**サマリー**

がん研究におけるトランスクリプトーム解析は、長年利用されてきたマイクロアレイ技術から遺伝子発現量だけではなく様々な他の情報が取得できるRNA-seqを利用したものが主流になってきている。本稿では我々開発したものを中心に、最近のがん研究におけるRNA-seqデータの解析手法について、1)融合遺伝子の検出、2)スプライシング異常を引き起こすゲノム変異の同定、及び3)遺伝子発現量データの解析の三項目に分けて紹介する。

**はじめに**

2000年初めのマイクロアレイ技術の登場以来、トランスクリプトーム解析はがんの分子レベルでの個性を明らかにしてきた。次世代シーケンス技術の登場により、近年、がん研究におけるトランスクリプトーム解析は、遺伝子発現量定量だけではなく、融合遺伝子、スプライシングバリアント、non-codingRNAの検出もできるRNA-seqを利用したものが主流になってきている。本稿では我々開発したものを中心に、最近のがん研究におけるRNA-seqデータの解析手法を紹介する。

**融合遺伝子の検出について**

ゲノムの構造異常により生じる融合遺伝子はがんの形成において重要な役割を果たすことが知られており，融合遺伝子の同定は，がんの病態の理解，診断，治療のために極めて重要である．近年のハイスループットシークエンス技術の発展により，融合遺伝子を網羅的に検出することが原理的に可能になった．その一方で，ゲノム上の高度に類似した配列の存在による，ショートリードのアラインメントエラーによるアーチファクトと真の融合遺伝子との区別が難しいことなどにより，多数の偽陽性が生じてしまうことが多い．

これまでに，RNAシークエンスから網羅的に融合遺伝子を検出するための多くのソフトウェアが開発されてきている．我々も、Genomon-fusion(http://genomon.hgc.jp/rna/)というソフトウェアを開発・公開し，各種がんゲノムシークエンスプロジェクトにおいて，多数の新規融合遺伝子の検出に貢献してきた[1]．しかし，Genomon-fusionを実行するためは膨大な計算量が必要であり，多数検体を解析するためには甚大な計算リソースが必要となってしまうという問題点があった．

一方で，特にSTAR, MapSplice2, TopHat2を始めとする，多くの有効なRNAシークエンスデータのアラインメントツールが開発されてきており，さらに，これらのツールはアラインメントの際の副産物として，キメラリードの候補リストを産出する．生のキメラリードのリストには，膨大なアラインメントのアーチファクトが含まれており，これをそのまま融合遺伝子のリストとすることはできないが，さらなるフィルタリング処理を加え，より信頼性のあるキメラリードのみを抽出して，信頼度の高い融合遺伝子の候補を抽出する試みが進んでおり，（STAR-Fusion; <https://github.com/STAR-Fusion/STAR-Fusion>）．我々もこうしたアプローチを採用した新たな融合遺伝子のツール，fusionfusionの開発を進めている．

現在のところ，fusionfusionはSTAR, MapSplice2, TopHat2の三種類のアラインメントツールに対応している．これらのツールが出力するキメラリードについて，例えば

1. 切断点付近のゲノム領域を候補chimeric readとそのpairが十分にカバーしているか
2. 同一切断点をサポートするキメラリードのアセンブリにより生成されるコンティグ配列と，他のゲノム領域との相同性

などをチェックして，フィルタリング処理を行っている．さらに，複数のアラインメントツールによるキメラリードのリストを組み合わせる場合には，最後に結果を統合する（図1参照）．

fusionfusionとSTAR-fusion, TopHat-fusion, MapSplice2の独自フィルタリングの結果の比較を行った．これらのツールで融合遺伝子を検出し，またそれに対応する構造異常が全ゲノムシークエンスデータから検出されているかを調べた．検出には，成人T細胞白血病[1]のRNAシークエンスデータ，全ゲノムシークエンスデータを用いた．結果は図２の通りであり，fusionfusionは他のツール群と比較して，感度・特異度のバランスが取れていることが示唆される．

fusionfusion自体は，最終的なフィルタリングをするだけであり，計算時間はさほど問題にならず（一つのアラインメントツールにつき，数十分ほど）、むしろその前のアラインメントに要する時間が問題になる．Mapsplice2とTopHat2に比べ，STARの計算速度は圧倒的に早く（１０倍以上），STARのみの使用であれば，一般的なデータの場合には，アラインメントされていないfastqシークエンスデータから，一時間ほどで最終的な融合遺伝子の候補を得ることが可能である．

fusionfusionはhttps://github.com/Genomon-Project/fusionfusionから自由に利用することが可能である．さらに，STAR及びfusionfusionを実行する一連の流れは，我々が新たに開発しているGenomon2において実装されている．

**スプライシング異常を引き起こすゲノム変異の同定**

シークエンス技術の発展により，がんゲノムにおける変異を網羅的に検出することが可能になり，多くの新規がん遺伝子が同定されたが，その一方で，これまでのアプローチは主にコーディング領域におけるアミノ酸配列を変化させる変異（ミスセンス・ナンセンス変異，欠失・挿入）のみに注目しているものがほとんどであり，その他の種類の変異から生じる機能的な変化についてはまだまだ解析が十分に進んでいない状態である．我々はこれまで，ゲノムシークエンスとトランスクリプトームシークエンスの比較解析により，ゲノム変異が転写にどういった影響をもたらすかについての解析に取り組んできた．これまでの解析結果から，様々ゲノム異常がスプライシングに影響を与えるということが明らかになった．これまで得られた知見を利用して現在は，ゲノムシークエンス解析から得られる変異データ，またトランスクリプトームシークエンス解析から得られるスプライシングのパターンのデータを統合して，スプライシングを変化させるゲノム異常を網羅的に同定するフレームワーク（GenomonSplicingMutation, 仮称）の開発に取り組んでいる．

これまで，スプライシングを変化させる変異としては，イントロン配列の両端2bp（多くの場合AG-GT）における変異（本解説では正準スプライシング変異と呼ぶ）のみを考量することが一般的であった，本解説では，これまでに我々の研究から得られた知見を加え，以下のタイプを考慮に入れる（図？）

1. エキソン-イントロン境界において，イントロンの両端の２塩基（GT-AGモチーフ）以外を巻き込む変異により，スプライシングモチーフが破壊されることによる，splicing site slipping, exon-skip
2. エキソン領域における，新たなスプライシングモチーフ形成による，splice site slipping．
3. イントロン領域における，スプライシングモチーフ形成による，pseudo-exon inclusion．

GenomonSplicingMutationによって最終的にスプライシングパターンとゲノム変異の組み合わせが得られるが，実際にそのゲノム変異が当該スプライシング異常を引き起こしているかを検証することは容易ではない．そこで，本解説においては，ゲノムシークエンスデータとトランスクリプトームシークエンスデータの組みをランダマイズして，全く異なる患者由来のゲノム変異・スプライシングパターンの組みからどのくらいの頻度で，スプライシングを変化させるゲノム変異が得られるかを検証する．

GenomonSplicingMutationを成人T細胞白血病[1]のシークエンスデータ（エキソームシークエンスデータ,RNAシークエンスデータの組みが34検体, コントロールRNAシークエンスデータが６検体）のデータに適用した．スプライシングを変化させると予想される一塩基変異，挿入・欠失は53個（false postiveの推定値: 10），構造異常は205個（false positiveの推定値： 38）と予想された．さらに，一塩基変異，挿入・欠失の中で，正準スプライシング変異でないものは？個あり，これらは通常のアノテーションでは捉えることのできないものである．

これまで腫瘍抑制遺伝子であるTP73において，がん細胞特異的にエキソン２，３のexon skipが生じているトランスクリプトが生じているということが知られており，成人T細胞白血病の全ゲノムシークエンス解析[1]により，これがエキソン２，３を巻き込むlong deletionにより生じているということが解明されていた．これに加えて，今回GenomonSplicingMutationにより，エキソン２，３を巻き込むlong deletionに加えて，スプライシングモチーフ付近のSNV, 83bp deletionがTP73の異常トランスクリプトと関連しており（図５），TP73のexon skipを引き起こすゲノム異常の新たなメカニズムが示唆された．

以上，ゲノム・トランスクリプトのデータを統合的に解析する試みの一つとして，スプライシングを変化させるゲノム変異を網羅的に検出するアプローチを紹介した．今後，他のスプライシング調節因子（スプライシングbranch pointやスプライシングエンハンサー・サイレンサーなど）におけるゲノム変異とスプライシングの関係なども明らかにしていきたい．

遺伝子発現量データの解析

ここからは、がんのRNA-seqで得られる各遺伝子の発現量データの解析手法について述べたい。RNA-seqの配列データから遺伝子の発現量を定量化するものとしてよく使われているものとして　Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/), RSEM (http://deweylab.github.io/RSEM/), eXpress (http://bio.math.berkeley.edu/eXpress/), 最近開発され、計算速度が速く我々のお薦めのsalmon　(http://combine-lab.github.io/salmon/), kalisto (http://pachterlab.github.io/kallisto/) 等が挙げられる。

がんの遺伝子発現量データは一般に遺伝子×サンプルの発現値行列として表現され (例えばある癌腫の分子サプタイプを明らかするために多数の臨床サンプルを解析する場合)、このような形のデータはこれまでにもマイクロアレイにより取得されてきたので様々な解析手法が開発されている。古典的な手法としてクラスタリング、PCA解析等があげられるが、Gene Ontologyやpathway情報等に基づいた遺伝子セットライブラリを利用した解析もよく行われる。遺伝子セットの情報を利用した手法は、個々の遺伝子に注目するより生物学的にも解釈しやすく統計的にも検出量が上がるという利点がある。例えば遺伝子セットライブラリと発現データ及び二群のサンプルラベル（例えば遺伝子Xの変異を持っているサンプルとそうでないサンプル）をインプットとして、二群間で差のある遺伝子セットを統計的にスクリーニングするツールとしてGene Set Enrichment Analysis (GSEA; http://www.broadinstitute.org/gsea/index.jsp)がよく使われる。GSEAでは二群間の発現差で遺伝子セットをランキングし、ランキングの上位または下位に各遺伝子セットが偏っているかを検定する。

GSEAのようなサンプルラベルを用いる遺伝子セット解析法は他にも多数発表されているが、我々はサンプルラベルを用いない、遺伝子セット解析法、Extraction of Expression Module (EEM; https://eem.hgc.jp/) を開発しているので紹介したい。EEMは共通の転写因子群によって制御される転写標的遺伝子群、発現モジュールを発現データから抽出することを目的としており、インプットとして遺伝子セットと発現値行列を必要とする。　遺伝子セットとしては共通のシス制御モチーフの存在やChIP-seq等の実験により発現モジュールとして予測された遺伝子セットを想定している。EEMはインプットの遺伝子セットが発現モジュールとして機能しているとすれば発現値行列中で共発現しているだろうという仮定のもとで、遺伝子セット中に有意に大きな共発現しているサブセットがあるかどうかを検定する。有意であれば共発現しているサブセットを発現モジュールとして抜き出し、サブセットに含まれる遺伝子の発現値の平均を発現モジュールの活性値として取得する。GSEAでよく行われるように遺伝子セットライブラリに適用することで、発現モジュールをスクリーニングすることができる。また発現データは最低でも30サンプル程度あることがのぞましいが、発現データはGEO (http://www.ncbi.nlm.nih.gov/geo/)、TCGA(https://tcga-data.nci.nih.gov/tcga/)等の公的なデータベースから取得し、遺伝子セットを自前の実験で作成するということもできる。例えば転写因子Xを細胞株で過剰発現させ、その下流標的遺伝子の遺伝子セットを発現解析により取得、興味のあるTCGAから得た癌腫Yの発現データを解析し、その癌腫Yに転写因子Xに制御される発現モジュールが存在するか、また存在した場合、どのサンプルで活性化しているかを調べることができる。EEMのがん研究への応用例として、in vitro実験に基づいたMYC標的遺伝子セットを用いて様々ながん種におけるMYC モジュールを予測して、更にゲノムデータと統合的に解析することでMYCの新規調節因子を同定した論文を共同研究で最近発表している。

また我々は発現データからの遺伝子ネットワーク推定手法の開発にも力を入れている。遺伝子ネットワーク推定手法としては、加重相関関係に基づいたWGCNA (http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/), 相互情報量に基づいたARACNE (http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE)等、他のグループからも多数の手法が公開されているが、我々はBayesian networkに基づいた、SiGN-BN (http://sign.hgc.jp/signbn/)を公開している。Bayesian networkは遺伝子制御の因果関係の推定に有用であるが、全遺伝子に適応するとなると計算コストが高いという問題がある。そこでSiGN-BNではスーパーコンピューター上で並列計算アルゴリズムを実装することでこの問題を解決している。現在東大医科研ヒトゲノム解析センターのスーパーコンピューター及び「京」コンピュータにおいて利用可能である。さに、あらかじめ公開データから「京」コンピュータとSiGN-BNをもちいて様々な癌腫の遺伝子ネットワークを推定した結果をデータベース The Cancer Network Galaxy (TCNG; http://tcng.hgc.jp/) として公開している。ま実際にSiGN-BNで推定したネットワーク中で多くの遺伝子を制御している遺伝子（ハブ遺伝子）が重要な機能を担っていることを実験的に検証している。以上よりSiGN-BN 及びTCNGは種々のがんでのmaster regulator探索に有用であると考えられる。

**おわりに**

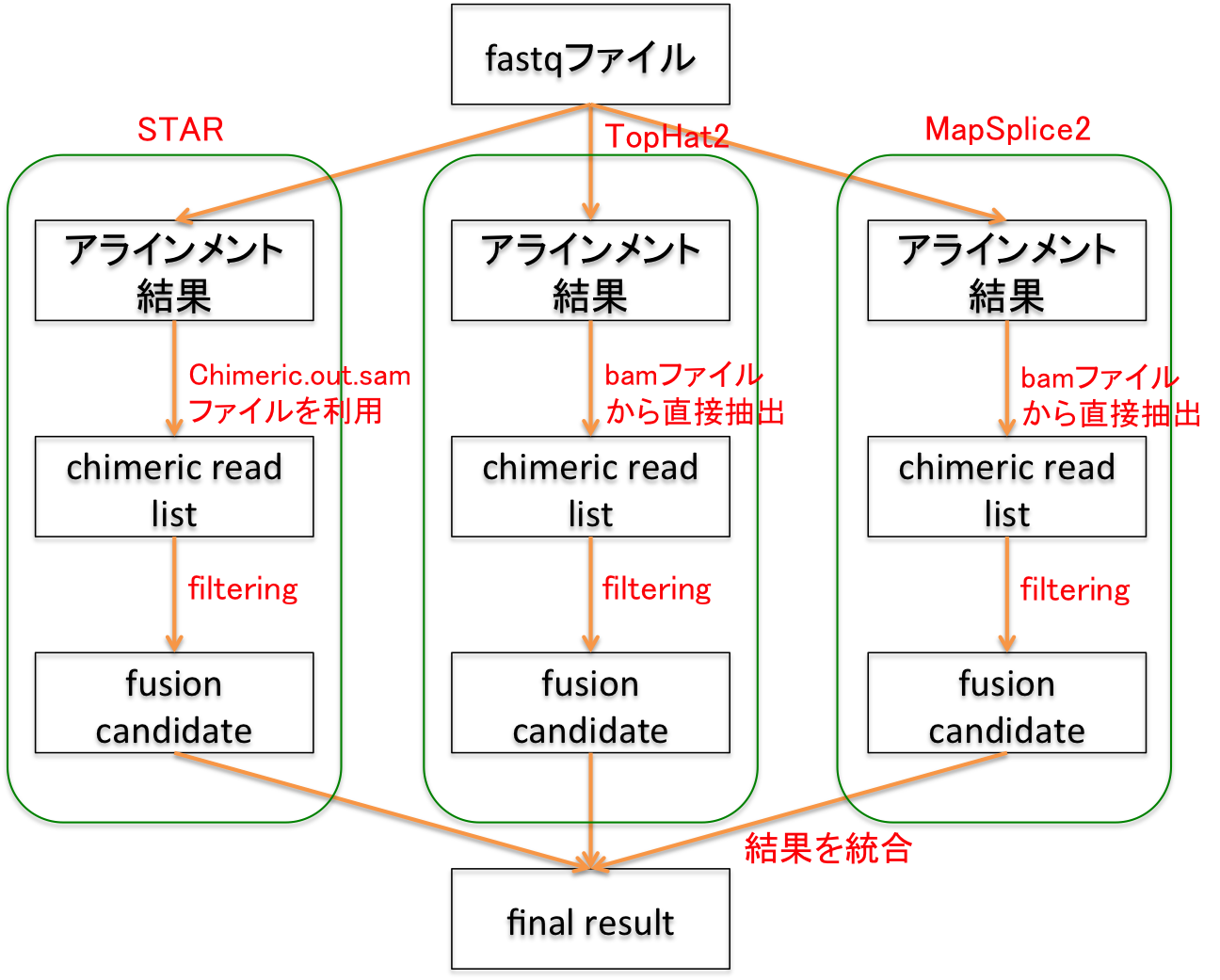
以上、がん研究におけるRNA-seqデータの解析手法を紹介してきた。近年の次世代シーケンスデータの爆発的増加によりのがん研究の中心はデータ解析になったと言っても言い過ぎではないだろう。本稿ではカバーしきれなかったが、RNA-seqデータを、ゲノム,　エピゲノムデータ、臨床情報等と組み合わせて統合的に解析することで、がんのシステム異常の更なる深部に切り込めると期待される。最後に、我々の開発した解析手法が少しでもがんの理解、そしてそれに基づいたがん医療の進歩に寄与することを願って筆をおきたい。

[参考文献]

1. Kataoka et al., Integrated molecular analysis of adult T cell leukemia/lymphoma, [Nature Genetics, **47**, 1304-1315 (2015)](http://www.ncbi.nlm.nih.gov/pubmed/26437031)
2. Niida, Atsushi, et al. "Gene set-based module discovery in the breast cancer transcriptome." BMC bioinformatics 10.1 (2009): 71.
3. Takahashi, Y., et al. "The AURKA/TPX2 axis drives colon tumorigenesis cooperatively with MYC." Annals of Oncology (2015): 26(5):935-42
4. TAMADA, YOSHINORI, et al. "SiGN: Large-scale gene network estimation environment for high performance computing." Genome Informatics 25.1 (2011): 40-52.
5. Arima, Chinatsu, et al. "Lung adenocarcinoma diversity definable by lung development-related miRNA expression profiles in association with clinicopathologic features." Carcinogenesis (2014): 35(10):2224-31
6. Affara, Muna, et al. "Vasohibin-1 is identified as a master-regulator of endothelial cell apoptosis using gene network analysis." BMC genomics 14.1 (2013): 23.

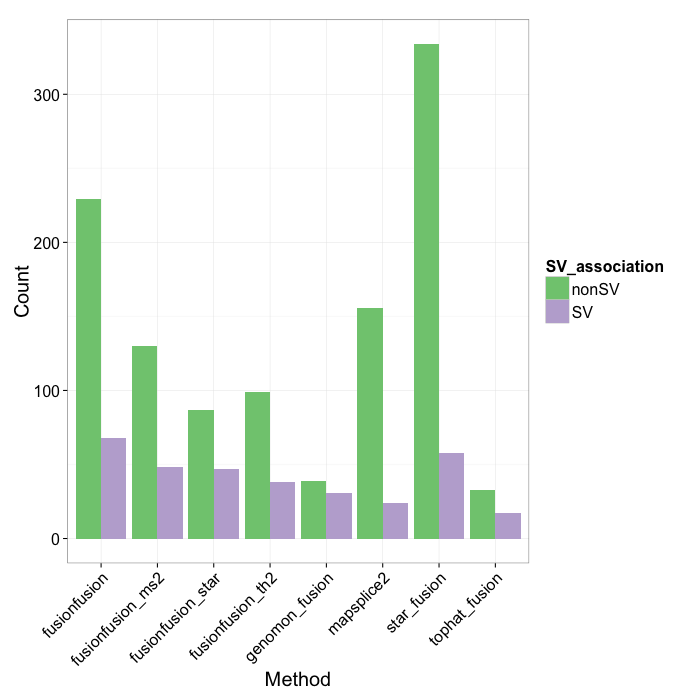
genomon-test2.png

図１

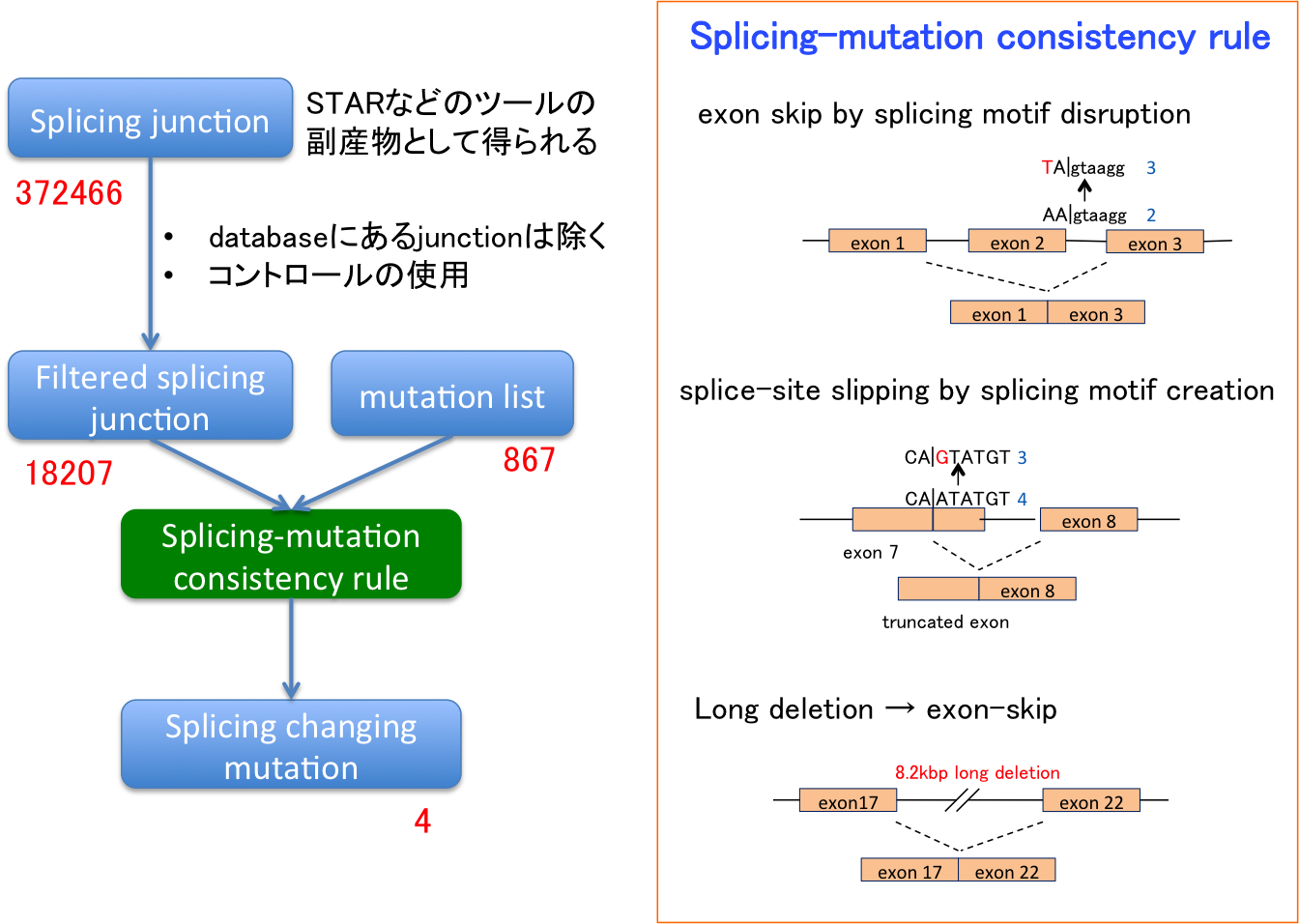


fusionsionのプログラムの概要図

図２

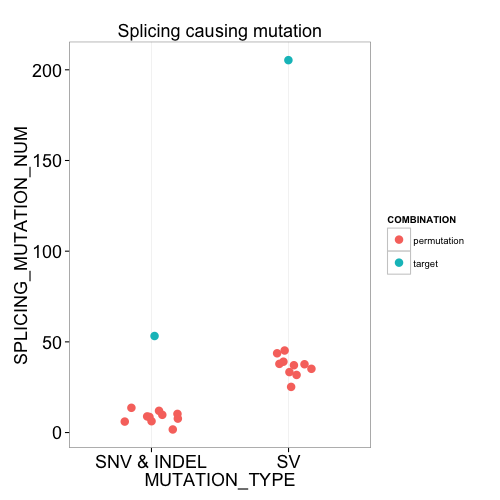


fusionfusionと他のツールの比較図．各々のツールで検出された融合遺伝子について，対応する構造異常(corresponding genomic structural variation, CGSV)が全ゲノムシークエンスのデータから検出できたかを検証．fusionfusion\_ms2, fusionfusion\_star, fusionfusion\_th2はそれぞれSTAR, MapSplice2, TopHat2による結果であり，fusionfusionはこれらの和集合である． STAR-fusion, map-splice2の独自フィルタリングの結果は，多くの候補を出すものの，その多くがnon CGSVであり，false positiveの割合が高いことが推測された．Tophat-fusionの結果は，出力される候補の個数が少なく，感度が低いことが推測される．fusionfusionの結果は，CGSVの割合も比較的高く，感度・特異度のバランスが取れている．また，複数のアラインメントツールの結果を組み合わせることにより，検出感度を上昇させることができることも示唆される．



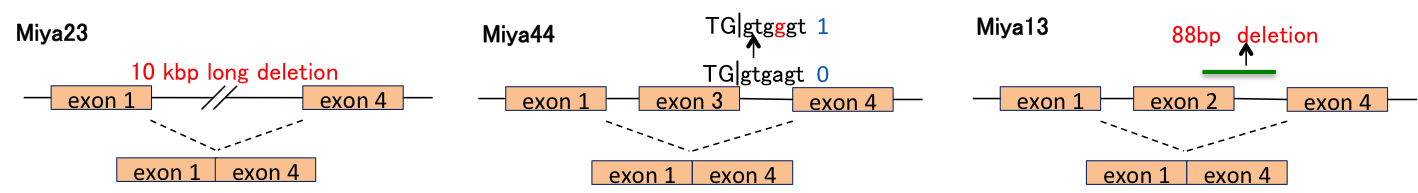
GenomonSplicingMutationの概念図

図４



ATLにおけるスプライシングを引き起こすゲノム変異（青点は正しい組み合わせによって推測された変異の個数，赤点はpermutationにより推測されたfalse positiveの個数のばらつきを表す）．

図５



TP73のexon skipを引き起こすと推測されるゲノム変異の例