

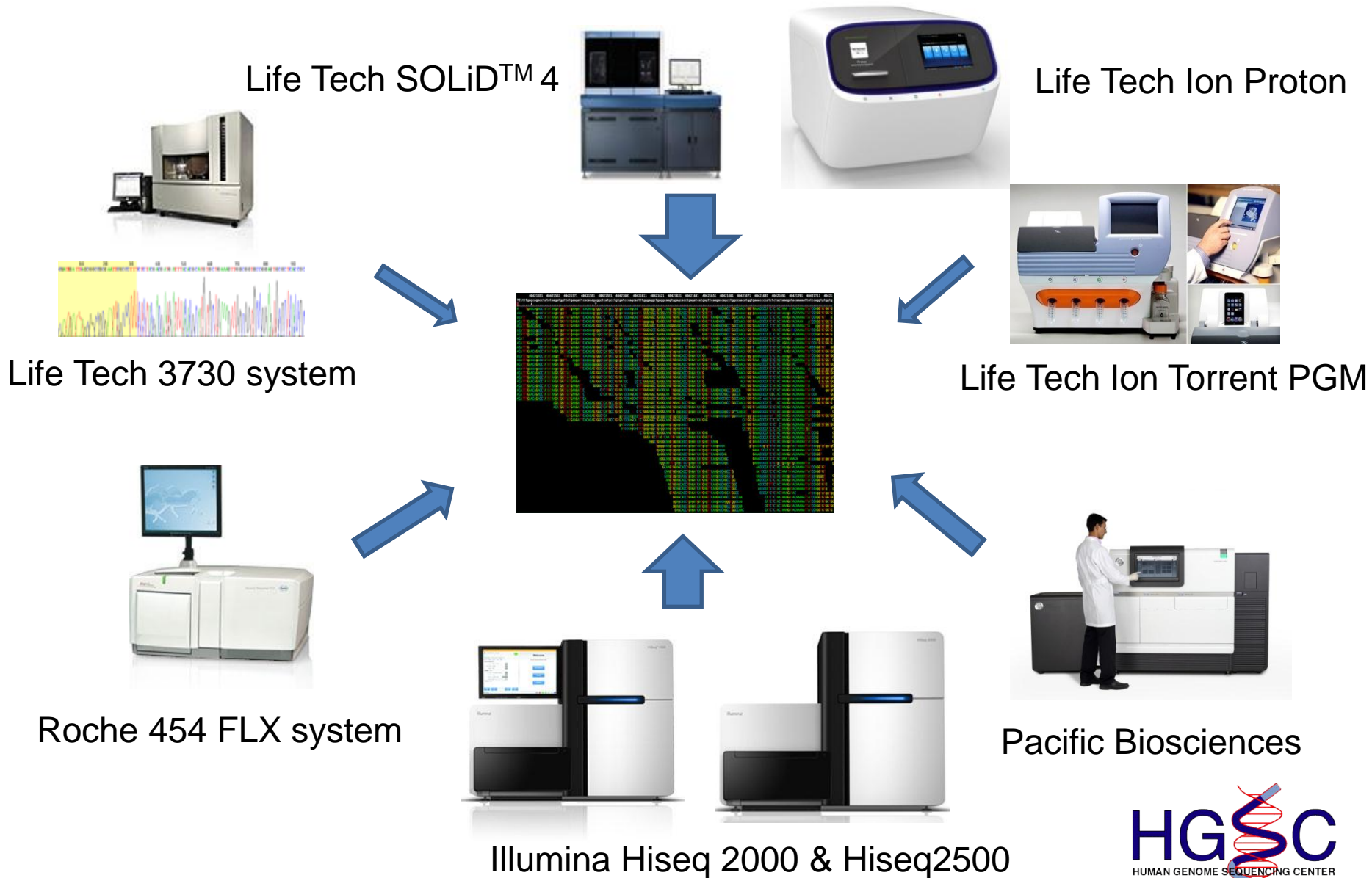
Recent Advances in Second and Third Generation Sequencing

ABRF 2013

Application of Next Generation
Sequencing Technologies for Whole
Transcriptome and Genome Analysis
Workshop

Steve Scherer, Ph.D.
Human Genome Sequencing Center
Baylor College of Medicine

BCM-HGSC Sequencer Fleet



Big Science Projects

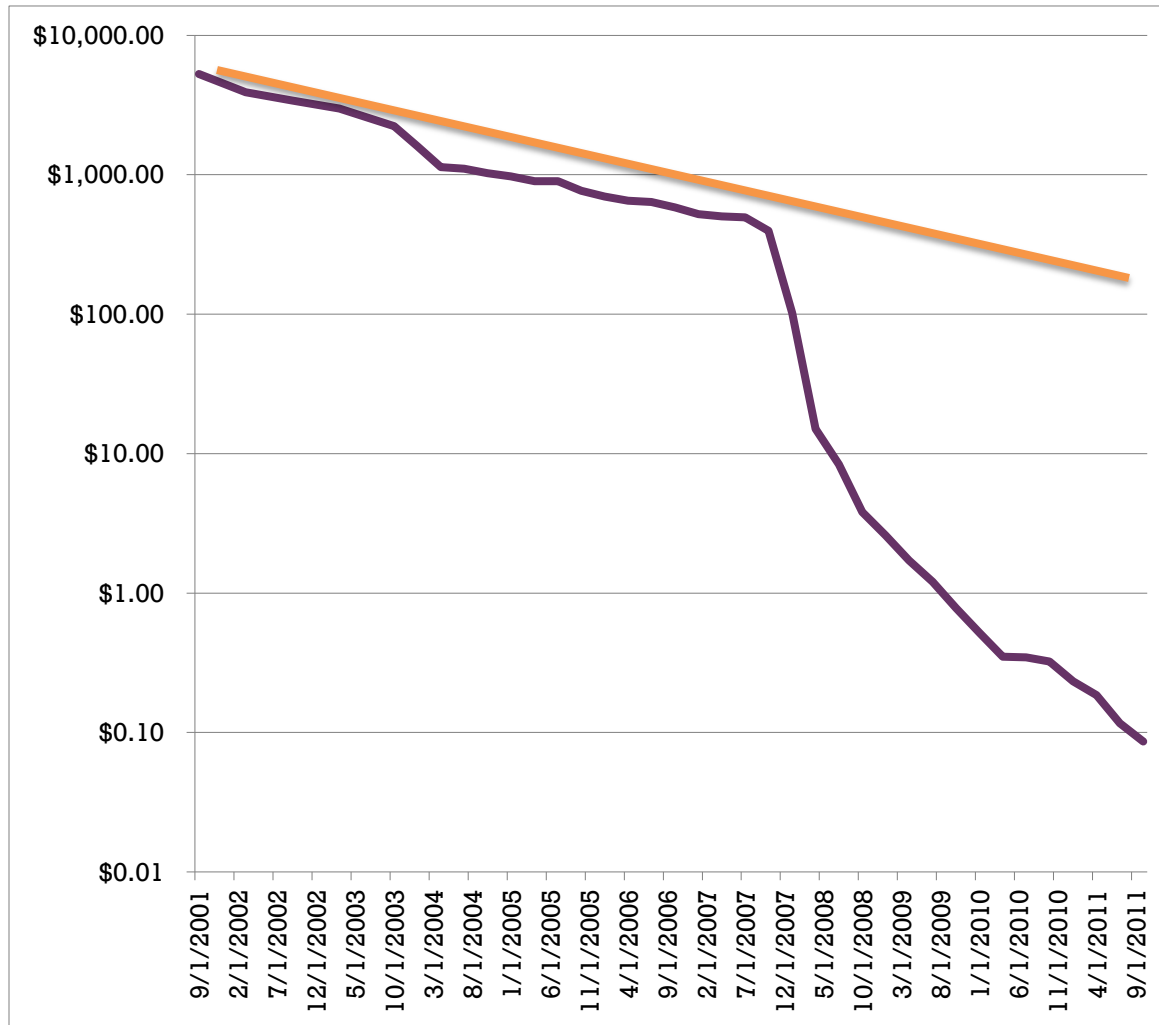
- 1,000 Genomes Project
 - Discovery of rare (1%) SNVs & SVs in normal genomes
- The Cancer Genome Atlas
 - Discovery of sequence variants in major cancers
- Personal Genome Project
 - Discovery of sequence variants associated with medical information
- Human Microbiome Project
 - Study of communities of mixed microbes within human niches
- The Exome Project
 - Discovery of sequence variants in protein coding regions
- Pharmacogenomics Research Network
 - Discovery of sequence variants involving drug-gene interactions

NGS Applications

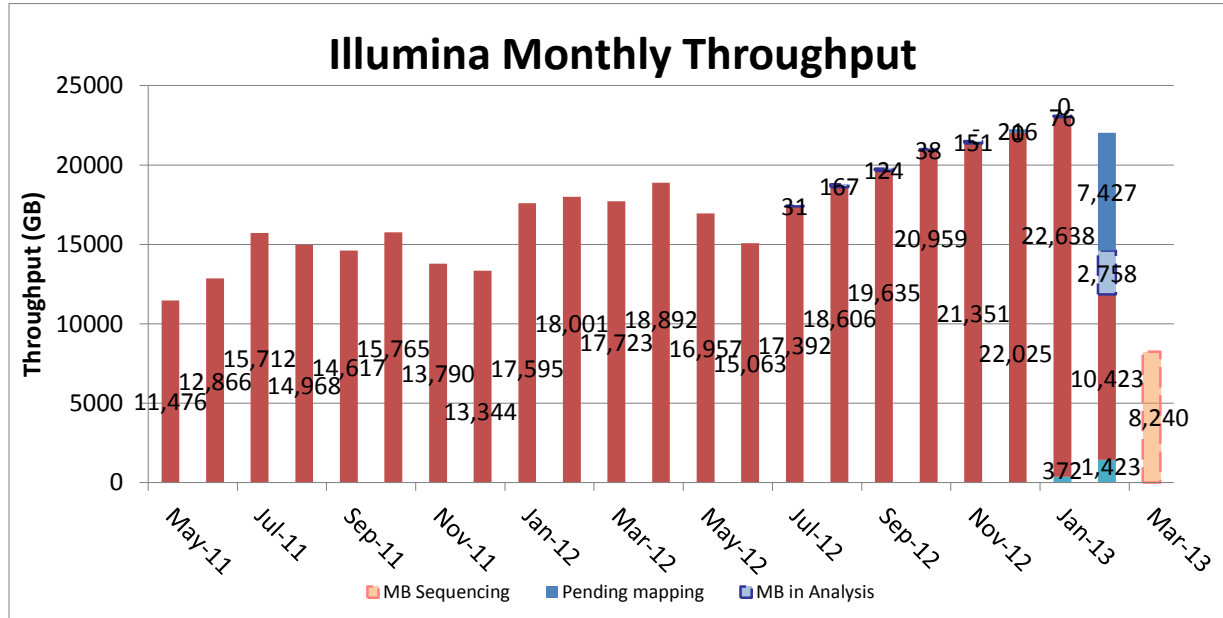
- Human biology
 - Genotype – phenotype interactions
- Pharmacogenomics
 - Drug – gene interactions
- Diagnostics
 - Actionable variants
- Human Microbiome
 - Study of microbe communities
- Forensics
 - Linking suspects to a crime scene
- Many more than time to cover...

Cost of DNA Sequencing (per Mb)

vs. Moore's Law

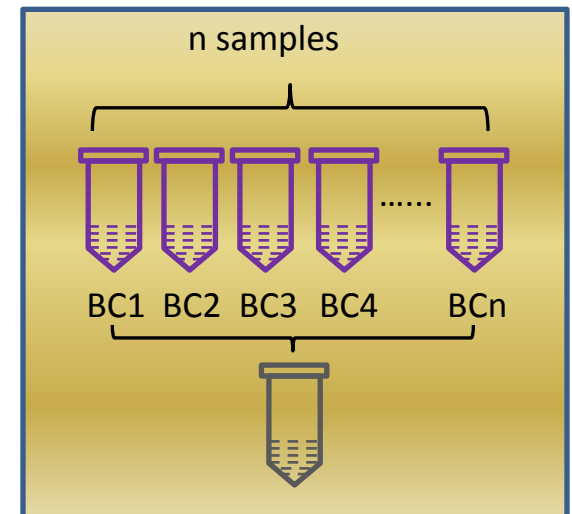


Illumina Sequencing at BCM-HGSC



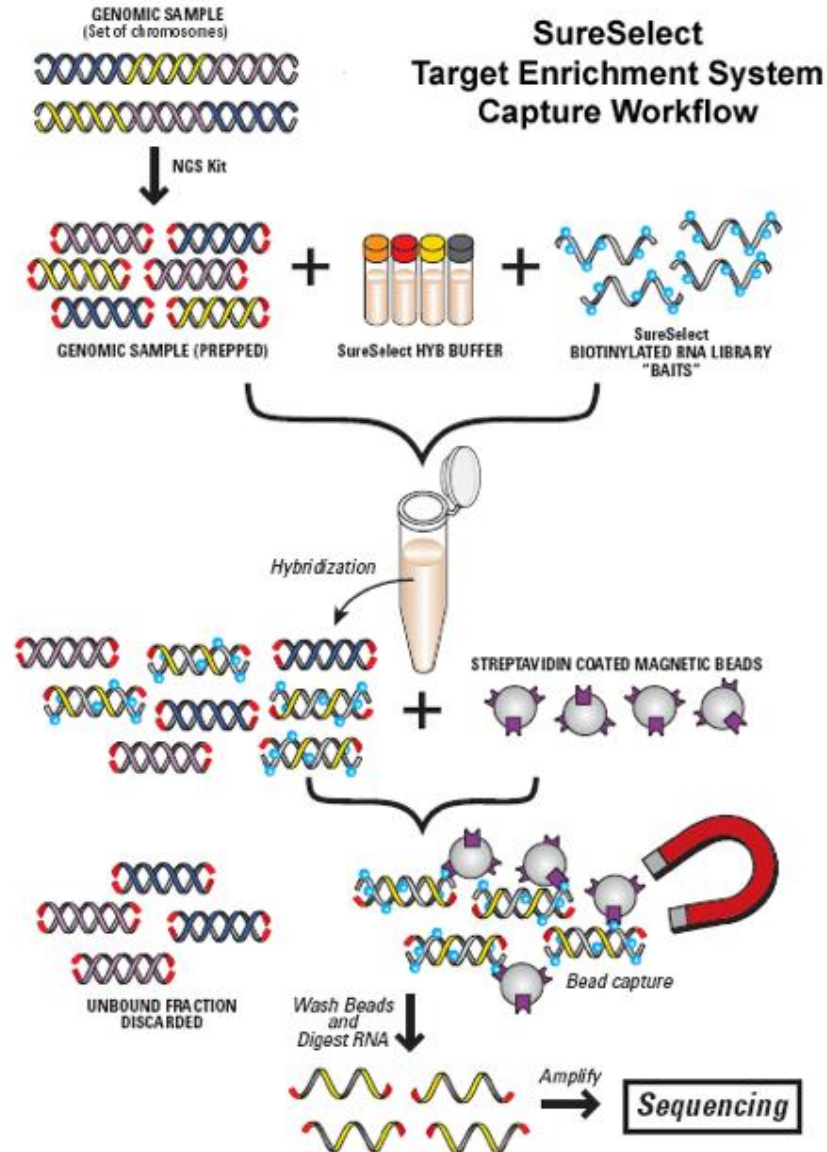
Library Automation

Decrease Reagent cost
 Decrease Labor cost
 Increase capture production



Multiplex Sequence Capture

Capture Sequencing

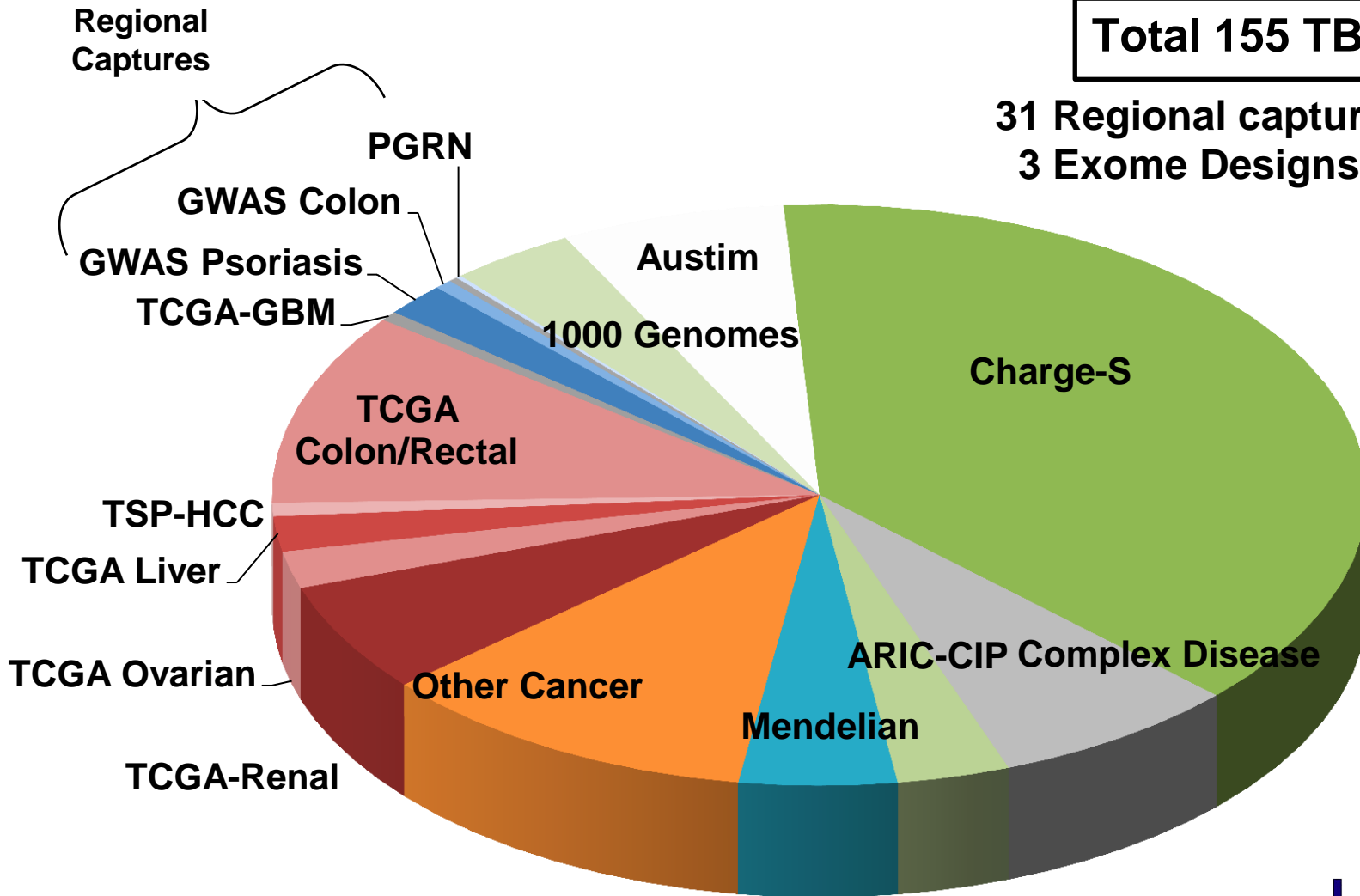


HGSC Capture Projects

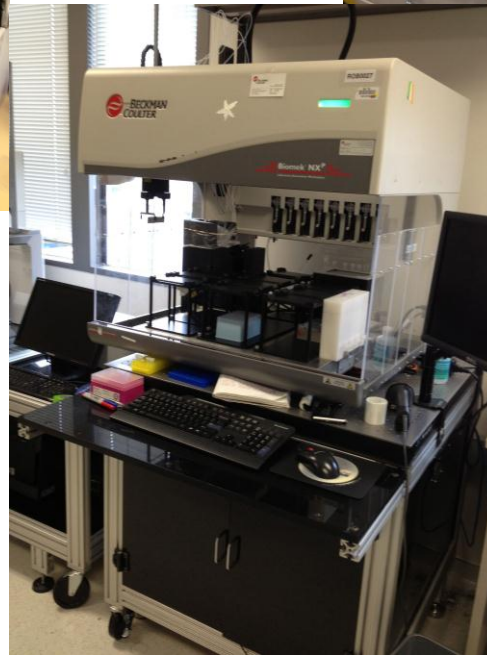
May 2008 to present

Total 155 TB

31 Regional capture designs
3 Exome Designs



Library Automation



Alkek Center for Metagenomics and Microbiome Research

Director: Joseph Petrosino

Mission

- Understand impact of human microbiome on health & disease
- Leverage understanding for therapeutics and diagnostics
- Serve as a hub for U.S. and international activities

Enrich established studies & develop new projects

- Advance sequencing, culturing, analysis technologies
- Enable feasibility/pilot studies
- Develop ties with clinicians to enhance translational impact

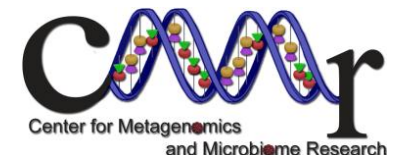
Develop means to study host-microbe interactions

- Advance animal and microbial model systems
- Host GWAS combined with microbiome analyses

Critical mass for systems biology approaches

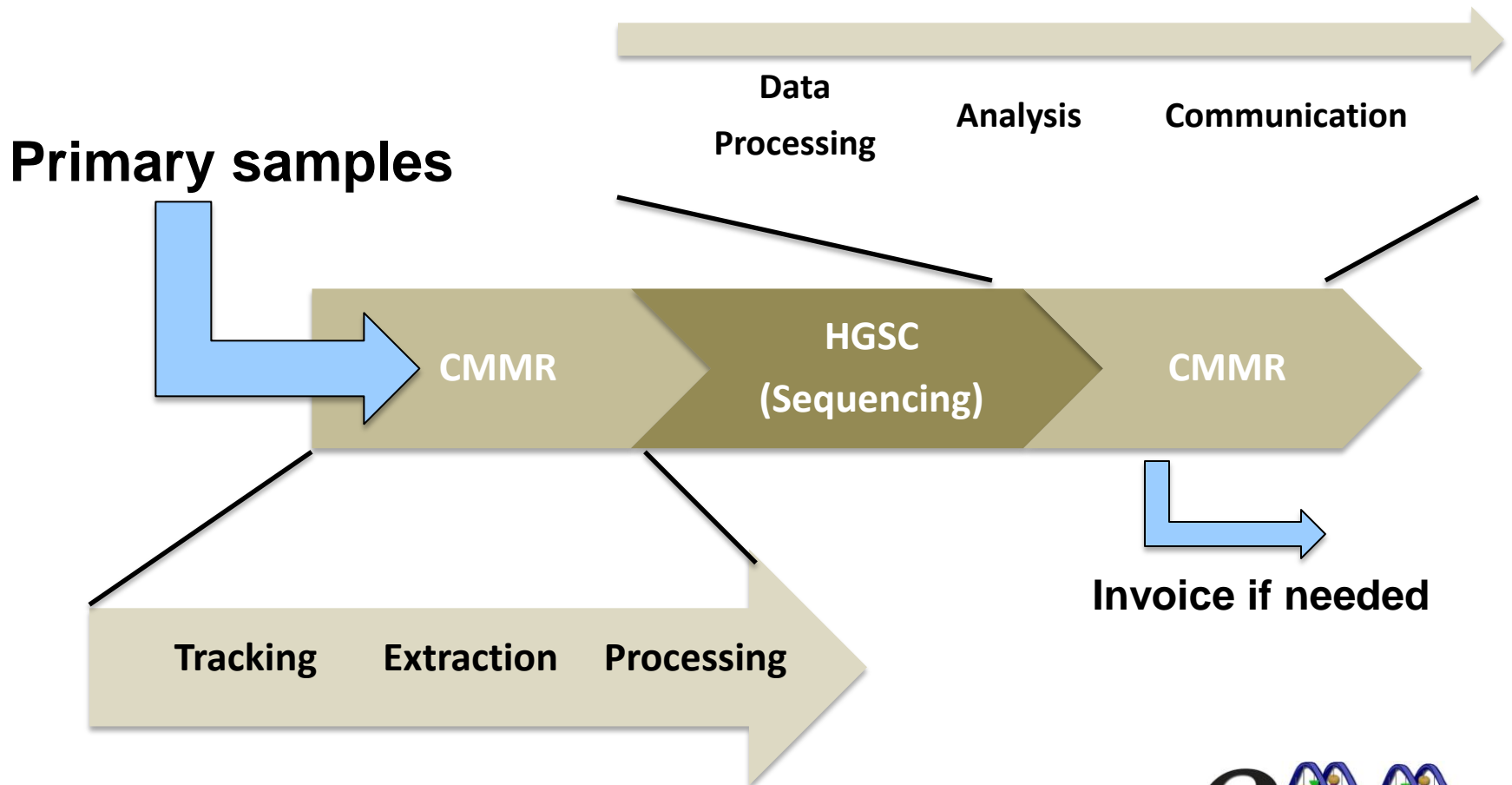
- Genetics, immunology, biochemistry, cell biology, metabolomics

Translate findings to the clinic



Organization and workflow

Metagenomic Sequencing and Analysis Pipeline



Disease and Disease Model Projects

>65 projects and growing...

- Travelers' Diarrhea (DuPont, BCM/St. Luke's Episcopal Hospital)
- Type 1 and Type 2 Diabetes (Fisher-Hoch, UTSPH, nPOD)
- IBS/IBD (DuPont (adults), Versalovic TCH/BCM (children))
- Leukemia (Javier Adachi, MD Anderson)
- Lung Cancer (Ming Hu, U of H)
- HIV-assoc periodontal and GI disease (Jim Katancik, Gena Tribble, UT Dental)
- Clostridium infections in hospitals (Dupont)
- Norovirus/Norwalk virus infection (Estes, BCM)
- Impact of IgA KO (murine) (Metzger, Albany MC)
- Crohn's disease (Britton, MSU)
- Cystic fibrosis (Lipuma, UMichigan)
- Murine GI microbiome (Schloss, UMichigan)

Comparative Genomics



Jeffrey Rogers



Kim Worley



Stephen Richards

Arthropod Genomics History



- With LBL and Celera, HGSC sequenced Chromosomes 3L and X.
 - ~15,000 genes
 - (~150 96 well plates)
- The best biology system?
- Extremely powerful genetic tools
- Sequence is the basis for high throughput molecular biology
- The Rosetta Stone for arthropod molecular biology



Arthropods

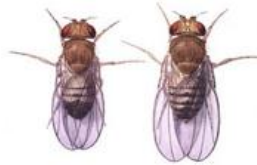


- 15K – pilot of 30
- Drosophila
 - D. melanogaster
 - D. pseudoobscura
 - Genetic Reference Panel
 - ModEncode
 - Additional species



- Agricultural pests and predators
 - Hessian fly
 - Centipede
 - Tobacco Hornworm
 - Cotton bollworm
 - Western Orchard Predatory Mite

- Bees
 - Honey bee
 - Dwarf honey bee
 - Bumble bee



- Vectors
 - Sandfly
 - Blackfly



- Historic
 - Aphid
 - Beetle
 - Wasp
 - Heliconious butterfly



Why Have an i5K Pilot?

LETTERS

edited by Jennifer Sills

Creating a Buzz About Insect Genomes

WHEN E. O. WILSON PROCLAIMED THAT INSECTS ARE THE “little creatures who run the world” (1), he was simply reaffirming the long-recognized dominance of the largest class of animals on our planet. Insects constitute approxi-



this unprecedented volume of data and derive meaning from these genomes.

GENE E. ROBINSON,^{1*} KEVIN J. HACKETT,²
MARY PURCELL-MIRAMONTES,³ SUSAN J. BROWN,^{4,5}
JAY D. EVANS,² MARIAN R. GOLDSMITH,⁶ DANIEL
LAWSON,⁵ JACK OKAMURO,² HUGH M. ROBERTSON,¹
DAVID J. SCHNEIDER⁷

¹Department of Entomology, University of Illinois at Urbana-

- Insects are more than ½ of all living species
- Ants alone are almost ¼ of terrestrial animal biomass
- Pollinate more than 75% of flowering plant species
- Consume or damage more than 25% of all agricultural, forestry and livestock production in the US (\$30 billion per year)
- Parasites and pathogens cause more deaths than all wars in history
 - Insect-borne diseases leading cause of death in children under 5

age of 5 (5). The annual cost of vector-borne diseases worldwide is estimated at almost \$50 billion (6). Clearly, our health and well-being depend on our ability to understand and manage arthropods of agri-

2007).

4. D. Pimentel, Ed., *Pest Management in Agriculture: Techniques for Reducing Pesticide Use: Environmental and Economic Benefits* (John Wiley & Sons, Chichester, UK, 1997).

Creating a Buzz About Insect Genomes

WHEN E. O. WILSON PROCLAIMED THAT INSECTS ARE THE “little creatures who run the world” (1), he was simply reaffirming the long-recognized dominance of the largest class of animals on our planet. Insects constitute approximately 53% of all living species, with one group alone (the ants), accounting for almost a quarter of terrestrial animal biomass (2). These tiny creatures also exert outsized impacts on human affairs. By serving as pollinators to more than 75% of flowering plant species (3), insects are essential to the maintenance and productivity of natural and agricultural ecosystems. But other insects consume or damage more than 25% of all agricultural, forestry, and livestock production in the United States, costing our economy more than \$30 billion annually (4). These losses occur despite more than 150 years of concerted efforts to prevent them. Insects and other arthropods not only affect our food supply, they also carry disease. Parasites and pathogens carried by insects and their relatives have led to more loss of human life than all wars in recorded history; even today, insect-borne diseases are a leading cause of death of children under the age of 5 (5). The annual cost of vector-borne diseases



this unprecedented volume of data and derive meaning from these genomes.

GENE E. ROBINSON,^{1*} KEVIN J. HACKETT,² MARY PURCELL-MIRAMONTES,³ SUSAN J. BROWN,^{4,5} JAY D. EVANS,² MARIAN R. GOLDSMITH,⁶ DANIEL LAWSON,³ JACK OKAMURO,² HUGH M. ROBERTSON,³ DAVID J. SCHNEIDER⁷

¹Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. ²USDA Agricultural Research Service, Beltsville, MD 20705, USA. ³USDA National Institute of Food and Agriculture, Washington, DC 20250, USA. ⁴Division of Biology, Kansas State University, Manhattan, KS 66506–4190, USA. ⁵Arthropod Genomics Consortium and European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. ⁶College of the Environment and Life Sciences, University of Rhode Island, Kingston, RI 02881, USA. ⁷USDA Agricultural Research Service, Ithaca, NY 14853, USA.

*To whom correspondence should be addressed. E-mail: generobi@illinois.edu

References and Notes

1. PBS, *Nova*. Transcripts: “Little creatures who run the world” (www.pbs.org/wgbh/nova/transcripts/2203crea.html).
2. E. O. Wilson, *The Diversity of Life* (W.W. Norton, New York, 1992).
3. National Research Council, *Status of Pollinators in North America* (National Academy of Sciences, Washington, DC, 2007).
4. D. Pimentel, Ed., *Pest Management in Agriculture: Tech-*

Why an i5K Pilot?

- Our aim is to identify the molecular components of arthropod life
- **“This project is aimed at sequencing and analyzing the genomes of all species known to be important to worldwide agriculture and food safety, medicine, and energy production; all species used as models in biology; the most abundant insects in world ecosystems; and, to achieve a deep understanding of arthropod evolution, representatives of insect relatives in every major branch of arthropod phylogeny.”**
- **5000 is a medium sized number**
- **Stephen Richards suggests is most likely to be 100 –500 projects of 10-50 species, or perhaps a small part of a much larger project?**

Which Species were Chosen?

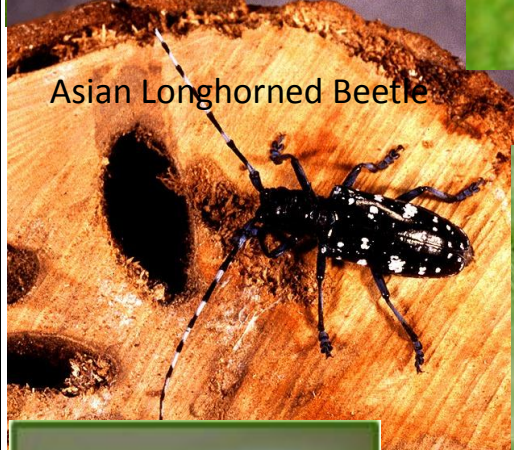
Brown Marmorated Stinkbug



Mediterranean Fruit fly



Asian Longhorned Beetle



Agricultural Pests

Western Flower Thrips



Glassy

Winged Sharpshooter



Sheep Blowfly



Mantis Impersonating lacewing



Hackberry
petiole gall psyllid



Bull-headed
Dung Beetle



Interesting Phenotypes

Which Species were Chosen?

House spider



Brown Recluse

Western Black Widow



Bark Scorpion



German Cockroach



Bedbug

Urban Pests



Parasitic wasp
(Tricogramma)



Parasitic wasps
And wasp out groups



Turnip Sawfly

Spiders

The Plan for the i5K Pilot

How to Sequence 10-50 Arthropod

- Reduce sequence polymorphism
 - Haploid individual
 - Sib-sib mating
 - Single individual
- Standardized Illumina HiSeq sequencing plan
- RNAseq – 3 tissues or lifestages, 5 Gb each
- Allpaths assembler, Atlas-Link and Atlas-GapFill assembly improvement tools (see poster P0968)
- Maker 2.0 automated annotation pipeline
- Building the community around these initial datasets



page discussion view source history



i5k Insect and other Arthropod Genome Sequencing Initiative

The *i5k* initiative plans to sequence the genomes of 5,000 insect and related arthropod species over the next 5 years. This project is transformative because it aims to sequence the genomes of all insect species known to be important to worldwide agriculture, medicine, and energy production; all those used as models in biology; the most abundant in world ecosystems; and represents a new branch of insect phylogeny so as to achieve a deep understanding of arthropod evolution and phylogeny.

The *i5k* initiative will be broad and inclusive and thus is seeking to involve scientists from around the world and obtain funding from academia, governments, industry, and private sources. To get involved please [sign up to this wiki](#), let people know which species you are interested in and maybe nominate some for sequencing as part of the *i5k* effort.

Introduction to the *i5k* Insect and other Arthropod Genome Sequencing Initiative - AGC Talk >> <http://arthropodgenomes.org/wiki/File:i5kFlyer010312.pdf> (Open this link in a separate tab or window, click the file name, then view or download the file.)

Please see attached: First Announcement for *i5k* Community Workshop May 30-31, 2012 >> <http://arthropodgenomes.org/wiki/File:i5kFlyer010312.pdf> (Open this link in a separate tab or window, click the file name, then view or download the file.)

Notice the newly added: Criteria for Prioritization of Arthropods/Pre-sequencing Informatics >> <http://arthropodgenomes.org/wiki/File:i5kFlyer010312.pdf>

Notice the newly added *i5k* Brochure of August, 2012 >> http://arthropodgenomes.org/w/images/b/b2/i5k_flier_Aug-2012.pdf

Use the links in the left-hand margin to navigate through the wiki to find out where much of the work is being accomplished and

i5k Coordinating Group*

*This list is formative: additional representatives will be recruited

- [Gene E. Robinson](#), University of Illinois at Urbana-Champaign
- [Kevin J. Hackett](#), USDA, Agricultural Research Service, Beltsville, Maryland
- [Susan J. Brown](#), Kansas State University and Arthropod Genomics Consortium



search

Go Search

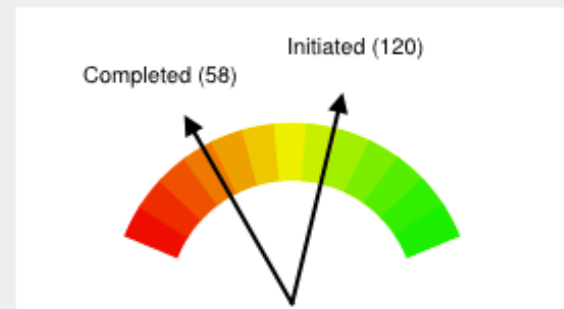
- navigation
- Species
 - People
 - Organisations
 - Resources & DBs
 - Collections
 - Documents
 - Help

- i5k
- i5K home
 - i5K working groups
 - i5K nominated species
 - Green/White papers

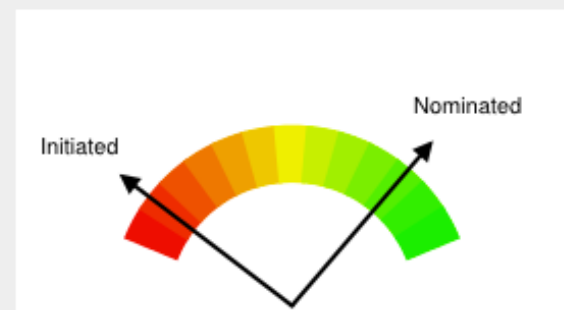
- data summaries
- Sequenced genomes



The first 200 genome projects



The first 1000 genome projects

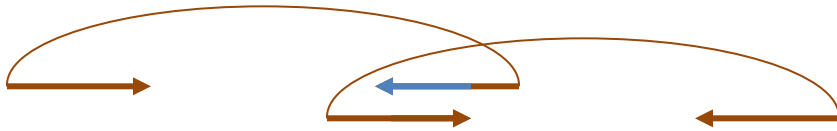


Sequence Assembly

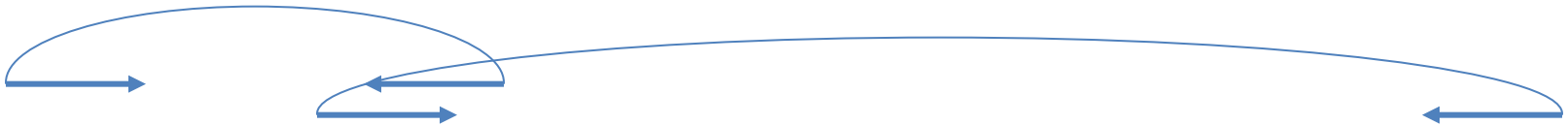
Basic unit - two end sequences with clone insert sized gap between



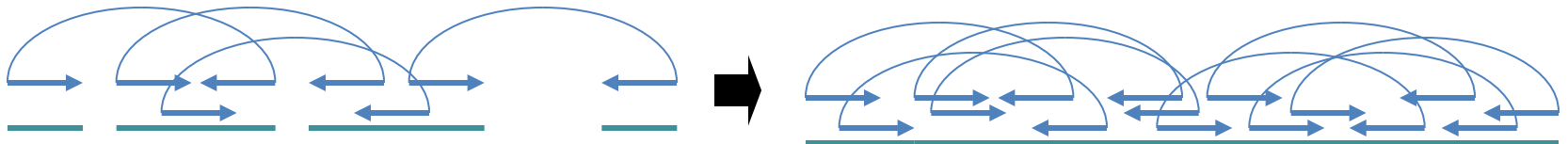
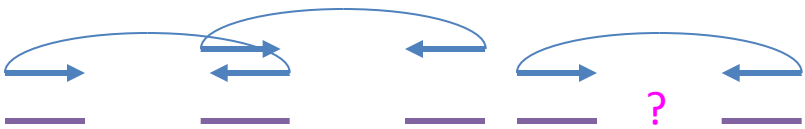
Overlaps - based on sequence similarity

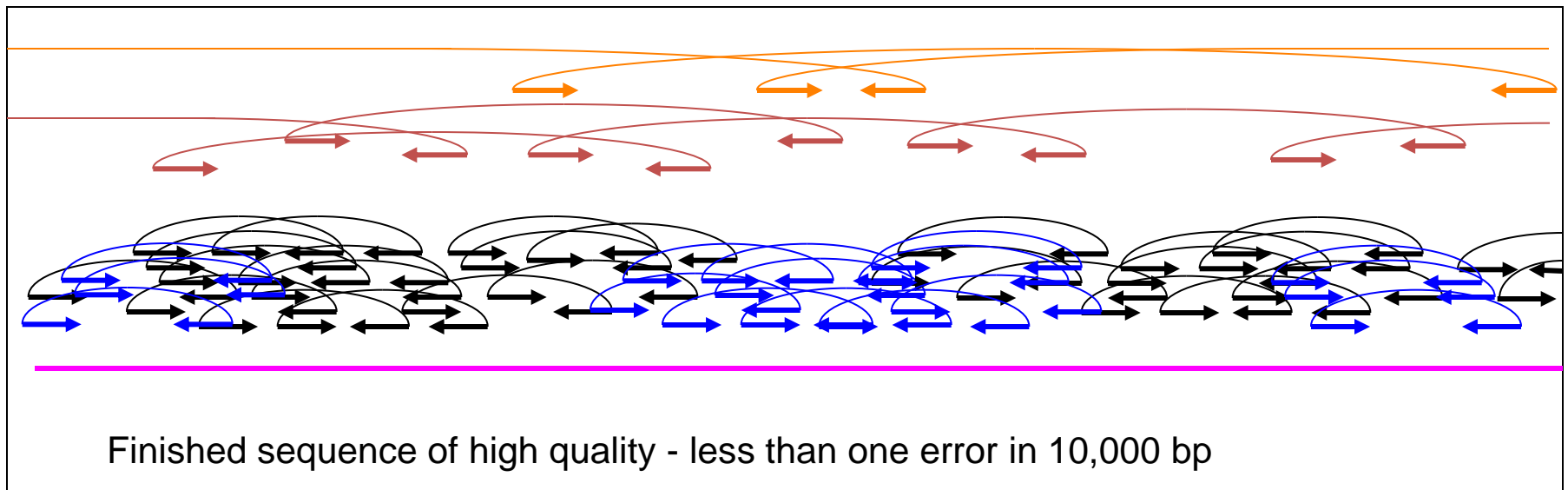
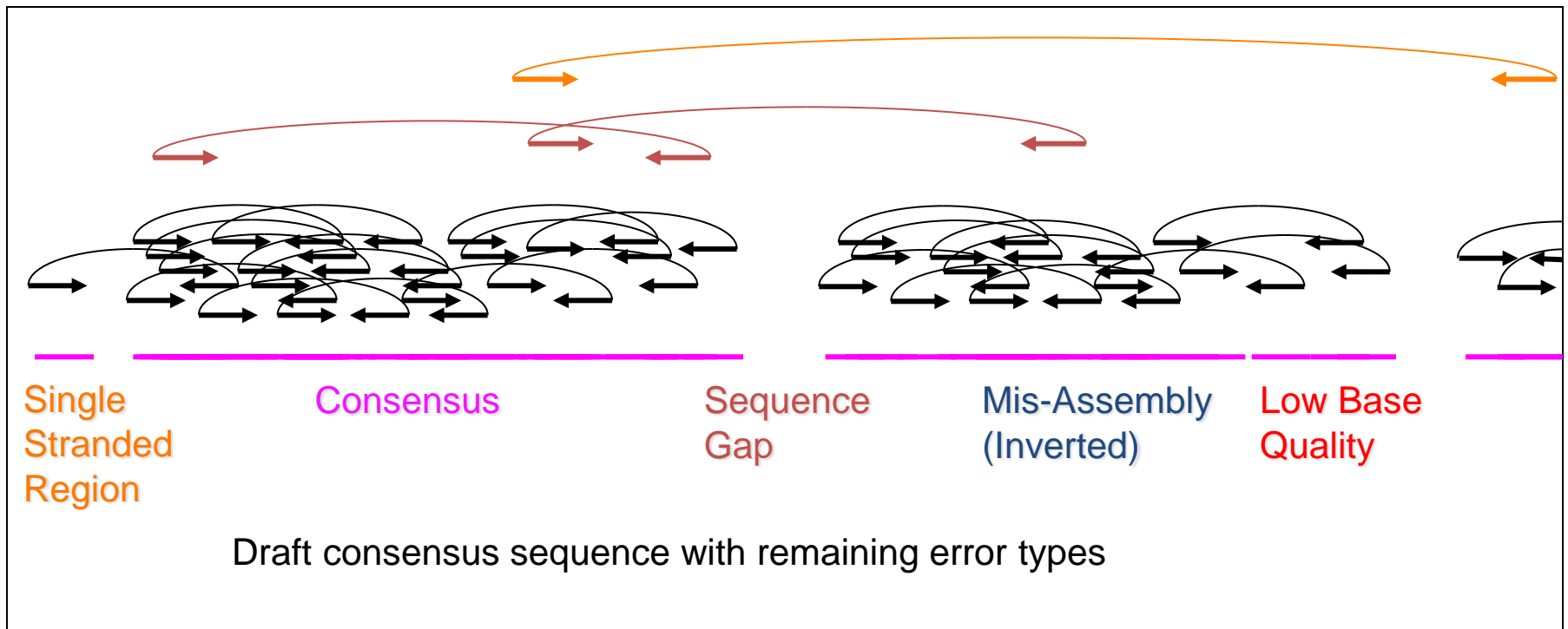


Different clone insert sizes step across features of different sizes



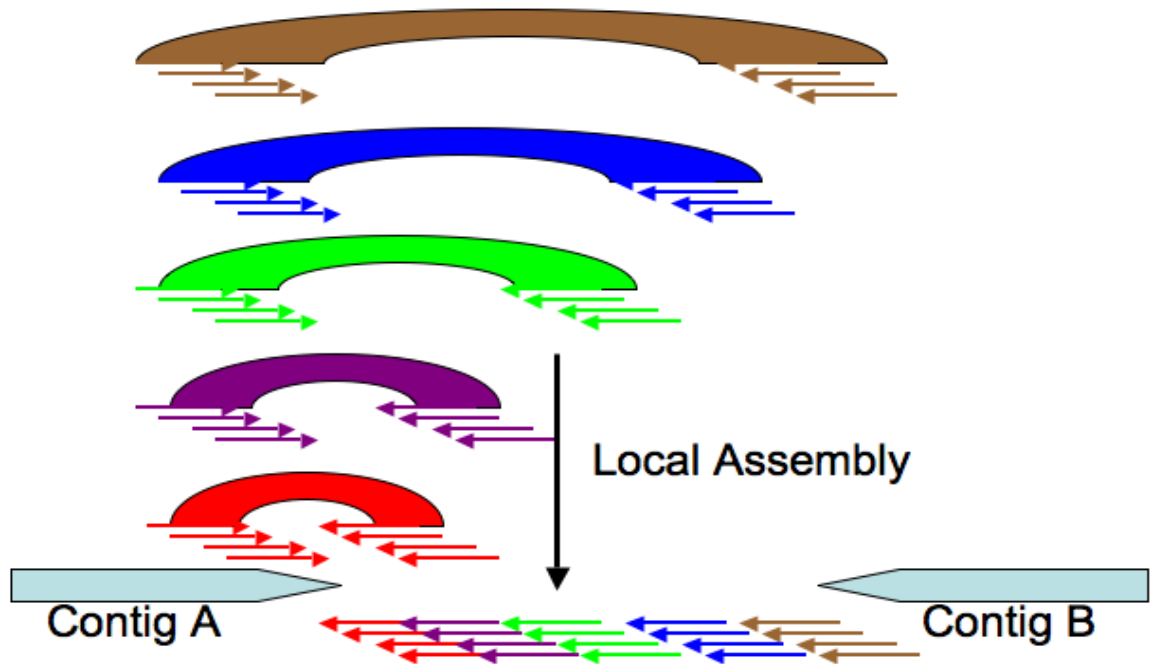
Greater coverage improves consensus sequence assembly





Assembly improvement

- Assemblies improve with more data, longer reads, and a larger variety of library insert sizes
- Using paired end data and local assembly we can improve some gaps, and improve the scaffolding.
- **BCM HGSC code:**
 - Atlas-gap-closer
 - Atlas-link



Improved Genome Honey bee



- Originally published in 2006
 - No nearby species
 - Bimodal GC content
 - Genes in AT rich regions
 - Little mRNA data
 - 10,000 OGS + 15,000 FgenesH *ab initio*
- Added 454 and SOLiD data
 - 3.6x fragment 454
 - 1.3x 2.75 kb mate pair 454
 - 20x SOLiD mate pair data
- Genome improved
 - Contig N50 40 to 45 kb
 - Scaffold N50 362 to 997 kb
 - 5.5% more anchored to linkage groups
- 5,000 more genes
 - 1/5 from assembly improvements
 - 4/5 from new RNAseq data
 - ? From more protein orthologues
 - ? From more comparative species data

Primate Project Summary



- Sanger projects with WU GI, in analysis
 - Marmoset
 - Gibbon
- Mixed platform projects
 - Baboon (Sanger, 454, Illumina)
 - Sooty mangabey (Illumina and PacBio)
 - Mouse Lemur (Sanger 2x and Illumina)
- Nine new Illumina *de novo* genomes
 - Pig-tailed macaque, Chinese rhesus macaque, white-fronted capuchin, buff-headed capuchin, gelada, drill, patas monkey, black and white colobus, and sifaka.

Rhesus monkeys
Photo by Shane Moore/Animals Animals, National Geographic

Baboon Panu_2.0 Assembly (*Papio anubis*)

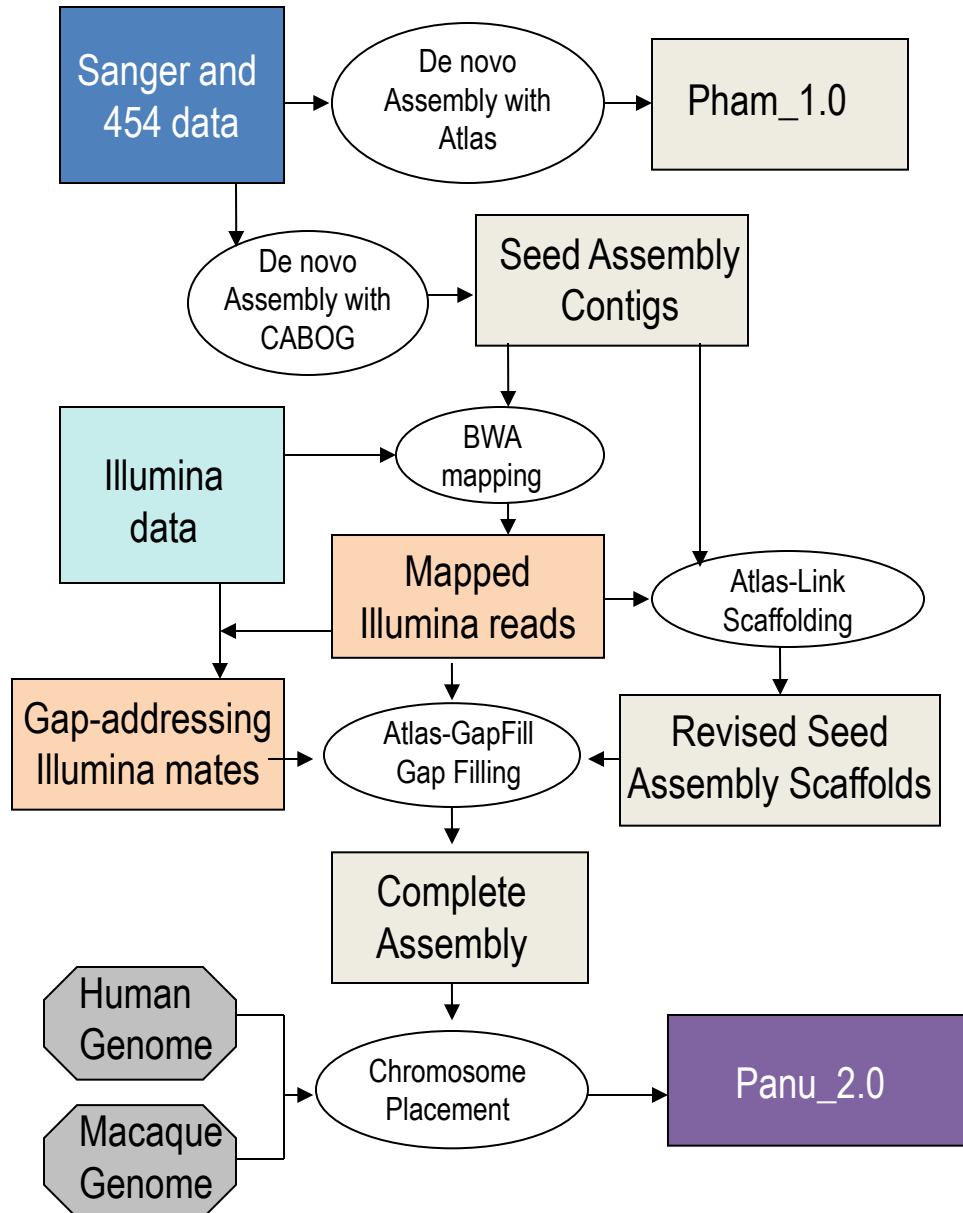


Photo by Muhammad Mahdi Karim *Papio anubis*,
Ngorongoro Conservation Area, Tanzania, June 2010

- Reads
 - 2x Sanger
 - 3x 454 frag.
 - Illumina 30x sequence in 3kb MP
 - Illumina 160x in 240 bp frag. PE
- Assembly
 - CABOG using Sanger & 454 data
 - Illumina data mapped using BWA
 - Atlas-Link
 - Atlas-GapFill
 - Placed by Mummer comparison to Rhesus macaque genome
 - Scaffolds split on discontinuities with low clone coverage
 - 323 total (0.45%)

Baboon Genome Assembly

- Sanger & 454 assembly (Pham_1.0)
- New Seed assembly
- Mapped Illumina data
- Atlas-Link improved scaffolding
- Atlas-GapFill identify gap adjacent reads and mates, local assembly of gaps, substitution of gap-filling contigs
- Chromosome placement using macaque and human assemblies without disagreeing mate pair data
- Version Panu_2.0



Sooty Mangabey Project



- Joint project with Guido Silvestri
- Illumina data (incomplete)
 - Preliminary assembly
 - Preliminary scaffolding
- Pac Bio data (~8x)
 - Preliminary PacBio gap filling
- Planned strand-specific RNA seq from liver, kidney, colon, spleen, lung, bone marrow, testis, hippocampus, cerebellum, frontal cortex, lymph node

Sooty Mangabey Production



Library Type	Total (Mb)	Sequence Coverage	Clone Coverage
1Kb	75,985	25	127
2Kb	56,366	19	188
3Kb	27,844	9	139
5Kb	75,730	25	631
8Kb	34,144	11	455
180bp	147,397	49	44
500bp	113,554	38	95
260bp	35,700	12	12
204bp	26,785	9	12

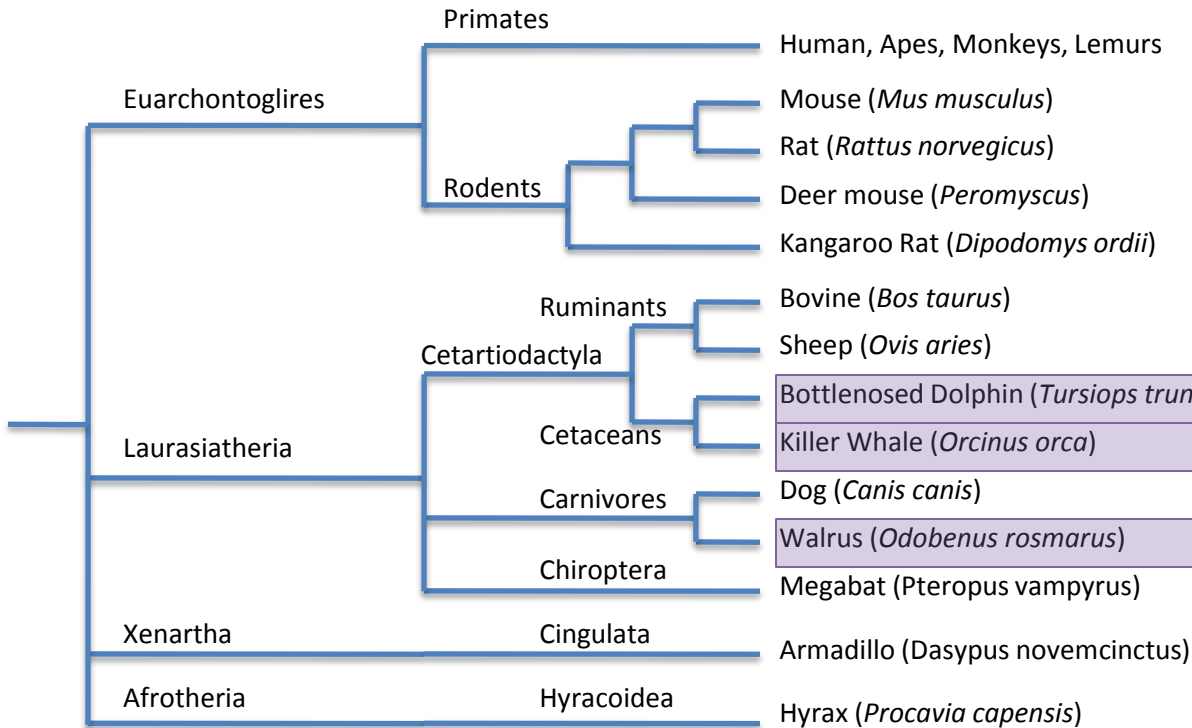
Sooty Mangabey Genome Assembly



Sooty mangabey photo by Yerkes NPRC

- Preliminary Assembly
 - Incomplete data
 - All-Paths-LG initial assembly
 - Contigs >300 bp N50 = 30.2kb
 - Scaffold N50 3.28 Mb
 - Total size 2.65 Gb, 2.84 Gb with gaps
- Atlas-Link
 - 5 days for mapping
 - 3-4 days for scaffolding on MrGAC
- Atlas-GapFill
 - 3 weeks

Marine Mammals



Tursiops truncatus, NASA photo.



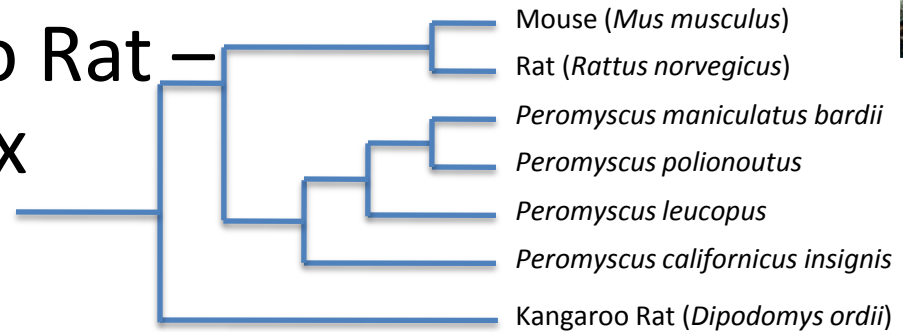
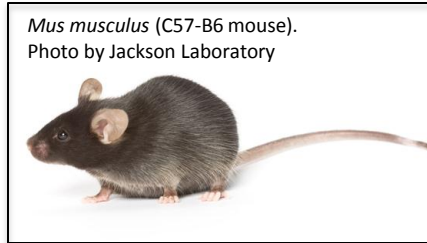
Orcinus orca. Photo by Robert Pittman, NOAA.



Odobenus rosmarus. Photo by Captain Budd Christman, NOAA.

Rodents

- Mouse – Finished
- Rat – HQ draft
- Peromyscus – ongoing
- Kangaroo Rat – Sanger 2x upgrade



Peromyscus

- Natural variation in wild mouse populations
- Two most abundant N. American mammals (*P. maniculatus* and *P. leucopus*)
- Better for study of phenotypes and variation than other rodents like squirrel and guinea pig
- Can be grown in lab colonies
- Movement disorders, autism, epilepsy, stereotypical behavior, cancer, alopecia, diabetes, alcohol metabolism, aging, genomic imprinting and placentation
- Partner fidelity, nests vs. burrows, coloration, photo period sensitivity, altitude adaptation
- Public health – hantavirus, Lyme disease and other tick-borne illnesses



BCM-HGSC Assembly Experience

One of the Best Genome Assembly Teams in the World

- Assemblathon 2 competition
 - 3 real data sets (bird, fish, snake)
 - 21 teams competing
 - BCM-HGSC Competed
 - One of only 5 teams that competed with all three data sets
 - One of only 2 teams that used all the available data types for the bird
 - The only team to do both
 - BCM-HGSC submissions ranked
 - First (and second) for the bird data (14 assemblies, 11 teams)
 - First for the fish data (16 assemblies, 11 teams)
 - Positively scoring for the snake data (11 assemblies, 11 teams)
- Experience
 - With different sequencing technologies, and with Illumina
 - Infrastructure to support

The Pac-Bio Story





PBJelly

OPEN ACCESS Freely available online

PLOS ONE

Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology

Adam C. English*, Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, Donna M. Muzny, Jeffrey G. Reid, Kim C. Worley, Richard A. Gibbs

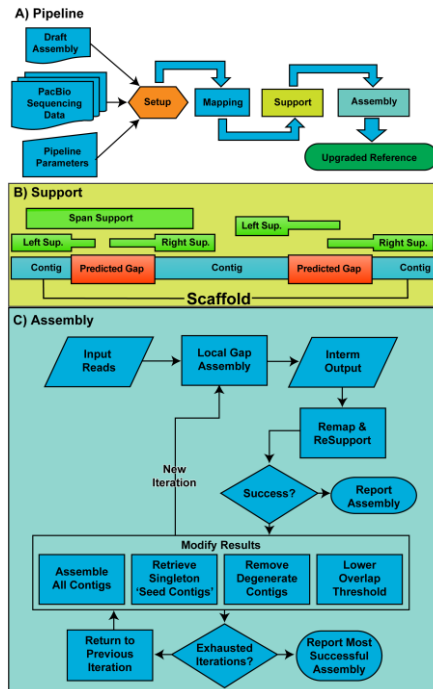
Department of Molecular and Human Genetics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, United States of America

Address Draft Assembly Issue Due To:

- Sequencing chemistry biases
- Genomic repeat structure
- Genome Polymorphism

Genomes Used for Testing

- *Drosophila melanogaster*: artificial Pac-Bio reads added to an artificial draft assembly
- *Drosophila pseudoobscura*: draft assembly containing > 6000 gaps
- Parakeet Genome: used as part of the Assemblathon
- Sooty Mangabey: initial mixed library assembly contained > 186K gaps



Results

Drosophila melanogaster: Closed 99% of the artificial gaps;
assessed accuracy versus the finished sequence

Drosophila pseudoobscura: 24x coverage addressed 99% of the
gaps, closed 69% and improved 12%

Parakeet Genome: 4x mapped coverage addressed 63% of the
gaps, closed 32% and improved 69%

Sooty Mangabey: 6.8x coverage addressed 97% of the gaps,
closed 66% and improved 19%



Microcebus murinus
Grey mouse lemur



Photo Credit: Duke Lemur Center

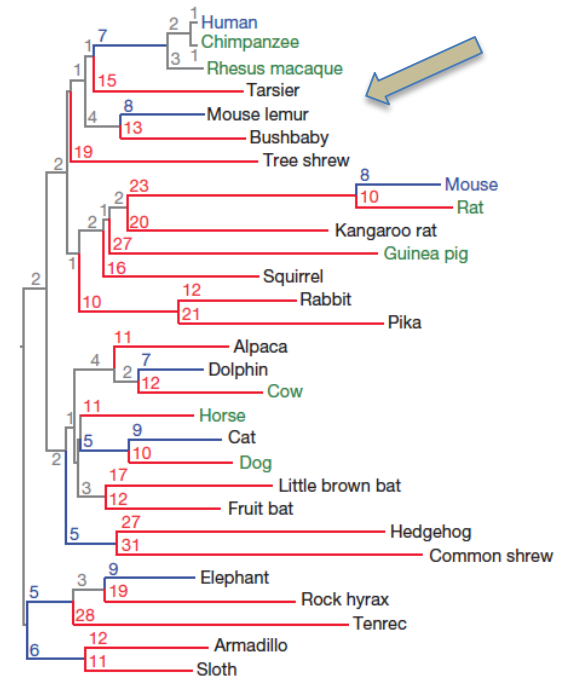
Mouse Lemur Genome Project

Previous genome analysis: 2x Sanger sequence assembly

A high-resolution map of human evolutionary constraint using 29 mammals

Kerstin Lindblad-Toh^{1,2}, Manuel Garber^{1*}, Or Zuk^{1*}, Michael F. Lin^{1,3*}, Brian J. Parker^{4*}, Stefan Washietl^{3*}, Pouya Kheradpour^{1,3*}, Jason Ernst^{1,3*}, Gregory Jordan^{5*}, Evan Mauceli^{1*}, Lucas D. Ward^{1,3*}, Craig B. Lowe^{6,7,8*}, Alisha K. Holloway^{9*}, Michele Clamp^{1,10*}, Sante Gnerre^{1*}, Jessica Alföldi¹, Kathryn Beal⁵, Jean Chang¹, Hiram Clawson⁶, James Cuff¹¹, Federica Di Palma¹, Stephen Fitzgerald⁵, Paul Flicek⁵, Mitchell Guttman¹, Melissa J. Hubisz¹², David B. Jaffe¹, Irwin Jungreis³, W. James Kent⁹, Dennis Kostka⁹, Marcia Lara¹, Andre L. Martins¹², Tim Massingham⁵, Ida Moltke⁴, Brian J. Raney⁶, Matthew D. Rasmussen³, Jim Robinson¹, Alexander Stark¹³, Albert J. Vilella⁵, Jiayu Wen⁴, Xiaohui Xie¹, Michael C. Zody¹, Broad Institute Sequencing Platform and Whole Genome Assembly Team†, Kim C. Worley¹⁴, Christie L. Kovar¹⁴, Donna M. Muzny¹⁴, Richard A. Gibbs¹⁴, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team‡, Wesley C. Warren¹⁵, Elaine R. Mardis¹⁵, George M. Weinstock^{14,15}, Richard K. Wilson¹⁵, Genome Institute at Washington University†, Ewan Birney⁵, Elliott H. Margulies¹⁶, Javier Herrero⁵, Eric D. Green¹⁷, David Haussler^{6,8}, Adam Siepel¹², Nick Goldman⁵, Katherine S. Pollard^{9,18}, Jakob S. Pedersen^{4,19}, Eric S. Lander¹ & Manolis Kellis^{1,3}

476 | NATURE | VOL 478 | 27 OCTOBER 2011



Mouse Lemur Genome Project

*Human Genome Sequencing Center, BCM
and Duke Lemur Center*

HGSC has used DLC sample to generate
>150x coverage using Illumina 100bp
reads

Reads from multiple paired-end and
mate-pair libraries

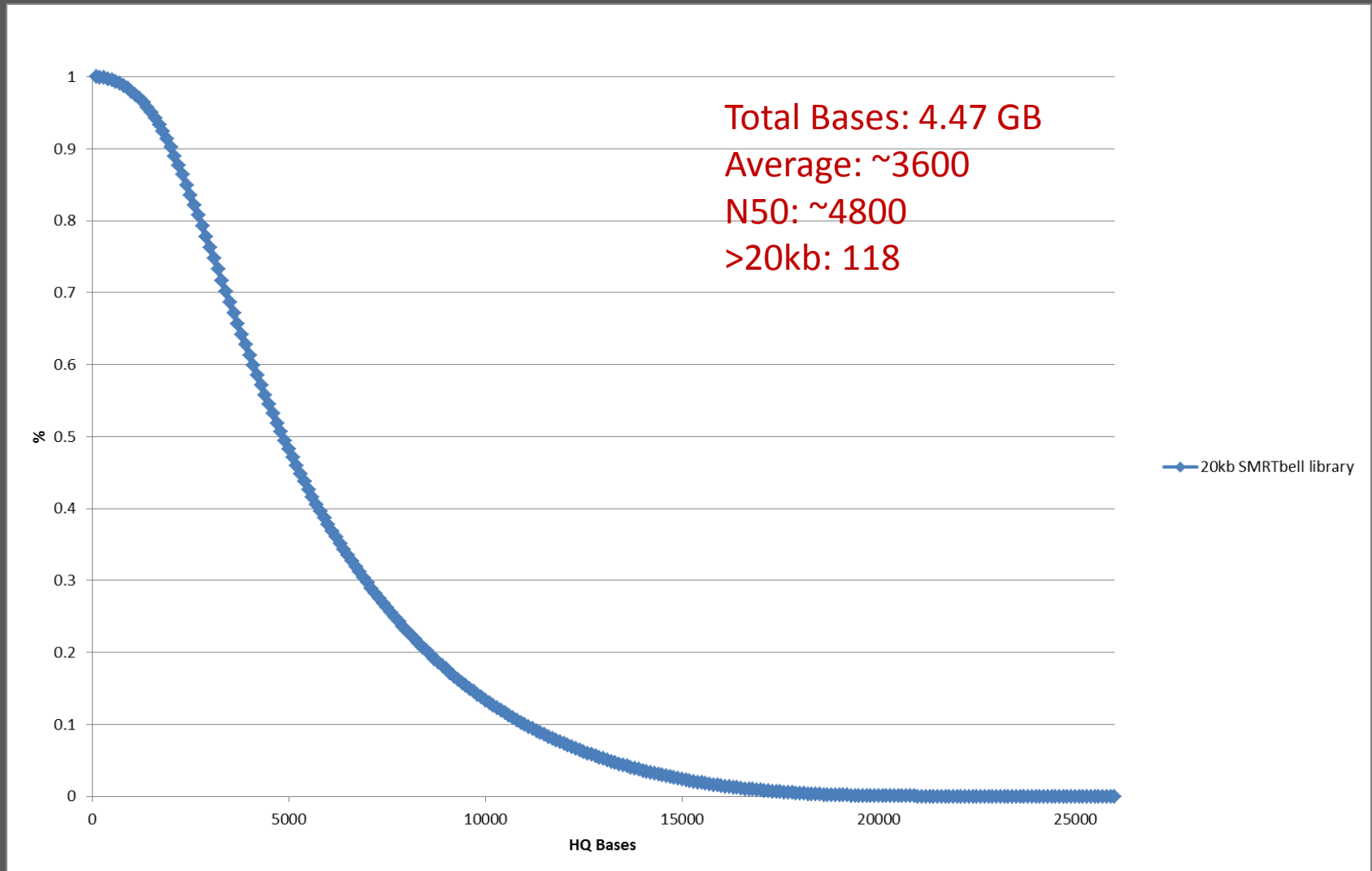
Contig N50: 56kb
Scaffold N50: 3.36 Mb
(includes 2x Sanger data)

PacBio and HGSC have produced about
8x coverage with RS long reads

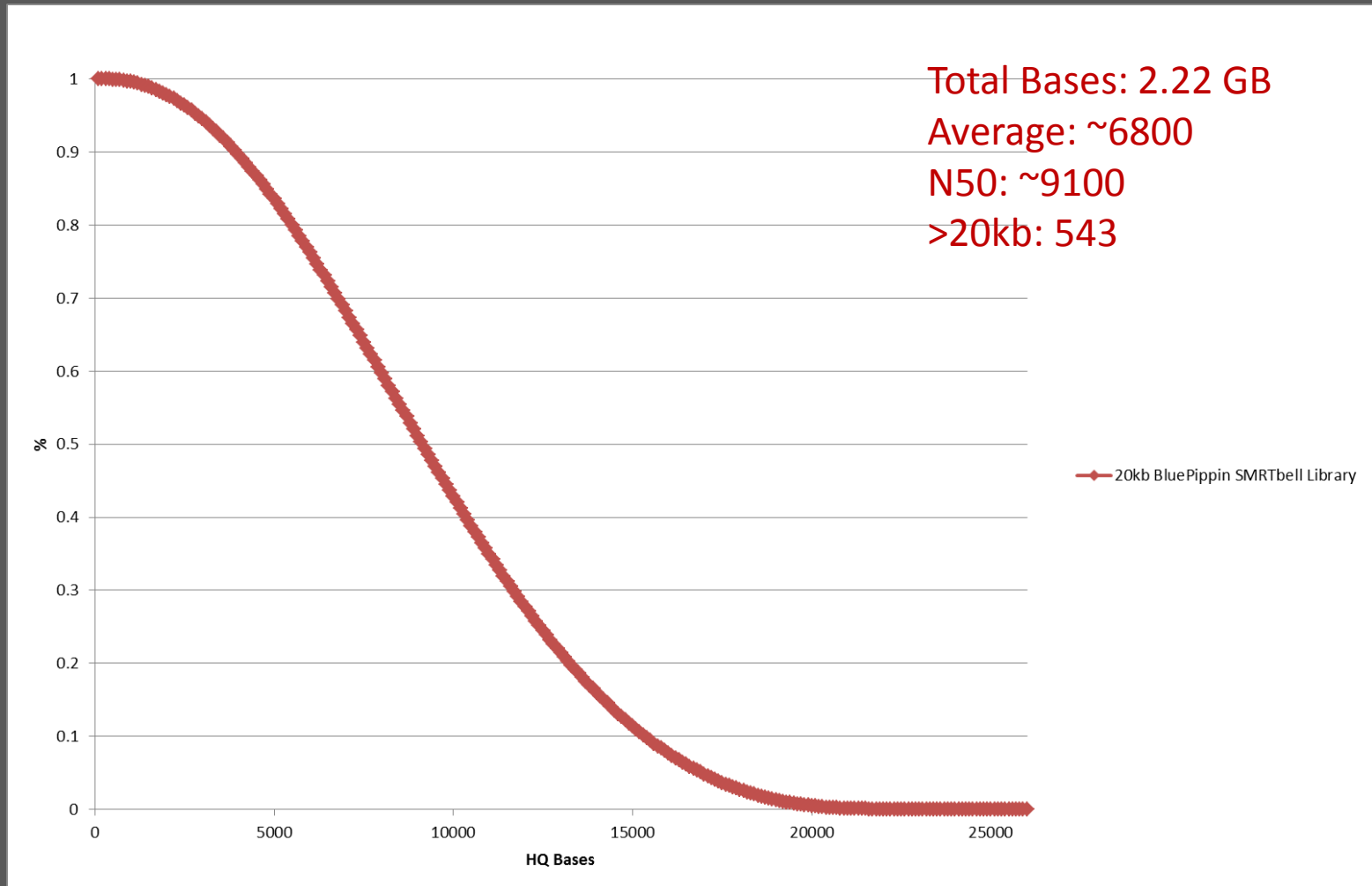


Photo credit: Duke Lemur Center

40 Cells of 20kb SMRTbell generated by PacBio

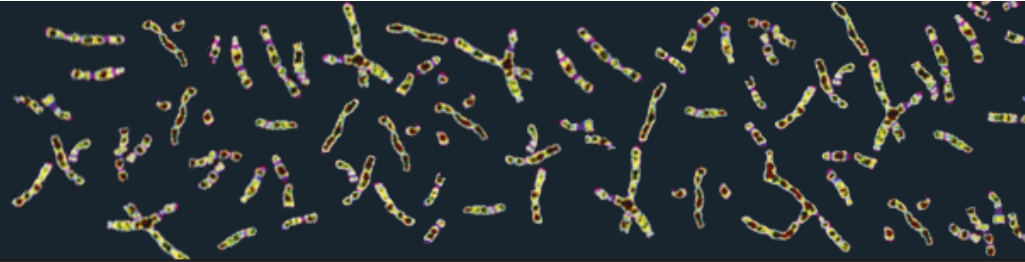


25 Cells of 20kb BluePippin SMRTbell generated by PacBio





1000 Genomes Project



1000 Genomes

A Deep Catalog of Human Genetic Variation

[Home](#) [About](#) [Data](#) [Analysis](#) [Participants](#) [Contact](#) [Browser](#) [Wiki](#)

LATEST ANNOUNCEMENTS

July 2010 Data Release

20 JULY 2010

Pilot Project Variant call release

Variant Calls from the three pilot projects are now available in VCF 4.0 format. This release includes SNPs, short indels and large scale structural variants. All 1000 genomes pilot project files reference the NCBI build 36 assembly of the human genome

Data access links: [EBI](#) / [NCBI](#)

Link to additional information: [README file](#)

Recent project announcements

4 AUGUST 2010 [New sequence data is available](#)

The latest release of sequence data from the 1000 Genomes full project is now available. The new sequence.index file can be found at: [20100804.sequence.index](#)

Data access links: [EBI](#) / [NCBI](#) / [Instructions for data download and Aspera](#)

Links to additional information: [List of new index and statistics files](#) / [Sequence index file format](#)

19 JULY 2010 [Release of full project alignment files](#)

The alignments based on the [20100611.sequence.index](#) have been released. There are both new BAM files and updated BAM files with more data were added. For the case of updated files, the older, redundant files have been withdrawn.

LOG IN

Username:

Password:

([Send me my password](#))

LINKS



[All Project Announcements](#)



[Sample and Project Information](#)



[Media Archive](#)

The enormity of background variation:

	Filter	Total variation	Known	Novel
Watson	Raw	14,829,087	3,283,273	11,545,814
	1	4,427,488	2,815,322	1,612,166
	2	3,971,513	2,752,991	1,218,522
	3	3,325,725	2,704,029	621,696
Venter	4	3,470,669	2,726,935	743,734

- ~ 25 Mb of DNA missing from reference, in JDW
- Sequence reads reveal CNVs
- 16% of Watson SNPs are novel
- 15% of Venter SNPs are novel
- ~10,500 ns variants
- ~1,500 novel ns variants !!
- Overall...more *previously novel* functional variants than expected

‘5 Guys from Africa’ – The African Genome Project

PERSONAL GENOMES:

Jim Watson

Jim Lupski

Desmond Tutu



Eliza Strickland's Exome on Ion Proton

The Gene Machine and Me - IEEE Spectrum

<http://spectrum.ieee.org/biomedical/devices/the-gene-machine-...>

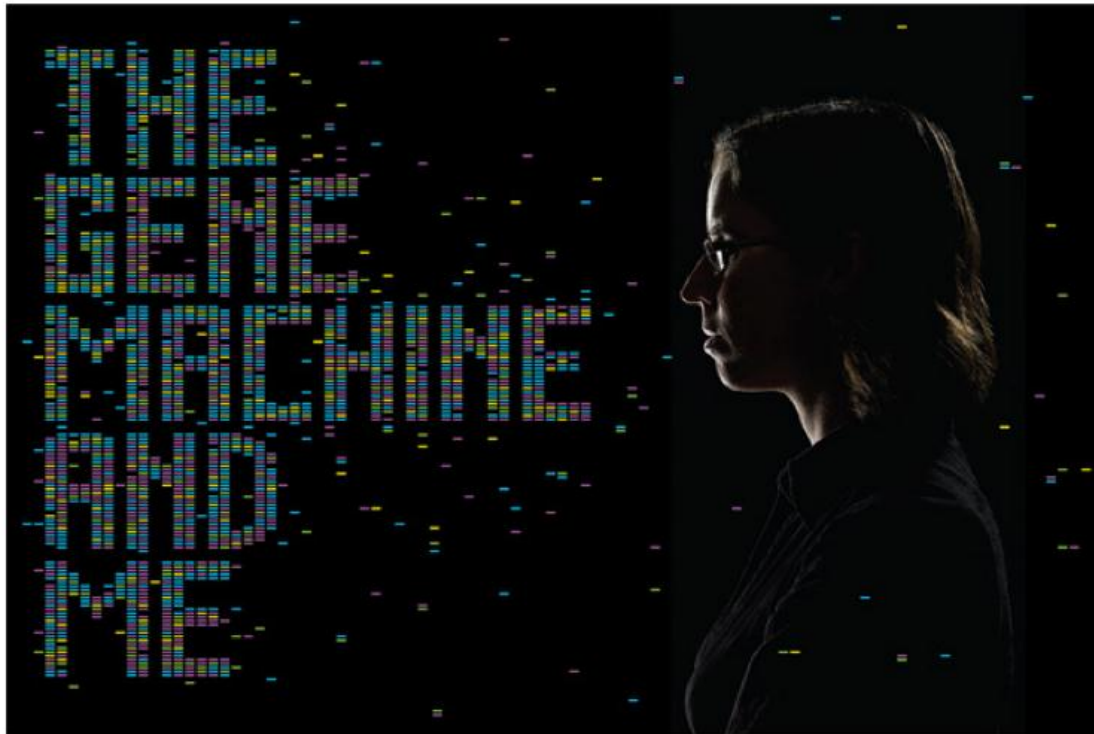
BIOMEDICAL / DEVICES

COVER

The Gene Machine and Me

Ion Torrent's chip-based genome sequencer is cheap, fast, and poised to revolutionize medicine

By ELIZA STRICKLAND / MARCH 2013



IEEE Spectrum 2-28-2013;

<http://spectrum.ieee.org/biomedical/ethics/the-gene-machine-and-me>

The Cancer Genome Atlas (TCGA)



National Cancer Institute

National Human Genome Research Institute



THE CANCER GENOME ATLAS

[Sign up for updates](#)

Search [GO](#)

[About TCGA](#)

[What We Do](#)

[Publications](#)

[News Center](#)

[Launch Data Portal](#)



The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate our understanding of the genetics of cancer using innovative genome analysis technologies.

News

NEW* Fostering Groundbreaking Medical Research: Investments in the National Institutes of Health

The White House reports on how Recovery Act funds are enabling groundbreaking research at NIH, including genomic mapping of 20 cancers through the TCGA project. [Read the Report.](#)

NEW* Francis Collins: One Year at the Helm

Francis Collins, Ph.D., marks his one-year anniversary as the National Institutes of Health (NIH) director. This *Nature* article addresses his accomplishments to date, including his investment in TCGA, and the challenges he faces ahead. [Read the article.](#)

NEW* Genomics Brochure Now Available

Learn more about what cancer genomics means for you. [Read the brochure.](#)

TCGA Identifies Novel Molecular Subtype in Brain Cancer Patients with Distinct

Looking for a Target on Every Tumor

[Science Article](#) | [Podcast](#)

How TCGA data could be translated to patient care.



TCGA Expanding to Study 20 or More Cancers

[Learn More](#) 



TCGA Data Portal

[Access TCGA Data Portal](#) 



Questions about cancer? Visit [Cancer.gov](#) 

1-800-4-CANCER

LiveHelp online chat 

Stay Connected

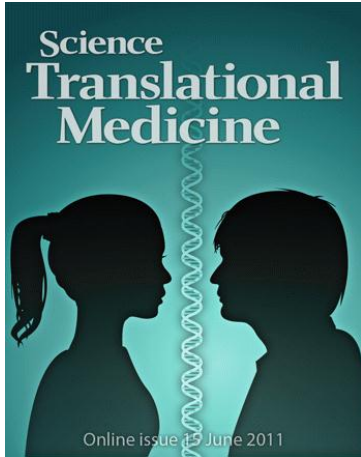
The Cancer Genome Atlas

- 20 cancer types, 500 patients each
- Discover and catalogue all somatic mutations
- DNA sequence, copy number alteration, structural variation, methylation
- Transcriptome (microarray → RNAseq)
- miRNA
- Data dissemination

Cancer Genomes currently undergoing sequencing:

- Glioblastoma
- Lung Adenocarcinoma
- Ovarian
- Colon
- Hepatic
- Oral
- Renal
- Pancreatic
- Bladder

Mendelian Diseases

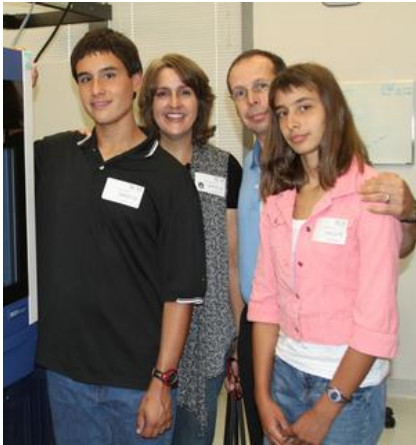


The Beery Family Story



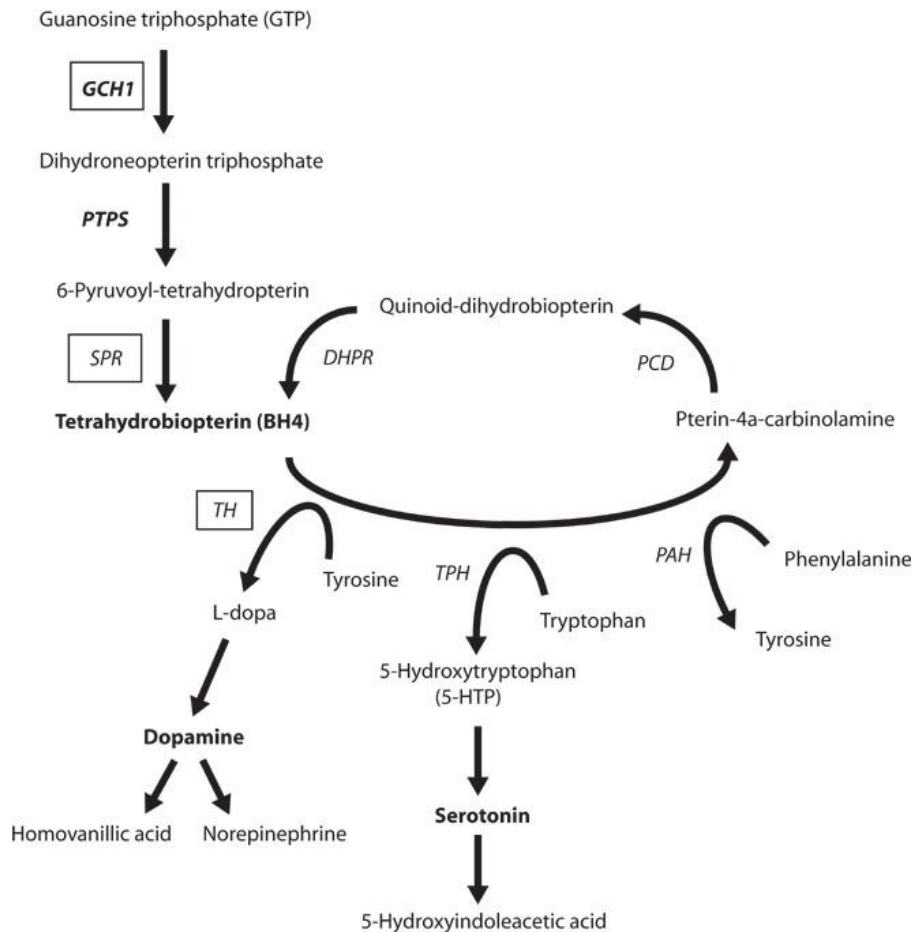
Matthew Bainbridge, et al. Science Translational Medicine

- Initially diagnosed with cerebral palsy
- Actually DRD = Dopa-Responsive Dystonia
- Prescribed L-Dopa; same as Parkinson's



<http://www.cbsnews.com/video/watch/?id=7374693n&tag=contentMain;contentBody>

Pathway



Handwriting pre and Post 5-HTP therapy

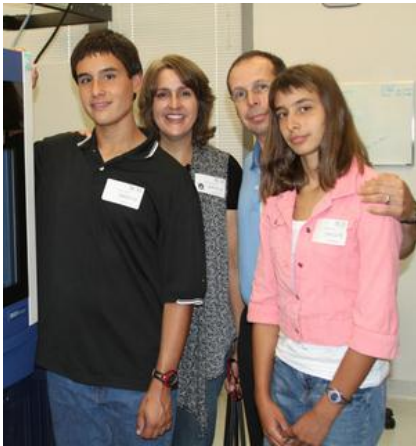
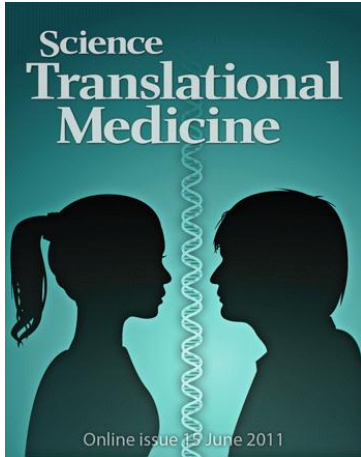
Pre-therapy	Post-therapy
<p>CRISTMAS CHRISTMAS CHRISTMAS</p>	<p>CRISTMAS CHRISTMAS</p>
<p>PARENTS</p>	<p>PARENTS PARENT PARENTS</p>
<p>week</p>	<p>weekend</p>

Mendelian Diseases

The Beery Family Story

Matthew Bainbridge, et al. Science Translational Medicine

- Initially diagnosed with cerebral palsy
- Actually DRD = Dopa-Responsive Dystonia
- Prescribed L-Dopa; same as Parkinson's
- But didn't relieve all symptoms – breathing problems
- Whole Genome Sequencing of both twins reveals mutation in SPR gene; involved in both dopamine and serotonin synthesis
- Supplemented L-Dopa with 5-hydroxytryptophan = Bingo!



<http://www.cbsnews.com/video/watch/?id=7374693n&tag=contentMain;contentBody>

Diagnostic Exome Sequencing



Medical Genetics Laboratories

[>BCM Home](#) [>BCM Centers](#) [>BCM Departments](#) [>Find a BCM person](#) [>Giving](#)

Houston, Texas



Whole Genome Laboratory (WGL)

[Home](#)

[Test Catalog](#)

[About MGL](#)

[Billing](#)

[Licenses](#)

The development and clinical implementation of the [Whole Exome Sequencing](#) test derives from a joint effort by Baylor's Human Genome Sequencing Center and the Medical Genetics Laboratories of the Department of Molecular and Human Genetics to establish a clinical laboratory dedicated to state-of-the-art next generation sequencing. The collaboration between these groups brings together genomic scientists, clinical laboratory scientists, and clinicians to provide reliable genome-wide analyses that are carefully annotated and interpreted for clinical significance by medical geneticists. [Whole Exome Sequencing](#) is the first test to be offered by the WGL and is focused on the evaluation of underlying genetic causes of disease. In the near future, the WGL will implement additional clinical tests, including Whole Genome Sequencing (WGS) that will bring this technology to other aspects of medical care and treatment.



Joint effort between BCM's HGSC and BCM's Medical Genetics Laboratories (MGL) to provide exome sequencing with clinical interpretation



PHARMACOGENOMICS

According to the CDC's Office of Public Health Genomics:

- 82% of the U.S. population takes at least 1 medication
- 29% take 5 or more medications
- 700,000 adverse reaction emergency room visits/year
- 120,000 hospitalizations/year
- Cost = \$3.5 billion

DRUG RESPONSE and the GENOME

COMMENTARY

Genetics and Variable Drug Response

Russell A. Wilke, MD, PhD

M. Eileen Dolan, PhD

ANNUAL HEALTH CARE EXPENDITURES CURRENTLY EXCEED \$2.5 trillion in the United States, a cost burden equivalent to more than \$8000 per person per year.

COMMENTARY

comes related to the use of these drugs can be strongly influenced by genetic variability in the cytochromes p450 (CYPs). For example, the biologically active form of warfarin is metabolized primarily by CYP2C9, and common variants in this enzyme alter warfarin dosing requirements. Further variance in warfarin dose can be explained by inheritable changes

nature publishing group

STATE OF THE ART

Facilitating Clinical Implementation of Pharmacogenomics

David A. Mrazek, MD, FRCPsych

Caryn Lerman, PhD

VARIABILITY IN DRUG RESPONSE CAN BE EXPLAINED, in part, by genetic differences among patients. A clear role in drug toxicity and efficacy has been

Evaluating Effectiveness

At the center of a debate on the clinical implementation of pharmacogenomics is the threshold of evidence required in practice. Consistent with the Clinical Pharmacogenomics Implementation Consortium of the Pharmacogenomics Research Network,⁴ the evidence threshold for implementation can be met by the existence of a strong biological r

The Emerging Role of Electronic Medical Records in Pharmacogenomics

RA Wilke¹, H Xu², JC Denny², DM Roden¹, RM Krauss³, CA McCarty⁴, RL Davis⁵, T Skaar⁶, J Lamba⁷ and G Savova^{8,9,10}

Health-care information technology and genotyping technology are both advancing rapidly, creating new opportunities for medical and scientific discovery. The convergence of these two technologies is now facilitating genetic association studies of unprecedented size within the context of routine clinical care. As a result, the medical community will soon be presented with a number of novel opportunities to bring functional genomics to the bedside in the area of pharmacotherapy. By linking biological material to comprehensive medical records, large multi-institutional biobanks are now poised to advance the field of pharmacogenomics through three distinct mechanisms: (i) retrospective assessment of previously known findings in a clinical practice-based setting, (ii) discovery of new associations in huge observational cohorts, and (iii) prospective application in a setting capable of providing real-time decision support. This review explores each of these translational mechanisms within a historical framework.



Pharmacogenomics
Research Network

Search

Home PGRN Network Resources Community News Graphics Publications & Reports

The eMERGE Network
electronic Medical Records & Genomics

A consortium of biorepositories linked to electronic medical records data for conducting genomic studies



For the PGRN

PGRN Calendar

Calls, Meetings,
Seminars, Workshops

Research Groups
&
Network Resources

Working Groups
&
Committees



Click [HERE](#) to view a larger PGRN map and list of Research Groups & Network Resources

Reports
from the PGRN

Publications

News, Media
& Blogs

For the Community

About the PGRN

Mission Statement,
Governance, Advisory Group

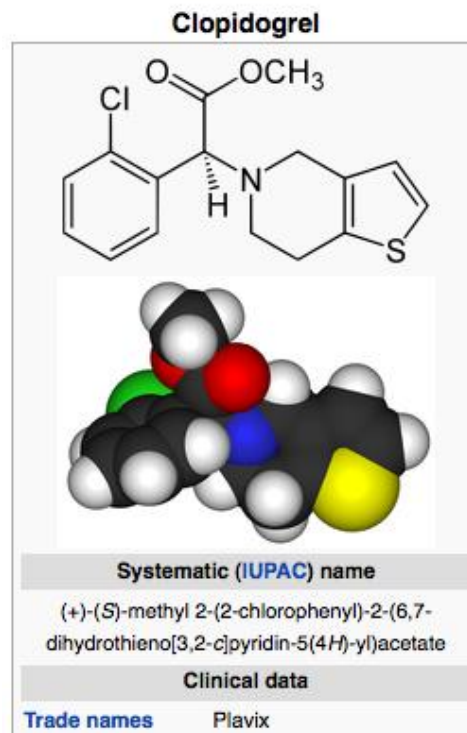
Consortia at
PharmGKB

Affiliate
Membership

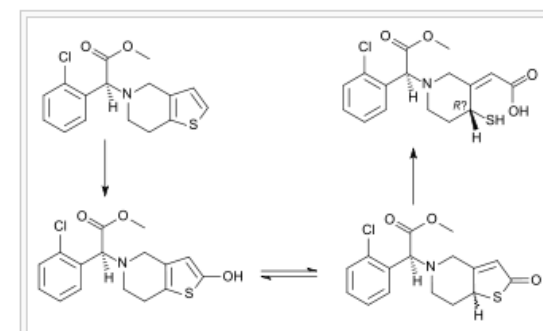
AN EXAMPLE . . .

CLOPIDOGREL

- antiplatelet agent used to inhibit blood clots in coronary artery disease, peripheral vascular disease, and cerebrovascular disease.
- 2nd most widely prescribed drug (2007)
- U.S. sales = \$3.8 billion (2008); worldwide = \$6.6 billion (2009)
- Prodrug requiring activation



Pharmacokinetics and metabolism



Clopidogrel (top left) being activated. The first step is an oxidation mediated (mainly) by **CYP2C19**, unlike the activation of the related drug **prasugrel**. The two structures at the bottom are **tautomers** of each other; and the final step is a hydrolysis. The active metabolite (top right) has **Z configuration** at the double bond C3–C16 and possibly **R configuration** at the newly asymmetric C4.^[10]

The Clinical Pharmacogenetics Implementation Consortium (CPIC) has published genotype-based dosing guidelines for CYP2D6 and codeine in Clinical Pharmacology and Therapeutics (CPT). Download the [article](#) and [supplement](#).

From Knowledge Acquisition to Clinical Applications

[our mission](#) ▶

Find Data By Type

Genomic Variations

VKORC1, G3673A
 Causative allele for the low dose phenotype
 Related drug: Warfarin
 rs9923231

- [Annotated SNPs by gene](#)
- [Annotated SNPs by drug](#)
- [Annotated SNPs by disease](#)
- [Genes with Haplotype Translations](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Clinical Interpretations



Azathioprine dosing

- [Clinical variant annotations](#)
- [Genotype-based dosing guidelines](#)
- [Drug labels](#)
- [Genetic tests for PGx](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Pathways



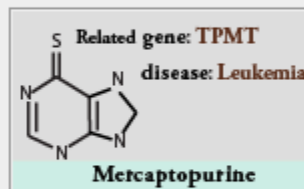
- [Pharmacokinetic pathways](#)
- [Pharmacodynamic pathways](#)
- [All pathways](#)
- [Pathways by therapeutic categories](#)



[examples](#)

hint: enter a gene, drug, disease

Drugs & Small Molecules



Related gene: TPMT
 disease: Leukemia
 Mercaptopurine

- [Drugs with genetic information](#)
- [Drugs with data](#)
- [Drugs by therapeutic categories](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Genes

■ Exon ■ Synonymous ■ UTR

- [Important PGx genes VIP](#)

Diseases

- [Diseases with genetic](#)



Tutorials

- [PharmGKB Overview](#)
- [Clinical PGx](#)
- [PGx Research](#)

Curators' Favorite Papers

- [Adoption of Pharmacogenomic Testing by US Physicians: Results of a Nationwide Survey](#)
- [Pharmacogenetic investigation of response to duloxetine treatment in generalized anxiety disorder](#)

- [The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement](#)

Updated 1/30/12.
 See the [archives](#) for more.



Pharmacogenomics
 Research Network
 PGRN

PGx in the News

The Clinical Pharmacogenetics Implementation Consortium (CPIC) has published genotype-based dosing guidelines for CYP2D6 and codeine in Clinical Pharmacology and Therapeutics (CPT). Download the [article](#) and [supplement](#).

From Knowledge Acquisition to Clinical Applications

[our mission](#) ▶

Find Data By Type

Genomic Variations

VKORC1, G3673A
Causative allele for the low dose phenotype
Related drug: Warfarin
rs9923231

- [Annotated SNPs by gene](#)
- [Annotated SNPs by drug](#)
- [Annotated SNPs by disease](#)
- [Genes with Haplotype Translations](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Clinical Interpretations



Azathioprine dosing

- [Clinical variant annotations](#)
- [Genotype-based dosing guidelines](#)
- [Drug labels](#)
- [Genetic tests for PGx](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Pathways



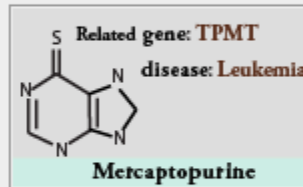
- [Pharmacokinetic pathways](#)
- [Pharmacodynamic pathways](#)
- [All pathways](#)
- [Pathways by therapeutic categories](#)



[examples](#)

hint: enter a gene, drug, disease

Drugs & Small Molecules



Related gene: TPMT
disease: Leukemia
Mercaptopurine

- [Drugs with genetic information](#)
- [Drugs with data](#)
- [Drugs by therapeutic categories](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Genes

■ Exon ■ Synonymous ■ UTR

- [Important PGx genes VIP](#)

Diseases

- [Diseases with genetic](#)



Tutorials

- [PharmGKB Overview](#)
- [Clinical PGx](#)
- [PGx Research](#)

Curators' Favorite Papers

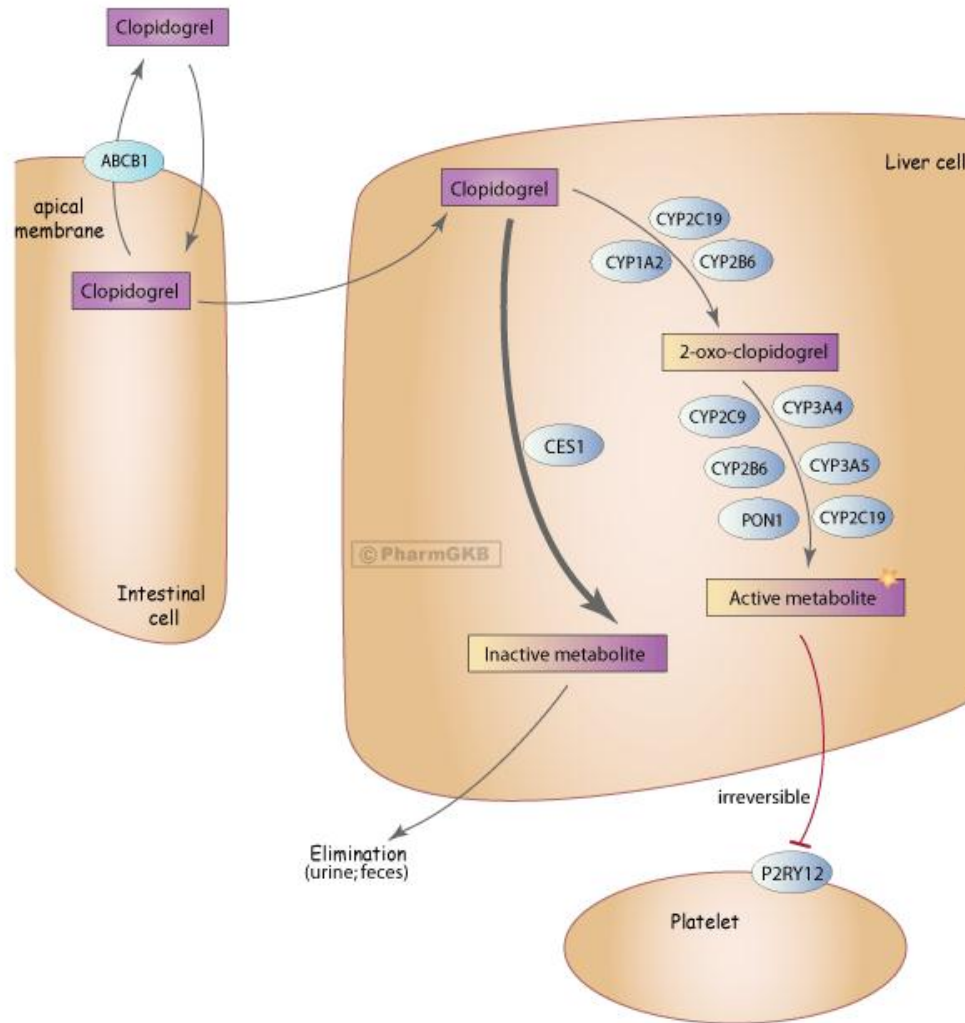
- [Adoption of Pharmacogenomic Testing by US Physicians: Results of a Nationwide Survey](#)
 - [Pharmacogenetic investigation of response to duloxetine treatment in generalized anxiety disorder](#)
 - [The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement](#)
- Updated 1/30/12.
See the [archives](#) for more.



Pharmacogenomics Research Network
PGRN

PGx in the News

Clopidogrel metabolism.



Approximately 14% of the population is $*2/*2$

The Clinical Pharmacogenetics Implementation Consortium (CPIC) has published genotype-based dosing guidelines for CYP2D6 and codeine in Clinical Pharmacology and Therapeutics (CPT). Download the [article](#) and [supplement](#).

From Knowledge Acquisition to Clinical Applications

our mission ▶

Find Data By Type

Genomic Variations

VKORC1, G3673A
Causative allele for the low dose phenotype
Related drug: Warfarin
rs9923231

- [Annotated SNPs by gene](#)
- [Annotated SNPs by drug](#)
- [Annotated SNPs by disease](#)
- [Genes with Haplotype Translations](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Clinical Interpretations



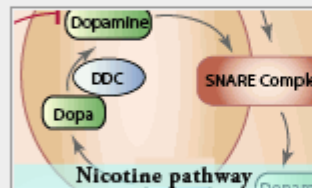
- [Clinical variant annotations](#)
- [Genotype-based dosing guidelines](#)
- [Drug labels](#)
- [Genetic tests for PGx](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Pathways



- [Pharmacokinetic pathways](#)
- [Pharmacodynamic pathways](#)
- [All pathways](#)
- [Pathways by therapeutic categories](#)



[examples](#)

hint: enter a gene, drug, disease

Drugs & Small Molecules

Related gene: TPMT
disease: Leukemia
Mercaptopurine

- [Drugs with genetic information](#)
- [Drugs with data](#)
- [Drugs by therapeutic categories](#)



[examples](#)

hint: enter a gene, rsid, drug, disease

Genes

■ Exon ■ Synonymous ■ UTR

- [Important PGx genes VIP](#)

Diseases

- [Diseases with genetic](#)



Tutorials

- [PharmGKB Overview](#)
- [Clinical PGx](#)
- [PGx Research](#)

Curators' Favorite Publications

- [Adoption of Pharmacogenomic Testing by US Physicians: Results of a Nationwide Survey](#)
- [Pharmacogenetic investigation of respiratory duloxetine treatment in generalized anxiety disorder](#)
- [The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement](#)

Updated 1/30/12. See the [archives](#) for more.



Pharmacogenomics Research Network
PGRN

PGx in the News

Dosing Guidelines

These dosing guidelines take into consideration patient genotype and have been published by the [Clinical Pharmacogenetics Implementation Consortium \(CPIC\)](#) or the Royal Dutch Association for the Advancement of Pharmacy - Pharmacogenetics Working Group (DPWG) (manually curated by PharmGKB).

Title	Drug - Gene Pair
Dosing Guidelines for abacavir	DPWG abacavir HLA-B
Dosing Guidelines for acenocoumarol	DPWG acenocoumarol CYP2C9 DPWG acenocoumarol VKORC1
Dosing Guidelines for amitriptyline	DPWG amitriptyline CYP2D6
Dosing Guidelines for aripiprazole	DPWG aripiprazole CYP2D6
Dosing Guidelines for atomoxetine	DPWG atomoxetine CYP2D6
Dosing Guidelines for azathioprine	CPIC azathioprine TPMT DPWG azathioprine TPMT
Dosing Guidelines for capecitabine	DPWG capecitabine DPYD
Dosing Guidelines for carvedilol	DPWG carvedilol CYP2D6
Dosing Guidelines for citalopram	DPWG citalopram CYP2C19
Dosing Guidelines for clomipramine	DPWG clomipramine CYP2D6
Dosing Guidelines for clopidogrel	CPIC clopidogrel CYP2C19 DPWG clopidogrel CYP2C19
Dosing Guidelines for clozapine	DPWG clozapine CYP2D6
Dosing Guidelines for codeine	CPIC codeine CYP2D6 DPWG codeine CYP2D6

Dosing Guidelines

These dosing guidelines take into consideration patient genotype and have been published by the [Clinical Pharmacogenetics Implementation Consortium \(CPIC\)](#) or the [Royal Dutch Association for the Advancement of Pharmacy - Pharmacogenetics Working Group \(DPWG\)](#) (manually curated by PharmGKB).

Title	Drug - Gene Pair
Dosing Guidelines for abacavir	DPWG abacavir HLA-B
Dosing Guidelines for acenocoumarol	DPWG acenocoumarol CYP2C9 DPWG acenocoumarol VKORC1
Dosing Guidelines for amitriptyline	DPWG amitriptyline CYP2D6
Dosing Guidelines for aripiprazole	DPWG aripiprazole CYP2D6
Dosing Guidelines for atomoxetine	DPWG atomoxetine CYP2D6
Dosing Guidelines for azathioprine	CPIC azathioprine TPMT DPWG azathioprine TPMT
Dosing Guidelines for capecitabine	DPWG capecitabine DPYD
Dosing Guidelines for carvedilol	DPWG carvedilol CYP2D6
Dosing Guidelines for citalopram	DPWG citalopram CYP2C19
Dosing Guidelines for clomipramine	DPWG clomipramine CYP2D6
Dosing Guidelines for clopidogrel	CPIC clopidogrel CYP2C19 DPWG clopidogrel CYP2C19
Dosing Guidelines for clozapine	DPWG clozapine CYP2D6
Dosing Guidelines for codeine	CPIC codeine CYP2D6 DPWG codeine CYP2D6

CPIC Dosing Guideline - [clopidogrel](#), [CYP2C19](#)

Guidelines regarding the use of pharmacogenomic tests in dosing for clopidogrel have been published in Clinical Pharmacology and Therapeutics by the [Clinical Pharmacogenetics Implementation Consortium \(CPIC\)](#).

Download: [article](#) and [supplement](#)

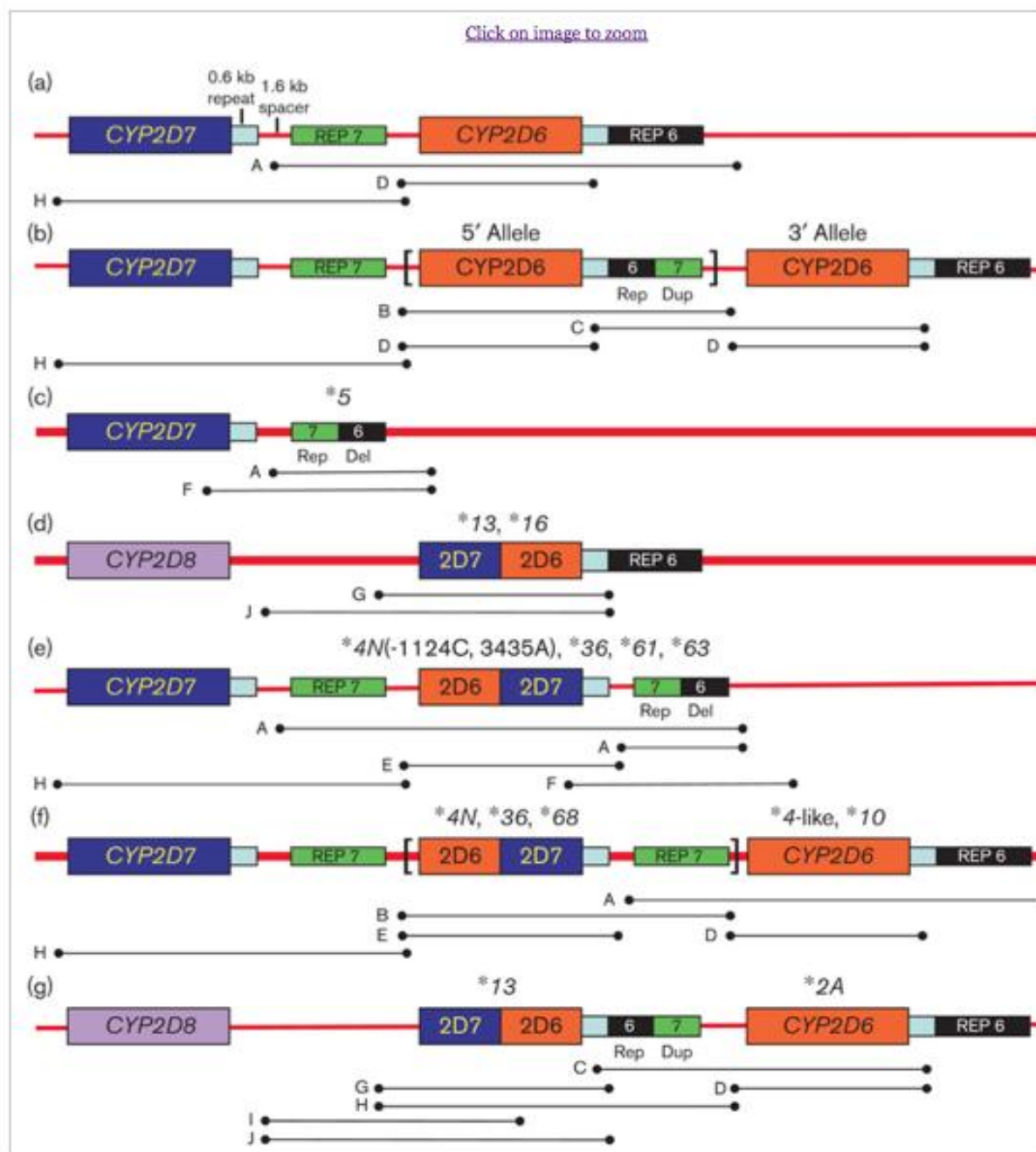
Excerpt from the clopidogrel dosing guidelines:

The table below summarizes the therapeutic CPIC guidelines for clopidogrel based on CYP2C19 phenotype for patients with acute coronary syndrome (ACS) and percutaneous coronary intervention (PCI) initiating antiplatelet therapy. These guidelines have been limited to the *CYP2C19**2 allele ([rs4244285](#)). At the time of writing these guidelines, only the *CYP2C19**2 allele has been adequately studied with respect to clinical outcomes on clopidogrel; other variants are too rare, have not been studied, or have resulted in inconclusive findings. In addition to the *CYP2C19**2 allele, many clinical genotyping platforms include other variant alleles (*3-*8, *17) that may alter the interpretation of a patient's predicted metabolizer phenotype. For some rare genotype combinations (e.g.*2/*17) metabolic phenotypes are difficult to predict.

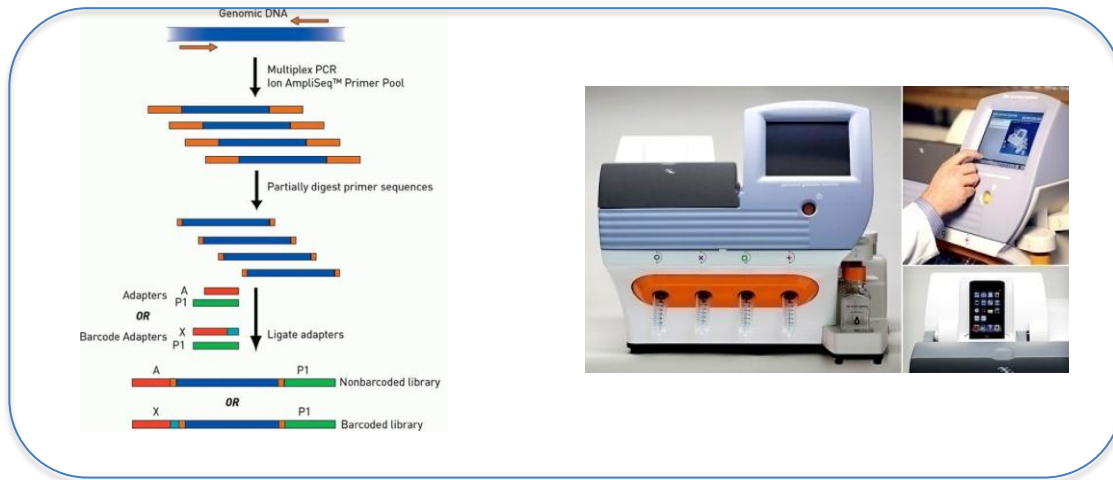
Clopidogrel therapy based on CYP2C19 phenotype for ACS/PCI patients initiating antiplatelet therapy:

Phenotype (Genotype)	Implications for clopidogrel	Therapeutic recommendations	Classification of recommendations
Ultrarapid metabolizer (UM) (*1/*17, *17/*17) and extensive metabolizer (EM) (*1/*1)	Normal (EM) or increased (UM) platelet inhibition; normal (EM) or decreased (UM) residual platelet aggregation ¹	Clopidogrel label-recommended dosage and administration	Strong
Intermediate metabolizer (IM) (*1/*2)	Reduced platelet inhibition; increased residual platelet aggregation; increased risk for adverse cardiovascular events	Prasugrel or other alternative therapy (if no contraindication)	Moderate
Poor metabolizer (PM) (*2/*2)	Significantly reduced platelet inhibition; increased residual platelet aggregation; increased risk for adverse cardiovascular events	Prasugrel or other alternative therapy (if no contraindication)	Strong

¹ The *CYP2C19**17 allele ([rs12248560](#)) may be associated with increased risk of bleeding (see [article](#) for reference).



Tough Pharmaco Region Mixed Platform

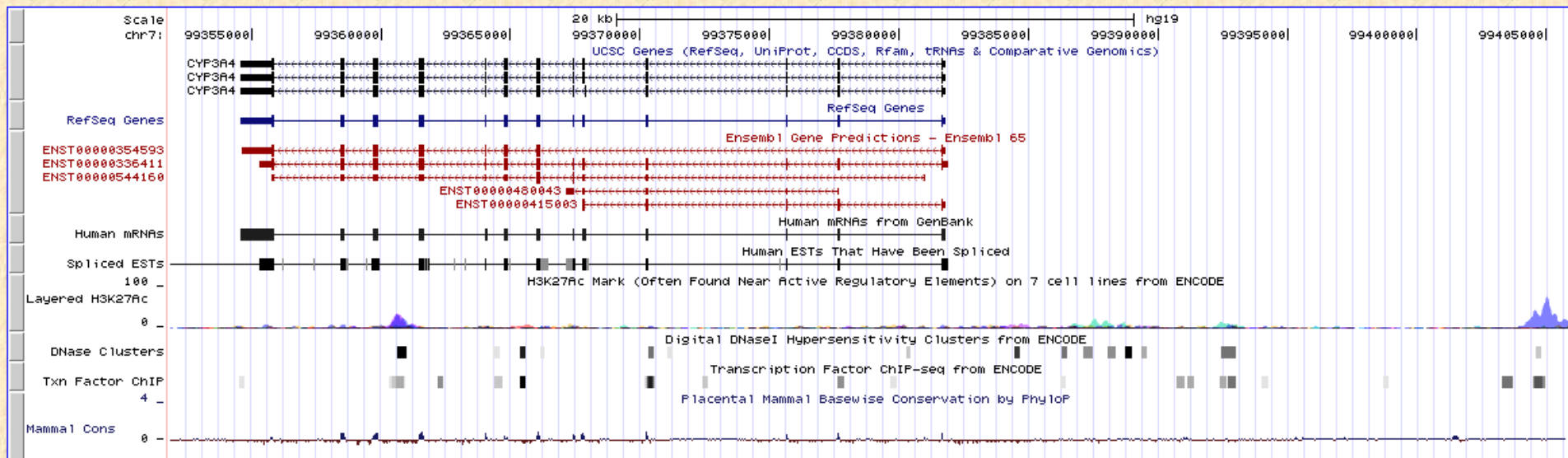


CHALLENGES

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr7:99,351,858-99,406,315 gene jump clear size 54,458 bp. configure



- Non-coding variants
- Pseudogenes
- Expression analysis
- Epigenetics

Acknowledgements

Richard Gibbs – Director

Eric Boerwinkle – Associate Director

Donna Muzny – Production Director

Jeff Reid – Sequence Analysis

Matthew Bainbridge – Mendelian Disease & Analysis

David Wheeler – Cancer Genomics

Harsha Doddapaneni – Library

Min Wang – Capture R&D

Michelle Rives – Administration

Michael Metzker – Production Research

Debra Murray – Minority Diversity Program

Xin Zhang – Data Analysis

Stephen Richards – Arthropod Genomes

Jeffrey Rogers – Primate Genomes

Kim Worley - Genome Assembly & Comparative
Annotation

Fuli Yu – 1000 Genomes and HapMap

