Evidence Against Homogeneity: Identifying Process Mismatch for Copy Number Variation Detection

Austin Talbot TALBOTA@PILLARBIOSCI.COM
Pillar Biosciences Inc
Natick, MA, USA

Alex V. Kotlar KOTLARA@PILLARBIOSCI.COM
Pillar Biosciences Inc
Natick, MA, USA

Yue Ke KEY@PILLARBIOSCI.COM
Pillar Biosciences Inc
Natick, MA, USA

Abstract

Batch effects represent a major confounder in genomic diagnostics. In copy number variant (CNV) detection from next-generation sequencing, many algorithms compare read depth between test samples and a reference derived from the processing batch, assuming samples are process-matched. When this assumption is violated, with causes ranging from reagent lot changes and sample quality differences to multi-site processing, the reference becomes inappropriate, introducing false CNV calls or masking true pathogenic variants. Detecting such heterogeneity before downstream analysis is critical for reliable clinical interpretation. Existing batch effect detection methods either cluster samples based on raw features, risking conflation of biological signal with technical variation, or require known batch labels that are frequently unavailable. We introduce a method that addresses both limitations by clustering samples according to their Bayesian model evidence. The central insight is that evidence quantifies compatibility between data and model assumptions, technical artifacts violate assumptions and reduce evidence. In contrast, biological variation, including CNV status, is anticipated by the model and yields high evidence. This asymmetry provides a discriminative signal that separates batch effects from biology. We formalize heterogeneity detection as a likelihood ratio test for mixture structure in evidence space, using parametric bootstrap calibration to ensure conservative false positive rates. We validate our approach on synthetic data demonstrating proper Type I error control, three clinical targeted sequencing panels (liquid biopsy, BRCA, and thalassemia) exhibiting distinct batch effect mechanisms, and mouse electrophysiology recordings demonstrating cross-modality generalization. Our method achieves superior clustering accuracy compared to standard correlation-based and dimensionality-reduction approaches while maintaining the conservativeness required for clinical usage.

1. Introduction

In next-generation sequencing (NGS)-based copy number variant (CNV) detection, widely used algorithms estimate copy number by comparing per-locus read depth in a test sample against reference samples (Talbot et al., 2026; Plagnol et al., 2012; Budczies et al., 2016). This comparison removes amplicon-specific amplification biases, but its validity depends

on a process-matching assumption: that test and reference samples were prepared with the same reagents, on the same instrument, and through the same bioinformatics pipeline (Krumm et al., 2012; Alkan et al., 2011; Fromer et al., 2012). In practice, this assumption is routinely violated by reagent lot changes, sample quality differences, and multi-site processing. The consequences are clinically significant: an undetected mismatch can produce false CNV calls that trigger unnecessary interventions, or mask true pathogenic variants in disorders such as thalassemia (Zhu et al., 2023) and cancer (Risinskaya et al., 2023; Becchi et al., 2023). Because these errors present as confident calls rather than flagged failures, they propagate silently through downstream clinical interpretation.

Existing methods for detecting batch heterogeneity are inadequate for this setting. Unsupervised approaches cluster samples on raw features such as principal components of coverage profiles (Leek and Storey, 2007). However, these approaches cannot distinguish technical artifacts from true biological signal: a subset of samples carrying genuine CNVs produces a distinctive coverage pattern that feature-space clustering will attribute to a batch effect. Supervised methods test the association between recorded batch labels and measurements (Johnson et al., 2007), but reliable batch metadata are frequently unavailable in clinical workflows. A practical detection method must therefore satisfy several criteria simultaneously: it must be label-free, since sample origin is often unrecorded; conservative, since false alarms trigger costly re-sequencing; sensitive to diverse technical effects beyond simple mean shifts; and, critically, it must separate technical artifacts from biological variation, so that batches containing CNV-positive patients are not erroneously flagged. No existing method addresses all four requirements.

We introduce a heterogeneity detection method, currently integrated into a commercial clinical NGS pipeline, that meets these criteria by operating in evidence space rather than feature space. For each sample, we compute the Bayesian model evidence (Kass and Raftery, 1995; MacKay, 2003), the marginal likelihood of the sample's data under the assumed CNV model, integrated over parameter uncertainty. The critical aspect of this approach is that model evidence measures compatibility between data and model assumptions. A CNV detection model anticipates copy-number changes; CNV-positive samples are well explained and yield high evidence. Technical artifacts, by contrast, introduce noise patterns the model does not anticipate, violating its distributional assumptions and systematically reducing evidence. This asymmetry provides a discriminative signal: low evidence indicates assumption violation, while high evidence, even from biologically unusual samples (i.e. clinical positives), indicates that the model remains well-specified. We formalize heterogeneity detection as a likelihood ratio test for mixture structure in the distribution of per-sample evidence scores, calibrated via the parametric bootstrap to maintain conservative false-positive control.

Because the method requires only a per-sample likelihood, it is not specific to genomics. Any domain in which a generative model is fit to individual samples can use this technique to answer the question: are all samples compatible with a single data-generating process, or does the cohort contain subsets whose data systematically deviate from model assumptions? In systems neuroscience, for example, combining electrophysiological recordings from different experimental cohorts or laboratories introduces unmeasured technical variation from differences in electrode impedance, recording hardware, and environmental conditions, which is a problem structurally identical to batch effects in sequencing. We validate our method

on synthetic data with controlled heterogeneity, three clinical targeted sequencing panels (liquid biopsy, BRCA, and thalassemia) representing distinct batch-effect mechanisms, and local field potential recordings from mouse experiments where cross-cohort batch structure is independently verifiable. Across all settings, the method detects heterogeneity when present and consistently outperforms correlation-based and dimensionality-reduction alternatives in clustering accuracy. Python code implementing all models and reproducing all results is available at `https://github.com/Pillar-Biosciences-Inc/BatchDetect`.

Generalizable Insights about Machine Learning in the Context of Healthcare

- **Per-sample likelihood is a label-free data-quality signal.** Any clinical ML system that computes a per-sample log-likelihood can repurpose that quantity as a principled indicator of data-quality degradation. Unlike held-out prediction error, the marginal likelihood requires no labeled reference set and does not conflate sample difficulty (a biologically unusual but valid sample) with sample contamination (a technically degraded sample that violates model assumptions).
- **Distribution-shift detection is effective in evidence space.** Monitoring data quality through the distribution of model-fit summaries, rather than input features, sidesteps the fundamental challenge of disentangling technical from biological variation, a problem that confounds feature-space approaches across modalities.
- **Bootstrap calibration is important for hypothesis testing in mixture models.** In clinical settings, false alarms carry direct operational costs: repeat sequencing, manual expert review, and diagnostic delay. The parametric bootstrap framework introduced here provides explicit control over the false-positive rate, which asymptotic alternatives do not guarantee for non-regular mixture models.

2. Related Work

Batch correction and reference selection in CNV detection. Batch effects have long been recognized as a major confounder in genomic analyses (Leek et al., 2010), and a substantial literature addresses their correction. In CNV detection from targeted panels, the dominant strategy is dynamic reference selection: ExomeDepth (Plagnol et al., 2012) selects reference panels via a correlation-guided greedy procedure that maximizes expected detection power under a beta-binomial depth model; CLAMMS (Packer et al., 2016) and CANOES (Backenroth et al., 2014) employ similar correlation- or distance-based strategies. These approaches can substantially improve CNV calling accuracy, particularly in targeted panels where sample-to-sample variability is high (Kuśmirek et al., 2019). More recent work, such as clearCNV (May et al., 2022), goes further by implementing unsupervised clustering to partition samples into batches before calling, though it provides no formal test for whether such partitioning is statistically warranted. In a complementary direction, grouping samples by known batch labels (e.g., library preparation date) prior to calling can reduce false positives (Krumm et al., 2012). General-purpose batch correction methods, such as ComBat (Johnson et al., 2007) and surrogate variable analysis (Leek and Storey, 2007), are widely used across genomic assay types. However, all of these approaches assume

that batch heterogeneity exists (or that correction is harmless) and attempt to remove it, but do not formally test whether heterogeneity is present.

Batch effect detection. In contrast to the mature correction literature, formally detecting batch heterogeneity has received less attention. The most common informal approach is visual inspection of principal component plots, examining whether samples separate along suspected batch variables, but this provides no calibrated false-positive guarantee and requires a suspected covariate to be available for stratification. Formal methods span several categories. Variance-attribution methods such as PVCA (Li et al., 2009) decompose total variance into components associated with known covariates, but require batch labels as input and, like any method that operates on a full sample covariance matrix, become unreliable when $p \gg n$ or batch sizes are small. Supervised PCA-based tests such as gPCA (Reese et al., 2013) test whether a batch covariate explains significant variance in the leading principal components, but again require labels and share the same high-dimensional covariance instability. Distributional tests such as quantro (Hicks and Irizarry, 2015) assess whether the global distributional assumptions underlying quantile normalization hold across pre-defined sample groups, but require those groups to be specified in advance and are designed for normalization guidance on count or intensity data rather than for the read-depth ratio setting of targeted CNV panels. In the single-cell domain, kBET (Büttner et al., 2019) measures local batch mixing via χ^2 tests on k-nearest-neighbor neighborhoods, and BatchQC (Manimaran et al., 2016) provides interactive diagnostics combining several of the above approaches; both require known batch labels.

Testing for mixture structure via the bootstrap. Our approach reformulates batch detection as a test for mixture structure in model-evidence space: if sample-level Bayesian evidence follows a unimodal distribution, the dataset is homogeneous; if it follows a mixture, heterogeneity is present. This reformulation motivates the following statistical background. The classical difficulty in testing for the number of components in a finite mixture is that the likelihood ratio test (LRT) statistic does not follow the standard χ^2 distribution under the null, because the null lies on the boundary of the parameter space. McLachlan (1987) introduced the parametric bootstrap as a practical solution: the null distribution of the LRT is approximated by repeatedly simulating data from the fitted single-component model and recomputing the statistic. This approach has been widely adopted for mixture model selection (McLachlan and Peel, 2000) and remains the standard method for calibrating mixture tests when asymptotic theory is unavailable. An alternative is to derive a modified asymptotic null distribution, as advocated by Lo et al. (2001).

3. Methods

Our principal focus is targeted amplicon sequencing for CNV detection. In this situation, predefined genomic loci (amplicons) are PCR-amplified and sequenced. The resulting read counts are approximately proportional to underlying copy number, with variability arising from PCR efficiency, sample effects, and sequencing noise. Preprocessing transforms raw counts into centered log ratios suitable for statistical modeling.

Preprocessing. Let J denote the number of targeted amplicons retained after quality control for a given sample. For amplicon $j \in \{1, \ldots, J\}$, we observe a read count s_j from

the test sample and a corresponding count n_j from a process-matched reference, such as a matched normal, a summary (e.g., median) across a panel of normals, or a batch-wide summary for normal-free calling. We form log ratios $\tilde{y}_j = \log(s_j) - \log(n_j)$; amplicons with low counts are removed during QC, so division by zero is not a concern. To remove global differences in sequencing depth, we center by subtracting a robust location estimate:

$$y_j = \tilde{y}_j - \text{median}_{j' \in \mathcal{S}} \, \tilde{y}_{j'}, \tag{1}$$

where $\mathcal{S}$ is either the full set of retained amplicons or a predefined subset of copy-neutral control amplicons. After centering, $y_j \approx 0$ indicates copy neutrality, negative values indicate deletions, and positive values indicate amplifications. We refer to $\{y_j\}_{j=1}^{J}$ as log copy-number ratios (lCNRs). The left panel of Figure 1 illustrates the contrast between two representative lCNR profiles after this centering step: one sample is noisy yet CNV-negative, its amplicons scattering broadly around zero; the other is low-noise and CNV-positive, with amplicons clearly displaced from zero. Importantly, the CNV-positive sample yields markedly higher model evidence, underscoring that elevated evidence scores reflect genuine copy-number signal and are not merely a surrogate for technical data quality.

3.1. CNV Models and Evidence Computation

We employ two Bayesian CNV models, each of which returns a log marginal likelihood (model evidence) quantifying how well it explains a given sample's lCNR profile. The first is a hierarchical Bayesian model for gene-level calling in liquid biopsy panels (Talbot et al., 2026): amplicons targeting the same gene share a latent gene-level log copy-number ratio, observations follow a heavy-tailed likelihood, and inference proceeds via HMC with the log marginal likelihood estimated by thermodynamic integration. The second is a state-space model for thalassemia panels (Talbot et al., 2025), which captures spatial correlations among amplicons and performs exon-level calling; inference proceeds via sequential Monte Carlo, yielding model evidence through importance weights. Full model specifications appear in Section A of the Appendix.

For each sample $i \in \{1, \ldots, N\}$ in a cohort, we compute the log marginal likelihood under the chosen CNV model and normalize by the number of amplicons J_i passing QC for that sample, yielding a per-amplicon evidence score ℓ_i. This normalization ensures comparability across samples that may differ in amplicon yield. As shown in the middle panel of Figure 1, the distribution of $\{\ell_i\}$ across a batch can exhibit pronounced bimodality, a pattern we exploit to detect technical heterogeneity, as described next.

3.2. Detecting Batch Heterogeneity via Mixture Modeling

Batch heterogeneity manifests as systematic differences in model compatibility across samples: if a cohort contains subsets processed under different technical conditions, one subset may be consistently better (or worse) explained by the CNV model, producing a multimodal distribution of $\{\ell_i\}$ (middle panel of Figure 1). We formalize this as a likelihood-ratio test (LRT) comparing a one-component model (homogeneous cohort) against a two-component mixture (heterogeneous cohort).

Let $f(\cdot \mid \boldsymbol{\theta})$ denote a location-scale density parameterized by $\boldsymbol{\theta}$, which collects the location, scale, and (where applicable) shape parameters of the chosen family. We consider

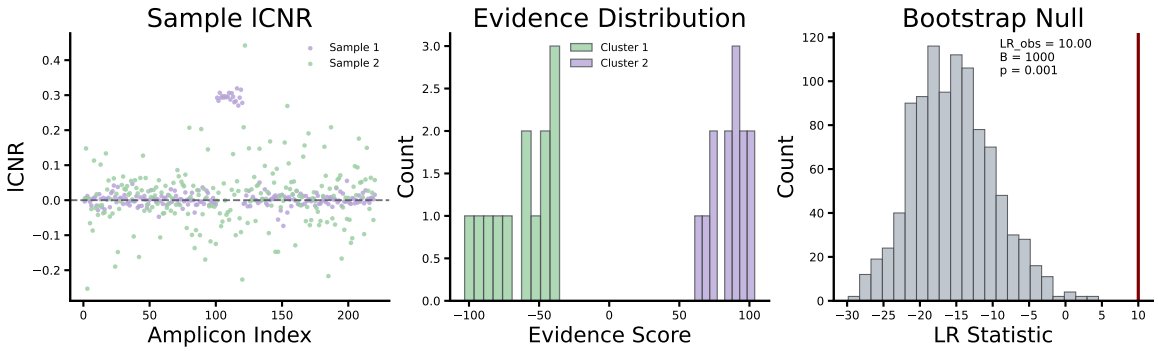

Figure 1: **Left:** Centered lCNR profiles (1) for two representative samples: a noisy, CNV-negative sample whose amplicons scatter broadly around zero, and a low-noise, CNV-positive sample whose amplicons are clearly displaced from zero. Crucially, the CNV-positive sample yields high model evidence, illustrating that elevated evidence scores reflect genuine copy-number signal rather than data quality alone. **Middle:** Distribution of per-sample evidence scores $\{\ell_i\}$ across the batch, exhibiting bimodality; colors indicate batch-group labels. **Right:** Parametric bootstrap null distribution of the LR statistic (Algorithm 1), with the observed $\mathrm{LR}_{\mathrm{obs}}$ marked.

five candidate families: Gaussian, Laplace, hypersecant, Student-t, and generalized normal. The competing hypotheses are

$$H_0: \quad \ell_i \sim f(\ell \mid \boldsymbol{\theta}_1), \tag{2}$$

$$H_1: \quad \ell_i \sim \pi\, f(\ell \mid \boldsymbol{\theta}_1) + (1-\pi)\, f(\ell \mid \boldsymbol{\theta}_2), \tag{3}$$

where $\pi \in (0,1)$ is the mixture weight. Parameters are estimated via expectation-maximization (EM); the two-component model is initialized from k-means centers over n_{init} random restarts, retaining the solution with the highest log likelihood. Convergence is declared when the change in mean log likelihood falls below 10^{-4} or after 1000 iterations.

3.3. Parametric Bootstrap Significance Test

Because finite mixtures are non-regular (the null lies on the boundary of the alternative's parameter space), the standard χ^2 approximation for the LRT does not hold (McLachlan and Peel, 2000). We therefore calibrate significance via a parametric bootstrap (Davison and Hinkley, 1997), summarized in Algorithm 1. The procedure fits both models to the observed scores and computes $\mathrm{LR}_{\mathrm{obs}}$, then generates B synthetic datasets from the fitted null $\widehat{M}_0$ to build an empirical reference distribution. The resulting null distribution and the placement of $\mathrm{LR}_{\mathrm{obs}}$ within it are shown in the right panel of Figure 1. The p-value is the fraction of bootstrap replicates whose LR equals or exceeds $\mathrm{LR}_{\mathrm{obs}}$, with a continuity correction that includes the observed statistic itself. Small p-values indicate that the evidence-score distribution is unlikely under a single homogeneous population, pointing to the presence of batch heterogeneity.

Algorithm 1: Parametric Bootstrap LRT for Batch Heterogeneity
Input: Evidence scores $\{\ell_i\}_{i=1}^N$, location-scale family f, bootstrap replicates B
Output: p-value and observed likelihood ratio $\mathrm{LR}_{\mathrm{obs}}$

Fit 1-component model $f(\cdot \mid \boldsymbol{\theta})$ to $\{\ell_i\} \to \widehat{M}_0$
Fit 2-component mixture of f to $\{\ell_i\} \to \widehat{M}_1$
$\mathrm{LR}_{\mathrm{obs}} \leftarrow 2\big(\mathcal{L}(\{\ell_i\} \mid \widehat{M}_1) - \mathcal{L}(\{\ell_i\} \mid \widehat{M}_0)\big)$
for $b \leftarrow 1$ **to** B **do**
  Sample $\{\ell_i^{(b)}\}_{i=1}^N \sim \widehat{M}_0$
  Fit 1-component model to $\{\ell_i^{(b)}\} \to \widehat{M}_0^{(b)}$
  Fit 2-component mixture to $\{\ell_i^{(b)}\} \to \widehat{M}_1^{(b)}$
  $\mathrm{LR}_b \leftarrow 2\big(\mathcal{L}(\{\ell_i^{(b)}\} \mid \widehat{M}_1^{(b)}) - \mathcal{L}(\{\ell_i^{(b)}\} \mid \widehat{M}_0^{(b)})\big)$
end
$p \leftarrow \dfrac{1 + \sum_{b=1}^{B} \mathbb{I}(\mathrm{LR}_b \geq \mathrm{LR}_{\mathrm{obs}})}{B+1}$
return p, LR_{obs}

A natural alternative to parametric bootstrap calibration is the Lo–Mendell–Rubin (LMR) adjusted likelihood-ratio test (Lo et al., 2001), which provides an asymptotic approximation to the null distribution and avoids the computational cost of repeated simulation. We initially evaluated the LMR test as a faster option for clinical deployment and include it as a baseline because it is well established in the finite-mixture literature. However, its approximation was developed primarily for normal-mixture settings, whereas our framework considers several non-Gaussian component families and relatively small clinical batch sizes. We therefore compare its empirical calibration with that of the parametric bootstrap in Appendix C.1.

4. Synthetic Results: A Conservative Null Distribution

4.1. Motivation and Setup

We first evaluate our method on synthetic data, focusing on two properties critical for clinical deployment: calibration under the null and robustness to single-sample outliers. Robustness to lone outliers is particularly important in practice, as customers frequently include an unmatched reference sample, such as a contrived sample with a known CNV, to benchmark caller performance alongside the test samples; such a sample can distort the batch evidence-score distribution if the heterogeneity test is not outlier-tolerant.

To assess both properties, we generate lCNR data from the hierarchical model in (6) using a 439-amplicon liquid-biopsy panel. We simulate 1,000 batches of $N = 20$ samples each, constructing a batch-mean reference, fitting the hierarchical CNV model, computing per-sample evidence scores, and applying the heterogeneity test to obtain a batch-level p-value. In Experiment 1 (no outlier) all samples share the same generative process; in Experiment 2 (outlier) the worst sample per batch is shifted by -2σ. We compare five evidence-space component densities: Gaussian, Student-t ($\nu = 3$), generalized normal (`gennorm`), Laplace, and hypersecant.

4.2. Calibration Diagnostics

Under a perfectly calibrated null, p-values follow Beta$(1, 1)$. We summarize departures via the one-parameter power family

$$p_i \sim \text{Beta}(a, 1), \qquad i = 1, \ldots, n, \tag{4}$$

with density $f(p) = a\, p^{a-1}$: $a > 1$ is conservative, $a \approx 1$ is calibrated, and $a < 1$ is anti-conservative (Hung et al., 1997). This scalar summary is preferable to omnibus goodness-of-fit tests, which have low power at $n = 1{,}000$ against the one-sided alternatives of clinical interest; it also admits a conjugate Gamma posterior, from which we report the posterior mean and 95% credible interval. We complement this with the one-sided Clopper–Pearson 95% upper bound on $\mathbb{P}(p < 0.05 \mid H_0)$ (Lehmann and Romano, 2005), providing a direct guarantee on the false-alarm rate at a common operating threshold.

4.3. Results

Table 1 summarizes calibration across all five distributions and both experimental conditions; Figure 2 shows the p-value histogram and Q–Q plot for the Laplace kernel under the null.

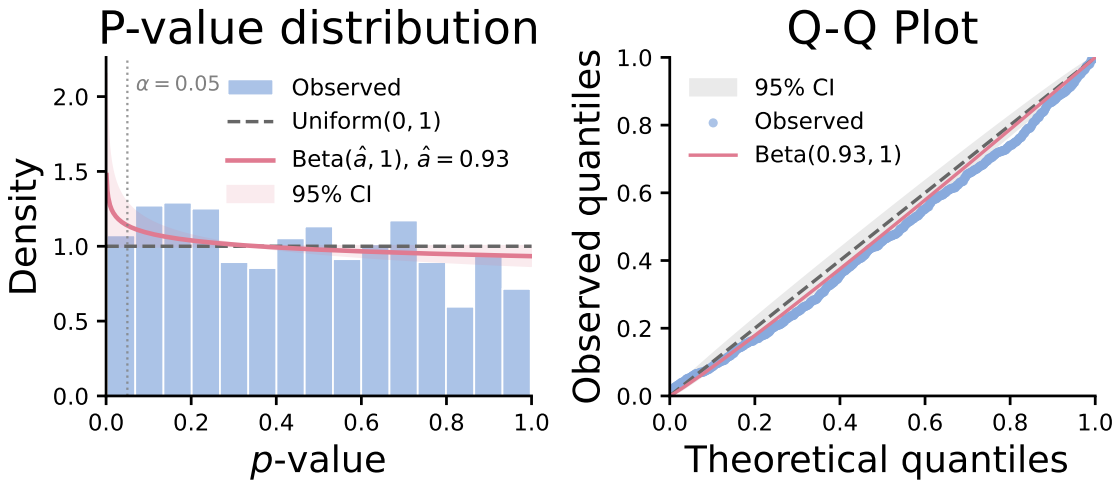

Figure 2: Calibration diagnostics for Experiment 1 (no outlier) using the Laplace kernel. **Left:** Histogram of observed p-values (blue) with the theoretical uniform density (dashed gray) and the fitted Beta$(\hat{a}, 1)$ density (red) overlaid; the shaded region represents the 95% posterior credible interval on $\hat{a}$. **Right:** Q–Q plot of observed p-values against uniform quantiles (dashed gray) and fitted Beta$(\hat{a}, 1)$ quantiles (red); the shaded band is the 95% confidence envelope for order statistics under the null.

Experiment 1: No outlier (null). Four of the five distributions are nearly calibrated or mildly conservative ($\hat{a} \in [0.93, 1.01]$, CP bounds ≤ 0.07). Laplace is the exception ($\hat{a} = 0.76$, CP bound $= 0.11$): its central spike and heavy tails inflate the mixture likelihood ratio even in the absence of a true outlier, producing mild anti-conservatism under the null.

Experiment 2: Outlier introduced (alternative). The results change substantially when an outlier is introduced. Gaussian is markedly anti-conservative ($\hat{a} = 0.25$, CP bound $= 0.74$): its thin tails assign extreme likelihood ratios to the shifted sample, generating a high rate of spurious detections. Hypersecant lies at the opposite extreme ($\hat{a} = 1.66$), its heavy tails absorbing the outlier so completely that the test becomes insensitive. Laplace, a less-calibrated distribution under the null, is the best calibrated under the alternative ($\hat{a} = 0.91$, CP bound $= 0.06$): its tails are heavy enough to accommodate the outlier score without producing an extreme likelihood ratio, yet not so heavy as to suppress sensitivity altogether. In practice we have found the Laplace distribution to be the most reliable, precisely due to this robustness.

Comparison with the Lo–Mendell–Rubin test. As motivated in Section 3.3, we also evaluated the computationally less expensive LMR approximation. Appendix C.1 shows that the LMR test is substantially less stable across component distributions and generally anti-conservative, supporting our use of parametric bootstrap calibration.

Table 1: Calibration results with and without an outlier. $\hat{a}$ is the Beta$(a, 1)$ posterior mean with 95% credible interval; values near 1 indicate calibration, < 1 anti-conservativeness, and > 1 conservativeness. CP bound is the one-sided 95% Clopper–Pearson upper bound on $\mathbb{P}(\text{reject} \mid H_0)$ at $\alpha = 0.05$; values ≤ 0.05 indicate that the false-alarm rate is controlled at the nominal level.

	No Outlier		**Outlier Included**	
Distribution	$\hat{a}$ **[95% CI]**	**CP bound**	$\hat{a}$ **[95% CI]**	**CP bound**
Gaussian	**0.96** (**0.90**, **1.01**)	**0.06**	0.25 (0.23, 0.26)	0.74
Hypersecant	0.92 (0.86, 0.97)	0.06	1.66 (1.56, 1.77)	0.05
Student-t	1.03 (0.97, 1.10)	0.02	0.57 (0.54, 0.61)	0.34
Laplace	0.76 (0.72, 0.81)	0.11	**0.91** (**0.86**, **0.97**)	**0.06**
Gennorm	0.94 (0.88, 1.00)	0.04	0.42 (0.40, 0.45)	0.41

5. Cohorts

A detailed description of preprocessing and quality control is provided in Appendix B. Here we summarize the datasets used for evaluation, emphasizing sample sizes, batch factors, and biological signals relevant to batch-effect detection. We consider two modalities: targeted amplicon sequencing (three clinical panels, $n = 24$–94) and mouse electrophysiology (a combination of two studies with $n = 54$). Key dataset characteristics are summarized in Table 2.

Table 2: Summary of datasets used for batch effect algorithm validation.

Dataset	N	Batch Factor	Bio. Signal	Validation Purpose
Liquid Biopsy	32	FFPE quality	None	Technical artifact robustness
BRCA Panel	24	Reagent lot	CNV+	Batch vs. biology discrimination
Thalassemia Panel	94	Laboratory	CNV+	Multi-site heterogeneity
Electrophysiology	54	Experiment	Behavior	Cross-modality generalization

5.1. Targeted Amplicon Sequencing Data

We evaluate three targeted panels spanning distinct, practically relevant batch-effect mechanisms; panel design details and preprocessing are provided in Appendix B.

5.1.1. Multi-cancer Liquid Biopsy Panel (LBX)

This dataset comprises 32 FFPE samples sequenced with a 173-amplicon panel designed for low-cost cancer screening (targeting five CNV-associated genes). The batch factor is FFPE preservation quality: 10 libraries were prepared from substantially fragmented FFPE DNA, and 22 from higher-quality clinical FFPE material. This cohort evaluates robustness to amplification-related technical heterogeneity induced by mixed sample quality.

5.1.2. Thalassemia Panel

The thalassemia dataset uses a 120-amplicon panel (including normalization controls) targeting the alpha- and beta-globin loci. The batch factor is laboratory origin in a two-site setting (Laboratory A: $n = 49$; Laboratory B: $n = 45$), reflecting multi-site heterogeneity from differences in protocols, reagents, instrumentation, and operator effects. Biological heterogeneity is substantial: 48 samples carry multiple types of CNVs affecting the alpha-globin locus, of multiple types, including alpha-3.7 deletions, alpha-SEA deletions, and alpha gains. Overall, this cohort evaluates batch-effect detection in a realistic clinical genetics scenario combining multi-site technical variation with strong biological signal.

5.1.3. BRCA Panel

The BRCA dataset targets *BRCA1* and *BRCA2* using 283 amplicons (including normalization controls). The batch factor is reagent lot: 24 samples were sequenced across two lots (lot A: $n = 16$; lot B: $n = 8$), introducing systematic shifts in amplification efficiency and baseline coverage profiles. Biological signal is present via CNVs in 12 samples (six with gene-level *BRCA2* gains/losses; six with exon-level *BRCA1* aberrations), enabling assessment of whether heterogeneity is correctly attributed to batch effects rather than true copy-number variation. This is the weakest of the heterogeneity effects measured, and is meant to evaluate the limits of performance.

5.2. Mouse Electrophysiology Data

Combining neural recordings across experimental cohorts is common in systems neuroscience but introduces unmeasured technical variation from differences in electrode impedance,

recording hardware, surgical implantation, and environmental conditions. These effects are structurally analogous to batch effects in sequencing: they corrupt the shared statistical model without being anticipated by it. We evaluate our method on local field potential (LFP) recordings from 54 mice aggregated from two experiments, in which the batch structure is known and independently verifiable.

The first study (SP) recorded LFPs from 26 mice across 8 brain regions during social (paired interaction) and non-social (isolated exploration) conditions (Mague et al., 2022). The second study (TST) recorded from 28 mice across 11 brain regions during stress-exposure and control conditions; 7 brain regions overlapped with SP (Talbot et al., 2023). Feature processing is described in Appendix B. Within each experiment, protocols were highly controlled, minimizing intra-experiment technical variation; however, the SP data were collected in two sequential cohorts separated by several years, introducing a potential source of instrumentation or environmental drift. We therefore expect no batch effects within the TST experiment, possible effects within SP, and clear effects when pooling data across the two experiments. This graded expectation, from absent to probable to near-certain, provides a natural test of both sensitivity and specificity.

6. Results

6.1. CNV Heterogeneity Detection

6.1.1. Overall Performance

We evaluate our method on each of the three CNV datasets using the Laplace mixture as the primary component distribution, motivated by its favorable calibration properties established in Section 4 (conservative under the null, with well-controlled Type I error). To assess sensitivity to this choice, we also report results for four additional component distributions (Gaussian, hypersecant, Student-t, generalized normal). All experiments use $B = 10{,}000$ bootstrap iterations. For each dataset we test the full cross-batch comparison (**Overall**) as well as the two within-batch subgroups where no batch effects are expected (**Subgroups**). Table 3 summarizes the results. Note that, in clinical/scientific usage, only one kernel is used, avoiding multiple hypothesis testing concerns.

Under the Laplace mixture, batch effects are significant in all three datasets: thalassemia ($p = 0.001$), LBX FFPE ($p = 0.011$), and BRCA ($p = 0.028$). This pattern is broadly confirmed across the other kernels: four out of five distributions detect significant batch effects in thalassemia ($p \leq 0.005$ except hypersecant, $p = 0.161$), and all five detect them in LBX FFPE ($p \leq 0.031$). The hypersecant mixture is the exception, failing to reach significance on thalassemia despite a large sample size; this is consistent with the ultra-conservative behavior of hypersecant observed in our synthetic experiments (Section 4). The BRCA panel presents the most challenging scenario: only the Laplace mixture yields a significant result ($p = 0.028$), with all other kernels producing $p \geq 0.055$. We attribute this to the high prevalence of large germline CNVs (50% of clinical samples harbor gene-level *BRCA2* deletions), which dominate the evidence scores and reduce the separation between batch-driven and biology-driven components. The distribution of likelihoods in each of the two groups for the three CNV panels is shown in Figure 4.

Table 3: P-values for batch heterogeneity detection on the three CNV panels. **Overall**: test comparing samples across the two known batches. **Subgroups**: tests within each batch (where no batch effects are expected). Laplace is the recommended kernel based on synthetic calibration results (Section 4); other kernels serve as a sensitivity analysis. Values of 0.001 indicate $p \leq 0.001$ (the resolution floor for $B = 10{,}000$ bootstrap iterations).

	BRCA		**Thalassemia**		**LBX FFPE**	
Distribution	**Overall**	**Subgroups**	**Overall**	**Subgroups**	**Overall**	**Subgroups**
Gaussian	0.070	0.130, 0.477	0.001	0.023, 0.032	0.013	0.352, 0.499
Hypersecant	0.800	0.177, 0.524	0.161	0.069, 0.155	0.007	0.682, 0.108
Student-t	0.077	0.122, 0.477	0.005	0.059, 0.180	0.031	0.303, 0.294
Laplace	0.028	0.057, 0.110	0.001	0.118, 0.650	0.011	0.233, 0.141
Gennorm	0.055	0.092, 0.743	0.001	0.039, 0.101	0.016	0.261, 0.264

Critically, the within-batch subgroup tests assess whether the method erroneously detects structure where none is expected. Under the Laplace mixture, our recommended kernel, no subgroup reaches significance: the smallest Laplace subgroup p-value is 0.057 (BRCA), with thalassemia subgroups at 0.118 and 0.650 and LBX subgroups at 0.233 and 0.141. This is consistent with the conservative calibration of Laplace demonstrated in Section 4.

The sensitivity analysis reveals that kernel choice does matter at the margins. Under the Gaussian mixture, two thalassemia subgroups yield nominally significant p-values ($p = 0.023$ and $p = 0.032$), and the generalized normal mixture produces one borderline subgroup ($p = 0.039$). These results are not unexpected: the thalassemia dataset has the largest sample size and a high CNV-positive rate, creating substantial biological heterogeneity within each batch. Moreover, the synthetic experiments in Section 4 showed that the Gaussian kernel tends toward anti-conservatism, particularly in the presence of outliers. Thus, the Gaussian subgroup significance likely reflects a combination of biological substructure and the known calibration weakness of thin-tailed kernels, rather than a true batch effect. The key observation is that the *magnitude* of the overall cross-batch signal ($p = 0.001$) dwarfs the within-batch signal ($p = 0.023$) even under Gaussian, and under the recommended Laplace kernel the within-batch tests are comfortably not significant.

6.1.2. Statistical Power and Size

We evaluate statistical power by repeatedly resampling from both batches at sample sizes ranging from 5 to 30, covering the batch sizes commonly encountered in clinical targeted sequencing. Statistical size (Type I error) is assessed by drawing clinical samples exclusively from a single batch at sizes of 12 and 20. All experiments use the Laplace mixture. Results are shown in Figure 3.

Power is limited at sample sizes below 15 across all datasets, but exceeds 0.8 by $n = 25$–30. Given that introductory clinical validation batches are typically of comparable size, this is sufficient to detect meaningful process mismatches before clinical samples proceed to

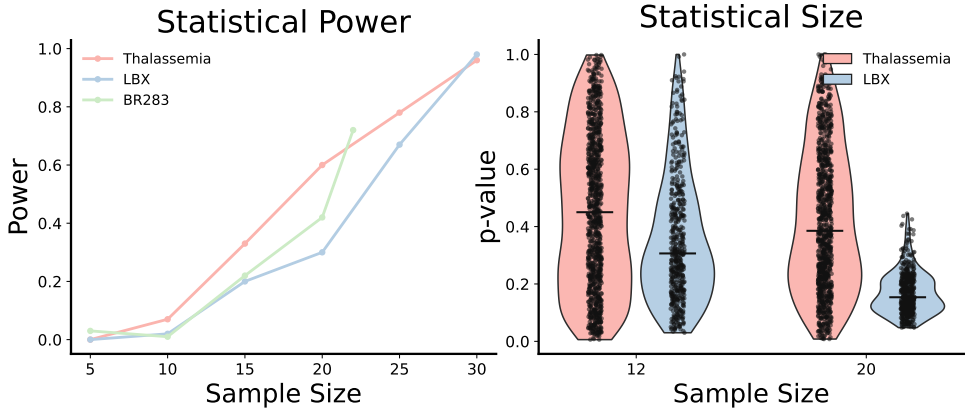

Figure 3: **Left:** Statistical power as a function of sample size for each CNV dataset (Laplace mixture, $\alpha = 0.05$). **Right:** Distribution of p-values under the null (within-batch resampling) at sample sizes of 12 and 20 for the thalassemia and LBX assays. Note that at 20 clinical samples, the bootstrap evaluation becomes poor, resulting in compressed p-values.

CNV calling. Statistical size is well-controlled: at most 4.4% of within-batch tests reject at $\alpha = 0.05$ across both sample sizes and both datasets, confirming the conservativeness of the Laplace-based test. The LBX p-value distribution appears compressed toward the center, likely an artifact of resampling redundancy (drawing 20 clinical samples from a pool of 24 introduces substantial overlap between replicates).

6.1.3. Clustering Accuracy

We compare our likelihood-based batch assignment against five unsupervised clustering baselines advocated in Kuśmirek et al. (2019). The first family operates on self-normalized read count matrices: **GMM** (Gaussian mixture models) and **PCA-Ward** (PCA followed by Ward's hierarchical clustering). The second family first computes pairwise sample correlations (Pearson or Spearman) and then applies **Hierarchical** clustering (average linkage on $1 - r_{ij}$), **Spectral** clustering (Gaussian affinity on thresholded correlations), or **PCA-KMeans** (eigendecomposition of the correlation matrix followed by k-means). All methods are provided the true number of clusters $k = 2$, which gives the baselines an advantage our method does not require.

Our method achieves the best or tied-best performance across all three datasets (Table 4). On LBX FFPE it achieves perfect classification (accuracy $= 1.0$, $\kappa = 1.0$), outperforming the next-best methods (Spectral, PCA-KMeans; $\kappa \leq 0.86$). On thalassemia, accuracy reaches 0.90 ($\kappa = 0.80$), a 5–14 percentage point improvement over all baselines, despite 55% of samples harboring CNVs. These gains demonstrate that projecting clinical samples into evidence space effectively disentangles biological signal from technical artifacts.

The BRCA dataset remains challenging for all methods. Twelve of 24 clinical samples harbor germline gene-level *BRCA2* CNVs spanning the majority of the panel, creating

Table 4: Batch classification performance. All baselines are given the true $k = 2$; our method infers cluster assignments from the mixture model under H_1. Bold indicates best performance.

	BRCA		Thalassemia		LBX FFPE	
Model	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**
GMM	0.63	0.0	0.78	0.55	0.91	0.76
PCA-Ward	0.63	0.0	0.76	0.51	0.91	0.76
Hierarchical	0.63	0.0	0.78	0.55	0.72	0.13
Spectral	**0.71**	0.27	0.84	0.66	0.94	0.85
PCA-KMeans	0.63	0.0	0.85	0.68	0.94	0.86
Likelihood (Ours)	0.63	**0.34**	**0.90**	**0.80**	**1.0**	**1.0**

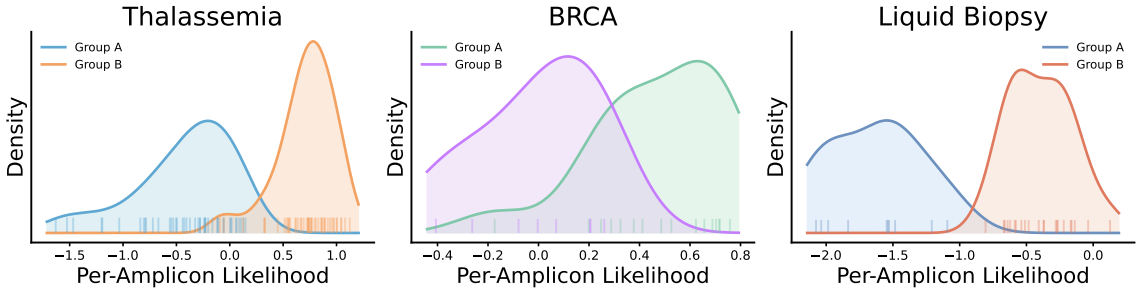

Figure 4: KDE of the per-amplicon likelihoods of each of the three panels. The separation is stronger in the LBX and thalassemia panels than in the BRCA panel, in part due to stronger heterogeneity (sample degradation and lab origin, as opposed to lot differences).

strong within-phenotype similarity that cuts across reagent lots. Most baselines predict the majority class ($\kappa = 0$). Spectral clustering achieves the highest raw accuracy (0.71) but with only moderate agreement ($\kappa = 0.27$). Our method matches the accuracy of most baseline methods (0.63) but achieves the highest κ (0.34), indicating that it captures meaningful batch structure rather than exploiting class imbalance. This result highlights both a strength, explicit generative modeling can partially disentangle biology from batch, and a limitation: when biological signal dominates the panel, the residual batch signal in evidence space is attenuated.

6.2. Mouse LFP Data

We evaluate cross-modality generalization on two mouse electrophysiology datasets using probabilistic principal component analysis (PPCA) as the generative model:

$$\begin{aligned} p(z) &= \mathcal{N}(0, I_L), \\ p(x \mid z, W, \sigma^2) &= \mathcal{N}(Wz, \sigma^2 I). \end{aligned} \tag{5}$$

Temporal correlation between observations violates the assumptions of standard dimensionality-selection criteria, making formal tests invalid. As a result, we set $L = 20$ components to approximately match the dimensionality used in the original studies (Mague et al., 2022; Gallagher et al., 2017). The model was fit on a training set of 10 mice (TST) and 5 mice (SP). Per-observation log-likelihoods in the held-out data were averaged within each mouse and condition, yielding two summary scores per mouse. This averaging step was necessary due to the substantial noise in the spectral features, particularly at low frequencies.

Table 5: Batch effect detection in mouse electrophysiology data. Significance probability is the fraction of bootstrap iterations yielding $p < 0.05$.

	Conditions Averaged		**Conditions Separated**	
Experiment	**Sig. Prob**	**Avg p**	**Sig. Prob**	**Avg p**
Social Preference	1.0	0.026	1.0	0.010
Tail Suspension	0.0	0.67	0.0	0.12
Both	1.0	0.011	1.0	0.004

Table 5 shows strong evidence of batch heterogeneity when combining mice from both experiments ($p = 0.011$, averaged over conditions). Within individual experiments, the tail suspension test (TST) shows no evidence of batch effects ($p = 0.67$), despite substantial biological differences between stressed and control mice—confirming that the method does not confuse biological signal with batch artifacts. In contrast, the social preference (SP) experiment yields significant batch effects ($p = 0.026$ averaged, $p = 0.010$ with conditions separated). These mice were recorded in two cohorts separated by several years, and the detected heterogeneity likely reflects instrumentation or environmental drift between cohorts. When conditions are separated within SP, significance is maintained but attenuated, consistent with the interpretation that splitting by condition reduces effective sample size without introducing additional cluster structure.

7. Discussion

Batch effects are a major source of error in clinical genomic diagnostics, and detecting them before CNV calling is critical. Existing methods are limited: unsupervised feature-space approaches cannot separate technical variation from biological signal, while supervised tests such as gPCA (Reese et al., 2013) require batch labels that are often unavailable. We address both problems by reducing each sample to a scalar Bayesian model evidence score and testing that one-dimensional representation for mixture structure. CNV-positive samples are captured by the generative model and retain high evidence, whereas artifacts

violate model assumptions and reduce it, separating biological from technical heterogeneity without labels.

The method has several limitations. First, some batch effects may not create detectable mixture structure in evidence space, and multiple heterogeneity sources can produce similar likelihoods. The BRCA dataset illustrates a related failure mode: when 50% of samples carry large gene-level CNVs, biological variance dominates technical signal, and only the Laplace mixture reaches significance ($p = 0.028$). Second, the choice of mixture distribution affects Type I error control: the Gaussian mixture is anti-conservative with outliers ($\hat{a} = 0.25$, Table 1) and yields borderline false positives in thalassemia ($p = 0.023$, $p = 0.032$). We therefore recommend the Laplace distribution as default. Finally, power is limited below $n \approx 15$, so the method should not be the sole quality control metric for very small batches.

References

Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nature reviews genetics*, 12(5):363–376, 2011.

Daniel Backenroth, Jason Homsy, Laura R Murillo, Joe Glessner, Edwin Lin, Martina Brueckner, Richard Lifton, Elizabeth Goldmuntz, Wendy K Chung, and Yufeng Shen. Canoes: detecting rare copy number variants from whole exome sequencing data. *Nucleic acids research*, 42(12):e97–e97, 2014.

Tommaso Becchi, Luca Beltrame, Laura Mannarino, Enrica Calura, Sergio Marchini, and Chiara Romualdi. A pan-cancer landscape of pathogenic somatic copy number variations. *Journal of Biomedical Informatics*, 147:104529, 2023.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Jan Budczies, Nicole Pfarr, Albrecht Stenzinger, Denise Treue, Volker Endris, Fakher Ismaeel, Nikola Bangemann, Jens-Uwe Blohmer, Manfred Dietel, Sibylle Loibl, et al. Ioncopy: a novel method for calling copy number alterations in amplicon sequencing data including significance assessment. *Oncotarget*, 7(11):13236, 2016.

Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. A test metric for assessing single-cell rna-seq batch correction. *Nature methods*, 16(1): 43–49, 2019.

Ben Calderhead and Mark Girolami. Estimating bayes factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*. Springer, 2020.

Leon Cohen. *Time-Frequency Analysis*. Prentice Hall Signal Processing Series. Prentice Hall, Englewood Cliffs, NJ, 1995. ISBN 978-0135945322.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997.

Menachem Fromer, Jennifer L Moran, Kimberly Chambert, Eric Banks, Sarah E Bergen, Douglas M Ruderfer, Robert E Handsaker, Steven A McCarroll, Michael C O'Donovan, Michael J Owen, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics*, 91(4): 597–607, 2012. doi: 10.1016/j.ajhg.2012.08.005.

Neil Gallagher, Kyle R Ulrich, Austin Talbot, Kafui Dzirasa, Lawrence Carin, and David E Carlson. Cross-spectral factor analysis. *Advances in neural information processing systems*, 30, 2017.

Stephanie C Hicks and Rafael A Irizarry. Quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome biology*, 16(1):117, 2015.

HM James Hung, Robert T O'Neill, Peter Bauer, and Karl Köhne. The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, pages 11–22, 1997.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Robert E Kass, Uri T Eden, and Emery N Brown. *Analysis of neural data*. Springer, 2014.

Niklas Krumm, Peter H Sudmant, Arthur Ko, Brian J O'Roak, Maika Malig, Bradley P Coe, Aaron R Quinlan, Deborah A Nickerson, Evan E Eichler, NHLBI Exome Sequencing Project, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Research*, 22(8):1525–1532, 2012.

Wiktor Kuśmirek, Agnieszka Szmurło, Marek Wiewiórka, Robert Nowak, and Tomasz Gambin. Comparison of knn and k-means optimization methods of reference set selection for improved cnv callers performance. *BMC bioinformatics*, 20(1):266, 2019.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

Jianying Li, Pierre R Bushel, Tzu-Ming Chu, and Russell D Wolfinger. Principal variance components analysis: estimating batch effects in microarray gene expression data. *Batch effects and noise in microarray experiments: sources and solutions*, pages 141–154, 2009.

Yungtai Lo, Nancy R Mendell, and Donald B Rubin. Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778, 2001.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Stephen D Mague, Austin Talbot, Cameron Blount, Kathryn K Walder-Christensen, Lara J Duffney, Elise Adamson, Alexandra L Bey, Nkemdilim Ndubuizu, Gwenaëlle E Thomas, Dalton N Hughes, et al. Brain-wide electrical dynamics encode individual appetitive social behavior. *Neuron*, 110(10):1728–1741, 2022.

Solaiappan Manimaran, Heather Marie Selby, Kwame Okrah, Claire Ruberman, Jeffrey T Leek, John Quackenbush, Benjamin Haibe-Kains, Hector Corrada Bravo, and W Evan Johnson. Batchqc: interactive software for evaluating sample and batch effects in genomic data. *Bioinformatics*, 32(24):3836–3838, 2016.

Vinzenz May, Leonard Koch, Björn Fischer-Zirnsak, Denise Horn, Petra Gehle, Uwe Kornak, Dieter Beule, and Manuel Holtgrewe. Clearcnv: Cnv calling from ngs panel data in the presence of ambiguity and noise. *Bioinformatics*, 38(16):3871–3876, 2022.

Geoffrey J McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):318–324, 1987.

Geoffrey J McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

Jonathan S Packer, Evan K Maxwell, Colm O'Dushlaine, Alexander E Lopez, Frederick E Dewey, Rostislav Chernomorsky, Aris Baras, John D Overton, Lukas Habegger, and Jeffrey G Reid. Clamms: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*, 32(1):133–135, 2016.

Michael K Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

Vincent Plagnol, James Curtis, Michael Epstein, Kin Y Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W Wood, Sophie Hambleton, Siobhan O Burns, Adrian J Thrasher, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754, 2012.

Raquel Prado and Mike West. *Time series: modeling, computation, and inference*. Chapman and Hall/CRC, 2010.

Sarah E Reese, Kellie J Archer, Terry M Therneau, Elizabeth J Atkinson, Celine M Vachon, Mariza De Andrade, Jean-Pierre A Kocher, and Jeanette E Eckel-Passow. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics*, 29(22):2877–2883, 2013.

Natalya Risinskaya, Maria Gladysheva, Abdulpatakh Abdulpatakhov, Yulia Chabaeva, Valeriya Surimova, Olga Aleshina, Anna Yushkova, Olga Dubova, Nikolay Kapranov, Irina Galtseva, et al. Dna copy number alterations and copy neutral loss of heterozygosity in adult ph-negative acute b-lymphoblastic leukemia: Focus on the genes involved. *International Journal of Molecular Sciences*, 24(24):17602, 2023.

Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Austin Talbot, David Dunson, Kafui Dzirasa, and David Carlson. Estimating a brain network predictive of stress and genotype with supervised autoencoders. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(4):912–936, 2023.

Austin Talbot, Alex Kotlar, Lavanya Rishishiwar, and Yue Ke. Classifying copy number variations using state space modeling of targeted sequencing data: A case study in thalassemia. *arXiv preprint arXiv:2504.10338*, 2025.

Austin Talbot, Alex Kotlar, Lavanya Rishishwar, Andrew Conley, Mengyao Zhao, Nachen Yang, Michael Liu, Zhaohui Wang, Sean Polvino, and Yue Ke. Bayescnv: A bayesian hierarchical model for sensitive and specific copy number estimation in cell free dna. *Diagnostics*, 16(2):280, 2026.

Jordy van Enkhuizen, Arpi Minassian, and Jared W. Young. Further evidence for ClockΔ19 mice as a model for bipolar disorder mania using cross-species tests of exploration and sensorimotor gating. *Behavioural Brain Research*, 249:44–54, jul 2013. ISSN 01664328. doi: 10.1016/j.bbr.2013.04.023.

Peter D Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

Dina Zhu, Linlin Xu, Yanxia Zhang, Guanxia Liang, Xiaofeng Wei, Liyan Li, Wangjie Jin, and Xuan Shang. Investigation of the mechanism of copy number variations involving the α-globin gene cluster on chromosome 16: two case reports and literature review. *Molecular Genetics and Genomics*, 298(1):131–141, 2023.

Appendix A. CNV Statistical Models and Evidence Computation

Our heterogeneity-detection method requires a scalar evidence score for each sample. We employ two CNV models suited to different panel designs, enabling evaluation of evidence-based heterogeneity detection under distinct modeling assumptions. The first model treats amplicons within each gene as conditionally independent replicates and targets gene-level CNVs, appropriate for panels with sparse targets spread across the genome. The second model induces spatial dependence among neighboring amplicons via a state-space formulation, appropriate for densely tiled panels covering a contiguous region (in this work, BRCA1/2 and thalassemia assays). The two models also differ in inference procedures and in how evidence is estimated.

A.1. Hierarchical Bayesian Model for Gene-Level CNV Detection

In the gene-level model, amplicons targeting the same gene are treated as noisy measurements of an underlying gene-level lCNR. Let $g \in \{1, \dots, G\}$ index genes, and let y_{ig} denote the lCNR for the ith amplicon targeting gene g. The latent gene-level lCNR is μ_g. We place hierarchical priors on gene-level means and gene-specific scales to share information across genes, and we use a SoftLaplace observation model, a smooth heavy-tailed alternative to the Laplace distribution, to improve robustness to outlier amplicons.

The generative model is

$$\begin{aligned} \mu_0 &\sim \mathcal{N}(0, 10), \\ \sigma^2 &\sim \text{InvGamma}(\alpha_\sigma, \beta_\sigma), \\ \tau_0^2 &\sim \text{InvGamma}(\alpha_{\tau_0}, \beta_{\tau_0}), \\ \mu_g \mid \mu_0, \sigma^2 &\sim \mathcal{N}(\mu_0, \sigma^2), \\ z_g &\sim \text{InvGamma}(\alpha_\tau, \beta_\tau), \\ y_{ig} \mid \mu_g, z_g &\sim \text{SoftLaplace}(\mu_g, \tau_0 z_g), \end{aligned} \tag{6}$$

where $\tau_0 z_g$ is a gene-specific scale parameter. The hyperparameters $\{\alpha_\sigma, \beta_\sigma, \alpha_{\tau_0}, \beta_{\tau_0}, \alpha_\tau, \beta_\tau\}$ are fixed a priori to values used in previous work.

We infer the model parameters $\theta = \{\mu_0, \sigma^2, \tau_0^2, \{\mu_g\}_{g=1}^G, \{z_g\}_{g=1}^G\}$ using Hamiltonian Monte Carlo (HMC) (Betancourt, 2017). To estimate the log marginal likelihood $\log p(y)$, we use thermodynamic integration (Calderhead and Girolami, 2009). Let $\tau \in [0, 1]$ denote the inverse temperature, and define the power posterior

$$p_\tau(\theta \mid y) \propto p(y \mid \theta)^\tau \, p(\theta). \tag{7}$$

Thermodynamic integration uses the identity

$$\ell = \log p(y) = \int_0^1 \mathbb{E}_{\theta \sim p_\tau(\theta|y)} \big[\log p(y \mid \theta) \big] \, d\tau. \tag{8}$$

We approximate the integral in (8) by evaluating the expectation at a discrete ladder of inverse temperatures $\{\tau_t\}_{t=1}^T$ spanning $[0, 1]$ and applying the trapezoidal rule. We use 100 temperatures with linear spacing.

A.2. State-Space Model for Spatially Correlated CNV Detection

When amplicons target a contiguous genomic region, lCNRs exhibit spatial correlation, and we instead use a state-space model (Chopin et al., 2020). Let x_j denote the latent true lCNR at amplicon j. Spatial dependence is induced via a heavy-tailed innovation model that favors local smoothness while permitting rare large transitions corresponding to CNV boundaries:

$$p(x_{j+1} \mid x_j) = (1-p)\,\mathcal{N}(x_j, \epsilon) + p\,\mathcal{N}(x_j, \sigma^2), \tag{9}$$

where $p \ll 1$ is the probability of a large transition, ϵ controls local smoothness, and $\sigma^2 \gg \epsilon$ permits large jumps. The initial state has a weakly informative prior reflecting the assumption of copy neutrality:

$$x_0 \sim \mathcal{N}(0, \tau_0^2). \tag{10}$$

Observations follow a Laplace likelihood for robustness to outliers:

$$y_j \mid x_j \sim \text{Laplace}(x_j, b). \tag{11}$$

The complete model is

$$\begin{aligned} x_0 &\sim \mathcal{N}(0, \tau_0^2), \\ x_{j+1} \mid x_j &\sim (1-p)\,\mathcal{N}(x_j, \epsilon) + p\,\mathcal{N}(x_j, \sigma^2), \\ y_j \mid x_j &\sim \text{Laplace}(x_j, b), \end{aligned} \tag{12}$$

with fixed transition and observation parameters $\theta = \{p, \epsilon, \sigma^2, \tau_0^2, b\}$ chosen a priori for each assay using domain knowledge of expected CNV sizes and noise characteristics (Talbot et al., 2025). Full Bayesian inference over θ is computationally prohibitive for our use case, and we therefore treat θ as fixed when computing evidence.

Inference over latent states $\{x_j\}_{j=0}^{J}$ proceeds via sequential Monte Carlo using an auxiliary particle filter (Pitt and Shephard, 1999). A standard byproduct of particle filtering is an unbiased estimator of the marginal likelihood. Let M denote the number of particles and let $w_j^{(m)}$ denote the (unnormalized) importance weight for particle m at position j. The particle filter yields the estimator

$$\ell = \log \widehat{p}(y \mid \theta) = \sum_{j=1}^{J} \log \left(\frac{1}{M} \sum_{m=1}^{M} w_j^{(m)} \right). \tag{13}$$

Appendix B. Data Processing

B.1. Targeted Amplicon Sequencing

All targeted amplicon sequencing datasets were processed through a common bioinformatics pipeline prior to feature extraction. Raw paired-end reads were aligned to the GRCh37/hg19 reference genome using BWA-MEM (Li, 2013). To improve alignment accuracy in regions of structural complexity, the initial alignments were refined through local realignment using the Smith–Waterman algorithm (Smith et al., 1981) in conjunction with a proprietary consensus-aware realignment procedure. Paired-end reads originating from the same DNA

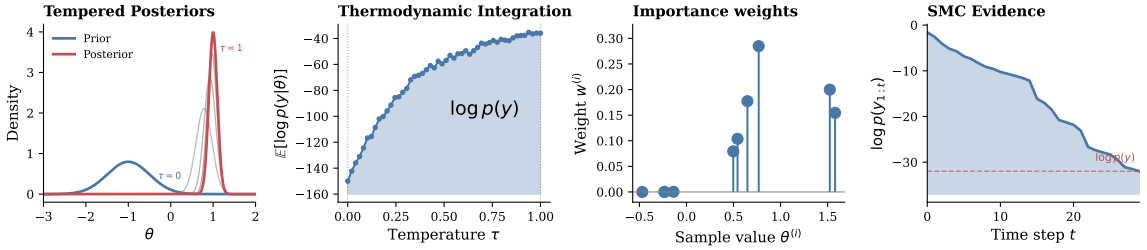

Figure 5: Overview of evidence computation. Left: a sequence of power posteriors indexed by inverse temperature $\tau \in [0, 1]$ used for thermodynamic integration. Second from left: thermodynamic integration estimates $\log p(y)$ by numerically integrating $\mathbb{E}_{p_\tau(\theta|y)}[\log p(y \mid \theta)]$ over τ (Eq. (8)). Middle right: particle-filter weights at a representative amplicon position; highly peaked weight distributions indicate that few particles explain the observation well. Right: particle-filter evidence accumulates across amplicons via the product form in Eq. (13).

fragment were then assembled into consensus reads, with base quality scores from both mates used as weights during assembly; this step reduces sequencing noise and improves per-base accuracy, particularly in low-complexity or repetitive regions.

Following consensus assembly, reads were filtered to retain only those mapping uniquely to intended amplicon positions. Reads mapping to pseudogene loci or to off-target locations (e.g., primer-dimer artifacts or non-specific amplification products) were excluded. For copy number variant (CNV) analysis, per-amplicon read depth was computed for each sample independently. Where amplicons overlap genomically, read counts for each amplicon were stored as separate entries to preserve the full resolution of the depth signal across the targeted region.

B.1.1. Liquid Biopsy Data

The first dataset was generated using a targeted liquid biopsy panel (referred to as LBX in the main text) designed to detect common somatic mutations associated with carcinogenesis. The panel comprised 173 amplicons spanning a broad set of oncologically relevant genes, of which 5 genes were designated as CNV targets for copy number assessment. Samples consisted of $n = 32$ formalin-fixed, paraffin-embedded (FFPE) specimens. FFPE preservation introduces characteristic DNA fragmentation and chemical modification artifacts that degrade library quality in a sample-dependent manner; accordingly, FFPE input quality was treated as a batch factor in the main analysis. Ten samples exhibited pronounced fragmentation consistent with poor FFPE preservation, while the remaining 22 samples were of comparatively higher quality, providing a natural source of technical variation for evaluating the robustness of the proposed quality control framework.

B.1.2. Thalassemia Data

The thalassemia dataset was assembled from two independent clinical laboratories to enable evaluation of inter-laboratory reproducibility as a batch factor. The first cohort comprised

49 de-identified remnant samples (36 positive, 13 negative) obtained from a clinical testing laboratory that had processed the samples for diagnostic purposes and would otherwise have discarded them. The second cohort comprised 45 samples obtained under analogous conditions from a separate laboratory. In total, the dataset contained $n = 94$ samples, with Laboratory A contributing $n = 49$ and Laboratory B contributing $n = 45$. Per FDA guidance on the research use of remnant clinical specimens, patient consent was not required for either cohort.

All samples were sequenced using a 131-amplicon panel covering alpha- and beta-thalassemia regions, as well as relevant pseudogene loci and internal control regions. Of the 131 amplicons, 11 are gap amplicons designed to span large deletion breakpoints; these were excluded from standard depth-based analyses but retained for structural variant detection. Sequencing was performed on an Illumina MiSeq™ platform targeting an average depth of approximately 2,800 paired-end reads per amplicon, providing sufficient coverage for reliable CNV calling across the panel. Ground-truth copy number status for alpha-thalassemia variants was established using an orthogonal, clinically validated workflow combining multiplex ligation-dependent probe amplification (MLPA) with reflex to GAP-PCR for the detection of alpha-globin deletions and duplications.

B.1.3. BRCA Data

The BRCA dataset was assembled to evaluate the quality control framework on a clinically important gene panel with a well-characterized set of reference samples. The panel comprised 283 amplicons tiling the coding regions and splice junctions of the *BRCA1* and *BRCA2* genes, of which 8 were control amplicons drawn from unrelated chromosomal loci to serve as normalization references for CNV calling.

Four mutation-positive DNA reference samples were obtained from the Coriell Institute for Medical Research, each harboring a distinct, well-characterized structural variant: NA18949 (BRCA1 exons 14–15 deletion), NA14626 (BRCA1 exon 12 duplication), NA03330 (BRCA2 whole-gene duplication), and NA02718 (BRCA2 whole-gene deletion). Four mutation-negative cell line samples (NA12878, NA19240, NA24385, and NA24143) were included as negative controls. To evaluate performance on chemically degraded DNA representative of clinical FFPE specimens, Horizon Discovery's Mimix™ Quantitative Multiplex, fcDNA (Moderate) Reference Standard was included as an additional sample. Libraries were prepared and sequenced on an Illumina MiSeq™ platform. All cell line samples were collected under informed consent.

The dataset comprised $n = 24$ samples in total. Reagent lot was treated as a batch factor, with lot A used for $n = 16$ samples and lot B used for $n = 8$ samples, enabling assessment of inter-lot technical variability.

B.2. Mouse Electrophysiology

Local field potential (LFP) recordings from two behavioral paradigms were processed through a shared pipeline to extract spectral features suitable for latent factor modeling. The two experiments — the Tail Suspension Test (TST) and the Social Preference (SP) assay — together constitute the electrophysiology component of the benchmark, with the experimental paradigm treated as a batch factor in the joint analysis.

B.2.1. Tail Suspension Test

LFP recordings for the Tail Suspension Test were obtained from mice of two genetic backgrounds: wild-type (WT, $n = 16$) and Clock-Δ19 ($n = 12$), for a total of 28 mice. The Clock-Δ19 genotype carries a dominant-negative mutation in the *Clock* gene and has been used as a translational model of bipolar disorder (van Enkhuizen et al., 2013). Each animal underwent a structured 20-minute recording session spanning three sequential behavioral contexts: 5 minutes in the home cage (low arousal, non-stressful baseline), 5 minutes in an open field arena (moderate arousal, mildly stressful), and 10 minutes suspended by the tail (high arousal, acutely stressful). This design provides a graded stress continuum within a single recording session.

LFPs were recorded simultaneously from 11 biologically relevant brain regions using a 32-electrode array, with multiple electrodes implanted per region to accommodate post-hoc exclusion of faulty or misplaced electrodes. To obtain a single representative signal per brain region, electrode signals within each region were averaged prior to feature extraction. This averaging strategy reduces measurement variance arising from electrode-to-electrode variability while preserving the biologically meaningful regional signal; the brain region represents the finest spatial scale at which signals can be reliably compared across animals.

The continuous LFP time series were segmented into non-overlapping 1-second windows. Windows containing saturated or artifactual signals — most commonly attributable to movement artifacts during the tail suspension phase — were identified and excluded prior to spectral analysis. Although continuous time-frequency methods exist for characterizing non-stationary neural signals (Prado and West, 2010), the behavioral states of interest evolve on timescales substantially longer than the underlying neural dynamics; discrete windowing therefore yields sharper spectral estimates (Cohen, 1995) that are better suited to the subsequent factor modeling framework.

Spectral power features were extracted from each 1-second window using Welch's method (Welch, 1967), as recommended for LFP analysis in the neuroscience literature (Kass et al., 2014). Power was estimated in 1 Hz bands spanning 1–56 Hz, yielding 56 spectral features per brain region per window. The upper frequency limit of 56 Hz was chosen to exclude the 60 Hz line noise artifact introduced by the recording environment, consistent with prior work demonstrating that the preponderance of behaviorally relevant LFP information is concentrated in lower frequency bands. Spectral features were computed prior to model fitting rather than within the modeling framework, as incorporating spectral estimation into the model would substantially increase the number of free parameters without a commensurate gain in interpretability.

B.2.2. Social Preference Dataset

LFP recordings for the Social Preference (SP) experiment were processed using the identical pipeline described above for the TST dataset, including the same windowing, artifact removal, and Welch's method spectral feature extraction procedure. The SP cohort comprised mice recorded across two sequential experimental cohorts, which represent a potential source of batch variation noted in the main text.

In the SP assay, mice were placed in a two-chamber apparatus and allowed to explore freely for 10 minutes while LFPs and animal position were recorded simultaneously. One

chamber contained a novel conspecific mouse (social stimulus), while the other contained an inanimate object (non-social control stimulus). The paradigm was designed to characterize the neural correlates of social approach behavior and to support causal manipulation of identified dynamics via optogenetic stimulation. LFPs were recorded from 8 brain regions, 7 of which overlap with the 11 regions recorded in the TST experiment, enabling cross-paradigm comparison of spectral features in shared circuits. A complete description of the experimental design, surgical procedures, and optogenetic stimulation protocol is provided in Mague et al. (2022).

B.3. Data Availability

Processed feature matrices and metadata for all targeted amplicon sequencing datasets (LBX, Thalassemia, and BRCA) are available in the GitHub repository associated with this work. The TST electrophysiology dataset is publicly available for download at https://research.repository.duke.edu/concern/datasets/zc77sr31x?locale=en. Access to the Social Preference electrophysiology dataset may be requested by contacting the principal investigator, Kafui Dzirasa.

B.3.1. Electrophysiology Data

We also evaluate the sensitivity of the results to the dataset used to train the PCA model. Tables 6 and 7 show similar trends.

Table 6: Batch effect detection in the electrophysiology data using tail suspension data as training set.

	Conditions Averaged		Conditions Separated	
Experiment	**Significance Prob**	**Avg P-Value**	**Significance Prob**	**Avg P-Value**
Social Preference	1.0	0.043	1.0	0.043
Tail Suspension	0.75	0.047	0.0	0.56
Both	1.0	0.013	1.0	0.005

Table 7: Batch effect detection in the electrophysiology data using social data as training set.

	Conditions Averaged		Conditions Separated	
Experiment	**Significance Prob**	**Avg P-Value**	**Significance Prob**	**Avg P-Value**
Social Preference	1.0	0.014	0.63	0.048
Tail Suspension	0.0	0.67	0.0	0.12
Both	1.0	0.013	1.0	0.005

Appendix C. Supplementary Synthetic Data Analyses

C.1. Comparison with the LMR P-Values

We also compared our parametric bootstrap likelihood ratio test to the Lo–Mendell–Rubin (LMR) adjusted likelihood ratio test (Lo et al., 2001), which provides an asymptotic approximation to the mixture LRT p-value without requiring simulation. Table 8 reports calibration results for both the no-outlier and outlier-included conditions across all five component distributions. The LMR test is substantially anti-conservative across nearly all settings. Under the null (no outlier), four of five distributions yield $\hat{a} < 1$ with Clopper–Pearson bounds well above the nominal 0.05 level: Gaussian ($\hat{a} < 0.48$, $CP = 0.19$), Student-t ($\hat{a} = 0.71$, $CP = 0.09$), Laplace ($\hat{a} = 0.36$, $CP = 0.37$), and generalized normal ($\hat{a} = 0.64$, $CP = 0.13$). Only hypersecant is conservative under the null ($\hat{a} = 18.5$), but this conservatism is so extreme as to render the test insensitive. When an outlier is introduced, the picture worsens: Gaussian becomes severely anti-conservative ($\hat{a} = 0.12$, $CP = 0.74$), and all distributions except hypersecant exhibit CP bounds far exceeding 0.05. In contrast, the parametric bootstrap achieves CP bounds at or below 0.07 for four of five distributions under the null (Table 1), and the recommended Laplace kernel maintains CP = 0.06 under the alternative.

Table 8: Effect of evidence-space component distribution on calibration with LMR p-values. **No Outlier**: all samples drawn from the null generative process. **Outlier Included**: one sample artificially shifted by $-2\,\sigma$. Beta$(a, 1)$ shape $\hat{a}$ with 95% credible interval; values near 1 indicate calibration, < 1 anti-conservativeness, > 1 conservativeness. Clopper–Pearson (CP) column reports the one-sided 95% upper bound on $\mathbb{P}(\text{reject} \mid H_0)$ at the $\alpha = 0.05$ operating point; values ≤ 0.05 confirm conservativeness. **Bold** rows indicate the best-performing distribution in each condition.

	No Outlier		**Outlier Included**	
Distribution	$\hat{a}$ **[95% CI]**	**CP bound**	$\hat{a}$ **[95% CI]**	**CP bound**
Gaussian	0.48 (0.45, 0.51)	0.19	0.12 (0.11, 0.12)	0.74
Hypersecant	18.5 (17.4, 19.7)	0.001	2.2 (2.07, 2.33)	0.004
Student-t	0.71 (0.67, 0.76)	0.09	0.22 (0.20, 0.23)	0.36
Laplace	0.36 (0.34, 0.39)	0.37	**0.49** (**0.46**, **0.52**)	**0.23**
Gennorm	0.64 (0.61, 0.69)	0.13	0.17 (0.16, 0.18)	0.48