原 著

エクソンレベルでのエンリッチメント解析における エクソン数に起因するバイアスの補正

Correction of the Bias Derived from the Number of Exons in Exon-Level Enrichment Analysis

法隆大輔・林 武司

Daisuke Horyu and Takeshi Hayashi

農業・食品産業技術総合研究機構
中央農業総合研究センター

NARO Agricultural Research Center

e-mail:horyu@affrc.go.jp

Enrichment analysis is an effective tool to interpret gene expression data generated by high throughput genomic analysis. Enrichment analysis can be applied to differential exon expression data generated by RNA-seq. However, the enrichment analysis based on differential exon expression introduced unknown exon number bias. Gene ontology terms related to the genes that have larger number of exons tend to be more significant in the enrichment analysis. To correct this exon number bias, we adopted two approaches. First approach is to apply non-central hypergeometric distribution to test significance of each term. Second approach is to count exon number instead of gene number. We confirmed that exon number bias had been adjusted by non-central hypergeometric distribution approach.

Key words: enrichment analysis, non-central hypergeometric distribution, exon number bias, RNA-seq.

1. はじめに

RNA-seq やマイクロアレイなどの技術によって、ゲノム全体にわたる遺伝子発現についての情報が容易に得られるようになった。このような技術によって得られた大規模な遺伝子発現データの生物学的な意味を解釈する手段の一つとして、GO (Gene Ontology) や KEGG (Kyoto Encyclopedia of Genes and Genomes) などのデータベースを利用したエンリッチメント解析 (Khatri and Drăghici 2005; Huang et al. 2009) がある。エンリッチメント解析は、データベースの中の特定のカテゴリに関連付けられた遺伝子に着目した時に、確率的に想定される以上に多数の遺伝子において発現量に差異が見られるかどうかを検定する方法である。ここで言う特定のカテゴリとは GO で言えばターム、KEGG で言えば代謝経路のことである。もし、特定のカテゴリに関連付けられた

遺伝子のうち、想定される以上に多くの数の遺伝子に発現量の差異が見られれば、そのタームや代謝経路と発現量の差異の間に偶然では考えられない何らかの関わりがあるとエンリッチメント解析では考える。これにより、単なる発現量の差異という現象に対して、GO のタームや KEGG の代謝経路を通じた意味づけが可能になる。発現量の差異が見られる遺伝子のリストから GO のタームや代謝経路に変換されることによって、それがどのような生物学的プロセスや分子的な機能と関係があるかが分かるようになり、人間による解釈が可能になる。

エンリッチメント解析は大規模な遺伝子発現データに解釈を加えることのできる有効な手段であるが、RNA-seq のような新しい技術によって得られたデータに適用しようとした場合に、技術的な特性に由来する問題が生じることもある。エンリッチメント解析を行うためには、遺伝子の発現量を推定し、その発現量を元に 2 つの条件の下で発現量に差異のある遺伝子を特定しておくが必要になる。しかし、RNA-seq を用いて遺伝子の発現量を推定する場合の問題として、サンプルごとにライブラリのサイズが異なることによる総リード数の違い、GC 含量によるリード数の変動、遺伝子の長さによる読み取り数の偏りなどが知られている (Dillies et al. 2012)。これらの問題の中でも、エンリッチメント解析を行う場合に特に問題となるのが、遺伝子の長さの影響である。

RNA-seq では mRNA の全体からシークエンシングのための cDNA を作成するため、シークエンサーで読み取られるリードの数は mRNA が長い遺伝子ほど多くなる。この結果として、リード数を元にして発現量を推定すると、同程度の発現量の遺伝子であっても mRNA が長い遺伝子ほど発現量が多いという誤った結論に到達してしまう。このような mRNA の長さ、すなわち遺伝子の長さの影響を取り除く簡単な方法として、シークエンサーで読み取られたリード数を遺伝子の長さで除すことが行われる。例えば、標準化された発現量の基準として頻繁に用いられる RPKM (reads per kilobase of exon model per million mapped reads; Mortazavi et al. 2008) は、リード数を遺伝子の長さとしリード数の総量で割ったものである。しかし、発現量の差異を検出する状況の下では、リード数を遺伝子の長さで割っただけでは、検定の際の検出力が遺伝子の長さの平方根に比例してしまい、結果として長い遺伝子の方が発現量の差異が検出されやすいという問題があることが知られており (Oshlack and Wakefield 2009)、遺伝子の長さで割るだけで十分とは限らない。エンリッチメント解析は、発現量の差異が検出された遺伝子の数を使って行う解析であるので、遺伝子の長さによるバイアスがそのままエンリッチメント解析にも反映されてしまう。すなわち、長さの長い遺伝子と関連付けられた GO のタームの方が、発現量の差異が検出された遺伝子の数が多くなりやすく、その数が想定される以上に多いと判断されやすくなるという結果を招く。こうした問題に対処するため、発現量の差異を検出する際に補正を行う方法 (Oshlack and Wakefield 2009; Bullard et al. 2010) のほか、エンリッチメント解析での検定に非心超幾何分布を用いて補正する方法 (Young et al. 2010; Gao et al. 2011) が提案されている。

ここまで述べてきた内容は全て遺伝子単位での発現量を元にした議論である。しかし、RNA-seq はゲノムシークエンス上にマップされたリードの数をもとに発現量の推定を行う手法であるため、RNA-seq のリード数からは遺伝子単位での発現量だけでなく、より細かいエクソンなどのレベルでの発現量の推定も可能である。実際に RNA-seq のデータから、エクソンのレベルで発現量の

差異を検出するプログラムも開発されるようになっていく。エクソンレベルでの発現量の差異をもたらす原因は、mRNA 前駆体に対する選択的スプライシングである。RNA のスプライシングは、DNA から転写された mRNA 前駆体の中からエクソンなどを切り出して結合する現象である。1 つの遺伝子からスプライシングを経て切り出された mRNA のパターンは 1 種類とは限らない。スプライシングが行われる位置や組み合わせの異なる複数の mRNA (アイソフォーム) とその転写産物を生じることがある。このような現象を選択的スプライシングという。選択的スプライシングは多細胞の真核生物では一般的に見られる現象であり、トランスクリプトームの多様性に寄与している。

エクソンレベルでの発現量の差異が検出できるようになったことで、エクソンレベルでの発現量の差異を元にエンリッチメント解析を行うこともできるようになった。例えば、Maxwell et al. (2012) では線虫が休眠状態から回復する際のエクソン等の発現量の差異をもとにエンリッチメント解析を行っている。また、Pimentel et al. (2014) では前赤芽球と生染性赤芽球の間での選択的スプライシングの差異をもとにエンリッチメント解析を行っている。他にも、骨髓異形成症候群に関わる遺伝子の発現の有無によるスプライシングの差異 (Dolatshad et al. 2015) や U2AF タンパク質の発現の有無によるスプライシングの差異 (Kralovicova et al. 2015) など、様々な場面でエクソンレベルでの発現量の差異を元にエンリッチメント解析が行われている。このようなエンリッチメント解析では、そこで起きている選択的スプライシングの変化がどのような意味を持つかという解釈を助けることができる。しかし、遺伝子の長さによるバイアスと同様のバイアスが、エクソンレベルでのエンリッチメント解析においても存在する可能性がある。1 個のエクソンあたりに換算した選択的スプライシングの数は、遺伝子あたりのエクソン数が少ない遺伝子で大きい傾向があるものの、遺伝子あたりのエクソン数が増加してもあまり低下しない (Pan et al. 2008)。このため、遺伝子あたりのエクソン数が多いほど、遺伝子あたりの起こり得る選択的スプライシングの数が多くなり、結果としてエクソンの発現量の差異が検出されるケースが多いのではないかと疑われる。これはすなわち、エンリッチメント解析で遺伝子を単位としてカウントする際に、エクソン数が多い遺伝子ほど発現量に差異のあるエクソンを含む遺伝子としてカウントされやすいことを意味する。こうしてエンリッチメント解析を行うと、エンリッチメント解析で検出されたデータベースの項目、すなわち GO のタームや KEGG の代謝経路が単にエクソン数の多い遺伝子と結び付けられている項目であるという可能性がある。しかし、このようなエクソン数の影響はこれまで認識されておらず、エクソンの発現量の差異をもとにしたエンリッチメント解析でエクソン数によるバイアスを考慮した例は無い。もし、エクソン数によるバイアスが存在するとすれば、それを考慮せずに解析することは誤った解釈を導くになりかねない。すなわち、エクソンの発現量の変異の持つ意味を解釈しようとしているのに、単にエクソン数の多い遺伝子に関連した生物学的プロセスや分子的な機能にすり替わる可能性がある。このような点を踏まえ、本稿ではエクソンの発現量の差異をもとに遺伝子数を用いたエンリッチメント解析を行う場合に、遺伝子あたりのエクソン数によるバイアスが存在するかどうかを検証する。また、そのような遺伝子あたりのエクソン数によるバイアスが存在する場合の対処法として、遺伝子の長さによるバイアスでも用いられる非心超幾何分布によるエンリッチメント解析を適用するなどして補

正を行い, その効果について検討する.

2. 方 法

2.1 発現量の差異の検出

エクソンレベルでの発現量の差異を元にしたエンリッチメント解析を行うためには, 最初にどのエクソンに発現量の差異があるかを明らかにする必要がある. このようなエクソンごとの発現量の差異を検出する方法として, DEXSeq (Anders et al. 2012)を用いた. DEXSeq では, エクソンごとの発現量が負の二項分布に従うものとして, 一般化線形モデルによる当てはめを行って, 個別のエクソンの発現量に相当するパラメータを入れたモデルがそれを含まないモデルと比べて有意に当てはまりが良いかどうかを検定する. 検定の結果は, エクソンごとの q 値として提示される. 本研究では, 過剰出現 (overrepresented) の場合の q 値を対象として, この q 値が 0.05 以下のエクソンを発現量の差異があるエクソンとした.

DEXSeq には留意すべき特性がいくつかある. まず, DEXSeq では線形モデルの中に条件ごとの遺伝子全体の発現量に相当するパラメータを個別のエクソンの発現量に相当するパラメータとは別に入れている. このため, アイソフォームの構成に変化が無く, 単に遺伝子全体の発現量が増えているだけの場合は, 個別のエクソンの発現量のパラメータはモデルに必要なく, エクソンに発現量の差異があるとみなされない. なお, エンリッチメント解析を行う場合に, 遺伝子全体の発現量に差異がある遺伝子と個別のエクソンの発現量にのみ差異がある遺伝子を一括して解析すべきという意見もある (Wang and Cairns 2014). しかし, そのような場合であっても, 部品としてエクソンの発現量の差異を組み込む方法を採用するのであれば, エクソンのレベルでの解析の段階でエクソン数の問題を考慮しておく必要があることに変わりはない. 次に, DEXSeq では, 手法の制約からエクソンが 1 個しかない遺伝子の場合, このエクソンの q 値は値が NA (統計パッケージ R における定数で Not Available を意味する) となってしまう, 有意にはならない. したがって, エクソンが 1 個しかない遺伝子を解析に含めることは, それ自体がエクソン数の少ない遺伝子は有意になりにくいというバイアスの原因となってしまうため, 今回の解析からは除外した. また, DEXSeq はゲノム中である転写産物と別の転写産物をコードしている領域が重複している場合に, これらを一つの転写産物として扱う場合がある. このような場合には, 両方の転写産物に関連付けられた GO のタームをこの一つの転写産物に対応するものとして扱った.

2.2 発現データおよびマッピング

バイアスを補正する効果の検証を行うため, 3 件の実際の RNA-seq による遺伝子発現データを使用した. 1 件目のデータとして, Tet1 というタンパク質が遺伝子発現に及ぼす影響を調べた結果 (Huang et al. 2014)を用いた. Tet1 は, DNA のメチル化を制御しているタンパク質の 1 つであり, 特にエピジェネティクスの観点から研究されている. このデータ (以下, Tet1kd とする) は, マウスの幹細胞において Tet1 に対応する遺伝子を働かないようにしたサンプルと標準のサンプルの間で, RNA-seq により遺伝子の発現量を比較したものである. 2 件目のデータとして, 主に哺乳類の組織ごとのトランスクリプトームの違いを検証したデータ (Merkin et al. 2012)を用いた. このデータは, 5 種類の生物の 9 種類の組織を対象として RNA-seq を行った結果を含んで

表 1. 使用したデータの概略
Tet1kd Brain-Liver L2L3

文献	Huang et al. (2014)	Merkin et al. (2012)	Garcia et al. (2013); Graveley et al. (2011)
生物種	Mus musculus	Mus musculus	Drosophila melanogaster
GEO ¹⁾ の ID	GSE50198	GSE41637	GSE49587
比較する組み合わせとデータ数	Tet1 ノックダウン(n=4)/標準(n=4)	脳(n=3)/肝臓(n=3)	2 齢幼虫(n=9)/3 齢幼虫(n=8)
解析対象とした遺伝子数 ²⁾	24,828	25,071	11,434
解析対象としたエクソン数	355,302	345,421	67,427

1) www.ncbi.nlm.nih.gov/geo.

2) 解析対象とした遺伝子とはエクソンが2個以上ある遺伝子で、ここで言う遺伝子には非コードRNAも含まれる。

いる。本研究では、このうちマウスの脳と肝臓のデータ(以下、Brain-Liver とする)を比較して、エクソンごとの発現量の差異を検出して検証に利用した。3 件目のデータとして、ショウジョウバエの 2 齢幼虫と 3 齢幼虫のトランスクリプトームを比較した結果を用いた。このデータ(以下、L2L3 とする)は、Graveley et al. (2011) において収集された RNA 配列の読み取り結果を利用して、Garcia et al. (2013) がマッピングと発現量の差異の検出を行ったものである。これら 3 件のデータの概略を表 1 に示した。

3 つのデータのうち、Tet1kd と L2L3 については DEXSeq によってエクソンごとに発現量の差異を検出した解析済みのデータが GEO (www.ncbi.nlm.nih.gov/geo) からダウンロードできるため、この解析済みのデータを利用した。2 件目のマウスの脳と肝臓の比較のデータについては、Tophat2.0.9 を用いてゲノム(mm9)にマッピングを行い、DEXSeq1.6.0 を用いて発現量に差異のあるエクソンを検出した。なお、解析済みのデータにおいても今回解析を行ったデータにおいても、microRNA や long intergenic non-coding RNA などの非コード RNA も DEXSeq での発現量の差異を検出する対象に含めている。ただし、次に述べるエンリッチメント解析では GO のタームと関連付けられていない転写産物は解析の対象とはならない。

2.3 エンリッチメント解析

エンリッチメント解析の対象となるデータベースとしては GO を利用した。GO のタームと遺伝子の間を関連付けるアノテーションのデータは、Tet1kd と Brain-Liver については MGI (Mouse Genome Informatics) による 2014 年 5 月 1 日のバージョンを利用した。また、L2L3 に関しては FlyBase による 2014 年 5 月 13 日のバージョンを利用した。エンリッチメント解析の方法としては、発現量の差異についての検定で p 値がしきい値以下の遺伝子またはエクソンのリストを作成した上で、GO のタームごとに検定を行う方法を用いた。これは Huang et al. (2009) の分類で言えば singular enrichment analysis にあたり、エンリッチメント解析としては最もシンプルな方法である。また、検定の方法としては、リストに含まれる GO のタームと関連した遺伝子またはエクソンの数が超幾何分布することを利用する方法を採用した。 N をゲノム中で対象となる全ての遺伝子またはエクソンの総数、 T をゲノム中において今調べようとしている GO のタームに関連

付けられた遺伝子またはエクソンの総数, n を発現量に差異がある遺伝子またはエクソンの数とする. このとき, 今調べようとしている GO のタームに関連付けられていて, かつ発現量に差異のある遺伝子またはエクソンの数が t である場合のエンリッチメント解析の p 値は,

$$p = \sum_{i=t}^{\min(n,T)} \frac{\binom{n}{i} \binom{N-n}{T-i}}{\binom{N}{T}} \quad (1)$$

となる (Eden 2009). また, エンリッチメント解析では, 多数のタームあるいは代謝経路についての検定を同時に行うことになるため, 多重検定の補正を行う必要がある. この補正は, Benjamini and Hochberg 法 (Benjamini and Hochberg 1995) による FDR (False discovery rate) によって行い, FDR を 5% 以下としたときに有意なタームをエンリッチメント解析で検出されたタームとした. これと同時にタームごとの q 値も計算した. なお, GO のタームは Biological Process, Cellular Component, Molecular Function の 3 つの区分に分けられているが, 全ての区分のタームを区別せず一括して解析の対象とした.

2.4 非心超幾何分布を用いたエンリッチメント解析

本研究で問題にしているエクソン数によるバイアスと同様の問題として, 遺伝子の長さの違いが原因となって, 遺伝子によって発現量の差異の検出されやすさにバイアスを生じる状況がある. このような場合において, これまで多く用いられている補正の方法は, エンリッチメント解析での超幾何分布による検定の代わりに非心超幾何分布による検定を用いる方法である (Leong and Kipling 2009; Young et al. 2010; Gao et al. 2011). 超幾何分布は壺から赤玉が取り出される確率と白球が取り出される確率が同じ場合の取り出された赤玉の数の分布であるのに対して, 非心超幾何分布は赤玉が取り出される確率と白玉が取り出される確率が異なる場合の取り出された赤玉の数の分布である. 非心超幾何分布を用いることで, 赤玉すなわち特定の GO のタームに関連付けられた遺伝子と, 白玉すなわちそれ以外の遺伝子の間で発現量の差異が検出される確率が異なる状況においても検定を行うことができる.

非心超幾何分布には, 標本の抽出を順次行う場合に相当する Wallenius 非心超幾何分布と標本の抽出を一括して行う場合に相当する Fisher 非心超幾何分布の 2 種類がある. バイアスがある場合の遺伝子を対象としたエンリッチメント解析の手法では, Young et al. (2010) と Gao et al. (2011) が Wallenius 非心超幾何分布を用いているほか, Leong and Kipling (2009) は Fisher 非心超幾何分布を用いており, どちらの分布も使われた例がある. 発現量に差異のあるエクソンや遺伝子を検出する場合は, それらのエクソンや遺伝子が一括して検出される状況に近いと考えられるため, 本研究では Fisher 非心超幾何分布を用いた. Fisher 非心超幾何分布を用いたエンリッチメント解析の p 値は, 非心パラメータを θ として,

$$p = \frac{\sum_{i=t}^{\min(n,T)} \binom{n}{i} \binom{N-n}{T-i} \theta^i}{\sum_{j=1}^{\min(n,T)} \binom{n}{j} \binom{N-n}{T-j} \theta^j} \quad (2)$$

である。なお、非心超幾何分布の p 値の計算には R の BiasedUrn パッケージを用いた。

非心超幾何分布を検定に利用するためには超幾何分布に必要なパラメータに加えて、非心パラメータを与える必要がある。非心パラメータは壺から赤玉が取り出される確率と白玉が取り出される確率のオッズ比に相当する。エンリッチメント解析の場合では、非心パラメータは特定の GO のタームに関連した遺伝子とそれ以外の遺伝子における発現量の差異の検出されやすさの比として理解できる。これまでの研究では、例えば Gao et al. (2011) では非心パラメータとして遺伝子の長さの平方根の比を用いている。具体的には、 N をゲノム中で対象とする全ての遺伝子の数、 M を対象とする GO のタームに関連付けられた遺伝子の数、 d をシーケンサーによる読み取りの長さとした時、 $\frac{\text{median}(\sqrt{L_i - d})_{1 \leq i \leq M}}{\text{median}(\sqrt{L_i - d})_{M < i \leq N}}$ を非心パラメータとするエンリッチメント解析を提案している。本研究では、まずどのエクソンも一定の確率 q で発現量の差異が検出されると仮定する。このとき、 k 個のエクソンを持つ遺伝子で少なくとも 1 個以上のエクソンで発現量の差異が検出される確率は $1 - (1 - q)^k$ となる。対象とする GO のタームに関連付けられた遺伝子の平均のエクソン数を k_1 、それ以外の遺伝子の平均のエクソン数を k_2 として、非心パラメータ θ を、

$$\theta = \left\{ \frac{1 - (1 - q)^{k_1}}{(1 - q)^{k_1}} \right\} / \left\{ \frac{1 - (1 - q)^{k_2}}{(1 - q)^{k_2}} \right\} \quad (3)$$

というオッズ比として定義した。

2.5 エクソン数を用いたエンリッチメント解析

バイアスの原因は、エクソンの数が多い遺伝子であっても少ない遺伝子であっても、遺伝子単位で式 (1) の N や T の値をカウントしていることにある。一つの考え方として、遺伝子を単位としてカウントするのではなくエクソンを単位としてカウントすることで、遺伝子あたりのエクソン数によるバイアスを解消できる可能性がある。このような考えのもと、 N や T の値としてエクソン数の平方根を用いてエンリッチメント解析を行うことを試みる。このとき、 N や T は整数でなければならないので、小数点以下は切り上げる。また、式 (1) の t の値には発現量の差異が検出されたエクソン数をそのまま用いる。ただし、個別の遺伝子について見たとき DEXSeq で発現量に有意な差があるエクソン数が、その遺伝子のエクソン数の平方根を上回る場合は、 t の値としてエクソン数の平方根を上限とした。エクソン数が多い遺伝子でも、発現量の変化しているエクソンが多くなければ、 N や T が増加するほどには t が増加せず、 p の値が小さくなりにくい。また、 N や T の値としてエクソン数の平方根を用いることには、次のような意味がある。起こり得る選択的スプライシングの数を、発現量の変化しているエクソン数の上限として捉えれば、上限は全エクソン数よりは小さいと考えられ、全エクソン数をそのまま用いるのは必ずしも適切ではない。そこで、1 遺伝子あたりのエクソン数と発現量に変化のあるエクソンの割合の関係をグラフにすると、図 1 のようになる。1 遺伝子あたりの発現量に差異のあるエクソン数の上限を n_{UL} とし、1 遺伝子あたりのエクソン数を n_{EX} とする。発現量に差異のあるエクソン数の上限を遺伝子あたりのエクソン数の平方根に設定するということは $n_{UL} = \sqrt{n_{EX}}$ とすることである。この式を発現量に差異のあるエクソンの比率とエクソン数の関係として見ると、

$$\frac{n_{UL}}{n_{EX}} = \frac{1}{\sqrt{n_{EX}}} \quad (4)$$

図 1. 1 遺伝子あたりのエクソン数と発現量に変異のあるエクソンの割合

図中のそれぞれの点は各遺伝子(非コード RNA も含む)に相当する. また, 曲線はエクソン数を用いる方法での発現量に差異のあるエクソン数の上限を示す. 図中の縦軸で用いた発現量に変異のあるエクソンとは DEXSeq で検出されたエクソンの割合であり, 遺伝子全体に発現量の変異がありアイソフォームの構成に変化が無い場合などは含まれない. また, データ間での遺伝子数の違いは, 生物種の違いおよび解析済みのデータであるか本研究で新たに解析したデータであるかの解析環境の違いによる.

という式になる. この (4) 式を図 1 の中に表示すれば曲線のようになる. 発現量に差異のあるエクソン数, すなわち図 1 中の点の上限に沿うように曲線が引かれている. このグラフから, エクソン数の平方根は発現量の変化しているエクソン数の上限として利用できることが分かる.

アイソフォームの種類によっては, エクソンの途中までを利用する場合や逆にエクソンを延長する場合, イントロンを除去しない場合などがある. このため, どこまでが 1 つのエクソンかを決めることができない場合も多く, 遺伝子ごとのエクソン数は自明ではない. 本研究では, 遺伝子ごとのエクソン数の定義として, DEXSeq が定義するエクソン数を用いた. DEXSeq ではなく何かのアイソフォームによってエクソン, イントロン境界として利用される境界を全て考慮した上で, 境界の間にあるエクソン領域は 1 つのエクソンとしてカウントされる. このため, エクソン数は比較的多くなる傾向にある. また, 発現量に差異のあるエクソン数は, 単純に DEXSeq で差異があるとされたエクソン数を用いた.

3. 結 果

3.1 エクソン数とエンリッチメント解析で有意なタームの割合の関係

初めに, エクソンの発現量の差異をもとにして遺伝子数によるエンリッチメント解析を行った時, 本当に遺伝子あたりエクソン数によるバイアスがあるのかどうかという点について検証しておく必要がある. このために, 補正を行っていない遺伝子数によるエンリッチメント解析, すなわち遺伝子中に 1 つでも発現量に差異のあるエクソンがある遺伝子の数を (1) 式の n および t としてエンリッチメント解析を行った. その結果について, エクソン数とエンリッチメント解析で検出されたタームの割合の関係を図 2 に示す. この図で横軸に示されている GO のタームあたりの平均エクソン数は, GO のタームごとにそのタームに関連づけられた遺伝子のエクソン数の平均を算出した上で, 等間隔のピンにまとめたものである. 平均のエクソン数で同じピンの中に入った GO のタームのうち, エンリッチメント解析で検出されたタームがどのくらいの割合を占めるかを示したのが縦軸である. もし, 遺伝子あたりのエクソン数の影響を受けていなければ, 縦軸すなわちエンリッチメント解析で検出された割合は横軸に示された平均のエクソン数に関わらず一定のはずである. ところが図 2 を見ると, 例えば Tet1kd の場合エンリッチメント解析で検出された割合はエクソン数 30 くらいまでは増加を続け, それ以上では大きく変動するようになる. 他の 2 つのデータでも傾向は同様である. これらの結果は, エンリッチメント解析で検出される GO のタームの割合は, 遺伝子あたりのエクソン数の影響を受けていることを示している. このように, エクソンごとの発現量の差異を元にして, 補正を行わないまま遺伝子数を用いてエンリッチメント解析を行うとエクソン数によるバイアスの影響を受ける, すなわちエクソン数の多い遺伝子と関連付けられたタームがエンリッチメント解析の結果として検出されやすいことが明らかとなった. 冒頭で挙げた Maxwell et al. (2012) や Pimentel et al. (2014) の例を含めて, エクソン数を意識せずに行ったこれまでのエンリッチメント解析では図 2 と同様にエクソン数によるバイアスが生じていることになる. このようなバイアスを補正するためには, 何らかの形でエクソン数の情報をエンリッチメント解析に反映させる必要がある.

図 2. エクソン数と検出されたタームの割合の関係(遺伝子数によるエンリッチメント解析)

GO のタームごとにそのタームに関連づけられた遺伝子のエクソン数の平均を算出して等間隔のビンにまとめ、同じビンの中に入った GO のタームのうち、エンリッチメント解析で有意と判定されたタームがどのくらいの割合を占めるかを示した。ただし、ビンに入るターム数が 50 個未満の点は表示していない。この図から、エンリッチメント解析で検出される GO のタームの割合は、遺伝子あたりのエクソン数とともに増加する傾向があることが分かる。

3.2 エクソン数と有意なタームの割合の関係における補正の効果

2つの方法によってバイアスを補正する効果について、まず図 2 と同様に GO のタームあたりの平均エクソン数とエンリッチメント解析で検出されたタームの割合との関係を見ることにする

図 3. 非心超幾何分布による補正を行った場合のエクソン数と検出されたタームの割合の関係
非心超幾何分布による補正を行ったエンリッチメント解析の結果について、図 2 と同様のプロットを行った結果である。図 2 で見られたようなエンリッチメント解析で検出される GO のタームの割合が遺伝子あたりのエクソン数とともに増加する傾向は、この図では見られなくなっている。なお、エンリッチメント解析で検出されたターム数はそれぞれの方法、それぞれのデータによって異なるため、これらの図の縦軸で示された割合を異なる方法、異なるデータの間で比較することには意味がない。

(図 3, 4). なお、エンリッチメント解析で検出されたターム数はそれぞれの方法、それぞれのデータによって異なるため、これらの図の縦軸で示された割合を異なる方法、異なるデータの間で比較することには意味がない。例えば、Tet1kd のデータセットでは遺伝子数を用いた場合に

図 4. エクソン数による補正を行った場合のエクソン数と検出されたタームの割合の関係
 エクソン数による補正を行ったエンリッチメント解析の結果について、図 2 と同様のプロットを行った結果である。図 2 と同じように、エンリッチメント解析で検出される GO のタームの割合が遺伝子あたりのエクソン数とともに増加する傾向がある。

エンリッチメント解析で検出されたターム数は 955 であるのに対して、非心超幾何分布による方法では 341 であった。全体のターム数はどちらも同じであるので、有意と判定されたタームの割合は両者で大きく異なる結果となっている。もし望ましい補正が行われているとすれば、平均のエクソン数の違いによって有意と判定されるタームの割合が影響を受けないということになり、これらの図では縦軸のエンリッチメント解析で有意と判定された割合は横軸に示された平均

図 5. 平均のエクソン数ごとに見た GO のターム数(データ: Tet1kd)

GO のタームごとにそのタームに関連づけられた遺伝子のエクソン数の平均を算出して等間隔のビンにまとめ、同じビンの中に入る GO のタームの個数を表示している。この図からは、平均のエクソン数が 20 前後の GO のタームが多いことが分かる。

のエクソン数に関わらず一定になるはずである。非心超幾何分布を用いてエンリッチメント解析を行った場合には、図 2 で見られたようなはっきりとした傾向は見られなくなっている(図 3)。一方で、エクソン数を単位としてエンリッチメント解析を行った場合には、平均のエクソン数が増加するにつれて有意なタームの割合が増加する傾向は無くなっていない(図 4)。

以上の結果からは、非心超幾何分布による方法でのみ補正の効果があり、エクソン数による方法では補正の効果が無いように見える。しかし、エクソン数による方法にも改善されている点がある。例えば、図 4 の Tet1kd の場合を見ると、タームあたりの平均のエクソン数が 20 くらいまでは検出されたタームの割合が大きく上昇するが、20 以上の範囲では、そのような上昇がほとんど見られない。一方で図 2 の遺伝子数の場合では、平均のエクソン数が 20 以上の範囲でも検出されたタームの割合が上昇している。ここで、対象となるデータの数すなわち GO のターム数が平均のエクソン数ごとにどのように分布しているかを見ると、その数が最も多いのは遺伝子あたりのエクソン数が 20 前後の範囲となっている(図 5)。このため、データ数の多いエクソン数が 20 以上の範囲で検出されたタームの割合が上昇しないという点で、エクソン数による方法にも補正による効果があると考えられる。

エクソン数による方法で 20 以上の範囲で検出されたタームの割合が上昇しない理由について考える。エクソン数を N や T の値として用いると、平均した遺伝子あたりのエクソン数が多い GO のタームでは、より多くのエクソンで発現量の差異が観測される必要が生じてくる。このため、エクソン数を N や T の値に用いることは、エクソン数の多い遺伝子の影響を抑制する効果を持っている。しかし、逆にエクソン数の少ない遺伝子の影響を大きくする効果はない。このことを考えれば、平均のエクソン数が小さい区間で有意と判定されたタームの割合が少ないのは予想できる結果である。また、エクソン数が 20 を超えた範囲で検出されたタームの割合の上昇が見られないのは、エクソン数の多い遺伝子の影響が抑制されている結果として理解できる。

3.3 q 値の水準ごとのエクソン数から見た補正の効果

エンリッチメント解析の結果がエクソン数の影響を受けているかどうか、また補正によって影響が緩和されるかどうかについて、今度は各タームの q 値の側から見てタームに関連づけられた遺伝子の平均のエクソン数が均等に分布しているかどうかという観点で検証する。もし、エンリッチメント解析の結果がエクソン数の影響を受けていなければ、 q 値に関わらずエクソン数の水準は一定になるはずであるので、この点について検討する。

遺伝子数によるエンリッチメント解析と2種類の補正を適用した場合のエンリッチメント解析の結果について、 q 値の水準とタームあたりの平均エクソン数の関係を図6に示した。横軸は q 値で、0.1 刻みのビンにまとめている。縦軸はタームあたりの平均エクソン数である。横軸のそれぞれのビンに入ったタームの平均エクソン数の分布を箱ひげ図としてプロットしてある。エンリッチメント解析による q 値の水準がタームあたりの平均エクソン数の影響を受けていないとすれば、このグラフは平坦になるはずである。補正を行っていない遺伝子数を用いたエンリッチメント解析では、中央値や四分位範囲が右に行くほど下がる傾向が見られる。これは、 q 値が小さいすなわち過剰出現であるほどタームあたりの平均エクソン数が多く、 q 値が大きいくすなわち過少出現であるほどタームあたりの平均エクソン数が少ないことを示している。このような関係は、図2で見た平均のエクソン数が増加するのにしたがってエンリッチメント解析で検出されたタームの割合が上昇するという結果とも矛盾せず、いずれもエンリッチメント解析の結果として平均エクソン数の多いタームが現れやすいというバイアスがあることを示している。

非心超幾何分布を用いた場合には、補正を行っていない遺伝子数を用いた場合とは逆に中央値や四分位範囲が右に行くほど上がる傾向が見られる。このことは、非心超幾何分布を用いる方法では補正が過剰な面があり、平均のエクソン数が多いタームでは q 値が大きくなりやすいことを示している。エクソンレベルでのエンリッチメント解析において超幾何分布の代わりに非心超幾何分布を用いる場合には、このように補正が過剰になる場合があることを理解しておく必要がある。エクソン数を用いたエンリッチメント解析の場合は他の2つの方法と比べると、中央値や四分位範囲が上下せず、一定に近い傾向となっている。このことから、エクソン数を用いる方法では、 q 値の水準が遺伝子あたりのエクソン数の影響を受けない方向に補正されていて、非心超幾何分布を用いた場合のように補正が過剰になることも無いことが分かる。以上のように、 q 値の水準ごとのエクソン数から見ると、エクソン数による方法で補正を行った場合に良好な結果が得られた。

3.4 q 値が上位のタームから見た補正の効果

ここまでの結果はGOのターム全体を対象として、それらの挙動を見るものであった。実際にエンリッチメント解析を利用する場面では、 q 値の小さいタームのみに関心がある場合が多い。これは、 q 値の小さい方から見て上位に位置づけられるタームを元に、発現量の差異がどのような意味を持つかという意味的な解釈を加えるためである。このような観点から検証を行うため、3件のデータについて q 値が小さい方の上位に位置づけられたGOのタームをまとめたものが表2である。なお、タームに関連付けられた遺伝子の数が少ない場合に、全ての遺伝子において発現量の差異が検出されると、 q 値が0になることがある。このような極端なデータは順位付けからは除外している。

図 6. q 値の水準とタームあたりの平均のエクソン数の関係

各タームを q 値によって 0.1 刻みのビンにまとめ、それぞれのビンに入ったタームに関してそのタームに関連付けられた遺伝子の平均のエクソン数を箱ひげ図として示した。もし q 値が平均のエクソン数の影響を受けていなければ、平坦なグラフとなるはずである。しかし、遺伝子数を用いたエンリッチメント解析では右に行くほど下がる傾向が現れている。

表 2. enrichment 解析で上位に位置づけられた GO のターム

a-1) 非心超幾何分布 (データ : Tet1kd)

順位	ID	Name	遺伝子数による順位	平均のエクソン数	p 値	q 値
1	GO:0005634	nucleus	1	20.95	0	0
1	GO:0044822	poly(A) RNA binding	2	22.69	0	0
1	GO:0003723	RNA binding	4	22.82	0	0
1	GO:0007049	cell cycle	7	22.83	0	0
1	GO:0006397	mRNA processing	10	23.32	0	0
1	GO:0007067	mitotic nuclear division	12	23.02	0	0
1	GO:0008380	RNA splicing	13	23.11	0	0
1	GO:0006351	transcription, DNA-templated	14	20.31	0	0
1	GO:0051301	cell division	15	22.63	0	0
1	GO:0005730	nucleolus	16	22.58	0	0

a-2) エクソン数 (データ : Tet1kd)

順位	ID	Name	遺伝子数による順位	平均のエクソン数	p 値	q 値
1	GO:0005634	nucleus	1	20.95	1.00E-305	9.78E-302
2	GO:0044822	poly(A) RNA binding	2	22.69	1.08E-240	5.30E-237
3	GO:0005737	cytoplasm	3	22.33	8.33E-155	2.71E-151
4	GO:0005515	protein binding	5	22.58	1.53E-136	3.75E-133
5	GO:0003723	RNA binding	4	22.82	4.00E-134	7.82E-131
6	GO:0006397	mRNA processing	10	23.32	1.64E-79	2.67E-76
7	GO:000166	nucleotide binding	6	25.37	7.08E-73	9.88E-70
8	GO:0003676	nucleic acid binding	18	21.50	1.33E-72	1.63E-69
9	GO:0008380	RNA splicing	13	23.11	3.89E-71	4.22E-68
10	GO:0070062	extracellular vesicular exosome	11	22.07	2.17E-69	2.12E-66

表 2. enrichment 解析で上位に位置づけられた GO のターム (続き)

b-1) 非心超幾何分布 (データ : Brain-Liver)

順位	ID	Name	遺伝子数による順位	平均のエクソン数	p 値	q 値
1	GO:0030835	negative regulation of actin filament depolymerization	4	21.75	1.67E-05	0.0126
1	GO:0016272	prefoldin complex	7	17.88	1.65E-05	0.0126
3	GO:0008427	calcium-dependent protein kinase inhibitor activity	74	2	2.49E-05	0.0175
4	GO:0045061	thymic T cell selection	10	19	4.06E-05	0.0265
5	GO:0033561	regulation of water loss via skin	16	14	5.63E-05	0.0344
6	GO:0071493	cellular response to UV-B	16	16.5	7.11E-05	0.0410
7	GO:0044309	neuron spine	16	23.6	1.18E-04	0.0643
8	GO:0008355	olfactory learning	16	17.67	1.31E-04	0.0675
9	GO:0010735	positive regulation of transcription via serum response element binding	74	4.5	1.63E-04	0.0696
9	GO:0010736	serum response element binding	74	4.5	1.63E-04	0.0696

b-2) エクソン数 (データ : Brain-Liver)

順位	ID	Name	遺伝子数による順位	平均のエクソン数	p 値	q 値
1	GO:1902083	negative regulation of peptidyl-cysteine S-nitrosylation	16	50.33	7.46E-21	3.65E-17
2	GO:0071493	cellular response to UV-B	16	16.5	9.57E-19	3.12E-15
3	GO:0070593	dendrite self-avoidance	-	33	7.64E-18	1.87E-14
4	GO:0001954	positive regulation of cell-matrix adhesion	74	26.33	2.04E-17	4.00E-14
5	GO:0034199	activation of protein kinase A activity	74	38	3.16E-16	5.16E-13
6	GO:0004415	hyaluronoglucosaminidase activity	74	15.78	5.59E-16	6.85E-13
6	GO:0030214	hyaluronan catabolic process	43	22.57	5.59E-16	6.85E-13
8	GO:0042592	homeostatic process	16	29.67	8.17E-16	8.00E-13
8	GO:0046811	histone deacetylase inhibitor activity	16	29	8.17E-16	8.00E-13
10	GO:0008104	protein localization	49	27.72	1.92E-15	1.71E-12

表 2. enrichment 解析で上位に位置づけられた GO のターム(続き)

c-1) 非心超幾何分布 (データ : L2L3)

順位	ID	Name	遺伝子数による順位	平均のエクソン数	p 値	q 値
1	GO:0017090	meprin A complex	4	5	7.68E-06	0.00604
2	GO:0004733	pyridoxamine-phosphate oxidase activity	4	6.33	2.23E-05	0.0125
2	GO:0005504	fatty acid binding	4	7	2.23E-05	0.0125
4	GO:0030239	myofibril assembly	2	11.33	3.40E-05	0.0149
4	GO:0043148	mitotic spindle stabilization	14	4.75	3.03E-05	0.0149
6	GO:0019216	regulation of lipid metabolic process	4	8.67	4.99E-05	0.0192
6	GO:0008615	pyridoxine biosynthetic process	14	5.5	5.36E-05	0.0192
8	GO:0045155	electron transporter, transferring electrons from CoQH2- cytochrome c reductase complex and cytochrome c oxidase complex activity	30	2	5.92E-05	0.0194
9	GO:0045214	sarcomere organization	1	16.76	1.22E-04	0.0295
9	GO:0051287	NAD binding	14	7.2	1.18E-04	0.0295

c-2) エクソン数 (データ : L2L3)

順位	ID	Name	遺伝子数による順位	平均のエクソン数	p 値	q 値
1	GO:0045214	sarcomere organization	1	16.76	3.87E-16	1.52E-12
2	GO:0005504	fatty acid binding	4	7	2.22E-14	4.36E-11
3	GO:0005471	ATP:ADP antiporter activity	-	11	1.33E-12	1.04E-09
3	GO:0015867	ATP transport	-	11	1.33E-12	1.04E-09
3	GO:0015866	ADP transport	-	11	1.33E-12	1.04E-09
3	GO:0004687	myosin light chain kinase activity	4	31.67	1.59E-12	1.04E-09
7	GO:0007498	mesoderm development	4	9.29	4.59E-12	2.58E-09
8	GO:0008523	sodium-dependent multivitamin transmembrane transporter activity	204	13.5	1.46E-11	7.19E-09
9	GO:0005811	lipid particle	135	6.39	2.49E-11	1.09E-08
10	GO:0030018	Z disc	3	12.16	3.78E-11	1.49E-08

これらの結果を見ると、いずれの方法でも遺伝子数を用いた場合と比較して、ほとんど順位が変化していなかった。相対的に見て順位の変化が大きかったのは、Brain-Liver のデータでエクソン数による方法を用いた場合である。表 2 の b-2 に示した Brain-Liver のデータでエクソン数による方法を用いた場合の表を見ると、例えば “hyaluronoglucosaminidase activity” と “hyaluronan catabolic process” という、ヒアルロン酸の分解に関わる 2 件のタームが順位を大きく上げている。ヒアルロン酸分解酵素は肝臓で多く発現している一方で、脳ではほとんど発現していないことが分かっている (Triggs-Raine et al. 1999)。このような発現量の違いは、アイソフォームの構成にも影響を与えていると考えるのは自然で、ヒアルロン酸の分解に関わるタームがエンリッチメント解析の結果の上位に位置づけられることは妥当な結果であると考えられる。

また、表 2 の a-1 と a-2 に示した Tet1kd のデータでは特に順位の変動が少ない。上位に位置づけられたタームを見ると、RNA の転写に関わるタームが多く見られ、Tet1 が働かなくなったことにより DNA のメチル化に影響が生じ、それが RNA の転写にも影響を与えたためであると考えられる。この場合のように、選択的スプライシングの変化が特定の現象に関係した遺伝子に偏って起きている場合には、補正を行った場合でも順位に変化が少ないのは当然の結果と考えられる。

4. 考 察

RNA-seq の出現とその解析手法の進展により、遺伝子単位や限られたエクソンだけでなく、全てのエクソンを対象として発現量の変化を検出できるようになった。そのような発現量の変化を解釈する手段として、エクソンの発現量の差異を元にしてエンリッチメント解析を行うことも可能になった。しかし、遺伝子の長さが遺伝子の発現の変化の検出とエンリッチメント解析にバイアスをもたらしたように、エクソン数によるエンリッチメント解析へのバイアスという新たな影響も生じた。本研究では、そのようなバイアスを補正する方法について検討し、効果を確認した。GO のタームあたりの平均エクソン数とエンリッチメント解析で有意と判定されたタームの割合との関係 (図 3, 4) で見ると、非心超幾何分布を用いる方法によってバイアスが補正できることが分かった。しかし、 q 値の水準とタームあたりの平均エクソン数の関係 (図 6) で見ると、非心超幾何分布を用いる方法の場合には補正が過剰になる面があるということも分かった。

本研究で扱ったエクソン数によるエンリッチメント解析におけるバイアスの問題は、遺伝子の長さによって生じるバイアスがその後のエンリッチメント解析にも影響を及ぼすという問題と類似している。遺伝子の長さによるバイアスの補正の方法としては、発現量の差異を検出する段階で補正を行う方法と超幾何分布によって特定のタームと関連する遺伝子が多く含まれるかどうかを検定する段階で補正する方法の 2 通りがある (Gao et al. 2011)。発現量の差異を検出する段階で補正する方法として、例えば Oshlack and Wakefield (2009) では検定の際に検定量を遺伝子の長さの平方根で除する方法を提案している。このように、発現量の差異を検出する段階で補正すれば、エンリッチメント解析の段階での補正は必要ない。一方、エンリッチメント解析の段階で補正する方法は、本研究でも用いた非心超幾何分布を利用する方法がある。このような遺伝子の長さによるバイアスの補正と異なり、本研究で扱ったエクソン数によるバイアスの補正の場合には発現量の差異を検出する段階で補正を行う選択肢が今のところ無い。エクソン数によるエンリッ

チメント解析の場合, もし発現量の差異を検出する段階で補正するとすれば, エクソン数が多い場合に発現量の差異を検出されにくくする必要がある. このような考え方もあり得ることはあるが, 現時点ではそのような手法は提唱されていない. このため, エクソン数によるバイアスを補正するためには, 本研究で示したようにエンリッチメント解析の段階で補正を加える方法を採用必要がある.

本研究ではバイアスへの対処として, 非心超幾何分布を用いる方法とエクソン数を用いる方法の2つの方法を検討した. 非心超幾何分布を用いる方法では, エクソン数によるバイアスを補正する効果が見られたが, 一方で補正が過剰になることも分かった. エクソン数を用いる方法では, GO のタームあたりの平均エクソン数が増加すると, エンリッチメント解析で有意と判定された割合が上昇する傾向は残った一方で, q 値の水準とタームあたりの平均エクソン数の関係ではバイアスが解消される傾向が見られた. 非心超幾何分布を用いる方法では, 非心パラメータの選択次第ではより良好な結果が得られる可能性がある. しかし, 非心パラメータが選択できる柔軟性があるということは, 一方でどのようなパラメータを用いることが妥当なのかという問題がつきまとうことを意味している. これに対して, エクソン数を用いる方法は直感的に理解しやすく, 適切なパラメータを選択するなどの手順が必要無いたことが利点と考えることもできる. エクソン数を用いる方法の問題として, エクソン間での相関の問題が挙げられる. 元々, 超幾何分布による検定を行う前提として抽出される遺伝子あるいはエクソンの間で独立性が仮定されている. しかしながら, 遺伝子間での相関によってそうした前提が成立しないことは以前から指摘されている (Goeman and Bühlmann 2007). 独立性の仮定が成立しないことによる影響として, 超幾何分布よりも裾の重い分布となり, 結果として検定が非保守的となってしまう. 同じアイソフォームに含まれるエクソン同士は連動して発現量が変化するため, エクソン数を用いる方法では検定的前提となる独立性が損なわれている懸念がある. 計算時間の面では, 両者に特段の優劣は無い. 元々, エンリッチメント解析では多数の GO のタームごとに計算を行う関係から計算時間を要している. 両者とも計算の過程は既存の遺伝子レベルでの発現量の差異にもとづく方法と本質的には同じであり, 計算時間の増加はほとんどない.

エクソンごとの発現量の差異は, 複数の mRNA のアイソフォームの間で相対的に発現量が変動していることの結果である. したがって, エクソンごとの発現量の差異を検出することは, mRNA のアイソフォームごとの発現レベルの差異を間接的に見ていることに相当する. このように間接的な指標であるエクソンごとの発現量の差異を利用する理由は, 第2世代シーケンサーを利用した RNA-seq のデータから, どのようなアイソフォームが発現しているかを推定することが困難な問題であるということと, それらの発現量を推定することもまた困難な問題であるということに原因がある. DEXSeq のようにエクソンごとの発現量の差異を検出する手法は, そうした困難を回避するための便宜的な方法と言える. したがって, アイソフォームの発現レベルの推定は困難な問題であるが, 仮にその精度が改善されたとすれば, アイソフォームの発現レベルの差異を検出すれば良い. それが可能であれば, 本研究で取上げたようなエクソン数の問題は生じなくなる. より新しい世代のシーケンサーでは, 長い配列の読み取りが可能なものもあり, そうしたシーケンサーを利用すればアイソフォームの発現量を正確に推定することも不可能ではない

かもしれない。

参考文献

- Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, **22**, 2008–2017.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, **57**, 289–300.
- Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, **11**, 94.
- Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, **14**, 671–683.
- Dolatshad, H., Pellagatti, A., Fernandez-Mercado, M., Yip, B. H., Malcovati, L., Attwood, M., Przychodzen, B., Sahgal, N., Kanapin, A. A., Lockstone, H., et al. (2015). Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. *Leukemia*, **29**, 1092–1103.
- Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, **10**, 48.
- Gao, L., Fang, Z., Zhang, K., Zhi, D., and Cui, X. (2011). Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*, **27**, 662–669.
- Garcia, E. L., Lu, Z., Meers, M. P., Praveen, K., and Matera, A. G. (2013). Developmental arrest of *Drosophila* survival motor neuron (Smn) mutants accounts for differences in expression of minor intron-containing genes. *RNA*, **19**, 1510–1516.
- Goeman, J. J., and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.
- Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature*, **471**, 473–479.
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, **37**, 1–13.

- Huang, Y., Chavez, L., Chang, X., Wang, X., Pastor, W. A., Kang, J., Zepeda-Martínez, J. A., Pape, U. J., Jacobsen, S. E., Peters, B., and Rao, A. (2014). Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, **111**, 1361–1366.
- Khatri, P., and Drăghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
- Kralovicova, J., Knut, M., Cross, N. C., and Vorechovsky, I. (2015). Identification of U2AF(35)-dependent exons by RNA-Seq reveals a link between 3' splice-site organization and activity of U2AF-related proteins. *Nucleic acids research*, **43**, 3747–3763.
- Leong, H. S., and Kipling, D. (2009). Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic acids research*, **37**, e79.
- Maxwell, C. S., Antoshechkin, I., Kurhanewicz, N., Belsky, J. A., and Baugh, L. R. (2012). Nutritional control of mRNA isoform expression during developmental arrest and recovery in *C. elegans*. *Genome research*, **22**, 1920–1929.
- Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593–1599.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, **5**, 621–628.
- Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, **4**, 14.
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, **40**, 1413–1415.
- Pimentel, H., Parra, M., Gee, S., Ghanem, D., An, X., Li, J., Mohandas, N., Pachter, L., and Conboy, J. G. (2014). A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. *Nucleic acids research*, **42**, 4031–4042.
- Triggs-Raine, B., Salo, T. J., Zhang, H., Wicklow, B. A., and Natowicz, M. R. (1999). Mutations in HYAL1, a member of a tandemly distributed multigene family encoding disparate hyaluronidase activities, cause a newly described lysosomal disorder, mucopolysaccharidosis IX. *Proceedings of the National Academy of Sciences*, **96**, 6296–6300.
- Wang, X., and Cairns, M. J. (2014). SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. *Bioinformatics*, **30**, 1777–1779.
- Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology*, **11**, R14.