Abstract

Omics studies aim to extract meaningful messages from a large-scale and high-dimensional data sets by treating the data sets as a whole. The attitude to treat the data sets as a whole plays an important role in every step of the data handling procedures; the pre-processing step of data records, the step of statistical analyses and machine learnings, translation of the outputs into human natural perceptions, and acceptance of the messages with uncertainty. In the preprocessing, how to control data quality and batch effects are discussed. For the main analyses, the approaches are divided into two types and their basic concepts are discussed; (1) evaluation of many items individually and followed by interpretation of individual items in the context of multiple testing and combination and (2) extraction of fewer important aspects from the whole data records. The outputs of main analyses are translated into natural languages with the technique, such as annotation and ontology. The other way of technique to make the outputs perceptionable is visualization of them. At the end of this review, one of the most important issues in the interpretation of omics data analyses is discussed; the omics studies have a large amount of information in their data sets and every approach only reveals a very restricted aspect of the whole data sets and the understandable messages from them have unavoidable uncertainty.

(225 words)

1. Omics studies

Omics studies are the studies where a particular type of molecules in the samples are measured for their characters and quantity as a whole and their patterns and/or relation with the samples attributes are investigated. Genomic studies measure DNA molecules and epigenomic, transcriptomic, proteomic and metabolomic studies measure chemical states of DNA and its binding proteins, RNA, proteins and metabolites, respectively. The idea of –omics, or collective measurement, is applied out of molecules but to various measurable targets, such as a set of traits (phenome), states of brain neural networks (connectome) and a bacterial flora (metagenome).

In omics studies each experimental instance generates huge amount of information at once. For example, a massive parallel sequencing (MPS) experiment produces billions of short reads for a genomic, epigenomic, or transcriptomic study. Another example is GC-MS (gas chromatography mass spectrometry) for metabolomic study that produces a spectrum that contains all the information of various metabolites.

1. Pre-processing step

Experiments are not perfect always both when they are manually handled small ones and when they are performed with expensive highly-automated high-throughput machines, that are the cases of omics experiments. The big difference between these two settings is that you can redo your manual experiments but that you would not repeat the Omics experiment even if the quality of a small fraction of outputs is unsatisfactory, because you cannot repeat the selected fraction separately. <https://www.ncbi.nlm.nih.gov/pubmed/28987912> (GLP)

One Omics experimental procedure corresponds to a large number of single experiments all at once and the quality among them vary; this is intra-experimental quality heterogeneity. Also a set of data records from one omics experimental procedure are affected by factors shared by all of them and another set of data records from another omics procedure are affected differently; this is inter-experimental heterogeneity and procedure dependent batch effect.

In a case of MPS, quality among the reads always vary, that is intra-experimental quality heterogeneity. When you run two MPS for two DNA samples, the first MPS’s set of reads may tend to be better than the second set of reads, for example because the DNA sample conditions are different. This is the inter-experimental heterogeneity.

<https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html>

Good one is good and bad one is bad, but…

Good data records should be used for data analysis and bad records should be removed, if possible. Unfortunately, quality of individual records is continuous and when records should be separated into two, the measure of quality has to be defined and its separating threshold value should be selected.

These raw sequences should be compared with the reference sequences. Again this raw output spectrum is not ready to be

ノイズ…ばらつきも情報

Quality and appropriateness for particular purpose

Single cell とノイズ

Omics approaches have been applied to single cells progressively. They enable us to understand chronological landscape of physiological differentiation and pathological progression of various diseases and particularly of cancer cell population.

GLP(Good Laboratory Principles [https://ja.wikipedia.org/wiki/Good\_Laboratory\_Practice](https://ja.wikipedia.org/wiki/Good_Laboratory_Practice?fbclid=IwAR1lfX1IeAsQLMXbMMQpYA22pmTNAYoK0qWIk9VEIMeTCgboIhNds8cY3hQ) )と言うのがありますが、それに準拠してのオミックスデータの質管理を問題にするとすれば、こんなペイパーが参考になるでしょうか。  
[https://www.ncbi.nlm.nih.gov/pubmed/28987912](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.ncbi.nlm.nih.gov%2Fpubmed%2F28987912%3Ffbclid%3DIwAR0zVYq5Je_ZfqzALiGjwXFWzhKbvH6dSyRhxQKIAeXQmUw1oTNySzE7ttU&h=AT3eNRQji4gLKj7xWlQSf_gwsx_IvM9r4OII0-D5RLKaus3kNlvpT0OnkQLa_Sl1Gr5ha-EWRN-iD9hTJ-DGqg4HTHfe-36KE6TU8qMq73rEDhMcaqPkf5_Qq3C5_0FSm6zfUSgc8JjecqtV5Y0RAVwk58xmXnXG7ANklkXpDmfZC0Oh8i-_25Y6Mn7Yz5-x4CqStFVxfyb_NXHgGrxuunNo3UsgEyr-LHa0Go7uW7DnLJxXkYtLehMh-t95XEODU8epAShQm76bVEDd8SfXFUhhZ1a0KtYM09bJdHuR8ENZ4A4-O7UK4bqpb3tZ_aCuZtJvhHyTpfmA90mMCxCc-I0koJo8d7W_EYsm6WfnSGAKUZd7ni0oETHNoq4C_vuZCEWaBneEGjaKfj3-kE4tPiMx_jBJA0kMJISDGaxgZ0RTRcKSrvpKQpt0a9r2gNueXNmw3_0y0WsoG9hW0fRV2-tduXmPSSIEp8kGY8E-zixEViSOowhhG1MKDTcdvH6DULb6odKInbg8y3WkHCger2DQfTOdFN2DqBThBQ6bkN8L5RoKw5DcMT3f904arA-xioNYcvsM2QvZ_47WY20ehBMbRoeErFbcF36bmunnHRnwklXeJin36Ai9bM9Ol7tBIIuneTy7osGflD94lvyWHnw)　  
アブストラクトには  
" challenges include (i) defining, storing, and archiving the  
raw data; (ii) transparent descriptions of data processing steps; (iii) software validation; and (iv)  
ensuring complete reproducibility of final results with respect to raw data. "  
とありますね。

[オミックスデータ解析とその結果の解釈：その３]  
データのQC：絶対的QCと相対的QC。  
「完璧にきれいでない情報」に「生物学的な情報は含まれる」

たくさんのデータが一度に得られるオミックス実験では個々のデータの質にばらつきが出ます。  
「ダメなデータは捨てる」という単純なやりかたももちろんありますが、手作業的な個別実験の積み重ねの場合と異なる点があり、最終的なデータ解釈方法とも関連があるので、留意点について把握しておき、QCのオプション選択に活かす必要があることを確認しておきましょう。

●　データには「絶対的な質」と「相対的な質」がある  
>>物理・化学測定量のみに依存して定量される質の良し悪しは「絶対的な質」  
>>そんな「絶対的な質」であっても、その「絶対的な質」の値だけをもって取捨選択するべきか否かは一度立ち止まって考える必要があります。一度の実験で取られた多数のデータの「絶対的な質」には分布がありますから、その情報を最終的解釈に使う可能性があるのであれば、「絶対的な質」も実験バッチを単位とした「相対的な質」に変じて、情報活用可能となる可能性があります  
>>例えば、物理測定量であるNGSのshort readsは、レファレンスゲノムにマップすることで、「生物学的なデータ」に変換されます。このとき、マップ対象であるレファレンスゲノムとの一致の良さを「データの質」の指標とすることがあり、この質が悪すぎたらフィルタリングアウトするとします。

●　生物学的な解析は、「尺度」からずれているものを拾うこと  
>>しかしながら、生物学的なデータは、「ばらついている様子を評価」することが「使命」なので、レファレンスとの不一致は「悪いこと」ではなくて、「貴重な情報」です。  
>>そのように考えなくては、ゲノムシークエンシングでのバリアントコールはできませんし、トランスクリプトーム解析でのオールターナティブ・スプライシングも見えてきません。とはいえ、レファレンスとのずれには「実験誤差」由来のものが含まれうるからこそ、「質」を問題にします。

●　「ばらつき」への興味の強さはスタディの性質によって変わる  
>>したがって、どんな「ばらつき」を知るために、どの程度「レファレンスとの違い」を是とするのかについて、方針を決める必要があります。それが、マッピングのオプション選択方針に直結します。「ばらつき」への興味が強ければ強いほど(ばらつきのありなしの興味だけでなく、ばらつきがあることを前提にして、その割合に興味があれば、それは「ばらつきへの興味」がより強いことになります。また、変異が大量にあることを前提とした癌ゲノムの配列解析は正常細胞のそれより「ばらつき」への興味が強いことになります)、「一見、質が悪そうなデータ」も捨てずに利活用する必要が生じます

●　マッピングには諸手法あり、それぞれにマッピング合否の特異度に影響する実行パラメタがある  
>>マッピング合否の特異度が上がれば、「ばらつきの検出感度」は下がります。  
>>マッピングのステップは計算機負荷がかかるので高速化アルゴリズムが提唱されることがありますが、その場合には、「高速化に当たって、何かを省略している」ことは必然ですから、それが自身のスタディの検出対象の感度・特異度・精度とどういう関係にあるのかは、新手法の使用の検討に際して、データの最終結果解釈側からの最大留意点になります

●　また、そもそも、レファレンスゲノムが1つの確定的なものではなくなっている  
>>ヒトゲノムは、全ゲノム配列を1列で表しえない、という立場から、複数の配列の入り組んだ様相全体を「ヒトゲノムレファレンス」とする方向性がある  
>>このような事情も「相対的な質」の定量に影響する

こんなあたりのことを書いているのが  
癌ゲノム関連では：  
[https://jmd.amjpathol.org/a…/S1525-1578(17)30598-6/fulltext…](https://jmd.amjpathol.org/article/S1525-1578(17)30598-6/fulltext?fbclid=IwAR1SLZsyaB-2kYDObvteJBGHcvDO-R40bednNOJqeyeXqRKJxmtH29Hi-nw#sec2.1)  
トランスクリプトーム関連では：  
[https://academic.oup.com/bioinformatics/…/33/13/2034/2995818](https://l.facebook.com/l.php?u=https%3A%2F%2Facademic.oup.com%2Fbioinformatics%2Farticle%2F33%2F13%2F2034%2F2995818%3Ffbclid%3DIwAR2-n-TNlDi88wq6wBmg19dFgPjxzgBtbSHoR0LCNe4OJzEkcR4XwhWzyH4&h=AT2UMRz_OhQZH4Uge-hp0cX1bNCMmWicbsiloc9Sl_UJPP3PQ_pN069QWeYLSghwiYB9evLvhfwgQFAi8x5l9KsMbR63_K38mz76ZWeuzEwblUekferqnOuRlGX9PzSqrD_InDDJqbHzx66tow3gM6R4t8bsTFptKb41vvsQQyNEuMkNGguV1oKQ0XWr6MO4Insmm0GhUFyxajbRb13lkXpJ-CIUS5B0Tl0PDBHfI2lvsG6h5oeIPd9CxVdBVkgU84n9dtBpf6TQ-mczxqVUWYEXYqRhQDBRv6jbR_RxLOY15E5HxnORjMm1lJY_k3h09FM7Nh5flT42_XXACxv2ZTsABnF44i7ELbjVZA7AzfXP-o8siCLCH-SsITmqp3Fx1JfMRegzAjflRBmHf3tV_1386UcC7JvgNC2sADaQemUpGRIjwyM-RU8hZ-hJFth6LHyo7xpz6YYwlmUyCLpol_5iBnD2EGRu8yvcNM8eUUJjsFlHX_SjNaDCXuJwjI48poChFEcKYdImpgsbmT26npHMp7CyxcowQbVw_vJhM8dF3iNNJw3ef_Ev3Hywhv1H7Or6BQEPi53Ij9cAOMMW86nkHbUdnBV1RipCtm9NzKdgemBNeTxSeU-XQyi7qihUwQkTpSjMyk3PWzB_)　  
といったあたりでしょう。  
Pangenomeについては：[https://www.ncbi.nlm.nih.gov/pubmed/32034321](https://www.ncbi.nlm.nih.gov/pubmed/32034321?fbclid=IwAR2gnijVrbicXw8ugEAu1t0zA4qPV5BXYQlrAAHu4XkHpCykuscKrcpbiUM)

NGS-mappingの基本パイプラインの図はこちらから：[https://genomics.sschmeier.com/ngs-mapping/](https://l.facebook.com/l.php?u=https%3A%2F%2Fgenomics.sschmeier.com%2Fngs-mapping%2F%3Ffbclid%3DIwAR19uYnQJlIQ5LYFbLzsf9KWzPDS6W7aeEl35oa_cv8Jg2Qr6z36r_50QqA&h=AT1cBPChVP0V6Q8CbFrgCYdmUtQCGm4NNinu4FEe7x0YuET93fCaVqksTYKB1eu9Shlg6U5c74xlfzE0vB_P2N_ckkuRXoFeOYYFcvgMW04mDF9OEeegLcAM2CSy8gvDheEm06DfMqCIzslCTdvjR-1B8G-KbgeVGs9qhMRepYDnDvBsAdim1MlnMBHE9p8iQMum06KZerO3kWZKs5CGX5ad_3v64RaXdcxDl-JZsJYt242fqOaUnFKBttuY-FyMw24P6U4PRUNq_mHuuuumSXkSCh4V7k8qwLTs6rVJffCXKHGcDZr7Hd4SBXr-lkJbDEYFkFYXArDvUNMvo8RkumHuxz2vTnih0dWtWb6Mi8xjJbSvxq4RTh0VoQ0LdelKsshy-MVzj_R9su_jOF7DIo0VLYxe8pkQuGb2t9JzWR5i-kQ7FMkJHBN-_gGcWDe8v3jaKtPAfc-_vzVyK5jYHnTRD9_uzp1W8LUIDSiid6J8HLl61_vlt3X_TGuCtVsjcxoKDhC7PGzkV0tn2eva7z0jmpBPcNaXRG8qshrSpyrNc-u6yowYiQW0O2w2EGcbXV65osxY7nEry5-atQfnvp9m9K70j2tl6IHU-KRmT5W-9ap9BoReOfEO2rMk-cf3M9dKeSLfZi02Ck-H)

1. Main analyses
   1. Two types of approaches

Every item is studied.

In omics studies, we treat many genetic variants, long sequential molecules, many genes, many molecules, many biomarkers many cells and many individuals. One way to manage these many items is to investigate each element one by one and to obtain many pieces of outputs. This approach includes GWAS where many SNVs are studies one by one and transcriptomic RNAseq analysis where all coding genes are quantified.

Whole items are evaluated for their patterns.

In the other approach, all items are handles together to extract meaningful messages. This approach includes clustering samples with expression profiles and principal component analysis to identify informative components that are consisted of weighted sum of individual items.

　個別の解析をたくさんやること

　　個々の解析には、検定・推定・学習・ベイズ

Four different approaches to extract messages from a data set コラムにしてもよいかも

There are four ways to handle an omics data set to say something meaningful. The contents of this section is true for any kind of data analyses and is not special for omics data analyses. The first approach is statistical test that reject the null hypothesis and to support the alternative hypothesis. The second is to estimate meaningful relation or to generate a predictive model. The third is to identify a pattern in a data-driven fashion. The fourth is to use a data set as information to update prior belief to posterior belief in Bayesian framework.

1. Statistical test

Scientific studies try to find novel things that we find difficult to believe without supportive evidences. In that case we use statistical tests that measure the probability to observe the evidences is too low if we don’t believe the novel finding but believe the hypothesis that deny the finding, null hypothesis. The rarity of the evidences when we believe the null hypothesis is quantified as p-value. When you test association between genotypes of a genetic variant and dichotomous phenotypes, the null hypothesis is that genotypes and phenotypes are mutually independent and p-value of its independence test measure the rarity of the table observation if the null hypothesis is true.

1. Statistical estimation and machine learning of predictive model

The independent test of the above-mentioned genotype-phenotype table can be used to estimate genotypic risk ratio (GRR), that should be one if no association between genotypes and phenotypes, but GRR deviates from 1 if null hypothesis should be rejected. GRR stands for relative risk to develop a particular phenotype of one genotype against other baseline genotype. Although we want to know the true value of GRR, it is always impossible as far as we try to know it with limited amount of information and we have to estimate the value. Sometimes one representative value for it is estimated, that is called “point estimate”, and sometimes a range within which the true value is believed to exist with residual uncertainty, that is called “interval estimate”. In this case, we estimate GRR, but when two variables, X and Y, are both continuous, the deviation from independence between X and Y is measured as linear regression slope coefficient in the simplest model. Again this coefficient should be estimated and its representative point value might be estimated and its interval might be estimated. Both GRR and slope coefficient are quantified effects or effect sizes. When we estimate effect size of something, we take the position where we believe there exists sizable effect rather than the position of no effect (null hypothesis). Of course the estimated effect is small enough and estimated effect size can be compatible with null hypothesis. Although these GRR and linear regression coefficient are simple examples of estimation based on a data set, they are based on an assumed model and enable us to predict value or probability of new samples who lack observation of a part of variables in the model. In this sense, they are simple cases of supervised machine learning.

1. Descriptive statistics and unsupervised learning to identify pattern

Supervised learning in the previous section is the machine learning task to generate a particular model to describe the relation of input and output with training samples and the generated model should work to predict output for new input. As stated here, it requires training samples who have both input and output that should be the answer of the prediction model. On the contrary, in some contexts, multiple variables of multiple samples are observed and all observed data records have some noises and no “answer” is available. In this situation, main interest in the data set is to extract particular patterns in the data set itself, or to identify deviation from randomness among the variables and samples.

1. Bayesian approach
2. ●個別解析を積み上げる vs. データ全体を押しつぶす
3. オミックスデータ解析では、多数のアイテム(ゲノム座位・遺伝子・分子等)を対象にして、多数のサンプルを調べるのが普通です。  
   必然的に高次元データが得られます。
4. この高次元に対して大きく2つのアプローチ。  
   （１）　個々のアイテムごとに調べて、アイテムの数だけ出力値を得てから考える  
   （２）　全体を圧縮機に放り込んで、解釈可能な要素での説明を抽出する
5. 具体例としては、（１）はＧＷＡＳにおける個別ＳＮＰ関連検定、遺伝子ごとの発現2群間比較検定が含まれ、（２）はトランスクリプトームデータ全体からのクラスタ検出、PCA/非線形次元圧縮視覚化(tSNE等)などの次元圧縮手法、既存知識に基づくアイテムのグループ化・構造化による対象アイテム数の縮減法、などがあります。
6. （１）の全アイテム個別解析アプローチには、必然的に２つの解釈上の課題が生じます  
   >>マルチプルテスティング問題。これには、  
   >>>>(i)閾値設定をして白黒２分する方法とそれに伴う解釈問題があります。多数アイテムのスクリーニング課題ともいえます  
   >>>>(ii)閾値設定をせずに、アイテム数の値の集まりを解釈する考え方があります。こちらは、スクリーニングするというよりも、現象全体をどうとらえるかという解釈課題になります  
   >>オミックス研究は本来、すべてのアイテムを合わせて全体としてどうなっているかを志向しているので、個々のアイテムに分離してしまったのでは、オミックス的とはいいがたいです。個別に分けてしまった上で、オミックス的に解釈するには、「組み合わせ」問題と取り組む必要があります
7. （２）は大きく「次元圧縮」とくくってもよいでしょう  
   >>次元圧縮が有効なのは、「データレコード」は高次元空間に置かれているが、それらはバラバラなわけではなく、何かしらより低次元な部分に収まっているはずだ、と考える場合です  
   >>高次元空間にあるけれども、回転しさえすれば低次元空間に収まるという期待があれば、PCAが有効ですしMDSも有効です  
   >>回転しただけでは収まらないが、伸ばしたり縮めたりすれば低次元的に捉えられるのであれば、非線形次元縮約が有用です。よい感じの回転を見つけるのは、方法として画一的なのに対して、非線形変換の方法は流儀次第なので、非線形次元縮約は方法の選択依存性があることを意識して解釈する必要があります。  
   >>また、「真のデータ全体を高次元のまま『見る』ことができない」ため、どの非線形変換が最適かということは、データからは解らないと覚悟を決めておくのが適当です。何度も何度も、あるタイプのデータを取ることによって、観測データに向いている方法が解ってくることはあります。  
   >>しかしながら、オミックススタディはそのようなことがわかる前に実験技術が進歩して、さらに大規模かつ複雑なデータを扱う方法の提案を求められる状況が続いているので、解析ツールは百花繚乱です　。ある程度の手法のパフォーマンス比較情報は発表されているので、それらを参考にしながら、手探りで解析手法を選ぶ・開発する必要が続いています。  
   >>なお、次元圧縮するときには、オリジナルのデータよりも少ない尺度を使って説明することになります。その尺度同士が、情報的にかぶりがない～独立である～ように取り出すことは「過不足のない分解」として数学的に好まれます。正規直交基底分解と呼ばれる方法がその例です。説明次元の縮約・データ分解の手法について考えるときには、情報的に過不足ない(キレイな)分解を目指したものなのか、そうではないのかは一つの大きな着眼点です。過不足のある(キレイでない)分解の代表例は、遺伝子を機能で分類して、遺伝子の亜集合ごとに重みを計算したりする方法です。このような遺伝子の分類・分解は数学的に「キレイ」である保証がないですが、それぞれの亜集合の生物学的な意味は解りやすいですから、総合的な解釈という観点からは有用性が高いことも多いです。このように分解・圧縮する場合でも、目的によって何が適切かは変わります。
8. ●　このほかに、次元は下げてみたものの、対象とするデータが、扱いやすい分布として低次元空間に存在しているとは限りません。そのような点を解析対象にするには、アイテム（遺伝子など）がどのような遠近関係になっているのか、その遠近関係は、連続的に広がっているのか、クラスタに分かれているのか、つながってはいるものの、木のような形なのか、ネットワークなのかなども、問題になってきます。対象が作る構造の「位相～トポロジー」の解釈が残っている、とも言いかえられます。
9. scRNAseq ：[https://link.springer.com/article/10.1186/s13059-019-1898-6](https://l.facebook.com/l.php?u=https%3A%2F%2Flink.springer.com%2Farticle%2F10.1186%2Fs13059-019-1898-6%3Ffbclid%3DIwAR1WJMgOHzUX91UUQue8Hb8aeCh6n29bbHqc1rsYqV8smACDoixbDTFDEoE&h=AT0AL0wT-Xk7odbjVQQvlf5BVuWp7FY_ibMji-iomQD5dkukuCG_WmVbbsQeVPTInArnWuyGh1wQMSLwFTQ7fjLlZ_zwUDDuAuxs3xOro4pt3hdp6qMGqJdTaKskMJqxXiBOSq2naaQVSR9lzXtQgI9aiVfEe8m5CS4oc6g4vDgnpeoPMoerclrQuVXLQhbdsxPz61Qc9TyA4BGXBqK8hxr97mlD4I7M-pegljnZ-vWYzlyCa1NdR-SBN5BKIrzbz-3wDuPsbbF6AqUiP6QQCatraHNkAad2jmf9w1PaLIlVBNzZTR5_2DB2oyNiAOWfPIaoPQ8WIvTSh4bduRosqiFRGbz5ReJ47XU5ZHrRJVyE0NUpPK63LmuDfOXwFyOmk24_FIn4nmUTh1BbzDp-sR1uiEzHyMWTm8KZ69qrYOnGYgdM7PhNndcjiT2Ja2HPmSq1BM6d6cobCuqhuyiF_5iXQxTdkOL_gAUQE4qFpvw-j56ErkY3XZrWZDOtkN9_hp2FjL-xLR6PL0FUYTJqGllAd3rNSxyySBZdzkC1fOH2ixHscZfjVjrftkhmPCaEd_V_4SxyQMrJKiEqQPDEELgHqABYu8wn3x2Bf-6e_mrvFQTVGbiZ_TEtn-I-_qj_yogq1Q" \t "_blank) [https://www.nature.com/articles/s41467-019-13056-x](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.nature.com%2Farticles%2Fs41467-019-13056-x%3Ffbclid%3DIwAR17msJFWogAuvlFzO7BPrAwOV7cNqUHMFb76FkRP1e76aS7-GRpyLo20bc&h=AT2R7iyToJt8o_zZlWevftsxllp99UvEteWqQr5O7yPEG7C2boFI7P6QKDpvcH0cVWdjeBuubMHiC1CaDkWenyyVUDog9Ylca9OzMohu2MeRelgTTS60WIPz-MFfJrgH-cu2JiT55YgXFtOxwD02kGUMXPYRTl7eFoED4B3BPoYaRcgszeRbFMDYlM-iIah6reW4SKEQloYetLdmcCQZcJvLOqzohLju1LScs-ah2UgedsfpMYxL47gZdxWrrj_UCf56rpovi3zsVqJq2cmb_ojpvck7RgK6WoWL24Je2S5EXL720DIgb1GejTueb9aoWKKwdBFZ679L25g1AQ5vwaOB72aU7-FwQAbYZ3_FyGorMtJzj7AJoG1kr1u80V0cvoERINRDkWOXhAIUoI_fecu_QvQwpEa6iyO9efBMZqnUrWiPth5umaRzBM7B4kuRJeaub7O7z5eZToPHi6F2BPwIuxzogcSxx01wZ9jEzADXzzS2Cac6_VqAx5K4kqBkkcFJ105YtE4Wgm-poYL15JYakuGImWxveuruIWgO5jm3t0RlwbwfdMA3v-rEY0Q7bUTqio5qfWkyWPlkmKYCDCb62mF12AqI8GJobiwRO7TFt7jmcX8e2goaPKvjgbyE_x-VWg) [https://www.embopress.org/doi/10.15252/msb.20188746](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.embopress.org%2Fdoi%2F10.15252%2Fmsb.20188746%3Ffbclid%3DIwAR27mryv4ybnXYwlecWOfbIn2nFD2udk-D3hmj9ZMCriumEPutMtbILh1j8&h=AT3mthXZA_Vo1qEMj3TEpOPiebNieJMom4FZv5k2C9Fn9DixfVxJKyQ2KQ_jfOn0W0INo0Imp2VwQ-bSnea4ORFkm1grIeO722vwEnJl9HRRhiWpbNFUiKucPDdRtLP1BSoZHqCX64793Tr8NZfmnHLCouvee64RU2OjXSXqCvM8yG_PVnVi-JVG-7kQxA0OlX9-l8H05UK1rwuKfw634yFHISgCrtpceJ6mzOHxPn_r9y88yTA2fPqJVFGIdWJ4pLJf7IJif2HBv9g_79ICjh4RbMGxdawyvBSd69bT97ddLfdbdCjcSS2qzrtMVha2TcF5_4e1knlPj-3jDr2FMvY-j58i3_8dHY7_Vt45wza2dpFDAF01Zc87CSbf9_jc_IRHXZpfl9g8h4litiItVrIvPbSzqxvfY8DXxWQ2PJlqQUCYuSspRp--R35V8zImjttkluYX-1wf3pY5RwBgCig6nVXfdNcy75uJBshsGicglMSW7D_qYjW8vAkbR0Aocj-Us0WaIYx_HSrfhXtjtUqi-hiCpfvZRRsZo4tLJgGswiSHwKENOwxNfhucqTs8P0QzOpU38I8lAtH3LJ4Fl8zXPepljkoA8lk8shU4DE6220zmvfdLcNty49LrnjkzUOS-oQ)  
   たくさんの手法の探索サイト：[https://omictools.com/](https://omictools.com/?fbclid=IwAR3pCvPQi5NMibdo4arQ0Wy7iUZaBwonz7Uq8sPYMDlDx-NFGCRzpzJzyVA)　　  
   トポロジー：[https://academic.oup.com/bioinformatics/…/34/17/i988/5093232](https://academic.oup.com/bioinformatics/article/34/17/i988/5093232?fbclid=IwAR0rOlSreiThScbZSlsCqz86wDzLB-w1WJDFS8ZD_fBO3AEGEb2j75O8sgY)
   1. Multiple testings
10. [オミックスデータ解析とその結果の解釈：その５]  
    マルチプルテスティング：たくさんのp値の解釈
11. たくさんの検定をしたら、「たまたま、そのうちの１個やそこらの検定」は、ふつうのタイプ１エラー基準での棄却水準を下回るp値が出ることは、まったく珍しくありません。  
    したがって、たくさんの検定をしたときには、どのくらい小さいp値ならば帰無仮説を棄却すればよいのかを考え直す必要があります。これがマルチプルテスティング補正です。
12. ● GWASでたくさんのSNVについて個別に検定して得られたp値は、ボンフェロニ補正・FWER補正などを用いて解釈します。100万SNVを検定したら、0.01 \* 10^(-8)程度のp値を得て初めて、タイプ１エラー0.01程度で棄却するという補正法です  
    ● 癌部と健常部の遺伝子発現の違いを検定すると、遺伝子の数だけp値が出ます。このときは、FDRを使うことが多いです
13. この違いは何なのか。
14. ● GWAS-ボンフェロニ/FWER補正の方では、「たくさんのSNVを検定するけれど、そのすべてが帰無仮説であるのではないか」という前提で補正をしています。そしてその補正は、最小p値を出したSNVにも、2番目、3番目・・・、何番目に小さいp値を出したSNVにも同じルールで補正をします。
15. ● 他方、癌部・健常部の遺伝子発現比較においては、「全遺伝子のうち、相当数の遺伝子に発現の差があるだろう」との前提で補正をしています。ですから、最小p値を出した遺伝子については、棄却検定の基準として厳しいものを適用しますが、いったん、その遺伝子の帰無仮説を棄却したなら、「この遺伝子には帰無仮説は成り立たない」ことを是として、2番目に小さいp値を出した遺伝子のp値の小ささを評価します。帰無仮説検定をするべき仮説の数（遺伝子の数）がすでに１個減っていますから、マルチプルテスティングの棄却基準は少し緩くなります。同様に、n番目に小さいp値を出した遺伝子にはn番目なりに緩い棄却基準を適用するのが妥当となります。この考え方を反映した方法がFDRです。FDRでp値を補正したものをq値と呼びます。
16. ● p値分布が一様からずれる場合
17. これとは別に、オミックス解析における多数のp値補正で知っておく必要があることがあります。GWASでは、集団構造化によるたくさんのp値が小さめに出てしまう場合の補正がその例になります。
18. 帰無仮説棄却検定のp値というのは、帰無仮説の下でデータを取ってp値を算出することを繰り返すときに、p値の分布が一様になるように作られています。
19. しかしながら、何かの要因があり、ほとんどすべての検定で帰無仮説が真であると信じるべき状況なのに、p値が小さめに偏ることがあります。集団が不均一で、不均一な集団からケースとコントロールをサンプリングしたものの、ケースがある亜集団の割合が高く、コントロールが別の亜集団の割合が高いと、亜集団間の違いのせいで、小さいp値が多発します。その影響は全検定・全p値に及びます。  
    このようなときに、「観測したp値セットは一様分布になっていないけれど、一様分布になっているべきなので、そのように補正する」ことがあります。
20. そして、このように補正した上で、さらにマルチプルテスティング補正をする必要があります。
21. ● p値/q値でポジティブシグナルを拾うこととは～スクリーニング
22. 状況によってp値を使うかq値を使うかの違いはあるものの、どちらも、棄却水準を定めて、帰無仮説を捨てて、違いがあるという対立仮説を信じてもよいだろうアイテムを選択するという作業に使います。  
    しかしながら、p値/q値とその閾値との設定は、必ず、ある割合でのFalse Positive(FP)を許容し、ある割合でのFalse Negative(FN)という取りこぼしを前提にしています。  
    したがって、p値/q値を用いてマルチプルテスティング補正した上で、アイテムのスクリーニングをするときには、目的さえはっきりしており、それがFP,FNの割合にどのように影響するのかを理解した上でであれば、棄却水準は目的に合わせて上下することは間違いではありません。  
    大事なことは、補正後p値/q値で主張したいことは何なのか、活用する理由は何なのかを明快にして、それに合わせて使うことです。
23. 実際、スクリーニング目的においては、p値/q値のみならず、エフェクトサイズとp値/q値を組み合わせて、取捨選択することもあります。ボルケーノプロットと呼ばれるプロットが意味していることはそういうことです。
24. [https://yimasato.com/r-volcanopot](https://l.facebook.com/l.php?u=https%3A%2F%2Fyimasato.com%2Fr-volcanopot%3Ffbclid%3DIwAR3pLToscdoicq_-j53NVkIBNfG-yWy7Cvssaq5nnBVQ3FdSW0eFb2S63g4&h=AT13gJSOOVYjRMlDZyqKII5izxNfc2umwdvXbgf8uo0gOTg3zWuNkHm2j9QBDEftsmqNoYites6_DgZpyh-GJ-0qu9zrM2wFYlFDE2Xa2q6Hv_3PeDyG_FbzJ_2xcG98KvAS0-qLhkqFSTOtl95KGKcwga_GKOWZFt8N5fMGjjNF_HGOEKv7-h0Cz1F41eRhxSKJKxMp58qboxtX6Zlpq6pMNdLURPdcU6VyNZ6JSYxS2DAyWakFOO0Wab6VBi_rtnBtuOkgMgrzGkqWfml1b6wRDbG2SIz1iAmMuavPInoh19IjKktGhQiYRO2SUICG0QCKPKUBG9qxKEHem09x1fU6pI0wwB434o6S9fuvS2tKaj_VXXCOaYIoJuSmPDD6EiPx7zJ-KwUSiqJ1W--UVN87ZYiV5Oc-PTkW--RVc0Sx-fFskLFFqPJnf2KUxi27pLa77Nr0anuUhhz2D5wLbaEYD3TXnD8D8jqXQTYB-v6UQJymbqAkdChPbk9TzoWmhyDavRfB6LApOjQionqBikxe6_mmprrxL5k9erZphleRj-QtQu-9dYbX7ptGmibKH66GAWB9R4xbIWqOGPByCjz9MvT8ahplX_PxmoH8sr5AOdqe2yCsl8c0JYPgePm5POOqXw)
25. オミックス研究で出てくる大きなデータの塊から、Take-home-messagesを取り出すときに必要なステップや考え方を整理する時間として使ってみることにします。

まずは、全体の構成がどうなるかをまとめてみます.  
大学院の講義でバラバラに扱っていることを整理しつつ、関連文献などに当たりながら進めることになりそうです。

●本解析前  
>>データの質管理  
>>取捨選択と捨てればよいわけではないこと  
>>バッチエフェクト  
●本解析  
>>アプローチ分類  
>>結果のOmics的補正  
>>データ全体の簡略化：次元削減・分類・構造化  
>>パラとノンパラ →　囲み記事  
●「人間的解釈」の道具  
>>言葉の置き換え：オントロジーとかパスウェイとか  
>>視覚化（という、強力だが読者任せな提示法)  
>>因果関係  
●たくさんあるツール・たくさんあるオプション  
>>どれも正解  
>>メッセージの「信じ方のコツ」

[オミックスデータ解析とその結果の解釈：その２]  
●本解析前  
>>データの質管理

[オミックスデータ解析とその結果の解釈：その４]

コラム  
>>パラとノンパラ  
ここで「ノンパラ」には２つの意味合いがあって、どちらも、オミックス解析では重要なので、その区別をしておくのがよい。

本当のところは「ノンパラ」とは「パラメタで(単純に)定義された分布関数を仮定しない」という意味では同じだが、使われ方に２通りあるので、その2通りを区別して、「ノンパラ」の名の下に整理しておくとよい。

（１）仮説検定をするときにノンパラ手法を使う、という意味のノンパラ。パラメタは使わない、という意味で、「ノンパラ」。パラメタを使わない代わりに順序などを使う。  
（２）モデル推定における、「パラ」と「ノンパラ」は意味合いが異なって見える。かなり単純なパラメトリックな分布であると仮定してその分布のパラメタ値を推定するのが「パラ」。そうではなくて、データ量・標本数が多くなるにつれて、複雑で柔軟なモデル推定をするのが「ノンパラ」。こちらは、パラメタの数をデータ量・サンプル数の増加につれて、モデルに使用するパラメタ数を増やせる。究極的には無限個のパラメタで複雑な表現を目指すという意味で「(少数の)(有限個の)パラメタに制約されない」という意味での「ノンパラ」。ちなみに、威力が華々しい深層学習などは、標本を増やすほどどんどん複雑なモデルも推定可能になるという点でノンパラ  
>>パラメトリック推定はそもそものモデルが単純なのでオーバーフィッティングの心配がほぼないのに対して、ノンパラメトリック推定はモデルがいくらでも複雑になりうるのでオーバーフィッティングが問題になる  
>>オーバーフィッティングは、モデルの単純さ・複雑さと、推定される具体的なモデルのバイアス・バリアンス問題と絡むので、それについては、項を改めて説明する必要が出るだろう

Rを使った、パラ推定・ノンパラ推定、バイアス・バリアンスの練習はこちらを：[http://ryamada22.hatenablog.jp/entry/20200218/1581987977](http://ryamada22.hatenablog.jp/entry/20200218/1581987977?fbclid=IwAR0_VdBo8ncp15LfEGYork8KP8mR41t_JxB9Hl7VlXNfjXrIuiNIq4351ZM)

（１）のノンパラ検定手法についてはこちらを：[https://kaigyou-turezure.hatenablog.jp/ent…/2018/…/16/141219](https://l.facebook.com/l.php?u=https%3A%2F%2Fkaigyou-turezure.hatenablog.jp%2Fentry%2F2018%2F08%2F16%2F141219%3Ffbclid%3DIwAR1SSl8Q7aMjlFicMqV8TZApm7w7Q2JYOV1yZDkU3S3Ip6_RlAUw1O6IQAU&h=AT35AX8wyd3KmqqwG5mMYDjBIjqPcfNFywp7MmTMciIYhmCn3lsaBVR5taEWJCdWs4QPEbuMR6AaNzw3PiuLvDewhBtCPid2UZJBsFFmHTopdS6fMW8AVLh8tbhJErtZ5SSC-wQZDDgaMrdSh5TyvnKktVGpwEadiEAWvZdM9RZjPG0vHaHCbywoRuYka9BXBOjSf_0VDqLKAhP8qANZhiGg4d-CBgHy2V67axNaQOyLk9VGmHhS12vLMyBglF3KEldfmbmDkczSoZJ5_GdhEIERq8SRE3U07yKcas-QqirRiwrz4OZk9LGoZOmdBvfao4sGn2JSZwNiRBP_AJcylbOQ0QuqMVNTjMggZkKtE_YAHmZgB9G0NCr6Eh88J5EKRbzsjjRO90urzDWLuY5QYWVEw9T3oEexfxNsiGbxI9nxiwc-v76PBHMMLnwQjtY8pQj3NjxU3dP8eFlUppafL79GAW14SGlGT4Ko8OqEZt7oMlT4DVjIXnw1zLjgyohtgDrr0g-qw-X4bxoGnDKfwPUAHhd_jH-vQmmL1JaVEC915u30IDcjOS4S4vyKLKFKm0L72CiJmkPdV0DNBJgiz8uwbbpVyY5LcSeaE8Vedm35aAPTqQi7zQ-xPzSHV6cCqhSGlaQvBro223hj)  
（２）のノンパラモデルについてはこちらを：[https://www.sciencedirect.com/…/mathema…/nonparametric-model](https://www.sciencedirect.com/topics/mathematics/nonparametric-model?fbclid=IwAR3H1TT2xXWFSo8MK9LVgKJIu9t3A-5BgWZ7mK5d0kmH67HxnWH3N-WuO-M)

1. Translation of the outputs into perceptionable forms

The raw outputs of statistical and machine learning methods are essentially in the form of numbers, formulae and symbols. The vast majority of people consider the expressions with numbers formulae and symbols are inappropriate as the ways of “interpretation”. Intellectual interpretation of inputs is rooted on the physical systems of perception. We, humans, have five senses and

オミックスデータ解析の結果を「人間的に解釈」するとき、数字と記号だけでは「わかった感じがしない」です。  
人間には五感がありますが、その中でも、(知的)コミュニケーションに使う感覚と言えば、聴覚と視覚です。  
聴覚的な解釈法は「言葉・文章」化することに相当します。  
視覚的な解釈法は「2次元の図」化することに相当します。

「言葉」にするために、オミックス・バイオインフォマティクスの世界ではアノテーションとか、オントロジーとか言うものを使います。ゲノム上の位置に遺伝子名を付与するのはアノテーションですし、化合物の名称やドメイン名もアノテーションです。複数の遺伝子をある機能を共有するという理由で亜集合にまとめるのは遺伝子オントロジーによる亜集合の作成です。アノテーションもオントロジーも人間界の用語として定義や説明がつけられる単語でできているので、オミックスのデータの結果は、アノテーション・オントロジーで意味を調べることができる単語での説明に変換されることによって「人間的解釈」をしたことにすることが多いです。

ただし、注意するべきなのは、実験とデータ解析によって示された事象をアノテーション・オントロジーの用語に変換することによって、ずれや誤解が生じる可能性がゼロではないことです。平易な例としては、あるSNVがある遺伝子構造上にあるとのアノテーションがついていたからと言って、そのSNVが関連している形質と、アノテーションで示された遺伝子とその形質が関連しているとは言い切れないことです。そのSNVはその遺伝子の5'上流の転写開始点から比較的近いところに存在していたからと言って、その遺伝子の発現量に影響を与えているとは限らず、なにか別の機能を持っているかもしれません。このようなとき、SNV->アノテーションで示された遺伝子という変換が、解釈上の誤解を生むことになります。  
そのような可能性に常に留意している限り、アノテーション・オントロジーを用いた人間的解釈は非常に有用です。

もう一つの視覚的解釈法はどうでしょうか？  
そもそも、2次元で表しえない多次元の情報を何らかの処理を介して、2次元の絵にすることは多いです。PCAで上位２軸を用いてひょじしたり、t-SNE表示をしたり、ネットワークというアイテム同士のつながりの様子を視覚的にグラフとして描図したりします。所詮2次元は高次元を表しえない非常に簡略化された空間なので、真実をすべて読み取ることはできません。2次元視覚化の手法は、高次元データの是非とも表したい部分情報を2次元の描図領域の位置や長さ、広がり、離れ具合として強調して描いた「デフォルメ絵画」のようなものです。したがって、何が強調され、何は描かれずに捨てられたのかを理解したうえで、2次元視覚化図は解釈するのが適切です。  
また、図の読み取り方は人によって千差万別です。それは絵画鑑賞の感想文が人によって違うのと同じ理由です。写実画の感想文は多くの人が同じように書くかもしれませんが、抽象絵画の感想文はバラバラになることは想像に難くないでしょう。  
したがって、図で解釈を伝えるとき、伝える側は、その図をどのように「読む」のかをコントロールして読み手に自身の伝えない内容が伝わるように説明する義務があります。逆に、読み手は、手法の特性(デフォルメされるルール)を理解し、その特性に合わせて、読み取る術を学んだうえで「読む」努力をする必要があります。

フォームの始まり

フォームの終わり

1. Many methods and many tools.

There are many tools for omics studies. Because of variation in methods of raw data production variation of purposes of studies, たくさんのツール・たくさんの手法

もしくは、みんな違ってみんないい

世の中にはデータ解析の統計手法・機械学習手法がたくさんあります。オミックスの解析手法もたくさんあります。

どのようなことを知るときに、どの手法がよいだろうか、との疑問に答えるために、手法を比較した論文もたくさんあります。

したがって、ある目的には、この手法がよさそうだ、というある程度のコンセンサスはあります。

しかしながら、どの手法も、データ全体を説明しない、という点では共通であり、その意味では、どの手法も不完全です。  
また、インプットを受け取り、アウトプットを返すことができますから、使えないわけではありません。「ある目的のためにとの設定の下で」メリットが多めでデメリットが少な目な手法を選ぶのがよいというだけのことであり、そのメリット・デメリットの多寡はあくまでも相対的なものです。

簡単な例では、明らかに、1次線形直線に乗らないように見えるデータが得られたときに、1次線形回帰をするのは誤りなのか、という例を取ってみましょう。

確かに、二次曲線に合致しそうなら、二次曲線回帰をする方がよさそうです。しかしながら、1次線形直線回帰をして傾きが正であるという結果を得ることに意味がないか、と言えばそんなことはありません。二次曲線回帰の結果から「傾きが正」という「言葉的な解釈」を引き出すには、二次曲線の増加部分のみにデータ点があることを付け加える必要がありますが、1次直線回帰であれば、係数の符号を見ただけで増加傾向であるという「言葉的な解釈」ができます。

実際、オミックススタディのようにデータの全体を俯瞰する方法すら持たずに行っている解析のときに、ある手法が、上述のように「曲がっているのに直線近似」をしているのか、そうでないのかを知ることは難しく、すべての方法は、かなり不確かな仮定に立脚して解釈を引き出しています。

そういう意味で、どのツール・どの手法がよいか悪いかは、特定の目的に照らして比較検討はできるものの、すべての目的のために絶対的によい手法があるとは思わずに、常に、手法には限界があることを知ったうえで用いることが大切で、そこから得られる「解釈」もその限界の下で引き出されたものであることを忘れないことが「解釈法」の中で最も重要なことと言えるかもしれません。

[https://omictools.com/](https://l.facebook.com/l.php?u=https%3A%2F%2Fomictools.com%2F%3Ffbclid%3DIwAR0tLBomY-1--JWwjajquye3iceeHxFEflpvDApZoKHBpsjRn7qyPhFfYhc&h=AT1r2lu2TDPWeK6K-oY0apz3U__EjpjZU_gA8uJJbCGESaCfvjGqnItC-j3MtbcPBHi7BvcjDW1CVDy5Kx0Zsyq9LDg9pDgFcGwT7RX_P1klFOdYfuwn4MvXUXfv1zsdLt6HMfjLWx4a11ZLkE1TdmQ9K3IBoHE-YY-KLHw1l01F8_ea5H0sB7bcehSvguHr6mubC4vNteT24Uhk4EGXkZnNGamUGBbUGKNf52ZVOVuGdxFrNLOZipIKoGvR840MX0drVssllktB_CeuNxk-jpFGN5WPHsLE9sj0jXH_DM-UXjp1QbEW-FPfLtLk8Bakmesdl7tXLRNNqW0IXk4psX2h_4ZjbEfe9VL9PFX5ZMB8TQe-nf-RyB16i9Jl1GrtrDBGhCGb4en5D53MOJhOXkxSDR-E6UWOkcNudsu8m2UgSwBI28Sk3xlXHaqjLckLvCOdnbGotbUSfTJ6yWNRMeeswB5hj60-2AHhEaz_xLAVln8uiYAteiE1orL7_G-6A7HkbMtIG9o7dp2LJilStc799fSD1cAOW6a34s4eiOQzc9XwQwyDqNCbt6Kf-EModniApEywdyc-kx5hv8vcRqnG6BKr1rygzslJDQr1goHYi7lpc4M)