Projekt

Java-Anwendung für die Sequenzanalyse (Metagenomik und Transkriptomik)

MHH – Prof. Tümmler, Dr. Davenport
FH – Prof. Sprengel, Prof. Ahlers

C. Davenport colindaven<at>gmail.com

Version 27.09.2010

Spezifikation

Programmiert in Java

Speicher optimiert (10000 < x < 100x10^6) Reads müssen eingelesen werden

Getestet in Windows, Linux

Sprache: Englisch (soweit möglich, CD wird an der Übersetzung teilnehmen)

Grundlagen

DNA – A, T, G, C, N chars, in einer Kette.

Genom – Zeichenkette (A,T,G,C, N), durchschnitt 3 x 10^6 chars

Read – Zeichenkette (A,T,G,C,N), 36-1000+ chars, meist 36-500 chars

Sequenzierung – Entzifferung des DNAs in einem Genom anhand von Reads, die von einer Maschine produziert werden

Metagenomik – Zuordnung von Reads zu 2+ Genome mit einem Aligner, Zusammenzählung, Statistik

Transkriptomik – Zuordnung von Reads zu einem Teil vom Genom, also ein Gen. Reads innerhalb und ausserhalb Gene werden gezaehlt

Teil 1: Metagenomik

Reads gemappt auf Spezies

Metatie

Unser Program heisst Metatie und läuft entweder lokal oder auf

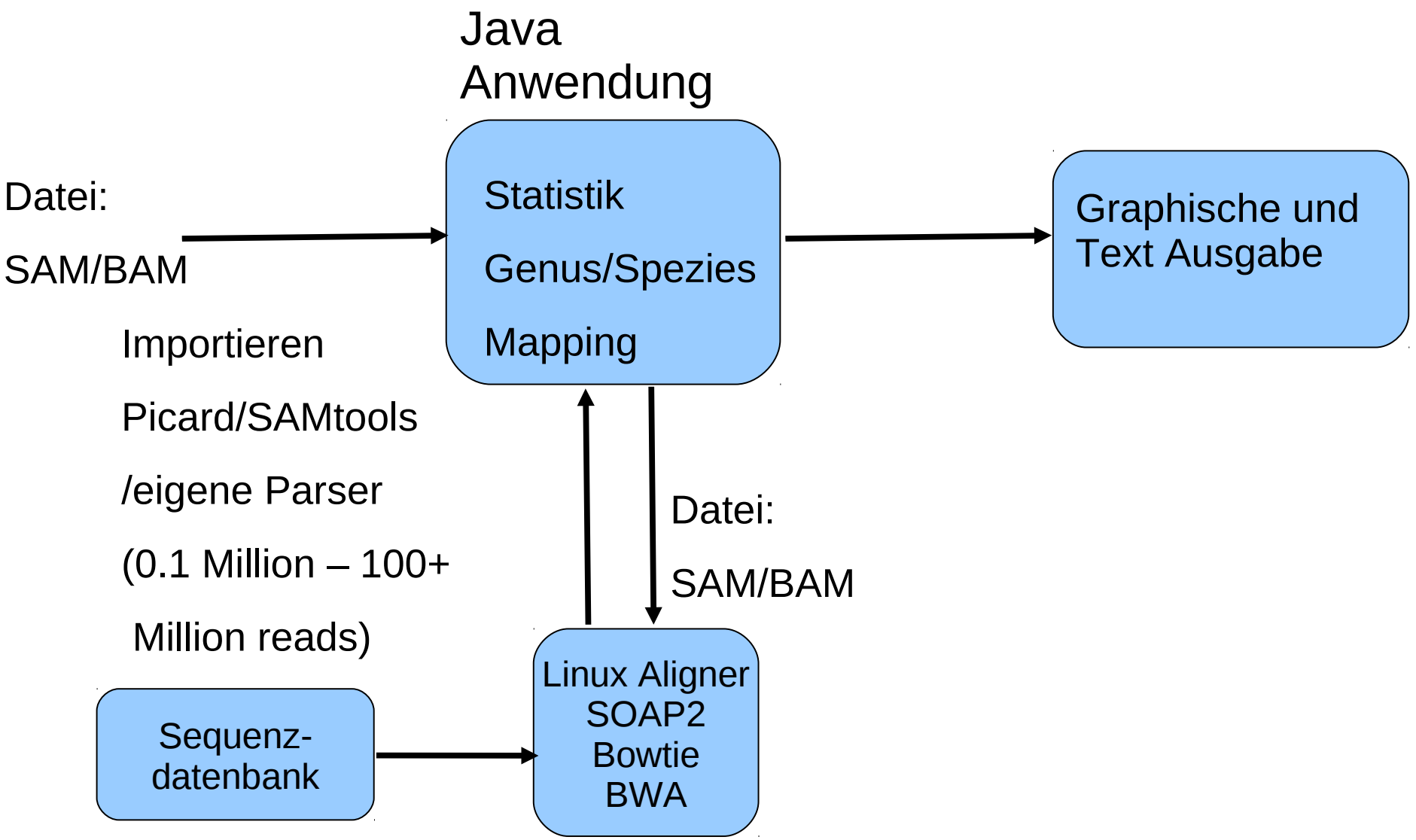
http://genomics1.mh-hannover.de/metatie/index.php

Es besteht aus einem Perl pipeline Skript – siehe Verzeichnis Metatie

Reads werden

- Aligned - *Bowtie*
- Gezählt und aufsummeriert - *Java*
- Graphisch dargestellt (PDF) mit *R*

Workflow 1


Java
Anwendung
Datei:
SAM/BAM
Importieren
Picard/SAMtools
/eigene Parser
(0.1 Million – 100+
Million reads)
Statistik
Genus/Spezies
Mapping
Graphische und
Text Ausgabe
Datei:
SAM/BAM
Sequenz-
datenbank
Linux Aligner
SOAP2
Bowtie
BWA

Workflow 2 – Metagenomik Ausgabe

Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Bakterien,
Wie viele Reads
wurden einem
Taxon eg.
Bakterium
zugeordnet

E. coli 5000 reads
Pseudomonas 85
Salmonella 53
H. sapiens 5

Graphische Ausgabe

Bar Grafik
X axis = 1 bis Genomlänge(z.B. 3X10^6)
Y axis = Anzahl von Reads

Überblick
Zoom-fähig
Features z.B. Gene einlesen von GFF Datei und als horizontale Balken darstellen

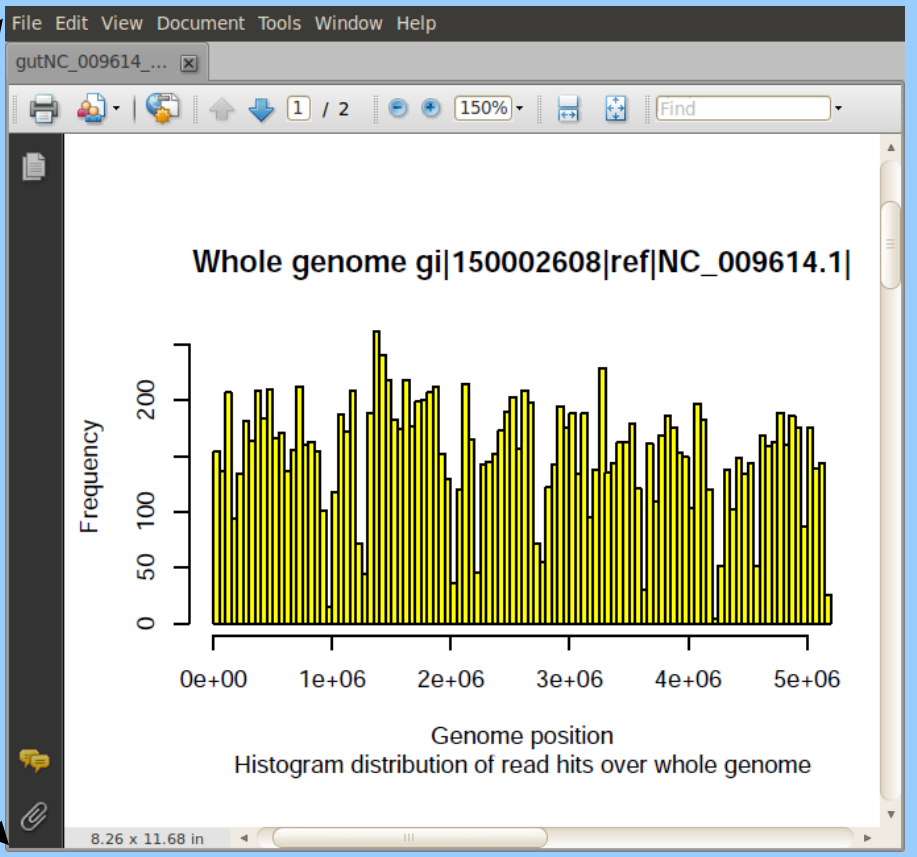
Workflow 2 – Metagenomik Ausgabe

Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Bakterien,
Wie viele Reads
wurden einem
Taxon eg.
Bakterium
zugeordnet

E. coli 5000 reads
Pseudomonas 85
Salmonella 53
H. sapiens 5

Beispiel



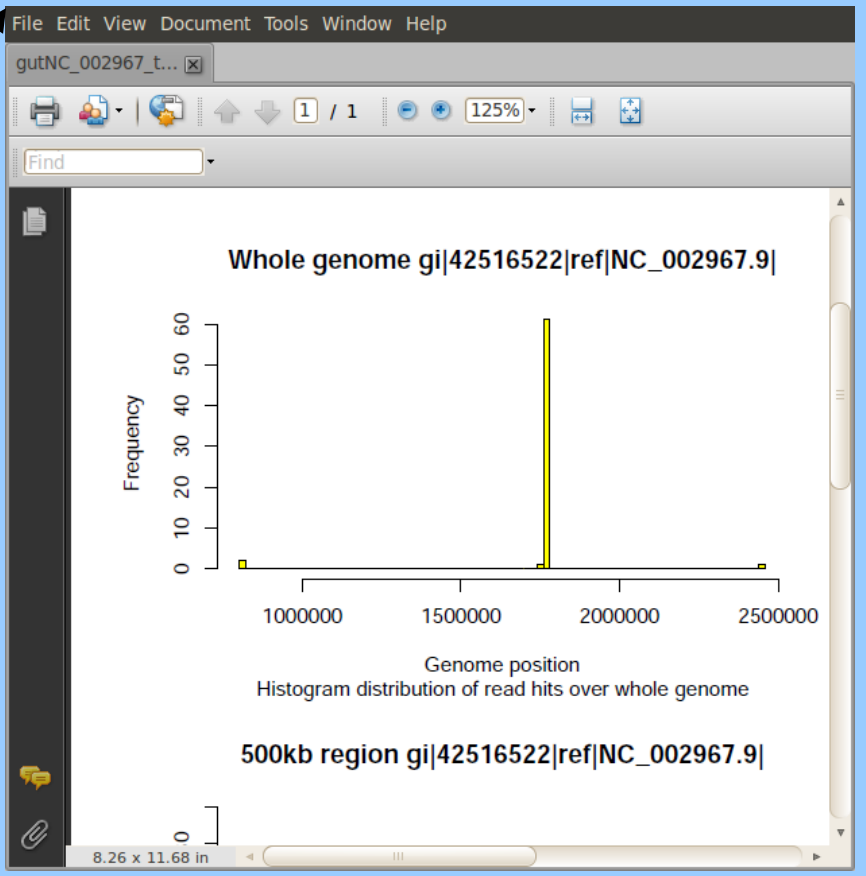
Workflow 2 – Metagenomik Ausgabe

Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Bakterien, Wie viele Reads wurden einem Taxon eg. Bakterium zugeordnet

E. coli 5000 reads
Pseudomonas 85
Salmonella 53
H. sapiens 5

Beispiel



Workflow 2 – Metagenomik Ausgabe

Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Bakterien,
Wie viele Reads
wurden einem
Taxon eg.
Bakterium
zugeordnet

E. coli 5000 reads
Pseudomonas 85
Salmonella 53
H. sapiens 5

Output

SVG/PDF
PNG

Listen (linker Kasten) als CSV txt

Datenformate: FASTA,

- Header + Zeichenketten
- .fasta, .fa, .fas, .fna

>gi|110645304|ref|NC_002516.2| Pseudomonas aeruginosa PAO1, complete genome

TTTAAAGAGACCGGCGATTCTAGTGAAATCGAACGGGCAGGTCAATTTCCAACCAGCGATGACGTAATAG

ATAGATACAAGGAAGTCATTTTTCTTTTAAAGGATAGAAACGGTTAATGCTCTTGGGACGGCGCTTTTCT

GTGCATAACTCGATGAAGCCCAGCAATTGCGTGTTTCTCCGGCAGGCAAAAGGTTGTCGAGAACCGGTGT

CGAGGCTGTTTCCTTCCTGAGCGAAGCCTGGGGATGAACGAGATGGTTATCCACAGCGGTTTTTTCCACA

CGGCTGTGCGCAGGGATGTACCCCCTTCAAAGCAAGGGTTATCCACAAAGTCCAGGACGACCGTCCGTCG

GCCTGCCTGCTTTTATTAAGGTCTTGA

Datenformate: multi FASTA

- Headers + Zeichenketten
- .fasta, .fa, .fas, .fna

>SRR032798.1

GGGGGACAGGAGCCACCCTGCCCACCCACAGCCCTCGTTGCCGTAAAGAGGCCACGGGTACCCGGGCT
GCGCCCAGGACCTGCCCTCGGGCCGGCGCGCCTTGGGGATAGGTTGGCGCCTTCCACGAGGTTGCGA
CCGCTCCGAAGTCTTCCGAGTCGCGCGCACACCCAATCTTGGGCCCCTGCGCGGCAACAAGTAAACTC
CGCCAACGATCTGGCCACCACCCGGGAACCTAACGTCCATGGGGCGGCGGTTGGTGTTACGGTGGGTT
TTTCTTTGAGGTTTAGGATTCATGCTCATGGTGCATGGTCTACGAGACCTCCCGGGG

>SRR032798.2

GCTGCCCAGGGCTATAAGGTGCTGGTGCTCAACCCCTCTGTAGCGGCTACTTTGGGCTTTGGGGCGTATA
TGTCCAAGGCACACAGGGGATAGGC

>SRR032798.3

CTCCGGGATACCGCATCCTGAGGCAATGTGGTGGTCTCTATGGAGAAGGTAGGGTCGAGGCTGAAGTCG
ATGGTTTGGACGACAGAAGTGTTGCAGTCTATGACCGAGTCAAAGTCACCGGTGAAGCC

>SRR032798.4

AATATGCCGACGGCATGACCCAGGGGGCACAGCAACGGACCACCCGAAGAGCCCTTAAGGGTAGAAATT
GGTTTGGGGCTCAAGAGAGCGCCCCGGGTGTCACCGCGTCTGCGAACAGGGATCACGTCTGCGTGCCT
GGTGACCAGGAATAAGTCCGATGCACCACAAGTGCATGGCGTAAGGGACTTTACTCCCGGGGGTGGCT
GGCCACCCCACCAGGTCTT

Datenformate: SAM

- Alignmentdaten
- Mit Zeichenketten (Reads)
- Mit Positionsangabe (von der Genomstring) wo

der Read auf dem Genomstring trifft

- Kann in binären BAM Format mit Samtools konvertiert werden
- Darstellung als Balken

KN-1065_01_1_2_628_85 16 NC_002947.3 171372 255 32M * 0 0
CAACTGAAGAGTTTGATCATGGCTCAGATTGA IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII XA:i:0 MD:Z:0G31
NM:i:1

KN-1065_01_1_64_198_561 16 NC_002947.3 9 255 32M * 0 0
CTCGGAAGTCGACCAACAAGTCAGCTATGACT IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII XA:i:0 MD:Z:32
NM:i:0

Datenformate: GFF

- Start, Stop, Richtung
- z.B. Genombereiche wie Gene
- Darstellung als Balken

##gff-version 3

##source-version NCBI C++ formatter 0.2

##date 2007-01-23

##Type DNA NC_002947.3

NC_002947.3 RefSeq gene 147 1019 . - . ID=NC_002947.3:parB;locus_tag=PP_0001;db_xref=GeneID:1043468

NC_002947.3 RefSeq CDS 150 1019 . - 0 ID=NC_002947.3:parB:unknown_transcript_1;Parent=NC_002947.3:parB;locus_tag=PP_0001;note=identified%20by%20match%20to%20PFAM%20protein%20family%20HMM%20PF02195;transl_table=11;product=chromosome%20partitioning%20protein%20ParB;protein_id=NP_742171.1;db_xref=GI:26986746;db_xref=GeneID:1043468;exon_number=1

NC_002947.3 RefSeq start_codon 1017 1019 . - 0 ID=NC_002947.3:parB:unknown_transcript_1;Parent=NC_002947.3:parB;locus_tag=PP_0001;note=identified%20by%20match%20to%20PFAM%20protein%20family%20HMM%20PF02195;transl_table=11;product=chromosome%20partitioning%20protein%20ParB;protein_id=NP_742171.1;db_xref=GI:26986746;db_xref=GeneID:1043468;exon_number=1

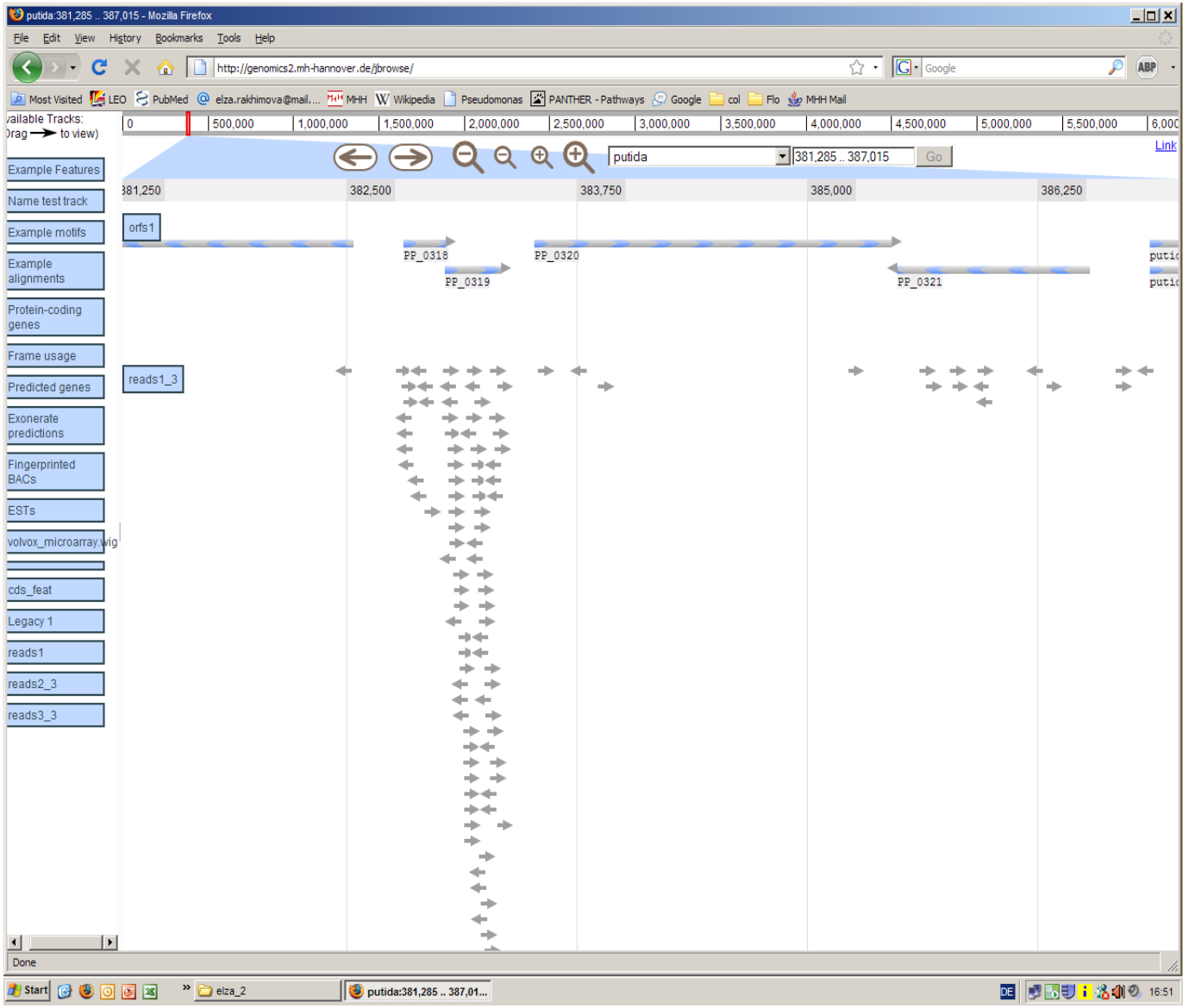
NC_002947.3 RefSeq stop_codon 147 149 . - 0 ID=NC_002947.3:parB:unknown_transcript_1;Parent=NC_002947.3:parB;locus_tag=PP_0001;note=identified%20by%20match%20to%20PFAM%20protein%20family%20HMM%20PF02195;transl_table=11;product=chromosome%20partitioning%20protein%20ParB;protein_id=NP_742171.1;db_xref=GI:26986746;db_xref=GeneID:1043468;exon_number=1

NC_002947.3 RefSeq gene 1029 1820 . - . locus_tag=PP_0002;db_xref=GeneID:1043469

NC_002947.3 RefSeq CDS 1032 1820 . - 0 locus_tag=PP_0002;note=similar%20to%20GB:V00540%2C%20GB:J00212%2C%20GB:M12350%2C%20GB:M28586%2C%20GB:M10201%2C%20SP:P01568%2C%20SP:P05014%2C%20PID:184655%2C%20PID:306905%2C%20PID:306906%2C%20PID:306912%2C%20PID:32717%2C%20PID:32725%2C%20PID:758078%2C%20and%20PID:825603%3B%20identified%20by%20sequence%20similarity%3B%20putative;transl_table=11;product=ParA%20family%20protein;protein_id=NP_742172.1;db_xref=GI:26986747;db_xref=GeneID:1043469;exon_number=1

Visualisierung - Beispiele

Visualisierung 1 – Jbrowse webbrowser



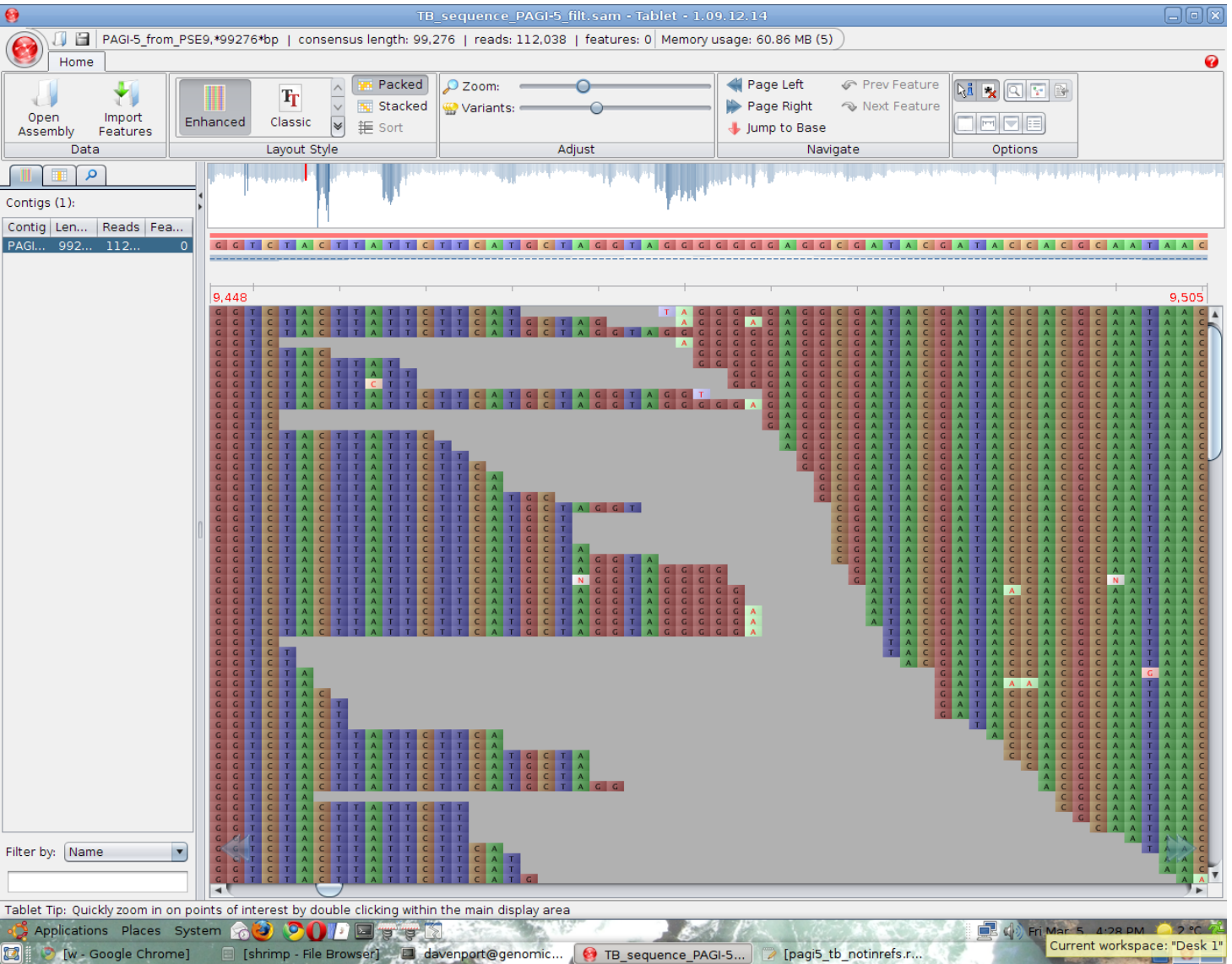
Annotation gezeigt

Farben etwas langweilig

Readposition punktgenau

Richtung von Reads angezeigt

Visualisierung 2 – Tablet (SAM)



Gute übersicht vom gesamten Chromosom (blaues Histogram)

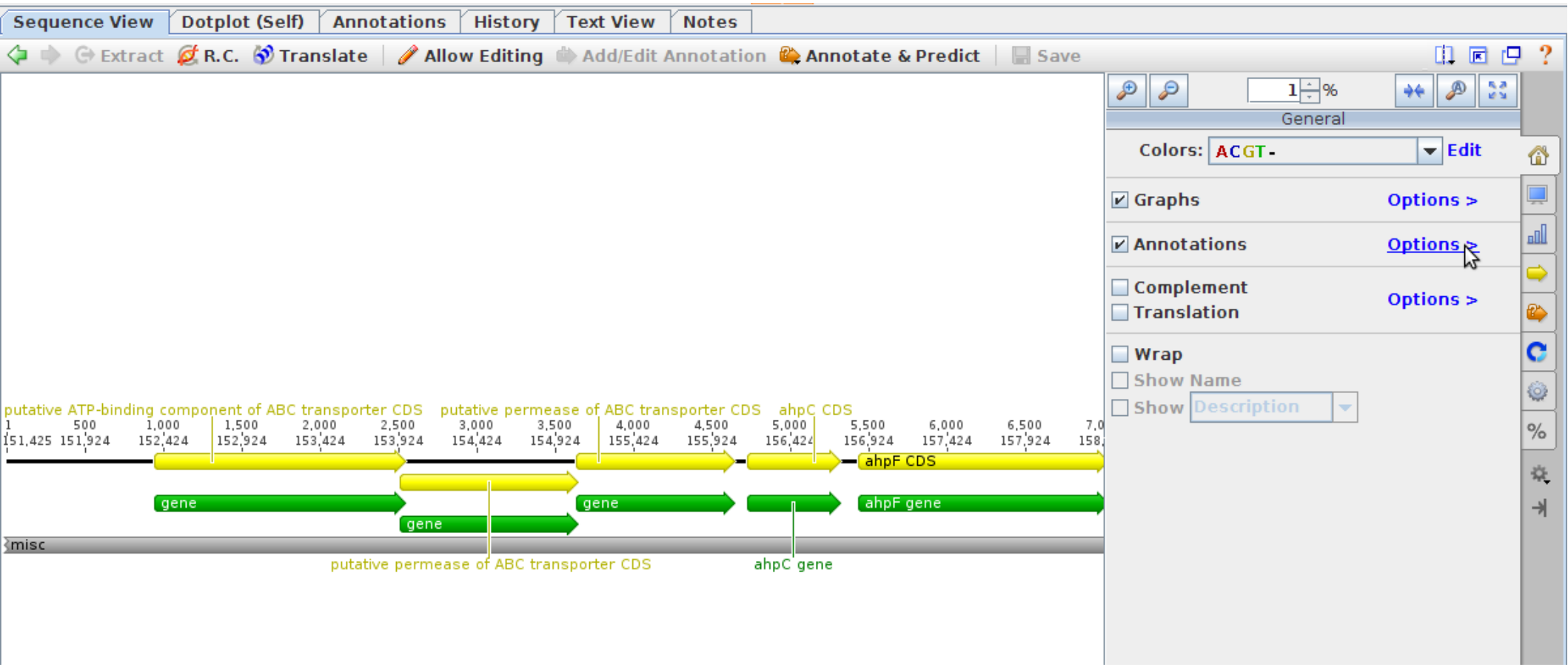
Sehr Read-zentrisch

Keine Annotationen (schlecht!)

Sehr schnelles Programm (Zoom, Skrollen)

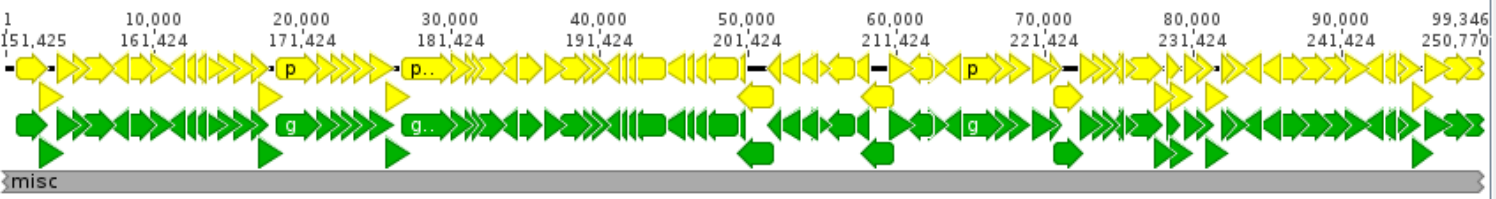
Visualisierung 3 – Geneious (genes, GFF)

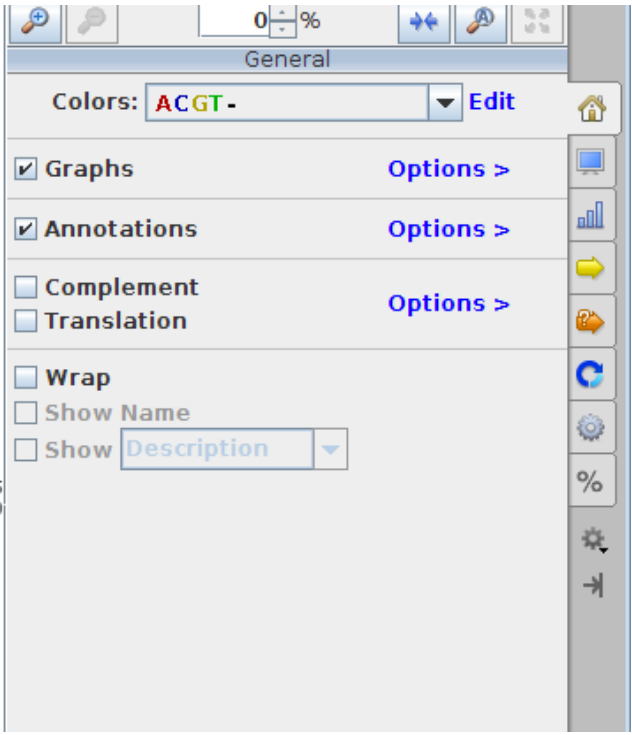
Schöne Darstellung
Verschiedene Farben für CDS (gelb), Gen (grün)
Nur Gene dargestellt, keine Reads sind hochgeladen
Annotationen auf der Grafik eingeblendet bei hohem Zoomfaktor



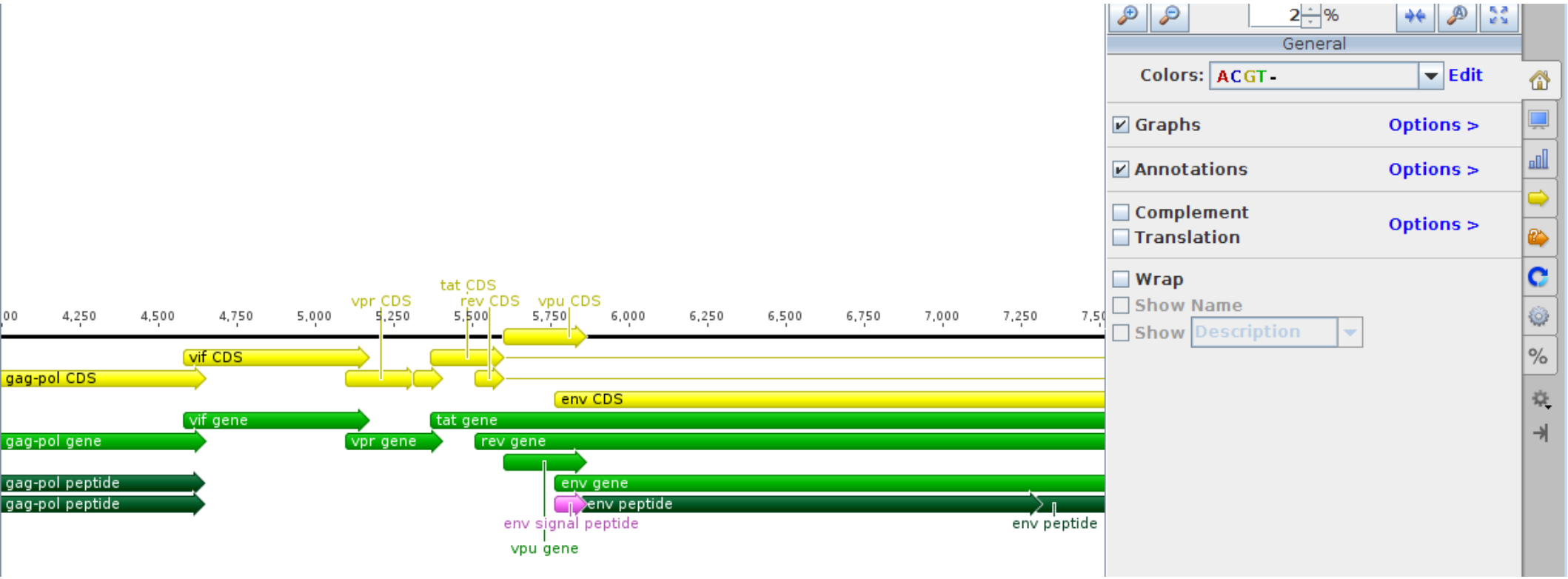
Visualisierung 3 – Geneious (genes, GFF)

Attraktive Grafik
Klar definiert
Optionen zum anzeigen von -
Annotationen ja/nein
Klassen von Annotationen zB gene, repeat, CDS





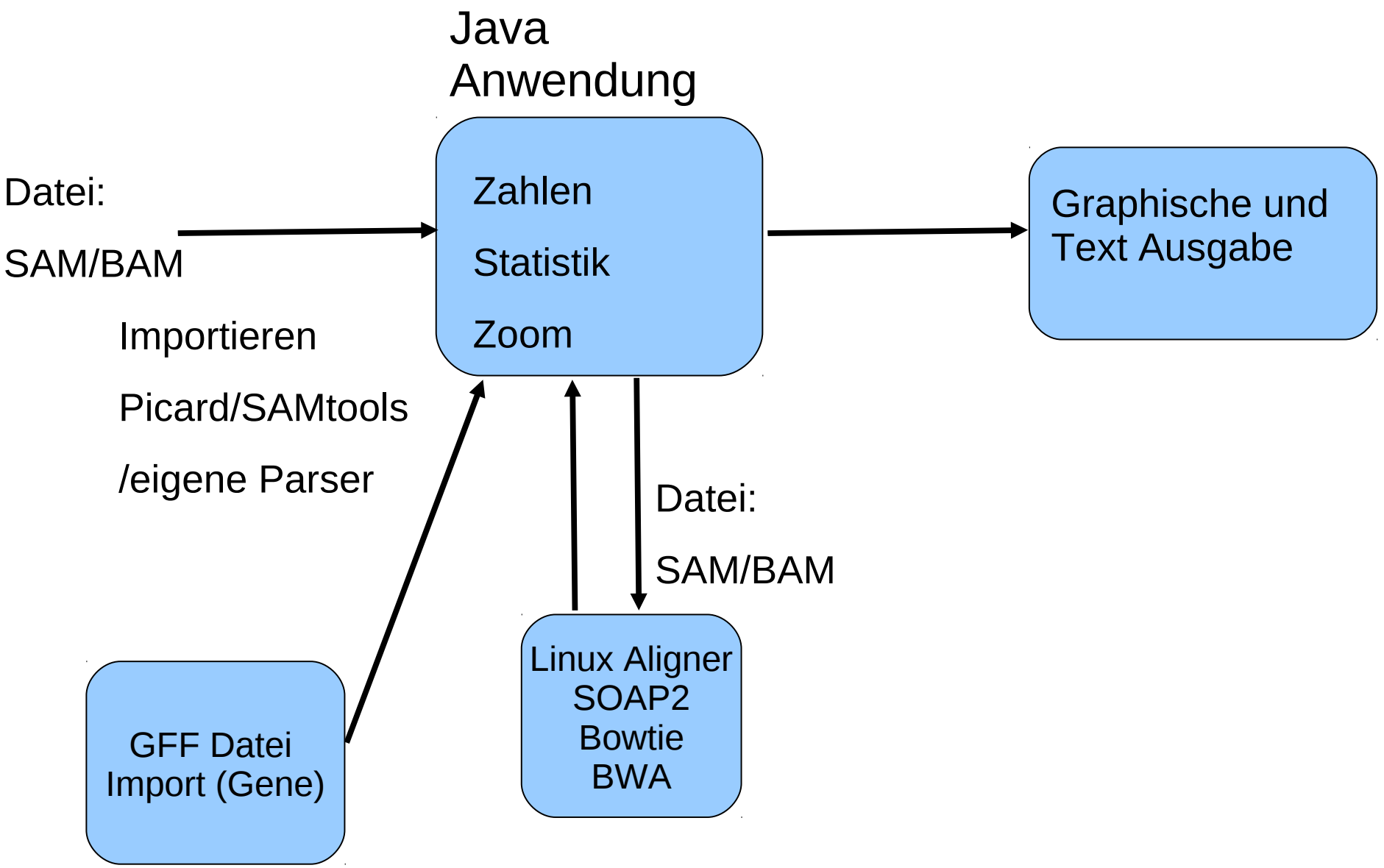
Visualisierung 3 – Geneious (genes, GFF)



Teil 2: Transkriptomik

Reads gemappt auf Gene (von einem Spezies)

Workflow 3


Java
Anwendung
Datei:
SAM/BAM
Importieren
Picard/SAMtools
/eigene Parser
Zahlen
Statistik
Zoom
Graphische und
Text Ausgabe
Datei:
SAM/BAM
Linux Aligner
SOAP2
Bowtie
BWA
GFF Datei
Import (Gene)

Workflow 4 – Transkriptomik Ausgabe

Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Gene,
Wie viele Reads
wurden ein
Gen zugeordnet

Gen1 50000 reads
Gen2 115
Gen3 60
Gen4 15

Graphische Ausgabe

Bar Grafik
X axis = 1 bis Genomlänge(z.B. 3X10^6)
Y axis = Anzahl von Reads

Überblick
Zoom-fähig
Features z.B. Gene einlesen von GFF Datei und als horizontale Balken darstellen

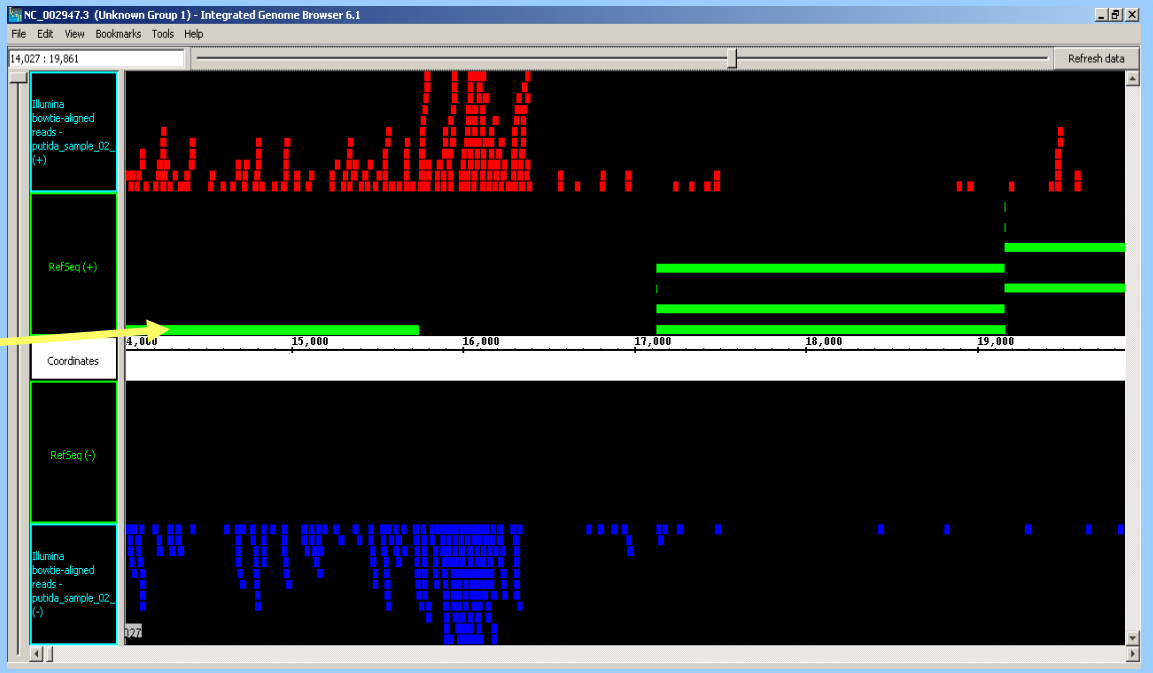
Workflow 4 – Metagenomik Ausgabe

Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Gene,
Wie viele Reads
wurden ein
Gen zugeordnet

Gen1 50000 reads
Gen2 115
Gen3 60
Gen4 15

Beispiel



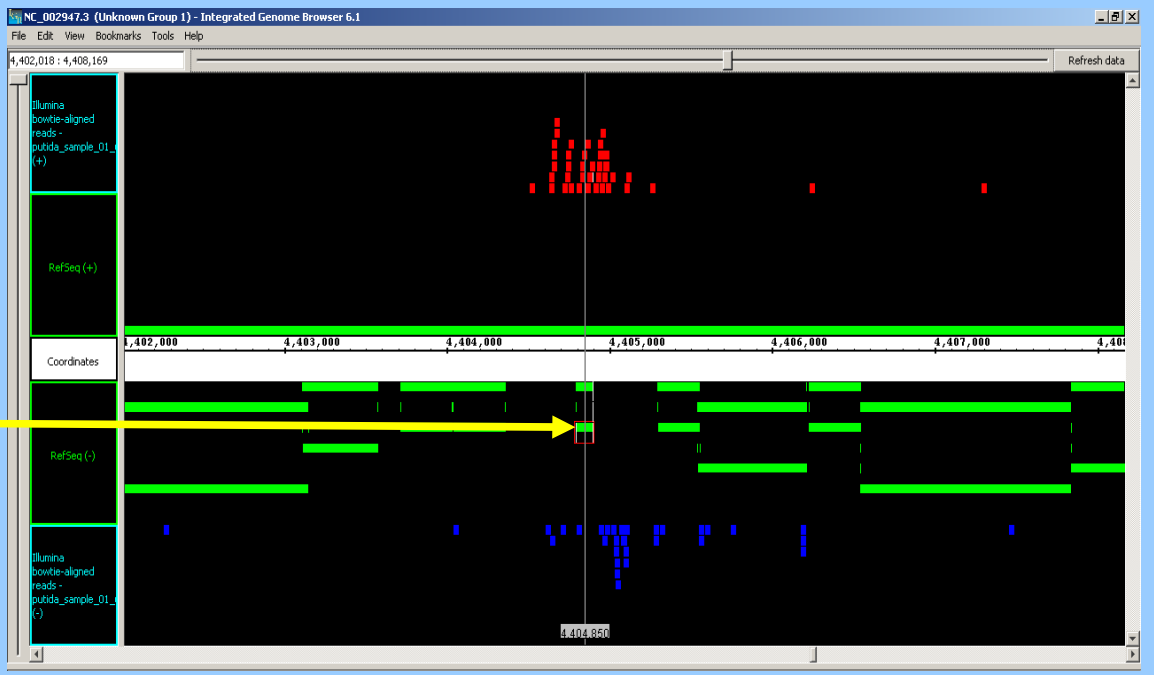
Workflow 4 – Metagenomik Ausgabe

Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Gene,
Wie viele Reads
wurden ein
Gen zugeordnet

Gen1 50000 reads
Gen2 115
Gen3 60
Gen4 15

Beispiel



Workflow 4 – Metagenomik Ausgabe

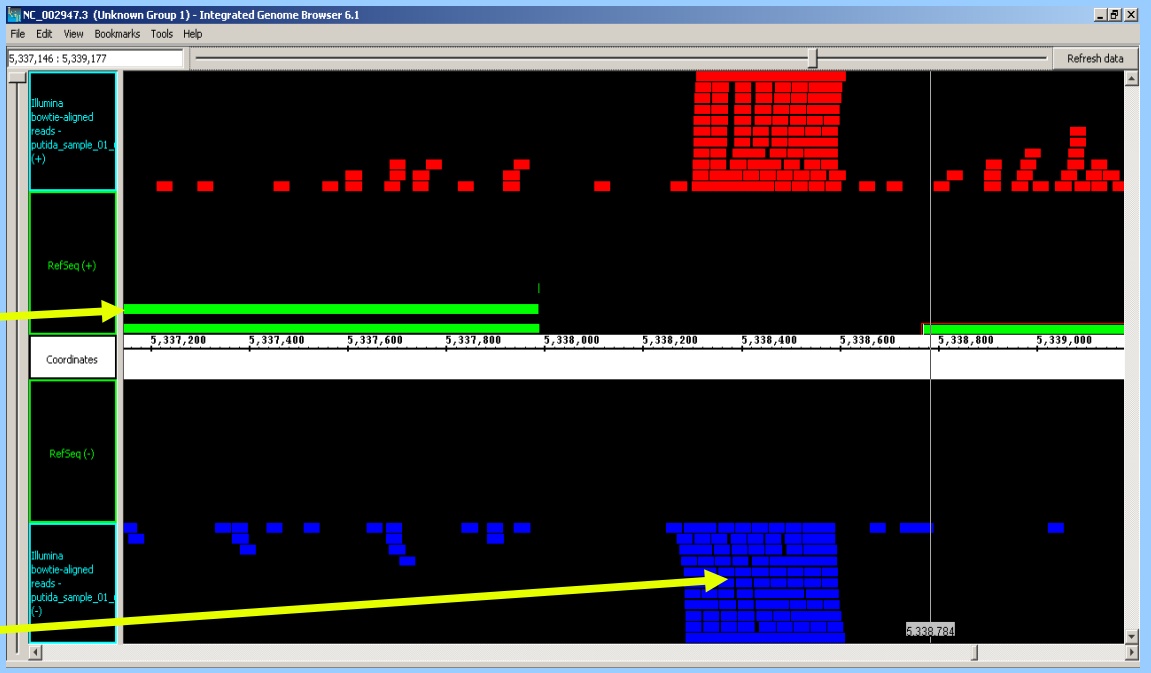
Menu – SAM Import, GFF Import, SVG Export, Tabelle Export

Liste von Gene,
Wie viele Reads
wurden ein
Gen zugeordnet

Gen1 50000 reads
Gen2 115
Gen3 60
Gen4 15

Reads nicht im Gen
Intergenischer
Bereich
IGR1 1500 reads

Beispiel



Architektur

Eigene Browser ?

Oder gebaut auf

IGB - http://www.bioviz.org/igb/

IGV - http://www.broadinstitute.org/igv/

Apollo (open source)-

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705230/?tool=pubmed

Picard (Java, open source, BAM/SAM lesen u prüfen)-

http://picard.sourceforge.net/index.shtml

Modularer Aufbau mit Plugins ist wahrscheinlich zu aufwändig für die geplanten Zeit

Weitere Visualisierung

Histogramme

Pie Grafiken

Andere Vorschläge?