Washington University in St. Louis
School of Medicine

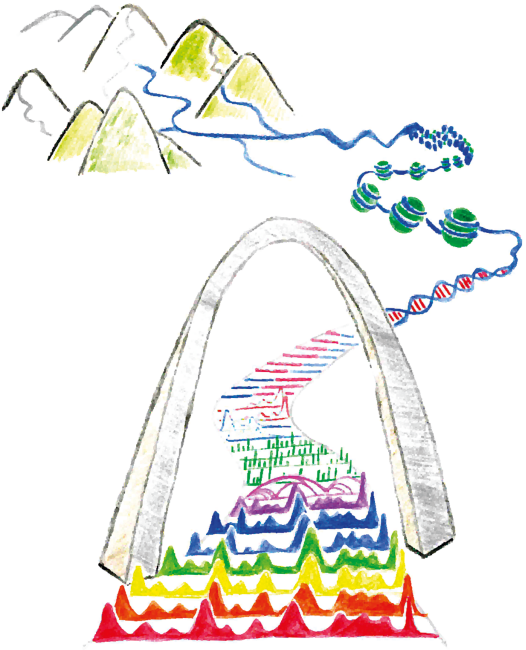
Department of Genetics

# 系统生物学与生物信息学

## 海外学者短期讲学系列课程

# Current Topics in Epigenomics

## 表观基因组学前沿



Ting Wang
Department of Genetics
Center for Genome Sciences and Systems Biology
Washington University School of Medicine

Tsinghua University
April 15-27

# Overview

- Lecturer:  Ting Wang, Ph.D.
  - http://wang.wustl.edu
  - twang@genetics.wust.edu
- Co-Lecturer: Jin Gu, Ph.D.
- TA: Aidi Tan
- Class website –
  - http://bioinfo.au.tsinghua.edu.cn/course/twang2016/

# Our lab

## Lab members

Xiaoyun Xing  Daofeng Li  Junchen Gu  Vasavi Sundaram
Hyung Joo Lee  Rebecca Lowdon  Renee Sears  Jennifer Flynn
Deepak Purushotham  Ye Hu  Nicole Rockweiler  Josh Jang
Mayank Choudhary  Mikayla Choi  Erica Pehrsson  Yiran Hou
Nakul Shah  Eileen Chen
Xin Zhou  Bo Zhang  Mingchao Xie  Michael Stevens
GiNell Elliott  Jia Zhou  Jing Li


## Funding:

NIH U01ES017154,
R01HG007354,
R01HG007175,
R01ES024992,
U01CA200060,
U24ES026699
American Cancer Society
RSG-14-049-01-DMC
March of Dimes
Mallinckrodt Foundation

# Outline

**http://bioinfo.au.tsinghua.edu.cn/course/twang2016/**

Topic 1 (day 1)
- **Introduction (Gene, Genome, and Epigenome)**

Topic 2 (day 1, 2, 3)
- **Epigenetic mechanisms**

Topic 3 (day 3, 4)
- **Genome and epigenome architecture**

Topic 4 (day 4)
- **The dark side of the epigenome: transposable elements**

Topic 5 (day 5)
- **Epigenome evolution**
- Group discussion

Topic 6 (day 6)
- **Big data in biology**
- Workshop and group discussion

**Biology**
**Technology**
**Informatics**
**Problem solving**

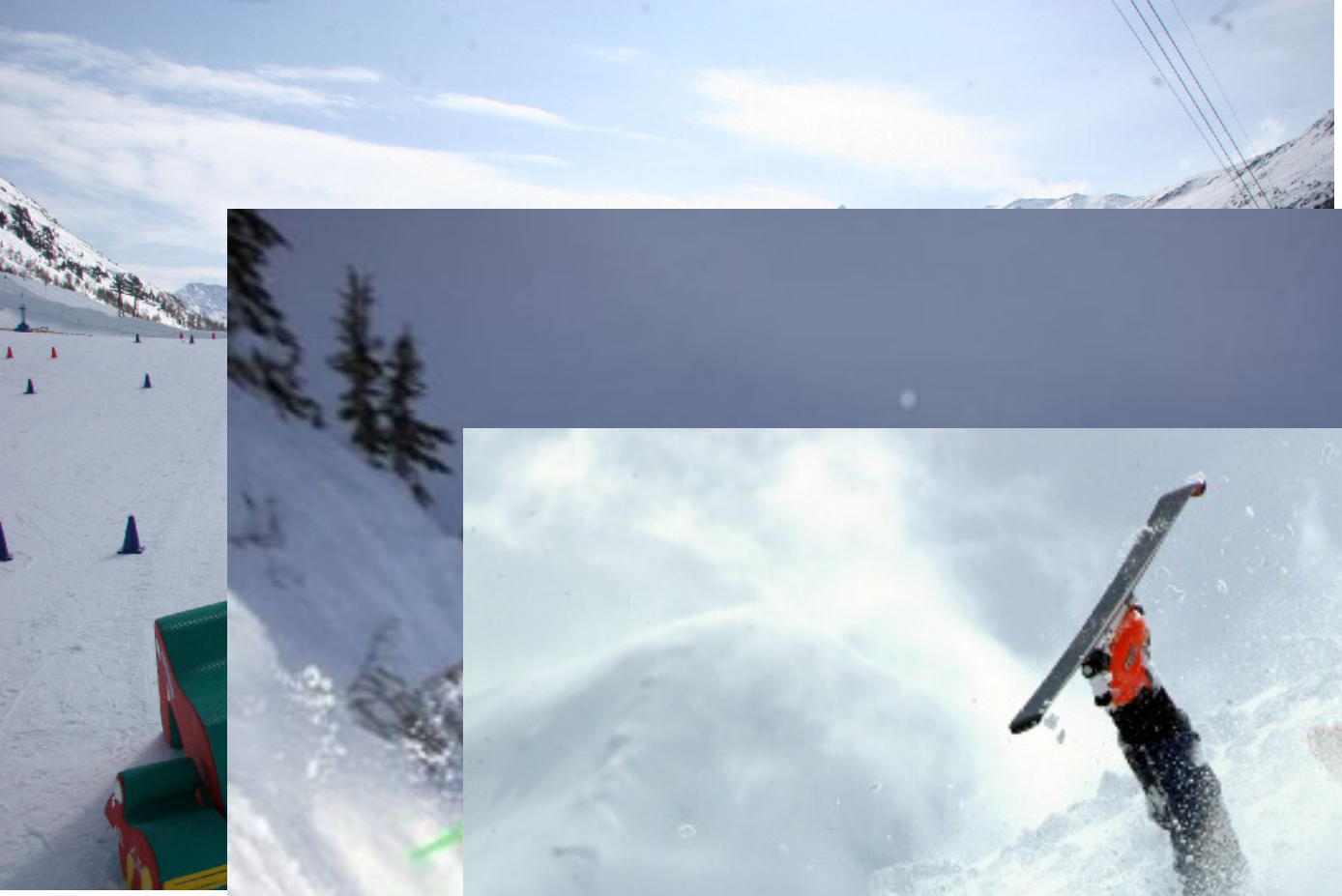
# Topics for discussion
# Challenges and opportunities

- Epigenome evolution

- Big bio data

# Outline of the day

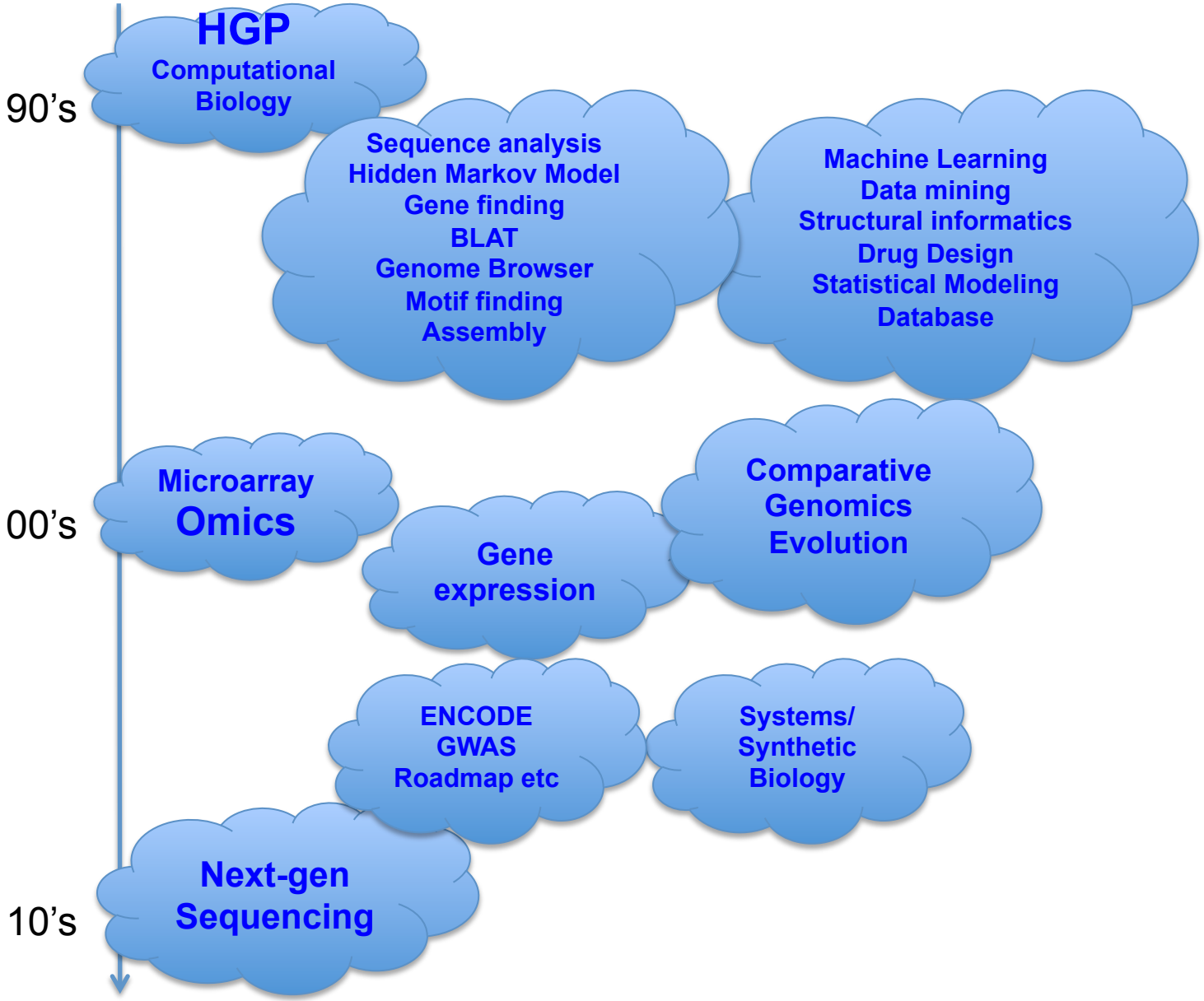
- Outline of the course
- What is genomics and epigenomics?
- A little history
- The simple principles of genomics
- Being quantitative
- From a student to an investigator



# History of Genomics and Epigenomics

| | |
|---|---|
| 1865 | Gregor Mendel: founding of genetics |
| 1953 | Watson and Crick: double helix model for DNA |
| 1955 | Sanger: first protein sequence, bovine insulin |
| 1970 | Needleman-Wunsch algorithm for sequence alignment |
| 1977 | Sanger: DNA sequencing |
| 1978 | The term “bioinformatics” appeared for the first time |
| 1980 | The first complete gene sequence (Bacteriophage FX174), 5386 bp |
| 1981 | Smith-Waterman algorithm for sequence alignment |
| 1981 | IBM: first Personal Computer |
| 1983 | Kary Mullis: PCR |
| 1986 | The term "Genomics" appeared for the first time: name of a journal |
| 1986 | The SWISS-PROT database is |
| 1987 | Perl (Practical Extraction Report Language) is released by Larry Wall. |
| 1990 | BLAST is published |
| 1995 | The *Haemophilus influenzea* genome (1.8 Mb) is sequenced |
| 1996 | Affymetrix produces the first commercial DNA chips |
| 2001 | A draft of the human genome (3,000 Mbp) is published |

# History of Genomics and Epigenomics



# Genome, genetics, and genomics

- What is a genome?
  - **The genetic material of an organism.**
  - **A genome contains genes, regulatory elements, and other mysterious stuff.**

- What is genetics?
  - **The study of genes and their roles in inheritance.**

- What is genomics
  - **The study of all of a person's genes (the genome), including interactions of those genes with each other and with the person's environment**

# What about genomes

- **Characterize the genome**
  - How big
  - How many genes
  - How are they organized
- **Annotate the genome**
  - What, where, and how
- **Modern genomics: “ChIPer” vs “Mapper”**
  - Direct measurement
  - Inference
  - Comparison
  - Evolution
- **From genome to molecular mechanisms to diseases**
  - Genomes/epigenomes of diseased cells
- **What do you want to learn from this class?**
  - Being quantitative
  - Concept/philosophy
  - Techniques/problem solving skills
  - Do not forget genetics!!!

# Motivation slides

# Genomes sequenced

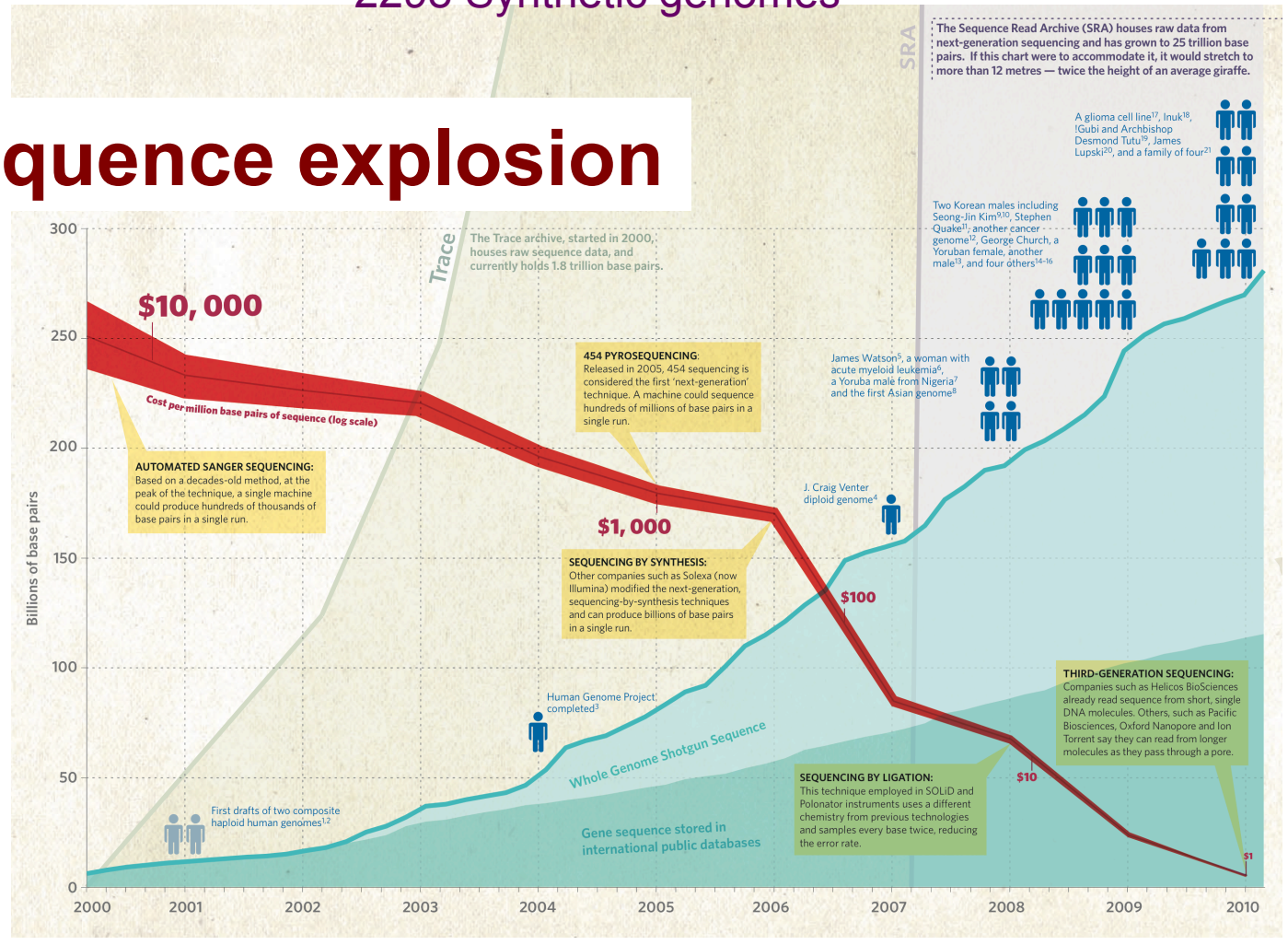
http://www.genomesonline.org/, (as of November 2015)

49559 Bacteria, 1136 Archaea, 11122 Eukarya, 4473 viruses

18385 metagenome samples

2298 Synthetic genomes

# The sequence explosion



# The Human Genome Project

## The human genome completed! – 2001


June 26, 2000
President Clinton, with Craig Venter and Francis Collins, announces completion of "the first survey of the entire human genome."

February 15, 2001

## The human genome completed, again! – 2010


April 1, 2010

**10 years after draft sequence: What have we learned?**

- Genome sequencing
- Functional elements
- Evolution of genome
- Basis of diseases
- Human history
- Computational biology

“ If I was a senior in college or a first-year graduate student trying to figure out what area to work in, I would be a computational biologist. ”

-- During AAAS Meeting Jan 2010

---

COLLINS: .... Computational biologists are having a really good time and it’s going to get better.

ROSE: Their day is coming?

COLLINS: Their day is here, but it’s going to be even more here in a few years.

COLLINS: They’re going to be the breakthrough artists ....

-- Mar 15, 2010, Charlie Rose Show

Francis Collins
NIH Director



# Program Initiatives

$200 million

## I. Facilitating Broad Use of Biomedical Big Data

- New Policies to Encourage Data & Software Sharing
- Catalog of Research Datasets to Facilitate Data Location & Citation
- Frameworks for the development of community-based standards
- Enabling Research Use of Clinical Data

## II. Developing and Disseminating Analysis Methods and Software

- Software to Meet Needs of the Biomedical Research Community, both analytic software and management/processing software
- The creation of a Catalog of NIH-funded Software
- Facilitating Data Analysis: Access to Large-scale Computing

## III. Enhancing Training for Biomedical Big Data

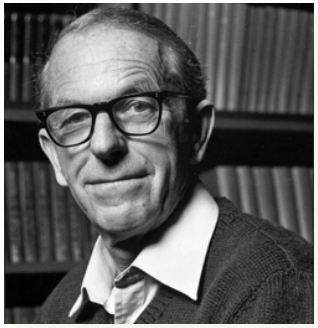
- Increase the Number of Computationally Skilled Biomedical Trainees
- Strengthen the Quantitative Skills of All Biomedical Researchers
- Enhance NIH Review and Program Oversight

## IV. Establishing Centers of Excellence for Biomedical Big Data

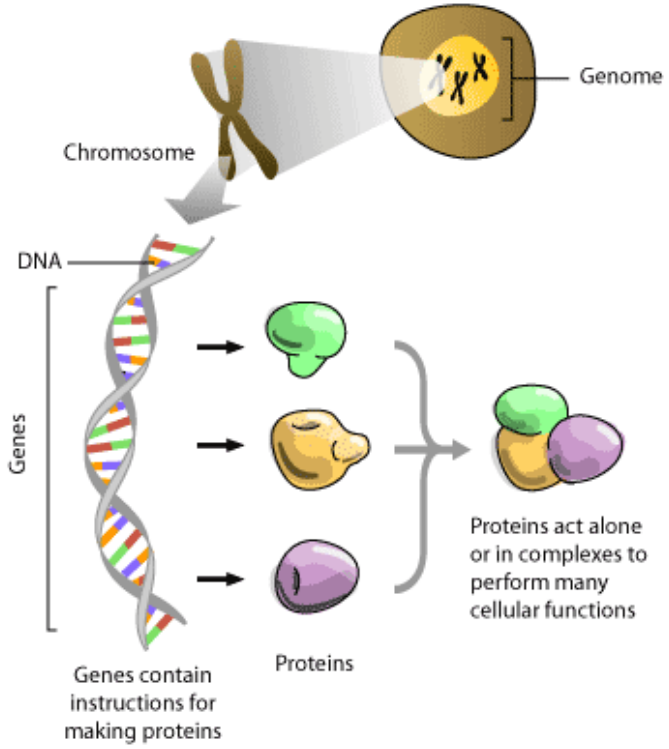
Advance the science of Big Data in the context of biomedical and behavioral research, and to create innovative new approaches, methods, software, and tools.

# The Human genome: the "blueprint" of our body

James Watson
Francis Crick



Fred Sanger




GTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCT
CCGAACAATAAAGATTCTACAATACTAGCTTTTATGGTTATG
AAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAA
ATTAACGAATCAAATTAACAACCATAGGATGATAATGCGATT
AGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG
ATGATTTTTGATCTATTAACAGATATATAAATGGAAAAGCTG
CATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTA
TTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAATT

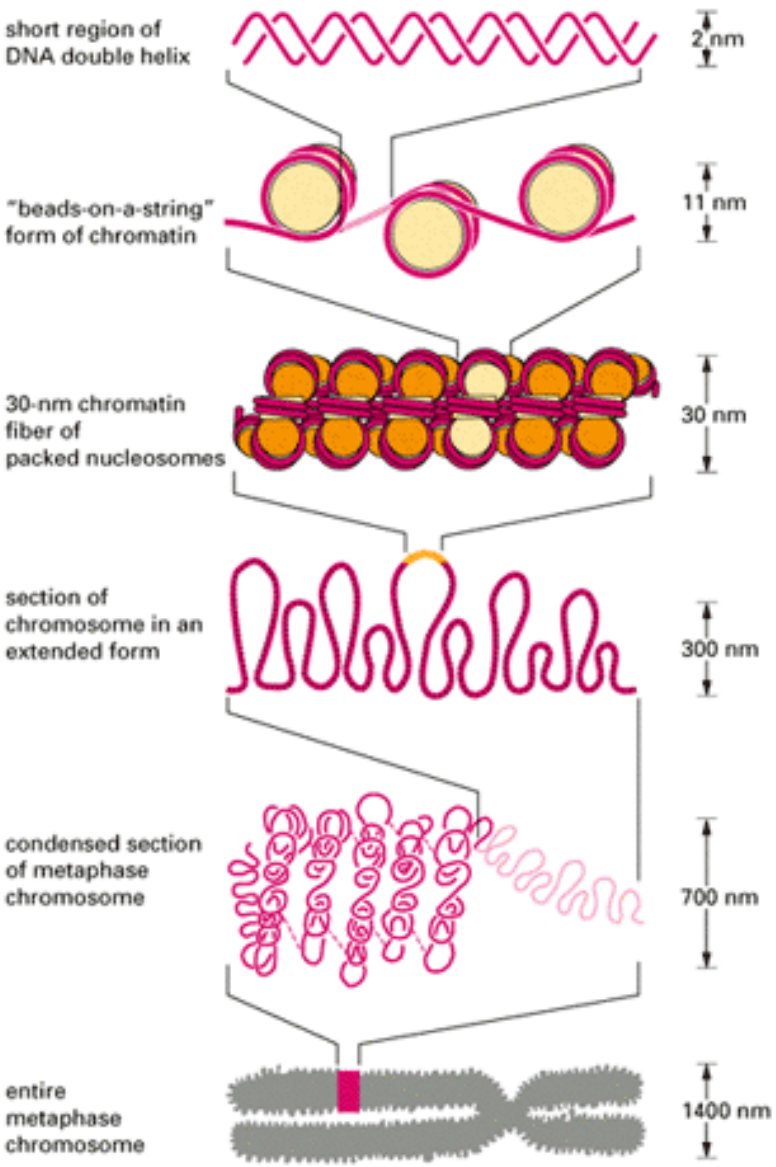
$10^{13}$ different cells in an adult human

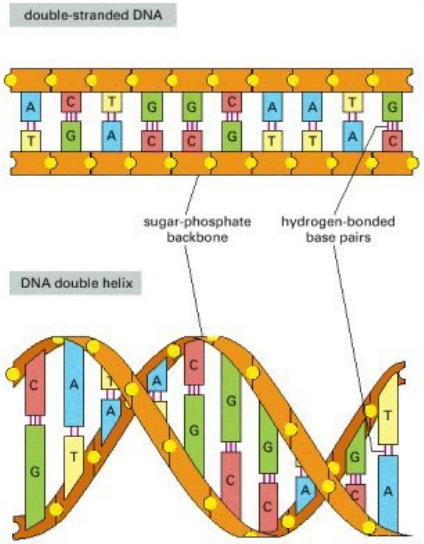
The cell is the basic unit of life

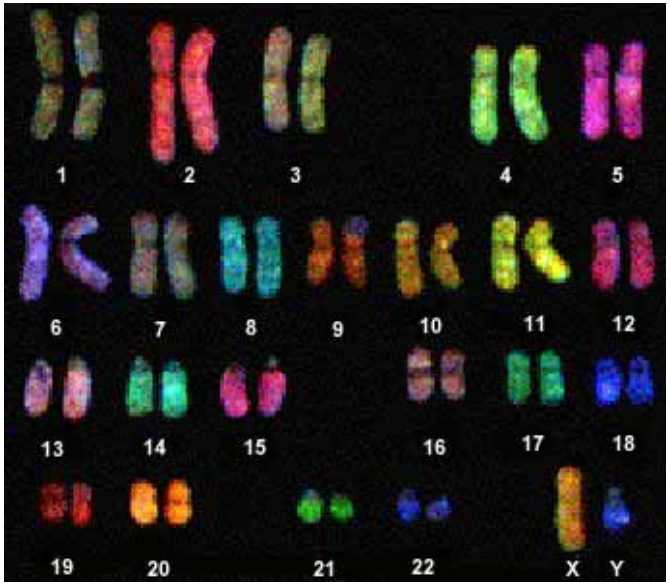
DNA = linear molecule inside the cell that carries instructions needed throughout the cell's life ~ long string(s) over a small alphabet

Alphabet of four (nucleotides/bases) {A,C,G,T}

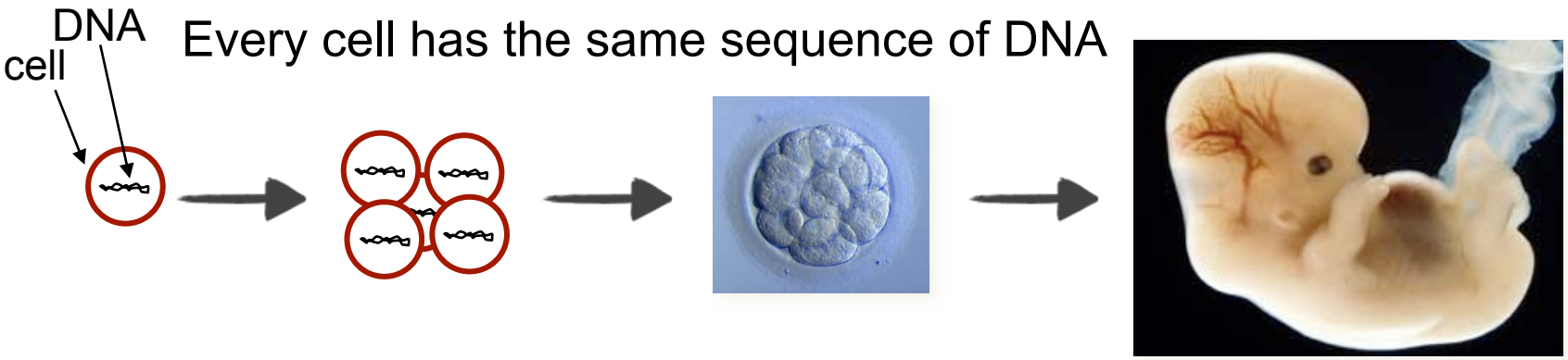
# DNA, Chromosome, and Genome



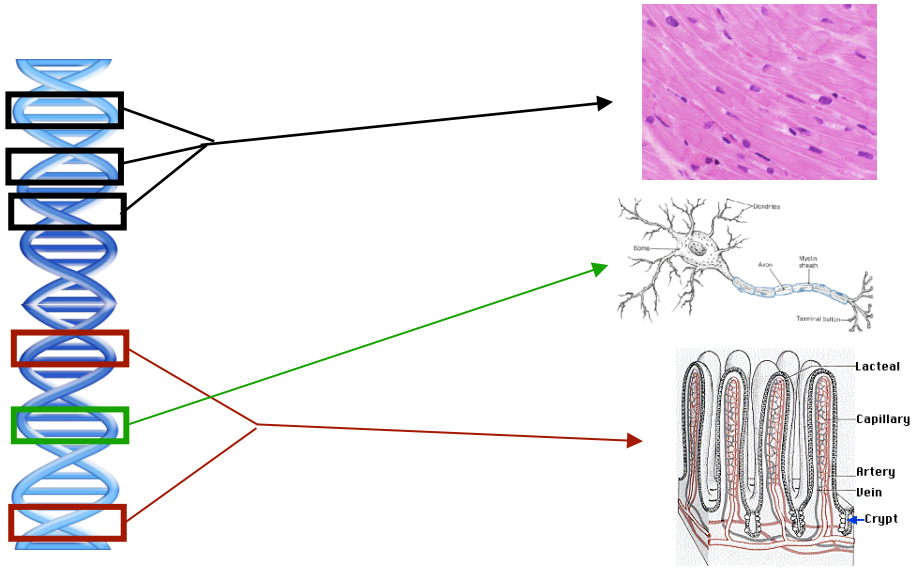



# Building an Organism




Subsets of the DNA sequence determine the identity and function of different cells

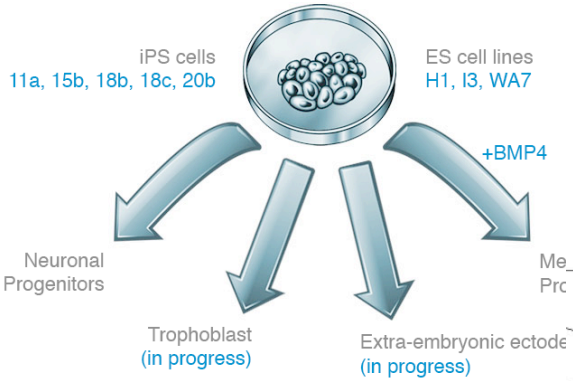


# What makes us different?

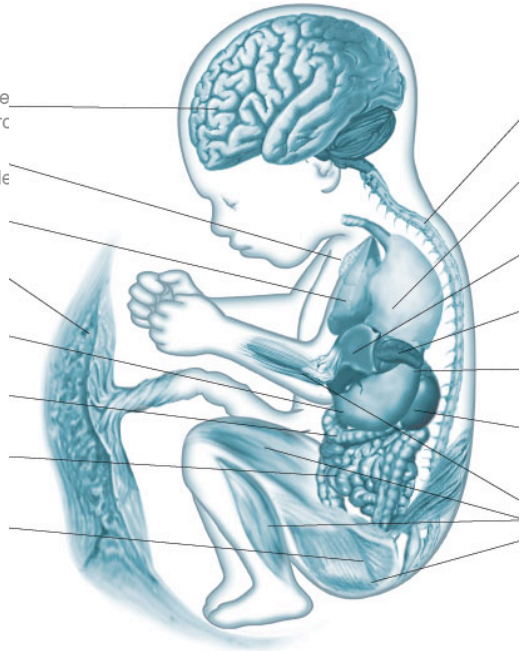
Differences between individuals?

Differences between species?

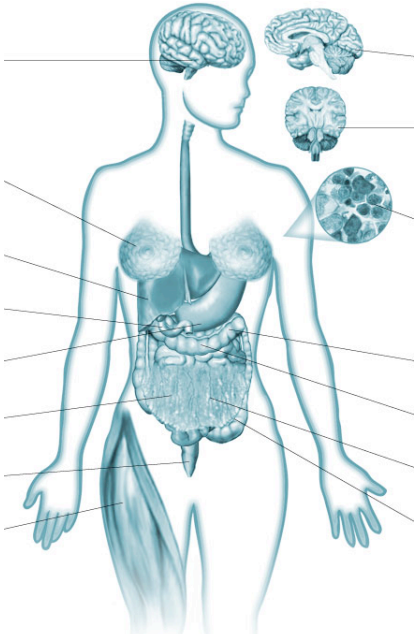
# One genome, thousands of epigenomes




**Embryonic stem cells**



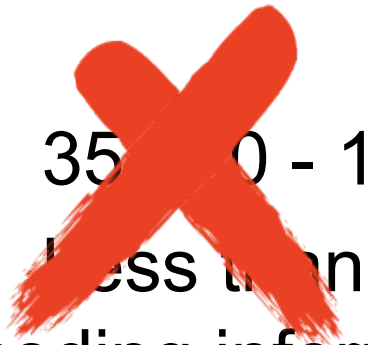
**Fetal tissues**



**Adult cells and tissues**

# Understanding the genome

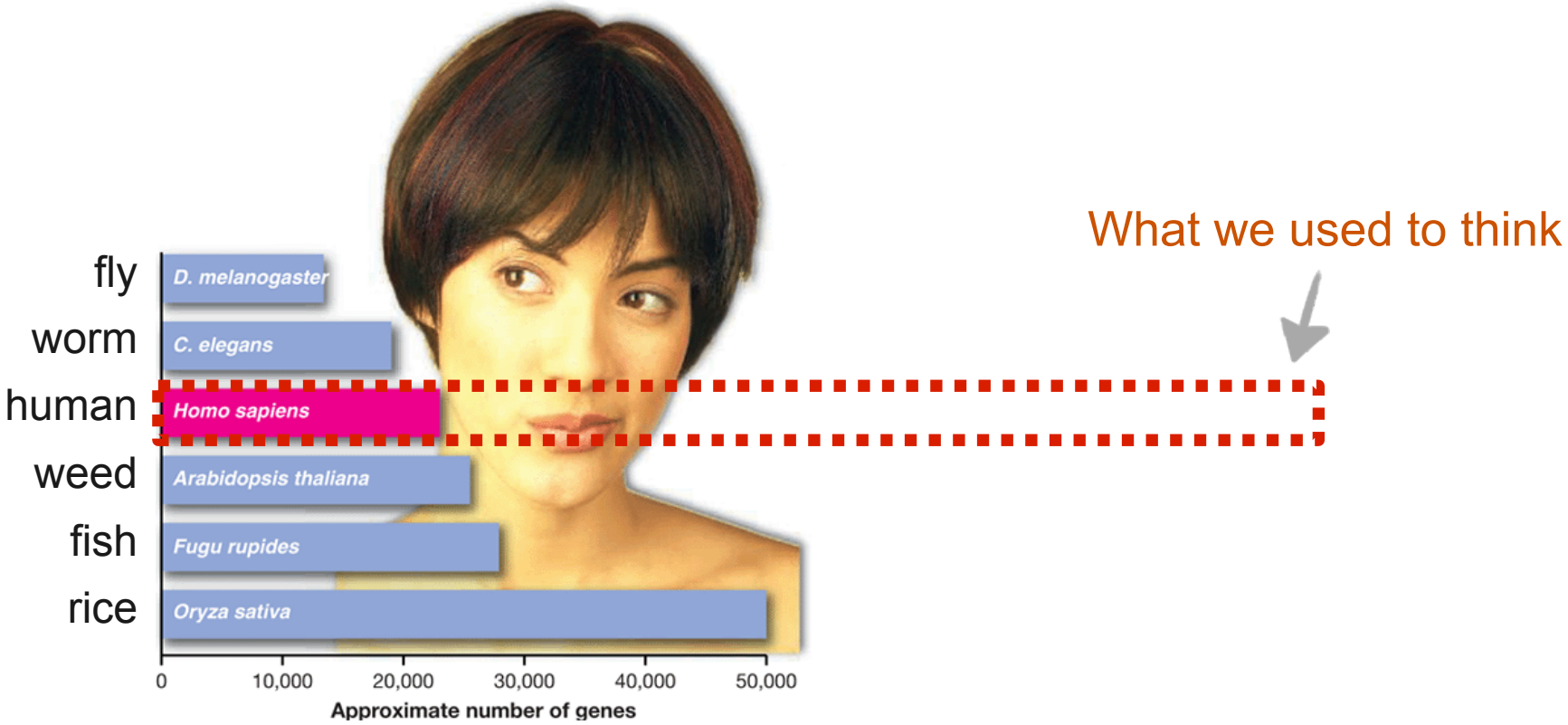
- View from 2000
  - Protein-coding genes 35,000 - 120,000
  - Regulatory sequence Less than protein-coding information
  - Transposons Junk DNA



All these are WRONG!

- We now know ...

# How many genes do we have?




Science 2005

Gene numbers do not correlate with organism complexity.
Many gene families are surprisingly old.

# Complexity, Genome Size and the C-value Paradox

| Organism | Genome Size (MB) |
|---|---|
| Amoeba | 670,000 |
| Fern | 160,000 |
| Salamander | 81,300 |
| Onion | 18,000 |
| Paramecium | 8,600 |
| Toad | 6,900 |
| Barley | 5,000 |
| Chimp | 3,600 |
| Gorilla | 3,500 |
| **Human** | **3,500** |
| Mouse | 3,400 |
| Dog | 3,300 |
| Pig | 3,100 |
| Rat | 3,000 |
| Boa Constrictor | 2,100 |
| Zebrafish | 1,900 |
| Chicken | 1,200 |
| Fruit fly | 180 |
| C. elegans | 100 |
| Plasmodium falciparum | 25 |
| Yeast, Fission | 14 |
| Yeast, Baker's | 12 |
| Escherichia coli | 4.6 |
| Bacillus subtilis | 4.2 |
| H. influenzae | 1.8 |
| Mycoplasma genitalium | 0.60 |

www.genomesize.com

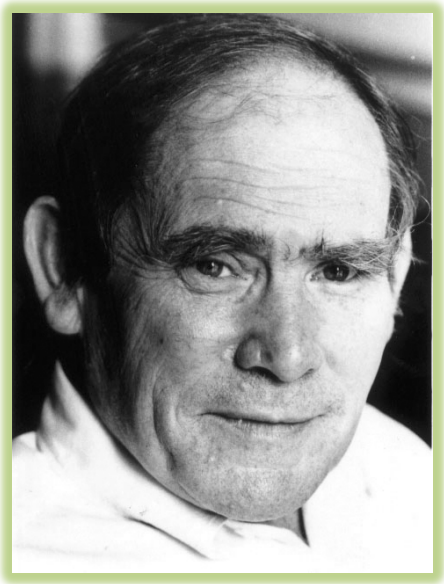
**C-value:** the amount of DNA contained within a haploid nucleus (e.g. a gamete) or one half the amount in a diploid somatic cell of a eukaryotic organism, expressed in picograms (1pg = $10^{-12}$ g).

# #protein-coding genes ≠cellular complexity

19,000

14,000

~20-25,000



~20-25,000

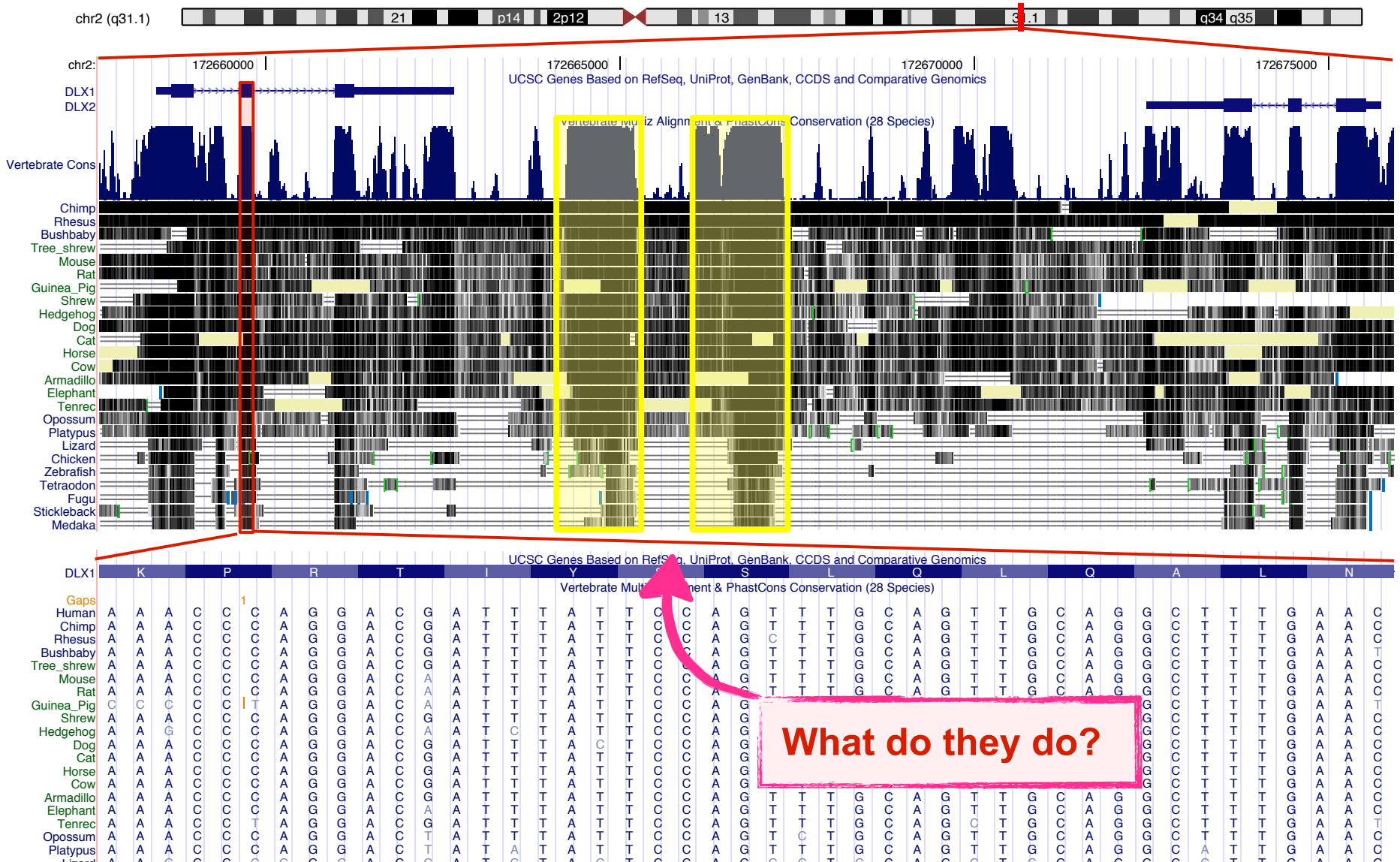
6,000

# Complexity and Organism Specific Genes

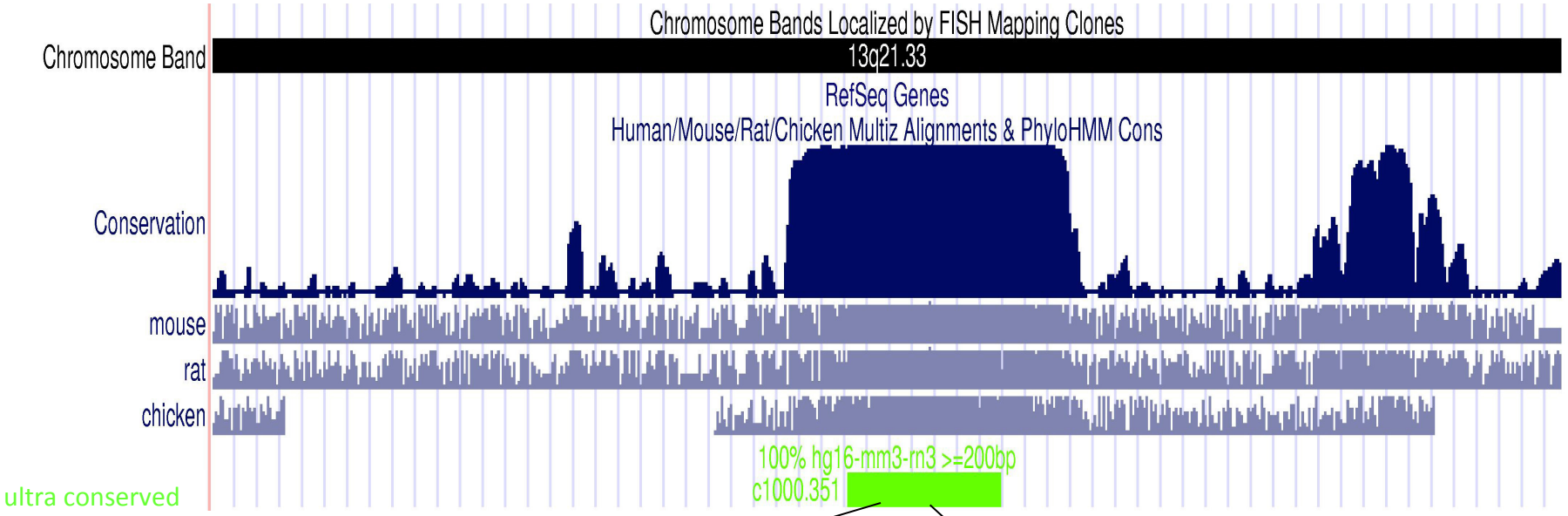
- Only 14 out of 731 genes on mouse chromosome 16 have no human homolog (Celera)
- Many genes specific to the mouse are olfactory receptors, also some differences in immunity and reproduction

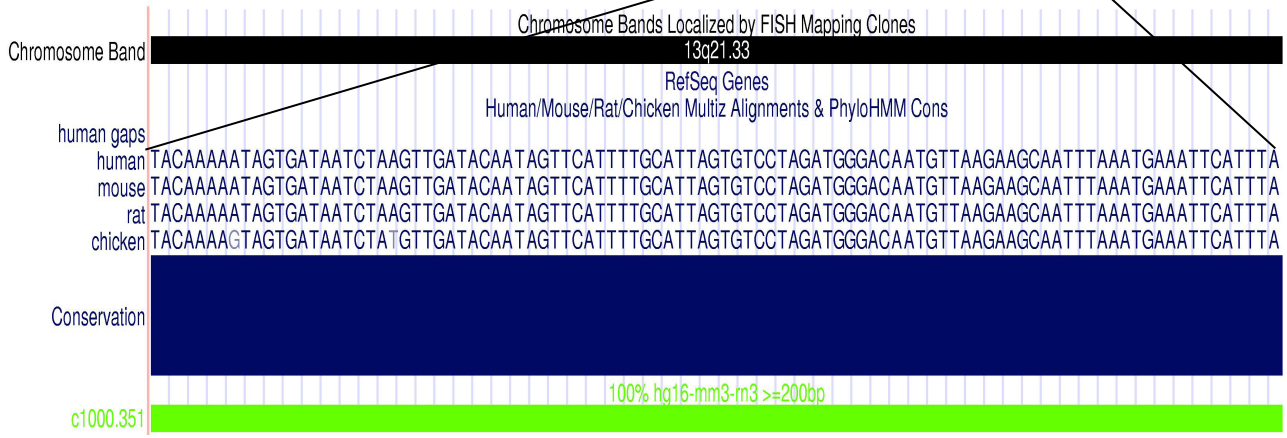
# Most functional information is non-coding

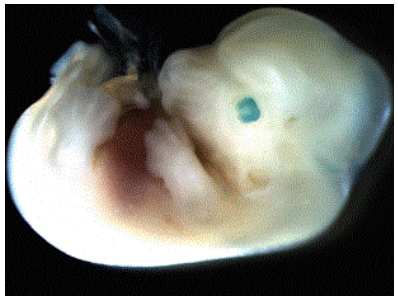
- 5% highly conserved, but only 1.5% encodes proteins



# Ultra conserved elements

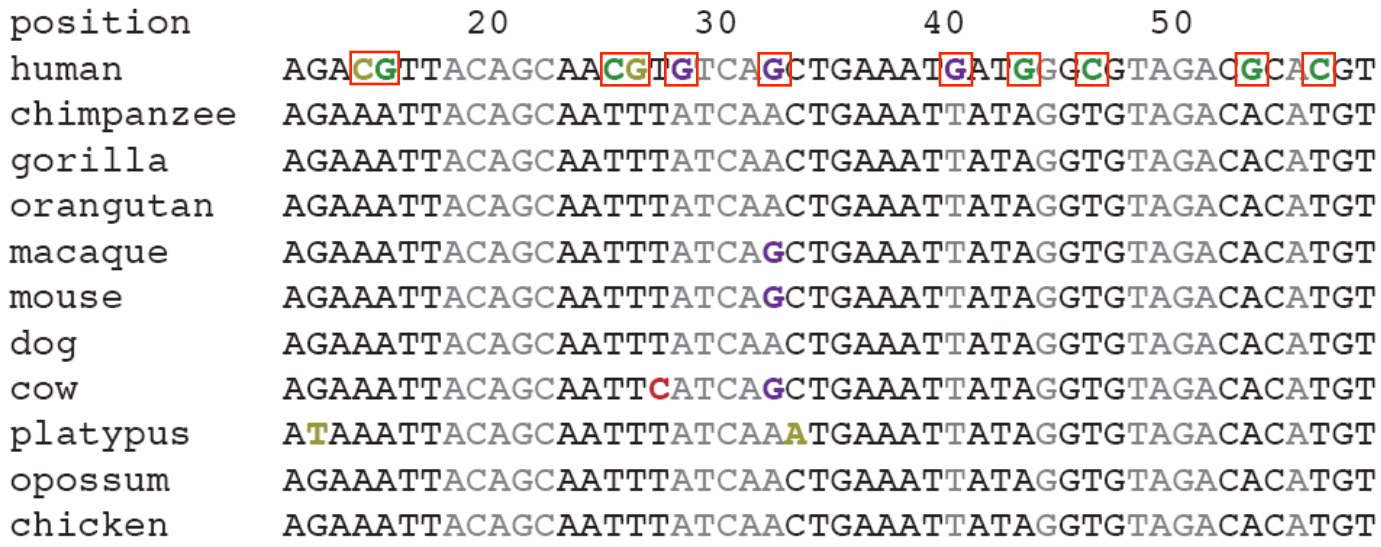


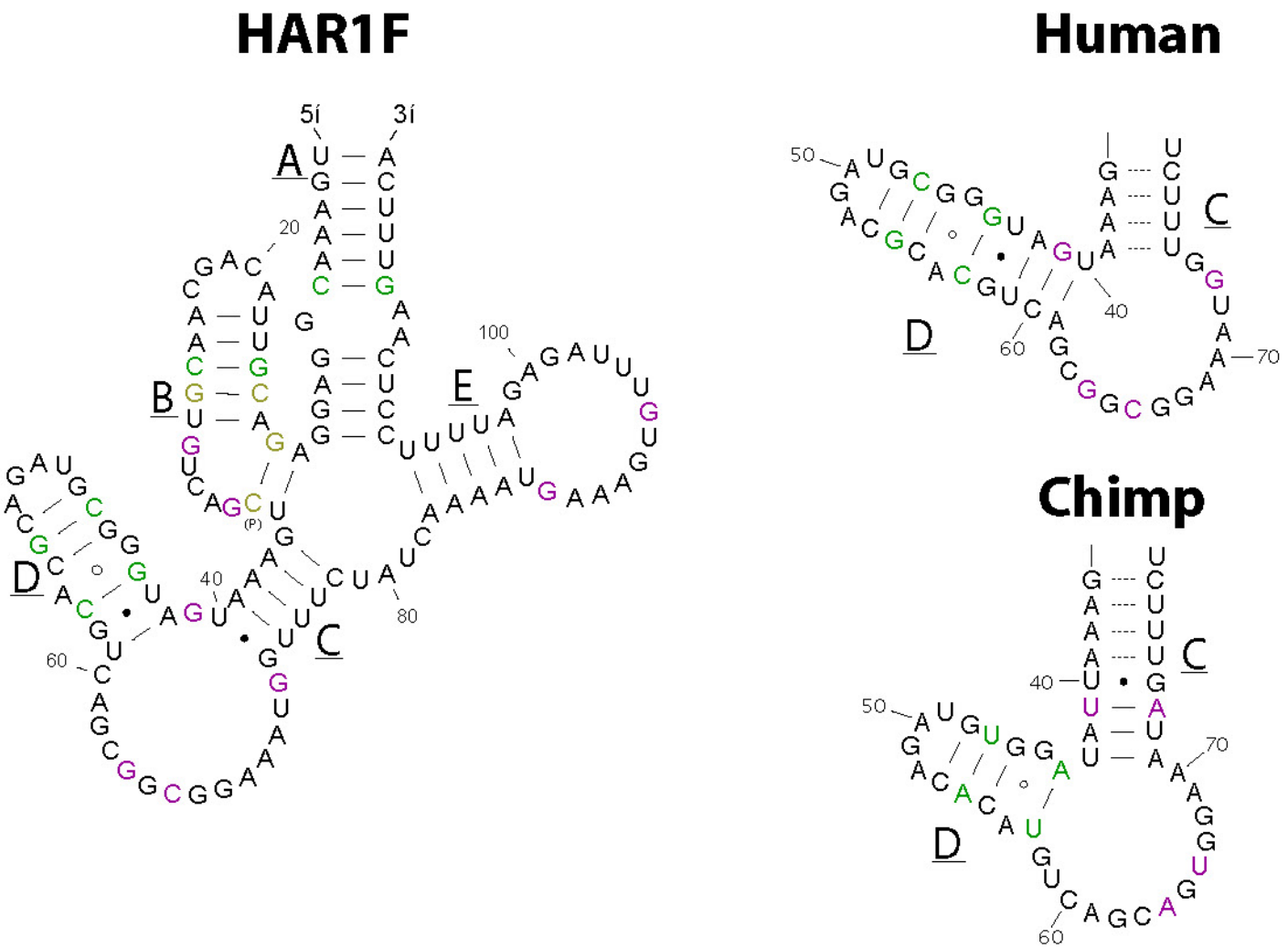
e.d 12.5

# HARs: Human accelerated regions

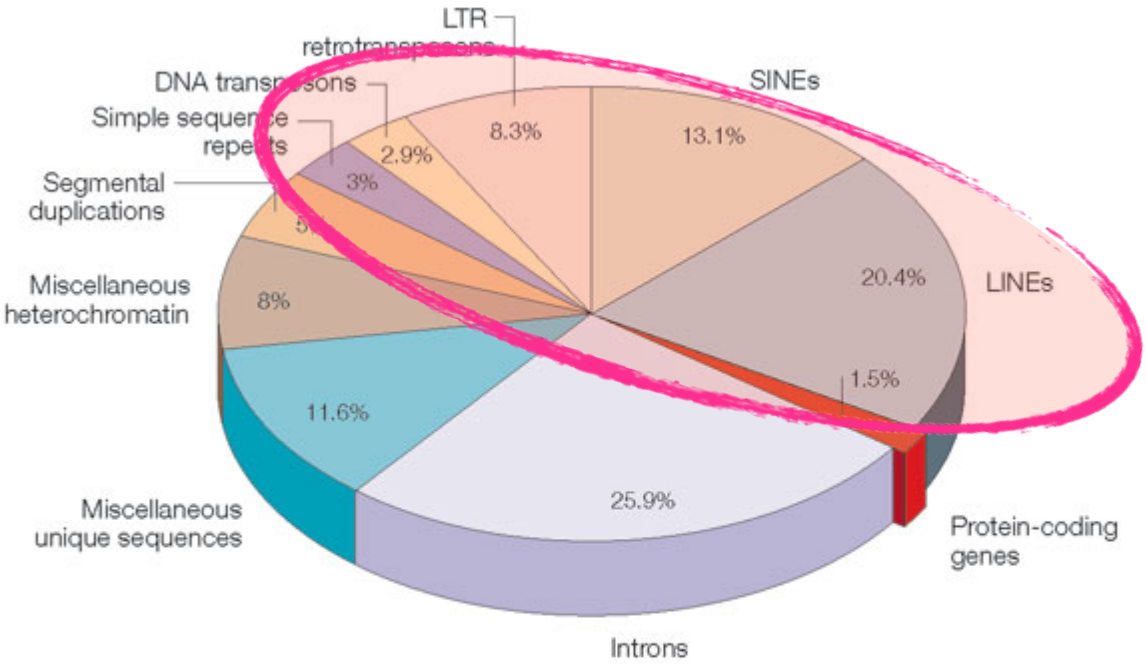



- 118 bp segment with 18 changes between the human and chimp sequences
- Expect less than 1

# Human HAR1F differs from the ancestral RNA structure



# Main components in the Human genome

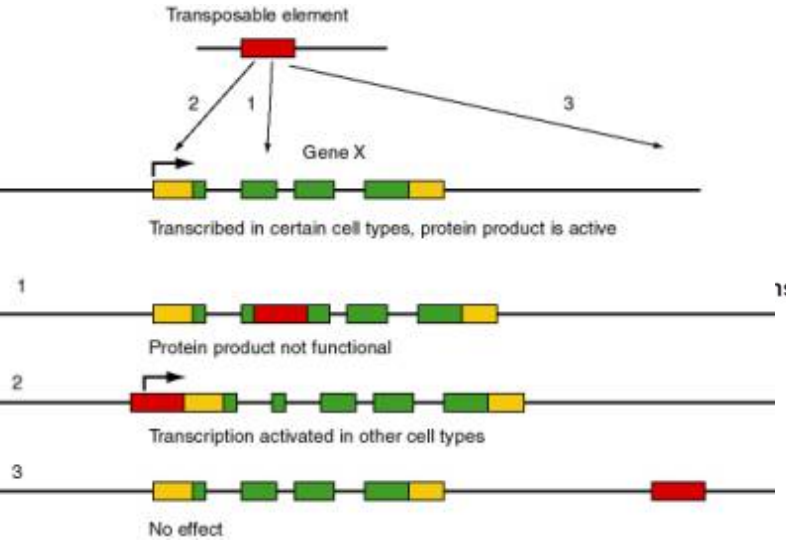


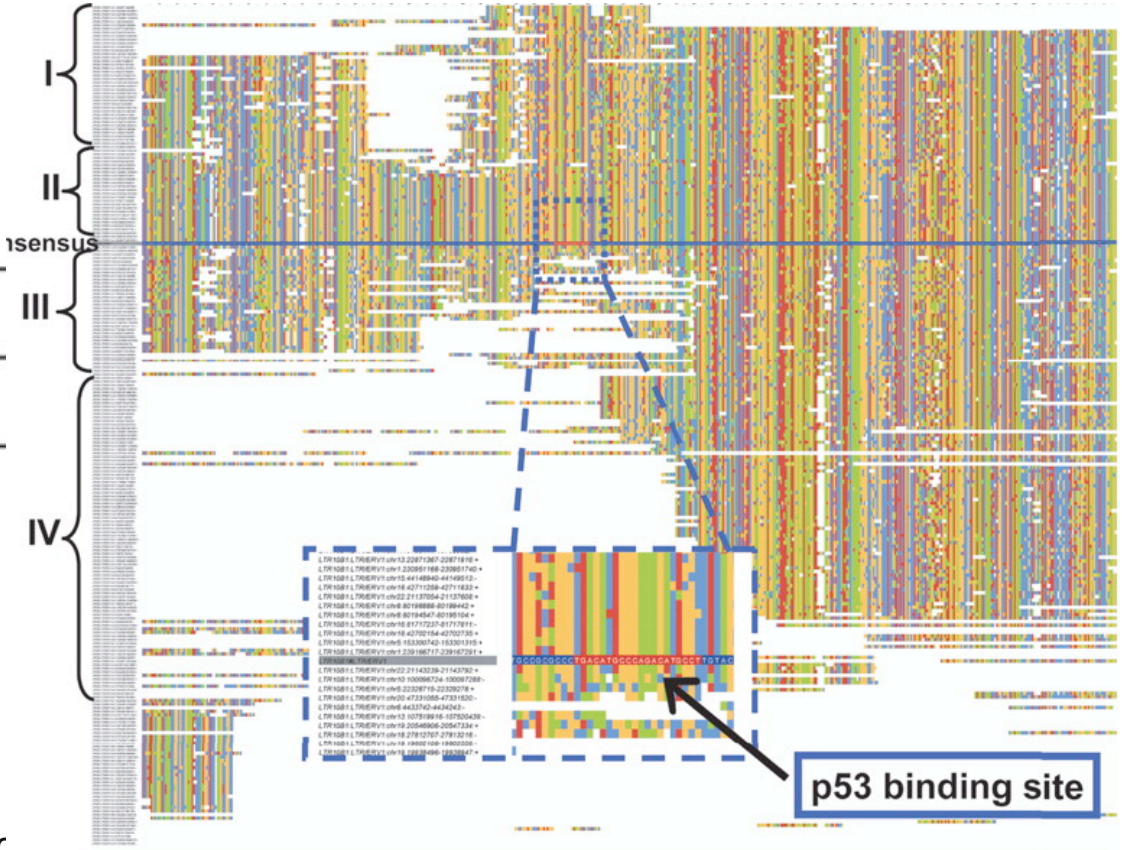



Barbara McClintock

Only 1.5% of the human genome are protein-coding regions
Transposable elements make up almost half of the human genome

# Transposable Elements (TEs)




TEs can shape transcriptional network

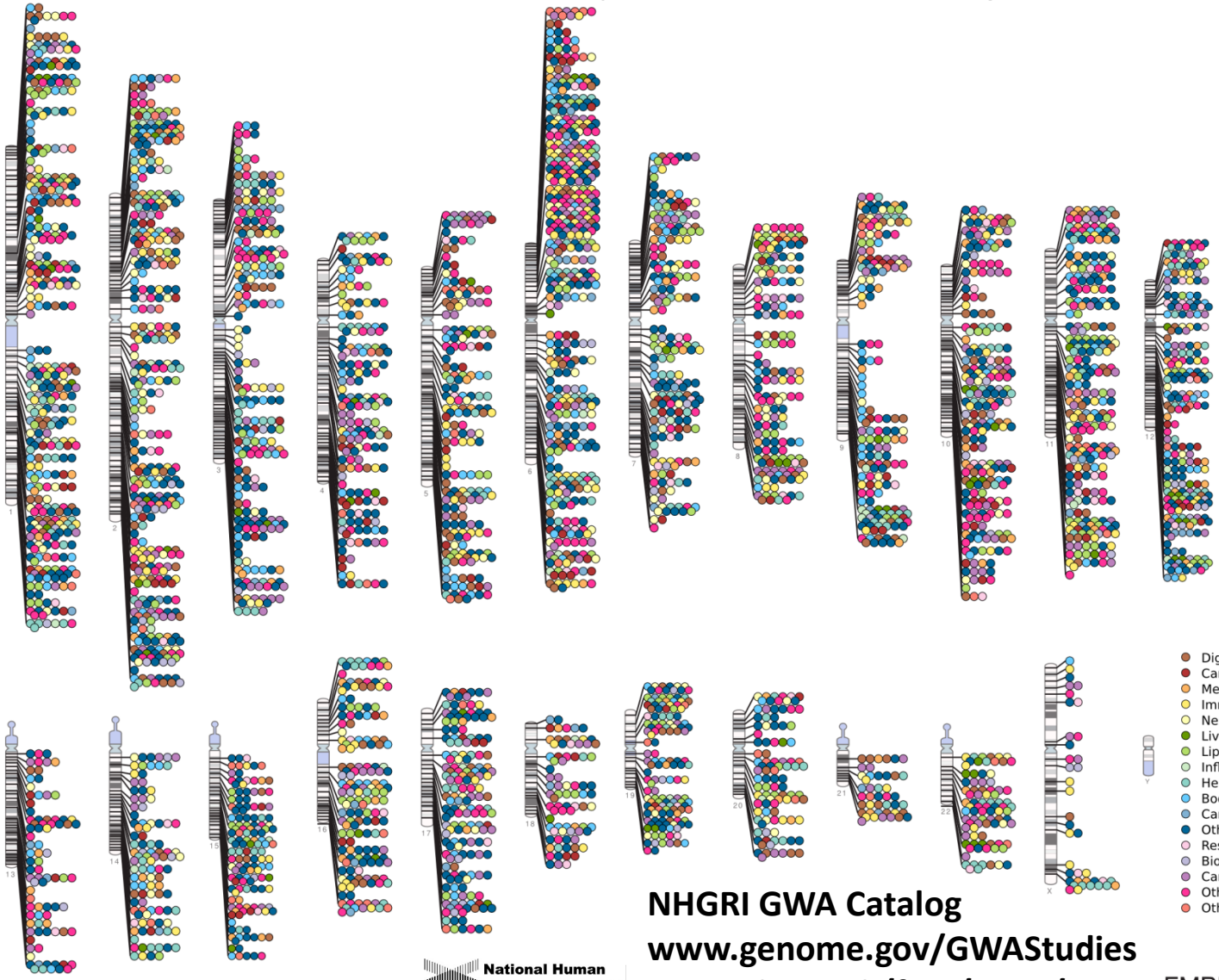
The LTR10 and MER61 families are particularly enriched for copies with a p53 site. These ERV families are primate-specific and transposed actively near the time when the New World and Old World monkey lineages split.

Evolutionary pattern of LTR10B1 genomic copies

Wang et al. PNAS 2007

# Understanding Human Disease

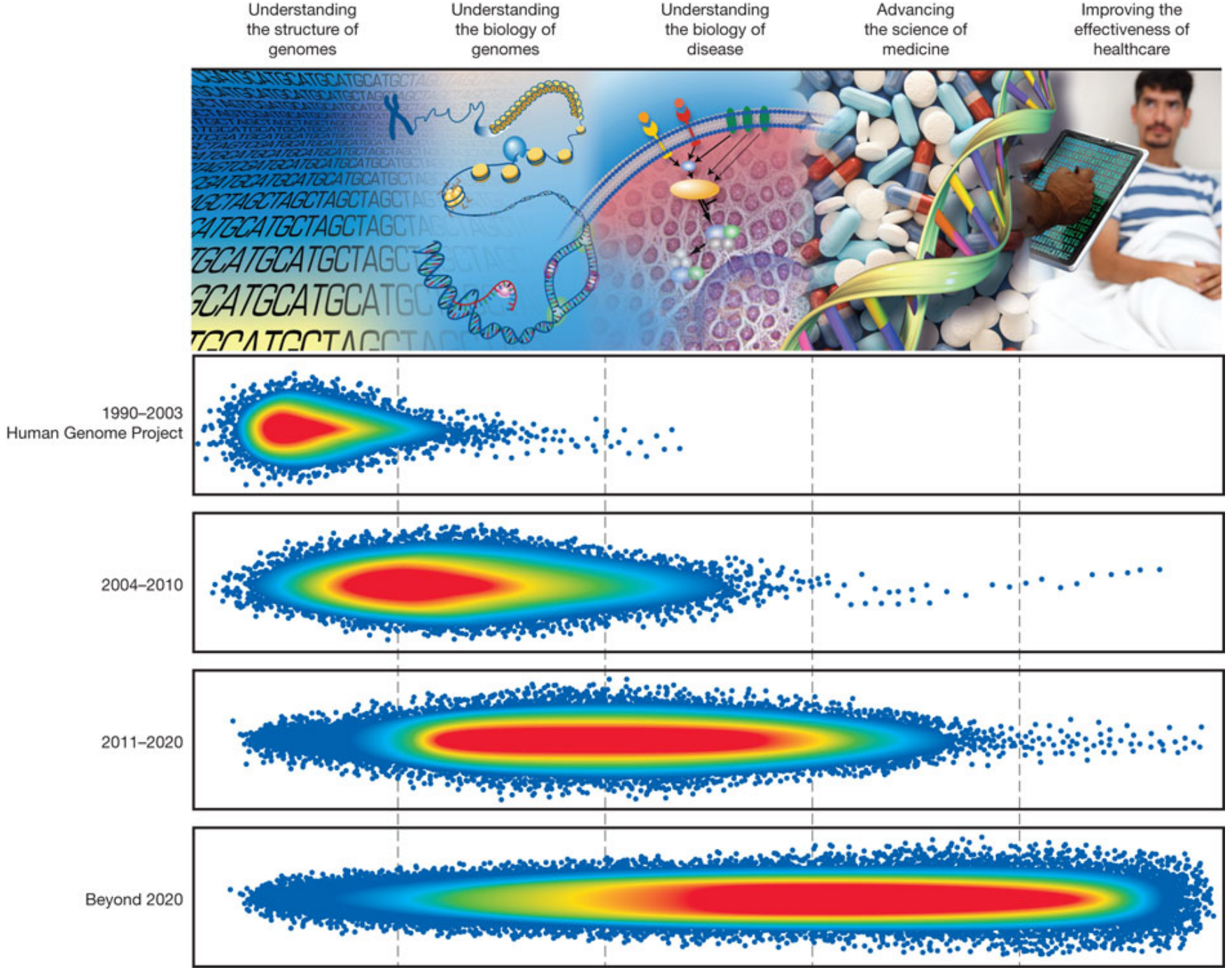
**Published Genome-Wide Associations through 12/2013**
**Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories**

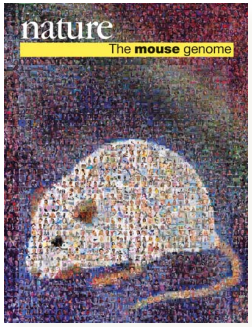



**NHGRI GWA Catalog**
**www.genome.gov/GWAStudies**
**www.ebi.ac.uk/fgpt/gwas/**

National Human Genome Research Institute

EMBL-EBI


Understanding the structure of genomes
Understanding the biology of genomes
Understanding the biology of disease
Advancing the science of medicine
Improving the effectiveness of healthcare
1990–2003 Human Genome Project
2004–2010
2011–2020
Beyond 2020











GENOME 10K©

Proposed in Nov 2009


ENCODE Project


HapMap

1000 Genomes

ROADMAP epigenomics PROJECT

Epigenome

THE CANCER GENOME ATLAS

TCGA

# Thinking Quantitatively

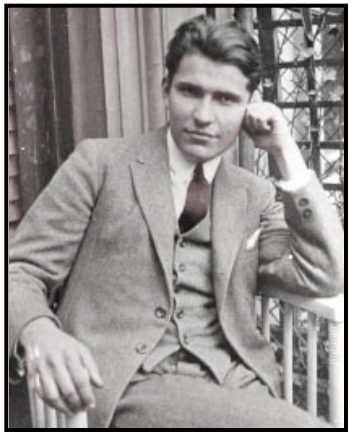
# Biology Is A Quantitative Science!!!

Gregor Mendel

1823-1884

1) Mendel's Laws
2) Chargaff's Rules

Erwin Chargaff



1929-1992

# Thinking Quantitatively

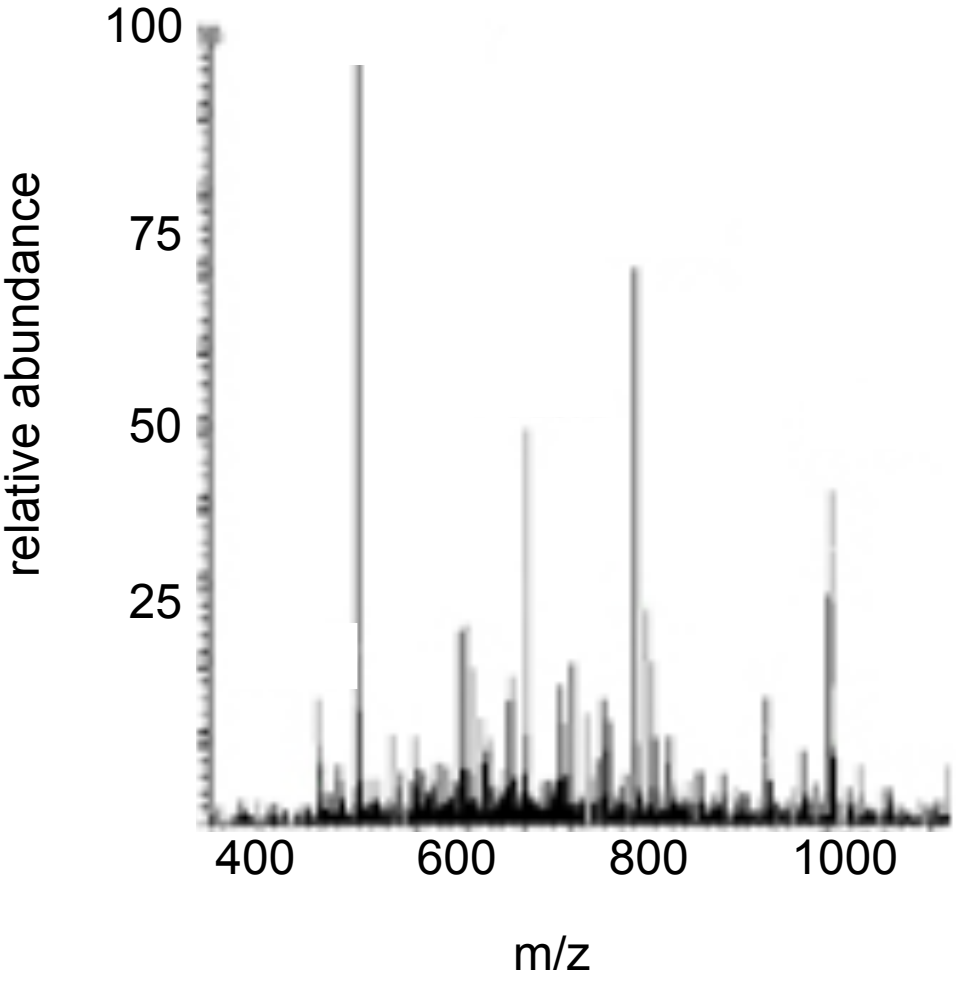
- Space
  - Be comprehensive
- Signal to Noise Ratio
  - Sensitivity, specificity, dynamic range
  - What is my background control?
- Distributions
  - Normal, Gaussian, Poisson, negative binomial, extreme value, hypergeometric, etc.
  - Discrete vs continuous
- The *P* value
- Statistics, Probability, Computation, and Informatics
- Don't forget genetics!!!

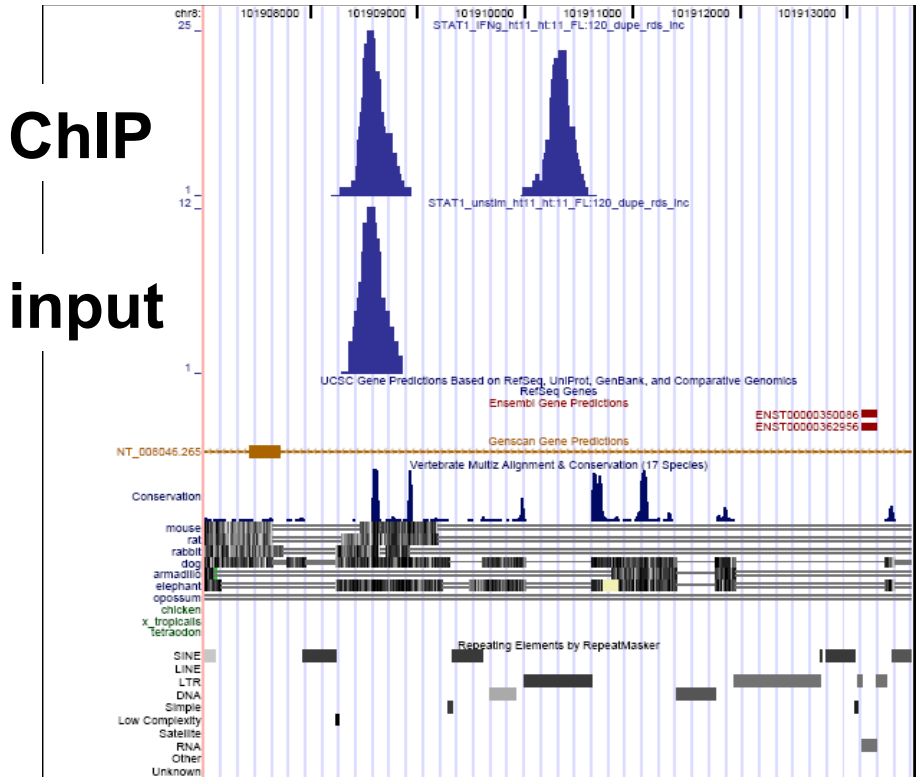
**Simple principle:**
**what is your expectation?**
**what is your observation?**

# Spaces: Be comprehensive

- Conditions: spatial, temporal, treatment – think about controlling for multiple variables
- Think globally – interaction between local features and global features (Placenta histone example)
- Be comprehensive about what assumptions are made – some we know, some we don't (genome assembly example)

# Signal to Noise




**Different sources of noise**

# Sensitivity, Specificity, and Dynamic Range

- **Sensitivity**
  - What is the smallest signal that can reliably be detected (signal to noise)?
  - True positive rate
- **Specificity**
  - How well can we discriminate between similar signals?
  - True negative rate
- **Dynamic Range**
  - What is the linear range of detection?
  - What is the range of natural variation?

# What is the reference human genome?

**Lander:** So the genes from which most of the work was done come from Buffalo, New York.

**Krulwich:** From Buffalo, New York?

**Lander:** Yes. **It's mostly a guy from Buffalo and a woman from Buffalo.** But that's because the laboratory that was making--

...

Eric Lander
NOVA interview, 2001

**Lander:** The laboratory that prepared the large DNA libraries that were used was a laboratory in Buffalo. And so they put an ad in the Buffalo newspapers, and they got random volunteers from Buffalo, and they got about 20 of them. They then erased all the labels and chose at random this sample and that sample and that sample. So nobody knows who they are. We don't have any links back to who they are, and that's deliberate.