平成 29 年度 修士論文

小型長鎖シークエンサーを用いた RNA-Seq 解析

(Analysis of RNA-seq with portable long-read sequencer)

2018 年 1 月 26 日提出

東京大学大学院新領域創成科学研究科

メディカル情報生命専攻

生命システム観測分野(指導教員: 鈴木穣 教授)

修士2年(47-166411) 勝俣恵理

目次

1）研究背景

近年、核酸の塩基配列決定技術は著しく発達している。次世代シークエンサーの登場により、生物の持つ遺伝情報をより高速かつ安価に解読できるようになってきている（Goodwin S, 2016）。次世代シークエンス技術が新たな研究分野を開拓した例は、ヒトを解析対象とした医療分野だけでなく、動植物を対象とした農産物や畜産物、微生物を対象とした新素材の開発等、多くの分野で報告されている。今後も、シークエンサーが生命科学の中で果たす役割はますます大きくなっていくと期待されている。

現在、次世代シークエンサーの主流となっている illumina 型シークエンサーは、いわゆるショートリードシークエンサーである。この型のシークエンサーは、塩基配列決定精度とスループットに優れる一方で、読み取れるリード長は最大 600 bp 程度である。近年の解析結果から、illumina 型シークエンサーでは 1 kb を超える構造多型の検出や反復配列の解読には不向きであることが示唆されている。そこで、リード長の長いロングリードシークエンサーを用いて、セントロメアやテロメア等の反復配列を多く含む領域の解読や、臨床サンプルを用いた数 kb 以上の構造異常の検出が試みられるようになってきている（Suzuki A, 2017）。

現在市販されているロングリードシークエンサーとして、Pacific Biosciences 社の一分子リアルタイムシークエンサーPacBio RS II/Sequel、Oxford Nanopore Technologies (ONT)社のナノポアシークエンサーMinION/PromethION が存在する。PacBio で用いられている SMRT DNA シークエンシング技術は、DNA ポリメラーゼによる伸長反応を行う際に生じる蛍光を検出して塩基を特定する。PacBio を用いた先行研究として、東北メディカルメガバンクの日本人標準ゲノムの作成が挙げられる（https://jrg.megabank.tohoku.ac.jp/）。SMART-Seq 法で合成した全長 cDNA をシークエンスして isoform を同定する Iso-seq (Sharon D, 2013)と呼ばれる RNA-Seq も開発されている。

ONT 社のナノポアシークエンサーは、膜に埋め込まれたタンパクナノポアを1分子の核酸が通過する際に生じる電流の変化を計測して塩基を決定する。ナノポアシークエンスでは、蛍光検出系を必要としないため、MinION のようなシークエンサーの小型化が実現されている。MinION はその携行性から、フィールドや臨床現場など様々な状況下での活用が期待されている。実際、フィールドでの MinION シークエンサーの活用は、ギニアでのエボラ出血熱患者のサンプル調査（Quick J, 2016）、国際宇宙ステーションでの大腸菌等の配列解析（Castro-Wallace S L, 2017)などが報告されている。また、ナノポアシークエンスのデータから核酸のメチル化修飾の検出に用いた報告もなされており（Simpson J, 2017, Rand A C, 2017）、MinION は幅広い分野での活用が可能だと考えられている。

本研究では、MinION を用いて RNA-Seq を行う一連のプロトコールの開発及び評価と、長鎖解読特性を活かしたアプリケーションの開発を行った。現在までに、MinION を用いた RNA-Seq の解析結果はいくつか報告されている（Oikonomopoulos S, 2016, Byrne A, 2017)が、本研究を開始した当時は報告がなかったため、手法の開発から行った。一細胞解析や臨床サンプル等微量の RNA に応用できる実用的なプロトコールを構築するため、全長 cDNA の合成には SMART-Seq2 法（Picelli S, 2014)を用いたタカラバイオ社の SMART-Seq v4 Ultra Low Input RNA Kit を使用した。RNA から変換した cDNA を読み取る RNA-Seq（以降、cDNA-Seq と呼

ぶ)を肺腺癌細胞株 6 株と同一個人由来の臓器7種に対して行った。cDNA に変換することなく、RNA 配列を直接読み取るRNA-Seq (direct RNA-Seq)を肺腺癌細胞株LC-2/adで行い、得られたリード長や精度、発現レベルを比較した。本研究によって、MinION は RNA-Seq 解析を行うために十分な性能を持ち、ショートリードシークエンサーでは難しかった解析も行うことができることを示した。

2) 研究手法

2A. cDNA-Seq / cell lines

2A-1. 使用サンプル

6 種の肺腺癌細胞株 H1975, H2228, LC-2/ad, PC-7, PC-9, VMRC-LCD を用いた(Suzuki A, 2014)。それぞれ培養したシャーレからスクレーパーで細胞をかきとり、4℃・2000xg・3 分間遠心して作成したペレットから RNeasy mini kit (QIAGEN)を用いて total RNA を抽出した。得られた RNA は Agilent Bioanalyzer (Agilent Technologies)の RNA nano キットを用いて測定した。RNA のクオリティを示す RNA Integrity Number (RIN)は 10, PC-7 のみ 9.4 とほとんど分解していない高品質の RNA であることを確認した。

2A-2. cDNA の合成と増幅

SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (タカラバイオ, 以下 SMARTer)を用いて total RNA から poly A tail を持つ RNA からの cDNA 合成と、生成した cDNA の PCR 増幅を行った。50 ng total RNA からスタートし、RNA, 10x Reaction Buffer 1 µl, 3' SMART-seq CDS primer ⅡA または dT primer (12 µM) 2 µl, Nuclease-Free Water で計 12.5 µl の反応液を 72℃ 3 min 反応させて cDNA 鎖の合成準備を行った。生成した mix に 5x Ultra Low First-Strand Buffer 4 µl, Smart2 TSO (48 uM) 1 µl, RNase inhibitor 0.5 µl, SMARTScribe Reverse Transcriptase 2 µl を加え、計 20 µl の反応液 (cDNA mix)で cDNA 合成反応を行った。42℃で 90 min 反応させて cDNA を合成してから、72℃ 10 min で酵素を失活させた。2x SeqAmp PCR Buffer 25 µl, PCR primer または UM PCR primer (12 uM) 1 µl, SeqAmp DMA Polymerase 1 µl, Nuclease-Free Water 3 µl の計 30 µl で master mix を作成し、cDNA mix に加えて合計 50 µl の PCR mix で PCR 反応を行った。95℃ 1 min プレヒートを行い、98℃ 10 sec, 65℃ 30 sec, 68℃ 3 min のサイクルを 16 回行ったあと、72℃ 10 min で処理した。PCR mix に 10x lysis buffer 1 µl を加えたあと、等量の AMPure XP beads (Beckman Coulter)で精製し、Nuclease-Free Water 17 µl に溶出した。溶出後、Agilent BioAnalyzer の DNA 7500 キットを使用して測定を行った。dT primer と PCR primer の配列は以下を用いた。

dT primer:

AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

PCR primer: AAGCAGTGGTATCAACGCAGAGT

dT primer (ビオチン修飾つき):

Bio-AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

PCR primer (ビオチン修飾つき): Bio-AAGCAGTGGTATCAACGCAGAGT

2A-3. MinION を用いたシークエンス

2A-2 で合成した cDNA 1 μg を用いてライブラリ調製を行った。cDNA, Ultra Ⅱ End-prep Buffer (New England Biolabs) 7 μl, Ultra Ⅱ End-prep enzyme mix (New England Biolabs) 3 μl, DNA CS 5 μl, の Nuclease-Free Water の計 60 μl の mix を作り、20℃ 5 min, 65℃ 5 min 反応させた。Mix に等量の AMPure XP beads (Beckman Coulter)を加えて精製し、Nuclease-Free Water 31 μl に End-Prepped DNA として溶出した。End-Prepped DNA 30 μl に Adapter Mix 10 μl, HP Adapter 2 μl, Blunt/TA Ligase Master Mix 50 μl, Nuclease-Free Water 8 μl を加えて室温で 10 min 静置した。HP tether 1 μl を加えてさらに 10 min 静置した。Bead Binding Buffer で wash した Dynabeads Myone streptavidin C1 (Life Technologies) 100 μl を用いて精製し、Elution Buffer 25 ul に再懸濁し、37℃ 10 min で処理した上清 25 μl を pre-seq mix として回収した。AMPure XP beads, Dynabeads Myone streptavidin C1 による精製の各段階で Qubit fluorometer DNA HS assay (Thermo Fisher Scientific)による測定を行い、核酸含有量を確認した。Loading buffer 1000 μl と loading library 150 μl を調製し、使用キットに対応するプログラムで 48 時間シークエンスを行った。使用キットとフローセルバージョン、ベースコールバージョンを Table 1 に示した。

Run number	CellType	FLOWCELL	PrepKit	Base caller
1	LC-2/ad	FLO_MAP104	SQK-NSK007	2D Basecalling RNN for SQK-NSK007
2	H1975	FLO_MAP104	SQK-NSK007	2D Basecalling RNN for SQK-NSK007
3	PC-9	FLO_MAP104	SQK-NSK007	2D Basecalling RNN for SQK-NSK007
4	PC-7	FLO_MIN105	SQK-NSK007	2D Basecalling RNN for SQK-NSK007
5	H2228	FLO_MAP104	SQK-NSK007	2D Basecalling RNN for SQK-NSK007
6	VMRC-LCD	FLO_MAP104	SQK-NSK007	2D Basecalling RNN for SQK-NSK007
7	VMRC-LCD	FLO_MIN105	SQK-NSK007	2D Basecalling RNN for SQK-NSK007
8-11	LC-2/ad	FLO_MIN106	SQK-LSK208	2D Basecalling for FLO-MIN106 250bps 1.125

Table 1. 各ランのバージョン

2A-4. HiSeq データの取得

LC-2/ad 細胞株の HiSeq RNA-Seq データのうち、TruSeq RNA Library Prep Kit でサンプル調製を行ったものは DNA Data Bank of JAPAN (DDBJ), DRA001846 より取得した

(ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/DRA001/DRA001846/DRX015062/)。 SMART-Seq (SMARTer + Nextera)は MinION の鋳型調製に用いた全長 cDNA より Nextera XT DNA Library Preparation Kit を用いて調製を行った。36 bp, SE, HiSeq2500 でシークエンスし、データを取得した。

2A-5. アライメント

アライメントには BWA-MEM (Li H, 2009), LAST (Kielbasa S M, 2011), BLAT (Kent W J, 2002)をそれぞれ用いた。BWA-MEM は default または-x ont2d のオプション設定で行った。LAST のバージョンは 833 で、last-train (Hamada M, 2017)をかけてパラメータを推定してからマッピングを行った。last-split をかけて構造異常の切断点やスプライスジャンクションをまたぐ配列を集約した。BLAT は default でマッピングした。それぞれ1本のリードが複数の箇所にマップされた場合はマップされた長さが最も長いパターンを選択した。リファレンスとしてトランスクリプトまたはゲノムを用いた。前者は RefSeq の全トランスクリプトの fasta ファイル、もしくは各遺伝子で最長の isoform のみを抽出した fasta ファイルを作成してリファレンスとした。後者は hg38 を用いた。HiSeq データの処理は BWA-MEM を用いた。パラメータは、TruSeq でサンプル調製したデータは PE、SMART-Seq データは SE を指定し、それ以外は default のまま処理した。

2A-6. identity, cover rate, 発現量の比較

RefSeq の全トランスクリプトをリファレンスとしてマッピングした結果を用いて identity, cover rate を計算した。また、発現レベルの比較には、最長の isoform をリファレンスとして用いた。アライメントした結果から、1つのリードに対して最も長くマップした単一 isoform のアクセションナンバーを付与した。identity はマップされたリード長のうち、リファレンスと同じ塩基の数の割合を計算した。cover rate はリファレンスの全長に対して各リードがカバーしている割合を計算した。発現量は、同じアクセションナンバーの数を数え、MinION データは RPM に、HiSeq データは TPM に補正した。得られた値に1を足し、底 10 で対数をとって散布図を描き、フリーソフト R によりピアソンの相関係数を求めて発現レベルを比較した。

2A-6. 融合遺伝子の探索

既知の融合遺伝子が検出できるか確認する場合、融合部分で配列をつなぎ合わせた fasta file を作成し、それをリファレンスとしてマッピングを行った。

新規の融合遺伝子は、以下の手順で検出した。BLAT を用いて hg38 をリファレンスとしてマッピングを行った。得られた結果のうち、2か所以上にマップされ、かつそれらが 5 bp 以内でつながるリードを抽出した。リードに遺伝子名を付与し、1本のリードが異なる染色体上にマップされた場合を融合遺伝子候補とした。同一染色体上でも、方向が異なるものや 1 Mb 以上離れた場合も候補とした。先行研究において、TruSeq でサンプル調製した HiSeq データを TopHat-Fusion (Kim D, 2011)で処理し、融合遺伝子候補のリストを得ている。MinION データと、HiSeq データのそれぞれの候補のうち、重複して検出できた遺伝子ペアを融合遺伝子候補とした。

2A-7. 融合遺伝子候補の検証

検出された融合遺伝子が本当に発現しているのか確認するため、サンガーシークエンスを行った。まず MinION でのシークエンスで用いたサンプルと同じ RNA で逆転写反応を行い、1st strand cDNA を得た。Total RNA 10 μg を用意し、Deoxynucleotide Solution Mix (New England Biolabs) 8 μl, 10 M dT primer (TTTTTTTTTTTTTTTTTTTT) 2.5 μl, RNasin Ribonuclease Inhibitors (Promega) 1 μl, SuperScript II Reverse Transcriptase (Thermo Fisher Scientific) 2 μl, 5X first-strand buffer (Thermo Fisher Scientific) 10 μl, 100 mM DTT (Thermo Fisher Scientific) 8 μl, Nuclease-Free Water (Thermo Fisher Scientific) を加えて合計 50 μl とした。12℃で 60 min、42℃ over night 後、RNase Free Water 50 μl を加えてピペッティングし、フェノール・クロロホルム 100 μl を加えて白く濁るまでピペッティングを行った。遠心 (室温・15000 rpm・5 min)し、上清を新しいチューブに移して 0.5 M EDTA (pH 8) 2 μl を加え、さらに 0.1 M 水酸化ナトリウム 15 μl を加えた。65℃で 40 min 処理し、1 M Tris-HCl (pH 7) 20 μl を加えた。99.5%エタノール 500 μl, 7.5 M 酢酸アンモニウム 70 μl, ethachinmate (ニッポンジーン) 1 μl を加えて遠心した (4℃・15,000 rpm, 10 min)。上清を除き、70% エタノール 700 μl でリンスした後、再度遠心した (4℃・15,000 rpm, 5 min)。上清を除いて風乾させ Nuclease Free Water 50 μl に溶出し、Nano drop (Thermo Fisher Scientific)で濃度の測定を行った。

得られた 1st strand cDNA を用いて PCR で標的の領域を増幅した。Template cDNA 1 μl, forward primer (5 μM) 2.5 μl, Reverse primer (5 μM) 2.5 μl, Phusion High-Fidelity PCR Master Mix with HF Buffer (Thermo Scientific) 10 μl, Nuclease Free Water 4 μl の mix を作成した。98℃ 30 sec, (98℃ 10 sec, 60℃ 30 sec, 72℃ 2 min)x40 cycle, 72℃ 10 min で PCR を行った。PCR 産物を QIAquick PCR purification kit (QIAGEN)で精製し、EB buffer 30 μl に溶出した。アガロースゲル電気泳動を行って標的のバンドを切り出し、QIAquick Gel Extraction Kit (QIAGEN)を用いて精製して EB buffer 30 μl に溶出した。得られたサンプルでサンガーシークエンスを行った。

fusion genes	forward primer	reverse primer	seq primer
CCDC6-RET	GCAGCAAGAGAACAAGGTGC	ACCATCCTAAGTTGCTGGGC	fwd
WAC-SFMBT2	AGCACAGGTCACAGTAAGGC	AACCTTAGTCACCGACGCAG	fwd
ZSCAN22-CHMP2A	AAGTTGGCTAGTCTCTGCGG	CCCAGCTCATCCAGAACCTG	fwd

Table 2. PCR とサンガーシークエンスに用いたオリゴプライマー

2A-8. Novel isoform の探索と検証

Martin Frith 博士より供与されたスクリプト(new-exon.sh)を用いて、さらに IGV 上での目視により候補を絞り込んだ。IGV 上での確認時に、「リードの端」「1リードしかない」「Gencode 上に記載されている」exon を持つ isoform を除いた。得られた候補から細胞増殖等に関与する遺伝

子を選び、novel exon を含む領域を PCR で増幅してからサンガーシークエンスを行った。用いたプライマーの配列を Table 3 に示した。

Gene symbol	Forward primer	Reverse primer	Sequence primer
MACC1	GTAATGGCGTGTGTTCTACC	TGACTGGCAGTCTTCACCTT	fwd
TRERF1	GAGAGTGAGGTGCCGAAGTC	ATGACACCTCCCAACTGCTG	fwd

Table 3. novel isoform の検証に用いたオリゴプライマー

2B. cDNA-Seq / tissue

2B-1. 使用サンプル

同一個人由来の7臓器から得られたRNA (Biochain Institute, Inc.)を購入し、同様に Agilent Bioanalyzer で測定したところ、RIN は Table 4 の通りになった。

Individual	Gender	Tissue	RIN	Individual	Gender	Tissue	RIN
#1	Male	Liver	7.5	#2	Female	Liver	8.6
#1	Male	Kidney	6.1	#2	Female	Kidney	8.1
#1	Male	Lung	7.1	#2	Female	Lung	8.4
#1	Male	Skeletal Muscle	6.7	#2	Female	Skeletal Muscle	6.8
#1	Male	Pancreas	8.6	#2	Female	Pancreas	6.3
#1	Male	Heart	6.4	#2	Female	Heart	8.6
#1	Male	Colon	5.6	#2	Female	Colon	6.9

Table 4. 正常組織由来 RNA の RIN

2B-2. cDNA の合成と増幅

2A-2 と同様に cDNA の合成と増幅を行った。SMARTer を用いて cDNA の生成・増幅を行った。AMPure XP beads で精製した。

2B-3. MinION を用いたシークエンスとアライメント

2B-2 で生成した cDNA を用いて、2A-3 と同様にライブラリ調製を行った。cDNA 1 μg からスタートし、精製の各段階で Qubit DNA HS kit (Thermo Fisher Scientific)による測定で核酸含有量を確認した。サンプル調製には SQK_LSK208 キットを用いた。ランには FLO_MIN106 (R9.4)フローセルを用いた。Albacore 2D 0.8.4 でベースコールを行った。得られたデータのうち 2D データのみを用いて LAST (last-833)でマッピングを行った。リファレンスは hg38 を用いた。last-train でパラメータを推定してからアライメントを行い、last-split をかけてスプライスジャンクションをまたいだ配列を集約した。

2B-4. WES データの取得

HiSeq 2500 (illumina)を用いて RNA と同一2検体の WES (Whole Exome sequence)を行っ

た。サンプル調製に SureSelect Human All Exon V5 (Agilent Technologies)を用いた。ランは PE, 100 bp で行った。SNP コールは GATK のベストプラクティスに従った。ヘテロ SNP データを以降の解析に用いた。

2B-5. アリル間で偏った発現を示す遺伝子の探索とフェージング

同一個人から得られた正常組織由来のトランスクリプトームから、アリル間で偏って発現する isoform を持つ遺伝子を探索した。各個人の持つゲノムに対して 2B-4 の通り WES を行い、それぞれの多型情報を取得した。MinION データの各リードに対して、RefSeq 上で最も近い構造を持つ isoform のアクセションナンバーを与えた。isoform 毎に、WES データで得られた SNP 情報を基に、同じ loci 上の SNP をカバーし、同じ SNP パターンを持つリードの数を数えた。Isoform 毎にカバーしているリード数で二項検定を行い、p 値を計算した。ボンフェローニ法に基づき p 値の補正を行った。補正した p 値から、有意に片アリルに偏って発現する isoform の有無を確認した。2つ以上のヘテロ SNP をまたいでいる isoform が性別・臓器毎にいくつ検出できたか数えた。

2C. direct RNA-Seq / cell line

2C-1. MinION を用いたシークエンスとアライメント

凍結しておいた培養細胞 LC-2/ad 株から μMACS mRNA アイソレーションキット, Small Scale (Miltenyi)を用いて total RNA を抽出し、polyA+ RNA を 500 ng 精製した。ONT 社の The Direct RNA Sequencing Kit (SQK-RNA001)を用いてライブラリ調製を行った。NEBNext Quick Ligation Reaction Buffer (New England Biolabs) 3 μl, 500 ng input polyA-tailed RNA 9 μl, RNA CS(内部コントロール) 0.5 μl, RT Adapter 1 μl, T4 DNA Ligase (New England Biolabs) 1.5 μl の mix1 を室温で 10 min インキュベーションした。Nuclease-free water 9 μl, 10 mM dNTPs 2 μl, 5x first-strand buffer (Thermo Fisher Scientific) 8 μl, 0.1 M DTT 4 μl の mix2 を mix1 に加えた。さらに SuperScript III reverse transcriptase (Thermo Fisher Scientific) 2 μl と共にピペッティングして混ぜた。50℃ 50 min, 70℃ 10 min インキュベーションを行って逆転写反応を行った。Agencourt RNAClean XP beads (Beckman Coulter) 72 μl を加えて精製し、nuclease-free water 20 μl に溶出した。逆転写した NEBNext Quick Ligation Reaction Buffer 8 μl, RNA Adapter 6 μl, Nuclease-free water 3 μl, T4 DNA Ligase 3 μl を加え室温で 10 min インキュベーションした。Agencourt RNAClean XP beads 40 μl を加えて精製し、Elution Buffer 21 μl に溶出した。Qubit fluorometer DNA HS assay (Thermo Fisher Scientific)で DNA の含有量を測定した。フローセル R9.5 (FLO-MIN107) でシーケンスを行った。ベースコールには albacore Ver 1.2.4 を用いた。1D リードとして得られたデータを、LAST (last-833)を用いて、リファレンストランスクリプトにマッピングした。last-train によるパラメータ推定と last-split による配列の集約を行った。

2C-2. Length distribution

LC-2/ad 株の cDNA-Seq, direct RNA-Seq それぞれの fasta ファイルや LAST でマッピング

した結果得られた各リードの長さを算出し、比較した。

2C-3. 発現レベルの比較
2A-6 と同様に発現レベルを比較した。

3) 結果と考察

3A. cell lines

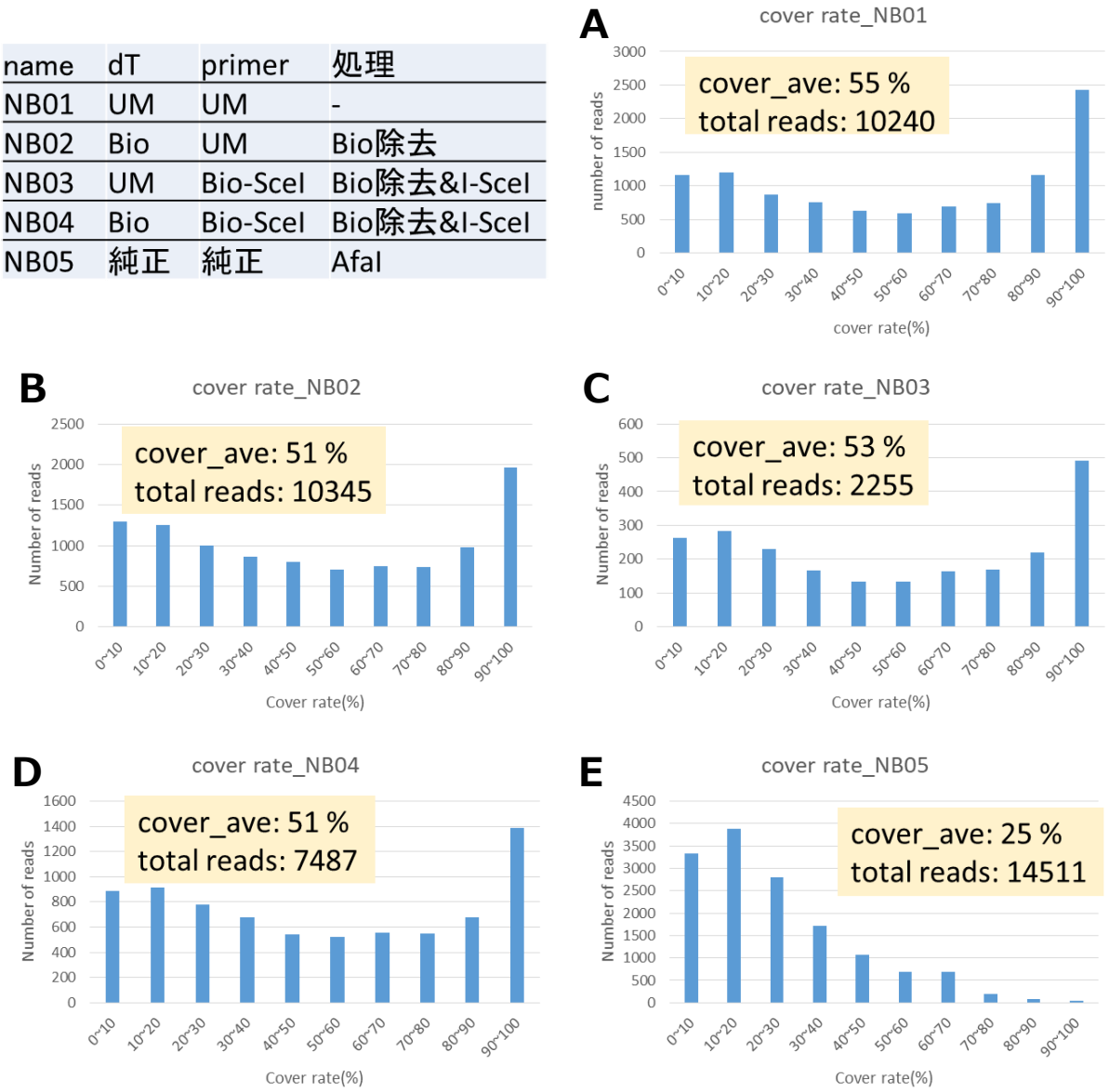
3A-1. MinION を用いた RNA-Seq のプロトコールの開発

LC-2/ad 細胞株から抽出した RNA を SMARTer (SMART-Seq2 法に基づいたキット)で全長 cDNA 生成・増幅を行った。得られたサンプルに対して、ONT 社のプロトコールの通りライブラリ調製を行った。リード数やマップ率ともに解析するのに十分なデータ量を得ることが出来なかった (Table 5, #1)。シングルセル解析にも用いられる SMARTer は、微量の RNA から効率良く cDNA の合成・増幅を行うために dT primer (dTp)や PCR primer (Pp)の 5' 末端に修飾が施されている。これらの修飾が MinION のシークエンスライブラリーを調製する際、アダプターのライゲーションを阻害する可能性が考えられた。そこで、修飾塩基を含まないプライマー (以下 UM)での反応や末端の修飾部位の制限酵素による切除を試みた。配列は SMART-Seq2 法 (Picelli S, 2014)で用いられている配列を参照した。また、5' 末端にビオチン修飾を施したプライマー(以下 M)も作成した。PCR プライマーはビオチン修飾に加え I-Sce I の 18 bp の認識配列を含めた。NB01 は dTp (UM), Pp (UM)で増幅し、そのままライブラリ調製に用いた。NB02 は dTp (M, ビオチン修飾), Pp (UM)で増幅後、ストレプトアビジンビーズでビオチン修飾された核酸やオリゴを除いた。NB03 は dTp (UM), Pp (M, ビオチン修飾+SceI の認識配列)で増幅後、I-SceI で制限酵素処理してビオチンを切除し、ストレプトアビジンビーズで精製した。NB04 は dTp (M, ビオチン修飾), Pp (M, ビオチン修飾+SceI の認識配列)で増幅後、I-SceI で制限酵素処理してビオチンを切除し、ストレプトアビジンビーズで精製した。NB05 は SMARTer キットに含まれる dT primer (M, 3' SMART-seq CDS primer ⅡA)と PCR primer (M)を用いて増幅した後、キットに付属した AfaⅠで制限酵素処理を行い、修飾を除いた。精製したそれぞれの cDNA の濃度を測定した。等量の cDNA を混合し、バーコード配列を利用したマルチプレックスのライブラリ調製を行い、MinION でのシークエンスを行った (Table 6)。

#	Flowcell version	dT primer	PCR primer	Number of pass_2D reads	mapped to genes (A)	Mapped to DNA_CS	mapped reads (B)	A / B (%)
1	R7.3	M	UM	153	60	70	130	46
2	R9	UM	UM	49,593	44,113	32	44,145	>99

Table 5. マッピング結果の比較
プライマー変更前(#1), 変更後(#2)のランのマッピング結果を示した。DNA_CS はライブラリ調製キットに含まれる内部コントロール。

name	dT	primer	処理
NB01	UM	UM	-
NB02	Bio	UM	Bio除去
NB03	UM	Bio-SceI	Bio除去&I-SceI
NB04	Bio	Bio-SceI	Bio除去&I-SceI
NB05	純正	純正	AfaI



左上：Table 6. 各サンプルの調製に用いたプライマーと増幅後の処理

Fig 1. 各サンプルの cover rata の分布

横軸に cover rate の分布、縦軸にリード数を示した。

リード数は NB05 が最も多く、次いで NB01, NB02 がほぼ同数であった（Fig 1）。最もリード数が多かったのはキットに含まれるプライマーを用いた NB05 であったが、cover rate が最も低かった。NB05 では Afa I による制限酵素処理を行っているが、Afa I の認識配列は 4 塩基であり、cDNA 中にも同じ配列が多数含まれている。cDNA の内部が切断され、cover rate が低くなったと考えられた。NB05 は cover rate が低いため、ロングリードを活かした解析に用いるのは難しい。NB01、NB02 は cover rate、リード数共にほぼ同じであるが、NB02 は煩雑な作業を必要とする。NB01 の dT primer、PCR primer どちらも修飾のないプライマーを用いた条件が最も単純なプロトコールであり、かつ長いリード長及び多くのリード数が得られる条件であった。Table

5 でプライマー変更前後のマッピング結果を示した。#1, #2 を LAST でリファレンストランスクリプト及びライブラリ調製の際に加える内部コントロール DNA_CS へのマッピングを行った。#1 ではマップされたリードは 130 リードだった。マップされたうち 54%にあたる 70 リードは DNA_CS にマップされていた。#2 ではマップされた 44,145 リードのうち DNA_CS にマップされたリードは 32 リードのみだった。ほとんどのリードはリファレンストランスクリプトにマップされており、#2 では高率で cDNA がシークエンスされたことが示された。なお、#1 と#2 ではフローセルのバージョンも異なるが、フローセルのアップグレードにより向上したのはアウトプット量や精度、塩基の決定スピードである。DNA_CS のマップ数は同等であることから、#1 と#2 の差はプライマーの変更によるものと考えられた。以降の実験では NB01 に用いたプライマーと同様の、修飾のない dT primer と PCR primer を用いた。

Run number	CellType	Version	Raw Reads (1D+2D)	2D		
				Total (pass+fail)	Pass	Fail
1	LC-2/ad	R9	169,401	70,508	49,593	20,915
2	H1975	R9	251,062	95,253	65,006	30,247
3	PC-9	R9	135,717	50,337	35,578	14,759
4	PC-7	R9	142,933	59,060	44,345	14,715
5	H2228	R9	240,567	88,364	58,134	30,230
6	VMRC-LCD	R9	84,733	29,986	19,313	10,673
7	VMRC-LCD	R9	104,863	38,228	27,359	10,869
8	LC-2/ad	R9.4	233,331	150,922	136,753	14,169
9	LC-2/ad	R9.4	252,951	159,117	142,245	16,872
10	LC-2/ad	R9.4	311,128	210,559	188,805	21,754
11	LC-2/ad	R9.4	410,826	280,851	255,947	24,904
Total (8-11)	LC-2/ad	R9.4	1,208,236	801,449	723,750	77,699

Run number	CellType	Ver.	Number of raw reads	Number of 1D reads
12	LC-2/ad	R9.5	737,230	556,195

Table 7. The sequencing summary (cDNA-Seq, cell lines)
各ランのシークエンス結果を示した。(ラン 1-11)左から順に、ランナンバー・株名・フローセルバージョン・総リード数・2D トータルリード数・2D パスリード数・2D フェイルリード数 を示した。(ラン 12) 左から順に、ランナンバー・株名・フローセルバージョン・総リード数・1D リード数 を示した。

6 種の細胞株から total RNA を抽出し、SMARTer と決定したプライマーで cDNA を合成・増幅してからシークエンスを行った (Table 7)。LC-2/ad 株は既知の融合遺伝子 CCDC6-RET を持つことが確認されている。CCDC6-RET 遺伝子は比較的高い発現を示すため、raw read が 100 万リードを超えるまで深く読んだ (ラン 8-11)。H2228 株も融合遺伝子を持っており、EML4-ALK 遺伝子の発現が確認されている。PC-9 株, H1975 株は EGFR に変異を持っている。PC-9 株は exon19 に 15 bp の欠失を持つ。H1975 株は exon21 の L858R 点変異と exon20 の T790M 点変異を持ち、EGFR のチロシンキナーゼ活性阻害剤ゲフィチニブに対して耐性を示す。VMRC-LCD 株はがん抑制遺伝子 STK11 にスプライスサイト変異がヘテロで入っている。さらに、MYCN の CNA(Copy Number Aberration, gain)が確認されている。PC-7 株は MYC の CAN (gain)と、がん抑制遺伝子 NF1 にスプライスサイト変異が入っていることが確認されている。各ランで得られたアウトプットはフローセルのバージョンに依存していた。R9 フローセル (ラン 1-7)では、raw read が 10-25 万リードだった。クオリティの高い pass 2D リードは 2-7 万リードで、30%程度だった。R9.4 フローセル (ラン 8-11)では raw read が 25-40 万リードだった。pass 2D リードは 13-25 万リード、50%以上が pass 2D リードだった。以降の cDNA-Seq の解析には主にラン 8-11を合算したデータを用いた。

LC-2/ad 細胞株については、polyA RNA に直接アダプターをライゲーションしてシークエンスを行う direct RNA-Seq も行った (Table 7, ラン 12)。フローセルは最新の R9.5(FLO-MIN107)を用いた。raw read が約 74 万リード、解析に使える 1D read は約 56 万リードで全体の 75%を占めていた。R9.4 に比べてアウトプットは格段に増加していた。

3A-2. シークエンス、マッピングの確認

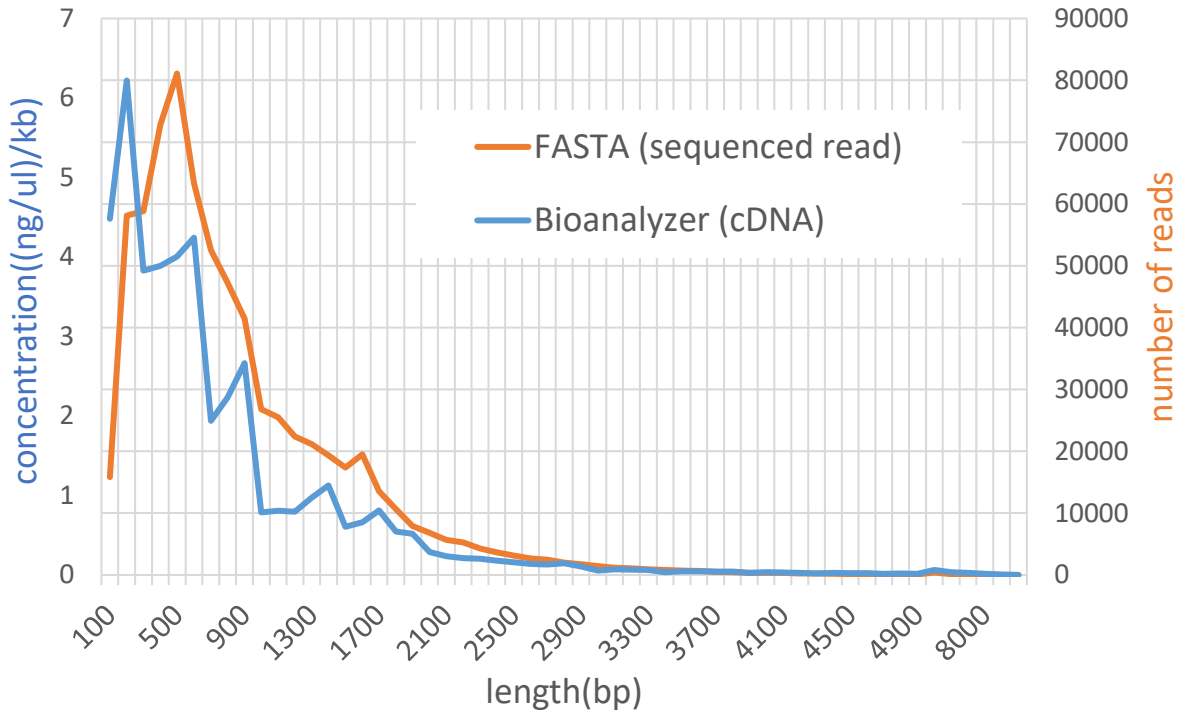


Fig 2. 全長 cDNA (アダプターライゲーション前)とリード長 (シークエンス後)の長さの分布 cDNA は Bioanalyzer の濃度 (ng/ul)の結果から、長さを補正してプロットした (青)。sequenced read は MinION によってシークエンスされたリード長を示した (橙)。

アダプターライゲーション前とシークエンス後の pass 2D リードを比較し、cDNA の長さによる読み取りバイアスが存在するのか確認した (Fig 2)。アダプターライゲーション前の cDNA 長の分布とシークエンス後のリード長は似た分布を示した。cDNA 長と fasta データのリード長を比較すると、リード長のほうが長いリードの割合が大きかった。Bioanalyzer は電気泳動と蛍光検出系を用いており、直接リード長を数える MinION でのシーケンスに比して誤差が大きいと考えられた。

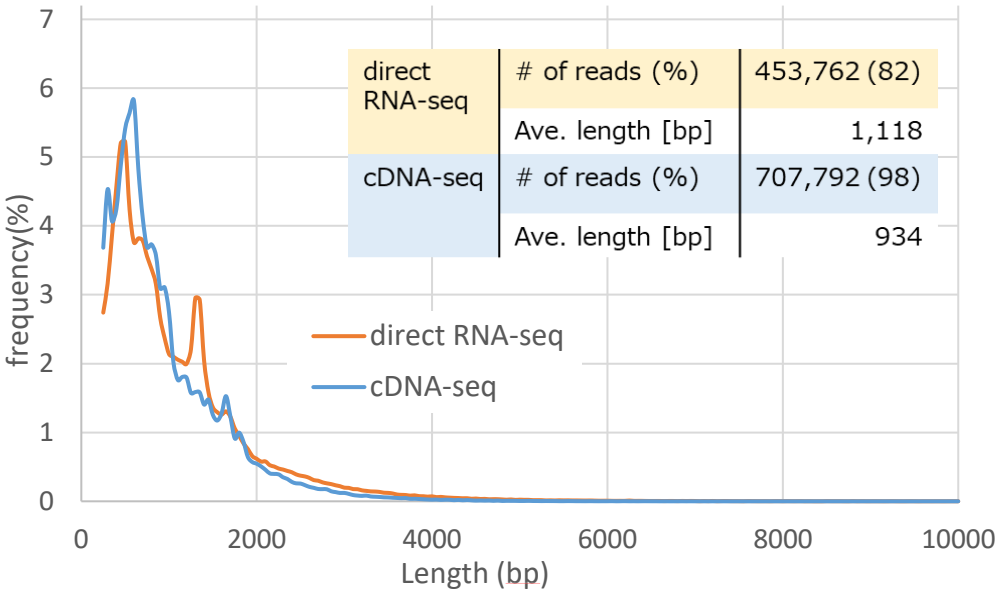


Fig 3. fasta ファイルでのリード長の分布比較
fasta ファイルのリード長をカウントし、横軸に 200 bp-10 Kb までの長さを、縦軸に各長さの頻度を示した。cDNA-Seq は pass 2D リードを、direct RNA-Seq は 1D リードの長さの分布を示した。

cDNA-Seq と、direct RNA-Seq によりシークエンスされたリード長を比較した(Fig 3)。全体では似たような分布を示す一方、direct RNA-Seq では約 1.5 kb の頻度が高くなっていた。平均長も direct RNA-Seq のほうが高かった。direct RNA-Seq も cDNA-Seq と同様にシークエンスが可能なだけでなく、大きな PCR バイアスをかけることなくシークエンスが可能であったと考えられた。一般に、PCR を行うことで長い配列や GC リッチな配列は増幅されにくい等のバイアスが生じると考えられている。Direct RNA-Seq は PCR を行わずにシークエンスするため、より元の RNA 状態に近いリード情報を得られる可能性がある。

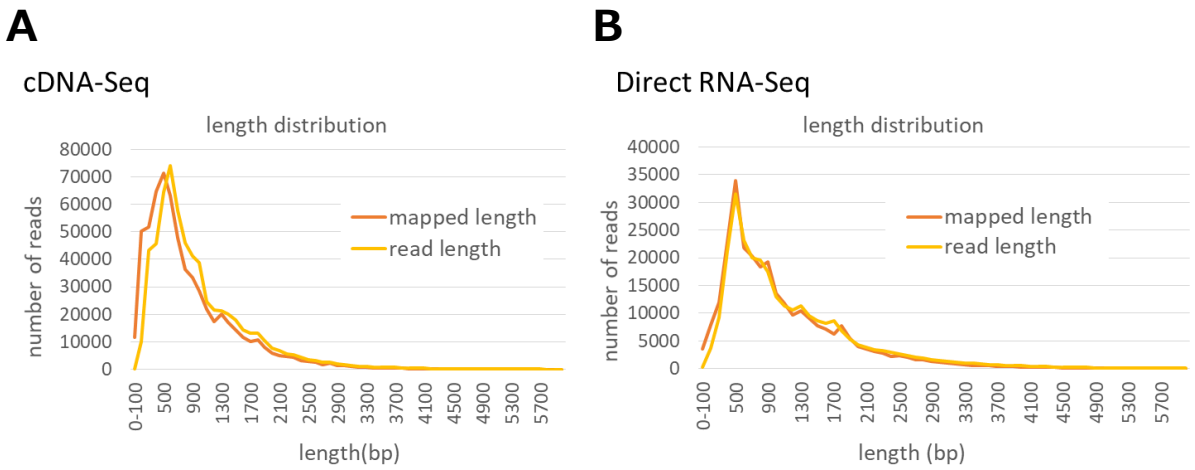


Fig 4. シークエンスリード長とマップされたリード長
fasta ファイルの塩基配列の長さをカウントしたもの（橙)と、LAST でマッピングしたリードの長さをカウントしたもの（黄)を比較した。cDNA-Seq (A), direct RNA-Seq (B)の分布をそれぞれ示した。横軸にリード長を、縦軸に各長さに該当するリードの数を示した。

シークエンスの raw read である fasta データと、LAST でマッピングしたデータにおいて、200 bp 以上のリードの長さの分布を解析した (Fig 4)。cDNA-Seq, direct RNA-seq それぞれのアダプターは約 100 bp, 30-50 bp である。cDNA-Seq (Fig 4A)の分布も direct RNA-Seq (Fig 4B)の分布もマッピング前後で大きな変化は見られなかったが、cDNA-Seq において多少の不一致が検出された。アダプターの長さだけでは説明出来ないエラー、PCR によって一部誤った配列が増幅されている等が発生しているのかもしれない。

3A-3. 解析に用いるアライメントツールの評価

	LAST	BWA-MEM (default)	BWA-MEM (-ont2d)	fasta data
mapped reads(%)	630,364(87)	634,493(88)	645,808(89)	723,750(100)
Average length[nt]	854	811	863	917

Table 8. 各マッピングソフトでマップしたときのマップ率とマップされた平均長

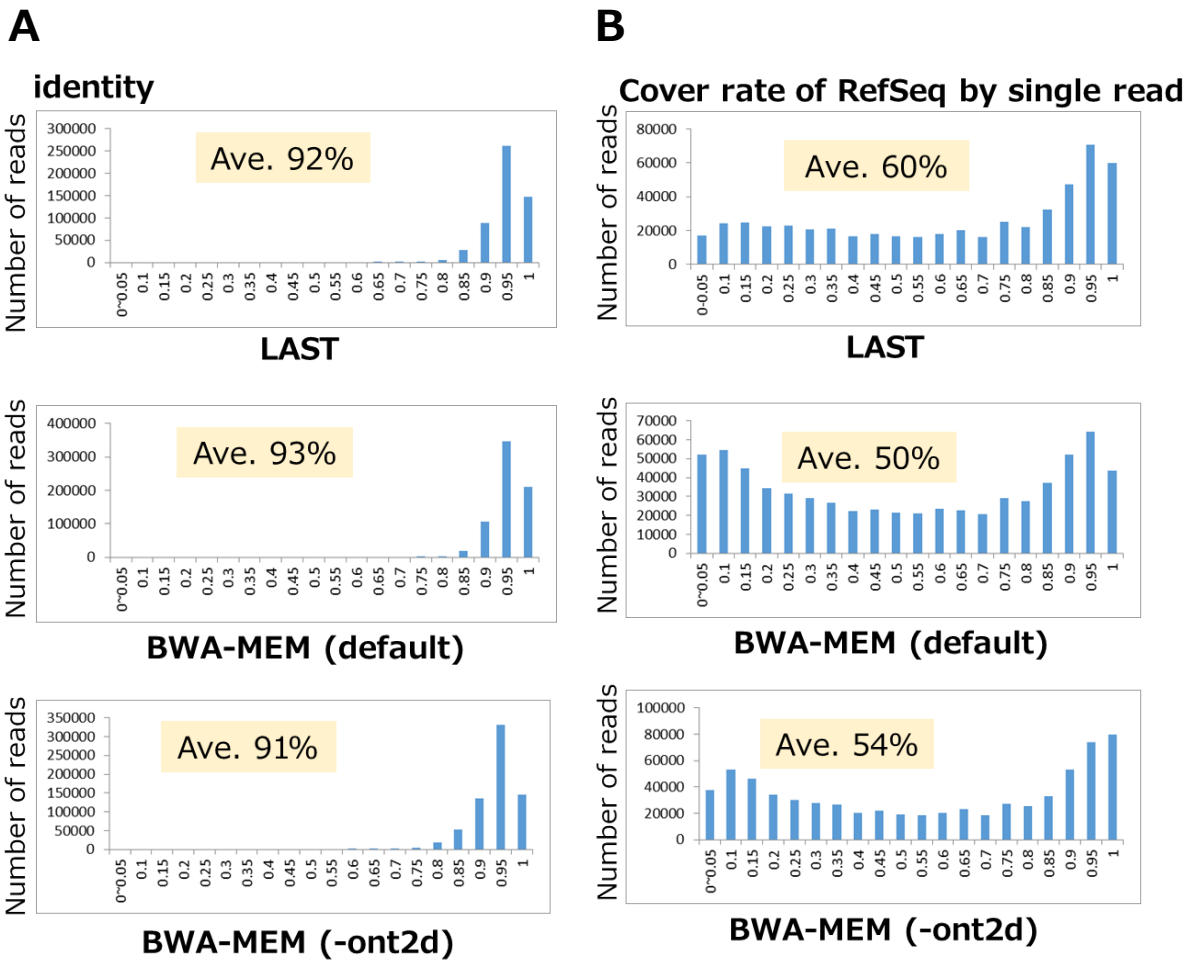


Fig 5. マッピングソフトの検討
上から順に LAST, BWA-MEM(default), BWA-MEM (-ont2d)で得られた結果から算出(A, B)。横軸に identity (A)または cover rate (B), 縦軸に該当するリード数を示した。グラフ上に平均値を示した。

Sample name	Flow cell version	Number of raw reads	Number of 1D reads	Mapped reads (%)	Average mapped length [bp]
LC-2/ad	R9.5	737,230	556,195	319,263 (57)	1,031

Table 9. The sequencing summary (direct RNA-seq)
direct RNA-Seq のシークエンス結果を示した。左から順に、使用した細胞株名・フローセルバージョン・総リード数・1D トータルリード数・LAST でマッピングされたリード数・LAST でマッピングされた各リードの平均長を示した。

MinION を用いた全長 cDNA-Seq の解析パイプライン検討のため、LAST と BWA-MEM を用いてマッピングを行った。リファレンスには RefSeq 上の全 isoform (計 58,196 本)の fasta ファイルを用いた。BWA-MEM は、パラメータを default のものと、オプション-ont2d の2種類を用いた。得られたデータを用いて、マップ率 (マップされたリード数 / fasta ファイル中のリード数)、マップされた平均長、identity (一致した塩基の数 / マップされた長さ)、cover rate (マップされた長さ / マップされたリードの長さ)を比較した。リードのマップ率は LAST が 87%、BWA-MEM (default)が 88%、BWA-MEM (-ont2d)が 89%であった (Table 8)。identity については、LAST と BWA-MEM (default), BWA-MEM (-ont2d)でそれぞれ 92%, 93%, 91%であった (Fig 5A)。 cover rate はそれぞれ 60%, 50%, 54%であった (Fig 5B)。BWA-MEM (-ont2d)は cover rate, identity で BWA-MEM (default)と LAST の中間値をとっているため、以降の解析には用いなかった。BWA-MEM は、LAST に比して identity は同等であるが、マップ率が高く、cover rate が低かった。これは、BWA-MEM は類似度が高い領域のみに短くアライメントしたためと考えられた。逆に、LAST はマップ率が低いが、BWA-MEM と同程度の identity でより長くアライメント出来るため、本研究の目的により適していることが示唆された。

direct RNA-Seq では、LAST を用いて cDNA-Seq と同じリファレンスに対してマッピングを行った (Table 9)。LAST によるマップ率は 57%, マップされたリード長は 1,031 bp であった。direct RNA-Seq リードのクオリティは cDNA-Seq のリードよりも低い。cDNA-Seq の結果に比して妥当なものであると考えられた。

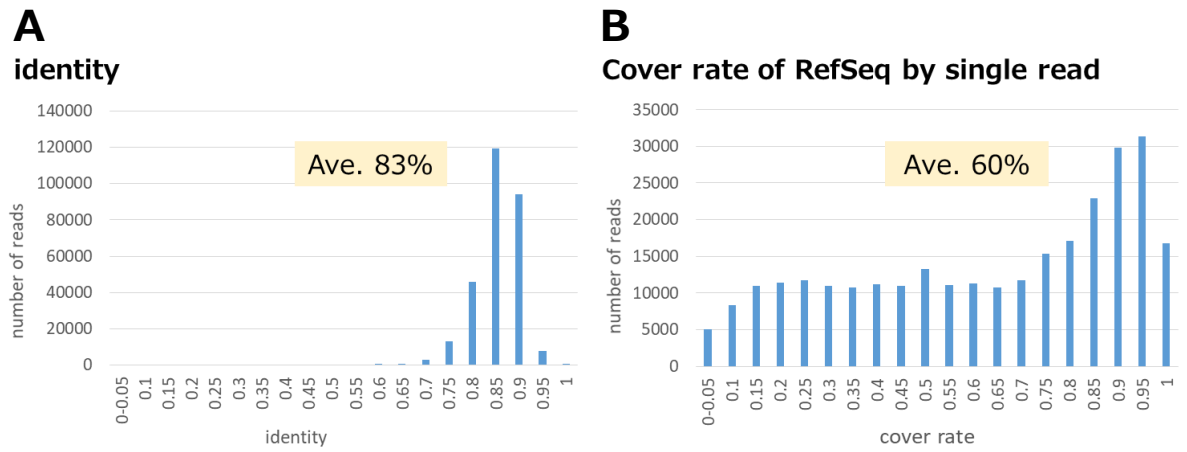


Fig 6. direct RNA-Seq の identity, cover rate
direct RNA-Seq のデータを LAST で処理し得られた結果から算出した (A, B)。横軸に identity (A)または cover rate (B), 縦軸に該当するリード数を示した。グラフ上に平均値を示した。

direct RNA-Seq のシークエンス精度を調べるため、identity と cover rate をそれぞれ算出した(Fig 6)。direct RNA-Seq は PCR を行わずに RNA をシークエンスすることができている。ただし、identity は平均 83%にとどまっていた。cDNA 配列の読み取りに比して、RNA の読み取りは十分ではなかった。cover rate は 60%で cDNA-Seq と同等であった。

3A-3. 発現レベルの比較

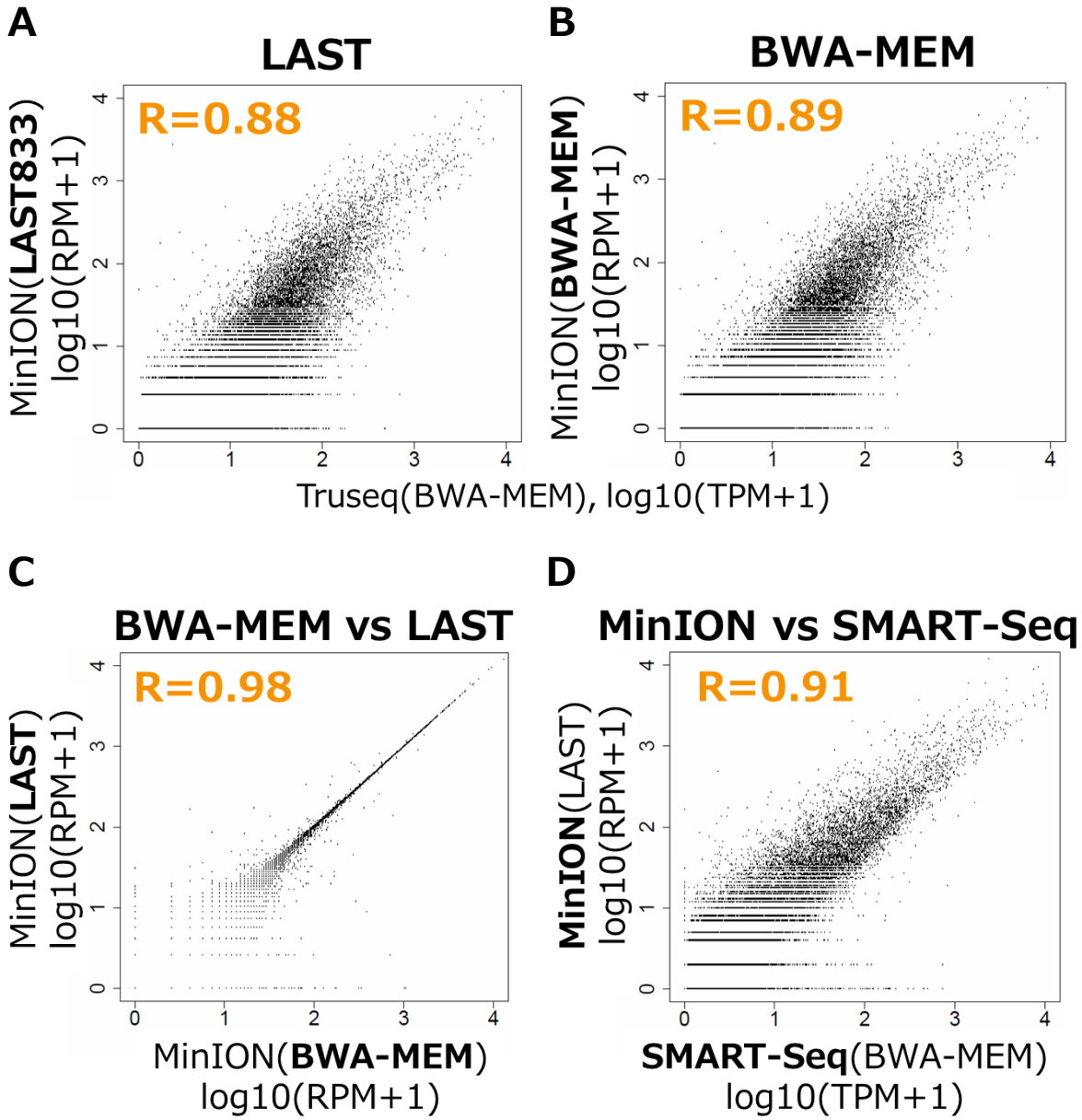


Fig 7. MinION での cDNA-Seq と HiSeq (illumina)での cDNA-Seq の発現レベルの比較
(A, B) 横軸に HiSeq データ(サンプル調製は TruSeq), 縦軸に MinION データを示す。マッピングソフトに(A)LAST, (B)BWA-MEM を用いた。(C)横軸に BWA-MEM でマッピングしたした MinION データ, 縦軸に LAST でマッピングした MinION データを示す。(D)横軸に HiSeq データ(サンプル調製は SMART-Seq), 縦軸に MinION データを示す。
RPM = raw counts / # total reads * 1M
(RPK = raw counts / gene length), TPM = RPK / sum all RPK * 1M

各遺伝子の発現量を、MinION を用いた cDNA-Seq のデータとショートリードシークエンサー HiSeq のデータを比較した。リファレンスには RefSeq 上の mRNA として登録されているもののうち、各遺伝子で最長の isoform (計 19,302 本)を選んで作成した fasta ファイルを用いた。リ

ファレンスに1つの遺伝子につき1つの isoform を用いることで、1つのリードがよりスコアの高い isoform にスプリットしてマップされる状況を回避した。illumina HiSeq による RNA-Seq のサンプル調製には TruSeq が一般的に用いられる。TruSeq は最初に RNA を断片化し、短くなった配列に対してランダムプライマーで逆転写、アダプターライゲーションを行うものである。TruSeq で調製してシークエンスを行ったデータと、LAST または BWA-MEM でマッピングした MinION の発現レベルを比較した。それぞれ 0.88, 0.89 のピアソン相関係数を示した（Fig 7A, 7B)。LAST でマッピングした MinION データと BWA-MEM でマッピングした MinION データを比較して、相関係数を調べた（Fig 7C)。LAST と BWA-MEM はアライメント方法が大きく異なるが、相関係数は 0.98 を示した。MinION のデータの発現量を調べる際には、LAST と BWA-MEM どちらを使っても差は無いと考えられた。HiSeq データにはサンプル調製に SMART-Seq（SMARTer + Nextera)を使う鋳型調製法もある。SMART-Seq では、最初に cDNA 全体を増幅し、のちに断片化とアダプターライゲーションを行う。これは断片化のステップ以外 MinION のサンプル調製と同じ順序である。SMART-Seq でサンプル調製を行った HiSeq データと MinION データの発現レベルを比較したところ、ピアソンの相関係数は 0.91 を示した（Fig 7D)。MinION とサンプル調製方法が類似しているため、TruSeq で調製したときよりも SMART-Seq で調製したほうが相関係数は高くなったと考えた。

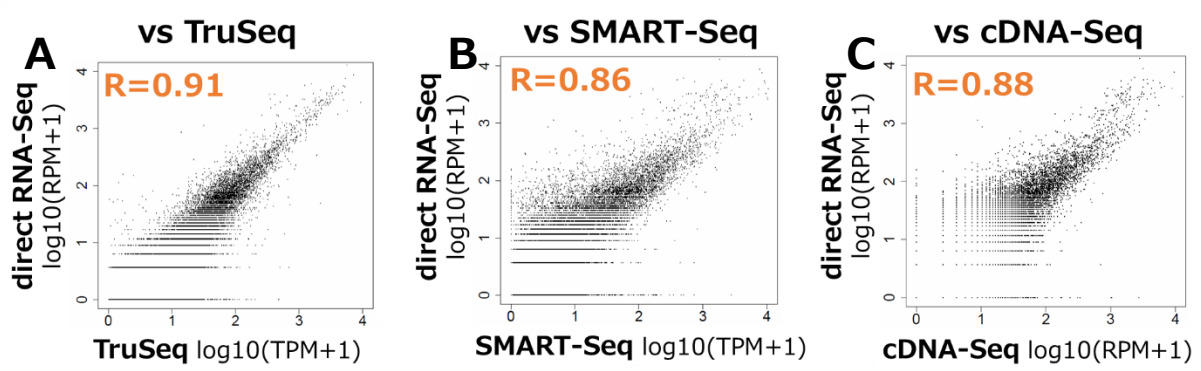
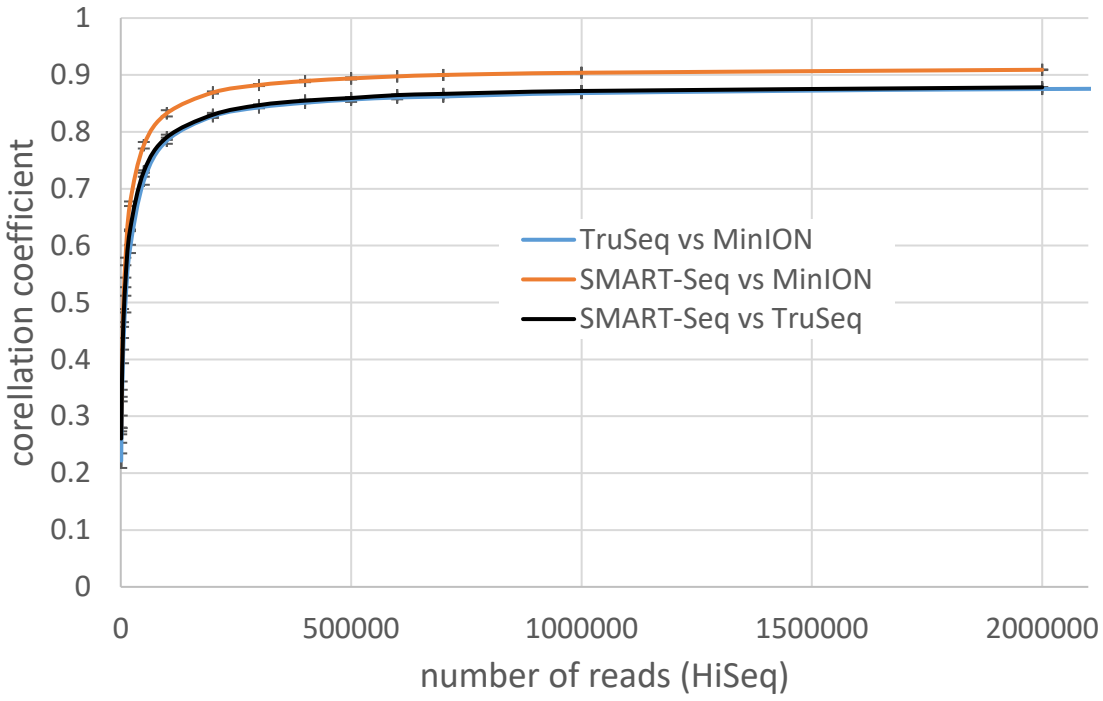


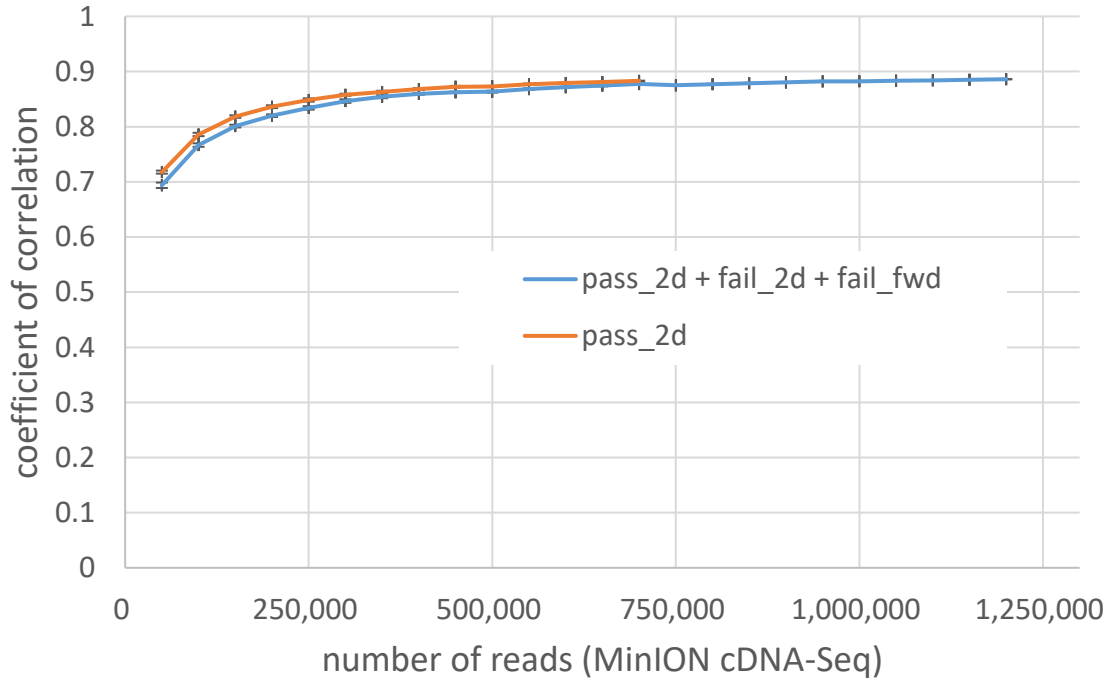
Fig 8. 各プラットフォーム間の発現レベルの相関
横軸に HiSeq (TruSeq)または HiSeq (SMART-Seq)または MinION (cDNA-Seq)の発現レベル、縦軸に MinION (direct RNA-Seq)の発現レベルを示した（A-C)。散布図の左上にピアソンの相関係数を示した。
RPM = raw counts / # total reads * 1M
(RPK = raw counts / gene length), TPM = RPK / sum all RPK * 1M

direct RNA-Seq により得られたデータを他法で得られたデータと比較した。direct RNA-Seq は MinION の cDNA-Seq とは強い相関を示した（Fig 8C)。興味深いことに、TruSeq でサンプル調製を行ったもののほうが、SMART-Seq でサンプル調製を行ったものよりも相関が高くなっていた（Fig 8A, B)。TruSeq でのサンプル調製は全長ではなく、断片化された cDNA を合成・増幅するため、SMART-Seq に比して PCR バイアスが少ない可能性が示唆された。

A


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
corellation coefficient
TruSeq vs MinION
SMART-Seq vs MinION
SMART-Seq vs TruSeq
0
500000
1000000
1500000
2000000
number of reads (HiSeq)

B


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
coefficient of correlation
pass_2d + fail_2d + fail_fwd
pass_2d
0
250,000
500,000
750,000
1,000,000
1,250,000
number of reads (MinION cDNA-Seq)

C

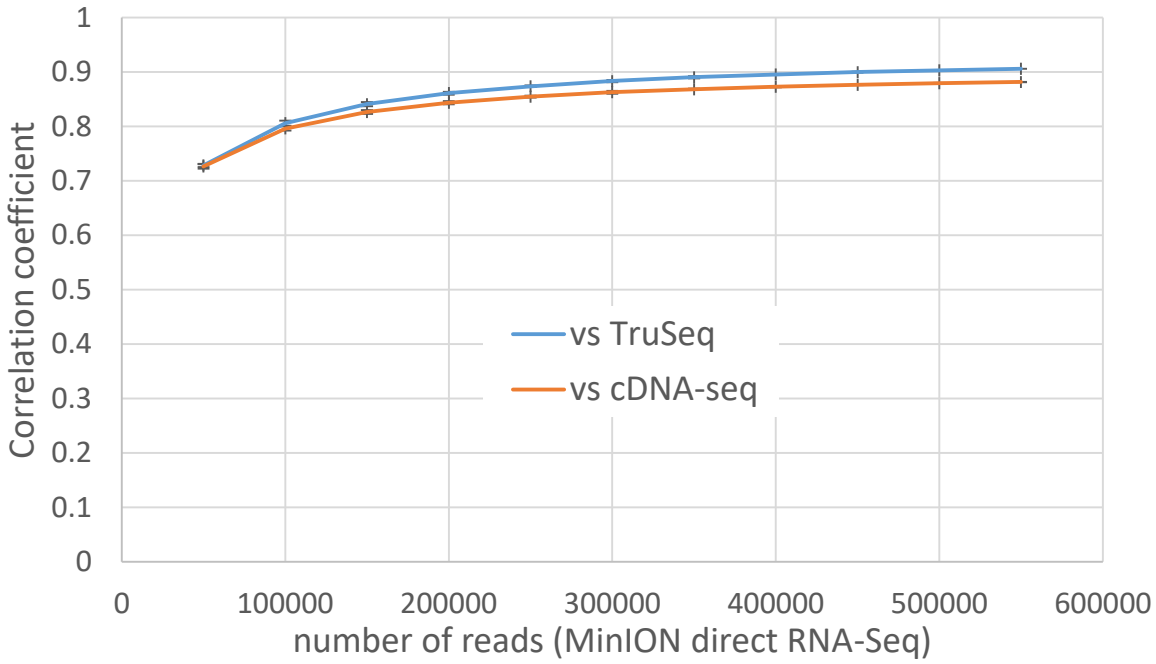


Fig 9. データ数を変化させたときの相関係数の推移
(A) HiSeq のリード数を横軸に、各リード数でのピアソンの相関係数を縦軸に示した。
(B) MinION のリード数を横軸に、各リード数でのピアソンの相関係数を縦軸に示した。
(C) direct RNA-Seq のリード数を横軸に、各リード数でのピアソンの相関係数を縦軸に示した。

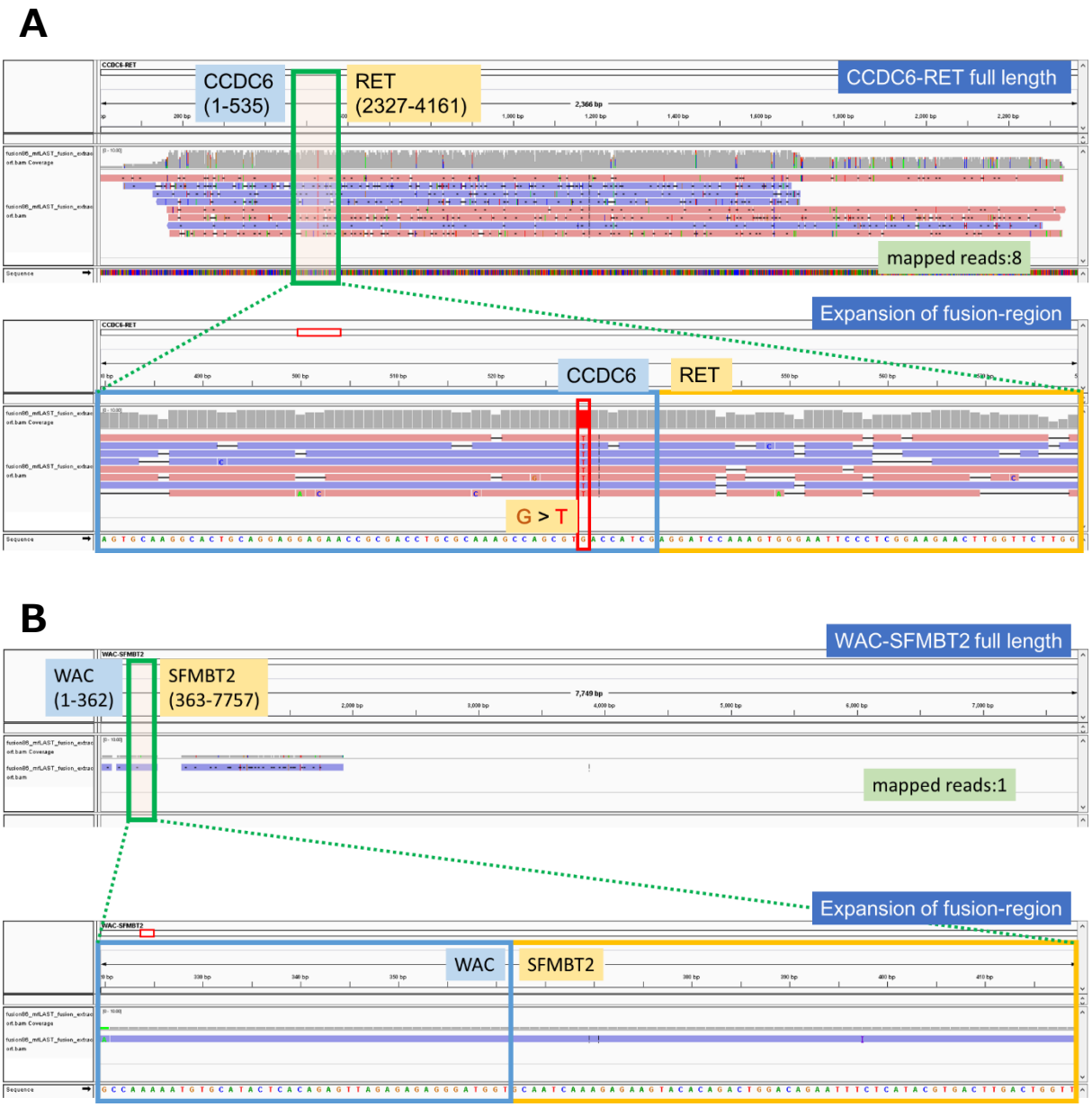
MinION もしくは HiSeq のリード数が、相関係数の算出に十分であるかを検討した。それぞれのデータから段階的にリードをランダムに取り出し、そのシークエンス深度での相関係数を求める作業を 10 回ずつ繰り返した。それぞれに、10 回の試行での相関係数の平均と標準偏差を求めて可視化した (Fig 9)。HiSeq のリード数を変化させたとき (Fig 9A), cDNA-Seq のリード数を変化させたとき (Fig 9B), direct RNA-Seq のリード数を変化させたとき (Fig 9C)ともに、ピアソンの相関係数は 0.9 前後でプラトーに達していた。シークエンス深度を上げても相関係数が 1 に到達するとは考えづらかった。

Fig 9A において、HiSeq のリード数を変化させた。SMART-Seq で調製した HiSeq データと MinION cDNA-Seq データの相関 (橙)が最も高くなった。TruSeq で調製した HiSeq サンプルデータと、SMART-Seq で調製した HiSeq データ (黒)及び MinION cDNA-Seq データ (青)は同等の相関を示した。発現レベルは用いた HiSeq あるいは MinION シークエンサーというよりも、ライブラリ調製方法に依存していた。Fig 9B では、クオリティの高い pass 2D リードの数を変化させてシミュレーションを行った。クオリティの高くない fail 2D リードや fail fwd リードを加えてリード数を増やしたデータを用いてもシミュレーションを行った。50 万リードを超えるとどちらも約 0.9 でプラトーに達していた。Fig 9C では、direct RNA-Seq のリード数を変化させてシミュレーション

を行った。TruSeq でサンプル調製した HiSeq データと比較した相関の方が、cDNA-Seq のデータと比較した相関より高かった。どちらもプラトーに達していると推測された。

MinION cDNA-Seq データと HiSeq データを比較したとき、相関係数は約 0.9 でリード数増加に対してプラトーに達した。この差は HiSeq と MinION のシークエンサー自体の特性に由来するものだと考えられた。MinION、 HiSeq それぞれで読めない、あるいは読みにくい配列が存在していれば、その配列を持つ遺伝子は発現量が低く見積もられる。どの配列が読めないのか明らかになっていないが、シークエンサー毎の特性が発現量の推測に影響を与えていると考えられた。

3A-4. 融合遺伝子の検出と検証



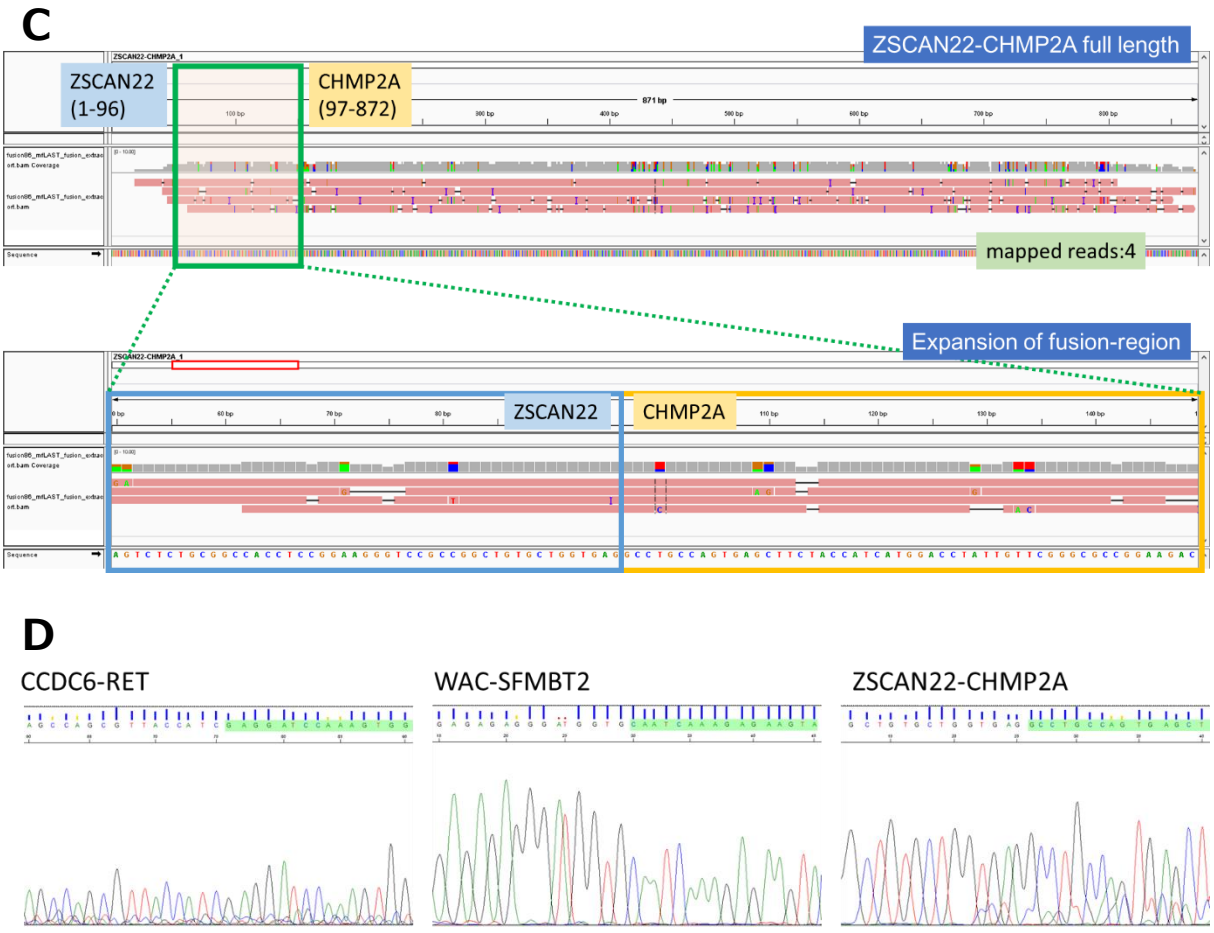


Fig 10. 検出された融合遺伝子
(A-C)IGV 上で見た MinION データ。上段が全長、下段が融合部分の拡大を示した。
(A) LC-2/ad 株で既知の融合遺伝子 CCDC6-RET, (B) LC-2/ad 株の新規候補融合遺伝子 WAC-SFMBT2, (C) PC-9 株の新規融合候補遺伝子 ZSCAN22-CHMP2A。
(D)サンガーシークエンスで確認できた融合部分を示した。

MinION データの解析から、融合遺伝子の発現が確認された。肺腺癌細胞株 LC-2/ad で既知の融合遺伝子 CCDC6-RET は融合する座位が知られている (Matsubara D, 2012)。融合した配列の fasta ファイルを作成し、それをリファレンスとして MinION データのマッピングを行った。ゲノムブラウザ IGV 上でマッピング結果を可視化した。約 72 万リードのうち 8 リードが CCDC6-RET の融合部分をカバーした (Fig 10A)。CCDC6-RET 融合部分の近くには SNP が知られている。この SNP を手掛かりに遺伝子融合が生じているアリルを同定することが可能であった。また、この遺伝子融合はサンガー法によるシークエンスにより確認された (Fig 10D)。

新規融合遺伝子を探索するため、各細胞のデータを、BLAT を用いてゲノムに対してアライメントを行った。アライメント結果のうち、2 カ所以上の遺伝子にマップしているリードを集計し、そのマップされた領域にある遺伝子ペア情報を集めた。融合候補の遺伝子ペアと、HiSeq RNA-Seq データから検出された融合遺伝子候補の遺伝子ペアとを照らし合わせて、重複したもの

を新規候補融合遺伝子とした。得られた新規候補融合遺伝子である LC-2/ad 株の WAC-SFMBT2 (Fig 10B), PC-9 株の ZSCAN22-CHMP2A (Fig 10C)について IGV 上で確認したところ、MinION データでそれぞれ 1 リードまたは 4 リードがカバーしていた。2つの候補は融合している状態での発現が示唆された。これらの融合遺伝子候補について、それぞれサンガーシークエンスを行った (Fig 10D)。2種の遺伝子ペアについても、融合が確かめられた。

3A-5. Novel isoform

A

B

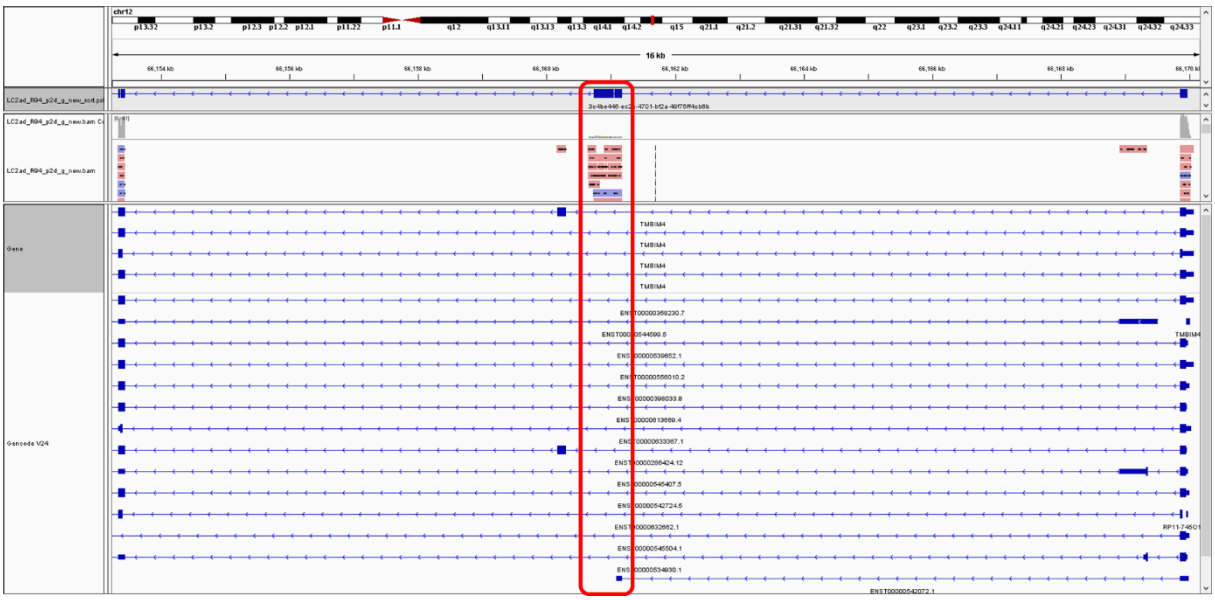
00000081 acccgggtttggtag. gtgactt.. att. actaaaaccatAAtctgaactgaagaagc. .gggctcactt. .acaaat 00000150 MinION
00000001 acggagcaaggcatgtttgaagagtacccgggtttggtagagtgacttctattcactaaaaccatgtgtctgaactgaagaagcttgggctcacttccacaaat 00000104 Sanger
20217402 acggagcaaggcatgtttgaagagtacccgggtttggtagagtgacttctattcactaaaaccatgtgtctgaactgaagaagcttgggctcacttccacaaat 20217299 hg38 reference

00000151 gagattgaaa. . .aatggcgtgtgttctaccctcacgattaaaaagg. aaagttctctgtgaatcataatcaaagaatcttttgacctgaa 00000237
00000105 gagattgaaaagtaatggcgtgtgttctaccctcacgattaaaaaggaaaagttctctgtgaatcataatcaaagaatcttt.gacctgaa 00000194
20185078 gagattgaaaagtaatggcgtgtgttctaccctcacgattaaaaaggaaaagttctctgtgaatcataatcaaagaatcttt.gacctgaa 20184989

00000238 ggatggaactcaactatttaaaggtgaagactgccagtcagtgtgaattgttt. aaaaagtccaacaattt. gggctggacaaccagtcaaatggttgaaataagaagatggg. aaaaatatgtcagg 00000362
00000195 ggatggaactcaactatttaaaggtgaagactgccagtcagtgtgaattgtttaaaaaagtccaacaatttgggggctggacaaccagtcaaatgg 00000289
20177930 ggatggaactcaactatttaaaggtgaagactgccagtcagtgtgaattgtttaaaaaagtccaacaatttgggggctggacaaccagtcaaatggttgaaataagaagatgggaaaaaatatgtcagg 20177803

00000363 atcag. aaaggcttctgctttctttaaaaa. taaaactac 00000400
20164399 atcagaaaaggcttctgctttctttaaaaaataaaactac 20164360

C



D

Fig 11. 検出された novel exon 候補
(A, B)MACC1 遺伝子、(C, D) TMBIM4 遺伝子。(A, C)IGV 上で見た MinION データ。赤枠で囲んだ部分ではデータベース上に exon がない。(B, D)サンガーシークエンスでデータベース上にない配列（赤枠)を検出した。

MinION データを用いて新規 isoform の探索を試みた。cDNA-Seq で取得した LC-2/ad 株データを LAST でゲノムにマップし、データベース（RefSeq, Genecode)上にはない exon を持つリードを探索した。IGV 上で確認し、両隣の exon をまたぐリードが複数存在するものを新規 isoform 候補として選んだ。72 の新規 isoform 候補でデータベース上にない exon が確認できた(Fig 11A, 11C)。このうち、がんとの関連が示唆される遺伝子の isoform についてサンガーシークエンスを行った(Fig 11B, 11D)。データベース上にはない exon が発現していることが示された。これらの新しい isoform の中には機能的にも重要な意味を持つものが含まれているかもしれない。

3B. tissue

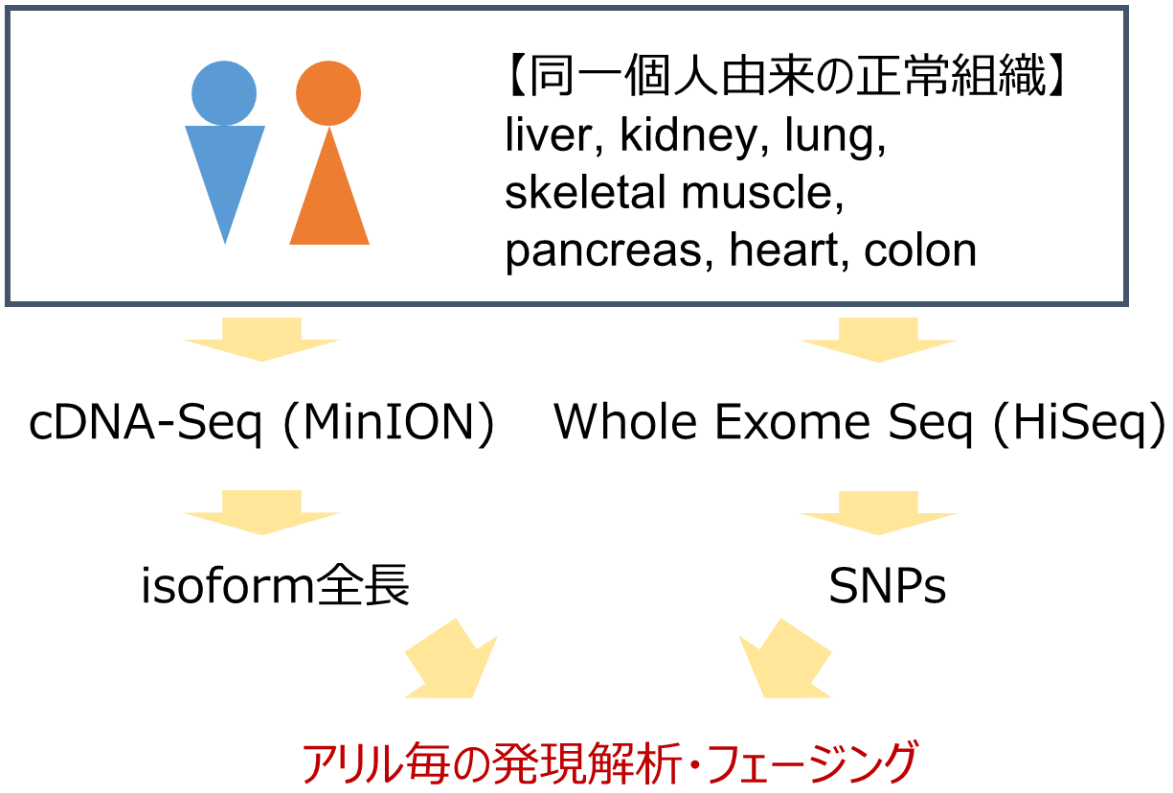


Fig 12. ヒト由来サンプルを用いた実験と解析の研究デザイン
同一個人由来の正常組織サンプルを用いて、各臓器に対する MinION cDNA-seq と HiSeq を用いた WES を行った。cDNA-Seq からは全長 isoform 情報を、WES からは SNP 情報を取得し、2つのデータを組み合わせてアリル毎の発現解析やフェージングを行った。

3B-1. Tissue のシークエンス結果

	Number of reads	Number of mapped reads	Number of detected SNPs	Number of hetero SNPs
Male	202,493,498	202,431,918	1,128,434	375,210
female	194,289,700	194,230,944	1,034,024	321,435

Table 10. The sequencing summary (tissue WES)
上段に男性心臓由来の、下段に女性心臓由来のサンプル WES の結果を示した。左から順に得られたリード数、マップされたリード数、検出されたリード数、解析に用いたヘテロ SNP の数を示した。

RIN score	Average length [bp]	Number of raw reads	Number of 2D reads	Tissue	RIN score	Average length [bp]	Number of raw reads	Number of 2D reads
7.5	383	1,877,492	950,526	Liver	8.6	575	2,408,040	1,704,811
6.1	300	2,169,687	1,038,648	Kidney	8.1	455	1,219,741	622,673
7.1	461	1,649,938	829,195	Lung	8.4	671	1,982,029	1,314,034
7.1	359	547,006	273,589	Lung2	-	-	-	-
6.7	1,377	967,999	714,931	Skeletal Muscle	6.8	524	1,140,199	756,317
8.6	429	1,136,345	563,300	Pancreas	6.3	496	750,299	436,896
6.4	348	1,596,992	776,182	Heart	8.6	523	1,235,122	781,718
5.6	300	809,755	434,140	Colon	6.9	395	2,030,195	1,215,465

Table 11. The sequencing summary (tissue cDNA-Seq)
中央（青い欄)よりも左側に男性由来、右側に女性由来の各組織のシークエンス結果を示した。各ブロックにおいて左から順に、RIN・平均リード長・総リード数・2D トータルリード数を示した。使用したフローセルは全て R9.4。

市販の同一個人に由来する各臓器の全エクソームシークエンス（WES)と、cDNA-Seq を行った。WES は男女それぞれの心臓由来サンプルを用いた。シークエンスして得られたデータは、BWA でマッピング後、picard による duplicate read 除去と GATK UnifiedGenotyper による snp call を行った（Table 10)。

cDNA-Seq は1人の男性由来の臓器 7 種（肝臓, 腎臓, 肺, 骨格筋, 膵臓, 心臓, 結腸）と、1人の女性由来の同じ臓器 7 種由来の RNA を各臓器で1枚のフローセルを用いてシークエンスランを行った（Table 11)。このうち、男性由来の lung については、2 回ランを行った。出発材料の RNA の RIN は 5-9 だった。RNA の分解が進んでいることが懸念された。シークエンス後の平均リード長も肺腺癌細胞株で行った結果よりも短いものが多く見られた。raw read の数は 50-200 万リードであり、クオリティの高い 2D リードは 25-180 万リードと、全体の 50%程度であった。

3B-2. アリル不均衡に発現する isoform

Male	p<0.05	p<0.01	補正 p<0.01	Tissue	Female	p<0.05	p<0.01	補正 p<0.01
2,356	59	38	15	**Liver**	3,758	185	125	38
2,695	35	24	8	**Kidney**	4,347	83	45	10
3,403	85	55	14	**Lung**	6,168	338	221	69
3,493	111	60	16	**Skeletal Muscle**	2,761	129	82	28
2,052	38	25	4	**Pancreas**	1,535	24	14	6
2,364	36	18	5	**Heart**	2,759	104	60	18
1,503	15	9	2	**Colon**	5,412	140	78	18

Table 12. アリル不均衡に発現している isoform の数
左側に男性同一個人由来、右側に女性同一個人由来の各臓器の cDNA-Seq を行い、アリル不均衡な発現を示す isoform の個数を示した。それぞれの性で左から順に、MinION リードでヘテロ SNP がカバーされた isoform の数、p<0.05 を満たす isoform 数、p<0.01 未満を満たす isoform 数、ボンフェローニ補正後に p<0.01 を満たす isoform 数を示した。

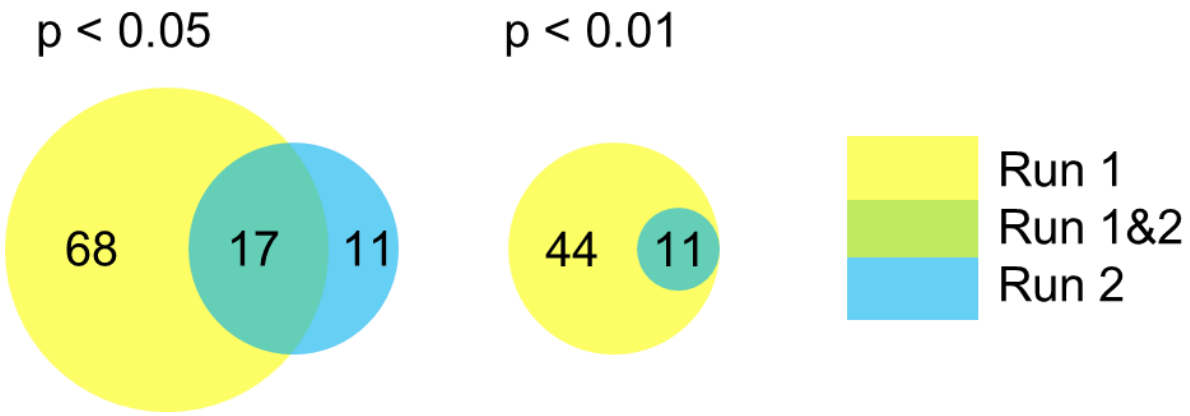


Fig 13. 閾値による重複 isoform 数の違い
黄色が男性由来肺サンプルのラン1のみで検出された isoform の数、青がラン2のみで検出された isoform の数、緑がラン1・ラン2の両方で検出された isoform の数。左側に p<0.05 で線引きしたときの数、右側は p<0.01 で線引きしたときの数。

染色体によって発現量が異なる isoform が存在することが知られている (Kim J, 2015)。得られた同一個人由来の MinION シークエンスデータとヘテロ SNP データの解析から、アリル間での発現に偏りを調べた。臓器ごとに各 isoform のリード数を数え、同一遺伝子の isoform 間で二項検定を行った。いくつかの閾値ごとに、p 値の条件を満たす isoform の数を調べた (Table 12)。男性の肺由来サンプル 2 ラン分のデータを用いて、閾値によって同一 isoform を検出できる数がどのように変化するかを調べた (Fig 13)。$p < 0.05$ のとき、重複が 17, ラン1のみが 68, ラン 2 のみが 11 であった。$p < 0.01$ のとき、重複が 11, ラン 1 のみが 44, ラン

2 のみは 0 と、ラン2で検出された isoform のすべてがラン1に含まれていた。p<0.01 のほうが、精度高く isoform の検出ができると考えた。また、ボンフェローニ法に基づく補正を行った。

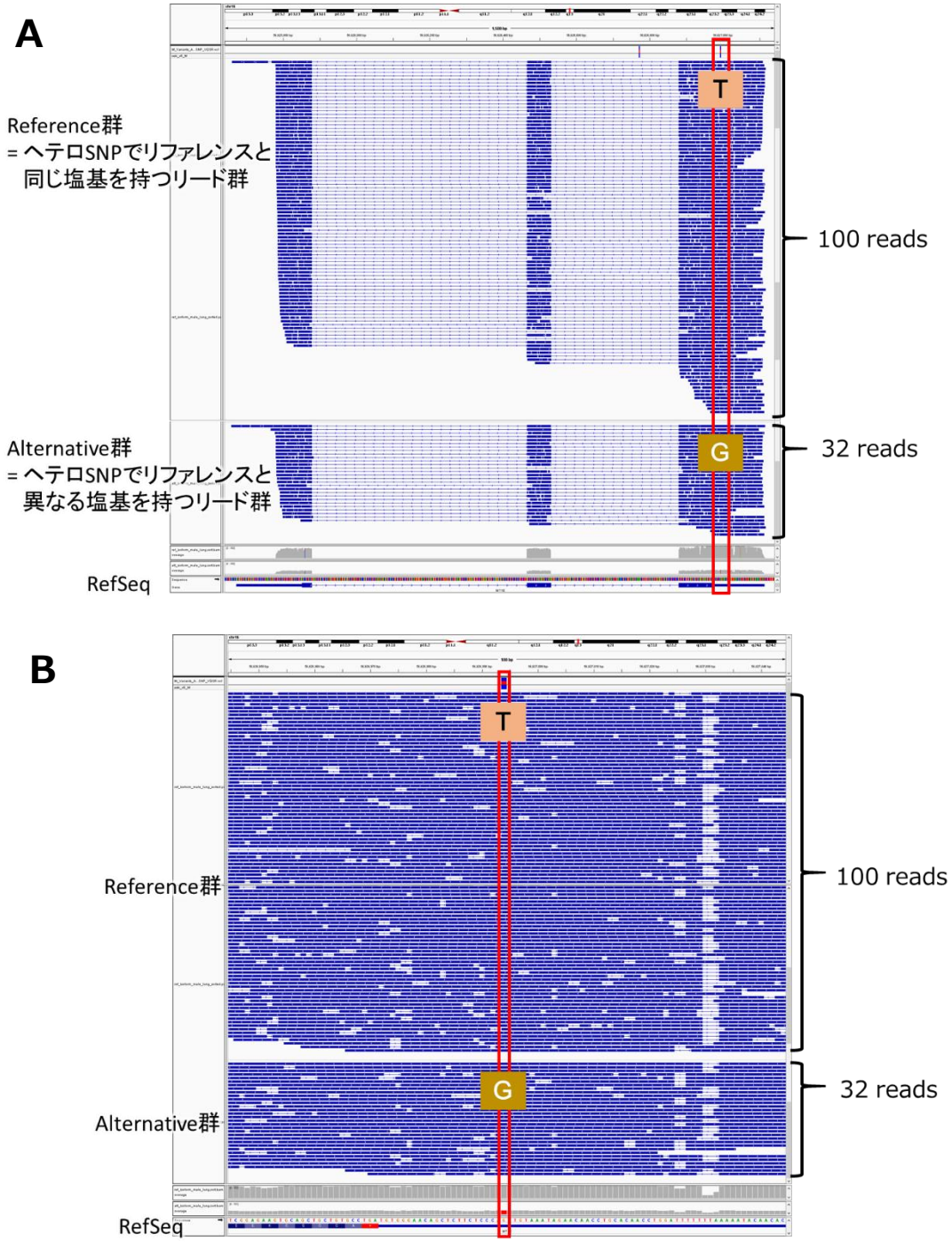


Fig 14. アリル不均衡に発現している isoform の検出
MT1E 遺伝子について、男性由来肺サンプル cDNA-Seq のリードと SNP データを、IGV 上で確認した。上段から順に WES による vcf ファイル, リファレンスと同じ塩基を持つリード群(Reference 群)の psl ファイル, リファレンスと異なる塩基を持つリード群 (Alternative 群)の psl ファイル, Reference 群 bam ファイルのカバレッジ, Alternative 群 bam ファイルのカバレッジ、RefSeq を示した。MT1E 遺伝子全長(A), SNP 付近の拡大図(B)を示した。

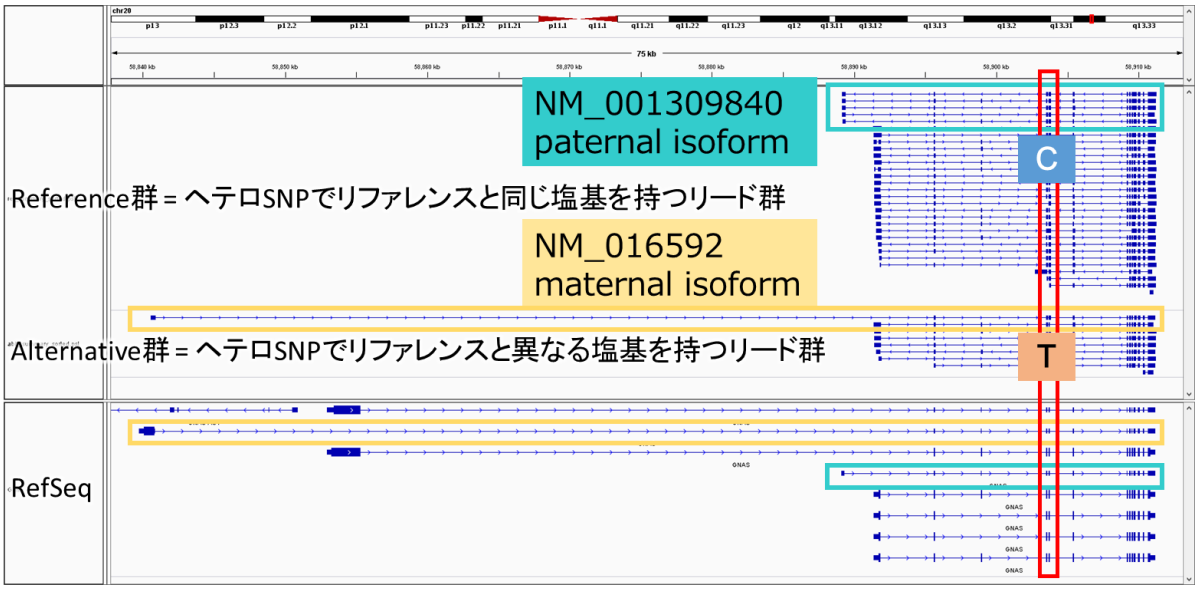


Fig 15. 片アリルのみに発現している isoform の検出
GNAS 遺伝子にマップされたリードの構造を IGV 上で確認した。上段から、hg38 と同じ塩基を持つリードの psl ファイル、hg38 とは異なる塩基を持つ psl ファイル、RefSeq isoform。

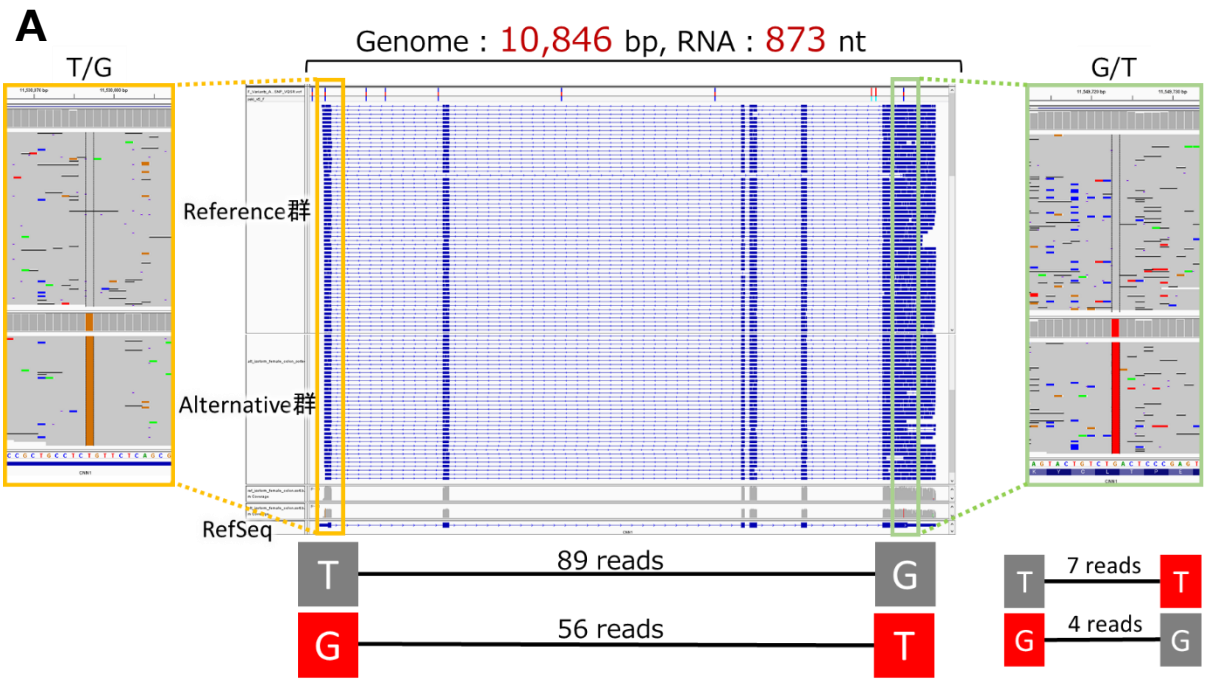
アリル間で発現が異なる isoform について、MinION データを IGV 上で可視化した。同一遺伝子間の isoform で二項検定を行い、補正した p 値が 0.01 未満のものを、アリル偏差遺伝子として数えた（Table 12)。例として、男性の肺由来データにおける MT1E 遺伝子を示した（Fig 14)。NM_175617 というアクセションナンバーを持つ isoform が、ヘテロ SNP 上でリファレンスと同じ塩基 T を持つ Reference 群では 100 リード、異なる塩基 G を持つ Alternative 群では 32 リード検出された。検出されたリード数で二項検定を行ったところ、ボンフェローニ補正を行った p 値は 2.94x10e-6 となった。

GNAS 遺伝子は、父方由来(paternal)の染色体から発現する isoform と、母方由来(maternal)の染色体から発現する isoform は違う構造をとることが知られている(Bastepe M, 2007)。MinION で得られた男性由来データから paternal な isoform (NM_001309840)と maternal な isoform (NM_016592)がそれぞれ検出できた(Fig 15)。これらの isoform レベルでの解析は MinION のデータを用いることなしでは一般に困難なものであったと考えられた。

3B-3. フェージング

Tissue / male	複数のヘテロSNPがカバーされているisoformの数	ゲノム上の距離 [bp] (最大距離の平均)	RNA上の距離 [bp] (最大距離の平均)
Liver	628	11,321	668
Kidney	660	10,884	695
Lung	935	9,499	664
Skeletal muscle	1,228	26,849	1,289
Pancreas	519	11,676	633
Heart	642	10,584	715
Colon	362	10,884	612
Tissue / female	**複数のヘテロSNPがカバーされているisoformの数**	**ゲノム上の距離 [bp] (最大距離の平均)**	**RNA上の距離 [nt] (最大距離の平均)**
Liver	1,032	11,107	805
Kidney	795	12,881	684
Lung	1,797	13,527	874
Skeletal muscle	653	10,137	613
Pancreas	376	9,673	577
Heart	704	12,217	664
Colon	985	10,310	748

Table 13. 各臓器上でのフェージングできた isoform の数



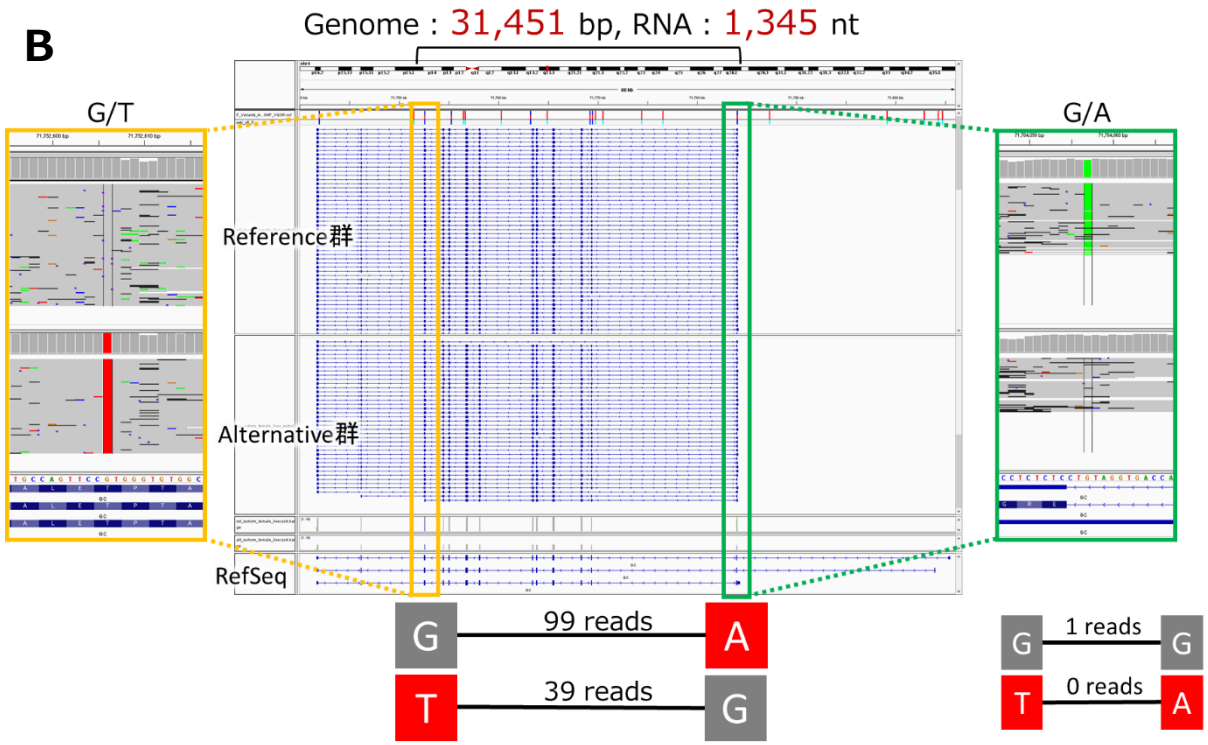


Fig 16. CNN1 遺伝子(A), GC 遺伝子(B)のフェージング
WES によって検出された SNP の位置と、Reference 群または Alternative 群がカバーしている塩基のパターン(灰色はリファレンスゲノムと同じ配列)及びリード数を示した。右下には対となる塩基のパターン及びそのリード数を示した。

MinION の長鎖解読特性を活かしてフェージング(haplotype phasing)を行った。1つの isoform 上に複数のヘテロ SNP が含まれているとき、ロングリードの情報を用いると各 isoform がどの塩基の組み合わせで発現しているのか容易に調べることができた。個人(性別)毎、臓器別にフェージングできた isoform の数を Table 13 に示した。ゲノム上の長い距離を1本の MinION リードでカバー出来ていると考えた。Table 12 の補正 $p<0.01$ のリストの中からいくつか選んで IGV 上で確かめた (Fig 16)。同定された塩基パターンと対になる塩基パターンを持つリード数も調べた。対になるパターンは、MinION の読みとりエラーの範囲内だと考えた。ゲノム上で 10kb 以上離れた SNP でも、MinION のデータを用いることでフェージング出来ることが示唆された。

4）結論と今後の展望

本研究では、MinION における RNA-Seq について、異なるプロトコール間での比較や、その長鎖解読特性を活かした解析結果を示してきた。cDNA-Seq では、cell line でプライマーの修飾について確認を行った。cDNA の合成・増幅の際には修飾なしのプライマーを用いることで解析に必要なリード数が得られることを示した。各細胞株で、少量の RNA から実施可能な RNA-Seq である cDNA-seq を行った。LC-2/ad 株では direct RNA-Seq を行った。MinION でのシークエンスはインプットしたサンプルの分子数に比例してシークエンス出来ることを示した。cDNA-Seq のデータについて、アライメントツールについての確認を行った。LAST でのアライメントが、ロングリードの解析を行う本研究の目的に合致していると考えた。direct RNA-Seq は PCR バイアスなしにシークエンスが可能であることを示した。cDNA-Seq、direct RNA-Seq について、HiSeq データと遺伝子発現情報についての比較を行った。発現情報は、サンプル調製の方法が途中まで同じ cDNA-Seq と SMART-Seq の相関が高かった。また、direct RNA-Seq と TruSeq でサンプル調製したデータにも高い相関が見られた。direct RNA-Seq は PCR バイアスなくシークエンスを行っているため、TruSeq で調製した HiSeq データ、SMART-Seq、MinION での cDNA-Seq よりも真に近い発現を示したのかもしれない。MinION の長鎖解読特性を活かして、融合遺伝子の検出を行った。LC-2/ad 株で既知の融合遺伝子 CCDC6-RET や新規候補融合遺伝子 WAC-SFMBT2, PC-9 株で ZSCAN22-CHMP2A が MinION データとサンガーシークエンスで検出できた。MinION データをゲノムにマップすることで、データベース上にない exon が存在することを示し、新たな isoform が存在する可能性を示唆した。tissue の cDNA-Seq では、補正 $p<0.01$ の確からしさでアリル不均衡に発現している遺伝子が 2-69 個検出することができた。362-1,797 個の isoform については、1本のリードで複数のヘテロ SNP をカバーしているため、ハプロタイプフェージングが行えることを明らかにした。

MinION のシークエンスは煩雑な手技を用いることもなく、空間的制約の大きい機器や特殊な機器も使わない。小規模の研究機関や臨床現場、フィールドでのシークエンスも可能である。本研究では特に、MinION を用いた RNA-Seq を行うことで様々な解析が可能であることを示した。1回のランで得られるデータだけで遺伝子発現量の推定や融合遺伝子の検出、isoform の同定、ハプロタイプフェージング等幅広い知見を得ることが可能であった。MinION シークエンサーの活用で、より幅広い分野の研究者が RNA-Seq 解析を行うことができるようになれば、様々なオーミクス解析が一層進展することが期待できる。

一方で、現段階の MinION では不十分な用途も存在する。現時点での MinION のシークエンス精度は最大 95%程度と HiSeq に劣る。1 度のランで得られるデータ量も 1 Gb 程である。MinION での精度、データ量の不足を補うために HiSeq データと組み合わた解析も行われている（Weirather J L, 2017）。direct RNA-Seq ではインプットに polyA 付 RNA を 500 ng も用いなければならないことから、微量サンプルでのシークエンスは不可能である。しかし、シークエンサーの改良は進展が著しく、これらの課題も暫時改善されていくことが期待される。

謝辞

本研究には、関真秀先生、鈴木穣先生を筆頭に、研究室内の皆様に大変お世話になりました。特に、実験においては今村聖実様、阿部佳澄様に、解析においては若栗浩幸様、鳥谷恵子様、堀内映実様に多くの手助けをして頂きました。また、国立がん研究センターの鈴木絢子先生には肺腺癌細胞等のデータや助言を頂きました。高次生命情報解析分野の Martin Frith 教授には、LAST を用いた解析を行うためのパイプラインを構築して頂きました。

参考文献

Bastepe, M. (2007). The GNAS Locus: Quintessential Complex Gene Encoding Gsalpha, XLalphas, and other Imprinted Transcripts. Curr. Genomics 8, 398-414.

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., and Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun. 8, 16027.

Castro-Wallace, S.L., Chiu, C.Y., John, K.K., Stahl, S.E., Rubins, K.H., Mcintyre, A.B.R., Dworkin, J.P., Lupisella, M.L., Smith, D.J., Botkin, D.J., et al. (2016). Nanopore DNA Sequencing and Genome Assembly on the International Space Station. bioRxiv Prepr. 1-12.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333-351.

Hamada, M., Ono, Y., Asai, K., Frith, M.C., and Hancock, J. (2017). Training alignment parameters for arbitrary sequencers with LAST-TRAIN. Bioinformatics 33, 926-928.

Kent, W.J. (2002). BLAT - The BLAST -Like Alignment Tool. Genome Res. 12, 656-664.

Kielbasa, S.M., Wan, R., Sato, K., Kiebasa, S.M., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison Adaptive seeds tame genomic sequence comparison. 487-493.

Kim, D., and Salzberg, S.L. (2011). TopHat-Fusion: An algorithm for discovery of novel fusion transcripts. Genome Biol. 12.

Kim, J., Bretz, C.L., and Lee, S. (2015). Epigenetic instability of imprinted genes in human cancers. Nucleic Acids Res. 43, 10689-10699.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Matsubara, D., Kanai, Y., Ishikawa, S., Ohara, S., Yoshimoto, T., Sakatani, T., Oguni, S., Tamura, T., Kataoka, H., Endo, S., et al. (2012). Identification of CCDC6-RET Fusion in the Human Lung Adenocarcinoma Cell Line, LC-2/ad. J. Thorac. Oncol. 7, 1872-1876.

Oikonomopoulos, S., Wang, Y.C., Djambazian, H., Badescu, D., and Ragoussis, J. (2016). Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci. Rep. 6, 31602.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 9, 171-181.
Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228-232.
Rand, A.C., Jain, M., Eizenga, J.M., Musselman-Brown, A., Olsen, H.E., Akeson, M., and Paten, B. (2017). Mapping DNA methylation with high-throughput nanopore sequencing. Nat. Methods 1-6.
Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. 31, 1009-1014.
Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nat Meth advance on, 1-7.
Suzuki, A., Makinoshima, H., Wakaguri, H., Esumi, H., Sugano, S., Kohno, T., Tsuchihara, K., and Suzuki, Y. (2014). Aberrant transcriptional regulations in cancers: Genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. Nucleic Acids Res. 42, 13557-13572.
Suzuki, A., Suzuki, M., Mizushima-Sugano, J., Frith, M.C., Makałowski, W., Kohno, T., Sugano, S., Tsuchihara, K., and Suzuki, Y. (2017). Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. Dna Res. 0, 1-12.
Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K.F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research 6, 100.