



# Data Acquisition

Dr. Matthieu-P. Schapranow  
Data Management for Digital Health  
Summer 2017

# Agenda

Real-world  
Use Cases

Oncology



Nephrology



Heart  
Insufficiency



Additional  
Topics



Data Management  
& Foundations



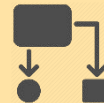
Biology  
Recap



Data  
Sources



Data  
Formats



Business  
Processes



Processing  
and Analysis

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
2

## Biology Recap: Class of 2017.



Biology  
Recap



### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

3



# Agenda

Real-world  
Use Cases

Oncology



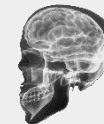
Nephrology



Heart  
Insufficiency



Additional  
Topics



Data Management  
& Foundations



Biology  
Recap



Data  
Sources



Data  
Formats



Business  
Processes



Processing  
and Analysis

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017

---

# Agenda

---

- Examples for data acquisition
  - Sequencing technologies
  - Longitudinal data
  - Sensor data
  - Text documents
- Data processing examples

## **Data Acquisition**

Data Management for  
Digital Health, Summer  
2017

---

## Our Motivation

### Turn Precision Medicine Into Clinical Routine

---



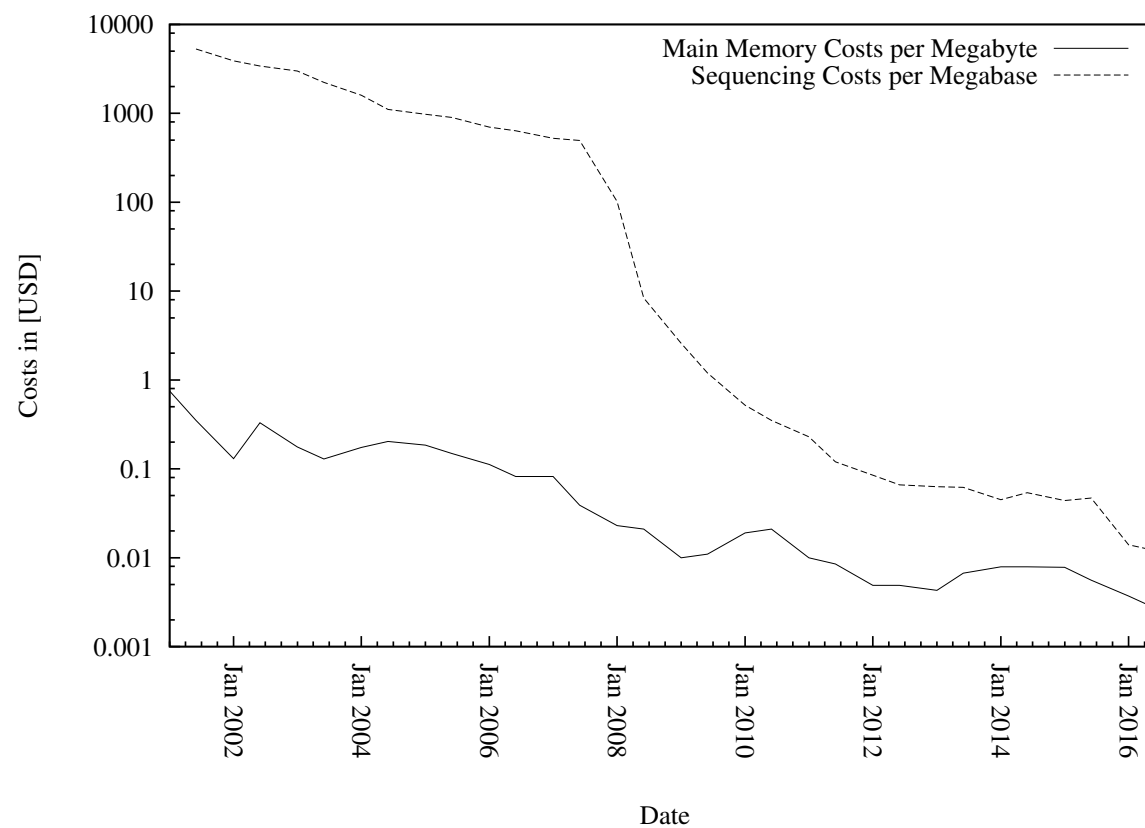
- Can we enable clinicians to take their therapy decisions:
  - Incorporating all available patient specifics,
  - Referencing latest lab results and worldwide medical knowledge, and
  - In an interactive manner during their ward round?

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

# Numbers You Should Know

## Comparison of Costs



### Data Acquisition

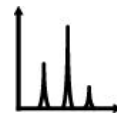
Data Management for  
Digital Health, Summer  
2017

# IT Challenges

## Distributed Heterogeneous Data Sources



**Human genome/biological data**  
600GB per full genome  
15PB+ in databases of leading institutes



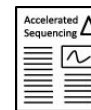
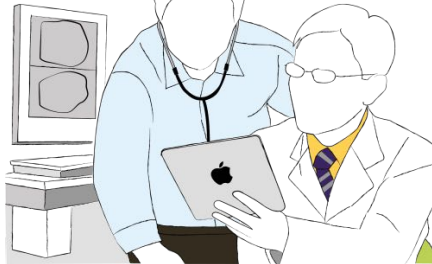
**Human proteome**  
160M data points (2.4GB) per sample  
>3TB raw proteome data in ProteomicsDB



**Hospital information systems**  
Often more than 50GB



**Cancer patient records**  
>160k records at NCT



**PubMed database**  
>23M articles



**Medical sensor data**  
Scan of a single organ in 1s  
creates 10GB of raw data



**Prescription data**  
1.5B records from 10,000 doctors and  
10M Patients (100 GB)



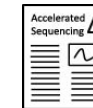
**Clinical trials**  
Currently more than 30k  
recruiting on ClinicalTrials.gov

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
8

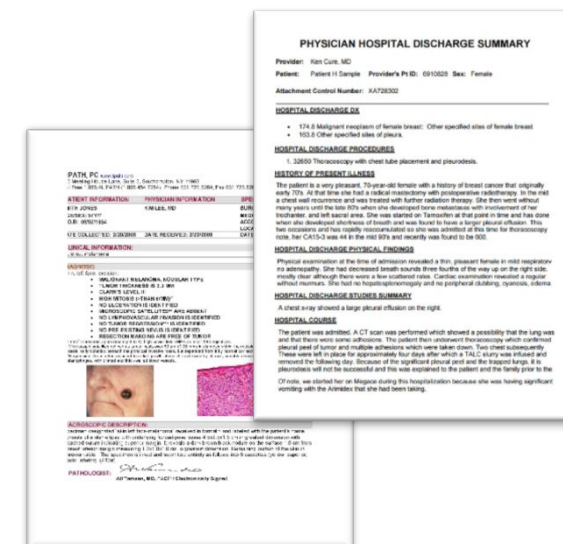


# Text Documents



Hasso  
Plattner  
Institut

- Unstructured data, i.e. not directly machine-readable / -processable
- Examples
  - Discharge / doctor letters
  - Pathology or radiology reports
  - Medical literature (PubMed, The Lancet, etc)



## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
9

# Discharge Letter



Hasso  
Plattner  
Institut

## PHYSICIAN HOSPITAL DISCHARGE SUMMARY

**Provider:** Ken Cure, MD

**Patient:** Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

**Attachment Control Number:** XA728302

### HOSPITAL DISCHARGE DX

- 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- 163.8 Other specified sites of pleura.

### HOSPITAL DISCHARGE PROCEDURES

1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

### HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that originally early 70's. At that time she had a radical mastectomy with postoperative radiotherapy. In the mid a chest wall recurrence and was treated with further radiation therapy. She then went without many years until the late 80's when she developed bone metastases with involvement of her trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done when she developed shortness of breath and was found to have a larger pleural effusion. This two occasions and has rapidly reaccumulated so she was admitted at this time for thoracoscopy note, her CA15-3 was 44 in the mid 90's and recently was found to be 600.

### HOSPITAL DISCHARGE PHYSICAL FINDINGS

Physical examination at the time of admission revealed a thin, pleasant female in mild respiratory no adenopathy. She had decreased breath sounds three fourths of the way up on the right side. mostly clear although there were a few scattered rales. Cardiac examination revealed a regular without murmurs. She had no hepatosplenomegaly and no peripheral clubbing, cyanosis, edema.

### HOSPITAL DISCHARGE STUDIES SUMMARY

A chest x-ray showed a large pleural effusion on the right.

### HOSPITAL COURSE

The patient was admitted. A CT scan was performed which showed a possibility that the lung was and that there were some adhesions. The patient then underwent thoracoscopy which confirmed pleural peel of tumor and multiple adhesions which were taken down. Two chest subsequently

## SAMPLE DISCHARGE SUMMARY

**Primary Diagnosis:** 40 week IUP with delivery of a liveborn infant

**Secondary Diagnosis:** Advanced Maternal Age; Prolonged second stage of labor with maternal exhaustion

### **Procedure Performed:**

1. Spontaneous Vaginal Delivery with delivery of live male infant weighing 7# 5oz at 1542 on January 3, 2012 with APGARS of 8 at one minute and 9 at five minutes.
2. Placement of Intrauterine Pressure Catheter.

**Reason for Hospitalization:** This 36yo G2P1001 presented at 40 weeks gestation by an LMP of 3/12/11 with an EDC of 1/3/12 in spontaneous labor. This pregnancy has been complicated by advanced maternal age. QS performed at 17 weeks was within normal limits and a genetic amniocentesis was offered and declined. Prenatal laboratory data showed blood type B+ with a negative antibody screen, Rubella Immune, VDRL nonreactive, HepBsAg negative, Diabetic Screen 120, HIV nonreactive. She remained normotensive throughout her pregnancy. At the time of admission she reported positive fetal movement and denied loss of fluid.

**Physical Exam on Admission:** Temperature 98.4. Pulse 94. Respirations 16. Blood pressure 128/78. Fetal Heart Rate 150's and reactive. Uterine contractions q 4 minutes. HEENT within normal limits. Heart regular. Lungs clear. Abdomen gravid with a fundal height appropriate for gestational age. Extremities 2+ DTR's and trace edema. Cervical exam 4 cm/80%/-1.

**Lab and X-Ray Data:** Predelivery H&H of 12.4 and 36.2 respectively. Platelets 221.

**Hospital Course:** The patient was admitted in spontaneous labor in the morning of January 3rd. was reactive and reassuring throughout the course of her stay in labor and delivery. Her labor progressed well and at 0900 hours, she had spontaneous rupture of membranes with a return of fluid. At that time, her cervix was dilated to 6 cm/90%/0. Epidural anesthesia was requested and obtained. Her labor then quickly progressed and the patient was noted to be completely dilated at +1 station at 1100 hours. She was then allowed to push. After pushing for 2 hours, the patient brought the vertex to the perineum, but was unable to continue her expulsive efforts. The infant delivered by outlet forceps over a midline episiotomy. **Please see operative report for full details.** The patient and infant did well. She is breast-feeding the infant well, and has remained afebrile with minimal lochia since delivery. The patient was voiding and ambulating without difficulty by the evening of PPD #0. She declined any contraception at the time of discharge, and was deemed stable for discharge on PPD 2.

# Pathology Report



Hasso  
Plattner  
Institut

2008-15

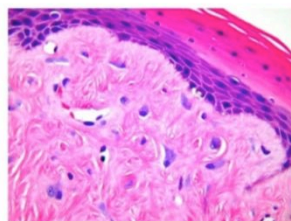
Patient [REDACTED]  
Date of birth [REDACTED] Sex Male  
Biopsy Date 1/3/2008  
Doctor [REDACTED]



## Part A: LEFT MAXILLARY SOFT TISSUE

### Gross description:

Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.



### Microscopic Description:

Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.

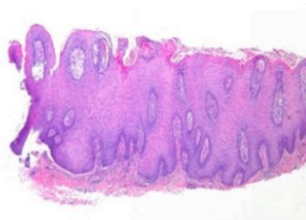
Diagnosis: **Fibroma, giant cell type**

ICD: 210.4  
CPT: 88305

## Part B: RIGHT LATERAL TONGUE

### Gross description:

Submitted is formalin fixed tissue, measuring 1.2x0.5x0.5cm., stated to be from the right lateral tongue. The specimen consists of one piece of tan soft tissue with suture. One section submitted.



### Microscopic Description:

Multiple sections show acanthotic, parakeratotic, verrucous stratified squamous epithelium covering a core of well-vascularized fibrous connective tissue. The interepithelial connective tissue papilla are filled with foamy histiocytes. Lymphocytes and plasma cells are also seen.

Diagnosis: **Verruciform xanthoma**

ICD: 210.4  
CPT: 88305

## DERMATOPATHOLOGY PATHOLOGY REPORT



PATIENT INFORMATION	PHYSICIAN INFORMATION	SPECIMEN INFORMATION
[REDACTED]	[REDACTED]	SURGICAL #: S08-02011 MEDICAL REC #: 0315961 ACCOUNT #: 409514 LOCATION: ASC
DATE COLLECTED: 2/20/2008	DATE RECEIVED: 2/20/2008	DATE REPORTED: 2/21/2008

### CLINICAL INFORMATION:

Rule out melanoma.

### DIAGNOSIS

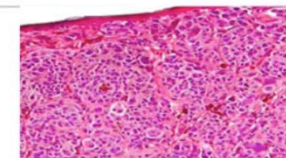
Skin, left face, excision:

- MALIGNANT MELANOMA, NODULAR TYPE
- TUMOR THICKNESS IS 3.0 MM
- CLARK'S LEVEL III
- HIGH MITOSIS (>THAN 6/MM)<sup>\*\*</sup>
- NO ULCERATION IS IDENTIFIED
- MICROSCOPIC SATELLITES<sup>\*\*</sup> ARE ABSENT
- NO LYMPHOVASCULAR INVASION IS IDENTIFIED
- NO TUMOR REGRESSION<sup>\*\*\*</sup> IS IDENTIFIED
- NO PRE-EXISTING NEVUS IS IDENTIFIED
- RESECTION MARGINS ARE FREE OF TUMOR

\* 1 mm<sup>2</sup> represents approximately 9 to 10 high power field (HPF) in most X40 objectives.

\*\* Microscopic satellites defined as tumor nests over 50 µm (0.05 mm) in diameter within the reticular dermis, fat tissue, blood vessels, or lymphatics beneath the principal invasive mass, but separated from it by normal connective tissue in serial sections.

\*\*\*Regression: Areas often adjacent to radial growth phase characterized by fibrosis, variable dense infiltrate of lymphocytes and melanophages, with dilated and thick-walled blood vessels.



### MACROSCOPIC DESCRIPTION:

Specimen designated "skin left face-melanoma" received in formalin and labeled with the patient's name consists of a skin ellipse with underlying fibroadipose tissue 4.5x2.5x1.5 cm in greatest dimension with attached suture indicating superior margin. It reveals a dark brown-black nodule on the surface 1.0 cm from

# Prescriptions German Muster 16



Krankenkasse bzw. Kostenträger		Hilfs- mittel 6		Impf- stoff 7	Sex - Stoff 8	Stoff- änderung 9	Reiz- Phosphat	Apotheken-Nummer / St.
<input checked="" type="checkbox"/> AOK Rheinland-Pfalz		Ausstellung		Gesamt-Brutto				
Name, Vorname des Versicherten		Geb. am						
Mustermann Erika		12.08.1964						
Heidestraße 17 51147 Köln		10/14						
Kassen-Nr.	Versicherten-Nr.	Status		Arzneimittel-Mittelwert-Id.				
106415300	A123456789	1000 1		1. Verordnung				
Betriebsstätten-Nr.	Arzt-Nr.	Datum		2. Verordnung				
271111100	654321161	10.07.2012		3. Verordnung				

**Rp.** (Bitte Leeräume durchstreichen)

☒ Antistressin Impfstoff Amp. 10 x 0.5 ml  
Muster Pharma GmbH  
\*\*\*\*\*  
\*\*\*\*\*

☐ auf ident  
☐ auf ident  
☐ auf ident

**bbbrl**       Abgabedatum in der Apotheke

Bei Arbeitsunfall auszufüllen: Unfalltag  Unfallbetrieb oder Arbeitgebernummer

271111100  
Psychologische Gemeinschaftspraxis  
**Dr. med. Markus Mustermann**  
**Dr. rer. nat. Erik Mustermann**  
Dortheidenstraße 1  
**51069 Köln**  
Tel. 02 21 16 87 65 43  
  
Unterschrift des Arztes  
(Muster 16 (7.2008))

**2711111004**

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
12

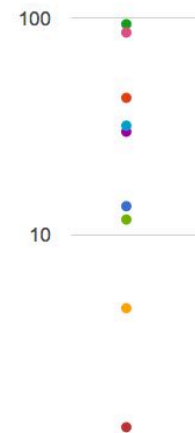
---

## Categories of Data

### Data Points

---

- Data points := Acquired once or multiple times in (non-)equidistant times
- Provides a single point in time impression
- Examples: Lab results
- Pros: Can provide just-in-time insights
- Cons: Does not provide holistic view



#### Data Acquisition

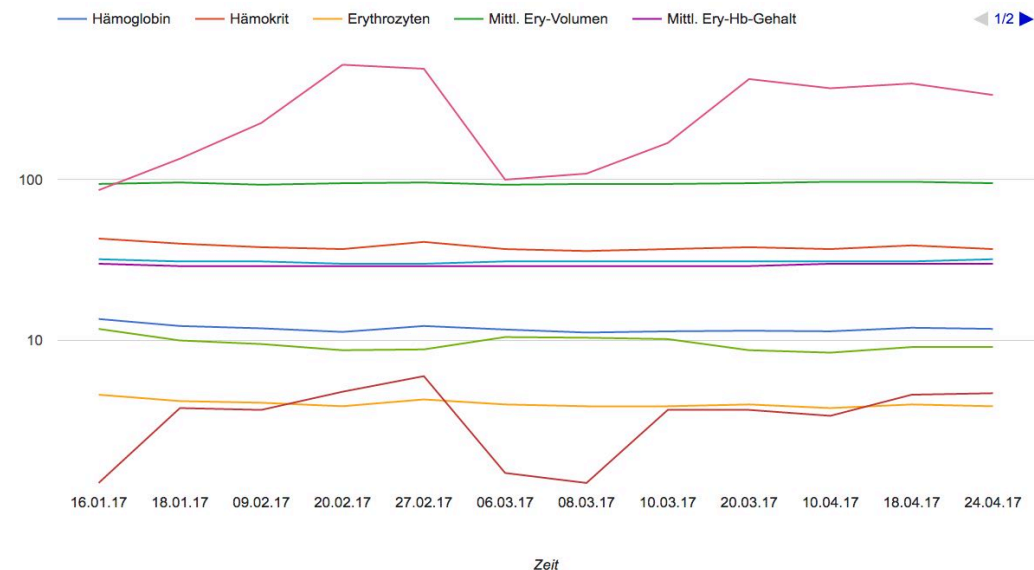
Data Management for  
Digital Health, Summer  
2017



# Categories of Data

## Longitudinal Data

- Longitudinal data := Multiple measurements over (equidistant) time spans
- Examples:
  - Lab values
  - Clinical studies
  - Observational studies
- Pros: Can provide a more holistic view on changes of data over time
- Cons: Requires time to acquire



### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
14



# Laboratory Tests (German Muster 10A)



Krankenkasse bzw. Kostenträger  
AOK Baden-Württemberg 61125

Name, Vorname des Versicherten  
Mustermann  
Max  
Testweg 1  
12345 Berlin 12/18

Kassen-Nr. 8018121  
Versicherten-Nr. 12345678901  
Status 1000 1

Betriebsstätten-Nr. 612345678  
Arzt-Nr. 619999900  
Datum 23.11.15

Anforderungsschein für Laboratoriums-  
untersuchungen bei Laborgemeinschaften

☐ Kurativ ☐ Präventiv ☐ bei belegärztl. Behandlung ☐ Unfall, Unfallfolgen

ggf. Kennziffer

Geschlecht  
W ☐ M ☒

10ABF

Abnahmedatum Abnahmezeit

Diagnosen

☐ Befund eilt 1

Serum Vollblut		Glukose	
<input type="checkbox"/> EDTA	<input type="checkbox"/> alkalische Phosphatase 13	<input type="checkbox"/> Eiweiß gesamt 26	<input type="checkbox"/> Kreatinin Clearance 40
<input type="checkbox"/> großes Blutbild 2	<input type="checkbox"/> Amylase 14	<input type="checkbox"/> Gamma GT 27	<input type="checkbox"/> Glukose 1 51
<input type="checkbox"/> kleines Blutbild 3	<input type="checkbox"/> ASL 15	<input type="checkbox"/> Glukose 28	<input type="checkbox"/> Glukose 2 52
<input type="checkbox"/> HbA1c 4	<input type="checkbox"/> Bilirubin direkt 16	<input type="checkbox"/> GOT 29	<input type="checkbox"/> Glukose 3 53
<input type="checkbox"/> Retikulozyten 5	<input type="checkbox"/> Bilirubin gesamt 17	<input type="checkbox"/> GPT 30	<input type="checkbox"/> Glukose 4 54
<input type="checkbox"/> Blutsenkung 6	<input type="checkbox"/> Calcium 18	<input type="checkbox"/> Harnsäure 31	<b>Urin</b>
<input type="checkbox"/> Diff. Blutbild (Ausstrich) 7	<input type="checkbox"/> Cholesterin 19	<input type="checkbox"/> Harnstoff 32	<input type="checkbox"/> Status 55
<b>Citrat</b>	<input type="checkbox"/> Cholinesterase 20	<input type="checkbox"/> HBDH 33	<input type="checkbox"/> Mikroalbumin 56
<input type="checkbox"/> Quick 8	<input type="checkbox"/> CK 21	<input type="checkbox"/> HDL-Cholesterin 34	<input type="checkbox"/> Schwangerschaftstest 57
<input type="checkbox"/> Quick unter Marcumar-Therapie 9	<input type="checkbox"/> CK-MB 22	<input type="checkbox"/> IgA 35	<input type="checkbox"/> Glukose 58
<input type="checkbox"/> Thrombinzeit 10	<input type="checkbox"/> CRP 23	<input type="checkbox"/> IgG 36	<input type="checkbox"/> Amylase 59
<input type="checkbox"/> PTT 11	<input type="checkbox"/> Eisen 24	<input type="checkbox"/> IgM 37	<input type="checkbox"/> Sediment 60
<input type="checkbox"/> Fibrinogen 12	<input type="checkbox"/> Eiweiß Elektrophorese 25	<input type="checkbox"/> Kalium 38	<input type="checkbox"/> Sonstiges 61
		<input type="checkbox"/> Kreatinin 39	
		<input type="checkbox"/> TSH basal 49	
		<input type="checkbox"/> TSH nach TRH 50	
		<input type="checkbox"/> Phosphat, anorganisches 46	
		<input type="checkbox"/> Lipase 43	
		<input type="checkbox"/> OP-Vorbereitung (32125) 45	
		<input type="checkbox"/> Natrium 44	
		<input type="checkbox"/> Transferrin 47	
		<input type="checkbox"/> Triglyceride 48	

<http://www.wenger.de/de/gesundheitswesen/blankoformularbedruckung/blankoformularbedruckung.php>

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
15

# Laboratory Values



- List of medical attributes and their “normal” thresholds
- Exceeded values are highlighted
- Standardized encoding using Logical Observation Identifiers Names and Codes (LOINC)
- LOINC was initiated 1994 in the U.S.

Untersuchungsparameter	Ergebnis	Ind.	Einheit	Normwerte bzw. therap. Bereich
<b>Blutstatus rot</b>				
Hämoglobin	11.8	-	g/dl	12.3 - 15.3
Hämatokrit	37	-	%	36 - 45
Erythrozyten	3.9	-	/pl	4.1 - 5.1
Mittl.Ery-Volumen	95	-	fl	80 - 96
Mittl.Ery-Hb-Gehalt	30	-	pg	28 - 33
Mittl.Ery-Hb-Konz.	32	-	g/dl	33 - 36
<b>Blutstatus Thrombozyten</b>				
Thrombozyten	336	-	/nl	150 - 400
Mittleres Thrombovol.	9.1	-	fl	7.4 - 11
<b>Blutstatus weiss</b>				
Leukozyten	4.7	-	/nl	4.3 - 10
<i>Kapillarblut : größere Streubreite der Messwerte insbesondere der Leukozyten</i>				
<b>mechanisches Diff.-BB</b>				
Blutstatus mechan. Diff				
Lymphozyten/mech.Diff.abs.	0.7	-	/nl	1.0 - 2.8
Monocyten/mech.Diff.abs.	0.7	-	/nl	0 - 0.8
Seg.Gran./mech.Diff.abs.	2.9	-	/nl	1.4 - 6.5
Basophile/mech.Diff.abs.	0.1	-	/nl	0 - 0.2
Eosinophile/mech.Diff.abs.	0.3	-	/nl	0 - 0.7
Lymphocyten/mech.Diff.%	15	-	%	20 - 55
Monocyten/Mech.Diff.%	14	+	%	2.5 - 10
Seg.Gran./mech.Diff.%	62	-	%	37 - 75
Basophile /mech.Diff.%	1.9	-	%	0 - 2
Eosinophile/mech.Diff.%	7.2	-	%	0.5 - 11

Vorläufiger Befund - noch nicht validiert.

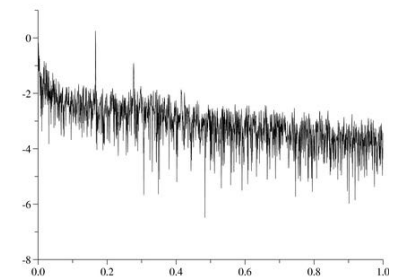
## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
16

# Sensor Data



- Data acquired by medical equipment in equidistant time
- Examples
  - Patient bedside monitoring
  - Electrocardiogram (ECG) monitors, pulse oximetry, blood pressure
  - Wearables, e.g. blood pressure, accelerator



SantaMedical's Finger Pulse Oximeter



Heal Force Portable ECG monitor



Affective Q Sensor

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017

17



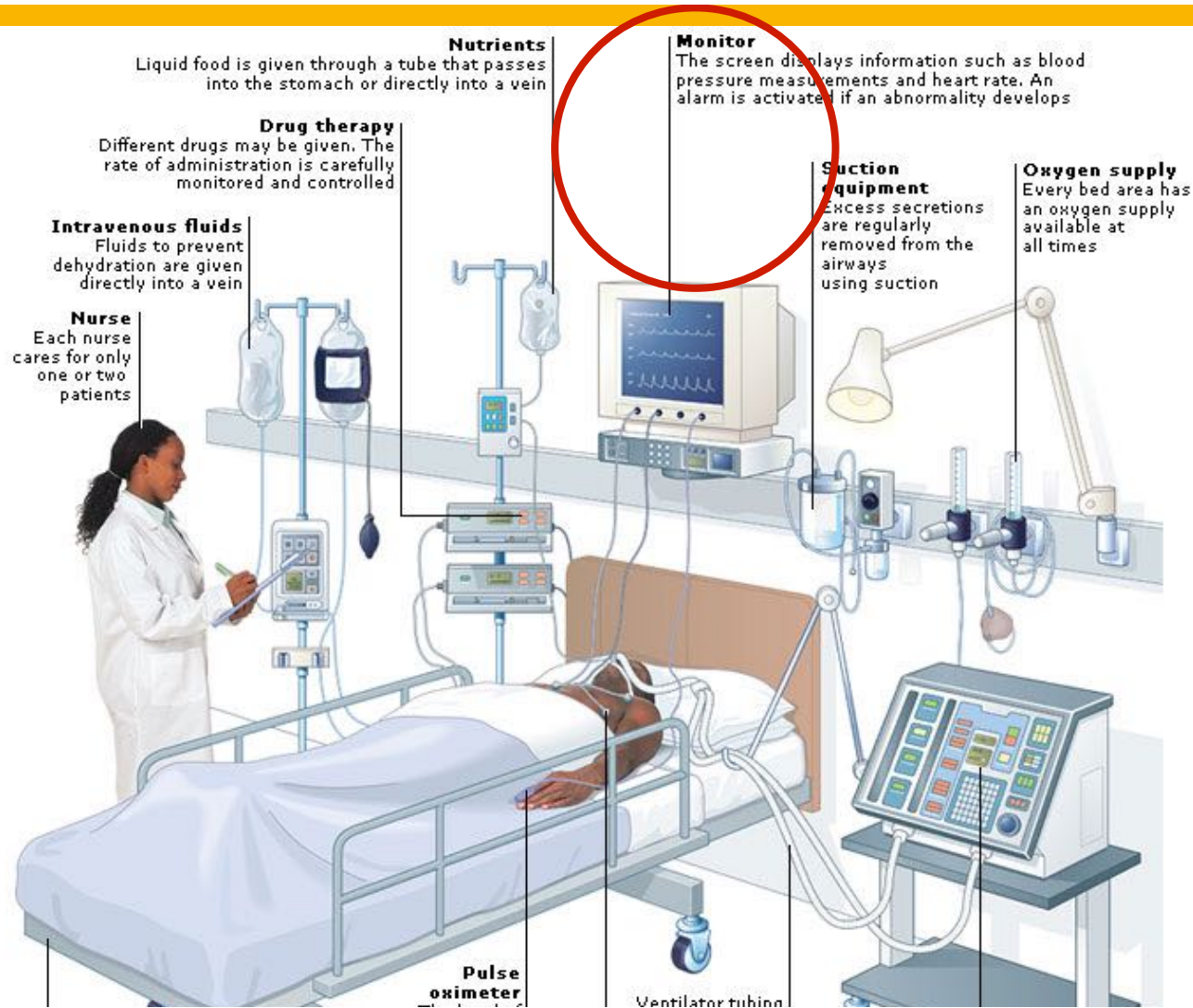
## Sensor Data Intensive Care Unit



Source: Armed Forces Institute of Cardiology & National Institute of Heart Diseases (Pakistan)

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
18



## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
19

---

## Sensor Data Intensive Care Unit: Physiologic Monitors

---



<https://www.copra-system.de/>

### Data Acquisition

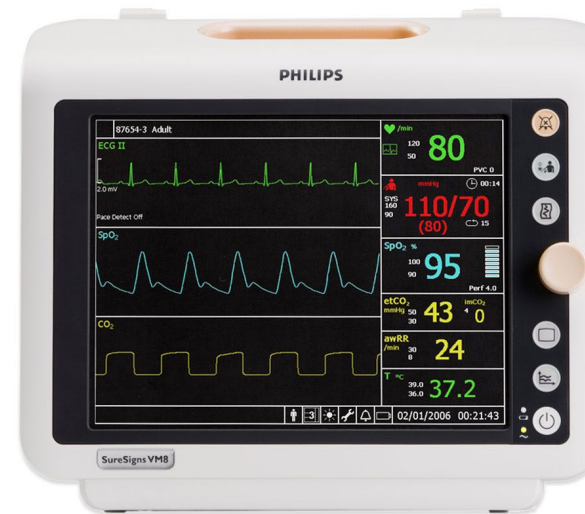
Data Management for  
Digital Health, Summer  
2017  
20



# Sensor Data

## Intensive Care Unit: Physiologic Monitors

- Main functions
  - Monitor vital signs
  - Provide alarms
- Main components
  - Central Station
  - Bedside Monitor
  - Telemetry Transmitter and Receiver
- ICU Patient Scores (monitoring)
  - APACHE (Acute Physiology And Chronic Health Evaluation)
  - SOFA (Sepsis-related organ failure assessment)



SureSigns VM6 Vital Signs Patient Monitor w/ECG

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
21

## Sensor Data

### Intensive Care Unit: Data Volume

- Assumptions
  - Capacity of Intensive Care Unit (ICU) per hospital: avg. 20 patients
  - Sensors per patient: avg. 8 signals
  - Data points per sensor: avg. 125 Hz
  
- Estimated data volume per ICU if all data would be persisted
  - 72M data points per hour
  - 1.7T data points per day

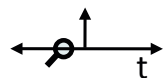
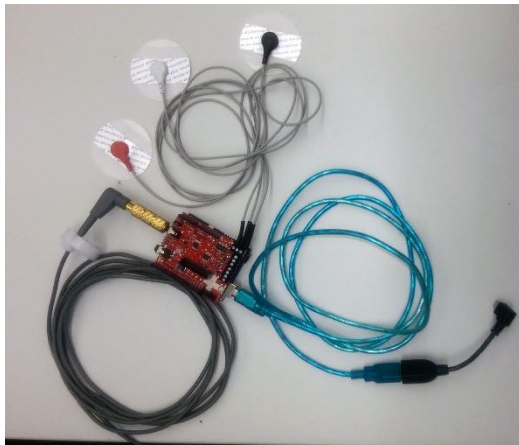


<http://www.physionet.org>

#### Data Acquisition

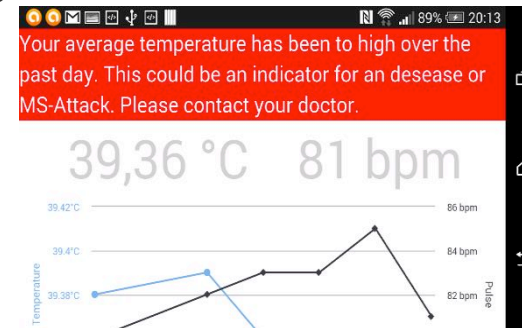
Data Management for  
Digital Health, Summer  
2017  
22

# Real-time Processing of Event Data from Medical Sensors



Comparison of waveform data  
with history of similar patients

- Processing of sensor data, e.g. from Intensive Care Units (ICUs) or wearable sensor devices
  - Multi-modal real-time analysis to detect indicators for severe events
- Incorporates machine-learning algorithms to detect severe events and to inform clinical personnel in time
- Successfully tested with 100 Hz event rate, i.e. sufficient for ICU use



Harvard-MIT  
Health Sciences & Technology

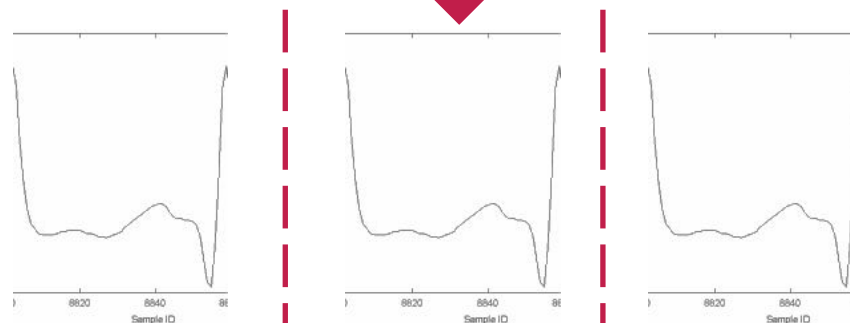
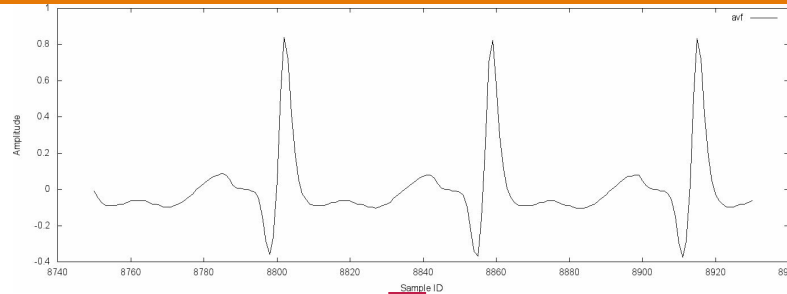


Future SOC Lab

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
23

# Heart Beat Segmentation



Harvard-MIT  
Health Sciences & Technology

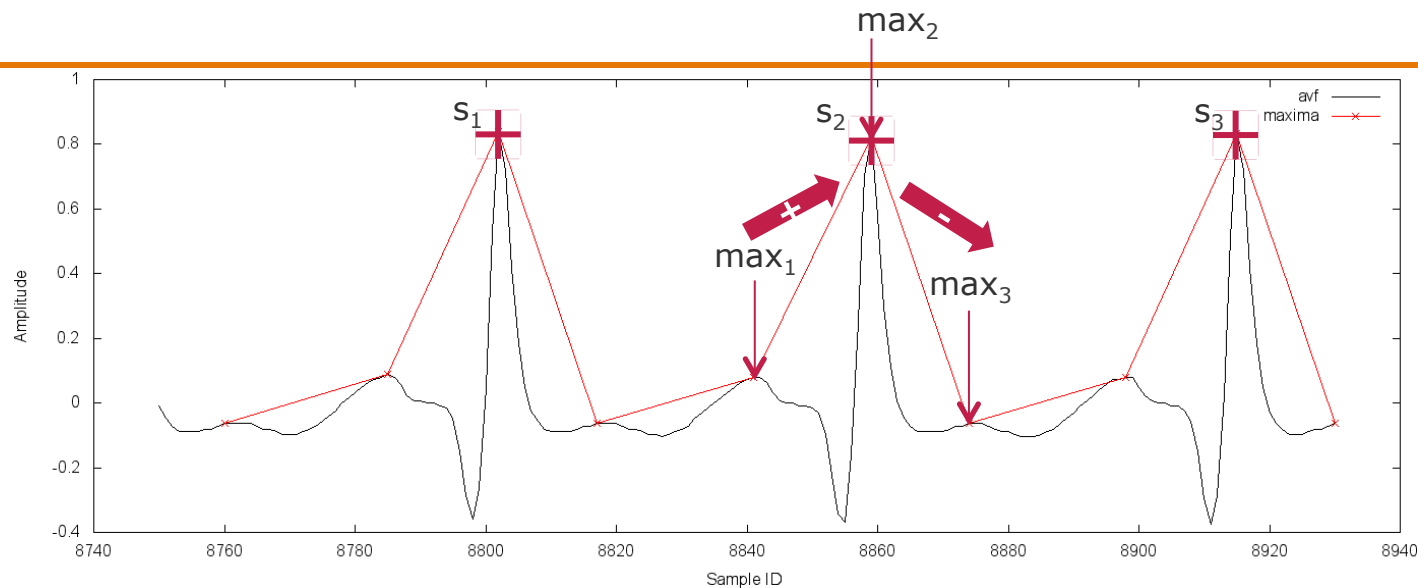


Future SOC Lab

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
24

# Heart Beat Segmentation



- Split data into single heartbeat segments
- Calculate heart rate from the number of segments per time span
- Extract minimum, maximum, and average amplitude
- Identification of change in slope

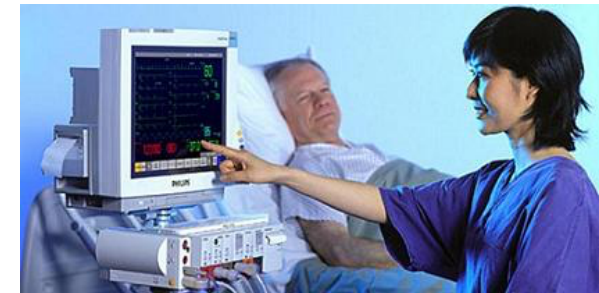
- Arrhythmia
- Tachycardia
- Bradycardia

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
25

# MIMIC III Database

- Clinical data of Hospital Information System (HIS)
- High-resolution waveform data incl. severe event annotations
- More than 42,000 patients and 58,000 ICU admissions
- Available data
  - Physiologic data
  - Demographics
  - Medications
  - Lab values
  - And more...



<http://www.physionet.org>

```
'sample interval','I','II','III','AVL','MCL1','ABP','PAP'  
'0.008 sec','mV','mV','mV','mV','mV','mmHg','mmHg'  
6669,-,-,-0.08510638,-,0.00000000,99.96003998,-  
6670,-,-,-0.08510638,-,0.00000000,97.56003902,-  
6671,-,-,-0.08510638,-,-0.03200000,95.16003806,-
```

## Data Acquisition

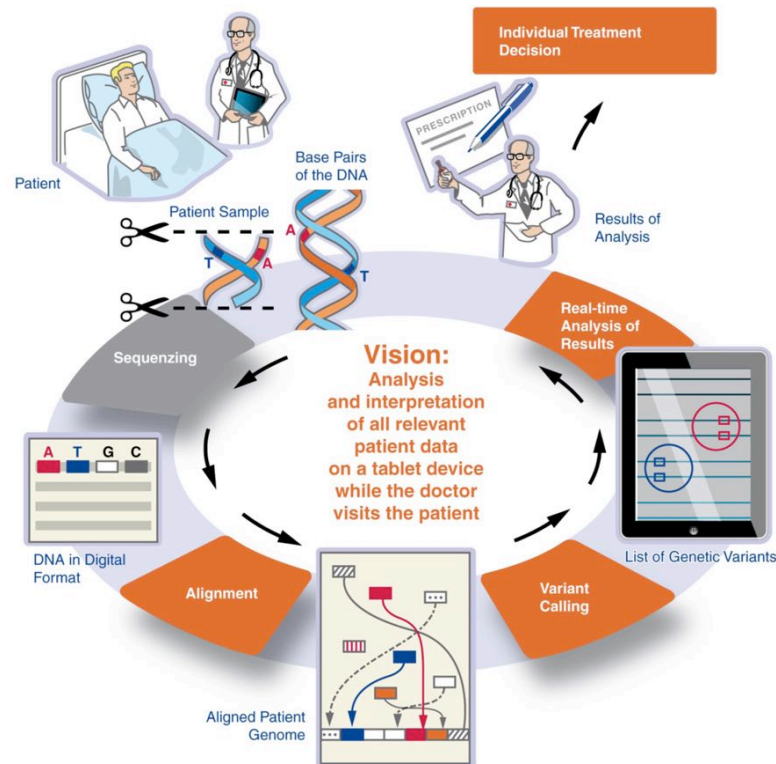
Data Management for  
Digital Health, Summer  
2017  
26

Access: <https://mimic.physionet.org/>



# 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



## Data Acquisition

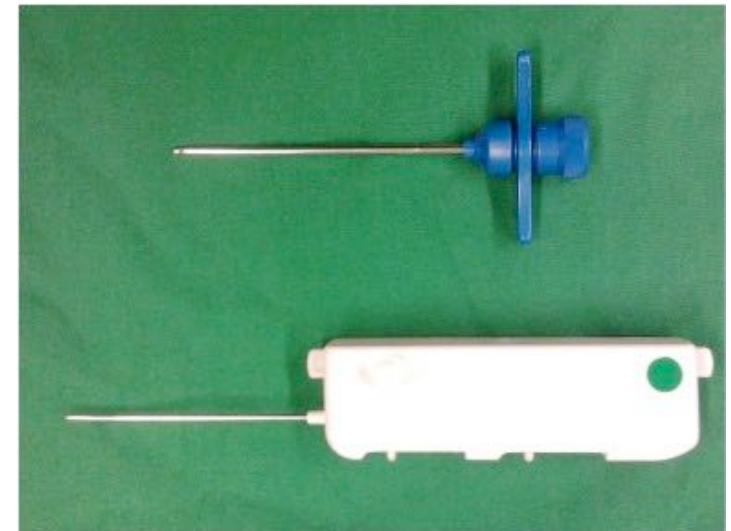
Data Management for  
Digital Health, Summer  
2017  
27

# Biopsy



Hasso  
Plattner  
Institut

- Biopsy := Extraction of tissue from body
- Purpose: Obtain tissue sample for analysis, e.g. abnormal vs. normal tissue
- Typically: Sample is processed by department of pathology and a report is created (duration: from minutes to days)
- Foundation for treatment decision
- Sample can be used for further tests, e.g. genome sequencing



<http://www.tumororthopaedie.org/biopsie.html>

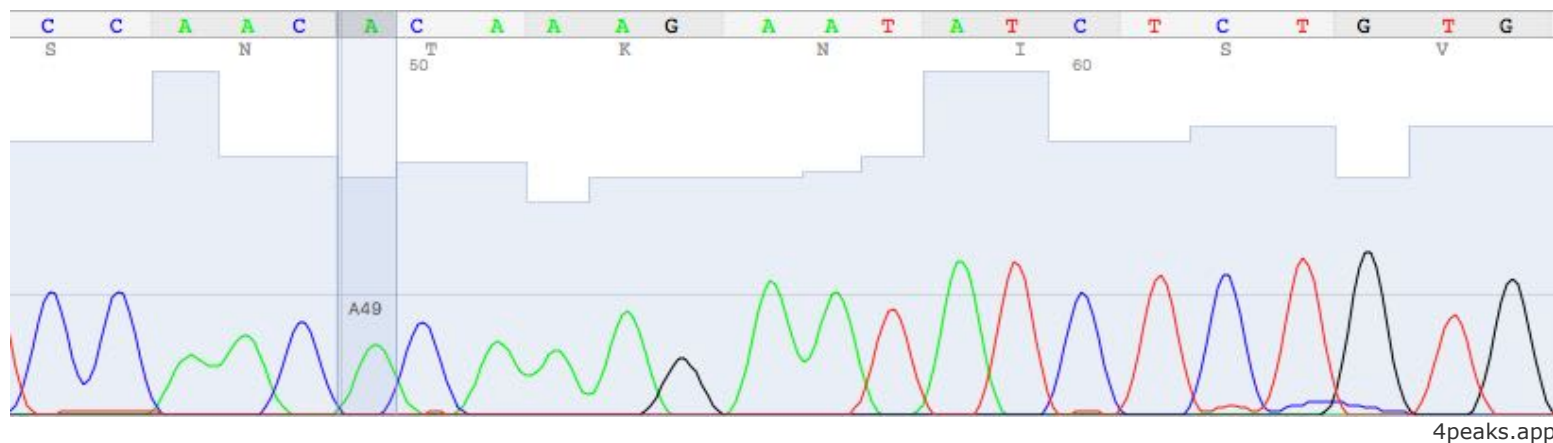
## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
28

# 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

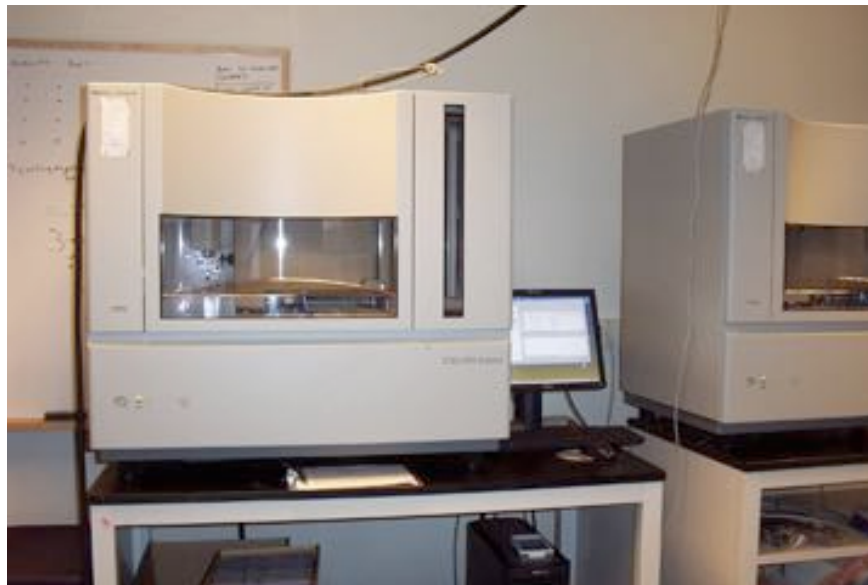


## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
29

## ABI Sequencing (1<sup>st</sup> gen)

- 2002: Sanger sequencing provides very high accuracy
- Accuracy: > 99.999%
- Throughput: 100 kbp / run (3hrs)
- Read length: 0.6-1 kbp
- Issues: time-intensive



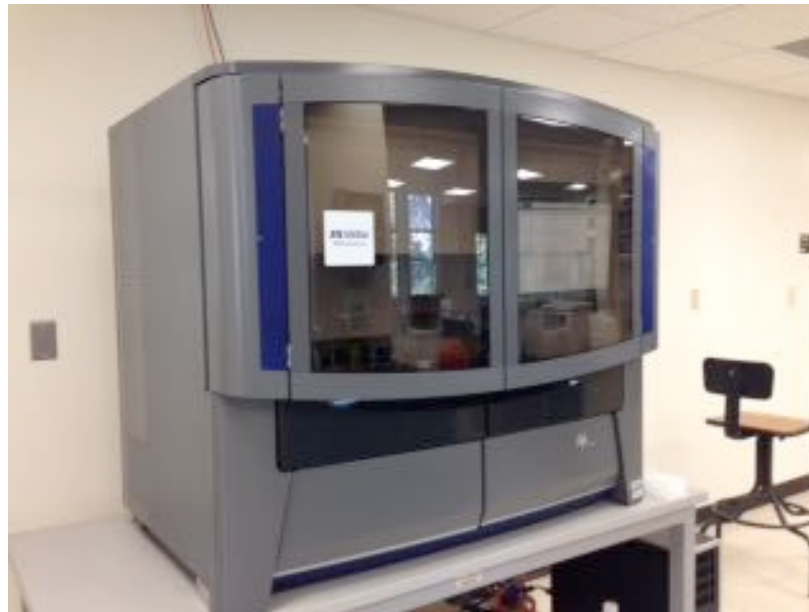
### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
30

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

## ABI Sequencing (2<sup>nd</sup> 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



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

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
31

## 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



### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
32

<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



## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
33

<http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html>

---

# 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



<https://nanoporetech.com/products>

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017

34

# 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



## Data Acquisition

Data Management for  
Digital Health, Summer  
2017

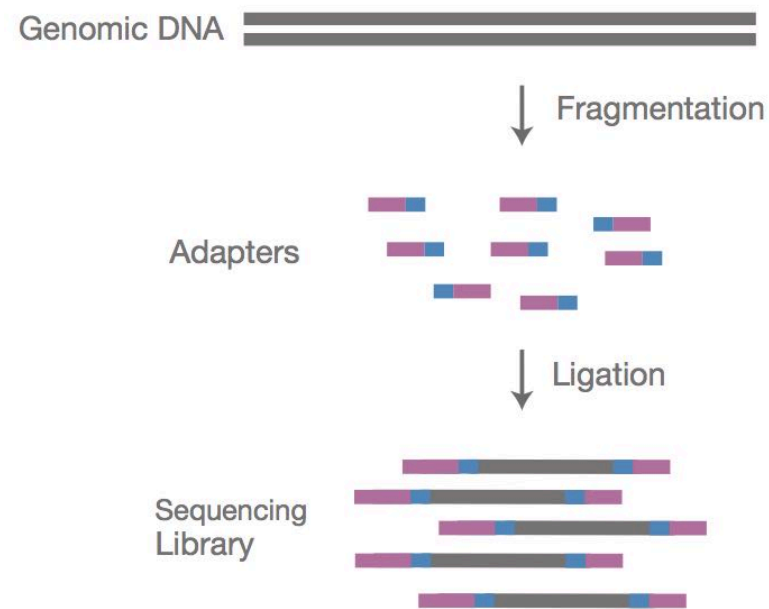
A. Rhoads and K. F. Au: PacBio Sequencing and Its Applications (2015)

<http://www.pacb.com/products-and-services/pacbio-systems/sequel/>

# Illumina Sequencing Process

## 1. Preparation

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



### Data Acquisition

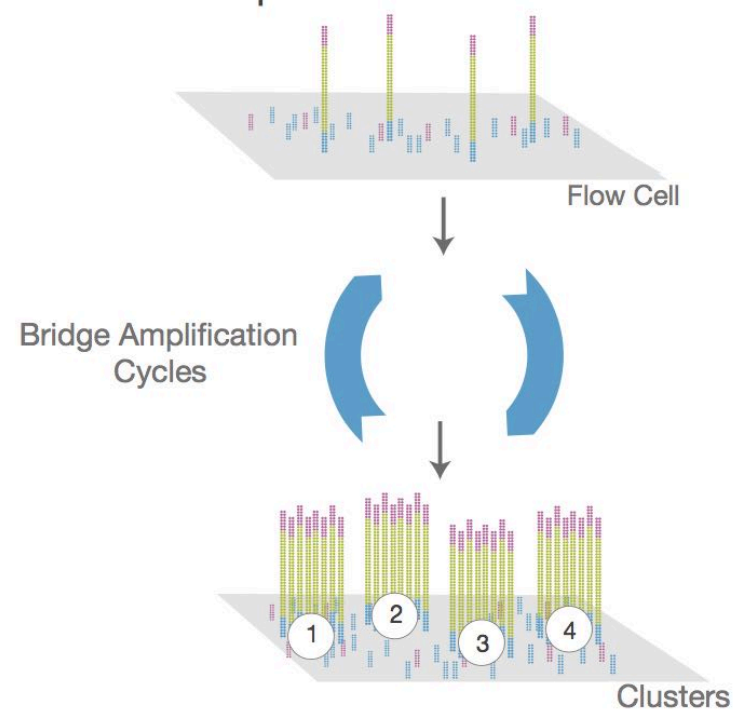
Data Management for  
Digital Health, Summer  
2017

36

# Illumina Sequencing Process

## 2. Amplification

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



### Data Acquisition

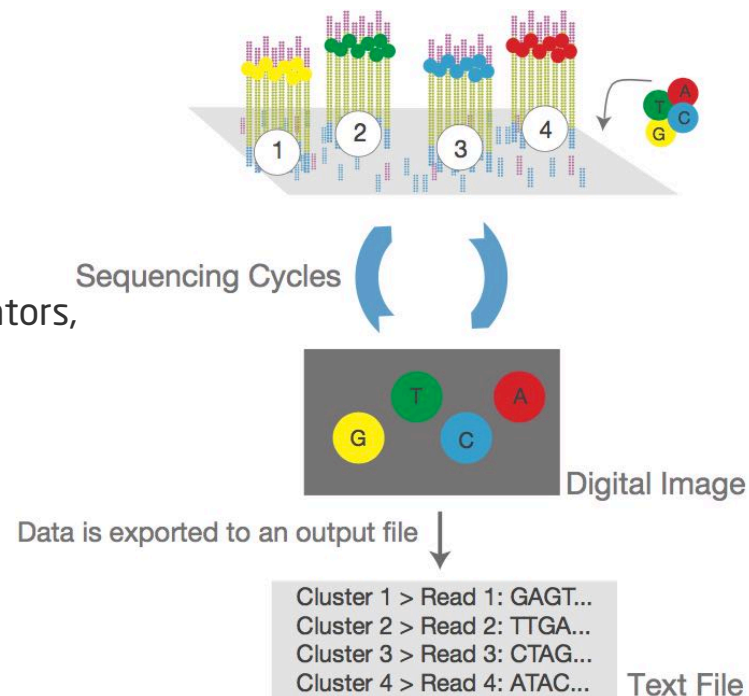
Data Management for  
Digital Health, Summer  
2017

37

# 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



### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

38



---

## Illumina Sequencing Process

---

- Double-stranded DNA is split into chunks of 200-800 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

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
39

- ```
@HJ40ITD02IGHKD rank=0016764 x=3351.0 y=603.5
CGTATCTACACAGGGTCAGGGTTCTGGATATTGGGAGAATATGGA
+
IIIIIIIIII=422:CA22///CFGGIIHHHBB>:/11::;2/4
@HJ40ITD02HBT0Z rank=0016788 x=2887.0 y=3969.0
CGTATCTACACAGGGTCGAGGTTCTGGAGTATCAGGTAACGAA
+
A@ADFDDBA?=8,,/,/,/——/11141428:7667...4200
@HJ40ITD02GKSZP rank=0016806 x=2580.0 y=819.0
CGTATCTACACAGGGTCAGGGTTCTGGATATAGGGCAGCACGGAC
+
FFFFFFFFFFFFFFD666ADD666??DFFFFHHHHHHHHHHHFFFFFFF
@HJ40ITD02F4FE5 rank=0016858 x=2393.0 y=2687.0
CGTATCTACACAGGGTCAGGGTTATGGATATCAGGTAACAGTCA
+
IIIIIIIIII@@@IIHHHHIIIIIGEEE@A<:5211121DDAD
@HJ40ITD02HNVGV rank=0017026 x=3025.5 y=893.0
CGTATCTACACAGGGTCAGGGTTCTGGATATTGGGAGAATATGA
+
IIIIIIIIIIIIHHHHIIHHHHIIIIIIIGG333390::C?@@@
@HJ40ITD02GIZMW rank=0017128 x=2559.5 y=2134.0
CGTATCTACACAGGGTCAGGGTTCTGGATATTGACCTAACTGCTG
+
IIIIIIIIIIIIHHHHIIHHHHIIIIIIIIIIIIIIIIIIIIIIIIII
```

Data Management for  
Digital Health, Summer  
2017  
40

## 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<br>(throughput mode) | 2x125 kbp<br>(paired) | >99.9 %                      | 800 Gbp (paired)     |
| 2013 | Pac Bio RS II                      | Up to 15 kbp          | >90% / >99 %<br>(multi-pass) | 6.75 Tbp             |
| 2014 | Oxford Nanopore<br>MinION          | Up to 5 kbp           | Up to 99%<br>(multi-pass)    | 115 Mbp              |

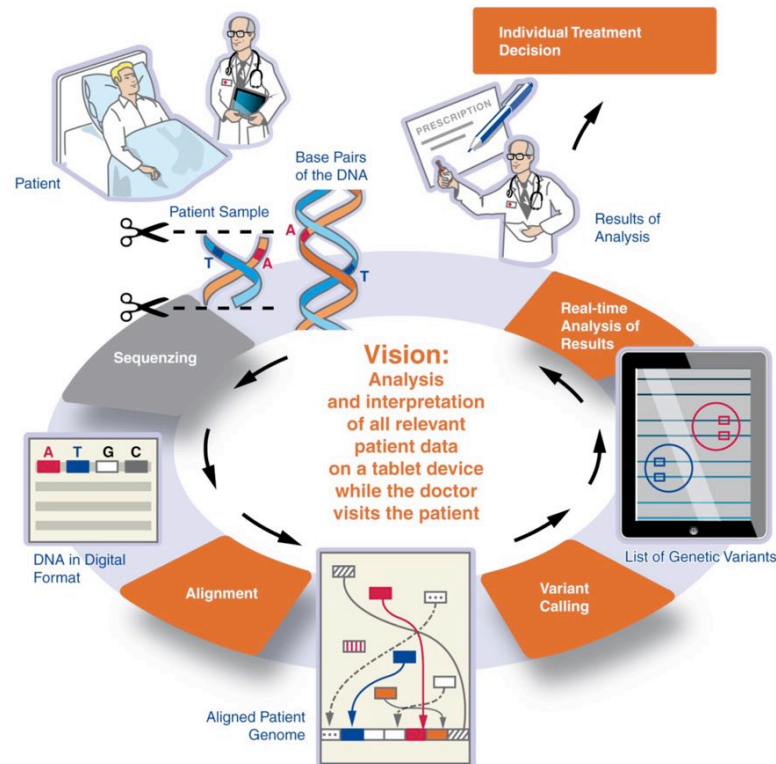
### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

41

# 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

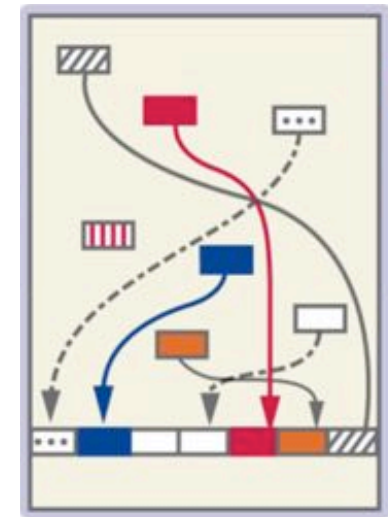


## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
42

# Alignment Overview

- Purpose: Mapping of DNA reads to a reference
- Input:
  - DNA reads := Sequence of nucleotides with a length of 100 bp up to some 1 kbp
  - Reference genome := Blueprint for alignment of DNA reads
- Output: Mapped DNA reads
  
- Bear in mind:
  - Less fraction in DNA reads, i.e. longer reads, allows more precise alignment
  - Reference from same origin improves mapping quality



## Data Acquisition

Data Management for  
Digital Health, Summer  
2017

---

## Selected Alignment Algorithms

### Needleman-Wunsch Algorithm

---

- Global alignment strategy
- Alignment score is defined by the value in the most lower right cell of the matrix
- Needleman, S. B. and Wunsch, C. D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins" in "Molecular Biology", 48(3): 443-53

#### **Data Acquisition**

Data Management for  
Digital Health, Summer  
2017



## Needleman-Wunsch Algorithm

- What is the best global alignment for sequences CTG and ACTGC?

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - |   |   |   |   |   |   |
| C |   |   |   |   |   |   |
| T |   |   |   |   |   |   |
| G |   |   |   |   |   |   |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

45

## Needleman-Wunsch Algorithm Matrix Initialization

- $M(0,0) = 0$

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - | 0 |   |   |   |   |   |
| C |   |   |   |   |   |   |
| T |   |   |   |   |   |   |
| G |   |   |   |   |   |   |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

46

## Needleman-Wunsch Algorithm Matrix Initialization

- $M(0,0) = 0$
- $M(i,0) = M(i-1,0) + \text{gap}()$ ,  $1 \leq i \leq m$
- $M(0,j) = M(0,j-1) + \text{gap}()$ ,  $1 \leq j \leq n$
- $\text{gap}() := -1$ , i.e. gap cost function

|   | -  | A  | C  | T  | G  | C  |
|---|----|----|----|----|----|----|
| - | 0  | -1 | -2 | -3 | -4 | -5 |
| C | -1 |    |    |    |    |    |
| T | -2 |    |    |    |    |    |
| G | -3 |    |    |    |    |    |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
47

## Needleman-Wunsch Algorithm

### Fill Matrix

- Weight function  $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$
- $D(i,j)$  defines value of matrix at coordinates  $(i,j)$

|   | -  | A  | C  | T  | G  | C  |
|---|----|----|----|----|----|----|
| - | 0  | -1 | -2 | -3 | -4 | -5 |
| C | -1 |    |    |    |    |    |
| T | -2 |    |    |    |    |    |
| G | -3 |    |    |    |    |    |

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
48

# Needleman-Wunsch Algorithm

## Fill Matrix

- Weight function  $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$
- $D(1,1) := \text{maximum of}$ 
  - ①  $D(0,0) + w(A,C) = 0 + (-1) = -1$
  - ②  $D(1,0) + \text{gap}() = -1 + (-1) = -2$
  - ③  $D(0,1) + \text{gap}() = -1 + (-1) = -2$

|   | -               | A               | C  | T  | G  | C  |
|---|-----------------|-----------------|----|----|----|----|
| - | 0 <sup>①</sup>  | -1 <sup>②</sup> | -2 | -3 | -4 | -5 |
| C | -1 <sup>③</sup> | D(1,1)          |    |    |    |    |
| T | -2              |                 |    |    |    |    |
| G | -3              |                 |    |    |    |    |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

# Needleman-Wunsch Algorithm

## Fill Matrix

■ Weight function  $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$

■  $D(1,1) := \text{maximum of}$

①  $D(0,0) + w(A,C) = 0 + (-1) = -1$

②  $D(1,0) + \text{gap}() = -1 + (-1) = -2$

③  $D(0,1) + \text{gap}() = -1 + (-1) = -2$

|   | -               | A               | C  | T  | G  | C  |
|---|-----------------|-----------------|----|----|----|----|
| - | 0 <sup>①</sup>  | -1 <sup>②</sup> | -2 | -3 | -4 | -5 |
| C | -1 <sup>③</sup> | -1              |    |    |    |    |
| T | -2              |                 |    |    |    |    |
| G | -3              |                 |    |    |    |    |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
50



## Needleman-Wunsch Algorithm

### Fill Matrix

- Repeat for all  $D(i,j)$  until matrix is filled

- Bear in mind: Filling the matrix can be performed in parallel

|   | -  | A  | C  | T  | G  | C  |
|---|----|----|----|----|----|----|
| - | 0  | -1 | -2 | -3 | -4 | -5 |
| C | -1 | -1 | 0  | -1 | -2 | -3 |
| T | -2 |    |    |    |    |    |
| G | -3 |    |    |    |    |    |

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
51

## Needleman-Wunsch Algorithm

### Fill Matrix

- Repeat for all  $D(i,j)$  until matrix is filled

- Bear in mind: Filling the matrix can be performed in parallel

|   | -  | A  | C  | T  | G  | C  |
|---|----|----|----|----|----|----|
| - | 0  | -1 | -2 | -3 | -4 | -5 |
| C | -1 | -1 | 0  | -1 | -2 | -3 |
| T | -2 | -2 | -1 | 1  | 0  | -1 |
| G | -3 | -3 | -2 | 0  | 2  | 1  |

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
52

## Needleman-Wunsch Algorithm

### Determine Best Global Alignment

- Trace path back from  $D(m,n)$  to origin  $D(0,0)$

|   | -  | A  | C  | T  | G  | C  |
|---|----|----|----|----|----|----|
| - | 0  | -1 | -2 | -3 | -4 | -5 |
| C | -1 | -1 | 0  | -1 | -2 | -3 |
| T | -2 | -2 | -1 | 1  | 0  | -1 |
| G | -3 | -3 | -2 | 0  | 2  | 1  |

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
53

# Needleman-Wunsch Algorithm

## Determine Best Global Alignment

- Reference: ACTGC
- Alignment: -CTG-
- Score of the alignment is: 1

- Match (diag)
- Mismatch (up)
- Gap (left)

- Bear in mind: Backtracing can be performed in parallel

|   | -  | A  | C  | T  | G  | C  |
|---|----|----|----|----|----|----|
| - | 0  | -1 | -2 | -3 | -4 | -5 |
| C | -1 | -1 | 0  | -1 | -2 | -3 |
| T | -2 | -2 | -1 | 1  | 0  | -1 |
| G | -3 | -3 | -2 | 0  | 2  | 1  |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
54

---

## Selected Alignment Algorithms

### Smith-Waterman Algorithm

---

- Determine optimal local alignments
- Adaption of Needleman-Wunsch algorithm
  - Initialize all cells within first row and column with zero
  - Alignment score is defined by highest value somewhere in the matrix
  - Backtracing from cell with alignment score to first cell containing zero
- Smith, T. F. and Waterman, M. S. (1981). "Identification of Common Molecular Subsequences" in "Molecular Biology", 147: 195-7.

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
55

## Smith-Waterman Algorithm

- What is the best local alignment for sequences CTG and ACTGC?

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - |   |   |   |   |   |   |
| C |   |   |   |   |   |   |
| T |   |   |   |   |   |   |
| G |   |   |   |   |   |   |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

56



## Smith-Waterman Algorithm Matrix Initialization

- $M(0,0) = M(i,0) = M(0,j) = 0 \mid 0 \leq i \leq m, 0 \leq j \leq n$

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 |   |   |   |   |   |
| T | 0 |   |   |   |   |   |
| G | 0 |   |   |   |   |   |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
57

## Smith-Waterman Algorithm

### Fill Matrix

- Weight function  $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$
- $\text{gap}() := -1$ , i.e. gap cost function
- $D(i,j)$  defines value of matrix at coordinates  $(i,j)$

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 |   |   |   |   |   |
| T | 0 |   |   |   |   |   |
| G | 0 |   |   |   |   |   |

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
58

# Smith-Waterman Algorithm

## Fill Matrix

■  $D(1,1) := \text{maximum of}$

- ①  $D(0,0) + w(A,C) = 0 + (-1) = -1$
- ②  $D(1,0) + \text{gap}() = 0 + (-1) = -1$
- ③  $D(0,1) + \text{gap}() = 0 + (-1) = -1$
- ④ 0

|   | -              | A                   | C | T | G | C |
|---|----------------|---------------------|---|---|---|---|
| - | 0 <sup>①</sup> | 0 <sup>②</sup>      | 0 | 0 | 0 | 0 |
| C | 0 <sup>③</sup> | D(1,1) <sup>④</sup> |   |   |   |   |
| T | 0              |                     |   |   |   |   |
| G | 0              |                     |   |   |   |   |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
59

# Smith-Waterman Algorithm

## Fill Matrix

■  $D(1,1) := \text{maximum of}$

①  $D(0,0) + w(A,C) = 0 + (-1) = -1$

②  $D(1,0) + \text{gap}() = 0 + (-1) = -1$

③  $D(0,1) + \text{gap}() = 0 + (-1) = -1$

④ 0

|   | -              | A              | C | T | G | C |
|---|----------------|----------------|---|---|---|---|
| - | 0 <sup>①</sup> | 0 <sup>②</sup> | 0 | 0 | 0 | 0 |
| C | 0 <sup>③</sup> | 0 <sup>④</sup> |   |   |   |   |
| T | 0              |                |   |   |   |   |
| G | 0              |                |   |   |   |   |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
60

## Smith-Waterman Algorithm Fill Matrix

- Repeat for all  $D(i,j)$  until matrix is filled

- Bear in mind: Filling the matrix can be performed in parallel

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 1 | 0 | 0 | 1 |
| T | 0 | 0 | 0 | 2 | 1 | 0 |
| G | 0 | 0 | 0 | 1 | 3 | 2 |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
61

## Smith-Waterman Algorithm

### Determine Local Alignments

- Trace path back from  $\max(D(i,j))$  to first  $D(i,j) = 0$

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 1 | 0 | 0 | 1 |
| T | 0 | 0 | 0 | 2 | 1 | 0 |
| G | 0 | 0 | 0 | 1 | 3 | 2 |

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
62



# Smith-Waterman Algorithm

## Determine Local Alignments

- Reference: ACTGC
- Local alignment: CTG
- Score of the alignment is: 3

- ← Match
- ← Mismatch
- ← Gap

- Bear in mind: Backtracing can be performed in parallel for multiple local optima

|   | - | A | C | T | G | C |
|---|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 1 | 0 | 0 | 1 |
| T | 0 | 0 | 0 | 2 | 1 | 0 |
| G | 0 | 0 | 0 | 1 | 3 | 2 |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
63

---

## Selected Alignment Algorithms

### Burrows-Wheeler Aligner

---

- Alignment of short read against long reference sequence
  - Uses Burrows-Wheeler Transform (BWT) to optimize search
  - BWT (aka block-sorting compression) := Rearrangement of character string, which tends to group similar characters by rotation and lexicographic ordering
  - Output of BWT supports compression as it contains repeated characters
- 
- Li, H. and Durbin, R. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform" in "Bioinformatics", 25(14): 1754-60

#### **Data Acquisition**

Data Management for  
Digital Health, Summer  
2017

---

## Burrows-Wheeler Transform Example

---

- BWT of character string: \*ENGINEERING#

# Burrows-Wheeler Transform

## 1. Create all Rotations

■ \*ENGINEERING#

|    |                |
|----|----------------|
| 1  | *ENGINEERING#  |
| 2  | ENGINEERING#*  |
| 3  | NGINEERING#*E  |
| 4  | GINEERING#*EN  |
| 5  | INEERING#*ENG  |
| 6  | NEERING#*ENGI  |
| 7  | EERING#*ENGINE |
| 8  | ERING#*ENGINEE |
| 9  | RING#*ENGINEER |
| 10 | ING#*ENGINEERI |
| 11 | NG#*ENGINEERIN |
| 12 | G#*ENGINEERING |
| 13 | #*ENGINEERING  |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
66

# Burrows-Wheeler Transform

## 2. Sort Rotations

■ \*ENGINEERING#

|    |                |
|----|----------------|
| 7  | EERING#*ENGINE |
| 2  | ENGINEERING#*  |
| 8  | ERING#*ENGINE  |
| 4  | GINEERING#*EN  |
| 12 | G#*ENGINEERIN  |
| 5  | INEERING#*ENG  |
| 10 | ING#*ENGINEER  |
| 6  | NEERING#*ENGI  |
| 3  | NGINEERING#*E  |
| 11 | NG#*ENGINEERI  |
| 9  | RING#*ENGINEE  |
| 1  | *ENGINEERING#  |
| 13 | #*ENGINEERING  |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
67

# Burrows-Wheeler Transform

## 3. Assemble Output from Last Character

- \*ENGINEERING#
- N\*ENNGRIEIE#G
- N\*EN<sup>2</sup>GR(IE)<sup>2</sup>#G (compressed after applying RLE)

|    |                |
|----|----------------|
| 7  | EERING#*ENGINE |
| 2  | ENGINEERING#*  |
| 8  | ERING#*ENGINE  |
| 4  | GINEERING#*E   |
| 12 | G#*ENGINEERIN  |
| 5  | INEERING#*ENG  |
| 10 | ING#*ENGINEER  |
| 6  | NEERING#*ENGI  |
| 3  | NGINEERING#*E  |
| 11 | NG#*ENGINEERI  |
| 9  | RING#*ENGINEE  |
| 1  | *ENGINEERING#  |
| 13 | #*ENGINEERING  |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
68

# Burrows-Wheeler Transform

## 1. Create all Rotations

■ \*DIGITAL#

|   |           |
|---|-----------|
| 1 | *DIGITAL# |
| 2 | DIGITAL#* |
| 3 | IGITAL#*D |
| 4 | GITAL#*DI |
| 5 | ITAL#*DIG |
| 6 | TAL#*DIGI |
| 7 | AL#*DIGIT |
| 8 | L#*DIGITA |
| 9 | #*DIGITAL |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

# Burrows-Wheeler Transform

## 2. Sort Rotations

■ \*DIGITAL#

|   |           |
|---|-----------|
| 7 | AL#*DIGIT |
| 2 | DIGITAL#* |
| 4 | GITAL#*DI |
| 3 | IGITAL#*D |
| 5 | ITAL#*DIG |
| 8 | L#*DIGITA |
| 6 | TAL#*DIGI |
| 1 | *DIGITAL# |
| 9 | #*DIGITAL |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
70



## Burrows-Wheeler Transform

### 3. Assemble Output from Last Character

- T\*IDGAI#L
- Does not always improve compressibility!

|   |           |
|---|-----------|
| 7 | AL#*DIGIT |
| 2 | DIGITAL#* |
| 4 | GITAL#*DI |
| 3 | IGITAL#*D |
| 5 | ITAL#*DIG |
| 8 | L#*DIGITA |
| 6 | TAL#*DIGI |
| 1 | *DIGITAL# |
| 9 | #*DIGITAL |

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

---

## Selected Alignment Tools

---

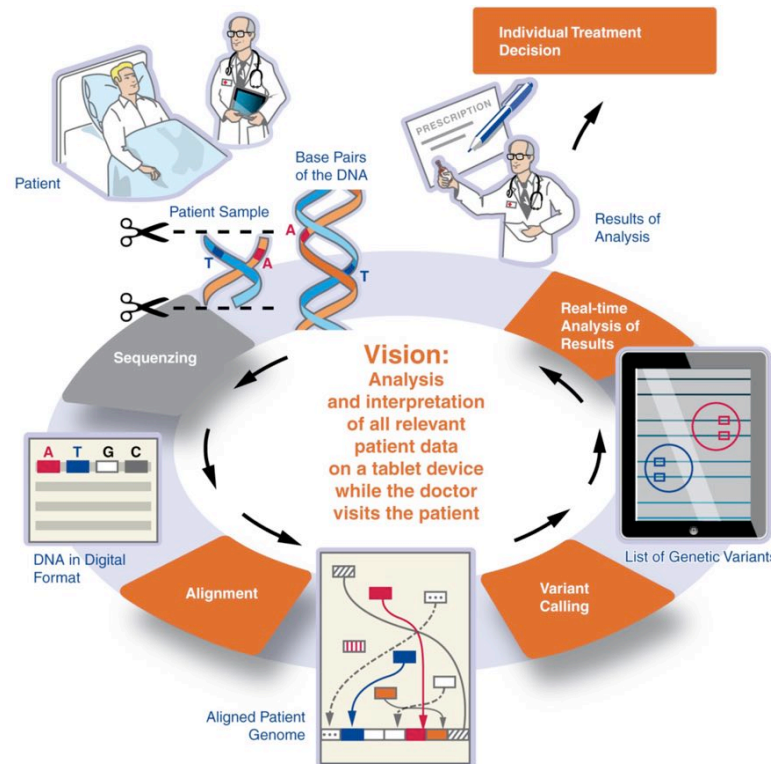
- BWA: Smith-Waterman + BWT to keep memory footprint low
- Bowtie: Similar to Smith-Water/Needleman-Wunsch + BWT
- HANA Aligner (based on IMDB): BWA + FM index/BWT to speed-up match detection
- Isaac (commercialized by Illumina): Smith-Waterman
- Torrent Mapping Alignment Program (TMAP) (commercialized by IonTorrent): Smith-Waterman + FM index/BWT

### **Data Acquisition**

Data Management for  
Digital Health, Summer  
2017  
72

# 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

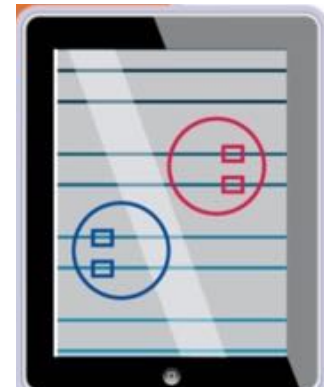


## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
73

# Variant Calling Overview

- Purpose: Variant Calling := Detect variations within a genome
- Input:
  - Mapped DNA reads, i.e. output of alignment process
  - Reference genome
- Output: List of variants

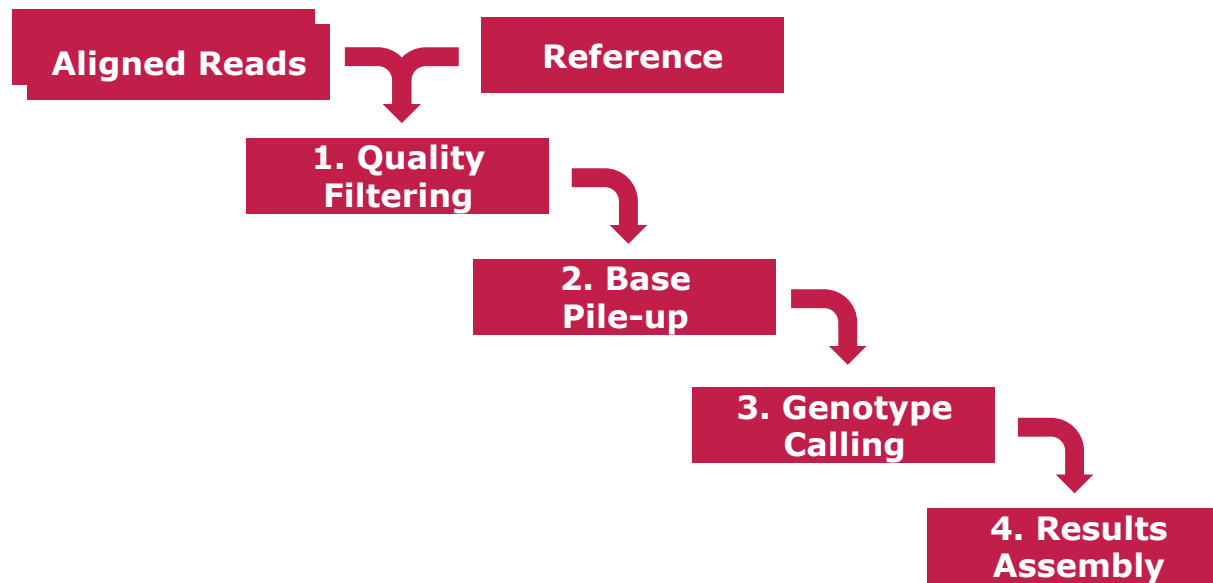


- Bear in mind:
  - Read depth at  $\text{pos}_i$  := Number of nucleotides storing information about  $\text{pos}_i$

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
74

# Variant Calling Process



## Data Acquisition

Data Management for  
Digital Health, Summer  
2017

# 1. Quality Filtering

- Extract locations from mapped reads where mapping issues were detected

- Reference

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| A | C | G | C | R | A | G | A | T | A |
|---|---|---|---|---|---|---|---|---|---|

- Read

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| - | - | G | C | A | T | G | A | T | A |
|---|---|---|---|---|---|---|---|---|---|

- CIGAR

|    |    |
|----|----|
| 2D | 8M |
|----|----|

| Op | BAM | Description                                           |
|----|-----|-------------------------------------------------------|
| M  | 0   | alignment match (can be a sequence match or mismatch) |
| I  | 1   | insertion to the reference                            |
| D  | 2   | deletion from the reference                           |
| N  | 3   | skipped region from the reference                     |
| S  | 4   | soft clipping (clipped sequences present in SEQ)      |
| H  | 5   | hard clipping (clipped sequences NOT present in SEQ)  |
| P  | 6   | padding (silent deletion from padded reference)       |
| =  | 7   | sequence match                                        |
| X  | 8   | sequence mismatch                                     |

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
76

## 2. Base Pile-up

- Reference (FASTA)
- Aligned read 1

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| A | C | G | C | R | A | G | A | T | A |
|   |   | G | C | A | T | G | A | T | A |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
77

## 2. Base Pile-up

■ Reference (FASTA)

■ Aligned read 1

...

■ Aligned read 8

■ Alleles

| A | C | G | C | R | A | G | A | T | A |
|---|---|---|---|---|---|---|---|---|---|
|   |   | G | C | A | T | G | A | T | A |
| A | C | G | C | G | T | G | A | T | A |
| A | T | G | C | G | T | G | A |   |   |
| A | C | G | C | G | A | G |   |   |   |
|   | C | G | C | A | T | G | A | T | A |
| A | C | G | C | G | T |   |   |   |   |
|   |   |   | C | A | T | G | A | T | A |
|   | C | G | C | A | T | G | A | T | A |

SNP or Read Error

| a | 4 | 5 | 7 | 8 | 4 | 1 | 7 | 6 | 5 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| b | 0 | 1 | 0 | 0 | 4 | 7 | 0 | 0 | 0 | 0 |



aa



ab



bb

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

78



## Reasons for Mismatches?

- Error(s) in the wet lab process



### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
79

---

## Reasons for Mismatches?

---

- Error during alignment phase, i.e. incorrect mapping of DNA chunk
- Error during base calling, i.e. algorithm only indicates probability
- Incorrect reference

- Bear in mind: Better references and algorithms may reduce the error!

### **Data Acquisition**

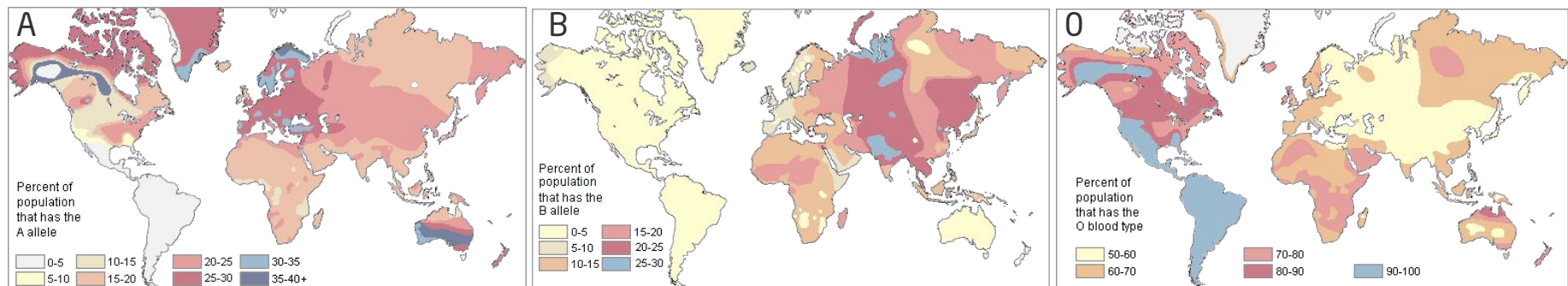
Data Management for  
Digital Health, Summer  
2017

80

## Reasons for Mismatches?

- Single Nucleotide Polymorphism (SNP) on the DNA strand

- Example: Worldwide distribution of blood types



[http://anthro.palomar.edu/vary/vary\\_3.htm](http://anthro.palomar.edu/vary/vary_3.htm)

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

81

## Friendly Reminder: Donate Blood to Save Lives

- When: May 9, 2017 (10.30am - 3.30pm)
- Where: HPI building A, 2<sup>nd</sup> floor



### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
82

### 3. Genotype Calling

- Purpose: Eliminate impact of noise and poor reading quality
- How: Compute probability for a genotype G given read context data D per sample
- Uses Bayes' theorem

- Recap Bayes' theorem: 
$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

- Given are two events A and B
- $P(A|B)$  defines the conditional probability for event A after event B
- $P(B|A)$  defines the conditional probability for event B after event A
- Relates conditional probability  $P(A|B)$  to  $P(B|A)$  for events A and B and  $P(B) > 0$ .

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

### 3. Genotype Calling

- Which genotype  $G$  has the highest posterior probability given the data  $D$ ?
- Therefore, calculate posterior probability  $P(G|D)$ .

- Bayes' Theorem:

- $D$ : All observation about current position  $i$   $\{D_i, \dots, D_n\}$
- $P(G)$ : Genotype probability  $\{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$
- $P(G_i)$ : Prior probability
- $P(D|G_i)$ : Genotype likelihood

$$\begin{aligned} P(G|D) &= \frac{P(D|G)P(G)}{P(D)} \\ &= \frac{P(D|G) P(G)}{\sum_{i=1}^n P(D|G_i) P(G_i)} \end{aligned}$$

- Heng Li (2011): "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from seq. data"

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
84

### 3. Genotype Calling

#### Computation of Prior Probability $P(G_i)$

- Affected by:
  - Number of (known) SNPs across the complete genome
  - Distribution of SNPs
  - Allele frequency

An example of prior probability for a dbSNP G/T site used in Li et al (2009)

|   | A                    | C                    | G                   | T                   |
|---|----------------------|----------------------|---------------------|---------------------|
| A | $4.55 \cdot 10^{-7}$ | $9.11 \cdot 10^{-8}$ | $9.1 \cdot 10^{-5}$ | $9.1 \cdot 10^{-5}$ |
| C |                      | $4.55 \cdot 10^{-7}$ | $9.1 \cdot 10^{-5}$ | $9.1 \cdot 10^{-5}$ |
| G |                      |                      | .454                | .0909               |
| T |                      |                      |                     | .454                |

R Li et al (2009) Genome Research 19:1124-132

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)}$$
$$= \frac{P(D|G) P(G)}{\sum_{i=1}^n P(D|G_i) P(G_i)}$$

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017



### 3. Genotype Calling

#### Computation of Genotype Likelihood $P(D|G_i)$

- Genotype likelihood depends on the surrounding data at position  $i$
- Includes present values and base quality scores from sequencing
- Where to find base quality score?
- Recap FASTQ file format
- Line 4 describes base quality score

```
@HJ40ITD02GKSZP rank=0016806 x=2580.0 y=819.0
CGTATCTACACAGGGTCAGGGTTCTGGATATAGGGCAGCACGGAC
+
FFFFFFFFFFFFD666ADD666??DFFFFHHHHHHHHHHIHHHHHHHH
@HJ40ITD02F4FE5 rank=0016858 x=2393.0 y=2687.0
CGTATCTACACAGGGTCAGGGTTATGGATATCAGGTAAACAGTCA
+
IIIIIIIIIIII@@@IIHHHHIIIIIGEEEE@A<:5211121DDAD
@HJ40ITD02HNVGV rank=0017026 x=3025.5 y=893.0
CGTATCTACACAGGGTCAGGGTTCTGGATATTGGGGGAGAATATGA
+
IIIIIIIIIIIIHHHHIIHHHHIIIIIIIIIGG333390::C?@@@
@HJ40ITD02GIZMW rank=0017128 x=2559.5 y=2134.0
CGTATCTACACAGGGTCAGGGTTCTGGATATTGACCTAACTGCTG
+
IIIIIIIIIIIIHHHHIIHHHHIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)}$$

$$= \frac{P(D|G) P(G)}{\sum_{i=1}^n P(D|G_i) P(G_i)}$$

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017



---

### 3. Genotype calling

#### Genotype consensus

---

- Select the genotype with the highest probability, i.e.
- $g_i = \operatorname{argmax} P(g_i|D)$  for  $g_i$  in  $(\langle a,a \rangle, \langle a,b \rangle, \langle b,b \rangle)$
- Assembly results for all positions  $i$ .

- Bear in mind: Variant calling can be performed in parallel for multiple loci.

#### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

## 4. Results Assembly

- Results are stored in Variant Calling Format (VCF)
- VCF is extensible, i.e. can store an arbitrary number of attribute/value pairs
- Result consists of:

- Header defining attributes

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

- One entry per variant (fixed number of attributes)

| CHROM | POS       | ID          | REF | ALT | QUAL | FILTER | INFO | FORMAT | SAMPLE |
|-------|-----------|-------------|-----|-----|------|--------|------|--------|--------|
| chr7  | 140753336 | rs113488022 | T   | A   | 61   | PASS   | NS=1 | GT     | 0/1    |

### Data Acquisition

Data Management for  
Digital Health, Summer  
2017

# Variant Callers

- Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores Genome Research 18:1851-1858
- Maq was the first widely used variant caller
- Latest examples:  
Broad's Genome Analysis Tool Kit
  - Unified Genotyper
  - Haplotype Caller
  - ...

## Run Maq Now

Follow these steps to try Maq. All you need is a reference sequence file in the FASTA format.

1. Prepare a reference sequence (ref.fasta). Better a bacterial genome.
2. Download maq, maq-data and maqview at the [download page](#).
3. Copy maq, maq.pl and maq\_eval.pl to the \$PATH or to the same directory.
4. Simulate diploid reference and read sequences, map reads, call variants and evaluate the results in one go:

```
maq.pl demo ref.fasta calib-30.dat
```

where *calib-30.dat* is contained in maq-data.

5. View the alignment:

```
cd maqdemo/easyrun;  
maqindex -i -c consensus.cns all.map;  
maqview -c consensus.cns all.map
```

**Even for advanced maq users, running `maq.pl demo' is recommended. You may find something helpful.**

<http://maq.sourceforge.net/>

## Data Acquisition

Data Management for  
Digital Health, Summer  
2017  
89


---

## What's next?

### 1<sup>st</sup> Exercise

---

#### ■ Procedure:

- Pass online questionnaire on The logo for the OPEN HPI project, with the word "OPEN" in large orange letters and "HPI" in a smaller red box.
- Conduct the exercise whenever it fits your schedule
- However, complete the exercise prior to its scheduled deadline (tba)

#### ■ Content you should review:

- Biology recap,
- Sequencing technology, and
- Alignment and variant calling algorithms.

#### **Data Acquisition**

Data Management for  
Digital Health, Summer  
2017