

Alexander Herrmann, Jochen Hampe

# Workshop Genomdaten <sup>1</sup>

	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

<sup>1</sup>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**

<b>Version</b>	<b>Datum</b>	<b>Name</b>	<b>Kurzbeschreibung</b>
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 .....	5
2	Material und Methoden .....	6
3	Ergebnisse und Diskussion .....	7
3.1	Langzeitarchivierung: Beweisfunktion vs. Sharing-Option .....	7
3.2	Qualitätsunterschiede der Sequenzierungstechniken .....	7
3.3	Archivierungsdaten und methoden .....	7
4	Ausblick.....	9
5	Anhang.....	10
5.1	WORKSHOP Agenda .....	10
5.2	Vorträge .....	11

## 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 Alternative 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



### 5.2.1.2 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: Genomexpression 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 Konzepten, 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?

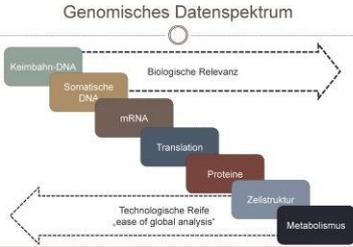
A: XtreamFS, abhängig von der Laborimplementierung, auch lokale Silos sind möglich, ggf. RAID-Konzept auf die unterschiedlichen Töpfe anwendbar, Laborimplementierung richtet sich nach den Anforderungen der Forscher, OneClickIngest ist gewagt, es kommt mehr Arbeit auf die Forscher zu (bzgl. Forschungsdatenmanagement).

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

<p>Herausforderung: Archivierung genomischer Hochdurchsatzdaten</p>  <p>PROF. DR. MED. JOCHEN HAMPE KLINIK FÜR INNERE MEDIZIN I UNIVERSITÄTSKLINIKUM SCHLESWIG-HOLSTEIN CHRISTIAN-ALBRECHTS-UNIVERSITÄT ZU KIEL</p>	<p>Genomisches Datenspektrum</p> 								
<p>Projektumfeld I</p> <table border="1"> <thead> <tr> <th>Datentypen</th> <th>Projektumfeld</th> </tr> </thead> <tbody> <tr> <td>           Genomweite Genotypdaten:            Für einen Datensatz:            N=700.000 – 2.500.000            - Rohdaten (color-codes/intensities)            - Kontextabhängiger Genotype-call (im Gesamtdatensatz)            - Imputierter Genotypdatensatz auf N=5.000.000 Varianten             Klinische Daten: Phänotyp, Überleben, epidemiologische Variablen         </td> <td>           Kartierung von Risikovarianten:            Gallensteinerkrankungen (DFG)            Gallenblasenkarzinom (DFG)            Kolonkarzinom (NGFN/BmBF)             Internationale Dimension            Datensätze aus UK / Chile / Belgien         </td> </tr> </tbody> </table>	Datentypen	Projektumfeld	Genomweite Genotypdaten: Für einen Datensatz: N=700.000 – 2.500.000 - Rohdaten (color-codes/intensities) - Kontextabhängiger Genotype-call (im Gesamtdatensatz) - Imputierter Genotypdatensatz auf N=5.000.000 Varianten  Klinische Daten: Phänotyp, Überleben, epidemiologische Variablen	Kartierung von Risikovarianten: Gallensteinerkrankungen (DFG) Gallenblasenkarzinom (DFG) Kolonkarzinom (NGFN/BmBF)  Internationale Dimension Datensätze aus UK / Chile / Belgien	<p>Projektumfeld II</p> <table border="1"> <thead> <tr> <th>Datentypen</th> <th>Projektumfeld</th> </tr> </thead> <tbody> <tr> <td>           Hochdurchsatzsequenzdaten:            Für einen Datensatz:            Gigabasen bis Terribasen von „finished“ Sequence            - Rohdaten (color-codes/intensities)            - Kontextabhängige Sequenz-calls            - Technologiewechsel 2nd auf 3rd Generation steht an             Klinische Daten: Phänotyp, Überleben, epidemiologische Variablen, Gewebetyp, Qualitätskriterien         </td> <td>           Funktionell Genomische Projekte:            SNP-abhängige Splicing (DFG)            Translationale Regulation des ER Stressnetz bei Leberentzündung (BmBF)            Virtual Liver (SysBio/BmBF)             IHEC – Leberdatensätze (beantragt)         </td> </tr> </tbody> </table>	Datentypen	Projektumfeld	Hochdurchsatzsequenzdaten: Für einen Datensatz: Gigabasen bis Terribasen von „finished“ Sequence - Rohdaten (color-codes/intensities) - Kontextabhängige Sequenz-calls - Technologiewechsel 2nd auf 3rd Generation steht an  Klinische Daten: Phänotyp, Überleben, epidemiologische Variablen, Gewebetyp, Qualitätskriterien	Funktionell Genomische Projekte: SNP-abhängige Splicing (DFG) Translationale Regulation des ER Stressnetz bei Leberentzündung (BmBF) Virtual Liver (SysBio/BmBF)  IHEC – Leberdatensätze (beantragt)
Datentypen	Projektumfeld								
Genomweite Genotypdaten: Für einen Datensatz: N=700.000 – 2.500.000 - Rohdaten (color-codes/intensities) - Kontextabhängiger Genotype-call (im Gesamtdatensatz) - Imputierter Genotypdatensatz auf N=5.000.000 Varianten  Klinische Daten: Phänotyp, Überleben, epidemiologische Variablen	Kartierung von Risikovarianten: Gallensteinerkrankungen (DFG) Gallenblasenkarzinom (DFG) Kolonkarzinom (NGFN/BmBF)  Internationale Dimension Datensätze aus UK / Chile / Belgien								
Datentypen	Projektumfeld								
Hochdurchsatzsequenzdaten: Für einen Datensatz: Gigabasen bis Terribasen von „finished“ Sequence - Rohdaten (color-codes/intensities) - Kontextabhängige Sequenz-calls - Technologiewechsel 2nd auf 3rd Generation steht an  Klinische Daten: Phänotyp, Überleben, epidemiologische Variablen, Gewebetyp, Qualitätskriterien	Funktionell Genomische Projekte: SNP-abhängige Splicing (DFG) Translationale Regulation des ER Stressnetz bei Leberentzündung (BmBF) Virtual Liver (SysBio/BmBF)  IHEC – Leberdatensätze (beantragt)								
<p>Zusammenfassung / Ausblick</p> <ul style="list-style-type: none"> <li>• Komplexe Datenstrukturen wie beschrieben</li> <li>• In keinem der Projekte Mittel oder Infrastruktur zur Archivierung</li> <li>• Klinische / Personenbezogene Daten             <ul style="list-style-type: none"> <li>○ Anbindung an Pseudonymisierungsservice POPGEN (external ID?)</li> <li>○ Datenreduktion – separate Datenhaltung?</li> </ul> </li> <li>• Genomische Daten             <ul style="list-style-type: none"> <li>○ „finished Data“ versus Rohdaten</li> <li>○ Rohdaten: Binär/Bilddatensätze, häufig proprietäre Formate</li> <li>○ Enddaten: Einfacher zu standardisieren, annotierte Textformate</li> <li>○ Enddaten: Genotype-Call existiert nur im Kontext eines Experiments.</li> </ul> </li> </ul>	<p>Projektgruppe</p> <p>UKSH Kiel - Statistics: Michael Krawczak, Michael Nothnagel          UKSH Kiel – Genomic Gastroenterology/Hepatology, IKMB          FLI Jena: Karol Szafranski, Matthias Platzer</p> 								

#### 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

#### Aufklärung der Genetik von Hämochromatose und Gallensteinleiden durch „Nachnutzung“ von GWAS-Daten

„Forschungsdatenmanagement biomedizinischer Genomdaten“  
27. März 2012  
Stephan Buch  
Genomische Gastroenterologie

**UK SH**  
UNIVERSITÄTSKLINIKUM  
Schleswig-Holstein

#### Proliferierende Nutzung von GWAS

- Für fast alle Laborparameter GWAS Daten und verantwortliche Gene vorhanden
- Pathophysiologische / klinische Relevanz?

*nature genetics*

Genome-wide association study identifies multiple loci influencing human serum metabolite levels  
Johannes Kettunen<sup>1,2,3\*</sup>, Taru Tikkanen<sup>1,4,5,7\*</sup>, Antti-Pekka Sarin<sup>1,3</sup>, Alfredo Ortega-Alonso<sup>6</sup>, Emmi Tikkanen<sup>1,5</sup>, Kettunen et al. 2012

#### Beispiel 1: Eisenstoffwechsel Loci & Hämochromatose

##### Evaluation of genome-wide loci of iron metabolism in hereditary haemochromatosis identifies XYZ as a predictor of liver cirrhosis

Felix Sickel<sup>1\*</sup>, Stephan Buch<sup>2\*</sup>, ... Christian Datzl, Jochen Hampe.  
*Hepatology* 2012 in review

#### Hämochromatose (Eisenspeicherkrankheit)

##### Epidemiologie, Pathologie, Risikofaktoren

- Am weitesten verbreitete Erbkrankung in der westlichen Welt in Deutschland über 200.000 Menschen mit einer Hämochromatose (Männer!)
- Vermehrte Eisenaufnahme im Dünndarm
- Eisenüberladung der Parenchymzellen der Leber und anderer Organe

#### HÄMOCHROMATOSE KRANKHEITSVERLAUF

Klinischer Endpunkt & Haupttodesursache: dekompenzierte Leberzirrhose und Hepatozelluläres Karzinom (HCC)

- Homozygotie für HFE Cys282Tyr Variante ist dominierender genetischer Risikofaktor (in 85% HH) (0.38% Bevölkerung homozygot)  
- Penetranz der Mutation ist gering: bis 30% bei Männern; 1% bei Frauen  
- Gibt es weitere genetische Modifikatoren der Erkrankung?

#### XYZ –1. krankheitsmodifizierende Gen der HH

Lokus	Initiale GWA-Studie			Hämochromatose-Patienten N=94 mit Zirrhose vs. N=474 ohne Zirrhose		
	GWAS P-Wert	GWAS-Typ	Eisen-Phänotyp	P	ORAllel	ORhom
1	3*10 <sup>-15</sup>	GWAS	Transferrin	0.72	1.06	0.92
2	2.2x10 <sup>-23</sup>	GWAS	Serum Eisen	0.22	0.80	0.62
3	1.5x10 <sup>-27</sup>	Meta-A. (N=5 GWAS)	XYZ	9.1x10 <sup>-4</sup>	2.57	10.94
4	1.0x10 <sup>-7</sup>	Meta-A. (N=2 GWAS)	Serum Eisen	0.03	0.67	0.58
5	5.9x10 <sup>-8</sup>	GWAS	Eisenbindekapazität	0.69	0.92	1.12

#### XYZ –Allelfrequenz rsXYZ (C) Hämochromatose (C282Y) Patienten

Potential für prädiktiven Test von C28Y HH-Patienten

RsXYZ Frequenz (CC) homozygot  
5.5% Zirrhose  
0.6% ohne Zirrhose

#### Beispiel 2: Bilirubin Loci & Gallensteinerkrankung

Genome-wide association meta-analysis for total serum bilirubin levels  
Asutava B. Johansson<sup>1\*</sup>, Maryam Kavousi<sup>1\*</sup>, Albert V. Smith<sup>1\*</sup>, Ming-Hui Chen<sup>1\*</sup>, Abbas Dehghan<sup>1</sup>, Thor Aspelund<sup>1</sup>, Jing-Ping Liu<sup>1</sup>, Cornelia M. van Duijn<sup>1</sup>, Tamara B. Harris<sup>1</sup>, E. Adrienne Cupples<sup>1</sup>, Andre G. Uitterlinden<sup>1</sup>, Lemari Launer<sup>1</sup>, Albert Hofman<sup>1</sup>, Fernando Rivadeneira<sup>1</sup>, Bruno Stricker<sup>1,11</sup>, George Yang<sup>1</sup>, Christopher J. O'Donoghue<sup>1,11</sup>, Vilmundur Gudnason<sup>1,11</sup> and Jacqueline C. Witteman<sup>1,11</sup>

Loci from a genome-wide analysis of bilirubin levels are associated with gallstone risk and composition  
Stephan Buch, Clemens Schaffmayer, ... Stefan Schreiber, Michael Krawczak, Jochen Hampe  
*Gastroenterology* 2010 Dec;139(6):1942-1951.e2

#### Gallensteinerkrankung Epidemiologie und Kosten

- Häufiges und relevantes Gesundheitsproblem:
  - ~190.000 Cholezystektomien jährlich in Deutschland
  - In den USA >\$6 Milliarden Kosten pro Jahr
  - Mortalität durch Komplikationen: Cholangitis, Pankreatitis
- Risikofaktoren: Alter, Geschlecht, BMI
- Familiäre Häufung bekannt seit 1936 (Körner et al.)

#### Konzept: „Komplexe Erkrankung“

Umweltfaktoren

z.B: Alter, BMI, Geschlecht, Bewegung...

Risikogene

-bisher nur ABCG8 bekannt (OR 2-3)

Krankheitsmanifestation

#### Rekrutierung operiertes Gallensteinleiden

- 10 chirurgische Kliniken operieren >95% der Patienten im POPGEN Einzugsgebiet (QS Daten 2004/2005)
- Rekrutierung durch POPGEN Rekrutierungsplattform und Studienzentrale
- Rekrutierung > 3000 Patienten  
N=1000 <50 Jahre bei OP

#### Analyse der Steinzusammensetzung

FTIR Spektroskopie

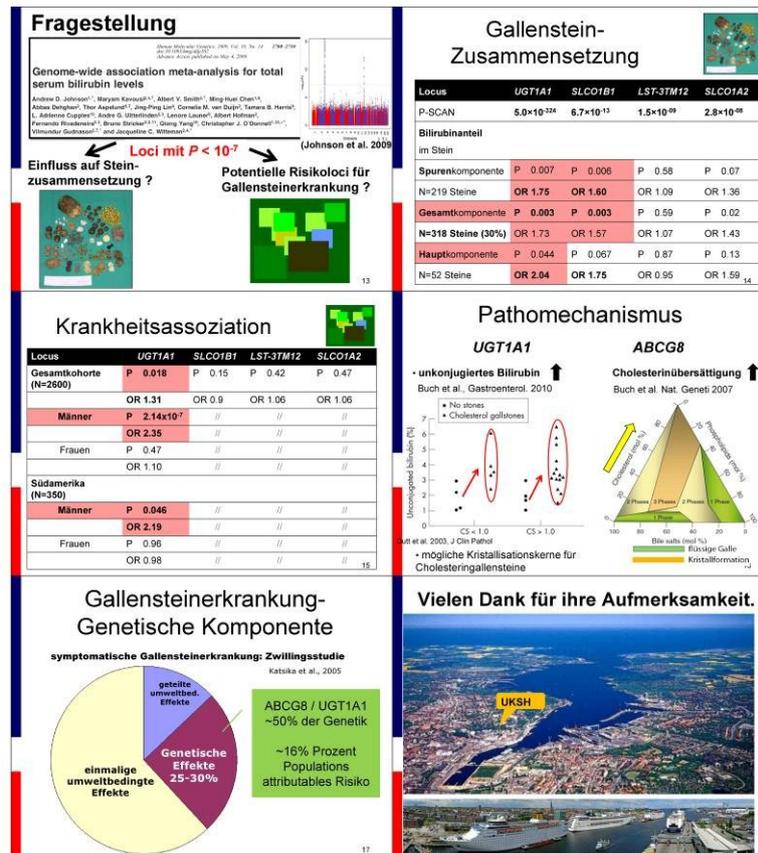
- > 90% Cholesterin-Gallensteine (>70% Cholesterin-Anteil)
- nur 2% reine Pigmentsteine
- in 30% der Steine Bilirubin nachgewiesen

Zusammensetzung N > 1000 analysiert

LABIMI/F

14

Deliverable 4.2



### 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 Comittees, wie lässt sich das klären?

A: Unterscheidung, worum es geht! Archivierung zur Prüfung (siehe oben)

Delegation an Access Committees 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 Loci  
 Gute Annotation zur Nachnutzung (ab call)

## 5.2.4 Vortrag – Nothnagel

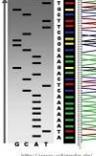
### 5.2.4.1 Vortragsfolien

#### 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-Sequenzierungstechniken
- 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 Tumorthherapie 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 Vorhinein 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.

#### 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

#### Übersicht über inferierte SNVs

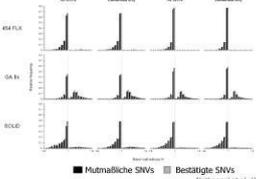
1000-Genome-Projekt, Pilot 2, Chromosomen 1-22

	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

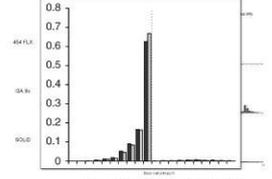
#### Read-Entropie pro SNV

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



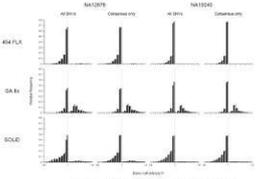
#### Read-Entropie pro SNV

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



#### Read-Entropie pro SNV

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



#### 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 <sup>-4</sup>	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 <sup>-4</sup>	11.1 (10.9-11.3)	3.9 (3.5-4.3)	<10 <sup>-4</sup>
SOLID™	17.1 (16.9-17.4)	0.8 (0.1-2.6)	<10 <sup>-4</sup>	7.3 (6.8-7.8)	4.0 (3.1-4.8)	<10 <sup>-4</sup>

P-Werte aus einem Permutationstest.

#### Fragen

- Hilft Qualitätskontrolle (QC)?
  - Minimaler Schwellenwert für die Abdeckung
  - Minimaler Schwellenwert für den Quality-Score
- Sind HapMap-Varianten ‚einfacher‘ zu sequenzieren? 
  - Untersuchung möglicher Unterschiede in der flankierenden Sequenz zwischen bestätigten und mutmaßlichen SNVs
- Sind die Ergebnisse spezifisch für den Datensatz? 
  - Analyse der July 2010 Release des 1000-Genome-Projekts
- Sind die Ergebnisse spezifisch für den SNV-Calling-Algorithmus? 
  - Analyse von SNVs, die mit einem alternativen Algorithmus gecallt wurden: GATK

<p><b>Fehler in NGS-basierter SNV-Detektion</b></p> <p><b>Frage:</b> Wie hoch ist der Anteil Falsch-Positiver unter neu detektierten (heterozygoten) SNVs?</p> <p><b>Detektiertes SNV ist</b></p> <ul style="list-style-type: none"> <li><b>bekannt</b>: SNV präsent in HapMap → <b>bestätigt</b> (konkordanter Genotyp zwischen NGS und HapMap [Gold-Standard])</li> <li><b>mutmaßlich</b>: SNV nicht präsent in HapMap</li> </ul> <p>Validierung durch Genotypisierung</p> <p>SNV: single-nucleotide variant</p>	<p><b>Analysierte Proben und Sequenzen</b></p> <p><b>CEU</b>: NA12891, NA12892, NA12878</p> <p><b>YRI</b>: NA19239, NA19238, NA19240</p> <ul style="list-style-type: none"> <li>Proben NA12878 und NA19240 wurden sequenziert mit je drei Technologien: 454 FLX™, GA IIx™ and SOLiD™ [zu verschiedenen Zeiten, verschiedene Algorithmen und QC-Ansätze]</li> <li>Download der Daten aller Sequenz-Reads von der 1000 Genome Projekt-Webseite (Pilot 2 data set, Mai 2010)</li> </ul>
<p><b>NGS: SNV-Inferenz</b></p> <p>Software: SAMtools, GATK, dBayes, CASAVA, etc.</p>	<p><b>Verteilung allelspezifischer Reads</b></p> <p>Allelspezifische Reads aus dem NGS</p> <p><b>Entropie:</b> <math>H = -\sum p_i \log_2(p_i)</math>, <math>p_i = c_i / \sum c_i</math></p>
<p><b>Entropie der allelspezifischen Reads</b></p> <p><b>Anzahl allelspezifischer Reads</b></p> <ul style="list-style-type: none"> <li>Gleichhäufige Reads für zwei Allele: <math>H = 1</math></li> <li>Ungleichhäufige Reads für zwei Allele: <math>H &lt; 1</math></li> <li>Präsenz von Reads für dritte / vierte Allele: <math>H &gt; 1</math></li> </ul> <p><b>Entropie-Histogramm (Beispiel)</b></p>	<p><b>Schätzung des Anteils falsch-positiver SNVs</b></p> <p>Beobachtete Dichte von <math>H</math> für mutmaßliche SNVs (Mischung der Dichten heterozygoter und homozygoter (falsch-positiver) Genotypen)</p> $f_{\text{mutmaßlich}} = (1-\alpha) \cdot f_{\text{het}} + \alpha \cdot f_{\text{homo}}$ <p>■ Mutmaßliche SNVs, ■ Bestätigte SNVs</p> <p><b>Schätzung für den Anteil falsch-positiver SNV-Detektionen:</b></p> $\hat{\alpha} = \min\{\alpha : f_{\text{mutmaßlich}}(x) \geq (1-\alpha) \cdot f_{\text{het}}(x), \forall x \in [0,1]\}$ <p>Bestätigte SNVs, Benutzung von Bins für die Schätzung</p>
<p><b>Fazit</b></p> <ul style="list-style-type: none"> <li>Unterschiedliche Fehlerprofile der Plattformen [spezifische Fehler und Anfälligkeiten]</li> <li>Risiko der Verfälschung von Analysen</li> <li>Konsens-Calls können Anteil irrtümlicher SNV-Detektionen reduzieren</li> <li>Public July 2010 Release erscheint in Teilen von schlechterer Qualität als Pilot2 Release (unklare QC?)</li> <li>NGS ist eine sich noch entwickelnde Technologie             <ul style="list-style-type: none"> <li>mit Fehlern (bekannter und unbekannter Form)</li> <li>ohne validierten Konsens über die Qualitätskontrolle der Daten</li> <li>ohne Konsens über die „richtige“ Form der Datenanalyse</li> <li>„work in progress“</li> </ul> </li> </ul>	<p><b>Datensicherung? Sicher!</b></p> <ul style="list-style-type: none"> <li>Datenarchivierung bietet Möglichkeiten zur Re-Analyse             <ul style="list-style-type: none"> <li>Post-hoc-Überprüfung nach Berichten über Fehlerquellen</li> <li>Erneute Analyse unterschiedlich gereinigter und analysierter Daten zur Herstellung von Vergleichbarkeit</li> <li>Erneute Analyse der ursprünglichen Daten (Read-Daten) mittels korrigierter, neuer und verbesserter Verfahren</li> <li>Detektion von Fälschungen</li> </ul> </li> <li>Archivierung bis zu einem Konsens über Qualitätskontrolle und Analyseform notwendig (bei Standard-Analysen)</li> <li>Archivierung daher in den meisten Fällen wünschenswert und notwendig (in den nächsten 5 Jahren)</li> </ul>
<p><b>Danksagungen</b></p> <p>Workshop: DNA 10. NOVEMBER 2010, 11. NOVEMBER 2010</p> <p>TECHNOLOGY-SPECIFIC ERROR SIGNATURES IN THE 1000 GENOMES PROJECT DATA</p> <p>Michael Nattagel · Alexander Herrmann · Andreas Wolf · Stefan Schreiber · Mathias Platzer · Rainer Siebert · Michael Krawczak · Jochen Hampe</p> <p><a href="https://www.ncbi.nlm.nih.gov/pubmed/21344269">https://www.ncbi.nlm.nih.gov/pubmed/21344269</a></p> <ul style="list-style-type: none"> <li>Andreas Wolf, Michael Krawczak (Christian-Albrechts-Universität zu Kiel)</li> <li>Alexander Herrmann, Stefan Schreiber, Rainer Siebert, Jochen Hampe (Universitätsklinikum Schleswig-Holstein, Campus Kiel)</li> <li>Mathias Platzer (Leibniz-Institut für Altersforschung, Jena)</li> </ul>	

**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.

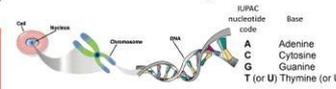
F: Wie lang ist die Analyse her? Filter haben Artefakte.

A: 1,5a, Reads sind länger geworden, vermutlich kleinere Fehlerrate -> Fehlerschätzung schon nicht mehr aktuell.

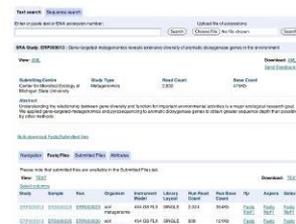
Unterstreicht die Notwendigkeit der Speicherung der Daten. Soll „nur“ Bewusstsein für diese Fehler schaffen

## 5.2.5 Vortrag – Herrmann

### 5.2.5.1 Vortragsfolien

 <p style="text-align: center;"><b>Datenstandards bei Sequenzierungsdaten</b></p> <p style="text-align: center;">A. Herrmann Arbeitsgruppe J.Hampe, UKSH Kiel</p>	<p style="text-align: center;"><b>Einleitung</b></p> <ul style="list-style-type: none"> <li>Datenformate:             <ul style="list-style-type: none"> <li>Sequenzen: FASTA, FASTQ</li> <li>Alignment: SAM/BAM</li> </ul> </li> <li>Metadaten eines Sequenzierungsexperimentes am Beispiel Sequence Read Archive (SRA)</li> </ul>																																		
<p style="text-align: center;"><b>Nukleotidsequenzen</b></p>  <p style="text-align: center;"><b>FASTA Format</b></p> <pre>&gt;sequence_name_1 ACGTACGT ACGTACGT &gt;sequence_name_2 ACGTACGT ACGTACGT</pre> <table border="1"> <thead> <tr> <th>IUPAC nucleotide code</th> <th>Base</th> </tr> </thead> <tbody> <tr><td>A</td><td>Adenine</td></tr> <tr><td>C</td><td>Cytosine</td></tr> <tr><td>G</td><td>Guanine</td></tr> <tr><td>T (or U)</td><td>Thymine (or Uracil)</td></tr> <tr><td>R</td><td>A or G</td></tr> <tr><td>Y</td><td>C or T</td></tr> <tr><td>S</td><td>G or C</td></tr> <tr><td>W</td><td>A or T</td></tr> <tr><td>K</td><td>G or T</td></tr> <tr><td>M</td><td>A or C</td></tr> <tr><td>B</td><td>C or G or T</td></tr> <tr><td>D</td><td>A or G or T</td></tr> <tr><td>H</td><td>A or C or T</td></tr> <tr><td>V</td><td>A or C or G</td></tr> <tr><td>N</td><td>any base</td></tr> <tr><td>- or -</td><td>gap</td></tr> </tbody> </table>	IUPAC nucleotide code	Base	A	Adenine	C	Cytosine	G	Guanine	T (or U)	Thymine (or Uracil)	R	A or G	Y	C or T	S	G or C	W	A or T	K	G or T	M	A or C	B	C or G or T	D	A or G or T	H	A or C or T	V	A or C or G	N	any base	- or -	gap	<p style="text-align: center;"><b>FASTQ</b></p> <p><b>Readformat:</b></p> <pre>@EAS54.6.R1.2.1.413.324 CCCTTCTGTCTTCAGCGTTTCGCC + !!!!!!!!!!!!!!!!!7!!!!!!!!!!B!</pre> <p><b>Qualitätswerte:</b> <span style="border: 1px solid black; padding: 2px;">26</span> <span style="border: 1px solid black; padding: 2px;">0</span></p> <p>0-40 Stufen: <math>-10 \log_{10}(\text{Pr}(\text{Die Base ist falsch}))</math></p>
IUPAC nucleotide code	Base																																		
A	Adenine																																		
C	Cytosine																																		
G	Guanine																																		
T (or U)	Thymine (or Uracil)																																		
R	A or G																																		
Y	C or T																																		
S	G or C																																		
W	A or T																																		
K	G or T																																		
M	A or C																																		
B	C or G or T																																		
D	A or G or T																																		
H	A or C or T																																		
V	A or C or G																																		
N	any base																																		
- or -	gap																																		
<p style="text-align: center;"><b>Alignmentformat SAM</b></p> <pre>Positionen 12345678901234 5678901234567890123456789012345 Referenz   AGCATGTTAGATAA--GATAGCTGTGCTAGTAGGGCAGTCAGCGCCAT Reads: +e001/1      TTAGATAAAGGATA-CTG +e002        ATAGCT.....TCAGC +e001/2      CAGCGCCAT</pre> <p><b>SAM format:</b></p> <pre>#HD VN:1.3 SO:coordinate #SQ SN:ref LN:45 e001 163 ref 7 30 BM214M1D3M - 37 39 TTAGATAAAGGATACTG + e002 0 ref 16 30 6M14NSM * 0 0 ATAGCTTCAGC + e001 83 ref 37 30 9M - 7 -39 CAGCGCCAT +</pre>	<p style="text-align: center;"><b>BAM: SAM Komprimierung</b></p> <ul style="list-style-type: none"> <li>BGZF Kompression Format             <ul style="list-style-type: none"> <li>Kleine Blöcke mit gzip komprimiert</li> <li>Zufallszugriff durch Index</li> <li>Schneller Positionszugriff bei vorsortierten SAM File</li> </ul> </li> <li>Entpacken (BAM → SAM) mit gunzip möglich</li> </ul>																																		

<h3>Variant Call Format (VCF)</h3> <p>Sequenzvariationen</p> <ul style="list-style-type: none"> <li>Header mit Metadaten</li> <li>Jede Zeile - eine Position in Genome</li> </ul> <pre> #CHR POS ID REF ALT QUAL FILTER INFO 20 14370 rs6054257 G A 29 PASS AF=0.5 20 17330 rs6054257 T A 3 Q10 AB=0.017 20 1110496 rs6040355 A G,T 67 PASS AF=0.333,0.667  ##FORMAT=NA00001 ##OT:QD:DP:HQ 010:48:1:51,51 ##OT:QD:DP:HQ 010:49:3:158,50 ##OT:QD:DP:HQ 312:23:1:6123,27                 </pre>	<h3>Sicht der Anwender</h3> <p>Archivierung durch Sequenzierzentren</p> <ul style="list-style-type: none"> <li>Sicherung der Sequenzen:             <ul style="list-style-type: none"> <li>FASTQ</li> <li>Proprietäre Formate der Gerätehersteller</li> </ul> </li> <li>Aufbewahrung: 10 Jahre / 30 Jahre ?</li> </ul>
<h3>European Nucleotide Archive</h3> <p>ENA</p>	<h3>EBI SRA</h3> <ul style="list-style-type: none"> <li>Primäre Archive für Next-Generation Sequenzierungsdaten und Alignments (BAM)</li> <li>Veröffentlichung der Daten mit Publikation</li> <li>Sperrfrist möglich</li> <li>Freier Zugriff auf die Daten</li> <li>Neue Algorithmen für Speicherung</li> </ul>
<h3>SRA: Wachstum</h3> <p>EMBL-Bank → Index 200M Sequenzen mit 600G Basen SRA → Kein Sequenzindex 1.3 Billion Sequenzen mit 133 Billion Basen</p> <p><small>Quelle: Commons, EMBL, EBI</small></p>	<h3>SRA Submission</h3> <p><a href="http://www.ebi.ac.uk/ena/about/sra_submissions">http://www.ebi.ac.uk/ena/about/sra_submissions</a></p> <ul style="list-style-type: none"> <li>Datenformats</li> <li>Metadaten Objekte</li> <li>Submission Account</li> <li>Übertragung</li> <li>Anpassen</li> <li>Dokumentation</li> </ul>
<h3>Submission Datenformat</h3> <ul style="list-style-type: none"> <li>Bevorzugte Format <b>BAM</b></li> <li>Andere: SRF, Fastq, SFF, SOLiD_native, illumina_native, PacBio_HDF5, CompleteGenomics_native</li> <li>Minimum Information: Basen und deren Qualität</li> <li>Sequenzierungen mit Barcode demultiplexen</li> <li>Technische Sequenzen eliminieren</li> </ul>	<h3>Metadaten Objekte</h3> <ul style="list-style-type: none"> <li>Submission</li> <li>Study</li> <li>Sample</li> <li>Experiment</li> <li>Run</li> <li>Analysis</li> </ul> <p>Für EGA andere:</p> <ul style="list-style-type: none"> <li>DAC, Policy, Dataset</li> </ul>
<h3>Metadaten Objekte (1)</h3> <ul style="list-style-type: none"> <li>Submission             <ul style="list-style-type: none"> <li>Neu oder Änderungen</li> <li>Veröffentlichungstermin</li> </ul> </li> <li>Study             <ul style="list-style-type: none"> <li>Projektziel mit Beschreibung</li> </ul> </li> <li>Sample             <ul style="list-style-type: none"> <li>Probenbeschreibung</li> <li>Organismus</li> </ul> </li> </ul>	<h3>Metadaten Objekte (2)</h3> <ul style="list-style-type: none"> <li>Experiment             <ul style="list-style-type: none"> <li>Library Information</li> <li>Platform Information</li> <li>Verbindet Study mit Sample und Run</li> </ul> </li> <li>Run             <ul style="list-style-type: none"> <li>Datenfiles mit primären Sequenzdaten</li> <li>md5sum</li> </ul> </li> <li>Analysis             <ul style="list-style-type: none"> <li>BAM files mit Referenzsequenzen</li> </ul> </li> </ul>
<h3>Metadaten Objekte (3)</h3> <ul style="list-style-type: none"> <li>Existierenden Study und Sample Objekte</li> <li>Eindeutiger Name für jeden Objekt</li> <li>Jeder Objekt bekommt Zugriffsnummer             <ul style="list-style-type: none"> <li>Submission: ERANNNNNNN</li> <li>Study: ERPNNNNNN</li> <li>Sample: ERSNNNNNN</li> <li>Experiment: ERXNNNNNN</li> <li>Run: ERRNNNNNN</li> <li>Analysis: ERZNNNNNN</li> </ul> </li> </ul>	<h3>SRA Webin: metadaten submission</h3>

 <p>FASTQ gezippt</p>	
	<p><b>Zusammenfassung</b></p> <ul style="list-style-type: none"> <li>• SRA – öffentliche Archivierungsstelle für Next-Generation Sequenzierungsdaten</li> <li>• Sequenzarchivierung mit Metadaten</li> <li>• Sequenzdateisuche über Metadatenindex</li> <li>• Verdoppelung der SRA Datenmenge pro Jahr</li> <li>• Wichtige Sequenzierungsformate: FASTQ, BAM</li> </ul>

### 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 Proben teil/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

- Sequence representations
  - Motivation
  - Requirements
- Efficient storage of genomic sequences
  - GTEncseq
  - Previous sequence representations
  - Results
- Efficient storage of large short read sets
  - Storage of sequencing data
  - Storage of resequencing data
- Future Work

Contents

- Sequence representations
  - Motivation
  - Requirements
- Efficient storage of genomic sequences
  - GTEncseq
  - Previous sequence representations
  - Results
- Efficient storage of large short read sets
  - Storage of sequencing data
  - Storage of resequencing data
- 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  $\alpha$
- 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

- Sequence representations
  - Motivation
  - Requirements
- Efficient storage of genomic sequences
  - GTEncseq
  - Previous sequence representations
  - Results
- Efficient storage of large short read sets
  - Storage of sequencing data
  - Storage of resequencing data
- Future Work

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://genome.tools.org>)

- written in portable C for POSIX compliant systems
- UNIX (Linux, BSD, Mac OS X, ...), Windows (with Cygwin)
- open source (BSD-license)
- components:
  - libgenomeTools shared library
  - collection of programs ("tools")

GenomeTools library

The diagram shows the GenomeTools library architecture. It is divided into several functional areas: Core (Identity management, Data structures, Index and access, Sequence parsers, Clustering, Hidden Markov models, Multireading support, Mark and coordinates, Stranded storage, Custom parser), Extended (Annotation parsers, Annotation handling, Stream processing, Clustering, Alignment, Annotation/Query), LTR (De novo LTR identification, LTR-comparative analysis, LTRharvest, LTRdigest), and Tools (Tallymer, LTRharvest, LTRdigest, Readjoinder, MetaGenomeThreader, Uniquesub). A central 'Annotation/Query' module connects the Extended and LTR sections. Below these are 'Tools' which include Tallymer, LTRharvest, LTRdigest, Readjoinder, sketch, and suffixsorter. A 'Storing/Retrieving' module is also shown at the bottom.

Tools using the *GTEncseq*

- Tallymer: fast and memory-efficient k-mer counting (S. Kurtz et al. BMC Genomics 9:507 (2008))
- LTRharvest: de novo detection of LTR retrotransposons (D. Ehrlich et al. BMC Bioinformatics 9:18 (2008))
- LTRdigest: annotation of internal features of LTR retrotransposons (S. Steinhilber et al. Nucleic Acids Research 37(11):7062-7013 (2009))
- Readjoinder: string graph-based short-read assembly (S. Gorenfeld and S. Kurtz, BMC Bioinformatics, accepted)
- MetaGenomeThreader: gene prediction in metagenome sequences (D. J. Schmidt Hübner and S. Kurtz, In: Metagenomics: Methods in Molecular Biology, 335-358, Humana Press, Totowa, NJ)
- Uniquesub: minimum unique substrings for designing tiling arrays (S. Graf et al. Bioinformatics 23(11):1495-1504 (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

The bar chart compares the size of representation in GB for different formats. The y-axis is 'size of representation (GB)' ranging from 0 to 40. The x-axis lists formats: GenBank, EST, short reads, human, and nr. For each format, four bars are shown: GenBank (black), BLAST (red), BLAT (green), and uncompressed (blue). GenBank and EST are the largest, around 37-38 GB. Short reads are significantly smaller, around 5.6 GB. Human and nr are the smallest, around 3-4.5 GB. The uncompressed format is consistently the largest for each category.

### 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<sup>6</sup> random access to single bases

Version	GtEncseq	SeqAn	BLAT enc.	BLAT acc.
1.3.8	1.2	v04	C	6.1
Implementation language	C	C++	Fortran	C
Sequence loading (s)	0.003	100.1	207.3	0.003
Extraction				
(1) Exonic sequences (s)	6.5	5.2	202.5	5.3
(2) Random access (s)	0.3	0.7	0.4	0.003
Memory peak (MB)	737	954	384	3,336
				1,456

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

- Sequence representations
  - Motivation
  - Requirements
- Efficient storage of genomic sequences
  - GtEncseq
  - Previous sequence representations
  - Results
- Efficient storage of large short read sets
  - Storage of sequencing data
  - Storage of resequencing data
- Future Work

### Data types

#### Sequencing data

- newly sequenced reads obtained using NGS technology
- often given in FASTQ format (P. Cook et al. *Nucleic Acids Res.* 30(6):1702-71 (2002))
  - 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...

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

### Sequencing data – FASTQ input format

```

@read.1 length=26
GAAACATTACCGAGTCTGTTGATT
+
IIIIIIIIII<ERDIIIIIEF--I
@read.2 length=26
TACAGATGACCGATTAAAGGCAATCT
+
BBBABBBBBS1B:DBB111###
    
```

#### Selection of specific encoding techniques

- description lines are highly repetitive
  - increasing numbers, equal strings, ...
- analyze structure and apply most appropriate encoding scheme (S. Demme et al. *Bioinformatics* 27(16):2021 (2011))
- sequences/quality pairs have characteristic occurrence frequencies
  - employ encoding scheme based on statistical measures (W. Tomke et al. *Bioinformatics* 26(6):730-34 (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

SRRO29844.1.filt, 1000 genomes proj. 5.7 × 10<sup>7</sup> reads, 76 bp, 1.14 GB

sampling distance	bits/pair	bits/desc	compr. ratio	compression speed	extraction time for 10 <sup>6</sup> 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	223 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.F. Fritz et al. *Genome Res.* 22(3):739-46 (2012))

#### 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

ACGATCTTAATGCCCTACTTGT--GG-CATTC reference

ATCGTAAT--TTAC mapped reads

ATGCCCTACTTGT

ACTTGTATGGCC

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

Acknowledgements

- Sascha Steinbiss (main developer, interfaces, integration)
- Dirk Willrodt (header compression, integration of code)
- Joachim Bonnet (short read compression)
- Giorgio Gonnella (large-scale testing)

Stefan Kurtz (ZIBB, Uni. Hamburg) Storage and retrieval of bioinformatics March 27, 2012 24 / 24

### 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

**Slide 1: Forschungsdatenmanagement biomedizinischer Genomdaten**  
SNP array und Next-Generation Sequenzierungs Daten  
Michael Wittig  
ikmb Institut für Klinische Molekulargenetik  
LABIMI

**Slide 2: SNP array Daten**  
Bsp: Affymetrix Genome Wide Human SNP array 6.0  
- 934,968 SNPs  
- 746,000 CNV probes

**Slide 3: SNP array Daten**  
Bsp: Affymetrix Genome Wide Human SNP array 6.0  
- 934,968 SNPs  
- 946,000 SV probe sets  
genotype AA  
genotype AG  
genotype GG

**Slide 4: SNP array Daten**  
Genotyp Speicherung  
SNP annotation  
0 – NoCall  
1 – Hom. A  
2 – Heterozygot  
3 – Hom. B

**Slide 5: SNP array Daten**  
Intensitäten Speicherung  
SNP annotation  
- 1 float/value  
-> 2 float/SNP

1.38	2.03	1.82	2.47	0.91	1.38	0.41	1.35	1.87	1.85
1.96	0.99	0.89	4.81	4.10	1.39	4.11	0.11	2.79	0.61
1.26	2.79	0.97	2.80	0.16	0.11	4.81	1.55	1.51	1.31
0.67	0.81	1.51	2.04	2.80	1.98	1.24	4.52	1.15	1.60
0.10	0.36	1.91	0.68	2.65	4.76	0.18	0.96	4.76	0.79
1.60	0.50	1.81	0.28	0.79	1.15	2.79	1.57	1.17	2.11
2.29	1.58	1.45	1.81	1.60	1.78	2.10	0.20	1.07	1.11
4.80	1.81	1.80	4.81	4.84	4.12	1.12	1.20	1.26	2.70
2.47	2.58	2.43	2.21	4.04	0.26	0.60	4.43	1.22	4.61
1.75	1.04	1.74	1.45	1.75	0.44	1.12	4.85	4.68	1.10
0.81	0.39	4.81	4.07	0.88	4.12	1.00	4.91	1.01	1.70
1.64	0.10	1.60	2.02	0.88	0.12	4.11	1.10	1.10	1.02
1.16	0.17	4.18	2.41	0.98	1.10	1.10	1.10	2.22	0.14
1.40	0.20	1.61	0.21	1.74	1.45	0.11	4.10	1.68	0.63
1.42	1.18	1.74	1.15	1.43	0.12	1.11	1.44	0.26	1.80
0.10	1.18	0.16	4.81	1.60	1.47	1.11	4.10	1.10	0.86
1.24	1.15	0.41	0.17	1.01	0.07	1.14	4.10	4.60	0.19
4.80	1.78	2.41	0.61	4.61	0.01	0.11	1.86	1.10	1.11
1.80	1.18	1.11	0.11	2.27	4.10	1.14	1.10	1.10	1.11
4.10	1.18	1.42	4.19	0.11	1.10	0.11	0.17	2.70	0.74

**Slide 6: SNP array Daten**  
SNP Array Data Interface  
8 MB/sample  
553 GB am IKMB

### SNP array Daten

SNP Array Data Interface

SNP annotation:  
 src\_juncmapchip.bed  
 src\_juncmapchip.bed  
 src\_juncmapchip.bed  
 Files are individual major mode. Conversion to default (SNP major mode), use:  
 -cfile  
 -makefile  
 -root  
 --flow-se-sets (optional)  
 If you get errors, try link v1.02 for conversion  
 11:31:15 START PLINK-FILE3 a florian@ikmb.mz.uni-bonn.de Zehn/1000-6466-487-ack4-88459031448  
 USE\_ORIGINAL\_NAMES: database\_clean\_juncmap\_v2\_broad/bedtools  
 Unknown individual Contigs skipped.  
 Strand information: 102115 plus strand SNPs, 83617 minus strand SNPs and 792 SNPs without strand information.  
 results written until Thursday, Mar 29 11:31 AM  
 11:31:29 END JOB

### SNP array Daten

Zusammenfassung

- Speicherung der Originaldaten (Aflly6 0 z Bsp 30MB/array)
- Genotypbestimmung durch aktuellen calling Algorithmus
- QC nach Herstellerangaben (Aflly6 0 z B. call rate, contrastQC)
- Verfügbarkeit über hausinterne Datenbank im Plink Format
- SNP spezifische Qualitätskontrolle über scatter plots

### NGS IT Infrastruktur

NextGeneration Sequencing Unit Backup server High Performance Compute Cluster Bioinformatics Workstations

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

### NGS Analysis Workflow

```

    graph LR
    A[Images] --> B[Primary reads base qualities]
    B --> C[Mapped reads - Configs - Variationen]
    C --> D[Biologischer Kontext Visualisierung]
    
```

### NGS quality control

fastQC

Basic Statistics

Measure	Value
Filename	A0850_ACTGA_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

### NGS quality control

fastQC

### NGS quality control

fastQC

Per sequence quality scores

### NGS quality control

fastQC

Per base sequence content

### NGS quality control

fastQC

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

### NGS quality control

Metadaten / Projektbezogen

```

run_date
machine
flowcell
flowcell_id
lane
barcode
project
sample
project_type
reference
description
#PF:reads
#mp:reads
mean_quality_PF
%Q30 bases
%PF
%raw cluster
%filtered index
%duplicates R1
%GC R1
avg_mean_quality_per_base_R1
avg_quality_per_base_R1
#flow qual positions R1
#fpos with %>=1 R1
max % N R1
%mapped R1
#duplicates R2
%GC R2
avg_mean_quality_per_base_R2
avg_quality_per_base_R2
#flow qual positions R2
#fpos with %>=1 R2
max % N R2
avg_quality_per_sequence_R2
%mapped R2
avg_insert_size
median_insert_size
    
```

### NGS quality control

Metadaten / Sequenzierungslauf Bezogen

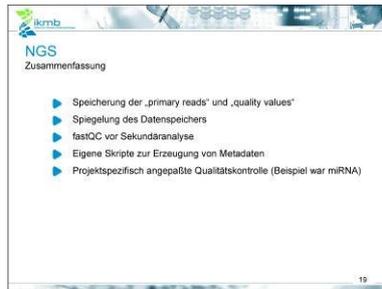
```

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
    
```

### NGS quality control

Anwendungsspezifische QC

	IGV1	IGV2	IGV3	IGV4	IGV5	IGV6	IGV7	IGV8	IGV9	IGV10	IGV11	IGV12	IGV13	IGV14	IGV15	IGV16	IGV17	IGV18	IGV19	IGV20
before	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000
after	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000	900000



### 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-Einrichtung?

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.