Read Mapping (ohne Varianten)

Vortrag im Rahmen der Projektgruppe

Marcel Bargull Janet Fiedler

PG583: Algorithmen zur Entdeckung krebsauslösender Genvarianten

Fakultät für Informatik, Technische Universität Dortmund

2. April 2014

Inhalt

1. Was ist Read Mapping?
2. Ablauf: Auswertung der Reads
3. Eingabeeigenschaften
4. Read Mapping: Indizierung
5. Read Mapper: Beispiele
6. Read Mapper: MAQ
7. Read Mapper: Bowtie
8. Read Mapper: Laufzeitvergleich
9. Ausgabe
10. PileUp
11. Variantenidentifizierung

Was ist Read Mapping?

Ausgabe der Sequenziermaschinen:

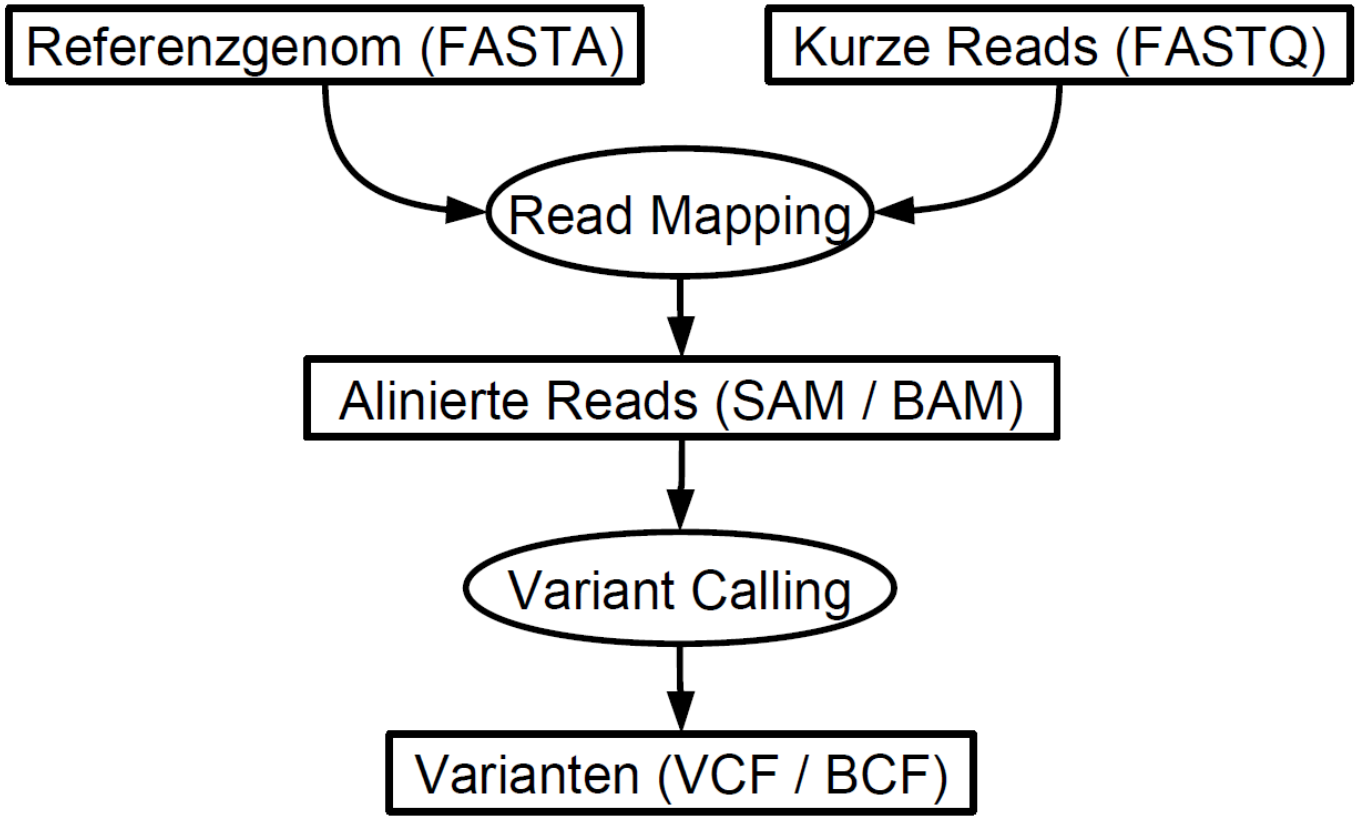
- Genom nicht als zusammenhängende Basensequenz
- Sondern: Reads
  - Millionen einzelner kleiner DNA-Schnipsel
- Keine Positionsangaben

Was ist Read Mapping?

Positionsbestimmung durch Read Mapping:

- Gegebenes Referenzgenom
- Suche den Reads ähnliche Teilsequenzen im Genom

Ablauf: Auswertung der Reads



Eingabeeigenschaften: Reads

Abweichung vom Referenzgenom

- einzelne Basen
  - Sequenzierfehler (Mismatch)
  - Echte SNV (Single Nucleotide Variation)
- Indels (Insertions / Deletions)
  - Sequenzierfehler
  - Echte Varianten

Eingabeeigenschaften

Zusätzliche Schwierigkeiten

- Mehrfachmatches
- Repeats im Genom

Erleichterungen

- Angabe zur Sequenzierqualität der Basen
- Paired-End Reads

Read Mapping: Indizierung

Indizierung für schnelles Mapping nötig, da

- Großes Genom
- Sehr viele Reads

Index von:

- Referenzgenom
- Reads

Verfahren:

- Burrows-Wheeler-Transformation
- Hashtabelle

Read Mapper: Beispiele

Zwei Beispiele für Read Mapper:

- MAQ (Hashtabellen auf Reads)
- Bowtie (BWT auf Referenzgenom)

MAQ

- Hashtabellen auf Reads
- Nur kurze Reads ≤ 63 bp (≤ 127 ab v. 0.7.0)
- Saat = Erste 28 bp vom "high quality end" (i.d.R. 5'-Ende)
- Lässt 2 Mismatches in der Saat zu
- Sucht pro Read ein Mapping ohne Gaps
- Benutzt Paired-End Reads
  - Korrektur falscher Mappings
  - Erkennung von Gaps
  - Unterstützung bei Repeats

MAQ

- Erstellt 6 Hashtabellen zu den Reads (28 bp Saat)
- Hashtabellen zugeordnet zu “noncontiguous seed templates”
- Indiziert nur Basen mit “1” im Template

Beispiel für 8 bp:

Ref.	acgttcga	
Read 1	acgttcga	0 Mismatches
Read 2	aggttcga	1 Mismatches
Read 3	tggttcga	2 Mismatches
Read 4	aggtacga	2 Mismatches
Template 1A	11110000	S1
Template 1B	00001111	S1 S2 S3
Template 2A	11000011	S1
Template 2B	00111100	S1 S2 S3
Template 3A	11001100	S1
Template 3B	00110011	S1 S2 S3 S4

MAQ

- Sucht für alle Referenz-28-Gramme nach Hashes
- Je 2 komplementäre Templates pro Durchlauf des Genoms
- Sechs Templates garantieren Finden von Saaten mit ≤ 2 Mismatches
- Bei gefundenem Hash
  - Mismatch Score = Summe der (Sequenzier-) Q-Werte der Mismatches (über ganzem Read)
- Wählt Mappingposition mit niedrigster Mismatch Score

MAQ

Weist Mappings Qualitätswerte (Q-Werte, phred-scaled) zu

- Hoher Q-Wert $\Leftrightarrow$ W'keit gering, dass Mapping falsch ist

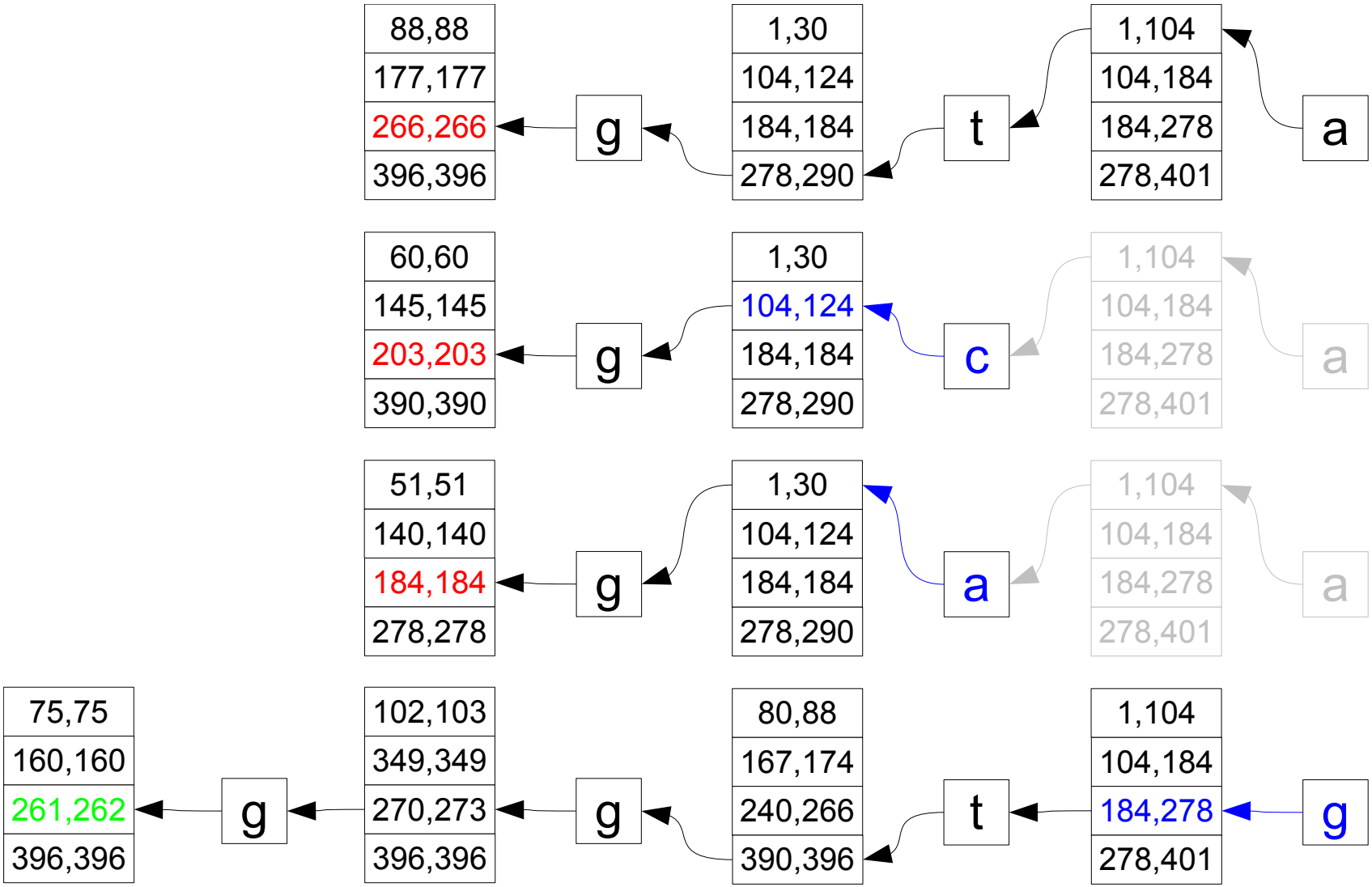
Bei Paired-Reads:

- 6 Tabellen je Richtung
- Speichert je 2 beste Positionen von Hits in Vorwärtsrichtung
- Markiere als Paar, falls Hit auf Rückwärtsstrang in erlaubtem Abstand zu den 2 Positionen
- Falls nur ein Read gemappt
  - Smith-Waterman gapped alignment

Bowtie

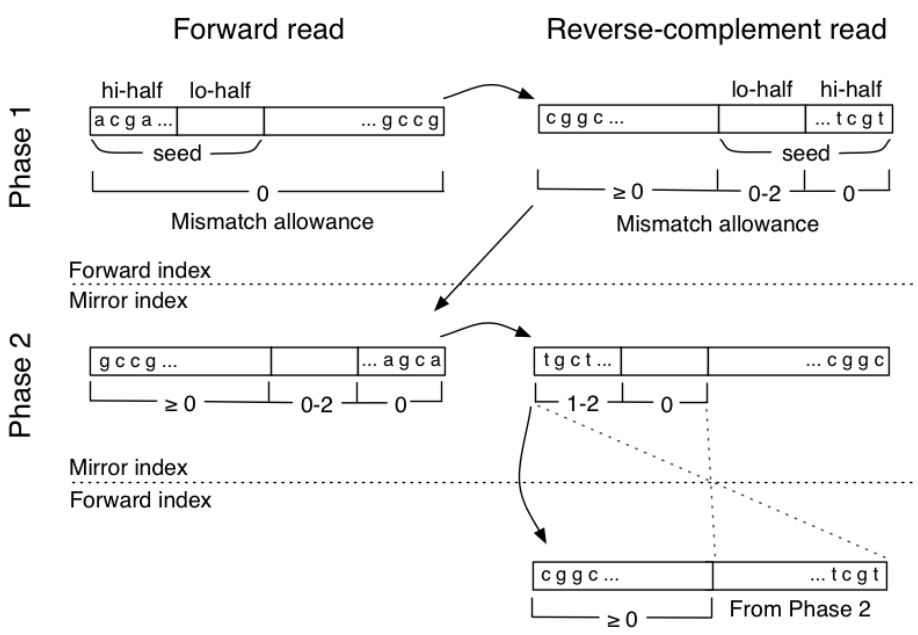
- Erstellt Index auf Referenzgenom mittels BWT
- Benutzt Backtracking bei Mismatches
  - Begrenzt Anzahl an Mismatches in der Saat (erste 28 Basen)
  - Wählt Substitution mit geringstem Basenqualitätswert Q
  - Begrenzt Summe von Q für Mismatches auf < 70
- Zusätzlich Index auf reversem Genom
  - Zur Vermeidung von Backtracking

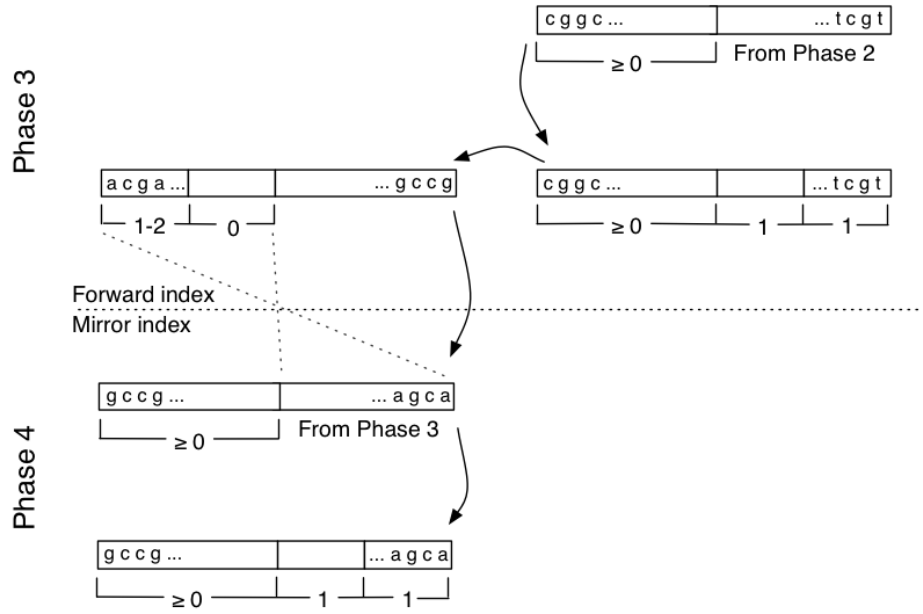
Bowtie – Backtracking



Bowtie – Algorithmus

- Wie MAQ bis zu 2 Mismatches in der Saat
- Backtracking vermeiden: 28bp-Saat in hi-half und lo-half teilen





Bowtie – Eigenschaften

+ Index kann vorberechnet werden
+ Deutlich schneller als MAQ
– Unterstützt keine Gaps
– Valide Alinierung für beide Paired-End Reads nötig
  - Auch hierdurch keine Gapunterstützung
- Nachfolger: Bowtie 2
  - Unterstützt Gaps und lokale Alinierung

Read Mapper: Laufzeitvergleich

	Single-end			Paired-end		
Program	**Time (s)**	**Conf (%)**	**Err (%)**	**Time (s)**	**Conf (%)**	**Err (%)**
Bowtie-32	1271	79.0	0.76	1391	85.7	0.57
BWA-32	823	80.6	0.30	1224	89.6	0.32
MAQ-32	19797	81.0	0.14	21589	87.2	0.07
SOAP2-32	256	78.6	1.16	1909	86.8	0.78
Bowtie-70	1726	86.3	0.20	1580	90.7	0.43
BWA-70	1599	90.7	0.12	1619	96.2	0.11
MAQ-70	17928	91.0	0.13	19046	94.6	0.05
SOAP2-70	317	90.3	0.39	708	94.5	0.34
bowtie-125	1966	88.0	0.07	1701	91.0	0.37
BWA-125	3021	93.0	0.05	3059	97.6	0.04
MAQ-125	17506	92.7	0.08	19388	96.3	0.02
SOAP2-125	555	91.5	0.17	1187	90.8	0.14

Abbildung: Ergebnisse

Ausgabe – Alinierung

```
Coor     12345678901234 5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1        TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004                    ATAGCT..............TCAGC
-r003                          ttagctTAGGC
-r001/2                                     CAGCGGCAT
```

Ausgabe – SAM-Datei

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT
```

http://samtools.github.io/hts-specs/SAMv1.pdf

Pileup

- Mappings aus Sicht der Referenz
- Pro Position / Base im Referenzgenom Auflistung der Überdeckenden Reads → Coverage
- Pileup-Format:

```
seq1 272 T 24 ,.$....,,.,...,,,,.,..^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23 ,.....,.,.,...,,,.,..A <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23 ,.$....,,.,...,,,,.,...    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23 ,$....,,.,...,,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22 ...T,,.,.,...,,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22 ....,,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq2 156 A 11 .$.....+2AG.+2AG.+2AGGG    <975;:<<<<<
seq3 200 A 20 ,,,,,,..,.-4CACC.-4CACC.....,,,.^~. ==<<<<<<<<<<<:<:<;2<<
```

http://samtools.sourceforge.net/pileup.shtml

Variantenidentifizierung

Variante erkennbar, falls

- Genügend Abdeckung / Coverage
- Mappings mit hoher Qualität (z.B. > 60) vorhanden
- Wenige Abweichungen von Referenz in Nachbarschaft
- Genügend Mappings mit geleicher Ausprägung

Variantenidentifizierung

Beispiel für weitere Probleme durch Alinierung:

- Fehlerhafte Mappings rund um Indels:

```
coor     12345678901234     5678901234567890123456
ref      aggttttataaaac----aattaagtctacagagcaacta
sample   aggttttataaaacAAATaattaagtctacagagcaacta
read1    aggttttataaaac****aaAtaa
read2     ggttttataaaac****aaAtaaTt
read3         ttataaaacAAATaattaagtctaca
read4             CaaaT****aattaagtctacagagcaac
read5               aaT****aattaagtctacagagcaact
read6                 T****aattaagtctacagagcaacta
```

http://samtools.sourceforge.net/mpileup.shtml

Read Mapping mit Varianten

Read Mapper können Fehler von Varianten schlecht unterscheiden

- Varianten evtl. als Fehler verworfen
- Verzerrung in Richtung Referenzgenom
- Read Mapping mit Variantenunterstützung nötig!