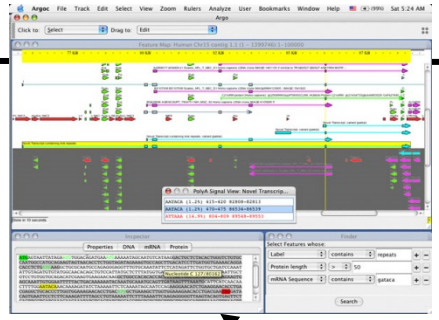
Semesterprojekt Verteilte Echtzeitrecherche in Genomdaten

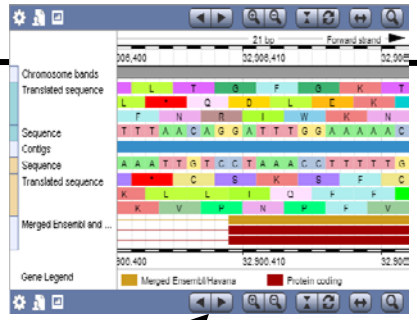
Ulf Leser, Marc Bux, Stefan Sprenger

Allgemein: Semesterprojekt

- Kernidee: Erstellen eines mittel-komplexen Programms im Team über ein komplettes Semester hinweg
- Ziele
  - Arbeiten in Teams: Gruppendynamik, Absprachen, Überflieger, ...
  - Softwareentwicklungsprozess von Anfang bis Ende miterleben
  - Anwenden von Tools der professionellen Softwareentwicklung
- Kein Ziel: Erstellen eines perfekten, umfangreichen Systems

Genombrowser





Distributed Mutation Data Store

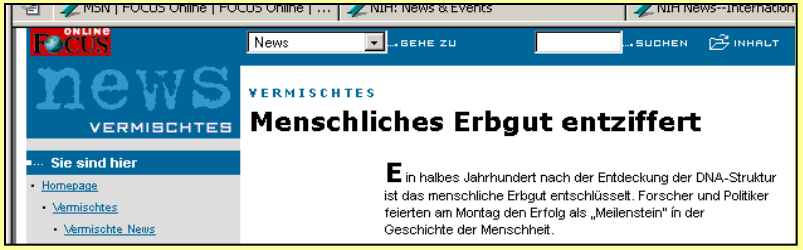
Middleware

HGMD®
The Human Gene Mutation Database
Cardiff

THE CANCER GENOME ATLAS

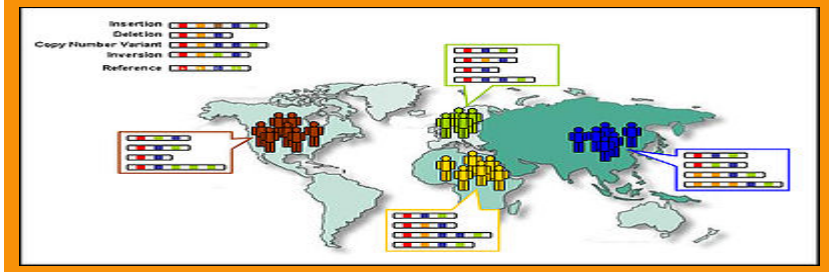
Fast Development

1953
Double helix structure of DNA, Watson/Crick



2003
First human genome sequenced
Took ~14 years, ~3 billion USD

2008
Genome of J. Watson finished
4 Months, 1.5 Million USD



2010
1000 Genomes Project

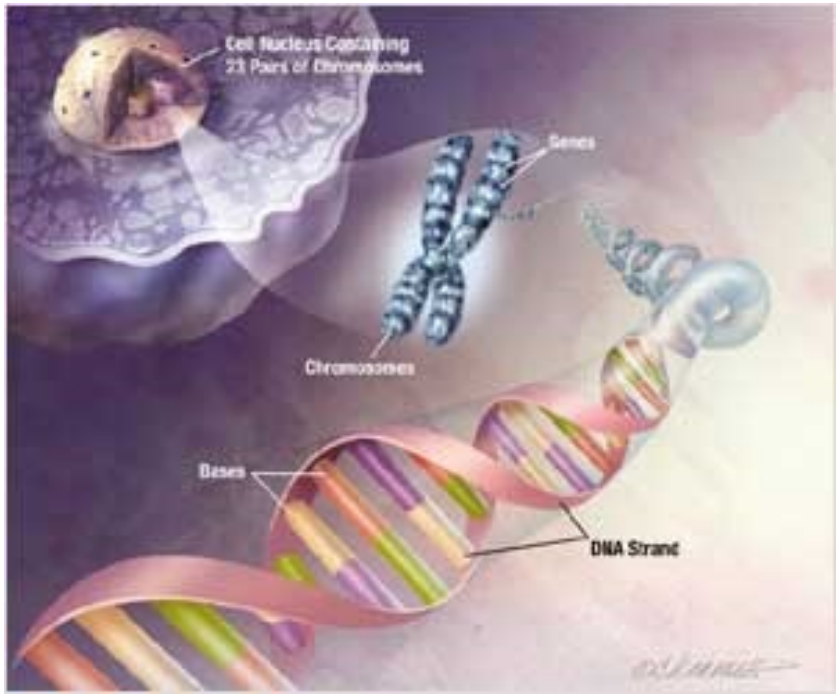
Large Scale Sequencing Projects

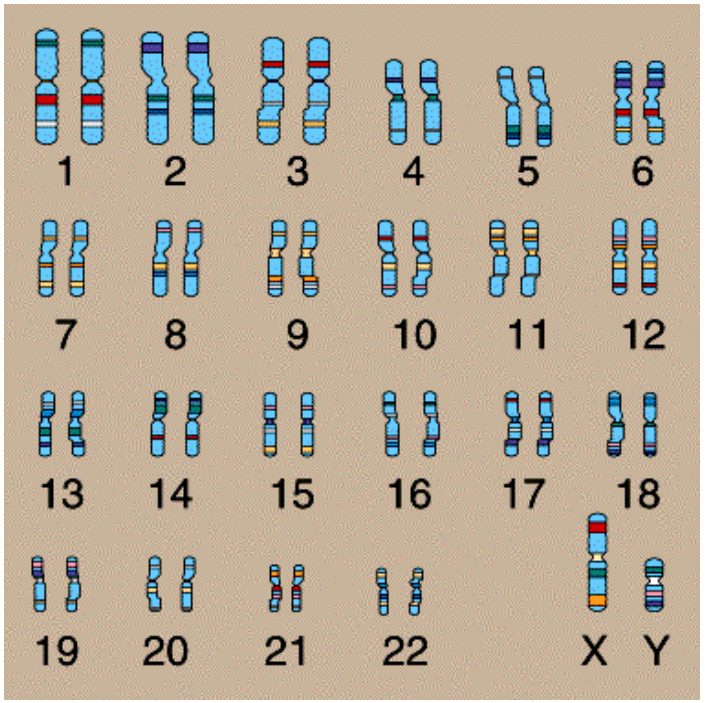
50,000 samples: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

Genomics England ... is creating a lasting legacy for patients, the NHS and the UK economy through the sequencing of 100,000 genomes: the 100,000 Genomes Project.

The Veterans Affairs (VA) Office of Research and Development is launching the Million Veteran Program (MVP) .... The goal of MVP is to better understand how genes affect health and illness in order to improve health care.

Genome ≈ String





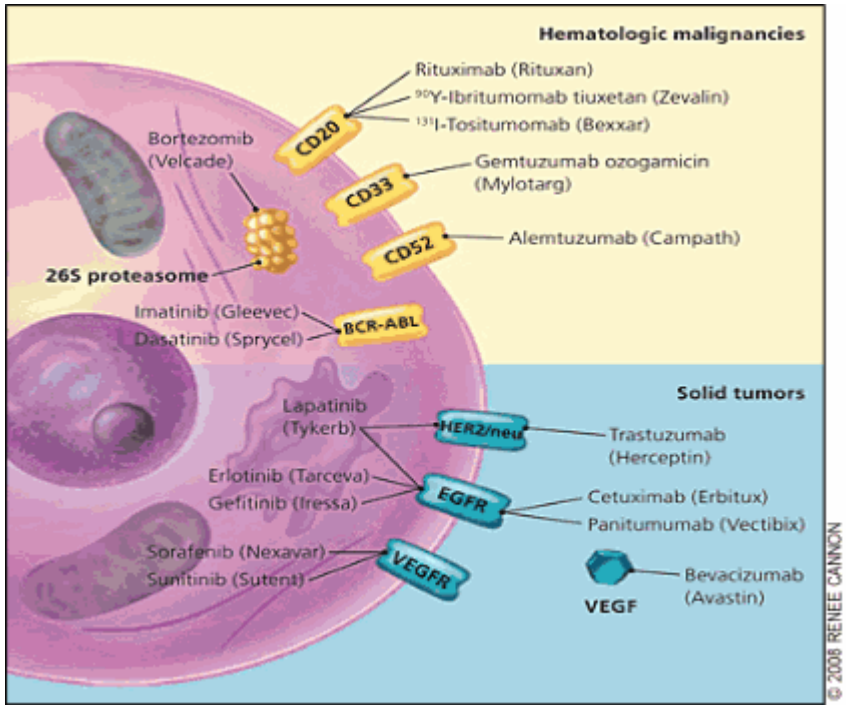
- Human genome: app. 3.000.000.000 letters ∈ {A,C,G,T}

Genomics in a Nutshell

- ~2% are coding – genes being translated into proteins
  - Whole Genome Sequencing – WGS
  - Whole Exome Sequencing - WES
- ~20.000 genes, forming maybe 2M different proteins
  - ~3000 genes are conserved since ever (yeast)
  - We share ~95% of our genes with mice, rats, dogs, ...
  - ~25% of our genes have a still unknown function
- It's not only genes: miRNA, enhancer, binding sites, chromatin structure, epigenomics, ...

Genomics for Medicine

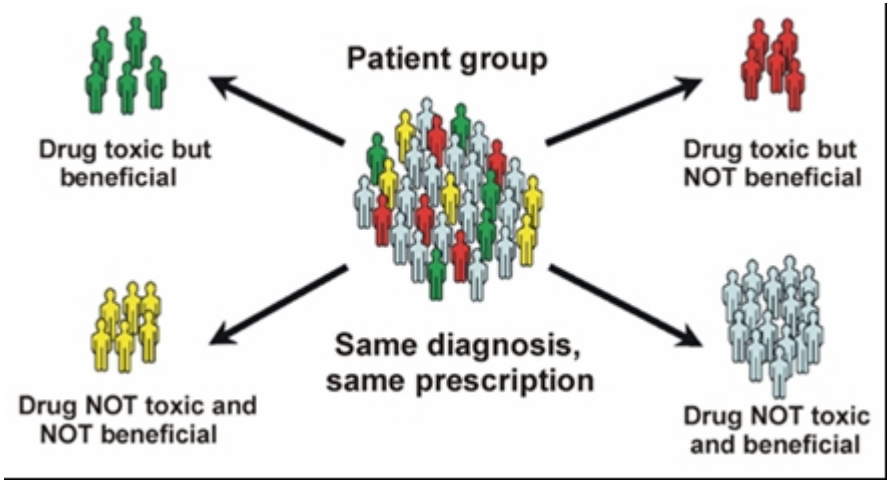
- Cancer, immunology, genetic diseases, infections
- Cancer
  - Cells proliferating uncontrolled, leaving their tissue
  - What goes wrong? Cell division, DNA repair, surface adhesion, cellular signaling
  - ~200 core cancer genes



- Targeted therapy: Drug attacking a mutated gene
- “Cancer is becoming a chronic diseases”

Precision Medicine, Personalized Medicine

- Tailor treatment to the individual patient's genome
- "Genome" – mutation profile
  - We know 10s of Millions of human mutations
  - Mutation – deviation from the norm?
  - Mutation – genomic subsequence rarely seen
- Requires many genomes
  - What is rare?
  - Often enough to obtain a statistically robust association
  - Most effects involve many mutations / genes
  - Combinatorial explosion

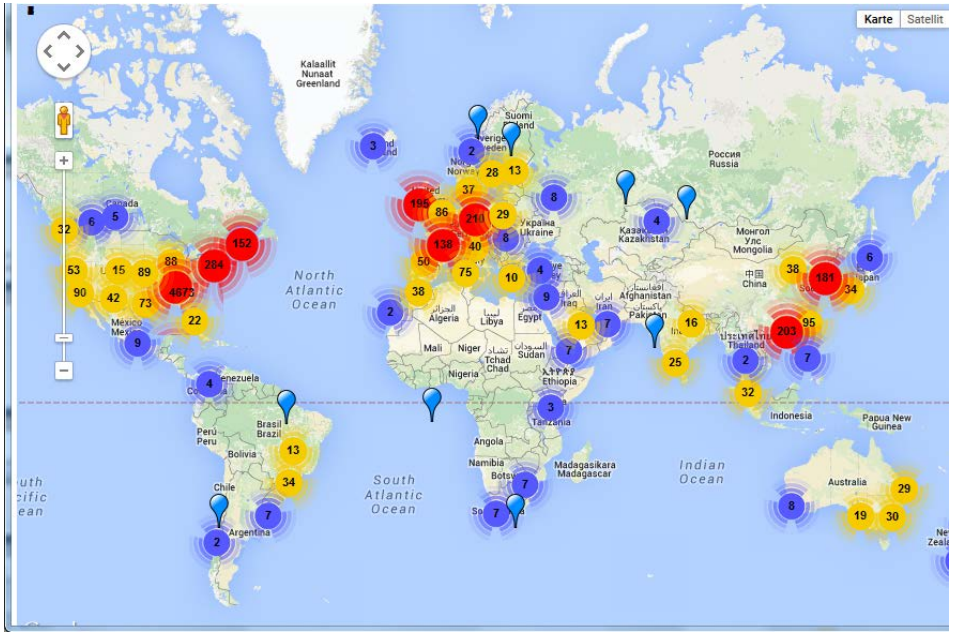


Next Generation Sequencing

- New generation of sequencers since ~2005
  - Illumina, Solexa, 454, Solid, ...
- Much higher throughput
  - ~15 TB raw data in 3-5 days
  - ~600 GB processed data/week
  - Cost for sequencing a genome down to ~2.000 USD
- 3rd generation sequencers
  - Single molecule sequencing
  - A (human) genome in a day
  - Sequence every human
  - Sequence different cells in every human

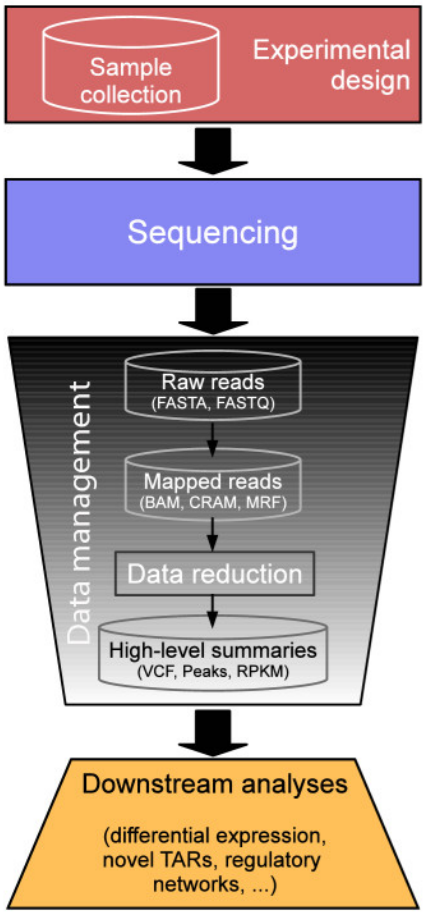
Illumina HiSeq 2000. DNAVision

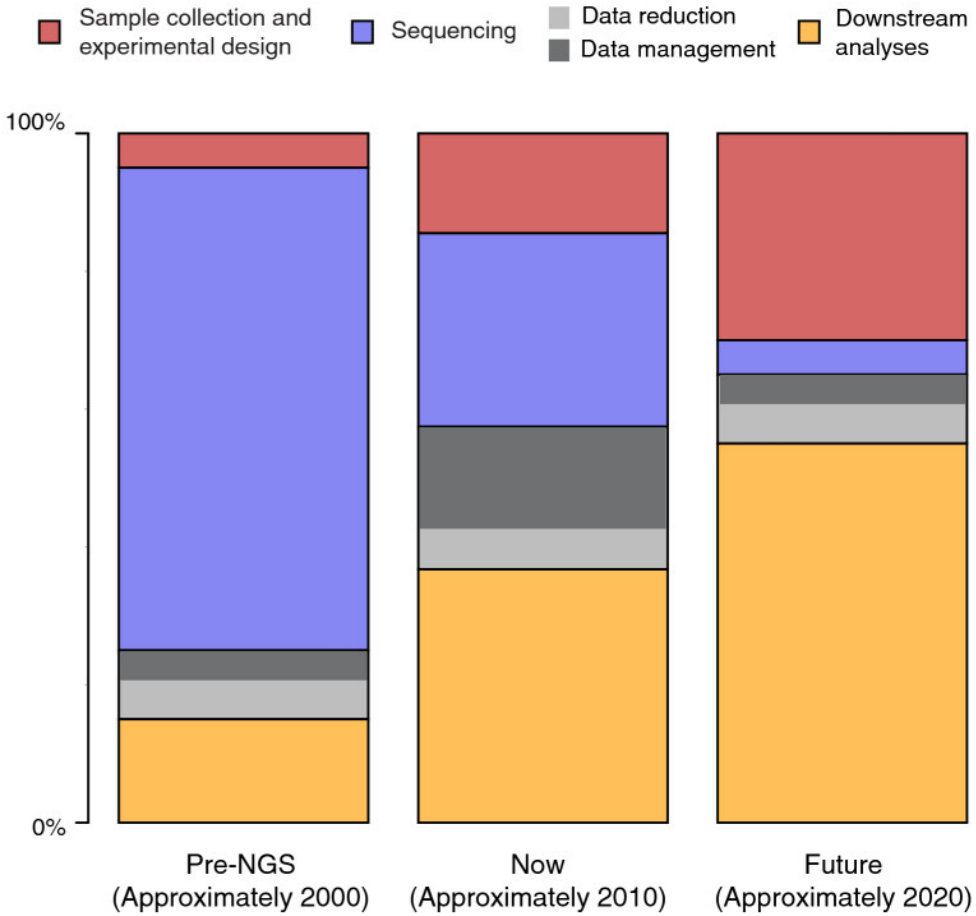
Sequencing becomes a commodity



- Sequencing dozens or hundreds of genomes is feasible (now!) for any mid-size research projects

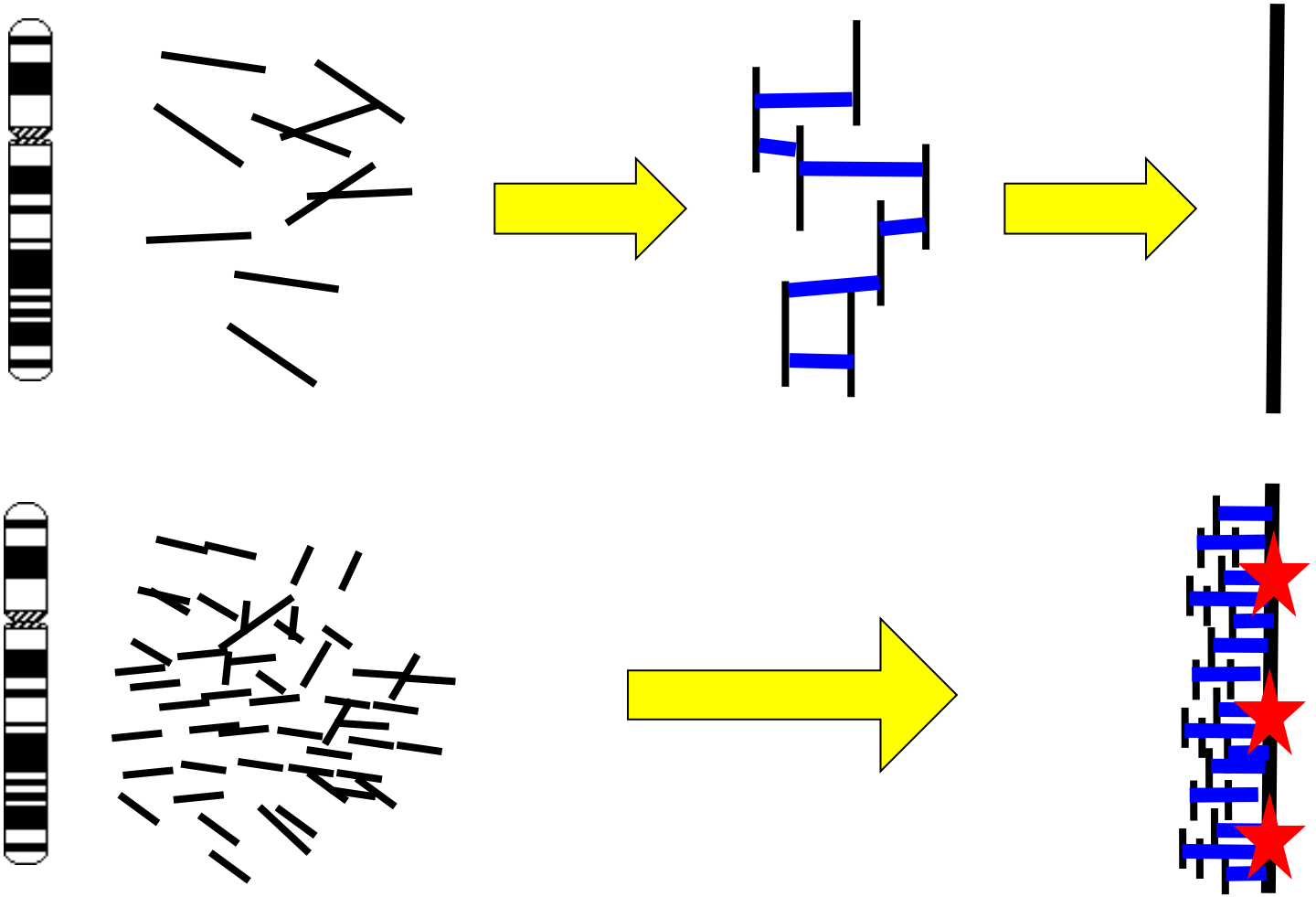
The „real“ Cost of Genomic Sequencing



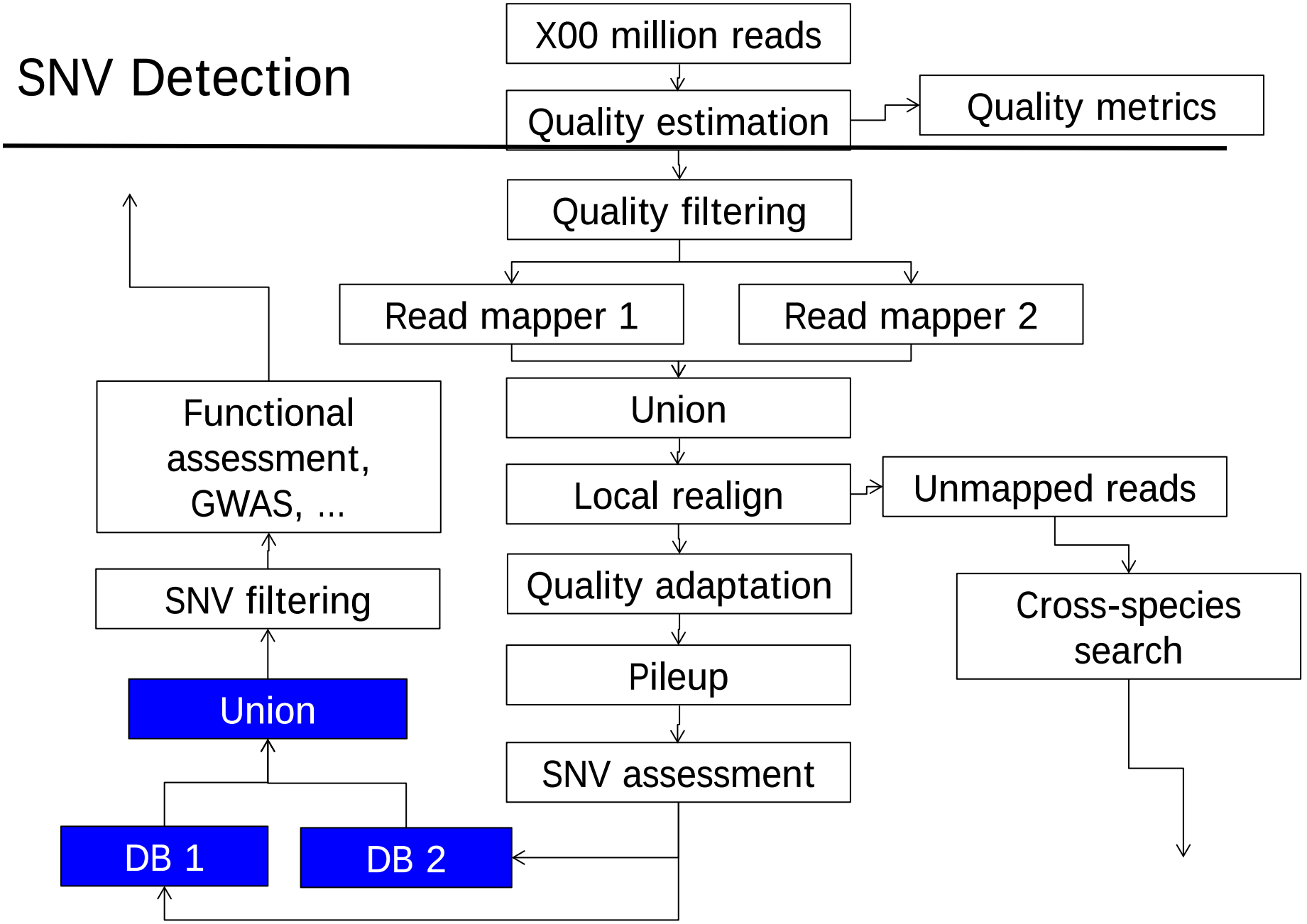


Sboner, A. (2011). The real cost of sequencing: higher than you think! Genome Biology 2011

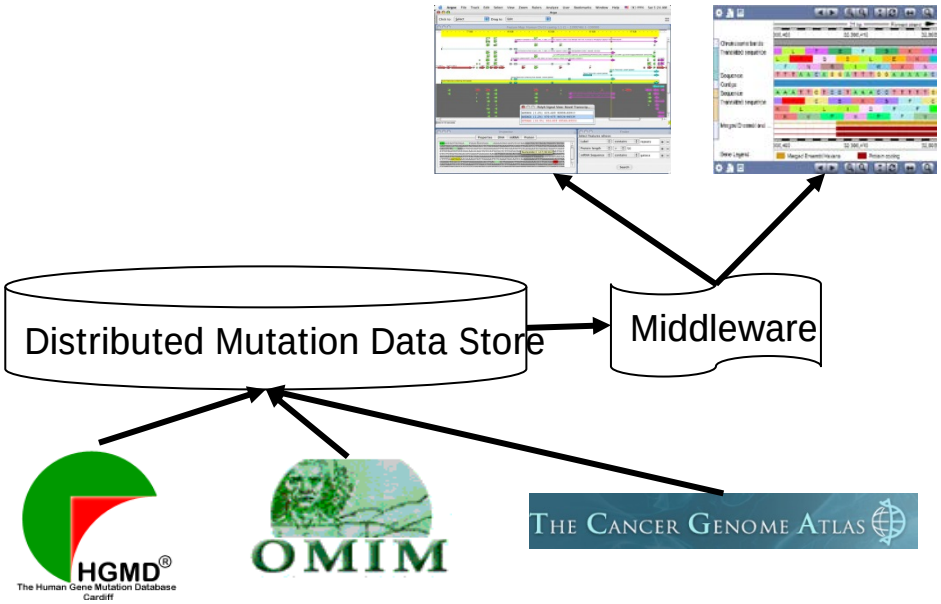
New Task: Read Mapping & SNV Detection



SNV Detection



Technischer Fokus



- (Kontinuierliche) Integration heterogener Daten in ein gemeinsames Schema
- Hochperformanter hauptspeicherbasierter verteilter Index zur Suchunterstützung
- Visualisierung in „Echtzeit“ (< 1sec für 95% aller Anfragen)

Ablauf

- Wir übernehmen die Kundenrolle: Angabe, *was* wir wollen, nicht, wie es umgesetzt wird
- Sie übernehmen die Anbieterrolle: Spezifikation des Funktionsumfangs, Aufteilung in Teilprojekte, Modellierung, Implementierung, Test, Abnahme
  - Wir moderieren
- Typischerweise: Drei Teilprojekte plus Projektleitung
- Engmaschiger Plan mit Abgaben, die diskutiert werden
- Begleitung durch SVN/Git und TRAC

Begleitprogramm

- An 5 Terminen werden zusätzlich Hintergrund und Anwendungswissen in 45-90 Minuten Vorträgen präsentiert
  - Benutzung von SVN/GIT und TRAC (Bux)
  - Human Mutation Databases und deren Integration (Leser)
  - Grundlagen verteilter Programmierung (Bux)
  - Hauptspeicherindexstrukturen (Sprenger)
  - Webprogrammierung (Starlinger)

Ablauf Überblick

- Oktober
  - Spezifikation des Funktionsumfangs, Formulierung von Use Cases
  - Aufteilung in Teilprojekte und Teams
  - Festlegung benutzter Technologien (alle)
  - Benennung zentraler Schnittstellen (alle)
  - Festlegung der Datenquellen
  - Definition Style-Guide (PL)
  - Projektplanung inkl. Meilensteine (Kunden und PL)
- November
  - Festlegung der Teilprojektschnittstellen (insb. PL)
  - Planung / Modellierung der Teilprojekte
  - Planung Testfälle (Teilprojekte und PL)
  - Mock-Ups der Benutzeroberfläche (entsprechendes TP)

Überblick 2

- Dezember
  - Implementierung der Teilprojekte
  - Implementierung und Durchführung der Testfälle
- Januar
  - Integrationstests
  - Fehlerdokumentation und Zuweisung
  - Bug Fixing
  - Code Dokumentation
  - Abnahme
- Februar
  - Präsentationen, Projektdokumentation (Wiki), Poster
  - Projektabschluss

Teilnehmer

Name	Vorerfahrung (Java, Web, DB, Projekte, Middleware)
Krause, Gabriel	
Schumacher, Ben	Java, C++
Wackerbauer, Martin	
Rose, Nikita	Haskell
Rebscher, Lucas	Java, App-Entwicklung
Papke, Robin	Java, Web
Löffler, Tobias	
Bauer, Martin	
Atanasov, Aleksanda	
Kruse, Malte	Java
Rosswog, Johannes	
Elisath, Erik	Java, App
Zyla, Daniel	
Lelis, Jan	Web, JS, Ruby
Siedlecki, Kacper	
Maass, Marius	
Sengün, Sonay	
Scholze, Felix	IOS, App
Naber, Bastian	Python

- Mindesttei

Phase 1: Spezifikation

Termin	
20.10	Benutzung von SVN/GIT und TRACK (Bux) Erfahrungen aus einem Semesterprojekt (Gross) Kundenpräsentation: Was wir wollen (Betreuer)
26.10, 9.00	Abgabe Projektstruktur/Teilprojekte/Use Cases
27.10	Webprogrammierung (Starlinger, 45min) Präsentation und Review Projektstruktur (alle)
2.11., 9.00	Abgabe Projektstruktur rev 2, Projektplan, Style Guide, Grobspezifikation aller Teilprojekt
3.11	Genomdaten und deren Integration (Leser, 90min)

Phase 2: Modellierung

Termin	
9.11., 9.00	Abgabe Schnittstellendefinition, Quellauswahl, Mock-Ups
10.11	Hauptspeicherindexierungsverfahren (Sprenger, 45min) Präsentation und Review Projektplan (PL) Erster Meilenstein: Grobspezifikation und Projektplan
17.11	Grundlagen verteilter Programmierung (Bux, 45min) Review
30.11., 9.00	Abgabe Modelle Teilprojekte inkl. Testfälle, Testdaten, Stresstests
1.12	Review Zweiter Meilenstein: Projektplanung

Phase 3: Implementierung

Termin	
15.12.	Zwischenreview: Stand alle Teilprojekte (Präsentation)
Weihnachtsferien	
4.1.2016, 9.00	Abgabe aller Teilprojekte
	...

Phase 4: Integration, Test, Abnahme

Termin	
...	...
	Test, Debugging
	Codereviews
	Abgabe Code + Dokumentation + Readme + Make-File
9.2.	Präsentation, Poster, Abnahme
16.2	Projektabschluss, Nachlese

Literatur lesen!

- Software
  - Design Pattern: Model-View-Controller
  - Interval-Indexstrukturen
  - Servlets, JSP, Java Server Faces
- Hintergrund (bis Ende Oktober)
  - Amberger et al. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®), Human Mutation
  - Stenson et al. The Human Gene Mutation Database (HGMD) and Its Exploitation ... Current Protocols in Bioinformatics
  - Karolchik et al. The UCSC Genome Browser database: 2014 update Nucl. Acids Res.
- Projekte (bis Ende November)
  - Lesk A. Introduction to Bioinformatics, 2014, Oxford University Press
  - Tom DeMarco und Doris Märtin (1998). Der Termin: Ein Roman über Projektmanagement. Hanser Fachbuch