IT and Bioinformatics Strategy of Sequencing the Whole Genome in Clinical Practice at the DKFZ

- Juergen Eils
  Data Management and Genomics IT
- @eilslabs.de

BioQuant
MODEL base of LIFE

IPMB

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

dkfz. GERMAN CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

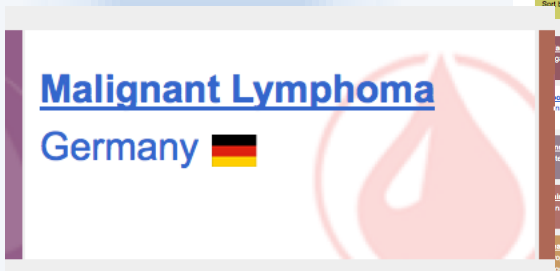
Motivation

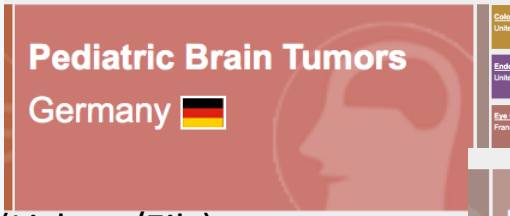
- **Development of sequencing technology at DKFZ**
- IT Infrastructure
- Big Data: Software für organisation and management of genomic data
- Visions

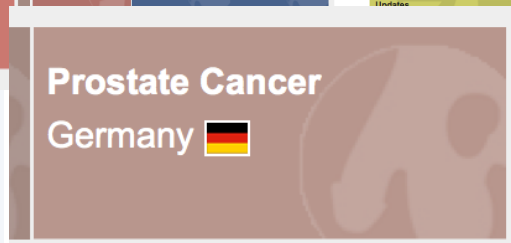
- DKFZ (German Cancer Research Center) is the largest biomedical research institute in Germany
- >3000 employees and >1000 scientists in more than 70 divisions, research groups and clinical cooperation groups
- DKFZ is member of the Helmholtz Association of National Research Centers (90% funding from German Federal Ministry of Education and Research (BMBF), 10% State of Baden-Württemberg
- Jointly with Heidelberg University Hospital, DKFZ has established the National Center for Tumor Diseases (NCT) Heidelberg where promising approaches from cancer research are translated into the clinic.

ICGC - big data project

ICGC Goal: 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.







1. **PedBrainTumor:** Coordinated at DKFZ (Lichter/Eils)
   - Pilocytic astrocytoma (most common pediatric brain tumor)
   - Medulloblastoma (most common malignant pediatric brain tumor)

2. **Prostate Cancer - Early Onset:** Coordinated at DKFZ & University Hospital Hamburg (Sültmann / Sauter)

3. **Malignant Lymphoma:** Coordinated at Univ. Kiel (Siebert), DKFZ responsible for data analysis and data management (Eils)

BioQuant MODEL base of LIFE

eils labs

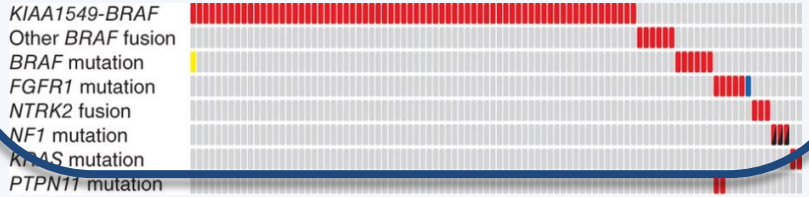
What have we learnt so far?

ICGC

PedBrain Tumor

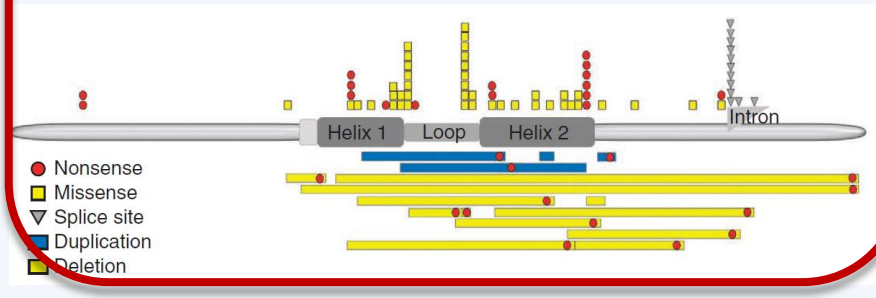
Jones, Jäger et al.: Dissecting the genomic complexity underlying medulloblastoma. **Nature 2012**

Jones, Hutter, Jäger et al.: Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma **Nature Genetics 2013**



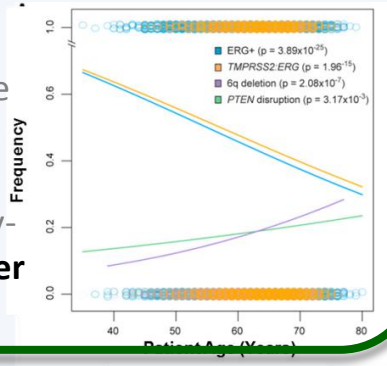
Malignant Lymphoma

Richter, Schlesner et al.: Recurrent mutation of the *ID3* gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. **Nature Genetics 2012**



Early Onset Prostate Carcinoma

Weischenfeldt, Simon, Feuerbach, et al.: Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. **Cancer Cell 2013**



Central Data Coordination for German Epigenome Program DEEP

IHEC
International Human Epigenome Consortium

DEEP
Deutsches Epigenom Programm

BioQuant
MODEL base of LIFE

eilslabs

dkfz.

Publications ICGC Pedbrain

Improved Tumor Classification

- Sturm *et al.* Cell 164, 1060-1072 (2016)
- Pajtler, K.W. *et al.* Cancer Cell, 27, 728-743 (2015)
- Korshunov, A. *et al.* Acta Neuropathol, 129, 669-78 (2015)
- Sturm *et al.* Nat Rev Cancer, 14, 92-107 (2014)
- Sturm *et al.* Cancer Cell, 22, 425-37 (2012)
- Kool *et al.* Acta Neuropathol, 123, 473-84 (2012)

Molecular Profiles, Tumor Patho-mechanisms, Therapeutic Targets

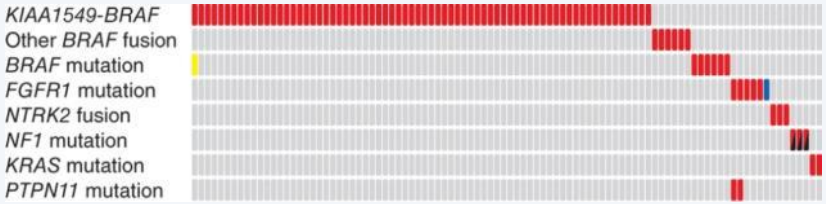
- Kool *et al.* Cancer Cell, 25, 393-405 (2014)
- Northcott *et al.* Nature, 511, 428-34 (2014)
- Mack *et al.* Nature, 506, 445-50 (2014)
- Jones *et al.* Nature Genetics, 45, 927-32 (2013)
- Bender *et al.* Cancer Cell, 24, 660-672 (2013)
- Lambert *et al.* Acta Neuropathol, 126, 291-301 (2013)
- Jones *et al.* Brain Pathol, 23, 193-9 (2013)
- Fontebasso *et al.* Brain Pathol, 23, 210-6 (2013)
- Northcott *et al.* Nature 488, 49-56 (2012)
- Jones *et al.* Nature, 488, 100-5 (2012)
- Pugh *et al.* Nature, 488, 106-10 (2012)
- Northcott *et al.* Nat Rev Cancer, 12, 818-34 (2012)
- Khuong-Quang et al. Acta Neuropathol, 124, 439-47 (2012)
- Schwartzentruber *et al.* Nature, 482, 226-31 (2012)

Genome Biology of Tumors

- Hovestadt *et al.* Nature, 510, 537-41 (2014)
- Jäger *et al.* Cell, 155, 567-81 (2013)
- Alexandrov *et al.* Nature, 500, 415-21 (2013)
- Rausch *et al.* Cell, 148, 59-71 (2012)

Methods Development

- Rieber *et al.* PLoS One, 8, e66621 (2013)
- Alioto *et al.* Nature Comm. 6, 10001 (2015)
- Hovestadt *et al.* Acta Neuropathol, 125, 913-6 (2013)



The ICGC and TCGA Pan-Cancer Project

Genomes (top 3 centers)

dkfz. wellcome trust sanger institute The Cancer Genome Atlas

Variant Calling Pipelines

BROAD INSTITUTE dkfz. EMBL wellcome trust sanger institute

Data and Computing Centers

Bionimbus cnag centre nacional d'anàlisi genòmica centro nacional de análisis genómico dkfz. EMBL-EBI ETRI RIKEN UCSC

Co-Lead of ICGC Pan-Cancer Working Groups

- Pathogens in cancer (R. Eils, P. Lichter, DKFZ and Xiaoping Su, MD Anderson)
- Integration of epigenome and genome (B. Brors, C. Plass, DKFZ, and Peter Laird, USC)

BioQuant MODEL base of LIFE eilslabs dkfz.

NATIONALES CENTRUM
FÜR TUMORERKRANKUNGEN

NCT HEIDELBERG

- **JOINT VENTURE BETWEEN DKFZ AND UNIVERSITY HOSPITAL**
- **10.000 CANCER PATIENTS PER YEAR**
- **FÜR 3.000 GENOME SEQUENZING AS AN OPTION**

dkfz.
NCT
UniversitätsKlinikum Heidelberg
50 Jahre – Forschen für ein Leben ohne Krebs

NCT CLINICAL CANCER PROGRAM: MOLECULAR SEQUENCING DIAGNOSTICS

Benedikt Brors

35 days

Patient enrolment	Sample assessment, asservation and processing	Molecular profiling and bioinformatics	Clinical interpretation of molecular data	Validation of immediately actionable lesions	Molecular tumor board	Treatment
- Diagnosis - NCT MASTER consent	- Biopsy and blood withdrawal - Pathological diagnosis - Biobanking - Analyte extraction and QC	- Exome and transcriptome high throughput sequencing - SNVs, CNVs, indels - Fusions, expression - Germline (e.g. TP53, BRCA1)	- Certified laboratory - Sanger sequencing, FISH, etc. - Target identification	- Literature research - Data quality assessment - Target identification	- Clinicians, translational oncologists, bioinformaticians, scientists, case management - Reporting of important lesions - Suggestion for clinical action - Secondary validation	- Targeted therapy - Combination therapy - NCT MASTER trial - NCT IITS - N of 1 Trial - SOC

Pilotphase: March 2014 – December 2015 : x00 patients, success rate* 56%

* actionable mutation validated by certified diagnostics methods

BioQuant MODEL base of LIFE

eils labs

dkfz.

Massive Genome Sequencing using Illumina HiSeq X Ten

The HiSeq X Ten contains 10 sequencing systems.

HiSeq X™ Ten

Population Power

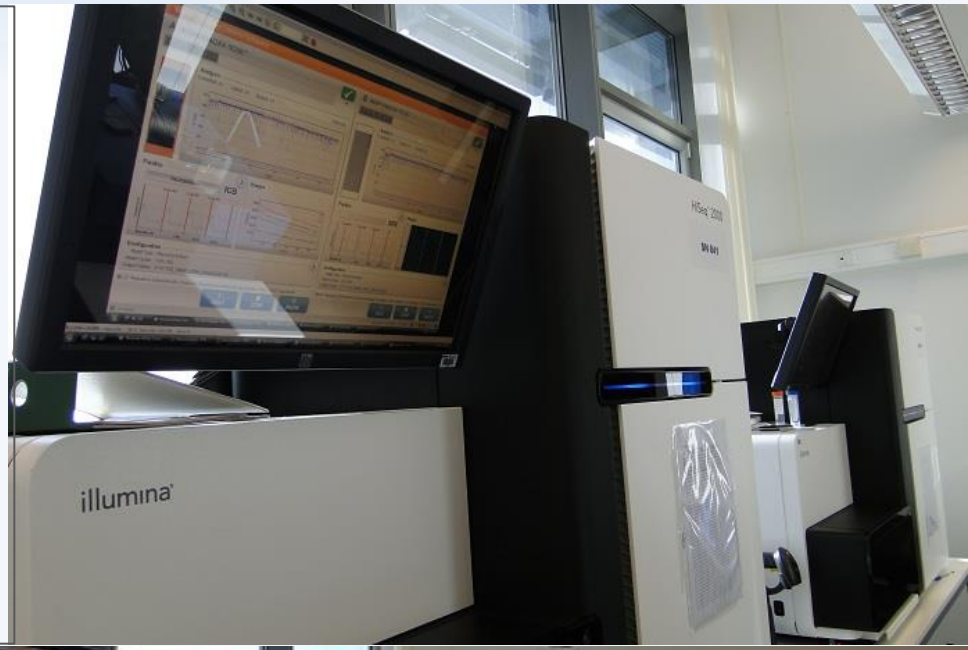
Composed of 10 HiSeq X Systems, the HiSeq X Ten is the first sequencing platform that breaks the $1000 barrier for a 30x human genome. The HiSeq X Ten System is ideal for population-scale projects focused on the discovery of genotypic variation to understand and improve human health. It can rapidly sequence tens of thousands of samples at high genome coverage, delivering a comprehensive catalog of human variation within and outside coding regions.

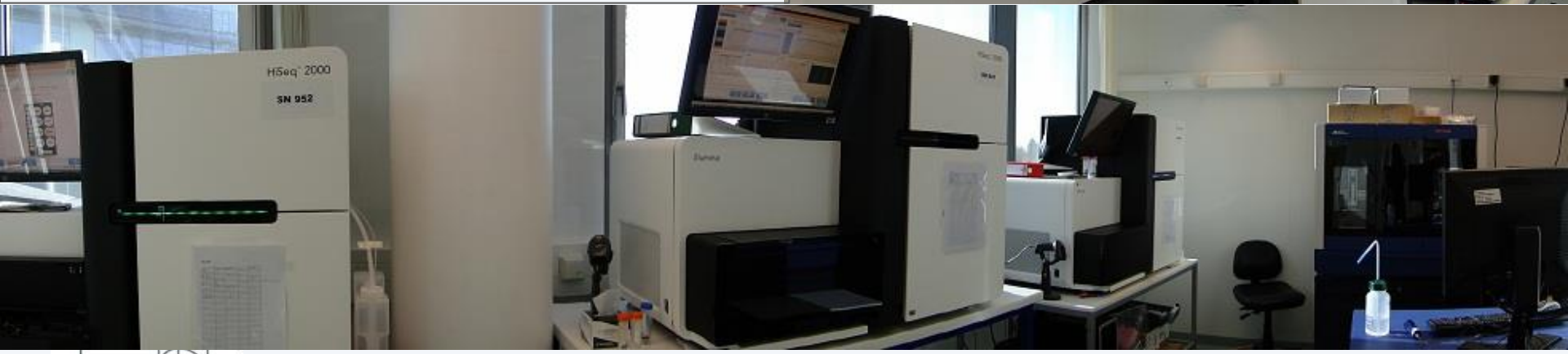
- Tens of thousands of whole human genomes per year
- $1000 human genome, including depreciation, sample preparation, and labor

Capacity:	4.500 patients / year (120x Coverage)
Raw Data:	1,8 PB / year (5 TBytes per day)
Total Data including Analysis Data (approx. 2x overhead)	4 PB / year (11 TBytes per day)
Required growth of storage incl. mirror storagy for 2015-2018:	~ 10 PB per year

Some Petabase / Petabyte numbers

Sequencer are the data producer
One Genome has roughly 3 Gbases
    3.000.000.000 Bases
The standard coverage
    rate is 30x to 40x
One sequenzed genome
    requires 100 Gbases
    1 genome = 100GB
    Analysis and mirror: factor 4



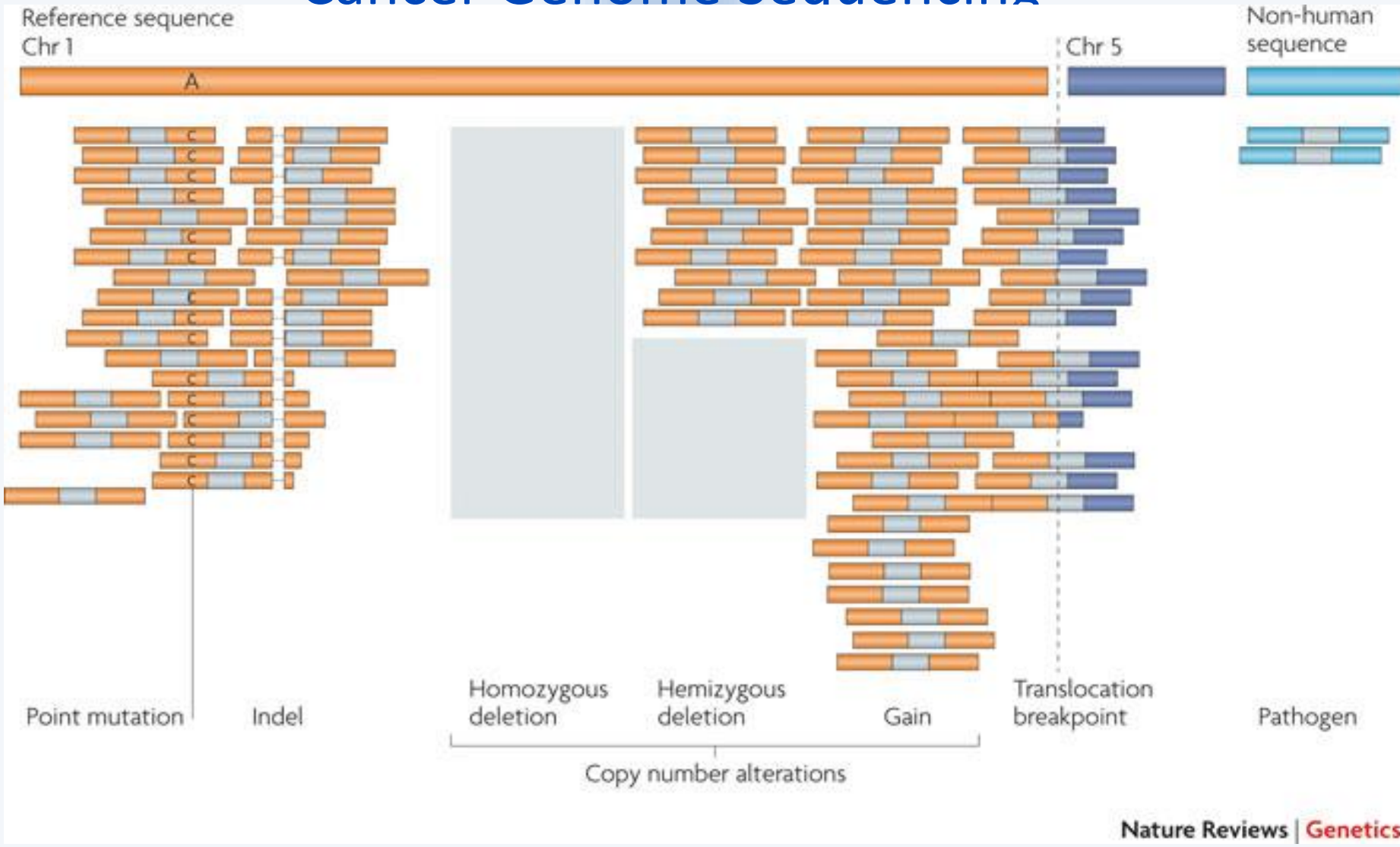


BioQuant MODEL base of LIFE

eilslabs

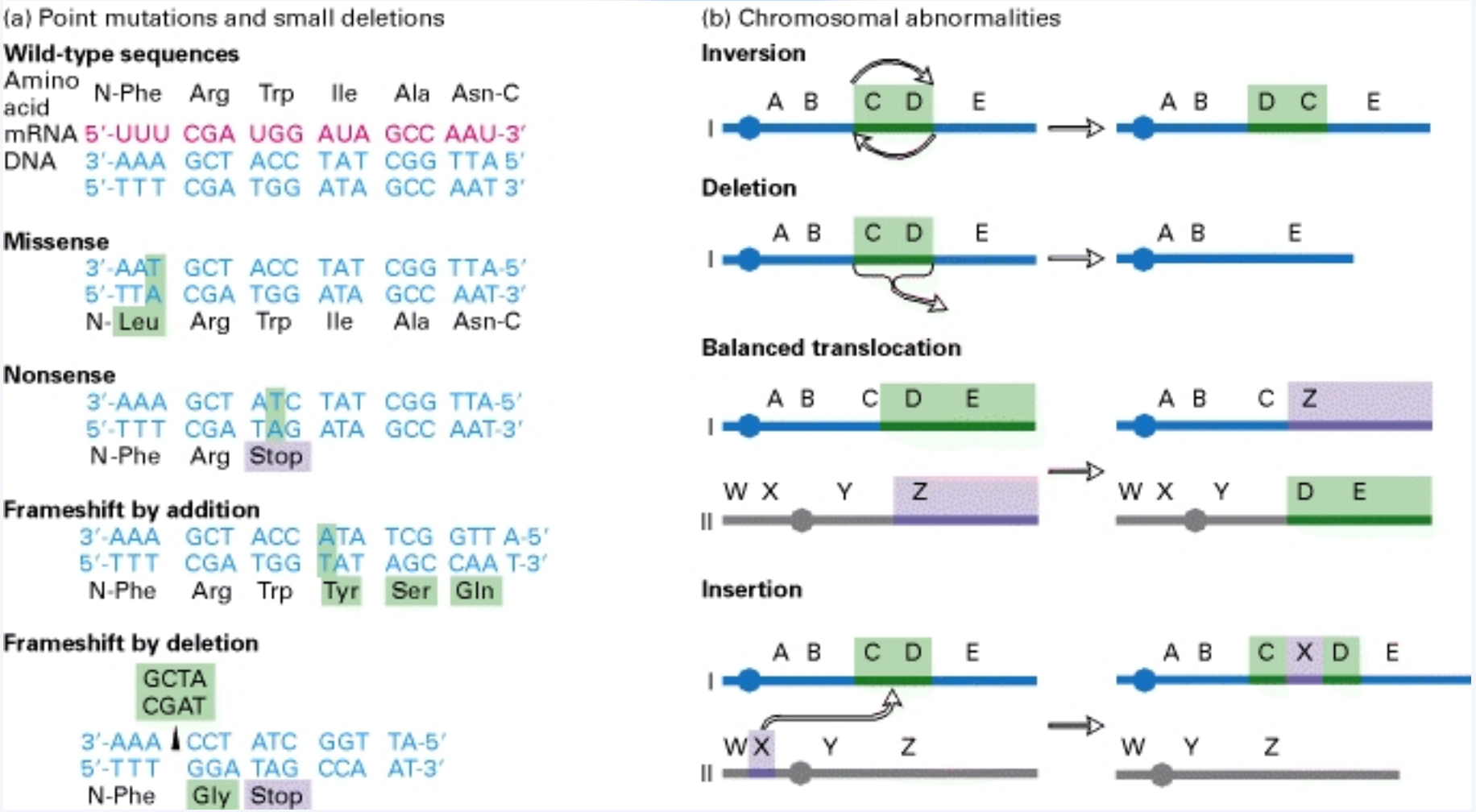
dkfz.

Cancer Genome Sequencing



Nature Reviews Genetics 11, 685-696

Big data approach: Somatic Variation in Cancer

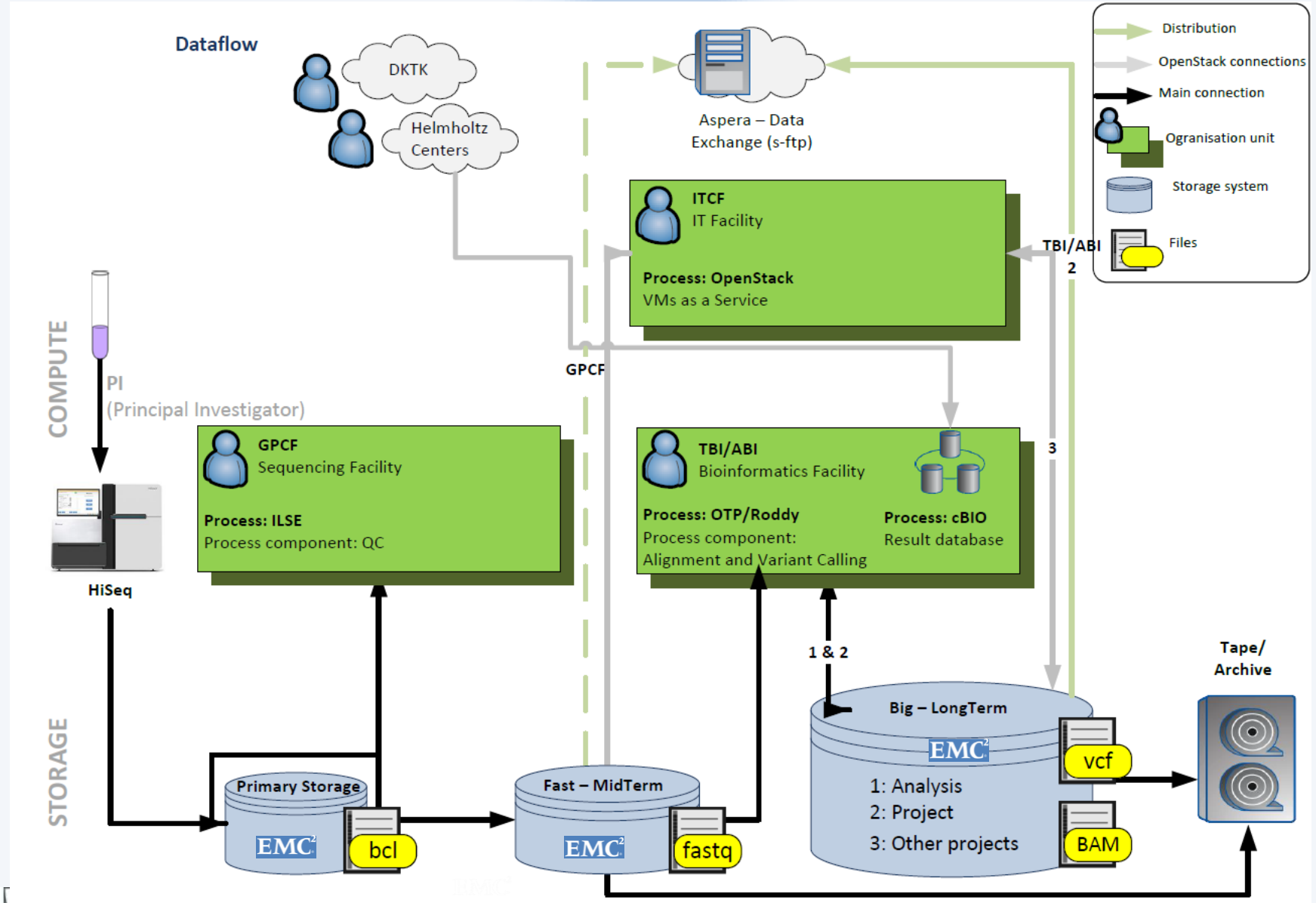


Baltimore et al., Mol. Cell Biol.

Motivation

- Development of sequencing technology at DKFZ
- **IT Infrastructure**
- Big Data: Software für organisation and management of genomic data
- Visions

Reduction of I/O

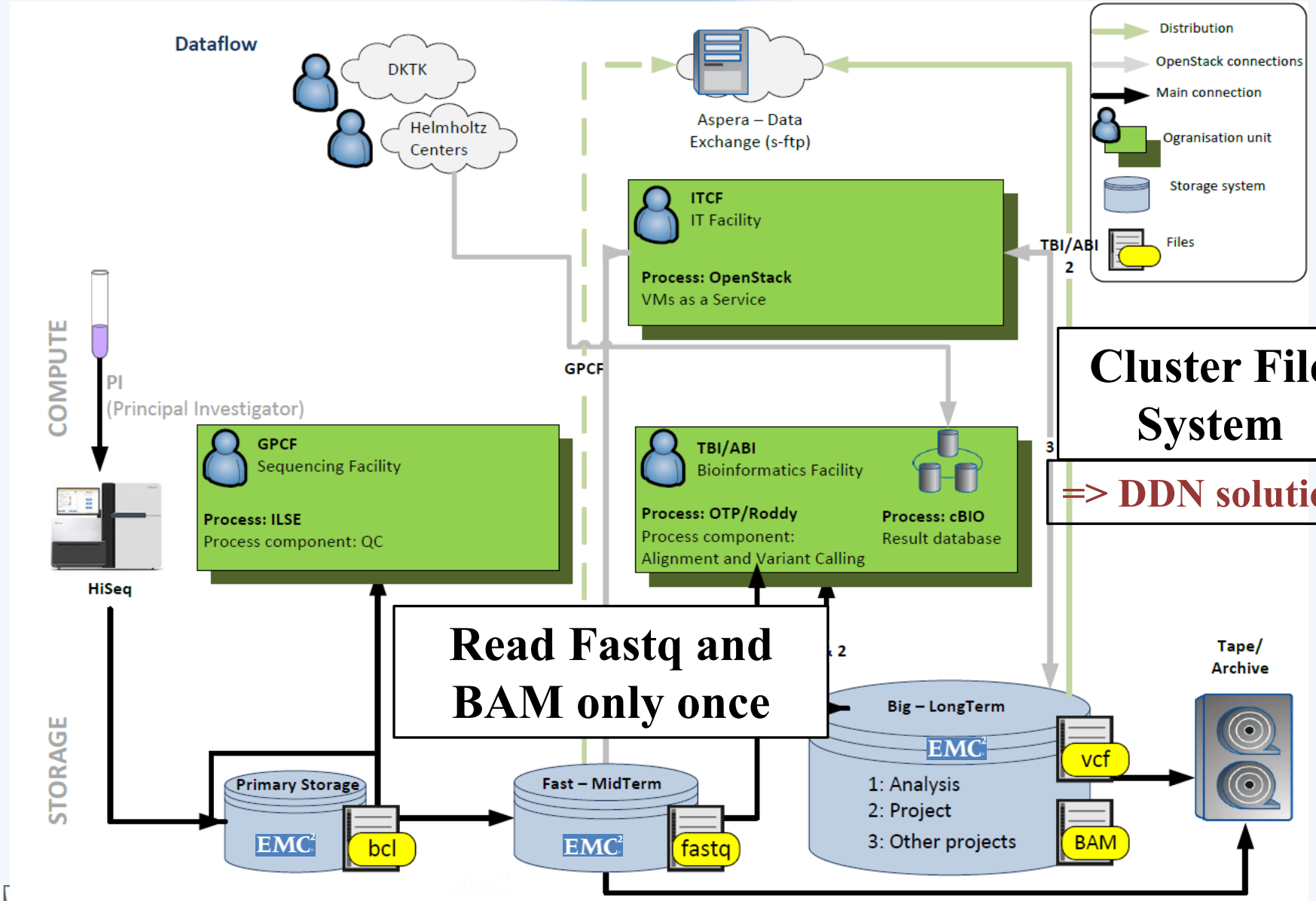


BioQuant MODEL base of LIFE

eilslabs

dkfz.

Acceleration II: reduction of I/O



BioQuant MODEL base of LIFE

eilslabs

dkfz.

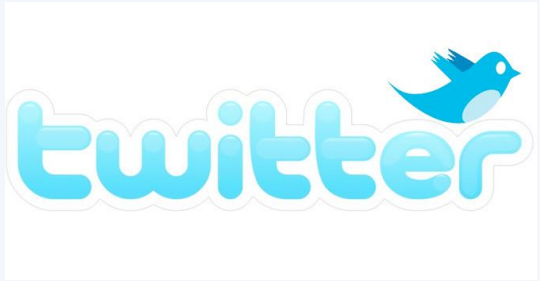
Motivation

- Development of sequencing technology at DKFZ
- IT Infrastructure
- **Big Data: Software für organisation and management of genomic data**
- Visions

Big Data in Genome Medicine: At equal level as Twitter

600 Terabytes per day

(Source: Vagata, P., & Wilfong, K. (2014). Scaling the Facebook data warehouse to 300 PB. https://code.facebook.com/posts/229861827208629/)



12 Terabytes per day

(Source: Zhao, L., Sakr, S., Liu, A., & Bouguettaya, A. (2014). Cloud Data Management, Springer)

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

50 Years – Research for A Life Without Cancer

Sequencing@DKFZ: 11 Terabytes per day

BioQuant MODEL base of LIFE

eils labs

dkfz.

Complexity

- Data from 15.000 samples
- 10 PB of data
- Find back your data
- Find back any data
  - Database
  - Structure the data

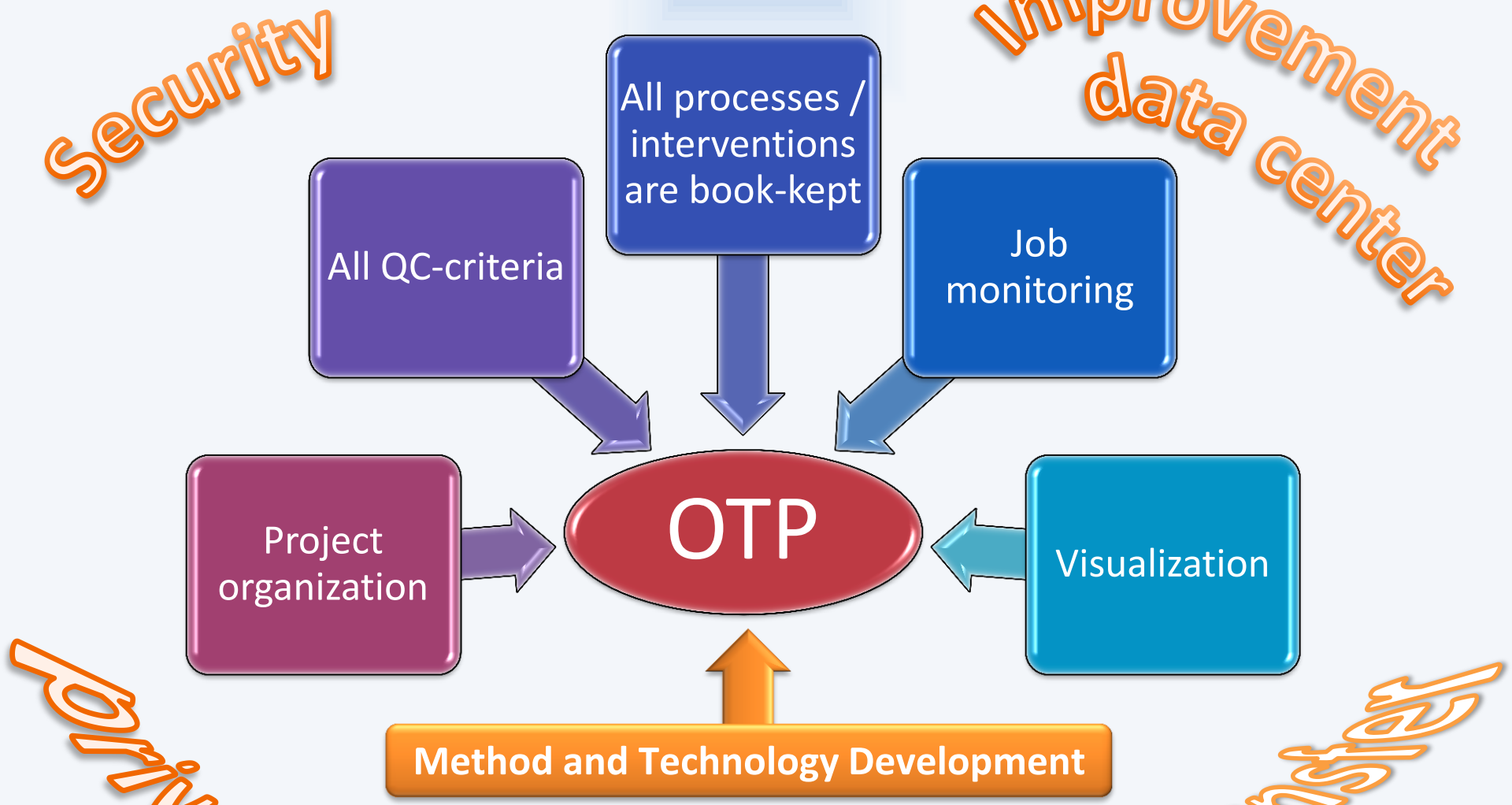
BioQuant MODEL base of LIFE

eilslabs

dkfz.

Automatisation

OTP: Central research platform



BioQuant MODEL base of LIFE

eils labs

dkfz.

Data flow at DKFZ

Major projects

DKFZ-HIPO

DKTK

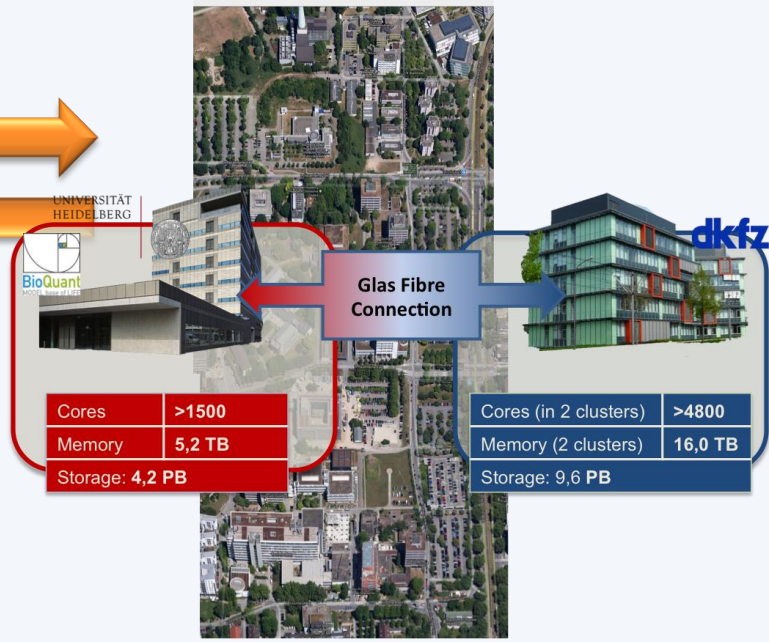
PanCancer

de.NBI

NCT POP

ICGC

OTP



DEEP

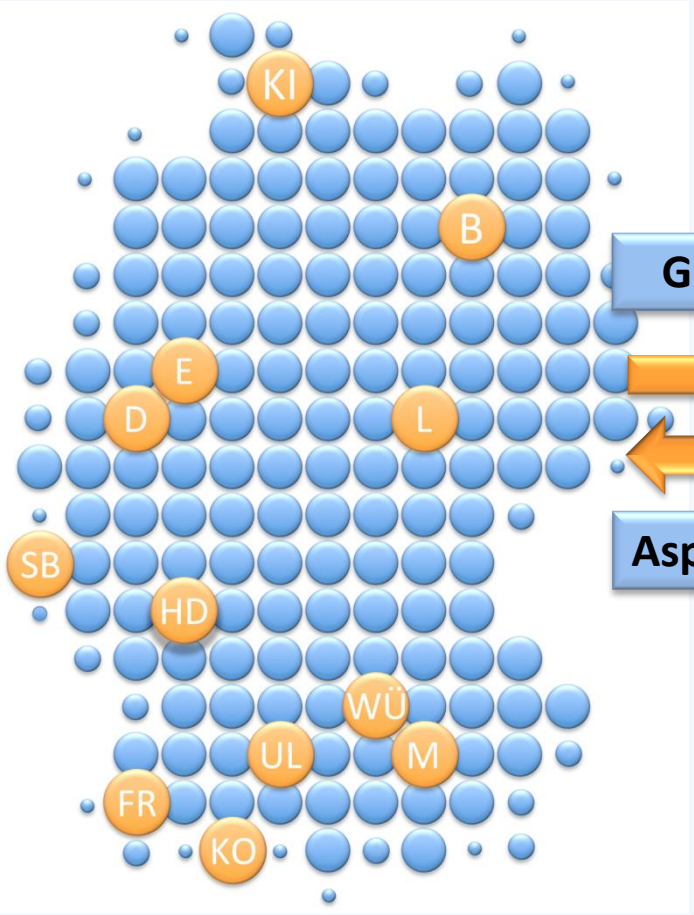
IHEC
International Human Epigenome Consortium

BioQuant
MODEL base of LIFE

eilslabs

dkfz.

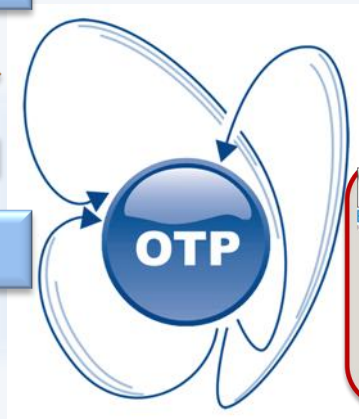
Data flow via OTP

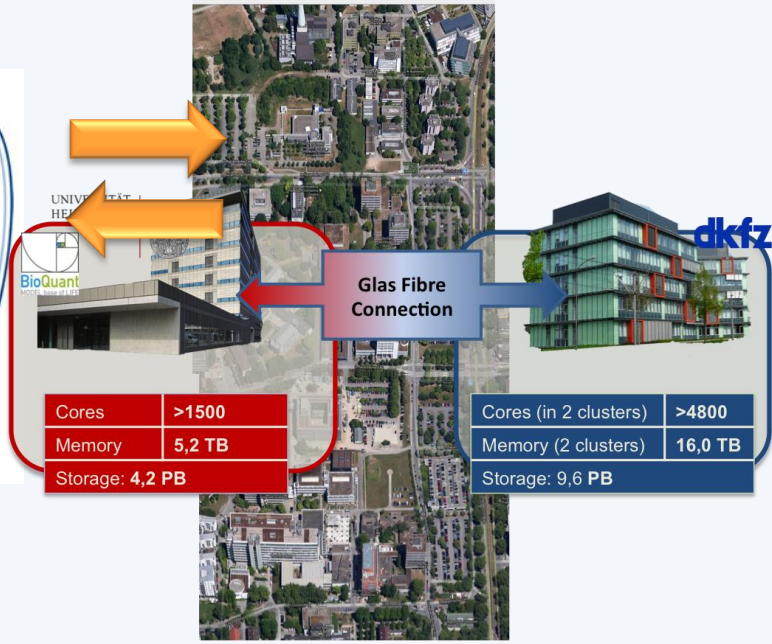


Sequence centers in Germany

GPCF

Aspera





Receiving

NGS fastq data

Processing

Alignment & merging & variant calling

Providing

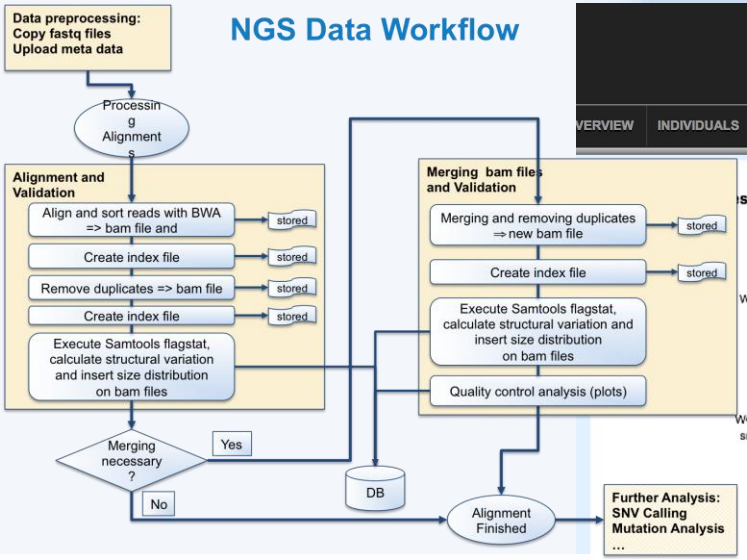
Structured data

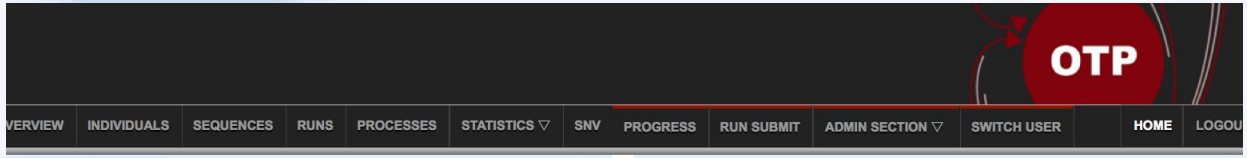
BioQuant
MODEL base of LIFE

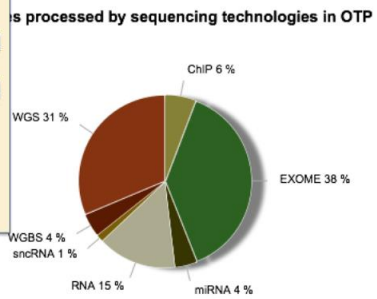
eilslabs

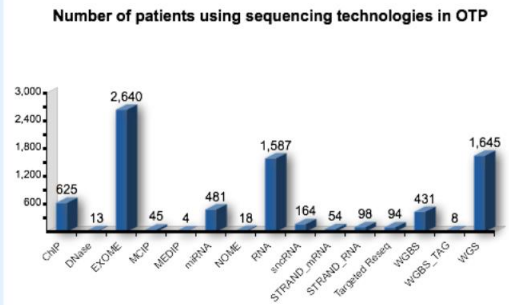
dkfz.

Automatisation: OTP - Processing framework

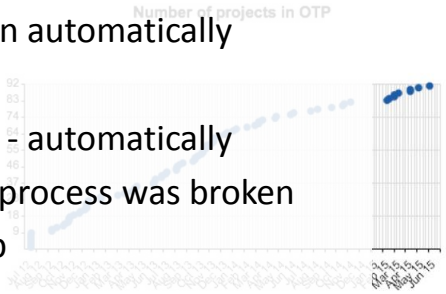


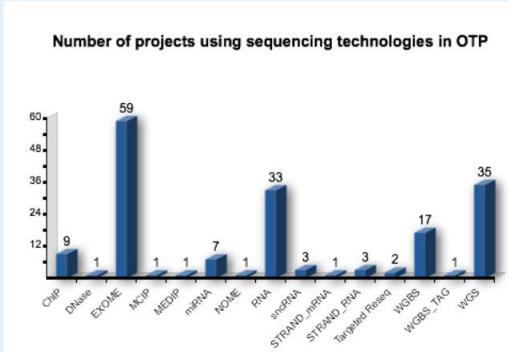






- Processing frameworks for huge NGS projects:
  - Project organization
  - To speed-up: All routine jobs run automatically
  - No more manual shell scripts
  - Registration, alignment, QC, VC - automatically
  - Automatic information when a process was broken
  - Restart each single process step





BioQuant
MODEL base of LIFE

eils labs

dkfz.

Motivation

- Development of sequencing technology at DKFZ
- **IT Infrastructure**
- Big Data: Software für organisation and management of genomic data
- **Visions**

Major cooperations in big data analytics

- IBM Watson health
- Sap HANA

Applications – Medical Research Insights

Design and build Analytics Software that allows doctors and researchers to access patient data from various systems in real-time with a single interface to improve cancer research.

NCT
NATIONAL CENTER FOR TUMOR DISEASES HEIDELBERG

supported by
German Cancer Research Center (DKFZ)
Heidelberg University Medical Center
Hospital for Thoracic Diseases
German Cancer Aid

INTERACTION AWARDS FINALIST OPTIMIZING

German Design Award WINNER

AWARD RECIPIENT C77 D A The Best of Design 2014

DESIGN AWARD 2015

Germany Land of Ideas Landmark 2015

reddot award communication design

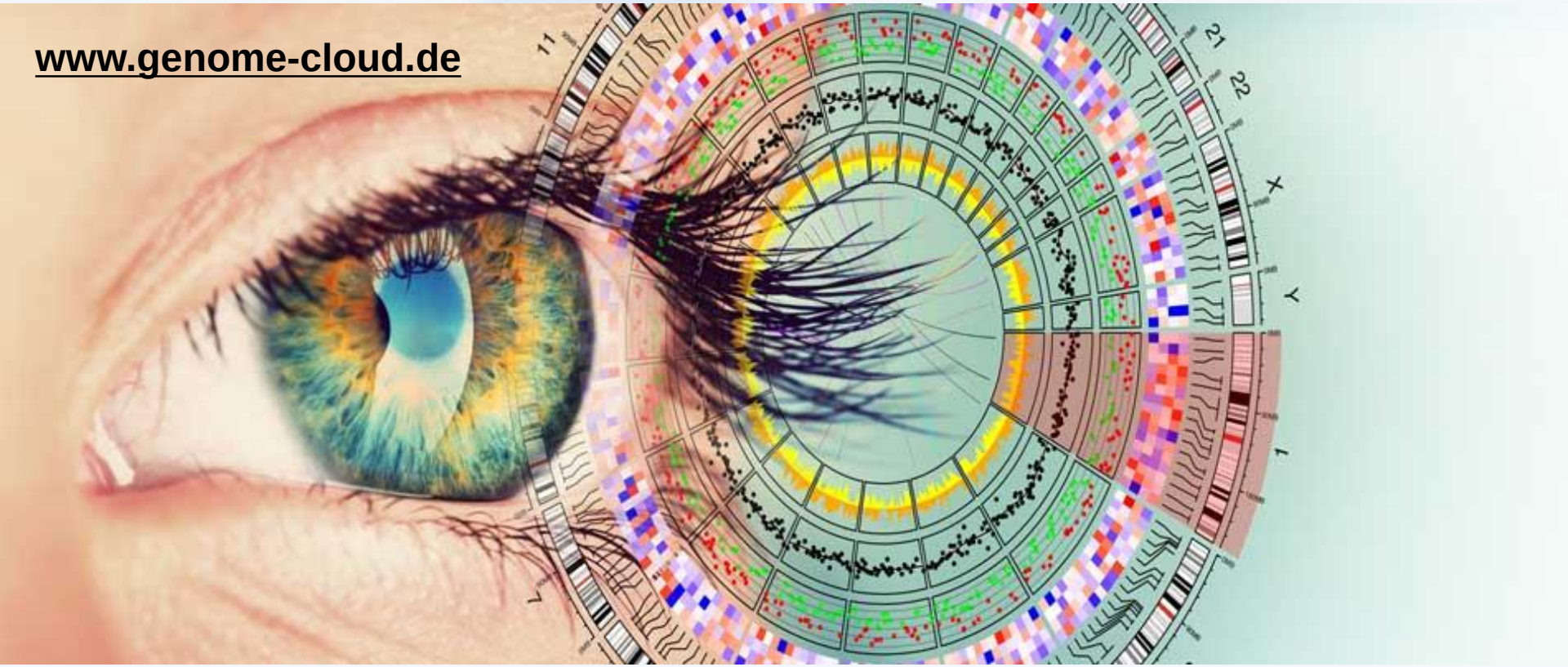
BioQuant MODEL base of LIFE

eils labs

dkfz.

Genomics Cloud:
Preparing Germany for 100,000 Genomes

www.genome-cloud.de



dkfz. EMBL NCT SAP

BioQuant MODEL base of LIFE

eilslabs

dkfz.

Possible solution: German genomics cloud

Infrastructure-as-a-Service

Preconfigured Pipelines Software-as-a-Service

Common standards and APIs

ICGC

PCAWG

amazon web services™

Google Genomics

BROAD INSTITUTE

IBM

Company X..

SAP

BioQuant MODEL base of LIFE

eils labs

dkfz.