



プロテオミクスデータにおける log2FC空間の構造的特徴

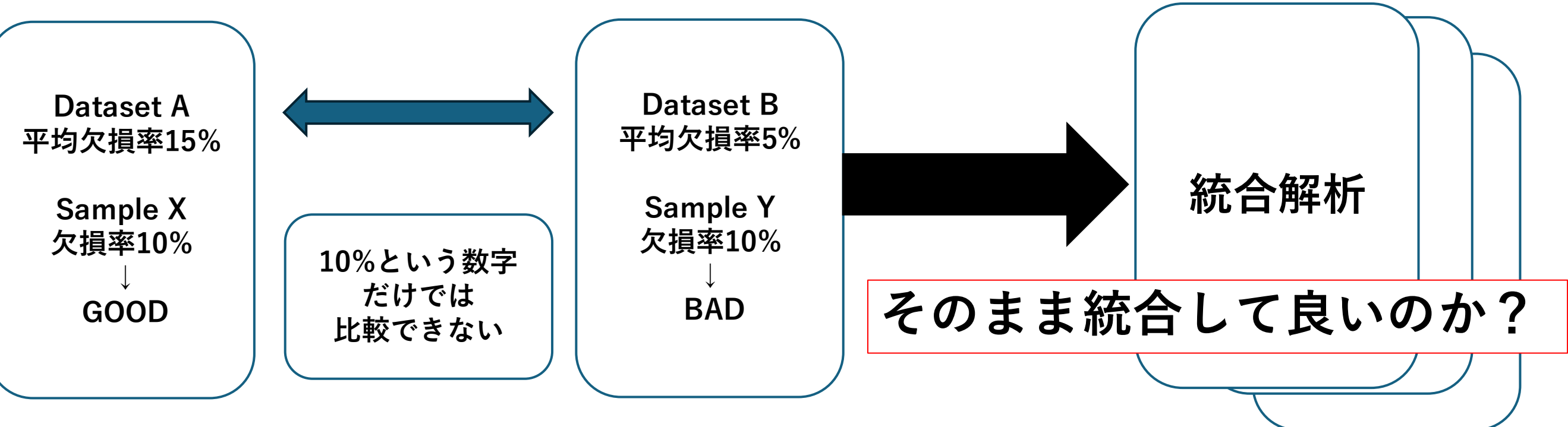
Structural Characteristics of the log2 Fold Change Space in Proteomics Data

西崎愛花¹, 荒木令江², 河野信^{1,3}

1. 北里大学大学院未来工学研究科, 2. 熊本大学大学院生命科学研究部(医), 3. 情報・システム研究機構 ライフサイエンス統合データベース部門

1 研究背景・目的

公開プロテオームデータ環境における統合解析の可能性拡張



jPOSTをはじめとする ProteomeXchange リポジトリから様々な実験で得られたプロテオームデータがオープンデータとして公開されている。これらの公開データを横断的に統合解析することで、新たな知見の創出が期待される。一方、Tandem Mass Tag (TMT) を用いたプロテオミクスでは、定量値は参照サンプルに対する相対値であるlog2 fold change (log2FC) として扱われることが多く、実験間のスケール差を補正できる反面、参照サンプルの変動や欠損構造の影響を受け、分布シフトや外れ値、符号反転などの構造的特徴を有する。そのため、公開データを統合的に利用するには、まずlog2FC空間の構造的特徴を把握することが重要である。本研究では、実験条件単位でこれらの特徴を定量化する多指標スコアリング手法を開発し、公開データに適用してlog2FC空間の構造を解析した。

2 研究方法

スコアリング手法

プロテオームlog2FCデータセット

	Sample 1	Sample 2	Sample 3	Sample 4	...
Protein A	0.5	1.2		0.2	...
Protein B	0.1	0.2	0.4		...
Protein C	-0.2	-1.3	1.1	1.5	...
Protein D	1.4	0.8			...
...	...	...	...	...	...

①タンパク質 フィルタリング

統計的に不安定なタンパク質
除去

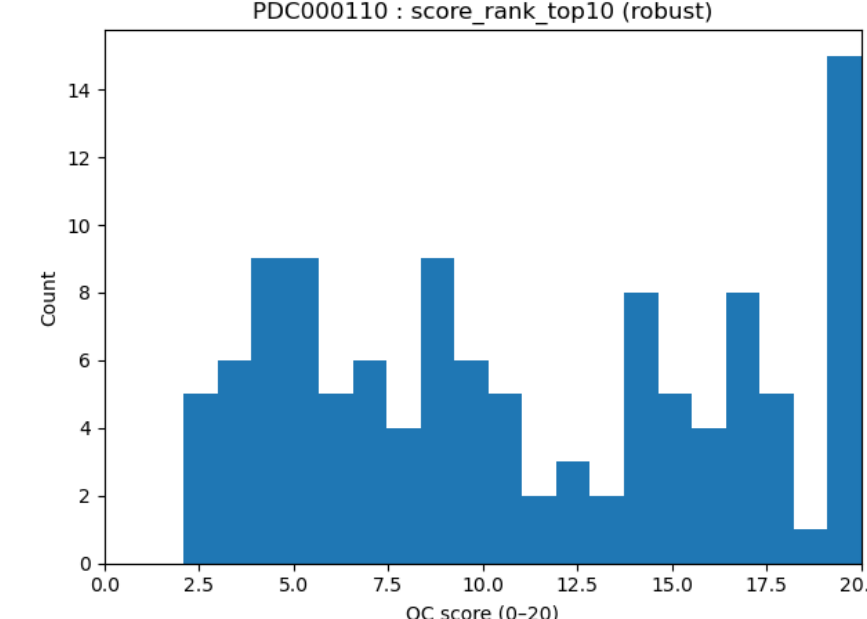
欠損率70%以上
CV $\geq$ 1.5
SD上位1%
MAD上位1%
IQR上位1%
これら条件に1つでも
当てはまるタンパク質

各実験条件（列）を様々な指標から評価する

②QC指標

	Sample 1	Sample 2	Sample 3	Sample 4	...
Protein A	0.5	1.2		0.2	...
Protein B	0.1	0.2	0.4		...
Protein E	-0.1		1.1	0.6	...
Protein G	1.0	0.8	1.2	0.9	...
...	...	...	...	...	...

- ・欠損率系指標
タンパク質フィルタリング前後の欠損率
フィルタリング前後の欠損改善率
- ・外れ値系指標
MAD/パーセンタイル外れ値
- ・タンパク質系指標
共通タンパク質の挙動検知
- ・中央値シフト系指標
実験条件の中央値ずれの検知



スコア分布の例

③QCスコア

0-20のスケール (0に近いほど良く、20に近いほど悪い)
各指標の中央値がスコア10となるようにMADベースでロバストに標準化

①タンパク質フィルタリング

統計的に不安定なタンパク質を除去するための前処理として欠損率および CV (Coefficient of Variation: 変動係数)、SD (Standard Deviation: 標準偏差)、MAD (Median Absolute Deviation: 中央値絶対偏差)、IQR (Interquartile Range: 四分位範囲) に基づき、タンパク質方向の品質フィルタリングを行った。フィルタリング前後の指標変化についてのスパマン相関散布図による可視化、また、metascape [https://metascape.org] を用いてエンリッチメント解析を行った。

②QC指標計算

QC指標 (欠損率系)

『missing_before』: タンパク質フィルタリング前の欠損率 (ももとの欠損率)
『missing_after』: タンパク質フィルタリング後の欠損率
『missing_improvement』: タンパク質フィルタリング前後の欠損率の改善を評価

QC指標 (中央値シフト系)

『score_median_shift』: 中央値が全体からどれくらいずれているか (絶対値)
『score_median_shift_directional』: 中央値が全体からどれくらいずれているか (符号考慮)

③QC指標からQCスコアへ変換

実験条件方向の品質指標として、欠損率およびタンパク質フィルタリング後の改善量、符号多数派からの逆符号頻度、パーセンタイルおよび MAD に基づく外れ値頻度、順位構造の一貫性、中央値シフトを算出した。各指標は中央値および MAD を用いてロバストに標準化し、**中央値10を基準とする0-20スケールの連続QCスコア**へ変換した。0に近いほどデータセット内で良い (例: 欠損率が低い、タンパク質一致率が高い) 20に近いほどデータセット内で悪い (例: 欠損率が高い、タンパク質一致率が低い)

QC指標 (外れ値系)

『outlier_5_95』: 5,95パーセンタイル外れ値になった回数
『outlier_mad』: MADの外れ値になった回数

QC指標 (タンパク質系)

『score_rank_top10』: 発現上位10%のタンパク質群の、他列との一致率
『score_rank_bottom10』: 発現下位10%のタンパク質群の、他列との一致率
『score_rank_consistency』: top10とbottom10の一致度を平均した総合指標
『score_sign_opposite』: 多数派の符号 (80%以上) と逆の振る舞いをしている回数

④妥当性検証

人工データを用いた指標反応性検証

プロテオームlog2FCを模した多変量正規分布に基づく合成データを作成し、欠損増加、分散拡大、分布シフトなどの技術的劣化を3段階 (Severity = 0.5, 1.0, 1.5) に分けて付与した。ロジスティック回帰を用いて指標反応性を確認した。

実データ検証用データセット

CPTAC [Edwards, NJ., *et al. Journal of Proteome Research* **14**, 2707-2713 (2015)] の49プロジェクト、5260実験条件のデータを用いた。

3 結果

1. 人工データ検証

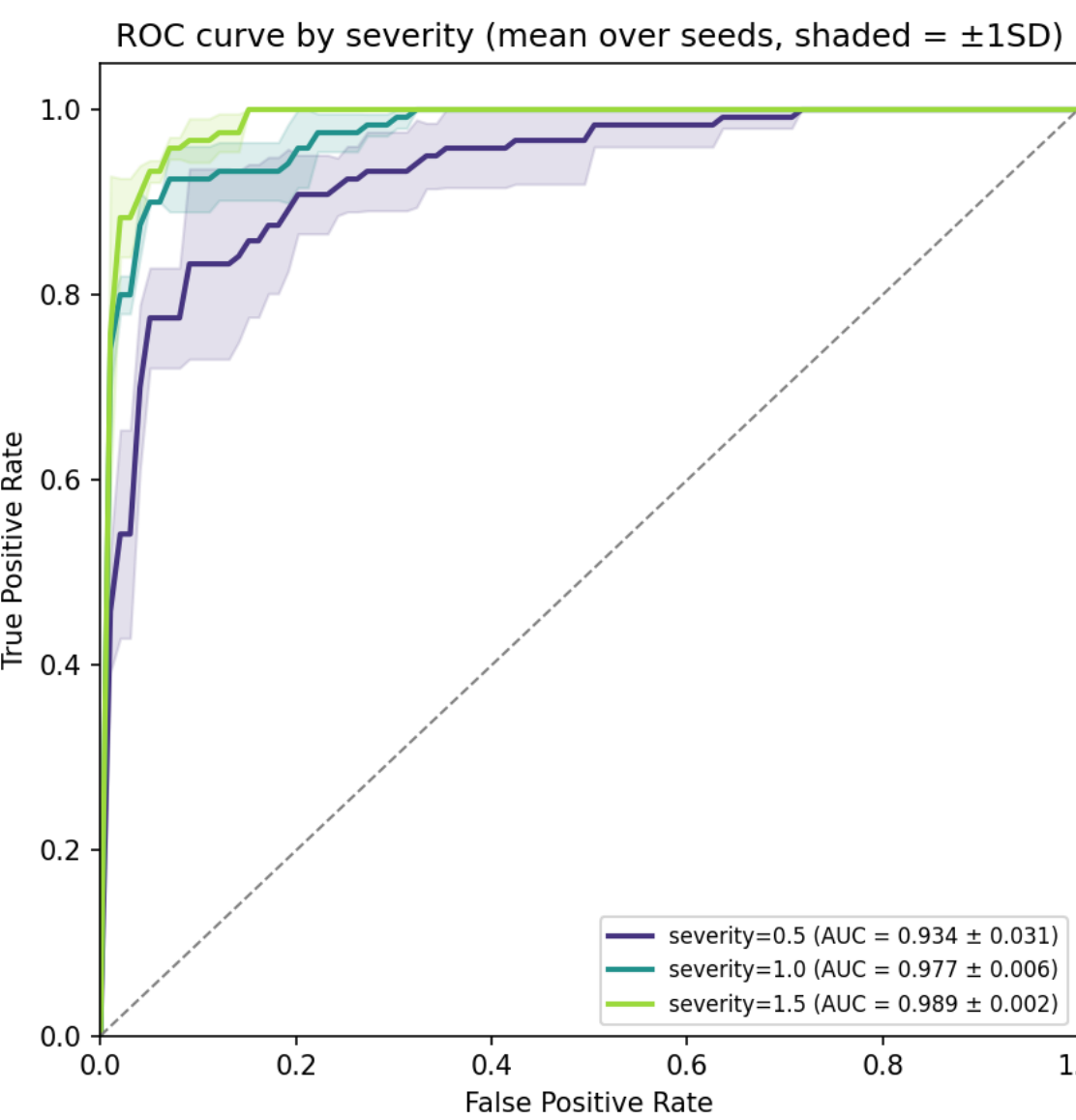


図1 ロジスティック回帰によるROC曲線

2. 実データを用いた主成分分析

多くの指標はまとまって同じ方向に分布していたが、**median_shiftは独立した軸を形成する傾向**が見られた。この指標は他のQC指標とは異なるデータの性質を反映している可能性がある。

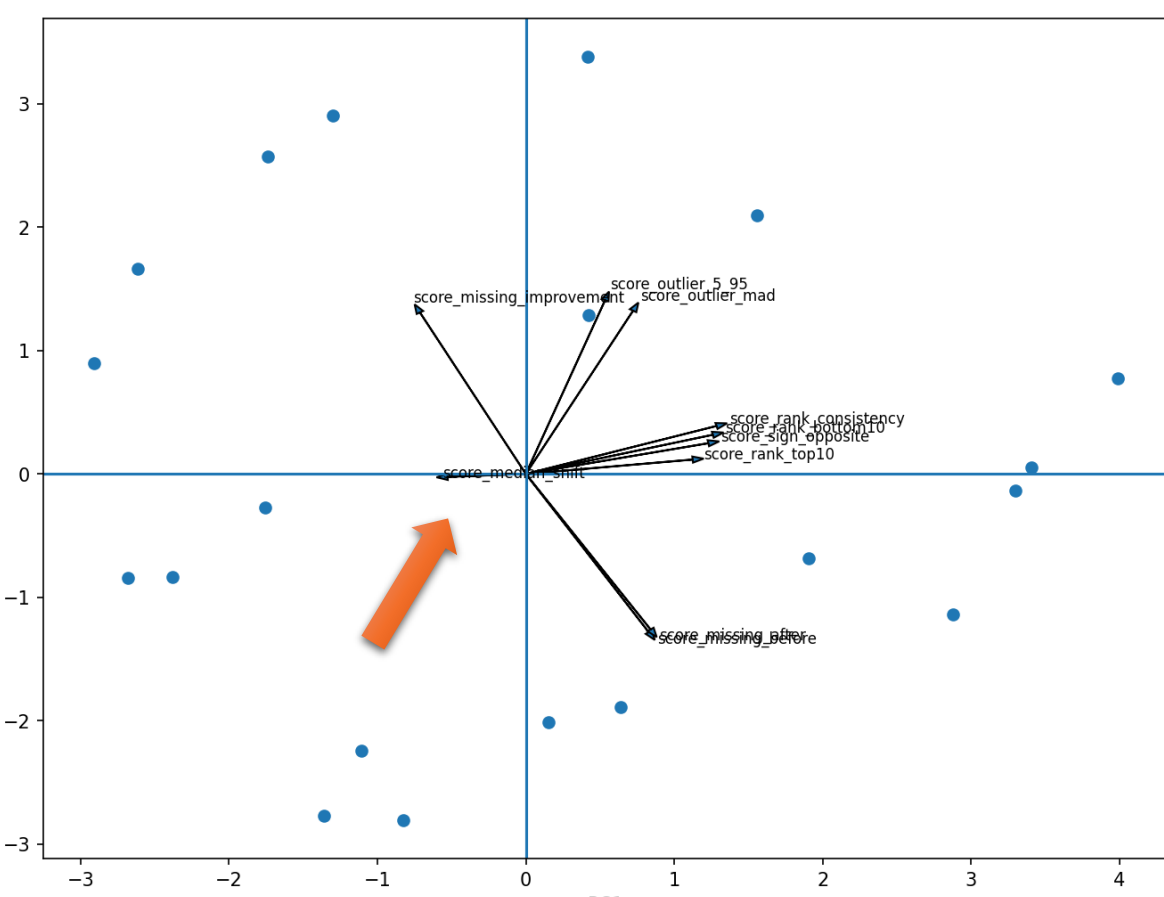


図2 QCスコアの主成分分析

3. タンパク質フィルタリング検証

タンパク質フィルタリング前後で、median_shiftが最も影響を受けていることがわかった。タンパク質フィルタリングは統計的なばらつきを対象としていて、中央値の値は確認していない。そのため現時点でmedian_shiftにこのような変化が起こる理由は現時点では不明である。

フィルタリングで除去されたタンパク質についてエンリッチメント解析を行った結果、細胞外マトリックス(ECM)/上皮/発生・分化関連などの機能が繰り返し出るが、全データセットで共通する強い機能分類は確認されなかった。

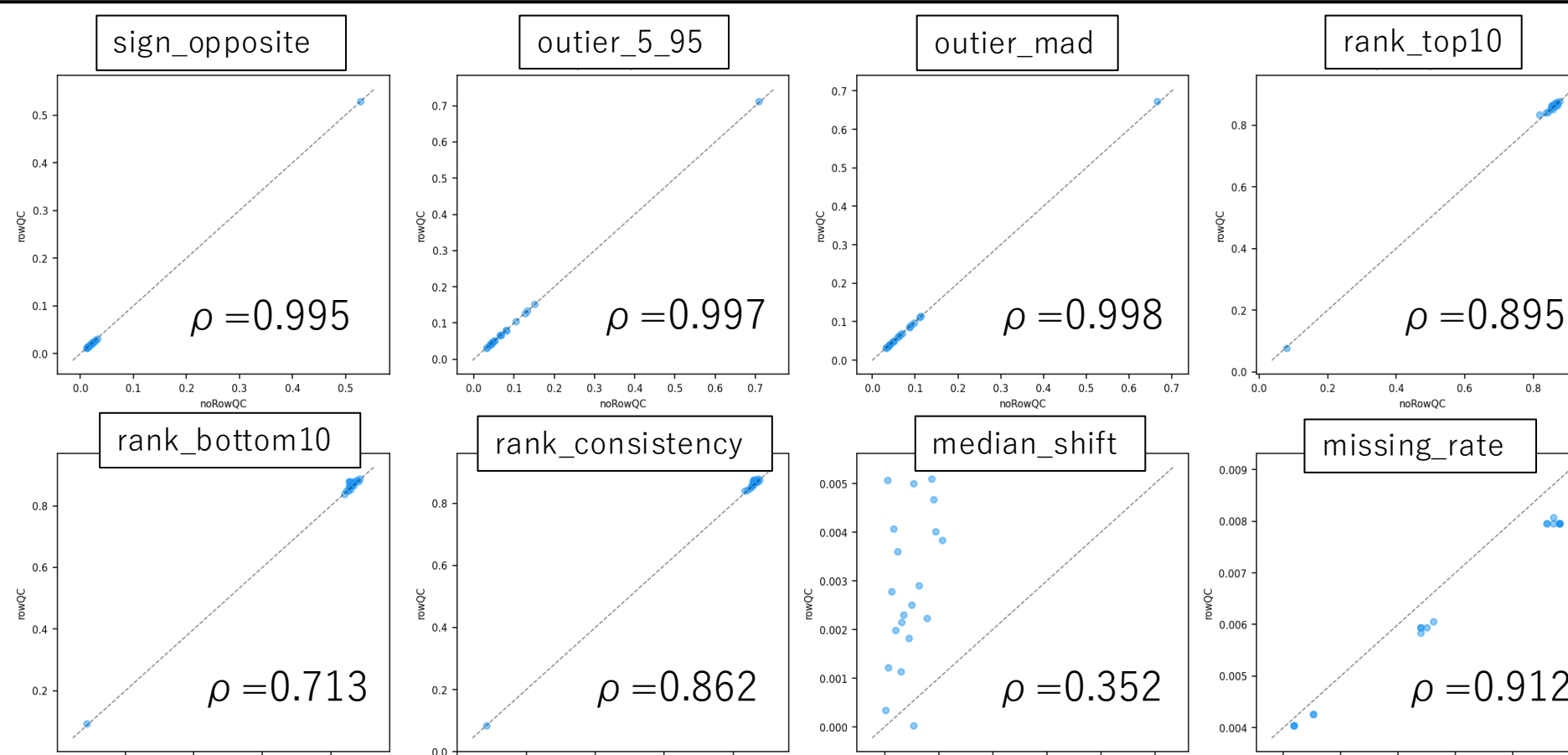


図3 フィルタリング前後の各指標のスパマン相関散布図

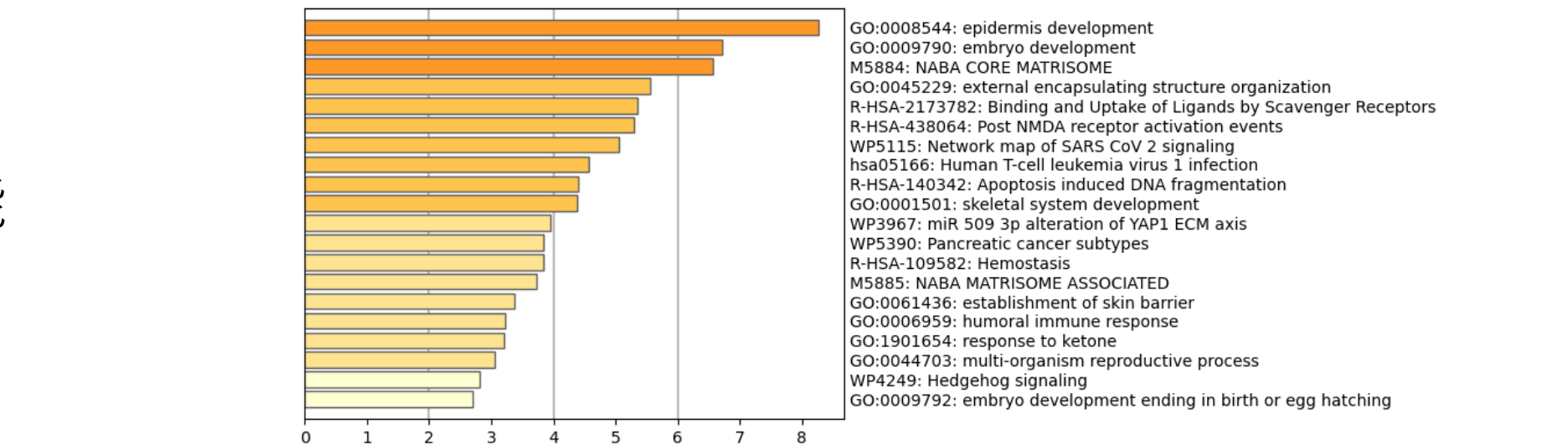


図4 除去されたタンパク質のエンリッチメント解析結果

4 まとめ・展望

公開プロテオームデータの統合解析に向けて、実験条件ごとのlog2FC空間の構造的特徴を多面的に評価するQCスコアリングフレームワークを構築した。本手法は人工データに対して高い識別性能を示し、データセット間の品質や構造の違いを比較可能な形で評価できることが示された。また、各指標はそれぞれ異なるlog2FC空間の構造的特徴を反映していることが示唆された。今後は各指標が表す構造的特徴の生物学的・統計学的意味づけを進めるとともに、公開プロテオームデータの統合解析における品質評価基盤としての活用を目指す。