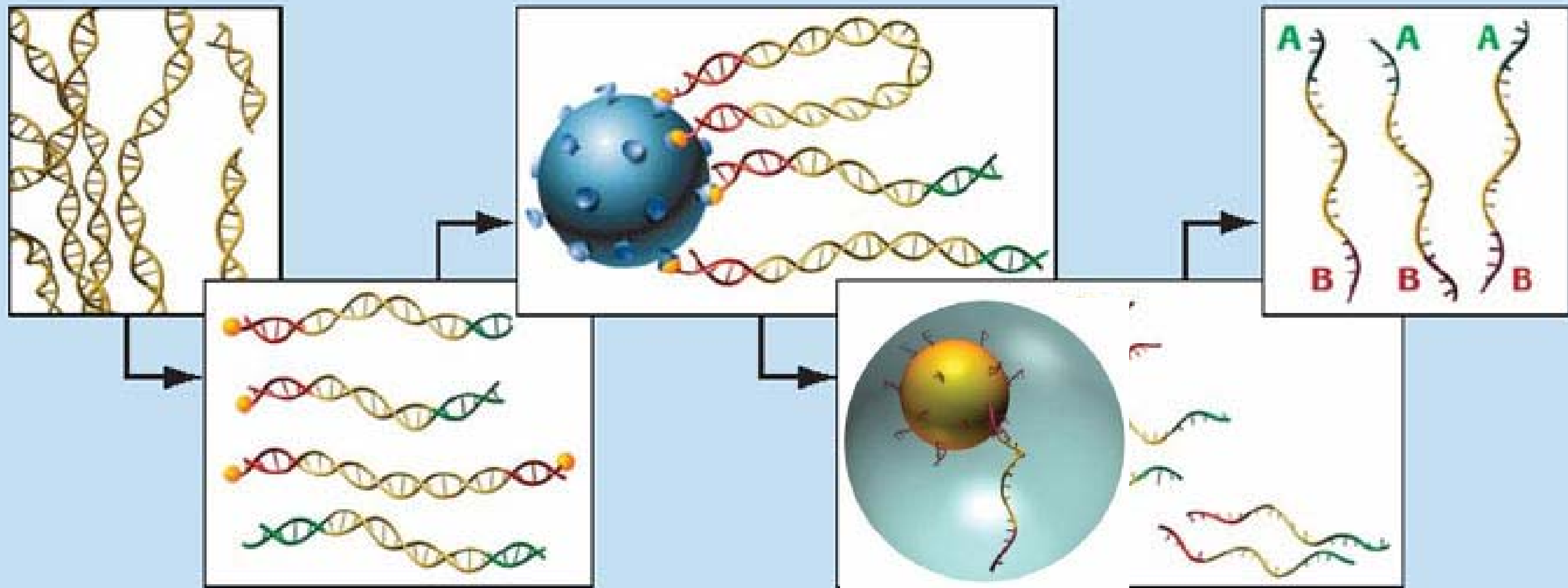
Einführung in die Gentechnologie
Erwin R. Schmidt

- Vorlesung # 11
- 24. 06. 2008

Massensequenzierung mit Roche 454/FLX Sequencer



DNA Library Preparation and Titration

4.5 hours

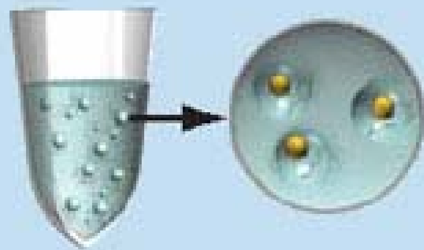
10.5 hours

emPCR

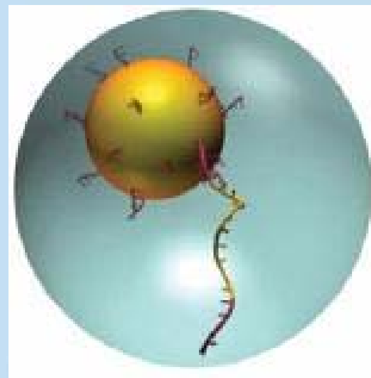
8 hours

Sequencing

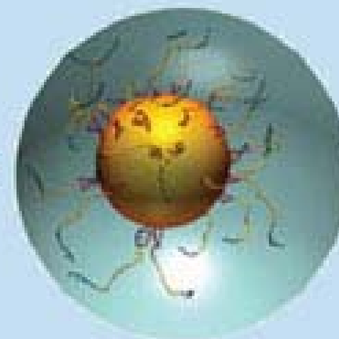
5.5 hours



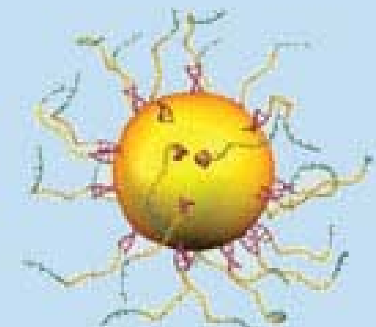
Anneal sstDNA to an excess of DNA Capture Beads



Emulsify beads and PCR reagents in water-in-oil microreactors



Clonal amplification occurs inside microreactors



Break microreactors, enrich for DNA-positive beads

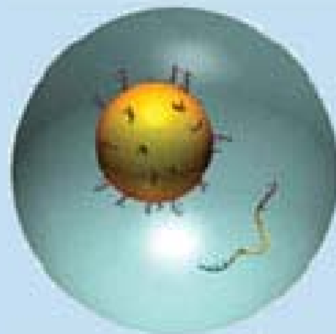
sstDNA library → **Clonally-amplified sstDNA attached to bead (millions of copies per bead)**

DNA Library Preparation and Titration — 4.5 hours … 10.5 hours

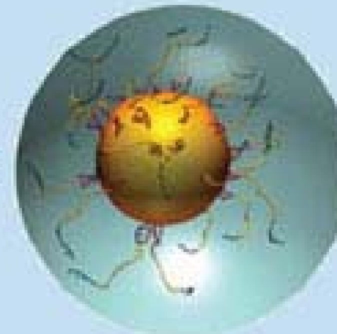
emPCR — 8 hours

Sequencing — 5.5 hours

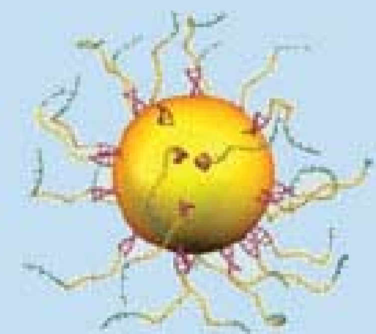
Anneal sstDNA to an excess of DNA Capture Beads



Emulsify beads and PCR reagents in water-in-oil microreactors



Clonal amplification occurs inside microreactors

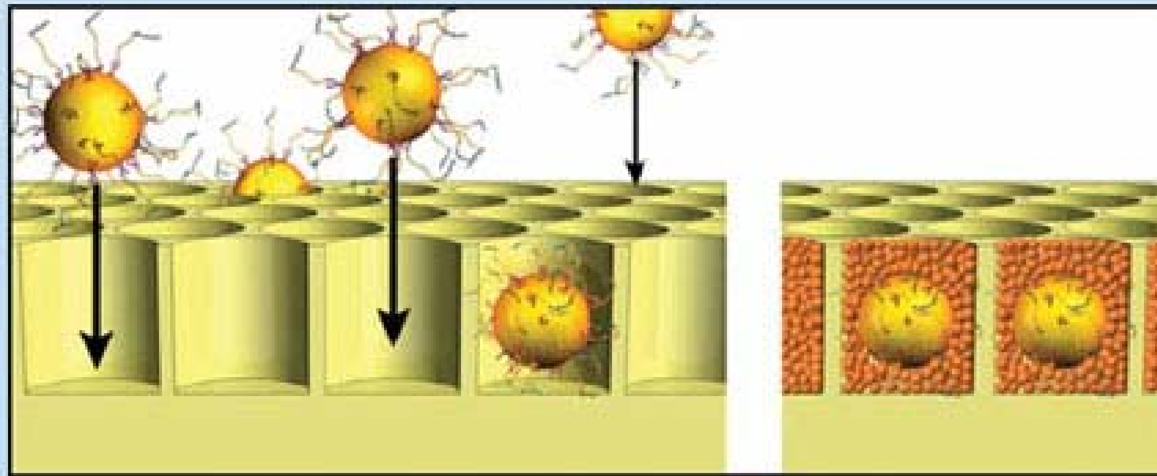


Break microreactors, enrich for DNA-positive beads

sstDNA library → **Clonally-amplified sstDNA attached to bead (millions of copies per bead)**

DNA Library Preparation and Titration — emPCR — **Sequencing**

4.5 hours — 10.5 hours — 8 hours — 5.5 hours

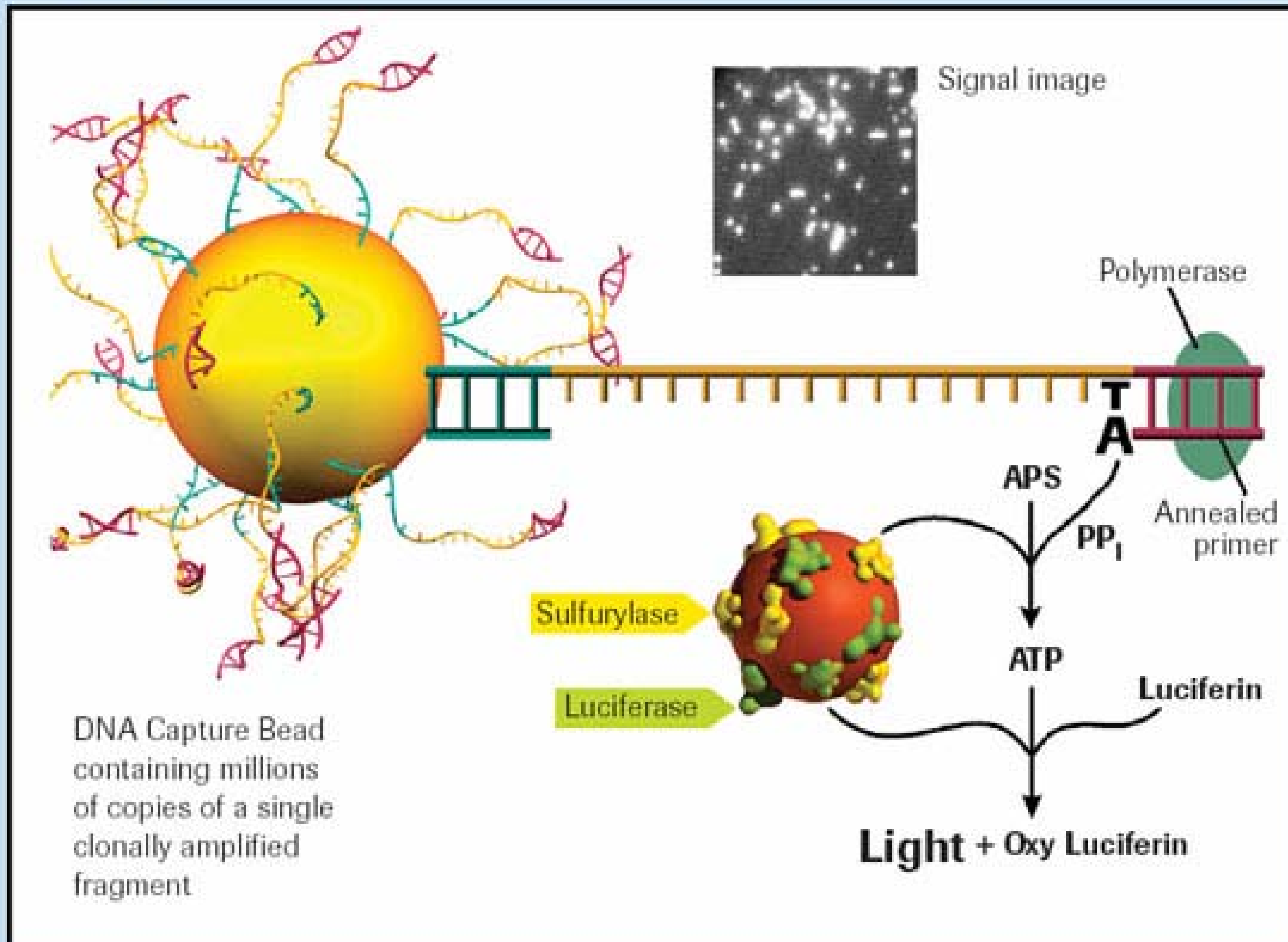


- Well diameter: average of 44 µm
- A single clonally amplified sstDNA bead is deposited per well
- 200,000 reads obtained in parallel on large-format PicoTiterPlate device

Amplified sstDNA library beads → Quality reads

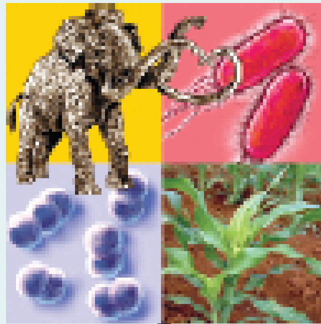
DNA Library Preparation and Titration → emPCR → **Sequencing**

4.5 hours — 10.5 hours — 8 hours — 5.5 hours



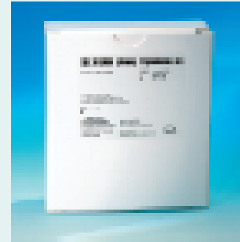
- Bases (TACG) are sequentially flowed (42 times)
- Chemiluminescent signal generation
- Signal processing to determine base sequence and quality score

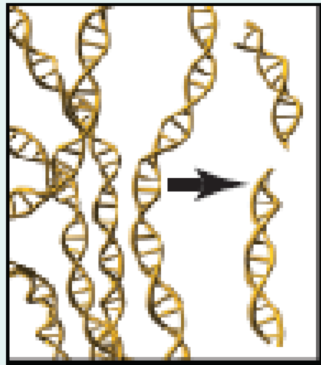
Amplified sstDNA library beads → **Quality reads**

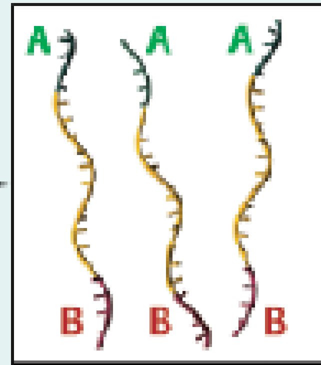


Sample Material

GS 20 DNA Library Preparation Kit





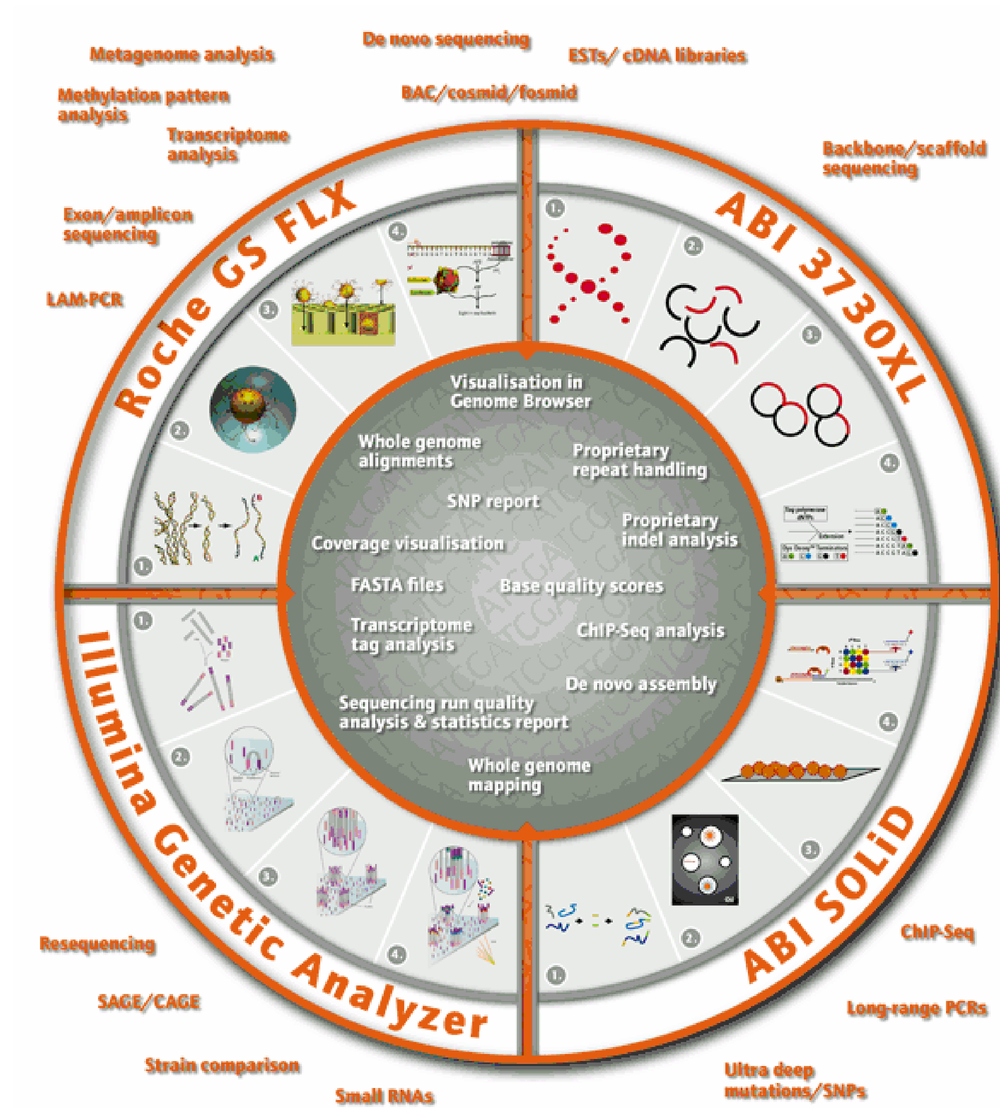


DNA Library Preparation and Titration

- Fragmentation
- Prepare sstDNA library with adaptors
- One library provides enough DNA for thousands of sequence runs
- emPCR (without enrichment step) and one sequencing run (4 small regions)
- Determine amount of sstDNA for the emPCR
- Only one titration per library is needed

20 Millionen Basenpaare
in 4,5 Std,
Ein Bakteriengenom
in einer Woche

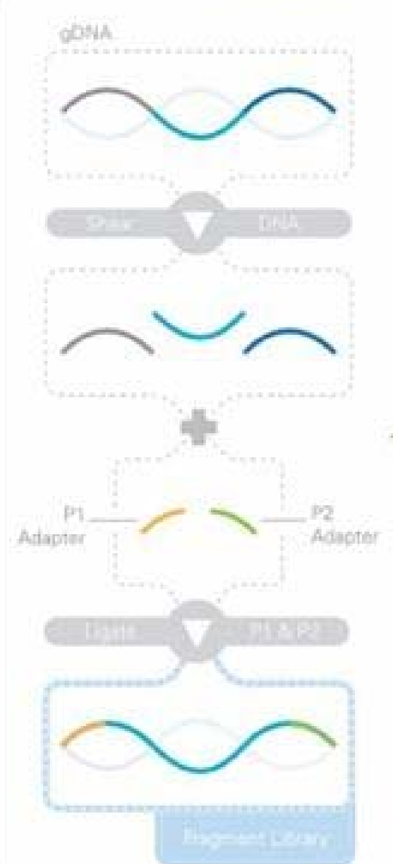
Next generation sequencing



LIBRARY PREPARATION

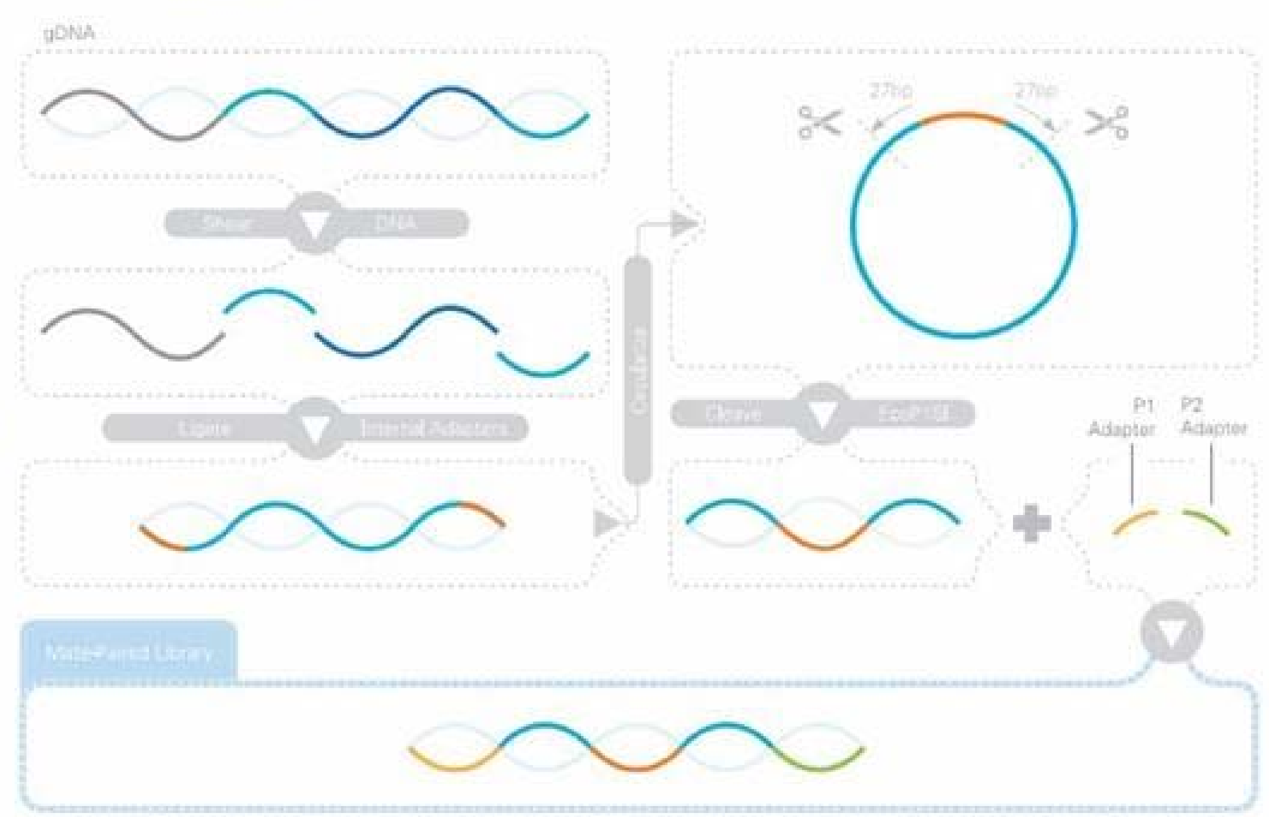
A.

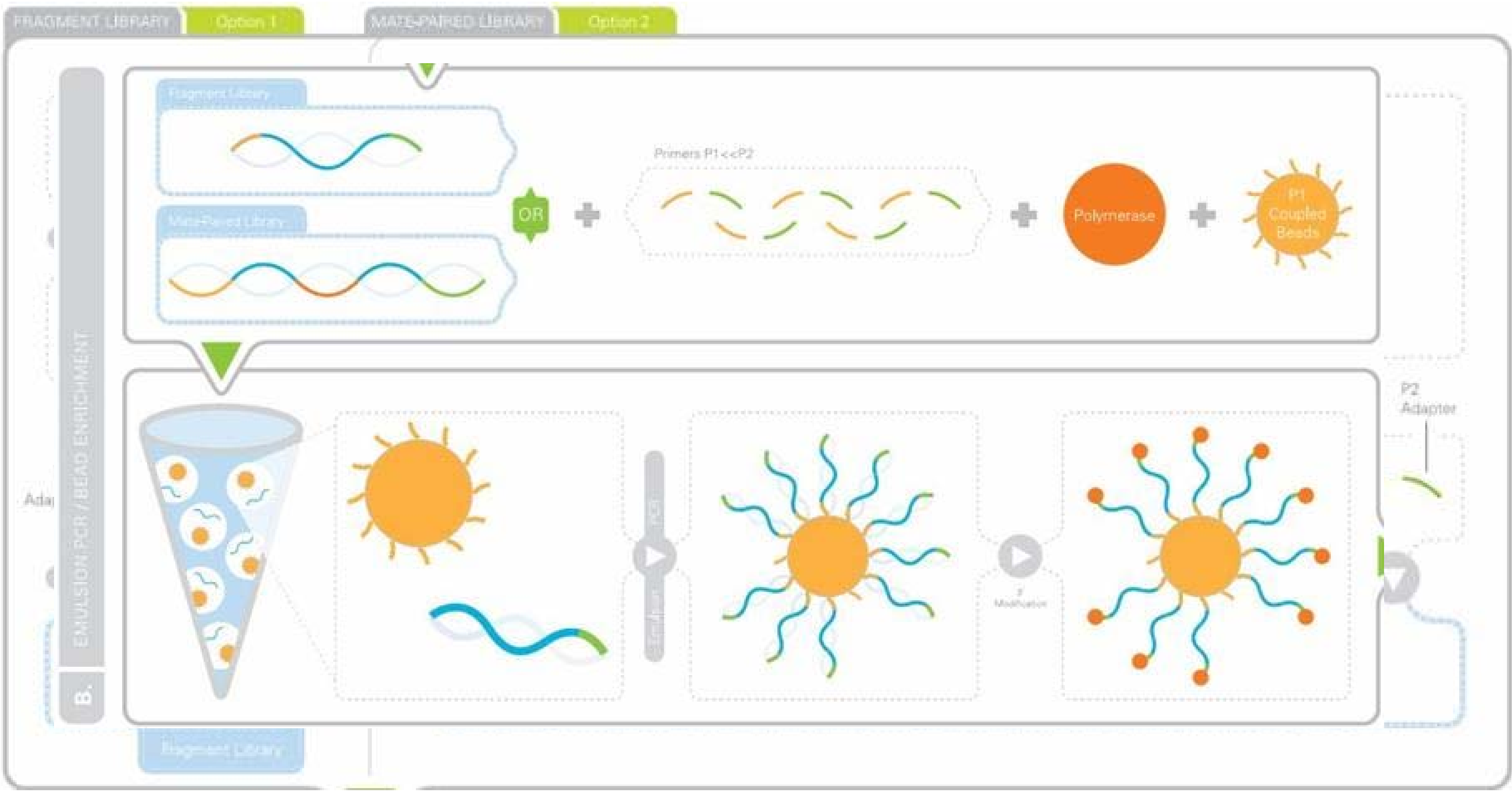
FRAGMENT LIBRARY | Option 1



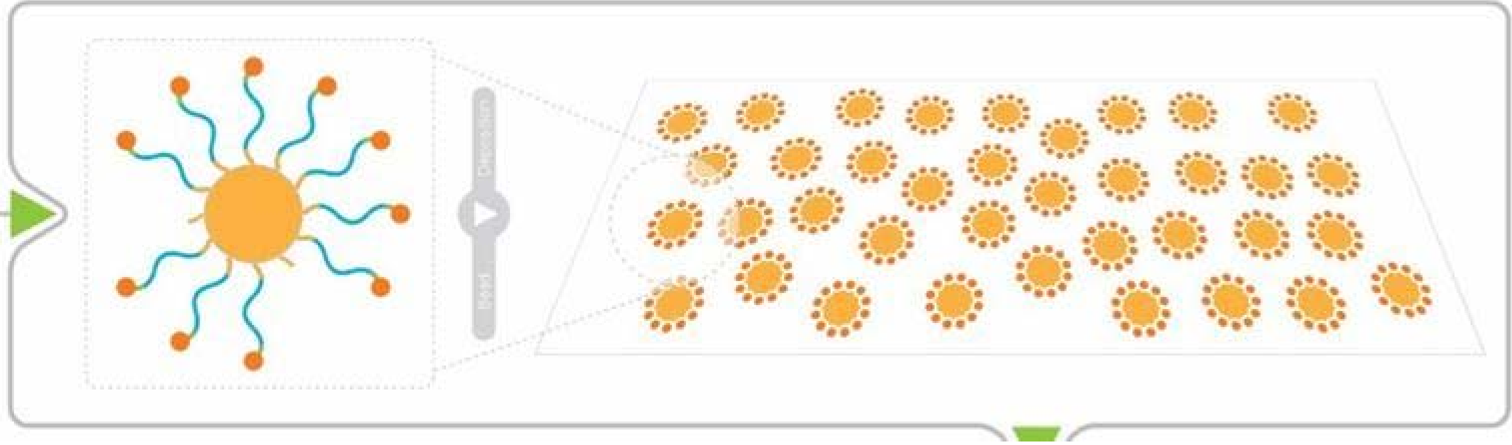
OR

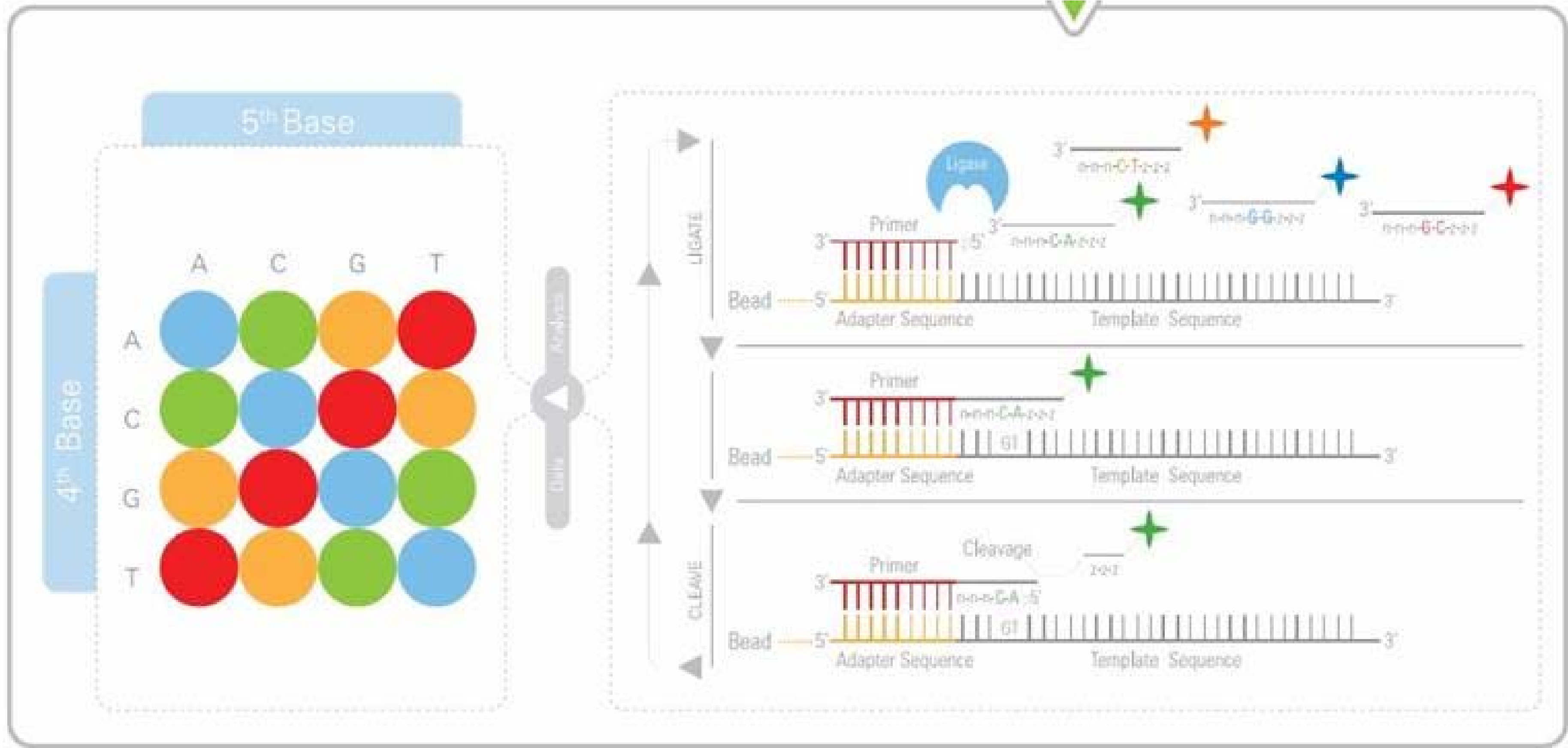
MATE-PAIRED LIBRARY | Option 2





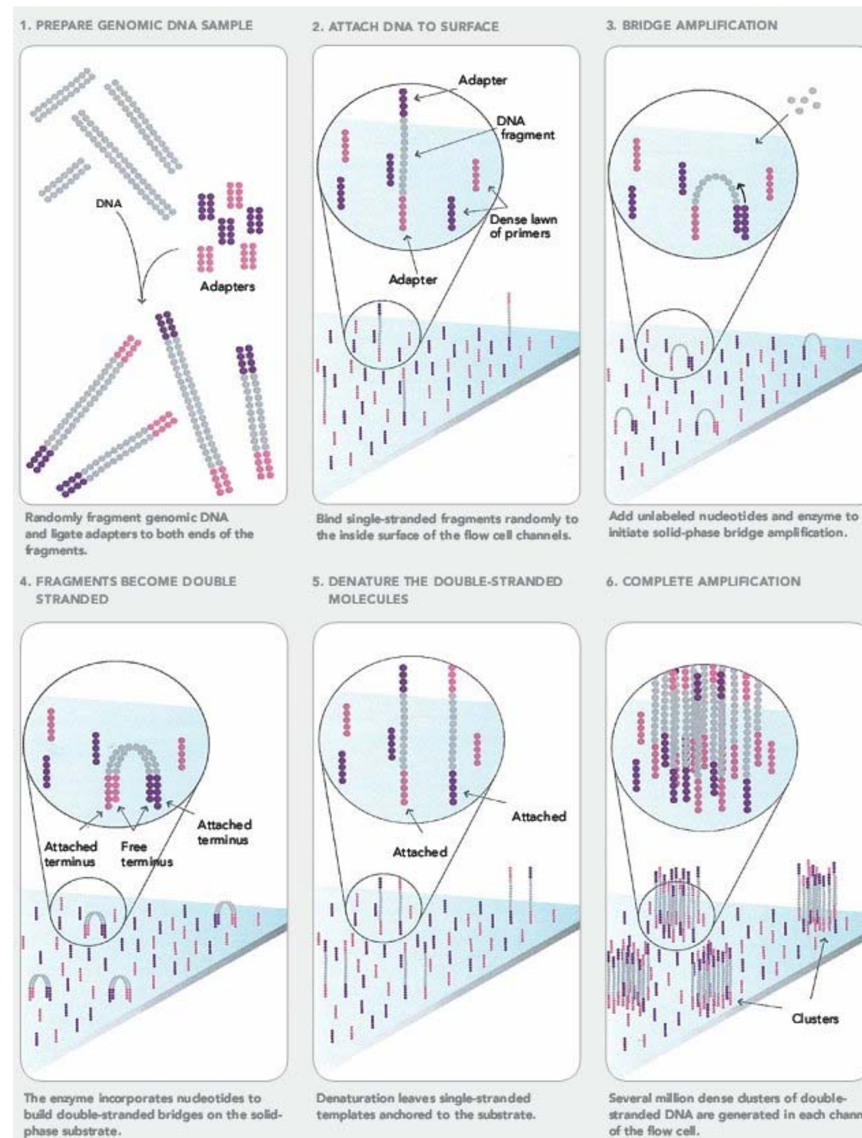
FRAGMENT LIBRARY
Option 1
MATE-PAIRED LIBRARY
Option 2
A.
LIBRARY PREPARATION
B.
EMULSION PCR / BEAD ENRICHMENT
Fragment Library
Mate-Paired Library
OR
Primers P1<<P2
Polymerase
P1 Coupled Beads
P2 Adapter
Fragment Library


C.
BEAD DEPOSITION

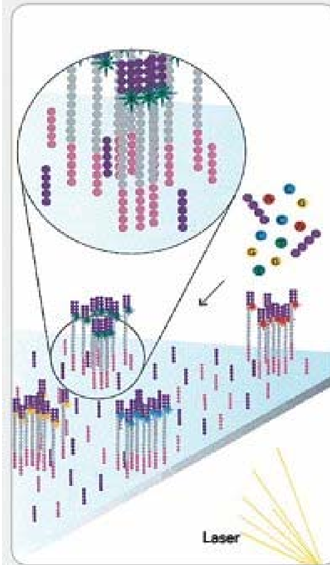


D. SEQUENCING BY LIGATION / DATA ANALYSIS

„paired end sequencing“ (Fa. Illumina



7. DETERMINE FIRST BASE

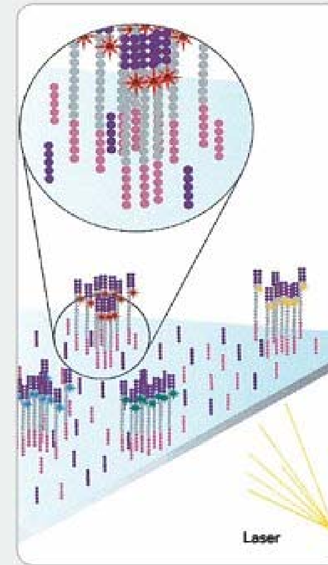


First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

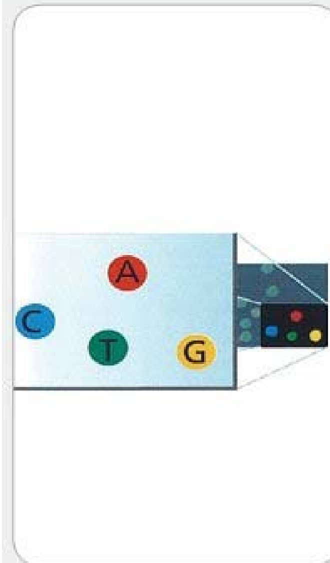
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



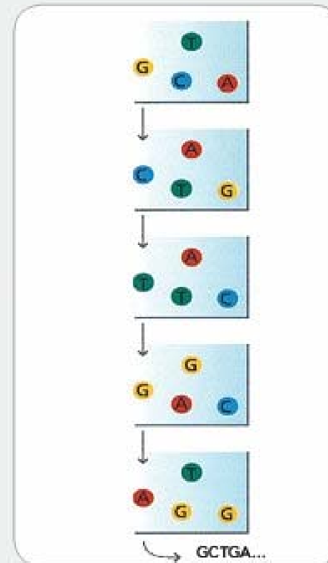
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



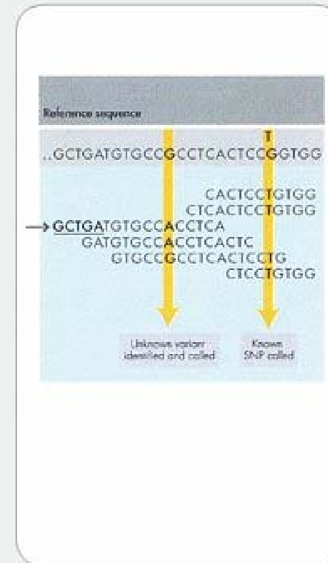
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA



Align data, compare to a reference, and identify sequence differences.

Die Ära der Genomik

„Even the smallest functional DNA varieties seen, those occurring in small phages, must have something like 5000 nucleotides in a row. We may, therefore, leave the task of reading the complete nucleotide sequence of a DNA for the next century, which will, however, have other worries.

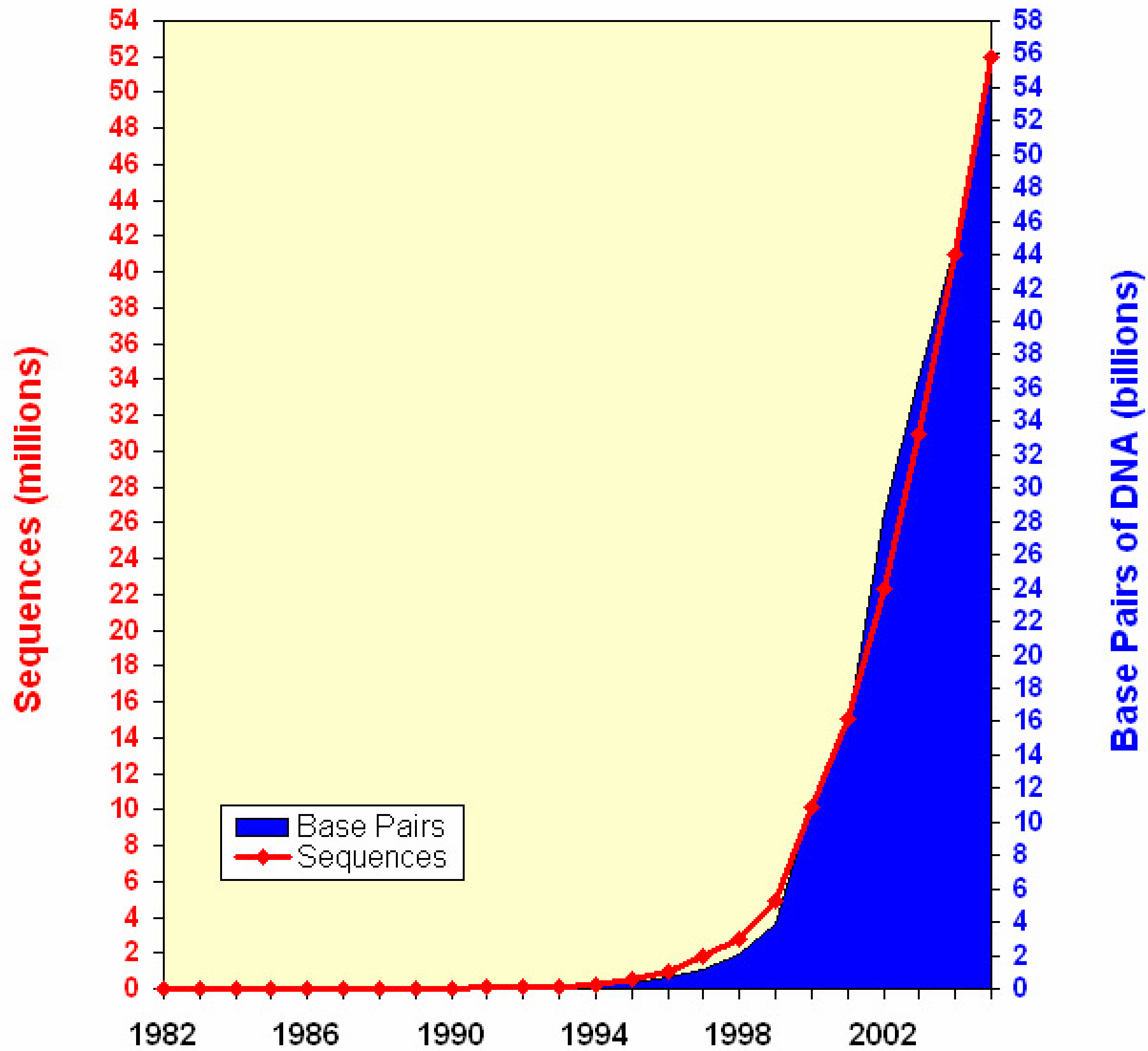
Progress in Nucleic Acid Research and Molecular Biology, 1968

Phi-X 174 sequenced, Nature 1977

Die Geschwindigkeit von Genomanalysen ist exponentiell gewachsen

Die Ära der Genomforschung:

Phi X 174	1977	5.386 bp
λ– Phage	1982	48.502 bp
M. genitalium	1995	580.000 bp
H. influenzae	1995	1.830.000 bp
M. jannaschii	1996	1.660.000 bp
S. cerevisae	1997	12.500.000 bp
E. coli	1997	4.654.000 bp
C. elegans	1998	97.000.000 bp
D. melanog.	1999	116.000.000 bp
A. thaliana	2000	115.000.000 bp
H. sapiens	2001	2.693.000.000 bp
Oryza sativa	2002	420.000.000 bp


Growth of GenBank
(1982 - 2005)
Sequences (millions)
Base Pairs of DNA (billions)
Base Pairs
Sequences
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
1982
1986
1990
1994
1998
2002

Stand der Genomik 12. 06. 08

- 819 komplette Genome sequenziert

  672 Eubakterien

  27 Archebakterien

  94 Eukaryoten

- 2874 aktuelle Genomprojekte

  1848 Bakterien

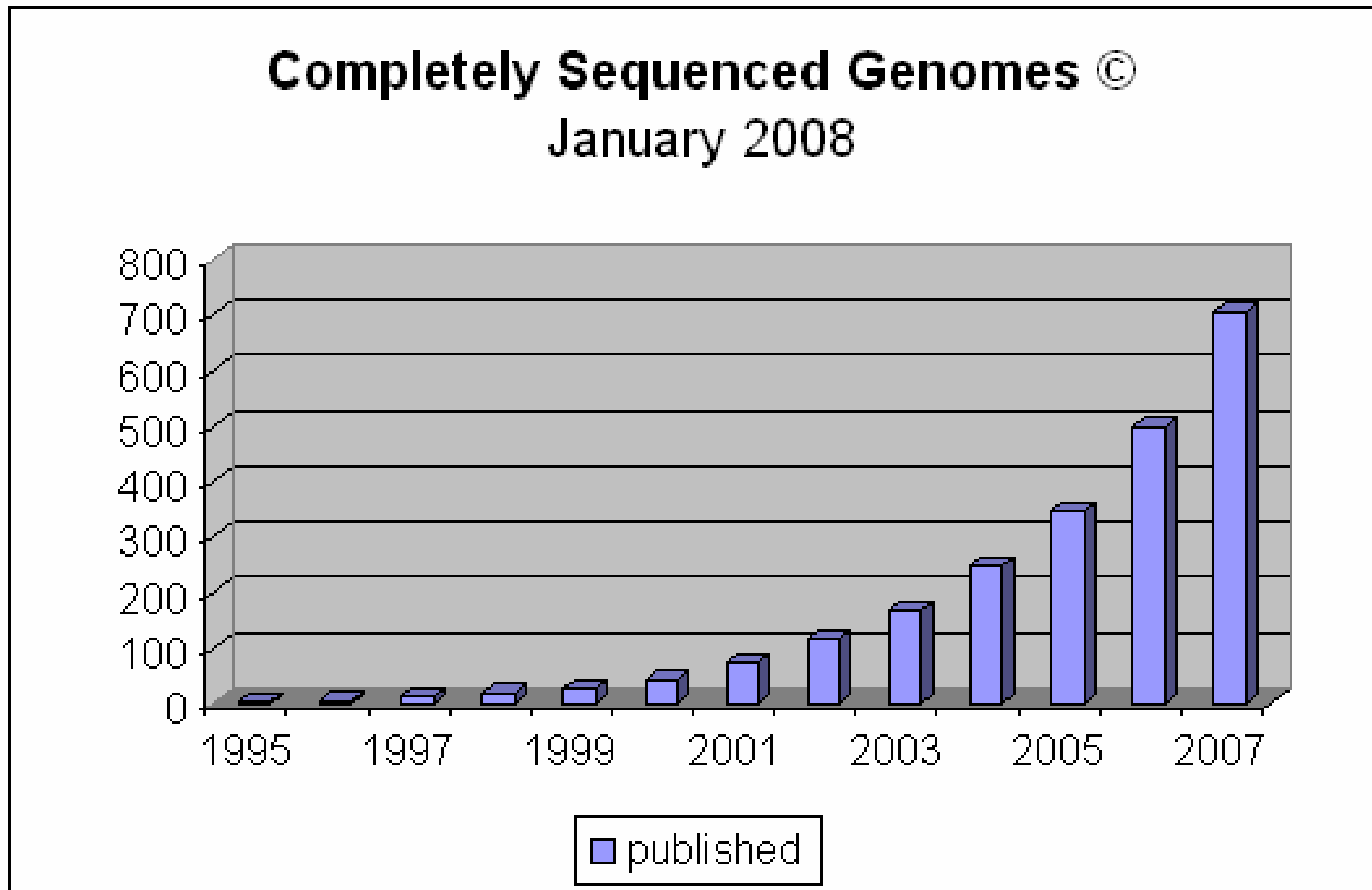
  90 Archebakterien

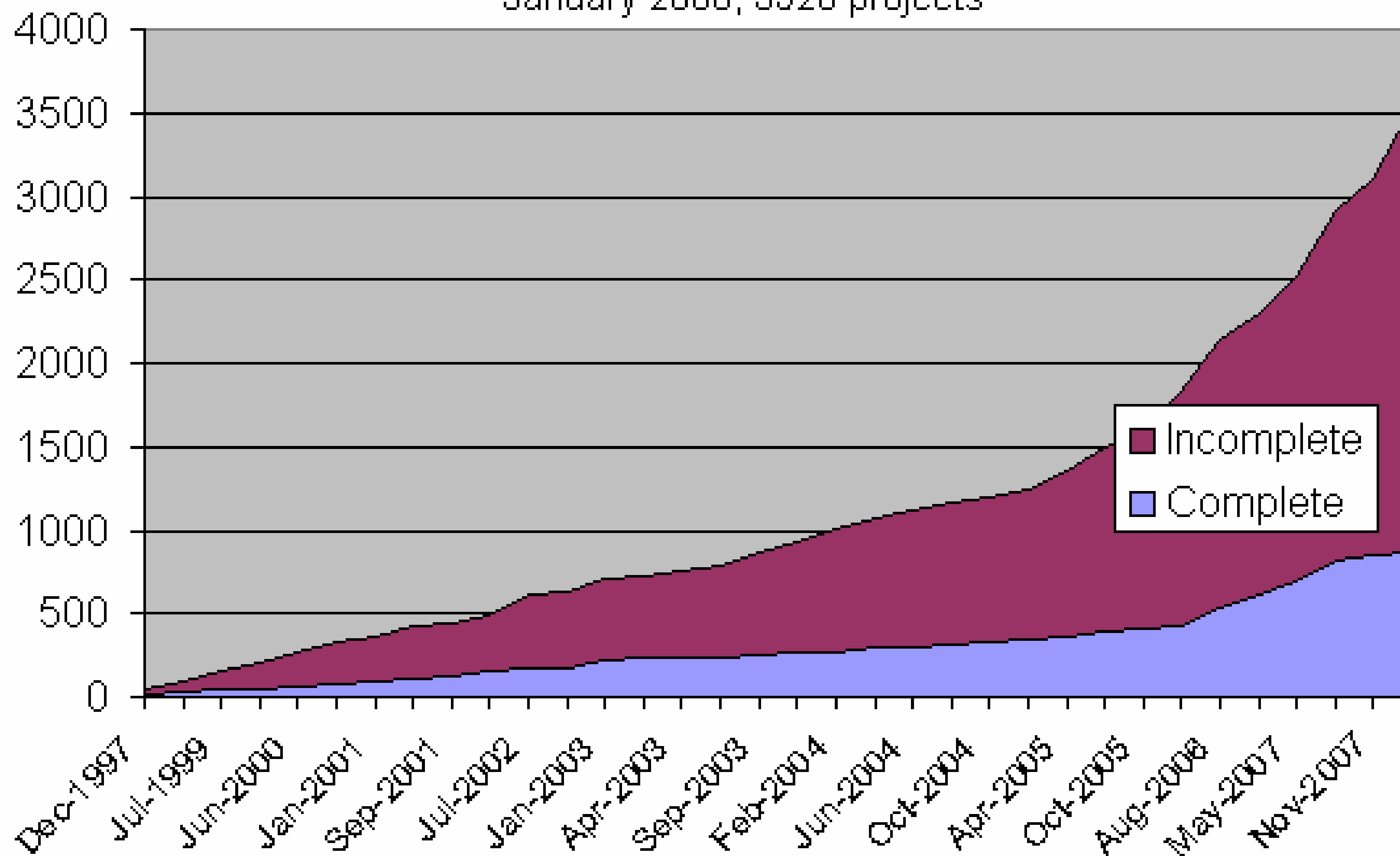
  936 Eukaryoten

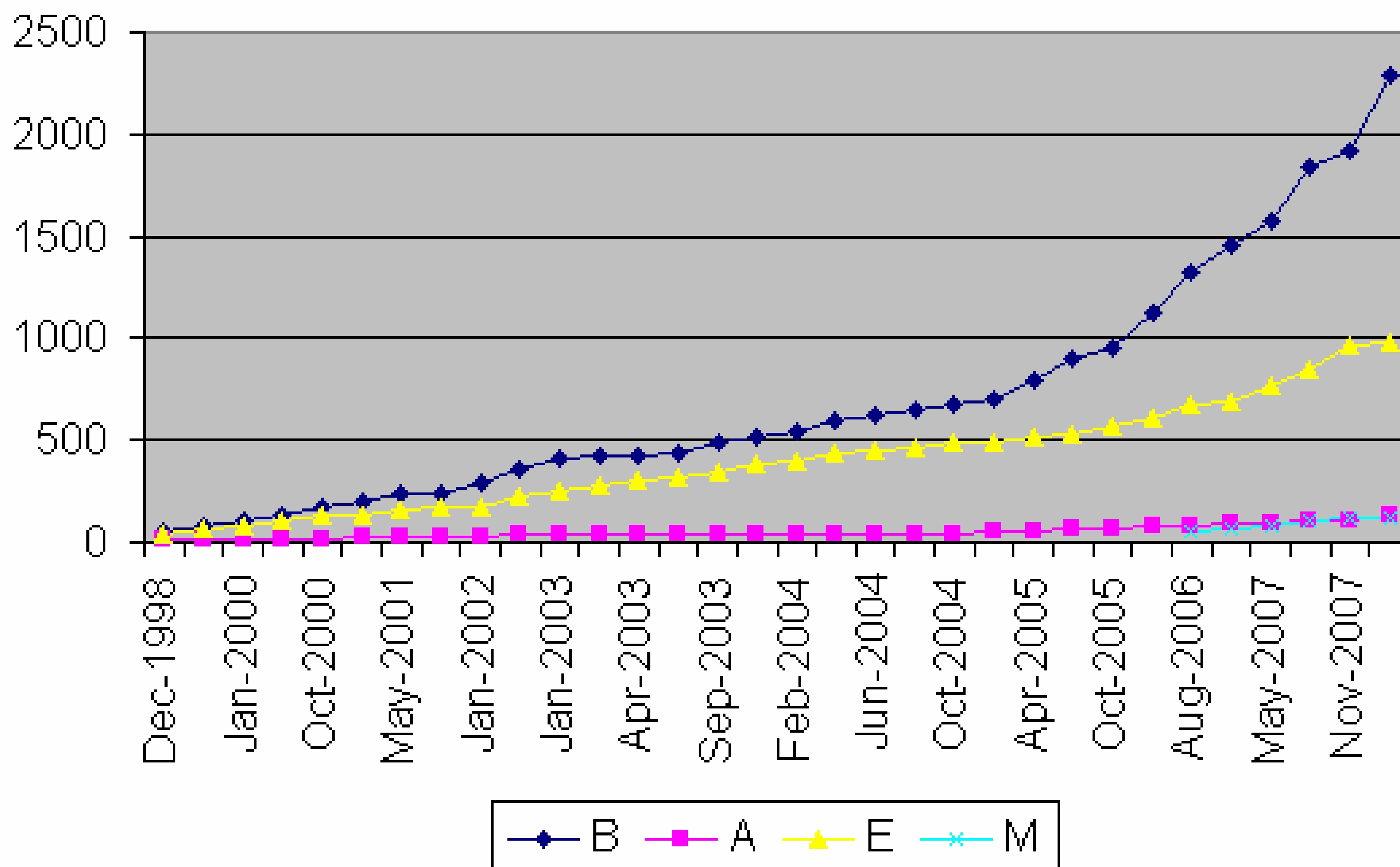
  130 Metagenome

http://www.genomesonline.org/gold.cgi

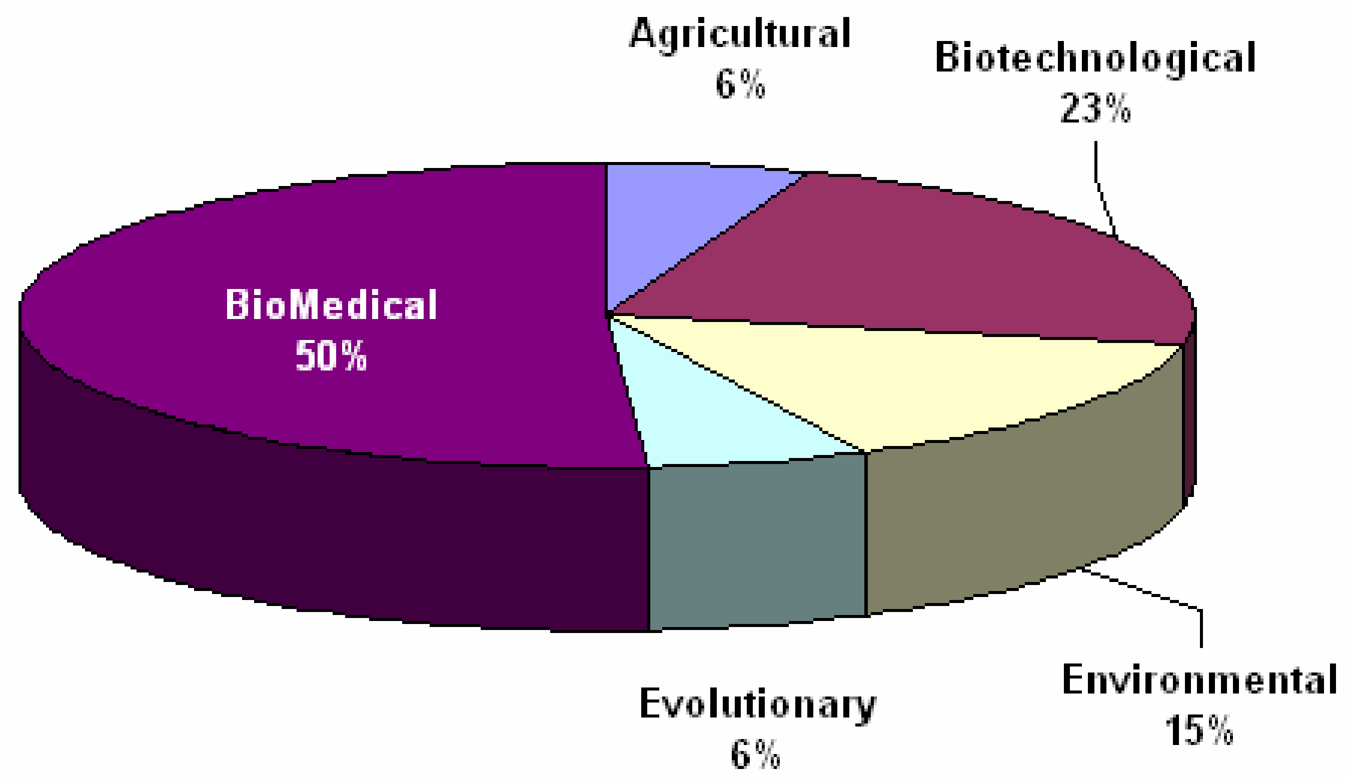
Vollständig sequenzierte Genome

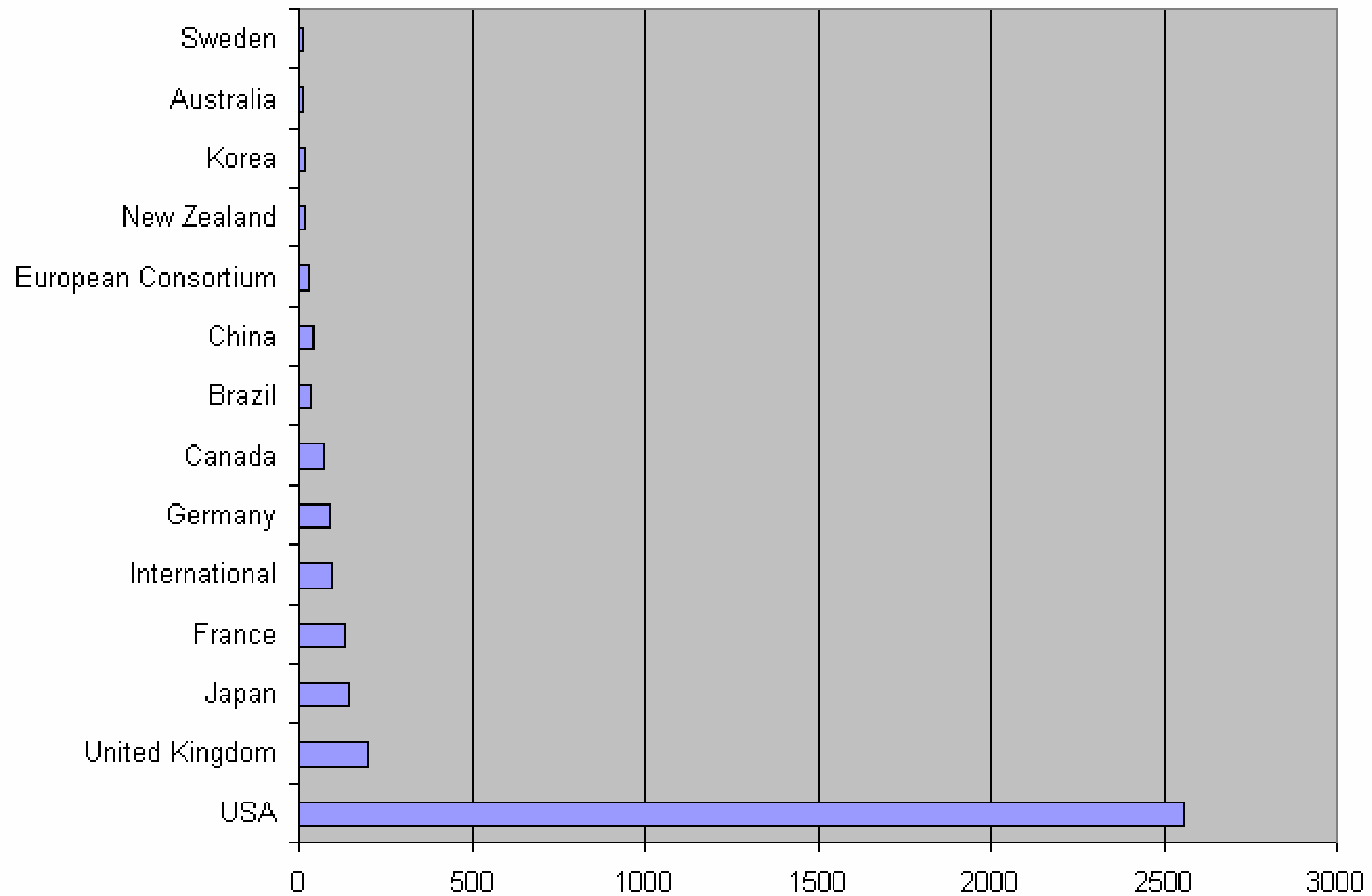



Genome Sequencing Projects on GOLD ©
January 2008, 3520 projects
4000
3500
3000
2500
2000
1500
1000
500
0
Dec-1997
Jul-1999
Jun-2000
Jan-2001
Sep-2001
Jul-2002
Jan-2003
Apr-2003
Sep-2003
Feb-2004
Jun-2004
Oct-2004
Apr-2005
Oct-2005
Aug-2006
May-2007
Nov-2007
Incomplete
Complete

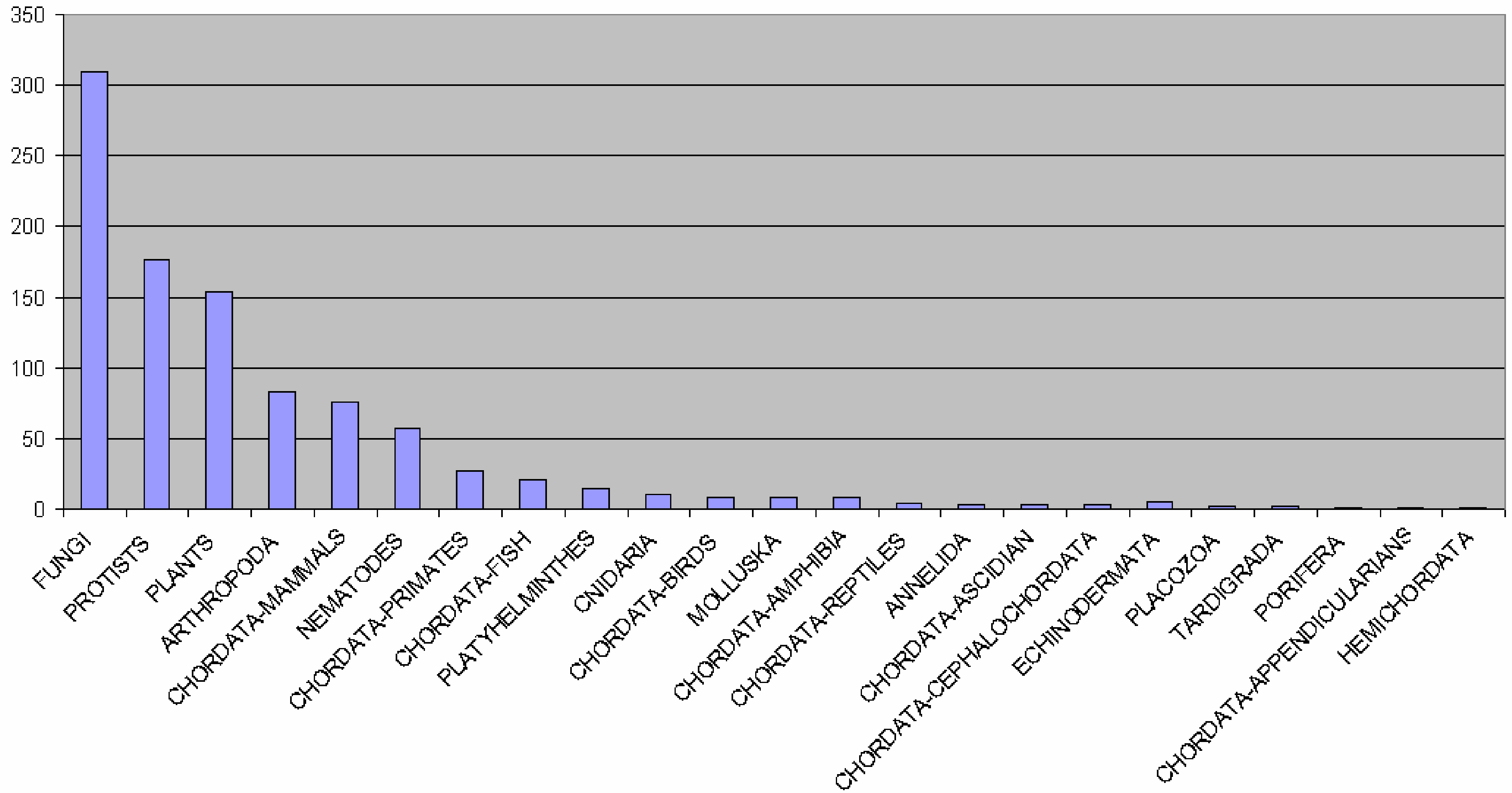


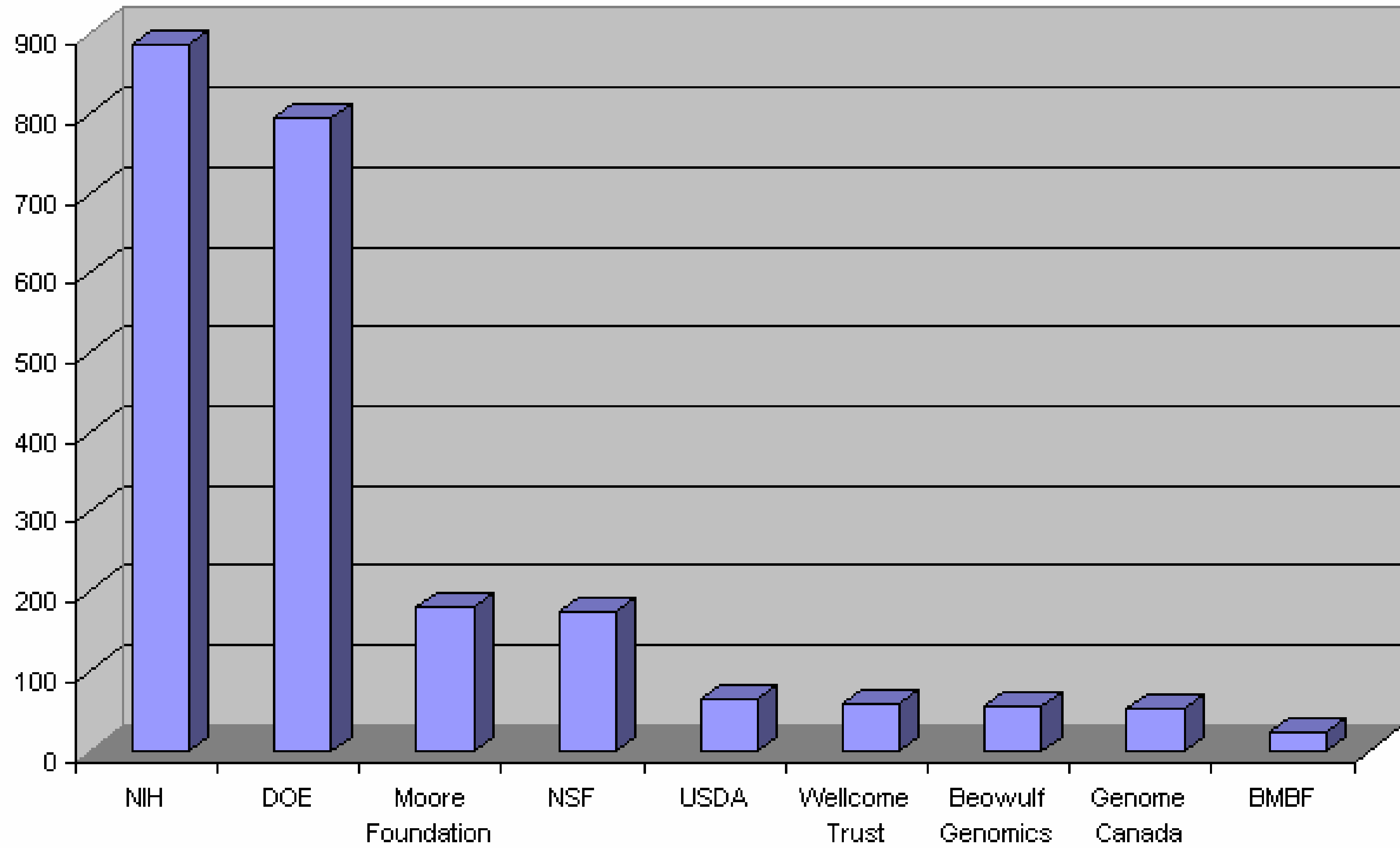
Genome Projects on GOLD according to Phylogenetic Groups ©
January 2008
2500
2000
1500
1000
500
0
Dec-1998
Jan-2000
Oct-2000
May-2001
Jan-2002
Jan-2003
Apr-2003
Sep-2003
Feb-2004
Jun-2004
Oct-2004
Apr-2005
Oct-2005
Aug-2006
May-2007
Nov-2007
B
A
E
M

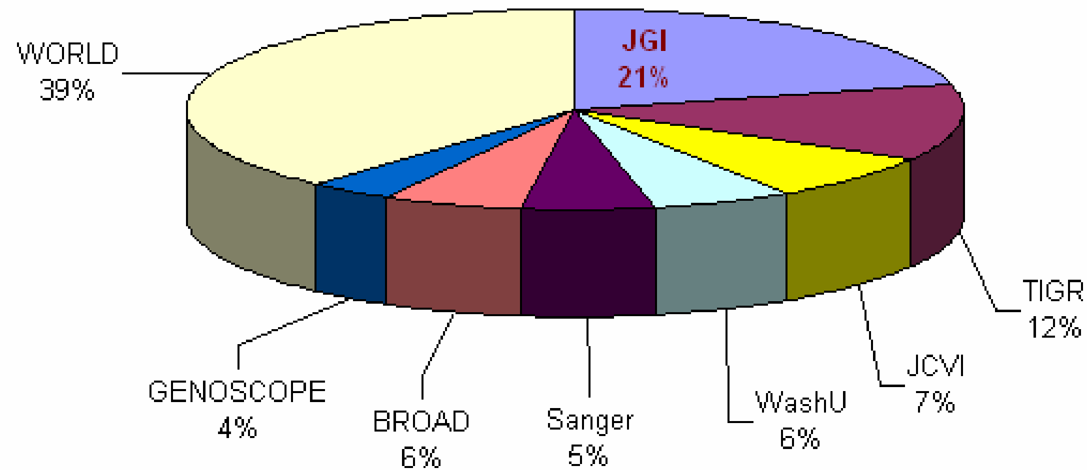

Funding Relevance of Bacterial Genome Projects ©
Agricultural
6%
Biotechnological
23%
BioMedical
50%
Evolutionary
6%
Environmental
15%



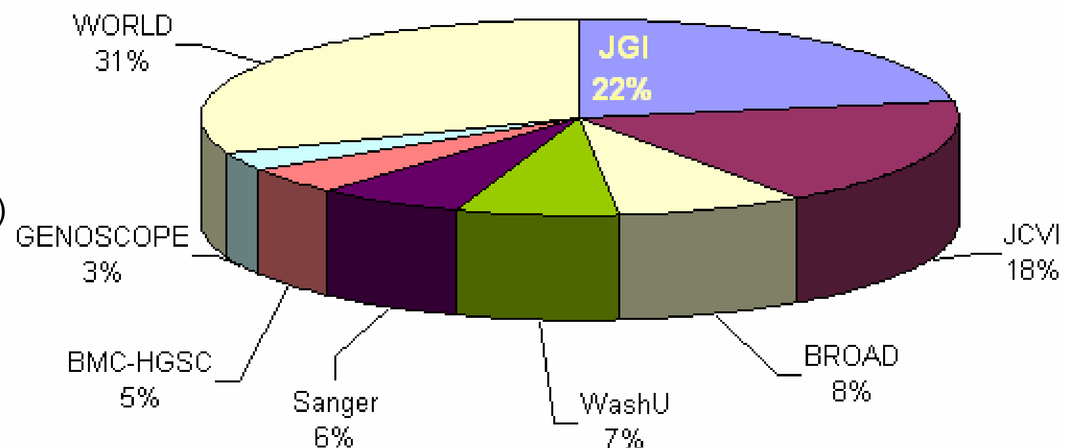
Top Countries with Genome Projects
January 2008
Sweden
Australia
Korea
New Zealand
European Consortium
China
Brazil
Canada
Germany
International
France
Japan
United Kingdom
USA
0
500
1000
1500
2000
2500
3000


Eukaryotic Phyla with Genome Projects
January 2008
350
300
250
200
150
100
50
0
FUNGI
PROTISTS
PLANTS
ARTHROPODA
CHORDATA-MAMMALS
NEMATODES
CHORDATA-PRIMATES
CHORDATA-FISH
PLATYHELMINTHES
CNIDARIA
CHORDATA-BIRDS
MOLLUSKA
CHORDATA-AMPHIBIA
CHORDATA-REPTILES
ANNELIDA
CHORDATA-ASCIDIAN
CHORDATA-CEPHALOCHORDATA
ECHINODERMATA
PLACOZOA
TARDIGRADA
PORIFERA
CHORDATA-APPENDICULARIANS
HEMICHORDATA


Top Funding Agencies
January 2008
900
800
700
600
500
400
300
200
100
0
NIH
DOE
Moore Foundation
NSF
USDA
Wellcome Trust
Beowulf Genomics
Genome Canada
BMBF



BMC: Baylor College of Medicine
Genoscope: Centre National de ´sequencage
Sanger: ThéWellcome Trust Sanger Institute
WashU: Washington University Genome
Sequencing Center
JGI: Joint Genome Institute,
Department of Energy, USA
BROAD: Eli & Edythe Broad Institute
(MIT, Whitehead, Harvard Univ.)
JCVI: J. Craig Venter Institute



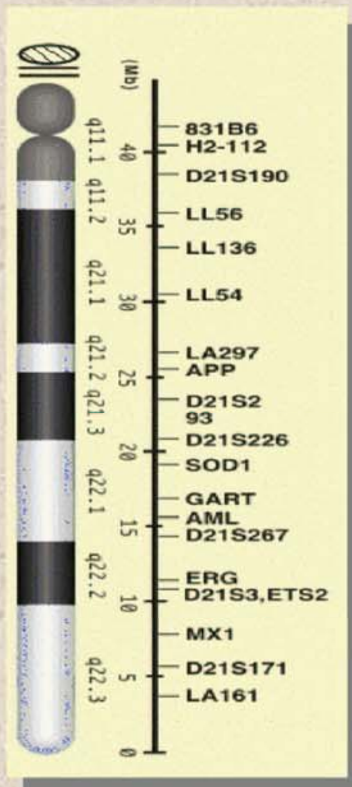
Nature vom 2. Dez. 1999:

„The DNA sequence of human chromosome 22“

- 33 464 000 Basenpaare
- noch 11 Lücken
- 545 Gene (davon 298 = 55% neu)
- 134 Pseudogene
- 39% der DNA von Genen besetzt (Exons + Introns)
- Genlängen: 1 kb min. bis 583 kb max
- 3% der DNA kodiert für Proteine
- 42% besteht aus repetitiver„junk“-DNA

Nature vom 10. Apr. 2000:

„The DNA sequence of human chromosome 21"

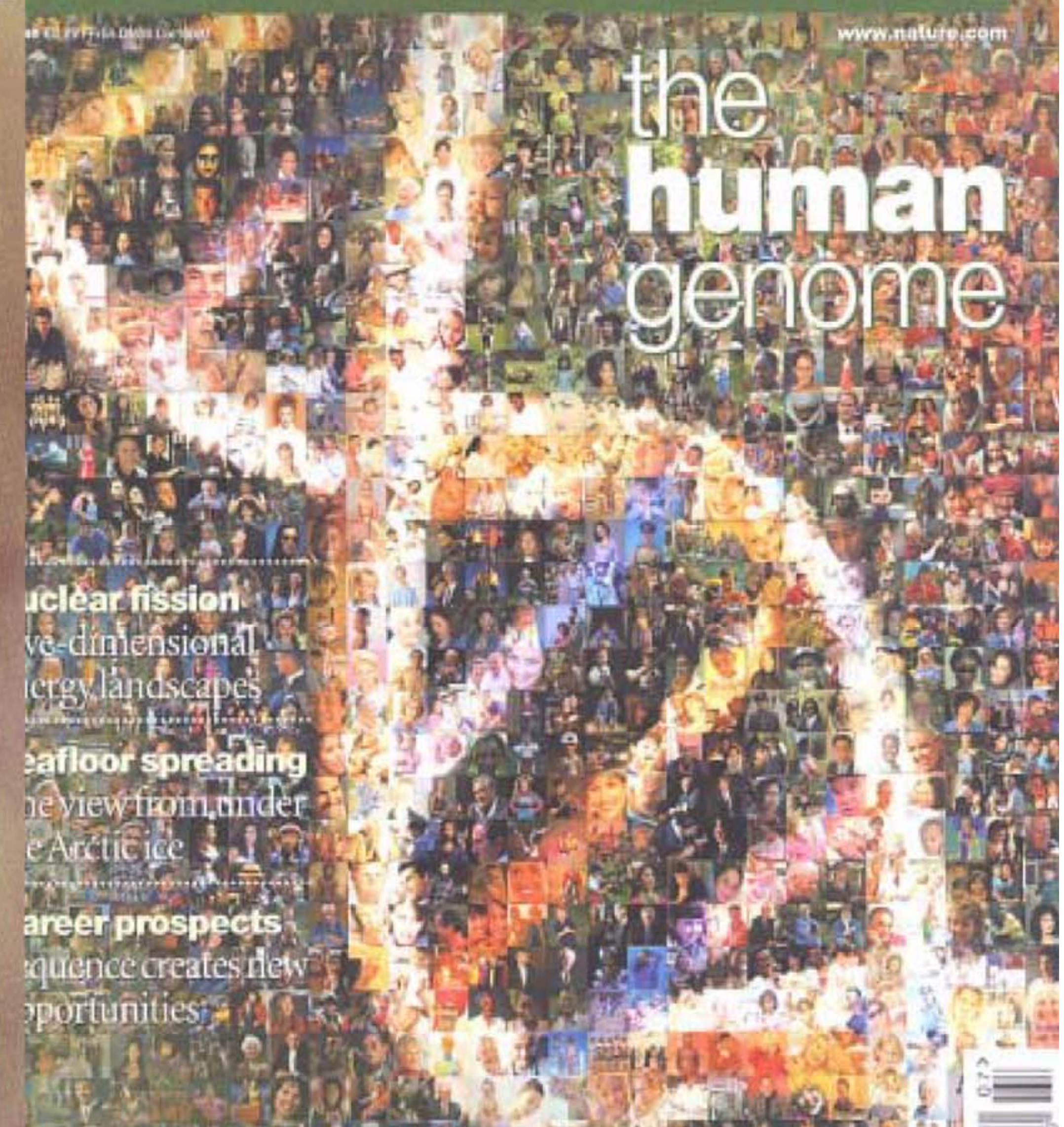


- **das kleinste menschliche Chromosom (1% der DNA)**
- **33 546 361 Basenpaare (3 Lücken, <100kb)**
- **225 Gene, 127 davon bekannt**
- **59 Pseudogene**
- **7 Mbp Abschnitt ohne einziges Gen!**

Maximal 40 000 Gene ??

(zum Vergl.: MHC Genkomplex auf Chr. 6 hat 128 Gene in nur 3,6 Mb)

2001



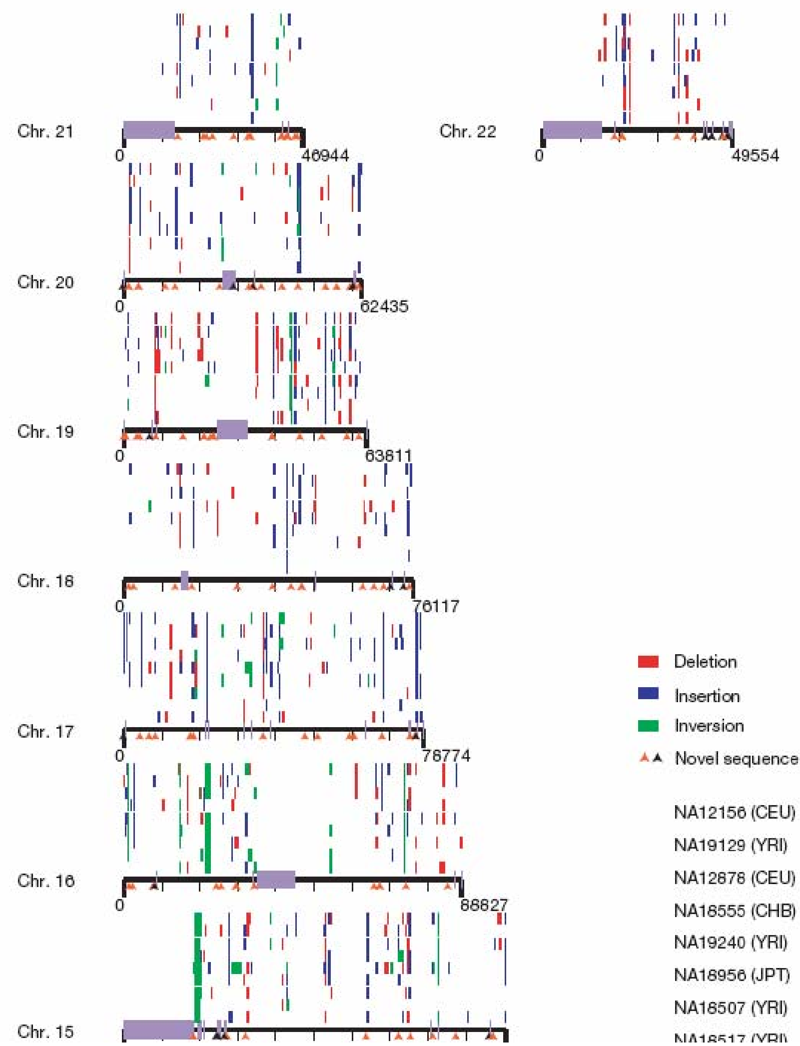
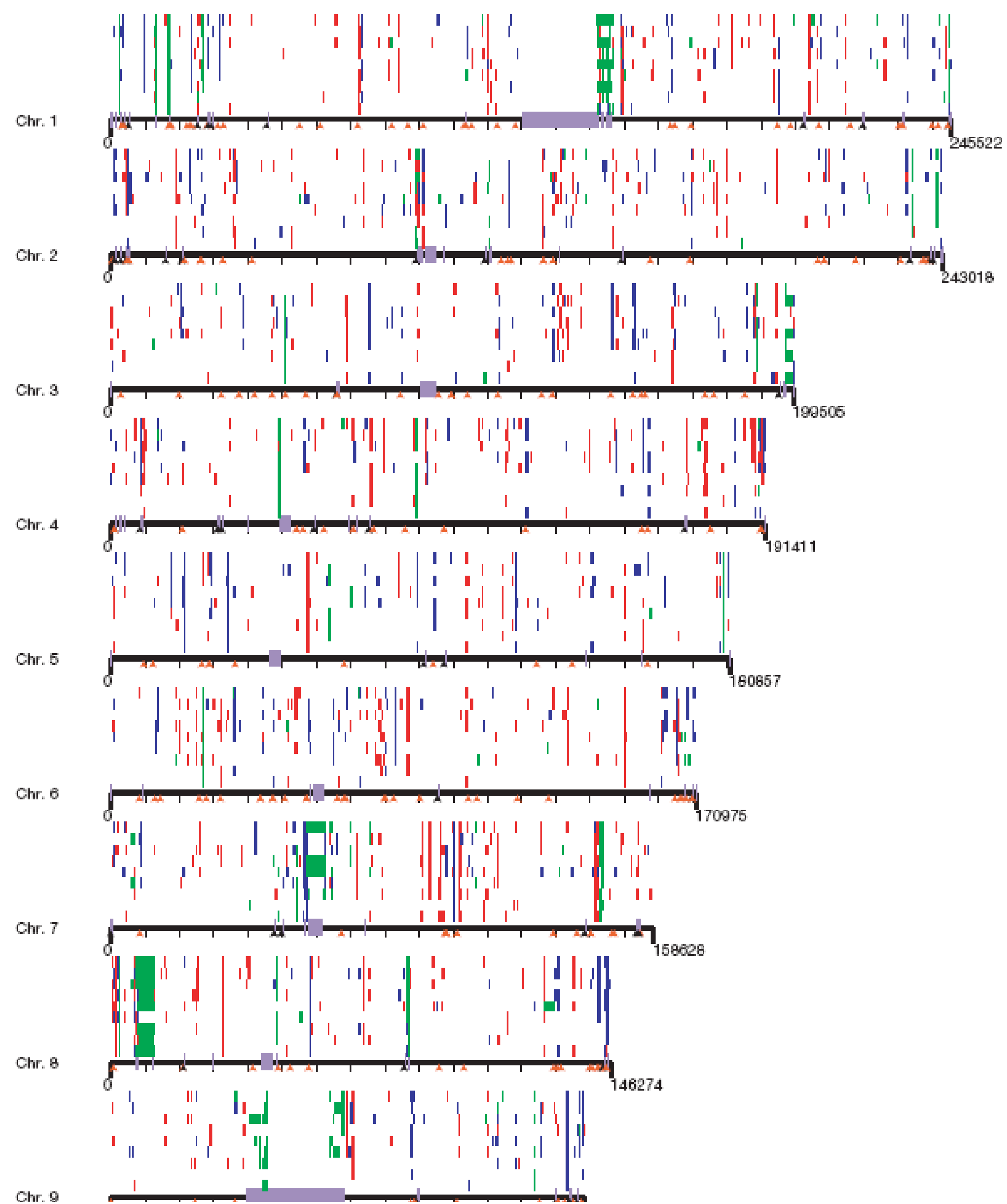


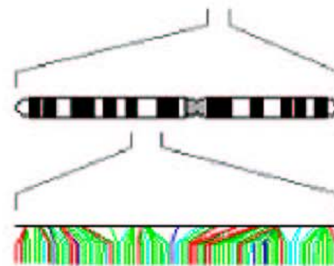
Figure 1 | Map of structural variation in the human genome. The location of 724 insertions (blue), 747 deletions (red) and 224 inversions (green) that have been experimentally validated are mapped onto the human genome (build35). Sites are arranged according to individuals in rows above each chromosome, in order of the nine individual genomic libraries (G248 (first row), then ABC7–ABC14); the Coriell IDs are listed in Table 1. All sites have been validated by array CGH, MCD analysis, or sequencing in at least one reference individual. The location of 525 novel sequence loci are depicted as arrows below each chromosome. Those mapping to gaps (black) are distinguished from those mapping to regions not associated with gaps (orange). The Y chromosome is not shown because samples were primarily from females.


Chr. 1
0
245522
Chr. 2
0
243018
Chr. 3
0
199505
Chr. 4
0
191411
Chr. 5
0
180857
Chr. 6
0
170975
Chr. 7
0
158628
Chr. 8
0
146274
Chr. 9

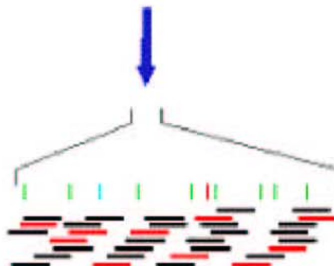
STRATEGIES FOR SEQUENCING THE HUMAN GENOME

BY MAPPED CLONES

1. Construction of maps of ordered landmarks (genetic markers, genes): provides long-range map and organisation into individual chromosomes.

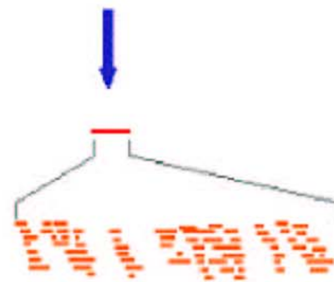


2. Physical maps of overlapping clones anchored to the landmark maps.



3. Selection of tile path (clones in red)

4. Shotgun sequencing and assembly (for working draft); subsequent directed finishing (for reference sequence).



BY WHOLE GENOME SHOTGUN

1. Shotgun sequencing of short-insert clones

2. Paired end sequencing of large-insert clones

3. Assembly of seed contigs (unitigs)

4. Incorporation of other sequences, and integration of long- range data.

Celera‘s „*whole genome shotgun*“

- Gesamt-Genom: 3, 12 Milliarden Basenpaare (99%)
- Zeitraum 8.9.99 bis 26.6.00
- 5 Individuen sequenziert (untersch. ethnische Gruppen)
- 26,4 Mio. Sequenzierungsreaktionen à 550 Bp
  = 14,5 x 10^9 Bp = 4,6 fache Abdeckung

HGP‘s „draft sequence“

- 350 000 BAC/PAC-Klone kartiert, ca. 30 000 Klone sequenziert
- 90% der 3,2 Milliarden Basenpaare in Datenbank, enthält 95% aller bekannten Gene
- ca. 7 fache Abdeckung des Genoms
- draft-Einträge bestehen aus 2-10 Teilsequenzen (Contigs), Reihenfolge oft unklar
- draft-Auswertung und -Alignment unter www.ensembl.org

Human Genome Project Centers

- 1.Baylor College of Medicine, Houston, Texas, USA
  2.Beijing Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing, China
  3.Gesellschaft für Biotechnologische Forschung mbH, Braunschweig, Germany
  4.Genoscope, Evry, France
  5.Genome Therapeutics Corporation, Waltham, MA, USA
  6.Institute for Molecular Biotechnology, Jena, Germany
- 7.Joint Genome Institute, U.S. Department of Energy, Walnut Creek, CA, USA
  8.Keio University, Tokyo, Japan
  9.Max Planck Institute for Molecular Genetics, Berlin, Germany
  10.RIKEN Genomic Sciences Center, Saitama, Japan
- 11.The Sanger Centre, Hinxton, U.K.
  12.Stanford DNA Sequencing and Technology Development Center, Palo Alto, CA, USA
  13.University of Washington Genome Center, Seattle, WA, USA
  14.University of Washington Multimegabase Sequencing Center, Seattle, WA,USA
- 15.Whitehead Institute for Biomedical Research, MIT, Cambridge, MA, USA
- 16.Washington University Genome Sequencing Center, St. Louis, MO, USA