**GTAC: Genotyping with the Assay for Transposase-Accessible Chromatin**

**Introduction: what is GTAC and why to use it?**

The accumulation of somatic mutations leads to aberrant clonal expansions, pre-cancerous clonal outgrowths, and neoplastic transformation. Further clonal evolution within malignant tissues, driven by serial and/or branched emergence of fit subclones, underlies tumor heterogeneity, with important implications for therapy resistance and disease relapse. In the context of cancer therapy, for example, the specific fitness pressures applied by treatment select for specific genetic subclones able to resist and overcome therapy. A deeper understanding of such heterogeneity within human samples is a key prerequisite for the development of more efficient treatment strategies.

Several genes mutated across cancers encode epigenetic regulators and transcription factors, which directly regulate chromatin organization and gene expression. Furthermore, mutations in genes encoding signalling molecules, metabolic enzymes or DNA damage regulators are expected to exert a less direct, yet important effect on chromatin regulation. Understanding how epigenetic landscape changes across clones within the same cancer would shed light on how genes are aberrantly regulated downstream of cooperating mutations.

Inspired by single-cell methods linking transcriptional landscapes to clonal identities, like TARGET-seq and Genotyping of Transcriptomes, we developed GTAC, a method linking somatic genotypes to chromatin accessibility profiles of single cells. GTAC is a plate-based approach capturing single cells from primary human samples, which leverages amplification of accessible chromatin fragments to simultaneously capture any genomic locus harbouring a mutation in a subset of cells. This allows for efficient linkage of clonal identity and epigenetic landscape at single-cell resolution.

While developing GTAC, we optimized an elegant and robust plate-based scATAC-seq approach published previously by Xi Chen and colleagues: <https://www.nature.com/articles/s41467-018-07771-0> and <https://www.nature.com/articles/s41596-021-00583-5>

Despite the generally lower throughput when compared to droplet-based strategies, GTAC comes with several crucial advantages:

* High-content scATAC-seq libraries capture a notably higher number of unique nuclear fragments per cell compared to droplet-based strategies, allowing for high-resolution study of accessible chromatin
* Low allelic drop-out rates for single-cell genotyping across loci allows for high-confidence genotype calling, a key aspect when comparing molecular landscapes of different clones
* Efficient multiplexed genotyping allows to reconstruct complex clonal hierarchies within cancers, thus being applicable to a vast spectrum of human malignancies
* Genotyping of genomic loci is independent of chromatin accessibility, allowing for uniform genotyping efficiency across different cell states
* The computational framework relies on widely used scATAC-seq tools, thus being accessible both to computational experts and to scientists with limited computational experience

**How does GTAC work?**

GTAC uses custom target-specific primers which amplify genomic loci of interest in parallel with classic scATAC-seq library amplification. For each locus, the primer pair needs to efficiently generate the genotyping amplicon, at PCR conditions used for scATAC-seq library amplification, without compromising the quality of the respective scATAC-seq libraries. For that reason, primer pairs are first tested in the simulation of the GTAC protocol, performed on cell lines, to confirm efficient single-cell amplification of genotyping amplicons and a concomitant high-quality ATAC trace. Notably, this step does not require sequencing. Once the efficient primer pairs are selected for each locus, GTAC is performed on the primary sample of interest. After amplification, the material from each well is split in two aliquots: one is used for fully barcoded scATAC-seq library purification and sequencing, while the other is used for genotyping library construction: at this step, primers nested to the original target-specific oligos are added in order to barcode genotyping fragments and attach sequencing adapters. After sequencing, genotyping fragments for each cell are linked to the respective scATAC-seq libraries via unique cell barcodes.

**Analysis overview**

scATAC-seq and genotyping libraries are sequenced and pre-processed separately.

The scATAC-seq pre-processing pipeline generates fragment files, which are used as input for downstream analyses, performed with published R packages like ArchR or Signac.

The genotyping pre-processing pipeline generates files with cell-specific numbers of sequencing reads mapping to the mutant and reference alleles. These files are processed with custom R pipelines which perform genotype calling for each locus and integrate genotype information across screened loci for each cell.

Finally, this information is merged to scATAC-seq metadata for each cell, allowing to perform all downstream chromatin accessibility analyses using cell-specific clonal identities.

For scATAC-seq data pre-processing, a previously published Python script is used. All details about cell demultiplexing, software, code, and running instructions, can be found here: <https://github.com/dbrg77/scATAC_snakemake>

For our initial analyses, we slightly modified the pre-processing pipeline, eliminating the code at lines 173-192, to skip the second duplicate removal step performed on the aggregated bam file. The key output produced by this pipeline is the aggregate\_fragments.tsv.gz file, used as input for downstream GTAC analysis.

For genotyping library pre-processing, a previously published Perl script is used. All details about cell demultiplexing, software, code, and running instructions, can be found here: <https://github.com/albarmeira/TARGET-seq>

Of note, this pipeline, originally used for TARGET-seq genotype calling, assumes that both genomic and mRNA targeting genotyping primers are used. Given that GTAC uses only genomic genotyping primers, the mPRIMER.bed file is left empty for GTAC data pre-processing. For each locus of interest, this pipeline will generate files containing information about sequencing reads for each cell. These files are used as input for custom R scripts.

**Downstream analysis**

All downstream analysis was performed in R. We provide seven different sections of code used to assign clonal identities and use them for scATAC-seq analysis

1. **01\_Genotype calling:** used to assign a genotype for each cell at each locus of interest
2. **02\_ADO estimation in cell lines:** this section is used to estimate the rate of allelic drop-out (ADO) in cell lines known to be heterozygous mutant at loci of interest. The code in this section can be used to benchmark ADO rates for a given set of primers.
3. **03\_Genotype integration:** once we have genotypes for each locus of interest, genotype information is merged across cells to assign each cell to a most likely clone and reconstruct clonal hierarchy
4. **04\_Integration of genotype and scATAC data:** once the ArchR project is generated using the fragment file as input, clone information is added to the project metadata. Standard QC is performed to exclude low-quality cells or cells for which the clone could not be confidently assigned
5. **05\_Dimensionality reduction\_peak calling\_cluster assignment:** this section performs LSI, UMAP visualization, evaluation of gene activities, peak calling on chosen pseudobulk aggregates, TF motif accessibility across populations, and enrichment of marker peaks. This allows to confidently label cell populations in the dataset
6. **06\_ Downstream clone-specific differential analysis:** most of this section analyses clone-specific chromatin accessibility: differential gene and motif accessibility, clone-specific dynamics of differentiation, variation of genomic elements in specific clones
7. **07\_Projection of data on healthy reference:** here, all the steps necessary for efficient mapping of data on a healthy reference are described.

**Future developments**

As we continue to use GTAC, we expect that several of these computational steps will be refined and optimized to incorporate more efficient analysis strategies. For example, novel peak calling algorithms, like LanceOtron (<https://github.com/LHentges/LanceOtron>), leverage machine learning to better distinguish *bona-fide* genomic peaks from noise, allowing for a more precise analysis of genomic data. Furthermore, future integration of clonal, chromatin, and transcriptomic data will add a new layer of power to the existing datasets.

We do appreciate that the analyses presented here are not exhaustive of what can be done by integrating clonal information with chromatin accessibility landscapes. For that reason, going forward, we will keep updating these sections with new potentially useful analyses that can be performed in order to best leverage such datasets.

**Contact**

Sven Turkalj

sven.turkalj@rdm.ox.ac.uk