課題番号　　：25指102
研究課題名　：「第三世代シークエンサーPacBio RSを活用した糖尿病・代謝疾患の病態解析」
主任研究者名：安田和基
分担研究者名：南茂隆生
キーワード　：ロングリード、ハプロタイプ決定、選択的スプライシング、DNA 修飾
研究成果　　：本研究では、第三世代シークエンサーである PacBioRS を用いて、技術開発とともに糖尿病関連疾患の病態解明を行う。このシステムは、A. 数 kb にわたる長いリード配列、B.高い GC 含量の領域も解析可能、C.Kinetics 情報（特に波形間の inter-pulse duration：IPD）の取得により塩基修飾情報も得られる可能性、など、これまでの次世代シークエンサー（NGS）にない特徴がある。そこで、A によりハプロタイプの決定や allelic expression の解析、ゲノム構造異常の解析、B によりプロモーター領域や繰り返し配列を含めたゲノム解析、C により細胞分化や糖尿病の病態や合併症に関する DNA 修飾検出、などを試みる。本技術はまだ発展途上の部分があり、特に哺乳類ゲノム・RNA についての報告はきわめて限定的であるため、本研究は既存の解析方法と比較する「feasibility study」と、新たなアプリケーションの開発を行う「potential の開拓」という、2 つの方向性をもつ。

初年度の進捗状況は以下のようになる。

［1］　本システムの性能を検証するために、ラムダファージ DNA からテンプレートを作成して、シークエンスを行った。10kb のサイズで DNA を断片化して、テンプレートを作成し、「Continuous Long Reads(CLR)」モードで解析を行なった。データは、1 次解析（ベースコール）の後、SMRT Portal 上で 2 次解析を行い、「BLASR（Basic Local Alignment with Successive Refinement）」というツールを用いてレファランス配列へマッピングを行った。その結果、「リード長」は 10～15kbp まで分布しており、両端の adaptor にはさまれた「サブリード」長は 2～5kbp 程度が多く、「ロングリード」の特性が確かめられた。マップされたサブリードについては、平均 accuracy が 81.6％と、本システムの特徴として知られている通り、第二世代 NGS よりやや低い。しかし、ケミストリーの特性に依存して特定の塩基でエラーを生じる他の NGS と異なり、いわゆるエラーはランダムに生じることから、depth を十分得られれば解決できる。

以上より、第三世代 NGS としての PacBioRS の特性および有用性を確認することができた。

［2］　他の方法（第二世代 NGS など）に対する優位性を軸とした研究計画の吟味

本機器の有効性と限界について、共同研究者であるトミーデジタルバイオロジーズ（以下 TDB）社（実験部門、解析部門）と情報交換を行った。その結果、本テクノロジーに適した DNA 領域濃縮法は、まだ開発途上であること、定量性が求められるプロジェクト（定量的発現解析のための RNA-Seq など）については、現時点で第二世代 NGS に対する優位性は乏しいこと、真核生物の DNA 修飾（特にメチル化）の IPD パターンは複雑でまだレファランスデータ収集が必要な段階であること、が判明した。以上をうけて、予定していた各サブテーマについて再吟味を行った結果、定量を目的とした RNA-Seq については当面 pending とし、目的ゲノム領域の濃縮方法についての検討を開始した。また、解析用サンプルとして、JCRB 生物資源バンクや米国の企業経由で市販されている初代培養ヒト肝細胞など、エピゲノム解析に耐えうるヒト組織の入手方法を検討した（詳細は分担南茂の項）。

本研究により、ヒト遺伝因子の機能解明、エピゲノム制御や DNA 修飾の検出が行われれば、糖尿病・代謝疾患の病態の解明や個別化医療の実現へ、一歩近づくことが可能になると期待される。

Subject No. : 25D02

Title : The application of the third generation sequencer, PacBioRS, to the researches on diabetes mellitus and metabolic disease.

Researchers : Yasuda K, Nammo T, and others.

Key word : the third generation sequencer, long read haplotype phasing, alternative splicing DNA modification

Abstract : In this project, we are going to utilize PacBioRS, a new platform called “the third generation sequencer”, for the diabetes research as well as to make some technical innovations. This system has several unique features over other next-generation sequencers(NGS), including extra-long reads(over several kb), robust base calls for GC-rich or repetitive sequences, and possible applications for the analysis of DNA modification. However, this technology is still on the progress, and so for there have been only a limited number of reports on its application to mammalian genomes and RNA. Therefore, we will perform a ‘feasibility study’ in which we will compare the performance of PacBioRS with existing technologies using known sequences, and an ‘innovative study’ in which we will try new applications of this technology to the genomic and epigenomic studies of diabetes mellitus.

1) In order to assess the performance of PacBioRS, we sequenced lambda phage DNAs. Templates were prepared from 10kb fragments on average, and run was done by Continuous Long Reads (CLR) modes. After base calling, data were analyzed on SMRT Portal, and mapped to the reference sequence using of the lambda phage BLASR（Basic Local Alignment with Successive Refinement）algorithm.

The read length was extended up to 10～15kb, and most of the subread length was 2～5kb, thus the striking ‘long read’ feature was confirmed. The average base accuracy was 81.6%, which was lower than commonly used ‘NGS’; however, since base call errors happened randomly and were not dependent on chemical or structural features of some specific bases, we think we can expect reliable base calls if the coverage was reasonably deep.

2) We made discussion several times with scientists from Tomy Digital Biology Ltd., over the advantages and disadvantages of PacBioRS, and reached a couple of conclusions. Although DNA enrichment of targeted regions is critical for the success of this technology when we deal with large genomes such as human, there are no established methods for this step yet. Quantitative experiments such as tag counting of RNA-Seq are better performed, at least for the present, by other NGS than PacBio. DNA modification of mammalian genomes is more complicated than previously supposed and does not exhibit consistent patterns of inter-pulse duration(IPD) by PacBio sequencing. Therefore, we modified the whole plan, and started to develop efficient methods for DNA enrichment of targeted regions, and to collect reference data for methylated DNA. Collaborator Nammo’s group collected the list of available human tissues or cells suitable for genomic (haplotype phasing), epigenomic (DNA modification) and transcriptomic (allelic expression and alternative splicing) analyses using PacBio.

第三世代NGS：PacBio RSの特性とサブテーマの吟味

（SMRT bell™ sequencing）

第二世代NGSにない特徴	Long read（数Kb以上）	GC rich領域も解析可能	塩基修飾検出可能性
ヒトを中心とした高等動物ゲノムへの応用（挑戦的分野）	ハプロタイプ決定 ミトコンドリアDNAの多型配列決定 Allelic expression 新規RNA同定（splice variants含む）	プロモーター領域を含めた配列変化の解析	DNAメチル化 メチル化以外のDNA修飾（8-oxo-グアニンなど）
糖尿病関連病態における未解決課題へのソリューション	ヒト糖尿病遺伝因子の機能解明 エピゲノム制御の寄与の検証 糖尿病発症・進展における塩基修飾の解明		

下線部分：本システムの特性を踏まえて、改めて優先課題と位置づけたもの

PacBioRSの特性の検証

（ラムダファージDNA、インサート10kbpで作成、
CLRモード、「120」分でデータ取得）

図1：リード長分布

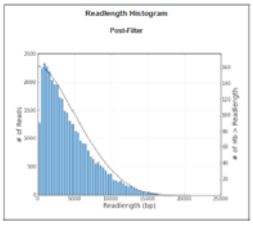


図2：サブリード長分布

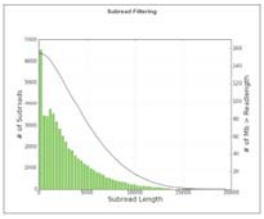


図3：Read Quality（RQ）

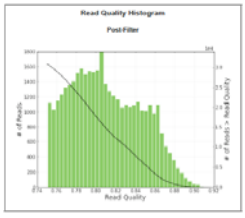


図4：Accuracy

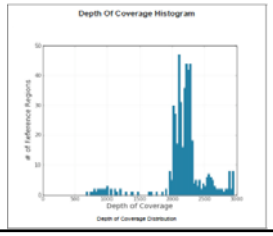
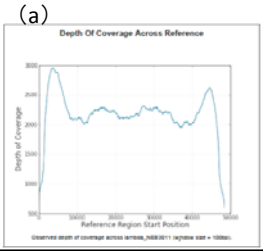
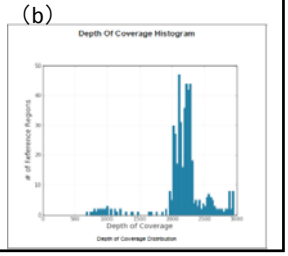


図5：カバー率分布

（a）



（b）



課題番号　　：25指102
研究課題名　：「第三世代シークエンサーPacBio RSを活用した糖尿病・代謝疾患の病態解析」
分担研究課題名：「研究の総括、及びPacBio RSの特性を活かしたアプリケーションの探索と2型糖尿病研究への応用」
分担研究者名：安田和基（協力研究者；西村渉、宇田川陽秀、舟橋伸昭、川口美穂、ほか）
キーワード　：ロングリード、DNA濃縮、DNA修飾
研究成果　　：

［1］　第三世代シークエンサーPacBioRS の性能を検証するために、ラムダファージ DNA から Template を作成して、シークエンスを行った。

PacBioRS で用いられる「SMRTbell テンプレート」については、「Continuous Long Reads(CLR)」、「Circular Consenseus Sequencing（CCS）」という2つのモードがある。前者は、大きなサイズのインサート（2kb 以上が推奨されている）を one pass で読む方法で、PacBioRS の long read の特徴を活かした方法である。一方、CCS は比較的短いサイズ（それでも 250bp-2kb と一般の NGS よりは長い）のインサートを multiple passes で読み込むことにより、エラーをカバーして正確度を上げる方法である。我々は長いインサートを用いた long read の特性を検証するために、10kb のサイズで DNA を断片化し、バイオアナライザーにより断片化したサイズを確認したのち、テンプレートを作成した。データは、1次解析（ベースコール）の後、SMRT Portal 上で2次解析を行い、レファランス配列へマッピングを行う、「BLASR（Basic Local Alignment with Successive Refinement）」というツールを用いて解析した。

「リード長」は、10～15kbp まで分布しており、「ロングリード」という本システムの有用性が示された。リードのうち、両端の adaptor にはさまれた「サブリード」も、2～5kbp 程度が多かった。リードごとの頻度をあらわす「Read Quality（RQ）」も良好であった。マップされたサブリードの accuracy は平均 81.6%であったが、いわゆるエラーはケミストリーに依存せず、depth を十分得られれば問題ないと期待される。

［2］　他の方法（第二世代 NGS など）に対する優位性を軸とした研究計画の吟味

共同研究者であるトミーデジタルバイオロジーズ（以下 TDB）社の本機器担当者（実験部門、解析部門）と会合を重ねた結果、以下のことが明らかになった。

1)　現時点では、他の NGS と比較してリード数が少ないため、ヒトゲノム全域を対象とするのは難しく、目的領域の十分な濃縮（エンチッリメント）が必要。ただし、本テクノロジーに適した DNA 領域濃縮法は、まだ開発途上である。
2)　リード数が多くないこととも関連して、定量性が求められるプロジェクト（RNA-Seq による定量的発現解析など）については、現時点では第二世代 NGS に対する優位性は乏しい。
3)　ほかの NGS に比較して、一度のリードではエラー率が高いと報告されているが、充分な depth が得られれば配列は確定できる。
4)　DNA 修飾（特にメチル化）の同定については、真核生物では DNA 修飾がこれまでの予想以上に多彩であり、もっとも頻度が高くかつ機能的に重要と思われる「シトシンメチル化」だけをとっても、単一の IPD パターンを示さず、その原因が前後の塩基配列によるのかどうかも未解決の課題で、まだ十分なデータが蓄積されていない。

以上をうけて、予定していた各サブテーマについて優先順位の再吟味を行い、目的ゲノム領域の濃縮に対して、既存の試薬のプロトコールをベースに、本テクノロジーに適したカスタマイズが可能かどうかの検討を開始した。

課題番号　　：25指102
研究課題名　：2型糖尿病感受性座位におけるハプロタイプと組織特異的遺伝子発現変化の意義の検討
主任研究者名：安田　和基
分担研究者名：南茂　隆生

キーワード　：2 型糖尿病、第三世代シークエンサー、ターゲットリシークエンシング、全長転写産物解析、疾患感受性多型、ヒト組織解析
研究成果　　：

ゲノムワイド相関解析（GWAS）によって、数年前から多数の 2 型糖尿病感受性一塩基多型（SNP）が同定されてきた。私たちは、2 型糖尿病と最も関連の強い *TCF7L2* 遺伝子の SNP（rs7903146）が、クロマチン構造の変化と共にエンハンサー活性の変化をともなって遺伝子発現に影響を与えることを示してきた。ところで、疾患感受性座位には複数の cis 調節領域がしばしば存在し、このような領域にも一塩基多型が認められる場合がある。よって、ハプロタイプ構造を基本とした疾患感受性領域内における cis 調節作用の総和を考慮すれば、遺伝子発現/疾患感受性と SNP との関連性をより正確に理解できるかもしれない。第三世代シークエンサー「PacBio RS」は長いリード配列が特徴であることから、ハプロタイプの決定や全長にわたる mRNA の配列決定が可能となり、有用であると期待される。

特に前者（ターゲットリシークエンシングによるハプロタイプ決定）を試みる場合には、目的のゲノム領域を濃縮することが必要であり、実際に 2010 年ころから開発が試みられてきた。しかし、①ライブラリのインサートサイズから期待されるシークエンス平均長が得にくい、②DNA サンプルの開始量がマイクログラムオーダーと多量である、といった理由から未だ発展途上の技術であり製品化には至っていない。50Mb までのヒトの任意のターゲットゲノム領域を抽出・濃縮する試薬が製品化されているものの、PacBio RS 用に最適化されておらず、ほ乳類を対象とした PacBio RS によるターゲットリシークエンシングは確立されていないといった現実がある。

後者、すなわち転写産物の全長シークエンシングについては、最近になり PacBio RS の有用性が報告されるようになってきた（PNAS 110:E4821-E4830, 2013 など）。ヒト遺伝子転写産物全長の中央値は約 2,500bp であるが、PacBio RS を用いれば多くの遺伝子の cDNA 全長をカバーできるようになってきたことが示されつつある。この結果、ショートリードを用いる次世代シークエンサーのみの解析と比較して、スプライシングアイソフォームや長鎖ノンコーディング RNA に関するより正確かつ詳細なデータを得られることが明らかとなっている。さらに、PacBio RS を用いて解析対象の細胞・組織ごとの正確なゲノムアノテーションを行った場合にのみ、ショートリードによるスプライシングアイソフォームごとの転写産物定量結果を正しく評価し得ることも明らかとなった。すなわち、真の意味で妥当な遺伝子発現解析を実現するためには、安定した「シークエンスの超ロングリード化」が不可欠であり、今後の技術発展に負うところが大きい。

上述のように、発展途上の事項については新しい成果が期待されるところであるが、2 型糖尿病の遺伝因子として同定された *TCF7L2*、*KCNQ1*、*CDKAL1* などについて、糖尿病関連臓器（膵、肝、筋、脂肪など）における詳細な基本情報の把握は重要である。JCRB 生物資源バンク（旧ヒューマンサイエンス研究資源バンク）には分譲を受けることのできるヒト肝組織が約 40 検体保管されている。ほとんどのサンプルが 60～70 歳代の日本人ドナー由来（男女比は約 2:1）であり、サンプル重量も 200 mg 以上あることから有用性が高い。また、初代培養ヒト肝細胞については、米国で調製された少なくとも 80 種類のロットが入手可能であり（男女比は約 1:1）、一検体につき 500～1000 万細胞/バイアルの細胞数となっている。人種の内訳は白人約 80%、白人以外が約 20%であるために、マイナーアリル頻度がアジア人に少ない多型について解析する場合には有用である。このように、解析目的に応じて検体の選択が可能であり、多様なデータを得ることも可能であるが、PacBio RS の現状でのスペックを考えると、cDNA 解析による網羅的なスプライシングアイソフォームの同定、およびこれらの発現レベル解析などが先ず以て重要かつ現実的であると考えられる。

研究発表及び特許取得報告について

課題番号： 25指102

研究課題名： 第三世代シークエンサーPacBio RSを活用した糖尿病・代謝疾患の病態解析

主任研究者名： 安田　和基

論文発表

論文タイトル	著者	掲載誌	掲載号	年
Genome-wide association study identifies three novel loci for type 2 diabetes.	Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, Peng C, Hu C, Ma RC, Imamura M, Iwata M, Tsunoda T, Morizono T, Shojima N, So WY, Leung TF, Kwan P, Zhang R, Wang J, Yu W, Maegawa H, Hirose Hi DIAGRAM Consortim, Kaku K, Ito C, Watada H, Tanaka Y, Tobe K, Kashiwagi A, Kawamori R, Jia W, Chan JC, Teo YY, Shyong TE, Kamatani N, Kubo M, Maeda S, Kadowaki T.	Hum Mol Genet	23(1): 239-246	2014
Poor responses to tyrosine kinase inhibitors in a child with precursor B-cell acute lymphoblastic leukemia with SNX2-ABL1 chimeric transcript.	Masuzawa A, Kiyotani C, Osumi T, Shiota Y, Iijima K, Tomita O, Nakabayashi K, Oboki K, Yasuda K, Sakamoto H, Ishikawa H, Hata K, Yoshida T, Matsumoto K, Kiyokawa N, Mori T.	Eur J Haematol	92(3): 263-267	2014

研究発表及び特許取得報告について

Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants.	Pasquali L, Gaulton KJ, Rodríguez-Seguí SA, Mularoni L, Miguel-Escalada I, Akerman I, Tena JJ, Morán I, Gómez-Marín C, van de Bunt M, Ponsa-Cobas J, Castro N, Nammo T, Cebola I, García-Hurtado J, Maestro MA, Pattou F, Piemonti L, Berney T, Gloyn AL, Ravassard P, Gómez-Skarmeta JL, Müller F, McCarthy MI, Ferrer J	Nat Genet	46(2): 163-143	2014

学会発表

タイトル	発表者	学会名	場所	年月
Type 2 diabetes-associated SNPs with in KCNQ1 gene modulate the affinity of the locus for DNA-binding factors	Hiramoto M, Udagawa H, Watanabe A, Kawaguchi M, Nishimura W, Nammo T, Yasuda K.	Annual Scentific Meeting, American Diebtes Association	Chicago, USA	2013年6月
ゲノム網羅的解析を用いた、高脂肪食摂取による膵島の代償機序の解明	南茂隆生、宇田川陽秀、川口美穂、衛藤弘城、上番増喬、平本正樹、西村渉、安田和基	第56回日本糖尿病学会年次学術集会	熊本	平成25年5月
教育講演：糖尿病診療に遺伝情報は役立つのか	安田和基	第13回糖尿病・情報学会	徳島	平成25年8月
ゲノム網羅的解析を用いた高脂肪食摂取による膵島の代償機序の解明	南茂隆生、宇田川陽秀、川口美穂、衛藤弘城、上番増喬、平本正樹、西村渉、安田和基	第3回NGS現場の会	神戸	平成25年9月
2型糖尿病遺伝因子はどこまでわかったか	安田和基	第24回新潟糖尿病臨床研究会	新潟	平成25年10月
糖尿病の遺伝子と臨床	安田和基	第28回糖尿病研修セミナー	神戸	平成25年12月

研究発表及び特許取得報告について

糖尿病の遺伝子解析の現状と課題−遺伝・環境相互作用も含めて	安田和基	Scientific Exchange Meeting 糖尿病における遺伝・環境の相互作用	宇都宮	平成26年2月
教育講演：遺伝子異常と糖尿病	安田和基	第57回日本糖尿病学会年次学術集会	大阪	平成26年5月
次世代シークエンサーを用いたゲノム、エピゲノム研究	安田和基	第57回日本糖尿病学会年次学術集会	大阪	平成26年5月
膵島のゲノム網羅的解析による糖尿病発症機序の考察	南茂隆生、宇田川陽秀、川口美穂、舟橋伸昭、上番増喬、平本正樹、西村渉、安田和基	第57回日本糖尿病学会年次学術集会	大阪	平成26年5月

その他発表(雑誌、テレビ、ラジオ等)

タイトル	発表者	発表先	場所	年月日
該当なし				

特許取得状況について　※出願申請中のものは(　　)記載のこと。

発明名称	登録番号	特許権者(申請者)(共願は全記載)	登録日(申請日)	出願国
該当なし				

※該当がない項目の欄には「該当なし」と記載のこと。
※主任研究者が班全員分の内容を記載のこと。