Next-generation DNA sequencing

Diana Le Duc, M.D.

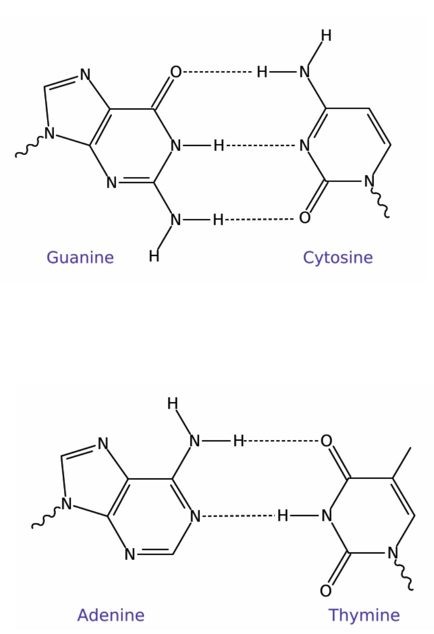
Biochemistry Institute, Medical Faculty, University of Leipzig

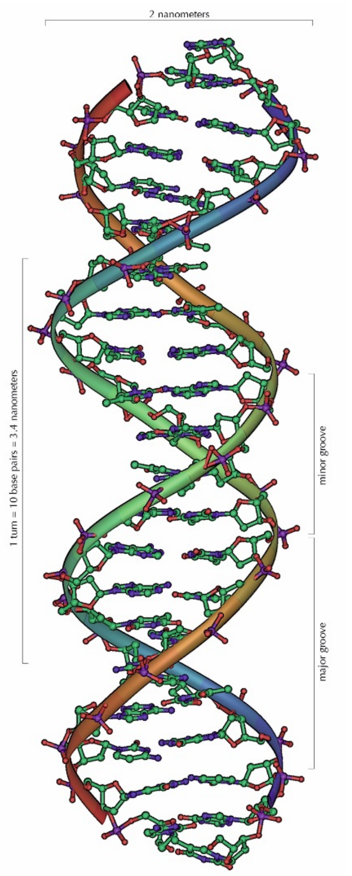
Statistical Analysis of RNA-Seq Data , University of Leipzig, 18th of April 2012

Gabriela-Diana.LeDuc@medizin.uni-leipzig.de

Deoxyribonucleic acid (DNA)

- Discovery (Miescher, 1869)
- Carrier of genetic information (Avery/MacLeod/ McCarty, 1944)
- Structural model (Watson/ Crick/Wilkins/Franklin, 1953)
- Replication using complementary base pairing
- Reading its information start early 1970s

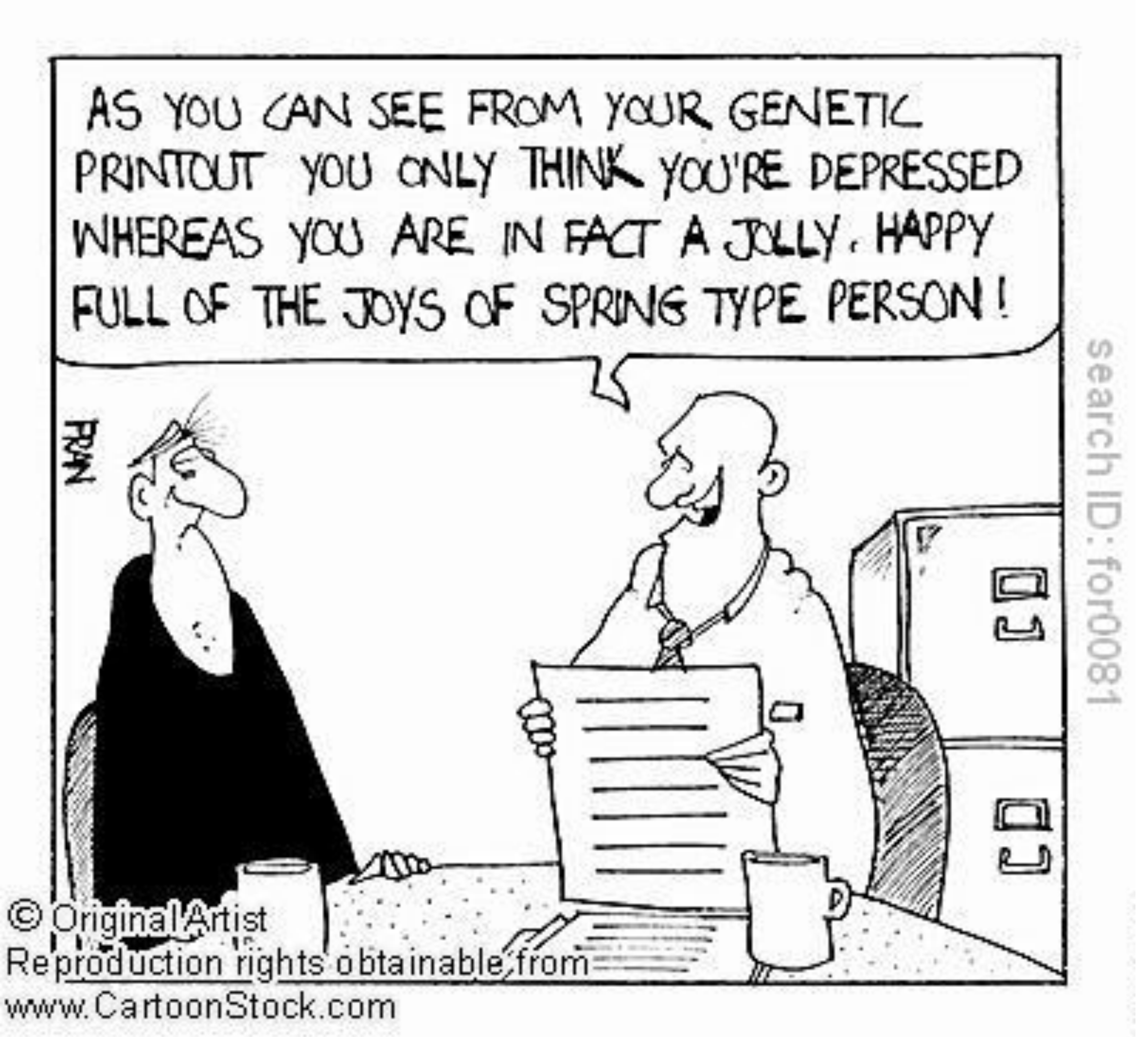




Picture from http://en.wikipedia.org/wiki/DNA

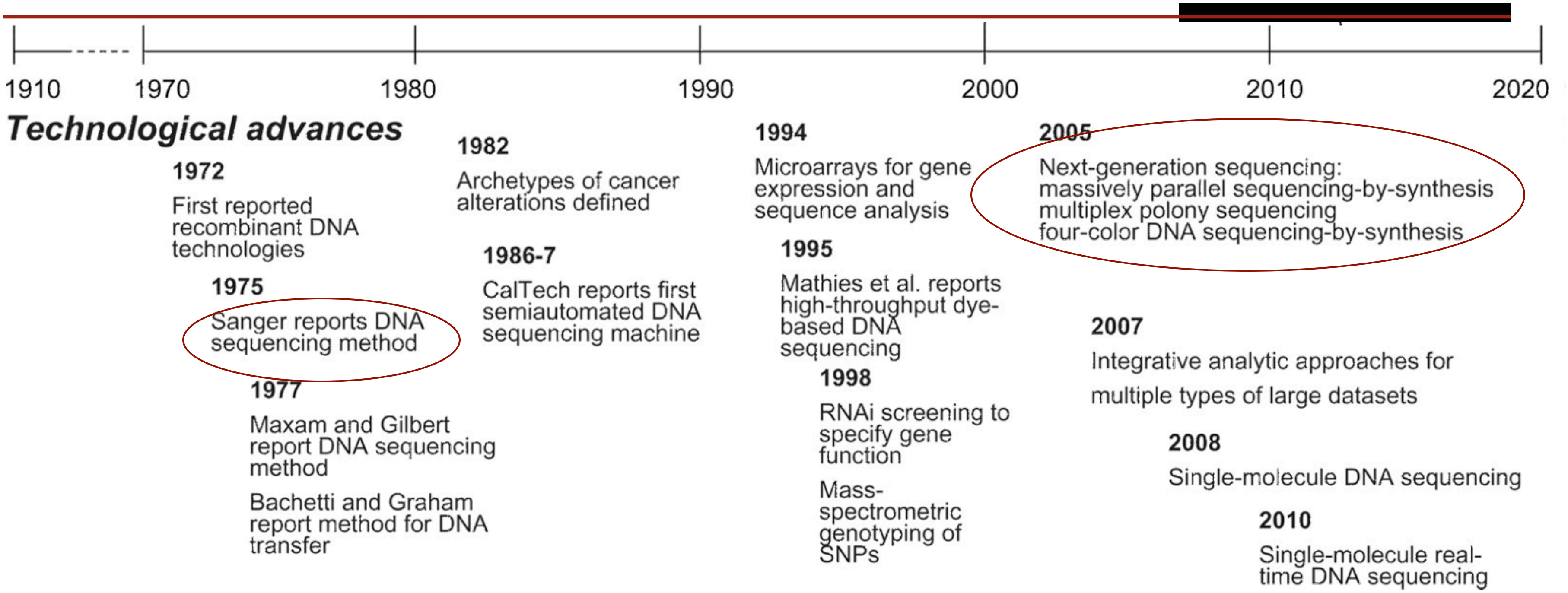
Why Sequencing?

- Medicine
- Forensics
- Biology
- Agriculture



© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
search ID: for0081

DNA Sequencing



Technological advances

1972

First reported recombinant DNA technologies

1975

Sanger reports DNA sequencing method

1977

Maxam and Gilbert report DNA sequencing method

Bachetti and Graham report method for DNA transfer

1982

Archetypes of cancer alterations defined

1986-7

CalTech reports first semiautomated DNA sequencing machine

1994

Microarrays for gene expression and sequence analysis

1995

Mathies et al. reports high-throughput dye-based DNA sequencing

1998

RNAi screening to specify gene function

Mass-spectrometric genotyping of SNPs

2005

Next-generation sequencing:
massively parallel sequencing-by-synthesis
multiplex polony sequencing
four-color DNA sequencing-by-synthesis

2007

Integrative analytic approaches for multiple types of large datasets

2008

Single-molecule DNA sequencing

2010

Single-molecule real-time DNA sequencing

cancerdiscovery.aacrjournals.org

Sanger sequencing

- DNA Sequencing = determining the order of the nucleotide bases
- single-stranded DNA template
- DNA primer
- DNA polymerase
- Normal dNTPs
- Terminating nucleotide

[Sanger Video]

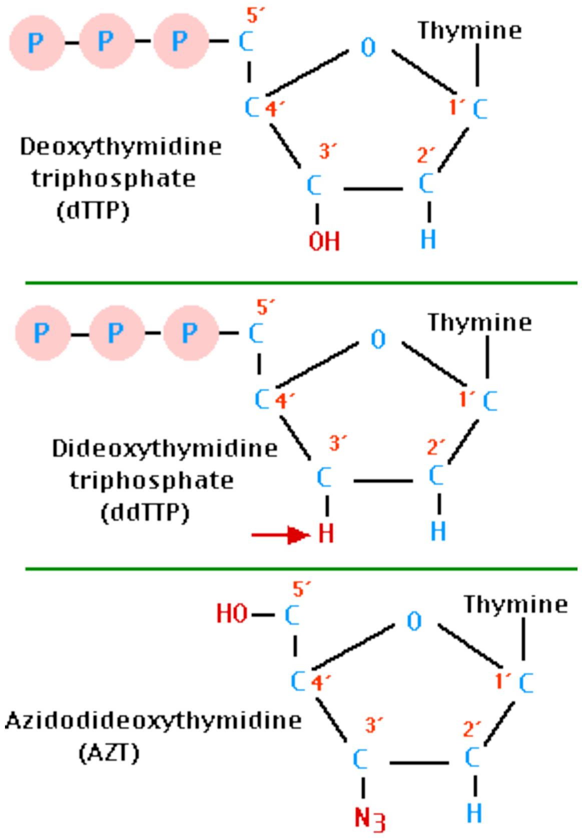


Image from http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/DNAsequencing.html

Sanger sequencing overview

- genomic DNA is fragmented
- cloned to a plasmid vector -> transform

E. coli

- a single bacterial colony is picked ->

plasmid DNA isolated

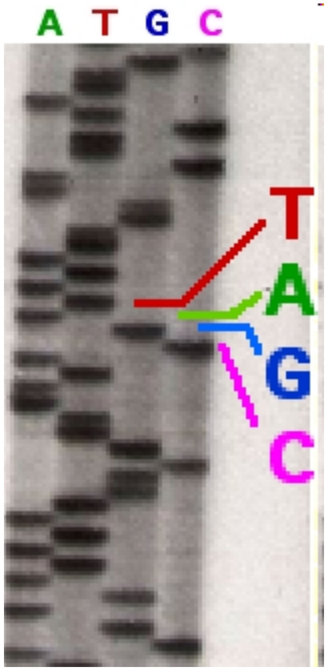
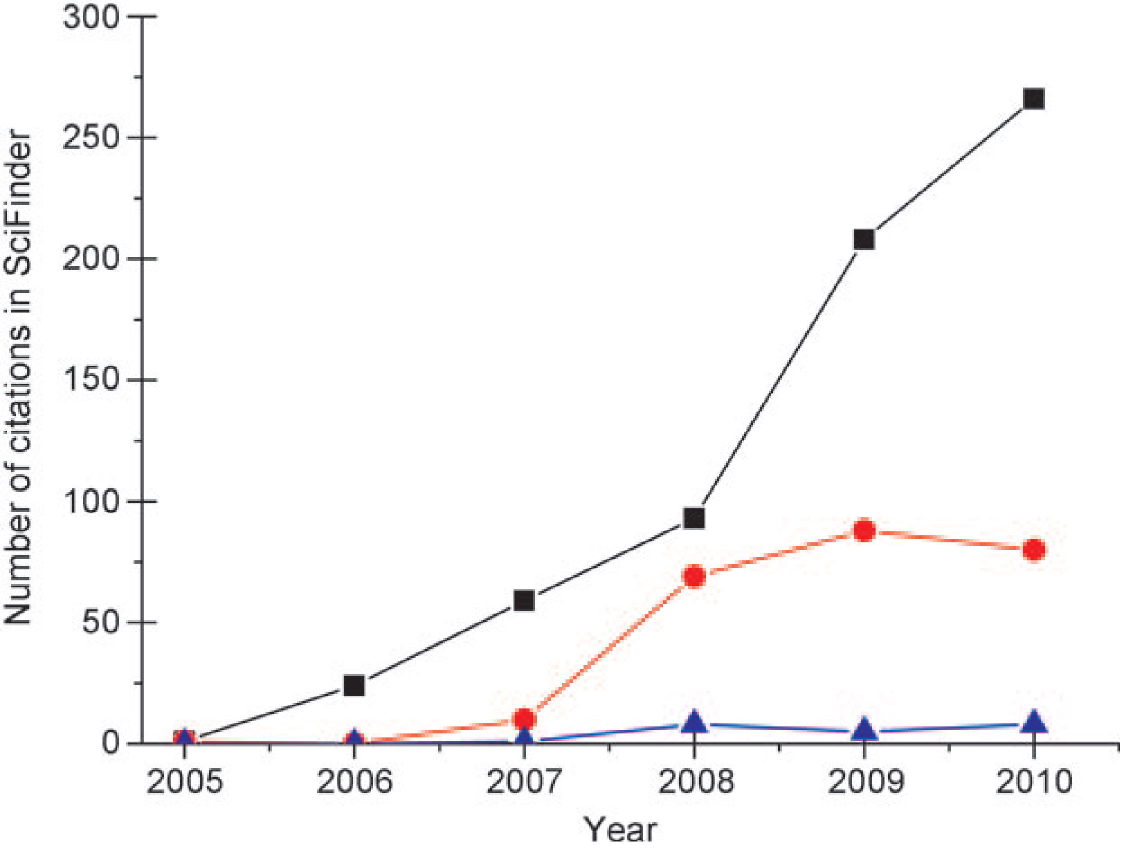


Image from http://en.wikipedia.org/wiki/DNA_sequencing

Sequencing technologies – Sequencing Revolution



DOI: 10.1002/anie.201003880

Improved technologies:

- Higher throughput 1500 x

- Reduced costs / Mb

- Common method: sequencing by extension

NGS – What Platforms are there?

- Illumina/Solexa reversible terminator chemistry
- Principle of SOLiD sequencing by ligation
- 454 Pyrosequencing
- Ion Torrent Personal genome Machine
- Single Molecule Sequencing

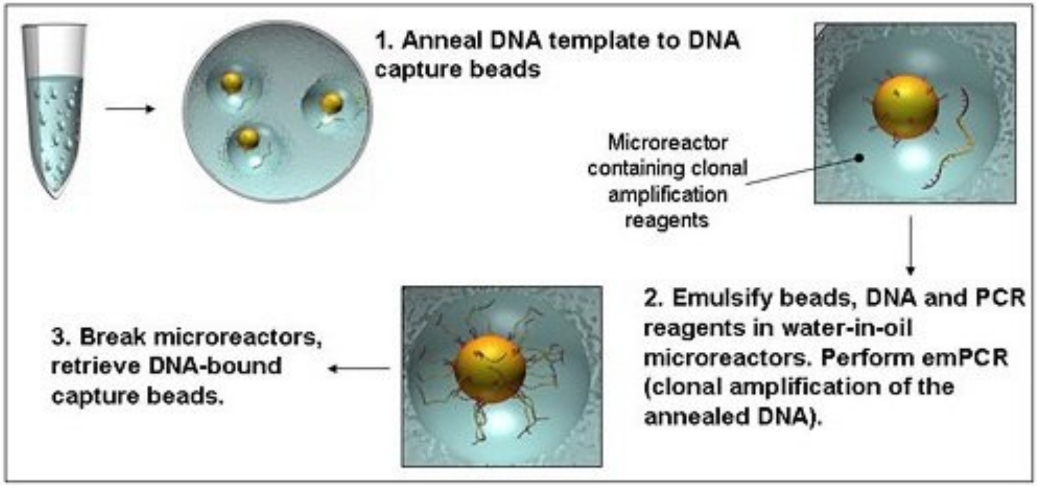
Sequencing technologies – shared attributes

- Template preparation
- Sequencing and imaging
- Data analysis

Sequencing technologies – NGS template preparation

A. Clonally amplified templates - cell free system:

Emulsion PCR [Emulsion PCR Video]



- standard microscope slide (Polonator)
- aminocoated glass surface (Life/APG; Polonator)
- PicoTiterPlate (PTP) wells (Roche/454)
- microchip sensor (Ion Torrent)

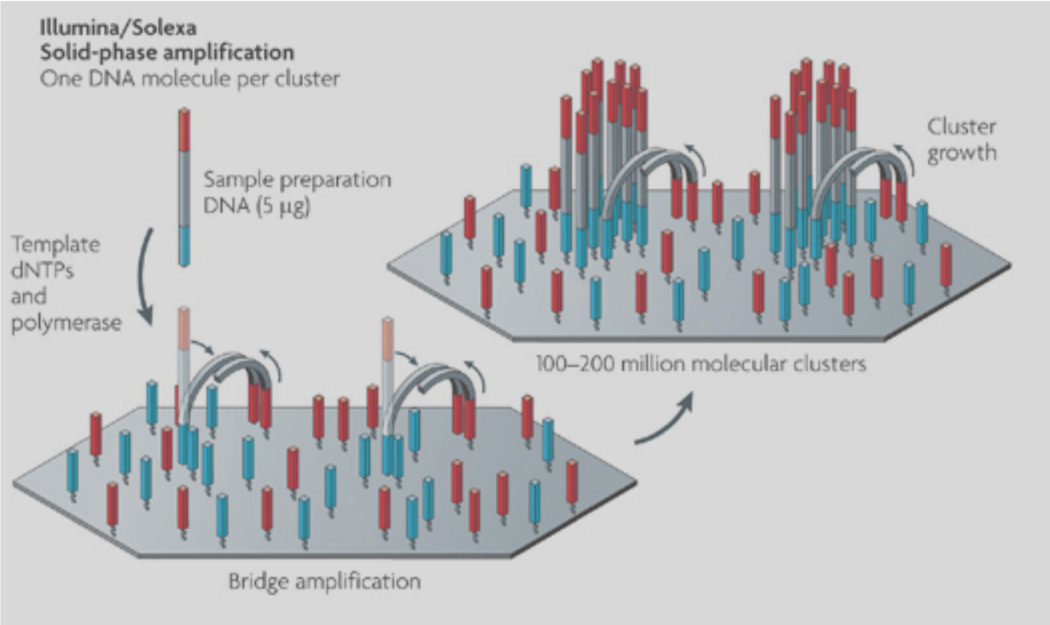
Metzker, M. L. Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Picture on http://www.seqtech.com/2011/11/08/454-life-sciences-2/

Sequencing technologies – NGS template preparation

A. Clonally amplified templates - cell free system:

Solid-phase amplification Bridge PCR Video



Metzker, M. L. Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Sequencing technologies – NGS template preparation

B. Single-molecule templates:

- Require less starting material
- Immobilized on the solid surface by

Primers: Helicos BioSciences

Template: Helicos BioSciences

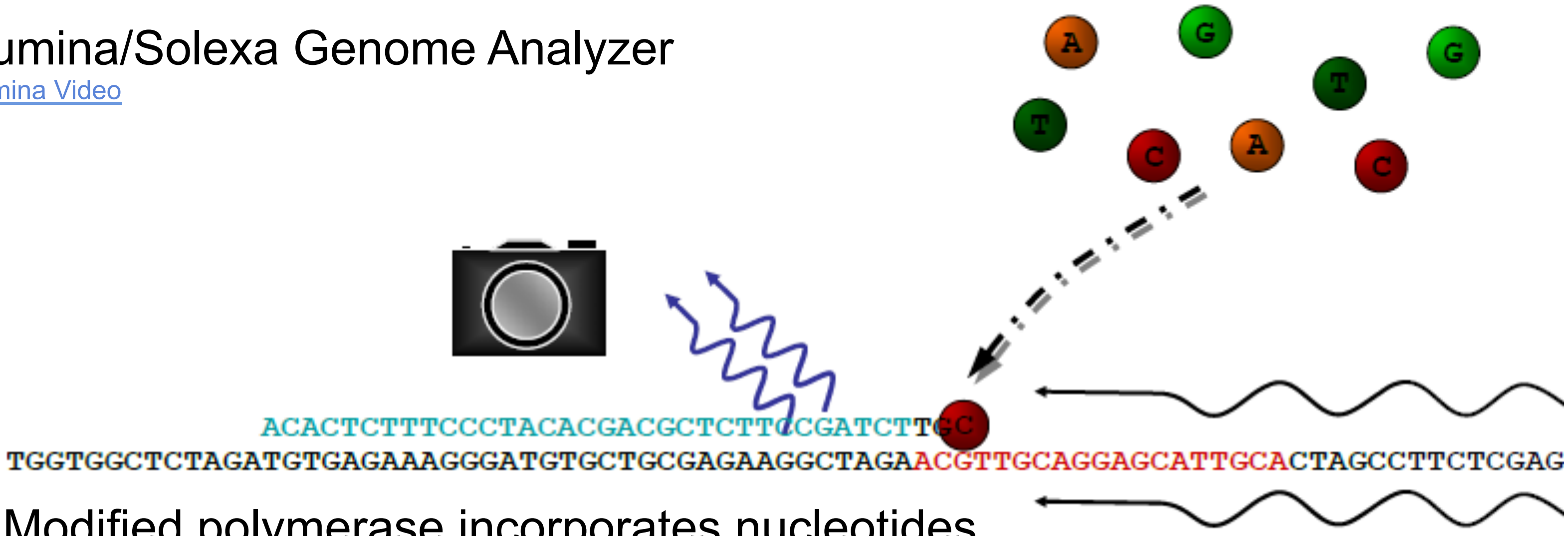
Polymerase: Pacific Biosciences, Life/Visigen, LI-COR Biosciences

Sequencing technologies – NGS sequencing and imaging

1. Cyclic reversible termination

Illumina/Solexa Genome Analyzer

[Illumina Video]

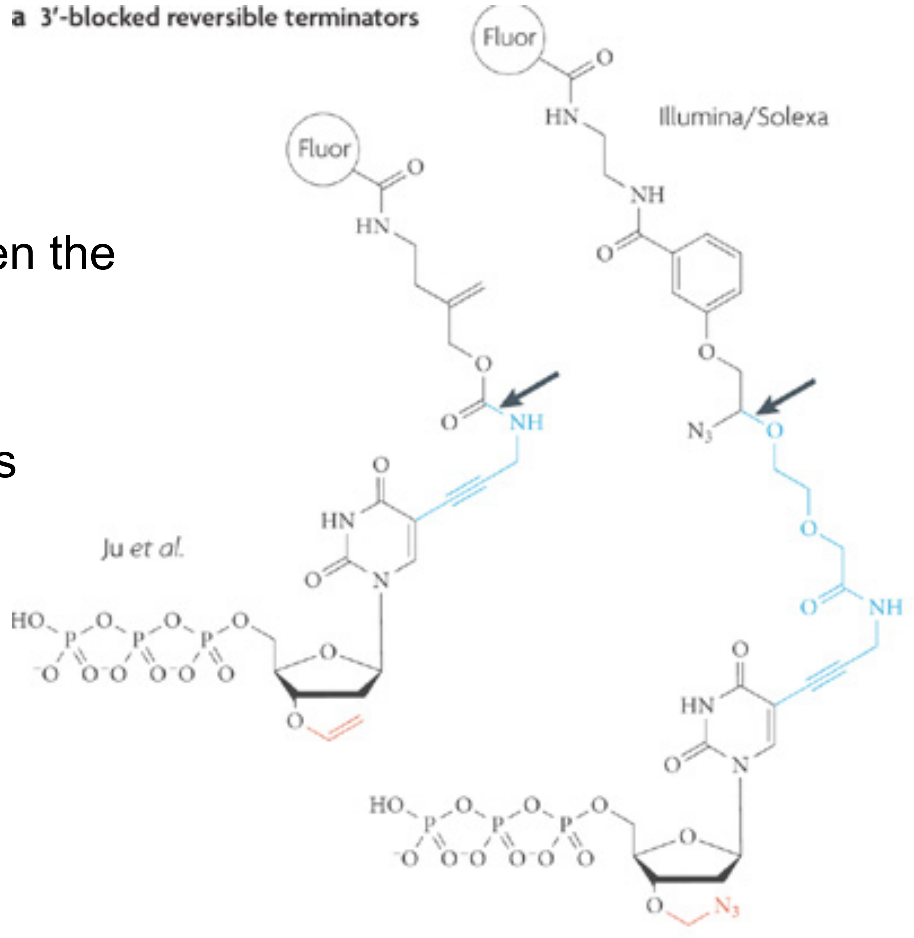


Modified polymerase incorporates nucleotides

- after each nucleotide incorporation process stops
- camera reads fluorophore signal (filter for each nucleotide type)
- terminator and labeling is removed and cycle starts again

Sequencing technologies – NGS sequencing and imaging

1. Cyclic reversible termination



- Substitutions with higher frequency when the previous base is ‘G’
- Underrepresentation of GC- rich regions

Metzker, M. L. Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Sequencing technologies – NGS sequencing and imaging

1. Cyclic reversible termination

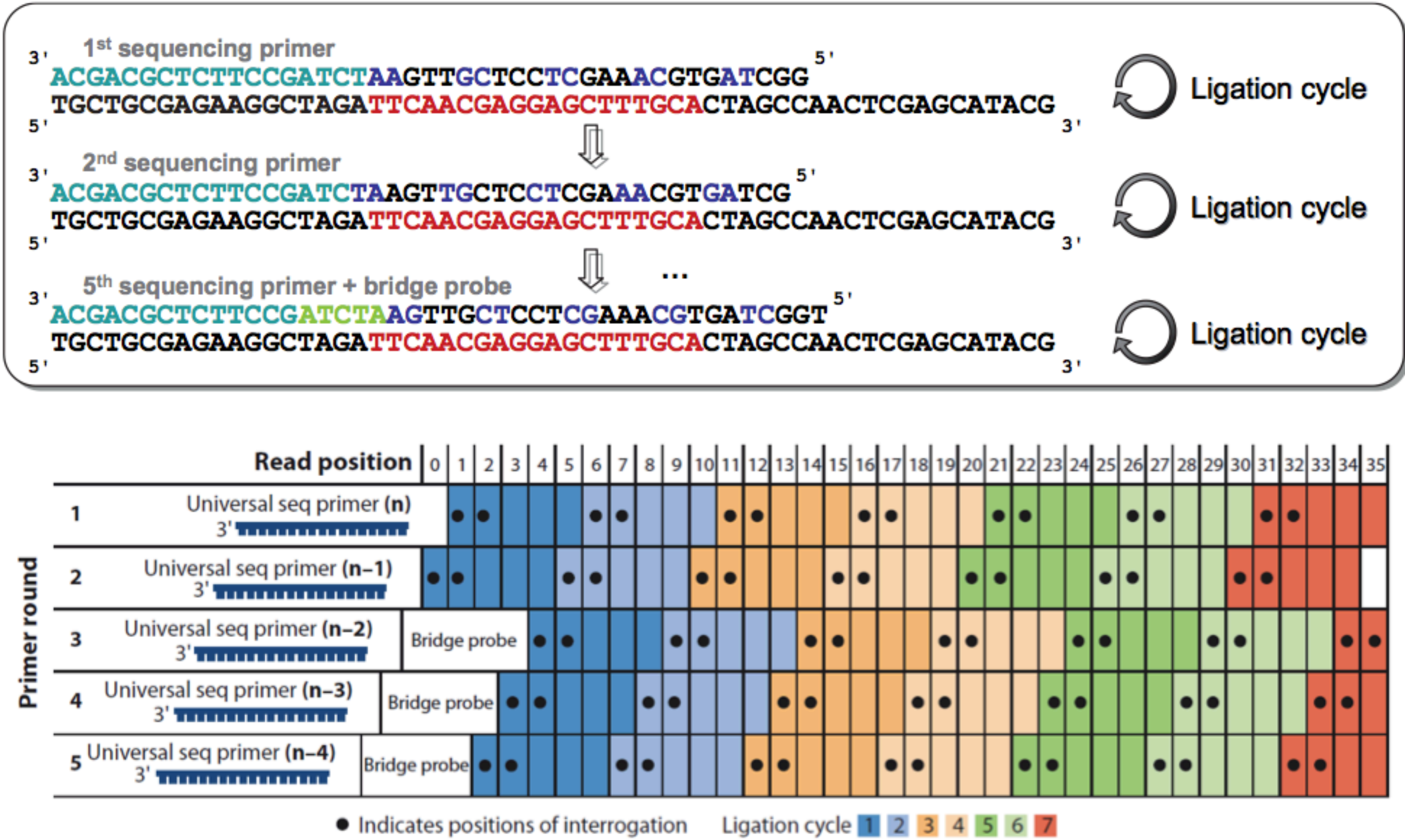
    - 3’-unblocked reversible terminators
    - LaserGen – Lightning Terminators
    - Helicos BioSciences – Virtual Terminators
    - Cleavage of only one bond

Sequencing technologies – NGS sequencing and imaging

2. Sequencing by ligation

- Difference – DNA ligase
- Hybridization of a fluorescently labelled probe
- SOLiD cycle of 1,2-probe hybridization

[SOLiD Video]



3' **1st sequencing primer** 5'
ACGACGCTCTTCCGATCTAAGTTGCTCCTCGAAACGTGATCGG
TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCAACTCGAGCATACG
5' 3'

Ligation cycle

3' **2nd sequencing primer** 5'
ACGACGCTCTTCCGATCTAAGTTGCTCCTCGAAACGTGATCG
TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCAACTCGAGCATACG
5' 3'

Ligation cycle

...

3' **5th sequencing primer + bridge probe** 5'
ACGACGCTCTTCCGATCTAAGTTGCTCCTCGAAACGTGATCGGT
TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCAACTCGAGCATACG
5' 3'

Ligation cycle

Primer round	Read position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1	Universal seq primer (n) 3'		1 ●	1 ●	1	1	1	2 ●	2 ●	2	2	2	3 ●	3 ●	3	3	3	4 ●	4 ●	4	4	4	5 ●	5 ●	5	5	5	6 ●	6 ●	6	6	6	7 ●	7 ●	7	7	7
2	Universal seq primer (n–1) 3'	1 ●	1 ●	1	1	1	2 ●	2 ●	2	2	2	3 ●	3 ●	3	3	3	4 ●	4 ●	4	4	4	5 ●	5 ●	5	5	5	6 ●	6 ●	6	6	6	7 ●	7 ●	7	7	7	
3	Universal seq primer (n–2) 3'	Bridge probe				1 ●	1 ●	1	1	1	2 ●	2 ●	2	2	2	3 ●	3 ●	3	3	3	4 ●	4 ●	4	4	4	5 ●	5 ●	5	5	5	6 ●	6 ●	6	6	6	7 ●	7 ●
4	Universal seq primer (n–3) 3'	Bridge probe			1 ●	1 ●	1	1	1	2 ●	2 ●	2	2	2	3 ●	3 ●	3	3	3	4 ●	4 ●	4	4	4	5 ●	5 ●	5	5	5	6 ●	6 ●	6	6	6	7 ●	7 ●	7
5	Universal seq primer (n–4) 3'	Bridge probe		1 ●	1 ●	1	1	1	2 ●	2 ●	2	2	2	3 ●	3 ●	3	3	3	4 ●	4 ●	4	4	4	5 ●	5 ●	5	5	5	6 ●	6 ●	6	6	6	7 ●	7 ●	7	7

● Indicates positions of interrogation Ligation cycle 1 2 3 4 5 6 7

Picture by ABI/Life Technologies

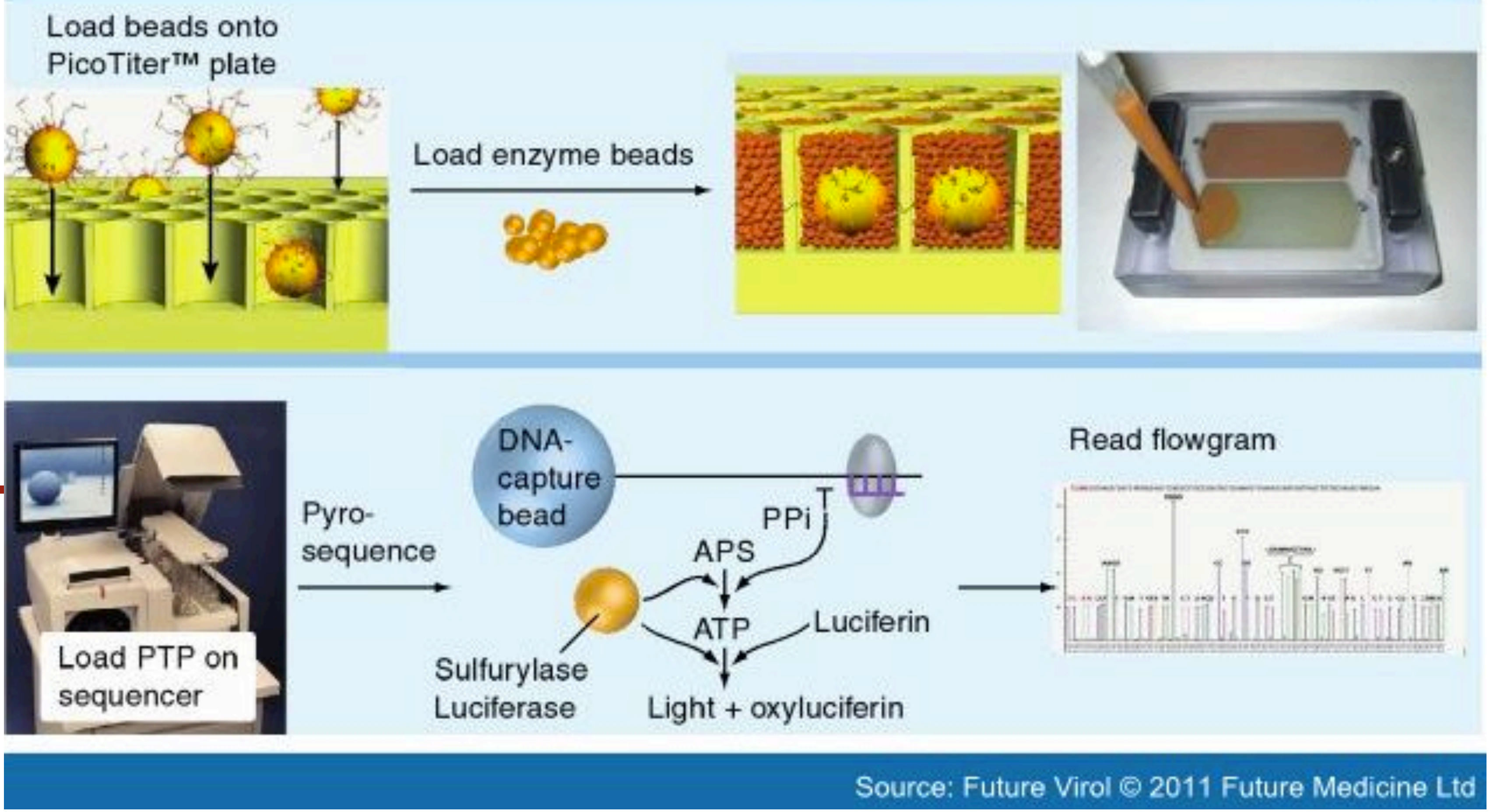
Sequencing technologies – NGS sequencing and imaging

2. Sequencing by ligation errors:

   - Substitutions
   - Underrepresentation of AT- and GC- rich regions

Sequencing technologies – NGS sequencing and imaging

3. Pyrosequencing [454 Video]



Source: Future Virol © 2011 Future Medicine Ltd

Sequencing technologies – NGS sequencing and imaging

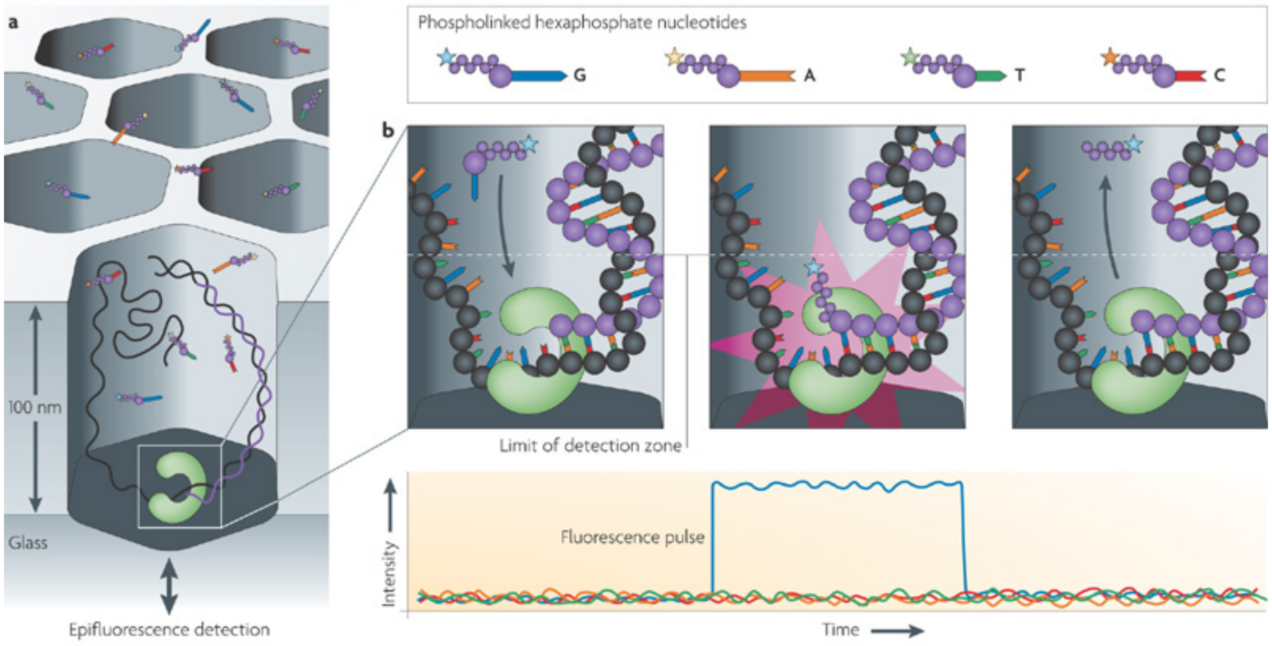
3. Pyrosequencing errors:

- For homopolymeric reads -> unreliable sequence
- Insertions
- Deletions

Sequencing technologies – NGS sequencing and imaging

4. Real-time sequencing:
   - Pacific Biosciences
   - Continuous imaging of dye-labelled nucleotides incorporation

Pacific Biosciences — Real-time sequencing

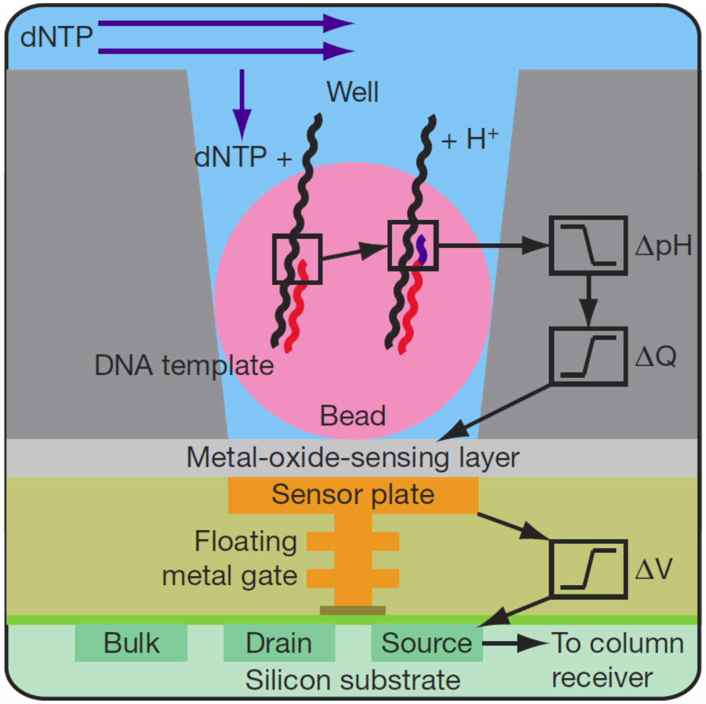


Nature Reviews | Genetics

Sequencing technologies – NGS sequencing and imaging

5. Ion Semiconductor Sequencing [Ion Torrent Video]

- incorporation of dNTP into DNA strand -> release of H^+
- ΔpH detected by an ion-sensitive field-effect transistor

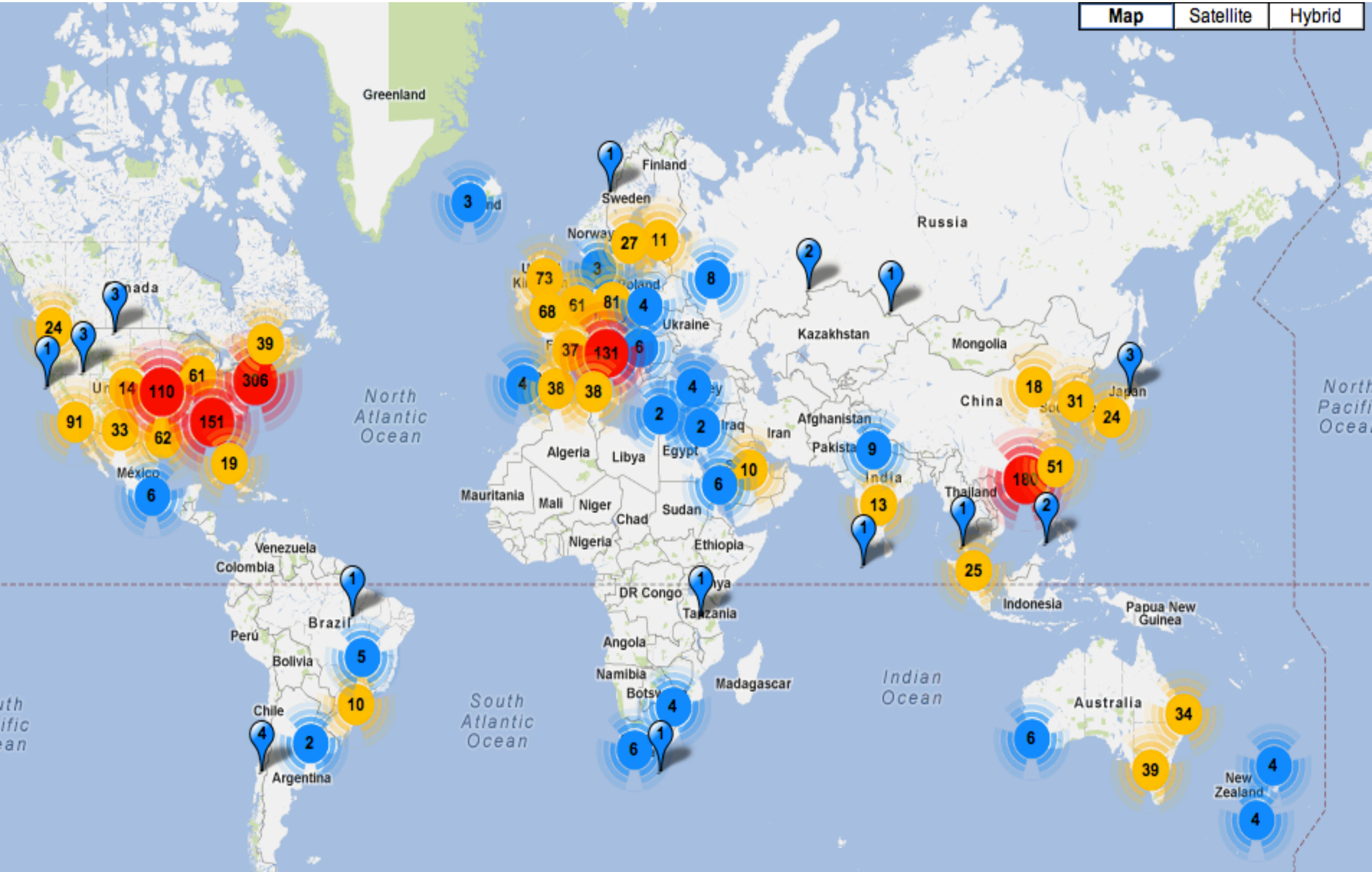


Comparison of different NGS platforms

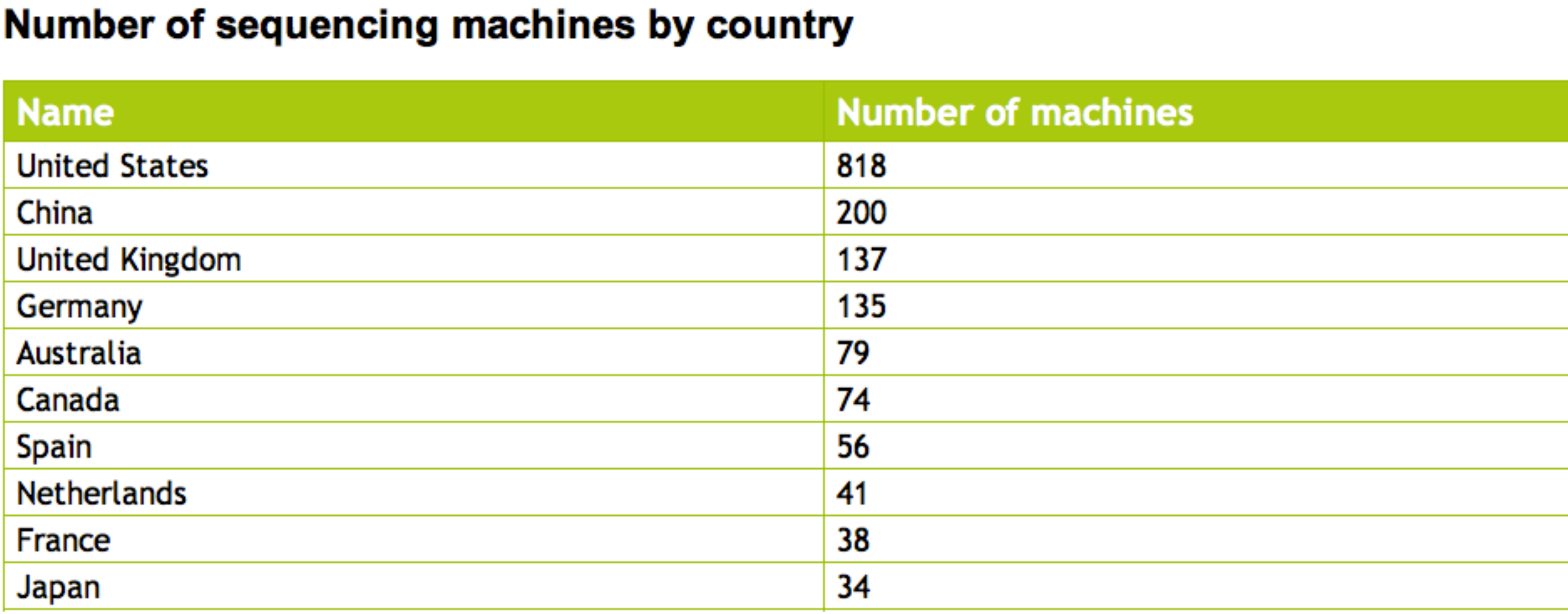
	Throughput	Length	Quality	Costs
Sanger	6 Mb/day	800nt	10^{-4} - 10^{-5}	500$/Mb
454	750Mb/day	400nt	10^{-3} - 10^{-4}	~20$/Mb
Ion Torrent	1600Mb/day	200nt	10^{-2} - 10^{-3}	~10$/Mb
Illumina	100000Mb/day	125nt	10^{-2} - 10^{-3}	~0.40$/Mb
SOLiD 4	100000Mb/day	125nt	10^{-2} - 10^{-3}	~0.40$/Mb
Helicos	5000Mb/day	32nt	10^{-2}	~0.40$/Mb

Adapted from IMPRS EVA Genetics Core Seminar Week – Janet Kelso, Martin Kircher

Sequencing around the World



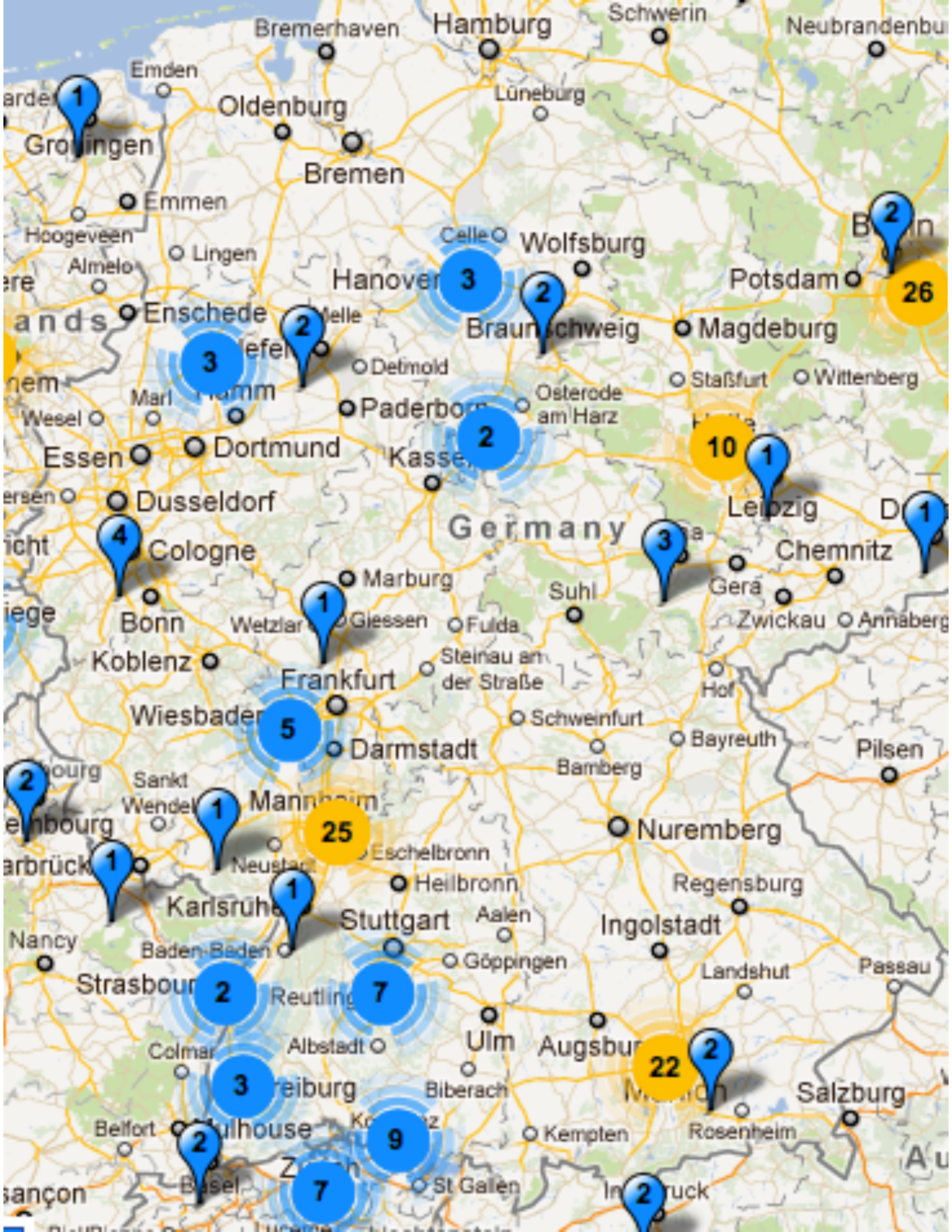
http://www.omicsmaps.com/



Number of sequencing machines by country

Name	Number of machines
United States	818
China	200
United Kingdom	137
Germany	135
Australia	79
Canada	74
Spain	56
Netherlands	41
France	38
Japan	34

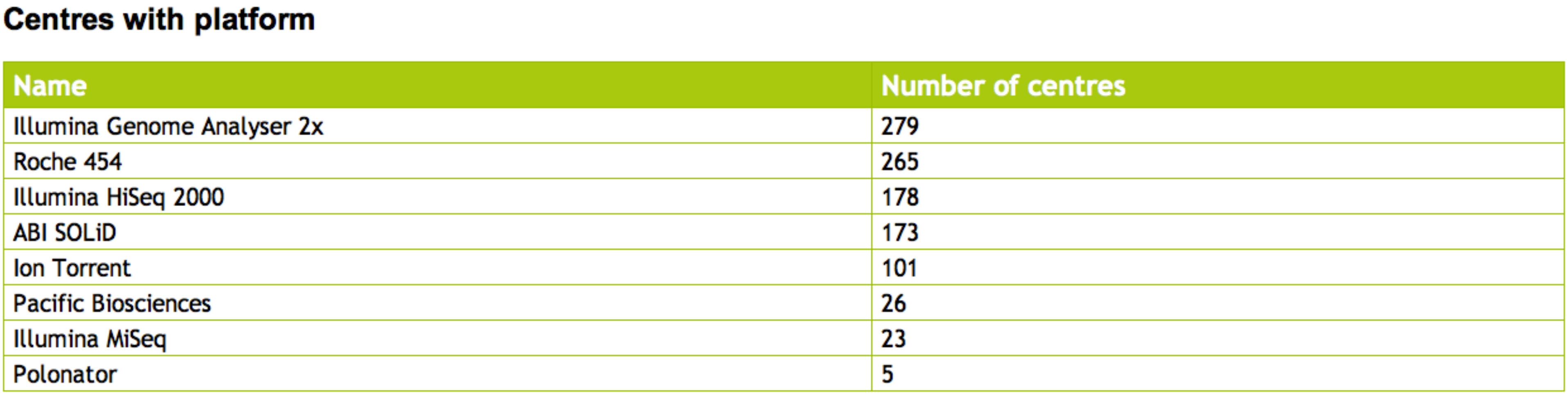
omicsmaps.com/**stats**



Leipzig

- 10 Sequencing Machines,
- 4th place in Germany

http://www.omicsmaps.com/



Centres with platform

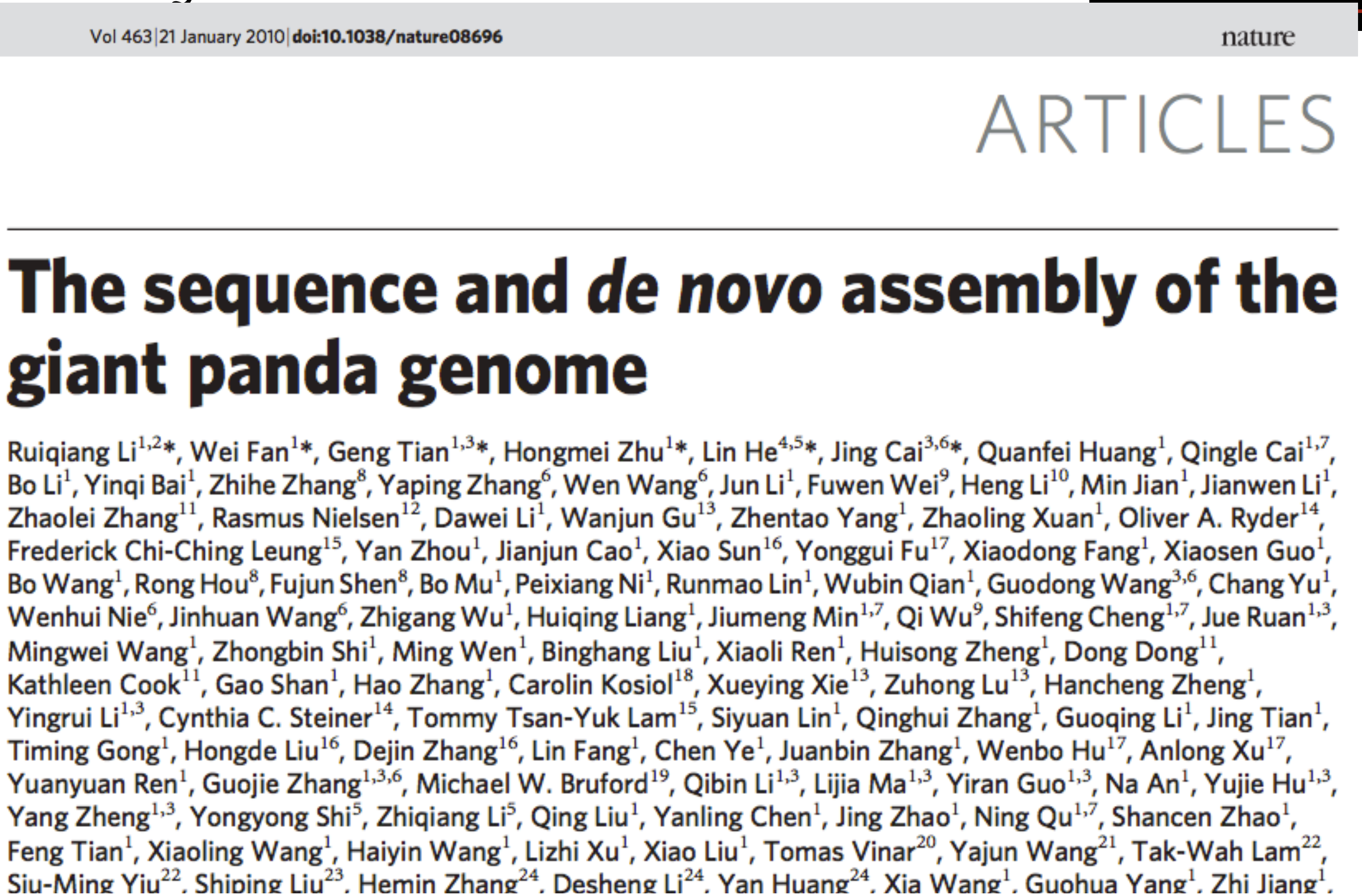
Name	Number of centres
Illumina Genome Analyser 2x	279
Roche 454	265
Illumina HiSeq 2000	178
ABI SOLiD	173
Ion Torrent	101
Pacific Biosciences	26
Illumina MiSeq	23
Polonator	5

omicsmaps.com/**stats**

Sequencing technologies – Data analysis

- Bioinformatics tools for:
  - Alignment
  - Base calling/polymorphism detection
  - *De novo* assembly
  - Genome browsing or annotation
- Challenging problems:
  - *De novo* assembly of short reads -> mate-paired libraries required
  - Reads in repetitive regions

Sequencing technologies – Data analysis



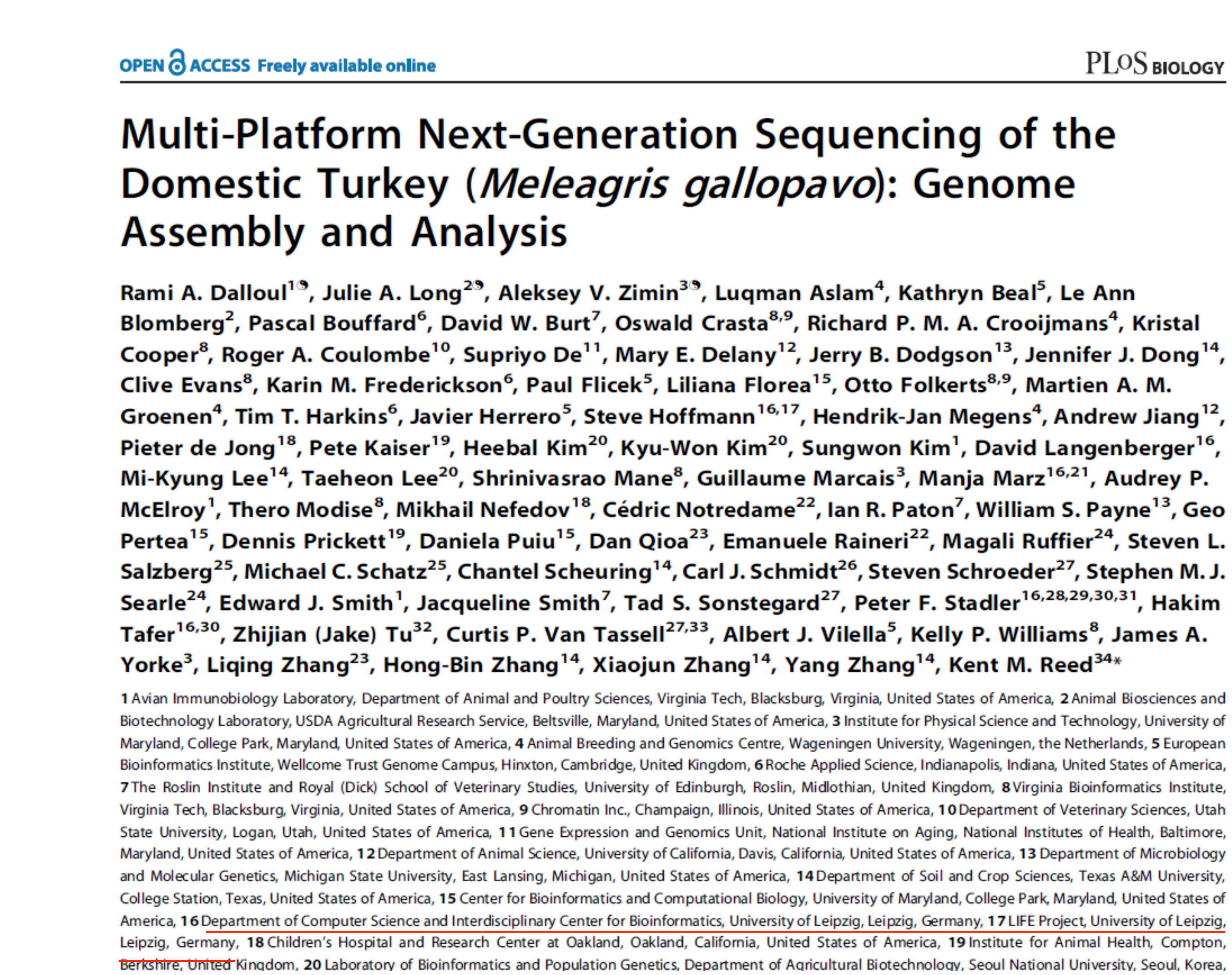
Vol 463|21 January 2010|doi:10.1038/nature08696 nature

ARTICLES

The sequence and *de novo* assembly of the giant panda genome

Ruiqiang Li[1,2]*, Wei Fan[1]*, Geng Tian[1,3]*, Hongmei Zhu[1]*, Lin He[4,5]*, Jing Cai[3,6]*, Quanfei Huang[1], Qingle Cai[1,7], Bo Li[1], Yinqi Bai[1], Zhihe Zhang[8], Yaping Zhang[6], Wen Wang[6], Jun Li[1], Fuwen Wei[9], Heng Li[10], Min Jian[1], Jianwen Li[1], Zhaolei Zhang[11], Rasmus Nielsen[12], Dawei Li[1], Wanjun Gu[13], Zhentao Yang[1], Zhaoling Xuan[1], Oliver A. Ryder[14], Frederick Chi-Ching Leung[15], Yan Zhou[1], Jianjun Cao[1], Xiao Sun[16], Yonggui Fu[17], Xiaodong Fang[1], Xiaosen Guo[1], Bo Wang[1], Rong Hou[8], Fujun Shen[8], Bo Mu[1], Peixiang Ni[1], Runmao Lin[1], Wubin Qian[1], Guodong Wang[3,6], Chang Yu[1], Wenhui Nie[6], Jinhuan Wang[6], Zhigang Wu[1], Huiqing Liang[1], Jiumeng Min[1,7], Qi Wu[9], Shifeng Cheng[1,7], Jue Ruan[1,3], Mingwei Wang[1], Zhongbin Shi[1], Ming Wen[1], Binghang Liu[1], Xiaoli Ren[1], Huisong Zheng[1], Dong Dong[11], Kathleen Cook[11], Gao Shan[1], Hao Zhang[1], Carolin Kosiol[18], Xueying Xie[13], Zuhong Lu[13], Hancheng Zheng[1], Yingrui Li[1,3], Cynthia C. Steiner[14], Tommy Tsan-Yuk Lam[15], Siyuan Lin[1], Qinghui Zhang[1], Guoqing Li[1], Jing Tian[1], Timing Gong[1], Hongde Liu[16], Dejin Zhang[16], Lin Fang[1], Chen Ye[1], Juanbin Zhang[1], Wenbo Hu[17], Anlong Xu[17], Yuanyuan Ren[1], Guojie Zhang[1,3,6], Michael W. Bruford[19], Qibin Li[1,3], Lijia Ma[1,3], Yiran Guo[1,3], Na An[1], Yujie Hu[1,3], Yang Zheng[1,3], Yongyong Shi[5], Zhiqiang Li[5], Qing Liu[1], Yanling Chen[1], Jing Zhao[1], Ning Qu[1,7], Shancen Zhao[1], Feng Tian[1], Xiaoling Wang[1], Haiyin Wang[1], Lizhi Xu[1], Xiao Liu[1], Tomas Vinar[20], Yajun Wang[21], Tak-Wah Lam[22], Siu-Ming Yiu[22], Shiping Liu[23], Hemin Zhang[24], Desheng Li[24], Yan Huang[24], Xia Wang[1], Guohua Yang[1], Zhi Jiang[1],

Sequencing technologies – Data analysis



OPEN ACCESS Freely available online

PLoS BIOLOGY

Multi-Platform Next-Generation Sequencing of the Domestic Turkey (*Meleagris gallopavo*): Genome Assembly and Analysis

Rami A. Dalloul[1¶], Julie A. Long[2¶], Aleksey V. Zimin[3¶], Luqman Aslam[4], Kathryn Beal[5], Le Ann Blomberg[2], Pascal Bouffard[6], David W. Burt[7], Oswald Crasta[8,9], Richard P. M. A. Crooijmans[4], Kristal Cooper[8], Roger A. Coulombe[10], Supriyo De[11], Mary E. Delany[12], Jerry B. Dodgson[13], Jennifer J. Dong[14], Clive Evans[8], Karin M. Frederickson[6], Paul Flicek[5], Liliana Florea[15], Otto Folkerts[8,9], Martien A. M. Groenen[4], Tim T. Harkins[6], Javier Herrero[5], Steve Hoffmann[16,17], Hendrik-Jan Megens[4], Andrew Jiang[12], Pieter de Jong[18], Pete Kaiser[19], Heebal Kim[20], Kyu-Won Kim[20], Sungwon Kim[1], David Langenberger[16], Mi-Kyung Lee[14], Taeheon Lee[20], Shrinivasrao Mane[8], Guillaume Marcais[3], Manja Marz[16,21], Audrey P. McElroy[1], Thero Modise[8], Mikhail Nefedov[18], Cédric Notredame[22], Ian R. Paton[7], William S. Payne[13], Geo Pertea[15], Dennis Prickett[19], Daniela Puiu[15], Dan Qioa[23], Emanuele Raineri[22], Magali Ruffier[24], Steven L. Salzberg[25], Michael C. Schatz[25], Chantel Scheuring[14], Carl J. Schmidt[26], Steven Schroeder[27], Stephen M. J. Searle[24], Edward J. Smith[1], Jacqueline Smith[7], Tad S. Sonstegard[27], Peter F. Stadler[16,28,29,30,31], Hakim Tafer[16,30], Zhijian (Jake) Tu[32], Curtis P. Van Tassell[27,33], Albert J. Vilella[5], Kelly P. Williams[8], James A. Yorke[3], Liqing Zhang[23], Hong-Bin Zhang[14], Xiaojun Zhang[14], Yang Zhang[14], Kent M. Reed[34*]

1 Avian Immunobiology Laboratory, Department of Animal and Poultry Sciences, Virginia Tech, Blacksburg, Virginia, United States of America, **2** Animal Biosciences and Biotechnology Laboratory, USDA Agricultural Research Service, Beltsville, Maryland, United States of America, **3** Institute for Physical Science and Technology, University of Maryland, College Park, Maryland, United States of America, **4** Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands, **5** European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **6** Roche Applied Science, Indianapolis, Indiana, United States of America, **7** The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Midlothian, United Kingdom, **8** Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, United States of America, **9** Chromatin Inc., Champaign, Illinois, United States of America, **10** Department of Veterinary Sciences, Utah State University, Logan, Utah, United States of America, **11** Gene Expression and Genomics Unit, National Institute on Aging, National Institutes of Health, Baltimore, Maryland, United States of America, **12** Department of Animal Science, University of California, Davis, California, United States of America, **13** Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, United States of America, **14** Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas, United States of America, **15** Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America, **16** Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany, **17** LIFE Project, University of Leipzig, Leipzig, Germany, **18** Children's Hospital and Research Center at Oakland, Oakland, California, United States of America, **19** Institute for Animal Health, Compton, Berkshire, United Kingdom, **20** Laboratory of Bioinformatics and Population Genetics, Department of Agricultural Biotechnology, Seoul National University, Seoul, Korea,

Sequencing technologies – Data analysis

- $ 1000 genome sequencing and
- $ 1000000 data analysis

NGS applications

- Genome resequencing: polymorphism and mutation discovery in humans (1000 Genomes Project)
- “Omics”: transcriptomics, proteomics, metabolomics, microbiomes

NGS applications

- Transcriptome sequencing:
  - Gene expression
  - Alternative splicing
  - Transcript annotation
  - SNPs
  - Somatic mutations

NGS applications Future

- Throughput and costs of sequencing will allow to characterize genetic variation within and between species in great detail
- Will become routine
- Greatest challenge is extracting biologically or clinically meaningful information

My Projects

1. Kiwi sequencing Illumina HiScan 2

2. Transcriptome analysis and comparison GPCR

   34 knock out – wild type C57BL/6

Kiwi

Goals:

1. Assessment of wing development genes:
    - Mutations
    - Signatures of selection
    - Functional assessment
2. G protein coupled receptors

Ensembl Gene ID	Associated Gene Name
ENSGALG00000001532	F1NPH2_CHICK
ENSGALG00000006379	SHH
ENSGALG00000007562	FGF4
ENSGALG00000007706	Q90696_CHICK
ENSGALG00000007834	SALL4
ENSGALG00000008253	TBX5_CHICK
ENSGALG00000009495	FGFR2
ENSGALG00000010863	TWISTNB
ENSGALG00000011630	GLI2
ENSGALG00000012329	GLI3
ENSGALG00000014872	FGF10
ENSGALG00000023904	FIBIN

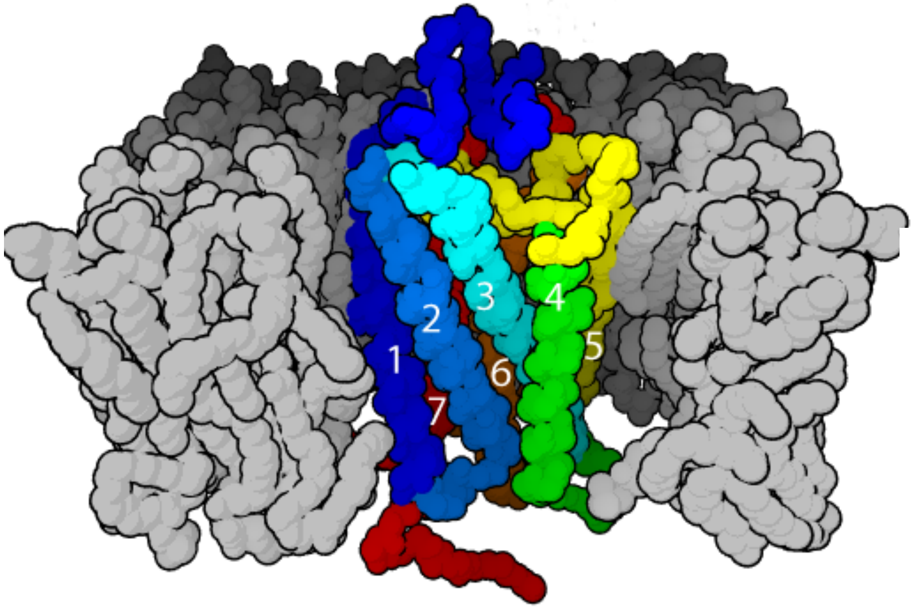
Kiwi

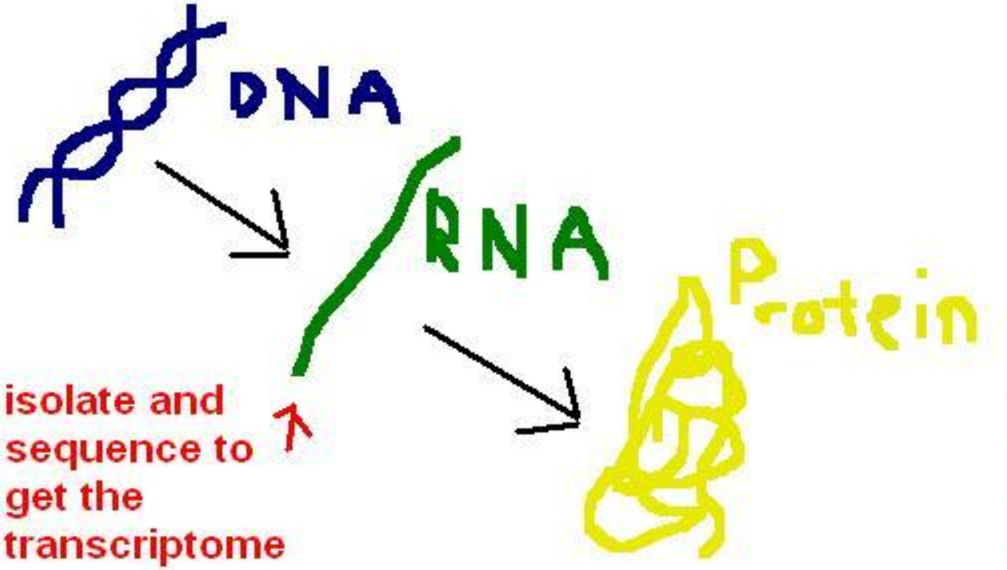
Further goals:

3. Phylogeny tree
4. Genome assembly

Scientific Partners:

- BGI-G10K: Prof. Guojie Zhang
- MPI EVA: Bioinformatics group Janet Kelso
- Allan Wilson Centre for Molecular Ecology and Evolution, School of Biological Sciences, University of Auckland, Auckland, New Zealand: Prof. David Lambert





isolate and sequence to get the transcriptome

G Protein Coupled Receptor Image from fossilmuseum.net

Image from http://labrat.fieldofscience.com

C57/Bl6 mouse image from http://www.criver.com

Transcriptome analysis

Goals:

- Differences in gene expression KO vs. WT
- Involved metabolic pathways
- Assess genes with immunologic involvement

Thank you!