Critique of Julia Salzman's Statistical Framework for Genomics

Bottom Line: Four Fundamental Problems

Julia Salzman's SPLASH/OASIS framework, while mathematically sophisticated, suffers from four fundamental conceptual problems:

1. **Treats technical replicates (sequencing reads) as the unit of statistical randomness** rather than biological replicates (individuals), conflating technical and biological variation
2. **Creates contingency tables as an artifact of her own method**, then critiques existing statistical tests for not handling these self-created tables well, rather than addressing problems that naturally arise in genomics
3. **Claims genomic tests use Pearson's χ^2 when most do not** - modern genomic methods predominantly use likelihood ratio tests, negative binomial models (DESeq2, edgeR), or linear models (limma), not χ^2 tests
4. **Mischaracterizes the K vs N problem** - conflates the legitimate multiple testing problem (testing K features requires correction) with the validity of individual statistical tests, which are unaffected by having K features when each is tested separately

---

Detailed Elaboration

1. The Technical vs Biological Replication Problem

What Salzman's model assumes:

From the OASIS paper (PNAS 2024), the null hypothesis is defined as:

> "Conditional on the column totals $n_1, n_2, ..., n_m$, each column of the contingency table is Multinomial(n_j, p), drawn independently for all $j \in [m]$, for some common vector of (unknown) row probabilities p."

Translation:

- Each individual j contributes n_j sequencing reads
- Each read is treated as an independent draw from a multinomial distribution
- **The statistical replication is at the read level, not the individual level**

The fundamental problem:

The framework treats sequencing reads from a single individual as independent statistical replicates, when in reality:

- **Sequencing reads from one person** = technical replicates of that person's biology
- **Different people** = biological replicates from a population

- **The biological question** is: "Do different individuals have different k-mer distributions?"
- **The statistical model** answers: "If all reads came from the same distribution, how likely is this pattern?"

These are fundamentally different questions.

Quote from Salzman (OASIS paper):

> "OASIS constructs a test statistic which is linear in the normalized data matrix, providing closed-form P-value bounds through classical concentration inequalities."

The concentration inequalities (Hoeffding's inequality) assume independent samples. But reads from the same individual are **not** independent samples from a population - they are dependent samples from that individual's genome/transcriptome.

What's missing:

No clear accounting of:

- Within-individual variation (technical + biological noise in that person)
- Between-individual variation (the actual biological signal)
- The hierarchical structure: reads nested within individuals

Implications:

P-values are answering: "What if all these reads (across all people) came from one multinomial distribution?"

But biologists care about: "What if these people came from populations with different k-mer frequencies?"

---

2. Self-Created Problems Presented as Solutions

The construction:

Salzman's SPLASH method:

1. Takes raw sequencing reads from N individuals
2. Identifies "anchor" k-mers (constant sequences)
3. For each anchor, identifies "target" k-mers that follow it
4. **Constructs a contingency table** where:
   - Rows = different target k-mers (K targets)
   - Columns = individuals (N samples)
   - Entries = counts of each target in each sample

The claimed problem:

From the OASIS paper:

> "Existing statistical tests are insufficient however, as none are simultaneously computationally efficient and statistically valid for a finite number of observations."

> "Today's genomics workflows typically require alignment to a reference sequence, which limits discovery."

The critique:

This is a **self-created problem**:

1. **Traditional genomic analyses don't naturally produce these contingency tables** - Salzman chooses to represent her data this way
2. **She then claims existing methods are inadequate for handling them**
3. This is circular reasoning: "I've formatted the data in a way that's hard to analyze, therefore we need a new method"

Alternative perspective:

Most genomic analyses handle the N samples, K features problem differently:

- Differential expression: Model each gene separately with appropriate error models
- GWAS: Test each variant with appropriate regression models
- They don't create giant sparse contingency tables and then complain about χ^2 tests

Quote from Salzman:

> "I tried to step back and ask if they're needed. I couldn't convince myself they were [reference genomes]." (Stanford interview, 2024)

But the contingency table framework isn't a natural representation of genomic data either - it's a choice that creates the very problems she then solves.

---

3. Mischaracterization of What Genomic Tests Actually Use

Salzman's claim:

From OASIS paper:

> "There is a rich literature that addresses testing for row and column independence in contingency tables beginning with the work of Pearson, who designed the widely used χ^2 test in the early 1900s."

> "Despite the prominence of Pearson's test, it suffers from multiple statistical drawbacks which limit its utility for scientific inference."

The reality:

Most modern genomic analyses do NOT use Pearson's χ^2 test:

Analysis Type	Actual Methods Used	Not χ^2
Differential gene expression (RNA-seq)	DESeq2, edgeR (negative binomial GLMs), limma-voom	✓
GWAS (genetic associations)	Logistic regression, score tests, linear mixed models	✓
eQTL analysis	Linear regression, mixed models	✓
Variant calling	Bayesian models, likelihood ratios	✓
ChIP-seq peaks	Negative binomial, Poisson models	✓

What genomics actually struggles with:

1. **Multiple testing**: Testing K features requires controlling family-wise error rate or FDR
2. **Appropriate error models**: Accounting for overdispersion, zero-inflation
3. **Batch effects and confounding**: Not solved by any test statistic
4. **Biological vs technical variation**: Requires hierarchical models

None of these are about χ^2 tests being invalid.

From Salzman's DEEPEST paper (PNAS 2019):

> "Current algorithms do not sufficiently identify false-positive fusions arising during library preparation, sequencing, and alignment."

This is a problem of:

- **Bioinformatics pipeline errors** (alignment artifacts)
- **Biological confounding** (germline polymorphisms misidentified as fusions)

Not a problem that χ^2 tests or their validity can solve.

---

4. The K vs N Problem Mischaracterized

Salzman's framing:

From OASIS:

> "However, in modern tables of interest, this is often the case; the biological tables which motivated this test's design have many rows (tens or hundreds) relative to the total number of observations per column (similarly in the tens or hundreds), violating use guidelines."

She implies:

- Having K features (rows) and N samples (columns) with K >> N creates invalid p-values
- This is a problem specific to finite samples that asymptotic theory doesn't handle

The reality:

K vs N affects multiple testing burden, not individual test validity.

For each feature k in 1:K:

- You test that feature across N samples
- The test is valid regardless of how many other features exist
- **The total number K doesn't enter the test statistic or p-value for feature k**

The actual problem:

If you test K features at $\alpha = 0.05$:

- You expect $K \times 0.05$ false positives by chance
- **Solution: Adjust for multiple testing** (Bonferroni, FDR control)
- This is well-established and not what Salzman addresses

Her contingency table setup:

When she creates a table with K rows (targets) and N columns (samples):

- She's testing ONE hypothesis: "Are row and column independent?"
- This is a single test, not K tests
- The problem is: **her test has K-1 degrees of freedom**, and with sparse data the χ^2 approximation fails

But this is a known problem with known solutions:

- Fisher's exact test for sparse 2×2 tables
- Permutation tests for larger sparse tables
- Aggregate rare categories

The issue is not "K >> N makes tests invalid" - it's that she's chosen a representation (contingency table) that's hard to test when sparse.

The Deeper Issue: Statistics vs Machine Learning Mindset

The Paradigm Shift in High-Dimensional Biology

The classical statistical paradigm:

- Measure **one or a few outcomes** per experimental unit
- N individuals, each contributing 1-10 measurements
- Focus on: parameter estimation, hypothesis testing, p-values, confidence intervals
- Model is interpretable, parameters have meaning

The modern genomics reality:

- Measure **thousands of outcomes** per experimental unit
- N individuals, each contributing 10,000-100,000 measurements
- This is fundamentally a **machine learning** problem, not a classical statistics problem

Key insight from your observation:

> "When statistical techniques developed in a context of analyzing one outcome measured on each analysis unit are used to analyze datasets in which each analysis unit contributes thousands of outcome measures... the analyst wants to keep thinking like a statistician when in fact they have entered the machine learning world."

Salzman's Framework is Wedded to Classical Statistical Thinking

Evidence from her work:

1. **From OASIS paper:** "OASIS provides P-value bounds that 1) are valid for a finite number of observations, 2) have a closed-form expression..."
2. **From Stanford interview:** "With math, I could... People told her, 'If you do statistics, you can do anything.'"
3. **Core claim:** Statistical validity (Type I error control, finite-sample p-values) is the cornerstone of her framework

The problem:

When you fit a model to thousands of features:

- The model is effectively a **black box prediction machine**
- Individual p-values lose interpretability
- You're in the domain of: prediction accuracy, cross-validation, regularization
- Not: "Is this one p-value exactly 0.049 or 0.051?"

Quote from OASIS:

> "Existing statistical tests fall short, however; none provide robust, computationally efficient inference and control type I error."

But in practice:

With K = 20,000 genes tested:

- Even with perfect Type I error control at $\alpha = 0.05$ per test
- And perfect FDR control at 5%

- You still have ~1,000 "significant" results to interpret
- **The statistical validity of individual p-values is not the bottleneck**

What Actually Matters in Modern Genomics

Prediction and validation:

- Does your method predict held-out samples accurately?
- Does it replicate in independent cohorts?
- Does it identify biologically interpretable patterns?

Not:

- "Is the p-value from a finite-sample bound exactly calibrated?"

Machine learning approaches that have succeeded:

- Deep learning on genomic sequences (predicts function without p-values)
- Random forests for variant prioritization
- Dimensionality reduction (PCA, UMAP) for visualization
- Clustering algorithms for cell type identification

None of these provide p-values. They work anyway.

The Irony

Salzman criticizes genomics for:

> "Downstream statistical analysis is therefore done on a signal convolved with an unknown noise source. Thus, references and annotations are scientifically and statistically problematic for reliable, sensitive, and [inference]." (NOMAD bioRxiv 2022)

But her own framework:

- Creates K × N contingency tables
- Treats reads as independent draws (ignoring individual-level structure)
- Performs data splitting for optimization (introducing its own noise)
- Aggregates across multiple tables with ad-hoc normalization
- **Also produces "a signal convolved with an unknown noise source"**

The difference:

- Reference-based methods acknowledge they're making approximations
- Salzman's methods claim exact finite-sample statistical validity

- **But both are fitting complex models to high-dimensional data**
- At that point, the distinction between "statistically valid" and "useful approximation" becomes philosophical

The Statistics Prison

By insisting on:

- Exact p-values
- Finite-sample validity
- Closed-form bounds
- Type I error control

Salzman constrains herself to:

- Simple test statistics (linear forms)
- Concentration inequalities (loose bounds)
- Heavy multiple testing burden (thousands of anchors tested)

Meanwhile, machine learning approaches:

- Use all the data jointly
- Learn complex nonlinear patterns
- Validate on prediction accuracy
- **Don't worry about p-values**

When P-Values Make Sense

Classical statistics works when:

- You have a clear null hypothesis (e.g., "drug has no effect")
- You measure **one or a few pre-specified** outcomes
- False positives have consequences (e.g., approving ineffective drug)
- Sample size is modest (hard to collect more data)

In this regime:

- Exact Type I error control matters
- P-values are interpretable
- Statistical validity is crucial

Modern genomics is different:

- Null hypothesis is vague ("no regulation difference anywhere in genome?")
- You measure **everything** (all genes, all variants)
- You're exploring, not confirming pre-specified hypotheses
- Data is plentiful (can sequence more samples)

In this regime:

- Prediction accuracy matters more than exact p-values
- Replication matters more than finite-sample validity
- Biological interpretability matters more than Type I error control

---

Summary: Where Salzman Goes Wrong

The four problems:

1. **Technical vs biological replication:** Treats reads as independent samples when they're dependent measurements from individuals
2. **Self-created problems:** Constructs contingency tables, then complains existing tests don't handle them well
3. **Mischaracterization:** Claims genomics uses χ^2 tests when it mostly doesn't
4. **K vs N confusion:** Conflates multiple testing burden with individual test validity

The deeper issue:

Salzman is a statistician trying to solve a machine learning problem with statistical tools. She's obsessed with p-value validity in a domain where p-values have limited utility.

Quote that reveals the mindset (Stanford interview):

> "With math, I could [study everything]... People told her, 'If you do statistics, you can do anything.'"

The reality:

With thousands of measurements per sample:

- You've left the domain where classical statistics rules apply
- Model interpretability is limited (too many parameters)
- P-values are one of many tools, not the gold standard
- Prediction, replication, and biological validation matter more than exact Type I error control

Salzman's contribution:

Her methods may be useful for specific applications (e.g., reference-free variant detection). But the framework is oversold as solving fundamental problems that either:

- Don't exist as stated (χ^2 tests aren't widely used)
- Are self-created (contingency tables are her choice)
- Are misunderstood (K >> N doesn't invalidate tests)
- Miss the point (in high dimensions, exact p-values matter less than she thinks)

References

1. Baharav, T.Z., Tse, D., and Salzman, J. (2024). "OASIS: An interpretable, finite-sample valid alternative to Pearson's X^2 for scientific discovery." *PNAS* 121(15).
2. Chaung, K., Baharav, T.Z., Henderson, G., et al. (2023). "SPLASH: a statistical, reference-free genomic algorithm unifies biological discovery." *Cell* 186(25):5440-5456.
3. Dehghannasiri, R., Freeman, D.E., Jordanski, M., et al. (2019). "Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers." *PNAS* 116(31):15524-15533.
4. Stanford Medicine interview (2024): "How one researcher flipped traditional genomics analysis on its head."

Appendix: What Genomics Actually Uses

For differential expression (RNA-seq):

- **DESeq2**: Negative binomial generalized linear model with empirical Bayes shrinkage
- **edgeR**: Negative binomial models with quasi-likelihood
- **limma-voom**: Linear models with precision weights

None use χ^2 tests.

For genetic association (GWAS):

- Logistic regression (case-control)
- Linear regression (quantitative traits)
- Score tests (computationally efficient)
- Linear mixed models (population structure)

None use χ^2 tests on contingency tables.

For variant calling:

- Bayesian models (GATK)
- Likelihood ratios
- Quality score recalibration

None use χ^2 tests.

The pattern: Modern genomics uses sophisticated models appropriate for count data, genetic structure, and biological variability. Salzman's focus on χ^2 tests attacks a straw man.