Alexander Herrmann, Jochen Hampe

Workshop Genomdaten [1]

	Workshop Genomdaten - Ergebnisprotokoll
Autor(en)	Alexander Herrmann, Jochen Hampe
Editor(en)	
Datum	22.06.2012
Version des Dokuments	1.0.6

A: Status des Dokuments
Version 1.0.4

B: Bezug zum Projektplan
Deliverable D4.2: Ergebnisprotokoll – Workshop Genomdaten, Version 1

C: Abstract
Am 27. März 2012 fand der Workshop „Forschungsdatenmanagement biomedizinischer Genomdaten“ mit 23 Teilnehmern aus Forschungseinrichtungen aus ganz Deutschland und Vertretern der TMF in Kiel statt. Ausgewählte Referenten gaben einen Überblick über die Datenlandschaft, den Entstehungsprozess komplexer Genomdaten, mögliche Standards und Nachnutzungsszenarien. Insbesondere die dezentralisierte deutsche Forschungsstruktur, in der die Datengenerierung an kleineren Zentren geleistet wird, braucht Konzepte für eine strukturierte Speicherung und Zugänglichkeit von Genomdaten. Die Nachnutzung von Genomdaten ist dabei prinzipiell möglich und wissenschaftlich ergiebig. Hierzu wurden Beispiele mit erfolgreichen Genidentifizierungen beispielsweise für die Hämochromatose und das Gallensteinleiden unter Nutzung bereits bestehender und publizierter genomweiter Assoziationsdatensätze gezeigt. Während im Genotypisierungsbereich die Datenstandardisierung weit vorangeschritten ist und

[1]Dieses Dokument wurde im Rahmen des Projekts LABIMI/F erstellt. Das Projekt LABIMI/F wird gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter dem Förderkennzeichen RI1000/2-1.

damit auch die Nachnutzbarkeit der Daten häufig möglich ist, besteht im Bereich der Hochdurchsatzsequenzierung noch großer Handlungsbedarf. Das Feld ist technologisch noch stark im Fluss, so dass im Moment nur industriell geprägte de-facto-Standards verfügbar sind. Probleme des Datenschutzes und der Re-identifizierbarkeit von Datensätzen wurden diskutiert und hierfür gestufte Zugangs- und Authentifizierungsmodelle als Lösungsmöglichkeit diskutiert. Insgesamt wurde der große Bedarf für strukturierte Archivierungs- und Nachnutzungskonzepte in der Genomforschungsszene deutlich und die Anforderungen an die Pilotimplementierungen im Rahmen des Workshops genauer gefasst.

D: Änderungen

Version	Datum	Name	Kurzbeschreibung
1.0.1	29.04.2012	Alexander Herrmann	Erste Dokumentversion
1.0.2	30.04.2012	Jochen Hampe	Überarbeitung und Freigabe
1.0.3	17.06.2012	Alexander Herrmann	Überarbeitung nach Rückfrage Projektkoordination
1.0.4	17.06.2012	Jochen Hampe	Überarbeitung und Freigabe
1.0.5	22.06.2012	Romanus Grütz	Überarbeitung
1.0.6	22.06.2012	Alexander Herrmann	Überarbeitung und Freigabe

E: Inhaltsverzeichnis

1 Einleitung

Um die Herausforderungen der Langzeitarchivierung von Forschungsdaten einer breiteren Community nahezubringen, wurde am 27. März 2012 der Workshop „Forschungsdatenmanagement biomedizinischer Genomdaten“ im Kiel mit entsprechendem Fachpublikum durchgeführt. Ausgewählte Referenten gaben dabei einen Überblick über die Datenlandschaft, den Entstehungsprozess komplexer Genomdaten, mögliche Standards und Nachnutzungsszenarien.
Die dezentralisierte deutsche Forschungsstruktur, in der die Datengenerierung an kleineren Zentren geleistet wird, braucht Konzepte für eine strukturierte Speicherung und Zugänglichkeit von Genomdaten. Dabei sollen die Probleme des Datenschutzes und der Re-identifizierbarkeit von Datensätzen beachtet werden.
Der Bereich der Hochdurchsatzsequenzierung ist technologisch noch stark im Fluss, so dass im Moment nur industriell geprägte de-facto-Standards verfügbar sind. Die Nachnutzung von Genomdaten ist prinzipiell möglich und wissenschaftlich ergiebig.

2 Material und Methoden

Am 27. März 2012 wurde von 14.00 bis 18:00 Uhr im Ärztekasino, 2. OG, Klinik für Innere Medizin I, Kiel der Workshop „Forschungsdatenmanagement biomedizinischer Genomdaten“ durchgeführt. Das detaillierte Workshop-Programm ist im Anhang 5.1 eingefügt. Die sieben Vortragenden haben in Kurzvorträgen (10-20 Minuten) mit anschließender Diskussion (10-15 Minuten) das Projekt und erste Lösungsansätze im Detail vorgestellt. Daneben berichteten ausgewählte Referenten über den Entstehungsprozess und die Archivierung digitaler Forschungsdaten. Abschließend nahmen die insgesamt 23 Anwesenden an einer offenen Diskussion über das Thema „Forschungsdatenmanagement“ teil.

3 Ergebnisse und Diskussion

3.1 Langzeitarchivierung: Beweisfunktion vs. Sharing-Option

Bei der Archivierung der Sequenzierungsdaten sind zwei prinzipielle Anwendungsfälle relevant: Einerseits kann die Archivierung zur Nachvollziehbarkeit von Daten, Analysen und Experimenten dienen, um später evtl. Fehler zu finden und den Richtlinien zur Aufbewahrung und Nachvollziehbarkeit der Projektförderer zu genügen. Ein anderer Anwendungsfall ist das Data Sharing, wo die Daten zur Nachnutzung anderen Forschern zur Verfügung gestellt werden. Jeder dieser Anwendungsfälle hat unterschiedliche Anforderungen beim Datenschutz, bei dem Detailgrad der zu speichernden Daten und in der Relevanz der Daten für die Nachnutzbarkeit. Die kontinuierliche Finanzierung der Datenarchivierung ist ein Infrastrukturproblem, das von den Universitäten/Klinika übernommen werden sollte und schwer über Projekte zu finanzieren ist.

Eine wichtige Anforderung an die Zielinfrastruktur zur Archivierung der Sequenzierungsdaten ist die Möglichkeit der automatischen Metadatenextraktion. Diese Metadaten sind für das Wiederauffinden der Forschungsdaten und weitere Analysen relevant. Wichtig ist dabei ein sinnvolles Verhältnis von Aufwand (Dateneingabe) für den Forscher und den späteren möglichen Nachnutzungen.

3.2 Qualitätsunterschiede der Sequenzierungstechniken

Im Vortrag von Dr. Nothnagel wurden die Qualitätsunterschiede zwischen aktuellen Sequenzierungstechniken am Beispiel der Genomsequenzierung im 1000-Genome Projekt dargestellt. Es wurde deutlich, dass jede Sequenzierungstechnik eine eigene Fehlersignatur zuzuordnen ist und evtl. mit angepassten Algorithmen die Fehlerrate reduziert werden kann. Als Empfehlung sollen die Rohdaten 5 Jahren archiviert werden um entsprechend mit neueren besseren Algorithmen ausgewertet werden zu können.

3.3 Archivierungsdaten und Methoden

Der Umfang der zu archivierenden Daten hängt von den Nachnutzungsszenarien ab: Die bei der Sequenzierung anfallenden Rohbilddaten werden aktuell in den meisten Zentren nicht archiviert. Die daraus via Primäranalyse generierten Sequenzen im FASTQ Format bilden im Moment bei allen Sequenzierzentren die Rohdaten für weitere Analysen und stellen einen de-facto-Standard dar. Bei der Langzeitarchivierung werden die Sequenzen mit Standardtools wie ZIP komprimiert gespeichert. Die speziell für Sequenzen entwickelte Komprimierungsalgorithmen können den Speicherbedarf bei der Archivierung erheblich minimieren. Einige neue Entwicklungen und Verallgemeinerungen der entsprechenden Algorithmen in einem nachnutzbaren Framework wurden im Vortrag von Prof. Kurtz aus Hamburg erläutert.

Die Ergebnisse der sekundären Analyse sind unter Umständen eine Alterative zu den Rohdaten. Da die analysierten FASTQ-Rohdaten lassen sich effizient im z.B: BAM-Alignment Format experimentbezogen speichern. Es könnten sowohl die FASTQ Sequenzen aus dem BAM Format erneut extrahiert werden als auch das Experimentwissen und der Analyseansatz erhalten bleiben.

Als Speichermedium werden am Standort Kiel zurzeit SATA-Festplatten als Archivierungsmedium benutzt. Wegen stetig wachsendem Datenaufkommen sind Bandlösungen in der Planung. Eine entsprechende GRID-Anbindung müsste demnach eine sehr hohe Bandbreite im Zugriff und Speicherkapazität bereitstellen.

4 Ausblick

Die Weiterentwicklung der neuen Komprimierungsmethoden der biomedizinischen Genomdaten soll weiterverfolgt werden. Die zwei unterschiedlichen Archivierungsszenarien erfordern weitere detaillierte Analysen. Probleme des Datenschutzes und der Re-identifizierbarkeit von Datensätzen wurden diskutiert und hierfür gestufte Zugangs- und Authentifizierungsmodelle als Lösungsmöglichkeit sollen im der Pilotimplementierung realisiert werden.

Der Workshop hat die Bedeutung der Pilotimplementierungen unterstrichen. Eine weitere Verfeinerung wird auf dem Workshop bei der TMF im Juni erfolgen.

5 Anhang

5.1 WORKSHOP Agenda

„Forschungsdatenmanagement biomedizinischer Genomdaten“

Die Archivierung der im Rahmen von genomischen Forschungsansätzen anfallenden Datenmengen, die zum einen die Anforderungen für Nachvollziehbarkeit der Forschung gemäß den DFG-Richtlinien erfüllen und gleichzeitig sinnvolle und rationelle Nachnutzung der Daten erlauben, ist eine große und bisher ungelöste Herausforderung. Im Rahmen des Workshops soll ein Eindruck aus der Praxis über die Datenstandards, Nutzungsmuster und Fragestellungen in der nachhaltigen Nutzung genomischer Daten gewonnen und diskutiert werden.

am 27. März 2012 von 14.00 bis 18:00 Uhr in Kiel

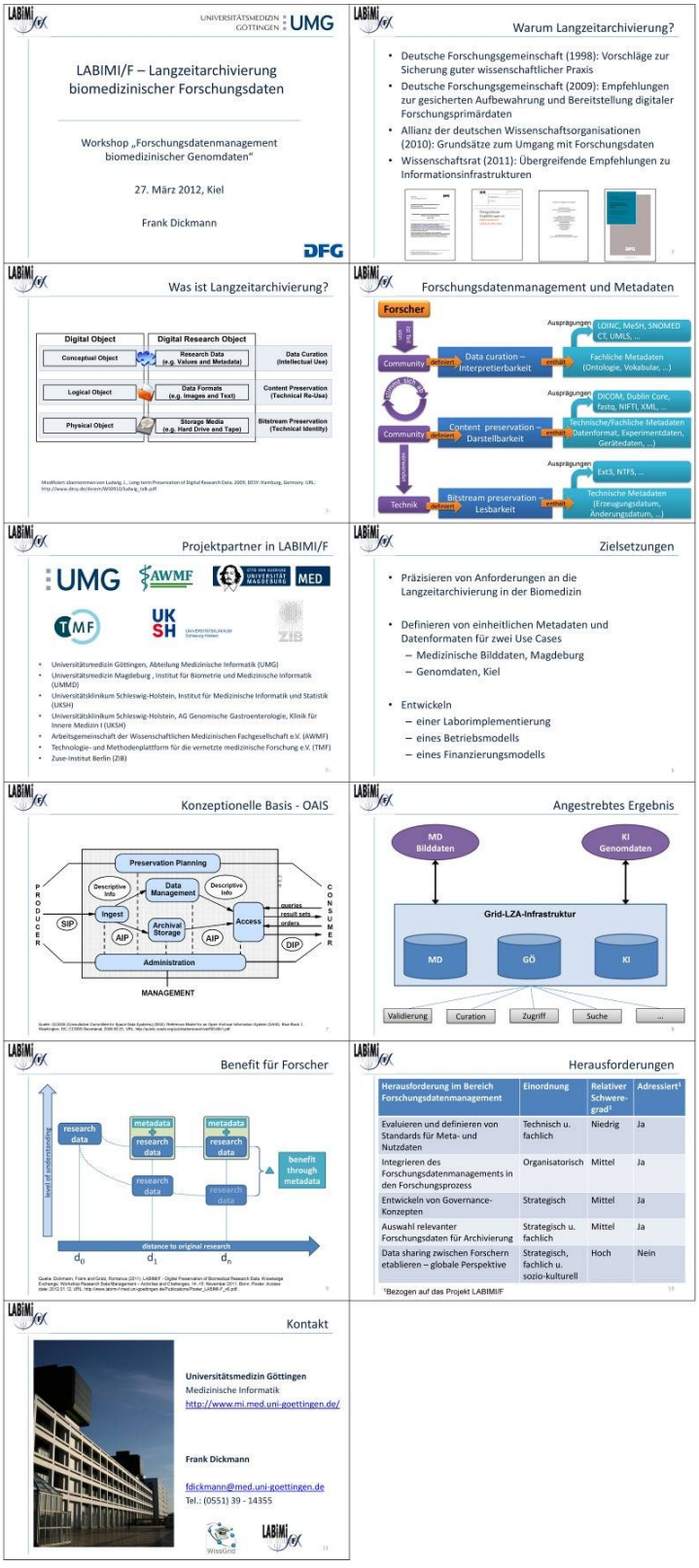
- Ärztekasino, 2. OG, Klinik für Innere Medizin I -

14.00 Uhr	Begrüßung	J. Hampe
14.15 Uhr	Gesamtvorstellung des Projektes LABIMI	F. Dickmann
14.35 Uhr	Problemstellung Datenarchivierung von Genomdaten: Sicherheit / Zugriff / Metadaten?	J. Hampe
14.45 Uhr	Praxisbeispiel: Aufklärung der Genetik von Hämochromatose und Gallensteinleiden durch „Nachnutzung“ von GWAS-Daten	S. Buch
15.15 Uhr	Praxisbeispiel: NG-Sequencingdaten – Artefaktsammlung oder Datenschatz – lohnt sich die Archivierung?	M. Nothnagel
15.45 Uhr	Kaffeepause	
16.15 Uhr	Datenstandards bei Sequenzierungsdaten	A. Herrmann
16:45 Uhr	Eine neue effiziente Datenstruktur für die Speicherung und Abfrage multipler Biosequenzen	S. Kurtz
17.00 Uhr	SNP array und Next-Generation Sequenzierungs Daten; Qualität und Analysestrategien	M. Wittig
17.30 Uhr	Diskussion und Lösungsstrategien	F. Dickmann / J. Hampe / M. Krawczak

5.2 Vorträge

5.2.1 Vortrag - Dickmann

5.2.1.1 Folien

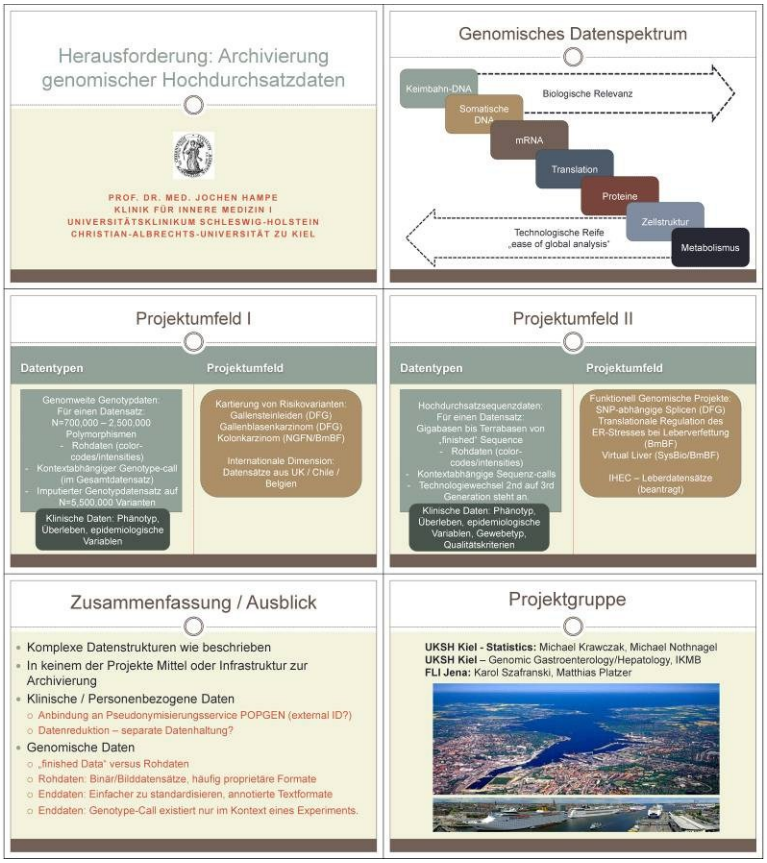


5.2.1.2 Diskussion zu den Folien

F: Klinische Daten usw. werden ausgeblendet?!
A: Ja, wegen der Größe des DFG-Projektes, klarer Rahmen, Projekt ist allerdings erweiterbar.
F: Ist im Hinblick auf das world data center (auch zitationsfähig) so etwas bereits vorhanden?
A: Genomexpresion Omnibus, webbasierte Lösung, allerdings keine übergreifende Interaktionen oder Möglichkeit Daten im Grid direkt an Programme zu schicken.
F: Wieso Grid-Technologie? Finanzierbarkeit / Nachhaltigkeit?
A: Verweis auf die Entwicklung von Governance Konzepte, Strategie siehe Workshop im Juni, Ergebniskonferenz ergab: keine weiteren Grid-Förderungen, Infrastruktur bleibt bestehen (Rechenzentrum), aber die Ressourcen müssen nach dem Projekt beschafft werden.
F: Kann ich mir aussuchen wo ich Daten speichere oder geschieht dies irgendwo?
A: XtreemFS, abhängig von der Laborimplementierung, auch lokale Silos sind möglich, ggf. RAID-Konzept auf die unterschiedlichen Töpfe anwendbar, Laborimplementierung richtet sich nach den Anforderungen der Forscher, OneClickIngest ist gewagt, es kommt mehr Arbeit auf die Forscher zu (bzgl. Forschungsdatenmanagement).
F: In welche Tiefe soll man dokumentieren? Forensisch / sinnvoll, was sind die Vorstellungen?
A: Wunsch: totale Überwachung, sinnvoll eher nicht, eher 80/20. Automatische Metadatenextraktion sollte aber implementiert werden. Ziel ist minimaler Aufwand (Dateneingabe) für den Forscher, vor allem ist der Erkenntnis des Problems wichtig. Keine 100% Eingabe. Wiederholbarkeit und Weiternutzen der Daten ist wichtig.

5.2.2 Vortrag – Hampe

5.2.2.1 Vortragsfolien – Hampe



5.2.2.2 Diskussion zu den Folien

F: Soll ich alles aufbewahren?

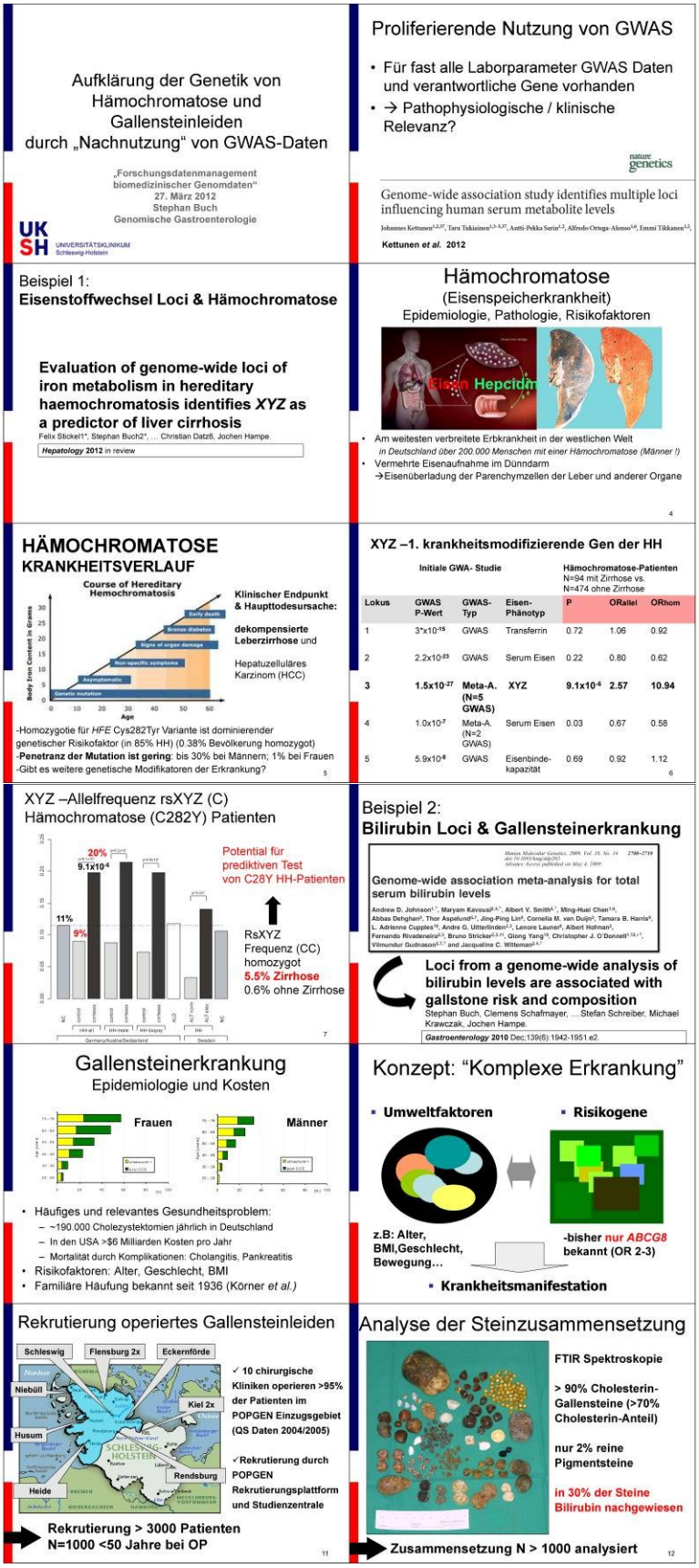
A: Bewertung in der Diskussion, ich würde Enddaten aufbewahren, Rohdaten bringen kontextabhängig Vorteile.

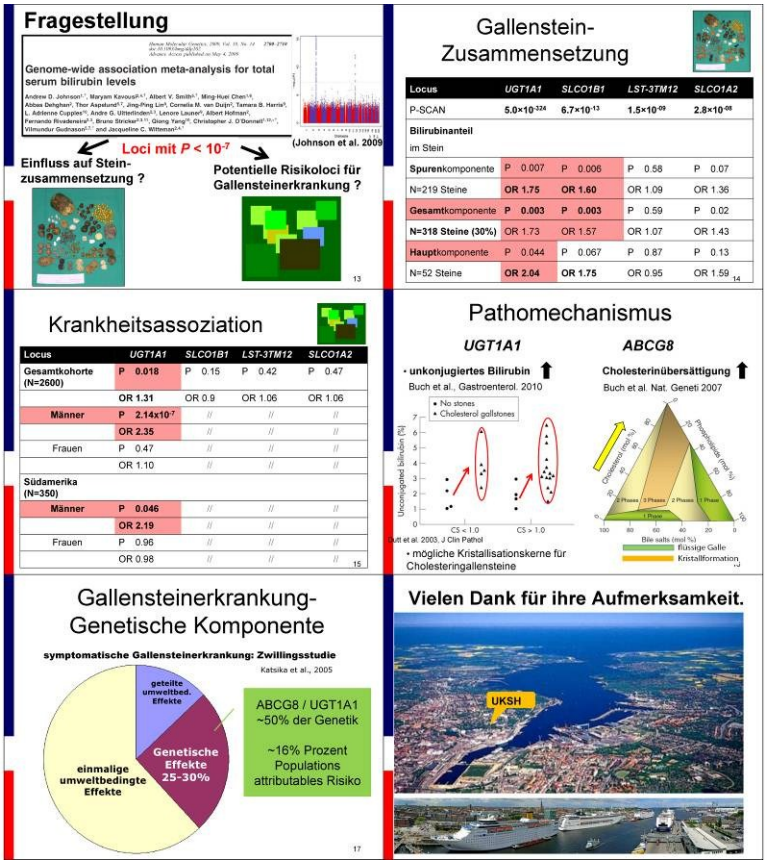
F: In dem Moment der Generierung kennt man den Wert ggf. nicht besser einfach die DNA aufheben? Mehr Informationen, kann sie aber zwischendurch nicht nutzen?

A: Differenzieren: 1xArchivieren zur Nachvollziehbarkeit um Fehler zu finden etc.; 2x Data Sharing, Nachnutzbarkeit; beides muss getrennt behandelt werden, da unterschiedliche Anforderungen: bspw. beim Datenschutz; Relevanz der Daten (für die Nachnutzbarkeit bspw. nichts/weniger Wert); Finanzierung der Infrastruktur (sollte von den Universitäten/Klinika übernommen werden) schwer über Projekte zu finanzieren.

5.2.3 Vortrag – Buch

5.2.3.1 Vortragsfolien





5.2.3.2 Diskussion zu den Folien

F: Viele Daten wurden verarbeitet, wurden diese veröffentlicht (Nachvollziehbarkeit / Excel Tabellen)?

A: Als Supplements. Genomweite Profile, gehören diese in ein öffentliches Repository? -> Rückschlüsse / Re-Identifizierung möglich bei vielen Daten?! Daher eingestellt. Bei NIH wird dies jedoch gemacht, weitere Gegenargumente bspw. Aufklärung der Patienten. Daten -> Vor-/Nachteile Public Domain (vs. Nachnutzen).
Problem: Qualitätssicherung der Daten, wie kann diese gewährleistet werden?

F: Journals verlangen wegen Transparenz die Daten, Unterscheidung public access / managed access, access Comites, wie lässt sich das klären?

A: Unterscheidung, worum es geht! Archivierung zur Prüfung (siehe oben)
Delegation an Access Comites in den USA ist ein großer Unterschied zur Speicherung hier, man muss dies den Patienten erklären!

F: Wie groß wären die Daten gewesen?

A: Beispiel für Nachnutzung von Daten, geringe Datenmengen, die Daten sind in der Public Domain verfügbar. Die meisten Studien veröffentlichen nur die Top Luci
Gute Annotation zur Nachnutzung (ab call)

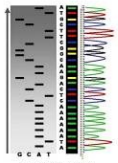
5.2.4 Vortrag – Nothnagel

5.2.4.1 Vortragsfolien

Archivierung von NGS-Daten – Artefaktsammlung oder Datenschatz?

Michael Nothnagel
Christian-Albrechts-Universität zu Kiel

Next-Generation Sequencing (NGS)



- Nach Sanger Sequencing (1977) zweite Generation von Hochdurchsatz-Sequenzierungs-techniken
- Entwicklung seit den 1990er Jahren
- Seit 2005 verbreitete Anwendung, wiederholte Veränderungen und Verbesserungen
- Durchbruch in der Bestimmung genomischer Sequenzen

⇒ immer noch relativ neue Technik

NGS als Hoffnungsträger

- Direkte Untersuchung kausaler Varianten; mögliche Ablösung des Ansatzes indirekter Assoziationsstudien
- Teilweise Aufklärung der ‚missing heritability' in häufigen Erkrankungen durch Identifizierung seltener genetischer Varianten
- Gezieltere Tumortherapie mit Hilfe somatischer Mutationsprofile
- Umfassendere Analyse genetischer Information, u.a. durch Quantifizierung epigenetischer Modifikationen (Epigenomics) und intermediärer Genprodukte (Transcriptomics)
- und mehr...

Neue Daten – Neue Fehler

- NGS-Daten sind fehlerbehaftet
- Natur der Fehler
  - im Vornherein häufig unbekannt (‚learning by doing', Erfahrungssammlung)
- Fehlerquellen für NGS-Daten (Auswahl):
  - Probleme beim Alignment von Reads (kurze Sequenzen teilweise unklaren genomischen Ursprungs)
  - einige Regionen sind nicht erreichbar (z.B. Pseudoautosomale Region der X/Y-Chromosomen, Repeats, Duplications etc.)
  - unterschiedlich hohe Abdeckung genomischer Regionen

Das HapMap-Referenz-Projekt

Ziel: Katalog häufiger genetischer Varianten in humanen Populationen (basierend auf Genotypisierung, Frequenz≥5%; 3.1 Mill. SNPs in Phase I+II)

Referenz für: tagSNP-Auswahl, Genotyp-Imputation, Qualitätskontrolle, etc.

The International HapMap Consortium (2003,2005,2007) Nature

Das 1000-Genome-Projekt

Ziel: Katalog der meisten genetischen Varianten mit einer Frequenz ≥1% in den untersuchten Populationen (basierend auf Next-Gen-Sequenzierung)

Pilot 2 Projekt:
Gesamtes Genom in 2 Trio-Familien (CEU, YRI) @ 20-60x Abdeckung

The 1000 Genomes Project Consortium (2010) Nature

Übersicht über inferierte SNVs

1000-Genome-Projekt, Pilot 2, Chromosomen 1-22

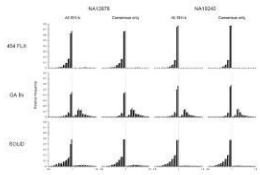
	NA12878 (CEU)		NA19240 (YRI)	
	Bekannte (Bestätigte) SNVs	Mutmaßliche SNVs	Bekannte (Bestätigte) SNVs	Mutmaßliche SNVs
454 FLX™	760,693 (724,548)	1,330,000	336,432 (319,106)	659,605
GA IIx™	821,017 (786,131)	1,126,727	892,372 (851,842)	1,816,994
SOLID™	686,686 (651,873)	1,219,584	812,710 (777,840)	1,544,714
Konsens	609,429 (587,348)	631,533	300,237 (288,818)	420,570

- SNV-Inferenz: SAMtools (Li et al. 2009) mit Standard-Optionen
- SNV-Filter: quality score ≥20, read coverage ≤100
- Consensus: SNVs, die konkordant durch alle drei Plattformen gecallt wurden

Nothnagel et al. (2011) Hum Genet

Read-Entropie pro SNV

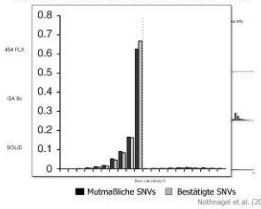
1000-Genome-Projekt, Pilot 2, Chromosomen 1-22



Nothnagel et al. (2011) Hum Genet

Read-Entropie pro SNV

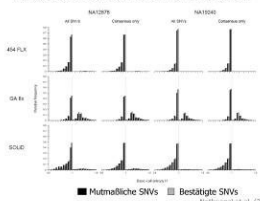
1000-Genome-Projekt, Pilot 2, Chromosomen 1-22



Nothnagel et al. (2011) Hum Genet

Read-Entropie pro SNV

1000-Genome-Projekt, Pilot 2, Chromosomen 1-22



Nothnagel et al. (2011) Hum Genet

Schätzungen des Anteils falsch inferierter SNVs

1000-Genome-Projekt, Pilot 2, Chromosomen 1-22

	NA12878 (CEU)			NA19240 (YRI)		
[%]	Alle SNVs	Konsens	P	Alle SNVs	Konsens	P
454 FLX™	6.3 (6.1-6.5)	0.7 (0.5-3.6)	$<10^{-4}$	2.9 (2.7-3.2)	2.6 (1.2-4.7)	0.08
GA IIx™	8.4 (8.0-8.7)	3.5 (3.1-3.9)	$<10^{-4}$	11.1 (10.9-11.3)	3.9 (3.5-4.3)	$<10^{-4}$
SOLID™	17.1 (16.9-17.4)	0.8 (0.1-2.6)	$<10^{-4}$	7.3 (6.8-7.8)	4.0 (3.1-4.8)	$<10^{-4}$

P-Werte aus einem Permutationstest.

Nothnagel et al. (2011) Hum Genet

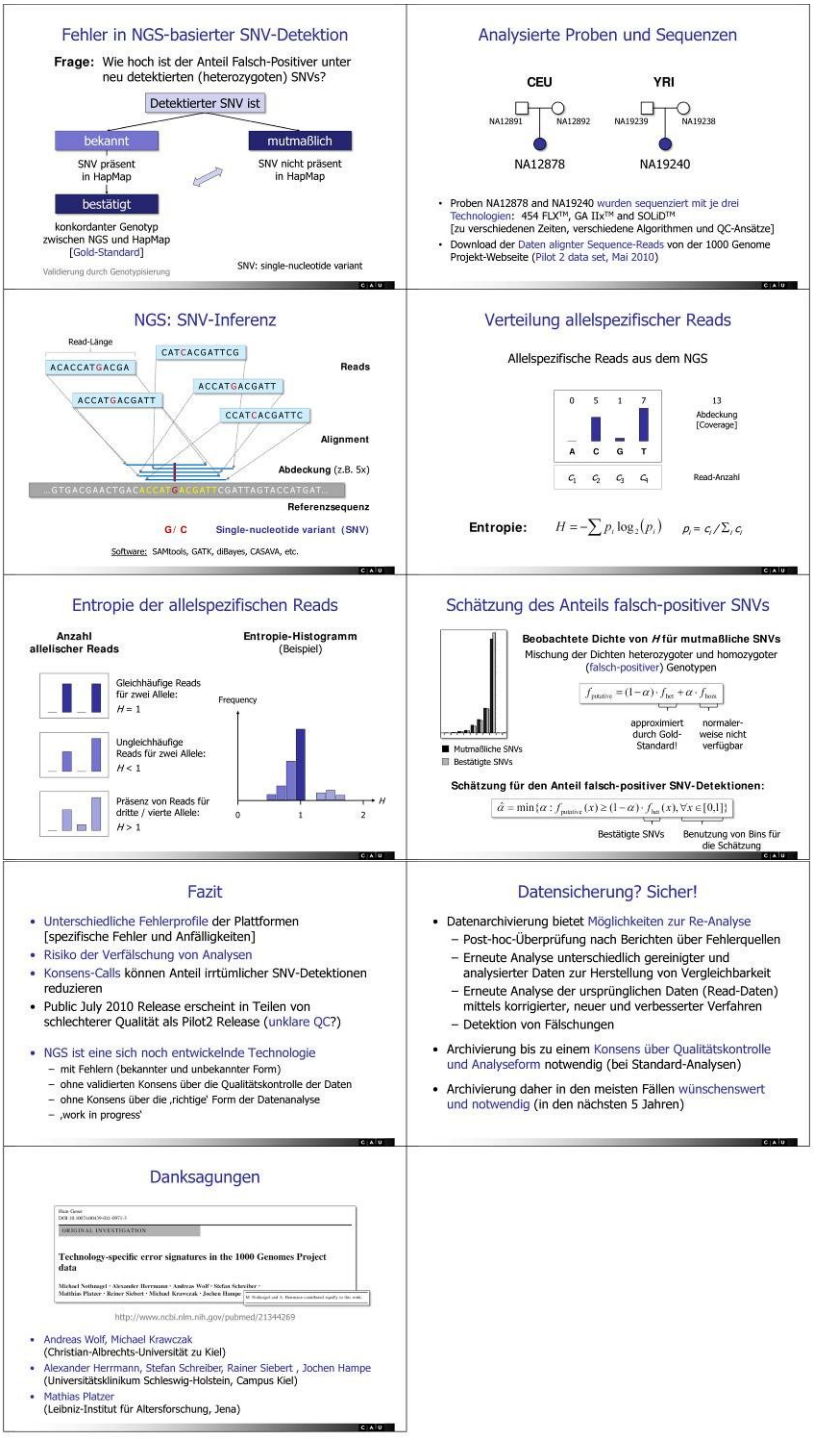
Fragen

I. Hilft Qualitätskontrolle (QC)?
- Minimaler Schwellenwert für die Abdeckung ✗
- Minimaler Schwellenwert für den Quality-Score ✓

II. Sind HapMap-Varianten ‚einfacher' zu sequenzieren? ✗
- Untersuchung möglicher Unterschiede in der flankierenden Sequenz zwischen bestätigten und mutmaßlichen SNVs

III. Sind die Ergebnisse spezifisch für den Datensatz? ✗
- Analyse der July 2010 Release des 1000-Genome-Projekts

IV. Sind die Ergebnisse spezifisch für den SNV-Calling-Algorithmus? ✗
- Analyse von SNVs, die mit einem alternativen Algorithmus gecalled wurden: GATK



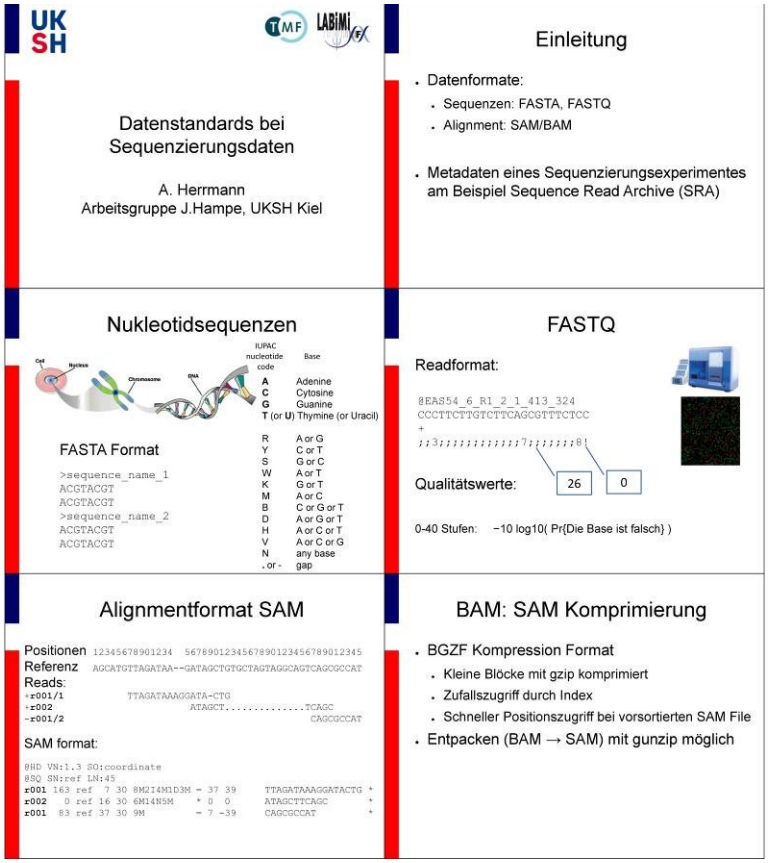
5.2.4.2 Diskussion zu den Folien

F: Lohnt es sich die Enddaten als die Rohdaten archivieren?
A: Erstmals auch Rohdaten, nach ca. 5 Jahren könnten sie gelöscht werden, dann wird es (hoffentlich) Tools / de-facto-Standards geben.
F: Qualitätsparameter von FASTQ?
A: SAM, öffentliche Daten aus dem Netz.
F: Wird immer mit 2-3 Sequenzierungstechniken gearbeitet?

A: Nein, es ist extrem teuer. Kommt aber auf das Experiment an. Wenn ich wirklich sicher gehen will sollte man 2 nehmen.
F: Optimale Sequenztiefe?
A: Sequenztiefe nicht unbedingt entscheidend, wenn systematischer Fehler drin sind hilft es nicht die Sequenztiefe zu erhöhen.
F: Wie lang ist die Analyse her? Filter haben Artefakte.
A: 1,5a, Reads sind länger geworden, vermutlich kleinere Fehlerrate -> Fehlerschätzung schon nicht mehr aktuell.
Unterstreicht die Notwendigkeit der Speicherung der Daten. Soll „nur“ Bewusstsein für diese Fehler schaffen

5.2.5 Vortrag – Herrmann

5.2.5.1 Vortragsfolien



Variant Call Format (VCF)

Sequenzvariationen

- Header mit Metadaten
- Jede Zeile - eine Position in Genome

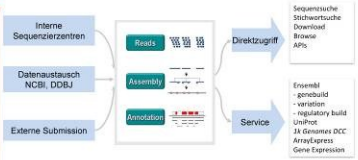
```
#CHR  POS     ID        REF ALT  QUAL FILTER INFO
20    14370   rs6054257 G   A    29   PASS   AF=0.5
20    17330   .         T   A    3    q10    AF=0.017
20    1110696 rs6040355 A   G,T  67   PASS   AF=0.333,0.667

...     FORMAT      NA00001
...     GT:GQ:DP:HQ 0|0:48:1:51,51
...     GT:GQ:DP:HQ 0|0:49:3:58,50
...     GT:GQ:DP:HQ 1|2:21:6:23,27
```

Sicht der Anwender

- Sicherung der Sequenzen:
  - FASTQ
  - Proprietäre Formate der Gerätehersteller
- Aufbewahrung
  10 Jahre / 30 Jahre ?

European Nucleotide Archive



ENA

EBI SRA

- Primäre Archive für Next-Generation Sequenzierungsdaten und Alignments (BAM)
- Veröffentlichung der Daten mit Publikation
- Sperrfrist möglich
- Freier Zugriff auf die Daten
- Neue Algorithmen für Speicherung

SRA: Wachstum

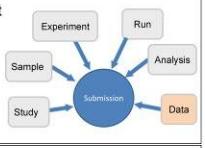
EMBL-Bank → Index
200M Sequenzen mit 600G Basen
SRA → Kein Sequenzindex
1.3 Billion Sequenzen mit 133 Billion Basen

Rasko Leinonen, EMBL-EBI

SRA Submission

http://www.ebi.ac.uk/ena/about/sra_submissions

- Datenformats
- Metadaten Objekte
- Submission Account
- Übertragung
- Anpassen
- Dokumentation

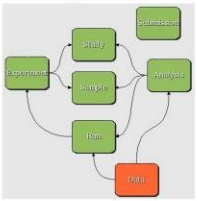


Submission Datenformat

- Bevorzugte Format **BAM**
- Andere:
  SRF, Fastq, SFF, SOLiD_native, Illumina_native, PacBio_HDF5, CompleteGenomics_native
- Minimum Information: Basen und deren Qualität
- Sequenzierungen mit Barcode demultiplexen
- Technische Sequenzen eliminieren

Metadaten Objekte

- Submission
- Study
- Sample
- Experiment
- Run
- Analysis



- Für EGA andere:
  - DAC, Policy, Dataset

Metadaten Objekte (1)

- Submission
  - Neu oder Änderungen
  - Veröffentlichungstermin
- Study
  - Projektziel mit Beschreibung
- Sample
  - Probenbeschreibung
  - Organismus

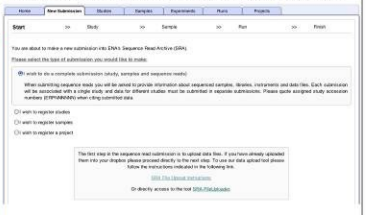
Metadaten Objekte (2)

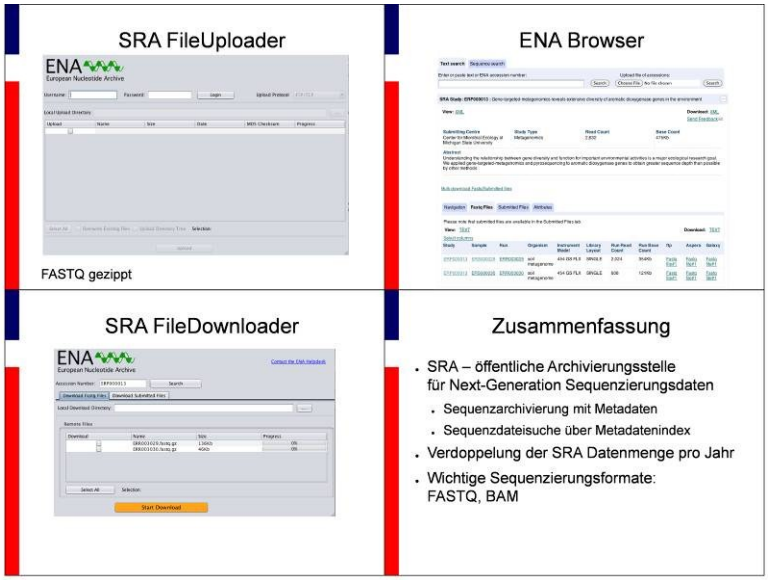
- Experiment
  - Library Information
  - Platform Information
  - Verbindet Study mit Sample und Run
- Run
  - Datenfiles mit primären Sequenzdaten
  - md5sum
- Analysis
  - BAM files mit Referenzsequenzen

Metadaten Objekte (3)

- Existierenden Study und Sample Objekte
- Eindeutiger Name für jeden Objekt
- Jeder Objekt bekommt Zugriffnummer
  - Submission: ERANNNNNN
  - Study: ERPNNNNNN
  - Sample: ERSNNNNNN
  - Experiment: ERXNNNNNN
  - Run: ERRNNNNNN
  - Analysis: ERZNNNNNN

SRA Webin: metadaten submission



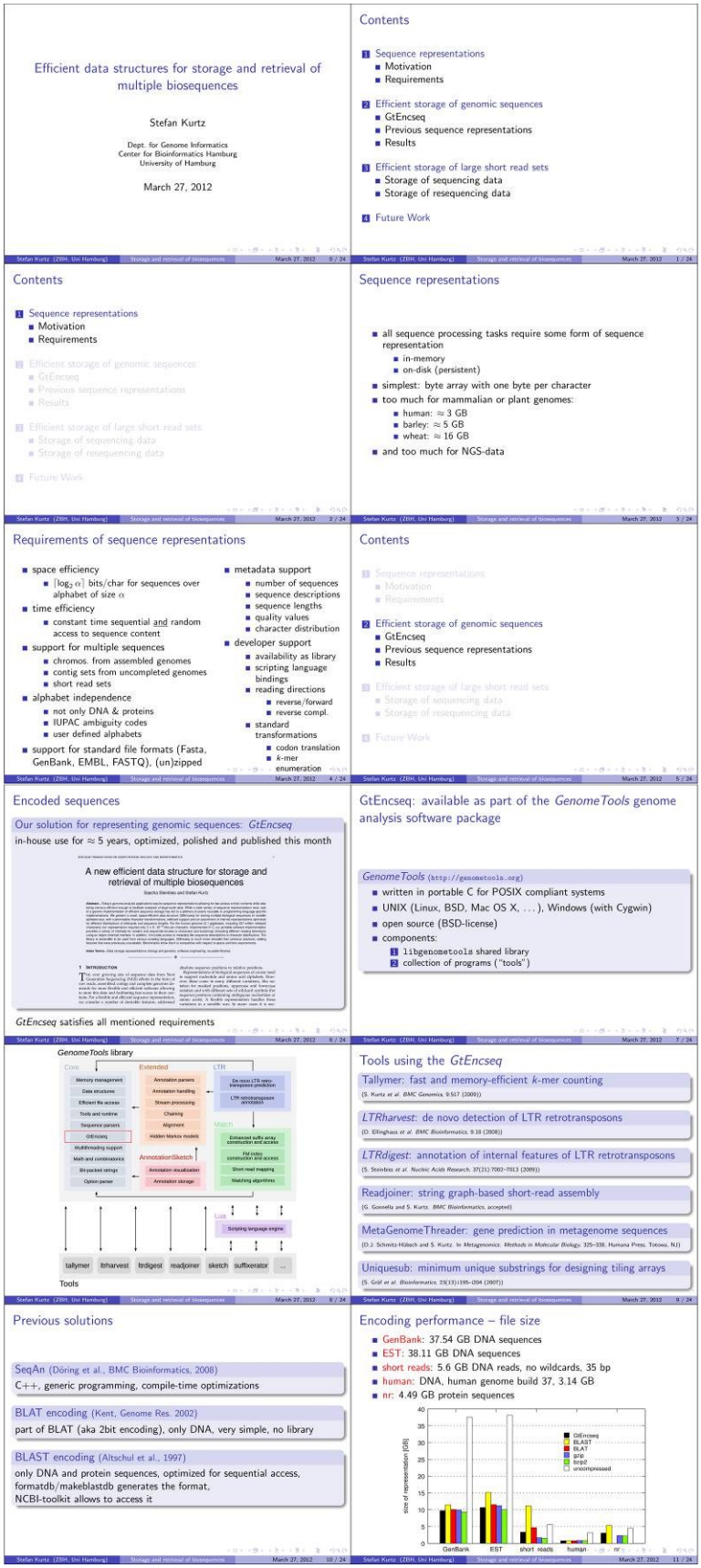


5.2.5.2 Diskussion zu den Folien

F: Datenschutz beim SRA-Archiv?
A: verwendete Patienten sind öffentlich.
F: Die zu vergebene Nummer auch als Handle?
A: Ja
F: Gibt es Journals die das Nachladen der Daten fordern?
A: Unbekannt ob es Pflicht ist, manche machen es.
F: Re-Identifizierung unmöglich?
A: Im Probenteil/Probenbeschreibung kann die Information eingetragen werden.
F: Mehrere Samples möglich?
A: Ja
F: Vorschriften für die Samples?
A: FASTQ-Datei von DNA, Kombination unterschiedlicher Geräte/Methoden muss gekennzeichnet werden. Unterschiedliche Geräte können durch mehrere Runs realisiert werden
F: Wieso nicht in EGA
A: Geschlossenes Archiv, strikter Datenschutz für Datenweitergabe. Für das Hochladen muss ein Account erstellt werden. Der Hochladende muss Datenschutz selber klären, das Archiv kümmert sich nicht drum.

5.2.6 Vortrag – Kurtz

5.2.6.1 Vortragsfolien



Efficient data structures for storage and retrieval of multiple biosequences

Stefan Kurtz

Dept. for Genome Informatics
Center for Bioinformatics Hamburg
University of Hamburg

March 27, 2012

Contents

1 Sequence representations
- Motivation
- Requirements

2 Efficient storage of genomic sequences
- GtEncseq
- Previous sequence representations
- Results

3 Efficient storage of large short read sets
- Storage of sequencing data
- Storage of resequencing data

4 Future Work

Contents

1 Sequence representations
- Motivation
- Requirements

2 Efficient storage of genomic sequences
- GtEncseq
- Previous sequence representations
- Results

3 Efficient storage of large short read sets
- Storage of sequencing data
- Storage of resequencing data

4 Future Work

Sequence representations

- all sequence processing tasks require some form of sequence representation
  - in-memory
  - on-disk (persistent)
- simplest: byte array with one byte per character
- too much for mammalian or plant genomes:
  - human: ≈ 3 GB
  - barley: ≈ 5 GB
  - wheat: ≈ 16 GB
- and too much for NGS-data

Requirements of sequence representations

- space efficiency
  - $\lceil \log_2 \alpha \rceil$ bits/char for sequences over alphabet of size α
- time efficiency
  - constant time sequential and random access to sequence content
- support for multiple sequences
  - chromos. from assembled genomes
  - contig sets from uncompleted genomes
  - short read sets
- alphabet independence
  - not only DNA & proteins
  - IUPAC ambiguity codes
  - user defined alphabets
- support for standard file formats (Fasta, GenBank, EMBL, FASTQ), (un)zipped
- metadata support
  - number of sequences
  - sequence descriptions
  - sequence lengths
  - quality values
  - character distribution
- developer support
  - availability as library
  - scripting language bindings
  - reading directions
    - reverse/forward
    - reverse compl.
  - standard transformations
    - codon translation
    - k-mer enumeration

Contents

2 Efficient storage of genomic sequences
- GtEncseq
- Previous sequence representations
- Results

Encoded sequences

Our solution for representing genomic sequences: *GtEncseq*

in-house use for ≈ 5 years, optimized, polished and published this month

A new efficient data structure for storage and retrieval of multiple biosequences

GtEncseq satisfies all mentioned requirements

GtEncseq: available as part of the *GenomeTools* genome analysis software package

GenomeTools (http://genometools.org)
- written in portable C for POSIX compliant systems
- UNIX (Linux, BSD, Mac OS X, ...), Windows (with Cygwin)
- open source (BSD-license)
- components:
  1. libgenometools shared library
  2. collection of programs ("tools")

GenomeTools library

Tools using the *GtEncseq*

Tallymer: fast and memory-efficient k-mer counting
(S. Kurtz et al. *BMC Genomics*, 9:517 (2009))

LTRharvest: de novo detection of LTR retrotransposons
(D. Ellinghaus et al. *BMC Bioinformatics*, 9:18 (2008))

LTRdigest: annotation of internal features of LTR retrotransposons
(S. Steinbiss et al. *Nucleic Acids Research*, 37(21):7002–7013 (2009))

Readjoiner: string graph-based short-read assembly
(G. Gonnella and S. Kurtz. *BMC Bioinformatics*, accepted)

MetaGenomeThreader: gene prediction in metagenome sequences
(D.J. Schmitz-Hübsch and S. Kurtz. In *Metagenomics. Methods in Molecular Biology*, 325–338. Humana Press, Totowa, NJ)

Uniquesub: minimum unique substrings for designing tiling arrays
(S. Gräf et al. *Bioinformatics*, 23(13):1595–1204 (2007))

Previous solutions

SeqAn (Döring et al., BMC Bioinformatics, 2008)
C++, generic programming, compile-time optimizations

BLAT encoding (Kent, Genome Res. 2002)
part of BLAT (aka 2bit encoding), only DNA, very simple, no library

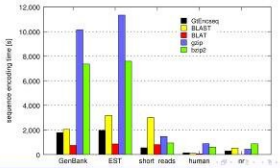
BLAST encoding (Altschul et al., 1997)
only DNA and protein sequences, optimized for sequential access, formatdb/makeblastdb generates the format, NCBI-toolkit allows to access it

Encoding performance – file size

- GenBank: 37.54 GB DNA sequences
- EST: 38.11 GB DNA sequences
- short reads: 5.6 GB DNA reads, no wildcards, 35 bp
- human: DNA, human genome build 37, 3.14 GB
- nr: 4.49 GB protein sequences

Encoding performance – encoding time

- GenBank: 37.54 GB DNA sequences
- EST: 38.11 GB DNA sequences
- short reads: 5.6 GB DNA reads, no wildcards, 35 bp
- human: DNA, human genome build 37, 3.14 GB
- nr: 4.49 GB protein sequences



GtEncseq access performance

Benchmarking scenario:

(1) Extraction of all exonic sequences from the human genome

(2) 10^6 random accesses to single bases

	GtEncseq	SeqAn		BLAT enc.	BLAT enc.
Version	1.3.0	1.2		v34	6.1
Implementation language	C	C++ packed	C++ external	C	C
Sequence loading (s)	0.003	126.1	207.3	0.001	0.005
Extraction					
(1) Exonic sequences (s)	6.5	5.2	307.1	5.3	362,044
(2) Random access (s)	0.3	0.2	0.4	739.0	seg fault
Memory peak (MB)	737	954	365	3,184	1,456

Poor behavior is shown in red

Conclusion for this part:

- Space and time requirements: *GtEncseq* is competitive with the best tool in each category
- *GtEncseq* is by far the most versatile and complete solution

Contents

Data types

Sequencing data

- newly sequenced reads obtained using NGS technology
- often given in FASTQ format (P. Cock *et al. Nucleic Acids Res.* 38(6):1767–71 (2009))
  - descriptions
  - sequences
  - quality values (e.g. encoded error probabilities)

Resequencing data

- short reads mapped to an established reference sequence
  ⇒ essentially set of alignments
- often given in BAM/SAM format (H. Li *et al. Bioinformatics*, 25:2078–9 (2009))
  - reference position
  - read number
  - alignment edit operations

Intention

Disclaimer

The methods of this part are not completely new

Intention

Our work aims at providing...

1. ...an integrated solution for sequence storage and access
   - without qualities (⇒ *GtEncseq*)
   - with qualities (ongoing work)
2. ...a unified interface
3. ...a set of reusable code modules for integrating previously isolated methods

Sequencing data – FASTQ input format

```
@read.1 length=26
GAAACATTACCAGTTCTGTTTCATTT
+
IIIIIIIIIII<IRDIIIIIEF--I
@read.2 length=26
TACAGATGACCAGTTAAGGGCAATCT
+
@BBABABBABB8:!B:DBB!!!###!
```

Selection of specific encoding techniques

- description lines are highly repetitive
  - increasing numbers, equal strings, ...
  - analyze structure and apply most appropriate encoding scheme (S. Deorowicz *et al. Bioinformatics*, 27(6):860–2 (2011))
- sequence/quality pairs have characteristic occurrence frequencies
  - employ encoding scheme based on statistical measures (W. Tembe *et al. Bioinformatics*, 26(6):2192–94 (2010))

Sequencing data – Decompression

- sequential access to whole read set
  - decode read set file from the start
  - fast, but inefficient for retrieval of single arbitrary reads
- random read access requires sampling
  - for every dth read store the starting position of its encoding
  - d is *sampling distance* providing time/space tradeoff
  - small sampling distance ⇒ fast random access, but large space requirement

Sequencing data – Results

SRR029844.1.filt, 1000 genomes proj, 5.7×10^7 reads, 76 bp, 1.14 GB

sampling distance	bits/pair	bits/desc	compr. ratio	compression speed	extraction time for 10^5 reads
∞	5.53	0.55	3.81	10.37 MB/s	78394 s
100000	5.53	0.56	3.80	10.55 MB/s	2080 s
10000	5.57	0.61	3.75	10.46 MB/s	221 s
1000	5.96	1.10	3.33	10.55 MB/s	24 s
100	9.85	6.04	1.56	10.27 MB/s	7 s

Storage of resequencing data

Basic approach

- store differences between reads and reference
- compress data by encoding edit operations (M.H.-Y. Fritz *et al., Genome Res* 21(5):734–40 (2011))

Input

- reference sequence
- sorted alignments of reads to reference ("mapped reads")
  - SAM (Sequence Alignment/Map, tab-delimited text file) or
  - BAM (binary version of SAM)

Output

- compressed representation allowing to extract all mapped reads (including quality values) given the reference sequence

Resequencing data – Example

```
ACGATCTTAATGCCTTACTTGTT--GG-CATTC     reference
  ATCGTAAT---TTAC
      ATGCCTTACTTGT                   mapped reads
           ACTTGTTATGGCC
```

position	mismatches	insertions	deletions
3	3: T → G	-	4; 3
6	-	-	-
7	-	7: AT, 3: C	-

Resequencing data – Encoding results

Options for quality storage

- none/of variations only/all qualities

Encoding results

- 6.4 GB Illumina reads from 1000 Genomes project mapped to human chromosome 20 reference
- reference stored as *GtEncseq*

preserved information	bits/base	compression speed	decompression speed
sequence, strand	0.63	13.58 MB/s	32.60 MB/s
sequence, strand, qualities of variations	1.16	13.25 MB/s	29.01 MB/s
sequence, strand, all qualities	5.21	11.12 MB/s	15.42 MB/s

Future Work

GtEncseq – storage of sequence variants

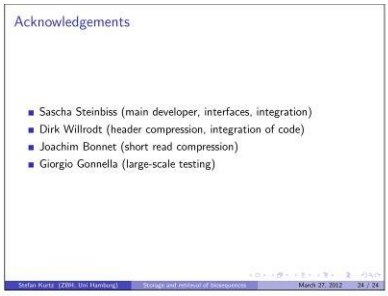
- nonredundantly store a set of similar genomic sequences
  - e.g. sets of individual genomes, strains, ...
  - access via virtual concatenation

GtEncseq – scripting language bindings

- improve efficiency of scripting language bindings
- introduce proper Perl bindings (*ctypes*)

GenomeTools integration

- unified object-oriented interface and command-line tools available for *GtEncseq* and short read processing
  - creation, loading, access

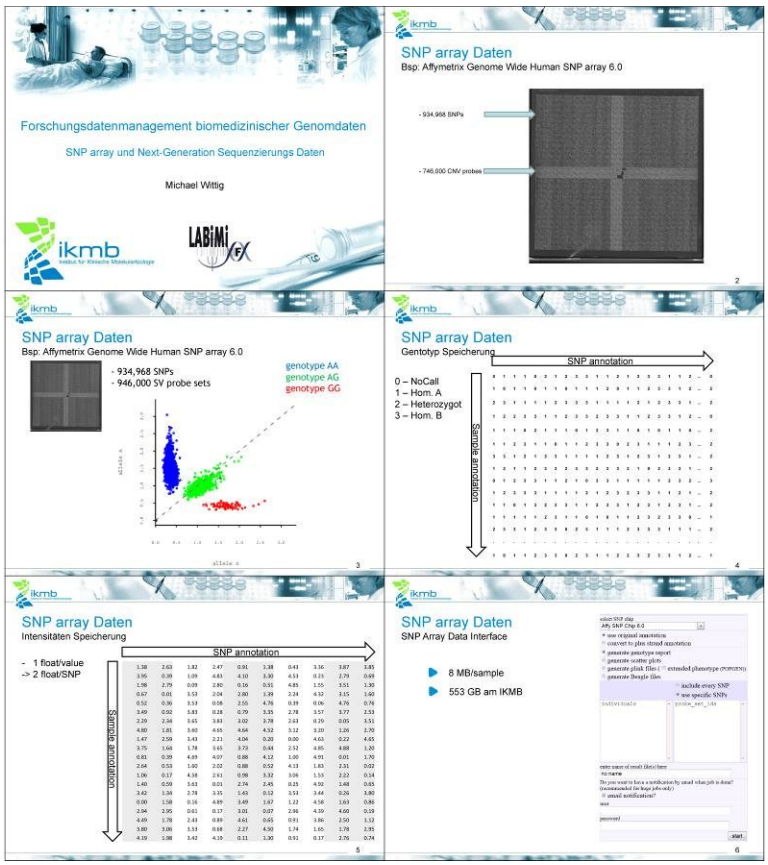


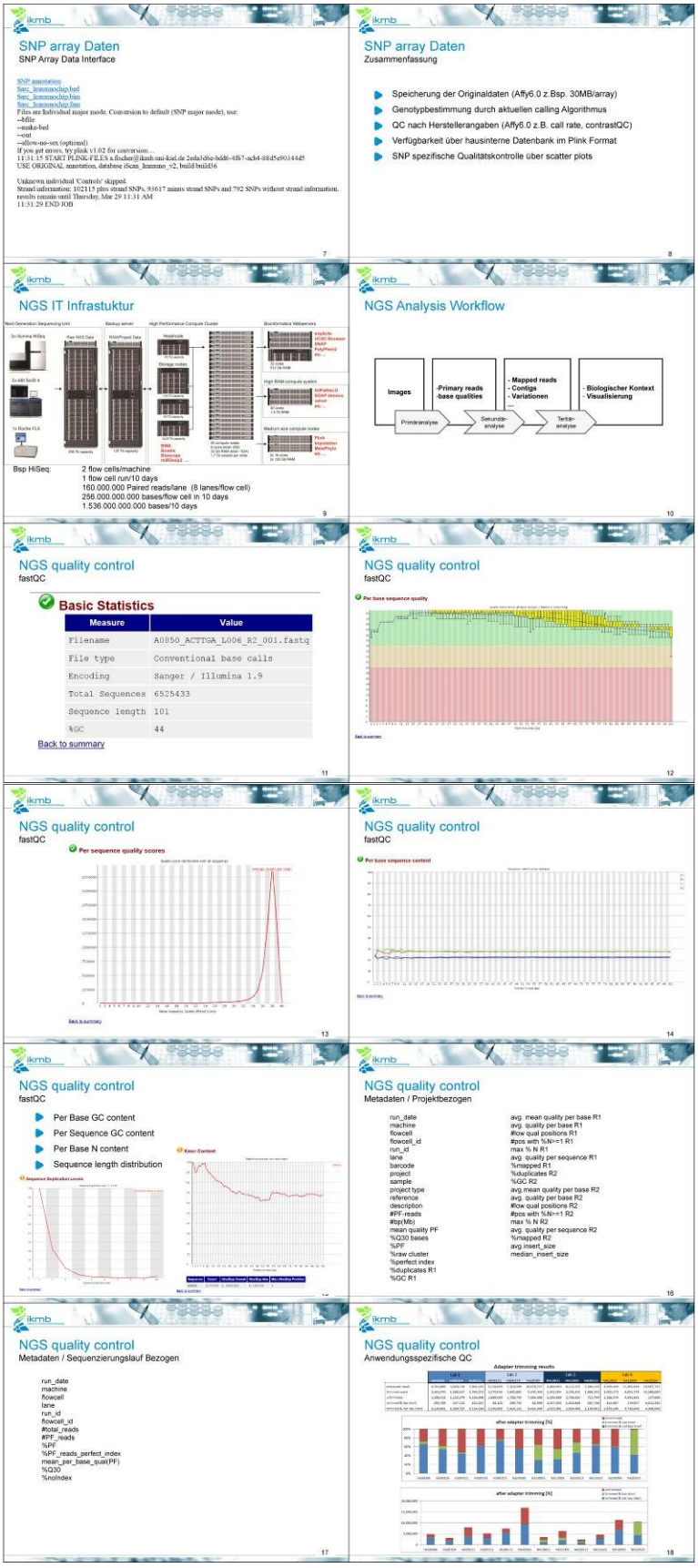
5.2.6.2 Diskussion zu den Folien

F: Genom-Browser, kennen Sie den GloveGenomeBrowser?
A: Nein.
F: GtencSec-Aufbau?
A: Format ist binär inkl. Metadaten, kann via C / Perl darauf zugegriffen werden.
F: Toolgebunden?
A: Ist offen, theoretisch können eigene Entwicklungen getätigt werden.
F: Abhängigkeit zu dem entwickelten Tool / nachhaltige Finanzierung?
A: Master/Doktor-Arbeiten / Lehrstuhlfinanziert, keine Projektförderung
Entwicklung nachhaltiger Lösung: keine Projektmittel.

5.2.7 Vortrag – Wittig

5.2.7.1 Vortragsfolien





SNP array Daten
SNP Array Data Interface

SNP annotation
Snrc_ilummoochip.bed
Snrc_ilummoochip.bim
Snrc_ilummoochip.fam
Files are fodividual major mode. Conversion to default (SNP major mode), use:
--bfile
--make-bed
--out
--allow-no-sex (optional)
If you get errors, try plink v1.02 for conversion ...
11:31:15 START PLINK-FILES a.fischer@ikmb.uni-kiel.de 2eda1d6e-bdd6-4f87-acb4-88d5e9034445
USE ORIGINAL annotation, database iScan_Ilummo_v2, build build36

Unknown individual 'Controls' skipped.
Strand information: 102115 ilmo strand SNPs, 83617 minus strand SNPs and 792 SNPs without strand information.
results remain until Thursday, Mar 29 11:31 AM
11:31:29 END JOB

7

SNP array Daten
Zusammenfassung

- Speicherung der Originaldaten (Affy6.0 z.Bsp. 30MB/array)
- Genotypbestimmung durch aktuellen calling Algorithmus
- QC nach Herstellerangaben (Affy6.0 z.B. call rate, contrastQC)
- Verfügbarkeit über hausinterne Datenbank im Plink Format
- SNP spezifische Qualitätskontrolle über scatter plots

8

NGS IT Infrastuktur

Bsp HiSeq:
2 flow cells/machine
1 flow cell run/10 days
160.000.000 Paired reads/lane (8 lanes/flow cell)
256.000.000.000 bases/flow cell in 10 days
1.536.000.000.000 bases/10 days

9

NGS Analysis Workflow

10

NGS quality control
fastQC

Basic Statistics

Measure	Value
Filename	A0850_ACTTGA_L006_R2_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6525433
Sequence length	101
%GC	44

Back to summary

11

NGS quality control
fastQC

Per base sequence quality

12

NGS quality control
fastQC

Per sequence quality scores

13

NGS quality control
fastQC

Per base sequence content

14

NGS quality control
fastQC

- Per Base GC content
- Per Sequence GC content
- Per Base N content
- Sequence length distribution

Kmer Content
Sequence Duplication Levels

15

NGS quality control
Metadaten / Projektbezogen

run_date
machine
flowcell
flowcell_id
run_id
lane
barcode
project
sample
project type
reference
description
#PF-reads
#bp(Mb)
mean quality PF
%Q30 bases
%PF
%raw cluster
%perfect index
%duplicates R1
%GC R1
avg. mean quality per base R1
avg. quality per base R1
#low qual positions R1
#pos with %N>=1 R1
max % N R1
avg. quality per sequence R1
%mapped R1
%duplicates R2
%GC R2
avg.mean quality per base R2
avg. quality per base R2
#low qual positions R2
#pos with %N>=1 R2
max % N R2
avg. quality per sequence R2
%mapped R2
avg.insert_size
median_insert_size

16

NGS quality control
Metadaten / Sequenzierungslauf Bezogen

run_date
machine
flowcell
lane
run_id
flowcell_id
#total_reads
#PF_reads
%PF
%PF_reads_perfect_index
mean_per_base_qual(PF)
%Q30
%noIndex

17

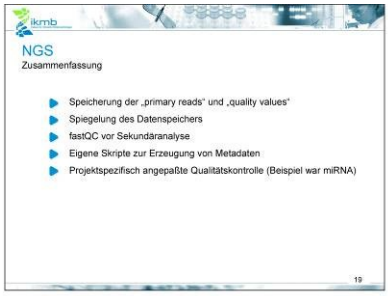
NGS quality control
Anwendungsspezifische QC

Adapter trimming results

after adapter trimming [%]

after adapter trimming [#]

18



5.2.7.2 Diskussion zu den Folien

F: Scheduler auf Cluster?
A: PBS.
F: Eigene Programme auf Cluster?
A: Alles was im User ist läuft auf den Knoten, manches via Rechenzentrum.
F: F: Eigene Forschungseinrichtung oder Service-Enrichtung?
A: Institut allerdings auch Auftragsarbeit, keine Analyse.
F: Permanente Geräteauslastung?
A: Im Prinzip schon.
F: Wie lang reicht der Speicher, wie lang werden die Daten aufgehoben?
A: 75% voll, es wird knapp -> Bandlösung.
F: Auftragsarbeit auch langzeitarchiviert?
A: Reines Service-Institut nicht, Nutzung Externer möglich, Vorteil wenn man selbst an den Systemen ist.