fig. 9). The accuracy of the consensus and subclone-specific CNV calls are robust to parameter variations (Supplementary Fig. 7c and Extended Data Fig. 9e). Similarly, the clonality predictions for most samples show high stability after the second iteration (Supplementary Figs. 11–13). The effect of the iterative update is most pronounced when the starting point is suboptimal (for example, initializing with one cluster or with random trees; Supplementary Figs. 12b,c and 13b,c).

Application of Numbat to TNBC and ATC datasets identified pronounced subclonal structures in four samples (TNBC1, TNBC5, ATC1 and ATC2; Fig. 4 and Extended Data Fig. 10). In particular, we found that allelic imbalances frequently contributed to the clonal complexity of tumors. For example, in TNBC1, Numbat inferred a branching phylogeny composed of two major subclonal lineages undergoing concurrent evolution (Fig. 4a). These two lineages share early CNLoH events on multiple chromosomes (for example, chromosomes 1p, 13, 14, 17 and 19; Fig. 4a). Numbat also identified subclonal CNLoH events on chromosomes 3p and 22q that are exclusive to the minor lineage (Fig. 4b,c). Such copy-neutral events do not exhibit deviations in expression magnitude and can be identified only through allele analysis (Supplementary Fig. 14a). In addition, Numbat revealed that the major lineage carries an imbalanced amplification on chr16 whereas the minor lineage carries an allelically balanced amplification on the same chromosome. Although both lineages carry an amplification on chr15 with similar increase in expression magnitudes (Fig. 4b), their haplotype frequencies appear to be mirrored (Fig. 4d), indicating that different homologous copies of the chromosome were duplicated in the evolutionary history of the two clones (Fig. 4e). Another example of an unusual clonal divergence pattern can be seen in ATC1. While the overall expression profile suggested that ATC1 harbors a relatively simple genome (Supplementary Fig. 14b), Numbat analysis revealed two diverging tumor lineages with reciprocal aberrations: while one subclone harbors an amplification on chr7 and a CNLoH on chr17, the other harbors a CNLoH on chr7 and an amplification on chr17 (Fig. 4f–i). Recent studies using scDNA-seq data revealed that such multiallelic and mirrored CNVs are prevalent sources of tumor heterogeneity[30,31]. These events, however, have not been previously inferred from scRNA-seq due to limited resolution in allele analysis and lack of signal in the overall expression profile. Such examples illustrate that the integration of phased allele data with expression signals can aid in the detection of subclonal alterations and lineage relationships, reflecting the dynamic clonal complexity of evolving tumors.

Earlier studies have shown that mitochondrial variants can also be used to detect subclonal populations in single-cell data[32,33]. We find that the distribution of the detected mitochondrial variants is consistent with the subclonal structure predicted by Numbat in the four samples examined above (Fig. 4a,f, Extended Data Fig. 10 and Supplementary Fig. 15). However, due to the sparse coverage of mitochondrial RNA from 3′ scRNA-seq protocols, we detected a low number of mutations (between one and nine) per sample, which were able to capture only a limited number of subclones.

Interplay between genetic and transcriptional heterogeneity. The decomposition of genetic subclones from scRNA-seq provides an opportunity to jointly characterize genetic and transcriptional heterogeneity during the course of tumor evolution. In particular, acquired copy number alterations can be used as natural genetic barcodes in conjunction with characteristic expression signatures to track the behaviors of clonal populations across time. We therefore applied Numbat to investigate the clonal evolutionary history of a therapy-resistant MM (patient no. 27522) with four sequential samples (primary, remission, first relapse and second relapse). Numbat identified three tumor subclones (g1–g3): one that harbors only

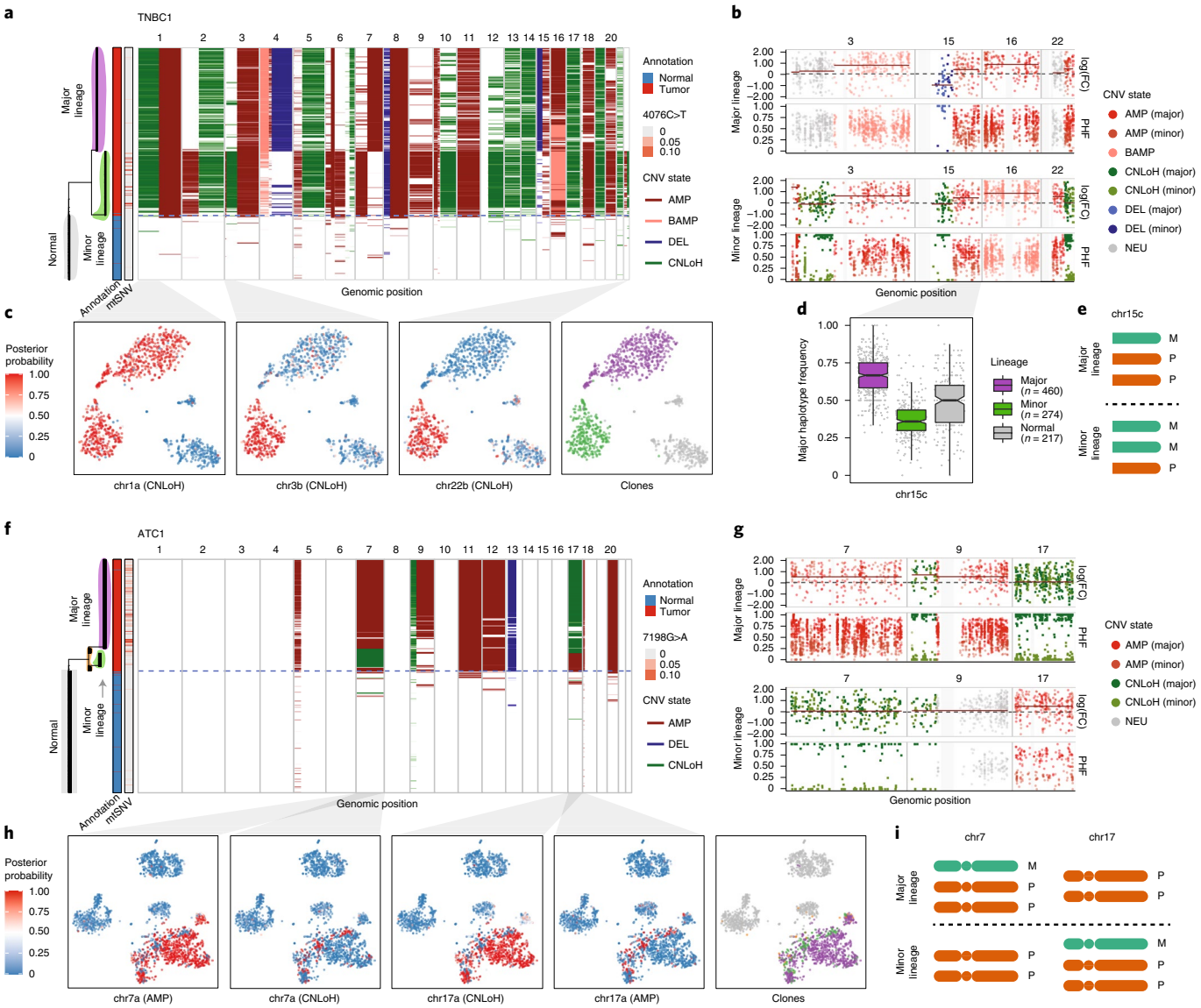


Fig. 4 | Numbat reveals additional complexity in tumor subclones through allele-specific copy number analysis. a, Single-cell CNV landscape and reconstructed phylogeny of TNBC1. Branch lengths correspond to the number of CNVs. Blue dashed line separates predicated tumor and normal cells. The first vertical bar on the left shows cell type ground truth while the second shows variant allele frequency of a clone-associated mitochondrial single-nucleotide variant (mtSNV). **b**, Pseudobulk CNV profile of major and minor lineages. Gray vertical bars represent centromeres and gap regions. **c**, Posterior CNV probability of shared and lineage-sepcific CNVs in t-SNE embedding of gene expression profiles. **d**, Major haplotype frequency in single cells. Only cells with at least five total allele counts in the region are shown. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5×IQR. **e**, Schematic of copy number state of chr15c in the major and minor lineages. M, maternal; P, paternal; the designation of maternal and paternal chromosomes is arbitrary. **f**, Single-cell CNV landscape and reconstructed phylogeny of ATC1. **g**, Pseudobulk CNV profile of major and minor lineages. **h**, Posterior CNV probability of subclonal multiallelic CNVs in t-SNE embedding of gene expression profiles. **i**, Schematic of copy number states of chr7 and chr17 in the major (top) and minor (bottom) lineages.

ancestral deletions on chromosomes 13 and 22 (g1), one that harbors an additional chr1p deletion (g2) and one that has acquired a chr16q deletion (g3; Fig. 5a–c). Both subclonal alterations are supported by DNA sequencing at the respective time points (Supplementary Fig. 16). At primary diagnosis the tumor comprised only clones g1 and g2, both of which appeared to be undetectable at the time of remission. However, clone g1 survived the therapy and reappeared at the first relapse. Furthermore, clone g1 also gave rise to clone g3, which continued to expand during subsequent therapy and became the dominant tumor subclone at the second relapse (Fig. 5c).

Tumor cells in the primary sample separated into two distinct expression-based clusters (e1 and e2; Fig. 5c). While the ancestral clone g1 is found in both e1 and e2, the derived subclone g2 appears to be restricted to cluster e1. This suggests that a large-scale shift in the transcriptional landscape gave rise to the two distinct tumor subpopulations (e1 and e2), which predated the chr1p deletion event within e1 (Fig. 5d). An alternative explanation is that, with the acquisition of chr1p deletion, g2 tumor cells lost the ability to enter transcriptional state e1. Integrating both aspects of heterogeneity, we resolved three main subpopulations in the primary sample: cells in expression cluster 1 with wild-type chr1 (e1g1), cells in expression cluster 1 with chr1p deletion (e1g2) and cells in expression cluster 2 (e2g1). Because g1 was the major cell population that re-emerged after remission, we asked whether it was derived from

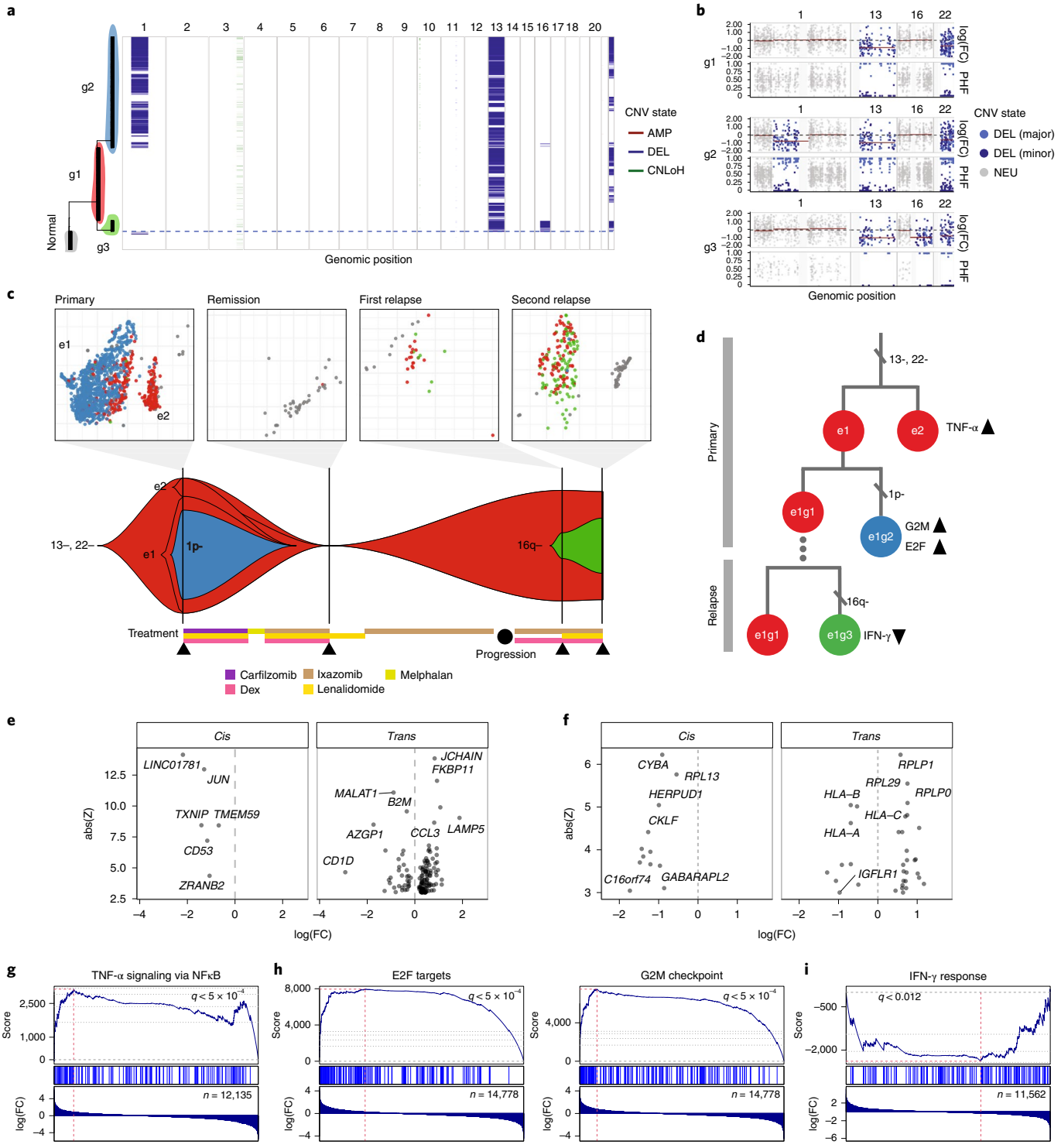


Fig. 5 | Tracking clonal evolution of a therapy-resistant MM using Numbat. a, Integrated single-cell CNV landscape and phylogeny of plasma cells from all four samples. **b**, Pseudobulk CNV profile of three main tumor subclones. Gray vertical bars represent centromeres and gap regions. **c**, Clonal evolutionary history integrating genetic and transcriptional alterations. Top: t-SNE embedding of gene expression profiles colored by genetic clones; embeddings are created separately for each sample. Only cells with >90% posterior classification confidence are shown. Bottom: change in tumor clonal composition over time. At each time point, only clones with >5% cellular fraction are shown. **d**, Genetic and transcriptional alterations in the proposed evolutionary history. **e**, Differentially expressed genes between e1g2 (observation) and e1g1 (reference) cells. **f**, Differentially expressed genes between e1g3 (observation) and e1g1 (reference) cells. **g**, Gene set enrichment analysis (GSEA) plot of the TNF-α signaling pathway in e2g1 relative to e1g1 cells. **h**, GSEA plots of E2F target and G2M checkpoint pathways in e1g2 relative to e1g1 cells. **i**, GSEA plot of the IFN-γ pathway in e1g3 relative to e1g1 cells. Dex, dexamethasone. abs(Z), absolute value of the differential expression Z score.

e1g1 or e2g1 cells in the primary sample. The g1 cells in the relapse sample carried the expression signatures of e1, as evidenced by the shared differentially expressed genes (Supplementary Fig. 17), indicating that the relapsed tumor probably originated from e1g1 cells in the primary sample (Fig. 5d).

We next investigated transcriptional differences between tumor subpopulations using differential expression and pathway enrichment analysis, separating probable *cis* (that is, genes residing within the CNV region) and *trans* (that is, genes residing outside of the CNV region) effects. Comparing e1 and e2 cells with the same copy number background (e2g1 versus e1g1) in the primary tumor, we found that e2 cells have higher activation of the tumor necrosis factor-α (TNF-α) signaling pathway (Fig. 5g and Supplementary Table 2). It has been shown that TNF-α triggers the release of interleukin 6, a myeloma growth factor, by activation of nuclear factor kappa B (NFκB)[34]. Comparing e1 cells with and without the chr1p deletion (e1g1 versus e1g2), we found that cells with the chr1p deletion have higher activation of pathways associated with cell cycle (G2M checkpoint and E2F targets), indicating a hyperproliferative phenotype (Fig. 5h and Supplementary Table 2). Differential gene expression analysis between e1g1 and e1g2 cells revealed six significantly differentially expressed genes in *cis* of the chr1p deletion event and 141 genes in *trans* (Fig. 5e). All six differentially expressed genes in *cis* of the deletion are significantly downregulated. The genes involved in the enriched pathways do not overlap significantly with the deleted region ($P=0.23$, E2F targets; $P=0.54$, G2M checkpoint, two-sided binomial test), indicating that those transcriptional changes may be driven by processes other than the CNVs we detected. The two genetic subclones in the second relapse sample (g1 and g3) do not separate into distinct expression clusters (Fig. 5c). Direct comparison of their expression patterns, however, revealed 12 significantly differentially expressed genes in *cis* and 34 in *trans* of the deletion (Fig. 5f), and showed that cells carrying the chr16q deletion have a significantly downregulated interferon gamma (IFN-γ) response pathway (Fig. 5i and Supplementary Table 2). Similar to the previous case, the genes involved in enriched pathways do not overlap significantly with the deleted region ($P=0.83$, two-sided binomial test). IFN-γ signaling plays an important role in tumor cell clearance by immune surveillance, and its dysregulation is associated with immune evasion and poor response to immunotherapy[35]. This is consistent with the more aggressive phenotype of clone g3, which achieved clonal dominance after several rounds of therapy (Fig. 5c).

Discussion

Tumor plasticity and the resulting therapy resistance can be driven by both genetic and nongenetic mechanisms, such as large-scale chromatin remodeling or aberrant activation of transcriptional programs[1,36]. The interplay between genetic and nongenetic mechanisms and their relative importance remains poorly understood. Methods that can reliably infer genetic alterations from a cell's transcriptome have the potential to illuminate these effects by characterizing both aspects of intratumoral heterogeneity at single-cell resolution.

Compared with DNA-based approaches, scRNA-seq provides limited coverage of alleles and suffers from transcriptional noise. Numbat attempts to address these challenges by incorporating previous haplotype information obtained from population-based phasing. We show that previous phasing information can be integrated with allele and expression signals in a HMM to enhance detection of subclonal copy number alterations from scRNA-seq data. The increasing availability of population-scale genetic data encompassing diverse ancestries should improve the power of this approach to patient samples from different genetic backgrounds[8,15,16]. The sensitivity of the Numbat haplotype-aware HMM can be further improved by more accurate haplotype information from other techniques, such as long-range haplotype phasing that takes advantage of individual relatedness[37] or experimental approaches that resolve haplotypes[38].

Reconstruction of the single-cell copy number profile from heterogenous cell populations requires simultaneous inference of clonal populations and genomic aberrations. Numbat solves this problem by iteratively inferring tumor phylogeny using detected aberrations and refining single-cell copy number estimates by exploiting the structure of the tumor phylogeny. Application to three tumor series (ATC, TNBC and MM) showed that Numbat precisely distinguished normal and malignant cells (marked by aneuploidy) in the tumor microenvironment and revealed additional subclonality within the tumor population. However, Numbat shares a common limitation with the existing methods in that determination of the number of confident subclones remains reliant on manual inspection of tumor phylogeny and copy number profile[2–5].

Tumor baseline ploidy estimation is a challenging problem in copy number analysis[22,39]. Existing methods infer copy number variations relative to the median ploidy, which can be confounded by hyperdiploidy or hypodiploidy[22]. Numbat attempts to address this problem by adopting a strategy previously developed for DNA analysis[22,40]. This approach proved effective, correctly identifying diploid regions in five tumor samples with WGS validation. However, challenges remain in tumors with genome-wide abberations (for example, TNBC1) or in those that have undergone whole-genome duplication, in which case manual curation is still necessary. Further improvements will be needed to robustly determine copy number baseline in tumors with complex copy number profiles.

Allele-specific CNV analysis has shown major advantages over total copy number analysis in studies of cancer genomes[30,31,41]. Although variations in chromosomal dosage are often discernible from large-scale gene expression changes, CNLoH events and haplotype-specific alterations can be detected only by using allele information. Numbat analysis of previously published tumor samples revealed additional subclonal complexity resulting from haplotype-specific alterations, highlighting the importance of allele-specific copy number analysis. Finally, to demonstrate the type of integrative analysis enabled by Numbat, we used it to characterize the genetic and transcriptional subpopulations in a serial MM sample. Comparison of the gene expression patterns of tumor subclones revealed that many of the transcriptional changes relevant to cancer progression and therapy resistance occur in *trans* and are not direct consequences of copy number aberrations. A variety of mechanisms, including other genetic mutations, epigenetic or regulatory changes, may mediate these effects. Dissection of their contribution to the expression state and overall phenotype of the cells remains a challenge. Among other advances, improved methods integrating genetic and epigenetic information will be needed to fully resolve the impact of genome instability on tumor cell states[30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01468-y.

Received: 23 January 2022; Accepted: 11 August 2022;
Published online: 26 September 2022

References

1. Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. & Baradaran, B. The different mechanisms of cancer drug resistance: a brief review. *Adv. Pharm. Bull.* **7**, 339–348 (2017).
2. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* **28**, 1217–1227 (2018).

3. Gao, R. et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**, 599–608 (2021).
4. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
5. Serin Harmanci, A., Harmanci, A. O. & Zhou, X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat. Commun.* **11**, 89 (2020).
6. Trinh, M. K. et al. Precise identification of cancer cells from allelic imbalances in single cell transcriptomes. *Commun. Biol.* **5**, 884 (2022).
7. Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.* **16**, 653–664 (2015).
8. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
9. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
10. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, e1007308 (2018).
11. Loh, P.-R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
12. Hujoel, M. L. A. et al. Influences of rare copy number variation on human complex traits. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.21.465308 (2021).
13. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
14. Vattathil, S. & Scheet, P. Haplotype-based profiling of subtle allelic imbalance with SNP arrays. *Genome Res.* **23**, 152–158 (2013).
15. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
16. The 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
17. Edsgärd, D., Reinius, B. & Sandberg, R. scphaser: haplotype inference using single-cell RNA-seq data. *Bioinformatics* **32**, 3038–3040 (2016).
18. Larsson, A. J. M. et al. Transcriptional bursts explain autosomal random monoallelic expression and affect allelic imbalance. *PLoS Comput. Biol.* **17**, e1008772 (2021).
19. Castel, S. E. et al. A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, 234 (2020).
20. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
21. Yau, C. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29**, 2482–2484 (2013).
22. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
23. Singer, J., Kuipers, J., Jahn, K. & Beerenwinkel, N. Single-cell mutation identification via phylogenetic inference. *Nat. Commun.* **9**, 5144 (2018).
24. Salehi, S. et al. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature*. 585–590 (2021)..
25. Dorri, F. et al. Efficient Bayesian inference of phylogenetic trees from large scale, low-depth genome-wide single-cell data. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.06.058180 (2021).
26. Wu, Y. Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics* **36**, 742–750 (2020).
27. Osta, W. A. et al. EpCAM is overexpressed in breast cancer and is a potential target for breast cancer gene therapy. *Cancer Res.* **64**, 5818–5824 (2004).
28. Guo, D. et al. Cytokeratin-8 in anaplastic thyroid carcinoma: more than a simple structural cytoskeletal protein. *Int. J. Mol. Sci.* **19**, 577 (2018).
29. Andor, N. et al. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. *NAR Genom. Bioinform.* **2**, lqaa016 (2020).
30. Wu, C.-Y. et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nat. Biotechnol.* **39**, 1259–1269 (2021).
31. Zaccaria, S. & Raphael, B. J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat. Biotechnol.* **39**, 207–214 (2021).
32. Kwok, A. W. C. et al. MQuad enables clonal substructure discovery using single cell mitochondrial variants. *Nat. Commun.* **13**, 1205 (2022).
33. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339 (2019).
34. Hideshima, T., Chauhan, D., Schlossman, R., Richardson, P. & Anderson, K. C. The role of tumor necrosis factor alpha in the pathophysiology of human multiple myeloma: therapeutic applications. *Oncogene* **20**, 4519–4527 (2001).
35. Castro, F., Cardoso, A. P., Gonçalves, R. M., Serre, K. & Oliveira, M. J. Interferon-gamma at the crossroads of tumor immune surveillance or evasion. *Front. Immunol.* **9**, 847 (2018).
36. Alekseyenko, A. A. et al. The oncogenic BRD4-NUT chromatin regulator drives aberrant transcription within large topological domains. *Genes Dev.* **29**, 1507–1523 (2015).
37. O'Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
38. Tourdot, R. W., Brunette, G. J., Pinto, R. A. & Zhang, C.-Z. Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol.* **22**, 139 (2021).
39. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14**, R80 (2013).
40. Zaccaria, S. & Raphael, B. J. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nat. Commun.* **11**, 4301 (2020).
41. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).

Methods

Preprocessing of scRNA-seq data. We used the Cell Ranger (v.6.0.2, 10x Genomics) software suite to process raw FASTQ or BAM files obtained from previously published studies. We included only those cell barcodes present in the gene expression count matrices or cell type annotation provided with the original publication. We used conos[42] (v.1.4.1) to perform multisample integration, clustering and generation of graph embeddings.

Genotyping and phasing from scRNA-seq data. To identify heterozygous and homozygous germline SNPs we used cellsnp-lite[43] (v.1.2.2) to generate allele counts for a panel of known common SNPs (population allele frequency >5%). SNPs with variant allele frequency (aggregating all cells) 0.1–0.9 were identified as heterozygous; SNPs with ten or more reads covering the alternate allele with variant allele frequency = 1 were identified as homozygous. We then used Eagle2 (v.2.4.1) to phase the identified heterozygous SNPs using the 1000G and TOPMed reference panels.

Coexpression-based phasing. To perform phasing using single-cell expression data we used the previously published *scphaser* package, which phases heterozygous alleles based on their coexpression patterns[17]. We ran *scphaser* with a minimum number of reads of one (min_acount = 1) and a fold-change cutoff of three (fc = 3) for genotyping and then phased the alleles using the exhaustive search mode (method = 'exhaust'), with allele counts as input (input = 'ac') and no weighting based on allele counts (weigh = FALSE).

Statistical modeling of expression and allele data. We formulated a generative model for the observed UMI counts per gene and observed allele counts per SNP site (Extended Data Fig. 2). This model generalizes to both pseudobulk and single-cell setting. We aim to infer the DNA state for each marker, denoted as $g = (c_p{:}c_m)$ where c_m is the number of maternal copies and c_p is the number of paternal copies. Note that, in single cells, c_p and c_m can take any non-negative integer value. For example, in diploid regions g = (1:1) whereas in a heterozygous loss of the maternal chromosome, g = (1:0). Since a pseudobulk can contain a mixture of cells in both diploid state and altered state, c_p and c_m can take any continuous value in the non-negative domain. For convenience, we reparameterize g as the change in total chromosome dosage relative to the diploid state (ϕ) and haplotype fraction (θ) as follows:

$$\phi = \frac{c_m + c_p}{2}, \quad \theta = \frac{c_p}{c_m + c_p}$$

which are the targets of inference. Note that, in single cells, ϕ and θ take on discrete values. In pseudobulks, $\phi \in [0,\infty]$ and $\theta \in [0,1]$, which depend on both the mixture proportion and underlying genotype (Supplementary Fig. 6).

We observe two types of marker: expression counts per gene and allele counts per SNP; gene expression counts are emitted only once per gene whereas allele counts are emitted at each SNP. Let N be the total number of genes measured in the transcriptome. For gene i we denote the gene expression count as X_i, which we model using a Poisson log-normal (PoisLogNorm) distribution:

$$X_i \sim \text{PoisLogNorm}\left(\mu + \log\left(l\lambda_i^*\right) + \log\phi, \sigma^2\right) \quad (1)$$

Here l is the total library size and λ_i^* is the baseline expression magnitude for gene i in the reference profile. Shared between all genes, μ and σ^2 are hyperparameters representing bias and variance, respectively, in log expression fold-change between the observation and reference profiles. The hyperparameters μ and σ^2 are unknown a priori and must be empirically estimated for each cell or pseudobulk with respect to a specific reference profile. Restricting to genes in diploid regions, the maximum-likelihood estimates of μ and σ are:

$$(\hat{\mu}, \hat{\sigma}) = \text{argmax}_{\mu,\sigma} \prod_{i=1}^{N} \text{p}\left(\text{X}_\text{i} | \lambda_\text{i}^*, l, \phi = 1, \mu, \sigma^2\right) \quad (2)$$

where argmax stands for arguments of the maxima. These baseline parameters are then used to configure the emission probabilities for CNV detection.

For allele data we use Y_j to denote the observed variant allele count of the jth SNP and m_j to denote the total allele count (sum of reference and variant allele counts). Once the variant alleles are phased, Y_j is the paternal allele count. We model paternal allele count for SNP j using a beta-binomial (BetaBinom) distribution:

$$Y_j \sim \text{BetaBinom}\left(m_j, \theta\gamma, (1-\theta)\gamma\right) \quad (3)$$

where γ is a hyperparameter that represents inverse overdispersion in allele counts.

Phase switch probabilities. We model the occurrence of phase switch errors from population-based haplotype phasing along the genome using a Poisson process with uniform rate ν. Between two adjacent SNPs with genetic distance (in cM) d, the number of phase switches W can be modeled by Poisson distribution:

$$\text{W} \sim \text{Poisson}(\nu d)$$

The probability of two SNPs being discordant in phase is therefore a function of genetic distance:

$$p_s(d) = \sum_{w=1,3,5,\ldots} \frac{(\nu d)^w e^{-\nu d}}{w!} = \frac{1 - e^{-2\nu d}}{2} \quad (4)$$

In practice, we fix $\nu = 1$ to predict phase switch probabilities based on genetic distance.

Haplotype-aware HMM. We designed an HMM that integrates expression deviation and haplotype imbalance signals to detect CNVs in cell population pseudobulk profiles. Depending on copy number configuration, cellular fraction and haplotype state (major or minor), each aberrant copy number state can exhibit a continuum of expression fold-changes ϕ and haplotype fractions θ (dashed lines in Supplementary Fig. 6). We therefore define a set of discrete hidden states $z \in \mathcal{Z} = \{1, 2, \ldots, 15\}$ to capture the joint behavior of (ϕ,θ) across the continuous space of CNV signals (black dots in Supplementary Fig. 6). Each of the 15 states emits a gene read count X_i and a paternal allele count Y_j according to the probability mass functions specified by equations (1) and (3), with the associated state parameters (ϕ_z,θ_z). That is,

$$X_i | Z_i = z \sim \text{PoisLogNorm}\left(\mu + \log\left(l\lambda_i^*\right) + \log\phi_z, \sigma^2\right)$$

$$Y_j | Z_j = z \sim \text{BetaBinom}\left(m_j, \theta_z\gamma, (1-\theta_z)\gamma\right)$$

The transition probabilities are specified by t and p_s, where t is the transition probability between copy number states and p_s is the transition probability between haplotype states (that is, phase switch probability between major and minor haplotypes; Extended Data Fig. 1). The term t is homogeneous in the Markov chain whereas p_s is site specific. To reflect LD decay, we model p_s as a monotonically increasing function of genetic distance from the previous SNP according to equation (4). The full transition matrix of the joint HMM can be found in Supplementary Table 3.

To infer hidden copy number states we use the Viterbi algorithm to identify the most probable copy number state for each marker position. Because contiguous genomic segments can occupy distinct copy number states, which cannot be captured by a single set of ϕ and θ, we use one set of minimum-threshold parameters ($\log\phi_{\min}$ and $\theta_{\min}$) to initially identify all detectable CNVs with various deviation magnitudes. Intuitively, lower threshold choices favor detection of more subclonal events. By default, we fix $\log\phi_{\min} = 0.25$ and $\theta_{\min} = 0.08$. To avoid oversegmentation resulting from large local deviations, we rejoin any segments containing fewer than ten genes with adjacent segments to obtain the final segmentation. The true underlying dosage ratio and haplotype frequency are event-specific and are estimated separately for each CNV segment by maximization of total model likelihood. Finally we obtain the haplotype classification of major/minor alleles based on posterior marginal probability at each SNP, computed from the forward–backward algorithm using the maximum-likelihood estimates of (ϕ,θ).

Testing for multiallelic CNVs. A CNV is determined as multiallelic if it is confidently (alpha level of 10^{-4}) assigned to distinct CNV types in different subclone pseudobulk profiles.

Single-cell CNV evaluation. We make inferences on the underlying genotype of individual cells jointly using the observed expression and allele counts. First, using the diploid regions identified in pseudobulk analysis, we estimate the cell-specific expression fold-change bias and variance (μ and σ^2) by maximum likelihood according to equation (2). In cases where diploid regions contain <5% of the genes, we include genes in CNLoH regions to estimate μ and σ^2. In a given genomic region of a given cell, the posterior probability of each genotype is obtained by

$$p(G = g|\mathbf{X}, \mathbf{Y}) = \frac{\prod_i p(X_i|G = g) \prod_j p\left(Y_j|G = g\right) p(G = g)}{\sum_{g \in \mathcal{G}} \prod_i p(X_i|G = g) \prod_j p\left(Y_j|G = g\right) p(G = g)}$$

where likelihood functions are defined according to the generative model described before. Posterior alteration type probabilities from the pseudobulk analysis are propagated as single-cell genotype priors. We note that $\mathbf{Y}$ represents phased allele counts using posterior haplotypes obtained from the HMM, which takes into account both previous phasing information and the observed allele frequencies of each SNP. The posterior haplotype should span the entire CNV event and allow aggregation of allele counts across the whole region. Because the effect of allele-specific expression is minimal when aggregating across a large number of genes, we simply use a binomial likelihood for allele counts (that is, $\gamma = \infty$ in the beta-binomial model). Although both maternal and paternal copy number can take any non-negative integer value in single cells, in practice we consider only seven possible genotypes: $g \in \{(1:1), (2:0), (1:0), (2:1), (3:1), (2:2), (0:0)\}$.

CNV filtering. To reduce the number of false-positive CNV calls, we filter the events called from the Numbat joint HMM based on statistical evidence. We define the log-likelihood ratio (LLR) of a CNV event in a pseudobulk profile as

$$\mathrm{LLR} = \mathrm{LLR}_x + \mathrm{LLR}_y = \log\left(\frac{p\left(\mathbf{X}|G=g\right)}{p\left(\mathbf{X}|G=(1:1)\right)}\right) + \log\left(\frac{p\left(\mathbf{Y}|G=g\right)}{p\left(\mathbf{Y}|G=(1:1)\right)}\right)$$

We define the entropy of the posterior distribution of a CNV event in single cells as

$$H\left(p\right) = -p\log_2\left(p\right) - \left(1-p\right)\log_2\left(1-p\right),$$

which captures the degree of uncertainty in the inference.

Maximum-likelihood phylogeny inference using uncertain genotypes. We implement a modified version of a recently described approach (ScisTree[26]) to infer a maximum-likelihood perfect phylogeny based on uncertain genotypes. Using the cell-by-CNV genotype probabilities previously obtained, we compute a distance matrix between cells using the Euclidean distance measure. We then construct two candidate trees using the neighbor-joining (NJ) and unweighted pair group method with arithmetic mean (UPGMA) algorithms. The candidate tree with the higher genotype likelihood (as defined in ref. [26]) is used as the initial tree. We then search for an optimal tree topology that maximizes the genotype likelihood using the NNI algorithm.

Posterior assignment of cells to copy number profiles and clades. Given K genomic segments, we denote copy number profile j by $\mathbf{C}_j = \left(g_j^1, g_j^2, \ldots, g_j^K\right)$. We can obtain the posterior probability that a given cell harbors copy number profile j by

$$p\left(\mathbf{C}_j|\mathbf{X},\mathbf{Y}\right) = \frac{\prod_k p\left(\mathbf{X}_k|g_j^k\right)\prod_k p\left(\mathbf{Y}_k|g_j^k\right)p\left(\mathbf{C}_j\right)}{\sum_j\prod_k p\left(\mathbf{X}_k|g_j^k\right)\prod_k p\left(\mathbf{Y}_k|g_j^k\right)p\left(\mathbf{C}_j\right)}$$

For example, the posterior probability that a cell is diploid in every region is $p\left(\mathbf{C}_0|\mathbf{X},\mathbf{Y}\right)$, where $\mathbf{C}_0 = \left(\left(1:1\right),\left(1:1\right),\ldots,\left(1:1\right)\right)$. The posterior probability that a cell belongs to a specific clade (in particular, the tumor lineage) in the phylogeny is then equal to the sum of the probabilities that it harbors each of the possible genotypes included in the clade.

WGS copy number analysis. We used hmftools[44] to perform unmatched CNV analysis of the WGS data from the MM dataset. The modules COLBALT (v.1.11) and AMBER (v.3.5) were used to obtain log read depth ratios (log(R)) and BAF profiles, respectively. The module PURPLE (v.3.2) was used to determine total copy number, tumor ploidy and purity. We performed resegmentation of log(R) data using the *pcf* function in the R package copynumber[45] (v.1.32.0), with a gamma parameter of 12,000. Significantly altered segments were determined by a threshold of $\log(R) > 0.25$, $\log(R) < -0.25$ and $\mathrm{BAF} > 0.75$ for amplifications, deletions and CNLoH, respectively.

Single-cell DNA-seq copy number analysis. We used CopyKit (v.0.1.1; https://github.com/navinlabcode/copykit) to perform preprocessing, quality control and analysis of scDNA-seq data. For each cell, read coverage was collected for variable-length genomic bins with a resolution of 220 kb[46]. Segmentation was performed using the CBS algorithm ($\mathrm{alpha} = 1\times10^{-9}$), and integer copy number calls were derived using a ploidy of 1.94 as reported in the original publication[29]. Using integer copy number calls, we performed hierarchical clustering using Manhattan distance and Ward2 linkage. A normal cell with diploid genome was added as an outgroup to root the tree.

Benchmarking the effect of population-based phasing on the detection of allelic imbalance. Using the cell annotations of TNBC4 from the original paper, we created subsampled datasets (total of 500 cells) comprising different tumor cell fractions. We defined chromosome arms with complete LoH using allele frequencies in the all-tumor pseudobulk ($\mathrm{MAF} > 0.95$; Supplementary Fig. 3). Using this setup, we performed three sets of benchmarking experiments. First, to benchmark the effect of previous phasing on the detection of subclonal allelic imbalance from heterogenous cell populations, we randomly sampled genomic segments of fixed length (10 Mb) from known aberrant regions for each mixture proportion. We additionally sampled segments from the all-normal pseudobulks to serve as true-negative examples. We then scored the allele profile of each sampled segment using haplotype-naive and -aware HMMs. Using these scores, we calculated an AUC for each tumor/normal mixture proportion. Second, to benchmark the effect of previous phasing on allele classification (major versus minor haplotype) from mixture pseudobulks, we defined ground truth haplotypes in known LoH regions using the observed BAFs in the all-tumor pseudobulk ($\mathrm{BAF} < 0.5$, minor; $\mathrm{BAF} \geq 0.5$, major). We classified alleles using haplotype-naive and -aware HMM for each cell mixture, then calculated the proportions of alleles correctly classified as a measure of model performance. Third, to benchmark the effect of previous phasing on single-cell event detection, we split the cells into training (70%) and testing sets (30%). We classified alleles using the two models in known aberrant regions with pseudobulk profiles created using cells from the training set, then used the obtained haplotypes to calculate CNV probabilities in single cells from the test set. The existing tumor-versus-normal annotations were used as ground truth labels for each cell and each event. We calculated an overall AUC (aggregating across events) for each tumor/normal mixture fraction. Pseudobulk and single-cell CNV detection benchmarks were also performed on the MM dataset, where LoH and amplification events were defined using the matched WGS for each sample.

Benchmarking CNV detection accuracy. We evaluated the overall copy number profile reconstruction quality by Numbat and three other methods (CopyKAT v.1.0.8, InferCNV v.1.8.1 and HoneyBADGER v.0.1) using five MM samples (from distinct patients) with sample-matched, flow-sorted WGS. Because Numbat, HoneyBADGER and InferCNV identify CNVs from pseudobulks, we supplied the pseudobulk profile from all tumor cells. For CopyKAT we summarized the consensus tumor copy number profile by averaging copy number intensities for each genomic bin across all tumor cells. Because CopyKAT does not explicitly call copy number events, we applied thresholds of +0.03 and −0.03 to identify amplified and deleted segments, respectively. For HoneyBADGER we used a minimum deviance threshold of 0.1 for expression HMM and included all heterozygous SNPs as input to the allele HMM. We took the union of events identified by the allele- and expression-based approaches. For InferCNV we used the recommended parameters for 10X (denoise = TRUE, cutoff = 0.1) and performed CNV calling using the 'consensus' i6 HMM mode. All other parameters were retained as the default setting otherwise, and the Human Cell Atlas (HCA) lung collection was used as diploid reference[47]. When cell-type-specific references could not be provided as input, we supplied the averaged expression profile. To evaluate CNV detection performance we computed precision and recall based on the extent of overlap between the predicted and true aberrant regions as defined by WGS. All types of event were considered (amplifications, deletions, CNLoH). To benchmark single-cell CNV testing accuracy, we first defined the boundaries and alteration types of individual CNV events from the DNA profile for each sample. We did not include regions that appeared to be affected by either complex events (for example, chr14 of 58408-primary; Extended Data Fig. 3) or subclonal events (for example, chr16q deletion in 27522-relapse-2; Extended Data Fig. 3) as judged from the DNA profiles. We then computed a score of each event for each individual cell using the four different methods. For Numbat and HoneyBADGER, the event posterior probability was used as the score; for InferCNV and CopyKAT, we defined the score as the average smoothed expression intensity in the region affected by the event. Scores of CNLoH events were set to 0 for all allele-agnostic approaches. As an approximation of the single-cell genotype ground truth, we assumed that CNV events are present in all tumor cells and absent in all normal cells. For each event, we calculated an AUC based on the single-cell event scores from each method.

Benchmarking tumor-versus-normal cell classification accuracy. We identified true tumor cells in the three datasets based on the combined evidence of expression-based clustering, cell type or tumor-specific marker expression and aneuploidy evidence. For the ATC and TNBC series, the tumor-versus-normal cell labels from the original publication were used and expression of tumor-specific markers (*EPCAM* for TNBC, *KRT8* for ATC) was used as visual reference in Extended Data Figs. 5 and 6. We excluded ATC5 from the benchmark due to the lack of clear expression of *KRT8*. For the MM series, we used the cell type annotation from the original study to identify malignant plasma cells and the expression of *MZB1* as visual reference in Extended Data Fig. 7. In one of the samples (27522-relapse-2), both normal and malignant plasma cells were present and the malignant plasma cell cluster was identified by upregulated *FGFR3* expression (due to t(4;14) translocation) as described in the original publication[48]. To evaluate performance, we calculated classification accuracy based on the ground truth labels and predictions made by the two methods. For Numbat, cells with aneuploidy probability >0.5 were designated as tumor and normal otherwise. For CopyKAT, the tumor/normal predictions from the original paper were used for the TNBC and MDA series; for the MM dataset, predictions were generated by running CopyKAT using the default parameters and the same expression reference supplied to Numbat.

Numbat run parameters. Numbat was run using default parameters unless otherwise specified ($\log\phi_{\min} = 0.25$, $\theta_{\min} = 0.08$, $\gamma = 20$, transition probability $t = 10^{-5}$, maximum cost $\tau = 0.3$, initial number of clusters $k = 3$, CNV overlap tolerance 0.45, minimum pseudobulk size 50 cells, LLR threshold 5, entropy threshold 0.5, maximum of two iterations). For TNBC1, because shared diploid regions could not be identified we manually supplied chromosomes 13, 14 and 19 (containing CNLoH) as baseline to Numbat. Because NCI-N87 is a cell line sample and does not contain normal diploid cells, we used the SNP density HMM (Supplementary Methods) to detect clonal LoH regions with a transition probability of $t = 10^{-4}$. For the longitudinal analysis of patient no. 27522 presented in Fig. 5, we used normal B cells from the same patient as expression reference. The HCA lung collection was used as the expression reference for all other analyses[47].

Gene set enrichment analysis. We used the LIGER R package (v.2.0.1) to perform gene set enrichment analysis between cell populations. Hallmark gene sets ($n = 50$) were obtained from MSigDB[49]. Only genes with at least one read count in at least five cells were used as input. A total of 10,000 random permutations were used to compute empirical P values. We used the Holm–Bonferroni method to adjust for multiple comparisons within each analysis. Significantly enriched gene sets were filtered by $q < 0.05$ and the sign of the edge value consistent with enrichment direction—that is, positive enrichment is consistent with a positive edge value and negative enrichment is consistent with a negative edge value.

Differential gene expression analysis. We used the Mann–Whitney U-test implemented in pagoda2 (v.1.0.9) (ref. [50]) to identify confident differentially expressed genes between subclones. We used default parameter settings, with a z-score threshold of 3.

Identification of transcribed mitochondrial mutations. We applied the MQuad method[32] (v.0.1.6) to identify mitochondrial RNA mutations from scRNA-seq samples. We used the default parameters recommended for 10X data (minimum depth = 5). We filtered variants by variant allele frequency >5% in more than five tumor cells.

Statistical analyses and visualization. Custom statistical analyses and visualizations were performed in R (v.4.1.2). The fishplot package[51] (v.0.5.1) was used to visualize tumor clonal structures.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The scRNA-seq and WGS validation data from the WASHU MM study can be accessed through SRA (PRJNA694128). The scRNA-seq data from the MDA CopyKAT study can be accessed through GEO (GSE148673) and SRA (PRJNA625321). The NCI-N87 scDNA-seq and scRNA-seq datasets are available on GEO (GSE142750) and SRA (PRJNA498809). The HCA collection of reference expression profiles can be obtained from Synapse under ID syn21041850. The 1000G phasing panel can be downloaded from the IGSR FTP site (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release). The TOPMed phasing panel can be accessed through the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/).

Code availability

The Numbat algorithm is available at https://github.com/kharchenkolab/numbat. The analysis scripts and notebooks used to reproduce results included in the paper are available at https://github.com/kharchenkolab/NumbatAnalysis.

References

42. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
43. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).
44. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
45. Nilsen, G. et al. Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
46. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
47. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
48. Liu, R. et al. Co-evolution of tumor and immune cells during progression of multiple myeloma. *Nat. Commun.* **12**, 2559 (2021).
49. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
50. Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241–244 (2016).
51. Miller, C. A. et al. Visualizing tumor evolution with the fishplot package for R. *BMC Genomics* **17**, 880 (2016).

Acknowledgements

P.V.K., R.S. and T.G. were supported by Synergy grant no. 85629 (KILL-OR-DIFFERENTIATE) from the European Research Council. P.-R.L. was supported by NIH grant no. DP2 ES030554, a Burroughs Wellcome Fund Career Award at the Scientific Interfaces and the Next Generation Fund at the Broad Institute of MIT and Harvard.

Author contributions

P.V.K. and T.G. formulated the study and the overall approach. A.K. carried out proof-of-concept tests of population-based phasing. T.G. developed the detailed algorithms with advice from P.V.K., R.S., H.S. and P.-R.L. T.G. implemented the Numbat package with help from E.B. T.G. and P.V.K. drafted the manuscript. All authors provided suggestions and corrections on the manuscript text.

Competing interests

P.V.K. is an employee of Altos Labs and serves on the Scientific Advisory Board to Celsius Therapeutics, Inc. and Biomage, Inc. The remaining authors declare no competing interests.

Additional information

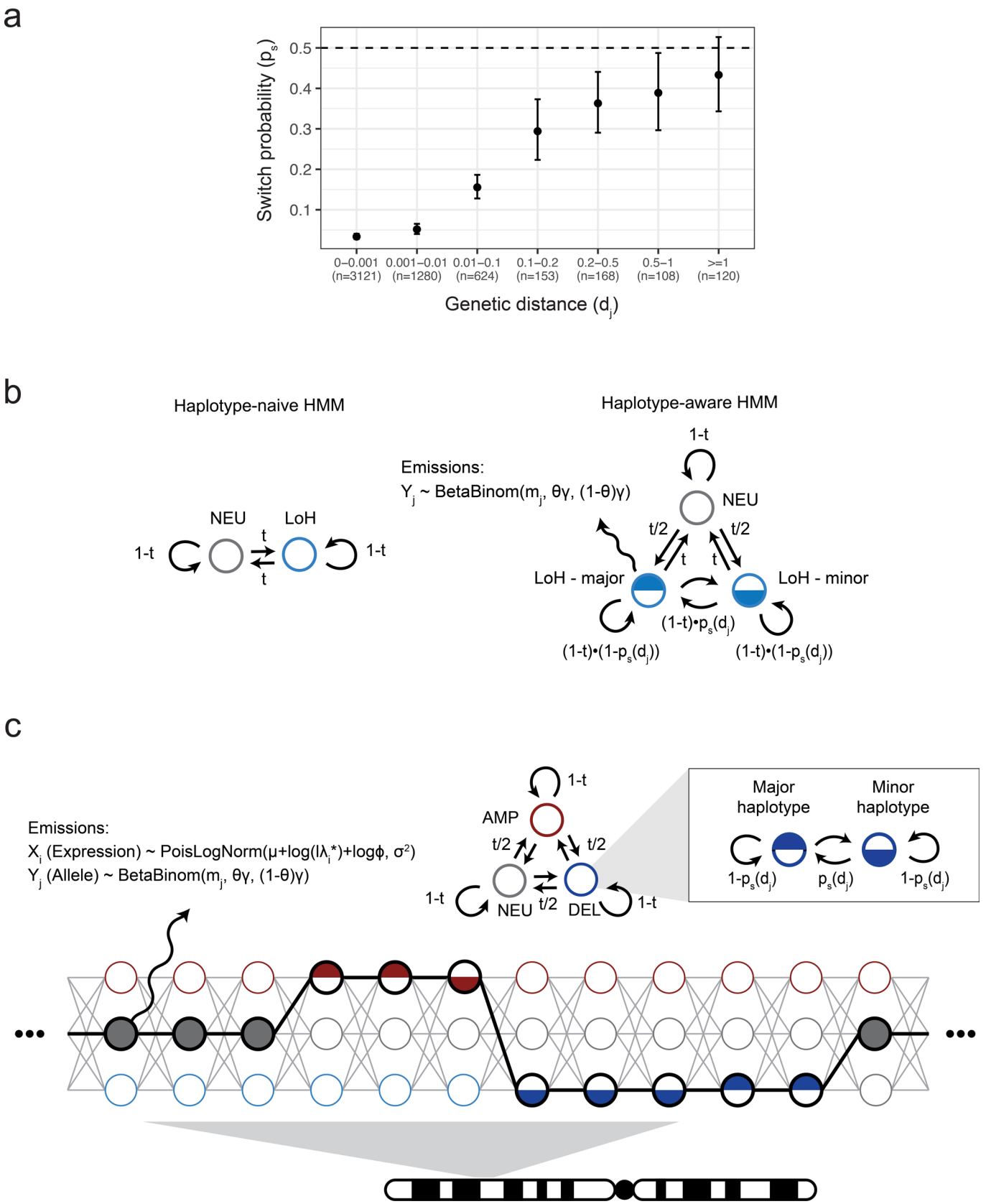
Extended data are available for this paper at https://doi.org/10.1038/s41587-022-01468-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01468-y.

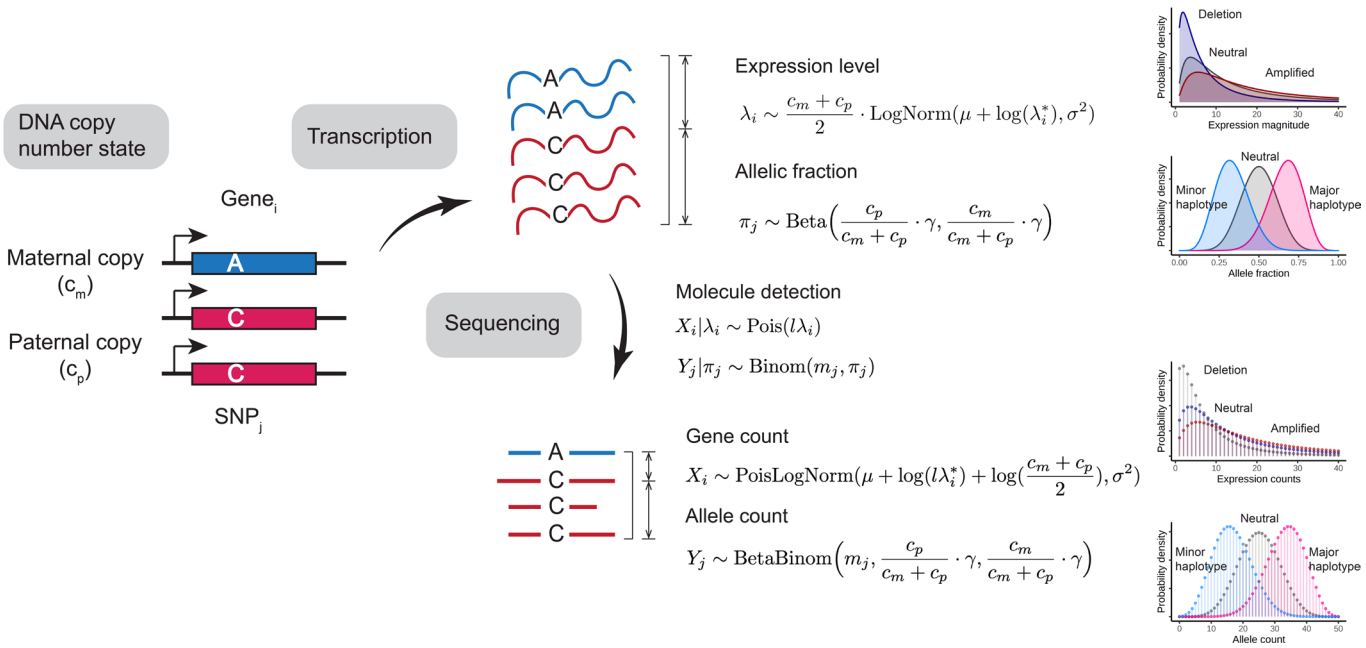
Correspondence and requests for materials should be addressed to Peter V. Kharchenko.

Peer review information *Nature Biotechnology* thanks Woong-Yang Park and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

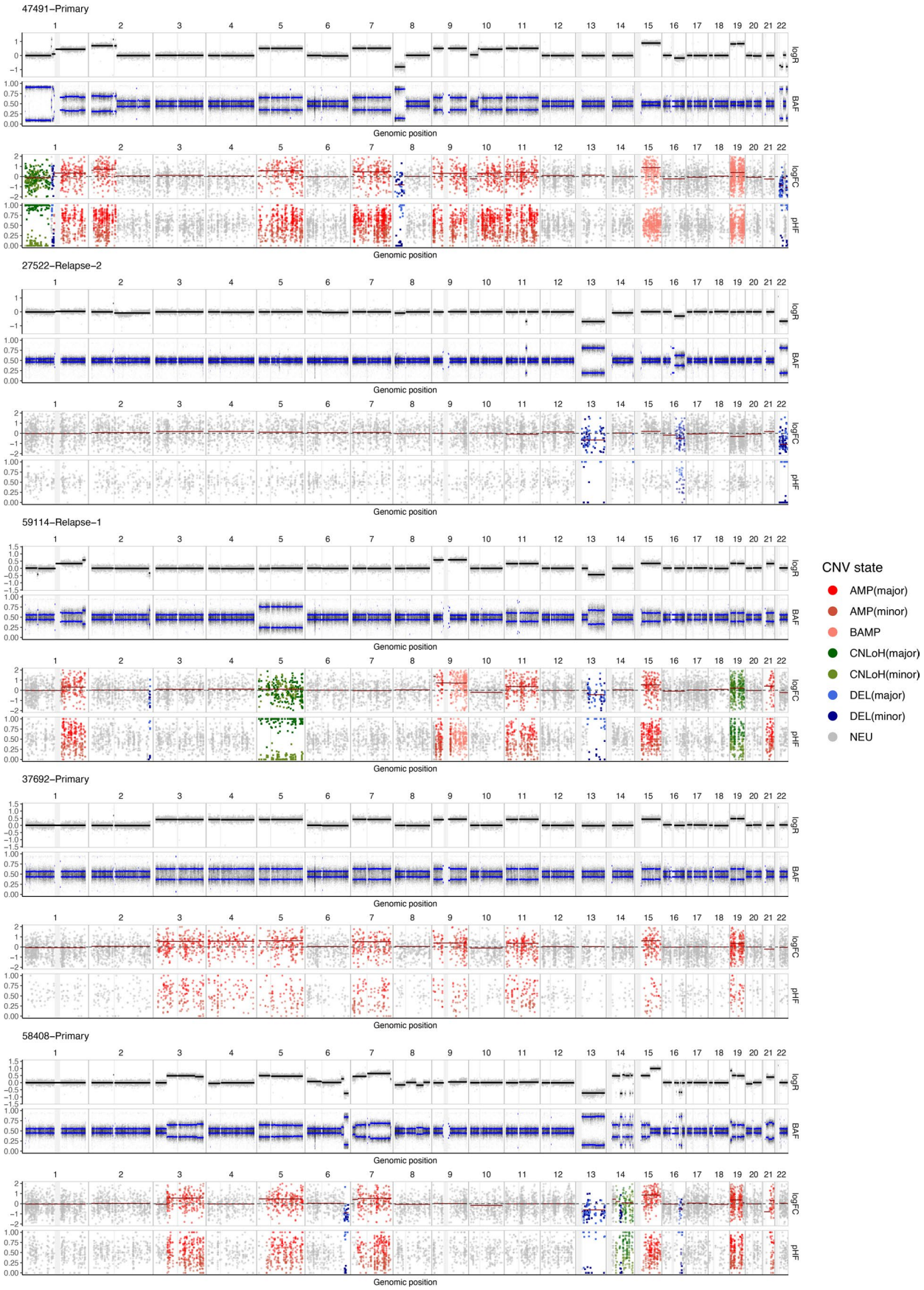
Reprints and permissions information is available at www.nature.com/reprints.



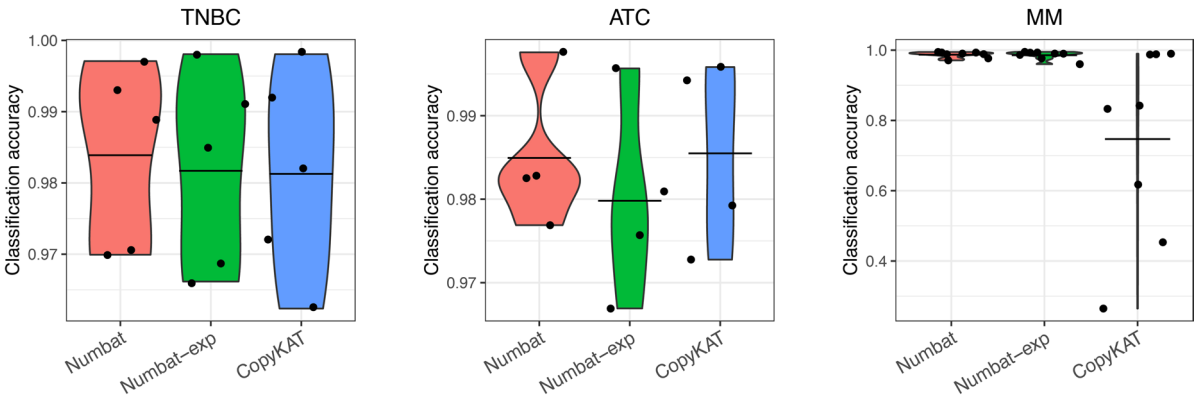
Extended Data Fig. 1 | Haplotype-aware Hidden Markov models. a, Phase switch probability as a function of genetic distance, estimated from alleles phased from LoH regions in TNBC4. Genetic distance is measured in centimorgan (cM). Error bar represents 95% CI derived from a binomial test. The center of the error bar represents the observed fraction of phase switches. **b**, Schematic of conventional and haplotype-aware allele HMM. *t*, copy number state transition probability. p_s, phase transition probability. **c**, Schematic of the Numbat joint HMM. Only three copy number states (neutral, deletion, amplification) are included for illustrative purposes.



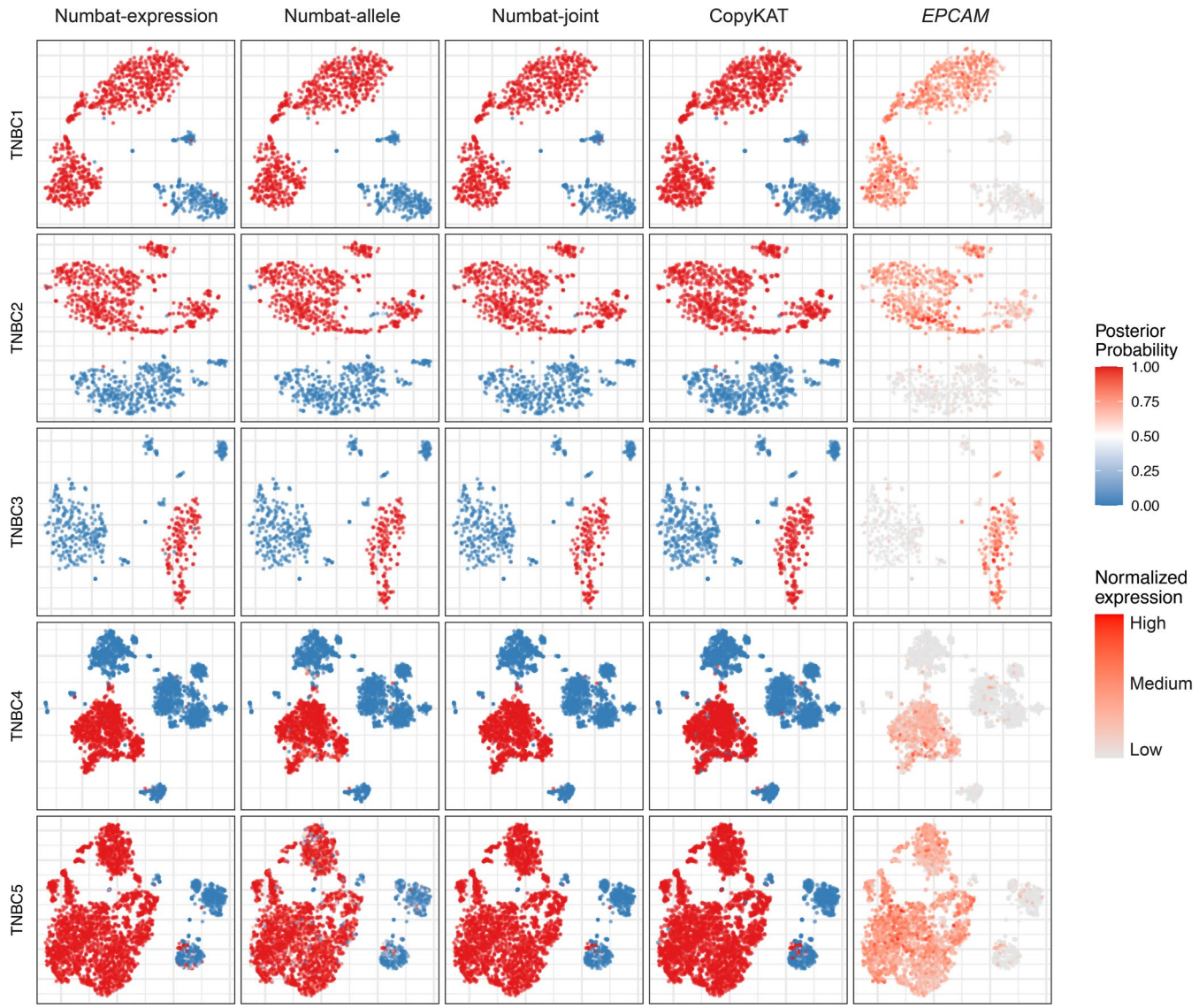
Extended Data Fig. 2 | Probabilistic model of gene expression and allele counts from transcriptome sequencing experiments. c_m, number of maternal chromosome copies. c_p, number of paternal chromosome copies. λ_i, observed gene expression magnitude of gene *i*. λ_i^*, reference gene expression magnitude of gene *i*. μ and σ^2, global bias and variance in gene expression. π_j, fraction of paternal alleles of SNP *j*. γ, global inverse overdispersion of allele-specific detection. *l*, library size. m_j, total allele count of SNP *j*. X_i, observed molecule counts for gene *i*. Y_j, observed paternal allele count for SNP *j*.



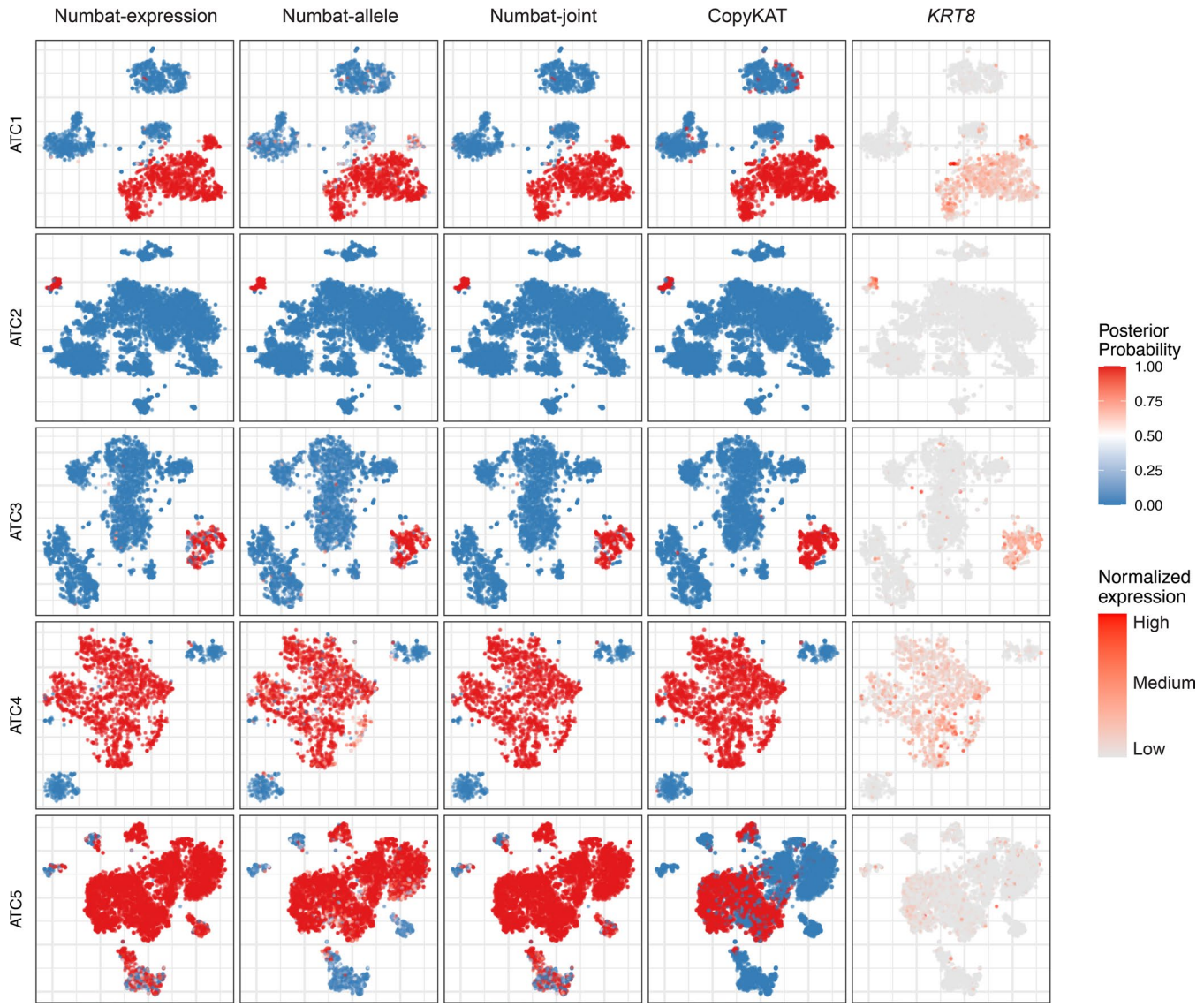
Extended Data Fig. 3 | WGS validation of Numbat CNV calls from scRNA-seq data. For each sample, the DNA profile (top) is juxtaposed with the copy number profile inferred by the Numbat joint HMM (bottom). Gray vertical bars represent centromeres and gap regions. logR, log coverage ratio. BAF, B-allele frequency. logFC, log expression fold-change. pHF, paternal haplotype frequency. BAMP, balanced amplification.



Extended Data Fig. 4 | Tumor versus normal cell classification accuracy of Numbat joint model, Numbat expression-only model, and CopyKAT. Each dot represents a distinct sample (TNBC, n = 5; ATC, n = 4; MM, n = 8). Center line, mean. ATC5 was excluded from the benchmark due to lack of clear expression of tumor marker *KRT8*.



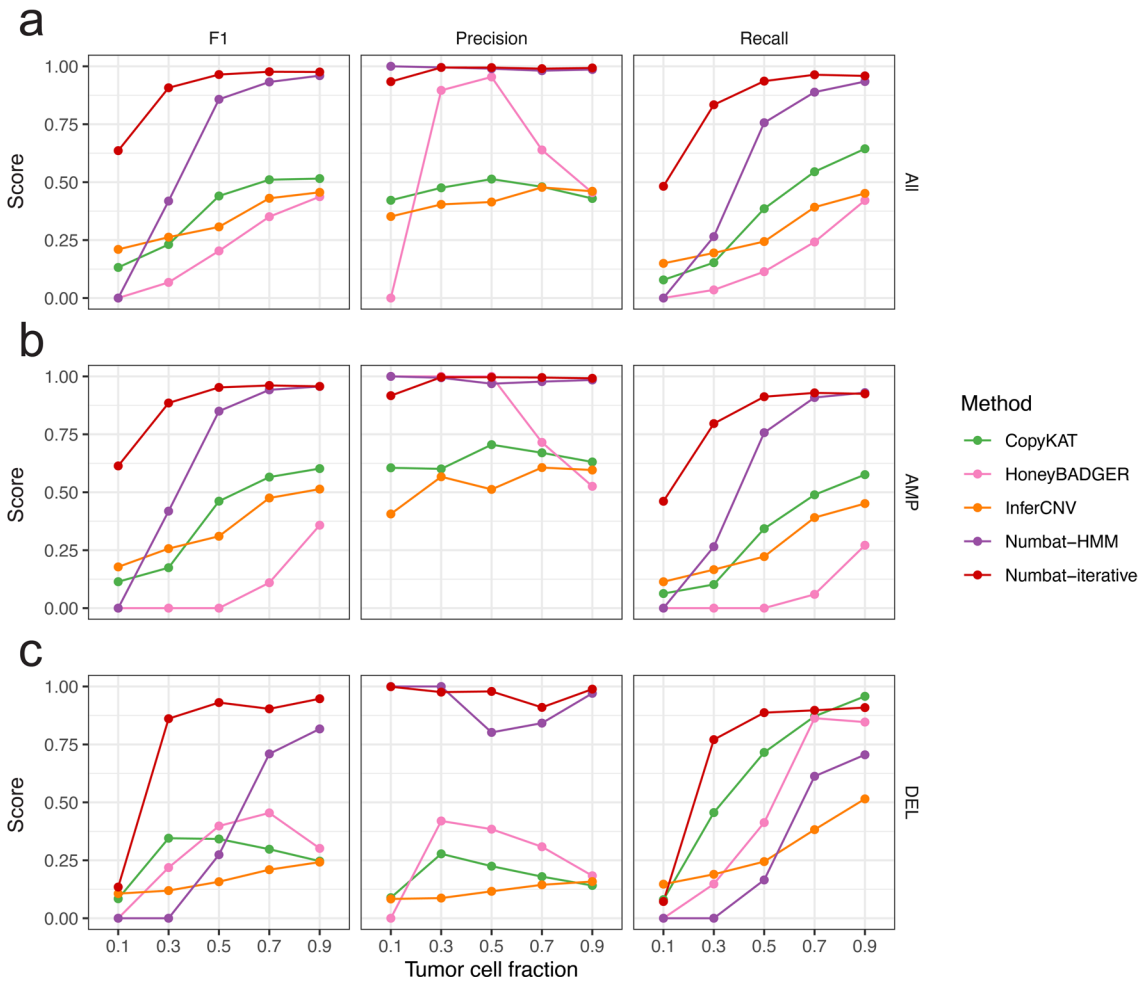
Extended Data Fig. 5 | Numbat reliably distinguishes tumor and normal cells (TNBC series). The aneuploidy probability is shown as a color gradient (red: high, blue: low). For each sample (row), the series of figures (columns) respectively show the aneuploidy probabilities by expression evidence, those by allele evidence, those by combined evidence, CopyKAT prediction (binary 0 or 1), and marker gene expression in a t-SNE embedding of gene expression profiles.



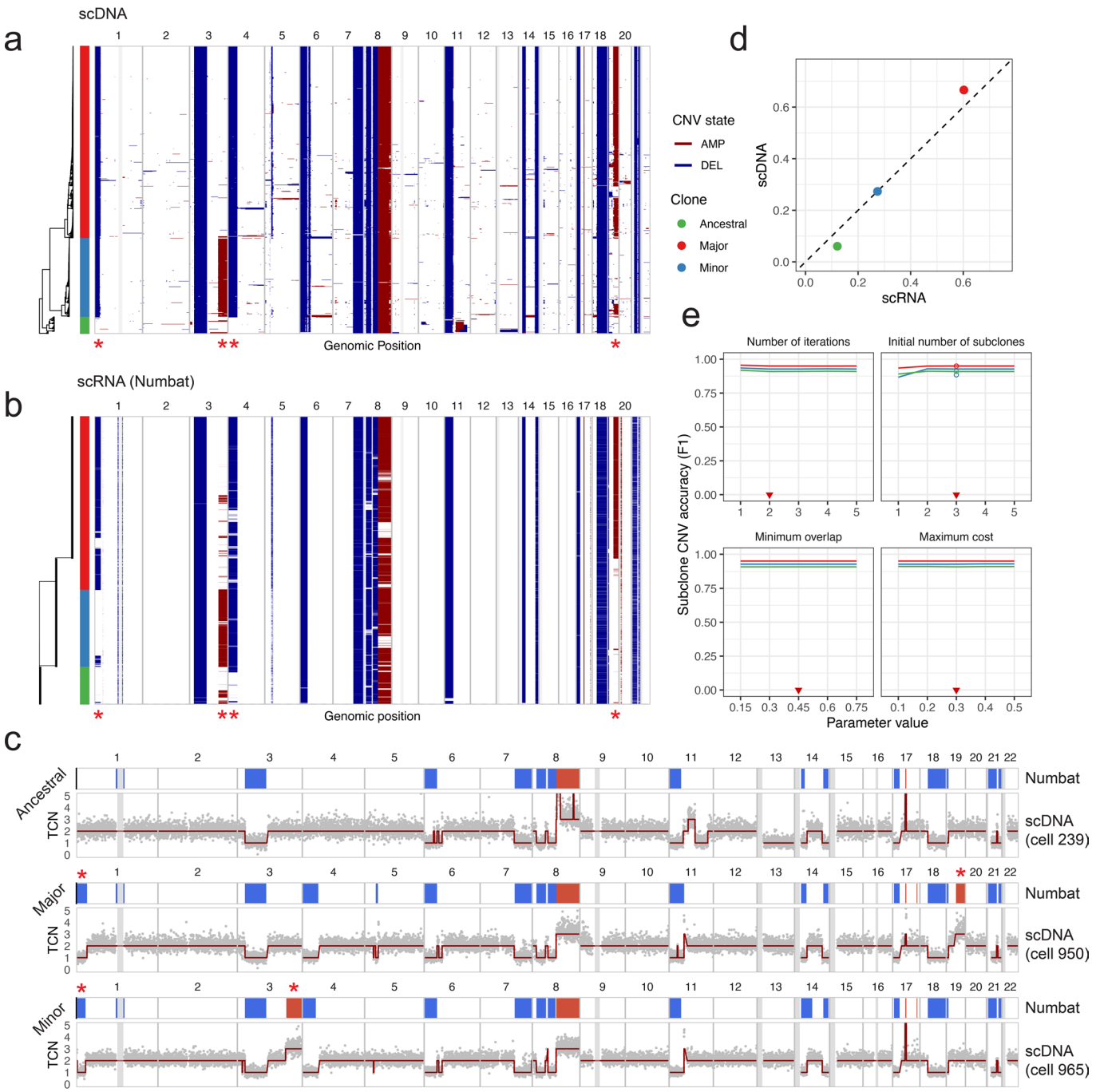
Extended Data Fig. 6 | Numbat reliably distinguishes tumor and normal cells (ATC series). The aneuploidy probability is shown as a color gradient (red: high, blue: low). For each sample (row), the series of figures (columns) respectively show the aneuploidy probabilities by expression evidence, those by allele evidence, those by combined evidence, CopyKAT prediction (binary 0 or 1), and marker gene expression in a t-SNE embedding of gene expression profiles.

Extended Data Fig. 7 | See next page for caption.

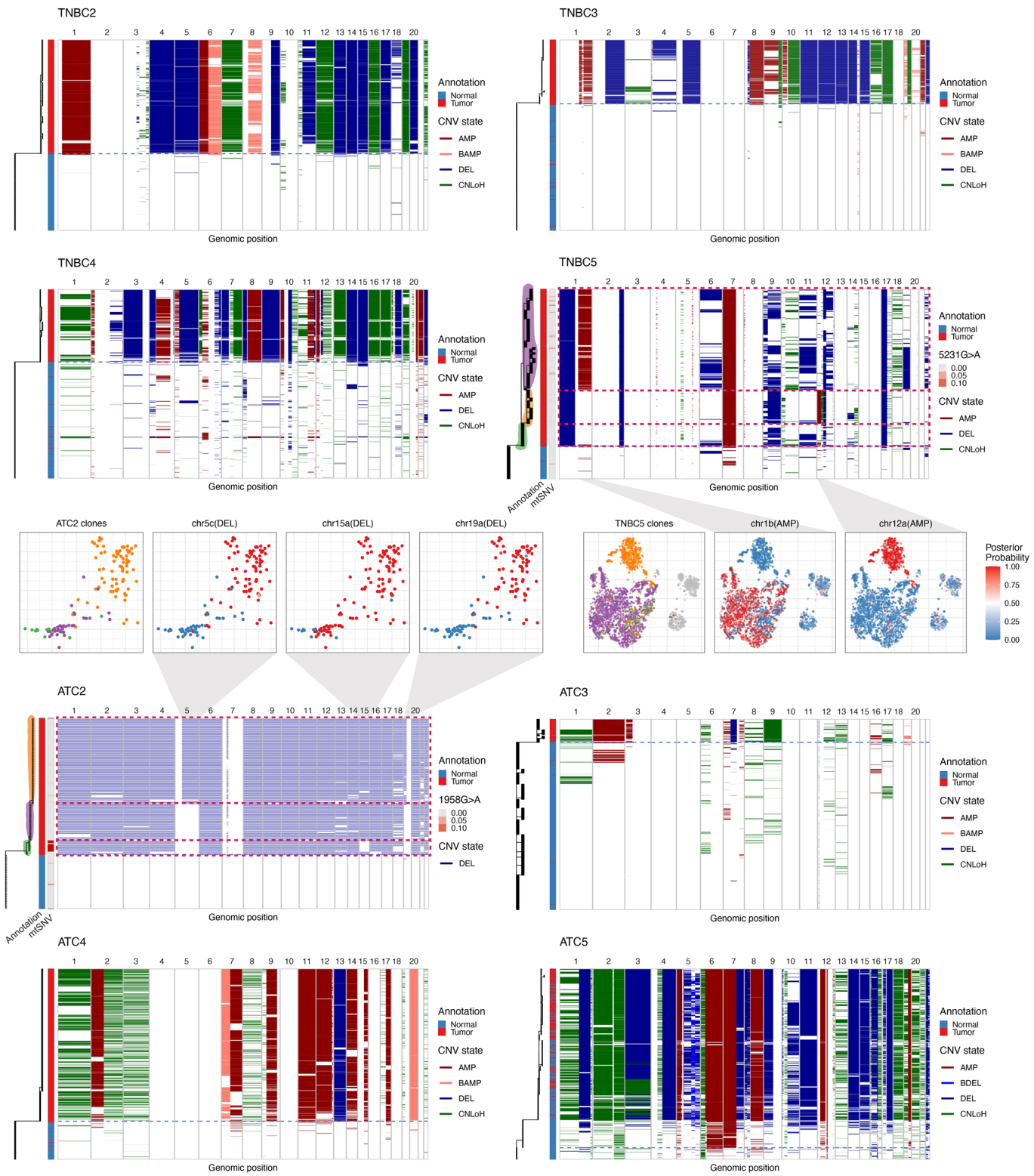
Extended Data Fig. 7 | Numbat reliably distinguishes tumor and normal cells (MM series). The aneuploidy probability is shown as a color gradient (red: high, blue: low). For each sample (row), the series of figures (columns) respectively show the aneuploidy probabilities by expression evidence, those by allele evidence, those by combined evidence, CopyKAT prediction (binary 0 or 1), and marker gene expression in a t-SNE embedding of gene expression profiles.



Extended Data Fig. 8 | CNV detection performance as a function of tumor cell fraction. At each tumor cell fraction, tumor cells were subsampled and mixed with randomly sampled normal cells at the corresponding proportion. Precision, recall and F1 scores were calculated based on the detected segments from scRNA-seq data and the ground truth copy number profiles (from WGS) in 5 multiple myeloma samples. For Numbat, two methods are compared: pseudobulk joint HMM (Numbat-HMM) and iterative optimization (Numbat-iterative) with no minimum pseudobulk size limit. **a**, Performance for all event types (amplification, deletion, and CNLoH). **b**, Performance for amplifications. **c**, Performance for deletions.



Extended Data Fig. 9 | Numbat analysis of gastric cell line (NCI-N87) scRNA-seq data and validation by scDNA-seq. a, Single-cell copy number landscape and subclonal structure reconstructed by scDNA-seq data. Gray vertical bars represent gap regions. A rooted hierarchical clustering tree is shown on the left. Three subclones were defined by cutting the tree with k=3. Red asterisks denote salient subclonal events. **b**, Single-cell CNV landscape and subclonal structure inferred from the paired scRNA-seq data by Numbat. The original prediction was composed of four subclones. The uppermost two clones were merged and denoted as the 'major' clone. Red asterisks denote validated subclonal events. **c**, Subclone-specific copy number profiles. For each subclone, the top track shows CNV calls made by clone-specific Numbat HMM; the bottom track shows DNA copy number profile of a representative cell from that subclone. Gray vertical bars represent gap regions. **d**, Numbat recapitulates clonal fractions measured by scDNA-seq. **e**, Stability and accuracy of Numbat CNV calls for each subclone with respect to parameter variations. F1 scores were computed by comparing DNA profiles for each subclone with the best-matching subclone CNV profiles predicted by Numbat. Circles denote F1 score from initialization with a random tree. Red triangles mark default parameter values.



Extended Data Fig. 10 | Single-cell copy number profile and phylogeny reconstructed by Numbat (TNBC and ATC). Branch lengths correspond to the number of CNV events. Blue dashed line separates predicted tumor and normal cells. Confident subclones are highlighted and marked by red dashed rectangles. The vertical bar on the left of each panel shows cell type ground truth. In TNBC5 and ATC2, the second vertical bar on the left of the panel shows variant allele frequency of a clone-associated mitochondrial mutation. For ATC2, results from the subsampled dataset (including aneuploid cells and 50 randomly sampled normal cells) are shown. In ATC5, some tumor cells were likely mis-annotated as normal in the original annotation.

nature research

Corresponding author(s): Peter Kharchenko

Last updated by author(s): Aug 10, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No custom software was used to collect the data.
Data analysis	We used the Cell Ranger (version 6.0.2, 10x Genomics) software suite to process the raw FASTQ or BAM files obtained from the previously published studies. We used conos (https://github.com/kharchenkolab/conos; version 1.4.1) to perform multi-sample integration, clustering, and generation of graph embeddings from scRNA-seq data. We used the LIGER R package (https://github.com/JEFworks/liger; version 2.0.1) to perform the gene set enrichment analysis between clonal cell populations, and the Mann Whitney U test implemented in pagoda2 (https://github.com/kharchenkolab/pagoda2; version 1.0.9) to identify confident differentially expressed genes. We applied the MQuad method (https://github.com/single-cell-genetics/MQuad; version 0.1.6) to identify mtRNA mutations from scRNA-seq samples. To perform phasing using single-cell expression data, we utilized the previously published scphaser R package (https://github.com/edsgard/scphaser; version 1.0.0). We used CopyKit (https://github.com/navinlabcode/copykit; version 0.1.1) to perform preprocessing, quality control, and analysis of scDNA-seq data. We used hmftools (https://github.com/hartwigmedical/hmftools; AMBER version 3.5, COLBALT version 1.11, PURPLE version 3.2) and R package copynumber (version 1.32.0) to perform unmatched CNV analysis of the WGS data from the MM dataset. We used cellsnp-lite (version 1.2.2) to obtain allele counts for each cell and Eagle2 (version 2.4.1) for population-based haplotype phasing. For performance benchmarking, we included HoneyBADGER (version 0.1), CopyKAT (version 1.0.8), and InferCNV (version 1.8.1). The fishplot R package (v0.5.1) was used to visualize tumor clonal structures. All other statistical analyses and visualizations were performed in R (version 4.1.2). The Numbat algorithm is available at https://github.com/kharchenkolab/numbat. The analysis scripts and notebooks to reproduce results included in the paper are available at https://github.com/kharchenkolab/NumbatAnalysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The scRNA-seq and WGS validation data from the WASHU multiple myeloma study can be accessed through SRA (PRJNA694128). The scRNA-seq data from the MDA CopyKAT study can be accessed through GEO (GSE148673) and SRA (PRJNA625321). The NCI-N87 scDNA-seq and scRNA-seq datasets are available on GEO (GSE142750) and SRA (PRJNA498809). The HCA collection of reference expression profiles can be obtained from Synapse under the ID syn21041850. The 1000 Genomes Project phasing panel can be downloaded from the IGSR FTP site (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release). The TOPMed phasing panel can be accessed through the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Since the aim of this study is to demonstrate the utility of the Numbat computational method through exploratory analysis of public datasets, we did not perform power calculation to determine sample size. The number of scRNA-seq samples analyzed for triple-negative breast cancer, anaplastic thyroid cancer, and ductal carcinoma in situ was based on the sample size of the original study (Gao et al. 2021). The number of multiple myeloma scRNA-seq samples analyzed was based on the number of patients with at least one high-quality sample-matched flow-sorted WGS (Liu et al. 2021). The number of analyzed samples with scDNA-seq validation was based on the availability of clonality analysis results in the original paper (Andor et al. 2020). The total number of samples analyzed (n=22) was sufficient to establish the reliability of the method in a diverse range of genomically unstable tumors.
Data exclusions	For the scRNA-seq samples from the Gao et al. and Andor et al. studies, we excluded cell barcodes that are absent in the gene expression count matrices downloaded from GEO (GSE148673, GSE142750) in order to recapitulate the quality control criteria of the original analyses. Similarly, for the scRNA-seq samples from the Liu et al. study, we excluded cell barcodes absent in the cell-type annotations from the original analysis. These exclusion criteria were pre-established.
Replication	Since we do not have access to patient samples included in the previously published studies, we did not perform replication.
Randomization	Randomization was not relevant to this study since no experimental groups were defined.
Blinding	Blinding was not relevant to this study since no experimental groups were defined.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Please refer to the original publications (Gao et al 2021, Liu et al 2021) for population characteristics.
Recruitment	Patients were recruited according to the original publications (Gao et al 2021, Liu et al 2021).
Ethics oversight	Study protocols were approved according to the original publications (Gao et al 2021, Liu et al 2021).

Note that full information on the approval of the study protocol must also be provided in the manuscript.