（平成 23 年度研究報告書）

23-A-8　がん研究分野における
生物統計学・情報学研究基盤の構築に関する研究

独立行政法人国立がん研究センター　研究所　がんゲノミクス研究分野分野長　　柴田　龍弘

研究の分類・属性

基礎系

研究の概要

本研究班は、次世代型解読技術に伴う情報解析方法論の調査・検討・開発とがんの基礎研究への応用・至適化に資する研究を実証的に行い、がん研究分野における情報解析基盤強化とセンター内の解析支援を目指す。構築したアルゴリズムは順次実用化し、積極的にセンター外へ発信すると共にセンター内の基礎・臨床研究の解析支援に活用する。統計・情報解析専門家グループが行うがん解析に適合した新規アルゴリズムの調査・開発研究とがん検体解析グループが行う検証研究を連携させ、実戦的な解析基盤構築の推進とその連携の中からがん情報解析分野における人材育成を計る。

研究経費

27,900 千円

研究班の組織

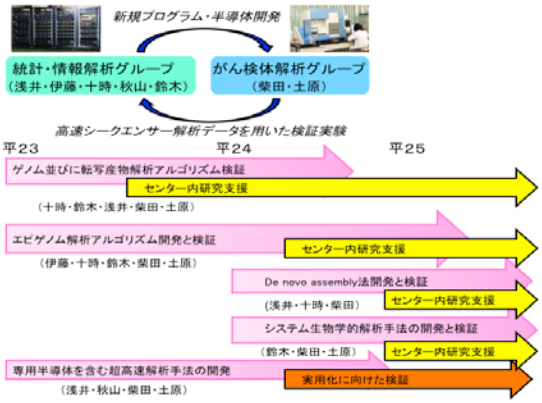
柴田　龍弘	国立がん研究センター・分野長	高速シークエンサーを用いた情報解析の検証
伊藤　隆司	東京大学大学院理学系研究科・教授	高速シークエンサーによるエピゲノム解析アルゴリズムの調査・検討・開発と検証
浅井　潔	産業技術総合研究所・センター長・東京大学大学院理学系研究科・教授	高速シーケンサーデータの情報解析とアルゴリズムの開発
十時　泰	国立がん研究センター・ユニット長	高速シーケンサーデータの情報解析とアルゴリズムの開発
秋山　泰	東京工業大学　大学院情報理工学研究科・教授	GPU等を用いた次世代シークエンサー向け情報解析の高速化
土原　一哉	国立がん研究センター東病院臨床開発センター・室長	高速シークエンサーを用いた情報解析の検証
鈴木　穣	東京大学大学院新領域創成化学研究科・准教授	高速シーケンサーデータの情報解析とアルゴリズムの開発

研究の目的と到達目標及び実績要点

全期間

（目的と到達目標）：

本研究は、数年以内に更に１０倍以上に増加すると予想される次世代型解読技術に必須の情報解析方法論の開発・検証と、がん研究への応用・至適化に資する研究を実証的に行い、同時に新たな人材確保によって、がん研究分野における情報解析基盤強化とセンター内の研究解析支援を目指す。具体的には、超高速配列アライメント・体細胞変異や染色体構造異常の高精度検出・新規 non-coding RNA の同定・染色体転座等による融合遺伝子同定・新規スプライシングなどヒトゲノム参照データベースに存在しない配列を決定する De novo assembly 法・全ゲノムメチレーション解析・システム生物学的手法等といった、高速塩基解読技術の革新に伴うがんゲノム・エピゲノム解析を支援することを目的とする。



第１年次（今年度）

（到達目標）

1　ゲノム並びに転写産物解析アルゴリズムの開発・検証

2　エピゲノム解析アルゴリズムの調査・検討・開発と検証

3　超高速ゲノム解析アルゴリズムの開発、がん情報解析分野における人材育成

（年次評価時点の実績要点）

１．ゲノム並びに転写産物解析アルゴリズムの開発・検証

a. 臨床検体を対象とした全エクソンシークエンスデータから体細胞変異を同定するアルゴリズムの開発と検証

b. 体細胞変異の予測アルゴリズムの更なる改良

c. 染色体構造異常の予測アルゴリズムの改良

２．エピゲノム解析アルゴリズムの調査・検討・開発と検証

a. RNAseq による non-coding RNA 解析並びに融合遺伝子の同定アルゴリズムの開発と検証

b. RNA の２次構造的なアクセサビリティを計算するプログラム並びに RNA のシュードノットを含む２次構造を予測するプログラムの開発

c. 全エクソンシークエンスデータと RNAseq/ChIPseq データとの統合を目指した解析手法の開発

d. バイサルファイト全ゲノムシークエンス解析について、マッピング・可視化・発見支援の観点から既存手法の調査。マッピングツールについてモデル生物データを用いて性能検証実験を行った。調査結果に基づき独自の解析ツール開発に着手し、一部については確認実験も行った。

３．超高速ゲノム解析アルゴリズムの開発、がん情報解析分野における人材育成

a. リードのクオリティスコアを考慮した高精度な配列マッピング手法の実装

b. シークエンサーから得られる大量の DNA リードを高速に参照ゲノム配列にマッピングするために、新規のソフトウェア GHOSTM を開発

c. がん情報解析分野における人材として新たに 2 名の生物統計学を専門とする研究員を獲得し、センターの解析支援増強に貢献

研究成果と考察

第１年次評価時点

１．ゲノム並びに転写産物解析アルゴリズムの開発・検証

a.　肺がん臨床検体 49 例についてシークエンスによる SNV の結果を Affymetrix GeneChip Human Mapping 250K SNP array の結果と参照したところ、平均で感度 74%、特異度 95%であった。更に 76 ヶ所の somatic SNV についてサンガー法による確認を行ったところ、全エクソンシークエンスで同定した変異を全て検証できた。大規模シークエンスデータから確実に変異と認められたものを抽出するパラメーター設定を行ったために、SNP アレイとの比較では感度がやや低下したものと

考えられるが、特異度については十分な結果が得られた。大腸がん検体の解析に関しては、凍結標本並びにホルマリン固定標本の両群とも 1 レーンあたりの冗長性の平均が 145 以上で、SNV の一致率も 95%以上であり、ホルマリン固定による悪影響は認められなかった。今後、多数例の検討を加えて、ホルマリン固定標本からの DNA サンプルの有用性を確認する。

既に全ゲノムシークエンスとサンガー法によって somatic mutation が確認済みの肝がん症例を使用して、exon capture 実験系の比較と mutation 検出に必要なエクソンシークエンスの最適な depth を検討した。Agilent Sureselect (Su)と Illumina Truseq (Tr)について、肝がんの腫瘍(HX5T)とリンパ球(HX5L)のサンプルをそれぞれ GAIIX の 2 レーン分のシークエンスを行った。参照ヒトゲノム配列にアライメントを行い PCR による duplication を取り除いた後、アライメントされたリード数を合わせて各 coding depth 以上の領域が占める coding 領域の割合を比較したところ、腫瘍とリンパ球の両方で、coding depth が低い 0-20 の狭い範囲では Truseq の方がわずかに高く、coding depth が高い 60 以上の広い範囲では Sureselect の方がはるかに高いことが解った。これは exon capture の設計が Truseq は coding 領域以外に UTR 領域を含んでいるのに対して、Sureselect は coding 領域だけに限定していることを反映している。次に両方とも同じ coding 領域だけを capture して同じ量だけシークエンスを行った場合を比較するために平均 coding depth を合わせて同じ比較を行ったところ、腫瘍とリンパ球の両方で、coding depth が低い 0-100 の範囲で Truseq の方が高く、coding depth が高い 120 以上の範囲で Sureselect の方が高いことが解った。つまり、Sureselect の方が coding depth の偏りが大きいことが解った。Truseq の方が coding depth の偏りが少なく、また mutation の検出で最低限必要な depth = 20 以上の割合は Truseq の方がはるかに高いので、特に同じ coding 領域だけを capture した場合は、Truseq の方が mutation 検出に適していると考えられた。次に、全ゲノムシークエンスと全エクソンシークエンスの depth の分布を比較した。全ゲノムシークエンスは平均 depth のところに狭いピークがくる正規分布を示すのに対して、全エクソンシークエンスの場合は平均 depth より低い depth が多い偏った分布になる。このことから全エクソンシークエンスの場合は、全ゲノムシークエンスよりも高い平均 depth が必要なことが想像される。そこで、今度は平均 coding depth = 50, 100, 150, 200 のデータセットを作成して、coding depth と coding coverage の関係と somatic mutation の検出精度(Sensitivity, Specificity)を調べた。coding depth と coverage の関係から平均 coding depth = 50 では急激に coding coverage が下がり、平均 coding depth = 100, 150, 200 と上がるにつれて、coding coverage もほぼリニアに上がることが解った。

somatic mutation の検出精度については、somatic SNV に関しては、平均 coding depth = 100, 150, 200 と上がるにつれて、Sensitivity と Specificity の両方ともほぼリニアに上がることが解った。平均 coding depth = 200 の false negative の 4 個を詳細に調べたところ、2 個はリンパ球の depth が低い(5 以下)ために検出されなかったことが解った。平均 coding depth = 200 の depth の分布を詳細に調べてみると、depth <= 5 が約 3%を占めており、depth が低いことが原因である false negative の率に一致する。Somatic indel に関しては、平均 coding depth = 100 で Sensitivity が 1.0 になった。

b. 予測された somatic SNV と somatic indel の repeat 領域と non-repeat 領域での頻度を比較すると、下記の表のように、somatic SNV は tandem repeat 領域で、somatic indel は tandem repeat 領域を含んだ repeat 領域で有意に予測数が多いことが解った。これは同じ塩基が繰り返して続く領域では、シークエンサーが繰り返し回数を間違えやすい性質を反映している。この結果より、somatic SNV は tandem repeat 領域、somatic indel は tandem repeat 領域を含んだ repeat 領域に false positive が多く現れることが解ったので、これらの領域を予測結果から取り除くフィルターを新たに開発した。このフィルターにより予測結果から false positive が大幅に取り除かれ、予測精度が改善した。

Prevalence (# / Mbp)

	Total	Repeat	Tandem	No repeat	Coding
Indels	0.98	1.52	15.63	0.43	0.10
SNVs	5.34	5.97	47.36	4.69	2.85

c. 肝がん全ゲノム解読データによる染色体構造異常の予測結果を PCR で確認したところ、予測された 477 個の染色体構造異常の内 350 個が確認された(Specificity = 350 / 477 = 0.73)。予測の間違えはミスアライメントが原因であるので、ミスアライメントによる false positive を取り除く必要がある。今までは染色体構造異常を支持する paired-end read について、本来アライメントされるべき領域(最大インサートサイズ内の領域)に類似配列がある場合は、その paired-end read が支持する染色体構造異常を取り除いていたが、それでは不十分なので、ゲノム全領域について類似配列を探すように拡張した。つまり、ゲノム全領域に類似配列が複数箇所ある場合は、アライメントのユニーク度が下がり、アライメントの信頼性も下がるという考え方である。このフィルターを加えることによって、Specificity = 348 / 450 = 0.77 に改善し

た。つまり新たに 27 個フィルターアウトされたがその内 false positive が 25 個、false negative が 2 個であった。正解が 2 個取り除かれてしまうが、不正解を 25 個も取り除くので有効であると判断して、新しいフィルターとしてパイプラインに組み入れることにした。

2　エピゲノム解析アルゴリズムの調査・検討・開発と検証

a. RNAseq のシークエンス結果から bowtie アライメントプログラムを使用して RefSeq, Ensembl, lincRNA など各種遺伝子データベースの配列へアライメントを行い、各種遺伝子データベースのアライメント結果を統合して、mapped read 数をカウントして発現量(RPKM: Read Per Kilo bases per Million reads)を算出するパイプラインプログラムを開発して、肝がん、胃がん、肺がんのサンプルに応用した。また、RNAseq を使用して融合遺伝子を検出するパイプラインプログラムを開発して、肝がん、胃がん、肺がんのサンプルに応用し、その結果、肺がん、胃がんにおいて新規キナーゼ融合遺伝子を同定できた。

b. ゲノム配列中の長い塩基配列が、RNA 分子として他の RNA 分子からアクセス可能となる確率（アクセサビリティ）を網羅的に計算する手法を開発することに成功した。siRNA と miRNA の計算機実験でその有効性を検証した。従来の手法と比べて最高の予測精度を示しつつ、圧倒的に高速な、シュードノットを含む RNA２次構造予測プログラムの開発に成功した。

c. 全エクソンシークエンスデータから見出された体細胞突然変異の生物学的意義の検証を行うことを目的に、同定された体細胞変異を RNAseq/ChIPseq データといった次世代シークエンス解析により得られた多様な実験データと統合する枠組みを整備した。肺がん由来細胞株を含む 15 種類の培養細胞より RNAseq/ChIPseq データを収集し、dbSNP、1000 人ゲノム計画等により見出された遺伝子多型ともあわせて、総計 20 億のシークエンスタグデータの可視化、検索が可能なデータベースを構築した。ChIPseq については、主として活性性のヒストンマーカー(H3K4me3, H3Ac)、抑制性のヒストンマーカー(H3K27me3)、RNA polymeraseII のパターンからクロマチンの活性化状態を記載することを目的に集計した。RNAseq については、total RNA の発現解析を主に、核、細胞質、ポリソームの各細胞画分についてデータを収集、統合を進めた。特に転写開始点の位置と転写活性を同定する TSS Seq 法を用いて、現在までにプロモーター、選択的プロモーターの転写開始点近傍に存在する変異を集約的に抽出し、その転写因子結合配列への影響を評価している。今後、肺がんエキソームデータセットの変異パターンを共有すると思われる培養細胞を選定、変異が誘起する機能異常の推定を行う。また同時に、臨床検体を用いて同様の解析が可能となるような実験手法の確立と評価を行い、より直接的なデータの収集と統合を試みる。

d. バイサルファイトシーケンスデータ解析アルゴリズムの開発

d-1: バイサルファイトシーケンスデータのマッピング既存ツールの問題点として以下の点が挙げられた。

① bismark は MethylC-seq のリードは容易に処理してくれるものの、PBAT のリードでは読み出しストランドやランダムプライマーに相当する読み出し直後の配列信頼性の問題からエラーを生じる。② bs seeker は MethylC-Seq のリードでもエラーが頻発する。　③ bsmap は遅すぎて問題外　④ 全てのアルゴリズムにおいて、マルチヒットの場合にどれか一つにリードをランダムに割り振るように設定されている。

そこでこれらの問題点を解決するため以下のような機能を実装した独自の専用ソフトウエアの開発を研究班内の共同研究として進めた。

① シード検索とアライメントをシームレスに結合する必要あり　② ランダム選択ではなく、全候補領域を SW 等で高精度に検証するのが理想　③ 高速な SW として GPGPU やビットパラレルアルゴリズムが候補。(秋山班員と共同開発)

d-2:メチロームデータの可視化既存ツールの問題点として以下の点が挙げられた。

① リード数等の重要な情報が得られない

そこでこれらの問題点を解決するため以下のような機能を実装した独自の専用ソフトウエアの開発を進めた。

① メチル化率とリード数を同時表示可能なトラックを開発

② 設定ツールの提供によりユーザーに必要コンテクストの選択肢を提供

d-3：メチロームデータからの発見促進

① フィッシャーの正確確率検定により各シトシンについてメチル化率変動の有意性をスコア化する機能について検討した。

3　超高速ゲノム解析アルゴリズムの開発

a. クオリティスコアを考慮し、アラインメント確率とギャップ確率の両方を導入した確率アラインメントを行うことにより、精度の高いリードマッピング手法の実装に成功した。この手法は、in/del の検出に効果的であることを検証した。

b. ベンチマーク試験においては、GHOSTM (GPU４枚使用) は、基準とした BLAST の 438 倍高速であり、GHOSTM(GPU を１枚のみ使用)でも 165 倍高速であった。BLAT アルゴリズムはこのベンチマークでは BLAST の 40 倍高速であったので、BLAT を基準

とするならばその約 11 倍の高速化を達成したと言える。しかし、精度についての検証では、GHOSTM は BLAT を圧倒的に凌駕する性能を示しており、現実的なケース（Bit スコア 50 以上）においては BLAST と遜色のない正確な結果を得られたのに対して、BLAT では無視できない量のクエリで誤った結果となった。BWT 等のより簡易な方法では、リファレンスゲノムに当たらない断片配列についても、超高速マッピングを行う可能性が拓けた。改良版の GHOSTX アルゴリズムについては現在も評価を継続中であるが、上述の a)の絞り込み過程を圧倒的に改善することにより、b)を必ずしも GPU で実装しなくても通常の高速な CPU を持つワークステーションでも実施が可能となりつつある。

倫理面への配慮

解析方法の検証研究に使用する解析データの取得に関しては、「ヒトゲノム・遺伝子解析研究に関する倫理指針」並びに「疫学研究に関する倫理指針」を遵守し、個人情報の保護等、試料等提供者の人権擁護上の配慮を十分に行い、また患者にとって不利益や危険性が起こらないように留意し、検体の使用について同意が得られた検体のみを用いる。検体については、国立がん研究センター中央病院並びに東病院個人情報管理者によって連結可能匿名化の作業を行った後に、解析を行う。解析データの取得に関する研究については、国立がん研究センター中央病院並びに東病院の倫理審査委員会の承認を得る。

本研究に関連する、本研究期間中の主な発表論文等

（雑誌論文）

1. **Totoki Y,** Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsustumi S, Sonoda K, Totsuka H, Shirakihara T, Sakamoto H, Wang L, Ojima H, Shimada K, Kosuge T, Okusaka T, Kato K, Kusuda J, Yoshida T, Aburatani H, **Shibata T.** High-resolution characterization of a hepatocellular carcinoma genome. **Nat Genet**, 2011, 43, 464-469.
2. Watanabe T, Tomizawa S, Mitsuya K, **Totoki Y,** Yamamoto Y, Kuramochi-Miyagawa S, Iida N, Hoki Y, Murphy PJ, Toyoda A, Gotoh K, Hiura H, Arima T, Fujiyama A, Sado T, **Shibata T**, Nakano T, Lin H, Ichiyanagi K, Soloway PD, Sasaki H. Role for piRNAs and a novel non-coding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. **Science**, 2011, 332, 848-852.
3. Kiryu H, Terai G, Imamura O, Yoneyama H, Suzuki K, **Asai K**. A detailed investigation of accessibilities around target sites of siRNAs and miRNAs. **Bioinformatics**, 2011 27: 13. 1788-1797.
4. Sato K, Kato Y, Hamada M, Akutsu T, **Asai K**. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. **Bioinformatics**, 2011 27: 13. i85-i93.
5. Hamada M, Wijaya E, Frith MC, **Asai K**. Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. **Bioinformatics**, 2011 27: 22. 3085-3092.

（学会発表）

1. Whole genome sequencing of virus-associated hepatocellular carcinoma
   Tatsuhiro Shibata, The U.S. –Japan Cooperative Cancer Research Program Workshop, “Cancer Genomics and Epigenomics: Towards Personalized Cancer Medicine” (2011)
2. Integrated genome and transcriptome sequencing analyses of virus-associated hepatocellular carcinoma
   Tatsuhiro Shibata, “Cancer Genomics”, EMBO/EMBL Symposium (2011)
3. What Can We Expect From Whole Genome Sequencing of HCC ?
   Tatsuhiro Shibata, International Liver Cancer Association Workshop (2011)
4. Integrated genome and transcriptome sequencing analyses of virus-associated hepatocellular carcinoma

Tatsuhiro Shibata, The 9th International Workshop on Advanced Genomics "Revolution of genome science" (2011)