

Von der DNA zur Datenbank: Sequenzierung & Assemblierung



Prof. Daniel Huson

ERHARD-KARLS
UNIVERSITÄT
TÜBINGEN

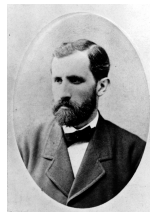


NANYANG
UNIVERSITY Singapore Center on Environmental
Life Sciences Engineering (SCELESE)

ZBIT Center for
Bioinformatics

ERHARD-KARLS
UNIVERSITÄT
TÜBINGEN
Center for Bioinformatics

Entdeckung der DNA



Friedrich Miescher
(1844-1895)

1869 Miescher entdeckte DNA in der
Küche des Schlosses Hohentübingen

Daniel Huson 2013

2

ERHARD-KARLS
UNIVERSITÄT
TÜBINGEN
Center for Bioinformatics

Biomolekulare Grundlage des Lebens



...
A-T
C-G
G-C
T-A
A-T
A-T
...

1953 Watson und Crick

- Struktur der DNA ist eine Doppelhelix
- Die **Reihenfolge** der Basen enthält Informationen, die in der Zelle verarbeitet werden

Daniel Huson 2013

3



4



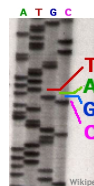
1955 Biochemiker Frederick Sanger bestimmt die erste vollständige Aminosäuresequenz eines Proteins, Insulin.

1 malwmrl1pl lallalwgpd paaafvnqhl cgshlvealy lvcgergffy tpktrreaed
61 lqvqgvelgg gpgagslqpl alegslqkrg iveqcctsic slyqlenycn

Daniel Huson 2013



1975 Frederick Sanger entwickelt die Kettenabbruchmethode zur Sequenzierung von DNA



Wikipedia.org

- Genomics ist das Studium der Genomsequenzen einzelner Organismen



- Sequenzierung und Assemblierung
- Annotation und Analyse
- Genomvergleich

Daniel Huson 2013

7

- **Genomgrößen:**
 - Virus 5-10 Kilobasen (kb)
 - Bakterien: 1-10 Millionen Basen (Mb)
 - *Arabidopsis thaliana*: 120 Mb
 - Fruchtfliege: 140 Mb
 - Mensch: 3 Milliarden Basen (Gb)

“Genetics is a way of thinking.
Genomics is a set of tools.”

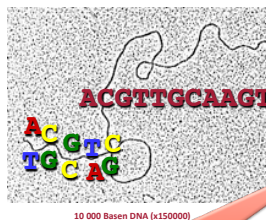
Mary-Claire King (Science, Feb 2011)

Daniel Huson 2013

8

[illegible]

Probe



10 000 Basen DNA (x150000)

**Gelesene DNA Sequenz
heißt „Read“**



Datenbank

9


Sanger Sequenzierung

DNA Polymerase
TCACAATCAACTGCGCTATAGACGGTTG
AGTGTTAGTTGACGCGATATCTGCC...

C G T A A C T C
 A C A T G C T G
 A T A C A G T G

Deoxynukleotide

Daniel Hoeser 2017

Diagram illustrating Sanger sequencing. A DNA template strand (3' to 5') is shown with the sequence XXXTCACAATCAACTGCGCTATAGACGGTTG. A DNA Polymerase III (Pol III) is shown synthesizing a new strand (5' to 3') starting from a primer (AGT). The primer is followed by the sequence AGTGTTA. The next bases to be added are AGTGTTAGTTGACGCGATATC. The bases are color-coded: A (green), G (red), T (blue), and C (yellow).

INTERNET-BASED
UNIVERSITÄT
TUBINGEN
Center for Bioinformatics

Kapillar-Technologie

A

C

G

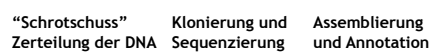
T

Daniel Huson, 2013

The sequencers create an image of the DNA samples being decoded. The four letters of the genetic code -- A, C, T, G -- each are assigned a color.



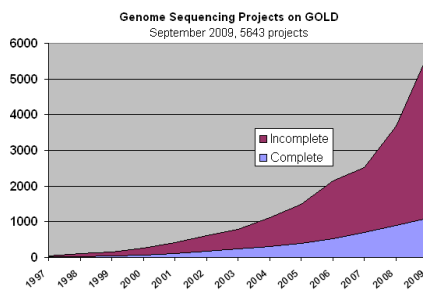
Die Sequenzierung des Genoms des Menschen



Assemblierung



Sequenzierungsprojekte



GOLD: Genomes online database
www.genomesonline.org

Second-Generation Sequenzierungstechnologien



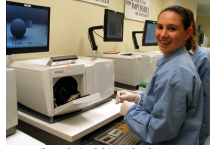
- Seit 2005 neue Hochdurchsatzmethoden

www.454.com

ARTICLES

Genome sequencing in microfabricated high-density picolitre reactors

Marcus Margolis, Michael Egholm¹, William E. Altmann², Said Adani³, José Sader⁴, Lisa A. Benham⁵, Jay Berka⁶, Michael S. Brubaker⁷, Wai-Chun Chan⁸, Zhihan Chen⁹, Scott A. Dowell¹⁰, Lei Dai¹¹, Joseph M. Fries¹², Xavier V. Gomes¹³, Brian C. Goodwin¹⁴, Wen Hu¹⁵, Thomas P. Hjeltnes¹⁶, Chun He Juo¹⁷, Gerard P. Jinks¹⁸, Sylvester C. Jordan¹⁹, Maria L. L. Alqueiroz²⁰, Thomas P. Jaeger²¹, Khamba B. Janga²², Jong-Bun Kim²³, James R. King²⁴, James R. Lanza²⁵, John H. Loomer²⁶, Steven M. McKenna²⁷, Ling Mei²⁸, Jing Li²⁹, Kenton L. Lohman³⁰, Hong Lu³¹, Vinod S. Mahalingam³², Keith E. Mohr³³, Michael F. McLeod³⁴, Eugene W. Myers³⁵, Elizabeth Nickerson³⁶, John R. Nisbels³⁷, Ramona Planic³⁸, Bernard P. Puc³⁹, Michael T. Ranan⁴⁰, George C. Roth⁴¹, Gary J. Sanks⁴², Jon Fredrik Semmen⁴³, John W. Simpson⁴⁴, Marianne Skovslien⁴⁵, Karrie R. Tartaro⁴⁶, Alexander Teyssie⁴⁷, Karl A. Vigg⁴⁸, Greg A. Volkmann⁴⁹, Shelly H. Wang⁵⁰, Yong Wang⁵¹, Michael P. Weiner⁵², Pengxiang Yu⁵³, Richard B. Bagley⁵⁴ and Jonathan M. Rothberg⁵⁵



Source: Stephan C. Schuster, Penn State.

Nature, 2005

Daniel Huson 2013

19

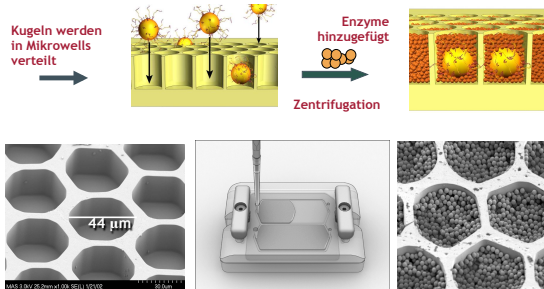
Emulsion Amplifikation



Daniel Huson 2013

Source: 454

Kugeln verteilen auf einer PicoTiterPlate™



Daniel Huson 2013

LIBRARY STATE
UNIVERSITÄT
TÜBINGEN
Center for Bioinformatics
Pyrosequenzierung

- Parallele Sequenzierung der Fragmente:

Fluss A
Fluss T

Daniel Huson 2013

LIBRARY STATE
UNIVERSITÄT
TÜBINGEN
Center for Bioinformatics
Pyrosequenzierung

- Parallele Sequenzierung der Fragmente:

Fluss A
Fluss T
Fluss G
usw...

Daniel Huson 2013

LIBRARY STATE
UNIVERSITÄT
TÜBINGEN
Center for Bioinformatics
Pyrosequenzierung

- Basenbestimmung durch Analyse einer Serie von Bildern

Sequencing and Basecalling Results for 191base Read

TCAGCGTAAG...

Source: 454

Daniel Huson 2013

Gegenüberstellung Sanger – Next Gen

- Erste Generation (Sanger Sequenzierung):
 - 100kb/Lauf, mittlere Read Länge 1000bp, Preis: 500\$/Mb
- Zweite Generation:
 - Roche(454): 600 Mb/Lauf, 600bp, 20\$/Mb
 - Illumina: 400 Gb/Lauf, 100bp, 0.50\$/Mb
 - SOLiD: 500 Gb/Lauf, 50bp, 0.50\$/Mb
- Dritte Generation:
 - PacBio SMRT: 25Gb/Lauf, >1000bp, ?\$/Mb
- Ganz neu:
 - Ion Torrent: 100bp/Lauf, 10x billiger als Roche(454)

Daniel Huson 2013

25

Assemblierung

Daniel Huson 2013

26

Assemblierung

- Sequenzierung produziert Millionen von kurzen DNA Reads
- Bioinformatik: Suche überlappende Sequenzen und füge sie zusammen

Daniel Huson 2013

27

UNIVERSITÄT TUBINGEN
Center for Bioinformatics

NCBI Datenbank

- Die Hauptresource für Sequenzdaten ist NCBI:

Daniel Huzon 2013

UNIVERSITÄT TUBINGEN
Center for Bioinformatics

NCBI Datenbank

Daniel Huzon 2013

UNIVERSITÄT TUBINGEN
Center for Bioinformatics

NCBI Datenbank

Hund-Genom?

Daniel Huzon 2013
