生物物理 53（6），290-294（2013）

解説

RNAシークエンシング

城口克之　独立行政法人理化学研究所統合生命医科学研究センター統合ジェノミクスラボ

The term “RNA sequencing (RNA-Seq)” often means the sequencing of cDNA generated from RNA as well as the direct sequencing of RNA. RNA-Seq has become well known as next generation sequencers emerge, particularly because their high throughput nature of data collection makes RNA-Seq a powerful tool for genome-wide gene expression analysis. In this review, I introduce, mainly from a technical point of view, a basic scheme of RNA sequencing, its development, and its application in several interesting studies.

RNA sequencing / Next generation sequencer / Gene expression / Genomics / Single molecule

1. はじめに

RNA シークエンシング（RNA sequencing, RNA-Seq）[1), 2)] とは，最終的に RNA の配列情報を決定するという意味で使われている．したがって，RNA の配列情報を DNA に逆転写し，その DNA の配列を読むことによりもとの RNA の配列を決定することも含む．この意味では，これまでも RNA-Seq は行われてきた．しかし，一度に大量の DNA の配列を決定できる次世代シークエンサの登場によって，RNA-Seq が遺伝子発現解析に多用されるようになり，“RNA-Seq” という言葉がより広く使われるようになってきている．本解説では，次世代シークエンサを用いた RNA-Seq の概要，課題と改良，さらに遺伝子発現解析とは異なる目的に利用されている RNA-Seq の発展と今後の可能性について解説・考察する．

2. 次世代シークエンサの登場

現在，市販されている次世代シークエンサ[3)] の中には，1 日でヒトの全ゲノム配列を決定できる容量をもつものもある．これは，最初のヒトゲノムを解読する際に，21 世紀初頭まで 14 年間程かかったことを考えれば，驚異的な技術革新といってよいだろう．

次世代シークエンサでは，容量が大きい装置だと，一度のランで試料内にある 10^9 程度の DNA 分子それぞれの配列を決定できる．その際，DNA 1 分子を用いて配列を決定しているわけではなく，各 DNA を 1,000 コピー程度に増幅してから配列を読んでいる（**図 1**）．次世代シークエンサでシークエンシングできるようにするためには，配列を決定したい DNA の両端に特定の DNA 配列を結合させる必要がある．この特定配列がシークエンシング直前の増幅に利用され，またシークエンシングの際に必要なプライマーが結合する部分としても使用される．多数の種類の DNA にこのような特定の DNA 配列が付加された試料は “ライブラリ” と呼ばれる．

次世代シークエンサは，主に illumina 社，Life Technologies 社，Roche Diagnostic 社から市販されているが，配列決定できる DNA の数や長さ，一回のランにかかる時間は，それぞれのメーカーや機種によって異なる．容量が多い装置では，10^9 程度の数の DNA を両端から 150 塩基ずつ配列を決定できるが，おおよそ 1 週間程度かかる．2 日で 10^7 程度の DNA の 600 塩基を決定できるもの，また，10^6 程度の DNA を 800 塩基決定できるものもある．これらの性能は日々進歩している．

3. 第三世代シークエンサ

1 分子を用いて配列を決定するシークエンサも Pacific Biosciences 社から市販されており，“第三世代” と呼ばれている．第三世代シークエンサは，数千塩基という長い DNA 配列を連続的にシークエンシングできるという長所をもつ．次世代シークエンサが複数の DNA のコピーから配列を読むときに生じる，各 DNA の反応の効率の差からくるシグナルの “ずれ” がないからである．しかし 1 つの分子の反応に頼るため，エ

RNA Sequencing
Katsuyuki SHIROGUCHI
Laboratory for Integrative Genomics, RCAI, RIKEN Center for Integrative Medical Sciences (IMS-RCAI)

ラーが多いことが課題である．長い配列を一度に決定できると，ゲノム DNA の参照情報がなくても，RNA の配列だけから RNA の全長が推定できる．次世代シークエンサと併用されることも多く，次世代シークエンサで同定された短い配列を統合する際に利用されることもある．

4. 次世代シークエンサを用いた RNA-Seq と遺伝子発現の定量化

次世代シークエンサを用いた RNA-Seq（**図 1**）が誕生し，2008 年には RNA-Seq を用いた多くの報告がされた．たとえば酵母（yeast）のゲノム全体の転写部位が同定され[4]，また，哺乳細胞の RNA の発現量をゲノムワイドに定量できることが示されている[5]．後者の例にもあるように，RNA-Seq の誕生により，次世代シークエンサは配列決定という役割とともに，核酸のカウンティングという定量計測装置としての役割も担うようになった．試料中に多数ある DNA 1 つ 1 つの配列を決定することは，どの配列をもつ DNA が何個存在するかを計測することになるからである．網羅的な RNA の発現量の解析は，遺伝子の発現ネットワークの同定，細胞集団を示すマーカー探索，さらには，分化における細胞の運命を決定づける遺伝子の同定などの研究に有効であり，現在，さまざまな分野で RNA-Seq が行われている[1),2)]．

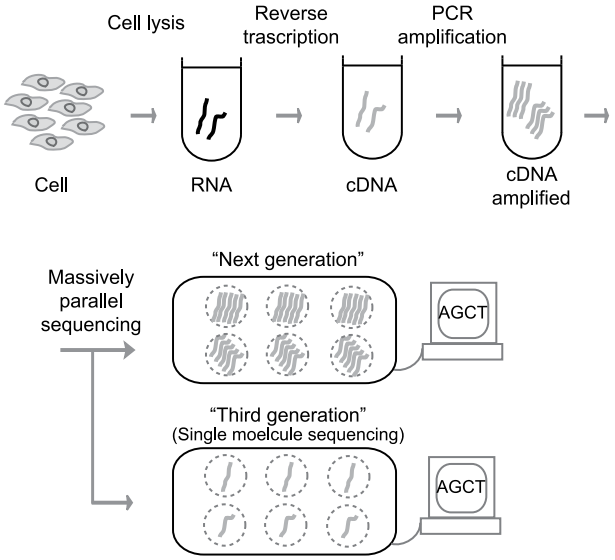

図 1
次世代シークエンサと第三世代シークエンサを用いた一般的な RNA-Seq．次世代シークエンサではシークエンシングの直前にも増幅を行い，シークエンシング自体には約 1000 コピーの DNA が用いられる．この増幅は，メーカーによってシークエンシングする際の基板上で行われるものや，装置に導入する前に溶液中で行われるものがある．第三世代のシークエンサでは，DNA 1 分子を用いて配列が決定される．この図では，点線で囲まれた円 1 つにつき 1 つの配列が決定される（ここでは，次世代，第三世代ともに 6 つ）．

これらの研究は DNA アレイを用いても行われてきているが，主に下記の理由から RNA-Seq にとって代わってきている．RNA-Seq では DNA アレイのようにプローブを作製する必要がないので解析するターゲットをあらかじめ決める必要がないこと，一塩基の分解能で RNA を同定できること，一度に配列を解読できる量が多いこと，さらには，DNA アレイのようにサンプル DNA とプローブの非特異的な結合からくるシグナルのバックグラウンドがないこと，などである．

5. RNA-Seq の課題と改良

RNA-Seq の発展に伴い，いくつかの課題が生じている．重要なものの 1 つとして定量の再現性・精度が挙げられる．RNA-Seq は一般に PCR 増幅を伴うが，PCR の増幅率は基本的に配列に依存するので，増幅産物の定量により異なる RNA の数を正確に比較することは難しい．さらに，PCR は増幅率が 100% ではないので（一回の PCR サイクルですべての DNA 分子が 2 倍に増幅されるわけではない），特に少数コピーから増幅された場合には，増幅後のコピー数の再現性が悪い[6]（条件によるが，10 倍程度の違いは十分にありうる）．そのため，少数コピーの場合は同じ配列をもつ RNA の試料間の相対比較も難しいし，標準曲線を用いる絶対数の推定も難しい．また，RNA から DNA に変換する逆転写反応の際にも，プライマー配列に依存したバイアスがあると考えられている．近年，これらの再現性の悪さを解決するためにいくつかの方法が開発・提案されている．

コンセプトとして最もシンプルだと思われる解決法は RNA を増幅しないことであり，複数の報告がある．cDNA を作製せずに RNA を直接シークエンシングする方法では，1 分子の RNA から配列を決定できるシステムを構築しており，シークエンシングできる長さは論文[7]の図から判断すると，シークエンシングした RNA のうち 50% 程度の RNA で 18-20 塩基以上となっている．シークエンシング時に用いられる基板上で RNA を捕捉して逆転写反応を行い，その後，各 cDNA の配列を決めた報告もある[8]．これらは第三世代の 1 分子でシークエンシングする技術を用いている．一方で，次世代シークエンサを用いた方法も報告されている[9]．ここでは RNA を直接次世代シークエンサの基板上に流しこみ，基板上で逆転写反応を行う．その後，基板上で各 cDNA を増幅してシークエンシング

している．ここで得られる 1 つの配列は，基板上の RNA 1つに対応する．シークエンシング自体は市販のシステムを使用しているので安定しているが，逆転写反応の部分は市販のシークエンサの内部に手を加えているので，ハードウェアを改良する技術が必要である．

6. “バーコード”を用いた 1 分子分解能デジタル定量法

筆者らは，先に挙げた課題を，シークエンサシステムに手を加えずにライブラリを準備する段階に改良を加えて解決する方法を報告した．そこでは“分子バーコード”という概念を利用している[10]．

ポイントは，同じ配列をもつ DNA 分子 1 つ 1 つを区別するために，定量したい増幅前の DNA 分子それぞれに，異なる DNA 配列（バーコード）を付加させることにある．たとえば試料内に同じ配列をもつDNA 分子が 3 つあるとしよう（**図 2** の DNA1）．この分子をそのまま PCR 増幅すると，増幅後の DNA の数を数えても増幅前の分子数はわからない．既知の個数の DNA 分子を同様に増幅して標準曲線から求める方法があるが，先に述べたノイズにより測定は不正確となる．これを解決するために，多数の種類のバーコードを用意し，定量したい増幅前の各 DNA 分子それぞれに確率的に異なるバーコードを結合させる．そして，増幅後に異なるバーコードの数を数えることにより，増幅前の DNA の数を 1 分子の分解能でデジタル定量する（**図 2** を参照）．この概念は 2003 年に提案されており[11]，次世代シークエンサの普及後，実際に蛋白質と RNA の相互作用解析に利用され[12]，さらに 1 種類の DNA 配列を 2 桁のダイナミックレンジで正確に定量できることが示された[13]．その後，フィンランドとスウェーデンのグループ[14]，そして筆者ら[10]が，ゲノムワイドな遺伝子発現解析に利用可能なことを示した．筆者らは，ゲノムワイドに正確に定量できるようにするため，下記に示すいくつかの工夫を加えている．

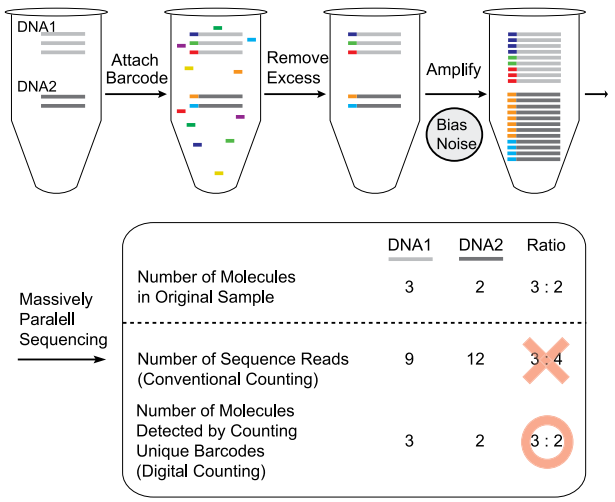

図 2
1 分子バーコード付加によるデジタル計測の概念図．ここでは 3 個の DNA1 は増幅後 9 個になり，2 個の DNA2 は増幅後 12 個になっている．これは，増幅率の違いは配列の違いや，各 PCR サイクルの増幅の成否が確率的に決まることによる（PCR 反応の増幅率は 100% ではない）．通常の計測により増幅後の数を定量すると（薄いグレー，濃いグレー），DNA1 と DNA2 の比は 3:4 (=9:12) となる．また元のチューブに存在した DNA の絶対量は，既知の数の DNA を用いて増幅後の数を定量した標準曲線により求める必要がある．一方で，増幅前に多数のバーコード配列を加え，それぞれの DNA が確率的に異なるバーコードが付加されるようにすると，増幅後にバーコードの種類の数を数えることにより，元のチューブに存在した DNA の絶対数がわかる．

この方法で注意が必要なことは，増幅エラーやシークエンシングエラーが起きると，分子数の計測に大きな影響がでることである．よく使われている，ランダム配列（すべての種類を含む配列）をもつ DNA をバーコードとして用いるとその影響が顕著であり，この場合はバーコード配列に一塩基のエラーが起きると，基本的にはバーコード部分にエラーが起きなかった分子と起きた分子の 2 つがもともと存在したと解釈される．筆者らはこの問題を解決するために，使用するバーコード配列を限定し，エラーが起きたときにそれをエラーだと同定できるようにした．この時，使用する任意のバーコード配列の組み合わせにおいて，同じ配列になってしまうのに必要なエラー（もしくは変異）の数が一定値以上になるように設計した．さらに，各 DNA に 2 つのバーコード配列を独立に付加させ，145 種類用意したバーコードを用いて，21,025（=145 × 145）種類のバーコードの組み合わせを実現し，4 桁のダイナミックレンジにより絶対定量ができることを示した[10]．この方法は，準備するバーコード配列の種類を増やすことによりその組み合わせが 2 乗で増加し，さらに定量したい DNA 1 つに 3 つ以上のバーコード配列を付加させて組み合わせの数を増やすことも原理的には可能なので，ダイナミックレンジは比較的容易に大きくできる．

このように筆者らは，バーコードを用いてデジタル定量を行うことで配列に依存したバイアスや増幅ノイズを排除し，これまで難しかったゲノムワイドかつ 1 分子の分解能をもつ定量法を実現した．これまでのゲノムワイドな定量には測定システムに配列依存性があることが多く，異なる遺伝子間の量の比較が難しかった．この方法を用いると，遺伝子間の発現量の正確な比を得ることができ，それを基に遺伝子発現定量ネットワーク解析が可能となる．また，増幅ノイズは特に

低コピーの RNA を同定する際に顕著であるため，ここで示したデジタル定量法は，転写因子など，分化などに重要な役割を担うが低コピーしか発現していない遺伝子の発現定量に効果をより発揮する．近年注目されている 1 細胞計測においては，RNA の量が限られること，また，各細胞の性質を同定するために細胞間で平均することが許されないことから，低コピー計測に効果的な精度の高いデジタル定量法は有効であると考えられる．尚，この方法はシークエンサの種類に依存しないので，汎用性も非常に高い．

7. RNA-Seq による 1 細胞全 RNA の網羅的解析

正確な定量法の開発とともに，少量サンプルの定量法の開発も進んでいる．そのゴールの 1 つが 1 細胞全 RNA の網羅的解析である．前セクションで少し触れたが，1 細胞計測の利点は，1 分子計測などからも広く知られているように，1 つ 1 つの細胞の状態が平均化されないことにある．細胞がヘテロな集団であっても 1 つ 1 つの細胞の特性を記述でき，また，（ほぼ）均一だと解釈されている細胞集団のバイオマーカーの探索にも直接つながる．

実際に複数のグループから，次世代シークエンサを用いた 1 細胞 RNA 網羅的解析の成果が報告されている．最初のグループは 2009 年に報告しており，RNA の量が多い卵割球の発現解析を行っている[15]．その後，2011 年[16] と 2012 年[17] にそれぞれ新しい方法が報告された．より詳しい記述がある 2011 年の報告の図から判断すると，100 コピーの RNA の検出からノイズが表面化して再現性が著しく減少している．ここでは，48 個のマウス胚性幹細胞と 44 個のマウス胎児繊維芽細胞それぞれにおいて 1 細胞 RNA 網羅的解析が行われたときに，既知のコピー数の RNA がコントロールとして加えられている．この時，たとえば10 コピーを加えられた RNA は，合計 92（=48 + 44）回の 1 細胞解析の結果，6-7 割の実験で検出されていない．少量の RNA からスタートする 1 細胞解析では，サンプルのハンドリングが難しいこともあり，ライブラリ作製時に増幅を積極的に行っているので，再現性が悪いと考えられる．これらの報告では先に挙げたデジタル定量法は用いられていない．

1 細胞全 RNA の網羅的解析は，生物の理解に向けた有効なツールであることは議論の余地がないといってよいであろう．高い検出効率と再現性をもち，簡便な方法の開発が期待されている．

8. RNA-Seq を用いた他の研究

RNA-Seq は，遺伝子の発現部位や発現量を調べることとは別の方向にも利用されており，例をいくつか紹介する（**図 3**）．

（I）出芽酵母にて，抗体を用いて RNA ポリメラーゼを単離し，一緒に単離された合成中の RNA の 3′ の配列を同定することにより，細胞中で転写中の RNA ポリメラーゼの位置を決定している[18]．これにより，ゲノム上の RNA ポリメラーゼの位置の分布や転写中に高頻度で一時停止をする配列などが明らかとなっている．

（II）出芽酵母にて，リボソームを単離した後にリボソームと直接結合していない RNA を除去し，残った（リボソームにカバーされていた）RNA の配列を同定することにより，細胞内におけるリボソームの位置を決めている[19]．リボソームの RNA 上の存在頻度に 3 塩基の周期が見られており，これはコドンを示していると考えられている．同じグループによる大腸菌を用いた報告では，遺伝子がコードされている部分のシャインダルガノ配列に似た配列でリボソームが高頻度に一時停止することが示されている[20]．

（III）酵素を利用して RNA の 2 本鎖部分を切断し，切断された部分を同定することにより，特に構造が機能発現に重要だと思われている Non-coding RNA（蛋白質をコードしていない RNA）の構造を推定している[21]．

9. RNA-Seq の展望

先にも述べたように，1 細胞内 RNA の 1 分子分解能による網羅的かつ高感度のコピー数解析が望まれ

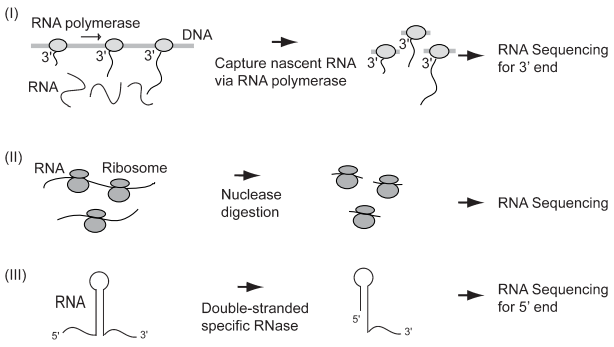

図 3
RNA-Seq の利用．（I）転写中の RNA ポリメラーゼの RNA 上の位置の決定．（II）翻訳中のリボソームの RNA 上の位置の決定．（III）RNA の二次構造の推定．

る．これにより，RNA 解析によるバイオマーカー探しはルーチン作業となる可能性がある．1 細胞・高感度の 1 つ手前であるが，少数細胞（10-100）での高精度の測定を安定して行うことにも意義があると思われる．生物研究において，またヒトのサンプルなどを用いる場合に，多数の細胞を準備できないことが多々あるからである．この場合でも低コピー RNA を含めた再現性のよい定量法は必要であろう．

1 細胞または少数細胞において，RNA の定量と同時に同じサンプルでのゲノム DNA シークエンス解析も望まれる．現在，ゲノム DNA も 1 細胞から増幅して 90% 以上のゲノム領域が次世代シークエンサの一度のランで検出・解析されているので[22]，近未来に実現するのではないだろうか．RNA の発現と DNA 変異との相関解析はがん研究などに強力な威力を発揮するであろう．

第三世代シークエンサのさらなる開発・改良も期待されているであろう．少ないエラーで長い配列を読むことができたら，リピートなどが多くて未だに決定できていない部位の配列を決定できる．また，ゲノム DNA の参照情報が存在しない，環境中にある微生物などの発現解析等に有効であろう．

1 細胞 RNA の網羅的解析が行われている今日，それらの細胞が作り出す分布を見てシステム全体がどのように振舞っているかを理解していくことが必要であろう．そのためにはたくさんの細胞を解析する必要があるが，現在のシークエンサでは 100 を超える細胞を一度のランで計測することは容易ではない．一方で，遺伝子の数を絞ってより多くの細胞の解析を行う方向も有効であろう．1,000 や 10,000 といった数の細胞を解析して分布を得ることにより見えてくる生物の新しいシステムがあるかもしれない．

DNA 解析においては，次世代シークエンサを用いて妊婦の血液に流れている少量の胎児のゲノム DNA のコピー数を決定した報告がされているなど[23]，診断などへの応用も期待されている．フェノタイプを同定できる RNA-Seq の診断への応用も同様に期待される．

10. さいごに

生物物理学とシークエンシングを用いた研究の親和性はどうであろうか．たとえばマイクロ流路と次世代シークエンサを用いて 1 細胞の全ゲノムハプロタイピングが行われたように，生物物理的手法を加えたユニークな研究も行われている[24]．また筆者が現所属に移る前に在籍した研究グループでは，以前にモーター蛋白質の細胞内ステップを観察していた研究者や RNA ポリメラーゼの DNA 上の動きを光ピンセットを用いて 1 塩基の分解能で観察していた研究者らが，生物物理学的手法を用いて新しい原理のシークエンサを作り上げている[25]．さらには先に挙げた例であるが，次世代シークエンサを用い，サンプルの調製法を工夫して 1 細胞から全ゲノム配列を決定できる方法を開発した研究者らは生物物理出身といえるだろう[22]．このように，定量計測に強い生物物理の研究者と，定量計測装置として発展しているシークエンサの親和性は高い．筆者は，シークエンサと生物物理学的な“技”の融合，そしてその融合による医学研究への貢献に高いポテンシャルを感じている．生物物理からこのような方向性に飛び込む研究者とともに切磋琢磨できたら幸いである．

謝　辞

RNA-Seq，そして Genomics という私にとって新しい分野に挑戦する機会を与えてくださった Harvard University の Prof. Sunney X. Xie に深く感謝します．

文　献

1） Wang, Z. *et al.*（2009）Nat. Rev. Genet. **10**, 57-63.
2） Ozsolak, F. *et al.*（2011）Nat. Rev. Genet. **12**, 87-98.
3） Metzker, M. L. *et al.*（2010）Nat. Rev. Genet. **11**, 31-46.
4） Nagalakshmi, U. *et al.*（2008）Science **320**, 1344-1349.
5） Mortazavi, A. *et al.*（2008）Nat. Methods **5**, 621-628.
6） Peccoud, J. *et al.*（1996）Biophys. J. **71**, 101-108.
7） Ozsolak, F. *et al.*（2009）Nature **461**, 814-818.
8） Mamanova, L. *et al.*（2010）Nat. Methods **7**, 130-132.
9） Ozsolak, F. *et al.*（2010）Nat. Methods **7**, 619-621.
10） Shiroguchi, K. *et al.*（2012）Proc. Natl. Acad. Sci. USA **109**, 1347-1352.
11） Hug, H. *et al.*（2003）J. Theor. Biol. **221**, 615-624.
12） König, J. *et al.*（2010）Nat. Struc. Mol. Biol. **17**, 909-915.
13） Fu, G. K. *et al.*（2011）Proc. Natl. Acad. Sci. USA **108**, 9026-9031.
14） Kivioja, T. *et al.*（2012）Nat. Methods **9**, 72-72.
15） Tang, F. *et al.*（2009）Nat. Methods **6**, 377-382.
16） Islam, S. *et al.*（2011）Genome Res. **21**, 1160-1167.
17） Hashimshony, T. *et al.*（2012）Cell Reports **2**, 666-673.
18） Churchman, L. S. *et al.*（2011）Nature **469**, 368-373.
19） Ingolia, N. T. *et al.*（2009）Science **324**, 218-223.
20） Li, G.-W. *et al.*（2012）Nature **484**, 538-541.
21） Wan, Y. *et al.*（2012）Cell **48**, 169-181.
22） Zong, C. *et al.*（2012）Science **338**, 1622-1626.
23） Fan, H. C. *et al.*（2012）Nature **487**, 320-324.
24） Fan, H. C. *et al.*（2011）Nat. Biotech. **29**, 51-57.
25） Sims, P. A. *et al.*（2011）Nat. Methods **8**, 575-580.


城口克之

城口克之（しろぐち　かつゆき）
自然科学研究機構岡崎統合バイオサイエンスセンター博士研究員，早稲田大学理工学術院客員講師，Harvard University Postdoc を経て現職．
研究内容：Single Cell & Single Molecule Integrative Genomics
連絡先：〒 230-0045 神奈川県横浜市鶴見区末広町 1-7-22
E-mail: katsuyuki.shiroguchi@riken.jp