PhD degree in Systems Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

MiTo: robust inference of mitochondrial phylogenies and clones

Settore disciplinare: MED/04

Andrea Cossa

Tutor: Pier Giuseppe Pelicci

Istituto Europeo di Oncologia

PhD Coordinator: Prof. Diego Pasini

Anno accademico 2023-2024

To hidden variables,
and to everyone chasing them

Table of content

List of Figures

Abstract

Somatic evolution, the process by which cells acquire genetic and epigenetic changes throughout an individual's lifetime, underlies both normal development and diseases like cancer. Single-cell lineage tracing (scLT) has emerged as a powerful approach to study these cellular dynamics, especially in primary tissues. In this scenario, mtDNA variants (MT-SNVs) have gained special attention recently, due to their low-profiling costs and compatibility with other informative cell-state modalities.

This thesis introduces MiTo, an novel toolkit for MT-SNV-based scLT. MiTo provides an integrated pipelines for flexible preprocessing of scLT data, lineage inference, and interactive exploration of MT-SNV-derived phylogenies and clonal structures. MiTo integrates seamlessly with popular single-cell analysis libraries, filling a significant gap in the scLT community. To benchmark MiTo, we generated a new single-cell multi-modal dataset, with simultaneous and longitudinal profiling of gene expression, expressed MT-SNVs, and lentiviral barcode labels. We used this dataset to benchmark several tasks in MT-SNVs based scLT (MT-scLT), demonstrating superior performance of MiTo compared to state-of-the-art tools.

Here, we found that informative MT-SNVs spaces may include even rare (i.e., 0.02-0.03 allelic frequency in at least 2 cells, and mean 1.2-1.5 alternative UMIs) detection events, but that statistically sound MT-SNVs genotyping methods are needed to handle the intrinsic noise of single-cell measurement, especially in high-clonal-complexity scenarios. Moreover, we show that MT-SNVs-based-phylogenies (MT-phylogenies) exhibit remarkable robustness to noise, even though the constrained number of available characters limits their resolution. Finally, by tracing clonally-enriched MT-SNVs, we show that multiple sub-clonal lineages within individual clones participate to the metastatic dissemination in Breast Cancer xenografts, implying stronger lineage-dependency of the metastatic phenotype in Breast Cancer that previously thought.

In summary, this work highlights opportunities and limitations of MT-scLT, providing novel data analysis tools and benchmarking datasets, and demonstrating the power of scLT to investigate complex cellular dynamics in somatic evolution.

Introduction

Cancer evolution

Malignant cells arise, adapt and survive through acquisition and selection of molecular modifications. Across decades, several attempts have been made to reduce the multi-faceted nature of this process into a reasonable number of universal features. Since 2001, the conceptualization of “Cancer hallmarks” by Hanahan and Weinberg [1,2] represented one of the most successful and recognized efforts in this direction. Originally six and more recently updated to eight, “Cancer hallmarks” currently comprise: the ability of cancer cell to 1) sustain proliferative signalling, 2) evade growth suppressors, 3) resist cell death, 4) unlock replicative immortality, 4) induce/access vasculature, 5) invade adjacent and distal tissue, 7) reprogram cellular metabolism, and 8) avoiding immune destruction. Genome instability and tumor-promoting inflammation have been recently added to this scheme as “enabling characteristics”, i.e., cellular molecular mechanisms by which hallmarks are acquired, rather than distinctive capabilities by themselves. Together, these traits fix in a defined number of abstract categories distinct features of cancer cells, and provide a unifying scaffold that holds together the overwhelming phenotypic heterogeneity observed both in the clinics and in experimental models. However, recent years have made increasingly clear that cancer should be considered more like a systemic disease rather than a localized, tissue-specific condition, as the role of non-cell-autonomous processes has been more and more appreciated. In particular, it has been demonstrated how cellular interactions within the Tumor Micro-Environment (TME) - defined as heterogeneous population of cancer stromal and immune – play a fundamental role in shaping the disease biology [3,4]. As previously recognized [1], despite its utility, any reductionist conceptualization scheme fails to address the full complexities of cancer pathogenesis, namely, the precise molecular and cellular mechanisms that allow evolving pre-neoplastic cells to develop the aberrant phenotypic capabilities that fuel malignant progression.

In this scenario, the most universal and defining trait of cancer is arguably its evolvability, i.e., the capacity to evolve, the ability of cancer cells to generate molecular variation that is selected and remodelled by the environment[5]. In essence, cancer is the evolutionary process through which a subset of “outlaw” cells in a multi-cellular organism breaks their contract with multi-cellularity to begin a new, de-regulated life[6]. In ecological terms, cancer

cells are endogenous, fast-evolving parasites attacking their host organism until stopped by environmental constraints, or by the organism collapse. Thus, it is not surprising that evolutionary concepts have been adopted to study and measures cancer biology and evolution since the very first days of cancer research. The evolution theory is one of the oldest paradigms used to study cancer biology. Since the seminal work of Nowell in 1976 [7], the evolution of cancer has been recognized as “clonal”. In cell biology, a ”clone” is the group of cells that stems from an single ancestor after a sequence of cellular division aancestor, while generating additional molecular changes at each cell division that can be potentially passed to the off-spring. Despite recent emerging evidence suggesting how cancer evolution might be initiated by multiple, independent cellular ancestors[9], the simple idea of multiple cancer clones (or lineages) competing for their ecological niche in their host naturally fits almost all the scattered biological knowledge that we have on such a disease [5,6].

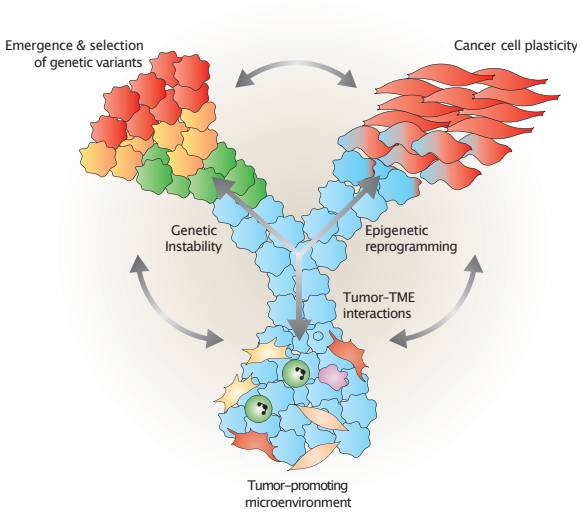


Fig 1. Cancer evolution*. Readapted from Cirello et al. [6].*

Currently, two unanswered questions dominate the current view of cancer evolution: i) What is the modality through which cancer evolve? Is cancer evolution Darwinian, Lamarkian, or a mix of both? ii) Are inheritable genetic (i.e., mutations) and epi-genetic (i.e., DNA methylation, histone-modifications, chromatin structure) or transiently acquired (i.e., non-heritable gene expression changes) traits the driving force of cancer? What is the most relevant molecular layer through which disease phenotypes are encoded? [5,6]
Dividing cells accumulate mutations throughout a lifetime producing genetic heterogeneity, the substrate for Darwinian evolution. Mutations are thought to emerge randomly, and are distributed across the genome as the result of DNA damage/repair or inaccuracies of the

DNA replication process. In cancer cells, mutations accumulate over time, gradually or in more punctuated or even catastrophic manner, with the resulting Intra-Tumoral-Heterogeneity (ITH)[10] recognized as a pervasive feature in human cancers. Whereas most somatic mutations have neutral phenotypic effects (i.e., passenger mutations), occasional mutations may confer a selective advantage to the cell (i.e., driver mutations), potentially leading to clonal expansion. These driver mutations can alter cell fitness through different mechanisms. “Classic” driver mutations were limited to well-characterized protein-coding genes, i.e., Loss of Function mutations in tumor suppressor and Gain of Function mutations in oncogenes, with easily interpretable links between genetic alterations and de-regulated protein functions and pathways. However, it is now more and more appreciated how somatic mutations in more broadly acting genes (i.e., epigenetic factors) may cause global de-regulation of expression programs sustaining the malignant phenotypes[10 12]. Indeed, this has been shown also for non-coding mutations and structural variants, more complex genetic aberrations for which direct mechanistic links to malignant phenotypes are more elusive. Even with a well-annotated catalogue of mutational processes and cancer drivers, the causal route to tumor initiation and progression still remains unclear. According to the old multi-step carcinogenesis theory, cancer development is the result of multi-step driver mutations accumulation, with the gradual acquisition of the malignant phenotype mirrored by single mutational hits[13]. However, this simplistic view was challenged by the recent discovery of ubiquitous clonal selection of cancer drivers in normal tissue mosaicism, which confirmed that human tumorigenesis requires specific combinations of genetic and, possibly, non-genetic alterations [10,12]. Considering only genetic changes, analyses of large tumor cohorts indeed showed that some mutation combinations can be non-random, i.e., specific sets of mutations are more frequently observed together than expected by chance (co-occurring mutations), while others are rarely or never found together (mutually exclusive mutations) [11] . These non-random co-mutation patterns highlighted evolutionary trajectories where the timing and occurrence of specific genetic lesions contribute to induce malignant phenotypic changes. However, the presence of cancer drivers in phenotypically normal cells [14], together with the existence of tumors with minimal presence of genetic aberrations[15], still suggests that additional, non-genetic mechanisms are required to induce overt malignancies.

In recent years, the notion of heterogeneous “cancer cell states” in otherwise genetically identical cells has gained traction. The existence of these cellular states has been discovered by transcriptional and/or epigenetic profiling of individual tumor cells, with the term “state” potentially integrating static (genetic) and dynamic and reversible (epigenetic and transcriptional) determinants[16,17]. Specifically, the vast majority of annotated cancer cell states have been identified from single cell data using methods that factorize high-

dimensional phenotypic read-outs (e.g., gene expression or chromatin accessibility counts) into a much smaller number of orthogonal regulatory programs. Such programs are either themselves defined as cell states, or clustered into cell states, consistent with the notion that cell phenotypes are the result of concurrent activation of different biological processes[18]. Interestingly, one of the cellular mechanisms underlying this major source of phenotypic heterogeneity is cell plasticity, a well-known concept in developmental biology[6].[19,20]

Despite sharing the same genome, any given cell must be able to respond to chemical and positional cues in its surroundings. This property enables the zygote to develop into a full multi-cellular organism, giving rise to a hierarchy of phenotypically distinct cells, specialized tissues and organs. Post-development, normal cells must still be able to respond to environmental perturbation (e.g., in wound healing) to ensure that tissue integrity and homeostasis are maintained. These known (de-)differentiation processes are regulated by epigenetic factors that are capable of reshaping the reading frame through which genome is expressed, switching on and off alternative expression programs. In the early days of epigenetics, genetic constraints regulating these phenotypic transitions have been depicted as an imaginary potential energy surface (i.e., the Waddington landscape) upon which cells committing to specialised cell types (or de-differentiating into high-potency states) roll up and down [21]. This “canalized” nature of cell identity and behaviour if fundamental for tissue homeostasis. Since all molecular interactions controlling gene expression are inherently noisy, higher-order organisms evolved complex genetic interactions able to repress harmful, uncontrolled regulatory switches. On the contrary, in cancer, these genetic constraints are extensively rewired[19]. Thus, when the fitness of a tumour cell becomes its ability to survive and proliferate in face of a changing environment, the molecular machinery controlling phenotypic plasticity can be co-opted as an extraordinary adaptation tool, opening the possibility for non-Darwinian evolution, deeply influenced by the environment. In summary, compared to normal developmental cell types, cancer cell states are much more transient, inducible, and potentially, reversible. Crucially, these properties have many fundamental clinical implications: i) genetically identical cells may be phenotypically heterogeneous, implementing different functions that are all needed to foster disease progression and treatment resistance. Therefore, targeting the right cell state/state transitions could become an important therapeutic strategy to control the disease; ii) treatment and micro-environmental conditions likely perturb cancer cell states, potentially driving state transitions among them. Rational, evolutionary-informed treatment designs could help predict such damaging phenotypic transitions, avoiding unwanted and potentially harmful treatment side-effects; (iii) there might be opportunities to revert cancer cell malignant states, or at least, to redirect them towards more harmless cellular

phenotypes. Precise identification of the molecular determinants sustaining cancer cell states is fundamental to achieve rational interception of each patient's disease trajectory; iv) Each human genome has its unique genetic make-up and predispositions to cancer. However, specific environmental conditions may be required for a full malignant transformation to take place. A comprehensive dissection of these additional enabling factors would be fundamental in cancer prevention and it the early stages of the disease [5,6,19].

Importantly, genetic and non-genetic mechanisms of cancer evolution are by no means mutually exclusive. Genotype and phenotype integrate with environmental cues to produce functional cell states and behaviours. This realization dramatically shifts the area of interest of cancer evolution as an entire field, from the simple characterization of genetic paths associated with specific clonal dynamics to a more integrated view of cancer as a complex cellular system, able to evolve through different molecular mechanisms, some of which reversible, and therefore, open to pharmacological modulation and/or prevention. This new conceptual framework translates the gene-centric view of a prime cancer evolution to an emerging, cell-centric and multi-omic view. This is how emerging methods in single-cell biology have met long-standing questions in cancer biology and evolution, as we will see in the next chapters.

Single-cell multi-omics

For more than two decades now, bulk DNA and RNA sequencing measured molecular information recording billions of digital from individual DNA (or cDNA) molecules. Despite being an incredible breakthrough in the early 00s, the output of a bulk sequencing experiment represents "average" read-outs across sampled cells, posing significant challenges in the precise deconvolution of cellular states from complex cellular mixtures[22]. Dedicated bioinformatic tools have been developed to estimate the relative proportions of cell types in complex tissues from their gene expression profiles[23], but low intensity signals from rare cell populations might remain undetectable with bulk, which precludes identification of rare (yet potentially relevant functionally) cell populations. Since the beginning of the early 10s, single-cell assays came to the stage. Based on different chemistries for molecular barcoding, these sequencing platforms allow profile molecular content of individual cells, providing high-resolution representations of cell types and states in development[24], immunity[25], ageing[26], and cancer[12]. Due to this unprecedented resolution, single-cell techniques have revolutionized how we understand cell heterogeneity, mechanisms of gene regulation, protein expression dynamics and epigenetic variation.

Different molecular layers can be probed with single-cell sequencing. In recent years, a great variety of protocols came out, each measuring one of more molecular layers (e.g., DNA, RNA, proteins, post-translational modifications, chromatin accessibility, DNA methylation, genome organization, to name a few). Here, we will focus primarily on single-cell RNA sequencing (scRNA-seq, the main single-cell sequencing technology adopted in this work) and single-cell DNA sequencing (scDNA-seq, to highlights challenges in cost-effectiveness, scalability, and reliability of molecular information retrieved from low input DNA protocols). Then, we will give a brief and general overview of existing multi-omics methods, and their application landscape.

In 2009, the first scRNA-seq method was reported[26]. The field has rapidly developed ever since, dividing into two main categories: plate-based and droplet-based transcriptomics. Plate-based scRNA-seq protocols isolate cells into unique micro-wells for single-cell library preparation[28]. Specifically, the first step usually entails cell sorting by, for example, fluorescent-activated cell sorting (FACS), where cells are sorted according to specific cell surface markers; or by micro pipetting. The selected cells are then placed into individual wells containing cell lysis buffers, where subsequently reverse transcription is carried out.

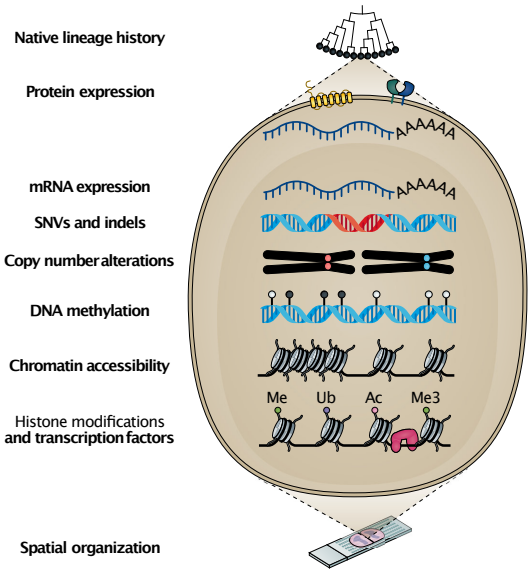


Fig.2 Single-cell multi-omics layers. Readapted from Nam et al., [10]

This allows for several hundreds of cells to be analyzed in a single experiment with 5000 to 10000 captured genes each. These protocols provide full-length information about RNA-molecules, with library preparation steps very similar to bulk RNA-seq sequencing. In spite of positional coverage biases of these protocols is that they: i) allow genotyping of variants

across the whole transcript length, and ii) allow detection of splice variants. However, these protocols are generally more expensive and have less throughput than other droplet-based protocols [29]. Droplet-based protocols isolate cells into droplets to create cell-specific reaction chambers into which individual transcripts are labeled (for their cell and molecular identity) before amplification and sequencing. Specifically, individual cells are encapsulated in nanoliter-sized hydrogel-droplets with specialized microbeads. Despite differences across protocols (e.g., inDrops[30], Drop-seq[31], 10x[31]) regarding the nature of the beads and the mechanics of this encapsulation process, these micro-beads have attached primers containing a PCR handle, a cell barcode (CB) and unique molecular identifier (UMI) and a poly-T tail (or in the case of a 5' kit, there will be a poly-T primer.). Upon lysis the cell's mRNA is instantaneously released and captured by the barcoded oligonucleotides that are attached on the beads. Next, the droplets are collected and broken to release single-cell transcriptomes attached to microparticles (STAMPs). This is followed by PCR and reverse transcription to capture and amplify the transcripts. Finally, tagmentation takes place where the transcripts are randomly cut and sequencing adaptors get attached. This process results in sequencing libraries that are ready for sequencing. In microfluidic based protocols only about 10% of the transcripts of the cell are recovered. However, this low sequencing is sufficient for robust identification of cell types. These protocols usually sequence only the 3' or 5' ends of the transcripts, making it difficult to unambiguously align reads to a transcript and distinguishing between different isoforms. However, the use of UMIs allows accurate identification of PCR duplicates, an important task for downstream analysis[33].

The typical scRNA-seq analysis workflow goes as follows[34]: raw reads from single-cell sequencing are aligned to a reference, and reads/UMI counts are recorded for each gene feature, to retrieve cell x genes counts matrices. To ensure that only high-quality cells are captured, count matrices are corrected for cell-free ambient RNA and filtered for doublets and low-quality or dying cells. The latter is done by removing outliers with respect to quality control metrics (the number of counts per barcode, called count depth or library size, the number of genes per barcode and the fraction of counts from mitochondrial genes per barcode (percentage mito.)). Then, in order to compare gene expression among cells raw counts are normalized to correct for cell-specific technical effects influencing capture efficiency. Single-cell RNA sequencing (scRNA-seq) data sets can contain counts for up to 30,000 genes for humans. However, most genes are not informative, with many genes having no observed expression. Therefore, the most variably expressed genes are selected, and, especially with complex experimental design or single-cell atlases, batches of data need to be integrated to get harmonized dataset with preserved biological variation and limited technical factors influencing cell-cell differences in gene expression[35,36]. This

corrected gene expression space (either batch-corrected counts, denoised latent space or kNN graph) is then used for downstream analyses. A gene expression space can then be organized into discrete clusters, which represent groups of cells with similar gene expression profiles, annotated by labels of interest such as cell type. This annotation can be conducted manually using prior knowledge or with automatic annotation approaches able to transfer biological labels with query-to-reference mapping[37]. Continuous processes, such as transitions between cell identities during differentiation or reprogramming, can be inferred to describe cellular diversity that does not fit into discrete classes[38]. Gene regulatory networks and gene modules can be inferred from gene-gene correlations across cells[39]. Depending on the question of interest and experimental set-up, conditions in the data set can be tested for upregulated or downregulated genes (differential expression analysis), effects on pathways (gene set enrichment) and changes in cell-type composition[34]. Perturbation modelling enables the assessment of the effect of induced perturbations and the prediction of unmeasured perturbations[40,41]. Moreover, expression patterns of ligands and receptors can reveal altered cell–cell communication[42].
scDNA-sequencing is a much more problematic deal than scRNA-sequencing, and despite technical advantages, the use of scDNA-seq remains limited[43]. Single-cell genome sequencing enables the elucidation of genetic heterogeneity; thus, it can be used for the analysis of de novo germline mutations and somatic mutations in normal and cancer cells, at the highest possible resolution. To uniformly amplify genomic DNA in individual cells, whole-genome amplification (WGA) methods have been developed [44], such as multiple displacement amplification (MDA)[45], multiple annealing and looping-based amplification cycles (MALBAC)[46] and degenerate oligonucleotide-primed PCR (DOP-PCR)[47] . However, WGA is challenging due to the presence of only two genomic copies DNA in human cells. Therefore, this strategy occasionally greatly suffer from allelic drop-out, and fails to achieve a uniform sequencing depth because of amplification bias[43]. For automatic library construction, the Fluidigm C1 system supports single-cell whole-genome and whole-exome sequencing. Additionally, 10× Genomics recently released a copy number variant (CNV) solution for the Chromium system to profile copy numbers in single cells. In cancer, researchers have attempted to identify intratumor genetic heterogeneity generated during cancer evolution. However, even if specialized bioinformatics methods (e.g., SCcaller[48], Monovar[49], LiRA[50], and Conbase[51]) have been developed to detect single-nucleotide variants (SNVs) considering allelic dropout and amplification artifacts, assess genetic heterogeneity through scDNA-seq is still very costly and technically challenging, and currently cannot scale to more than 10s-100s cells per experiment.
scRNA- and scDNA-seq provide valuable information from a single molecular layer, i.e., gene expression and genetics, respectively. Thus, in recent years, a number of works

attempted to complement this information by coupling RNA/DNA measurements with other single-cell read-outs, to achieve higher phenotyping depth[52]. Simultaneous measurement of genome and transcriptome in single cell was reported for the first time in 2014[52], followed by other protocols. In G&T-seq[54], oligo(dT)-coated magnetic beads are used to separate genomic DNA from full-length mRNAs in single cells, followed by a modified Smart-seq2[55] protocol in which cells are isolated and lysed to release genomic DNA and mRNA for whole-genome and whole-transcriptome analyses, respectively. In 2019, TARGET-seq[56] , a plate-based method, demonstrated increased throughput of ~5,000 single cells profiled per run through the use of barcoding and pooling of libraries in reduced reaction volumes. Another plate-based method, simultaneous isolation of genomic DNA and total RNA (SIDR), involves incubation of single cells with antibody-conjugated magnetic beads, which are subsequently sorted into a microplate in which hypotonic selective lysis produces a separated supernatant and pellet solution that distinguishes between DNA and total RNA[56].

Coupled genome and transcriptome profiling cannot address the question of how the same DNA sequence can have varying expression patterns in different cells. The epigenome layer is needed in tandem with gene expression to elucidate mechanistic relationships among as DNA methylation, DNA accessibility and histone modifications and the gene expression patterns they produce. DNA methylation profiling methods (e.g., reduced-range bisulfite sequencing, RRBS, and whole-genome bisulfite sequencing, WGBS) have been re-adapted for single-cell multi-omics (e.g., the scM&T-seq[58] protocol, enabling single-cell genome-wide methylome and transcriptome sequencing). The same has been done for accessibility profiling methods using the assay for transposase-accessible chromatin sequencing (ATAC-seq), which was integrated with RNA (ASTAR-seq, SNARE-seq, SHARE-seq, 10x Multiome[59–61]) and DNA measurements (TEA-seq[62]). Protocols detecting bulk-level histone modifications ChIP–seq, and CUT&RUN, CUT&Tag have been modified further to achieve single-cell level resolution (scCUT&Tag[63,] scCUT&Tag2for1[64], scCUT&TAG-pro[63]). Chormosome capture C technologies has also move to the single-cell worlds (scHi-C[65]).

While extensive characterization of gene expression regulation is of pivotal importance, all cellular processes and functions revolve around proteins, as they contribute to the structure of cells and perform biochemical processes by functioning as enzymes. Thus, protein abundance post-translational modifications and interactions are essential to dissect cellular mechanisms that does not depend on differences in gene expression regulation. Sequencing protein is much more difficult than DNA and RNA molecules, due

to the lack of amplification. Thus, existing single-cell proteomic studies uses curated antibody panels and high-dimensional cytometry[66] (FACs, or mass cytometry). For the vast majority, multi-omics protocols in which protein and mRNA are co-profiled, followed the same strategy. For instance CITE-seq[67] and REAP-seq[68], combine highly multiplexed protein-marker detection with unbiased transcriptome profiling for thousands of single cells. Cell-surface proteins are detected by antibodies conjugated to oligonucleotides containing a PCR handle that can be captured by oligo(dT) or probe-specific primers compatible with Drop-seq and 10x Genomics microfluidic systems, making this method adaptable for integration with most scRNA-seq methods. Recently, ATAC with select antigen profiling by sequencing (ASAP-seq[69]) and DOGMA-seq[69] demonstrated simultaneous profiling of chromatin accessibility. Incorporating even more modalities is NEAT-seq[70], which co-profiles the abundance of nuclear protein epitopes, chromatin accessibility and the transcriptome in single cells.

Additionally, since the advent of CRISPR-screens and the massive use of single-cell sequencing to characterized the immune system, emerging protocols couple previously described data modalities with sgRNAs (i.e., marking specific genetic perturbations individual cells[71]) and T-/B-cell receptor sequencing (to resolve T- and B-cell clonotypes[72,73]). Finally, the last decade has seen extraordinary efforts to add the spatial dimension to single-cell measurements[74].

Depending on the experimental question, design, and data modalities, analysis of single-cell multi-omics data may consist of different tasks, including data integration, multi-modal gene-regulatory network, perturbation and trajectory inference, with statistical models more and more physically informed about the relationships between single-cell molecular layers[34].

In summary, these technological advancements produced a tremendous data-driven shift in the way we think about cell biology[75]. Single-cell multi-omics was fundamental to uncover previously unknown cellular types and to characterize cellular phenotypes and molecular mechanisms underlying virtually all aspects of human health and disease. In the next chapter we will focus on a specific branch of single-cell biology, single-cell lineage tracing, detailing scope, methods, and areas of application.

Lineage tracing at single-cell resolution

The term lineage tracing refers to a wide range of techniques developed to identify and characterize individual cell progenies. From an historical perspective, the beginnings of lineage tracing date back to very ancient times: 19th century developmental biology[76].

The first pioneer of lineage tracing was Charles O Whitman. He and colleagues, inspired by the observation that cells arise from pre-existing cells, rather than through spontaneous generation, began investigating early cleavages in invertebrate embryos. Whitman studied leech development, which involves stereotypical, invariant cell divisions. Tracing individual cell fate through direct observation under a light microscope, he discovered that from the earliest cleavages, individual cell fates were developmentally distinct, with each cell division giving rise to cells with specialized roles in later development. A century after these initial studies, another pioneer, Sulston[77], successfully took the challenge of determining the fate of every cell in the Caenorhabditis Elegans embryo, "by eye". Development in C. elegans, as in the leech, is highly determinate, involving cell fate decisions that are entirely autonomous. As one could imagine, tracing individual cell lineages (i.e., progenies) by eye has severe limitations in whole embryos/organisms, as it feasible only for transparent embryos (or organisms) and a small number of cells. However, these limitations do not apply to lineage analysis of single cells in culture, and cell biologists have long used time-lapse microscopy to determine lineage relationships. For example, isolated precursor cells of the central nervous system from rat embryos[78]. By imaging single cells over a number of days, Temple determined whether cells divided and formed clones assessed cell progenies differentiation status. In this rudimental way, she discovered the stochasticity in cell fate decisions that distinguish mammals from C. elegans and leech early development. When direct observation was not possible, 19th and 20th century scientists came up with different strategies to test hypothesis about cell fate and evolutionary relationships, including: i) embryos chimeras, ii) tissue transplants and iii) cellular labelling, namely, the process through which one attach to individual cells inheritable molecular "labels", or "barcodes" (i.e., dyes, or exogenously inserted genetic markers). Tracing these "labels" in time provided precious information about the proliferative capacity individual lineages, and their ability to acquire different fates. Importantly, regardless of their physical nature, these "labels" must strictly satisfy the following properties to be considered reliable lineage markers: i) they need to be passed to the entire progeny of a founder cell, ii) they must be retained over time, and iii) they should never be transferred to unrelated cells from independent lineages[76].

While lineage tracing roots in developmental biology, another discipline used "markers" with the same properties to trace the clonal origin and cellular fate in cancer: cancer phylogenetics[79]. Cancer phylogenetics deals with the inference of cancer lineage histories (i.e., trees) from somatic mutations (i.e., mostly SNVs and CNVs) that "mark" newly

generated cancer lineages across cancer evolution. Similar to species phylogenetics, these graphical models (i.e., trees) encode for the relationships between biological entities (i.e., samples, deconvolved clones, or single-cells), with internal nodes representing either inferred putative ancestors or evolutionary steps in a chain of mutation of ordered mutation events, and branches representing (qualitatively or quantitatively) some notion of time between inferred events. Within this basic framework, tumour phylogenetics can be remarkably heterogeneous. Different sources of data have been used to reconstruct cancer trees, considering both the study design (i.e., cross-cohort studies of many tumours, single-patient studies of regional bulk genomic assays, or studies of single-cell variability in single tumours) (FIG. 1) and the type of genomic data profiled (initially, pre-sequencing marker types, such as large-scale CGH or fluorescence *in situ* hybridization[80]; now, predominantly next-generation sequencing (NGS)-derived SNVs or CNVs[80], and sometimes more exotic variant types such as gene expression, DNA methylation, or histone marks[82]). Moreover, different models have been developed to model the evolution of different kinds of mutations (for example, SNVs versus structural variants), each with different assumption regarding underlying selection processes[79]. Furthermore, a number of different algorithms have been employed to for cancer phylogeny reconstruction. While most works in the filed adapted standard algorithms from species phylogenetics (e.g., maximum parsimony[82], minimum evolution[84], Neighbour Joining[84], UPGMA[85], Maximum Likelihood and Bayesian inference[87]) to infer cancer trees, novel algorithms were designed to suit the peculiar properties of cancer evolution[79], such as the high occurrence of mutations, the balance between strong purifying selection and neutral phases of evolution. As introduced previously (see Cancer evolution section), these cancer tree can represent either ordered combinations of mutations appearing at certain tumor progression stages, as inferred from cross-sectional studies, or individual histories of tumor clones inferred from (multi-regional) bulk DNA sequencing data. In this latter case, each sequenced sample is treated as a separate species. Alternatively, clonal deconvolution algorithms are used to before tree building (or jointly integrated in this step) to infer clones, i.e., clusters of co-occurring mutations with similar prevalence that are assumed to occur in the same fraction of cells in the tumor[79].

Whichever the case, the great majority of cancer trees available to date have been inferred from retrospective, bulk-DNA sequencing data. Only recently the rise of single-cell biology provided new ground for both lineage tracing and cancer phylogenetics.

Single-cell experiments provide high-dimensional “snapshots” of dynamic processes[88]. Depending on the experimental design, these snapshots can be single, or ordered in longitudinal time series. Assuming: i) complete and error-free measurement of all relevant

molecules governing a dynamic process of interest, and ii) that cell-fate decisions are governed deterministically only by interactions among measured molecules, one could infer the regulatory mechanisms underlying these cellular dynamics, and use it to predict future cell states starting from arbitrary initial conditions[89,90].

Unfortunately, single-cell measurements are noisy and largely incomplete, most cell-fate decision are inherently stochastic (at least to some extent), and single-cell molecular spaces include poor information about cell-extrinsic factors influencing cell fate decisions[91]. In addition, very differently from direct observation in prime lineage tracing studies, sequencing-based detection of the molecular content of a cell currently require the cell lysis (with a single, notable exception: a proof-of-concept study demonstrated very recently the possibility of "puncturing" cell cytoplasm to recover transcripts for scRNA-seq sequentially and in the very same cells[92]. However, this is still very far from being adopted in common practice, even in pure research settings). Thus, even when longitudinal measurements of the same process are provided in replicate experiments "photographed" at different timepoint, cellular dynamics inference is a non-trivial task[90].

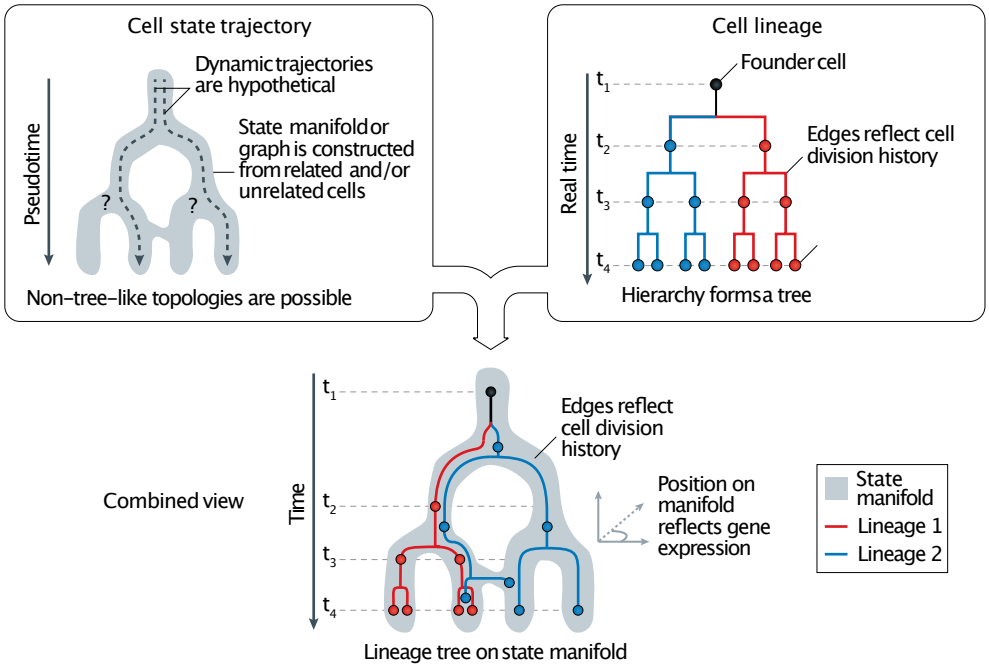


Fig.3 Single-cell lineage tracing. *Readapted from Wagner et al.,* [87]

Different approaches have been employed to extract approximate (but meaningful) cellular dynamics inferences from static cell state snapshots. For instance, leveraging the asynchrony through which cells progress across dynamic processes, pseudotime methods

are able order cell states according to their diffusion distance from a “source” state that represent the starting point of the cellular dynamics[93]. More complete descriptions of cellular dynamics used the Fokker-Planck equation, integrating drift (i.e., the ideal potential energy surface regulating cellular dynamics) with cell growth and similarity in cell state space[94,95]. Other methods took advantage of unspliced and spliced mRNA measurements to infer cellular dynamics in the form of “RNA velocity”[96], which providing a notion of directionality to cell state change. Finally, the optimal transport (OT) [89,97] framework has been used to learn cellular dynamics. Specifically, OT has been used to infer optimal transport maps that match similar cell states in adjacent time points, providing inferred ancestor-descendant cell state transition probabilities.

Importantly, all of the afore-mentioned methods try to make inferences only from the observed cell states, without any other external evidence guiding such inference. Thus, all of these methods suffer from the same limitation: they all assume that the closer two states are in some high-dimensional space, the closer would be in physical time. This assumption may be exceptionally wrong, leading to inaccurate inferences [88,90,91].

To circumvent these fundamental limitations, a variety of tools has recently emerged to measure ancestor-descendant relationships in single-cell experiments. These systems, collectively referred to as single-cell lineage tracing (scLT) systems, integrate “old-school” lineage tracing and phylogenetics techniques with the unprecedented phenotypic depth of single-cell sequencing.

Currently, there are two major paradigms for scLT: prospective and retrospective scLT [88,98].

As the name suggest, prospective lineage tracing attempts to establish lineage relationships across cells forwards in time. Different prospective methods have been developed. For instance, fate mapping, the practice of associating the position of a cell in the early embryo with the ultimate positions and fates of that cell’s descendants, can be considered one of the first prospective lineage tracing systems. Recombination-based systems such as early cellular barcoding methods based on CRE or FLIP recombinases, inducing permanent genetic labelling of progenitor cells based on the activity of a transgenic promoter, are also included among prospective scLT methods[76]. Other examples of prospective lineage tracing systems include multi-color reporter constructs, such as double markers[99] (MADM, e.g., GFP and LacZ) or the ‘‘Brainbow’’ mice, which, through independent recombination of four fluorescent proteins, can be used to stably mark multi-color clones[100]. Despite their historical and scientific value, these approaches are intrinsically limited in their throughput, i.e., the number of independent lineages that can be simultaneously labelled[88]. Moreover, these methods are not readily coupled with other phenotypic assays, and thus, they only inform about tissues clonal dynamics, without

providing any information about the underlying cellular phenotypes. The advent of DNA-barcoding methods was a major step forward in this sense[98]. With current DNA-barcoding strategies, high complexity retroviral or lentiviral libraries (e.g., ~10^6-10^7 unique viral particles) can be used to mark complex population of cells[100–102]. These static DNA barcodes stably integrate in a cell genome and inherited by descendant cells at each subsequent mitotic division. Targeted DNA sequencing of these barcodes can be used to measure abundance of individual clones at the bulk level. However, if these barcodes are expressed as exogenous poly-adenilated transcripts, simultaneous measurement of clonal identity and cell state can be achieved by scRNA-seq[104–106]. While these methods allow detection of static clonal labels, other recently developed DNA-barcoding strategies leverage evolving DNA barcodes. These dynamic barcoding strategies can be achieved through either continuous *in vivo* transposition of the same DNA construct across cell genomes [107], or continuous CRISPR-Cas9 editing of a pool of target arrays[108–112]. Detection of these random transposition sites or CRISPR-Cas9 induced genetic "scars" may be used to infer complete cell phylogenies, giving temporal resolution to lineage speciation events. Despite technical challenges, that needs to be overcome, these methods provide unambiguous and tightly controlled definition of cell lineage in single-cell experiments. However, their main limitation is the requirement for cell engineering, which limits their application to model systems.

By contrast, retrospective scLT methods seek to map the history of lineage relationships with respect to the cell states usually sampled at a single end point[88,98]. With these methods, state and lineage features are measured only at the end of the experiment, and lineage relationships are mapped backward in time in order to infer earlier lineage speciation events and associate phenotypic changes decisions. Importantly, retrospective scLT use natural genetic markers as genetic barcodes, and therefore, they can be applied to human patient samples and in other cases in which experimental intervention is not possible. Different endogenous lineage markers have been tested in recent years[113], including nuclear SNVs[114] and CNVs[114], T[116]- and B[117] cells receptor sequences, micro-satellite instability[118], DNA methylation and chromatin accessibility states[118]. All of these markers have potential pros and cons. For instance, nuclear SNVs represents the richest source of somatic genetic alterations in higher-order organism, but these genetic mutations have very low mutation rate and are scattered all the cross nuclear genome, requiring high-cost and technically challenging scDNA Whole Genome Sequencing approaches. T- and B- receptor sequence have demonstrated great value in distinguish T- and B-cell clonotypes, but unfortunately, this cost-effective lineage markers are restricted to these cell types, while other epigenetic markers either require very high

sequencing coverage to detect base-specific epi-alleles, or can be dynamically remodelled as a consequence of regulatory changes, confounding lineage reconstruction[43].
In the next chapter we will see how another, recently emerged natural lineage marker (i.e., mitochondrial single nucleotide variants, or MT-SNVs, hold the promise for cost-effective and phenotypic informed retrospective scLT).

Mitochondrial genetics

The mitochondrial DNA (mtDNA) in humans and other animals has retained several characteristics of its bacterial ancestry. Unlike nuclear DNA, mtDNA is a small, circular dsDNA molecule approximately 16.6 kilobases in length, containing 37 genes that are critical for cellular respiration and energy production. These genes include 13 protein-coding genes necessary for oxidative phosphorylation (OXPHOS), 22 transfer RNAs (tRNAs), and 2 ribosomal RNAs (rRNAs). The protein-coding genes contribute to components of the electron transport chain (ETC), specifically complexes I, III, IV, and V, and are crucial for the cell's ability to generate ATP through oxidative phosphorylation. Mitochondria have a unique double-membrane structure, consisting of an inner and outer membrane. The inner membrane houses the ETC complexes and creates a proton gradient essential for ATP synthesis. Within the mitochondria, mtDNA is organized in nucleoid structures, each containing several copies of mtDNA and associated proteins. Depending on the cell type, human cells contain …-… mitochondria, each containing …-… mtDNA copies, leading to 100s-to-1000s mtDNA copies per cell. This high copy number acts as a buffer, diluting the effect of deleterious mutations and helping maintain the essential role of mitochondria in cellular bioenergetics. mtDNA is tightly packed, with minimal non-coding regions. The D-loop region, one of the few non-coding sections, serves as the primary regulatory site for mtDNA replication and transcription. The mtDNA lacks nucleosomal structures and introns, and has limited non-coding sequences. Given the lack of protective histones and the proximity of mtDNA to the ETC (a significant source of reactive oxygen species, or ROS), mtDNA is more vulnerable to damage than nuclear DNA[120,121].

Differently from nuclear DNA, mtDNA replication is characterized by both "relaxed" and "stringent" control modes, alongside dynamic anterograde and retrograde signaling that coordinates mitochondrial function with cellular demands[122]. In relaxed replication, mtDNA replicates independently of the cell cycle, allowing for multiple mtDNA copies within each mitochondrion . This unregulated replication ensures a sufficient supply of functional mtDNA to meet the high ATP production demands of various cell types. By contrast,

stringent replication is tightly controlled, occurring at specific stages, such as during oogenesis or early embryogenesis. During stringent replication, a bottleneck effect reduces mtDNA diversity, favoring the inheritance of intact mtDNA by limiting the number of copies transmitted. In addition to replication modes, mtDNA relies on anterograde and retrograde signaling to maintain mitochondrial and cellular homeostasis. Anterograde signaling involves nuclear-encoded genes that regulate mitochondrial functions, influencing processes like mtDNA replication, transcription, and repair in response to cellular energy needs. Conversely, retrograde signaling is initiated by mitochondrial stress or dysfunction, sending signals from the mitochondria back to the nucleus. This communication triggers adaptive responses, including changes in gene expression to address mitochondrial damage or bioenergetic deficits. Together, relaxed and stringent replication modes, coupled with anterograde and retrograde signaling pathways, enable mitochondria to adapt to cellular conditions while preserving mtDNA integrity across generations [120,123].

mtDNA undergoes a variety of mutations, including point mutations, deletions, duplications, and rearrangements. The high mutation rate in mtDNA is largely attributed to low fidelity of mtDNA polymerase (PolG), close proximity to ROS generated by the ETC, lack of histones, and limited DNA repair capabilities. ROS can damage nucleotide bases, leading to base mispairing and ultimately causing mutations if not repaired before replication. Common mutation hotspots in mtDNA include the D-loop, where replication and transcription are initiated, resulting in high exposure to ROS and replication errors. Unlike nuclear DNA, which possesses a range of robust repair mechanisms, mtDNA repair is very limited, relying mainly on base excision repair (BER). In BER, damaged bases are excised and replaced; however, this process is not as efficient in mitochondria as in the nucleus. There is some evidence for mismatch repair and homologous recombination in mtDNA, but these processes are less efficient and not fully understood. Because of the limited repair options, mtDNA mutations accumulate faster than their nuclear counterpart, contributing to aging-related cellular dysfunction and the development of mitochondrial diseases [124,125].

Across generations, the transmission of mtDNA is predominantly maternal, with several key mechanisms controlling for the quality of mtDNA passed to the offspring [120,122]. Key processes in mtDNA inheritance include genetic bottlenecks during oogenesis, mechanisms to prevent paternal mtDNA transmission, and purifying selection. mtDNA is inherited maternally as sperm mitochondria are usually degraded after fertilization. Mechanisms ensuring the degradation of paternal mitochondria include ubiquitin-mediated

tagging for proteasomal degradation and autophagy. Studies in various organisms, including *Caenorhabditis elegans* and mice, have demonstrated that paternal mitochondria are eliminated via mitophagy and proteolytic pathways in the zygote, ensuring that only maternal mtDNA is passed on. This uniparental inheritance helps maintain homoplasmy and avoids potential incompatibilities between nuclear and mtDNA from different lineages. During oogenesis, a dramatic reduction in the mtDNA population occurs, creating a genetic bottleneck that significantly alters heteroplasmy levels across generations. This bottleneck reduces the number of mtDNA molecules passed to each oocyte, leading to a rapid genetic drift in mtDNA populations. As a result, siblings from the same mother can inherit vastly different proportions of mutant versus wild-type mtDNA. In cases where pathogenic mutations are present, offspring may inherit a high mutation load if the bottleneck amplifies the mutant mtDNA variant. Studies indicate that a purifying selection mechanism may be present during oogenesis, which eliminates oocytes with a high burden of deleterious mtDNA mutations. However, some mutations still bypass this selection, explaining why certain mitochondrial diseases persist across generations[123]. The inheritance of mtDNA can also be influenced by environmental factors that affect mitochondrial dynamics and turnover.

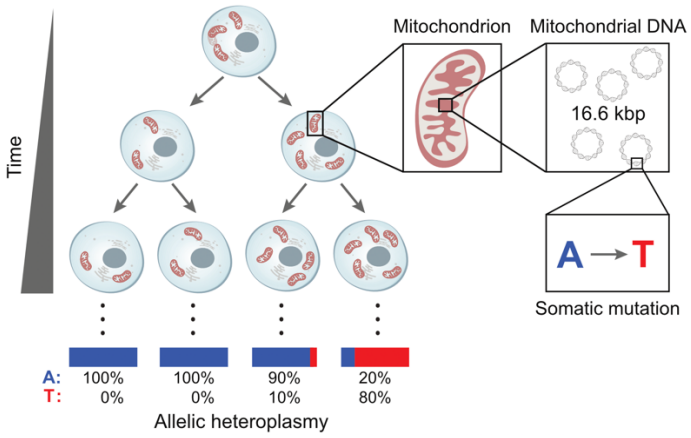


***Fig. 4 Mitochondrial heteroplasmy across cell divisions**. Readapted from Ludwig et al.* [125]

Mitochondria are highly dynamic organelles, undergoing continuous structural and genetic reshaping between cellular division events. These processes include replication, fusion, fission, degradation, and selective removal through mitophagy. Collectively, these dynamics are crucial for managing mitochondrial DNA (mtDNA) quality, adapting to metabolic demands, and maintaining cellular health. mtDNA “relaxed replication,” produce high number of copies of mtDNA per cell, buffering against the potential impact of mutations, regulating the relative ratio of mutant vs wild-type mtDNA copies, i.e., the allelic

frequency of mtDNA mutations. Fusion of mitochondrial membranes allows mitochondria to share their contents, creating a more homogeneous mitochondrial network and preventing the isolation of dysfunctional mitochondria. In contrast, fission, allows the segregation of defective mtDNA molecules, effectively targeting them for degradation. Mitochondrial fusion and fission are coordinated with the cell cycle. For example, during the G1-to-S phase transition, fusion is upregulated, allowing mitochondria to form elongated, interconnected networks. As cells progress into mitosis, fission increases, generating numerous discrete mitochondrial units that can be evenly distributed between daughter cells. This regulation ensures each new cell inherits a sufficient number of mitochondria and helps prevent the unequal distribution of mutated mtDNA. Mitophagy is a selective autophagy process that degrades dysfunctional mitochondria, serving as a crucial quality control mechanism. When mitochondria experience a loss of membrane potential or other signs of damage, they are tagged for degradation by mitophagy-related proteins, primarily through the PINK1-Parkin pathway. The efficiency of mitophagy varies across cell types and can decline with age, leading to an increased accumulation of damaged mtDNA in older cells.

Heteroplasmy, the coexistence of wild-type and mutant mtDNA alleles within a single cell, is central to mitochondrial genetics. mtDNA heteroplasmy can shift over time due to random genetic drift, selective pressures, and cellular processes that either amplify or reduce specific variants. Random genetic drift, or "mtDNA drift," refers to the stochastic fluctuation of mtDNA variant allelic frequencies within a cell. This drift can be particularly evident at low number of mtDNA molecules, as small numbers increase stochasticity of mtDNA dynamics. Over time, mtDNA drift can lead to the loss of a mtDNA variant, increased heteroplasmy, or even fixation[126].

mtDNA allelic frequencies are influenced by selective pressures both at the intra-cellular and at the population level. Within individual cells, individual mtDNA molecules compete for access to the cell's limited replication machinery. In general, wild-type mtDNA molecules with intact replication origins and promoter regions tend to replicate more efficiently. Mutated mtDNA molecules, particularly those with deletions or point mutations that affect replication origins or other regulatory sequences, generally have lower replication efficiency, and are be recognized by the cell mitochondrial quality control system (i.e., mitophagy). However, there are instances where mutated mtDNA molecules may possess replicative advantages. Some mutations, especially large deletions, may remove non-essential regions of mtDNA resulting in shorter mtDNA molecules that can replicate faster than full-length mtDNA. In some cases, mutations mtDNA control regions

of mtDNA can increase the replication rate of specific mtDNA molecules. At the level of cell populations, cells with high mutation loads in their mtDNA may be selectively removed through mechanisms like apoptosis or senescence, ensuring that cell populations retain an overall functional mtDNA pool. However, under certain conditions, such as metabolic stress or hypoxia, specific mtDNA mutations may confer a selective advantage, allowing these mutations to expand within the cell population[123].

Human mitochondrial genetics studies have historically focused on mitochondriopathies. However, advances in mtDNA genotype determination (also referred to as variant calling), and the refinement of quality control analyses, including in whole-genome sequencing (WGS) data, have accelerated large-population genetic studies. These studies have established links between genetic variations within the mitochondrial genome and complex human phenotypes . For example, by determining genotype–phenotype associations from biobank-scale genome-wide association studies in cohorts such as the UK Biobank, recent studies have linked the contribution of mtDNA variation to common disease and human traits, including haplotype-defining variants with large effect sizes, such as variants that affect human height and are mediated by variation in ATP synthesis. In addition to analyzing germline variation in these biobank studies, the utility of mitochondrial genetics has emerged in other settings. Specifically, cancer genome sequencing has provided a rich resource of mtDNA genotype associations with patient phenotypes. For example, pathogenic mtDNA mutations are associated with increases in overall survival in colorectal cancer and broadly appear to modulate transcriptional programs[127].

Mitochondrial variants as natural lineage markers

The properties of mtDNA made it a fundamental tool to study species evolution[128]. Maternal inheritance, high mutation rate, and lack of recombination[129,130] make mtDNA and its variants effective lineage markers[122]. Specifically, maternal inheritance creates a direct lineage record that avoids the complexity of nuclear recombination, while high mutation rates generate distinct haplogroups that can be retrospectively identified even over short timescales.

Different mathematical have been used to quantifying mtDNA mutation rates and evolutionary dynamics. A foundational model is the coalescent. The coalescent models mtDNA ancestry within populations, allowing for the estimation of mutation rates and divergence times based on genetic drift and mutation events[131]. This model allowed researchers to reconstruct mtDNA lineage histories by simulating the ancestral

relationships among individuals in a sample, providing insight into genetic structure and variation within populations. Other important frameworks are the infinite sites model, which assumes that each mutation occurs at a unique site and a single time in the history of an evolving population. When applied alongside the molecular clock hypothesis, which assumes a constant mutation rate over time, the infinite sites model can estimate divergence times with high resolution[132]. These models were critical in dating evolutionary events and reconstructing phylogenies. Selection and genetic drift were also incorporated stochastic simulations, providing a dynamic view of mtDNA evolution under various evolutionary scenarios [133].

In phylogenetics, mtDNA variation was used to build species trees. Techniques such as maximum likelihood**,** Bayesian inference, and Neighbor-Joining are commonly applied to mtDNA data. These methods leverage the high mutation rate of mtDNA to capture recent evolutionary relationships at a fine scale[134]. Maximum likelihood methods calculate the probability of observing a given set of mtDNA sequences given a specific tree structure, while Bayesian inference estimates the likelihood of tree structures based on prior probabilities and observed data. Both methods have been widely used in constructing phylogenies for closely related species[135]. Thus, the study of mtDNA variants has significantly advanced our understanding of species evolution by providing a tool for tracing lineage history, documenting genetic drift, and identifying adaptive responses to environmental changes.

Since the seminal paper from Ludwig and colleagues in 2019[125], mtDNA variants (hereafter referred to as MT-SNVs, for simplicity) have been used to trace cell lineages in human cell populations. This dramatic shift, from species to somatic evolution, was largely due to the advent of single-cell technologies, providing unprecedented sensibility in the detection of MT-SNVs from either mitochondrial DNA or RNA. As previously discussed, (see single-cell multi-omics) single-cell biology have dramatically reshaped the way we investigate cellular and molecular processes in health and disease. However, efforts to capture genetic variation are relatively underdeveloped, compared to other phenotypic profiling assays. scDNA-sequencing, and in particular single-cell WGS (scWGS), has been successfully leveraged to identify somatic variants and construct phylogenetic trees of developmental processes. However, scWGS remains limited by high-costs and technical challenges in confident detection of single-nucleotide variants. On the contrary, most scRNA-seq and scATAC-seq techniques inherently cover large proportions of the mitochondrial genome as a ‘by-product’ [137].

As >90% of mtDNA is transcribed, full-length RNA-seq techniques (for example, Smart-seq), capturing the entire sequence of MT-stranscripts are particularly attractive for MT-SNVs detection, as demonstrated in [125]. However, this plate-based methods are limited by their cost and throughput. On the contrary, scalable droplet base scRNA-seq protocols only sequence 3′- or 5′ transcript ends, and thus, cannot yield the necessary MT-coverage. Primer-based tiling of mitochondrial transcripts (i.e., the 'MAESTER'[138] protocol, coupled with *maegatk*, its companion pipeline for MT-SNVs calling), has demonstrated successful detection of MT-SNVs from these peculiar cDNA libraries. Importantly, this protocol starts from cDNA from the popular 10x commercial protocol, and therefore can be applied retrospectively without the need for a full scRNA-seq library preparation, starting from fresh cell suspensions. Moreover, the standard gene expression library from 10x scRNA-seq contains precious cell state information. In spite of these advancements, these, RNA-based methods have limitations, including: lack of MT-coverage for un-transcribed regions, MT-coverage expression and strand biases, and difficulties to discern bona fide variants from transcriptional errors and/or RNA-editing events.

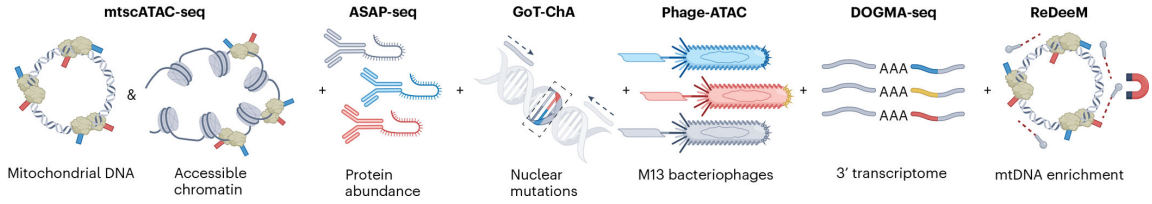


Fig 5. Single-cell landscape of MT-SNVs profiling. *Readapted Nitsch et al., [136]*

On the other side, Tn5-based transposition of mtDNA as used in mitochondrial scATAC-seq (mtscATAC-seq[139]) achieves similarly scalable but much more uniform MT-coverage, and despite not providing transcriptional information, accessible chromatin profiles may inform about cell states (that is, accessible chromatin profiling). Here, the main limitation lies in the absence of UMIs to de-duplicate sequence information from individual mtDNA molecules (thus enable UMI-based error-correction strategies, as in MAESTER).

Sophisticated multi-omic variants of these latter protocols have been developed (e.g., ASAP-seq[69], PHAGE-ATAC[140], GoT-ChA[141], DOGMA-seq[69]) yielding up to four single-cell data modalities, including, transcriptome, accessible chromatin, surface markers, targeted nuclear SNVs and mtDNA mutations. Among these, regulatory multi-omics with deep mitochondrial mutation profiling ('ReDeeM' protocol[142]) represents the latest advancement in phenotypic-informed MT-SNVs detection protocols. Building from the commercial 10x

Multiome kit, Weng and colleagues added a ('MAESTER-like') target enrichment step for mtDNA (yielding higher mtDNA coverage than previous ATAC-based protocols) coupled by introduction of "endogenous UMIs" (eUMIs), i.e., endogenous molecular barcodes that can be spotted in ATAC reads by considering unique Tn5 transposition sites. These innovations brought together benefits from MAESTER (i.e., target enrichment, UMI-based error-correction) and mtscATAC-seq (i.e., uniform and unbiased mtDNA coverage), coupling it to the multi-omic definition of cell state (i.e., gene expression + chromatin accessibility) from 10x Multiome. Together, these technologies enabled cost-effective and phenotypic informed retrospective scLT in primary human tissue samples.

In this scenario, there are still lots of unknowns regarding MT-SNVs-based scLT[137,143].

Considering retrospective scLT (i.e., endogenous barcodes), it has been estimated that, in spite of lower mutation rates, the vast character space of the nuclear genome is able to generate SNVs at virtually all cell divisions[14,143]. Thus, even if in principle WGS data can be used solve deep cell phylogenies, previously discussed scWGS challenges (i.e., extremely high costs and allelic drop-out) still constitute a severe limitation in practice. With a little side step, recent works leveraged shallow WGS from single-cell-derived colonies (e.g., tens-to-hundreds colonies) to build deeply resolved somatic cell phylogenies[143,144]. Importantly, profiling colonies (instead of single-cells) mitigated allelic drop-out. However, this procedure still requires remarkable costs and time-consuming labor, to achieve arguably low throughput. Moreover, WGS *per se* does not provide any phenotypic information. Thus cell (colony, in this case) state information has to be retrieved from either non-disruptive (but less informative) technologies (e.g., flow cytometry) or orthogonal assays from aliquots of the same samples, which complicates analyses and interpretation of results.

Considering prospective lineage tracing, instead (i.e., exogenous barcodes), it has been shown that evolving lineage recorders (e.g., Cas9-based) may produce high-resolution phylogenies[146,147], but this resolution greatly depend on the editing frequency of the base editor, and the growth dynamics of the cellular population under investigation. Despite seminal works establishing elegant state-fate associations in developmental and tumor biology, these methods can be applied only to model organisms, as previously discussed.

Considering MT-SNVs-based scLT, the resolution limit of MT-SNVs-based cell phylogenies (hereafter, MT-phylogenies) is still unclear, and very actively debated. In principle, high mtDNA mutational rate, high coverage, and high number of mtDNA copies, suggest feasibility of high-resolution phylogenetic reconstruction from MT-SNVs. However,

stochastic processes altering MT-SNVs allelic frequencies within cell divisions, non-deterministic inheritance of mtDNA copies at mitosis, and limited sensibility of current sequencing protocols, constitute significant challenges.

This works contributes to assess this unkown resolution limit, focusing on expressed MT-SNVs (i.e., MAESTER protocol).

Materials and Methods

MiTo benchmark sample preparation

We generated data from 3 high-quality biological specimens, all derived from the MDAMB231 Breast Cancer cell line: i) a mixture of single-cell derived colonies generated *in vitro* (MDA_clones sample), and a ii) pair of matched primary tumor- (PT) lung metastasis samples grown in vivo (MDA_PT and MDA_lung samples, respectively).

Plasmids:

The Perturb-seq[71] GBC library (pBA57117,18) was purchased from Addgene and used for lineage tracing experiments. This vector contains a random 18-nt guide barcode (GBC) between the blue fluorescent protein (TagBFP) and polyadenylation signal sequences. This vector contains puromycin and ampicillin resistance genes and the reporter gene TagBFP constitutively expressed under the control of EF1a promoter. The pLenti CMV Puro LUC (w168–1) was purchased from Addgene. This vector also contains puromycin and ampicillin resistance genes.

Cell lines:

All cells were cultured in adhesion in a 20% O2, 5% CO2 incubator at 37º C. HEK293T and the metastatic human TNBC cell line MDA-MB-231 were purchased from the ATCC and cultured in DMEM (EuroClone), supplemented with 10% South American FBS, 2 mmol/L L-glutamine, and 100 U/mL penicillin–streptomycin. All the cell lines were tested for Mycoplasma contamination routinely. All the cell lines were split once they reached approximately 80% confluence and cultured in vitro for no more than 10 passages after thawing. Puromycin selection for GBC+ cells was given 2 µg/mL. Transient transfection with Lipofectamine TM was performed uniquely to transfect the Perturb-seq GBC library in HEK293T.

Murine models:

Female NOD/SCID Il2-Rg null (NSG) mice were purchased from Charles River Laboratory and housed under pathogen-free conditions at 22º C±2º C, 55%±10% relative humidity, and with 12 hours d/light cycles in mouse facilities at the European Institute of Oncology–Italian Foundation for Cancer Research Institute of Molecular Oncology (Milan, Italy) campus. In vivo studies were performed after approval from our fully authorized animal facility and our institutional welfare committee and notification of the experiments to the Ministry of Health (as required by the Italian Law (D.L.vo 26/14 and following amendments); IACUC numbers: 833/2018, 679/2020), in accordance with EU directive 2010/63.

In vivo scLT experiment:

MDA-MB-231 cells (3 x 10^6 /plate) were infected with the both the Perturb-seq GBC library at MOI = 0.3, puromycin selected (2.5 mg/mL for 72 hours), and cultured for 72 hours. For each mouse, 200,000 MDA-MB- 231 cells were resuspended 1:1 in 30-µL PBS and growth-factor reduced Matrigel, and injected in the ninth mammary gland of female NSG mice. Organ infiltration was monitored by IVIS-Lumina and mice sacrificed 21 days after PT resection in the control group. Chemotherapy *in vivo* consisted of adriamycin (A, doxorubicin, 1 mg/kg) and cyclophosphamide (50 mg/kg). Chemotherapy was administered every week, for three cycles in neoadjuvant setting, and for one cycle in adjuvant setting. Mice that received both neoadjuvant and adjuvant treatment were the "double-treated", while the "adjuvant-treated" just received the adjuvant chemotherapy. Primary tumor was monitored via caliper measurement three times a week.

MDA-MB-231 clonal mixture (MDA_clones):

Single-cells were isolated into 96-well plates by limiting dilution and expanded for ~30 days. The resulting cellular colonies were then infected with unique barcodes (i.e., 8 distinct colonies were infected with 8 distinct barcodes) Infected cells were selected with puromycin for three days. Barcoded clones were mixed at known ratios, FACS-sorted for the blue-fluorescent protein (BFP) expression (the BFP coding sequence is present in the GBC construct, and therefore marks lentivirally barcoded cells), and subjected to library preparation and sequencing.

In vivo longitudinal PT-lung couple (MDA_PT and MDA_lung):

For this work, we selected a PT-lung metastasis couple from the scLT experiment described above. Specifically, we selected a treatment naïve PT-lung couple. Both lesions were single-cell dissociated, FACS-sorted for the blue-fluorescent protein (BFP) expression, and subjected to library preparation.

MiTo benchmark library preparation and sequencing

FACS-sorted cells were counted with vital count by mixing cell suspension and erythrosin B in 1:1 ratio, cells were then pelleted at 2000 rpm for 5 minutes at 4 degrees, resuspended in a volume that allows ~1000 cells/µL, and finally counted again. Cell suspensions (~5000-6000 cells per sample) were submitted to the 10X Chromium, following the conventional protocol for cDNA production and gene expression (GEX) library preparation (10x v3 kit). cDNA and the resulting GEX libraries were quality-controlled through BioAnalyzer (Agilent).

Perturb-seq Bulk DNA sequencing:

Bulk sequencing of DNA-integrated GBC was performed. 200 ng of genomic DNA from approximately 10^6 cells per sample (mixed with spike-in controls) were PCR-amplified using specific primers and sequenced on NovaSeq 6000 (30 million reads/sample).

Perturb-seq sub-library preparation:

Starting from the GEX libraries obtained for each sample, we used a 20 ng aliquot to generate Perturb-seq lentiviral barcodes sub-libraries. A semi-nested PCR approach was used to ensure maximum yield. This PCR employs two specific primers (targeting the adapter sequences inserted during the GEX library production, one couple for each sample), and one reverse non-specific primer (annealing to a constant portion of the BFP sequence), which is constant in all barcode fragments. Presence of the specific ~400 bp fragment was assessed for each sample using 2% agarose (Canvax Biotech) gel electrophoresis. The PCR product was diluted 1:1000 to reduce the amount of other non-specific fragments. Then, we performed a 2nd reaction with the setting and primers reported in Suppl. Table 1. The presence of the diagnostic ~400bp product was checked for every sample using 2% agarose (Canvax Biotech) gel electrophoresis. PCR products were purified with QIAquick PCR purification kit (Qiagen) and 2ng per sample sequenced on NovaSeq 6000 Sequencing System (Illumina), with a depth of sequencing equal to 30 million reads/sample. See Roda, Cossa et al., 2023 for the primer sequences and the PCR cycles.

MAESTER sub-library preparation:

From the 10X cDNA production, we generated the mitochondrial transcripts (MT-) sub-library following the MAESTER[138] protocol. Also in this case, we employed a semi-nested PCR approach, by using specific P5 and P7 adapter-annealing primers, together with MAESTER primer mixes. Briefly, 20 ng of cDNA were used for each of the reactions with the 12 MAESTER primer mixes (each primer mix targets different regions of the mitochondrial genome, thus allowing a full coverage), thus meaning 240 ng of cDNA for each sample.

Set up the PCR reactions (12 per sample +1 negative control) in a 96-well plate as in follows:

Reagents	**Volume (ul)**	**concentration**
cDNA + H20	15	20ng
Primer P5	1	10uM
Mix Primer mitochondrial	4	1uM
KAPA Hifi	20	

Tot volume	40	

Reaction	Temperature	Time	Cycles
Initial denaturation	95 °C	3min	X 1
	98 °C	20 sec	X 6
	65 °C	15 sec	
	72 °C	3min	
Final extension	72 °C	5min	X 1

We then performed PCR purification with the AMPure Bead kit, with a 0.8X ratio, in order to capture the amplicons of interest.

2) PCR 2

Reagents	Volume (ul)	concentration
cDNA amplified (PCR1)	18	
Primer P5	1	5uM
Primer P7	4	5uM
KAPA Hifi	20	
Tot volume	40	

Reaction	Temperature	Time	Cycles
Initial denaturation	95 °C	3min	X 1
	98 °C	20 sec	X 6
	60 °C	30 sec	
	72 °C	3min	
Final extension	72 °C	5min	X 1

MAESTER samples were then quality-controlled by BioAnalyzer and submitted to sequencing, by using the Illumina flow cell, requesting ~250 million reads per sample.

Data pre-processing

We developed the *mito_preprocessing* Nextflow pipeline, which implements workflows for preprocessing any combinations of standard 10x scRNA-seq (i.e., GEX) plus lentiviral

barcoding (GBC) and/or MAESTER (MT) data. *mito_preprocessing* pipeline includes the following processes:

- Alignment of GEX, GBC and MT-libraries reads: *STARSolo*[148]
  Paired end .fastq files from all libraries are aligned to a custom GRCh38 genome reference. The hg38 reference genome downloaded from Cell Ranger Downloads page is complemented with a custom sequence to retrieve GBC-containing reads (a constant sequence of the transcribed Perturb-seq vector, with two constant flanking a stretch of 18 Ns, the 18bp lentiviral barcode). Additionally, a blacklist of potentially confounding nuclear mitochondrial DNA segments (NUMTs) sites is masked from this custom reference genome, as described in [138] and [149]. From this reference, *STAR* is used to generate a custom index for alignment of all libraries (GEX, GBC, MT).

- Cellular barcodes (i.e., CBs) and Unique Molecular Identifiers (i.e., UMIs) correction:
  GEX, GBC and lentiviral reads are separately aligned with *STARSolo,* enabling *EmptyDrops*[150] CB correction and UMI deduplication, given the following command line arguments:

  - *--soloType: CB_UMI_Simple*
  - *–soloCBmatchwhitelist: 1MM_multi_Nbase_pseudocounts;*
  - *–soloCellFilter: Empty_Drops_CR*
  - *--soloUMIdedup: 1MM_CR*

  Alignment and counting of the GEX library by *STARSolo* provides a list of “putatively good” CBs (from *STARSolo* cell calling algorithm) that is used as input for GBC and MT-reads pre-processing after alignment (see below). The 10x v3 CBs whitelist is been used as a reference for STARSolo CBs correction.

- GBC-containing UMIs consensus sequence generation: *samtools*[151], *fgbio*[152], *bwa-mem*[153]
  In order to get accurate GBC species and molecule counts, we introduced consensus sequence generation with *fgbio.* Specifically, reads aligned to the lentiviral cassette are splitted by CB (“putatively good” CBs from *STARSolo* cell calling algorithm) and grouped by UMI tag (and query alignment starting point) with fgbio GroupReadsByUmi. *fgbio CallMolecularConsensusReads* is used to

generate consensus sequences for each read group, considering bases with *--min-input-base-quality 30 and –min_reads 10.* Consensus sequences are realigned to the lentiviral cassette, and *fgbio FilterConsensusReads* is used with the following parameters:

- *--min-reads: 10*
- *–min-base-quality: 30*
- *–max-base-error-rate: 0.2*

to filter good quality consensus sequences (i.e., all bases with consensus error <0.2), masking (i.e., replacing with "N"s) consensus bases with quality <30 and/or consensus depth <10. From these filtered consensus sequences, CB-UMI-GBC triplets are extracted (i.e., given the costant "anchor" sequence TAGCAAACTGGGGCACAAGCTTAATTAAGAATT, each expressed lentiviral barcode consists of the first 18 random nucleotides after the anchor sequence), gathered across cells, and used for clonal assignment of individual cells.

- Clonal assignment: custom script.
  Different strategies have been implemented to assign individual cells to lentiviral clones[71,101]. Our current workflow entails:
  - GBC count and (further) correction with an external reference. Unique CB-UMI-GBC triplets are counted, and GBCs species supported by a single consensus UMI are discarded. For remaining triplets, GBCs are corrected further using an external reference. For MDA_clones, each GBC is mapped to a whitelist of 8 known GBC sequences, while for MDA_PT and MDA_lung the reference consists of (error-corrected) GBCs from bulk DNA sequencing. In both cases, a (single-cell detected) GBC is mapped to a reference GBC if their pairwise hamming distance is <=3, and discarded otherwise. After this last correction, UMIs supporting CB-GBC combinations are counted again.
  - CB-GBC combination filtering, similar to [71]. Specifically, for each CB-GBC combinations two metrics are calculated: a) the relative_abundance, i.e., the number of UMIs supporting a given CB-GBC normalized by the total number of UMIs of the cell, and b) the max_ratio, i.e., the ratio between a given CB-GBC UMI counts and the UMI counts of the most abundant CB-GBC of the cell. CB-GBC combinations with >=0.75 max_ratio and

relative_abundance are filtered and used for final clone assignment. After these procedures, the vast majority of CB possesses a single CB-GBC combination (as expected both for the *in vitro* experiment, where each single-cell derived colony has been infected with only one GBC species, and for the *in vivo* experiment, where low MOI infection should guarantee minimal occurrence of multiple infection events). CBs not mapped to a single, well supported GBC are discarded from further analysis.

- MT-genome aligned reads retrieval: *samtools, picard*[154]
  After alignment, reads aligned to the MT-genome are filtered from both GEX and MT library, and merged into a single-bam file with CB and UB tags. Reads with CBs from *STARSolo* cell calling are filtered with *picard FilterSamReads* and splitted into cell-specific bams (*samtools split*) for consensus sequence generation.

- MT-UMIs consensus sequence generation: *samtools*, *fgbio*.
  In order to get accurate basecalls, cell specific MT-reads are grouped by UMI tag (and query alignment starting point) with fgbio GroupReadsByUmi, as described for GBC-containing reads. Then, *fgbio CallMolecularConsensusReads* is used to generate consensus sequences for each read group with *--min-input-base-quality 30 and –min_reads 3.*

- Re-alignment of MT-UMIs consensus sequence to MT-genome reference: *bwa-mem*

- MT-basecall pileup and Allele Frequency Matrix (AFM) generation: *custom scripts.*
  Consensus sequences realigned to the MT-genome are parsed to filter single consensus bases and record their statistics (i.e., average quality, consensus score, and depth, both in forward and reverse orientation). Throughout this work, MT-SNVs from the *mito_preprocessing* pipeline come from consensus sequences of >=30 alignment quality, >=0.75 consensus score, >=30 base quality and >=3 read depth (i.e., UMI group size). Finally, these tables are gathered across cells and parsed to obtain an Allele Frequency Matrix (AFM), i.e., an *AnnData*[155] object, an annotated matrix storing (raw, unfiltered) cell x variant relevant information (i.e., allelic frequency, AD, DP, quality and site_coverage). Cell and variant coverage statistics are recorded in cell and variant meta data, respectively.

mito_preprocessing takes input .fastqs from 10x GEX, lentiviral (GBC) and MAESTER (MT-) library sequencing (plus a configuration file, in .json format), and depending on the chosen entry-point, returns one (or all) of the following main outputs: i) a CB-GBC UMI counts table and a cell-clone assignment table; ii) CB-GEX UMI counts matrix from GEX data processing; and iii) MT-pileup tables, and AFM. These inputs are subjected to further Quality Control (custom scripts, see below) to retain only cells with good GEX and MT-library quality, and robustly assigned to a single GBC.

To enable flexible benchmaking of different preprocessing tools, *mito_preprocessing* provides the BENCH entrypoint. This workflow takes a .bam file with MT-reads and a list of cell barcodes of interest for each sample, and implements *cellsnp-lite*[155], *samtools*, *freebayes*[157] and *maegatk*[138] (default parameters*)*. Outputs from these tools are used to build Allele Frequency Matrices. Of note, while *maegatk* (the original oneSample_maegatk.py, script, with default parameters values taken from *maegatk* command line, except for –min-reads 3, instead of 1) outputs a very similar output format compared to *mito_preprocessing* (except for additional statistics recorded only by *mito_preprocessing*, i.e., average basecall consensus score and UMI groups size), *cellnsp-lite*, *samtools* and *freebayes* do not provide complete coverage and pileup information across all MT-genome sites. These latter tools perform variant selection and pileup from raw MT-reads alignment, yielding a filtered MT-SNV callset with only AD and DP (reads or UMI) counts. Thus, AFMs from these tools can be different from the one os *mito_preprocessing*/*maegatk* not only for detected MT-SNVs, but also for the information that is attached to each MT-SNV (i.e., not just AD and DP layers, but also quality, site coverage, group size, … etc.).

Quality Control

GEX library: Standard cell Quality Control (QC) was performed to remove residual empty droplets, not viable cells and/or cell doublets. Cell doublets were filtered out using *scroublet*. Then, we excluded CBs with <500 transcripts, <250 genes, at >3 median absolute deviations (MADs) above the median of expressed transcripts and genes, and with >15%of their total UMI counts from MT-genes.

GBC library: We included all CBs assigned to a single GBC, as previously described[158].

MT-library: the *filter_cells* function from *mito_utils* provides several cell filters. Throughout this study, we used 'filter2', which selects cells with median target site coverage >25 and % 75 of MAESTER target site coverage (i.e., n consensus UMIs>0). Of note, since this filter requires cell coverage information across nearly all MT-genome, it can be applied only to *mito_preprocessing* and *maegatk* inputs, for which this information is available.

CBs passing both GEX and MT- QC and assigned to a single lentiviral clone are retained for subsequent analysis.

mito_preprocessing-maegatk comparison

Across this work, the starting point for mutation selection is the total possible pool of deviations from the MT-genome revised Cambridge Reference Sequence (rCRS, [159]) (n total characters = 3 x length rRCS sequence, 16569 bp: 49707 total characters, hereafter defined as the set of all possible MT-SNVs). Both *mito_preprocessing* and *maegatk* collect information about all MT-sites, and all candidate MT-SNV basecalls are filtered in subsequent stages. Since the oneSample_maegatk.py script (implemented as pre-processing option within *mito_preprocessing*) does not collect the average UMI group size and consensus error per se, to compare *mito_preprocessing* and *maegatk* basecalls we verified that by setting the same parameters (i.e., the ones controlling UMI consensus sequence generation and consensus base filtering), *mito_preprocessing* is able to reproduce nearly identical variant basecalls compared to the original *maegatk* script. Having verified that, we compared *mito_preprocessing* (default parameters) with *mito_preprocessing* with *maegatk*-like parameters. Specifically, we used:

- *--min-reads*=3 (instead of the *maegatk* default=1, as in [138])
- *--base-qual*=0 (*maegatk* default)
- *--alignment-quality*=0 (*maegatk* default)

from maegatk command line interface (CLI), translated as *mito_preprocessing* arguments:

- *--fgbio_min_reads_mito*=3 (*mito_preprocessing* default)
- *--fgbio_base_quality*=0 (instead of the *mito_preprocessing* default=30)
- *--fgbio_min_alignment_quality*=0 (instead of the *mito_preprocessing* default=0)

with the addition of:

- *--fgbio_base_error_rate_mito*=1 (i.e., no filtering on consensus score)

These maegatk-like parameters are compared to *mito_preprocessing* defaults:

- *--fgbio_min_reads_mito*=3

- *--fgbio_base_quality=3*0
- *--fgbio_min_alignment_quality*=30
- *--fgbio_base_error_rate_mito*=0.25

Statistics in Fig. 10 are computed from "raw" basecalls (i.e., no additional filtering of MT-genome site, ecc), to quantify differences in cellular coverage and in the average number of candidate MT-SNVs per cell. All MT-reads form CBs passing QC checks for GEX, MT and GBC-modalities from the MDA_clones sample were considered for this experiment. This choice was made to compare different pipelines "raw" outputs, given very similar pre-processing tasks, but different choices for the "default" options that govern their behavior. It has to be acknowledged that: i) *mito_preprocessing* main functionalities were borrowed directly from *maegatk*, ii), despite not considering consensus scores for filtering individual bases and using all bases for consensus generation, *maegatk* allows remarkable flexibility in other base filtering parameters, and iii) these two pre-processing tools interface with different final objects to store pre-processing output: *maegatk* outputs are stored as either plain text tables or in a *SummarizedExperiment* R class, while *mito_preprocessing* uses plain text tables or and the *AnnData* python class. Thus, the main difference between *mito_preprocessing* and *maegatk* lies in the different way through which very similar functionalities are assembled into a pipeline with certain inputs and outputs. *mito_preprocessing* takes as input either: i) GEX and MT (optionally GBC) library *.fastqs*, handling pre-processing of all sequencing libraries in a single call, or ii) a *.bam* file storing MT-aligned reads, and a list of cell barcodes of interests. *maegatk* only performs MT-reads consensus generation and pileup, and therefore, MT-reads alignment and GEX library pre-processing are external to the tool. Also, *mito_preprocessing* allows multi-sample inputs, and has unique functionalities for lentiviral barcoding data pre-processing that are not shared with *maegatk.*

MT-SNVs selection

Raw AFM matrices (see Pre-processing and Quality Control section), from the *mito_preprocessing* pipeline contains typically thousands of MT-SNVs per cell that need to be filtered to obtain an "informative", denoised MT-SNVs space. The *filter_afm* function from *mito_utils* provides surgical MT-SNVs annotation and selection capabilities. *filter_afm*: i) annotate MT-SNVs summary statistics as in [138] and [142]; ii) apply a very loose, "baseline" MT-SNVs filter (i.e., mean site coverage>=5, mean quality>=30, and n cells with non-zero AF >=2); iii) apply one of either 4 previously published MT-SNVs filters (i.e., Coefficient of Variation, CV[139], miller2022[138], weng2024 [142], MQuad [159]), the *MiTo* filter

(which re-adapt and refine filtering criteria from both miller2022 and weng2024); iv) Filter common dbSNP variant and annotate RNA-editing events as done in [161]; v) Call cell MT-genotypes (i.e., employs one between *vanilla* and *MiTo* binarization strategy to assign binary MUT/WT genotype status at each cell-variant combination); vi) Filter cells positive for at least one MT-SNV. vi) Compute (optional) additional MT-SNVs space metrics (e.g., MT-SNVs lineage bias, "MT-SNVs quality", "Connectedness" and "Variation" metrics); vii) Reconstruct (optional) the cell phylogeny of selected cell x variant character matrix (see *build_tree,* below).

filter_afm is used to filter all MT-SNVs subsets characterized through this work, either interactively or automatically via the *phylo_inference* pipeline. For *MiTo benchmarking*, AFMs from *samtools* and *freebayes* were minimally filtered (given the few number of MT-SNVs, and the lack of MT-SNVs quality and depth statistics); for cellsnp-lite AFMs (we processed both the entire *cellsnp-lite* callset or its filtered version, using the *MQuad* filter as recommended by the authors); *mito_preprocessing* and *maegatk* AFMs were filtered with the *MiTo* filter, which select variants according to:

- Mean variant quality >= min_var_quality
- Fraction of negative cells >= min_frac_negative
- n of +cells >= min_n_positive
- AF of confident detection >= af_confident_detection
- Mean number of UMIs supporting the ALT allele in +cells >= min_mean_AD_in_positives
- Mean number of total UMIs on the variant site >= min_mean_DP_in_positives

For the data presented in Chapter *"Phylogenetic signal in expressed MT-SNVs spaces: the accuracy-cellular yield trade-off",* we tested all filtering combinations from the following hyper-parameter space:

- min_var_quality: 30
- min_frac_negative: 0.2
- min_n_positive: [2,5]
- af_confident_detection: [0.01,0.02,0.03,0.05,0.07,0.1]
- min_n_confidently_detected: [2,3]
- min_mean_AD_in_positives: [1,1.25,1.5]
- min_mean_DP_in_positives: 5

More details about the *phylo_inference* pipeline can be found in section: “Phylogeny inference”.

MT-SNVs genotyping

Assuming diploid status across all cells in a population, the genotype of a nuclear SNVs in a given cell cells can be either homozygous WT (00 state, AF of the mutated allele=0), homozygous MUT (11 state, AF ot the mutated allele=1), or heterozygous (01 state, AF ot the mutated allele=0.5). On the contrary, MT-SNVs display a nearly continuous spectrum of allelic frequencies, that can be binarized with two genotype states: 0 → absence of a MT-SNVs, 1→ presence of the MT-SNVs. It has been previously shown that discrete metrics are more robust to scRNA-seq technical noise than continuous oneswith regards to quantifying pairwise cell-cell genetic (dis-)similarity[142]. However, these former metrics need accurate MT-SNVs genotyping (i.e., AFM binarization). In this work, we implemented two binarization methods: *vanilla* and *MiTo*. For a given MT-SNV site, one can retrieve alternative and total UMI counts observed across the cell population, ad and dp (ad , $dp \in \mathbb{N}^{1xN}$, with $N = n\,cells$). For a given MT-SNV, the *vanilla* method assigns genotype 1 to cell i if ad_i / dp_i > $t_{vanilla}$ and ad_i >= $minAD$, as in [138] and [142]. Throughout this work: $t_{vanilla}$ has been set to 0, while $minAD$ has been set to either 1 or 2. On the other hand, *MiTo genotyping* leverages the statistical modelling of ad and dp counts to assign binary genotypes to each cell. In particular, we re-adapted the probabilistic approach introduced by *MQuad* [159] to MT-SNVs genotyping. To do this, we assumed as in [159] that, ad , dp counts are generated by weighted sampling of two binomial distributions, representing the background (component 0) and the true positive signal (component 1), respectively. Under this assumption, denoting with ad_i anddp_i the alternative and total UMI counts for cell *i* respectively, the probability P of observing exactly ad_i alternative UMI counts for cell i is defined as:

$$P\colon \mathbb{N}^{+} \rightarrow [0,1];\ ad_i \rightarrow f(ad_i \mid dp_i, \Theta)$$

$$P(ad_i \mid dp_i, \Theta) := \sum_{k=0}^{K} \pi_k \cdot Binomial(ad_i \mid dp_i, p_k) \quad \text{Eq. 1}$$

with:

$$Binomial(x \mid n, p) := \binom{n}{x} p^x (1-p)^{n-x}$$

$K = 1$, with $k = 0$ background component and $k = 1$ true positive signal component

$\Theta = [p_k, \pi_k] \in \mathbb{R}^{2(K+1)}$, parameter vector of the model

$p_k \in [0,1]$, success rate for component *k*

$\pi_k \in [0,1]$, mixing weight for component *k*

$$\sum_{k=0}^{1} \pi_k = 1$$

This probabilistic model can be fitted to the observed values of the ad , dp UMI counts by maximizing the total likelihood $L(\Theta \mid ad, dp)$:

$$L : \mathbb{R}^{2(K+1)} \rightarrow \mathbb{R}^{+};$$
$$L := \prod_{i=1}^{N} P(ad_i \mid dp_i, \Theta) \qquad \text{Eq. 2}$$

By taking the logarithm of both sides and unfolding the definition of the binomial distribution probability mass function, Eq. 2 becomes:

$$logL = log(\prod_{i=1}^{N} P(ad_i \mid dp_i, \Theta))$$
$$= \sum_{i=1}^{N}(log \binom{ad_i}{dp_i} + log(\sum_{k=0}^{K} \pi_k \cdot p_k^{ad_i} \cdot (1 - p_k)^{dp_i - ad_i})) \qquad \text{Eq. 3}$$

Maximization of this $logL$ ((equivalent to likelihood maximization given the monotonicity of the logarithm function) can be nicely achieved via Expectation-Maximization (EM), a Bayesian optimization approach that is capable of handling hidden variables through Expectation (E) and Minimization (M) steps. First, at each E-step, Bayes Theorem is used to compute the posterior probabilities of latent (*hidden*) variables $Z_i \in \mathbb{N}^{1xN}$, i.e., the cell memberships to either background ($k = 0$, $Z_i = 0$) or true positive signal ($k = 1$, $Z_i = 1$) components (i.e., cell genotypes). Given the definition of posterior probability of A given B, $P(A \mid B)$, from Bayes Theorem:

$$P(A \mid B) := \frac{P(B \mid A)\, P(A)}{P(B)}$$

Considering:

$$A = P(Z_i = k)$$
$$B = [ad_i, dp_i, \Theta^{\star}]$$
$\Theta^{\star}$ current estimate for Θ

the posterior probability of cell i assignment to the k-th component is defined as:

$$\gamma_{ik} := P(Z_i = k \mid ad_i, dp_i, \Theta^\star) = \frac{P(\, ad_i \mid dp_i, Z_i = k, \Theta^\star)\, P(\, Z_i = k \mid \Theta^\star)}{P(\, ad_i \mid dp_i, \Theta^\star)} \quad \text{Eq. 4}$$

With:

$$P(Z_i = k \mid \Theta^\star) = \pi_k^\star$$
$$P(ad_i \mid dp_i, Z_i = k, \Theta^\star) = Binomial(ad_i \mid dp_i, p_k^\star)$$
$$P(\, ad_i \mid dp_i, \Theta^\star) = \sum_{k=0}^{K} \pi_k \cdot Binomial(ad_i, dp_i, p_k^\star), \text{ from Eq. 1}$$

Eq. 4 becomes:

$$\gamma_{ik} = \frac{\pi_k^\star \cdot Binomial(\, ad_i \mid dp_i, p_k^\star)}{\sum_{k=0}^{1} \pi_k \cdot Binomial(ad_i, dp_i, p_k^\star)} = \frac{\pi_k^\star \cdot p_k^{ad_i} \cdot (1-p_k^\star)^{dp_i - ad_i}}{\sum_{k=0}^{K} \pi_k^\star \cdot p_k^{ad_i} \cdot (1-p_k^\star)^{dp_i - ad_i}} \quad \text{Eq.5}$$

These estimated γ_{ik} are then used to update $\Theta^\star$ in the following M-step. Specifically, Eq. 3 describes the $logL$ of the mixture model, *without* including unknown cell membership, explicitly. *If* these memberships were known (omitting terms not dependent on model parameters Θ) Eq. 3 would simplify from:

$$logL(\Theta)_{complete} = \sum_{i=1}^{N} log(\sum_{k=0}^{K} \pi_k \cdot\, p_k^{ad_i} \cdot (1 - \, p_k)^{dp_i - ad_i})$$

to:

$$logL(\Theta)_{complete} = \sum_{i=1}^{N} \sum_{k=0}^{K} \delta_{ik}\, (log\pi_k + ad_i log p_k + (dp_i - ad_i)(1 - p_k)) \quad \text{Eq. 6}$$

where δ_{ik} is the Kroenecker delta defined as:

$$\delta_{ik} = 1 \text{ if } Z_i = k, \text{ and 0 otherwise}$$

While Z_i (and therefore δ_{ik}) are unknown, the expected values $\gamma_{ik} := E[\, Z_i = k \mid ad_i, dp_i, \Theta^\star]$ are estimated in the E-step. Thus, substitution of δ_{ik} with γ_{ik} in Eq. 6, gives:

$$logL(\Theta)_{complete} = \sum_{i=1}^{N}\sum_{k=0}^{K}\gamma_{ik}\,(log\pi_k + ad_i log p_k + (dp_i - ad_i)(1-p_k)) \quad \text{Eq. 7}$$

$logL(\Theta)_{complete}$ ais not a sum over all N observations and K model components without involving log-sum terms and with nicely separated model parameters. Thus, analytic methods for direct maximization can be readily used to update π_k and p_k separately. New π_k values, π_k^{new}, are computed by solving:

$$argmax\ logL(\pi_k)_{complete} = argmax\ \sum_{i=1}^{N}\sum_{k=0}^{K}\gamma_{ik} log\pi_k + const \quad \text{Eq.8}$$
$$subjected\ to: \sum_{k=0}^{1}\pi_k = 1$$

which gives:

$$\pi_k^{new} = \frac{1}{N}\sum_{i=1}^{N}\gamma_{ik}\text{ , with } k \in \{0,1\} \quad \text{Eq.9}$$

Instead, new p_k values, p_k^{new}, are computed by solving:

$$argmax\ \ logL(p_k)_{complete} =$$
$$= argmax\ \sum_{i=1}^{N}\sum_{k=0}^{K}\gamma_{ik}(ad_i log p_k + (dp_i - ad_i)(1-p_k)) + const \quad \text{Eq.10}$$

which gives:

$$p_k^{new} = \frac{\sum_{i=1}^{N}\gamma_{ik}\cdot ad_i}{\sum_{i=1}^{N}\gamma_{ik}\cdot dp_i}\text{ , with } k \in [0,1] \quad \text{Eq.11}$$

Crucially, the EM steps alternate between estimating posterior probabilities for genotypes Z and updating model parameters until convergence of log-likelihood values is reached. Across this optimization path, Z plays the fundamental role of bridging the gap between the observed data and the structure of the probabilistic model. However, since *MQuad* only uses the likelihood of this mixture model to rank individual MT-SNVs, the cell genotypes Z and their expected values γ_{ik} are "lost" in the fitting process. To genotype individual MT-SNV, the *MiTo genotyping* method fits the same mixture model employed by *MQuad* (implemented in the *MixtureBinomial* class from the *bbmix* package, https://github.com/StatBiomed/BBMix), but then, for each cell, *MiTo* calculates the $Z_{ik} = 0$ and $Z_{ik} = 1$ genotypes posterior probabilities, γ_{i0} and γ_{i1} . For each MT-SNV, genotype 1 is assigned to cells with $\gamma_{i1} > t_{prob}$ and $\gamma_{i0} < 1 - t_{prob}$, while genotype 0 is assigned to all the other cells. By estimating the background signal associated with the observed alternative UMI counts *MiTo* achieves accurate genotyping, especially for challenging *low-*

detection/high-prevalence MT-SNVs. Since the binomial mixture model performs best with relatively high numbers of detection events, the *call_genotypes* function in *mito_utils* allows flexible tuning of t_{prob} and the minimal cell prevalence required for a MT-SNVs variant to be genotyped this way. All the other low-prevalence variants are genotyped with the simpler *vanilla* method. In this work, t_{prob} and *min_cell_prevalence* were set at 0.75 and 0.1, respectively.

MT-SNVs kNN, distances and embeddings

mito_utils include functions to: i) compute cell-cell pairwise distances (*compute_distances)*; ii) perform k-Nearest Neighbors (k-NN) searches (*build_kNN*) and compute kNN-based metrics (i.e., Shannon Entropy, purity and kBET[162] of cellular neighborhoods, with respect to some set of cellular labels, e.g., lentiviral clones); iii) and reduce the dimensionality of a MT-SNVs space (PCA, UMAP[163] and diffusion maps[164] are implemented). *compute_distances* support custom metrics provided by the user. The *draw_embeddings* function allows flexible visualization of cell embeddings in 2D scatter plots.

Phylogeny inference

The *phylo_inference* Nextflow pipeline (https://github.com/andrecossa5/phylo_inference) builds upon *Cassiopeia*[165], the leading python library for scLT tree reconstruction and manipulation. This pipeline support two main entrypoints: *tuning and phylo.*

The *tuning* entrypoint performs efficient exploration of alternative MT-SNVs spaces that can from one or more "raw", unfiltered AFMs (i.e., the main output from *mito_preprocessing* the pipeline). Specifically, a grid of hyper-parameters (specified by the user through a single *.json* configuration file) is used to generate a Nextflow channel of jobs (hyperparameter combinations producing unique MT-SNV spaces, i.e., filtered and genotyped AFMs), marked by unique alpha-numeric IDs. This channel is consumed in massively parallel fashion by a single DSL2 *Nextflow* module, ONESAMPLE, calling the homonimous python script *onesample.py*. With >20 command line options controlling the behaviour of cell and variant filtering, MT-SNV genotyping, cell-cell distances computation and tree building (all processes implemented within the *filter_afm* function from mito_utils), *onesample* records metrics (see MiTo benchmarking metrics and ranking system section) for the evaluation of each tested MT-SNVs space that are stored as separate .pickle objects and a unique .csv file for interactive exploration.

The *phylo* entrypoint, instead, is designed to finalize lineage inference on specific MT-SNV spaces selected after extensive hyper-parameters tuning (i.e., *tuning* entrypoint). This

entry-point is structured in modular subworkflows for AFM preprocessing, tree building, annotation and scoring. Given a set of unique job IDs, matching hyper-parameters options are retrieved to perform the same filtering and genotyping of "raw" AFMs from the *tuning* entrypoint, followed by: i) character matrix bootstrapping; ii) cell-cell distance computation and distance metric scoring (i.e., correlation between cell-cell distances across bootstrap replicates and AUPRC[125], if ground truth lineage labels are provided); iii) tree building, for each bootstrap replicate character matrix (all Cassiopeia tree solvers can be specified, along with *mpboot* and *iqtree*[166]; iv) internal nodes support calculation via *booster*[167], which implements both Felsestein Bootstrap Proportions (FBP) and Transfer Bootstrap Expectation (TBE) methods, the latter providing accurate and fair support evaluation (i.e., does not over-penalize missing identical matching between bootstrapped clades, as FBP does) support evaluation for large scale phylogenies; iv) final tree creation, i.e., creation of the final CassiopeiaTree object storing raw and binarized characters, cell-cell distances, cell metadata and individual tree nodes and branches attributes; v) final tree annotation, i.e., assignment of MT-SNVs to individual tree clades and annotation of these clades into discrete MT-clones with the *MiToTreeAnnotator* (see the MT-phylogenies annotation section); vi) final tree metric scoring (see MiTo benchmark metrics and ranking system section and MT-phylogenies robustness section below).

The final *phylo_inference* outputs are: i) a filtered, genotyped and annotated AFM (*AnnData* object stored .h5ad format); ii) an annotated *CassiopeiaTree;* and iii) a set of tree diagnostic metrics (.csv format). If "lineage_column" is specified within the configuration file, PATH[168] is used to calculate phylogenetic correlations among cell labels attached to each cell in AFM cell metadata (i.e., afm.obs["lineage_column"]). All downstream analyses involving these MT-SNVs spaces and phylogenies can be performed interactively, leveraging available *mito_utils* and *cassiopeia* functionalities (i.e., cell tree visualization). The *AnnData* and *CassiopeiaTree* formats facilitate seamless integration with existing and newly developed single-cell libraries for complex multi-omics analyses in python.

MiTo benchmark metrics, rankings and meta-analysis

To systematically evaluate MT-SNVs space retrieval hyper-parameters, the *tuning* entrypoint (see Phylogeny inference section) records 19 different metrics. These metrics evaluate different properties of filtered AFMs, cell-cell distances, reconstructed trees and inferred (see MT-phylogenies annotation section) MT-clones. Here is the complete description of individual metrics grouped by metric type (with reference to names in :

-Mut Quality:

- n dbSNP: number of filtered (i.e., after baseline and *MiTo* filter) MT-SNVs flagged as common mutation event in the dbSNP database: https://ngdc.cncb.ac.cn/databasecommons/database/id/1622

- n REDIdb: number of filtered MT-SNVs flagged as common RNA-editing event in the REDIdb database: http://srv00.recas.ba.infn.it/redidb/

- Mut signature: transition vs transversion ratio, used to quantify the deviation from the expected MT-mutational signature (i.e., transitions >> transversion)

-GBC:

- Clonal biased MT-SNVs: % of filtered MT-SNVs that is significantly enriched (Fisher's exact test, FDR <= 0.05) within at least one lentiviral clone
- AUPRC: Area Under Precision Recall Curve as described in [125]
- ARI: Adjusted Rand Index, to quantify concordance between GBC labels and inferred MT-clones (see MiTo tree annotator)
- NMI: Normalized Mutual Information score, to quantify concordance between GBC labels and inferred MT-clones (see MiTo tree annotator)

-Tree structure:

- CI: mean Consistency Index of tree characters. Ths measure assessing how well a cofinguration of character is explained by a phylogenetic tree assuming minimal evolutionary changes (i.e., Camin-Sokal parsimony[165])
- Tree- vs char- based distance correlation: Pearson's correlation between tree-based (i.e., minimum number of nodes connecting two leaves) and character-based (i.e., throughout this work jaccard distance between cell-cell MT-SNVs genotypes) cell-cell distances.

-Connectedness:

- Density: % of non-zero entries in the binarized AFM
- Transitivity: transitivity (i.e., clustering coefficient) of the cells shared-MT-SNVs graph
- Mean path length: average cell-cell path-length on the cells shared-MT-SNVs graph

- Average degree: average cell degree across the cells shared-MT-SNVs graph
- LCC: largest connected component of the cells shared-MT-SNVs graph

Variation:

- Haplotype redundancy: % of unique MT-haplotypes (i.e., beared by single-cells) considering all MT-hatplotypes observed in a population of cells
- Median n of MT-SNVs per cell

Yield:

- n GBC clones
- n cells
- n MT-SNVs

To produce hyper-parameters combination rankings, all jobs were grouped according to 5 hyper-parameters of interest: i) pre-processing method, ii) binarization methods, iii) min confident AF, iv) min number of cofident detection events and v) min AD to assign a cell genotype.

Median metric values each combinations were rescaled with min-max normalization. resulting values were averaged across metric type, and weighted sum of these values is used to produce a final "Overall" score. Only n dbSNP and n REDIdb where sign-inverted before min-max normalization. Metric types were chosen to maximize both lineage"'GBC" and "Yield". Accordingly to produce a final ranking, we assigned 0.4 to both "Association with GBC" and "Yield", and 0.1 to both "Tree structure" and "Mutation Quality", to control for potential errors in lentiviral barcoding and variant calling, and weights for other metrics vase set to zero). These scores and rankings were visualized with funkyheatmap[37]. All across the MT-SNV space chapter, these 19 metrics are grouped and presented as average across josn, within 108 unique parameters combinations. Individual jobs were used only to establish relative feature importances. For this latter analysis, we used *lightgbm.LGBMRegressor* (https://lightgbm.readthedocs.io/en/stable/) to regress all tested hyperparameters (i.e., the ones tested with >1 unique value) against metrics selected of interest, and reported estimated feature importances.

MT-phylogenies robustness

Robustness of inferred MT-phylogenies was quantified through different parameters, described in the above sections (i.e., support → bootstrap TBE, CI consistency index ecc.). Specifically (median) internal nodes support was quantified for all caldes, only largest

clades, (>95th percentile n cells per clade), and MT-SNVS-assigned clades. See the “Clonal reconstruction benchmark” to get info abuto the final choice of n=10 jobs per sample.

MT-phylogenies annotation

To annotate MT-phylogenies with discrete cellular clones, we developed *MiToTreeAnnotator*. Given an input *CassiopeiaTree* and its binary character matrix, *MiToTreeAnnotator*:

1. Assigns each MT-SNV to a unique internal node of the tree, treating each internal node as a bipartition of the leaves. Specifically, given an internal node x, all leaves in a tree can be bi-partitioned into set1 (i.e., all leaves that share x as most recent common ancestor, the the clade identified by x) and set 0 (set1 complement). *MiToTreeAnnotator* computes Fisher’s Exact test statistics for all internal node-MT-SNV combinations, evaluating the “enrichment” of each MT-SNV across each tree clade. Each MT-SNV is assigned to the internal node with lower Fisher’s Exact test FDR.
2. Clusters MT-SNVs co-occurrence matrix into an optimal number of MT-SNVs clusters. First the binary character matrix of the *CassiopeiaTree* is transposed, and variant-variant pairwise *jaccard* distances (i.e., MT-SNVs co-occurrence matrix) are computed. Then, hierarchical clustering (i.e., *scipy.hierarchy.linkage*) is employed to group MT-SNVs into an optimal number of clusters. This is achieved heuristically, by choosing the MT-SNV partitioning (i.e., the distance threshold in *scipy.hierarchy.fcluster*) that maximizes the average silhouette score across MT-SNV clusters.
3. Uses MT-SNVs clusters and MT-SNV-internal node assignments to cut the tree into clades supported by MT-SNVs clusters (i.e., MT-clones). I.e., in the last step, *MiToTreeAnnotator* iterates across MT-SNVs clusters, locating co-occuring MT-SNVs on the tree (using MT-SNV-internal node assignments). Within this MT-SNV-assigned nodes *MiToTreeAnnotator* choose the most ancestor node MRCA of a MT-clone, and labels all cells under this node as a discrete “MT-clone”. All cells from clades without MT-SNVs assigned are annotated as “Unassigned”.

This rather simple and fast procedure effectively use the hierarchical structure encoded in the cell tree topology to “cut off” clades supported by co-occuring MT-SNVs. As shown in

the results, this method works best with high numbers of MT-SNVs (otherwise it's hard to find stable/meaningful MT-SNVs clusters). After visual inspection of MT-SNVs co-occurrence matrix, the user can also input its chosen "target" number of MT-SNVs clusters, from which MT-clones directly derive. *MiToTreeAnnotator* returns a tree with MT-SNVs assigned to internal nodes, and cells annotated with the "MT_clone" categorical label (*CassiopeiaTree*.cell_meta).

Clonal reconstruction benchmark

To benchmark the discrete output of *phylo_inference,* (i.e. the set of labels produced by MiToTreeAnnotator), we selected 10 unique hyper-parameters combinations (and therefore MT-SNVs spaces) for each sample. To make this test fair, we selected the most "informative" MT-SNVs spaces that: i) showed a discrete amount of phylogenetic signal, quantified with metrics that are agnostic with respect to the method used to infer discrete clones; and ii) included high number of cells and clones. Specifically, we used the following sample-specific criteria:

- MDA_clones: AUPRC>0.5, char-based vs tree-based distance correlation >0.6, n_cells>300, n_GBC_groups==7 and n_vars>10
- MDA_PT: AUPRC>0.3, char-based vs tree-based distance correlation >0.5, n_cells>1000, n_GBC_groups>30 and n_vars>10
- MDA_lung: AUPRC>0.5, char-based vs tree-based distance correlation >0.5, n_cells>1000, n_GBC_groups>8 and n_vars>10

To select a first set of sample-specific MT-SNVs spaces. Then, for each sample, we binned MT-SNVs spaces according to their number of MT-SNVs (n=5 bins) and for each bin we selected the top2 MT-SNVs spaces according to char-based vs tree-based distance correlation. This procedure yielded 10 unique MT-SNVs spaces (i.e., AFMs) for each sample.

These MT-SNV spaces were used to benchmark both MT-phylogenies robustness (see MT-phylogenies robustness section) and clonal reconstruction performance. For the latter, for each sample MT-SNVs space, we inferred MT-clones with:

- *MiToTreeAnnotator,* using UPMGA, NJ, iqtree, and mpboot trees as input
- *leiden* [169] clustering, using as input kNN (k=15) graphs build on cell-cell jaccard distances. For each MT-SNVs space, we selected the optimal resolution parameter

that maximized the average silhouette score across n=50 different resolution values

- *vireoSNP*[170], taking AD and DP counts as input. For each MT-SNVs space, the optimal k (i.e., number of clusters) was selected as recommended by the authors: taking the k value at which the Evidence Lower Bound (ELBO) curve stops increasing dramatically across n trials with different k values. We tried 2:maxK values, with maxK=50 for MDA_PT and MDA_lung and maxK=15 for MDA_clones
- *Cclone*[161], taking AD and DP counts as input. For each MT-SNVs space, we selected the k value (i.e., number of clusters) producing the best orthogonality score among wNMF components

Fig. 36, shows ARI and NMI scores computed using ground truth lentiviral clones and inferred MT-clones from all described MT-SNVs spaces and methods. Fig. 37 shows UMAP cell embeddings colored for ground truth and inferred discrete clonal labels. This latter visualization include only a single MT-SNVs space per sample (i.e., row): the one with better average ARI across methods.

Longitudinal dynamics of MT-SNVs

MDA_PT and MDA_lung samples were processed as described in the data preprocessing section. Other details regarding preprocessing and downstream analysis of the full MDAMB231 scLT tracing data are out of the scope of this manuscript, and will be provided in that specific project context. To investigate the longitudinal dynamics of clones MT-SNVs, we started by merging MDA_clones and MDA_PT AFMs (*maegatk* pre-processing). *filter_cells* ("filter2") was applied before filtering the merged AFMs to retain the union of all MT-SNVs showing up in "informative" MDA_PT and MDA_lung MT-SNVs spaces (n=20 MT-SNVs spaces). The resulting AFM was genotyped with the *MiTo genotyping* method and cells with at least one MT-SNV were retained for further analysis. Clones with at least 10 cells in both PT and lung were selected (n=6). The *compute_lineage_bias* function from *mito_utils* was used to select (for each longitudinal clone at PT and lung, respectively) "clonally enriched MT-SNVs" (FDR Fisher's exact test <0.1, cell prevalence within the clone >1%, fraction of clone cells positive for the MT-SNV over the total number of positive cells > 75%). Dotplots are used to visualize median AF and cellular pravalence of these clonally enriched MT-SNVs across longitudinal clones (PT and lung, separately).

scLT experiment transcriptional characterization

All samples (n=12) from the *in vivo* scLT experiment described in "MiTo benchmark sample preparation" were preprocessed with *mito_preprocessing,* using the "TENX_GBC" entrypoint (i.e., only pre-processing of GEX and GBC libraries). QC on GEX and GBC modalities was performed as described in "Quality Control". GEX matrices from *STARSolo* (i.e., CBs passing both GEX and GBC QC steps) were processed within the standard *scanpy* pipeline[171]. Cell states were manually curated, and gene modules were inferred with *Hotspot*[172]*.*

Statistics

Boxplots highlights medians, with box margins representing the IQ range and whiskers extending to the 10-90th percentiles.

Data availability

Raw data (.fastq) files will be submitted to the European Nucleotide Archive (ENA) at manuscript submission. Processed data will be submitted to Zenodo for reproducibility at manuscript publication.

Code availability

mito_preprocessing: https://github.com/andrecossa5/mito_preprocessing
phylo_inference: https://github.com/andrecossa5/phylo_inference
mito_utils: https://github.com/andrecossa5/mito_utils
Reproducibility code for this work: https://github.com/andrecossa5/MI_TO_analysis_repro

Results

The MiTo benchmarking dataset

To quantitate phylogenetic signals associated with expressed MT-SNVs, orthogonal lineage markers are needed.

Lentiviral barcoding has been recently used for single-cell lineage-tracing (scLT) across multiple fields in developmental and cancer biology [88]. Of note, this technology was previously used to benchmark the reliability of expressed MT-SNVs as lineage markers [125]. However, this seminal work provided two benchmarking datasets with limited number of cells (80 and 180, respectively) and clones (3 and 18, respectively), due to the low scalability of the plate-based, full-length Smart-seq protocol. More recently, several groups developed protocols able to obtain full-length sequence information of individual MT-transcripts using target enrichment of the cDNA produced with the 10x technology. Importantly, this droplet-based scRNA-seq platform is much more scalable than the older Smart-seq2, and effectively leverages UMIs (Unique Molecular Identifiers) to mark PCR duplicates of individual RNA molecules. However, validation of these new protocols was more focused on the quality of recovered MT-transcripts and SNVs rather than the recovery of ground-truth clonal structures, limited by the analysis of species-mixing experiment data. Accordingly, while computational methods for raw mitochondrial-sequence data pre-processing, variant filtering, genotyping, and lineage inference have been developed to extract clonal structures from data of increasingly higher throughput, benchmarking of these methods has been limited by the lack of complex labelled data outside of the datasets in [125] and [142].

To fill these gaps, here we generated a novel high-quality dataset encompassing ~4k cells and >200 ground-truth clones, as defined by expressed lentiviral barcoding. To this end, we extended the original MAESTER protocol in [138] to include target enrichment of exogenous lentiviral transcripts. The resulting single-cell multi-omics protocol starts from a 10x-generated 3' cDNA to produce 3 sequencing libraries and associated data modalities (Methods): i) the Gene Expression (GEX) library, with the standard full transcriptome single-cell read-out, ii) the lentiviral-barcode (GBC) library, enriched for barcodes-containing transcripts marking the clonal origin of each cell, and iii) the mitochondrial (MT) library, enriched for mitochondrial transcripts as in the original MAESTER publication [138].

The MAESTER protocol was chosen for mitochondrial enrichment (**Fig. 6**) due to its straightforward use and flexibility (necessary for the addition of the lentiviral data modality), the highly-optimized yield in terms of MT-transcript, higher scalability and cost-effectiveness, compared both to other droplet-based [173] and plate-based [125] protocols.

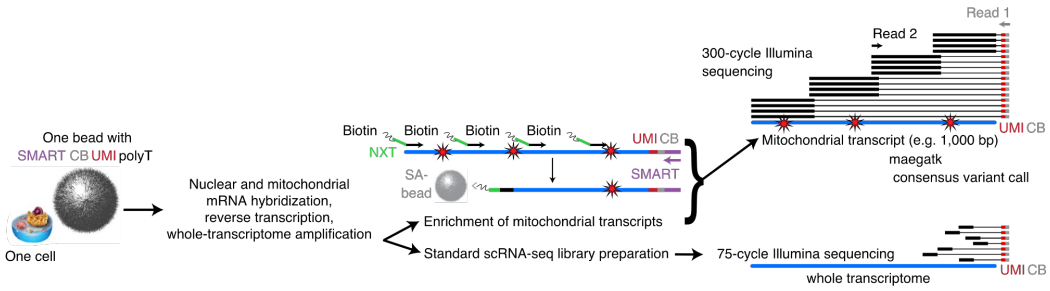


Fig. 6. ***MAESTER protocol target enrichment****. Readapted from Miller et al.,* [137]*.*

Importantly, starting from 10x 3' cDNA, this protocol produces sequencing reads that can be de-multiplexed for their cellular of origin (using the 10x cellular barcode, CB) and molecular identity (using the UMI barcode), allowing for UMI-based consensus error-correction strategies before variant calling.

To cover multiple real-world clonal-complexity scenarios, we generated data from 3 high-quality biological specimens, all derived from the MDA-MB-231 Breast Cancer cell line: i) a mixture of single-cell derived colonies generated *in vitro* (the MDA_clones sample), and a ii) pair of matched primary tumor- (PT) lung metastasis samples grown in vivo (the MDA_PT and MDA_lung samples, respectively), chosen from a larger longitudinal scLT experiment (Methods). For the *in vitro* samples, single cells from MDA-MB-231 cells were sorted and expanded *in vitro* to obtain single-cell derived colonies. After ~30 days in culture, each colony was infected with a single lentiviral species (i.e., lentiviral particles all bearing a unique barcode) and after a short (~7 days) selection period, 8 barcoded colonies were mixed and subjected to library preparation and sequencing (Methods).

For the *in vivo* samples, MDA-MB-231 cells were infected with a high complexity lentiviral library (n~10^6 unique random barcodes), selected (as for the *in vitro* samples) and orthotopically injected into immunodeficient mice. After ~30 days, the PT was surgically removed and. After approximately one month, mice were sacrificed, and lung metastasis collected (Methods). Cellular suspensions from all specimens were generated and subjected to MiTo protocol sequencing.

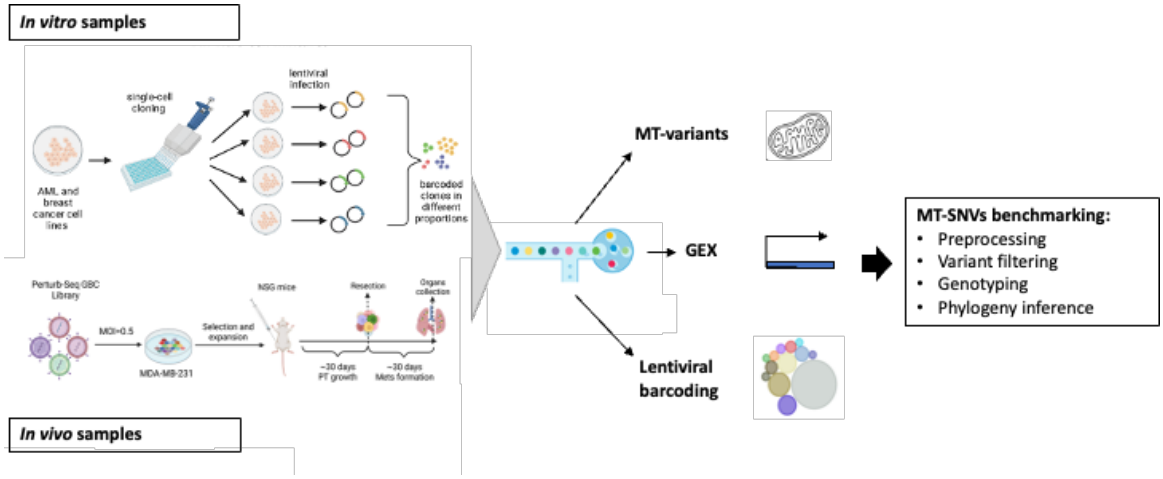


Fig. 7. ***MiTo benchmarking, overview.***

With sequencing data at end, particular attention was given to pre-processing and quality control to ensure that only good quality cells for all data modalities were included in the final dataset.

To achieve this, we developed the *mito_preprocessing* pipeline, a feature-rich workflow that can handle pre-processing of up to 3 sequencing libraries/data modalities per sample (i.e., GEX, GBC and MT, or combinations of them) (Methods). This pipeline streamlines several operations for MiTo data pre-processing, (Methods) producing 3 main outputs: i) a per cell, per gene UMI-counts matrix, ii) a table of CB-GBC UMI counts, from which clonal assignments are derived, and iii) several tables of *per* base, *per* site and *per* cell MT-genome UMI counts, with statistics used for downstream MT-SNVs filtering and genotyping.

Compared to available tools and implementations, the *mito_preprocessing* pipeline introduces two main novelties. First, it leverages consensus-sequence correction for individual UMIs from the GBC library followed by robust CB-GBC combinations filtering and cell-clone assignment (Methods). Together, these two steps effectively remove spurious and/or poorly detected CB-GBC combinations and generate more accurate clonal labelling of individual cells (**Fig. 9**), a fundamental step for our benchmarking study.

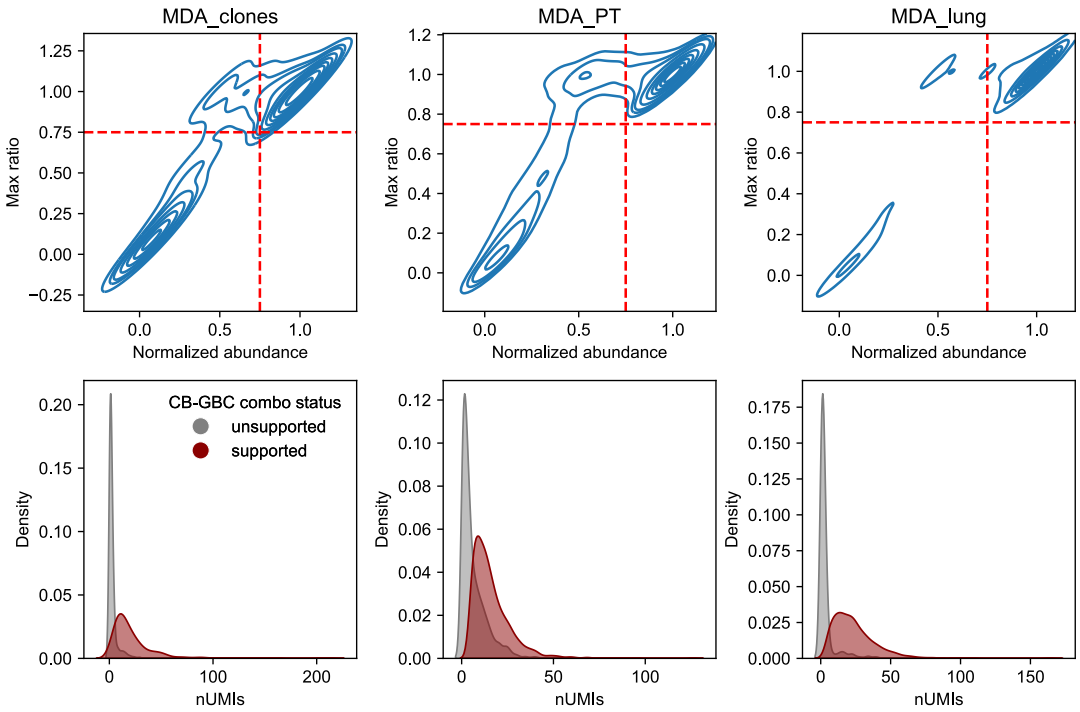


Fig. 9. *CB-GBC combinations filtering for robust clonal labelling*. *Top row: Density plot showing the normalized abundance (x-axis) (Methods) vs max ratio (y-axis) of each CB-GBC combination, for each sample. Clonal assignment (Methods) considers only the most abundant top right been combinations, clearly separated from the noisy one in the bottom-left corner. Bottom row: Density plot of the n of consensus UMIs from the GBC library (Methods) detected for unsupported and supported CB-GBC combinations (each column represents a different sample, ordered as in the top row). These statistics include CB-GBC combinations from all CBs qualified as putative cells from STARsolo.*

Second, *mito_preprocessing* performs more stringent base filtering than previously developed in the original *maegatk* pipeline, i.e., the state-of-the-art MAESTER data pre-processing toolkit developed in [138]. In the original publication, *maegatk* introduced consensus error correction for MT sequences according to the following workflow: *i)* PCR replicates (i.e., reads) from the same RNA molecule (UMI) are used to call a single "molecular consensus" sequence; ii) consensus sequences are re-mapped to the MT-genome.; iii) for each cell, consensus bases (filtered for their base calling quality and for the alignment quality of the entire consensus sequence) are recorded into forward and reversed pile-up tables. These tables are then used for downstream analyses (i.e., data quality checks, variant filtering, etc.). In this implementation, *maegatk* uses *fgbio CallMolecularConsensus* [152] for consensus sequence calling, without filtering any base at this step (i.e., all observed bases in a group of reads assigned to the same RNA molecule are used to generate the consensus sequence). In addition, *fgbio CallMolecularConsensus* produces unaligned reads with tags annotating two base-

specific consensus metrics: i) *cd:* the "consensus depth" or "UMI group size" [142], the number of good quality reads that were used for *fgbio* sequence consensus, and ii) *ce:* the "consensus error", the number of discordant bases from the one registered as final "consensus base". In its original implementation, *maegatk* enable filtering of good quality bases and consensus reads, but does not consider base-specific UMI-group size and consensus error as additional indicators of reliable molecular information. Since recent works attempted to make single-cell MT-SNV calls with single-molecule detection[142] (i.e., basecalls, supported by a single consensus UMI), we hypothesized that even a small fraction of low-quality/weakly-supported consensus UMIs could lead to sub-optimal lineage inference accuracy. To test this hypothesis, we modified the original *maegatk* implementation in order to: i) use only Q30 bases for consensus sequence calling (*fgbio CallMolecularConsensus --min-input-base-quality* parameter*)*; ii) include filtering of individual consensus bases according to their UMI group size and consensus error (Fig….); iii) leverage STARSolo[148] for MT reads alignment and CB,UB de-duplication *before* UMI consensus sequences generation, base filtering and pile-up (instead of bulk-mode STAR as in [138], which does not make any CB and UB de-duplication).

This modifications produced very similar outputs overall, as shown in **Fig. 10**. For instance, considering MDA_clones, the cell coverage (i.e., the median number of consensus UMIs across MT-genome target sites) was extremely similar across pipelines (Pearson's r>0.99), with *mito_preprocessing* yielding fewer counts (median 66 vs 80 counts per site and cell, respectively), a smaller fraction of target sites covered (median 85% vs 89%), and smaller UMI group sizes (median 10.2 vs 10.5) compared to *maegatk.* Indeed, *maegatk* recorded ~83k basecalls with average consensus score (i.e., 1-consensus error of supporting consensus UMIs) <0.7 (1.81% of total consensus basecalls, ~4.6M) and ~120k basecalls with average base-calling quality <30 (2.61% of total consensus basecalls) that were absent in *mito_preprocessing* basecalls. Strikingly, these apparently minor differences resulted in dramatically different (i.e., ~3-fold difference) numbers of variant basecalls (i.e., raw basecalls with at least one UMI supporting an alternative allele for some cell at some MT-genome position) across pipelines, with median 142 (+-71) and 442 (+-160) variant basecalls across cell for *mito_preprocessing* and *maegatk*, respectively. We will see in the next chapter how these different choices in data pre-processing coupled with different approaches for MT-SNVs filtering and genotyping may influence the "informativeness" of detected MT-SNVs.

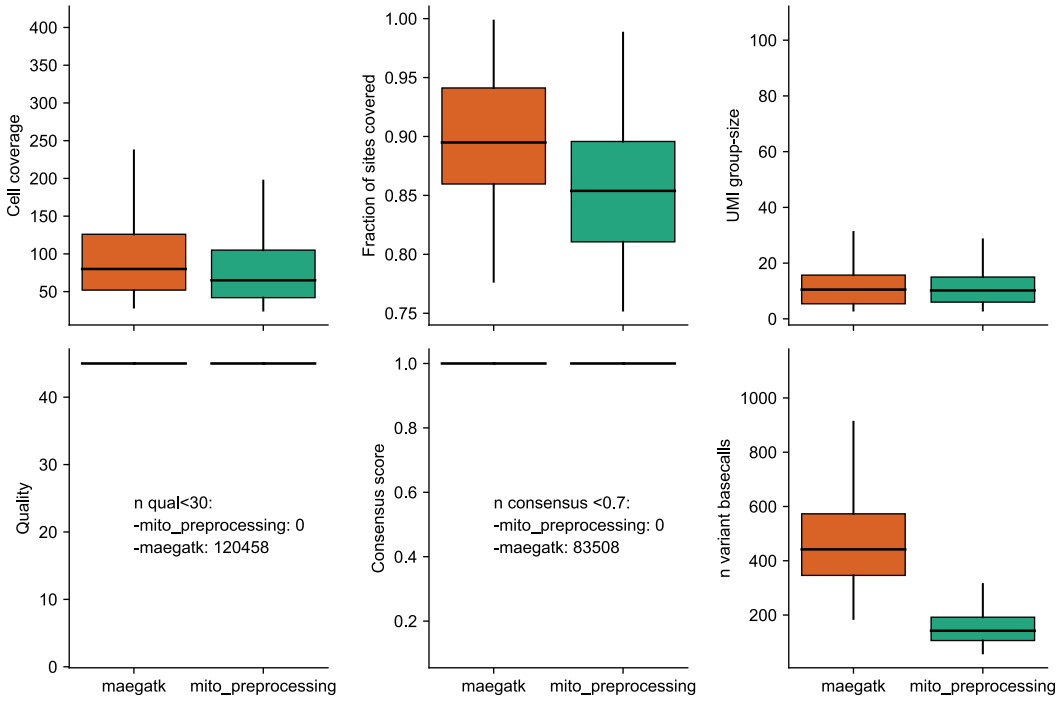


Fig. 10. ***mito_preprocessing-maegatk comparison****. mito_preprocessing removes low quality and consensus basecalls from MAESTER data, compared to maegatk (default parameters, except for –min_reads=3). Cell coverage: median n UMIs on MT-genome target sites, per cell. Fraction of sites covered: Fraction of MT-target sites covered (i.e., n UMIs>0), per cell. UMI group-size: mean group size of supporting UMIs, per basecall.* Quality: *mean quality of supporting UMIs, per basecall.* Quality: *consensus score of supporting UMIs, per basecall. n variant basecalls: n of raw, unfiltered basecalls supporting alternative alleles, per cell. All statistics have been computed considering the same subset of MDA_clones cells (n=345) qualified for all 3 modalities (i.e., GEX, GBC, MT, see Methods).*

The *mito_preprocessing* pipeline is under active development (Methods). *mito_preprocessing* is implemented in a highly modular and extendable fashion thanks to the flexible Nextflow DSL2 synthax. The pipeline uses Docker/Singularity containerization for cross-platform and High-Performance Computing cluster compatibility. Proper tuning of individual process resources and pre-processing options is possible via dedicated configuration files. Moreover, to facilitate downstream benchmarking tasks, *mito_preprocessing* comes with a benchmarking subworkflow for MT- data pre-processing (-entry BENCH). This subworkflow can be used for MT-SNV pile-up/genotyping with 4 alternative pre-processing tools, namely: *samtools*[151], *freebayes*[157], *maegatk*[138] and *cellsnp-lite*[155] (Methods).

After pre-processing (with *mito_preprocessing, ndr*) and Quality Control on individual data modalities (Methods), the final MiTo benchmarking dataset, included 3 high quality samples with a total of 3795 qualified cells (Fig.). These samples displayed variable numbers of cells (345-2069), clones (8-194 considering all clones, 7-31 considering clones

with >=10 cells, representing 81-98% of each sample) and clonal complexity (Shannon Entropy: 0.77-1.66) (**Fig. 11**), from the simple MDA_clones mixture to the highly multi-clonal PT (MDA_PT) of the matched PT-lung couple.

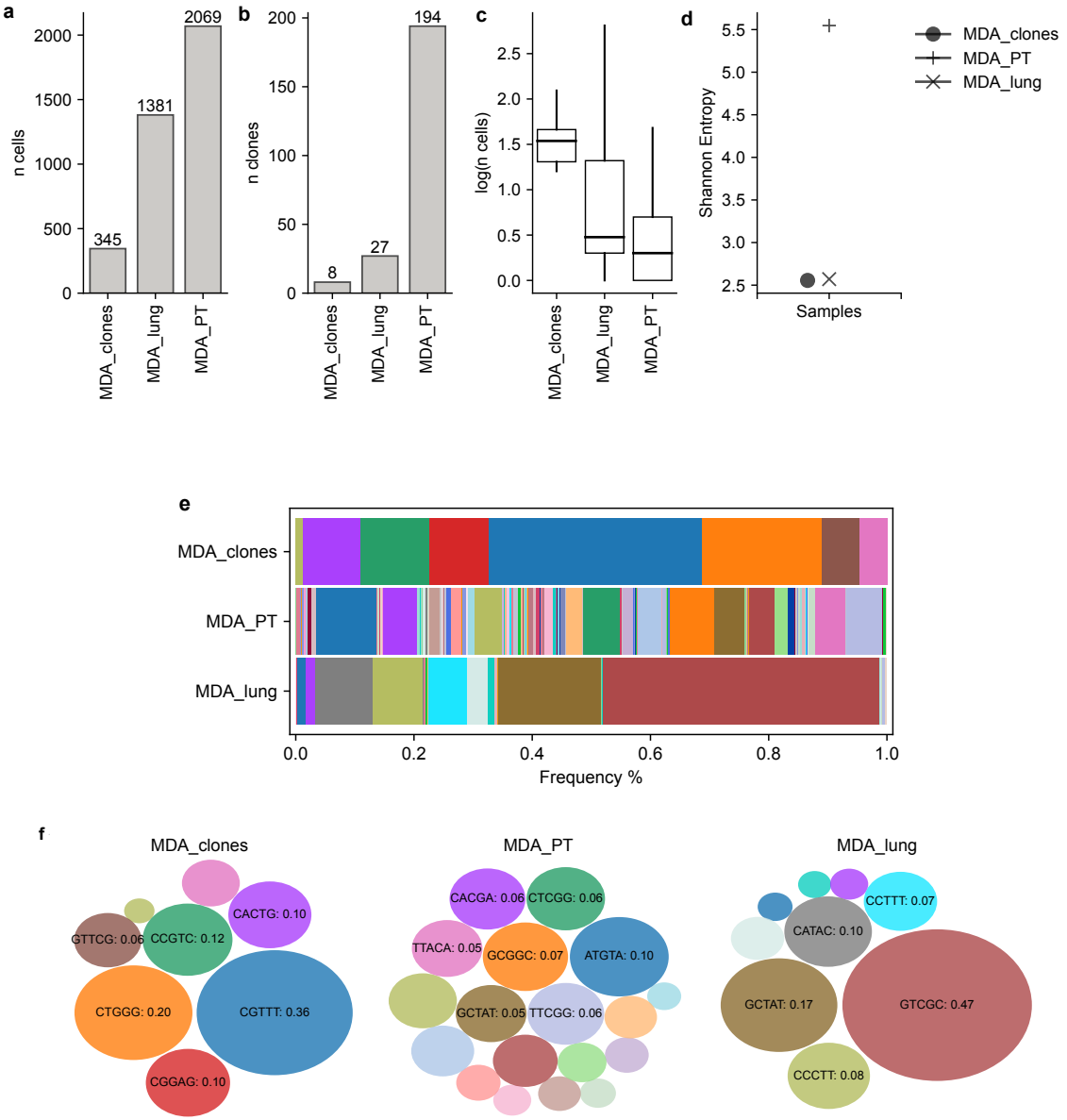


Fig. 11. ***The MiTo benchmarking dataset****.* ***a-d.*** *Number of lentiviral clones, cells, clone size and Shannon Entropy, for each sample.* ***e.*** *Stacked barplot of clonal prevalences for each sample (all clones).* ***d.*** *Circle packed plot showing only lentiviral clones with >= 5% prevalence (same color code as in* ***e****).*

	MDA_clones	MDA_PT	MDA_lung
n reads total	245631788	283041484	447696382
n genes (cell)	8307	4718	6011
n UMIs per (cell)	108654	18884	34235
n Solo cells	954	6555	5864
n qualified cells	862	6203	5253
n reads total	105228371	28514982	35710231
n consensus UMIs (cell)	11	13	17
observed MOI (cell)	1	1	1
Pearson's r with ref	0.59	0.88	0.42
n qualified cells	587	4845	2460
n reads total	207931846	302964967	283057136
n consensus UMIs (cell)	2887	1226	1483
Group size (UMI)	10	6	7
Q30 (UMI)	45	45	45
Consensus score (UMI)	1	1	1
nUMI per target site (cell)	67	40	49
Target/untarget log-ratio	7.83	7.60	7.69
n qualified cells	557	2885	2839

Table 1. ***The MiTo benchmarking dataset****. Pre-processed (mito_preprocessing) data statistics for each samples and sequencing libraries (blue: GEX, green: GBC, orange: MT) (Methods). All statistics in this table are calculated for CBs passing quality control of the related library (Methods).*

While the *in-vitro* grown MDA_clones specimen was specifically generated for *MiTo benchmarking*, the PT-lung couple was derived from a larger scLT experiment, whose complete analyses will be presented elsewhere. This experiment included >50k good-quality cells in 12 PT-lung longitudinal couples randomized across 4 chemotherapy arms, and was performed to investigate *in vivo* Breast Cancer clonal and phenotypic behaviors upon chemo-therapy at different stages of cancer progression (Methods). Here we will provide a glimpse of the transcriptional characterization obtained so far for that dataset, to provide complementary information about the identity of the cells included in the *MiTo benchmarking* dataset. Specifically, the MDA_PT and MDA_lung cells used in this study derive from a single-untreated mouse chosen for its good transcriptional quality-control metrics and manageable clonal complexity (both at the PT and lung level). Phenotypic heterogeneity is pervasive across these cells, considering both the tissue of origin, treatment condition and proliferation status (**Fig. 12**, top). Unsupervised clustering and gene modules (Methods) were used to annotate cells into 11 distinct cell states (**Fig. 12**,

bottom) including several well-known cancer cell phenotypic traits (i.e., proliferation, EMT, Interferon signaling, glycolysis, hypoxia and Stress).

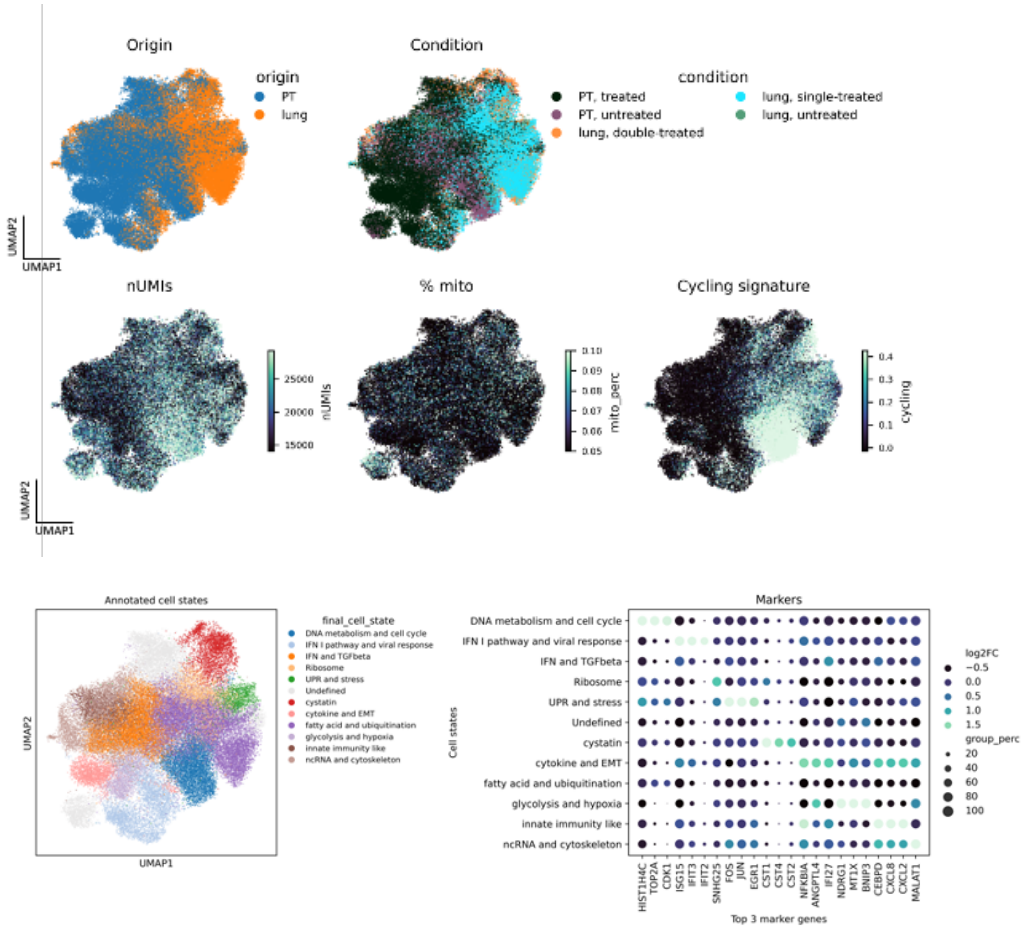


Fig. 12. ***Transcriptional charachterization of the full scLT BC_chemo dataset. a.*** *UMAP plots with the full dataset (n=58k cells) colored by tissue of origin, treatment condition and standard diagnostics covariates: the number of UMIs (nUMIs), the percentage of total UMIs assigned to MT-genes (% mito) and a cycling signature (Methods).* ***b.*** *Left: UMAP plot with cells annotated for their cell state. Right: Dot plot showing markers genes for each cell state (n=3 markers, if available, for each cell state).*

mito_preprocessing site coverage was in line with previous studies [138,142]: across samples, we detected a median (across cells) site coverage of 40-71 consensus UMIs (UMI group size>=3 mean, base calling quality>=30 and mean consensus score >=0.7) for targeted loci (~13k bases across ~16k total MT-genome size), and ~0 in regions untargeted by MAESTER PCR primers (**Fig. 13**).

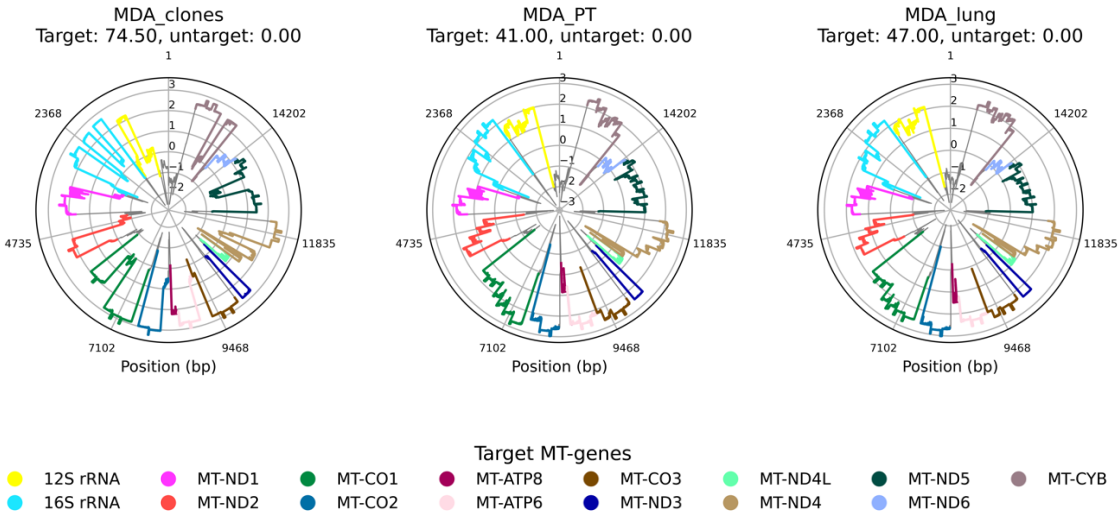


Fig. 13. ***MT-genome coverage from MAESTER target enrichment.*** *Each plot represents in radial coordinates the median (across cells) log-number of consensus UMIs over at each MT-genome position (polar coordinate). Genomic regions are colored by gene loci. Grey regions are untargeted MT-regions.*

Further inspection revealed other important features of MAESTER data (**Fig. 14**). First, the vast majority of basecalls are derived from sequences aligned in reversed orientation. Second, the vast majority of these bases were observed either in the forward or reversed strand (~96% total observed basecalls). This is consistent with pervasive mono-allelic expression of MT transcripts enriched by MAESTER protocol. Third, we observed a very strong correlation (Pearson's r=0.93) between the mean expression of MT-genes (as assessed by gene-tag UMI quantification, standard scRNA-seq UMI counts from the GEX library) and the median site-coverage (aggregated across sites of distinct MT-genes) of the same genes quantified after pre-processing of the MAESTER library, highlighting gene expression-biases in MT-genome coverage that is linearly amplified by MAESTER targeted enrichment. These three properties also illustrate the main differences between RNA- and DNA-based single-cell MT-SNVs genotyping [137,142], regardless of PCR enrichment steps, variation in sequencing depths and pre-processing pipelines. DNA-based workflows yield uniform coverage across MT-genome sites and paired-strand molecular evidence for alternative/reference allele calls [139,142], which does not apply to RNA-based workflows, where MT-genome site coverage is unevenly distributed and basecalls are supported almost exclusively by single-stranded molecular evidence, posing additional challenges in error detection and MT-SNVs filtering.

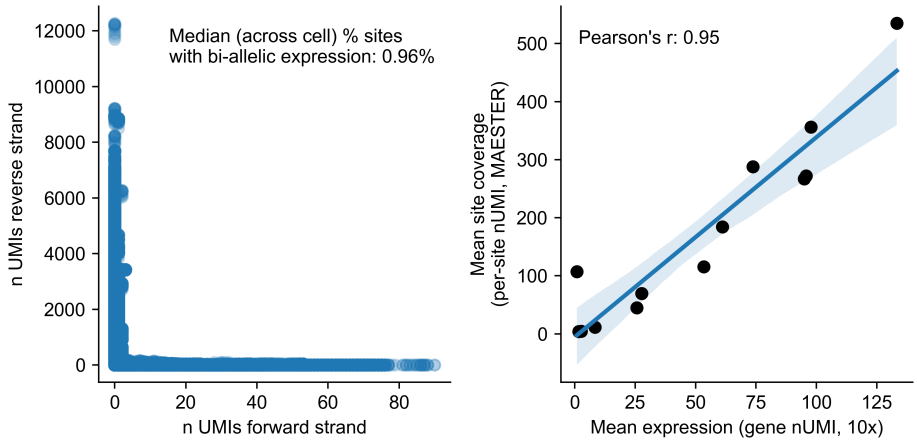


Fig. 14. ***Strand and expression biases in MAESTER data****. Left: Scatterplot showing the number of forward (x-axis) and reversed (y-axis) UMIs supporting MDA_clones basecalls (each dot is a basecall). Right: Regplot showing linear relationship between the mean (across cell) expression of each targeted MT-gene (each dot) as detected in the GEX library, and its (average) site coverage as detected in the MAESTER library. Shaded area highlights 95% confidence intervals.*

Together, building on previous single-cell multi-omics protocols and bioinformatic pipelines, we demonstrated the feasibility of joint lentiviral-barcodes and expressed MT-SNVs profiling using the 10x technology. Introducing several innovations in error correction, the MiTo protocol and the flexible *mito_preprocessing* pipeline provided an high-quality dataset for lineage inference benchmarking.

Phylogenetic signal in expressed MT-SNVs spaces: the accuracy-cellular yield trade-off

Collecting raw MT-allele basecalls and associated statistics is the first step towards MT-based lineage inference, and requires automated, scalable and flexible bioinformatic pipelines. However, after this first step, the single-cell analyst is given the responsibility to select an “informative”[138] set of MT-SNVs, i.e., a set of phylogenetic characters holding relevant information about the evolutionary relationships within a population of cells. Unfortunately, giving the lack of extensive benchmarks with ground truth lineages, there is currently no clear and context-independent definition of what “informative” means for a MT-SNV. Thus, standard practice relies on extensive exploratory analyses and *ad hoc*

filtering strategies to empirically select the best candidate MT-SNVs "space", before attempting lineage inference and follow-up analyses.

To facilitate both interactive and automated exploration of alternative MT-SNVs spaces, we developed the *mito_utils* python package and *phylo_inference* Nextflow pipeline (Methods).

The *mito_utils* package implements several utilities and APIs for the interactive exploration of MT-SNVs data, including cell and variant filtering, cell- and character distance, kNN-graph calculations, dimensionality reduction, phylogenetic tree building, diagnostics and visualization. *mito_utils* includes two novel algorithms: i) the *MiTo genotyping* method, which uses the posterior probabilities of the two-component binomial mixture introduced by Kwock et al., 2022 to assign binary genotypes at each cell-MT-SNVs combination (Methods), and ii) the *MiTo tree annotator*, a general purpose tree post-processing method that: a) assigns individual MT-SNVs to internal tree nodes; b) finds the optimal clustering of co-occurring MT-SNVs; and c) cuts the input tree into discrete, MT-SNVs-supported clades, interpretable as "cellular clones" (Methods).

To facilitate interoperability with other single-cell and scLT frameworks, *mito_utils* includes two core widely popular data objects: i) the *AnnData* class from the *anndata*[155] package, the gold-standard for efficient storage and access of sparse, annotated data matrices in single-cell genomics, and ii) the *CassiopeiaTree* class from the *cassiopeia*[165] *package,* the leading python library for single-cell lineage tracing.

phylo_inference instead integrates the main features of *mito_utils* and other popular phylogenetic tools[166–168] into two modular and scalable Nextflow DSL2 workflows: i) the *tuning* workflow, allowing automatic and flexible selection of MT-SNVs spaces with user-defined values of more than 20 hyper-parameters for cell and variant filtering, MT-genotyping, and phylogenetic tree building. Importantly, this workflow reports a comprehensive set of metrics to rank candidate MT-SNVs spaces, cell-cell distance matrices and resulting cell phylogenies, allowing fast and efficient prioritization of the most likely "informative" MT-SNVs spaces at hand. Once inormative MT-SNVs have been found, ii) the *phylo* workflow provides advanced functionalities for tree building, bootstrapping and post-processing, producing annotated trees for fine-tuned downstream analyses (Methods).

mito_preprocessing, *phylo_inference* and *mito_utils* are under active development (Methods). They all stems from the same *MiTo* project, and while efforts have been put to ensure a certain level of flexibility in expected inputs and outputs, they are all designed to work together as a single, integrated solution for MT-based single-cell multi-omics, from raw .fastq sequences to annotated lineage trees.

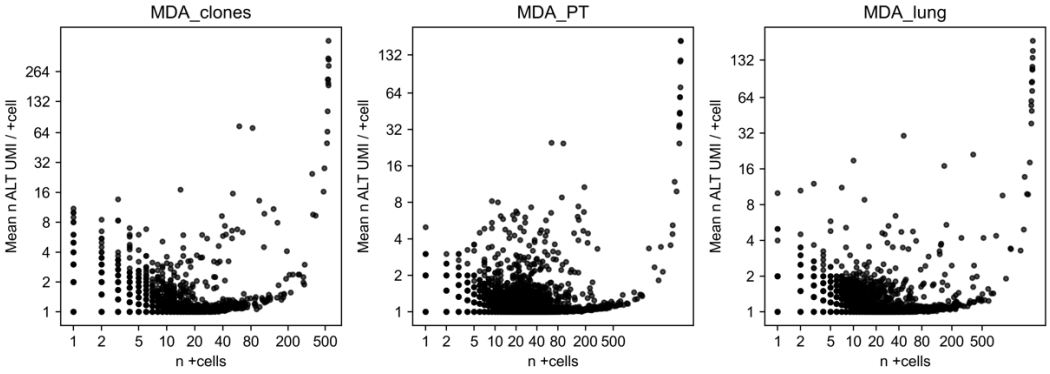


Fig. 15. ***Un-filtered MT-SNVs spaces for MiTo benchmarking samples****. These scatterplot represent the raw, unfiltered space of MT-SNVs that needs to be filtered to retrieve high quality lineage markers (if any).*

With this toolkit at hand, we first wanted to assess the defining properties of "informative" MT-SNVs in our MiTo benchmarking dataset, given the ground truth definition of clonal structure from lentiviral barcoding.

Starting from the basecalls of 5 different pre-processing pipelines, we tested 1764 (n=588 per sample) unique combinations of cell and MT-SNVs filtering strategies, binarization methods, distance metrics and tree reconstruction algorithms (Methods). For each combination, we recorded a set of 19 metrics measuring: 1) the quality of filtered MT-SNVs ("Mutation Quality"), 2) the association of cell-to-cell distances and reconstructed trees with ground truth labels from lentiviral barcoding ("Association with GBC"), 3) the consistency between reconstructed tree topologies and underlying genetic characters ("Tree structure"), 4) the cell "connectedness" in MT-SNV spaces ("Connectedness"), 5) the variation of MT haplotypes ("Variation"), and 6) the cellular yield, i.e., the number of clones, cells and variants obtained for downstream analysis ("Yield") (Methods) (**Fig. 15-18**).

Then, for each sample we grouped all tested combinations by 5 hyper-parameters of particular interest (i.e., pre-processing pipeline, binarization method, minimum Allelic Frequency of confident detection, minimum number of confidently detected cells, and minimum number of ALT alleles required to assign a MUT genotype. n=108 groups for each sample), aggregated each metric value by median, and ranked overall performance through a single summary score, following the same approach adopted in [35] (Methods). Specifically, we rescaled each metric value with min-max normalization, averaged these values across each metric type, and use a weighted sum of these values to produce a final "Overall" score. The weight of each individual metric type was chosen to maximize both

lineage inference accuracy and cellular yield (i.e., we assigned 0.4 to both “Association with GBC” and “Yield”, and 0.1 to both “Tree structure” and “Mutation Quality”, to control for potential errors in lentiviral barcoding and variant calling, and 0 to other scores). These procedure was repeated separately for each sample to maximize insights coming from different cellularity and clonality scenarios.

For the simplest lineage inference task (**Fig. 16**, MDA_clones, low number of cells and clones, relatively balanced clonal prevalences, high cell and site coverage) the top performing combinations made use of the *mito_preprocessing* pipeline coupled with the simplest binarization method (i.e., *vanilla*, where AF>some threshold is sufficient to score the presence of a MT-SNVs), filtering variants that were confidently detected in at least 2-3 cells within a narrow high AF range (0.02-0.03). On the contrary, bulk methods readapted for MT-SNVs variant (i.e., *freebayes* and *samtools*) calling were consistently bottom-ranked. Under this scenario, joint optimization of “Association with GBC” and “Yield” was feasible without compromising too much any of the two measures (i.e., the top 5 ranked hyper-parameter combinations recovered all seven lentiviral clones and >300 cells, while achieving 0.66+-0.07 ARI, 0.76+-0.03 NMI and 0.75+-0.02 AUPRC).
Interestingly, top ranked combinations showed superior MT-SNVs quality, more supported tree structures, and intermediate values for “Variation” and “Connectedness” scores, with less clear cut separation between top and bottom ranked combinations considering these latter metric types (**Fig. 16**, MDA_clones, low number of cells and clones, relatively balanced clonal prevalences, high cell and site coverage).

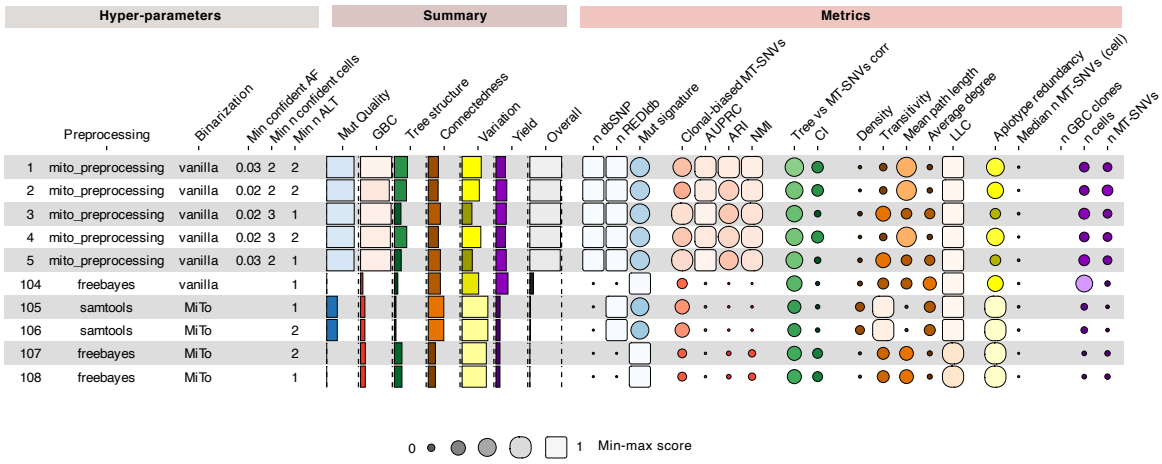


Fig. 16*. MiTo benchmarking results overview, MDA_clones*. *This funkyheatmap represent the top- and bottom-5 hyper-parameters combination (n=108 total) for the MDA_clones lineage inference task. See Methods for a comprehensive description of hyper-parameters, metrics and scoring and ranking methods.*

At the opposite side of the spectrum (Fig…, MDA_PT sample, higher number of cells and clones and and lower cell and site coverage compared to MDA_clones), the balance between cellular yield and GBC association was more difficult to achieve, particularly considering the final number of recovered cells (i.e., the top 5 ranked hyper-parameter combinations recovered >31 lentiviral clones and >1000 cells, while achieving 0.72+-0.06 ARI, 0.81+-0.04 NMI and 0.43+-0.07 AUPRC). Here, the best preprocessing-binarization combination was consistently *maegatk-MiTo,* with top ranked MT-SNVs spaces including variants in the 0.01-0.05 confident AF range, and at least 2 ALT UMIs required to assign the MUT genotype. Interestingly, even if all top scoring combinations consistently showed high tree structural support, the worse performing combination (i.e., *freebayes,* with either *MiTo* or *vanilla* genotyping) showed similar structural properties (n.d.r., the association between these structural metrics and the “GBC association score” is analyzed further at the and of the chapter). Indeed, MT-SNVs spaces “connectedness” was low in top-ranked combinations, as indicated by the high average path length and the low average degree of cells in their shared MT-SNVs graph (Methods).

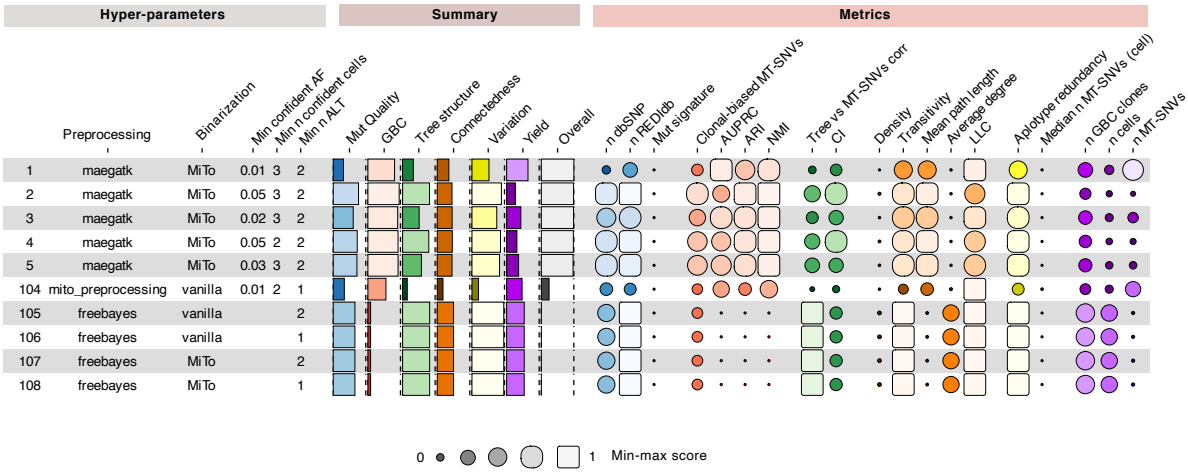


Fig. 17*. MiTo benchmarking results overview, MDA_PT*. *This funkyheatmap represent the top- and bottom-5 hyper-parameters combination (n=108 total) for the MDA_PT lineage inference task, (see **Fig. 16**. and Methods).*

Very similar patterns were observed for the intermediate clonal-complexity sample (Fig…, MDA_lung, high number of cells, low number of umbalanced clones, lower coverage compared

to MDA_clones). In this case, top 5 ranked hyper-parameter combinations recovered >9 lentiviral clones and >1000 cells, while achieving 0.82+-0.07 ARI, 0.73+-0.05 NMI and 0.70+-0.07 AUPRC. Here, the best performing preprocessing-binarization method was *maegatk-vanilla,* with variants selected at higher values of confident AF (0.03-0.1) compared to MDA_clones and MDA_PT.

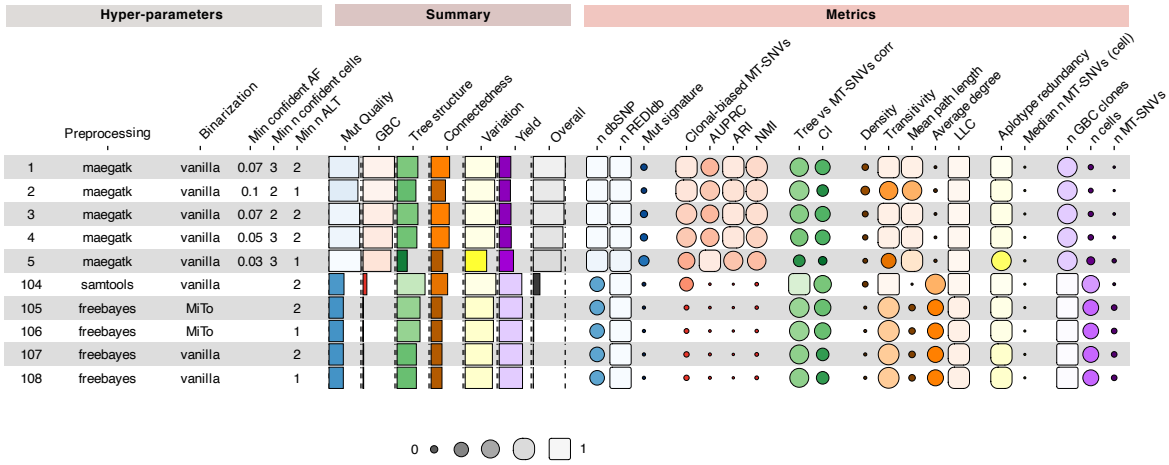


Fig. 18*. MiTo benchmarking results overview, MDA_lung*. *This funkyheatmap represent the top- and bottom-5 hyper-parameters combination (n=108 total) for the MDA_lung lineage inference task, (see **Fig. 16**. and Methods).*

To gain a more detailed understanding on these variable rankings and performances, we performed a meta-analysis of collected metrics (Methods).

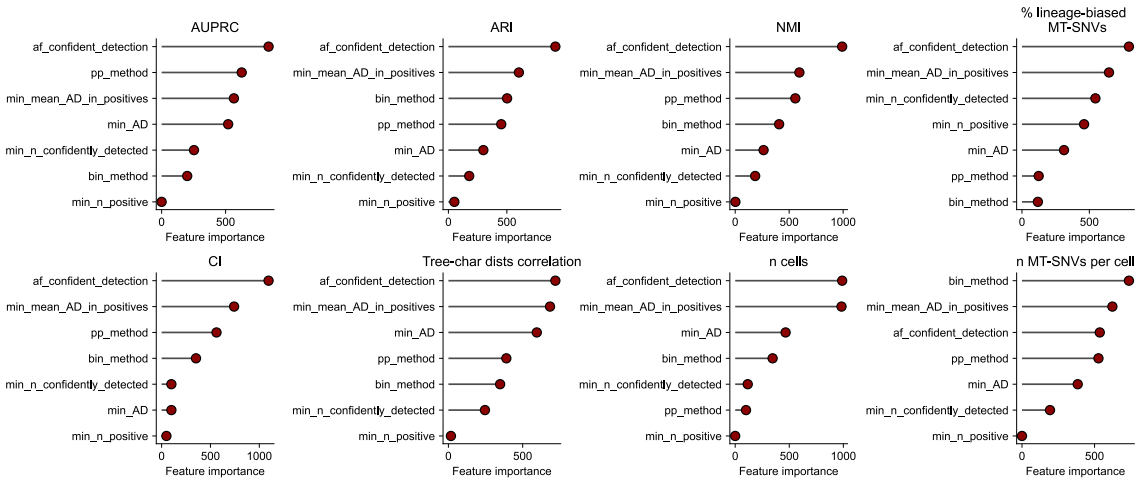


Fig. 19. MiTo benchmarking hyper-parameters feature importance. *Association (i.e., lightgbm feature importance) between tested hyper-parameters and select metrics (Methods).*

Regression analysis of tested hyper-parameters against metrics of interest (**Fig. 19**) revealed that the AF of confident detection and the minimum mean number of alternate (ALT) alleles in positive cells are by far the most important features influencing "informativeness" of MT-SNV spaces, with more modest and variable contribution of other hyper-parameters (e.g., pre-processing and binarization methods). Therefore, we investigate the contribution of these hyper-parameters separately.

As observed in **Fig. 20** and **Fig. 21**, MT-SNVs spaces including progressively more confident detection events (i.e., higher AF and higher mean number UMIs supporting the ALT allele in positive cells) gave MT-clones (Methods) that are much more concordant with ground truth lentiviral labels, as measured by Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) scores (**Fig. 21**, second row). Depending on the sample, this increase in ARI and NMI saturates within the 0.02-0.05 AF range.

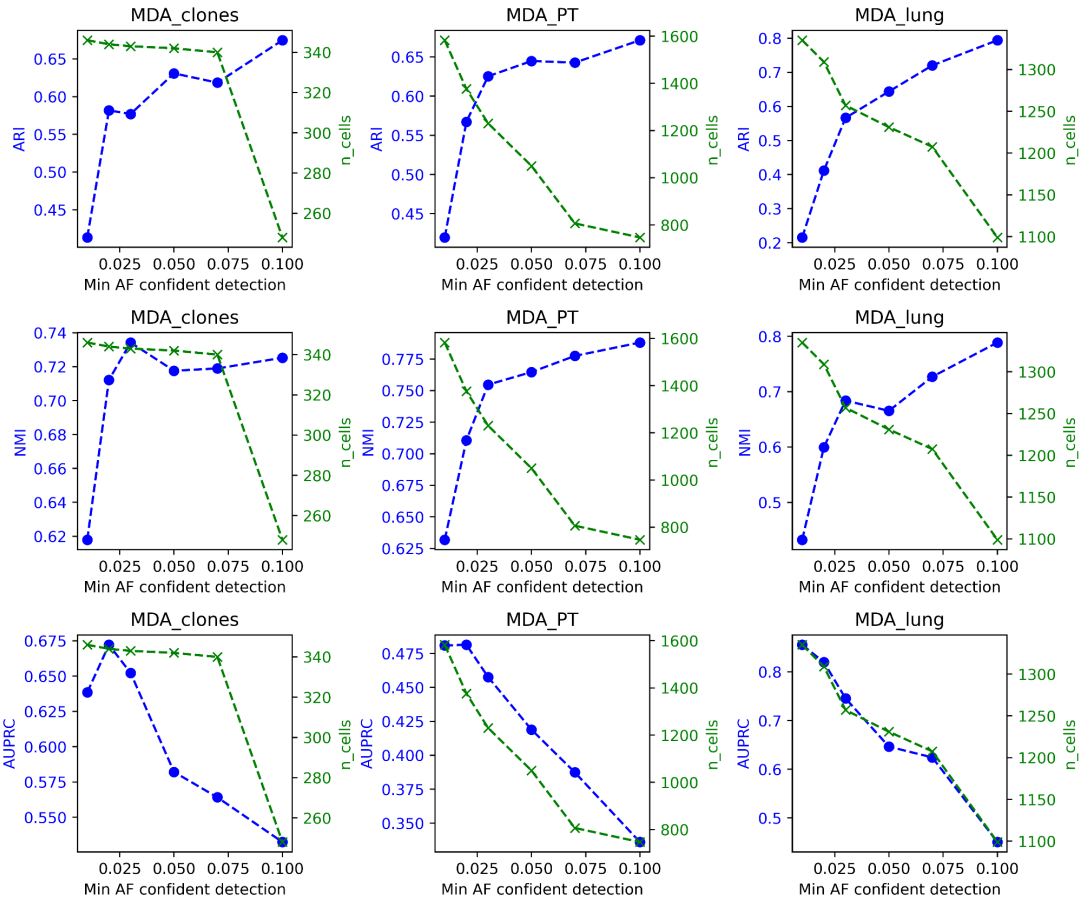


Fig. 20. ***Accuracy-cellular yield trade-off.*** *GBC association metrics and number of recovered cells at different values of the threshold for minimum AF for the confident detection of a MT-SNV (Methods).*

Consistently, requiring at least 2 ALT UMIs instead of 1 as minimal molecular detection evidence to assign mutant MT-SNV genotypes resulted in more accurate clonal reconstruction (**Fig. 21**, fourth column). Interestingly, higher values of confident AF tresholds are mirrored by higher AUPRC values (i.e., the Area Under Precision-Recall Curve obtained using cell-cell distances as a binary classifier to detect same-clone/different-clone cell-pairs[138]) up to AF values of ~0.02. At higher AF values, cell-cell distances are less and less discriminative. Thus, differently from other GBC association metrics (i.e., NMI, ARI), the AUPRC metric is extremely sensible to all losses of MT-SNVs, even the ones that might be confounded with errors due to extremely low AF of detection (i.e., AF<0.02 in at least 2-3 cells). In any case, progressively higher AF thresholds are mirrored by a dramatic loss of cells retained for lineage inference, especially for MDA_PT (**Fig. 20**).

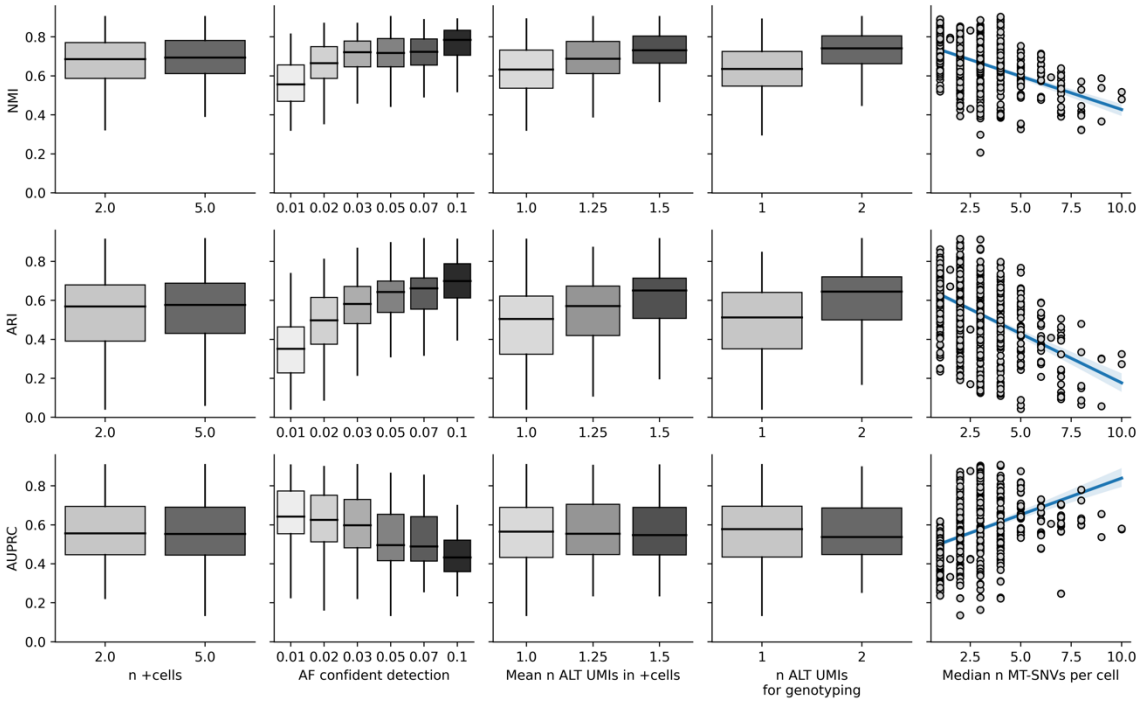
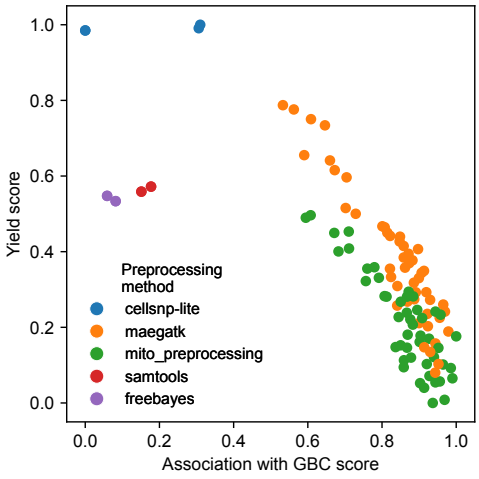


Fig. 21*. GBC association metrics and MT-SNVs filtering/genotyping options* *(Methods).*

Considering pre-processing pipelines (**Fig. 22**), on average (i.e., across samples) *maegatk* and its upgraded version mito*_preprocessing* outperformed the other tools. Interestingly, more polished basecalls (i.e., *mito_preprocessing* vs *maegatk)* resulted in slightly better

clonal reconstructions, worse cellular yield, fewer MT-SNVs and putative artifacts (dbSNP flagged common variants or REDIdb annotated RNA editing events), but higher ratios of transitions vs transversions. These data suggest that: i) the transitions *vs* transversion ratio by itself does not necessarily informs about the quality of a subset of MT-SNVs, as “low quality and consensus” basecalls still give MT-SNVs distributed according to the expected MT- molecular signature; ii) “low quality and consensus” basecalls might introduce errors that are unrecognizable from true MT-SNVs variants in terms of expected substitution pattern, and iii) “low quality and consensus” basecalls recorded by *maegatk* and not by mito_*preprocessing* does not add enough noise to confound the true biological signal in the data, and recover slightly more cells for downstream analysis.



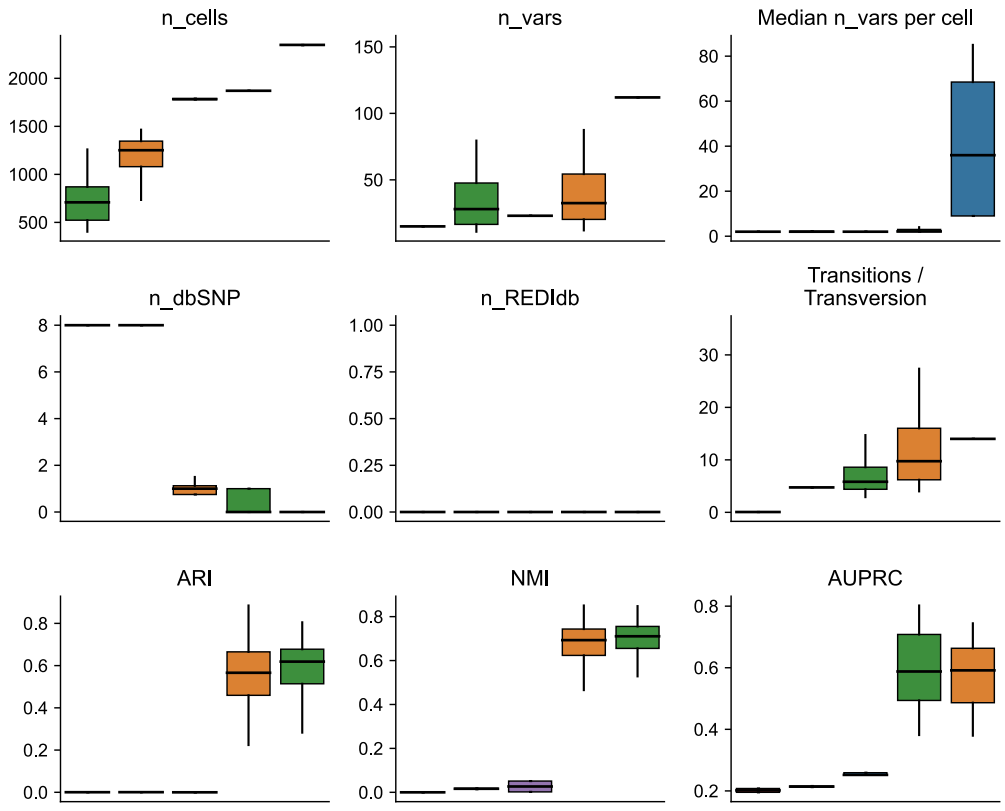


Fig. 22. ***Properties of MT-SNVs spaces from different pre-processing pipelines.*** *(Methods).*

We performed the same assessment for the choice of binarization method (**Fig. 23**). Here, the two tested strategies tested are: i) *vanilla*, i.e., simple thresholding on AF values and ALT alleles UMI counts, to assign binary genotypes (i.e., 1 MUT, 0 WT); and ii) the *MiTo* binarization method. For the development of the latter, we hypothesized that modeling the background distribution of observed ALT UMI counts could better discriminate between true positive and negative cells, especially for high-prevalence/low-AF MT-SNVs[173]. Based on this hypothesis, we re-adapted the model introduced by *Mquad*[159] to rank and select "informative" MT-SNVs. Specifically, *MQuad* uses Bayesian Information Criterion (BIC) differences (i.e., deltaBIC) between single and two-component binomial mixture models fitted on MT-SNVs AD and DP counts to prioritize MT-SNVs with higher statistical evidence of both negative and positive cell populations (i.e., the first and second components of the binomial mixture model, representing the background and the true positive population, respectively). Building upon this, we re-adapted the Bayesian inference scheme used by MQuad (Methods) to obtain, for each cell-MT-SNV combination, the two binomial-mixture-components cell assignment posterior probabilities. Thresholding on these posteriors, rather than directly using AF and AD values (i.e., *vanilla* method), produce binary genotypes (Methods) that can be used for lineage inference (Methods).

On average, *MiTo* produced more accurate clonal reconstructions than *vanilla genotyping* (higher ARI and NMI values), removing noise in genotype assignment (higher average CI of selected MT-SNVs), with minimal impact on the number of cells and MT-SNVs selected. *MiTo*-derived cell phylogenies displayed slightly lower correlation between tree and character based distances, and lower AUPRC.

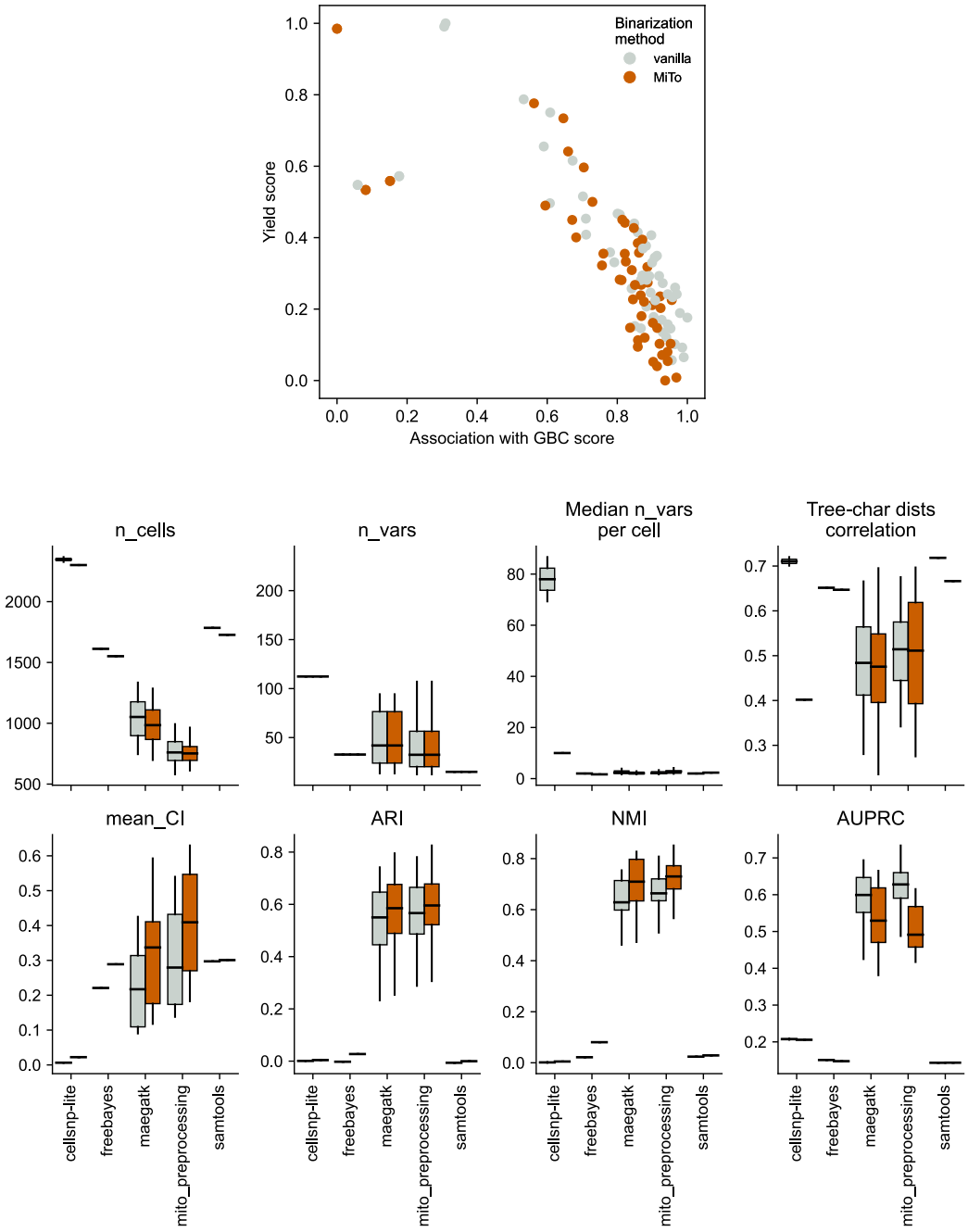


Fig. 23. ***Properties of MT-SNVs spaces from different binarization strategies.*** *(Methods)*

We conclude this meta-analysis by further inspecting relationships between “Association with GBC” score and other metrics of interest.

First, we assessed the relationship between GBC recovery and cell-cell “connectedness”, a concept introduced in Weng et al., 2024 that has been recently debated [173,174] (**Fig. 24**, Methods).

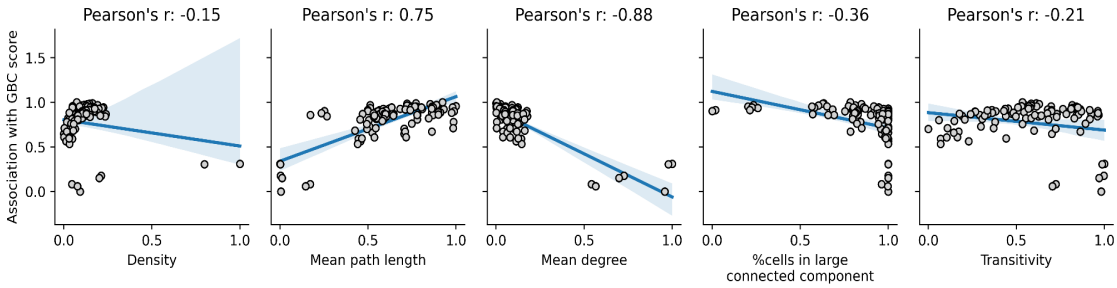
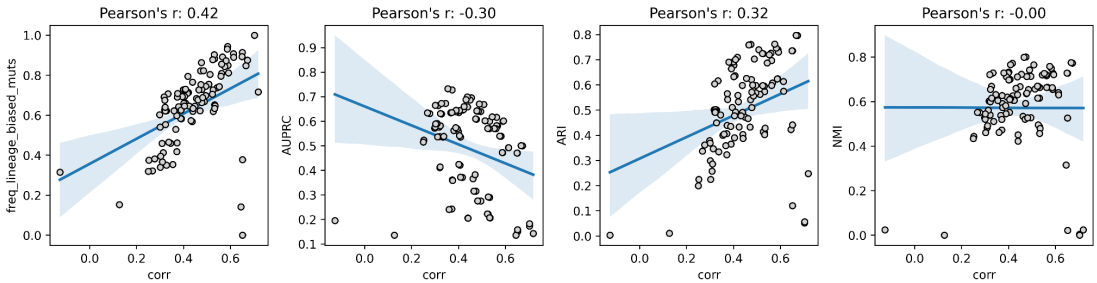


Fig. 24. ***Association between cell connectedness and lineage inference accuracy.*** *Cell-cell connectedness in MT-SNVs spaces and relationship with accuracy in clonal reconstruction.*

As observed for MDA_PT and MDA_lung, we found substantial anti-correlation between GBC recovery and metrics quantifying high cell-cell connectedness. This implies that overly connected MT-SNVs spaces either: i) selected False Positives MT-SNVs or ii) inaccurately assigned the alternative genotype to cells that did not share MT-SNVs, resulting in noisy and spurious connections. However, it is interesting to note how small increases in the density of the binarized AF matrix (0-0.25 density range) (**Fig. 24**) may give better association with GBC labels, suggesting that there is a small window of “right” connectedness that is actually beneficial for accurate clonal reconstruction.



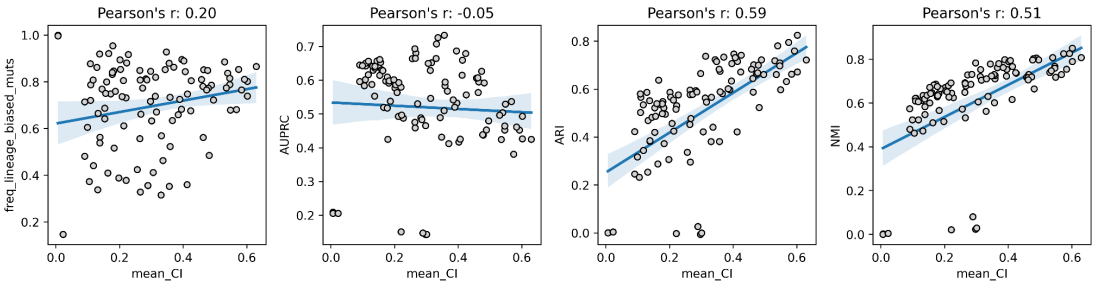


Fig. 25. Association between label- dependent and -independent metrics. *Relationship between "Tree structure" metrics and "GBC association" metrics.*

Second, we investigated relationships between GBC recovery and the two tree structural metrics included in MiTo benchmarking: i) the correlation between cell-cell tree- and character-based distances, which measures how much distances on an inferred tree (e.g., the number of nodes/edges connecting leaves) actually reflect disssimilarity among leaves, considering their characters; and ii) the Consistency Index (CI), which quantify the level of homoplasy and noise of the characters (Methods) used to reconstruct a tree. Crucially, good association between metrics explicitly using GBC labels (e.g., AUPRC, NMI, ARI, % of lineage-biased MT-SNVs) and metrics quantifying properties of "good" trees/cell distances independently of any ground truth (e.g., tree- vs character based distance correlation and CI) is fundamental for the identification of "informative" MT-SNVs spaces in real-world scenarios. Association between these 2 sets of metrics is not trivial: MT-based phylogenies are fundamentally different from standard trees in evolutionary biology and phylogenetics, as the ratio of available genetic characters vs phyla (i.e., the leaves on the tree) significantly smaller.

As it can be appreciated in **Fig. 25**, tree structural metrics are positively correlated with all GBC metrics except for AUPRC (which, as previously discussed, does not necessarily follow other GBC metrics). Thus, the correlation between character and tree cell-to-cell distances can be used effectively to prioritize MT-SNVs with higher phylogenetic signal, and, on the opposite, starting from a MT-SNVs space with enough phylogenetic signal, a simple baseline like Neighbor Joining can generate phylogenies that accurately represents the ground truth clonal structure of a population of cells, as we will demonstrate in the next chapter.

In summary, we discovered several peculiar properties of (expressed) MT-SNVs spaces. First and foremost, our analysis highlight remarkable variability between "informative" MT-SNVs spaces, showcasing the need for extensive data exploration and sample-specific analyses. In spite of this, common patterns can be found. The "sweet spot" for optimal MT-

SNVs selection and genotyping needs to balance a severe trade-off between lineage inference accuracy and cell recovery, and by far, hard thresholds controlling the sensibility of MT-SNVs detection events are the most important hyper-parameters that needs to be tuned to get “informative” MT-SNVs spaces. Our data suggests that the most “informative” MT-SNVs spaces allowing analyses of a reasonable number of cells include MT-SNVs detected with an allelic frequency of ~0.02-0.03 in at least 2-3 cells and at least 1.25-1.5 (mean) ALT UMIs in positive cells. Indeed, regardless of the pre-processing pipeline of choice, accepting single-molecule evidence of MT-SNVs presence rather than 2 or more UMIs leads to noisier lineage inferences, at least in our setting. Moreover, the optimal choice of pre-processing pipeline/genotyping strategy is heavily sample specific, and ultimately depends on clonal complexity and coverage: for high-coverage and low-complexity samples, more stringent pre-processing pipelines (i.e., *mito_preprocessing*) are preferred, in combinations with simple genotyping strategies (i.e., *vanilla*). On the contrary, more forgiving pre-processing pipelines (i.e., *maegatk*) are best suited to extract all available (but potentially dirtier) information from less covered and highly clonal samples. However, to achieve optimal results, more principled strategies for MT-SNVs genotyping (i.e., *MiTo*) are needed. Finally, our data demonstrate that “Informative” MT-SNVs spaces are not over-crowded with spurious cell-cell connections, but can benefit from carefully assigned alternative genotypes that increase the observed density and variation of character matrices. Indeed, caution is warranted when evaluating the quality of selected MT-SNVs: over-relying onto criteria that do not include confidence in molecular detection might include unwanted technical artifact, as “low quality and consensus” basecalls (and potentially other technical artifacts) might still give MT-SNVs that are distributed according to the expected C>T / T>C mutational signature.

MiTo: robust inference of mitochondrial phylogenies and clones

Inheritable molecular markers can trace historical events in evolving populations[14]. MT variants have been used to reconstruct species trees and infer population dynamics since decades[120]. Nuclear SNVs have been fundamental to trace the clonal evolution of cancer cell populations since the advent of NGS[79]. More recently, molecular recorders (e.g., Cas9-based evolving lineage tracers [88] have been engineered to recover division events in dividing cell populations with single-cell sequencing. Compared to other “static” markers (e.g., lentiviral barcoding), encoding flat, independent cell groups (i.e., cell clones), dynamically accumulating genetic variants can be used to infer cell phylogenies (i.e., cell trees, encoding the evolutionary relationships in a cell population)[113]. Recent scLT studies demonstrated the power of joint analysis of cell state and lineage, with “lineage” defined

as either cell clones or cell phylogenies [98]. However, the phylogeny of a cell population can be more informative than its cellular clones as: i) the topology of a cell tree can be used by itself to study the evolution of cell phenotypes with phylogenetics and phylodynamics methods[8], and ii) a cell tree can always be "cut" into discrete entities (i.e., cell clones), with the tree topology providing additional information about the evolutionary relationships between these cellular groups.

Direct analysis of cell state and lineage in human primary cell populations has great potential to build cell state-fate maps in normal development and disease. In [125], Ludwig and collegues validated for the first time the feasibility to trace cellular clones with expressed MT-SNVs using hierarchical clustering. Since then, several methods have been proposed to infer cellular clones from MT-SNVs (expressed or not) [161,169,170]. In a very recent work [142], Weng and colleagues built the first ever large-scale (>1000 cells) MT cell-phylogenies with a weighted jaccard distance and the Neighbor Joining algorithm. The authors provided compelling evidence for: i) the robustness to noise of cell-cell distances, ii) the visual support of specific tree sub-structures by MT-SNVs, and iii) the local correspondence between MT-based and Cas9-based cellular neighborhoods, assessed in an extremely elegant (and very debated henceforth [173,174]) scLT validation experiment. The authors also showed that, using their MT-SNVs filters, the vast majority of MT-SNVs detected by the MAESTER protocol is also detected by the RedeeM[142] protocol (based on MT-DNA enrichment, and detecting 5x more MT-SNVs). Here, we hypothesized that: i) "informative" subsets of expressed MT-SNVs (i.e., detected by the MAESTER protocol) could also be used to build robust cell phylogenies, and ii) that optimal "cutting" of these trees into discrete, MT-SNVs-supported clades could accurately represent discrete, ground-truth clonal structures. We developed the *MiTo* tree annotator (Methods). This tool takes an arbitrary cell phylogeny and a (binarized) character matrix (annotating tree leaves, i.e., cells) and:

1. Assigns each MT-SNV to a unique internal node of the tree, treating each internal node as a bipartition of the leaves;
2. Clusters MT-SNVs co-occurrence matrix into an optimal number of MT-SNVs clusters;
3. Uses MT-SNVs clusters and MT-SNV-internal node assignments to cut the tree into clades supported by MT-SNVs clusters (i.e., MT-clones)

Benchmark MT-phylogenies and *MiTo* clonal reconstruction performance, we took advantage of our *MiTo benchmarking* dataset. For each sample, we selected 10 "informative" MT-SNVs spaces (Methods) (**Fig. 26**). To make these assessments as fair

as possible, we selected MT-SNVs spaces with reasonable phylogenetic signal (regardless of the correspondence between *MiTo* clones and lentiviral clones described in the previous chapter) and variable numbers of MT-SNVs. Specifically, we filtered the MT-SNVs spaces: i) with reasonably high numbers of cells and clones; and ii) high AUPRC and character- vs tree-based cell-cell distances correlation (Methods). Then, we leveraged 4 different tree reconstruction algorithms to build cell phylogenies, and evaluated i) the structural support provided by MT-SNVs and ii) the robustness of tree individual clades to perturbation of original character matrices (i.e., bootstrapping).

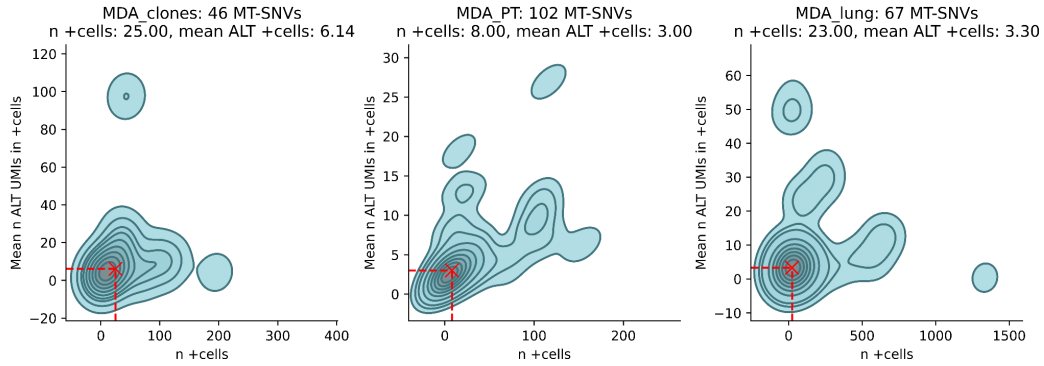


Fig. 26. Informative MT-SNVs spaces. *Cell prevalence and number of ALT alleles in positive cells of MT-SNVs selected to benchmark MiTo performance (n=10 MT-SNVs subsets for each sample).*

We selected 4 independent tree solvers with different working principles: two distance based-solvers (UPMGA and Neighbor Joining), commonly used for tree building in scLT[125,141,164], a parsimony based solver (mpboot[166]), and a maximum likelihood solver (iqtree[176]), more general purpose algorithms that have been widely used for species tree and/or single-colonies tre building. On average (**Fig. 27**), all solvers produce trees with a similar average depth, number of MT-clones and number of MT-SNVs assigned to each clone (Methods). However, distance-based solvers produced trees with higher character support (i.e., higher correlation between tree- and character- based cell-cell distances, higher Consistency Index) and much more robust to bootstrapping, with UPMGA showing median bootstrap support (Transfer Bootstrap Expectation, TBE) >0.7 considering all clades, only the largest ones (top 5 percentile) or the ones with MT-SNVs assigned.

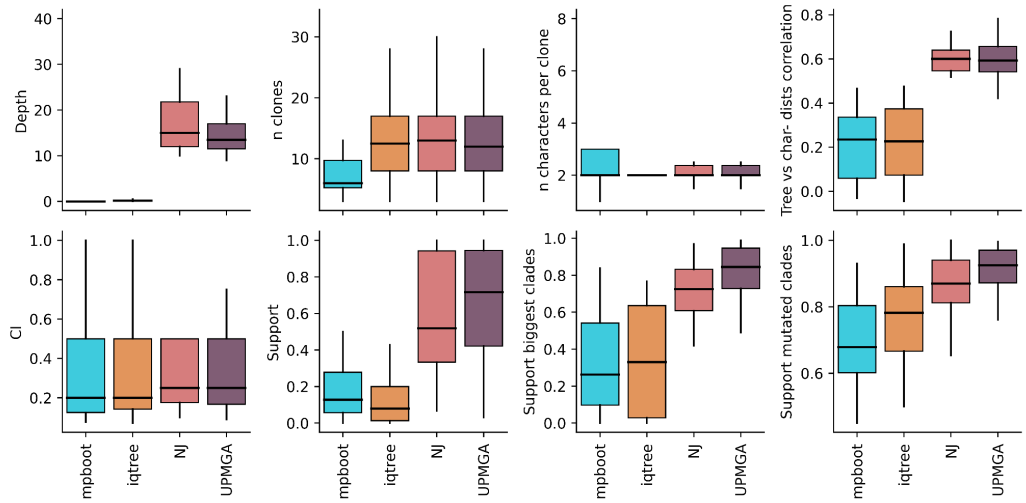


Fig. 27. Properties of mitochondrial phylogenies*. Cell prevalence and number of ALT alleles in positive cells of MT-SNVs selected to benchmark MiTo performance.*

Then, we benchmarked *MiTo* clonal reconstruction performance. We tested *MiTo* against 3 other clonal reconstruction methods: i) *leiden*[169] clustering, a fast community-detection algorithm with wide-spread use in single-cell genomics, ii) *vireoSNP*[170], a specialized Bayesian clustering methods developed for scRNA-seq demultiplexing and re-adapted for MT-clones inference; iii) *CClone*[161] (Methods) a weighted Non-Negative Matrix Factorization method recently introduced for MT-clones inference. Notably, these methods output either discrete clonal labels (i.e., *leiden*, *CClone*) or clone assignment probabilities (*vireoSNP*) that can be subsequently converted into discrete clonal labels. None of these methods output single-cell phylogenies, or give any additional information about the evolutionary relationship between cells and inferred MT-clones. We fed these methods with selected MT-SNVs spaces and tuned key hyperparameters of each method (i.e., the resolution for *leiden* clustering and the number of target clones, k, for *vireoSNP* and *CClone*) to guarantee their optimal performance (Methods). We used ARI and NMI to measure correspondance between inferred clonal labels and ground truth lentiviral clones (**Fig. 28, top**). Strikingly, *vireoSNP* and *MiTo* (i.e., MT-clones cut from *UPMGA*, *NJ*, *mpboot* and *iqtree* trees) outperformed all the other algorithms. Considering both ARI and NMI, *vireoSNP* placed 1st in low-complexity samples (i.e., *MDA_clones* and *MDA_lung,* 7-9 lentiviral clones, respectively), while *MiTo* ranked 1st with higher number of clones and cells (i.e., MDA PT clones, ARI and NMI >=0.75 with ~30 clones). Remarkably, in this latter scenario all alternative tools were unable to solve a consistent number of ground truth clones (**Fig. 28, bottom**).

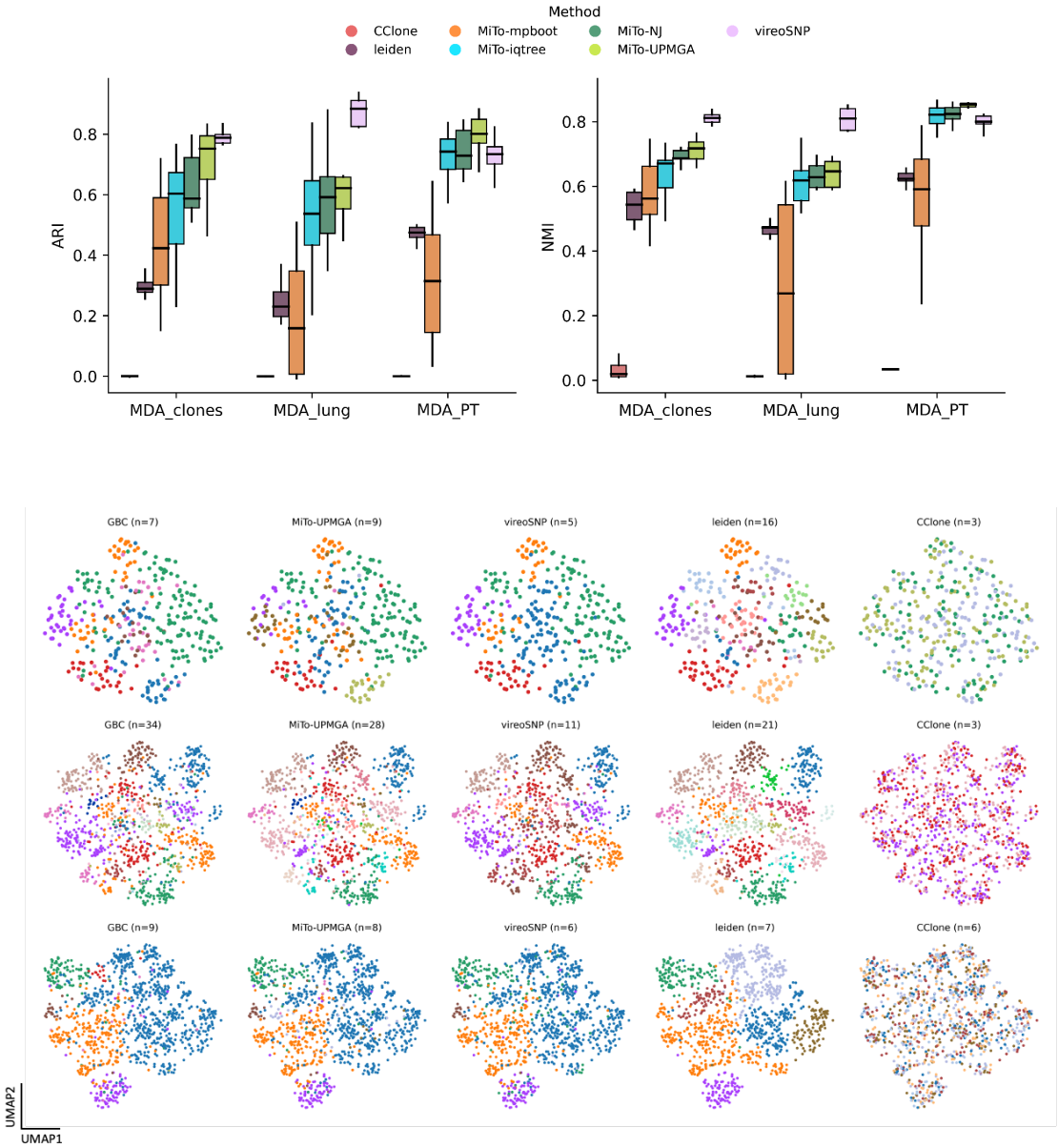


Fig. 28. Clonal reconstruction benchmarking. *Top panel: ARI and NMI of 10 lineage inference tasks (i.e., informative MT-SNVs spaces) per sample, for 7 alternative methods. Bottom panel: UMAP visualization of the most informative (i.e., highest average ARI across methods) MT-SNVs space for each sample (rows are MDA_clones, MDA_PT, MDA_lung, respectively). For each row, the ground truth lentiviral annotation is represented in the first column (dots represent cells, and the color-coding refers to individual GBC clones). All the other columns represent the respective inferences. For each reconstruction, n refers to the number of predicted clones.*

In the following paragraphs, we will illustrate properties of representative MT-SNVs spaces, MiTo inferred-clones and phylogenies for each *MiTo benchmarking* sample.

Fig. 29 shows essential properties of a representative MT-SNVs space (314 cells x 18 MT-SNVs and 7 lentiviral clones) for MDA_clones the low-complexity *in vitro* clonal-mixture

sample. In the first row (first two columns), we can appreciate the AF spectrum and molecular detection evidence for selected vs all, unfiltered MT-SNVs. On average, each of these MT-SNVs mark ~30 cells with ~5 UMIs.

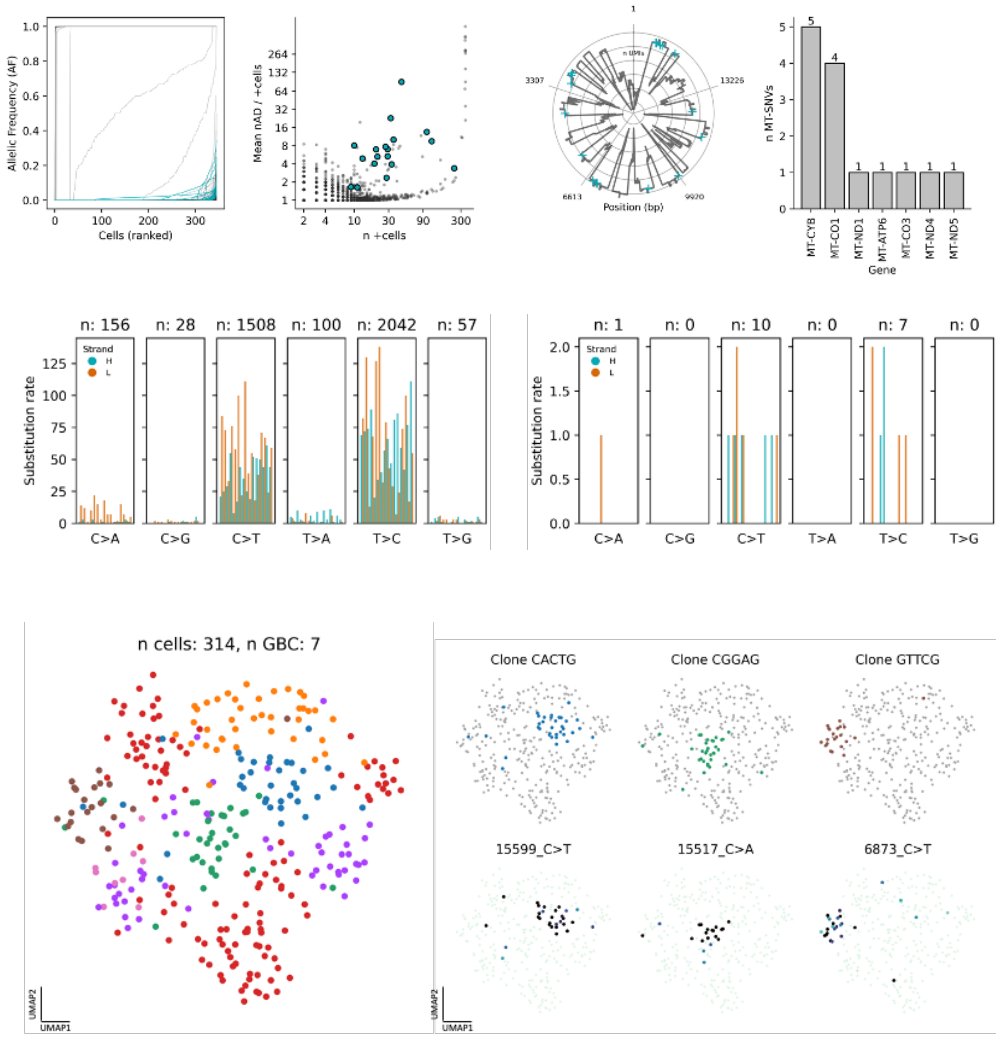
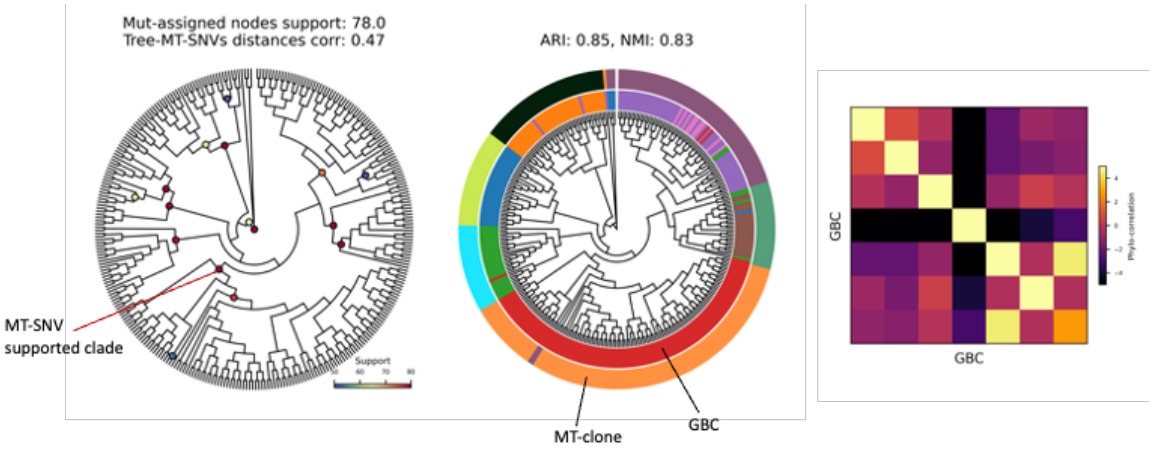


Fig. 29. Representative MDA_clones MT-SNV space.

These variants are scattered across the MT-genome (**Fig. 29** first row, columns three and four) with an enrichment (5-18) for the MT-CYB gene. The mutational signature of selected MT-SNVs includes 17/18 C>T/T>C substitutions (**Fig. 29** second row, second column) as expected [137,141], but, as discussed previously, the very permissive baseline filtering of MT-SNVs (**Fig. 29** second row, first column)) yields thousands of MT-SNVs enriched for the same substitution patterns. **Fig. 29** (i.e., third row) shows UMAP plots colored by covariates of interest. Specifically, we build a kNN graph (k=15) from the complete cell-cell pair-wise jaccard distance matrix, and embedded this graph in a two-dimensional space with the UMAP method[163] (Methods). These visualizations show clear separation of lentiviral labels within this MT-SNVs

space, and specificity and sensitivity of representative MT-SNVs as ground-truth clonal markers.

Fig. 30 shows a visualization of the main outputs from the *phylo_inference* pipeline. The first and second column represent the Neighbor Joining cell tree with internal nodes dots marking MT-SNV-supported clades (colored by Transfer Bootstrap Expectation support, TBE, Methods) and colorstrips on leaves annotating ground-truth GBC and inferred MT-clonal assignments. Here we can diagnose several properties of the cell tree and derived cellular clones: i) MT-SNVs assigned clades are very robust to bootstrapping (median TBE=78); ii) the topology of the tree is fairly supported by characters of the leaves (distances correlations=0.47); and iii) there is an almost perfect one-to-one mapping between ground truth and MiTo inferred cellular clones (ARI=0.85 and NMI 0.83). As an orthogonal diagnostic metric to evaluate significant association between ground truth cell labels and the tree structure, the *phylo_inference* pipeline computes *PATH*[168] phylogenetic correlations, (**Fig. 30**, third column, Methods) a bivariate derivation of Moran's I statistic to quantify plasticity vs heritability of arbitrary cell phenotypes on a phylogeny. If we assume (see Discussion) that GBC labels are strongly inheritable and mutually independent, we should expect to observe very high auto-correlations (diagonal values) and very weak cross-correlations (off-diagonal values), as we observe in this case. High off-diagonal entries may represent either noise



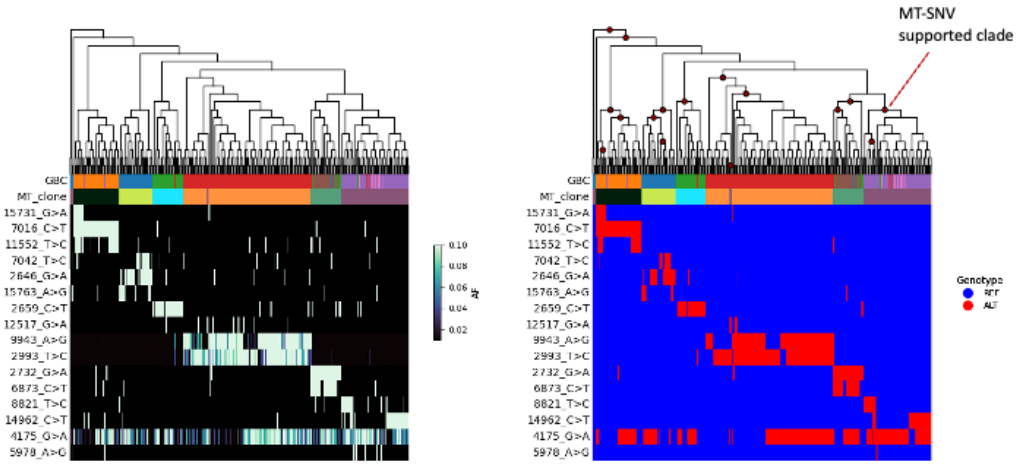
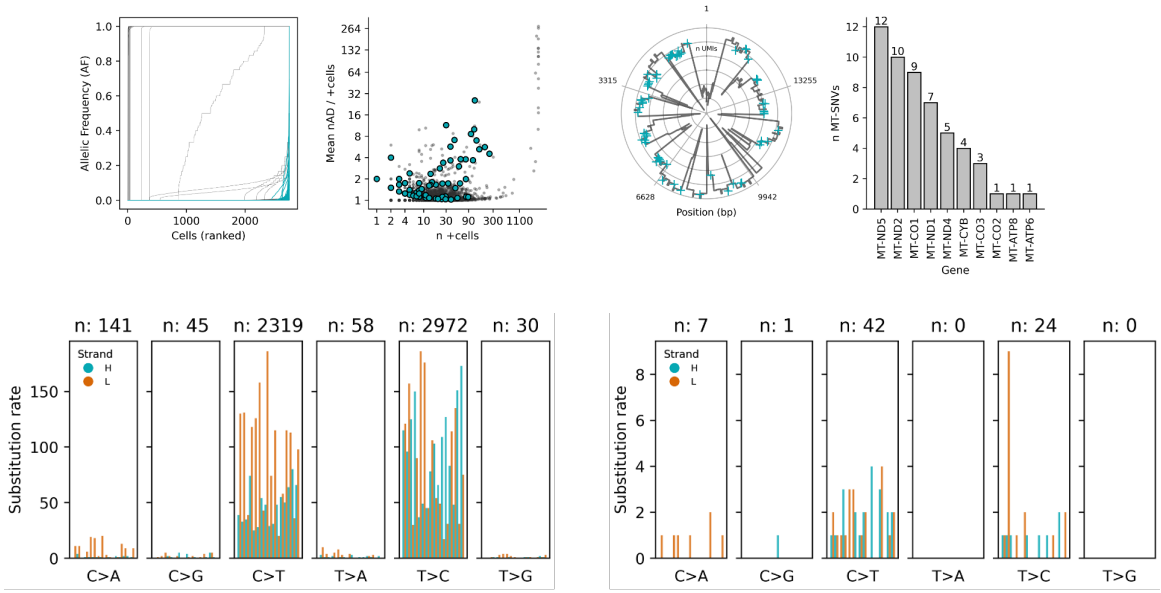


Fig. 30. Representative MDA_clones MiTo phylogeny and clones.

in lineage inference or true evolutionary relationships between lentiviral clones that can be further investigated using the cell tree and its characters.

Fig. 31 and **32** show the same properties of an informative MT-SNVs space (1220 cells x 74 MT-SNVs and 36 lentiviral clones) and its associated cell phylogeny and MiTo clones for the high-complexity MDA_PT sample.



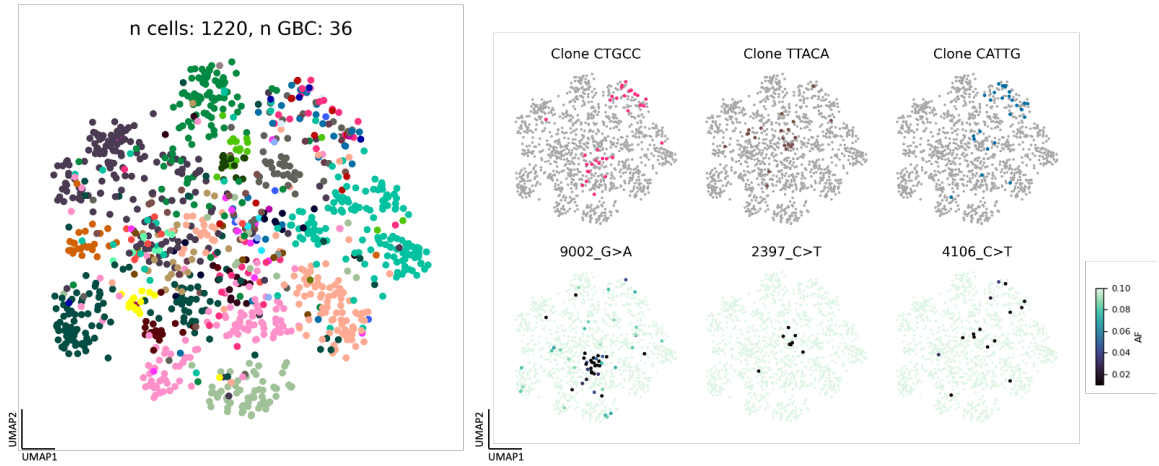
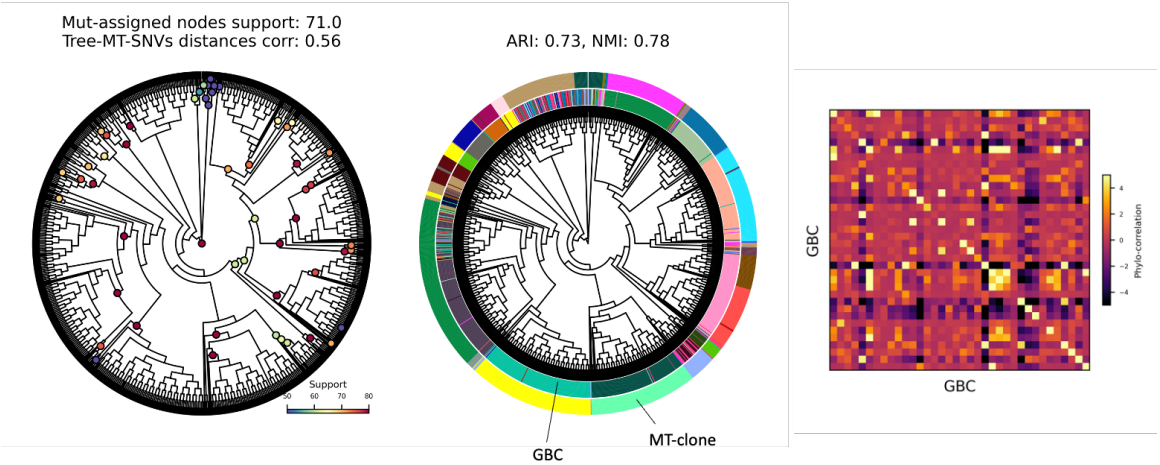


Fig. 31. Representative MDA_PT MT-SNV space.

The same trends observed for MDA_clones clonal mixture can be appreciated here, with some differences. MDA_PT "informative" MT-SNVs are more sensibly detected (mean number of ALT UMIs per positive cell <2, on average, compared MDA_clones), and are enriched for MT-ND5 gene. Importantly, high-support and MT-SNV-assigned tree clades identify ground truth clones. All major lentiviral clones and all MT-SNV-identifiable lentiviral clones (even small ones) were accurately inferred as robust, MT-SNV-assigned tree clades by MiTo. Conversely, small lentiviral clones without exclusive MT-SNVs clustered either: i) within other robust, MT-SNV-assigned tree clades or ii) as separate, noisy clades that could be easily spotted examining tree clades and their characters (**Fig. 32**). This pattern can be diagnosed even with PATH phylogenetic correlations, where a small fraction of ground truth clones did not show higher auto-than-cross-correlation values.



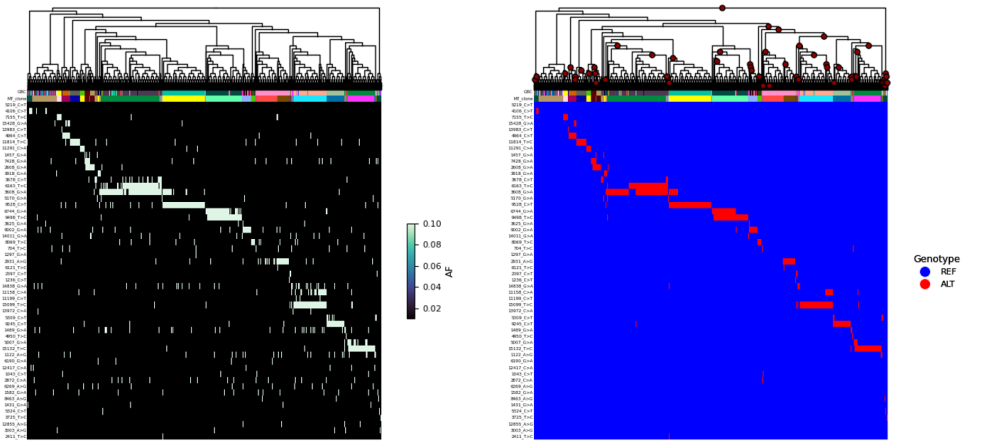
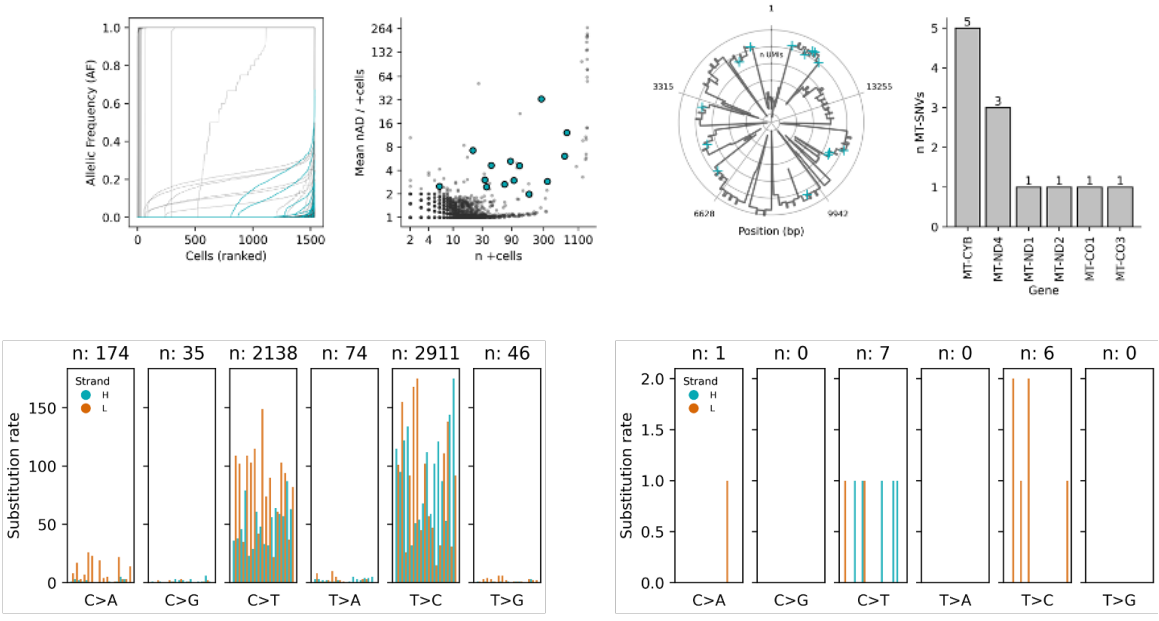


Fig. 32. Representative MDA_PT MiTo phylogeny and clones.

The last sample of our MiTo benchmarking dataset (**Fig. 33** and **34**) represents and intermediate-complexity sample with respect tothe two presented above. Here, our ground-truth clonal analyses showed the presence of relatively-low numebr of clones with remarkably different metastatization potential and clonal prevalence, as previously reported for metastatic clones in Breast Cancer[102,103]. Consistently, we found fewer and higher prevalence MT-SNVs compared to the matched primary tumor lesion.



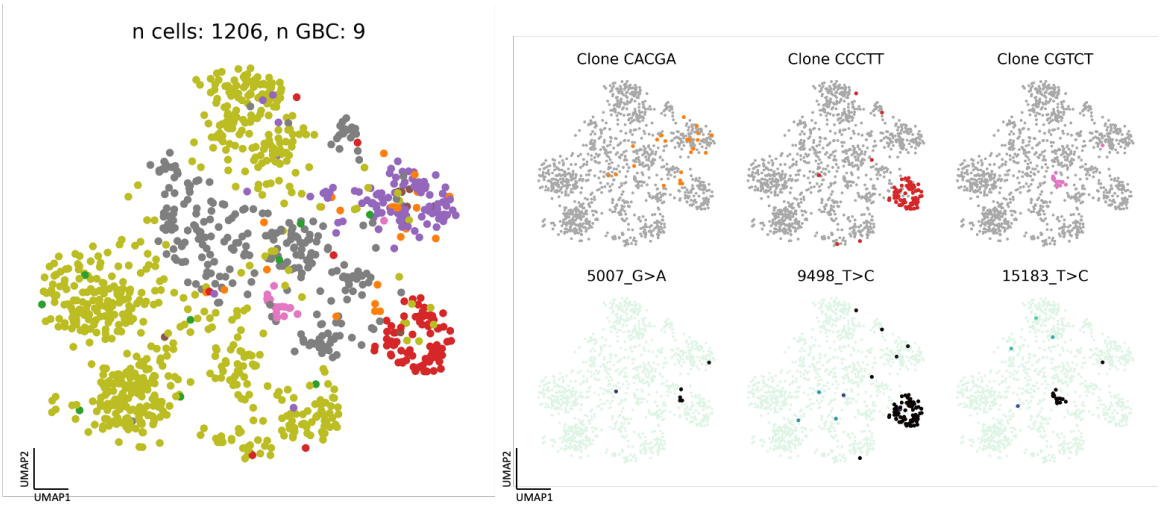


Fig. 33. Representative MDA_lung MT-SNV space.

The top3 most expanded clones, representing > 80% of the total cells, can be accurately identified with associated marker MT-SNVs, their assigned clades and respective *MiTo* clones, while this is much more challenging for very rare lentiviral clones.

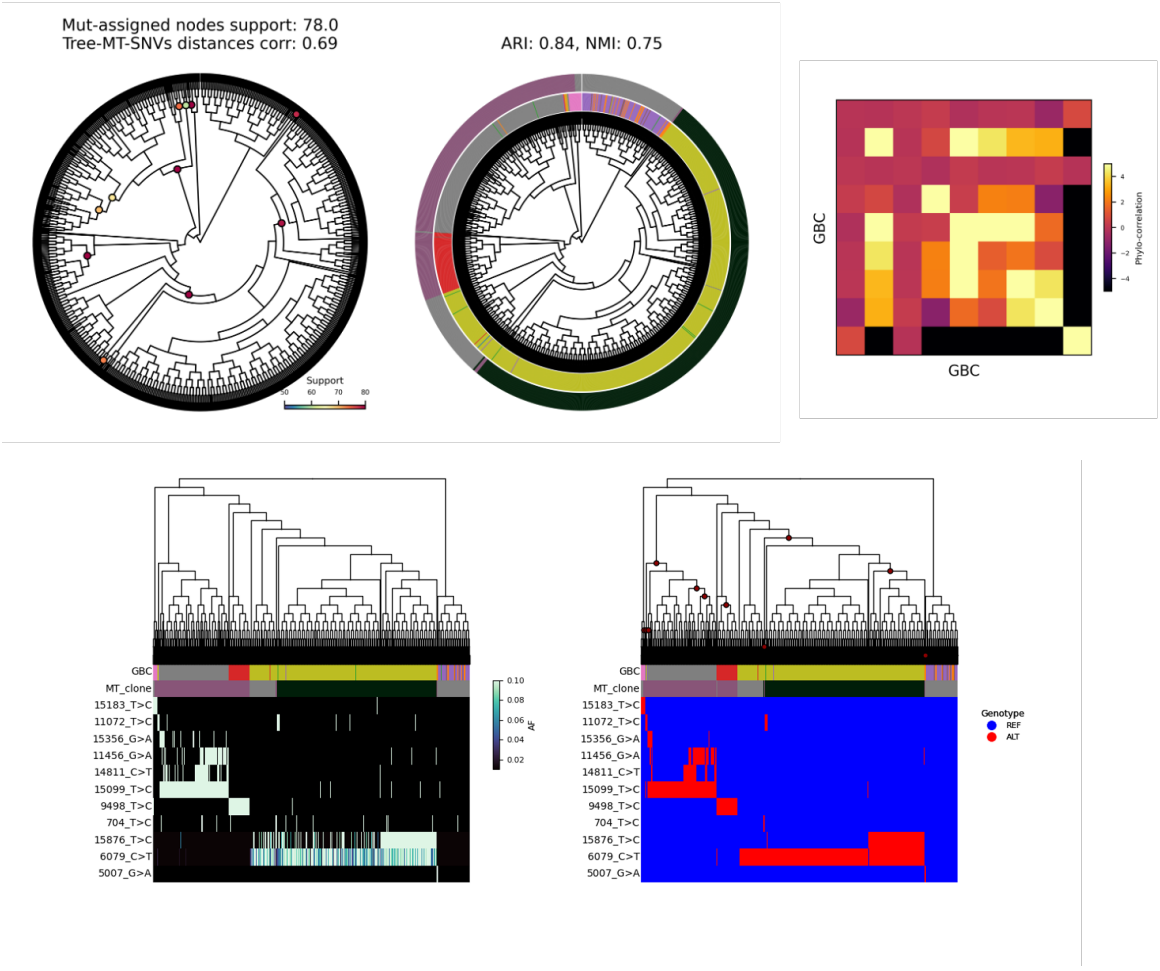


Fig. 34. Representative MDA_lung phylogeny and MiTo clones.

Together, these data demonstrates: i) feasibility of robust cell phylogeny inference from expressed MT-SNVs, and ii) accurate clonal reconstruction performance achieved by *vireoSNP* and *MiTo* compared to other state-of-the art lineage inference methods, with *MiTo* being particularly accurate for large phylogenies and high-complexity clonal structures. Importantly, *MiTo* is very simple and fast, and is general enough to accomodate any kind of tree reconstruction method and data source (any kind of tree and binary character matrix on the leaves can be used as input). Moreover, *MiTo* annotations can be used together with other tree diagnostic metrics to identify poorly solved, noisy and/or unsupported regions of a cell phylogeny, while to get additional information on the evolutionary relationships of individual cells and clones. Thus, in spite of inherent limitations from MT-SNVs dynamics (see the next chapter, and the Discussion), our toolkit facilitate the retrospective study of evolving cellular phenotypes leveraging expressed MT-SNVs as natural and cost-effective lineage marker.

Longitudinal dynamics of MT-SNVs assessed by lentiviral single-cell lineage tracing

Despite extensive experimental- and simulation-based evidence of MT-SNVs dynamics[125,142], the mutation rate and stability of MT-SNVs inheritance at cell division is still under active debate. Recent works produced contrasting data on this matter. On one hand, Campbell et collegues [143], demonstrated through elegant stochastic simulations that MT-SNVs can be used to trace ground-truth lineages only for relatively brief periods of time (2-5 yrs) and only if the AF of these MT-SNVs starts relatively high (>0.1). Consistently, Wang and collegues[177] demonstrated how, given the currently expected MT-genome mutation rate, much of the observed somatic variation in MT-genome haplotypes within a cell population is generated in very long periods with rare and sporadic mutational events. On the other hand, Weng and collegues [142] observed close agreement between cellular nearest neighbors in MT-SNVs and Cas9-induced INDELs spaces. Importantly, this local similarity implies that, over the course of an *in vivo* experiment (i.e., ~4-8 months, according to the tumor growth kinetics of the KP-Tracer mice[178] model) *newly generated* MT-SNVs co-evolved with INDELs artificially inserted in the nuclear genome through fast and continuous Cas9-induced DNA dsbreak and repair. Since the range of applications of an evolving lineage marker is dictated by the its mutation rate[146], resolving this apparent

contrast is of pivotal importance to establish safe and reliable use cases for mitochondrial scLT. Here, we leveraged our single-cell lineage traced longitudinal dataset to investigate short-period MT-SNVs dynamics (i.e., approximately one month, Methods) in individual breast cancer clones, *in vivo* **(Fig. 35**).

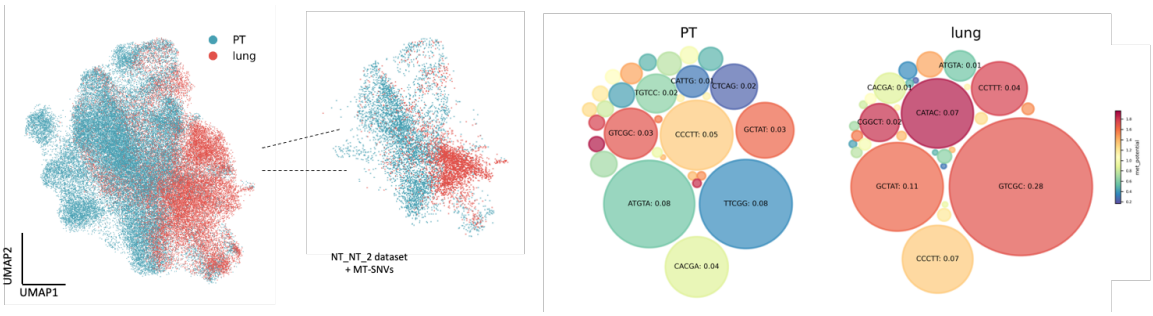


Fig. 35. In vivo breast cancer clonal dynamics assessed by lentiviral scLT. *Longitudinal breast cancer clones sampled at PT and lung sites, annotated for their prevalence and metastatic potential (Methods).*

As previously mentioned, our longitudinal dataset consists of clones with very heterogeneous cellular prevalence and metastatization potential (i.e., the lung prevalence of a given cellular clone, normalized for its PT prevalence). To maximize reliability, we focused on 6 lentiviral clones for which >10 cells were detected at both PT and lung sites. **Fig. 36** shows the evolution of 22 clonally-enriched MT-SNVs at two clonal sampling timepoint (i.e., PT, ~40 days after cellular barcoding, and lung, ~70 days of barcoding, Methods).

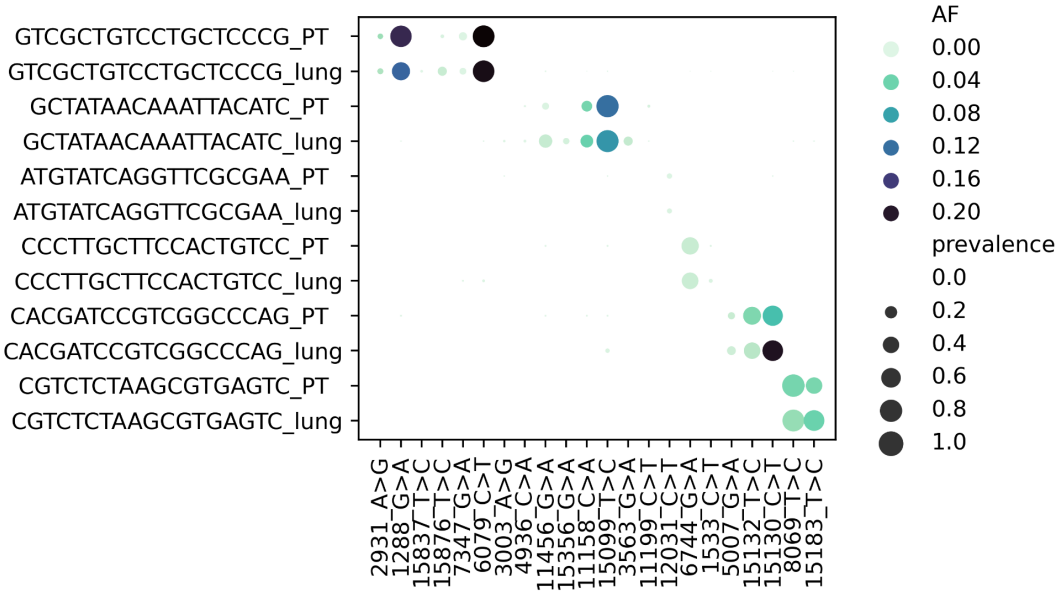


Fig. 36. MT-SNVs dynamics in 6 longitudinal Breast Cancer clones, in vivo *(Methods).*

Each lentiviral clone showed 1-6 clonally enriched MT-SNVs, with larger clones (n.d.r., clones are ordered in descending order for their mean number of cells across timepoints) showing more MT-SNVs than smaller ones. Strikingly, we detected 18 out of 22 of these clonally enriched MT-SNVs: i) at both timepoints, and ii) with very similar prevalence, regardless of their rarity. All of these mutations were detected almost exclusively in a single lentiviral clone, with nearly absent detection outside of it. Only 3 out of 22 MT-SNVs were detected in the metastatic population of an individual clone, but not in its primary tumor counterpart. These newly acquired MT-SNVs have moderate-to-low prevalence and mean AF. On the contrary, even more strikingly, we did not find any evidence of MT-SNVs loss across the same clone, from PT to metastasis. Importantly, these observations imply that: i) most clone specific MT-SNVs were generated *before* cellular barcoding, ii) in spite of the non-mendelian genetics of MT-SNVs, inheritance of these genetic trait at cell division is extremely stable, iii) the variation in MT-haplotypes that is generated within a cell population in such a short time span is very limited, and iii) all (MT-SNVs-) identifiable sub-clones within a pro-metastatic clone contribute to metastatic seeding (regardless of the underlying kinetics and directionality of the process).

Importantly, relaxing thresholds of clonal enrichment statistical significance did not alter these results.

Then, we looked for MT-SNVs selected at the PT or metastatic level, considering also non-clonally enriched MT-SNVs (i.e., MT-SNVs detected in multiple, independent lentiviral clones).

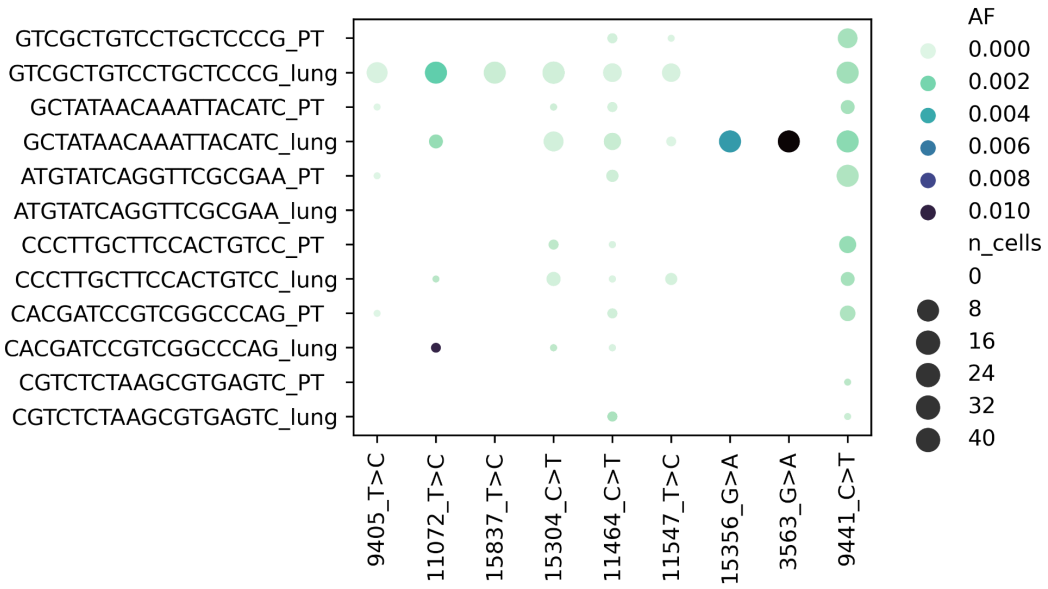


Fig. 37. Selection of MT-SNVs across 6 longitudinal Breast Cancer clones, in vivo *(Methods).*

Here, we detected n=9 MT-SNVs (out of n=121 across the joint MDA_PT/MDA_lung dataset) with uneven PT-lung detection: 3 MT-SNVs (e.g., 15356_G>A, 3563_G>A, 15837_T>C) were unique somatic events within single lentiviral clones (see also Fig., …); 2 MT-SNVs were detected in all clones and timpoints, with only sporadic exceptions (e.g., 9441_C>T, 11464_C>T); while 2 MT=SNVs were detected exclusively (11072_C>T) or predominantly (9305_T>C, 11547_T>C) at PT or lung sites. In spite of the very low mean AF and very low number of detection events requiring cautious interpretation, these data suggest how specific MT-SNVs might confer selective (dis-)advantage across breast cancer growth and dissemination.

Thus, coupling lentiviral scLT and MT-SNVs profiling demonstrated the reliability of MT-SNVs as lineage markers for longitudinally sampled cellular clones and revealed previously unappreciated features of Breast Cancer clonal dynamics.

Discussion

The terrific complexity of human body originates from a single, massive branching process [14,21]. This process, named somatic evolution, starts early with embryogenesis and continues through adulthood. In physiological conditions, stem cells respond to environmental stimuli to differentiate into specialized cells and replenish damaged tissues. Deviation from this normal trajectory may translate into diseases, such as ageing and cancer. Somatic evolution has been studied for centuries, with researchers inherently limited by the technological possibilities of their time. With the advent of Next Generation Sequencing, -omics technologies, and single-cell biology, cellular phenotypes have been abstracted into high-dimensional "states", i.e., the integrated functioning of cell constituent molecules, profiled by unbiased sequencing assays[29,75]. This data-driven revolution renewed interests in cell state-fate relationships and predictive models of cellular dynamics, both in healthy and diseased contexts. Indeed, single-cell lineage tracing (scLT) has emerged as a powerful technique to map cellular decision making, adding the "lineage" dimension to otherwise static single-cell high-dimensional "snapshots"[88,91,98]. In particular, remarkable efforts have been devoted to find lineage markers that could inform about cellular ancestries in their native environment (i.e., primary tissues). Among these markers, mitochondrial variants (MT-SNVs) have recently gained special attention, due to their high accumulation rates, scalability, and compatibility with other state-informative single-cell modalities[125,136]. In spite of their peculiar inheritance mode, MT-SNVs have been used to trace species evolution for decades[126]. However, due to fast improvements in the protocols and computational strategies to detect these variants, it is presently unclear what is the role of MT-SNVs in the realm of single-cell biology[137,149]. Specifically, we still do not know whether these variants can solve only cellular "clones" (i.e., coarse-grained cellular ancestries, identifying cells with "recent" common ancestors), or complete cell phylogenies (i.e., cell trees, graphical models approximating the true sequence of cellular division events giving rise to the observed cell population). Remarkably, while both types of information would provide valuable context to other phenotypic data, the cell phylogeny paradigm would allow repurposing of well-established phylodynamics methods to single-cell studies[8], opening entirely new possibilities for the study of somatic evolution.

In this work, we developed *MiTo*, a set of tools to facilitate automatic and interactive exploration of MT-SNVs data. This toolkit include: i) a flexible pre-processing pipeline for single-cell multi-omic datasets including the (expressed) MT-SNVs modality; ii) a comprehensive pipeline for MT-SNVs-based lineage inference; and iii) a *python* package for interactive exploration of MT-SNVs, phylogenies and clones. This toolkit streamlines a number of operations from raw *.fastqs* to annotated cell phylogenies, filling an important gap in the scLT community (particularly for *python* users, for which no off-the-shelf solution is available). Importantly, *MiTo* builds upon the *AnnData*[155] and *CassiopeiaTree*[165] data structures, enabling straightforward interoperability with other popular single-cell libraries from the *scverse*[179] ecosystem. Compared to existing approaches, MiTo includes refined data pre-processing (e.g., UMI-based consensus sequence generation for lentiviral- and MT- reads), a statistically sound method for MT-SNVs genotyping, and a tree "cutter" to annotate a phylogeny into MT-SNVs-supported clades. Importantly, these novel features are implemented alongside state-of-the-art tools to facilitate robust benchmarking.

To systematically assess the phylogenetic signal associated with expressed MT-SNVs and benchmark *MiTo* performance, we generated a new single-cell multi-omic dataset with simultaneous profiling of gene expression, expressed MT-SNVs (MAESTER protocol), and ground truth clonal labels derived from lentiviral scLT. This tri-modal (i.e., gene expression, lentiviral barcodes, MT-SNVs) dataset encompass three samples with thousands (~5k) high-quality cells and widely different number of clones, faithfully representing real world lineage inference scenarios. To the best of our knowledge, only two other (comparable) datasets were available prior to this work: i) the dataset published by Ludwig et al. in [125], including two samples from the TF1 cell line, with limited number of cells (70 and 158), clones (3 and 11, identified with lentiviral barcoding), and "informative" variants (9 and 20, detected with the poorly scalable Smart-seq scRNA-seq protocol, Fig. 3 and Supp Fig. 2); and ii) the dataset published earlier this year by Weng et al. [142]. This dataset was generated with a dual scLT experiment leveraging the KP-tracer mice[165] (a TP53-driven lung cancer mouse model, engineered for dynamic, Cas9-based scLT), and comprise 10 multi-modal (i.e., RNA, chromatin accessibility, MT-SNVs and CRISPR-induced INDELS) samples. Despite the unprecedented richness of these scLT read-outs, these samples show low average number of cells (~500 per sample, Supp Fig. 3 and Extended data Fig.5) and "informative" variants (17-40, Supp Fig.2). We believe that, since all of these datasets have unique features (e.g., single-cell protocol and layer of choice for MT-SNVs detection, organism, orthogonal scLT labels, etc.), they all provide fundamental reference for new scLT method development. However, the *MiTo benchmarking* dataset: i) includes 2 samples (i.e., MDA_PT and MDA_lung) with substantially higher number of cells, clones, and MT-SNVs, compared to all the other samples described; and ii) is the only one that

include samples derived from a clonal resampling experiment, where the same ground truth clones were profiled longitudinally, *in vivo*. While we acknowledge limitations of the *MiTo benchmarking* dataset (e.g., limited number of samples and static definition of ground truth lineage) these features enabled us to assess MT-SNVs properties that could not have been assessed otherwise.

Our analysis unfolds into three main phases.

In the first phase, we extensively explored alternative strategies for MT-SNVs pre-processing, cell and MT-SNVs filtering, and MT-SNVs genotyping, to evaluate their impact on lineage inference accuracy. To do this, we defined a set of metrics to evaluate different facets of resulting "MT-SNVs spaces" (i.e., a certain selection of cells and their MT-SNVs genotypes). With systematic meta-analyses of these metrics, we discovered several peculiar properties of (expressed) MT-SNVs spaces. For instance, we discovered that the most "informative" MT-SNVs spaces allowing analysis of reasonable cell numbers include MT-SNVs detected with an allelic frequency of ~0.02-0.03 in at least 2-3 cells and with at least 1.25-1.5 (mean) ALT UMIs in positive cells, with 2 UMIs better than 1 to accurate assignment of MT-SNVs alternative genotypes. Indeed, we found that the optimal choice of pre-processing pipeline/genotyping strategy is heavily sample specific, and ultimately depends on sample clonal complexity and coverage: for high-coverage and low-complexity samples, more stringent pre-processing pipelines (i.e., *mito_preprocessing*) are preferred, in combinations with simple genotyping strategies (i.e., *vanilla*). On the contrary, more forgiving pre-processing pipelines (i.e., *maegatk*[138]) are best suited to extract all available (but potentially dirtier) information from less covered and highly clonal samples. However, to achieve optimal results, more principled strategies for MT-SNVs genotyping (i.e., *MiTo*) are needed. Furthermore, our data suggest that "informative" MT-SNVs spaces are not over-crowded with spurious cell-cell connections, but can benefit from carefully assigned alternative genotypes that increase the observed density and variation of character matrices, and that "low quality and consensus" basecalls (and potentially other technical artifacts) might still give MT-SNVs that are distributed according to the expected C>T / T>C mutational signature.

In the second phase, we benchmarked MT-SNVs -derived cell phylogenies (MT-phylogenies) and cellular clones (MT-clones) from different lineage inference algorithms. Here, we found remarkable robustness to noise of MT-phylogenies reconstructed with simple, distance-based tree inference algorithms (i.e., Neighbors Joining and UPMGA), as measured by Transfer Bootstrap Expectations[167] (a modified version of classic Falsestein's Bootstrap Proportions that evaluates how similarly tree clades are found across bootstrap replicates, rather than quantifying how many times these clades are identically found across replicates). Consistently, we found that leveraging these tree structures to infer

discrete cellular clones (i.e., *MiTo* tree annotator) results in accurate clonal inferences, especially for high clonal complexity scenarios, where the challenge is to accurately detect the highest possible number of ground truth clones.

In the third phase, we leveraged the longitudinal nature of our dataset to assess: i) stability of MT-SNVs, and ii) *de novo* generation of MT-SNVs, in short time periods (i.e., ~1 month). Here we found that, without exceptions, all clonally-enriched MT-SNVs are stably inherited *within* ground truth clones resampled across timepoints, with nearly absent evidence of parallel evolution. Strikingly, this pattern was conserved for all enriched MT-SNVs, i.e., both high prevalence MT-SNVs (marking entire lentiviral clones) and low prevalence MT-SNVs (marking sub-populations *within* individual lentiviral clones), regardless of observed allelic frequencies. We also detected few (i.e., n=3, across 6 independent clones) putative *de novo* MT-SNVs, exclusively detected in the metastatic population of individual clones. We also detected interesting MT-SNVs that could mirror the action selective pressures on MT-phenotypes during tumor growth and metastatization cascade.

Together, these data need careful interpretation, considering recent evidence in the field. Our data suggest that, in some way, “less is more”, when it comes to MT-SNVs “informative” spaces. In other terms, too loose MT-SNVs genotyping produces False Positive variant calls leading to sub-optimal lineage inference. However, too strict MT-SNVs genotyping causes dramatic cell and MT-SNVs loss, preventing scLT analysis*, in toto*. Thus, according to these data, one should include all the MT-SNVs that is able to genotype, even low-AF MT-SNVs (i.e., 0.02-0.03 AF), but this should be done with a careful assessment of the confidence with which these MT-SNVs are detected, and possibly, with statistical methods that are able to discriminate between background, noisy detection events, and true positive signals. In spite of MT-SNVs detection differences (i.e., RNA vs DNA), this view reconcile the need for stringent filtering strategies adopted in [137,148,173] with the benefit of added phylogenetic characters demonstrated in [142] . Indeed, our data demonstrate high robustness to noise of MT-phylogenies, and their utility for MT-clone inference. The former property is difficult to compare with other scLT studies, as reporting single-cell phylogenies bootstrapping results is not common practice in scLT, and robustness-to-noise of cell-cell distances is much easier to achieve than robustness of inferred ancestries. However, compared to other phylogenies[47,142–144,166,180] (i.e., from species phylogenetics or single-cell colonies WGS), MT-phylogenies internal branches are supported by orders-of-magnitude less characters, and thus, the molecular evidence supporting these inferred ancestral events (and hence, the resolution of MT-phylogenies) remains limited, at least considering expressed MT-SNVs from the MAESTER protocol. Importantly, challenging MT-phylogenies from 10-fold higher number of MT-SNVs (i.e., RedeeM[142] protocol) could already provide better guarantees about MT-phylogenies

branch support, and molecular dating[8] methods could reveal additional opportunities and limitations of MT-scLT. Indeed, a detailed comparison with previously published scLT phylogenies (e.g., derived from scWGS or Cas9 lineage recorders) would give better perspective to the properties of MT-phylogenies and retrospective scLT in general. These analyses will be part of future work. Thus, despite inherent limitations, MT-SNVs resolve useful (albeit approximate) cellular genealogies that can be leveraged to study the evolution of cellular phenotypes, at reasonable costs, and in native contexts. We anticipate that further experimental and computational advancements will improve current MT-phylogenies resolution, reducing the temporal scale at which MT-SNVs can resolve somatic evolutionary events. Finally, results from our longitudinal experiment have three main implications: i) MT-SNVs are exceptionally stable across cell divisions, in spite of the known stochasticity of mtDNA inheritance; ii) the mutation rate of mtDNA is low, and newly generated MT-SNVs take time to drift at high AF in a substantial number of cells; iii) metastatic seeding of individual clones is sustained by multiple sub-clonal lineages rather than being restricted to one or few sub-clones selectively (or stochastically) acquiring the metastatic phenotype. Further analyses and data will be required to get more precise estimates of mtDNA mutation rate, but, qualitatively, these results are concordant with experiments and simulations from [143] and [177]. The interesting seeding pattern that we observed in Breast Cancer xenografts, instead, needs to be further corroborated with additional experimental evidence.

In summary, this work highlights opportunities and limitations of MT-scLT, providing novel data analysis tools and benchmarking datasets.

Our hope is that these methods and data will be useful to understand cancer somatic evolution, to intercept it, and ultimately, to counteract it.

References

1. Hanahan, D. (2022). Hallmarks of cancer: New dimensions. Cancer Discov. *12*, 31–46.

2. Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

3. Rabas, N., Ferreira, R.M.M., Di Blasio, S., and Malanchi, I. (2024). Cancer-induced systemic pre-conditioning of distant organs: building a niche for metastatic cells. Nat. Rev. Cancer. https://doi.org/10.1038/s41568-024-00752-0.

4. Harris, M.A., Savas, P., Virassamy, B., O'Malley, M.M.R., Kay, J., Mueller, S.N., Mackay, L.K., Salgado, R., and Loi, S. (2024). Towards targeting the breast cancer immune microenvironment. Nat. Rev. Cancer *24*, 554–577.

5. Vendramin, R., Litchfield, K., and Swanton, C. (2021). Cancer evolution: Darwin and beyond. EMBO J. *40*, e108389.

6. Ciriello, G., Magnani, L., Aitken, S.J., Akkari, L., Behjati, S., Hanahan, D., Landau, D.A., Lopez-Bigas, N., Lupiáñez, D.G., Marine, J.-C., et al. (2024). Cancer evolution: A multifaceted affair. Cancer Discov. *14*, 36–48.

7. Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

8. Parsons, B.L. (2018). Multiclonal tumor origin: Evidence and implications. Mutat. Res. Rev. Mutat. Res. *777*, 1–18.

9. Black, J.R.M., and McGranahan, N. (2021). Genetic and non-genetic clonal diversity in cancer evolution. Nat. Rev. Cancer *21*, 379–392.

10. Nam, A.S., Chaligne, R., and Landau, D.A. (2021). Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat. Rev. Genet. *22*, 3–18.

11. Caravagna, G. (2020). Measuring evolutionary cancer dynamics from genome sequencing, one patient at a time. Stat. Appl. Genet. Mol. Biol. *19*. https://doi.org/10.1515/sagmb-2020-0075.

12. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122–128.

13. Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science *349*, 1483–1489.

14. Sweet-Cordero, E.A., and Biegel, J.A. (2019). The genomic landscape of pediatric cancers: Implications for diagnosis and treatment. Science *363*, 1170–1175.

15. Hahn, W.C., Bader, J.S., Braun, T.P., Califano, A., Clemons, P.A., Druker, B.J., Ewald, A.J., Fu, H., Jagu, S., Kemp, C.J., et al. (2021). An expanded universe of cancer targets. Cell *184*, 1142–1155.

16. Gargiulo, G., Serresi, M., and Marine, J.-C. (2024). Cell states in cancer: Drivers, passengers, and trailers. Cancer Discov. *14*, 610–614.

17. Rambow, F., Marine, J.-C., and Goding, C.R. (2019). Melanoma plasticity and phenotypic diversity: therapeutic barriers and opportunities. Genes Dev. *33*, 1295–1318.

18. Whiting, F.J.H., Household, J., Baker, A.-M., Sottoriva, A., and Graham, T.A. (2024). Phenotypic noise and plasticity in cancer evolution. Trends Cell Biol. *34*, 451–464.

19. Sánchez Alvarado, A., and Yamanaka, S. (2014). Rethinking differentiation: stem cells, regeneration, and plasticity. Cell *157*, 110–119.

20. Waddington, C.H. (1942). Canalization of development and the inheritance of acquired characters. Nature *150*, 563–565.

21. Li, X., and Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing. Int. J. Oral Sci. *13*, 36.

22. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

23. Griffiths, J.A., Scialdone, A., and Marioni, J.C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. Mol. Syst. Biol. *14*, e8046.

24. Ginhoux, F., Yalin, A., Dutertre, C.A., and Amit, I. (2022). Single-cell immunology: Past, present, and future. Immunity *55*, 393–404.

25. He, X., Memczak, S., Qu, J., Belmonte, J.C.I., and Liu, G.-H. (2020). Single-cell omics in ageing: a young and growing field. Nat. Metab. *2*, 293–302.

26. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

27. Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. *50*, 1–14.

28. Chow, A., and Lareau, C.A. (2024). Concepts and new developments in droplet-based single cell multi-omics. Trends Biotechnol. https://doi.org/10.1016/j.tibtech.2024.07.006.

29. Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A.M., and Mazutis, L. (2017). Single-cell barcoding and sequencing using droplet microfluidics. Nat. Protoc. *12*, 44–73.

30. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

31. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

32. You, Y., Tian, L., Su, S., Dong, X., Jabbari, J.S., Hickey, P.F., and Ritchie, M.E. (2021). Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. Genome Biol. *22*, 339.

33. Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. Nat. Rev. Genet. *24*, 550–572.

34. Argelaguet, R., Cuomo, A.S.E., Stegle, O., and Marioni, J.C. (2021). Computational principles and challenges in single-cell data integration. Nat. Biotechnol. *39*, 1202–1215.

35. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50.

36. Clarke, Z.A., Andrews, T.S., Atif, J., Pouyabahar, D., Innes, B.T., MacParland, S.A., and Bader, G.D. (2021). Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat. Protoc. *16*, 2749–2764.

37. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554.

38. Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154.

39. Bunne, C., Stark, S.G., Gut, G., Del Castillo, J.S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. (2023). Learning single-cell perturbation responses using neural optimal transport. Nat. Methods *20*, 1759–1768.

40. Kamimoto, K., Stringa, B., Hoffmann, C.M., Jindal, K., Solnica-Krezel, L., and Morris, S.A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. Nature *614*, 742–751.

41. Wilk, A.J., Shalek, A.K., Holmes, S., and Blish, C.A. (2024). Comparative analysis of cell-cell communication at single-cell resolution. Nat. Biotechnol. *42*, 470–483.

42. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 31.

43. Biezuner, T., Raz, O., Amir, S., Milo, L., Adar, R., Fried, Y., Ainbinder, E., and Shapiro, E. (2021). Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. Sci. Rep. *11*, 17171.

44. Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. (2006). Whole-genome multiple displacement amplification from single cells. Nat. Protoc. *1*, 1965–1970.

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

46. Blagodatskikh, K.A., Kramarov, V.M., Barsova, E.V., Garkovenko, A.V., Shcherbo, D.S., Shelenkov, A.A., Ustinova, V.V., Tokarenko, M.R., Baker, S.C., Kramarova, T.V., et al. (2017). Improved DOP-PCR (iDOP-PCR): A robust and simple WGA method for efficient amplification of low copy number genomic DNA. PLoS One *12*, e0184507.

47. Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A.Y., Wang, T., and Vijg, J. (2017). Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat. Methods *14*, 491–493.

48. Zafar, H., Wang, Y., Nakhleh, L., Navin, N., and Chen, K. (2016). Monovar: single-nucleotide variant detection in single cells. Nat. Methods *13*, 505–507.

49. Bohrson, C.L., Barton, A.R., Lodato, M.A., Rodin, R.E., Luquette, L.J., Viswanadham, V.V., Gulhan, D.C., Cortés-Ciriano, I., Sherman, M.A., Kwon, M., et al. (2019). Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat. Genet. *51*, 749–754.

50. Hård, J., Al Hakim, E., Kindblom, M., Björklund, Å.K., Sennblad, B., Demirci, I., Paterlini, M., Reu, P., Borgström, E., Ståhl, P.L., et al. (2019). Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing. Genome Biol. *20*, 68.

51. Baysoy, A., Bai, Z., Satija, R., and Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. Nat. Rev. Mol. Cell Biol. *24*, 695–713.

52. Liu, L., Liu, C., Quintero, A., Wu, L., Yuan, Y., Wang, M., Cheng, M., Leng, L., Xu, L., Dong, G., et al. (2019). Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat. Commun. *10*, 470.

53. Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods *12*, 519–522.

54. Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

55. Rodriguez-Meira, A., Buck, G., Clark, S.-A., Povinelli, B.J., Alcolea, V., Louka, E., McGowan, S., Hamblin, A., Sousos, N., Barkas, N., et al. (2019). Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Mol. Cell *73*, 1292-1305.e8.

56. Han, K.Y., Kim, K.-T., Joung, J.-G., Son, D.-S., Kim, Y.J., Jo, A., Jeon, H.-J., Moon, H.-S., Yoo, C.E., Chung, W., et al. (2018). SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. Genome Res. *28*, 75–87.

57. Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat. Methods *13*, 229–232.

58. Xing, Q.R., Farran, C.A.E., Zeng, Y.Y., Yi, Y., Warrier, T., Gautam, P., Collins, J.J., Xu, J., Dröge, P., Koh, C.-G., et al. (2020). Parallel bimodal single-cell sequencing of transcriptome and chromatin accessibility. Genome Res. *30*, 1027–1039.

59. Chen, S., Lake, B.B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat. Biotechnol. *37*, 1452–1457.

60. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell *183*, 1103-1116.e20.

61. Swanson, E., Lord, C., Reading, J., Heubeck, A.T., Genge, P.C., Thomson, Z., Weiss, M.D.A., Li, X.-J., Savage, A.K., Green, R.R., et al. (2021). Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. Elife *10*. https://doi.org/10.7554/eLife.63632.

62. Zhang, B., Srivastava, A., Mimitou, E., Stuart, T., Raimondi, I., Hao, Y., Smibert, P., and Satija, R. (2022). Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. Nat. Biotechnol. *40*, 1220–1230.

63. Janssens, D.H., Otto, D.J., Meers, M.P., Setty, M., Ahmad, K., and Henikoff, S. (2022). CUT&Tag2for1: a modified method for simultaneous profiling of the accessible and silenced regulome in single cells. Genome Biol. *23*, 81.

64. Chang, L., Xie, Y., Taylor, B., Wang, Z., Sun, J., Armand, E.J., Mishra, S., Xu, J., Tastemel, M., Lie, A., et al. (2024). Droplet Hi-C enables scalable, single-cell profiling of chromatin architecture in heterogeneous tissues. Nat. Biotechnol. https://doi.org/10.1038/s41587-024-02447-1.

65. Saeys, Y., Van Gassen, S., and Lambrecht, B.N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat. Rev. Immunol. *16*, 449–462.

66. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868.

67. Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. *35*, 936–939.

68. Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzetto-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B.Z., Papalexi, E., et al. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat. Biotechnol. *39*, 1246–1258.

69. Chen, A.F., Parks, B., Kathiria, A.S., Ober-Reynolds, B., Goronzy, J.J., and Greenleaf, W.J. (2022). NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. Nat. Methods *19*, 547–553.

70. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853-1866.e17.

71. Song, L., Cohen, D., Ouyang, Z., Cao, Y., Hu, X., and Liu, X.S. (2021). TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nat. Methods *18*, 627–630.

72. Ainciburu, M., Morgan, D.M., DePasquale, E.A.K., Love, J.C., Prósper, F., and van Galen, P. (2022). WAT3R: recovery of T-cell receptor variable regions from 3' single-cell RNA-sequencing. Bioinformatics *38*, 3645–3647.

73. Vandereyken, K., Sifrim, A., Thienpont, B., and Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. Nat. Rev. Genet. *24*, 494–515.

74. Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nat. Biotechnol. *34*, 1145–1160.

75. Kretzschmar, K., and Watt, F.M. (2012). Lineage tracing. Cell *148*, 33–45.

76. Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. *100*, 64–119.

77. Temple, S. (1989). Division and differentiation of isolated CNS blast cells in microculture. Nature *340*, 471–473.

78. Schwartz, R., and Schäffer, A.A. (2017). The evolution of tumour phylogenetics: principles and practice. Nat. Rev. Genet. *18*, 213–229.

79. Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., and Schäffer, A.A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. J. Comput. Biol. *6*, 37–51.

80. Zhao, Z.-M., Zhao, B., Bai, Y., Iamarino, A., Gaffney, S.G., Schlessinger, J., Lifton, R.P., Rimm, D.L., and Townsend, J.P. (2016). Early and multiple origins of metastatic lineages within primary tumors. Proc. Natl. Acad. Sci. U. S. A. *113*, 2140–2145.

81. Pennington, G., Smith, C.A., Shackney, S., and Schwartz, R. (2007). Reconstructing tumor phylogenies from heterogeneous single-cell data. J. Bioinform. Comput. Biol. *5*, 407–427.

82. Zhang, J., Fujimoto, J., Zhang, J., Wedge, D.C., Song, X., Zhang, J., Seth, S., Chow, C.-W., Cao, Y., Gumbs, C., et al. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science *346*, 256–259.

83. Brocks, D., Assenov, Y., Minner, S., Bogatyrova, O., Simon, R., Koop, C., Oakes, C., Zucknick, M., Lipka, D.B., Weischenfeldt, J., et al. (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell Rep. *8*, 798–806.

84. Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, 886–895.

85. de Bruin, E.C., McGranahan, N., Mitter, R., Salm, M., Wedge, D.C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A.J., et al. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science *346*, 251–256.

86. Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science *294*, 2310–2314.

87. Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. Nat. Rev. Genet. *21*, 410–427.

88. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell *176*, 1517.

89. Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., and Klein, A.M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proc. Natl. Acad. Sci. U. S. A. *115*, E2467–E2476.

90. Haghverdi, L., and Ludwig, L.S. (2023). Single-cell multi-omics and lineage tracing to dissect cell fate decision-making. Stem Cell Reports *18*, 13–25.

91. Chen, W., Guillaume-Gentil, O., Rainer, P.Y., Gäbelein, C.G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., et al. (2022). Live-seq enables temporal transcriptomic recording of single cells. Nature *608*, 733–740.

92. Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods *13*, 845–848.

93. Cho, H., Ayers, K., DePills, L., Kuo, Y.-H., Park, J., Radunskaya, A., and Rockne, R. (2018). Modelling acute myeloid leukaemia in a continuum of differentiation states. Lett. Biomath. *5*, S69–S98.

94. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. Nat. Methods *19*, 159–170.

95. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

96. Mittnenzweig, M., Mayshar, Y., Cheng, S., Ben-Yair, R., Hadas, R., Rais, Y., Chomsky, E., Reines, N., Uzonyi, A., Lumerman, L., et al. (2021). A single-embryo, single-cell time-resolved model for mouse gastrulation. Cell *184*, 2825-2842.e22.

97. Sankaran, V.G., Weissman, J.S., and Zon, L.I. (2022). Cellular barcoding to decipher clonal dynamics in disease. Science *378*, eabm5874.

98. Espinosa, J.S., Tea, J.S., and Luo, L. (2014). Mosaic analysis with double markers (MADM) in mice. Cold Spring Harb. Protoc. *2014*, 182–189.

99. Livet, J., Weissman, T.A., Kang, H., Draft, R.W., Lu, J., Bennis, R.A., Sanes, J.R., and Lichtman, J.W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature *450*, 56–62.

100. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*, eaaw3381.

101. Rodriguez-Fraticelli, A.E., Weinreb, C., Wang, S.-W., Migueles, R.P., Jankovic, M., Usart, M., Klein, A.M., Lowell, S., and Camargo, F.D. (2020). Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. Nature *583*, 585–589.

102. Berthelet, J., Wimmer, V.C., Whitfield, H.J., Serrano, A., Boudier, T., Mangiola, S., Merdas, M., El-Saafin, F., Baloyan, D., Wilcox, J., et al. (2021). The site of breast cancer metastases dictates their clonal composition and reversible transcriptomic profile. Sci. Adv. *7*, eabf4408.

103. Roda, N., Cossa, A., Hillje, R., Tirelli, A., Ruscitto, F., Cheloni, S., Priami, C., Dalmasso, A., Gambino, V., Blandano, G., et al. (2023). A rare subset of primary tumor cells with concomitant hyperactivation of extracellular matrix remodeling and dsRNA-IFN1 signaling metastasizes in breast cancer. Cancer Res. *83*, 2155–2170.

104. Wild, S.A., Cannell, I.G., Nicholls, A., Kania, K., Bressan, D., CRUK IMAXT Grand Challenge Team, Hannon, G.J., and Sawicka, K. (2022). Clonal transcriptomics identifies mechanisms of chemoresistance and empowers rational design of combination therapies. Elife *11*. https://doi.org/10.7554/eLife.80981.

105. Oren, Y., Tsabar, M., Cuoco, M.S., Amir-Zilberstein, L., Cabanos, H.F., Hütter, J.-C., Hu, B., Thakore, P.I., Tabaka, M., Fulco, C.P., et al. (2021). Cycling cancer persister cells arise from lineages with distinct programs. Nature *596*, 576–582.

106. Bowling, S., Sritharan, D., Osorio, F.G., Nguyen, M., Cheung, P., Rodriguez-Fraticelli, A., Patel, S., Yuan, W.-C., Fujiwara, Y., Li, B.E., et al. (2020). An engineered CRISPR-Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. Cell *181*, 1693–1694.

107. Zafar, H., Lin, C., and Bar-Joseph, Z. (2020). Single-cell lineage tracing by integrating CRISPR-Cas9 mutations with transcriptomic data. Nat. Commun. *11*, 3055.

108. Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. *36*, 442–450.

109. McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

110. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. Nat. Biotechnol. *36*, 469–473.

111. Simeonov, K.P., Byrns, C.N., Clark, M.L., Norgard, R.J., Martin, B., Stanger, B.Z., Shendure, J., McKenna, A., and Lengner, C.J. (2021). Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. Cancer Cell *39*, 1150-1162.e9.

112. Abyzov, A., and Vaccarino, F.M. (2020). Cell lineage tracing and cellular diversity in humans. Annu. Rev. Genomics Hum. Genet. *21*, 101–116.

113. Chaligne, R., Gaiti, F., Silverbush, D., Schiffman, J.S., Weisman, H.R., Kluegel, L., Gritsch, S., Deochand, S.D., Gonzalez Castro, L.N., Richman, A.R., et al. (2021). Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. Nat. Genet. *53*, 1469–1479.

114. Ribera, J., Zamora, L., Morgades, M., Mallo, M., Solanes, N., Batlle, M., Vives, S., Granada, I., Juncà, J., Malinverni, R., et al. (2017). Copy number profiling of adult relapsed B-cell precursor acute lymphoblastic leukemia reveals potential leukemia progression mechanisms. Genes Chromosomes Cancer *56*, 810–820.

115. Penter, L., Dietze, K., Ritter, J., Lammoglia Cobo, M.F., Garmshausen, J., Aigner, F., Bullinger, L., Hackstein, H., Wienzek-Lischka, S., Blankenstein, T., et al. (2019). Localization-associated immune phenotypes of clonally expanded tumor-infiltrating T cells and distribution of their target antigens in rectal cancer. Oncoimmunology *8*, e1586409.

116. Papavasiliou, F.N., and Schatz, D.G. (2002). Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. Cell *109 Suppl*, S35-44.

117. Wasserstrom, A., Frumkin, D., Adar, R., Itzkovitz, S., Stern, T., Kaplan, S., Shefer, G., Shur, I., Zangi, L., Reizel, Y., et al. (2008). Estimating cell depth from somatic mutations. PLoS Comput. Biol. *4*, e1000058.

118. Xiao, Y., Jin, W., Ju, L., Fu, J., Wang, G., Yu, M., Chen, F., Qian, K., Wang, X., and Zhang, Y. (2024). Tracking single-cell evolution using clock-like chromatin accessibility loci. Nat. Biotechnol. https://doi.org/10.1038/s41587-024-02241-z.

119. Bernt, M., Braband, A., Schierwater, B., and Stadler, P.F. (2013). Genetic aspects of mitochondrial genome evolution. Mol. Phylogenet. Evol. *69*, 328–338.

120. Dowling, D.K., and Wolff, J.N. (2023). Evolutionary genetics of the mitochondrial genome: insights from Drosophila. Genetics *224*. https://doi.org/10.1093/genetics/iyad036.

121. Birky, C.W., Jr (2001). The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu. Rev. Genet. *35*, 125–148.

122. Sanchez-Contreras, M., and Kennedy, S.R. (2022). The complicated nature of somatic mtDNA mutations in aging. Front. Aging *2*. https://doi.org/10.3389/fragi.2021.805126.

123. Lawless, C., Greaves, L., Reeve, A.K., Turnbull, D.M., and Vincent, A.E. (2020). The rise and rise of mitochondrial DNA mutations. Open Biol. *10*, 200061.

124. Johnston, I.G., and Burgstaller, J.P. (2019). Evolving mtDNA populations within cells. Biochem. Soc. Trans. *47*, 1367–1382.

125. Ludwig, L.S., Lareau, C.A., Ulirsch, J.C., Christian, E., Muus, C., Li, L.H., Pelka, K., Ge, W., Oren, Y., Brack, A., et al. (2019). Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. Cell *176*, 1325-1339.e22.

126. Wilton, P.R., Zaidi, A., Makova, K., and Nielsen, R. (2018). A population phylogenetic view of mitochondrial heteroplasmy. Genetics *208*, 1261–1274.

127. Yuan, Y., Ju, Y.S., Kim, Y., Li, J., Wang, Y., Yoon, C.J., Yang, Y., Martincorena, I., Creighton, C.J., Weinstein, J.N., et al. (2020). Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat. Genet. *52*, 342–352.

128. Birky, C.W., Jr, Maruyama, T., and Fuerst, P. (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics *103*, 513–527.

129. Fu, Q., Mittnik, A., Johnson, P.L.F., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., et al. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. Curr. Biol. *23*, 553–559.

130. Posth, C., Wißing, C., Kitagawa, K., Pagani, L., van Holstein, L., Racimo, F., Wehrberger, K., Conard, N.J., Kind, C.J., Bocherens, H., et al. (2017). Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. Nat. Commun. *8*, 16046.

131. Tavaré, S., Balding, D.J., Griffiths, R.C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. Genetics *145*, 505–518.

132. Coalescent Theory, and the Analysis of Genetic Polymorphisms.

133. Tam, Z.Y., Gruber, J., Halliwell, B., and Gunawan, R. (2013). Mathematical modeling of the role of mitochondrial fusion and fission in mitochondrial DNA maintenance. PLoS One *8*, e76230.

134. Kapli, P., Yang, Z., and Telford, M.J. (2020). Phylogenetic tree building in the genomic age. Nat. Rev. Genet. *21*, 428–444.

135. Han, L., Zi, X., Garmire, L.X., Wu, Y., Weissman, S.M., Pan, X., and Fan, R. (2014). Co-detection and sequencing of genes and transcripts from the same single cells facilitated by a microfluidics platform. Sci. Rep. *4*, 6485.

136. Nitsch, L., Lareau, C.A., and Ludwig, L.S. (2024). Mitochondrial genetics through the lens of single-cell multi-omics. Nat. Genet. *56*, 1355–1365.

137. Miller, T.E., Lareau, C.A., Verga, J.A., DePasquale, E.A.K., Liu, V., Ssozi, D., Sandor, K., Yin, Y., Ludwig, L.S., El Farran, C.A., et al. (2022). Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations. Nat. Biotechnol. *40*, 1030–1034.

138. Lareau, C.A., Ludwig, L.S., Muus, C., Gohil, S.H., Zhao, T., Chiang, Z., Pelka, K., Verboon, J.M., Luo, W., Christian, E., et al. (2021). Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. Nat. Biotechnol. *39*, 451–461.

139. Fiskin, E., Lareau, C.A., Ludwig, L.S., Eraslan, G., Liu, F., Ring, A.M., Xavier, R.J., and Regev, A. (2022). Single-cell profiling of proteins and chromatin accessibility using PHAGE-ATAC. Nat. Biotechnol. *40*, 374–381.

140. Izzo, F., Myers, R.M., Ganesan, S., Mekerishvili, L., Kottapalli, S., Prieto, T., Eton, E.O., Botella, T., Dunbar, A.J., Bowman, R.L., et al. (2024). Mapping genotypes to chromatin accessibility profiles in single cells. Nature *629*, 1149–1157.

141. Weng, C., Yu, F., Yang, D., Poeschla, M., Liggett, L.A., Jones, M.G., Qiu, X., Wahlster, L., Caulier, A., Hussmann, J.A., et al. (2024). Deciphering cell states and genealogies of human haematopoiesis. Nature *627*, 389–398.

142. Campbell, P., Chapman, M.S., Przybilla, M., Lawson, A., Mitchell, E., Dawson, K., Williams, N., Harvey, L., Ranzoni, A.M., Cvejic, A., et al. (2023). Mitochondrial mutation, drift and selection during human development and ageing. Research Square. https://doi.org/10.21203/rs.3.rs-3083262/v1.

143. Fabre, M.A., de Almeida, J.G., Fiorillo, E., Mitchell, E., Damaskou, A., Rak, J., Orrù, V., Marongiu, M., Chapman, M.S., Vijayabaskar, M.S., et al. (2022). The longitudinal dynamics and natural history of clonal haematopoiesis. Nature *606*, 335–342.

144. Mitchell, E., Spencer Chapman, M., Williams, N., Dawson, K.J., Mende, N., Calderbank, E.F., Jung, H., Mitchell, T., Coorens, T.H.H., Spencer, D.H., et al. (2022). Clonal dynamics of haematopoiesis across the human lifespan. Nature *606*, 343–350.

145. Salvador-Martínez, I., Grillo, M., Averof, M., and Telford, M.J. (2019). Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? Elife *8*. https://doi.org/10.7554/eLife.40292.

146. Wang, R., Zhang, R., Khodaverdian, A., and Yosef, N. (2023). Theoretical guarantees for phylogeny inference from single-cell lineage tracing. Proc. Natl. Acad. Sci. U. S. A. *120*, e2203352120.

147. Kaminow, B., Yunusov, D., and Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv. https://doi.org/10.1101/2021.05.05.442755.

148. Lareau, C.A., Liu, V., Muus, C., Praktiknjo, S.D., Nitsch, L., Kautz, P., Sandor, K., Yin, Y., Gutierrez, J.C., Pelka, K., et al. (2023). Mitochondrial single-cell ATAC-seq

for high-throughput multi-omic detection of mitochondrial genotypes and chromatin accessibility. Nat. Protoc. *18*, 1416–1440.

149. Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., participants in the 1st Human Cell Atlas Jamboree, and Marioni, J.C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63.

150. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

151. Liu, D. (2019). Algorithms for efficiently collapsing reads with Unique Molecular Identifiers. https://doi.org/10.1101/648683.

152. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

153. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

154. Virshup, I., Rybakov, S., Theis, F.J., Angerer, P., and Wolf, F.A. (2021). anndata: Annotated data. https://doi.org/10.1101/2021.12.16.473007.

155. Huang, X., and Huang, Y. (2021). Cellsnp-lite: an efficient tool for genotyping single cells. Bioinformatics *37*, 4569–4571.

156. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv [q-bio.GN]. https://doi.org/10.48550/ARXIV.1207.3907.

157. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational identification of cell Doublets in Single-cell transcriptomic data. Cell Syst. *8*, 281-291.e9.

158. Bandelt, H.-J., Kloss-Brandstätter, A., Richards, M.B., Yao, Y.-G., and Logan, I. (2014). The case for the continuing use of the revised Cambridge Reference

Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. J. Hum. Genet. *59*, 66–77.

159. Kwok, A.W.C., Qiao, C., Huang, R., Sham, M.-H., Ho, J.W.K., and Huang, Y. (2022). MQuad enables clonal substructure discovery using single cell mitochondrial variants. Nat. Commun. *13*, 1205.

160. Marot-Lassauzaie, V., Beneyto-Calabuig, S., Obermayer, B., Velten, L., Beule, D., and Haghverdi, L. (2024). Identifying cancer cells from calling single-nucleotide variants in scRNA-seq data. Bioinformatics *40*. https://doi.org/10.1093/bioinformatics/btae512.

161. Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., and Theis, F.J. (2019). A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods *16*, 43–49.

162. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. *3*, 861.

163. Coifman, R.R., and Lafon, S. (2006). Diffusion maps. Appl. Comput. Harmon. Anal. *21*, 5–30.

164. Jones, M.G., Khodaverdian, A., Quinn, J.J., Chan, M.M., Hussmann, J.A., Wang, R., Xu, C., Weissman, J.S., and Yosef, N. (2020). Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. Genome Biol. *21*, 92.

165. Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A., von Haeseler, A., and Minh, B.Q. (2018). MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol. Biol. *18*, 11.

166. Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., and Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature *556*, 452–456.

167. Schiffman, J.S., D'Avino, A.R., Prieto, T., Pang, Y., Fan, Y., Rajagopalan, S., Potenski, C., Hara, T., Suvà, M.L., Gawad, C., et al. (2024). Defining heritability, plasticity, and transition dynamics of cellular phenotypes in somatic evolution. Nat. Genet. *56*, 2174–2184.

168. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

169. Huang, Y., McCarthy, D.J., and Stegle, O. (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biol. *20*, 273.

170. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*. https://doi.org/10.1186/s13059-017-1382-0.

171. DeTomaso, D., and Yosef, N. (2021). Hotspot identifies informative gene modules across modalities of single-cell genomics. Cell Syst. *12*, 446-456.e9.

172. Beneyto-Calabuig, S., Merbach, A.K., Kniffka, J.-A., Antes, M., Szu-Tu, C., Rohde, C., Waclawiczek, A., Stelmach, P., Gräßle, S., Pervan, P., et al. (2023). Clonally resolved single-cell multi-omics identifies routes of cellular differentiation in acute myeloid leukemia. Cell Stem Cell *30*, 706-721.e8.

173. Lareau, C.A., Chapman, M.S., Penter, L., Nawy, T., Pe'er, D., and Ludwig, L.S. (2024). Artifacts in single-cell mitochondrial DNA mutation analyses misinform phylogenetic inference. bioRxiv. https://doi.org/10.1101/2024.07.28.605517.

174. Weng, C., Weissman, J.S., and Sankaran, V.G. (2024). Robustness and reliability of single-cell regulatory multi-omics with deep mitochondrial mutation profiling. bioRxivorg. https://doi.org/10.1101/2024.08.23.609473.

175. Stadler, T., Pybus, O.G., and Stumpf, M.P.H. (2021). Phylodynamics for cell biologists. Science *371*, eaah6266.

176. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534.

177. Wang, X., Wang, K., Zhang, W., Tang, Z., Zhang, H., Cheng, Y., Zhou, D., Zhang, C., Zhong, W.-Z., Ma, Q., et al. (2024). Clonal expansion dictates the efficacy of mitochondrial lineage tracing in single cells. bioRxiv. https://doi.org/10.1101/2024.05.15.594338.

178. Yang, D., Jones, M.G., Naranjo, S., Rideout, W.M., III, Min, K.H. (joseph), Ho, R., Wu, W., Replogle, J.M., Page, J.L., Quinn, J.J., et al. (2021). Lineage recording reveals the phylodynamics, plasticity and paths of tumor evolution. https://doi.org/10.1101/2021.10.12.464111.

179. Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Scverse Community, Berger, B., et al. (2023). The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat. Biotechnol. *41*, 604–606.

180. Coorens, T.H.H., Spencer Chapman, M., Williams, N., Martincorena, I., Stratton, M.R., Nangalia, J., and Campbell, P.J. (2024). Reconstructing phylogenetic trees from genome-wide somatic mutations in clonal samples. Nat. Protoc. *19*, 1866–1886.