博士論文

論文題目　ヒト脳神経組織における体細胞変異の探索

氏名　西岡 将基

目次

要旨

統合失調症などの精神疾患において様々な関連候補遺伝子が報告されてきたが、病態を十分には説明できていない。脳奇形を伴う脳神経疾患を中心に、生殖系列ゲノムの変異に加え、体細胞変異も病態の一部を担っている可能性が提起されている。本研究では、全ゲノム・全エクソンシークエンスデータからの体細胞変異検出系を確立し、健常者脳神経組織において実際に脳部位・細胞種に特徴的な体細胞変異を同定した。また、精神疾患について不一致な一卵性双生児ペアの血液試料で、罹患者に特徴的な体細胞変異を同定した。結果に基づき、体細胞変異検出の技術的・生物学的な側面を議論し、今後の精神疾患研究への意味合いを考察した。

1. 序文

統合失調症は、幻聴・被害妄想・被注察感をはじめとする陽性症状、無為・自閉・感情鈍麻といった陰性症状、ワーキングメモリの低下などの認知機能障害を特徴とした精神疾患であり、思春期前後で発症することが多く、人口の 1%程度が罹患していると報告されている。本人自体の苦痛に加え、社会生活が障害されることなどからの社会的損失も大きく、世界的に解決すべき大きな課題である[1]。

統合失調症は、一卵性双生児と二卵性双生児の診断一致率の比較から高い遺伝率が指摘されており[2]、遺伝学的研究が歴史的に数多くなされてきた。遺伝要因の影響が強く罹患者の多い疾患については、相関解析や連鎖不平衡を利用し原因遺伝子が特定できるであろうという想定のもと、一塩基多型（single nucleotide polymorphism, 以下 SNP）の頻度情報を利用したゲノムワイド関連解析（genome-wide association study, 以下 GWAS）が行われており、統合失調症と有意に関連する SNP が多数報告されている[3]。また、数 Mb にわたる遺伝子コピー数の変化（copy number variation, 以下 CNV）に注目した関連研究[4-6]や、一部の家系で患者のみが持つ稀な遺伝子変異に注目した研究も多数報告されている[7,8]。これらの研究により、オッズ比が高い変異が報告されている。

近年、罹患者と非罹患両親の 3 名（孤発例のトリオ）のゲノムを比較することで、配偶子形成段階で生じた新規変異（新生突然変異）を検出するトリオ解析が盛んに行われている。トリオ解析にて有力な統合失調症関連変異が同定されており[9-12]、シナ

プス関連遺伝子の変異が多く検出されるなど、統合失調症の病態解明に重要な知見を提供している。新生突然変異として検出された変異の中には、血液細胞でモザイクに存在する変異も含まれており、このような変異は、罹患者の発生過程において生じた体細胞変異と考えられ[13]、疾患発症を説明する因子となる可能性が考えられる。特に脳神経系に特徴的に存在する体細胞変異は、生殖系列ゲノムの変異を補完する形で、精神疾患をはじめとした精神・神経機能の異常に関連する可能性が考えられる。また、同一の生殖系列ゲノムを持つ一卵性双生児も、体細胞変異のパターンは異なると予想され、診断不一致の背景に体細胞変異が存在する可能性を考えることができる。

体細胞変異の脳神経組織内の局在によって、障害を受ける脳部位・脳機能が異なり、表現型が多様となることが予想される。これまでの遺伝学的研究により、症候学的に統合失調症と分類される集団は遺伝的多様性を大きく含んだ集団であると考えられているが[14]、体細胞変異の脳神経組織内局在が表現型の多様性に寄与している可能性も想定できる。以上の経緯から、本研究は、統合失調症をはじめとする精神疾患の発症を説明する機序、疾患内の多様性を説明する機序の候補として、体細胞変異に注目した。

近年のゲノム解析技術の進歩に伴い、ヒト脳神経組織における体細胞変異の網羅的解析が報告されている。単一神経細胞ゲノム解析により、一神経細胞あたり約 1500 個の一塩基変異（single nucleotide variant, 以下 SNV）が存在すると推定されている[15]。また、脳組織に特徴的な CNV も、単一細胞単位の大規模なゲノム解析で報告されて

いる[16,17]。ヒトを含むほ乳類のゲノムには、自身の転写産物が逆転写されることでゲノム内に DNA 配列が複製・新規挿入される因子（転移因子）があり、中でもレトロトランスポゾンと呼ばれる転移因子は転移頻度が高く、生殖系列ゲノムにおける変異の主たる要因の一つとみなされている[18]。近年、神経前駆細胞においてレトロトランスポゾンが転移活性を保持していることが見出され、脳神経細胞におけるレトロトランスポゾンの新規挿入（レトロトランスポジション）が検出されている[19-21]。レトロトランスポジションはごく一部の脳神経細胞で見出される体細胞変異であり、単一脳神経系細胞を用いた複数の研究においても新規挿入の検出が報告されている[22-24]。

疾患との関連に関しては、一側性巨脳症や皮質形成異常を伴うてんかんにおいて、特定の脳部位に特徴的な体細胞変異が検出されており、前者については *AKT3* (v-akt murine thymoma viral oncogene homolog 3)関連遺伝子群の SNV、後者については *MTOR* (mechanistic target of rapamycin)遺伝子の体細胞 SNV と疾患との関連が示唆されている[25-27]。Lim らは、皮質形成異常及びてんかんを持つ患者の罹患脳部位においてアレル割合 1.3~12.6%の *MTOR* 遺伝子体細胞変異を検出し、モデル動物を用いててんかんとの因果関係を示した[27]。Jamuar らは、二重皮質症候群や脳回肥厚症を伴う患者由来の血液試料を用いて、標的遺伝子群の SNV 解析を行い、*DCX* (doublecortin)遺伝子などで、アレル割合 5~35%の SNV を見出している[28]。Rett 症候群において、血液細胞から *MECP2* (methyl-CpG binding protein 2)遺伝子の体細胞変異を検出したという報告もある[29,30]。これらの研究では脳神経組織を直接対象とはしていないものの、既に因果

関係が強く示唆されている遺伝子群の体細胞変異が検出されており、体細胞変異と疾患との有力な関連を示唆している。また、レトロトランスポゾン LINE1 (Long interspersed nuclear element 1)については、患者死後脳でのコピー数増大が、血管拡張性失調症、レット症候群、統合失調症で報告されている[31-33]。特に Bundo らは、統合失調症患者死後脳試料のほか、精神疾患動物モデルや患者から樹立した iPS 細胞から誘導した神経細胞においてもコピー数の増大を認めており、脳神経系における体細胞レトロトランスポジションが患者群で特徴的に認められることを示している[33]。

明らかな解剖学的特徴を伴わない統合失調症において体細胞変異が病因・病態に関与している可能性が示唆されたことで、体細胞変異が、精神疾患の病因・病態を説明する機序の一つとして想定できると考え、脳神経組織及び一卵性双生児血液試料における体細胞変異の解析を行った。精神疾患発症メカニズム解明への道筋をつけることを目的とし、体細胞変異として報告のあるレトロトランスポゾン LINE1 に加え、SNV の解析を行った。現在のゲノム解析技術では、挿入・欠失（insertion/deletion, 以下 INDEL）や構造変異はアラインメントが難しく、SNV に比べ相対的に解析困難であり、低割合に存在する体細胞変異の解析として SNV が適当であると考えたためである。

本研究では、まず、1) 健常者由来脳組織や肝臓組織での全ゲノムシークエンシング解析データを用い、既存ソフトウェアにて脳神経組織の体細胞 SNV を高感度に解析する方法を確立した。体細胞変異検出のためのソフトウェアは既に複数開発されており、体細胞変異が頻発するガン研究分野で使用されているものの、脳神経系試料や

精神疾患患者試料において効果的な検出法は確立されていないため、脳神経試料解析に適切な手法の開発を行った。確立した解析手法を基に、脳組織組織の生理的な体細胞 SNV（脳・脳部位に特徴的な体細胞 SNV を含む）を同定し、体細胞変異の生物学的な特徴を解析した。

次に、2) 統合失調症圏の疾患について不一致な計 5 組の一卵性双生児由来血液試料の全ゲノムまたは全エクソームシークエンシング解析データを用い、体細胞 SNV の探索を行った。統合失調症は、一卵性双生児での診断一致率が約 50%であり、これまで不一致の部分は環境要因によるものと考えられてきた。しかし、双生児間で体細胞変異の状態が異なると予想され、体細胞変異が診断不一致の背景にある可能性が考えられる。健常者同士の 1 組であるが、全ゲノムシークエンスデータから一卵性双生児における体細胞変異を検出したという報告もある [34]。この報告ではサンガー法のクロマトグラム波形から体細胞変異の存在を確認しているが、本研究ではより定量的な手段を用い、複数の手法により体細胞変異の存在を確認した。本研究の結果、妄想性障害罹患者に特徴的な体細胞 SNV を血液試料から検出した。

また、3) レトロトランスポゾン LINE1 の転移部位同定法の確立と健常者死後脳を用いた解析を行った。Rett 症候群や統合失調症において LINE1 コピー数増加との関連が指摘されているが [29,30,33]、LINE1 挿入位置の決定が課題となっており、疾患試料における LINE1 挿入位置を同定するための手法の確立を行った。LINE1 配列の中でも、唯一自律的な転移活性を持つとされている L1Hs 配列のゲノム上の転移部位解析法

（L1Hs-seq）が先行研究で報告されている[35]。L1Hs-seq は、L1Hs に特異的なプライマーを使用して L1Hs の 3'末端と下流のゲノム配列を含む領域を増幅し、網羅的に L1Hs 配列挿入部位を決定する方法である。L1Hs-seq 原法は、低割合に存在する新規挿入検出について評価されておらず、新規挿入した位置の正確な同定も難しい。疾患試料への適用を目標に、一塩基レベルの新規挿入位置同定と低割合に存在する新規挿入検出が行えるよう L1Hs-seq に対する改良を行い、健常者死後脳での挿入部位探索を行った。

以上の研究課題にて、脳神経試料から体細胞 SNV を検出する手法及び、LINE1 挿入位置決定法の開発を行った。開発した手法を用い、脳神経組織における体細胞 SNV、一卵性双生児における体細胞 SNV を同定・解析した結果を報告する。結果に基づき、本研究で用いた手法の妥当性及び限界点、脳神経組織・一卵性双生児における体細胞 SNV の生物学的意味について議論する。最後に、体細胞変異による精神疾患発症の説明モデルを提示し、今後の精神疾患研究に対する示唆を考察する。

2. 方法

本研究で行った実験と解析の方法を記す。データ解析にあたっては、RHEL 6.2 (Linux kernel 2.6.32)、または CentOS 6.5 (Linux kernel 2.6.32)の OS 上でソフトウェアを作動させた。前者は Xeon X5675 3.06 GHz/ 12 MB, 6.40 GT/s 6-core CPU と 96 GB のメモリを搭載した計算機上で、後者は Xeon E5-2690 2.90 GHz/ 20 MB, 8.0 GT/s 8-core CPU と 264 GB のメモリを搭載した計算機上で稼働した。統計解析とグラフ作成は R-3.2.0（https://www.r-project.org/）にて行った。

本研究の計画・実験・解析は、東京大学医学部、理化学研究所脳科学総合研究センター、東北大学医学部、札幌医科大学医学部において倫理委員会より承認を受けて実行されており、試料提供者または家族（死後の場合）に対して十分な説明を経て試料提供及び研究における使用の同意を得た。東京大学医学部における倫理承認審査番号は G0639-(32)である。共同研究者が行った旨の記載のない実験・解析は全て筆者自身が行った。

2.1 高深度全ゲノムシークエンスによる脳特徴的な体細胞一塩基変異の探索

死後脳組織・肝臓組織由来の高深度全ゲノムシークエンス（whole genome sequence, 以下 WGS）データを、同一個体間で互いに比較することで、各組織に特徴的な体細胞 SNV 候補の探索を行った。体細胞 SNV 候補に対し、独立して超高深度ターゲットアンプリコンシークエンスを行い、変異の割合を定量的に確認した。

2.1.1 試料の準備

68 歳男性より提供された死後脳（前頭葉、サンプル ID：AL30_cortex）と死後肝臓（AL30_liver）、78 歳男性より提供された死後脳（前頭葉）から後述の手順によって神経細胞核・非神経細胞核を分離した試料と死後肝臓（それぞれ Y8763_NeuN+、Y8763_NeuN-、Y8763_liver）、84 歳男性から提供された死後脳皮質（前頭葉、S6_cortex）と死後小脳（S6_cerebellum）を試料として用いた。試料と試料提供者のデータについては、表 1 に記載した。68 歳男性の提供者のみコーカシアンであり、他の試料提供者は全て日本人である。

78 歳男性より提供された死後脳前頭葉皮質から以下の方法によって、神経細胞核と非神経細胞核の分離を行った。STKM buffer (50 mM Tris-HCl pH 7.4, 25 mM KCl, 5 mM $MgCl_2$, 250 mM sucrose)内で組織片のホモジナイズを行ったのち、パーコール密度勾配遠心法を用いて、粗細胞核画分を生化学的に精製した。2% bovine serum albumin (BSA) で 4 度 2 時間ブロッキングを行った後、Alexa Fluor 488 でラベルされた抗 NeuN 抗体 (Millipore, #MAB377X)で一晩静かに震盪させて染色した。BD FACS Aria II (BD Biosciences) セルソーターを用いて、抗 NeuN 抗体が結合した核（NeuN+核）と結合しない核（NeuN-核）のソートを行った。核画分からのゲノム DNA は、標準的なフェノール・クロロホルム法にて抽出した。小脳、前頭葉、肝臓からのゲノム DNA 抽出も、標準的なフェノール・クロロホルム法にて行った。以上の実験は共同研究者で

ある文東美紀が行った。

2.1.2 シークエンス方法

サンプル AL30_cortex, AL30_liver, Y8763_NeuN+, Y8763_NeuN-, Y8763_liver に対しては、抽出した全ゲノム DNA に対して TruSeq DNA PCR-Free Sample Prep Kit (Illumina)を用い、製造者のプロトコルに従ってライブラリ調整を行った。PCR エラーの可能性を除外するべく、PCR を含まない方法でシークエンスライブラリを調整した。調整したライブラリは、HiSeq 2500 (Illumina)を用い、Rapid Run モード・ペアエンド法にて、理論的深度約 100、各ペア 162bp で配列決定を行った。各サンプルにインデックス配列を用意することで、サンプル間のクロスコンタミネーションや機械に残留した DNA によるコンタミネーションを予防した。HiSeq 2500 の稼働・データ取得は、共同研究者である安田純、長﨑正朗、勝岡史城、佐藤行人、黒木陽子が行った。

サンプル S6_cortex, S6_cerebellum に対しては、抽出した全ゲノム DNA に対して TruSeq Nano DNA sample Prep kit (Illumina)を用い、製造者のプロトコルに従ってライブラリ調整を行った。調整したライブラリは、HiSeq X (Illumina)を用い、ペアエンド法にて、理論的深度約 120、各ペア 150bp で配列決定を行った。構造上、HiSeq X は機械に残留した DNA によるコンタミネーションがないため、インデックスは使用していない。サンプル調整、HiSeq X の稼働は、理研ジェネシス社の受託業務として行った。

2.1.3 シークエンスデータのアラインメントとクオリティコントロール

得られた WGS データに対して、FastQC-0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Trimmomatic-0.32 にてシークエンスデータのクオリティコントロール（以下 QC）を行い、BWA-0.7.12 によりリファレンスゲノム build37 + decoy 配列に対するアラインメントを行った [36,37]。アラインメントデータの QC は、まずレーンごとに、PCR や光学系による重複リードを除去するための Picard Deduplication (Picard-1.102)、INDEL 領域のアラインメントを正確に行うための GATK Indel Realignment (GATK-3.2-2)、ベースコールのベースクオリティを補正する GATK Base Recalibration (GATK-3.2-2)を、記載の順で行った [38,39]。レーンごとのアラインメントデータをマージした後、シークエンスライブラリ調整に PCR を使用した試料に関しては Picard Deduplication を再度行い、全ての試料について GATK Indel Realignment を再度行い、Samtools-0.1.19 にてマッピングクオリティ（mapQ）による選択を行った [39]。各ソフトウェアのパラメーターは、セクション 3.1 に記載したように、得られたシークエンスデータのクオリティに応じて設定し、最終的なアラインメントが体細胞 SNV 探索に最適となるようパラメーターの調整を行った。各サンプルのパラメーター設定の詳細は図 1 に記載した。

2.1.4 体細胞一塩基変異候補の検出

得られたアラインメントデータ（BAM ファイル）に対して、MuTect-1.1.5[40], Strelka-1.0.14[41] を用い、同一個体由来の各組織間で比較を行い、体細胞一塩基変異（SNV）の候補を検出した。ガン組織を対象とした体細胞解析において、これらのソフトウェアが感度・特異度に最も優れていると報告されている [40,42]。比較によって、一方の組織で数%の割合で認めるが、対照組織では存在を認めない（あるいは割合が小さい）といったベースコール（体細胞 SNV 候補）を検出することを想定し、CL_WGS_set (AL30_cortex, AL30_liver), NeuN_WGS_set (Y8763_NeuN+, Y8763_NeuN-, Y8763_liver), CC_WGS_set (S6_cortex, S6_cerebellum)の 3 セット内で比較を行った。

MuTect は、解析対象サイトにおけるリファレンス上の塩基（以下、Ref）と異なるベースコール（以下、Alt）に対し、1) Alt が全てシークエンスエラーである確率及び、2）Alt が全て生殖系列ゲノムにおける SNV である確率を、対象サイトにアラインメントされた全リードの各ベースクオリティから計算し、それら体細胞変異であることの帰無仮説を棄却することで、体細胞変異候補を検出する。いずれの確率計算も全ゲノムの塩基数（約 30 億塩基）を考慮した設定となっており、生殖系列ゲノム SNV による偽陽性がかなりの頻度で予想されるため、帰無仮説 2 の棄却閾値は更に厳しく設定されている。体細胞変異候補は、プラス鎖・マイナス鎖に偏りのないこと、リード末端などの特定の位置に集積していないこと、Alt としてコールされた塩基が 2 種類以上ないこと、などの更なるフィルタリングを経て、最終的な体細胞変異候補がアウトプットされる。実際の MuTect による解析は、デフォルトモード（デフォル

トのパラメーター）と高感度検出モード（パラメーター：--minimum_normal_allele_fraction 0.15 --max_alt_alleles_in_normal_count 20 --max_alt_alleles_in_normal_qscore_sum 600）の 2 パターンの解析を行った。デフォルトモードでは、対照組織に体細胞変異が原則ない状態を仮定して解析を行うモードであり、発生段階がある程度進んでから生じた体細胞変異候補を検出する。高感度検出モードでは、対照組織に一定以上（本解析では Alt 割合 15%までであり、常染色体であれば全細胞の 30%までの変異）の同じ体細胞変異がある状態を仮定して解析を行うモードであり、発生段階初期に生じた体細胞変異候補も含めて検出することができる。デフォルトモードでの体細胞変異候補は、全て高感度検出モードに含まれるため、以降このセクションでの MuTect 解析結果という記載は高感度検出モードの結果である。許容 Alt 割合は値が大きすぎると、生殖系列ゲノムの変異が偽陽性として混入しやすくなり、値が小さすぎると、体細胞変異の見落とし（偽陰性）が増えるため、上記の値に設定した。

一方 Strelka では、対照試料での Alt 割合から予想される確率分布から、対象試料での割合がどの程度外れているかという計算により体細胞変異候補を検出する。Strelka での解析はデフォルトのパラメーターを用いた。

2.1.5 体細胞一塩基変異候補のフィルタリング

ヒトゲノム上には、トランスポゾンやマイクロサテライトなど、ゲノム上に多数の

相同配列が存在する領域があるが、現在のショートリードシークエンス技術では正確なアラインメントが難しく解析が困難である[43]。そのような領域上の体細胞 SNV 候補は、偽陽性が頻出すると考え、以降の解析からは除外した。具体的には、UCSC ゲノムブラウザ[44]上で RepeatMasker, Interrupted repeats, Segmental Duplication, Microsatellites, Self chain, Simple Repeats とアノテーションされている領域（総計 1.63 Gbp、以下多コピー領域）で検出された体細胞 SNV 候補を以降の解析から除外した。数十 bp 以下の INDEL の前後も正確なアラインメントが難しく SNV の解析が困難であるため、GATK-3.2-2 UnifiedGenotyper でコールされた INDEL の前後 10bp 上の体細胞 SNV 候補は除外した。

得られた候補リストの中には、同一個体由来のデータを合わせると、数百 Kb 内に MuTect で検出された候補が密集している領域があり、平均より高深度な領域もあった。それらは多コピー疑い領域として以降の解析から除外し、残った候補から対象試料での Alt 割合が、解析コントロール試料での Alt 割合より大きいものを選択した。

フィルタリング基準を数値的に決定するため、まず CL_WGS_set に関して上記の候補から IGV-2.3.40[45] による BAM ファイル可視化にて、ソフトクリップ（一部のみがアラインメントされたリード中、対象ゲノム領域にアラインメントされなかった残りの部分）近傍に集積した候補や近くにミスマッチが頻出した候補を、配列決定精度・アラインメント精度が低い領域上の候補として除外し、残りを最終的な体細胞 SNV 候補とした（マニュアル法)。この際、多コピー疑いとした領域からもランダムに 1

カ所ずつ選び、最終候補と合わせ、セクション 2.1.7 のバリデーション対象とした。マニュアル法のバリデーション結果を基に、ベースクオリティや深度など複数のフィルターの閾値を数値的に設定し、客観的なフィルタリング方法を設定した（操作的方法)。操作的方法のフィルタリング閾値の詳細は、セクション 3.1.3 に記載した。

2.1.6 超高深度ターゲットアンプリコンシークエンスによる体細胞一塩基変異検出の検討

脳神経組織のおける体細胞 SNV を検出するための系の開発を行った。2 名の日本人男性血液細胞由来ゲノム DNA を用意し、ジェノタイプが異なる SNP 位置を同定した上で、2 名のゲノム DNA を一定の割合で混合することにより、1~10%の体細胞 SNV をシミュレートした試料を作成した。作成した混合試料に対して、深度 20 万超のターゲットアンプリコンシークエンス（target amplicon sequence, 以下 TAS）を行い、理論的な体細胞 SNV 割合と実験的に検出された体細胞 SNV 割合を比較し、系の妥当性を確認した。

2.1.6.1 体細胞一塩基変異シミュレーション用試料のジェノタイピング

2 名の日本人男性血液細胞由来ゲノム DNA（JM1, JM2）に対して SNP 解析を行い、dbSNP[46] に登録されている SNP であり、かつ 2 名で異なるジェノタイプである SNP サイトを選択した。ゲノム DNA は、血液試料から三菱化学メディエンスの受託業務

として抽出し、対象とする dbSNP を含むよう PCR プライマーを Primer-3plus[47] で設計した。プライマーは、アダプター配列含めた最終フラグメントが、MiSeq (Illumina) によるシークエンスに最適な 345bp 程度になるよう設計した。使用するプライマー配列選択のため、まず試験的に設計したプライマー候補にて、NEBNext HighFidelity 2X PCR MasterMix (NEB)による試験的な PCR を行った。ゲノム DNA 5 ng, Q5 High-Fidelity DNA Polymerase, dNTP 200 uM/each, Mg++Cl 2 mM, 各プライマー 0.5 uM, 全体容量 15 ul となるよう PCR サンプルを調整した。サーマルサイクラーは「98ºC にて 30 秒」、「98ºC を 10 秒、62ºC を 10 秒、72ºC を 10 秒」を 30 回繰り返し、「72ºC を 5 分」の設定にて稼働した。複数のプライマー候補から、シングルバンドが得られたものを選択した。

選択した 12 種類のプライマーセットの配列（表 2）を用い、図 2 のように 2 回の PCR にてシークエンス用のフラグメントを作成し、理論的深度 20 万前後で MiSeq にてシークエンスを行った。2 回の PCR ともに NEBNext HighFidelity 2X PCR MasterMix を使用した。1 回目の PCR は、対象試料のゲノム DNA 10ng, Q5 High-Fidelity DNA Polymerase, dNTP 200 uM/each, Mg++Cl 2 mM, 各プライマー 0.5 uM, 全体容量 15 ul となるよう調整し、サーマルサイクラーを「98ºC にて 30 秒」、「98ºC を 10 秒、62ºC を 10 秒、72ºC を 10 秒」を 20 回繰り返し、「72ºC を 5 分」の設定にて稼働した。SPRIselect (Beckman Coulter)で PCR プロダクトの精製を行い、10 ul に濃縮した。2 回目の PCR は、濃縮した PCR プロダクト 3 ul を用い、Q5 High-Fidelity DNA Polymerase, dNTP 200

uM/each, Mg++Cl 2 mM, 各プライマー 0.5 uM, 全体容量 15 ul となるよう調整し、サーマルサイクラーを「98ºC にて 30 秒」、「98ºC を 10 秒、62ºC を 10 秒、72ºC を 10 秒」を 11 回繰り返し、「72ºC を 5 分」の設定にて稼働した。2 回目の PCR でサンプル固有のインデックスのついたアダプター配列を付加し、SPRIselect で PCR プロダクトの精製を行った。表 3 に、1 回目のプライマーに付加した配列及び、2 回目の PCR に使用したプライマーセットとインデックスを記載した。得られたプロダクトは、Qubit 2.0 Fluorometer, Qubit dsDNA HS Assay Kit (Life Technologies)で濃度測定し、フラグメント長を考慮して等モル濃度になるようアンプリコンを混合したのち、SPRIselect にて 250-700bp のプロダクトを中心に精製した。Bioanalyzer (Agilent)で分布を確認し、Qubit 2.0 Fluorometer, Qubit dsDNA HS Assay Kit で正確に濃度測定した上で、アンプリコン混合産物が最終的に 4nM になるよう buffer EB (Qiagen)で調整した。4nM の混合産物 5 ul を、HT1 buffer (Illumina)にて最終濃度 10-11pM のシークエンスライブラリ 600ul に調整し、Illumina 社のマニュアルに従って、MiSeq v3 600cycle 試薬 (Illumina)にて MiSeq によるシークエンスを行った。

得られたリードデータに対しては、Trimmomatic-0.32 を用い、SE -threads 8 -phred33 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:20 MINLEN:150 のパラメーターでクオリティコントロールを行い、BWA-0.7.12[37] を用いて build37 + decoy 配列（ブロード研究所がウェブ上に公開: ftp://ftp.broadinstitute.org/）にアランメントを行った。得られた BAM ファイルは、マ

ッピングクオリティ 60 以上のデータのみ選択し、GATK-3.2-2 UnifiedGenotyper（パラメーターはデフォルト）を用いて、SNP のジェノタイピングをした。SNP タイピングには、リードペアの内、よりシークエンスクオリティが高いリードを採用した（原則フォワードリード)。

2.1.6.2 試料の混合による体細胞一塩基変異のシミュレーション

JM1 における SNP が Ref/Ref（リファレンス配列ホモ)、JM2 における SNP が Ref/Alt（ヘテロ SNP）のサイトで、Alt の割合が理論的に 1%, 2.5%, 5%, 10%となるように、JM1:JM2 を、49:1, 19:1, 9:1, 4:1 で混合し、体細胞 SNV をシミュレートした試料を作成した（サンプル ID: Mix1, Mix2.5, Mix 5, Mix10)。ゲノム DNA は、5ng/ul 前後に希釈した上で Qubit 2.0 Fluorometer, Qubit dsDNA HS Assay Kit (Qiagen)を用いて正確な濃度を求め、トータル 200-300ul になるよう 2 名のゲノム DNA を混合した。

2.1.6.3 ターゲットアンプリコンシークエンスによる解析

Mix1, Mix2.5, Mix 5, Mix10 に対して TAS を行い、対象 SNP サイトにおける Alt の割合を実験的に求め、理論的な Alt 割合と比較した。混合試料 DNA 7 ng を使用して PCR からシークエンスライブラリ調整を行った。使用したプライマーセット、シークエンスライブラリの調整、シークエンサーの稼働、シークエンスデータのクオリティコントロールはセクション 2.1.6.1 と同様である。対象 SNP サイトでベースクオリテ

ィ 20 以上の塩基数を、bam-readcount (https://github.com/genome/bam-readcount)で数え、体細胞 SNV をシミュレートした Alt の割合を求めた。割合の計算は、リードペアの内、よりシークエンスクオリティが高いリードを採用した（原則フォワードリード）。

2.1.7 ターゲットアンプリコンシークエンスによる体細胞一塩基変異候補の確認

WGS データから MuTect, Strelka にて検出され、フィルタリングにて選択された体細胞 SNV 候補を含むよう PCR プライマーを Primer-3plus で設計した。同時に、dbSNP に登録のある SNP サイトのうち、比較組織間で共通したヘテロ SNP とホモ SNP をランダムに選び、SNP サイトを含むように PCR プライマーを設計し、体細胞 SNV 候補に対する確認実験のポジティブコントロールとした（表 2 に示した dbSNP サイトとプライマーセット）。プライマーは、解析対象サイト以外の SNP, INDEL を避け、アダプター配列含めた最終フラグメントが、MiSeq (Illumina)によるシークエンスに最適な 345bp 程度になるよう設計した。まず、セクション 2.1.6.1 と同様に試験的 PCR を行い、複数のプライマー候補から、原則シングルバンドが得られるものを選択し、シングルバンドでない場合、1000bp 以下に別バンドがなく、かつ別バンドが 1 つまでのプライマーセットを選択した（図 3）。

選択したプライマーセットの配列を用い、セクション 2.1.6.1 と同様に、理論的深度 20 万前後の TAS（ターゲットアンプリコンシークエンス）を行った。使用したプライマーセットは表 4~7 に記載した。シークエンスライブラリの作成には、比較組織間

で等量（1 カ所につき 5~10ng）の対象試料ゲノム DNA を用い、WGS に用いたゲノム DNA と同じロットを用いた。1 カ所につき 5~10ng の DNA を用いた根拠は、ヒト 1 細胞における DNA が 6.6pg とすると、6.6ng で 1000 細胞となり、アレルの割合を定量するのに最小限に十分であると考えたからである。シークエンスライブラリ調整、シークエンサーの稼働、得られたリードデータに対する QC とアラインメントもセクション 2.1.6.1 と同様に行った。マッピングクオリティ 60 以上のアラインメントデータをから、bam-readcount を用いて体細胞 SNV 候補位置におけるベースクオリティ 20 以上の塩基数を数え、体細胞 SNV 候補の割合（Alt 割合）を計算した。割合の計算は、リードペアの内、よりシークエンスクオリティが高いリードを採用した（原則フォワードリード）。ベースクオリティ 25 がエラー率 $10^{-2.5}$ = 0.316%を意味することから、0.316%以下の割合のものはシークエンスエラー（体細胞変異ではない）と保守的に想定し、深度 5000 以下を深度不足と定義した。

脳神経組織でのバリデーションの際、同時に dbSNP サイト（体細胞 SNV 候補が存在していない dbSNP サイト）でも TAS を行い、実験上のポジティブコントロールとして、実験的に計算した Alt 割合と理論的 Alt 割合の比較を行った。ターゲットサイトとして用いた dbSNP 及びプライマーセットは、セクション 2.1.6.1 で用いたものである（表 2）。

2.1.8 ターゲットアンプリコンシークエンス結果の再現性の確認

TAS で得られた結果の信頼性を確認するため、TAS にて確認された体細胞 SNV 候補に対し、再度 PCR からシークエンスライブラリ調整を行い、TAS による再確認実験を行った。各サイトに対するプライマーセットは表 4~7 と同じ配列を用い、各試料同じロットのゲノム DNA を同じ量で使用した。シークエンスライブラリ調整、シークエンサーの稼働、シークエンスデータの解析はセクション 2.1.7 と同様に行った。

2.1.9 体細胞変異の生物学的特徴の解析と ROC 解析

最終的に確認された体細胞 SNV 候補サイトにおける WGS での Alt 割合と TAS での Alt 割合を比較した。SnpEff-4.1[48] による機能推定、ToppGene[49] による gene ontology 解析を行った。TAS にて確認された体細胞変異について、ベースクオリティ閾値による ROC 解析を、R 用パッケージの ROCR package[50] を用いて行った。

2.2 高深度全ゲノムシークエンスによる一卵性双生児不一致例における体細胞一塩基変異の探索

高深度全ゲノムシークエンスデータを一卵性双生児間で互いに比較することで、各個体に特異的な体細胞 SNV の探索を行った。体細胞 SNV 候補に対し、独立して超高深度ターゲットアンプリコンシークエンスを行い、Alt 割合を定量的に確認した。

2.2.1 試料の準備

1 人が統合失調感情障害に罹患しており、もう 1 人に精神疾患の既往歴のない 27 歳の男性一卵性双生児不一致ペアより提供された血液試料から、ゲノム DNA を採取した。被験者のデータについては表 1 に示した。試料提供者のリクルートは、共同研究者である加藤忠史、澤田知世が行い、血液細胞からのゲノム DNA 抽出は、三菱化学メディエンスの受託業務として行った。

2.2.2 シークエンス方法

血液細胞から抽出した全ゲノム DNA に対して TruSeq Nano DNA sample Prep kit (Illumina)を用い、製造者のプロトコルに従ってライブラリ調整を行った。調整したライブラリは、HiSeq X を用い、ペアエンド法にて理論的深度 80 で、各ペア 150bp で配列決定を行った。HiSeq X には構造上、機械に残留した DNA によるコンタミネーションがないため、インデックスは使用していない。HiSeq X の稼働は、理研ジェネシスでの受託業務として行った。

2.2.3 シークエンスデータのアラインメントとクオリティコントロール

得られたシークエンスデータのアラインメントとクオリティコントロールは、セクション 2.1.3 で記載した方法と同様であり、使用したパラメーターについては図 2 に記載した。HiSeq X での配列決定は、HiSeq 2500 Rapid Run モードより精度が劣るが、この解析では理論的深度が脳神経組織の解析よりも小さいため、閾値設定は若干緩め、

fastx_toolkit-0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html)の mask 機能を使用し、各リードのベースクオリティ 10 以下の塩基を全て N に置換することで、低クオリティの塩基による偽陽性を予防した。

2.2.4 体細胞一塩基変異候補の検出

得られたアラインメントデータ（BAM ファイル）に対して、MuTect, Strelka を用い、一卵性双生児ペア間の血液由来ゲノムの比較を行うことで、体細胞 SNV 候補を検出した。一方の組織では数%の割合で認めるが、対照組織では存在を認めない体細胞 SNV を探索した。Strelka はデフォルトのパラメーターを用い、MuTect はデフォルトモード（デフォルトのパラメーター）で解析を行った。デフォルトモードは、セクション 2.1.4 で記載した通り対照組織に体細胞変異が原則ない状態を仮定して解析を行うモードであり、一卵性双生児で生じる体細胞は原則共通しないと考えられるので、この形で解析を行った。以降この方法による MuTect の解析結果の記載はデフォルトモードの結果である。

2.2.5 体細胞一塩基変異候補のフィルタリング

この解析では、多コピー領域上の候補を除外した領域の解析と、多コピー領域上の候補に分けて解析を行った。まず、セクション 2.1.5 と同様に、MuTect 結果から多コピー領域に存在する体細胞 SNV 候補を除外したリストを作成し、GATK-3.2-2

UnifiedGenotyper でコールされた INDEL の前後 10bp 上の体細胞 SNV 候補は除外した。得られた候補リストの中には、同一個体由来のデータを合わせても、数百 Kbp 内に検出された候補が密集している領域は認めず、本解析では、多コピー疑い領域として除外した領域はなかった。

セクション 3.1 の結果を踏まえて、体細胞 SNV 候補のフィルタリングを行うための閾値を以下のように設定した：体細胞 SNV を示唆するベースコールの平均ベースクオリティが 20 以上、候補サイトの深度が 30 以上、候補サイト前後 150bp（トータル 300bp）の UCSC BLAT スコアが 150 未満。Strelka では平均ベースクオリティの代わりに、QSS (Quality Score for Somatic SNV)という値が独自に計算されており、Strelka でのフィルタリングは、QSS が 20 以上とした。今回得られた双生児の WGS データは理論的深度がより小さいため、深度のフィルタリング閾値は緩めに設定した。セクション 3.1.3 で定義した SBC (Supporting Basecall Count)を用い、下記のように信頼性の高い候補（HC, high confidence）、信頼性のより低い候補（LC, low confidence）に分けた。一卵性双生児同士で、同一サイトに同じ SNV が生じる可能性は無視できるほど小さいと考えられるため、脳神経組織での解析と異なり、対照試料では同じ体細胞 SNV がまったく存在しないものと仮定した。

NonRepeat HC: SBC が 2 個以上であり、STR（short tandem repeat, 以下 STR）の切り替わりや poly-A 領域の両端ではなく、対照双生児ゲノムに同一のコールが存在しない。

NonRepeat LC: SBC が 1 個であり、STR の切り替わりや poly-A 領域の両端ではない。または、SBC が 2 個以上だが、STR の切り替わりや poly-A 領域の両端にある。いずれも対照双生児ゲノムに同一のコールが存在しない。

多コピー領域上の候補は、相同配列が頻出すると考えられるが、多コピー領域末端の候補や相同配列の多様性が高いゲノム領域の候補に関しては、特異的にアラインメントされるものもあると考えられ、そのような特異的にアランメントされている候補は、別途解析を行った。BAM ファイルからマッピングクオリティ 60 以上のリードのみ選択し、MuTect の解析を行い（デフォルトモード)、多コピー領域（1.63Gbp）上の候補を選択した。その後、GATK-3.2-2 の UnifiedGenotyper でコールされた INDEL の前後 10bp 上の体細胞 SNV 候補は除外した。ほとんどの多コピー領域は対象にならないため、この部分は網羅的な解析とはなっておらず、探索的な手法である。操作的なフィルタリングの方法は上述と同様で、下記のように HC, LC に分けた。なお、考察で後述するように、Strelka のみで検出された SNV がこれまでの解析でなく、これまでの解析よりも特異性が求められる解析であるため、この解析では Strelka は用いていない。

Repeat HC: SBC が 2 個以上であり、STR の切り替わりや poly-A 領域の両端ではなく、対照双生児ゲノムに同一のコールが存在しない。

Repeat LC: SBC が 1 個であり、STR の切り替わりや poly-A 領域の両端ではない。または、SBC が 2 個以上だが、STR の切り替わりや poly-A 領域の両端にある。対照双

生児ゲノムに同一のコールが存在しない。

2.2.6 ターゲットアンプリコンシークエンスによる体細胞一塩基変異候補の確認

同じプラットフォームを用いた脳神経組織での解析（CC_WGS_set）において、LC でのバリデーション率がゼロまたは解析困難であったため、この解析では NonRepeat HC, Repeat HC をバリデーションの対象とし、体細胞 SNV 候補サイトを含むよう PCR プライマーを設計した。プライマー設計、シークエンスライブラリ調整、シークエンサーの稼働、シークエンスデータの解析の方法はセクション 2.1.7 と同様である。使用したプライマーセットは表 8 に記載した。

2.3 全エクソームシークエンスによる双生児不一致例における体細胞一塩基変異の探索

高深度全エクソームシークエンス（whole exome sequence, 以下 WES）データを、一卵性双生児間で互いに比較することで、各個体に特異的な体細胞一塩基変異候補の探索を行った。独立して超高深度ターゲットアンプリコンシークエンスを行い、体細胞一塩基変異候補の割合を定量的に確認した。

2.3.1 試料の準備

統合失調症圏の疾患について不一致である一卵性双生児ペア 4 例より、共同研究者

である西村文親、吉川茜、垣内千尋、佐々木司が血液試料を採取した。被験者のデータについては表 1 に示した。血液細胞からのゲノム DNA 抽出は、三菱化学メディエンスの受託業務として行った。

2.3.2 シークエンス方法

血液細胞から抽出したゲノム DNA に対して、SureSelect V4/V5+UTR (Agilent)を用いて、エクソン領域を濃縮した。濃縮した DNA 試料をサンプル調整し、HiSeq 2000 にて各ペア 100bp で配列決定を行った。HiSeq 2000 は構造上、機械に残留した DNA によるコンタミネーション（キャリーオーバー）はないため、シークエンスによるサンプルコンタミネーションはないものと考えた。以上は、共同研究者の西村文親、吉川茜、垣内千尋、佐々木司が、東京大学医学部附属病院ゲノム医学センターのゲノム支援事業受託業務として行ったものである。

2.3.3 シークエンスデータのアラインメントとクオリティコントロール

得られたシークエンスデータのアラインメントとクオリティコントロールは、セクション 2.1.3 と同様であり、使用したパラメーターについては図 1 に記載した。

2.3.4 体細胞一塩基変異候補の検出

得られたアラインメントデータ（BAM ファイル）に対して、MuTect, Strelka を用い、

一卵性双生児ペア間の血液由来ゲノムの比較を行うことで、体細胞 SNV 候補を検出した。一方の組織では数%の割合で認めるが、対照組織では存在を認めない塩基を探索することを想定した。Strelka はエクソーム用に許容深度を調整した（isSkipDepthFilters = 1）以外はデフォルトのパラメーターを用い、MuTect はデフォルトモード（デフォルトのパラメーター）で解析を行った。デフォルトモードは、前述の通り対照組織に体細胞 SNV が原則ない状態を仮定して解析を行うモードであり、一卵性双生児で生じる体細胞は原則共通しないと考えられるので、この形で解析を行った。以降この方法による MuTect の解析結果という記載はデフォルトモードの結果である。

2.3.5 体細胞一塩基変異候補のフィルタリング

エクソーム解析は、原則エクソンを対象としており、リピート配列など他のゲノム領域に相同配列が存在するケースは少ないと考えられるため、多コピー領域の除外は行わなかった。INDEL の前後は正確なアラインメントが難しく SNV の解析が困難であるため、GATK-3.2-2 UnifiedGenotyper でコールされた INDEL の前後 10bp 上の体細胞 SNV 候補は除外した。また多コピー疑い領域の候補はなかったため、これらの除外も行っていない。

体細胞 SNV 候補の操作的なフィルターを行うための閾値を設定した。体細胞 SNV を示唆するベースコールのベースクオリティが 20 以上、候補サイトの深度が 30 以上、

候補サイト前後 100bp（トータル 200bp）での UCSC BLAT スコアが 160 未満の閾値を用いた。Strelka では平均ベースクオリティの代わりに、QSS (Quality Score for Somatic SNV)という値が独自に計算されており、Strelka でのフィルタリングは、QSS が 20 以上とした。エクソン周辺に限られた解析のため、より広く候補を挙げるため、脳神経組織おける解析よりも緩めた閾値を使用した。フィルタリングされた候補から、セクション 3.1.3 で定義した SBC とシークエンス文脈を用い、下記のように信頼性の高い候補（HC, high confidence）、信頼性のより低い候補（LC, low confidence）に分けた。

HC: SBC が 2 個以上であり、STR の切り替わりや poly-A 領域の両端ではない。また対照双生児でのコールが存在しない。

LC: SBC が 1 個であり、STR の切り替わりや poly-A 領域の両端ではなく、対照双生児でのコールが存在しない。または、SBC が 2 個以上だが、STR の切り替わりや poly-A 領域の両端にあり、対照双生児でのコールが存在しない。

2.3.6 ターゲットアンプリコンシークエンスによる体細胞一塩基変異候補の確認

脳神経組織での解析結果（セクション 3.1）では、PCR を含んだサンプル調整によるシークエンスデータの解析で LC のバリデーション率がゼロまたは解析困難であったため、この解析では HC の体細胞 SNV 候補のみを選択し、PCR プライマーを設計した。プライマー設計、シークエンスライブラリ調整、シークエンサーの稼働、シー

クエンスデータの解析はセクション 2.1.7 と同様である。使用したプライマーリストは表 9 に記載した。確認された体細胞 SNV に対し、SnpEff-4.1 による機能推定を行った。

2.3.7 パイロシークエンスによる体細胞一塩基変異候補の確認

セクション 2.3.6 の方法にて確認された体細胞 SNV に対し、パイロシークエンス法による確認実験を行った。PCR に使用したプライマーセットの配列は表 9 と同様で、片方の 5'末端にビオチン修飾を付加したプライマーセットを用いた（表 10）。NEBNext HighFidelity 2X PCR MasterMix を用い、ビオチン修飾つきプライマーセットにて PCR を行った。ゲノム DNA 10 ng, Q5 High-Fidelity DNA Polymerase, dNTP 200 uM/each, Mg++Cl 2 mM, 各プライマー 0.5 uM, 最終容量 30 ul となるようサンプルを調整した。サーマルサイクラーは「98ºC にて 30 秒」、「98ºC を 10 秒、64ºC を 10 秒、72ºC を 10 秒」を 33 回繰り返し、「72ºC を 5 分」の設定にて稼働した。よりプライマー結合の特異性を高める厳密な解析とするため、アニーリング温度を、セクション 2.1.6 の設定である 62ºC から 64ºC に変更した。

PCR プロダクト 30 ul に対し、Streptavidin Sepharose High Performance (GE Healthcare) 1.5 ul, PyroMark Binding Buffer (Qiagen) 40 ul, Milli-Q 水 8.5 ul を加え、10 分震盪させてビオチン付加した PCR プロダクトを不動化した。処理したサンプルをバキュームプレップツール (Qiagen)で吸引し、70%エタノールで 5 秒、0.2N 水酸化ナトリウム

水溶液で 5 秒、10mM Tris-HCl (pH 7.6)で 10 秒洗浄し、PyroMark Annealing Buffer (Qiagen) 38.5ul とシークエンスプライマー(10uM) 1.5ul の溶液に吸引物を溶かした。シークエンスプライマーの配列は表 10 に記載した。調整したサンプルは、PyroMark Q96 (Qiagen)をメーカーのマニュアル通りに稼働し、AQ (allele quantification)モードでパイロシークエンスを行った。

2.4 高感度 L1Hs-seq の開発と体細胞新規挿入検出能力の検討

LINE1 配列を特異的に増幅する方法として L1Hs-seq が報告されている [35]。LINE1 配列の中でも唯一自律的な転移活性を持つとされている L1Hs 配列に注目し、L1Hs3’末端配列に特異的なプライマーを片側に用いることで L1Hs 及びその転移後の配列を含む領域を増幅する方法である。この L1Hs-seq を改善し、体細胞変異のようにアレル割合の低い LINE1 配列のレトロトランスポジション検出法を開発した（改良型 L1Hs-seq）。開発した改良型 L1Hs-seq に対しては、リファレンス上の L1Hs に対する感度と、リファレンスにない位置での L1Hs 配列挿入の検出、体細胞新規転移をシミュレーションした人工遺伝子配列の検出能力を評価した。

2.4.1 試料の準備

31 歳日本人男性（健常者）由来の血液を採取し、ゲノム DNA を抽出した。ゲノム DNA の抽出は三菱化学メディエンスの受託業務として行った。体細胞新規転移のシ

ミュレーションとして、図 4 に記載した人工配列（L1Hs 配列＋ランダム配列）を pMK-RQ プラスミドに挿入した全長 3777-3778bp の人工遺伝子を 3 種類合成し(Life Technologies)、一定の割合（0.5~50%）で上記のヒトゲノム DNA と混合した（サンプル ID: IC1, IC2, IC3）。また、表 2 に記載した S6_cerebellum の試料も本実験に使用した。

2.4.2 シークエンスライブラリ調整・シークエンス方法

L1Hs 3’末端特異的なプライマーを片側に用い、片側に 8 種の縮重プライマーを用いることで、まず L1Hs 3’末端と隣接領域を網羅的に増幅した。8 種の縮重プライマーごとに PCR を行い、1 試料につき 8 反応用意した。次に、シークエンス用のアダプターを付加した L1Hs 3’末端に特異的な第 2 のプライマーを用いて PCR することにより、L1Hs 3’末端の隣接領域を特異的かつ網羅的に増幅し、シークエンスライブラリを作成した（図 5）。プライマーのリストは表 11 に示した。

2 回の PCR ともに NEBNext High-Fidelity 2X PCR MasterMix にて行った。1 回目の PCR は、まずフォワードプライマーL1HsTAILSP1A2 のみ用い（最終濃度 0.8 uM）、1 チューブにつき、DNA 200ng (8 チューブで 1600ng), Q5 High-Fidelity DNA Polymerase, dNTP 200 uM/each, Mg++Cl 2 mM, 最終容量 25ul となるよう調整し、「98ºC を 30 秒」、「98ºC を 10 秒、60ºC を 1 分、72ºC を 45 秒」を 5 回繰り返した後、リバースプライマー（縮重プライマー、最終濃度 0.2 uM）を添加して、「98ºC を 10 秒、58ºC を 30 秒、

72ºC を 30 秒」を 14 回繰り返し、「72ºC を 5 分」の条件でサーマルサイクラーを稼働した。PCR 後のプライマー除去・精製は、SPRIselect によって行い、1 チューブにつき 10 ul となるようプロダクトを濃縮した。2 回目の PCR は、濃縮した PCR プロダクト 3ul を用い、Q5 High-Fidelity DNA Polymerase, dNTP 200 uM/each, Mg++Cl 2 mM, 各プライマー0.5 uM, 最終容量 25ul となるよう調整し、「98ºC を 30 秒」、「98ºC を 10 秒、62ºC を 30 秒、72ºC を 30 秒」を 14 回繰り返し、「72ºC を 5 分」の条件でサーマルサイクラーを稼働した。PCR 後のプライマー除去・精製は、SPRIselect によって行い、最終的なシークエンスライブラリが、270~1500bp となるように精製した。8 種の縮重プライマーごとに得られたプロダクトを、Qubit 2.0 Fluorometer, Qubit dsDNA HS Assay Kit で濃度測定し、等濃度で混合した。混合産物は Bioanalyzer (Agilent)で分布を確認し（図 6）、Qubit 2.0 Fluorometer, Qubit dsDNA HS Assay Kit で正確に濃度測定した上で、最終的に 4nM になるよう buffer EB (Qiagen)で調整した。混合産物 5 ul (4nM)を MiSeq v3 600cycle 試薬 (Illumina)にて準備し、Illumina 社のマニュアルに従って、最終濃度 11-12 pM のシークエンスライブラリ 600ul にて MiSeq を稼働した。

2.4.3 シークエンスデータ解析

まず、得られたシークエンスデータを図 7 のように処理を行い、クオリティコントロール後のリードデータを得た。まず Cutadapt[51] を用いて、フォワードリード 5’末端が L1Hs の配列にマッチするものを選択し、アダプター配列やクオリティの低いベー

スコールを除去した後（後述の閾値）、ペアリードとして残るものとシングルリードになったものを分離した。前者のペアリードに関しては、FLASh-1.2.11[52] を用い、厳しい閾値（-M 300 -m 20 -x 0.1）でスティッチング（フォワードリードとリバースリードに重複する配列が十分な長さで存在する場合、重複した部分を縫い合わせるように重ねてフォワードリードとリバースリードを結合し、1 本のリードとする作業）することで、ペアリードから元のフラグメントを構成し直した。この操作で、スティッチングした長いリード、スティッチングされなかったペアリード、シングルリードの 3 種類が分割され、各々に対し BWA-0.7.12 により hg19 リファレンスゲノム及び 3 種の人工遺伝子配列にアラインメントを行った。非リファレンス位置での挿入を検出するため、LINE1 配列を含んだ decoy 配列はリファレンスとして使用していない。アラインメントデータの QC は、PCR や光学系による重複リードを除去するための Picard Deduplication (Picard-1.102)、INDEL 領域のアラインメントを正確に行うための GATK Indel Realignment (GATK-3.2-2)、ベースコールのベースクオリティを補正する GATK Base Recalibration (GATK-3.2-2)を、記載の順で行った。Samtools-0.1.19 にてマッピングクオリティ（mapQ）による選択を行い、3 種のリードデータから生成されたアラインメントデータをマージした。参照した SNP サイトや INDEL 領域のファイルは 1000G_phase1.indels.hg19.vcf, dbsnp_137.hg19.vcf, Mills_and_1000G_gold_standard.indels.hg19.vcf である（ブロード研究所がウェブ上に公開: ftp://ftp.broadinstitute.org/）。

マージしたBAMファイルから、bedtools[53]によって、bedgraphを生成し、アラインメント距離が1000bp以内のリードデータをブロックとしてまとめ、以下の2種類の閾値設定（S: stringent, R: relaxed）で、ブロックごとにアラインメントでカバーされたゲノム領域の長さ(bp)を計算した。

<閾値設定S> Cutadaptにおいて-q 28 -m 30で解析を行い，マッピングクオリティ30以上のリードのブロックがゲノム領域にして350 bp以上をカバー

<閾値設定R> Cutadaptにおいて-q 20 -m 20で解析を行い，マッピングクオリティ0以上のリードのブロックがゲノム領域にして100 bp以上をカバー

以上の閾値設定で得られたブロック（領域）を用い、リファレンス上のL1Hsに対する検出感度、非リファレンス位置でのL1Hs配列の挿入、人工遺伝子配列の検出の有無を解析した。ヒトリファレンスゲノムhg19上にL1Hsは1544ヶ所アノテーションされているが[44]、L1HS-seqの2回目のPCRに用いたプライマーの3’末端Gに相当する配列は813ヶ所であり[22]、活性を持つすべてのL1Hsが含まれていると考えられている[22,35]。この813ヶ所をリファレンスL1Hsとし、本実験にて検出できた割合を感度として計算した。また、非リファレンスL1Hs（リファレンス上に存在しないL1Hs）の挿入検出数を計算し、既に文献上報告されている非リファレンスL1Hs（Known NonReference L1Hs, 以下KNR L1Hs）[22,54]を除外した数も計算した。なお、文献として使用した1000ゲノムプロジェクトの解析対象2,504名には、104名の日本人試料が含まれており、LINE1新規挿入として全体で3,060カ所報告されている。

シークエンスデータのリード数による解析結果の違いを調べるため、IC1 のデータから、get_subset.py (https://github.com/happykhan/nfutil)を使用してランダムにリードを抽出し、擬似的に低リード数のシークエンスデータを生成した。上記と同様の解析にてリファレンス上の L1Hs に対する検出感度、非リファレンス位置での L1Hs 配列の挿入検出数を解析した。

2.4.4 非リファレンス L1Hs 挿入部位のバリデーション

非リファレンス位置での L1Hs 挿入を認めたゲノム領域をランダムに抽出し、挿入位置を挟むように L1Hs の 3'末端配列と挿入位置のリファレンスゲノム配列から PCR プライマーを設計し、Nested PCR にて増幅したプロダクトをサンガー法により配列決定した。

まず、L1Hs の 3'末端配列に対してフォワードプライマーが、挿入位置のリファレンスゲノム配列に対してリバースプライマーが結合するよう Primer-3plus で複数設計した（表 12)。1 回目の PCR は、対象試料のゲノム DNA 10ng, Q5 High-Fidelity DNA Polymerase, dNTP 200 uM/each, Mg++Cl 2 mM, 各プライマー 0.5 uM, 全体容量 10 ul となるよう調整し、サーマルサイクラーを「98ºC にて 30 秒」、「98ºC を 10 秒、61ºC を 30 秒、72ºC を 15 秒」を 15 回繰り返し、「72ºC を 5 分」の設定にて稼働した。SPRIselect にて PCR プロダクトの精製を行い 10 ul とした。2 回目の PCR は、精製 PCR プロダクト 1 ul を用い、Q5 High-Fidelity DNA Polymerase, dNTP 200 uM/each, Mg++Cl 2 mM,

各プライマー 0.5 uM, 全体容量 10 ul となるよう調整し、サーマルサイクラーを「98ºC にて 30 秒」、「98ºC を 10 秒、56ºC を 30 秒、72ºC を 15 秒」 を 25 回繰り返し、「72ºC を 5 分」の設定にて稼働した。2 回の PCR とも NEBNext HighFidelity 2X PCR MasterMix を使用した。

2 回目の PCR プロダクト 10 ul を 2% agarose gel で電気泳動し、標的長さのバンドを含むようゲルを切り出した。切り出したゲルを 95ºC で 10 分、65ºC で 10 分インキュベートし、Thermostable β-Agarase (Nippon Gene)をゲル 100mg に対し 3 ul となるよう加え、65 ºC で 10 分インキュベートした。氷上に置いた後、常温にて 20,000G で 15 分遠心し、沈渣物を除いた液に 1/10 量の 3M 酢酸ナトリウム、等量のイソプロパノール、Dr. GenTLER Precipitation Carrier (Takara) 4ul を加え混合した後、常温で 10 分インキュベートした。常温にて 12,000G で 5 分遠心し上清を捨てた後、70%エタノール 1 ml 加え、常温にて 12,000G で 5 分遠心し上清を捨てた。ペレットを常温で 15 分インキュベートした後、buffer TE を加えた。ゲルから精製した PCR プロダクトに対し、2 回目の PCR で使用したフォワードプライマー、リバースプライマーを用いてサンガー法による配列決定を行った。サンガー法による配列決定はユーロフィンジェノミクス社の受託業務として行った。

3. 結果

脳神経組織の WGS データ、一卵性双生児の WGS/WES データから検出・フィルタリングされた体細胞 SNV 候補に対し超高深度 TAS を用いて、体細胞 SNV 候補のバリデーションを行った。結果、脳神経組織において 31 カ所（CL_WGS_set で 6 カ所、NeuN_WGS_set で 12 カ所、CC_WGS_set で計 13 カ所）、MZ_Exome_set において 7 カ所の体細胞 SNV が確認された。超高深度 TAS による体細胞 SNV 検出の妥当性は、シミュレーション試料（2 名の DNA を一定割合で混合した試料）を用いて確認した。

3.1 高深度全ゲノムシークエンスによる脳特徴的な体細胞一塩基変異の探索

脳神経組織における WGS データの要約を表 13 に示した。WGS データに対して、FastQC によってリードデータのクオリティチェックを行い、シークエンスプラットフォームに応じて体細胞 SNV 探索に最適となるよう、各ソフトウェアのパラメーターを設定した（図 2）。HiSeq X によるシークエンスデータを FastQC で確認したところ、ベースクオリティの低いリードが多かったため、QC 基準を HiSeq 2500 によるシークエンスデータより厳しくして偽陽性を予防した（図 8）。QC が厳しすぎると深度が減少するため、そのトレードオフを勘案したバランスを取った。

3.1.1 全ゲノムシークエンスデータからの体細胞一塩基変異候補の検出（マニュアル法）

CL_WGS_set、NeuN_WGS_set で MuTect 高感度検出モードによる体細胞 SNV 候補の検出を行ったところ、平均 20,073 個の候補が検出された（表 14）。これらの大部分（平均 99.3%）は、多コピー領域、INDEL 領域、多コピー疑いの領域に位置し、以降の解析から除外した。多コピー領域は計 1.63 Gb であり、多コピー領域を除外した 1.46 Gb が体細胞 SNV 候補サイト検出の対象領域となった。

多コピー疑いとなる領域は 4 領域（表 15）であり、後述する CC_WGS_set を含め脳神経組織の解析で使用した 3 個体でほぼ共通して出現した。多コピー疑いの 4 領域における MuTect による候補検出数は平均 8.2 kb に 1 つであり、多コピー疑い除外後の体細胞 SNV 候補検出数が平均 1.36 Mb に 1 つであることを考えると明らかに密集していた。CL_WGS_set における多コピー疑い領域内では、AL30_cortex と AL30_liver で偏りなく候補が検出されており、NeuN_WGS_set でも多コピー疑い領域内で各試料からの候補が偏りなく検出されていた。いずれも、生殖系列ゲノムの特徴としての多コピー疑い領域であること（相同配列が他のゲノム領域に存在すること）に支持的であり、除外は適当であると考えた。CL_WGS_set には、chr6:167607240-167781635 の領域の記載はないが、当該領域に MuTect による候補検出が 6 個あり、これを含めると 4 領域は全て共通した。

CL_WGS_set、NeuN_WGS_set で Strelka による体細胞 SNV 候補の検出を行ったところ、平均 469 個の候補が検出された（表 14）。これらの大部分（平均 95.9%）も多コピー領域、INDEL 領域であり、多コピー領域、INDEL 領域上の候補は、以降の解

析から除外した。なお、本実験では脳神経組織に特徴的な体細胞 SNV の検出を目指しており、解析対象試料を肝臓（AL30_liver, Y8763_liver）として検出された体細胞 SNV 候補は、フィルタリングまで行ったが、TAS によるバリデーションの対象としなかった。

まず、AL30_cortex を解析対象試料とし、AL30_liver を比較対照試料として体細胞 SNV 候補の検出を行った。MuTect にて 21,312 個、Strelka にて 354 個検出されたが、多コピー領域、INDEL 領域、多コピー疑い領域上の候補を除外したところ、MuTect では 125 個、Strelka では 14 個に絞られた（表 14）。絞られた計 139 個を IGV で目視して確認し、ソフトクリップ近傍に集積した候補や近くにミスマッチが頻出した候補を、配列決定精度・アラインメント精度が低い領域上にあるものとして除外し、最終的な体細胞 SNV 候補を 35 個選択した（マニュアル法）。

3.1.2 超高深度ターゲットアンプリコンシークエンスの妥当性の検証

脳神経試料における体細胞 SNV 候補の確認実験を行うために、まず TAS（ターゲットアンプリコンシークエンス）による体細胞 SNV 検出の妥当性を確認した。JM1, JM2 に対して、計 12 カ所の dbSNP サイトの SNP タイピングを、TAS にて行った（表 16）。JM1 と JM2 で計 7 カ所の dbSNP サイトが異なるジェノタイプであり、JM1, JM2 を混合した試料にて体細胞 SNV をシミュレートし、7 カ所の dbSNP サイトに対して解析を行った。

JM1 と JM2 を一定割合で混合することで、1~10%の体細胞 SNV をシミュレートした DNA を 4 試料（Mix1, Mix2.5, Mix5, Mix10）調整した（表 15）。混合した 4 試料に対し 7 カ所の TAS を行い、dbSNP サイトにおける Alt 割合を計算した。理論的な Alt 割合と実験により測定された Alt 割合の比較を図 9 に示した。理論的 Alt 割合と実験的 Alt 割合は、高い相関（Pearson's r = 0.969, $p < 2.2\times10^{-16}$）を示した。理論的 Alt 割合（%）を説明変数 x、実験的 Alt 割合（%）を目的変数 y として回帰分析（最小二乗法）を行うと、y = 1.39x + 0.62 の関係が得られ、残差の標準誤差は 1.27 (%)であった。残差は最小で-1.94%、最大で 2.56%であった。1.27 (%)の標準誤差から、本実験における TAS の Alt 割合の数字が示す精度については、95%信頼区間が± 2.49%であると言える。

脳神経組織における体細胞変異候補の TAS バリデーションの際に、実験上のポジティブコントロールとして同時に測定した dbSNP サイトでの Alt 割合は、理論値 50%のヘテロ SNP サイト（のべ 35 カ所、平均深度 20.6 万）では、49.61% ± 0.955 (平均±標準偏差)であり、理論値 100%のホモ SNP サイト（のべ 23 カ所、平均深度 20.9 万）では、平均 99.98% ± 0.014 (平均±標準偏差)であった。TAS にてコントロールサイトとして用いた dbSNP で得られた実験的割合は、理論的割合に近似しており、かつ標準偏差も小さいと言える。

のべ 23 カ所の理論値 100%のホモ SNP サイトにおける名目上のエラー率は 0.028%であった。ベースクオリティの値が完全に信頼できると仮定して、Alt ベースコールの数（Alt_BC_count）と平均ベースクオリティ（Alt_BQ）、それ以外のベースコール

の数（Others_BC_count）と平均ベースクオリティ（Others_BQ）から、Alt_BC_count * $10^{[-Alt\_BQ/10]}$ + Others_BC_count * (1 - $10^{[-Others\_BQ/10]}$)の計算にて、Alt ベースコール以外の数を求めそのアレル割合を計算すると、23 カ所で平均 0.046%であった。ベースクオリティの値が完全に信頼できるものと仮定すると、これは全ての塩基に対する PCR エラーの割合に相当すると考えられる。

3.1.3 CL_WGS_set における体細胞一塩基変異候補の確認

AL30_cortex の解析（マニュアル法）にて最終的な体細胞 SNV 候補として選択された 35 候補に対し、TAS でバリデーションを行い、当該サイトにおける体細胞 SNV 候補（Alt）の割合を計算した（表 17）。加えて、多コピー疑いの 3 領域から試験的 PCR によりシングルバンドの得られた候補を 1 個ずつ選び、バリデーション対象とした。しかし、多コピー疑いの 3 カ所はマッピングクオリティ 60 以上のリードがなく、ゲノム上に相同配列の存在が示唆された。35 ヶ所の体細胞 SNV 候補のうち、6 カ所で Alt 割合差を高い信頼性で認め、6 個の体細胞 SNV が存在するものと考えた。確認された体細胞 SNV 候補サイトを IGV にて可視化した例を図 10 に示した。バリデーションされた体細胞 SNV は、体細胞 SNV を示唆するベースコールの平均ベースクオリティ 30 以上であり、候補サイトの深度が 60 以上、マッピングクオリティが 60 以上であった。うち 1 カ所（chr6: 164440297）は、ヘテロ SNP 近傍に位置し、WGS データ上でも TAS データ上でも、体細胞 SNV 候補が片方の SNP のみとリードを共有してお

り、体細胞 SNV であることと矛盾はなかった（図 10）。解析の難しい候補は、STR（STR）の切り替わりサイト（例えば、TGTGTG<T/A>GAGAGAG の<T/A>）や poly-A 領域など、配列決定や正確なアラインメントが難しい領域に位置していた。特に、STR の切り替わりサイトはリードごとに異なる INDEL が出現し、正確なアラインメントが難しく、Alt 割合を求めるのが困難であった（図 10）。

マニュアル法で抽出された体細胞 SNV 候補のうち、TAS でバリデーションされた候補・されなかった候補・解析の難しい候補の比較を行うことで、客観的なフィルタリングを行うための閾値を設定した。網羅的な検出を目指し、以下のような方法での操作的なフィルタリング法を設定した（操作的方法）: 体細胞 SNV を示唆するベースコールの平均ベースクオリティが 21 以上、候補サイトの深度が 40 以上 （Hiseq X ではシークエンスクオリティが低いため 50 とした）、候補サイト前後 150 bp（トータル 300 bp）の配列が UCSC BLAT スコアで 150 未満。ベースクオリティ 21 は閾値として緩めであるが、本解析では感度を優先し緩い閾値とした。Strelka では平均ベースクオリティの代わりに、QSS (Quality Score for Somatic SNV)という値が独自に計算されており、Strelka でのフィルタリングは、QSS が 21 以上とした。これにより、体細胞 SNV を示唆する塩基が明らかなミスコールではなく、生殖系列ゲノム SNV が偽陰性とならない深度があり、他のゲノム領域に相同配列がないものと期待できる。以上のフィルタリングを行った後、体細胞 SNV を支持するベースコールについて、(1) ベースクオリティが 25 以上、(2) リードのマッピングクオリティが 30 以上、(3) リードの両

端 10 bp 内に位置しない、(4) 同じリードの前後 15 bp にマイナーミスマッチ（当該サイトにおいて 20%以下の割合で存在するミスマッチ）や INDEL がない、(5) リードに XA タグがない（代替的なアラインメントがない）、(6) 同じリードに 10 bp 以上のソフトクリップがない、(7) INDEL リアラインメントにより Ref と一致しない、という条件を満たすものを Supporting Basecall と定義し、その数を Supporting Basecall Count (SBC)とした。また、体細胞 SNV 候補サイトが STR（ショートタンデムリピート）の切り替わりや poly-A 領域末端であるかどうか（シークエンス文脈）を考慮した。最終的に体細胞 SNV 候補を、SBC が 2 個以上であり、STR の切り替わりや poly-A 領域末端ではない High Confidence (HC)と、SBC が 1 個であり、STR の切り替わりや poly-A 領域末端ではない、または、SBC が 2 個以上だが、STR の切り替わりや poly-A 領域末端にある Low Confidence (LC)とに分けた。

CL_WGS_set の比較結果について、操作的方法と SBC・シークエンス文脈にて再度検討したところ、6 個の HC、4 個の LC が同定された（表 14, 18）。この基準で選択された候補のうち、TAS 法でバリデーションに成功した確率を再計算すると、HC で 100%（6/6）、LC で 0%（0/3）となった。ただし、操作的方法で同定した LC サイト 4 ヶ所のうち 1 カ所（chr8: 106342766）はマニュアル法による TAS のバリデーション対象から漏れており、操作的方法による LC を十分には評価できなかった。バリデーションの対象としなかったものの、AL30_liver を解析対象試料とした解析では、操作的方法と SBC・シークエンス文脈考慮後の HC が 22 個、LC が 4 個と AL30_cortex にお

ける HC, LC より多数の体細胞 SNV 候補が検出された。

3.1.4 NeuN_WGS_set における体細胞一塩基変異候補の確認

続いて CL_WGS_set と同じプラットフォームである Hiseq 2500 Rapid Run モードを用いた NeuN_WGS_set を解析した。脳組織における体細胞 SNV の探索を目的としているため、Y8763_liver を解析対象試料とする比較の結果はバリデーション対象から除外し、Y8763_NeuN+, Y8763_NeuN-を解析対象試料とする 4 パターンの解析で得られた体細胞 SNV 候補を TAS によるバリデーションの対象とした。この解析では、MuTect にて平均 20,352 個の体細胞 SNV 候補が検出され、多コピー領域、INDEL 領域、多コピー疑い領域上の候補を除外すると 107 候補となった。Strelka では平均 395 個の体細胞 SNV 候補が検出され、多コピー領域、INDEL 領域上の候補を除外すると平均 15.8 候補となった。MuTect 解析結果で平均 99.5%、Stelka 解析結果で平均 96.0%が多コピー領域上に存在しており、CL_WGS_set の結果と同様に多コピー領域上の体細胞 SNV 候補が多いという結果であった（表 14）。

その後、セクション 3.1.3 で設定した操作的方法によるフィルタリングを行い、SBC・シークエンス文脈を考慮して、MuTect, Strelka による解析ごとに HC, LC を選択した。4 パターンの解析の総計で、MuTect 解析により 26 個の HC と 9 個の LC、Strelka 解析により 3 個の HC が、最終的な体細胞 SNV 候補として選択された（表 14）。MuTect と Strelika で、サイトが重複する候補が複数あり、ゲノム上の位置としては計 24 カ所

（MuTect による HC サイト 16 カ所・LC サイト 7 カ所、Strelka による HC サイト 1 カ所）が体細胞 SNV 候補の存在するサイトの候補となった。1 カ所（HC）は特異的なプライマーセットが設計できず、23 カ所のサイトを含むよう TAS を行った。バリデーションの対象としなかったものの、Y8763_liver を解析対象試料とした解析では、操作的方法と SBC・シークエンス文脈考慮後の HC が平均 88 個、LC が平均 7 個であり、Y8763_NeuN+, Y8763_NeuN-における HC, LC より多数の体細胞 SNV 候補サイトが検出された。

TAS データの QC 後平均深度は 19.5 万となり、12 カ所で体細胞 SNV 候補（Alt）の割合差が認められ、体細胞 SNV であることが確認できた（表 19）。一方、poly-A 領域（LC）においてシークエンスクオリティの落ち込みが認められ、poly-A 領域での体細胞 SNV 候補サイトで、QC 後深度が 1-11 万と低下したものが認められた。これら候補サイトでは、数字上は体細胞 SNV が存在していることを示す Alt 割合となっているが、シークエンスクオリティが急激に低下するサイトであり、バリデーションされたものとは考えられず、解析は困難な領域であると考えた。

このセットにおいて、バリデーションされた 12 カ所の体細胞 SNV 候補サイトは、いずれも MuTect で検出されたものであった。Strelka のみで検出された体細胞 SNV 候補サイトは 1 カ所であり、バリデーションされなかった。MuTect に限定したバリデーション率は、HC で 73.3% (11/15), LC で 14.3% (1/7)である。

3.1.5 CC_WGS_set における体細胞一塩基変異候補の確認

次に、死後皮質・小脳の比較を行った。死後皮質・小脳のセット（CC_WGS_set）は、Hiseq X を用いており、シークエンスデータが十分な質ではなかったため（図 8）、WGS データの QC とアラインメントは図 2 のパラメーターで解析した。WGS のアラインメントデータから、MuTect, Strelka で体細胞 SNV 候補を検出した。

MuTect（高感度検出モード）では、S6_cortex を解析対象試料とした解析で 29,197 個、S6_cerebellum を解析対象試料とした解析で 24,188 個の体細胞 SNV 候補が検出された。多コピー領域、INDEL 領域、多コピー疑い領域上の候補を除外し、深度 50 以上で選択したところ、それぞれ 466 候補、439 候補が選択された（表 20）。深度によるフィルタリング後も候補が多数となったため、更なるフィルタリングを設定した。最初のフィルタリングでは、S6_cortex と S6_cerebellum で Alt 割合差が 5%以上であるサイト上の候補を選択した。Alt 割合差を 5%とした根拠は、MuTect のデフォルトモード（比較対照試料で体細胞変異がないと仮定して解析を行うモード）でも検出された体細胞 SNV 候補（S6_cortex で 81 個、S6_cerebellum で 69 個。全て高感度検出モードの検出に含まれる）の Alt 割合が 5%以上であり、確度がより高いものと推定されたからである。Alt 割合差 5%以上で選択したところ、S6_cortex を解析対象試料とした解析で 257 候補、S6_cerebellum を解析対象試料とした解析で 229 候補が選択され、操作的方法及びセクション 3.1.3 で記載した SBC・シークエンス文脈にて HC, LC を選択したところ、総計で HC が 13 個、LC が 44 個となった（表 20）。

Strelka による解析では、S6_cortex を解析対象試料とした解析で 1627 個、S6_cerebellum を解析対象試料とした解析で 724 個の体細胞 SNV 候補が検出された。多コピー領域、INDEL 領域上の候補を除外したところ、それぞれ、61 候補、32 候補が選択され、操作的方法及びセクション 3.1.3 で記載した SBC・シークエンス文脈にて HC, LC を選択したところ、総計で HC 2 個、LC 1 個となった（表 20）。Strelka 解析の HC サイトは 2 カ所とも MuTect 解析の HC サイトと重複しており、Strelka 特異的な HC サイトはなかった。LC サイト 1 カ所は Strelka 特異的な検出であった。

以上の解析から選択された LC サイト計 45 カ所から、システマティックな評価を行うためシークエンス文脈ごとにランダムに 20 カ所抽出し、HC サイト 13 カ所と合わせて TAS による Alt 割合を計算した（表 21）。ランダムに抽出した LC サイト 20 カ所の内訳は、STR の切り替わりサイトが 6 カ所、poly-A 領域末端が 8 カ所、SBC が 1 個であり STR や poly-A ではない候補のサイトが 6 カ所であった。HC は 7 カ所でバリデーションされ、バリデーション率は 53.8% (7/13)であった。HC の中でバリデーションされなかった候補のうち、1 カ所（chr3: 125258109）は Alt 割合が 48.786%, 50.94%であり、dbSNP に登録されているサイト（rs34563051）であることから、生殖系列ゲノムにおける SNP であったと考えられる。別な 1 カ所（chr9: 109101822）は、Alt 割合が 0.566%, 0.625%であり、想定したエラー率 0.316%を超えていたが、Ref 以外の他の塩基も同程度の割合であり、シークエンスエラーと考えた。

LC のバリデーション率は 0% (0/20)であった。13 カ所は体細胞 SNV を示唆する Alt

割合は認めず、体細胞 SNV はないものと考えた。LC の中でも STR が切り替わる候補サイトを 6 カ所バリデーション対象としたが、5 カ所はアラインメント精度が十分ではなく、各リードに前後に INDEL が頻出するため、解析困難であった。 poly-A 領域 8 カ所中 6 カ所は、TAS による Alt 割合がエラー率以下であり、体細胞 SNV が存在しないものと考えられたが、残り 2 カ所は、数字上体細胞 SNV が存在していることを示す値となっていた。この 2 カ所は、シークエンスクオリティが急激に低下する部分となっており、解析は困難であると考えた。SBC が 1 個でシークエンス文脈が良好な LC 6 カ所のバリデーションでは、Alt 割合は全て想定エラー率以下であり、体細胞 SNV は存在しないものと考えた。

この結果を受けて、MuTect 解析において S6_cortex と S6_cerebellum の Alt 割合差が 5%未満の体細胞 SNV 候補の中から、更なるフィルタリング閾値を設定し、TAS の対象を選択した。表 21 結果では、バリデーションされた体細胞 SNV 候補を示唆するベースコールの平均ベースクオリティが 25 以上であったことから、平均ベースクオリティ 25 以上にてフィルタリングを行い、IGV 可視化の下、操作的方法と SBC・シークエンス文脈にて HC, LC を選択した（表 20）。最終候補は HC が 12 個、LC が 6 個（うち 4 個は STR の切り替わりサイトにある）となったが、前述の通り CC_WGS_set における LC のバリデーション率が 0%であったのを鑑み、本回の TAS は HC のみを対象とした。HC に対する TAS による確認結果を表 22 に示した。HC は 6 カ所でバリデーションされ、バリデーション率は 50% (6/12)であった。バリデーションできなか

った候補のうち、2 カ所（chr12: 39773148, chr18: 28039112）は理論的深度 20 万から著明に低い深度（それぞれ 33130, 1635）であった。

3.1.6 超高深度ターゲットアンプリコンシークエンス実験の再現性確認

TAS より求めた Alt 割合の再現性を確認するため、TAS でバリデーションされた脳神経組織における体細胞 SNV 31 カ所（CL_WGS_set で 6 カ所、NeuN_WGS_set で 12 カ所、CC_WGS_set で計 13 カ所）に対して、同じロットの DNA を用い、再度、独立した PCR からの TAS による追試実験を行った。2 回目の確認実験では、QC 後 18~27 万の深度が確保されており、十分な深度において表 23 に示した Alt 割合となった。1 回目と 2 回目の確認実験における体細胞 SNV (Alt)の割合は、図 11 で示したように高い相関（Pearson's r = 0.987, $p < 2.2\times10^{-16}$）を示した。

31 カ所での 1 回目と 2 回目の Alt 割合の差（TAS_1st_FA – TAS_2nd_FA）は 0.04% ± 0.60（平均±標準偏差）であった。仮に 2 つの試料に同じサイトの体細胞変異が同じ Alt 割合で存在した場合、95%は Alt 割合差が± 1.20%に収まるものと考えられる。31 カ所の体細胞変異サイトのうち、1.20%以上の Alt 割合の差を認めたサイトは 26 カ所であった。この 26 カ所については、同じ変異の誤差では説明できない水準の Alt 割合差があると言え、実際の体細胞変異割合にも差があるものと推定される。1 回目で Alt 割合が想定エラー率以下であった候補は、2 回目の Alt 割合も想定エラー率以下であった。

3.1.7 脳神経組織における体細胞変異の特徴

本研究においてバリデーションされた体細胞 SNV は 31 カ所（CL_WGS_set で 6 カ所、NeuN_WGS_set で 12 カ所、CC_WGS_set で計 13 カ所）であり、いずれも WGS データにおいて、体細胞 SNV を示唆するベースコールの平均ベースクオリティが 25 以上、マッピングクオリティが 60 以上の体細胞 SNV 候補から確認された。TAS での深度は、31 カ所全てのサイトで 10 万以上確保されていた。脳神経組織を含む 3 セット（CL_WGS_set、NeuN_WGS_set、CC_WGS_set）から選択された HC 候補（総計 46 カ所）について、候補選択に用いるベースクオリティの値を変化させて ROC 解析を行ったところ、図 12 で示す特異度と感度の関係が得られた（AUC = 0.876）。HC 候補のみであるが、ベースクオリティ 25 以上で選択すると、感度 100%、特異度 18.8%となり、ベースクオリティ 30 以上で選択すると、感度 73.3%、特異度 87.5%となった。真陽性率と偽陽性率の差が最大となったのは、ベースクオリティ 28 以上で選択した時で、感度 90%、特異度 75%、真陽性率と偽陽性率の差は 65%であった。

バリデーションされた 31 カ所において、WGS にて検出された Alt 割合と、TAS での Alt 割合を比較したところ、緩やかな相関（Pearson's r = 0.696, $p = 5.87\times10^{-12}$）を示した（図 13）。TAS によりバリデーションされた体細胞 SNV 31 個のうち 21 個（67.7%）が C>T（G>A）の transition であった。全ての一塩基置換が等しい確率で生じるとすると期待値は 16.7%であることから、C>T（G>A）の transition に著名な偏りがあると

言える（χ^2 検定 $p = 2.34\times10^{-14}$）。21 個の C>T (G>A) transition のうち、11 個（52.4%）は CG dinucleotide 上に存在していた。変異が生じるパターンに偏りがないとすると CG dinucleotide 上に存在する期待値は 25%であり、CG dinucleotide 上のシトシン塩基に変異が生じやすいと言える（χ^2 検定 $p = 0.0038$）。

SnpEff で解析したところ、脳神経組織で検出された 31 個の体細胞 SNV のうち、アミノ酸変化を伴う変異は 1 つであり、機能欠失に至る変異は認めなかった（表 24）。遺伝子がアノテーションされた変異は 17 個あり、これら 17 個を含む 17 遺伝子で ToppGene による gene ontology 解析を行ったところ、biological process, molecular function では特徴はなかったが、cellular component で神経細胞関連の gene ontology が得られた（表 25）。

3.2 高深度全ゲノムシークエンスによる一卵性双生児不一致例における体細胞一塩基変異の探索

一卵性双生児不一致例 1 組（MZ_WGS_set）における WGS データの要約を表 13 に示した。WGS には、CC_WGS_set と同じ Hiseq X（同バッチ）を用いており、CC_WGS_set のシークエンスデータが十分な質ではなかったため（図 8）、図 1 のような解析パラメーター設定を行った。MuTect と Strelka にて体細胞 SNV 候補の検出を行い、体細胞 SNV 候補のフィルタリングを行った。本解析では、SBT1（罹患者由来）を解析対象試料とした解析と、SBT4（非罹患者由来）を解析対象試料とした解析の 2 パターン

にて体細胞 SNV 候補の検出を行った。非多コピー領域、多コピー領域双方における体細胞 SNV 候補の解析を行った。

非多コピー領域における解析では、MuTect にて平均 5375.5 個、Strelka では平均 891 個の体細胞 SNV 候補が得られ、多コピー領域・INDEL 領域を除外したところ、それぞれ平均 369.5 候補、49.5 候補が選択された（表 26）。比較対照試料に存在する体細胞 SNV 候補を除外し、操作的方法、SBC・シークエンス文脈により HC, LC を選択したところ、MuTect では総計 4 個の HC、24 個の LC、Strelka では SBT1 にのみ 2 個の LC が選択された。

多コピー領域における解析では MuTect のみを用い、MuTect にて平均 2142.5 個の体細胞 SNV 候補が得られた。多コピー領域上の候補は平均 1863.5 個であり、INDEL 領域を除外したところ、平均 1203.5 候補が選択された（表 26）。比較対照試料にも存在する体細胞 SNV 候補を除外し、操作的方法・SBC、シークエンス文脈により HC, LC を選択したところ、総計 14 個の HC、31 個の LC が選択された。

同じプラットフォームでシークエンスした CC_WGS_set の解析では、LC サイトのバリデーション率がゼロであったため LC を除外し、HC サイトでのみ TAS によるバリデーションを行った（表 27）。非多コピー領域の HC サイトは 4 カ所と少なく、TAS では体細胞 SNV 候補は確認されなかった（バリデーション率 0%）。うち 2 カ所は、SBT1, SBT4 ともに TAS の深度が 1 万以下であった。多コピー領域の HC サイトは 14 カ所あったが、1 カ所（chr20: 58016089）を除き、いずれもエラー率以下の割合であ

った。chr20: 58016089 は SBT4 に特異的な体細胞 SNV 候補サイトとして WGS データから検出されたものであるが、TAS による Alt 割合が SBT1 で 1.986%、SBT4 で 2.134%であり、解析対象試料のみならず比較対照試料でも同程度検出された。多コピー領域上であることを考えると、ゲノム上の相同領域によるアーティファクトの可能性が高い。多コピー領域は、いずれも BLAT スコアが 150 未満であるものの、試験的 PCR でシングルバンドとなるものが少なく、非特異的な増幅が多い傾向であった。今回の TAS でも 5 カ所は十分な深度は得られず、他の解析と比較して著名に欠損データが多い結果であった。深度が少ない 5 カ所も、PCR プロダクト自体は他の候補と等モル存在していることから、相同配列の増幅が一定以上あると考えられる。

3.3 全エクソームシークエンスによる双生児不一致例における体細胞一塩基変異の探索

一卵性双生児不一致例 4 組（MZ_Exome_set）における WES データの要約を表 13 に示した。MuTect と Strelka にて双生児間の比較を行うことで、体細胞 SNV 候補の検出を行い、候補のフィルタリングを行った。エクソーム解析では、エクソン領域にシークエンスが限られているため、できるだけ広く候補を挙げるためフィルタリングの閾値は緩めに設定した。

MuTect では、1 試料につき平均 98.5 個の体細胞 SNV 候補が検出され、INDEL 領域を除外し、比較対照試料にも存在する候補を除外すると、平均 60 候補が選択された

（表 28）。操作的方法、SBC・シークエンス文脈により HC, LC を選択したところ、8 試料で計 28 個（平均 3.5 個）の HC、 計 8 個（平均 1 個）の LC が選択された。Strelka では、1 試料につき平均 34.75 個の体細胞 SNV 候補サイトが検出され、INDEL 領域を除外し、比較対照試料にも存在する候補を除外すると、平均 25 候補が選択された。操作的方法、SBC・シークエンス文脈により HC, LC を選択したところ、8 試料で計 22 個（平均 2.75 個）の HC、計 4 個（平均 0.5 個）の LC が選択された。

MZ_Exome_set の WES では、シークエンス用ライブラリ調整にあたり PCR 増幅を用いており、PCR エラーの問題があるため LC のバリデーション率は低いものと想定し、HC サイトのみ TAS の対象とした。MuTect による 28 カ所の HC サイトと、Strelka による 22 カ所の HC サイトは、7 カ所重複しており、最終的に合計 43 カ所の体細胞 SNV 候補サイトが TAS の対象となった。43 カ所の HC サイトの内、7 カ所で体細胞 SNV 候補が確認された（表 29）。いずれも 60 歳の妄想性障害不一致例に限局しており、他の 3 組では体細胞 SNV 候補は確認できなかった。確認された体細胞 SNV 候補サイトは、いずれも MuTect でコールされたサイトであり、変異を示唆するベースコールの平均ベースクオリティが 30 以上、候補サイトの深度が 50 以上であった。罹患者側では 3 ヶ所、*CDHR3*（cadherin-related family member 3）, *P2RY2*（purinergic receptor P2Y, G-protein coupled, 2）, *ABCC9*（ATP-binding cassette, sub-family C, member 9）の遺伝子上で体細胞 SNV が確認された。健常者側で 4 ヶ所、*ECE1* (endothelin converting enzyme 1), *BMP8A* (bone morphogenetic protein 8a), *KIF26B* (kinesin family member 26B),

NAV3 (neuron navigator 3)の遺伝子上で体細胞 SNV が確認された（表 30）。バリデーションされた 7 個の変異のうち、4 個はアミノ酸配列の変化を伴う変異であった。この解析では、閾値を緩め、できるだけ多く体細胞 SNV 候補を選択する方針で候補を挙げたため､バリデーション率が 16.3% (7/43)と低くなった｡ベースクオリティ 30 以上、深度 50 以上の基準で事後的に体細胞 SNV 候補を選択すると、計 10 個となり、この基準でのバリデーション率は 70% (7/10)である。

これら 7 カ所の体細胞 SNV 候補に対し、TAS 法による確認実験に加え、パイロシークエンシング法による独立した確認実験を行った。パイロシークエンスにより、4 カ所で体細胞 SNV の存在を確認することができ、図 14 に Alt 割合の結果及びパイログラムの一例を示した。TAS での Alt 割合が 4~5%を超えているサイトはいずれもパイロシークエンシングで確認できたが、4%以下のサイトは確認されたサイトとされなかったサイトが混在していた。

3.4 改良型 L1Hs-seq の確立

LINE1 挿入部位を検出する方法として L1Hs-seq が報告されている [35]。これは LINE1 配列の中でも唯一自律的な転移活性を持つとされている L1Hs 配列に注目し、L1Hs3’末端配列に特異的なプライマーを片側に用いることで L1Hs 及びその転移後の配列を含む領域を増幅し、次世代シークエンサーにて網羅的に解析する方法である。本研究では、体細胞変異の検出を目的として、低アレル割合で存在する L1Hs 新規挿入を検

出できるよう L1Hs-seq を改善し、より正確かつ網羅的な LINE1 配列のレトロトランスポジション検出法の開発を試みた（改良型 L1Hs-seq）（図 5）。実験は Miseq を用いて独立に 4 回行い、人工配列の検出リード数、リファレンスゲノム上の L1Hs（リファレンス L1Hs）の検出感度、非リファレンス位置での L1Hs（非リファレンス L1Hs）配列の挿入検出数を調べた。

血液由来ゲノム DNA に、L1Hs 配列＋ランダム配列となるよう人工的に設計した配列（図 4）を 0.1~10%の割合で混合し、体細胞変異をシミュレートした試料（IC1, IC2, IC3）として改良型 L1Hs-seq 解析を行った。IC1, IC2, IC3 の解析では、平均 1860 万リードペアのリードが得られ、Cutadapt にて L1Hs 配列をミスマッチ率 15%で選択したところ、平均して全リードの 96.4%が選択された。L1Hs 配列は挿入部位によって配列が多様なため、ミスマッチ率を 15%と高めに設定した。既報の L1Hs-seq ではシングルエンドのシークエンスであり、L1Hs 配列を含まない偽陽性リードを除外できていないことから、方法として特異度の上昇が得られたものと考える。

リファレンス L1Hs は、緩い閾値設定 R で 96.3~97.4%、厳しい閾値設定 S で 82.4~83.9%で検出できており、高い感度を示した（表 31）。シークエンス方法が異なるため単純な比較はできないが、L1Hs-seq 原法の感度は平均 78.4%であり [35]、改善が得られたと言える。非リファレンス L1Hs の挿入検出数は、82~142 個、本解析で新規に同定できたと考えられる挿入数（KNR L1Hs を除いた数）は 31~88 個であった（表 31）。ペアエンドシークエンスのデータにスティッチングを適用することで、アライ

ンメント精度の上昇と、非リファレンス L1Hs 挿入領域のジャンクション配列決定が可能となった（図 15)。体細胞変異をシミュレートした人工配列は、混合率 0.5~1%からアラインメントが確認できており（表 31、図 16)、アレル割合 1%程度から新規挿入が検出できると推定された。

IC1 で検出された非リファレンス L1Hs 挿入位置から、ランダムに 4 カ所選択し、Nested PCR とサンガー法による配列決定にてバリデーション実験を行った。選択した 4 カ所のジャンクションサイト（chr1: 119553351, chr5: 33797557, chr8: 75723721, chr13: 61462344）に対するサンガーシークエンス結果を図 17 に示した。4 カ所とも、リファレンスゲノム配列及び poly-A（リファレンスゲノム配列に存在しない）の配列がサンガー法にて確認され、ジャンクションサイトは一塩基レベルで確認できた。chr1: 119553351 では 6 パターン、chr5: 33797557 では 4 パターン、chr8: 75723721 では 4 パターン、chr13: 61462344 では 4 パターンの Nested PCR によるサンガーシークエンスのアラインメントが確認できた。リファレンスゲノム配列はサンガー法により明瞭に配列決定できたが、poly-A より 5’側の L1Hs 配列は、明瞭に配列決定できなかった。また、明瞭に配列決定できたサンガーシークエンスはいずれもリバースプライマー（リファレンスゲノム配列に結合するプライマー）を使用したものであり、フォワードプライマー（L1Hs 配列に結合するプライマー）では明瞭な結果が得られなかった。選択した 4 カ所であるが、解析当時は chr1: 119553351, chr8: 75723721, chr13: 61462344 が文献上報告されているサイト [22] であり、chr5: 33797557 は本実験で確認された新規

挿入であった。しかし、1000 ゲノムプロジェクトの報告[54]に chr5: 33797557 は含まれており、1000 ゲノムプロジェクトを含めた解析を行ったところ、結果として 4 ヶ所とも文献上報告されているサイトとなった。

血液試料を用いた検討で良好な結果が得られたため、次に健常者小脳試料（S6_cerebellum）を用いた検討を行った。1 回の Miseq の解析により約 1700 万リードペアが得られ、リファレンス L1Hs が閾値 R で 96.2%、閾値 S で 81.3%検出できた。非リファレンス L1Hs が 159 個、うち新規挿入箇所が 122 個検出できた（表 31）。

リード数と、リファレンス L1Hs 検出感度及び非リファレンス L1Hs 検出数の関係を調べるため、元のリードデータをダウンサンプルして解析した結果を図 18 に示した。リファレンス L1Hs への感度は 1000 万リードペア前後から横ばいになる一方で、非リファレンス L1Hs 検出数がリード数に対し単調に増加しており、これは体細胞新規挿入の検出を示唆していると考えられた。、

4. 考察

今後の精神疾患研究に向けて、WGS/WES データを用いた体細胞変異候補の同定及びその確認、LINE1 新規挿入位置を同定する改良型 L1Hs-seq の開発を行った。WGS データを用いた体細胞 SNV 解析により、健常者由来脳神経組織において 31 カ所の体細胞変異を、WES データを用いた体細胞 SNV 解析により、一卵性双生児血液において 7 カ所の体細胞変異を確認した。改良型 L1Hs-seq は、リファレンス L1Hs や非リファレンス L1Hs に加え、体細胞変異として生じたアレル割合の低い新規挿入も検出できる可能性を示した。以下に、体細胞 SNV 解析と LINE1 解析の技術的妥当性と限界点を考察し、検出した体細胞 SNV の生物学的意義を議論する。最後に、今後の精神疾患研究に対する示唆を述べる。

4.1 体細胞一塩基変異解析の技術的考察

4.1.1 技術的妥当性と今後の応用可能性

WGS/WES データは、Hiseq 2000, Hiseq 2500, Hiseq X という異なる大規模並列シークエンサーを用いている。TAS と改良型 L1Hs-Seq に用いた Miseq を加えると、シークエンスクオリティは、Miseq > Hiseq 2500 (Rapid Run mode) > Hiseq 2000 (HighThroughput mode) > Hiseq X の順となっており（製造元 Illumina 社より情報提供）、それぞれのシークエンサーのリードクオリティに応じた解析の閾値設定が必要であった。また、死後皮質・肝臓のセット、神経細胞・非神経細胞・肝臓のセットはシー

クエンス用のライブラリ調整の際に PCR を使用していないため、PCR エラーによる体細胞変異の偽陽性はないものと想定できるが、死後皮質・小脳のセットや一卵性双生児の比較では、PCR を含むライブラリ調整を行っているため、PCR エラーによる体細胞変異の偽陽性検出に対する慎重な防止対策が必要であった。そのため本解析では、2 回の Picard Deduplication を行った。また、Hiseq 2500 Rapid Run mode や Miseq には、キャリーオーバーと呼ばれる前ランからの DNA コンタミネーションが 0.1-0.5%程度生じるとされているため、キャリーオーバーによる誤解析を防ぐためシークエンスライブラリのサンプル毎にインデクシングを行い、厳密な体細胞変異の解析を行った。

TAS による確認実験は、後述する定量性の限界点があるものの、以下の 4 点から体細胞確認実験としての妥当性が支持されるものと考えた。(1) 体細胞シミュレーション試料にて理論的 Alt 割合と実験的 Alt 割合が高い相関を示した（Pearson’s r = 0.969）。(2) 2 回の独立した再確認実験で Alt 割合が高い相関を示した（Pearson's r = 0.987）。(3) 実験上のコントロールとして計算した dbSNP における Alt アレル割合が、理論値と実験値でよく一致しており、小さな標準偏差であった。(4) 一部の候補はパイロシークエンスでも体細胞変異の存在を確認した。TAS によるバリデーション実験に用いる DNA は 5~10 ng と少量であり、死後脳試料など、限られた試料に使用する際には利点があると考えられる。本実験では、体細胞変異候補の操作的フィルタリングで感度を優先し、ベースクオリティ 21 以上という緩めの閾値で候補を選択した。HC 候補における ROC 解析からは、感度を重視する場合はベースクオリティ 25 以上、特異度を重

視する場合はベースクオリティ 30 以上、真陽性率と偽陽性率の差を最大化するにはベースクオリティ 28 以上を候補として選択することが今後の解析に有用であろうと考えられる。

TAS 実験のポジティブコントロールとして使用したのべ 23 カ所の理論値 100%のホモ dbSNP サイトにおける名目上のエラー率は 0.028%であったが、ベースクオリティの値を考慮して Alt 以外のベースコールの割合を計算すると、23 カ所で平均 0.046% (4.6×10^{-4})であった。ベースクオリティの値が完全に信頼できるものと仮定すると、これは PCR エラーの割合に相当すると考えられる。PCR エラー割合を 0.046%とすると、Alt 割合のバリデーション基準に用いた想定シークエンスエラー率 0.316%より一桁低い値であり、PCR エラーによる偽陽性の可能性は高くはないと考える。シークエンスライブラリ調整に使用した NEB Q5 DNA polymerase は複製時のエラー率は 1.0×10^{-6} とされており（NEB 資料より）、計 31 回の PCR を用いた本実験の一塩基あたりの最終的なエラー割合の期待値は 3.1×10^{-5} であった。実際のエラー割合は、期待値より一桁高い値であり、これはアニール温度などの PCR 条件が Q5 DNA polymerase に最適なものではないという可能性が一因として考えられる。また、Phred スケールのベースクオリティ値が完全には正確ではなく、このような計算に使用できる水準の精度ではないという可能性なども考えられるであろう。

解析の結果、Strelka でのみ検出された体細胞変異候補は、TAS でバリデーションされなかった。Strelka で検出された体細胞変異でかつバリデーションされたものは、い

ずれも MuTect により検出されており、本解析では MuTect のみでの解析としても結果は変わらなかったと言える。MuTect が感度・特異度において最も信頼性が高いとしている報告を支持する結果であった[40]。MuTect の開発グループによると、比較対照組織に存在しない体細胞変異を検出する際、深度 80 では、割合 10%の体細胞変異が感度 99%で検出され、割合 3%の体細胞変異を感度 99%で検出するには、深度 330 が必要になるとのことである[40]。本実験では脳神経組織においては、クオリティコントロール後の最終深度 74~85 の WGS を行ったが、シークエンスコストの低下とともに、更なる深度での実験・解析が可能となり、同様の方法にてより低い割合の体細胞変異を網羅的に検出できるようになるであろう。一方で、必ずしも比較対照となる組織が得られるとは限らないため、比較組織がない状態での体細胞変異の検出は今後の課題である。

本論文を準備中に、神経細胞の単一細胞ゲノム解析による SNV の網羅的なプロファイリングが報告された[15]。3 個体から計 36 個の単一神経細胞を準備し、単一神経細胞ゲノム DNA から φ29DNA ポリメラーゼにて全ゲノム増幅し、WGS データと MuTect にて体細胞変異を検出している。一神経細胞あたり 1500 程度の SNV があると報告しており、科学的知見として意義が高いものと言える。しかし、単一細胞解析において、φ29DNA ポリメラーゼのエラー率は 10^{-7} とされており、Taq Polymerase の約 1,000 分の 1 と少ないものの、全ゲノム増幅中の初期のエラーは実際の体細胞変異と区別がつきにくく、また全ゲノム増幅中のコンタミネーションリスクが非常に高いという技術

的難点やシークエンスコストが莫大になるという経済的難点もある。本研究では、単一細胞ではなくバルクの組織を用いた解析であるが、単一細胞解析と比較して、以下の利点がある。(1) ライブラリ調整で PCR を介さず行う場合は、PCR によるエラー（偽陽性）がない、(2) 1 リードあたり 1 アレル由来と考えると、リードデータに無駄がなく、対象部位での体細胞変異の割合を評価しやすい、(3) 単一細胞を分離するための高価な機器の必要性や技術的難点が少なく汎用性が高い。一細胞に集積している変異の組み合わせを解析する目的でなければ、各遺伝子の体細胞変異が当該組織でどの程度の割合で存在しているかというデータの方が、疾患をはじめとしたフェノタイプとの関連への理解には重要と考えられ、疾患研究やフェノタイプとの相関を解析する場合には、本研究の方法が有用と言えるであろう。

本研究の SNV 解析では、ヒト由来試料に対する適用を主眼として解析を行ってきたが、モデル動物で行う体細胞ゲノム編集の評価方法としても想定できる。脳神経組織に部位特異的なゲノム編集を行う技術が普及しつつあるが [55]、ゲノム編集による人工的な体細胞変異の割合を TAS で定量化することで、ゲノム編集の程度と生化学的性質や動物の行動特徴との関連を考えるといった実験が可能になる。より精密なモデル動物での実験を行うための方法を提供できるであろう。

4.1.2 技術的限界点

体細胞シミュレーション試料における Alt 割合の理論値と実験値は高い相関を示し

たが、実験的 Alt 割合は理論値と完全には一致しておらず、残差の標準誤差は 1.27 (%)であった。実験的 Alt 割合は、理論値に対し傾き 1.39 となったが、これは試料 JM1, JM2 を混合する際に理論値より多めに JM2 を混合していたという手技的な問題に起因するものと考えられる。実験値の分散については、PCR 増幅の際に、Ref アレルと Alt アレルにターゲットサイトごとの増幅バイアスがかかったために生じたことが主因であろう。このキャリブレーション実験に使用した DNA は 7 ng（約 1000 細胞に相当）であり、1%に相当するアレルが常染色体で 20 個であった。本研究では、脳神経組織試料の DNA が限られていたため、TAS を 5-10ng の DNA から開始する必要があり、ごく少量の DNA からの評価を行った。少量の DNA から開始したことにより、PCR 増幅バイアスが大きくなった可能性があり、TAS に使用する DNA を増量することが対策になるであろう。1.27 (%)の標準誤差から、本研究における TAS の Alt 割合の数字が示す定量性については 95%信頼区間が± 2.49%であると言える。より高い感度の解析や低割合に存在する体細胞変異のより確実な解析には、使用 DNA 量の増量やデジタル PCR が有用であると考える。

TAS 再現実験における 1 回目と 2 回目の Alt 割合の差の標準偏差は 0.60%である一方、キャリブレーション実験における標準誤差は 1.27%であった。異なる解析のため単純な比較はできないが、後者での分散が大きくなったのは、TAS 再現実験においては同じサイトの同じ Alt 割合の変異を再実験している一方、キャリブレーション実験では比較したサイトが 7 種で異なっており、かつ 2 種類の DNA（JM1、JM2）の非生

理的なアレルの組み合わせであることから、PCR バイアスが大きくなったものと考えられる。JM1 と JM2 という 2 種類の DNA の混合過程において、染色体・ゲノム領域ごとに混合割合のバラつきが生じ、増幅バイアスが増大した可能性が高い。生理的な試料における Alt 割合の計測は、このキャリブレーション実験よりは標準誤差が小さくなるものと考えられるが、数%の範囲での定量性には課題が残る。

一卵性双生児における体細胞変異 7 ヶ所のうち、TAS による Alt 割合が 4~5%以上の候補はパイロシークエンスによっても変異の存在が確認されたが、4%以下の候補は確認結果にばらつきがあった。また、TAS による Alt 割合とパイロシークエンスによる Alt 割合は完全に一致しているとは言えない。本実験のみでは、この結果が TAS の定量性の問題によるものか、パイロシークエンスの感度・定量性不足によるものかは、断定することはできない。Alt 割合の小さな体細胞変異候補に更なるバリデーション実験を行う場合は、デジタル PCR などのより感度の高い手法が必要になるであろう。

複数の手法により確認された体細胞変異サイトに関しても、リファレンスゲノムや現在のデータベースでは十分にカバーされていない相同配列が他のゲノム領域にある可能性があり、相同配列のミスアラインメントによるアーティファクトである可能性は完全には否定できない。また、Hiseq と Miseq はともに同じ Illumina 社のシークエンサーであるため、同じ傾向のシステマティックなシークエンスエラーを持つ可能性がある。生殖系列ゲノムに存在する相同配列によるエラーやシークエンサー由来の

システマティックなエラーであれば試料間で大きな Alt 割合差はつかないと仮定すると、組織間で一定以上の Alt 割合差を認めた体細胞変異 26 ヶ所はより確度の高いものと考えられる。しかし、この議論は仮定に基づいており、相同配列の可能性は完全には除外できず、パイロシークエンスで確認された変異以外はシステム上のエラーである可能性も完全には否定できない。相同配列に対しより厳密な解析を行うには、シークエンスライブラリ調整における PCR のアニール温度を上げることがひとつの対策になるであろう。シークエンサーに特徴的なシステマティックなエラーを除外するには、パイロシークエンスやデジタル PCR など他の原理による配列決定を行うことが対策になるであろう。

本解析において HC として選択された候補には、体細胞変異の存在が確認されなかった候補が存在した。これは、操作的フィルタリングの数値設定において感度を優先し、BQ をはじめとした閾値を比較的緩めに設定したためと考えられる。体細胞変異候補となりながら TAS によりバリデーションされなかったサイトの解釈としては、1) WGS でのシークエンスエラー、2) ライブラリ調整時の PCR エラー、3) サンプル調整時のコンタミネーション、4) 他のゲノム領域に存在する相同配列のミスアラインメントが考えられる。1)のシークエンスエラーである可能性は、シークエンスエラーである確率を棄却するプロセスを取る MuTect の解析原理から考えにくいが、シークエンスデータに示された Phred スケールのベースクオリティが完全に信頼できるものではないとすると否定できない。2)に関しては、ライブラリ調整に PCR を伴うセッ

ト（CC_WGS_set）では特に LC の体細胞変異候補が多く、これらの LC は TAS によりバリデーションされなかったことから、PCR エラーは主因の 1 つと考えられる。ゲノム上の特徴を見ると、STR 領域周辺や poly-A 領域での体細胞変異候補同定が PCR フリーのサンプルよりも多く見られ、poly-A 領域に関しては伸長している傾向が認められた。3)の可能性として、これまでの大規模並列シークエンサーを用いたゲノム研究において、試料採取時やシークエンスライブラリ作成時のサンプルコンタミネーションが体細胞変異の偽陽性にもなりうると報告とされており [56,57]、サンプルコンタミネーションも主因の 1 つとして考えられる。4)の相同配列の可能性については、セクション 3.2 の結果に顕著であるが、確認されなかったサイトの一部は、適切なフラグメント長の特異的なプライマーセットのデザインや、PCR によるシングルバンドを得るのが困難なサイトが多かった。TAS で確認された体細胞変異候補は、全て適切なフラグメント長の PCR プライマーが速やかにデザインでき、ほとんどの場合でシングルバンドが得られた。また、確認されなかったサイトの一部は、深度 10 万以下の低深度のサイトであった。DNA 量としては他のサイトと等モルでシークエンスしていることから、他の領域の増幅が多く含まれていた可能性がある。これらは、確認されなかったサイトの一部にはゲノムの別領域に相同配列が存在する可能性を示唆しており、WGS データ解析の時点で、相同配列が誤ってアラインメントされた可能性が考えられる。

特に、レトロトランスポゾンなどコピー数の多い配列は、挿入位置や配列の個体差

が大きいため、リファレンスにおける位置・配列が必ずしも信頼できるものではない。生殖系列ゲノムでも特異的なアラインメントが難しく[43]、アレル割合の少ない体細胞変異の解析はより困難であると考えられる。別個体の血液細胞ゲノム DNA での試験的 PCR の結果に対する判断（シングルバンドないしシングルバンドに近いものが得られたという判断）が、そのまま実試料に適用できるわけではない可能性が他のゲノム領域より高いであろう。本研究では、試料が十分になく、試験的 PCR を他個体のゲノム DNA にて代替的に行ったが、試料が十分にある場合は、実試料にて試験的 PCR を行うことが、このような領域の解析には望ましいと考える。また STR 周辺領域や poly-A 末端は TAS による評価が困難であった。これらの領域はデジタル PCR やサンガー法、パイロシークエンスなど他の手法でも評価が難しく、現在の技術で正確に検出するのは困難であり、今後の技術的課題であろう。

本解析では、MuTect の高感度検出モードで比較対照組織での許容 Alt 割合を 15%としたが、このパラメーターでは比較対照組織に 15%を超えて存在する体細胞 SNV は検出できない。発生の最初期に生じた体細胞変異は、比較対照組織にも 15%を超える割合で存在する可能性もあり、このような体細胞 SNV をするために、比較対照組織の許容 Alt 割合を高くし、感度を更に高めることは可能であろう。体細胞変異候補として検出した中で、生殖系列ゲノムの SNP と判明した偽陽性は、全解析を通じて 1 カ所のみであったことから、本解析の許容 Alt 割合が高すぎるということはなく、今後の解析で更に高い数値とすることは可能と考える。

4.2 レトロトランスポゾン解析の考察

4.2.1 解析の妥当性と今後の応用可能性

Bundo らは、統合失調症患者由来死後脳試料において、神経細胞ゲノムに特徴的な LINE1 コピー数増大が見られることを報告した。この報告では、患者由来 iPS 細胞から分化させた神経前駆細胞や、polyI:C による統合失調症様動物モデルの神経細胞においても同様のコピー数増大を認めた[33]。統合失調症と LINE1 レトロトランスポジションの関連が疑われるが、LINE1 新規挿入が生じているゲノム領域の特定が課題であった。

本研究で開発した改良型 L1Hs-seq は、感度良く一塩基レベルの解像度で LINE1 挿入位置を検出できる系であり、コピー数増大を認めた患者試料に適用することで、LINE1 新規挿入位置を決定することができると期待できる。体細胞における新規挿入は、アレル割合にして 1%前後から検出可能であるため、組織試料に対して適用する場合は、発生初期の割合が比較的高い新規挿入の検出が現実的な目標になる。本解析では 2 種類の閾値で解析を行ったが、同一個体の組織間で比較することで体細胞変異を示す場合、対象組織での LINE1 新規挿入検出に閾値 S を用い（ポジティブフィルタリング)、対照組織に同位置の新規挿入が存在しないことを示すのに閾値 R を用いる（ネガティブフィルタリング）ことが有用となる可能性がある。

発生・発達が進んだ段階で生じた、1%以下の割合と想定される LINE1 新規挿入を

検出する場合は、リード量を増加させていくのも一つの方法であるが、単一細胞ゲノムを全ゲノム増幅することで、一細胞ごとに解析する方法が望ましいと考えられる。一細胞内に存在する新規挿入はアレル割合にして理論的には 50%であるが、全ゲノム増幅をした場合、ゲノム領域によって増幅効率が異なるため領域ごとに相対的な量にばらつきが生じる。改良型 L1Hs-seq は、アレル割合にして 1%に相当する程度のフラグメントから検出可能なため、全ゲノム増幅にて増幅が不良であった領域の検出にも優れていると考えられ、より安定的な LINE1 配列の検出ができると期待できる。今後は、患者由来試料での一細胞ゲノム解析と組み合わせることで、脳神経組織における LINE1 の新規挿入位置と疾患の関連を解析することができるであろう。

4.2.2 解析の限界点

本実験により新規に同定したと考えられる非リファレンス L1Hs（文献上の報告のないもの）には、一部で poly-A の長さに多様性があった（図 15）。新規に挿入された L1Hs は有糸分裂により poly-A の短縮が生じやすく、進化的に古い LINE1 には生じにくいという報告がある [58]。Poly-A の長さを正しく配列決定できていると仮定すると、これは新規挿入を支持する特徴である。しかし、Miseq による poly-A 配列決定は十分な精度とは言えず、シークエンサーによるアーティファクトの可能性も考えられる。

改良型 L1Hs-seq で検出された非リファレンス L1Hs は、4 ヶ所について挿入位置のバリデーションを行った。いずれも挿入位置のゲノム配列はサンガー法により決定で

きたが、poly-A より 5’側の L1Hs 配列は明瞭に決定できず、poly-A の長さにも多様性を認めた。サンガー法では poly-A 領域の配列決定が難しく、L1Hs 部分の配列決定は十分ではないと言える。また、L1Hs 側に結合するプライマーによるサンガー法では、配列決定が困難であった。リファレンスゲノム hg19 上には、2 つの LINE1 配列 3’末端が互いに向き合って存在する領域があり、例えば chr8: 92534081 - 92534363 には L1Hs 配列 3’末端が互いに向き合って存在する。L1Hs 側に結合するプライマー1 種のみで、このような領域の増幅が生じ、Nested PCR でも除外することができない。Nested PCR プロダクト中に、LINE1 配列 3’末端が互いに向き合った領域の増幅があると考えられ、L1Hs 側に結合するプライマーによるサンガー法が困難になったものと考えられる。

また、バリデーションした 4 ヶ所は、結果的にいずれも文献に報告されている L1Hs であった。解析時点では、1 ヶ所が本実験で新規に同定した挿入位置であったが、新規に同定した位置については別途評価する必要があるだろう。バリデーションも 4 ヶ所と少数であるため、非リファレンス L1Hs をシステマティックに評価するためには、より多数の挿入位置についてバリデーションを行う必要があるだろう。

4.3 体細胞一塩基変異の生物学的意味

セクション 4.1.2 で述べた技術的限界点があるものの、TAS で確認された体細胞変異候補を実際の体細胞変異と考えると、アレル割合（Alt 割合）は体細胞変異が生じ

た発生段階と相関するものと考えられる。例えば、サイト A で 13%、サイト B で 5%、サイト C で 1%の体細胞変異が確認された場合、発生過程において A, B, C の順で変異が生じたものと想定される。

完全に均等な細胞分裂により全ての体細胞が生じると仮定すると、アレル割合の期待値は、受精卵で生じた変異が 50%、2 細胞期に生じた変異が 25%、4 細胞期に生じた変異が 12.5%というように、細胞分裂に応じてアレル割合の期待値が半減していくことになる。脳神経組織における体細胞変異のアレル割合は 0.7~14%に分布していたが、この仮定に基づくと、アレル割合 14%の変異は 4 細胞期前後、アレル割合 0.7%の変異は細胞数 70 前後の時期に生じた変異と考えられる。しかし実際は、発生過程で完全に均等な細胞分裂が生じているわけではなく、組織分化が進んだ後もクローナルな増殖が存在する。本研究でも、脳または脳部位に特徴的と考えた体細胞変異には、アレル割合が 1%を超えているものが存在した。低割合に存在している変異に関しては、検出における技術的限界があるとともに、アレル割合からの発生時期推定は困難と言える。アレル割合は、傾向として発生時期を反映しているものと想定されるが、発生時期の詳細な推定を行うには、様々な組織における体細胞変異の解析と比較が必要になるであろう。

4.3.1 脳神経組織における体細胞一塩基変異の生物学的意味

低割合に存在する体細胞変異の検出には技術的限界点があるものの、本研究により

脳または脳部位に特徴的と考えられる体細胞変異を検出した。また、脳部位や細胞種ごとに体細胞変異のサイト・アレル割合が異なることが示された。前頭葉と肝臓の比較にて前頭葉に特異的に認めた体細胞変異は、外胚葉と内胚葉の分化後に生じた体細胞変異である可能性が高い。脳神経組織はクローナルな増殖が少ないが、神経幹細胞により神経新生が行われており、体細胞変異の中には、その領域における神経幹細胞のクローナルな増殖により生じたものも含まれるであろう。前頭葉・小脳の比較では、各部位で特徴的に検出された体細胞変異が 1 カ所ずつあり、いずれも脳の発達過程で、部位が分かれた後に生じたものと考えられる。神経細胞・非神経細胞の比較では、片方にのみ特異的な体細胞変異は認めなかったが、深度を高めることで、特異的な体細胞変異が検出される可能性はある。

Lim らは、皮質形成異常を伴うてんかんにおいて、罹患脳部位での *MTOR* 体細胞変異アレル割合が、1.3~12.6%であったと報告している [27]。*MTOR* 変異とてんかんとの因果関係もモデル動物にて示しており、てんかんにおいてはこのようなアレル割合の体細胞変異が原因になると考えられる。本研究で検出された脳神経組織体細胞変異のアレル割合は 0.7~14%であり、Lim らの報告におけるアレル割合と同様の範囲であった。てんかんで成り立つ事象がそのまま統合失調症などの精神疾患に適用できるとは限らないが、このようなアレル割合の体細胞変異が精神疾患患者死後脳で検出された場合、体細胞変異が精神疾患に寄与している可能性を考えることができるであろう。

体細胞変異は有糸分裂の際に生じることが多いとされているが、神経細胞では他の

メカニズムとして神経活動によりゲノム DNA に double strand break が生じ、修復の際に変異が導入される可能性が指摘されている [59,60]。神経活動依存的な double strand break は、海馬歯状回に特に多く、神経変性疾患との関連も示唆されている [59]。Lodato らは、単一神経細胞における体細胞 SNV 解析により、神経細胞では C>T transition が極めて高頻度（80~90%）であり、メチル化されているシトシン塩基上に多いことから、DNA メチル化が体細胞変異に関連している可能性を提起している [15]。メチル化シトシンは、ハイドロキシメチル化シトシンを経由する能動的脱メチル化過程において塩基除去修復エラーといった変異の機会があり、また AID/APOBEC による脱アミノ反応にてチミン塩基になるという変異の機会がある [61]。ハイドロキシメチルシトシンは、他の体細胞に比べ脳神経組織に多く検出されていることから [62]、脱メチル化過程にて脳神経組織に特徴的な体細胞変異が導入される可能性は十分に考えられる。また、脱メチル化は発生初期にも高頻度に生じており [63]、発生初期に生じる体細胞変異の主要なメカニズムのひとつとしても想定できるであろう [64]。本研究でも、脳神経組織から検出された 31 個の変異の方向として C>T に大きな偏りがあり、CG dinucleotide 上に多かったことは、メチル化シトシンとの関連可能性に支持的な結果であった。単一神経細胞における SNV の報告は、より高頻度の C>T transition を報告しているが [15]、本研究で検出した体細胞変異は発生初期に生じたものが主である一方、単一神経細胞解析で検出した体細胞変異は、単一神経細胞のみで生じた一細胞特異的な変異が多数含まれており、変異のメカニズムのレパートリーが異なるためと考えられる。

脳神経組織において体細胞 SNV が生じた遺伝子群の gene ontology 解析では、神経細胞関連の ontology が多く検出され、神経細胞の変異は神経関連遺伝子に多いという Lodato らの報告[15]と一致した。Lodato らは、神経細胞の転写活動に伴う変異を想定しているが、上記のような神経活動依存的な double strand break や脱メチル化過程における変異の導入といったメカニズムが想定される。脳神経組織では神経関連遺伝子に体細胞変異が生じやすいとすると、精神疾患の発症メカニズムとしての体細胞変異の可能性について示唆的な結果である。本研究で検出した健常者脳神経組織における体細胞変異は、アミノ酸変化を伴う変異も 1 つのみと少なく、機能欠失に至る変異はなかった。これは、精神疾患罹患のない提供者由来試料に対する解析であることと矛盾しない結果であった。

4.3.2 一卵性双生児における体細胞一塩基変異の生物学的意味

本研究で検出された一卵性双生児血液ゲノムにおける体細胞変異に対しては、2 つの解釈がありえる。1 つは、発生の初期で生じたものであり、中胚葉由来の血液細胞と外胚葉由来の脳神経組織が共通して同じ体細胞変異を持っているという解釈である。発生初期の体細胞変異はフェノタイプへの寄与が大きいと予想され、理論的にも血液細胞 DNA から検出可能と考えられる。実際に Rett 症候群をはじめ、血液細胞 DNA から脳神経系に関与する可能性の高い遺伝子の体細胞変異を検出した報告がある[28,29,65]。Jamuar らは、二重皮質症候群などの脳奇形疾患において、血液試料中から

既に強い関連が報告されている遺伝子の体細胞変異をアレル割合5~35%で検出した[28]。強い関連が報告されている遺伝子の変異であることから、同じ体細胞変異が脳神経組織にも存在し、疾患の原因となっている可能性を示唆している。

本研究において、妄想性障害罹患者の血液試料から検出した *ABCC9* 上の体細胞変異のアレル割合は 7.3%であったが、Jamuar らの報告のアレル割合の範囲に収まっており、また脳神経組織の解析において脳・肝臓にともに存在した体細胞変異のアレル割合の範囲（0.5~11.3%）にも収まっていた。後述する血液幹細胞のクローナルな増殖の可能性は否定できないものの、脳神経組織にも同じ体細胞変異が存在する可能性は十分考えられる。本研究では、妄想性障害に関して不一致の一卵性双生児ペアにおいて、罹患者側のみに *CDHR3, P2RY2, ABCC9* の体細胞変異が検出された。*CDHR3, P2RY2* は精神疾患との関連の報告はないが、*ABCC9* は睡眠障害との関連が GWAS にて報告されている[66]。*ABCC9* では、体細胞変異の割合が比較的高いこと、パイロシーケンスでも確認されたこと、アミノ酸配列の変化を伴うミスセンス変異であることから、疾患との関連を検討することは可能であろう。脳神経組織の解析は死後にのみ可能となるため、生前に試料提供者に役立つことはできないが、血液細胞ゲノム情報から脳神経組織の体細胞変異が推測できれば、生前に情報として役立てる可能性が出てくる。

2 つめの解釈は、血液の幹細胞で生じた体細胞変異がクローナルに増殖した結果を検出したという解釈である。本研究では、5 組の双生児のうち 60 歳という最高齢のペ

アでのみ体細胞変異が検出された。体細胞変異が罹患者、健常者ともに同程度認められたことはこちらの解釈を支持する。加齢とともに、特定の血液幹細胞のクローナルな増殖を認め、アレル割合 10%前後の体細胞変異が検出されやすくなるという報告[67]も、こちらの解釈に支持的である。血液幹細胞のクローナルな増殖かどうかを調べるには、血液以外の組織で同様の体細胞変異が存在するかどうかを調べることが必要になる。

4.3.3 生物学的考察にあたっての限界点

体細胞変異のアレル割合と発生時期との関連を考察したが、実際の発生時期特定のためには、アレル割合だけではなく、様々な組織試料の比較が必要である。本研究では、脳と肝臓以外の試料は用いておらず、発生過程の詳細な解析とはなっていない。体細胞変異の発生時期の詳細を明らかにするには、様々な組織試料を用いたシステマティックな体細胞変異解析が必要になるであろう。

脳神経組織の体細胞変異の特徴から、脳神経組織に特徴的な変異のメカニズムやシトシンメチル化との関連、神経細胞関連の遺伝子群で脳神経組織の体細胞変異が認められやすいことを、本実験の結果と文献から議論した。いずれも実験結果と文献の報告は一致していたが、本実験で検出された脳神経組織体細胞変異は 31 カ所と限られた数であり、確定的に結論を出すには不十分である。確定的な結論を得るには、より多数例の試料から網羅的に脳神経組織体細胞変異を検出する必要がある。試料数を増

やすとともに、より高い深度での WGS/WES を行うことが望ましい。また、本研究で使用した脳神経組織試料は健常者由来であり、精神疾患との関連を直接示せるものではない。

一卵性双生児における体細胞一塩基変異は、血液幹細胞のクローナルな増殖の可能性をまず検討する必要があり、血液以外の組織における同じ体細胞変異の有無の検討を要する。仮に、血液幹細胞のクローナルな増殖ではなく、脳神経組織にも存在する発生初期の体細胞変異であるとしても、本研究では一例のみで疾患と体細胞変異の関連を考察しており、確定的な関連ではない。妄想性障害と *ABCC9* との関連を検討するには、サンプル数を増やした解析を行うことが必要になるであろう。

4.4 精神疾患研究への示唆

本研究の脳神経試料提供者は、いずれも精神疾患の既往歴がなく、フェノタイプと体細胞変異との関連は現時点では考察できないものの、一卵性双生児不一致例に関しては、疾患保有者に特異的な体細胞変異を検出することができた。現時点では、精神疾患との関連は明確ではないが、脳神経組織試料を含め、精神疾患患者由来試料に対して同様の解析を大規模に行っていくことが、精神疾患との関連解明に有用であると考える。同時に、脳神経組織における体細胞変異のメカニズムや発症との因果関係を探求していくことも必要である。

統合失調症や自閉症をはじめとした精神疾患は、神経発達障害（Neurodevelopmental disorder）として理解されることが多い[68]。序文で触れたように、脳奇形を伴う疾患で体細胞変異と疾患の関連が報告されており[25-27]、統合失調症において神経細胞ゲノムにおける LINE1 コピー数の増大が報告されている[33]。精神疾患においても発生・発達過程における体細胞変異が発症に寄与している可能性は十分に考えられる。生殖系列ゲノムの大規模な解析や多発家系解析により、遺伝的要因の実体が更に解明されていくものと考えられるが、体細胞変異の解析を並行して進めることで、疾患発症のメカニズムの詳細がより明らかになるものと期待される。

前頭葉・小脳や神経細胞・非神経細胞の比較で示したように、脳部位や細胞種ごとに体細胞変異のサイト・割合が異なることがある。脳部位や細胞種ごとに精神・神経機能への寄与が異なることから、同じ遺伝子の体細胞変異としても、生じる脳部位や細胞種や割合ごとにフェノタイプが異なると想定される。これは、精神・神経疾患の多様性に対する説明（仮説）の一つとして考えられる。ゲノム情報という一次元の情報に対して、時間（発生）・空間（脳部位）的に異なる体細胞変異という別次元の要素が加わることで、ヒトの精神・神経的特徴（フェノタイプ）の多様性に寄与することが想定できる。

以上を踏まえ、体細胞変異により精神疾患の発症を説明するモデルを 2 つ提示する（図 19）。1 つ目は、生殖系列ゲノムの変異に加え、各発達段階において複数の体細胞変異が加わることで精神疾患の発症を説明する「多段階変異モデル」である。家族

歴など遺伝負因があるケースにおいて、複数の体細胞変異が発生・発達段階で重なることにより polygenic（多遺伝子的）に発症が説明されるとともに、生殖系列ゲノムの変異と体細胞変異の時間・空間的多様性によりフェノタイプの多様性・個体ごとの違いを説明する一般モデルである。2 つ目は、特殊な形として稀な体細胞変異を主因と想定した「体細胞変異モデル」である。胎生致死や重篤な先天性疾患をもたらす強力な変異は、生殖系列ゲノム上では安定的に存在できないが、一部の組織において一定のアレル割合以下であれば、体細胞変異として存在できる可能性がある。そのような稀ではあるが効果の大きい変異が脳神経組織に存在し、精神・神経機能の障害をもたらす可能性を想定したモデルである。

本研究は健常者由来死後脳試料における体細胞 SNV の存在を示したが、いわゆる「健常者」にも体細胞変異が存在することは、精神疾患という概念の捉え方の問題も提起する。DSM-IV から DSM-5 へと、精神疾患を、カテゴリカルに断続的に分類することから、スペクトラムとして連続的に捉える方向に変わっていく傾向にある。臨床の現場でも、カテゴリカルな概念では捉えきれない個体ごとのフェノタイプの多様性や個性があることは言うまでもない。その分子的・生物学的背景として、生殖系列ゲノムの多様性・個性に加え、体細胞変異の時間的・空間的な多様性・個性も想定できるであろう。解剖学的特徴を伴う脳奇形やてんかんの患者において体細胞変異が存在し、かつ疾患との関連が強く考えられる一方で、いわゆる「健常者」において別な遺伝子での体細胞変異が存在することは、「健常」と「精神疾患」の区別も、以前に

考えられていたほど明瞭ではなく、どこに体細胞変異が生じるかという確率的な要素がある可能性を示唆している。仮に脳神経組織の体細胞変異が精神疾患に寄与しており、それが確率的に生じる交通事故のようなものであるとすると、体細胞変異は、いわば「遺伝学的交通事故 (genetic car accident)」と喩えることができ、精神疾患・神経疾患の罹患という出来事を倫理的にどう捉えるかという視点の問題を提起し、その社会的な処遇を検討するための知見ともなるであろう。

5. 結論

本研究では、WGS/WES データを用い、ヒト脳神経組織由来のゲノム DNA、一卵性双生児血液由来ゲノム DNA における体細胞変異の探索を行った。脳神経組織に特徴的な体細胞変異や、脳部位ごとの体細胞変異割合の差異、神経細胞・非神経細胞での体細胞変異割合の差異があることが、複数のゲノム領域で確認できた。体細胞変異割合の違いは、体細胞変異が生じた発生・発達段階の違いを反映しているものと考えられる。一卵性双生児間で異なる体細胞変異が検出でき、妄想性障害に関する不一致との関連を考察した。また、疾患試料適用に向けて、体細胞における LINE1 新規挿入位置を検出する方法の開発を行った。

脳神経組織における体細胞変異の存在は、精神・神経疾患の発症や、疾患を含めた個体の行動・認知の特性・個性を説明するメカニズムのひとつとして想定できる。今後は精神疾患患者における脳組織での体細胞変異のパターンと疾患との関連、脳神経組織における体細胞変異のメカニズムを明らかにしていく必要がある。

謝辞

本研究を学位論文の形にまとめるにあたり、多くのご支援とご指導を賜りました。指導教員の笠井清登先生をはじめ、岩本和也先生（東京大学大学院医学系研究科・分子精神医学講座)、文東美紀先生(東京大学大学院医学系研究科･分子精神医学講座)、加藤忠史先生（理化学研究所・脳科学総合研究センター）に丁寧なご指導を賜り、実験･解析･データ解釈のみならず、医学･科学に対する姿勢を学ぶことができました。今後に学びを活かしていきたいと思います。上田順子様（理化学研究所・脳科学総合研究センター）には、解析の初歩から親身なご指導を賜りました。

村山繁雄先生（東京都健康長寿医療センター研究所・神経病理学研究部)、橋本恵理先生（札幌医科大学・神経精神医学教室)、鵜飼渉先生（札幌医科大学・神経精神医学教室)、石井貴男先生（札幌医科大学・神経精神医学教室)、スタンレー財団（アメリカ合衆国)、澤田知世先生（理化学研究所・脳科学総合研究センター)、佐々木司先生（東京大学大学院教育学研究科)、垣内千尋先生（東京大学医学部・精神医学教室)、西村文親先生(東京大学医学部･精神医学教室)、吉川茜先生(東京大学医学部･精神医学教室）から貴重な試料をご提供頂いたことで、本研究を行うことができました。

安田純先生（東北大学・東北メディカルメガバンク機構)、長﨑正朗先生（東北大学・東北メディカルメガバンク機構)、勝岡史城先生（東北大学・東北メディカルメガバンク機構)、佐藤行人先生（東北大学・東北メディカルメガバンク機構)、黒木陽

子先生（国立成育医療研究センター・ゲノム医療研究部)、東京大学医学部附属病院ゲノム医学センターの皆様にシークエンス頂いたデータのクオリティが極めて高く、本研究を円滑に進めることができました。

ご指導賜りました先生方、本研究にご協力頂いた先生方に深い感謝の意を表して、謝辞といたします。

引用文献

1 Whiteford, H. A. *et al.* Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* 382, 1575-1586, doi:10.1016/S0140-6736(13)61611-6 (2013).

2 Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of general psychiatry* 60, 1187-1192, doi:10.1001/archpsyc.60.12.1187 (2003).

3 Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421-427, doi:10.1038/nature13595 (2014).

4 International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455, 237-241, doi:10.1038/nature07239 (2008).

5 Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* 455, 232-236, doi:10.1038/nature07229 (2008).

6 Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* 505, 361-366, doi:10.1038/nature12818 (2014).

7 Millar, J. K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Human molecular genetics* 9, 1415-1423 (2000).

8 Blackwood, D. H. *et al.* Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *American journal of human genetics* 69, 428-433 (2001).

9 Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular psychiatry* 17, 142-153, doi:10.1038/mp.2011.154 (2012).

10 Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 154, 518-529, doi:10.1016/j.cell.2013.06.049 (2013).

11 Takata, A. *et al.* Loss-of-function variants in schizophrenia risk and SETD1A as a candidate susceptibility gene. *Neuron* 82, 773-780, doi:10.1016/j.neuron.2014.04.043 (2014).

12 Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506, 179-184, doi:10.1038/nature12929 (2014).

13 Acuna-Hidalgo, R. *et al.* Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *American journal of human genetics* 97, 67-74, doi:10.1016/j.ajhg.2015.05.008 (2015).

14 Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. *Science* 349, 1489-1494, doi:10.1126/science.aaa8954 (2015).

15 Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and

transcriptional history. *Science* 350, 94-98, doi:10.1126/science.aab1785 (2015).

16 McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science* 342, 632-637, doi:10.1126/science.1243472 (2013).

17 Cai, X. *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell reports* 8, 1280-1289, doi:10.1016/j.celrep.2014.07.043 (2014).

18 Kazazian, H. H., Jr. Mobile elements: drivers of genome evolution. *Science* 303, 1626-1632, doi:10.1126/science.1089670 (2004).

19 Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* 435, 903-910, doi:10.1038/nature03663 (2005).

20 Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* 460, 1127-1131, doi:10.1038/nature08248 (2009).

21 Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479, 534-537, doi:10.1038/nature10531 (2011).

22 Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 151, 483-496, doi:10.1016/j.cell.2012.09.035 (2012).

23 Evrony, G. D. *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* 85, 49-59, doi:10.1016/j.neuron.2014.12.028 (2015).

24 Upton, K. R. *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* 161, 228-239, doi:10.1016/j.cell.2015.03.026 (2015).

25 Poduri, A. *et al.* Somatic activation of AKT3 causes hemispheric developmental brain malformations. *Neuron* 74, 41-48, doi:10.1016/j.neuron.2012.03.010 (2012).

26 Lee, J. H. *et al.* De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nature genetics* 44, 941-945, doi:10.1038/ng.2329 (2012).

27 Lim, J. S. *et al.* Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nature medicine* 21, 395-400, doi:10.1038/nm.3824 (2015).

28 Jamuar, S. S. *et al.* Somatic mutations in cerebral cortical malformations. *The New England journal of medicine* 371, 733-743, doi:10.1056/NEJMoa1314432 (2014).

29 Armstrong, J., Pineda, M., Aibar, E., Gean, E. & Monros, E. Classic Rett syndrome in a boy as a result of somatic mosaicism for a MECP2 mutation. *Annals of neurology* 50, 692 (2001).

30 Topcu, M. *et al.* Somatic mosaicism for a MECP2 mutation associated with classic Rett syndrome in a boy. *European journal of human genetics : EJHG* 10, 77-81, doi:10.1038/sj.ejhg.5200745 (2002).

31 Coufal, N. G. *et al.* Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proceedings of the National*

Academy of Sciences of the United States of America 108, 20382-20387, doi:10.1073/pnas.1100273108 (2011).

32 Muotri, A. R. *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 468, 443-446, doi:10.1038/nature09544 (2010).

33 Bundo, M. *et al.* Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* 81, 306-313, doi:10.1016/j.neuron.2013.10.053 (2014).

34 Dal, G. M. *et al.* Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *Journal of medical genetics* 51, 455-459, doi:10.1136/jmedgenet-2013-102197 (2014).

35 Ewing, A. D. & Kazazian, H. H., Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome research* 20, 1262-1270, doi:10.1101/gr.106419.110 (2010).

36 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

37 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

38 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20, 1297-1303, doi:10.1101/gr.107524.110 (2010).

39 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

40 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* 31, 213-219, doi:10.1038/nbt.2514 (2013).

41 Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28, 1811-1817, doi:10.1093/bioinformatics/bts271 (2012).

42 Xu, H., DiCarlo, J., Satya, R. V., Peng, Q. & Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC genomics* 15, 244, doi:10.1186/1471-2164-15-244 (2014).

43 Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics* 13, 36-46, doi:10.1038/nrg3117 (2012).

44 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome research* 12, 996-1006, doi:10.1101/gr.229102. Article published online before print in May 2002 (2002).

45 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* 29, 24-26, doi:10.1038/nbt.1754 (2011).

46 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research*

29, 308-311 (2001).

47 Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research* 35, W71-74, doi:10.1093/nar/gkm306 (2007).

48 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80-92, doi:10.4161/fly.19695 (2012).

49 Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* 37, W305-311, doi:10.1093/nar/gkp427 (2009).

50 Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940-3941, doi:10.1093/bioinformatics/bti623 (2005).

51 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10-12 (2011).

52 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

53 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

54 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* 526,

68-74, doi:10.1038/nature15393 (2015).

55 Swiech, L. *et al.* In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nature biotechnology* 33, 102-106, doi:10.1038/nbt.3055 (2015).

56 Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 27, 2601-2602, doi:10.1093/bioinformatics/btr446 (2011).

57 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American journal of human genetics* 91, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).

58 Grandi, F. C., Rosser, J. M. & An, W. LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. *Molecular biology and evolution* 30, 503-512, doi:10.1093/molbev/mss251 (2013).

59 Suberbielle, E. *et al.* Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-beta. *Nature neuroscience* 16, 613-621, doi:10.1038/nn.3356 (2013).

60 Madabhushi, R. *et al.* Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *Cell* 161, 1592-1605, doi:10.1016/j.cell.2015.05.032 (2015).

61 Nishioka, M., Bundo, M., Kasai, K. & Iwamoto, K. DNA methylation in schizophrenia: progress and challenges of epigenetic studies. *Genome medicine* 4, 96, doi:10.1186/gm397

(2012).

62 Guo, J. U., Su, Y., Zhong, C., Ming, G. L. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* 145, 423-434, doi:10.1016/j.cell.2011.03.022 (2011).

63 Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* 511, 606-610, doi:10.1038/nature13544 (2014).

64 Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nature genetics*, doi:10.1038/ng.3469 (2015).

65 Bourdon, V. *et al.* Evidence of somatic mosaicism for a MECP2 mutation in females with Rett syndrome: diagnostic implications. *Journal of medical genetics* 38, 867-871 (2001).

66 Allebrandt, K. V. *et al.* A K(ATP) channel gene effect on sleep duration: from genome-wide association studies to function in Drosophila. *Molecular psychiatry* 18, 122-132, doi:10.1038/mp.2011.142 (2013).

67 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *The New England journal of medicine* 371, 2477-2487, doi:10.1056/NEJMoa1409405 (2014).

68 Rapoport, J. L., Giedd, J. N. & Gogtay, N. Neurodevelopmental model of schizophrenia: update 2012. *Molecular psychiatry* 17, 1228-1238, doi:10.1038/mp.2012.23 (2012).

図表

表 1. 本研究で使用した試料及び試料提供者のデータ

セット名	サンプル ID	試料	性別	年齢	精神疾患	死因	死後経過時間
CL_WGS_set	**AL30_cortex**	前頭葉	男性	68	罹患なし	肺塞栓	13 時間
	AL30_liver	肝臓	〃	〃	〃	〃	〃
NeuN_ WGS_set	**Y8763_NeuN+**	神経細胞核	男性	78	罹患なし	敗血症	9.5 時間
	Y8763_NeuN-	非神経細胞核	〃	〃	〃	〃	〃
	Y8763_liver	肝臓	〃	〃	〃	〃	〃
CC_ WGS_set	**S6_cortex**	前頭葉	男性	84	罹患なし	胃癌	6 時間
	S6_cerebellum	小脳	〃	〃	〃	〃	〃
MZ_WGS_set	**SBT1**	血液	男性	27	統合失調感情障害	-	-
	SBT4	血液	男性	27	罹患なし	-	-
MZ_Exome_set	**FT11**	血液	女性	41	統合失調症	-	-
	FT12	血液	女性	41	罹患なし	-	-
	JT11	血液	女性	46	統合失調症	-	-
	JT12	血液	女性	46	罹患なし	-	-
	TT21	血液	女性	28	統合失調症	-	-
	TT22	血液	女性	28	罹患なし	-	-
	TT11	血液	女性	60	妄想性障害	-	-
	TT12	血液	女性	60	罹患なし	-	-

図 1. アラインメント及びクオリティコントロールで使用したソフトウェアとパラメーター

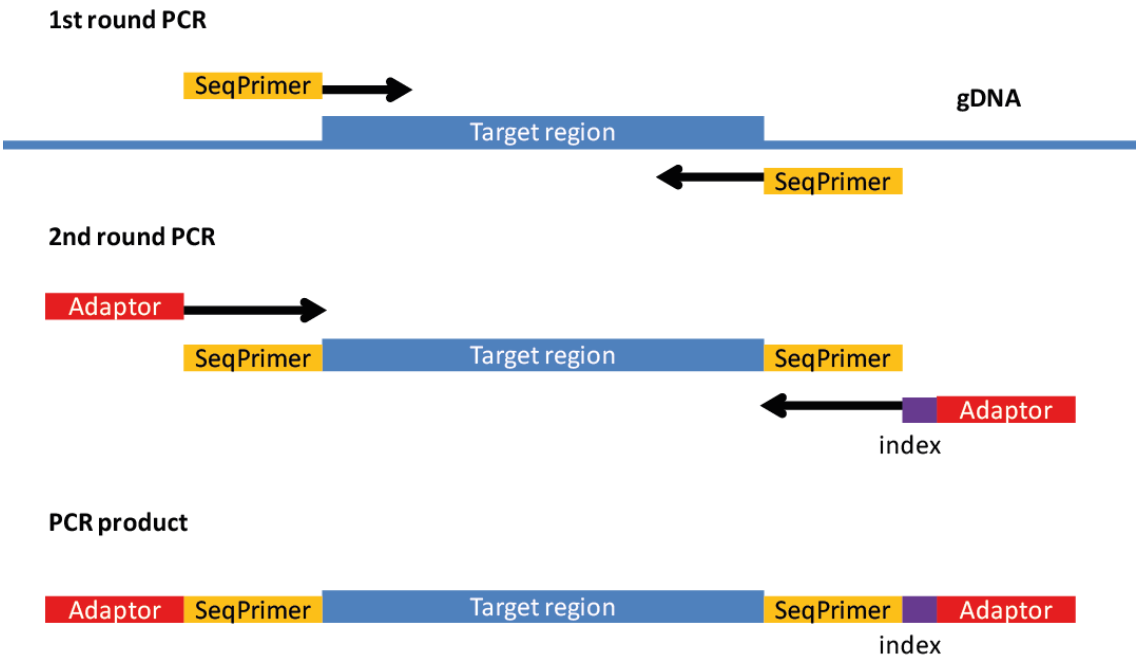
<table>
<tr><th>ソフトウェア</th><th>パラメーター</th><th>CL_WGS_set</th><th>NeuN_WGS_set</th><th>CC_WGS_set</th><th>MZ_WGS_set</th><th>MZ_Exome_set</th></tr>
<tr><td>Fastx_toolkit
Fastq_masker</td><td>q (baseQ)</td><td colspan="3">-</td><td>10</td><td>-</td></tr>
<tr><td rowspan="5">Trimmomatic</td><td>ADAPTOR</td><td colspan="5">TruSeq3-PE.fa:2:30:10 (イルミナアダプター配列)</td></tr>
<tr><td>TRAILING</td><td colspan="2">3</td><td colspan="2">4</td><td>5</td></tr>
<tr><td>LEADING</td><td colspan="2">3</td><td colspan="2">4</td><td>-</td></tr>
<tr><td>SLIDINGWINDOW</td><td colspan="2">4:15</td><td>4:18</td><td colspan="2">4:15</td></tr>
<tr><td>MINLEN</td><td colspan="2">30</td><td colspan="2">36</td><td>30</td></tr>
<tr><td>BWA</td><td>reference</td><td colspan="5">NCBI build37 + decoy (broad institute)</td></tr>
<tr><td>Picard
Deduplication (lane)</td><td></td><td colspan="5">yes</td></tr>
<tr><td rowspan="3">GATK
RealignerTargetCreator
IndelRealigner</td><td>known</td><td colspan="5">1000G_phase1.indels.b37.vcf (broad institute)
Mills_and_1000G_gold_standard.indels.b37.vcf (broad institute)</td></tr>
<tr><td>targetIntervals</td><td colspan="5">each bam files (each lane)</td></tr>
<tr><td>mode</td><td colspan="5">USE_READS</td></tr>
<tr><td>Picard
FixMateInformation</td><td></td><td colspan="5">yes</td></tr>
<tr><td>GATK
BaseRecalibrator
PrintReads</td><td>known_sites</td><td colspan="5">dbsnp_138.b37.vcf (broad institute)
1000G_phase1.indels.b37.vcf (broad institute)
Mills_and_1000G_gold_standard.indels.b37.vcf (broad institute)</td></tr>
<tr><td>Picard
Deduplication (merged)</td><td></td><td colspan="2">no</td><td colspan="2">yes</td><td rowspan="4">no need</td></tr>
<tr><td>GATK
RealignerTargetCreator
IndelRealigner (2nd)</td><td>Known</td><td colspan="4">1000G_phase1.indels.b37.vcf
Mills_and_1000G_gold_standard.indels.b37.vcf</td></tr>
<tr><td>targetIntervals</td><td colspan="4">merged bam files</td></tr>
<tr><td>Mode</td><td colspan="4">USE_READS</td></tr>
<tr><td>Samtools</td><td>q (mapQ)</td><td colspan="3">1</td><td>1, 60</td><td>1</td></tr>
</table>

ソフトウェアとパラメーターを、解析に使用した順に図に示した。MZ_Exome_set は最初に全てのレーンの fastq を結合して解析を始めた。その他のサンプルはレーンごとの fastq に対して解析をした後に、最後に BAM ファイルを結合した。

表 2. dbSNP サイトとシークエンスライブラリ作成用プライマーセット

Chr	Position	dbSNP ID	Primer_forward	Primer_reverse	Length (bp)
1	19945888	rs1770491	AAATACATTGCTGCCCAAAGAC	AGGGGGAAAATGTATGTTGTTG	214
2	40391789	rs1005213	TCAGATTAAACCTTCCCTTTGG	AGTCTCACCTTGGGACAGAAAC	247
2	159363774	rs1125662	TGCTCCATATATTTTGCAGTGG	TGGCTTTGAATTTACCTTCGAC	239
5	161381930	rs115725937	TGCTCAAATAAGTGAAAAGCAG	TGAACAATCTTAAAAACAAAGCAA	217
7	18993249	rs11505418	GATTTGCAAGGTAGGAGTTTGG	CACAGTGTTTTTACATCCAACCA	223
7	127583275	rs113453543	TGGTAGCTTTGGTTCCTTCAAC	GATTGAGGTGAGGTGGGAACTA	239
8	9472445	rs11249930	GATGTGGCTGATAAATGCTTTG	GTGGAGCTGCAAAAAGGAACT	229
10	4433008	rs10904247	TCAATCTGCTTTCTACCTTCAGC	TCAGAGGAGGGAGGACATTTAG	247
13	33896531	rs3848097	ACATGAGCCTTCCAAACAATG	AAAACATGCCTTTCCTACTTGC	207
14	39295782	rs4902177	CTTCCTAACCAACGCTTAGTGC	AGAGGTTCTCACTGATCGCTTC	226
17	35572413	rs3848462	TGATGCTTCTTATTTGAATCACAG	CTGGAAACAAATTCTTCAGTCAAC	240
18	10551782	rs10207	CTGCGGGAGAGCTTAGAATAAC	CTGCAATGTCCTTAAATGCAAA	228

図 2. 超高深度ターゲットアンプリコンシークエンスにおけるアンプリコン作成手順



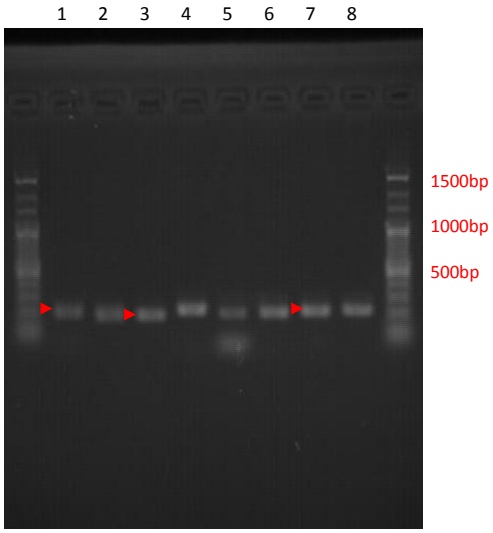
1 回目の PCR で対象領域を増幅するとともに、MiSeq 用のシークエンスプライマー配列を 5’末端に付加した。シークエンスプライマー配列に対して相補的なプライマーの 5’末端に MiSeq 用のアダプター配列を付加し、2 回目の PCR を行った。結果、図のように対象領域にシークエンスプライマー配列とアダプター配列が付加されたプロダクトが増幅された。2 回目のプライマーセットでインデクシングを行った。

表 3. 超高深度ターゲットアンプリコンシークエンスにおける 2 回目の PCR プライマーセット

Primer_name	Sequence
SeqPrimer_F	TCTTTCCCTACACGACGCTCTTCCGATCT
SeqPrimer _R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq_R_idxA002	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq_R_idxA003	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq_R_idxA004	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq_R_idxA005	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq_R_idxA006	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq_R_idxA007	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq_R_idxA008	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

SeqPrimer_F, SeqPrimer_R が 1 回目の PCR で 5’末端に付加したアダプター配列である。TruSeq_F, TruSeq_R が 2 回目の PCR で使用したプライマーセットであり、idxA00X がインデックスナンバーを示す。

図 3. TAS バリデーション用プライマー設計に当たっての試験的 PCR



CC_WGS_set の試験的 PCR 8 例を示した。chr14: 26673636 (lane 1,2), chr16: 65094255 (lane 3,4), chr3: 160478510 (lane 5,6), chr3: 180147593 (lane 7,8)を含むように 2 パターンずつプライマーセットを設計し試験的 PCR を行った。いずれも対象領域の長さのシングルバンドが得られた。この 8 例からは、赤い三角で示したプロダクトのプライマーセット 3 組を超高深度ターゲットアンプリコンシークエンスに使用した。

表 4. CL_WGS_set における体細胞 SNV 候補サイトとバリデーション用プライマーセット

Chr	Position	Primer_forward	Primer_reverse	Length (bp)
1	47016199	GATGAGCATCTGTGCCAGAAC	AAAAGGGTTTCGTCTTCTTTCC	227
1	110642378	GACATTTTTGTCCAAGAGAGGA	AAATCCTAGCTCTGCACATTCC	347
1	145004314	TTTTCCCCTAACACTCAGAAGC	CATCTGGTTATTCCAGGATTGG	228
1	206662735	GATGCTTTCCCATCACTTCATT	TGCACACAGCTCCATGTATGT	245
1	211248507	GAGATGTAGGAGAGGAGGTTTTTC	CAGAAGCTTTTGTCAACAGTGG	211
2	65847302	ATGGCATAAGCCTCATCTCCT	TCCCAGTTATTCTTCCCACTTC	229
2	76444676	AAACACACTTCAGAGAACAAAACTATT	TGTCCAAATTTTATCCTCAGTTC	246
3	105152628	GGAAAGGGCTTTCTATCACCTT	ACCTTTGTCTTATATTTTCTGACCA	336
3	142718948	AGCCTTGTGTCAGATCCTGTTC	TTAAATTGATGCTGCAAGGAAG	218
4	117374794	GCGACTGGACTTTATCCAGAAC	TGTGTATTGGATATGACGGTTTG	234
4	190637837	TCACACTGTGGGTTACAAGAGG	GGGCTACTGGTGAGTTTATGGA	242
5	21542513	CAAACGTGTTTTAAGCAATAAGGA	ACATAAAATGTACCGATGAATGC	218
5	137203810	TTTTTGTCCAAATCAGATGAGTG	ATCAAGTGTCCCTTTTCAAGGA	214
5	173320041	TGATTATGGGAAAAGGATGGTC	CTTTCAAATCACCCCACTGCTA	209
6	57413745	TGATGTCTCTTTACCAACAAAGGA	CCTTGAGATCATTCTCATTAGTGAAC	219
6	164440297	GGTTGTCCATGAGCATACTTGA	CAATGGTTTCTAAATTGCAAGAAG	231
6	169095224	CTTCCAAATACCCGACCTTG	GTGTAACATTCAGCCGAGCTAC	248
7	28777126	TTAAGAAGCAAGGTGAGGCATT	CTTCCTGAGACTATGCACCTTG	381
8	56888381	TTTCCTTCTGAAAAAGCCAAAG	CACCACCTCTGCCATACACTAA	217
8	124186999	ACTGGGATAGGCACTTGGAGT	TGTTGCATATACTCTGCATCCTG	234
9	32635670	TCGCATAAACCGGAAATAAAAC	GAACCCAGAAGGCAGAGCTT	235
9	129278657	TGCAGGCTAATTAACTCTGAGC	TTGAGGAAGCAACAAGTTTGAG	233
9	132122408	CTTTCTGGATTTTCTGGAGAGG	ATCACCAGAAGGACTGGGATTA	244
13	70176335	AAACCATATCAAGGCCATTCAA	TCCTGTCCAGAAATACGTCTCC	222
13	72252357	TCGTTTTCAAATGTTACTTCTCAG	ATGCCTTCATATACTGGGGAAC	237
14	42831852	TGCTACAACTTTTTGATGGCTATG	ATCAGTCAGGGCTAACCAGAAA	225
15	65477364	TTTACTTCCAATTTGAGCAACG	GCGCAGAGAGGTAACGAAATAG	218
16	3119020	GACTCTGGGAAAAGTCCCTCTT	CCTCAACATCCGGGACAG	244
16	7911650	AGGGAGTACAAGGCAACTTTTG	GGGTGGCATCATACTGACCTAC	242
16	31114885	TTGGTGACATGAGCGATATCTT	CAGAGGGAGACCGTGGAAG	222
16	64119939	GACCGAGGGGCTTCAATATAC	CACAACCTGGGCCTAAGTTAAA	528
17	2623125	TTCCTTTAGGAGGAAGGAAGAGA	AAAGAGGGTTATCCTGGGACTG	228
17	21208187	GTGATAAGCTCATGGACCACCT	CTCCCAATCCCCAAACTGAG	223
18	10196501	GCCCAAGTGGATTCTGAAGAC	CAGGACTCCACAGAGTTCACAG	244
18	61596240	AAGCAATTGTGAATGGGAGTTC	GCCAAGTATTGTCACAAAAGCA	229

18	62002768	AAAAATCTTTCCTGCATTGGTG	CACAGAGGGACAAAGGAAAGTC	243
	62002769			
21	47062881	AGCCAGCCCAGAGAAAGG	GAGCGAGACCTTGGTGGAC	365

表 5. NeuN_WGS_set における体細胞 SNV 候補サイトとバリデーション用プライマーセット

Chr	Position	Primer_forward	Primer_reverse	Length (bp)
2	20562894	GCAAAGGCCCATCTCTAAAAACT	TCATAAACAAATGGTGAATTACTTGTG	344
2	30504790	ACCCTTAGTCTGGAATGAAGGC	TTTTTGATGCAGGCTGTTGGG	229
2	80253532	ACAAGCATTCATCGATCCAAGG	CTTAACTCTTTTGCTGCAGGCC	221
2	144010826	GTGTGCTTCTGAGACTTGCAAC	TGACTATTGAGATTTTTGGTTCAAGGA	212
2	206520964	TGCAATCAATGAGTATGTGCAG	ATAAATCAGCAGTGCAAAAGCA	215
3	192234799	ACAGCATTGTCCTGTTCATTTG	AATTGGCTCATCTGGGATTCTA	215
4	40343190	ATAAATGAATTGGGGCTCAAAG	TAGGTGACAGAGCAAACTACCG	226
4	138417312	AGGTCTGAGTTCAATTTAAAAATAGGT	CCTATTCGGCCATCTTCCAGAA	230
4	166316873	TGTGCATGATATGTAGTTCAAAGGG	TTAGCCCAGTCTATACAGGCGA	221
7	141447818	TGGGCAATAGAGGAAGACTCTC	TGAATGTTTATGAAGTTATCAAGGA	374
7	147447647	CATGCTATCCTGAAGCAATCTG	TTTCTCATTTGGAGAACATTGG	226
11	14952526	TCCTCTCTTTGTTGCTAAGCCA	AAGTGCTGATAACTGGGTTCCA	217
11	112875224	GTATGATCCTCATCCTGGCAAA	TTCTTATTTTTCTTCTGCTGTGTG	385
11	117661790	GCAAGCATCATGTAGAGACAGG	AGAGGTCAAGTGCTGCCTATG	229
15	99259587	ACATGATGATCAGTGTGCCATT	TGTGCCGTTTCCTCAAACTCTA	218
17	72282109	ACAGGAAGAAGGAGATGAGTGC	GAGATGGGGTTTTGCTATGTTG	210
18	65440162	GCAGTGATTTCTCATTCTTGGTTTTG	GCAGAAGGTGACCTGTAACAAG	230
20	44667433	TGTCACAAACAGCATTGTAACTGA	TTCTCCCTCATCATTTGCCTCC	228
20	45395811	TCATGCCCAATACTCTACTGGA	GGAGTCCTTTCTCCCTGGAATG	217
21	23955109	TCAGTGGGATGATTGTTCAGAG	GCGAACATTTTATTTGTTGACAG	223
22	41257815	AGAAGTAGCTGAGGCATACGAG	CGATTTAACAGGTCCTCAAGCG	225
22	45393311	GTCACCTCTCCTTCCAAGCTC	TGCAAAGTGACAGGTAGCATTT	227
22	48883859	GCCTTTCTTTGCTCAGTTTCTC	AGGTGAAGGGATGTGGACTGT	214

表 6. CC_WGS_set（初回の解析）における体細胞 SNV 候補サイトとバリデーション用プライマーセット

Chr	Position	Primer_forward	Primer_reverse	Length (bp)
1	27190857	TTGAGAGGAAACATGGTCACAC	AGGCTACTTCTCCCCTCCTCTT	221
1	47281838	TGATTTTTCTCTCGCTGTTTGA	GATCCACTTGAAAGGGACTGAC	228
1	109167096	TCTATGGAGACAATGCCCTTTT	CATTTACGAACTGGAGTTGTGG	234
1	239023569	GCACCTAGGTCTCCTCTTCAAA	ATGGCTCTGAGCATCTAGGAAC	235
2	20562894	GCAAAGGCCCATCTCTAAAA	TCCTTGATATCATAAACAAATGGTG	353
2	102612321	ATCAATCCACTCACTTGCTCCT	ACAACATGAGCTGATTTCTTGC	229
2	151776679	GGTAATGATCACATCCAGGCTA	GACAGGGTGAATTCCACATAGG	236
2	170156074	ATCTTGTTGCTCGGTCACTTTT	TCTTAAGCCTGAGATTCCTTGC	238
2	202878979	CTAGGCAGGGCATGTTTAAGAG	GTACACCTGGGACTTCTTCGAC	222
3	29527382	TGGGTGGCATAGTTGAAGAAAG	CACGAGACCCACGAATGTAATA	249
3	66073554	ATGGTTCCTGATCCTACCAGTG	GAAACCCCAGCTCTACACAAAA	233
3	125258109	GCTGTGGTGACTCTTCATTAAAC	TTCTGCATACATCAATCCCGTA	236
3	180147593	TTTTCCCATGTCAAATTCCAAC	ATATCAGGGATGGGGGAGTATC	237
4	58560503	ACCCCATTTTTCCTATTTGCTT	ATCCTGGATCCTCCTACTTTCC	202
4	189822464	ATGTTTCCTCAACACACTTGGA	AAACCACCAGTGTGAAAATATCC	233
4	190633484	AACATGTATGCCAGTTTCCTCA	GGCTGGATTGTGCTATGTGTG	244
5	173320041	CTTTCAAATCACCCCACTGCTA	TGATTATGGGAAAAGGATGGTC	209
6	67384535	AACTTCCATCTGCCAAACTTTC	CATGAATACCAGAAGCATCATTG	215
7	152105445	GGGAACCCTGAGTACGATAAAC	CTGAATCCAGGGAGGTAGAGG	245
8	109128812	TTTGGAAGCAGAATGTCTAACG	CTGAGCCAGTCCCTAACTCAAG	229
9	10743921	ACTGCACTCCTTGACTTCCTGT	GAAAAAGCCAGAAGAGAAGAGC	248
9	109101822	GCTTGTCAACAGTGGAGTGAAG	CTTCACCACAGGCTCTACTTCC	394
11	110559709	ACATGTCACTTGTTGGGATTTG	TCTATGACATATCTTCAAATTGTGC	383
11	118320941	GACTTGGAACCTGAGGATTCTG	CAAAGCAAATGAGACAGGTGAC	235
12	86774226	CTTTCTACCAAACGAAGTCATCA	CCAAATTTCTGAAAGGCTGTTT	223
13	51922947	GCCCAAATCTTATCCCATCC	TGCCACCACTTCACTTTTTG	242
13	97895034	GTCAATCAAGCAACGTACAGGA	CCATGTGGAACACAGAAACATC	222
14	26673636	TTGGATGGTTTTCACATTATGG	ACCATGTCCCAACAGAAATACC	224
14	93953919	TAGCATTTGTATGCCAGGACAC	CATGGATGATGCAGAGAGGTTA	231
16	65094255	AAGGGGAAATATCTAAGCACACC	CTTCTCACCTCCACCCTAAAAA	210
17	21208979	GAAGTCCGAAGTCCAGGTGTAG	GATGGCATCGTCCTTCCAC	223
17	25623385	AGCTACGCAGAGAAATGTTCAA	CACTAAGTGTTCCCCCACTAGC	235
18	26648231	AACTTGAGTAGCATTGCCAAATC	AATTCTCAGCCAAACCTTGGTA	205
18	76867501	GGAGGCAGGAGAGTGAAGAC	AAAAATGAATCACTTGTGAGCAG	234
20	46250020	TCAAGATGCTGATCAAACTTCAA	AGCCTTTCCATTTTACAACCAA	207

21	41284253	AGGGCAACAGGATCAAAAATTA	TGACATCAGCCTCTGTTAATGC	216
X	29270668	TACACAGCGTCTCAAGTGTGG	CAGGGCATACAGGGAAATAAAA	227

表 7. CC_WGS_set（二回目の解析）における体細胞 SNV 候補サイトとバリデーション用プライマーセット

Chr	Position	Primer_forward	Primer_reverse	Length (bp)
1	240935724	AAGAGAAGTGAGCAGTCAACCA	TCTCCTCTGAGCGATGTTACAA	225
2	137814327	ATTCAGAAGCTGAACCGAACTG	CTCAGTCAGATTTGGACATTGC	231
4	126987958	ATAACAGCAACGTGGTCTCAAG	GATGAGCCGTTTCATATTACCC	217
6	170209984	GAAATACACAAGGCTGTTCACC	GTGTAGCATGCTGTCTTTCCAA	228
7	83812741	TGAGTAGCTATGTATGAAAACTCCA	CAGACACTGATCTGCCTTGCTA	219
12	39773148	GGTGGTGTCATAACATTCACAAA	GGGCAGAAGCACTAGCATAAAA	372
12	43522425	ACTCTTTTTCTGTGACCTTGCAG	CCTGCAATTTCCAAGCTTCTTA	210
12	48380047	CCCCTGGAGAAAGAGTTAAGTG	TGGGAAGGAGCTCAGACTATTT	222
12	132819457	AGTGGCCTCAGTGTCCAGTC	GAGGCGGGAGAAATAGAAATAGA	258
13	74859312	CCAAATTCTCCATGGTGTAAAAA	GCCTTGAGATTGGAGGTTTATG	220
18	28039112	CTCACAGACCAGAACTGGATGG	GACACATCCAAGATCACAATGG	216
18	66437389	TGACTATAGTGGCTCCTCTGCAT	TTGAAATGTTTGTTAGGCAGCA	212

表 8. MZ_WGS_set における体細胞 SNV 候補サイトとバリデーション用プライマーセット

Chr	Position	Primer_forward	Primer_reverse	Length (bp)
2	169920685	ACAAAGAGAGACTCCGTCTCCA	CACAGGTATCGGACCTGAAGTT	227
4	27468649	TTACTGAGTCAAATGGCGGTAA	TCTTCCATCTTAGCAACAGCAA	224
5	543501	AGCTCCGACACTGTTGTCCTAT	GGTCCAGTGAGTGTGCTCAAG	341
5	175021794	CTAGAAATGGAGACAGGGAAGG	CCAACACTCTCTTCCCTAGAGC	236
7	67737975	GTCTTCAAAGCGCTATTCCACT	TTCCCTGCAAAGAAAACTTCTG	227
8	24548901	ATCCCATCCACTGCTAACTCTC	GCACACGCCTGTAATCTTAGC	227
8	53091419	TGAGTGAGCACAAATCTCAGTTC	TTGATCAGTTCAAGGAAACAAAAA	222
10	5895525	AGACTGCGTGAAACTCATCTCA	AGTAACTCAGCCAAGGGAAACA	224
12	121114214	AGATCACATCATCATGCTCCAG	CATGGCAGAATGCTCTTGTCT	223
15	50631622	AGTGAGCCAAGATCACACTGCT	GGGACACTGATAAGCACAGATAG	372
15	65402075	TGGGCAAGAGACTTTTAATATGC	TTTCTGACCAACAGTGAATGAGA	228
16	85394970	CATTTGGAATGTGTGTCCTGTG	CATACACAGAGTACCCCCACAA	367
17	53577074	TCATCCAAATTAAGCCAAAACA	AACTGCAAGGTTCCACCTAATG	222
17	67718603	CTTCCTCACTCAGTCCTTAATCC	TGCTAAAGCTATCATTTTCATGC	500
20	58016089	ACCCCCTAGTCTCTCCTTTCTG	AAGTGGGAGATGATGATCCAAG	227

表 9. MZ_Exome_set における体細胞 SNV 候補サイトとバリデーション用プライマーセット

Chr	Position	Primer_forward	Primer_reverse	Length (bp)
1	21605869	TCACCTGCAGGGAAGGAG	TGCTTGACTCTCTGATGTTTGG	226
1	28607676	GGTGAATGGCAAGTAGGAGGTA	ACTTGGAGATCCTAAGGGACCA	247
1	28833911	TGTGGAAAGGGACTTGTACATC	TTTTTAACACCCCACTGTGGAC	226
1	39991592	ACAGCTCAAGCAGGAGTGTCA	CCACAGTCTGACAGGTCCTAAG	222
1	52821154	CCTGGAGAGTTTGTGTTCTTCA	AACTCGGTGAGGTTCAGGTAGA	217
1	78603073	CCAAAAAGGAAGGAGAAACTGA	AGTGAGTTCGTTTTGGAGGTTC	230
1	201010615	AAGAGACAGAGACGCCTGCTAC	GTAGCTGGTCCTGATGGTTTTC	229
1	205242142	AGTGTCTTCCCTGCTGAGTACC	ACAACCTCCTCCACACGAATAC	225
1	245849059	GCAAGTCAGAAAGGGACTGC	TTCGGACCCATTATCTTCCTTA	215
2	54871421	GTTTCTCCTTCACCACAGCTTC	GGGGTCGTTGTTGATTTTATTG	222
2	114718299	CTTTTCCCCCTTTTCTTACCAG	AGAACATTTTGCTCCCAGAGTC	211
2	175664545	CATAGACCCCAAAAGCAAAGTC	AACAGTGAACACACGCTTTCTG	223
3	42252628	CTTTCCTGCCCTGTATTTCAAG	TGCTCTGTCTAGACTGGACCTC	220
4	113110009	GCTCTGGTTCCACCAGTACCTA	TTTTTCCCATATCAAAGCCAAA	214
7	29440490	CCTATAGTTAACCCGCCTGTTG	ACAAAAGCTGCCGAGTACATTT	236
7	48413830	TAGTGTTTTGATTGGGACCACA	GACACTTCAAGCATCTCGGTAA	226
7	105641974	ACAGTCCTGGAGGAACTGAGTC	ACCAATTCTCATGGGTAACGAG	229
7	142637430	CTAGGCTTCAACGTGTTTTTCC	TGGAAAGAGTGCTTGTGAACAG	225
9	113341741	CGACCTTTCAGCAGATGTCC	AATCATCCACCAGGAAGACAAG	216
9	116779003	CAGAAACTGGCAATTACACCTG	AAAGACTGACCCAGAACTGAGC	224
9	117014809	AGGGACCCTCTAGAGACCTTGA	GGCAAGAGGATGCAGACTTACT	214
9	139849022	GGACCCAGAGGGAGGAGA	GAGTTTGGTTGACCCTAGAGGA	349
10	50960631	AAACTGGGTAAGACCCTTCCTC	CTCTCAGGCTTGAACACATGG	224
11	66307068	GCCAGCACTCCTATGACCTG	ACCCGACTTCCTTACTGAGTCT	239
11	72947061	GACAGCTAGTGAGAAGGCAGGT	AAGAGATGAACATCTGGGGACT	225
12	22040794	AGCACTTACAACTCCAGTGTGC	CACAGGCATCCTACTCACCATA	230
12	42491817	GATCACCCCATGTGATGTTTAG	TTGTGAGGAGTTTGCATTTCTC	214
12	54070004	CCCTGTAGTTTCTCCTCTCGAA	CCCACTGCAGAGAAGACAGAG	212
12	78571018	CAACTTACGCTGAAGCAGATTG	CCAAGTGCAAATAGACAAGAAGG	226
12	123752524	AAGGGCTCTGTATCCCAAGAC	GTTACCGCTGTTTGTCATTCTG	224
14	64591770	AATTAACATCAGCAGGGTCTAAA	CACATTGTGTGCAGCTTCTACC	210
14	65560458	TCAGGAAACTCACCTTCTCTCC	ACAGTGCTGACAGCGTGTACTT	220
16	75728247	TTTTCTTGCTCACATTCTTCCA	CTTTGTCAGAATTGGCAGTGTC	210
17	27381715	ACTGGCATCTACCTGCCTTATC	TCTTCATTTCCCACTTCCTCTC	219
19	5208010	TAGCATCTTCACCGTCTGAAAG	AAGCTTAGGCTGTCCCATCTG	214

19	42352925	GAGACAGCAGGTTCTTGGTACA	CCGTATCCCAGAATACACTCAA	237
19	44739117	TTTGCAGGTTATTTTTCACATCA	AGGTGTCTTCAGGGGAAGAGTT	223
19	49621802	TCTTGAGATGTGGAGGGAATG	CTCAGACGCTCTTAAGGGTAGG	230
20	58411402	AGGAGAGAAAGTGAACCTTTGG	ACAGATCTCCTCGCTGCTTAAA	218
21	31864375	TAGTGAGTTTGGTTTGCTGCTC	TGTTAAAGGGACAGCATGAGAA	229
22	43253229	AGATGGTCAAGATGTCCTGCTT	AAACGTTGGTGGATCCTATGG	395
X	37587307	ATCAACATGTTCTGCTGGTCTG	TGAAGAGAATGGCTGTGCAGTA	229
X	41077657	ATGCATTGTGTATTCTCCTTCG	TGACTAGAAGGTCCAGGAGAGG	220
X	50119242	GCAAGGACAGATAGATGAGTTGG	GGTGAGACAAGGATGGGTAAAG	213

表 10. パイロシークエンスに使用したプライマー配列

Chr	Position	PCR_Primer1	PCR_Primer2_biotinylated	Pyrosequence_primer
1	21605869	TCACCTGCAGGGAAGGAG	TGCTTGACTCTCTGATGTTTGG	GCGGGGAAGACGTGAGCCCC
1	39991592	ACAGCTCAAGCAGGAGTGTCA	CCACAGTCTGACAGGTCCTAAG	AGGGGCCCTCACTCTCGGTG
1	245849059	GCAAGTCAGAAAGGGACTGC	TTCGGACCCATTATCTTCCTTA	GGACTGCCTGAAGTGCAACA
7	105641974	ACAGTCCTGGAGGAACTGAGTC	ACCAATTCTCATGGGTAACGAG	ATCGTGGCCAATATCACAGC
11	72947061	AAGAGATGAACATCTGGGGACT	GACAGCTAGTGAGAAGGCAGGT	CAAGGGTCCTTTCTCCAATC
12	22040794	AGCACTTACAACTCCAGTGTGC	CACAGGCATCCTACTCACCATA	TCTTACAGGACTCAAAAGGA
12	78571018	CCAAGTGCAAATAGACAAGAAGG	CAACTTACGCTGAAGCAGATTG	GCTTACTGATTCATCCCTTC

PCR_Primer1: PCR に使用したプライマー1（ビオチン標識なし）

PCR_Primer2_biotinylated: PCR に使用したプライマー2（ビオチン標識あり）

Pyrosequence_primer: パイロシークエンスに使用したプライマー

図 4. 改良型 L1Hs-seq 評価のために設計した人工遺伝子配列

>Internal_Control_v1

CACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGCATTGGGAGATATACCTAATGCTAGATGACACGTTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACA

TATGTAACTAACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAGCGCAGGAACTTTAAGTCGAGTCAATAAAACTCGCATTACAGTGTTTACC

GCATCTTGTCGTTACTCACAAACTGTGATTCACCACAAGTCAAGCCATTGCCTCTTTGATACGCCGTAAGAATTAATATGTAAACTTTGCGCGGGTTA[CAAAG][ACAGA]TAC

TTAGGTTAGATCTTCTCCCGTGAATTTTAATCCTCATCATATATCAAGTATAAGGTAAGTCAAAAAAGCACGTAGTGGCGCTCTCCGACTGTTCCCAAATTGTAACTTATTGTTC

TGTGAAGGCCAGAGTTACTTCCCGGC[TTCAA]G[AGAAG]CCTTTCTATGTGCGCACCATATCCTCCTAATTCTTTGGTTATCTTTCCGAAGTAGGAGTGAACGAACTTTCGTTT

ACGTCTTATTACCAATGATATAGCTATGCACTTTGTATAGGGTACCAACAGGTTTCACAATTCCAAGAT[AGGCA][TGCAG]AGTGGGGATCCCGGCAAAAGACCTATATTTGC

GGTTACACTTAGCGATAAACCTCGATGCTACCTACTCAGACCTACTTTGCACGAAGTAAATAAGCGATTCATCCCAGCTGGTTCTTGGCGTTCTACGCAGCGACATGTATAT[T

GAGA][CTCTG]TACAAGTTGTTGTGTAGCACAAAACTTTTACCATAGTCGTAGAAATTCTCCAGAGTTAATTCATACCTAATGTCACAAATGTGATAGAACGCCAATGAGTATT

AGACTTTAGGTCGAGTACAGTTCGGTAACGGAGAAACCTTG[CTCTG][TGAGA]TACAAGTTGTATGGTAGCACAAAACTTTTACCATAGTCGTAGAAATTCTCCAGAGTTAA

TTCATACCTAATGTCACAAATGTGATAGAACGCCAATGAGTATTAGACTTTAGGTCGAGTACAGTTCGGTAACGGAGAAACCTTG[TGCAG][AGGCA]CGGCATACTTTATTAT

ATATATGAAACGTGCCCAAGTGATGCTAAACAAGCTTATGCTGTGTCCTGTGTTAGTTTAAGGGTAAACATACAAGTTGATTGAACATGGGTTGGGGGCTTACAATCATCGAA

GACTCTATAGTATCT[AGAAG][TTCAA]CGAAGACCAAGTAGGGCAGCCATTAGTTGATACAAGAACTATTTCGAAGGGCGAGCCCTTATCGTCTCTTTTGCGGATGACTTAAC

ACGTTAGGAACGTGAAATTGATTCCTTCGATGGTTATAAATCAAAAATTCAGAATGCT[ACAGA][CAAAG]GTCTGGAGCATGAATCTAACGGTATGTATCTCGATTACTTAGTC

GCTTTTCGTACTCCGCGAAAGTTCGTACCGCTCATTCACTAGATTGCGAAGCCTATGCTGATATATGAATTACAACTAGAGCACTGGGCCTCATGGGCCTTCCTTTCACTGCCC

GCTTTCCAG

>Internal_Control_v2

CACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGCATTGGGAGATATACCTAATGCTAGATGACACGTTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACA

TATGTAACTAACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAGCGCTGACCTAACTGACAAATTACCATAGATGACTAGCCATGCCATTAGC

TCTTAGATAACCCGATACAGTGATTATGAAAGGTTTGTGGGGTATAGCTATGACTTGCTTAGCTACGTATGTGAGTAGAAACTTTTCCGTATTTGTAT[CAAAG][ACAGA]ATAA

GGTTAGCCGAAAATGCACGTGGTGGCGTCCGTCGACTGCTCCCAGAGTGTGGCTCTTTGTTCTGTCAAGATCCAACCTTTATCACGATCGATTCTTTCTGCGGACCATGTTGT

TCTGATACTTTGGTCATATTTCCGTTG[TTCAA]G[AGAAG]TAGGAGTGAATTCACTTAGCTTTGCGCCATAATTTTAATGAAAAACCTATGCACTTTGTTTAGGGTACCATCAG

GAATCTGAACCTTCAGATAGTGAAGATCCCGAGTATAGACCTTTATCTGCGGTACAACTTAAGCATAAAC[AGGCA][TGCAG]CTGCATGCTATCTTGTTAGACCTACTGTGCA

CGAAGTAAATATAGGATGCGTCCGACCTGGCTCCTGGCGTTCCACATCGTCACGTATTCGTTAACTGTTAATTGGTGACACATAAGTAATATTATAGTCCCTGAAATTCAGCT[T

GAGA][CTCTG]CACTTATTTTGAGCGTAATGTCTCAAATGGCGTAGAACAGCAATGACTGTTTGACACATAGTGGTGTTCAGTTCGGTAACGGAGAATCTATGCGGCATTCTTA

TTAATACATTTGAAACACGCCCAATTGATACTAAACAAATC[CTCTG][TGAGA]AATGCTGACTCCCATGTTAGAATAAGGATAAACATACAAGTCGATAGTAGATGGGTAGGG

GCTTTTAATTCATCCAACACTCTACGGTTTCTTGAAGAGACAGTAGGGCACCCTGTAGTTCGAAAAGAAATTATTTCGTAAGGC[TGCAG][AGGCA]ATGCTCATACCGTCTTT

TTTGCGGAAGACTTAACACGTTAGGAAGATGGAATAGTTTCGAACGATGGTTATTAATCGTAATAACGGAACGCTGTCTGAAGGATGAGTCTGACGGTGTGTAACTCGATCA

GTATCTCACTATTCG[AGAAG][TTCAA]AACTGCGCGAAAGATCCCAGCGCTTATGCACTTAATTCCGAGGCCTGTCCCGATATATGAACTTAAACTAGAGCGGGACTGTTGAC

GTTTGGAGTTGAAAAAATCTATTATACTAATCGGCTTGAACGTGCTCTACAGCAGCCA[ACAGA][CAAAG]CCTGACGAGGGGCCCACACCGAGGAAGTAAACTGTTATACGT

TGGGGATAGTGCTAACTAATAAAGATGCTTGCTTGAACAAAAGTATCAAACCCGTATAAAGAGAACATCCACACTTTAGTGAATTGCTGGGCCTCATGGGCCTTCCTTTCACT

GCCCGCTTTCCAG

>Internal_Control_v3

CACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGCATTGGGAGATATACCTAATGCTAGATGACACGTTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACA

TATGTAACTAACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAGCGCCATGTTAAACTGCTAGTAAACCGCGATTCTACGACTGGTGCATAAT

TTAATTACGCTGACGTGATGACATTCCTGCTAATGCCTCACCTGTCGGATCGCTCTCGTGATAGGGTAGTTGGACATGACCTTGTAACATATAACAAG[CAAAG][ACAGA]GTT

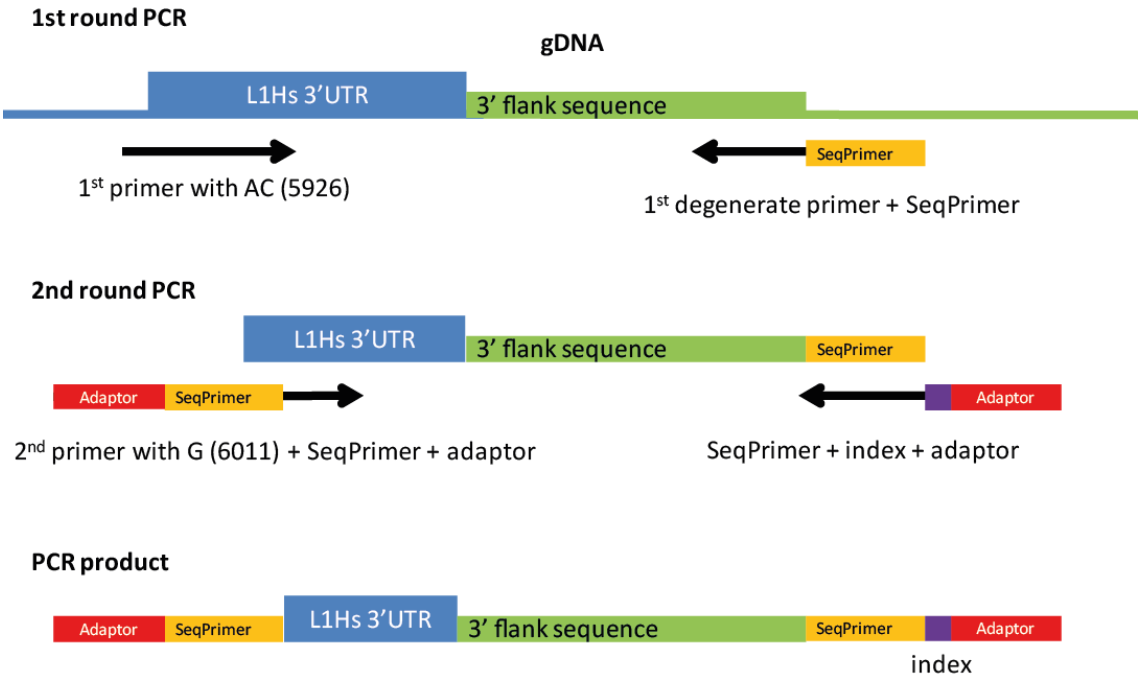
CACCTATCTATTACCCATGCCCGAAGATTATGTAGGTTGTGTGATGCTGGAGTAGTTCTCGATCTTCTCGTAGGACGTCAACCTTTTCTTTAATAAAGCATTCCATTCGAGTATG

GCAGTAAGTACGCTTTCTGAATTGTG[TTCAA]G[AGAAG]CTAATCTTCATCCTTATCATAGTTTGCTGTCAATGATTAGGATTATTGCCTTGCAATAGACTTCTTATTCACACTC
GCTCACATTGAGCTACTCGATGGGTTATCAGCTTGACCCGCTGTGTAAGGTCGCGATTACGTGAGTTA[AGGCA][TGCAG]GGGCTGTGAACTGCGCTGTATAGTCGAATCTGA
TTTCGCCCTCACAACTGCAAACCCCAACTTATTTAGATAACATGATTAGCCGAAGTTGTACGGGATATCCACCGTGAACTCCTCCCCGGGTGTCGCTCCTTCATTTGATAA[TG
AGA][CTCTG]TATGCTGCCGCTACCATTATTGATTAATACAACGAACGGTGATATTATCATAGATTCGGCACATTTCTCTTGTAGGTGTGAAATCACTTAACTTCGTGCCGAAGT
CTTATGGCAAAATCGATGGACTATGTTTCGGGTAACACT[CTCTG][TGAGA]TCGCATCAACGTGTATTCGAATATTGTTAATTGTTCACACATGAACAAAATAGTAGACTGTCA
ATTTCAGCCCTATTATCCTCGGCGTTGTGTGTTAAATGGCGTAGATCTGAATTGACTCTATAATGGTATTTACTGATGGGT[TGCAG][AGGCA]ACTCCCAATAAGGGATCCATAT
TTAAAGAATAAGTGTCAGATAGTATAACCCGATGAGGTATCCAAAAGCTGAACTGAGCCAGACAATCCGGTGCTAACGCACTCATAGCCGGGACACGACGCGACATATCGGC
TAAGAGTAAG[AGAAG][TTCAA]CCGGGAGTATAGACCTTTGGGGTTGAATAAATCTGTCGTAGTAACCGGCTACAACAACCCGTATAAGTGGCATTTCAGGAGGGGCCCGC
AGGGAGGAAGTTTTCTACTATTCGTGGCCGTTCGTAATAACTAGTTGCGTTCCTA[ACAGA][CAAAG]GCCACTATAATTGTTTCTAAGCCGTGTAATGTGAACAACCACACCA
TAGCGAATTAATGCACCGCCTCGGAATACCGTTTTAGCAACCCCTTACTAAGACCATTACGATTTTCAGGTATCGTGTATGTCTGGGCCTCATGGGCCTTCCTTTCACTGCCCG
CTTTCCAG

1500 bp 前後となるよう人工的に合成した DNA 配列を示した。赤字部分が L1Hs 3’末端の配列であり、下線部が L1Hs に特異的なプライマーが結合する配列である。[緑字]で示した 5 塩基（8 種）が縮重プライマーの結合配列であり、各人工遺伝子で 8 種が 2 回ずつ出現するよう均等に割り振っている（計 16 ヶ所）。その他の配列は、Random DNA Sequence Generator*にて生成したランダムな配列である。緑字以外に縮重プライマー結合部位がないこと、UCSC BLAT にてヒトリファレンスゲノムに相同配列がないことを確認し合成を行った (Life Technologies)。1500 bp 中の縮重プライマー結合配列の出現頻度の期待値は 11.7 ヶ所なので、期待値より若干多い設定となっている。

* http://www.faculty.ucr.edu/~mmaduro/random.htm

図 5. 改良型 L1Hs-seq におけるアンプリコン作成手順



L1Hs 3'末端に特異的なプライマー（L1Hs 標準化配列において 5'末端から数えて 5925, 5926 番目の塩基が A, C であり、その部位に結合するよう設計された配列：表 11 の L1HsTAILSP1A2）と縮重プライマー8 種を用い、1 回目の PCR で、L1Hs 3'末端及び隣接領域を増幅した。縮重プライマーには MiSeq 用のシークエンスプライマー配列を 5'末端に付加した。L1Hs 3'末端に特異的な別なプライマー（L1Hs 標準化配列において 5'末端から数えて 6011 番目の塩基が G であり、その部位に結合するよう設計された配列：表 11 の Adap1L1HsG）と、シークエンスプライマー配列に対して相補的なプライマーの 5'末端に MiSeq 用のアダプター配列を付加したプライマーで 2 回目の PCR を行った。2 回の PCR により、L1Hs 3'末端と隣接領域に、シークエンスプライマー配列とアダプター配列が付加されたプロダクトが増幅された。2 回目のプライマーセットでインデクシングを行った。

表 11. 改良型 L1Hs-seq で使用したプライマーリスト

Primer_name	Sequence
L1HsTAILSP1A2 (1st forward)	GGGAGATATACCTAATGCTAGATGACAC
DEGSeq1N5TCTGT (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNTCTGT
DEGSeq1N5CTTCT (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNCTTCT
DEGSeq1N5CTGCA (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNCTGCA
DEGSeq1N5TGCCT (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNTGCCT
DEGSeq1N5TCTCA (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNTCTCA
DEGSeq1N5CAGAG (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNCAGAG
DEGSeq1N5TTGAA (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNTTGAA
DEGSeq1N5CTTTG (1st reverse)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNCTTTG
Adap1L1HsG (2nd forward)	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCACATGTACCCTAAAACTTAG
Adap2_idxA001 (2nd reverse)	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Adap2_idxA002 (2nd reverse)	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Adap2_idxA003 (2nd reverse)	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Adap2_idxA004 (2nd reverse)	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

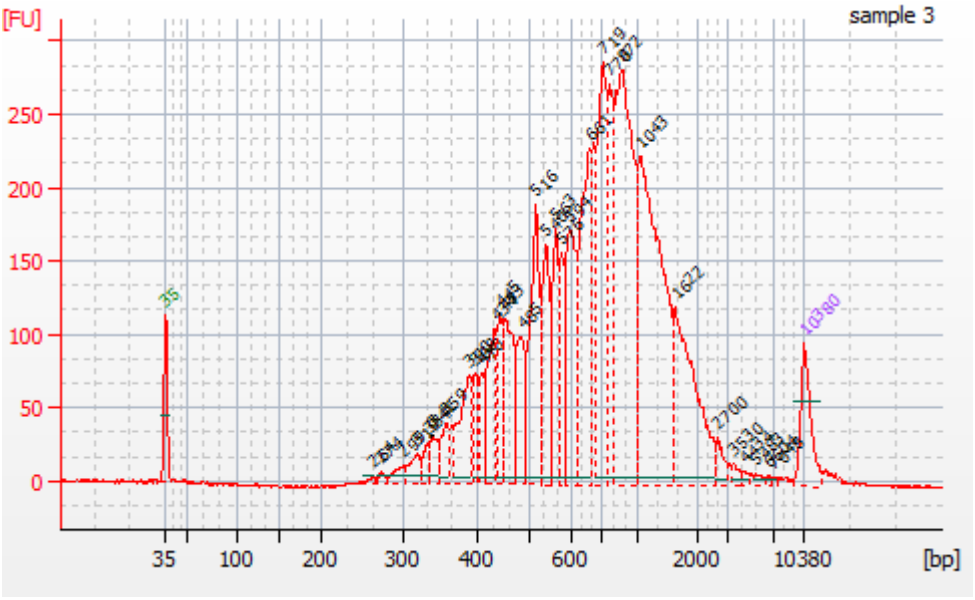
L1HsTAILSP1A2: L1Hs 標準化配列において 5’末端から数えて 5925, 5926 番目の塩基が A, C であり、その部位に結合するよう設計された配列

DEGSeq1N5: 縮重プライマー（8 種）

Adap1L1HsG: L1Hs 標準化配列において 5’末端から数えて 6011 番目の塩基が G であり、その部位に結合するよう設計された配列

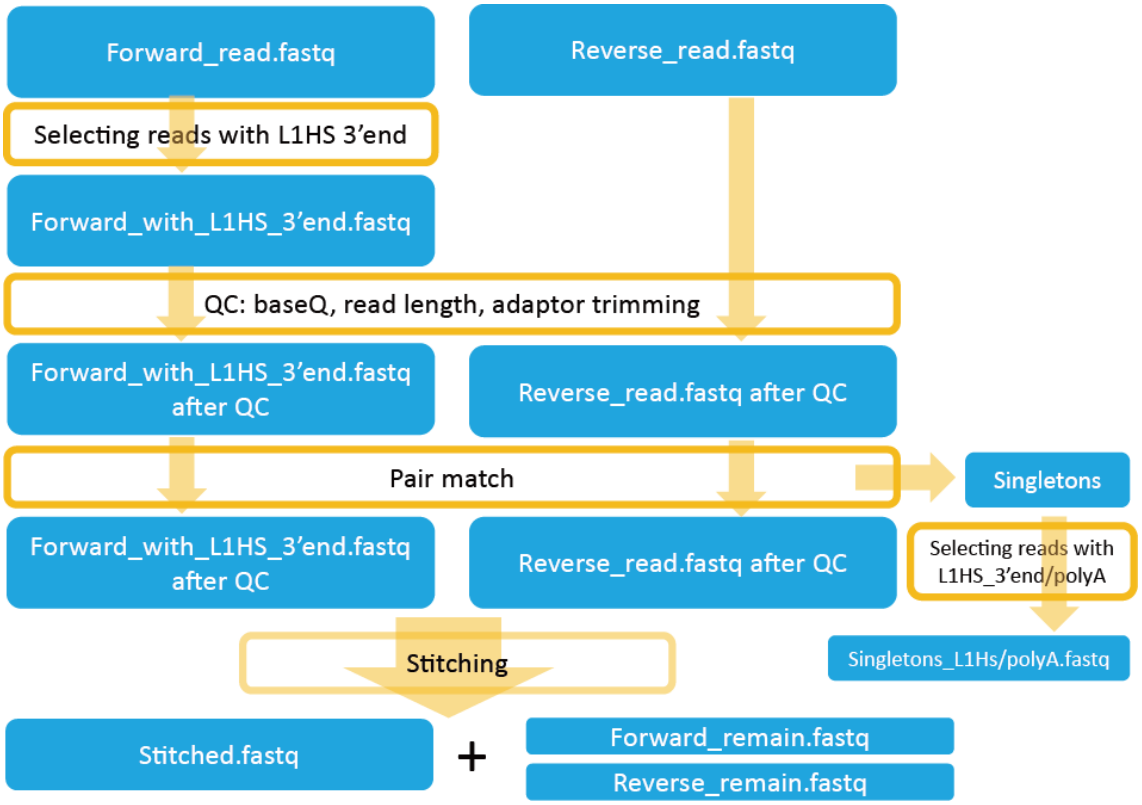
Adap2_idxA00X: Miseq 用アダプターの配列を含むプライマー。A00X がインデックス番号を示す。

図 6. 改良型 L1Hs-seq プロダクトにおけるフラグメント長の分布



Bioanalyzer による解析により、270bp-1600bp のプロダクトが確認された。

図 7. 改良型 L1Hs-seq におけるシークエンスリードのクオリティコントロール



Cutadapt を用いて、L1Hs 配列を含むフォワードリードを選択し、アダプター配列やクオリティの低いベースコールを除去した。ペアエンドとして残るリードとシングルエンドのリードを分離し、前者に関しては FLASh を用い、フォワードリードとリバースリードを厳密な閾値（-M 300 -m 20 -x 0.1）でスティッチングすることで、ペアリードから元の長いフラグメントを構成し直した。シングルエンドのリードは、L1Hs 配列またはシークエンスクオリティの高い poly-A 配列を含むリードのみ選択した。この操作で、スティッチングした長いリード、スティッチングされなかったペアリード、残りのシングルリードの 3 種類が、重複なく分割された。

表 12. 非リファレンス L1Hs 挿入位置バリデーションに用いたプライマーセット（Nested PCR）

Chr	Position	1st Forward Primer	1st Reverse Primer	2nd Forward Primer	2nd Reverse Primer	Length (bp)
1	119553351	TGCACATGTACCCTAAAACTTAG	AAGGAAGAGCCCTGCCTAGT	ATGTACCCTAAAACTTAGAGTAT	GCGTGTATGGCTGAGATAGA	393
1	119553351	TGCACATGTACCCTAAAACTTAG	AAGGAAGAGCCCTGCCTAGT	ATGTACCCTAAAACTTAGAGTAT	ATCAAACCCAGCATTGTGTG	346
1	119553351	TGCACATGTACCCTAAAACTTAG	AAGGAAGAGCCCTGCCTAGT	ATGTACCCTAAAACTTAGAGTAT	AAGAAGGCCAGCATTCCATA	228
1	119553351	CACATGTACCCTAAAACTTAGAG	GAGCCCTGCCTAGTTACAGT	ATGTACCCTAAAACTTAGAGTAT	GCGTGTATGGCTGAGATAGA	393
1	119553351	CACATGTACCCTAAAACTTAGAG	GAGCCCTGCCTAGTTACAGT	ATGTACCCTAAAACTTAGAGTAT	ATCAAACCCAGCATTGTGTG	346
1	119553351	CACATGTACCCTAAAACTTAGAG	GAGCCCTGCCTAGTTACAGT	ATGTACCCTAAAACTTAGAGTAT	AAGAAGGCCAGCATTCCATA	228
5	33797557	TGCACATGTACCCTAAAACTTAG	GTCCAGCCTTTGGGAATGGA	TACCCTAAAACTTAGAGTATAAT	CGTCAACAAATCTAGATGCC	354
5	33797557	TGCACATGTACCCTAAAACTTAG	GTCCAGCCTTTGGGAATGGA	TACCCTAAAACTTAGAGTATAAT	AAGGCATCTCTAAATCCAGG	429
5	33797557	TGCACATGTACCCTAAAACTTAG	AAGGCAGGACTGAGAGGACT	TACCCTAAAACTTAGAGTATAAT	CGTCAACAAATCTAGATGCC	354
5	33797557	TGCACATGTACCCTAAAACTTAG	AAGGCAGGACTGAGAGGACT	TACCCTAAAACTTAGAGTATAAT	AAGGCATCTCTAAATCCAGG	429
8	75723721	CACATGTACCCTAAAACTTAGA	TAGGGGAGGGAAGTGAAAGG	ATGTACCCTAAAACTTAGAGTAT	CATTTCTGCTGGTGTCCAAG	238
8	75723721	CACATGTACCCTAAAACTTAGA	GTCAGCCTCACATTTACCACA	TACCCTAAAACTTAGAGTATAAT	TATTAGGGGAGGGAAGTGAA	517
8	75723721	CACATGTACCCTAAAACTTAGA	GTCAGCCTCACATTTACCACA	ATGTACCCTAAAACTTAGAGTAT	CATTTCTGCTGGTGTCCAAG	238
8	75723721	CACATGTACCCTAAAACTTAGAG	TTTTCCAGGAGATGGTGGCA	TACCCTAAAACTTAGAGTATAAT	CACATATAGGGCATTTCTGC	246
13	61462344	TGCACATGTACCCTAAAACTTAG	TGGGGCTTTTGCAGTCATGT	ATGTACCCTAAAACTTAGAGTAT	TCCTTCTGTTTCCTGTCTGG	390
13	61462344	TGCACATGTACCCTAAAACTTAG	TGGGGCTTTTGCAGTCATGT	ATGTACCCTAAAACTTAGAGTAT	TGTCTTTCCCTTCTCCTTGG	475
13	61462344	ATGTACCCTAAAACTTAGAGTAT	TGTCTTTCCCTTCTCCTTGG	ATGTACCCTAAAACTTAGAGTAT	TCCTTCTGTTTCCTGTCTGG	390
13	61462344	CACATGTACCCTAAAACTTAAAG	GTGGGGAGAAAAGGAGGAATG	ATGTACCCTAAAACTTAAAGTAT	GGCTTAGTTCACAAGTCTCC	407
13	61462344	CACATGTACCCTAAAACTTAAAG	GTGGGGAGAAAAGGAGGAATG	ATGTACCCTAAAACTTAAAGTAT	GGCCAATGTCATAACACTCA	329

表 13. 本研究で得られたシークエンスデータ

セット名	サンプル ID	対象領域	シークエンサー	PCR	リード長	リードペア数	理論的深度	QC 後深度
CL_WGS_set	**AL30_cortex**	WG	HiSeq 2500	-	162	945708090	97.7	84.2
	AL30_liver	WG	HiSeq 2500	-	162	951672483	98.3	81.5
NeuN_WGS_set	**Y8763_NeuN+**	WG	HiSeq 2500	-	162	940935271	97.2	83.5
	Y8763_NeuN-	WG	HiSeq 2500	-	162	943326695	97.4	85.1
	Y8763_liver	WG	HiSeq 2500	-	162	915540856	94.5	80.6
CC_WGS_set	**S6_cortex**	WG	HiSeq X	+	150	1282279777	122.6	76.3
	S6_cerebellum	WG	HiSeq X	+	150	1257772546	120.3	74.1
MZ_WGS_set	**SBT1**	WG	HiSeq X	+	150	855613506	81.8	56.1
	SBT4	WG	HiSeq X	+	150	888176560	84.9	59.6
MZ_Exome_set	**FTW_11**	V4+UTR	HiSeq 2000	+	100	53940007	153.3	92.1
	FTW_12	V4+UTR	HiSeq 2000	+	100	59159933	168.1	99.7
	JT11	V5+UTR	HiSeq 2000	+	100	54100195	145.1	74.8
	JT12	V5+UTR	HiSeq 2000	+	100	60187660	161.4	85.4
	TT21	V5+UTR	HiSeq 2000	+	100	43965857	117.9	64
	TT22	V5+UTR	HiSeq 2000	+	100	48817259	130.9	69.4
	TT11	V5+UTR	HiSeq 2000	+	100	56239238	150.8	79.9
	TT12	V5+UTR	HiSeq 2000	+	100	65675687	176.1	92.5

PCR: ライブラリー作製時における PCR の有無

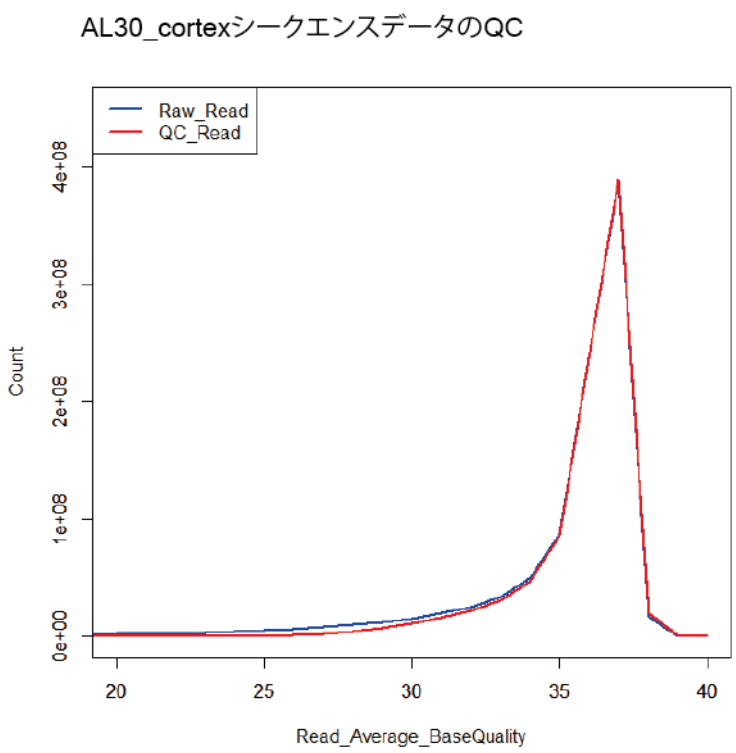
QC: クオリティコントロール

WG: 全ゲノム領域

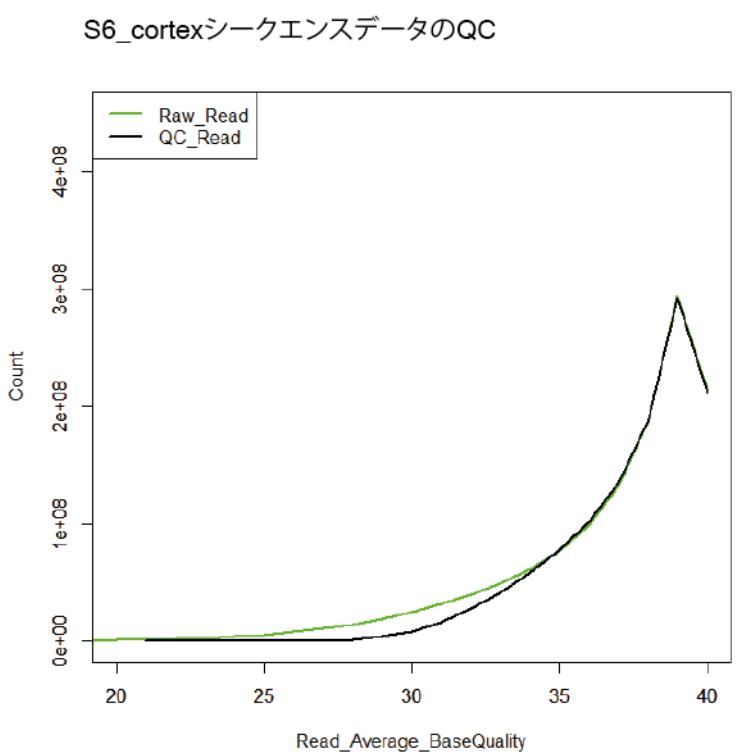
V4/V5+UTR: Agilent SureSelect Human All Exon V4/V5+UTR でエンリッチされたエクソン領域

図 8. FastQC による WGS データクオリティの比較

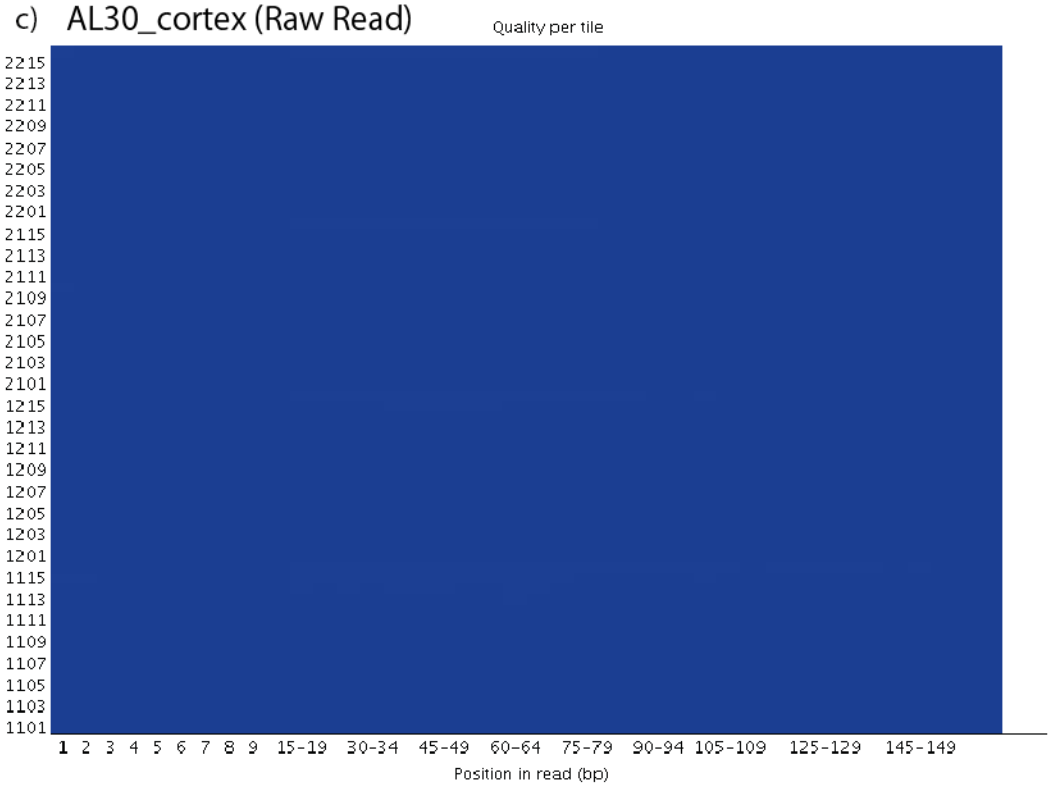
a)



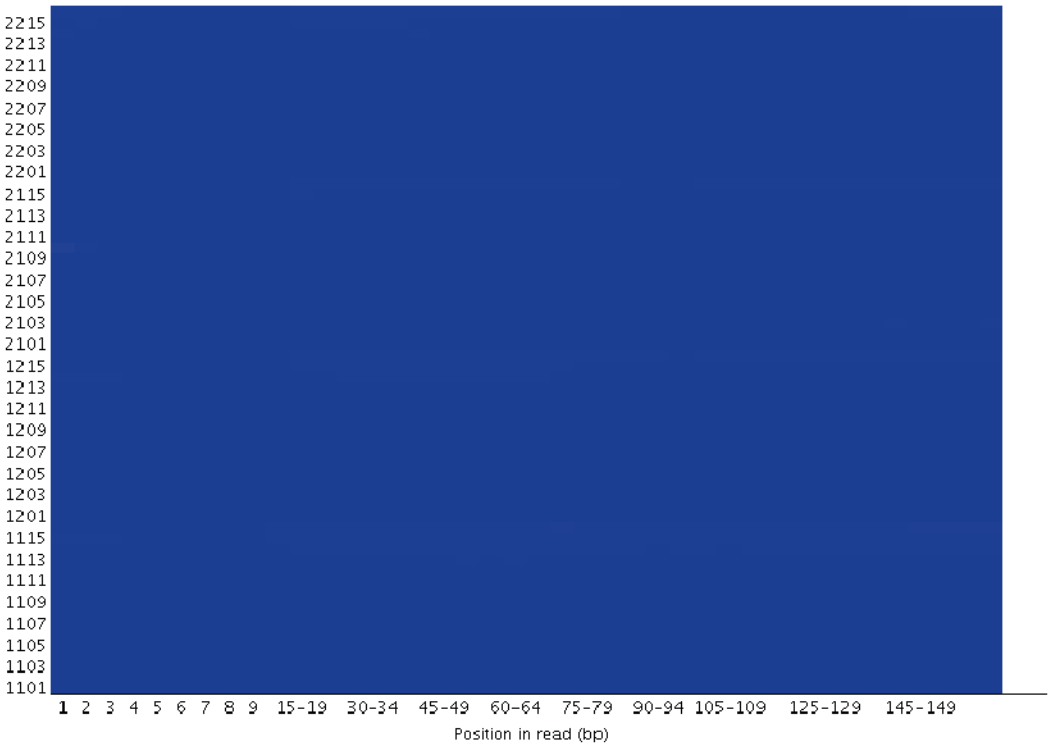
b)



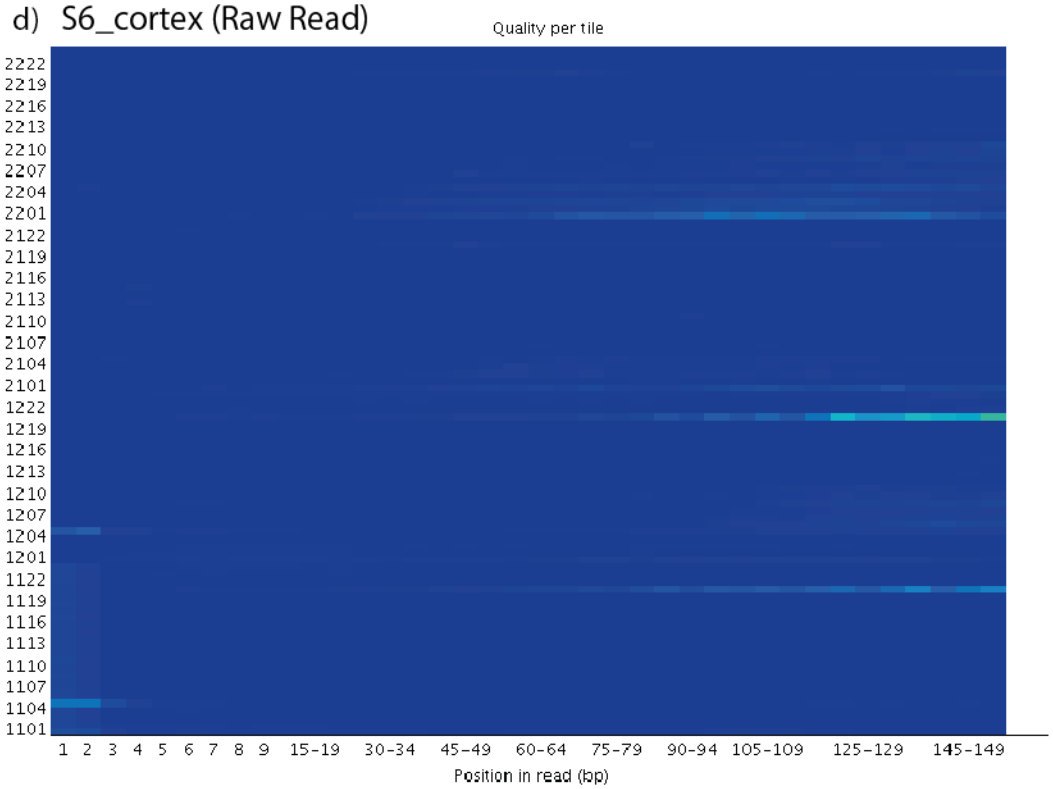
c) AL30_cortex (Raw Read)



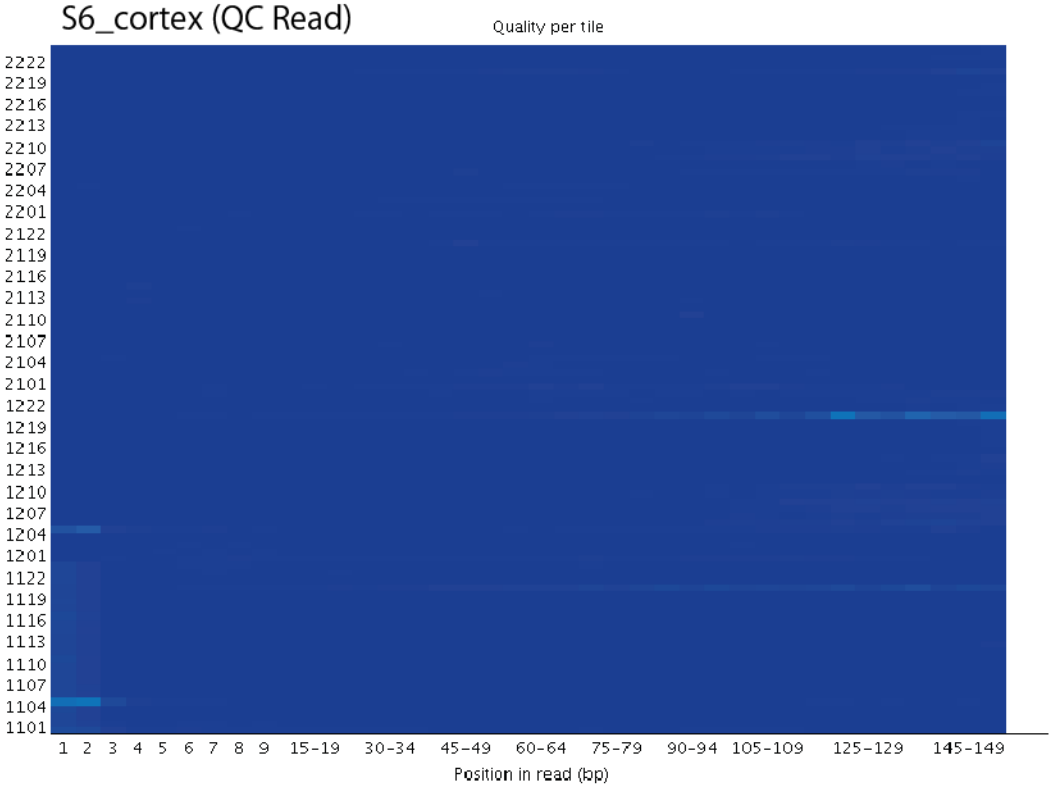
AL30_cortex (QC Read)



d) S6_cortex (Raw Read)



S6_cortex (QC Read)



a) AL30_cortex の WGS フォワードリードにおけるシークエンスクオリティ（リードの平均ベースクオリティ）の分布を示した。青線が QC 前のリード、赤線が Trimmomatic による QC 後のリードを示す。AL30_cortex は QC 前からリードのクオリティが高く、厳しい QC は不要であった。
b) S6_cortex の WGS フォワードリードにおけるシークエンスクオリティ（リードの平均ベースクオリティ）の分布を示した。緑線が QC 前のリード、黒線が Trimmomatic による QC 後のリードを示す。S6_cortex は QC 前のリードに低クオリティのものがあり、体細胞変異の偽陽性を予防するため厳しい QC を必要とした。QC により、分布が左に寄り、ほとんどのリードが平均ベースクオリティ 30 以上となった。
c) AL30_cortex の WGS フォワードリードにおけるシークエンスタイルごとのベースクオリティを示した。X 軸がリード内ポジション、Y 軸がシークエンスタイルであり、青が濃いほどクオリティが高いことを示す。上段が QC 前リード、下段が QC 後リードである。AL30_cortex のデータは、QC 前からほぼ濃い青であり、クオリティが高いことが示されている。
d) S6_cortex の WGS フォワードリードにおけるシークエンスタイルごとのベースクオリティを示した。X 軸がリード内ポジション、Y 軸がシークエンスタイルであり、青が濃いほどクオリティが高いことを示す。上段が QC 前リード、下段が QC 後リードである。S6_cortex の QC 前リードデータには水色が散見され、一部のタイルでのクオリティが低いことが示されている。QC により水色が少なくなり、低クオリティのリードが除去されたことがわかる。

表 14. CL_WGS_set、NeuN_WGS_set における体細胞 SNV 候補の同定

MuTect 解析	**CL_WGS_set**		**NeuN_WGS_set**					
解析対象試料	AL30_cortex	AL30_liver*	Y8763_NeuN+	Y8763_NeuN+	Y8763_NeuN-	Y8763_NeuN-	Y8763_liver*	Y8763_liver*
比較対照試料	AL30_liver	AL30_cortex	Y8763_NeuN-	Y8763_liver	Y8763_NeuN+	Y8763_liver	Y8763_NeuN+	Y8763_NeuN-
MuTect 結果	21312	19289	19662	20417	20366	20964	19422	19155
多コピー領域除外	196	184	175	180	180	167	271	258
INDEL 除外	165	153	135	140	122	129	224	212
多コピー疑い領域除外	125	126	109	117	96	106	204	191
BLAT < 150	53	88	74	81	62	72	155	149
BQ ≥ 21, DP ≥ 40	45	58	45	37	32	37	119	127
HC	6	22	9	5	5	7	87	89
LC	4	4	4	3	1	1	7	7
Strelka 解析	**CL_WGS_set**		**NeuN_WGS_set**					
解析対象試料	AL30_cortex	AL30_liver*	Y8763_NeuN+	Y8763_NeuN+	Y8763_NeuN-	Y8763_NeuN-	Y8763_liver*	Y8763_liver*
比較対照試料	AL30_liver	AL30_cortex	Y8763_NeuN-	Y8763_liver	Y8763_NeuN+	Y8763_liver	Y8763_NeuN+	Y8763_NeuN-
Strelka 結果	354	441	352	401	382	444	700	680
多コピー領域除外	17	25	18	23	14	17	88	76
INDEL 除外	14	25	16	21	12	14	86	72
多コピー疑い領域除外	-	-	-	-	-	-	-	-
QSS ≥ 21, DP ≥ 40	10	13	9	10	3	4	55	50
BLAT < 150								
HC	2	9	1	1	1	0	53	42
LC	1	0	0	0	0	0	1	1

INDEL: 挿入・欠失（insertion/deletion）

BQ: 体細胞 SNV 候補を支持するベースコールの平均ベースクオリティ（base quality）
DP: 体細胞 SNV 候補サイトの深度（depth）
HC: High Confidence
LC: Low Confidence
* 肝臓を解析対象試料とした解析は HC, LC の選択まで行ったが、バリデーション実験は行っていない。

表 15. CL_WGS_set, NeuN_WGS_set, CC_WGS_set における多コピー疑い領域

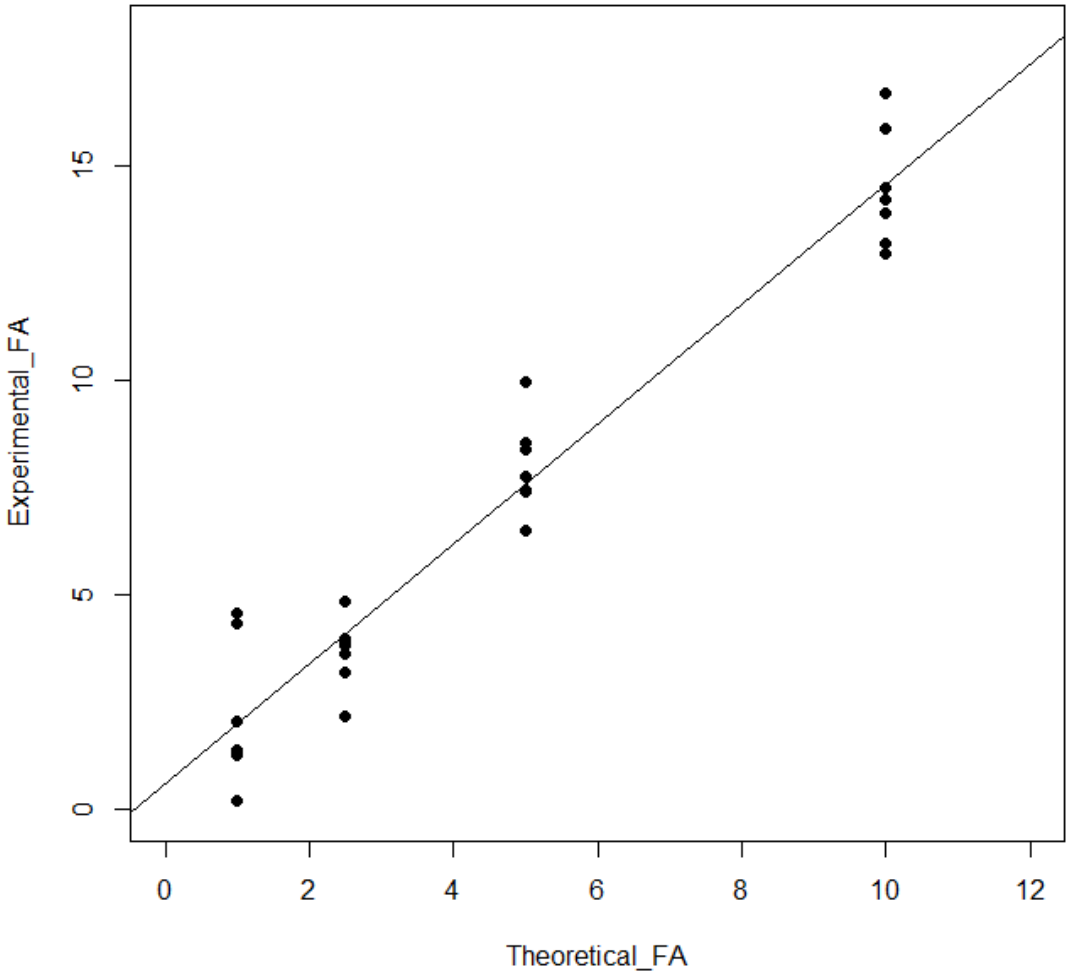
Set	Chr	Start Position	End Position	Length (bp)	MuTect Candidate
CL_WGS_set	1	144988011	145038954	50943	16
	4	190540169	190650521	110352	14
	6	57231654	57493331	261677	19
NeuN_WGS_set	1	144989517	145064529	75012	17
	4	190540426	190650468	110042	19
	6	57217859	57494949	277090	32
	6	167636425	167781635	145210	9
CC_WGS_set	1	144995309	145027115	31806	18
	4	190540811	190650384	109573	13
	6	57219213	57457523	238310	24
	6	167607240	167773885	166645	11

表 16. SNP タイピング結果と混合試料における理論的 Alt 割合

					Genotype		Theoretical Alt Fraction			
Chr	**Position**	**dbSNP ID**	**Ref**	**Alt**	**JM1**	**JM2**	**Mix1**	**Mix2.5**	**Mix5**	**Mix10**
2	159363774	rs1125662	G	C	C/C	G/C	99	97.5	95	90
7	18993249	rs11505418	T	C	T/T	T/C	1	2.5	5	10
8	9472445	rs11249930	G	A	A/A	G/A	99	97.5	95	90
10	4433008	rs10904247	T	C	T/C	T/T	49	47.5	45	40
13	33896531	rs3848097	C	T	C/T	T/T	51	52.5	55	60
14	39295782	rs4902177	G	T	G/T	G/G	49	47.5	45	40
18	10551782	rs10207	C	T	C/C	C/T	1	2.5	5	10
1	19945888	rs1770491	A	T	A/A	A/A	0	0	0	0
2	40391789	rs1005213	G	T	G/G	G/G	0	0	0	0
5	161381930	rs115725937	A	G	A/A	A/A	0	0	0	0
7	127583275	rs113453543	A	G	A/A	A/A	0	0	0	0

12 カ所の dnSNP サイトで SNP タイピングをしたところ、7 カ所（上段）で JM1, JM2 のジェノタイプが異なり、体細胞 SNV シミュレーション試料での評価サイトとした。

図 9. TAS による理論的 Alt 割合と実験的 Alt 割合の比較



X 軸が、体細胞変異をシミュレーションした試料の理論的 Alt 割合 (FA: %で表記)であり、Y 軸が実験で得られた Alt 割合である。Alt 割合は、JM2 の混合により JM1 のジェノタイプから偏位した割合（例えば、JM1 のジェノタイプが Ref/Alt で Alt 割合が理論的に 50%であり、JM2 の混合により Alt 割合が 47.5%になった場合、2.5%分を偏位とみなす）で示した。両者は高い相関（Pearson's r = 0.969, $p < 2.2\times10^{-16}$）を示した。理論的 Alt 割合（説明変数 x）と実験的 Alt 割合（目的変数 y）は、y = 1.39x + 0.62 の関係となり、残差の標準誤差は 1.27 (%)であった。

表 17. CL_WGS_set における体細胞 SNV 候補（マニュアル法）と TAS による確認結果

					AL30_cortex (WGS)			AL30_liver (WGS)		AL30_cortex (TAS)		AL30_liver (TAS)		
Soft	**Chr**	**Position**	**Ref**	**Alt**	**BQ**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**Results**
M	1	211248507	G	A	32	91	7.7	68	5.9	220414	9.657	241711	11.28	Validated
M	5	137203810	G	A	31	104	4.9	105	0	220589	1.897	253286	0.5	Validated
M	6	164440297	G	A	33	123	4.1	99	0	225210	1.71	235770	0.028	Validated
M	13	72252357	G	A	32	105	13.3	102	5.9	212364	14.038	245040	10.32	Validated
M, S	1	47016199	G	A	33	60	8.3	63	0	186823	3.05	212514	4.574	Validated
M, S	4	117374794	C	T	30	102	5.9	103	0	208229	0.736	239646	0.028	Validated
M	2	65847302	T	G	14	60	18.3	60	15	184394	0.043	199760	0.065	Not validated
M	3	105152628	G	A	22	96	8.3	86	5.8	175110	0	197687	0	Not validated
M	6	169095224	G	A	32	87	6.9	83	1.2	73272	0	78342	0	Not validated
M	8	56888381	T	G	10	50	24	47	19.1	110717	0.425	119253	0.438	Not validated*
M	8	124186999	A	C	19	49	12.2	66	9.1	61903	0.023	65458	0.014	Not validated
M	9	32635670	A	G	10	62	24.2	55	18.2	241033	0.052	296048	0.057	Not validated
M	9	132122408	T	G	16	32	31.3	32	15.6	194419	0.003	215334	0.002	Not validated
M	16	31114885	A	C	32	33	9.1	46	0	42877	0.047	55621	0.058	Not validated
M	17	2623125	C	G	29	43	11.6	76	6.6	183433	0.022	214138	0.033	Not validated
M	18	61596240	A	T	31	120	9.2	94	6.4	215192	0.009	252037	0.005	Not validated
S	13	70176335	A	G	(39)	88	7.3	83	0	163944	0.01	254395	0.006	Not validated
M	1	110642378	T	G	18	43	16.3	37	10.8	8	NA	5	NA	Low depth
M	3	142718948	A	C	12	31	25.8	36	13.9	9	NA	7	NA	Low depth
M	5	21542513	G	A	30	131	5.3	108	3.7	2451	NA	2823	NA	Low depth
M	7	28777126	G	C	31	115	5.2	102	2	322	NA	1045	NA	Low depth

M	9	129278657	A	C	11	33	31.3	28	14.3	4849	NA	7530	NA	Low depth
M	16	64119939	A	T	22	67	9	46	4.3	7932	3.392	7877	3.335	Low depth
M	17	21208187	C	T	32	78	7.7	74	5.4	0	NA	0	NA	Low depth
M	21	47062881	A	C	11	25	36	28	3.6	94	NA	115	NA	Low depth
M	1	206662735	C	T	26	73	8.2	90	6.7	72099	0.355	99865	0.413	Not Conclusive (STR)
M	2	76444676	A	G	30	76	6.6	94	2.1	118283	21.898	130210	21.94	Not Conclusive (STR)
M	5	173320041	A	T	30	70	7.1	85	5.9	143011	4.795	157206	4.667	Not Conclusive (STR)
M	14	42831852	C	T	31	80	12.5	83	4.8	98712	5.376	108664	5.654	Not Conclusive (STR)
M	15	65477364	A	C	12	63	25.4	69	11.6	114744	2.147	130242	2.259	Not Conclusive (STR)
M	16	7911650	T	A	32	74	5.4	84	1.2	215069	7.084	239712	6.922	Not Conclusive (STR)
M	18	10196501	C	A	24	69	18.8	100	5	31506	17.115	48916	17.33	Not Conclusive (STR)
M	18	10196502	T	C	25	69	17.4	103	3.9	31209	17.261	48447	17.46	Not Conclusive (STR)
M	18	62002768	A	C	28	67	11.9	76	1.3	96779	12.734	87160	12.72	Not Conclusive (STR)
M	18	62002769	A	C	24	69	8.7	79	0	98500	4.927	88813	4.98	Not Conclusive (STR)
M	1	145004314	C	T	31	155	5.8	158	3.2	0	NA	0	NA	Low depth (SMC)
M	4	190637837	C	T	33	91	7.7	94	3.2	0	NA	0	NA	Low depth (SMC)
M	6	57413745	T	C	29	114	8.8	118	5.9	0	NA	0	NA	Low depth (SMC)

Soft: 使用したソフトウェア。M = Mutect, S = Strelka。

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）。（括弧）で示した値は、Strelka のみで検出された候補の QSS である。

DP: 体細胞 SNV 候補サイトの深度（depth）

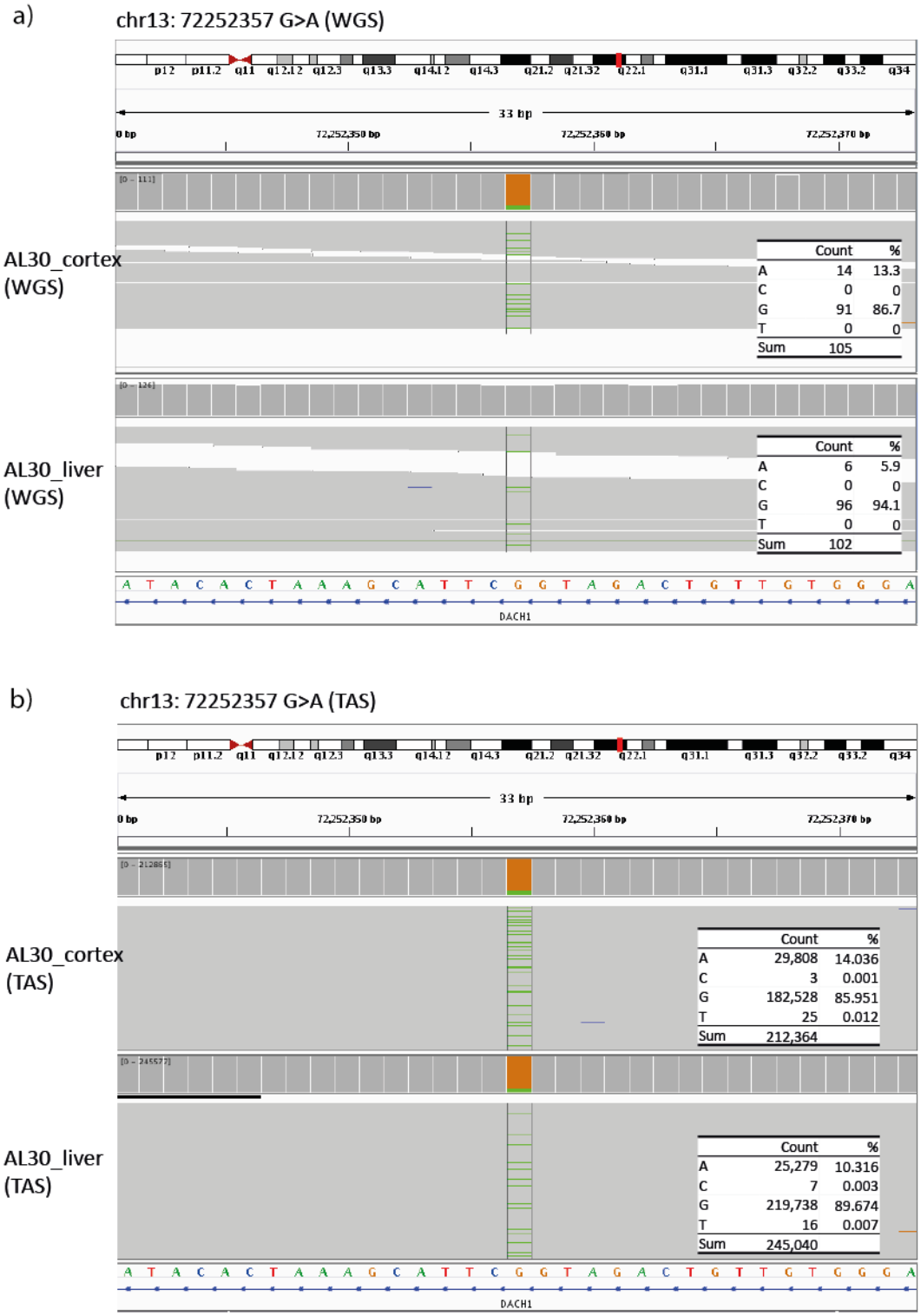
FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)

* BQ, DP, FA は MuTect アウトプットファイル（VCF 形式）の表現と定義を用いた（以下の図表でも同様である）。

Not Conclusive: 解析が難しく評価困難（STR: Short Tandem Repeat）

Low depth: 深度 5000 以下のため評価困難（SMC: suspected multi-copy region（多コピー疑い領域））

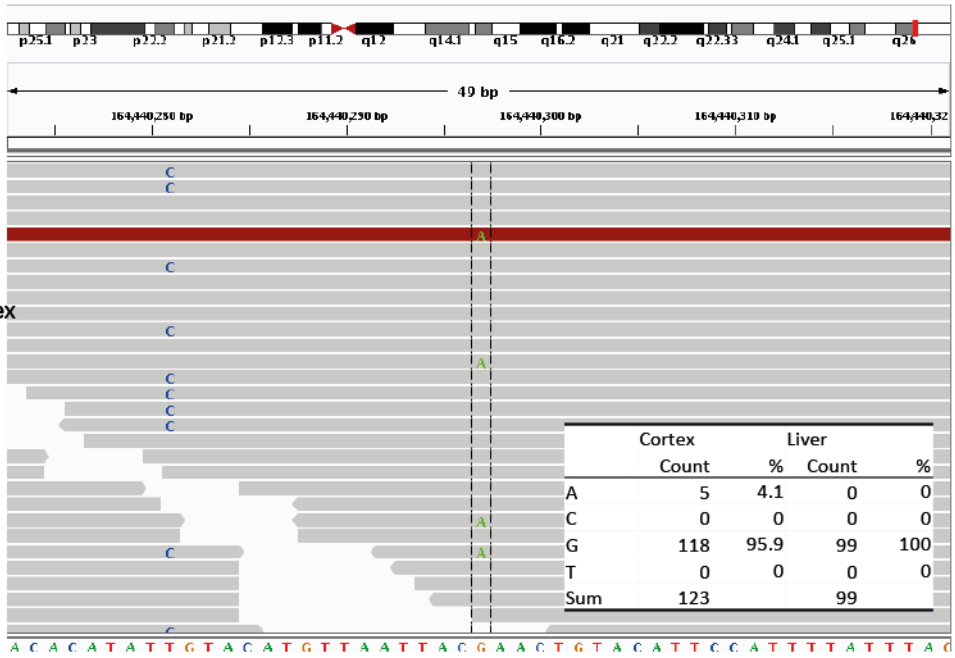
図 10. 体細胞 SNV（候補）の IGV による可視化



c)

chr6: 164440297 G>A (WGS)

AL30_cortex (WGS)

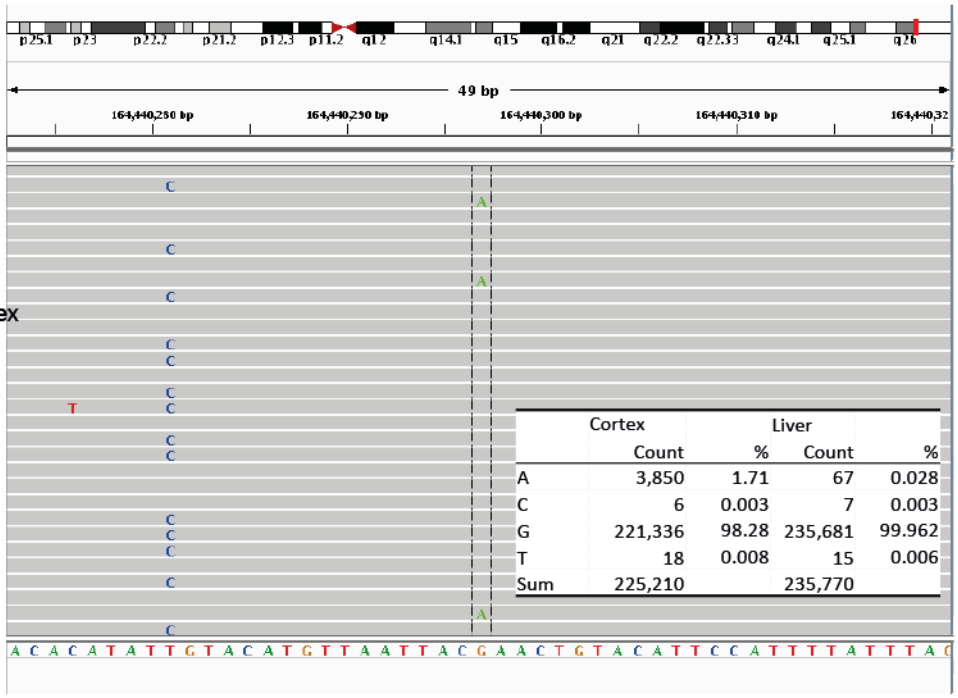


	Cortex		Liver	
	Count	%	Count	%
A	5	4.1	0	0
C	0	0	0	0
G	118	95.9	99	100
T	0	0	0	0
Sum	123		99	

d)

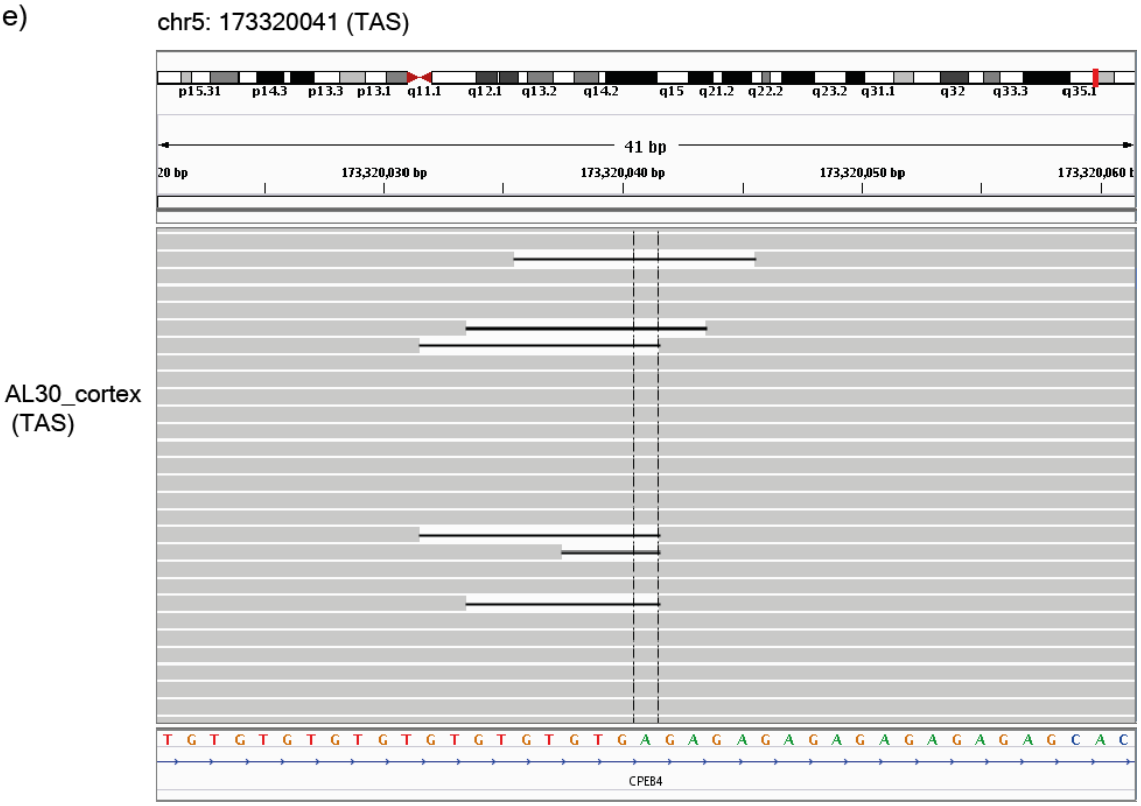
chr6: 164440297 G>A (TAS)

AL30_cortex (TAS)



	Cortex		Liver	
	Count	%	Count	%
A	3,850	1.71	67	0.028
C	6	0.003	7	0.003
G	221,336	98.28	235,681	99.962
T	18	0.008	15	0.006
Sum	225,210		235,770	

e)



a) CL_WGS_set においてバリデーションされた体細胞 SNV（chr13: 72252357 G>A）の WGS データを示した。AL30_cortex, AL30_liver ともに、緑（変異）がまばらに存在し、AL30_cortex で多い。

b) CL_WGS_set においてバリデーションされた体細胞 SNV（chr13: 72252357 G>A）の TAS データを示した。AL30_cortex, AL30_liver ともに、緑（変異）がまばらに存在し、AL30_cortex で多い。WGS より割合の差のあることが、可視化されている。

c) CL_WGS_set においてバリデーションされた SNP 近傍の体細胞 SNV（chr6_164440297 G>A）の WGS データを示した。体細胞 SNV を示唆するベースコールは、SNP 一方のみとリードを共有している。

d) CL_WGS_set においてバリデーションされた SNP 近傍の体細胞 SNV（chr6_164440297 G>A）の TAS データを示した。WGS と同様に、体細胞 SNV を示唆するベースコールは、SNP 一方のみとリードを共有している。この体細胞 SNV は、WGS, TAS ともに AL30_cortex 特異的に存在していた。

e) ショートタンデムリピートの切り替わりサイトにて体細胞 SNV 候補として WGS データから検出された chr5: 173320041 A>T に対し、TAS を行ったところ、当該サイトにおいてリード毎に異なる INDEL が頻出し、解析困難であった。

表 18. CL_WGS_set における体細胞 SNV 候補（操作的方法）と TAS による確認結果

						AL30_cortex (WGS)			AL30_liver (WGS)		AL30_cortex (TAS)		AL30_liver (TAS)		
Soft	**Conf.**	**Chr**	**Position**	**Ref**	**Alt**	**BQ**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**Results**
M, S	**HC**	1	47016199	G	A	33	60	8.3	63	0	186823	3.05	212514	4.574	Validated
M	**HC**	1	211248507	G	A	32	91	7.7	68	5.9	220414	9.657	241711	11.277	Validated
M, S	**HC**	4	117374794	C	T	30	102	5.9	103	0	208229	0.736	239646	0.028	Validated
M	**HC**	5	137203810	G	A	31	104	4.9	105	0	220589	1.897	253286	0.5	Validated
M	**HC**	6	164440297	G	A	33	123	4.1	99	0	225210	1.71	235770	0.028	Validated
M	**HC**	13	72252357	G	A	32	105	13.3	102	5.9	212364	14.038	245040	10.317	Validated
S	**LC**	13	70176335	A	G	(39)	88	7.3	83	0	163944	0.01	254395	0.006	Not validated
M	**LC**	2	76444676	A	G	30	76	6.6	94	2.1	118283	21.898	130210	21.937	Not Conclusive (STR)
M	**LC**	5	173320041	A	T	30	70	7.1	85	5.9	143011	4.795	157206	4.667	Not Conclusive (STR)
M	**LC**	14	42831852	C	T	31	80	12.5	83	4.8	98712	5.376	108664	5.654	Not Conclusive (STR)
M	**LC**	8	106342766	C	A	32	91	6.6	130	4.7	NA	NA	NA	NA	NA

Soft: 使用したソフトウェア。M = Mutect, S = Strelka。

Conf.: 候補の信頼度。HC = High Confidence, LC = Low Confidence

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）。（括弧）で示した値は、Strelka のみで検出された候補の QSS である。

DP: 体細胞 SNV 候補サイトの深度（depth）

FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)

Not Conclusive: 解析が難しく評価困難

STR: Short Tandem Repeat

Low depth: 深度 5000 以下のため評価困難

NA: データが存在しない

表 19. NeuN_WGS_set における体細胞 SNV 候補と TAS による確認結果

						Y8763_NeuN+ (WGS)			Y8763_NeuN- (WGS)			Y8763_liver (WGS)			Y8763_NeuN+ (TAS)		Y8763_NeuN- (TAS)		Y8763_liver (TAS)		
Soft	**Conf.**	**Chr**	**Position**	**Ref**	**Alt**	**BQ**	**DP**	**FA**	**BQ**	**DP**	**FA**	**BQ**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**Results**
M	**HC**	2	80253532	C	T	30	72	11.1	31	81	7.4	30	94	6.4	231229	10.112	247774	10.772	213923	7.9	Validated
M	**HC**	2	206520964	T	G	26	92	6.5	26	89	4.5	27	79	7.6	217913	8.261	234970	6.786	206105	8.921	Validated
M	**HC**	3	192234799	C	T	31	94	5.3	30	96	7.3	30	75	5.3	225529	8.041	249859	7.403	208052	4.223	Validated
M	**HC**	4	138417312	G	A	31	82	4.9	NA	89	2.2	NA	65	0	226823	1.592	252948	1.169	219794	0.01	Validated
M	**HC**	7	147447647	G	C	30	110	5.5	30	119	5	30	97	9.3	245536	8.26	260308	7.072	214337	9.636	Validated
M	**HC**	15	99259587	G	A	30	104	14.4	30	89	4.5	30	89	10.1	203123	8.328	214236	6.578	180034	9.303	Validated
M	**HC**	18	65440162	G	A	31	101	8.9	31	102	4.9	30	86	10.5	201940	7.066	226591	6.649	188342	5.318	Validated
M	**HC**	20	44667433	G	A	32	85	4.7	NA	108	0.9	30	88	6.8	206561	1.265	228462	1.461	198882	3.818	Validated
M	**HC**	20	45395811	G	T	30	79	7.6	30	85	4.7	NA	85	1.2	212954	5.484	246753	4.517	213072	1.006	Validated
M	**HC**	22	41257815	G	A	29	81	3.7	30	83	8.4	30	59	5.1	252768	2.718	257417	4.113	231004	3.629	Validated
M	**HC**	22	45393311	G	A	30	74	6.8	30	75	6.7	28	73	9.6	197269	7.062	201880	7.324	170990	7.172	Validated
M	**LC**	21	23955109	T	C	25	99	8.1	26	115	7.8	26	130	16.9	225810	7.052	259748	7.484	223402	9.856	Validated
M	**HC**	2	20562894	A	G	NA	72	5.6	22	89	7.9	NA	66	4.5	195811	0.006	207708	0.012	166917	0.007	Not validated
M	**HC**	2	30504790	G	A	22	77	9.1	18	103	6.8	NA	77	5.2	222414	0.012	249816	0.012	192969	0.013	Not validated
M	**HC**	2	144010826	T	G	27	110	15.5	25	105	19	22	97	10.3	222091	0.005	243980	0.005	204759	0.007	Not validated
M	**HC**	11	14952526	C	A	NA	80	2.5	27	81	6.3	NA	52	1.9	204616	0.008	229411	0.014	192445	0.006	Not validated
M	**LC**	4	40343190	T	G	24	46	10.9	NA	59	10.2	NA	54	3.7	214106	0	247881	0	214031	0	Not validated
M	**LC**	11	117661790	T	G	22	64	9.4	NA	84	8.3	NA	89	5.6	220687	0.001	246513	0.001	212975	0.001	Not validated
M	**LC**	17	72282109	T	G	29	48	8.3	NA	74	0	NA	69	0	144945	0.009	135668	0.015	133667	0.011	Not validated
S	**HC**	4	166316873	T	A	(29)	96	4.2	NA	99	0	NA	103	0	225991	0.005	236345	0.007	210033	0.006	Not validated
M	**LC**	7	141447818	T	A	NA	75	1.3	29	84	6	NA	83	1.2	12323	1.485	14583	1.694	9527	1.574	NC (poly-A)

M	**LC**	11	112875224	T	A	29	94	5.3	28	88	3.4	NA	86	2.3	103680	2.647	107690	2.711	83334	2.442	NC (poly-A)
M	**LC**	22	48883859	A	T	24	40	15	NA	60	0	NA	48	0	42763	3.038	46727	2.964	27068	2.797	NC (poly-A)
M	**HC**	20	59904421	G	A	29	194	8.2	28	315	9.5	25	247	8.9	NA	NA	NA	NA	NA	NA	No Primers

Soft: 使用したソフトウェア。M = Mutect, S = Strelka。

Conf.: 候補の信頼度。HC = High Confidence, LC = Low Confidence

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）。NA は当該試料で体細胞変異候補が検出されていないことを示す。BQ の数値表示があるものは、体細胞変異候補が検出されていることを示す。（括弧）で示した値は、Strelka のみで検出された候補の QSS である。

DP: 体細胞 SNV 候補サイトの深度（depth）

FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)

NC: 解析が難しく評価困難 (Not Conclusive)

NA: データが存在しない

表 20. CC_WGS_set における体細胞 SNV 候補の同定

MuTect 解析	**CC_WGS_set (1st)**		**MuTect 解析**	**CC_WGS_set (2nd)**	
解析対象試料	S6_cortex	S6_cerebellum	**解析対象試料**	S6_cortex	S6_cerebellum
比較対照試料	S6_cerebellum	S6_cortex	**比較対照試料**	S6_cerebellum	S6_cortex
MuTect 結果	29197	24188		-	-
多コピー領域除外	1429	1428		-	-
INDEL 除外	1365	1362		-	-
多コピー疑い領域除外	1332	1329		-	-
DP ≥ 50	466	439		-	-
dFA ≥ 0.05	257	229	**dFA < 0.05**	209	210
BQ ≥ 21	157	139	**BQ ≥ 25**	112	108
BLAT < 150	139	120	**BLAT < 150**	59	60
HC	7	6	**HC**	7	4
LC	24	20	**LC**	3	4
Stelka 解析	**CC_WGS_set**				
解析対象試料	S6_cortex	S6_cerebellum			
比較対照試料	S6_cerebellum	S6_cortex			
Strelka 結果	1627	724			
多コピー領域除外	73	42			
INDEL 除外	61	32			
多コピー疑い領域除外	-	-			
QSS ≥ 21, DP ≥ 50	22	7			
BLAT < 150					
HC	1	1			
LC	0	1			

INDEL: 挿入・欠失（insertion/deletion）

DP: 体細胞 SNV 候補サイトの深度（depth）

dFA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)の差

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）

HC: High Confidence

LC: Low Confidence

表 21. CC_WGS_set（1 回目）における体細胞 SNV 候補と TAS による確認結果

							S6_cortex (WGS)			S6_cerebellum (WGS)			S6_cortex (TAS)		S6_cerebellum (TAS)		
Soft	**Conf.**	**SeqContext**	**Chr**	**Position**	**Ref**	**Alt**	**BQ**	**DP**	**FA**	**BQ**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**Results**
M	**HC**	Good	1	47281838	G	C	30	83	14.5	NA	72	7	184920	12.376	137318	9.528	Validated
M	**HC**	Good	1	239023569	G	T	NA	87	6.9	29	79	17.7	179001	8.648	140954	12.313	Validated
M, S	**HC**	Good	2	102612321	C	T	NA	90	0	25	76	9.2	194985	0.022	148125	1.772	Validated
M	**HC**	Good	2	202878979	G	T	NA	100	2	28	90	8.9	223533	4.857	163338	5.422	Validated
M	**HC**	Good	3	29527382	C	T	31	104	4.8	NA	77	0	233578	1.51	180110	0.16	Validated
M	**HC**	Good	8	109128812	A	T	NA	74	5.4	30	83	10.8	167442	10.276	214726	13.863	Validated
M, S	**HC**	Good	18	76867501	C	T	26	87	6.9	NA	74	0	163264	2.771	216981	1.383	Validated
M	**HC**	Good	3	125258109	A	G	NA	53	5.7	30	52	11.5	217388	48.786	169818	50.94	Germline SNP
M	**HC**	Good	2	20562894	A	G	28	72	12.5	NA	82	7.3	172724	0.008	122939	0.01	Not validated
M	**HC**	Good	9	109101822	A	G	NA	93	1.1	25	90	6.7	40811	0.566	53797	0.625	Not validated*
M	**HC**	Good	12	86774226	C	A	22	74	8.2	NA	83	1.2	179386	0.013	233256	0.012	Not validated
M	**HC**	Good	13	51922947	T	C	29	83	6	NA	67	1.5	166792	0.215	200732	0.189	Not validated
M	**HC**	Good	X	29270668	C	T	30	144	18.8	NA	140	7.1	181331	0.007	223209	0.004	Not validated
M	**LC**	Good	1	109167096	C	T	NA	71	1.4	21	66	9.1	128063	0.012	181535	0.01	Not validated
M	**LC**	Good	2	170156074	T	A	NA	80	1.3	28	62	8.1	112076	0.003	137421	0.006	Not validated
M	**LC**	Good	6	67384535	T	A	NA	61	1.7	21	66	11.1	113590	0.01	142294	0.01	Not validated
M	**LC**	Good	9	10743921	G	T	21	52	13.5	NA	60	1.7	104507	0.085	158085	0.049	Not validated
M	**LC**	Good	11	110559709	A	T	NA	79	2.6	23	50	12.5	40375	0.012	24070	0.004	Not validated
M	**LC**	Good	17	25623385	A	T	24	87	7	NA	77	0	118890	0.013	157994	0.02	Not validated
M	**LC**	STR	2	151776679	T	A	29	88	8.1	NA	74	1.4	180900	0.022	127207	0.038	Not validated
M	**LC**	poly-A	3	180147593	T	A	NA	69	2.9	22	78	9.2	221068	0	142750	0.001	Not validated

M	**LC**	poly-A	4	58560503	G	T	NA	72	11.1	23	68	17.6	144703	0.113	93760	0.124	Not validated
M	**LC**	poly-A	4	189822464	C	T	NA	75	2.7	27	78	7.8	204938	0.044	122482	0.042	Not validated
M	**LC**	poly-A	11	118320941	A	T	NA	57	1.8	26	87	6.9	182114	0.178	235444	0.172	Not validated
M	**LC**	poly-A	14	93953919	A	T	NA	88	0	30	87	5.7	133691	0.137	178823	0.15	Not validated
M	**LC**	poly-A	20	46250020	A	T	NA	64	1.6	26	73	6.8	174030	0.051	105129	0.043	Not validated
M	**LC**	STR	1	27190857	C	T	29	61	10	NA	77	3.9	85533	1.966	49413	1.987	Not Conclusive
M	**LC**	STR	5	173320041	A	T	29	68	14.7	NA	73	6.8	88280	13.307	62888	12.145	Not Conclusive
M	**LC**	STR	14	26673636	T	C	26	62	8.3	NA	74	1.4	65036	12.013	43650	11.991	Not Conclusive
M	**LC**	STR	16	65094255	T	G	29	86	7	NA	82	1.2	243362	0.428	160317	0.399	Not Conclusive
M	**LC**	STR	18	26648231	A	T	29	82	13.6	NA	76	6.6	207174	3.916	133327	4.129	Not Conclusive
M	**LC**	poly-A	13	97895034	C	T	21	58	10.5	NA	67	1.5	31899	5.649	59509	4.883	Not Conclusive
M	**LC**	poly-A	21	41284253	A	T	NA	71	1.4	30	66	9.1	143701	13.754	111458	11.686	Not Conclusive

Soft: 使用したソフトウェア。M = Mutect, S = Strelka。

Conf.: 候補の信頼度。HC = High Confidence, LC = Low Confidence

SeqContext: シークエンス文脈。STR は Short Tandem Repeat を意味し、Good は STR や poly-A 領域ではないことを示す。

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）。NA は当該試料で体細胞変異候補が検出されていないことを示す。BQ の数値表示があるものは、体細胞変異候補が検出されていることを示す。

DP: 体細胞 SNV 候補サイトの深度（depth）

FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)

Not Conclusive: 解析が難しく評価困難

NA: データが存在しない

* 0.316%を超えているが、他のベースコールも同程度あり、エラーと判断した。

表 22. CC_WGS_set（2 回目）における体細胞 SNV 候補と TAS による確認結果

						S6_cortex (WGS)			S6_cerebellum (WGS)			S6_cortex (TAS)		S6_cerebellum (TAS)		
Soft	**Conf.**	**Chr**	**Position**	**Ref**	**Alt**	**BQ**	**DP**	**FA**	**BQ**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**Results**
M	**HC**	1	240935724	C	T	30	96	6.3	NA	92	4.3	196405	5.325	199631	5.875	Validated
M	**HC**	2	137814327	C	T	NA	75	4	30	79	8.9	193294	5.419	214767	6.356	Validated
M	**HC**	4	126987958	G	A	NA	101	4	30	68	7.4	226918	3.609	254774	5.273	Validated
M	**HC**	12	43522425	C	A	NA	79	5.1	28	86	8.1	226865	3.89	267615	6.36	Validated
M	**HC**	12	48380047	C	G	28	76	6.6	NA	56	5.4	225344	9.068	243091	13.12	Validated
M	**HC**	18	66437389	C	T	NA	92	7.6	30	102	11.8	213025	9.891	239215	12.954	Validated
M	**HC**	6	170209984	G	C	27	87	6.9	NA	73	5.5	206034	0.001	229696	0.003	Not validated
M	**HC**	7	83812741	C	T	27	81	6.2	NA	63	3.2	227509	0.007	245037	0.005	Not validated
M	**HC**	12	39773148	T	C	26	76	6.7	NA	71	7.4	28936	0	37325	0.011	Not validated
M	**HC**	12	132819457	C	G	27	62	9.7	NA	95	7.4	103457	0.084	120194	0.03	Not validated
M	**HC**	13	74859312	A	T	26	78	6.7	NA	68	3	71977	0.007	80601	0.005	Not validated
M	**HC**	18	28039112	T	G	26	56	9.1	NA	57	11.1	1584	NA	1686	NA	Low depth

Soft: 使用したソフトウェア。M = Mutect, S = Strelka。

Conf.: 候補の信頼度。HC = High Confidence, LC = Low Confidence

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）。NA は当該試料で体細胞変異候補が検出されていないことを示す。BQ の数値表示があるものは、体細胞変異候補が検出されていることを示す。

DP: 体細胞 SNV 候補サイトの深度（depth）

FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)

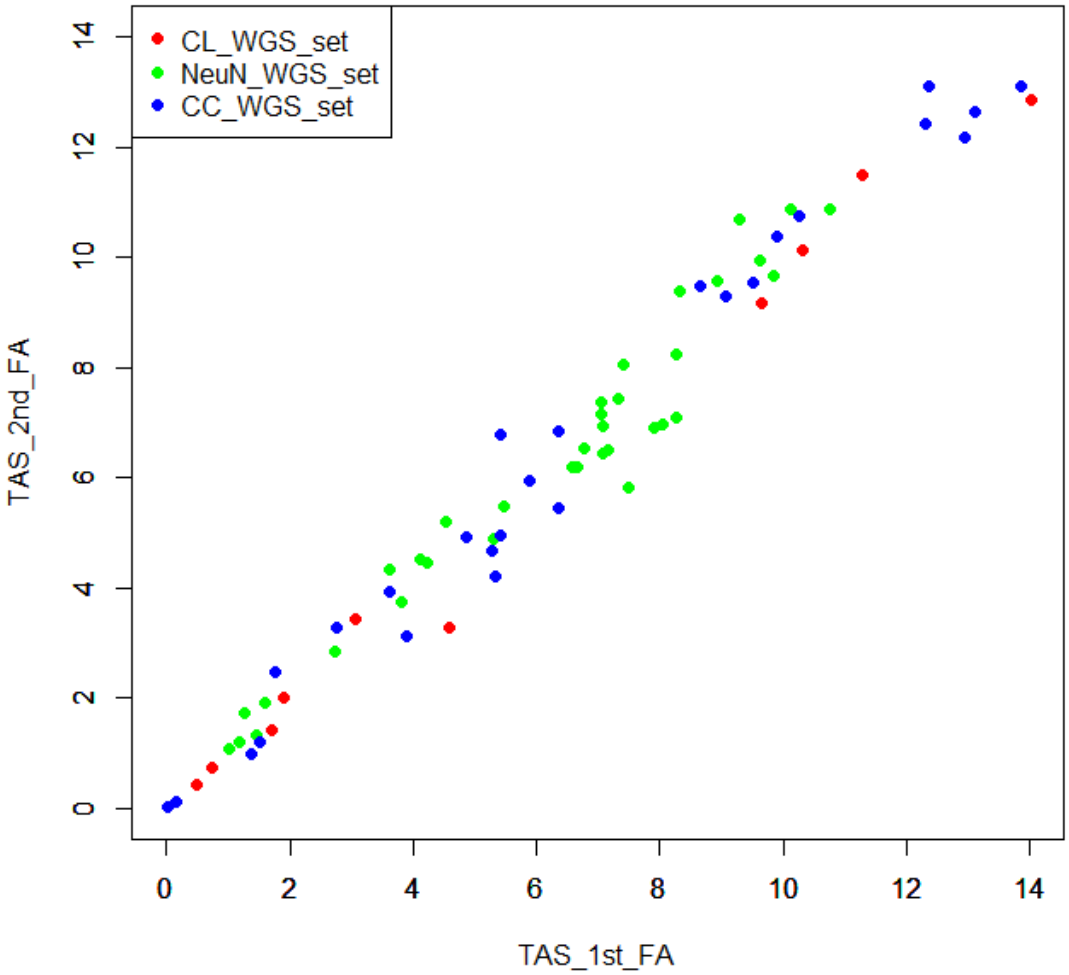
Low depth: 深度 5000 以下のため評価困難

NA: データが存在しない

表 23. 脳神経組織での TAS 再確認結果（Alt 割合の比較）

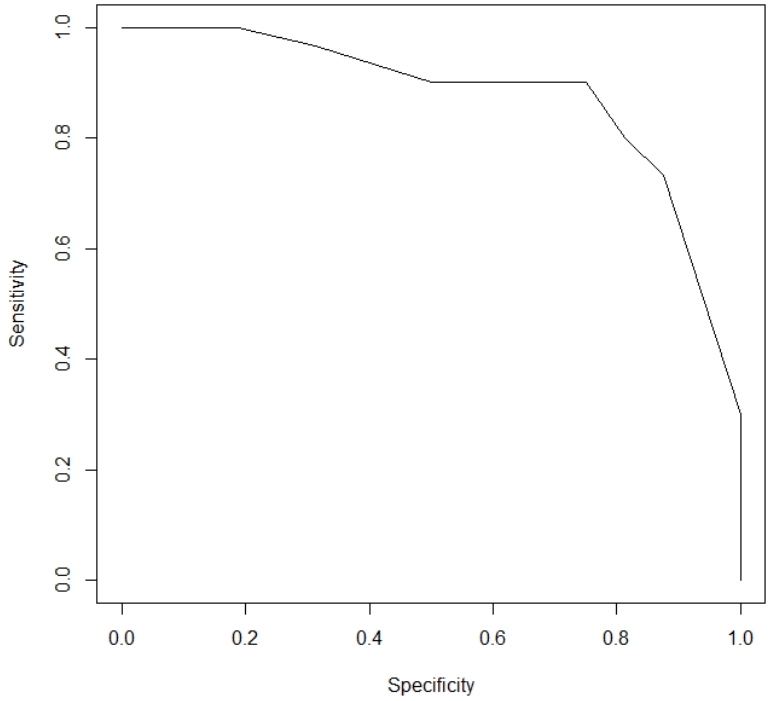
CL_WGS_set				**Cortex**		**Liver**		
Chr	**Position**	**Ref**	**Alt**	**1st**	**2nd**	**1st**	**2nd**	
1	47016199	G	A	3.05	3.425	4.574	3.288	
1	211248507	G	A	9.657	9.182	11.277	11.501	
4	117374794	C	T	0.736	0.742	0.028	0.008	
5	137203810	G	A	1.897	2.004	0.5	0.419	
6	164440297	G	A	1.71	1.418	0.028	0.013	
13	72252357	G	A	14.038	12.861	10.317	10.115	
NeuN_WGS_set				**NeuN+**		**NeuN-**		**Liver**	
Chr	**Position**	**Ref**	**Alt**	**1st**	**2nd**	**1st**	**2nd**	**1st**	**2nd**
2	80253532	C	T	10.112	10.866	10.772	10.888	7.9	6.91
2	206520964	T	G	8.261	8.253	6.786	6.54	8.921	9.564
3	192234799	C	T	8.041	6.975	7.403	8.052	4.223	4.441
4	138417312	G	A	1.592	1.902	1.169	1.18	0.01	0.011
7	147447647	G	C	8.26	7.08	7.072	6.949	9.636	9.933
15	99259587	G	A	8.328	9.373	6.578	6.202	9.303	10.683
18	65440162	G	A	7.066	6.44	6.649	6.198	5.318	4.885
20	44667433	G	A	1.265	1.716	1.461	1.305	3.818	3.725
20	45395811	G	T	5.484	5.492	4.517	5.192	1.006	1.07
21	23955109	T	C	7.052	7.137	7.484	5.831	9.856	9.655
22	41257815	G	A	2.718	2.842	4.113	4.528	3.629	4.326
22	45393311	G	A	7.062	7.376	7.324	7.445	7.172	6.489
CC_WGS_set				**Cortex**		**Cerebellum**		
Chr	**Position**	**Ref**	**Alt**	**1st**	**2nd**	**1st**	**2nd**	
1	47281838	G	C	12.376	13.099	9.528	9.533	
1	239023569	G	T	8.648	9.465	12.313	12.419	
1	240935724	C	T	5.325	4.192	5.875	5.936	
2	102612321	C	T	0.022	0.016	1.772	2.478	
2	137814327	C	T	5.419	4.939	6.356	5.442	
2	202878979	G	T	4.857	4.909	5.422	6.775	
3	29527382	C	T	1.51	1.209	0.16	0.115	
4	126987958	G	A	3.609	3.925	5.273	4.656	
8	109128812	A	T	10.276	10.739	13.863	13.12	
12	43522425	C	A	3.89	3.117	6.36	6.834	
12	48380047	C	G	9.068	9.294	13.12	12.632	
18	66437389	C	T	9.891	10.369	12.954	12.193	
18	76867501	C	T	2.771	3.282	1.383	0.991	

図 11. 脳神経組織でのターゲットアンプリコンシークエンス再確認結果（Alt 割合の比較）



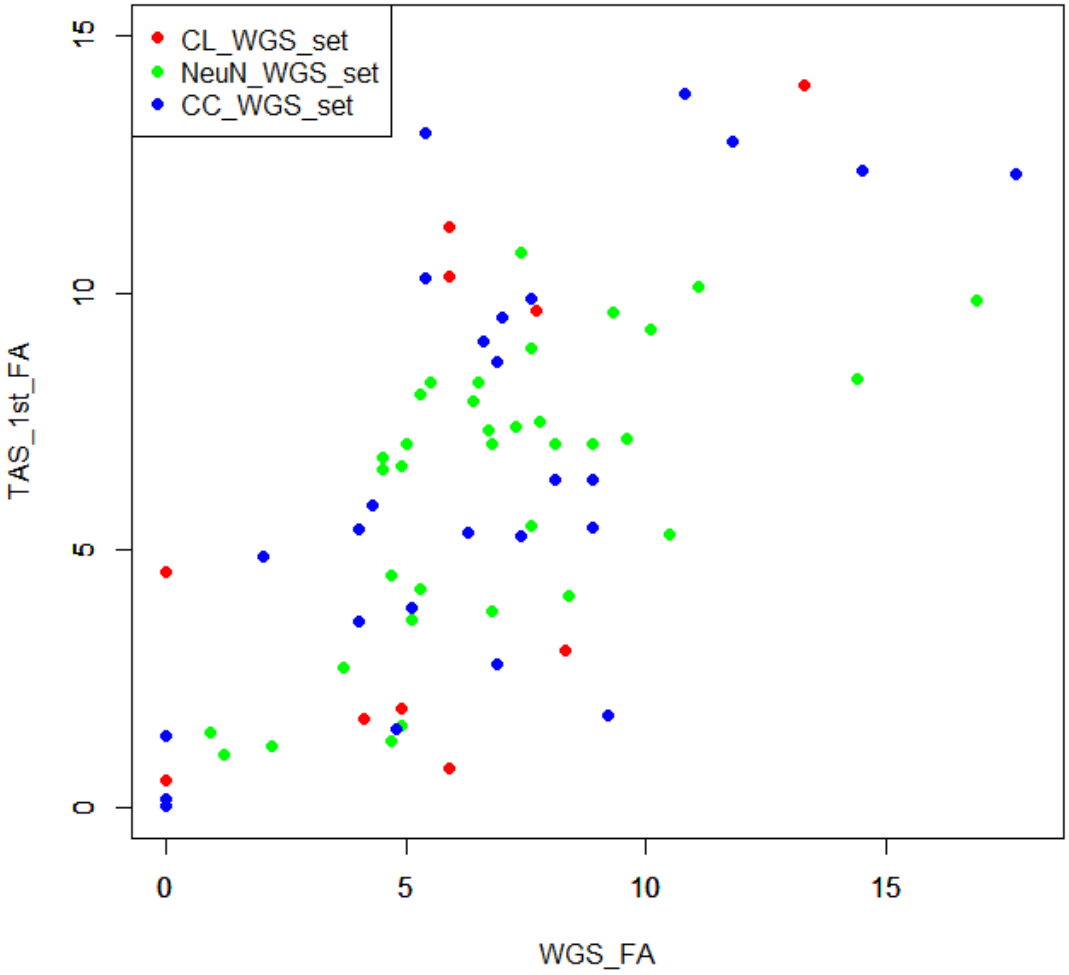
脳神経組織に対するターゲットアンプリコンシークエンス（TAS）1 回目と、ゲノム DNA からの独立した PCR 実験から行った 2 回目の TAS 結果を比較した。両者は高い相関（Pearson’s r = 0.987, $p < 2.2\times10^{-16}$）を示した。X 軸は 1 回目の TAS での Alt 割合、Y 軸は 2 回目の TAS での Alt 割合を示し、%で表記した。

図 12. 脳神経組織 HC 候補のベースクオリティによる ROC カーブ



脳神経組織 3 セット（CL_WGS_set、NeuN_WGS_set、CC_WGS_set）で検出された HC 候補に対し、ベースクオリティでのカットオフ値とバリデーション結果の関係（感度・特異度の）を示した（AUC = 0.876）。

図 13. WGS における Alt 割合と TAS における Alt 割合の相関



バリデーションされた体細胞 SNV 計 31 カ所について、WGS での Alt 割合（X 軸）と 1 回目のターゲットアンプリコンシークエンス（TAS）での Alt 割合（Y 軸）を比較したところ、緩やかな相関（Pearson's r = 0.696, $p = 5.87\times10^{-12}$）が得られた。

表 24. 脳神経組織における体細胞 SNV の機能推定

CL_WGS_set

Chr	Position	dbSNP_ID	Ref	Alt	Gene	SO	Impact	Transcript	HGVS.p
1	47016199	.	G	A	KNCN	sequence_feature	LOW	protein_coding	
1	211248507	.	G	A	KCNH1	sequence_feature	LOW	protein_coding	
4	117374794	.	C	T		intergenic_region	MODIFIER		
5	137203810	.	G	A	MYOT	intron_variant	MODIFIER	protein_coding	
6	164440297	rs140408129	G	A		intergenic_region	MODIFIER		
13	72252357	.	G	A	DACH1	sequence_feature	LOW	protein_coding	

NeuN_WGS_set

Chr	Position	dbSNP_ID	Ref	Alt	Gene	SO	Impact	Transcript	HGVS.p
2	80253532	.	C	T	AC016716.2	downstream_gene_variant	MODIFIER	antisense	
2	206520964	.	T	G		intergenic_region	MODIFIER		
3	192234799	.	C	T	FGF12-AS3	splice_region_variant non_coding_exon_variant	LOW	antisense	
4	138417312	.	G	A		intergenic_region	MODIFIER		
7	147447647	.	G	C	CNTNAP2	sequence_feature	LOW	protein_coding	
15	99259587	.	G	A	IGF1R	intron_variant	MODIFIER	protein_coding	
18	65440162	.	G	A	RP11-638L3.1	intron_variant	MODIFIER	lincRNA	
20	44667433	rs375582719	G	A	SLC12A5	upstream_gene_variant	MODIFIER	processed_transcript	
20	45395811	.	G	T		intergenic_region	MODIFIER		
21	23955109	.	T	C		intergenic_region	MODIFIER		
22	41257815	rs149771105	G	A	DNAJB7	missense_variant	MODERATE	protein_coding	p.Arg62Trp
22	45393311	.	G	A	RP4-753M9.1	downstream_gene_variant	MODIFIER	sense_intronic	

CC_WGS_set									
Chr	**Position**	**dbSNP_ID**	**Ref**	**Alt**	**Gene**	**SO**	**Impact**	**Transcript**	**HGVS.p**
1	47281838	.	G	C	CYP4B1	upstream_gene_variant	MODIFIER	processed_transcript	
1	239023569	.	G	T		intergenic_region	MODIFIER		
1	240935724	.	C	T	PRKRIRP8	upstream_gene_variant	MODIFIER	processed_pseudogene	
2	102612321	rs145338146	C	T	IL1R2	upstream_gene_variant	MODIFIER	protein_coding	
2	137814327	.	C	T	THSD7B	synonymous_variant	LOW	protein_coding	p.Cys159Cys
2	202878979	.	G	T		intergenic_region	MODIFIER		
3	29527382	.	C	T	RBMS3	sequence_feature	LOW	protein_coding	
4	126987958	.	G	A	RP11-318I4.1	upstream_gene_variant	MODIFIER	lincRNA	
8	109128812	.	A	T		intergenic_region	MODIFIER		
12	43522425	.	C	A		intergenic_region	MODIFIER		
12	48380047	rs143809736	C	G	COL2A1	sequence_feature	LOW	protein_coding	
18	66437389	.	C	T	CCDC102B	intron_variant	MODIFIER	protein_coding	
18	76867501	.	C	T	ATP9B	sequence_feature	LOW	protein_coding	

SO: sequence ontology (http://www.sequenceontology.org/)にて定義されたカテゴリー

HGVs.p: Human Genome Variation Society の推奨するアミノ酸配列変化の表記法による

表 25. 脳神経組織で検出された体細胞 SNV の Gene ontology 解析

Cellular Component					
GO ID	**GO**	**p 値**	**FDR**	**解析対象群の遺伝子数**	**GO 内の全遺伝子数**
GO:0032589	neuron projection membrane	5.42×10^{-4}	3.15×10^{-2}	2	47
GO:0005585	collagen type II trimer	7.43×10^{-4}	3.15×10^{-2}	1	1
GO:0008076	voltage-gated potassium channel complex	1.41×10^{-3}	3.15×10^{-2}	2	76
GO:0034705	potassium channel complex	1.45×10^{-3}	3.15×10^{-2}	2	77
GO:0043204	perikaryon	1.89×10^{-3}	3.28×10^{-2}	2	88
GO:0032437	cuticular plate	2.97×10^{-3}	4.30×10^{-2}	1	4
GO:0031256	leading edge membrane	4.37×10^{-3}	4.62×10^{-2}	2	135
GO:0060091	kinocilium	4.45×10^{-3}	4.62×10^{-2}	1	6
GO:0043025	neuronal cell body	4.78×10^{-3}	4.62×10^{-2}	3	478
GO:0034703	cation channel complex	5.79×10^{-3}	4.83×10^{-2}	2	156
GO:0044297	cell body	6.57×10^{-3}	4.83×10^{-2}	3	536
GO:0044224	juxtaparanode region of axon	6.66×10^{-3}	4.83×10^{-2}	1	9

GO: Gene Ontology (http://geneontology.org/)にて定義されたカテゴリー

FDR: False Discovery Rate

表 26. MZ_WGS_set における体細胞 SNV 候補の同定

MuTect 解析	**MZ_WGS_set (非多コピー領域)**		**MuTect 解析**	**MZ_WGS_set (多コピー領域)**	
解析対象試料	SBT1*	SBT4	**解析対象試料**	SBT1*	SBT4
比較対照試料	SBT4	SBT1*	**比較対照試料**	SBT4	SBT1*
MuTect 結果	4567	6184	**MuTect 結果**	1816	2469
多コピー領域除外	359	432	**多コピー領域上**	1558	2169
INDEL 除外	340	399	**INDEL 除外**	489	1918
対照 FA=0	144	182	**対照 FA=0**	103	516
BQ ≥ 20, DP ≥ 30	56	68	**BQ ≥ 20, DP ≥ 30**	79	353
BLAT < 150	45	62	**BLAT < 150**	50	191
HC	2	2	**HC**	3	10
LC	10	14	**LC**	4	27
Stelka 解析	**MZ_WGS_set (非多コピー領域)**				
解析対象試料	SBT1*	SBT4			
比較対照試料	SBT4	SBT1*			
Strelka 結果	570	1212			
多コピー領域除外	36	85			
INDEL 除外	24	75			
対照 FA=0	12	38			
QSS ≥ 20, DP ≥ 30	5	26			
BLAT < 150					
HC	0	0			
LC	2	0			

INDEL: 挿入・欠失（insertion/deletion）

FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）

DP: 体細胞 SNV 候補サイトの深度（depth）

HC: High Confidence

LC: Low Confidence

* 罹患者

表 27. MZ_WGS_set における体細胞 SNV 候補と TAS による確認結果

							SBT1* (WGS)			SBT4 (WGS)			SBT1* (TAS)		SBT4 (TAS)		
Soft	**Region**	**Conf.**	**Chr**	**Position**	**Ref**	**Alt**	**BQ**	**DP**	**FA**	**BQ**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**Results**
M	**NonRepeat**	**HC**	2	169920685	C	A	22	64	9.5	NA	54	0	7702	0.013	8550	0	Not validated
M	**NonRepeat**	**HC**	8	53091419	G	A	NA	46	0	30	59	6.8	237165	0.022	269086	0.018	Not validated
M	**NonRepeat**	**HC**	12	121114214	A	G	25	45	11.4	NA	64	0	7709	0.013	8392	0.048	Not validated
M	**NonRepeat**	**HC**	17	53577074	A	G	NA	60	0	29	68	7.5	234582	0.051	261399	0.052	Not validated
M	**Repeat**	**HC**	20	58016089	G	A	31	53	7.5	NA	67	0	221571	1.986	231893	2.134	Homologous?
M	**Repeat**	**HC**	4	27468649	G	A	NA	37	0	32	39	7.7	201923	0.007	206862	0.007	Not validated
M	**Repeat**	**HC**	5	175021794	C	T	NA	53	0	30	58	6.9	133628	0.015	180230	0.016	Not validated
M	**Repeat**	**HC**	7	67737975	G	T	31	51	7.8	NA	55	0	211020	0.01	236472	0.008	Not validated
M	**Repeat**	**HC**	8	24548901	C	A	NA	36	0	30	41	12.2	211579	0.01	244288	0.011	Not validated
M	**Repeat**	**HC**	15	50631622	G	T	NA	49	0	27	46	8.7	99316	0.014	113684	0.017	Not validated
M	**Repeat**	**HC**	15	65402075	G	T	NA	53	0	32	60	6.7	245329	0.006	269035	0.006	Not validated
M	**Repeat**	**HC**	17	67718603	A	G	NA	56	0	25	61	9.8	84696	0.011	99804	0.015	Not validated
M	**Repeat**	**HC**	5	543501	G	A	NA	41	0	29	40	10	1002	NA	1020	NA	Low depth
M	**Repeat**	**HC**	10	5895525	C	A	31	31	9.7	NA	31	0	285	NA	254	NA	Low depth
M	**Repeat**	**HC**	16	85394970	G	T	NA	46	0	22	44	11.4	54	NA	77	NA	Low depth
M	**Repeat**	**HC**	9	138276825	T	C	NA	50	0	27	48	8.7	NA	NA	NA	NA	No Primers
M	**Repeat**	**HC**	15	34602425	A	G	NA	49	0	23	45	13.3	NA	NA	NA	NA	No Primers

Soft: 使用したソフトウェア。M = Mutect, S = Strelka。

Conf.: 候補の信頼度。HC = High Confidence, LC = Low Confidence

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）。NA は当該試料で体細胞変異候補が検出されていないことを示す。BQ の数値表示があるものは、体細胞変異候補が検出されていることを示す。

DP: 体細胞 SNV 候補サイトの深度（depth）
FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)
Low depth: 深度 5000 以下のため評価困難
NA: データが存在しない
* 罹患者

表 28. MZ_Exome_set における体細胞 SNV 候補の同定

MuTect 解析	**MZ_Exome_set**							
解析対象試料	FT11*	FT12	JT11*	JT12	TT21*	TT22	TT11*	TT12
比較対照試料	FT12	FT11*	JT12	JT11*	TT22	TT21*	TT12	TT11*
MuTect 結果	70	101	108	116	90	99	94	110
多コピー領域除外	-	-	-	-	-	-	-	-
INDEL 除外	60	96	102	111	83	92	89	104
多コピー疑い領域除外	-	-	-	-	-	-	-	-
対照 FA=0	38	66	64	70	58	68	53	63
BQ ≥ 20, DP ≥ 30	19	25	16	17	18	17	31	21
BLAT < 160	12	12	4	5	4	0	10	8
HC	4	6	1	2	3	0	6	6
LC	2	1	2	1	0	0	2	0
Stelka 解析	**MZ_Exome_set**							
解析対象試料	FT11*	FT12	JT11*	JT12	TT21*	TT22	TT11*	TT12
比較対照試料	FT12	FT11*	JT12	JT11*	TT22	TT21*	TT12	TT11*
Strelka 結果	12	17	33	60	31	31	40	54
多コピー領域除外	-	-	-	-	-	-	-	-
INDEL 除外	10	17	31	57	31	31	39	45
多コピー疑い領域除外	-	-	-	-	-	-	-	-
対照 FA=0	9	11	23	45	24	23	33	32
QSS ≥ 20, DP ≥ 30	3	2	5	8	0	4	8	4
BLAT < 160								
HC	1	1	3	5	0	2	6	4
LC	1	0	1	0	0	0	2	0

INDEL: 挿入・欠失（insertion/deletion）

FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）

DP: 体細胞 SNV 候補サイトの深度（depth）

HC: High Confidence

LC: Low Confidence

* 罹患者

表 29. MZ_Exome_set における体細胞 SNV 候補と TAS による確認結果

							Subject (WGS)			Control (WGS)		Subject (TAS)		Control (TAS)		
Subject	**Soft**	**Conf.**	**Chr**	**Position**	**Ref**	**Alt**	**BQ**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**DP**	**FA**	**Results**
FT11*	M	HC	1	205242142	C	A	32	30	10	35	0	175596	0.008	187992	0.007	Not validated
FT11*	M	HC	9	113341741	C	T	33	46	6.5	40	0	143991	0.044	151078	0.037	Not validated
FT11*	S	HC	9	116779003	C	T	(35)	141	4.3	151	0	180142	0.024	195770	0.021	Not validated
FT11*	M	HC	19	49621802	C	T	34	45	6.7	63	0	181984	0.026	183934	0.012	Not validated
FT11*	M	HC	21	31864375	G	A	33	47	6.4	48	0	184995	0.012	177311	0.017	Not validated
FT12	M	HC	7	142637430	C	A	33	40	7.5	40	0	168609	0.007	165539	0.007	Not validated
FT12	M	HC	9	117014809	G	A	32	45	6.7	38	0	181804	0.023	160056	0.008	Not validated
FT12	M, S	HC	9	139849022	C	G	33	78	7.8	75	0	44171	0.113	42346	0.085	Not validated
FT12	M	HC	14	65560458	G	A	33	41	7.3	45	0	187233	0.054	180378	0.062	Not validated
FT12	M	HC	16	75728247	C	T	33	50	6	47	0	177706	0.011	160548	0.016	Not validated
FT12	M	HC	20	58411402	C	T	33	33	9.1	31	0	177480	0.026	175440	0.017	Not validated
JT11*	M	HC	2	54871421	A	G	36	39	7.7	66	0	170480	0.008	148382	0.008	Not validated
JT11*	S	HC	2	175664545	G	A	(22)	46	8.7	73	0	164701	0.044	138271	0.012	Not validated
JT11*	S	HC	7	48413830	G	A	(22)	56	7.1	58	0	187957	0.015	166505	0.019	Not validated
JT11*	S	HC	22	43253229	G	A	(22)	43	9.3	48	0	119623	0.011	100607	0.01	Not validated
JT12	S	HC	2	114718299	G	A	(23)	53	7.5	56	0	189889	0.008	156027	0.006	Not validated
JT12	S	HC	3	42252628	T	C	(20)	54	7.4	41	0	186149	0.01	158963	0.007	Not validated
JT12	M	HC	7	29440490	G	A	32	60	6.7	59	0	175586	0.017	158943	0.011	Not validated
JT12	S	HC	12	123752524	G	T	(20)	34	11.8	39	0	180015	0.007	151686	0.005	Not validated
JT12	M	HC	17	27381715	C	A	32	32	9.4	31	0	183143	0.037	154613	0.067	Not validated
JT12	S	HC	19	42352925	G	A	(27)	31	12.9	51	0	172060	0.006	211555	0.01	Not validated

JT12	S	HC	X	37587307	C	A	(23)	166	3.6	119	0	180906	0.003	163703	0.003	Not validated
TT21*	M	HC	1	28607676	G	A	32	30	10	37	0	217106	0.014	215669	0.011	Not validated
TT21*	M	HC	4	113110009	G	A	33	38	7.9	35	0	212589	0.011	222123	0.009	Not validated
TT21*	M	HC	X	41077657	G	A	32	35	8.6	46	0	172232	0.012	169733	0.013	Not validated
TT22	S	HC	1	28833911	C	A	(21)	48	8.3	44	0	208262	0.011	205429	0.009	Not validated
TT22	S	HC	12	54070004	G	A	(20)	114	4.4	122	0	179211	0.012	175962	0.013	Not validated
TT11*	M	HC	7	105641974	G	T	37	93	4.3	112	0	177376	2.417	180806	0.012	Validated
TT11*	M, S	HC	11	72947061	C	T	35	65	9.2	77	0	188628	5.77	186399	0.01	Validated
TT11*	M, S	HC	12	22040794	A	C	31	74	8.1	78	0	179669	7.32	182798	0.007	Validated
TT11*	M	HC	1	52821154	G	A	32	33	9.1	39	0	185818	0.048	180707	0.063	Not validated
TT11*	S	HC	1	78603073	G	A	(22)	53	7.5	86	0	192305	0.014	205610	0.01	Not validated
TT11*	S	HC	10	50960631	T	C	(25)	48	8.3	55	0	196603	0.008	198006	0.006	Not validated
TT11*	M	HC	12	42491817	G	A	34	46	6.5	63	0	212885	0.017	200265	0.017	Not validated
TT11*	S	HC	19	5208010	G	A	(26)	106	4.7	109	0	182979	0.028	178617	0.07	Not validated
TT11*	M, S	HC	19	44739117	C	T	34	73	5.5	73	0	191009	0.027	192628	0.019	Not validated
TT12	M	HC	1	21605869	G	A	33	64	6.3	55	0	191426	3.83	191942	0.014	Validated
TT12	M, S	HC	1	39991592	C	T	33	80	11.3	102	0	179781	6.588	173062	0.006	Validated
TT12	M, S	HC	1	245849059	C	T	34	91	5.5	106	0	207113	1.12	204793	0.012	Validated
TT12	M, S	HC	12	78571018	C	T	31	104	5.8	78	0	201424	3.092	196370	0.016	Validated
TT12	M	HC	1	201010615	G	A	31	40	7.5	44	0	192446	0.006	193493	0.01	Not validated
TT12	M	HC	14	64591770	G	A	34	47	6.4	52	0	185043	0.219	186252	0.213	Not validated
TT12	S	HC	X	50119242	G	A	(20)	41	9.8	40	0	169088	0.005	190112	0.007	Not validated

Soft: 使用したソフトウェア。M = Mutect, S = Strelka。

Conf.: 候補の信頼度。HC = High Confidence, LC = Low Confidence

BQ: 体細胞 SNV を支持するベースコールの平均ベースクオリティ（base quality）。（括弧）で示した値は、Strelka のみで検出された候補の QSS である。
DP: 体細胞 SNV 候補サイトの深度（depth）
FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)
上から解析ペアごとに示した。
* 罹患者

表 30. MZ_Exome_set における体細胞 SNV の機能推定

Subject	Chr	Position	dbSNP_ID	Ref	Alt	FA	Gene	SO	Impact	Transcript	HGVS.p
TT11*	7	105641974	rs377249040	G	T	2.417	CDHR3	synonymous_variant	LOW	protein_coding	p.Ala260Ala
TT11*	11	72947061	.	C	T	5.77	P2RY2	3_prime_UTR_variant	MODIFIER	protein_coding	
TT11*	12	22040794	.	A	C	7.32	ABCC9	missense_variant	MODERATE	protein_coding	p.Leu626Arg
TT12	1	21605869	.	G	A	3.83	ECE1	missense_variant	MODERATE	protein_coding	p.Pro20Leu
TT12	1	39991592	.	C	T	6.588	BMP8A	3_prime_UTR_variant	MODIFIER	protein_coding	
TT12	1	245849059	rs375857224	C	T	1.12	KIF26B	missense_variant	MODERATE	protein_coding	p.Thr925Met
TT12	12	78571018	rs148932437	C	T	3.092	NAV3	missense_variant	MODERATE	protein_coding	p.Pro1741Leu

FA: 全ベースコールに対する、リファレンスと異なるベースコールの割合 (%, Allele fraction of the alternate allele)。比較対照試料での FA は 0%である。

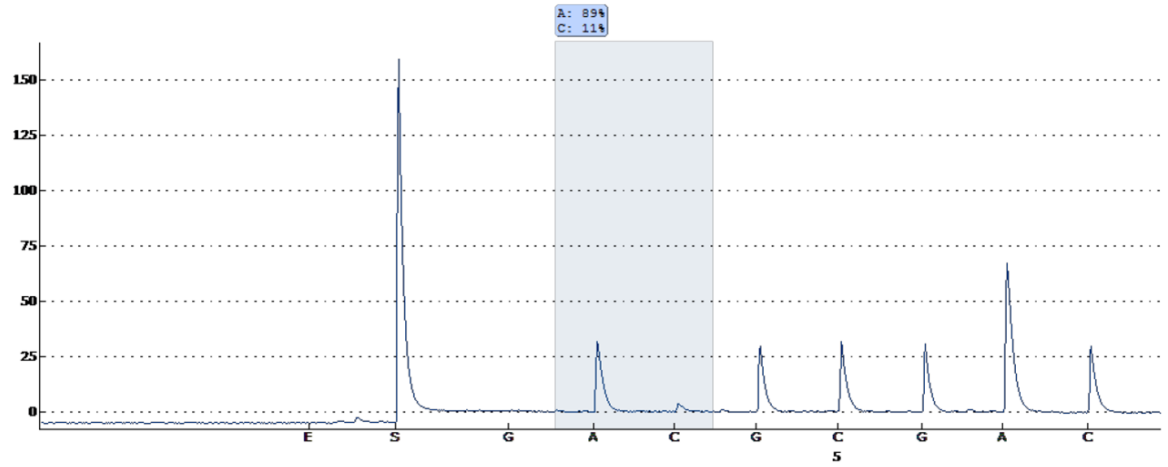
SO: sequence ontology (http://www.sequenceontology.org/)にて定義されたカテゴリー

* 罹患者

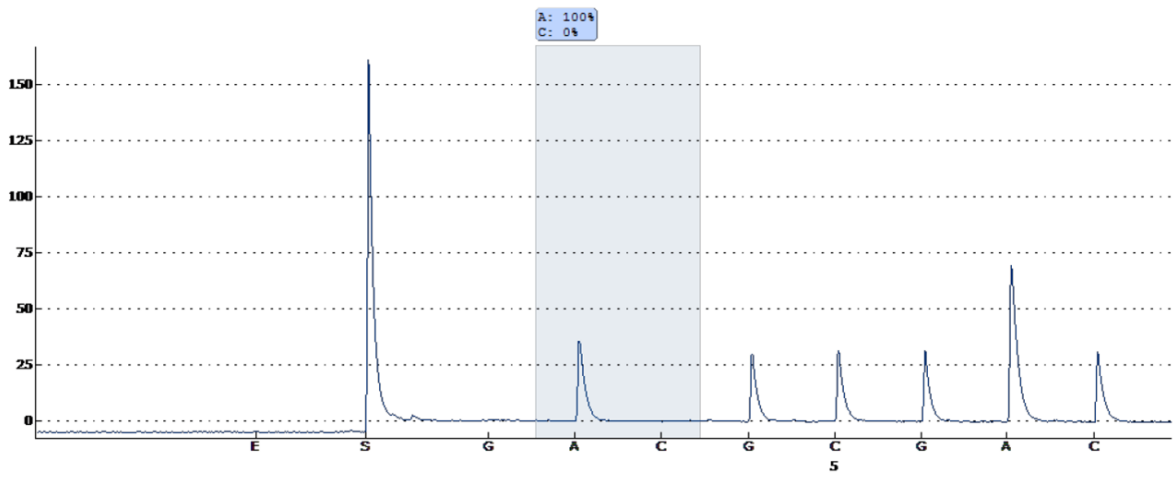
図 14. MZ_Exome_set で検出された体細胞 SNV に対するパイロシークエンス結果

				TT11*		TT12	
Chr	**Position**	**Ref**	**Alt**	**TAS_FA**	**TAS_FA**	**Pyro_FA**	**Pyro_FA**
11	72947061	C	T	5.8%	0	11%	0
12	22040794	A	C	7.3%	0	11%	0
1	39991592	C	T	0	6.6%	0	8%
12	78571018	C	T	0	3.1%	0	3%

TT11*のパイログラム（chr12: 22040794）



TT12 のパイログラム（chr12: 22040794）



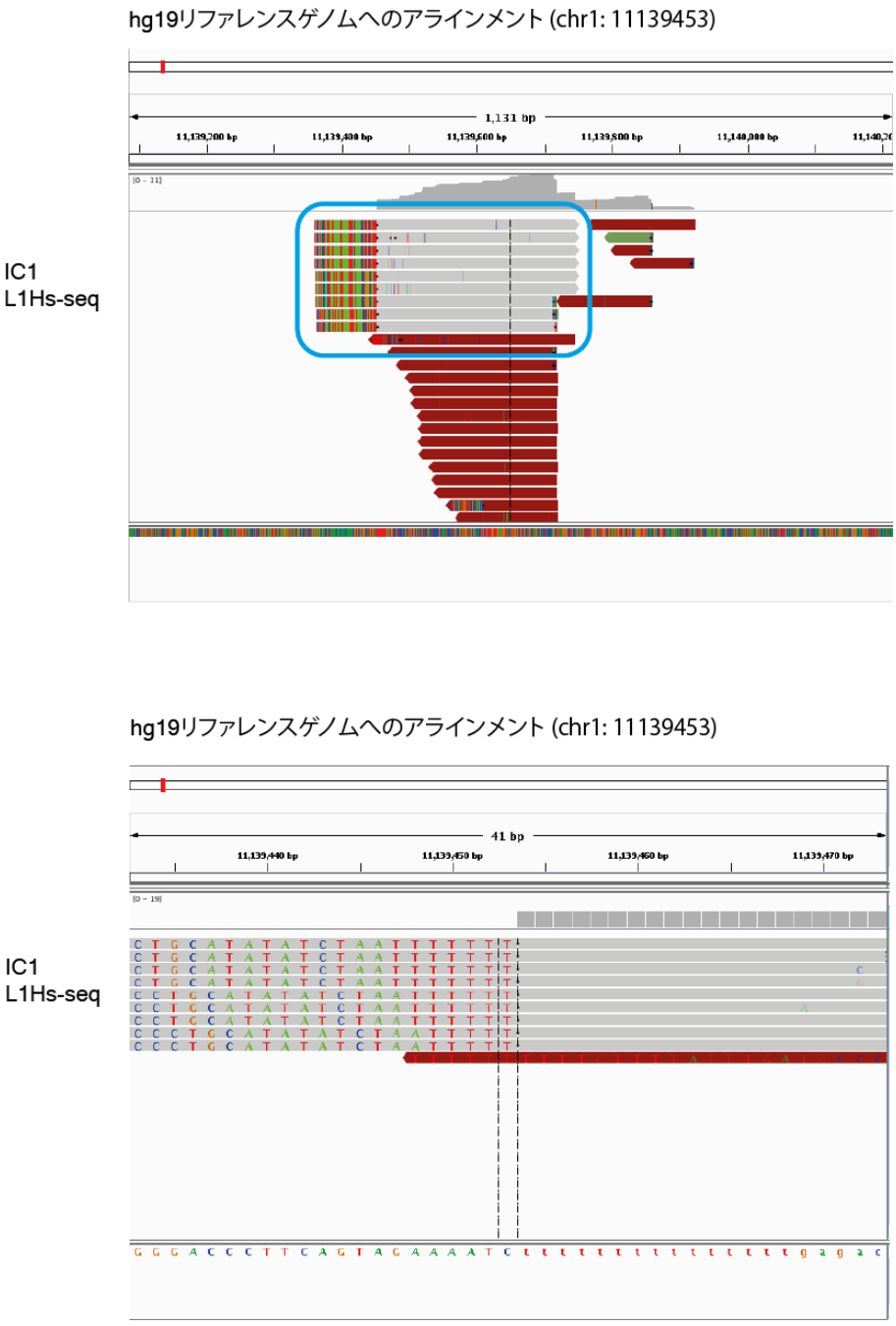
パイロシークエンスで検出された 4 ヶ所について、パイロシークエンスでの Alt 割合（Pyro_FA）と、TAS で計算した Alt 割合（TAS_FA）を表に示した。図に、chr12: 22040794 A>C (*ABCC9*, p.Leu626Arg)におけるパイログラムを示した。罹患者（TT11*）でのみ A>C の一塩基変異を 11%の割合で認めた。

表 31. 改良型 L1Hs-seq によるリファレンス L1Hs 検出感度・非リファレンス検出数・人工遺伝子検出リード数

		IC1	IC2	IC3	S6_cerebellum
リード数		18,847,018	18,774,594	18,249,161	17,194,633
感度（閾値 S）		83.9%	82.4%	82.8%	81.3%
感度（閾値 R）		96.3%	97.4%	97.4%	96.2%
非リファレンス数	**（閾値 S）**	142	92	82	159
	KNR 除く	88	44	31	122
人工遺伝子配列	**50%**	964	NI	NI	NI
検出リード数	**10%**	NI	154	NI	NI
	2.5%	NI	17	NI	NI
	1%	NI	NI	27	NI
	0.5%	NI	16	0	NI
	0.1%	NI	NI	0	NI
pMK-RQ		0	0	0	NI

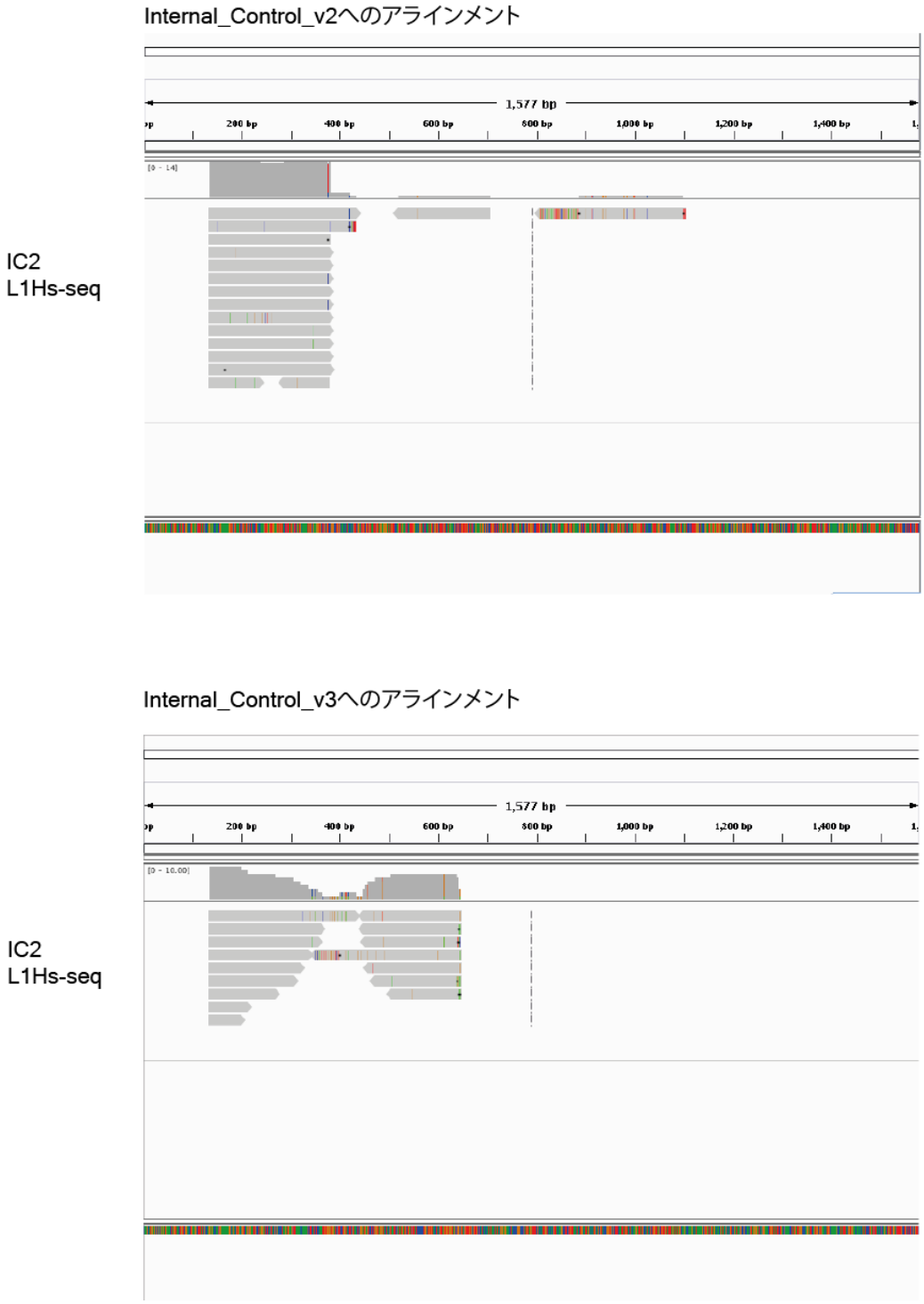
閾値 S: cutadapt において-q 28 –m 30 で解析を行い，マッピングクオリティ 30 以上のリードのブロックがゲノム領域にして 350bp 以上をカバーしている。
閾値 R: cutadapt において-q 20 –m 20 で解析を行い，マッピングクオリティ 0 以上のリードのブロックがゲノム領域にして 100bp 以上をカバーしている。
KNR: 既知の非リファレンス位置での L1Hs 挿入位置 (Known NonReference)
NI: 該当する割合での人工遺伝子の混合はしていない (Not Included)

図 15. 非リファレンス位置での L1Hs 新規挿入の検出とジャンクション配列の決定



試料 IC1 に対して改良型 L1Hs-seq を行い、L1Hs 新規挿入 88 カ所を同定した。上段は chr1: 11139453 における新規挿入を IGV に可視化したもので、下段は挿入位置をズームしたものである。青四角で囲った部分のリードは、スティッチングしたリードで約 500 bp のリードも認めた。スティッチングにより、L1Hs 配列と挿入ゲノム領域配列のジャンクションサイトが、一塩基レベルの解像度で検出することができた。挿入位置では poly-A の長さの多様性を認めている。

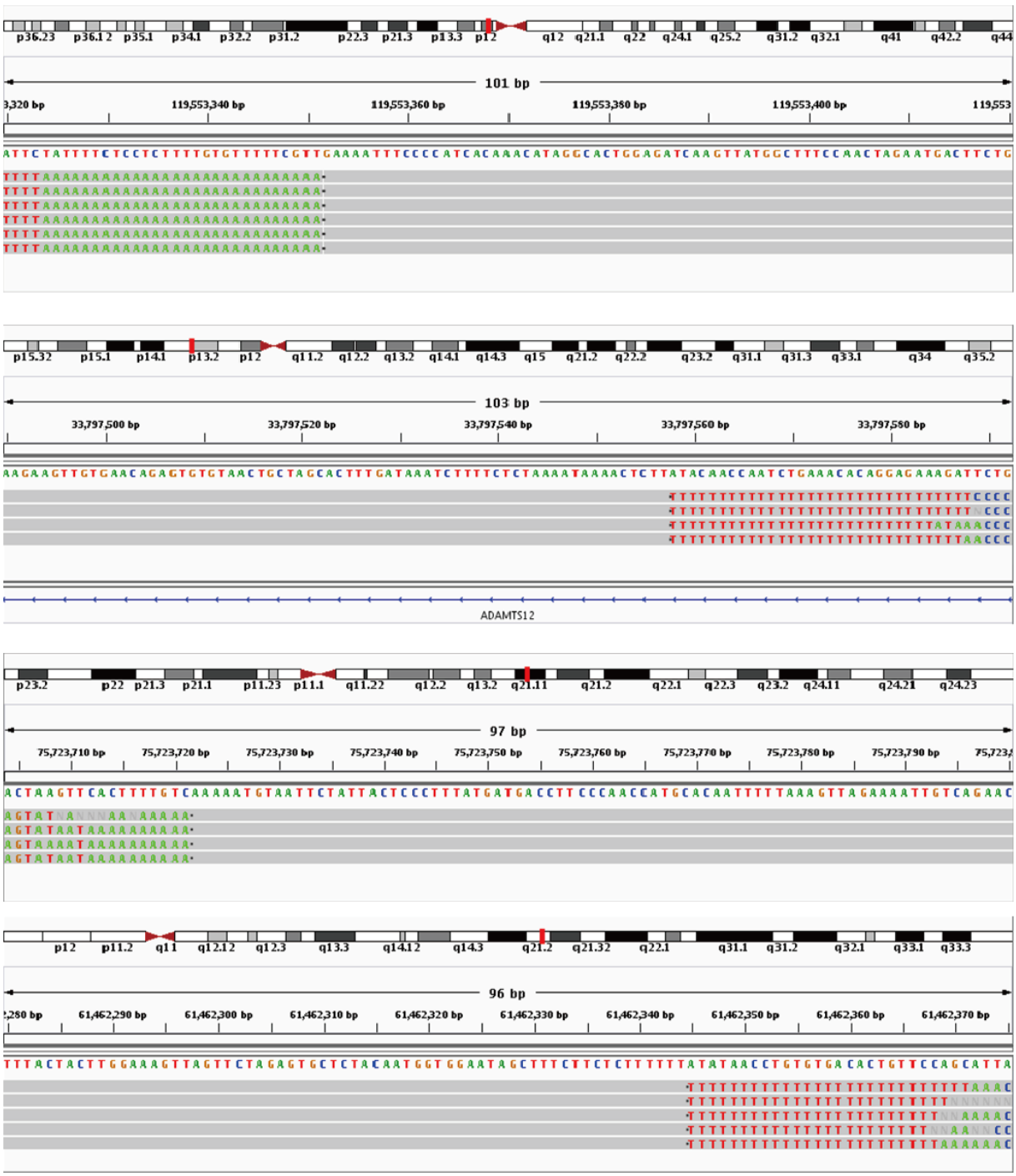
図 16. 人工遺伝子配列へのアラインメントの確認



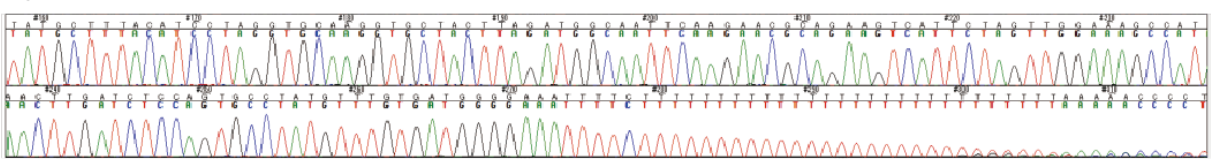
試料 IC2 に対して改良型 L1Hs-seq を行った。Internal_Control_v2 はアレル割合 2.5%を、Internal_Control_v3 はアレル割合 0.5%をシミュレートして混合されたものであり、ともにアラインメントが認められる。IC2 に対する実験では、0.5%, 2.5%の体細胞変異をシミュレートした人工遺伝子配列の検出ができた。

図 17. 非リファレンス L1Hs に対するサンガーシークエンス結果のアラインメント

a)



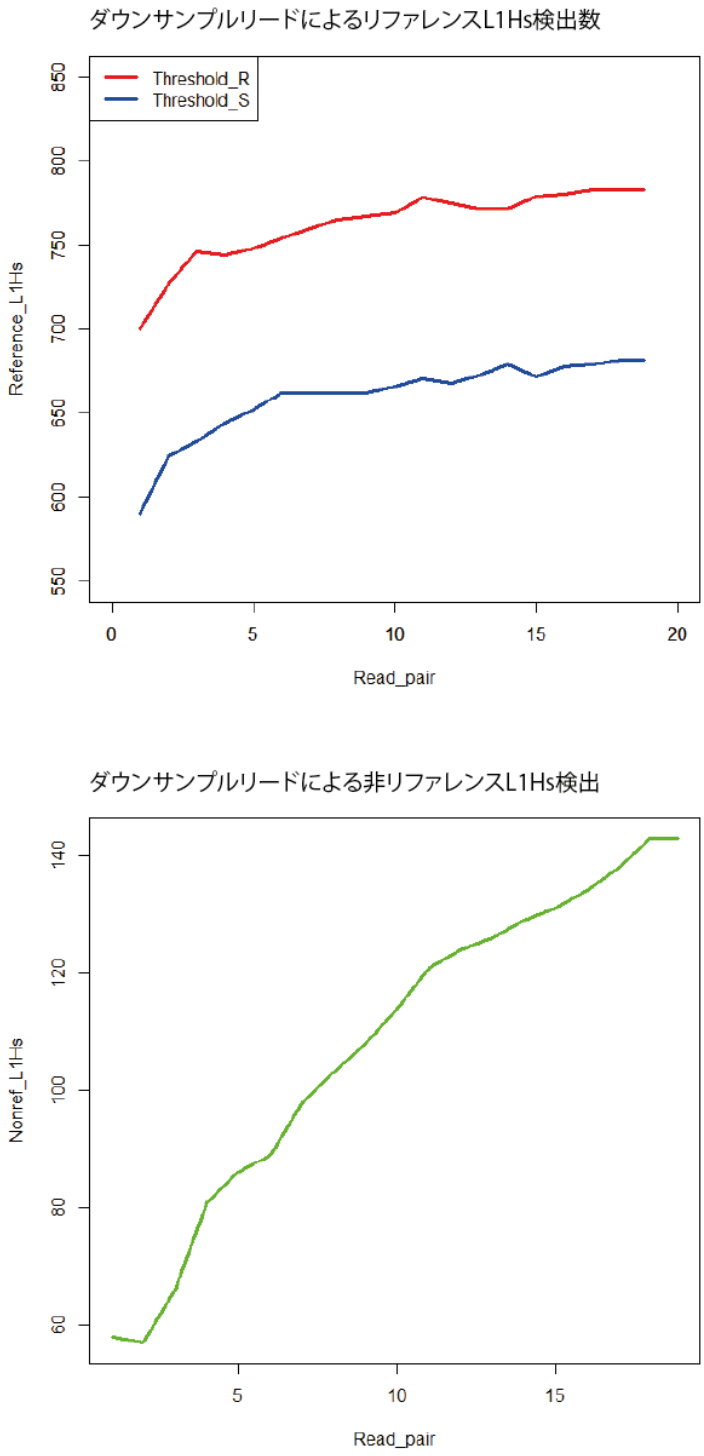
b)



a) 非リファレンス L1Hs 挿入が検出されたサイトに対する Nested PCR・サンガー法によるシークエンス結果のアラインメントを示した。BWA にてサンガーシークエンス結果を hg19 にアラインメントし、IGV にて可視化した。上から、chr1: 119553351, chr5: 33797557, chr8: 75723721, chr13: 61462344 におけるアラインメントを示しており、いずれもリファレンスゲノム配列と、リファレンスゲノム配列には存在しない poly-A が確認でき、ジャンクションサイトが一塩基レベルで確認できた。灰色で示されたリードは、表 12 の各プライマーセットによる Nested PCR プロダクトであり、chr1: 119553351 では 6 パターン、chr5: 33797557 では 4 パターン、chr8: 75723721 では 4 パターン、chr13: 61462344 では 4 パターンでのアラインメントが確認された。サンガー法により図で示された配列以外のリファレンスゲノム配列も明瞭に配列決定されたが、poly-A より 5’側の L1Hs 配列は明瞭に配列決定できなかった。

b) サンガーシークエンスの波形の一例（chr1: 119553351 における挿入）を示した。

図 18. リードペア数とリファレンス L1Hs 感度・非リファレンス検出数



試料 IC1 から百万リード単位でダウンサンプルし、リード数に対してリファレンス L1Hs 検出数と非リファレンス検出数がどう変化するかを解析した。X 軸がダウンサンプルしたリードペア数（単位：百万リード）であり、Y 軸がリファレンス L1 検出数（上段、上限 813 カ所）、リファレンスにない L1 配列を認めた領域の検出数（下段）を示す。上図の赤線が閾値 R での解析、青線が閾値 S での解析である。

図 19. 精神疾患発症における体細胞変異の位置づけのモデル

上段に、生殖系列ゲノムの変異に加え、発生・発達過程で生じる体細胞変異により、精神疾患の発症を説明する「多段階変異モデル」を示した。このモデルは発症を説明するとともに、疾患の多様性も説明するモデルとして想定した。下段に、体細胞変異を主因として精神疾患の発症を説明する「体細胞変異モデル」を示した。このモデルは、耐性致死をもたらす変異など、生殖系列ゲノム上には安定的に存在できないが、体細胞変異であれば存在し得る、効果の強い稀な変異を想定している。