Biomedical Informatics for Clinical Decision Support: A Vision for the 21st Century

Aravinda Chakravarti, Ph.D.
McKusick-Nathans Institute of Genetic Medicine
Johns Hopkins University School of Medicine

Bioinformatics: the science of organization, storage, dissemination and analysis of biological data

Computational Biology: computational (model-based) inference of function from sequence

Bioinformatics: the science of organization, storage, dissemination and analysis of biological data…

we are quite good at this

Computational Biology: computational (model-based) inference of function from sequence

Bioinformatics: the science of organization, storage, dissemination and analysis of biological data…

we are quite good at this

Computational Biology: computational (model-based) inference of function from sequence…

but we are lousy at this

The genomic data deluge…

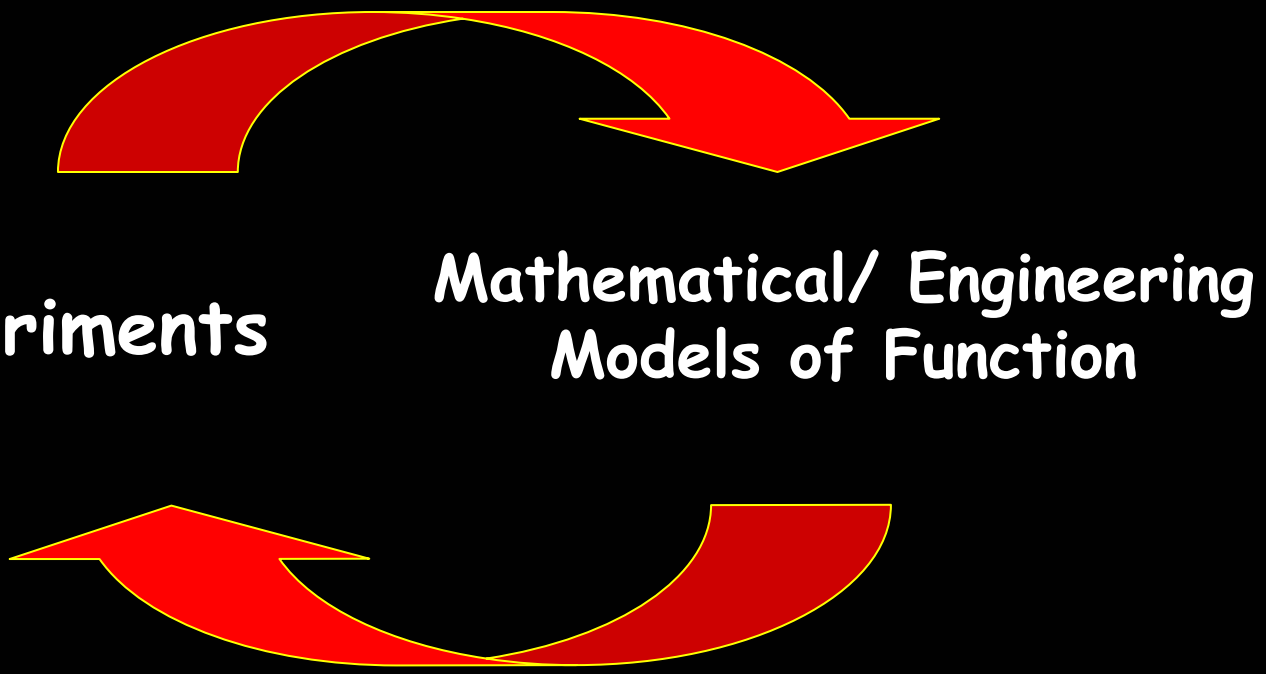
we have perhaps less than 1% of the genomic data we are likely to collect and analyze within the next 5 years

Biology and medicine are unlikely to benefit simply from having large amounts of data…

there is an absolute need to model biological function…

we need a theory of genomic and biological function

A Significant Innovation: The Wet-Dry Cycle



Four perspectives

Importance of genome sequence to identify genes

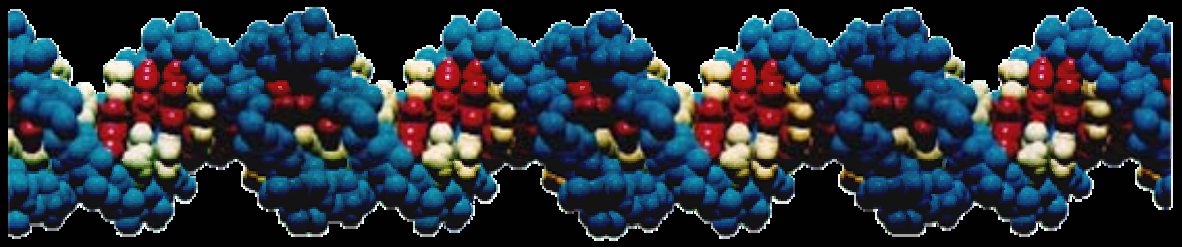
Identifying function through comparative genomics

Identifying disease genes/processes through large-scale association studies

Proving function by chemical genetics


The
Human
Genome
Project

The Human Genome



Type	Proportion of genome	Features
Unique or Single-copy DNA	55%	Includes genes and control sequences; dispersed in genome
Repetitive DNA	45%	Rarely involved in functional features (centromeres, telomeres); interspersed in genome

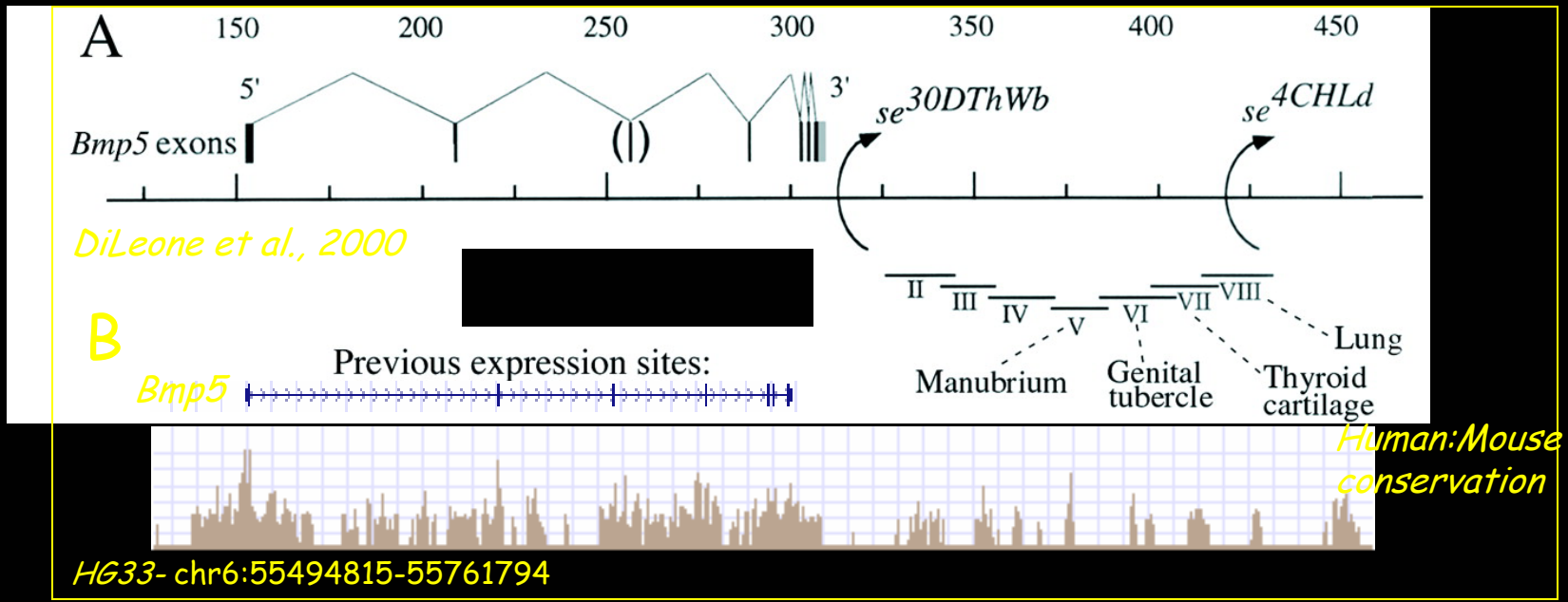
Human Genome ~ 22,500 genes encoded in 3×10^9 DNA units (nucleotides)

Mutations in non-coding DNA regulate gene function

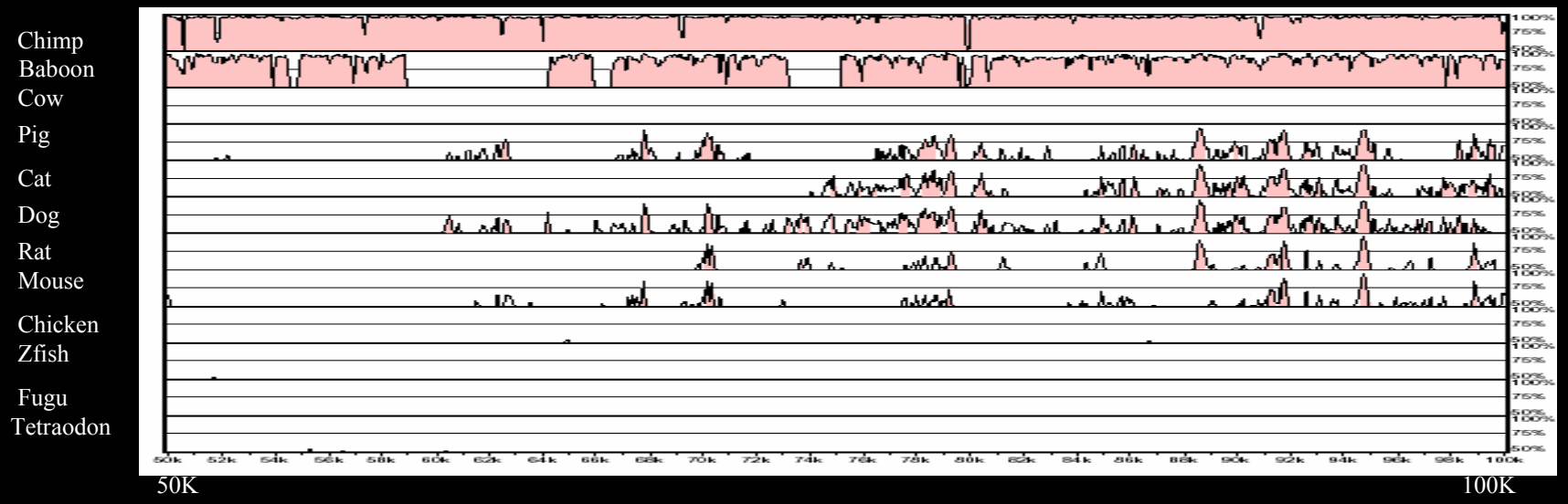
mouse transgenes recover expression with long range elements

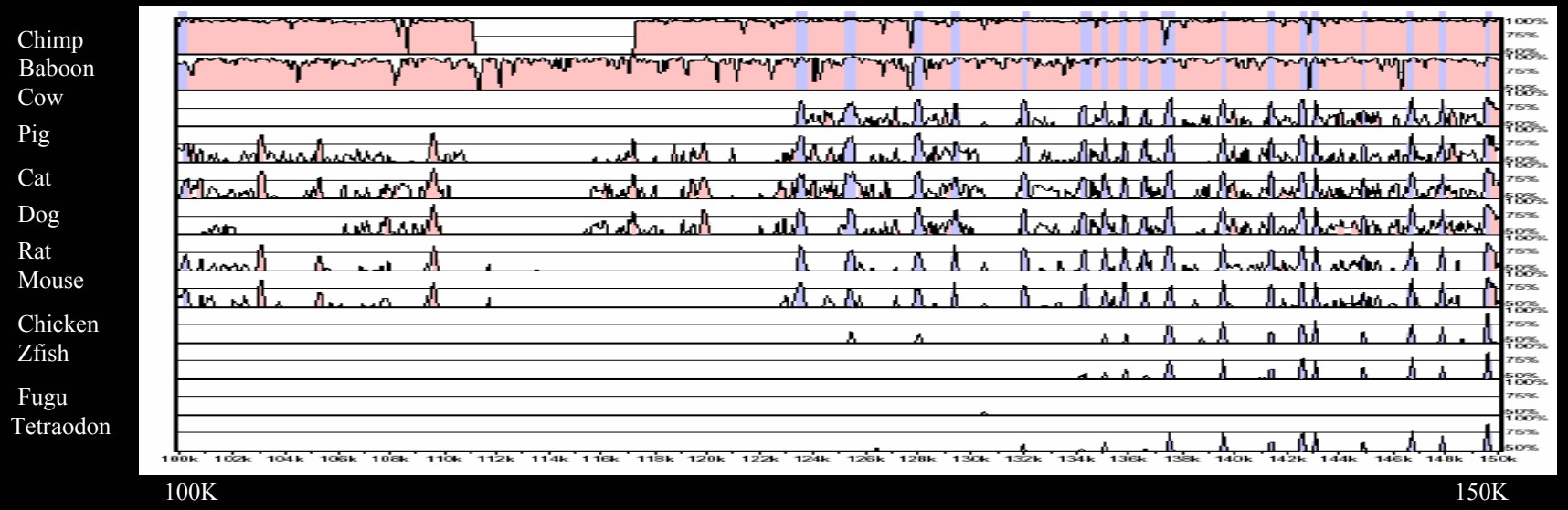
mutations from local rearrangements keeping coding sequence intact

15-20% of Mendelian disorders have no identifiable coding/regulatory motif mutations



Comparative sequencing of genomic *RET* in 13 species…comparisons to human





Identifying functions in the genome

- protein-encoding genes (1.5%)
- RNA genes
- protein (transcription factor) binding sites
- Non-coding conserved elements (3%) or "dark matter"...other regulatory sites

Identifying functions by comparative genome sequencing

sequence throughout the tree of life ?
(<40% sequence in alignments)

sequence many primates ?
(minimizes multiple substitutions;
preserves regulatory apparatus)

sequence even more humans ?

What have we learned ?

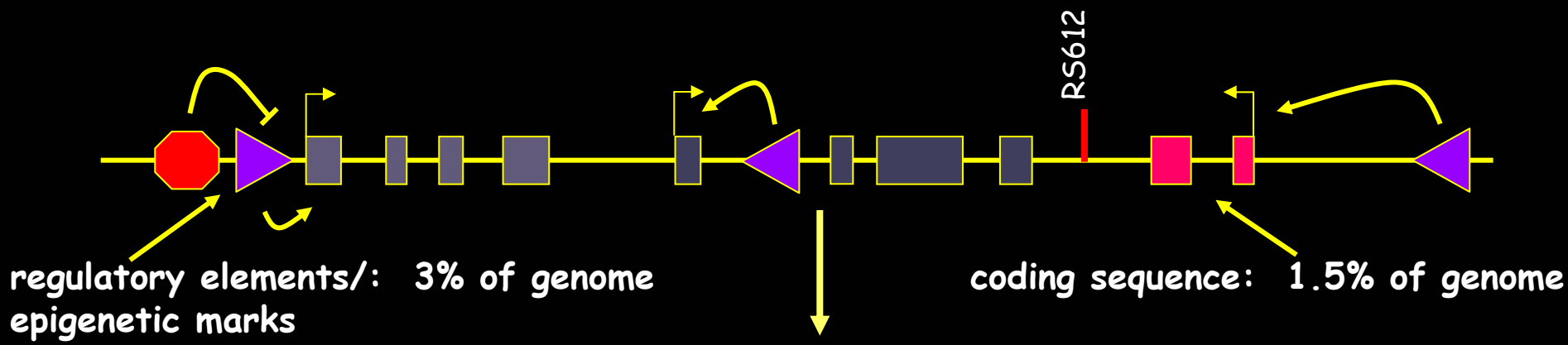
- Genes: 1.5% of genome; most vertebrate genomes have similar gene repertoire; distinct expansions/contractions of genes
- Non-coding: 3.0% of genome; possibly regulatory but largely of unknown function

The future of gene hunting

Genome-wide association mapping will be indispensable in finding genes of smaller effect…effect is "gene sized"

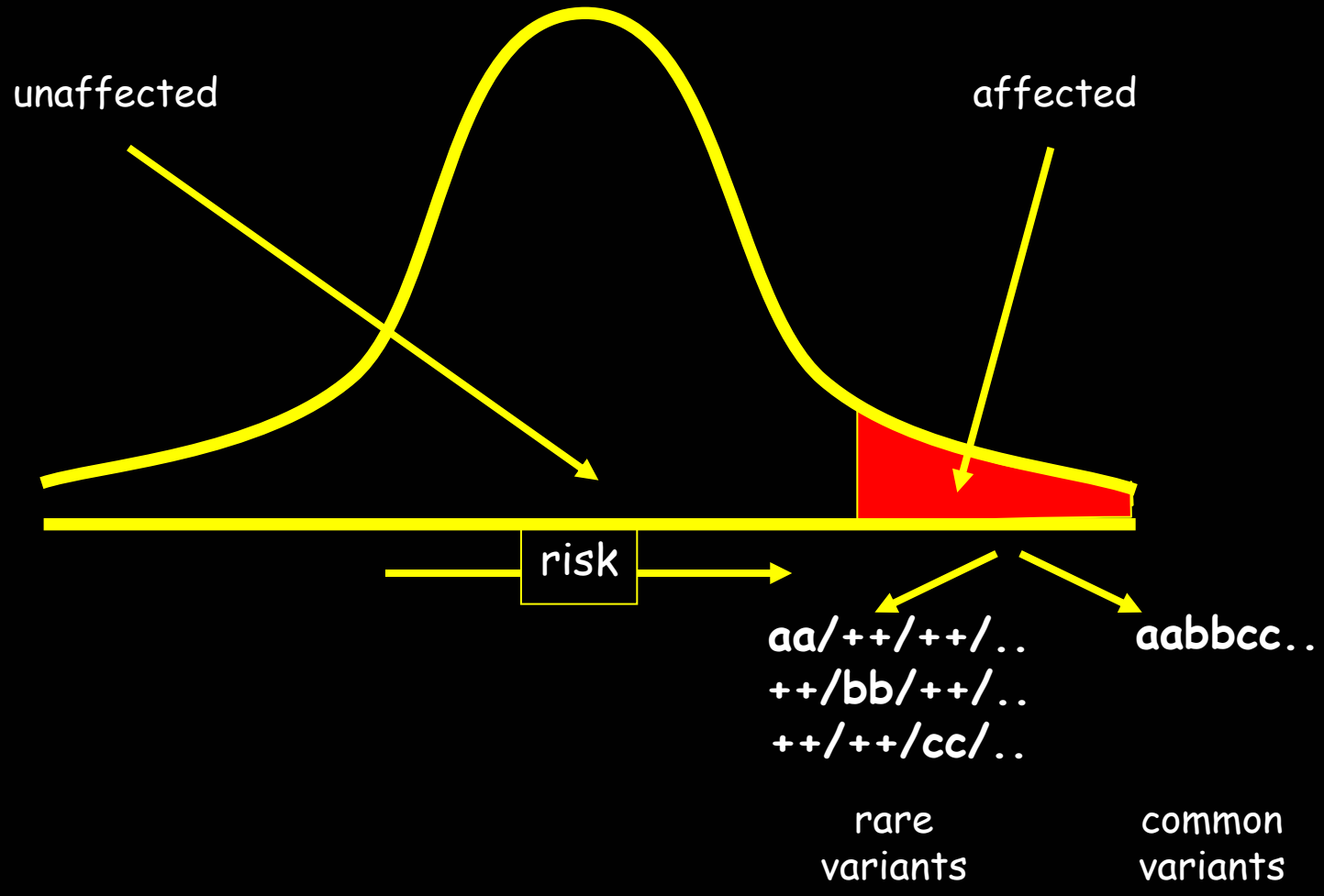
Given an association sequencing of the 5% conserved/functional regions may be the best route to identifying the relevant molecular changes…

Technologies to enable these experiments are now critical



Molecular variant ACGGGAGATGATGTACGTCC

Competing Theories for Common Disease



The HapMap Project

GENOMICS

New Mapping Project Splits the Community

A new type of genomic map, known as the haplotype map, promises to speed the search for elusive genes involved in complex diseases. But some geneticists question whether it will work

It will upend the practice of medicine and save lives the world over, asserts Francis Collins, the director of the National Human Genome Research Institute (NHGRI). "This is the single most important genomic resource for understanding human disease, after the sequence," he says. "We needed it yesterday, as far as I am concerned." Indeed, his institute is leading the National Institutes of Health's (NIH's) $40 million downpayment on the project.

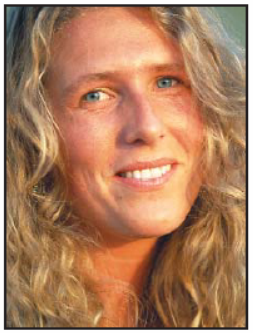
But to many biologists, it's an untested concept hardly worthy of the $110 million it will consume. "The whole thing is a big waste of taxpayer money," says Joseph Terwilliger, a statistical geneticist at Columbia University in New York City.

Welcome to the haplotype map, a new type of genome map that, depending on where you look, is eliciting exuberance or exasperation.

Proponents of the map, who include Collins, Eric Lander of the Whitehead Institute Center for Genome Research in Cambridge, Massachusetts, Panos Deloukas of the Sanger Institute in Hinxton, U.K.—and just about every big name from the Human Genome Project—say it's the best hope for tracking down the genes involved in common diseases, such as heart disease and cancer, that claim so many lives and have eluded most gene-hunting strategies. As an added perk, they say, it provides a tantalizing glimpse at human evolution and migrations.

On the other side, many researchers—mostly population geneticists—say the map's promise is inflated and it may fail to deliver, adding that its proponents are forging ahead with too little data on how best to proceed. "I think there was lots of good, pure, scientific motivation for wanting to do this haplotype map," says Nancy Cox, a human geneticist at the University of Chicago who's not involved in the project. Still, she adds, "people fell into a trap that we should have been smart enough to avoid: that was saying, 'If we have this, we'll get the genes for complex common disorders.' I think it's premature to know how much that will help."

The most virulent critics even contend that the HapMap, as it is known, is nothing more than a full-employment enterprise for the big sequencing labs that might find themselves out of a job once the human genome is completed. "Is the HapMap being designed to satisfy the need [to get] another big project going?" asks Kenneth Weiss, an anthropologist and geneticist at Pennsylvania State University, University Park. Terwilliger minces no words. "The haplotype map is just an excuse for Lander and others to keep funding coming for large factories set up for the sequencing of the genome"—an idea Lander dismisses as sour grapes.



Not everyone's smiling. A plan to study haplotypes in these populations is prompting angry words.

CREDITS (CLOCKWISE FROM LEFT): BOB ROWAN, PROGRESSIVE IMAGE/CORBIS; GWEN FRANKLIN/CORBIS; PAUL A. SOUDERS/CORBIS

Despite these misgivings, the HapMap is well under way. In addition to NIH's $40 million, Canada recently kicked in CAN$9.5 million to fund one of its own researchers. Still some $60 million short, NHGRI will award the first grants this fall.

Unexpected architecture

The idea for the HapMap emerged from the gradual realization that the genome has a surprisingly structured architecture. Rather than being thrown together randomly, thousands of DNA bases—as well as patterns of single-base variations found among them—line up in roughly the same order in many different people. Like an interior decorator debating among four kitchen designs, a person's genome has just one of a few potential blocks of DNA to slap into a defined space on a chromosome. Each DNA block—or kitchen design—is a haplotype.

Mark Daly, a computational biologist at Whitehead, stumbled upon these blocks a couple of years ago as he was scouring chromosome 5 for susceptibility genes for Crohn's disease. Stretches of DNA from 129 families affected by the disease kept falling out in one of about four patterns, Daly wrote in the October 2001 *Nature Genetics*.

His Whitehead colleagues David Altshuler, Lander, and others wasted no time investigating. In a paper published last spring in *Nature*, they described common haplotypes in a group of northern Europeans and a Nigerian population called Yorubans, arguing that haplotypes varied somewhat between the two—a vestige of evolutionary history. A paper in *Science* (23 November 2001, p. 1719) by a group led by David Cox of Perlegen Sciences in Mountain View, California, found just a few haplotypes on chromosome 21.

Presented at a Cold Spring Harbor Laboratory meeting last spring, these findings initially generated skepticism. They began to win converts, however, and by summer, Collins made the HapMap—a map identifying haplotypes across the genome—a priority and was rustling up money and collaborators. In March, NIH began soliciting grants for the first of the map's two stages. The first will be to create haplotype maps of the genomes of three populations: those of northern and western European ancestry, Japanese and Chinese, and Yorubans. In the second stage, scientists will test whether the haplotypes they find in those very large populations also appear in about 10 others.

From SNPs to haplotypes

Haplotypes gained popularity as scientists realized that mining the genome was much tougher than expected. A few years ago, geneticists heralded single-nucleotide polymorphisms (SNPs) as the long-sought answer to finding genes involved in complex

International HapMap Project

Home | About the Project | Data

中文 | English | Français | 日本語 | Yoruba

[Genotype Access] [Register]

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap Project**" for more information.

Project Information

About the HapMap Project
HapMap Project Data
HapMap mailing list
HapMap Project Participants

Useful Links

HapMap Project Press Release
NHGRI HapMap Page
NCBI Variation Database (dbSNP)
Japanese SNP database (JSNP)

News

- *2004-05-04 :* ***Public data release #7***
  Genotypes, frequencies and assays for 557,083 SNPs (50,137,470 genotypes) released for **bulk download** and **graphical browsing**.
- *2004-04-09 :* ***Public data release #6***
  Genotypes, frequencies and assays for 462,670 SNPs (41,640,300 genotypes) released for **bulk download** and **graphical browsing**.
- ***Old News***

Genome-wide association studies are efficient for disease gene discovery

- SNPs (HapMap); resolution < 2 kb

- high efficiency of association studies for common disorders

- resolution of association in the human genome is < 20 kb

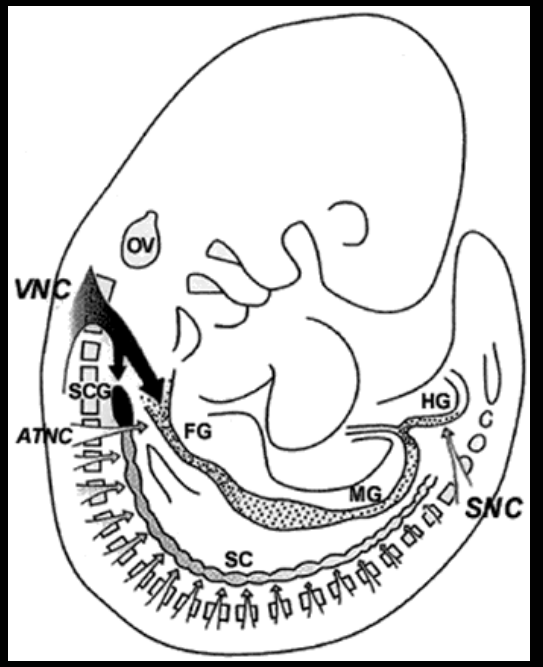
- cleft lip/palate and IRF6; leprosy and PARK/PARG

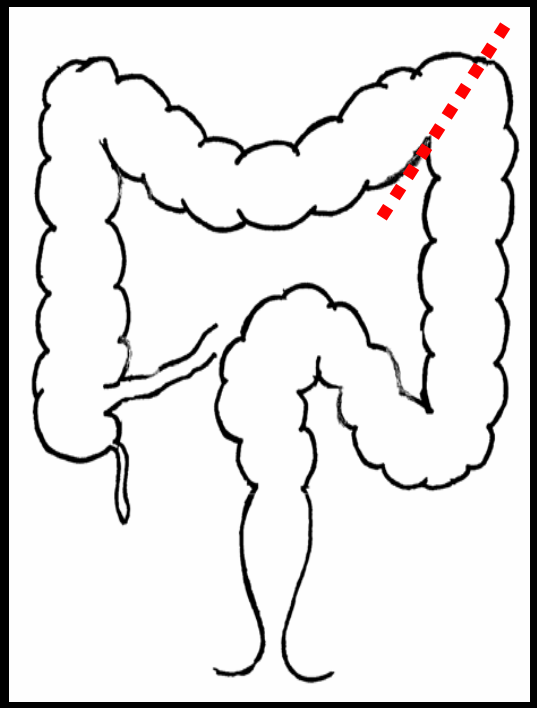
How do we find the molecular mutations ?

One simple example: DNA analysis in Hirschsprung disease

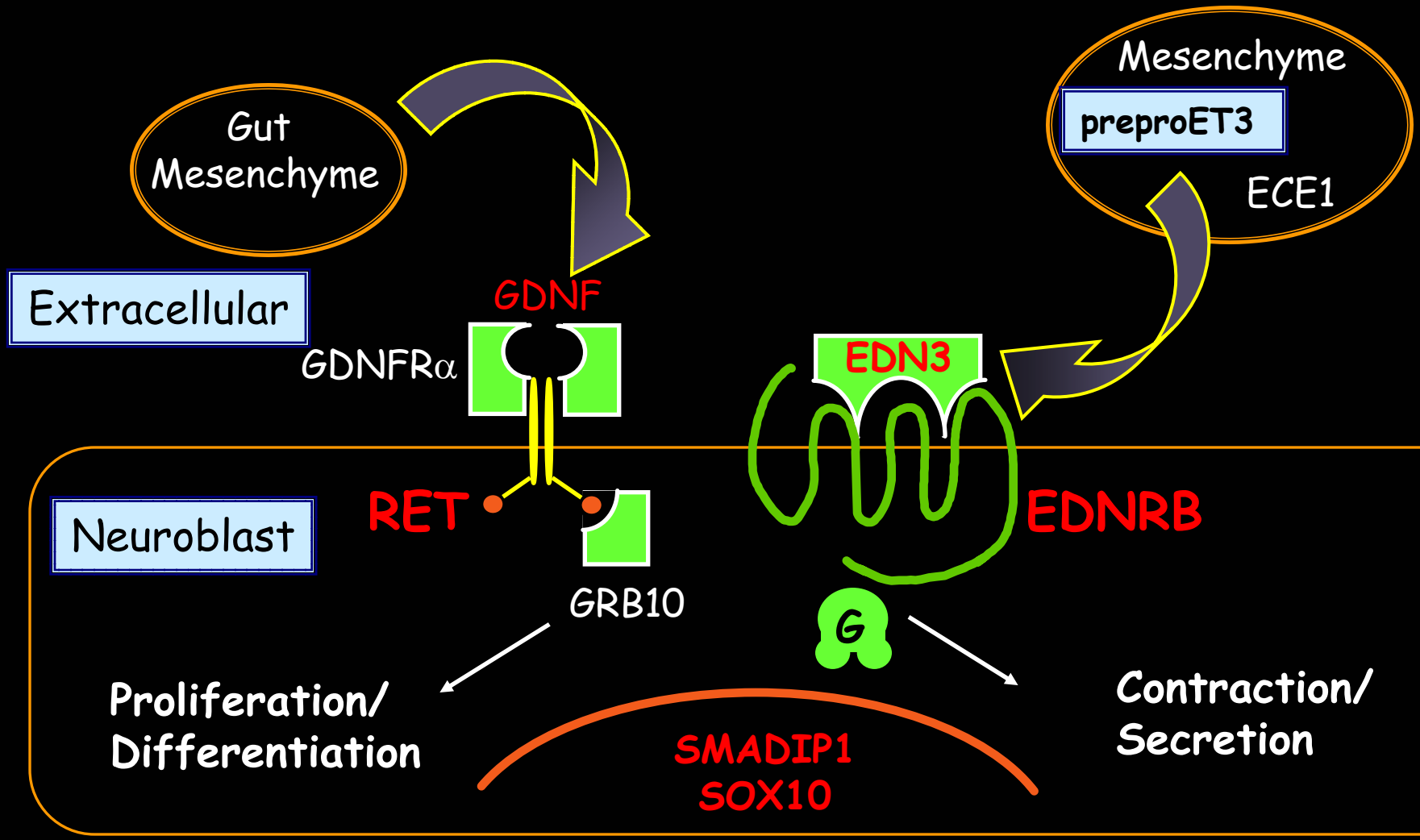
HSCR genetic epidemiology

- Absence of enteric ganglia; neurocristopathy
- Long & short segment HSCR (20%:80%)
- Mutations: Long >> Short
- Incidence ~ 1/5,000
- Affected males:females = 4:1
- Syndromic associations e.g. WS, MEN2; Trisomy 21
- Non-mendelian inheritance





HSCR: disrupted interaction between enteric neuroblasts and gut mesenchyme



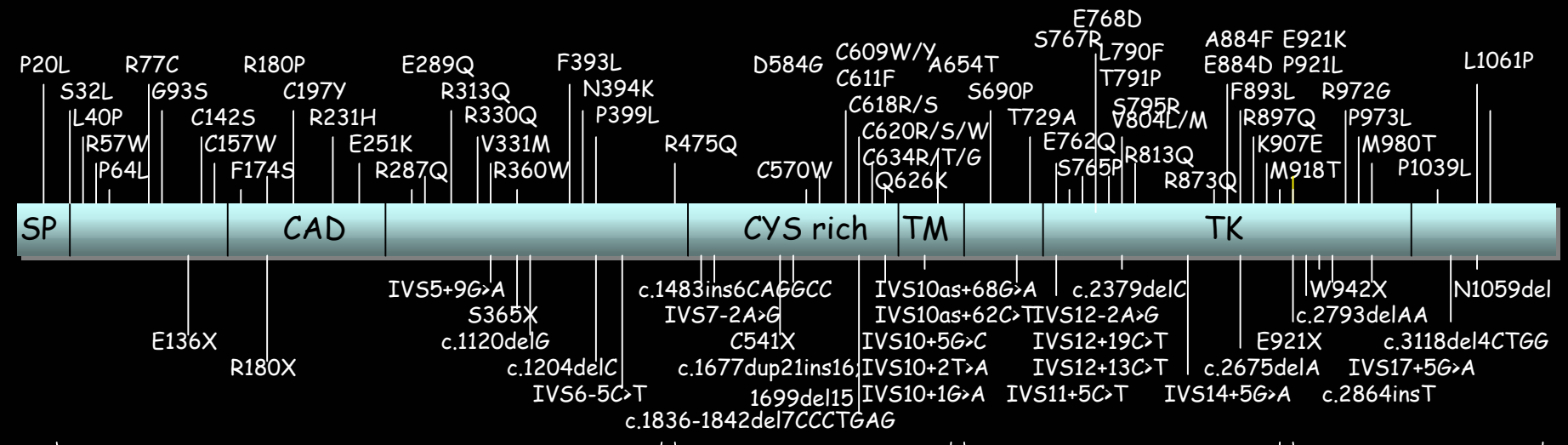
Also - 3p21, 9q31, 19q12, 16q23, 21q21-tel

RET mutations in syndromic & L-HSCR

RET-linked families - 90%

RET mutations

- 50% familial cases
- 35% sporadic cases



HSCR mutations

Class I
Reduced RET to cell membrane

Class II
Reduced transport to cell membrane / constitutive dimerization

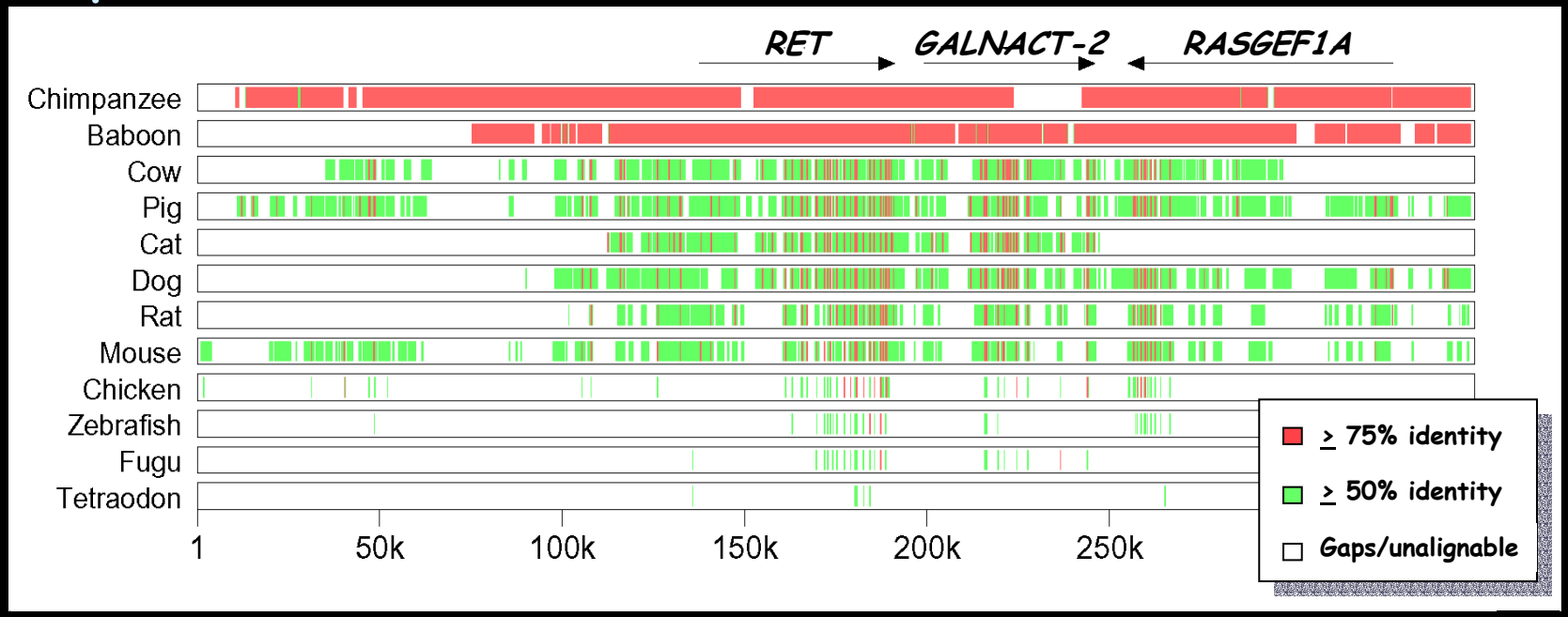
Class III
Abrogated /reduced kinase activity

Class IV
Compromised adaptor protein interaction

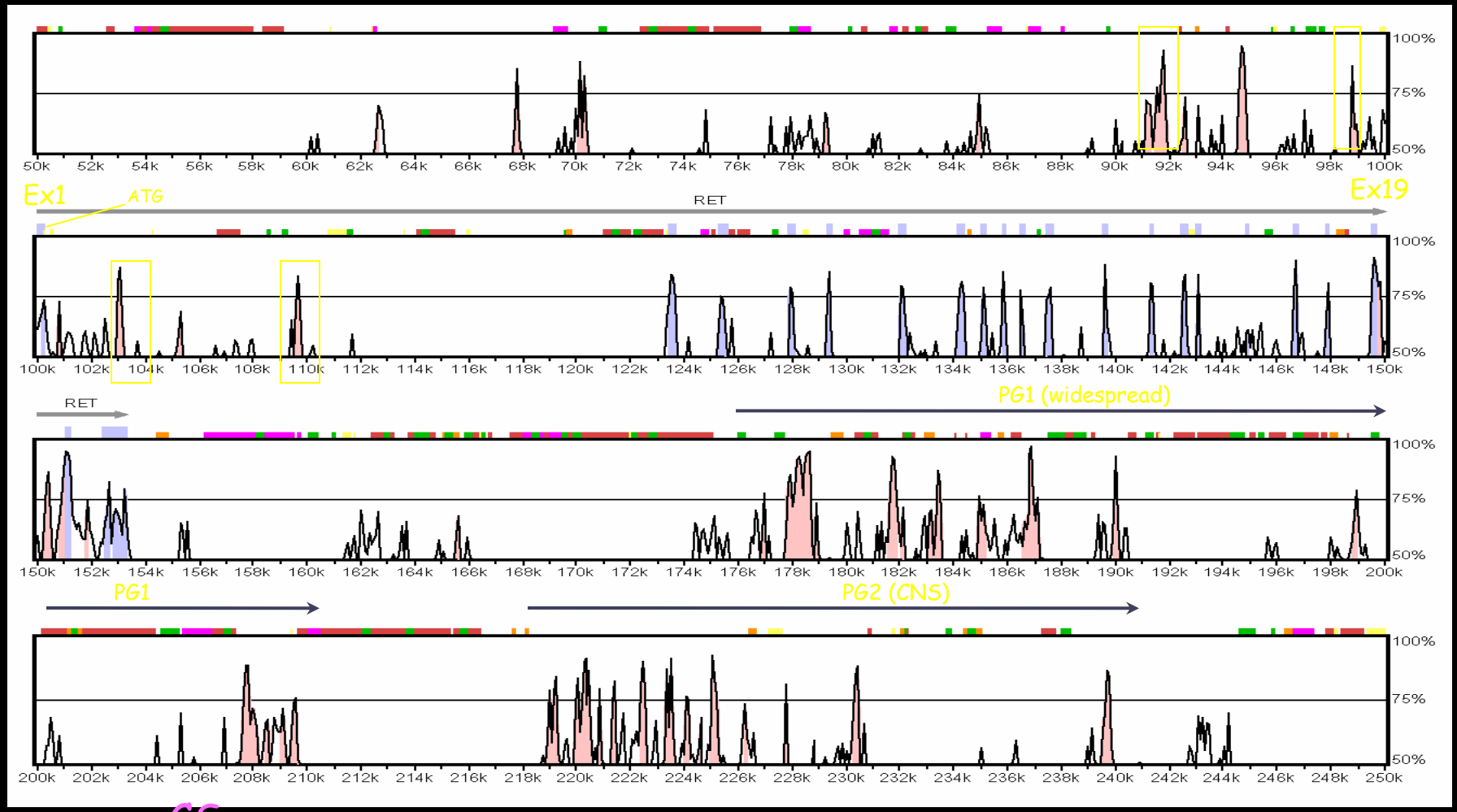
Where are the other mutations?

- *RET* non-coding sequence?
- tightly linked gene?

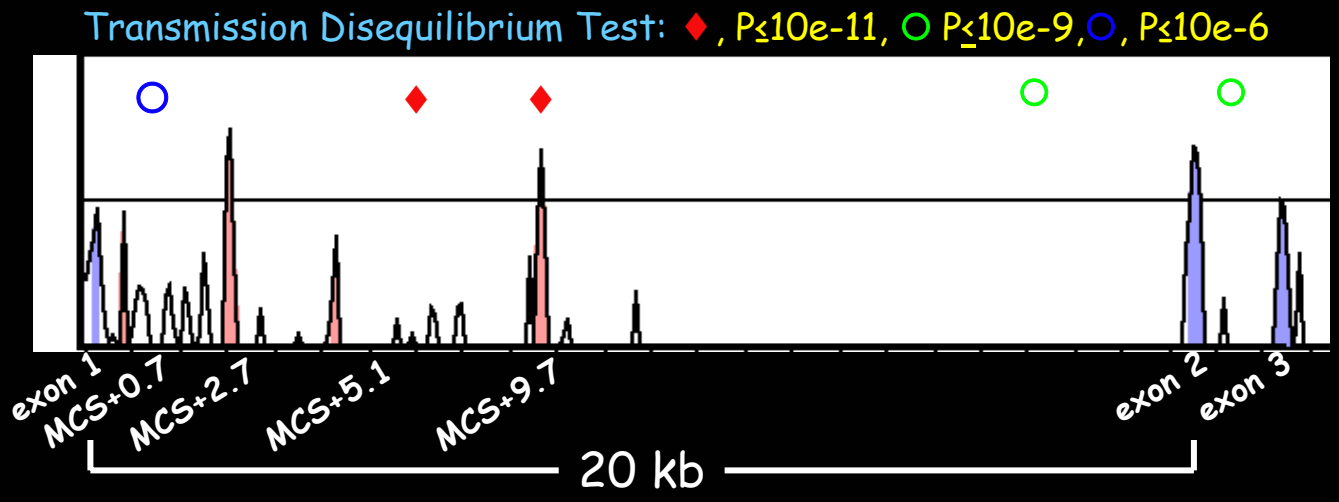
Identification of multi-species conserved sequences (MCSs) within 350 kb at human *RET*

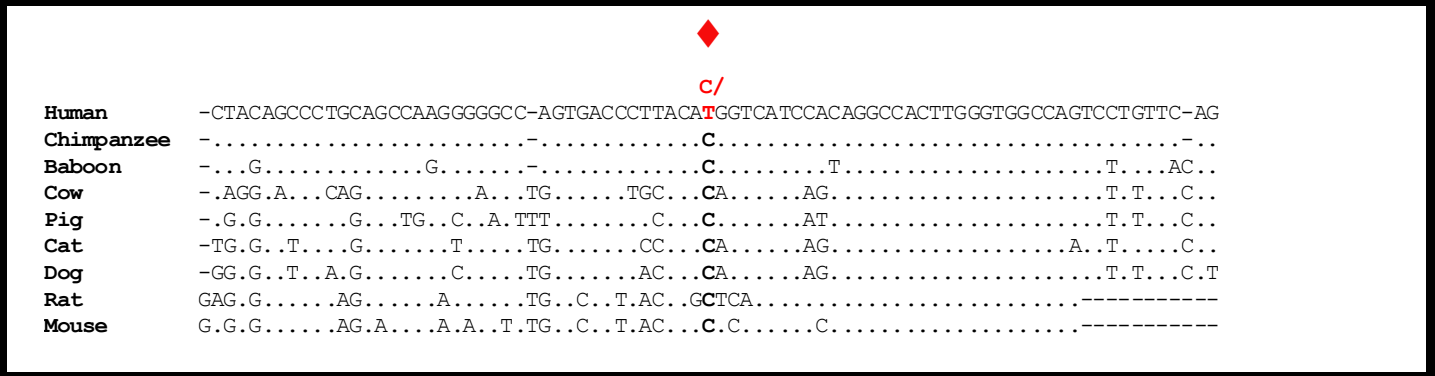


RET is flanked by conserved, non-transcribed sequences and novel genes

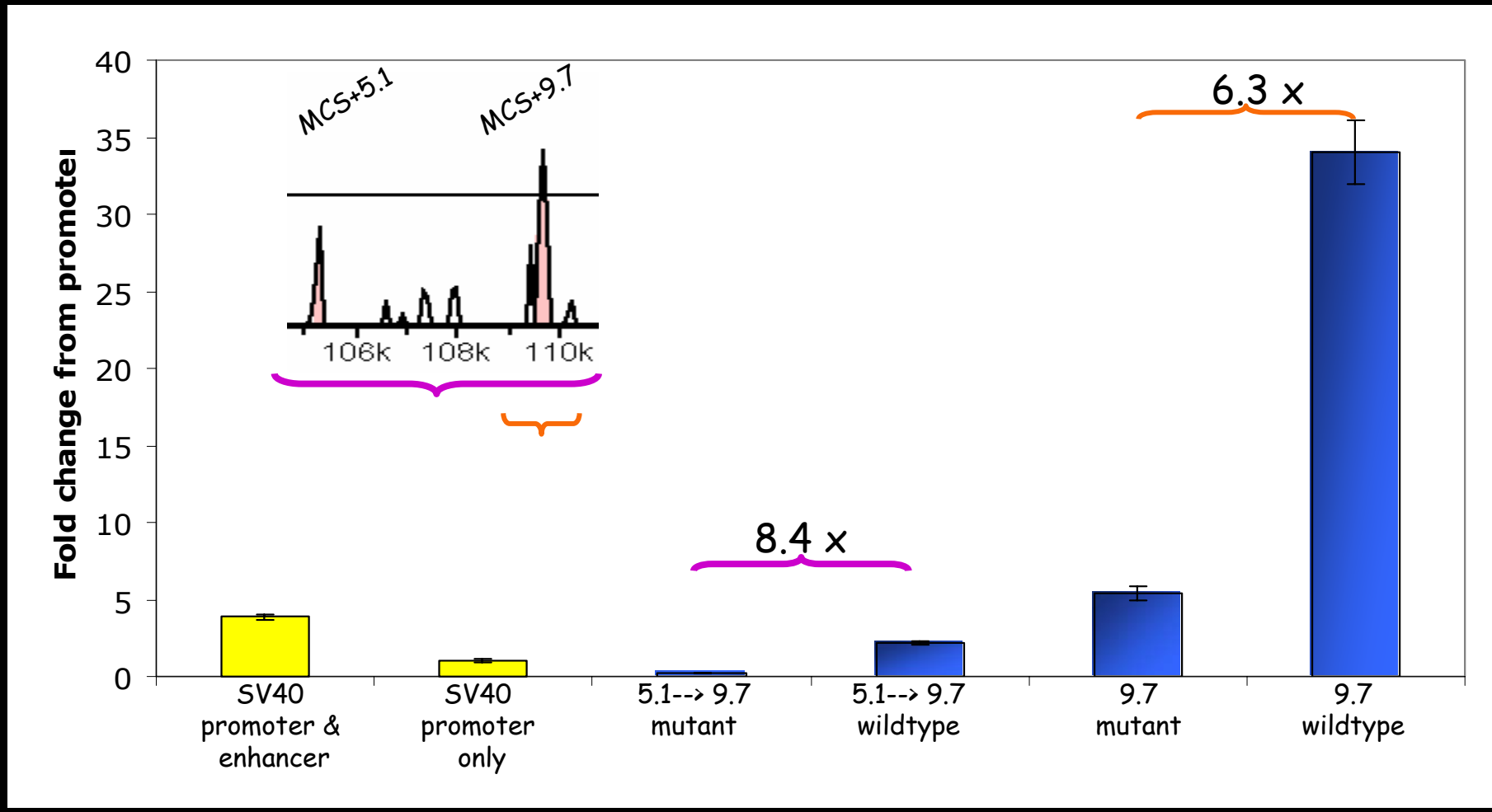


Identification of non-coding MCSs within *RET* intron 1





MCS+9.7 functions as an enhancer *in vitro*

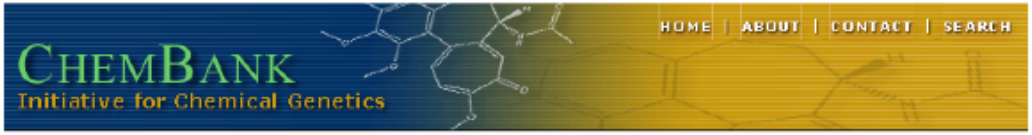


Sequence-based Genetics

- Sequence genomes of affecteds & controls
- Identify sequence changes
- Predict functional effect

Validating genetic variants

- Functional studies
- Replication in other samples
- Challenge tests in extreme homozygotes
- Chemical genetics



ChemBank is a freely available collection of data about small molecules and resources for studying their properties, especially their effects on biology. It is being developed to assist biologists who wish to identify small molecules that can be used to perturb a particular biological system and chemists designing novel compounds or libraries, and serve as a source of data for cheminformatic analyses. The project is still in an early stage and we will be adding tools and data frequently.

Newly Added

- A Structure Database of over 900,000 reference structures with full search capabilities.
- A resource for searching, retrieving, and analyzing Biological Assay Data.
- Significant data and tools — including substructure searching — have been added to the Bioactives Database.

Available Resources

Structure Database - contains all chemical structures and biological activity data in the ChemBank database. Structures are available for more then 900,000 compounds, including a reference set of over 700,000 commercially available drug-like compounds.

Small Molecule Bioactives Database - contains chemical structures and biological activity data for over 5900 known bioactive compounds. This represents a subset of the compounds in the Structure Database.

Biological Assay Data - a resource for searching, analyzing, and downloading publicly available assay data from high throughput chemical screens.

Data Downloads - downloadable files of compound structures and other useful information.

Resources - Links to outside resources relevent to small molecules, cheminformatics, and high throughput screening.

Coming Soon

- Enhanced bioactivity data including many new compound structures and updated annotations.
- A reaction database and other tools for populating and navigating chemical descriptor space.
- Access to more high throughput screening data including signature datasets (see J. Am. Chem. Soc., 125, 10543-10545(2003)).

ChemBank is intended as a resource for the entire scientific community, and we welcome feedback and suggestions. To enquire about participating in the development of ChemBank, please contact us.

ChemBank is supported by the National Cancer Institute's Initiative for Chemical Genetics.

Finding drug targets is related to identifying the genetic factors in complex human disease…

…factors modifying susceptibility, duration, severity and course

…and Computational Biology is critical to this exercise