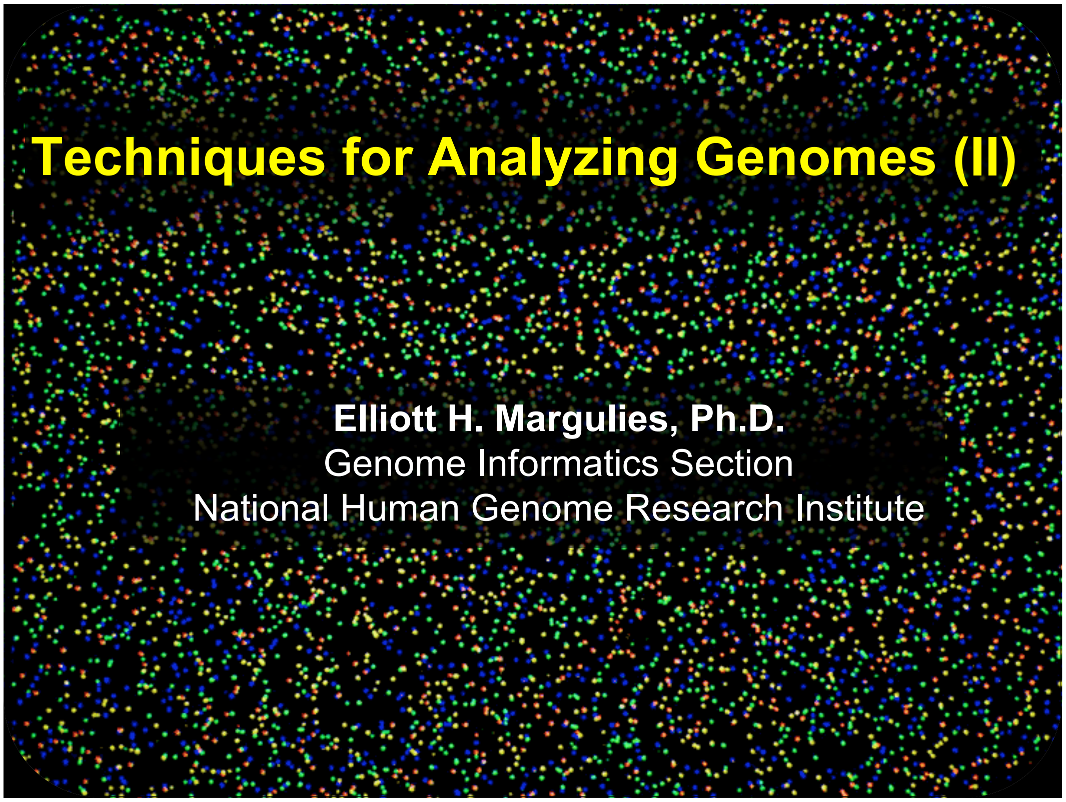
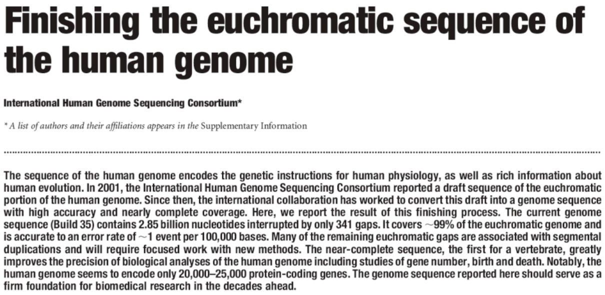


Techniques for Analyzing Genomes (II)

Elliott H. Margulies, Ph.D.
Genome Informatics Section
National Human Genome Research Institute

Sequencing Complete



Finishing the euchromatic sequence of the human genome

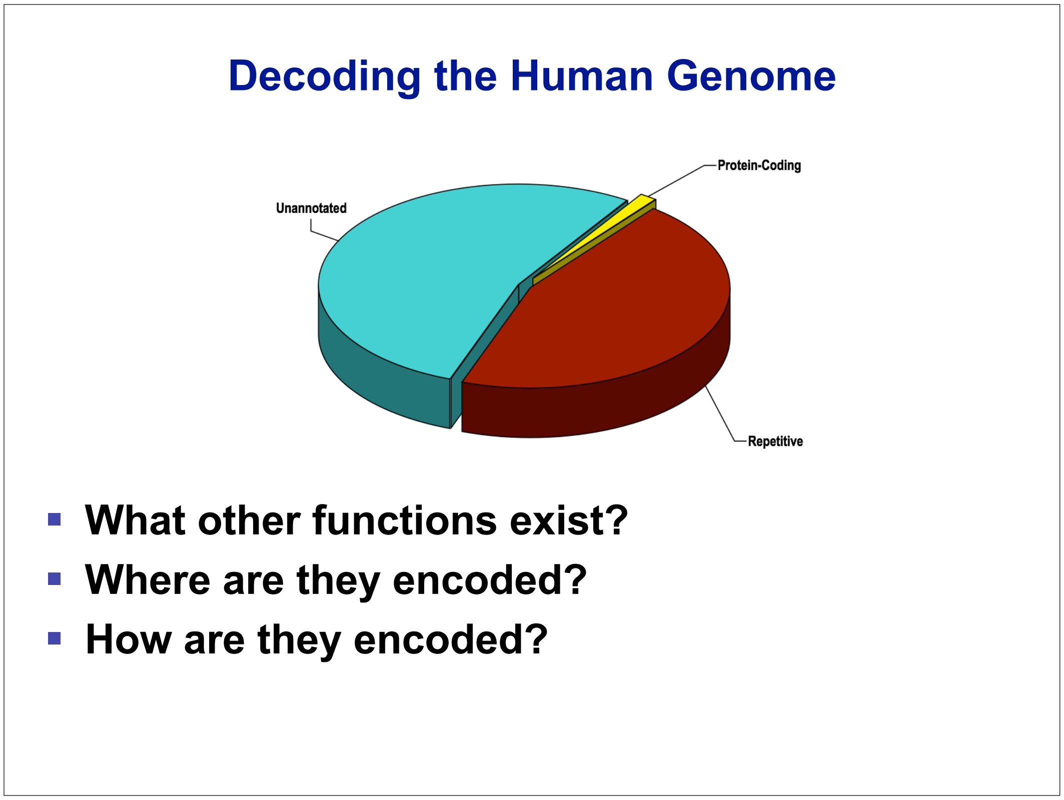
International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

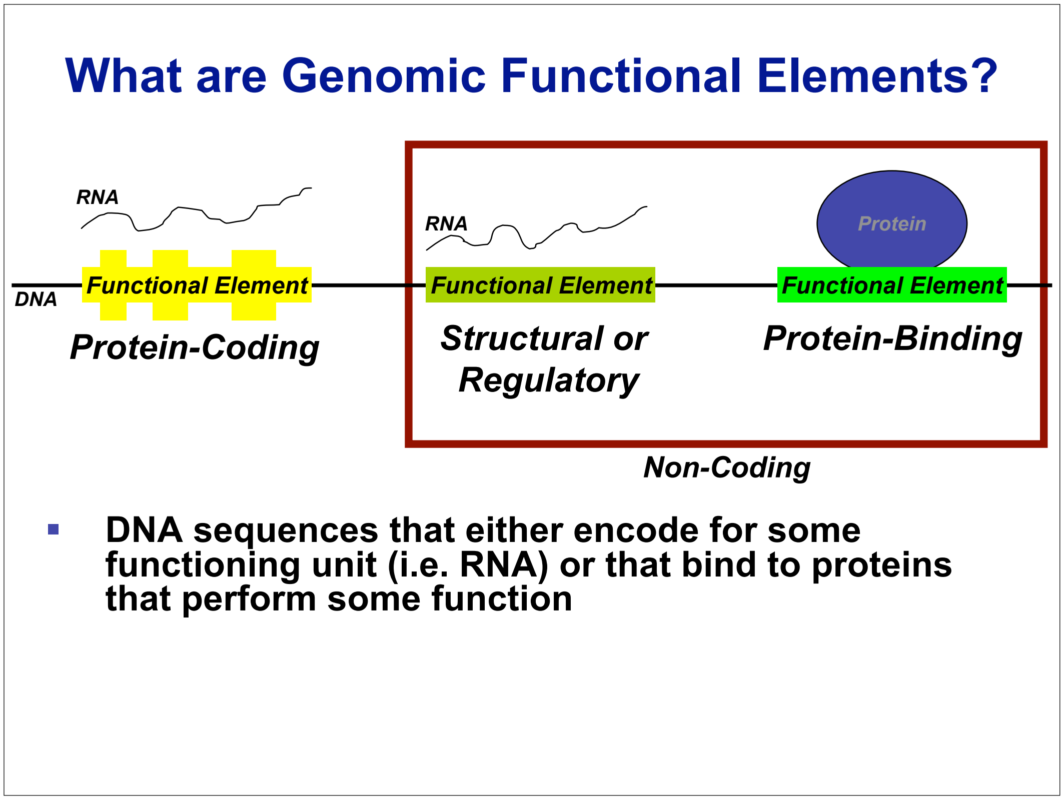
International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.



Decoding the Human Genome

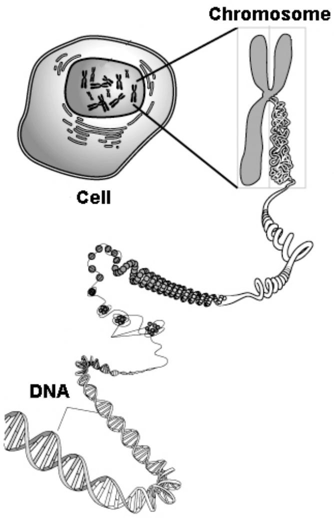
- What other functions exist?
- Where are they encoded?
- How are they encoded?

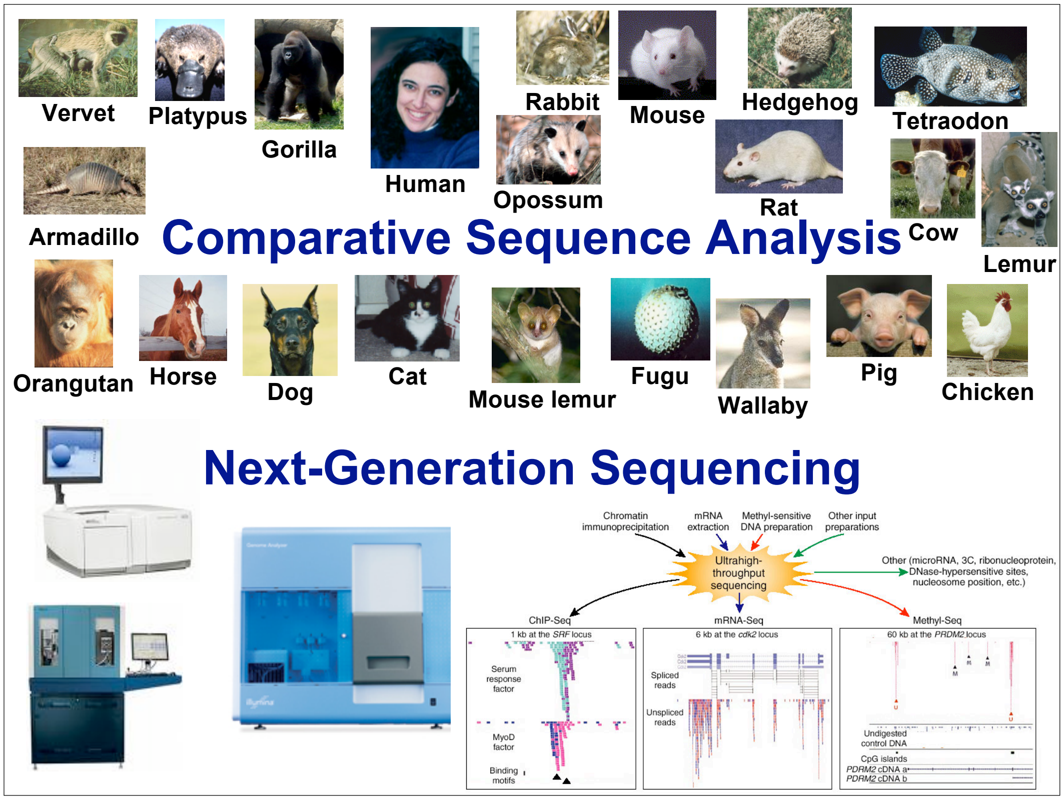


What are Genomic Functional Elements?

- DNA sequences that either encode for some functioning unit (i.e. RNA) or that bind to proteins that perform some function

How can we analyze genomes to find functional elements?




Vervet
Platypus
Gorilla
Human
Rabbit
Mouse
Hedgehog
Tetraodon
Opossum
Rat
Armadillo
Cow
Lemur
Orangutan
Horse
Dog
Cat
Mouse lemur
Fugu
Wallaby
Pig
Chicken

Comparative Sequence Analysis

Next-Generation Sequencing

Chromatin immunoprecipitation
mRNA extraction
Methyl-sensitive DNA preparation
Other input preparations
Ultrahigh-throughput sequencing
Other (microRNA, 3C, ribonucleoprotein, DNase-hypersensitive sites, nucleosome position, etc.)
ChIP-Seq
1 kb at the *SRF* locus
Serum response factor
MyoD factor
Binding motifs
mRNA-Seq
6 kb at the *cdk2* locus
Spliced reads
Unspliced reads
Methyl-Seq
60 kb at the *PRDM2* locus
M
U
Undigested control DNA
CpG islands
PDRM2 cDNA a
PDRM2 cDNA b

Comparative Sequence Analysis

OUTLINE

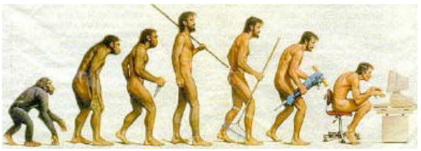
- **How comparative genomics “works”**
- **Steps involved involved**
  - Sequence Generation
  - Homologous Co-linearity prediction (synteny)
  - Base-pair alignment
  - Identification of constrained sequences
- **Lessons learned from comparative analyses**

Comparative Genomics to Decode the Genome

```
TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGCANAYOUGFINDAMEGGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCCACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTTGGGGCCGGACCAGGCATHISAISAIMPORTANTGSTUFFGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGGTGGGTGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGT
TGTCTCCAAACTTTTTTTCAGGTGAGAAGGTGGCCAACCGAGCTTCSUPERCALAFRAGALISTICEXPEALADOTIOUSAGTATGGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTTTAAAAAGATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTTGTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGG
ACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATT
TTGGGGTAGGTAGAAAATATATGOTBLUEGTATTTATTGTTATGAGACTGGATATATCTAGTATTTGTCACAGGTAAATGATTCTTCAAAAATTG
AAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGGAATAGATACCGAAGTTATATCCAA
CTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAATAGCATCTAAGTTCGG
AATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGT
TATAAATAGCTCATATITMADEGTHISTSLIDEAONGMYABIRTHDAYGSEPTEMBERGTWENTYEIGHTHAGCATGTGCAGTTAATCCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGT
TGATGTTATCCACCTTTTGTTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCAT
AAATTATAGACTTTTTTTAGCAGAGAACTTTGAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATA
GTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTGATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTT
AATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGEEKSAMAKEAGOODCHUSBANDSTGTTCTCC
AAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTACTATTTATAGGAACTGACAATCACCTAAAATACCAATGA
TTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATTAAGTTAGGCAGTATTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTTAACATCTCTGTG
TGTGACTTGACCTTTAAAAONETDAYALEFTIESTWILLARULEATHEGWORLDTATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATG
TTTCTATTGTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT
```

>target1:1308901-1311845

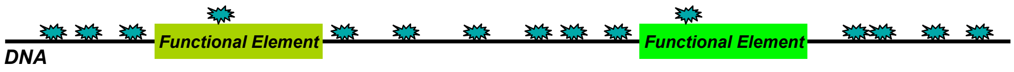
Rationale Behind Comparative Genomics



- **DNA represents a "blueprint" for the structure and physiology of all living things**
- **Mutations occur randomly throughout the genome**
  - Neutral theory of evolution (M. Kimura, 1983)
- **Mutations in *functional* DNA are less likely to be tolerated**

Kimura M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge [Cambridgeshire]; New York.

Fewer Mutations are Found in Functional DNA

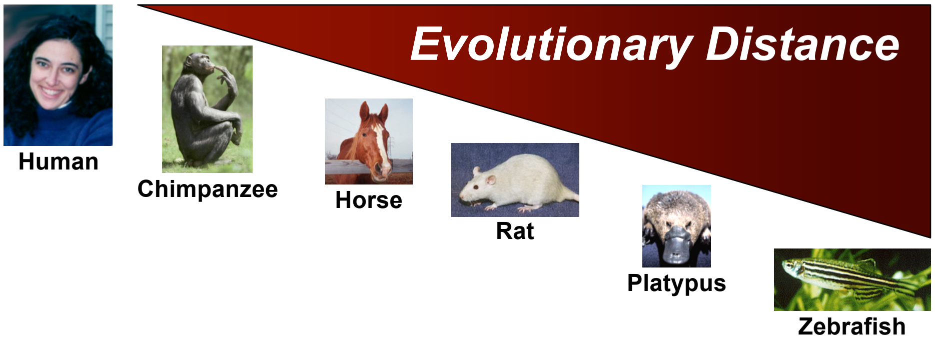


- **Functional sequences will be "more similar" when compared between different species**

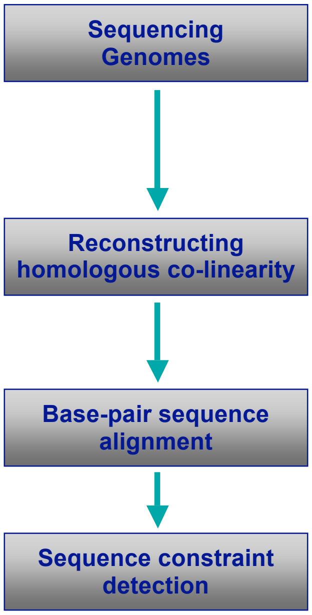
Comparative Sequence Analysis Provides an Unbiased Approach for Detecting Non-Coding Functional Elements

Comparative Genomics

- **Find sequences that have diverged less than we expect**
  These sequences are likely to have a functional role
- **Our expectation is related to the time since the last common ancestor**



Four Major Components of Comparative Genomics

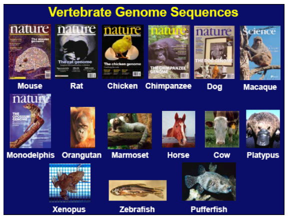


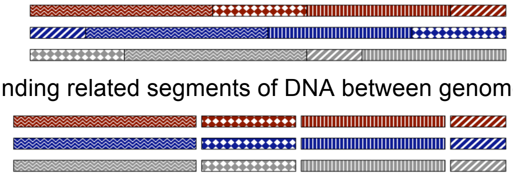
Sequencing Genomes

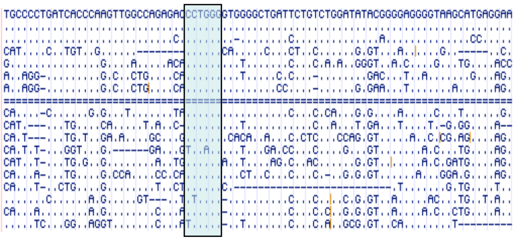
Reconstructing homologous co-linearity

Base-pair sequence alignment

Sequence constraint detection

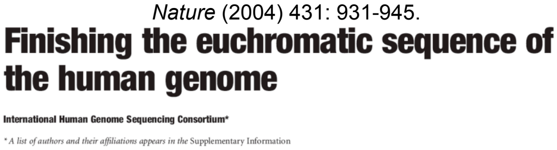






Sequencing Genomes

Nature (2004) 431: 931-945.



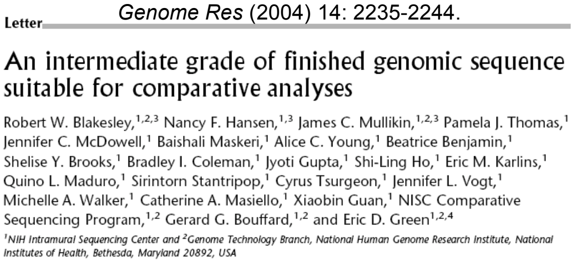
Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

"Finnished"
Essentially Complete
High Contiguity

Genome Res (2004) 14: 2235-2244.



Letter

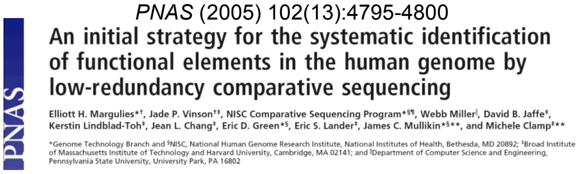
An intermediate grade of finished genomic sequence suitable for comparative analyses

Robert W. Blakesley,[1,2,3] Nancy F. Hansen,[1,3] James C. Mullikin,[1,2,3] Pamela J. Thomas,[1] Jennifer C. McDowell,[1] Baishali Maskeri,[1] Alice C. Young,[1] Beatrice Benjamin,[1] Shelise Y. Brooks,[1] Bradley I. Coleman,[1] Jyoti Gupta,[1] Shi-Ling Ho,[1] Eric M. Karlins,[1] Quino L. Maduro,[1] Sirintorn Stantripop,[1] Cyrus Tsurgeon,[1] Jennifer L. Vogt,[1] Michelle A. Walker,[1] Catherine A. Masiello,[1] Xiaobin Guan,[1] NISC Comparative Sequencing Program,[1,2] Gerard G. Bouffard,[1,2] and Eric D. Green[1,2,4]

[1]NIH Intramural Sequencing Center and [2]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

"Comparative Grade" or "Draft"
Majority of Genome Represented
Contiguity varied

PNAS (2005) 102(13):4795-4800



An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing

Elliott H. Margulies*†, Jade P. Vinson†‡, NISC Comparative Sequencing Program*§¶, Webb Miller‖, David B. Jaffe‡, Kerstin Lindblad-Toh‡, Jean L. Chang‡, Eric D. Green*§, Eric S. Lander‡, James C. Mullikin*§**, and Michele Clamp‡**

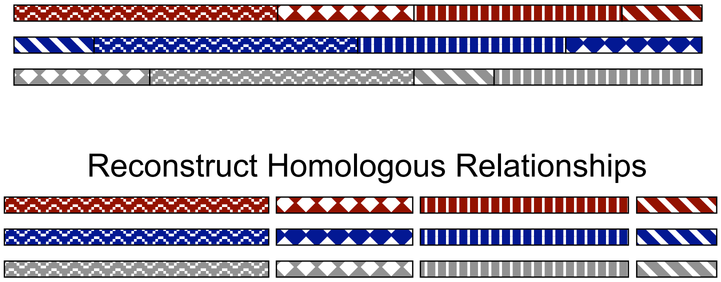
*Genome Technology Branch and §NISC, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; ‡Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02141; and ‖Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802

"Low Redundancy"
60-80% of Genome Represented
Contiguity low

Reconstructing Homologous Co-linearity (Synteny Mapping)

- **Chromosomes do not evolve as single co-linear segments**

Sequenced Genomes



Reconstruct Homologous Relationships

Approaches to Reconstructing Homologous Co-linearity among Related Genomes

- **“Chains and Nets”**
  - Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100**, 11484-11489 (2003).
- **GRIMM**
  - Tesler, G. GRIMM: genome rearrangements web server. Bioinformatics 18, 492-3 (2002).
- **Mercator**
  - Dewey, C.N. Aligning Multiple Whole Genomes with Mercator and MAVID. Methods Mol Biol 395, 221-36 (2007).
- **Infinite Sites**
  - D. Haussler group, UC Santa Cruz
- **Ortheus**
  - E. Birney group, EBI, Hinxton UK

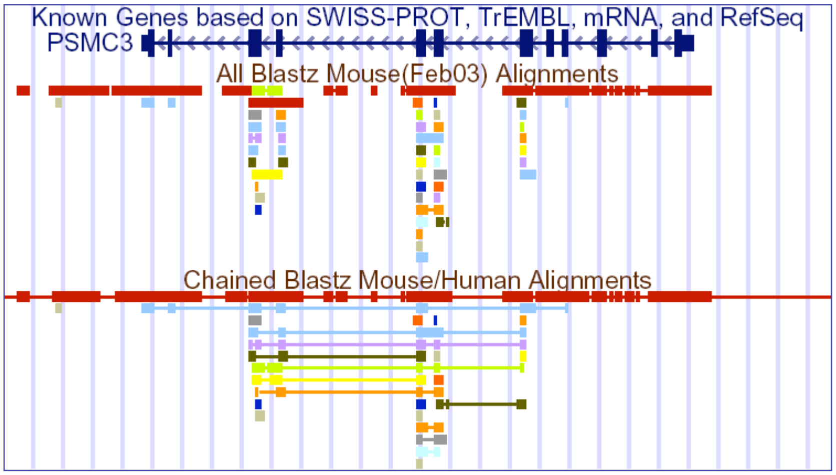
“Chains and Nets” – The UCSC Way

Chaining Alignments

- **Chaining bridges the gulf between large syntenic blocks and base-by-base alignments**
- **The Challenge:**
  - Local alignments tend to break at transposon insertions, inversions, duplications, etc.
  - Global alignments tend to force non-homologous bases to align.
- **The Solution:**
  - Chaining is a rigorous way of joining together local alignments into larger structures.

Slide (though modified) Courtesy of Jim Kent

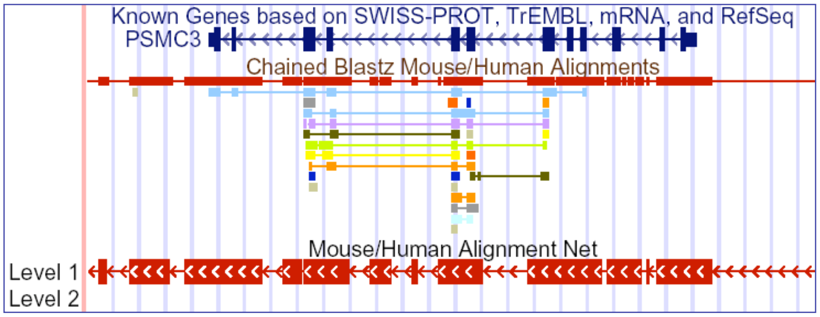
Chains join together related local alignments


Known Genes based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
PSMC3
All Blastz Mouse(Feb03) Alignments
Chained Blastz Mouse/Human Alignments

Protease Regulatory Subunit 3

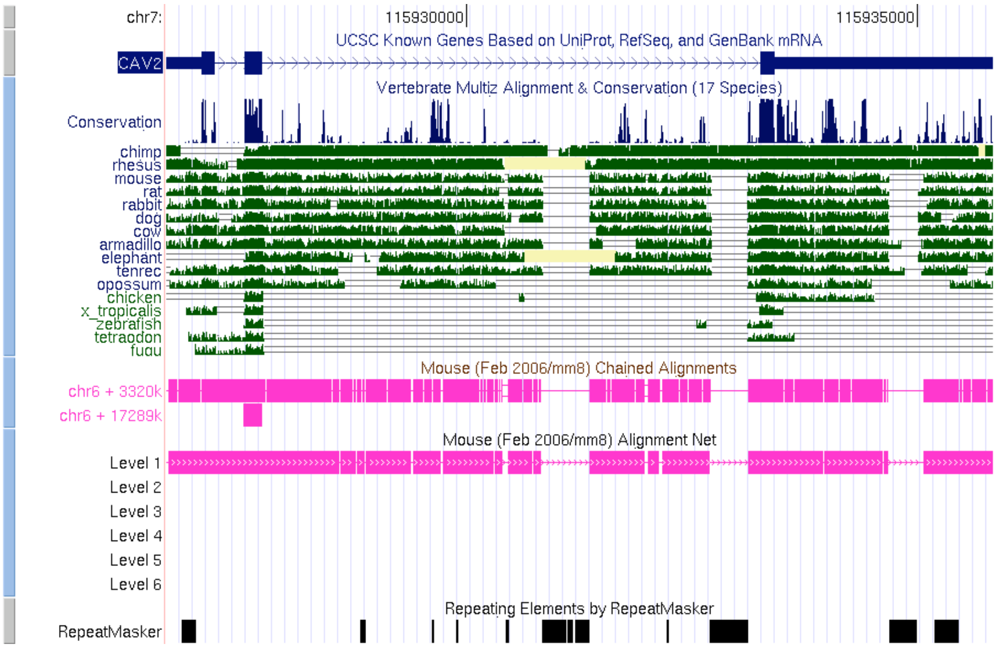
Slide Courtesy of Jim Kent

Net Alignments: Focus on Orthology


Known Genes based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
PSMC3
Chained Blastz Mouse/Human Alignments
Mouse/Human Alignment Net
Level 1
Level 2

- **Frequently, there are numerous mouse alignments for any given human region, particularly for coding regions.**
- **Net finds best mouse match for each human region.**

Slide (though modified) Courtesy of Jim Kent


chr7:
115930000
115935000
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
CAV2
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
chimp
rhesus
mouse
rat
rabbit
dog
cow
armadillo
elephant
tenrec
opossum
chicken
x_tropicalis
zebrafish
tetraodon
fugu
Mouse (Feb 2006/mm8) Chained Alignments
chr6 + 3320k
chr6 + 17289k
Mouse (Feb 2006/mm8) Alignment Net
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Repeating Elements by RepeatMasker
RepeatMasker

Click here for a more complicated example

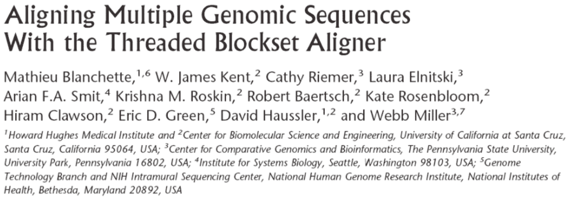
Genome-wide Multi-sequence Alignments

This is not a "solved problem"

Significant challenges:

- **Finding the correct sequences to align**
- **Not all sequences should align**
- **Dealing with insertions/deletions**
- **Handling duplications and rearrangements**
- **Missing data challenges (i.e., sequencing gaps)**

Base-pair Sequence Alignment

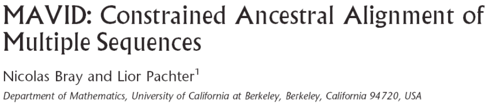


Aligning Multiple Genomic Sequences
With the Threaded Blockset Aligner

Mathieu Blanchette,[1,6] W. James Kent,[2] Cathy Riemer,[3] Laura Elnitski,[3] Arian F.A. Smit,[4] Krishna M. Roskin,[2] Robert Baertsch,[2] Kate Rosenbloom,[2] Hiram Clawson,[2] Eric D. Green,[5] David Haussler,[1,2] and Webb Miller[3,7]

[1]Howard Hughes Medical Institute and [2]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA; [3]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [4]Institute for Systems Biology, Seattle, Washington 98103, USA; [5]Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

***Genome Research* (2004) 14:708-715**

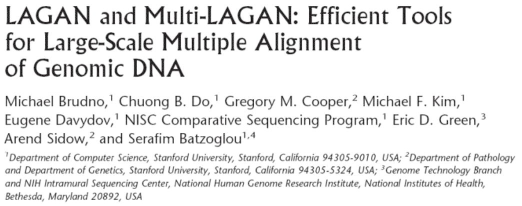


MAVID: Constrained Ancestral Alignment of
Multiple Sequences

Nicolas Bray and Lior Pachter[1]

Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA

***Genome Research* (2004) 14:693-699**



LAGAN and Multi-LAGAN: Efficient Tools
for Large-Scale Multiple Alignment
of Genomic DNA

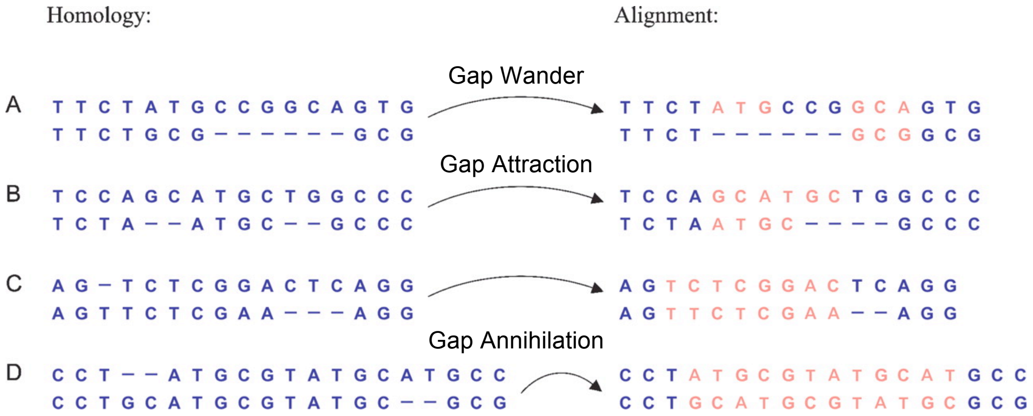
Michael Brudno,[1] Chuong B. Do,[1] Gregory M. Cooper,[2] Michael F. Kim,[1] Eugene Davydov,[1] NISC Comparative Sequencing Program,[1] Eric D. Green,[3] Arend Sidow,[2] and Serafim Batzoglou[1,4]

[1]Department of Computer Science, Stanford University, Stanford, California 94305-9010, USA; [2]Department of Pathology and Department of Genetics, Stanford University, Stanford, California 94305-5324, USA; [3]Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

***Genome Research* (2003) 13:721-31**

Types of Alignment Artifacts

Lunter et al. *Genome Res.* 18:298-309, 2008


Homology:
Alignment:
Gap Wander
A
T T C T A T G C C G G C A G T G
T T C T G C G — — — — — — G C G
T T C T A T G C C G G C A G T G
T T C T — — — — — — G C G G C G
Gap Attraction
B
T C C A G C A T G C T G G C C C
T C T A — — A T G C — — G C C C
T C C A G C A T G C T G G C C C
T C T A A T G C — — — — G C C C
C
A G — T C T C G G A C T C A G G
A G T T C T C G A A — — — A G G
A G T C T C G G A C T C A G G
A G T T C T C G A A — — A G G
Gap Annihilation
D
C C T — — A T G C G T A T G C A T G C C
C C T G C A T G C G T A T G C — — G C G
C C T A T G C G T A T G C A T G C C
C C T G C A T G C G T A T G C G C G

Genome Res (2008) 18:199-200

Commentary

Confidence in comparative genomics

Elliott H. Margulies[1]

Genome Informatics Section, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

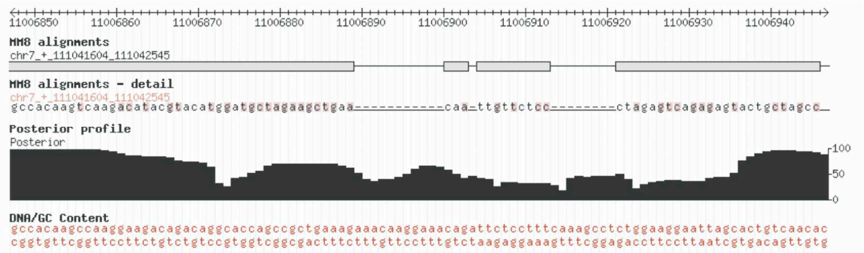
Genome Res (2008) 18: 298-309

Methods

Uncertainty in homology inferences: Assessing and improving genomic sequence alignment

Gerton Lunter,[1,3] Andrea Rocco,[2] Naila Mimouni,[2] Andreas Heger,[1] Alexandre Caldeira,[2] and Jotun Hein[2]

[1]MRC Functional Genetics Unit, University of Oxford, Department of Physiology, Anatomy, and Genetics, Oxford OX1 3QX, United Kingdom; [2]Department of Statistics, University of Oxford, Oxford Centre for Gene Function, Oxford, OX1 2TG, United Kingdom

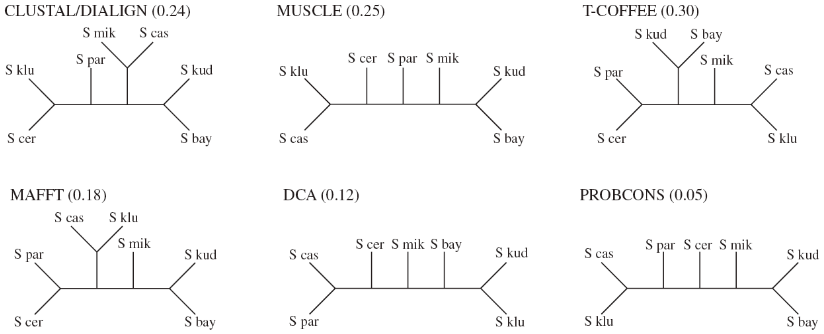


```
MM8 alignments - detail
chr7_+_111041604_111042545
gccacaagtcaagacatacgtacatggatgctagaagctgaa-----------caa-ttgttctcc--------ctagagtcagagagtactgctagcc_
```

```
DNA/GC Content
gccacaagccaaggaagacagacaggcaccagccgctgaaagaaacaaggaaacagattctcctttcaaagcctctggaaggaattagcactgtcaacac
cggtgttcggttccttctgtctgtccgtggtcggcgactttctttgttcctttgtctaagaggaaagtttcggagaccttccttaatcgtgacagttgtg
```

Alignment Uncertainty and Genomic Analysis

Karen M. Wong,[1] Marc A. Suchard,[2] John P. Huelsenbeck[3]*



Science **319**, 473-6 (2008)

Genome Browsers

UCSC Genome Bioinformatics

http://genome.ucsc.edu

e! project Ensembl

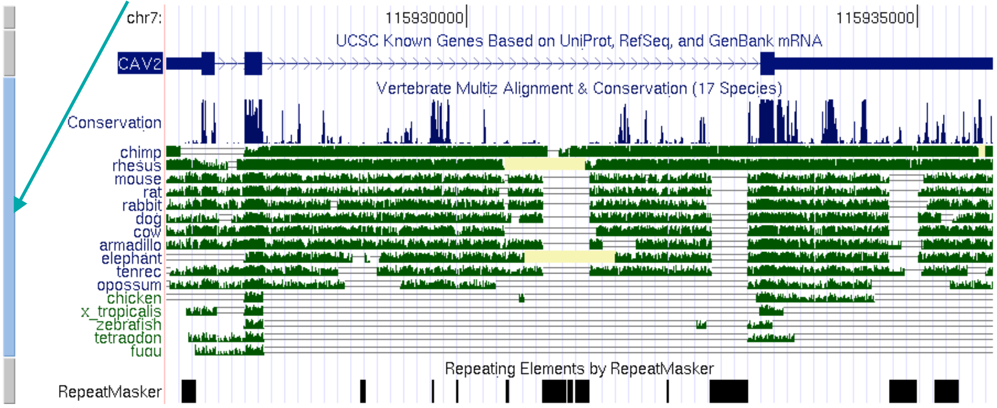
http://www.ensembl.org

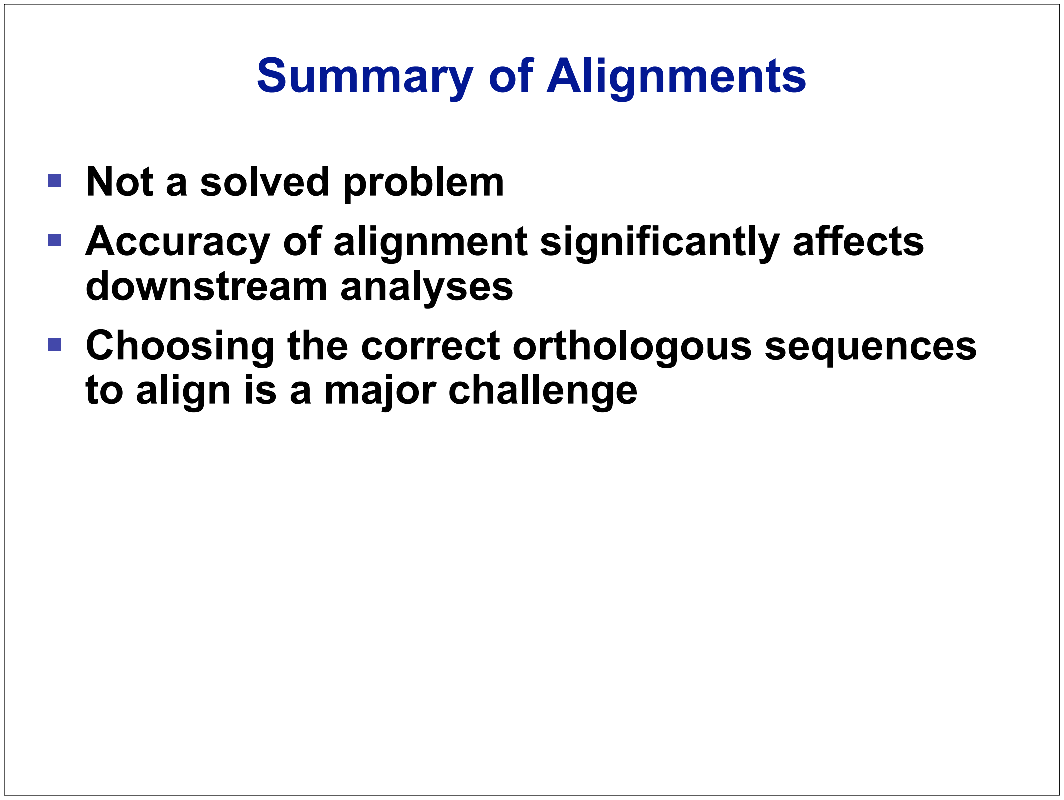
NCBI Map Viewer

http://www.ncbi.nlm.nih.gov/mapview/

Multi-sequence Alignments at UCSC

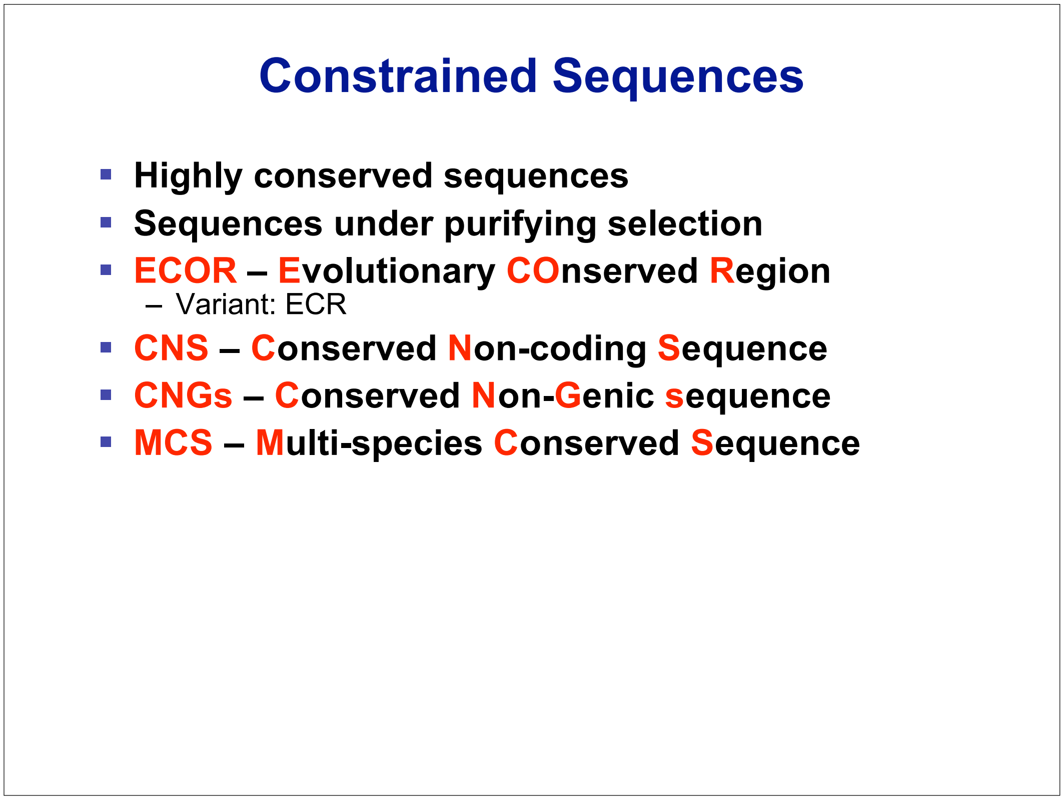
Click here for track details page





Summary of Alignments

- Not a solved problem
- Accuracy of alignment significantly affects downstream analyses
- Choosing the correct orthologous sequences to align is a major challenge



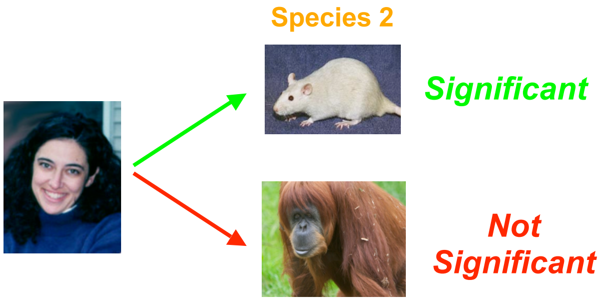
Constrained Sequences

- Highly conserved sequences
- Sequences under purifying selection
- ECOR – Evolutionary COnserved Region
  - Variant: ECR
- CNS – Conserved Non-coding Sequence
- CNGs – Conserved Non-Genic sequence
- MCS – Multi-species Conserved Sequence

Finding Constrained Sequences

85% Identical

```
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
          || |||| |||| |||||||||| |||||| |||| ||||
Species 2 CACGGGCTAATTCGCCCATTGGCTATGGGG-CCCAGCGTA
```



Compare to some measure of neutral evolution

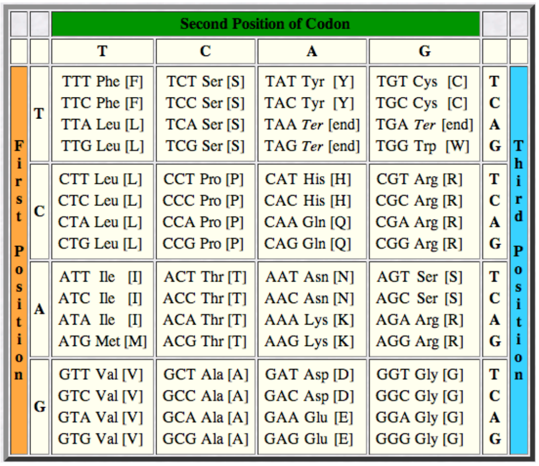
Neutral Evolution

- **No selective pressure/advantage to keep or change the DNA sequence**
- **Amount of observed variation correlates with:**
  - Rate of mutation
  - Length of breeding cycle
  - Amount of time since the last common ancestor
- **The neutral rate can vary across the genome**

Types of Neutrally Evolving DNA

- **4-Fold Degenerate Sites**

  Third position of codons which can be any base and code for the same amino acid



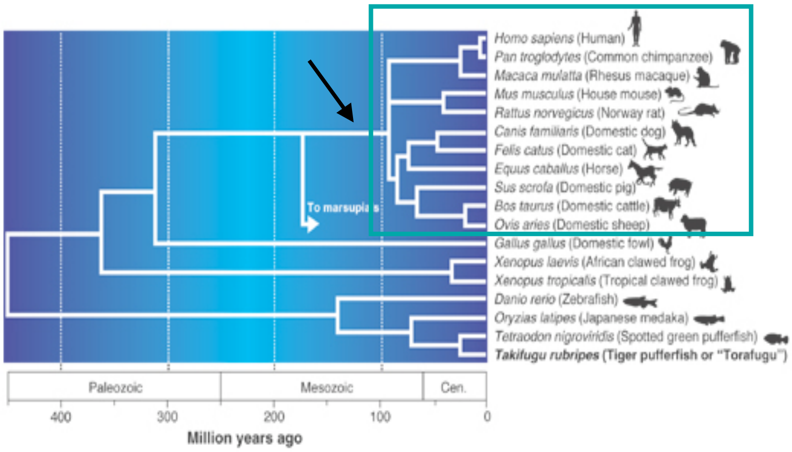
		Second Position of Codon					
		T	C	A	G		
First Position	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	Third Position
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
		TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A	
		TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G	
	C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T	
		CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C	
		CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A	
		CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G	
	A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T	
		ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C	
		ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A	
		ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G	
	G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T	
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C	
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A	
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G	

http://psyche.uthct.edu/shaun/SBlack/geneticd.html

Types of Neutrally Evolving DNA

- **Ancestral Repeats**

  Ancient Relics of Transposons Inserted Prior to the Eutherian Radiation


Homo sapiens (Human)
Pan troglodytes (Common chimpanzee)
Macaca mulatta (Rhesus macaque)
Mus musculus (House mouse)
Rattus norvegicus (Norway rat)
Canis familiaris (Domestic dog)
Felis catus (Domestic cat)
Equus caballus (Horse)
Sus scrofa (Domestic pig)
Bos taurus (Domestic cattle)
Ovis aries (Domestic sheep)
Gallus gallus (Domestic fowl)
Xenopus laevis (African clawed frog)
Xenopus tropicalis (Tropical clawed frog)
Danio rerio (Zebrafish)
Oryzias latipes (Japanese medaka)
Tetraodon nigroviridis (Spotted green pufferfish)
Takifugu rubripes (Tiger pufferfish or "Torafugu")
To marsupials
Paleozoic
Mesozoic
Cen.
400
300
200
100
0
Million years ago

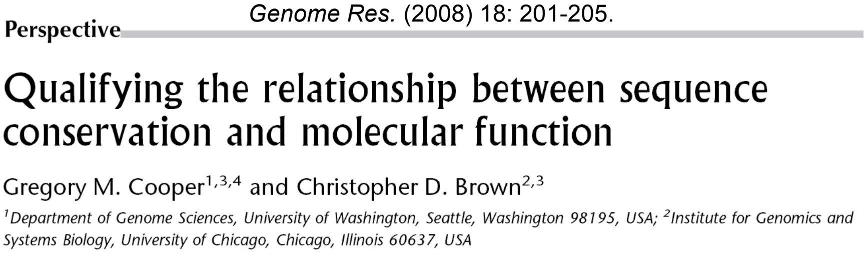
Adapted from Hedges & Kumar, *Science* **297:**1283-5

Conservation *vs.* Constraint

- **Conservation is simply a measure of similarity**
- **Constraint implies *purifying selection***

"Conservation, when observed to be in excess of the levels predicted by a neutral model, can be used to infer constraint"

Genome Res. (2008) 18: 201-205.



Perspective

Qualifying the relationship between sequence conservation and molecular function

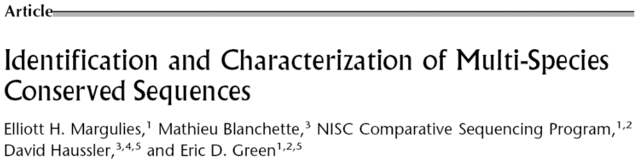
Gregory M. Cooper[1,3,4] and Christopher D. Brown[2,3]

[1]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;* [2]*Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA*

Major Approaches used for Sequence Constraint Detection

Binomial-based Method

binCons



Article

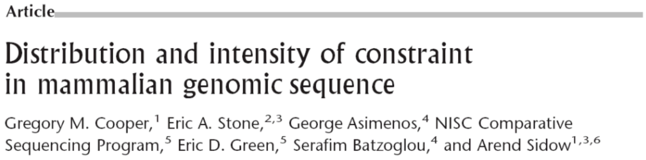
Identification and Characterization of Multi-Species Conserved Sequences

Elliott H. Margulies,[1] Mathieu Blanchette,[3] NISC Comparative Sequencing Program,[1,2] David Haussler,[3,4,5] and Eric D. Green[1,2,5]

***Genome Research* (2003) 13:2507-2518**

Genomic Evolutionary Rate Profiling

GERP



Article

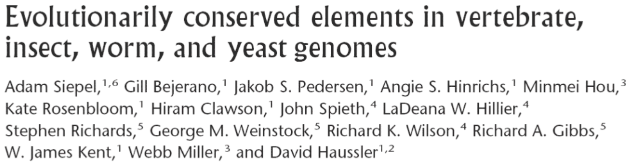
Distribution and intensity of constraint in mammalian genomic sequence

Gregory M. Cooper,[1] Eric A. Stone,[2,3] George Asimenos,[4] NISC Comparative Sequencing Program,[5] Eric D. Green,[5] Serafim Batzoglou,[4] and Arend Sidow[1,3,6]

***Genome Research* (2005) 15:901-913**

PHylogenetic Analysis with Space/Time models

phastCons

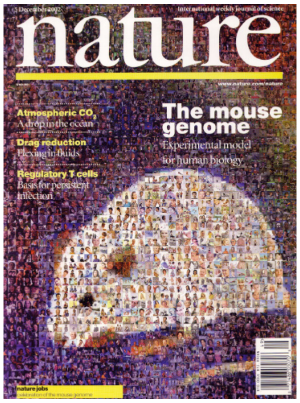


Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes

Adam Siepel,[1,6] Gill Bejerano,[1] Jakob S. Pedersen,[1] Angie S. Hinrichs,[1] Minmei Hou,[3] Kate Rosenbloom,[1] Hiram Clawson,[1] John Spieth,[4] LaDeana W. Hillier,[4] Stephen Richards,[5] George M. Weinstock,[5] Richard K. Wilson,[4] Richard A. Gibbs,[5] W. James Kent,[1] Webb Miller,[3] and David Haussler[1,2]

***Genome Research* (2005) 15:1034-1050**

Insights from Human-Rodent Sequence Comparisons



Nature 420:520, 2002

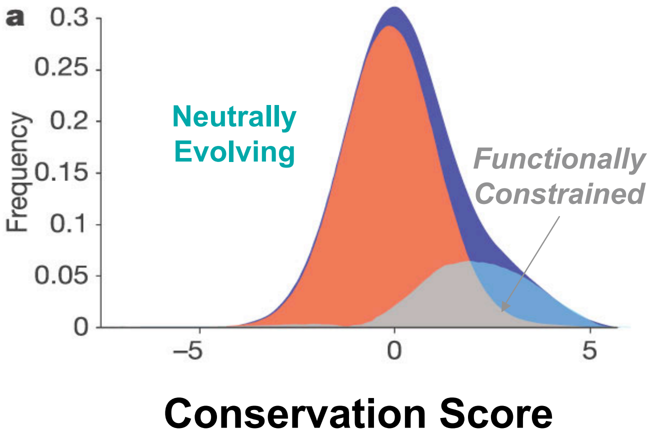
Nature 428:493, 2004

- **Sequence Conservation**
  - ~40% in Alignments
  - ~5% Under "Selection"
    - ~1.5% Protein Coding
    - ~3.5% Non-Coding

Determining the Fraction of Sequence Under Purifying Selection

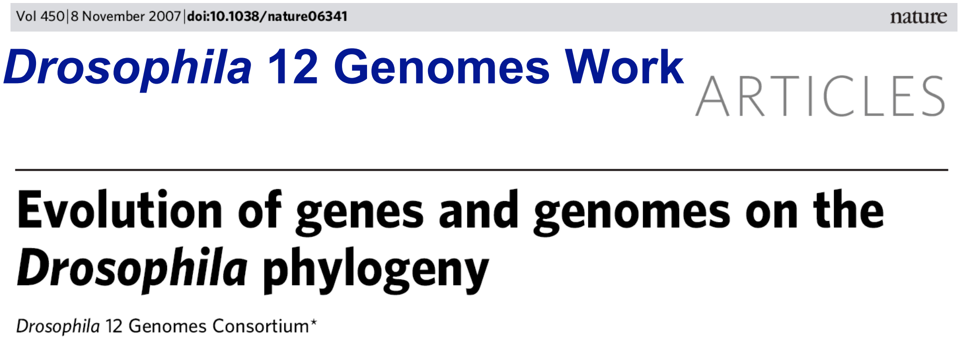
Neutral + Functional = Genome-Wide

Genome-Wide – Neutral = Functional



Adapted From Figure 28, *Nature* **420**:553

Drosophila 12 Genomes Work



Vol 450|8 November 2007|doi:10.1038/nature06341 nature

ARTICLES

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

nature
GENOMICS ON A PHYLOGENY
Twelve Drosophila sequences traverse 400 million years of evolution

The ENCODE Project

- ENCODE:

  ENCyclopedia Of DNA Elements
- Goal: Compile a *comprehensive encyclopedia* of all functional elements in the human genome
- Initial pilot project: 1% of human genome
- Apply multiple approaches to study and analyze that 1% in an international consortium

Integration of ENCODE Data

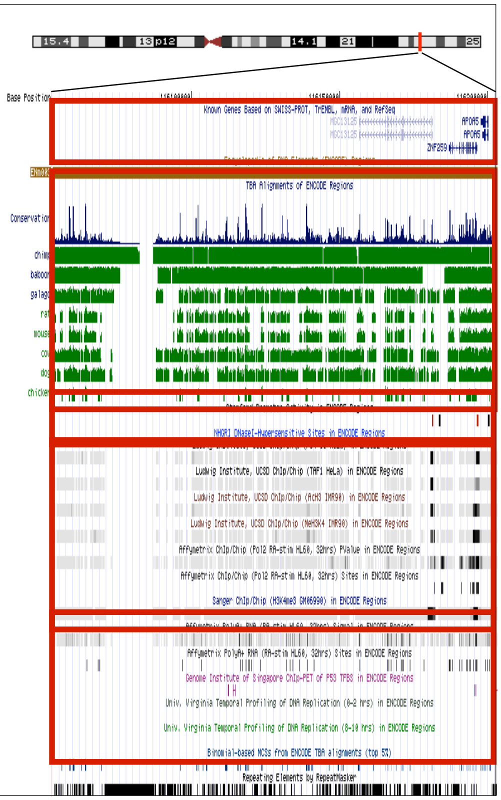
Gene Annotation

Comparative Sequence Analysis

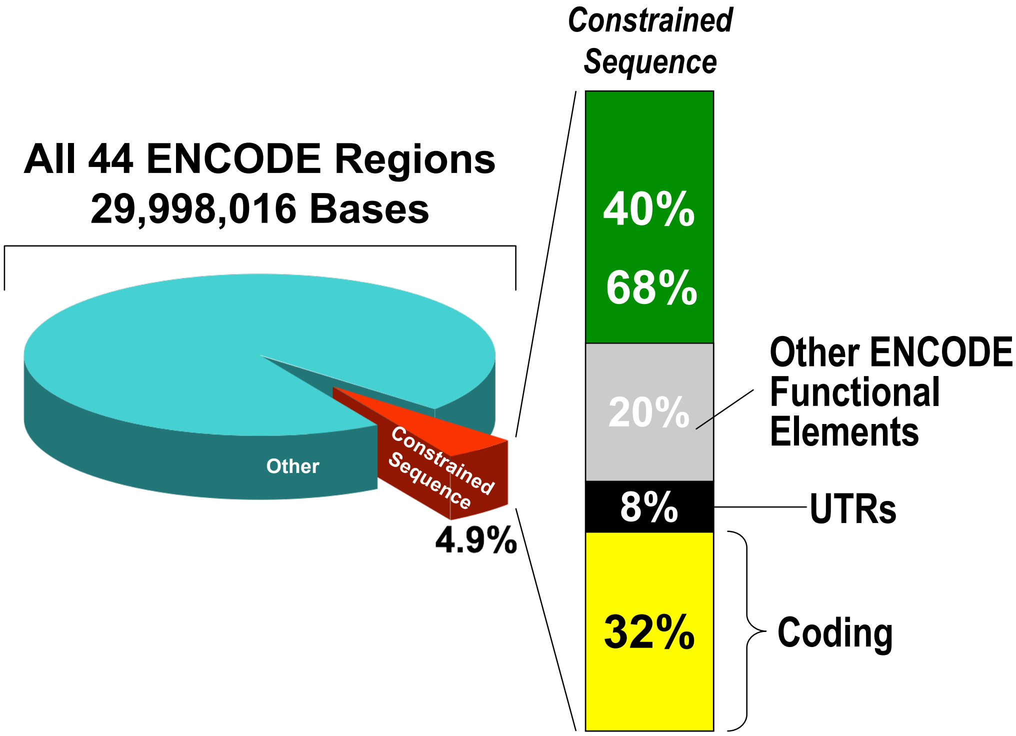
Promoter Identification

DNA-Protein Interactions

RNA Expression

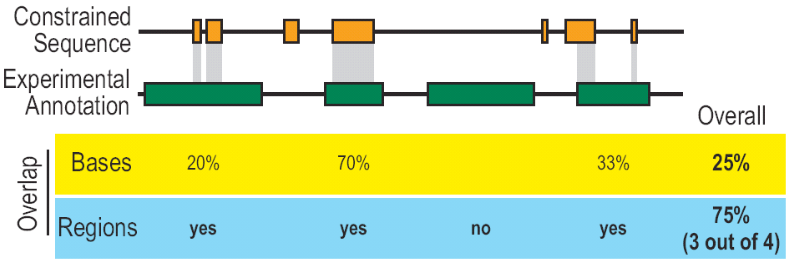

15.4
13 p12
14.1
21
25
Base Position
116100000
116150000
116200000
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
MGC13125
MGC13125
APOA5
APOA5
ZNF259
ENm003
TBA Alignments of ENCODE Regions
Conservation
chimp
baboon
galago
rat
mouse
cow
dog
chicken
NHGRI DNaseI-Hypersensitive Sites in ENCODE Regions
Ludwig Institute, UCSD ChIp/Chip (TAF1 HeLa) in ENCODE Regions
Ludwig Institute, UCSD ChIp/Chip (AcH3 IMR90) in ENCODE Regions
Ludwig Institute, UCSD ChIp/Chip (MeH3K4 IMR90) in ENCODE Regions
Affymetrix ChIp/Chip (Pol2 RA-stim HL60, 32hrs) PValue in ENCODE Regions
Affymetrix ChIp/Chip (Pol2 RA-stim HL60, 32hrs) Sites in ENCODE Regions
Sanger ChIp/Chip (H3K4me3 GM06990) in ENCODE Regions
Affymetrix PolyA+ RNA (RA-stim HL60, 32hrs) Sites in ENCODE Regions
Genome Institute of Singapore ChIp-PET of P53 TFBS in ENCODE Regions
Univ. Virginia Temporal Profiling of DNA Replication (0-2 hrs) in ENCODE Regions
Univ. Virginia Temporal Profiling of DNA Replication (8-10 hrs) in ENCODE Regions
Binomial-based MCSs from ENCODE TBA alignments (top 5%)
Repeating Elements by RepeatMasker

All 44 ENCODE Regions
29,998,016 Bases

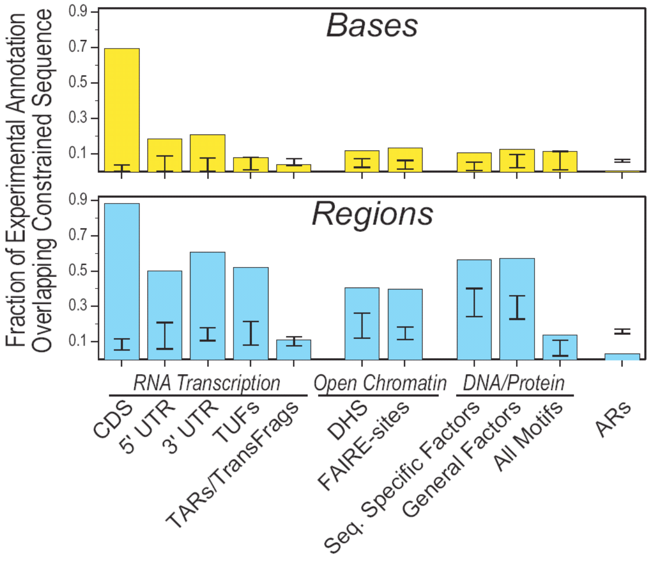


01/06/2006 MSA-Compiled Dataset

Assessing the Overlap between Constrained Sequences and Experimental Annotations



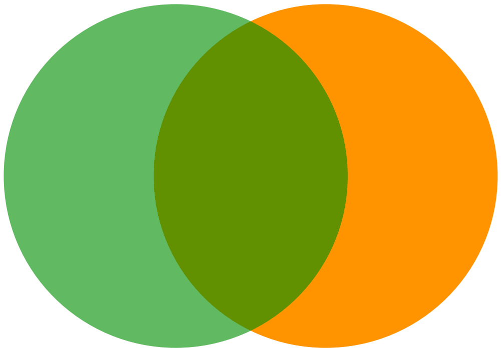
Overlap between Constrained Sequences and Experimental Annotations



Margulies et al. (2007) Relationship between Evolutionary Constraint and Genome Function in 1% of the Human Genome. *Genome Res*, 17:760-774.

The ENCODE Consortium (2007) The ENCODE Pilot Project: Functional Annotation of 1% of the Human Genome, *Nature*, 447: 799-816

Current Understanding of Relationship between *Constrained* and *Functional* *Sequences?*

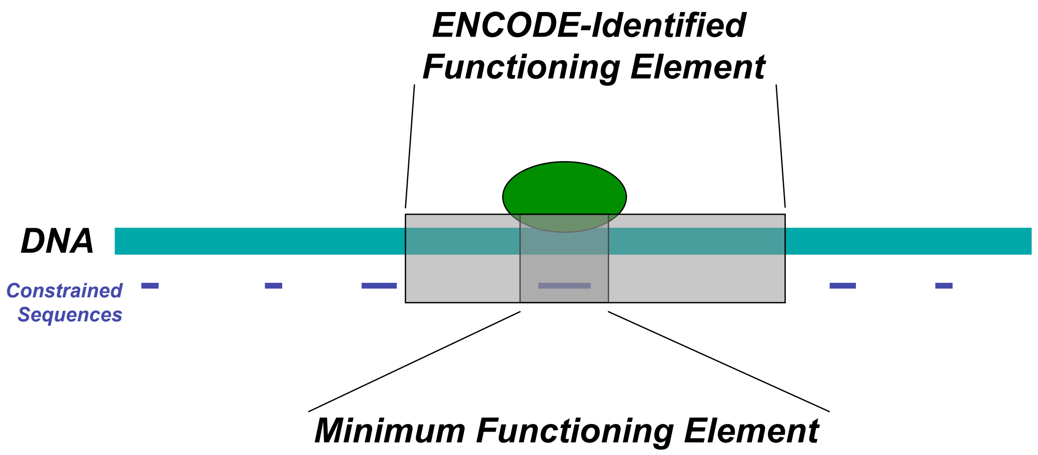


- **40% of all constrained sequences do not correspond to functional annotations**
- **Many functional annotations fail to overlap at least some constrained sequence**

Why not a Complete Correlation Between Sequence Constraint and Sequence Function?

- **Likely not due to false positive experimental annotations**
- **Did not ascertain all functions at all time-points**
- **Reproducible biochemical events with no biological consequence to the organism**
- **Annotation is larger than the functioning unit**

Resolution Issue



ENCODE-Identified Functioning Element

DNA

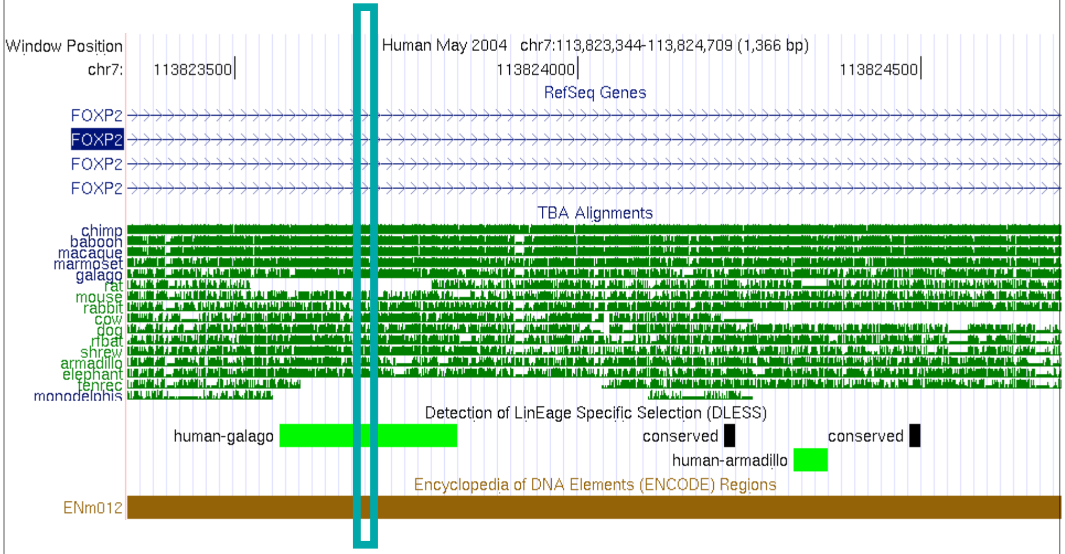
Constrained Sequences

Minimum Functioning Element

Why not a Complete Correlation Between Sequence Constraint and Sequence Function?

- Likely <u>not</u> due to false positive experimental annotations
- Did not ascertain all functions at all time-points
- Reproducible biochemical events with no biological consequence to the organism
- Annotation is larger than the functioning unit
- **Not constrained throughout all mammals**
  Lineage-specific constraint beyond this 5%

Lineage-Specific Sequence Conservation



Window Position — Human May 2004 chr7:113,823,344-113,824,709 (1,366 bp)

chr7: 113823500 | 113824000 | 113824500

RefSeq Genes

FOXP2
FOXP2
FOXP2
FOXP2

TBA Alignments

chimp, baboon, macaque, marmoset, galago, rat, mouse, rabbit, cow, dog, rfbat, shrew, armadillo, elephant, tenrec, monodelphis

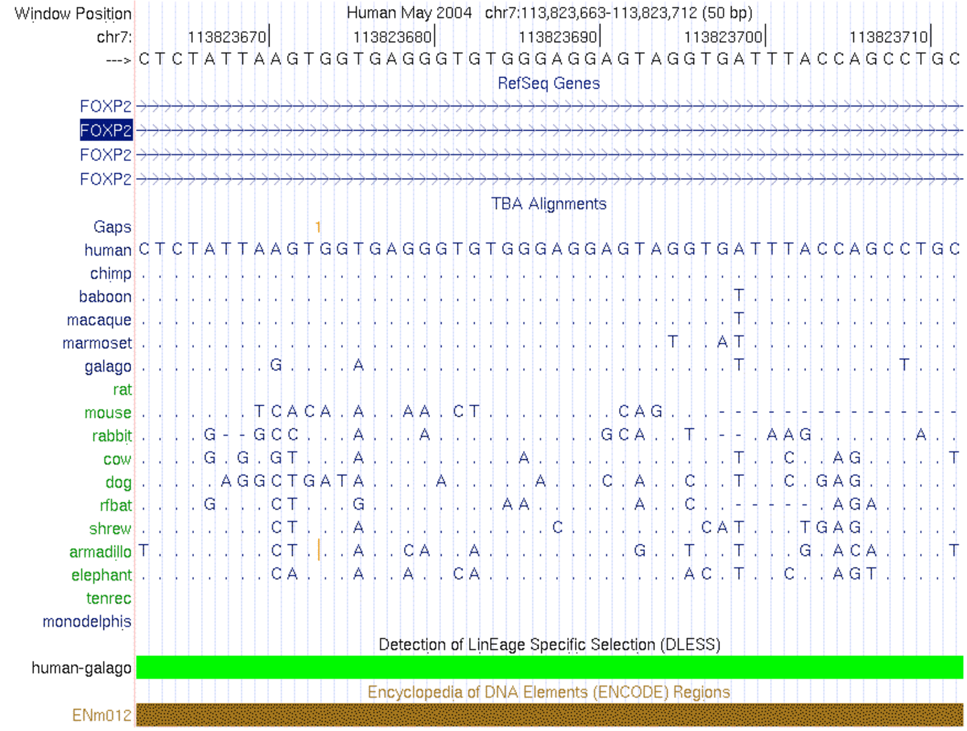
Detection of LinEage Specific Selection (DLESS)

human-galago — conserved — conserved — human-armadillo

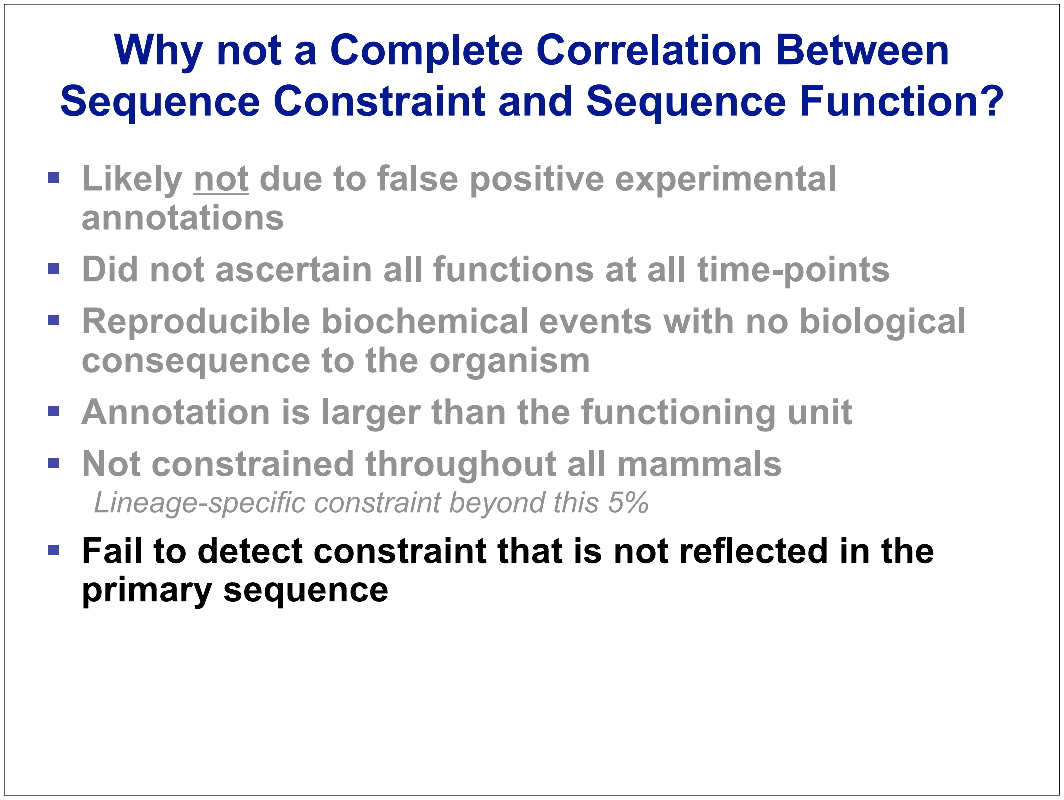
Encyclopedia of DNA Elements (ENCODE) Regions

ENm012

Adam Siepel
Cornell University

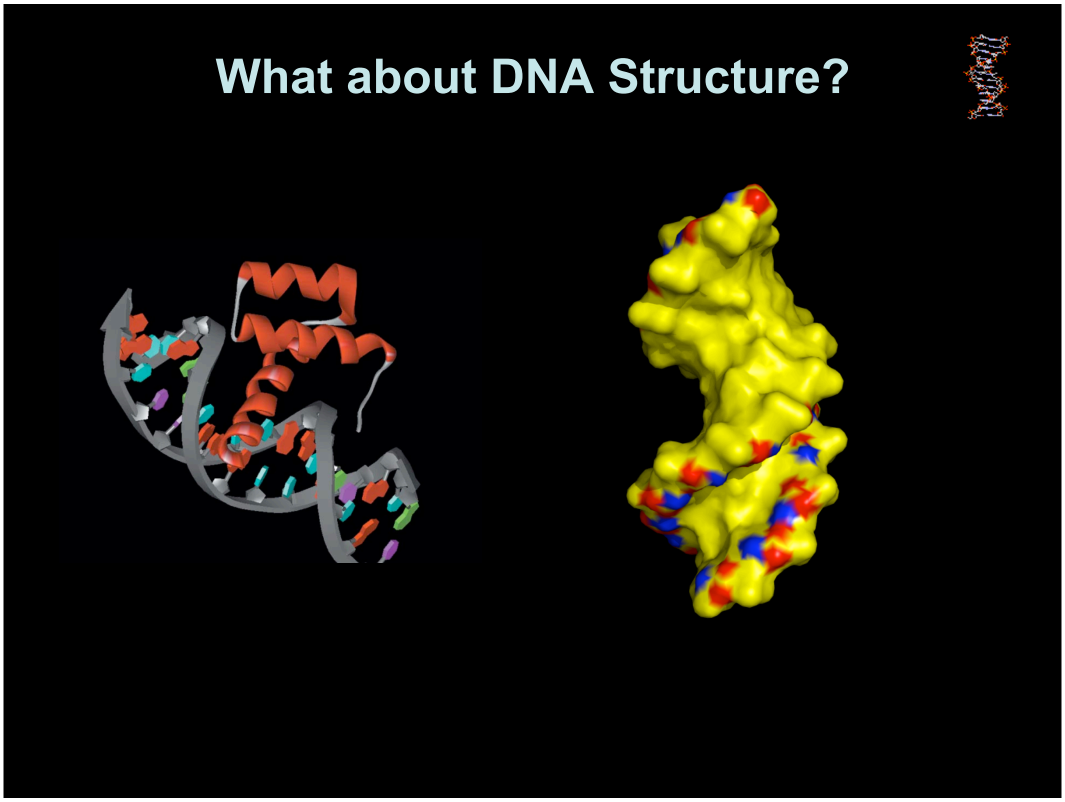


```
Window Position   Human May 2004   chr7:113,823,663-113,823,712 (50 bp)
           chr7:       113823670|    113823680|    113823690|    113823700|    113823710|
           ---> CTCTATTAAGTGGTGAGGGTGTGGGAGGAGTAGGTGATTTACCAGCCTGC
                                     RefSeq Genes
          FOXP2
          FOXP2
          FOXP2
          FOXP2
                                     TBA Alignments
           Gaps           1
          human CTCTATTAAGTGGTGAGGGTGTGGGAGGAGTAGGTGATTTACCAGCCTGC
          chimp ..................................................
         baboon ....................................T.............
        macaque ....................................T.............
       marmoset ................................T..AT.............
         galago ........G....A......................T.........T...
            rat
          mouse .......TCACA.A..AA.CT........CAG...---------------
         rabbit ....G--GCC...A...A........GCA..T.--.AAG......A..
            cow ....G.G.GT...A.........A............T..C..AG.....T
            dog .....AGGCTGATA....A.....A...C.A..C..T..C.GAG......
          rfbat ....G...CT...G........AA......A..C..-----.AGA.....
          shrew ........CT...A...........C........CAT...TGAG......
      armadillo T.......CT.|..A..CA..A.........G..T..T...G.ACA....T
       elephant ........CA...A..A..CA............AC.T..C..AGT.....
         tenrec
    monodelphis
                   Detection of LinEage Specific Selection (DLESS)
   human-galago
                   Encyclopedia of DNA Elements (ENCODE) Regions
        ENm012
```



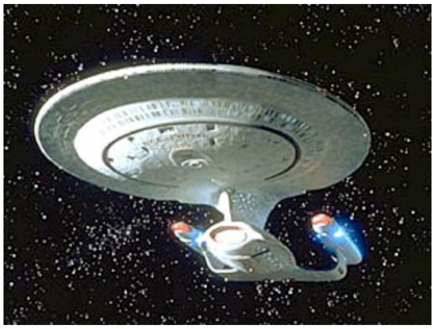
Why not a Complete Correlation Between Sequence Constraint and Sequence Function?

- Likely not due to false positive experimental annotations
- Did not ascertain all functions at all time-points
- Reproducible biochemical events with no biological consequence to the organism
- Annotation is larger than the functioning unit
- Not constrained throughout all mammals
  Lineage-specific constraint beyond this 5%
- **Fail to detect constraint that is not reflected in the primary sequence**



What about DNA Structure?

Next Generation Sequencing



Why Sequence DNA?

1) *De novo* Sequencing

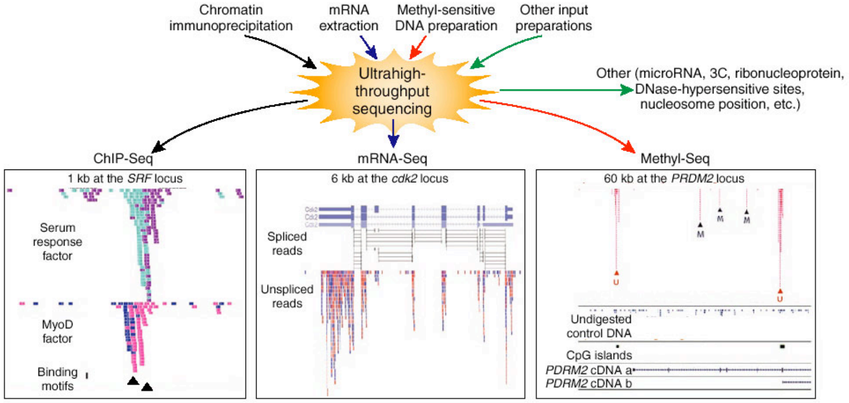
2) Variation (SNP) Detection

3) “Counting” Experiments

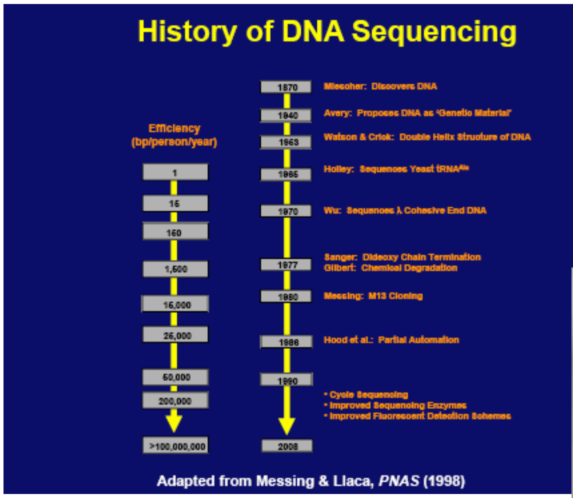
NATURE METHODS | VOL.5 NO.1 | JANUARY 2008 | 19

Sequence census methods for functional genomics

Barbara Wold & Richard M Myers



Plateau in Sequencing Technology



History of DNA Sequencing

Efficiency (bp/person/year)

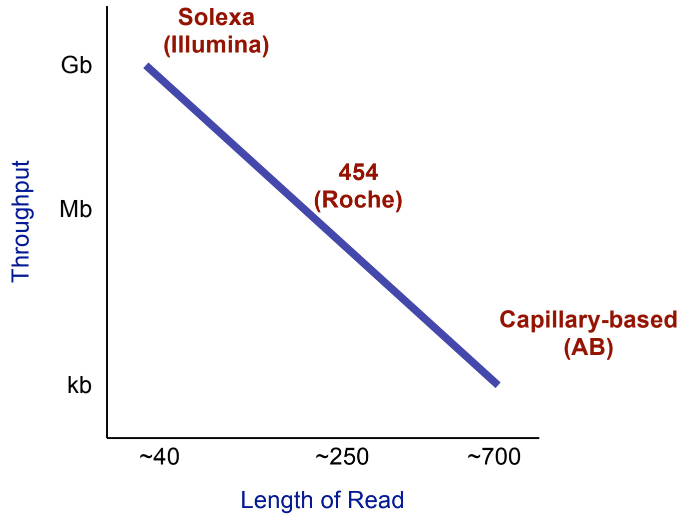
Efficiency (bp/person/year)	Year	Event
	1870	Miescher: Discovers DNA
	1940	Avery: Proposes DNA as 'Genetic Material'
	1953	Watson & Crick: Double Helix Structure of DNA
1	1965	Holley: Sequences Yeast tRNA^{Ala}
15	1970	Wu: Sequences λ Cohesive End DNA
150		
1,500	1977	Sanger: Dideoxy Chain Termination Gilbert: Chemical Degradation
15,000	1980	Messing: M13 Cloning
25,000	1986	Hood et al.: Partial Automation
50,000	1990	• Cycle Sequencing • Improved Sequencing Enzymes • Improved Fluorescent Detection Schemes
200,000		
>100,000,000	2008	

Adapted from Messing & Llaca, *PNAS* (1998)

Current Topics in Genome Analysis, E. Green, Lecture 1

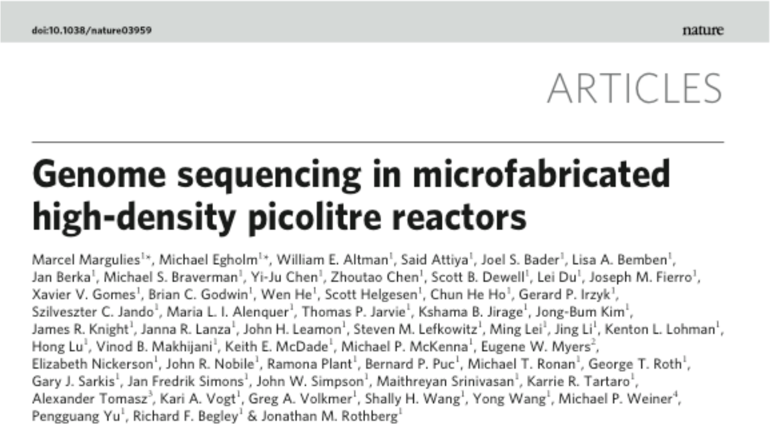
AB 3730 xl

Trade-offs with Newer Sequencing Technologies



Throughput = $\frac{\text{Amount of Sequence Generated}}{\text{Unit of Time or Cost}}$

454 Sequencing Technology



doi:10.1038/nature03959 nature

ARTICLES

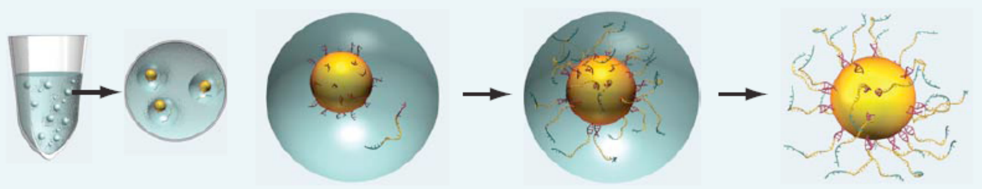
Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1*], Michael Egholm[1*], William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

Nature 31st July 2005

454 LIFE SCIENCES

Emulsion PCR (Template Prep)



Anneal sstDNA to an excess of DNA Capture Beads

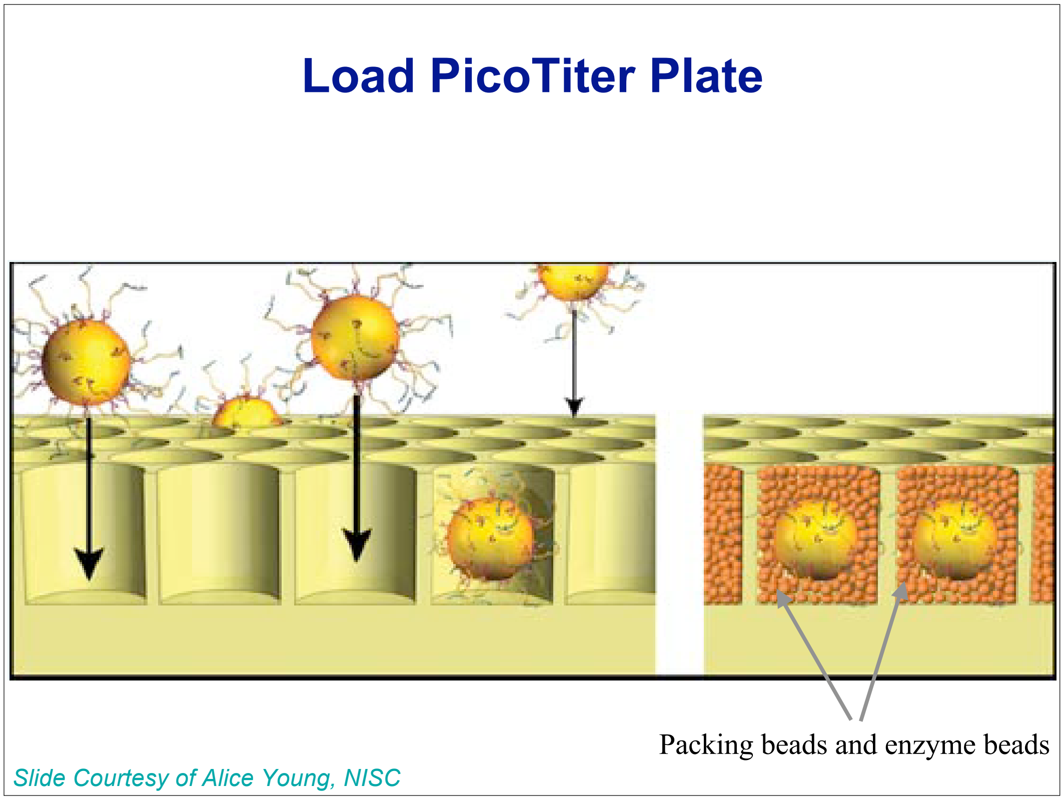
Emulsify beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

Break microreactors, enrich for DNA-positive beads

Each bubble in the emulsion will potentially contain a different fragment.

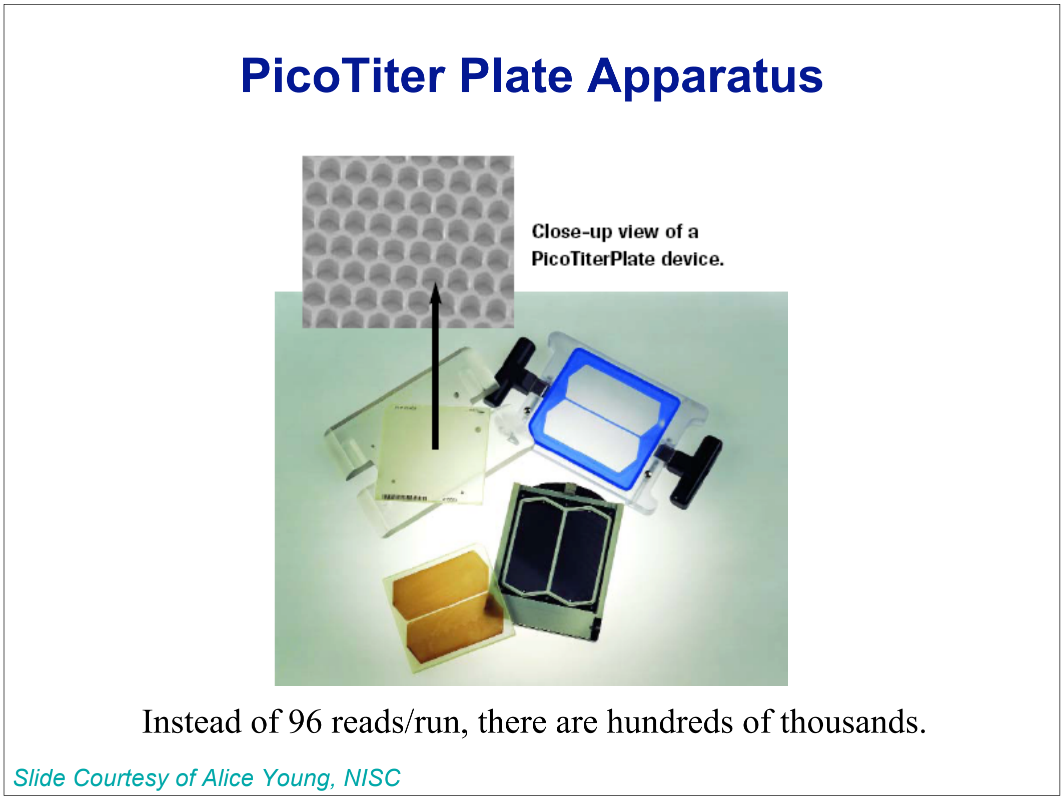
Slide Courtesy of Alice Young, NISC



Load PicoTiter Plate

Packing beads and enzyme beads

Slide Courtesy of Alice Young, NISC

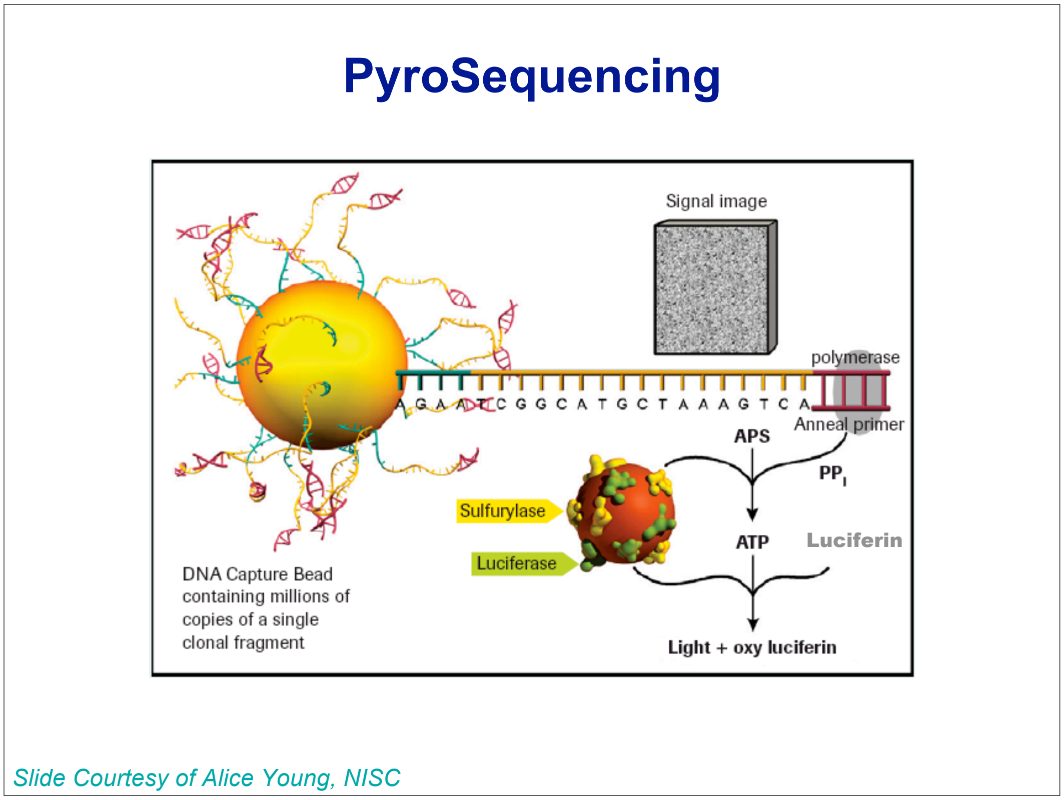


PicoTiter Plate Apparatus

Close-up view of a PicoTiterPlate device.

Instead of 96 reads/run, there are hundreds of thousands.

Slide Courtesy of Alice Young, NISC



PyroSequencing

Signal image
polymerase
A G A A T C G G C A T G C T A A A G T C A
Anneal primer
APS
PP$_i$
Sulfurylase
Luciferase
ATP
Luciferin
Light + oxy luciferin
DNA Capture Bead containing millions of copies of a single clonal fragment

Slide Courtesy of Alice Young, NISC

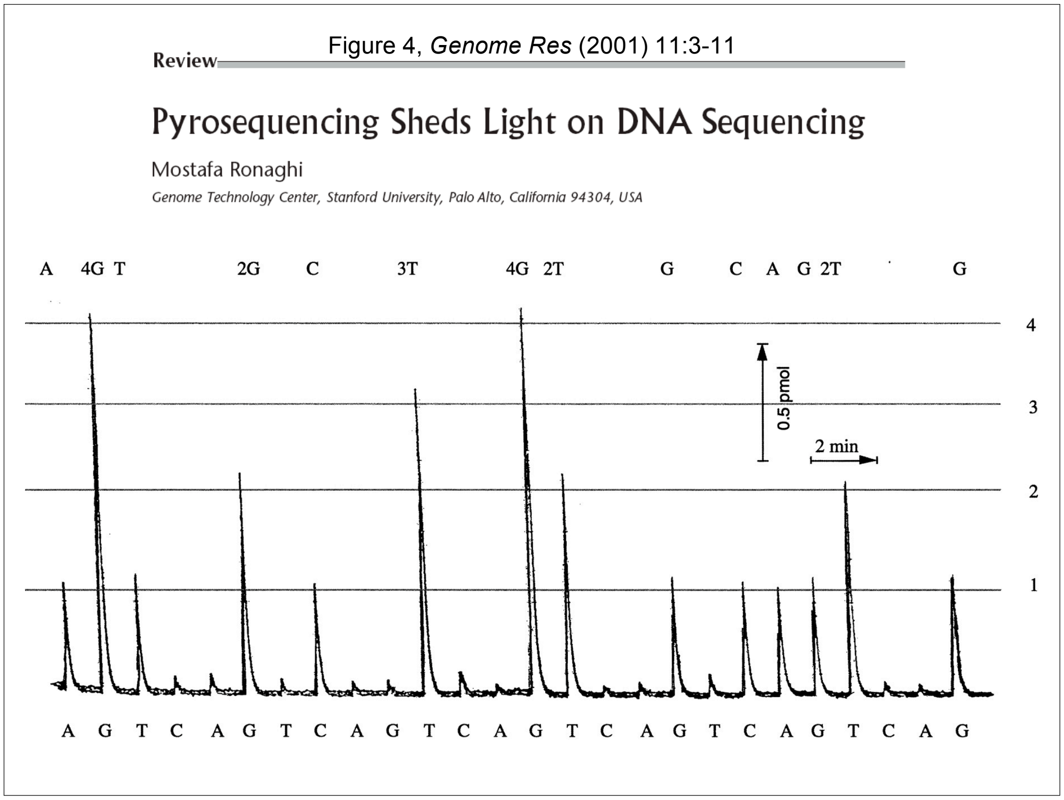


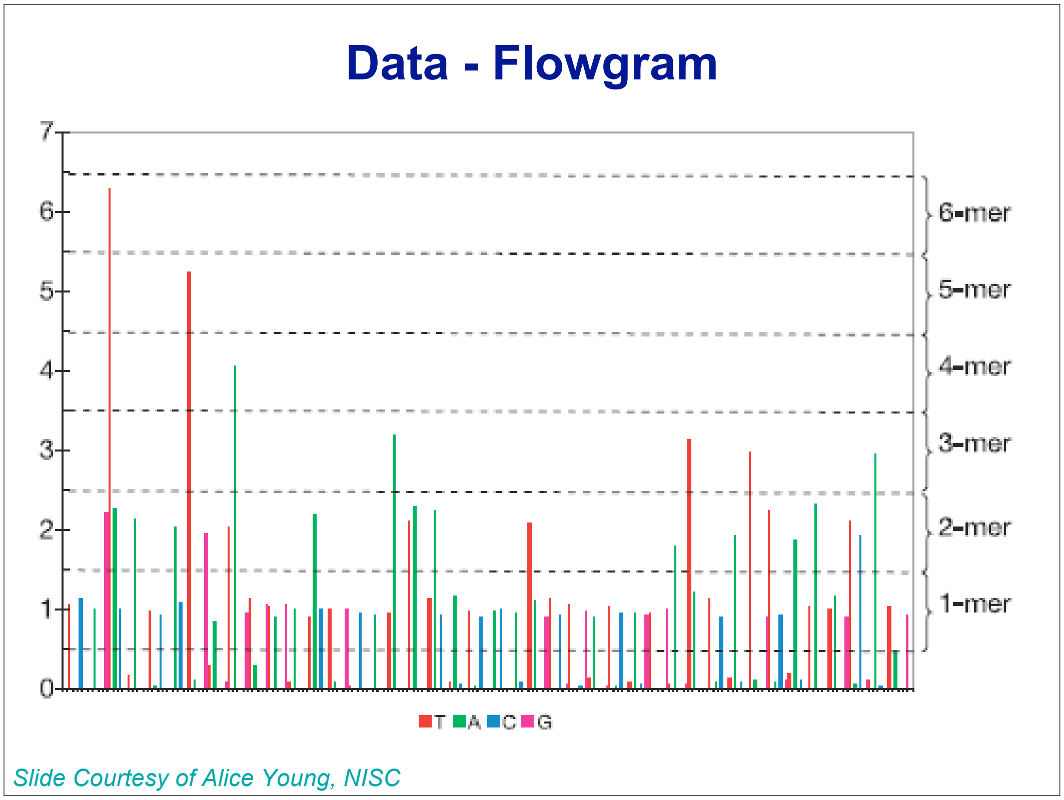
Figure 4, *Genome Res* (2001) 11:3-11

Review

Pyrosequencing Sheds Light on DNA Sequencing

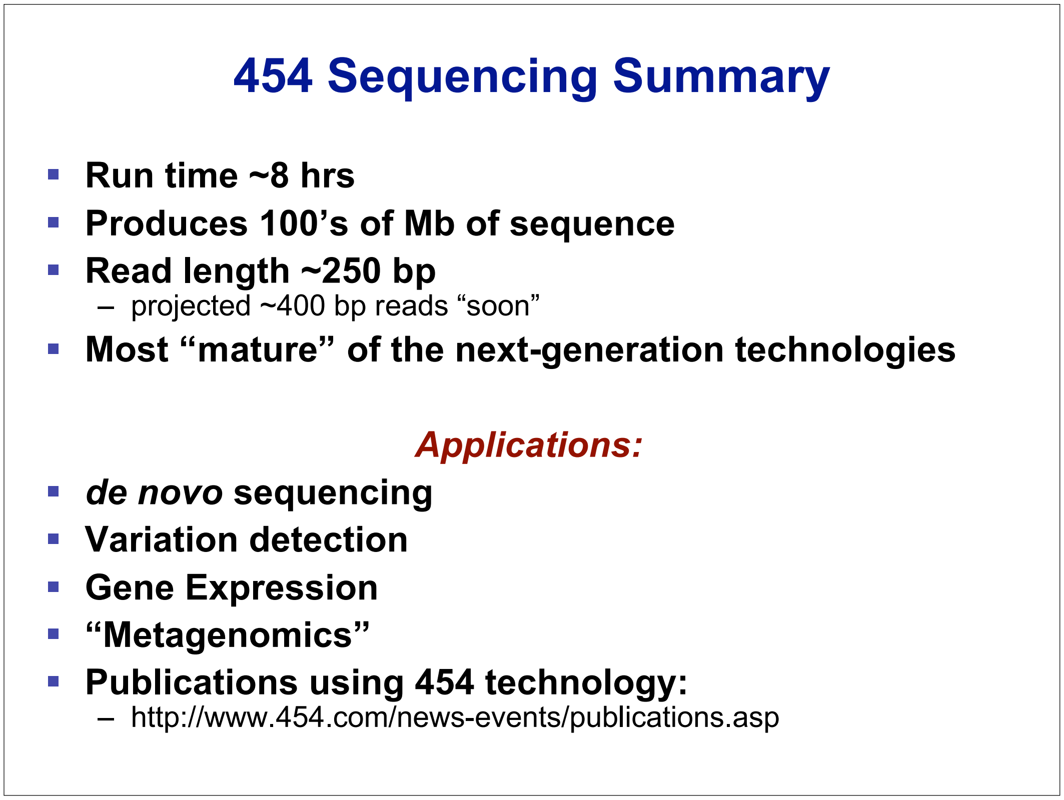
Mostafa Ronaghi

Genome Technology Center, Stanford University, Palo Alto, California 94304, USA



Data - Flowgram

Slide Courtesy of Alice Young, NISC



454 Sequencing Summary

- **Run time ~8 hrs**
- **Produces 100's of Mb of sequence**
- **Read length ~250 bp**
  - projected ~400 bp reads "soon"
- **Most "mature" of the next-generation technologies**

Applications:

- ***de novo* sequencing**
- **Variation detection**
- **Gene Expression**
- **"Metagenomics"**
- **Publications using 454 technology:**
  - http://www.454.com/news-events/publications.asp

ARTICLES

Nature, 2006 November 16; vol. (7117), 444 330-336

Analysis of one million base pairs of Neanderthal DNA

Richard E. Green[1], Johannes Krause[1], Susan E. Ptak[1], Adrian W. Briggs[1], Michael T. Ronan[2], Jan F. Simons[2], Lei Du[2], Michael Egholm[2], Jonathan M. Rothberg[2], Maja Paunovic[3]‡ & Svante Pääbo[1]

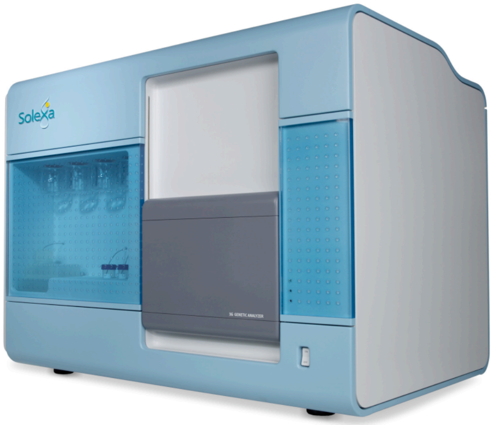
Science, 2006 November 17 ; vol. 314, 1113-111

Sequencing and Analysis of Neanderthal Genomic DNA

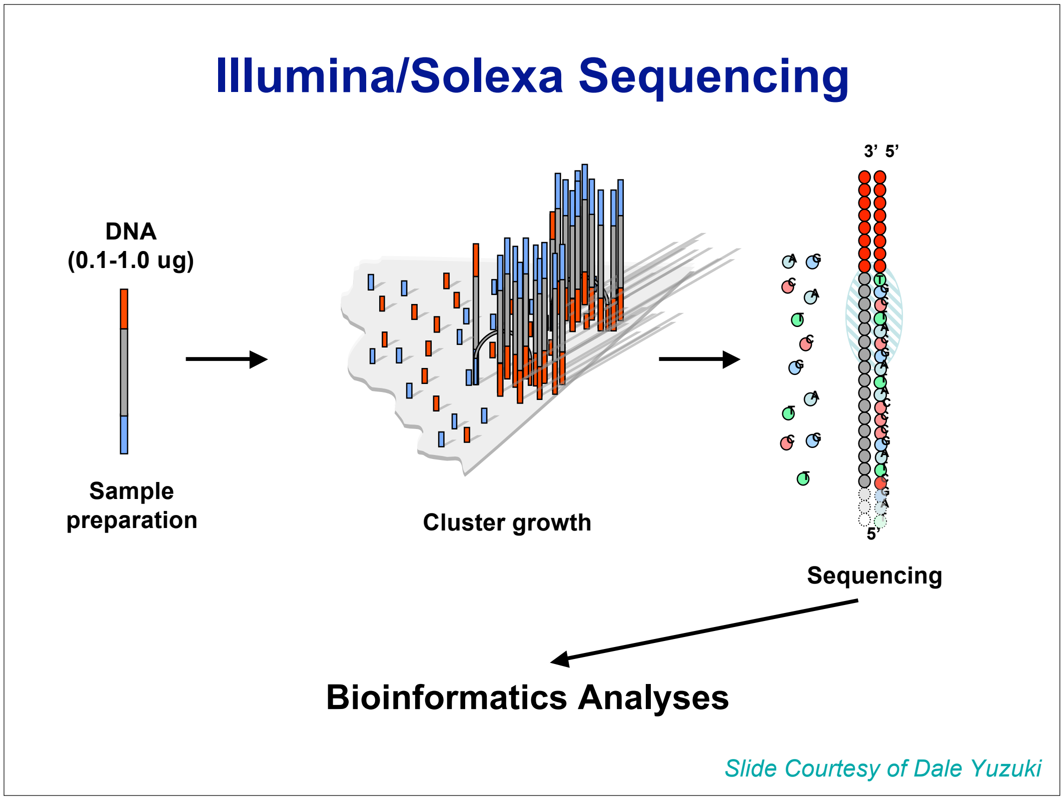
James P. Noonan,[1,2] Graham Coop,[3] Sridhar Kudaravalli,[3] Doug Smith,[1] Johannes Krause,[4] Joe Alessi,[1] Feng Chen,[1] Darren Platt,[1] Svante Pääbo,[4] Jonathan K. Pritchard,[3] Edward M. Rubin[1,2]*

http://popsci.typepad.com/photos/uncategorized/2007/10/25/laluezafox1lr.jpg

Illumina/Solexa 1G Genome Analyzer



illumina



Illumina/Solexa Sequencing

DNA
(0.1-1.0 ug)

Sample preparation

Cluster growth

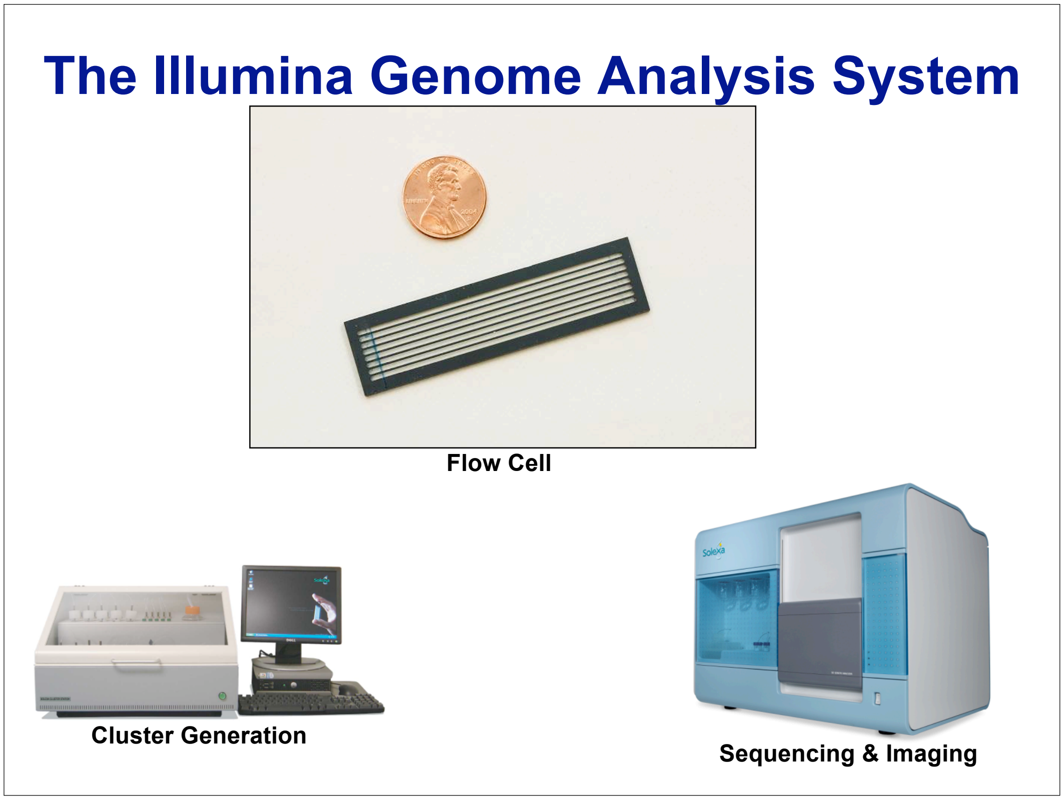
3' 5'

5'

Sequencing

Bioinformatics Analyses

Slide Courtesy of Dale Yuzuki

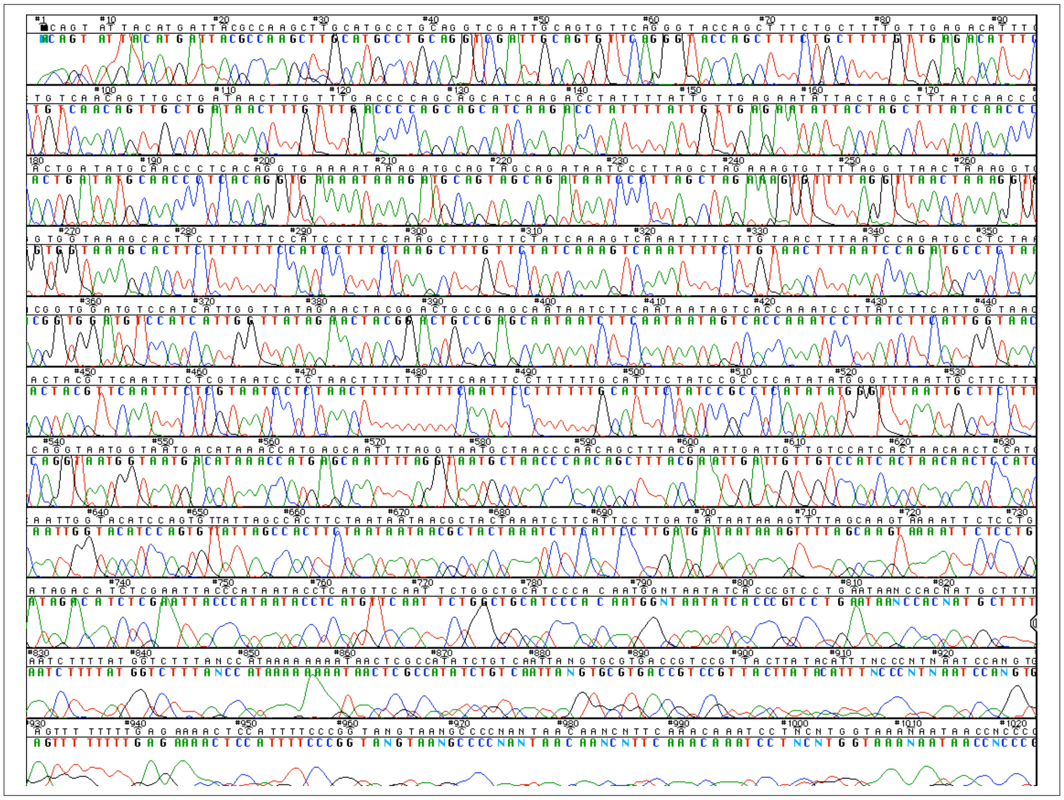


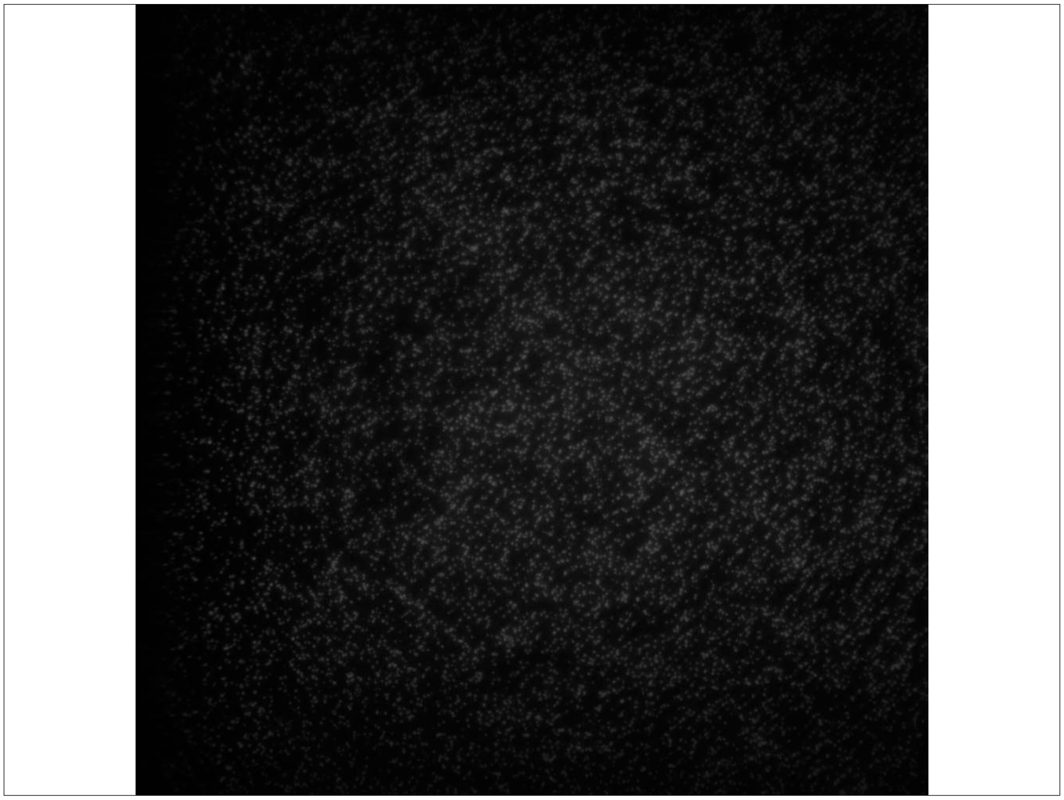
The Illumina Genome Analysis System

Flow Cell

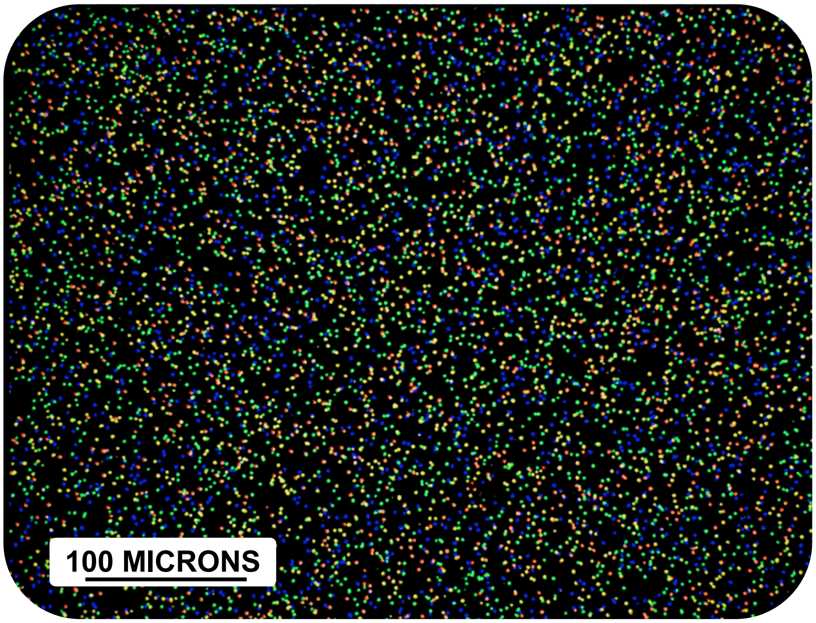
Cluster Generation

Sequencing & Imaging



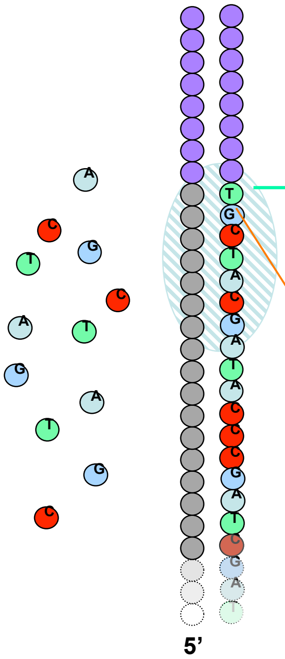


Pseudo-color Enhanced Image



Slide (though modified) Courtesy of Dale Yuzuki

Sequencing By Synthesis (SBS)



Cycle 1: Add sequencing reagents

First base incorporated

Remove unincorporated bases

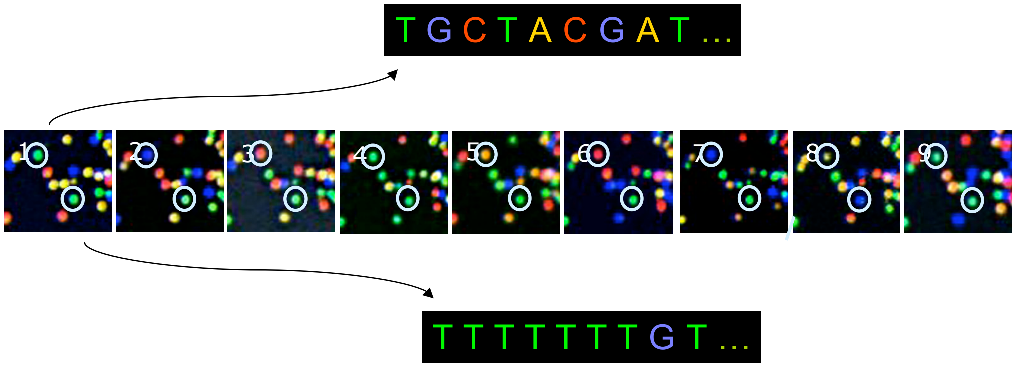
Detect signal

Cycle 2-n: Add sequencing reagents and repeat

All four labeled nucleotides in one reaction
Base-by-base sequencing
No problems with homopolymer repeats

Slide (though modified) Courtesy of Dale Yuzuki

Base Calling from Raw Data


T G C T A C G A T …

T T T T T T T G T …

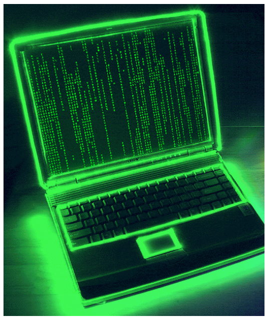
The identity of each base of a cluster is read off from sequential images.

Slide (though modified) Courtesy of Dale Yuzuki

Bioinformatics

- **~3 days per run**
- **~1Tb of “raw” data per run**
- **>1Gb of sequence**
  – 25-40 million reads

- **Significant computing horsepower needed for primary analyses**
  – Image analysis to base-calling
  – Alignment
  – Assembly



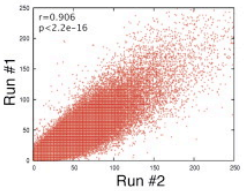
Illumina/Solexa Summary

- **Well-suited for "counting" based experiments**
- **Alternate approaches to alignment**
- **Quality of individual reads *vs.* depth of coverage**
  - De novo genome sequencing
  - Variation detection
- **Cheap sequence fast!**

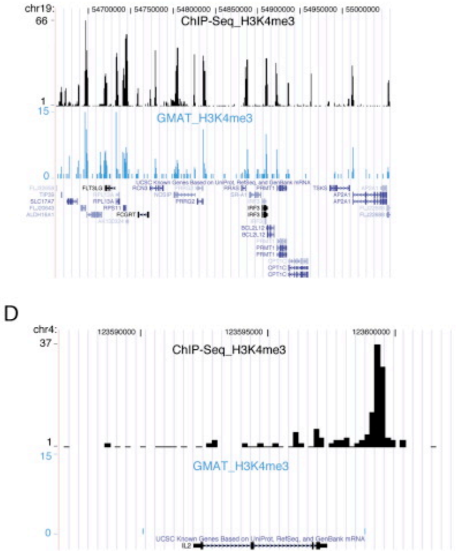
Cell (2007) May 18;129(4):823-37.

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1,3] Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]
[1] Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA
[2] Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3] These authors contributed equally to this work and are listed alphabetically.
*Correspondence: zhaok@nhlbi.nih.gov
DOI 10.1016/j.cell.2007.05.009



- **One of the first publications using Solexa data**
- **Reproducible data production**
- **Correlates with other sequence-based counting experiments**
- **Identify biologically-relevant patterns of histone methylation**
  - Transcription
  - Enhancers
  - Insulators
- **Stay tuned for Laura Elnitski's lecture!**



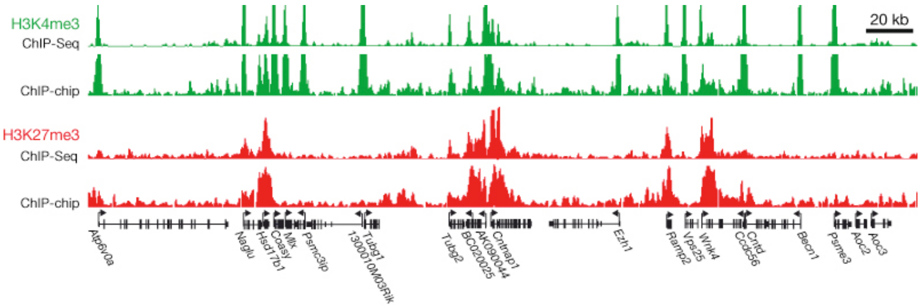
Sequencing-based methods equivalent to Microarray-based methods

Nature. 2007 Aug 2;448(7153):553-60

ARTICLES

Genome-wide maps of chromatin state in pluripotent and lineage-committed cells

Tarjei S. Mikkelsen[1,2], Manching Ku[1,4], David B. Jaffe[1], Biju Issac[1,4], Erez Lieberman[1,2], Georgia Giannoukos[1], Pablo Alvarez[1], William Brockman[1], Tae-Kyung Kim[5], Richard P. Koche[1,2,4], William Lee[1], Eric Mendenhall[1,4], Aisling O'Donovan[4], Aviva Presser[1], Carsten Russ[1], Xiaohui Xie[1], Alexander Meissner[3], Marius Wernig[3], Rudolf Jaenisch[3], Chad Nusbaum[1], Eric S. Lander[1,3]* & Bradley E. Bernstein[1,4,6]*

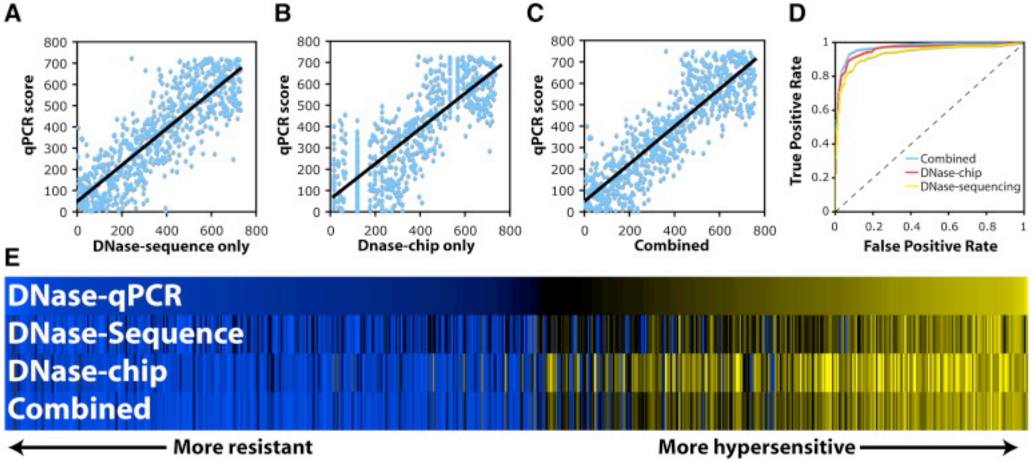


Cell. (2008) Jan 25;132(2):311-22.

High-Resolution Mapping and Characterization of Open Chromatin across the Genome

Alan P. Boyle,[1] Sean Davis,[3] Hennady P. Shulha,[2] Paul Meltzer,[3] Elliott H. Margulies,[4] Zhiping Weng,[2] Terrence S. Furey,[1,*] and Gregory E. Crawford[1,*]
[1]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA
[2]Biomedical Engineering Department, Boston University, Boston, MA 02215, USA
[3]Center for Cancer Research, National Cancer Institute
[4]National Human Genome Research Institute
National Institutes of Health, Bethesda, MD 20892, USA
*Correspondence: terry.furey@duke.edu (T.S.F.), greg.crawford@duke.edu (G.E.C.)

DOI 10.1016/j.cell.2007.12.014



Future Horizons

Applied Biosystems

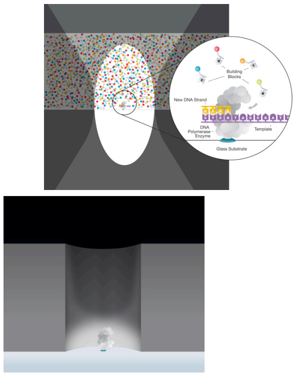
SOLiD

Ligation-based extension

Helicos BioSciences Corporation

HeliScope

PACIFIC BIOSCIENCES



SMRT Technology

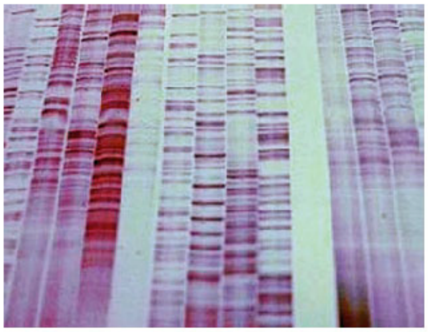
True Single Molecule Sequencing

Nature Methods, January 2008 Issue

METHOD OF THE YEAR | SPECIAL FEATURE

The year of sequencing

In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.



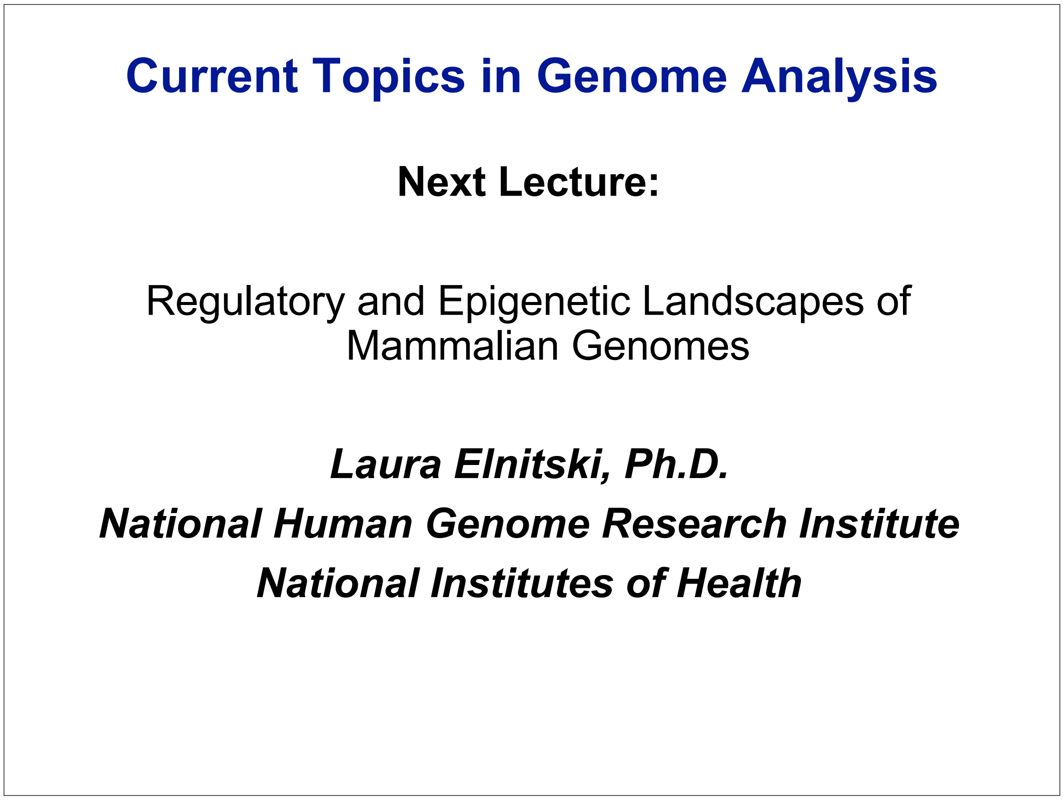
Sanger Sequencing becomes the 'old' generation

Primer: Sequencing—the next generation

Different sequencing technologies, at a glance.

Nicole Rusk and Veronique Kiermer

Good overview of latest-generation sequencing technologies currently available



Current Topics in Genome Analysis

Next Lecture:

Regulatory and Epigenetic Landscapes of
Mammalian Genomes

Laura Elnitski, Ph.D.

National Human Genome Research Institute

National Institutes of Health