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

Nolan H. Hamilton[1,2,4], Yu-Chen E. Huang[1,2,4], Benjamin D. McMichael[1,2], Michael I. Love[1,3], Terrence S. Furey[1,2,*]

[1]Department of Genetics, University of North Carolina at Chapel Hill
[2]Department of Biology, University of North Carolina at Chapel Hill
[3]Department of Biostatistics, University of North Carolina at Chapel Hill
[4]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill

[*]Correspondence to: tsfurey@email.unc.edu

Abstract

We present Consenrich, a simple but principled technique for genome-wide estimation of signals hidden in noisy multi-sample sequencing-based functional genomics datasets. Consenrich appeals to a sequential prediction-correction framework and models both the spatial dependencies between proximal loci and regional, sample-specific noise processes that corrupt sequencing data. Experiments reveal distinct improvement compared to benchmarks in a series of challenging estimation problems, where noisy functional genomics data samples must be reconciled. We further highlight the immediate practical appeal of this refined signal extraction for differential analyses between disease conditions and identification of functionally enriched genomic regions. A complete implementation of Consenrich is hosted at https://github.com/nolan-h-hamilton/Consenrich.

Keywords: functional genomics, multi-sample data integration, chromatin profiling, regulatory element discovery, signal quantification, uncertainty quantification, differential analysis, genomic signal processing

Background

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 [1, 2]. Data generated using these assays are translated to genome-wide annotations by first mapping sequenced reads to specific genomic positions. Targeted assay-specific genomic features such as accessible chromatin (ATAC-seq, DNase-seq) and transcription factor binding sites (ChIP-seq, CUT-N-RUN) can then be defined by determining variably-sized discrete regions with a significant enrichment in aligned read counts, referred to as *peaks* [3]. These peaks offer practical utility for general genomic annotation tasks and can serve as candidate regions over which to test for significant differences in the target molecular phenotype between groups of samples.

However, the quantification of signals that are used to determine peaks—particularly in the multi-sample setting—has received little attention in prior work [4]. Indeed, multi-sample consensus peak annotations are generally determined after first calling peaks on individual samples' quantified signals that, in isolation, do not reinforce consistent, reproducible patterns across samples. After deriving sample-specific peaks, heuristic *post hoc* merging strategies are commonly employed [5, 6]. These methods impose discretionary, rigid thresholds on the minimum degree of overlap across samples and/or number of samples in which a candidate peak must be present 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, and peak boundaries may be affected by sample-specific noise resulting in overly broad annotations.

We aimed to rectify the above-mentioned issues by developing an algorithm for molecular-level state estimation, capable of elucidating well-evidenced signal features and attenuating noise. The proposed method, *Consenrich*,

jointly integrates multiple samples' sequence alignment data to define a robust, mutually reinforced signal profile representing the molecular phenotype of interest. Two fundamental challenges impeding this effort were given special consideration. First, samples often vary significantly in data fidelity due to a number of factors such as inconsistent preparation and preservation of biosamples, data generation across multiple time periods and technologies, and other artifacts. Second, region-specific factors can provoke uncertainty at distinct loci, even in samples that are reliable elsewhere. Previous statistically sound methods have been proposed to account to address some of these challenging aspects [4] but are catered to a specific assays and/or smaller genomes ($<$ 50mb). Efficient *ad hoc* techniques are commonly applied, such as pooling samples' sequenced reads [7, 8], but these approaches may fail to accurately characterize joint signal profiles in imbalanced, mixed-quality datasets.

In a wide array of analytical experiments that include noisy, heterogeneous input datasets, we find that Consenrich effectively separates signal from noise to yield strong enrichment at validated targets and elucidate spatial features. This increased enrichment likewise improves discrete annotation tasks such as consensus peak calling and subsequent differential analyses based on consensus peaks. Applied to a class-imbalanced Alzheimer's Disease (AD) chromatin accessibility dataset, peaks called on the Consenrich-estimated signal track reveal substantially more differentially accessible regions (DARs) than conventional alternatives. Further these DARs show substantially greater overlap at well-established AD risk genes and quantitative trait loci affecting gene expression (eQTL) than the alternative methods.

Results

Consenrich provides an efficient, explainable model that integrates biological reasoning with observed functional genomics data from multiple samples. Consenrich shifts focus to the quantification stage of enrichment analyses, as opposed to the discrete annotation stage, aiming to characterize high-resolution, genome-wide profiles of the epigenetic landscape. These quantitative profiles are informative in themselves, but also enhance discrete annotation tasks downstream for multi-sample analyses that hinge on accurate overall signal estimates.

At each genomic interval, an initial *a priori* prediction step is carried out that appeals to a basic notion of spatial dependence between proximal loci: The initial *state estimate* step is based on forward-and-backward-propagated signals from nearby genomic intervals. Subsequently, in the *correction* step, Consenrich integrates alignment coverage data from each sample to obtain the *a posteriori* estimate. The magnitude of this correction is determined through the *gain*, discussed in Methods. Informally, the gain represents a trade-off between the reliability of the initial propagation-based estimate and the reliability of observed data—the latter of which is approximated individually for each sample using a custom uncertainty model to address the multitude of region-and-sample-specific sources of noise in functional genomics data such as GC-content, low sequence complexity and mappability, bottlenecking due to sonication bias, and so on [9]. A simplified schematic is presented in Figure 1, where data uncertainty within specific genomic intervals and samples is depicted.

While these features provide technical appeal for the task of estimating real-valued molecular states, more concretely, they enhance practical downstream analyses such as consensus peak calling and differential analysis. Note, unless otherwise stated, Consenrich was run using its default configuration. Several experiments addressing scalability for growing sample sizes and sensitivity to parameter configurations are diverted to Supplementary Material.

Consenrich yields robust signal estimates despite heterogeneous, noisy input samples

A key challenge in combining data across samples is accounting for varying data quality, where noisy, low quality data samples can dilute valuable information conveyed in less noisy, high quality data samples. This challenge is exacerbated in HTS data, where noise processes affecting data are nonstationary due to region-dependent technical artifacts. Therefore, we first evaluated the estimation performance of Consenrich using mixed-quality ATAC-seq data from lymphoblastoid cell lines and intentionally noisy samples. We used the transcription start site (TSS) enrichment score—a signal-to-noise QC metric defined by ENCODE for evaluation of ATAC-seq data—to rank and define the *Best5* and *Worst5* lymphoblastoid samples from a set of fifty-six total samples. Note, the Best5 and Worst5 are collectively referred to as the Het10 lymphoblastoid samples. Additionally, variable numbers of intentionally-

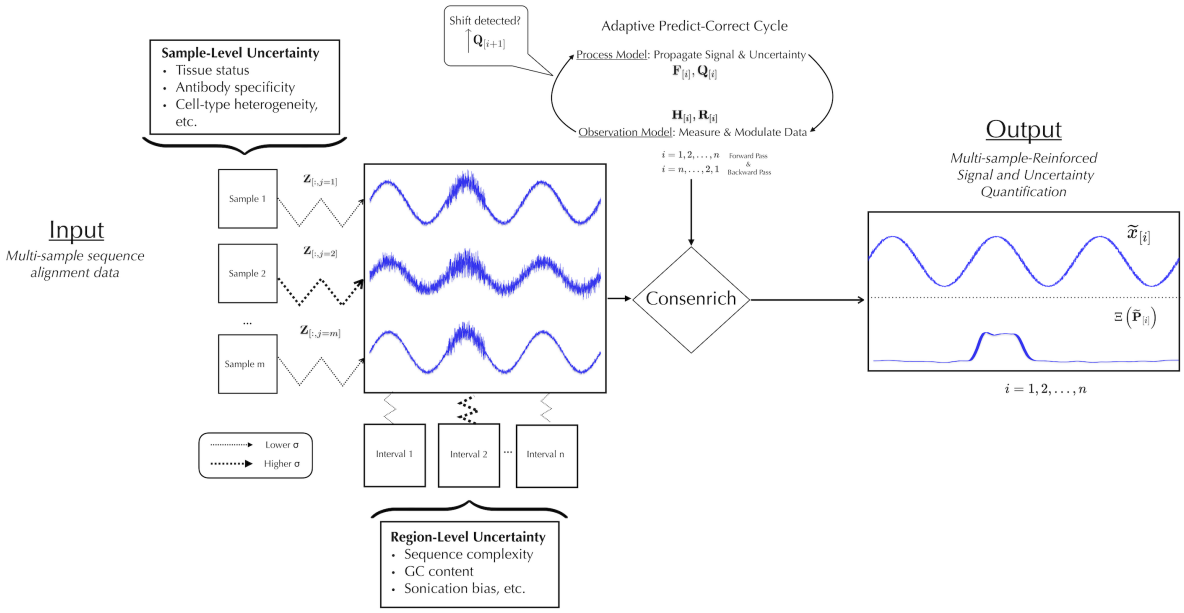


Figure 1: *Conceptual Schematic.* Given sequence alignments from multiple independent samples (BAM format), Consenrich outputs genome-wide annotations of consensus signal estimates and corresponding uncertainty measures (bigWig format). Underlying Consenrich are simple but explicit models constructed to leverage both (i) spatial dependencies between genomic regions and (ii) noise processes arising from both regional artifacts (e.g., 'Interval 2' in the schematic) and sample-specific artifacts (e.g., 'Sample 2' in the schematic). Note, $\widetilde{x}_{[i]}$ denotes the estimated state at genomic interval i, and Ξ denotes some scalar uncertainty function (the trace of the estimated state covariance, sum of standardized residuals, etc.).

noised samples, largely devoid of true signal (Noisy10, Noisy20), were appended to the Het10 dataset in evaluations (Supplementary Material A.1)

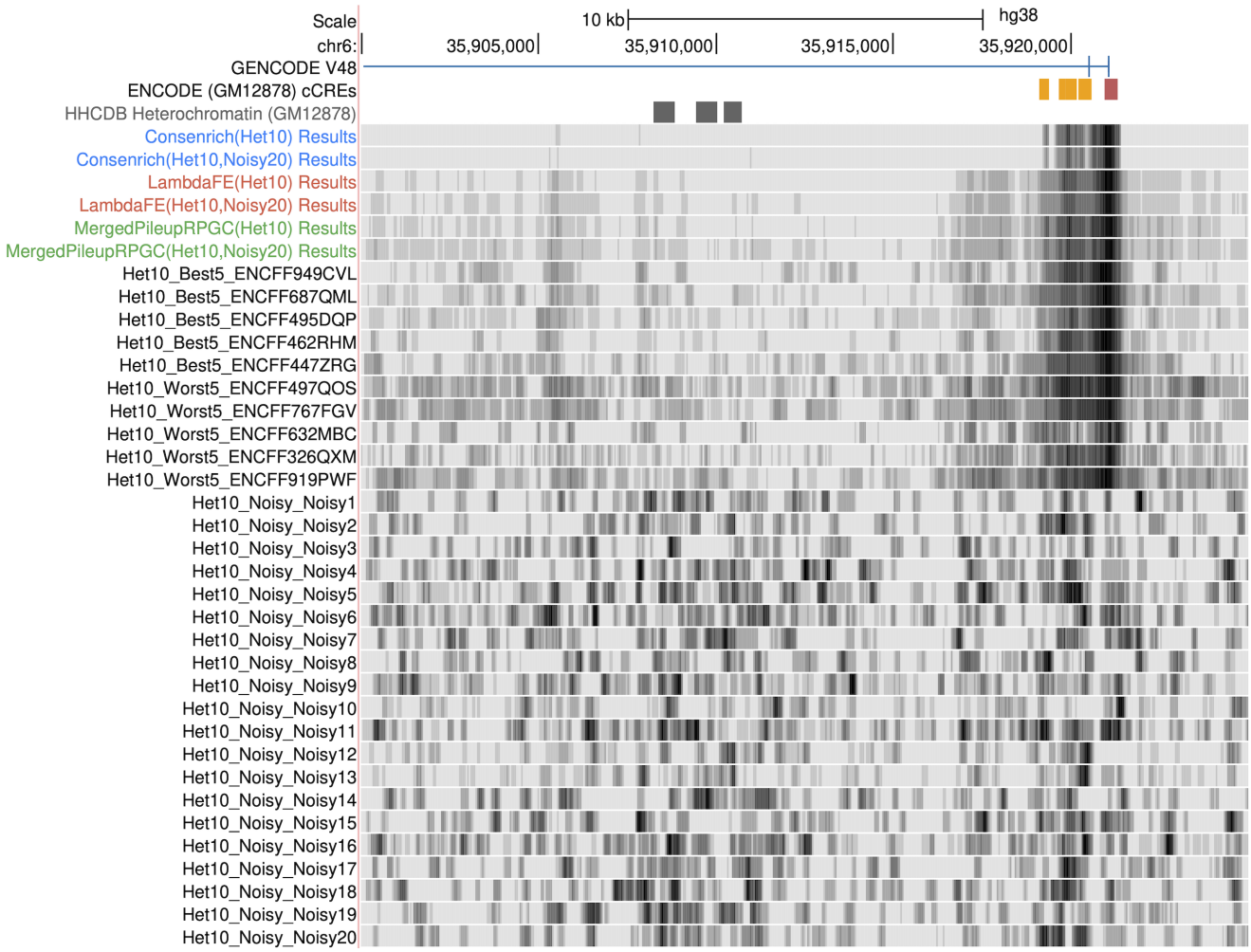
For benchmarking, we include two signal quantification methods that are widely used in practice for downstream peak calling and general exploratory data analysis: LambdaFE computes "fold change over control" signal tracks using the same protocol employed by ENCODE for ATAC-seq[1] and ChIP-seq experiments[2] experiments; and `MergedPileupRPGC` that merges reads from all samples and applies `deeptools`' "reads per genomic coverage" (RPGC) normalization to scale counts according to the "effective genome size" [10]. Both benchmarks provide a genome-wide, track-like quantitative summary of the same input alignments as Consenrich.

We first measured each method's output over positive control regions, referred to in Figure 2 as *Active*. This set consists of 14,834 ubiquitous DNase1 peak clusters confirmed in at least $\frac{50}{95}$ cell lines by ENCODE. For reference to an empirical background (*Null*), corresponding null distributions were approximated for each method using 1000 segmented block bootstrap samples [11] of the *Active* regions, where the segmentation was based on gene density (See Section Datasets and Evaluation). We likewise measured each method's output over 4,876 GM12878-specific heterochromatin regions from the Human Heterochromatin Database (HHCB) [12] (*Inactive*).

Figure 2a includes a snapshot from the UCSC Genome Browser [13] encompassing the relevant aspects of this experiment and demonstrating qualitative behavior of the methods given varying inputs. To quantitatively assess the genome-wide signal enrichment at relevant regulatory regions, we calculated log-ratios between *Active* and *Null* regions (Figure 2b). Consenrich showed the strongest evidence of significant enrichment over *Active* genomic regions compared to those representing the null hypothesis—that is, the `nullranges::bootRanges` generated *Null* genomic regions. We also calculated the average signals in heterochromatic regions, where accessible chromatin should be largely absent. Here, Consenrich signals are an order of magnitude smaller than other methods (Figure 2b).

To relate the methodological basis of Consenrich with its improved performance, we examined the gains (Equation 13) of each sample (i) in aggregate in Figure 3a, and (ii) sequentially over an exemplary region in Figure 3b. Recall,

[1] encodeproject.org/pipelines/ENCPL344QWT
[2] encodeproject.org/pipelines/ENCPL367MAS



(a)

Method	Input Dataset (ATAC)	Mean ± SE (Active)	Mean ± SE (Null)	$\log_2$-ratio ± SE	Lower Quartile (HHCDB)	Median (HHCDB)	Upper Quartile (HHCDB)
`Consenrich`	Het10	24.52 ± 0.04	0.338 ± 0.004	6.18 ± 0.016	0.00	0.00	0.039
	Het10+Noisy10	20.96 ± 0.03	0.301 ± 0.003	6.12 ± 0.015	0.00	0.002	0.060
	Het10+Noisy20	18.60 ± 0.03	0.286 ± 0.003	6.02 ± 0.014	0.00	0.007	0.088
`LambdaFE`	Het10	19.81 ± 0.03	0.959 ± 0.003	4.37 ± 0.005	0.46	0.62	0.90
	Het10+Noisy10	18.17 ± 0.04	0.965 ± 0.004	4.24 ± 0.007	0.52	0.68	0.95
	Het10+Noisy20	16.88 ± 0.02	0.966 ± 0.003	4.13 ± 0.004	0.56	0.72	0.98
`MergedPileupRPGC`	Het10	25.70 ± 0.04	1.167 ± 0.004	4.46 ± 0.005	0.55	0.76	1.10
	Het10+Noisy10	23.59 ± 0.03	1.196 ± 0.003	4.30 ± 0.004	0.66	0.86	1.20
	Het10+Noisy20	21.98 ± 0.03	1.218 ± 0.003	4.17 ± 0.004	0.73	0.94	1.27

(b)

Figure 2: *Robust signal extraction from noisy, multi-sample functional genomics data.* (**a**) Density graphs for each method's output given the *Het10* and *Het10+Noisy20* lymphoblastoid-based ATAC-seq input datasets are shown using the same visualization settings. For reference, ENCODE cCREs and HHCDB heterochromatin annotations specific to the GM12878 lymphoblastoid cell line are included. We note distinct qualitative improvements in signal-to-noise and spatial resolution in the Consenrich density graphs compared to benchmarks LambdaFE and MergedPileupRPGC. In particular, Consenrich-estimated signals are non-zero—almost exclusively—at genomic regions overlapping the GM12878 cCREs. (**b**) In the left table panel, to empirically evaluate genome-wide enrichment at DNase1 hypersensitive loci, we compute average log-ratios of values over *Active* regions and 1000 *Null* bootstrap replicates generated using `nullranges::bootRanges`. Consistent with the qualitative behavior observed in the UCSC browser depiction, Consenrich achieved strongest signal intensity versus baseline, with log-ratios > 6 for all instances, even in the presence of a 2:1 majority of Noisy samples (Het10+Noisy20). In the right panel, we further evaluate each method's signal estimates over 4,874 HHCDB-annotated heterochromatin regions. Given the inverse relationship between heterochromatin and open chromatin, we would expect lower ATAC-seq signals at these regions, and the methods' signal estimates should reflect this. The lower quartile, median, and upper quartile of signal estimates for each method/dataset are listed. Notably, Consenrich exhibits near-zero median signal estimates at the annotated heterochromatin regions for all three input datasets.

the gain scales each sample's deviation with the initial state estimate—or *residual* as defined in Equation 9—to determine the additive updates yielding the *a posteriori* state estimate. Thus, samples with large gains contribute more to the posterior estimate than samples with small gains.

In the aggregate case, we observed sample gains decreased almost monotonically over Best5 → Worst5 → Noisy20 samples (Figure 3a). This consistency at the genome-wide level suggests `Consenrich`'s inferred uncertainties align with conventional interpretations of data quality in functional genomics. The interval-specific departures observed in Figure 3b (marked ×), however, underscore an important capability of the dynamic uncertainty model—to upweight data from generally lower-quality samples that offer informative signal locally. Data from generally high-quality samples can be downweighted in the same manner in specific intervals. These fluctuations illustrate a key aspect of `Consenrich`'s locally-oriented uncertainty model that is not limited in perspective to a single genome-wide policy and is motivated by well-known regional and sample-specific sources of noise [9, 14].

Effective multi-sample estimation of TF binding affinities with Consenrich

ChIP-seq assays are used to determine locations of transcription factor (TF) binding and histone modifications. Many experiments quantify aligned reads in both a "treatment sample" that captures presence of the molecular characteristic of interest and a non-specific "control input" accounting for baseline read coverage that may be non-random due to experimental artifacts. Ideally, after removing the effects captured by the control input, the corrected treatment signal reflects true molecular activity. Several mature approaches for analyzing these data in individual samples exist [15, 16] but principled methods for joint extraction of signals in data from multiple treatment-control pairs are not as well-established.

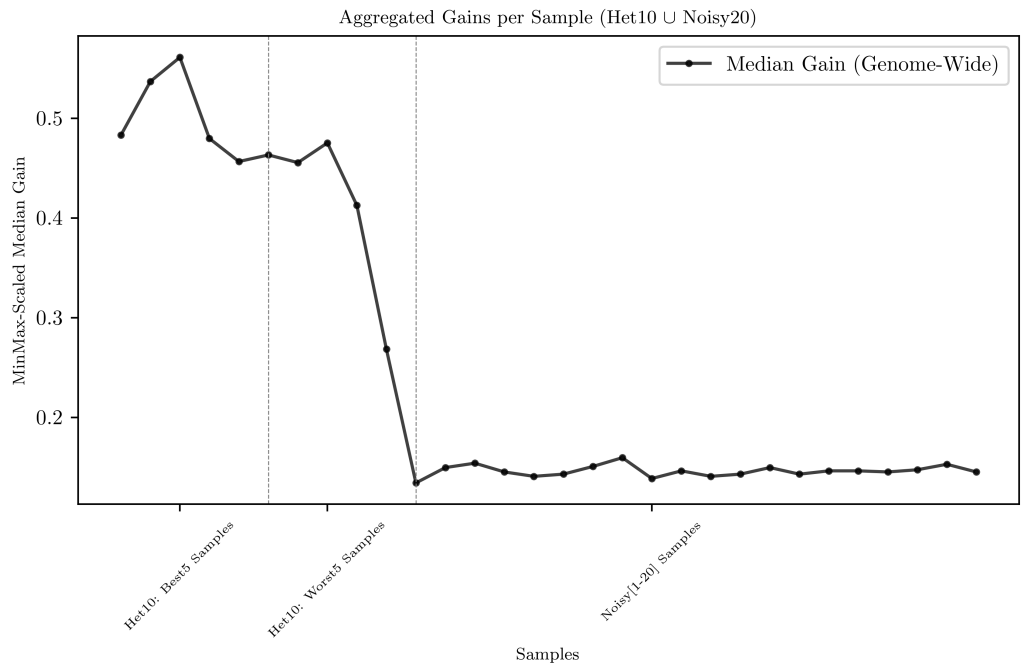
POLR2A binding is known to be highly enriched in promoter regions of actively transcribed genes. We therefore evaluated signal estimates of POLR2A ChIP-seq data measured in colon at a set of ENCODE-annotated promoter-like signatures (PLS) that overlapped genes known to be expressed in colon cell types [17]. As a benchmark, we computed the `AvgLambdaFE` signal by directly averaging the available fold-change tracks from ENCODE using `deeptools bigwigAverage`. Average signal estimates over 1000 `nullranges::bootRanges` iterations were used to represent the null distribution (See Datasets and Evaluation). Consenrich returned an average signal estimate (± SE) of 7.25 ± 0.28 in the PLS regions and 0.23 ± 0.09 over the null sets. `AvgLambdaFE` output values averaged 2.65 ± 0.06 over the promoter regions versus 0.771 ± 0.008 at null regions. These empirical results are consistent with the qualitative performance observed in Figure 4, where Consenrich precisely annotates transcription start sites of several ZNF genes in a 100kb exemplar region.

Enhanced Consensus Peak Calling and Differential Analyses in Complex Human Disease

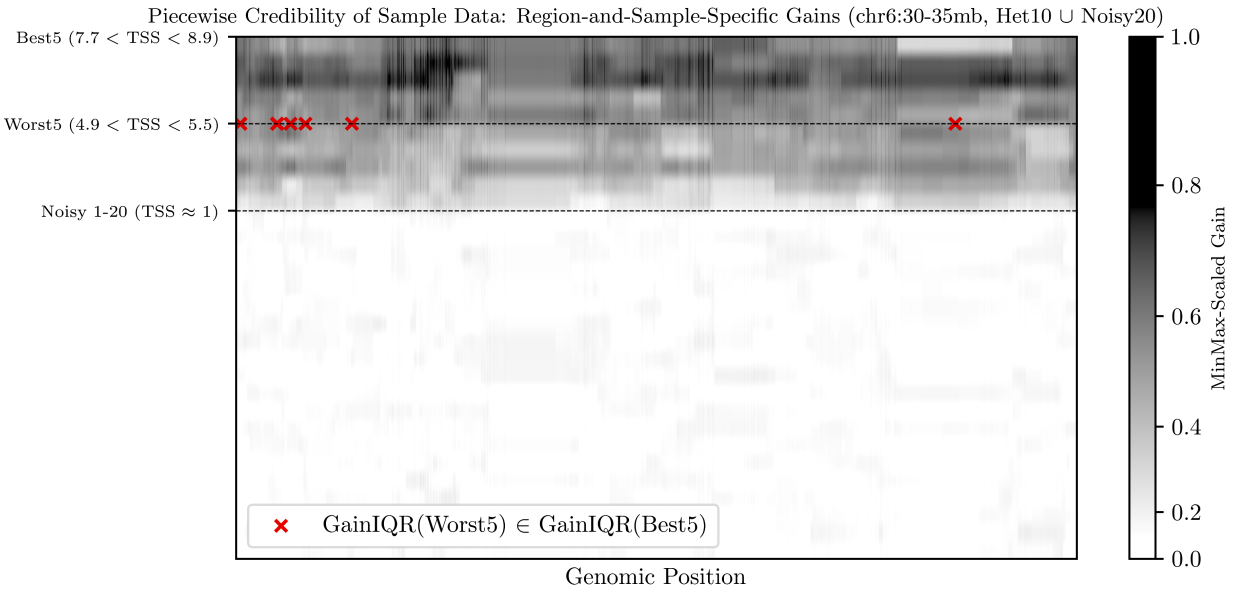
We wanted to determine whether improved multi-sample signal estimates would have positive effects on identifying discrete candidate loci where condition-specific sample groups may differ with respect to a molecular phenotype. Specifically, we analyzed the chromatin accessibility profiles (DNase-seq) derived from dorsolateral prefrontal cortex (DLPFC) tissues of $N = 20$ female donors above 75 years of age. Of these, $N_{\text{AD}} = 5$ had been diagnosed with Alzheimer's Disease (AD) while the remaining $N_{\text{nonAD}} = 15$ patients did not have AD. Using these data, we sought to identify peak regions of enriched signal and determine which of these are differentially accessible regions (DARs). Note, the class imbalance between 'AD' and 'nonAD' is intentional for stress-testing each methods' ability to retrieve disease-specific signal within noisy, heterogeneous datasets.

For experiments aiming to establish consistent, reproducible chromatin accessibility signatures that are indicative of complex disease, *consensus peaks* that are evidenced in numerous samples are preferable as candidate regions to prevent one-off, sample-specific trends from distorting the broader, trait-group-level results. Given multiple samples' sequencing data, reliable detection of consensus peak regions is nontrivial [5, 6, 18]. As a robust estimator of signals in noisy large-scale HTS datasets, we first sought to evaluate Consenrich's utility in this setting.

Identification of candidate, consensus peaks was performed using four different approaches. All methods were applied without leveraging information related to the class distribution (AD vs nonAD) of samples. This is a conservative approach employed to mitigate false positives that might arise due to prematurely leveraging knowledge of samples' respective conditions or "data snooping". Alternatively, candidate consensus peak regions could be



(a) *Averaged gains (Genome-wide).* The gains for each sample were recorded at $N = 1{,}000{,}000$ randomly selected genomic intervals. Their observed medians are plotted such that their order on the x-axis is decreasing with respect to their TSS-enrichment scores—a genome-wide SNR-like measure commonly used to evaluate ATAC-seq data quality.



(b) *Region-dependent influence of sample data.* The gains for each sample in Het10+Noisy20 were recorded at each genomic interval in an exemplar 5 Mb genomic region and min–max-scaled to $[0, 1]$ for visibility. Darker cells indicate greater gains. Rows correspond to samples, columns to genomic intervals. The gains broadly track the samples' TSS-enrichment, yet several regions (marked ×) show overlap between Best5 and Worst5 despite large TSS-enrichment differences.

Figure 3: Consenrich accounts for both region- and sample-specific noise processes. Uncertainty is quantified at each genomic interval and each input sample included in a *weighted average* of global uncertainties to determine data reliability genome-wide.

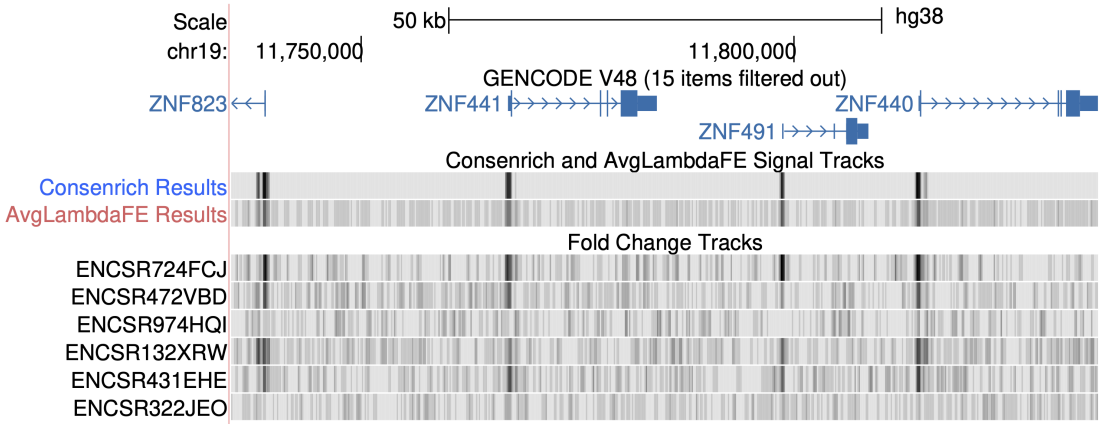


Figure 4: *Promoter-specificity in Consenrich-estimated POLR2A binding signals.* We evaluated the performance of Consenrich for genome-wide multi-sample TF-binding estimation for RNA polymerase 2 (POLR2A) in ChIP-seq experiments performed on colon tissues from six distinct donors. Signal intensity is represented by shades of gray, where higher signal corresponds to darker coloring. In this ZNF-dense 100kb exemplar, we observe accurate localization in Consenrich-estimated binding signals at genes' transcription start sites.

defined separately in each disease group and then combined to form a union set, but previous studies have shown this may compromise control of the false discovery rate (FDR) at nominal levels [19]. In experimental settings where an abundance of condition-exclusive signals may exist—rather than differential intensity in signals that are generally enriched above-baseline in each condition—this homogeneous policy may underperform due to dilution across conditions.

`Consenrich+ROCCO` first applies Consenrich to generate a single track of positional signal estimates from the multiple DNase-seq sample alignments. ROCCO [18] is then used to identify discrete peaks in the Consenrich signal track as reliable consensus-enriched candidate regions. To evaluate the increased performance obtained by using the Consenrich-estimated signal representation of sample data, we included `ROCCO` applied to the original sample data and using its default enrichment measures without Consenrich as a benchmark. The third method, `Genrich` is a popular option for multi-sample studies and has been successfully applied to data from several experiments including ChIP-seq, DNase-seq and ATAC-seq [20, 21, 22, 23, 24, 25]. Genrich first identifies enriched regions in individual samples before applying Fisher's Method [26, 27] to combine p-values for detection of consensus-enriched candidate regions. The fourth method, `ENCODE+UnionThreshold`, represents *decision-level fusion* of multiple results. Briefly, methods invoking this paradigm count overlaps across each sample's independent peak results [1, 5, 6, 7]. That is, given a genomic region $\mathcal{G}$ with flanks extended by some tolerance $\pm W$ bp, if $T \geq 1$ samples' peak sets overlap $\mathcal{G}$, then the region is retained as a consensus peak. In our analysis, the independent sample-level peak results were obtained directly from ENCODE. We considered two peak sets generated with this approach using criteria $T = 2, W = \pm 50$bp and $T = 5, W = \pm 50$bp. Note that this approach is practically limited to $T \leq N_{\text{AD}}$, otherwise an AD-specific region supported by 100% of AD samples could be definitively ignored.

Next, a rigorous differential accessibility testing protocol was applied to determine differentially accessible chromatin regions (DARs) between the AD and non-AD patients (Supplementary Material: C.2). Both explicit covariates (e.g., donor age) and latent, orthogonal sources of variation were included in the final DESeq2 design formula. In Table 1 we list the number of consensus peak regions returned by each method (Consensus Peaks) and the number of these regions found to be differential with respect to Alzheimer's disease status (Significant DARs).

The `Consenrich+ROCCO` consensus peaks result in nearly twice the number of differential regions obtained by any of the listed alternatives. Despite downstream corrections for multiple testing that should, in concept, ensure significance of DARs regardless of the number of candidate consensus peaks, we emphasize that an increased number of legitimate differential results does not necessarily confirm enhanced detection of informative, disease-relevant results. Thus, we assessed whether the greater number of DARs obtained using `Consenrich+ROCCO` reinforced AD-related insights. First, a gene-region enrichment analysis using GO-defined gene sets [28] was performed using all DARs for each method. For `Consenrich+ROCCO`, we find that the top 25 most enriched gene sets are relevant to AD, including ones related to axonogenesis, neurogenesis, and synapse organization and activity (Figure 5). Similar analyses using

Method	Consensus Peaks	Significant DARs	DAR Size $\hat{\mu}$ (bp)	DAR Size $\hat{\sigma}$ (bp)
`Consenrich+ROCCO`	470,601	11418	428.55	278.58
`ROCCO`	232,087	6043	561.96	388.48
`Genrich`	749,350	1508	821.27	689.42
`ENCODE+UnionThreshold(T5)`	83,396	2155	191.84	88.29
`ENCODE+UnionThreshold(T2)`	278,428	4533	102.25	59.70

Table 1: For every method in the leftmost column, the total number of consensus peaks and the number of differentially accessible regions (DARs; AD vs nonAD) are listed, with the average sizes and standard deviations of the size of DARs.

DARs from the other methods identified similar GO terms, but with reduced gene-set coverage and diminished significance of the enriched GO terms. For example, axonogenesis was included in the top 3 results for all methods. In `Consenrich+ROCCO` DARs, however, this term was supported by 251 unique genes, receiving an adjusted p-value $p_{\text{adj}} < 5 \times 10^{-30}$. The second greatest enrichment for this term was attained by `ENCODE+UnionThreshold(T2)`, with 138 unique supporting genes and an adjusted p-value $p_{\text{adj}} \approx 5 \times 10^{-18}$. Similar patterns followed for terms related to neurogenesis, gliogenesis, and synaptic structure. Broadly, using an adjusted p-value cut-off $p_{\text{adj}} < 0.01$, `Consenrich+ROCCO` yielded 303 total significant GO: Biological Process terms. Alternative methods yielded < 200 significant terms at this level of significance.

To further assess whether the increased number of DARs obtained using `Consenrich+ROCCO` remained disease-relevant, we conducted a separate GO enrichment analysis on the 4,729 DARs exclusive to `Consenrich+ROCCO` that do not share any overlap with results from any alternative method (Figure 5). We find that these Consenrich-specific DARs maintained significant enrichments for many of the same disease-related GO terms, supporting their relevance. Together, these results provide suggest that the `Consenrich`-based protocol identifies more differential regions that are strong putative regulatory elements with altered activity in AD. Complete data for each method generated in analysis are included as supplementary material.

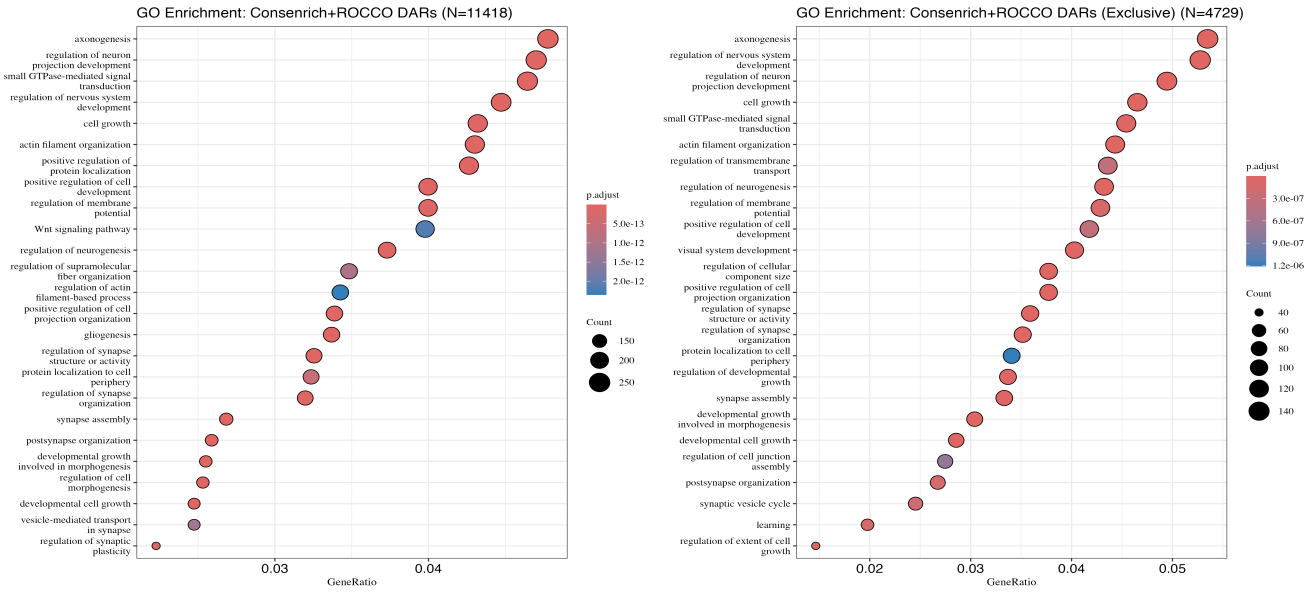


Figure 5: *GO-based enrichment analysis of significant DARs.* `ChIPseeker::enrichGO` was run with an FDR-adjusted p-value threshold (`p.adjust < 0.01`) and semantic similarity cutoff (`simplify(cutoff=0.70)`) [29, 30] to determine significantly enriched, nonredundant GO: Biological Process terms related to the differentially accessible regions. Two sets of DARs were used as input in separate analyses. The first consisted of *all* significant `Consenrich+ROCCO` DARs ($N = 11{,}418$, left plot). The second consisted of DARs *exclusive* to `Consenrich+ROCCO` ($N = 4729$, right plot), without any intersection in the alternative methods'. Disease specificity is preserved in the DARs unique to `Consenrich+ROCCO`, with AD-relevant GO terms comprising the most-enriched results in each case.

Improved capture of risk loci and cis-eQTL in class-imbalanced cohorts To more directly support relevance of altered regulatory regions in Alzheimer's disease, we further compared DARs to expression quantitative loci (eQTL) and risk loci associated with AD. Prior work suggests an enrichment of eQTLs in differentially accessible regions (DARs), and eQTLs enriched in DARs can serve to identify functional regulatory variants that influence gene expression [31, 32]. We evaluated `Consenrich+ROCCO` DARs for enrichment with lead cis-eQTLs identified in the ROSMAP Alzheimer's disease cohort that consisted of $N = 560$ dorsolateral prefrontal cortex tissue samples and identified 16,962 lead cis-eQTL [33]. Each method's DARs were tested for significant enrichment for lead cis-eQTL.

DAR Set	Observed eQTL Overlaps	Mean Overlaps in Null Regions (Est.)	Fold Enrichment	SE	p-value
`Consenrich+ROCCO`	149	39.0	3.82	0.17	1.05E−64
`ROCCO`	96	26.8	3.58	0.20	3.79E−38
`Genrich`	29	9.3	3.11	0.34	7.28E−10
`ENCODE+UnionThreshold` (T2)	10	3.6	2.76	0.54	0.0011
`ENCODE+UnionThreshold` (T5)	7	3.1	2.28	0.59	0.0295

Table 2: *Enrichment of lead cis-eQTLs in DARs.* Differentially accessible chromatin regions identified by each approach were evaluated for enrichment with respect to lead cis-eQTL identified by ROSMAP ($N = 560$ dorsolateral prefrontal cortex tissue samples). To test significance, a null distribution of 10,000 bootstrap replicates was constructed using `nullranges::bootRanges`, and standard errors were estimated with the delta method [34]. A z-test was then performed to compute the p-values in the last column. `Consenrich+ROCCO` DARs showed the strongest fold enrichment and smallest p-value. `ROCCO` DARs likewise showed compelling evidence of eQTL enrichment but a smaller fold-change in overlaps and a greater standard error.

The Consenrich-based DARs revealed the highest fold enrichment with respect to lead cis-eQTL overlaps (Table 2). The 149 corresponding lead cis-eQTL observed in Consenrich-based DARs are given in Supplementary Material S4.

We also considered the number of previously identified AD-associated risk genes/loci overlapping differentially accessible regions. The Alzheimer's Disease Sequencing Project (ADSP) lists a total of $N = 76$ Tier I risk loci[3] supported by "genome-wide evidence" of affecting Alzheimer's risk. In addition to expert review by ADSP's Gene Verification Committee, these loci were determined according to methodology and standards discussed in [35]. To evaluate the breadth of AD-risk loci encompassed by each method's DARs, we counted the number of *unique* risk loci overlapped by at least one DAR. By this measure, `Consenrich+ROCCO` DARs overlap 31 risk loci, `ROCCO` DARs overlap 20 risk loci, `Genrich` DARs overlap 4 risk loci, and `ENCODE+UnionThreshold(T2,T5)` overlap 13 and 5 risk loci, respectively. Together, these analyses further support the increased ability of Consenrch-estimated signals to identify disease relevant regulatory elements.

Conclusion

We have introduced Consenrich, a coherent state estimator for multi-sample functional genomics data based on a simple but principled adaptive linear filter. As an algorithm, Consenrich is explainable, efficient, and effective. By concentrating on integrating sample information at the signal level rather than at the determination of consensus discrete regions of enrichment, Consenrich represents a novel approach to the challenge of multi-sample studies that results in a genome-wide consensus annotation rather than at just the peak level.

Thorough analyses were conducted across multiple assays to demonstrate that Consenrich improved signal extraction capabilities in noisy functional genomics datasets, and that this improved annotation directly led to an increased ability to annotate and analyze data in the more traditional discrete consensus peak setting. In particular, downstream utility was presented in the context of peak calling and differential accessibility testing in a class-imbalanced cohort of Alzheimer's patients and controls. Our analyses clearly show that Consenrich-based consensus peak regions identified more disease relevant regulatory regions compared to other methods. These empirical results clarify the value of sound multi-sample signal estimation as more than a technical consideration—it translates immediately into richer biological discovery.

[3] adsp.niagads.org/gvc-top-hits-list/

We imagine two areas of future work that can take advantage of this new paradigm. First, while here we have applied Consenrich to bulk tissue data, we believe a similar strategy can be used to improve annotation of regulatory regions in single-cell data. The current sparsity of these data at the single-cell level provides an additional challenge, but the underlying characteristics of these data are not changed. Second, we note that the state-space formulation embraced by Consenrich opens a wide range of possibilities for extension. One particularly promising direction involves estimating linear combinations of signals corresponding to complementary assay targets. This can be brought to fruition through practicable augmentations of the process and/or observation models. For instance, to target potential pioneering activity in a given transcription factor, estimating the state and covariance of the difference between signals from multiple assays, such as ChIP-seq for a pioneer factor and ATAC-seq for chromatin state, may highlight distinct regions where factor binding occurs in inaccessible chromatin regions—reflecting nascent pioneer activity. Similar strategies may be developed to incorporate Hi-C contact probabilities with other chromatin data to elucidate three-dimensional regulatory landscapes. As the number of complementary assays and data increases, the strategy implemented by Consenrich provides flexible data assimilation framework for their principled integration at the genome-wide signal level, potentially uncovering novel functional activity.

Methods

Table 3: Common Notation Reference

Notation	Domain	Description
α_D	$\mathbb{R}$	Threshold for adjustment of the process noise covariance $\mathbf{Q}_{[i]}$.
Δ	$\mathbb{R}$	Scales the trajectory to obtain the *a priori* estimate: $x_{[i-1\|i-1]} + \Delta\dot{x}_{[i-1\|i-1]} \rightarrow x_{[i\|i-1]}$ and covariance.
L	$\mathbb{N}$	Length in base pairs of each genomic interval
m	$\mathbb{N}$	Number of samples/sequence alignments used for estimation
n	$\mathbb{N}$	Number of genomic intervals at which the signal is estimated
$\mathbf{F}$	$\mathbb{R}^{2\times 2}$	State transition matrix of the process model
$\mathbf{H}$	$\mathbb{R}^{m\times 2}$	Observation model matrix: maps from the state dimension (i.e., $\mathbb{R}^2$) to observation space (i.e., $\mathbb{R}^m$)
$\mathbf{Q}_{[i]}$	$\mathbb{R}^{2\times 2}$	Process noise covariance matrix at genomic interval i
$\mathbf{R}_{[i]}$	$\mathbb{R}^{m\times m}$	Observation noise covariance at interval i
$\mathbf{K}_{[i]}$	$\mathbb{R}^{2\times m}$	The forward gain matrix at genomic interval i
$\mathbf{A}_{[i]}$	$\mathbb{R}^{2\times 2}$	The backward "smoother gain" matrix at genomic interval i
$\mathbf{E}_{[i]}$	$\mathbb{R}^{m\times m}$	Residual covariance matrix at genomic interval i
$\mathbf{P}_{[i]}$	$\mathbb{R}^{2\times 2}$	True state covariance at genomic interval i
$\mathbf{P}_{[i\|i-1]}$	$\mathbb{R}^{2\times 2}$	*a priori* state covariance at genomic interval i
$\mathbf{P}_{[i\|i]}$	$\mathbb{R}^{2\times 2}$	*a posteriori* state covariance at genomic interval i
$\widetilde{\mathbf{P}}_{[i]}$	$\mathbb{R}^{2\times 2}$	Final state covariance estimate at genomic interval i after the backward pass
$x_{[i]}$	$\mathbb{R}$	True state of the target signal at genomic interval i
$\dot{x}_{[i]}$	$\mathbb{R}$	True rate-of-change in $x_{[i]}$ at genomic interval i
$\mathbf{x}_{[i]}$	$\mathbb{R}^2$	True state vector
$\mathbf{x}_{[i\|i-1]}$	$\mathbb{R}^2$	*a priori* state vector estimate at genomic interval i
$\mathbf{x}_{[i\|i]}$	$\mathbb{R}^2$	*a posteriori* state estimate at genomic interval i
$\widetilde{\mathbf{x}}_{[i]}$	$\mathbb{R}^2$	Final state vector estimate at genomic interval i after the backward pass
$\mathbf{y}_{[i]}$	$\mathbb{R}^m$	Residual or "innovation" vector at genomic interval i

Our aim is to integrate multiple samples' functional genomics sequencing data to track a genome-wide signal—or some proxy for this signal—as it varies positionally across successive *genomic intervals*. For instance, Tn5 transposase insertion or DNaseI hypersensitivity indicated by 'pileups" in sequence alignment files can be treated as genome-wide, quantifiable proxies for chromatin accessibility [5, 14]. Consenrich is an adaptive linear filter with domain-specific features designed to account for the unique challenges posed by high throughput functional genomics data. Consenrich falls within an adaptive Kalman Filtering (KF) framework [36, 37, 38].

We begin by defining an index set for positional estimation over sequences of genomic intervals. See Table 3

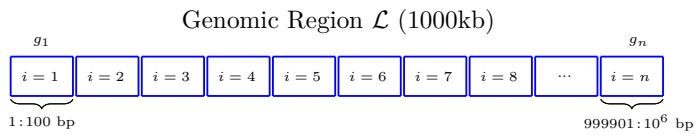


Figure 6: *The genomic region $\mathcal{L}$ is partitioned into n fixed-width intervals. In this* 1000*kb segment, each genomic interval g_i is indexed by $i = 1, 2, \ldots, 10{,}000$ and has length $L = 100$ base pairs.*

for a summary of notation used throughout this manuscript. Note, several additional method and implementation details, including data preprocessing, initialization, and integration of control samples are diverted to Supplementary Material B.

Let $\mathcal{L}$ denote an arbitrary genomic region (e.g., a chromosome) over which the signal of interest varies. To define positional signal estimates, we first partition $\mathcal{L}$ into a sequence of smaller *genomic intervals*. Each genomic interval is fixed in width and consists of L base pairs (Default $L = 25$bp). These intervals are indexed by $i = 1, 2, \ldots, n$ so that:

$$i \mapsto \textsf{Genomic Interval } g_i : \ \{(i-1) \cdot L + 1, \ldots, i \cdot L\} \tag{1}$$

as in Figure 6.

Genome-Wide, Multi-Sample Signal Estimation

The unknown signal at index i is treated as a *real-valued* quantity $x_{[i]} \in \mathbb{R}$. We emphasize the distinction between "data" (sequence alignment counts) and "signal". Though the observed data is indeed discrete in many functional genomics contexts, modeling the signal as a real-valued variable does not substantially compromise an already limited representation of the targeted biological process/signal and confers analytic advantages for capturing trajectories and spatial trends of the signal.

A second state variable, $\dot{x}_{[i]} \in \mathbb{R}$, is included to account for the trajectory in the primary state of interest at each genomic interval i. This second state variable is not directly observed in the data but instead is inferred as a latent variable during iteration across states (Equation 13) Note that the use of Newton-like notation for the second state variable is for convenience and is not intended to imply a literal kinematic interpretation but rather it reflects the basic notion that functional genomic signals at successive genomic positions may be partially inferred from the local context assuming sufficiently small intervals (base pair length L).

The *state vector* is two-dimensional, encompassing both the primary and secondary states

$$\mathbf{x}_{[i]} = \begin{pmatrix} x_{[i]} \\ \dot{x}_{[i]} \end{pmatrix} \tag{2}$$

We assume these quantities cannot be known with exact certainty due to a lack of complete information about the underlying biological process.

We therefore track both the state and its uncertainty, with the latter referred to hereafter as the *state covariance* $\mathbf{P}_{[i]}$:

$$\mathbf{P}_{[i]} = \begin{pmatrix} P_{[i,11]} & P_{[i,12]} \\ P_{[i,21]} & P_{[i,22]} \end{pmatrix} \tag{3}$$

where the diagonal elements correspond to the variance in the primary and secondary states and $P_{[i,12]} = P_{[i,21]}$ convey the relationship between them as covariances. This genome-wide tracking of the covariance permits locally-calibrated uncertainty quantification.

Process and Observation Models

Two interacting models are leveraged for estimation at each interval $i = 1 \ldots n$.

1. The *process model* specifies how the underlying biological process transitions between successive genomic intervals. That is, given $\mathbf{x}_{[i-1]}, \mathbf{P}_{[i-1]}$, the process model yields initial *a priori* state estimates at interval i:

$$\mathbf{x}_{[i-1]}, \mathbf{P}_{[i-1]} \longrightarrow \mathbf{x}_{[i|i-1]}, \mathbf{P}_{[i|i-1]} \tag{4}$$

We employ a simple linear model $\mathbf{F} \in \mathbb{R}^{2\times 2}$ to reflect the basic notion that genomic signals at successive intervals may be partially inferred from the local context assuming sufficiently small L (interval length).

$$\mathbf{F} = \begin{pmatrix} 1 & \Delta \\ 0 & 1 \end{pmatrix}, \tag{5}$$

$$\mathbf{x}_{[i|i-1]} = \mathbf{F}\mathbf{x}_{[i-1]} = \begin{pmatrix} x_{[i-1]} + \Delta \dot{x}_{[i-1]} \\ \dot{x}_{[i-1]} \end{pmatrix}, \tag{6}$$

$$\mathbf{P}_{[i|i-1]} = \mathbf{F}\,\mathbf{P}_{[i-1]}\,\mathbf{F}^{\top} + \mathbf{Q}_{[i]}, \tag{7}$$

where $\Delta > 0$ defines the extent of propagation from the previous state, and the *process noise covariance*, $\mathbf{Q}_{[i]}$, accounts for unmodeled/unknown system aspects—in other words, the uncertainty due to deterministic noise in the state transition model.

2. The *observation model* facilitates comparisons between the *a priori* state estimate and observed data from m samples. This requires a simple mapping of the state vector $\mathbf{x}_{[i|i-1]} \in \mathbb{R}^2$ and covariance $\mathbf{P}_{[i|i-1]} \in \mathbb{R}^{2\times 2}$ to $\mathbb{R}^m$ and $\mathbb{R}^{m\times m}$, respectively.

More concretely, recall that we observe measurements only for the first state in vector $\mathbf{x}_{[i|i-1]}$, that is, $x_{[i|i-1]}$. Because we can only compare the m sample observations to this scalar value, the observation matrix is defined as follows:

$$\mathbf{H} \in \mathbb{R}^{m\times 2}, \mathbf{H} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \end{pmatrix} \tag{8}$$

Then,

$$\mathbf{H}\mathbf{x}_{[i|i-1]} = \mathbf{1}_m \cdot x_{[i|i-1]} \in \mathbb{R}^m.$$

and

$$\mathbf{H}\mathbf{P}_{[i|i-1]}\mathbf{H}^{\top} = \mathbf{1}_{m\times m} P_{[i|i-1,11]} \in \mathbb{R}^{m\times m}$$

We define the vector difference between the m-dimensional representation of the *a priori* estimate $\mathbf{1}_m x_{[i|i-1]}$ and the observed data $\mathbf{z}_{[i]} \in \mathbb{R}^m$ as the *residuals* and encoded in the m-dimensional vector $\mathbf{y}_{[i]}$, defined as:

$$\mathbf{y}_{[i]} = \mathbf{z}_{[i]} - \mathbf{H}\mathbf{x}_{[i|i-1]} \tag{9}$$

$$= \mathbf{z}_{[i]} - \mathbf{1}_m x_{[i|i-1]}. \tag{10}$$

As will be seen, these residuals *update* the *a priori* state estimate and covariance; However, the extent to which residuals can influence the *a posteriori* estimates is modulated by the uncertainty due to each sample's observation. This uncertainty is encoded in the *observation noise covariance*, $\mathbf{R}_{[i]} \in \mathbb{R}^{m\times m}, \mathbf{R}_{[i]} \succ 0$. This covariance is approximated efficiently at each interval $i = 1, 2, \ldots n$ and can therefore account for both sample-specific and region-specific measurement noise inherent in functional genomics data [9]. By augmenting the mapped state covariance with the observation noise covariance, we obtain the *total system uncertainty*, $\mathbf{E}_{[i]} \in \mathbb{R}^{m\times m}$,

$$\mathbf{E}_i = \mathbf{H}\mathbf{P}_{[i|i-1]}\mathbf{H}^{\top} + \mathbf{R}_{[i]}. \tag{11}$$

Note, where the process model is accurate, this expression arises as the covariance of residuals, $\mathbb{E}\left[\mathbf{y}_{[i]}\mathbf{y}_{[i]}^{\top}\right]$. This is fitting given that the residuals are influenced by both the process and observation models' respective uncertainties.

Filter Gain

To integrate the process and observation models, the *gain* $\mathbf{K}_{[i]} \in \mathbb{R}^{2\times m}$ accounts for the respective uncertainty in each sample in relation to the *a priori* estimate. If the uncertainty in the jth sample's data is small (large) relative to the uncertainty in the process model's *a priori* estimate, the gain will be large (small).

More explicitly, the gain can be derived to minimize the mean squared error of the posterior estimate:

$$\arg\min_{\mathbf{K}_{[i]}} \ \mathbb{E}\left[\|\mathbf{x}_{[i]} - \mathbf{x}_{[i|i]}\|^2\right]. \tag{12}$$

Rewriting the mean squared error as the trace of the *a posteriori* state covariance

$$\mathbb{E}\left[\|\mathbf{x}_{[i]} - \mathbf{x}_{[i|i]}\|^2\right] = \text{Trace}\left(\mathbf{P}_{[i|i]}\right),$$

it is straightforward to show that the closed-form

$$\mathbf{K}_{[i]} = \mathbf{P}_{[i|i-1]}\mathbf{H}^\top \underbrace{\left(\mathbf{H}\mathbf{P}_{[i|i-1]}\mathbf{H}^T + \mathbf{R}_{[i]}\right)^{-1}}_{\mathbf{E}_{[i]}^{-1}} \tag{13}$$

is optimal for the minimization in (12). Note, the inversion of $\mathbf{E}_{[i]} \in \mathbb{R}^{m\times m}$ in the expression above is optimized in the software implementation of Consenrich for scaling to large sample sizes m (Supplementary Material B.2).

With the gain computed for interval i, we update *a priori* estimates to obtain *a posteriori* estimates as follows:

$$\mathbf{x}_{[i|i]} = \mathbf{x}_{[i|i-1]} + \mathbf{K}_{[i]}\mathbf{y}_{[i]} \tag{14}$$

$$\mathbf{P}_{[i|i]} = \left(\mathbf{I}_{2\times 2} - \mathbf{K}_{[i]}\mathbf{H}\right)\mathbf{P}_{[i|i-1]}. \tag{15}$$

This two step "predict-correct" cycle is carried out for all intervals $i = 1, 2, \ldots, n$. We refer to this phase of the algorithm as the "forward pass".

Retrospective Estimate Refinement: The Backward Pass

After performing the forward pass, we can employ a second "backward" pass [39] to improve estimates retrospectively given information at later intervals. Starting at index n, we initialize the first "smoothed" estimates in the backward pass from the forward *a posteriori* estimates:

$$\widetilde{\mathbf{x}}_{[n]} = \mathbf{x}_{[n|n]} \tag{16}$$

$$\widetilde{\mathbf{P}}_{[n]} = \mathbf{P}_{[n|n]}. \tag{17}$$

Define

$$\begin{aligned}\mathbf{A}_{[i]} &= \mathbf{P}_{[i|i]}\,\mathbf{F}^\top\,\mathbf{P}_{[i+1|i]}^{-1},\\ &= \mathbf{P}_{[i|i]}\mathbf{F}^\top\left(\mathbf{F}\mathbf{P}_{[i|i]}\mathbf{F}^\top + \mathbf{Q}_{[i+1]}\right)^{-1},\end{aligned}$$

which has a form analogous to the forward gain in Equation 13. Note that the inverted matrix is precisely the *a priori* state covariance from the forward pass at interval $i+1$. Akin to how the forward gain controls the extent to which samples' data can affect the *a posteriori* estimate, this "smoother gain" controls the extent to which information from the subsequent genomic interval $(i+1)$ affects the current interval i.

The backward-recursive smoothed state estimate is defined as

$$\widetilde{\mathbf{x}}_{[i|i]} = \mathbf{x}_{[i|i]} + \mathbf{A}_{[i]}\left(\widetilde{\mathbf{x}}_{[i+1]} - \mathbf{x}_{[i+1|i]}\right), \tag{18}$$

with covariance

$$\widetilde{\mathbf{P}}_{[i]} = \mathbf{P}_{[i|i]} + \mathbf{A}_i\big(\widetilde{\mathbf{P}}_{[i+1]} - \mathbf{P}_{i+1|i}\big)\mathbf{A}_i^\top. \tag{19}$$

Intuitively, as uncertainty due the "next-step" intervals $(i+1)$ increases arbitrarily, the effect of the backward pass approaches zero.

Noise models

Strict theoretical optimality of the Kalman filter (KF) as the minimum mean square error estimator (MMSE) depends on a strict characterization of the system and noise processes: Namely, independent additive white Gaussian noise must be assumed for the process and observations [40, 41]. Practical scenarios violate these assumptions on the noise, especially in the presence of regime shifts, transient dynamics, and data outliers. This does not preclude effective application of domain-tailored, adaptive KF-variants which are the subject of a dedicated literature in systems and control theory [42, 43, 44, 45]. Indeed, robust KF-based approaches routinely match or exceed the filtering performance of resource-intensive, black-box neural architectures [38]. Though theoretical constructions (existence) of transformer-based filters capable of ϵ-approximating classical Kalman filters have been presented for time-invariant systems [46], the premium computational resources, sacrifices in interpretability, and abundance of in-house training data that such an approach would require for cohort-specific signal estimation in HTS data reinforces the appeal of a lightweight linear filter. We discuss the dynamic noise models developed for Consenrich in the following two subsections.

Adaptive Process Noise Covariance

A bounded, adaptive procedure to adjust the process noise covariance $\mathbf{Q}_{[i]}$ is introduced to reduce lag at regime shifts/transients and ensure samples' data is never discarded entirely in favor of the process model. During iteration $i = 1 \dots n$, if we encounter residuals $\mathbf{y}_{[i]}$ that are substantially larger than expected given their modeled covariance,

$$\mathbf{E}_{[i]} = \mathbf{H}\mathbf{P}_{[i|i-1]}\mathbf{H}^\top + \mathbf{R}_{[i]},$$

this suggests a "mismatch" between the process model and the true system. At each interval, we compute the squared residuals divided by their respective uncertainty as modeled:

$$D_{[i,j]} = \frac{y_{[i,j]^2}}{E_{[i,jj]}}.$$

Briefly, if we determine that the values $D_{[i,j]}$ suggest a model mismatch, we can temporarily "inflate" the process model noise covariance by scaling up $\mathbf{Q}_{[i]}$. This has the effect of increasing the gain (Equation 13) such that greater influence is granted to the data when calculating *a posteriori* estimates until the magnitude of residuals better aligns with the expected uncertainty. For brevity, a more thorough treatment of this procedure is moved to Supplementary Material B.1.

Regional, Sample-Specific Observation Noise Covariance

We approximate region-and-sample-specific observation noise levels with an approach that mitigates conflation of signal and noise in the data. This is in contrast to conventional moving-average based approximations of local background, which may fail to discern noise and signal when estimating the former over broadly sparse or enriched regions and regime shifts. For a given sample j and interval i, the local noise level, $R^{\mathsf{local}}_{[i,jj]}$, is approximated by averaging local variances at the k-nearest "sparse regions" (Default $k = 25$). Generic annotations of sparse regions, defined within large "gaps" between previously-established regulatory regions are supplied in the software implementation of Consenrich. Alternatively, users analyzing data in less-studied genomes can apply an annotation-free approach based on the average of local variances (ALV) [47].

Local versus Global Weighting After computing local noise levels in each sample, a weighted average of $R^{\mathsf{local}}_{[i,jj]}$ and an analogous genome-wide noise statistic, $R^{\mathsf{global}}_{[:,jj]}$, are used to determine the observation noise covariance $\mathbf{R}_{[i]}$ (Supplementary Material B.2).

Datasets and Evaluation

Benchmark methods for comparison are included where possible, but we note the lack of well-established like-for-like comparators that similarly emphasize scalable signal and uncertainty quantification from multi-sample HTS data. Experiments evaluating robustness and biological insight with reference to real data from well-profiled biosamples. Additionally, we incorporate null comparisons to evaluate results in an absolute sense.

ATAC-seq From a set of fifty-six total previously published lymphoblastoid ATAC-seq samples, we collated the best five ("Best5") and worst five ("Worst5") samples, ranked by their respective transcription start site (TSS) enrichment scores[4], into a single dataset, referred to hereafter as "Het10". In experiments evaluating robustness to noisy data, intentionally low-quality synthetic samples were constructed to reflect realistic noise processes observed in genome-wide DNA enrichment assays. Rather than perturbing count data with draws from an arbitrary probability distribution, we introduce noise at the sequence alignment level: Unrelated, low-quality alignments were merged with ChIP-seq control alignments and subsampled to generate $m = 20$ total "Noisy" samples with both low signal-to-noise ratios (SNR) and piecewise artifacts capable of distorting true signals (Supplementary Material A.1). These noisy samples were gradually appended to the Het10 dataset to form the Het10+Noisy10 and Het10+Noisy20 datasets.

To establish a frame of reference for lymphoblastoid-based results, we use the rich set of available annotations in the GM12878 cell line generated by ENCODE. The Het10 ATAC-seq dataset used as input to Consenrich for various analyses in this manuscript is not exclusive to the GM12878 lymphoblastoid cell line. Nonetheless, GM12878 is one of the most deeply-profiled [2], with an abundance of publicly available annotations that are valuable as proxies for ground truth in relative performance comparisons involving lymphoblastoid data.

ChIP-seq Consenrich's ability for genome-wide multi-sample TF-binding estimation were evaluated using ENCODE ChIP-seq data derived from colon tissues in six distinct donors. Both treatment and control alignment files were utilized (Supplementary Material A.2).

DNase-seq In experiments evaluating downstream benefits for genome-wide differential analyses, we used ENCODE DNase-seq alignments from dorsolateral prefrontal cortex (DLPFC) tissue. Twenty samples were extracted from female donors above 75 years of age with-and-without Alzheimer's disease (AD): $m_{\text{AD}} = 5$, $m_{\text{nonAD}} = 15$ (Supplementary Material A.3). Differential analysis was performed using a multi-step workflow involving RUVseq [48] and DESeq2 [49] (Supplementary Material C).

Evaluating Signal Enrichment For a set of genomic regions $\mathcal{G}$, where some molecular-level phenotype is assumed, let T denote a statistic used to estimate relevant signal from the corresponding quantified sequence alignment data. To evaluate the significance of observations in $\mathcal{G}$, we make frequent use of `nullranges::bootRanges` hosted on BioConductor [11, 50]. We configure `nullranges::bootRanges` to apply a segmented block-bootstrapping technique, where the segmentation is with respect to gene density, that implicitly accounts for structural properties of genomic regions in $\mathcal{G}$. The output consists of B bootstrap-sampled sets of genomic regions $\mathcal{N}_b$, $b = 1, 2, \ldots, B$. By computing the statistic T over these sets, we gain a useful baseline to evaluate a generic null hypothesis—that the observed statistics over $\mathcal{G}$ are no more extreme than we would expect by chance over sets with similar structural relationships (e.g., feature clustering) in genome segments of comparable gene density.

Declarations

Funding This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (grant numbers **R01DK138462** and **R01DK136262**) and by the National Human Genome Research Institute of the National Institutes of Health (grant number **R01HG009937**).

[4] encodeproject.org/data-standards/terms/#enrichment

Data Availability The sequence alignment files used as input to evaluate Consenrich are publicly available from ENCODE [2]. Full details and accession codes can be found in Supplementary Material A.1, A.2, A.3.

Software A complete software implementation of Consenrich is hosted at `github.com/nolan-h-hamilton/Consenrich`.

Competing Interests The authors declare that they have no competing interests.

Author Contributions **NHH**: Conceptualization, Methodology, Formal analysis, Software, Validation, Writing – original draft, Writing – review and editing. **YCH**: Data Curation, Formal analysis, Experiment design, Validation, Writing – original draft, Writing – review and editing. **BDM**: Data Curation, Experiment Design, Validation, Writing – review and editing. **MIL**: Experiment design, Supervision, Writing – review and editing. **TSF**: Experiment design, Funding acquisition, Resources, Supervision, Writing – review and editing.

References

[1] M. Ryan Corces et al. "The chromatin accessibility landscape of primary human cancers". In: *Science* 362.6413 (Oct. 2018).

[2] Yunhai Luo et al. "New developments on the Encyclopedia of DNA Elements (ENCODE) data portal". In: *Nucleic Acids Research* 48.D1 (2019), pp. D882–D889.

[3] Yong Zhang, Tao Liu, Clifford A Meyer, et al. "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (Sept. 2008).

[4] Jeremy W. Schroeder and P. Lydia Freddolino. "Enricherator: A Bayesian Method for Inferring Regularized Genome-wide Enrichments from Sequencing Count Data". In: *Journal of Molecular Biology* (Apr. 2024).

[5] Feng Yan, David R. Powell, David J. Curtis, et al. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis". In: *Genome Biology* 21.1 (Feb. 2020).

[6] Jason P Smith et al. "PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments". In: *NAR Genomics and Bioinformatics* 3.4 (Oct. 2021).

[7] Yajie Yang et al. "Leveraging biological replicates to improve analysis in chip-seq experiments". In: *Computational and Structural Biotechnology Journal* 9.13 (2014).

[8] Anshul Kundaje et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (Feb. 2015), pp. 317–330.

[9] Clifford A. Meyer and X. Shirley Liu. "Identifying and mitigating bias in next-generation sequencing methods for chromatin biology". In: *Nature Reviews Genetics* 15.11 (Sept. 2014).

[10] Fidel Ramírez et al. "deepTools2: a next generation web server for deep-sequencing data analysis". In: *Nucleic Acids Research* 44.W1 (Apr. 2016), W160–W165.

[11] Wancen Mu et al. "bootRanges: flexible generation of null sets of genomic ranges for hypothesis testing". In: *Bioinformatics* 39.5 (Apr. 2023), btad190. ISSN: 1367-4811.

[12] Hongli Wang et al. "HHCDB: a database of human heterochromatin regions". In: *Nucleic Acids Research* 52.D1 (Oct. 2023), pp. D145–D153. ISSN: 1362-4962.

[13] D. Karolchik. "The UCSC Genome Browser Database". In: *Nucleic Acids Research* 31.1 (Jan. 2003), pp. 51–54. ISSN: 1362-4962.

[14] Alan P. Boyle et al. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". In: *Cell* 132.2 (Jan. 2008), pp. 311–322. ISSN: 0092-8674.

[15] John M Gaspar. "Improved peak-calling with MACS2". In: *bioRxiv* (2018).

[16] Nanxiang Zhao and Alan P Boyle. "F-Seq2: improving the feature density based peak caller with dynamic statistics". In: *NAR Genomics and Bioinformatics* 3.1 (Feb. 2021).

[17] Yalong Wang et al. "Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine". In: *Journal of Experimental Medicine* 217.2 (Nov. 2019). ISSN: 1540-9538.

[18] Nolan H Hamilton and Terrence S Furey. "ROCCO: a robust method for detection of open chromatin via convex optimization". In: *Bioinformatics* 39.12 (Dec. 2023).

[19] Aaron T.L. Lun and Gordon K. Smyth. "De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly". In: *Nucleic Acids Research* 42.11 (May 2014).

[20] Jakob Hofvander et al. "Undifferentiated pleomorphic sarcomas with PRDM10 fusions have a distinct gene expression profile". In: *The Journal of Pathology* 249.4 (2019), pp. 425–434.

[21] Mazdak Salavati et al. "Profiling of open chromatin in developing pig (Sus scrofa) muscle to identify regulatory regions". In: *G3 Genes|Genomes|Genetics* 12.2 (2021). Ed. by J Holland.

[22] Lindsey N. Guerin, Kelly R. Barnett, and Emily Hodges. "Dual detection of chromatin accessibility and DNA methylation using ATAC-Me". In: *Nature Protocols* 16.12 (2021), pp. 5377–5397.

[23] Roman Tsaryk et al. "Shear stress switches the association of endothelial enhancers from ETV/ETS to KLF transcription factor binding sites". In: *Scientific Reports* 12.1 (2022).

[24] Susu Guo et al. "VGLL2 and TEAD1 fusion proteins drive YAP/TAZ-independent tumorigenesis by engaging p300". In: *eLife* (Apr. 2025).

[25] Brandon M. Wenz et al. "Genotype inference from aggregated chromatin accessibility data reveals genetic regulatory mechanisms". In: *Genome Biology* 26.1 (Mar. 2025). ISSN: 1474-760X.

[26] Frederick Mosteller. "Questions and Answers". In: *The American Statistician* 2.5 (Oct. 1948), pp. 30–31. ISSN: 1537-2731.

[27] R. A. Fisher. "Statistical Methods for Research Workers". In: *Breakthroughs in Statistics*. Springer New York, 1992, pp. 66–70. ISBN: 9781461243809.

[28] Suzi A Aleksander, Balhoff, et al. "The Gene Ontology knowledgebase in 2023". In: *GENETICS* 224.1 (Mar. 2023). Ed. by A Baryshnikova.

[29] Guangchuang Yu, Li-Gen Wang, and Qing-Yu He. "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization". In: *Bioinformatics* 31.14 (Mar. 2015), pp. 2382–2383.

[30] Qianwen Wang et al. "Exploring Epigenomic Datasets by ChIPseeker". In: *Current Protocols* 2.10 (Oct. 2022).

[31] Jacob F. Degner et al. "DNaseI sensitivity QTLs are a major determinant of human expression variation". In: *Nature* 482.7385 (Feb. 2012), pp. 390–394. ISSN: 1476-4687.

[32] Kaur Alasoo et al. "Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response". In: *Nature Genetics* 50.3 (Jan. 2018), pp. 424–431. ISSN: 1546-1718.

[33] Alejandra P. Pérez-González et al. "The ROSMAP project: aging and neurodegenerative diseases through omic sciences". In: *Frontiers in Neuroinformatics* 18 (Sept. 2024). ISSN: 1662-5196.

[34] Ting Qi et al. "Genetic control of RNA splicing and its distinct role in complex trait variation". In: *Nature Genetics* 54.9 (Aug. 2022), pp. 1355–1363. ISSN: 1546-1718.

[35] Céline Bellenguez et al. "New insights into the genetic etiology of Alzheimer's disease and related dementias". In: *Nature Genetics* 54.4 (Apr. 2022), pp. 412–436. ISSN: 1546-1718.

[36] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (Mar. 1960).

[37] Paul Zarchan and Howard Musoff. *Fundamentals of Kalman Filtering: A Practical Approach, Fourth Edition*. American Institute of Aeronautics and Astronautics, Inc., Jan. 2015.

[38] Ido Greenberg, Netanel Yannay, and Shie Mannor. "Optimization or Architecture: How to Hack Kalman Filtering". In: *Advances in Neural Information Processing Systems (NeurIPS '23)*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 50482–50505.

[39] Dan Simon. *Optimal State Estimation*. Wiley, Jan. 2006. Chap. 9.

[40] Dan Simon. *Optimal State Estimation*. Wiley, Jan. 2006.

[41] Jeffrey Uhlmann and Simon J. Julier. "Gaussianity and the Kalman Filter: A Simple Yet Complicated Relationship". In: *Journal de Ciencia e Ingeniería* 14.1 (June 2022), pp. 21–26.

[42] Z.M. Durovic and B.D. Kovacevic. "Robust estimation with unknown noise statistics". In: *IEEE Transactions on Automatic Control* 44.6 (June 1999), pp. 1292–1296. ISSN: 0018-9286.

[43] Dan Simon. *Optimal State Estimation*. Wiley, Jan. 2006. Chap. 6.

[44] Dan Simon. *Optimal State Estimation*. Wiley, Jan. 2006. Chap. 7.

[45] Lei Zou et al. "A survey on outlier-resistant state estimation and its applications". In: *Systems Science and Control Engineering* 13.1 (Mar. 2025). ISSN: 2164-2583.

[46] Gautam Goel and Peter Bartlett. "Can a transformer represent a Kalman filter?" In: *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*. Vol. 242. Proceedings of Machine Learning Research. PMLR, July 2024, pp. 1502–1512.

[47] P.K. Bocher and K.R. McCloy. "The fundamentals of average local variance - part I: detecting regular patterns". In: *IEEE Transactions on Image Processing* 15.2 (2006), pp. 300–310.

[48] Davide Risso et al. "Normalization of RNA-seq data using factor analysis of control genes or samples". In: *Nature Biotechnology* 32.9 (Aug. 2014), pp. 896–902. ISSN: 1546-1696.

[49] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (Dec. 2014). ISSN: 1474-760X.

[50] Robert C Gentleman et al. "Bioconductor: open software development for computational biology and bioinformatics". In: *Genome Biology* 5.10 (Sept. 2004). ISSN: 1474-760X.

[51] Charles R. Harris, K. Jarrod Millman, et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362.

[52] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (June 2009), pp. 2078–2079. ISSN: 1367-4803.

[53] William K. Pratt. "Digital Image Processing". In: 4th ed. Hoboken, NJ: John Wiley & Sons, 2007. Chap. 10.

[54] Pavel Raykov. "An Optimal Algorithm for Sliding Window Order Statistics". In: *26th International Conference on Database Theory, ICDT 2023, March 28-31, 2023, Ioannina, Greece*. Ed. by Floris Geerts and Brecht Vandevoort. Vol. 255. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023, 5:1–5:13.

[55] Daniel J. Wilson. "The harmonic mean p -value for combining dependent tests". In: *Proceedings of the National Academy of Sciences* 116.4 (Jan. 2019), pp. 1195–1200. ISSN: 1091-6490.

[56] Nozie D. Aghaizu, Hanqing Jin, and Paul J. Whiting. "Dysregulated Wnt Signalling in the Alzheimer's Brain". In: *Brain Sciences* 10.12 (Nov. 2020), p. 902. ISSN: 2076-3425.

[57] Courtney M Kloske et al. "Inflammatory pathways are Impaired in Alzheimer Disease and Differentially Associated With Apolipoprotein E Status". In: *Journal of Neuropathology & Experimental Neurology* 80.10 (Sept. 2021), pp. 922–932. ISSN: 1554-6578.

[58] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. 4th ed. Baltimore, MD: Johns Hopkins University Press, 2013. Chap. 2, pp. 63–104. ISBN: 978-1421407944.

A Supplementary Material: Datasets

A.1 ATAC-seq: Het10

Sample/Data	Datasets	TSS
ENCFF632MBC	Worst5,Het10	4.954
ENCFF919PWF	Worst5,Het10	5.196
ENCFF497QOS	Worst5,Het10	5.203
ENCFF767FGV	Worst5,Het10	5.311
ENCFF326QXM	Worst5,Het10	5.532
ENCFF495DQP	Best5,Het10	7.744
ENCFF949CVL	Best5,Het10	7.766
ENCFF687QML	Best5,Het10	7.814
ENCFF447ZRG	Best5,Het10	7.944
ENCFF462RHM	Best5,Het10	8.919

Table S1: *The Het10 dataset.* The five best-and-worst-scored ATAC-seq alignments (TSS enrichment) were selected from a set of fifty-six total lymphoblastoid experiments available from ENCODE.

"Noisy" samples are introduced to evaluate Consenrich's performance in the presence of low-SNR, distorted input (Section). Low quality ChIP-seq control inputs and ATAC-seq samples from distinct experiments were merged to create a single alignment file `noisy_template.bam`. This alignment file was then randomly subsampled to draw each `noisy[k].bam`.

Note, this procedure was designed to introduce background but also distort true signals in a piecewise fashion that is consistent with real noise processes in genome-wide functional genomics data. A stationary/baseline noise could be filtered with a simple detrending operation and would not provide a challenging or realistic experiment. Similarly, we opted to add noise at the sequence alignment level so that the challenges due to sequencing artifacts are well-represented. In contrast, adding noise at the count/coverage track level involves selecting a distribution for the noise that is both discretionary and ignores practical concerns due to alignment quality, fragment length, indels, etc.

Two ChIP-seq control alignment files (ENCFF415JON, ENCFF140JFH) and two ATAC-seq alignment files (ENCFF589DFL, ENCFF946URK) were obtained from ENCODE. These alignments were merged to generate `noisy_template.bam`, referenced in Algorithm 1.

Algorithm 1 Noisy[K] Dataset Generation

1: **Inputs:** `noisy_template.bam`: Merged ChIP-seq control and low-quality ATAC-seq alignments, `prop`: percent of reads to subsample (10)
2: **for** $k = 1$ **to** 20 **do**
3: Compute `seed` $\leftarrow (k \times 100 + k)$`prop`
4: `samtools view -b -f 2 -F 3840 -s seed noisy_template.bam > noisy[k].bam`
5: **end for**

A.2 ChIP-seq: Colon/POLR2A

ChIP-Seq experiments targeting TF *POLR2A* in six independent donors' colon tissue samples (4 sigmoid, 2 transverse), obtained from ENCODE.

For each ENCODE accession in the above table, both a treatment and control alignment file were used (in total: $m = 12$ alignments).

A.3 DNase-seq: AD vs. nonAD

Twenty dorsolateral prefrontal cortex tissue (DLPFC) samples were collected from ENCODE. Each experiment and corresponding DNase-seq alignment file listed are listed in Table S3 with relevant metadata.

Sample/Data	Tissue	Age	Gender
ENCSR322JEO	sigmoid colon	53	female
ENCSR472VBD	sigmoid colon	54	male
ENCSR431EHE	sigmoid colon	37	male
ENCSR724FCJ	sigmoid colon	51	female
ENCSR974HQI	transverse colon	51	female
ENCSR132XRW	transverse colon	54	male

Table S2: Data used in POLR2A/colon ChIP-seq experiments (Section)

Alignment	Experiment	Status	Age
ENCFF054OWB.bam	/experiments/ENCSR921PPJ/	No_Ad	83
ENCFF066YOO.bam	/experiments/ENCSR617SVO/	No_Ad	≥ 90
ENCFF096RBN.bam	/experiments/ENCSR865CYW/	No_Ad	≥ 90
ENCFF175CMR.bam	/experiments/ENCSR518WRP/	No_Ad	78
ENCFF195KCN.bam	/experiments/ENCSR661AMM/	No_Ad	83
ENCFF211LAD.bam	/experiments/ENCSR751VMX/	No_Ad	≥ 90
ENCFF266UVR.bam	/experiments/ENCSR386XPD/	No_Ad	82
ENCFF306XXE.bam	/experiments/ENCSR800KVI/	No_Ad	84
ENCFF340VDR.bam	/experiments/ENCSR720FVJ/	No_Ad	≥ 90
ENCFF502IOV.bam	/experiments/ENCSR879XUH/	No_Ad	≥ 90
ENCFF505OFL.bam	/experiments/ENCSR173YSO/	No_Ad	79
ENCFF879QFZ.bam	/experiments/ENCSR819GAF/	No_Ad	≥ 90
ENCFF884NRT.bam	/experiments/ENCSR001QTZ/	No_Ad	≥ 90
ENCFF984QXK.bam	/experiments/ENCSR995YPS/	No_Ad	87
ENCFF990EJE.bam	/experiments/ENCSR800MBE/	No_Ad	82
ENCFF375NXF.bam	/experiments/ENCSR426KLJ/	AD	87
ENCFF446JDN.bam	/experiments/ENCSR428ZCP/	AD	≥ 90
ENCFF617VLY.bam	/experiments/ENCSR329FAP/	AD	89
ENCFF647TPO.bam	/experiments/ENCSR570BWW/	AD	89
ENCFF669HSH.bam	/experiments/ENCSR340XQX/	AD	≥ 90

Table S3: DNase-seq alignment files from twenty donors' dorsolateral prefrontal cortex tissues. $m = 5$ 'AD' patients and $m = 15$ 'No_AD' patients. All tissues were extracted from the dorsolateral prefrontal cortex (DLPFC) of female donors above 75 years of age.

B Additional Method Details

B.1 Process Noise Covariance

At each interval $i = 1 \dots n$, we consider whether the modeled uncertainty sufficiently explains the observed residuals $\mathbf{y}_{[i]}$ (Eqn. 9). Define

$$D_{[i,j]} = \frac{y_{[i,j]^2}}{E_{[i,jj]}}. \tag{20}$$

This is the ratio between the square of the jth residual and the respective diagonal term in the residual covariance matrix $\mathbf{E}_{[i]}$,

$$E_{[i,jj]} = \mathsf{diag}\left(\mathbf{H}\mathbf{P}_{[i|i-1]}\mathbf{H}^\top + \mathbf{R}_{[i]}\right)_{[j]} = P_{[i|i-1,11]} + R_{[i,jj]}, \tag{21}$$

Where the system model is accurate, we have

$$\mathbb{E}[y_{[i,j]}] = 0$$

such that Eqn. (20) resembles a standardized z-score.

Thus, at a given interval $i = 1 \dots n$, large values

$$\widetilde{D}_{[i]} = \mathsf{Median}\{D_{[i,j]}\}_{j=1}^{m}$$

suggest a mismatch between the model and data that prompts additional uncertainty. These cases are typical of abrupt shifts in dynamics that prompt rescaling of model parameters to avoid lagging signal estimates.

In such cases, the process noise covariance $\mathbf{Q}_{[i]}$ is multiplied (scaled up) by a function monotonic in $|\widetilde{D}_{[i]} - \alpha_D|$ strictly above 1. Scaling up $\mathbf{Q}_{[i]}$ has the effect of allowing greater relative influence from the observed data (Eqn. 13).

By default, for $\widetilde{D}_{[i]} > \alpha_D$, we scale up

$$\mathbf{Q}_{[i]} \leftarrow \left(\sqrt{d \cdot |\widetilde{D}_{[i]} - \alpha_D| + 1}\right) \mathbf{Q}_{[i]}$$

.

At following iterations $i + k, k \geq 1$, whenever $\widetilde{D}_{[i+k]} \leq \alpha_D$,

$$\mathbf{Q}_{[i+k]} \leftarrow \left(\sqrt{d \cdot |\widetilde{D}_{[i+k]} - \alpha_D| + 1}\right)^{-1} \mathbf{Q}_{[i+k]}$$

Note that during iteration, the diagonal elements (in our case, eigenvalues) of $\mathbf{Q}_{[i]}$ are *bound* in $[Q_{\min}, Q_{\max}]$ to preserve positive definiteness of covariances and to prevent either the process or observation model from becoming dominant and determining estimates exclusively.

B.2 Observation Noise Covariance

Local noise levels $R^{\mathsf{local}}_{[i,jj]}$ are computed from sample data over "sparse" regions devoid of candidate regulatory regions and evidence of signal. Consenrich is packaged with default sparse annotations for convenience, but users can supply alternative references or apply built-in routines to generate an annotation from their own data. Alternatively, a one-dimensional "average of local variances" (ALV) [47] approach is supported and may be ideal for users analyzing under-profiled genomes.

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

$$R_{[i,jj]} = \omega^{\mathsf{local}} R^{\mathsf{local}}_{[i,jj]} + \left(1 - \omega^{\mathsf{local}}\right) R^{\mathsf{global}}_{[:,jj]} \tag{22}$$

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. Note, similar to the bounds on the process noise covariance (Section B.2, B.1), the diagonal elements in $\mathbf{R}_{[i]}$ are restricted to $[R_{min}, R_{max}]$.

Algorithmic Details Computing the sequence of 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.

1. Store and sort the locations of sparse regions in an array, `s`, `length`$(s) = s \ll n$. This is $\mathcal{O}(s \log s)$ and only needs to be performed once per contig.
2. For each interval i, determine its insertion point in `s`: Find `h` such that `s[h] < intervals[i] <= s[h+1]` via bisection. Note, this is $\mathcal{O}(\log s)$ and can be vectorized across intervals using NumPy [51] `np.searchsorted` instead of an explicit loop.

3. Return the array slice `s[h-k:h+k]`

In practice, runtime is significantly reduced with several optimizations not reflected in 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 the noise variance.

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. 13). General matrix inversion is $\mathcal{O}(m^3)$ 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 $\mathbf{E}_{[i]}$:

$$\mathbf{E}_{[i]} = \mathbf{v}\mathbf{v}^\top + \mathbf{R}_{[i]} \tag{23}$$

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:

$$\mathbf{E}_{[i]}^{-1} = \mathbf{R}_{[i]}^{-1} - \frac{P_{[i|i-1,11]} \cdot \mathsf{diag}\left(\mathbf{R}_{[i]}^{-1}\right) \mathsf{diag}\left(\mathbf{R}_{[i]}^{-1}\right)^\top}{1 + P_{[i|i-1,11]} \cdot \mathsf{Trace}(\mathbf{R}_{[i]}^{-1})} \tag{24}$$

This reduces runtime to $\mathcal{O}(m^2)$, 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. (24) 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 Section .

B.3 Calculating Read Coverage

Several preprocessing techniques are applied to initial sequence alignment data to obtain the sample-by-count matrix

$$\mathbf{Z} \in \mathbb{R}^{m \times n}.$$

In a given sample, the initial alignment/count data track is generated by counting the number of sequencing mapping to each interval subject to filtering based on SAM mapping quality scores and flags [52].

Background Removal and Detrending

During preprocessing, we measure local background levels in each sample's read count tracks using a *percentile filter*. The percentile filter is a generalization of the median filter [53, 54] and, in contrast to moving-averages or general least-squares approaches, effectively captures low-frequency baseline content while remaining robust to local enrichments that could inflate background estimates. By default, Consenrich applies a 75th-percentile filter to each sample's data in windows of 10kb. The low-frequency, baseline content captured in this representation is then subtracted from the original data.

B.4 Initialization

As a recursive state estimator, Consenrich relies on the previously inferred states at intervals $i-1, i-2, \ldots$ to infer quantities at interval i. We therefore require initial estimates of the state and covariance. Due to refinement during the secondary "backward pass" and adaptive scaling of the process noise covariance (Section B.1) these initial estimates are of minimal practical consequence. Arbitrarily, we set $\mathbf{x}_{[1|1]} = \begin{bmatrix}\mathsf{Median}(Z[:,1]) & 0\end{bmatrix}^\top$, and $\mathbf{P}_{[1|1]} = 100 \cdot \mathbf{I}_{2\times 2}$.

B.5 Integration of Control Samples

Consenrich accounts for genome-wide control inputs (e.g. ChIP-seq IgG data) by scaling treatment-control pairs to each other. To avoid amplifying noise in samples with lower sequencing depth, we follow MACS' default protocol, scaling values to the smaller of the pair's sequencing depths.

Subsequently, the control input is subtracted from the treatment to obtain a "corrected" treatment data track. In the default implementation, corrected treatment data tracks are then normalized with respect to the sequencing depth of the original treatment sample to facilitate between-experiment comparisons (e.g., counts per million mapped reads (CPM), 1x genome coverage (RPGC), etc.) Alternatively, explicit weights for each sample (including controls) can also be specified at the command line.

C Alzheimer's DARs: Additional Details

C.1 Consensus Peaks

ENCODE+UnionThreshold

`ENCODE+UnionThreshold` represents *decision-level fusion* of multiple sample-specific results. Briefly, methods invoking this paradigm count overlaps across each sample's independent peak results [5, 6, 7]. That is, given a genomic region $\mathcal{G}$ with flanks extended by some tolerance $\pm W$ bp, if $T \geq 1$ samples' peak sets overlap $\mathcal{G}$, then the region is retained as a consensus peak.

In our analysis, the independent sample-level peak results were obtained from the $m = 20$ experiments' data available directly from ENCODE. Overlaps were counted by applying `bedtools multiinter` on $\pm$50bp extended features. We considered two peak sets generated with this approach using criteria $T = 2, W = \pm 50$bp and $T = 5, W = \pm$ 50bp. Note that this approach is practically limited to $T \leq N_{\text{AD}}$, otherwise an AD-specific region supported by 100% of AD samples could be definitively ignored despite its relevance for downstream differential analyses.

Genrich

Genrich is a method for detection of enrichment (peaks) in general HTS/seq data. Genrich provides explicit support for reconciling consensus peaks via Fisher's Method [26, 27] and is a popular option for studies with multiple samples [20, 21, 22, 23, 24, 25]

```
Genrich \
  -t <comma-separated alignments>
  -o genrich_out.narrowPeak -v \
  -k genrich_out.bdg \
  -z \
  -e chrY
  -p <0.01 (suggested default), 0.001, 0.0001, 0.00001* (selected)>
```

The independence assumption underlying Fisher's Method may become impractical as heterogeneity and sample size grow in typical functional genomics experiments, especially those without genome-wide control inputs for each sample[5]. Future work may consider alternative approaches that are more robust to unknown dependencies between p-values [55] and do not treat each as independent supporting evidence.

ROCCO

```
rocco \
  -i <space-separated alignments>
  -g hg38 --skip_chroms chrY --threads 8
```

C.2 Differential Accessibility Testing

A multi-step differential accessibility testing protocol was then employed to determine differentially accessible chromatin regions (DARs) between the AD and nonAD patients. DESeq2 [49] was applied to compute likelihood ratio statistics based on estimates derived with regularized negative binomial regression. In the DESeq2 design formula, we explicitly specified donor age as a covariate and up to $K \leq 3$ latent, orthogonal sources of variation from a reduced-rank representation of count data at negative control regions (`RUVSeq::RUVg`) [48]. The following design formula was supplied to DESeq2:

$$\mathsf{design} \ = \sim \mathbf{U}[1 : m, K \leq 3] + \mathsf{Age} + \mathsf{Status}.$$

[5]Note, this approach based on Fisher's Method is not exclusive to Genrich. For instance MACS' `cmbreps` documentation `macs3-project.github.io/MACS/docs/cmbreps.html` suggests this approach to combine results across biological/technical replicates

An absolute fold-change threshold $|\log_2 \mathsf{FC}| > 0.50$ and FDR-corrected $p_{\mathsf{adj}} < 0.05$ were specified to determine significant disease-specific changes in chromatin state. Full details and code used to implement this workflow accompany this manuscript.

C.3 DAR-eQTL overlaps

First, we estimate the variance of fold enrichment. Let r_{DAR} denote the observed overlap ratio between a given method's DARs and the lead cis-eQTLs in each dataset, and let $\overline{r_{\mathsf{NULL}}}$ denote the average overlap ratio across m null replicates. Given uncorrelated r_{DAR}, $\overline{r_{\mathsf{NULL}}}$ of equal variance, a consistent estimator of the fold-enrichment variance is

$$\mathrm{var}\left(\frac{r_{\mathsf{DAR}}}{\overline{r_{\mathsf{NULL}}}}\right) \approx \left(\frac{r_{\mathsf{DAR}}}{\overline{r_{\mathsf{NULL}}}}\right)^2 \left[\frac{\widehat{\mathrm{var}}\left(r_{\mathsf{NULL}}\right)}{r_{\mathsf{DAR}}^2} + \frac{\widehat{\mathrm{var}}\left(r_{\mathsf{NULL}}\right)}{m\,\overline{r_{\mathsf{NULL}}}^2}\right].$$

We defined z-scores as the difference between the observed fold enrichment for each method and its expected value under the null. Fold enrichment is determined empirically with 10,000 block-bootstrap replicates obtained using `nullranges::bootRanges`.

UBR7	TYMP	TNK2	ISOC2
HMG20B	CLTCL1	SPEG	CDHR2
FSCN1	ZBTB25	BLVRB	SPAG7
MYO9B	PICK1	KIAA0930	OAZ1
RNF32	LMBR1	DPH1	WNT5B
ATN1	CUTA	DNAJC27	NDUFS7
CDK18	CAMSAP2	CPSF3	KIAA1217
KDM3B	B4GALT4	VPS26A	GIPC1
TBC1D20	AIF1L	SRD5A3	ATP8B3
CALY	LLGL1	CD164	HEY2
HINT2	GCSH	MTNAP1	SF3B4
SUSD4	HES6	TMF1	TNIP1
PCDHB18P	ANO7	SH3GLB2	SETD9
B3GNT7	LDLRAP1	CFAP410	RALGDS
MAML1	COX7A1	BICDL2	DRAXIN
MELTF	CASP3	SLU7	TRMT61A
NEMP1	CCDC178	GPX4	TK1
BMP1	CNNM3	TCTN2	IRS1
LRRC45	CENPX	DCXR	MFN1
HINFP	SNX32	ZNRD2	RCE1
HSD11B2	SHMT1	DPP7	MAN1B1
PIDD1	RPLP2	ZC3H12D	PLEC
KCTD12	SERTAD2	RCC1	CLRN3
AEN	FAM87A	ADI1	COL18A1
ATL3	CA13	SV2B	GABRA5
TERF1P2	ZFTRAF1	FBF1	MYO18A
ENSG00000197376	RAB40C	MAFK	C6orf89
OXCT2	MFAP3L	ZNF204P	CPT1B
PDPK2P	LINC01089	HAGLR	
TRAM2-AS1	ENSG00000226849	LMCD1-AS1	RPS18
RPS12P3	LINC00940	ENSG00000236209	B3GNT9
ABCA17P	MINDY2-DT	ENSG00000255581	LINC02352
LINC01220	ENSG00000260273	CORO7	ENSG00000262528
ENSG00000264548	SYPL1P2	FXYD1	ENSG00000270055
FAM193B-DT	ENSG00000272720	BNAT1	ENSG00000274367
ENSG00000274528	STPG3-AS1	Metazoa_SRP	ENSG00000277192
ENSG00000278995	C18orf15	ENSG00000280422	ENSG00000281350
GTF2IP12			

Table S4: List of genes with ROSMAP-identified lead eQTLs overlapping `Consenrich+ROCCO` DARs. Among these genes, WNT5B (Wnt Family Member 5B) encodes a ligand in the non-canonical WNT signaling pathway. Its elevated expression reflects a pathological shift from canonical to non-canonical signaling, which has been associated with reduced synaptic stability in Alzheimer's disease. [56]. The gene FSCN1 (Fascin Actin-Bundling Protein 1) is involved in actin filament bundling and can affect cytoskeletal dynamics. Additionally, its expression related to microglial motility can delineate APOE-ϵ4/4-AD individuals and APOE-ϵ3/3-AD individuals [57].

D Computational Efficiency

The time complexity of the estimation procedure is linear in the number of genomic intervals, n. Computing the gain in the forward pass at each interval requires inversion of the m-dimensional residual covariance, $\mathbf{E}_{[i]} \in \mathbb{R}^{m \times m}$ where m is the number of input samples. We leverage the the diagonal-plus-rank-one structure of the residual covariance to apply a Woodbury inversion [58] that side-steps the traditional computational overhead associated with dense matrices (cubic runtime). In theory, this only reduces the worst-case complexity to $\mathcal{O}(nm^2)$, but practical reuse of previous results via a simple hashing scheme applied during iteration effectively ensures this quadratic scaling in the sample size is not realized.

Several additional measures are taken to efficiently traverse and process data where possible: vectorized operations on numeric arrays, precomputing frequently referenced entities in loops, etc. Because these optimizations are not evident in big O notation, an empirical analysis of runtime is warranted. We run Consenrich for gradually increasing sample sizes $m = 10, 15, 20, 30$ and profile memory usage and wallclock runtime. Note that results were generated using default parameters limiting the number of simultaneous processes (Figure S1). At the cost of increased memory use, runtime can be substantially reduced with more aggressive multiprocessing.

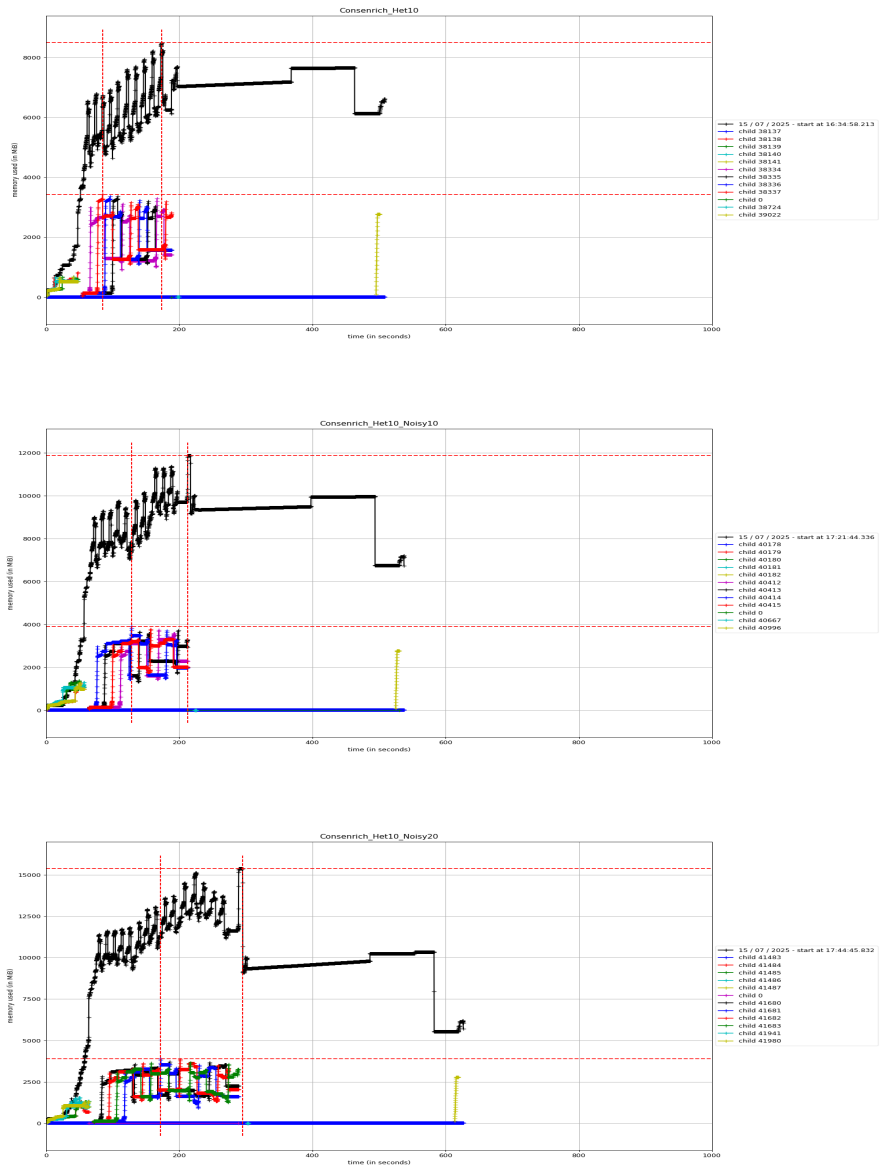


Figure S1: Runtime and memory use for increasing sample sizes $m = 10 \ldots 30$. Both the main process and its child processes are tracked. Despite a 3-fold increase in sample size, runtime shows only a 20% increase in this experiment.

D.1 Computing Environment

Software

Exact Python package versions used to run the experiments are included for reference but are not required in general.

```
numpy=2.2.5
pandas=2.2.3
scipy=1.15.2
pywavelets=1.8.0
pybigwig=0.3.24
pybedtools=0.11.0 # wraps bedtools v2.31.1
pysam=0.23.0 # wraps htslib v1.21
consenrich=0.1.13b1
```

Hardware

```
% sysctl -n hw.model kern.osrelease  hw.physicalcpu hw.physmem
Mac16,5
24.3.0
16
3343040512
```

Case I

```
mprof run --multiprocess consenrich -t ../het10/*.bam \
-g hg38 --chroms chr1 -p 4 --save_args
```

Case II

```
mprof run --multiprocess consenrich -t ../het10/*.bam  \
  ../het10/noisy/noisy_bams/noisy1.bam [...] ../het10/noisy/noisy_bams/noisy10.bam \
-g hg38 --chroms chr1 -p 4 --save_args
```

Case III

```
mprof run --multiprocess consenrich -t ../het10/*.bam \
../het10/noisy/noisy_bams/*.bam \
-g hg38 --chroms chr1 -p 4 --save_args
```

E Parameters

Basic default parameters, given in Table S5, are used throughout the manuscript unless otherwise specified. The genomic interval size L, minimum allowed eigenvalue for the process noise covariance P_{min}, and the model mismatch detection threshold α_D are particularly consequential parameters that can be used for tuning. In Figure S2, the qualitative behavior of Consenrich is evaluated for various choices of these parameters. Additional genome-wide empirical sensitivity analyses are included with this submission.

Table S5: Default Parameters

Parameter	Domain	Default Value
L	$\mathbb{N}$	25
$Q_\text{min}, Q_\text{max}$	$\mathbb{R}^+$	$5.0 \times 10^{-}2, \quad 5.0 \times 10^2$
$R_\text{min}, R_\text{max}$	$\mathbb{R}^+$	$\frac{1}{m} + 10^{-3}, \quad 5.0 \times 10^2$
$\omega^{\text{local}}, \omega^{\text{global}}$	$\mathbb{R}^+$	$\frac{1}{3}, \frac{2}{3}$
α_D	$\mathbb{R}^+$	2.0

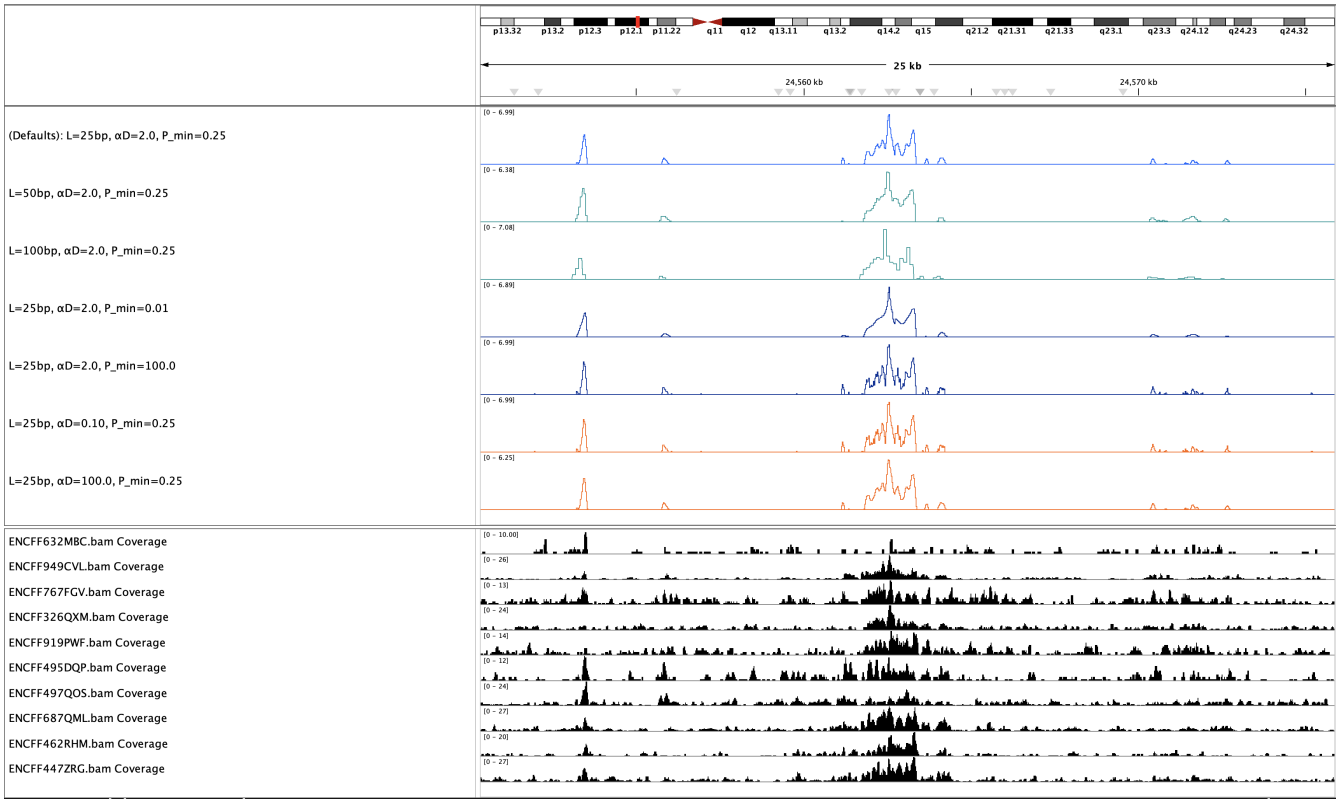
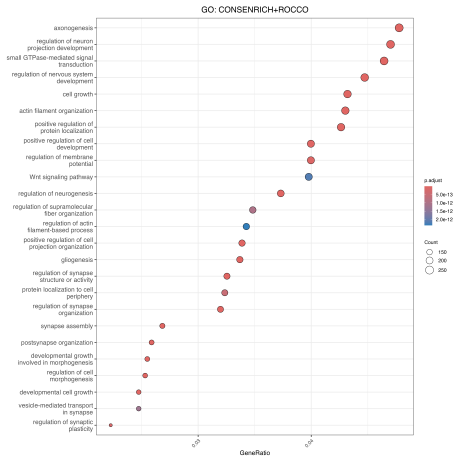
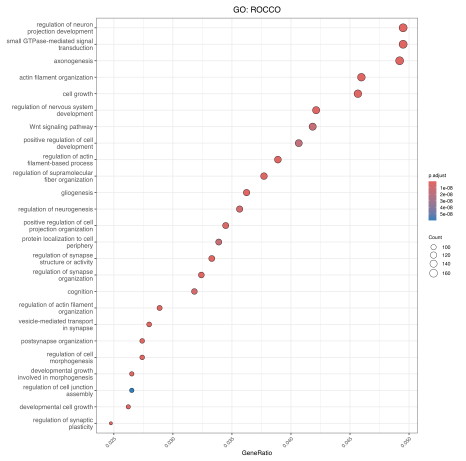


Figure S2: Consenrich signal estimates as several fundamental parameters vary: genomic interval length (L), minimum process noise covariance eigenvalue (P_{min}), and the mismatch threshold (α_D). For reference, signal estimates using default parameters are also displayed in the first row.

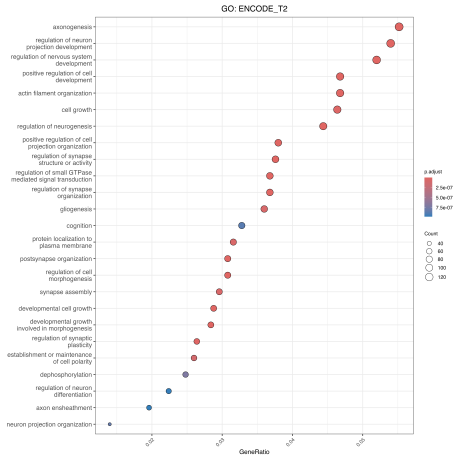
F DARs: GO Enrichment Analyses



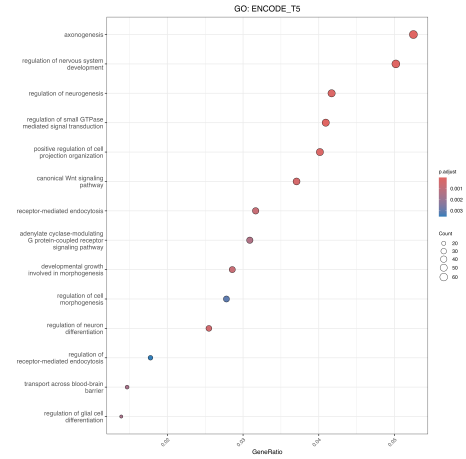
(a) DARs: Consenrich+ROCCO



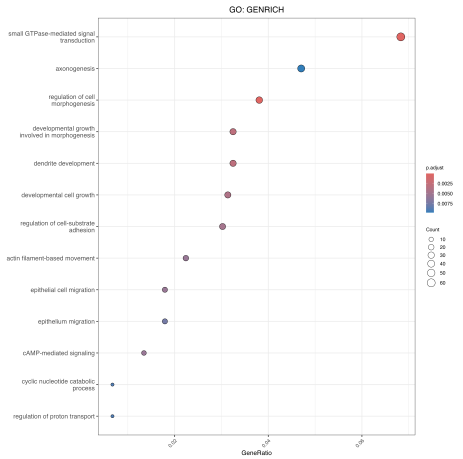
(b) DARs: ROCCO



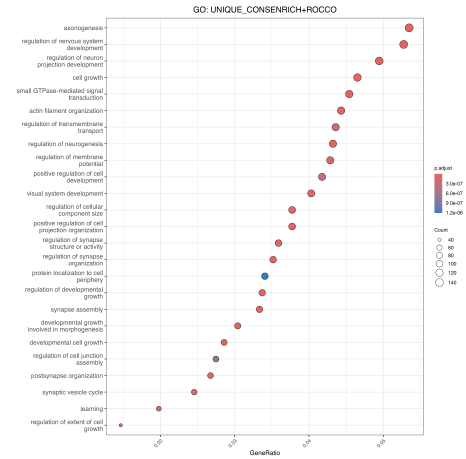
(c) DARs: ENCODE/UnionThreshold $T = 2$



(d) DARs: ENCODE/UnionThreshold $T = 5$



(e) DARs: Genrich



(f) DARs: Consenrich+ROCCO Unique

Figure S3: GO enrichment results for each method. An adjusted p-value of 0.01 and semantic similarity cutoff of 0.70 were used to identify top hits (up to 25) with minimal redundancy.