JHG review

Title

Characteristics and interpretation of human omics data

Abstract 250 words

Max 5000 words, including abstract and excluding refs and figs and tables)

Tables and Figs max 8

Refs max 100

オミックスとは、オミックスのデータ解析とは

データが一時にたくさん取れる

　そのために実解析に入る前の前処理が必要

　前処理にも統計的に考えることがある

　　捨てればよいというものではない

　　たくさん、一度に取ったがために使える情報がある

　　使える情報は、結果の解釈に確率的に効いてくる

　ＱＣ、バッチエフェクト

解析には大きく２つ

　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

　　個々の解析をたくさんやったら、個々の解析結果の解釈とは異なる考え方

個々の結果の全部を見渡して考え直すこと

マルチプルテスティング

個々の結果は、全体の分布の影響を受けて解釈すること

順序をつけること（ボルケーノ）

　全体を解析すること

　　高次元・広い空間

　　次元削減・軸を見つける・説明のための組み合わせを見つける

　　　何かを捨てている、捨てている部分が「観測誤差」であれば成功

　　分類する

　　木にする

　　ネットワークにする

パラとノンパラ

解釈の道具

　人間の言葉への置き換え　ＧＯ、パスウェイ

　視覚化

　因果関係

ツールがいっぱい・オプションもいっぱい

　「正解」はない

　これも正解、あれも正解

　単純か、複雑か。単純なら、「そのものが真」と信じるのではなく、「その傾向」は多分本当、と信じる。複雑なら、「精度よく信じてもよい」かもしれないけれど、大きく踏み外しているかもしれない、との陥穽がある可能性込みで信じる。

　　オミックスとは、オミックスのデータ解析とは

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).

図１　オミックス層

データが一時にたくさん取れる

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.

　そのために実解析に入る前の前処理が必要

　前処理にも統計的に考えることがある

　　捨てればよいというものではない

　　たくさん、一度に取ったがために使える情報がある

　　使える情報は、結果の解釈に確率的に効いてくる

　ＱＣ、バッチエフェクト

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.

解析には大きく２つ

Two approaches to analyze omics data

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

　　個々の解析をたくさんやったら、個々の解析結果の解釈とは異なる考え方

個々の結果の全部を見渡して考え直すこと

マルチプルテスティング

個々の結果は、全体の分布の影響を受けて解釈すること

順序をつけること（ボルケーノ）

　全体を解析すること

　　高次元・広い空間

　　次元削減・軸を見つける・説明のための組み合わせを見つける

　　　何かを捨てている、捨てている部分が「観測誤差」であれば成功

　　分類する

　　木にする

　　ネットワークにする

パラとノンパラ

解釈の道具

　人間の言葉への置き換え　ＧＯ、パスウェイ

　視覚化

　因果関係

ツールがいっぱい・オプションもいっぱい

　「正解」はない

　これも正解、あれも正解

　単純か、複雑か。単純なら、「そのものが真」と信じるのではなく、「その傾向」は多分本当、と信じる。複雑なら、「精度よく信じてもよい」かもしれないけれど、大きく踏み外しているかもしれない、との陥穽がある可能性込みで信じる。

オミックス解析とは

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).

図１　オミックス層

Collection of raw data and pre-processing of them before data analysis

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.

Quality control and batch effect

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.

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.

図２　ＦＡＳＴＱでのリードのクオリティ、リード集合でのクオリティ図

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

オミックスデータの特徴。その取得実験との関係も含めて

　生データから、解析用データへの変換

高次元

　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.

データの塊を解釈するということそのものについて

　全体を知る

　個々を知る

　個々を全体の文脈で知る

解釈の流儀、３体

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

オミックスだと、たくさんの仮説、たくさんの推定をするので…

　仮説の検定と推定

　データ・ドリブン

　事前仮説を使う

　パラとノンパラ、ノンパラには２つの意味合いがある

　バイアス・バリアンス

　白黒をつける

　グレーに解釈する・分布を答えにする

　状態全体が「答え」

スクリーニングとｐ値とｑ値と点推定と区間推定

ボルケーノプロットの意味すること

解釈の方法論

　個別を評価して、それらを組み合わせる

　　単純に組み合わせる

　　意味を考慮して組み合わせる（パスウェイ）

　個別の組み合わせを抽出する

　　分解する・きれいに分解する

　　きれいに分解するために工夫する

　　意味を大きく取り出す・固有値

　　ノイズとして切り捨てること

一筋縄で行かない情報を捨てていないかを気にする

　位置・順序・時間・ペア

解析オプションがあるということは、解析に正解がないということ

解釈の補助

言葉・機能　ＧＯ

　視覚化・見て安心する

　　木とネットワーク

わかることと、わかった気持ちになること

予測モデルの良しあし

　同じ評価値でも、中身は違う。誰が外れるかは変わってくる

　予測性能の評価にも立場がある。感度・特異度

　手法がたくさんある

ＱＣ

バッチエフェクト

現実的なことも書く。具体例を図表で示す

フローを示す

留意点・take-home messagesをつける

ワード数例　１２２ワード

Only articles that have been published or accepted and waiting for publication (listed as ‘in press’ following digital object identifier number) should be in the reference list. Reference to ‘unpublished data’, ‘personal communications’ and database should not appear in the list but should be cited in the text parenthetically only (e.g. Smith A, 2007, unpublished data). Authors must obtain permission from the individual concerned to quote his/her unpublished work. Written proof for ‘personal communication’ and preprint for ‘in press’ may be requested for review. Abstracts may be cited only if they are the sole sources, and must be identified in the reference list as ‘(Abstract)’. Journal names are abbreviated (with full stops) according to common usage; refer to Index Medicus for details.