AIとスパコンを活用したがんゲノムデータ解析手法の開発

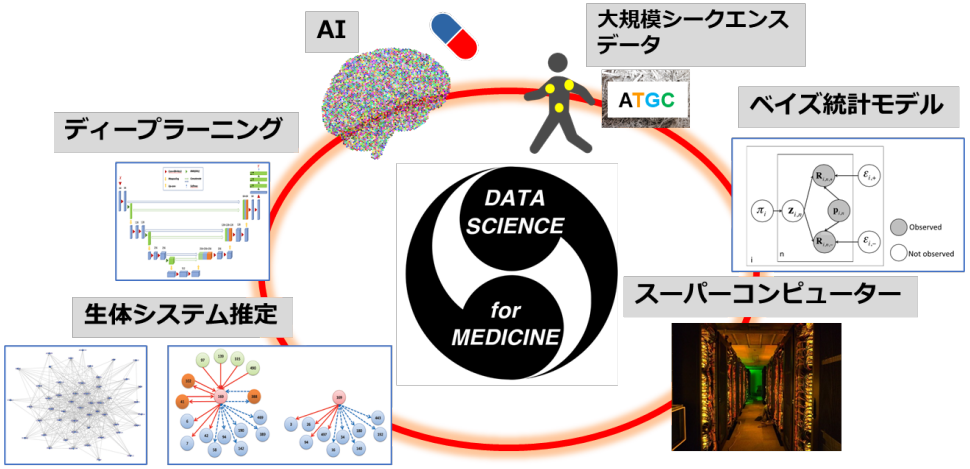
愛知県がんセンター
システム解析学分野

はじめに

- がん細胞は、正常な細胞中の**DNA（ゲノム）配列**に変異が起きることにより、細胞が異常に増殖する能力などを獲得したものです
- 同種類のがんであっても、ゲノム上で変異の起きる場所や種類の頻度にバリエーションがあります
- 現在**ゲノム上の変異**に応じて治療法の選択を考える、**がんゲノム医療**が、本格化しつつあります
- ここではがんゲノムの変異を見つけるための**元となる観測データ**はどのようなものか、**どのような原理で変異を検出している**のか、また実際にはどのような難しさがあるのかなどを、
- システム解析学分野での研究を交えて概説します

システム解析学分野について

データ科学の力で医療へ貢献！



- 2019年2月に研究所にできた新しい分野です
- がん細胞から得られた**ゲノムデータ**などの様々な**生体ビッグデータ**を**解析する方法**の研究を行っています
- 患者さんのデータから、**がん細胞の複雑なシステム**に関わる情報を抽出し、**一人ひとりに合わせた医療**へつなげることを目指しています

研究室メンバー

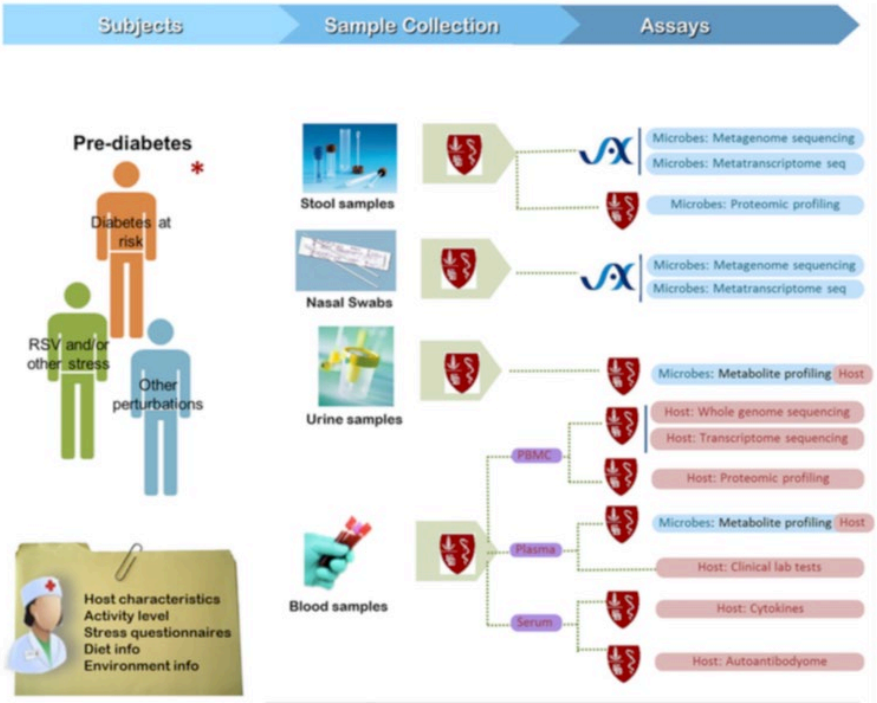
分野長：	山口　類
研究員：	郭　中梁
リサーチレジデント：	武藤　理
研究補助：	鈴木一基
事務員：	竹中亜由美

研究室HP
https://www.pref.aichi.jp/cancer-center/ri/11shisutemu_kaiseki/index.html01bumon/

様々なデータの取得が可能になってきています

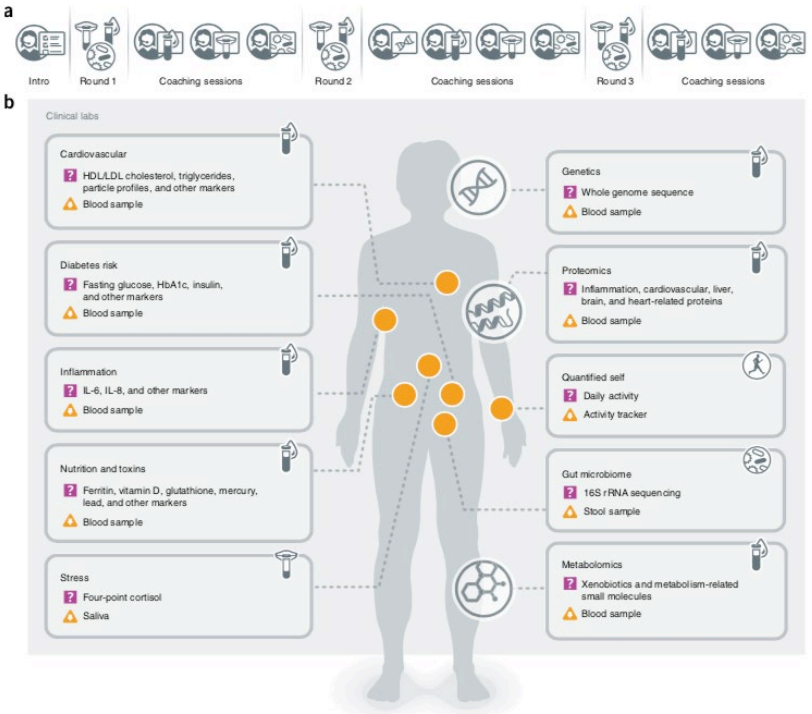
DNA, RNA, タンパク、メタゲノム（腸内細菌叢なども）、、、

Integrative Human Microbiome Project



Integrative HMP (iHMP) Research Network Consortium, *Cell Host & Microbe* , (2014).

Pioneer 100 wellness project



Price *et al., Nat Biotechnol,* (2017)

様々な生体データを大量に取得できるようになった背景

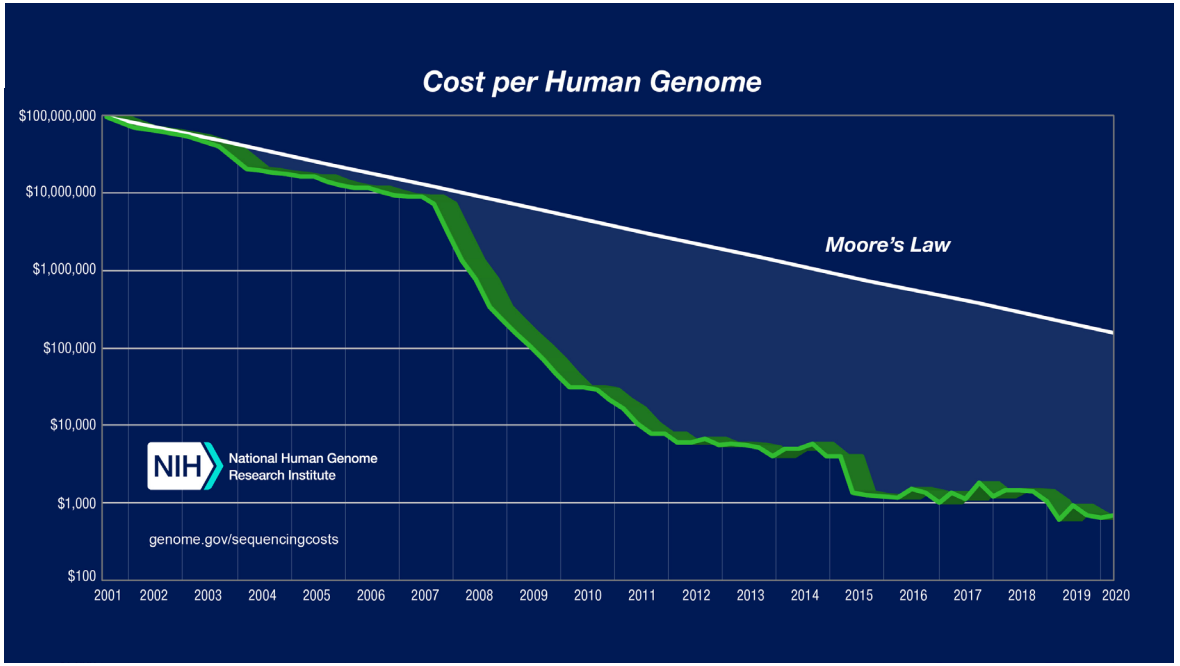
計測装置の飛躍的性能向上・計測コスト低下

一人分の全ゲノム配列を得るのにかかる費用の変化

20年前：100億円

1/100,000

現在：10万円



様々な生体データが大量に取得できるようになった背景には、**DNA**の塩基配列（**シークエンス**）などを計測する**次世代シークエンサー(NGS)**の飛躍的性能向上があります。**一人分の全ゲノム情報を得るのにかかるコストは20年で10万分の1**になりました。**RNA**などもNGSで計測できます。**腸内細菌叢のゲノム**（メタゲノム）なども読むことができます。**これらのシークエンスデータからがん細胞特有の変異・変化を見つけることで、がんの原因や治療法に関する情報を得ることができます。**

シークエンスデータって、どんなデータ？

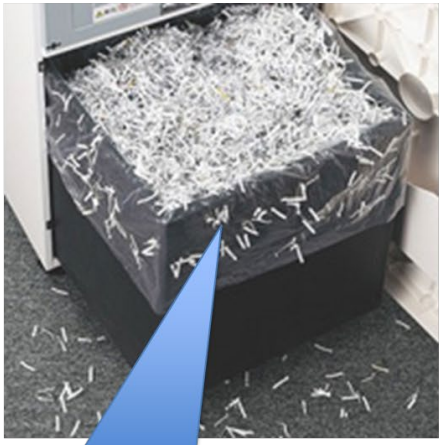
生のサンプル

DNAとして抽出
ATCGの4種の文字

次世代シークエンサー

得られるデータ ゲノムシュレッダー

100文字ぐらいの断片になった
21億ピース
の文字列断片がコンピュータに
吐き出される



※全ゲノムシークエンスを想定

一人分のゲノムは、**4種類の塩基**（A、T、G、C）からなる塩基対、**約30億個分**からなります。現在主流のシークエンサーでは、一度にはDNA中の**約100塩基対分**しか読めません。これは**ATGCの4文字からなる30億文字分の文章が書かれた書類**の束が、シュレッダーにかけられ**100文字ずつの断片**になって出てくるようなものです。通常、がん細胞特有の変異を見つける場合、**がん細胞由来の書類を40コピー分**、**正常細胞由来の書類を30コピー分**、シークエンサーで読み取ります。つまり**21億ピースの文字列断片**が得られます。**ここから各ピースがゲノム上のどこにあったかを正確に決定し、がん細胞と正常細胞の違いを見つけ出す（変異を検出する）必要があります。**

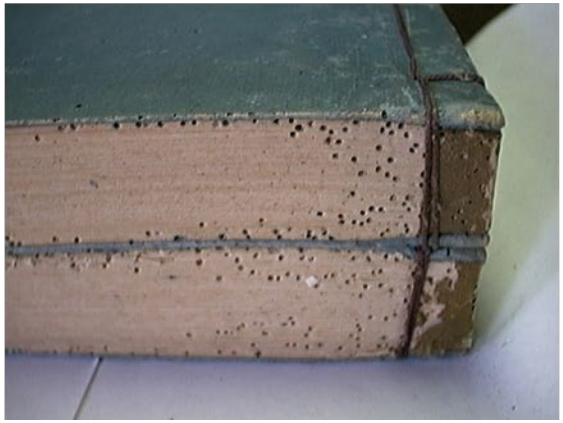
「変異」を探し出せ

1文字変異

挿入や欠失

転座、逆位・・・

本にたとえれば



変異検出にチャレンジ

TP53遺伝子（最も有名ながん抑制遺伝子）の一部分。**一文字だけ配列が異なります。**

正常細胞のDNA配列

```
gatgggattg gggttttccc ctcccatgtg ctcaagactg gcgctaaaag ttttgagctt
ctcaaaagtc tagagccacc gtccagggag caggtagctg ctgggctccg gggacacttt
gcgttcgggc tgggagcgtg ctttccacga cggtgacacg cttccctgga ttgggtaagc
tcctgactga acttgatgag tcctctctga gtcacgggct ctcggctccg tgtattttca
gctcgggaaa atcgctgggg ctgggggtgg ggcagtgggg acttagcgag tttgggggtg
agtgggatgg aagcttggct agagggatca tcataggagt tgcattgttg ggagacctgg
gtgtagatga tggggatgtt aggaccatcc gaactcaaag ttgaacgcct aggcagagga
gtggagcttt ggggaacctt gagccggcct aaagcgtact tctttgcaca tccacccggt
gctgggcgta gggaatccct gaaataaaag atgcacaaag cattgaggtc tgagactttt
ggatctcgaa acattgagaa ctcatagctg tatattttag agcccatggc atcctagtga
aaactggggc tccattccga aatgatcatt tgggggtgat ccggggagcc caagctgcta
aggtcccaca acttccggac ctttgtcctt cctggagcga tctttccagg cagccccggg
ctccgctaga tggagaaaat ccaattgaag gctgtcagtc gtggaagtga gaagtgctaa
accaggggtt tgcccgccag gccgaggagg accgtcgcaa tctgagaggc ccggcagccc
tgttattgtt tggctccaca tttacatttc tgcctcttgc agcagcattt ccggtttctt
tttgccggag cagctcacta ttcacccgat gagaggggag gagagagaga gaaaatgtcc
tttaggccgg ttcctcttac ttggcagagg gaggctgcta ttctccgcct gcatttcttt
ttctggatta cttagttatg gcctttgcaa aggcaggggt atttgttttg atgcaaacct
caatccctcc ccttctttga atggtgtgcc ccaccccgcg ggtcgcctgc aacctaggcg
gacgctacca tggcgtgaga cagggaggga aagaagtgtg cagaaggcaa gcccggaggt
attttcaaga atgagtatat ctcatcttcc cggaggaaaa aaaaaaagaa tgggtacgtc
tgagaatcaa attttgaaag agtgcaatga tgggtcgttt gataatttgt cggaaaaaca
atctacctgt tatctagctt tgggctaggc cattccagtt ccagacgcag gctgaacgtc
gtgaagcgga aggggcgggc ccgcaggcgt ccgtgtggtc ctccgtgcag ccctccggcc
cgagccggtt cttcctggta ggaggcggaa ctcgaattca tttctcccgc tgccccatct
cttagctcgc ggttgtttca ttccgcagtt tcttcccatg cacctgccgc gtaccggcca
ctttgtgccg tacttacgtc atcttttcc taaatcgagg tggcatttac acacagcgcc
agtgcacaca gcaagtgcac aggaagatga gttttggccc ctaaccgctc cgtgatgcct
accaagtcac agacccttt catcgtccca gaaacgtttc atcacgtctc ttcccagtcg
attcccgacc ccaccttat tttgatctcc ataaccattt tgcctgttgg agaacttcat
atagaatgga atcaggctgg gcgctgtggc tcacgcctgc actttgggag gccgaggcgg
gcggattact tgaggatagg agttccagac cagcgtggcc aacgtggtga atccccgtct
ctactaaaaa atacaaaaat tagctgggcg tggtgggtgc ctgtaatccc agctattcgg
gagggtgagg caggagaatc gcttgaaccc gggaggcaga ggttgcagtg agccaagatc
```

がん細胞のDNA配列

```
gatgggattg gggttttccc ctcccatgtg ctcaagactg gcgctaaaag ttttgagctt
ctcaaaagtc tagagccacc gtccagggag caggtagctg ctgggctccg gggacacttt
gcgttcgggc tgggagcgtg ctttccacga cggtgacacg cttccctgga ttgggtaagc
tcctgactga acttgatgag tcctctctga gtcacgggct ctcggctccg tgtattttca
gctcgggaaa atcgctgggg ctgggggtgg ggcagtgggg acttagcgag tttgggggtg
agtgggatgg aagcttggct agagggatca tcataggagt tgcattgttg ggagacctgg
gtgtagatga tggggatgtt aggaccatcc gaactcaaag ttgaacgcct aggcagagga
gtggagcttt ggggaacctt gagccggcct aaagcgtact tctttgcaca tccacccggt
gctgggcgta gggaatccct gaaataaaag atgcacaaag cattgaggtc tgagactttt
ggatctcgaa acattgagaa ctcatagctg tatattttag agcccatggc atcctagtga
aaactggggc tccattccga aatgatcatt tgggggtgat ccggggagcc caagctgcta
aggtcccaca acttccggac ctttgtcctt cctggagcga tctttccagg cagccccggg
ctccgctaga tggagaaaat ccaattgaag gctgtcagtc gtggaagtga gaagtgctaa
accaggggtt tgcccgccag gacgaggagg accgtcgcaa tctgagaggc ccggcagccc
tgttattgtt tggctccaca tttacatttc tgcctcttgc agcagcattt ccggtttctt
tttgccggag cagctcacta ttcacccgat gagaggggag gagagagaga gaaaatgtcc
tttaggccgg ttcctcttac ttggcagagg gaggctgcta ttctccgcct gcatttcttt
ttctggatta cttagttatg gcctttgcaa aggcaggggt atttgttttg atgcaaacct
caatccctcc ccttctttga atggtgtgcc ccaccccgcg ggtcgcctgc aacctaggcg
gacgctacca tggcgtgaga cagggaggga aagaagtgtg cagaaggcaa gcccggaggt
attttcaaga atgagtatat ctcatcttcc cggaggaaaa aaaaaaagaa tgggtacgtc
tgagaatcaa attttgaaag agtgcaatga tgggtcgttt gataatttgt cggaaaaaca
atctacctgt tatctagctt tgggctaggc cattccagtt ccagacgcag gctgaacgtc
gtgaagcgga aggggcgggc ccgcaggcgt ccgtgtggtc ctccgtgcag ccctccggcc
cgagccggtt cttcctggta ggaggcggaa ctcgaattca tttctcccgc tgccccatct
cttagctcgc ggttgtttca ttccgcagtt tcttcccatg cacctgccgc gtaccggcca
ctttgtgccg tacttacgtc atcttttcc taaatcgagg tggcatttac acacagcgcc
agtgcacaca gcaagtgcac aggaagatga gttttggccc ctaaccgctc cgtgatgcct
accaagtcac agacccttt catcgtccca gaaacgtttc atcacgtctc ttcccagtcg
attcccgacc ccaccttat tttgatctcc ataaccattt tgcctgttgg agaacttcat
atagaatgga atcaggctgg gcgctgtggc tcacgcctgc actttgggag gccgaggcgg
gcggattact tgaggatagg agttccagac cagcgtggcc aacgtggtga atccccgtct
ctactaaaaa atacaaaaat tagctgggcg tggtgggtgc ctgtaatccc agctattcgg
gagggtgagg caggagaatc gcttgaaccc gggaggcaga ggttgcagtg agccaagatc
```

変異検出にチャレンジ（答え）

TP53遺伝子（最も有名ながん抑制遺伝子）の一部分

正常細胞のDNA配列

```
gatgggattg gggttttccc ctcccatgtg ctcaagactg gcgctaaaag ttttgagctt
ctcaaaagtc tagagccacc gtccagggag caggtagctg ctgggctccg gggacacttt
gcgttcgggc tgggagcgtg ctttccacga cggtgacacg cttccctgga ttgggtaagc
tcctgactga acttgatgag tcctctctga gtcacgggct ctcggctccg tgtattttca
gctcgggaaa atcgctgggg ctgggggtgg ggcagtgggg acttagcgag tttgggggtg
agtgggatgg aagcttggct agagggatca tcataggagt tgcattgttg ggagacctgg
gtgtagatga tggggatgtt aggaccatcc gaactcaaag ttgaacgcct aggcagagga
gtggagcttt ggggaacctt gagccggcct aaagcgtact tctttgcaca tccacccggt
gctgggcgta gggaatccct gaaataaaag atgcacaaag cattgaggtc tgagactttt
ggatctcgaa acattgagaa ctcatagctg tatattttag agcccatggc atcctagtga
aaactggggc tccattccga aatgatcatt tgggggtgat ccggggagcc caagctgcta
aggtcccaca acttccggac ctttgtcctt cctggagcga tctttccagg cagccccccgg
ctccgctaga tggagaaaat ccaattgaag gctgtcagtc gtggaagtga gaagtgctaa

accaggggtt tgcccgccag gc[c]aggc accgtcgcaa tctgagaggc ccggcagccc
tgttattgtt tggctccaca tttacctcttgc agcagcattt ccggtttctt
tttgccggag cagctcacta ttcacccgat gagaggggag gagagagaga gaaaatgtcc
tttaggccgg ttcctcttac ttggcagagg gaggctgcta ttctccgcct gcatttcttt
ttctggatta cttagttatg gcctttgcaa aggcagggtt atttgttttg atgcaaacct
caatccctcc ccttctttga atggtgtgcc ccacccccgcg ggtcgcctgc aacctaggcg
gacgctacca tggcgtgaga cagggaggga aagaagtgtg cagaaggcaa gccccggaggt
attttcaaga atgagtatat ctcatcttcc cggaggaaaa aaaaaaagaa tgggtacgtc
tgagaatcaa attttgaaag agtgtcgttt gataatttgt cggaaaaaca
atctacctgt tatctagctt tgccgtccagtt ccagacgcag gctgaacgtc
gtgaagcgga aggggcgggc ccgtgtggtc ctccgtgcag ccctccggcc
cgagccggtt cttcctggta ggagaattca tttctcccgc tgccccatct
cttagctcgc ggttgtttca ttcctcccatg cacctgccgc gtaccggcca
ctttgtgccg tacttacgtc atcatcgagg tggcatttac acacagcgcc
agtgcacaca gcaagtgcac aggttggccc ctaaccgctc cgtgatgcct
accaagtcac agacccttt catacgtttc atcacgtctc ttcccagtcg
attcccgacc ccaccttat tttaccattt tgcctgttgg agaacttcat
atagaatgga atcaggctgg gcgcgctgc actttgggag gccgaggcgg
gcggattact tgaggatagg agtcgtggcc aacgtggtga atccccgtct
ctactaaaaa atacaaaaat tacgggtgc ctgtaatccc agctattcgg
gagggtgagg caggagaatc gctaggcaga ggttgcagtg agccaagatc
```

答え

このような変異を、大量の文字列ピースデータから見つけ出す必要があります。

でも、どうやって見つけるのでしょうか？

人が手で並べて、目で見つける？

いえ、コンピューターで検出します。

gcc>gac

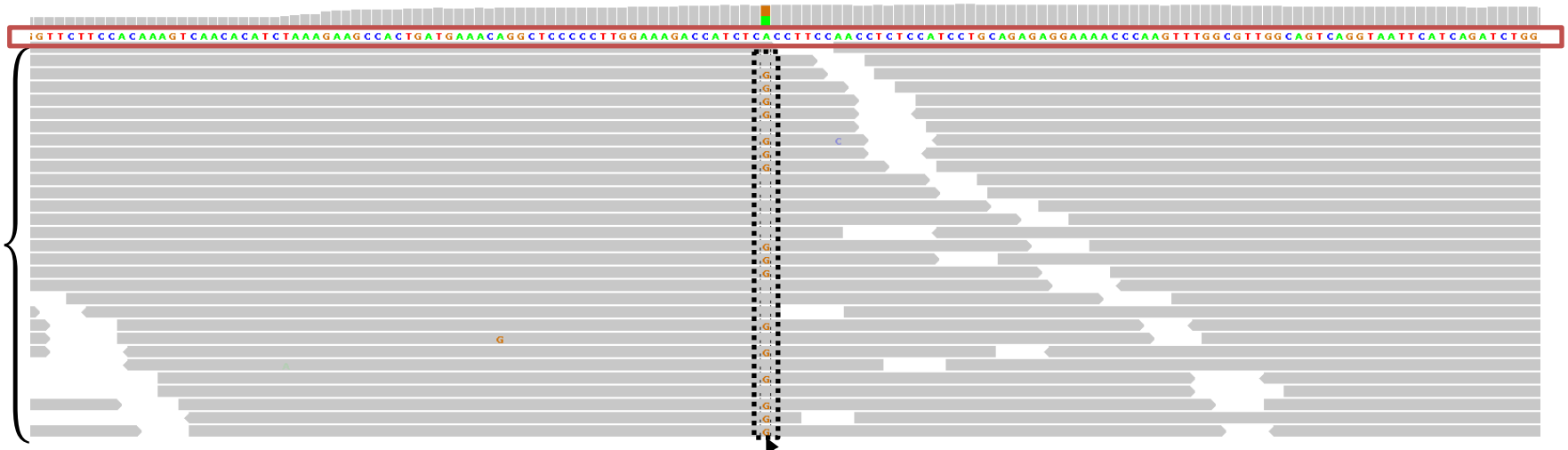
アラニン>アスパラギン

DNA文字列ピースデータからの変異検出原理

文字列ピースデータ
（約100文字；21億ピース）

ステップ1：
スーパーコンピュータを使って
ヒトゲノム標準配列（30億文字）
上でマッチする場所を探索します
（アライメント）。

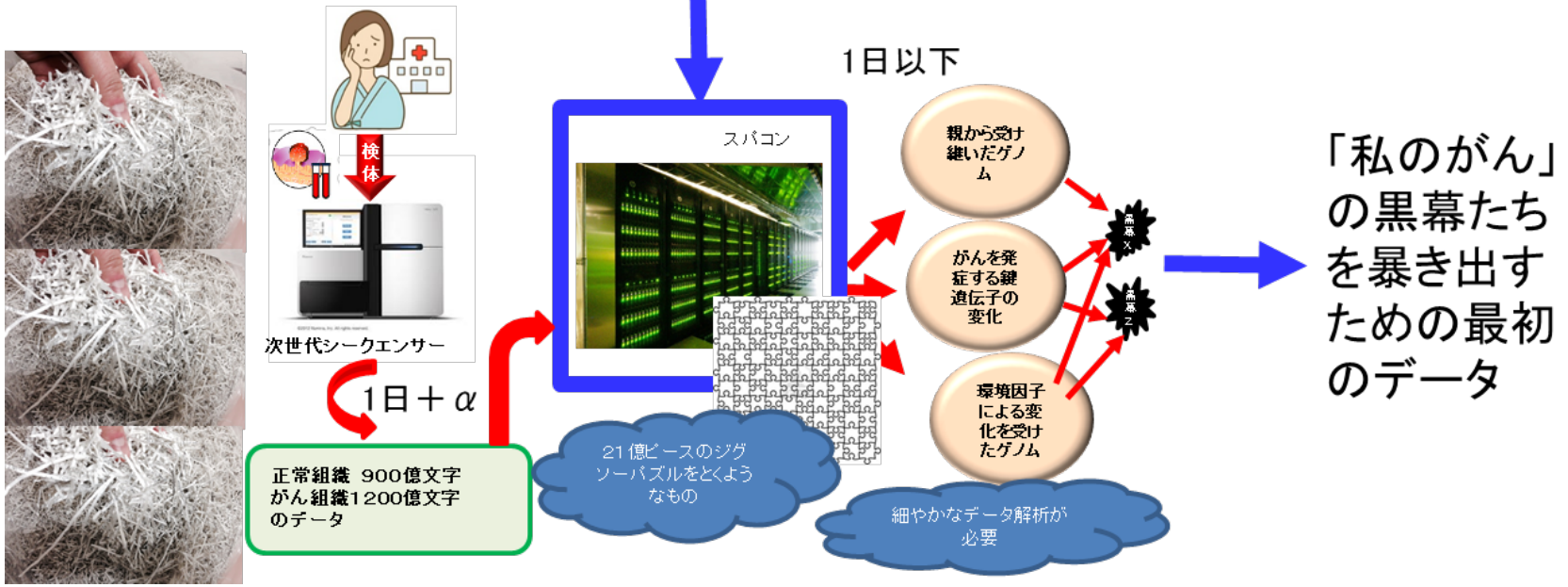
ヒトゲノム標準配列（30億文字）



アライメントされた文字列ピースデータ。
上段の標準配列と同じ文字（塩基）で
あれば灰色で表示。

ステップ2：
アライメントされた文字列ピースデータ中で、
標準配列と異なる文字が複数回観測されている場所を検出します。この例では標準配列がAのところが、Gに置き換わっています。
（一塩基変異）

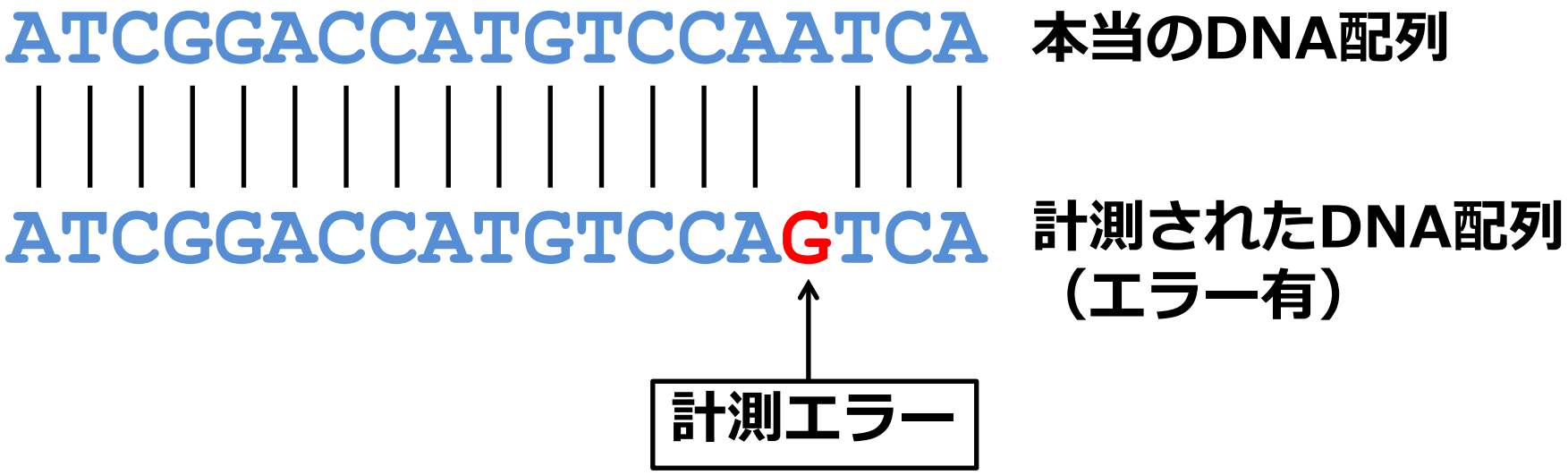
スパコンで21億ピースのジグソーパズルを解き、がんのシステム異常の原因を暴き出さねばならない！



一塩基変異だけでなく、欠失、増幅、転座などの変異もあります。それらをスーパーコンピュータと変異検出アルゴリズム（多くは統計的手法に基づく）を用いて、高精度かつ高速に検出する必要があります。
最近ではGPUと呼ばれる演算装置を活用した計算の高速化も進んでいます。

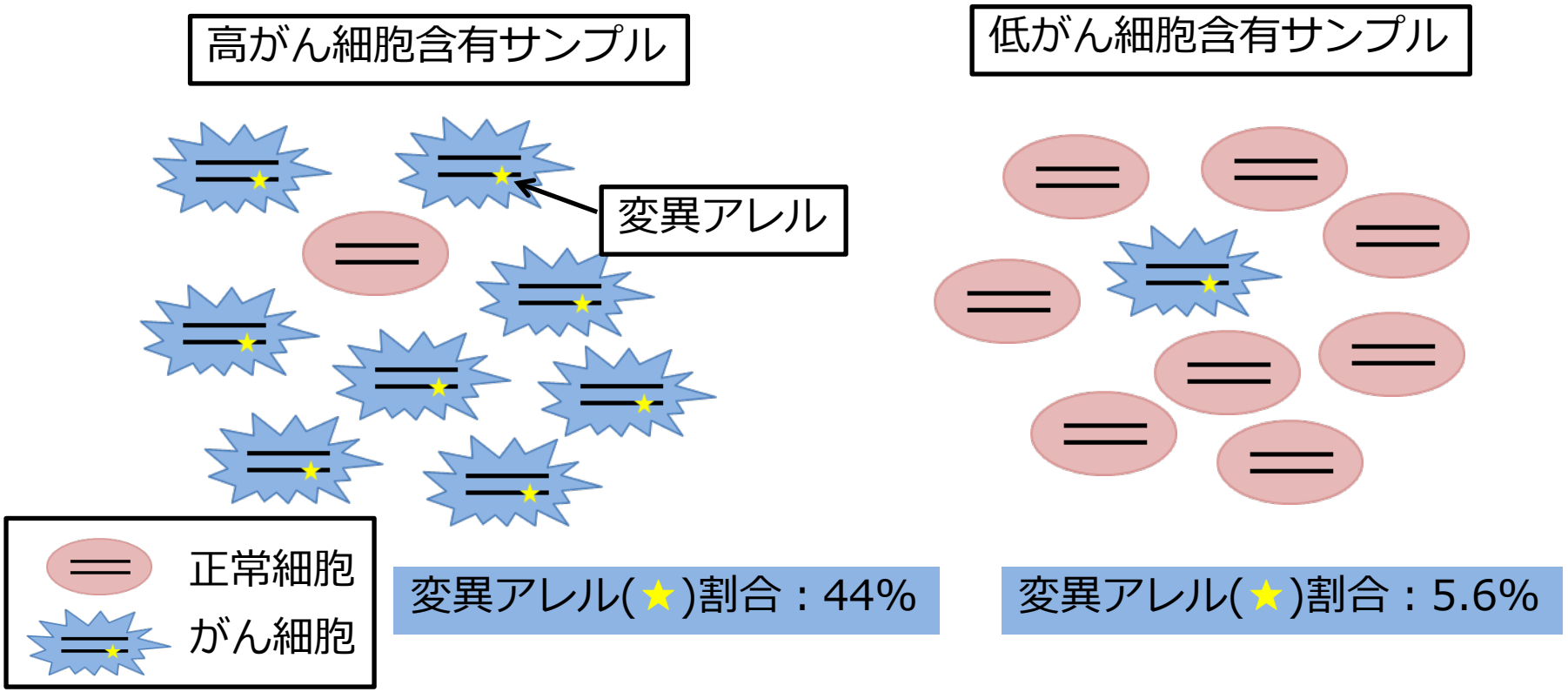
変異検出の難しい点（１）
計測エラーの混入

- DNA配列のシークエンサーでの計測時に**計測エラー**が入ることがあります
  - 本当の変異がある位置を検出する際のノイズとなります



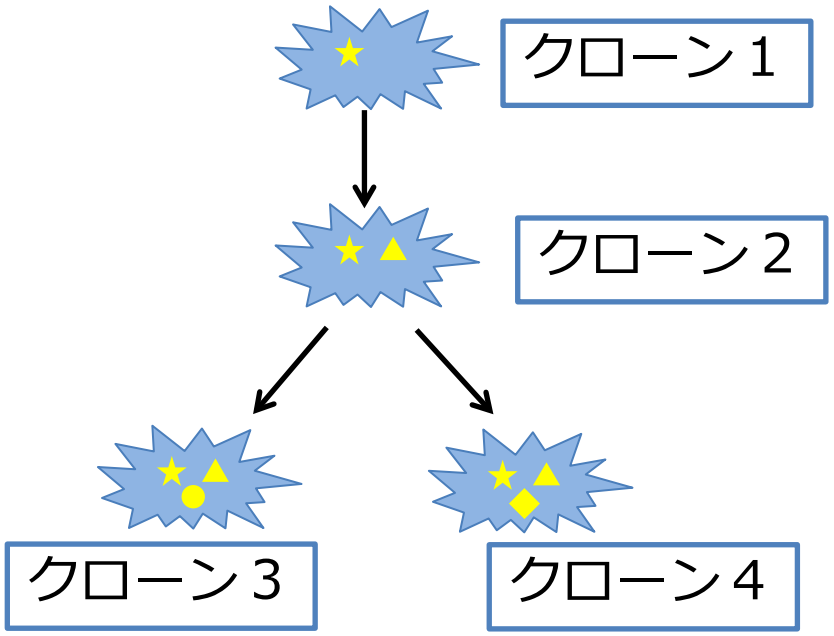
変異検出の難しい点（２）
がん細胞含有率が低い場合

- サンプル中の**がん細胞含有量が少ない場合**があります
  - がん組織中には正常な細胞も含まれます
  - **変異アレル観測割合**が相対的に少なくなり、エラーとの区別が難しくなります



変異検出の難しい点（３）
がん細胞クローンの存在

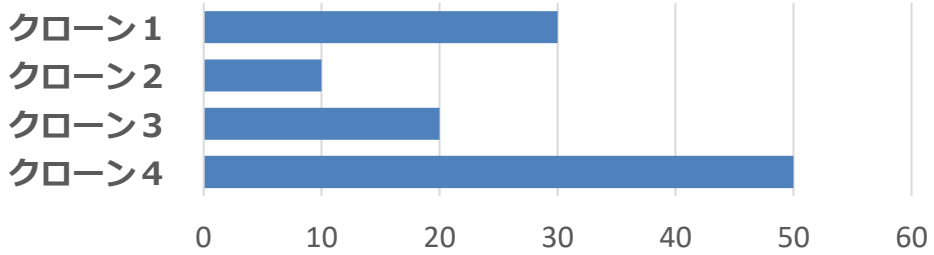
- がん細胞は**進化**します。新しい変異を獲得したがん細胞クローンが出現することがあります
  - 変異の獲得時期、クローンの存在量に応じて変異の観測頻度が変わります
  - 存在量の少ないクローン固有の変異の検出は難しい



変異

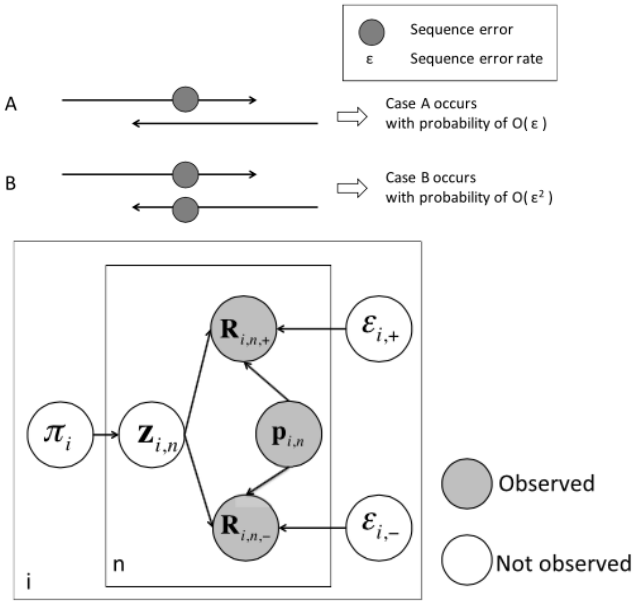
	★	▲	●	◆
クローン１	✓			
クローン２	✓	✓		
クローン３	✓	✓	✓	
クローン４	✓	✓		✓

クローン存在量（例）



データから有用な情報を抽出する手法の開発が重要

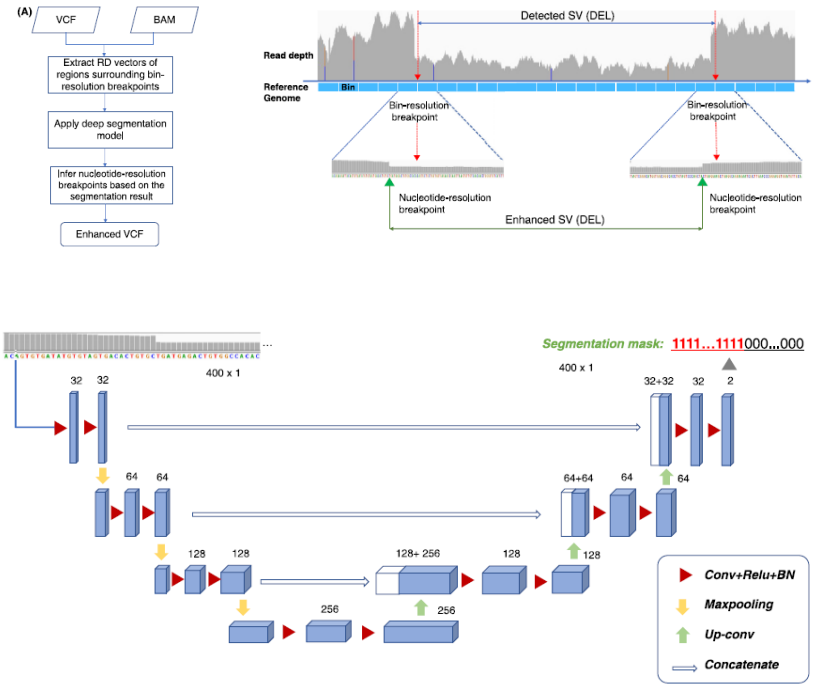
ベイズ統計モデル × **深層学習モデル**



$$p(\mathcal{R}_i, \mathcal{Z}_i|\boldsymbol{\gamma}_i, \boldsymbol{\alpha}_{i,+}, \boldsymbol{\alpha}_{i,-}) = p(\pi_i|\boldsymbol{\gamma}_i)p(\epsilon_{i,+}|\boldsymbol{\alpha}_{i,+})p(\epsilon_{i,-}|\boldsymbol{\alpha}_{i,-}) \cdot \prod_n p(\boldsymbol{R}_{i,n,+}, \boldsymbol{R}_{i,n,-}|z_{i,n}, \epsilon_{\pm,i}, \boldsymbol{\pi}_{i,n}, \boldsymbol{p}_{i,n})p(z_{i,n}|\pi_i)$$

短い塩基変異の検出法

Moriyama et al., Bioinformatics, 2019

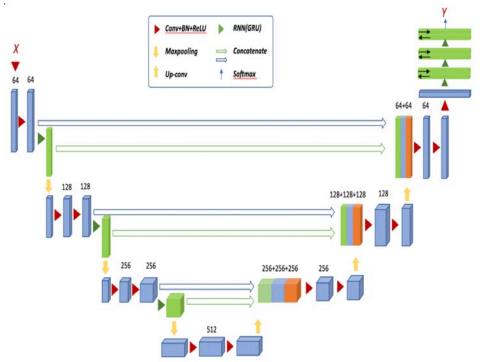


長い塩基変異の検出法

Zhang et al., PLoS Compt Biol, 2021

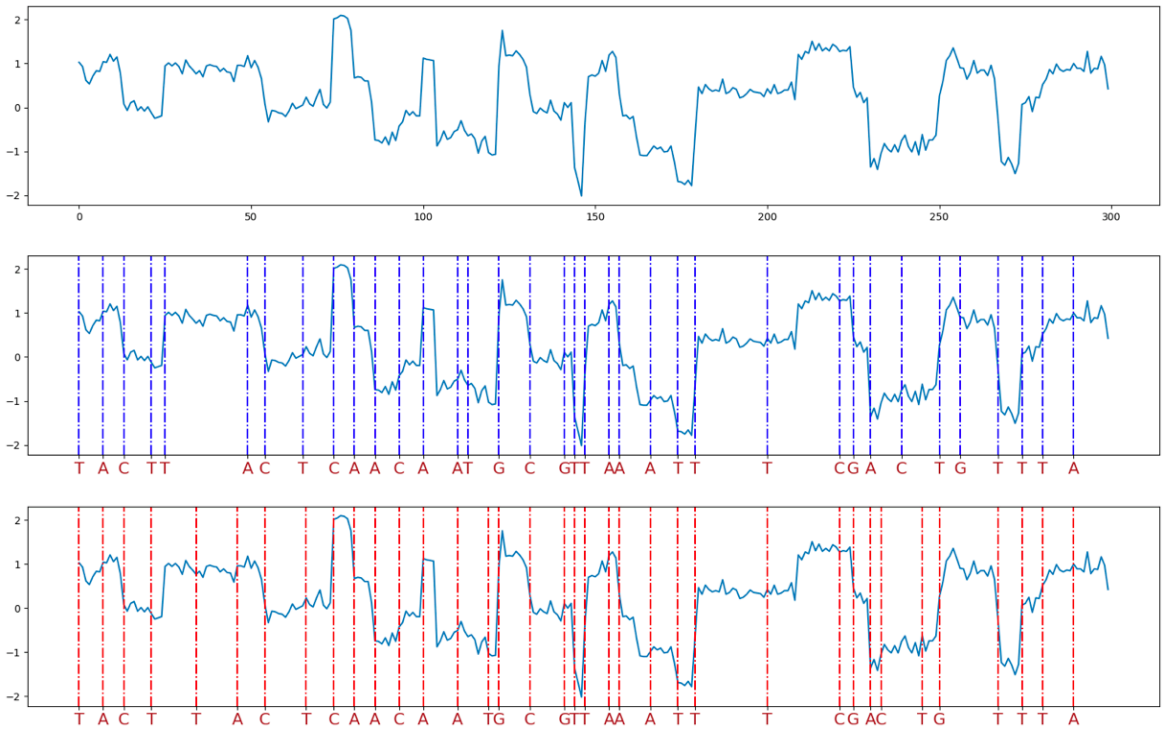
前のページで説明した困難を克服して、高精度に変異を検出する必要があります。そのために、**ベイズ統計モデル**に基づく手法や、**深層学習モデル（AI）**に基づく手法などの開発研究を行っています。

深層学習モデル(AI)によるDNA配列推定

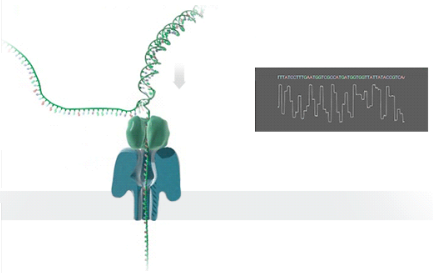


- 最後にナノポアシークエンサーと呼ばれる、新しい計測装置から得られるデータの解析手法の紹介をします
- 従来の装置に比べて、長いDNA断片(1000塩基以上)を計測できる反面、まだエラーが大きい問題があります
- 計測データ（電流値）の複雑なパターンからDNAの塩基を推定するのが難しいからです
- 我々は新しい深層学習モデル(URnano)を開発しました
- AIは大量のデータを学習することで、複雑な問題を解決することができます

Zhang et al., BMC Bioinformatics, 2019



計測データ（電流値）



AIの予測塩基ラベル

正解塩基ラベル

同じ塩基であっても得られる電流パターンは長さも形も様々

おわりに

がんゲノムの変異を検出するため用いる、観測データ、変異検出の原理、計算資源、困難な点、それを克服するためのデータ解析手法などについて概説しました

次世代シークエンサーによる計測データは、がんのゲノム変異の検出に加えて、個人の免疫遺伝子型決定や、免疫細胞クローンの多様性の推定にも用いられ、がん免疫療法の開発にとっても不可欠なものとなっています

今後もシークエンサーの性能向上に伴い、新たな種類のデータが産生されると期待されます。それに応じて、新たな解析手法の開発を行う必要があります。ここではナノポアシークエンサーからのデータ解析手法についても紹介しました

また大量のデータ解析結果を、いかに医療に有用な情報へ翻訳し還元するかということも課題となっており、人工知能を活用する研究が進みつつあります