Methoden vor und Nachteile

1. Prüfung immer gut 2. Prüfung schwer

Viel wert auf Sprache gelegt [Sprache lernen]

Wie arbeite ich mit Biologen zusammen

Wie werden Proben im Labor bearbeitet

Ansatz schon vor dem Experiment

Wir bewegen uns im Petabyte bereich

Orga: Mail schreiben, nach 3 Tagen nochmal

Praktikum: Wenn wichtig und es verpasst sollte, gibt es Lösung

Prüfung: Mündlich.

Rollenspiel in der Präs.

Kritische Fragen von der Pipeline zu erfragen

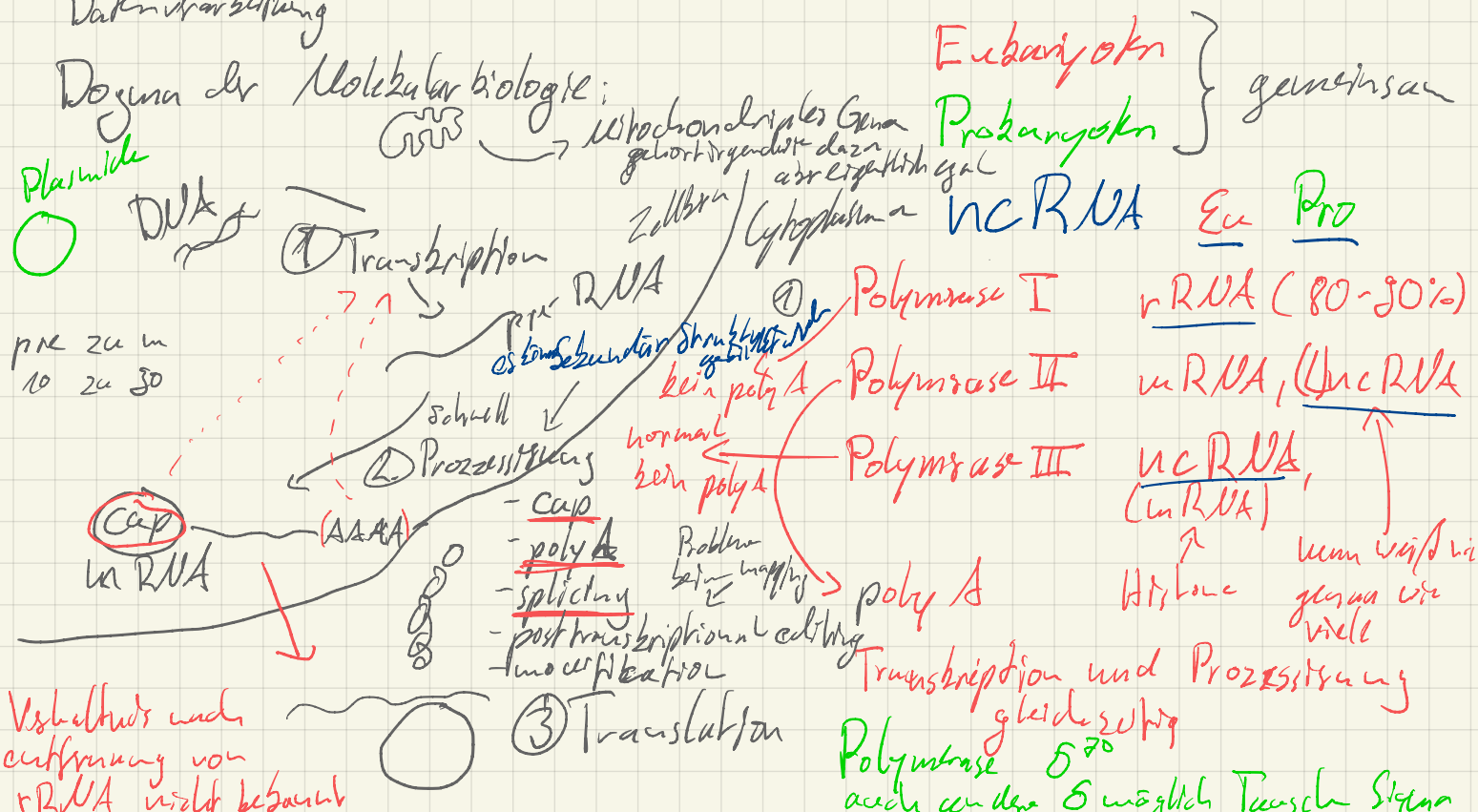
Genom interessant → genom aus Zelle → DNA muss seq werden →

DNA zerschnippelt → wieder zusammen Bauen ⇒ Assemblierung

Bei RNA große Frage → Mapping

Datenverarbeitung

Dogma der Molekularbiologie:

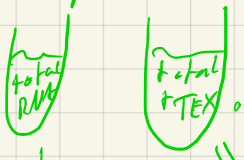


Der großteil von nicht mRNA codierende Bereiche, codieren ncRNA

Prokaryoten haben kein capping, aber Phosphat am 5' ende "dual Seq"

Genom = gesamtheit aller Gene + Rest.

Transkriptom = Alles was RNA ist [Total RNA] inclusive Ribosome.



oder TAB fressen alles was nur ein Ph haben

↓ pre-RNA ≪ RNA

⇒ pre-RNA

freie RNA ⊆ total RNA kann aber einzelnd gezogen werden

↓ muss später gemappt werden

Vorbereitung zur Sequenzierung:

Es geht jetzt um Transkriptom

dRNA-Seq (differential) → unterschied zwischen zwei Proben / Kohorte
Bsp. infiziert vs. nicht infiziert / auch zeitlicher verlauf

- total RNA —Seq→ 90-95% rRNA (schlecht)
- poly A - Seq (nur mRNA aus Pol II, günstig, effektiv)
- trennung über sonden (binden an ziel RNA) [Ribozero → 5% - 60% rRNA] ↓ schlecht kleine RNA ≤ 50nt
  selbst herstellen 3-4 wochen für key-Organismen
- small RNA → alles kleine micro RNA, piRNA

[CLIP-Seq: alles was an target Protein bindet]

→ wenn gar kein Plan.

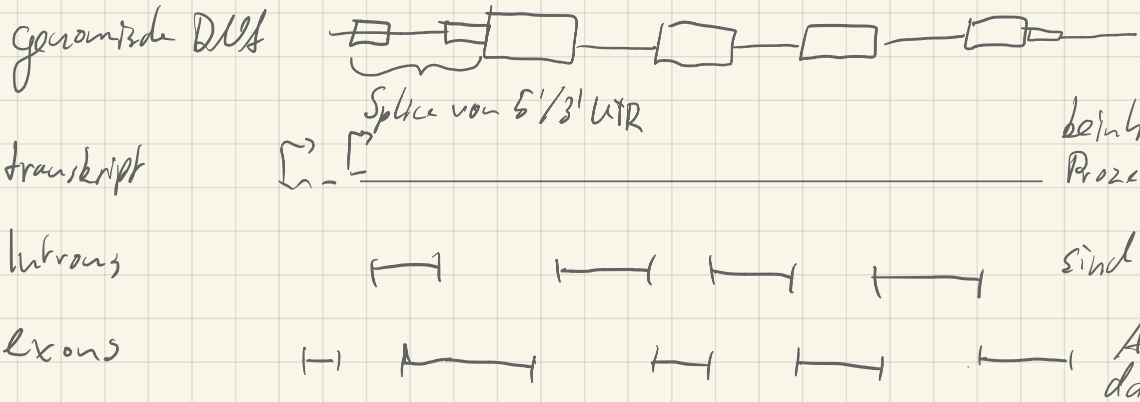
Verschieden warum aber irgendwann immer weg

total RNA, warum? kostengünstiger geworden als Ribozero zuerst. (Bei Bakterien, Archaeen, kleine Einzeller)

Man braucht fast immer 1µg pro Methode.

Wir sollten Protokolle kennen um zu empfehlen

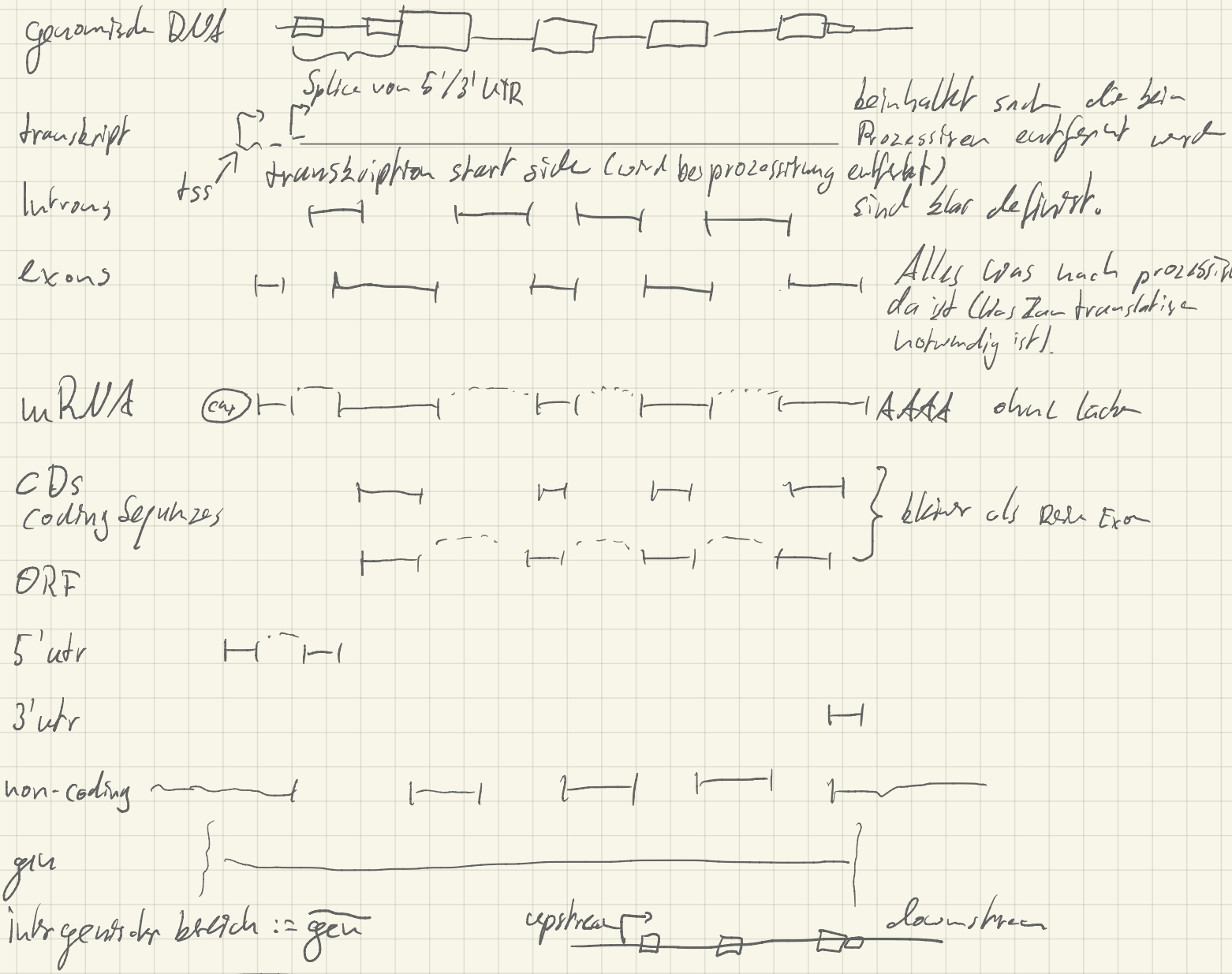
Gen := [führende Genetiker sind sich nicht sicher] alles damit es funktioniert



transkript: beinhaltet sachen die beim Prozessieren entfernt wird

introns: sind klar definiert.

exons: Alles was nach prozessieren da ist (was zur translation notwendig ist).



Ähnlichkeiten:

Mensch Mensch	99,9%	} eigentlich alles
Mensch Schimpanse	98,4%	
Mensch Maus	90 %	
Mensch Fliege	60 %	
Mensch Banane	50 %	→ nur auf beiden (Prok) die beide Hals
Mensch Hefe	30 %	

Output der Sequenzierung: reads → (zusammenbauen) contigs → scaffold → chromosom

Alles sehr durcheinander. gibt auch zwischen stufen.

keine genauen values

→ kein Genom ist fertig Assembliert!

Es gibt etwa 100 (nicht Bakterielle) Genome im chromosomalen Bereich. Bsp.: Bos Taurus (Kuh/Rind)

Die selben reads in verschiedenen Tools ergeben verschiedene Assemblies mit etwa 60% ld (Problem sind repetitive regionen)

Genome die man downloaded sind oft nur durch einen Assembler gelaufen (kritisch sehen)

Homologe gene: = Gene des selben Ursprungs (Übergriff) hsa mmu hom. sapiens mus musculus

Paralogegene: = kopien von genen im gleichen Genom, müssen nicht gleiche funktion haben, aber gleicher Ursprung

Orthologegene: = haben den selben vorfahren, in 2 spezies, machen das selbe

Pseudogen: paralogegene die oft nicht exprimiert werden.
In Datenbanken sind manche falsch annotiert, früher so gedacht und dann das gegenteil entdeckt und nicht verbessert

Promotor: Region des Gens meist up- (oder downstream).

Terminator: Bei Bakterien 2 terminatoren rho abhängig oder nicht. ↑ hairpin
Bei Eukaryoten nicht verstanden.
Theorie: Polymerase liest einfach weiter bis irgendwann abfällt.

Splicing:

Alternatives Splicing:

Intron retention: (Intron zurückhaltung)

Enhancer

Silencer

Est: Expressed Sequenz tag

EST: Einzelne gefundene Expremierte Sequenzen

~Viren: 1 kilo Basenpaar

Human Genom Projekt: 3,2 mrd BP ~Bakteriengenom: 3 Mega Basenpaar

Unterscheidung zwischen De novo und Referenz sequenzierung

Ziele: Genom Sequenzierung
Annotation
Geeignete Speicherung [Fasta]
Tools entwickeln, die mit Datensatz umgehen können
Ethik → will man wissen woran man stirbt
Genregulation verstehen
Funktion jedes Gens verstehen (prot. codierende 50% ... Fasta
Gen Interaktion (auf DNA und nach Expremierung)
Chromosomenübergreifende Interaktion
Konservierung des Gene im Genom
SNPs verstehen (Ähnlichkeit)

gibt nur 1 ref Genom
mittlerweile mehrere aber wieder in einer Datenbank

Pangenome keine sinnvolle Datenstruktur

und gff
Chr. +/- start stop ...

genome feature file
storing gebunden an
Fasta

Beim Sequenzieren ist jede 100000ste Base falsch

Fehlerrate 10^{-5}

Mensch hat etwa 22000 gene (Proteincodierende gene)

(gibt aber mehr, da nicht-codierende auch dabei)

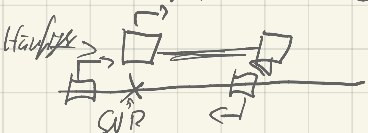
Korrelation zwischen Komplexität der Organismen und menge nicht-codierende regionen es gibt.

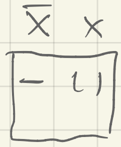
20000 behandeln nicht Isoforme (alternative formen des Gens)

von 20000 Genen konnte man 1500 zu Krankheiten zuordnen.

Pro Zeiteinheit etwa 6000 Gene exprimiert. Viele Gene sind für extreme cases

Forensik





SNPs die nur in wenige % vorhanden

Nutzen: Krankheitsdiagnose, Gentherapie, Frühdiagnose, Medikamente an DNA anpassen, Forensik [konzentration auf SNPs], Versicherungen, Arbeitgeber, Militär,

Genom: Protein-codierend: 2,7% (exons), Proteincodierende introns: 25,9%, Lines 20,4%, Sines 13,1% (... — vielleicht ursprung in viren), Retrotransposons (zu viren zuzuordnen) 8,3%, Zwangs

man weiß so gut wie nix über nicht proteine

Von prokine Transkriptionsfaktor 12

Trans fressen 8,8

Nukleinsäure bindende 8,5
prokine

Rezeptor Promotor 6,5

Transportprot. 6,4

keine Angabe 23,6%

Sequenzierung
- Maxam-Gilbert
- Sanger (Kette-Abbruch)
→ vorbereiten.

---

ncbi full genome Download googlen

ftp.ncbi.nih.gov/genomes/

fna fasta nucleotide

faa fasta amino acid

Genomnummer m39

Annotationsnummer normalerweise im Readme
→ gff file lesen können

ucsc genome browser
hgdownload.soe.ucsc.edu/download.html

masked → repeats durch definition ersetzen (hard mask)
→ Uppercase durch lowercase ersetzen (soft mask)

Verschiedene tools oben genannt nutzen RepeatMasker
centromere und telomere region meist simple repeats
Sines Lines prozentual am meisten
k-mere based analysis

animal genome size database
genomesize.com

C value gewicht der nucleotide V in pikogramm

Sequenziermethoden

Illumina: [Illumina Sequencing by Synthesis] Fehler 10^{-3}

Firma	Methode	Device
Illumina		HighSeq / MiSeq / NovaSeq (Unterschiede in Throughput (linear))

PyroSeq

Kommentare: Länge 200-300 Nucleotide — four five four Reads

Nicht gleich komplettes Gen, zerstückelte Teile

Adaptoren verändern ist wichtig, letzter Teil wirklich relevant

Clustering Oligos ⇒ Flowcell

Oligos auf Flowcells sind fest und exakt complementär zu letzten Teil an Adaptor

Nur 5'enden auf Flowcell

Da Licht nicht stark genug → Spots (Cluster von gleichen Seq)

Oligos in Spots nah genug damit Brücken bauen

Weiter (Waschritt) um nur 1 Bindung zu erlauben

Sequenzierung durch Synthese

Cycle = 1 Waschrunde + Enzymatische

Base calling = Lichtsignal übersetzung

2x150 Hilft beim Assemblieren → Paired end Sequenzierung

Barcodes in Adaptoren hilft beim aufteilen

Pacbio Sequencing — Fehler 10^{-4}

Pacific bioscience — SMRT Sequencing
Single Molecule Real Time Sequencing

relativ viele Fehler → Circularisieren und 6 mal lesen.

Multiplexing → Mehrere Sequenzen gleichzeitig analog Barcoding
Grundlage für Single cell Sequencing

Ont minion Sequencing

Oxford nanopore technologies   Device MinION
PromethION (8x MinION)

5 Gleichzeitig in Pore (200 Poren pro chip)

5-Mer analysiert 10%.

Hohe Fehlerrate weil ID etc schlecht, oder zu viele Modifikationen.

Viele Modifikationen kennen ist aktuelle Forschung

Vorteile: RNA und Proteine direkt Sequenzierung, realtime, keine Assemblierungsmethode vorhanden, direkt im Feld, günstig

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000	ONT
Instrument Cost*	$128 K	$80 K**	$695 K	$256 K	$654 K	1000 $ (nach 48h tot; Waschen gibt es 24h)
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb	22 GB DNA, 15 GB RNA
Sequencing cost per Gb*	$502	$1000 (318 chip)	$2000	$148	$41	75 $ (~50 €)
Run Time	27 hours***	2 hours	2 hours	10 days	11 days	live (Poren werden müde)
Reported Accuracy	Mostly > ~~Q30~~ Q60	Mostly Q20	~~[illegible]~~ über Q20	Mostly > Q30	Mostly > ~~Q30~~ Q60	~Q20 DNA 2%, >Q10 RNA 15%
~~Observed Raw Error Rate~~	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %	
Read length	~~up to~~ 2x 150 bases, auch 2x500	~200 bases	Average 1500 bases**** (C1 chemistry) mittlerweile 10000	up to 150 bases	~~up to~~ 2x 150 bases, auch 2x500	keine Beschränkung, Pore und länge 450kB
Paired reads	Yes	Yes	No	Yes	Yes	—
Insert size	up to 700 bases	up to 250 bases -350	up to 10 kb	up to 700 bases	up to 700 bases	
Typical DNA requirements	50-1000 ng bis 10 µg	100-1000 ng	~1 µg	50-1000 ng	50-1000 ng bis zu 10µg	1 µg +

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated.
** System price including PGM, server, OneTouch and OneTouch ES.
*** Includes two hours of cluster generation.
**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter.

1 Mb Zellen => 1 µg

Coverage: reads auf Genom und schauen wie viel ist abgedeckt (Wikipedia lesen)
60 Fach bei Illumina ist schon
Bei Ref reicht schon 10-20 Fach
Bei langen reads braucht man nicht so hohe Coverage

und NovaSeq (Tabelle vervollständigen)   iontorrent nachschauen

1 Sequenziermethode in Prüfung erkennen

Warum Fehler

Kosten berechnung: Für Menschl. genom

3200 Mb —Transkriptomanalyse 2,34%→ 75 Mb) 80 X (Coverage)
6000 Mb

Kostenberechnung: Für Menschl. genom

3200 Mb —Transkriptom analyse→ 2,34% (Pilz 80%) → 75 Mb) 80 x (Coverage) 60000 Mb) 2·100 bp → 30 Mio reads

Abfrage:
Genomgröße
Wie viel davon
Coverage
Read output

=> OUT für 3 TK

15 Gb ~ 500 € Flowcell 1x
250 € Libprep 3x

Illumina
1x 3000 € Flowcell (36 Gb)
6x ~200 € Libprep

Fehler Ignorieren Rechnung in Prüfung

Assemblierungsverfahren:

Wie gut ist das Tool? Welche Probleme hat dieses Tool (immer/oft welche kann das Tool lösen)? Laufzeiten der Methoden (wenn Alignments) beim GREEDY-Algorithmus (bei 5 Reads 10 Alignments)

1.) pairwise Alignments
2.) max Score nehmen A|B
3.) Merge A|B → AB
4.) repeat

Kommt klar mit ①, ④ klar

Algorithmus kommt mit wenige an Reads und längen nicht klar

Probleme der Assemblierung

① Sequenzierfehler
② Repeats
③ Orientierung
④ Lücken (Bug oder Feature)
⑤ Chimera (Zusammenschluss von Sequenzen die nichts miteinander zu tun haben, Barcode).
⑥ Zeit
⑦ Speicher

Shortest Superstring Problem mit DBA

Geg: Set von Strings — von 0 0 0 / 1 1 1

Ges: kürzester Superstring

Ansatz: Euler Pfad, Hamiltonian Pfad

Assemblierung:

~ das Zusammenführen von Nukleotidsequenzen zu längeren Sequenzfragmenten (Contigs oder Sequenz alignments)

Reads → contigs → (super contig) → scaffold → (super scaffold) → chromosome

Nachteile fehlt zusammenschluss. Mehrere Contigs, aber wissen über diese

Greedy

1. Paarweise Alignment
2. Suche max Score von Alignment
3. Merge max Score-Alignment
4. Wiederhole

Nachteile:

- kann mit Repeats nicht umgehen
- Sehr hohe Laufzeit, viel Speicher
- Ordnerstörung

Gütekriterien für Genom-Assemblierung:

N50-Wert / N90 ~ Länge des kürzesten Contigs, das in Summe mit allen längeren Contigs die Hälfte des Genomassembly abdeckt.

→ Contigs der Länge nach sortiert, dann von lang nach kurz die Contig Länge aufsummiere bis 50% der Basen des Gesamtassemblies erreicht sind → Nimm die Länge des letzten addierten Contigs

L50-Wert / L90 ~ Wie N50 aber anstatt der Länge des zuletzt addierten Contigs betrachte man anzahl der Summierten Contigs.

Bsp.: 10, 9, 8, 7, 6, 5, 4, 3, 2 Σ54

10+9+8 = 27 ⇒ N50 = 8 L50 = 3

Für Bakterien sind die Werte gut (Nur kein ein Genom (Chromosom))
Und sagen nichts über Richtigkeit aus.

NG50 ~ wie N50, aber nicht im Bezug auf die im Assembly vorhandenen Nucleotide, sondern der echten Genomgröße.

C-Value: 1pg ≈ 978 Mb

$$\text{Coverage} = \frac{n \cdot L}{G}$$

n = # Reads
L = durchschnittliche Readlänge
G = Genomlänge

bis neuen und. 80

Strategien: Greedy, Hamilton, Euler, Overlapping Consensus (Graph, De Bruijn
↳ SSAKE ↳ Newbler ↳ (Velvet, Trinity, (Spades), (Soap denovo)

(↳ Newbler; ↳ Velvet, Spades, Soap denovo)

Hamilton:

Shortest-Superstring-Problem: ~ SS von zwei Strings X und Y ist ein kürzester String Z von dem sowohl X als auch Y substrings sind.

→ einfach erweiterbar auf n strings.

- Das SSP ist NP-schwer (NP-vollständig)
- Es gibt nicht immer eine eindeutige Lösung für das SSP
- Kann in das Traveling-Sales-Person-problem umgewandelt werden (NP-Vollständig). ↳ aber effiziente Heuristiken.

→ Reduzierung von SSP zum TSP-Hamiltonpfad.

Hamilton-Pfad: ~ ist ein Geschlossener Kreis in einem Graphen, der jeden Knoten genau einmal enthält

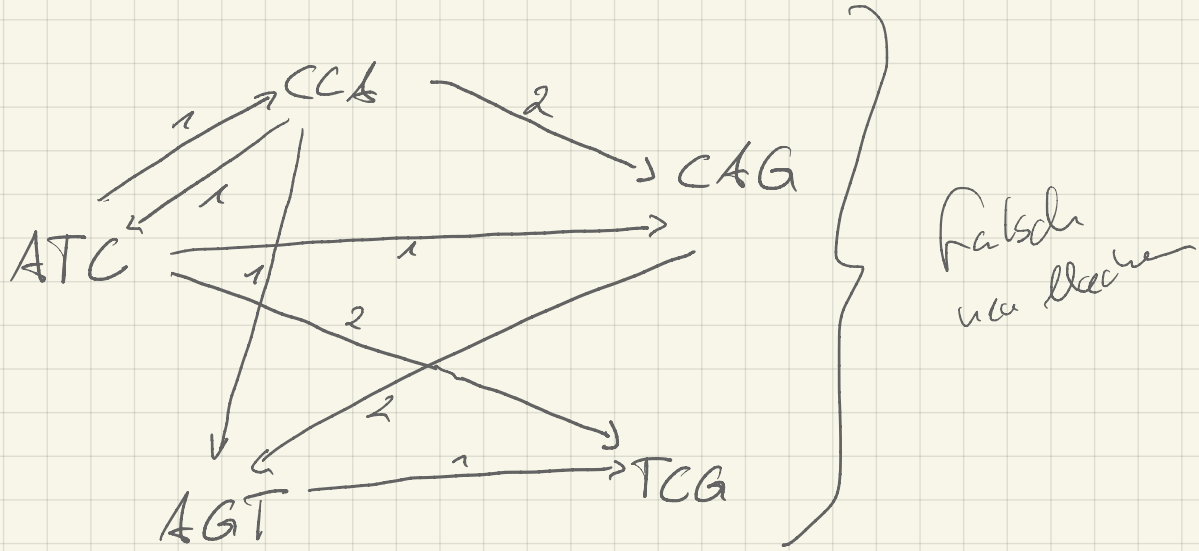
Wir definieren einen Overlap (s_i, s_j) als die Länge des längsten Präfixes von s_j das auch suffix von s_i ist. ⇒ Reihenfolge ist entscheidend.

Algorithmus - Idee:

- jeder Read ist ein Knoten im neuen Graph
- jedes Kantengewicht zwischen zwei beliebigen Knoten ist Overlap der beiden Knoten
  → Graph ist gerichtet!
- finde den schwersten Hamilton-Pfad.

Große Probleme: ① Repeats
② Lücke

$S = \{ATC, CCA, CAG, TCG, AGT\}$



falsch rum Kanten

1. ATC (2) → TCG
   CCA (2) → CAG
   AGT ⇒ ges 4

ATC
CCA 1
CAG 2
TGT 0
TCG 1 ⇒ 4

Euler-Algorithmus

Euler-Kreisproblem: Gegeben ein Graph G, finde einen Kreis, der jede Kante von G genau einmal enthält.

Euler-Pfad: Wie Kreis nur nicht an selben Endknot.

Algorithmusidee

① Zerlege alle Reads in ein Fragmentspektrum F der Länge k

② Erstelle einen Graphen G_{DB} (gerichtet)

$G_{DB}(S) = (V, E, Le)$

V = alle Präfix/Suffix der Länge k-1 aus F

$E = \{(a,b) \mid \exists f_i \text{ in } F : \langle a,b \rangle = f_i\} \quad a,b \in V$

beide Knoten sind Teil eines Kmers (Fragments)

$Le = E \to \Sigma$ Labeling $(a,b) = b_n$

③ Finde Eulerpfad

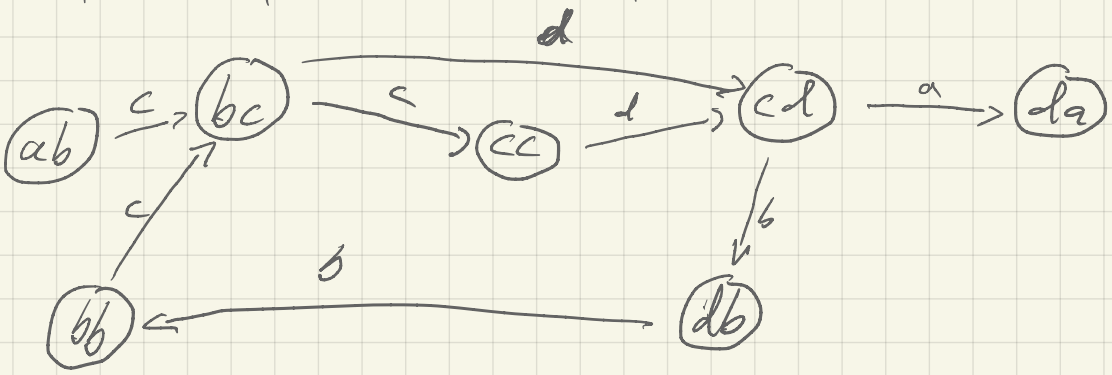
Vorteile

- keine Berechnung paarweise Alignments oder explizit Overlap zwischen Reads
- Es gibt schon sehr effiziente Algorithmen einen Eulerpfad zu finden $O(n)$ (Alg. von Hierholzer) n = # Reads
- Ein Eulergraph lässt sich leicht in Kanten disjunkte Kreise zerlegen.

Nachteile:

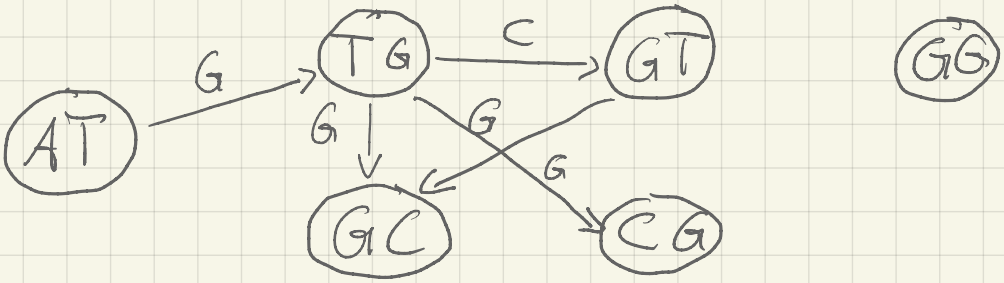
- (Assembly) nicht robust gegen Sequenzierfehler
- # Eulerpfade $\sim e^{\#\text{Repeats}}$
- Zerteilung unserer Reads in k-mere → Informationsverlust?

$S = \{abc, bbc, bcc, bcd, ccd, cda, cdb, dbb\}$



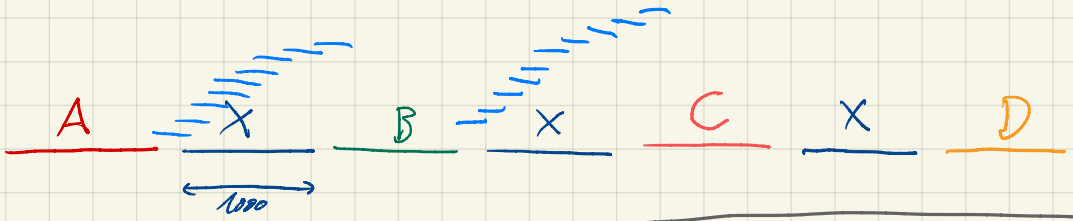
$S = \{ATG, TGC, GTG, GGC, GCA, GCG, CGT, TGG\}$

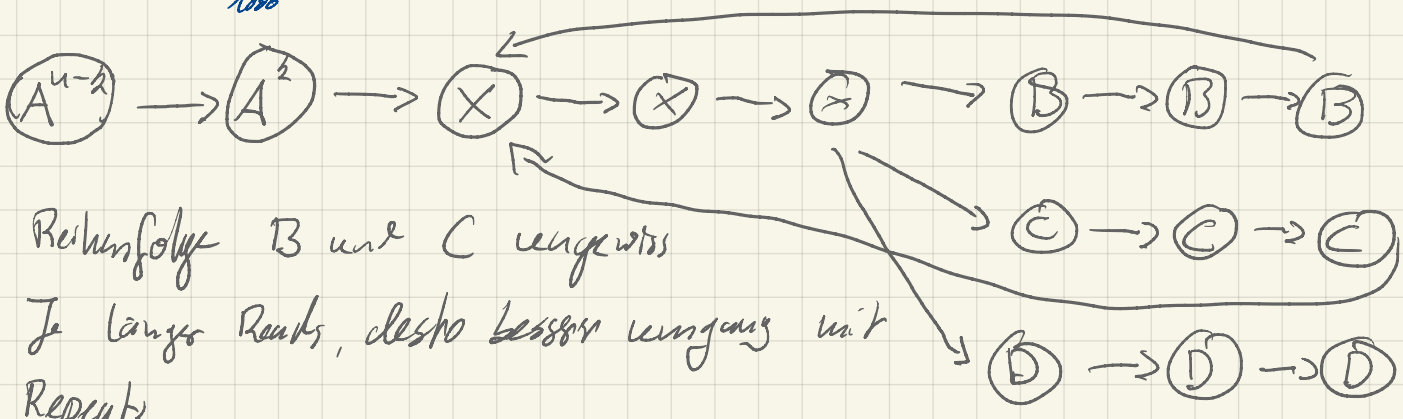
$k = 2$



Paired end für infos über Lücke

Euler und Hamilton Lösungen
↳ Hauptsächlich für Illumina





Reihenfolge B und C vertauscht

Je länger Reads, desto besser Umgang mit Repeats

Euler/Hamilton kann nicht mit Repeats umgehen

Menschliches Genom und ähnliche Eukaryoten hatte keine rRNAs drin (Nor 2013), da Assembler nicht mit Repeats klar kommen. Hefe meist 1, aber nicht alle. Assembler kommt auch nicht mit Telo-, Centromer klar

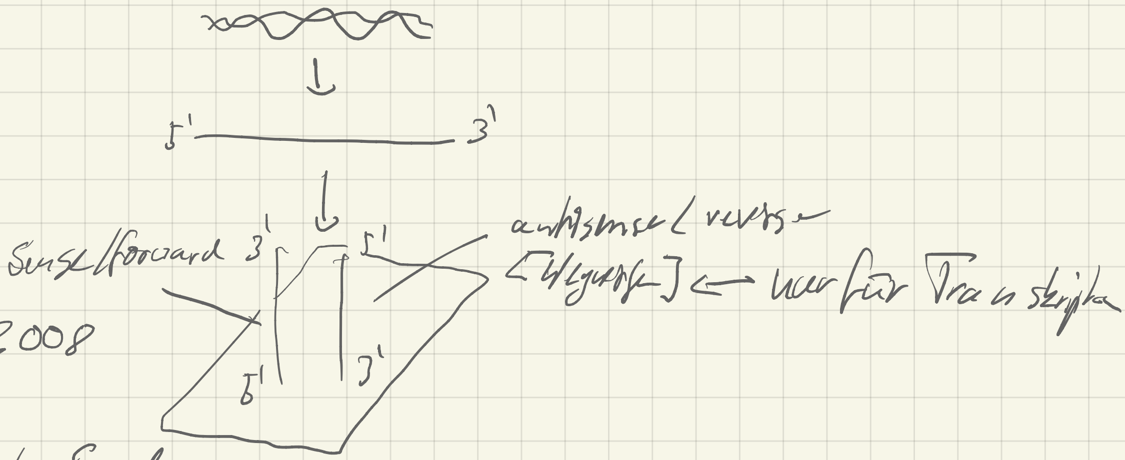
Bakterien repeats: 16 Operon sonst keine

Velvet

Paper:

Velvet Algorithms for denovo ....

Zerbino, Birney, 2008



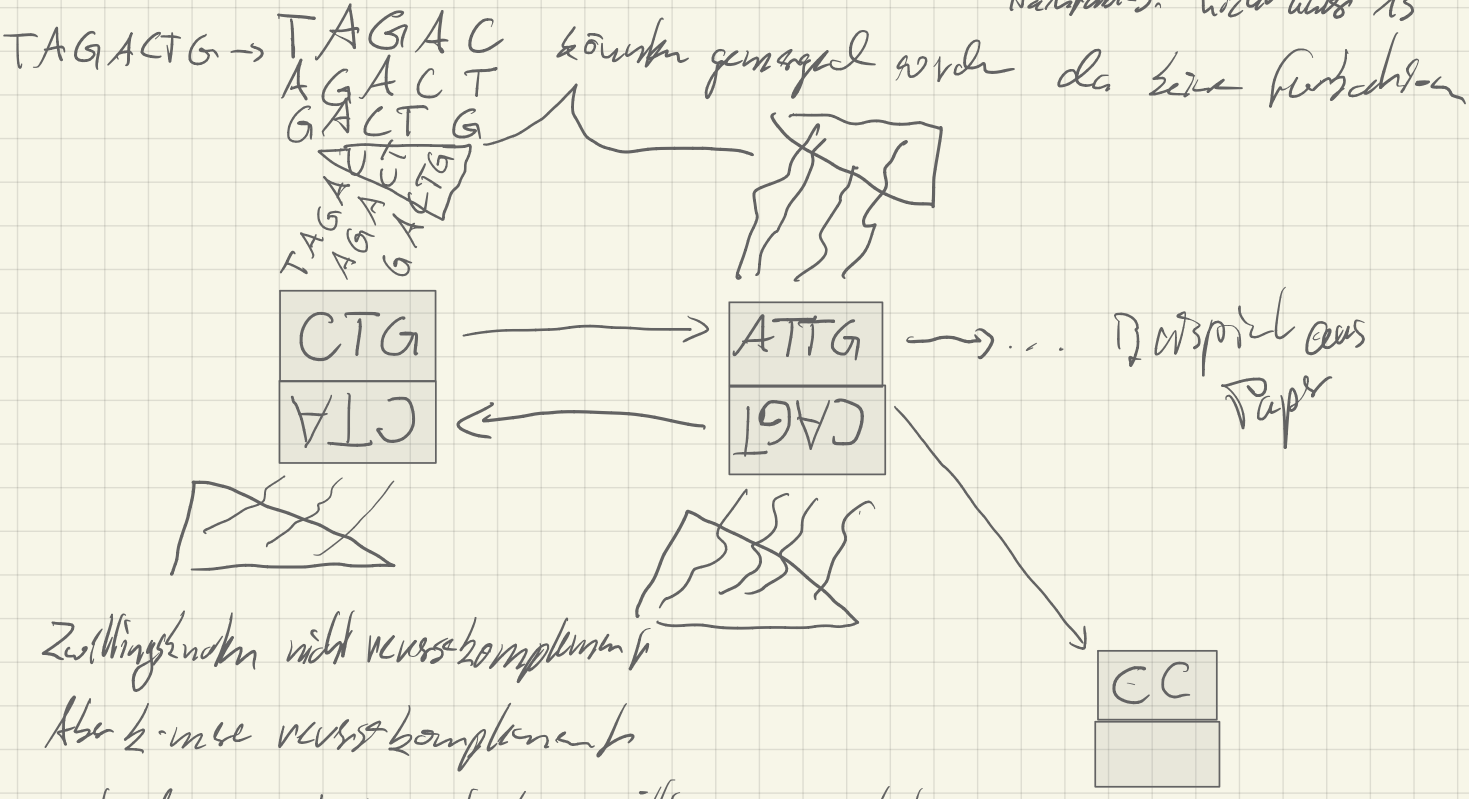
Velvet abgelöst durch Spades

Besteht auf K-mer und Ecke

Alle Knoten haben Zwillingsknoten

für 25-50 Bp reads

k-länge wählen, kleine k's kleiner graph
große k-länge zu viel rechenleistung
Hashtables. nicht unter 13



Zwillingsknoten nicht reverse komplement

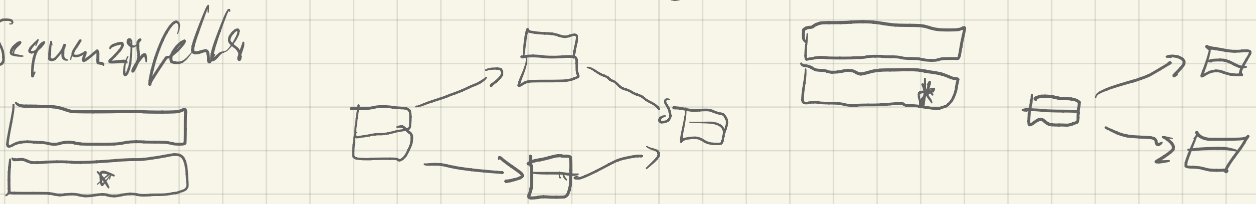
Aber k-mer reverse komplement

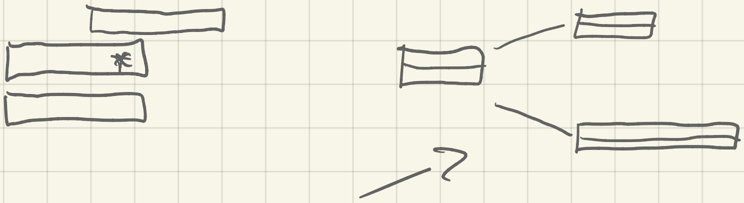
Verbindungen sind auch bei Zwilling äquivalent.

Graphen aufbauen eigentlich unabhängig vom Tool gleich.

Tools unterscheiden sich in Konfliktlösung.

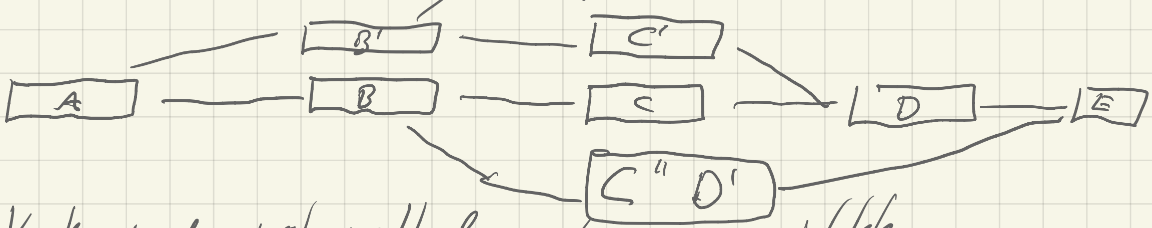
Sequenzierfehler





ⓐ TIP   Sequenzierfehler

TIP wird entfernt wenn Tip länge max 2k-1



Kanten werden mit anzahl der vorkommen gewichtet
man geht meist die stärkere Kante

ⓑ Tourbus (mehr oder weniger Dijkstra)

⇒ Mehr reads heißt auch mehr Fehler. Dijkstra bekommt probleme
→ Heuristiken → ungenauer...

Mapping Verfahren   Maq, soap, cufflinks

In Review vor einiger Zeit hat zwei tools ergeben.

Bowtie → Bowtie 2 → Tophat → Hisat ← State of the Art

SEGEMEHL (Steve Hoffmann, Jena)

Bowtie: Paper: Ultrafast and Memorie efficient alignment of short DNA sequences to the human genome, 2009

Steven Salzberg

Entwickelt für Illumina Sequenzierung

2 Arten von Mappern

1. Genom in Speicher → Reads über genom [Soap, Bowtie, Segem.]
2. Reads in Speicher → Genom über reads [MAQ, Zoom]

Burrow-Wheeler   Index(?) zum schnellen suchen in Genom.

gegeben: Text T, terminalsymbol $

gesucht: BWT(T)

1. $acaacg — Terminalsymbol vorne anhängen

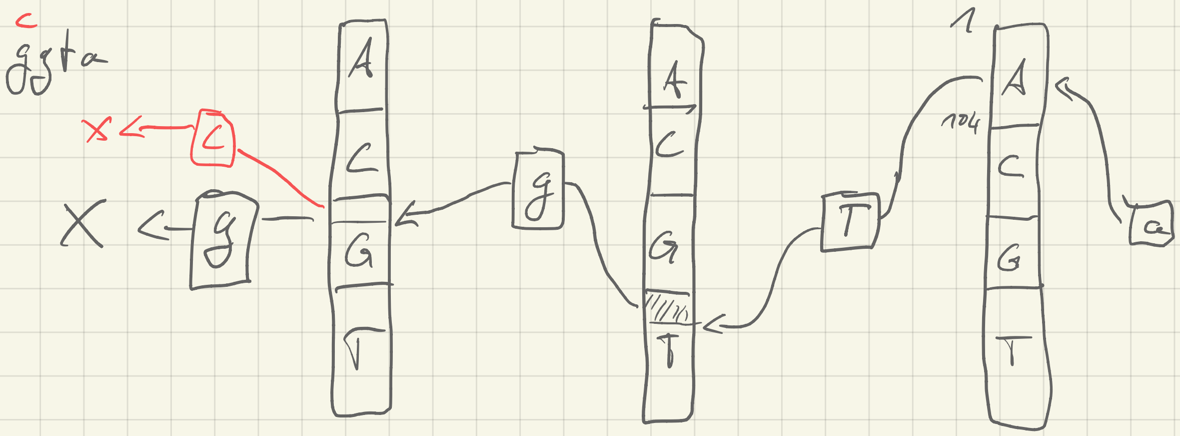
2. Lexikografisch sortieren und vorige original position in Hashmap speichern

0	$acaac	g
1	aacg$a	c
	acaacg	$
3	acg$ac	a
4	caacg$	a
5	cg$aca	a
6	g$acaa	c

← BWT

3. aac (←) — suchen der Reads

Wenn read sequenzierfehler ist schon Problematisch. Wie Handlen?



* .fastq
>id
ACTCGATG

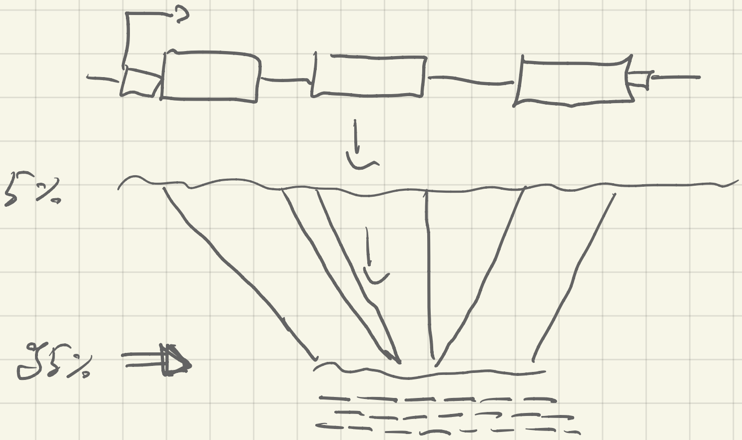
Hinten meiste Fehler

Bowtie 2 dreht reads um.

Wenn Bowtie einen read nicht zuordnen kann wird das read verworfen

Laufzeit erhöht sich stark mit erhöhung der Fehler

„Seedregionen"



Bowtie im original kann nicht mit splitreads umgehen.

-> Bowtie2 schon (= splitreads

SEGEMEHL:

- splitreads
- Insertion / Deletions
- ≥ 3 Mutationen

Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures

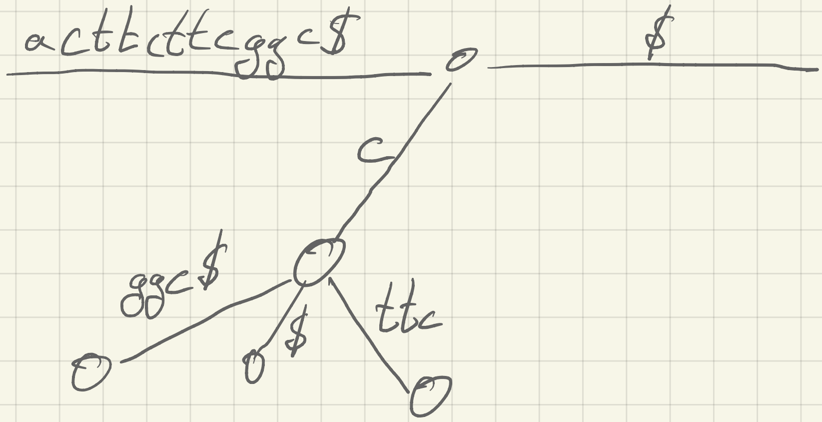
2009

Fehler in figure caption finden, für prüfung wichtig

genom: S := acttcttcggc

Suffix-Baum   terminalsymbol $

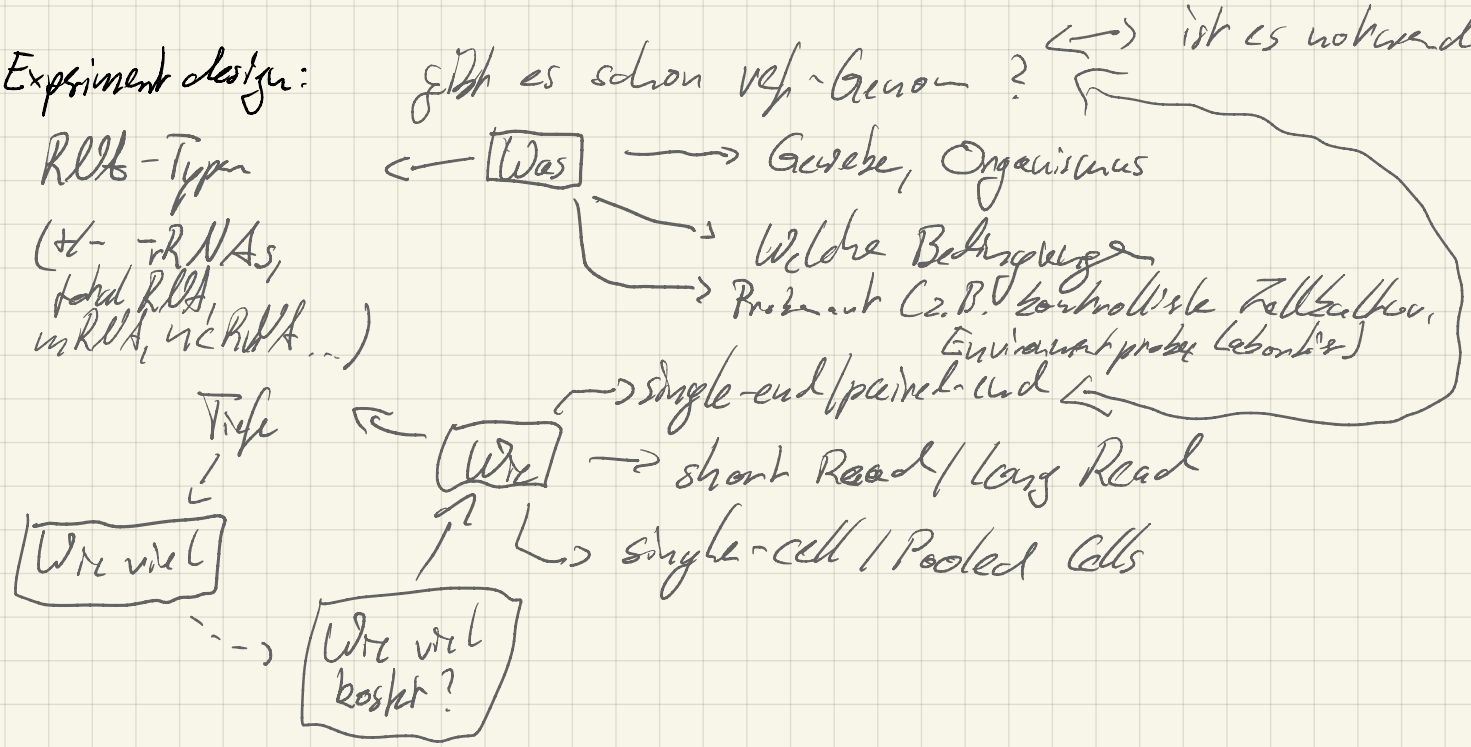
Trie ist speziell



Bowtie sehr picky vs Segemehl will mappen

Klassische Genexpressionspipline

1. Experiment design
2. Probenentnahme
3. Sequenzierung
4. Raw Reads
5. Qualitätskontrolle der Rohdaten
6. Readprozessierung
6.5 Assemblierung
7. Mapping
8. Postprocessing
9. Readcounting Expressions-quantifizierung
10. differenzielle Genexpressions Analyse



Probenentnahme + Library Prep:

1. Probenentnahme -> Lysieren -> RNA isolieren
2. RNA evtl. filtern
3. RNA in cDNA umschreiben
4. cDNA spalten (Enzyme oder Ultraschall)
5. Adapter ligieren

Sequenzierung: siehe vorheriger Skript

Rohdaten:

Erster Check: #Reads · Ø Readlänge ≈ Cov. / Tiefe

Fastq-Format: - Reiner Ascii-Text
- 1 Read => 4 Zeilen
1.: Header beginnt mit @
2.: Sequenz des Reads im 5 Buchstabencode
3.: Spacer / Trennsymbol beginnt mit +
4.: Sequenzqualität -> Phred-Score

vergleichen mit der theoretischen Output der Flowcell

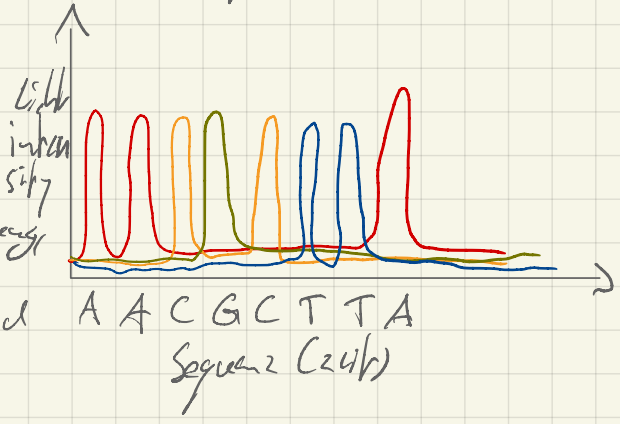
Phred-Score: Q ist ein Wert, der sich logarithmisch zur Sequenzierfehlerwahrscheinlichkeit P einer einzelnen Base verhält.

$$Q = -10 \cdot \log_{10}(P)$$
$$P = 10^{-1 \cdot \frac{Q}{10}}$$

-> Prüfung! und was macht Phred

Q = 10 -> P = 0.1 ≈ Accuracy 90%
Q = 20 -> P = 0.01 ≈ Accuracy 99%
Q = 30 ... 99,9%
⋮
Q = 60 ---- 99,9999%

Phred ist eigentlich ein Base-Caller (-> Software, welche Lichtsignale in Rohdaten umwandelt)
Anhand bestimmter Parameter wie z.B. Peakfarbe, Peakbreite, Peakauflösung, Peaküberlappung, ... ordnet Q wert und Base zu



Qualitätskontrolle:

Bekanntes Tools: FastQC

- Analysiert eine zufällige Teilmenge an Reads eines Fastq-Files und erstellt aus diesen Sequenzen und Quality-Scores verschiedene statistische Metriken. → Sollen einem helfen die Gesamtqualität des Fastq-Files abzuschätzen.
- Wurde eigentlich für Whole-Genom Sequenzierprojekte entwickelt, deshalb sind ein paar Metriken für RNA-Seq irrelevant und können ignoriert werden.

Read Prozessing:

① Entfernen von Adaptersequenzen (evtl. Barcodes entfernen)
② Abschneiden von 5'- und 3'-Enden der Basen mit schlechter Qualität
③ Entfernen ganzer Reads, wenn z.B. Ø Qualität zu niedrig, zu kurz, zu viele N Basen

Tools: fastp, prinSeq, Cutadapt, trimmomatic

Mapping:

~ Zuordnung der Reads zur ursprünglichen Transkriptionsstelle in der Referenzsequenz.

Tools: Segemehl, Hisat2, STAR 2nd
minimap2 3rd Gen (Nanopor, PacBio)

Split-Read Aligner:

Warum Splitten: Introns und Splicing können Lücken im Sequenzalignment
⇒ Infos über expression und evtl. neue (unbekannte) exons/Isoformen

Wenn Split nicht betrachtet: Wenn zu wenig Splits an pos. oder zu langer Split (bei Hsa ≈200000 max

Mehrere Treffer: bis zu einem gewissen Punkt alle Positionen merken später entscheiden wo am Wahrscheinlichsten.

- Output: SAM-Format
  - Besteht aus mehreren Header zeilen (@)
  - Für jedes Read-Alignment eine Zeile aus mind. 11 Spalten

Wenn schlechtes Mapping:
- Probe könnte kontaminiert sein
- Falsches/schlechtes Referenz genom
- Falsche Mapping-Strategie (Parameter oder Tools)
- Biologische gründe (Mutationen etc.)

Postprocessing:
-> Konvertierung in BAM (Binär) für Speicher — IGV
-> Filterung bestimmter Alignments — IGB tablet
-> Sortierung (ID, pos, ...), Indexierung → Visualisierung
SAM Tools

Read Counting:
~ Ziel des Countings ist es jedem (bekannten) genomischen Feature eine korrekte Anzahl an gemappten Reads zuzuordnen

Genomisches Feature: alle genomischen Regionen, die in irgendeiner Art exprimiert werden, aber es gibt auch Ausschlüsse, Ausnahmen! Sequenzierungs methode bedingt.

- Was wollen wir zählen?
- Ist die zugrundeliegende Annotation vollständig (genug)?
- Wie gehen wir mit multimapped Reads um?
- Reads die auf Feature mappen die sich überlappen

Differenzielle Expressionsanalyse:

Begründung der Strategie wichtig!

- Wie stabil das Verhältnis (Foldchange) der Expressionsänderung eines Gens in unterschiedlichen Bedingungen aus?

→ Wir vergleichen die Counts jeweils eines Gens zwischen den Bedingungen

vgl. Slides.

↳ Counts normalisieren!

Nanopore

Software zu MinION ist MinKNOW

RAW-Signal in HDF5 (.fast5) → "Squiggle" genannt

↳ Array von Integer

RNA braucht poly-A und wird von 3' nach 5' seq (DNA andersrum)

Base caller ist Guppy

Anwendung

Projekt mit Gynäkologie (Plazenta labor)

Spülen Plazenta durch (Blut raus)

Schicken von Material an befreundete Seite was rein und schaut was ankommt.

Trophoplastschicht (dort treffen sich Mutter und Embryo)

Transkripten aus nabelschnur

Transkripten von Trophoplastschicht > Exosome

↳ was da drin?

→ Embryonale Exosome kommen zu Mutter

→ Exosome für Kommunikation?

Exosomen unterscheiden sich von Mikrovesikeln in größe und oberfläche

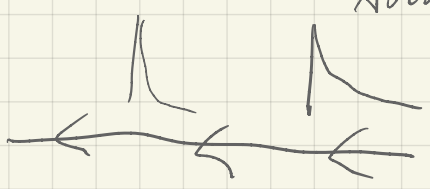
=> Artefakte finden! mit IGV und UCSC

Artefakt microRNA expression anschalt Proben

Transcriptome genome browser → google image search

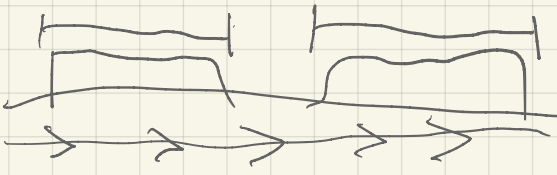
Single Nucleotide → „Fehler" bei reverse Transkript, falsche Annotation ↑ post Transkript geändert.

Artefakt



→ Splicing / Sashimiplot → lang distance Splicing

→ nicht das Gen oberhalb DESEQ



etwa 22 bp lang 2 sind festgelegt 1 auch möglich

Lesen von Genombilder:
Transkription Profile:
In Prüfung bild lesen können

Prüfungstermin: 17.08.
8^{45}

---

microRNA
Eigener Promotor -> Poly A
(intronisch Clarint) -> kein Poly A
weitere Degredation
DICER teilt rest
Wegen shift auf Hairpin zwei Micro Übereinander
Jede non-coding RNA hat eigene Signatur

Wellen in Proteinen sind auf Sekundärstruktur und Sequenziermethode zurückzuführen