AI 特許紹介(17)
AI 特許を学ぶ！究める！
～ゲノム解析パイプライン～

2020 年 6 月 18 日
河野特許事務所
所長　弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要
特許権者　Sentieon
出願日　2016 年 3 月 24 日
登録日　2019 年 9 月 24 日
登録番号 US10424396
発明の名称　位置依存変異コールの計算パイプライン

396 特許は、計算パイプラインを使用して複数の核酸配列リードから、変異を分析するものであり、ゲノム内の複数のリード位置に応じて、異なる事前確率を使用することにより変異コール[1]精度を大幅に向上させるアイデアである。

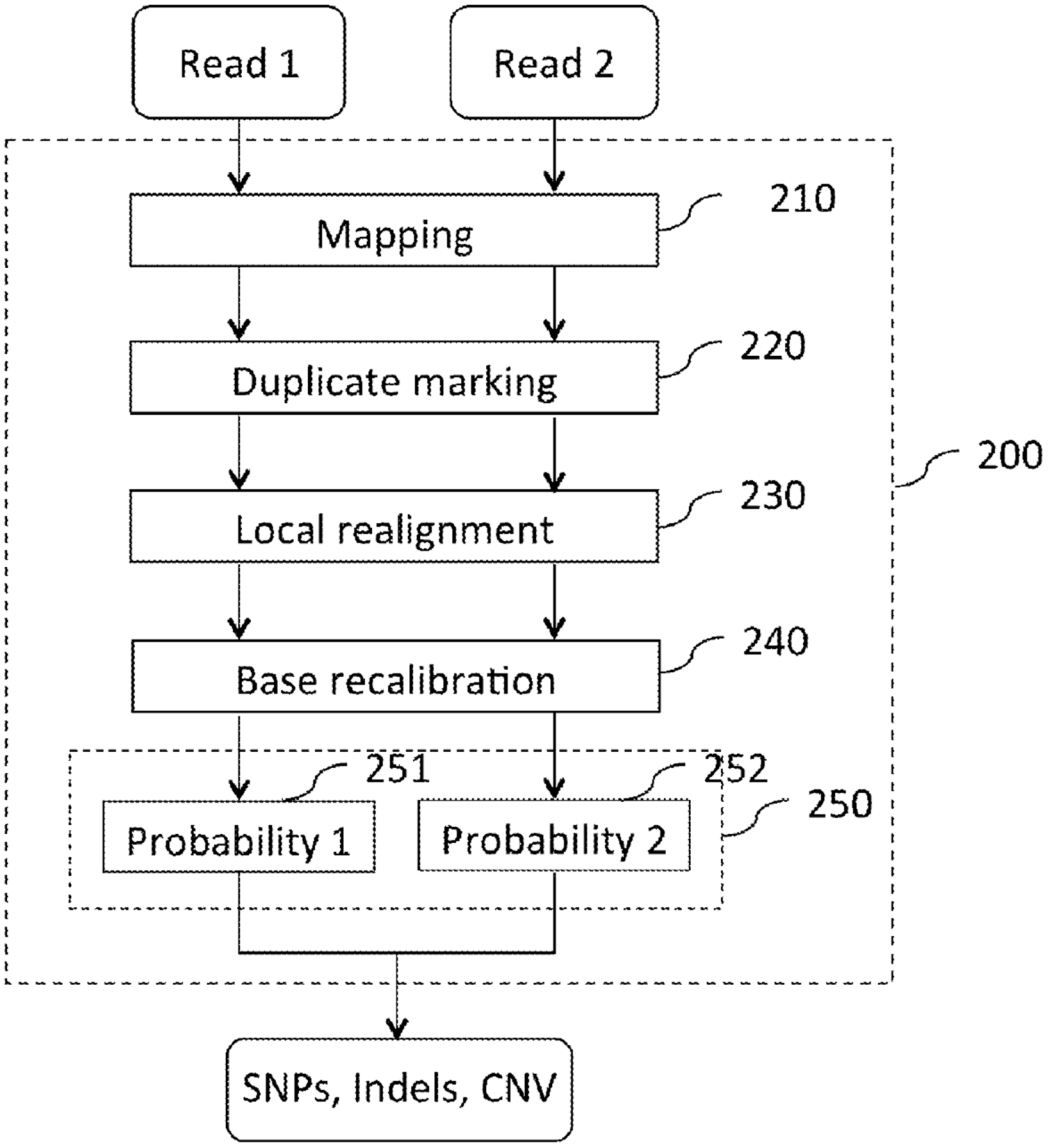
2.特許内容の説明
次世代ゲノムシーケンシング（NGS : Next-generation sequencing）技術により、個別化医療を可能とする膨大な量の生物学的データを取得することができるようになった。ハイスループットゲノムシーケンスのコストは、単にシーケンスデータを取得する

[1] 変異コールとは解析対象の塩基配列中、どの位置にどのような変異が生じているかを特定することをいう。

という点で減少したが、依然としてこれらの大規模なシーケンスデータの分析が精度及び速度の面で、大きな課題となっている。

396 特許は計算パイプラインを用いて、解析対象の塩基配列中のどの位置にどのような変異が発生しているかを特定する。具体的には、SNP(Single Nucleotide Polymorphism 一塩基多型)、Indel(insertion/deletion:塩基の挿入または欠損による遺伝的変異)、CNV（Copy Number Variation：コピー数多型。ある集団のなかで 1 細胞あたりのコピー数が個人間で異なるゲノムの領域）等の変異を検出する。

396 特許の計算パイプライン 200 は以下の通りである。



通常のパイプラインと異なり、異なる位置にマッピングされる２つの核酸配列リードを用いる。そして複数の事前確率により変異コールが行われる。詳細なフローは以下の通りである。

リード１及びリード２を含む生のリードデータは、短いシーケンシングリードを参照ゲノムにアライメントさせるためにマッピングモジュール 210 に供給される。リード１は参照ゲノムの位置１（たとえば、エクソン）にマップされ、リード２は参照ゲノムの位置２（たとえば、イントロン[2]）にマップされる。

マッピングモジュール 210 の出力は複製マーキングモジュール 220 に供給され、その出力はローカル再アライメントモジュール 230 に供給される。複製マーキングモジュール 220 は、PCR(Polymerase Chain Reaction)複製を除去する。シーケンシング用の DNA サンプルの準備中に、PCR を使用してフラグメントを増幅し、複製を生成することがよくあるためである。

ローカル再アライメントモジュール 230 は、リードのアラインメントを改善する。通常、再アライメントは、参照に対するリードの挿入と削除（Indel）の周囲の領域で行われ、リードを、Indel の一方の端に、残りの端をもう一方の端にマッピングする。ローカル再アライメントモジュール 230 の出力は、ベース再キャリブレーションモジュール 240 に供給される。

ベース再キャリブレーションモジュール 240 は、すべてのリードにおいて、各ベースの経験的に正確なベース品質スコアを提供する。ベース再キャリブレーションモジュール 240 は、機械学習を適用して品質スコアのエラーを経験的にモデル化し、それに応じて品質スコアを調整する。ベース再キャリブレーションモジュール 240 の出力は、変異コールモジュール 250 に供給される。

入力データが変異コールモジュール 250 を介して渡されるとき、値は、リードの位置に基づいて、例えば、事前確率(prior probabilities)等のパラメータに設定される。

変異コールモジュール 250 は、SNP、ショートインデル、CNV を含むリード間に存在する代替対立遺伝子の統計的証拠を持つすべての位置を検出する。通常、変異コールモジュール 250 には、アルゴリズムまたは統計モデルが使用される。これは通常、１つ

[2] エクソンは遺伝子中においてタンパク質を作るための情報持つ部分、イントロンはタンパク質を作るための情報を持たない部分をいう。

以上のランダム変数と、場合によっては他の非ランダム変数を関連付ける数式によって特定される。たとえば、リード深度と変異カウントに基づいて、特定の位置の特定の変異が真陽性であるという信頼レベルを示す確率値は、統計モデルに基づく方法と参照サンプルを使用するローカライズされた方法を使用して計算される。例えば、GATK(Genome Analysis Toolkit)またはAtlas２等の機械学習モデルが採用される。

GATKは、単純なベイジアンモデリングの並列プログラミングにMapReduceの考えを取り入れている。Atlas2は、通常の尤度計算ではなく、検証済みの全エクソームキャプチャシーケンスデータでトレーニングされたロジスティック回帰モデルを採用している。

図示されているように、リード１が変異コールモジュール250を通過すると、第１の値が事前確率(例えばエクソン領域のSNP確率0.0005)に設定される。リード2が変異コールモジュール250を通過すると、第2の値が事前確率（たとえば、イントロン領域のSNP確率0.0008）に設定される。その後、変異コールが生成される。

現在使用されている典型的な変異コールパイプラインでは、単一の事前確率（たとえば、全体の平均変異頻度）がゲノム全体で使用される。その結果、事前確率が低い領域（例えば、エクソン）で余分な不要な変異が呼び出され、事前確率の高い領域（例えば、イントロン）での変異コールが失われる。

ゲノム全体にわたって値のセットが１つしかない場合、事前確率パラメータを調整しても、ゲノム全体で同時に最適な呼び出しを行うことができない。

しかしながら396特許では、ゲノム内のリード位置に応じて、異なる事前確率を使用することにより、各変異は、エクソンまたはイントロン領域などのゲノム内の位置に依存する最適な事前確率を使用して呼び出される。このような位置に依存する変異コールにより、変異コールの精度を大幅に向上することができる。

3.クレーム

396特許のクレーム1-5は以下の通りである。

1.　計算パイプラインを使用して複数の核酸配列リードから、変異を分析するためのコンピュータ実装方法において、前記計算パイプラインは、SNP、または、挿入および削除（Indelインデル）確率のパラメータを含み、前記方法は、プロセッサ上で以下のステップを実行することを含む。　：

少なくとも第１の核酸配列リードおよび第２の核酸配列リードを含む少なくとも

1000 個の核酸配列リードを受け取り、

少なくとも 1000 の核酸配列リードをそれぞれゲノム内の位置にマッピングし、

第 1 の核酸配列リードは第 1 の位置にマッピングされ、

第 2 の核酸配列リードは前記第 1 の位置とは異なる第 2 の位置にマッピングされ、

ゲノム内の位置に基づいて、それぞれ SNP またはインデル確率のパラメータの値を設定し、第 1 の位置に設定された第 1 の値は、第 2 の位置に設定された第 2 の値とは異なり、

SNP またはインデル確率のパラメータに設定された値のセットを使用して、少なくとも 1000 の核酸配列リードを計算パイプラインに渡し、

変異コールを生成する。

2. クレーム 1 の方法において、第 1 および／または第 2 の位置は、エクソンまたはイントロン内にある。

3. クレーム 1 の方法において、ゲノムは民族集団または地域集団から得られる。

4. クレーム 1 の方法において、ゲノムは健康な対象または疾患の対象から得られる。

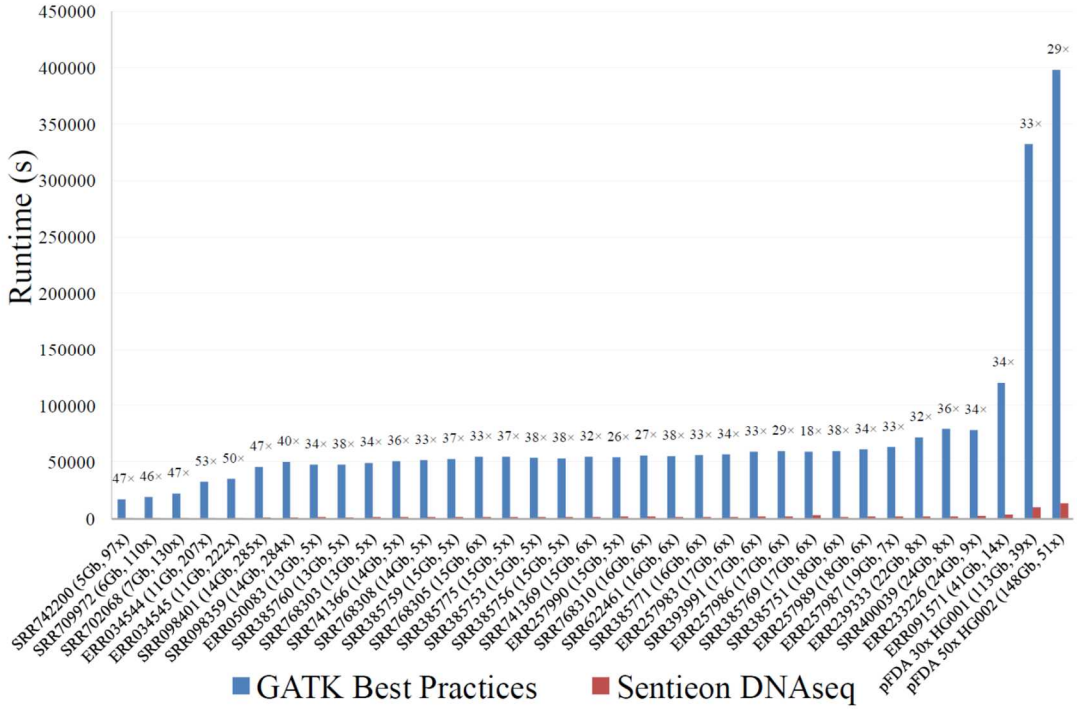
5. クレーム 4 の方法において、疾患はガンである。

4. ゲノム解析パイプラインに関する論文

Sentieon 社は 2014 年米国カリフォルニア州に設立された次世代シーケンサーデータを解析するソフトウェア企業である。機械学習及び統計手法を用い解析の高速化、解析データの品質向上を図っている。

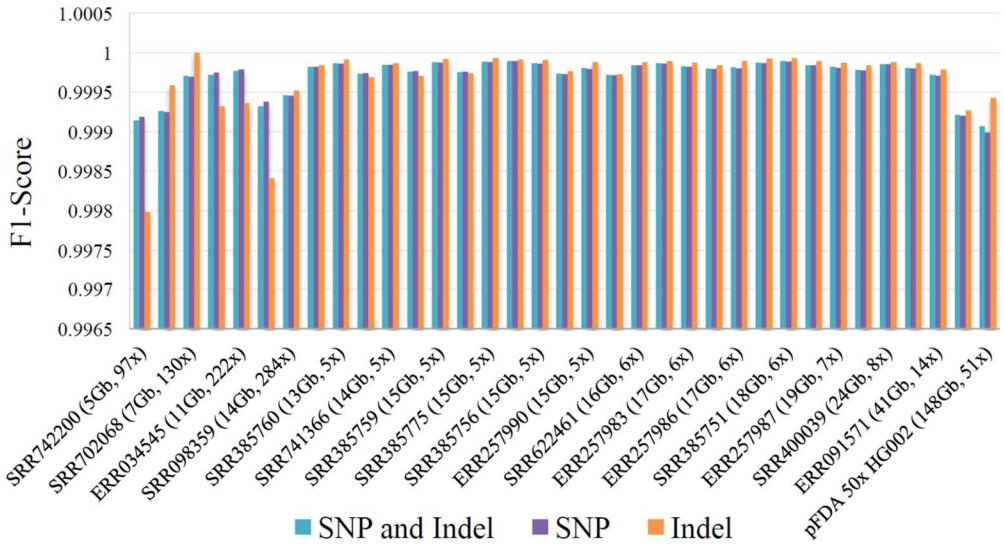
本特許に関連する論文[3]が Sentieon 社の Donald Freed 氏らにより発表されている。

[3] Donald Freed, Rafael Aldana, Jessica A. Weber, Jeremy S. Edwards “The Sentieon Genomics Tools – A fast and accurate solution to variant calling from next-generation sequence data” https://doi.org/10.1101/115717



上記図は DNS シーケンスパイプラインのランタイムの比較を示すグラフである。青色が GATK によるランタイム、赤色が Sentieon のランタイムである。GATK と比較して 18 倍～53 倍に高速化されている。

また精度に関しても GATK とほとんど変わりないことが確認されている。



上記図は、SNP/インデル、SNP 及び SNP についての F スコアを示すグラフである。GATK と比較した平均 F スコアは 0.9996、F スコアの範囲は 0.9974 ～1.0000 を達成している。

以上

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウェア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 2.0」がある。