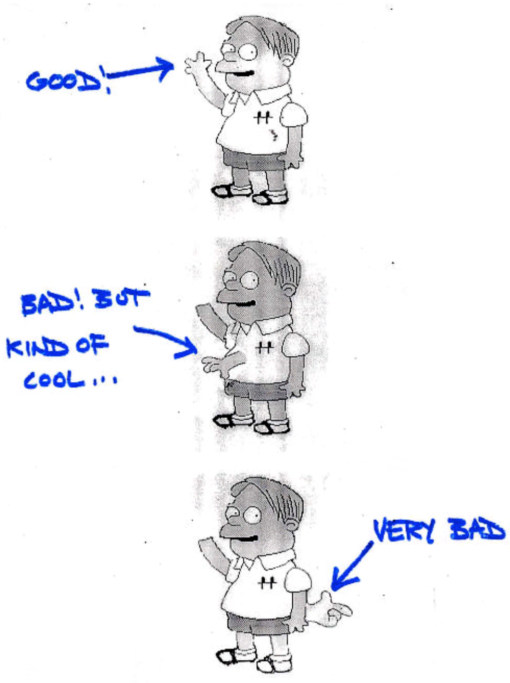
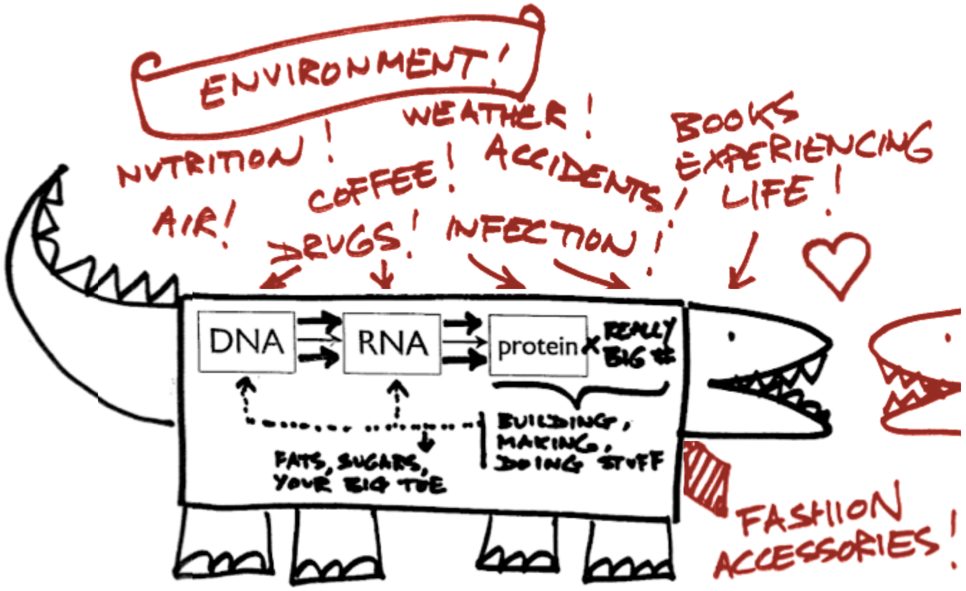
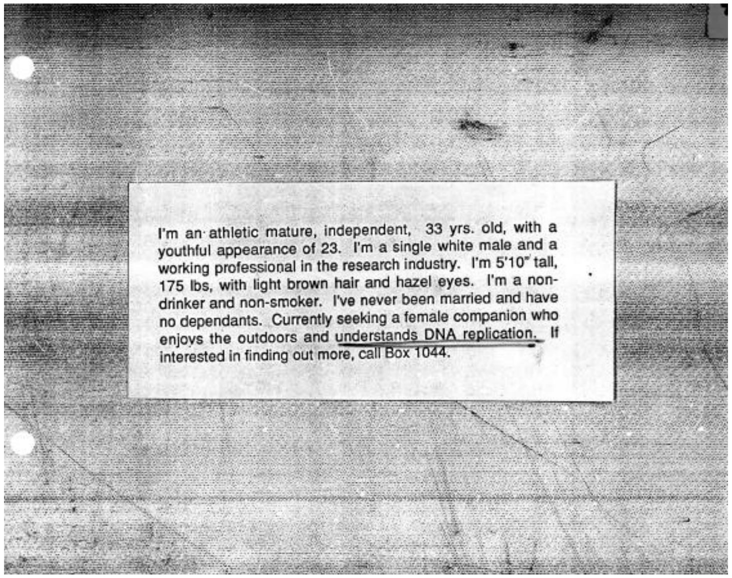





I'm an athletic mature, independent, 33 yrs. old, with a youthful appearance of 23. I'm a single white male and a working professional in the research industry. I'm 5'10" tall, 175 lbs, with light brown hair and hazel eyes. I'm a non-drinker and non-smoker. I've never been married and have no dependants. Currently seeking a female companion who enjoys the outdoors and understands DNA replication. If interested in finding out more, call Box 1044.

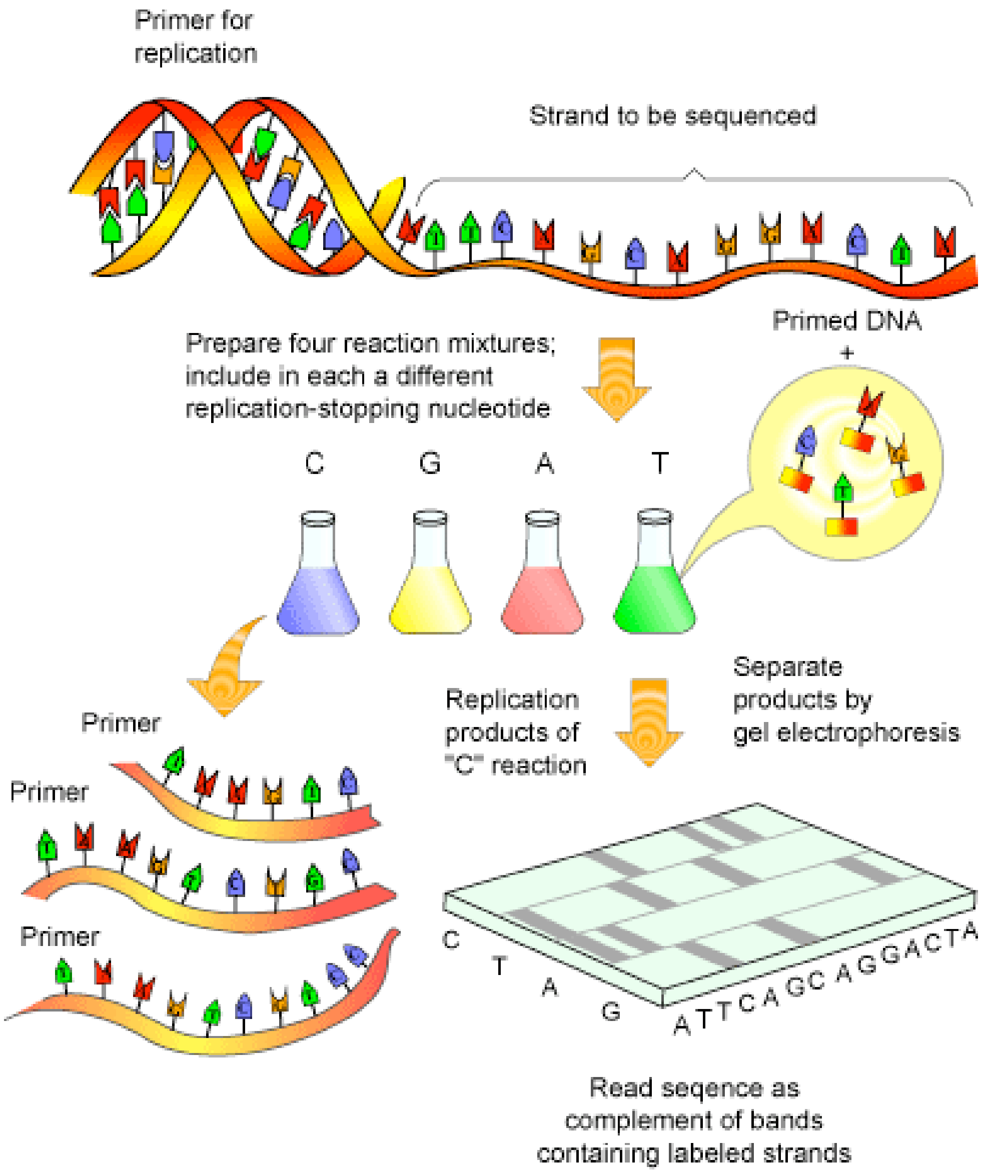
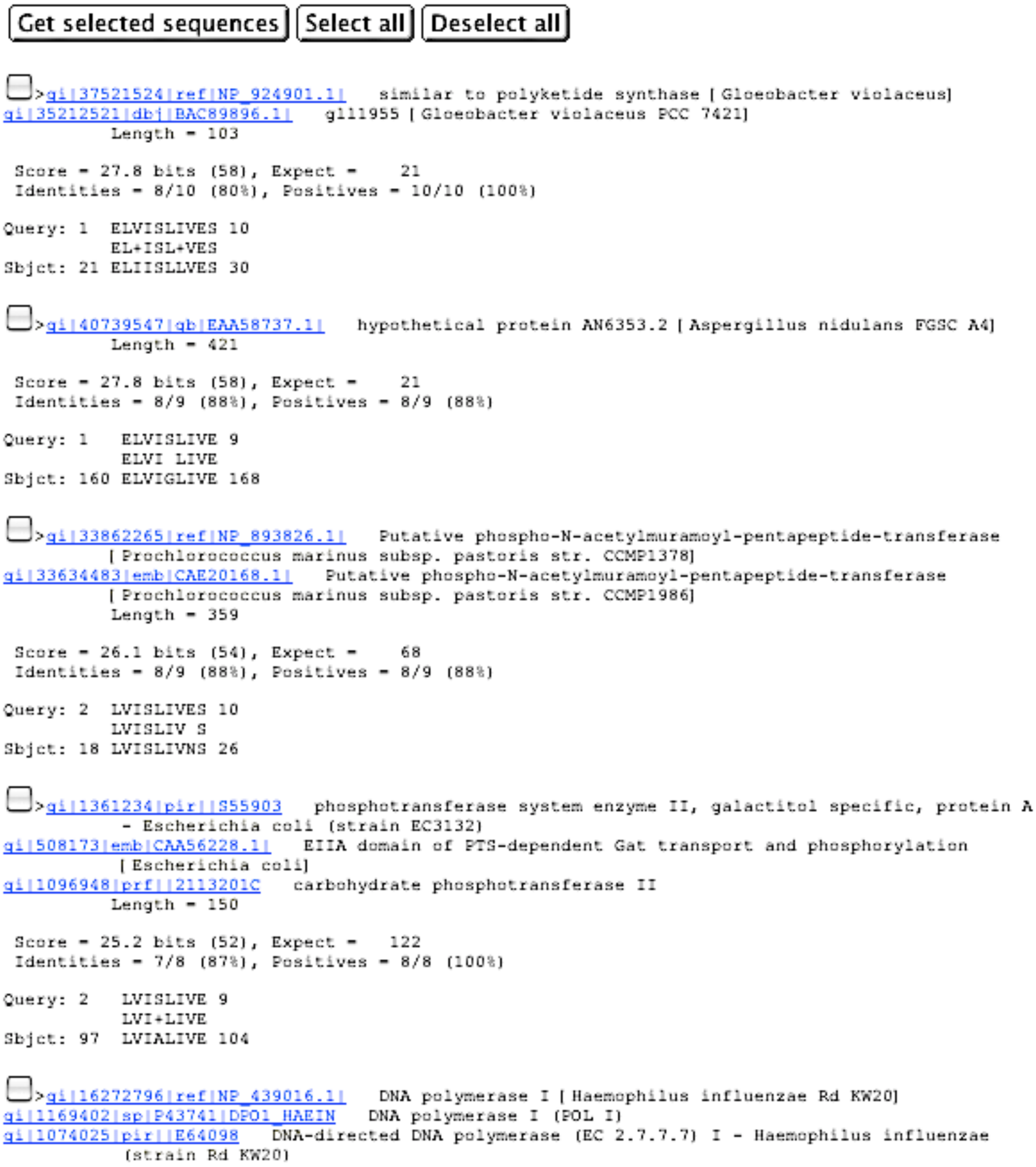
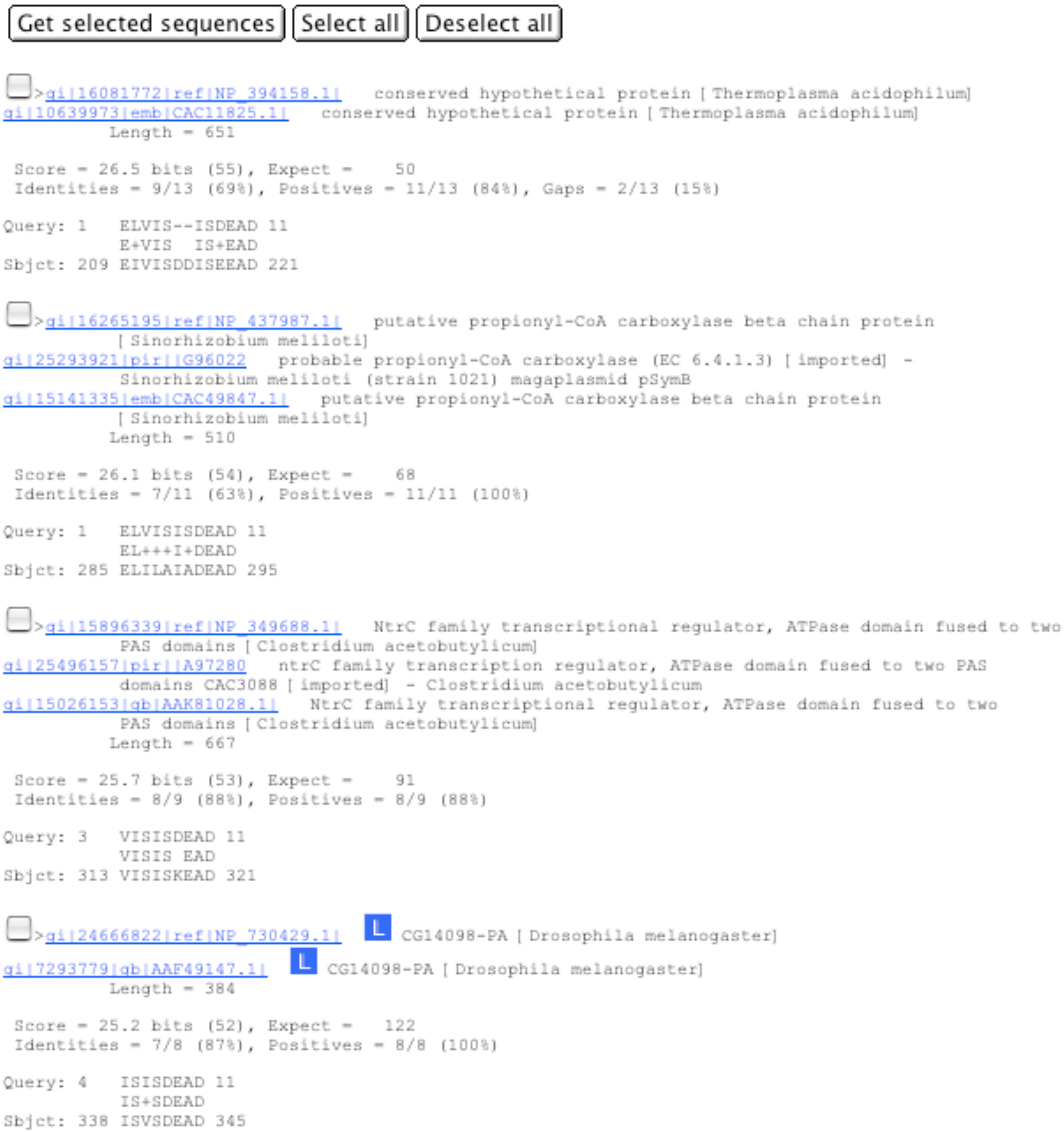

Primer for replication
Strand to be sequenced
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA
+
C
G
A
T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
Read seqence as complement of bands containing labeled strands

Figure 1. *The Sanger sequencing reaction. Single stranded DNA is amplified in the presence of fluorescently labelled ddNTPs that serve to terminate the reaction and label all the fragments of DNA produced. The fragments of DNA are then separated via polyacrylamide gel electrophoresis and the sequence read using a laser beam and computer.*

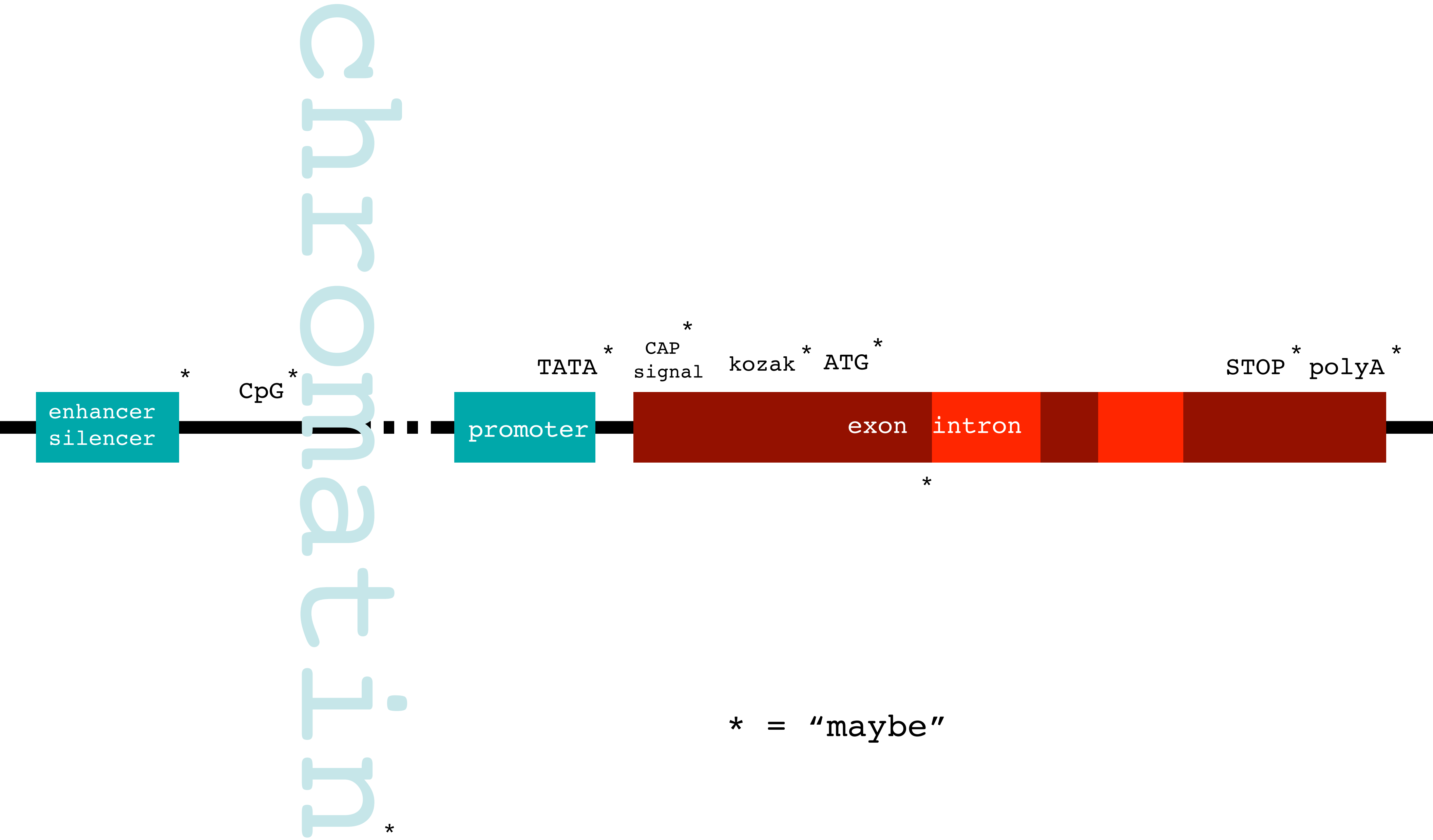
source: http://www.scq.ubc.ca/genome-projects-uncovering-the-blueprints-of-biology/



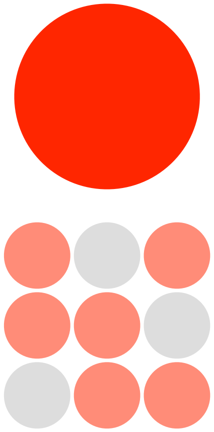
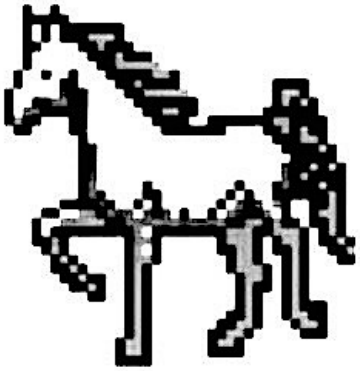
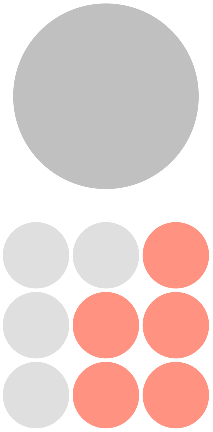
```
□>gi|37521524|ref|NP_924901.1|   similar to polyketide synthase [Gloeobacter violaceus]
gi|35212521|dbj|BAC89896.1|   gll1955 [Gloeobacter violaceus PCC 7421]
          Length = 103

 Score = 27.8 bits (58), Expect =    21
 Identities = 8/10 (80%), Positives = 10/10 (100%)

Query: 1  ELVISLIVES 10
          EL+ISL+VES
Sbjct: 21 ELIISLLVES 30

□>gi|40739547|gb|EAA58737.1|   hypothetical protein AN6353.2 [Aspergillus nidulans FGSC A4]
          Length = 421

 Score = 27.8 bits (58), Expect =    21
 Identities = 8/9 (88%), Positives = 8/9 (88%)

Query: 1   ELVISLIVE 9
            ELVI LIVE
Sbjct: 160 ELVIGLIVE 168

□>gi|33862265|ref|NP_893826.1|   Putative phospho-N-acetylmuramoyl-pentapeptide-transferase
          [Prochlorococcus marinus subsp. pastoris str. CCMP1378]
gi|33634483|emb|CAE20168.1|   Putative phospho-N-acetylmuramoyl-pentapeptide-transferase
          [Prochlorococcus marinus subsp. pastoris str. CCMP1986]
          Length = 359

 Score = 26.1 bits (54), Expect =    68
 Identities = 8/9 (88%), Positives = 8/9 (88%)

Query: 2  LVISLIVES 10
          LVISLIV S
Sbjct: 18 LVISLIVNS 26

□>gi|1361234|pir||S55903   phosphotransferase system enzyme II, galactitol specific, protein A
           - Escherichia coli (strain EC3132)
gi|508173|emb|CAA56228.1|   EIIA domain of PTS-dependent Gat transport and phosphorylation
           [Escherichia coli]
gi|1096948|prf||2113201C   carbohydrate phosphotransferase II
          Length = 150

 Score = 25.2 bits (52), Expect =   122
 Identities = 7/8 (87%), Positives = 8/8 (100%)

Query: 2   LVISLIVE 9
            LVI+LIVE
Sbjct: 97  LVIALIVE 104

□>gi|16272796|ref|NP_439016.1|   DNA polymerase I [Haemophilus influenzae Rd KW20]
gi|1169402|sp|P43741|DPO1_HAEIN   DNA polymerase I (POL I)
gi|1074025|pir||E64098   DNA-directed DNA polymerase (EC 2.7.7.7) I - Haemophilus influenzae
           (strain Rd KW20)
```



```
□>gi|16081772|ref|NP_394158.1|   conserved hypothetical protein [Thermoplasma acidophilum]
gi|10639973|emb|CAC11825.1|   conserved hypothetical protein [Thermoplasma acidophilum]
          Length = 651

 Score = 26.5 bits (55), Expect =    50
 Identities = 9/13 (69%), Positives = 11/13 (84%), Gaps = 2/13 (15%)

Query: 1   ELVIS--ISDEAD 11
           E+VIS  IS+EAD
Sbjct: 209 EIVISDDISEEAD 221

□>gi|16265195|ref|NP_437987.1|   putative propionyl-CoA carboxylase beta chain protein
           [Sinorhizobium meliloti]
gi|25293921|pir||G96022   probable propionyl-CoA carboxylase (EC 6.4.1.3) [imported] -
           Sinorhizobium meliloti (strain 1021) magaplasmid pSymB
gi|15141335|emb|CAC49847.1|   putative propionyl-CoA carboxylase beta chain protein
           [Sinorhizobium meliloti]
          Length = 510

 Score = 26.1 bits (54), Expect =    68
 Identities = 7/11 (63%), Positives = 11/11 (100%)

Query: 1   ELVISISDEAD 11
           EL+++I+DEAD
Sbjct: 285 ELILAIADEAD 295

□>gi|15896339|ref|NP_349688.1|   NtrC family transcriptional regulator, ATPase domain fused to two
           PAS domains [Clostridium acetobutylicum]
gi|25496157|pir||A97280   ntrC family transcription regulator, ATPase domain fused to two PAS
           domains CAC3088 [imported] - Clostridium acetobutylicum
gi|15026153|gb|AAK81028.1|   NtrC family transcriptional regulator, ATPase domain fused to two
           PAS domains [Clostridium acetobutylicum]
          Length = 667

 Score = 25.7 bits (53), Expect =    91
 Identities = 8/9 (88%), Positives = 8/9 (88%)

Query: 3   VISISDEAD 11
           VISIS EAD
Sbjct: 313 VISISKEAD 321

□>gi|24666822|ref|NP_730429.1|  L  CG14098-PA [Drosophila melanogaster]
gi|7293779|gb|AAF49147.1|  L  CG14098-PA [Drosophila melanogaster]
          Length = 384

 Score = 25.2 bits (52), Expect =   122
 Identities = 7/8 (87%), Positives = 8/8 (100%)

Query: 4   ISISDEAD 11
           IS+SDEAD
Sbjct: 338 ISVSDEAD 345
```

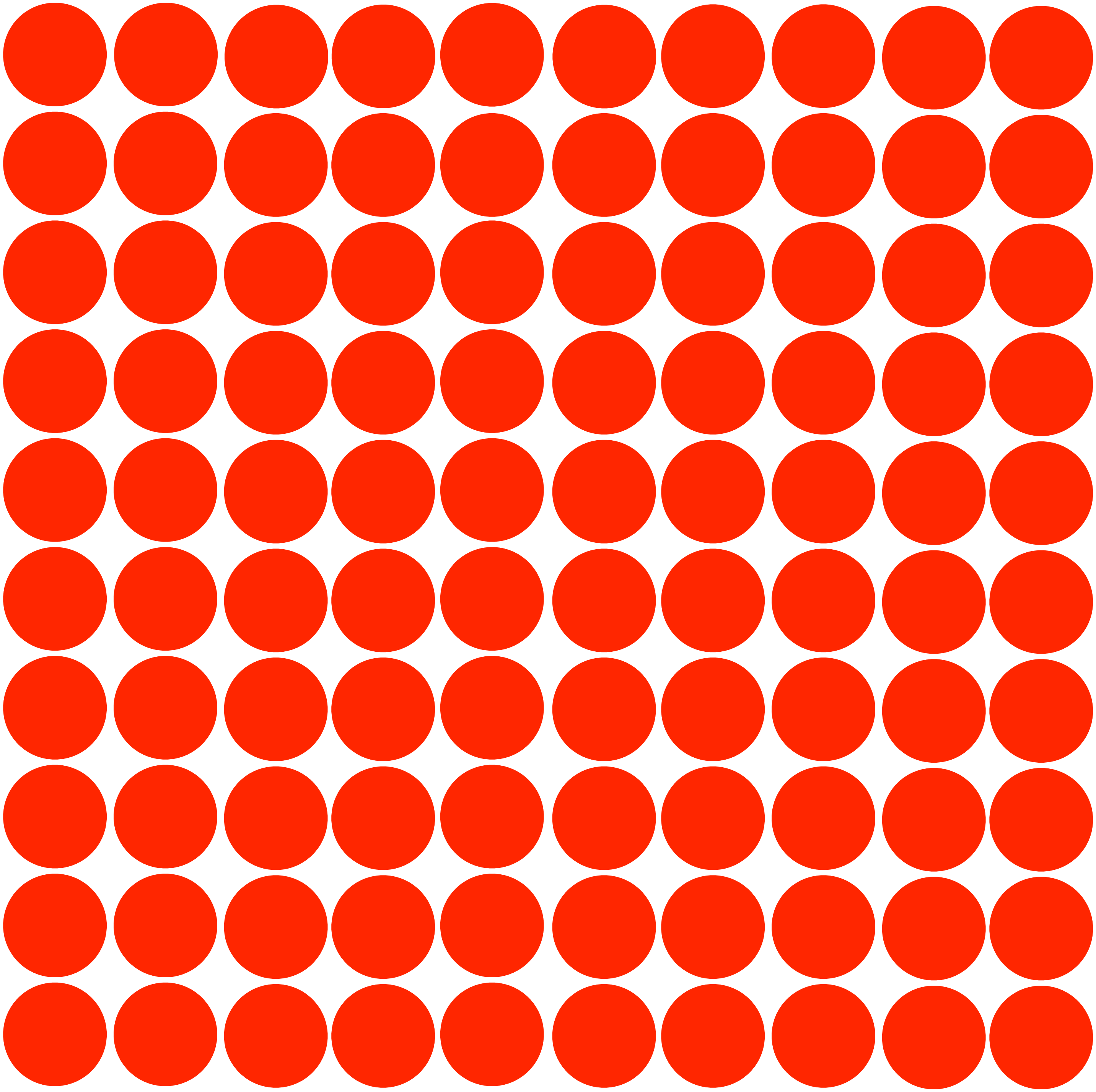

enhancer
silencer
*
CpG *
promoter
TATA *
*
CAP
signal
kozak *
ATG *
exon
intron
*
STOP *
polyA *
chromatin
*

* = “maybe”

PHENOTYPE	PROTEINS	GENES	DNA SEQ
	+ + dQ –		ATCG TAGC
	– – dQ +		ATGC TACG

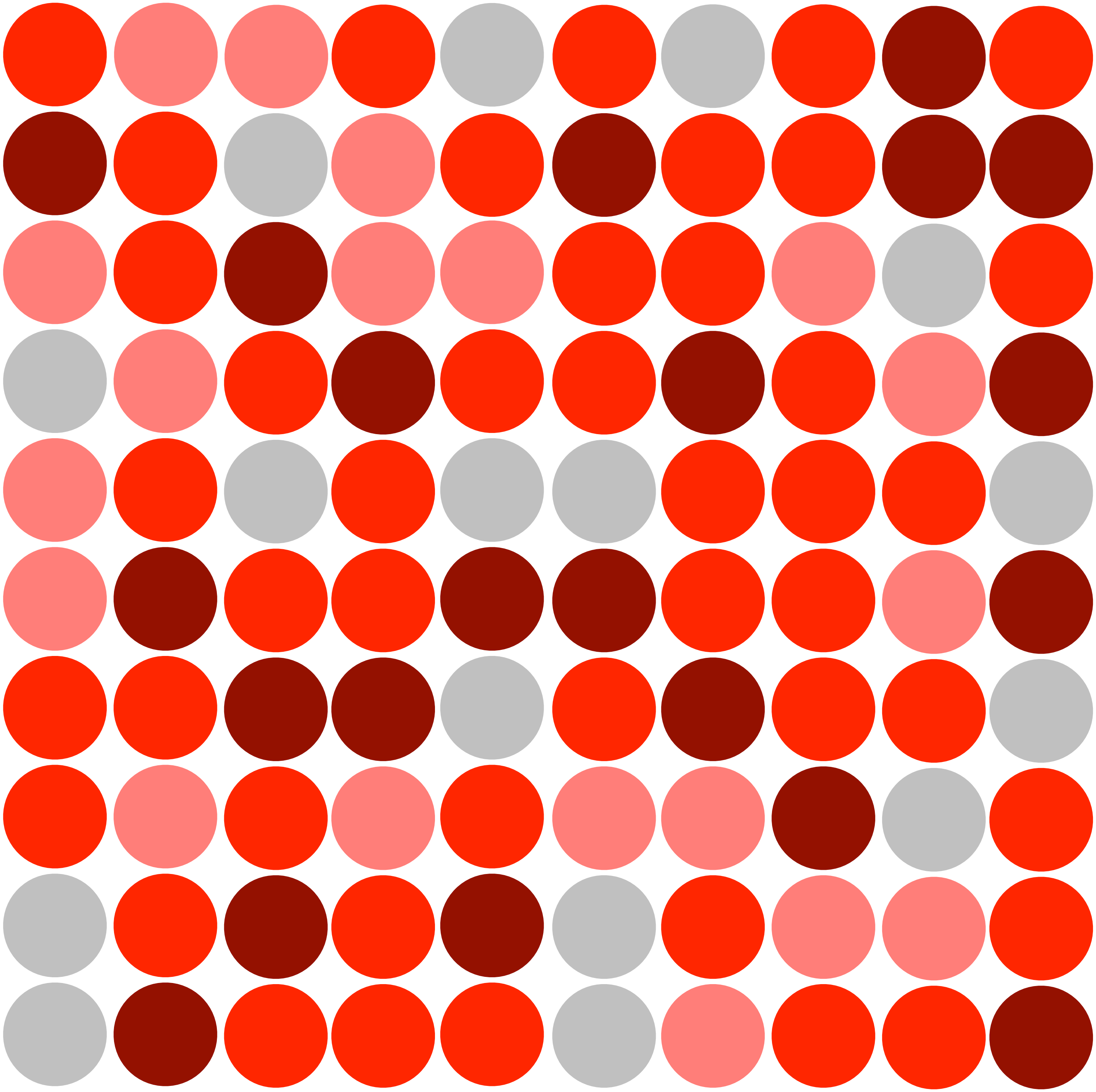
http://miserycat.deviantart.com/art/Free-Unicorn-Icon-180216460?q=&qo=

MOLECULAR BIOLOGY

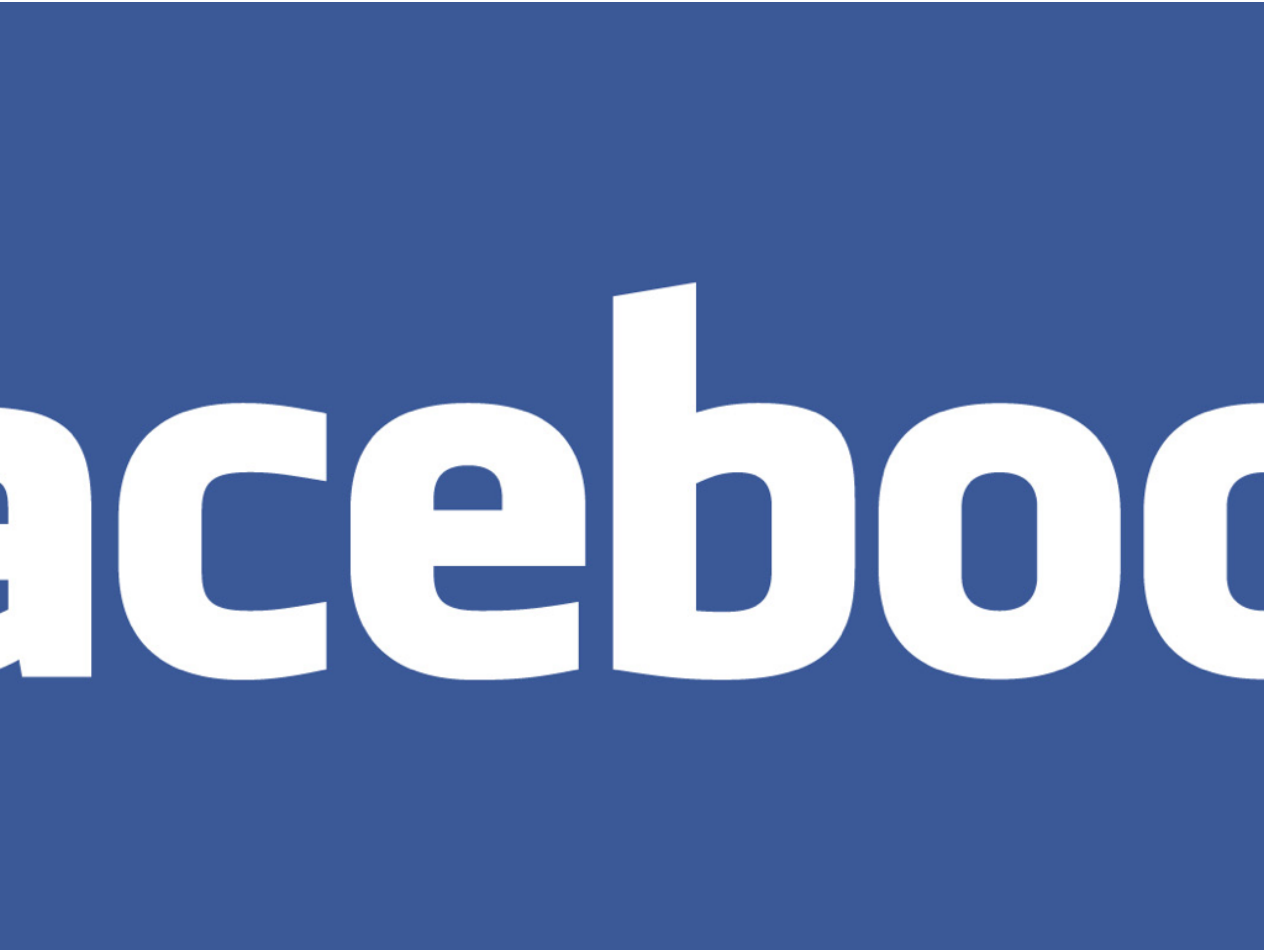


http://devisional.com/

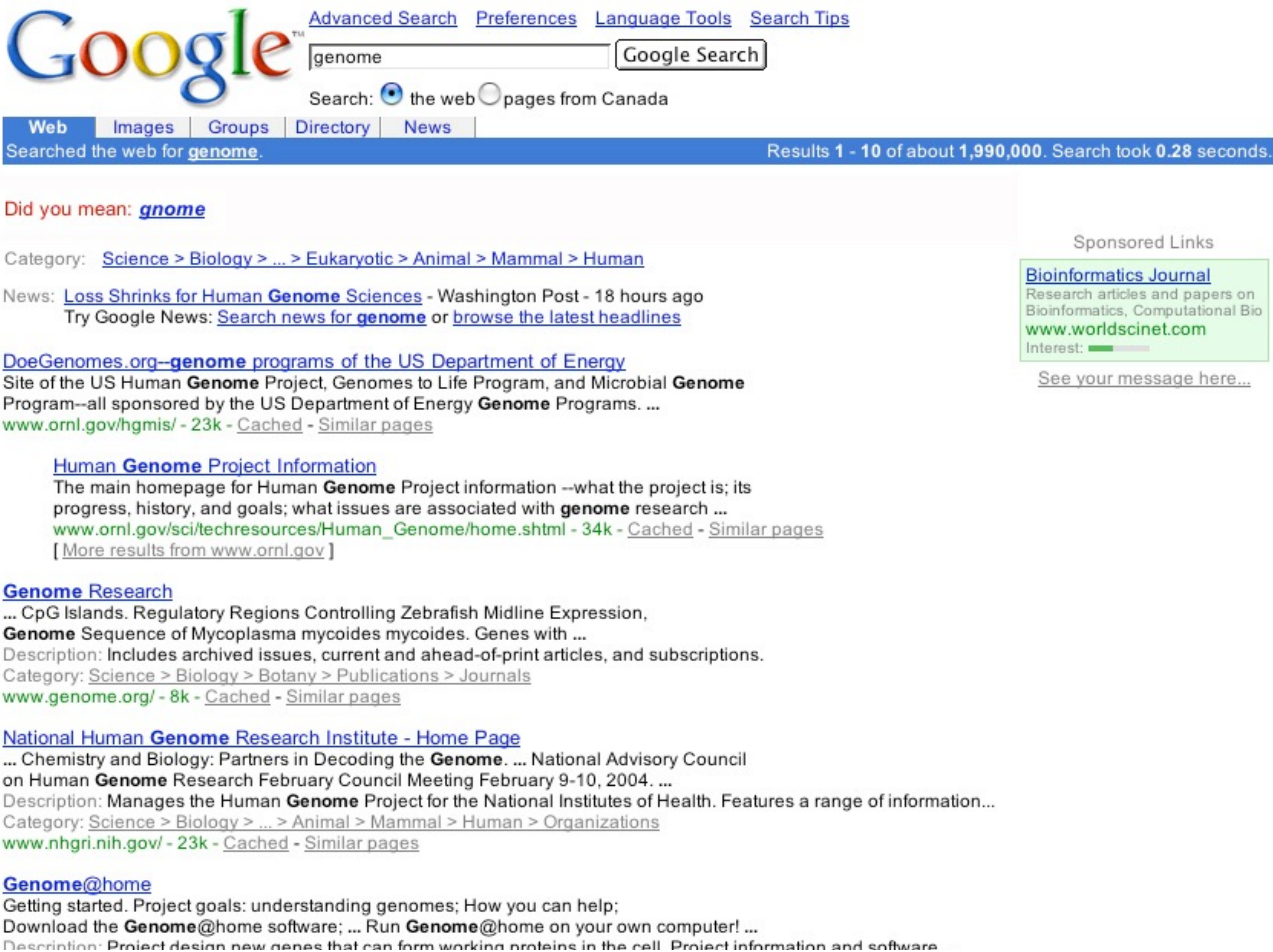
MOLECULAR BIOLOGY - VARIATION IS KEY!



http://devisional.com/


aceboo

THE (HUMAN) GENOME



Searched the web for **genome**. Results **1** - **10** of about **1,990,000**. Search took **0.28** seconds.

Did you mean: ***gnome***

Category: Science > Biology > ... > Eukaryotic > Animal > Mammal > Human

News: Loss Shrinks for Human **Genome** Sciences - Washington Post - 18 hours ago
Try Google News: Search news for **genome** or browse the latest headlines

DoeGenomes.org--**genome** programs of the US Department of Energy
Site of the US Human **Genome** Project, Genomes to Life Program, and Microbial **Genome** Program--all sponsored by the US Department of Energy **Genome** Programs. ...
www.ornl.gov/hgmis/ - 23k - Cached - Similar pages

Human **Genome** Project Information
The main homepage for Human **Genome** Project information --what the project is; its progress, history, and goals; what issues are associated with **genome** research ...
www.ornl.gov/sci/techresources/Human_Genome/home.shtml - 34k - Cached - Similar pages
[More results from www.ornl.gov]

Genome Research
... CpG Islands. Regulatory Regions Controlling Zebrafish Midline Expression, **Genome** Sequence of Mycoplasma mycoides mycoides. Genes with ...
Description: Includes archived issues, current and ahead-of-print articles, and subscriptions.
Category: Science > Biology > Botany > Publications > Journals
www.genome.org/ - 8k - Cached - Similar pages

National Human **Genome** Research Institute - Home Page
... Chemistry and Biology: Partners in Decoding the **Genome**. ... National Advisory Council on Human **Genome** Research February Council Meeting February 9-10, 2004. ...
Description: Manages the Human **Genome** Project for the National Institutes of Health. Features a range of information...
Category: Science > Biology > ... > Animal > Mammal > Human > Organizations
www.nhgri.nih.gov/ - 23k - Cached - Similar pages

Genome@home
Getting started. Project goals: understanding genomes; How you can help; Download the **Genome**@home software; ... Run **Genome**@home on your own computer! ...
Description: Project design new genes that can form working proteins in the cell. Project information and software

Sponsored Links

Bioinformatics Journal
Research articles and papers on Bioinformatics, Computational Bio
www.worldscinet.com
Interest:

See your message here...

GENOME: In modern molecular biology and genetics, the **genome** is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and the non-coding sequences of the DNA/RNA

GENOMICS: is a discipline in genetics concerning the study of the genomes of organisms

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

The sequence of the human genome is of interest in several respects. It is the largest genome to be extensively sequenced so far, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, it is the genome of our own species.

Much work remains to be done to produce a complete finished sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

- The genomic landscape shows marked variation in the distribution of a number of features, including genes, transposable elements, GC content, CpG islands and recombination rate. This gives us important clues about function. For example, the developmentally important HOX gene clusters are the most repeat-poor regions of the human genome, probably reflecting the very complex coordinate regulation of the genes in the clusters.
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark G-bands' in karyotypes.
- Recombination rates tend to be much higher in distal regions (around 20 megabases (Mb)) of chromosomes and on shorter chromosome arms in general, in a pattern that promotes the occurrence of at least one crossover per chromosome arm in each meiosis.
- More than 1.4 million single nucleotide polymorphisms (SNPs) in the human genome have been identified. This collection should allow the initiation of genome-wide linkage disequilibrium mapping of the genes in the human population.

In this paper, we start by presenting background information on the project and describing the generation, assembly and evaluation of the draft genome sequence. We then focus on an initial analysis of the sequence itself: the broad chromosomal landscape; the repeat elements and the rich palaeontological record of evolutionary and biological processes that they provide; the human genes and proteins and their differences and similarities with those of other

Genome Sequencing Centres (Listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors at each centre is available as Supplementary Information.)

Whitehead Institute for Biomedical Research, Center for Genome Research: Eric S. Lander[1]*, Lauren M. Linton[1], Bruce Birren[1]*, Chad Nusbaum[1]*, Michael C. Zody[1]*, Jennifer Baldwin[1], Keri Devon[1], Ken Dewar[1], Michael Doyle[1], William FitzHugh[1]*, Roel Funke[1], Diane Gage[1], Katrina Harris[1], Andrew Heaford[1], John Howland[1], Lisa Kann[1], Jessica Lehoczky[1], Rosie LeVine[1], Paul McEwan[1], Kevin McKernan[1], James Meldrim[1], Jill P. Mesirov[1]*, Cher Miranda[1], William Morris[1], Jerome Naylor[1], Christina Raymond[1], Mark Rosetti[1], Ralph Santos[1], Andrew Sheridan[1], Carrie Sougnez[1], Nicole Stange-Thomann[1], Nikola Stojanovic[1], Aravind Subramanian[1] & Dudley Wyman[1]

The Sanger Centre: Jane Rogers[2], John Sulston[2]*, Rachael Ainscough[2], Stephan Beck[2], David Bentley[2], John Burton[2], Christopher Clee[2], Nigel Carter[2], Alan Coulson[2], Rebecca Deadman[2], Panos Deloukas[2], Andrew Dunham[2], Ian Dunham[2], Richard Durbin[2]*, Lisa French[2], Darren Grafham[2], Simon Gregory[2], Tim Hubbard[2]*, Sean Humphray[2], Adrienne Hunt[2], Matthew Jones[2], Christine Lloyd[2], Amanda McMurray[2], Lucy Matthews[2], Simon Mercer[2], Sarah Milne[2], James C. Mullikin[2]*, Andrew Mungall[2], Robert Plumb[2], Mark Ross[2], Ratna Shownkeen[2] & Sarah Sims[2]

Washington University Genome Sequencing Center: Robert H. Waterston[3]*, Richard K. Wilson[3], LaDeana W. Hillier[3]*, John D. McPherson[3], Marco A. Marra[3], Elaine R. Mardis[3], Lucinda A. Fulton[3], Asif T. Chinwalla[3]*, Kymberlie H. Pepin[3], Warren R. Gish[3], Stephanie L. Chissoe[3], Michael C. Wendl[3], Kim D. Delehaunty[3], Tracie L. Miner[3], Andrew Delehaunty[3], Jason B. Kramer[3], Lisa L. Cook[3], Robert S. Fulton[3], Douglas L. Johnson[3], Patrick J. Minx[3] & Sandra W. Clifton[3]

US DOE Joint Genome Institute: Trevor Hawkins[4], Elbert Branscomb[4], Paul Predki[4], Paul Richardson[4], Sarah Wenning[4], Tom Slezak[4], Norman Doggett[4], Jan-Fang Cheng[4], Anne Olsen[4], Susan Lucas[4], Christopher Elkin[4], Edward Uberbacher[4] & Marvin Frazier[4]

Baylor College of Medicine Human Genome Sequencing Center: Richard A. Gibbs[5]*, Donna M. Muzny[5], Steven E. Scherer[5], John B. Bouck[5]*, Erica J. Sodergren[5], Kim C. Worley[5]*, Catherine M. Rives[5], James H. Gorrell[5], Michael L. Metzker[5], Susan L. Naylor[6], Raju S. Kucherlapati[7], David L. Nelson, & George M. Weinstock[8]

RIKEN Genomic Sciences Center: Yoshiyuki Sakaki[9], Asao Fujiyama[9], Masahira Hattori[9], Tetsushi Yada[9], Atsushi Toyoda[9], Takehiko Itoh[9], Chiharu Kawagoe[9], Hidemi Watanabe[9], Yasushi Totoki[9] & Todd Taylor[9]

Genoscope and CNRS UMR-8030: Jean Weissenbach[10], Roland Heilig[10], William Saurin[10], Francois Artiguenave[10], Philippe Brottier[10], Thomas Bruls[10], Eric Pelletier[10], Catherine Robert[10] & Patrick Wincker[10]

GTC Sequencing Center: Douglas R. Smith[11], Lynn Doucette-Stamm[11], Marc Rubenfield[11], Keith Weinstock[11], Hong Mei Lee[11] & JoAnn Dubois[11]

Department of Genome Analysis, Institute of Molecular Biotechnology: André Rosenthal[12], Matthias Platzer[12], Gerald Nyakatura[12], Stefan Taudien[12] & Andreas Rump[12]

Beijing Genomics Institute/Human Genome Center: Huanming Yang[13], Jun Yu[13], Jian Wang[13], Guyang Huang[14] & Jun Gu[15]

Multimegabase Sequencing Center, The Institute for Systems Biology: Leroy Hood[16], Lee Rowen[16], Anup Madan[16] & Shizen Qin[16]

Stanford Genome Technology Center: Ronald W. Davis[17], Nancy A. Federspiel[17], A. Pia Abola[17] & Michael J. Proctor[17]

Stanford Human Genome Center: Richard M. Myers[18], Jeremy Schmutz[18], Mark Dickson[18], Jane Grimwood[18] & David R. Cox[18]

University of Washington Genome Center: Maynard V. Olson[19], Rajinder Kaul[19] & Christopher Raymond[19]

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu[20], Kazuhiko Kawasaki[20] & Shinsei Minoshima[20]

University of Texas Southwestern Medical Center at Dallas: Glen A. Evans[21]†, Maria Athanasiou[21] & Roger Schultz[21]

University of Oklahoma's Advanced Center for Genome Technology: Bruce A. Roe[22], Feng Chen[22] & Huaqin Pan[22]

Max Planck Institute for Molecular Genetics: Juliane Ramser[23], Hans Lehrach[23] & Richard Reinhardt[23]

Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center: W. Richard McCombie[24], Melissa de la Bastide[24] & Neilay Dedhia[24]

GBF—German Research Centre for Biotechnology: Helmut Blöcker[25], Klaus Hornischer[25] & Gabriele Nordsiek[25]

* **Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):** Richa Agarwala[26], L. Aravind[26], Jeffrey A. Bailey[27], Alex Bateman[2], Serafim Batzoglou[1], Ewan Birney[28], Peer Bork[29,30], Daniel G. Brown[1], Christopher B. Burge[31], Lorenzo Cerutti[28], Hsiu-Chuan Chen[26], Deanna Church[26], Michele Clamp[2], Richard R. Copley[30], Tobias Doerks[29,30], Sean R. Eddy[32], Evan E. Eichler[27], Terrence S. Furey[33], James Galagan[1], James G. R. Gilbert[2], Cyrus Harmon[34], Yoshihide Hayashizaki[35], David Haussler[36], Henning Hermjakob[28], Karsten Hokamp[37], Wonhee Jang[26], L. Steven Johnson[32], Thomas A. Jones[32], Simon Kasif[38], Arek Kaspryzk[28], Scot Kennedy[39], W. James Kent[40], Paul Kitts[26], Eugene V. Koonin[26], Ian Korf[3], David Kulp[34], Doron Lancet[41], Todd M. Lowe[42], Aoife McLysaght[37], Tarjei Mikkelsen[38], John V. Moran[43], Nicola Mulder[28], Victor J. Pollara[1], Chris P. Ponting[44], Greg Schuler[26], Jörg Schultz[30], Guy Slater[28], Arian F. A. Smit[45], Elia Stupka[28], Joseph Szustakowki[38], Danielle Thierry-Mieg[26], Jean Thierry-Mieg[26], Lukas Wagner[26], John Wallis[3], Raymond Wheeler[34], Alan Williams[34], Yuri I. Wolf[26], Kenneth H. Wolfe[37], Shiaw-Pyng Yang[3] & Ru-Fang Yeh[31]

Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins[46]*, Mark S. Guyer[46], Jane Peterson[46], Adam Felsenfeld[46]* & Kris A. Wetterstrand[46]; **Office of Science, US Department of Energy:** Aristides Patrinos[47]; **The Wellcome Trust:** Michael J. Morgan[48]

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]
Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1]
George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5]
Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9]
Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1]
Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1]
Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1]
Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1]
Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1]
Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1]
Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1]
Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1]
Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1]
Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12]
Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1]
Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1]
Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1]
Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1]
Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1]
Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1]
Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1]
Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1]
Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1]
Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1]
Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1]
Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1]
Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1]
Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1]
Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1]
Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1]
Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1]
Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1]
Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1]
Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1]
Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1]
Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1]
Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1]
Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1]
Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1]
Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1]
Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1]
Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1]
Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1]
Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1]
Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ∼12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1] Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2] GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3] Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4] Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5] Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6] Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7] Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8] New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9] Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10] Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11] Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13] Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14] Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

SOME KEY FINDINGS

- The HGP has revealed that there are probably about 25,000 to 40,000 (since updated to a count of ~20,500 human genes)

- Human genome is remarkably similar to other genomes in terms of total gene humbers and gene functions, although most genes are more complex. (Comparitive Genomics)

- Between 1.1% to 1.4% of the genome's sequence codes for proteins. Nonfunctional regions appear to account for ~97%. 12% of human genomic DNA is due to copy number variations - CNVs

- ~2 million single nucleotide polymorphisms - SNPs (~0.1 to 0.3% of total genome)

LOOKING AT SNPs

SOME KEY FINDINGS

- The HGP has revealed that there are probably about 25,000 to 40,000 (since updated to a count of ~20,500 human genes)

- Human genome is remarkably similar to other genomes in terms of total gene numbers and gene functions, although most genes are more complex. (Comparitive Genomics)

- Between 1.1% to 1.4% of the genome's sequence codes for proteins. Nonfunctional regions appear to account for ~97%. 12% of human genomic DNA is due to copy number variations - CNVs

~2 million single nucleotide polymorphisms - SNPs (~0.1 to 0.3% of total genome).

- ~2 million single nucleotide polymorphisms - SNPs (~0.1 to 0.3% of total genome)

is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between (human) members.

AGCTTAGCGAGTGACCGGTCAGCTTACGCAGATCGAGGAGCTTACG

AGCTTAGCGAGTGCCCGGTCAGCTTACGCAGATCGAGGATCTTACG

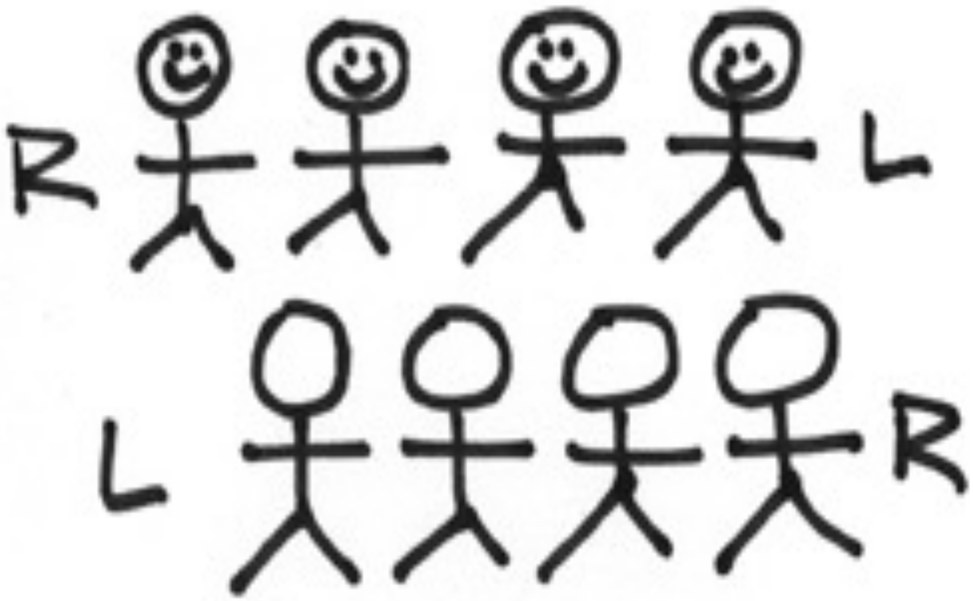
AGCTTAGCGAGTGCCCGGTCAGCTTACGCAGATCGAGGATCTTACG

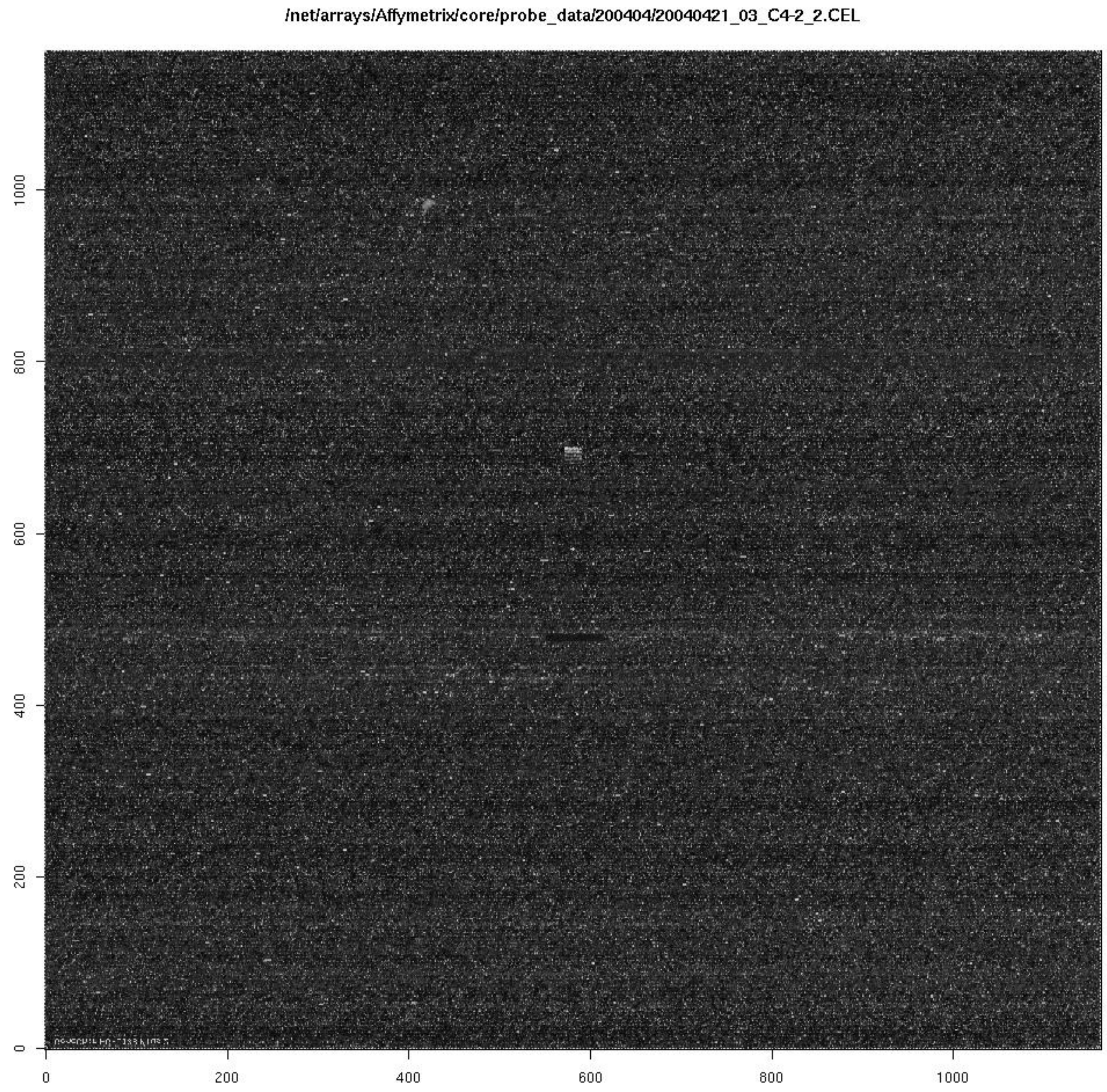
AGCTTAGCGAGTGACCGGTCAGCTTACGCAGATCGAGGAGCTTACG

AGCTTAGCGAGTGCCCGGTCAGCTTACGCAGATCGAGGATCTTACG

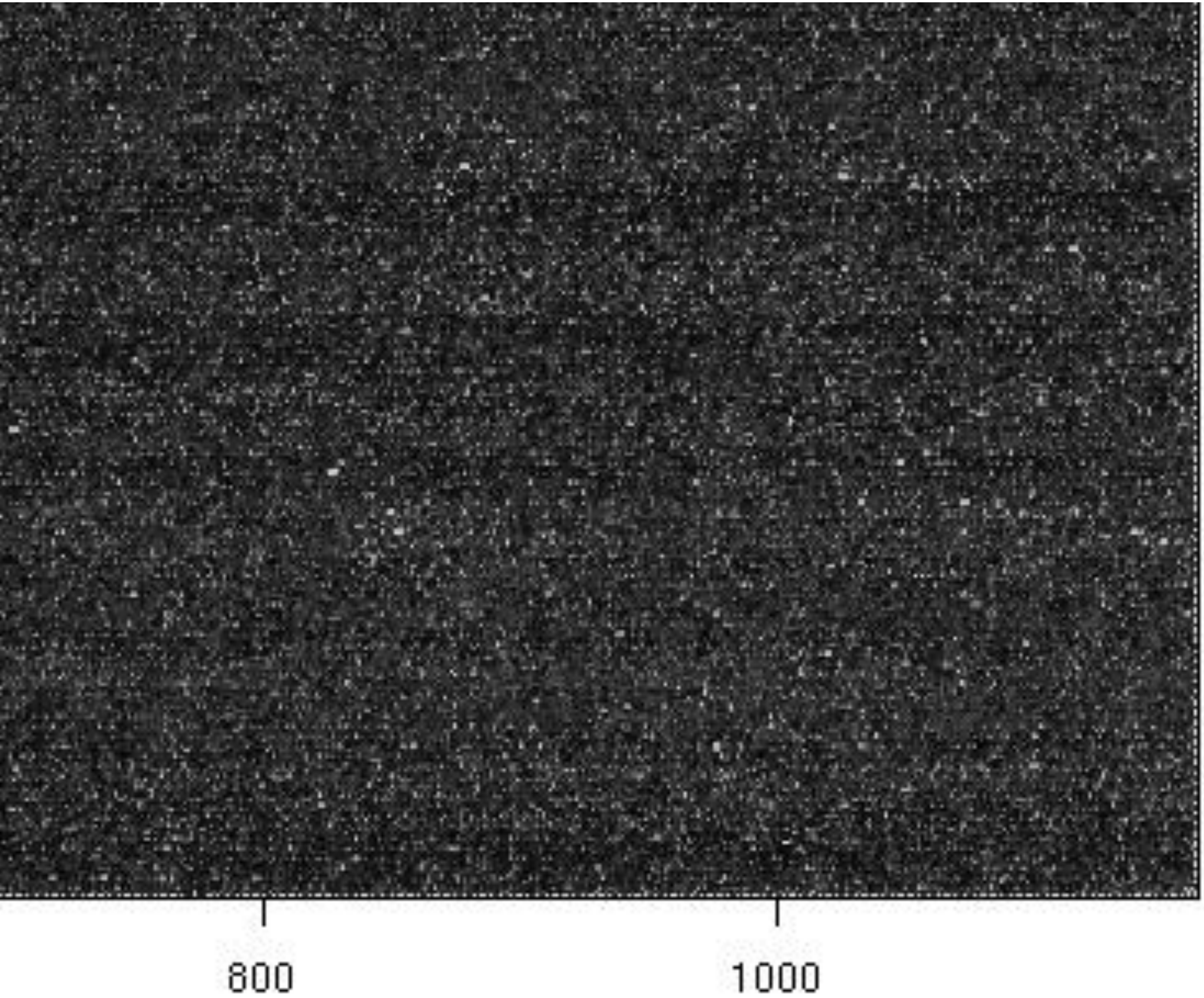
2 ALLELES A vs C in GENE X

2 ALLELES G vs T in GENE Y


R
L
L
R

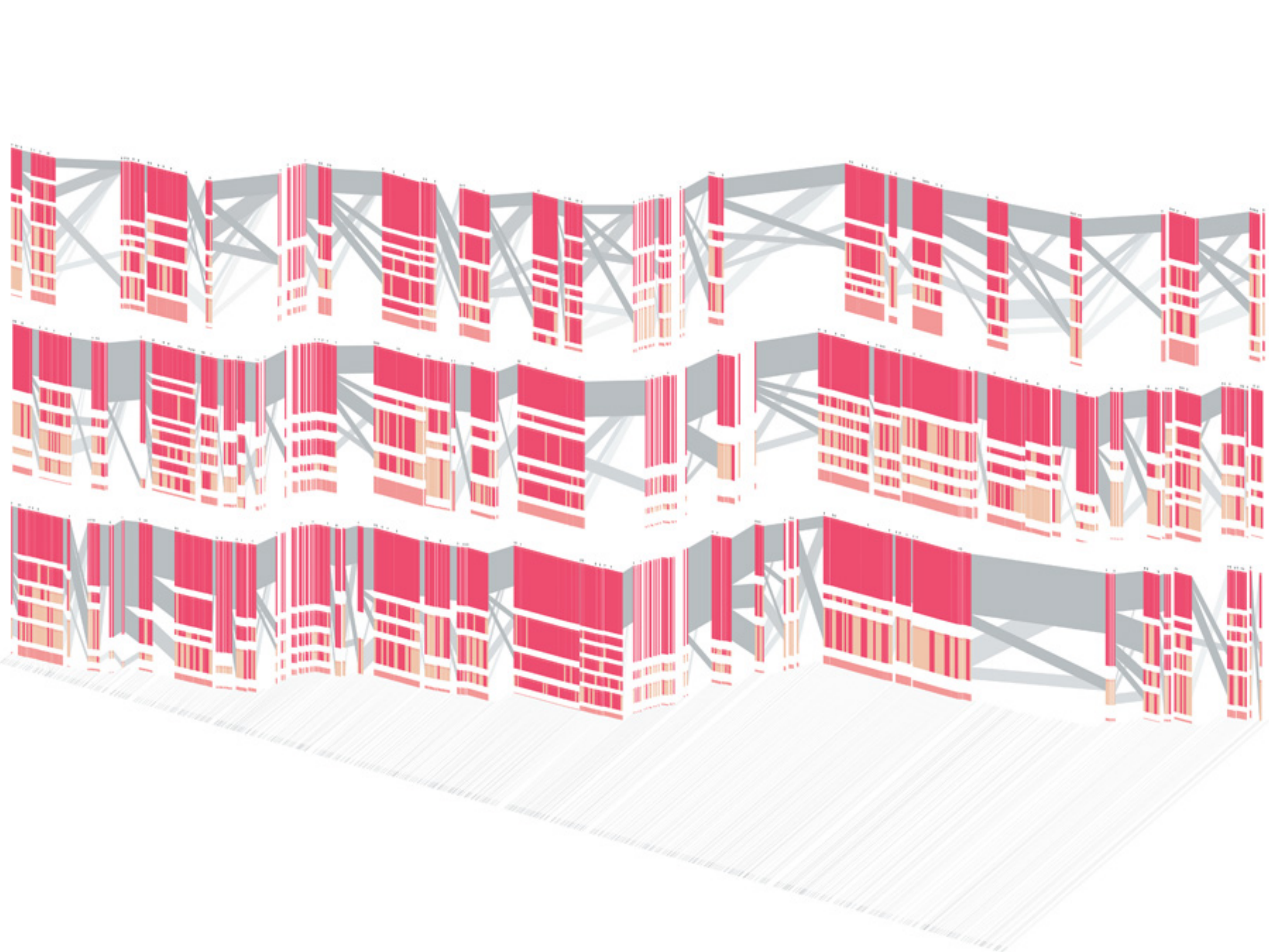

/net/arrays/Affymetrix/core/probe_data/200404/20040421_03_C4-2_2.CEL
0
200
400
600
800
1000
0
200
400
600
800
1000

MICROARRAY DNA CHIP


800
1000

~10,000,000 SNPs

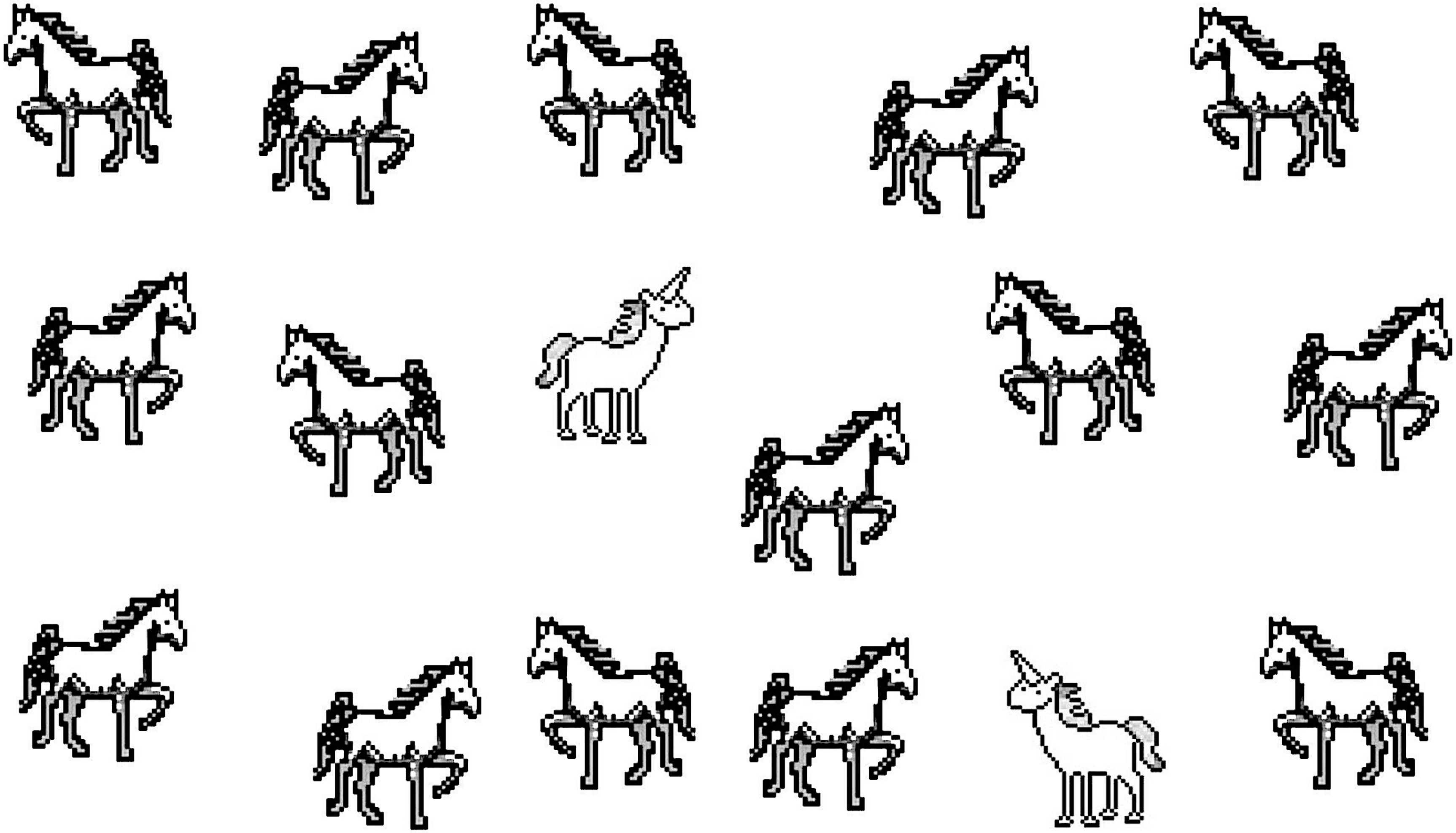
27 October 2005 · www.nature.com/nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE HAPMAP PROJECT
Chapter and verse on human genetic variation



23andMe
Ancestry
Edition

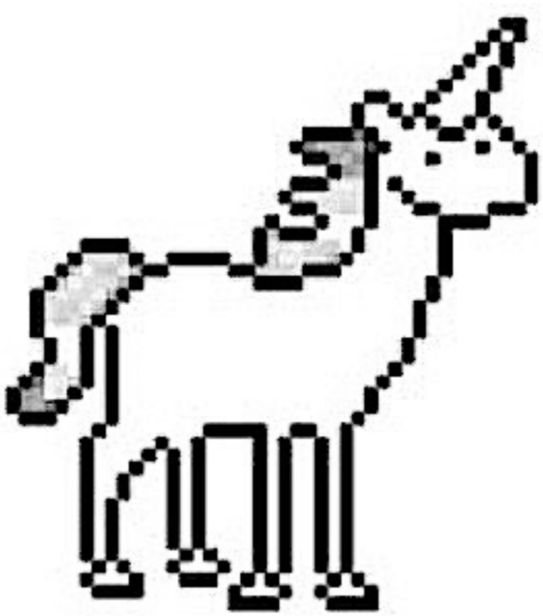
JUST SEQUENCE THE ENTIRE THING BETTER?

HYPOTHETICAL EXAMPLE



http://miserycat.deviantart.com/art/Free-Unicorn-Icon-180216460?q=&qo=

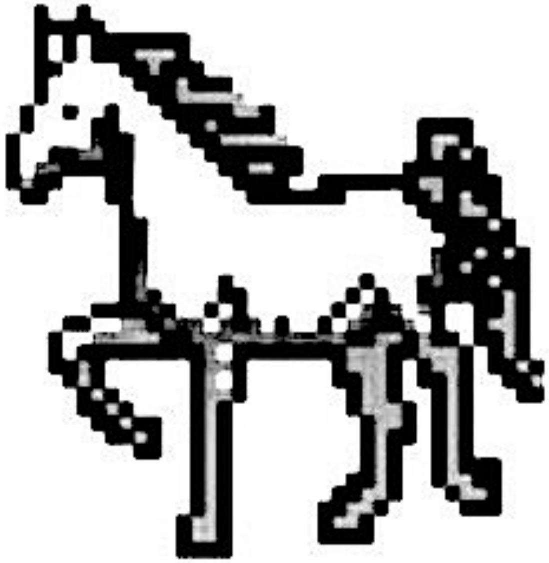
UNICORN



🔴 1	⚪ 2	🔴 3
🔴 4	🔴 5	⚪ 6
⚪ 7	🔴 8	🔴 9

http://miserycat.deviantart.com/art/Free-Unicorn-Icon-180216460?q=&qo=

HORSE



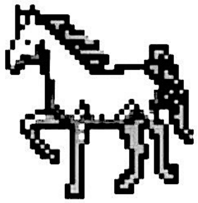
⚪ 1	⚪ 2	🔴 3
⚪ 4	🔴 5	🔴 6
⚪ 7	🔴 8	🔴 9

UNICORN

●○● ●●○ ○●●	●○● ●●● ●○○

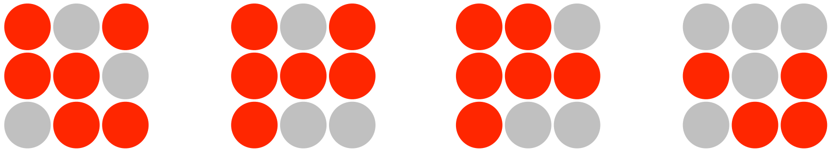
http://miserycat.deviantart.com/art/Free-Unicorn-Icon-180216460?q=&qo=

HORSE



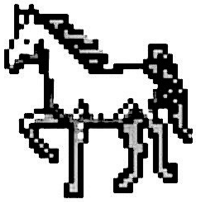
○○● ○●● ○●●	○○● ○●○ ●○●	●●○ ○●● ●○○	○○○ ○○● ○●●
○●○ ○○● ○●●	●○● ○●○ ○○●	●●○ ○●● ●●●	●○○ ○●● ●○○
○●○ ○●○ ●●●	●○○ ○●● ●●○	●●○ ○○● ○○○	○●● ○○○ ○○○
●●● ○●○ ○●○	●●○ ○●● ●○○	○○○ ○○● ●●○	○○○ ○○○ ○●●
○●● ○○● ○●○	○○● ○●○ ●○●	●●● ○●● ●●○	●○○ ○●● ●○○

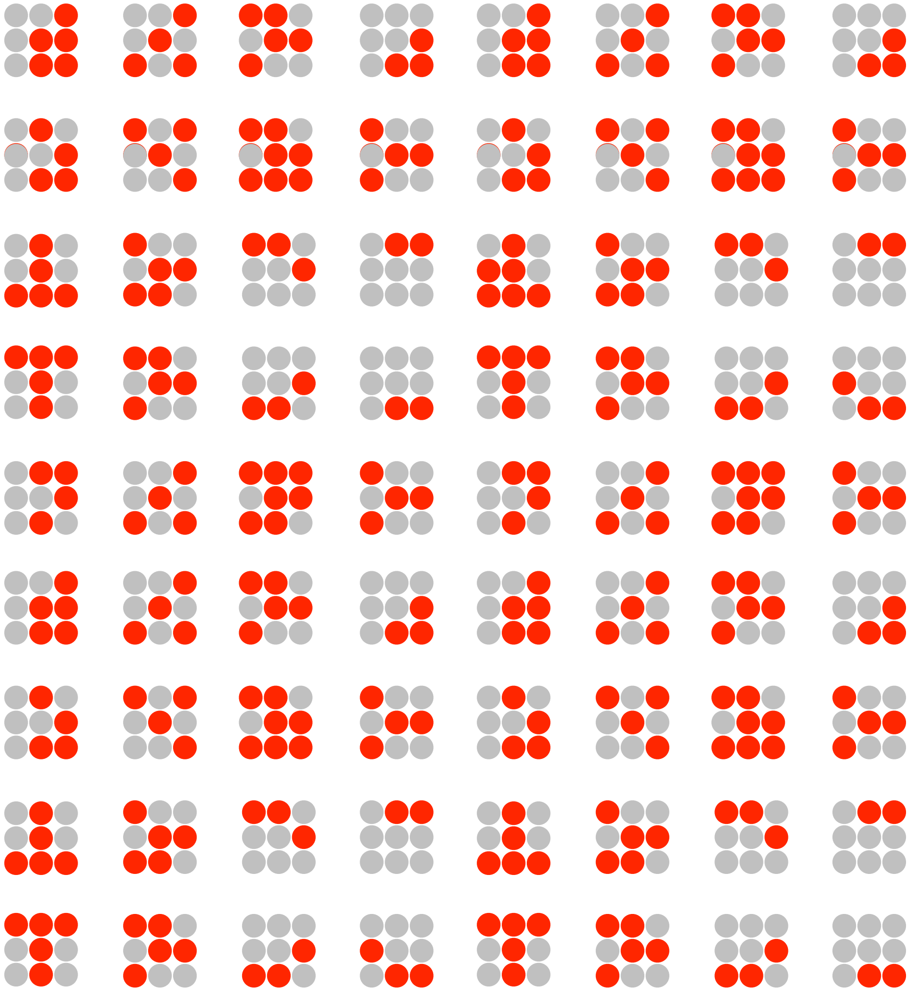
UNICORN



http://miserycat.deviantart.com/art/Free-Unicorn-Icon-180216460?q=&qo=

HORSE

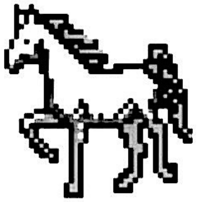


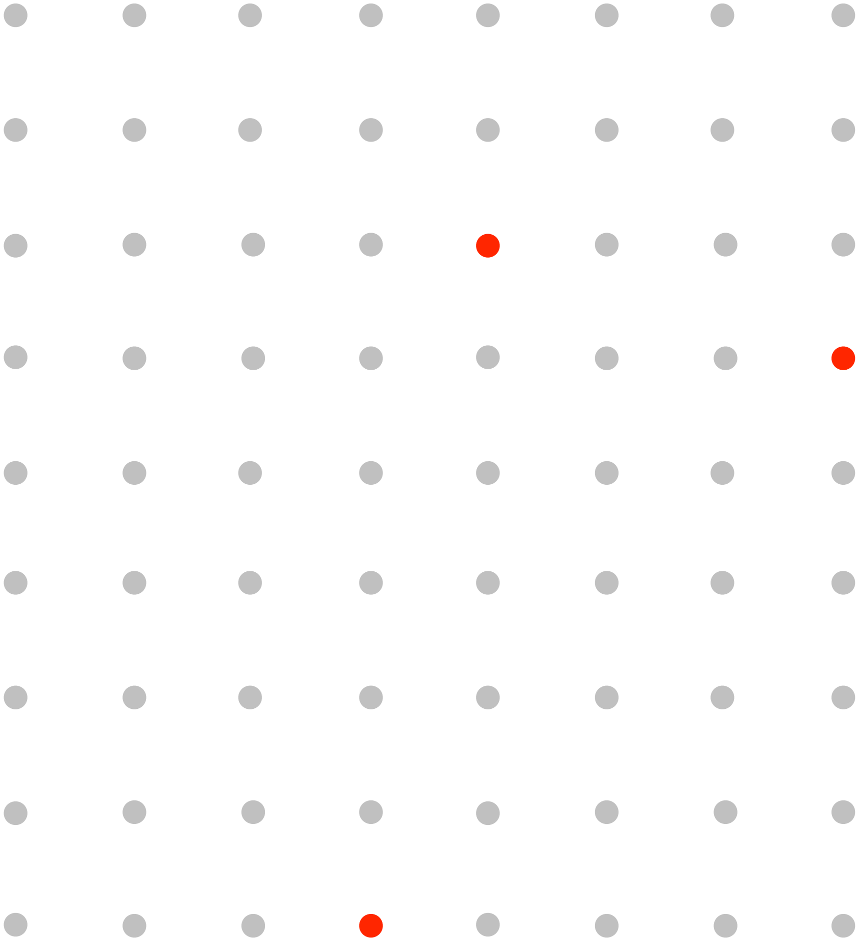


UNICORN

http://miserycat.deviantart.com/art/Free-Unicorn-Icon-180216460?q=&qo=

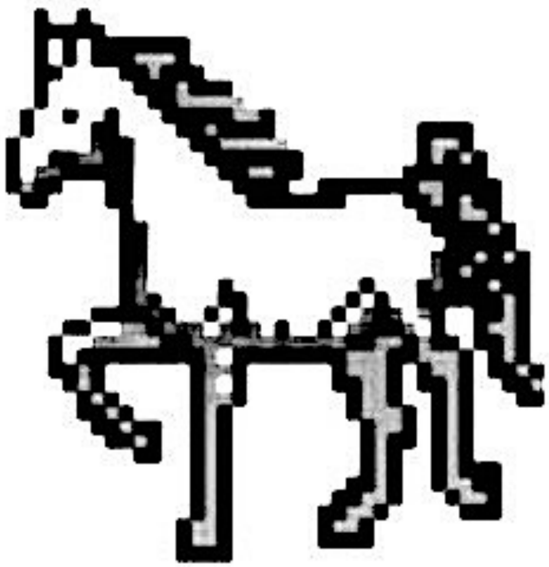
HORSE





UNICORN

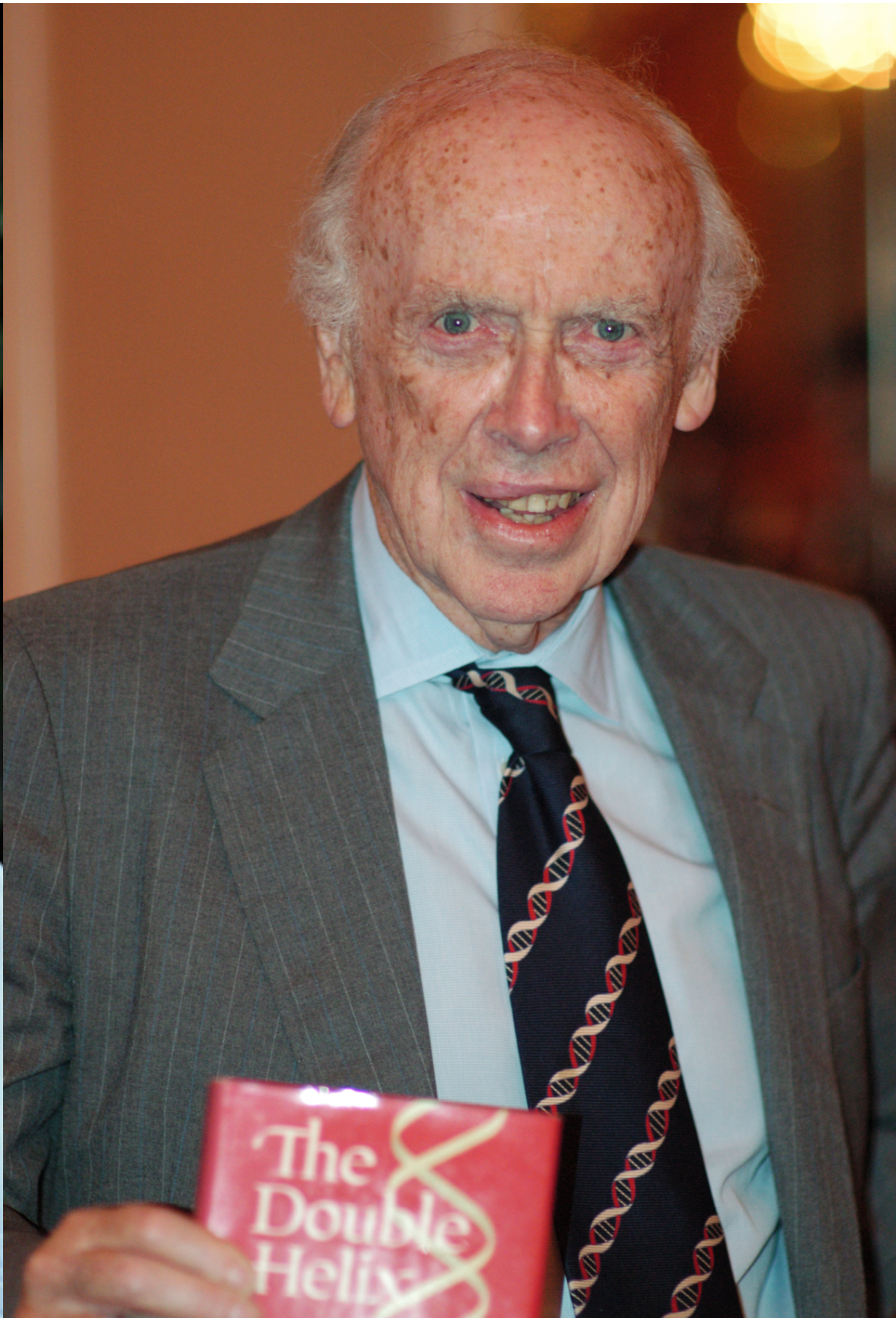
HORSE



4

4

http://miserycat.deviantart.com/art/Free-Unicorn-Icon-180216460?q=&qo=


The Double Helix

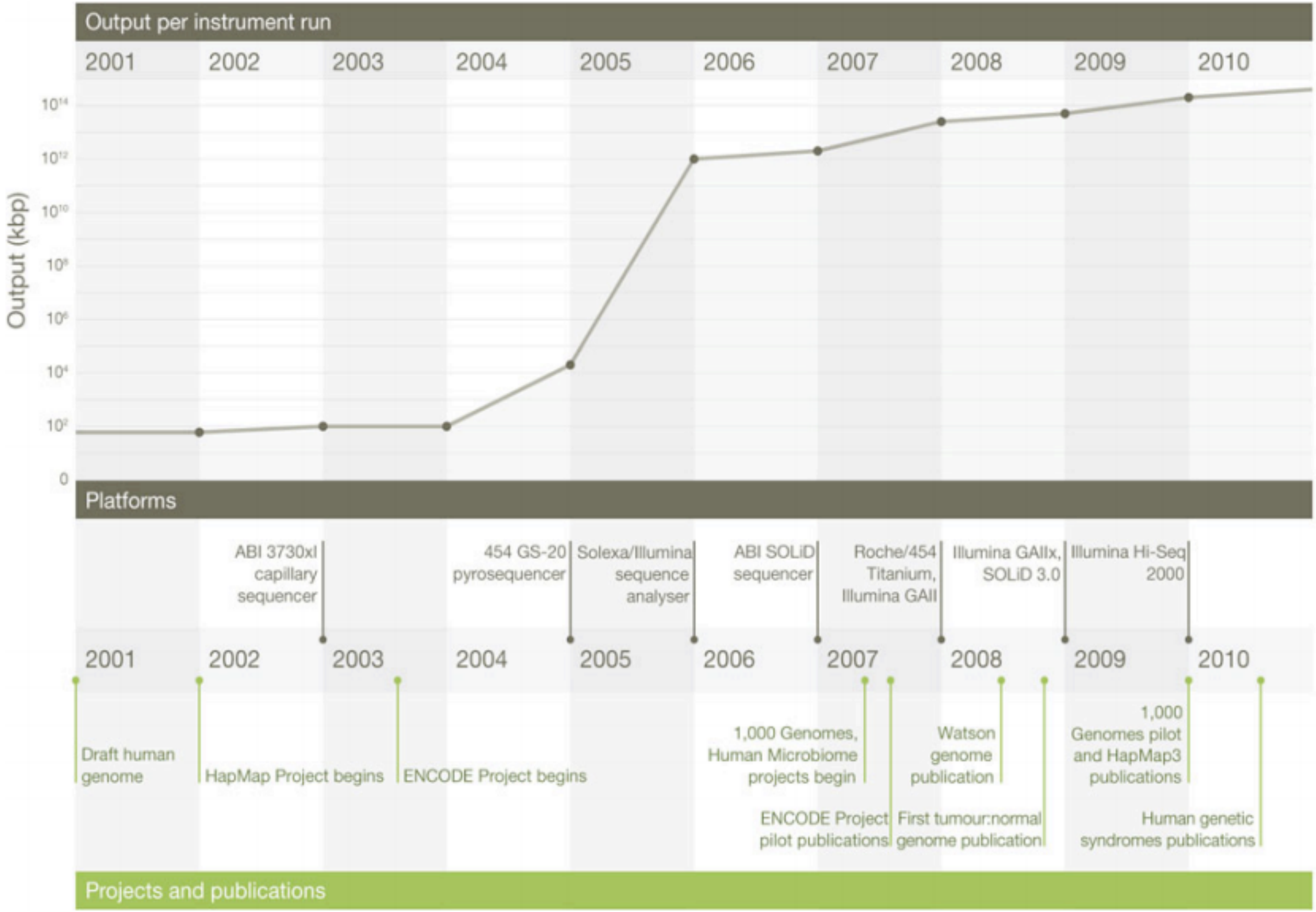
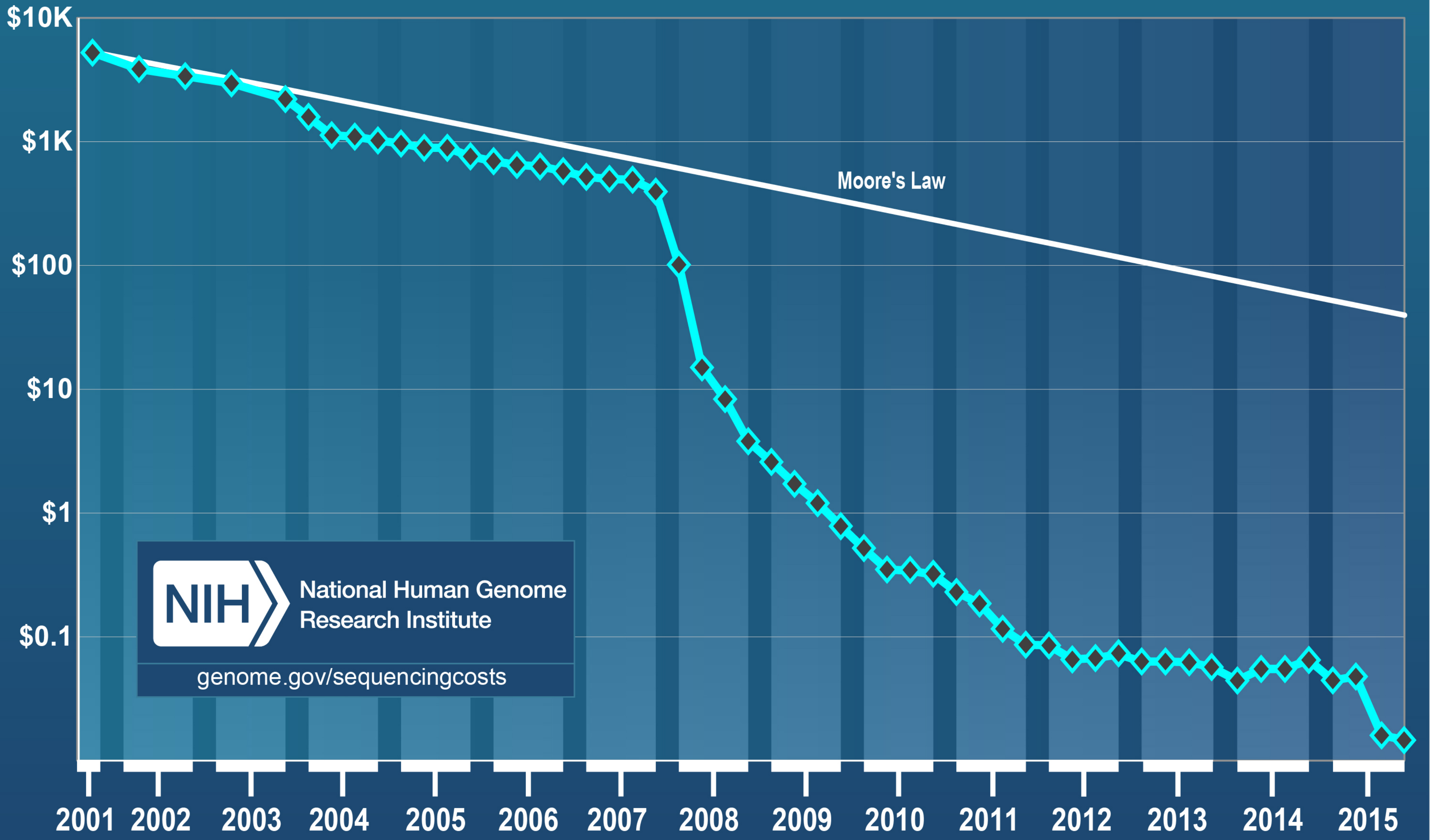


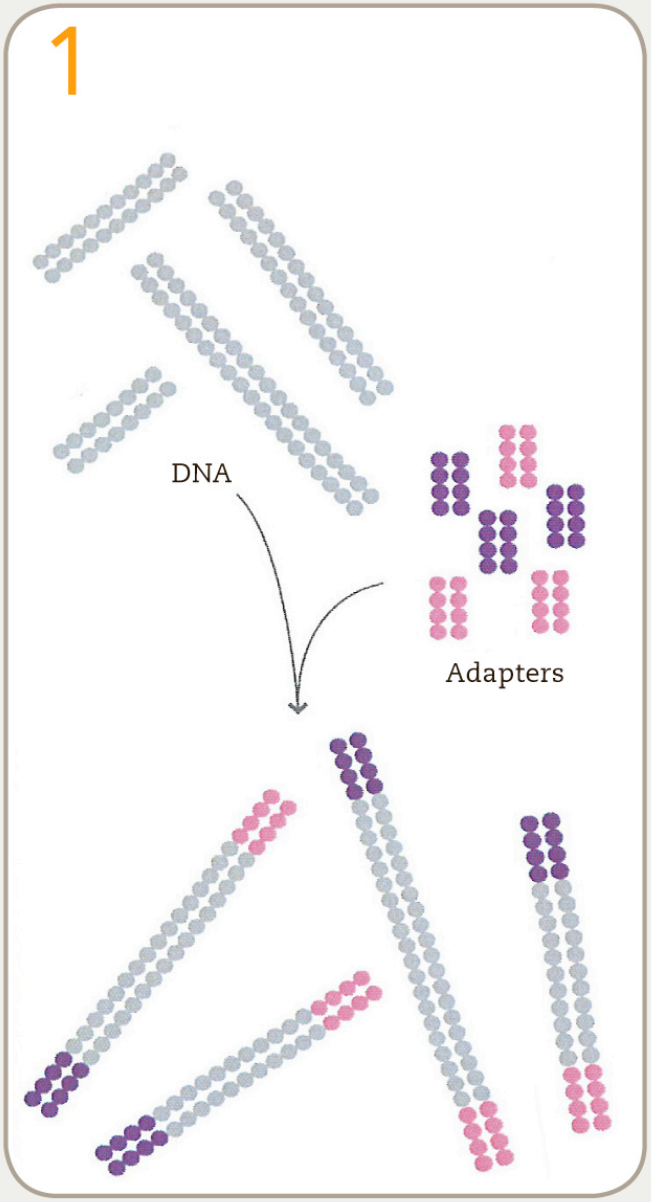
Figure 1. Changes in instrument capacity over the past decade, and the timing of major sequencing projects (ER Mardis. *Nature* **470**, 198-203 (2011) doi:10.1038/nature09796)

Cost per Raw Megabase of DNA Sequence



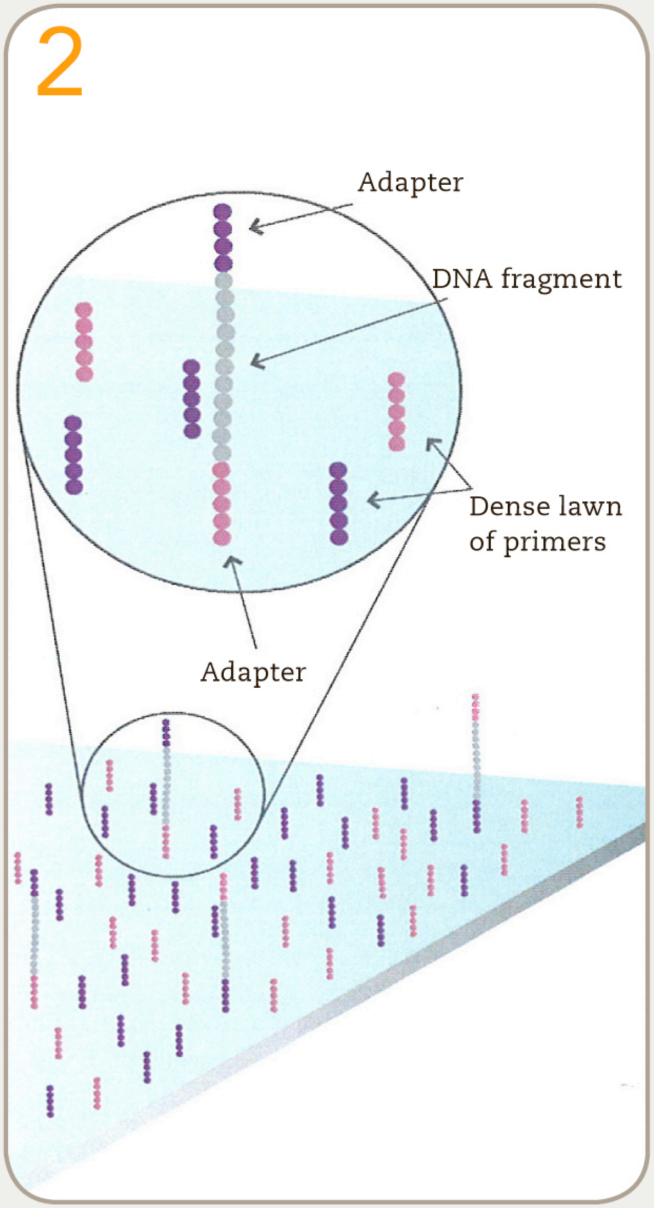
NIH National Human Genome Research Institute

genome.gov/sequencingcosts



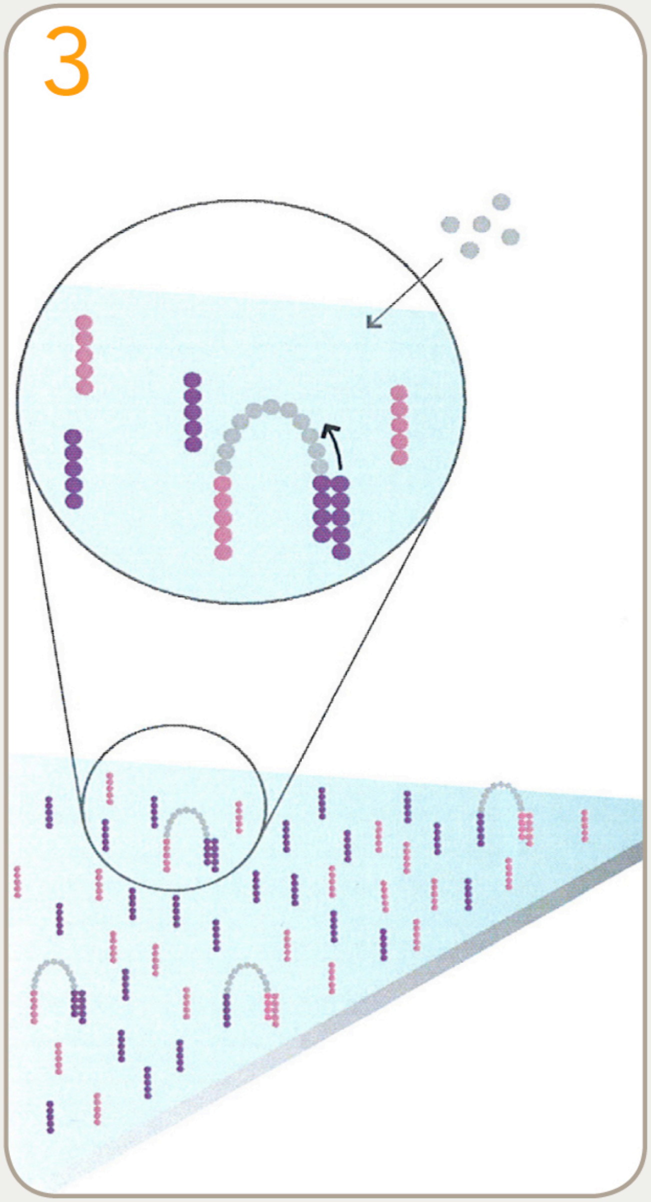
PREPARE GENOMIC DNA SAMPLE

Randomly fragmented genomic DNA and ligate adaptors to both ends of the fragments



ATTACH DNA TO SURFACE

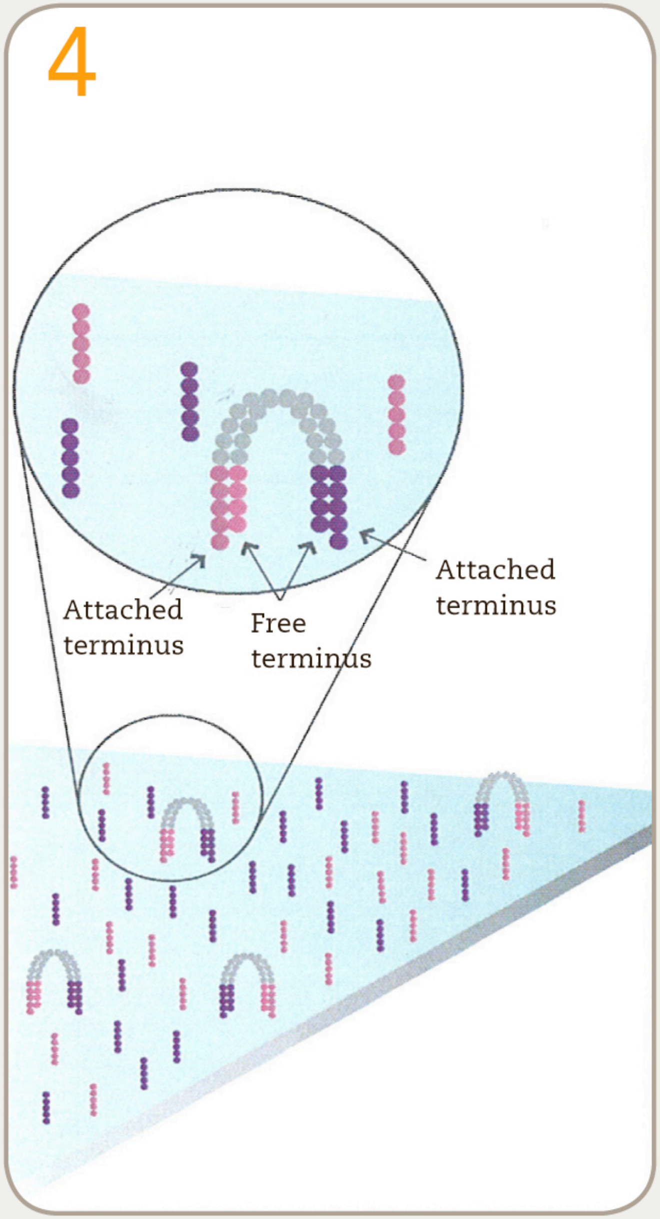
Bind single stranded fragments randomly to the inside surface of the flow cell channels.



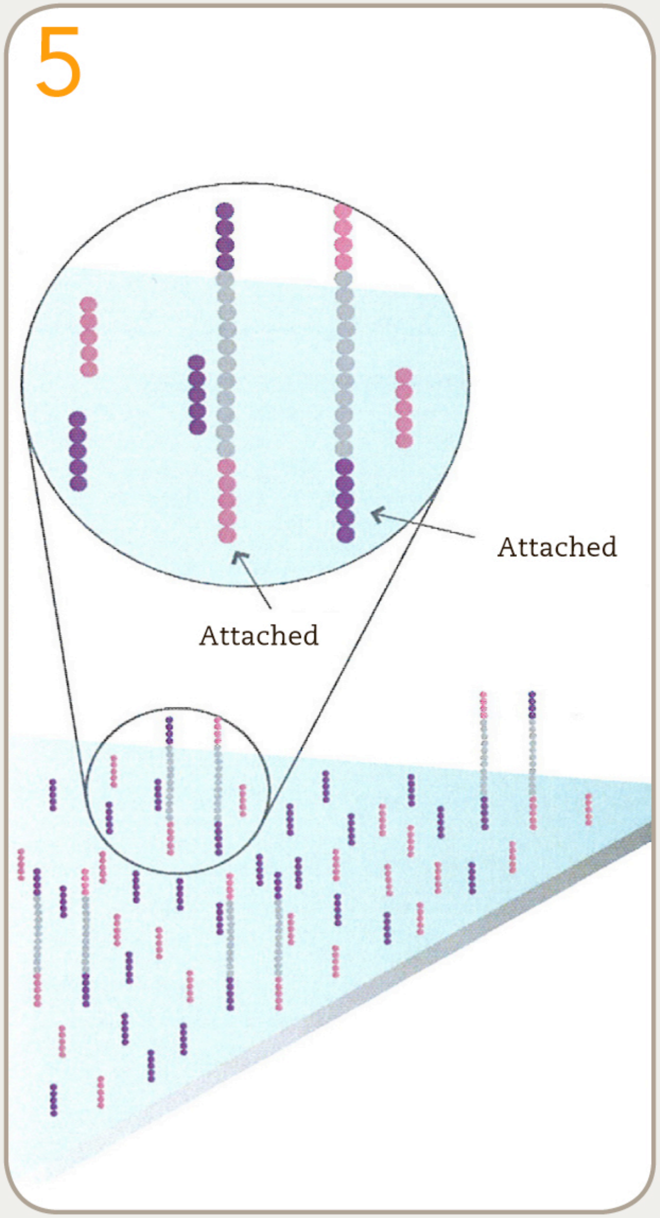
BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

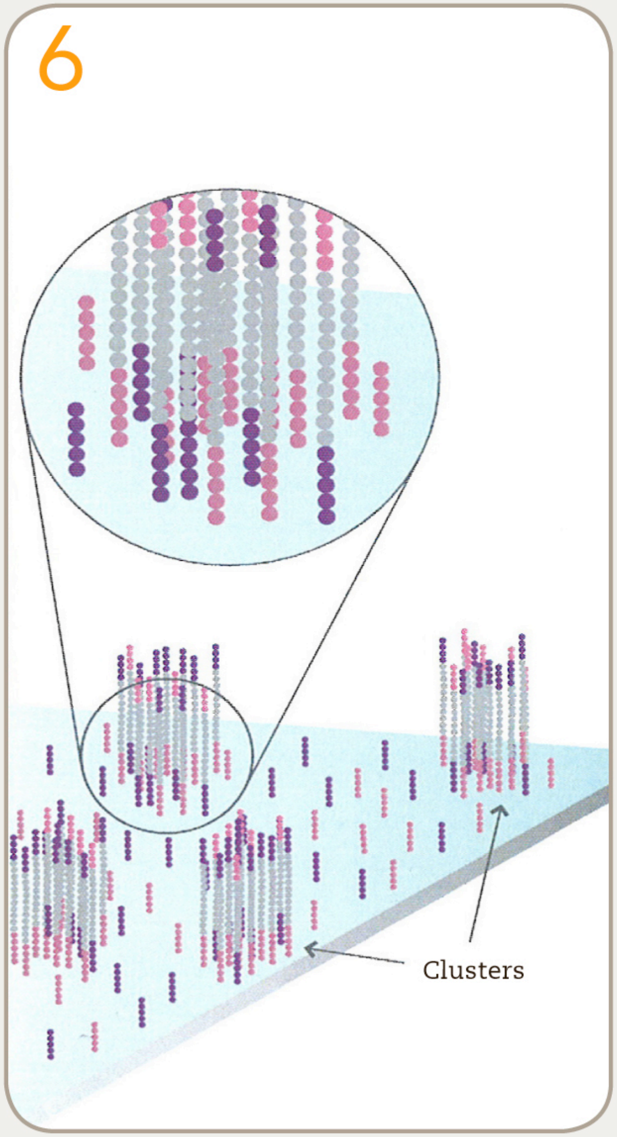
source: http://www.illumina.com/



FRAGMENTS BECOME DOUBLE STRANDED



DENATURE THE DOUBLE STRANDED MOLECULES

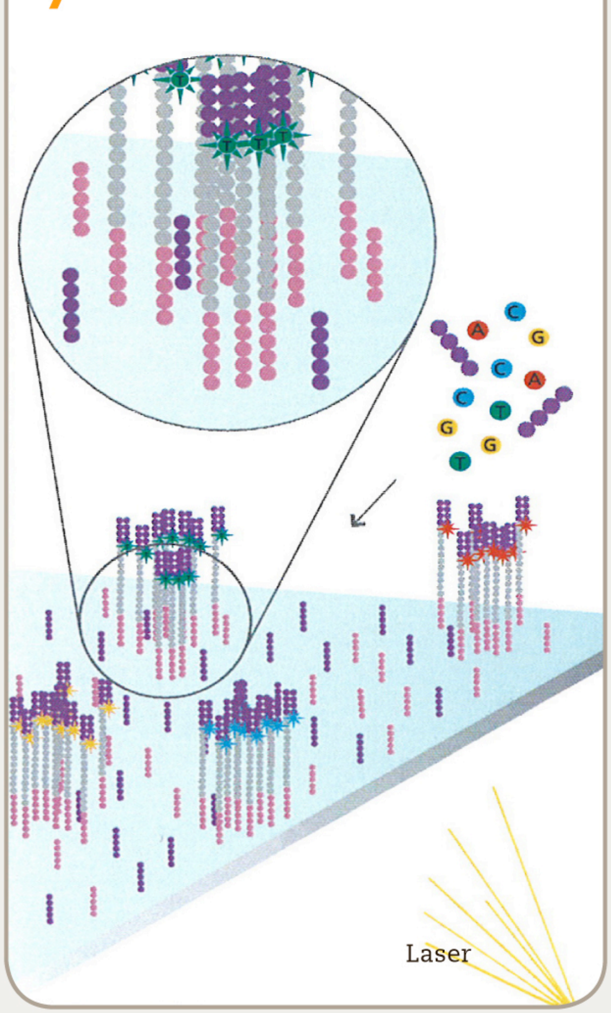


COMPLETION OF AMPLIFICATION

On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.

source: http://www.illumina.com/

7



FIRST CHEMISTRY CYCLE: DETERMINE FIRST BASE

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

8

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.

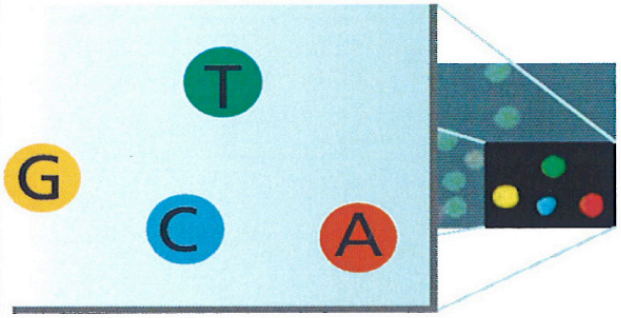
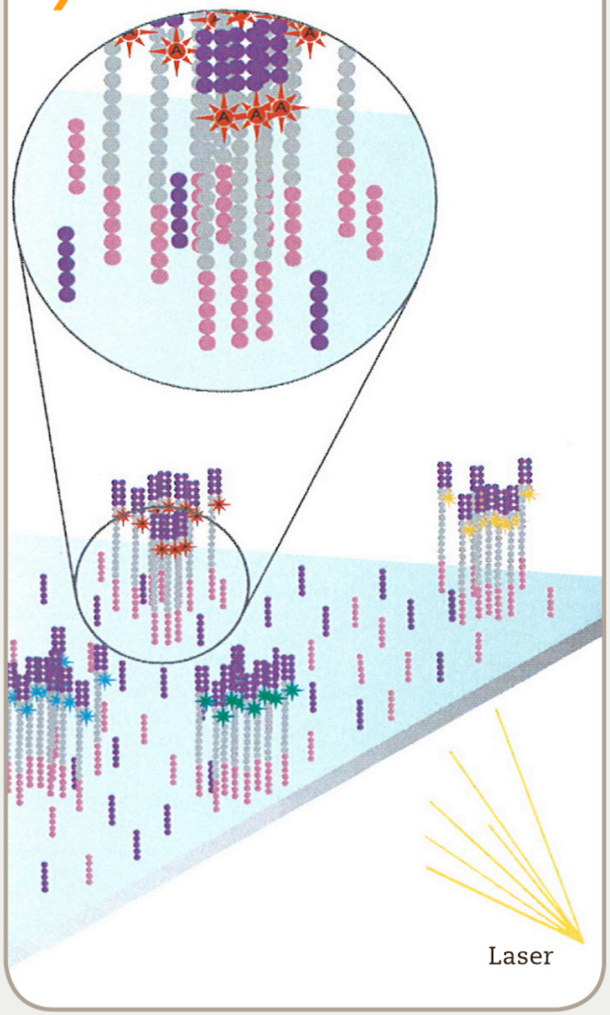


IMAGE OF FIRST CHEMISTRY CYCLE

9



SECOND CHEMISTRY CYCLE: DETERMINE SECOND BASE

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

source: http://www.illumina.com/

10

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

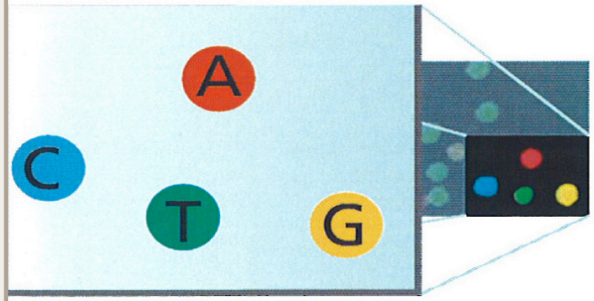
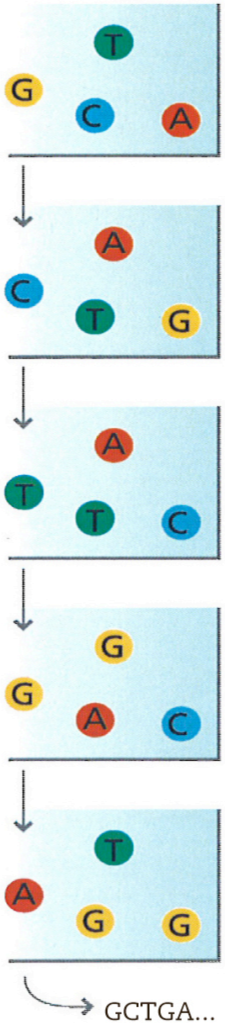


IMAGE OF SECOND CHEMISTRY CYCLE IS CAPTURED BY THE INSTRUMENT.

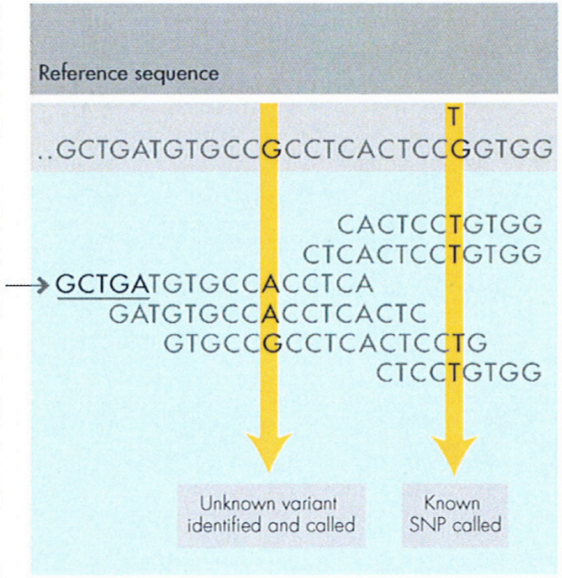
11

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



SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

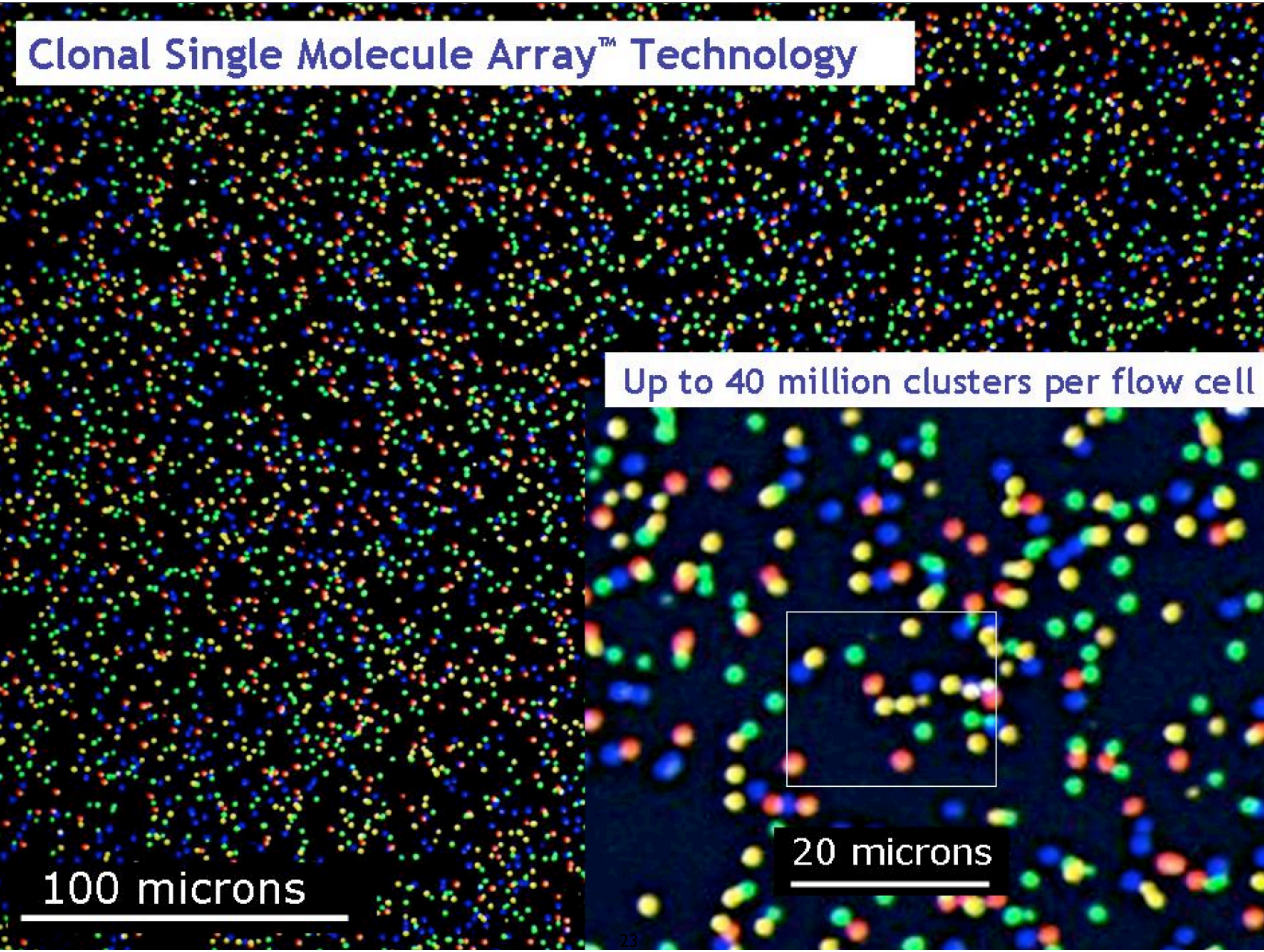
12



ALIGN THE NEW DATA TO A REFERENCE AND IDENTIFY SEQUENCE DIFFERENCES.

source: http://www.illumina.com/

Clonal Single Molecule Array™ Technology



Up to 40 million clusters per flow cell

AT LAUNCH OF HUMAN GENOME PROJECT (1990)
Several machines to sequence the human genome. Est. time and cost:
15 years and $3 billion

4 years ago (2012):
One machine can sequence an entire genome in about 8 days at a cost of about $10,000

3 years ago (2013):
One machine can sequence an entire genome in about 3 days at a cost of about $5,000

2 years ago (2014):
One large scale set up (HiSeq X Ten) can sequence an entire genome's worth of data in about 1 day at a cost of $1,000 (set up is tens of millions of dollars).

Latest (October 2015):
One machine can sequence an entire genome in about 1 day at a cost of about $1,200

THE PRICE IS RIGHT!

Illumina MiSEQ
$125,000

Ion Proton
$80,000

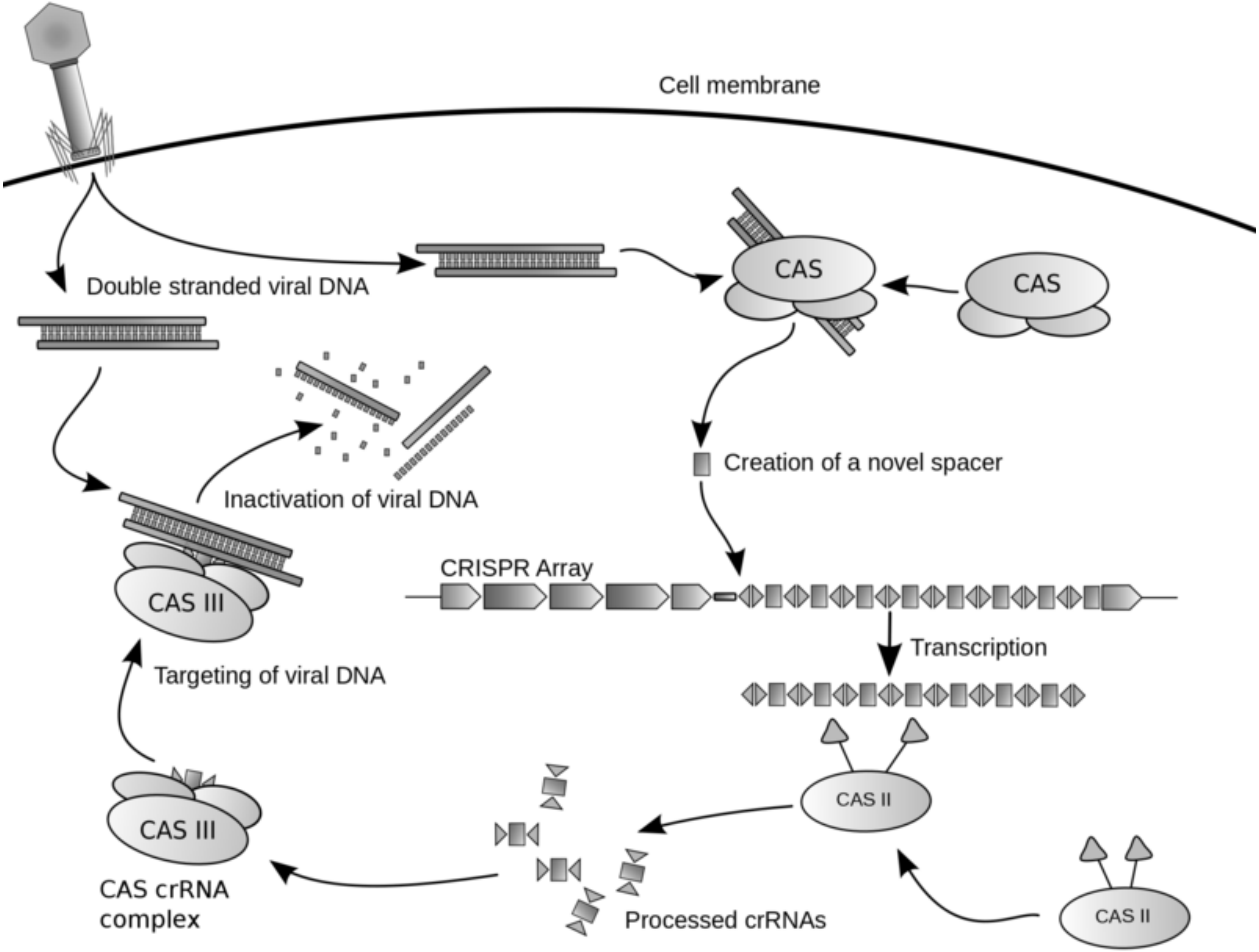
MinION
~$1,500

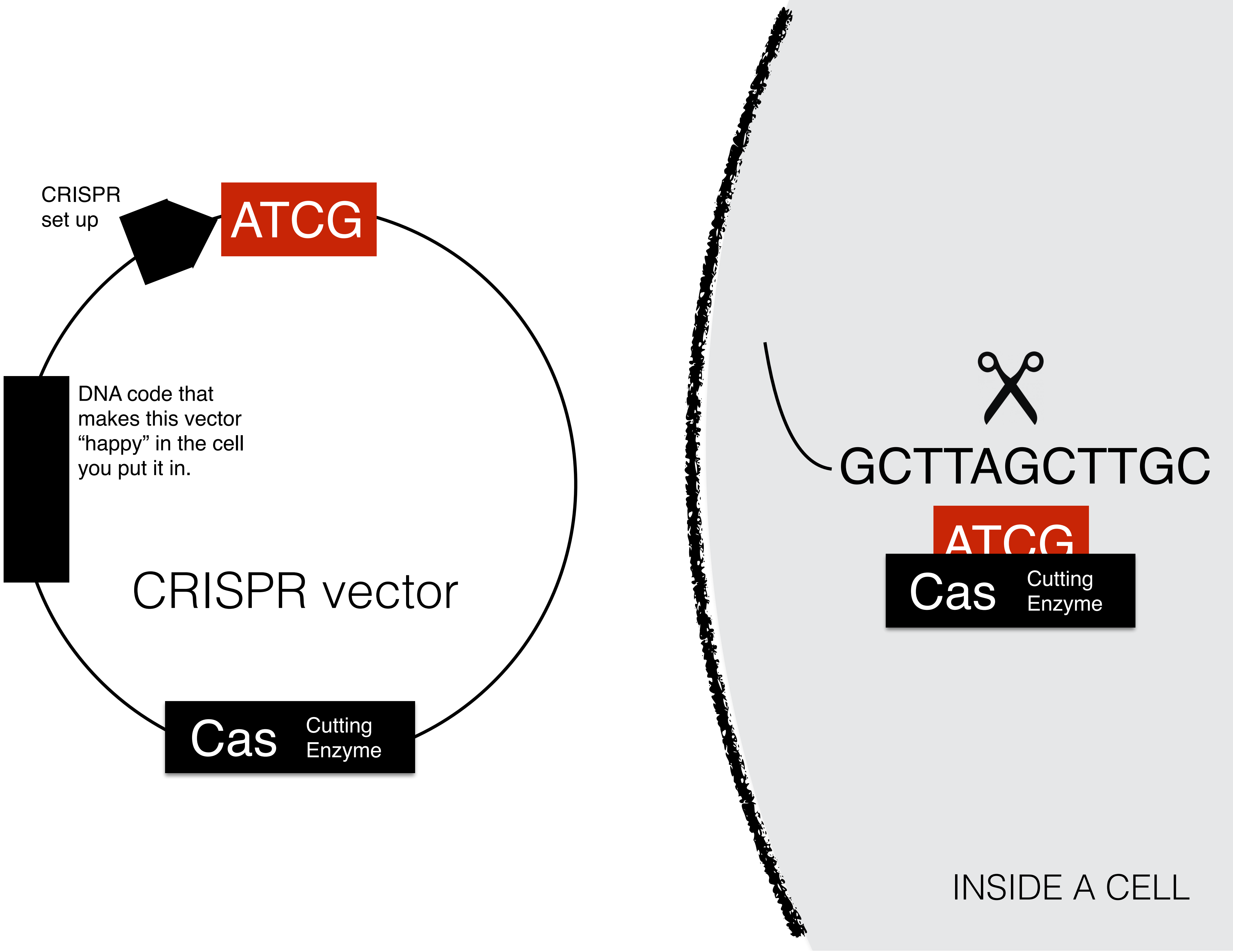
http://vimeo.com/77246565

ARMED WITH INFORMATION ON VARIATIONS OF CODE, WHAT IF YOU COULD EDIT THAT CODE IN YOUR GENOME?

↓

CRISPR/Cas


Cell membrane
Double stranded viral DNA
CAS
CAS
Creation of a novel spacer
Inactivation of viral DNA
CRISPR Array
CAS III
Transcription
Targeting of viral DNA
CAS II
CAS III
CAS II
CAS crRNA
complex
Processed crRNAs



CRISPR set up

ATCG

DNA code that makes this vector “happy” in the cell you put it in.

CRISPR vector

Cas Cutting Enzyme

GCTTAGCTTGC

ATCG

Cas Cutting Enzyme

INSIDE A CELL