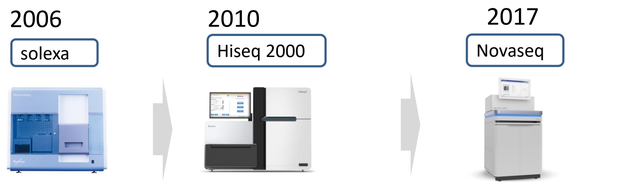
D8 がんを中心としたゲノムビッグデータ解析

○ 上田 宏生　先端科学技術研究センター 生命データサイエンス分野

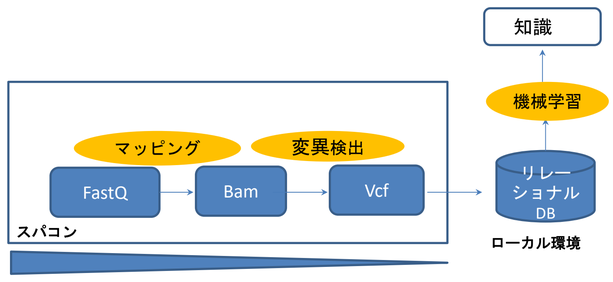
近年、次世代シーケンサ(NGS)の性能向上や実験技術の進展、ゲノム医学の発展により生物学で扱うデータ量が増大し、大量ではあるが、疎（スパース）なデータから如何に知識発見を行うかが重要な課題となっている。 この課題の解決には、ビッグデータ解析と機械学習を反復して行うことが必要であり、今日データサイエンスと呼ばれる、ビッグデータ関連技術や深層学習の技術を応用することが欠かせない。大規模ゲノムデータと深層学習をシームレスに連携させるソフトウェア基盤が充実しておらず研究の多くがデータ量や計算時間の制約を受ける結果になっている。そこで、オープンソース技術を利用したゲノム解析基盤を開発しており、(1)大規模がんゲノム解析、(2)ナノポア型シーケンサのデータ解析、および(3)質量分析器を用いたヒストン修飾の解析に応用を始めている。

クラウド環境で動作する次世代シーケンサ・質量分析器ソフトウェアを開発

１．ゲノム解析技術の変遷



２．現状のデータ解析の問題点



ビッグデータに対する問題点

※　ウォーターホールモデル　(後戻りできない、情報損失が起こる)
※　分散処理に向かない
※　同一環境にすべてのデータをUPLOADすることが難しい。

機械学習（深層学習）に対する問題点

※　大量データの機械学習を同一環境で実行しにくい。
※　GPUやFPGAを用いた処理と統合しにくい。

ビッグデータ・深層学習が必要とされる分野

大規模コホート、公開データとの統合解析
- VUS (Variant of Unknown Significance) の同定
- 新規がん遺伝子、変異、耐性品位の発見
- オミックス解析　　など

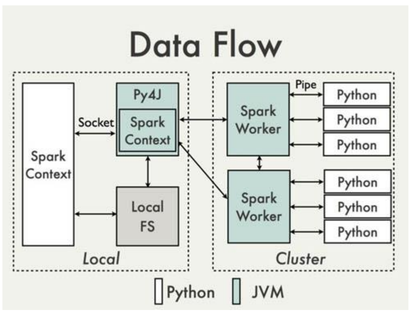
新規解析アルゴリズムへの応用
- がん変位検出
- nanopore シーケンサベースコール　など

３．オープンソース技術の必要性

クラウド（分散処理、仮想化）
AI (深層学習)
のオープンソースを使用することが
汎用性のにおいて現実的。

大手クラウドやオープンソース
を利用した環境で動作させることができる。

オープンソース技術
- 分散基盤：Hadoop/Spark,HDFS,YARN
- 深層学習　Tensorflow, KERAS
- データ構造：カラム型データ構造
  Apache　Parquet
  など



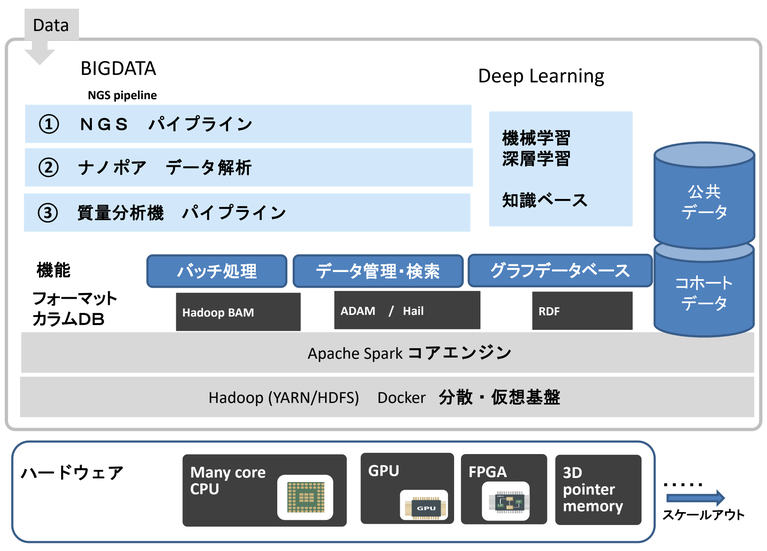
PySpark　の構成

機械学習のライブラリは
Python

ビッグデータのライブラリは
Scala/Java が充実しているが、
両者の統合が進んでる。

（出典：https://cwiki.apache.org/confluence/display/SPARK/PySpark+Internals）

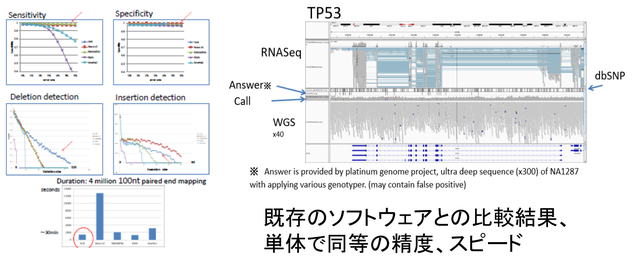
４．クラウド技術を活用した NGS・質量分析 解析パイプライン



５．開発ソフトウェア

マッピングソフト

Volt: アルゴリズム：Hash-DP
Java で作成されておりSparkに組み込まれている。



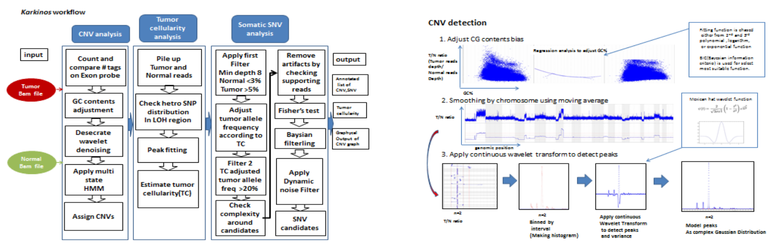
※ Answer is provided by platinum genome project, ultra deep sequence (x300) of NA1287 with applying various genotyper. (may contain false positive)

既存のソフトウェアとの比較結果、
単体で同等の精度、スピード

SparkMappy: (開発中)
既存ソフト　minimap2 のpySpark 実装

変異検出ソフト

Karkinos：
体細胞ー塩基変異、Indel　コピー数変異、
腫瘍率を同時検出。



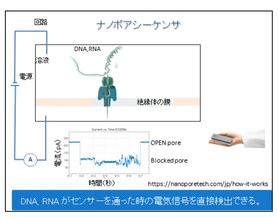
ー塩基変異はheuristic 処理で候補をフィルタリング
コピー数はHMMを使うことにより算出。
腫瘍率は理論値とのベクトルマッチングにより算出

本ソフトウェアを用いた解析例

・Aihara, K. et al.　Genetic and epigenetic stability of oligodendrogliomas at　recurrence.
Acta neuropathologica 7;5(1):18. (2017).
・Totoki, Y. et al. Trans-ancestry mutational landscape of hepatocellular carcinoma　genomes.
Nature genetics 46, 1267-1273, (2014).
・Kakiuchi, M. et al.　Recurrent gain-of-function mutations of RHOA in diffuse-type　gastric carcinoma.
Nature genetics 46, 583-587, (2014).

先端研ゲノムサイエンス分野との共同研究

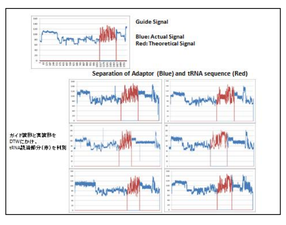
Nanopore シーケンサの解析プログラム



問題1：　Short Read から増加するデータ量

DNA／RNAの塩基修飾情報を
含めた解析が可能であるが、
そのためには、配列情報のBAMファイルに比較して、約100倍のraw データ
ファイル
をリファレンスに対してアラインメントする必要がある。

→　Spark　を用いた　分散処理基盤に統合、
新規のナノポアシグナル用のフォーマットを作成



問題2：　Adaptor Trimming

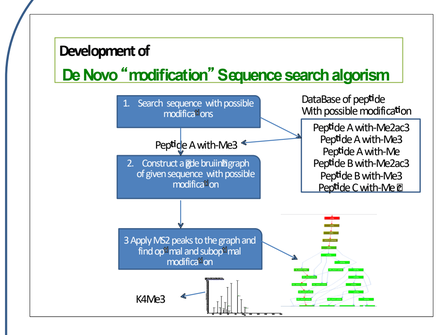
Basecall のミス、PolyA の長さの違い、
Quality のばらつきにより、
Adaptor Trimming が正確にできない。

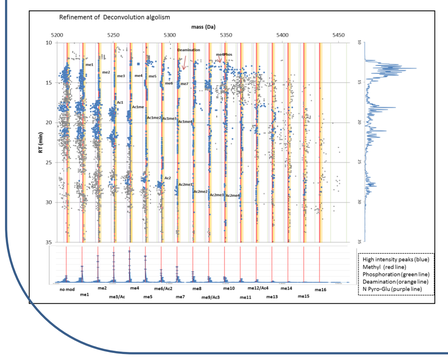
→　動的計画法で横方向の移動を許してシグナルを並べ替えるアルゴリズム
Dynamic Time Warping を用いて
Adaptorをトリミングし、リファレンスにアラインメントする
手法を開発している。

問題3：深層学習とビッグデータの反復的解析が必要
Spark 上の解析基盤と深層学習ライブラリ　KERAS を統合して
11TBのシグナルデータを分散して深層学習(CNN) で解析できる
ソフトウェアを構築中。修飾の有り無しで、noncoding –RNA
を分類することに成功している。

化学生命工学
鈴木研との共同研究

質量分析機　パイプライン





Denovo
ヒストン修飾解析
プログラムの開発

ヒストン middle down
MS/MS データから、
組成のマッチングを行い、
ヒストン修飾部位の解析を行う。
精度の要求される
デコンボリューション
デアイソトープ処理をSpark基盤に移行。
同位体分布に対して、
厳密に連続ウェーブレット変換を
行うことでO価、monoisotopic mass
に変換している。

アイソトープセンター
川村研との共同研究

６．ゲノム　大規模計算において留意する点

- **データの集約**
  Amazon/Google などのクラウドでは
  TCGA / ICGC , 1000 genome,
  nanopore-wgs-project など
  多くの公共データが集積されている。
  同様のデータの集積が必要。
- **ディスク領域の確保**
  テラ・ペタオーダーの
  ディスク領域が必要。
- **分散処理・仮想化**
  仮想化・コンテナ技術を用いて異なった環境でも同様に動作させることが必要。
  Hadoop/Spark などでオープンソースの分散処理・データベース検索の仕組みを作ることが課題。
- **多様なインスタンス**
  GPU/FPGA などのインスタンスを使用し処理を高速化する必要がある。
- **オープンソース**
  産業的に応用の進むオープンソース技術を活用する事が必要。