Introduction to Human Genomics

Bioinformatics and Disease Gene Identification

Peter K. Rogan, Ph.D.
Laboratory of Human Molecular Genetics
Children's Mercy Hospital
Schools of Medicine & Computer Science and Engineering
University of Missouri- Kansas City

http://www.sce.umkc.edu/~roganp

Rationale for the human genome project

- To acquire fundamental information concerning our genetic make-up which will further our basic scientific understanding of human genetics and of the role of various genes in health and disease.

Steps in determining the genome sequence

- High-resolution genetic maps were constructed - achieved by 1994
- These were used as a *framework* for constructing high-resolution physical maps - large clone contigs had been assembled for much of the genome by 1999
- Large-scale DNA sequencing was very much underway, with the first draft appearing in 2000
- Finished sequence is available. However, certain regions have been ignored because they cannot be unequivocally assembled

Medical and scientific benefits

Knowing the structure of each human gene will enable:

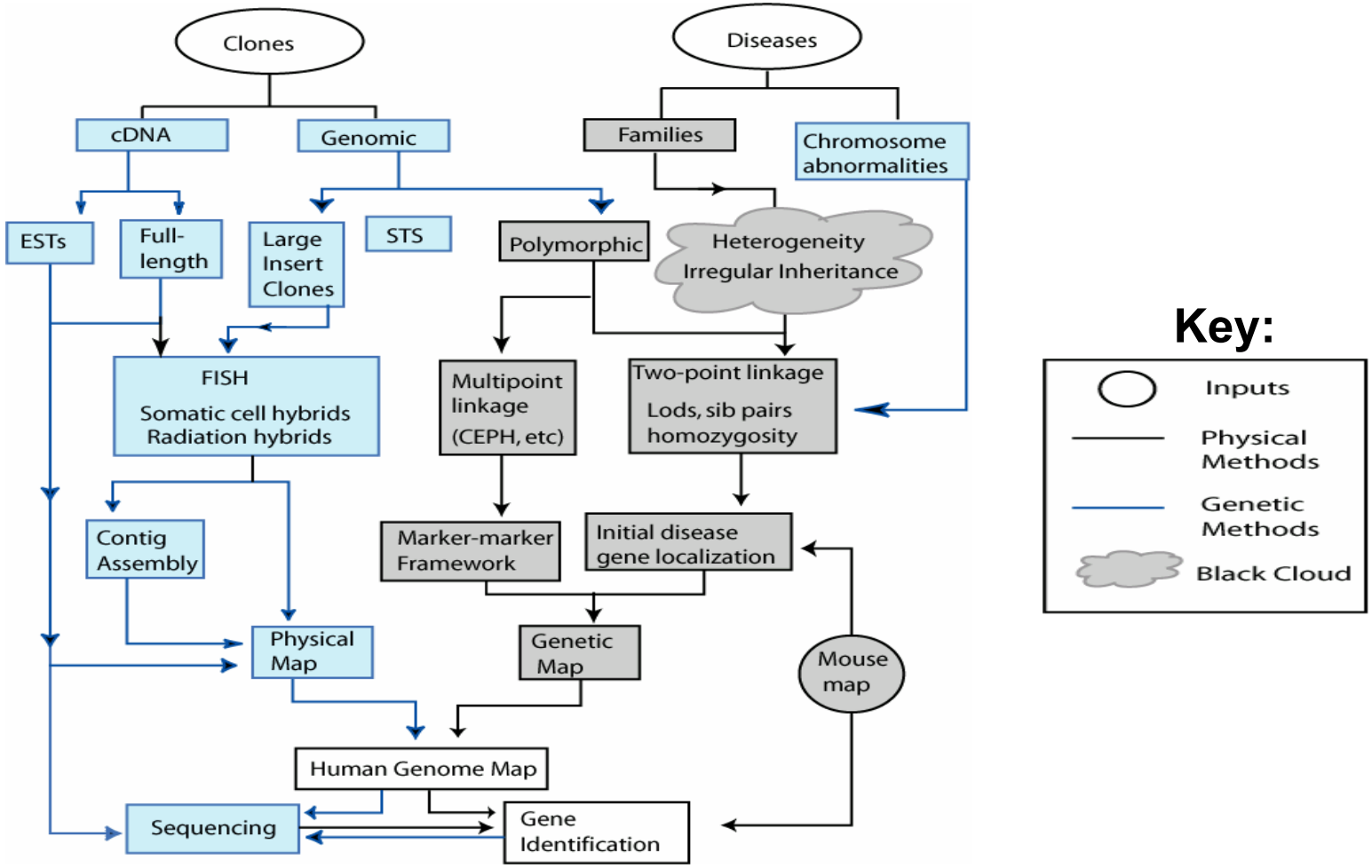
- More comprehensive prenatal and presymptomatic diagnoses of disorders in individuals judged to be at risk of carrying a disease gene
- Information on gene structure will also be used to explore how individual genes function and how they are regulated, which will provide sorely needed explanations for biological processes in humans
- A framework for developing new therapies for diseases, in addition to simple gene therapy approaches
- Better molecular diagnostics which will alter radically the current approach to medical care, from one of treating advanced disease to preventing disease based on the identification of individual risk.

Challenges

- Difficulties in understanding precisely and comprehensively how some genes function and are regulated.
- Single-gene disorders, which should be the easiest targets for developing novel therapies, are very rare.
- The most common disorders are multifactorial, and they are more difficult to diagnose and treat via genetic approaches.

Different Types of physical map can be used to map the human nuclear genome

Type of Map	Examples/ methodology	Resolution
Cytogenetic	Chromosome banding maps	An average band has several Mb of DNA
Chromosome breakpoint maps	Somatic cell hybrid panels containing human chromosome fragments derived from natural translocation or deletion chromosomes Monochromosomal radiation hybrid (RH) maps Whole genome RH maps	Distance between adjacent chromosomal breakpoints on a chromosome is usually several Mb Distance between breakpoints is often many Mb Resolution can be as high as 0.5 Mb
Restriction map	Rare-cutter restriction maps, e.g. *Not*l maps	Several hundred kb
Clone contig map	Overlapping YAC clones Overlapping contig clones	Average YAC insert has several hundred kb of DNA Average cosmid insert is 40 kb
Sequence-tagged site (STS) map	Requires prior sequence information from ordered clones so that STSs can be ordered	A desired goal is an average spacing of 100 kb
Expressed sequence tag (EST) map	Requires cDNA sequencing then mapping cDNAs back to other physical maps	Highest possible average spacing is ~40 kb
DNA sequence map	Complete nucleotide sequence of chromosomal DNA	1 bp



Major scientific strategies and approaches being used in the Human Genome Project

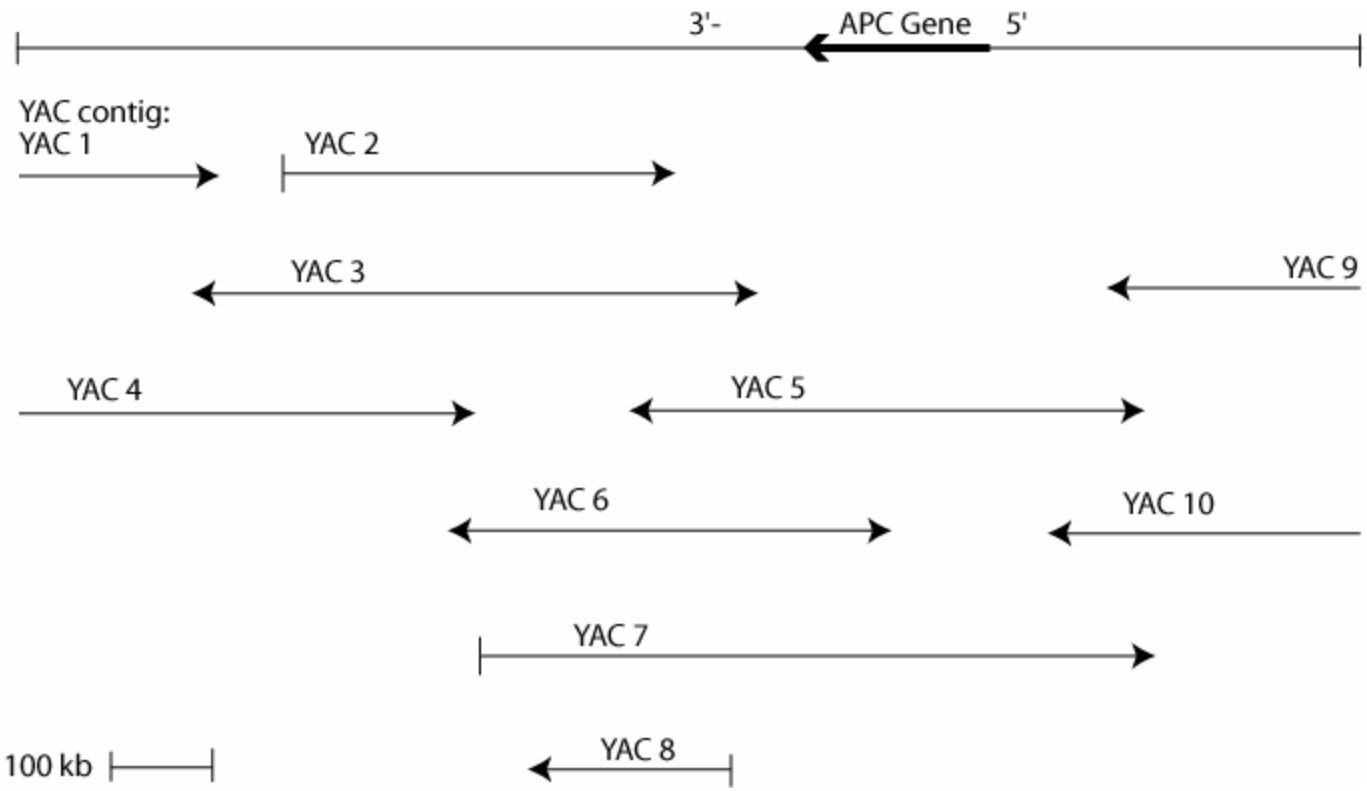
The major scientific thrust of the Human Genome Project began with the isolation of human genomic and cDNA clones (by cell-based cloning or PCR-based cloning). These are then used to construct **high-resolution genetic and physical maps** prior to obtaining the ultimate physical map, the complete nucleotide sequence of the 3300 Mb nuclear genome. *Inevitably, the project interacts with research on mapping and identifying human disease genes.* In addition, projects include studying genetic variation; genome projects for model organisms, and research on ethical , legal and social implications. The data produced are being channeled into mapping and sequence databases permitting rapid electronic access and data analysis. Abbreviations: EST, expressed sequence tag; STS sequenced tagged site.

Human Gene and DNA Segment Nomenclature

The nomenclature used is decided by the HUGO nomenclature committee. Genes and pseudogenes are allocated symbols of usually two to six characters; a final P indicated a pseudogene. For anonymous DNA sequences, the convention is to use D (+DNA) followed by 1-22, X or Y to denote the chromosomal location, then S for a unique segment, Z for a multilocus DNA family, and finally a serial number. The letter E following the number for an anonymous DNA sequence indicates that the sequence is known to be expressed.

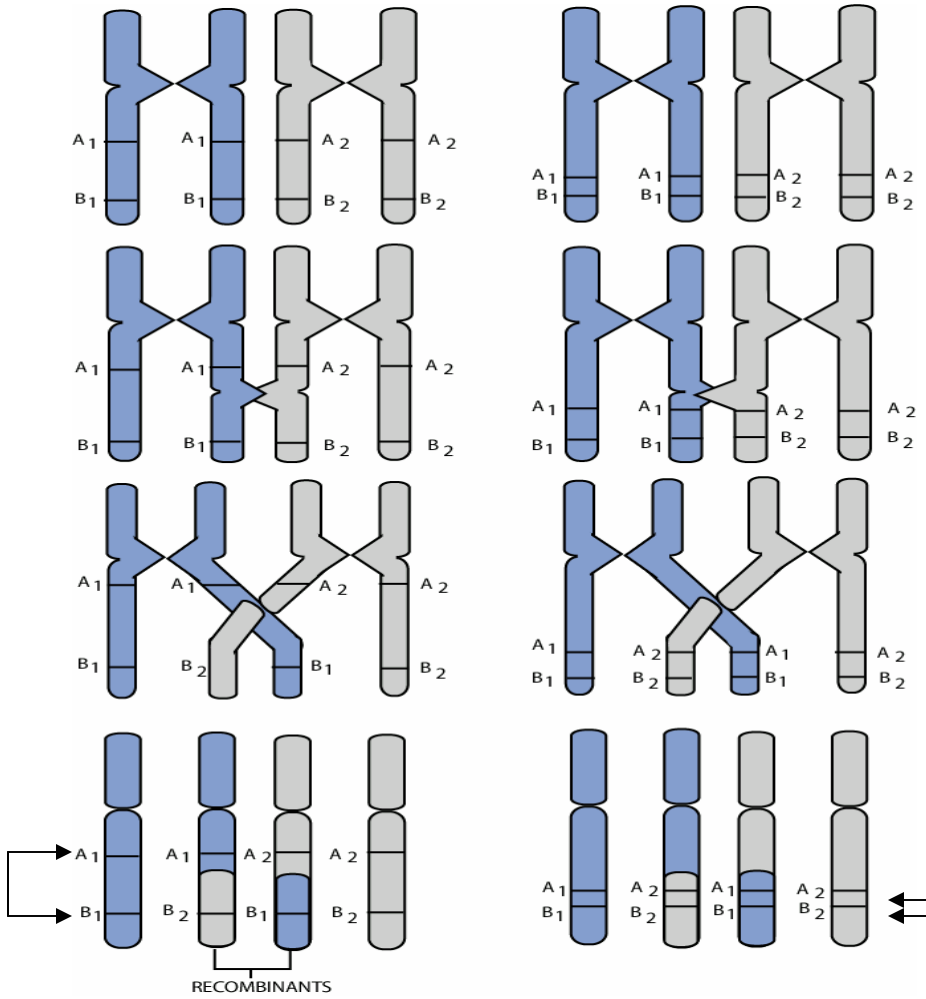
Symbol	Interpretation
CRYBA1	Gene for crystalline, beta A1 polypeptide
GAPD	Gene for glyceraldehydes-3-phosphate dehydrogenase
GAPDL7	Gene-like gene 7, functional status unknown
GAPDP1	GAPD pseudogene 1
AK1	Gene for adenylate kinase, locus 1
AK2	Gene for adenylate kinase, locus 2
*PGK1*2*	Second allele at PGK1 locus
B3P42	Breakpoint number 43 on chromosome 3
DYS29	Unique DNA segment number 29 on the Y chromosome
D3S2550E	Unique DNA segment number 2550 on chromosome 3, known to be expressed
D11Z3	Chromosome 11-specific repetitive DNA family number 3
DXYS6X	DNA segment found on the Y chromosome, with a known homologue on the Y chromosome, and representing the 6^{th} XY homologue pair to be classified
DXYS44Y	DNA segment found on the Y chromosome, with a known homologue on the X chromosome, 44^{th} XY homologue pair
D12F3S1	DNA segment on chromosome 12, first member of multilocus family 3
DXF3S2	DNA segment on chromosome X, second member of multilocus family 3
FRA16A	Fragile site A on chromosome 16

Physical mapping of genes and DNA segments



An example of a physical map used in “walking” along chromosome 5 in the region of the adenomatous polyposis coli (APC) gene. Each horizontal line below the top line corresponds to a different yeast artificial chromosome containing 100-1000 kb of human genomic DNA.

Principle of Genetic Mapping

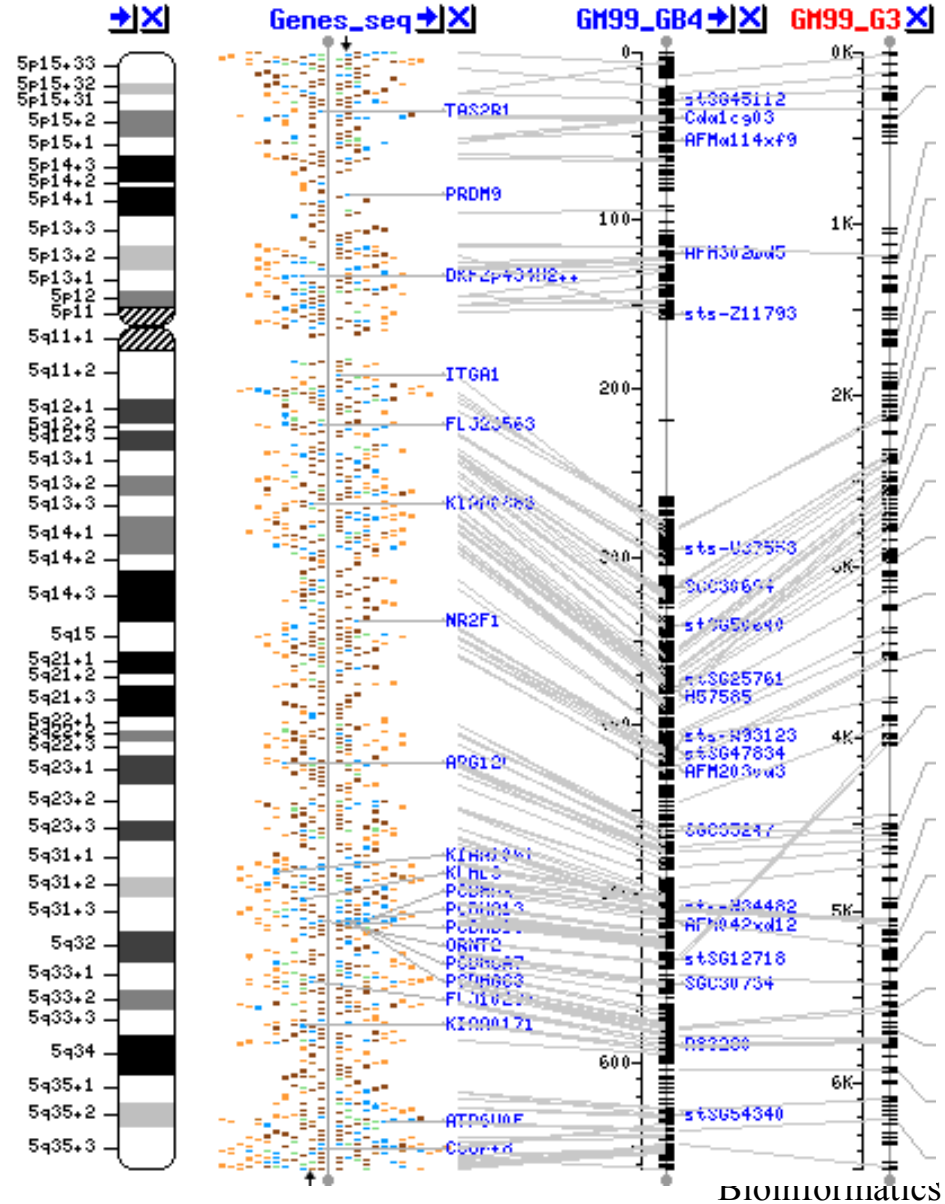


Crossover has occurred Between A and B

No crossover Between A and B

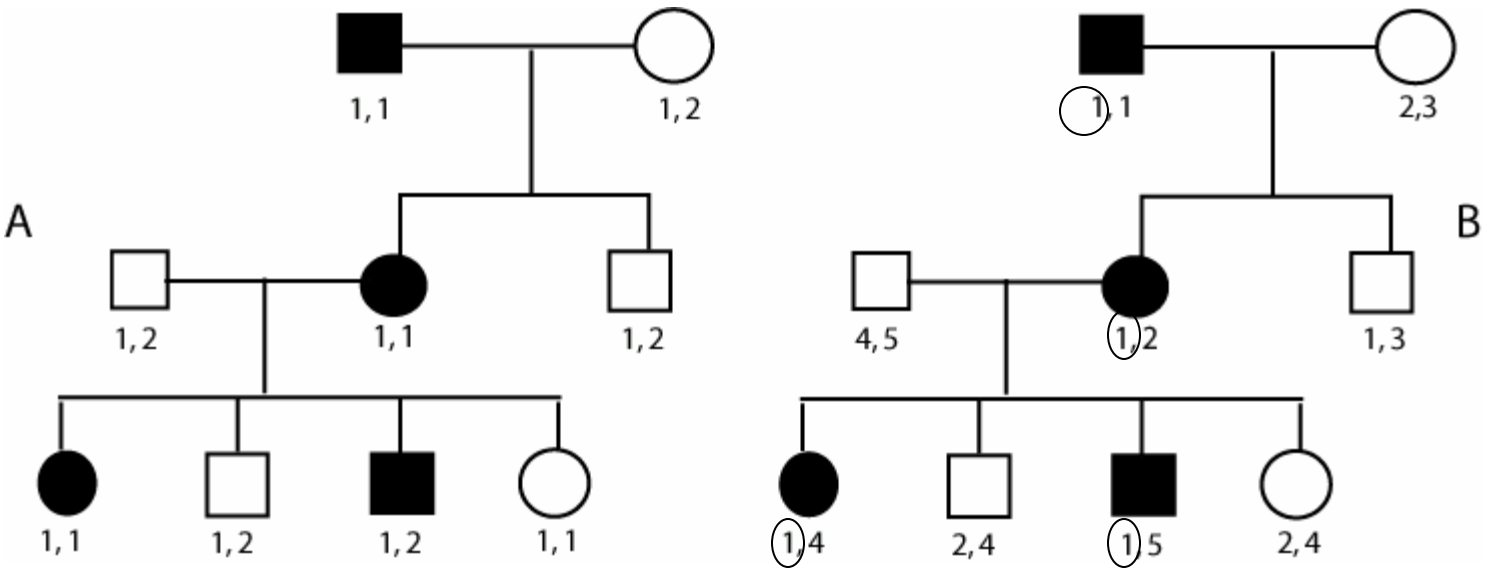
Crossover is more frequent when loci are far apart on chromosomes (*left*) than those that are close together (*right*). **Genetic maps show the frequencies of crossovers between chromosomal loci.**

Integrated genetic & physical map of Chromosome 5



Loci	Type	RH pos'n	STS bin	cM bin
SHGC-12958	cdna	371.00	D5S630-D5S478	18.6-21.4
SHGC-36633	cdna	1206.00	D5S477-D5S651	44.6-46.6
AFM276vb1	poly	1617.00	D5S634-D5S628	59.9-61.1
AFM207yg1	poly	1940.00	D5S634-D5S628	59.9-61.1
SHGC-10654	cdna	2136.00	D5S491-D5S427	67.2-69.6
SHGC-36384	cdna	2395.00	D5S427-D5S647	69.6-74.7
AFM294wd1	poly	2566.00	D5S637-D5S1977	76.5-82.2
SHGC-36700	cdna	2770.00	D5S1977-D5S428	82.2-95.4
SHGC-35966	cdna	2934.00	D5S1977-D5S428	82.2-95.4
AFM238xf4	poly	3223.00	D5S428-D5S401	95.4-97.8
AFM203va3	poly	3532.00	D5S484-D5S495	105.2-107.8
SHGC-12330	cdna	4028.00	D5S495-D5S492	107.8-116.3
SHGC-10332	cdna	4523.00	D5S656-D5S471	121.6-129.6
AFMc006xf9	poly	4930.00	D5S471-D5S393	129.6-140.8
AFM066xf11	poly	5137.00	D5S471-D5S393	129.6-140.8
SHGC-9824	cdna	5339.00	D5S500-D5S436	141.4-147.2
SHGC-11908	cdna	5554.00	D5S436-D5S470	147.2-152.8
AFM234yh8	poly	5700.00	D5S470-D5S487	152.8-157.6
AFMc033xh	poly	5942.00	D5S1955-D5S496	163.0-172.6
SHGC-11995	cdna	6239.00	D5S498-D5S408	184.7-195.8

Genetic Linkage Analysis in Families with Inherited Diseases



Uninformative Mating

Informative Mating

Figure 8-10. An autosomal dominant disease gene is segregating in this family. **A**, A closely linked two-allele RFLP has been typed for each member of the family, but linkage phase cannot be determined (uninformative mating). **B**, A closely linked six-allele microsatellite polymorphism has been typed for each family member, and linkage phase can now be determined (informative mating).

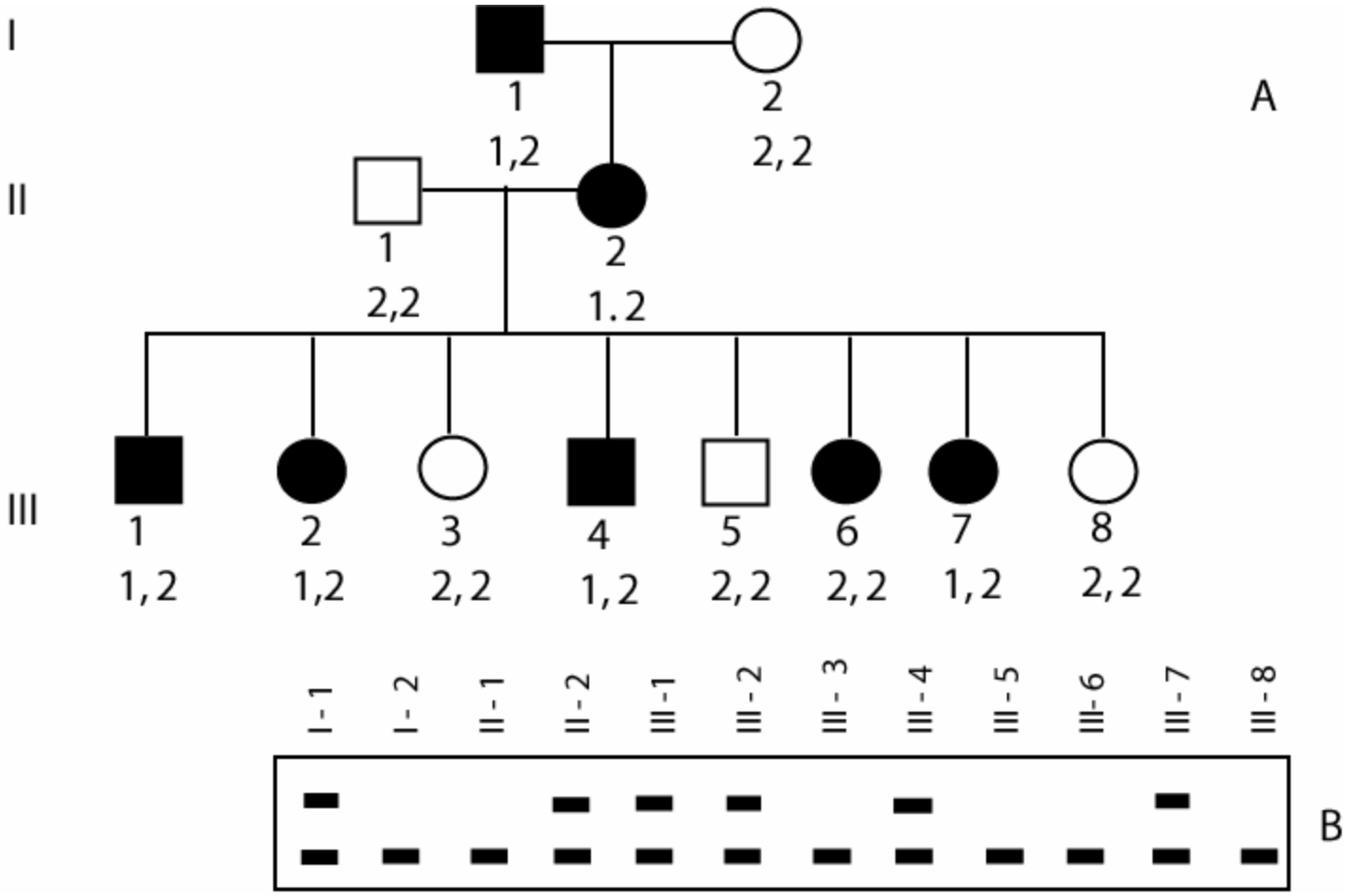
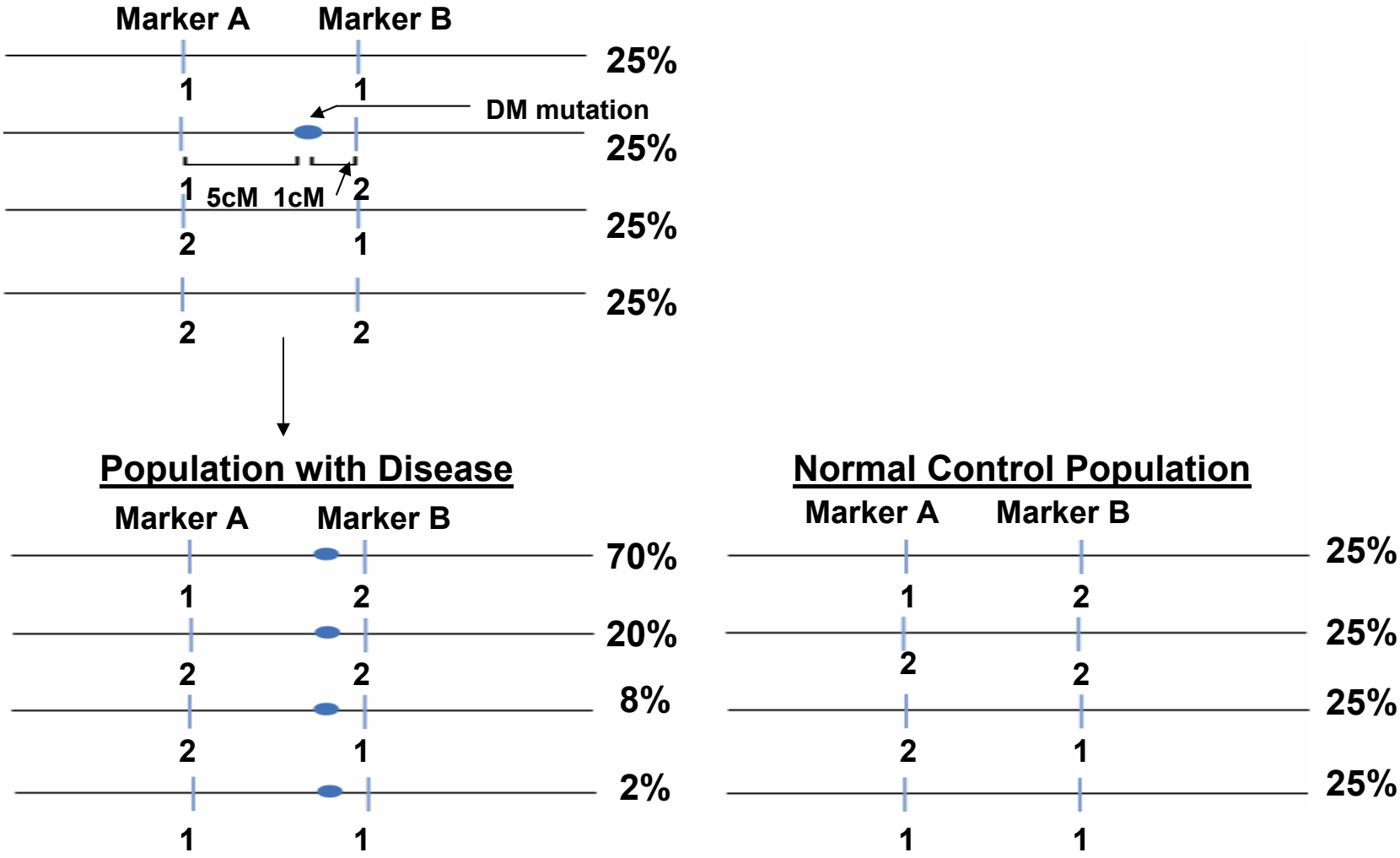


Fig 8-5. **A,** an NF1 pedigree in which each member has been typed for the *1F10* polymorphism. Genotypes for this two-allele marker locus are shown below each individual in the pedigree. Affected pedigree members are indicated by a shaded symbol. **B,** An autoradiogram for the *1F10* polymorphism in this family.



Linkage disequilibrium between the myotonic dystrophy (*DM*) locus and two linked loci, *A* and *B*. The *DM* mutation first arises on the chromosome with the A_1, B_2 haplotype. After a number of generations have passed, most chromosomes carrying the *DM* mutation still have the A_1B_2 haplotype, but, as a result of recombination, the *DM* mutation is also found on other haplotypes. Because the A_1B_2 haplotype is seen in 70% of *DM* chromosomes but only 25% of normal chromosomes, there is linkage disequilibrium between *DM* and loci *A* and *B*. Since locus *B* is closer to *DM*, it had greater linkage disequilibrium with *DM* than does locus *A*.

Examples of databases relevant to the human genome project which store globally produced data

Database	Description and Electronic Address
GenBank	DNA and protein sequences. One of many databases distributed by the US National Center for Biotechnology Information (NCBI), NIH (http://www.ncbi.nlm.nih.gov)
EMBL	DNA sequences. Distributed by the European Bioinformatics Institute (EBI) at Cambridge, UK, together with 30 other molecular biology databases (http://www.ebi.ac.uk)
DDBJ	DNA database of Japan (Mishima) http://www.nig.oc.jp/home.html
PIR (protein information resource)	Protein sequences (single database distributed as a collaboration by: the US National Biomedical Research Foundation, Washington; the Martinsreid Institute for Protein Sequences, Germany; the Japan International Protein Information Database, Tokyo). Accessible at various centers including the US NCBI (see above)
SWISS-PROT	Protein sequences. Maintained collaboratively by the University of Geneva and the EMBL data library. Distributed by several centers, such as the NCI and EBI. Access via NCBI or EBI (see above)
Genome sequence assembly	Draft human genome sequence annotated with known and predicted genes. Updated quarterly (genome.ucsc.edu)
GDB (genome database)	The major human mapping database. Permits interaction with the OMIM database (see below) (http://gdbwww.gdb.org)

Candidate gene analysis

- Establish regional localization of gene in pedigrees, sib-pairs, or inbred or populations with founder effects.
- Examine known and putative genes in region whose boundaries are delineated by probability of gene residing there. Region should be <10 cM.
- To identify candidates in region, consider (1) known function of gene, (2) similar genes in other species whose functions are known, (3) other members of same gene family that are mutated in related disorders (in human and animal models), (4) synteny to mutant loci in other species. (5) changes in expression patterns of genes, biochemical activity, or modification patterns of gene products in region in patients vs. controls.
- Sequence/screen the candidates in patients vs. controls to identify mutations.

Example: Familial Hypertrophic Cardiomyopathy

- One of the most common forms of inherited heart disease leading to abnormal hypertrophy, congestive heart failure and sudden death (especially, in adolescents)
- Heterogeneous: 9 different genes encoding for various components of the sarcomere have been implicated in the disease process.
- However these genes only account for 60% of all cases of FHCM.

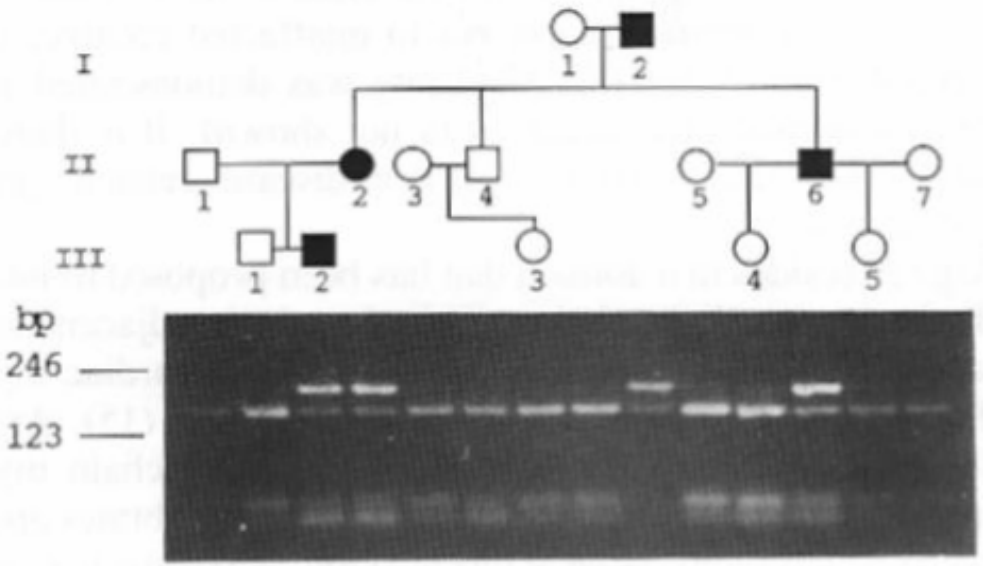
Sarcomeric genes causing (or suspected to cause) FHCM.

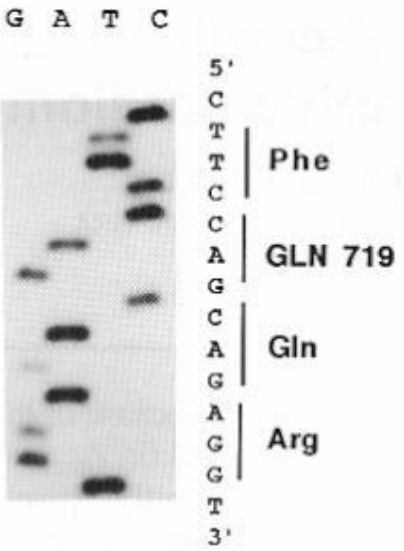
Gene	Location	FHCM	Sarcomeric Gene	Location	FHCM
Tropomyosin 3	1q22-q23	+	Tropomodulin	9q22	-
Troponin T2	1q3	-	Vinculin	10q22.1-q23	-
Myosin Binding Protein H	1q32.1	+	Actinin -3	11q13-q14	-
Troponin I1	1q32	+	Myosin Binding Protein C	11p11.2	+
Actin (skeletal)	1q42.1-q42.3	-	Myosin Light Chain 2	12q23 - q24.3	+
Actinin Binding Protein	1q42-q43	-	β-cardiac Heavy Chain Myosin	14q11	+
Titin	2q31	-	Actinin -1	14q24.1 - q24.2	-
Nebulin	2q31-q32	-	Tropomyosin	15q22.2	-
Desmin	2q35	?	Actin, cardiac	15q14	-
Troponin C	3p21.3-p14.3	-	Cadherin 15	16q22.1	-
β-1 Catenin	3p22-p21.3	-	Plakoglobin	17q21	-
Myosin Light Chain 3	3p21.3 - p21.2	+	N-Cadherin	18q12.1	-
α- Catenin	5q31	-	Desmocollin	18q12.1 - q12.2	-
Phospholamban	6q22.1	-	Troponin T1, Cardiac	19q13.3 - q13.4	+
Filamin	7q32-q35	-	Troponin C	19q13.3 - q 13.4	-
Tropomyosin 2	9p13	-			

A new missense mutation, Arg719Gln, in the β-cardiac heavy chain myosin gene of patients with familial hypertrophic cardiomyopathy

Mild

Michael W.Consevage, Grant C.Salada, Barry G.Baylen, Roger L.Ladda and Peter K.Rogan



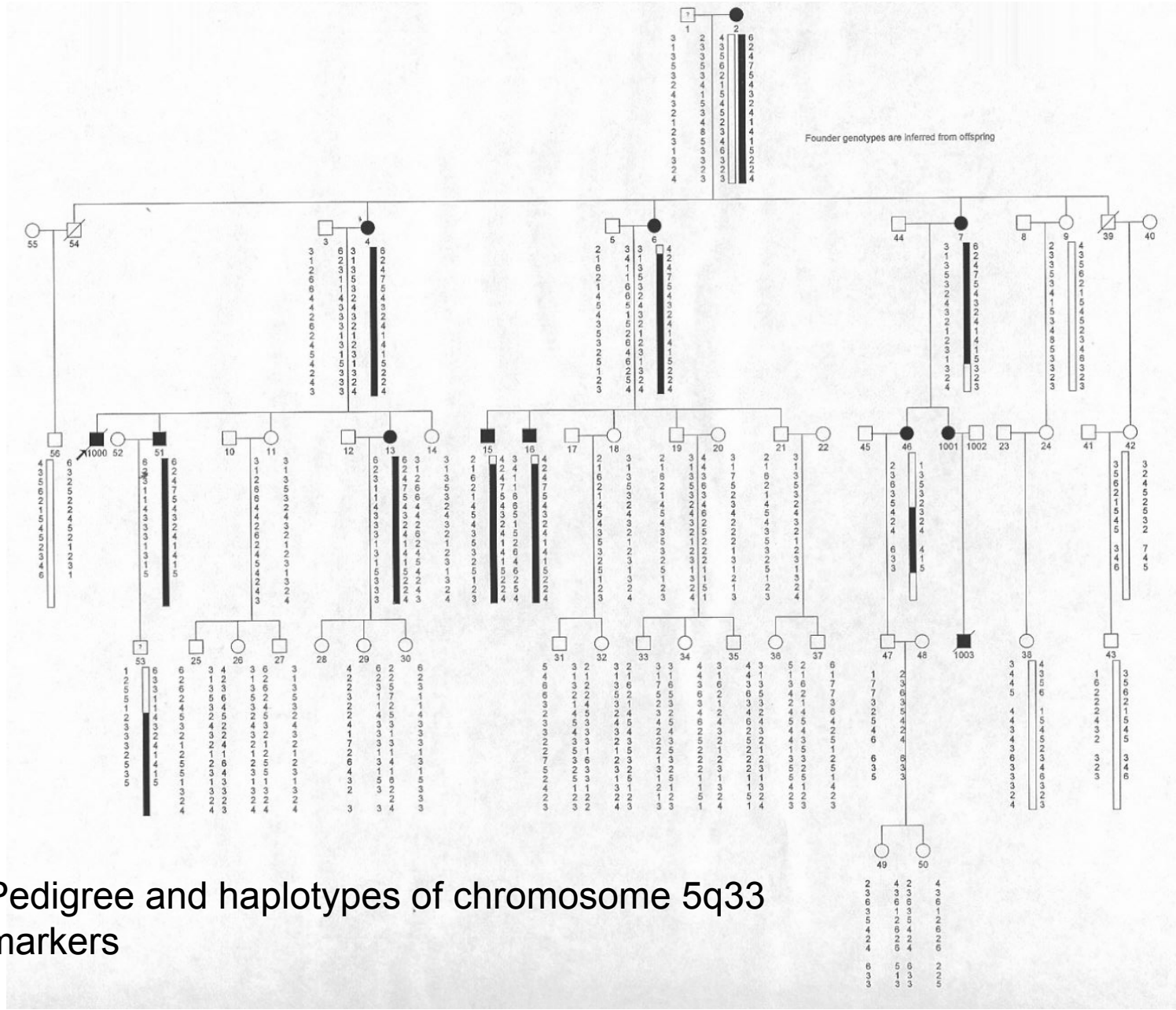


719

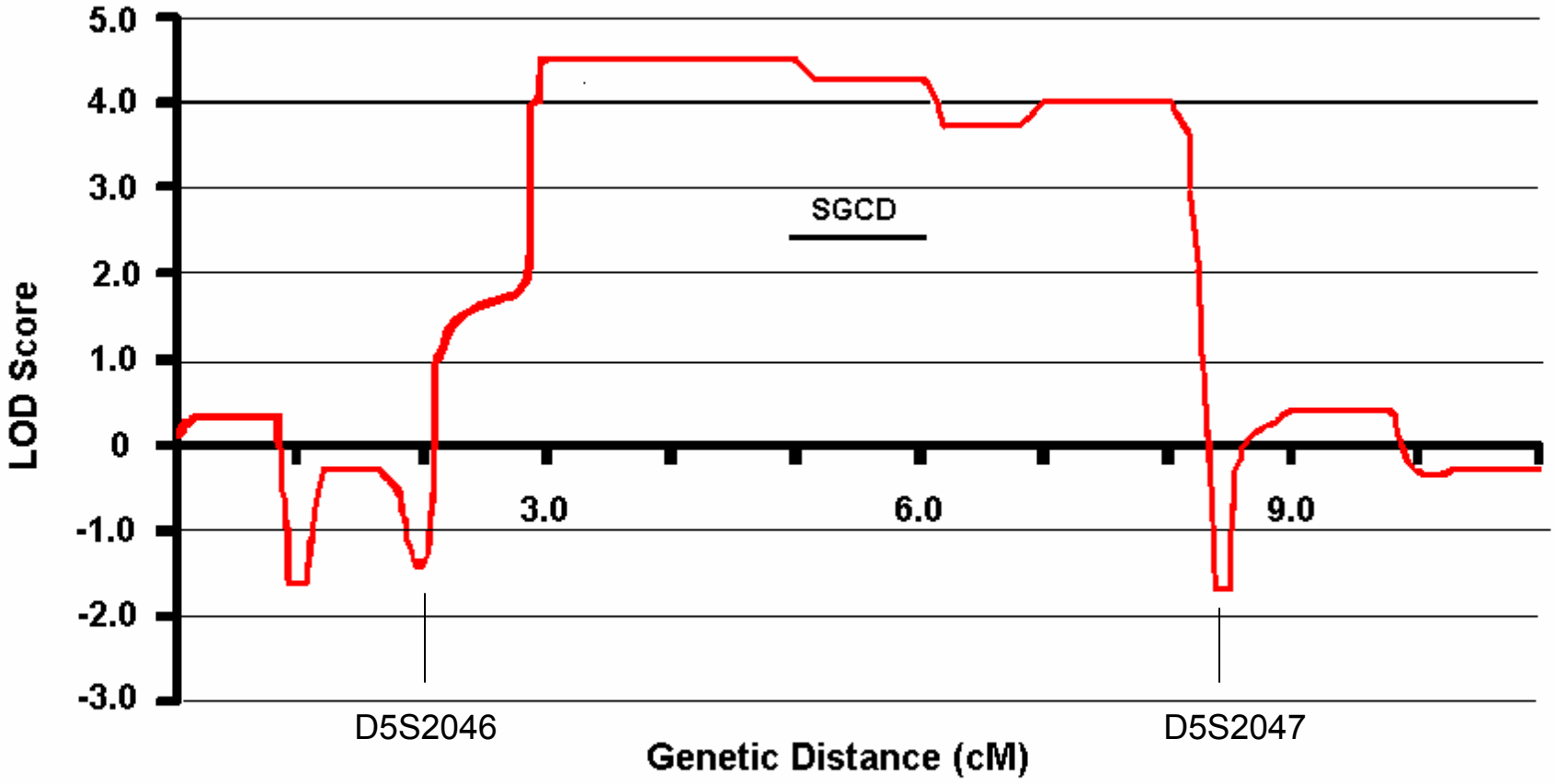
	Phe	Arg	Gln	Arg
Normal ...	TTC	CGG	CAG	AGG ...
HCM		CAG		
		Gln		

- Arg719 resides in a domain that interacts with myosin light chains.
- This amino acid is invariant in skeletal and cardiac myosin in humans and other organisms; some vertebrates avian muscle contain lysine at this position -> positively charged side chain is important for normal function.
- Different point mutation (Arg719Trp) at the same position has much more severe prognosis – whereas Arg719Gln mutation does not cause arrhythmia or sudden death.
- The indole ring of Trp perturbs the structure of the myosin head domain to a greater extent than the ethyl amide side chain of Gln?

New Locus for Familial Hypertrophic - Restrictive Cardiomyopathy Maps to Chromosome 5q33. Peter K. Rogan* Michael W. Consevage, Steven Kasarda, Darrin W. Sabol. Human Genetics, in press

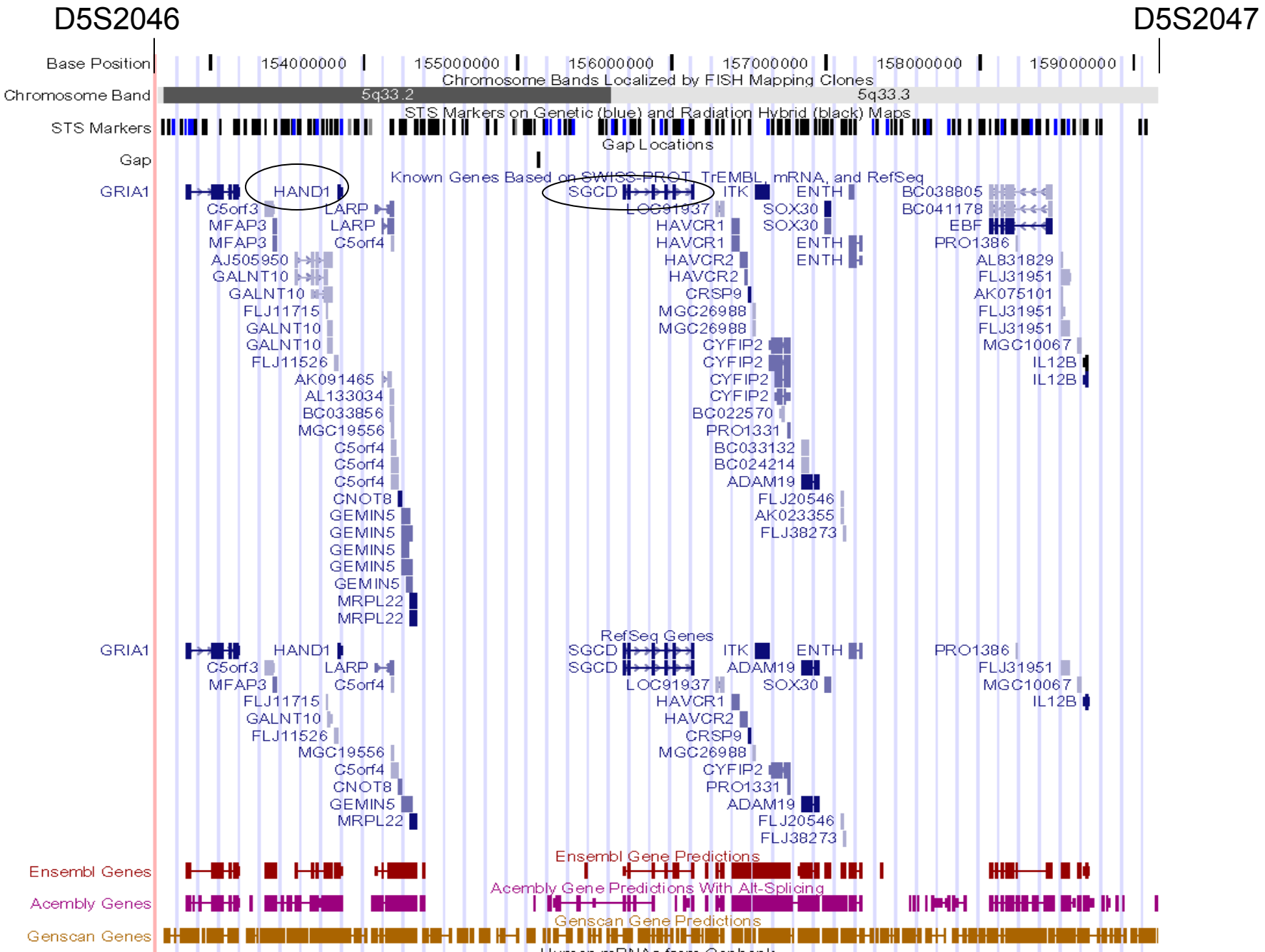


Pedigree and haplotypes of chromosome 5q33 markers



Multipoint linkage results for 5q33 in RCM-FHCM Pedigree

Candidate Genes in the 5q33 RCM-FHCM Linkage Region



Selection of candidate genes

SGCD (Sakamoto et al. PNAS 94(25):13873. 1997)
“Both hypertrophic and dilated cardiomyopathies are caused by mutation of the same gene, delta-sarcoglycan, in hamster: an animal model of disrupted dystrophin-associated glycoprotein complex. A breakpoint causing genomic deletion was found to be located at 6.1 kb 5' upstream of the second exon of delta-SG gene, and its 5' upstream region of more than 27.4 kb, including the authentic first exon of delta-SG gene, was deleted. This deletion included the major transcription initiation site, resulting in a deficiency of delta-SG transcripts with the consequent loss of delta-SG protein in all the CM hamsters, despite the fact that the protein coding region of delta-SG starting from the second exon was conserved in all the CM hamsters.”

HAND1 (Thattaliyath BD et al. BBRC 297(4):870. 2002)
“Human HAND genes are expressed in the adult heart and HAND1 expression is downregulated in cardiomyopathies. Induction of cardiac hypertrophy shows modulation of HAND expression, corresponding with observations in human cardiomyopathy. The downregulation of HAND expression observed in rodent hypertrophy and human cardiomyopathy may reflect a permissive role allowing, cardiomyocytes to reinitiate the fetal gene program and initiate the adaptive physiological changes that allow the heart to compensate (hypertrophy) for the increase in afterload”

Functional genomics

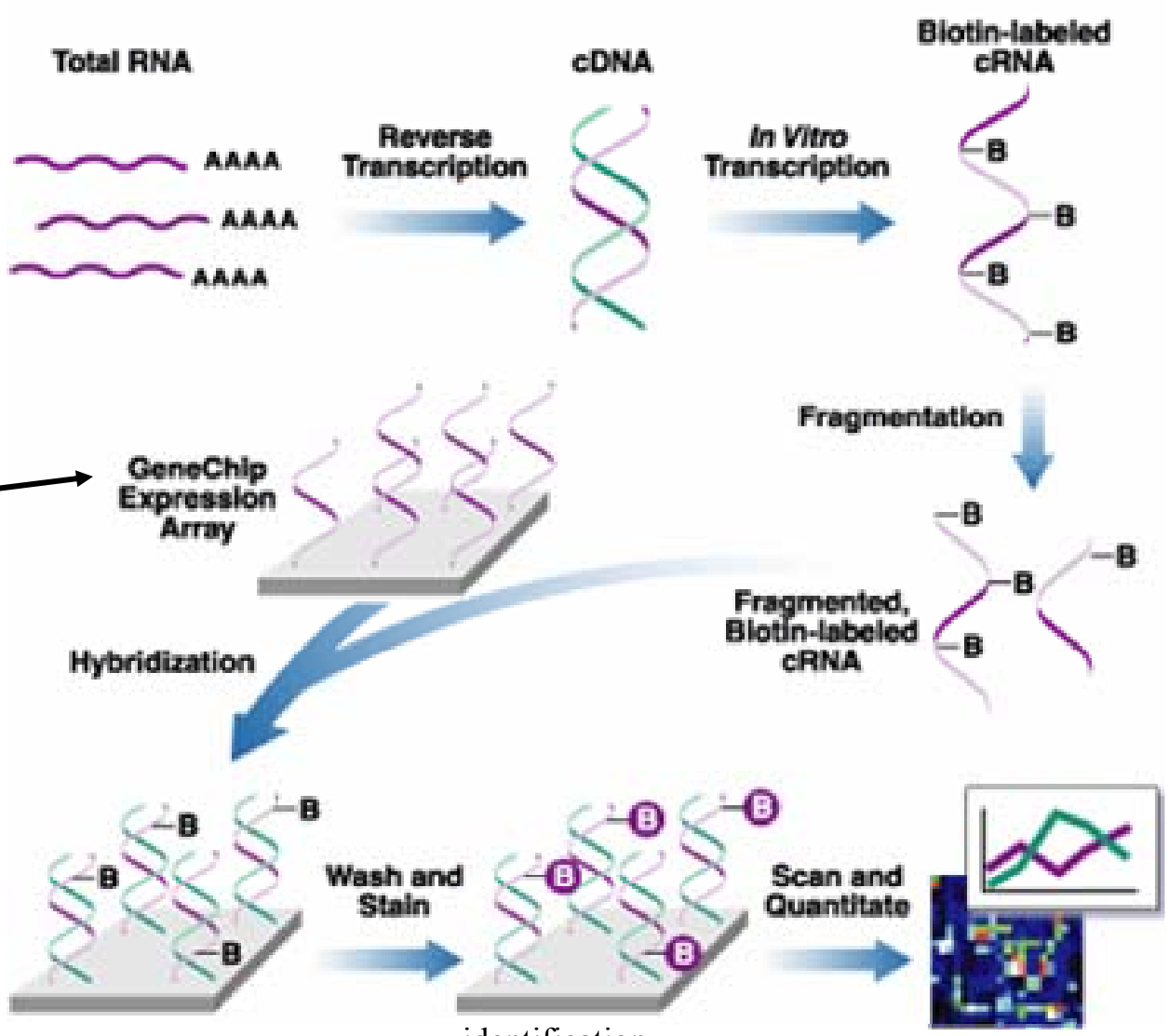
- Functional genomics refers to large scale or global investigations of gene function
- Genome-wide analyses of gene expression and function will become a major area of investigation, ie. the way a cell responds to a particular signal or environmental stimulus can be monitored by simultaneously analyzing the expression patterns of every single gene.
- Once all human genes are known, we can know all the products of those genes. Functional analysis of these products, includes studies of:
  - the transcriptosome (the total collection of RNA transcripts in a cell).
  - the proteome (the total collection of polypeptides/ proteins expressed in a cell).
  - proteomics is devoted to the study of global changes in protein expression and the systematic study of protein-protein interactions.

Gene expression arrays

Affymetrix

from abnormal cell, or cell exposed to a drug (compared to control)

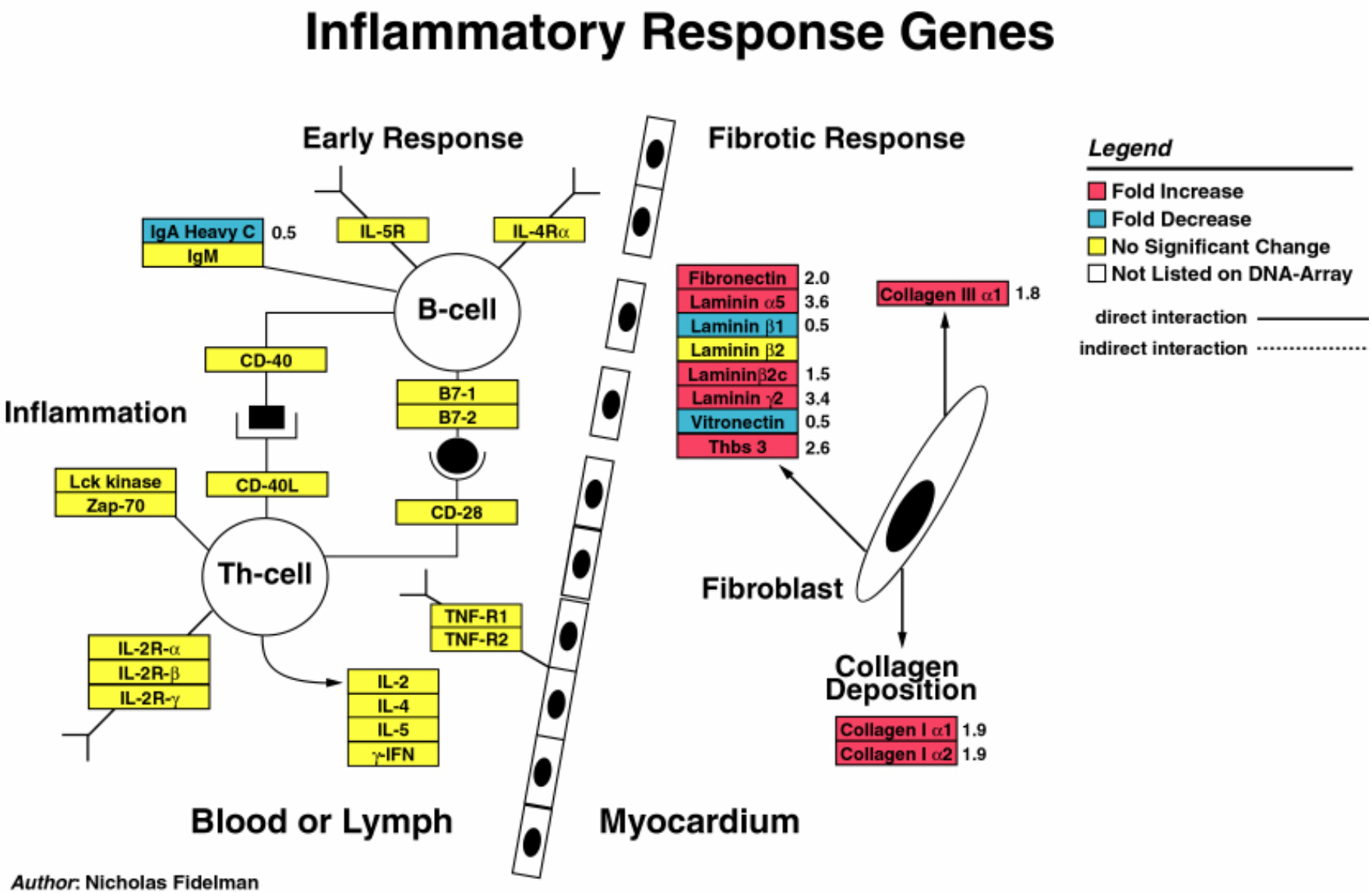
66,000 Human cDNAs



identification

Correlate expression and pathways with pathology

Example: murine hypertrophic cardiomyopathy



Comparative genomics

- Comparative genomics involves analysis of two or more genomes to identify the extent of similarity of various features, or large- scale screening of a genome to identify sequences present in another genome.
- Comparison of archaeal genomes with eubacterial and eukaryotic genomes to infer evolutionary relatedness, and ultimately function.
- We have very limited information about regulatory elements in complex eukaryotic genomes. Using databases of known regulatory element sequences, computer programs can inspect new genomic sequences for the presence of regulatory elements.
- Electronic screening of EST databases can identify homologs of biologically interesting genes in other species. Systematic screening of the dbEST database has revealed many potentially interesting human homologs of Drosophilia genes known to be loci for mutant phenotypes.

Comparison of *FOXP2* in human and mouse:

A gene encoding the "speech" phenotype http://pipeline.lbl.gov/

The post-genome (sequencing) era

- Genetic testing will become widely available, not just for genetic disorders, but also in terms of genetic susceptibility to a variety of different conditions, including infectious diseases.
- New types of genetic testing will become available.
- Improved treatments can also be expected. Gene therapy approaches may prove technically difficult, but the new information will undoubtedly assist the development of novel therapies.

Summary

- Mapping and sequencing the human genome has led to the identification of the genes responsible for thousands of genetic disorders and accelerated the process of finding others
- Genetic maps led to physical maps, which prepared the ground to determine the sequence(s) of the human genome
- There are many useful web based resources that reveal the medical relevance the genome sequence in as much detail as desired.
- The genome sequence has revolutionized candidate gene determination of the causes of many inherited and acquired disorders.
- In the post-genome era, comparative and functional genomics will help us to interpret the genome sequence, resulting in faster diagnoses and improved therapies.