Supplementary Note: Models for Converting ATAC-seq Data into Weighted Gene Sets

Summary

This note derives and summarizes a family of models that transform ATAC-seq data into gene-level weights suitable for downstream analyses that consume (i) gene lists or (ii) weighted gene lists. ATAC-seq measures chromatin accessibility at genomic loci. Any gene-level score is therefore a model-based projection from locus space to gene space. We organize the models under a shared mathematical framework: assign each ATAC feature (typically a peak) a peak-level weight and a peak-to-gene linkage distribution, then aggregate peak contributions to obtain a gene-weight vector. We provide biological motivations, explicit equations, and practical implementation heuristics for both bulk and single-cell ATAC-seq.

1 Notation and inputs

Let:

- Peaks (or regions) be indexed by $p \in \{1, \ldots, P\}$. Each peak p has genomic coordinates (c_p, s_p, e_p) (chromosome, start, end).
- Genes be indexed by $g \in \{1, \ldots, G\}$. Each gene g has chromosome c_g, strand $\sigma_g \in \{+, -\}$, transcription start site (TSS) position t_g, and promoter interval $\mathcal{P}_g$ (defined below).
- The observed ATAC signal for peak p is x_p. This could be raw counts, normalized accessibility, or a differential accessibility statistic (possibly signed).
- The desired gene-level weight for gene g is w_g. We may further normalize $\mathbf{w} = (w_1, \ldots, w_G)$ to form a distribution over genes.

Throughout, we will distinguish:

- **Peak weights** (how important each peak is in the dataset/contrast): α_p.
- **Peak-to-gene link weights** (how peak p is attributed to gene g): $L_{pg} \geq 0$.

1.1 Coordinate conventions

Implementations must be explicit about coordinate systems:

- BED is 0-based, half-open: interval $[s, e)$.
- GTF/GFF is 1-based, closed: interval $[s, e]$.

A common approach is to convert all inputs into a single internal convention (typically 0-based half-open) before overlap computations.

1.2 TSS and promoter definition

Given gene interval $[a_g, b_g]$ in 0-based coordinates:

$$t_g = \begin{cases} a_g, & \sigma_g = + \\ b_g, & \sigma_g = - \end{cases}$$

Define promoter window parameters (u, d) (upstream and downstream in bp). A strand-aware promoter interval is:

$$\mathcal{P}_g = \begin{cases} [t_g - u,\ t_g + d), & \sigma_g = + \\ [t_g - d,\ t_g + u), & \sigma_g = - \end{cases}$$

Clipping to chromosome bounds is recommended in practice.

2 ATAC-seq gene set extraction: general modeling framework

2.1 Biological rationale

ATAC-seq captures chromatin accessibility at regulatory DNA: promoters, enhancers, insulators, and other elements. Accessibility is a proxy for regulatory activity and/or potential: regions must typically be accessible to be bound by transcription factors and co-factors. Gene regulation is driven by both promoter state and distal regulatory elements connected through 3D genome organization and regulatory wiring. Thus, mapping ATAC signal to genes requires a *peak-to-gene* model.

2.2 A generic linear projection model

We represent the gene-level weight as a linear projection of peak-level weights:

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

Here:

- α_p encodes how strongly peak p is implicated by the ATAC dataset (or contrast).
- L_{pg} encodes how strongly peak p regulates or is associated with gene g.

Equation (1) can be motivated as a pragmatic approximation to an unknown cis-regulatory model. For example, suppose a (latent) gene regulatory response y_g obeys:

$$y_g \approx \sum_{p=1}^{P} \beta_{pg} \, x_p,$$

where β_{pg} are unknown peak-to-gene effects. If we have a prior or proxy for β_{pg} in the form of nonnegative link weights L_{pg} (possibly rescaled), and we choose a peak statistic α_p derived from x_p (or from a comparison), then the induced gene score (1) is an interpretable, reproducible surrogate for regulatory potential or regulatory change.

2.3 Peak-weight transforms

ATAC peak statistics can be signed (e.g., differential accessibility) or nonnegative (e.g., counts). Downstream enrichment methods often expect nonnegative weights, or separate up and down signatures. Define a transform $\phi(\cdot)$ mapping a raw peak statistic x_p to a peak weight α_p:

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

Common choices:

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

The `pos` and `neg` transforms allow constructing separate gene sets for opening vs closing chromatin (or case up vs case down).

2.4 Normalization to a gene-weight distribution

Many downstream tools benefit from comparable scaling across samples or groups. A common normalization is L1 normalization:

$$\tilde{w}_g = \frac{w_g}{\sum_{h=1}^{G} w_h}, \qquad \text{assuming } w_g \geq 0 \text{ and not all zero.} \tag{2}$$

This yields a distribution over genes: $\sum_g \tilde{w}_g = 1$. If signed weights are retained, an alternative is to normalize the positive and negative parts separately or to report both.

3 ATAC-seq peak-to-gene linkage models

This section defines the linkage weights L_{pg} used in Eq. (1). Each model encodes different biological assumptions about cis-regulatory wiring.

3.1 Model family A: promoter-only linkage

Biological justification

Promoter accessibility is tightly coupled to transcriptional competency: an inaccessible promoter often indicates repression or low transcriptional activity. Promoter-only models focus on the most direct and least ambiguous association between peaks and genes.

Definition

Define an overlap indicator:

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

Then:

$$w_g = \sum_{p : p \cap \mathcal{P}_g \neq \emptyset} \alpha_p.$$

Practical heuristics

- Choose promoter window defaults such as $(u, d) = (2000, 500)$ bp, but expose as parameters.
- Handle multiple TSS per gene by either: (i) using a canonical transcript definition, or (ii) taking the union of promoter intervals across transcripts and aggregating.
- If multiple peaks overlap the same promoter, summing is typical; max is an alternative if peaks are highly redundant.

3.2 Model family B: nearest-TSS hard assignment

Biological justification

A substantial fraction of enhancers regulate the nearest gene, and nearest-TSS assignment is a simple heuristic that includes distal signal while preserving a one-peak-to-one-gene mapping.

Definition

Let $d(p, g)$ be the distance from peak center to gene TSS:

$$m_p = \frac{s_p + e_p}{2}, \qquad d(p, g) = |m_p - t_g|.$$

Then assign each peak to the nearest gene within a maximum distance D:

$$g^*(p) = \arg \min_{g : c_g = c_p,\, d(p,g) \le D} d(p, g).$$

Define:

$$L_{pg} = \mathbf{1}\{g = g^*(p)\}.$$

Then:

$$w_g = \sum_{p : g^*(p) = g} \alpha_p.$$

Practical heuristics

- Use chromosome-specific sorted lists of TSS positions to compute $g^*(p)$ in $O(\log G_c)$ via binary search.
- Choose D (e.g., 100 kb) based on expected cis range.
- Tie-breaking: if two genes are equidistant, pick the one with smallest gene ID for determinism.
- Consider promoter-prioritized nearest assignment: if peak overlaps any promoter, assign to that promoter gene(s) before applying nearest rule.

3.3 Model family C: distance-decay soft assignment

Biological justification

3D contact probability and regulatory influence often decrease with genomic distance on average. A soft assignment allows a distal peak to contribute to multiple nearby genes with weights that decay with distance, reflecting uncertainty and long-range regulation.

Definition

Let $d(p, g)$ be distance as above, and let D be a maximum distance and λ a decay length scale. Define an unnormalized affinity:

$$A_{pg} = \begin{cases} \exp\left(-\dfrac{d(p,g)}{\lambda}\right), & \text{if } c_p = c_g \text{ and } d(p,g) \leq D \\ 0, & \text{otherwise.} \end{cases}$$

Two common variants:

Variant C1: unnormalized links Set $L_{pg} = A_{pg}$ and compute w_g via Eq. (1). This causes peaks in gene-dense regions to contribute more total mass across genes.

Variant C2: per-peak normalized links Normalize affinities across genes for each peak:

$$L_{pg} = \frac{A_{pg}}{\sum_{h=1}^{G} A_{ph} + \epsilon},$$

where ϵ is a small constant to avoid division by zero if a peak has no nearby genes. This ensures each peak contributes the same total mass (in expectation) regardless of gene density.

Practical heuristics

- Typical defaults: $D = 500$ kb, $\lambda = 50$ kb; but expose both.
- Cap the number of genes per peak to reduce computation: keep only top K genes by A_{pg} (e.g., $K = 5$).
- Optionally boost promoters: if peak overlaps a promoter, multiply A_{pg} by a factor (e.g., 5 to 20) for that gene.
- Efficient implementation: for each chromosome, maintain genes sorted by TSS; for each peak center, expand outward from the insertion index until distance exceeds D.

3.4 Model family D: gene-body activity models (gene activity / gene score)

Biological justification

In single-cell contexts, sparse peak calls and weak signal-to-noise can make distal peak-to-gene assignment difficult. A robust proxy for gene regulatory state is aggregated accessibility in the promoter and gene body (and sometimes flanking regions), which can correlate with transcriptional activity and chromatin state.

Definition (continuous accessibility)

Let $a(z)$ be an accessibility signal along the genome. Define a gene locus interval $\mathcal{G}_g$ (gene body plus an upstream extension). Then:

$$w_g = \int_{\mathcal{G}_g} a(z)\, dz.$$

In practice, ATAC is discrete (reads/fragments) or summarized in peaks. If peaks partition the genome, then:

$$w_g \approx \sum_{p:[s_p, e_p) \cap \mathcal{G}_g \neq \emptyset} \alpha_p \cdot \ell_{pg},$$

where ℓ_{pg} is the overlap length (or an indicator), and α_p is a peak accessibility summary.

Practical heuristics

- Define $\mathcal{G}_g$ as gene body plus u bp upstream (e.g., 2 kb) and optionally a downstream extension.
- Control gene-length bias:
  - normalize by gene length, or
  - restrict to promoter-only for comparability, or
  - use TF-IDF-like normalization at the peak/cell level before aggregation in scATAC.
- When using a peak-by-cell matrix, define peak-level summaries (per group or per cell) as described in the single-cell section.

3.5 Model family E: co-accessibility based linkage

Biological justification

In single-cell ATAC, peaks that are co-accessible across cells often reflect coordinated regulatory programs and can proxy enhancer-promoter communication. Co-accessibility scores can be used to link distal peaks to promoters more specifically than distance alone.

Definition

Let X_{pc} be accessibility of peak p in cell c, and let $q(g)$ denote a promoter peak (or set of promoter peaks) representing gene g. Define a co-accessibility score $C_{p,q(g)}$ (e.g., correlation after smoothing/aggregation). Then set:

$$L_{pg} = \begin{cases} C_{p,q(g)} \cdot \mathbf{1}\{d(p,g) \leq D\}, & \text{if } C_{p,q(g)} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Gene weights follow Eq. (1).

Practical heuristics

- Co-accessibility estimation typically requires many cells and careful handling of sparsity (kNN smoothing or aggregation).
- Restrict candidate pairs to within D (e.g., 250 kb to 500 kb) to reduce computation.
- Threshold $C_{p,q(g)}$ to reduce noise (e.g., keep top edges or $C > \tau$).
- Combine with distance decay by setting $L_{pg} \propto C_{p,q(g)} \exp(-d/\lambda)$.

3.6 Model family F: external enhancer-gene maps (priors)

Biological justification

Large consortia and prior models provide predicted enhancer-gene links based on multi-assay evidence (chromatin marks, accessibility, eQTL, 3D contacts, etc.). Using a precomputed map can improve biological specificity compared to purely distance-based heuristics, at the cost of possible context mismatch.

Definition

Let $\mathcal{E}$ be a set of regulatory elements (e.g., cCREs) and a precomputed link matrix M_{eg} mapping element e to gene g (nonnegative). We map peaks to elements (by overlap), then elements to genes:

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

If M_{eg} is probabilistic, one may normalize per peak or per element.

Practical heuristics

- Provide sensible defaults: pick a broadly applicable map for human and mouse; document version and source.
- If multiple biosample-specific maps exist, choose: (i) the closest matching tissue/cell type, or (ii) an aggregate union map, but record the choice in metadata.
- If peaks overlap multiple elements, sum their contributions or take the maximum to avoid double counting.

3.7 Model family G: Activity-by-Contact style linkage

Biological justification

Enhancer influence can be modeled as a product of enhancer activity and enhancer-promoter contact probability. This is a mechanistic approximation: enhancers must be active and physically contact promoters to regulate transcription.

Definition

Assume each peak p is a candidate enhancer with activity $A_p \geq 0$, and each gene g has a promoter contact probability $K_{pg} \geq 0$. Define an ABC-like score:

$$S_{pg} = A_p\, K_{pg}. \tag{3}$$

Then define link weights by optional normalization:

$$L_{pg} = \frac{S_{pg}}{\sum_{h=1}^{G} S_{ph} + \epsilon}.$$

Finally compute gene weights by:

$$w_g = \sum_{p=1}^{P} \alpha_p\, L_{pg}.$$

Practical heuristics

- If only ATAC is available, set A_p proportional to ATAC signal in the peak; if histone marks are available, combine (e.g., ATAC and H3K27ac).
- If no sample-specific contact map is available, approximate K_{pg} by a distance-based contact kernel, e.g., $K_{pg} \propto \exp(-d(p,g)/\lambda_c)$ or a power-law.
- Thresholding: keep links with S_{pg} above a cutoff to reduce spurious long-range assignments.

3.8 Model family H: TF-centric gene sets via motif accessibility

Biological justification

ATAC-seq is sensitive to transcription factor (TF) binding and regulatory program activation. Instead of mapping peaks to genes directly, one may infer TF activity from motif accessibility and then project TF activity onto genes via a TF-target network, producing a gene-weight vector that emphasizes putative regulatory targets.

Definition

Let a_k be an activity score for TF (or motif) $k \in \{1, \dots, K\}$ inferred from ATAC (e.g., motif deviation scores). Let $T_{kg} \geq 0$ be a TF-to-gene prior weight (e.g., regulatory network edge weight). Define:

$$w_g = \sum_{k=1}^{K} a_k \, T_{kg}. \tag{4}$$

Optionally combine with direct peak-to-gene scoring by convex combination:

$$w_g^{\text{combo}} = \eta \, w_g^{\text{peak}} + (1-\eta) \, w_g^{\text{TF}}, \qquad \eta \in [0,1].$$

Practical heuristics

- Choose a TF-target prior appropriate for the organism and context.
- Motif redundancy and shared motifs can blur specificity; consider grouping motifs by TF family.
- Report TF activities alongside gene weights for interpretability.

4 ATAC-seq: Bulk models

Bulk ATAC typically yields a set of peaks with per-peak accessibility (single sample) or differential accessibility statistics (two-group contrasts, regression). Bulk models differ mainly in how α_p is defined and whether signed information is retained.

4.1 Bulk peak-weight choices

Single-sample bulk accessibility

Let x_p be normalized accessibility (e.g., counts per million or normalized coverage). Then set $\alpha_p = x_p$ (nonnegative), optionally after log transform.

Differential accessibility

Let x_p be a signed statistic (e.g., log fold change or Wald z). Then choose transform ϕ:

- A nonnegative "magnitude" gene set: $\alpha_p = |x_p|$.
- An "opening" gene set: $\alpha_p = \max(x_p, 0)$.
- A "closing" gene set: $\alpha_p = \max(-x_p, 0)$.
- A signed gene set for methods that accept signed ranks: $\alpha_p = x_p$.

Constructing separate opening and closing gene sets is often the most interpretable.

4.2 Bulk model B1: promoter overlap

Use Model family A with bulk α_p.

4.3 Bulk model B2: nearest TSS

Use Model family B with bulk α_p.

4.4 Bulk model B3: distance-decay

Use Model family C with bulk α_p.

4.5 Bulk model B4: external maps or ABC-style

Use Model family F or G if an external enhancer-gene prior is available.

4.6 Bulk model B5: TF-centric

Use Model family H when the scientific question emphasizes upstream regulators and their targets rather than direct peak-to-gene wiring.

4.7 Bulk implementation heuristics

- **Peak file formats:** narrowPeak is common; decide which column to use as x_p (score, signalValue, or a separate differential weight file).
- **Chromosome naming:** ensure peak and GTF chromosomes match (e.g., "chr1" vs "1").
- **Filtering:** remove peaks on nonstandard contigs unless needed; record filters in metadata.
- **Diagnostics:** report number of peaks assigned to at least one gene and fraction assigned; extreme low assignment rates often indicate genome build mismatch.

5 ATAC-seq: Single-cell models

Single-cell ATAC provides cell-by-peak accessibility X_{pc} that is sparse. It is often useful to compute gene weights:

- per cell (for integration), or
- per group/cluster (for enrichment and interpretation).

This note focuses on producing one gene-weight vector per group, which is typically the most stable for enrichment.

5.1 Single-cell peak summarization (pseudo-bulk)

Given a group of cells $c \in \mathcal{C}$, define a per-peak summary statistic $\bar{x}_p$:

- **Sum of counts:** $\bar{x}_p = \sum_{c \in \mathcal{C}} X_{pc}$.
- **Mean counts:** $\bar{x}_p = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} X_{pc}$.
- **Fraction nonzero:** $\bar{x}_p = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{1}\{X_{pc} > 0\}$.

Then set $\alpha_p = \bar{x}_p$ (or apply a transform). Fraction nonzero is often robust to outliers and library size variation.

5.2 scATAC model S1: group-level promoter overlap

Compute α_p as above for a group and apply promoter-only linkage (Model A). This yields a promoter accessibility signature for the group.

5.3 scATAC model S2: group-level distance-decay

Compute α_p for the group and apply distance-decay linkage (Model C), capturing distal regulatory programs. This is computationally heavier but often more sensitive to cell-type identity enhancers.

5.4 scATAC model S3: gene activity (gene body + upstream)

Define gene locus $\mathcal{G}_g$ (gene body plus upstream). Compute:

$$w_g = \sum_{p : p \cap \mathcal{G}_g \neq \emptyset} \alpha_p.$$

Optionally normalize by locus length or by number of peaks to mitigate gene-length bias.

5.5 scATAC model S4: co-accessibility enhanced linkage

Estimate co-accessibility between distal peaks and promoter peaks, then use Model E to define L_{pg}. Compute gene weights by Eq. (1). This model is often the most biologically specific but requires sufficient cells and careful estimation.

5.6 scATAC model S5: multiome-guided peak-to-gene

If paired RNA (multiome) is available, estimate peak-to-gene links by correlation or regression of accessibility and expression across cells:

$$L_{pg} \propto \operatorname{corr}(X_{p\cdot}, Y_{g\cdot}) \cdot \mathbf{1}\{d(p,g) \leq D\},$$

then compute gene weights using peak weights. This is highly context-specific and can be used as a reference to evaluate simpler models.

5.7 Single-cell implementation heuristics

- **Grouping labels:** use existing clustering labels if available; otherwise define coarse groups (e.g., by known markers or using scATAC clustering pipeline).
- **Matrix I/O:** 10x peak-by-cell matrices may be large; implement streaming accumulation per group to avoid repeated scans.
- **Normalization:** L1 normalize gene weights per group to compare distributions across groups.
- **Stability:** compare top-N gene overlap across models (promoter vs decay) as a robustness diagnostic.

6 Practical implementation guidance (cross-cutting)

6.1 A minimal recipe

A practical pipeline for any ATAC dataset:

1. Choose peak statistic x_p (counts, signal, differential statistic).
2. Transform to peak weights $\alpha_p = \phi(x_p)$.
3. Choose a peak-to-gene linkage model L_{pg} (A/B/C/D/E/F/G/H).
4. Aggregate gene weights: $w_g = \sum_p \alpha_p L_{pg}$.
5. Optionally normalize gene weights to a distribution (Eq. (2)).
6. Export as a weighted gene list and record full metadata (parameters, genome build, annotation version).

6.2 Computational considerations

- Avoid quadratic peak-gene loops by indexing genes per chromosome by TSS and using binary search to find candidate genes within D.
- For promoter overlaps and gene-body aggregation, use interval indexing or sorted sweeps.
- Determinism: sort outputs by decreasing weight with a stable tie-break (e.g., gene ID).

6.3 Diagnostics and failure modes

- **Low assigned fraction:** often indicates genome build mismatch or chromosome naming mismatch.
- **Gene-length bias:** gene-body models can overweight long genes; mitigate via normalization or promoter-only variants.
- **Over-smoothing:** large D and large λ in distance-decay can blur specificity; evaluate sensitivity.
- **Context mismatch in external maps:** priors may not match the sample; consider reporting multiple models and comparing concordance.

7 Recommended defaults (pragmatic)

For a general-purpose, reproducible starting point:

- Bulk ATAC:
  - promoter overlap with $(u, d) = (2000, 500)$ bp
  - distance-decay with $D = 500$ kb, $\lambda = 50$ kb, cap $K = 5$
  - separate opening and closing gene sets if differential statistics are available
- Single-cell ATAC (group-level):
  - summarize peaks by fraction nonzero per group
  - promoter overlap as a stable baseline
  - distance-decay or gene-body gene activity as distal-aware extensions