


METHOD

Open Access



CaClust: linking genotype to transcriptional heterogeneity of follicular lymphoma using BCR and exomic variants

Kazimierz Oksza-Orzechowski^{1†}, Edwin Quinten^{2†}, Shadi Shafiqhi^{1,3}, Szymon M. Kielbasa⁴, Hugo W. van Kessel², Ruben A. L. de Groen², Joost S. P. Vermaat², Julieta H. Sepúlveda Yáñez^{2,5}, Marcelo A. Navarrete⁶, Hendrik Veelken², Cornelis A. M. van Bergen^{2†} and Ewa Szczurek^{1,7*†} 

[†]Kazimierz Oksza-Orzechowski, Edwin Quinten, Cornelis A. M. van Bergen, and Ewa Szczurek contributed equally to this work.

*Correspondence: szczurek@mimuw.edu.pl

¹ Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

² Department of Hematology, Leiden University Medical Center, Leiden, Netherlands

³ Cancer Research UK, Cambridge Institute, Cambridge, UK

⁴ Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

⁵ Facultad de Ciencias de la Salud, Universidad de Magallanes, Punta Arenas, Chile

⁶ Escuela de Medicina, Universidad de Magallanes, Punta Arenas, Chile

⁷ Institute of AI for Health, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Abstract

Tumours exhibit high genotypic and transcriptional heterogeneity. Both affect cancer progression and treatment, but have been predominantly studied separately in follicular lymphoma. To comprehensively investigate the evolution and genotype-to-phenotype maps in follicular lymphoma, we introduce CaClust, a probabilistic graphical model integrating deep whole exome, single-cell RNA and B-cell receptor sequencing data to infer clone genotypes, cell-to-clone mapping, and single-cell genotyping. CaClust outperforms a state-of-the-art model on simulated and patient data. In-depth analyses of single cells from four samples showcase effects of driver mutations, follicular lymphoma evolution, possible therapeutic targets, and single-cell genotyping that agrees with an independent targeted resequencing experiment.

Keywords: Cancer genetics, Tumour heterogeneity, Statistical methods, Follicular lymphoma

Background

From their onset, cancers are subject to continuous evolutionary processes, during which tumour cells acquire mutations in their genomes, forming clones and giving rise to genotypic heterogeneity [1–3]. At the same time, transcriptional heterogeneity is common, with various tumour cell subpopulations having different transcriptional programs. Cancer cell phenotype is thought to be predominantly driven by genetic alterations. However, recent studies suggest that distinct cancer cell states may emerge due to non-genetic factors [4–8]. In this context, a fundamental question arises: to what extent the observed transcriptional heterogeneity in tumours is explainable by the genotypic



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

differences between clones, and which remaining variance should rather be attributed to other effects? A better understanding of the genotype-phenotype link in cancer could guide personalised treatment and prompt development of novel therapies [6, 9–11].

Follicular lymphoma (FL) is a malignancy of mature B-cells that are arrested at the germinal centre stage and usually presents as pathological lymph nodes [12, 13]. FL is a paradigmatic disease with a clinical course that can range from spontaneous regression or stable disease for years, to transformation into aggressive diffuse large B-cell lymphoma. In addition to acquisition of mutations in BCR loci, aberrant somatic hypermutation in non-BCR loci causes accumulation of somatic variants that can be clonal (early events, present in all clones) or subclonal (later events, specific to a subset of clones) [14]. Under the assumption that acquisition of a potential oncogenic subclonal mutation occurs alongside BCR diversification, BCRs can serve as markers in the study of clonal evolution in FL tumours [15]. At the same time, FL cells may display transcriptional heterogeneity, with different transcriptional subpopulations displaying varying drug responses [16]. An example of a genotype to transcriptional phenotype link in FL is the presence of N-linked glycosylation motifs in BCR, which drives FL cells from a more light zone-like gene expression profile toward a dark zone-associated transcriptional program [17]. Still, the genotype to phenotype maps of FL have not been so far comprehensively studied.

The major obstacle in investigating the relation between genotypic and transcriptional heterogeneity in tumours is the fact that simultaneous DNA and RNA profiling of single cells is not possible using widespread experimental protocols. Indeed, innovations in this area emerged only very recently [18–20] and only limited throughput methods such as ResolveOME are commercially available [21]. To address this, computational methods were proposed that probabilistically match genomic alterations between bulk DNA sequencing and single cell RNA sequencing (scRNA-seq) or spatial transcriptomics data [15, 22–26]. However, we see an unmet need for methods dedicated to the specific case of FL. Indeed, standard phylogenetic methods cannot account for the parallel evolution of the exome and of the BCR sequences, where the latter proceeds with a high mutation rate. Dedicated methods are needed to utilise the statistical signal in the exome, the BCR sequences, and in the scRNA-seq to infer the evolutionary structure of the clones and their transcriptional phenotypes. In particular, our approach, CACTUS was previously applied to FL data by clustering cells by BCR sequences and performing cluster-to-clone assignment by mutation matching, benefiting from BCR information in this task [15]. However, the clustering of cells in CACTUS was effectively limited to grouping of cells with identical BCRs and required defining a hyperparameter corresponding to the (unknown) number of clusters. As an alternative to computational approaches, targeted DNA resequencing of single cells previously sequenced using scRNA-seq can also be used for genotyping cells [27]. Unfortunately, due to technical limitations it can only be performed with a very small number of variants, and is subject to noise in the scRNA data due to bursty gene expression.

Here, we use deep whole exome sequencing (WES), single cell RNA sequencing, single cell BCR sequencing (BCR-seq) and probabilistic modelling to deeply investigate the nature of tumour evolution as well as genotype to transcriptional phenotype interactions in four FL samples. To this end, we propose CaClust, a nonparametric Bayesian

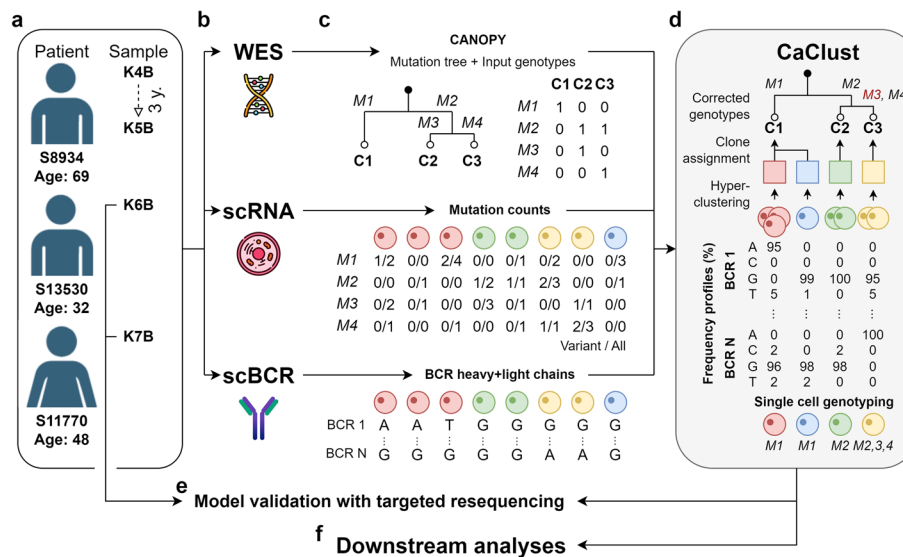


Fig. 1 Application of CaClust in this study: **a** 4 samples from 3 FL patients were chosen for inclusion in the study; **b–c** data collection and preprocessing; **d** model application to infer the clone genotypes, BCR hyperclustering with assignment, and clone clusters, output cell genotypes are obtained with the mapping of cells to their clone of origin; **e** after model application an additional resequencing experiment was performed on samples K6B and K7B to validate the output cell genotypes; **f** the output cell genotypes and clone structure were used in downstream analyses

extension of CACTUS [15], which is able to find a confident cell-to-clone mapping, improve estimation of clone genotypes, and infer genotypes for single cells. CaClust identifies the number of BCR clusters from the data and models probabilistic profiles of BCR sequences characteristic of every cluster. Applying CaClust to FL samples we find a range of genotype to phenotype links of varying strength. Focusing on the sample with the largest strength, we investigate mutations that could trigger specific transcriptional phenotypes. Additionally, we uncover the evolutionary link between two time-separated samples from another FL patient. In summary, our model enables comprehensive analysis of FL cell phenotypes in the context of their clonal origin, uncovering underlying tumour evolutionary mechanisms and targetable dependencies.

Results

Approach overview

We performed comprehensive molecular profiling of three FL patients: K4B and K5B, two samples from 69 year old male subject S8934 separated by 3 years, sample K6B from 32-year-old male subject S13530, as well as K7B from 48 year old woman S11770 (Fig. 1a). Each sample was profiled using WES at 1500× coverage. scRNA-seq and heavy and light chain scBCR-seq resulted in a total of 22,492 single cells sequenced at average 1620 transcriptome-wide genes, and average 24 full length umis of the expressed BCR genes per single cell.

In the CaClust model application pipeline, we first perform variant calling and copy number alteration (CNA) analysis on WES data, and next we use their output to estimate the input phylogenetic tree of the clones and their genotypes. From the scRNA-seq data, after standard alignment and mapping, we extract for each cell the variant and total

read counts at the single nucleotide variant (SNV) positions. From the BCR-seq data we extract the BCR sequence for each cell, comprising of the concatenated heavy and light BCR chains (Fig. 1c).

From this input, CaClust aims to simultaneously reconstruct *BCR hyperclusters* of cells while assigning those hyperclusters to tumour clones (Fig. 1d). Each BCR hypercluster is represented by its frequency profile of BCR nucleotides at different positions. We assume all cells in a hypercluster must come from the same clone. The tumour clones are represented by the SNVs present in the clone's genotype. The BCR hyperclusters are assigned to the clones by a probabilistic matching of variant reads at the SNV positions. Finally, *clone clusters* of cells are identified by tracking the clone that the BCR hypercluster of each cell is assigned to. Effectively, the cells are mapped to clones based on their shared BCR frequency profile characteristic of their hypercluster, as well as the genotype of its assigned clone, which then we also use to perform single cell genotyping (Fig. 1d). In this way, CaClust marries genotypes with phenotypes and enables detailed gene expression analysis of clones.

The output cellular genotypes obtained with our genotype-to-phenotype mapping were validated using simulated data and an independent resequencing experiment (Fig. 1e) and next used for downstream analyses of the evolutionary structure and the genotype to phenotype links in the FL samples (Fig. 1f).

CaClust model performance is validated on simulated data and independent targeted resequencing experiment

Before performing detailed downstream analyses, we validated the performance of the CaClust model on simulation scenarios with known ground truth and evaluated its quality on the FL samples, comparing against a predecessor model, as well as checked the correctness of its single cell genotyping with an independent targeted resequencing experiment.

We devised eight simulation scenarios (see the “[Methods](#)” section) varying three properties of a FL dataset: the scRNA read depth, the number of BCR hyperclusters, and the variance of BCR sequences within a hypercluster. A scenario with parameter values resembling the FL patient datasets used in this study was established as a baseline (referred to as *Basic*) and by varying one of these properties eight further scenarios were created with (i) high and (ii) low number of scRNA reads; referred to as *High reads* and *Low reads*, respectively; (iii) sparse BCR clustering (resulting in more BCR hyperclusters, referred to as *Sparse clusters*); (iv) high variance BCR sequences within hyperclusters (*High variance BCR*); and finally, four scenarios with centroid behaviour, where a portion x of cells within a fraction y of hyperclusters BCR sequence identical to the most probable of its hypercluster's BCR profile (in four versions of (x, y) v–viii): (0.8, 0.8), (0.8, 0.2), (0.2, 0.8), (0.2, 0.2); referred to as *x centroids in y clusters*). The simulation process and all scenario details are described in the “[Methods](#)” section. Both CaClust and its predecessor model CACTUS were applied to ten datasets simulated per each scenario and their performance was evaluated using three metrics: cell to clone assignment accuracy, accuracy of clone genotype reconstruction, and the quality of the hyperclustering reconstruction (see the “[Methods](#)” section).

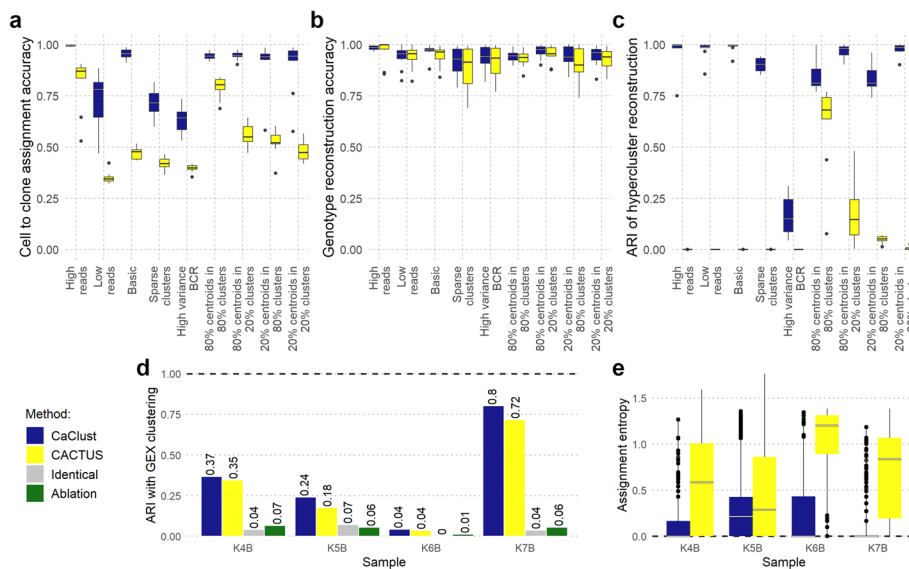


Fig. 2 CaClust validation results: **a–c** performance comparison on simulation scenarios vs. predecessor model CACTUS; **d–e** comparison on experimental data: **d** ARI agreement with gene expression clustering for CaClust clones, CACTUS clones, clusters of cells with identical BCR sequences, and hyperclusters from an ablation study that does not use scRNA variant data; **e** the assignment entropy of cells to clones in CaClust and CACTUS results. Bold dashed line: the value for the best possible clustering agreement as measured with ARI (**d**) and the value of the least entropy corresponding with the highest model certainty of assignment (**e**)

CaClust outperformed CACTUS in all scenarios and showed near perfect scores in the Basic type, High reads, as well as centroid scenarios (Fig. 2a–c). In terms of cell to clone assignment accuracy, the only scenarios that affected its performance were Low reads, sparse clusters and High variance BCRs. However, in all these scenarios CaClust significantly improved over CACTUS and kept median accuracy over 0.64 (Fig. 2a). In the task of genotype reconstruction CaClust showed a less pronounced improvement over CACTUS (Fig. 2b). Both models performed near perfect in reconstructing the true clone genotypes, with better reconstruction accuracy in scenarios with high scRNA read counts and worse accuracy in scenarios with degraded (sparse or high variance) BCR clustering. Hyperclustering reconstruction comparison again demonstrated a major advantage of CaClust (Fig. 2c). It achieved high adjusted rand index (ARI) scores in all but the high variance BCR scenario, which is specifically designed to give almost no BCR information. CACTUS performed poorly in all scenarios with basic and degraded BCR structure, since it did not manage to reconstruct hyperclusters with varied BCR sequences. However, in the scenarios with centroid behaviour, which can happen for real BCR data, its performance improves, albeit not to the level of CaClust.

Next, we compared performance of CaClust and CACTUS methods on the FL patient sample datasets by their sensitivity to the possible link between clone genotypes and transcriptional heterogeneity, measured using ARI against an independent cell clustering by gene expression (see the “Methods” section and Additional File 1: Fig. S1). We used two naive approaches as baselines: grouping cells with identical BCR sequences; and an ablation study with a stripped-down version of the CaClust model, which only produces hyperclusters with no further grouping into clones. CaClust achieved higher agreement with gene expression clustering as compared to CACTUS for all FL samples

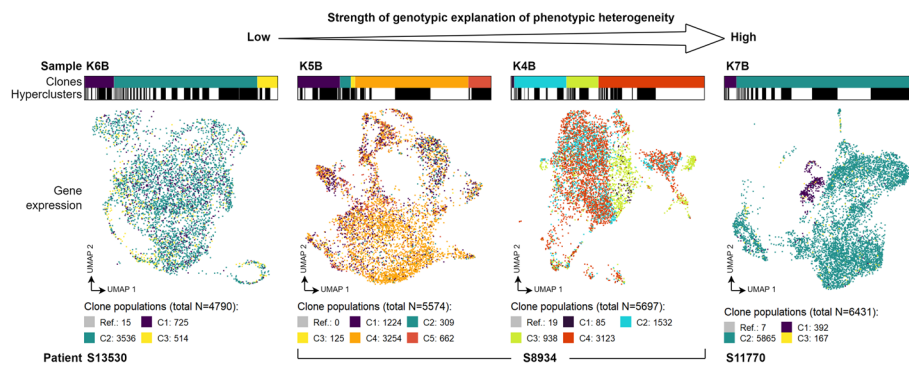


Fig. 3 Summary of observed strength of genotype-phenotype link in studied samples. Bars above UMAP plots show clone assignment of cells (colour coded) grouped by their hypercluster assignment (each solid white or black bar is one hypercluster). Cells in UMAP plots of gene expression are also colour coded by clone assignment. Clone populations are provided in the legend. The clones are not shared between samples. The samples are ordered by apparent influence of the clone genotypes on their phenotypic heterogeneity

(Fig. 2d). This indicates that the clustering of cells into clones found using CaClust identifies clones that are more distinct on the phenotypic level. Both models obtained higher ARI than the baselines, showing that the integration of both BCR and scRNA variant data is more accurate in finding phenotypically distinct populations of cells.

Further, we compared CaClust and CACTUS by their confidence of cell to clone assignment, measured with entropy. CaClust assigned cells to clones with lower entropy than CACTUS, indicating that the model was better able to extract information from the integrated data and reduce noise to make confident predictions (Fig. 2e, Additional File 1: Fig. S2).

Finally, we performed an additional targeted resequencing of samples K6B and K7B for independent experimental validation of cell genotyping performed by CaClust (see the “Methods” section). CaClust single cell genotypes show > 90% agreement for 4/8 resequenced variants, which increases to 6/8 variants when accounting for random monoallelic expression in the resequencing data (see Additional File 2: Supplementary Notes and Tables S1--2).

FL samples show different strengths of genotype-phenotype influence

To inspect the level of genotypic and transcriptional heterogeneity, and to assess the strength of genotype to phenotype link in the four analysed FL samples, we investigated the number and proportion of inferred clones, BCR hyperclusters, as well as visual agreement between the gene expression similarity and clone assignment (Fig. 3).

Sample K6B from patient S13530 displayed relatively uniform gene expression across its cells. The clustering of cells into the three clones with confidently assigned mutation profiles (see Additional File 3: Output profiles) and 97 identified BCR hyperclusters did not coincide with gene expression similarity (Fig. 3 left). This is in agreement with the lowest $ARI = 0.04$ obtained for that sample (compare Fig. 2d).

In time-separated samples K4B (with four clones and 82 BCR hyperclusters) and K5B (five clones and 32 BCR hyperclusters) coming from patient S8934, we observed higher phenotypic variance, with K4B showing more agreement between the transcriptional subpopulations and the inferred clones ($ARI = 0.37$ vs. $ARI = 0.24$, Figs. 2d and 3

middle). This could be the effect of specialisation in K5B, where its cells are more genetically homogeneous after genetic selection from K4B.

Sample K7B from patient S11770 (three clones and 92 BCR hyperclusters) displayed the strongest genotype-phenotype link ($ARI = 0.8$, Figs. 2d and 3 right), with clone C1 forming a separate expression cluster from clones C2 and C3. This points at subclonal variants highly affecting their carriers expression profiles.

To investigate whether some expression differences between clones are washed out by expression differences coming from cycling cells we performed cell cycle scoring and analysed the expression of cells from phase G1 in all samples. No additional expression cluster for a clone of any sample was found and overall expression structure was preserved as compared to the results obtained for all cells (Additional File 1: Fig. S3). Additionally, to detect whether there were any effects of subclonal mutations in sample K6B, we performed a targeted search in REACTOME C2 pathways containing K6B subclonal pathogenic variants, but found none of them to be enriched.

As samples K5B and K6B showed little to no link between the clonal genotypic structure and expression variation, we investigated other potential sources behind the observed transcriptional heterogeneity. Using Gene Ontology term enrichment of the top 250 most variably expressed genes in each sample we found the top 10 terms in each sample were tied to B-cell functions such as lymphocyte activation, positive regulation of immune system process, and immune response (Additional File 1: Fig. S4, Fig. S5). It is also possible that the observed transcriptional heterogeneity was caused by differences in the microenvironment, epigenetic factors or plasticity; however, these could not be analysed in this study's experimental setup.

In summary, our analysis revealed the clonal and BCR hypercluster structure of the FL tumour samples, and allowed ranking them by the strength of genotypic explanation of transcriptional phenotypic heterogeneity.

In depth investigation identifies four potential mutations driving clone phenotypes in patient sample K7B

To showcase the usefulness of our approach in the study of the genotype-phenotype link, we performed a detailed analysis of the results in patient sample K7B.

CaClust identified three tumour clones, with phenotypes showing the effects of known mutations. The output clone genotypes (Fig. 4a) included four subclonal mutations for which we predicted a pathogenic effect (see the “Methods” section): (i) *MYD88*(L265P) mutation, (ii) a variant in the *TSPAN33* gene (both shared between clones C2 and C3), (iii) a missense variant in the HAT domain of the *CREBBP* gene (specific to clone C2), and (iv) an early stop variant *VMA21*(R93X) (specific to clone C1). Clones C2 and C3 shared a larger number of SNVs and appeared to be evolutionally closer to each other than to clone C1.

The model assigned the cells to clones with a very high confidence (Fig. 4b). Only a small fraction of cells had their assignment probability mixed between C2 and C3. This was in line with the aforementioned higher genotypic similarity of C2 and C3. Among the resulting clone clusters, clone C2 was the most prevalent (5865 cells), with C1 and C3 being smaller in size (392 and 167 cells, respectively).

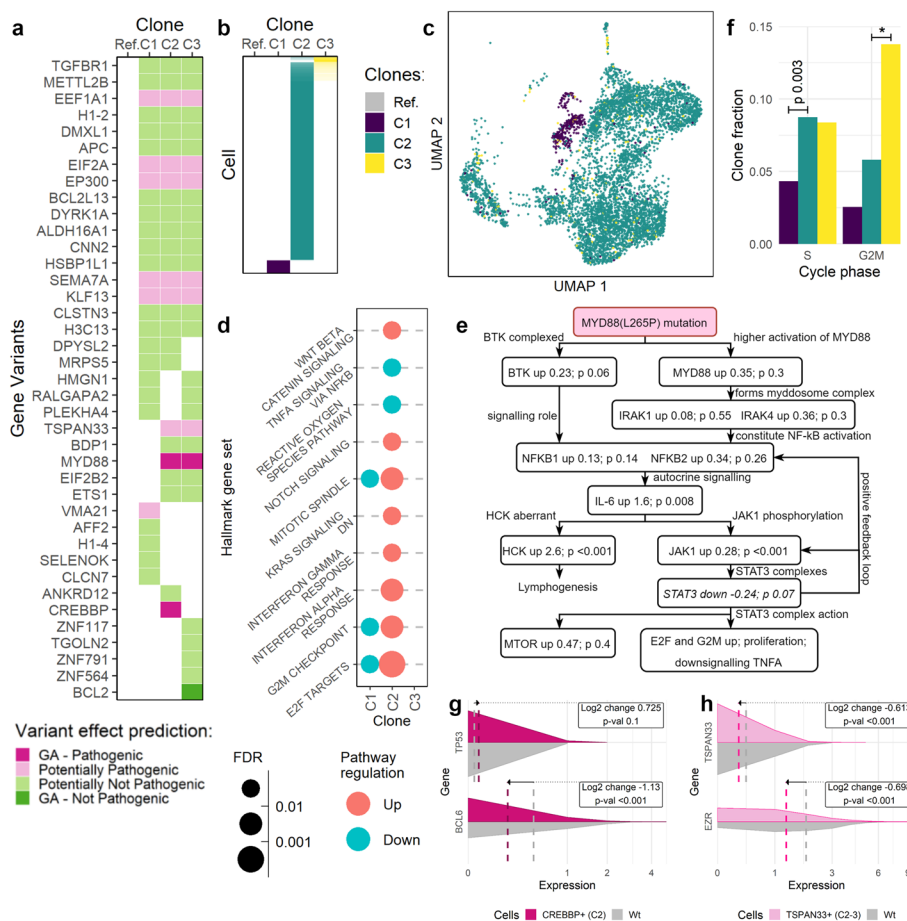


Fig. 4 Analysis of CaClust results on sample K7B: **a** output clone genotypes; **b** output cell-to-clone assignment probabilities, higher probability denoted with higher opacity; **c** UMAP reduction of cells' gene expression, coloured by clone assignment; **d** results of GSEA analysis; **e** graph of predicted MYD88(L265P) effects compared with observed DE results; **f** distribution of cells in clones across cell cycle phases; **g** estimated expression of TP53 and BCL6 genes in CREBBP+ cells vs. wildtype cells; **h** estimated expression of TSPAN33 and EZR genes in TSPAN33+ cells vs. wildtype cells. *p < 0.001; Wt, wildtype

When the UMAP reduction of cells' gene expression was overlaid with clone clusters (Fig. 4c; reproduced from Fig. 3), clone C1 formed its own clear expression cluster, whereas clones C2 and C3 appeared more mixed. This hints at gene expression being highly affected by a subclonal variant, either one that is characteristic to C1 or shared between C2 and C3.

To study the effects of subclonal variants, we first performed differential gene expression analysis between each clone and the rest (see the "Methods" section), finding 918 differentially expressed genes. Next, to check for enriched pathways we performed gene set enrichment analysis (Fig. 4d). Clone C2 showed upregulated pathways linked to cell proliferation (E2F targets, G2M checkpoint, mitotic spindle (respective FDRs: < 0.001, 0.004, 0.007; all FDRs estimated with gene label reshuffling in GSEA) and down-regulated tumour necrosis factor alpha signalling (FDR 0.09), both hinting at gained advantages over the other clones.

Next, we investigated whether the known effects of the *MYD88*(L265P) mutation could be observed. Based on a comprehensive study [28], we created the graph of the predicted effects of that mutation in tumour cells as compared to healthy B-cells, and indicated their agreement with our own data (Fig. 4e). Most observed effects agreed with the known ones, with the exception of the downregulated *STAT3* expression (log-foldchange -0.236 , adj. p -val 0.07 , all log fold-change p -values obtained with LRT from DESeq2, see the “Methods” section). However, JAK1 affects the formation of the *STAT3* complex through protein-protein interaction, and the *STAT3* downregulation in expression may be independent of that process. Moreover, the known effects relate to the comparison of healthy B-cells with *MYD88*(L265P) carriers, whereas our analysis compares tumour cells with multiple additional mutations against each other, which can introduce confounding effects. Since the ultimate effect of the *MYD88*(L265P) is increased tumour proliferation and cell survival, we also analysed the cell cycle distribution of clones (Fig. 4f, Additional File~2: Table S3). As expected, cells in C2 and C3 that harbour this mutation proliferated faster, as measured by the fraction of the cells in those clones entering the S phase (hypergeometric test: $p < 0.001$, see the “Methods” section). Additionally, since increased proliferation should require higher energy production, we analysed the enrichment of the beta-oxidation pathway in clones C2 and C3, and found 20/25 genes to be upregulated (see Additional File 1: Fig. S6).

We next investigated whether cells assigned to clone C2 follow the behaviour known for carriers of the CREBBP variant. CREBBP proteins with a missense variant in the HAT domain are unable to acetylate the tumour suppressor TP53 and BCL6 oncogene, preventing the tumour suppression mechanisms [29]. As expected, cells assigned to clone C2, even though they express more *TP53* than cells in clones C1 and C3 (log fold-change: 0.725 , $p_{adj} = 0.1$) and less *BCL6* (-1.13 , $p_{adj} < 0.001$; Fig. 4g), which should lead to cell cycle arrest, pass through the G2M checkpoint normally. In contrast, cells from clone 3 are stuck in in the G2M phase (hypergeometric test: $p < 0.001$; Fig. 4f).

The TSPAN33 protein has been linked to a migratory phenotype in the B-lymphocytes. It forms complexes on the B-cell membrane with the EZR protein, and its overexpression was linked to an increase in B-cell migration by Navarro-Hernandez et al. [30]. In clones C2 and C3 showing the *TSPAN33* missense variant, we observed a decrease in the expression of both *TSPAN33* and *EZR* (log fold-changes: -0.613 , -0.698 , $p_{adj} < 0.001$, Fig. 4h), which could point to a decrease in migratory capabilities over C1.

The observed *VMA21*(R93X) mutation and was shown by Wang et al. [31] to result in a targetable survival dependency. Specifically, VMA21 is a chaperone protein that takes part in the V-ATPase assembly and the mutation p.R93X results in a premature stop and a loss of the C terminus at AA93-101. Consequently, VMA21 is mislocated to the lysosomes, leading to impaired V-ATPase ability to acidify lysosomes that is compensated by an increase in autophagic flux. In the Wang et al. study, treatment of *VMA21*(R93X) B-cells with an inhibitor of autophagy regulating ULK1 kinase complex led to their death, while wildtype B-cells remained mostly unaffected; thus, the clone C1 exhibiting *VMA21*(R93X) could also be targetable by such a therapy.

In summary, our analysis firstly identified which effects of known subclonal mutations can be observed in the phenotypes of specific clones; secondly, pointed to the effects of potentially pathogenic subclonal mutations, which have not been studied yet; lastly, can

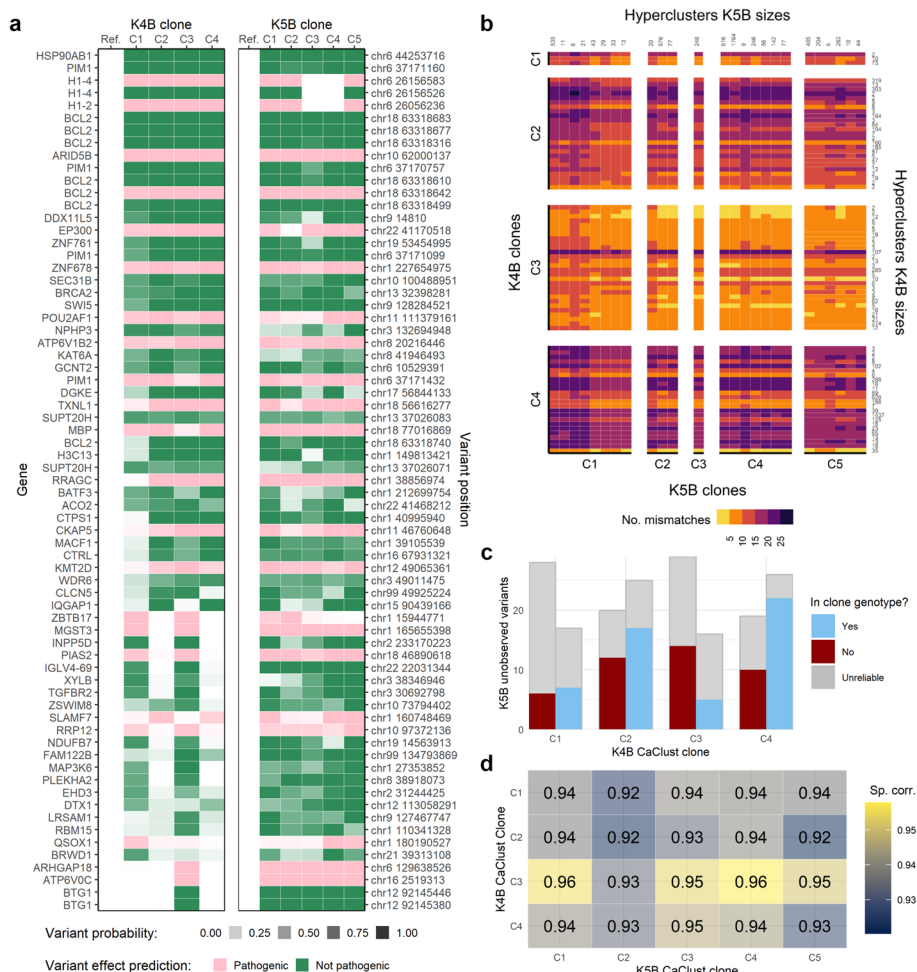


Fig. 5 Analysis of the link between K4B and K5B samples using CaClust results: **a** output probabilities of variants common between the samples to be present in clone genotypes; **b** Hamming distance between BCR hyperclusters in K4B and K5B, sorted by their clone assignments; **c** K4B variants unobserved in K5B, by their presence or absence in K4B clone genotypes; **d** Spearman correlation between gene expression of K4B and K5B clones. Sp. corr., Spearman correlation

potentially guide the treatment choice by identifying clones with mutations known to be susceptible to targetable therapy.

Inferred clone genotypes suggest evolutionary history of the time-related samples

To further showcase the usefulness of our method in studying tumour evolution, we investigated the relation between samples K4B and K5B. Both of those samples were taken from patient S8934, 3 years apart. In CaClust results for sample K4B, we found evidence for a possible founder clone of the tumour in sample K5B, with both genetic and phenotypic similarity.

Firstly, we investigated the 69 SNV variants that were common for both samples and were included in the model analysis (Fig. 5a). Variant probabilities for each clone of K4B indicated that clone C3 contained almost all of the common variants with a high probability, while the other K4B clones were missing many of them. Most of those variants

in the K5B sample were clonal (i.e., present in all clones), which is expected for a later, possibly descendant tumour sample. This result suggests the possible descent of all clones found in the later K5B sample from clone C3 or its close ancestor in the evolution of the earlier K4B sample. However, three variants (two in *HI-4* gene and one in *HI-2*) were absent from clones C3 and C4. Those SNVs belonged to a region (chr6 positions 26056K–26156K) that, based on the WES CNA analysis of sample K5B, was found to be affected by a copy number alteration (deletion), which could have happened in the course of evolution of clones C3 and C4 in K5B.

Secondly, we investigated the BCR similarity of cells in hyperclusters in K4B to hyperclusters in K5B (Fig. 5b). It should be noted, that no exact same BCR sequence was shared between samples; therefore, there is no immediate candidate hypercluster that could be the founder of K5B. However, we again found that cells in hyperclusters belonging to clone C3 in K4B showed the highest degree of similarity (number of BCR nucleotide mismatches) to hyperclusters in all clones from K5B, further demonstrating the evolutionary similarity of K4B clone C3 to sample K5B.

Thirdly, WES data of sample K4B contained SNVs that were not found in the WES data of sample K5B, which suggests some parallel evolution between the samples. We investigated the subclonality of 45 SNVs that were used in the K4B model inference but have not been observed in K5B WES data. For each clone, we checked how many of those 45 SNVs were inferred to be present in its genotype and how many were not. We considered a genotyping call on a variant position in a clone as reliable if the scRNA data of cells mapped to that clone contained at least one mutated read or three reference reads (Fig. 5c). Clone C3 has the fewest reliably called SNVs (5) not observed in K5B. This again highlights the similarity of clone C3 to the K5B sample, and while the presence of additional mutations may not make it the exact predecessor in evolution, their common ancestor could be very close up the evolution tree. This finding is in line with [14], where time-related FL samples were shown to come from a common progenitor clone (CPC), rather than being direct descendants (for divergent evolution in FL see also [32]).

We also checked the correlation of gene expression between the clones of K4B and K5B. We took the union of the top 100 most variably expressed genes in each sample and calculated the Spearman correlation coefficient between their average expression in clones of K4B and clones of K5B (Fig. 5d). Here again clone C3 from K4B had the highest correlation of expression to clones from K5B; however, it has to be noted that the overall correlation scores were very high. That is expected, since it was a comparison of malignant B-cells, which had a single ancestor cell and a high overall similarity could be expected.

It is important to mention that although CANOPY is equipped with a multi-sample deconvolution method, it was unable to produce a phylogeny that would explain the evolutionary dependency between the samples. One would expect such a phylogeny to contain clones that have K4B specific mutations and others that have K5B specific mutations, with K4B clones being potentially related to K5B ancestors. However, output CANOPY phylogenies for both 5 and 10 clones (5 being chosen by BIC and 10 being the number of clones expected by single sample outputs: 1 reference clone, 4 tumour clones in *K4B* and 5 tumour clones in *K5B*) put sample specific mutations at all levels of the

evolution tree, which indicates that deconvolution based on allele frequencies is not sufficient in this complex case (see Additional File 1: Fig. S7, Fig. S8).

In summary, thanks to accurate clonal mapping using CaClust, we could resolve the clonal structure and ancestry connections between the time-related samples, even in the case when the multi-sample CANOPY model was unable to reliably combine the two samples in a unified phylogenetic inference.

Higher confidence assignment of cells to clones by CaClust facilitates more informative downstream analysis as compared to CACTUS

As shown in Fig. 2e and Additional File 1: Fig. S2, CaClust assigned cells with much higher confidence. The uncertainty in cell to clone assignment, characteristic for CACTUS, is particularly destructive, as noise in model assignment can mix clones and harm downstream analyses. This is evident in the case of CACTUS results for sample K7B. Its assignment of cells to clones shows much higher degree of noise than CaClust, and its clones are more mixed by expression (Additional File 1: Fig. S9). This in turn affects DE analysis, where CACTUS finds fewer differentially expressed genes with smaller statistical significance (Additional File 1: Fig. S10), and its clones no longer show significant differences in cell cycle distributions (see Additional File 1: Fig. S11). In the worst case scenario, the noise in CACTUS assignment can completely wash-out the results, as is the case with sample K6B, where based solely on CACTUS results we would be unable to tell whether a clonal structure is present at all, and what its relation to the expression heterogeneity is (Additional File 1: Fig. S12). These comparisons demonstrated that due to more confident assignments of cells to clones, CaClust is able to deliver more biological findings and clearer conclusions from the data than CACTUS.

Analysis of CaClust clonal genotypes shows agreement with input phylogenies

During the inference procedure CaClust performs corrections of the genotypes from input trees obtained with CANOPY to match the observations in the scRNA data of the cells assigned to clones (see the “Methods” section). Since CANOPY infers the phylogenies and the genotypes solely based on deconvolution of WES allele frequencies, it may provide erroneous genotypes at input, and the corrections are an attempt to take into account the additional evidence in scRNA-seq data. However, no tree structure is imposed during those corrections and as a result the output genotypes may not form a phylogeny.

To check how much the corrections changed the input phylogenies behind the genotypes for the analysed samples, we first performed hierarchical clustering (Manhattan distance, single linkage) on the corrected genotypes, and next compared the obtained trees with the input CANOPY phylogenies (see Additional File 1: Fig. S13). The phylogenies and the tree structures matched in all samples (up to clone reindexing, which can change during inference), with clone fractions at corresponding levels of the phylogenies differing moderately from those estimated by Canppy. This indicates that although some changes to the clonal structure were made in CaClust inference, the underlying phylogenetic relationships between clones stayed the same.

We next considered the possible reasons behind those genotype corrections made by CaClust that resulted in potential violations to the phylogeny.

To this end, we investigated in detail the corrected genotypes of sample K7B (Fig. 4a). For this sample, the tree structure over corrected genotypes cannot be established due to *DPYSL2*, *MRPS5*, *HMGNI*, *RALGAPA2*, and *PLEKHA4* variants. Analysis of read coverage for the corrected positions revealed that those positions have extremely low read counts in the small clone C3 (1, 3, 0, 0, and 5 reference reads respectively, no alternate reads), and it is possible that the model cannot make a confident and correct genotype call on these positions. For such low read count positions, we hypothesise that enforcing the phylogenetic tree structure would introduce prior knowledge to the model, helping it to make better calls.

As another example we investigated the output genotypes for sample K5B on the positions common with sample K4B for *H1-4* and *H1-2* variants (Fig. 5). These variants are called as absent in C3 and C4, thus breaking a phylogeny. However, as mentioned in the previous section, these positions are affected by a CNA in sample K5B and are thus probably absent due to a deletion. Therefore, the genotypes corrected by CaClust for this example are likely correct and the phylogenetic tree behind the clones could be reconstructed by accounting for the CNAs in addition to SNVs.

Discussion

In this work we combined in-depth molecular profiling of patient samples with probabilistic modelling to investigate the evolutionary histories and to explain the relationship between genomic and transcriptional heterogeneity in FL. To this end, we performed WES, scRNA-seq, BCR-seq and targeted resequencing and introduced CaClust, a novel method for clonal phenotype profiling with single cell genotyping in FL.

CaClust integrates BCR, WES, and scRNA information for increased accuracy and confidence. Since we consider that the evolution of BCR sequences proceeds much faster than and in parallel with the evolution of the rest of the genome, the CaClust model makes an important assumption that cells with similar BCR sequences belong to the same genetic clone. This assumption, combined with the use of nonparametric Bayesian clustering allow the new CaClust model to efficiently pool scRNA information on the clonal assignment of FL cells based on their BCR similarity. By pooling the single cells together into BCR hyperclusters, the model circumvents the biggest problem in single-cell genotyping based on scRNA sequencing data, which is the sparsity of reads in each cell. Moreover, the approach used is flexible in that it infers the optimal number of hyperclusters along with their BCR profiles from the data. As we have demonstrated both on simulated and experimental datasets, this greatly improves the clonal profile reconstruction, cell assignment, and genotyping accuracy over a rigid BCR clustering, which was previously shown to perform best in those tasks for FL [15]. The newly improved clonal and single-cell genotypes obtained with CaClust enable multiple downstream analyses that can shed light on the effects of driver mutations, possible therapeutic targets, and parallel evolution of time-related FL samples, as demonstrated on data from 4 samples from 3 patients.

While CaClust was specifically developed to model the concurrent evolution of BCR loci and mutations in other parts of the genome, it could easily be adapted to other data tied to the clonal evolution. That is, any other information source can be used for hyperclustering in the model, provided it meets the assumption that cells in one hypercluster

should belong to the same evolutionary clone. Additionally, the CaClust model could be extended to enforce a phylogenetic structure during its genotype correction procedure, which could further improve the results in the case of variant positions with low coverage in small clones.

However, limitations of our method still exist. Firstly, the used input scRNA data is based on 5' end sequencing, which does not capture the full transcriptome and thus some variants could be potentially missed. Secondly, even despite the high depth sequencing and pooling effect that the hyperclusters provide for scRNA reads, some variants still did not have sufficient coverage to make reliable genotyping calls in clones. Thirdly, not all variants will result in a major phenotypic difference in their clones, thus in some cases of more homogeneous FL, the analysis will bring less discoveries. Lastly, to carry out an adequately powerful study of FL biology for novel discoveries with CaClust would require collecting hundreds of patient samples, with each needing the specific combination of WES, scRNA, and scBCR data.

Despite these limitations, our approach brings important insights into the ongoing debate on the sources of intratumour heterogeneity. Comparison to previous model CACTUS and simpler baselines showed that with a model that is more robust to noise and smarter in data integration, more transcriptional heterogeneity can be explained by genetic causes (Fig. 2d). Only having established the more likely genotype to transcriptional phenotype link should the remaining phenotypic variance be attributed to other effects, for which the mechanisms are less clear. With its excellent performance and rich output for downstream analysis, CaClust proved highly useful in the study of heterogeneity in FL, by extracting the phenotype-to-genotype mapping from high throughput sequencing data into an easily interpretable structure of hyperclusters and clone clusters.

Conclusions

In this work, we proposed CaClust, a novel method for clonal phenotype profiling with single cell genotyping in FL and demonstrated its potential use in the study of evolutionary histories and the relationship between genomic and transcriptional heterogeneity in FL. To our knowledge, our approach is the first to enable the joint study of these two types of heterogeneity in FL and to evaluate the strength of genotype-to-phenotype links in the evolutionary context of BCR hypermutation. Our in-depth analysis of 22,492 single cells and whole exomes from four FL samples using CaClust gives insights into effects of driver mutations, possible therapeutic targets, and FL evolution.

Firstly, as model validation we showed that CaClust outperforms a state-of-the-art model on simulated and patient data. Secondly, we demonstrated that CaClust single-cell genotyping agrees with genotypes observed in an independent targeted resequencing experiment. Additionally, our investigation of CaClust clones identified potential mutations driving clone phenotypes in patient sample K7B, which include two known pathogenic variants of *MYD88* and *CREBBP*, a *VMA21* variant causing a targetable dependency, and a novel *TSPAN33* variant; for mutations with known pathogenicity, their effects were observed in the expression phenotypes of their carriers. Lastly, the inferred clone genotypes and BCR hypercluster profiles of the time-related samples K4B and K5B gave hints of the evolutionary history of their clones, that agree with the findings on CPCs from [14].

Altogether our results illustrate that CaClust greatly facilitates an effective study of the extensive genomic and transcriptional heterogeneity in FL and their link, by providing the first method for their joint analysis utilising the context of BCR hypermutation.

Methods

Data collection

Patient sample collection

Samples with histologically confirmed infiltration of follicular lymphoma (FL) grade 1–2 were collected according to the Declaration of Helsinki and under authorization of applicable biobank regulations of Leiden University Medical Center and the Ethical Committee of Leiden University Medical Center (reference HEM 008/SH/sh). Excisional lymph node biopsies were processed immediately by gentle mechanical disruption and mesh filtration. Single cell suspensions were frozen in 10% DMSO and remaining tissue fractions were cultured in low-glucose (1 g/L) DMEM with 8% foetal bovine serum to obtain adherent cell cultures for isolation of DNA of cells representing normal counterpart.

Cell processing, library preparation and sequencing

FL cells were thawed and purified by flowcytometry using anti-CD19-APC (Becton Dickinson, Franklin Lakes, NJ) and anti-CD10-PECy7 (Becton Dickinson) followed by removal of dead cells (MACS Dead Cell Removal Kit, Miltenyi Biotec, Bergisch Gladbach, Germany). For whole exome sequencing, DNA was isolated from $1 \cdot 10^6$ purified FL cells and from $0.5 \cdot 10^6$ cultured adherent cells (Allprep DNA/RNA Mini Kit, Qiagen, Hilden, Germany). Whole exomes were sequenced using SureSelect Human All Exon V7 baits (Agilent, Santa Clara, CA). Adherent cells representing normal cells were sequenced at $50\times$ coverage. To discriminate between early clonal variants and more recently acquired subclonal variants as putative drivers of distinct clones, FL bulk DNA was sequenced at $1500\times$ coverage to allow reliable calling of rare variants down to a variant allele frequency (VAF) of 0.02. For 5'-based single cell transcriptome sequencing, $1 \cdot 10^5$ similarly purified viable cells were loaded on a Chromium X single cell device to generate cDNA libraries for an expected $6 \cdot 10^3$ – $8 \cdot 10^3$ cells per sample. (10X Genomics, San Francisco, CA) Inside the 10X Genomics chip, single cells and oligonucleotide-covered beads are simultaneously captured as aqueous droplets in oil. Per bead, all oligonucleotides share an identical single cell barcode (scbc), and every single oligonucleotide molecule carries a unique molecular identifier (UMI). In the droplet, cells are lysed and cDNA synthesis is 3' primed with oligo-dT. After amplification of the primary cDNA library by using universal primers, the amplified library is split in 3 fractions for (1) full transcriptome sequencing, (2) enrichment of BCR transcripts followed by full length sequencing, (3) targeted resequencing of subclonal somatic variants. Sequencing for WES and single cells was performed on HiSeq2500 or HiSeq4000 devices (Illumina, San Diego, CA).

Variant calling

FASTQ files from whole exome sequencing (WES) were processed using the Sarek workflow v2.7 and aligned to the human reference genome GRCh38 using Burrows Wheeler Algorithm (BWA) v0.7.17. [33, 34] Duplicated mapped reads were marked,

local realignment of regions flanking indels and recalibration of base quality scores were performed to obtain more accurate bases according to the Genome Analysis ToolKit (GATK) best practices version v4.1.7.0. [35] Single nucleotide variants (SNV) and short insertions and deletions (INDELS) were called using Strelka2 v2.9.10. [36] Only high confidence variants defined by quality scores (GQX) of at least 15 for SNV and 30 for INDELS were kept. Pathogenicity of variants was determined using the Geneticist Assistant NGS Interpretive Workbench (SoftGenetics) based on publicly available variant-databases (dbSNP, ClinVar, and COSMIC) and literature, into class 1 (benign), class 2 (likely benign), class 3 (unknown significance), class 4 (likely pathogenic), or class 5 (pathogenic) [37–39].

Variant selection for CaClust modelling

We chose only somatic single nucleotide variants called from Strelka that could also be observed in the scRNA data of the cells with complete BCR heavy and light chain sequences, and that showed at least one alternate read across the cells.

Some germline variants were present in the Strelka output, as they had an increased frequency of the alternate nucleotide over the normal sample. We chose to include only those variants that showed over 3-fold increase in the frequency of the alternate nucleotide between the normal and tumour sample. The full table of included variants per sample can be seen in Additional File 3: Output profiles.

Copy number inference (FalconX)

We use FalconX [40] for the inference of copy number alteration (CNA) events. In the getASCN.x method we use a threshold of 0.1; later, in the quality-filtering falconx.qc we set the CNA length.threshold of 10^7 basepairs and the delta copynumber threshold of 0.1.

Inference of input clonal profiles

We use CANOPY [41] for the estimation of input clonal profiles for the model. We run 5 CANOPY chains for each clone number $K \in 3, \dots, 6$, choosing the number of clones for each sample with the highest BIC score. The minimum number of model iterations is set to 20,000 and the maximum to 100,000. The input CNAs for CANOPY inference are obtained with FalconX.

The CaClust model formulation

CaClust can be seen as a significant extension to our previous model, CACTUS, with the functionality of non-parametric Bayesian clustering applied to BCR sequences (Fig. 6).

We assume we are given a cancer tissue sample with WES, scRNA, and BCR receptor profiling. Let $i \in \{1, \dots, N\}$ denote a position of a SNV that can be found both in WES and scRNA data. We assume that a set of $k \in \{1, \dots, K\}$ clones is given, each with a distinct genotype. We describe the given clone genotypes with a matrix Ω , where an entry $\Omega_{i,k}$ is 1 if variant i is present in the given genotype of clone k and 0 if it is not.

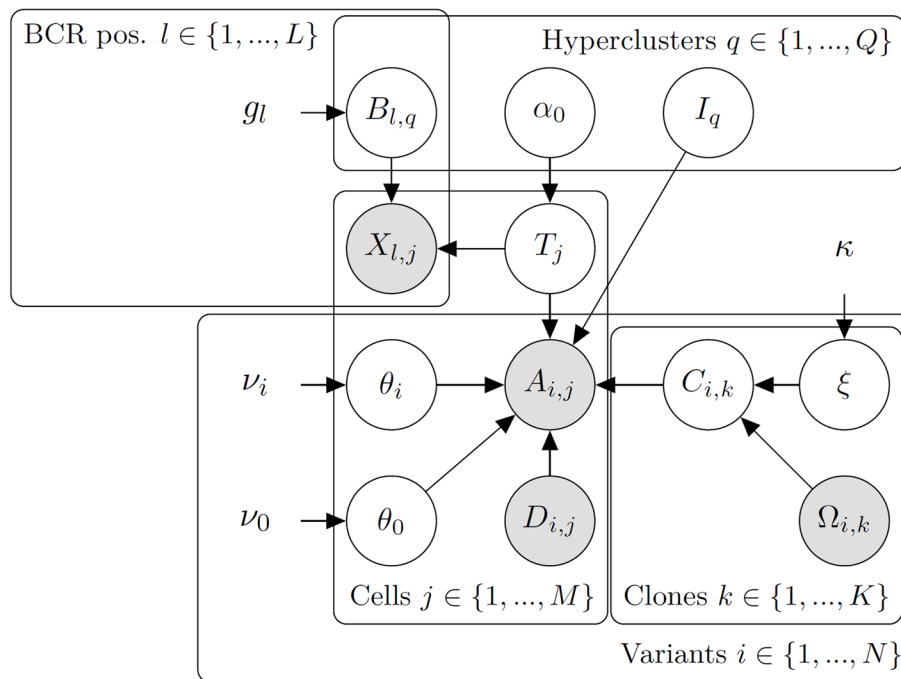


Fig. 6 Graph of the CaClust model. White vertices represent hidden variables, grey vertices represent observed variables. Directed edges show probabilistic dependencies between variables. Vertices with no outline are model parameters. $\Omega_{i,k}$ is the observed genotype of clone k at position i , $C_{i,k}$ being the true genotype: $C_{i,k} = 1$ if mutation i is present in clone k and 0 otherwise. ξ is the error rate between $\Omega_{i,k}$ and $C_{i,k}$. $A_{i,j}$ is the count of observed unique molecules with an alternate nucleotide at position i in cell j , with $D_{i,j}$ being the total number of unique molecules observed at position i in cell j . θ_0 is the probability of observing an alternate read if a cell does not carry a mutation at that position, θ_i is the probability of observing an alternate read if a cell does carry mutation i . I_q is the assignment of hypercluster q to one of the tumour clones. T_j is the assignment of a cell j to one of the hyperclusters and α_0 is the concentration parameter of the CRP prior of T_j . $B_{l,q}$ is the vector of nucleotide frequencies at position l in BCR sequences of cells from hypercluster q . $X_{l,j}$ is the nucleotide present at position l in the BCR sequence of cell j . κ , ν_0 , ν_1 , g are the hyperparameters of the model

As in CACTUS, the input matrix of clone genotypes is assumed to be imperfect, containing errors with rate ξ . We take ξ with a prior distribution Beta with parameters $\kappa = (\kappa_0, \kappa_1)$, obtaining $\mathbb{P}(\xi|\kappa) = \text{Beta}(\xi; \kappa_0, \kappa_1)$. We then introduce the matrix C , where $C_{i,k}$ are the hidden variables representing the true genotype of clone k at variant position i , such that:

$$\mathbb{P}(C_{i,k} = 1|\Omega_{i,k}, \xi) = \begin{cases} 1 - \xi, & \Omega_{i,k} = 1 \\ \xi, & \Omega_{i,k} = 0 \end{cases}$$

The main advantage of CaClust over its predecessor CACTUS is the way the clustering of cells by their BCR receptors is modelled. Let $j \in \{1, \dots, M\}$ be the cell indices. T_j denotes the BCR hypercluster to which cell j is assigned. Contrary to its predecessor, CaClust does not consider a fixed number of BCR hyperclusters, but rather allows it to be inferred, using the Chinese Restaurant Process (CRP) for the prior, i.e.:

$$\mathbb{P}(T_j|T_{-j}, \alpha_0) = \text{CRP}(T_{-j}, \alpha_0),$$

where T_{-j} is the hypercluster assignment of all cells but j to the BCR hyperclusters in the model, and α_0 is the concentration parameter.

We characterise each BCR hypercluster q of cells with its BCR frequency profile B_q and with \mathbf{B} we denote all those profiles in the model. Let $l \in \{1, \dots, L\}$ index BCR positions. Then the vector $B_{l,q} = [B_{A,l,q}, B_{C,l,q}, B_{G,l,q}, B_{T,l,q}]$ describes the probabilities, with which a cell belonging to hypercluster q has at position l a nucleotide A, C, G or T , respectively. We set a Dirichlet prior on $B_{l,q}$ with parameters $g = (g_A, g_C, g_G, g_T)$:

$$\mathbb{P}(B_{l,q}|g) = \text{Dirichlet}(B_{l,q}|g_A, g_C, g_G, g_T).$$

For each cell j we are given its BCR sequence as X_j , where $X_{j,l}$ is the nucleotide at position l ; with \mathbf{X} we denote the matrix of all cells' BCR sequences. Given the hyperclustering and its BCR frequency profiles we treat each cell's BCR as coming from a categorical distribution with probabilities described by its hypercluster's profile. So for a cell j , its hypercluster T_j and the BCR frequency profile, we have:

$$\mathbb{P}(X_{j,l} = Nuc|\mathbf{B}, T_j = q) = B_{Nuc,l,q},$$

where Nuc is one of the four nucleotides, $Nuc \in \{A, C, G, T\}$.

We make the assumption that cells from the same hypercluster belong to the same tumour clone, so we want to find the hypercluster to clone assignment. By I_q we denote the tumour clone that hypercluster q is assigned to and with \mathbf{I} the assignment of all hyperclusters in the model. We make no prior assumptions on that assignment and so we set a uniform prior distribution: $\mathbb{P}(I_q = k) = \frac{1}{K}$.

From the scRNA data we create matrices \mathbf{A} and \mathbf{D} , where $A_{i,j}$ and $D_{i,j}$ are the numbers of alternate and total reads respectively, that map to position i in cell j . We then define observation probabilities $\theta = (\theta_0, \theta_i)$: θ_0 is the probability of a read being mutated if it comes from a cell mapping to a clone that does not have that mutation in its genotype, θ_i is the probability of a read being mutated if it comes from a cell mapping to a clone that does have mutation i in its genotype. Then, the likelihood of observing $A_{i,j}$ mutated reads from $D_{i,j}$ total reads is:

$$\mathbb{P}(A_{i,j}|D_{i,j}, I_q, C_{i,I_q}, \theta, T_j = q) = \begin{cases} \text{Binom}(A_{i,j}|D_{i,j}, \theta_0), & C_{i,k} = 0 \\ \text{Binom}(A_{i,j}|D_{i,j}, \theta_i), & C_{i,k} = 1 \end{cases}.$$

We pick beta priors for θ_0 and θ_i , with parameters (a_0, b_0) and (a_i, b_i) , respectively:

$$\mathbb{P}(\theta_0|v_0) = \text{Beta}(\theta_0|a_0, b_0)$$

$$\mathbb{P}(\theta_i|v_i) = \text{Beta}(\theta_i|a_i, b_i),$$

where we denote $v_0 = (a_0, b_0)$, $v_i = (a_i, b_i)$ and $v = (v_0, v_i)$.

Let $A_q = \{A_{i,j}\}_{j \in q}$, $D_q = \{D_{i,j}\}_{j \in q}$ be scRNA reads from cells in hypercluster q . Since we assume that scRNA reads at different positions or from different cells are conditionally independent, then the total likelihood of these reads is:

$$\mathbb{P}(A_q|D_q, I_q, \mathbf{C}, \mathbf{T}, \theta) = \prod_{j \in q} \prod_i \mathbb{P}(A_{i,j}|D_{i,j}, I_q, C_{i,I_q}, T_j = q, \theta),$$

Gibbs sampling

Inference in CaClust is performed using a Gibbs sampler, where each variable is iteratively sampled from its conditional probability given the current values of the other variables in the model. Since CaClust is a probabilistic graphical model, this conditional probability is equivalent to the conditional probability of the variable given its Markov Blanket [42]. The sampling of variables related to the CRP is performed using a dedicated procedure described below. The inference is carried out until convergence, as measured by the Gelman-Rubin diagnostic; afterwards, the samples from iterations after burn-in approximate the true posterior distribution of the variables.

Conditional probabilities of variables

In the Gibbs sampler, we sample the variables from their conditional probabilities given their Markov Blankets (MB). Using Bayes' rule we factor these probabilities as follows.

For the error rate ξ , we have:

$$\mathbb{P}(\xi|MB(\xi)) \propto \mathbb{P}(\xi|\kappa) \cdot \mathbb{P}(\mathbf{C}|\xi, \Omega).$$

The prior on ξ is a Beta distribution with parameters (κ_0, κ_1) and the likelihood $\mathbb{P}(\mathbf{C}|\xi, \Omega)$ is a product of Binomial distribution functions over variables $C_{i,k}$ for $i \in 1, \dots, N$ and $k \in 1, \dots, K$, where a success is defined as a disagreement (since ξ is an error rate) between $\Omega_{i,k}$ and $C_{i,k}$. Therefore, from the Beta-Binomial conjugacy, we obtain:

$$\mathbb{P}(\xi|MB(\xi)) = \text{Beta} \left(\xi; \kappa_0 + \sum_{i,k} \mathbf{1}(\Omega_{i,k} \neq C_{i,k}), \kappa_1 + \sum_{i,k} \mathbf{1}(\Omega_{i,k} = C_{i,k}) \right).$$

For the true genotypes $C_{i,k}$, we have:

$$\mathbb{P}(C_{i,k}|MB(C_{i,k})) \propto \mathbb{P}(C_{i,k}|\Omega_{i,k}, \xi) \cdot \mathbb{P}(\mathbf{A}|\mathbf{D}, \mathbf{I}, C_{i,k}, \mathbf{T}, \theta).$$

Since we assume reads at different variant positions i and reads in different cells j are conditionally independent, the above probability factorises as:

$$\mathbb{P}(C_{i,k}|MB(C_{i,k})) \propto \mathbb{P}(C_{i,k}|\Omega_{i,k}, \xi) \cdot \prod_{q, I_q=k} \prod_{j \in q} \left\{ \text{Binom}(A_{i,j}|D_{i,j}, \theta_1)^{C_{i,k}} \cdot \text{Binom}(A_{i,j}|D_{i,j}, \theta_0)^{1-C_{i,k}} \right\}.$$

For the hypercluster-clone assignment variable I_q , we have:

$$\mathbb{P}(I_q = k|MB(I_q)) \propto \mathbb{P}(I_q = k) \cdot \mathbb{P}(A_{i,j}|D_{i,j}, I_q = k, C_{i,k}, T_j, \theta).$$

We use an uninformative prior on I_q , so the posterior probabilities of I_q are proportional to the likelihoods of scRNA reads, which again we assume to be conditionally independent across positions i and cells j :

$$\mathbb{P}(I_q = k | MB(I_q)) \propto \frac{1}{K} \cdot \prod_{j \in q} \prod_i \left\{ \text{Binom}(A_{i,j} | D_{i,j}, \theta_1)^{C_{i,k}} \times \text{Binom}(A_{i,j} | D_{i,j}, \theta_0)^{1-C_{i,k}} \right\}.$$

For the hyperclustering variable T_j , we have:

$$\mathbb{P}(T_j = q | MB(T_j)) \propto \mathbb{P}(T_j = q | \alpha_0) \cdot \prod_i \mathbb{P}(A_{i,j} | D_{i,j}, I_q, C_{i,k}, T_j = q, \theta) \cdot \prod_l \mathbb{P}(X_{l,j} | B_{l,q}, T_j = q).$$

The observations of BCR sequences across positions l and cells j are also conditionally independent, so the above factorises to:

$$\begin{aligned} \mathbb{P}(T_j = q | MB(T_j)) &\propto \text{CRP}(T_j = q | \alpha_0) \cdot \\ &\prod_i \left\{ \text{Binom}(A_{i,j} | D_{i,j}, \theta_1)^{C_{i,k}} \times \text{Binom}(A_{i,j} | D_{i,j}, \theta_0)^{1-C_{i,k}} \right\} \cdot \\ &\prod_l \text{Categorical}(X_{l,j} | B_{l,q}, T_j = q). \end{aligned}$$

We can calculate the above for all hyperclusters q that are non-empty. However, since we are dealing with a CRP, then during the sampling of cell j 's hypercluster assignment T_j we need to include the possibility of joining a new hypercluster. So during the sampling of T_j , if Q is the number of non-empty hyperclusters in clustering T_{-j} , we add a hypercluster $Q + 1$ with parameters I_{Q+1}, B_{Q+1} sampled from their prior distributions. Then we can calculate the above for hyperclusters $1, \dots, Q + 1$ and sample T_j with appropriate probabilities.

For the hypercluster BCR frequency profiles $B_{l,q}$ we have:

$$\mathbb{P}(B_{l,q} | MB(B_{l,q})) \propto \mathbb{P}(B_{l,q}) \cdot \mathbb{P}(X_{l,j} | B_{l,q}, T_j = q).$$

Since the $B_{l,q}$ variable has a Dirichlet prior and the observations of nucleotides $X_{l,j}$ come from a categorical distribution with probabilities $B_{l,q}$, then from the Dirichlet-Multinomial conjugacy, we get:

$$\mathbb{P}(B_{l,q} | MB(B_{l,q})) = \text{Dirichlet}[B_{l,q}; g_{A,l} + n_{A,l,q}, g_{C,l} + n_{C,l,q}, g_{G,l} + n_{G,l,q}, g_{T,l} + n_{T,l,q}],$$

where $n_{Nuc,l,q}$ is the number of occurrences of nucleotide Nuc at position l in BCR sequences of cells belonging to hypercluster q .

For the variant read observation probabilities θ , we have:

$$\mathbb{P}(\theta | MB(\theta)) \propto \mathbb{P}(\theta | \nu) \cdot \mathbb{P}(\mathbf{A} | \mathbf{D}, \mathbf{C}, \mathbf{I}, \mathbf{T}, \theta)$$

Since θ_0 and θ_i have a Beta prior and the likelihoods of \mathbf{A} are Binomial distributions with sizes \mathbf{D} , then from the Beta-Binomial conjugacy, we get:

$$\mathbb{P}(\theta | MB(\theta)) = \text{Beta}[\theta_0; \nu'_0] \times \prod_i \text{Beta}[\theta_i; \nu'_i],$$

where $\nu'_0 = (a'_0, b'_0)$ and $\nu'_i = (a'_i, b'_i)$ are defined by:

$$\begin{aligned}
 a'_0 &= a_0 + \sum_q \sum_{ij \in q} (1 - C_{i,I_q}) \cdot A_{ij} \\
 b'_0 &= b_0 + \sum_q \sum_{ij \in q} (1 - C_{i,I_q}) \cdot (D_{ij} - A_{ij}) \\
 a'_i &= a_i + \sum_q \sum_{j \in q} C_{i,I_q} \cdot A_{ij} \\
 b'_i &= b_i + \sum_q \sum_{j \in q} C_{i,I_q} \cdot (D_{ij} - A_{ij}),
 \end{aligned}$$

The conditional probability of α_0 is a special case and its sampling is described in the following section.

Updating the concentration parameter α_0

During model inference we also sample α_0 , the concentration parameter of the Chinese Restaurant Process behind our clustering. For sampling α_0 we use a method described in paper [43]. It applies to mixture models in general, and in this section we explain how it is implemented in the CaClust model.

Firstly, we assume a Gamma prior over α_0 ,

$$\alpha_0 \sim \text{Gamma}(a, b).$$

We have M as the number of cells, and let Q be the number of non-empty hyperclusters in current sampling iteration. Since Q is not fixed, we consider Q as another random variable. In the Chinese Restaurant Process (CRP), we have:

$$\begin{aligned}
 \mathbb{P}(\alpha_0 | \mathbf{T}) &= \mathbb{P}(\alpha_0 | Q, M) \\
 &\propto \mathbb{P}(\alpha_0) \cdot \mathbb{P}(Q | \alpha_0, M).
 \end{aligned}$$

Then from the probability density of CRP, we have:

$$\mathbb{P}(Q | \alpha_0, M) = |s(Q, M)| \cdot (\alpha_0)^Q \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + M)},$$

where $s(Q, M)$ are Stirling numbers of the first kind and, more importantly, they are independent from α_0 . For $\alpha_0 > 0$ (as in our case), we can rewrite the gamma functions from above as:

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + M)} = \frac{(\alpha_0 + M) \cdot \beta(\alpha_0 + 1, M)}{\alpha_0 \cdot \Gamma(M)},$$

where β is the beta function. So, since $\Gamma(M)$ is constant w.r.t. α_0 , the conditional probability of α_0 takes the form:

$$\begin{aligned}
 \mathbb{P}(\alpha_0 | Q, M) &\propto \mathbb{P}(\alpha_0) \cdot (\alpha_0)^{Q-1} \cdot (\alpha_0 + M) \cdot \beta(\alpha_0 + 1, M) \\
 &\propto \mathbb{P}(\alpha_0) \cdot (\alpha_0)^{Q-1} \cdot (\alpha_0 + M) \cdot \int_0^1 x^{\alpha_0} (1-x)^{M-1} dx.
 \end{aligned}$$

This shows that $\mathbb{P}(\alpha_0 | Q, M)$ is a marginal probability distribution of a joint distribution of pairs (α_0, σ) , where:

$$\mathbb{P}(\alpha_0, \sigma | Q, M) \propto \mathbb{P}(\alpha_0) \cdot (\alpha_0)^{Q-1} \cdot (\alpha_0 + M) \cdot (\sigma)^{\alpha_0} (1 - \sigma)^{M-1} \quad \sigma \in (0, 1). \quad (1)$$

With the *Gamma*(a, b) prior on α_0 , we obtain the following probabilities:

$$\begin{aligned} \mathbb{P}(\alpha_0 | \sigma, Q, M) &\propto \frac{b^a}{\Gamma(a)} \alpha_0^{a-1} e^{-b\alpha_0} \cdot \alpha_0^{Q-1} (\alpha_0 + M) e^{\ln \sigma \alpha_0} \\ &\propto \alpha_0^{a+Q-1} e^{-(b-\ln \sigma)\alpha_0} + M \alpha_0^{a+Q-2} e^{-(b-\ln \sigma)\alpha_0} \\ &\propto \frac{\Gamma(a+Q)}{(b-\ln \sigma)^{a+Q}} \text{Gamma}(\alpha_0; a+Q, b-\ln \sigma) \\ &\quad + M \frac{\Gamma(a+Q-1)}{(b-\ln \sigma)^{a+Q-1}} \text{Gamma}(\alpha_0; a+Q-1, b-\ln \sigma), \end{aligned}$$

which is a mixture of two gamma densities:

$$\begin{aligned} \mathbb{P}(\alpha_0 | \sigma, Q, M) &\propto \pi_\sigma \text{Gamma}(\alpha_0; Q+a, b-\ln \sigma) \\ &\quad + (1-\pi_\sigma) \text{Gamma}(\alpha_0; Q+a-1, b-\ln \sigma), \end{aligned}$$

with weights such that $\frac{\pi_\sigma}{1-\pi_\sigma} = \frac{a+Q}{M(b-\ln \sigma)}$.

Lastly, by marginalising Eq. 1 w.r.t. α_0 , we get the conditional distribution of σ :

$$\mathbb{P}(\sigma | \alpha_0, Q, M) \propto \sigma^{\alpha_0} (1 - \sigma)^{M-1} \propto \text{Beta}(\sigma; \alpha_0 + 1, M).$$

Therefore, we sample α_0 at each iteration as follows:

1. Pick a new value for σ with previous values of Q and α_0
2. Pick a new value for α_0 with previous value of Q and new σ .

Data simulation methods

In this section, we describe the process of simulating data for evaluation of model performance, specifying the parameter settings for different simulation runs and evaluation metrics.

Simulation process

The input data required by the model for inference are: matrix \mathbf{D} of read counts over mutations in the cells, matrix \mathbf{A} of alternate read counts over mutations in the cells, BCR sequences of cells, and matrix $\mathbf{\Omega}$ of observed clone genotypes.

Assume we are given: the number of clones K , the number of cells M , the number of variant positions N , the length of BCR sequences L . Then the simulation steps are as follows:

- Step 1: Hyperclustering simulation.
 1. Generate the concentration parameter α_0 from its prior distribution.
 2. Using the CRP with concentration parameter α_0 cluster the M cells.
 3. Assign each of the Q resulting hyperclusters to one of the K clones with uniform probability.

- Step 2: BCR sequence (\mathbf{X}) simulation.
 1. For each hypercluster q and each BCR position l , sample its BCR frequency profile $B_{l,q}$ from the Dirichlet distribution with prior parameters g_l .
 2. For each cell j simulate its BCR sequence at position l from a categorical distribution with the frequencies $B_{l,q}$ that we obtain for its hypercluster q from the previous step. In simulation scenarios with centroid behaviour (see simulation types), we additionally randomly select hyperclusters with rate r_{clust} and within them r_{cell} cells to express identical BCR sequences, equal to the sequence that is most probable given their hypercluster's profile $B_{l,q}$.

- Step 3: $\mathbf{\Omega}$ and \mathbf{C} simulation.
 1. Simulate θ_0, θ_i, ξ from their prior distributions with desired parameters.
 2. Pick a variant rate ν .
 3. For each clone simulate its mutation profile C_k : at each variant position i we have a chance of ν that clone k exhibits a variant at that position, i.e., $C_{i,k} = 1$.
 4. Simulate the observed matrix $\mathbf{\Omega}$ by randomising \mathbf{C} with error rate ξ .

- Step 4: scRNA-seq variant data (\mathbf{A} and \mathbf{D}) simulation:
 1. For each cell j and variant position i sample the total number of observed reads mapping to that position ($D_{i,j}$) from a Poisson distribution with mean μ_D .
 2. For each cell j and each position i sample the number of observed variant reads $A_{i,j}$ from its conditional probability distribution.

Choosing simulation parameters

In the simulation process, we have control over the values of several parameters, which influence the simulated structure in different ways. We can divide those parameters into three groups: data dimensions, simulation type agnostic, and simulation type specific.

Data dimension parameters

Data dimension parameters are fixed for all simulation types and affect primarily the computational times.

We set the number of tumour clones $K = 3$, the number of variant positions $N = 100$ in the tumour genotypes (fixed according to the numbers of variant positions after filtering in the analysed real patient data), the number of cells $M = 1000$. Finally, we set the number of mutated BCR positions $L = 300$, since in the experimental data we observed between 200 and 400 positions that contain at least one alternate nucleotide. Note that this number is smaller than the combined length of the BCR heavy and light chain sequences, which is around 600–700 nucleotides.

Simulation scenario-agnostic parameters

These parameters are shared between all simulation scenarios. Firstly, the variant rate ν used to generate the clone genotypes; for each clone k and each variant position i we set

a probability $\nu = 0.3$ that $C_{i,k} = 1$. Secondly, the parameters $\nu = (\nu_0, \nu_1)$ of error rate ξ 's beta prior distribution; we set $\nu_0 = 1$, $\nu_1 = 19$.

Lastly, the prior parameters of alternate nucleotide observation probabilities θ_0 and θ_i . In the model we use only somatic variants, so θ_0 reflects the probability that an alternate nucleotide from a different position was mismapped. We assume high quality of position analysis, thus θ_0 should be low. Secondly, we assume we are dealing with heterozygotic somatic variants; therefore, the probability of observing a variant nucleotide from the mutated allele is on average 50%, but can vary due to the random nature of mRNA expression and sequencing. To reflect the above, we choose the prior parameters of θ_0 to be $a_0 = 0.2$, $b_0 = 99.8$, and the prior parameters of θ_i to be $a_i = 4.5$, $b_i = 5.5$.

Simulation scenarios

We wanted to test model performance on datasets with three varying characteristics: scRNA read depth, controlled with μ_D ; number of BCR hyperclusters, controlled with α_0 ; and intracluster variance of BCR sequences, controlled with s_g, r_{clust}, r_{cell} . To do this, we created a base scenario modelling medium read depth ($\mu_D = 0.01$), low number of BCR hyperclusters ($\alpha_0 = 5$), and medium intracluster BCR variance ($s_g = 0.01, r_{clust} = 0, r_{cell} = 0$). Then, in each simulation type we changed one of these parameters, while keeping the rest at base values.

Apart from the base scenario we created eight following simulation scenarios. Two scRNA read depth scenarios with high, or low average numbers of reads per variant in a cell; these scenarios reflect a high and low quality sequencing experiments with high and low scRNA information respectively. One sparse hyperclustering scenario with numerous BCR hyperclusters in the data; this models a tissue with multiple BCR clusters evolving in parallel. A scenario with high variance BCR sequences within hyperclusters; this models cells with highly mutated BCR sequences that give low information. And lastly, four scenarios with $x\%$ of cells within $y\%$ of hyperclusters sharing their hyperclusters' prevalent BCR sequence (with $x, y \in \{20, 80\}$); this accounts for BCR hypercluster tendencies observed in real data, in which most hyperclusters contain a large subset of cells with identical BCR sequences. All simulation types are shown in Additional File 2: Table S4 along with the values of the varied parameters.

Performance metrics

To measure the performance of CaClust and CACTUS on simulated datasets we define three performance metrics: cell to clone assignment accuracy, genotype reconstruction accuracy, and hyperclustering reconstruction agreement.

We calculate the cell to clone assignment accuracy as the fraction of cells that were assigned to their correct clone in the MLE assignment after model inference.

Secondly, the genotype reconstruction accuracy measures how well the model corrects the input matrix of clonal profiles Ω , which contains errors with rate ξ . To calculate it, we take the MLE of the true clone genotypes \mathbf{C} from the model and compute the fraction of entries that it agrees on with the hidden clonal profiles from the simulation.

Finally, with the hyperclustering reconstruction agreement we measure the similarity between the hyperclustering from the data generation process and the hyperclustering reconstructed by the model. For this, we calculate the adjusted Rand index between the

final hyperclustering of cells in the model and the hyperclusters from the data generation process.

Model application process

Inference parameters

For model inference, we need to choose the following prior parameters.

Firstly, the parameters of the beta priors of θ_0, θ_i , which are $v_0 = (a_0, b_0), v_i = (a_i, b_i)$. Since θ_0 should be the small probability of an observed read being variant in a clone not containing that mutation in the genotype, we set the prior parameters a_0, b_0 such that $\frac{a_0}{a_0+b_0} = 0.002$ and that their magnitude $a_0 + b_0$ is equal to the number of total reference reads in the sample. θ_i is the probability of an observed read at position of SNV i being variant in a clone containing SNV i . To account for bursty expression and harder mapping of variant fragments, we set a_i, b_i such that $\frac{a_i}{a_i+b_i} = 0.45$ and their magnitude $a_i + b_i$ to be equal the total number of variant reads of SNV i in the cells.

Secondly, the prior parameters of the BCR frequency profiles. At each BCR position we use a low strength uninformative prior of $g = (0.01, 0.01, 0.01, 0.01)$, which influences more data-driven frequency profiles.

For the beta prior on the input genotype error rate ξ , we use parameters κ_0, κ_1 with values such that $\frac{\kappa_0}{\kappa_0+\kappa_1} = 0.8$ and their magnitude $\kappa_0 + \kappa_1 = N \cdot K$, where N is the number of SNVs and K is the number of clones used for the sample. In that way, the genotype at each clonal position is a priori weighed $\kappa_0 : \kappa_1$ in favour of the input genotype call; during the inference that ratio changes with the numbers of agreements and disagreements between Ω and \mathbf{C} in the model, which are on the level of $N \cdot K$.

For the gamma prior of α_0 , we use a non-informative prior with parameters (1, 1).

Model initialisation

We initialise the hyperclustering in the model (T variable) with the clustering of cells with identical BCR sequences. This is to promote faster convergence and is in line with the assumption, that hyperclusters of cells with identical BCR sequences should come from the same evolutionary clone. The model can also be initialised with random hyperclusters or hyperclusters containing singular cells.

After hypercluster initialisation we obtain the initial assignment of hyperclusters to clones by performing initial Gibbs sampling iterations only for $\mathbf{I}, \mathbf{C}, \xi, \theta_0, \theta_i$ variables, while keeping the initial BCR hyperclustering constant. This is done to ensure that before relaxing the BCR hyperclustering in full model iterations we resolve any major disagreement between the input clonal profiles and the scRNA variants observed in the initial BCR hyperclustering.

Afterwards, the α_0 and \mathbf{B} variables are initialised from their conditional probabilities and the full model sampling iterations are ready to be performed.

Convergence assessment

The model is run for a set number of initial and full sampling iterations; then, for assessing the model convergence we use the Gelman-Rubin diagnostic with stable variance estimators [44] on the variables with continuous values in the model ($\theta_0, \theta_i, \alpha_0$). The model is assumed to have converged if between the chains in a sample we observe the

multivariate potential scale reduction factor (MPSRF) lower than 1.001. If not converged, the next set of full sampling iterations is carried out and the convergence is reassessed. We use 5000 initial and 500 full sampling iterations.

Result extraction

For each sample, we choose the model chain with the highest likelihood; then, from that chain we take the maximum a posteriori (MAP) of cell-clone assignment and clone genotypes, and the output hyperclusters are taken to be the hyperclustering with maximum likelihood through the iterations.

Independent gene expression clustering

Gene expression clusters were obtained by PCA dimensionality reduction to 30 dimensions and next Leiden clustering on the reduced profiles with resolution parameter of 0.3 m using Seurat package [45].

Pathogenicity prediction

All variants were annotated within the Geneticist Assistant NGS Interpretive Workbench (SoftGenetics) by public variant-databases (dbSNP, ClinVar, and COSMIC) and available literature, into class 1 (not pathogenic), class 2 (potentially not pathogenic), class 3 (unknown significance), class 4 (potentially pathogenic), or class 5 (pathogenic). [37–39]

Differential gene expression analysis

We use the DESeq2 package [46] for differential gene expression analysis in the samples. For each clone, we use its cells as a test group and the rest of the cells as a reference group. The analysis is carried out on the SCT counts, with size factors estimated using the scran package [47]. For significance we use the likelihood-ratio test (LRT) implemented in the package, with Benjamini-Hochberg correction for multiple testing.

From the DE analysis, we exclude genes with a total read count < 1% of the cell count in the data; the ribosomal protein L and S genes; the immunoglobulin IG[HKL] genes; and the mitochondrial MT- genes.

Gene set enrichment analysis

We perform Gene set enrichment analysis (GSEA) as described in [48] on the 50 hallmark genesets [49]. For each clone we use as input the list of genes ranked by their fold-change found in the DE analysis. For FDR in the GSEA we use gene label reshuffling with 10,000 permutations.

Cell cycle analysis

Cell cycle phase was assigned to cells using the CellCycleScoring function from the Seurat package. To analyse the cell cycle distribution differences in sample K7B, we used Pearson's chi-squared test. The minimal expected value in the contingency tables was 10.1 for the differences between cycles of C2 and C3, which is in line with best practices for the test, where the minimal expected value cannot be lower than 5 [50].

Targeted resequencing

Variant selection for targeted resequencing

Variants with subclonal VAF ($0.05 < \text{VAF} < 0.4$ or $0.6 < \text{VAF} < 0.8$) in genes that were detectable in $\geq 2.5\%$ of single cells based on single cell transcriptome data were selected. Variants that were synonymous, located outside exons, in immunoglobulin genes or more than 2000 bp from start of transcripts were excluded.

Targeted resequencing procedure

Pools of 10X Genomics cDNA that were remaining after GEX and BCR sequencing were used as template for targeted resequencing. A semi-nested 2-step amplification strategy was designed based on a reverse outer and a reverse nested inner primer both located 3' of the target variant position (see Additional File 4: Primer information). For primer validation, an additional forward primer in the 5' region of the gene was designed and used on cDNA that was generated from bulk-sorted FL cells as described previously with modifications: initiation of reverse transcription using oligo-dT and 5' extension with an alternative 5' template switching oligo [51]. Validation PCRs were performed with 10 μL bulk FL derived cDNA template using Phusion Flash High Fidelity PCR Master Mix (Thermo Fisher Scientific, Waltham, MA) with the reverse outer gene specific primers and enrichment primer I both at 1 μM . As PCR program was used: melting 45 s 98 °C, amplification for 20 cycles: melting 20 s 98 °C, annealing 30 s 67 °C, elongation 120 s 72 °C, followed by a final elongation step of 120 s 67 °C. Aliquots of 10 μL PCR product were run on 1% agarose gel and visualised. If bands were visible, remaining 40 μL PCR products were purified using AMPure XP Beads (BeckmanCoulter, Indianapolis, IN). The nested PCR was performed under identical conditions with the inner gene specific primer and enrichment primer II. If no bands were visible, alternative primers were designed and tested. Using validated primer sets, aliquots of 1.5 μL of 10X Genomics single cell cDNA pools were amplified with enrichment primers I and II and the validated outer and inner 3' reverse primers under identical conditions as in primer validation. After the second amplification, PCR products were run on preparative 1% agarose gel, visible bands were excised, purified using Promega Wizard PCR Preps DNA Purification System (Thermo Fisher Scientific) and run on Bioanalyzer (Agilent) for accurate quantification of obtained PCR products. Equimolar amplicon pools were generated and single molecule sequencing was performed on PacBio Sequel II platform (PacBio, San Diego, CA).

PacBio full length sequence data processing

Circular consensus sequence fastq files were used as input for filtering and genotype calling. PacBio polymerizes circularised single strand DNA templates and dependent on the start of the reaction results in either the forward or reverse sequence. To obtain all reads in the forward direction, a copy of every read was reverse complemented and added to the data. Using the 5' end PCR adapter sequence AATGATAcg, also allowing deletion of 1–3 nt thus accepting aATGATAcg, aaTGATACg, and aatGATACG, were used to keep only reads in the forward direction. For reads that were not correctly split during circular consensus generation and thus consisted of the forward linked to the reverse sequence, mean quality score of nt 5–100 of each sequence end was calculated.

The read part with lower quality was clipped. Next, alignment scores were obtained for read nucleotides (nt) 47 to 60 with the distal PCR adapter motif CTTCCGATCT, and read nt 82 to 100 with TSO+G motif TTTCTTATATG. The space between adapter and TSO comprises the single cell barcode (scbc) of 16 nt and the unique molecular identifier (umi) of 10 nt, and thus should be 26 nt. Reads with adapter motif alignment score ≥ 9 and, TSO+G alignment score ≥ 10 and a scbc-umi space of 24 to 28 nt were accepted. The latter filter taking into account potential insertions and deletions within polyhomologous stretches. In the second step, reads were aligned with wildtype and mutant reference sequences of 41 nt with 20 up- and downstream nt flanking the nt substitution or insertion/deletion. Reads that aligned with a single reference with alignment score ≥ 30 of max 41 and nt quality score ≥ 120 of max 126 at the variant position were accepted. In the third step, reads were assigned to cells using as reference scbc from valid cells from Cell Ranger default output for gene expression profiling and BCR VDJ/VJ sequencing. Nt from end of adapter - 1 to + 18, thus a stretch of 19 nt was aligned with all reference scbc. Matches with a single scbc and alignment score ≥ 14 and an average quality score of ≥ 110 of max 126 were accepted. Two potential umis were extracted, one of exactly 10 nt downstream of the scbc (umi10), and the second one between the end of the scbc until the last nt before the TSO start (umiTso). No reference for umis is available, and we expect due to amplification and sequencing errors an overestimation of the repertoire of umis. We therefore collapsed highly similar umis as follows: Per scbc and gene, identical umis were counted and arranged by decreasing count. With the most dominant umi as reference, for all other umis a pairwise alignment score was calculated. A discrepancy of max 3 nt was allowed for a umi to be collapsed into the more dominant umi. Independently for umi10 and umiTso, the process was iterated over all umis. If the resulting umi of both umi10 and umiTso was identical, collapsed umis were corrected towards the dominant umi. If the resulting umi of umi10 and umiTso were not identical but their Levenshtein distance was ≤ 3 , umi10 was chosen as the final umi. At Levenshtein distance >3 , the read was rejected. Genotyping of umis and scbc was performed as follows: In case more than 1 umi was detected per single cell and gene, we counted the number of duplicates and evaluated if all umis had the same genotype. If all reads had identical genotypes, the umi genotype was called accordingly. As a result of amplification errors however, chimeric PCR products can be formed resulting in mixed wildtype and mutant reads within 1 umi. Umi genotypes were called based if mutant or wildtype read counts were at least 0.67 of the total umi read count. Single cells were called mutant if at least 1 mutant umi was detected. Wildtype was called only if no mutant umis and at least 4 wildtype umis were detected in the absence of mutant umis.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03417-1>.

Additional file 1: Supplementary Figures S1-S14.

Additional file 2: Supplementary tables S1-S4 and supplementary notes on targeted resequencing and a correction to account for random monoallelic expression [58–61].

Additional file 3: Output profiles for all samples.

Additional file 4: Information on primers used.

Additional file 5: Review history.

Acknowledgements

We thank Leiden University Medical Center Flow Core Facility and Leiden Genome Technology Center for excellent purification of lymphoma cells and cell preparation and sequencing. We wish to express our deepest gratitude to all patients who allow us to use samples that are essential for our research.

Review history

The review history is available as Additional file 5.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

K.O.O., S.S., C.A.M.v.B., and E.S. conceived the project and methodology. C.A.M.v.B and S.M.K. curated the data. K.O.O. implemented the model, performed the computational analysis and prepared the figures. E.Q., C.A.M.v.B, H.W.v.K, R.A.L.d.G., J.S.P.V., J.H.S.Y., and M.A.N. performed wet lab experiments. E.S. supervised the study. K.O.O, C.A.M.v.B, and E.S. wrote the manuscript with feedback from H.V.

Authors' X handles

X handles: @ShadiShafighi (Shadi Shafighi); @marcelnavarrete (Marcelo A. Navarrete); @ewa_szczurek (Ewa Szczurek).

Funding

This work received funding from the Polish National Science Centre SONATA BIS grant No. 2020/38/E/NZ2/00305 and KWF Dutch Cancer Society grant No. 13104. M.N. and J.S. are funded by Anillo ATE220016 and Fondecyt 1230298(ANID, Chile).

CaClust model code and example input data available at [57] under the GPL3 licence.

Data availability

Single cell gene expression and B-cell receptor transcriptome data is available in NCBI GEO repository identifiers GSE252344 (K4B [52]), GSE252416 (K5B [53]), GSE252642 (K6B [54]), and GSE252687 (K7B [55]). Whole exome sequencing data is available in NCBI Sequence Research Archive BioProject id PRJNA1062119 [56].

Declarations**Ethics approval and consent to participate**

Samples with histologically confirmed infiltration of follicular lymphoma (FL) grade 1–2 were collected according to the Declaration of Helsinki and under authorization of applicable biobank regulations of Leiden University Medical Center and the Ethical Committee of Leiden University Medical Center (reference HEM 008/SH/sh). All patients gave written, informed consent for participation and publication.

Consent for publication

All patients gave written, informed consent for participation and publication.

Competing interests

Projects at Szczurek lab are co-founded by Merck Healthcare. The other authors declare no competing interests. Ewa Szczurek is an Editorial Board Member for *Genome Biology* but had no input into the assessment or peer review of this manuscript.

Received: 17 April 2024 Accepted: 8 October 2024

Published online: 05 November 2024

References

1. Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. *Nat Rev Genet.* 2019. <https://doi.org/10.1038/s41576-019-0114-6>.
2. Nowell PC. The Clonal Evolution of Tumor Cell Populations. *Science.* 1976. <https://doi.org/10.1126/science.959840>.
3. Yap TA, Gerlinger M, Futreal AP, Pusztai L, Swanton C. Intratumor Heterogeneity: Seeing the Wood for the Trees. *Sci Transl Med.* 2012. <https://doi.org/10.1126/scitranslmed.3003854>.
4. Lenz G, Onzi GR, Lenz LS, Buss JH, dos Santos JA, Begnini KR. The Origins of Phenotypic Heterogeneity in Cancer. *Cancer Res.* 2022;82:3–11. <https://doi.org/10.1158/0008-5472.CAN-21-1940>.
5. Sharma A, Merritt E, Hu X, Cruz A, Jiang C, Sarkodie H, et al. Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung Tumors. *Cell Rep.* 2019;29(8):2164–2174.e5. <https://doi.org/10.1016/j.celrep.2019.10.045>.
6. da Silva-Diz V, Lorenzo-Sanz L, Bernat-Peguera A, Lopez-Cerda M, Muñoz P. Cancer cell plasticity: Impact on tumor progression and therapy response. *Semin Cancer Biol.* 2018;53:48–58. <https://doi.org/10.1016/j.semcancer.2018.08.009>.
7. Qin S, Jiang J, Lu Y, Nice EC, Huang C, Zhang J, et al. Emerging role of tumor cell plasticity in modifying therapeutic response. *Signal Transduct Target Therapy.* 2020;5. <https://doi.org/10.1038/s41392-020-00313-5>.
8. Gavish A, Tyler M, Greenwald AC, Hoefflin R, Simkin D, Tschernichovsky R, et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature.* 2023;618:598–606. <https://doi.org/10.1038/s41586-023-06130-4>.

9. Fittall MW, Van Loo P. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Med.* 2019. <https://doi.org/10.1186/s13073-019-0632-z>.
10. Ding L, Ley T, Larson Dea. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature.* 2012. <https://doi.org/10.1038/nature10738>.
11. McGranahan N, Swanton C. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell.* 2015. <https://doi.org/10.1016/j.ccell.2014.12.001>.
12. Kridel R, Sehn LH, Gascoyne RD. Pathogenesis of follicular lymphoma. *J Clin Investig.* 2012. <https://doi.org/10.1172/JCI63186>.
13. Pasqualucci L. Molecular pathogenesis of germinal center-derived B cell lymphomas. *Immunol Rev.* 2019. <https://doi.org/10.1111/imr.12745>.
14. Okosun J, Bödör C, Wang J, Araf S, Yang C, Pan C, et al. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat Genet.* 2014;46:176–81. <https://doi.org/10.1038/ng.2856>.
15. Shafiqhi SD, Kielbasa S, Sepúlveda-Yáñez J, Monajemi R, Cats D, Mei H, et al. CACTUS: integrating clonal architecture with genomic clustering and transcriptome profiling of single tumor cells. *Genome Med.* 2021;13:891–921. <https://doi.org/10.1016/j.jocms.2021.101523>.
16. Roider T, Seufert J, Uvarovskii A, Frauhammer F, Bords M, Abedpour N, et al. Dissecting intratumour heterogeneity of nodal B-cell lymphomas at the transcriptional, genetic and drug-response levels. *Nat Cell Biol.* 2020;22:896–906. <https://doi.org/10.1038/s41556-020-0532-x>.
17. van Bergen CAM, Kloet SL, Quinten E, Sepúlveda Yáñez JH, Menafra R, Griffioen M, et al. Acquisition of a glycosylated B-cell receptor drives follicular lymphoma toward a dark zone phenotype. *Blood Adv.* 2023. <https://doi.org/10.1182/bloodadvances.2023010725>.
18. Olsen TR, Talla P, Furnari J, Bruce JN, Canoll P, Zha S, et al. Scalable co-sequencing of RNA and DNA from individual nuclei. *bioRxiv.* 2023. <https://doi.org/10.1101/2023.02.09.527940>.
19. Yu L, Wang X, Mu Q, Tam SST, Loi DSC, Chan AKY, et al. scONE-seq: A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors. *Sci Adv.* 2023;9. <https://doi.org/10.1126/sciadv.abp8901>.
20. Han K, Kim K, Joung J, Son D, Kim Y, Jo A, et al. SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res.* 2018;28:75–87. <https://doi.org/10.1101/gr.223263.117>.
21. Blackinton J, Morozova T, Zawistowski J, Salas-Gonzalez I, Arvapalli D, Velivela S, et al. The ResolveOME Platform for Comprehensive Analysis of Multi-omic Layers of Single-cell Biology. <https://www.bioskryb.com/resolveome-comprehensive-multi-omic-single-cell-analysis/>. Accessed 13 Aug 2024.
22. McCarthy DJ, Rostom R, Huang Y, Kunz DJ, Danecek P, Bonder MJ, et al. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nat Methods.* 2020. <https://doi.org/10.1038/s41592-020-0766-3>.
23. Jun SH, Toosi H, Mold J, Engblom C, Chen X, O'Flanagan C, et al. Reconstructing clonal tree for phylo-phenotypic characterization of cancer using single-cell transcriptomics. *Nat Commun.* 2023. <https://doi.org/10.1038/s41467-023-36202-y>.
24. Shafiqhi SD, Geras A, Jurzysta B, Naeini AS, Filipiuk I, Rączkowski Ł, et al. Tumorscope: a probabilistic model for mapping cancer clones in tumor tissues. *bioRxiv.* 2022. <https://doi.org/10.1101/2022.09.22.508914>.
25. Erickson A, He M, Berglund E, Marklund M, Mirzazadeh R, Schultz N, et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature.* 2022;608:360–7. <https://doi.org/10.1038/s41586-022-05023-2>.
26. Fan J, Lee HO, Lee S, Eun Ryu D, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* 2018;28:1217–27. <https://doi.org/10.1101/gr.228080.117>.
27. 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. <https://doi.org/10.1038/s41586-019-1367-0>.
28. de Groen R, Schrader A, Kersten M, Pals S, Vermaat J. MYD88 in the driver's seat of B-cell lymphomagenesis: from molecular mechanisms to clinical implications. *Haematologica.* 2019;2337–48. <https://doi.org/10.3324/haematol.2019.227272>.
29. Pasqualucci L, Dominguez-Sola D, Chiarenza A, Fabbri G, Grunn A, Trifonov V, et al. Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature.* 2011;471:189–95. <https://doi.org/10.1038/nature09730>.
30. Navarro-Hernandez IC, López-Ortega O, Acevedo-Ochoa E, Cervantes-Díaz R, Romero-Ramírez S, Sosa-Hernández VA, et al. Tetraspanin 33 (TSPAN33) regulates endocytosis and migration of human B lymphocytes by affecting the tension of the plasma membrane. *FEBS J.* 2020;287:3449–71. <https://doi.org/10.1111/febs.15216>.
31. Wang F, Yang Y, Boudagh G, Eskelinen E, Klionsky D, Malek S. Follicular lymphoma-associated mutations in the V-ATPase chaperone VMA21 activate autophagy creating a targetable dependency. *Autophagy.* 2022;18:1982–2000. <https://doi.org/10.1080/15548627.2022.2050663>.
32. Haebe S, Shree T, Sathe A, Day G, Czerwinski D, Grimes S, et al. Single-cell analysis can define distinct evolution of tumor sites in follicular lymphoma. *Blood.* 2021. <https://doi.org/10.1182/blood.202009855>.
33. Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res.* 2020. <https://doi.org/10.12688/f1000research.16665.2>.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009. <https://doi.org/10.1093/bioinformatics/btp324>.
35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010. <https://doi.org/10.1101/gr.107524.110>.
36. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods.* 2018. <https://doi.org/10.1038/s41592-018-0051-x>.
37. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* 1999;9(8):677–9.

38. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018. <https://doi.org/10.1093/nar/gkx1153>.
39. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Cancer Nucleic Acids Res.* 2019. <https://doi.org/10.1093/nar/gky1015>.
40. Chen H, Jiang Y, Maxwell K, Nathanson K, Zhang N. Allele-specific copy number estimation by whole exome sequencing. *Ann Appl Stat.* 2017;11:1169–92. <https://doi.org/10.1214/17-AOAS1043>.
41. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *PNAS.* 2016;(113). <https://doi.org/10.1073/pnas.1522203113>.
42. Bishop CM. *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Heidelberg: Springer-Verlag; 2006.
43. Escobar MD, West M. Bayesian Density Estimation and Inference Using Mixtures. *J Am Stat Assoc.* 1995;430:577–88.
44. Vats D, Knudson C. Revisiting the Gelman-Rubin Diagnostic. *Stat Sci.* 2021;36(4):518–29. <https://doi.org/10.1214/20-STS812>.
45. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multi-modal and scalable single-cell analysis. *Nat Biotechnol.* 2023. <https://doi.org/10.1038/s41587-023-01767-y>.
46. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014. <https://doi.org/10.1186/s13059-014-0550-8>.
47. Lun A, McCarthy D, Marioni J. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2; peer review: 3 approved, 2 approved with reservations]. *F1000Research.* 2016. <https://doi.org/10.12688/f1000research.9501.2>.
48. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* 2005. <https://doi.org/10.1073/pnas.0506580102>.
49. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011. <https://doi.org/10.1093/bioinformatics/btr260>.
50. Wikipedia contributors. Pearson's chi-squared test — Wikipedia, The Free Encyclopedia. 2024. https://en.wikipedia.org/w/index.php?title=Pearson%27s_chi-squared_test&oldid=1224191275. Accessed 13 Aug 2024.
51. Koning MT, Kielbasa SM, Boersma V, Buermans HPJ, van der Zeeuw SAJ, van Bergen CAM, et al. ARTISAN PCR: rapid identification of full-length immunoglobulin rearrangements without primer binding bias. *Br J Haematol.* 2016. <https://doi.org/10.1111/bjh.14180>.
52. Oksza-Orzechowski K, Quinten E, van Bergen CAM, Szczurek E, Shafighi S, Kielbasa SM, et al. CaClust: linking genotype to transcriptional heterogeneity of follicular lymphoma using BCR and exomic variants. Dataset K4B. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE252344>. Accessed 4 Oct 2024.
53. Oksza-Orzechowski K, Quinten E, van Bergen CAM, Szczurek E, Shafighi S, Kielbasa SM, et al. CaClust: linking genotype to transcriptional heterogeneity of follicular lymphoma using BCR and exomic variants. Dataset K5B. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE252416>. Accessed 4 Oct 2024.
54. Oksza-Orzechowski K, Quinten E, van Bergen CAM, Szczurek E, Shafighi S, Kielbasa SM, et al. CaClust: linking genotype to transcriptional heterogeneity of follicular lymphoma using BCR and exomic variants. Dataset K6B. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE252642>. Accessed 4 Oct 2024.
55. Oksza-Orzechowski K, Quinten E, van Bergen CAM, Szczurek E, Shafighi S, Kielbasa SM, et al. CaClust: linking genotype to transcriptional heterogeneity of follicular lymphoma using BCR and exomic variants. Dataset K7B. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE252687>. Accessed 4 Oct 2024.
56. Oksza-Orzechowski K, Quinten E, van Bergen CAM, Szczurek E, Shafighi S, Kielbasa SM, et al. CaClust: linking genotype to transcriptional heterogeneity of follicular lymphoma using BCR and exomic variants. WES dataset. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1062119>. Accessed 5 Oct 2024.
57. Oksza-Orzechowski K, Quinten E, van Bergen CAM, Szczurek E, Shafighi S, Kielbasa SM, et al. CaClust: linking genotype to transcriptional heterogeneity of follicular lymphoma using BCR and exomic variants. GitHub. 2023. <https://doi.org/10.5281/zenodo.13861131>.
58. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science.* 2014. <https://doi.org/10.1126/science.1245316>.
59. Reinius B, Mold J, Ramsköld D, Deng Q, Johnsson P, Michaëlsson J, et al. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat Genet.* 2016;48:1430–5. <https://doi.org/10.1038/ng.3678>.
60. Naik HC, Hari K, Chandel D, Mandal S, Jolly MK, Gayen S. Semicoordinated allelic-bursting shape dynamic random monoallelic expression in pregastrulation embryos. *iScience.* 2021;24(9):102954. <https://doi.org/10.1016/j.isci.2021.102954>.
61. Larsson A, Ziegenhain C, Hagemann-Jensen M, Reinius B, Jacob T, Dalessandri T, et al. Transcriptional bursts explain autosomal random monoallelic expression and affect allelic imbalance. *PLoS Comput Biol.* 2021. <https://doi.org/10.1371/journal.pcbi.1008772>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.