Big data in Next Generation Sequencing (NGS): Requirements and Challenges

Juergen Eils
Data Management Group @ eilslabs

- DKFZ (German Cancer Research center) is the largest biomedical research institute in Germany
- In 2008, Professor Harald zur Hausen awarded the Nobel Prize in Medicine for discovering that human papillomaviruses (HPV) cause cervical cancer.
- More than 70 divisions and research groups,
- About 80 employees are working in the Bioinformatics division eilsLabs
- About 20 employees are working in the data management team

NGS in personalized oncology
Structure of the talk

– NGS projects

– Infrastructure, cloud

– Pipelines and software

BioQuant MODEL base of LIFE

eilslabs

dkfz.

Data flow at DKFZ

Major projects

DKFZ-HIPO

NCT POP

ICGC

DEEP

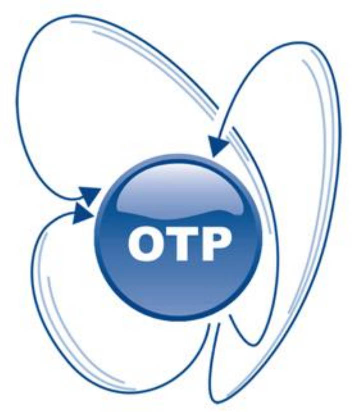
IHEC
International Human Epigenome Consortium

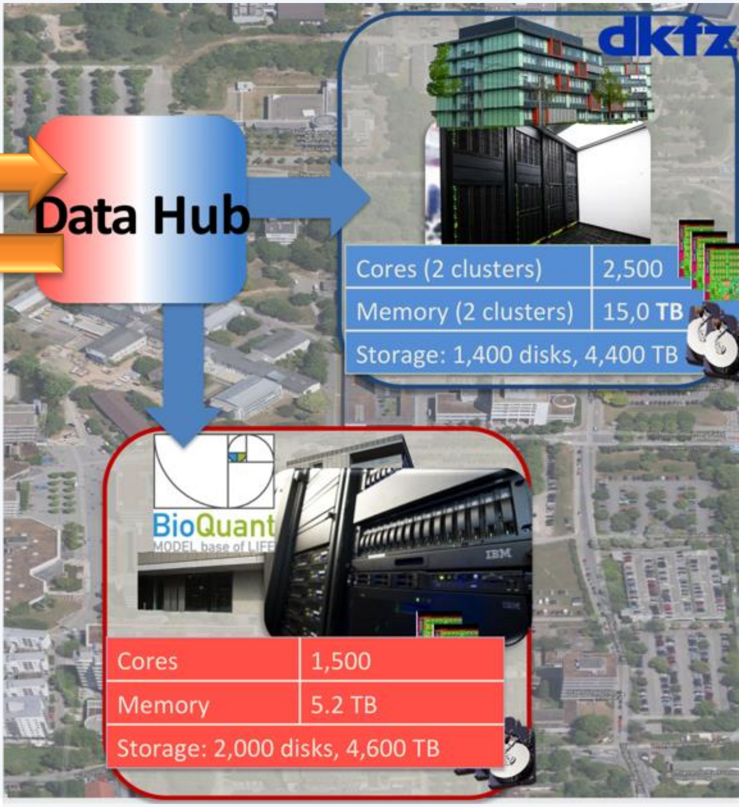
BioQuant
MODEL base of LIFE

eilslabs

PanCancer

de.NBI





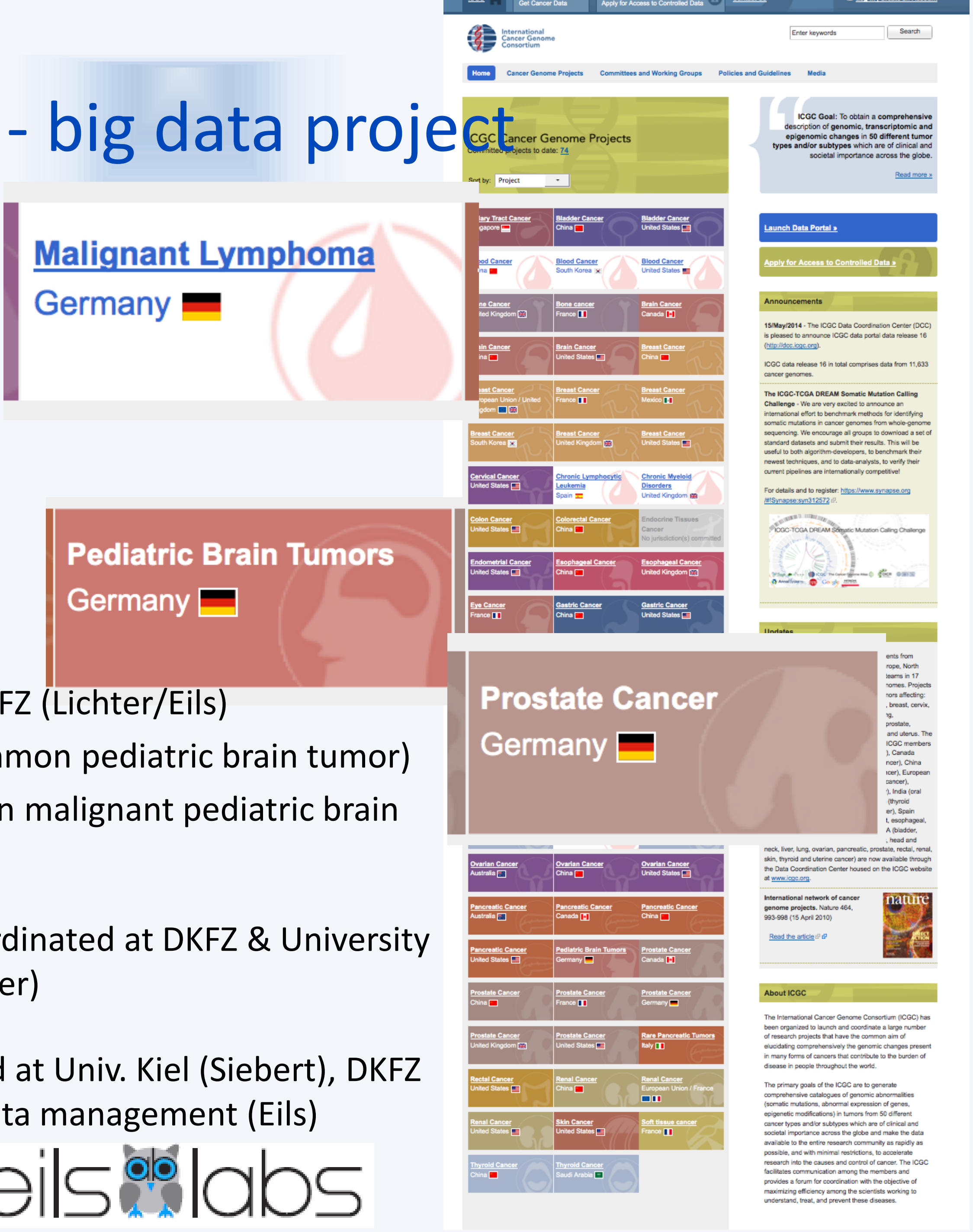
Cores (2 clusters)	2,500
Memory (2 clusters)	15,0 TB
Storage: 1,400 disks, 4,400 TB	

Cores	1,500
Memory	5.2 TB
Storage: 2,000 disks, 4,600 TB	

dkfz.

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.



Malignant Lymphoma
Germany

Pediatric Brain Tumors
Germany

Prostate Cancer
Germany

Get Cancer Data | Apply for Access to Controlled Data

International Cancer Genome Consortium

Enter keywords | Search

Home | Cancer Genome Projects | Committees and Working Groups | Policies and Guidelines | Media

CGC Cancer Genome Projects
Committed projects to date: 74

Sort by: Project

lary Tract Cancer gapore	Bladder Cancer China	Bladder Cancer United States
od Cancer ina	Blood Cancer South Korea	Blood Cancer United States
ne Cancer ited Kingdom	Bone cancer France	Brain Cancer Canada
in Cancer ina	Brain Cancer United States	Breast Cancer China
ast Cancer ropean Union / United gdom	Breast Cancer France	Breast Cancer Mexico
Breast Cancer South Korea	Breast Cancer United Kingdom	Breast Cancer United States
Cervical Cancer United States	Chronic Lymphocytic Leukemia Spain	Chronic Myeloid Disorders United Kingdom
Colon Cancer United States	Colorectal Cancer China	Endocrine Tissues Cancer No jurisdiction(s) committed
Endometrial Cancer United States	Esophageal Cancer China	Esophageal Cancer United Kingdom
Eye Cancer France	Gastric Cancer China	Gastric Cancer United States
Ovarian Cancer Australia	Ovarian Cancer China	Ovarian Cancer United States
Pancreatic Cancer Australia	Pancreatic Cancer Canada	Pancreatic Cancer China
Pancreatic Cancer United States	Pediatric Brain Tumors Germany	Prostate Cancer Canada
Prostate Cancer China	Prostate Cancer France	Prostate Cancer Germany
Prostate Cancer United Kingdom	Prostate Cancer United States	Rare Pancreatic Tumors Italy
Rectal Cancer United States	Renal Cancer China	Renal Cancer European Union / France
Renal Cancer United States	Skin Cancer United States	Soft tissue cancer France
Thyroid Cancer China	Thyroid Cancer Saudi Arabia	

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.

Read more »

Launch Data Portal »

Apply for Access to Controlled Data »

Announcements

15/May/2014 - The ICGC Data Coordination Center (DCC) is pleased to announce ICGC data portal data release 16 (http://dcc.icgc.org).

ICGC data release 16 in total comprises data from 11,633 cancer genomes.

The ICGC-TCGA DREAM Somatic Mutation Calling Challenge - We are very excited to announce an international effort to benchmark methods for identifying somatic mutations in cancer genomes from whole-genome sequencing. We encourage all groups to download a set of standard datasets and submit their results. This will be useful to both algorithm-developers, to benchmark their newest techniques, and to data-analysts, to verify their current pipelines are internationally competitive!

For details and to register: https://www.synapse.org/#!Synapse:syn312572.

ICGC-TCGA DREAM Somatic Mutation Calling Challenge

Updates

ents from
rope, North
teams in 17
nomes. Projects
nors affecting:
, breast, cervix,
ng,
prostate,
and uterus. The
ICGC members
), Canada
ncer), China
ncer), European
cancer),
), India (oral
(thyroid
er), Spain
t, esophageal,
A (bladder,
, head and
neck, liver, lung, ovarian, pancreatic, prostate, rectal, renal, skin, thyroid and uterine cancer) are now available through the Data Coordination Center housed on the ICGC website at www.icgc.org.

International network of cancer genome projects. Nature 464, 993-998 (15 April 2010)

Read the article

About ICGC

The International Cancer Genome Consortium (ICGC) has been organized to launch and coordinate a large number of research projects that have the common aim of elucidating comprehensively the genomic changes present in many forms of cancers that contribute to the burden of disease in people throughout the world.

The primary goals of the ICGC are to generate comprehensive catalogues of genomic abnormalities (somatic mutations, abnormal expression of genes, epigenetic modifications) in tumors from 50 different cancer types and/or subtypes which are of clinical and societal importance across the globe and make the data available to the entire research community as rapidly as possible, and with minimal restrictions, to accelerate research into the causes and control of cancer. The ICGC facilitates communication among the members and provides a forum for coordination with the objective of maximizing efficiency among the scientists working to understand, treat, and prevent these diseases.

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)

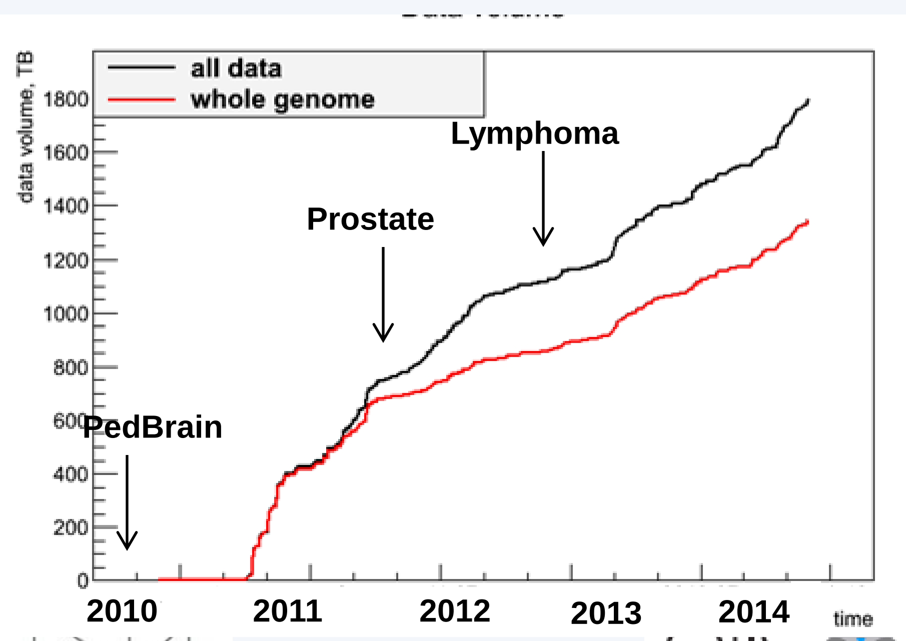
3. **Malignant Lymphoma:** Coordinated at Univ. Kiel (Siebert), DKFZ responsible for data analysis and data management (Eils)

BioQuant MODEL base of LIFE

eilslabs

The NGS data flood from 3 German ICGC projects
Status end of 2014

	WGS*	WES*	RNAseq	Mate-pair*	WGBS*
PedBrain-Medulloblastoma	599	53	174	232	62
PedBrain-Astrocytoma	316	-	94	10	-
Early Onset-Prostate	99	38	39	97	-
Malignant Lymphoma	232	12	106	8	35
Glioblastoma	101	10	29		8

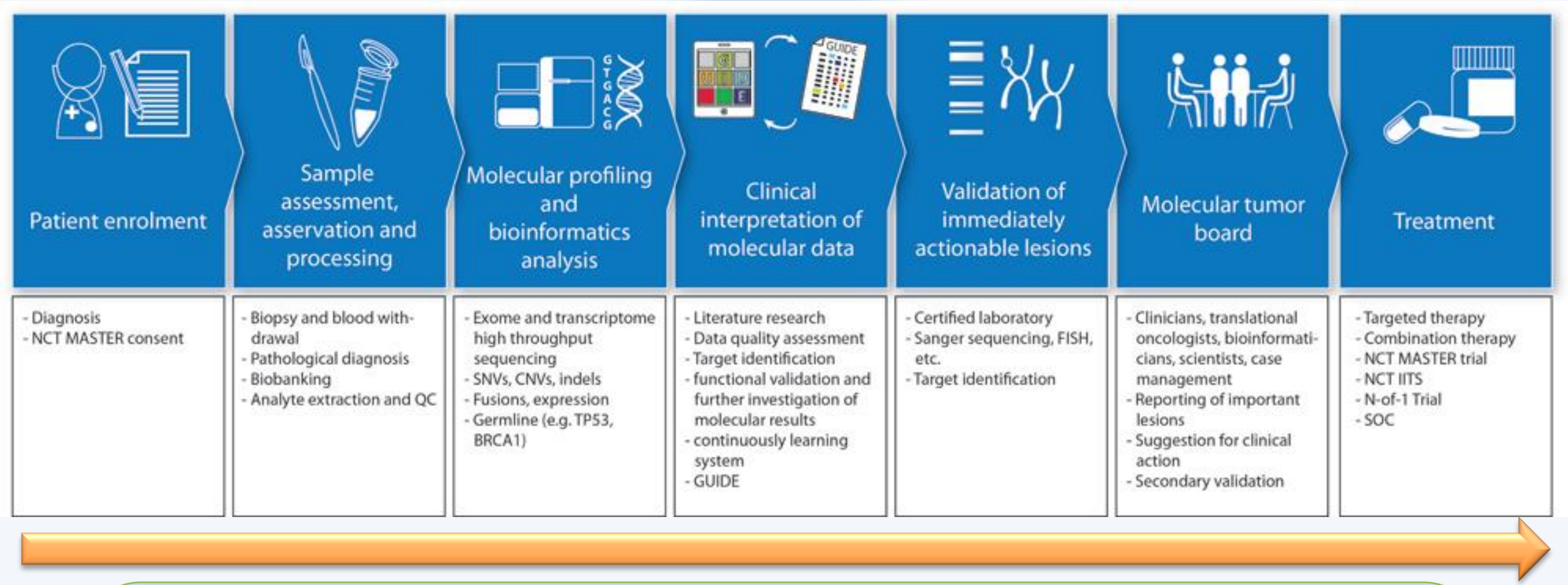


*Tumor and normal counted separately

- only main data types shown

- combined for all 3 ICGC projects

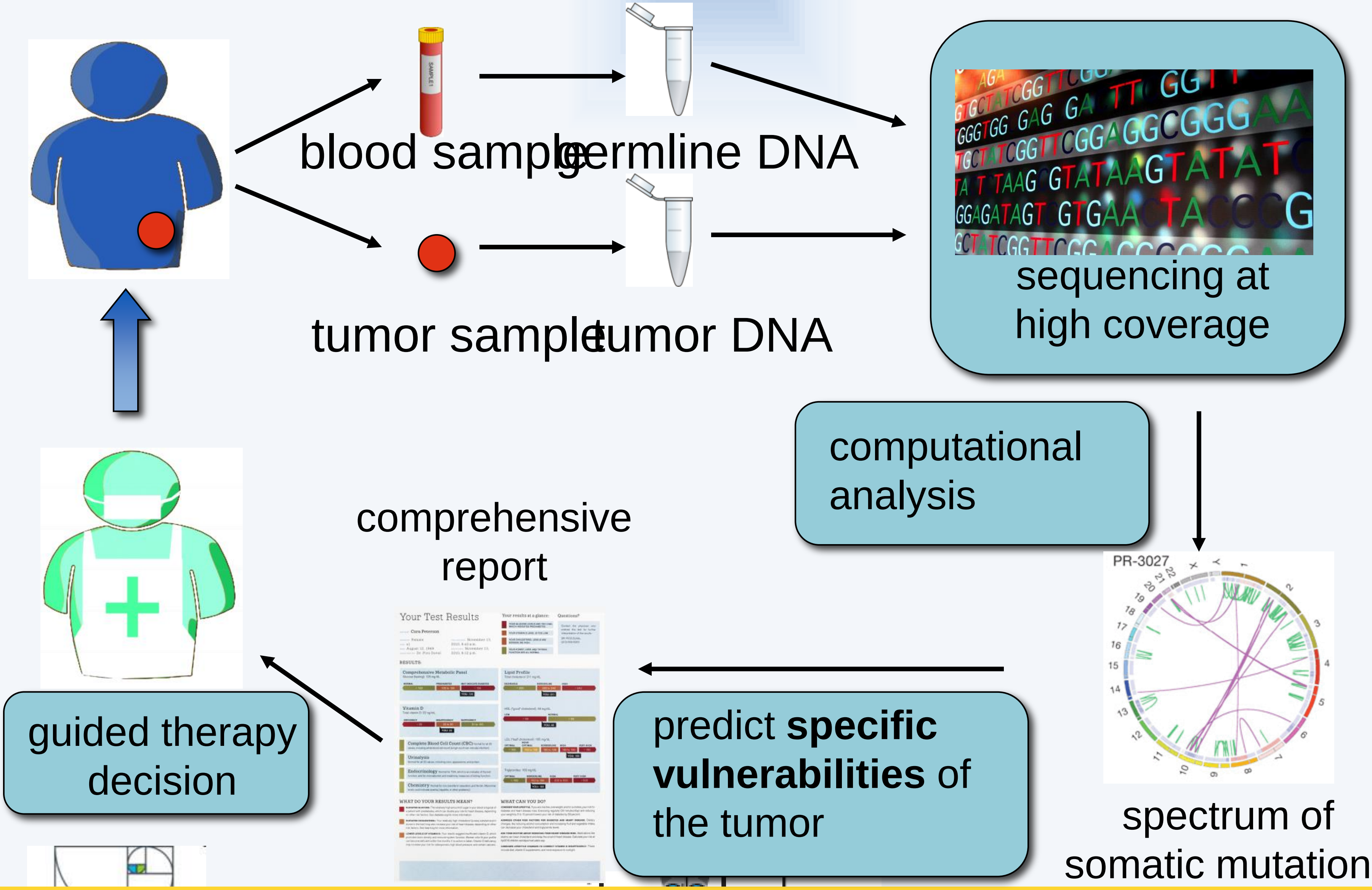
dkfz.hipo: Precision Oncology



Patient enrolment	Sample assessment, asservation and processing	Molecular profiling and bioinformatics analysis	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)	- Literature research - Data quality assessment - Target identification - functional validation and further investigation of molecular results - continuously learning system - GUIDE	- Certified laboratory - Sanger sequencing, FISH, etc. - 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

- Mission: Bringing Genome Sequencing to the Patient
- Currently 50 projects selected including glioblastoma, pediatric cancers, CLL, sarcoma, gastric, colon, prostate, pancreatic, lung, breast and head/neck cancer
- 2015 1500 pat. /year, 2016 2500 p. /y, 2017 3500 p. /y,
- Goal: providing sequencing profile to each cancer patient (20.000 p.a.)

Goal: Genomic Cancer Medicine



blood sample

germline DNA

tumor sample

tumor DNA

sequencing at high coverage

computational analysis

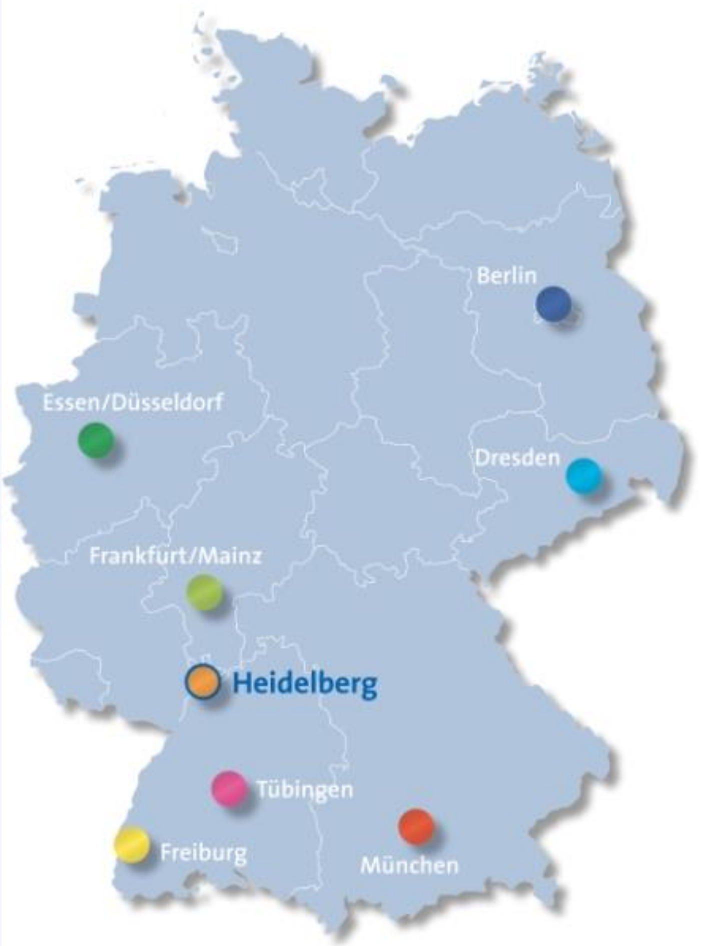
spectrum of somatic mutations

predict **specific vulnerabilities** of the tumor

comprehensive report

guided therapy decision

Inform INdividualized Therapy FOr Relapsed Malignancies in Childhood: 250-300 cases for feasibilty study



Berlin	Charité Universitätsmedizin Berlin
Dresden	Universitätsklinikum Carl Gustav Carus der Technischen Universität Dresden, Max-Planck-Institut für Molekulare Zellbiologie und Genetik, Helmholtz-Zentrum Dresden-Rossendorf
Essen/ Düsseldorf	Westdeutsches Tumorzentrum am Universitätsklinikum Essen, Heinrich-Heine-Universität Düsseldorf
Frankfurt/ Mainz	Universitätsklinikum Frankfurt am Main, Johann-Wolfgang-Goethe-Universität Frankfurt, Georg-Speyer-Haus, Krankenhaus Nordwest Frankfurt, Universitätsmedizin der Johannes Gutenberg-Universität Mainz
Freiburg	Tumorzentrum Ludwig Heilmeyer – Comprehensive Cancer Center Freiburg, Albert-Ludwigs-Universität Freiburg, Universitätsklinikum Freiburg und Max-Planck-Institut Freiburg
Heidelberg	Deutsches Krebsforschungszentrum (Kernzentrum), Universitätsklinikum Heidelberg, Ruprecht-Karls-Universität Heidelberg, Nationales Centrum für Tumorerkrankungen Heidelberg
München	Ludwig-Maximilians-Universität München, Technische Universität München
Tübingen	Eberhard Karls Universität Tübingen, Medizinische Fakultät, Universitätsklinikum Tübingen, Mathematisch-Naturwissenschaftliche Fakultät

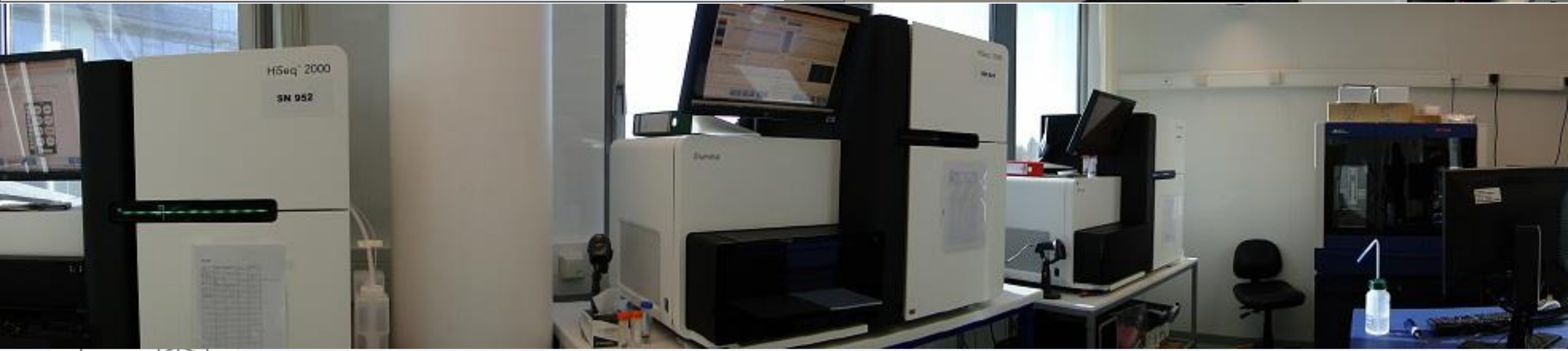
NGS in personalized oncology
Structure of the talk

- NGS projects
- Infrastructure, cloud
- Pipelines and software

Genome Profiling Core Facility (GPCF)

Equipment

- 14 Illumina HiSeq 2000 / 2500
- 2 Illumina MiSeq
- 1 454 FLX
- 2 HiSeq X, 8 more in 2015 (some old HiSeq will go...)



Some Petabase numbers

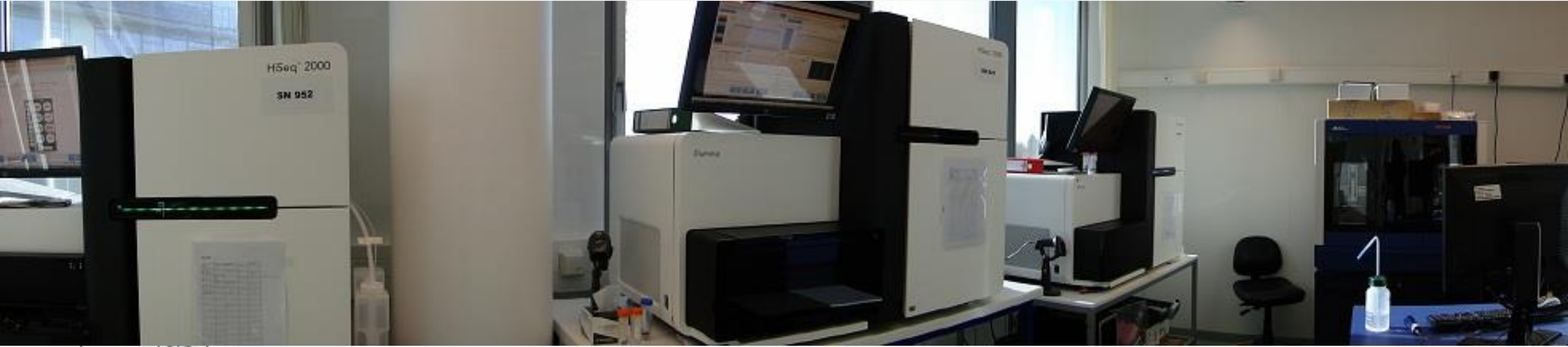
Sequencer are the data producer

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

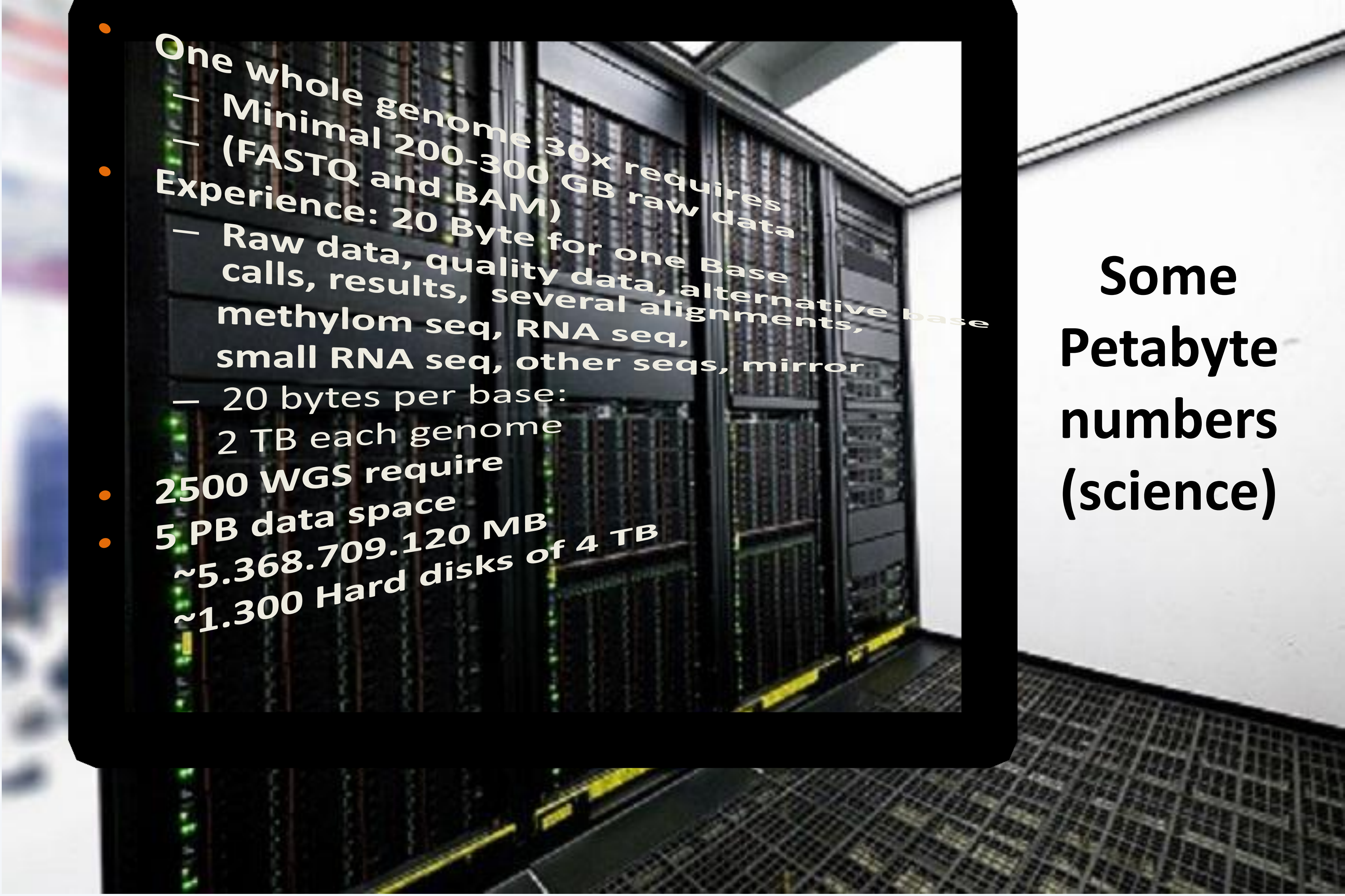


BioQuant MODEL base of LIFE

eils labs

dkfz.

Some Petabyte numbers (science)

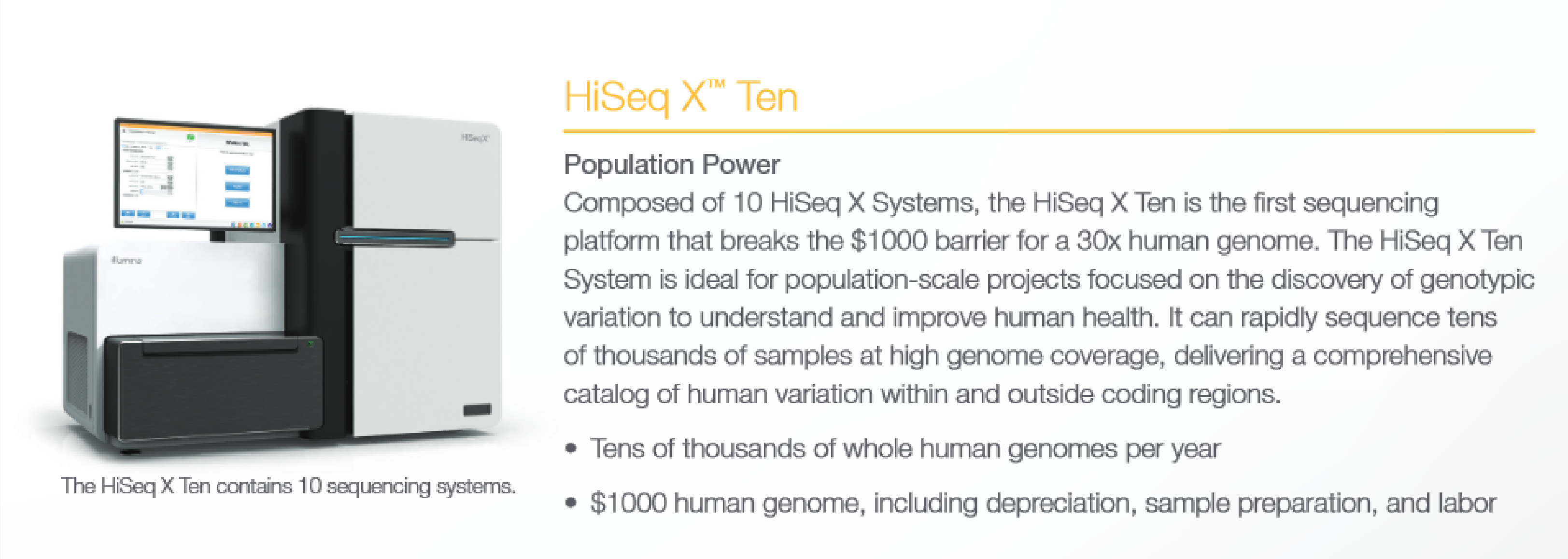


- One whole genome 30x requires
  - Minimal 200-300 GB raw data
  - (FASTQ and BAM)
- Experience: 20 Byte for one Base
  - Raw data, quality data, alternative base calls, results, several alignments, methylom seq, RNA seq, small RNA seq, other seqs, mirror
  - 20 bytes per base: 2 TB each genome
- 2500 WGS require
- 5 PB data space

~5.368.709.120 MB

~1.300 Hard disks of 4 TB

X Ten technology



The HiSeq X Ten contains 10 sequencing systems.

HiSeq X™ Ten

Population Power

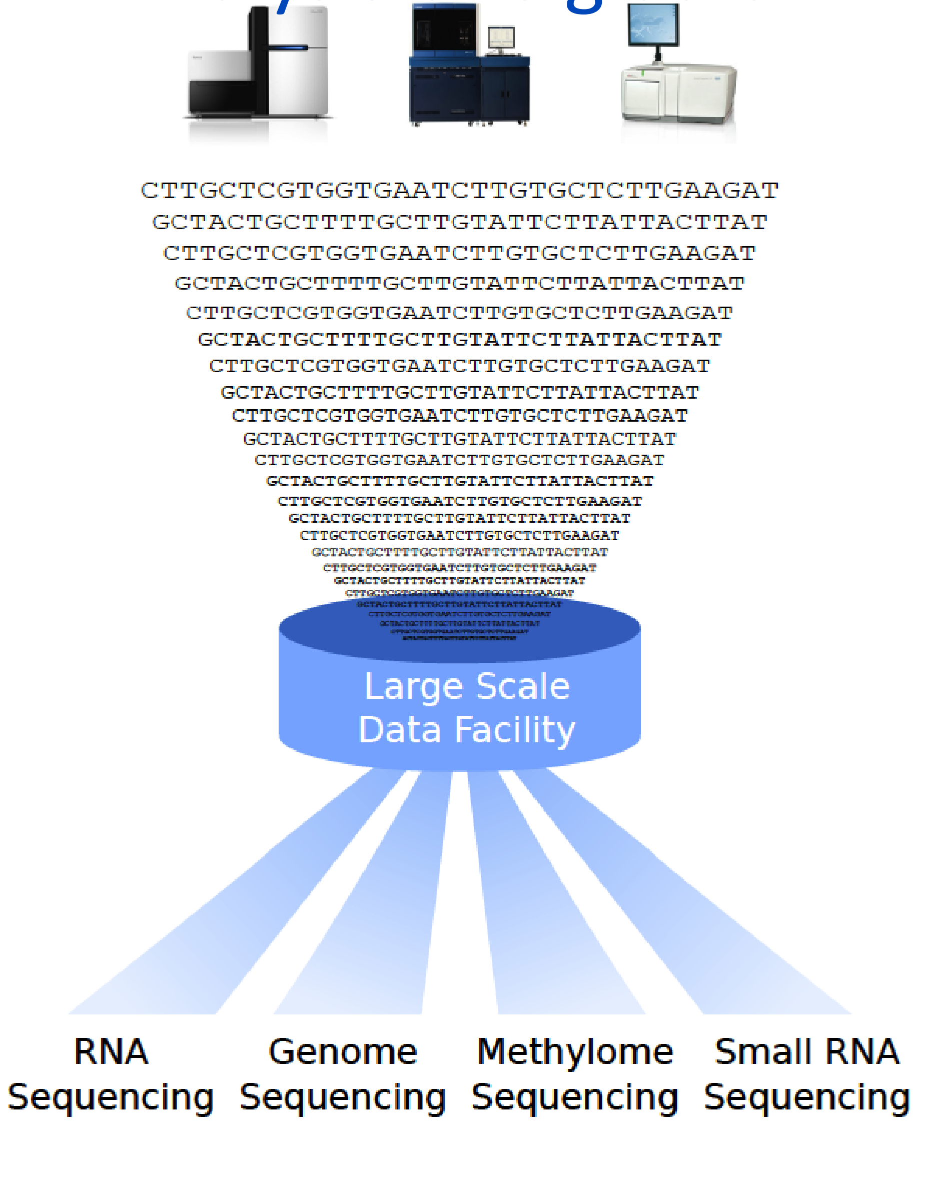
Composed of 10 HiSeq X Systems, the HiSeq X Ten is the first sequencing platform that breaks the $1000 barrier for a 30x human genome. The HiSeq X Ten System is ideal for population-scale projects focused on the discovery of genotypic variation to understand and improve human health. It can rapidly sequence 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

- First capacity genome facility in WGS
- Investment of 10 million Euros / Dollars requested
- It's still research experience
- **Closing gap between clinical research and practice**

- 18.000 patients / year (30x)
- 1,8 PB / year of raw data once
- 3,6PB / year backup
- **Disk space ~5-10 PB per year expectation**

Analysis of big data


CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
CTTGCTCGTGGTGAATCTTGTGCTCTTGAAGAT
GCTACTGCTTTTGCTTGTATTCTTATTACTTAT
Large Scale
Data Facility
RNA
Sequencing
Genome
Sequencing
Methylome
Sequencing
Small RNA
Sequencing

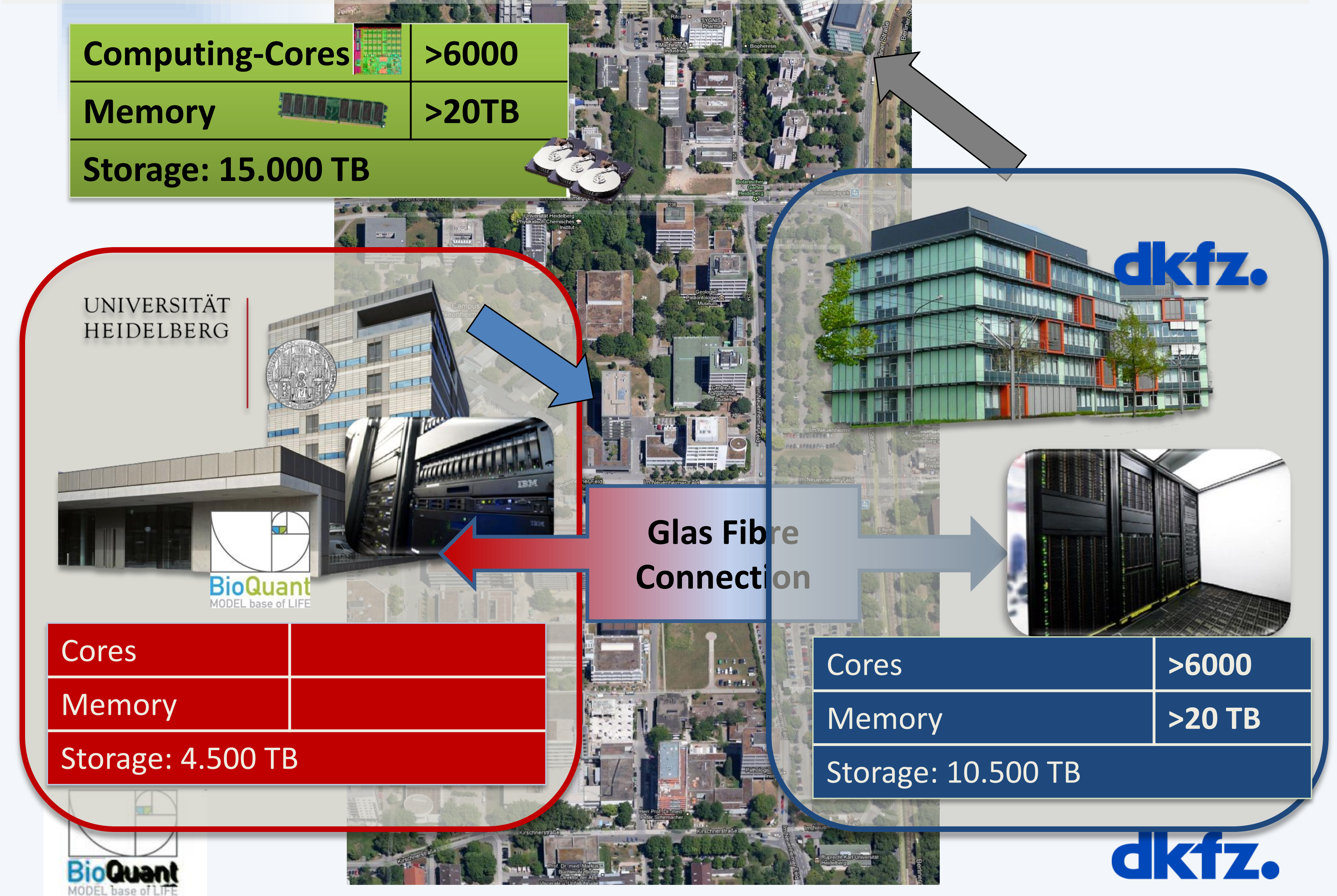
BioQuant MODEL base of LIFE

eilslabs

dkfz.

Usable Computing Capacities Heidelberg Campus

Computing-Cores	**>6000**
Memory	**>20TB**
Storage: 15.000 TB	



Glas Fibre Connection

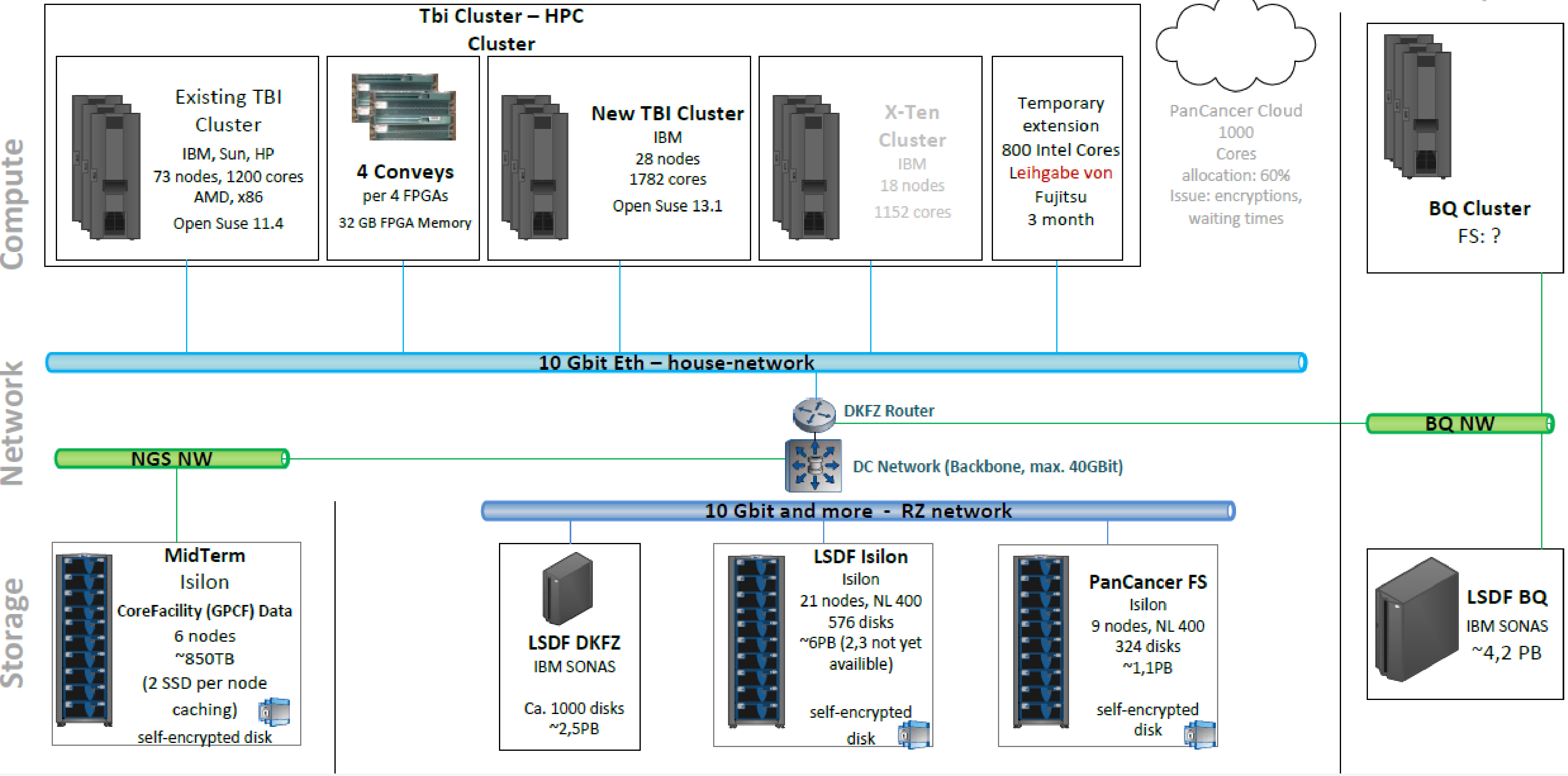
Cores	
Memory	
Storage: 4.500 TB	

Cores	**>6000**
Memory	**>20 TB**
Storage: 10.500 TB	

Bioinformatics Infrastructure without details

DKFZ

BQ



Compute

Tbi Cluster – HPC Cluster

Existing TBI Cluster
IBM, Sun, HP
73 nodes, 1200 cores
AMD, x86
Open Suse 11.4

4 Conveys
per 4 FPGAs
32 GB FPGA Memory

New TBI Cluster
IBM
28 nodes
1782 cores
Open Suse 13.1

X-Ten Cluster
IBM
18 nodes
1152 cores

Temporary extension
800 Intel Cores
Leihgabe von
Fujitsu
3 month

PanCancer Cloud
1000
Cores
allocation: 60%
Issue: encryptions, waiting times

BQ Cluster
FS: ?

Network

10 Gbit Eth – house-network

DKFZ Router

BQ NW

NGS NW

DC Network (Backbone, max. 40GBit)

10 Gbit and more - RZ network

Storage

MidTerm
Isilon
CoreFacility (GPCF) Data
6 nodes
~850TB
(2 SSD per node caching)
self-encrypted disk

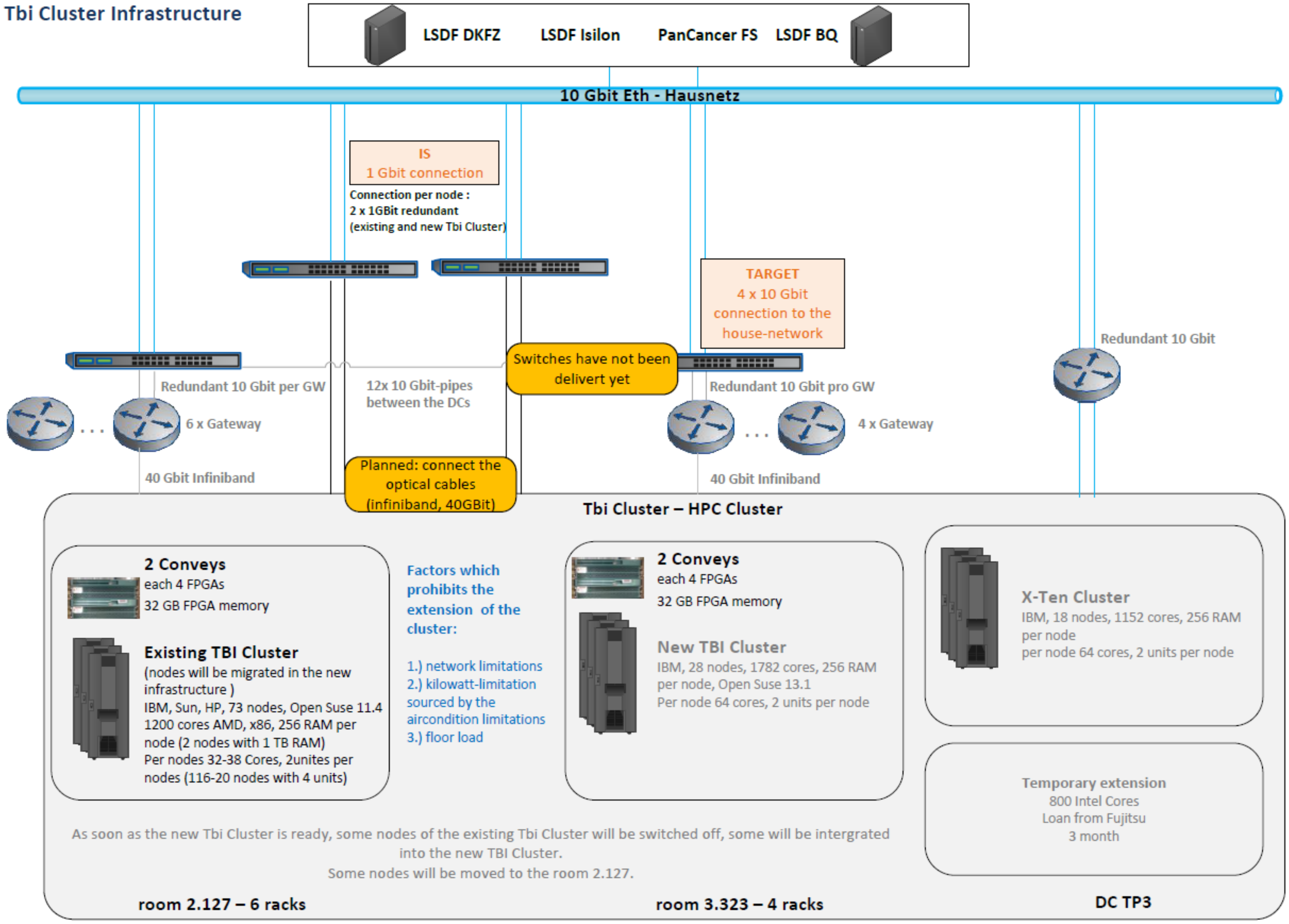
LSDF DKFZ
IBM SONAS
Ca. 1000 disks
~2,5PB

LSDF Isilon
Isilon
21 nodes, NL 400
576 disks
~6PB (2,3 not yet availible)
self-encrypted disk

PanCancer FS
Isilon
9 nodes, NL 400
324 disks
~1,1PB
self-encrypted disk

LSDF BQ
IBM SONAS
~4,2 PB

Tbi Cluster Infrastructure



LSDF DKFZ

LSDF Isilon

PanCancer FS

LSDF BQ

10 Gbit Eth - Hausnetz

IS
1 Gbit connection

Connection per node :
2 x 1GBit redundant
(existing and new Tbi Cluster)

TARGET
4 x 10 Gbit
connection to the
house-network

Switches have not been
delivert yet

Redundant 10 Gbit

Redundant 10 Gbit per GW

12x 10 Gbit-pipes
between the DCs

Redundant 10 Gbit pro GW

. . . 6 x Gateway

. . . 4 x Gateway

40 Gbit Infiniband

Planned: connect the
optical cables
(infiniband, 40GBit)

40 Gbit Infiniband

Tbi Cluster – HPC Cluster

2 Conveys
each 4 FPGAs
32 GB FPGA memory

Existing TBI Cluster
(nodes will be migrated in the new infrastructure)
IBM, Sun, HP, 73 nodes, Open Suse 11.4
1200 cores AMD, x86, 256 RAM per node (2 nodes with 1 TB RAM)
Per nodes 32-38 Cores, 2unites per nodes (116-20 nodes with 4 units)

Factors which prohibits the extension of the cluster:

1.) network limitations
2.) kilowatt-limitation sourced by the aircondition limitations
3.) floor load

2 Conveys
each 4 FPGAs
32 GB FPGA memory

New TBI Cluster
IBM, 28 nodes, 1782 cores, 256 RAM per node, Open Suse 13.1
Per node 64 cores, 2 units per node

X-Ten Cluster
IBM, 18 nodes, 1152 cores, 256 RAM per node
per node 64 cores, 2 units per node

Temporary extension
800 Intel Cores
Loan from Fujitsu
3 month

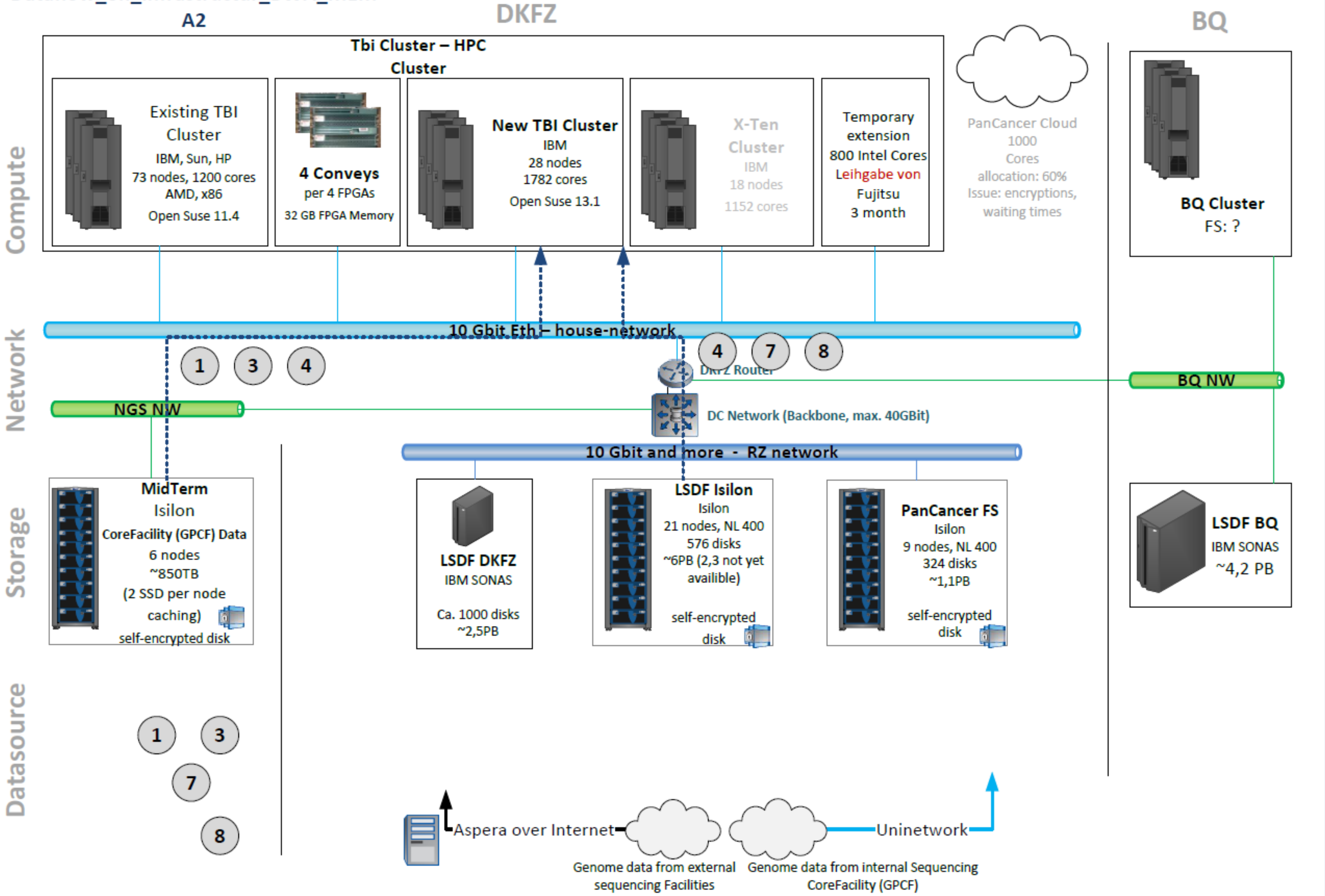
As soon as the new Tbi Cluster is ready, some nodes of the existing Tbi Cluster will be switched off, some will be intergrated into the new TBI Cluster.
Some nodes will be moved to the room 2.127.

room 2.127 – 6 racks

room 3.323 – 4 racks

DC TP3

Dataflow_on_infrastructur_BWA_MEM
A2



DKFZ

BQ

Tbi Cluster – HPC Cluster

Existing TBI Cluster
IBM, Sun, HP
73 nodes, 1200 cores
AMD, x86
Open Suse 11.4

4 Conveys
per 4 FPGAs
32 GB FPGA Memory

New TBI Cluster
IBM
28 nodes
1782 cores
Open Suse 13.1

X-Ten Cluster
IBM
18 nodes
1152 cores

Temporary extension
800 Intel Cores
Leihgabe von Fujitsu
3 month

PanCancer Cloud
1000 Cores
allocation: 60%
Issue: encryptions, waiting times

BQ Cluster
FS: ?

Compute

Network

Storage

Datasource

10 Gbit Eth – house-network

1 3 4

4 7 8

DKFZ Router

BQ NW

NGS NW

DC Network (Backbone, max. 40GBit)

10 Gbit and more - RZ network

MidTerm Isilon
CoreFacility (GPCF) Data
6 nodes
~850TB
(2 SSD per node caching)
self-encrypted disk

LSDF DKFZ
IBM SONAS
Ca. 1000 disks
~2,5PB

LSDF Isilon
Isilon
21 nodes, NL 400
576 disks
~6PB (2,3 not yet availible)
self-encrypted disk

PanCancer FS
Isilon
9 nodes, NL 400
324 disks
~1,1PB
self-encrypted disk

LSDF BQ
IBM SONAS
~4,2 PB

1 3 7 8

Aspera over Internet

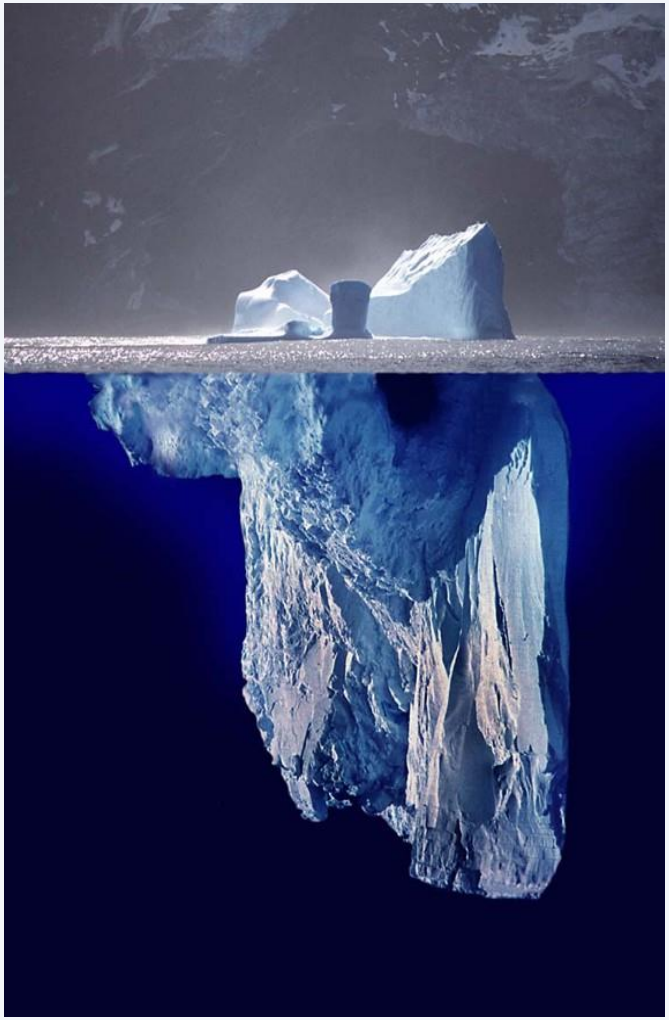
Uninetwork

Genome data from external sequencing Facilities

Genome data from internal Sequencing CoreFacility (GPCF)

Cloud approaches
The PanCancer Project

- Goals:
  - Characterization of commonalities, differences of cancer types
  - Understand what's going on in the 95% of the cancer genome that isn't protein-coding
    - Non-coding RNAs
    - Regulatory elements
    - Amplifications/deletions & other structural changes



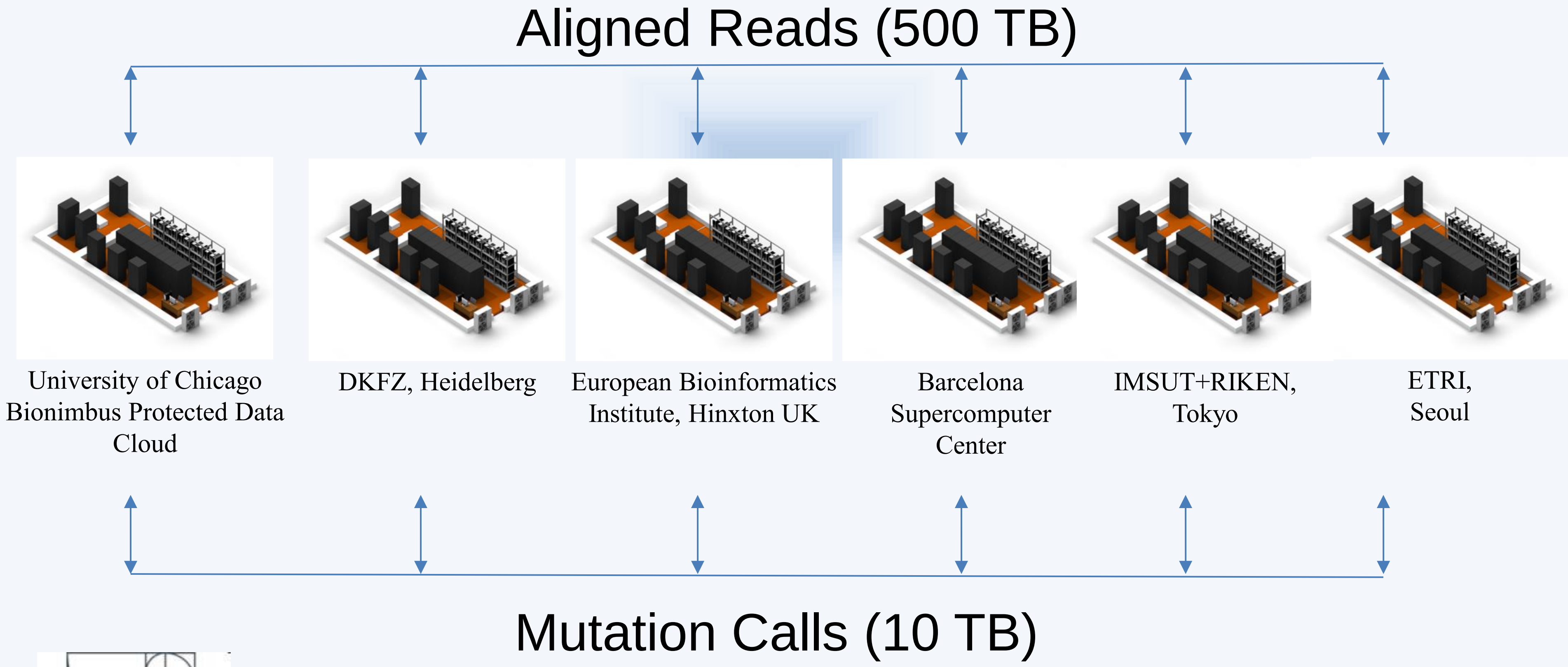
- Resources:
  - >2500 whole genome tumor/normal pairs from ICGC and TCGA
  - 15 working groups
  - 130 research subprojects

BioQuant MODEL base of LIFE

eils labs

dkfz.

Phase II: Synchronize Alignments & Mutation Calls



Aligned Reads (500 TB)

University of Chicago Bionimbus Protected Data Cloud

DKFZ, Heidelberg

European Bioinformatics Institute, Hinxton UK

Barcelona Supercomputer Center

IMSUT+RIKEN, Tokyo

ETRI, Seoul

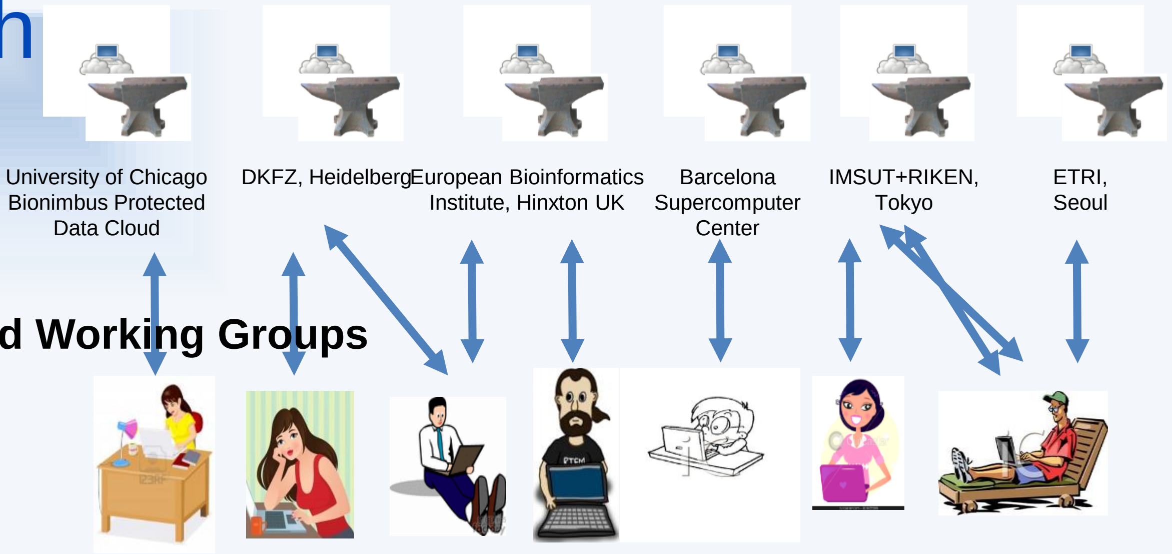
Mutation Calls (10 TB)

BioQuant MODEL base of LIFE

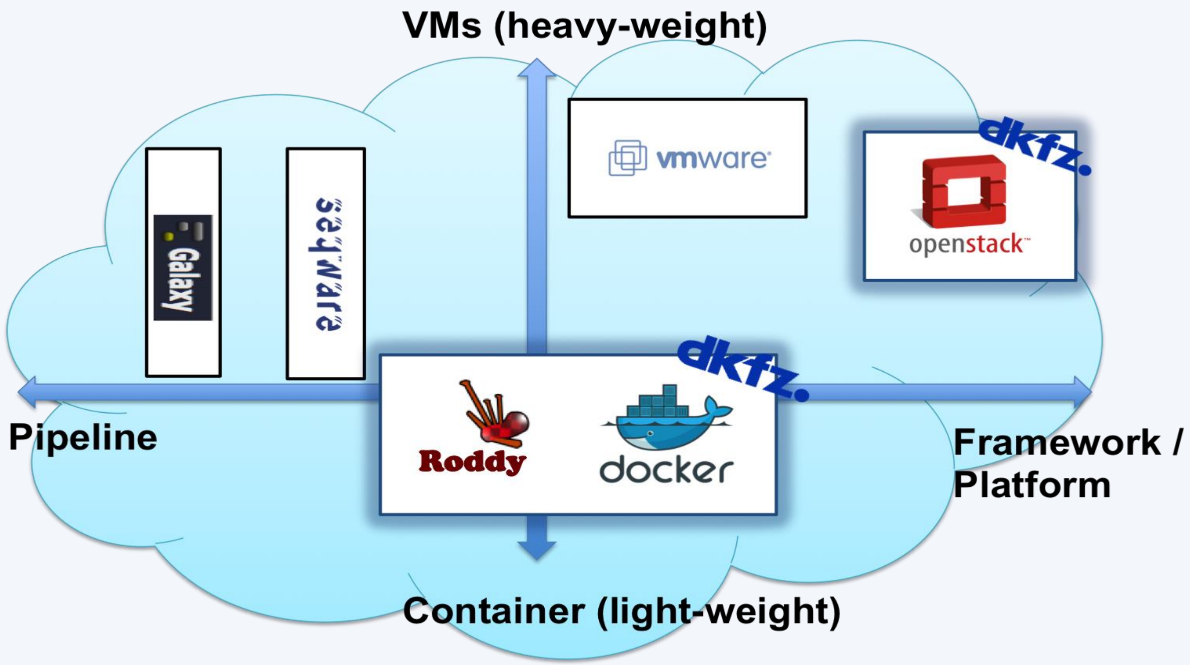
eilslabs

dkfz.

Cloud Approach PanCancer


University of Chicago Bionimbus Protected Data Cloud
DKFZ, Heidelberg
European Bioinformatics Institute, Hinxton UK
Barcelona Supercomputer Center
IMSUT+RIKEN, Tokyo
ETRI, Seoul

ICGC Researchers and Working Groups



The ICGC Pan-Cancer Project: Participation of eilslabs / DKFZ

ICGC

Genomes (top 3 centers)

dkfz. wellcome trust sanger institute The Cancer Genome Atlas

Variant Calling Pipelines

BROAD INSTITUTE dkfz. EMBL wellcome trust sanger institute

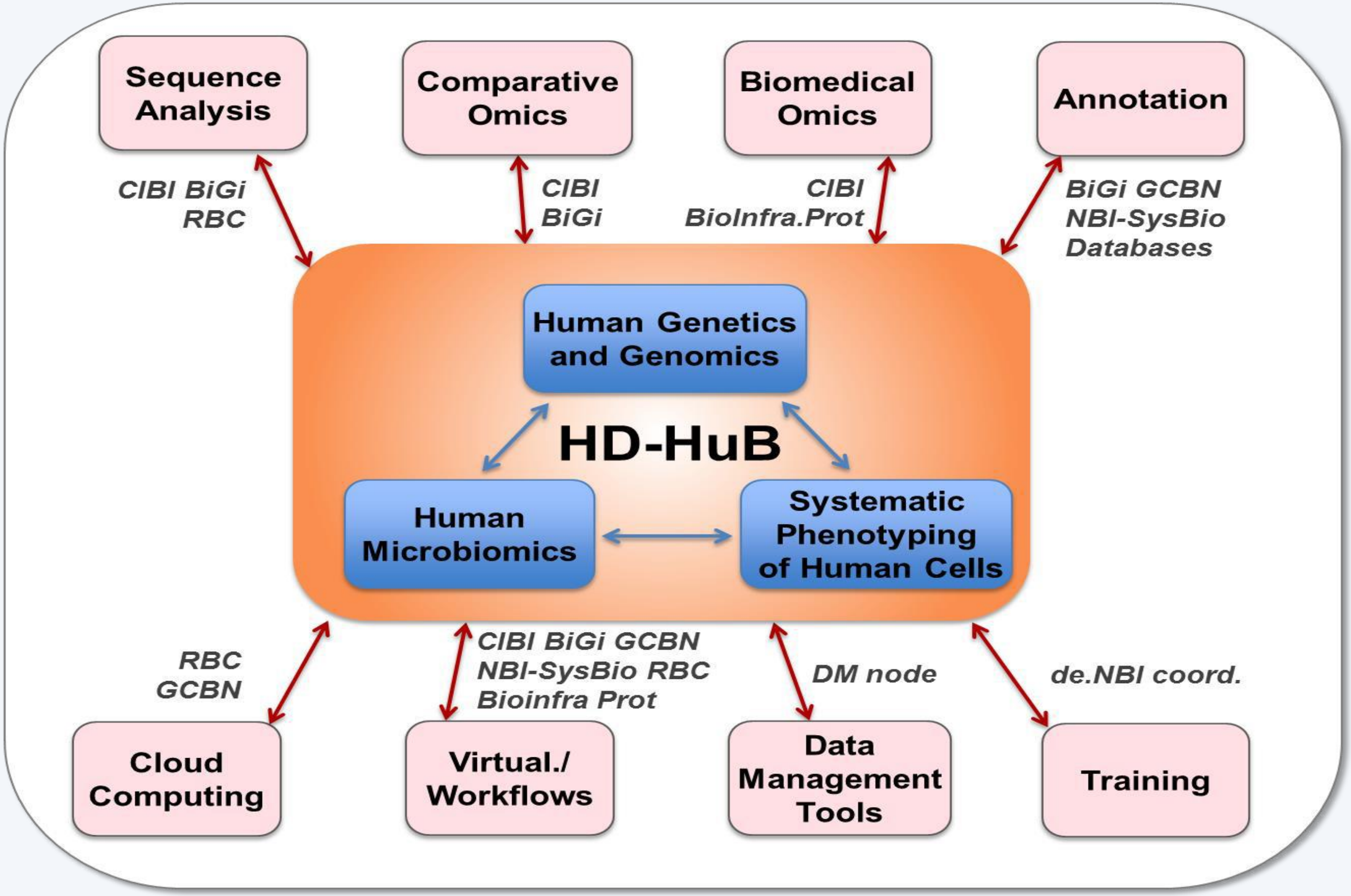
Data and Computing Centers

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)

Cloud Approach de.NBI

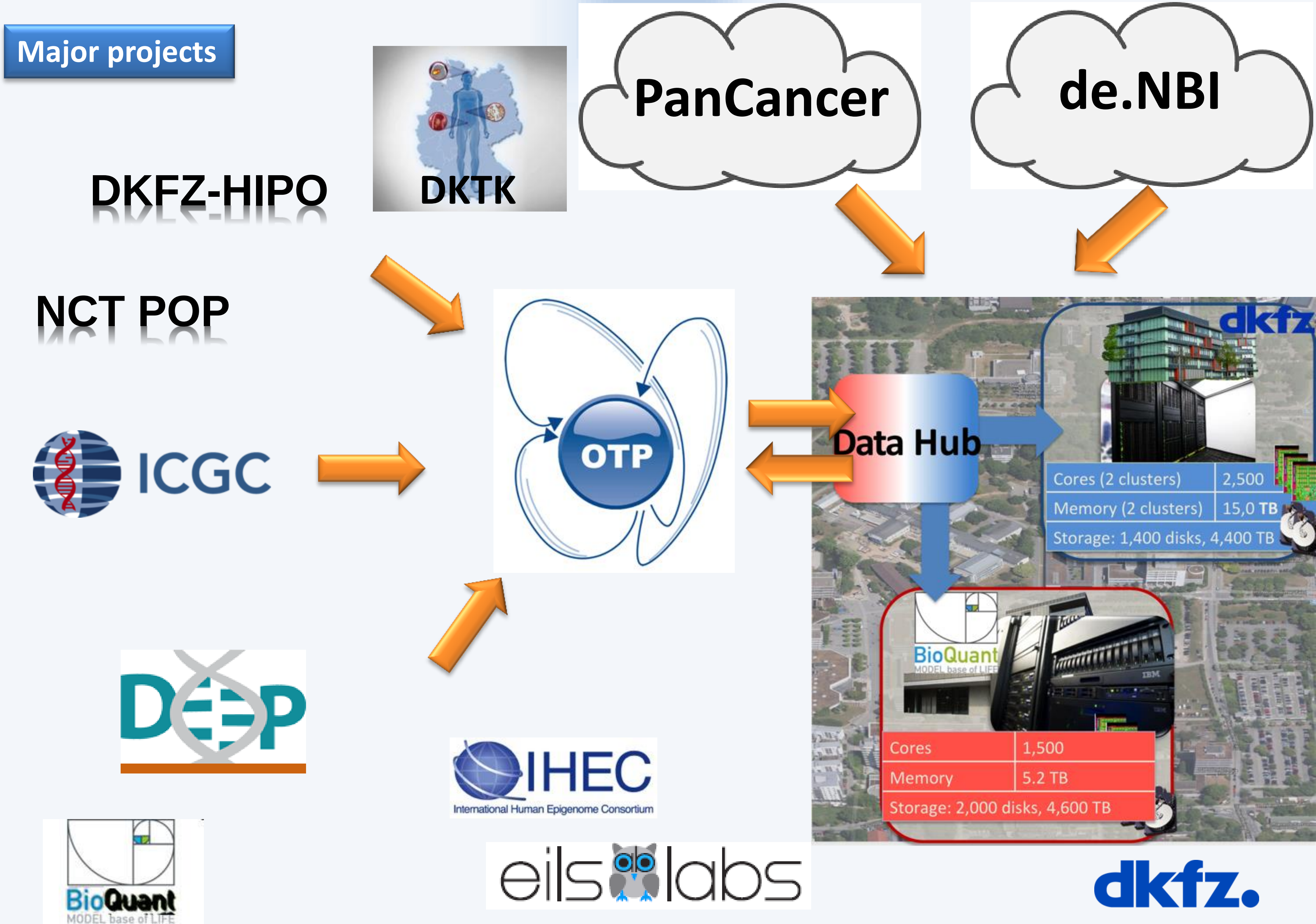


BioQuant MODEL base of LIFE

eils labs

dkfz.

Data flow at DKFZ



Major projects

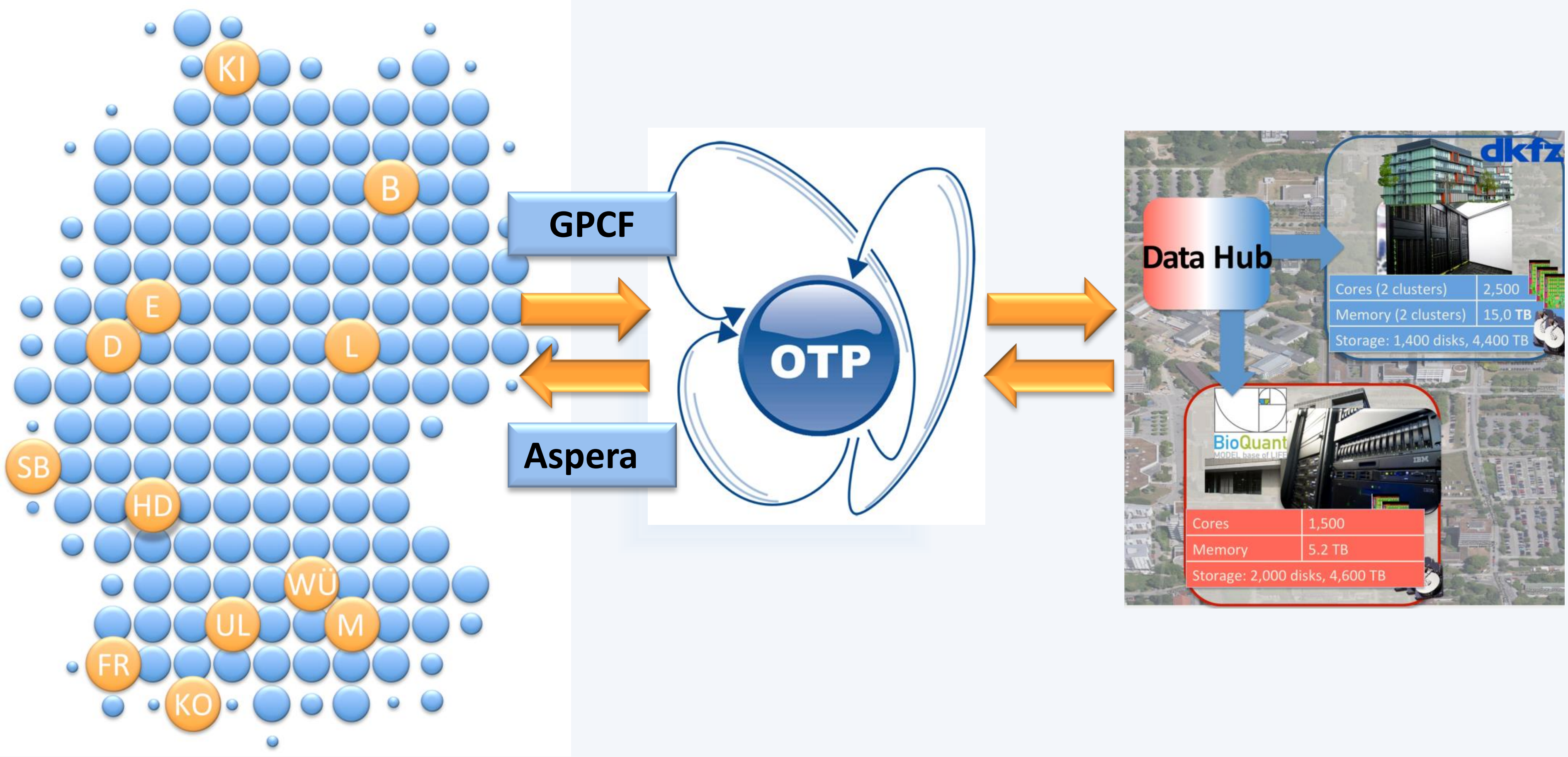
DKFZ-HIPO

NCT POP

Cores (2 clusters)	2,500
Memory (2 clusters)	15,0 TB
Storage: 1,400 disks, 4,400 TB	

Cores	1,500
Memory	5.2 TB
Storage: 2,000 disks, 4,600 TB	

Sequence centers in Germany



Cores (2 clusters)	2,500
Memory (2 clusters)	15,0 TB
Storage: 1,400 disks, 4,400 TB	

Cores	1,500
Memory	5.2 TB
Storage: 2,000 disks, 4,600 TB	

Receiving

NGS fastq data

→

Processing

Alignment & Merging

→

Providing

Structured data

NGS in personalized oncology
Structure of the talk

– NGS projects

– Infrastructure, cloud

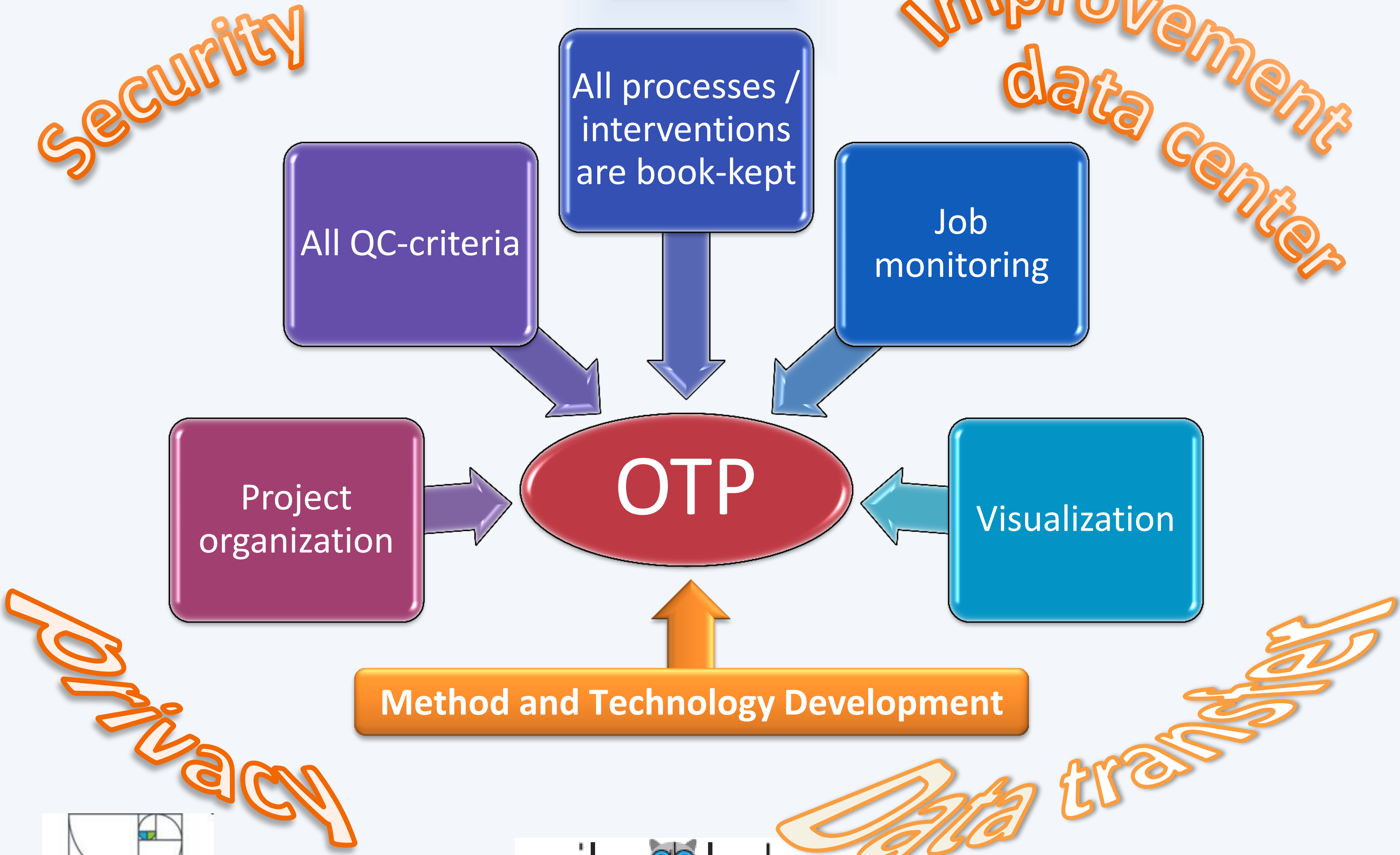
– Pipelines and software

BioQuant MODEL base of LIFE

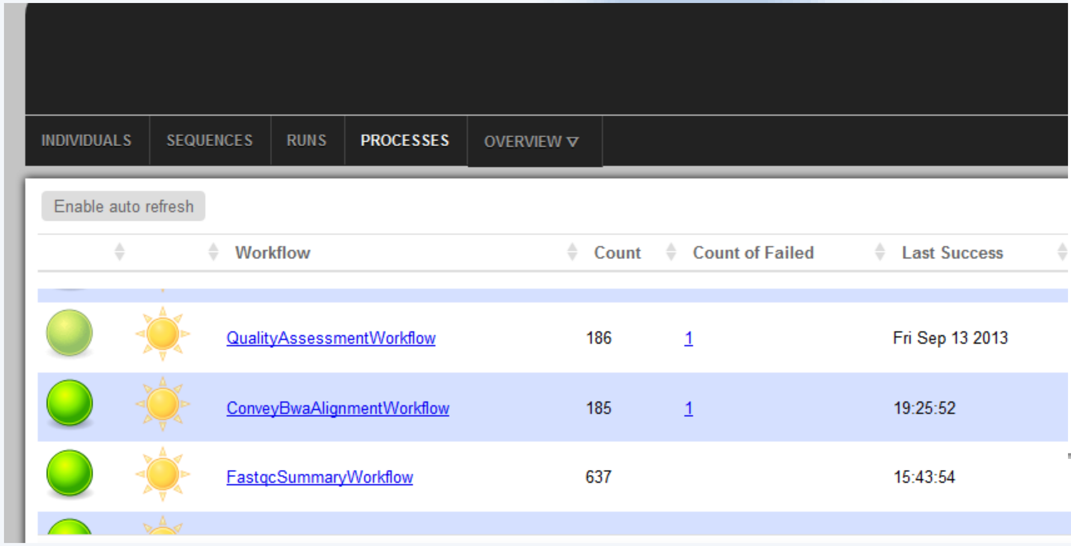
eils labs

dkfz.

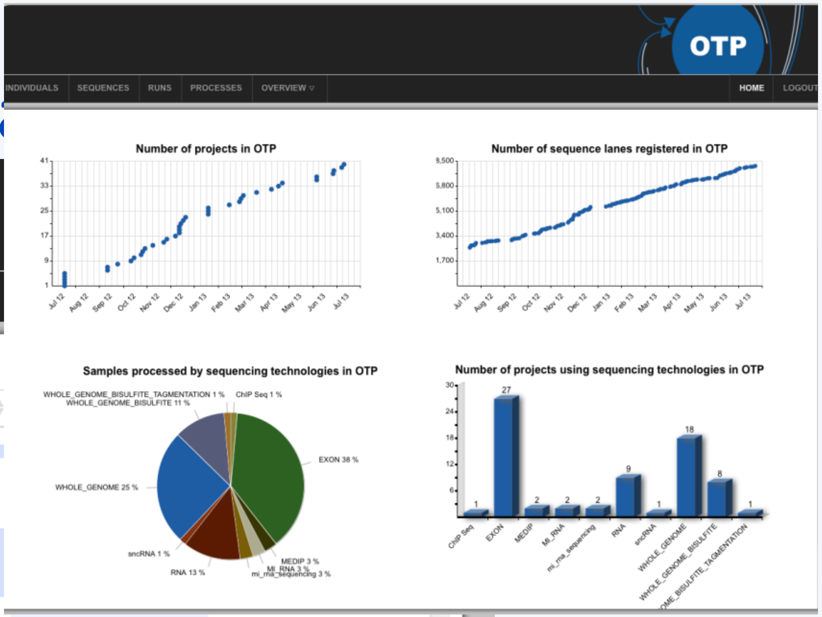
OTP: Central research platform



OTP: processing fr

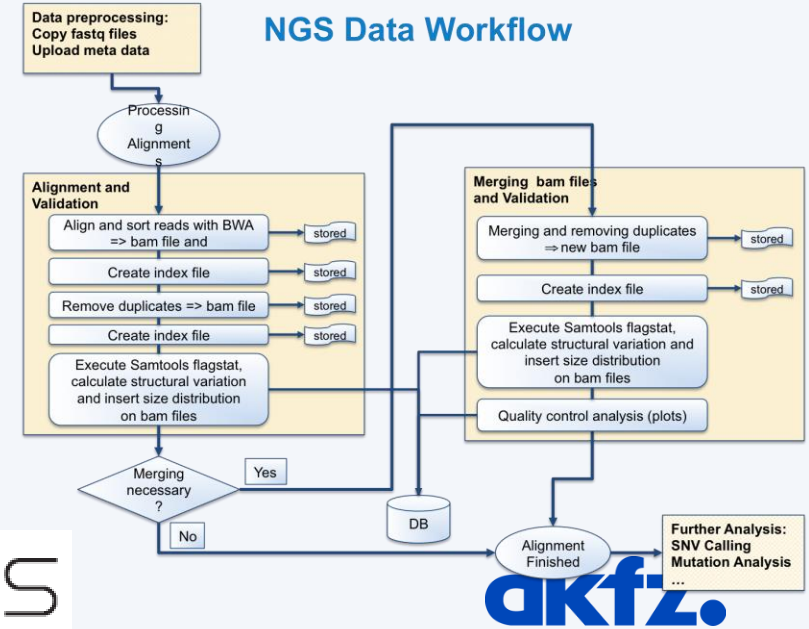


		Workflow	Count	Count of Failed	Last Success		
●	☀	QualityAssessmentWorkflow	186	1	Fri Sep 13 2013		
●	☀	ConveyBwaAlignmentWorkflow	185	1	19:25:52		
●	☀	FastqcSummaryWorkflow	637		15:43:54	Thu Aug 01 2013	1 sec 248 msec



- Processing frameworks for huge NGS projects:
  - Project organization
  - To speed-up: All routine jobs run automatically
  - No more manual shell scripts
  - Alignment and QC done by pushing a buttom
  - Automatic information when a process was broken

NGS Data Workflow



Data preprocessing:
Copy fastq files
Upload meta data

Processing Alignments

Alignment and Validation

- Align and sort reads with BWA => bam file and → stored
- Create index file → stored
- Remove duplicates => bam file → stored
- Create index file → stored
- Execute Samtools flagstat, calculate structural variation and insert size distribution on bam files

Merging necessary? Yes / No

DB

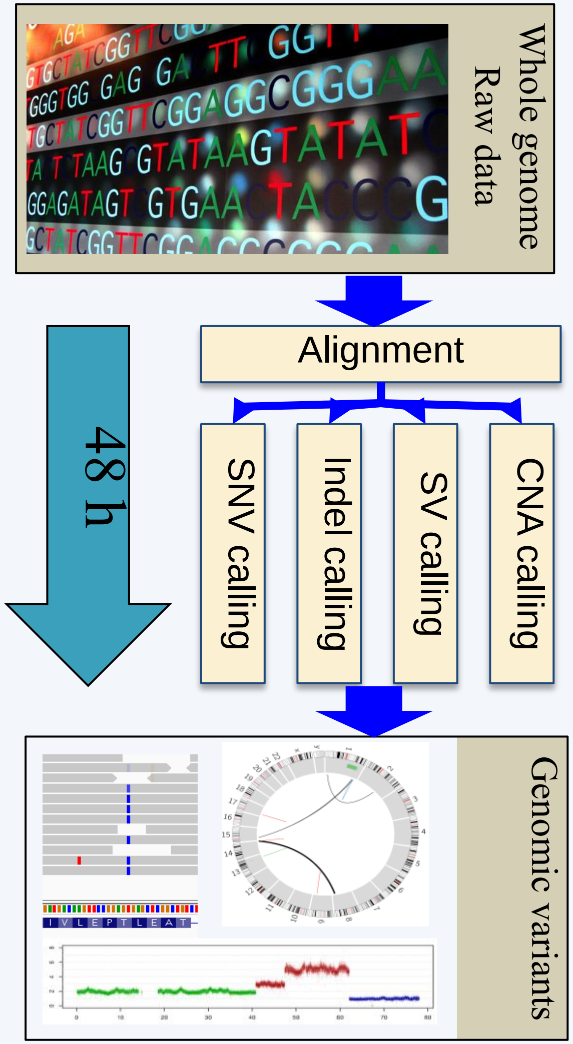
Merging bam files and Validation

- Merging and removing duplicates ⇒ new bam file → stored
- Create index file → stored
- Execute Samtools flagstat, calculate structural variation and insert size distribution on bam files
- Quality control analysis (plots)

Alignment Finished

Further Analysis:
SNV Calling
Mutation Analysis
...

Processing of NGS data



Whole genome
Raw data

Alignment

48 h

SNV calling

Indel calling

SV calling

CNA calling

Genomic variants

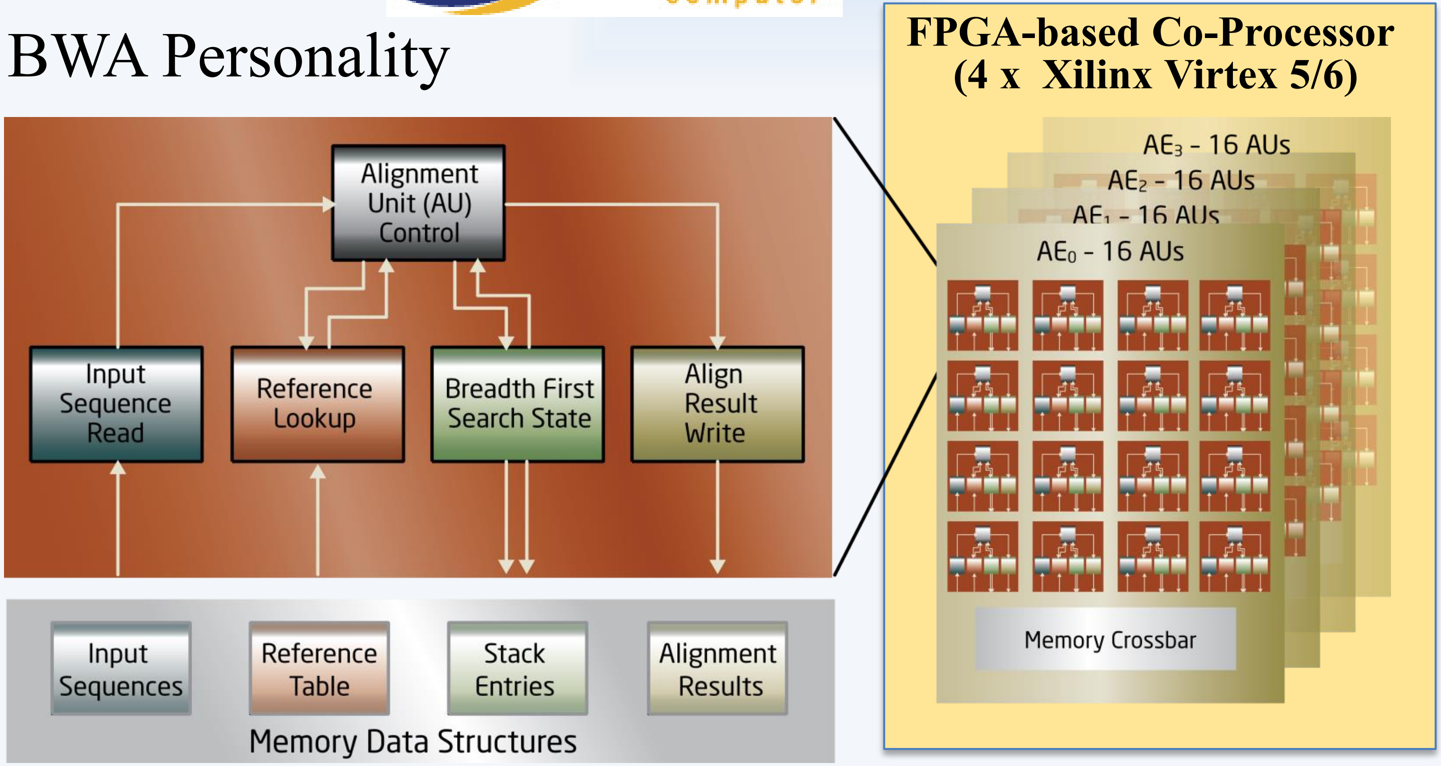
- processing of 30x whole genome from raw data to variant calls in < 48 h

- accelerated by:
  - streamlined process (e.g. merge + mark duplicates in one step)
  - use of pipes to avoid I/O (input/output: here writing to / reading from disks)
  - hardware-accelerated alignment (Convey)

Variants	Tool
SNVs	DKFZ SNV pipeline
Small Indels	Platypus pipeline (Rimmer et al.)
CNVs	ICGC ACEseq

Example of acceleration : Convey HC-2

BWA Personality



- Implemented in hardware on coprocessor FPGAs
- 64 alignment units with 32-stage pipeline = 2048 simultaneous alignment operations
- **20x speed-up** compared to alignment with 8 cores
- **saves 800 CPU-hours** per whole genome pair (tumor + control)
- reduces start-to-end-time of QC-pipeline from 62 hours to 38 hours on average

Overview about projects, samples
Organizational issues

- Typical obvious questions:
  - Which sequencing type was done on my sample?
  - Was the sequencing deep enough, how big is the coverage?
  - Where is my smaple actually processed?
  - Where is my data?
  - What‘s going on?
  - What are the results of my NGS experiment?

More important issues

- Typical not obvious questions:
  - Who has the permission to distribute the data?
  - Who can be asked?
  - Whom has the data given at which time?
  - Has person xyz inhouse the permission to access the data?
  - Who is responsible: the coordinator, the PI, the Professor
  - Is sequencing data personalized data?

- These aspects are often underestimated
- => Big or many projects lead to communication stress
- => Data privacy, policies and ethic rules

DMG Heads	Chris Lawerenz, Jürgen Eils
DMG Developer Team	Manuel Prinz, Alexander Balz Pavel Komardin, Phillip Kensche Eva Reisinger, Stefan Borufka Gideon Zipprich, Charles Imbusch Florian Kärcher, Amal Mertens Andreas Kling, Jonas Stadter Jan Matuschek, Jules Kerssemakers
DMG Support Team	Ingrid Scholz Serkan Oelmez, Bärbel Felder Christina Jaeger-Schmidt