Supplementary Note: From ATAC-seq to Gene Programs and GMT Gene Sets

Scope and motivation

ATAC-seq measures chromatin accessibility at genomic loci. Any gene-level representation is therefore a model-based projection from locus space to gene space. The objective in this project is to extract compact gene programs (typically 100–500 genes) that are suitable for downstream enrichment and mechanism-level interpretation. Converters may emit a gene-score distribution over many genes, but downstream workflows usually benefit from a smaller set.

This note provides:

- a unified mathematical framework to compute gene scores from ATAC peaks,
- biologically motivated peak-to-gene linkage models,
- explicit operators to turn a gene-score distribution into one or more gene sets of size 100–500,
- and a specification for exporting these gene sets to a `.gmt` file.

1 Notation

- Peaks (regions) indexed by $p \in \{1, \ldots, P\}$ with coordinates (c_p, s_p, e_p).
- Genes indexed by $g \in \{1, \ldots, G\}$ with chromosome c_g, strand $\sigma_g \in \{+, -\}$, TSS position t_g, promoter interval $\mathcal{P}_g$, and optionally a gene locus interval $\mathcal{G}_g$.
- Peak-level observed statistic x_p (counts, coverage, or differential statistic).
- Peak weight α_p derived from x_p using a transform ϕ.
- Peak-to-gene linkage weight $L_{pg} \geq 0$.
- Raw gene score s_g and (optional) normalized gene weight w_g.

2 ATAC-seq to gene scores: a unified projection model

2.1 Peak weights

Define a transform ϕ that maps a peak statistic x_p to a peak weight α_p:

$$\alpha_p = \phi(x_p).$$

Common transforms:

$$\begin{aligned}\phi_{\text{signed}}(x) &= x,\\ \phi_{\text{abs}}(x) &= |x|,\\ \phi_{\text{pos}}(x) &= \max(x,0),\\ \phi_{\text{neg}}(x) &= \max(-x,0).\end{aligned}$$

For differential accessibility (opening vs closing), ϕ_{pos} and ϕ_{neg} naturally yield direction-specific programs.

2.2 Gene scores as a linear projection

Compute raw gene scores by:

$$s_g \;=\; \sum_{p=1}^{P} \alpha_p \, L_{pg}. \tag{1}$$

Equation (1) is an interpretable surrogate for regulatory influence: peaks with large α_p contribute more, and L_{pg} encodes how peak p is attributed to gene g.

2.3 Optional normalization to a distribution

If desired, convert scores to a nonnegative distribution:

$$w_g \;=\; \frac{s_g}{\sum_{h=1}^{G} s_h}, \quad \text{when } s_g \geq 0 \text{ and not all zero.} \tag{2}$$

This creates a simplex-valued vector $\sum_g w_g = 1$. Important: this is a *relative allocation* across genes, not a per-gene probability of activity. For program-sized outputs (100–500 genes), global L1 normalization across all genes often yields small and visually uniform weights when many genes have nonzero score. Therefore, selection is required.

3 Peak-to-gene linkage models L_{pg}

The linkage model encodes biological assumptions about regulatory wiring. All linkage models below can be combined with any program extraction operator (Sections 4–5).

3.1 Model A: promoter overlap

Biology. Promoter accessibility is a direct indicator of local transcriptional competence and has minimal ambiguity for gene assignment.

Definition.

$$L_{pg} = \mathbf{1}\{c_p = c_g \wedge [s_p, e_p) \cap \mathcal{P}_g \neq \emptyset\}.$$

Heuristics. Typical promoter window defaults $(u, d) = (2000, 500)$ bp; expose as parameters. Consider union across transcripts or canonical TSS only; record the choice.

3.2 Model B: nearest TSS assignment

Biology. Many enhancers regulate nearby genes; nearest TSS is a simple heuristic for distal inclusion.

Definition. Let $m_p = (s_p + e_p)/2$ and $d(p, g) = |m_p - t_g|$. With max distance D:

$$g^*(p) = \arg \min_{g:d(p,g) \leq D} d(p, g), \quad L_{pg} = \mathbf{1}\{g = g^*(p)\}.$$

Heuristics. Use per-chromosome sorted TSS arrays for $O(\log G)$ nearest lookup; deterministic tie-break by gene ID.

3.3 Model C: distance-decay soft assignment

Biology. Regulatory influence and contact probability tend to decrease with genomic distance on average.

Definition. For distance $d(p, g)$, max distance D, and decay length λ:

$$A_{pg} = \begin{cases} \exp(-d(p, g)/\lambda), & d(p, g) \leq D \\ 0, & \text{otherwise} \end{cases}$$

Either use $L_{pg} = A_{pg}$ (unnormalized), or normalize per peak:

$$L_{pg} = \frac{A_{pg}}{\sum_h A_{ph} + \epsilon}.$$

Heuristics. Defaults such as $D = 500$ kb, $\lambda = 50$ kb; cap links per peak to top $K_{\max}$ genes.

3.4 Model D: gene locus activity (gene body plus promoter)

Biology. Aggregated accessibility across promoter and gene body can correlate with transcriptional activity and provides robustness in sparse scATAC.

Definition. Define a gene locus interval $\mathcal{G}_g$ and compute:

$$s_g = \sum_{p:[s_p, e_p) \cap \mathcal{G}_g \neq \emptyset} \alpha_p.$$

Heuristics. Mitigate gene-length bias via standardized windows, length normalization, or promoter-only baselines.

3.5 Model E: co-accessibility linkage

Biology. Peaks that co-vary across cells can reflect regulatory coupling and enhancer-promoter communication.

Definition. Let $C_{p,q(g)}$ be a co-accessibility score between distal peak p and promoter peak(s) $q(g)$:

$$L_{pg} = \max(C_{p,q(g)}, 0) \cdot \mathbf{1}\{d(p, g) \leq D\}.$$

Heuristics. Requires many cells; restrict candidates within D; threshold edges; optionally combine with distance-decay.

3.6 Model F: external enhancer-gene priors

Biology. External maps integrate multi-assay evidence and can improve specificity.

Definition. Overlap peaks to external elements e with link matrix M_{eg}:

$$L_{pg} = \sum_{e} \mathbf{1}\{p \cap e \neq \emptyset\} M_{eg}.$$

Heuristics. Choose a default map per genome build; record version; handle biosample mismatch explicitly.

4 Program extraction from a gene-score distribution

4.1 Why extraction is required

A converter can produce scores for many genes. For enrichment and interpretation, we prefer compact programs (100–500 genes). Therefore, we apply an explicit extraction operator that maps $\{s_g\}$ (or $\{w_g\}$) to one or more gene sets.

4.2 Operator P1: fixed-size top-K program

Sort genes by decreasing score: $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(G)}$ with corresponding genes $g_{(1)}, \ldots, g_{(G)}$. Define the top-K program:

$$\widehat{\mathcal{G}}_K = \{g_{(1)}, \ldots, g_{(K)}\}.$$

This set maximizes captured total score among all size-K sets:

$$\widehat{\mathcal{G}}_K \in \arg\max_{|\mathcal{G}|=K} \sum_{g \in \mathcal{G}} s_g.$$

Within-program weights (optional) can be computed by normalizing within the selected set:

$$\tilde{w}_g = \begin{cases} \dfrac{s_g}{\sum_{h \in \widehat{\mathcal{G}}_K} s_h}, & g \in \widehat{\mathcal{G}}_K \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This avoids the near-uniformity induced by global normalization across all genes.

4.3 Operator P2: highest-probability-mass (HPD) program for a target mass

If scores are converted to a distribution w_g (Eq. 2), we may want the smallest gene set that captures a target cumulative mass $\tau \in (0,1)$. Let $w_{(1)} \geq \cdots \geq w_{(G)}$ be the sorted weights. Define:

$$K^*(\tau) = \min\left\{K : \sum_{i=1}^{K} w_{(i)} \geq \tau\right\}, \qquad \widehat{\mathcal{G}}_\tau = \{g_{(1)}, \ldots, g_{(K^*(\tau))}\}.$$

Minimality property. For any set $\mathcal{A}$ of size K, $\sum_{g\in\mathcal{A}} w_g \le \sum_{i=1}^{K} w_{(i)}$. Therefore, $K^*(\tau)$ is the minimal cardinality required to achieve mass at least τ, and $\widehat{\mathcal{G}}_\tau$ is a minimal-size set achieving that mass.

Size constraint. If the goal is 100–500 genes, one can clamp the size:

$$K = \min(\max(K^*(\tau), K_{\min}), K_{\max}),$$

with $K_{\min} = 100$ and $K_{\max} = 500$, then take $\widehat{\mathcal{G}}_K$. In practice, fixed-size top-K is simpler and more reproducible across datasets; HPD sets are useful when one wants to capture a chosen fraction of total mass.

4.4 Operator P3: multiple programs per distribution (multi-resolution)

A single distribution can yield more than one gene set. Two practical multi-set constructions are:

Nested top-K sets. Choose K values (e.g., 100, 200, 500):

$$\widehat{\mathcal{G}}_{100} \subset \widehat{\mathcal{G}}_{200} \subset \widehat{\mathcal{G}}_{500}.$$

This provides a multi-resolution view of the same program.

Disjoint tiers (optional). Partition the top-$K_{\max}$ genes into tiers of size B (e.g., 100):

$$\mathcal{T}_1 = \{g_{(1)}, \dots, g_{(B)}\},\ \mathcal{T}_2 = \{g_{(B+1)}, \dots, g_{(2B)}\},\ \dots$$

This yields multiple disjoint gene sets capturing successively weaker parts of the signal. This is less biologically guaranteed than nested sets but can be useful for sensitivity analyses.

4.5 Signed scores: positive and negative programs

If gene scores can be signed, construct two nonnegative score vectors:

$$s_g^+ = \max(s_g, 0), \qquad s_g^- = \max(-s_g, 0),$$

and apply any extraction operator separately to obtain opening and closing (or up and down) programs.

5 Exporting extracted programs to GMT format

5.1 GMT line format used here

We output gene sets in a one-line format:

```
<gene_set_name>\t<gene1> <gene2> ... <geneN>.
```

That is, a single tab after the gene set name, followed by genes separated by single spaces. (Notes: Some tools expect the classical GMT with tab-delimited gene entries and an explicit description column. This project uses the simplified single-tab plus space-delimited gene list described above.)

5.2 Name and gene identifier choice

Gene set name. A robust default name concatenates provenance fields:

`<dataset_id>__<converter>__<group>__<link>__<score>__<method>`.

Names must be sanitized to avoid whitespace; recommended replacements are space to underscore and removal of path separators.

Gene tokens. Prefer gene symbols when available and non-empty; otherwise use stable gene IDs. Within a gene set, enforce uniqueness by retaining the first occurrence and skipping duplicates (while preserving order).

5.3 Deriving gene sets from a distribution

Let D denote a single gene-score distribution output from a converter. Two canonical derivations are:

Derivation 1 (top-K).

1. Compute scores s_g (Eq. 1) or use weights w_g if provided.
2. Sort genes by decreasing score, tie-break by gene ID.
3. Choose $K \in [100, 500]$ (default 200), and take the top-K genes.
4. Emit one GMT line, where the gene list is the ordered top-K genes.

Derivation 2 (HPD mass).

1. Convert scores to a distribution w_g by Eq. 2.
2. Choose a mass target τ (e.g., 0.5 or 0.8).
3. Compute $K^*(\tau)$ and the HPD set $\widehat{\mathcal{G}}_\tau$ (Section 4.3).
4. Clamp to $[100, 500]$ if necessary and emit the GMT line.

Multiple sets per distribution. If a converter emits one distribution per group (e.g., scATAC clusters), then one can emit one or more GMT lines per group:

- one set per group: top-K only,
- nested sets per group: top-100, top-200, top-500,
- and optionally signed pos/neg sets if applicable.

5.4 Implementation complexity

For a distribution with G genes, extraction is dominated by sorting: $O(G \log G)$ time and $O(G)$ memory. This is negligible compared to peak-to-gene linking for typical ATAC datasets.

6 Bulk and single-cell ATAC usage notes (high level)

6.1 Bulk ATAC

For bulk contrasts, the most interpretable programs are directional:

- opening program: use positive peak statistics and extract top-K genes,
- closing program: use negative peak statistics and extract top-K genes.

6.2 Single-cell ATAC

For scATAC clusters, absolute activity programs can be broad; cluster-specific programs are typically more useful:

- compute a group vs rest contrast at the peak level,
- map peaks to genes (promoter baseline and distance-decay alternative),
- extract top-K genes (100–500) and export to GMT.

7 Practical recommendations

- Default to fixed top-K extraction with $K = 200$; it is reproducible and ensures the desired set size.
- Optionally emit nested sets at sizes (100, 200, 500) for sensitivity analysis.
- If scores are nearly uniform, the extracted set may be unstable; in that case prefer differential or specificity scoring (e.g., group vs rest for scATAC).
- Export GMT using gene symbols when available; always record genome build and annotation version in metadata.