AI 特許紹介(18)
AI 特許を学ぶ！究める！
～DeepVariant～

2020 年 7 月 10 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　Verily Life Sciences
出願日　2017 年 4 月 18 日
登録日　2019 年 7 月 16 日
登録番号 US10354747
発明の名称　次世代シーケンシング用の深層学習分析パイプライン

747 特許は、次世代シーケンサから読み取った対立遺伝子中の候補変異(バリアント)についてパイルアップ画像を生成し、当該画像をディープラーニングを用いて解析することにより、SNP(single nucleotide polymorphism)[1]、インデル[2]等の変異を検出するアイデアである。

[1] SNP(一塩基多型)とは個人間の遺伝情報のわずかな違いのことであり、１つの塩基だけが別の塩基に置き換わっていることをいう。
[2] インデルとは、insertion/deletion の略であり、塩基の挿入または欠損による遺伝的変異のことをいう。

2.特許内容の説明

次世代シーケンシング（NGS : Next-generation sequencing）テクノロジーは、DNA フラグメントリードを並列化することによって、DNA シーケンシングのコストを大幅に低減した。一部の NGS 方法では、数百万のシーケンスリードを同時に実行し、数時間で数百万の塩基対のデータを生成することが可能である。

適切に遺伝的変異を検出することができれば、遺伝的障害及びメンデル疾患等を理解するためのフレームワークを構築することができる。747 特許ではディープラーニングを用いることにより精度良く変異を検出することを目的としている。

下記図 2 は、変異コーラーパイプライン 200 を示す説明図である。

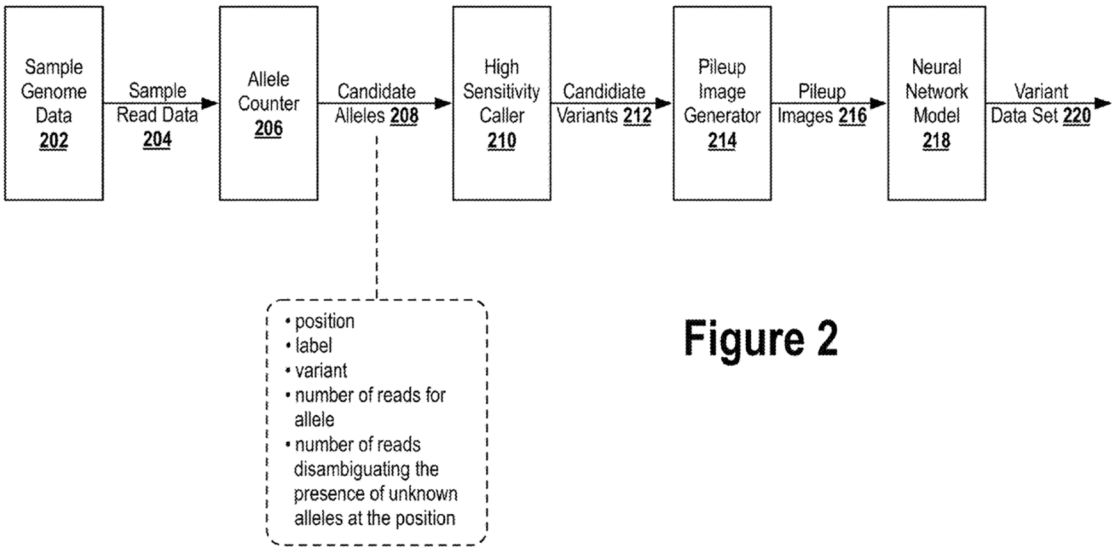


Figure 2

サンプルゲノムデータ 202 には、一連の配列リードを含むサンプルリードデータ 204(例えば、DNA シーケンサからの DNA フラグメントのヌクレオチド配列を示すデータ)が含まれる。サンプルリードデータ 204 は、対立遺伝子カウンター206 へ入力される。

(1)候補対立遺伝子の特定

ヒトゲノムにおける所定の位置は、1 つまたは複数の既知の対立遺伝子と関連している可能性があり、それぞれが染色体内に独自のヌクレオチド配列を持っている。対立遺伝子カウンター206 は、サンプルリードデータ 204 を受け取り、サンプルリードデータ 204 内の各ヌクレオチド位置について、それらのヌクレオチド位置に存在し得る候補対立遺伝子を決定する。

対立遺伝子カウンター206 は、3 つ以上の配列リードのヌクレオチド塩基が特定の対

立遺伝子（例えば、ヒトゲノムにおけるその位置での既知の対立遺伝子）と一致すると判断した場合、対立遺伝子カウンター206 は、特定の対立遺伝子が、候補対立遺伝子であると決定する。

(2)候補変異の特定

次に、候補対立遺伝子 208 中の各候補対立遺伝子について、高感度コーラー210 は、高度に許容的な閾値を有する対立遺伝子を呼び出す(calling)ことにより、独立してかつ並行して突然変異の候補部位を選択し、候補変異 212 を出力する。

(3)パイルアップ画像の生成

次に、候補変異 212 の各候補変異について、パイルアップ画像生成器 214 は、元のリードデータセットからすべての重複するリードを独立して抽出し、候補変異および関連するリードデータ(すなわち、候補変異と整列したリードを含むリードパイルアップウィンドウ)を表す画像を解釈する。

データを画像内のピクセル位置とピクセルカラー値にマッピングすると、ニューラルネットワークがパターンを識別することが可能なトレーニングデータを生成することができる。パイルアップ画像ジェネレータ 214 は、１組のパイルアップ画像 216 をレンダリングし、それらをニューラルネットワークモデル 218 に出力する。

(4)ニューラルネットワークへのパイルアップ画像の適用

読み取られた各パイルアップ画像について、訓練されたニューラルネットワークモデル 218 を使用して、候補変異部位が真の突然変異である確率として、遺伝子型構成の確率分布を決定し、変異データセット 220 を出力する。

(5)ニューラルネットワークの学習

ディープラーニングネットワークをトレーニングするために使用されるトレーニングデータは、ゲノム全体の各ヌクレオチド位置のリードパイルアップ画像と、ラベルとを含む。各変異サイトが既知のリファレンスゲノムの場合、各リードパイルアップウィンドウは、「変異」または「リファレンス」（つまり、非変異）としてラベル付けされる。

ディープラーニングネットワークのトレーニングでは、既知の変異と非変異の両方を含む各リードパイルアップ画像をディープラーニングネットワークに提供する。

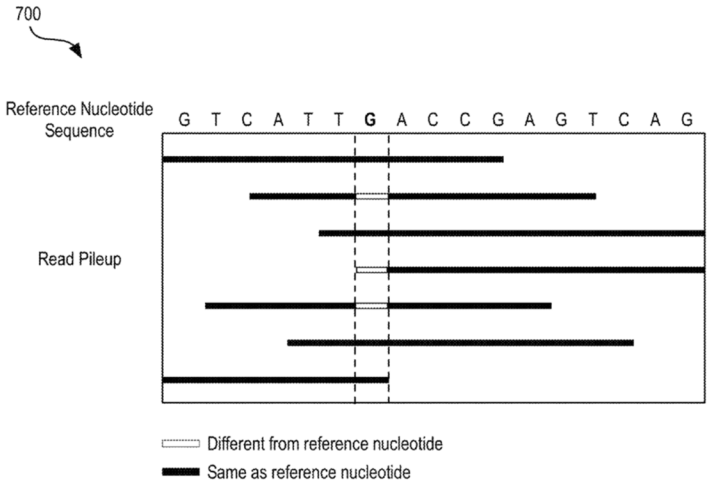


Figure 7A

図 7A は、変異コール[3]リードパイルアップウィンドウ 700 を示す。この例では、リードパイルアップウィンドウには、特定のヌクレオチド位置（点線で表示）で重複する 7 つのリードフラグメントが含まれている。

パイルアップウィンドウは、17 塩基のヌクレオチドにまたがっており、黒色のリードフラグメントは、その位置のヌクレオチドが、関連する参照ヌクレオチドシーケンスのヌクレオチドと同じであることを示す。一方、白色のリードフラグメントは、その位置のヌクレオチドが関連する参照ヌクレオチド配列のヌクレオチドとは異なることを示す。

図 7A の例では、2 番目、4 番目、5 番目のリードフラグメントが、参照ゲノム内の関連するヌクレオチド塩基とは異なるヌクレオチドを持っている。変異コーラーによっては、このリードパイルアップウィンドウが SNP 変異であると検出される。

[3] 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。

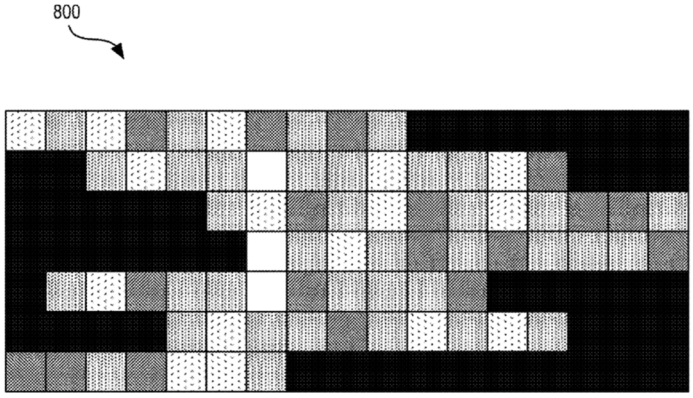


Figure 8A

図 8A は図 7A に対応する変異コールの画像表現を示す説明図である。各ピクセルには、水平位置（参照ゲノムと整列したリードフラグメント上のヌクレオチド位置）と垂直位置（特定のリードフラグメント）が含まれます。図 8A の例では、黒色のピクセルはその位置にリードフラグメントデータが存在しないことを示し、影付きのピクセルはその位置にリードフラグメントが存在することを示している。

陰影の程度は、情報の組み合わせを表す（たとえば、ヌクレオチド塩基、塩基が参照ゲノムと一致するかどうか、品質スコア等）。ピクセルに陰影が存在しない場合、その場所のヌクレオチド塩基は、参照ゲノムのその場所のヌクレオチドとは異なる。

陰影に基づいて、ニューラルネットワークは、画像内のパターンを識別する。例えば、画像 800 中の 3 つの陰影のない（白色）正方形は、関連する参照ヌクレオチドとの不一致を示す。多くのサンプル画像（画像 800 など）を含むデータセットを使用してトレーニングされたニューラルネットワークは、画像の陰影または色の特徴を検出し、誤検出変異または真陽性変異を特定する。

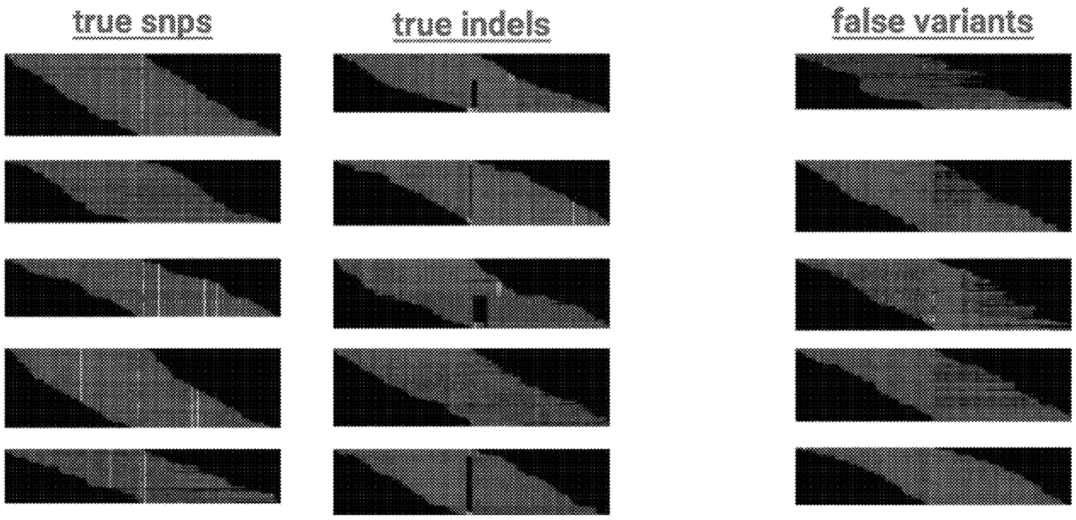


Figure 8B

図 8B における「true snps」の画像は、SNP を含むリードパイルアップウィンドウを表し、「true indels」の画像は挿入または削除を含むリードパイルアップウィンドウを表し、「false variants」の画像は偽陽性変異コールのリードパイルアップウィンドウを表す。

3.クレーム

747 特許のクレーム 1 は以下の通りである。

1. それぞれがサンプルゲノム内のヌクレオチド位置に整列したヌクレオチドを含む複数の配列リードを取得し、

ヌクレオチド位置に関連する複数の対立遺伝子を取得し、

複数の対立遺伝子のヌクレオチド位置に位置する特定の対立遺伝子が、複数の配列リードのうちの 1 つまたは複数の配列リードと一致することを決定し、

複数の配列リードに関連する情報に基づいて画像を生成し、

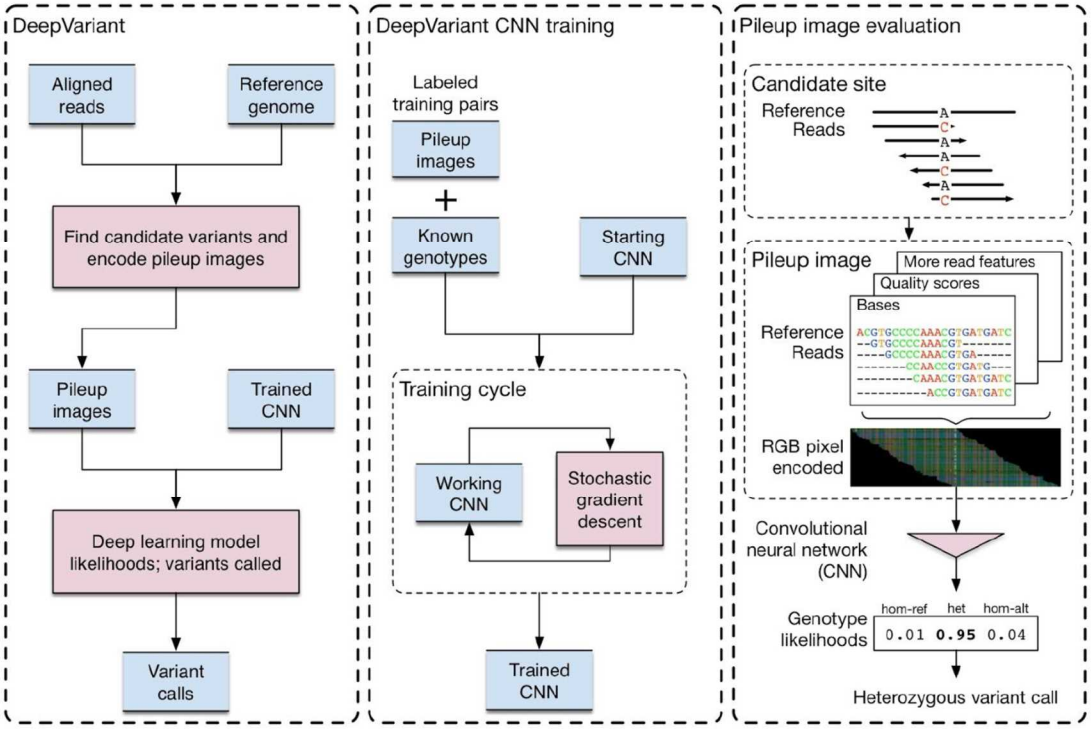
生成された画像を訓練されたニューラルネットワークに提供することにより、サンプルゲノムに特定の対立遺伝子が含まれている尤度を決定し、

決定された尤度を示す出力信号を提供する。

4. DeepVariant に関する論文

Verily Life Sciences 社は米国 Alphabet 傘下の研究組織であり、生命科学の研究を中

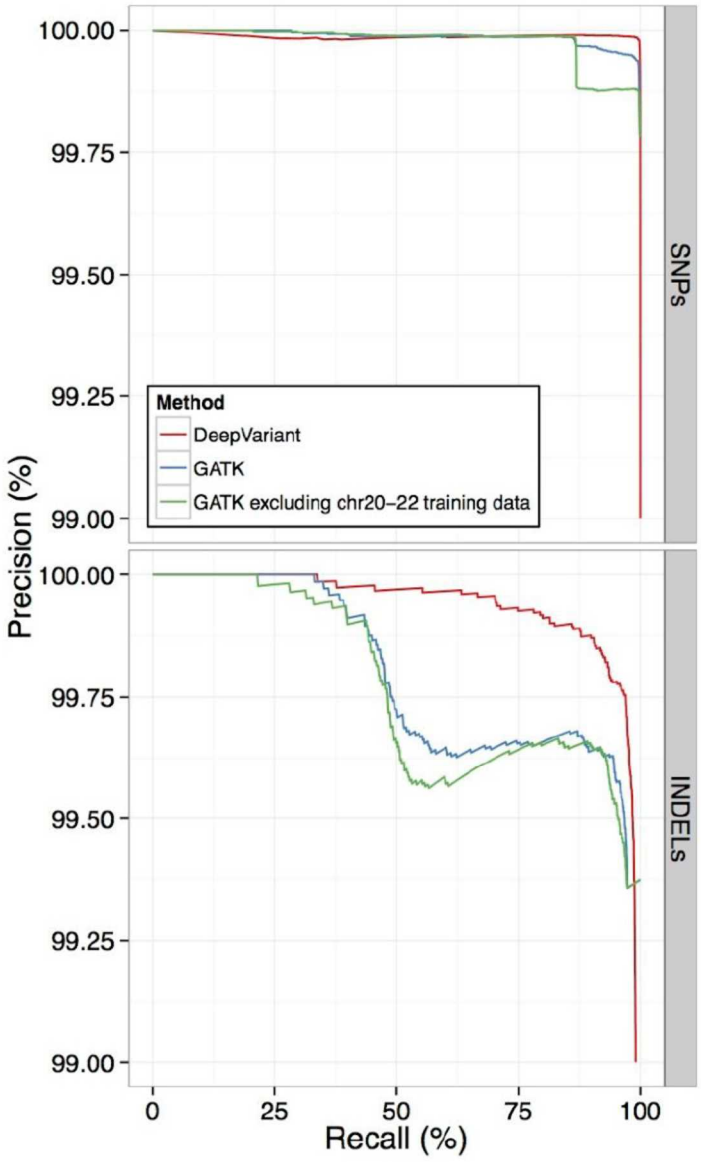
心に行っている。DeepVariant に関しては、Ryan Poplin 氏らが論文[4]を発表している。



上記図は DeepVariant のワークフローを示す。リードデータセットからパイルアップ画像が生成される。ヌクレオチド塩基及び品質スコアを含むパイルアップ画像はエンコードされ、エンコードされた画像が CNN に入力され、変異コールが行われる。

図の例では、このエンコードされた画像は、CNN に提供され、ホモ接合参照（hom-ref）、ヘテロ接合（het）、またはホモ接合代替（hom-alt）の 3 つの 2 倍体遺伝子型状態の遺伝子型尤度を計算する。この例では、最も可能性の高い遺伝子型の可能性が「het」であるため、ヘテロ接合バリアントコールが発行される。

[4] Ryan Poplin, Pi-Chuan Chang , David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, Mark A. DePristo "Creating a universal SNP and small indel variant caller with deep neural networks" https://doi.org/10.1101/092890



上記グラフは従来の GATK((Genome Analysis Toolkit))と比較した DeepVariant の精度を示すグラフである。DeepVariant（赤）と GATK（緑、青）の精度再現プロットは、Platinum Genomes プロジェクトの 2x101 Illumina HiSeq データを使用したボトルベンチマークサンプル NA12878 のゲノムである。従来の GATK と比較して SNP 及びインデル共に、DeepVariant の方が精度で上回っている。

DeepVariant は、Google Life Sciense(ライフ サイエンスのデータを管理、処理、変換するための一連のサービスとツール)のサービスの一つとして提供されている。

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「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 2.0」がある。