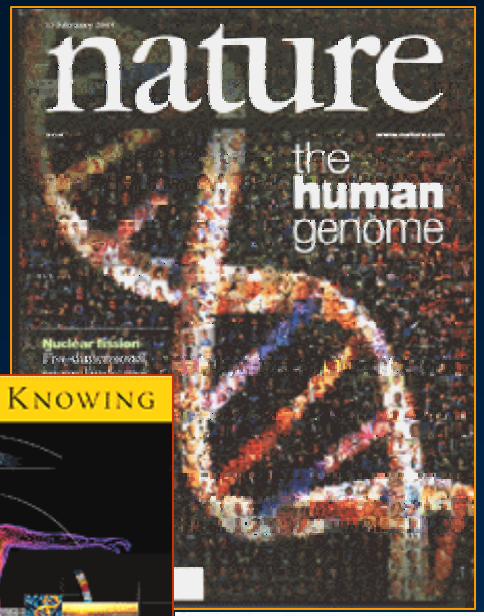
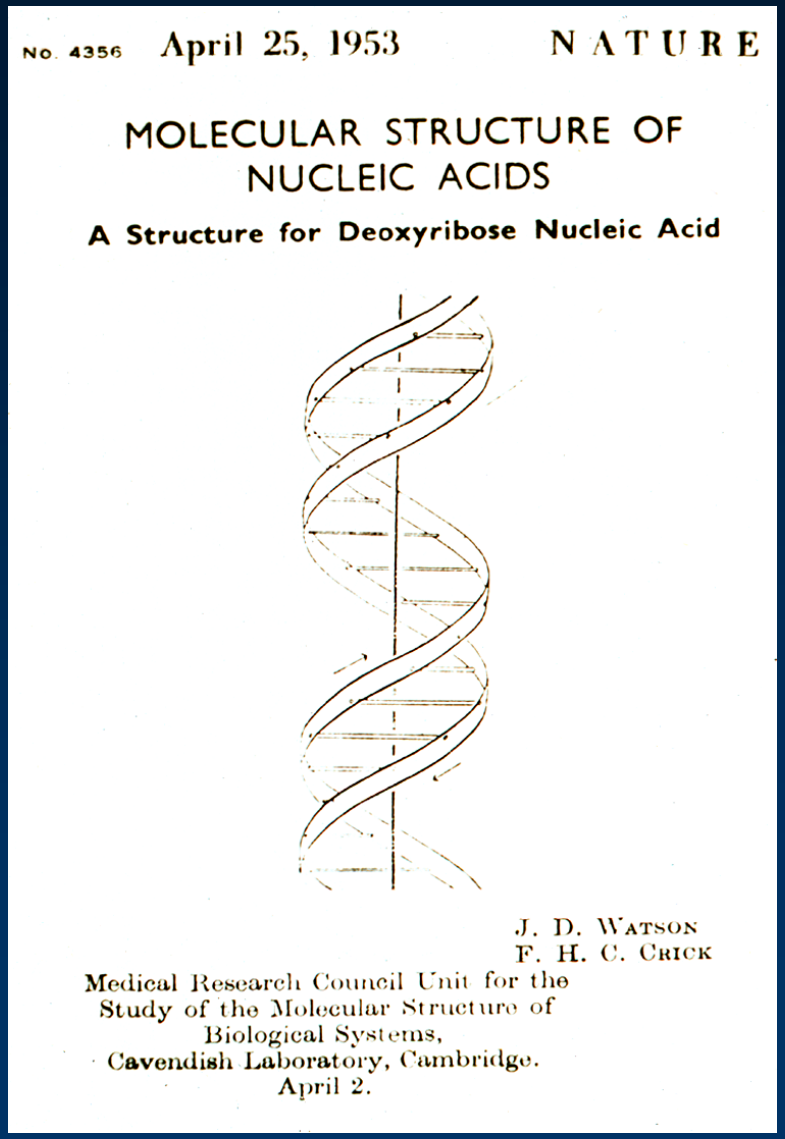
Public Health Genomics Series
January 18, 2007



"Omics" 101 for Medicine & Public Health

Stephen Chanock MD
National Cancer Institute

April, 1953 → April, 2001



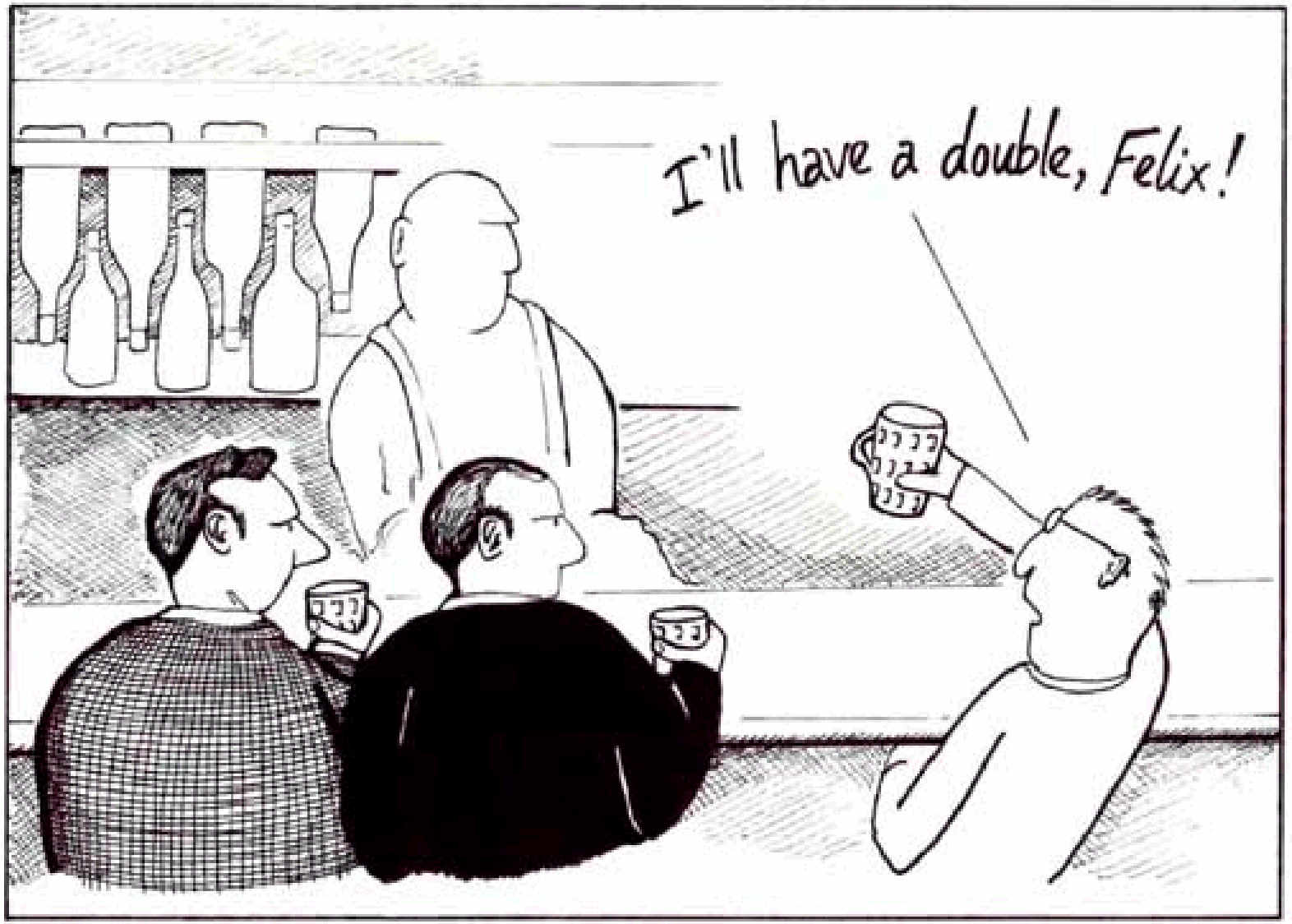
No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

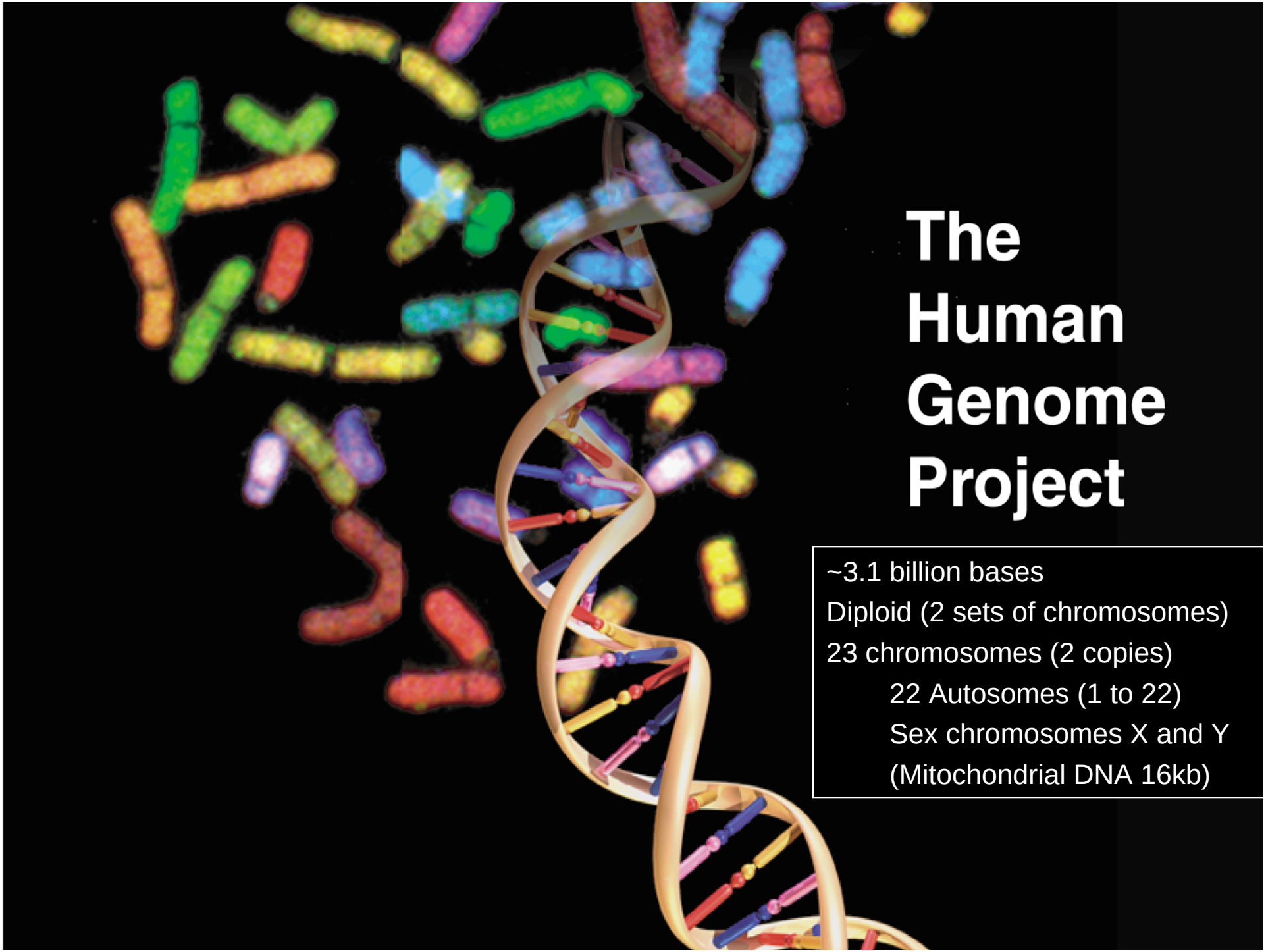
A Structure for Deoxyribose Nucleic Acid

J. D. WATSON
F. H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge.
April 2.



Cambridge, 1953. Shortly before discovering the structure of DNA, Watson and Crick, depressed by their lack of progress, visit the local pub.



The Human Genome Project

~3.1 billion bases
Diploid (2 sets of chromosomes)
23 chromosomes (2 copies)
- 22 Autosomes (1 to 22)
- Sex chromosomes X and Y
- (Mitochondrial DNA 16kb)

“Omics”

Genomics

Proteomics (>1 million proteins, cell specific)

Transcriptomics (set of all transcripts)

Metabolomics (chemical fingerprints behind)

Epigenomics (alterations of DNA-methylation)

Glycomics (structure and function of sugars)

Pharmacogenomics (variation and response)

“Veritagenomics” (truth)…….

Genomics

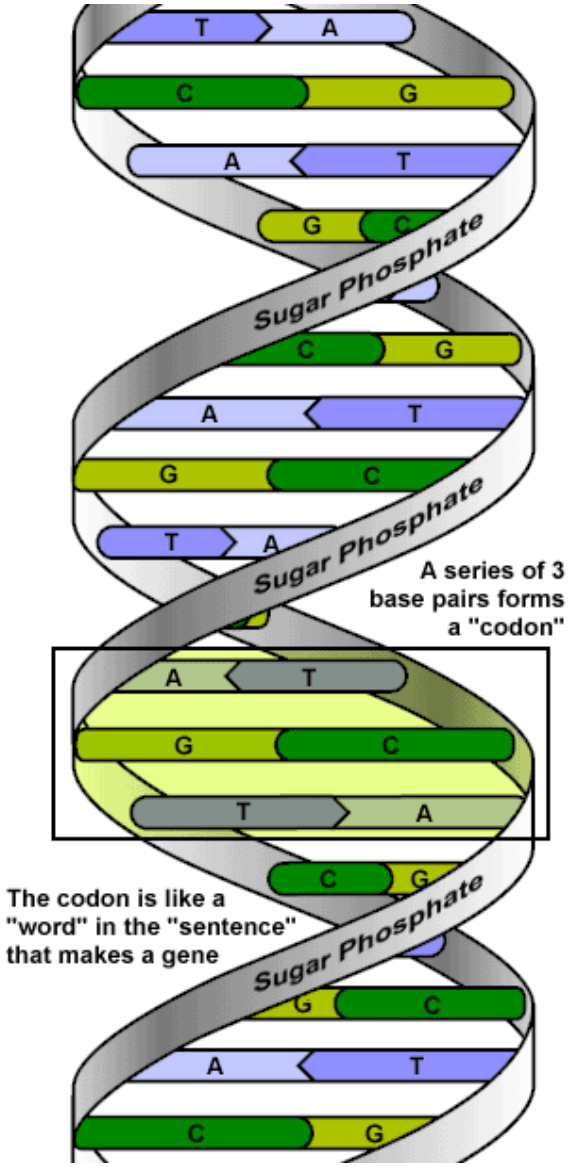
"The near simultaneous study of all nucleotide sequences, including structural genes, regulatory sequences and non-coding DNA segments, in the chromosome of an organism"

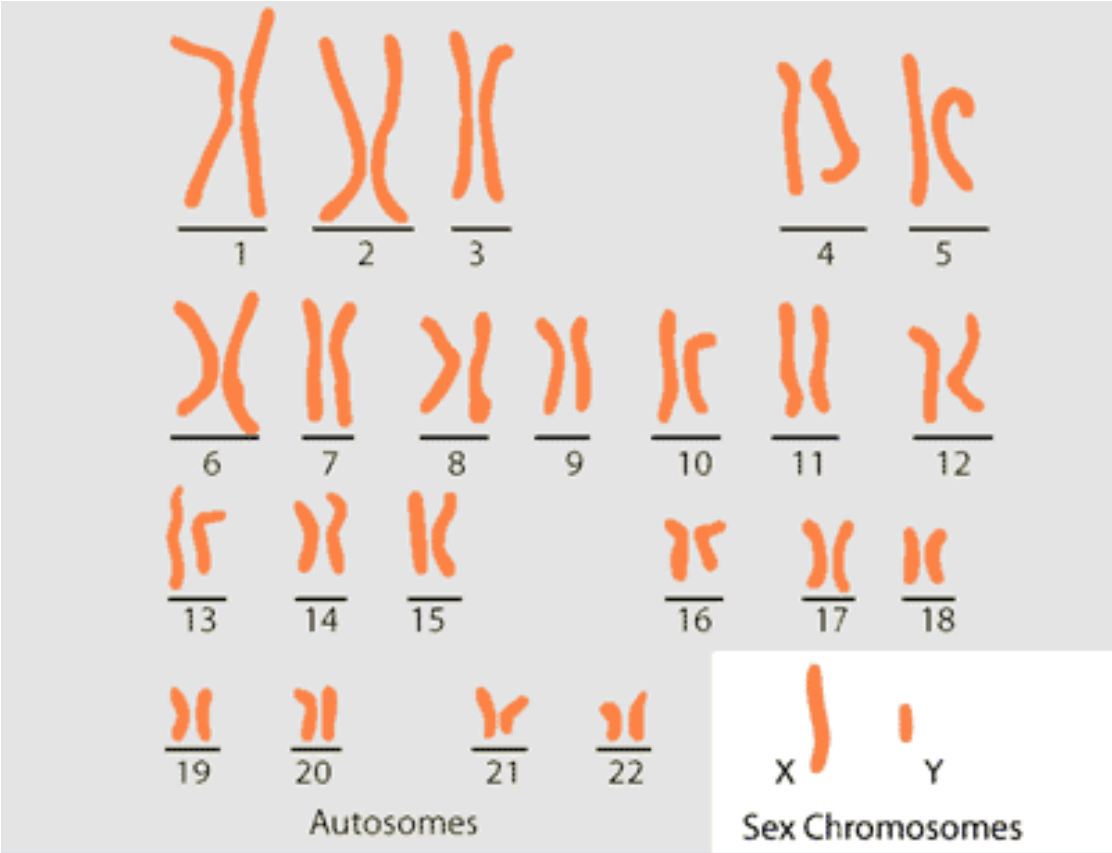
Completion of Genome Sequences of Many Organisms

Evolutionary Ladder: Are More Genes Better?

Organism	Number of Genes
Rice	50,000
Poplar Tree	45,000
Puffer Fish	30,000
Human	**<25,000**
Mouse	25,000
Sea Urchin	23,300
Nematode	19,000
Sea Squirt	16,000
Drosophila	13,600
Yeast	6,300
E. coli	4,000
HIV-1	9-10

From Small to Large





The Human Genome

The Human Genome is the total of the genetic information that is held in each human cell. It is usually made up of 46 chromosomes: 22 pairs of autosomes and 1 pair of sex chromosomes, which are usually X and X for females and X and Y for males.

Definitions

bp= base pair

Kb= 1000 base pairs

Mb= 1000 kilobases

Genetic Map:

Placement of genes by recombination mapping.

Position of gene or markers relative to each other

Physical Map:

Placement of genes by nucleotide sequence

Basic Definitions

Locus: Place on a chromosome where a specific gene or set of markers reside

Gene: Contiguous piece of DNA that can contain information to make or modify ‘expression’ of specific protein(s)

Polymorphism: Variation in the sequence of DNA among individuals

Allele: A variant form of a DNA sequence at a particular locus on a chromosome

Note: *these terms were previously defined when we did not have access to complete DNA sequence*

Chromosome Anatomy

Telomere Centromere Telomere

GC rich

Recombinogenic

Gene-rich

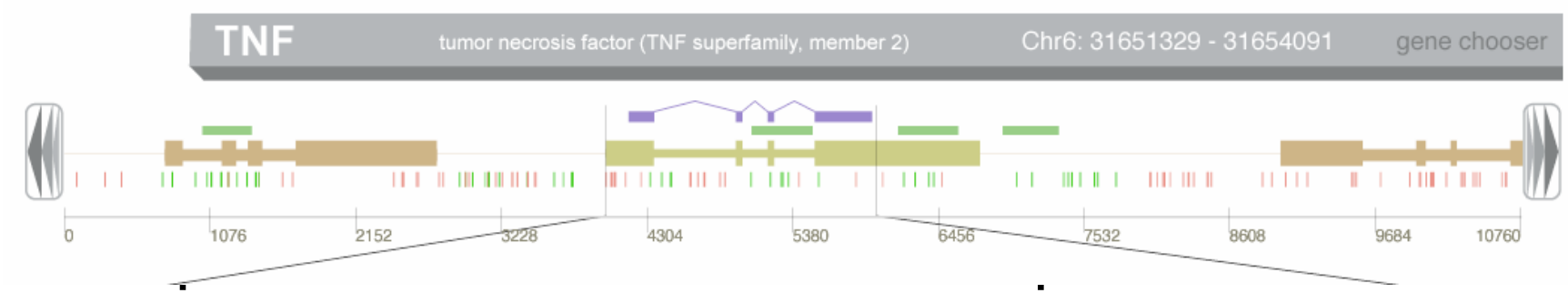
TTAGGG repeats

Genes

Official name and symbol designated by The Human Genome Organization (HUGO)

Symbols for Humans

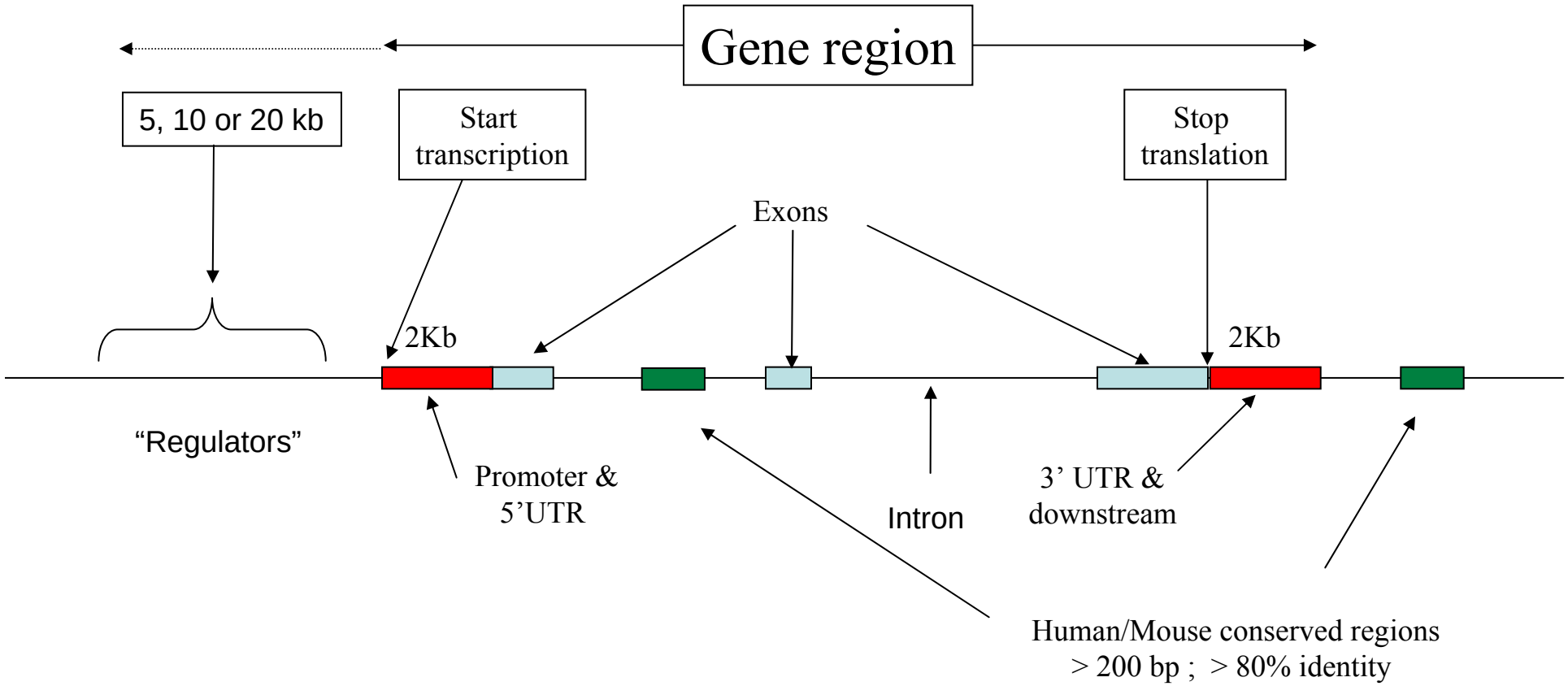
ALL CAPITALIZED & ITALICS



Gene= *TNF* (mouse= *tnf*)

Protein= TNF-α

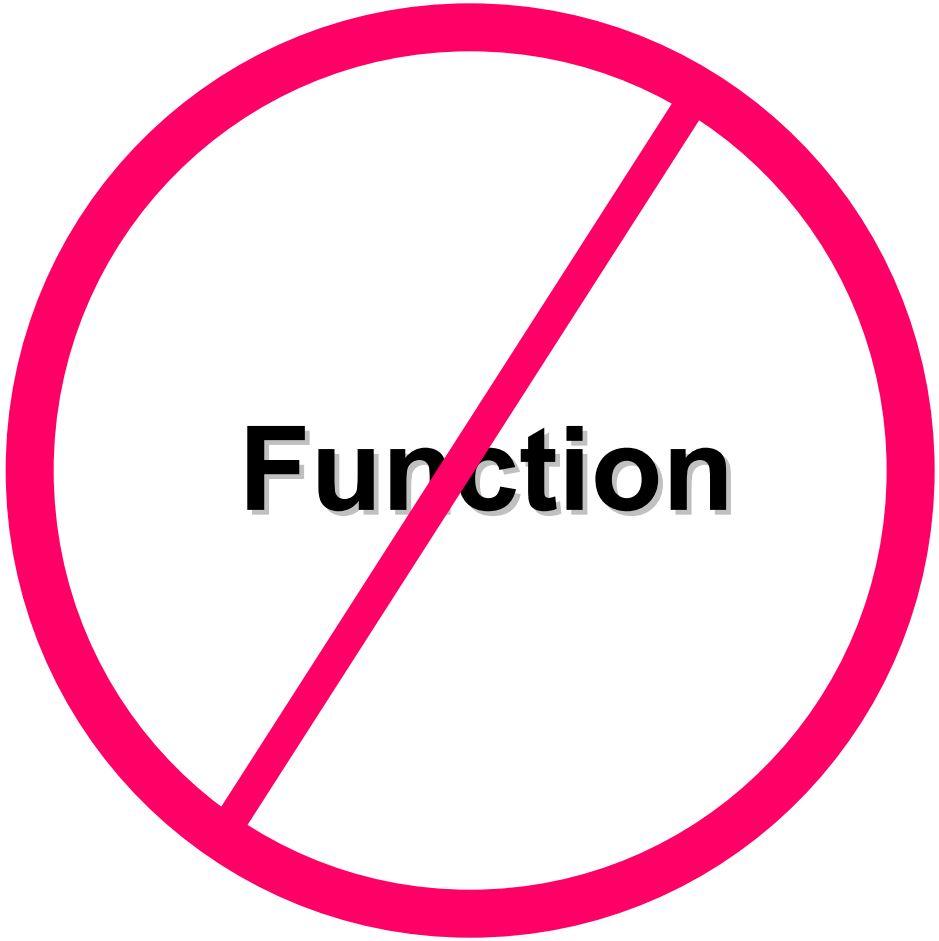
What is a ‘Classical’ Gene?



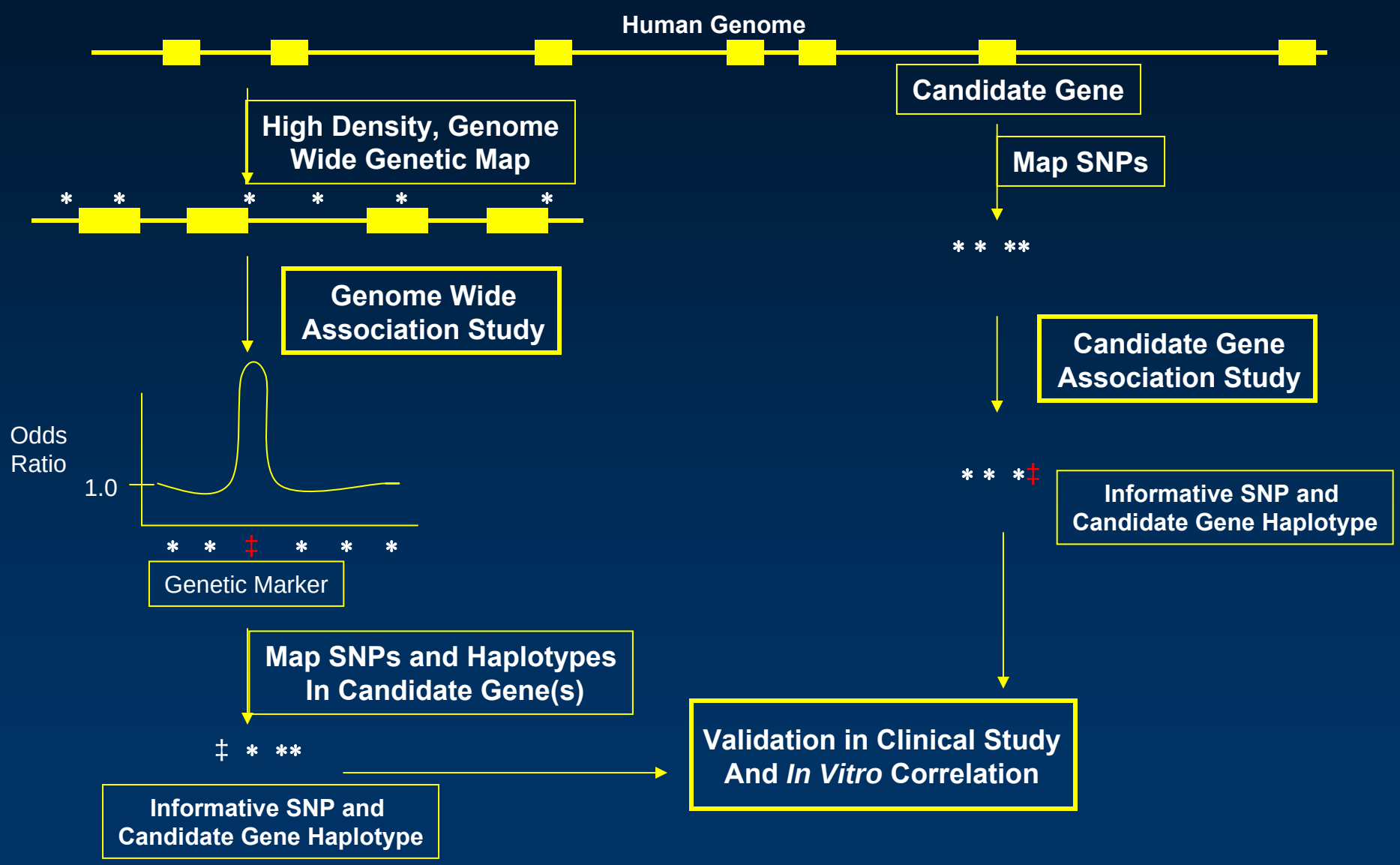
"What ever will we think about now that the genome project is almost complete?"

Fundamentals of Genetic Variation

1. Different types of variation throughout genome
2. Genetic Studies *search* for Genetic Markers
3. Causal significance MUST be established in corollary studies
   - In vitro
   - Animal models
4. Function is established separately…..


Function

Parallel Approaches to Identifying Genetic Markers for Disease



What happens when things change?

Mutation=change in bp sequence

Point substitution- most frequent in genome

Transition= purine to purine (e.g., A->G)

Transversion= purine to pyrimidine (e.g., A->T)

Mutational events:

Occur approximately one every 10^8 replication events for point mutations

Types of Polymorphisms I

- Single nucleotide polymorphisms (SNPs)
  - Most common SNPs are defined as >1% in at least one population
  - Rare SNPs are hard to identify and validate
    - *But*, it is estimated that there are a large number per individual
- MAF= minor allele frequency

Coding SNPs

Synonymous:

No change in amino acid

Previously termed “silent” but…..

Can alter mRNA stability

Nonsynonymous

Changes amino acid

Conservative and radical

Nonsense

Insertion of stop codon

Indel

Disrupts codon sequence

Rare but disruptive

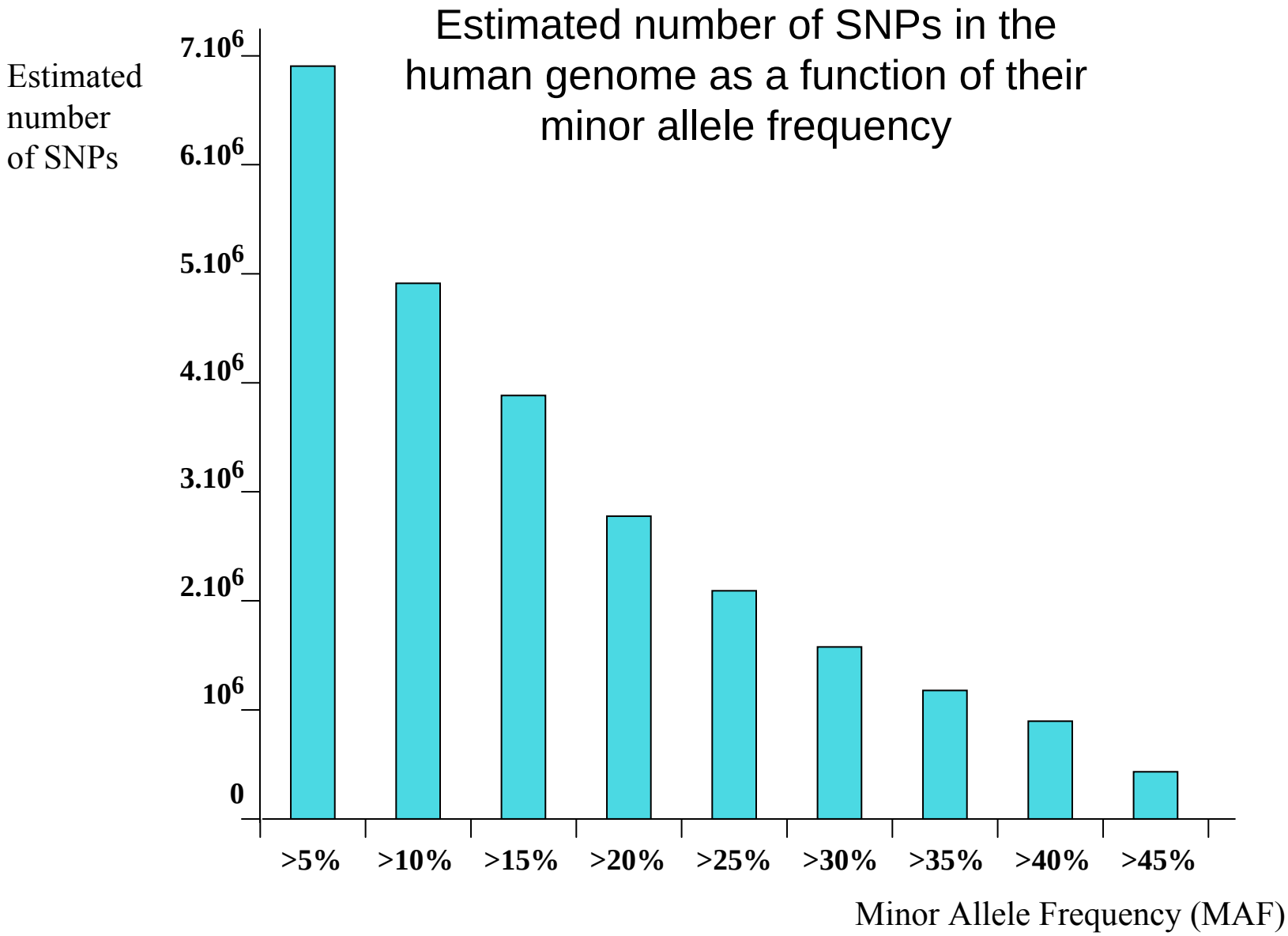
SNPs Outside Genes:
Too many….

Majority distributed throughout genome are "silent"

No function by predictive models or analysis

Excellent as markers

'Hitchhikers'



Common SNP : a SNP with MAF > 0.05 ; frequency of heterozytotes >≈ 10%

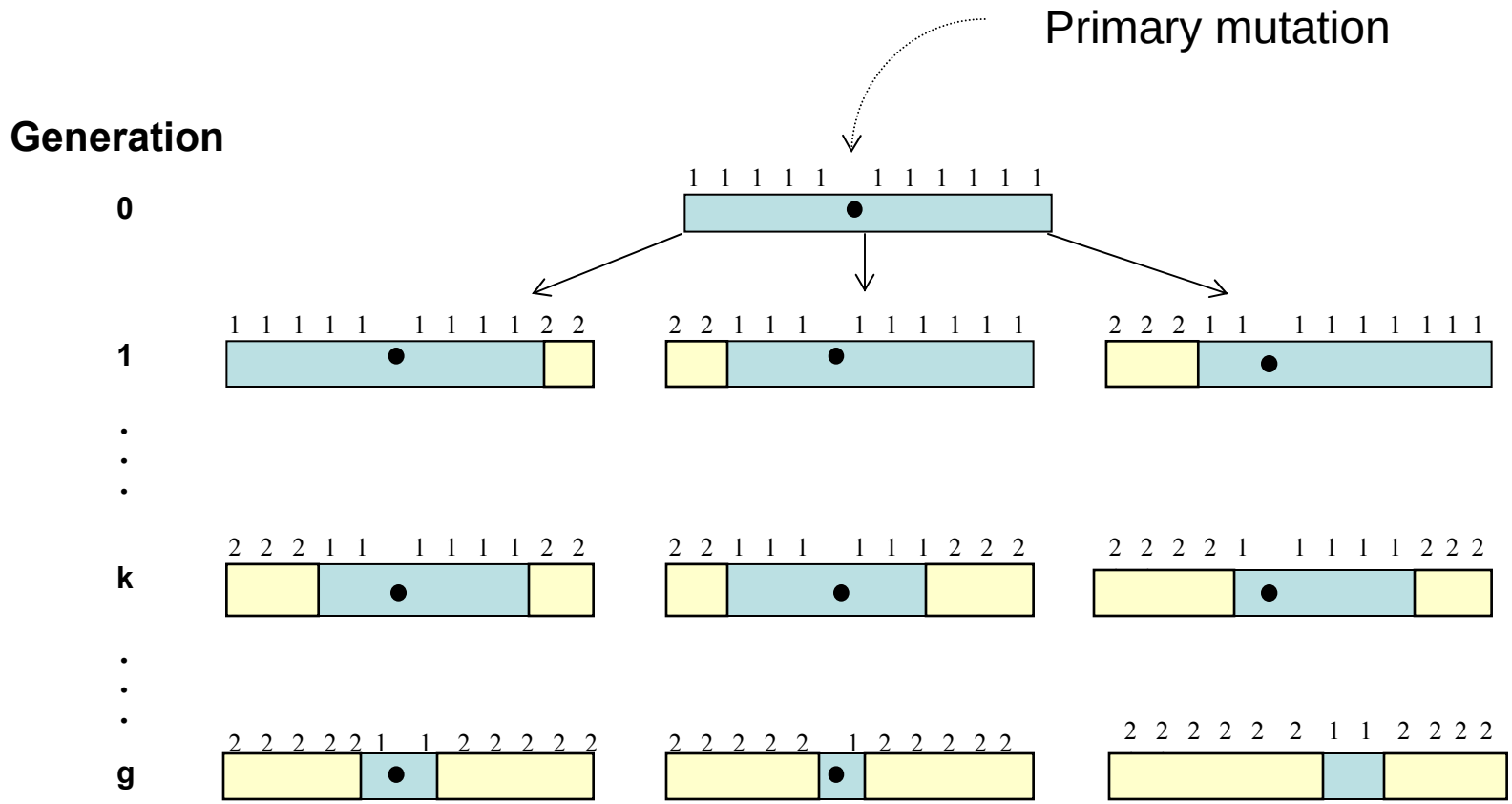
G Thomas 2005

Adapted from Reich et al. Nat Genet (2003)

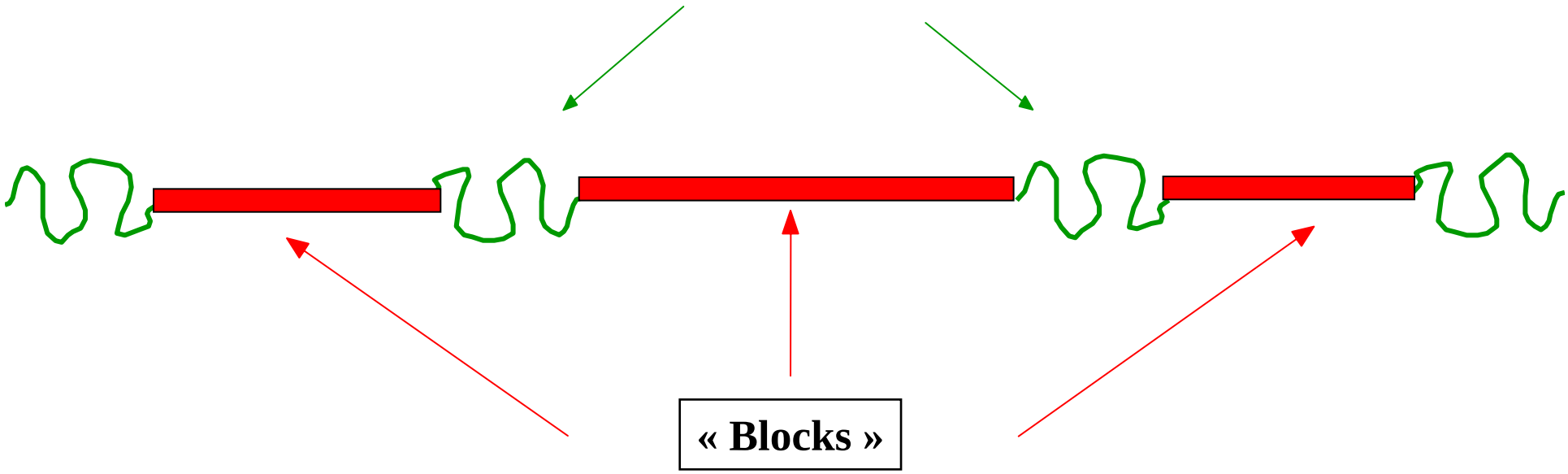
Linkage disequilibrium (LD)

- The non-random association of alleles in the population
- Alleles at neighboring loci tend to co-segregate
- Linkage disequilibrium implies population allelic association

LD around an ancestral mutation

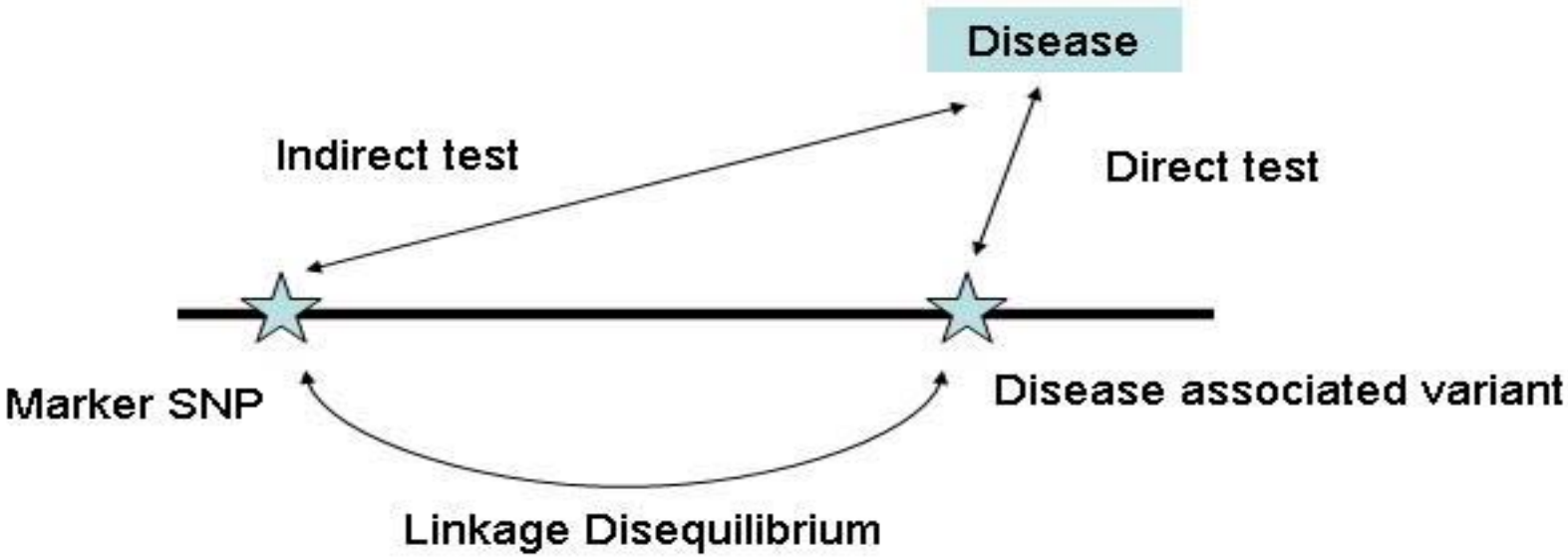


"Hot" spots of meiotic recombination



Each block has a limited number of haplotypes

Genetic Association Testing: *Finding Markers*

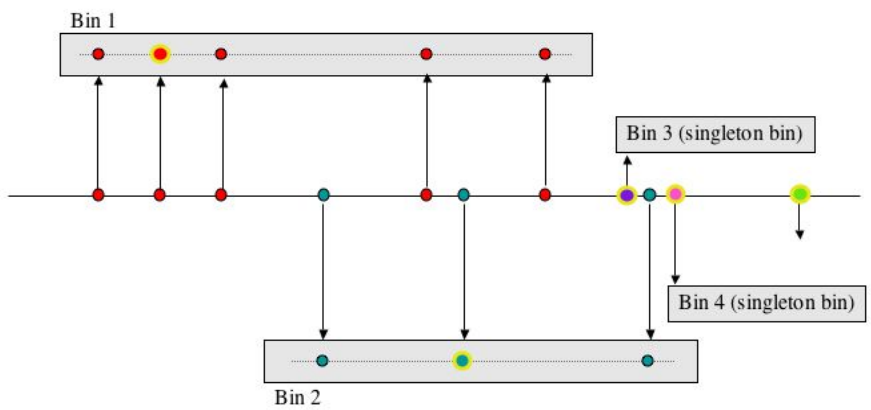


Strategy for SNP Selection

To test all SNPs is presently too costly

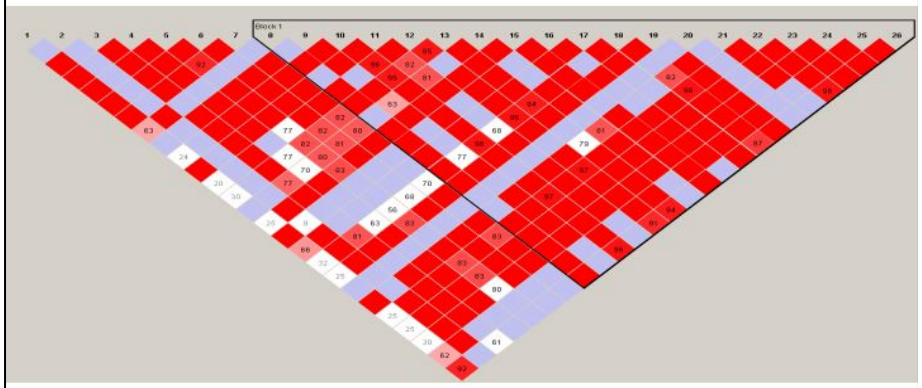
Utilize a strategy that capitalizes on linkage disequilibrium between SNPs

Grouping of SNPs into bins based on pairwise r^2.



Carlson et al. AJHG 74:106 (2004)

Haplotype blocks defined by Gabriel et al
Based on D' values for linkage disequilibrium



www.hapmap.org

Vol 437|27 October 2005|doi:10.1038/nature04226 nature

ARTICLES

A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

International HapMap Consortium, *Nature* 2005; 437:1299-1320.

(www.hapmap.org)

Goal: To construct a haplotype map across the entire genom in 270 individuals (Yoruba trios, Japanese, Chinese and European Caucasian trios)

Phase 1: Completed 03/01/2005

1,000,000 common SNPs (≥ 5%) genotyped: 1 per ~5 kb

Phase 2: Completed 10/28/05

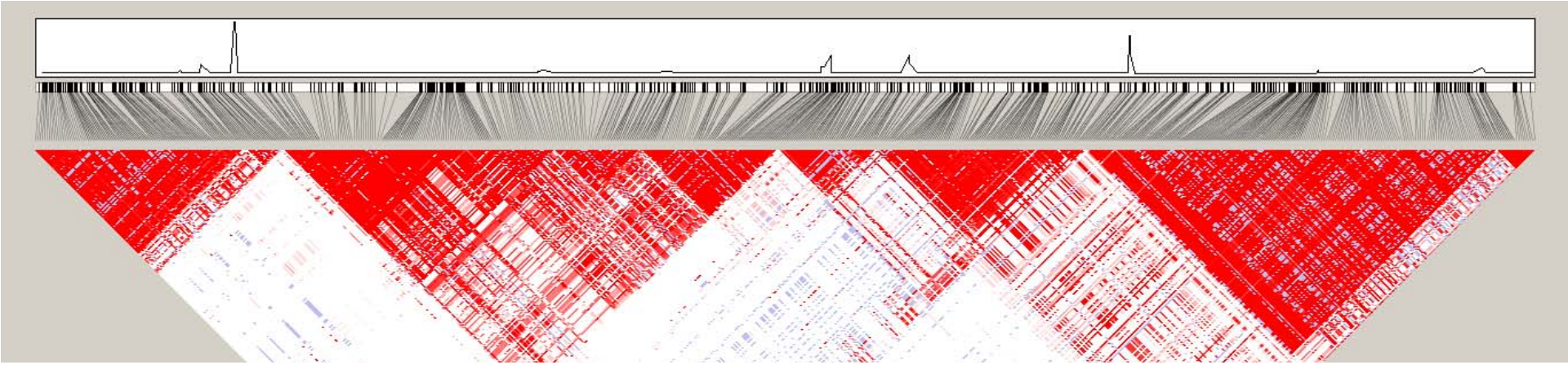
~4,000,000 common SNPs (>5%) genotyped: 1 per ~1.5 kb

A framework for comprehensive candidate gene and genome-wide association studies

Between 500,000 and 1.000,000 for common SNPs (MAF > 5%)

Block Structure in the Genome: *What's it all about?*

Estimated recombination rates: Donnelly lab (Oxford)



Pairwise LD: red is strong D' and LOD > 3.0

Courtesy D Altshuler

SNP500Cancer: Search by SNP - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address https://snp500cancer.nci.nih.gov/snp.cfm Go Links

NCI > CGAP > SNP500Cancer > Search by SNP

Welcome, YEAGERM

Home

Search by Gene/ Chromosome/ Pathway

Search by SNP

Links

Login Tasks

Log Out

SNPs matching: **adh1c-02**

dbSNP ID: **rs1693482**

SNP500Cancer ID: ADH1C-02
Gene: ADH1C
Amino acid change: R272N

dbSNP
NCBI map
Ensembl map
LocusLink

Sequence of Analyzed Amplicon

```
CTATCTGTTGTTATGGGCTGTAAAGCAGCTGGAGCAGCCAGAATCATTGCTG
TGGACATCAAYAAGGACAAATTTGCAAAGGCTAAAGAGTTGGGTGCCACTGA
ATGCATCAACCCTCAAGACTACAAGAAACCCATTCAGGAAGTGCTAAAGGAA
ATGACTGATGGAGGTGTGGATTTTTCGTTTGAAGTCATCGGTC(A/G)GCTT
GACACCATGGTATGWWCCRTGACATGCCCTGAAATTTCTGCCTCTGCAACCT
GGAGGATRCATTTAGGCAGYAGAATATACGTATTATGTATAAAGGATATTTT
TAATGATGAATGGAAATTTCCCRTCATCTTTTTGTTACCTGGCTTGTTTAAT
TTA
```

Frequency Data (102 anonymized subjects):

Total Completed	Genotypic			Allelic	
	AA	AG	GG	A	G
101	14/101 **(0.139)**	34/101 **(0.337)**	53/101 **(0.525)**	62/202 **(0.307)**	140/202 **(0.693)**
View Subpopulation Frequencies					

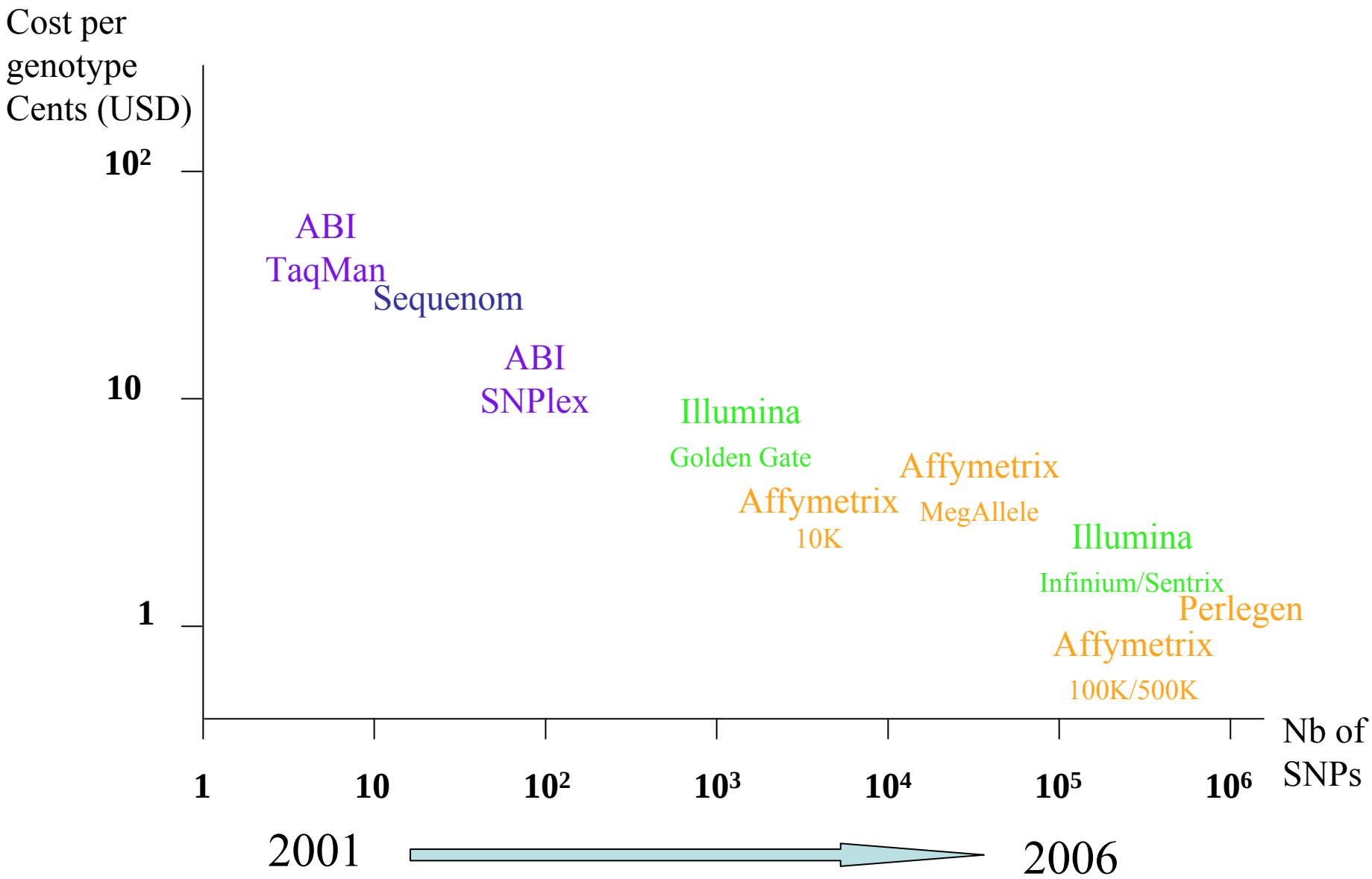
Assays - these frequency results were validated on the following platforms - click to view primers, probes, and conditions:
Sequencing TaqMan

Search by SNP

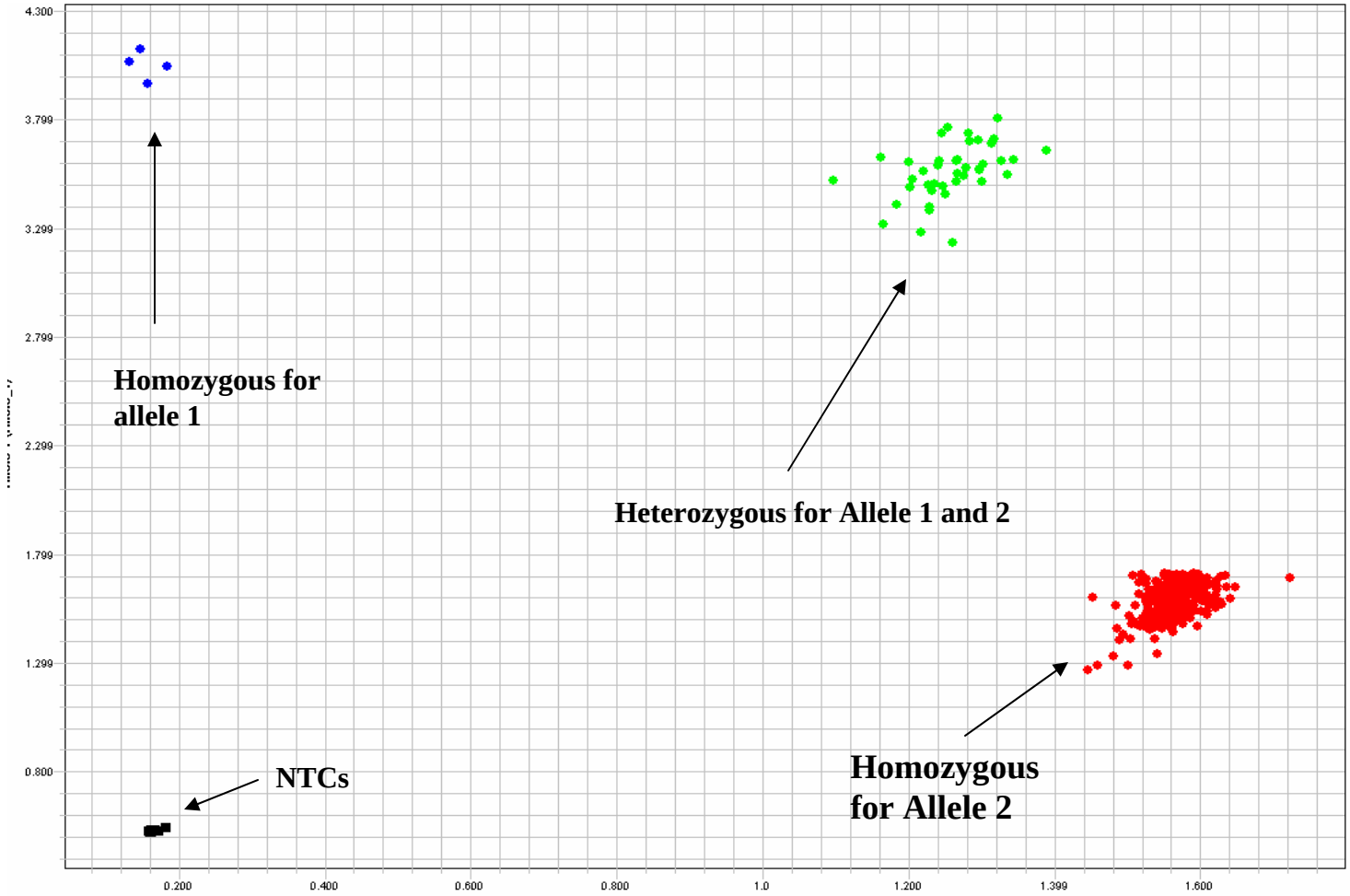
Internet

Start 5:21 PM

Progress in genotyping technology



Excellent Taqman genotyping assay



Extreme Genotyping

Genotype thousands of markers in one reaction

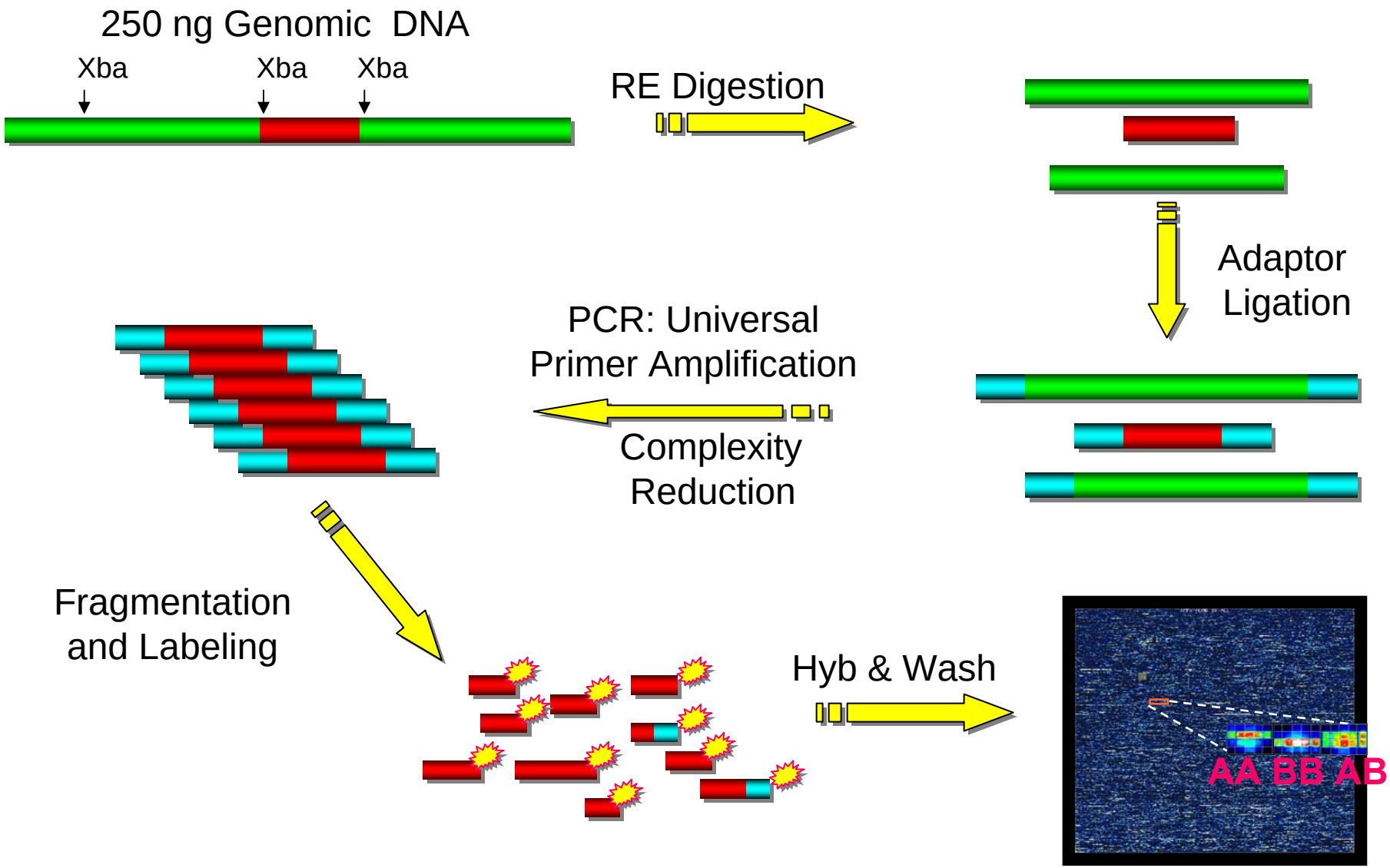
- **Preset for**
  - **Candidate Genes**
  - **Across Genome or Chromosome**

Simplify Genome

Highly Parallel Analysis

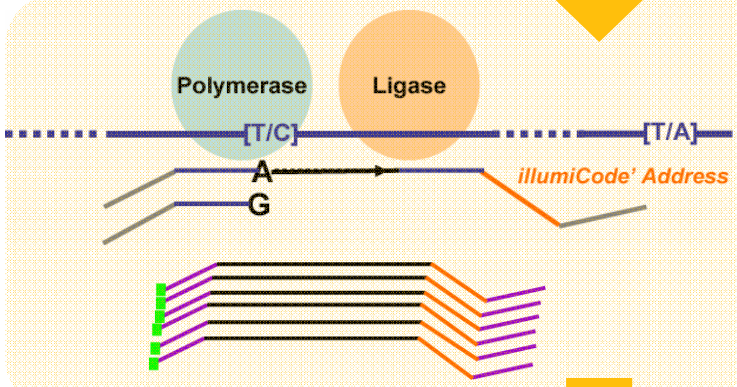
You get what they give you…..

Affymetrix GeneChip® Mapping Assay



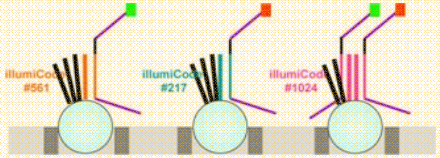
Illumina GoldenGate™ Assay

Genomic DNA

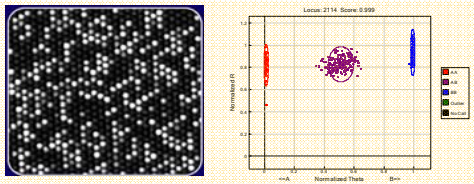


GoldenGate™ Assay
Multiplexed at 1536 loci

Universal Amplification



Sentrix™ BeadArray Hybridization



Automated Array Scan & Analysis
Genotypes with quality score

Human-1 Genotyping BeadChip

25mm x 82.5mm

- Proven BeadArray™ technology
  - 100% QC on 100% of arrays
  - Average 30-fold redundancy
- Exon-centric content emphasis
- >100,000 SNPs
- Flexible BeadChip design
  - High density architecture
  - Easily configured for different content and sample numbers

2006 What is Available for Whole Genome Scans

Coverage analysis based on HapMap II Data

Build 20 MAF $\geq$5%, $r^2 \geq 0.8$ (pair-wise)

		CEU	YRI	JPT/CHB
Illumina	**HumanHap300**	**80%**	**35%**	**40%**
Illumina	**HumanHap500**	**91%**	**58%**	**88%**
Affymetrix*	**500k Mapping**	**63*%**	**41%**	**63%**

77% (with 50k MegA)

Issues for GWAS Genotyping

- Establishing Pipeline
- Analytic Framework
  - Data management
  - Quality Control
- Platform
- Genomic Coverage
- Changing Costs

NIH-Wide Efforts: Genes & Environment Initiative

- RFAs Issued
  - Genotype Facilities (3)
  - Analysis Coordinating Centers (1)
  - Study Investigators (9 or 10 in FY07)
- NCI Actively Involved on Environmental RFAs
  - Technology Development (5)
  - NCI Leadership on Diet, Physical Activity

Candidate Genes and Loci
vs
Genome-wide Association Studies

Discovery vs Characterization

Never leave candidate genes behind…..

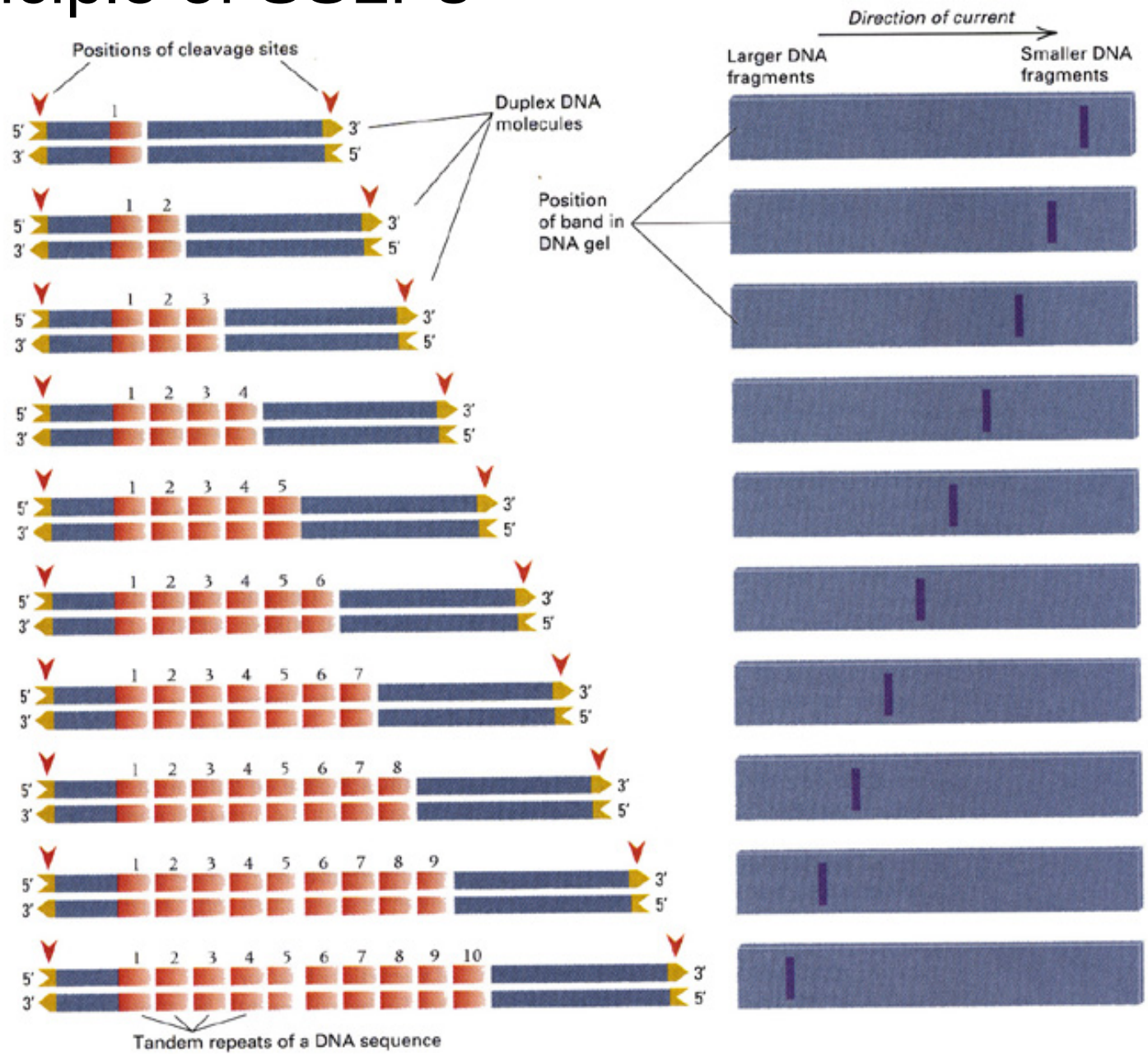
Follow-up of GWAS: Steps to Clinical Implementation

- Fine mapping of notable regions
- Functional determination of causal variants
- Design issue for analysis in clinical studies
  - Population-based studies
  - Sequence of clinical studies
- Validation criteria

Types of Polymorphisms II

- Simple Sequence Length Polymorphisms (SSLPs)
  - Mini-satellites=VNTRs (variable number of tandem repeats)
    - Repeat size is large- e.g., 25 to 100s of bp
  - Micro-satellites=STRs (simple tandem repeats)
    - Repeat size is small- e.g., 2 to 7 bp
    - Used for Linkage mapping of highly penetrant genes

Principle of SSLPs



Types of Polymorphisms III

- **Copy Number of Polymorphisms**

  Regional “repeat” of sequence

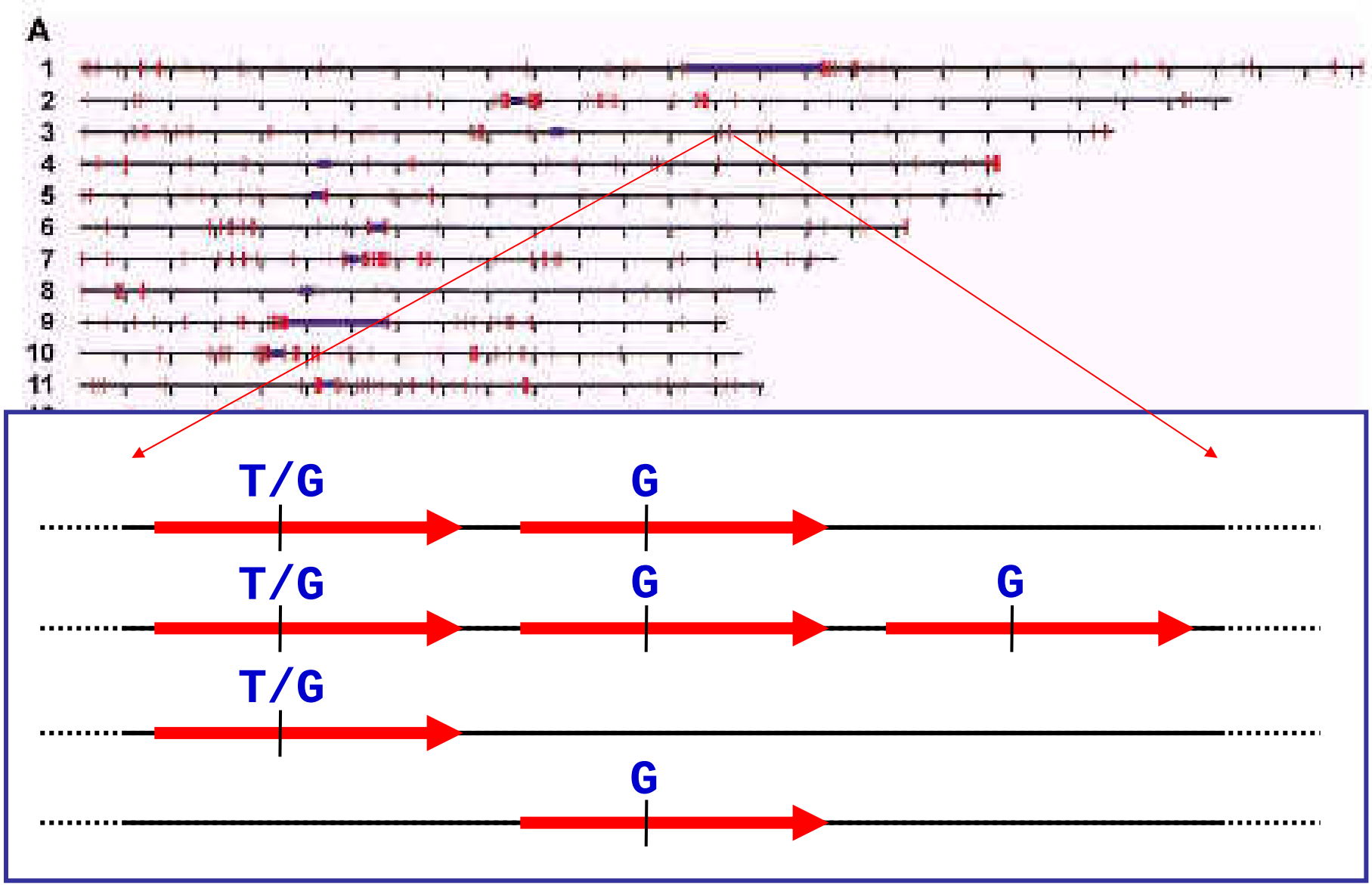
  10s to 100s kb of sequence

  Estimate of >10% of human genome

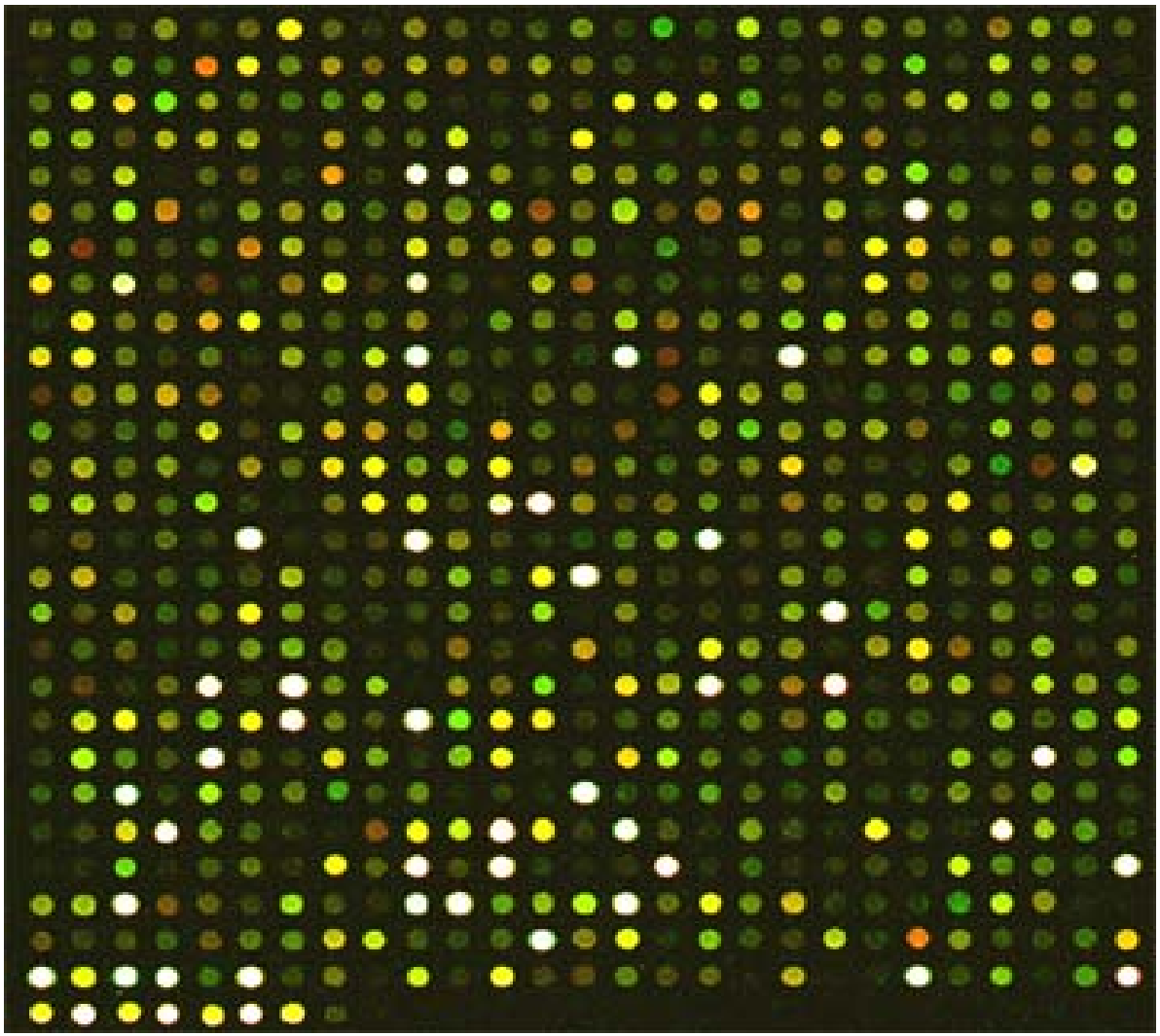
  Multi-copy in many individuals

- **International Database**

MSV: Multi-Site Variants

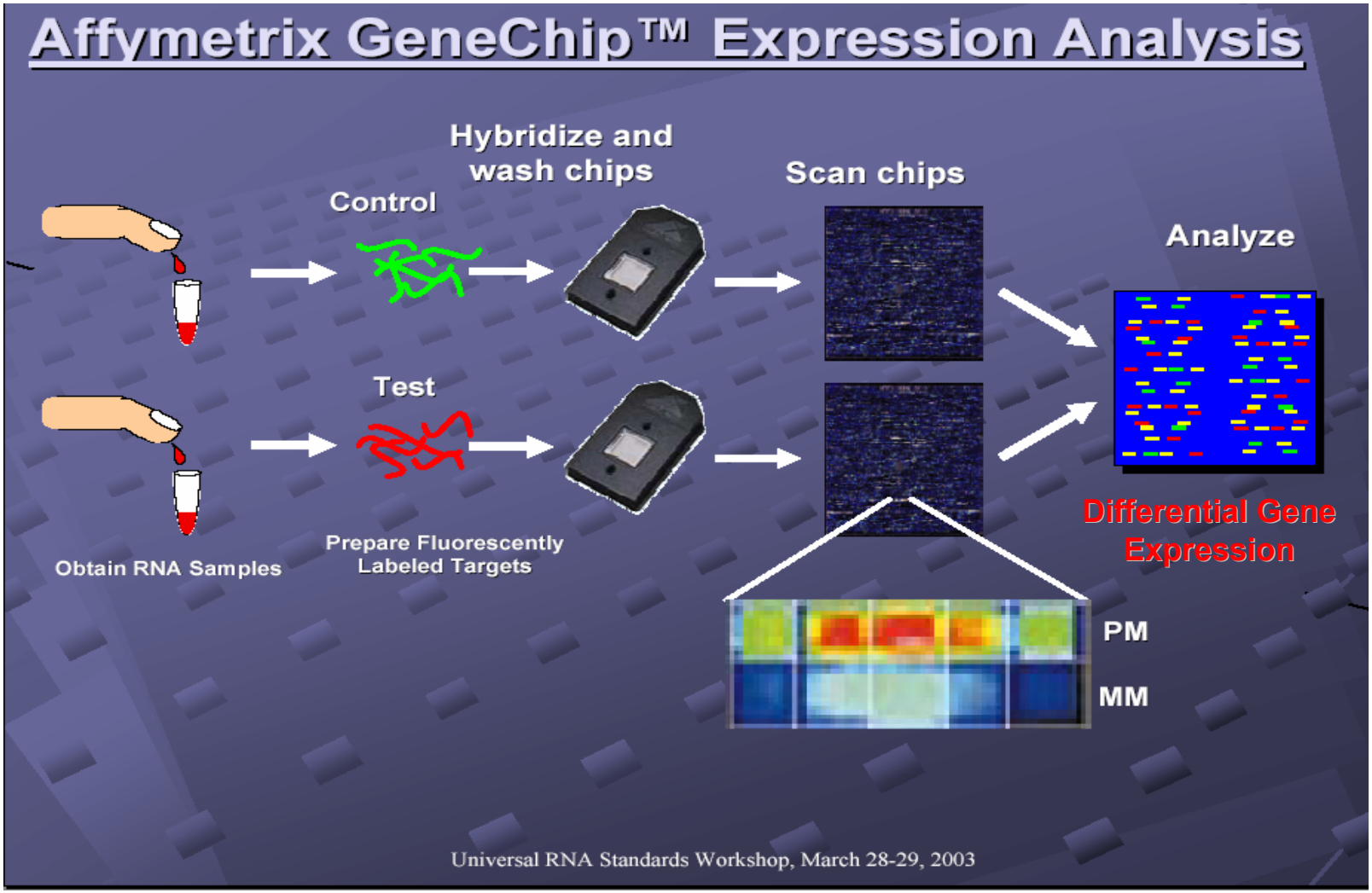


Looking at the Transcriptosome

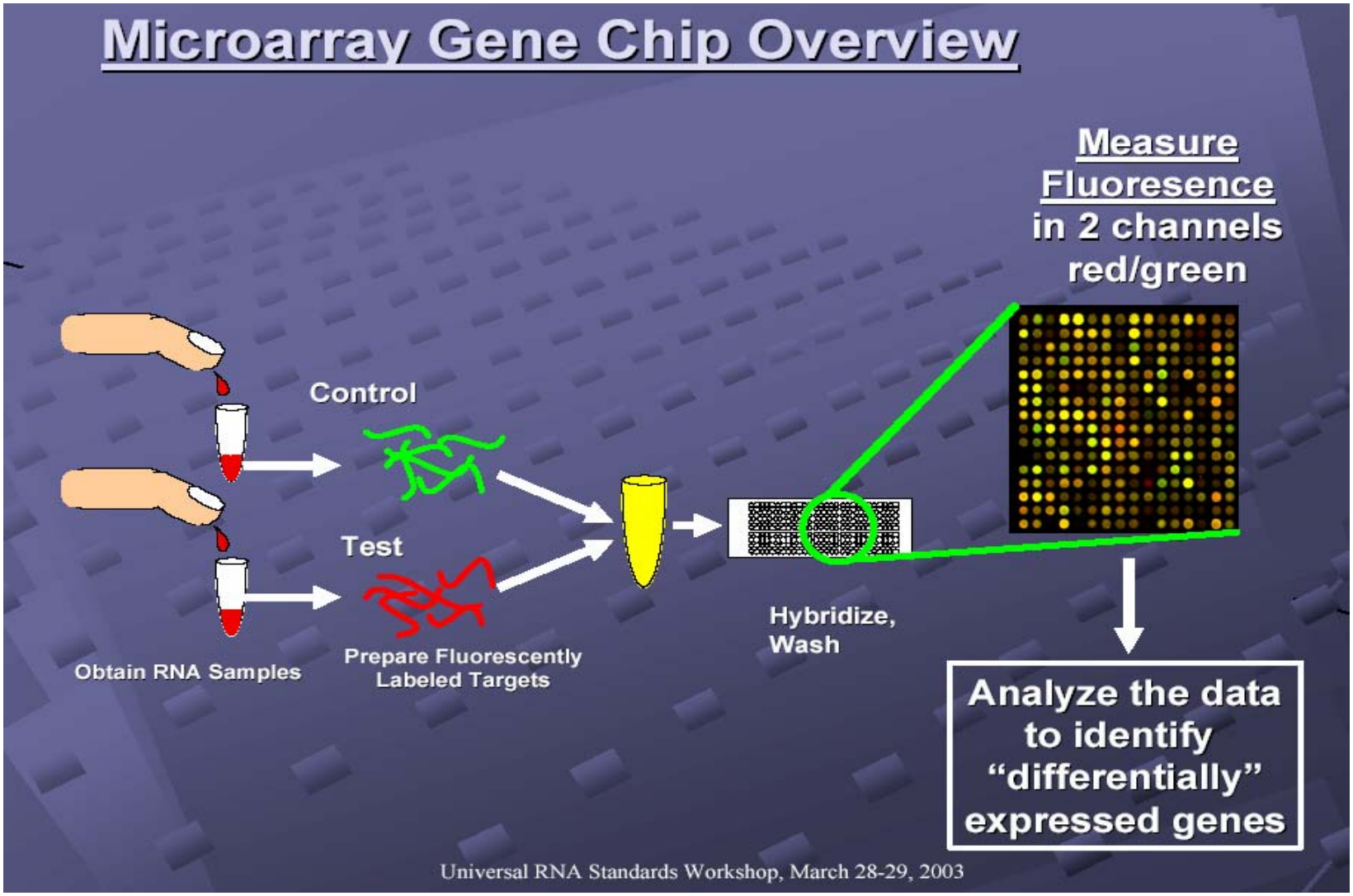


Genes & Transcripts

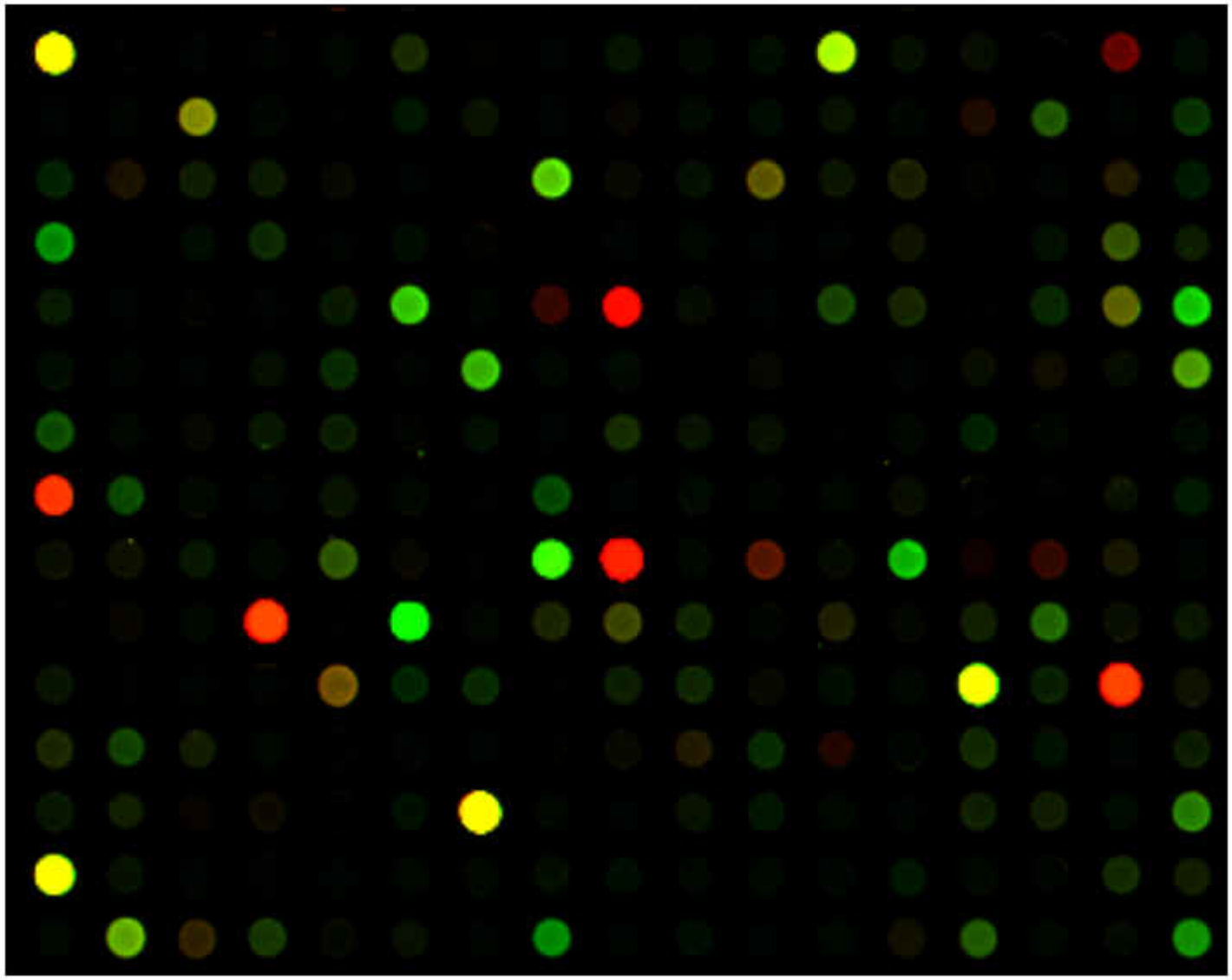
Single Color Arrays



Two Color Arrays

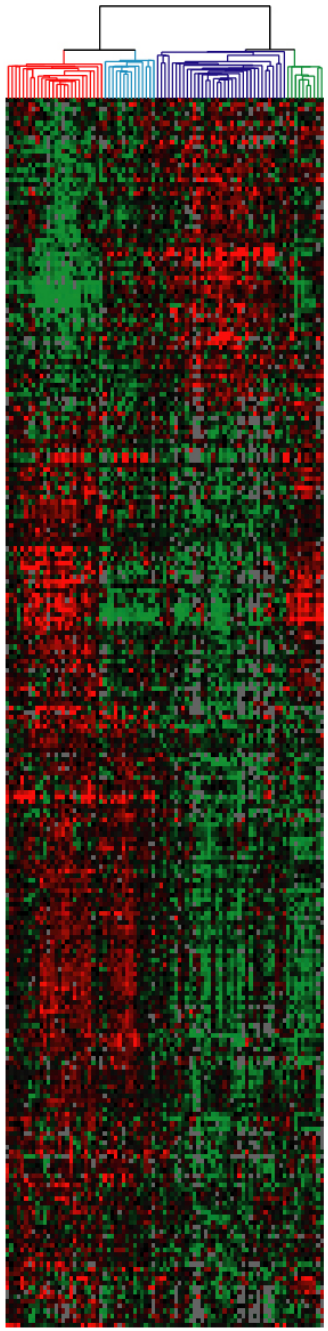


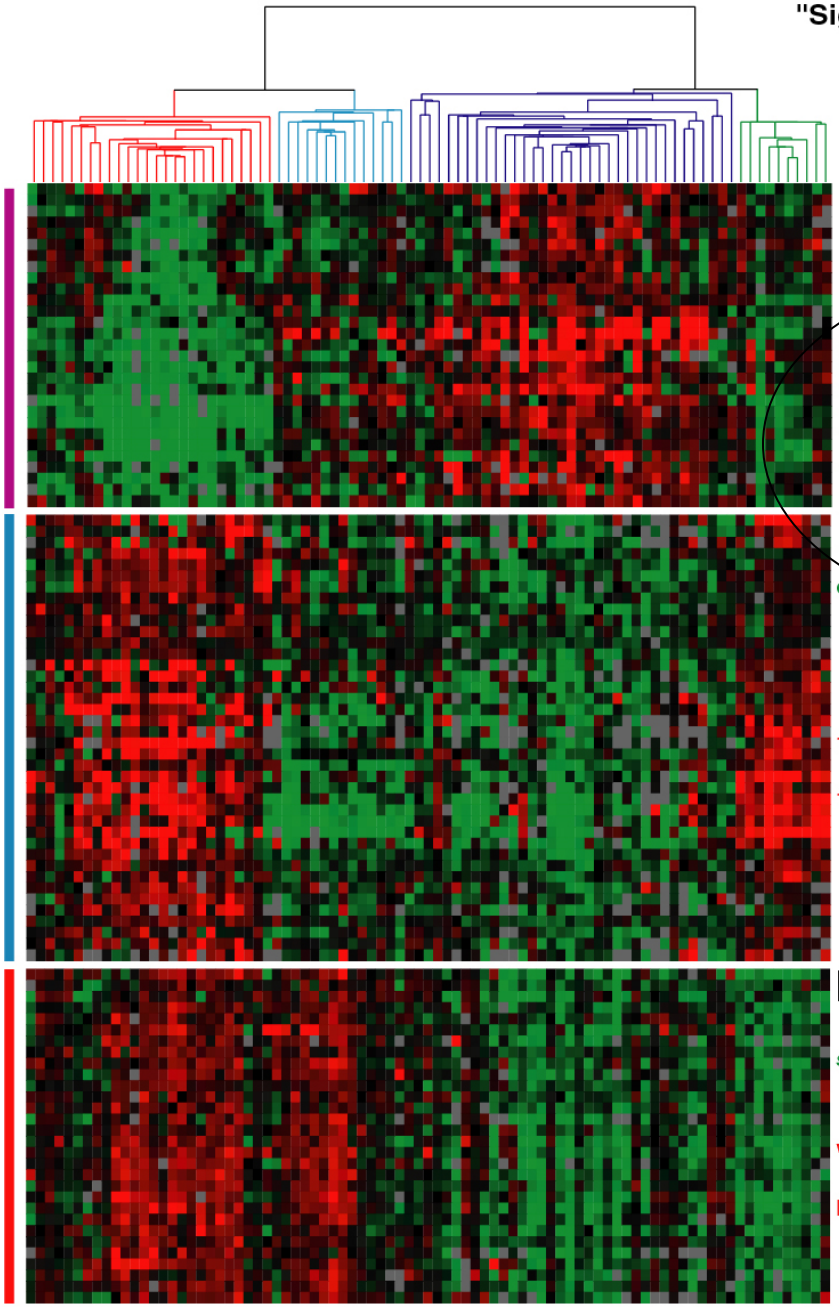
John Quackenbush Harvard University



Agilent Two Color Oligo Array Exhibiting Differential Gene Expression

CD, SD & E





"Significance Analysis of Microarrays"
SAM264 survival set
with genes for
DNA sequence analysis

Luminal/ER+ Gene set

estrogen receptor 1
GATA binding protein 3
X-box binding protein 1
hepatocyte nuclear factor 3, alpha
RAS-like, estrogen-regulated, growth-inhibitor
retinoic acid receptor responder (tazarotene induced) 3

Basal Epithelial Gene set

epidermal growth factor receptor

forkhead box C1

frizzled homolog 7 (Drosophila)

Proliferation Gene set

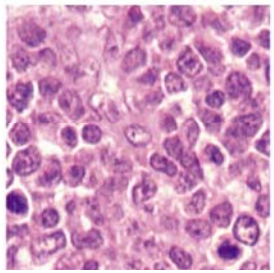
serine/threonine kinase 15

v-myb myeloblastosis viral oncogene homolog avian)-like 2

polo (Drosophia)-like kinase

Sorlie et al 2001

Dissecting a Cancer into Molecularly and Clinically Distinct Subgroups by Gene Expression Profiling



<u>Diffuse large B cell lymphoma</u>

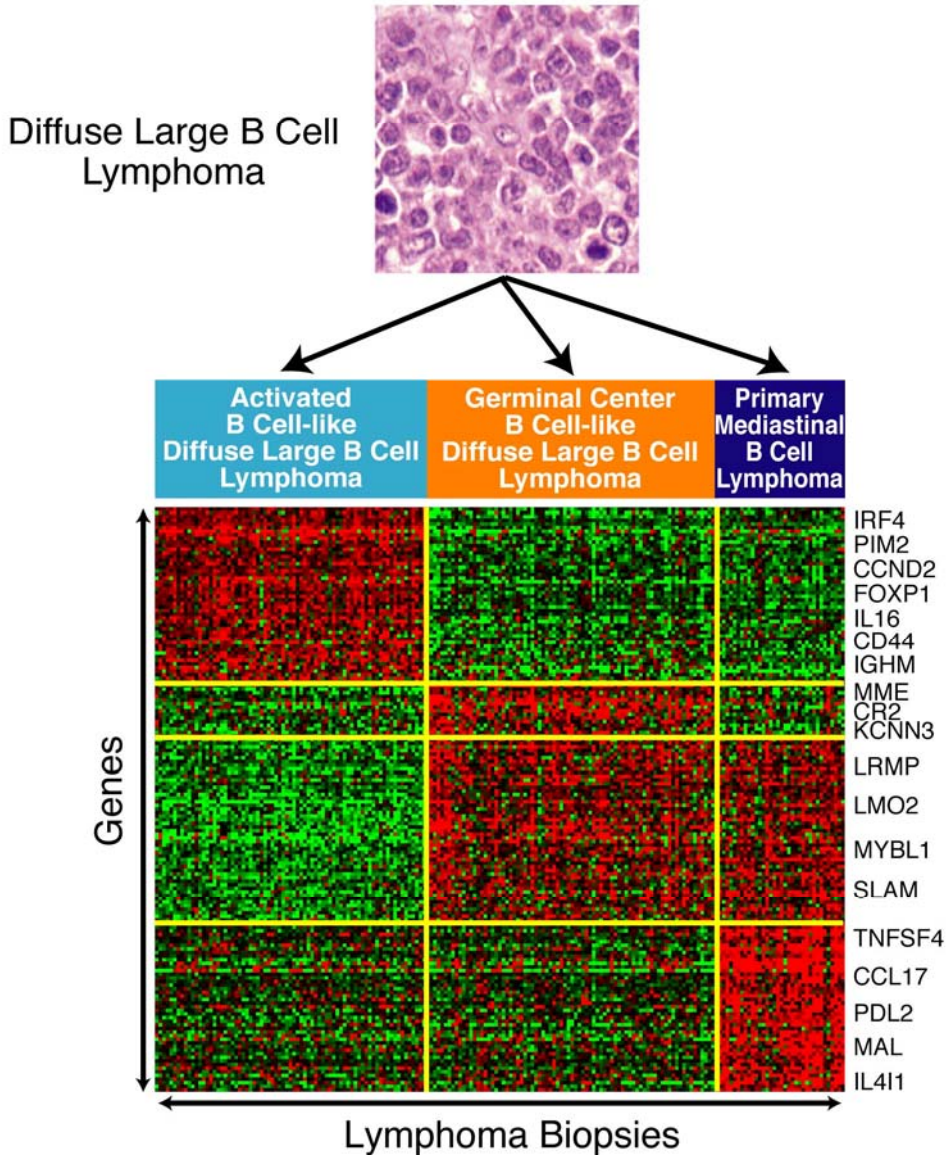
40% of Non-Hodgkin lymphomas

~23,000 new diagnoses/yr

~40% cure rate

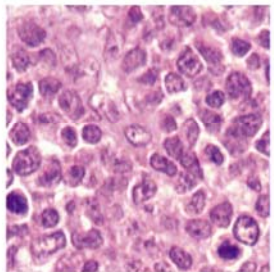
~10,000 deaths/yr

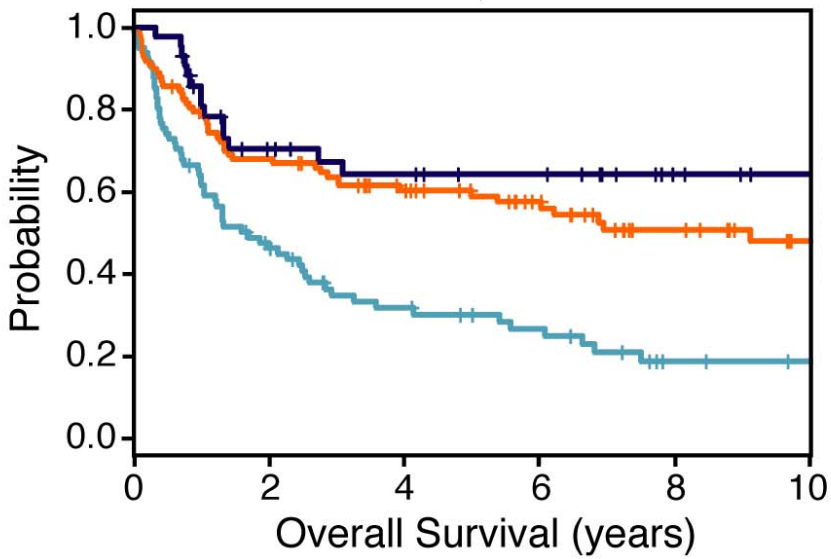
Dissecting a Cancer into Molecularly and Clinically Distinct Subgroups by Gene Expression Profiling



Dissecting a Cancer into Molecularly and Clinically Distinct Subgroups by Gene Expression Profiling

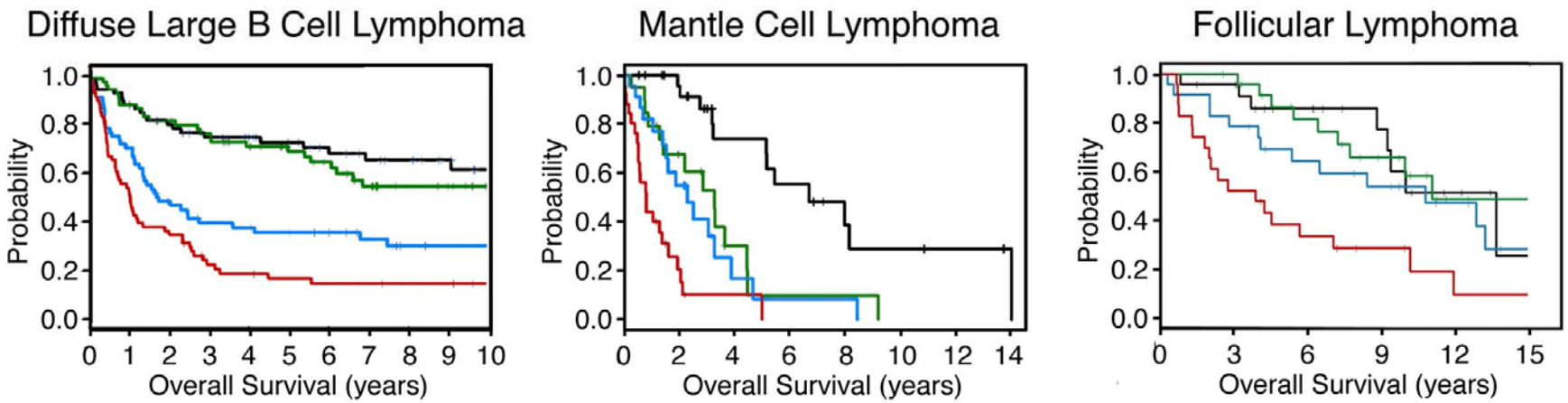
Diffuse Large B Cell Lymphoma

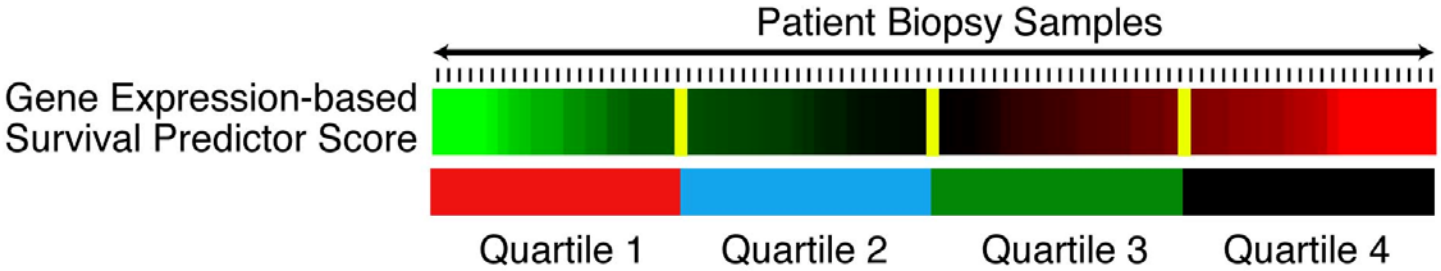




	5-year survival
PMBL	64%
GCB DLBCL	59%
ABC DLBCL	30%

Survival Prediction Based on the Gene Expression Profile of the Diagnostic Biopsy

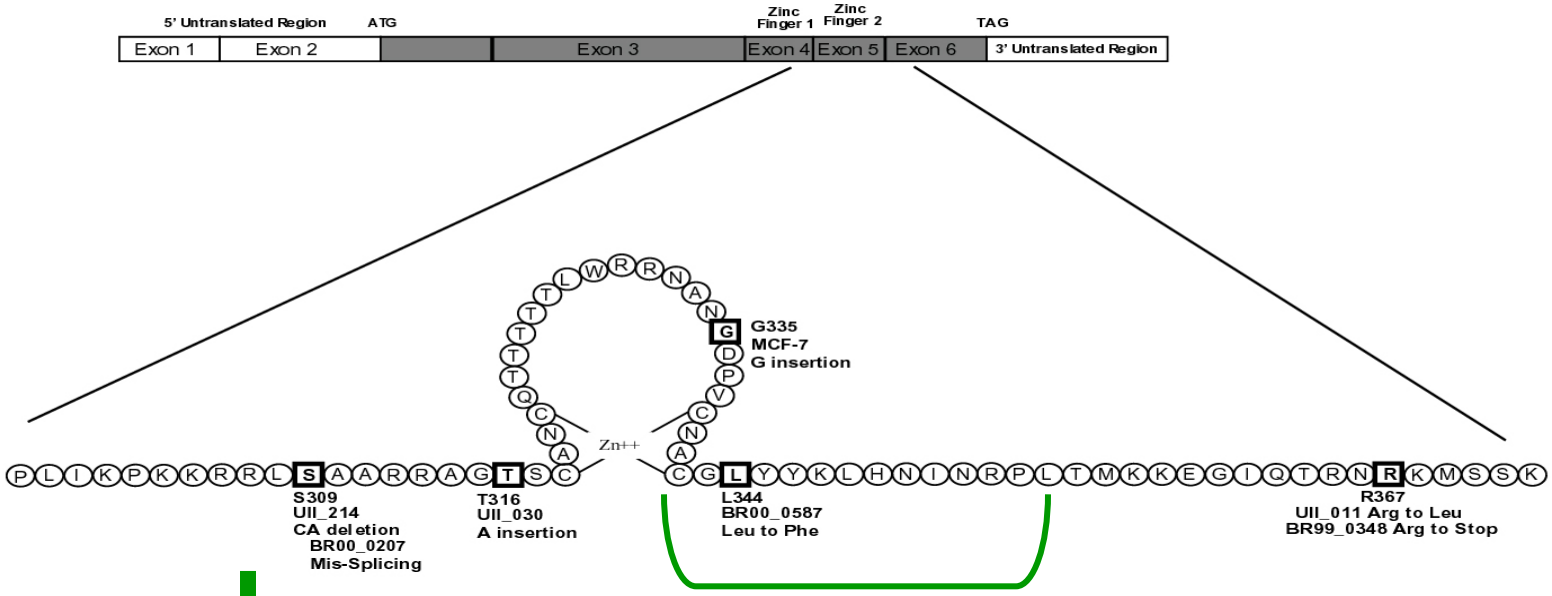




Mutation of GATA3 in Human Breast Tumors

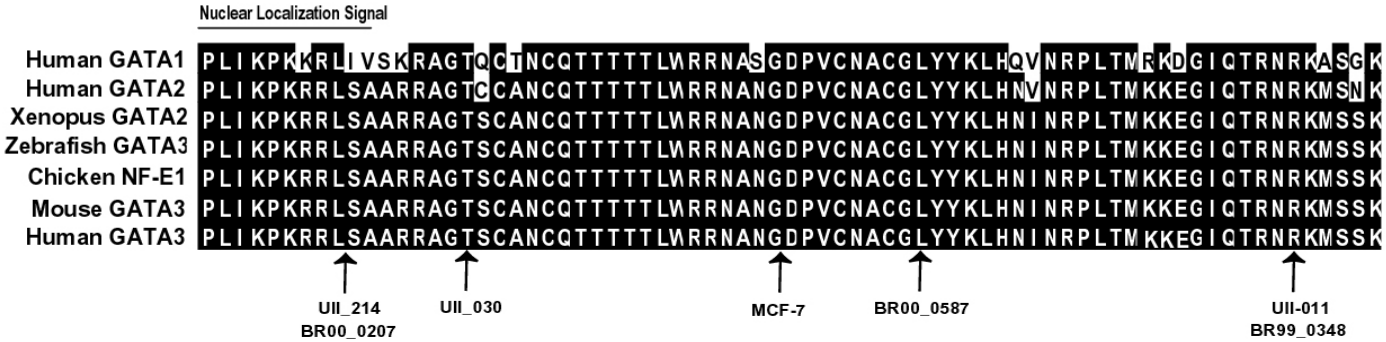
(Usary et al., Oncogene, 2004)

6/70 ER+ tumors mutated (0/35 ER- mutated)



Deletion of 343-355 destroys DNA binding and transactivation (303-347 is required) Yang et al.

Nuclear Localization is 249-311: Yang et al.



Comprehensive Analysis

Re-sequence analysis of germ-line and tumor DNA

- Discover catalog of driver and hitchhiker mutations

Analysis of Expression Profiling in Tumor

Analysis of Copy Number Changes in Tumor

- Loss of Heterozygosity
- Amplification of Region

Public Resource for

- Discovery
- Validation

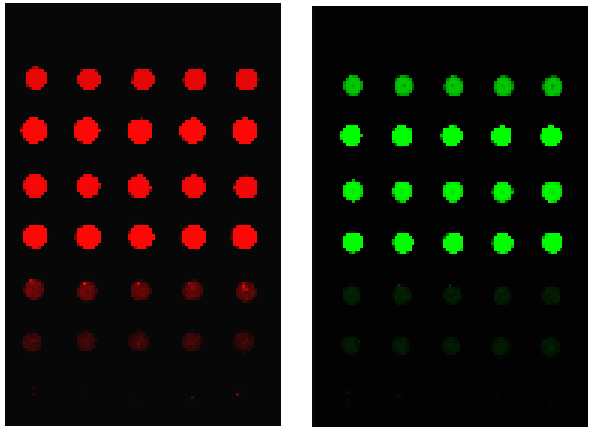
Proteomics

- Large scale study of proteins
  - Structure
  - Function
- Differs cell to cell
- >1 million proteins
  - Post-translational modification
  - Alternative splicing
- More challenging for analysis
  - Stability and quality of sample

Analysis of Proteomics

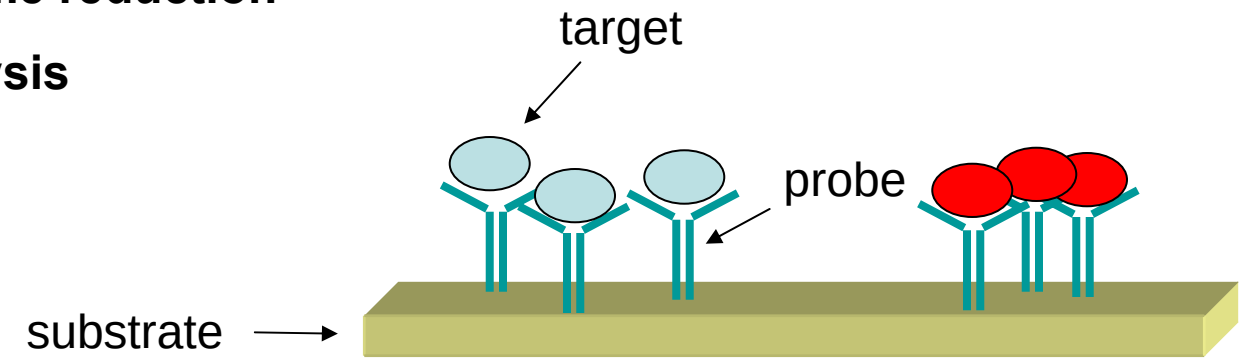
- Separation
- Identification
- Quantification
- Sequence Analysis
- Structural assessment
- Interaction
- Protein modification
  - Phosphoproteomics
  - Glycoproteomics

Why microarrays for proteins?



<u>Advantages</u>

- **multiplexing and miniaturization**
  - **↑ throughput**
  - **sample volume reduction**
  - **parallel analysis**



Protein Micro-array Applications . . .

- **Protein - protein interactions**
- **DNA - protein interaction**
- **Small molecule screening**
- **Protein profiling**
- **Antibody characterization**
- **Enzyme-substrate analysis**

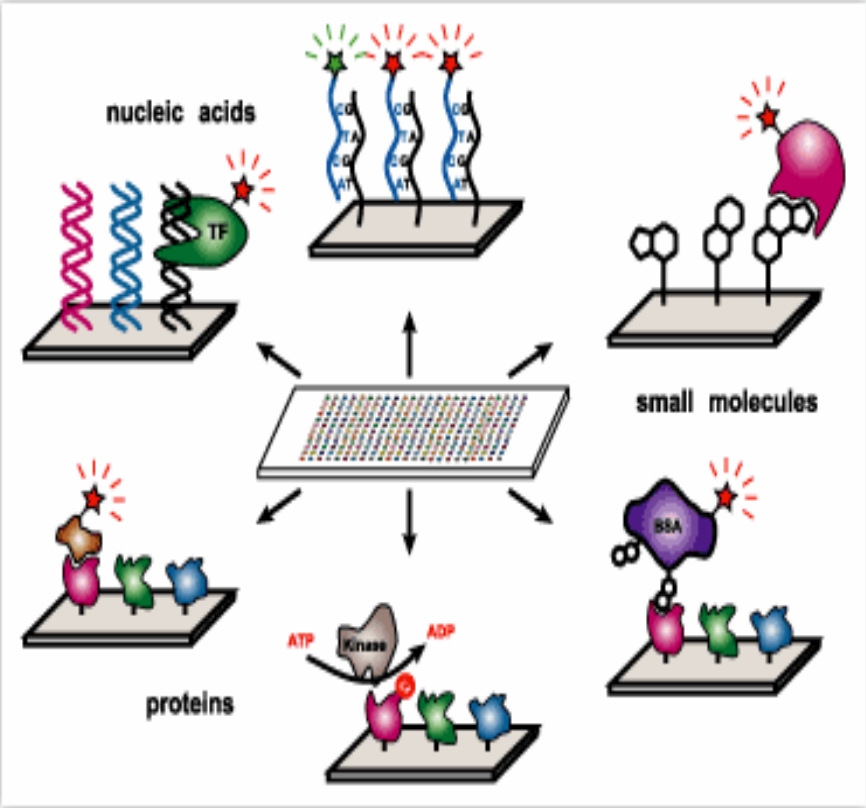


Image courtesy of Dr. Gavin MacBeath, Bauer Center for Genomics Research, Harvard University

Proteomic Techniques

One- and two-dimensional gel electrophoresis (Mass and isoelectric point)
3-D structure
- X-ray crystallography
- Nuclear magnetic resonance

2-D electrophoresis
- Tandem mass spectrometry
- Reverse phase chromatography or 2-D electrophoresis

Mass spectrometry
- MALDI-TOF for peptide mass fingerprinting
- SELDI-TOF chip analysis

Protein-protein & Protein-DNA interactions
- Affinity chromatography
- Yeast two hybrid
- Fluorescence resonance energy transfer (FRET)
- Surface Plasmon Resonance (SPR)

X-ray Tomography
Software based image analysis

Pointers

- We are searching for genetic markers
- Function (e.g., plausibility) comes later
- We are capitalizing on ancient relationships between common genetic variants
- SNPs are for common variants
- Sequencing is for rare variants

What is down the road?

1-2 Year Forecast

- **Cheaper and denser SNP technologies**
  - **Better coverage of genome but**
    - ***Power vs coverage….***

3-6 Year Forecast

- **Whole Genome Sequencing**
  - **Replace SNPs**
  - **Magnification of Challenge of Confidentiality**
  - **Challenge to Epidemiologic Rigor**

Search for Genetic Contribution to Complex Diseases

Well positioned for

- **Common SNPs (>5%)**
- **High throughput technology**

Not as well positioned for

- **Uncommon variants**
- **Structural variants (copy number variants)**
- **Populations not in the "BIG 3"**
  - **CEU, Yoruba, East Asia**

What Tools Do We Have?

Extensive data base of common SNPs (MAF>5%)

Technologies for small to large (1 to 10^6 SNPs)

Analytical programs for simple analyses

- **Main effect**
- **Population structure**

2007

What Tools Do We Need?

Extensive data base of <u>un</u>common SNPs (MAF<5%)

***Flexible* Technologies for small to large (1 to 10^6 SNPs)**

- **Targeted to different populations**

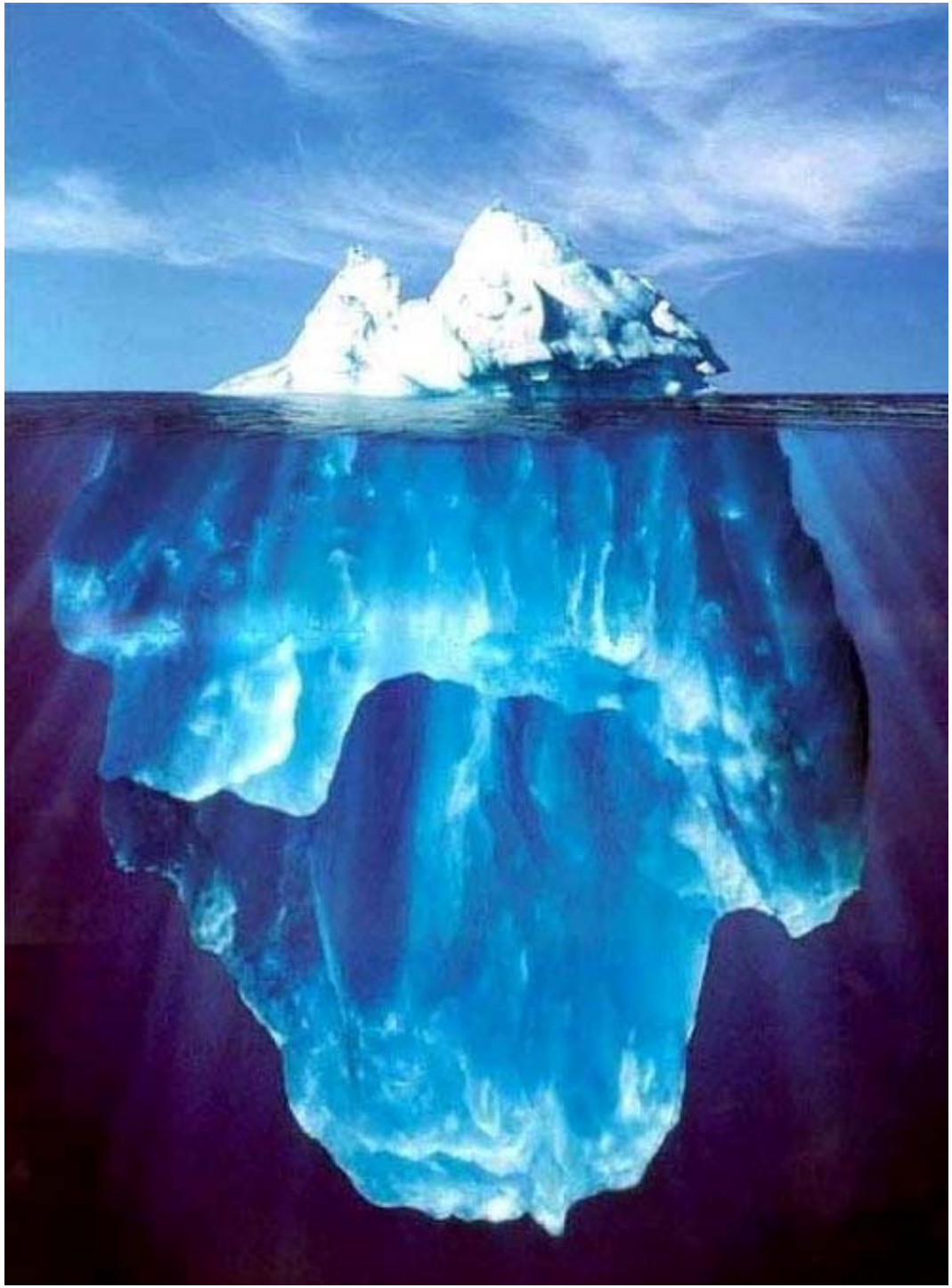
Analytical programs for complex analyses

- **Gene-gene interaction**

Environmental measurements

Complete genome sequence technology

Post 2007



Genomics
and
Public Health

Navigating the deep waters of genetics…..

More complex than we Imagined……

Still, worth it……

The Future of Medicine??

Major Challenges

Genome

Patriotism

Personalized Health