



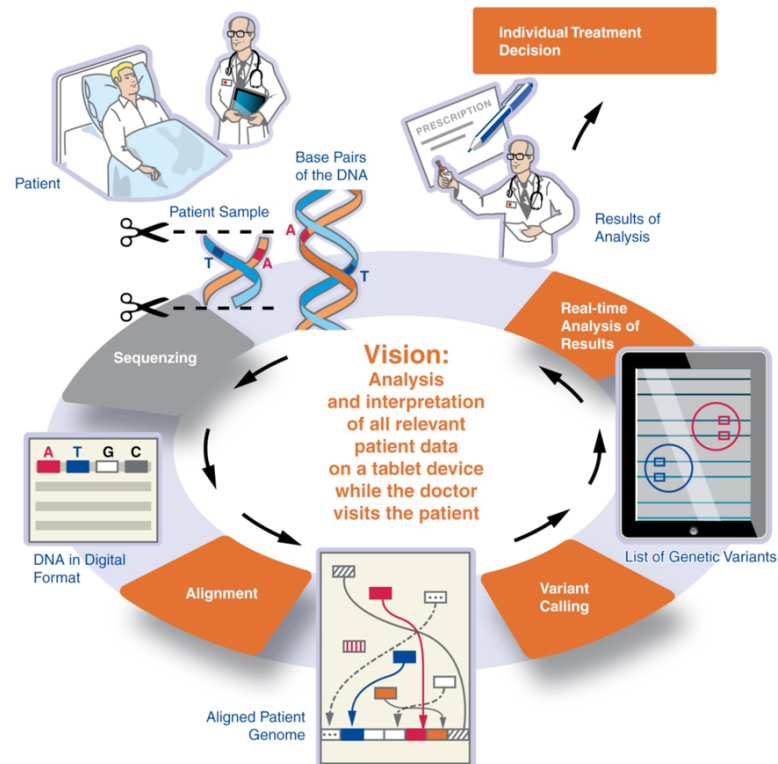
Digital Transformation: From DNA and RNA to Data

Dr. Matthieu-P. Schapranow
Code of Life

November 2016

From Raw Genome Data to Analysis

- **DNA Sequencing:** Transformation of analogues DNA into digital format
- **Alignment:** Reconstruction of complete genome with snippets
- **Variant Calling:** Identification of genetic variants
- **Data Annotation:** Linking genetic variants with research findings

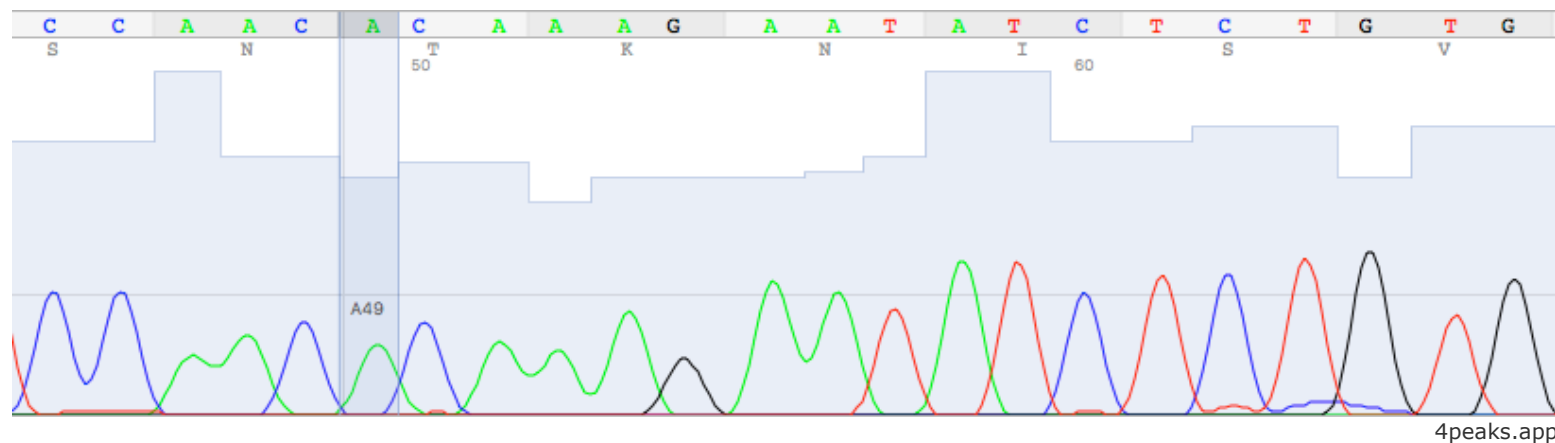


**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

DNA Sequencing

- Purpose: Transformation of analogous DNA into digital format (A/D converter)
- Input: Chunks of DNA
- Output: DNA reads in digital form, e.g. in FASTQ format

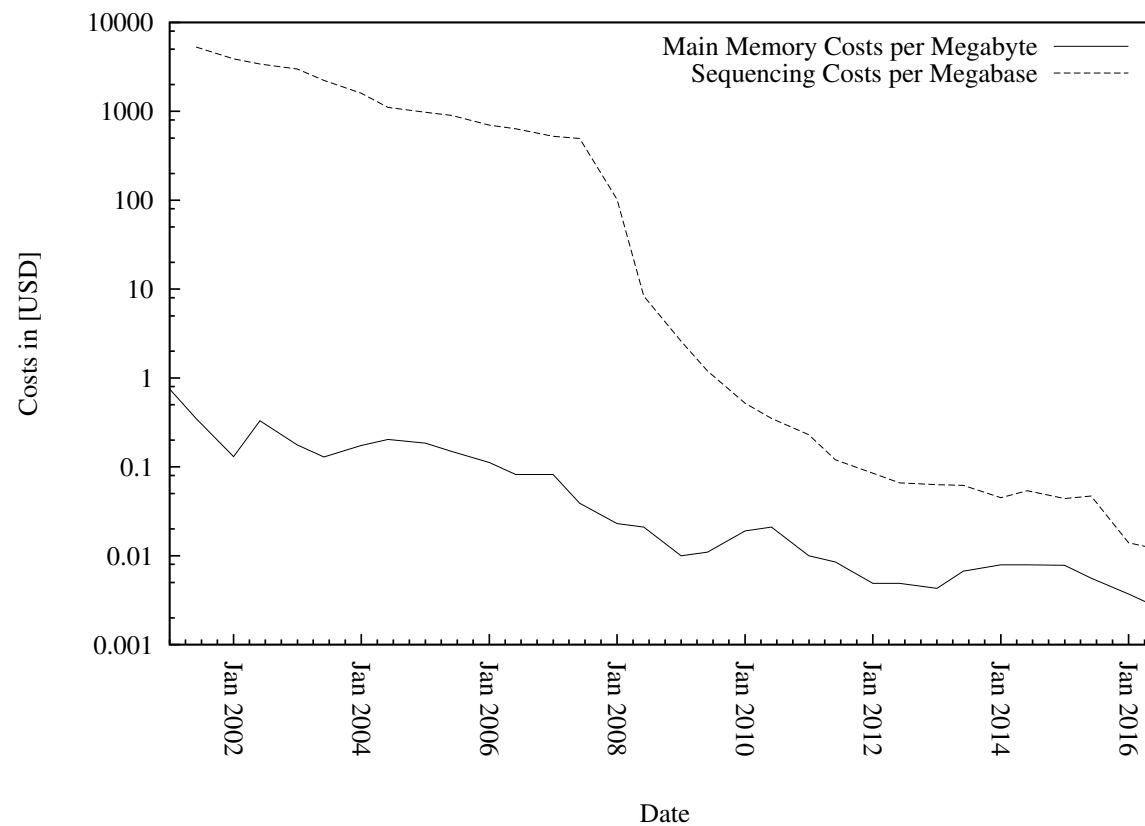


**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

Numbers You Should Know

Comparison of Costs

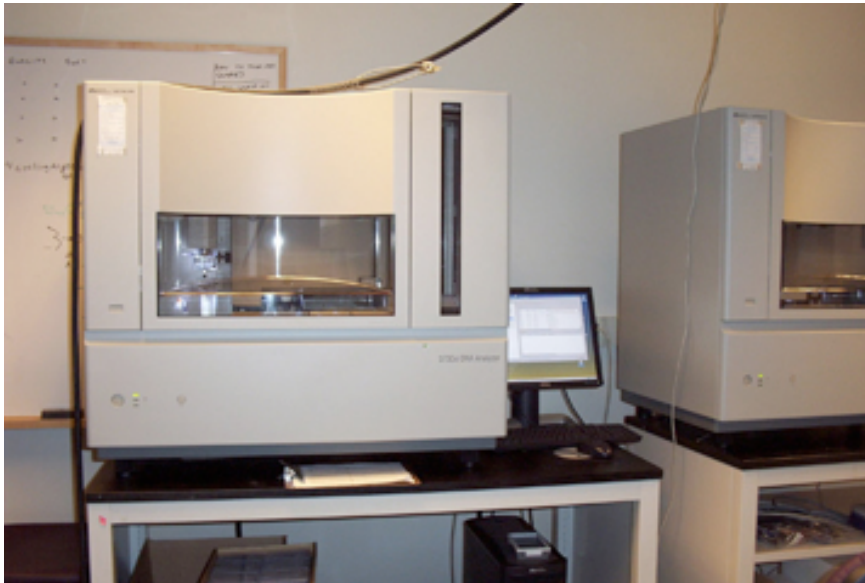


**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

ABI Sequencing (1st gen)

- 2002: Sanger sequencing provides very high accuracy
- Accuracy: > 99.99%
- Throughput: 100 kbp / run (3hrs)
- Read length: 0.6-1 kbp
- Issues: time-intensive



**Digital Transformation:
From DNA and RNA to
Data**

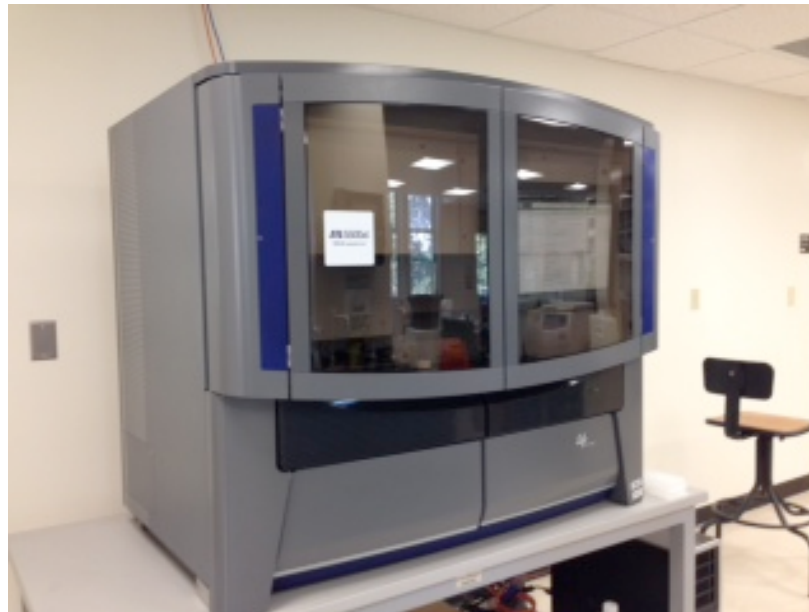
Schapranow, Nov 2016

5

<https://dnaseq.med.harvard.edu/aboutus.html>

ABI Sequencing (2nd gen)

- 2006: Sequencing by Oligonucleotide Ligation and Detection (SOLiD)
- Accuracy: > 99.99%
- Throughput: 60 Gbp / run (5-10 days)
- Read length: 35-100 bp
- Issues: time-intensive



**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

6

<http://cgi.uconn.edu/applied-biosystems-solid-5500xl/>

Roche-454 Sequencing

- 2005-2013: Roche-454 Life Sciences launched first NGS device using pyrosequencing / sequencing by synthesis approach
- Accuracy: >99.9%
- Throughput: 400-600 Mbp / run
- Read length: 200-400 bp (2009) later up to 700 bp
- Issues: Homopolymer repeat regions



**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

7

<http://454.com/products/gs-flx-system/>

Illumina Sequencing

- 2006: Solexa introduced **Genome Analyzer**
- 2007: Illumina acquired Solexa
- Accuracy: >99.9%
- Throughput:
 - 2006: 1 Gbp / run (2006),
 - 2016: up to 1 Tbp / run (6 days)
- Read length: 200-600 bp
- Issues: cheap but less accurate



**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

Oxford Nanopore

- Vision: Very cheap and mobile long-read alternative
- Accuracy: up to 99%
- Throughput: approx. 10 Gbp / run (<48hrs)
- Read length: 230-300 kbp
- Issues: still early phase and behind expectations



**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

<https://nanoporetech.com/products>

Pacific Biosciences

- 2013: PacBio introduces long-read sequencer supporting innovative sequence assembling
- Accuracy: >99% (at high coverage)
- Throughput: 0.5-1 Gbp/run
- Read length: up to 60 kbp (→ DeNovo Alignment)
- Issues: still comparable slow and lacks precision



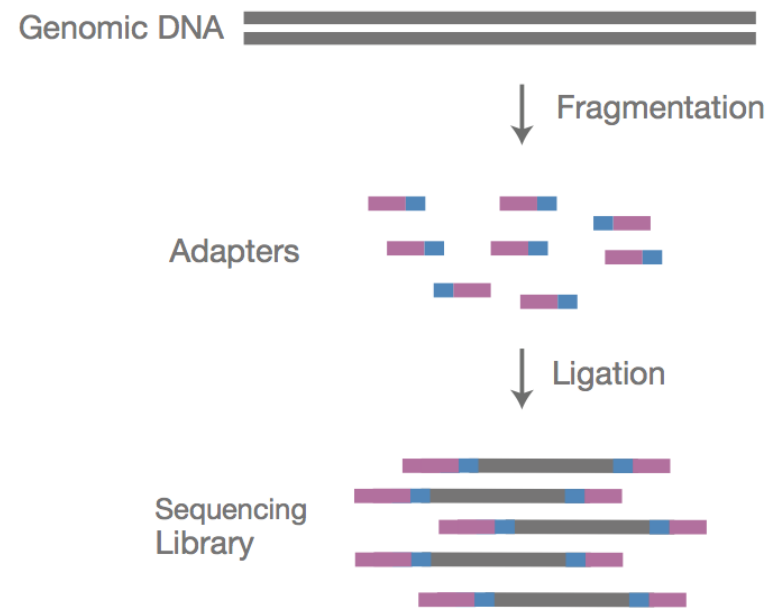
**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

Illumina Sequencing Process

1. Preparation

- Double-stranded DNA is split into chunks of 200-800 bp
- Adapters are ligated to chunks



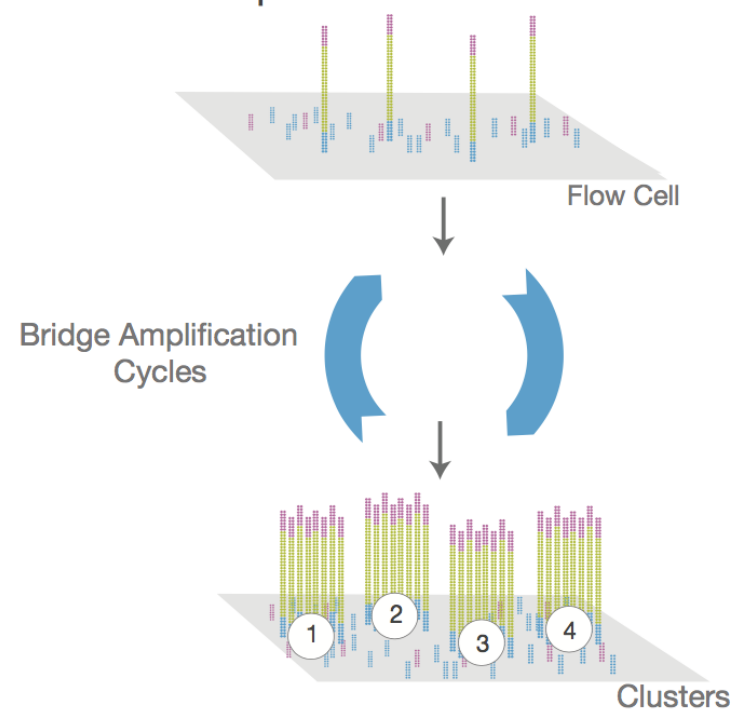
**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

Illumina Sequencing Process

2. Amplification

- Polymerase Chain Reaction (PCR) is used for amplification of DNA chunks



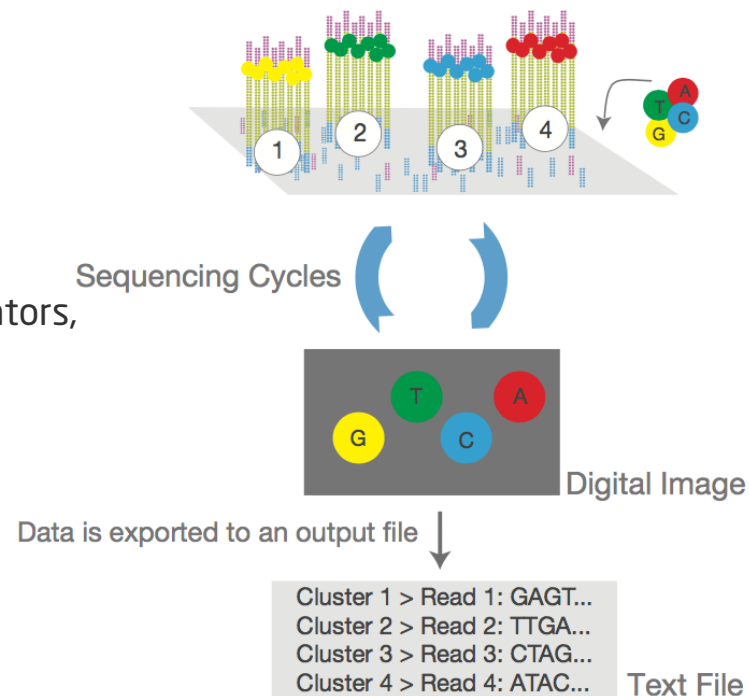
**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

Illumina Sequencing Process

3. Sequencing

- pos = 0
- While (pos < read length) do
 - pos++
 - Wash-off terminators
 - Add primers with fluorescently terminators, i.e. A, C, G, T + stop codon
 - Record laser light reflection image
 - Process image to write textual output
- Done



**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

Illumina Sequencing Process

- Double-stranded DNA is split into chunks of 200-600 bp length
- Adaptors attached to DNA chunks
- Separation of double-strand into two strands using sodium hydroxide
- DNA chunks are washed across flowcell, i.e. DNA not binding to primers is removed
- Polymerase Chain Reaction (PCR) is used for amplification of DNA chunks
- Nucleotide bases and DNA polymerase are added to build bridges b/w primers
- Double strand is split-up using heat → dense clusters of identical DNA sequences
- Primers with fluorescently terminators are added, e.g. A, C, G, T + stop codon
- Primers attach to DNA chunks and DNA polymerase attaches to terminator
- Laser passes flowcell, i.e. each terminator type emits unique light
- Terminators are removed and new terminators are added to next DNA position

Schapranow, Nov 2016

What to take Home?

- Sample preparation results in chunks of DNA
- DNA sequencing is highly automated and results in FASTQ file
- Throughput increased over the past decade

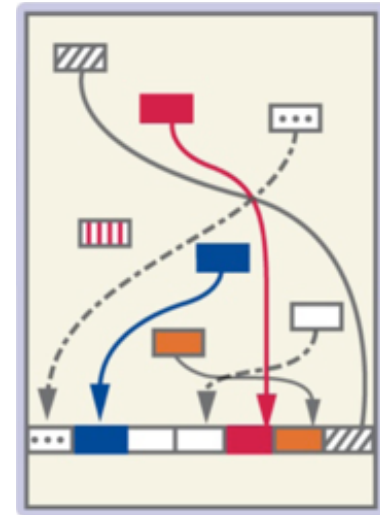
Year	Method	Read Length	Accuracy	Throughput (per day)
2002	Sanger ABI 3730xl	Up to 1 kbp	>99.999 %	400 kbp
2008	Roche 454 GS FLX+	700 bp	>99.9 %	700 Mbp
2012	Illumina 2500 (throughput mode)	2x125 kbp (paired)	>99.9 %	800 Gbp (paired)
2013	Pac Bio RS II	Up to 15 kbp	>90% / >99 % (multi-pass)	6.75 Tbp
2014	Oxford Nanopore MinION	Up to 5 kbp	Up to 99% (multi-pass)	115 Mbp

**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

What is Next?

- How to use DNA in digital format?
- What kind of processing is required prior to interpretation?



**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016

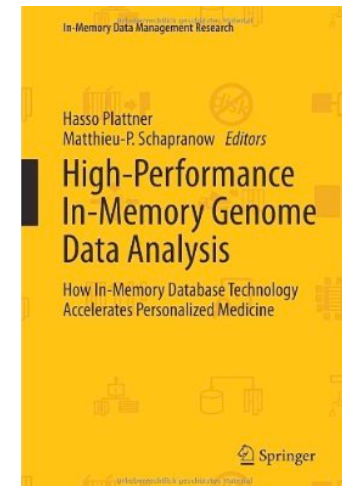
Keep in contact with us!



Dr. Matthieu-P. Schapranow
Program Manager E-Health & Life Sciences

Hasso Plattner Institute
August-Bebel-Str. 88
14482 Potsdam, Germany

schapranow@hpi.de
<http://we.analyzegenomes.com/>



**Digital Transformation:
From DNA and RNA to
Data**

Schapranow, Nov 2016