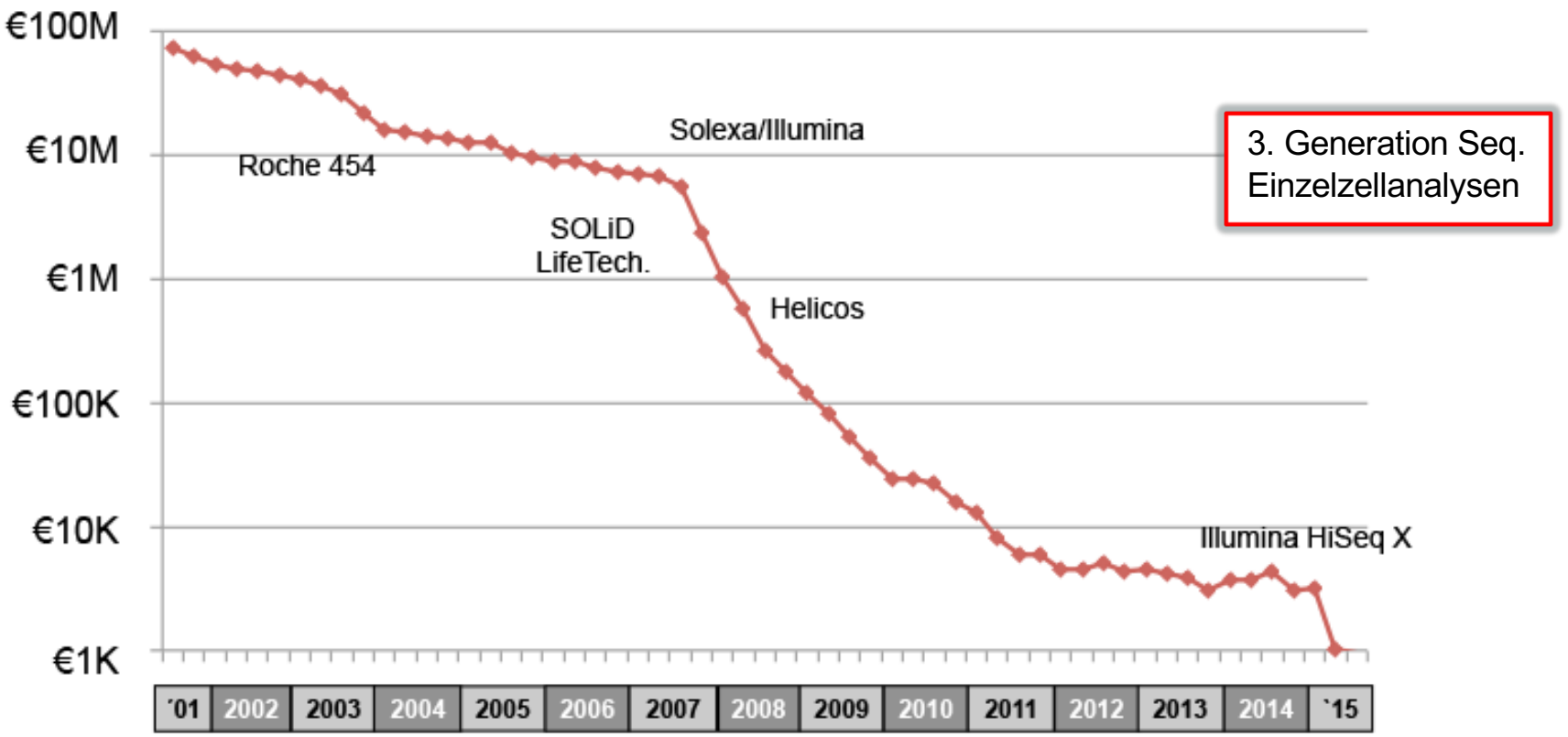
Genomic Sequencing & „Big Data“

Prof. Dr. med.Dr.rer.nat. Michal R. Schweiger
Uniklinik Köln

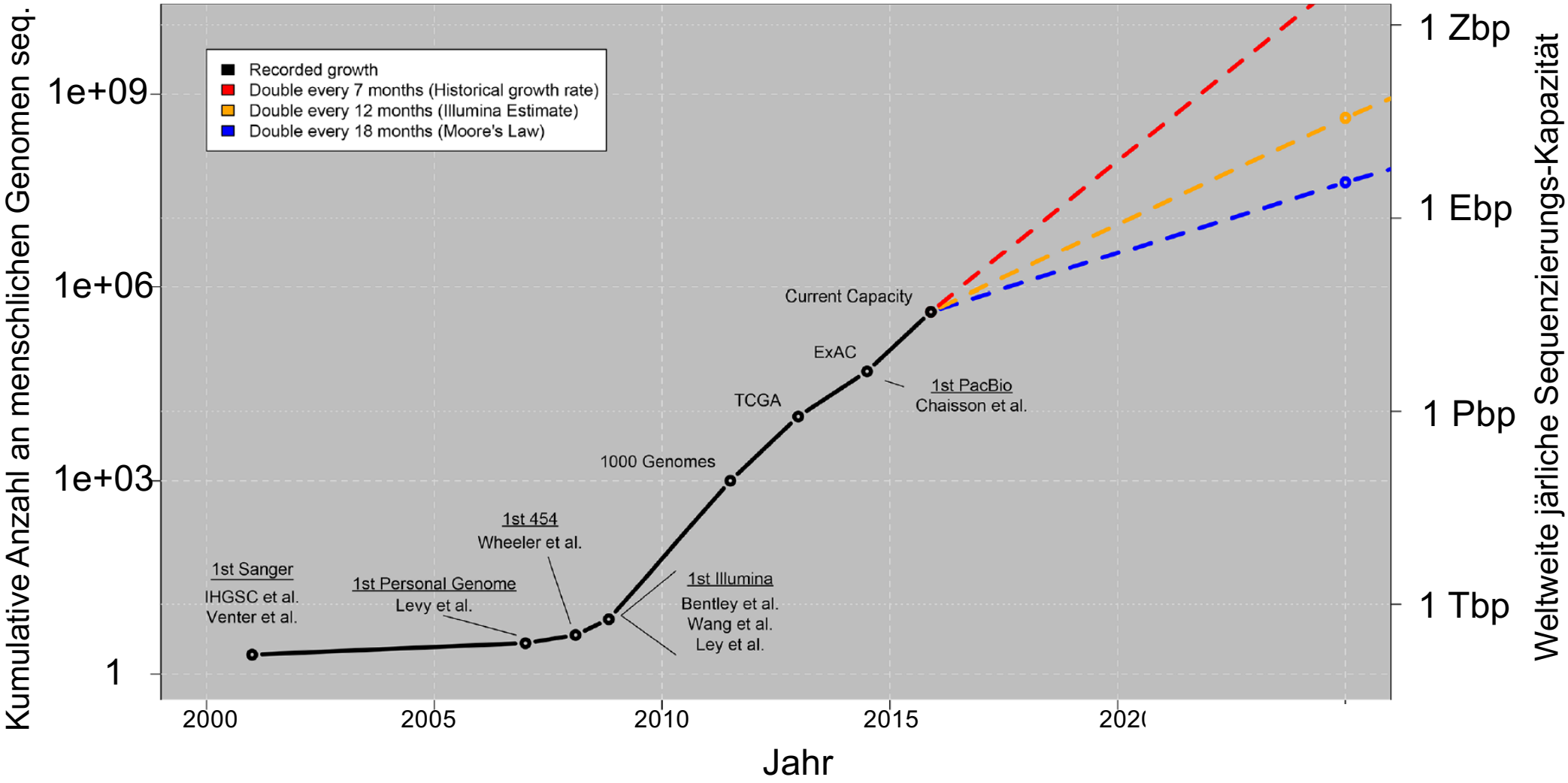
UNIKLINIK KÖLN

Universität zu Köln
Medizinische Fakultät

Entwicklung der Sequenzierungskosten in den letzten Jahren



Sequenzierungsleistung der letzten Jahre



Stephens ZD. PLOS Biology 2015

Phase der Datengenerierung	Astronomie	Twitter	YouTube	Genomik
Gewinnung	25 zetta-bytes/Jahr	0,5 - 15 Milliarden Tweets/Jahr	500 - 900 Mio Stunden / Jahr	1 zetta-basen /Jahr
Speicherung	1 EB / Jahr	1-17 TB / Jahr	1-2 EB / Jahr	2 - 40 EB / Jahr
Analyse	in situ Datenreduktion Echtzeit Prozessierung riesige Volumen	Schlagwortsuche Metadaten Analyse	eingeschränkt	heterogene Daten Variant calling, notwendig mit ca 2 Trillionen central processing units (CPU) Stunden
Verbreitung	spezielle Wege zwischen Antennen und Server (600 TB/s)	kleine Einheiten der Verbreitung	großer Anteil an moderner Nutzung (10 MB/s)	viele kleinere (10MB/s) und einige riesige (10 TB/s) Bewegungen

1 kilobyte	1.000	10^3
1 megabyte	1.000.000	10^6
1 gigabyte	1.000.000.000	10^9
1 terabyte	1.000.000.000.000	10^12
1 petabyte	1.000.000.000.000.000	10^15
1 exabyte	1.000.000.000.000.000.000	10^18
1 zettabyte	1.000.000.000.000.000.000.000	10^21

Stephens ZD. PLOS Biology 2015

Phase der Datengenerierung	Astronomie	Twitter	YouTube	Genomik
Gewinnung	25 zetta-bytes/Jahr	0,5 - 15 Milliarden Tweets/Jahr	500 - 900 Mio Stunden / Jahr	1 zetta-basen /Jahr
Speicherung	1 EB / Jahr	1-17 TB / Jahr	1-2 EB / Jahr	2 - 40 EB / Jahr
Analyse	in situ Datenreduktion Echtzeit Prozessierung riesige Volumen	Schlagwortsuche Metadaten Analyse	eingeschränkt	heterogene Daten Variant calling, notwendig mit ca 2 Trillionen central processing units (CPU) Stunden
Verbreitung	spezielle Wege zwischen Antennen und Server (600 TB/s)	kleine Einheiten der Verbreitung	großer Anteil an moderner Nutzung (10 MB/s)	viele kleinere (10MB/s) und einige riesige (10 TB/s) Bewegungen

Es existieren mehr als 2.500 Hochdurchsatz Instrumente in ca 1.000 Sequenzierungs-Zentren in 55 Ländern

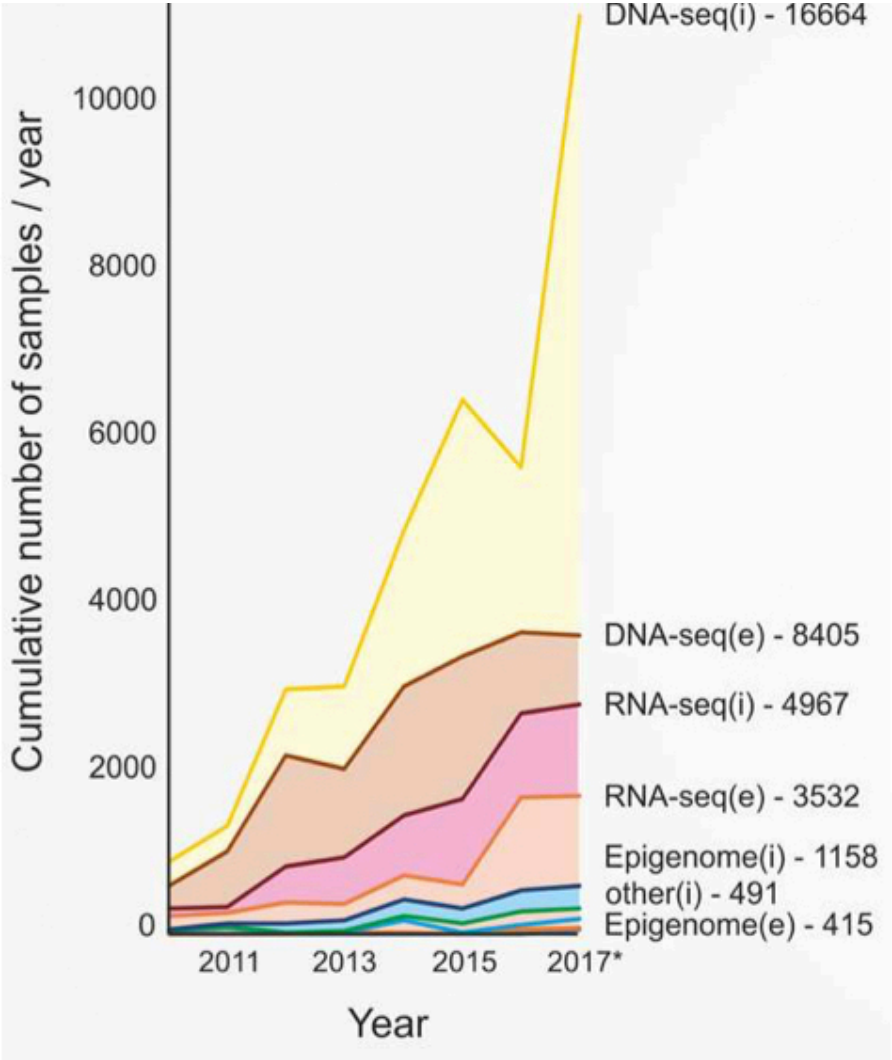
Es wurden bereits mehr als
>32,000 Mikroorganismen
>5,000 Pflanzen
>250,000 menschliche Genome sequenziert

Es wird erwartet, dass mehr als 2,5 Mio Genome von Planzen und Tiere sowie ca. 100 Mio bis 2 Mill. Genome bis zum Jahr 2025 sequenziert sein werden

1 kilobyte	1.000	10^3
1 megabyte	1.000.000	10^6
1 gigabyte	1.000.000.000	10^9
1 terabyte	1.000.000.000.000	10^12
1 petabyte	1.000.000.000.000.000	10^15
1 exabyte	1.000.000.000.000.000.000	10^18
1 zettabyte	1.000.000.000.000.000.000.000	10^21

Stephens ZD. PLOS Biology 2015

NGS Daten Produktion in Köln: ‚Cologne Center for Genomics' und das Regionale Rechenzentrum der Uni Köln



2 Illumina NovaSeq 6000 werden die Sequenzierungs-Leistung auf mehr als 480 Tb / Jahr 2019 erhöhen

Die Anforderungen an die Datenanalysen steigen von 1,5 Mio CPU Stunden / Jahr auf 15 Mio CPU Stunden / Jahr ab 2019

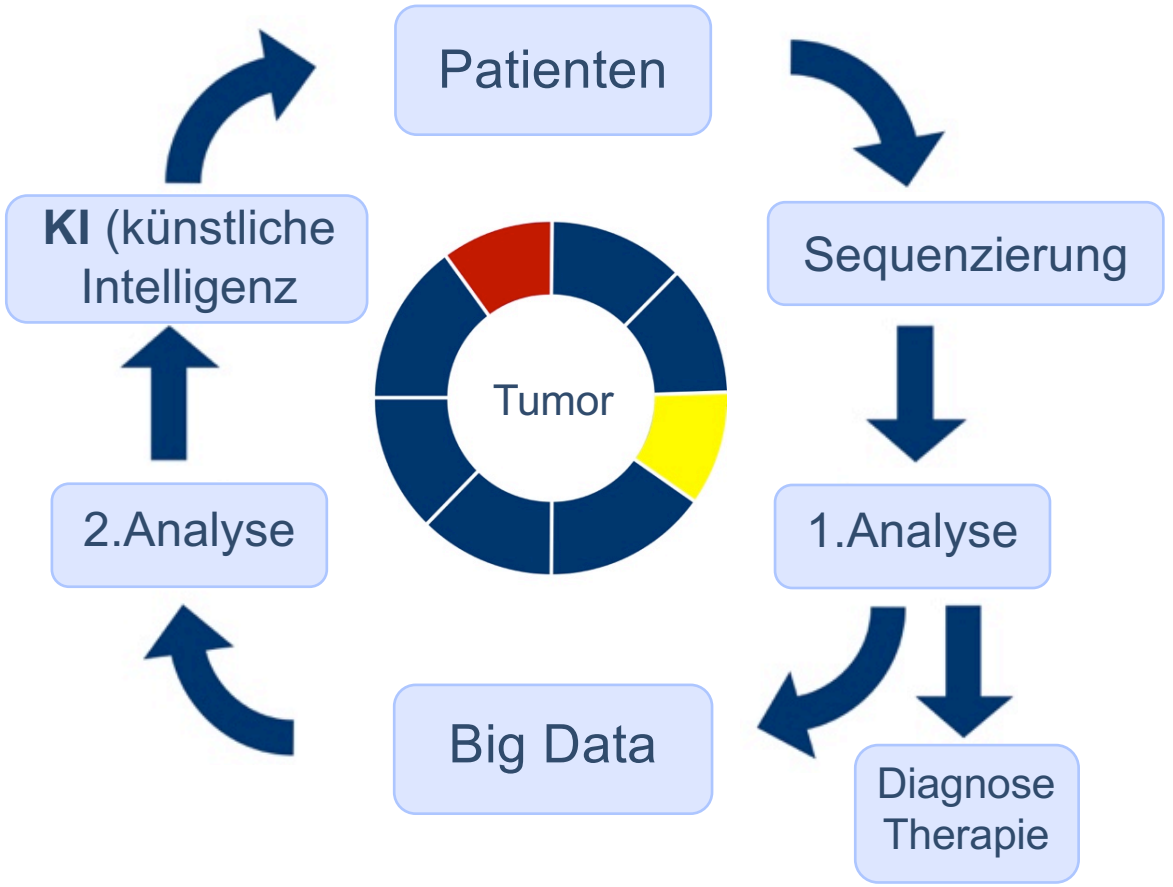
WEST GERMAN GENOME CENTER

DFG NGS-Competenccenter

- **University of Cologne**
- **University Bonn**
- **Heinrich Heine University Düsseldorf**

3 NovaSeq6000, 1 PacBio Sequel
Total Capacity: >750 Tb / year

S.Motameny, V.Achter, U.Lang


Patienten
Sequenzierung
1.Analyse
Diagnose
Therapie
Big Data
2.Analyse
KI (künstliche Intelligenz
Tumor