



A Bayesian method to infer copy number clones from single-cell RNA and ATAC sequencing

Lucrezia Patruno^{1,2}, Salvatore Milite^{2,3}, Riccardo Bergamin², Nicola Calonaci², Alberto D'Onofrio², Fabio Anselmi², Marco Antoniotti¹, Alex Graudenzi¹, Giulio Caravagna^{2*}

1 Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy

2 Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy

3 Centre for Computational Biology, Human Technopole, Milan, Italy.

 These authors contributed equally to this work.

* Corresponding author: gcaravagna@units.it.

Abstract

Single-cell RNA and ATAC sequencing technologies allow one to probe expression and chromatin accessibility states as a proxy for cellular phenotypes at the resolution of individual cells. A key challenge of cancer research is to consistently map such states on genetic clones, within an evolutionary framework. To this end we introduce CONGAS+, a Bayesian model to map single-cell RNA and ATAC profiles generated from independent or multimodal assays on the latent space of copy numbers clones. CONGAS+ can detect tumour subclones associated with aneuploidy by clustering cells with the same ploidy profile. The framework is implemented in a probabilistic language that can scale to analyse thousands of cells thanks to GPU deployment. Our tool exhibits robust performance on simulations and real data, highlighting the advantage of detecting aneuploidy from two distinct molecules as opposed to other single-molecule models, and also leveraging real multi-omic data. In the application to prostate cancer, lymphoma and basal cell carcinoma, CONGAS+ did retrieve complex subclonal architectures while providing a coherent mapping among ATAC and RNA, facilitating the study of genotype-phenotype mapping, and their relation to tumour aneuploidy.

Author summary

Aneuploidy is a condition caused by copy number alterations (CNAs), which brings cells to acquire or lose chromosomes. In the context of cancer progression and treatment response, aneuploidy is a key factor driving cancer clonal dynamics, and measuring CNAs from modern sequencing assays is therefore important. In this framing, we approach this problem from new single-cell assays that measure both chromatin accessibility and RNA transcripts. We model the relation between single-cell data and CNAs and, thanks to a sophisticated Bayesian model, we are capable of determining tumour clones from clusters of cells with the same copy numbers. Our model works when input cells are sequenced independently for both assays, or even when modern multi-omics protocols are used. By linking aneuploidy to gene expression and chromatin conformation, our new approach provides a novel way to map complex genotypes with phenotype-level information, one of the missing factors to understand the molecular basis of cancer heterogeneity.

Introduction

Cancer is a disease in which cell subpopulations with enhanced functional capabilities emerge, evolve and undergo selection against the immune system response and treatment [1]. The investigation of tumour evolution in terms of omics layers – e.g., genome, transcriptome, proteome, epigenome, metabolome – has key translational repercussions [2], and can benefit from widespread single-cell sequencing technologies [3] that probe, among many, RNA (scRNA-seq) and ATAC (scATAC-seq) from biopsies and patient-derived model systems [4]. With the current technologies, the most recent protocols can also extract multiple measurements from the very same cell (e.g., G&T [5] or GoT [6] for matched RNA/ DNA, or 10x multiome [7] for ATAC/ RNA), even if “multimodal” technologies have still limited diffusion because they are very expensive and relatively low throughput. For this reason, a much more common single-cell design is based on separating cells before sequencing, with many computational efforts focused on integrating, a posteriori, data generated from different data modalities. In this second scenario, sometimes referred to as diagonal integration [8], the general idea is to map the measurements in a latent space, using some unsupervised integration method. If we do not need the latent space to reflect any specific biological quantity, variational autoencoders or factor analysis can be adopted [9–12]. Otherwise, when it is required that the latents are biologically interpretable, other approaches should be preferred. In cancer genomics and tumour evolution studies, one possibility is to reconcile RNA and ATAC from the observation that both capture distinct aspects of the same DNA molecule. In this sense, RNAs are products of the transcriptional processes that initiate from DNA, and ATAC is an assessment of chromatin conformation, a physical feature of DNA. Therefore, an interesting attempt – which is the one we follow in this paper – is mapping RNA and ATAC on latent DNA states. Notably, an extra layer of complexity is introduced by observing that latent DNA configurations can differ between subgroups of cells characterised by multiple types of genetic mutations. Here we define these as tumour clones with associated Copy Number Alterations (CNAs), potent modifications of the chromosome conformation and copies. While point mutations are difficult to characterise and link to RNA and ATAC, the opportunity of modelling latent CNAs seems more feasible. The possibility of inferring latent tumour subclones from scRNA-seq has been already widely investigated [13–17], and some preliminary attempts at working with scATAC-seq are also present [18–20]. In this framing, we recently introduced CONGAS [13], a method to perform CNA-based integration from scRNA-seq. Starting from a genome segmentation (set of breakpoints), CONGAS uses a Bayesian probabilistic model to infer latent total CNAs (i.e., per segments ploidy estimates) while clustering input cells. CONGAS was the first model to join signal deconvolution (i.e., clustering), while detecting subclonal patterns of aneuploidy, and worked better than methods like InferCNV [14], HoneyBADGER [21], CopyKAT [17] and Numbat [22] that decoupled CNA inference from clustering. The solution achieved with CONGAS was however only partially satisfactory, because the statistical signal of CNAs in scRNA-seq is generally affected by strong confounders such as allele-specific expression and post-transcriptional regulation, two biological phenomenon that are only partially understood and play an important role in cancer [23–25]. In practice, the distribution of read counts in RNA space (the inverse of the latent mapping), is not a perfect predictor of CNAs, and a better-quality signal can instead be achieved by examining chromatin conformation, a direct measurement of DNA. We reasoned that, as in CONGAS, one can postulate that the more alleles (i.e., copies) of a chromosome region are open, the stronger the signal of ATAC on the region should be. In this sense, a model a-la-CONGAS could be developed to link the latent CNA to the observable ATAC peaks. As far as we understand, such an intuition has never been leveraged before, missing the possibility of integrating independent

scRNA-seq and scATAC-seq measurements using a biology-informed latent model of DNA copy numbers.

Building on this intuition, in this paper we present CONGAS+, a Bayesian model to map single-cell RNA and ATAC measurements in latent CNAs, clustering cells across the two data modalities, and predicting clones with a well-defined discrete copy number profile. Doing so, the CONGAS+ framework opens the possibility to compare both gene expression and chromatin accessibility across copy number clones. As a byproduct, the tool can readily separate tumour from normal cells when the former are characterised by aneuploidy, a very common situation in cancer. The model is unsupervised, and the likelihoods of RNA and ATAC are conditionally independent given the latent CNAs, but combined thanks to a shrinkage statistics to weight the contribution of the data modalities unevenly, which helps when one of the two modalities (usually RNA) is a worst predictor of CNAs. The overall model uses stochastic variational inference and gradient descent to learn parameters from data, and enjoys a fast implementation via probabilistic programming in Pyro [26]. This allows deploying CONGAS+ on GPUs seamlessly, analysing datasets with tens of thousands of cells in matters of minutes thanks to the massively parallel architectures offered by graphical devices. In this paper we show its capacity to extract complex genotype/ phenotype information from a B-cell lymphoma (~ 6400 cells RNA/ATAC multimodal), a basal cell carcinoma (~ 1200 cells RNA, ~ 1200 cells ATAC) and a prostate cancer cell line (~ 7600 cells RNA, ~ 8800 cells ATAC).

Materials and methods

The CONGAS+ statistical model

CONGAS+ is a Bayesian model to infer copy number clones from scRNA-seq and scATAC-seq data, while simultaneously grouping cells into clusters characterised by similar Copy Number Alterations (CNAs). The model can also process data for which only one modality is available, as well as multimodal data generated by a multi-omics assay. Applied to cancer, CONGAS+ can i) separate tumours from normal cells (if the tumour has aneuploidy), ii) and detect tumour subclones associated with distinct CNAs. By default, the model scans for large CNAs at the resolution of chromosome arms but can be run also with a custom segmentation/ ploidy, for instance obtained by an independent sequencing assay. The model encodes this information as categorical priors on discrete copy number values, whose posteriors are obtained from a likelihood of scRNA and scATAC counts pooled at the resolution of the segments. The process of integrating two distinct measurements reflects also on input data resolutions: e.g., for ATAC we use either a fixed-width binning or open chromatin peaks, while for RNA we employ transcript counts. Notably, CONGAS+ is inspired by its predecessor CONGAS [13], the first method to attempt this inference from scRNA-seq alone. CONGAS+ has two main advantages: first, it builds a stronger statistical signal by joining ATAC on top of RNA, and second it models discrete CNAs, whereas the original model was approximating continuous copy numbers, making it difficult to compare the difference among clones in a solid statistical way.

CONGAS+ is a parametric Dirichlet mixture for $K \geq 1$ clusters, determined from scRNA/ scATAC counts data X of $N > 0$ and $M > 0$ cells, respectively, in $I > 0$ segments. The model can process both discrete raw counts, as well as their continuous version normalised by library size and, interestingly, it can also work for a multimodal RNA/ATAC assay where $N = M$. As in other methods [13,16], the total genome ploidy of a segment (sum of the latent major and minor alleles) is a linear predictor of the

observed counts, i.e., if the latent number of DNA copies of a segment i is c , if we found $r > 0$ RNA transcripts and $a > 0$ open chromatin peaks mapped to the segment, then the predictors are of the form $r \propto c\theta_i^{\text{RNA}}$ and $a \propto c\theta_i^{\text{ATAC}}$ for certain parameterizations of the rate at which we observe RNA transcripts, and ATAC peaks.

The CONGAS+ likelihood of every modality (denoting its data broadly as X) is defined as

$$p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\pi}, \Phi) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \prod_{i=1}^I f(x_{n,i}|\boldsymbol{\theta}_i, \Phi) \quad (1)$$

where X is a $N \times I$ matrix if there are N cells for the modality, π are the K -dimensional mixing proportions, $x_{n,i}$ the counts for the n -th cell on the i -th segment, and Φ is a $K \times I \times H$ tensor that models the probability distribution over latent CNAs, which here have $H = \{1, 2, \dots\}$ possible states (usually capped at 5 copies). To map independent scRNA-seq/ scATAC-seq on the same set of clusters, Φ does not change across modalities. Note that f is a generic observational model that depends on counts either being discrete or normalised. The model is also available in two configurations, one with π independent across RNA/ATAC, and one where mixing is shared, which helps if signal strength is uneven across datasets.

For scRNA-seq/ scATAC-seq integer counts $x_{n,i}$ mapped to the i -th segment of the n -th cell, the function f associated with the k -th mixture component is a Negative Binomial parameterized by mean and overdispersion, i.e.,

$$f_k(x_{n,i}|\boldsymbol{\theta}_i, \Phi_{k,i}) = \text{NegBin}\left(x_{n,i} \left| \frac{\mu_{k,i,n}}{\mu_{k,i,n} + r_i}, r_i \right.\right) \quad (2)$$

In CONGAS+ the mean depends on the expected counts per allele (θ_i , learnt), the library size factor of the cell (ρ_n , observed), and the linear combination (dot product) of the latent CNAs, i.e.,

$$\mu_{k,i,n} = \underbrace{(\rho_n \cdot \theta_i)}_{\text{Normalisation}} \cdot \underbrace{\left(\sum_h \Phi_{k,i,h} \cdot h\right)}_{\text{CNAmixture}} \quad (3)$$

where we are using a nested mixture of CNAs with value h weighted by their probability $\Phi_{k,i,h}$ – i.e., the probability to detect CNA value h , in segment i and cluster k . The overdispersion of this segment is instead learnt from data. At the level of priors, the probability for each CNA $\Phi_{k,i,h}$ is a Dirichlet distribution with parameter an $|H|$ -dimensional vector α , which can be set to any input predefined CNA profile: if we expect a ploidy p for a segment, we assign to the p -th entry value 0.6, and to the remaining 0.1 to obtain skewed samples from the Dirichlet. The joint scRNA/scATAC CONGAS+ likelihood for $\mathbf{X} = [X^{\text{RNA}}, X^{\text{ATAC}}]$ has a shrinkage form

$$p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\pi}, \Phi) = \lambda p(\mathbf{X}^{\text{RNA}}|\boldsymbol{\theta}^{\text{RNA}}, \boldsymbol{\pi}^{\text{RNA}}, \Phi) + (1 - \lambda)p(\mathbf{X}^{\text{ATAC}}|\boldsymbol{\theta}^{\text{ATAC}}, \boldsymbol{\pi}^{\text{ATAC}}, \Phi) \quad (4)$$

where $0 \leq \lambda \leq 1$ is a fixed hyperparameter to weight the likelihood of both modalities.

For a given configuration with K clusters, parameters are learnt using a stochastic variational inference schema, giving in input segments that, from an isolated run, are fit to at least two clusters (i.e., multi-modal segments). Moreover, in learning the categorical Φ , we perform a reparameterization based on the Gumbel-softmax to draw low variance differentiable samples [27]. The inference returns the posterior distribution over copy number profiles and cell clustering assignments, while for all the other parameters CONGAS+ outputs the Maximum A Posteriori (MAP) estimate. The

number of clusters $K > 0$ is optimised via model selection of standard score functions [28]: the Bayesian or Akaike Information Criteria (BIC, AIC), as well as the Integrated Completed Likelihood (ICL) [28]. In particular, given the complete log-likelihood $L(\mathbf{X}) = \ln p(\mathbf{X}|\theta, \pi, \Phi)$ and the number of parameters v for a model with n samples, the scores are $\text{BIC}(\mathbf{X}) = v \ln(n) - 2L(\mathbf{X})$, $\text{AIC}(\mathbf{X}) = 2v - 2L(\mathbf{X})$ and $\text{ICL}(\mathbf{X}) = \text{BIC}(\mathbf{X}) + H(Z)$ where Z are the latent variables for cell clustering assignments, and $H(Z)$ their entropy [29]. To choose the optimal number of clusters we minimise any of those (BIC by default).

The full formulation of CONGAS+, including the alternative Gaussian continuous likelihood are provided as Supplementary Material. CONGAS+ is implemented in 2 open-source packages: one, in Python, using the probabilistic programming language Pyro [26], while the other, in R, to preprocess data and visualise inputs/ outputs, interacting with Python through reticulate.

Results

Model validation and parameterization

Performance and comparison to alternative methods.

We performed simulations to assess the performance of CONGAS+ and other methods. To obtain an unbiased and realistic performance assessment, we simulated scRNA-seq and scATAC-seq data outside of CONGAS+, even if our method can generate data.

We simulated RNA and ATAC of normal cells via simATAC [30] and SPARsim [31], two tools for synthetic data generation calibrated on real sequencing technologies. Then, we added CNAs for $2 \leq K \leq 10$ clones assembled from a random clone tree [32] with ≤ 5 extra CNAs every new subclone added to the tree. Finally, we generated mixing proportions from a Dirichlet distribution with uniform concentration, and assigned cells to clusters in 10 replicas for each K , for 90 total datasets, each one with 1500 scRNA-seq and 1500 scATAC-seq cells.

We applied CONGAS+ to each dataset searching for $K \leq 10$, and measured i) the Adjusted Rand Index (ARI), i.e. the similarity between the known and inferred cluster memberships, and ii) the mean absolute error (MAE) between cluster-specific CNAs, and simulated ones. Across all tests we observed a very good performance (median ARI > 0.8 and MAE consistently lower than 1) suggesting that CONGAS+ can retrieve the cells from each clonal population, as well as their copy number profiles (Figure 1I,K). The same test has been carried out against alternative CNA detection algorithms that can work with either RNA or ATAC single-cell data (Figure 1J). In particular, we measured ARI for i) CONGAS+ ii) copyKAT [17], which uses only RNA data, and iii) copy-scAT [20], which uses only ATAC data. In this test we observed that CONGAS+ outperforms both tools. We explain this because our method can work even in the presence of pure tumour samples and multiple subclones, whereas the other two callers can only separate tumour from normal cells, therefore their usage is more limited.

The importance of using a joint assay

CNA-associated signal quality is not necessarily even across ATAC and RNA, with the latter showing higher overdispersion due to difference in sample preparation, library size, gene expression variability, and sequence-specific biases [33,34]. High overdispersion can act as a confounder, making it difficult to infer poorly separable clusters from scRNA-seq

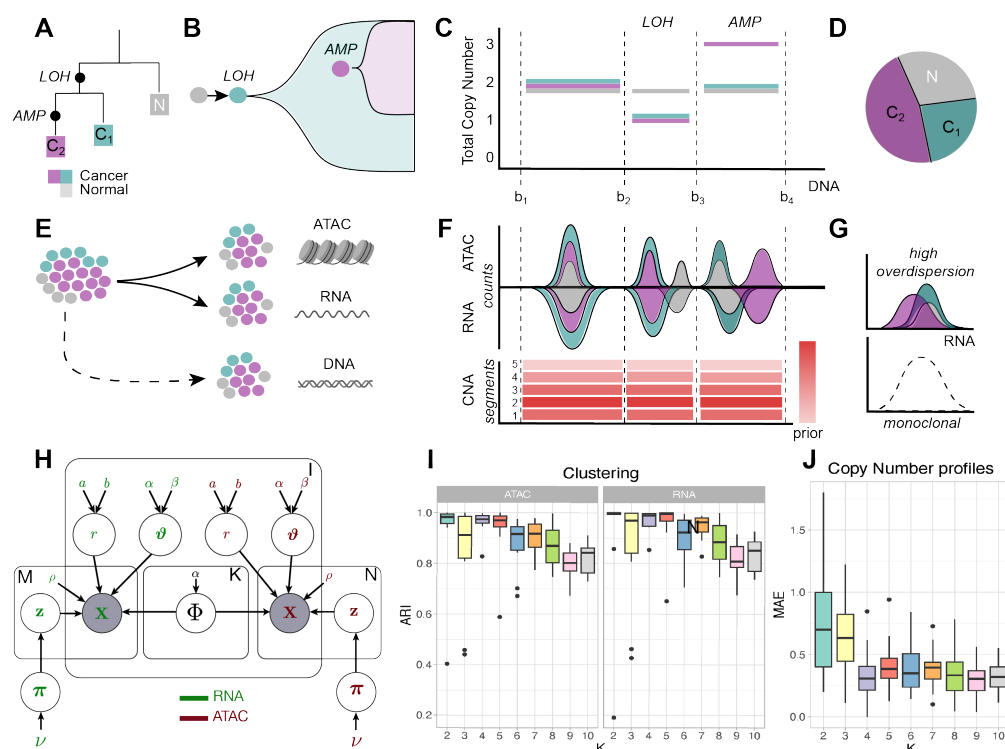


Fig 1. CONGAS+ framework A-D: Given a tumour sample characterised by an accumulation of Copy Number events (A-B), CONGAS+ aims at inferring the total Copy Number profiles (C) and the clonal composition of each sample (D). E-F: CONGAS+ takes in input the single-cell RNA, single-cell ATAC and, optionally, a copy number segmentation obtained from a bulk whole genome assay, which is used as a prior for the copy number state of each segment. In case it is not provided, the tool employs a diploid arm-level segmentation. G: CONGAS+ infers the copy number state in each segment. H: Probabilistic Graphical Model. I: Adjusted Rand Index (ARI) computed comparing the ground truth labels of simulated cells with the clustering assignments returned by CONGAS+. 90 datasets with 1500 cells for scRNA and 1500 cells for scATAC were simulated. J: CONGAS+ performance compared with copyKAT [17] and copy-scAT [20], computing the ARI obtained on the same simulated datasets as in I. K: For the same data in panel (I), Mean Absolute Error (MAE) between the ground truth and the inferred copy number profiles. L: Boxplot showing the ARI for copyKAT, CONGAS and CONGAS+ (computed on scRNA and scATAC separately) obtained running the tools on bootstrap samples characterised by a bimodal signal poorly evident in RNA.

data alone, eventually leading to failures in detecting subclones. Using ~ 1800 scRNA and ~ 600 scATAC profiles from the Basal Cell Carcinoma (BCC) sample SU008 [35,36], we created a dataset of tumour and normal cells in even proportions, subsetting the genome to two diploid and two with aneuploidy, with bimodal signal poorly evident in RNA. Then, we performed non-parametric bootstrapping for the genes in each segment, and compared 30 inferences with CONGAS+ (RNA plus ATAC), CONGAS (RNA) and copyKAT (RNA). Using a joint ATAC-RNA assay, CONGAS+ with $\lambda = 0.1$ detected CNAs that distinguish tumour from normal cells, obtaining a median ARI ~ 0.7 on ATAC but a lower ARI on RNA (Figure 1L). In general, due to the weaker RNA signal, all tools that looked only at RNA struggled separating tumour

and normal cells, with copyKAT and CONGAS unable to detect the split (Figure 1L). In this test, copy-scAT failed to execute with standard parameters. Overall, this shows that with a joint inference on the ATAC and RNA modalities we can detect the clonal structure of the dataset also when one data modality has a weak signal.

Shrinkage effect with Basal Cell Carcinoma data

Since signal quality can be uneven across data modalities, CONGAS+ is equipped with a shrinkage hyperparameter that can be used to weigh the evidence differently across RNA/ATAC. This serves as a natural hyperprior to decrease the importance given to a modality that we believe is more noisy or affected by some consistent bias. A natural question is therefore how does this affect the inference, and what value for λ should be suggested in the general case.

Building on the test shown in the previous section, we used data from SU008 and from sample SU006 [35,36] in order to test a signal present in just ATAC (SU008), against a signal present in both RNA and ATAC (SU006). For SU006, we selected 2 diploid segments and 2 segments with monosomy loss of heterozygosity (LOH) (fig. 2F,G). To investigate the effect of λ on the inference, we scanned $\lambda = 0.05, 0.15, \dots, 0.95$, set $K = 2$ and performed 10 runs to compare the ARI for cluster assignments against tumour/normal labels from [35,36]. We observed RNA/ATAC inferences stable against λ , with tumour and normal cells always separated (Figure 2C,H). For SU008, instead, only ATAC exhibits a neat bimodal distribution (fig. 2B), and this time we observed (fig. 2C) that for $\lambda < 0.5$ the ARI for ATAC is stable at ≈ 0.75 , whereas it decreases as λ approaches 0.95. Inferences for the best/ worst ARI (fig. 2D-E) show discordant tumour and normal assignments. With $\lambda \geq 0.25$ CONGAS+ did not fit the ATAC bimodality, merging 63% of tumour and 90% of normal cells together. Instead, for $\lambda < 0.25$ – more weight assigned to ATAC – assignments retrieved are perfect. As expected, the model is never able to separate tumour and normal from RNA, due to its unimodal distribution.

Overall, these tests show that if the quality of ATAC/ RNA are different, λ can be used to favour one assay over the other. CONGAS+ offers a principled approach based on likelihoods to inspect the optimal λ , and a final decision has to be taken on each dataset, also inspecting fits quality.

Phasing ATAC and RNA profiles in B-cell lymphoma multimodal data

The ideal data to be integrated in CONGAS+ is a multimodal ATAC/RNA assay, where joint measurements are available for all cells. We gathered data of a B-cell lymphoma [37] sequenced with the 10x multiome kit [7], and tested if CONGAS+ did identify clusters across the two modalities, and assign cells consistently. In this test we phased cells across modalities to exploit the one-to-one correspondence between ATAC/RNA barcodes. The expectation was to cluster together cells in both RNA and ATAC, even without any a priori imputation. We processed ~ 6400 cells with manually annotated (cfr. [37]) cell types after quality control. Cell types were distinguishable in a joint ATAC/RNA UMAP [38] low-dimensional representation (fig. 3A): two tumour cell populations (B and B-cycling) cluster together, whereas normal cells split into Monocytes, T and B cells. Note that, while CNAs could tell apart normal from tumour cells, the distinction among B and B-cycling tumour subpopulations is more likely linked to cell cycle entry dynamics, a byproduct of complex transcriptional regulation not necessarily linked to CNAs [39].

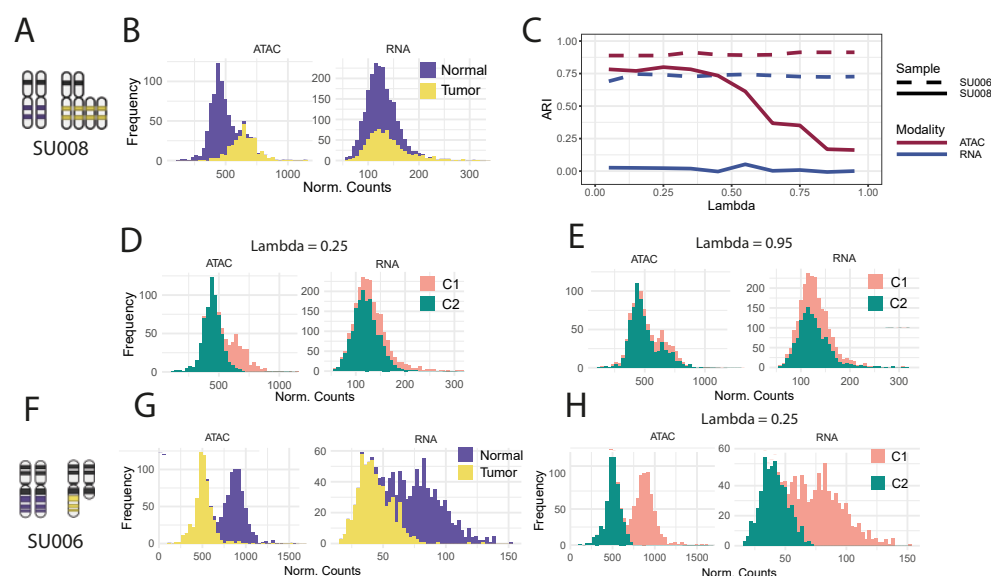


Fig 2. Impact of lambda variation on CONGAS+ performance using Basal Cell Carcinoma data. A,B,C,D,E: Basal Cell Carcinoma sample SU008 from [35] and [36], where we selected segments with bimodal signal in both scRNA-seq and scATAC-seq profiles. F,G,H: BCC sample SU006 from [35] and [36], where we selected segments in which ATAC signal is bimodal and RNA unimodal. C: ARI value for each modality, computed for every lambda ranging from 0.05 to 0.95. B,G: normalised counts distribution coloured by the ground truth cell labels for chromosomes chr9q and chr20q in samples SU008 and SU006 respectively. D,H: distributions coloured according to clustering assignments obtained from the solution showing the highest ARI for samples SU008 and SU006 respectively. E: normalised distribution for the worst solution in terms of ARI for sample SU008.

We found $K = 3$ clusters using an arm-level segmentation and diploid priors (Figure 3D). Comparing cell labels from [37], we observed a single population of tumour cells, but two clusters composed of normal cells (Figure 3B-C). While the tumour/normal split is reasonably explained by CNAs which characterise neoplastic cells on chromosomes chr2, chr3q (Figure 3E), chr4, chr5p, chr7p, chr8p, chr9, chr18q and chr22q, the distinction among normals was unexpected and linked to distinct ATAC profiles for chromosomes chr 16p and chr 19 (Figure 3F). We begin to verify that CONGAS+ was not splitting normal cells due to differences in cell type composition. Then, we performed differential expression analysis across the RNA profiles of all populations. We did find – as expected – differences in the expression (Figure 3H) of genes that distinguish normal from lymphoma cells, as measured from absolute log fold change (LFC) above 0.5 and p-values below 0.01 (Wilcoxon test), but did not find differences across the two normal subpopulations (Figure 3I). Among the tumour-associated genes we find LCP2, a prognostic gene for metastatic melanoma-infiltrating CD8+ T cells [40], MEF2C a gene which has been linked with the lymphoma pathogenesis [41], and MACF1, a gene which has been associated with cancer development [42]. Instead, a differential analysis of ATAC peaks between the two normal subpopulations showed marked ($|LFC| > 0.5$ and p-values < 0.01 ; Wilcoxon test) ATAC differences across chromosomes chr16 and chr 19 (Figure 3G,J).

The possible explanation to this outcome could be that the subset of normals identified

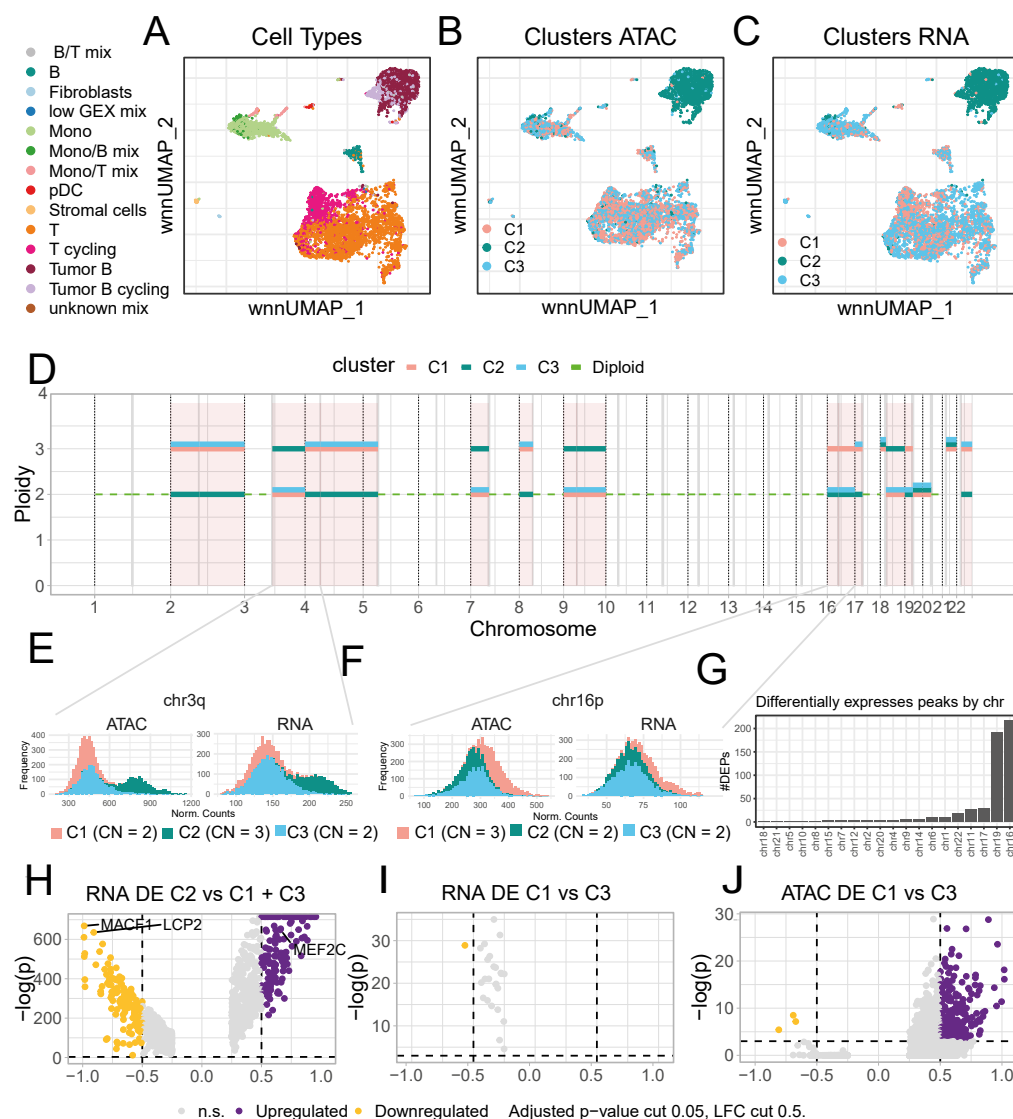


Fig 3. Application of CONGAS+ to B-cell lymphoma multimodal data. A: UMAP low-dimensionality representation of ~ 6400 RNA and ATAC single-cell profiles from the 10x multiome dataset from [37]. B,C: The same UMAP plots coloured according to the clusters inferred by CONGAS+ for ATAC and RNA, respectively. D: copy number profiles for each cluster inferred by CONGAS+ E-F: normalised counts distribution coloured according to cluster assignments, shown for segment chr3q (E) where an amplification characterises tumour cells and chr16p (F) where an amplification is observed for cluster C1. G: distribution across chromosomes of the differentially expressed peaks between C1 and C3. H-I: volcano plot for differentially expressed genes between C1 and C3 (E) and between the tumour cluster C2 and normal cells. J: volcano plot for differentially expressed peaks between normal cells clusters C1 and C3.

by CONGAS+, which cannot be otherwise identified by RNA-based tools, are an ancestor of the tumour population, describing the evolutionary link between normal cells and lymphoma cells. This will require further investigations also involving lineage-tracing via genetic polymorphisms to be confirmed, but this explanation would fit with biological observations that B-cell lymphomas originate from neoplastic

transformation of germinal centre B cells [43]. Overall, this case study shows the combined power of ATAC and RNA, a joint framework that, to the best of our extent, has not yet been exploited before to study copy number alterations.

CNA-associated drug-resistance clones in a prostate cancer cell line

Chromosomal instability and aneuploidy can generate potent phenotypes, sometimes capable of resisting negative selection induced by anticancer drugs [44]. We tested if CONGAS+ could identify, from scRNA-seq and scATAC-seq data, CNA-associated tumour subclones that resist treatment. We used data from [45], where ATAC/RNA data was generated for the untreated prostate cancer cell line LNCaP (parental), and then for one line treated 48 hours with AR antagonist enzalutamide (ENZ), and two resistant lines (RES-A and RES-B) derived after long-term exposure to ENZ and diarylthiohydantoin RD-162, respectively (fig. 4A). To search for high-resolution subclonal CNAs we downloaded LNCaP cytogenetics data from the DepMap portal [46], and used it to obtain breakpoint coordinates and priors for CNAs. We merged the 4 samples (parental, ENZ-48, RES-A, RES-B), and filtered out segments with more than 10% of cells showing zero counts.

CONGAS+ identified 3 clusters present in both ATAC and RNA across all samples (fig. 4C-F). The parental and ENZ-48 lines cluster together while RES-A and RES-B split in two clusters. This is consistent with the experimental design of [45]: ENZ48 has not yet acquired resistance due to its short-term exposure to ENZ, and is expected to cluster with parental cells. The two other clusters are composed of almost fully-resistant cells, with a good partition of the RES-A and RES-B cells. These two clusters share one amplification of the q-arm of chromosome 6 and 21 (fig. 4F), indicating evolution from sensitive cells through a common ancestor (fig. 4B). Moreover, the two resistant populations cluster separately and CONGAS+ finds, for RES-B, an amplification on the p-arm of chromosome 6 (fig. 4E), suggesting further evolution in that clone (fig. 4B).

Overall, this analysis shows that lineage relations associated with CNA-associated subclones can be effectively detected by CONGAS+ and longitudinal data, posing the bases for more systematic investigations on the causal roles of CNAs in promoting therapy resistance.

Discussion

The relation between somatic mutations and cancer phenotypes is extremely complex and intimately related to the underlying evolutionary dynamics of cancer cells and the environment. To understand this genotype-phenotype mapping, single-cell technologies can be adopted to achieve a fine-grained resolution of the measurements, but methods are required to resolve signals in such noisy data. In this paper, we approached this problem from RNA and ATAC single-cell sequencing, inferring latent tumour subclones associated with CNAs, a specific type of complex genomic mutation. CONGAS+ is the first Bayesian model that can jointly analyse RNA and ATAC, inferring CNAs while clustering cells through variational inference. The model has a shrinkage formulation to weigh the evidence between the two modalities, a feature motivated by our experience where scATAC-seq has a cleaner signal while scRNA-seq is overdispersed. This phenomenon could be explained considering that ATAC is a direct measurement of DNA, while RNA is a byproduct and is therefore more subject to biases.

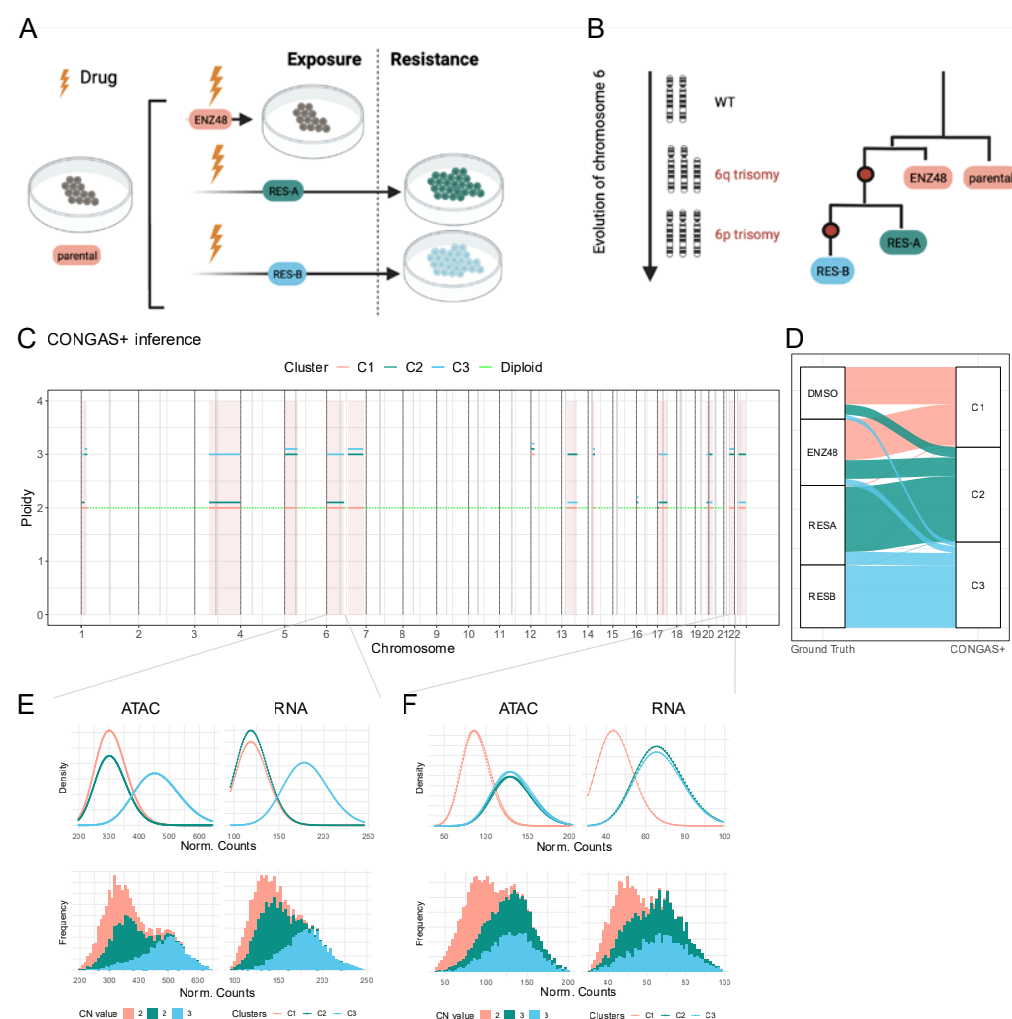


Fig 4. Application of CONGAS+ to prostate cancer cell line LNCaP with independent assays. CONGAS+ application to a prostate cancer dataset from [45], composed of a mixture of four cell lines with 7600 scRNA-seq cells and 8800 scATAC-seq cells. A-B: Cartoon representing the design of the drug resistance experiment and the corresponding sample tree. C: copy number profiles inferred by CONGAS+ for each cluster. D: Sankey plot showing the overlap between the sample of origin and the cluster inferred by the model. E-F: density plot and histogram of normalised counts coloured according to cluster assignments for chromosome 6p (D) where an amplification event is private to cluster C3, and chromosome chr21q where an amplification is shared by clusters C2 and C3.

Using simulations, we assessed that CONGAS+ is robust and accurate in retrieving both the clonal composition and corresponding CNAs. This was further confirmed with real data, where the method showed the capacity to extract evolutionary relations that are difficult to retrieve with tools that analyse just RNA or ATAC. Moreover, we did appreciate the possibility of running CONGAS+ also on multimodal data, where RNA and ATAC are measured from the same cell. Even in this case, our model could correctly phase the sequenced cells across the modalities, suggesting that CONGAS+ is also ready to process multi-omics data.

In the future, following the stream of work on CONGAS and CONGAS+, we plan to further data modalities, e.g., methylations, which require ad hoc methods to be processed [47]. Moreover, another possibility is to include finer-resolution information such as B-Allelic Frequency (BAF) and Depth Ratio (DR) profiles, as commonly used to detect CNAs from bulk [48]. This would allow one to infer copy-neutral losses of heterozygosity which, otherwise, are diploid and therefore poorly identifiable. Moreover, BAF/DP profiles might also be exploited in conjunction with read counts data to implement an algorithm for de novo genome segmentation and copy number calling, without requiring any input bulk DNA data.

Code and Data Availability

The CONGAS+ Python implementation and the R wrapping package are available at

- [Python] <https://github.com/caravagnalab/CONGASp>.
- [R] <https://github.com/caravagnalab/rcongas/tree/categorical> (*categorical* branch)

The code to reproduce the analyses in the text will be made available upon publication.

Acknowledgments

This work was funded by AIRC under MFAG 2020 - ID. 24913 project – P.I. Caravagna Giulio, and by the CRUK/AIRC Accelerator Award #22790, “Single-cell Cancer Evolution in the Clinic” (MA, AG and GC).

References

1. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer discovery*. 2022;12(1):31–46.
2. Acar A, Nichol D, Fernandez-Mateos J, Cresswell GD, Barozzi I, Hong SP, et al. Exploiting evolutionary steering to induce collateral drug sensitivity in cancer. *Nature communications*. 2020;11(1):1923.
3. Lim B, Lin Y, Navin N. Advancing cancer research and medicine with single-cell genomics. *Cancer cell*. 2020;37(4):456–470.
4. Hou X, Du C, Lu L, Yuan S, Zhan M, You P, et al. Opportunities and challenges of patient-derived models in cancer research: patient-derived xenografts, patient-derived organoid and patient-derived cells. *World Journal of Surgical Oncology*. 2022;20(1):1–9.
5. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods*. 2015;12(6):519–522.
6. Nam AS, Kim KT, Chaligne R, Izzo F, Ang C, Taylor J, et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature*. 2019;571(7765):355–360.

7. Belhocine K, DeMare L, Habern O. Single-Cell Multiomics: Simultaneous Epigenetic and Transcriptional Profiling: 10x Genomics shares experimental planning and sample preparation tips for the Chromium Single Cell Multiome ATAC+ Gene Expression system. *Genetic Engineering & Biotechnology News*. 2021;41(1):66–68.
8. Argelaguet R, Cuomo AS, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. *Nature biotechnology*. 2021;39(10):1202–1215.
9. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*. 2018;14(6):e8124.
10. Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC genomics*. 2019;20:1–11.
11. Tan K, Huang W, Hu J, Dong S. A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. *BMC Medical Informatics and Decision Making*. 2020;20:1–9.
12. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*. 2020;21(1):1–17.
13. Milite S, Bergamin R, Patruno L, Calonaci N, Caravagna G. A Bayesian method to cluster single-cell RNA sequencing data using copy number alterations. *Bioinformatics*. 2022;38(9):2512–2518.
14. Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 2019;.
15. Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nature communications*. 2020;11(1):89.
16. Campbell KR, Steif A, Laks E, Zahn H, Lai D, McPherson A, et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome biology*. 2019;20(1):1–12.
17. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature biotechnology*. 2021;39(5):599–608.
18. Wu CY, Lau BT, Kim HS, Sathe A, Grimes SM, Ji HP, et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nature biotechnology*. 2021;39(10):1259–1269.
19. Guilhamon P, Chesnelong C, Kushida MM, Nikolic A, Singhal D, MacLeod G, et al. Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival. *Elife*. 2021;10:e64090.
20. Nikolic A, Singhal D, Ellestad K, Johnston M, Shen Y, Gillmor A, et al. Copy-scAT: Deconvoluting single-cell chromatin accessibility of genetic subclones in cancer. *Science Advances*. 2021;7(42):eabg6045.

21. Fan J, Lee HO, Lee S, Ryu De, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome research*. 2018;28(8):1217–1227.
22. Gao T, Soldatov R, Sarkar H, Kurkiewicz A, Biederstedt E, Loh PR, et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nature Biotechnology*. 2023;41(3):417–426.
23. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews genetics*. 2008;9(2):102–114.
24. Robles-Espinoza CD, Mohammadi P, Bonilla X, Gutierrez-Arcelus M. Allele-specific expression: Applications in cancer and technical considerations. *Current opinion in genetics & development*. 2021;66:10–19.
25. Abramov S, Boytsov A, Bykova D, Penzar DD, Yevshin I, Kolmykov SK, et al. Landscape of allele-specific transcription factor binding in the human genome. *Nature communications*. 2021;12(1):2751.
26. Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, et al. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*. 2019;20(1):973–978.
27. Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:161101144*. 2016;.
28. Bishop CM, Nasrabadi NM. *Pattern recognition and machine learning*. vol. 4. Springer; 2006.
29. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*. 2000;22(7):719–725.
30. Navidi Z, Zhang L, Wang B. simATAC: a single-cell ATAC-seq simulation framework. *Genome biology*. 2021;22:1–16.
31. Baruzzo G, Patuzzi I, Di Camillo B. SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics*. 2020;36(5):1468–1475.
32. Ramazzotti D, Angaroni F, Maspero D, Ascolani G, Castiglioni I, Piazza R, et al. Lace: inference of cancer evolution models from longitudinal single-cell sequencing data. *Journal of Computational Science*. 2022;58:101523.
33. Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature genetics*. 2021;53(6):770–777.
34. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome biology*. 2022;23(1):27.
35. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nature medicine*. 2019;25(8):1251–1259.
36. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature biotechnology*. 2019;37(8):925–936.

37. 10x Genomics. Flash-Frozen Lymph Node with B Cell Lymphoma (14k sorted nuclei);. <https://www.10xgenomics.com/resources/datasets/fresh-frozen-lymph-node-with-b-cell-lymphoma-14-k-sorted-nuclei-1-standard>
38. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. 2018;.
39. Richards S, Watanabe C, Santos L, Craxton A, Clark EA. Regulation of B-cell entry into the cell cycle. *Immunological reviews*. 2008;224(1):183–200.
40. Wang Z, Peng M. A novel prognostic biomarker LCP2 correlates with metastatic melanoma-infiltrating CD8+ T cells. *Scientific reports*. 2021;11(1):1–12.
41. Jingjing Z, Lei M, Jie Z, Sha C, Yapeng H, Weimin Z, et al. A novel MEF2C mutation in lymphoid neoplasm diffuse large B-cell lymphoma promotes tumorigenesis by increasing c-JUN expression. *Naunyn-Schmiedeberg's Archives of Pharmacology*. 2020;393:1549–1558.
42. Miao Z, Ali A, Hu L, Zhao F, Yin C, Chen C, et al. Microtubule actin cross-linking factor 1, a novel potential target in cancer. *Cancer science*. 2017;108(10):1953–1958.
43. Scott DW, Wright GW, Williams PM, Lih CJ, Walsh W, Jaffe ES, et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood, The Journal of the American Society of Hematology*. 2014;123(8):1214–1217.
44. Lukow DA, Sheltzer JM. Chromosomal instability and aneuploidy as causes of cancer drug resistance. *Trends in Cancer*. 2022;8(1):43–53.
45. Taavitsainen S, Engedal N, Cao S, Handle F, Erickson A, Prekovic S, et al. Single-cell ATAC and RNA sequencing reveal pre-existing and persistent cells associated with prostate cancer relapse. *Nature communications*. 2021;12(1):5307.
46. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012;483(7391):570–575.
47. Shahryary Y, Hazarika RR, Johannes F. MethylStar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data. *BMC genomics*. 2020;21:1–8.
48. Househam J, Bergamin R, Milite S, Cross WC, Caravagna G. Integrated quality control of allele-specific copy numbers, mutations and tumour purity from cancer whole genome sequencing assays. *bioRxiv*. 2021; p. 2021–02.