Genome-Wide Uncertainty-Moderated Extraction of Signal Annotations from Multi-Sample Functional Genomics Data

Nolan H. Hamilton[*,†,1,3], **Benjamin D. McMichael**[†1,3], **Michael I. Love**[2,3], **Terrence S. Furey**[*1,3]

University of North Carolina at Chapel Hill, [1]Department of Biology, [2]Department of Biostatistics, [3]Department of Genetics

[*]Corresponding authors. nolanh@email.unc.edu, tsfurey@email.unc.edu

[†]Nolan H. Hamilton and Benjamin D. McMichael should be regarded as joint first authors.

Abstract

Motivation: Most genome-wide consensus profiles of regulatory features consist of discrete binary annotations, or "peaks", identified as enriched in functional genomics data targeting loci with a particular molecular phenotype, e.g., accessible chromatin or a transcription factor binding site. Despite the established practical utility of this paradigm, these annotations can limit the representation of quantitative signal data to rigid boundaries and mask informative spatial and structural features in signals. We reasoned that a more robust integration of numeric data from multiple samples could elucidate higher-resolution genome-wide profiles for improved molecular state characterizations. This integration poses several fundamental challenges due to dynamic, sample-specific noise patterns that vary in their effects across the genome and render straightforward data aggregation approaches vulnerable to signal dilution or distortion.

Results: We propose Consenrich, a genome-wide state-estimation scheme for extraction of consistent underlying signals given functional sequencing data from multiple samples. Consenrich addresses fundamental challenges inherent in this task by implementing a biologically-informed model for signal and variance propagation that treats nonstationary noise processes to yield a genome-wide signal track of positional estimates and measures of uncertainty. We demonstrate improved multi-sample signal estimation and noise attenuation of Consenrich on real, heterogeneous functional genomics data compared to least-squares smoothing filters and straightforward "alignment merging" strategies. We highlight immediate practical appeal of this improved signal extraction performance for both traditional consensus peak calling tasks and address the potential for unique analyses not possible with simple binary peak calls.

Availability and Implementation: A software implementation of Consenrich accompanied by several practical usage examples and documentation is hosted at `https://github.com/nolan-h-hamilton/Consenrich`.

1. Introduction

High-throughput sequencing (HTS) functional genomics experiments have broad utility in studying diverse molecular processes and cellular states, such as gene regulatory programs or epigenomic states, by characterizing targeted molecular phenotypes in single or multiple samples [Corces et al., 2018; Luo et al., 2019].

Data generated using these assays are translated to genome-wide annotations by first mapping sequenced reads to specific genomic positions and then quantifying read density within genomic intervals. Targeted assay-specific features such as accessible chromatin (ATAC-seq) and transcription factor binding sites (ChIP-seq, CUT-N-RUN) can then be defined by determining regions with a significant enrichment in aligned read counts, referred to as peaks [Zhang et al., 2008]. While these peaks offer practical utility in downstream analyses, such as the identification of regions with differential activity or signal between two sets of phenotypically distinct samples, they provide a limited granularity, discarding significant amounts of potentially biologically meaningful spatial patterns. In addition, multi-sample, consensus peak annotations are generally determined with heuristic *post hoc* merging strategies [Yang et al., 2014]. These methods can impose discretionary, rigid thresholds on the minimum degree of overlap and/or number of samples a candidate peak must be present in to qualify as a consensus peak. As a result, consensus peak regions may be included or excluded based on small differences in signal or total overlap. While we previously developed a method for improved consensus peak determination across many samples [Hamilton and Furey, 2023], final consensus annotations are still restricted to binary state profiles where genomic loci have or lack a particular molecular phenotype.

We sought to create higher resolution molecular state profiles that both provided truly genome-wide measurements and could integrate information across multiple samples to define a more accurate, generalized profile for a sample phenotype of interest, such as a specific tissue or disease state (Figure 1). Two fundamental challenges impeding this effort were given special consideration. First, functional genomics experiments often vary significantly in signal quality due to a number of factors such as inconsistent sample quality, data generation across multiple time periods and technologies, and other experimental artifacts. Second, noise processes affecting data can vary across genomic regions due to local molecular structure such as chromatin compaction and the intrinsic stability of that molecular state at

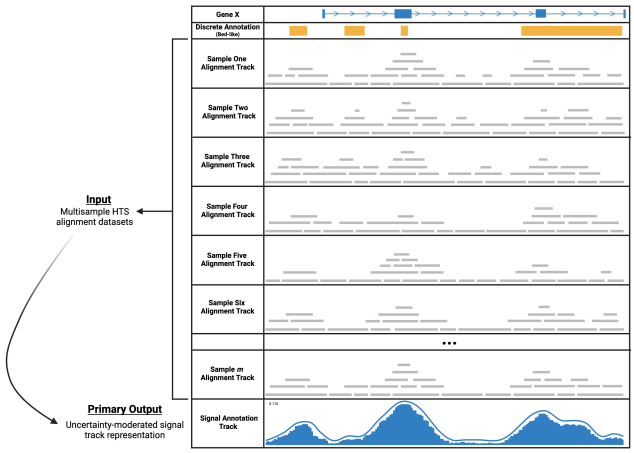


Fig. 1: An illustrative depiction of a real-valued signal annotation (bottom row, blue) generated from multiple samples' sequence alignments.

a particular locus across time and individuals [Meyer and Liu, 2014; Zhao and Boyle, 2021]. In related work, principled statistical methods aimed at this task are geared specifically toward ChIP-seq in small genomes (< 50mb) [Schroeder and Freddolino, 2024]. In the other direction, crude but efficient *ad hoc* approaches, such as pooling data across all samples, may fail to accurately characterize consensus profiles in the presence of heterogeneity that is common in larger sample sizes [Yang et al., 2014].

Our solution, implemented within our software package Consenrich, addresses these challenges with an efficient, explainable model that integrates biological reasoning with observed data from multiple samples. Emphasis is placed on the estimation of real-valued molecular states that provide sound representation of sample groups while remaining robust to heterogeneous data quality. In particular, we treat *noise processes*, which obscure true signals in the data, as nonstationary across the genome. The reliability of observed data is evaluated dynamically in each sample to allow for positional comparisons and calibrated uncertainty quantification. Consenrich dynamically recognizes and weighs more heavily information from high quality data while minimizing the influence of noisy data and accounting for local trends with a simple process model. These features provide appeal not only for estimating real-valued molecular state profiles in signal track formats—the results may also be supplied as input to peak callers for improved accuracy in representing commonly enriched regions of a particular sample group (consensus peak calling). Finally, there is the inherent computational burden of synthesizing and analyzing positional data spanning entire genomes from large numbers of independent samples. Consenrich thus relies primarily on fast matrix-vector operations and incorporates several additional efficiency optimizations for scalability to these large sample sizes that are becoming more common as the abundance of genomics data increases.

2. Method

Let $x_i \in \mathbb{R}$ denote the epigenomic signal or "state" of interest and $\dot{x}_i \in \mathbb{R}$ denote the rate of change or "trend" in this state at the ith genomic interval. Each interval is fixed in length L with a default of 50bp. We denote the respective "state vector" of these two variables at each interval $i = 1 \dots n$ as

$$\mathbf{x}_i = [x_i \quad \dot{x}_i]^\top \tag{1}$$

We likewise consider the "estimate covariance" to model uncertainty at each interval i:

$$\mathbf{P}_i = \begin{bmatrix} P_{i,11} & P_{i,12} \\ P_{i,21} & P_{i,22} \end{bmatrix}. \tag{2}$$

To estimate $\mathbf{x}_i, \mathbf{P}_i$, we construct two interacting models to leverage simple principles common to functional genomics data signals observed from multiple samples, respectively: (i) A recursive "process model" that propagates previous state/covariance estimates at intervals $i-1$ to intervals i, yielding the *a priori* estimate, and (ii) An "observation model" to process sequence alignment count data (or a transformation thereof) in the sample-by-interval matrix $\mathbf{Z} \in \mathbb{R}^{m \times n}$ at interval i to "update" the *a priori* estimate, yielding the *a posteriori* estimate. Uncertainty due to both models is considered and determines their respective influence on the *a posteriori* estimates. The "process model noise covariance", $\mathbf{Q}_i$, and "observation model noise covariance", $\mathbf{R}_i$ are computed as region and sample specific entities due to artifacts inherent in HTS experiments [Meyer and Liu, 2014] that affect different genomic regions and samples variably.

2.1. Initialization

The initial state vector and covariance are set as $\mathbf{x}_{init} = \begin{bmatrix} x_0 & 0 \end{bmatrix}^\top$ and $\mathbf{P}_0 = P_{init}\mathbf{I}$, where default $P_{init} = 1000$ to conservatively represent initial uncertainty. The default for x_{init} is the sample mean of values in the first column of $\mathbf{Z}$. Due to several robust procedures incorporated by Consenrich and its underlying paradigm addressed in Section 2.4, these initial values are generally of little consequence for the scale of our problem setting.

2.2. Process Model

Consenrich applies a linear, recursive process model $\mathbf{F} = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix}$ to obtain the current state estimate and estimate covariance from the previous interval at each $i = 1, 2, \dots, n$:

$$\mathbf{x}_{i|i-1} = \mathbf{F}\mathbf{x}_{i-1} = [(x_{i-1} + \Delta\dot{x}_{i-1}) \quad \dot{x}_{i-1}]^\top \tag{3}$$

$$\mathbf{P}_{i|i-1} = \mathbf{F}\mathbf{P}_{i-1}\mathbf{F}^\top + \mathbf{Q}_i \tag{4}$$

where $\mathbf{Q}_i \in \mathbb{R}^{2\times 2}$ denotes the assumed positive-definite "process noise covariance" and we treat each interval as a unit-step by default. As we cannot expect to capture all aspects of the true system in our process model, $\mathbf{Q}_i$ is used to augment the estimate covariance $\mathbf{P}_{i|i-1}$ to reflect uncertainty in the model itself. $\mathbf{Q}_i$ is determined with a dynamic procedure during iteration as discussed in Section 2.4.1.

2.3. Observation Model

After the *a priori* state estimate and covariance are computed, we consider the measurements from each sample. The residuals between the *a priori* state estimate and each sample's

measurements, $\mathbf{y}_i \in \mathbb{R}^m$, are taken as:

$$\mathbf{y}_i = \mathbf{z}_i - \mathbf{H}\mathbf{x}_{i|i-1} = \begin{bmatrix} Z_{i,1} - x_{i|i-1} \\ Z_{i,2} - x_{i|i-1} \\ \cdots \\ Z_{i,m} - x_{i|i-1} \end{bmatrix}. \tag{5}$$

where $\mathbf{H} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m \times 2}$ is the "observation model matrix". The residuals provide insight for evaluating filter performance, especially in the absence of ground truth. Given an accurately modeled underlying process and effective preprocessing of raw data as necessary, residuals will resemble zero-mean noise in the absence of abrupt changes in system dynamics [Zarchan and Musoff, 2015]. These properties of the residuals are later addressed empirically in Section 3.4. Further, the observed magnitude of residuals relative to the modeled uncertainty can be used to indicate deviations that warrant dynamic adjustments as discussed in Section 2.4.1.

In the proposed system model, the covariance of residuals reduces to the sum of (i) the *a priori* estimate covariance projected into $\mathbb{R}^{m \times m}$ and (ii) the observation noise covariance $\mathbf{R}_i \in \mathbb{R}^{m \times m}$:

$$\mathbf{E}_i = \mathbf{H}\mathbf{P}_{i|i-1}\mathbf{H}^T + \mathbf{R}_i \tag{6}$$

Elements $j = 1 \ldots m$ in $\mathbf{R}_i \succ \mathbf{0}$ represent the noise levels in each sample's alignment count data. Though a common reductive approach is to assume static $\mathbf{R}$, we define this matrix for each interval using the procedure presented in Section 2.4.2.

The residuals in $\mathbf{y}_i$ are scaled by the gain, $\mathbf{K}_i \in \mathbb{R}^{2 \times m}$,

$$\mathbf{K}_i = \mathbf{P}_{i|i-1}\mathbf{H}^\top \mathbf{E}_i^{-1} \tag{7}$$

to update *a priori* estimates and obtain *a posteriori* estimates.

$$\mathbf{x}_{i|i} = \mathbf{x}_{i|i-1} + \mathbf{K}_i \mathbf{y}_i \tag{8}$$

$$\mathbf{P}_{i|i} = (\mathbf{I} - \mathbf{K}_i \mathbf{H})\, \mathbf{P}_{i|i-1}. \tag{9}$$

where $\mathbf{I}$ in Eqn. 9 is the two-dimensional identity.

Remark 1 *Informally, the behavior of the gain (Eqn. 7) used to obtain the a posteriori estimates (Eqn. 8, 9) can be understood as a moderator that balances uncertainty arising from (i) the process model's a priori estimate, and (ii) the observation model's representation of the data (in our case, this is the data itself). If the uncertainty ascribed to the data is small (large) relative to the uncertainty ascribed to the process model's a priori prediction, the gain scales the data to allow a larger (smaller) effect on the a posteriori estimate. This is most clearly seen in the case of a single state variable x_i and a single observation z_i at each index, where Eqn. (7) reduces to $K_i = \frac{P_{i|i-1}}{P_{i|i-1}+R_i}$. For observation noise $R_i > 0$, as uncertainty in the process model's a priori estimate decreases toward zero,*

$$\lim_{P_{i|i-1} \to 0} K_i = 0 \implies \lim_{P_{i|i-1} \to 0} x_{i|i-1} + (0 \cdot y_i) = x_{i|i} = x_{i|i-1},$$

the a posteriori estimate approaches the a priori estimate. In contrast, for arbitrarily increasing uncertainty in the process model's a priori estimate,

$$\lim_{P_{i|i-1} \to \infty} K_i = 1 \implies \lim_{P_{i|i-1} \to \infty} x_{i|i-1} + (1 \cdot y_i) = x_{i|i}$$
$$= x_{i|i-1} + (z_i - x_{i|i-1}) = z_i,$$

the a posteriori estimate approaches the observation exclusively.

That we have two sources to leverage for estimation is a valuable feature for treating sparse or noisy regions where sample data may be error prone and the true state is more effectively determined by the *a priori* prediction. In contrast, where the sample data are more reliable, they can contribute substantially to refine the process model's initial prediction. Note, after the initial "forward" pass ($i = 1, 2, \ldots, n$), a second "backward" pass ($i = n, n-1, \ldots, 1$) is executed for retrospective smoothing of estimates. Several approaches for the backward pass, e.g., [Rauch et al., 1965], are available and mitigate genomic directional biases and further reduce estimate uncertainty.

2.4. Noise Covariances and Additional Method Details

Strict theoretical optimality of the Kalman Filter as a minimum mean squared error estimator relies on a completely-modeled linear system and independent white Gaussian noise processes [Kalman, 1960]. These conditions are not expected to hold precisely in practical settings, but this does not preclude the Kalman Filter's effective application, especially if suitable adaptive checks are incorporated to adjust the noise covariances $\mathbf{Q}_i, \mathbf{R}_i$ and domain-specific knowledge is incorporated [Simon, 2006; Uhlmann and Julier, 2022]. Indeed, the performance of "adaptive KFs" is often competitive with more complex, computationally expensive algorithms, including those based in modern neural architectures [Greenberg et al., 2023]. In the following subsections, we discuss the default protocols employed by Consenrich to compute $\mathbf{Q}_i$ and $\mathbf{R}_i$ dynamically.

2.4.1. Adaptive Process Noise

A *bounded* adaptive process noise procedure (hereafter abbreviated "APN") is adopted. APN falls loosely within the adjustable level process noise paradigm [Bar-Shalom et al., 2002, Chapter 11.2] but considers the median (or more general q^{th} quantile) of uncertainty-scaled residuals for increased robustness to outliers and to avoid overactive adjustments to $\mathbf{Q}_i$.

At a given interval i, if we observe $\mathbf{z}_i$ that differs substantially from the *a priori* estimate $x_{i|i-1}$, it suggests a mismatch between our process model and the true system and should elicit increased scrutiny. This reasoning is carried out in practice by scaling up the process noise covariance $\mathbf{Q}_{i+1}$ to represent additional uncertainty and temporarily favor observations rather than $x_{i|i-1}$. We define

$$D_i = \mathsf{median}\left\{ \frac{y_{i,j}^2}{E_{i,jj}} \right\}_{j=1}^{j=m}, \tag{10}$$

that is, the median of squared, inverse-variance-weighted residuals. Let α_D denote a user-specified threshold on D_i to scale the process noise (Default: $\alpha_D = 2$). By default, if $D_i > \alpha_D$, we set $\mathbf{Q}_{i+1} = \left(\sqrt{d \cdot |D_i - \alpha_D| + 2}\right) \mathbf{Q}_i$. In contrast, if $D_i < \alpha_D$ but the process noise covariance was scaled up at a previous iteration, we scale down: $\mathbf{Q}_{i+1} = \left(\frac{1}{\sqrt{d \cdot |D_i - \alpha_D| + 2}}\right) \mathbf{Q}_i$. However, as a practical measure, the eigenvalues of $\mathbf{Q}_i$ are bounded in $[Q_{min}, Q_{max}]$ to prevent ill-conditioning and ensure estimates are not driven by the data or process model, exclusively.

2.4.2. Observation Noise Covariance

We evaluate "sparse segments" of contiguous intervals $r_{\mathsf{start}} \ldots r_{\mathsf{end}}$ that are (i) distant from highly enriched regions and/or previously annotated regulatory elements and (ii) exhibit low structural complexity (can be fit and detrended with a trivial signal model

via least squares regression). Reasoning that in a given sample j, counts in alignment count sequences $\mathbf{Z}[r_{\mathsf{start}} : r_{\mathsf{end}}, j]$ will often indicate technical artifacts, we accordingly approximate the local observation noise level at each interval $i = 1 \ldots n$ as follows:

1. Identify k proximal sparse segments from the *complement* of "widened" candidate active regulatory elements or data-specific enriched regions.
2. Take $R^{\mathsf{local}}_{i,jj}$ as the sample mean, median, or q^{th} quantile of squared values observed in the k proximal regions

We then compute the complete (global and local) observation noise as a weighted sum:

$$R_{i,jj} = \omega^{\mathsf{local}} R^{\mathsf{local}}_{i,jj} + \left(1 - \omega^{\mathsf{local}}\right) R^{\mathsf{global}}_{:,jj} \tag{11}$$

where default $\omega^{\mathsf{local}} = \frac{1}{3}$ and the global term can be computed as the average of local values genome-wide unrestricted by k. Organism-specific sparse segment annotations constructed using multiple encompassing annotations (e.g., [Luo et al., 2019; Gerstein et al., 2010]) are built-in to the software implementation of Consenrich for convenience.

These comprehensive annotations are useful for narrowing the search space for sparse, stationary subregions to the gaps between identified elements, but we emphasize Consenrich can empirically determine suitable sparse segments for effective approximation of sample-specific local noise variances genome-wide: Users can specify an annotation-free procedure in which initial regions of strong enrichment are quickly identified in their data with a simple peak detection heuristic (scipy.signal.find_peaks) [Virtanen et al., 2020]. The "gap regions" between these strongly enriched peaks are then partitioned and detrended. Following this step, a second check is performed for each gap region to verify the corresponding data is suitable for approximating the noise variances. Informally, if the autocorrelation of a near-baseline discrete-time signal, $A_{xx}(\tau)$, is maximal at $\tau = 0$ by a large margin compared to all time lags $\tau = 1, 2, \ldots$, 'true signals', which exhibit spatial dependencies and trends, are less likely to be a driving factor in the data. In these cases, the data will instead resemble zero-mean noise [Oppenheim et al., 2009]. This property enables straightforward approximation of local noise variances as the averages of squared data and mitigates conflation of noise and true signal. We thus check suitability of each preprocessed gap region described above for the purposes of approximating $R^{\mathsf{local}}_{i,jj}$ by verifying their respective autocorrelation functions experience a maximum at time lag $\tau = 0$ and that

$$A_{xx}(\tau = 0) > a \left(\max_{\tau=1,2,\ldots} A_{xx}(\tau) \right), \tag{12}$$

where default $a = 2$. Regions passing this check are then used as the 'sparse segments' to compute Eqn. (11).

We emphasize the distinction between the described procedure for mitigating effects of region-specific technical artifacts, which restricts noise level approximation to "sparse" genomic intervals, versus straightforward convolutions that can inflate approximations since they do not exclude data from proximal enriched regions. Though these approaches are often effective if supplemented with genome-wide control data to model a dynamic background, in the absence of such data, simultaneous extraction of both broad and narrow signals is particularly challenging [Zhang et al., 2008].

Algorithmically, computing observation noise covariances is not a bottleneck in our implementation, as a map from each genomic interval i to proximal sparse segments can be constructed efficiently: (i) Store and sort the locations of all sparse segments in an array, `s`, $len(s) = s \ll n$, which is $\mathcal{O}(s \log s)$ and only needs to be performed once; (ii) For each interval i, determine the insertion point of genomic interval i in `s`. That is, find `h` such that `s[h] < intervals[i] <= s[h+1]` via bisection ($\mathcal{O}(\log s)$); and (iii) Scan `s[h-k:h+k]` to determine the nearest sparse segments ($\mathcal{O}(k)$). This yields a modest time complexity $\mathcal{O}\left(n(\log(s) + k)\right)$. In practice, runtime is significantly reduced with several optimizations not reflected in the above Big O notation. For genomes spanning several billion base pairs, such as human and mice, we propose $\log(s)$ need not exceed 20 and the default $k = 25$ provides sufficient data for dynamic approximation of technical noise. Note, as for APN (Section 2.4.1), eigenvalues in $\mathbf{R}_i$ are likewise restricted to an interval $[R_{min}, R_{max}]$.

2.4.3. Scalable Inversion of the Residual Covariance Matrix

To compute the gain $\mathbf{K}_i$ requires inversion of the residuals' covariance matrix $\mathbf{E}_i \in \mathbb{R}^{m \times m}$ (see Eqn. 7). General matrix inversion is $O(m^3)$ with straightforward approaches and could incur prohibitive computational expense for large sample sizes. Fortunately, our problem setup offers several opportunities for computational reduction to ensure scalability.

Let $\mathbf{v} = \left(\sqrt{P_{i|i-1,11}}\right) \cdot \mathbf{1} \in \mathbb{R}^m$ and rewrite Eqn. (6) as

$$\mathbf{E}_i = \mathbf{v}\mathbf{v}^\top + \mathbf{R}_i \tag{13}$$

where $\mathbf{R}_i$ is positive-definite and can be quickly inverted. The outer product $\mathbf{v}\mathbf{v}^\top$ is rank-one, and the inverse of $\mathbf{E}_i$ can be obtained by applying the Woodbury Identity to write:

$$\mathbf{E}_i^{-1} = \mathbf{R}_i^{-1} - \frac{P_{i|i-1,11} \cdot \mathsf{diag}\left(\mathbf{R}_i^{-1}\right) \mathsf{diag}\left(\mathbf{R}_i^{-1}\right)^\top}{1 + P_{i|i-1,11} \cdot \mathsf{Trace}(\mathbf{R}_i^{-1})} \tag{14}$$

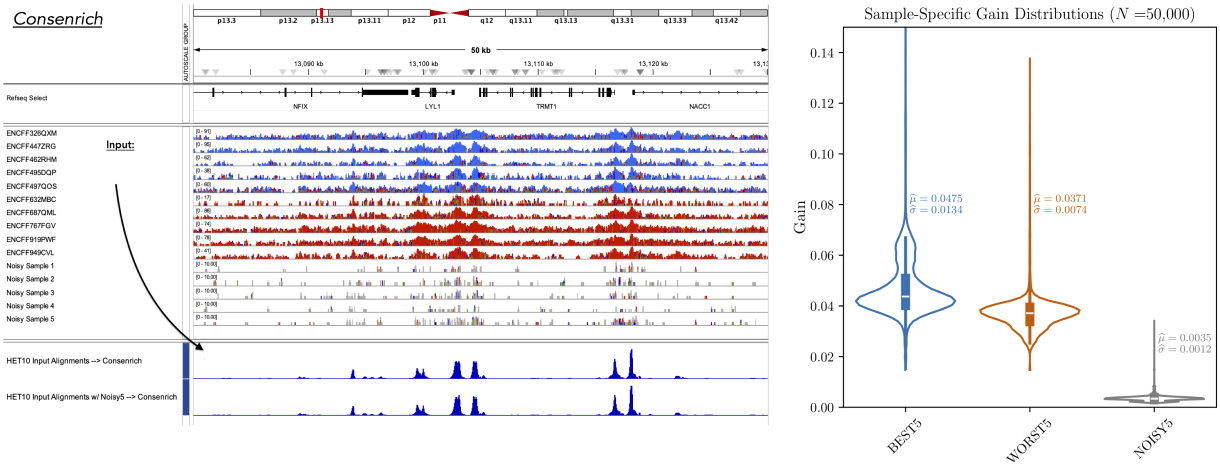
This reduces runtime to $O(m^2)$, with the outer product $\mathsf{diag}\left(\mathbf{R}_i^{-1}\right) \mathsf{diag}\left(\mathbf{R}_i^{-1}\right)^\top$ as the source of quadratic complexity in m. We see further computational reduction in practice by storing $\mathbf{E}_i^{-1}$ in the form of Eqn. (14) such that if $\mathbf{R}_i = \mathbf{R}_{i-1}$, we only need to substitute the $P_{i|i-1,11}$ term to obtain the inverse at the current interval. Note, the set of intervals such that $\mathbf{R}_i = \mathbf{R}_{i-1}$ is computed efficiently by storing and later comparing hashed sums of the matrices while running the procedure discussed in in Section 2.4.2.

2.5. Length of Genomic Intervals

The specified length of genomic intervals (default $L = 50$) has consequences for the spatial resolution of extracted signals and computational expense. The runtime of Consenrich is linear in the number of genomic intervals which is inversely related to L:

$$n \approx \frac{\text{Base-Pair Length of Whole Genomic Region}}{L}.$$

We have found the default $L = 50$ to provide reasonable performance across a range of ATAC-seq, ChIP-seq, and CUT-N-RUN experiments while offering computational efficiency such that Consenrich can be run genome-wide in lightweight personal computing environments; However, users can tune L to their specific experiments and available computational resources.



(a) Consenrich signal extraction with and without Noisy5

(b) Data integration from samples of varying (net) quality

Fig. 2: (**a**) Consenrich is used to analyze fifteen ATAC-seq samples of highly varying data quality in total (As defined in Section 3: Best5 - Blue, Worst5 - Orange, Noisy5 - Gray). Consenrich is run twice—with and without Noisy5 inputs—and results displayed in the bottom panel ('→ Consenrich). In each case, Consenrich yields consistent, spatially-resolved signal tracks. (**b**) The violin plots of gains (Eqn. 7) for each of the three sample quality categories reflect a flexible integration of data from samples with region-dependent informative value that is not reflected in net summary statistics.

3. Experiments and Applications

We posit that the greatest factors limiting an accurate genome-wide estimation of molecular states across multiple samples is the variability in overall signal quality across samples and the variability in noise profiles at different genomic locations. Thus, we designed experiments to determine Consenrich's ability to effectively scale and integrate data from high-and-low noise samples and genomic regions. We further evaluated how well Consenrich recovered spatial features and enriched regions indicative of potential regulatory patterns. In all experiments, publicly available ATAC-seq and ChIP-seq datasets from ENCODE [Luo et al., 2019] were used. Note, unless otherwise stated, Consenrich was run with default parameters.

3.1. Dynamic, Uncertainty-Moderated Integration of Multi-Sample HTS Data

We first evaluated how Consenrich integrated information from multiple samples of varying data quality. To do this, we analyzed ATAC-seq data from lymphoblastoid cell lines derived from individual donors (Figure 2a). Samples were ranked based on their transcription start site (TSS) enrichment score; an average of genome-wide SNR-like measures near ubiquitously accessible TSSs that is a standard quality control measure of ATAC-seq data. We defined a 'Het10' dataset, constructed from the five lowest-ranked samples ('Worst5') and five highest-ranked samples ('Best5') in the total set of $N = 56$ lymphoblastoid ATAC-seq samples made available publicly by ENCODE. Additionally, we simulated five noisy samples with uniformly low SNR across the genome ('Noisy5') by merging randomly shuffled reads with subsampled alignments from the set of $N^* = 56 - 10$ alignments from ENCODE that were excluded from Het10. The 'Het10 with Noisy5' dataset combined both sets for a total sample size of fifteen.

As described previously, we reasoned that information from high quality samples should have greater influence on the final signal estimation (in total). Therefore, we compared Consenrich results using the Het10 dataset alone and Het10 with Noisy5 dataset (Figure 2a, bottom tracks). The influence of each groups' samples on *a posteriori* estimates was quantified by their respective gains (Eqn. 7). Recall, for given *a priori* estimate uncertainty, high-valued gains represent greater weight/confidence ascribed to the respective sample in determining the *a posteriori* estimates (Eqn. 8). As desired, we see that samples from the Best5 group had greater influence than samples in the Worst5 group overall. Consistently noisy samples should exhibit a minimal effect on the estimates, however. We accordingly observed average gains and 99%-confidence intervals of ($.0475 \pm .0134$), ($.0371 \pm .001$), ($.0035 \pm .0012$) for samples in Best5, Worst5, and Noisy5, respectively (Figure 2b). We note, however, multiple instances where samples with low TSS enrichment scores offered greater certainty *locally* than higher TSS enrichment samples (For example, see TSS enrichment scores of Best5 and Worst5 in Figure 3). When performing HTS experiments, subregions of the genome are often consistently noisy across a majority of samples, while other regions of high uncertainty may be unique to individual replicates (Examples of both cases are enclosed in the white, dashed regions of Figure 3). The dynamic observation noise model (Section 2.4.2) of Consenrich leverages these instances to extract information from samples of varying overall quality where they are most informative and does not forfeit potential insight by uniformly downweighting or discarding samples deemed unreliable with crude summary statistics. For instance, we note the overlap between tails in the gain distributions of Best5 and Worst5 observed in Figure 2b.

Consenrich allows for adjustment of the genomic interval size (default $L = 50$bp) to obtain the desired spatial signal resolution or account precisely for factors related to sequencing and alignment. Here, the benefits of smaller intervals did not appear to

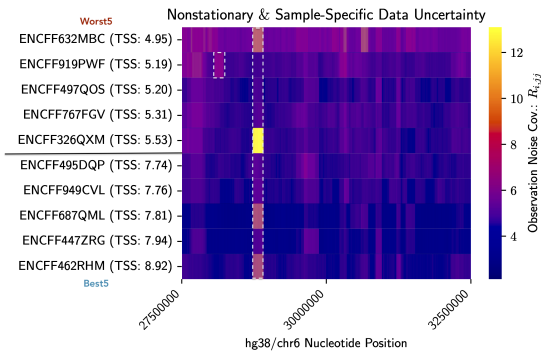


Fig. 3: The Consenrich-approximated observation noise variances (Section 2.4.2) are displayed over chr6:27,500,000-32,500,000. Certain subregions are more consistently noisy across a majority of samples (e.g., region-specific noise enclosed in the rightmost dashed rectangle near 29mb). Other regions of high uncertainty are unique to individual samples (e.g., white/dashed rectangle in row ENCFF919PWF near 28.20mb).

significantly improve signal capture (Figure 4) justifying additional computational expense.

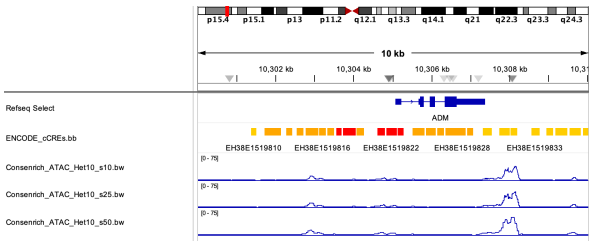


Fig. 4: Consenrich results for the Het10 (Best5,Worst5) ATAC-seq dataset over a 10kb region in chr11 are displayed. Three Consenrich signal tracks generated with $L = 50$bp, $L = 25$bp, and $L = 10$bp are shown.

Next we qualitatively assessed Consenrich's ability to robustly estimate signals and preserve spatial features thereof in the presence of noisy data. In Figure 5, we further compared the estimated signal returned by Consenrich from datasets with and without Noisy5 samples. Over this region, we observed a negligible influence from noisy samples as evidenced by a mean squared difference of 0.043 in the signal across the region. In contrast, the estimate sequences returned by a least-squares smoothing filter exhibited a mean squared difference of 1.675. We compared the estimated signals in their approximated time-frequency representations in Figure 5 via consecutive discrete Fourier transforms [Oppenheim et al., 2009, Chapter 10]. We note that Consenrich effectively tempered the influence of noisy data while preserving subtle but legitimate high frequency components that were spatially informative.

We then sought to exhaust the estimation capabilities of Consenrich despite the presence of uniformly low-quality inputs. In this experiment, additional noisy samples were generated iteratively and appended to the Het10 dataset until a threshold on the minimum 'similarity' between the output Consenrich signal tracks with-and-without these $k = 1, 2, \ldots$ noisy samples was no longer satisfied. We defined the threshold on acceptable similarity as a Pearson correlation coefficient greater than .95, computed genome-wide with respect to all positions yielding a non-zero estimate. Consenrich produced signal tracks satisfying $\text{Pearson}\Big(\text{Consenrich(Het10)}, \text{Consenrich(Het10 + Noisy_k)}\Big) > 0.95$ up to $k = 10$ (half of total samples) as recorded in Table 1.

Table 1. Genome-Wide Pearson correlations between Consenrich signal track outputs for increasing numbers of noisy input samples.

	Het10	Het10 + Noisy5	Het10 + Noisy10
Het10	1.00	0.98	0.96
with Noisy5	.	1.00	0.98
with Noisy10	.	.	1.00

Together, these properties of Consenrich allow for the confident inclusion of samples without concerns that poor quality samples would negatively affect annotations—while still being able to contribute information at a subset of intervals where they are deemed reliable.

3.2. Increased Relative Enrichment in ATAC-seq

To demonstrate that Consenrich's extracted signal shows greater enrichment in discrete genomic regulatory elements than a method simply merging alignments, we analyzed the relative enrichment at $N = 500$ randomly-selected promoter regions (with nonzero, above-median coverage values after outlier removal) in the Het10 dataset between Consenrich and merged alignments (Figure 6). The Merged Alignments signal track was generated by combining multiple samples' sequence alignments into a single BAM file and subsequently quantifying coverage over genomic intervals $i = 1 \ldots n$ with deepTools' bamCoverage utility. The RPGC (1x-genome coverage) normalization method was used to obtain interpretable, scaled coverage values for comparison with Consenrich estimates [Ramírez et al., 2016] at the same resolution.

Due to the large dynamic range of signal tracks evident in their empirical cumulative distribution functions (ECDFs) displayed in Figure 6, the median is used for comparison of central tendency. We find that Consenrich recorded a median promoter enrichment of 42.25 compared to Merged Alignments' 22.56.

Many standard analyses require discrete, consensus peak annotations. We aimed to qualitatively assess whether Consenrich signal tracks offered potential to reliably convey multiple samples' features for determining consensus peaks. For visualization, we called peaks on Consenrich (blue) and Merged Alignments (red) representations of the Het10 inputs (Figure 7, bottom panel). To mitigate dependence on a single peak calling method, we included results from two peak callers that operate on signal tracks and appeal to different inference paradigms, LanceOtron [Hentges et al., 2022] and ROCCO [Hamilton and Furey, 2023]. In general, a robust encompassing ground truth sets of regulatory regions cannot be established with high confidence. We stress that this qualitative analysis is not for comparing peak callers but their relative performance as a function of the provided input. We provide further biological context for the results by displaying a union set of conservative, Irreproducible Discovery Rate (IDR)-thresholded ChIP-seq peak results [Li et al., 2011] from 109 high quality experiments across a wide range of transcription factor binding loci in a lymphoblast cell line (black) and the ENCODE cCREs annotation [Luo et al., 2019]. Over the displayed

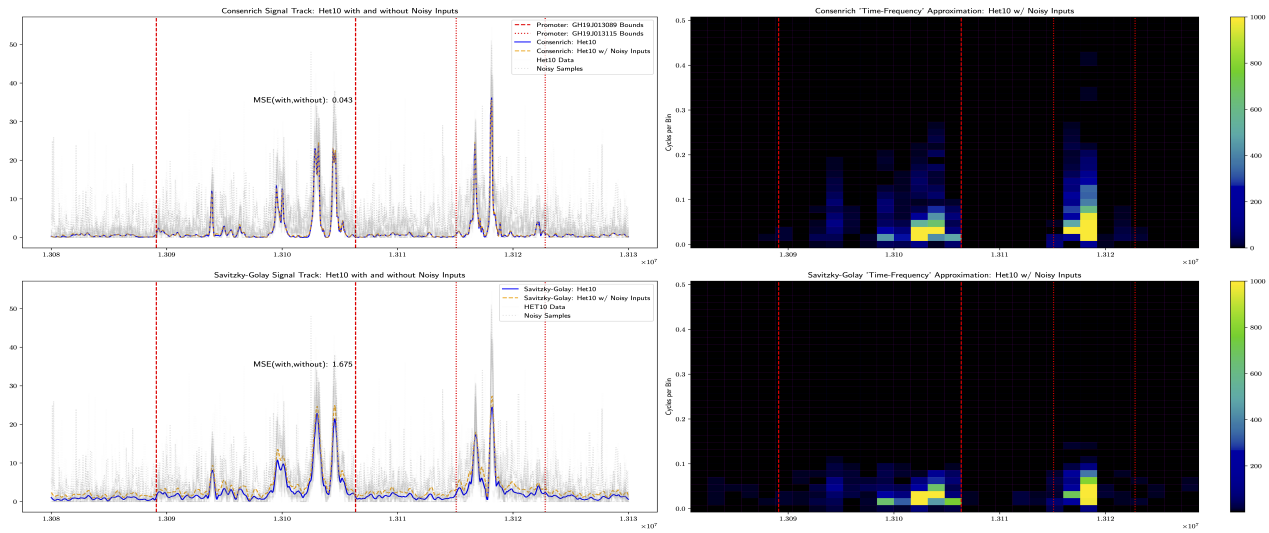


Fig. 5: *Simultaneous Noise Reduction and Preservation of Signal Structure*. We compare performance with a generic smoothing filter (Savitzky-Golay: polynomial degree = 5, window size = 1kb). Both the original (left) and time-frequency (right, colored according to power spectral density) representations of estimates are considered. Consenrich (first row) is negligibly affected by the inclusion of noisy samples and successfully identifies legitimate signals across a greater range of frequencies for improved capture of spatial signatures. In contrast, a generic Savitzky-Golay filter yields values uniformly shifted by Noisy5 samples and discards high frequency components as noise.

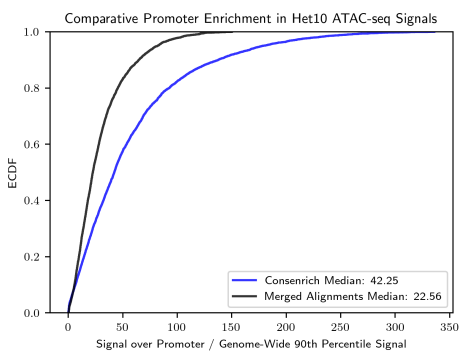


Fig. 6: The empirical cumulative distribution functions (ECDFs) of each method's signal values over five-hundred promoter regions, divided by their genome-wide 90th percentile scores are displayed.

35kb region in human chromosome 6, both peak calling methods showed improved detection abilities using the Consenrich signal track as input. For instance, both ROCCO and LanceOtron better subdivided the distinct enhancer-like regions within (A, B) when using the Consenrich signal track (blue) as input. In contrast, given the Merged Alignments signal track (red) as input, ROCCO crudely annotated the elements in (B) as a single peak, and LanceOtron returns only one peak in total. Together, these analyses provide support for Consenrich's ability to better integrate data across multiple samples in defining regions of enrichment. We note the current implementation of Consenrich provides convenient pairing with ROCCO: Users can directly supply the Consenrich signal track output or an uncertainty-penalized measure of enrichment (`--ratio_bigwig`) as input to ROCCO.

3.3. Multi-Sample ChIP-seq with Input Control Data

Several HTS experiments generate sample control data to account for sample-specific biases. We evaluated the ability of Consenrich to integrate paired treatment and control data from multiple ChIP-seq samples. We ran Consenrich using ChIP-seq treatment/control pairs from $N = 6$ biological replicates (individual human donors/colon tissues) targeting POL2RA. For reference, a measure of enrichment was computed for the same data by supplying treatment and control pileup files to `macs3 bdgcmp -m logFE`. Both approaches first scale experiments' treatment and control data to each other before applying their respective techniques. Since POL2RA is known to show the strongest binding signal in promoter regions, we quantitatively analyzed genome-wide enrichment patterns across 500 randomly selected promoter-like signatures (PLS) as identified by ENCODE [Luo et al., 2019] (Figure 8). After exponentiating the logFE signal values for comparable scaling and dividing each method's promoter-overlapping signal values by their 90th percentile values genome-wide, Consenrich recorded a median promoter enrichment of 10.80, while the log-ratio-based approach (`macs3 bdgcmp -m logFE`) recorded a median promoter enrichment of 4.68. Altogether, these indicate that Consenrich can successfully integrate control data into its model and suggest it can increase signal enrichment allowing for use with multiple functional genomics assays.

3.4. Empirical Distribution of Residuals

As discussed in Section 2, a well-constructed system model and filter will yield *a priori* residuals that resemble a zero-mean noise. This property is not expected to hold genome-wide due to abrupt shifts in epigenetic activity. Such cases are accounted for with Consenrich's APN heuristic (Section 2.4.1), and in net, we can still rely on this property of residuals as a cursory validation of the process model's *a priori* estimates (Eqn. 3). Using the ChIP-seq

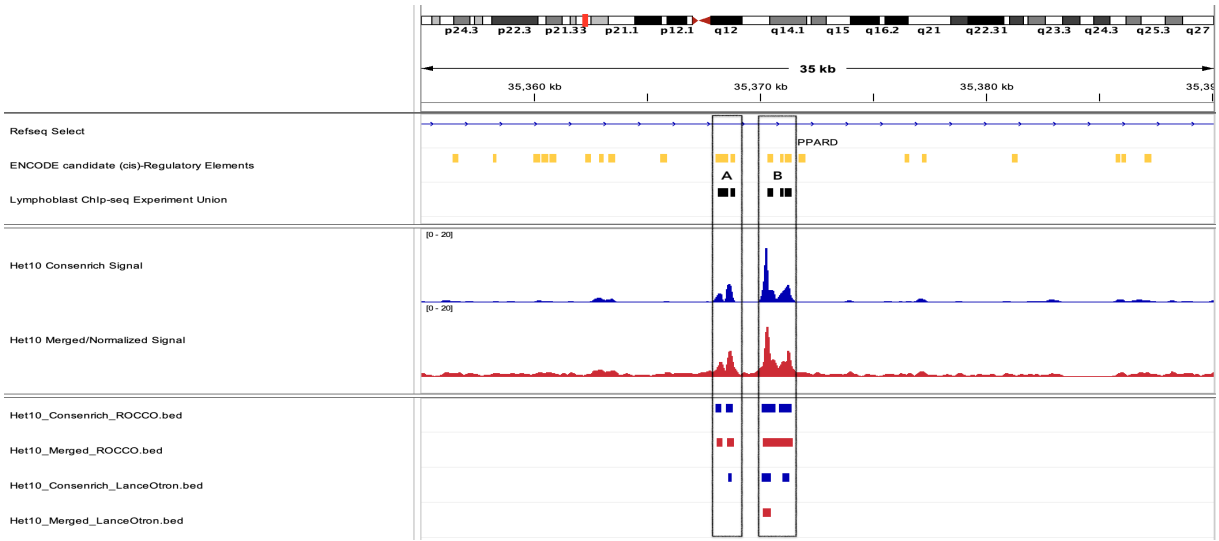


Fig. 7: The Het10 ATAC-seq samples (lymphoblastoid) are supplied as input to Consenrich and the 'Merged Alignments' aggregation/normalization procedure discussed in Section 3.2. The resulting signal tracks are used as input to peak callers ROCCO and LanceOTron. Peak annotations are colored based on the inputs from which they were derived: Consenrich (blue) or Merged Alignments (red).

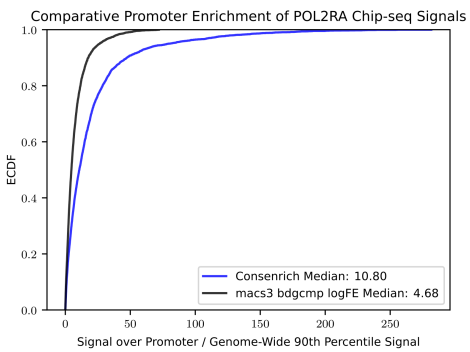


Fig. 8: Six donor/tissue-level POL2RA ChIP-seq signals, normalized by their respective controls were analyzed in with Consenrich and MACS3 `bdgcmp -m LFE` (fold-enrichment based on MACS3's treatment pileups and λ control/background estimates). The empirical cumulative distribution functions (ECDFs) of each method's signal values over five-hundred randomly-selected nonzero promoter regions, divided by their genome-wide 90th percentile scores are displayed.

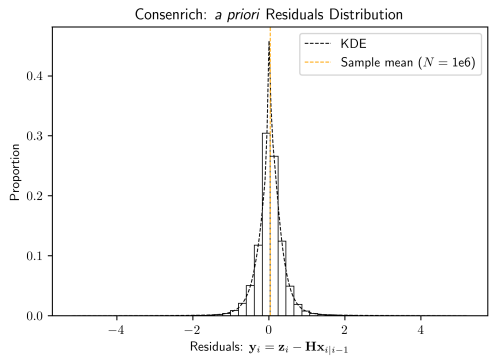
data from Section 3.3, we evaluated *a priori* residuals over one-million randomly selected intervals genome-wide. A sample mean $y_{\hat{\mu}} = 0.034$ and standard deviation $y_{\hat{\sigma}} = 0.492$ with a symmetric empirical distribution (Figure 9) were observed.

The distribution of residuals also provides for calibrated uncertainty quantification (UQ) [Zarchan and Musoff, 2015, Chapter 14]. Several UQ metrics are supported in the software implementation. Both *a priori* and post-fit-based measures offer information and can be specified for tracking. For example, the *post hoc*, inverse variance weighted (IWW) residual track can be used as a reference for comparison with the Consenrich signal estimates themselves to evaluate errors in light of observation noise and the final estimate covariance.

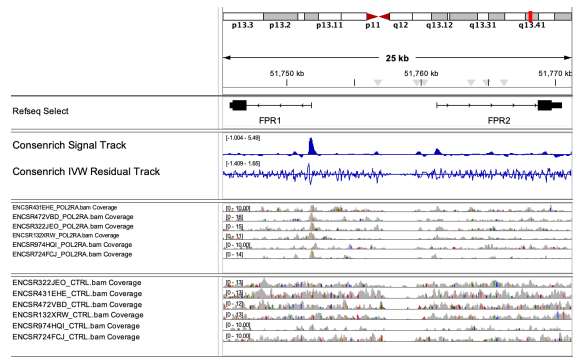
4. Conclusion

Consenrich represents substantial methodological novelty as an effective estimator of real-valued signals from multi-sample functional genomics data. This capability offers general appeal for deriving genome-wide annotations that reflect both enrichment and informative spatial/structural patterns in signals and also for conventional multi-sample enrichment analyses such as consensus peak calling. For future work, we note that the model underlying Consenrich can be extended to include support for single-cell analyses. Notably, the process model of Consenrich, which propagates forward previous states along an inferred trajectory, bears considerable potential in the single-cell setting where data is notoriously sparse [Granja et al., 2021], and significant patterns in low-coverage regions are difficult to capture from the data alone. In addition to extracting consensus signal annotations in one assay across many samples, Consenrich is well-suited to to integrate data from multiple related assays (e.g., ATAC-seq and DNase-seq) given its sample-specific observation model and noise approximation. For highly heterogeneous datasets, a "filter bank" approach can be implemented that runs distinct Consenrich instances in parallel and 'switches' at each iteration to the the instance yielding the most confident estimates [Zarchan and Musoff, 2015, Chapter 16].

Producing high-resolution genome-wide annotations also presents opportunities for developing alternative methods for identifying regions with altered molecular characteristics. Instead of defining discrete peaks prior to a differential analysis based on variation in signal enrichment, Consenrich signals derived from two sets of phenotypically distinct samples could be used to better pinpoint regions of significant change using a combination of spectral and conventional enrichment-based analyses. Furthermore, the ability of Consenrich to capture spatial signal patterns from multiple noisy samples with high confidence makes it particularly agreeable to statistical signal processing and wavelet analysis techniques targeted to precise structural patterns of biological interest [Du et al., 2006].



(a) Histogram and kernel density estimate of *a priori* residuals.



(b) Consenrich signal estimates and post-fit IVW residuals.

Fig. 9: *Empirical Distribution of Residuals.* ChIP-seq data from $N = 6$ experiments targeting POL2RA in independent donors' colon tissues (Section 3) are analyzed with Consenrich. (**a**) The distribution of *a priori* residuals (Eqn 5) are depicted in aggregate: Over $N = 1e6$ randomly selected intervals genome-wide, a sample mean $\hat{\mu} = 0.034$ and standard deviation $\hat{\sigma} = 0.492$ were observed. (**b**) A subset of post-fit inverse-variance-weighted residuals are displayed as a sequence over 25kb.

References

Y. Bar-Shalom, X. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software.* Wiley, Jan. 2002.

M. R. Corces et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), Oct. 2018.

P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, July 2006.

M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, et al. Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330(6012):1775–1787, Dec. 2010.

J. M. Granja, M. R. Corces, Pierce, et al. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3):403–411, Feb. 2021.

I. Greenberg, N. Yannay, and S. Mannor. Optimization or architecture: how to hack kalman filtering. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2023.

N. H. Hamilton and T. S. Furey. ROCCO: a robust method for detection of open chromatin via convex optimization. *Bioinformatics*, 39(12), Dec. 2023.

L. D. Hentges, M. J. Sergeant, C. B. Cole, D. J. Downes, J. R. Hughes, and S. Taylor. Lanceotron: a deep learning peak caller for genome sequencing experiments. *Bioinformatics*, 38(18): 4255–4263, July 2022.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), Mar. 1960.

Q. Li, J. B. Brown, H. Huang, et al. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3), 2011.

Y. Luo et al. New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Research*, 48 (D1):D882–D889, 2019.

C. A. Meyer and X. S. Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11), Sept. 2014.

A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-Time Signal Processing.* Pearson Education, 3rd edition, 2009. ISBN 978-0131988422.

F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, and T. Manke. deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, Apr. 2016.

H. E. Rauch, F. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8), Aug. 1965.

J. W. Schroeder and P. L. Freddolino. Enricherator: A bayesian method for inferring regularized genome-wide enrichments from sequencing count data. *Journal of Molecular Biology*, Apr. 2024.

D. Simon. *Optimal State Estimation.* Wiley, Jan. 2006.

J. Uhlmann and S. J. Julier. Gaussianity and the kalman filter: A simple yet complicated relationship. *Journal de Ciencia e Ingeniería*, 14(1):21–26, June 2022.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, Mar. 2020.

Y. Yang et al. Leveraging biological replicates to improve analysis in chip-seq experiments. *Computational and Structural Biotechnology Journal*, 9(13), 2014.

P. Zarchan and H. Musoff. *Fundamentals of Kalman Filtering: A Practical Approach, Fourth Edition.* American Institute of Aeronautics and Astronautics, Inc., Jan. 2015.

Y. Zhang, T. Liu, C. A. Meyer, et al. Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9), Sept. 2008.

N. Zhao and A. P. Boyle. F-Seq2: improving the feature density based peak caller with dynamic statistics. *NAR Genomics and Bioinformatics*, 3(1), 02 2021.