1 Statistical Analysis: Data Structure and Categorical Variables

1.1 Gene Selection

We selected 8 genes from the extracellular matrix (ECM) and collagen pathway, identified by Phillips et al. (2013) as enriched in the all-trans-retinoic acid (ATRA) signaling network activated by resistance training.

Table 1: Target Genes and Their Functions

Probe ID	**Symbol**	**Gene Name**
211980_at	COL4A1	Collagen type IV alpha 1 chain
203477_at	COL15A1	Collagen type XV alpha 1 chain
204114_at	NID2	Nidogen 2 (ECM glycoprotein)
212013_at	PXDN	Peroxidasin (ECM crosslinking)
204115_at	GNG11	G protein subunit gamma 11
205656_at	PCDH17	Protocadherin 17
204008_at	DNAL4	Dynein axonemal light chain 4
218429_s_at	SHFL	Shiftless antiviral inhibitor

COL4A1, COL15A1, NID2, and PXDN are directly involved in extracellular matrix remodeling—a key molecular response to resistance exercise. This selection allows us to test whether ECM-related genes show consistent training-induced changes.

1.2 Paired vs Unpaired Comparison

To demonstrate the importance of accounting for data structure, we compared two analytical approaches on the same gene expression data:

- **Paired Analysis**: Model `GeneDiff ~ 1` tests whether the mean within-subject change $\neq 0$. Each subject serves as their own control, removing between-subject variability.
- **Unpaired Analysis**: Model `Gene ~ Pre_Post` compares Post vs Pre as independent groups, ignoring the paired structure.

1.2.1 Results

The unpaired analysis showed standard errors approximately 1.12× larger on average than the paired analysis. However, both approaches detected all 8 genes as significant ($p < 0.05$), indicating the training effect was strong enough to overcome the variance inflation.

Table 2: Paired vs Unpaired Analysis Comparison

Gene	**Estimate**	**SE (Paired)**	**SE (Unpaired)**	**SE Ratio**	**Sig.**
203477_at	1852.98	357.19	407.44	1.14	* / *
204008_at	28.11	6.13	5.74	0.94	* / *
204114_at	181.84	29.74	35.62	1.20	* / *
204115_at	539.05	88.49	122.00	1.38	* / *
205656_at	98.81	17.98	21.53	1.20	* / *
211980_at	1588.24	291.71	308.95	1.06	* / *
212013_at	146.27	29.26	33.47	1.14	* / *
218429_s_at	192.07	37.08	40.11	1.08	* / *

1.2.2 Pre-Post Correlation Analysis

The SE ratio between unpaired and paired analyses follows approximately:

$$\text{SE ratio} \approx \frac{1}{\sqrt{1-r}}$$

where r is the Pre-Post correlation within subjects. We computed empirical correlations to verify this relationship.

Table 3: Pre-Post Correlation and SE Ratio Comparison

Gene	**Pre-Post** r	**Predicted SE Ratio**	**Observed SE Ratio**
203477_at	0.197	1.116	1.141
204008_at	−0.231	0.901	0.937
204114_at	0.307	1.201	1.198
204115_at	0.413	1.305	1.379
205656_at	0.309	1.203	1.198
211980_at	0.093	1.050	1.059
212013_at	−0.020	0.990	1.008
218429_s_at	0.065	1.034	1.068

The predicted and observed SE ratios show strong agreement, validating the theoretical relationship. The modest SE ratios (averaging 1.12) reflect relatively low Pre-Post correlations in this dataset, suggesting that between-subject variability is not dramatically larger than within-subject variability for these genes.

1.2.3 Negative Pre-Post Correlations

Two genes (204008_at and 212013_at) exhibited negative Pre-Post correlations, which is unexpected—baseline and follow-up measures are typically positively correlated. When $r < 0$, the predicted SE ratio falls below 1, meaning the unpaired analysis would actually have *smaller* standard errors than the paired analysis. This counterintuitive pattern could reflect:

- **Compensatory regulation**: High baseline expression leading to stronger downregulation post-training
- **Ceiling/floor effects**: Constraints on expression range producing negative associations
- **Measurement noise**: Particularly in low-expression genes where technical variability dominates

The biological meaning of negative Pre-Post correlation warrants further investigation, as it may indicate distinct regulatory mechanisms for these genes.

1.2.4 Interpretation

The training effect in this dataset is sufficiently strong that both approaches yield identical conclusions regarding statistical significance. Nevertheless, the paired design remains theoretically preferable because it explicitly models within-subject correlation. In datasets with weaker effects or higher between-subject variability, the difference in statistical power would be more pronounced.

1.3 Categorical Variable Analysis

We examined whether categorical groupings influenced gene expression changes.

1.3.1 Age Group Analysis

Subjects were divided into High and Low age groups based on the median age. A t-test compared gene expression changes between groups.

Result: 0 of 8 genes showed significantly different responses between age groups ($p < 0.05$). Age does not appear to moderate the training effect—both age groups respond similarly.

1.3.2 Responder Analysis

Subjects were classified as High or Low responders based on median expression change for gene 211980_at (COL4A1).

Result: 6 of 8 genes showed significantly different expression changes between responder groups. This indicates correlated gene responses: individuals who respond strongly on the reference gene also tend to show stronger responses on other genes. Note that this analysis is partly circular since responder status was defined using one of the target genes.

1.4 Summary

Table 4: Analysis Summary

Analysis	Key Finding	Implication
Paired vs Unpaired	SE inflation: 1.12×; Power difference: 0 genes	Strong effects detectable either way; paired still theoretically preferable
Pre-Post Correlation	Two genes show negative r; predicted and observed SE ratios match well	Validates theoretical framework; negative correlations warrant biological investigation
Age Group Effect	0 genes differ by age group	Age does not moderate training response
Responder Effect	6 genes differ by responder status	Correlated individual response patterns across genes