

## 論文の内容の要旨

### 論文題目

# ヒト免疫細胞を用いた選択的スプライシング QTL 解析による GWAS 候補領域の解析

氏名 山口 健介

## 1. 序論

これまでゲノムワイド関連解析 (GWAS; genome wide association study)により、関節リウマチ (RA; rheumatoid arthritis) や全身性エリテマトーデス (SLE; systemic lupus erythematosus) など様々な自己免疫疾患と遺伝子多型との関連が示された。同時に、末梢血細胞などの発現解析によって、遺伝子多型が遺伝子発現量を制御する遺伝子領域 (eQTL; expression quantitative trait locus) が網羅的に同定され、自己免疫疾患の疾患感受性遺伝子領域の多くが eQTL に集積していることが明らかにされた。一方で、遺伝子多型は、選択的スプライシングにも影響を与える可能性があり、これらの遺伝子領域をスプライシング QTL (sQTL; splicing QTL) と呼ぶ。最近の報告では、これまで eQTL と評価されてきた遺伝子領域の一部は sQTL と再定義されることが明らかになっており、GWAS 候補領域において、sQTL の占める割合が大きいことが示されている。sQTL では、isoform の量的変化に加えて、アミノ酸配列の変化を介した遺伝子機能の質的変化をもたらす可能性があり、これを解析することで病態理解を深めることが可能であると期待できる。

sQTL の同定は、eQTL 解析と同様に各 isoform の発現量の比と遺伝子多型のジェノタイプとの関連解析 (sQTL 解析) により可能である。近年、次世代シーケンサー技術の発達により、RNA シーケンシング (RNA-seq) の手法を用いてスプライシングの網羅的な定量が可能となった。本論文では、免疫疾患の病態解析に有用であると考えられる日本人の健常人末梢血白血球各分画 (B 細胞、CD4 陽性 T 細胞、CD8 陽性 T 細胞、Monocyte、NK 細胞、PBL) を用いた sQTL 解析を行なった。sQTL の同定には、trQTL (transcript ratio QTL) 解析とジャンクションリードに着目した LeafCutter を併用し、より感度・特異度の高い sQTL 解析を目指した。また、isoform の ORF (open reading frame) 配列に着目し、遺伝子機能を質的に変化させる sQTL に力点を置いた。さらに、各疾患の GWAS 解析結果を統合することで、疾患感受性遺伝子領域における選択的スプライシングの役割を評価した。アジア人集団における初めての RNA-seq を用いた sQTL 解析であり、アジア人における自己免疫疾患の病態理解に寄与することが期待される。

## 2. 方法

日本人の健常人ボランティア 110 人 (女性 80%、平均 39.0 ± 10.6 歳) から末梢血を採取した。FACS (Moflo XDP) を用いて、B 細胞、CD4 陽性 T 細胞、CD8 陽性細胞、Monocyte、NK 細胞

を分画し、この 5 細胞種と PBL を解析した。ジェノタイピングは Infinium OmniExpressExome BeadChip (Illumina)を用いた。imputation を行ない、常染色体上の  $MAF \geq 0.05$  である計 5,600,101 の SNP と InDel を解析に用いた。ライブラリ調整は TruSeq Stranded mRNA Library Prep Kits (Illumina)を用いて、Hiseq 2500 (Illumina)でシーケンシング (125 bp paired-end)を行なった。マッピングは多型情報を考慮に入れることが可能な HISAT2 を用いた。ゲノム配列は GRCh38、トランスクリプトのレファレンスは Gencode v25 annotation、遺伝子多型情報は UCSC common SNPs 147 ( $MAF \geq 0.01$ )を用いた。発現量定量は Cufflinks を用いた。

eQTL 解析は Matrix eQTL を用いて細胞種ごと独立に行なった。遺伝子から 1MB 内の多型を解析対象とした。遺伝子を対象とした gene eQTL 解析、各 isoform を対象とした isoform eQTL 解析を行なった。

sQTL 解析は、各 isoform 発現量を定量した上で isoform 間の発現量の比を解析する trQTL 解析、およびエクソン間を跨ぐジャンクションリードのみに着目しその比を検討する LeafCutter を用いた解析の 2 種を比較・統合した。trQTL 解析では、同一遺伝子でかつ同一の ORF 配列を持つ isoform を集約し、同遺伝子全体の発現量で除した値を transcript ratio とした。LeafCutter を用いた解析では、ジャンクションリードを用いてクラスタリングを行ない、あるジャンクションのリード数をクラスター内のリード総数で除した値を junction read ratio とした。これらの ratio を用いて、eQTL 解析と同様に相関解析を行なった。最終的に trQTL 解析において FDR (false discovery rate)  $\leq 0.05$  かつ LeafCutter で FDR  $\leq 0.1$  であり、さらに beta の正負が両解析で一致したものを有意な sQTL とした。

sQTL 解析と GWAS の統合として、免疫疾患の GWAS データベースである ImmunoBase の 20 疾患の GWAS SNP から連鎖不平衡にある ( $r^2 \geq 0.8$ )の多型について、sQTL effect を評価した。

### 3. 結果

gene eQTL 解析にて、B 細胞 2,570; CD4 陽性 T 細胞 2,796; CD8 陽性 T 細胞 2,280; Monocyte 3,634; NK 細胞 2,573; PBL 3,196 遺伝子が有意な gene eQTL と判定され、isoform eQTL 解析にて、B 細胞 2,574; CD4 陽性 T 細胞 2,776; CD8 陽性 T 細胞 2,227; Monocyte 3,423; NK 細胞 2,526; PBL 3,247 遺伝子が有意な isoform eQTL と判定された (FDR  $\leq 0.05$ )。

trQTL 解析と LeafCutter を用いた解析の両方で有意な sQTL と判定された isoform-遺伝子多型ペア (FDR  $\leq 0.05$ )の sQTL effect (beta)は 93.68 %で正負が一致し、同一の sQTL effect を検出していると示唆された。trQTL 解析と LeafCutter による sQTL 解析を統合した結果、B 細胞 307; CD4 陽性 T 細胞 343; CD8 陽性 T 細胞 268; Monocyte 475; NK 細胞 325; PBL 465 遺伝子が有意な sQTL と判定された。

さらに、細胞種特異的なスプライシング、sQTL effect を評価するため、遺伝子発現量、isoform

発現量、isoform ratio のクラスタリングを行なった。遺伝子発現量や isoform 発現量と比較し、isoform ratio は、細胞種ごとのクラスタ化はされにくい、Monocyte では一部クラスタ化され、細胞種特異的なスプライシングの存在が示唆された。実際に、細胞種特異的な sQTL effect が *CD300A* 遺伝子など一部の遺伝子で確認された。eQTL/sQTL の機序を明らかにするため、これらの effect を示す SNP を解析したところ、いずれも TSS (transcription start site) を中心に分布するが、gene body との位置関係では異なる分布を示した。

sQTL 解析と GWAS 結果の統合として、ImmunoBase に登録された GWAS SNP と強い連鎖不平衡 ( $r^2 > 0.8$ ) にある多型の sQTL effect を評価したところ、B 細胞 30; CD4 陽性 T 細胞 31; CD8 陽性 T 細胞 17; Monocyte 31; NK 細胞 26; PBL 39 遺伝子で有意な sQTL effect がみられた (FDR  $\leq 0.05$ )。個別の領域として、*SNRPC* 遺伝子では、リスクアレルにより機能的な *SNRPC* isoform の発現量が増加していた。*IL6R* 遺伝子では、リスクアレルにより可溶性 *IL6R* isoform の発現量が減少していた。*HLA-DQB1* 遺伝子では、リスクアレルにより可溶性 *HLA-DQB1* isoform の発現量が減少していた。

#### 4. 考察

本研究は、RNA-seq を用いた末梢血免疫細胞の sQTL 解析により、GWAS で同定された免疫疾患感受性遺伝子多型 (GWAS SNP) による疾患発症メカニズムの解明を目的とした。細胞表面マーカーによりソートされた細胞分画を用いることで、GWAS SNP がどの細胞種において重要な役割を果たしているかを評価可能と考えられる。これまで細胞種特異的な sQTL の報告は少ないが、本解析では細胞種特異的な eQTL より少ないながら細胞種特異的な sQTL の存在を確認した。この eQTL と sQTL の細胞特異性の違いについて両者の分布から考察した。eQTL は TSS 付近に大きなピークと gene body の 3' 側に小さなピークの 2 峰性に分布した。これは eQTL がプロモーター領域、エンハンサー領域、3'-UTR 領域に集積するという知見と一致する。また、エンハンサーの活性化と調整には細胞種特異性があることから、細胞種特異的な eQTL が生じていると考えられる。一方の sQTL は TSS 付近に小さなピークと gene body の 3' 側に大きなピークの 2 峰性に分布した。後者は、splicing site などに存在する多型によると考えられる。TSS のピークは、eQTL effect を介して生じた sQTL と考えられる。すなわち、遺伝子多型による転写活性の変化が発現量とスプライシングの両方に影響を与える可能性がある。

本研究では、GWAS 候補領域における sQTL の寄与についても評価した。まず、SLE の疾患感受性遺伝子領域に存在する *SNRPC* 遺伝子はスプライソソームの構成因子である U1-snRNP をコードする。U1-snRNP に対する自己抗体 (抗 U1-RNP 抗体) が SLE や混合性結合組織病で出現し、U1-U6 snRNP に共通に存在する抗原に対する自己抗体 (抗 sm 抗体) は SLE の特異抗体とされる。本解析において、GWAS SNP (rs2764208) のリスクアレルは、機能的な *SNRPC* isoform の発現量を増やすため、この機能亢進が病態に関与しているものと考えられた。

次に RA および強直性脊椎炎の感受性遺伝子 *IL6R* は、膜型 *IL6R* と可溶性 *IL6R* をコードするが、可溶性 *IL6R* の生成機序として、膜型 *IL6R* のタンパクレベルでの Shedding と、選択的ス

プライシングの2つが想定されてきた。GWAS SNP (rs2228145)はアミノ酸置換を伴ない、エクソン9上の proteolytic cleavage site に存在するため、切断効率に変化を与えることで、可溶性 IL6R を介して疾患に関わるとされてきた。本解析において rs2228145 は sQTL effect を持ち、リスクアレルにより可溶性 IL6R の mRNA 発現量が減少することが示された。遺伝子多型による可溶性 IL6R 発現への影響として、これまで一般的に想定されていたタンパクレベルでの制御とは異なり、スプライシングを介した mRNA レベルでの制御が確認された。

最後に *HLA-DQB1* 遺伝子は、HLA-DQ 抗原の  $\beta$  鎖をコードし、抗原提示を行なうことで免疫において中心的な役割を果たす。通常型の isoform に加えて、cytoplasmic tail (エクソン5)が追加された isoform、および transmembrane domain (エクソン6)を欠く可溶性 isoform が存在する。splice site に存在する rs28688207 のアレルにより、エクソン5の有無が決定することが知られていたが、興味深いことに、エクソン5を持つ isoform とそれ以外の isoform とは排他的な発現様式を示すことが本解析で明らかになった。可溶性の MHC class II 分子は膜型 MHC 分子と競合し、CD4 陽性 T 細胞に対して抑制的に働くとの報告もあるため、*HLA-DQB1* における sQTL effect がタンパク機能に総体を与える影響としては、通常型および cytoplasmic tail をもつ HLA-DQ  $\beta$  鎖の機能とともに、可溶性 HLA-DQ  $\beta$  を介した機能変化を考慮する必要がある。

これまでに見てきたように、自己免疫疾患の GWAS SNP の一部は、sQTL effect を介して病態に寄与していることが明らかになった。本研究では同じ ORF をもつ isoform を集約することによりタンパク機能に質的影響を与える sQTL を効率的に同定することを試みたが、これらのタンパクの機能解析などを行なうことにより、自己免疫疾患の本質に迫ることができると期待される。