Universität Heidelberg
Institut für Medizinische Biometrie und Informatik
Sektion Medizinische Informatik

Studiengang Medizinische Informatik
Bachelorthesis

Auswahl und Implementierung eines Scientific Workflow Management Systems zur Analyse von Next-Generation Sequencing Daten

Referentin: Frau Prof. Dr. Petra Knaup-Gregori
Korreferentin: Frau Dr. rer. nat. Stephanie Rössler

Verfasser: Moritz Juchler

23. September 2013

Inhalt

1 Einbettung und Einleitung

1.1 Einbettung der Bachelorarbeit in den SFB/TRR 77

In dem transregionalen Sonderforschungsbereich SFB/TRR 77 untersuchen Heidelberger und Hannoveraner Wissenschaftler Entstehungsmechanismen und neue Therapieansätze des Leberzellkarzinoms, einer der tödlichsten Tumorerkrankungen unserer Zeit.

Die IT-Plattform Pelican, die ein Teil des Gebiets Z2 ist, soll dem Forschungsverbund die softwaregestütze Analyse und die nachhaltige Bereitstellung von Leberkrebs-Forschungsdaten ermöglichen [Ganzinger et al. 2011].

Ein Teil von Pelican soll eine gemeinsame Informationsplattform anbieten, die die biomedizinischen Daten der verschiedenen medizinischen und biologischen Projekte integriert und den beteiligten Projektgruppen biostatistische Programme und projektübergreifende Auswertungen zur Verfügung stellt. Die Integration von Gewebe-, Molekül-, Genetik- und Klinikdaten in eine gemeinsame Plattform ermöglicht Datenerhaltung und umfassende Analysen. Die integrierte Analyse begegnet durch die Verknüpfung verschiedener Forschungsprojekte des SFB/TRR 77 den Herausforderungen der Multidisziplinarität klinischer Forschung und Genforschung.

1.2 Gegenstand und Bedeutung

Mit dem Next-Generation DNA Sequencing ist durch Kostenreduzierung und immenser Zeiteinsparung die DNA Sequenzierung einem breiten Spektrum an Wissenschaftlern zugänglich geworden und hat Kompetenzen zur Sequenzierung von zentralen Stellen in die Hände vieler individueller Forscher gelegt [Shendure and Ji 2008, Ding et al. 2010, Wetterstrand 2011]. Die Kombination dieser hochentwickelten Technologien aus der Gentechnik und rechnerbasierten Werkzeugen erlaubt die Beantwortung biologischer Fragestellungen in erheblich umfangreicherer Art und Weise als dies bisher möglich gewesen ist [Shaer et al. 2013]. Die rasche Entwicklung des Next-Generation Sequencing beinhaltet auch das Konstruieren neuer Ansätze zur bioinformatischen Datenanalyse, ohne die kein Informationsgewinn, wie beispielsweise die Entdeckung von Genvariationen, möglich wäre. Das dabei neu gewonnene Wissen kann zu erheblichen Fortschritten in der Krebsforschung führen, beispielsweise wenn es um das Identifizieren der Genomveränderungen einer Tumorzelle geht [Ding et al. 2010]. Anstatt Sequenzierungen in kleinem Maßstab durchzuführen, können Forscher inzwischen Sequenzierungen in weit umfangreicherem Ausmaß realisieren, in denen Informationen von multiplen Genen und Genomen vermessen, dokumentiert und in Datenbanken gespeichert werden können. Die DNA Sequenzen werden nach der Sequenzierung in einer Kette aus vielen Prozessschritten – eine bioinformatische Pipeline – analysiert und verarbeitet. Zu den Einzelschritte, wie zum Beispiel Alignment oder die Entfernung von Duplikaten, gibt es oftmals viele Alternativen.

1.3 Problematik

Indes bleibt die Weiterverarbeitung der durch das Next-Generation Sequencing gewonnenen Daten wegen deren enormen Umfangs hinter den Erwartungen der Forscher zurück. Die Datenanalyse ersetzt die Datengenerierung als limitierenden Faktor in der Genomforschung [Nielsen et al. 2010]. Die Interpretation von Daten in größerem Ausmaß erfordert neue computergestützte Werkzeuge, die die Analyse unterstützen. Allerdings zeigen die heutigen Rechenwerkzeuge deutliche Einschränkungen in Langlebigkeit, Benutzbarkeit und Unterstützung für Zusammenarbeiten und Gedankengänge auf hoher Ebene. Bisherige Herangehensweisen sind oft noch sehr langsam und benötigen eine komplexe Installation und Konfiguration [Shaer et al. 2010].

1.4 Motivation

Die Informationsflut, die aufgrund der neuen Technologien in der Genomforschung entstanden ist, treibt nicht nur den Umfang an Erforschung voran, sondern auch die Weiterentwicklung der von den Wissenschaftlern benutzten Werkzeuge. Neben einer Pipette und einem Stift ist heute der Webbrowser ein Werkzeug, ohne das kaum noch Informationsgewinn möglich ist. Ein einfach zu konfigurierender Workflow von der Dateneingabe über die Verarbeitung bis hin zur Ergebnisausgabe erlaubt es Forschern, sich auf das Wesentliche, also biomedizinischen Vorgänge und Analysen, zu konzentrieren. Bei einer entsprechenden komfortablen Vorgehensweise werden möglichst viele Schritte automatisiert. Dies beginnt bereits bei der Benutzerregistrierung und erfordert keine speziellen Kenntnisse über die Installation und die Konfiguration der eingesetzten Software [Otto et al. 2008, Shaer et al. 2010]. Durch die Kapselung können Anwender sich auf die inhaltlichen Entscheidungen konzentrieren, zum Beispiel welcher Algorithmus zum Einsatz kommt oder welche Datenbank durchsucht werden soll um sog. Single Nucleotide Polymorphisms (oder kurz SNPs, siehe Beschreibung unten) zu identifizieren.

1.5 Problemstellung

Derzeit ist noch nicht bekannt, welches Programm die oben erwähnten Anforderungen an die computergestützten Werkzeuge bestmöglich erfüllt und die bereits genannten Einschränkungen am wirkungsvollsten zurückdrängt.

In dem SFB/TRR 77 ‚Leberkrebs – von der molekularen Pathogenese zur zielgerichteten Therapie‘ Projekt gibt es noch kein einfach zu bedienendes Werkzeug, mit dem eine Pipeline erstellt und automatisiert ausgeführt werden kann. Es bleibt bisher nur die Möglichkeit die einzelnen Schritte sequentiell nach einem zuvor erstellten Skript durchzuführen, was eine Reihe von Problemen mit sich bringt. Hierzu gehören das Risiko des Datenverlusts bei einem Programmabbruch und die Notwendigkeit der Auswahl verschiedener Parameter vor der Durchführung einer Analyse.

1.6 Zielsetzung

1 Auswahl eines geeigneten Pipeline Management Tools zum komfortablen Ausführen und Verwalten von bioinformatischen Pipelines.

2 Exemplarische Etablierung einer Pipeline, die computergestützt Genomdaten analysiert und zur Identifikation von Genvariationen führt.

1.7 Fragen und Aufgaben

1 Auswahl eines geeigneten Pipeline Management Tools zum komfortablen Ausführen und Verwalten von bioinformatischen Pipelines.

1.1 Welche Programme stehen zur Auswahl?

1.2 Welche funktionalen und nicht-funktionalen Kriterien muss das Programm erfüllen?

1.3 Welches Programm erfüllt die Kriterien am besten?

1.4 Installation des Programms.

2 Exemplarische Etablierung einer Pipeline, die computergestützt Genomdaten analysiert und zur Identifikation von SNPs führt.

2.1 Wie soll die Pipeline aufgebaut und konfiguriert sein?

2.2 Pipeline Ausführung auf den Leberkrebsdaten.

2 Grundlagen

2.1 Hintergrundinformationen zu DNA Sequenzierung

Sequenzierung. Den größten Einfluss hat das Next Generation Sequencing auf die Krebsforschung durch die Möglichkeit des Sequenzierens, des Analysierens und vor allem des Vergleichs von paarigen Tumor- und Referenzgenomen eines einzelnen Patienten [Mardis and Wilson 2009]. So können Genomveränderungen wie einzelne Nukleotidaustausche, strukturelle Umstrukturierung oder Sequenzvervielfältigungen herausgefunden werden [Meyerson et al. 2010]. Das Prinzip des Next Generation Sequencing beruht darauf, dass Millionen von kurzen fragmentierten DNA-Abschnitten auf eine kleine Oberfläche gebunden und gleichzeitig ausgelesen werden. Mittels Hybridisierung werden die fragmentierten DNA-Abschnitte (b) (siehe Abbildung 1), welche kodierende Sequenzen enthalten, an komplementäre Sequenzen gebunden. Danach werden Introns, nichtkodierende, zwischen den Genen liegende DNA-Abschnitte aus-gewaschen. Die exonischen, also die kodierenden DNA-Fragmente werden mittels Adaptoren, die an die Enden der Fragmente andocken, an eine Oberfläche wie beispielsweise Glas gebunden (c) und dort vervielfältigt (d). Nun liegen Milliarden DNA-Abschnitte, die an die Oberfläche gebunden und vervielfältigt wurden, zur Sequenzierung vor. Diese startet, indem Nukleotide (Arginin, Guanin, Cytosin und Thymin), welche farblich unterschiedliche Fluoreszenzmarker tragen, über diese Oberfläche geleitet werden (e). Die eingeleiteten Nukleotide binden der Abfolge der Nukleinbasen entsprechend an die zu sequenzierenden DNA-Abschnitte (f) und emittieren dabei das für die eingeleitete Base spezifische Farb-signal. Dieses spezifische Farbsignal wird von einem Laserscanner erfasst und abgespeichert (g). Das Einfluten von Nukleotiden, deren Bindung

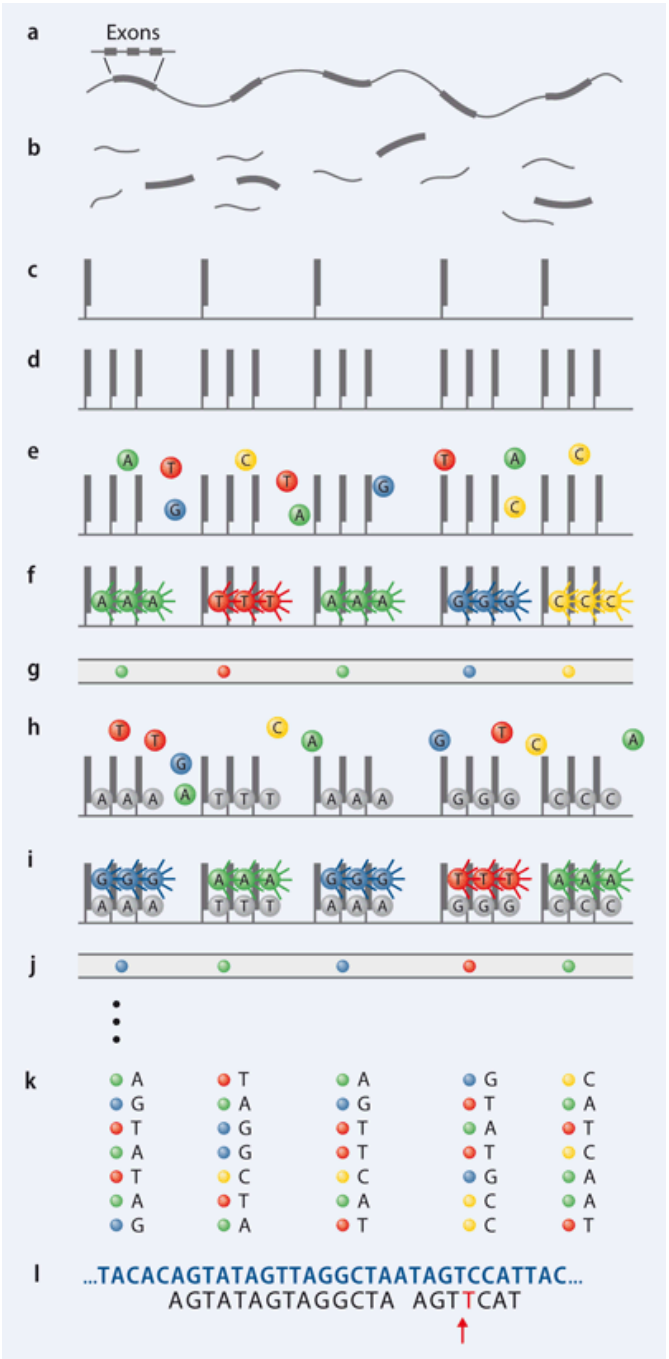


Abbildung 1: Prinzip der Exomsequenzierung aus [Hempel et al. 2011]

an die DNA-Fragmente und das Freisetzen und Auslesen der Farbsignale wird mehrfach wiederholt (h,j,k). So liest der Laserscanner für jeden DNA-Cluster eine bestimmte Abfolge von Fluoreszenzsignalen aus, anhand derer für jeden einzelnen an der Oberfläche gebundenen DNA-Abschnitt die Basenabfolge bestimmt werden kann [Hempel et al. 2011]. Die Farbsignale werden für jedes DNA-Fragment in eine Sequenz übersetzt (Read). In Abb. 1 ist der erste Read beispielsweise AGTATAG (l).

Qualitätsbeurteilung. Der erste Schritt der Analyse nach dem Durchlauf des Sequenzierers ist die Feststellung der Qualität der DNA-Abschnitte und deren Entfernung, Kürzung oder Korrektur, falls Qualitätskriterien nicht eingehalten werden. Die rohen Daten beinhalten Sequenzartefakte die beim Übersetzen der Fluoreszenzsignale in die Basenfolge auftreten, beispielsweise falsch positive INDELs also Insertion und Deletion von Mutationen oder Reads geringer Qualität. Solche Fehler sind in Sequenzdaten üblich, weil die Plattformen anfällig für eine Reihe chemischer oder instrumentenspezifischer Störungen sind. Diese Schritte der Qualitätssicherung sind deswegen nötig, um das Ziehen falscher Schlüsse zu vermeiden [Pabinger et al. 2013].

Alignment. DNA-Abschnitte, die die Qualitätskriterien erfüllen, werden mit einer Referenzsequenz verglichen und ihrem Platz im Genom zugeordnet (Alignment). Ein menschliches Referenzgenom in verschiedenen Versionen bietet beispielsweise die University of Santa Cruz (UCSC) an. UCSC bietet die Version hg18 (human genome 18) und hg19, die neueste Version von 2009, an, während das Genome Reference Consortium GRCh36 und GRCh37 anbieten. Zusammen sind diese die meistverwendeten menschlichen Referenzgenome [Pabinger et al. 2013]. Es gibt viele Programme die zur Anordnung der DNA Abschnitte genutzt werden können. Das Ergebnis das Alignments ist in Abb. 1 (l) zu sehen. Bei einer erfolgreichen Sequenzierung können mehr als 95% der DNA Abschnitte der Referenzsequenz zugeordnet werden.

Varianten-Identifizierung. Das Ziel der Sequenzierung ist das Entdecken von Abweichungen der DNA Abschnitte von der Referenzsequenz, sogenannte Varianten [Hempel et al. 2011]. In der Krebsforschung werden hierzu durch den Vergleich von Tumorgewebe mit gesundem Gewebe somatische Mutationen identifiziert. Somatische Mutationen entstehend während der embryonalen oder fetalen Phase oder aber auch im Laufe des Lebens eines Menschens [Schaaf and Zschocke 2013]. Da ein Großteil der Sequenzvarianten für den Phänotyp eines Organismus nicht relevant sind, müssen, um die krankheitsrelevanten Sequenzvarianten herauszufiltern, nur Varianten betrachtet werden, die zu Veränderungen auf Proteinebene führen. Dies können sein:

- Nicht-synonyme Missense-Mutationen, also Punktmutationen, die in ein Kodon resultieren, das für eine andere Aminosäure kodiert [Richards and Hawley 2010],
- Deletitionen, das bedeutet die Entfernung einzelner Basen,
- Insertionen, das bedeutet das Einfügen einzelner Basen,
- Nonsense-Mutationen, die durch Bildung eines Stopkodons zum vorzeitigen Abbruch bei der Proteinbiosynthese führt [Heinritz 2004, Knust and Janning 2008].

Varianten-Annotation. Durch die großen Datenmengen ist es zunehmend wichtig geworden, automatisch die funktionelle Auswirkung von Varianten abzuschätzen. Computerunterstützte Annotation ermöglicht es Forschern krankheitserregende Mutationen zu filtern und einzuordnen. Meistens liegt der Fokus der Annotation auf SNPs, da diese schnell identifiziert und analysiert werden können. Single Nucleotide Polymorphism (SNP) bezeichnet Variationen einzelner Basenpaare in einem DNA-Strang, wie in Abb. 1 (l) zu sehen. Dort wird aus einem Cytosin eine Thymin. Sie stellen etwa 90% aller genetischen Varianten im menschlichen Genom dar [Schaaf and Zschocke 2013]. Manche Programme finden auch INDELs, während die Annotation von strukturellen Varianten auf Copy Number Variations (CNV), also Abweichungen der Anzahl der Kopien eines bestimmten DNA-Abschnittes, beschränkt ist [Pabinger et al. 2013]. Die gebräuchlichste Form der Annotation ist der Abgleich der Varianten mit öffentlichen Variantendatenbanken wie dbSNP, der Single Nucleotide Polymorphism Database [Sherry et al.

2001]. Es wird beispielsweise nach nicht-synonymen Varianten und nach Varianten, die nicht in der Normalbevölkerung vorkommen, gesucht. Hinsichtlich der Funktion der Varianten verfolgen die Programme unterschiedliche Ansätze, die von einer simplen Sequenzanalyse bis zur Untersuchung des Einflusses der strukturellen Veränderungen auf Proteine reichen. So wird die Pathogenität der neuen Varianten untermauert [Hempel et al. 2011, Pabinger et al. 2013].

2.2 Hintergrundinformationen zu ‚Scientific Workflow Management Systemen'

Ein Workflow ist eine abstrakte Beschreibung einzelner Schritte, die zur Ausführung eines einzelnen realen Prozesses benötigt werden, und dem Informationsfluss zwischen diesen Schritten. Jeder Einzelschritt wird durch eine Reihe von auszuführenden Aktionen definiert. Innerhalb eines Workflows durchlaufen die Daten die verschiedenen Schritte in einer festgelegten Reihenfolge und die Aktionen in jedem Schritt werden entweder durch eine Person oder durch eine Systemfunktion, zum Beispiel ein Computerprogramm, ausgeführt [Curcin and Ghanem 2008]. Somit bieten Workflows eine deklarative Art und Weise um eine Anwendung zu spezifizieren, während dem Benutzer die zugrundeliegenden Details verborgen bleiben [Talia 2013].

Oftmals sammeln und analysieren Benutzer Daten von mehreren Quellen und benutzen zur Analyse selbst mehrere Programme, so dass stufenweise das Ergebnis eines Schrittes als Eingabe des nächsten dient [Abouelhoda et al. 2010]. Dies wird als Pipeline definiert [Hohpe and Woolf 2012].

‚Scientific Workflow Management Systeme' (SWfMS) beinhalten Software Systeme, die computergestützt Pipelines handhaben und ausführen. Es gibt eine ganze Reihe solcher Tools, die sich in Umfang und Umsetzung unterscheiden. Die Kategorie der Systeme, auf die der Fokus dieser Arbeit gelegt wird, ermöglicht Biologen, Daten in einer Weise zu handhaben, ohne dass tiefere Programmierkenntnisse wie Skriptsprachen nötig sind.

Dabei verrichten SWfMS Tätigkeiten wie das Abrufen von Sequenzen von öffentlichen Beständen/Sammlungen, die Extraktion von Sequenzausschnitten, die Konvertierung zwischen verschiedenen Dateiformaten und die Durchführung von Mengenoperationen auf Ergebnissen [Orvis et al. 2010, Bux and Leser 2013]. Da es für die einzelnen Schritte eine Fülle an verschiedenen Tools gibt, kann deren Auswahl sehr komplex werden [Pabinger et al. 2013].

Aufgabe der vorliegenden Arbeit ist es unter anderem, die Anforderungen an ein Scientific Workflow Management System für das SFB/TRR 77 zu ermitteln, ein für das TRR Projekt geeignetes System auszuwählen und für die zur Verfügung gestellten Genomdaten prototypisch zu implementieren.

3 Methodik

Da es bereits eine ganze Reihe von Pipeline Management Tools gibt, erscheint die Möglichkeit ein solches für unsere Zwecke neu zu programmieren nicht notwendig.

Geeigneter erscheint der Weg ein bereits auf dem Markt befindliches Programm auszuwählen, welches unseren Erfordernissen entspricht. Falls ein solches Programm nicht gefunden werden kann, soll das geignetste Tool an unsere Erfordernisse angepasst werden.

Für eine solche Auswahl muss zunächst eine systematische Suche nach Programmen erfolgen, die dann vor dem Hintergrund unserer Erfordernisse analysiert werden müssen. Hierfür stellt zunächst eine Anforderungsanalyse einen Kriterienkatalog auf, der für das Projekt relevante Eigenschaften des Tools erfasst. Auf Basis dieser Übersicht werden die verschiedenen Management Systeme miteinander verglichen und eine entsprechende Auswahl für die Verwendung im SFB/TRR 77 getroffen.

Um einen Überblick über die verfügbaren Werkzeuge zu erhalten, wurde eine Internet- und Literaturrecherche zum Thema Workflow Management Systeme betrieben, auf deren Basis wir auf die im Ergebnisteil aufgeführten zur Auswahl stehenden Programme gestoßen sind. Folgende Suchbegriffe wurden für die Recherche benutzt:

- Bioinformatic pipelines
- Managing bioinformatic pipelines
- Scientific workflow management system

Speziell www.ncbi.nlm.nih.gov/pubmed und www.scholar.google.com liefert bei diesen Suchbegriffen wissenschaftliche Dokumente, die die Programme beschreiben. Die ersten 50 Artikel, die die Suchanfragen identifizieren, wurden im Rahmen dieser Bachelorarbeit untersucht. Dazu wurden die Zusammenfassungen gelesen und eingeordnet, ob die Artikel zur Beantwortung der Fragestellungen dienen können. Artikel, die keine Informationsaussage lieferten, wurden ausgeschlossen.

Auch gibt es eine Reihe von Quellen, die eine Übersicht bestehender Programme liefern und diese unter bestimmten Gesichtspunkten vergleichen. Dazu hilft die Suche nach

- Comparison scientific workflow management system
- Overview scientific workflow management system

Aufgrund der Rahmenkriterien und der großen Anzahl an Tools haben wir eine Vorauswahl entsprechend des späteren Einsatzgebietes des Programms getroffen. Bei dieser projektspezifischen Auswahl wurde darauf Wert gelegt, dass das Management System Open Source ist und nicht kommerziell vertrieben wird, sondern frei verfügbar ist. Ein weiterer Gesichtspunkt war eine hohe Verbreitung des Programms. Dies wurde durch das Vorhandensein eines Supportforums oder einer ähnlichen Einrichtung validiert. Damit geht die Unterstützung bei Fragen und Problemen während der Installation und der Benutzung einher.

Die Kriterien der SWfMS finden sich durch eine Suche nach

- (key) requirements scientific workflow management
- requirements bioinformatic pipeline

Ein Ansatz, der in abgewandelter Form bei der Evaluation möglicher Scientific Workflow Management Systeme in dieser Arbeit verfolgt werden soll, wurde von Lin et al. gewählt [Lin et al. 2009]. Dort wählen die Autoren fünf repräsentative Programme aus und evaluieren diese gegen sieben Anforderungen, indem sie den Programmen je Anforderung ein „+“ bei erfüllter Anforderung, ein „−“ bei fehlender Unterstützung des Kriteriums und ein „+/−“ bei teilweiser

oder zweifelhafter Erfüllung. Dieses Verfahren zeigt sowohl Stärken als auch Schwächen jedes Programms und passt auf unsere Anforderungen, weswegen wir es mit weniger Programmen und verringerter Kriterienanzahl übernehmen. Im Gegensatz zu der Analyse von Lin et al. konnte ich mich in dieser Arbeit auf die vier im Ergebnisteil beschriebenen Anforderungen in vier Programmen beschränken.

Zur Beantwortung der Fragestellung, inwieweit die Programme die Kriterien erfüllen, ergeben eine Suche auf www.scholar.google.com und www.ncbi.nlm.nih.gov/pubmed mit den Suchbegriffen „<Programmname> <Kriterienname>“ und die bei der Programmsuche verwendete Literatur die benötigten Ergebnisse.

Für die prototypische Implementierung wird das am besten geeignete Tool testweise auf dem TRR Server installiert und so konfiguriert, dass die Pipeline durchgeführt werden kann. Um zu evaluieren, inwieweit das Tool in der Praxis geeignet ist, wird abschließend die Pipeline auf den Daten ausgeführt.

4 Ergebnisse

4.1 Welche Programme stehen zur Auswahl?

Es gibt eine große Auswahl an zur Verfügung stehenden Produkten. Diesen gemein ist lediglich das Ziel, wissenschaftliche Workflows komfortabel zu erstellen und auszuführen. Die Programme verschiedener Hersteller unterscheiden sich neben weiteren Aspekten im Funktionsumfang, der Bedienung und den verwendeten Technologien. Deshalb wird eine systematische Programmsuche und –auswahl nötig. Wir beschränken uns im Rahmen dieser Arbeit auf Open Source Produkte, also Software deren Quelltext öffentlich zugänglich ist und deren Verbreitung und Modifikation nicht beschränkt ist [Perens 1999]. Des Weiteren beschränken wir uns auf frei am Markt verfügbare Produkte, die keine Lizenz- oder Nutzungskosten verursachen.

Die Literaturrecherche mit den im Methodikteil genannten Begriffen hat zu untenstehenden Ergebnissen geführt (Stand: 14.08.2013):

Datenbank / Suchbegriff	www.ncbi.nlm.nih.gov/pubmed	www.scholar.google.com
Bioinformatic pipelines	282	16800
Managing bioinformatic pipelines	5	16300
Scientific workflow management system	318	17700

Tabelle 1 Anzahl an Suchergebnissen zur Programmauswahl

Tabelle 1 zeigt die Anahl an Suchergebnissen zweier Suchmaschinen bei unterschiedlichen Suchgebriffen. Wie im Methodikteil beschrieben, werden die ersten 50 Abstracts gelesen. Einige herausstechende Artikel sind besonders zielführend und werden deswegen im Folgenden kurz erläutert.

In „Parallelization in Scientific Workflow Management Systems“ findet sich eine Aufteilung der Management Systeme in die drei Klassen textuelle Eingabe, grafische Darstellung oder anwendungsspezifische Webportale [Bux and Leser 2013], in die zum besseren Verständnis auch die vorgestellten Programme eingeteilt werden sollen.

Die Webseite https://code.google.com/p/bpipe/wiki/ComparisonToWorkflowTools bietet eine Übersicht bestehender Scientific Workflow Management Systeme. Bux und Leser bieten trotz Themenfokus auf die Parallelisierung von Workflows eine gute Übersicht über alle Workflow Management Systeme [Bux and Leser 2013]. Die Wikipedia Seite zu Scientific Workflow Systems en.wikipedia.org/wiki/Scientific_workflow_system zitiert Curcin und Ghanem, die in „Scientific workflow systems - can one size fit all?“ diverse Management Tools vorstellen [Curcin and Ghanem 2008].

Die im folgenden vorgestellten Programme sind die vier am besten geeigneten Programme für die Erstellung und der wiederholten Ausführung von bioinformatischen Pipelines. Es werden Gründe für ihre Nominierung genannt. Ausgeschlossene Programme haben teilweise einen anderen Schwerpunkt für Workflows, wie YAWL oder BPEL, die den Schwerpunkt im Unternehmensbereich haben oder Triana, dessen Schwerpunkt auf Bild- und Audioanalyse liegt.

Bpipe. Bpipe ist eine einfache Skriptsprache zur Definierung einzelner Programmschritte und deren Verkettung um bioinformatische Pipelines auszuführen und zu managen. Es spezialisiert sich vor allem darin, vorhandene Shell-Skripte mit möglichst geringem Aufwand in flexible, anpassungsfähige und wartungsfreundliche Workflows umzuwandeln [Sadedin et al. 2012].

Galaxy. Ursprünglich im Jahr 2010 nur für die Genomforschung entwickelt, hat sich Galaxy stark ausgebreitet und wird mittlerweile als gebräuchliches bioinformatisches Workflow Management System eingesetzt. Es ist eine webbasierte Open Source Plattform, die auch ohne Programmiererfahrung leicht zugänglich ist. Galaxy ist als Webportal sowohl mit einer webbasierten Benutzerschnittstelle als auch mit verschiedenen vorgefertigten Komponenten für gängige Aufgaben ausgestattet [Goecks et al. 2010, Bux and Leser 2013].

Taverna. Entwickelt 2004 an der University of Manchester, ist Taverna ein Open Source, auf Java basierendes Workflow Management System, dessen vorrangiges Ziel die Unterstützung der Biowissenschaften durch die Erstellung und Ausführung von wissenschaftlichen Workflows ist. Als sehr allgemein gehaltenes Programm legt Taverna Workbench, allgemein Taverna, einen starken Fokus auf die Benutzerfreundlichkeit und bietet unter anderem eine graphische Benutzeroberfläche zur Modellierung von Workflows aus zahlreichen unterschiedlichen Webservices an [Bux and Leser 2013].

Kepler. Basierend auf Java, bietet Kepler eine graphische Benutzeroberfläche und eine Runtime Engine, die Workflows entweder innerhalb der Benutzeroberfläche oder innerhalb einer Kommandozeile ausführen kann. Kepler wird seit 2005 durch ein Team entwickelt und gepflegt, das sich aus Mitarbeitern von sieben Institutionen der University of California zusammensetzt [Talia 2013]. Es stellt eine Auswahl an vorgefertigten Komponenten mit Fokus auf statistischer Analyse zur Verfügung. Kepler funktioniert nach dem Konzept von Direktoren, die vorschreiben, welche Schritte durchzuführen sind. Einzelne Schritte im Workflow werden als wiederverwendbare Aktoren eingebaut und können Datenquellen, Datensenken, Datenumwandlungen, Untersuchungsschritte oder andere beliebige Schritte sein. Workflows können sowohl innerhalb einer graphischen Oberfläche (siehe Anhang 3) als auch einer Kommandozeile ausgeführt werden [Bux and Leser 2013, Talia 2013].

4.2 Welche funktionalen und nicht-funktionalen Kriterien muss das Programm erfüllen?

Die technischen Rahmenbedingungen sind der TRR-Server, der unter dem Betriebssystem Linux OpenSuse 11.3 läuft.

Die Autoren von Artikeln wie „Examining the challenges of scientific workflows“ oder „A reference architecture for scientific workflow management systems and the view soa solution“ [Gil et al. 2007, Lin et al. 2009] haben eine ganze Reihe Kriterien herausgearbeitet anhand derer Scientific Workflow Management Systeme beurteilt werden können. Im Rahmen dieser Bachelorarbeit beschränken wir uns auf die vier wichtigsten, die uns zum Ziel führen tatsächlich unsere Daten analysieren zu können und SNPs herauszufiltern.

Zusätzlich zu generellen Anforderungen wie Skalierbarkeit, Zuverlässigkeit, Erweiterbarkeit, Verfügbarkeit und Sicherheit sind die folgenden Kriterien wesentlich für den Erfolg.

K1: Reproduzierbarkeit. Das vielleicht wichtigste Kriterium an ein ‚Scientific Workflow Management System‘ (SWfMS) ist die Reproduzierbarkeit wissenschaftlicher Analysen, Prozesse und Ergebnisse. Nur wenn diese gewährleistet ist, können Wissenschaftler gegenseitig die Gültigkeit ihrer Hypothesen beurteilen. Eine Untersuchung hat festgestellt, dass weniger als die Hälfte aller Microrray Experimente, die in Nature Genetics veröffentlicht worden sind, von zwei Analystenteams reproduziert werden konnten. Gründe dafür waren fehlende Rohdaten, fehlende Details der eingesetzten computergestützten Methoden sowie fehlende Informationen

über die Software und die Hardware [Ioannidis et al. 2008]. Fehlende Standards, immer größer werdende Datenmengen, wachsende Komplexität der eingesetzten Methoden sowie der Einsatz von zahlreichen Datenquellen und mehreren computergestützten Methoden erschweren die Reproduzierbarkeit weiter. Reproduzierbarkeit erfordert aussagekräftige Information zur Provenienz, so dass Forscher die eingesetzten Methoden kopieren können um wissenschaftlich vergleichbare Ergebnisse zu erhalten [Gil et al. 2007].

Für die Reproduzierbarkeit ist es unabdingbar, dass Unterbrechungen, beispielsweise durch einen Programmabsturz, nicht dazu führen, dass Zwischenergebnisse verschwinden, sondern es sollte möglich sein nach einer Unterbrechung am aktuellen Stand weiterzuarbeiten, ohne die vorangegangenen Schritte wiederholen zu müssen.

K2: Performanz. Heutzutage benötigen viele wissenschaftliche Probleme die Unterstützung von Hochleistungsrechenkapazitäten wie Grid oder Cloud Computing. Ein SWfMS sollte deswegen die wissenschaftlich orientierte, technologieunabhängige Umgebung von der zugrundeliegenden Computerinfrastruktur trennen. Auf diese Weise können sich Wissenschaftler auf ihre Domäne konzentrieren und gleichzeitig transparent den neusten Stand der Technik benutzen [Lin et al. 2009].

K3: Audittrail: Die Verfolgung und Aufzeichnung aller relevanten Aktivitäten und Fortschritte ist gerade für Workflows wichtig, die über einen großen Zeitraum hinweg laufen. Außerdem sind wissenschaftliche Workflows, da sie oftmals von Wissenschaftlern ad hoc entworfen oder modifiziert werden und verschiedene verteilte Aufgaben enthalten können, die wiederum per Netzwerkkommunikation aufgerufen werden, anfällig für Ausnahmen im Programmablauf und unerwartete Störungen von außen. Letztendlich stellt die Komplexität und der Umfang an Datenanalyse und Berechnungen in wissenschaftlichen Workflows eine zusätzliche Herausforderung für die Überwachung des Workflows, dessen Aufzeichnung und die Fehlerbehandlung dar [Lin et al. 2009].

K4: Konfiguration: Wissenschaftler benötigen einfach zu benutzende Programme, die für derart komplexe Möglichkeiten zur Gestaltung von Workflows intelligente Hilfestellungen anbieten. Die Automatisierung von simplen Aufgaben und Interaktionsmöglichkeiten, die unnötige Komplexität verbergen und die Sprache der Wissenschaftler sprechen, sind für den Erfolg ausschlaggebend [Gil et al. 2007].

4.3 Welches Programm erfüllt die Kriterien am besten?

Die im Methodikteil beschriebene Programmanalyse führte zu folgenden Ergebnissen:

Bpipe. Pipelines, die mit Bpipe aufgesetzt sind, ermöglichen durch die Textform der Skripte ein einfaches Wiederholen aller einmal durchgeführten Schritte mitsamt gleichen Parametern. Bei Unterbrechungen kann Bpipe von der Fehlerstelle neu gestartet werden, da Zwischenergebnisse gespeichert werden. Bpipe lässt sich nicht auf verteilten Rechenkapazitäten auslagern und benötigt deswegen unter Umständen längere Zeit für die Ausführung. Roller- oder Sicherheitssysteme sind nicht implementiert. Eine Trennung der wissenschaftlichen Umgebung von der Computerinfrastruktur findet kaum statt, es werden beispielsweise Dateiendungen versteckt und es ist möglich, Variablen anstatt festgelegter Namen und absoluter Pfade für Ein- und Ausgabe einzelner Schritte zu benutzen (siehe Anhang 2). Bpipe speichert den Verlauf aller ausgeführten Kommandos sowie deren Ein- und Ausgabe in einem automatisch erstellten Textdokument. Die Installation erfolgt per Shell und ist dem geringen Umfang entsprechend einfach gehalten [Sadedin et al. 2012].

Galaxy. Durch die Möglichkeit Workflows in einem Online oder lokalen Portal in einem Browser zu erstellen, auszuführen und freizugeben, ohne lokal Software installieren zu müssen, schafft es Galaxy, Ergebnisse eines Workflows reproduzieren zu können [Bux and Leser 2013].

Allerdings nutzt Galaxy im Gegensatz zu beispielsweise Taverna oder Bpipe keine explizite Darstellung, beispielsweise in XML, um Workflows zu spezifizieren. Stattdessen wird ein Workflow direkt auf der passenden Datenbank gespeichert. Der Austausch von Workflows zwischen zwei Nutzern wird dennoch durch passende Freigaben in den Benutzerkonten ermöglicht [Abouelhoda et al. 2010]. Galaxy speichert Metadaten um die Reproduzierbarkeit zu gewährleisten [Byelas et al. 2012]. Jede Benutzeraktion wird in der „History", einem zentralen Element in Galaxy, gespeichert. Dadurch entsteht die Möglichkeit, unabhängige Abfragen und Analyseschritte durchzuführen und diese danach in Galaxy zu kombinieren. Auch ist es möglich die erstellten Workflows in Galaxy weiterzuentwickeln, einzelne Schritte zu wiederholen oder Ergebnisdaten zu visualisieren. Operationen wie das Joinen oder die Schnittmenge oder Vereinigung zweier oder mehr Datentupel können über eine einfache Schnittstelle aufgerufen werden. Sowohl Grid- als auch Cloudcomputing werden unterstützt [Afgan et al. 2010]. Galaxy kann unter https://main.g2.bx.psu.edu/ online in vollem Umfang genutzt werden. So entfällt jegliche Konfiguration. Des Weiteren stehen dort umfangreichen Tools zur Datenanalyse zur Verfügung, so dass auch deren Installation wegfällt. Um Galaxy lokal zu nutzen, um beispielsweise sehr große oder sensible Dateien zu verarbeiten, müssen dennoch die entsprechenden Tools installiert werden.

Taverna. Der Ablauf von Workflows wird überwacht und aufgezeichnet [Shendure and Ji 2008]. Workflows von Benutzern können über die Webseite www.myexperiment.org veröffentlicht und heruntergeladen werden [McLennan et al.]. Auch im Programm selber gibt es die Möglichkeit, die Workflows von myExperiment in den eigenen Workflow zu integrieren. Taverna bietet nur begrenzt Möglichkeiten zur Parallelisierung und Benutzung von verteilten Rechenresourcen. Grid oder Cloud Computing werden ohne externe Programme nicht unterstützt [Bux and Leser 2013]. Die Möglichkeiten zur Parallelisierung und Nutzung von verteilten Computer-Ressourcen sind begrenzt, anspruchsvollere Methoden als eine Aufgabenwarteschlange und Nutzung von mehr als einer einzelnen lokalen Ressource fehlen [Hull et al. 2006, Curcin and Ghanem 2008, Bux and Leser 2013]. Jedes Experiment beinhaltet eine detaillierte Methodenbeschreibung, die sowohl zur Ergebnisvalidierung als auch der Wiederverwendung der Methoden dient. So kann ein Workflow wie ein *in silico* Experiment genutzt werden. Web Services können über die Angabe der URL des WSDL Dokuments eingebunden werden [Oinn et al. 2006, Talia 2013].

Kepler. Kepler unterstützt die parallele Ausführung von Pipelines und Multithreading auf einem lokalen Client [Bux and Leser 2013]. In Kepler repräsentieren geordnete Bäume die Ergebnisse von Workflows [Anand et al. 2009]. Diese Bäume werden im XML Format aufgezeichnet und Kepler erlaubt nun das Navigieren in der Ergebnishistorie [Byelas et al. 2012].

SWfMS	Bpipe	Galaxy	Taverna	Kepler
Reproduzierbarkeit	+	+	+	+
Performanz	−	+	−	+
Audittrail	+	+/−	+	+/−
Konfiguration	+/−	+	+/−	+

Tabelle 2 Erfüllung des Kriterienkatalogs durch die einzelnen Programme

Auswahl. Tabelle 2 stellt das Ergebnis der Evaluation dar. Es ergibt sich ein heterogenes Gesamtbild, da die Programme stark unterschiedliche Ursprünge und Ziele haben. Leider konnte kein Programm alle Kriterien in vollem Maße erfüllen. Auf der anderen Seite weist kein Programm gravierende Unzulänglichkeiten auf, was auf die grundsätzliche Eignung aller Programme schließen lässt. Betrachtet man die Erfüllung der Kriterien der evaluierten Tools fällt auf, dass sich diese nur gering unterscheiden. Galaxy und Kepler erreichen sogar die gleiche

Bewertung. Ein Grund dafür könnte die ähnliche Ausrichtung der Funktionalität sein. Auch liegen die Unterschiede meist im Detail. Beispielsweise bietet Galaxy schon ohne zusätzliche Installation viele Algorithmen zur Analyse der Sequenzdaten an. Es fällt deshalb schwer eine objektive Auswahl zwischen den Produkten zu treffen. Bpipe und Taverna werden jedoch wegen der mangelhafter Performanz und Konfiguration nicht in die engere Auswahl einbezogen.

Kepler bietet ohne zusätzliche Module wie auch bioKepler keine native Unterstützung bioinformatischer Analysen [Altintas 2011]. Zwar ist es sehr variabel und breit gefächert in den Anwendungsgebieten, doch die große Abstrahierung eines Workflows macht es unnötig komplex einen Workflow für den Zweck einer bioinformatischen Analyse zu erstellen.

Da Galaxy speziell für biomedizinische Forschung entwickelt wurde, ist es das für unsere Anforderungen am ehesten geeignete Tool. Es ist Kepler in vieler Hinsicht überlegen. Beispielsweise lassen sich Datenbankabfragen direkt in Galaxy absetzen ohne komplexe SQL-Befehle schreiben zu müssen oder überhaupt eine Datenbank aufrufen zu müssen. Gegenwärtig gibt es in Galaxy drei Kategorien im Umgang mit Daten: Anfrageoperationen, Sequenzanalysetools und Datenanzeige. Die erste Kategorie umfasst übliche Mengenoperationen wie beispielsweise Vereinigung, Schnittmenge, Subtraktion und Komplementbildung. Sequenzanalysetools sind eigenständige Module, die biologisch orientierte Analyseschritte wie das Finden von orthologen Regionen, das Alignment von zwei oder mehr Sequenzen oder das Berechnen des GC-Gehalts, da ein hoher GC-Gehalt auf ein DNA-Abschnitt hinweist, die für Gene kodieren. Schließlich erlaubt die grafische Darstellung von Daten in verschiedenen Formen einen schnellen Überblick und lässt so Rückschlüsse auf die Qualität zu. Beispielhaft sei hier die Betrachtung des Alignments (siehe Kap. 4.5) genannt. Mittels des UCSC Table Browsers kann der Nutzer aus vielen verfügbaren Genomen und Regionen auswählen und die gesuchte Sequenz in die Galaxyhistory laden.

Die Integration neuer Tools ist abhängig vom Tool selbst. Tools, die in Toolshed, dem Onlineverzeichnis für Tools in Galaxy (siehe Kap. 4.6), verfügbar sind, können sehr simpel in wenigen Schritten in Galaxy integriert werden. Für viele weitere Programme oder andere als die in Toolshed angebotenen Versionen gibt es Installationsanleitungen im Wiki der Galaxy Homepage wiki.galaxyproject.org. Um ein komplett neues Programm in Galaxy zu integrieren, reicht eine Spezifikationsdatei, die beschreibt, wie das Programm aufgerufen wird und Parameter angegeben werden. Durch die abstrakte Arbeitsweise von Galaxy kann die Schnittstelle des jeweiligen Tools automatisch erstellt werden. Durch diese Verfahrensweise von Galaxy wird das Toolverhalten so spezifiziert, dass Transparenz und Reproduzierbarkeit eingehalten werden können [Goecks et al. 2010].

Obwohl die von Galaxy gespeicherten Metadaten bei der Ausführung der Tools ausreichen um eine Analyse zu wiederholen, reichen sie nicht aus, um den Sinn einer Analyse zu verstehen. Dazu gibt es die Möglichkeit, Annotationen an Analyseschritte anzufügen. Diese Notizen stellen einen wichtigen Aspekt in der Reproduzierbarkeit dar, da sie dem Schöpfer des Workflows die Möglichkeit geben, zu erklären, weshalb ein bestimmter Schritt wichtig ist. Damit zeigen die Metadaten an, was getan wurde, während die Annotationen das Warum erklären.

Mit der Entscheidung für Galaxy als Scientific Workflow Management System wurde die Basis für das Durchführen einer bioinformatischen Pipeline geschaffen. Der folgende Teil dieser Arbeit widmet sich deswegen mit der Installation des Programms und nötigen Komponenten sowie der Konzeption und Durchführung der Pipeline.

4.4 Installation des Programms

Um Galaxy auf einem lokalen Server zu installieren, muss dieser Python Version 2.6 oder 2.7 anbieten. Die aktuellste stabile Version kann danach folgendermaßen auf einer Kommandozeile heruntergeladen werden:

```
trr@portalmoritz:~> hg clone https://bitbucket.org/galaxy/galaxy-dist/
trr@portalmoritz:~> cd galaxy-dist
trr@portalmoritz:~/galaxy-dist> hg update stable
0 files updated, 0 files merged, 0 files removed, 0 files unresolved
```

Galaxy benötigt zum Starten Konfigurationsdateien und pythonspezifische Module. Diese werden allerdings beim ersten Starten des Servers erstellt. Deswegen reicht es den folgenden Befehl im Ordner der Galaxy Distribution auszuführen um den Server zu starten:

```
trr@portalmoritz:~/galaxy-dist> sh run.sh
```

Nach der erfolgreichen Durchführung des Startskripts kann Galaxy über einen Internetbrowser aufgerufen werden. Ohne Änderung an den Einstellungen wird der Server auf localhost und Port 8080 gestartet, also kann Galaxy unter http://localhost:8080 erreicht werden. Der Startbildschirm ist in Anhang 1 zu sehen. Um den Galaxy Server zu beenden, reicht es die Tastenkombination `Strg+c` in der Kommandozeile zu betätigen. Die auf dem TRR Server installierte Version von Galaxy ist die neueste Version „release_2013.08.12“ die verfügbar ist.

4.5 Wie soll die Pipeline aufgebaut und konfiguriert sein?

Wie in Abbildung 2 gezeigt, ist der nächste Schritt der Analyse nach dem Durchlauf des Sequenzierers die Beurteilung der Qualität der Reads und deren Entfernung, Kürzung oder Korrektur – falls Qualitätskriterien nicht eingehalten werden.

Zur Qualitätsbeurteilung gibt es verschiedene Programme. Die Exomdaten liegen im FastQ Format vor [Cock et al. 2010]. Dieses wird benutzt um neben der Nukleotidsequenz die korrespierende Qualität und Metadaten zur Sequenzierung zu speichern. FastQC [Andrews 2010] produziert aus den Reads einen Bericht anhand dem die Qualität der Reads feststellbar ist.

Anhang 4 FastQC Report für Samplesequenz zeigt die Qualität der ersten 10.000 Reads eines Exoms aus Tumorgewebe. Der Report wurde auf dem Testserver in der lokalen Ausführung von Galaxy mit einem integrierten Tool, das FASTQ Zusammenfassungen erstellt, unter Version 0.52 erstellt. Die Qualität ist in diesem Fall sehr gut, eine Qualität nach dem Phred-

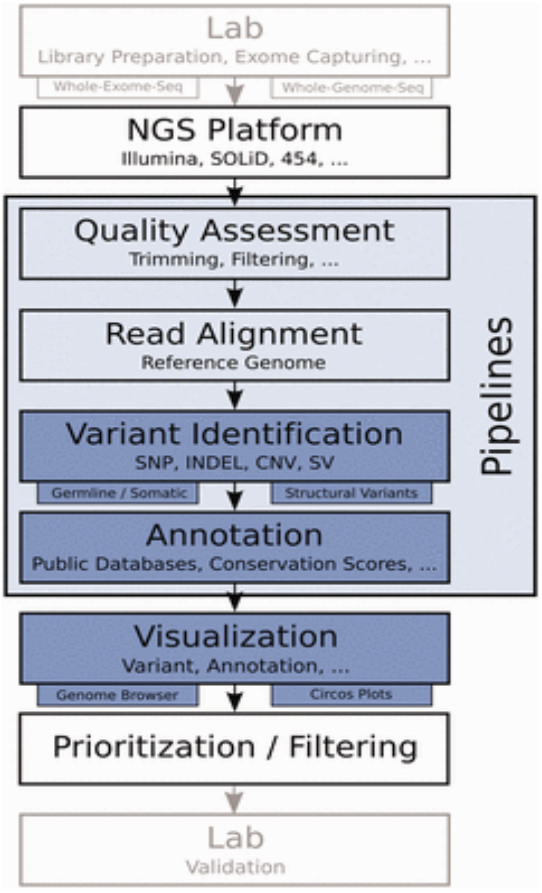


Abbildung 2: Workflow zur Analyse des gesamten Exoms/Genoms aus [Pabinger et al. 2013]

Qualitätsscore von 30 bedeutet, dass diese Base in 1 aus 1000 Fällen falsch abgelesen worden ist.

Danach werden die Reads an ein menschliches Referenzgenom aligniert. Hierzu wurde BWA [Li and Durbin 2009] installiert und damit das Refernzgenom indexiert. Das Referenzgenom, in diesem Fall grCh37.fa, liegt im FASTA Format vor. Dies ist ein simples Format, das nach einer Zeile an Metadaten die Sequenzdaten als Zeichenfolge enthält. Die Indexierung wird folgendermaßen gestartet:

```
bwa index –a bwtsw grCh37.fa
```

Um ein nicht vorinstalliertes Tool, in diesem Fall BWA, in Galaxy nutzen zu können, muss dieses entweder nach einer Anleitung von der Galaxy Webseite (wiki.galaxyproject.org/Admin/Config/Tool%20Dependencies) manuell in die lokale Galaxyinstanz eingefügt werden oder per Toolshed (toolshed.g2.bx.psu.edu/), ein Speicherort für den Austausch von Galaxytools zwischen den weltweit verteilten lokalen Galaxyinstanzen, automatisch heruntergeladen und installiert werden. Beide Möglichkeiten wurden für BWA erfolgreich für die lokale Instanz auf dem TRR Server durchgeführt. Bei der manuellen Installation muss darauf geachtet werden, dass Galaxy das Tool nicht ohne Weiteres aufrufen kann. Hierzu muss die Datei tool_conf.xml im Ordner der Galaxy Distribution folgendermaßen bearbeitet werden:

```
<section name="NGS: Mapping" id="solexa_tools">
  <tool file="sr_mapping/bwa_wrapper.xml"/>
```

Ein Nachteil der Benutzung von Toolshed besteht darin, dass manchmal keine Auswahl in der Version des benötigten Tools besteht. Dies führte im Beispiel von BWA dazu, dass trotz der komfortablen Nutzung von Toolshed auf die manuelle Installation von BWA zurückgegriffen werden musste, da die von Toolshed angebotene Version von BWA 0.5.9 einige Bugs enthielt, die nach dem Alignieren zu Fehlern in der weiteren Analyse führten.

Damit BWA das Referenzgenom aufrufen kann, müssen die Dateien bwa_index.loc und bwa_index_color.loc im /galaxy-dist/tool-data/ Verzeichnis an die lokale Galaxyinstanz angepasst weden, indem nach den jeweiligen Anweisungen in den Dokumenten eine Zeile pro Referenzgenom hinzugefügt werden muss, damit BWA den Speicherort des Referenzgenoms aufrufen kann.

Hilfreich zur Lösung der aufgetretenen Schwierigkeiten bei der Installation ist die Support Homepage von Galaxy galaxyproject.org/Support und speziell die Verteilerlisten von Galaxy wiki.galaxyproject.org/MailingLists.

Das eigentliche Alignment mit BWA erfolgt in Galaxy. Dazu muss nach dem Upload der Sequenzdaten darauf geachtet werden, dass die Dateien mit dem „Edit Attributes“ (Stift Symbol) Menü in das „fastqsanger“ Format umgewandelt werden müssen, damit BWA die Dateien erkennt. Unter „NGS:Mapping“, „Map with BWA for Illumina“ links im Menü, findet sich das Alignment Programm. Das Referenzgenom sollte als „built-in index“ markiert sein, falls dies nicht der Fall ist, ist die Installation von BWA unvollständig oder fehlgeschlagen. Nun muss noch die zu analysierende Sequenzdatei als „FASTQ Datei“ ausgewählt werden. Auch ist es hier möglich Parameter für das Programm anzugeben. Im Fall von BWA ist das beispielsweise „-O INT Gap open penalty [11]“, also Strafpunkte für das Öffnen einer Lücke. Die Erhöhung dieses Werts führt in der Ergebnisdatei zu weniger Lücken, allerdings kann dann das Alignment verfälscht werden. Der Wert in den eckigen Klammern ist der Defaultwert dieses Parameters. In diesem Fall ist er elf. In dieser Form gibt es für das Alignment achtzehn verschiedene Paramter, die alle sorgfältig gewählt werden müssen, da sonst die Qualität des Alignments nicht optimal wird.

Das Ergebnis des Alignments ist eine SAM, Sequence Alignment/Map, Datei [Li et al. 2009]. Das SAM-Format ist ein Textformat, in dem pro Zeile ein Alginment festgehalten wird. Eine Zeile enthält 11 Pflichtrubriken, die durch Tabstopps getrennt werden, wie beispielsweise der Name, die Mappingqualität, die gemappte Sequenz und Angaben über eventuelle Insertionen oder Deletionen. Diese Datei kann mit den Integrative Genomics Viewer (IGV) [Thorvaldsdóttir et al. 2013] betrachtet werden, indem in Galaxy auf das Diskettensymbol geklickt wird und die Datei heruntergeladen und manuell geöffnet wird oder indem IGV geöffnet wird, und dann in Galaxy auf „IGV“, „display with local“ geklickt wird. Es werden die gemappten Reads angezeigt und eventuell auftretende Nukleotidpolymorphismen hervorgehoben. Die Visualisierung der Daten hilft die Qualität der Analyse zu überblicken, ermöglicht einen schnellen Blick auf bestimmte Gene sowie den Vergleich mehrerer Datensätze. Des Weiteren ist es möglich SNPs zu betrachten. Erkennbar in Abbildung 3 ist auch noch das Prinzip der „Paired End Reads“, also zwei Reads, die in unterschiedlicher Richtung gelesen worden sind (entgegengesetzte Richtung der Pfeilspitzen) aber über den bekannten Abstand der beiden zusammengehören. Auch zu sehen ist, dass sich die Reads überlappen, idealerweise gibt es eine mindestens 20-fache Überlappung, diese ist in der Abbildung nur zweifach, da die Datenmenge reduziert wurde. Darüber hinaus sind einige SNPs zu betrachten, beispielsweise ist an Position 110.045.740 bp die Base im grünen Read eine andere als im blauen und im Referenzgenom. Zudem steht am unteren Rand der Name des betrachteten Gens „RAD23B“, welches zu Saccharomyces cerevisiae Rad23 kodiert, ein Protein das in der Nukleotid-Reparatur involviert ist. Nicht im Bild zu sehen ist die Angabe über das Chromosom auf dem das Gen liegt, in diesem Fall Chromosom 9.

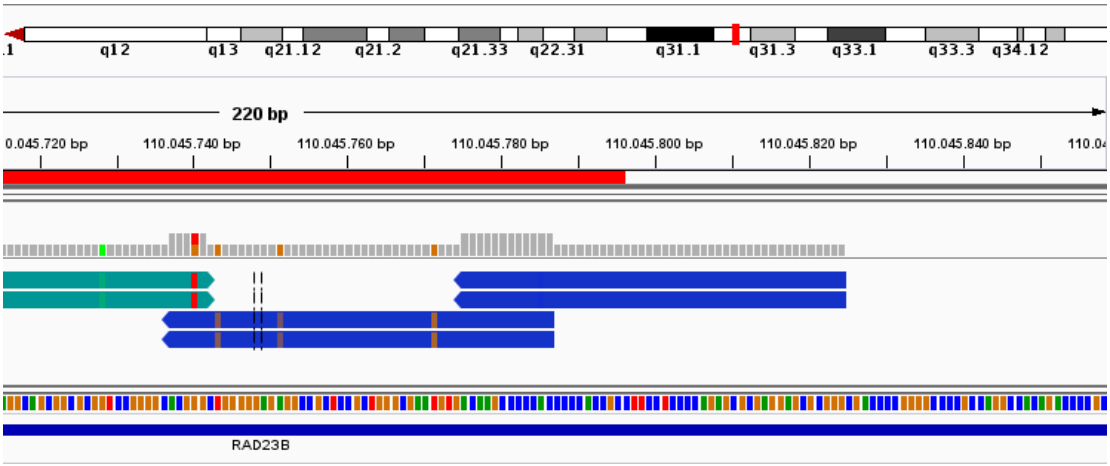


Abbildung 3 Sequenz Alignment im Integrative Genomics Viewer

Wegen inhärenten Fehlern der Sequenzierung kann es zu Duplikaten in den Reads kommen. Diese PCR Duplikate werden mit Hilfe von Picard (ww.picard.sourceforge.net) entfernt, da sie sonst zu überrepräsentierten Allelen führen könnten. Dieses Program ist in Galaxy vorinstalliert und unter „NGS: Picard (beta)“, „Mark Duplicate reads“ zu finden.

Zur Detektion von Varianten eignen sich Tools wie das Genome Analysis Toolkit [McKenna et al. 2010] oder SAMtools. Diese Programme sind nicht vorinstalliert, müssen also mit obiger Anleitung nachträglich installiert werden.

Die häufigste Form der Variantenannotation ist der Abgleich mit öffentlichen Datenbanken wie dbSNP, auf deren Grundlage eine Vorhersage der Funktion gestellt wird. snpEFF [Cingolani et al. 2012] ist ein bekanntes Programm, das in Galaxy integriert ist.

Weiter wurde die Pipeline nicht getestet, da weitere Schritte, beispielsweise Tools wie der Unified Genotyper des Genome Analysis Toolkits weitere Installationen und Konfiguration benötigten, aus denen kein weiterer Informationsgewinn möglich wäre.

4.6 Pipeline Ausführung auf den Leberkrebs Daten.

Galaxys zentrale Oberfläche ist die „Analyze Data“ Seite (siehe Anhang 1). Auf der linken Seite ist eine Liste aller verfügbaren Tools und Datenverarbeitungsschritte. Das Hauptfenster in der Mitte wird zum Setzen von Toolparametern oder zur Dateneingabe sowie zum Absenden von Tools benutzt. Auf der rechten Seite ist das Historyfenster in dem hochgeladene Daten und die Ergebnisse durchgeführter Schritte in chronologischer Reihenfolge als einzelne Ereignisse aufgelistet sind. Über die Navigationsleiste sind die Galaxys wesentliche Komponenten wie der Analyse-Arbeitsbereich, die Workflowsliste, veröffentlichtes Material, das Hilfemenü und Benutzereinstellungen erreichbar.

Das Erstellen und wiederholte Ausführen eines Workflows erfordert die folgenden Schritte:

1. Erstellen und Bearbeiten einer History, die das Ziel der Analyse erfüllt;

2. Automatisches Erstellen eines Workflows anhand der History;

3. Ausführung des erstellten Workflows zur wiederholten Analyse für verschiedene Eingabedaten oder variable Parameter.

Die Analyseschritte sind nachdem die gewünschten Parameter eingetragen worden sind oder die Defaulteinstellungen übernommen worden sind nach dem Abschicken rechts im Ereignisfenster zu sehen. Danach kann per „Extract Workflow“ aus einer Abfolge von Eingabedaten und Analyseschritten ein Workflow erstellt werden. Ein beispielhafter Workflow, der zwei Sequenzdateien einliest, diese mit BWA an ein Referenzgenom aligniert, dann die SAM Ausgabe in ein binäres Format konvertiert und abschließend Duplikate entfern ist in Abbildung 4 (kompletter Screenshot in Anhang 5 Extrahierter Workflows) zu sehen.

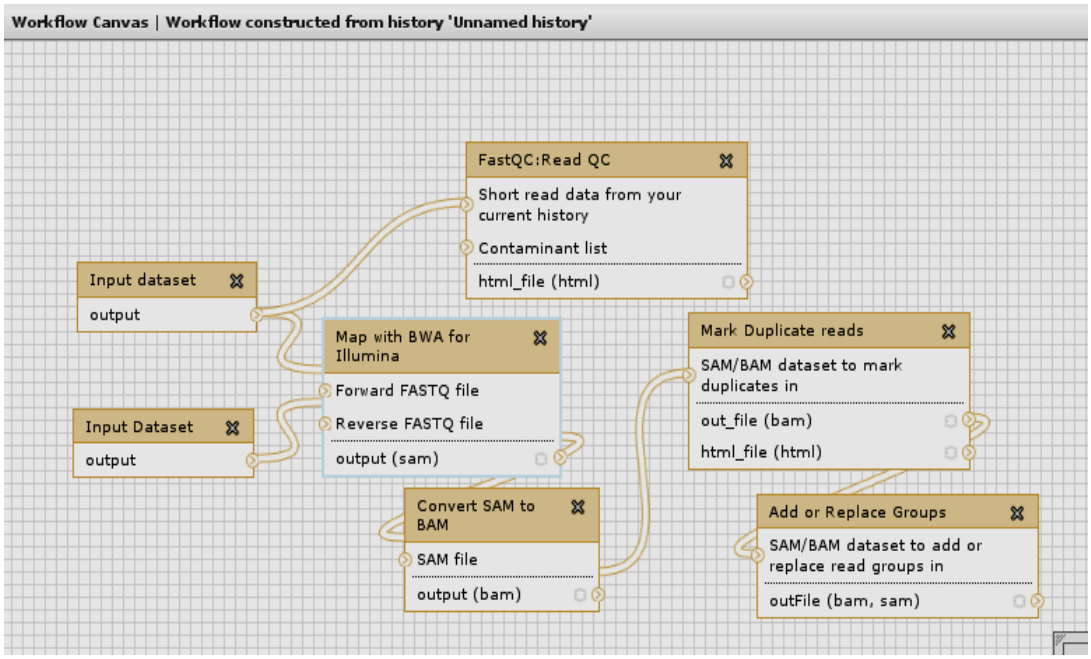


Abbildung 4 Automatisch erstellter Workflow aus History

5 Diskussion

Im Folgenden wird das Erreichen der Ziele dieser Arbeit untersucht und hinsichtlich der Vorgehensweise kritisch betrachtet.

5.1 Diskussion des Vorgehens

Die Literaturrecherche dient der Suche nach Scientific Workflow Management Systemen, der Suche nach möglichen Auswahlkriterien für diese und der Beantwortung inwieweit die Programme die Kriterien erfüllen. Die Festlegung auf bereits entwickelte Programme erleichtert die Auswahl auf ein Programm massiv, da die Entwicklung eines neuen Programms sehr aufwendig ist. Nachteilig ist bei einem fertigen Programm die fehlende oder unvollständige Anpassung an die projektspezifischen Anforderungen. Eine andere Herangehenseweise ist beispielsweise eine grafische Oberfläche für ein skriptbasiertes System.

Die benutzten Suchbegriffe könnten erweitert oder verfeinert mehr Ergebnisse liefern. Es wurde sich im Rahmen dieser Arbeit aber auf wenige wichtige beschränkt.

5.2 Diskussion der Ergebnisse

1.1. Welche Programme stehen zur Auswahl?

Erste Aufgabe der vorliegenden Arbeit war es, zur Verfügung stehende Scientific Workflow Management Systeme zu finden. Dazu wurde zunächst eine Internet- und Literaturrecherche durchgeführt um potenzielle Lösungen zu finden. Aus diesen Programmen wurden die vier am ehesten geeigneten in Kapitel 4.1 im Einzelnen und Argumente für ihre Berücksichtigung vorgestellt. Andere Programme eignen sich für unsere Ansprüche nicht, da sie beispielsweise andere Zielvorgaben haben.

1.2. Welche funktionalen und nicht-funktionalen Kriterien muss das Programm erfüllen?

Kapitel 4.2 beschreibt die vier genutzten Kriterien und deren Inhalte im Detail. Die Auswahl der Kriterien orientiert sich an vorhergehenden Untersuchungen, berücksichtigt aber die spezifischen Anforderungen für das Leberkrebsprojekt. Insbesondere finden Anforderungen an die Interoperabilität und an die Benutzerschnittstelle keine Berücksichtigung, dafür wurde mit der Konfiguration ein weiteres Kriterium eingeführt.

1.3. Welches Programm erfüllt die Kriterien am besten?

Als Folge des nicht eindeutigen Ergebnisses musste eine Entscheidung außerhalb der genannten Kriterien erfolgen. In Zukunft könnte dies durch die Einführung neuer Kriterien oder die Differenzierung der bestehenden Kriterien vermieden werden.

Es wurde eine Tabelle aufgestellt, in der die Bewertung der vier Softwarelösungen in übersichtlicher Form dokumentiert ist. In der ersten Spalte sind die Kriterien aufgelistet. In den nachfolgenden Spalten werden die jeweiligen Kandidaten in den Grad der Erfüllung eingeteilt. Die Auswertung der Tabelle allein erlaubte noch keine Entscheidung für eine Software, da zwei von ihnen die gleiche Bewertung erreichen konnten. Es wurden daher weitere praktische Erwägungen, wie die Unterstützung für bioinformatische Aufgaben hinzugezogen. Aus diesen Überlegungen wurde Galaxy als zu verwendendes Scientific Workflow Management System ausgewählt.

Galaxy ist gezielt für die benötigten Schritte geeignet. Der Grund hinter der Entwicklung von Galaxy war die ‚informatics crisis' in den Biowissenschaften: Computergestützte Ressourcen sind mitunter schwer zu verwenden und die Kommunikation mit computergestützten Experimenten und deren Reproduzierbarkeit eine große Herausforderung. Zur Bewältigung dieser Krise bietet Galaxy eine offene, webbasierte Plattform für die Durchführung nachvollziehbarer, reproduzierbarer und transparenter Genomforschung [Goecks et al. 2010].

Äußerst positiv ist demnach zu vermerken, dass Galaxy praktisch ohne Schulungen genutzt werden kann. Es ist in der Verwendung größtenteils selbsterklärend und oftmals intuitiv benutzbar.

1.4. Installation des Programms.

Der Installationsprozess von Galaxy ist in Kapitel 4.4 nachzulesen. Es konnte gezeigt werden, dass Galaxy recht leicht installiert und benutzt werden kann. Es ist frei verfügbar und Open Source und kann bei lokaler Installation und Ausführung an die speziellen Anforderungen des jeweiligen Servers angepasst werden. Die Funktionalität von Galaxy ist in einem Webbrowser verfügbar. Prinzipiell kann Galaxy ohne Anpassung sofort genutzt werden. Allerdings müssen einige Programme installiert und Einstellungen für den produktiven Einsatz angepasst werden. Hierfür gibt es eine Online Dokumentation, durch die der Nutzer geleitet wird.

Standardmäßig verwendet Galaxy SQLite damit die Nutzung von Galaxy ohne die Installation und Konfiguration einer Datenbank funktioniert. Um Galaxy deutlich effizienter und mit verringerten Risiken, beispielsweise eines Deadlocks, zu verwenden, wird das Wechseln zu PostgreSQL empfohlen. Auch ist der integrierte HTTP Server mit Unsicherheiten verbunden, die durch das Umstellen auf einen Apache Proxy unterbunden werden. Des Weiteren wird im Projekt auf Dauer die Nutzung eines Rechenclusters geplant, der die Rechenleistung deutlich verbessern würde.

1.5. Wie soll die Pipeline aufgebaut sein?

Die Hintergrundinformationen zur DNA Sequenzierung und der Analyse liefern die theoretischen Grundlagen während Kapitel 4.5 den Aufbau unserer bioinformatischen Pipeline vorstellt. Des Weiteren wird hier ein Leitfaden zur Verfügung gestellt, der die Installation neuer Tools in Galaxy erläutert. Es konnte gezeigt werden, dass sich neue Programme in Galaxy integrieren lassen und somit in den Workflow eingebaut werden können.

Eine besondere Herausforderung war die Einbindung des Referenzgenoms. Im Kapitel 4.5 werden dieses und andere Probleme identifiziert und praktische Lösungsansätze geliefert.

1.6. *Pipeline Ausführung auf den Leberkrebsdaten.*

Galaxys Hauptaufgabe ist die Unterstützung von Ärzten und Biologen bei der Untersuchung von biomedizinischen Daten. Durch die Komplexität der DNA-Sequenzanalyse und Sequenzvergleichen kann Galaxy bei der Vermeidung von Fehler helfen, die bei der manuellen Durchführung auftreten könnten. Auch wird der Aufwand erheblich reduziert, der normalerweise von Nöten wäre, um die Schritte einzeln auszuführen und um eine Pipeline wiederholt auszuführen. Eine prototypische bioinformatische Pipeline wurde erstellt und auf den Leberkrebsdaten ausgeführt. Die Ergebnisse beispielsweise das Alignment wurden in Kapitel 4.5 dargestellt, die Pipeline, die in der lokalen Galaxyinstanz erstellt wurde, ist in Kapitel 4.6 abgebildet.

5.3 Dikussion der formulierten Ziele

Ziel 1: .Auswahl eines geeigneten Pipeline Management Tools zum komfortablen Ausführen und Managen von bioinformatischen Pipelines.

Das Scientific Workflow Management System Galaxy wurde nach sorgfältiger Evaluation von vier potenziellen Lösungen für am geeignetsten erklärt, biomedizinische Pipelines im Projekt SFB/TRR 77 zu organisieren und auszuführen. Im Bereich der Open Source SWfMS findet sich noch eine Vielzahl weiterer Anbieter, die jedoch ausgeschlossen wurden.

Das in der Arbeit verwendete Verfahren einer Kriterientabelle ist im Bereich der Softwareevaluation gängig. Das Auffinden von geeigneten Programmen ist aus einer Literaturrecherche entstanden. Die Kriterien der Evaluation sind spezifisch für das TRR77 Projekt ausgewählt worden. Diesbezüglich bereits publizierte Auswahlverfahren wie von [Lin et al. 2009] konnten dabei lediglich als Orientierung dienen und mussten angepasst werden. Die ausgewählten Kriterien sind deswegen außerhalb des Projekts nicht oder nur in modifizierter Form gültig. Die persönliche Überprüfung der Erfüllung aller Kriterien war nicht möglich, weswegen die Informationen aus Publikationen über die Produkte entnommen werden mussten. Dies ist gegenüber Informationen aus den Produkdokumentationen ein Vorteil, da die Meinungen dort verzerrt sein können.

Im Endergebnis wiesen die ermittelten Bewertungen der Produkte eine geringe Varianz auf: Es gab zwei erste und zwei zweite Plätze. Als Folge des nicht eindeutigen Ergebnisses musste eine Entscheidung außerhalb der genannten Kriterien erfolgen. In Zukunft könnte dies durch die Einführung neuer Kriterien oder die Differenzierung der bestehenden Kriterien vermieden werden. Kriterien die letztlich zur Auswahl von Galaxy als Scientific Workflow Management System führten, wie Benutzerfreundlichkeit, Umfang an vorinstallierten Tools und Anbindung an biologische Datenbanken sollten in die Ermittlungsmatrix aufgenommen werden.

Ziel 2: Exemplarische Etablierung einer Pipeline, die computergestützt Genomdaten analysiert und zur Identifikation von SNPs führt.

Nach der Installation wurde Galaxy so konfiguriert, dass es für die Ausführung der Pipeline genutzt werden konnte. Dazu gehörte die Installation von spezifischen Tools wie BWA und grundlegende Einstellungen wie das Hinzufügen eines Adminstrators als Benutzerrolle. Die erfolgreiche Durchführung der Pipeline hat gezeigt, dass das Programm in der Praxis einsetzbar ist.

Kritisch zu bewerten sind die Schwierigkeiten im Zusammenhang mit der Integration neuer Tools. Die Versionsabhängigkeiten machen es schwierig, direkt die richtige Lösung zu finden und so dauert die Auswahl und Integration neuer Programme teilweise sehr lange.

Damit das System produktiv eingesetzt werden kann, müssen unter anderem Konzepte zur Datensicherheit, Benutzerprofile- und gruppen, Performanzverbesserung durch die Nutzung von Servercluster sowie eventuelle projektübergreifende Anforderungen umgesetzt werden.

Grenzen der Arbeit.

Die vorliegende Arbeit hatte nicht zum Ziel, Genomforschung zu betreiben, die zu therapeutischen Maßnahmen führen könnte. Es sollte ein Programm ausgewählt werden, das die Bedienung und Ausführung von bioinformatischen Pipelines vereinfacht. Des Weiteren sollte eine einfache Pipeline prototypisch durchgeführt werden.

Auch wurde nicht untersucht, ob es neben Scientific Workflow Management Systemen eine bessere Plattform zur Untersuchung von Next Generation Sequencing Daten gibt. Obwohl in Kapitel 3.2 einige Argumente aufgeführt wurden, die für die Nutzung von SWfMS sprechen, gibt es durchaus Alternativen, zum Beispiel vorgefertigte Skripts, die die einzelnen Schritte nacheinander aufrufen. Allerdings hat dies Nachteile die in Kapitel 1 angesprochen wurden.

Generell gibt es im Bereich der biomedizinischen Informatik noch viele Herausforderungen, die mit der vorliegenden Arbeit Berührungspunkte aufweisen. Galaxy sowie die meisten Scientific Workflow Management Systeme befinden sich noch immer in der Entwicklung, es gibt keinerlei fertiggestellte Komplettsoftware für die Zwecke der Genomforschung. Auch im Bereich der Genomforschung gibt es zwar viele Herangehensweisen und Ansätze, doch ist die Forschung in diesem Bereich weit davon entfernt, aus Daten direkt Ergebnisse ableiten zu können.

6 Ausblick

Die Möglichkeit Workflows und deren Ergebnisse im SFB/TRR 77 Projekt zu veröffentlichen fördert das gegenseitige Verständnis der jeweiligen Arbeit. Darüber hinaus können Workflows auch der gesamten Wissenschaft zur Verfügung gestellt werden, wodurch es möglich wird, dass Forscher der gesamten Welt, die sich für die jeweilige Fragestellung interessieren, die veröffentlichten Vorgehensweisen und Ergebnisse betrachten, nachvollziehen oder modifzieren können.

Das Scientific Workflow Management System ist ein Basisprodukt zur Beantwortung bioinformatischer Fragen und kann in absehbarer Zeit in den vielen medizinischen Einzelprojekten im SFB Projekt zur Anwendungen kommen. Immer dann, wenn bioinformatische Datenanalysen benötigt werden, dient Galaxy als einfach zu bedienendes Werkzeug zur Erstellung und automatisierten Durchführung von Pipelines. Auch projektübergreifende Anforderungen können mit Galaxy gelöst werden, da Forscher ihre erstellten Workflows teilen können und dann weitere Forscher den Workflow weiter ausbauen können. Neue Tools oder aktualisierte Versionen von Programmen können meist in wenigen Schritten in die lokale Instanz von Galaxy eingebunden werden.

Mit Galaxy geht die großangelegte Genomforschung neue Wege und bietet erstmals umfangreiche Analysen ohne vorher eingehende Programmierkenntnisse und Fähigkeiten im Umgang mit Datenbanken besitzen zu müssen. In absehbarer Zukunft können Forscher des SFB/TRR 77 ‚Leberkrebs – von der molekularen Pathogenese zur zielgerichteten Therapie' Projekts in der Galaxy Arbeitsumgebung mit einem einfach zu bedienenden Werkzeug bioinformatische Pipelines erstellen und ohne Programmierkenntnisse automatisiert ausführen und aus den so gewonnenen Erkenntnissen einen wertvollen Beitrag zur Erforschung einer der tödlichsten Krankheiten unserer Zeit liefern.

Literatur

Abouelhoda M, Alaa S, Ghanem M (2010): *Meta-workflows: pattern-based interoperability between Galaxy and Taverna*. Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science, ACM.

Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J (2010): Galaxy CloudMan: delivering cloud compute clusters. *BMC bioinformatics* **11**(Suppl 12): S4.

Altintas I (2011): *Distributed workflow-driven analysis of large-scale biological data using biokepler*. Proceedings of the 2nd international workshop on Petascal data analytics: challenges and opportunities, ACM.

Anand MK, Bowers S, McPhillips T, Ludäscher B (2009): *Efficient provenance storage over nested data collections*. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, ACM.

Andrews S (2010): FASTQC. A quality control tool for high throughput sequence data. *URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc*.

Bux M, Leser U (2013): Parallelization in Scientific Workflow Management Systems. *arXiv preprint arXiv:1303.7195*.

Byelas H, Dijkstra M, Swertz MA (2012): *Introducing Data Provenance and Error Handling for NGS Workflows within the MOLGENIS Computational Framework*. BIOINFORMATICS.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X (2012): Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in genetics* **3**.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010): The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* **38**(6): 1767-1771.

Curcin V, Ghanem M (2008): *Scientific workflow systems-can one size fit all?* Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, IEEE.

Ding L, Wendl MC, Koboldt DC, Mardis ER (2010): Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum Mol Genet* **19**(R2): R188-96.

Ganzinger M, Noack T, Diederichs S, Longerich T, Knaup P (2011): Service oriented data integration for a biomedical research network. *Studies in health technology and informatics* **169**: 867.

Gil Y, Deelman E, Ellisman M, Fahringer T, Fox G, Gannon D, Goble C, Livny M, Moreau L, Myers J (2007): Examining the challenges of scientific workflows. *Computer* **40**(12): 24-32.

Goecks J, Nekrutenko A, Taylor J, Team TG (2010): Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**(8): R86.

Heinritz W (2004). *Molekulargenetische Mutationsanalyse des APC-Gens mittels DHPLC bei Patienten mit Familiärer Adenomatöser Polyposis (FAP)*, Tenea.

Hempel M, Haack T, Eck S (2011): „Next generation sequencing "–neuer Zugang zur molekularen Aufklärung und Diagnostik von Stoffwechseldefekten. *Monatsschr Kinderheilkd* **9**.

Hohpe G, Woolf B (2012). *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*, Pearson Education.

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T (2006): Taverna: a tool for building and running workflows of services. *Nucleic acids research* **34**(suppl 2): W729-W732.

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G (2008): Repeatability of published microarray gene expression analyses. *Nature genetics* **41**(2): 149-155.

Knust E, Janning W (2008). *Genetik: Allgemeine Genetik - Molekulare Genetik - Entwicklungsgenetik*, Thieme.

Li H, Durbin R (2009): Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009): The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Lin C, Lu S, Fei X, Chebotko A, Pai D, Lai Z, Fotouhi F, Hua J (2009): A reference architecture for scientific workflow management systems and the VIEW SOA solution. *Services Computing, IEEE Transactions on* **2**(1): 79-92.

Mardis ER, Wilson RK (2009): Cancer genome sequencing: a review. *Human molecular genetics* **18**(R2): R163-R168.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**(9): 1297-1303.

McLennan M, Clark S, Deelman E, Rynge M, Vahi K, McKenna F, Kearney D, Song C Bringing Scientific Workflow to the Masses via Pegasus and HUBzero. *parameters* **13**: 14.

Meyerson M, Gabriel S, Getz G (2010): Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11**(10): 685-696.

Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T (2010): Visualizing genomes: techniques and challenges. *Nat Methods* **7**(3 Suppl): S5-S15.

Oinn T, Greenwood M, Addis M, Alpdemir MN, Ferris J, Glover K, Goble C, Goderis A, Hull D, Marvin D (2006): Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience* **18**(10): 1067-1100.

Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V (2010): Ergatis: a web interface and scalable software system for bioinformatics workflows. *Bioinformatics* **26**(12): 1488-1492.

Otto TD, Vasconcellos EA, Gomes LH, Moreira AS, Degrave WM, Mendonca-Lima L, Alves-Ferreira M (2008): ChromaPipe: a pipeline for analysis, quality control and management for a DNA sequencing facility. *Genet Mol Res* **7**(3): 861-71.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z (2013): A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*.

Perens B (1999): The open source definition. *Open sources: voices from the open source revolution*: 171-85.

Richards JE, Hawley RS (2010). *The Human Genome: A User's Guide, Third Edition*, Elsevier Science.

Sadedin SP, Pope B, Oshlack A (2012): Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28**(11): 1525-1526.

Schaaf CP, Zschocke J (2013). *Basiswissen Humangenetik*, Springer-Verlag GmbH.

Shaer O, Kol G, Strait M, Fan C, Grevet C, Elfenbein S (2010). G-nome surfer: a tabletop interface for collaborative exploration of genomic data. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**:** 1427-1436.

Shaer O, Mazalek A, Ullmer B, Konkel M (2013). From big data to insights: opportunities and challenges for TEI in genomics. Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction**:** 109-116.

Shendure J, Ji H (2008): Next-generation DNA sequencing. *Nat Biotechnol* **26**(10): 1135-45.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001): dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**(1): 308-311.

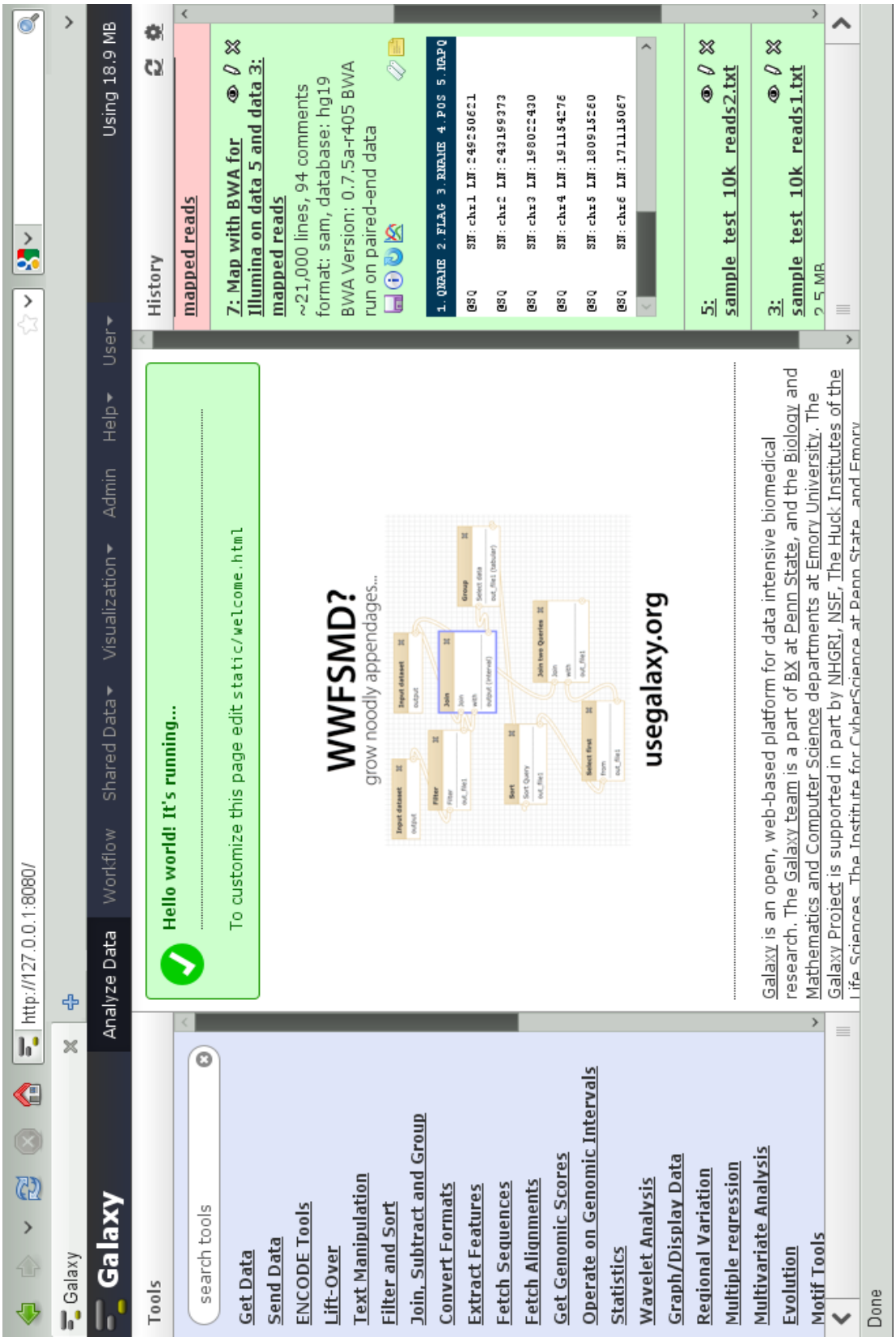
Talia D (2013): Workflow Systems for Science: Concepts and Tools. *ISRN Software Engineering* **2013**.

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013): Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**(2): 178-192.

Wetterstrand KA (2011): DNA sequencing costs: data from the NHGRI large-scale genome sequencing program. *Accessed November* **20**: 2011.

Anhang

Anhang 1 Galaxy Start

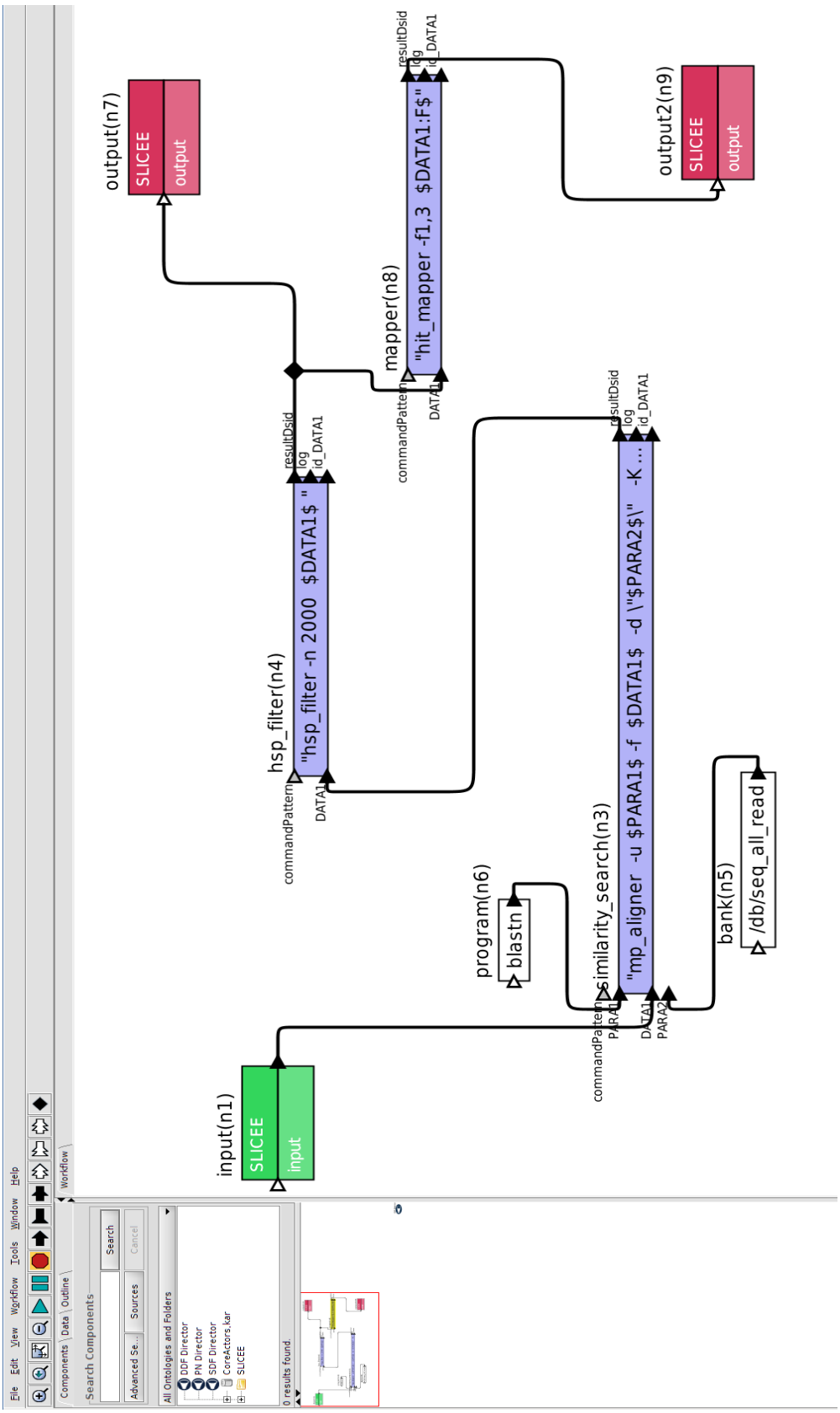


Anhang 2 Bpipe Sample Pipeline

```
REFERENCE="/genedata/human genome GRCh37/hg19.fa"
PICARD_HOME="/home/trr/picard-tools-1.93/picard-tools-1.93"
BWA_HOME="/home/trr/bwa-0.7.5a"

seq1="/genedata/sample_test_10k_reads/sample_test_10k_reads1.txt.gz"
seq2="/genedata/sample test 10k reads/sample test 10k reads2.txt.gz"

//readgroup information:
rg_id="lane712s006433"
rg_lb="nextera"
rg pl="ILLUMINA"
rg pu="flowcell-barcode.lane"
rg_sm="GDNA"

//###############################################################
//Alignment
//###############################################################

//BWA
@Transform("sai")
align bwa = {
      exec "$BWA HOME/bwa aln -t 8 -q 10 $REFERENCE $input > $output.sai"
}

@Transform("sam")
sampe_bwa = {
      exec "$BWA HOME/bwa sampe -P -r
'@RG\tID:$rg id\tLB:$rg lb\tPL:$rg pl\tPU:$rg pu\tSM:$rg sm' $REFERENCE $input1.sai
$input2.sai $seq1 $seq2>$output"
}

//###############################################################
//SAM to BAM
//###############################################################

@Transform("bam")
sort = {
        exec "samtools view -bS $input | samtools sort -o - - > $output"
}

//###############################################################
//Remove Duplicates
//###############################################################

@Filter("dedupe")
dedupe = {
        exec """
            java -Xmx1g -jar $PICARD_HOME/MarkDuplicates.jar
                            MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
                            METRICS FILE=out.metrics
                            REMOVE DUPLICATES=true
                            ASSUME SORTED=true
                            VALIDATION_STRINGENCY=LENIENT
                            INPUT=$input
                            OUTPUT=$output
        """
}
//###############################################################
//Run Pipeline
//###############################################################

Bpipe.run {
    "sample_test_10k_reads%" * [align_bwa] + sampe_bwa + sort + dedupe
}
```

```
bpipe run sample_test_pipeline.pipe sample_test_10k_reads*
```

Anhang 3 Kepler Workflow (www.vapor.gforge.inria.fr/)



Anhang 4 FastQC Report für Samplesequenz

FastQC Report

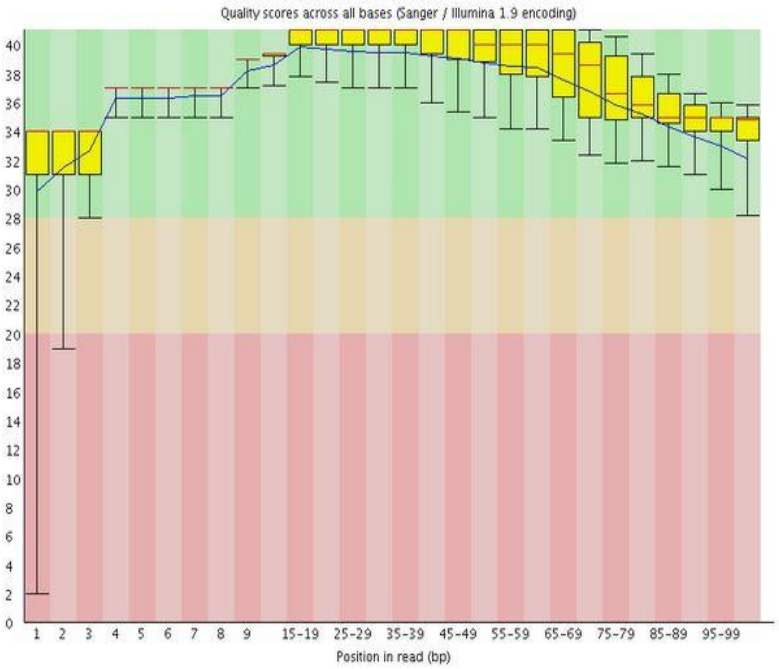
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

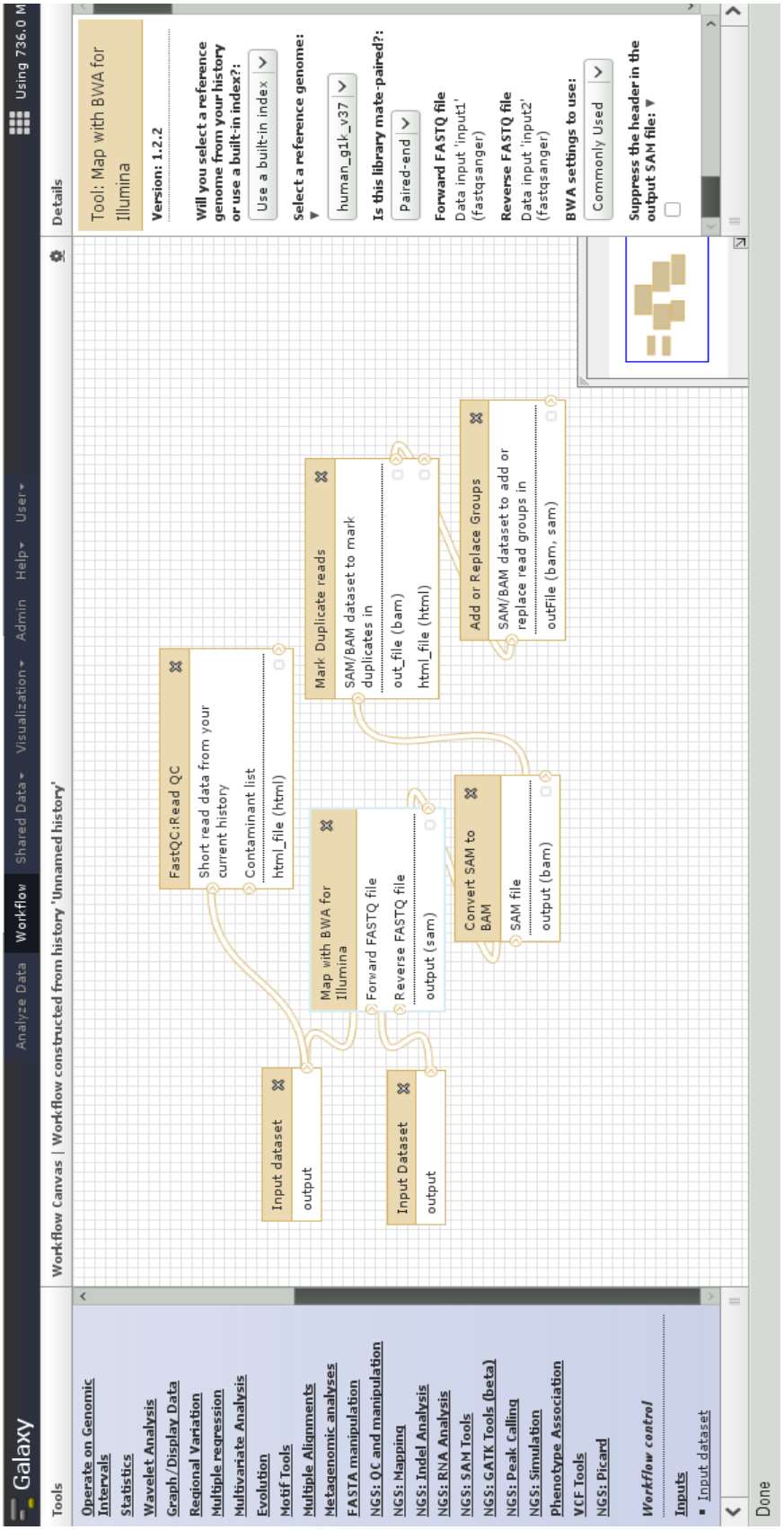
Basic Statistics

Measure	Value
Filename	sample_test_10k_reads1.txt
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10000
Filtered Sequences	0
Sequence length	104
%GC	39

Per base sequence quality



Anhang 5 Extrahierter Workflows



Tabellenverzeichnis

Abbildungsverzeichnis

Eidesstattliche Versicherung

Hiermit erkläre ich eidesstattlich, dass die vorliegende Arbeit von mir selbständig und ohne unerlaubte Hilfe angefertigt worden ist und dass ich alle Stellen, die wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen sind und alle Informationen, die aus interviewähnlichen Gesprächen gewonnen wurden, als Zitate gekennzeichnet habe.

Ort, Datum Unterschrift