National Science Foundation
Tokyo Celebration

W. Richard McCombie
Cold Spring Harbor Lab
Watson School of Biological Sciences

October 6, 2010

articles

Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*

The Arabidopsis Genome Initiative*

* *Authorship of this paper should be cited as 'The Arabidopsis Genome Iniative'. A full list of contributors appears at the end of this paper*

The flowering plant *Arabidopsis thaliana* is an important model system for identifying genes and determining their functions. Here we report the analysis of the genomic sequence of *Arabidopsis*. The sequenced regions cover 115.4 megabases of the 125-megabase genome and extend into centromeric regions. The evolution of *Arabidopsis* involved a whole-genome duplication, followed by subsequent gene loss and extensive local gene duplications, giving rise to a dynamic genome enriched by lateral gene transfer from a cyanobacterial-like ancestor of the plastid. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of *Drosophila* and *Caenorhabditis elegans*— the other sequenced multicellular eukaryotes. *Arabidopsis* has many families of new proteins but also lacks several common protein families, indicating that the sets of common proteins have undergone differential expansion and contraction in the three multicellular eukaryotes. This is the first complete genome sequence of a plant and provides the foundations for more comprehensive comparison of conserved processes in all eukaryotes, identifying a wide range of plant-specific gene functions and establishing rapid systematic ways to identify genes for crop improvement.

© 2000 Macmillan Magazines Ltd

Genome Sequencing Groups

Samir Kaul, Hean L. Koo, Jennifer Jenkins, Michael Rizzo, Timothy Rooney, Luke J. Tallon, Tamara Feldblyum, William Nierman, Maria-Ines Benito, Xiaoying Lin, Christopher D. Town, J. Craig Venter & Claire M. Fraser

The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA

Satoshi Tabata, Yasukazu Nakamura, Takakazu Kaneko, Shusei Sato, Erika Asamizu, Tomohiko Kato, Hirokazu Kotani & Shigemi Sasamoto

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292, Japan

Joseph R. Ecker[1]*†, Athanasios Theologis[2]*, Nancy A. Federspiel[3]*†, Curtis J. Palm[3], Brian I. Osborne[2], Paul Shinn[1], Aaron B. Conway[3], Valentina S. Vysotskaia[2], Ken Dewar[1], Lane Conn[3], Catherine A. Lenz[2], Christopher J. Kim[1], Nancy F. Hansen[3], Shirley X. Liu[2], Eugen Buehler[1], Hootan Altafi[3], Hitomi Sakano[2], Patrick Dunn[1], Bao Lam[3], Paul K. Pham[2], Qimin Chao[1], Michelle Nguyen[3], Guixia Yu[2], Huaming Chen[1], Audrey Southwick[3], Jeong Mi Lee[2], Molly Miranda[3], Mitsue J. Toriumi[2] & Ronald W. Davis[3]

*1, Plant Science Institute, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104 USA; 2, Plant Gene Expression Center/USDA-U.C.Berkeley, 800 Buchanan Street, Albany, California 94710, USA; 3, Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA. * These authors contributed equally to this work. † Present addresses: The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA (J.R.E.); Exelixis, Inc., 170 Harborway, P.O. Box 511, South San Francisco, California 94083-0511, USA (N.A.F)*

European Union Chromosome 4 and 5 Sequencing Consortium: R. Wambutt[1], G. Murphy[2], A. Düsterhöft[3], W. Stiekema[4], T. Pohl[5], K.-D. Entian[6], N. Terryn[7] & G. Volckaert[8]

1, AGOWA GmbH, Glienicker Weg 185, D-12489 Berlin, Germany; 2, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 3, QIAGEN GmbH, Max-Volmer-Str. 4, D-40724 Hilden, Germany; 4, Greenomics, Plant Research International, Droevendaalsesleeg 1, NL 6700, AA Wageningen, The Netherlands; 5, GATC GmbH, Fritz-Arnold Strasse 23, D-78467 Konstanz, Germany; 6, SRD GmbH, Oberurseler Str. 43, Oberursel 61440, Germany; 7, Department for Plant Genetics, (VIB), University of Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium; 8, Katholieke Universiteit Leuven, Laboratory of Gene Technology, Kardinaal Mercierlaan 92, B-3001 Leuven, Belgium

European Union Chromosome 3 Sequencing Consortium: M. Salanoubat[1], N. Choisne[1], M. Rieger[2], W. Ansorge[3], M. Unseld[4], B. Fartmann[5], G. Valle[6], F. Artiguenave[1], J. Weissenbach[1] & F. Quetier[1]

1, Genoscope and CNRS FRE2231, 2 rue G. Crémieux, 91057 Evry Cedex, France; 2, Genotype GmbH Angelhofweg 39, D-69259 Wilhemlsfeld, Germany; 3, European Molecular Biology Laboratory, Biochemical Instrumentation Program, Meyerhoftstr. 1, D-69117 Heidelberg, Germany; 4, LION Bioscience AG, Im Neuenheimer Feld 515-517, 69120 Heidelberg, Germany; 5, MWG-Biotech AG, Anzinger Strasse 7a, 85560 Ebersberg, Germany; 6, CRIBI, Università di Padova, via G. Colombo 3, Padova 35131, Italy

The Cold Spring Harbor and Washington University Genome Sequencing Center Consortium: Richard K. Wilson[1], Melissa de la Bastide[2], M. Sekhon[1], Emily Huang[2], Lori Spiegel[2], Lidia Gnoj[2], K. Pepin[1], J. Murray[1], D. Johnson[1], Kristina Habermann[2], Neilay Dedhia[2], Larry Parnell[2], Raymond Preston[2], L. Hillier[1], Ellson Chen[3], M. Marra[2], Robert Martienssen[4] & W. Richard McCombie[2]

1, Washington University Genome Sequencing Center, Washington University in St Louis School of Medicine, 4444 Forest Park Blvd., St. Louis, Missouri 63108 USA; 2, Lita Annenberg Hazen Genome Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 3, Celera Genomics, 850 Lincoln Center Drive, Foster City, California 94494, USA; 4, Plant Biology Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Genome Analysis Group

Klaus Mayer[1]*, Owen White[2]*, Michael Bevan[3], Kai Lemcke[1], Todd H. Creasy[2], Cord Bielke[2], Brian Haas[1], Dirk Haase[1], Rama Maiti[2], Stephen Rudd[1], Jeremy Peterson[2], Heiko Schoof[1], Dimitrij Frishman[1], Burkhard Morgenstern[1], Paulo Zaccaria[1], Maria Ermolaeva [2], Mihaela Pertea[2], John Quackenbush[2], Natalia Volfovsky[2], Dongying Wu[2], Todd M. Lowe[4], Steven L. Salzberg [2] & Hans-Werner Mewes[1]

*1, GSF-Forschungszentrum f. Umwelt u. Gesundheit, Munich Information Center for Protein Sequences, am Max-Planck-Institut f. Biochemie, Am Klopferspitz 18a, D-82152, Germany; 2, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; 3, Molecular Genetics Deartment, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 4, Dept Genetics, Stanford University Medical School, Stanford, California 94305-5120, USA. * These authors contributed equally to this work*

Vol 436|11 August 2005|doi:10.1038/nature03895

nature

ARTICLES

The map-based sequence of the rice genome

International Rice Genome Sequencing Project*

Rice, one of the world's most important food plants, has important syntenic relationships with the other cereal species and is a model plant for the grasses. Here we present a map-based, finished quality sequence that covers 95% of the 389 Mb genome, including virtually all of the euchromatin and two complete centromeres. A total of 37,544 non-transposable-element-related protein-coding genes were identified, of which 71% had a putative homologue in *Arabidopsis*. In a reciprocal analysis, 90% of the *Arabidopsis* proteins had a putative homologue in the predicted rice proteome. Twenty-nine per cent of the 37,544 predicted genes appear in clustered gene families. The number and classes of transposable elements found in the rice genome are consistent with the expansion of syntenic regions in the maize and sorghum genomes. We find evidence for widespread and recurrent gene transfer from the organelles to the nuclear chromosomes. The map-based sequence has proven useful for the identification of genes underlying agronomic traits. The additional single-nucleotide polymorphisms and simple sequence repeats identified in our study should accelerate improvements in rice production.

International Rice Genome Sequencing Project (Participants are arranged by area of contribution and then by institution.)
Physical Maps and Sequencing: Rice Genome Research Program (RGP) Takashi Matsumoto[1], Jianzhong Wu[1], Hiroyuki Kanamori[1], Yuichi Katayose[1], Masaki Fujisawa[1], Nobukazu Namiki[1], Hiroshi Mizuno[1], Kimiko Yamamoto[1], Baltazar A. Antonio[1], Tomoya Baba[1], Katsumi Sakata[1], Yoshiaki Nagamura[1], Hiroyoshi Aoki[1], Koji Arikawa[1], Kohei Arita[1], Takahito Bito[1], Yoshino Chiden[1], Nahoko Fujitsuka[1], Rie Fukunaka[1], Masao Hamada[1], Chizuko Harada[1], Akiko Hayashi[1], Saori Hijishita[1], Mikiko Honda[1], Satomi Hosokawa[1], Yoko Ichikawa[1], Atsuko Idonuma[1], Masumi Iijima[1], Michiko Ikeda[1], Maiko Ikeno[1], Kazue Ito[1], Sachie Ito[1], Tomoko Ito[1], Yuichi Ito[1], Yukiyo Ito[1], Aki Iwabuchi[1], Kozue Kamiya[1], Wataru Karasawa[1], Kanako Kurita[1], Satoshi Katagiri[1], Ari Kikuta[1], Harumi Kobayashi[1], Noriko Kobayashi[1], Kayo Machita[1], Tomoko Maehara[1], Masatoshi Masukawa[1], Tatsumi Mizubayashi[1], Yoshiyuki Mukai[1], Hideki Nagasaki[1], Yuko Nagata[1], Shinji Naito[1], Marina Nakashima[1], Yuko Nakama[1], Yumi Nakamichi[1], Mari Nakamura[1], Ayano Meguro[1], Manami Negishi[1], Isamu Ohta[1], Tomoya Ohta[1], Masako Okamoto[1], Nozomi Ono[1], Shoko Saji[1], Miyuki Sakaguchi[1], Kumiko Sakai[1], Michie Shibata[1], Takanori Shimokawa[1], Jianyu Song[1], Yuka Takazaki[1], Kimihiro Terasawa[1], Mika Tsugane[1], Kumiko Tsuji[1], Shigenori Ueda[1], Kazunori Waki[1], Harumi Yamagata[1], Mayu Yamamoto[1], Shinichi Yamamoto[1], Hiroko Yamane[1], Shoji Yoshiki[1], Rie Yoshihara[1], Kazuko Yukawa[1], Huisun Zhong[1], Masahiro Yano[1], Takuji Sasaki (Principal Investigator)[1]; **The Institute for Genomic Research (TIGR)** Qiaoping Yuan[2], Shu Ouyang[2], Jia Liu[2], Kristine M. Jones[2], Kristen Gansberger[2], Kelly Moffat[2], Jessica Hill[2], Jayati Bera[2], Douglas Fadrosh[2], Shaohua Jin[2], Shivani Johri[2], Mary Kim[2], Larry Overton[2], Matthew Reardon[2], Tamara Tsitrin[2], Hue Vuong[2], Bruce Weaver[2], Anne Ciecko[2], Luke Tallon[2], Jacqueline Jackson[2], Grace Pai[2], Susan Van Aken[2], Terry Utterback[2], Steve Reidmuller[2], Tamara Feldblyum[2], Joseph Hsiao[2], Victoria Zismann[2], Stacey Iobst[2], Aymeric R. de Vazeille[2], C. Robin Buell (Principal Investigator)[2]; **National Center for Gene Research Chinese Academy of Sciences (NCGR)** Kai Ying[3], Ying Li[3], Tingting Lu[3], Yuchen Huang[3], Qiang Zhao[3], Qi Feng[3], Lei Zhang[3], Jingjie Zhu[3], Qijun Weng[3], Jie Mu[3], Yiqi Lu[3], Danlin Fan[3], Yilei Liu[3], Jianping Guan[3], Yujun Zhang[3], Shuliang Yu[3], Xiaohui Liu[3], Yu Zhang[3], Guofan Hong[3], Bin Han (Principal Investigator)[3]; **Genoscope** Nathalie Choisne[4], Nadia Demange[4], Gisela Orjeda[4], Sylvie Samain[4], Laurence Cattolico[4], Eric Pelletier[4], Arnaud Couloux[4], Beatrice Segurens[4], Patrick Wincker[4], Angelique D'Hont[5], Claude Scarpelli[4], Jean Weissenbach[4], Marcel Salanoubat[4], Francis Quetier (Principal Investigator)[4]; **Arizona Genomics Institute (AGI) and Arizona Genomics Computational Laboratory (AGCoI)** Yeisoo Yu[6], Hye Ran Kim[6], Teri Rambo[6], Jennifer Currie[6], Kristi Collura[6], Meizhong Luo[6], Tae-Jin Yang[6], Jetty S. S. Ammiraju[6], Friedrich Engler[6], Carol Soderlund[6], Rod A. Wing (Principal Investigator)[6]; **Cold Spring Harbor Laboratory (CSHL)** Lance E. Palmer[7], Melissa de la Bastide[7], Lori Spiegel[7], Lidia Nascimento[7], Theresa Zutavern[7], Andrew O'Shaughnessy[7], Sujit Dike[7], Neilay Dedhia[7], Raymond Preston[7], Vivekanand Balija[7], W. Richard McCombie (Principal Investigator)[7]; **Academia Sinica Plant Genome Center (ASPGC)** Teh-Yuan Chow[8], Hong-Hwa Chen[9], Mei-Chu Chung[8], Ching-San Chen[8], Jei-Fu Shaw[8], Hong-Pang Wu[8], Kwang-Jen Hsiao[10], Ya-Ting Chao[8], Mu-kuei Chu[8], Chia-Hsiung Cheng[8], Ai-Ling Hour[8], Pei-Fang Lee[8], Shu-Jen Lin[8], Yao-Cheng Lin[8], John-Yu Liou[8], Shu-Mei Liu[8], Yue-Ie Hsing (Principal Investigator)[8]; **Indian Initiative for Rice Genome Sequencing (IIRGS), University of Delhi South Campus (UDSC)** S. Raghuvanshi[11], A. Mohanty[11], A. K. Bharti[11,13], A. Gaur[11], V. Gupta[11], D. Kumar[11], V. Ravi[11], S. Vij[11], A. Kapur[11], Parul Khurana[11], Paramjit Khurana[11], J. P. Khurana[11], A. K. Tyagi (Principal Investigator)[11]; **Indian Initiative for Rice Genome Sequencing (IIRGS), Indian Agricultural Research Institute (IARI)** K. Gaikwad[12], A. Singh[12], V. Dalal[12], S. Srivastava[12], A. Dixit[12], A. K. Pal[12], I. A. Ghazi[12], M. Yadav[12], A. Pandit[12], A. Bhargava[12], K. Sureshbabu[12], K. Batra[12], T. R. Sharma[12], T. Mohapatra[12], N. K. Singh (Principal Investigator)[12]; **Plant Genome Initiative at Rutgers (PGIR)** Joachim Messing (Principal Investigator)[13], Amy Bronzino Nelson[13], Galina Fuks[13], Steve Kavchok[13], Gladys Keizer[13], Eric Linton Victor Llaca[13], Rentao Song[13], Bahattin Tanyolac[13], Steve Young[13]; **Korea Rice Genome Research Program (KRGRP)** Kim Ho-Il[14], Jang Ho Hahn (Principal Investigator)[14]; **National Center for Genetic Engineering and Biotechnology (BIOTEC)** G. Sangsakoo[15], A. Vanavichit (Principal Investigator)[15]; **Brazilian Rice Genome Initiative (BRIGI)** Luiz Anderson Teixeira de Mattos[16], Paulo Dejalma Zimmer[16], Gaspar Malone[16], Odir Dellagostin[16], Antonio Costa de Oliveira (Principal Investigator)[16]; **John Innes Centre (JIC)** Michael Bevan[17], Ian Bancroft[17]; **Washington University School of Medicine Genome Sequencing Center** Pat Minx[18], Holly Cordum[18], Richard Wilson[18]; **University of Wisconsin-Madison** Zhukuan Cheng[19], Weiwei Jin[19], Jiming Jiang[19], Sally Ann Leong[20]

Annotation and Analysis: Hisakazu Iwama[21], Takashi Gojobori[21,22], Takeshi Itoh[22,23], Yoshihito Niimura[24], Yasuyuki Fujii[25], Takuya Habara[25], Hiroaki Sakai[23,25], Yoshiharu Sato[22], Greg Wilson[26], Kiran Kumar[27], Susan McCouch[26], Nikoleta Juretic[28], Douglas Hoen[28], Stephen Wright[29], Richard Bruskiewich[30], Thomas Bureau[28], Akio Miyao[23], Hirohiko Hirochika[23], Tomotaro Nishikawa[23], Koh-ichi Kadowaki[23] & Masahiro Sugiura[31]

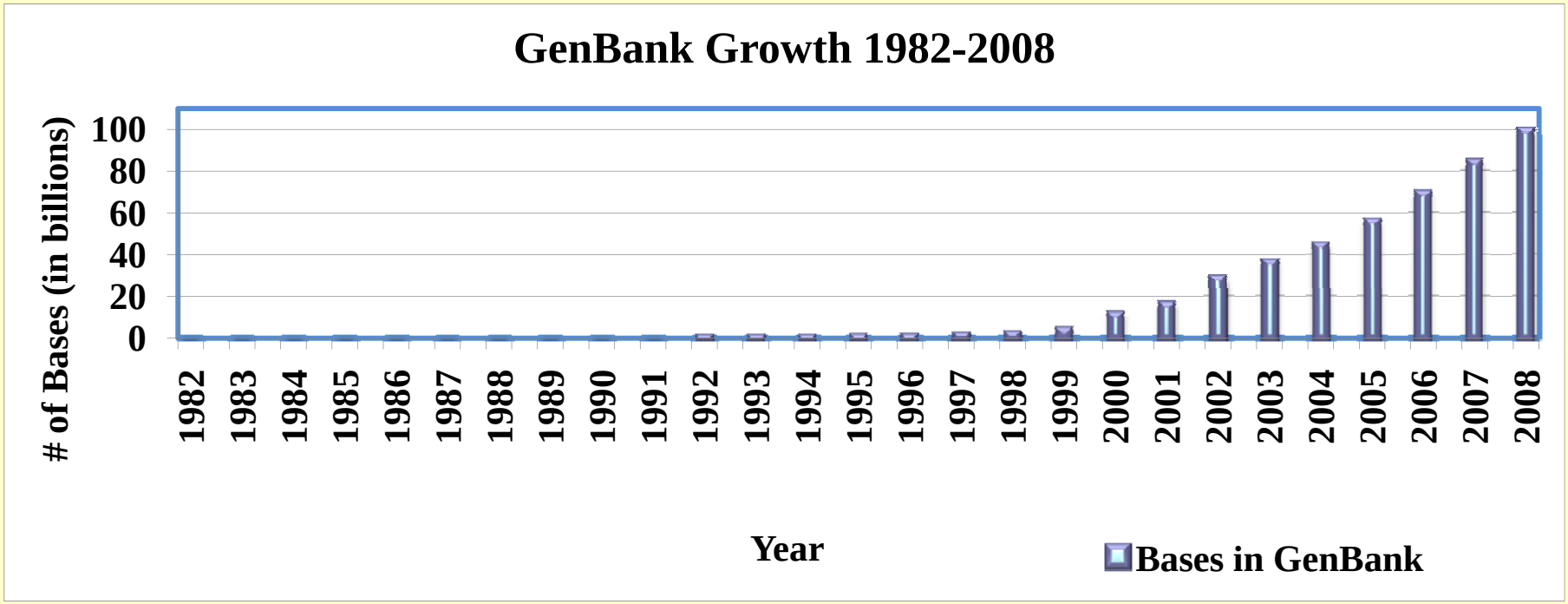
Coordination: Benjamin Burr[32]

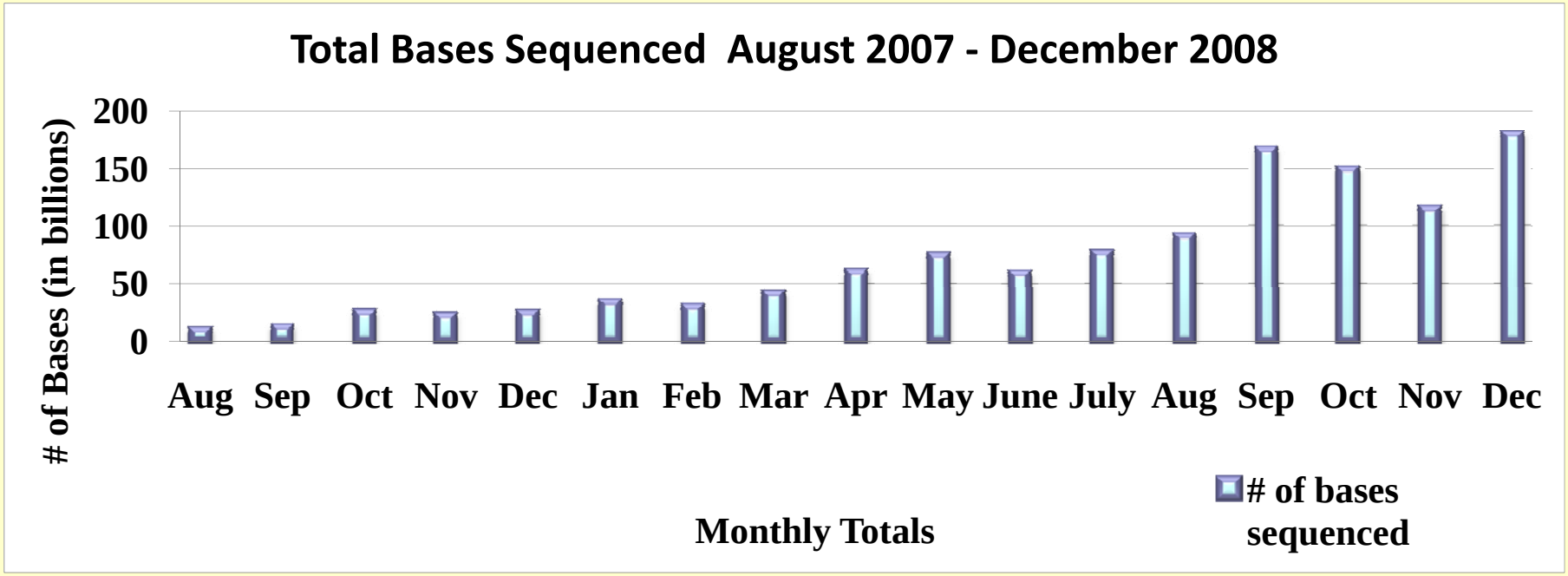
DNA sequence variants among individual organisms of a species are responsible for much of the variation in those organisms' phenotypes

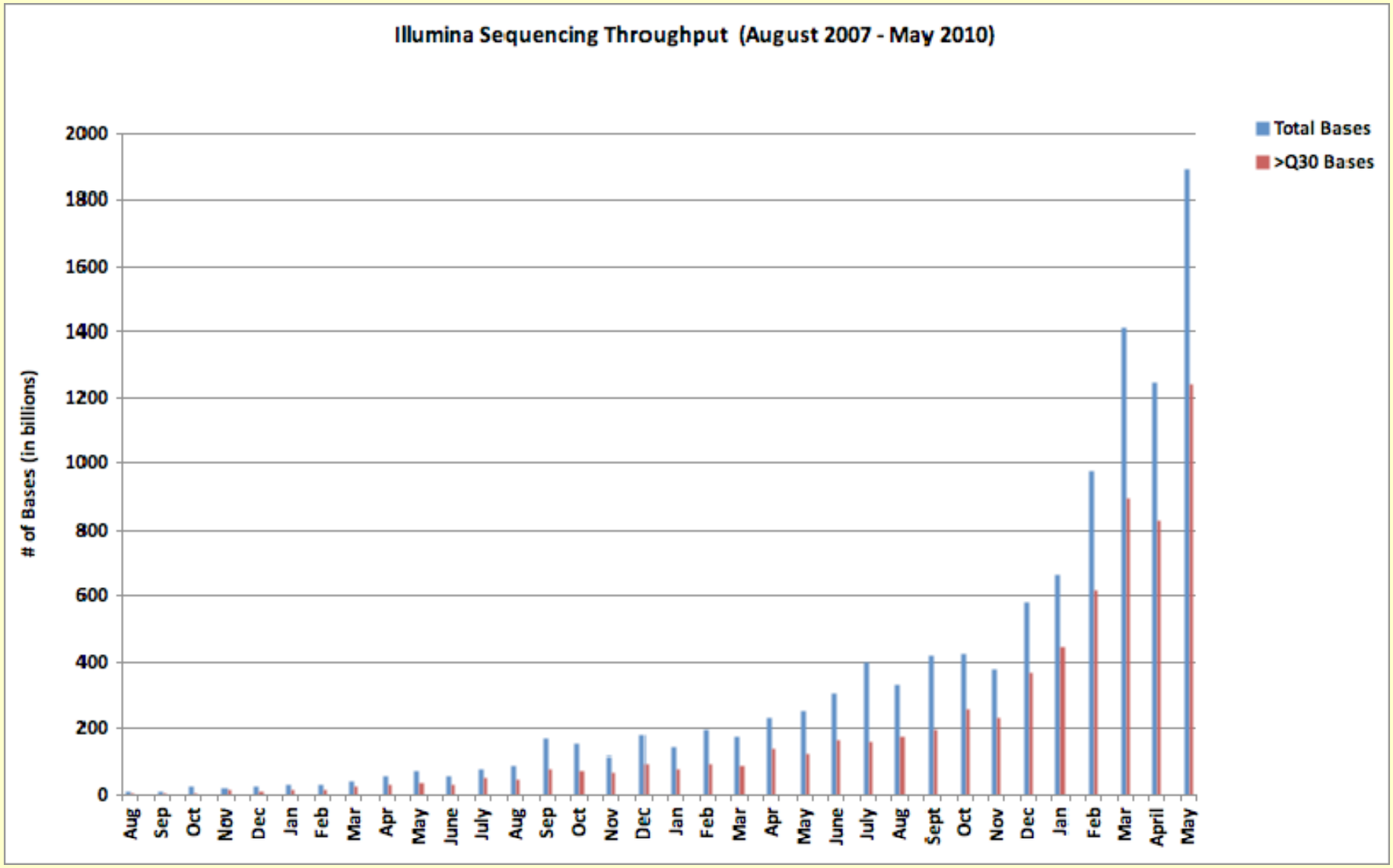
- Susceptibility to disease
- Susceptibility to environmental insults
- Longevity
- Behavior
- Biological timing

DNA Sequencing Capability (per person)

- 1980: 0.1-1 kb per year
- 1985: 1-5 kb per year
- 1990: 25-50 kb per year
- 1996: 100-200 kb per year
- 2000: 500-1000 kb per year
- 2007: 30,000,000 -60,000,000 kb per year
- 2009: 100,000,000-1,000,000,000 kb per year
- 2011: ~4,000,000,000 kb per year?? Or more?


GenBank Growth 1982-2008
# of Bases (in billions)
100
80
60
40
20
0
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
Year
Bases in GenBank


Total Bases Sequenced August 2007 - December 2008
# of Bases (in billions)
200
150
100
50
0
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
June
July
Aug
Sep
Oct
Nov
Dec
Monthly Totals
# of bases sequenced


Illumina Sequencing Throughput (August 2007 - May 2010)
Total Bases
>Q30 Bases
# of Bases (in billions)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
June
July
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
June
July
Aug
Sept
Oct
Nov
Dec
Jan
Feb
Mar
April
May

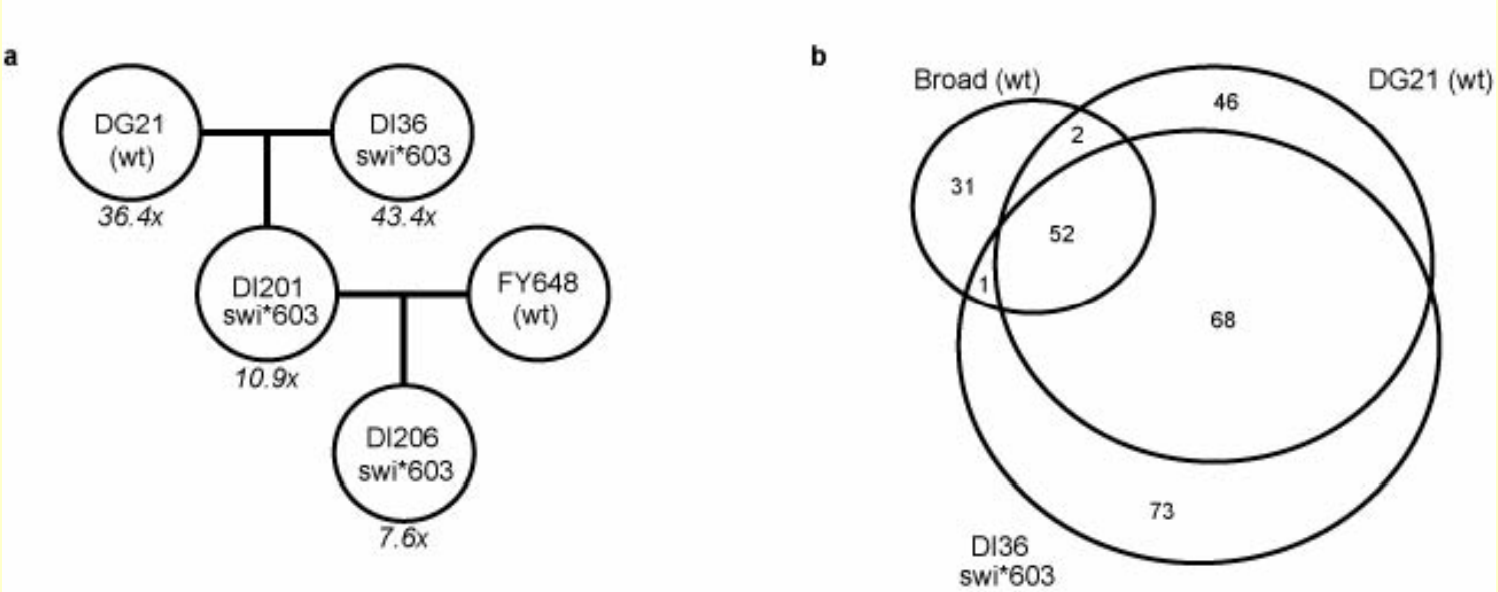
How much is the change in technology impacting the way we do biology?

Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing

Danielle V. Irvine, Derek B. Goto, Matthew W. Vaughn, Yukinobu Nakaseko, W. Richard McCombie, Mitsuhiro Yanagida, and Rob Martienssen

Genome Res. 2009 June; 19(6): 1077–1083.

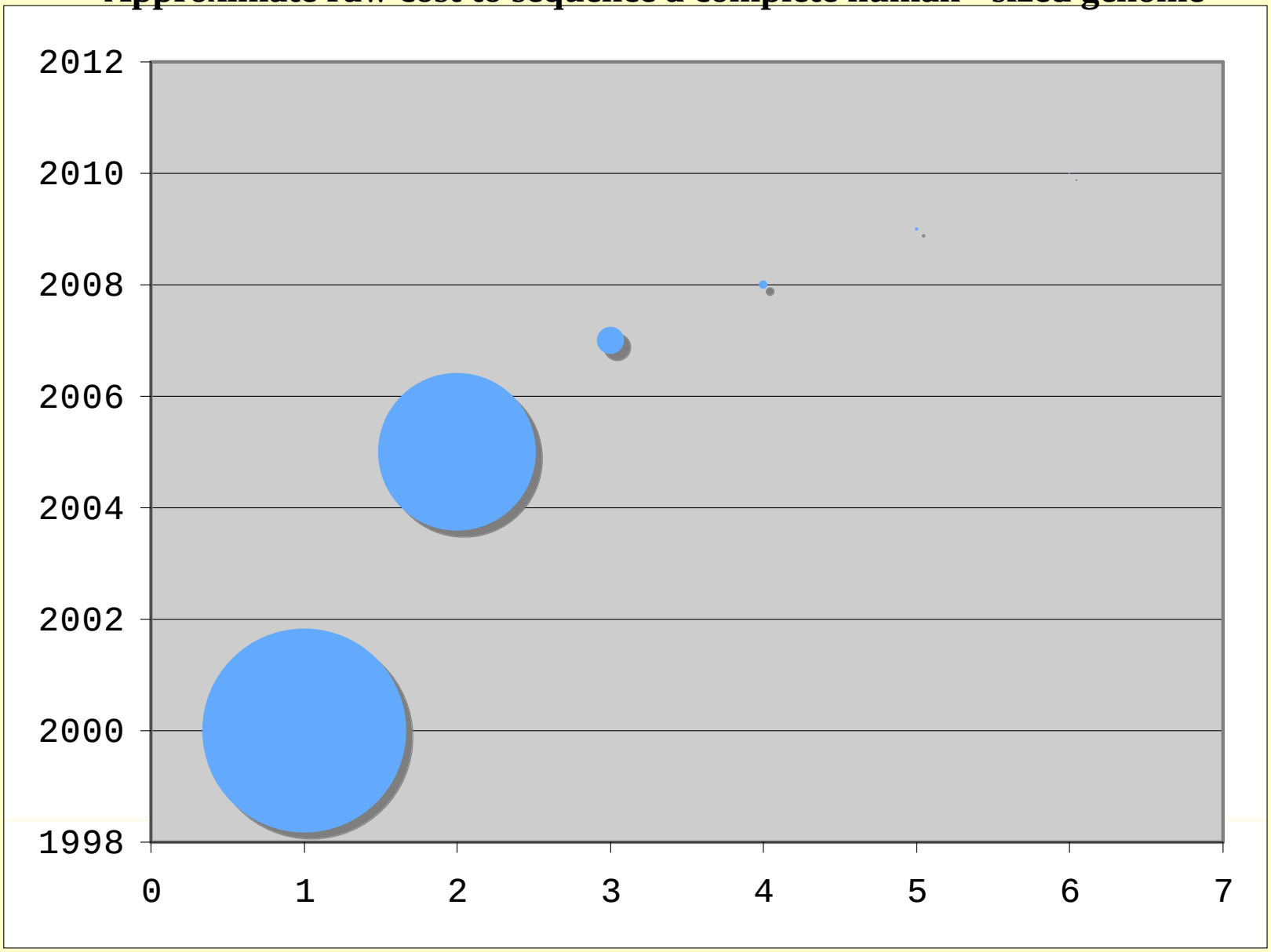
SNP detection by Illumina sequencing



- 73 candidate SNPs, only 22 inherited in backcross, 6 non-silent substitutions in genes

To put things in perspective…….

Approximate raw cost to sequence a complete human - sized genome



Genetics has become genomics: An example two years later

- Bipolar disorder effects 1%-2% of the population -- highly heritable
- Bipolar and related disorders are second to cardiovascular diseases in terms of years of lost productive life
- Family with 135 members (5 generations), 27 of whom have bipolar or major recurrent depression

While re-sequencing is the first major application of the new technology - de novo sequencing is being more frequently attempted

Current Status of plant projects

Project	Coverage
Nipponbare	43X
Peach	30X
Pigeon pea	79X
Chinese spring wheat arm 3DL	75X
Lemna gibba 131 (duckweed)	42X
Cocoa	72X
L. pimpinellifolium	38X

Pure Illumina assembly results (no mate pairs yet)

Project	N50
Nipponbare	~3 kb
Pigeon pea	~3.5 kb
Lemna gibba 131 (duckweed)	~0.5 kb !!
L. pimpinellifolium	~3.5 kb

What will we be seeing in the future?

Comparison with Genome Analyzer

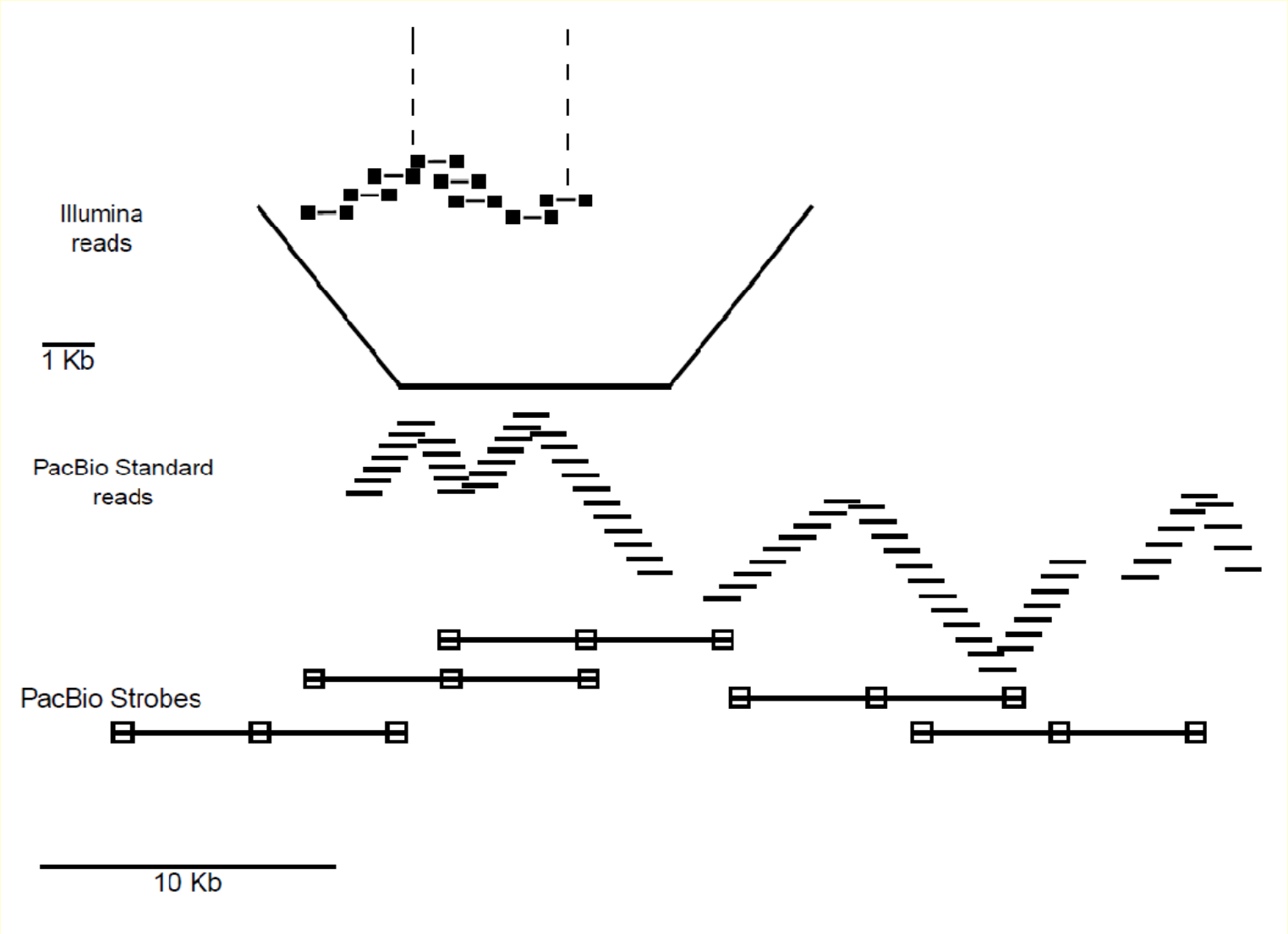
	HiSeq 2000 (at launch)	GA_{IIx} (at 50G)	GA_{IIx} (at 95G)
Gb per run	150-200	50	95
Gb per day	20-25	5	7
Cluster density in KClusters/mm²**	260-350	490	620
Read length	2 x100	2 x100	2 x150
Available surface area (mm²)*	2880	510	510

*GA_{IIx} with single surface, single FC, HiSeq 2000 with dual surface, dual FC
**Clusters passing filter

Gene Discovery in Wheat

W. Richard McCombie, Doreen Ware and Robert Martienssen

NSF Award: IOS-1032105


Illumina
reads
1 Kb
PacBio Standard
reads
PacBio Strobes
10 Kb

We need to think about how we will do genetics - and biology in general - when it costs less than $1000 to sequence a maize sized genome or $100 to sequence a rice strain (in 18-24 months)?

Conclusions

- Next generation sequencing technology is certain to rapidly and fundamentally change our understanding of the salient variability associated with complex phenotypes
- International collaborations to exhaustively characterize the genomes of a large number of individual plants will have an enormous impact on our ability to correlate genome structure with phenotype (disease resistance, ability to grow in poor soil, etc)
- This will soon allow us to do predictive plant breeding by computer
- This has the capability to vastly improve our agricultural capabilities, particularly on marginal land
- Continued and expanded international sharing of raw data is essential to maximize progress and benefits
- People need this - but they aren’t ready for it a almost any level

Acknowledgements

- Melissa Kramer
- Jennifer Parla
- Maureen Bell
- Danea Rebolini
- Laura Cardone
- Stephanie Muller
- Elena Ghiban
- Darrin Lewis
- Thomas Caulfield
- Magda Gierszewska
- Senum Mavruk
- Patricia Mocombe
- Sneh Lata
- JC Yao
- Ian Grabill

- University of Edinburgh
  - David Porteous
  - Douglas Blackwood
  - Walter Muir
  - Ian Deary
  - Pippa Thomson
  - William Hennah
  - Katherine Evans

- Greg Hannon
- Emily Hodges
- Sydney Gary
- Jim Watson
- Doreen Ware
- Robert Martienssen

- Stanley Family and Stanley Medical Research Institute
- National Institutes of Health
- Illumina, Inc.

- The Arabidopsis Genome Initiative

- The International Rice Genome Sequencing Project

- National Science Foundation

- Mary Clutter
- DeLill Nasser
- Machi Dilworth
- Jane Silverthorne