第８回アプリFS全体ミーティング
ミニアプリ化進捗報告

鈴木惣一朗、滝澤真一朗、丸山直也(理研AICS)

2013年8月6日

目次

ミニアプリ化作業進捗報告

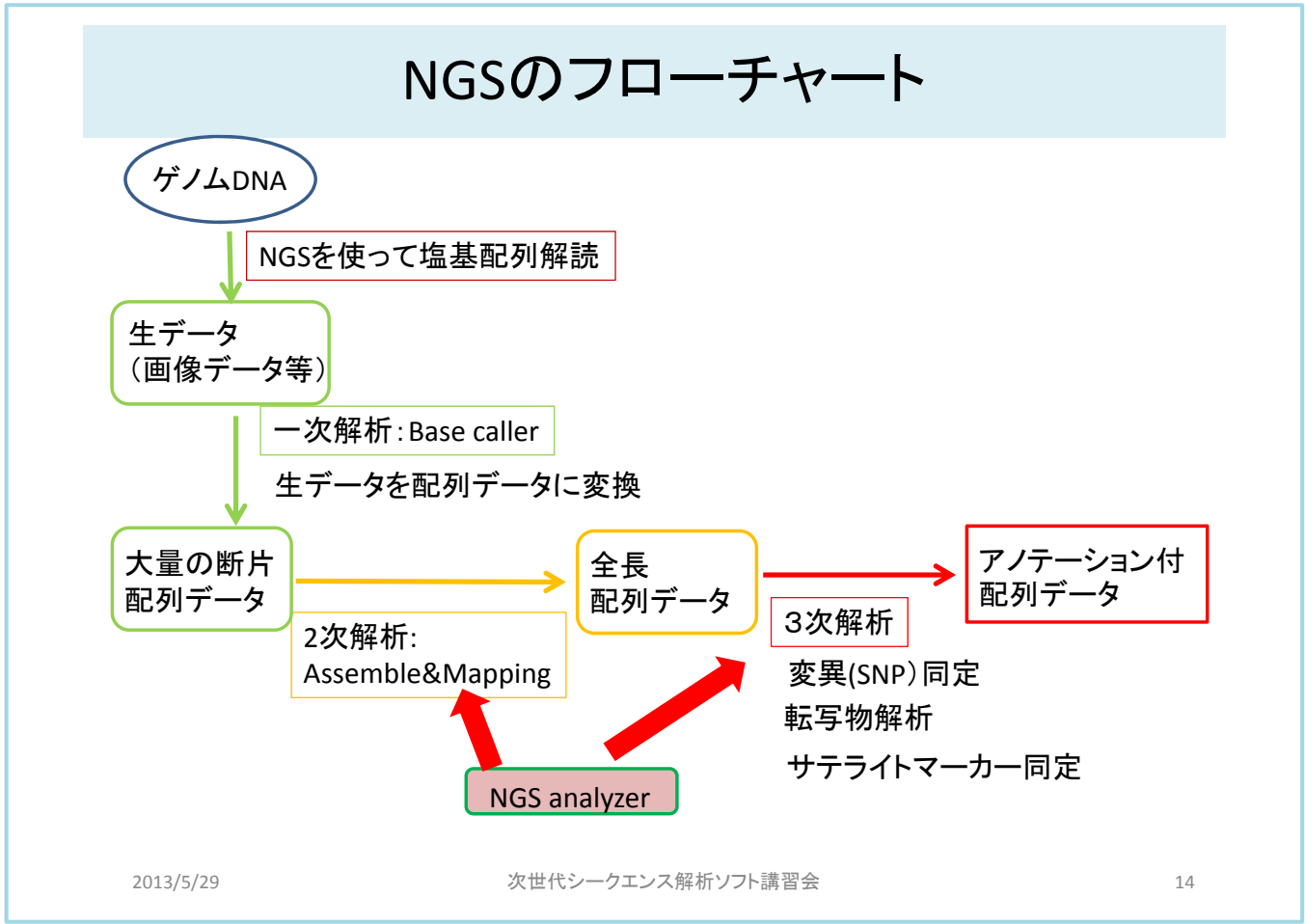
- システムFSチームへ引き渡し
  - FFVC → 東大FS
  - CCS-QCD → 筑波大FS
- ミニアプリ化作業中
  - NGS Analyzer (8月中完了予定)
  - 物質科学系アプリ
    - 密度汎関数法アプリ(STATE, CONQUEST, …)
    - ALPS/looper
    - …

NGS Analyzerのミニアプリ化

NGS Analyzer

次世代シークエンス解析プログラム

次世代シークエンサーの出力データを高速に解析し、ヒト個人間の遺伝的差異やがんゲノムの突然変異を高い精度で同定



NGS Analyzer プログラム構成

- ５つのシェルスクリプト
  1. リファレンスデータ作成
  2. シークエンスデータ（解析対象データ）分割
  3. アライメント処理
  4. アライメント結果からの重複除去
  5. 変異同定

  シークエンスデータ出力に合わせて、繰り返し実行
- 並列化手法はembarrassingly parallel
  - アライメント処理のみノード内スレッド並列
  - 他は、フラットMPI

1. リファレンスデータ作成

- アライメント処理のために、参照（リファレンス）ゲノムデータを変換
- 入力
  - 参照ゲノムデータ
  - NCBI配布の2.9GBのデータを利用予定
- 出力
  - アライメント処理ツール（BWA）で使用するフォーマット
  - 入力2.9GBの時、4.2GBの出力

2. シークエンスデータ分割

- 解析対象のシークエンスデータをアライメント処理の単位毎に分割
- 入力
  - 日本人初全ゲノムシークエンスデータ（圧縮時122GB）の一部を利用予定
  - 1ファイルセットあたり30～90GB
- 出力
  - 入力ファイルを25万シークエンス単位で分割したファイル
  - 出力サイズ合計は入力サイズと同程度

3. アライメント処理

- シークエンスデータをリファレンスデータにマッピング
- 入力
  - 1, 2の出力（リファレンスデータと解析データ群）
  - サイズは解析対象シークエンスデータ量に依存
- 出力
  - アライメント結果（sam形式）
  - 出力サイズと入力サイズの関連は調査中

4.アライメント結果からの重複除去

- フォーマット変換（sam => bam）し、PCR重複を除去
- 入力
  - 3の出力（アライメント結果, sam形式）
  - リファレンスデータのインデックス
    - 対象リファレンスデータの場合、22KB
- 出力
  - 重複除去結果（bam形式）
  - ファイルサイズは sam形式 > bam形式
  - 出力ファイル数はアライメント結果ファイル数に等しい

RCP（ポリメラーゼ連鎖反応）
＝DNAポリメラーゼ酵素を用いて連鎖反応的にDNAを増幅させる方法

5. 変異同定

- pileup（ゲノム座標毎に塩基を積み上げて出力し、解析に適した情報形式に変換）し、尤度推定に基づいて、遺伝的多様性の検出を行う
- 入力
  - 4の出力
  - リファレンスデータ（2.9GB）
- 出力
  - 4の出力ファイル1つにつき、4ファイル
    - pileup形式ファイル、解析サマリーファイル、変異（SNP、INDEL）ファイル
  - 出力サイズと入力サイズの関連は調査中

SNP（一塩基多型）＝一塩基だけ違っている変異
INDEL（挿入欠失）＝挿入または欠損による変異

ミニアプリ化方針案

- プログラムの特徴
  - ステップ1は1度のみ実行すれば十分
  - 全ステップにおいて、逐次プログラムをembarrassingly parallelに実行（ステップ3のみノード内スレッド並列、他はフラットMPI）
  - 1ファイルを処理するMPIプロセスは1つ
    - 入力ファイル数以上にはスケールしない
- ミニアプリ化方針案
  - ステップ3以降を逐次プログラムとして提供
  - 並列実行した時の計算時間、I/Oサイズ等を評価するための性能モデル式を別途提供
  - ステップ3の入力となるシーケンスデータセットを複数提供予定