HiGHmed
Medical Informatics

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Datenqualität in der Bioinformatik / Genomdaten

03.05.2018

Roland Eils

Heidelberg
University Hospital

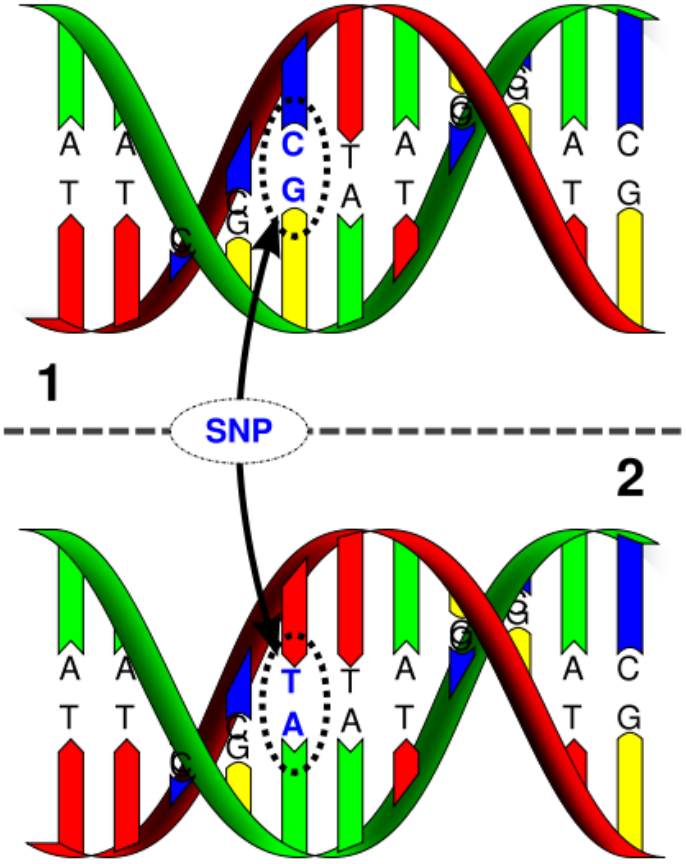
BERLINER
INSTITUT FÜR
GESUNDHEITS
FORSCHUNG
Charité & Max-Delbrück-Centrum

Wir sind alle (fast) gleich

HiGHmed
Medical Informatics

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Variationen im humanen Genom



http://www.science.marshall.edu/murraye/341/Images/416px-Dna-SNP_svg.png

Indivuuen unterschieiden sich an jeder 1000. Position im Genom

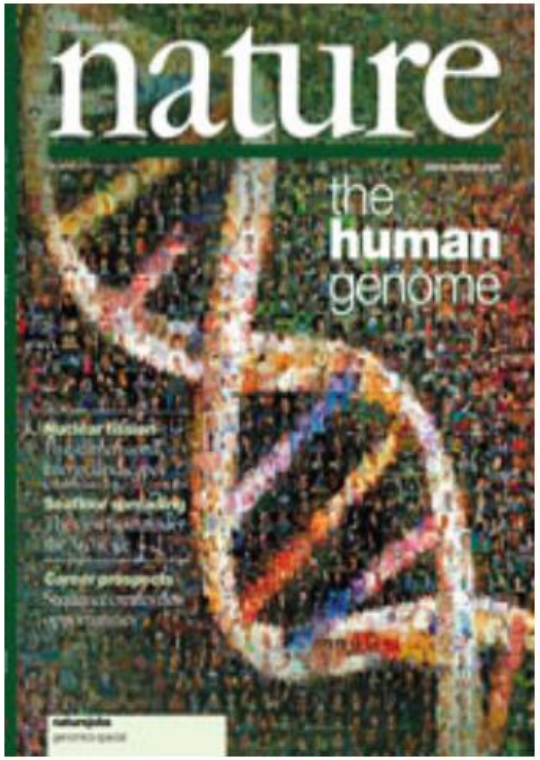
Polymorphismen
(vererbte Punktmutationen)

Somatische Vara**nten**
(erworbene Punktmutationen)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Evolution of Large-scale Genome Analysis

- 2000: Human genome working drafts
  - All data freely released
- Project took about 10 years and cost about $3 billion



- 2008: Major genome centers can sequence the same number of base pairs as were produced for the HGP

Every 16 hours

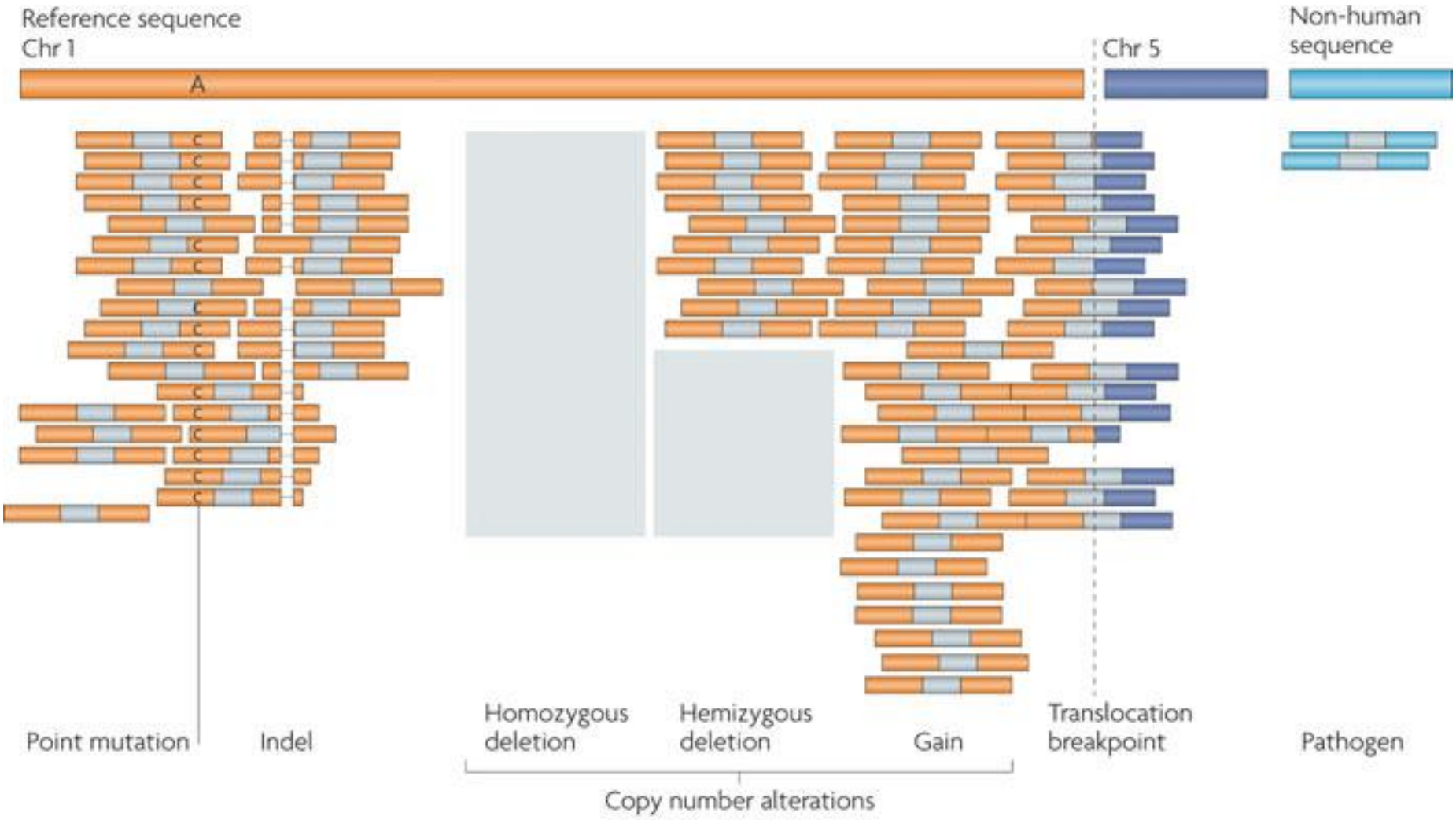
~~***Every day***~~

~~***Every 2.5 days***~~

- ~~***Every 4 days***~~

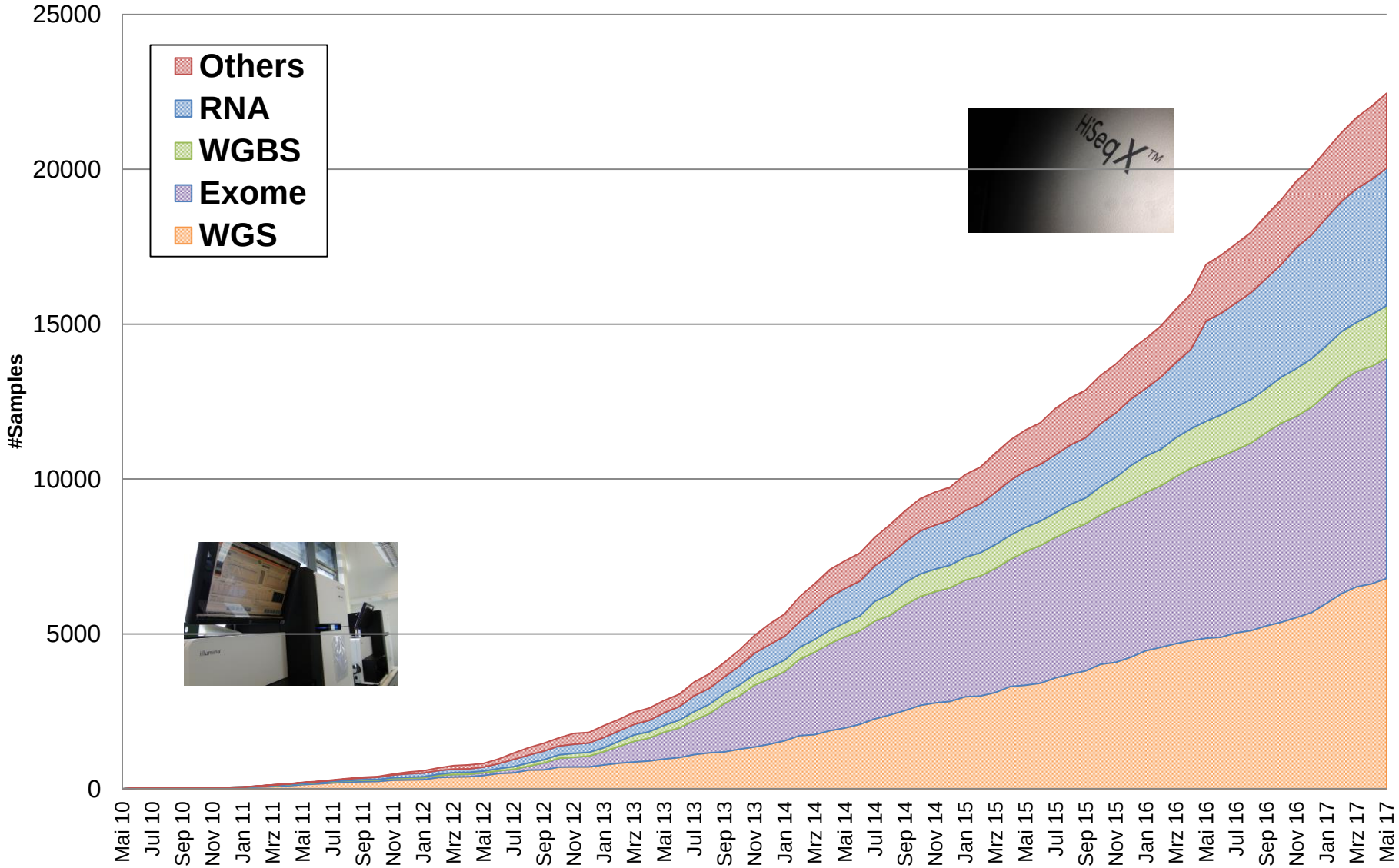
Varianten im Genom

HiGHmed
Medical Informatics



Nature Reviews | Genetics

Number of Samples (approx. 25,000) sequenced in Heidelberg (clinical) sequencing program



Big Data in Genomics: eilslabs vs. Facebook

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)

11 Terabytes per day

International Cancer Genome Consortium

Login | Create your ICGC account

ICGC Search

Overview | Cancer Genome Projects | Committees | Policies and Guidelines | Media | Contacts

International Cancer Genome Consortium

- Brain Cancer — United States
- Breast Cancer — European Union / United Kingdom
- Breast Cancer — France
- Breast Cancer — United Kingdom
- Chronic Lymphocytic Leukemia — Spain
- Colon Cancer — United States
- Gastric Cancer — China
- Leukemia — United States
- Liver Cancer — France
- Liver Cancer — Japan
- Lung Cancer — United States

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.

90 projects committed

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

HTML

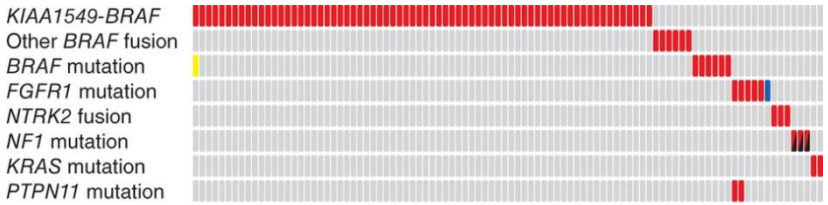
- Lung Cancer — United States
- Malignant Lymphoma — Germany
- Oral Cancer — India
- Ovarian Cancer — Australia
- Ovarian Cancer — United States
- Pancreatic Cancer — Australia
- Pancreatic Cancer — Canada
- Pediatric Brain Tumors — Germany
- Prostate Cancer — Canada
- Prostate Cancer — Germany
- Rare Pancreatic Tumors — Italy
- Renal Cancer — European Union / France

What have we learnt so far?

PedBrain Tumor

Dissecting the genomic complexity underlying medulloblastoma.
Nature 2012

Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma
Nature Genetics 2013

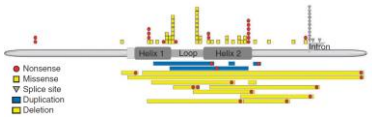


Decoding the regulatory landscape of meulloblastoma by methylation sequencing
Nature 2014

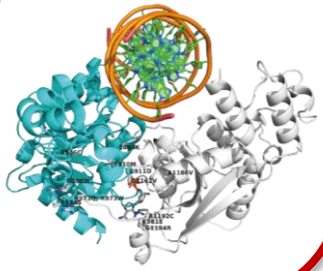
Comprehensive molecular profiling identifies novel drivers and subtypes underlying medulloblastoma
Nature 2017

Malignant Lymphoma

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



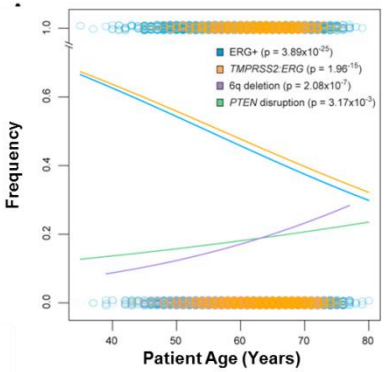
Linking differential DNA methylation in gene regions with the mutational landscape of Burkitt lymphoma
Nature Genetics 2015



Early Onset Prostate Carcinoma

Androgen driven somatic alteration landscape
Cancer Cell 2013

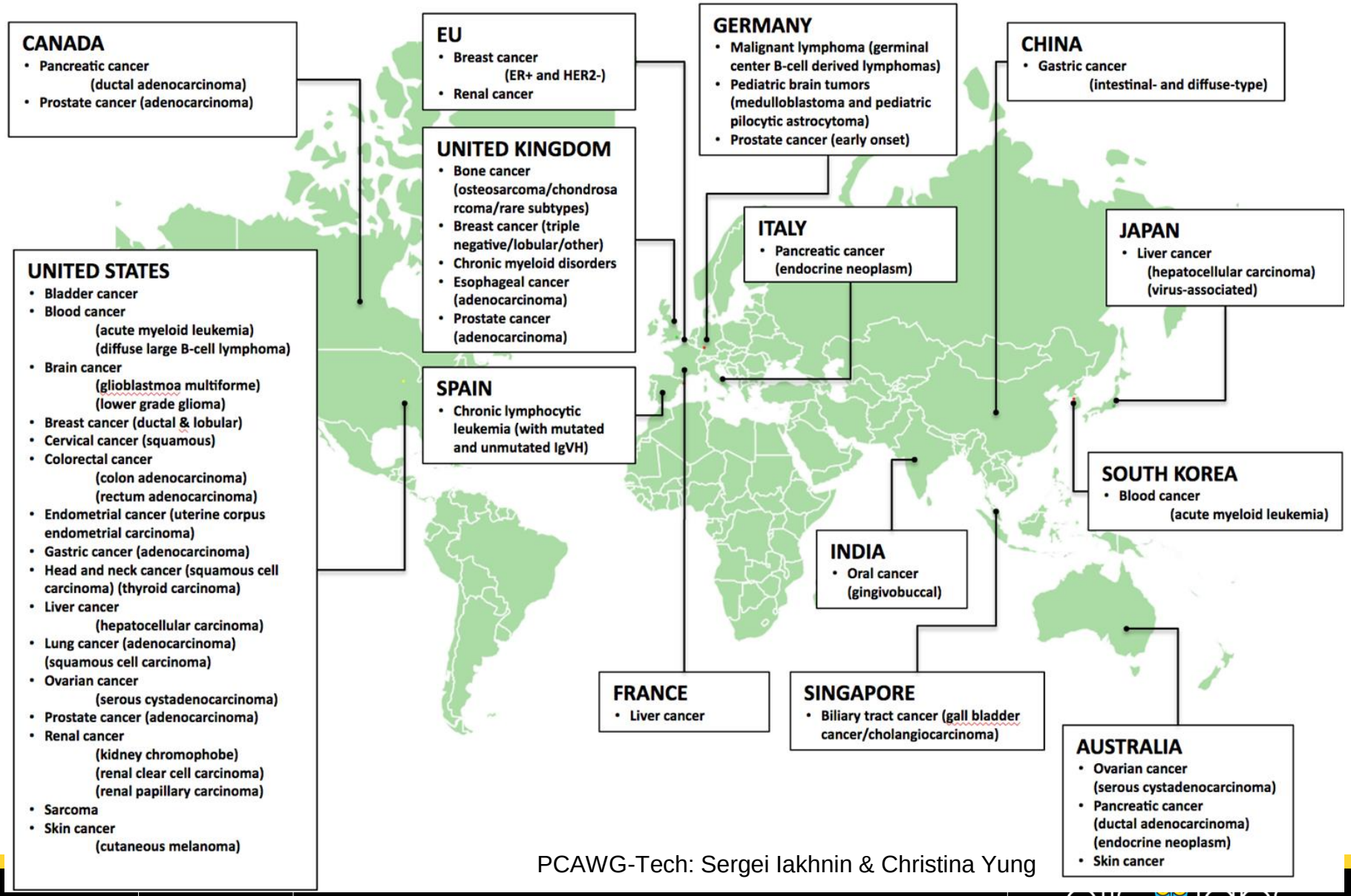
BAZ2A is involved in epigenetic alterations in prostate cancer
Nature Genetics, 2015



Impact des ICGC-Projekts: >200 Publikationen

Jurisdiction	Publications
Germany	74
Spain	65
Japan	47
Canada	20
UK	8
France	4
Australia	4
Saudia Arabia	2
Singapore	1

Weltweites PanCancer – Projekt: 2834 Tumorgenome aus 48 Projekten in 14 Rechtsgebieten, 20 primäre Krebstypen



PCAWG-Tech: Sergei Iakhnin & Christina Yung

Technische Herausforderungen

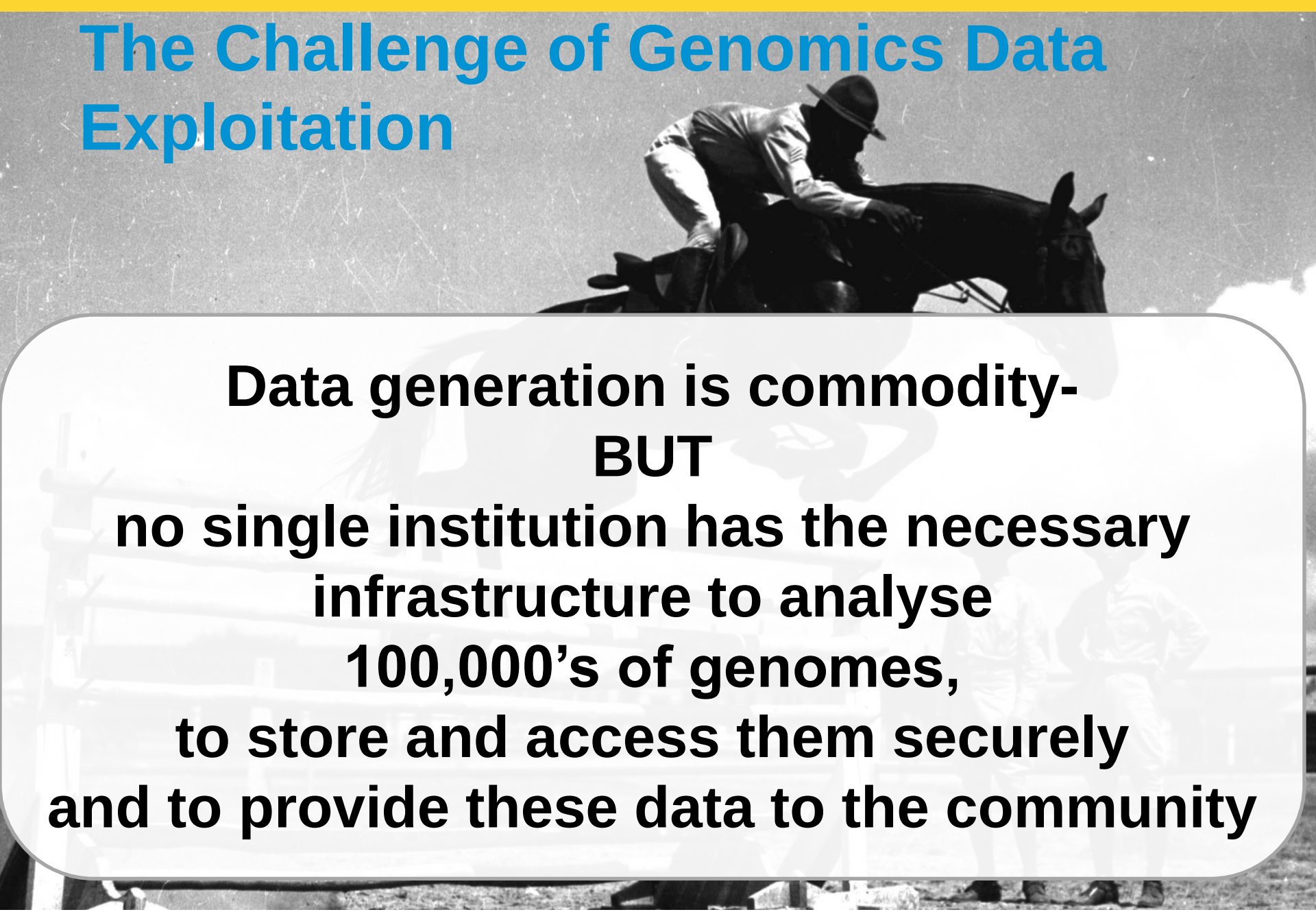
- 2800 pairs of whole genomes amount to **~800TB raw data**
  - 1 WGS alignment & 3 core variant calling workflows

Workflow	Compute (Cores / RAM)	Average runtimes	Storage per donor
BWA-MEM alignment	8 /16GB	5 days/specimen x 2	240GB
Sanger	8 / 32GB	4 days / donor	2GB
DKFZ/EMBL	16 / 64GB	2 days / donor	5GB
Broad	32 / 256GB	3 days / donor	35GB
Total per donor		19 days	282GB
Total for 2800 donors		**>53,000 days (145 years)**	**~800TB (30 years HD movie)**

Over **700** researchers and **130** projects organized into **16** Research Working Groups.

PCAWG-Tech: Junjun Zhang

The Challenge of Genomics Data Exploitation



Data generation is commodity-
BUT
no single institution has the necessary infrastructure to analyse 100,000's of genomes,
to store and access them securely
and to provide these data to the community

Ethical and legal issues related to genome clouds

UNIVERSITÄT HEIDELBERG | ZUKUNFT SEIT 1386

MARSILIUS KOLLEG

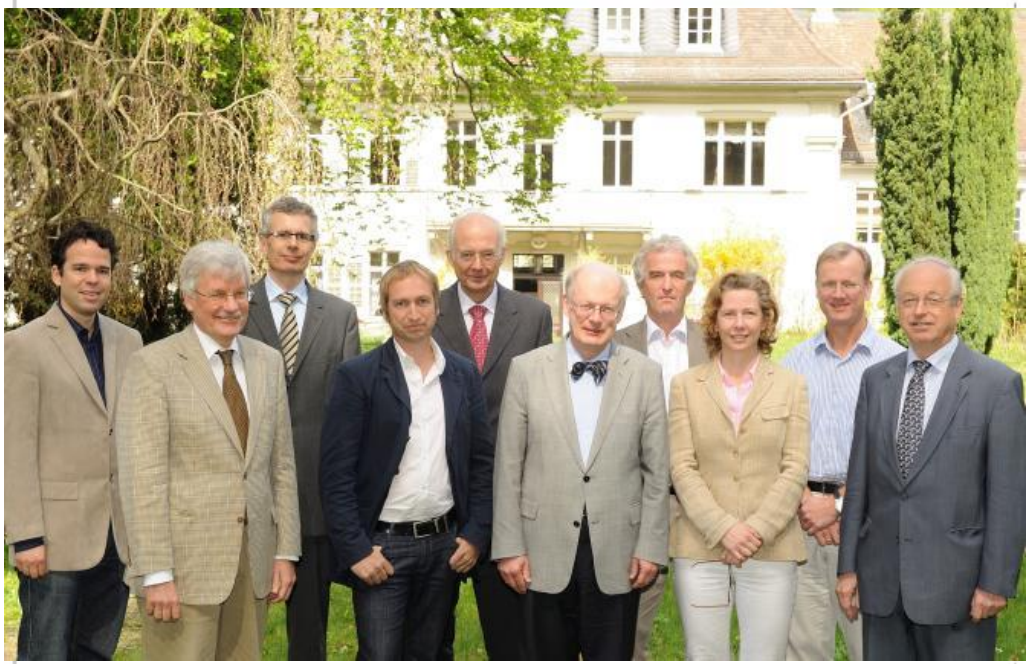
ETHISCHE UND RECHTLICHE ASPEKTE DER TOTALSEQUENZIERUNG DES MENSCHLICHEN GENOMS

- **Home**
- Activities
  - Workshop
  - Press Conference
  - Symposium
  - Lecture with Prof. Knoppers
  - Opening with Prof. Bartram
- Publications
- Project Description
- Project Group
- Project Speaker
- Project Management
- Research Staff
- Associated Project Members
- Further Information
- Ethical and Legal Issues
- Reports
- Research Centres and Networks

Contact | Deutsch

Home > Marsilius-Kolleg > EURAT Home >

EURAT - Ethical and Legal Aspects of Whole Genome Sequencing



NEWS

Comments on the Code

Prof. Rehmann-Sutter and Dr. Mahr discuss the EURAT-Code in a book chapter.

Hyperlink (August 2015)

On incidental findings

Members of the EURAT-group and other researchers register an absence of incidental findings in genomic research and discuss the consequences for the ethical debate.

Hyperlink (July 2015)

On big data

Genomics will be one of the largest data producers - EURAT-members Prof. von Kalle and Prof. Eils comment this development.

Hyperlink (July 2015)

On cloud-computing

MARSILIUS KOLLEG

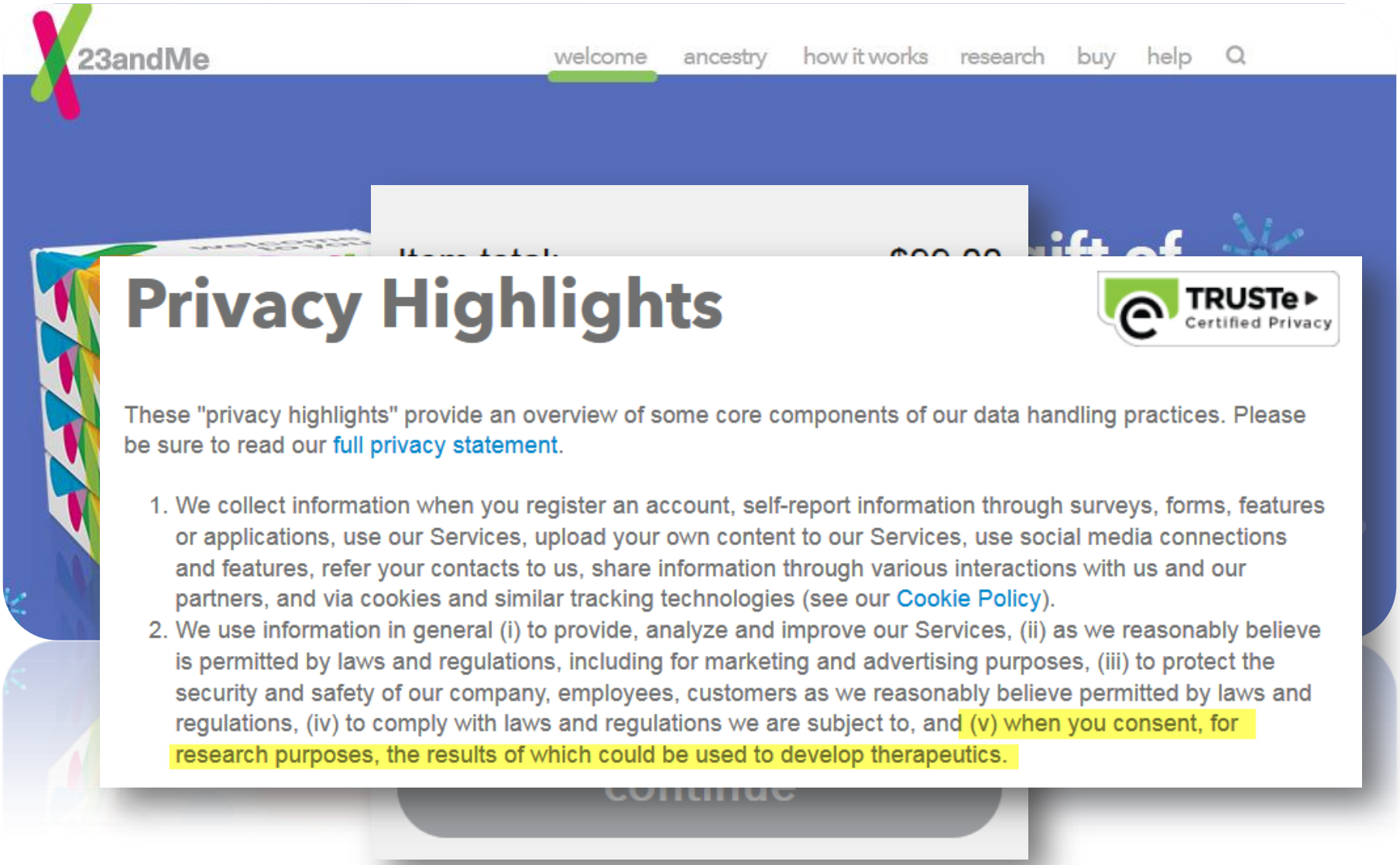
ETHISCHE UND RECHTLICHE ASPEKTE DER TOTALSEQUENZIERUNG DES MENSCHLICHEN GENOMS

EURAT

Position Paper (12.6.2013)

Cornerstones for an ethically and legally informed practice of whole genome sequencing

http://www.uni-heidelberg.de/totalsequenzierung/english.html



Privacy Highlights

TRUSTe Certified Privacy

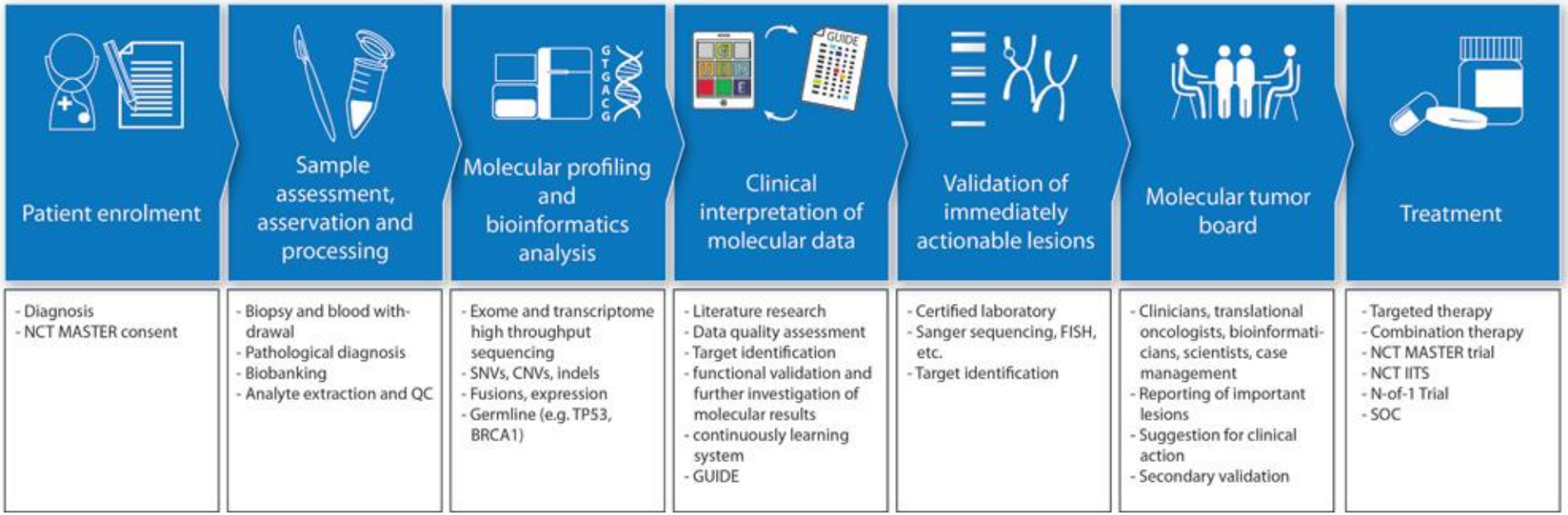
These "privacy highlights" provide an overview of some core components of our data handling practices. Please be sure to read our full privacy statement.

1. We collect information when you register an account, self-report information through surveys, forms, features or applications, use our Services, upload your own content to our Services, use social media connections and features, refer your contacts to us, share information through various interactions with us and our partners, and via cookies and similar tracking technologies (see our Cookie Policy).
2. We use information in general (i) to provide, analyze and improve our Services, (ii) as we reasonably believe is permitted by laws and regulations, including for marketing and advertising purposes, (iii) to protect the security and safety of our company, employees, customers as we reasonably believe permitted by laws and regulations, (iv) to comply with laws and regulations we are subject to, and (v) when you consent, for research purposes, the results of which could be used to develop therapeutics.

Increased Significance for Personalized Medicine

HiGHmed Medical Informatics

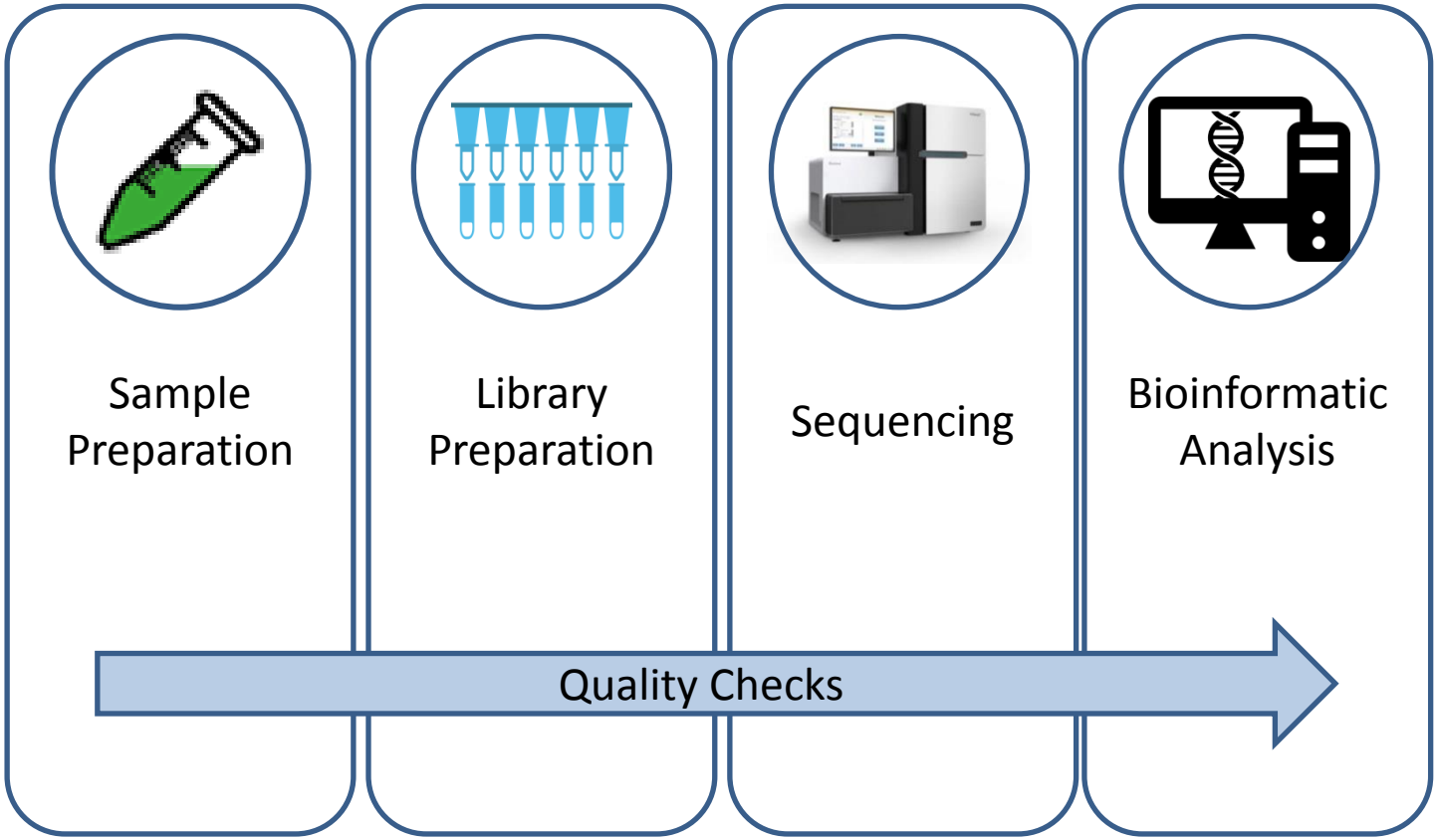
NCT Master Workflow



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Quality of NGS Data

HiGHmed
Medical Informatics



All steps of the NGS workflow impact the data quality

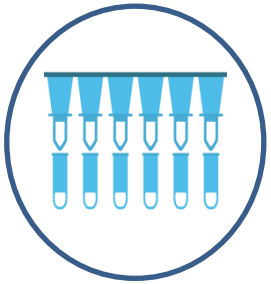
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Impact on Data Quality

HiGHmed
Medical Informatics

e.g.

- Tumor content, purity
- DNA quantity and quality
- Contamination
- Metadata labelling



e.g.

- Library concentration
- Barcode/adaptor errors
- PCR amplification errors

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Impact on Data Quality

HiGHmed
Medical Informatics

e.g.

- Sequencing depth (coverage)
- Length of reads
- Duplication rate

e.g.

- Variant allele frequency
- Filtering (specificity, sensitivity)
- Reference genome used
- Analysis pipelines/versions used
- Including world knowledge into the results

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Quality Assurance & Control

HiGHmed Medical Informatics

- Standard Operation Procedures (SOPs)
- Quality Checks:

Individual	SampleType	QC status	Cov. w/o N	ChrX Cov. w/o N	ChrY Cov. w/o N	Lib Prep Kit	Mapped Reads %	Duplicates %	Properly Paired %	Single %	Insert Size Median	Diff Chrom %
K20K-1YUYYE	BLOOD1	✔	38.10	19.51	14.55	Il TruSeq Nano DNA	99.97	10.96	92.72	0.04	394.00	10.64
K20K-1YUYYE	PATIENT-DERIVED-CULTURE1	✔	41.27	47.51	1.09	Il TruSeq Nano DNA	100.00	10.60	95.55	0.00	395.00	5.73
K20K-27ANNP	BLOOD1	✔	41.65	21.49	17.58	Il TruSeq Nano DNA	100.00	9.18	96.88	0.01	347.00	3.75
K20K-27ANNP	PATIENT-DERIVED-CULTURE1	✔	42.70	42.99	0.78	Il TruSeq Nano DNA	100.00	13.23	95.91	0.00	351.00	5.21
K20K-27ANNP	PATIENT-DERIVED-CULTURE2	✔	41.12	42.16	0.94	Il TruSeq Nano DNA	100.00	14.19	96.65	0.00	324.00	4.21
K20K-27ANNP	PATIENT-DERIVED-CULTURE3	✔	39.44	40.93	0.91	Il TruSeq Nano DNA	99.99	9.09	95.73	0.02	330.00	4.87
K20K-27ANNP	PATIENT-DERIVED-CULTURE4	✔	39.82	39.85	0.88	Il TruSeq Nano DNA	99.99	11.27	96.62	0.02	324.00	4.02
K20K-27ANNP	PATIENT-DERIVED-CULTURE5	✔	39.91	41.24	0.92	Il TruSeq Nano DNA	100.00	10.78	96.22	0.00	319.00	4.63

- Standards to ensure comparability and exchange (fastq, bam, vcf)
- Professional software development in an agile framework (scrum)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Metadata for NGS

HiGHmed Medical Informatics

- Patient ID, Sample, Tissue Type, Cell Type, Disease, library, devices used, reference genome…

→ Essential for interpretability, reproducibility, comparability

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Challenges of Data Quality

HiGHmed Medical Informatics

- Lack of standardization of data generation
- Bioinformatic Workflows not fully standardised
- NGS Workflow not yet clinically “certified“
- **But**: community has long-standing tradition (and willingness!) to successfully address data quality/ standardization issues

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung