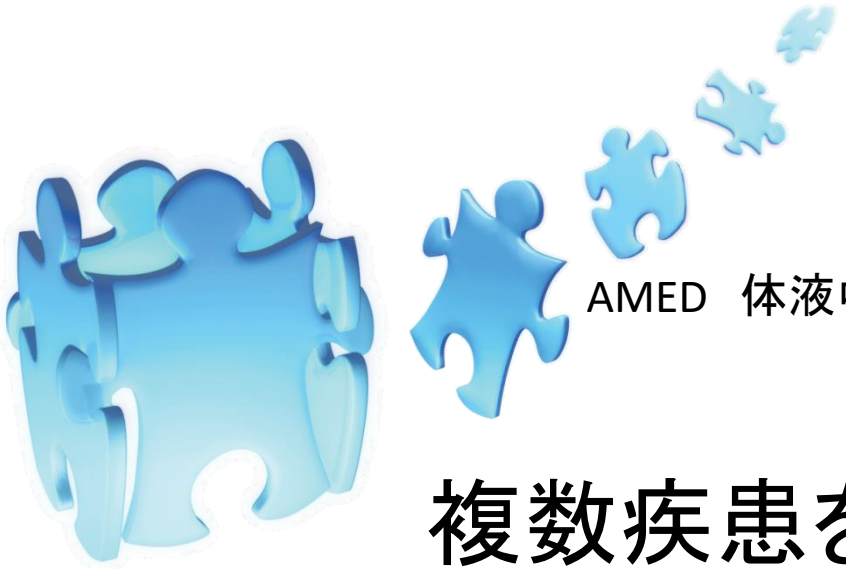


第6回 JMACシンポジウム

AMED　体液中マイクロRNA測定技術基盤開発プロジェクト
成果報告会

複数疾患を対象にした組み合わせマーカー探索アルゴリズムの開発

2019/01/24

(株)ダイナコム
小川、青木、中村、三浦、藤宮

会社紹介

- 株式会社ダイナコム
  - 設立1995
- 本社:千葉市
- 神戸オフィス:神戸市

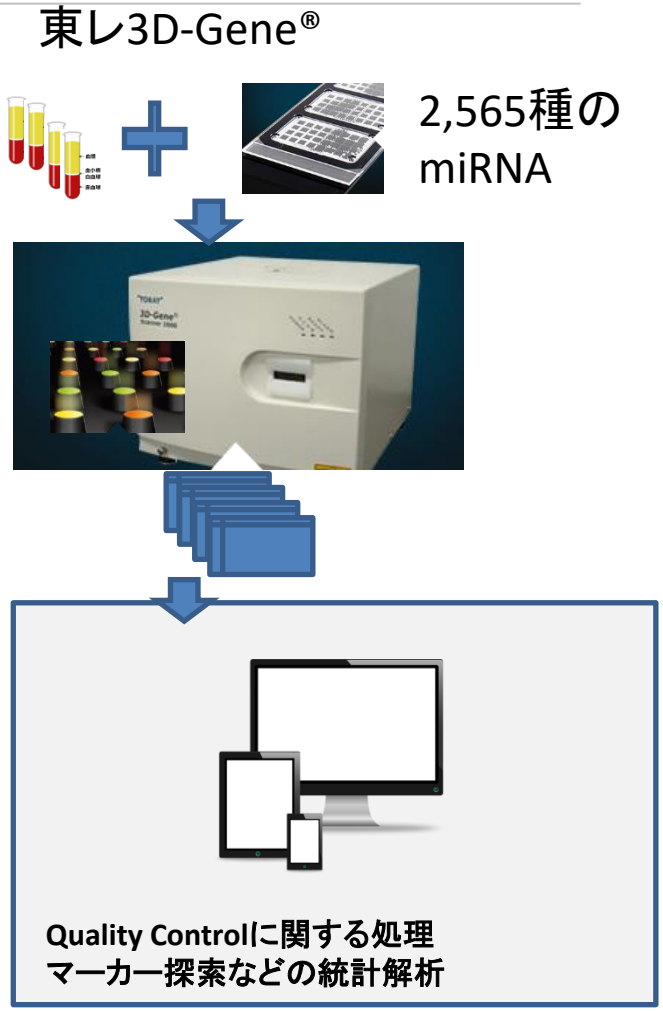
- 主な業務
  - バイオインフォマティクス関連ソフトウェア開発
  - 健康医療データ解析
  - 学会などの会員情報管理、論文投稿管理システムなど

目次

1. miRNAデータの概要と統計解析の担当部分
2. 健常人に対するがん種ごとのmiRNAマーカー探索
3. 疾患横断的な組み合わせ判別マーカー探索
4. miRNA情報収集ツール

miRNAデータの概要と統計解析の担当部分

- 東レ社製 DNAチップ(3D-Gene®)のデータ前処理のR言語スクリプト開発
- miRNAマーカー探索
  - 参画企業の多様な検査プラットフォームに対応できるように複数の候補を得る
  - できるだけ少ないマーカーで判別モデルを構築
  - 複数の疾患の横断的に診断できるような組み合わせマーカー候補を得る
- 得られたマーカーに関する情報収集などのサポートツール
- その他の解析
  - 標準血清に対するmiRNAデータ解析
  - 付随研究でのマーカー絞込支援
  - 予後予測などの時間軸を加味した解析



測定されているmiRNAの概要

- 収集経緯:
  - 国立長寿医療研究センター
  - 横浜みのるクリニック: 男女35歳～80歳代まで年齢のボランティア検体
  - 国立がんセンターバイオバンク:　腫瘍マーカーまたは感染症検査の血清検体
- 対象miRNAの数:
  - miRBase release 21対応　配列数にして2,565種

データクレンジングおよびDNAチップデータのQC状況

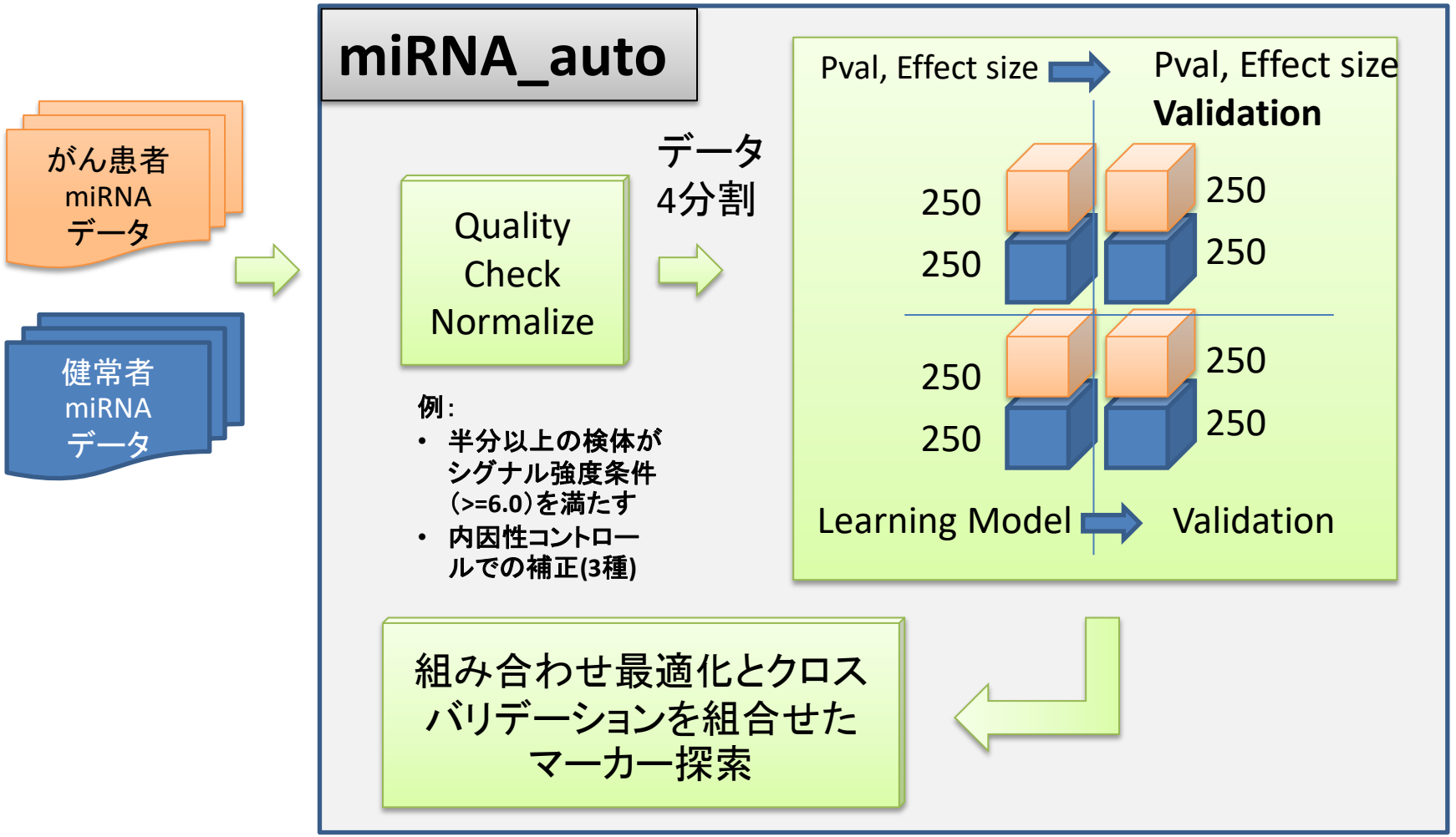
- データクレンジング
  - がんの重複・再発/転移患者データの除外
  - 読み取り不良チップデータの除外
- DNAチップデータのQC
  - 低レベルシグナルのmiRNA除外（ケース、コントロールのいずれかの半数以上で対数化シグナルが基準値以上となるmiRNAを選別）

内在性コントロールとして使用しているmiRNAのセット

検体間でのmiRNAの比較解析を正確に行うために利用

1. 3種のmiRNAを内在性コントロール
2. コントロールの平均値でmiRNAの値を補正

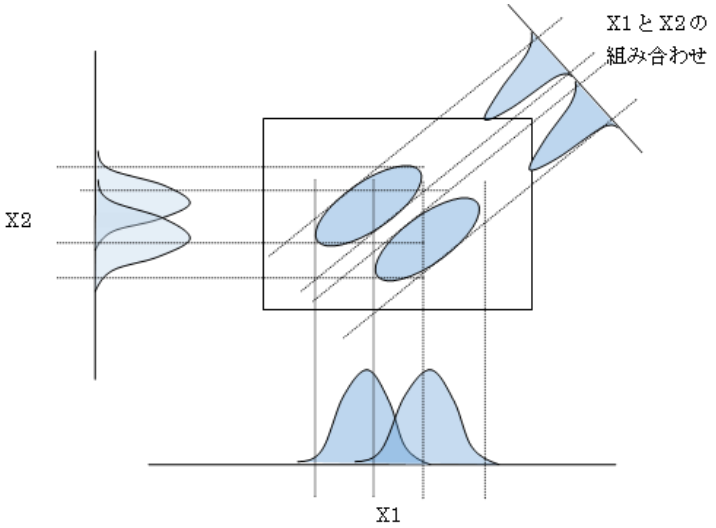
健常人に対するがん種ごとのmiRNAマーカー探索



線形判別＋組み合わせ最適化法の意義

要求条件に（最も合致する）最適な解を求める

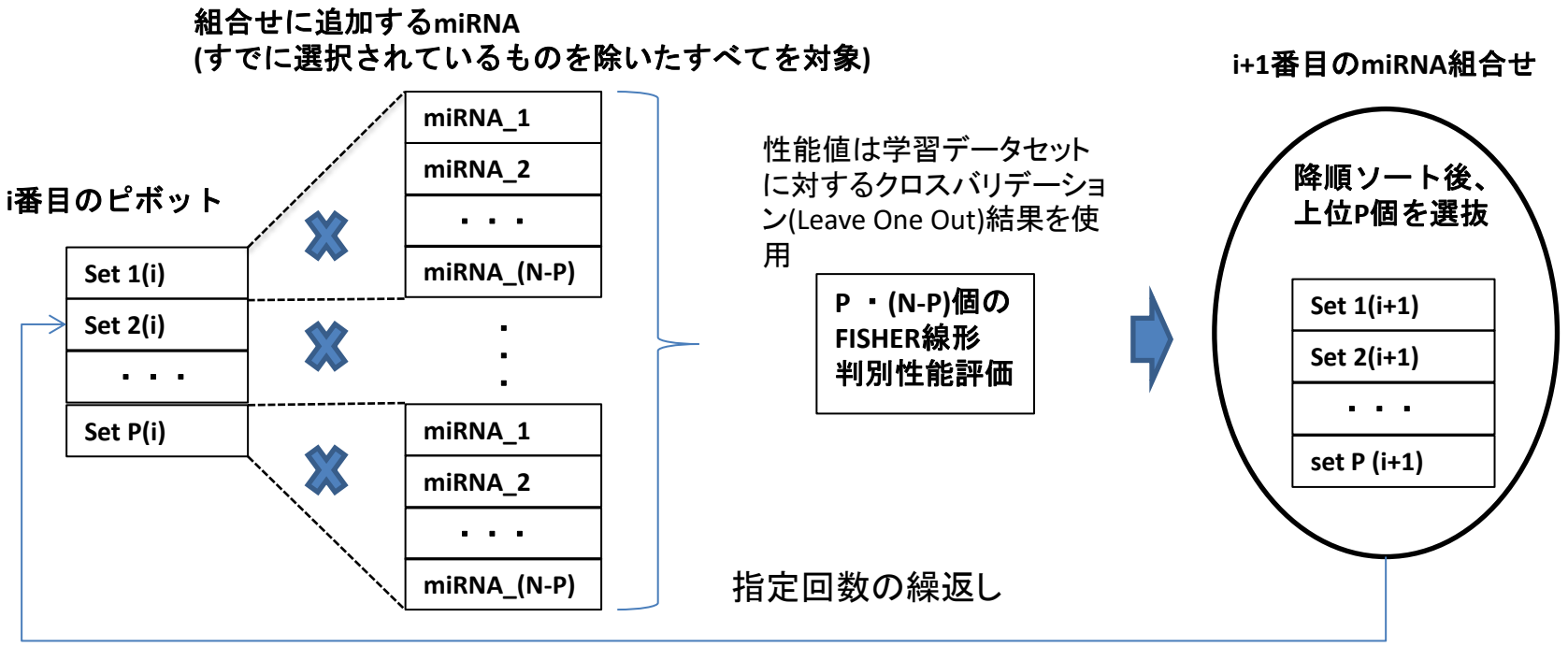
- 個々にp値（または効果量など）が有意でなくても、miRNAの組合せによって力を発揮するものがある
- miRNAがm個あったとするとmの階乗オーダーの組み合わせがある
- m=100の場合 → 組み合わせの数は「9.332622e+157」になる。今回のPJでは300～600のmiRNAのデータがあり天文学的組合せ数である
- 最近はスパース推定による変数絞り込み方法を利用することが多くなっているが、一つの最適解を得るのみ



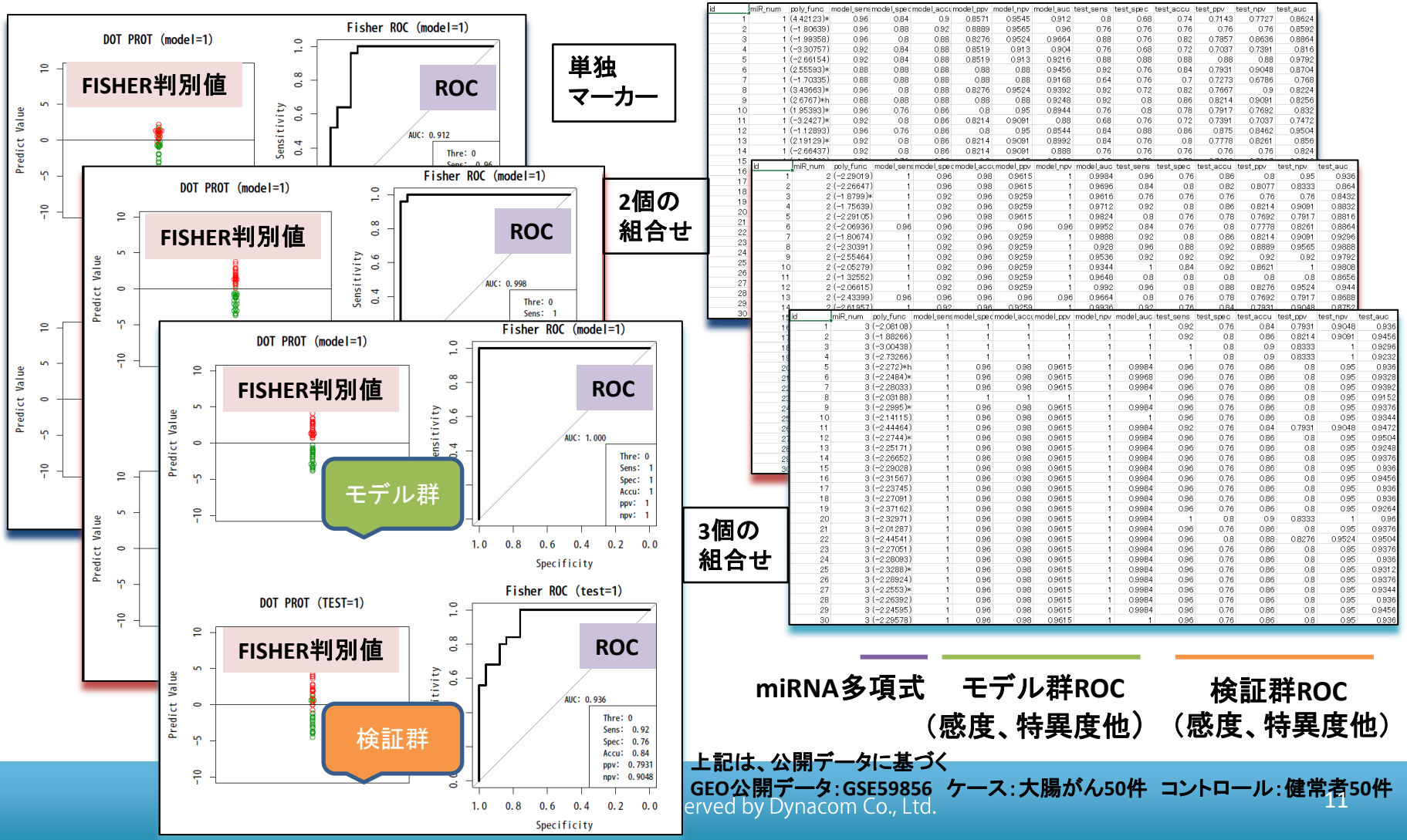
例：単独で有意でないものでも組合せると有効な例

線形判別+組合せ最適化の処理フロー

修正グリーティアルゴリズム:
最初に単独で識別性能が良いP個のmiRNAをピボットにする。ピボットを除く残りのmiRNA (N-P)個と組合せ(P･(N-P)個)の中から、性能の良い組合せ上位P個を次の段階の候補として残す。これを指定された組合せ数になるまで繰り返す。



結果出力例（線形判別+組合せ最適化）



Id	miR_num	poly_func	model_sens	model_spec	model_accu	model_ppv	model_npv	model_auc	test_sens	test_spec	test_accu	test_ppv	test_npv	test_auc
1	1	(4.42123)*	0.96	0.84	0.9	0.8571	0.9545	0.912	0.8	0.68	0.74	0.7143	0.7727	0.8624
2	1	(−1.80639)	0.96	0.88	0.92	0.8889	0.9565	0.96	0.76	0.76	0.76	0.76	0.76	0.8592
3	1	(−1.99358)	0.96	0.8	0.88	0.8276	0.9524	0.9664	0.88	0.76	0.82	0.7857	0.8636	0.8864
4	1	(−3.30757)	0.92	0.84	0.88	0.8519	0.913	0.904	0.76	0.68	0.72	0.7037	0.7391	0.816
5	1	(−2.66154)	0.92	0.84	0.88	0.8519	0.913	0.9216	0.88	0.88	0.88	0.88	0.88	0.9792
6	1	(2.55593)*	0.88	0.88	0.88	0.88	0.88	0.9456	0.92	0.76	0.84	0.7931	0.9048	0.8704
7	1	(−1.70335)	0.88	0.88	0.88	0.88	0.88	0.9168	0.64	0.76	0.7	0.7273	0.6786	0.768
8	1	(3.43663)*	0.96	0.8	0.88	0.8276	0.9524	0.9392	0.92	0.72	0.82	0.7667	0.9	0.8224
9	1	(2.6767)*h	0.88	0.88	0.88	0.88	0.88	0.9248	0.92	0.8	0.86	0.8214	0.9091	0.8256
10	1	(1.95393)*	0.96	0.76	0.86	0.8	0.95	0.8944	0.76	0.8	0.78	0.7917	0.7692	0.832
11	1	(−3.2427)*	0.92	0.8	0.86	0.8214	0.9091	0.88	0.68	0.76	0.72	0.7391	0.7037	0.7472
12	1	(−1.12893)	0.96	0.76	0.86	0.8	0.95	0.8544	0.84	0.88	0.86	0.875	0.8462	0.9504
13	1	(2.19129)*	0.92	0.8	0.86	0.8214	0.9091	0.8992	0.84	0.76	0.8	0.7778	0.8261	0.856
14	1	(−2.66437)	0.92	0.8	0.86	0.8214	0.9091	0.888	0.76	0.76	0.76	0.76	0.76	0.824

Id	miR_num	poly_func	model_sens	model_spec	model_accu	model_ppv	model_npv	model_auc	test_sens	test_spec	test_accu	test_ppv	test_npv	test_auc
1	2	(−2.29019)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.936
2	2	(−2.26647)	1	0.96	0.98	0.9615	1	0.9696	0.84	0.8	0.82	0.8077	0.8333	0.864
3	2	(−1.8799)*	1	0.92	0.96	0.9259	1	0.9616	0.76	0.76	0.76	0.76	0.76	0.8432
4	2	(−1.75639)	1	0.92	0.96	0.9259	1	0.9712	0.92	0.8	0.86	0.8214	0.9091	0.8832
5	2	(−2.29105)	1	0.96	0.98	0.9615	1	0.9824	0.8	0.76	0.78	0.7692	0.7917	0.8816
6	2	(−2.06936)	0.96	0.96	0.96	0.96	0.96	0.9952	0.84	0.76	0.8	0.7778	0.8261	0.8864
7	2	(−1.80674)	1	0.92	0.96	0.9259	1	0.9888	0.92	0.8	0.86	0.8214	0.9091	0.9296
8	2	(−2.30381)	1	0.92	0.96	0.9259	1	0.928	0.96	0.88	0.92	0.8889	0.9565	0.9688
9	2	(−2.55464)	1	0.92	0.96	0.9259	1	0.9536	0.92	0.92	0.92	0.92	0.92	0.9792
10	2	(−2.05279)	1	0.92	0.96	0.9259	1	0.9344	1	0.84	0.92	0.8621	1	0.9808
11	2	(−1.32552)	1	0.92	0.96	0.9259	1	0.9648	0.8	0.8	0.8	0.8	0.8	0.8656
12	2	(−2.06615)	1	0.92	0.96	0.9259	1	0.992	0.96	0.8	0.88	0.8276	0.9524	0.944
13	2	(−2.43399)	0.96	0.96	0.96	0.96	0.96	0.9664	0.8	0.76	0.78	0.7692	0.7917	0.8688
14	2	(−2.61957)	1	0.92	0.96	0.9259	1	0.9836	0.92	0.76	0.84	0.7931	0.9048	0.8752

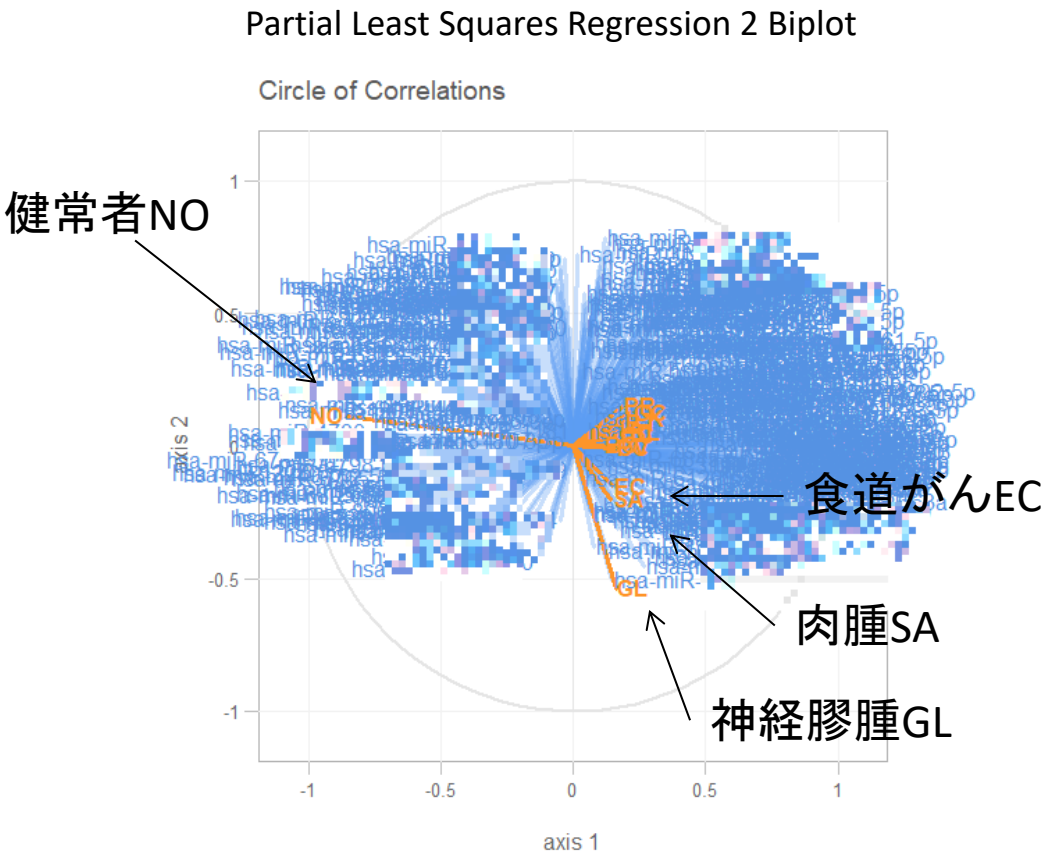
Id	miR_num	poly_func	model_sens	model_spec	model_accu	model_ppv	model_npv	model_auc	test_sens	test_spec	test_accu	test_ppv	test_npv	test_auc
1	3	(−2.06108)	1	1	1	1	1	1	0.92	0.76	0.84	0.7931	0.9048	0.936
2	3	(−1.88266)	1	1	1	1	1	1	0.92	0.8	0.86	0.8214	0.9091	0.9456
3	3	(−3.00438)	1	1	1	1	1	1	1	0.8	0.9	0.8333	1	0.9296
4	3	(−2.73266)	1	1	1	1	1	1	1	0.8	0.9	0.8333	1	0.9232
5	3	(−2.272)*h	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.936
6	3	(−2.2484)*	1	0.96	0.98	0.9615	1	0.9968	0.96	0.76	0.86	0.8	0.95	0.9328
7	3	(−2.28033)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9392
8	3	(−2.03188)	1	1	1	1	1	1	0.96	0.76	0.86	0.8	0.95	0.9152
9	3	(−2.2995)*	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9376
10	3	(−2.14115)	1	0.96	0.98	0.9615	1	1	0.96	0.76	0.86	0.8	0.95	0.9344
11	3	(−2.44464)	1	0.96	0.98	0.9615	1	0.9984	0.92	0.76	0.84	0.7931	0.9048	0.9472
12	3	(−2.2744)*	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9504
13	3	(−2.25171)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9248
14	3	(−2.26652)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9376
15	3	(−2.29028)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.936
16	3	(−2.31567)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9456
17	3	(−2.23745)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.936
18	3	(−2.27091)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.936
19	3	(−2.37162)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9264
20	3	(−2.32971)	1	0.96	0.98	0.9615	1	0.9984	1	0.8	0.9	0.8333	1	0.96
21	3	(−2.01287)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9376
22	3	(−2.44541)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.8	0.88	0.8276	0.9524	0.9504
23	3	(−2.27051)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9376
24	3	(−2.28093)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.936
25	3	(−2.3288)*	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9312
26	3	(−2.28924)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9376
27	3	(−2.2553)*	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9344
28	3	(−2.26392)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.936
29	3	(−2.24595)	1	0.96	0.98	0.9615	1	0.9984	0.96	0.76	0.86	0.8	0.95	0.9456
30	3	(−2.29578)	1	0.96	0.98	0.9615	1	1	0.96	0.76	0.86	0.8	0.95	0.936

miRNA多項式　**モデル群ROC（感度、特異度他）**　**検証群ROC（感度、特異度他）**

上記は、公開データに基づく
GEO公開データ：GSE59856　ケース：大腸がん50件　コントロール：健常者50件

疾患横断的な組み合わせ判別マーカー探索

- データの概要
  - がん13種(BC, BL, BT, CC, EC, GC, GL, HC, LK, OV, PC, PR, SA):各200件
  - 健常者(NO): 2,600件
  - QC基準を満たしたmiRNA: 366本
- 多目的変数とmiRNAの全体的な関係(右図)
  - PLS2(Partial Least Squares Regression 2)によるBiplot



疾患横断的な組み合わせ判別マーカー探索: 判別モデルの例

- 多群判別モデル: Multinomial Model は Logistic回帰の組み合わせ
- 2群判定の組み合わせは、$\{2^{(n-1)}-2\}$の組み合わせが可能
  - N個のモデル(例: 右図)
  - 最小数のモデル: $\log_2 N$

単純な直交行列を用いた組み合わせの一例(Multinomial Modelの典型例)

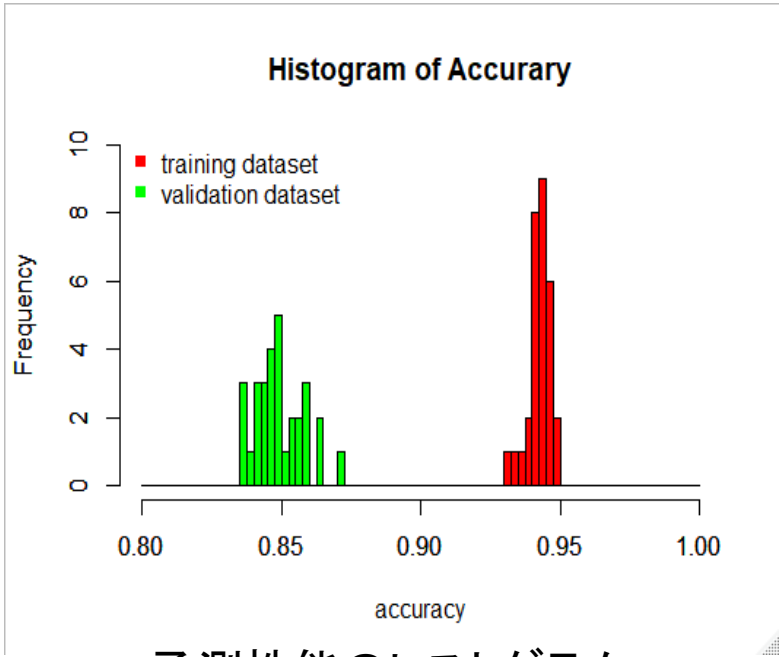
横軸：モデル1～5, 縦軸：疾患A～E

	1	2	3	4	5
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1

14群に対するMultinomial Model

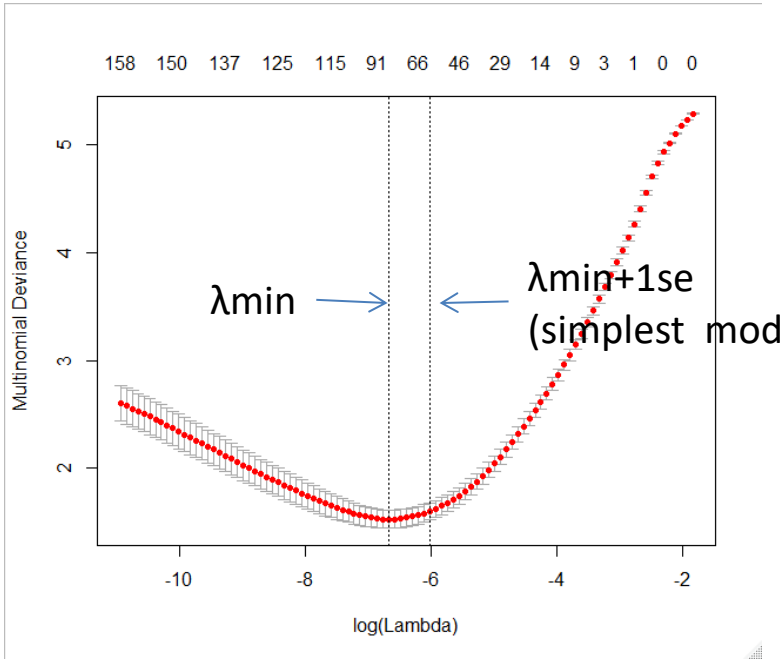
LASSO: Least Absolute Shrinkage and Selection Operator

- Repeated Cross-Validation
  - the number of repetitions : 30
    - Training set: 80%, validation set: 20%
    - Training set: 5-fold cross validation



予測性能のヒストグラム

↑予測残差



クロスバリデーション時の予測残差例

情報エントロピーを用いた組み合わせ

情報エントロピー：バランス良く小分けできるほど大きな値になる性質を持つ
（4に分ければ2bit, 8に分ければ3bit）

	CC	EC	GC	NC
No. 4	0	0	0	1
No. 2	0	1	0	0
No. 5	1	1	0	0

行単位のエントロピー

$E=-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}=0.8$ bit

$E=-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}=1.0$ bit

全体（4つに別れているので）

$E=-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}=2.0$ bit

組合せアルゴリズムは、評価関数として**正規化したエントロピーとAUCの積**を用い、N個の候補を残すGreedy Algorithmを改良した方法

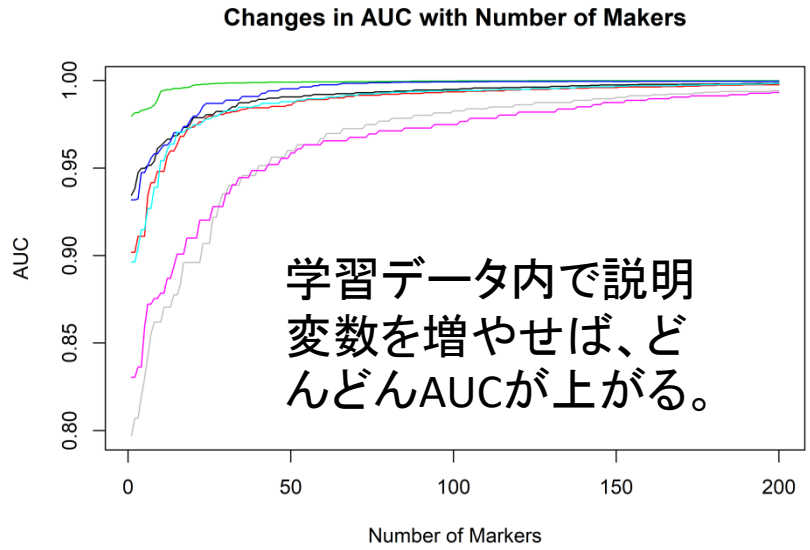
別の組み合わせ解の例

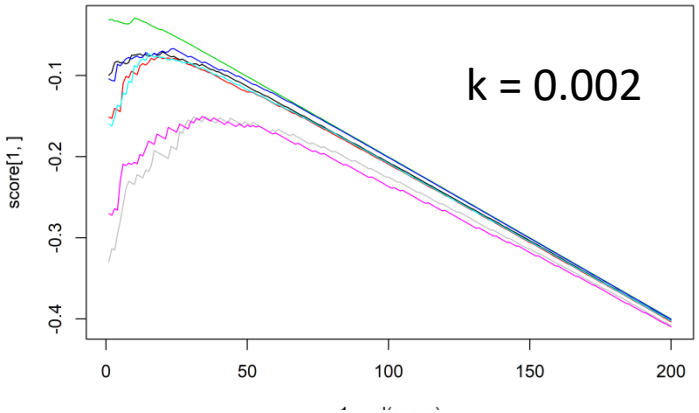
	CC	EC	GC	NC
No. 4	0	0	0	1
No. 2	0	1	0	0
No. 3	0	0	1	0

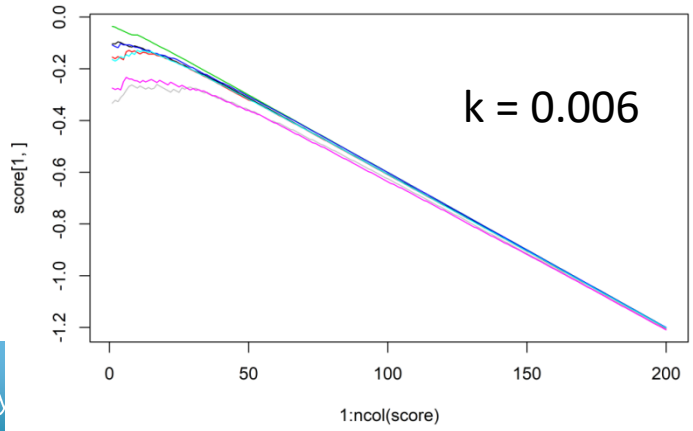
※No.4はNCとその他を区別するモデルで単一で高性能のため、固定で入れている
※0/1は2群を区別する名義尺度なので、陽性・陰性は反転しても構わない
※原理的には(2<)N分割でlogN()とすることが可能。

情報量によるマーカーの絞込み

クロスバリデーションでは、予測残差を最小とするλminの位置を決定できるが、数が多すぎる場合絞込の別な基準が必要である。→情報量規準の考えを導入





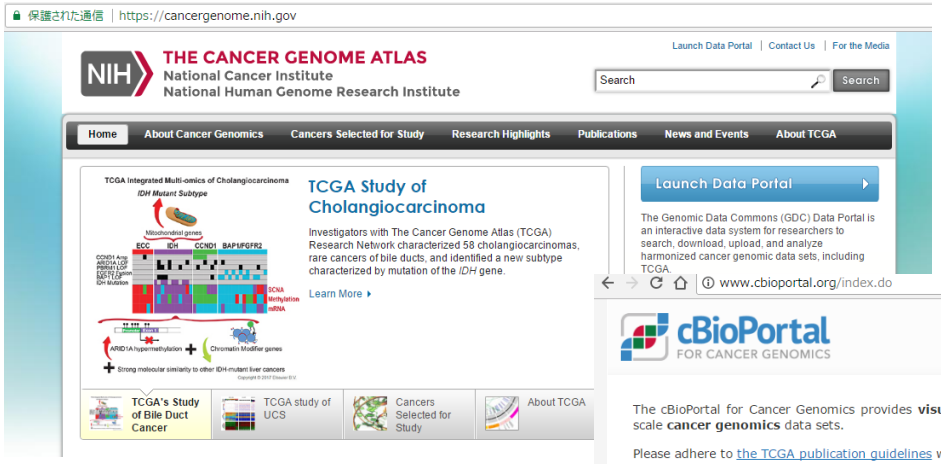


情報量＝$\log_2$(AUC) – k * miRNA数

マーカー数によるペナルティ項

miRNA情報の取得Tool

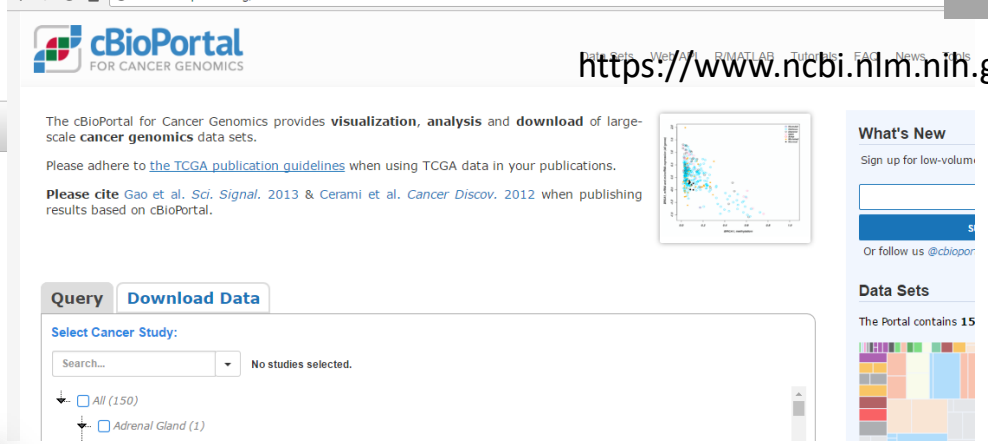
miRNA関連遺伝子情報探索→共発現遺伝子情報まで



cBioPortal

TCGAのポータルサイト

https://www.cbioportal.org/



PubMed

https://www.ncbi.nlm.nih.gov/pubmed/

TCGA

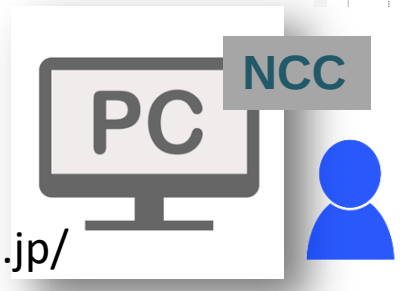
がん関連総合データサイト

https://cancergenome.nih.gov/

COXPRESdb

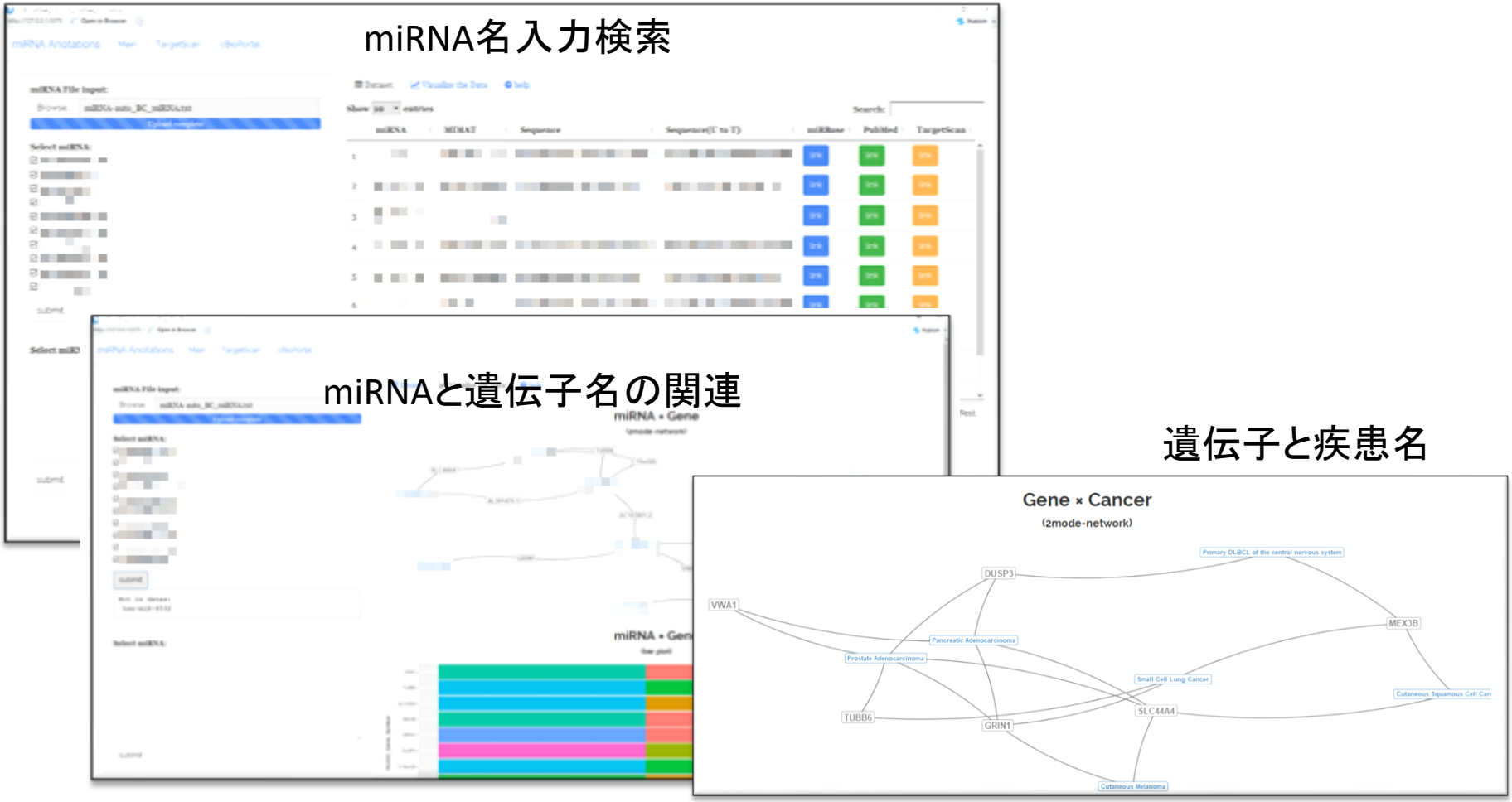
共発現遺伝子DB

https://coxpresdb.jp/



がんマーカー
miRNA関連
データ取得ス
クリプト

miRNA情報の自動取得結果



弊社が関係している主な案件

- 標準血清データ解析
- 乳がん転移予測
- 膀胱がん早期診断
- 前立腺がんの早期診断
- 卵巣がん新規診断
- 食道がんの診断および治療効果予測
- 大腸がんステージII外科治療の予後予測
- 脳腫瘍の判別（3群判別、IDH: Isocitrate Dehydrogenase予測）
- その他（標準血清のデータ解析）

まとめ

- 複数のマーカー候補を得るために、修正グリーディーアルゴリズムによる組合せ最適化プログラムを開発した
  - 線形判別による複数マーカー組み合わせ最適化アルゴリズム
  - 複数疾患を対象とするEntropy組み合わせ最適化アルゴリズムを開発した
- 判別モデルの性能
  - がん種と健常者を一対一でモデル化した場合、90%台
  - 複数疾患横断的な予測モデルでは全体で0.85程度のAccuracy
  - Entropy + AUCによる総合的な評価基準を用い、がんの種類を3種程度に絞った場合、20数種のmiRNAの組み合わせで80%～90%台の判別モデルを構築が可能
- miRNA 情報の取得
  - miRNAと遺伝子、関連疾患情報を公開DBから自動収集するスクリプトを作成