平成 26 年度 先端的計算科学研究プロジェクト 成果報告書
大規模情報処理パイプラインによる次世代シークエンサーによるエピゲノム解析

大川恭行（九州大学医学研究院先端医療医学部門エピジェネティクス分野）

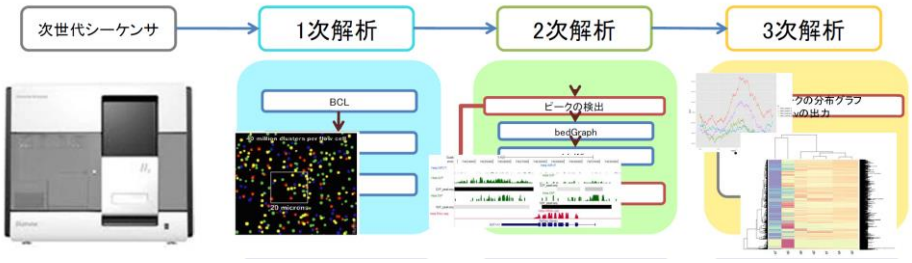
1. 背景

ゲノム DNA は、デオキシリボ核酸で構成された巨大分子である一方で、4 つの塩基によって描かれた巨大な生命の設計図である。このゲノム上の DNA やヒストン等のタンパク質がメチル化、アセチル化等の化学修飾を受けることで、遺伝子の発現状態が調節される。生物個体を構成する個々の細胞がもつゲノム修飾の総体をエピゲノムと呼ぶ。受精から発生を経て老化に至るまで遺伝子が適切に働くためには、エピゲノムの制御が重要である。ゲノムプロジェクトの完了により DNA の塩基配列情報の解読がなされた今、エピゲノム情報解析の時代が到来している。これまでは、ゲノム情報はヒト固有の 27 億文字の塩基配列、テキストファイルにして 2GByte 程度の情報をもとに、エピゲノム解析では、その情報がいかに使われているかという視点での解析が行われ、27 億文字のどの箇所が、いつ、どのように使われているかという、多面的かつ網羅的な解析が行われていた。現在明らかになっているエピゲノム情報だけでもヒストン修飾が 150 種類以上、DNA 修飾が 6 種類以上と多岐にわたり、60 兆個の細胞がこれらのさまざまな修飾を受けているといわれる。そこで個々の配列、また細胞毎の異なる修飾の組み合わせ数を考慮するとなれば、エピゲノム情報を構成する空間はきわめて広大であることが理解できる。これまでは、この広大な情報の空間のごく一部分しか見ることができなかったが、次世代シークエンサーの登場により、ゲノム情報に匹敵するデータサイズの個別のエピゲノム情報を短時間で取り出すことが可能になった。一方で現在、得られる情報量の爆発的な増加に伴う計算資源や解析データを保持するためのストレージ資源の不足は、本分野において大きな問題となっている。特に、九州大学も参画している国際エピゲノムデータのデータ公開が３月より本格化し、その件数は一気に５０００件を突破した。本データ量は、概算で 2.5PB に達する。本プロジェクトでは、医学生物学研究者が優れた計算資源であるスパコンを活用することで効率の良いエピゲノム解析を行うモデル系の構築を目指した。

2. 計算パイプラインの構築とその実行環境の整備

4 年目にあたる本年度では、外部データとの対応解析並びにデータ解析パイプラインの安定稼働、計算パイプラインのバージョンアップを図った。九州大学内には現在 5 台の次世代シークエンサーと呼ばれるゲノム、エピゲノム解析装置が設置されている。今年度はこのうち３基に解析が集中して行われた。。これら装置は P&P プロジェクト A タイプ（ゲノム、エピゲノム解析拠点構築プロジェクト）により稼働支援が行われ、公募研究も 7 件展開された。今年度は独自の試みとして、新学術領域「クロマチン動構造」、「幹細胞老化」

「性差」「転写サイクル」の４領域の解析が主に解析され国内において屈指の解析拠点となっている。海外からは、米国、シンガポール、韓国より解析依頼を受けており、現在解析を行っている段階である。また、九州大学における大規模情報処理解析拠点の構築へのウェット面での支援は継続的に行われており、ドライから実験科学であるウェット面まだ一貫したサポートを得ている。一方で、我々医学研究院シークエンスセンターとして、外部解析の依頼が今年度は大きく増加し、拠点の受け入れ可能解析件数を大幅に超えているため受け入れ件数の制限を導入せざるを得ない状況になっている。次世代シークエンサーが生み出すデータ量は昨年度は、解析パイプラインの見直しによりデータ量の圧縮、要約化を行い、トータルでは 500TB を突破したているが、受け入れ制限により大きなデータ量の変動はない。また、昨年同様に、解析後のデータの元データを削除する等を行っているため常時保存に必要なストレージ容量は 50TB 程度に抑制されている。また、これらのデータは、同様に、国際データベースに、本年度よりほぼ全て登録の義務付けを行っており、ストレージの有効利用については一定の目途が立っている。次世代シークエンサーのデータフォーマットについても、FASTQ あるいは BAM フォーマットの導入で一般化の目途が立っており、最近では塩基配列データではなく、染色体座標に対するマトリクスデータとしての登録が進んでおり、データサイズの縮小化が進みつつある。この縮小については現在議論がなされており、一般化されるまではいましばらく時間がかかることが予想されている。これをうけて本プロジェクトにおけるデータ解析では、1)FASTQ データの作出、2)マッピングによる BAM データ、3)染色体座標に対するマトリクスデータ(bw フォーマット)の３つの作製を行うことに特化している。これら解析データは圧縮され、前述の国際データベース(DNA DATA BANK OF JAPAN, DDBJ)に登録される元データ FASTQ に加えて、米国国立生物学情報研究所で公開されている Gene Expression Omnibus(GEO)で解析の詳細についての説明と、解析結果とともに公開が促進されている。特に今年度からは多くの生物系学術誌が元データの公開のみならず、解析途中そして、最終解析結果データを GEO 経由での公開を義務付けたことから今後３つのデータ解析が本拠点解析での主流となることが考えられる。本プロジェクトでは、学内で産出されたデータを hakozaki,tatara 上に転送し、1 次解析による FASTQ の作製から最終解析までをサンプル毎・条件毎の処理単位で並列実行を行うパイプライン構築を行った。情報解析パイプラインは、主に Ruby および Parallel Job Organizer を用いて、オープンソースのソフトウェアである bowtie/tophat/macs/samtools および独自に開発した解析プログラムを適宜目的に応じて組み合わせることで構築した。多くのデータは国際データベース登録されているが、論文未発表の段階が多く、今後の追加解析等の要望に応えていく必要がある。



3. 解析パイプラインによって得られた成果と現在抱える問題について

本パイプラインの構築により、現在までに 3000 解析を超え、配列情報のデータ総量は 50TB を超える。また国際データベース登録件数は平成 27 年 4 月現在で 2300 解析を超えている。一方で、学術論文としての公開規定が大幅に変更されデータ解析パイプラインを変更すること、そして、論文執筆に当たっても、日に日に増す外部データとの比較解析が査読者より求められることが増えており、外部データとの比較を如何に効率的に行っていくかが課外となっている。特に、一回に比較すべき外部データの量がペタバイトクラスを超えており今後どのような形で解析を行っていくか、また、どのようなアウトプットで解析依頼に応えていくか新たな模索を開始している段階である。

成果論文

Harada A, Mallappa C, Okada S, Butler JT, Baker SP, Lawrence JB, Ohkawa Y, Imbalzano AN.
Spatial re-organization of myogenic regulatory sequences temporally controls gene expression.
Nucleic Acids Res. 2015 Feb 27;43(4):2008-21

Takahashi H, Takigawa I, Watanabe M, Anwar D, Shibata M, Tomomori-Sato C, Sato S, Ranjan A, Seidel CW, Tsukiyama T, Mizushima W, Hayashi M, Ohkawa Y, Conaway JW, Conaway RC, Hatakeyama S.
MED26 regulates the transcription of snRNA genes through the recruitment of little elongation complex.
Nat Commun. 2015 Jan 9;6:5941.

・Harada A, Maehara K, Sato Y, Konno D, Tachibana T, Kimura H, Ohkawa Y.
Incorporation of histone H3.1 suppresses the lineage potential of skeletal muscle.
Nucleic Acids Res. 2015 Jan 43(2):775-86.

Nakamura M, Shibata K, Hatano S, Sato T, Ohkawa Y, Yamada H, Ikuta K, Yoshikai Y.
A Genome-Wide Analysis Identifies a Notch-RBP-Jκ-IL-7Rα Axis That Controls IL-17-Producing γδ T Cell Homeostasis in Mice.
J Immunol. 2015 Jan 1;194(1):243-51.

Matsumoto M, Baba A, Yokota T, Nishikawa H, Ohkawa Y, Kayama H, Kallies A, Nutt SL, Sakaguchi S, Takeda K, Kurosaki T, Baba Y.
Interleukin-10-producing plasmablasts exert regulatory function in autoimmune inflammation.
Immunity. 2014 Dec 18;41(6):1040-51.

Oki S, Maehara K, Ohkawa Y, Meno C.
SraTailor: Graphical user interface software for processing and visualizing ChIP-seq data.
Genes Cells.2014 Oct ;19(12):919-26

Kobayakawa K, Kumamaru H, Saiwai H, Kubota K, Ohkawa Y, Kishimoto J, Yokota K, Ideta R, Shiba K, Tozaki-Saitoh H, Inoue K, Iwamoto Y, Okada S.
Acute hyperglycemia impairs functional improvement after spinal cord injury in mice and humans.
Sci Transl Med.2014 Oct 1;6(256):256ra137.

Stasevich TJ, Hayashi-Takanaka Y, Sato Y, Maehara K, Ohkawa Y, Sakata-Sogawa K, Tokunaga M, Nagase T, Nozaki N, James G. McNally JG, Kimura H
Regulation of RNA polymerase II activation by histone acetylation in single living cells.
Nature. 2014 Sep 11;516(7530):272-5

Tanaka M, Yamaguchi M, Shiota M, Kawamoto Y, Takahashi K, Inagaki A, Osada-Oka M, Harada A, Wanibuchi H, Izumi Y, Miura K, Iwao H, Ohkawa Y.
Establishment of neutralizing rat monoclonal antibodies for fibroblast growth factor-2.
Monoclon Antib Immunodiagn Immunother. 2014 Aug;33(4):261-9.

Tamura I, Ohkawa Y, Sato T, Suyama M, Jozaki K, Okada M, Lee L, Maekawa R, Asada H, Sato S, Yamagata Y, Tamura H, Sugino N.
Genome-wide analysis of histone modifications in human endometrial stromal cells.
Mol Endocrinol. 2014 Jul 28(10):1656-69.

Yokoyama A, Igarashi K, Sato T, Takagi K, Otsuka I M, Shishido Y, Baba T, Ito R, Kanno J, Ohkawa Y, Morohashi KI, Sugawara A.
Identification of Myelin Transcription Factor 1 (MyT1) as a Subunit of the Neural Cell Type-specific Lysine-specific Demethylase 1 (LSD1) Complex.
J Biol Chem. 2014 Jun 27;289(26):18152-18162.