Archivierungserfordernisse und Datenstrukturen in genetischen High Throughput-Analysen

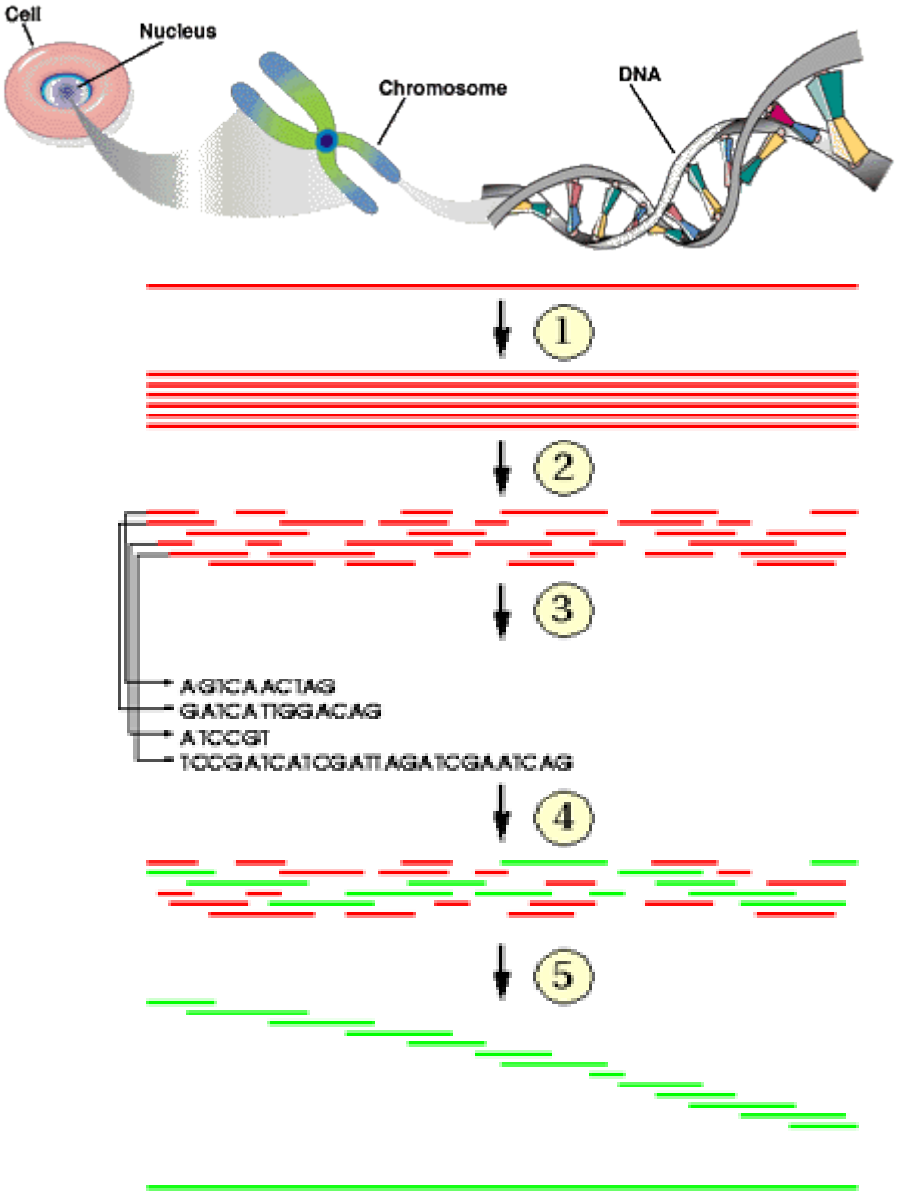
A. Herrmann
Arbeitsgruppe J.Hampe, UKSH Kiel

Einleitung

- Genome Sequenzierung
- Aktuelle Hochdurchsatz-Technologien
  - 454
  - SLX
  - SOLiD
- Datenaufkommen
- Datenmanagement
- Sequenzierzentrum Kiel

Genom Sequenzierung

- Sequenzierung des menschlichen Genoms: HGP 1990-2003
- Hochdurchsatz-Sequenziermethoden: Individual Genome



Sequenzierungsmethoden

1990-2003 → 2005-jetzt → In wenigen Jahren

Sequenzierung	Sanger	„Next Generation“	Einzehln Molekül
Datenmenge / Run	100 Reads	10M Reads	1G Reads
Probenanzahl / Run	einige	einige	hunderte
Sequenzqualität	analog	analog	digital
Kosten pro menschlichen Genom	1000000$	10000$	1000$

„Next-Generation“ Sequenzierungstechnologien

- Solexa
  - Firma Illumina, Gerät ca. 500T€
  - ~30-70 nt/read, 30GB/run, 5-10T€/Run, 3-10Tage/Run
- SOLiD
  - Firma Applied Biosystems, Gerät ca. 500T€
  - ~20-70 nt/read, 25GB/Run, 5-10T€/Run, 3-10Tage/Run
- 454
  - Firma 454 Life Sciences von Roche
  - ~400 nt/read, 400-500 MB/Run, ca. 2T€/Run, 1Tag/Run

Sequenzierungsprojekte

- 1000 Genome Projekt
  - Start 2008
  - Ein Individuum
    ~ 200Gb Sequenzen und ca. 300Gb Alignments
- Cancer Genome Projekt
- Größtes Europäisches Sequenzierzentrum:
  - Sanger Institut, Hinxton UK
    37 Solexa
    theoretisch 5000 Humangenome pro Jahr

Sequenzierungspipeline

Archivierung durch Sequenzierzentren

Gerätedaten
(Bilder)

Gerätehersteller
Software

Sequenzen
Format: FASTQ

Sekundäre
Analysen

Alignment
Format: SAM

Archivierung durch Anwender

Sicht der Anwender

- Sicherung der Sequenzen:
  - Format:
    - FASTQ
    - Proprietäre Formate der Gerätehersteller
  - Aufbewahrung
    10 Jahre / 30 Jahre - medizinische Daten
- Analysedaten

  Alignmentformat: SAM
- Sicherheitsaspekt

Datenformat

- Sequenzierungsdaten, FASTQ

<fastq> := <block>+
<block> := @<seqname>\n<seq>\n+[<seqname>]\n<qual>\n
<seqname> := [A-Za-z0-9_.:-]+
<seq> := [A-Za-z\n\.~]+
<qual> := [!-~\n]+

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

- Alignment, SAM

```
@HD VN:1.0
@SQ SN:chr20 LN:62435964
@RG ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
    AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
    NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M = 28701 -168 \
    ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
    MF:i:18 RG:Z:L2
```

Sequenzierzentrum Kiel

- Universitätsklinikum Schleswig-Holstein
  - 3 SOLiD Maschinen und Sanger-Sequenziergerät
  - 30 Runs/Gerät/Jahr
  - Archiv (komplettes Pipeline) ca. 500GB-1TB/Run
- Kooperation mit Rechenzentrum Kiel
  - Rechencluster mit 100 Knoten
  - Festplattenkapazität ca. 90TB RAID
- Lokale Speicherung auf USB-Platten
  - Vollständige Pipeline Archivierung
  - Notlösung ca. 50 1TB USB Platten

Sequenzierzentrum Jena

- Fritz Lipmann Institute
  - 2 Solexa Geräte
  - Durchsatz ca. 40 Runs/Gerät/Jahr
  - Archiv (Sequenzen + Alignment) ca. 50GB/Run
- Datenmanagement
  - 80 TB Raid System,
    zur Zeit 5 TB mit Archiven belegt
  - Parallele Speicherung auf Band
  - Keine Archivierung der Bilderdaten

Ausblick

- stetiger Wachstum der Sequenzdatenmengen
- Langzeitarchivierung
  - Wird nicht ausreichend beachtet
  - Keine eindeutige Regelung
  - Sicherstellung der 10 oder 30 Jahre Speicherung
  - Spätestens mit „Einzel-Molekül“ Sequenzierung ein akutes Problem