

生命情報解析サービス

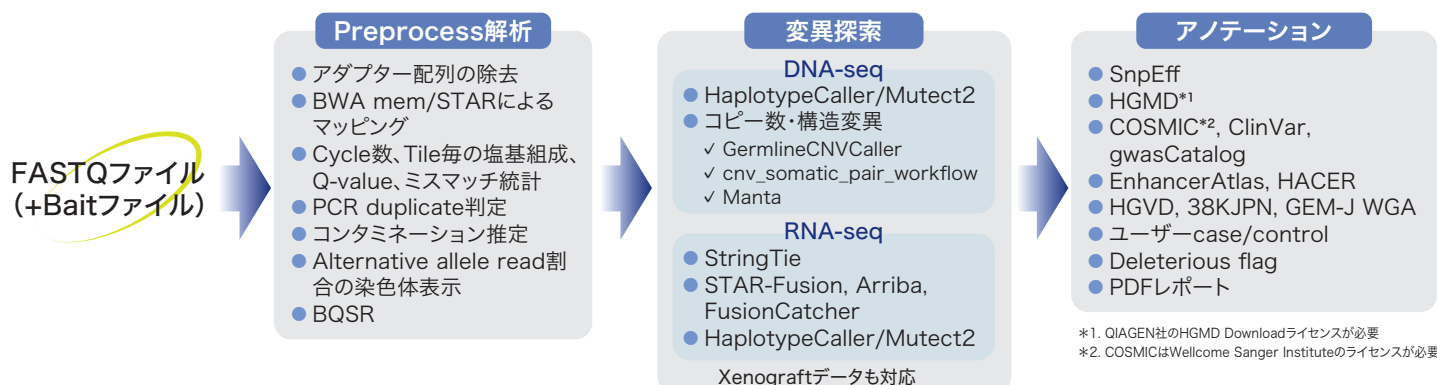
生物・医療統計、バイオインフォマティクス、ケモインフォマティクス、
AIを用いたデータ解析・コンサルティングをご提供

主なデータ解析項目

- 次世代シーケンサー (NGS) データ解析: WGS、WES、WTS、TS
- オミックスデータ解析: SNPs、メチル化、遺伝子発現、代謝物データ解析、パスウェイ解析
- 多層オミックスデータ解析: Loss of function解析、多層パスウェイ変動解析
- AIモデル構築: 機械学習、deep learning
- ケモインフォマティクス: 構造活性相関、有害性予測

次世代シーケンサー (NGS) データ解析

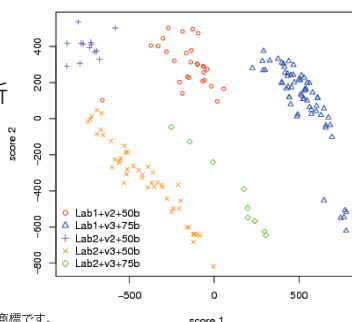
GATK Best Practices® に準拠した解析を提供



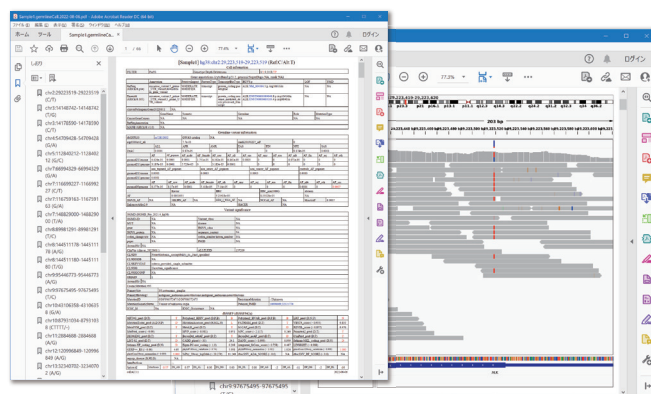
ポスト解析

- バイアス解析
- Case-control study
- Deleterious変異遺伝子解析
- Loss of function解析
- エキスパートパネル支援

Principal components scatter plot: FPKM



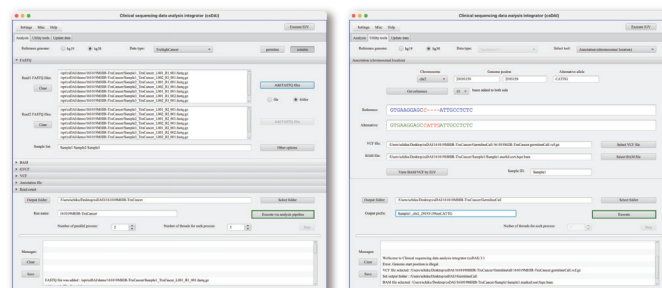
※「GATK BEST PRACTICES」は、The Broad Institute, Inc.の登録商標です。
※HGMDは米国QIAGENの米国およびその他の国における登録商標または商標です。



Clinical sequencing data analysis integrator (csDAI®)

csDAIは次世代シーケンサー (NGS) データ解析の統合パッケージであり、GUIからGATK Best Practicesに準拠した解析を可能にしたシステムです。GUIはmacOS®及びLinux®上で動作し、Linux版ではPCクラスター等のバッチシステムに対応したコマンドライン版も用意されています。

※「csDAI」は、みずほリサーチ&テクノロジーズ株式会社の登録商標です。
※macOSは、米国およびその他の国で登録されたApple Inc.の商標です。
※Linuxは、Linus Torvaldsの米国およびその他の国における登録商標または商標です。



オミックスデータ解析

サービス概要

- ゲノム多型データ解析: genome wide association study (GWAS)
 - ▶ SNPsチップやDNA-seqから得られる多型データの解析
- メチレーションデータ解析
 - ▶ メチル化DNAアレイやbisulfite sequencing (BS-seq)から得られるDNAメチル化データの解析
- トランスクリプトームデータ解析
 - ▶ DNAマイクロアレイやRNA-seq発現量データ解析から得られる遺伝子発現量データの解析
- メタボロームデータ解析
 - ▶ ガスクロマトグラフィー質量分析法 (GC/MS) 等から得られる代謝物データの解析

サービスの流れ

実験データのQC、正規化

分散分析、傾向性検定等の統計的分析

単層/多層でのエンリッチメント解析

既に取得済のデータであっても再度データのquality control (QC) からパスウェイ解析まで、一連のデータ解析を実施

ゲノム多型データ解析: genome wide association study (GWAS)

◆データのQC

測定系のエラーやバイアスに加えて検体の確認も行う必要があります。

- 遺伝子型一致割合による血縁関係の同定
- 近交係数とrare allele数による適合基準の確認
- 性別の同定
- Minor allele frequency (MAF) によるSNPsの足切り
- HWE検定によるSNPs graphの確認と不良call SNPsの同定

◆集団の構造化解析

関連を解析する集団間に分集団の不均衡が存在すると、分集団に特徴的なSNPsを多数検出してしまいます。そのため分集団の存在をチェックし、帰無仮説がどの程度の精度で成立しているのかを予め明らかにしておく必要があります。

EIGENSTRAT (Price et al. 2006)



・上図は、ヒトの遺伝子解析における集団存在相関の例であり、人種や民族に対する優劣などを表現するものではありません。

◆関連解析

- 遺伝子型trend test (exact)
- Matched case-control study
- Logistic回帰 (バイアス調整)
- Haplotype推定を用いた関連解析

メチレーションデータ解析

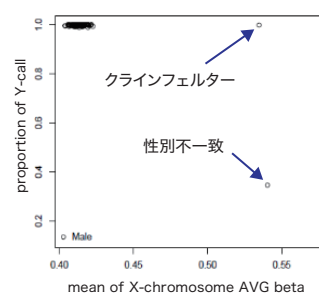
メチル化解析アレイのデータは、遺伝子発現と同様に定量値のため様々な実験バイアスが生じます。様々な品質管理解析に加え、バイアス探索解析を行い、適切にバイアスを補正した統計解析が必要になります。

◆データのQC

手法例:

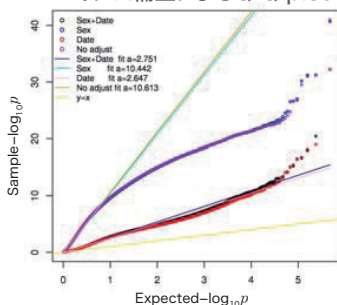
- プローブcall割合ヒストグラム
- サンプルcall割合ヒストグラム
- 性別の確認

性染色体上のプローブによる性別確認

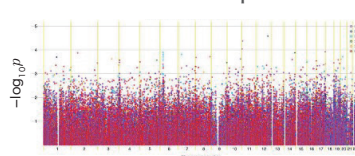


◆バイアス補正統計解析

バイアス補正によるQ-Q plot

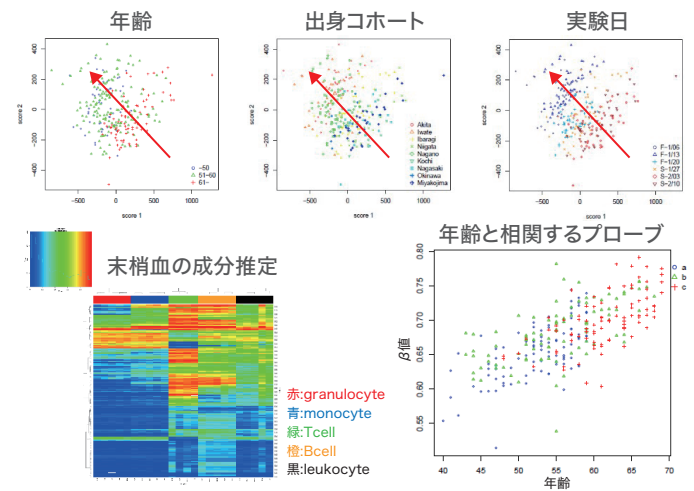


Manhattan plot



◆バイアス解析

主成分分析の結果を様々な情報で色づけ、臨床情報と実験バイアスとの交絡を探索します。検体の細胞種含有割合もバイアスになります。



■トランスクリプトームデータ解析

◆データのQC

実験データの性質を把握するためにデータを可視化
実験データのバイアスやラベルのつけ間違い等がない等をチェック

手法例:

- 発現量のヒストグラム
- ボックスプロット
- 主成分分析

◆遺伝子(プローブ)毎の統計解析・多変量解析

2郡以上のデータに対して各プローブ・遺伝子の有意差を評価
検体数やデータの分散から適当な手法を選定

手法例:

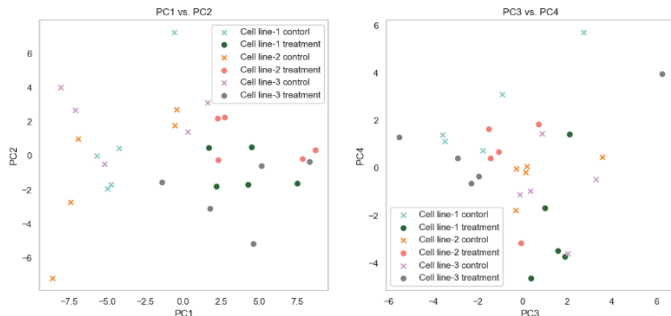
- 分散分析
ノンパラメトリック検定: Kruskal-Wallis検定、Mann-WhitneyのU検定
パラメトリック検定: 1元配置分散分析、Studentのt検定、Welchのt検定
- 傾向性検定
Jonckheere-Terpstra検定、線形回帰モデルを用いた検定
- 多変量解析
主成分分析、クラスター分析、因子分析、判別分析、重回帰

■メタボロームデータ解析

◆データのQC

実験データの性質を把握するためにデータを可視化
実験データのバイアスやラベルのつけ間違い等がない等をチェック
化学物質名称の修正等も必要に応じて実施

例:主成分分析によるデータの可視化



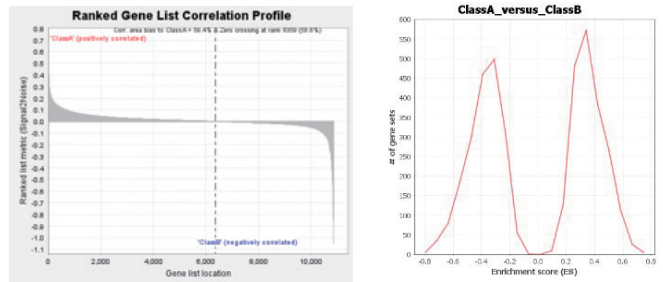
◆Flux balance analysis(FBA)

代謝物のパスウェイに対して化学反応の量論的關係を適用し、代謝回路の挙動範囲(代謝量の制限)と経路の特徴を分析します。代謝経路と最終生成物の量に関する影響を定量的に予測できます。

◆パスウェイ変動解析

2郡間のデータに対して遺伝子セット単位で有意差を評価
データの確認や考察のために解析結果にはgene setの説明やリンク等もまとめてスプレッドシート形式で整理

Gene Set Enrichment Analysis(GSEA)*を用いた解析例



*Subramanian, Aravind, et al. *Proceedings of the National Academy of Sciences* 102.43 (2005): 15545-15550.

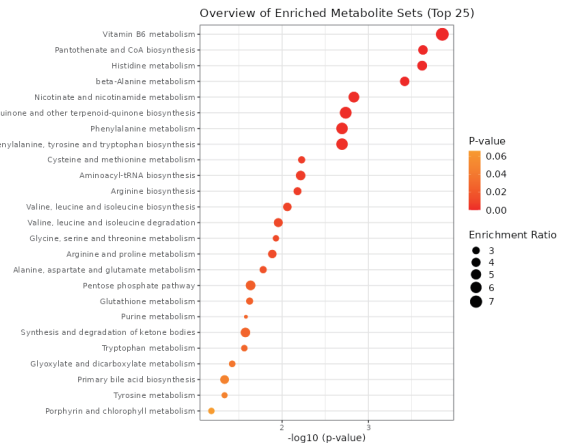
◆時系列データ解析

マイクロアレイの時系列データから遺伝子発現量の変化を解析することにより、遺伝子間の相関パターンを構成します。転写因子やリン酸化酵素などの機能情報を組込んだモデルの推論を行います。

◆パスウェイ変動解析

2郡間のデータに対して遺伝子セット単位で有意差を評価
データの確認や考察のために解析結果にはgene setの説明やリンク等もまとめてスプレッドシート形式で整理

Metabolite Set Enrichment Analysis(MSEA)*を用いた解析例



*Xia, Jianguo, and David S. Wishart. *Nucleic acids research* 38.suppl_2 (2010): W71-W77.

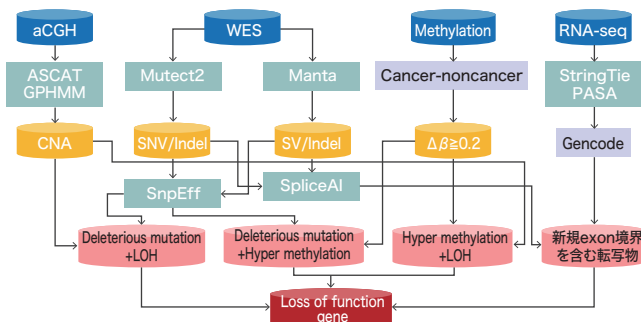
多層オミックスデータ解析

◆コピー数・変異・メチル化・発現量を組み合わせた loss of function解析

両アレルの機能失活によりがん化が進むことが考えられます。機能失活には幾つかの原因が考えられ、これらを組み合わせた解析が有効になると考えられます。

多層オミックスデータを組み合わせパイプライン例

- ゲノムコピー数をSNPsアレイデータから解析
- ゲノム上の変異をwhole exome sequencingデータから解析
- メチレーションチップデータからメチル化による転写制御推定
- RNA-seqデータから新規exon境界を確認



◆多層パスウェイ変動解析

マイクロアレイデータと代謝物データがある場合、両方のデータを利用して多層のパスウェイ変動解析を実施
多層で評価することで解析結果の信頼性の向上や考察の深化に貢献

- mRNA発現量データのパスウェイ変動解析
- miRNA発現量データとmRNA発現量データを組み合わせた解析
- 代謝物データのパスウェイ変動解析
- mRNA発現量データと代謝物データを組み合わせたパスウェイ変動解析

MetaboAnalyst5.0*のJoint Pathway-Analysisを用いた解析例

パスウェイとその概要

パスウェイに関連した実験データの遺伝子リスト

パスウェイに関連した実験データの代謝物名

Pathway	Class	Description	Total	Expected	Hit	Raw p	Adj. p	FDR	Score	matched_chemicals	matched_genes
1 Drug metabolism - cytochrome P450	Metabolism	Metabolism: Xenobiotics biodegradation	39	25.248	19	3.1E-08	3E-08	1.23E-07	0.713	1	1
2 alpha-Linolenic acid metabolism	Metabolism	Lipid metabolism:BRITe	22	14.242	24	3.1E-08	3E-08	1.23E-07	0.702	1	1
3 Drug metabolism - other enzymes	Metabolism	Metabolism: Xenobiotics biodegradation	69	44.669	80	3.1E-08	3E-08	1.23E-07	0.626	1	1
4 Drug metabolism - other enzymes	Metabolism	Metabolism: Glycan O-glycans are a	22	14.242	27	3.1E-08	3E-08	1.23E-07	0.61	1	1
5 Mucin type O-glycan biosynthesis	Metabolism	Metabolism: Lipid metabolism:BRITe	39	25.248	42	3.1E-08	3E-08	1.23E-07	0.5	1	1
6 Ether lipid metabolism	Metabolism	Metabolism: Lipid metabolism:BRITe	79	51.143	84	3.1E-08	3E-08	1.23E-07	0.100	1	1
7 Arachidonic acid metabolism	Metabolism	Metabolism: Amino acid metabolism	49	31.722	56	3.1E-08	3E-08	1.23E-07	0.100	1	1
8 Lysine degradation	Metabolism	Metabolism: Carbohydrate metabolism	17	11.095	22	3.1E-08	3E-08	1.23E-07	0.100	1	1
9 Acetate and acetate metabolism	Metabolism	Metabolism: Lipid metabolism:BRITe	86	55.675	89	3.1E-08	3E-08	1.23E-07	0.123	1	1
10 Glycophosphatidyl metabolism	Metabolism	Metabolism: Metabolism of cofactors	14	9.0533	16	3.1E-08	3E-08	1.23E-07	0.123	1	1
11 Thiamine metabolism	Metabolism	Metabolism: Carbohydrate metabolism	69	44.669	73	3.1E-08	3E-08	1.23E-07	0.126	1	1
12 Inositol phosphate metabolism	Metabolism	Metabolism: Biosynthesis of other a	4	2.5895	5	3.1E-08	3E-08	1.23E-07	0.133	1	1
13 Neurexin, neuromin and gartenmannin biosynthesis	Environmental Information Processing	Environmental Information Processing	74	47.906	83	3.1E-08	3E-08	1.23E-07	0.134	1	1
14 Phosphatidylinositol signaling system	Metabolism	Metabolism: Lipid metabolism:BRITe	35	22.658	52	3.1E-08	3E-08	1.23E-07	0.135	1	1
15 Glycolipid metabolism	Metabolism	Metabolism: Energy: The biological a	10	6.4738	19	3.1E-08	3E-08	1.23E-07	0.144	1	1
16 Nitrogen metabolism	Metabolism	Metabolism: Energy: The biological a	10	6.4738	19	3.1E-08	3E-08	1.23E-07	0.144	1	1

緑色セル:弊社付加情報部分

*<https://www.metaboanalyst.ca/>

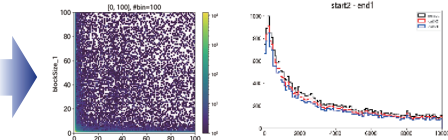
サービス概要

データの量や性質によって適切な機械学習・深層学習モデルを選択する必要があります。また学習を成功させるにはモデルに応じて適切な前処理も重要になります。

◆データのの前処理

- データの収集支援、データクリーニング
- 機械学習・深層学習モデルに特化したデータ成型(前処理)の支援
- 学習にとって有益と考えられる特徴量の抽出

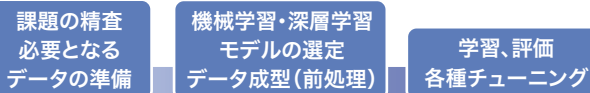
各種統計処理から
データの持つ性質を
調査し、前処理の実施



◆機械学習・深層学習モデルの学習、構築の支援

- 目的用途に合わせたモデルの提案と実装
- 各種ツール(scikit-learn/TensorFlow/PyTorch)を解析目的に合わせて選択
- 深層学習モデルの出力結果の解釈支援

サービスの流れ



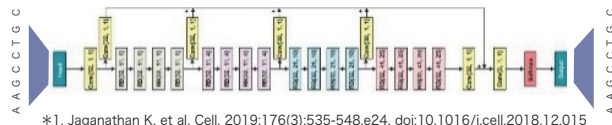
◆チューニング

- 学習の成否を決める初期値範囲の探索
- 深層学習ハイパーパラメーター探索
- 各種ライブラリの特性調査、最適な設定の探索

実績紹介

SpliceAI*1モデルを利用したエンハンサー領域の学習*2

SpliceAIは長距離の塩基配列の影響を組み込んだモデルであり、タンパク質との結合関係を学習していると思われます。



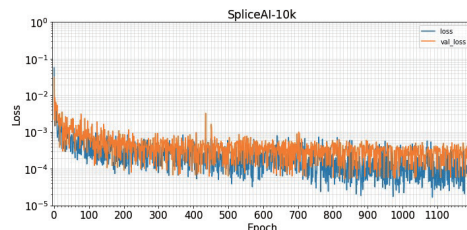
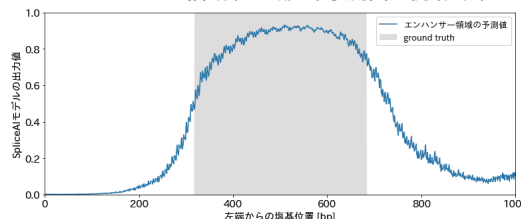
*1. Jaganathan K, et al. Cell. 2019;176(3):535-548.e24. doi:10.1016/j.cell.2018.12.015

*2. 国立がん研究センターとの共同研究

◆エンハンサー領域予測モデルの構築

- エンハンサー領域データの作成
- SpliceAIモデルのエンハンサー予測のための修正

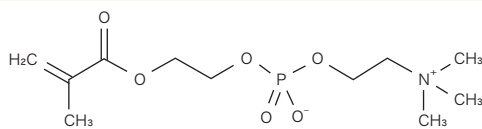
エンハンサー領域(灰色部分)予測出力例(青線)



ケモインフォマティクス解析

サービス概要

分子の化学構造と物性や活性等との関係を情報学的に解析することで分子の選択やデザインを行うことが可能になります。



化学構造式

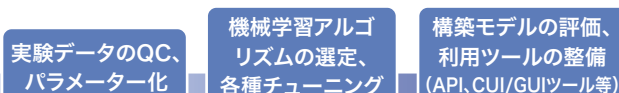
データの性質に合わせた
機械学習アルゴリズムを適用

目的値の予測

- 物性値/活性値
- 表現型/有害性

予測対象についても
実験バイアスや目的を考慮し
パラメーター化

サービスの流れ



◆構造活性相関

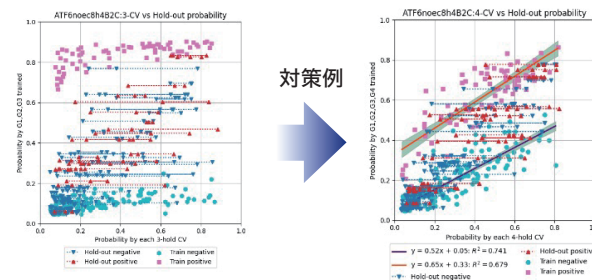
生物・医療系データはデータサイズが小さく、活性等の有無が不均衡なデータとなることが多くあります。

課題

- データが小さすぎてundersamplingやoversamplingが機能しない
- Cross-validation(CV)時の分割による揺らぎが大きい
- CVの評価を含めてoverfittingしてしまう
- 予測対象は多数の生理活性イベントが重なっていると予想される

- 対策
- ✓ Imbalanceデータに強い指標を用いた交差検定
 - ✓ 予測確率を活用したoverfitting対策
 - ✓ 特徴量重要度に基づく記述子選択
 - ✓ Multiple partitionsを用いたhold-out評価

予測確率を活用したoverfitting対策



CV学習時とhold-out評価時のprobabilityの相関をチェック

問い合わせ先

みずほリサーチ&テクノロジーズ

〒101-8443 東京都千代田区神田錦町 2-3
TEL : 03-5281-5289 FAX : 03-5281-3457
e-mail : icd-joho@mizuho-rt.co.jp.jp
URL : <https://www.mizuho-rt.co.jp>

情報通信研究部社会基盤技術チーム

ご相談やお見積もり等の詳細については、
左記までお気軽にお問い合わせください。