Computational Biology

Lecture #2: A Vision…

Bud Mishra

Professor of Computer Science, Mathematics, & Cell Biology

Sept 19 2005

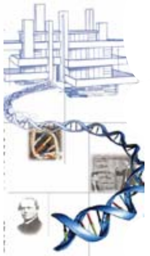
Challenges

NATURE | VOL **422** | 24 APRIL 2003
lwww.nature.com/nature

A vision for the future genomics research

A blueprint for the genomic era.

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute*



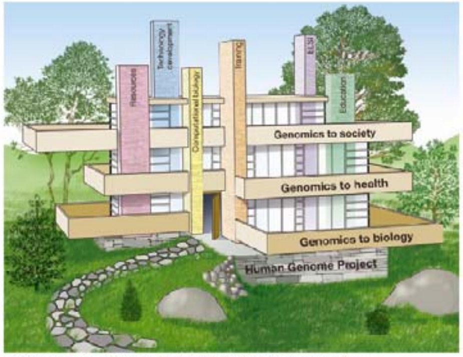
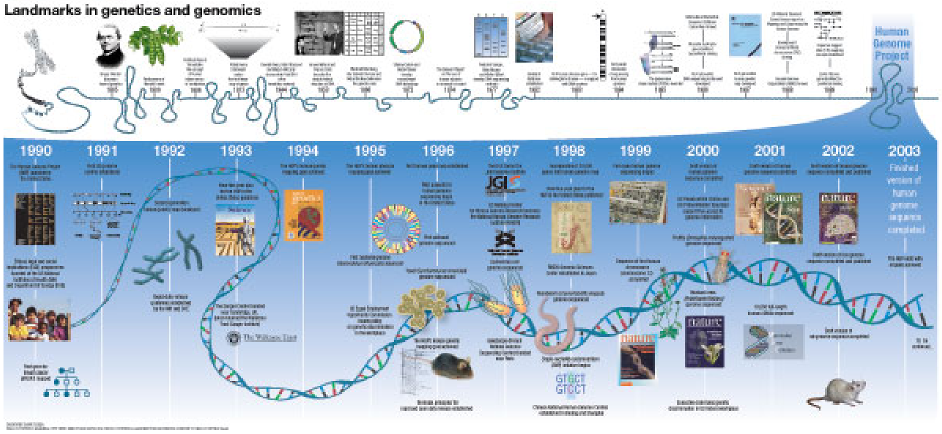


Fig 2 The future of genomics rests on the foundation of the Human Genome Project.

TimeLine



Context

- Human Genome Project (HGP) beginning in 1990.
- The completion of a high-quality,comprehensive sequence of the human genome, in the fiftieth anniversary year of the discovery of the double-helical structure of DNA, is a landmark event. The genomic era is now a reality.
- The remarkable path to Human Genome:
  - Gregor Mendel's discovery of the laws of heredity and their rediscovery in the early days of the twentieth century
  - Recognition of DNA as the hereditary material
  - Determination of its structure
  - Elucidation of the genetic code
  - Development of recombinant DNA technologies
  - Establishment of increasingly automatable methods for DNA sequencing

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

What's Next?

- Now, genomics has become a central and cohesive discipline of biomedical research.
  - The genomic approach of technology development
  - Large-scale generation of community resource data sets
  - Interwoven advances in *genetics*, *comparative genomics*, *high throughput biochemistry* and *bioinformatics*
- Improved repertoire of research tools to allow the functioning of organisms in health and disease to be analyzed and comprehended at an unprecedented level of molecular detail.

What's now possible?

- "Identification of the genes responsible for human mendelian diseases,once a Herculean task requiring large research teams, many years of hard work, and an uncertain outcome, can now be routinely accomplished in a few weeks by a single graduate student with access to DNA samples and associated phenotypes, an Internet connection to the public genome databases, a thermal cycler and a DNA-sequencing machine."

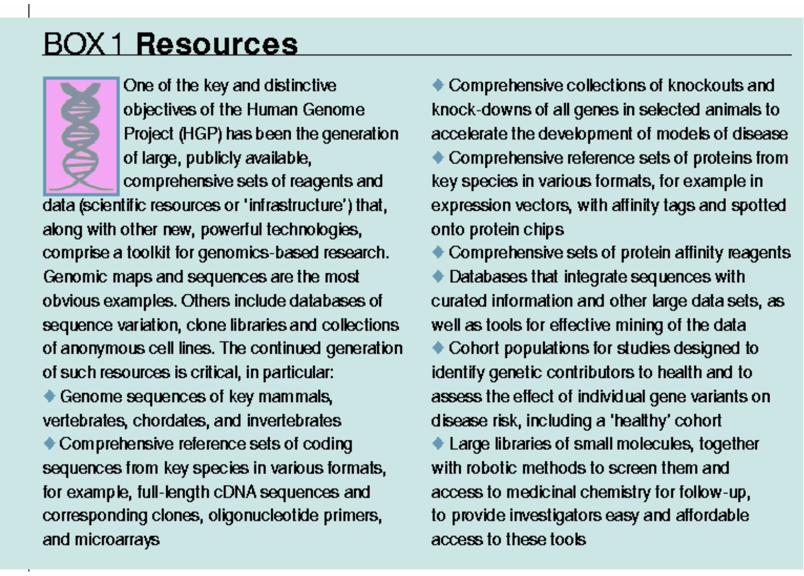
Major Themes

Fig 1 The future of genomics rests on the foundation of the Human Genome Project.

- From genomic information to improved human health
  - US National Institutes of Health (NIH, www.nih.gov)and numerous national and international governmental and charitable organizations supporting medical research
  - The National Human Genome Research Institute (NHGRI)
  - The National Institute for General Medical Sciences (NIGMS)
- The vision: three major themes
  - *genomics to biology*
  - *genomics to health*, and
  - *genomics to society* and
  - six crosscutting elements.
- See www.genome.gov/About/Planning.

Six Crosscutting Elements

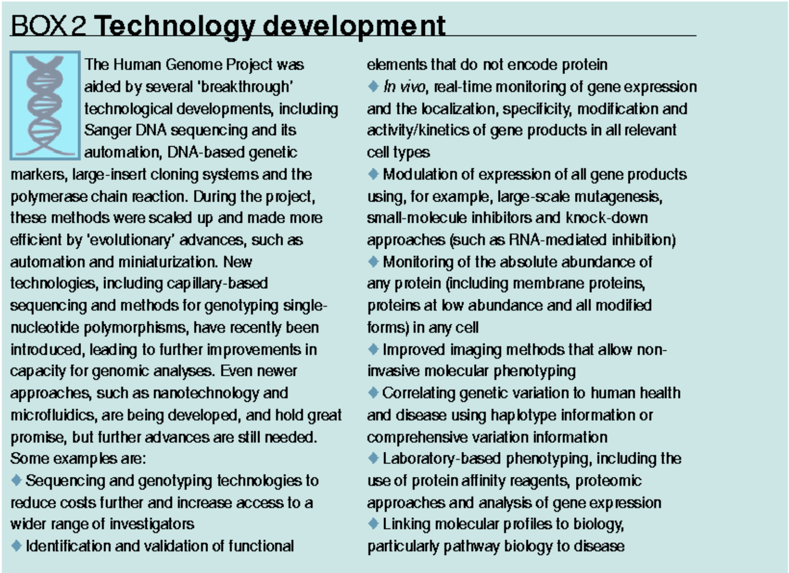
- Resources
- Technology
- Development
- Computational biology
- Training
- Ethical legal and social implications (ELSI)
- Education



BOX 1 Resources

One of the key and distinctive objectives of the Human Genome Project (HGP) has been the generation of large, publicly available, comprehensive sets of reagents and data (scientific resources or 'infrastructure') that, along with other new, powerful technologies, comprise a toolkit for genomics-based research. Genomic maps and sequences are the most obvious examples. Others include databases of sequence variation, clone libraries and collections of anonymous cell lines. The continued generation of such resources is critical, in particular:

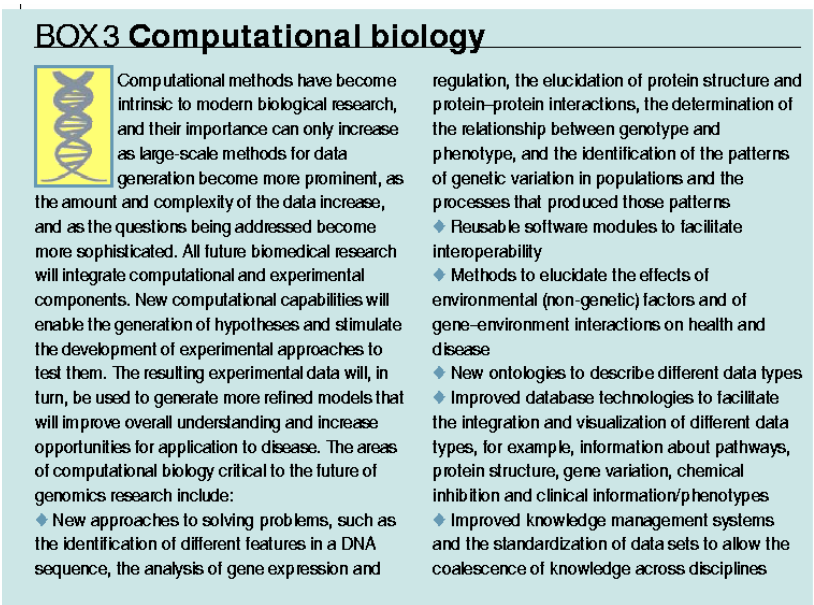
- Genome sequences of key mammals, vertebrates, chordates, and invertebrates
- Comprehensive reference sets of coding sequences from key species in various formats, for example, full-length cDNA sequences and corresponding clones, oligonucleotide primers, and microarrays
- Comprehensive collections of knockouts and knock-downs of all genes in selected animals to accelerate the development of models of disease
- Comprehensive reference sets of proteins from key species in various formats, for example in expression vectors, with affinity tags and spotted onto protein chips
- Comprehensive sets of protein affinity reagents
- Databases that integrate sequences with curated information and other large data sets, as well as tools for effective mining of the data
- Cohort populations for studies designed to identify genetic contributors to health and to assess the effect of individual gene variants on disease risk, including a 'healthy' cohort
- Large libraries of small molecules, together with robotic methods to screen them and access to medicinal chemistry for follow-up, to provide investigators easy and affordable access to these tools



BOX 2 Technology development

The Human Genome Project was aided by several 'breakthrough' technological developments, including Sanger DNA sequencing and its automation, DNA-based genetic markers, large-insert cloning systems and the polymerase chain reaction. During the project, these methods were scaled up and made more efficient by 'evolutionary' advances, such as automation and miniaturization. New technologies, including capillary-based sequencing and methods for genotyping single-nucleotide polymorphisms, have recently been introduced, leading to further improvements in capacity for genomic analyses. Even newer approaches, such as nanotechnology and microfluidics, are being developed, and hold great promise, but further advances are still needed. Some examples are:

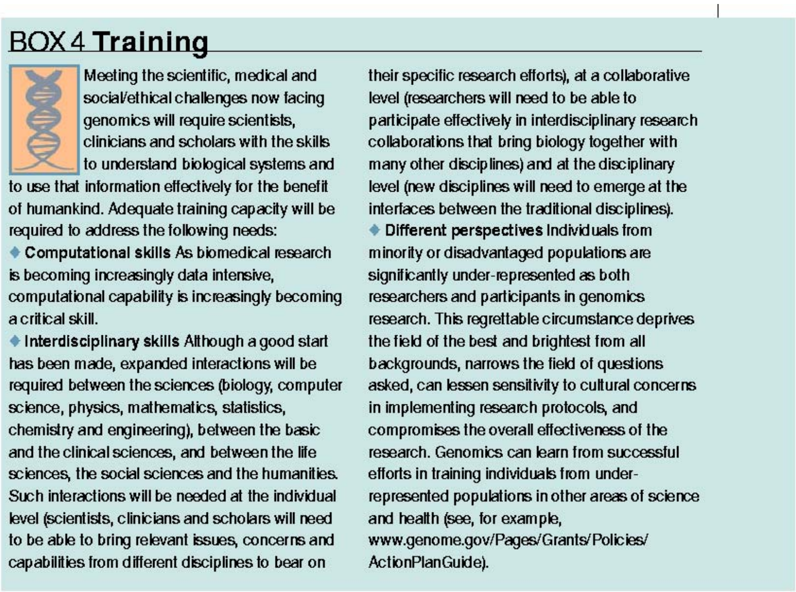
- Sequencing and genotyping technologies to reduce costs further and increase access to a wider range of investigators
- Identification and validation of functional elements that do not encode protein
- *In vivo*, real-time monitoring of gene expression and the localization, specificity, modification and activity/kinetics of gene products in all relevant cell types
- Modulation of expression of all gene products using, for example, large-scale mutagenesis, small-molecule inhibitors and knock-down approaches (such as RNA-mediated inhibition)
- Monitoring of the absolute abundance of any protein (including membrane proteins, proteins at low abundance and all modified forms) in any cell
- Improved imaging methods that allow non-invasive molecular phenotyping
- Correlating genetic variation to human health and disease using haplotype information or comprehensive variation information
- Laboratory-based phenotyping, including the use of protein affinity reagents, proteomic approaches and analysis of gene expression
- Linking molecular profiles to biology, particularly pathway biology to disease



BOX 3 Computational biology

Computational methods have become intrinsic to modern biological research, and their importance can only increase as large-scale methods for data generation become more prominent, as the amount and complexity of the data increase, and as the questions being addressed become more sophisticated. All future biomedical research will integrate computational and experimental components. New computational capabilities will enable the generation of hypotheses and stimulate the development of experimental approaches to test them. The resulting experimental data will, in turn, be used to generate more refined models that will improve overall understanding and increase opportunities for application to disease. The areas of computational biology critical to the future of genomics research include:

- New approaches to solving problems, such as the identification of different features in a DNA sequence, the analysis of gene expression and regulation, the elucidation of protein structure and protein–protein interactions, the determination of the relationship between genotype and phenotype, and the identification of the patterns of genetic variation in populations and the processes that produced those patterns
- Reusable software modules to facilitate interoperability
- Methods to elucidate the effects of environmental (non-genetic) factors and of gene–environment interactions on health and disease
- New ontologies to describe different data types
- Improved database technologies to facilitate the integration and visualization of different data types, for example, information about pathways, protein structure, gene variation, chemical inhibition and clinical information/phenotypes
- Improved knowledge management systems and the standardization of data sets to allow the coalescence of knowledge across disciplines



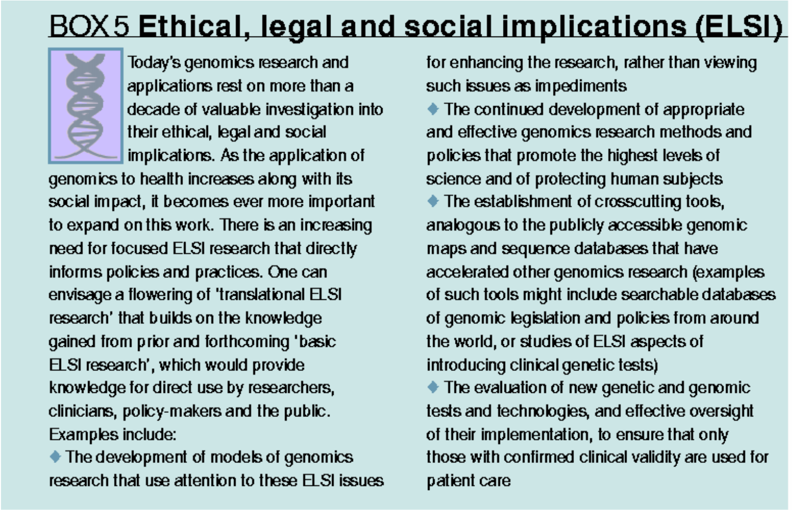
BOX 4 Training

Meeting the scientific, medical and social/ethical challenges now facing genomics will require scientists, clinicians and scholars with the skills to understand biological systems and to use that information effectively for the benefit of humankind. Adequate training capacity will be required to address the following needs:

◆ **Computational skills** As biomedical research is becoming increasingly data intensive, computational capability is increasingly becoming a critical skill.

◆ **Interdisciplinary skills** Although a good start has been made, expanded interactions will be required between the sciences (biology, computer science, physics, mathematics, statistics, chemistry and engineering), between the basic and the clinical sciences, and between the life sciences, the social sciences and the humanities. Such interactions will be needed at the individual level (scientists, clinicians and scholars will need to be able to bring relevant issues, concerns and capabilities from different disciplines to bear on their specific research efforts), at a collaborative level (researchers will need to be able to participate effectively in interdisciplinary research collaborations that bring biology together with many other disciplines) and at the disciplinary level (new disciplines will need to emerge at the interfaces between the traditional disciplines).

◆ **Different perspectives** Individuals from minority or disadvantaged populations are significantly under-represented as both researchers and participants in genomics research. This regrettable circumstance deprives the field of the best and brightest from all backgrounds, narrows the field of questions asked, can lessen sensitivity to cultural concerns in implementing research protocols, and compromises the overall effectiveness of the research. Genomics can learn from successful efforts in training individuals from under-represented populations in other areas of science and health (see, for example, www.genome.gov/Pages/Grants/Policies/ActionPlanGuide).



BOX 5 Ethical, legal and social implications (ELSI)

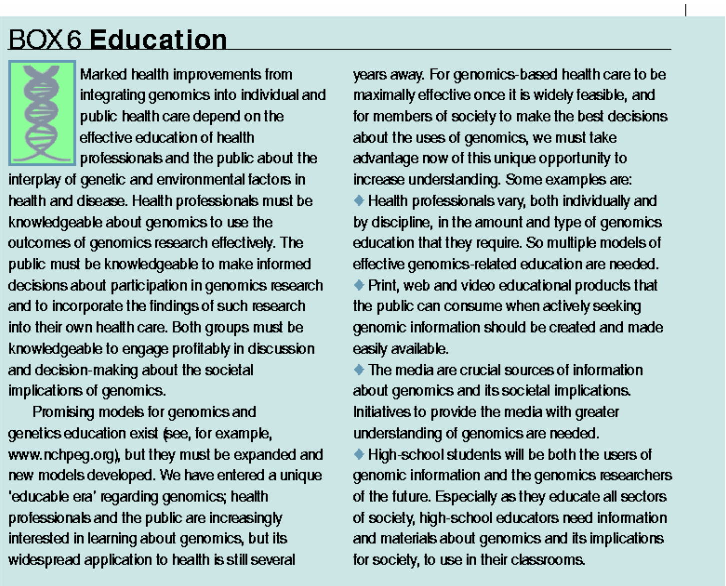
Today's genomics research and applications rest on more than a decade of valuable investigation into their ethical, legal and social implications. As the application of genomics to health increases along with its social impact, it becomes ever more important to expand on this work. There is an increasing need for focused ELSI research that directly informs policies and practices. One can envisage a flowering of 'translational ELSI research' that builds on the knowledge gained from prior and forthcoming 'basic ELSI research', which would provide knowledge for direct use by researchers, clinicians, policy-makers and the public. Examples include:

◆ The development of models of genomics research that use attention to these ELSI issues for enhancing the research, rather than viewing such issues as impediments

◆ The continued development of appropriate and effective genomics research methods and policies that promote the highest levels of science and of protecting human subjects

◆ The establishment of crosscutting tools, analogous to the publicly accessible genomic maps and sequence databases that have accelerated other genomics research (examples of such tools might include searchable databases of genomic legislation and policies from around the world, or studies of ELSI aspects of introducing clinical genetic tests)

◆ The evaluation of new genetic and genomic tests and technologies, and effective oversight of their implementation, to ensure that only those with confirmed clinical validity are used for patient care



BOX 6 Education

Marked health improvements from integrating genomics into individual and public health care depend on the effective education of health professionals and the public about the interplay of genetic and environmental factors in health and disease. Health professionals must be knowledgeable about genomics to use the outcomes of genomics research effectively. The public must be knowledgeable to make informed decisions about participation in genomics research and to incorporate the findings of such research into their own health care. Both groups must be knowledgeable to engage profitably in discussion and decision-making about the societal implications of genomics.

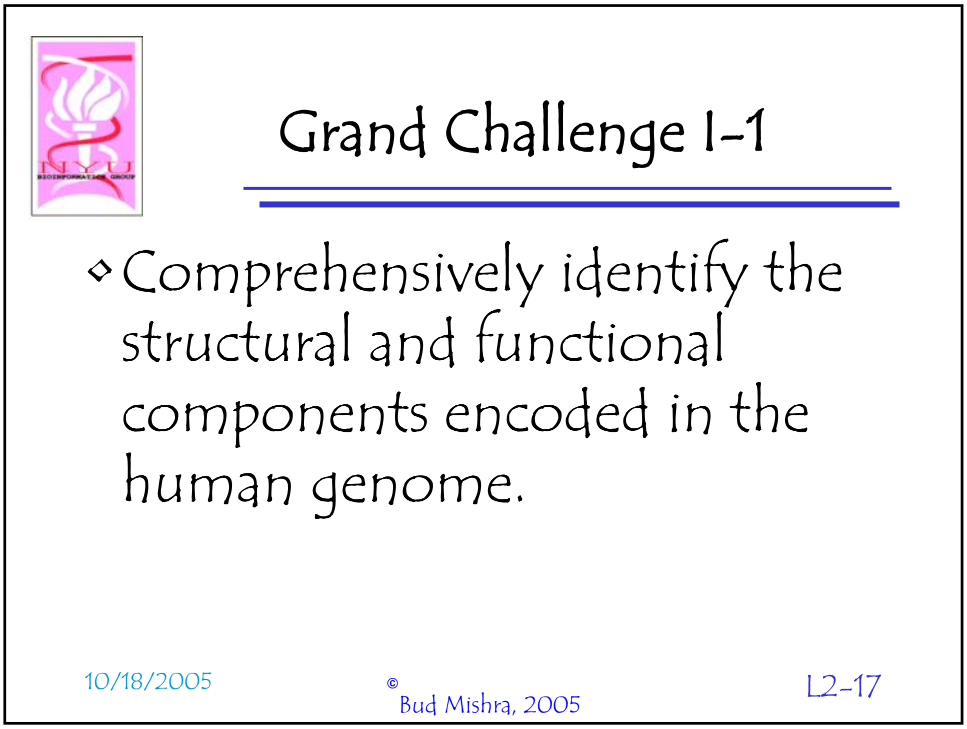
Promising models for genomics and genetics education exist (see, for example, www.nchpeg.org), but they must be expanded and new models developed. We have entered a unique 'educable era' regarding genomics; health professionals and the public are increasingly interested in learning about genomics, but its widespread application to health is still several years away. For genomics-based health care to be maximally effective once it is widely feasible, and for members of society to make the best decisions about the uses of genomics, we must take advantage now of this unique opportunity to increase understanding. Some examples are:

- Health professionals vary, both individually and by discipline, in the amount and type of genomics education that they require. So multiple models of effective genomics-related education are needed.
- Print, web and video educational products that the public can consume when actively seeking genomic information should be created and made easily available.
- The media are crucial sources of information about genomics and its societal implications. Initiatives to provide the media with greater understanding of genomics are needed.
- High-school students will be both the users of genomic information and the genomics researchers of the future. Especially as they educate all sectors of society, high-school educators need information and materials about genomics and its implications for society, to use in their classrooms.

Genomics to biology

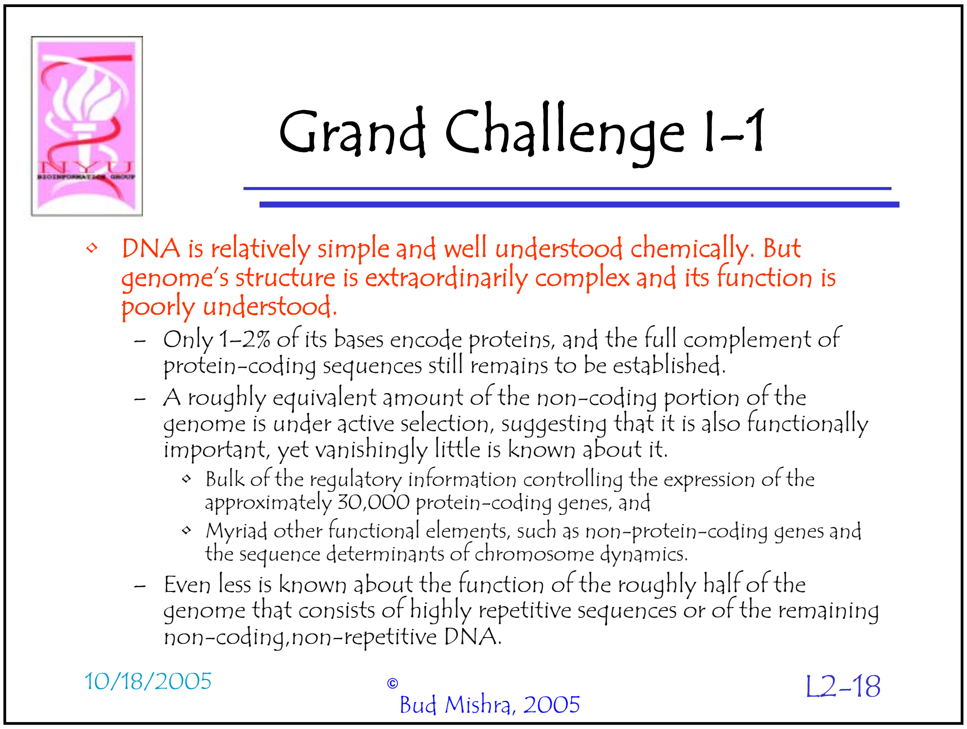
Elucidating the structure and function of genomes

- PART-LISTS: Embedded within a genome, as-yet poorly understood code, are the genetic instructions for the entire repertoire of cellular components.
  - Develop a comprehensive and comprehensible catalogue of all of the components encoded in the human genome.
  - Determine how the genome-encoded components function in an integrated manner to perform cellular and organismal functions.
  - Understand how genomes change and take on new functional roles.



Grand Challenge I-1

- Comprehensively identify the structural and functional components encoded in the human genome.



Grand Challenge I-1

- DNA is relatively simple and well understood chemically. But genome's structure is extraordinarily complex and its function is poorly understood.
  - Only 1-2% of its bases encode proteins, and the full complement of protein-coding sequences still remains to be established.
  - A roughly equivalent amount of the non-coding portion of the genome is under active selection, suggesting that it is also functionally important, yet vanishingly little is known about it.
    - Bulk of the regulatory information controlling the expression of the approximately 30,000 protein-coding genes, and
    - Myriad other functional elements, such as non-protein-coding genes and the sequence determinants of chromosome dynamics.
  - Even less is known about the function of the roughly half of the genome that consists of highly repetitive sequences or of the remaining non-coding,non-repetitive DNA.

Grand Challenge I-1

- Catalogue, characterize and comprehend the entire set of functional elements encoded in the human and other genomes.
- Compiling this genome '*parts list*' will be an immense challenge.
- Well-known classes of functional elements, such as protein-coding sequences, still cannot be accurately predicted from sequence information alone.
- Other types of known functional sequences, such as genetic regulatory elements, are even less well understood; undoubtedly new types remain to be defined, so we must be ready to investigate novel, perhaps unexpected, ways in which DNA sequence can confer function.
- Similarly, a better understanding of epigenetic changes (for example, methylation and chromatin remodelling) is needed to comprehend the full repertoire of ways in which DNA can encode information.

Grand Challenge I-1

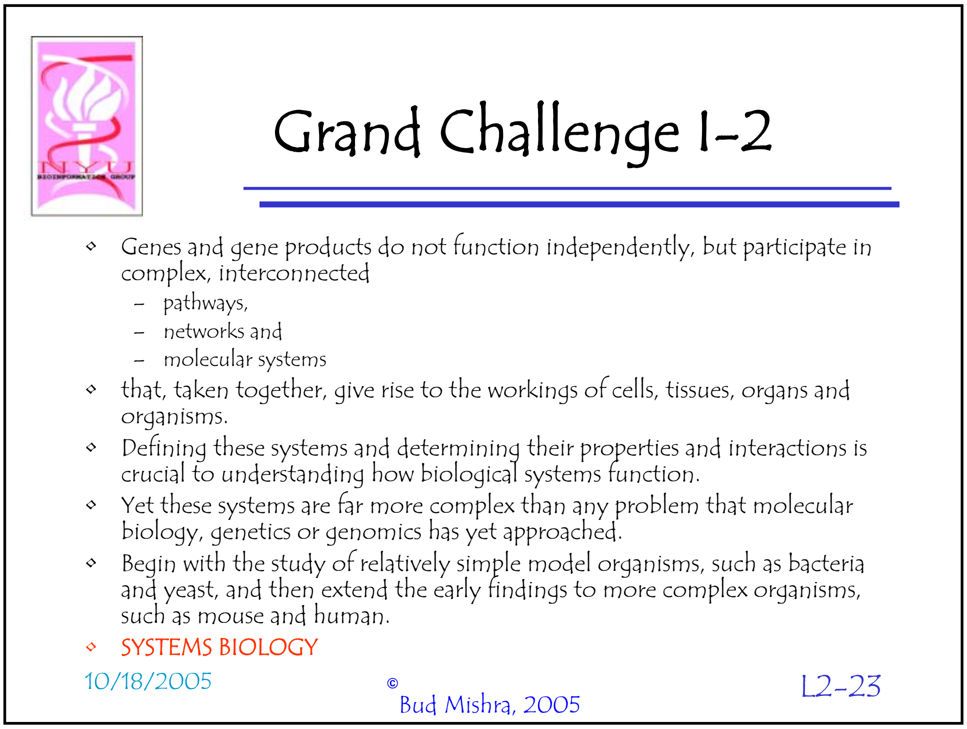
- Comparison of genome sequences from evolutionarily diverse species has emerged as a powerful tool for identifying functionally important genomic elements.
- Initial analyses of available vertebrate genome sequences have revealed many previously undiscovered protein-coding sequences.
- Mammal-to-mammal sequence comparisons have revealed large numbers of homologies in non-coding regions, few of which can be defined in functional terms.
- Further comparisons of sequences derived from multiple species,especially those occupying distinct evolutionary positions, will lead to significant refinements in our understanding of the functional importance of conserved sequences.

Grand Challenge I-1

- Additional genome sequences from several well-chosen species is crucial to the functional characterization of the human genome.
- Effective identification and analysis of functional genomic elements will require increasingly powerful computational capabilities,
  - New approaches for tackling ever-growing and increasingly complex data sets
  - A suitably robust computational infrastructure for housing, accessing and analyzing those data sets
  - Investigators will need to become increasingly adept in dealing with this treasure trove of new information
- The NHGRI recently launched the Encyclopedia of DNA Elements (ENCODE) Project (www.genome.gov/Pages/Research/ENCODE) to identify all the functional elements in the human genome.
  - Systematic strategies for identifying all functionally important genomic elements will be developed and tested using a selected 1% of the human genome.
  - Parallel projects involving well-studied model organisms,for example,yeast,nematode and fruit fly

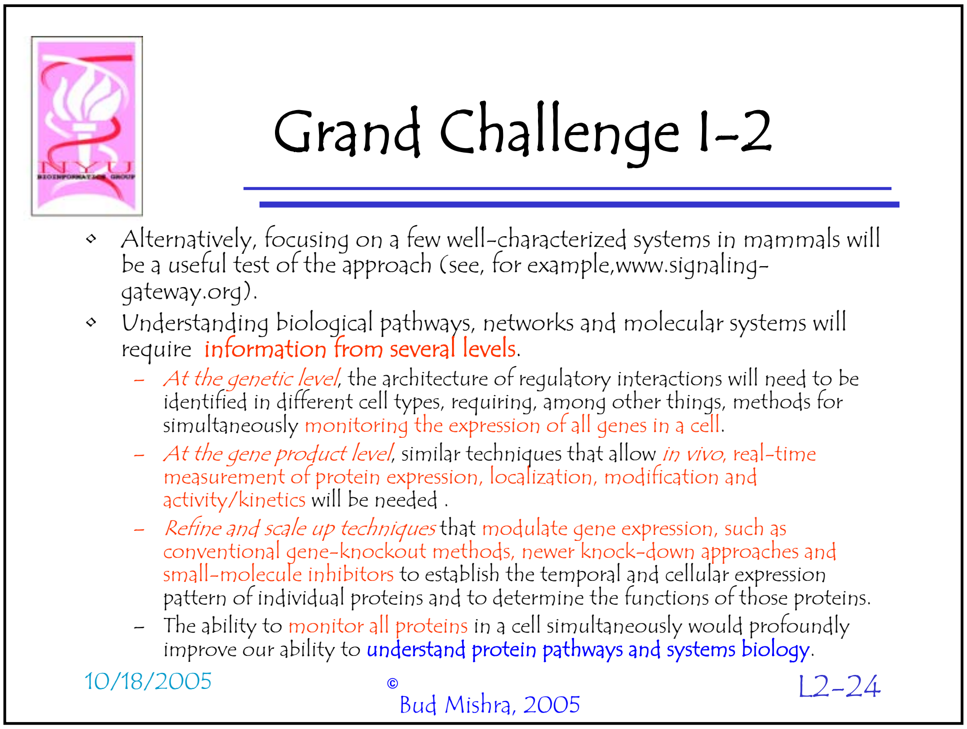
Grand Challenge I-2

- Elucidate the organization of genetic networks and protein pathways and establish how they contribute to cellular and organismal phenotypes
- What is the difference between a 'bag of molecules' and a functioning biological system?



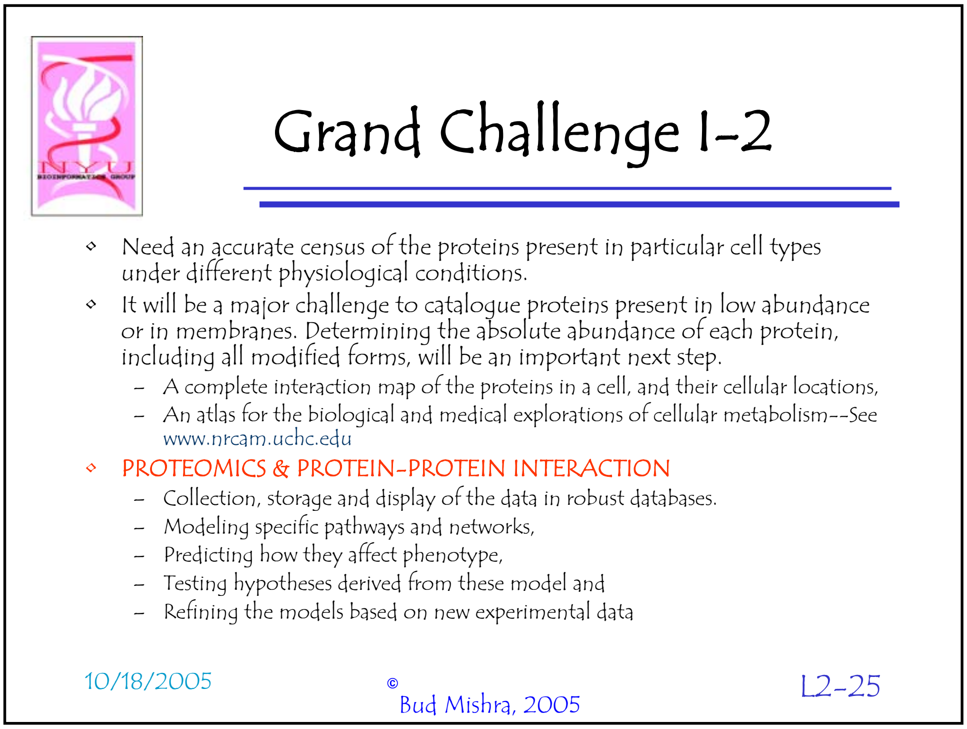
Grand Challenge I-2

- Genes and gene products do not function independently, but participate in complex, interconnected
  - pathways,
  - networks and
  - molecular systems
- that, taken together, give rise to the workings of cells, tissues, organs and organisms.
- Defining these systems and determining their properties and interactions is crucial to understanding how biological systems function.
- Yet these systems are far more complex than any problem that molecular biology, genetics or genomics has yet approached.
- Begin with the study of relatively simple model organisms, such as bacteria and yeast, and then extend the early findings to more complex organisms, such as mouse and human.
- SYSTEMS BIOLOGY



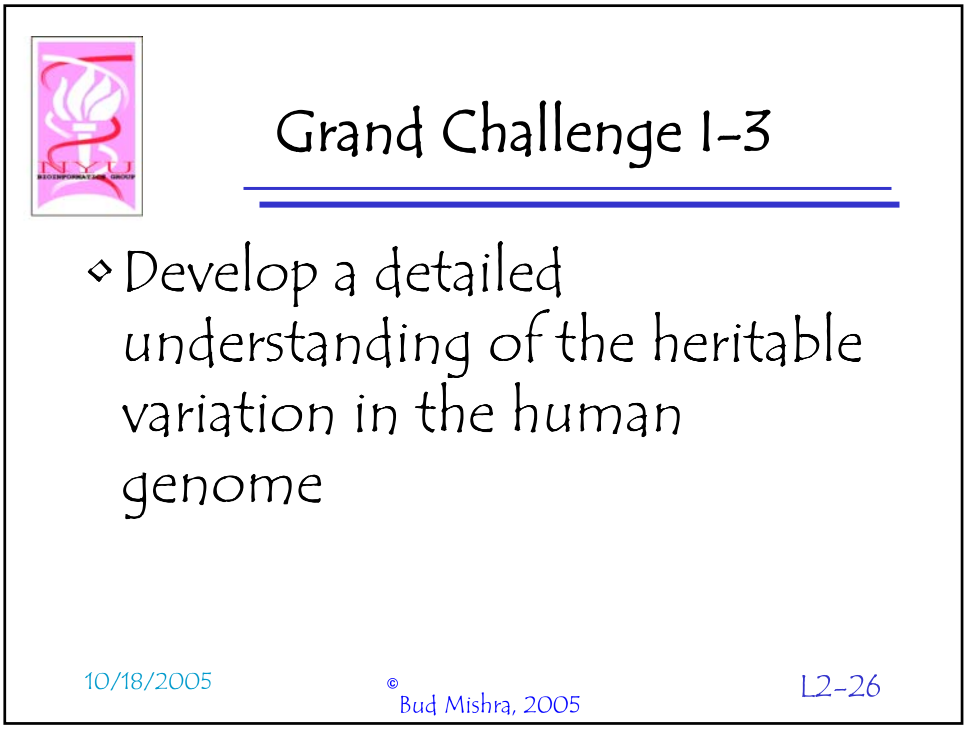
Grand Challenge I-2

- Alternatively, focusing on a few well-characterized systems in mammals will be a useful test of the approach (see, for example,www.signaling-gateway.org).
- Understanding biological pathways, networks and molecular systems will require information from several levels.
  - *At the genetic level*, the architecture of regulatory interactions will need to be identified in different cell types, requiring, among other things, methods for simultaneously monitoring the expression of all genes in a cell.
  - *At the gene product level*, similar techniques that allow *in vivo*, real-time measurement of protein expression, localization, modification and activity/kinetics will be needed .
  - *Refine and scale up techniques* that modulate gene expression, such as conventional gene-knockout methods, newer knock-down approaches and small-molecule inhibitors to establish the temporal and cellular expression pattern of individual proteins and to determine the functions of those proteins.
  - The ability to monitor all proteins in a cell simultaneously would profoundly improve our ability to understand protein pathways and systems biology.



Grand Challenge I-2

- Need an accurate census of the proteins present in particular cell types under different physiological conditions.
- It will be a major challenge to catalogue proteins present in low abundance or in membranes. Determining the absolute abundance of each protein, including all modified forms, will be an important next step.
  - A complete interaction map of the proteins in a cell, and their cellular locations,
  - An atlas for the biological and medical explorations of cellular metabolism--See www.nrcam.uchc.edu
- PROTEOMICS & PROTEIN-PROTEIN INTERACTION
  - Collection, storage and display of the data in robust databases.
  - Modeling specific pathways and networks,
  - Predicting how they affect phenotype,
  - Testing hypotheses derived from these model and
  - Refining the models based on new experimental data



Grand Challenge I-3

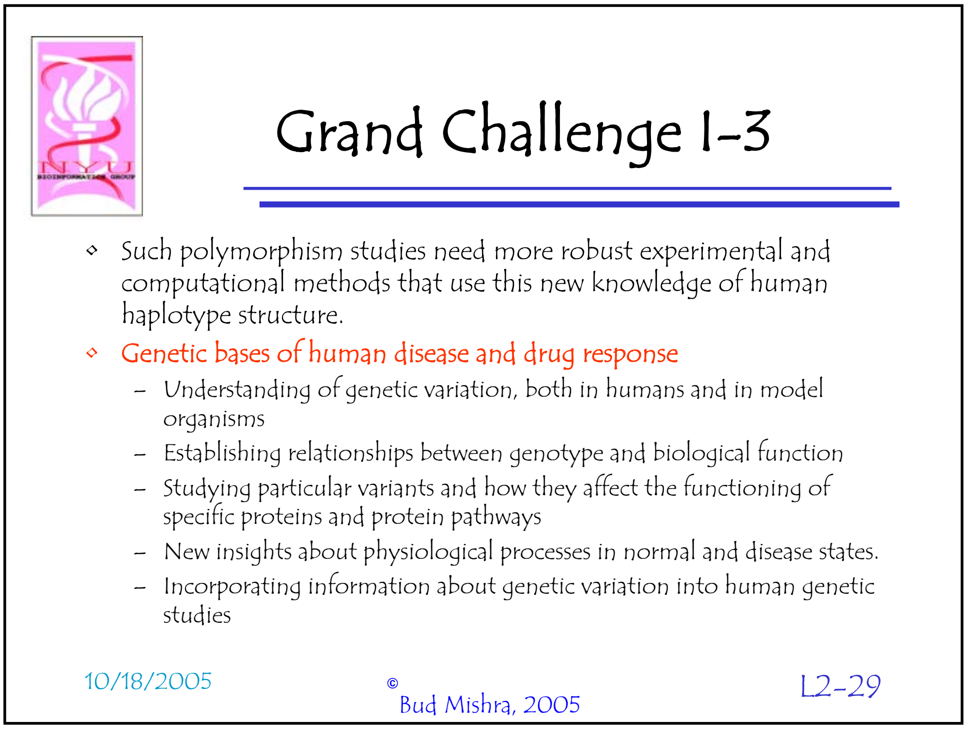
- Develop a detailed understanding of the heritable variation in the human genome

Grand Challenge I-3

- *Genetics*: Correlate variation in DNA sequence with phenotypic differences (traits).
- The greatest advances in human genetics have been made for traits associated with variation in a single gene (Mendelian Trait).
- Complex Traits: But most phenotypes, including common diseases and variable responses to pharmacological agents, have a more complex origin, involving the interplay between multiple genetic factors (genes and their products) and nongenetic factors (environmental influences).
- Unraveling such complexity will require both a complete description of the genetic variation in the human genome and the development of analytical tools for using that information to understand the genetic basis of disease.

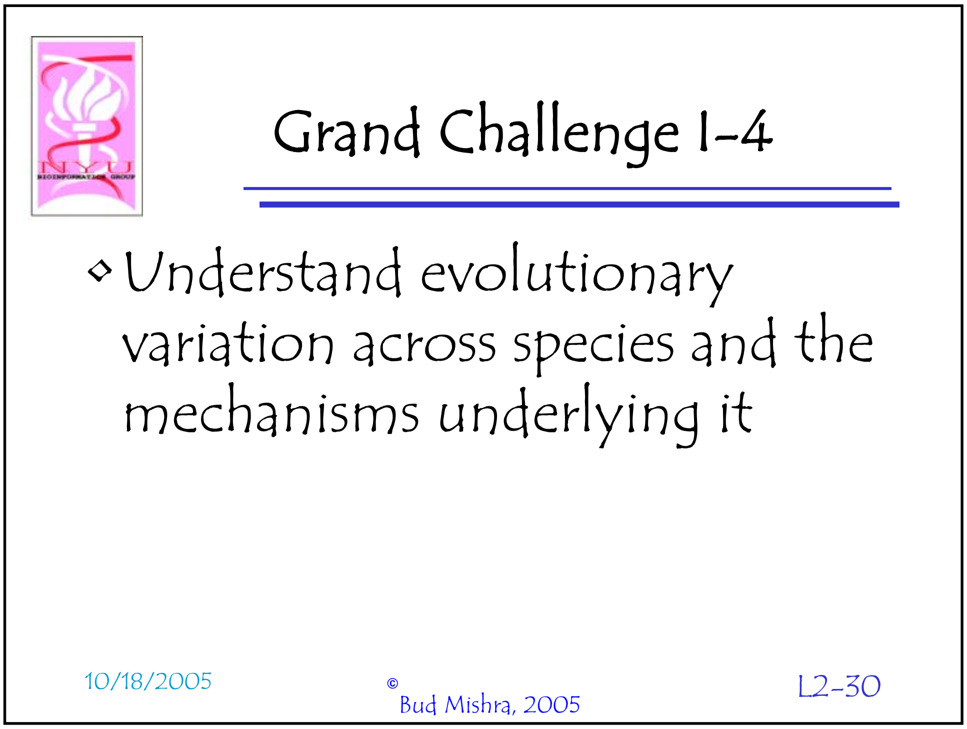
Grand Challenge I-3

- A catalogue of all common variants in the human population,
  - Single-nucleotide polymorphisms (SNPs)
  - Small deletions and insertions
  - Copy number polymorphisms and
  - Other structural differences
- Many SNPs have been identified, and most are publicly available (www.ncbi.nlm.nih.gov/SNP).
- A public collaboration, the International HapMap Project (www.genome.gov/Pages/Research/HapMap), was formed in 2002
  - To characterize the patterns of linkage disequilibrium and haplotypes across the human genome and
  - To identify subsets of SNPs that capture most of the information about these patterns of genetic variation to enable large-scale genetic association studies.



Grand Challenge I-3

- Such polymorphism studies need more robust experimental and computational methods that use this new knowledge of human haplotype structure.
- Genetic bases of human disease and drug response
  - Understanding of genetic variation, both in humans and in model organisms
  - Establishing relationships between genotype and biological function
  - Studying particular variants and how they affect the functioning of specific proteins and protein pathways
  - New insights about physiological processes in normal and disease states.
  - Incorporating information about genetic variation into human genetic studies



Grand Challenge I-4

- Understand evolutionary variation across species and the mechanisms underlying it

Grand Challenge I-4

- The genome is a dynamic structure, continually subjected to modification by the forces of evolution.
  - A parallel understanding of the sequence differences across species
  - Inter-species sequence comparisons — for identifying functional elements in the genome
  - Insight into the distinct anatomical, physiological and developmental features of different organisms — genetic basis for speciation
  - The fundamental processes that have sculpted their genomes
  - The characterization of mutational processes
  - Understanding of DNA mutation and repair

Grand Challenge I-5

- Develop policy options that facilitate the widespread use of genome information in both research and clinical settings

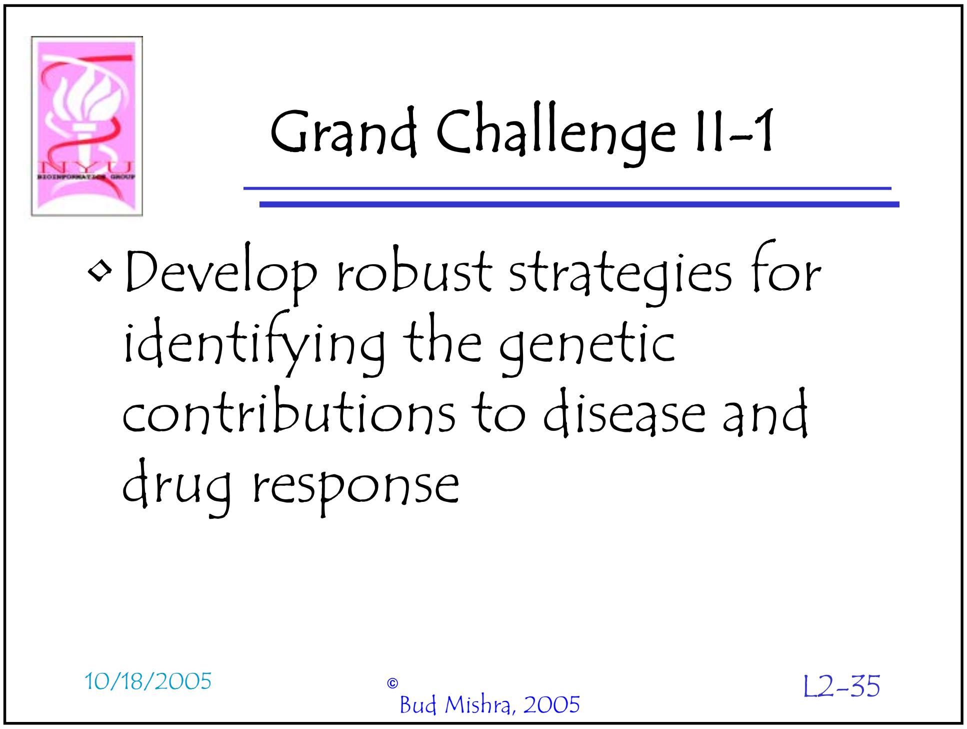
Grand Challenge I-5

- Access to the data about
  - genes, gene variants, haplotypes,
  - protein structures,
  - small molecules and
  - computational models.

II Genomics to health

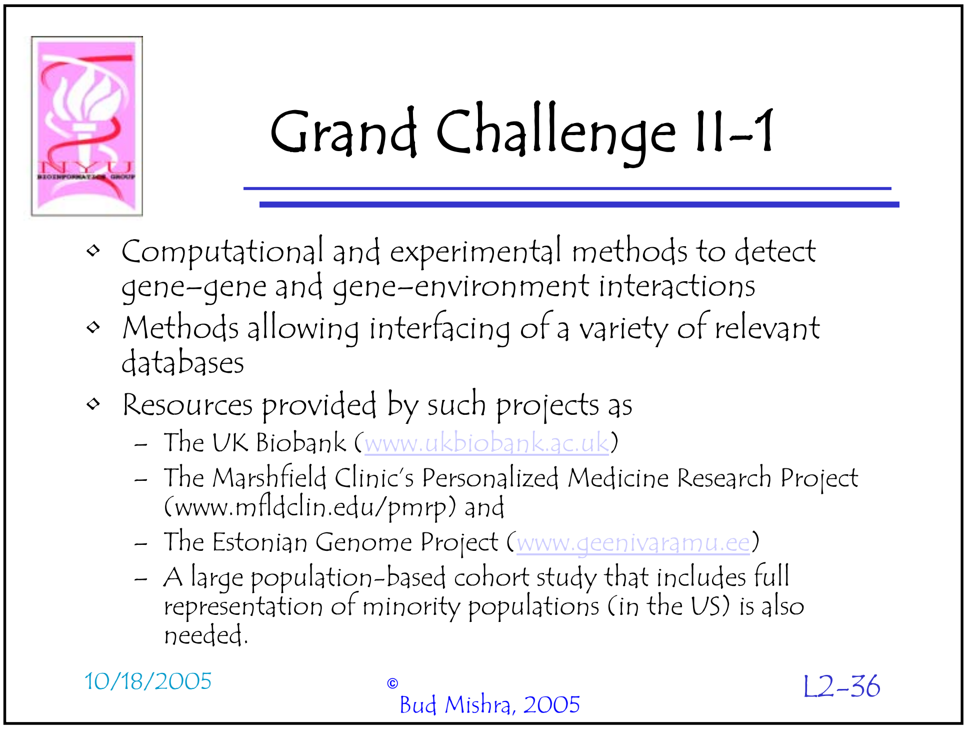
Translating genome-based knowledge into health benefits

- Identify genes and pathways with a role in health and disease, and determine how they interact with environmental factors.
- Develop, evaluate and apply genome-based diagnostic methods for the prediction of susceptibility to disease, the prediction of drug response, the early detection of illness and the accurate molecular classification of disease.
- Develop and deploy methods that catalyze the translation of genomic information into therapeutic advances.



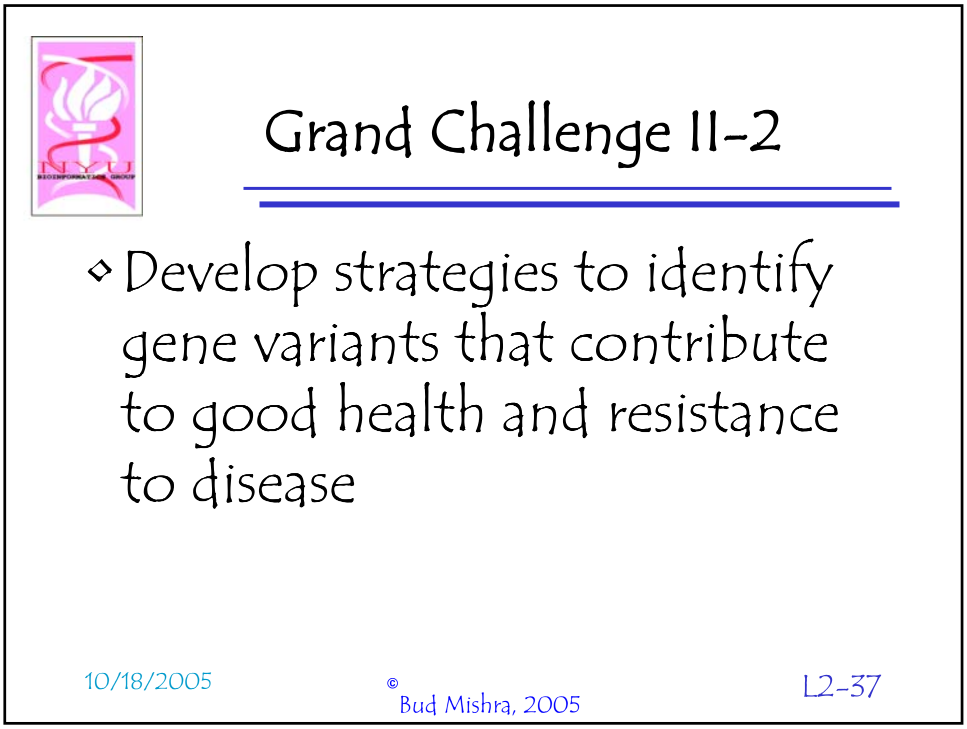
Grand Challenge II-1

- Develop robust strategies for identifying the genetic contributions to disease and drug response



Grand Challenge II-1

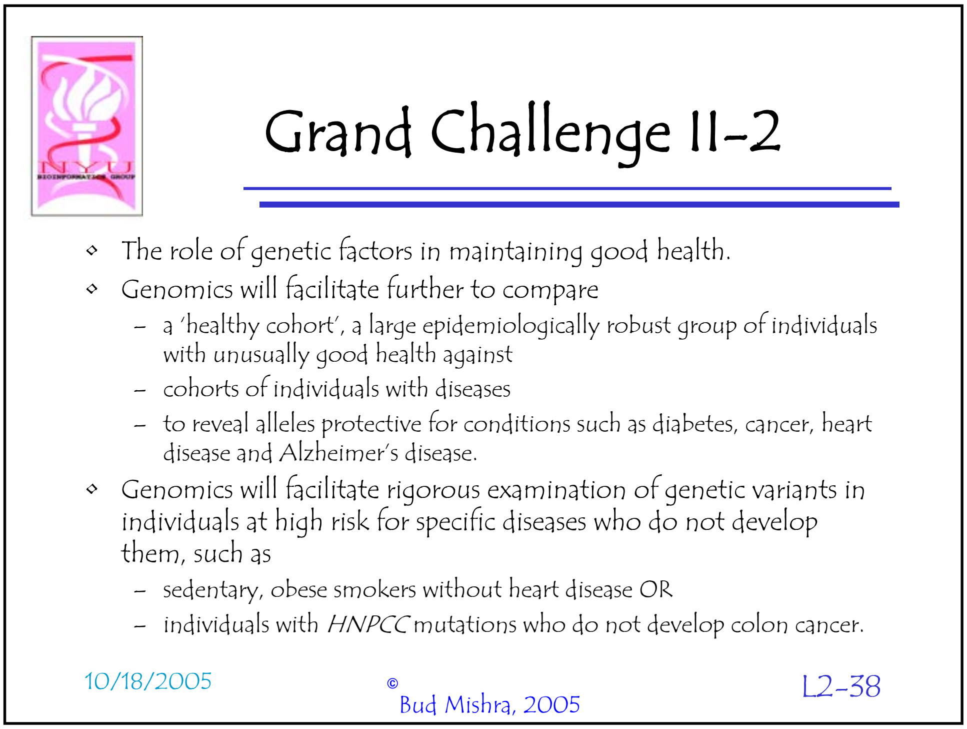
- Computational and experimental methods to detect gene-gene and gene-environment interactions
- Methods allowing interfacing of a variety of relevant databases
- Resources provided by such projects as
  - The UK Biobank (www.ukbiobank.ac.uk)
  - The Marshfield Clinic's Personalized Medicine Research Project (www.mfldclin.edu/pmrp) and
  - The Estonian Genome Project (www.geenivaramu.ee)
  - A large population-based cohort study that includes full representation of minority populations (in the US) is also needed.



Grand Challenge II-2

- Develop strategies to identify gene variants that contribute to good health and resistance to disease

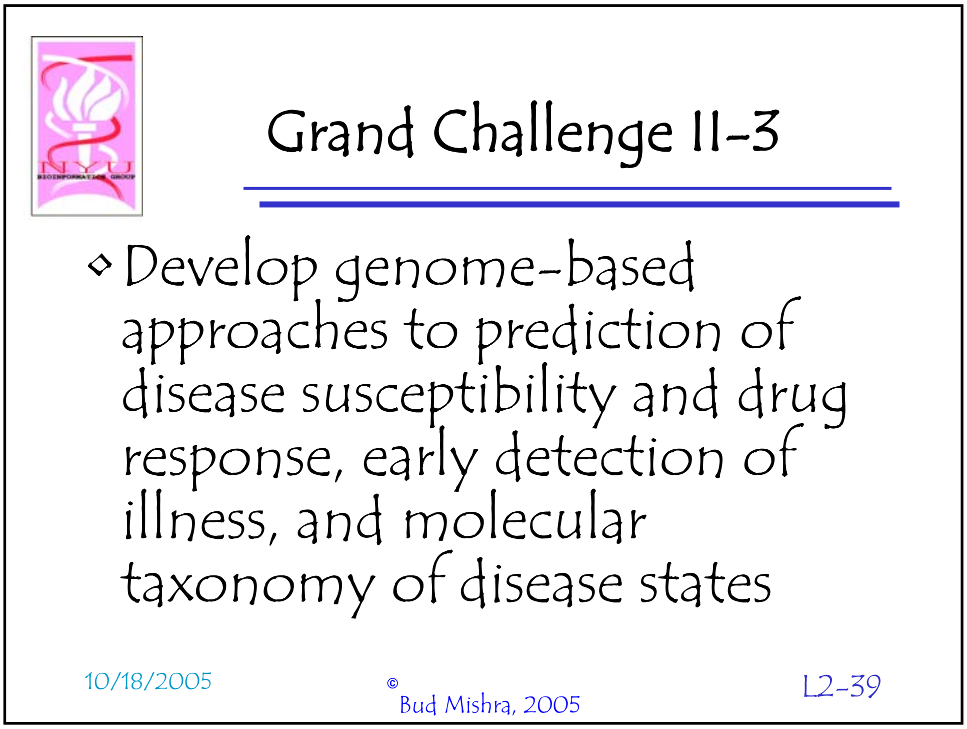
10/18/2005 © Bud Mishra, 2005 L2-37



Grand Challenge II-2

- The role of genetic factors in maintaining good health.
- Genomics will facilitate further to compare
  - a 'healthy cohort', a large epidemiologically robust group of individuals with unusually good health against
  - cohorts of individuals with diseases
  - to reveal alleles protective for conditions such as diabetes, cancer, heart disease and Alzheimer's disease.
- Genomics will facilitate rigorous examination of genetic variants in individuals at high risk for specific diseases who do not develop them, such as
  - sedentary, obese smokers without heart disease OR
  - individuals with *HNPCC* mutations who do not develop colon cancer.

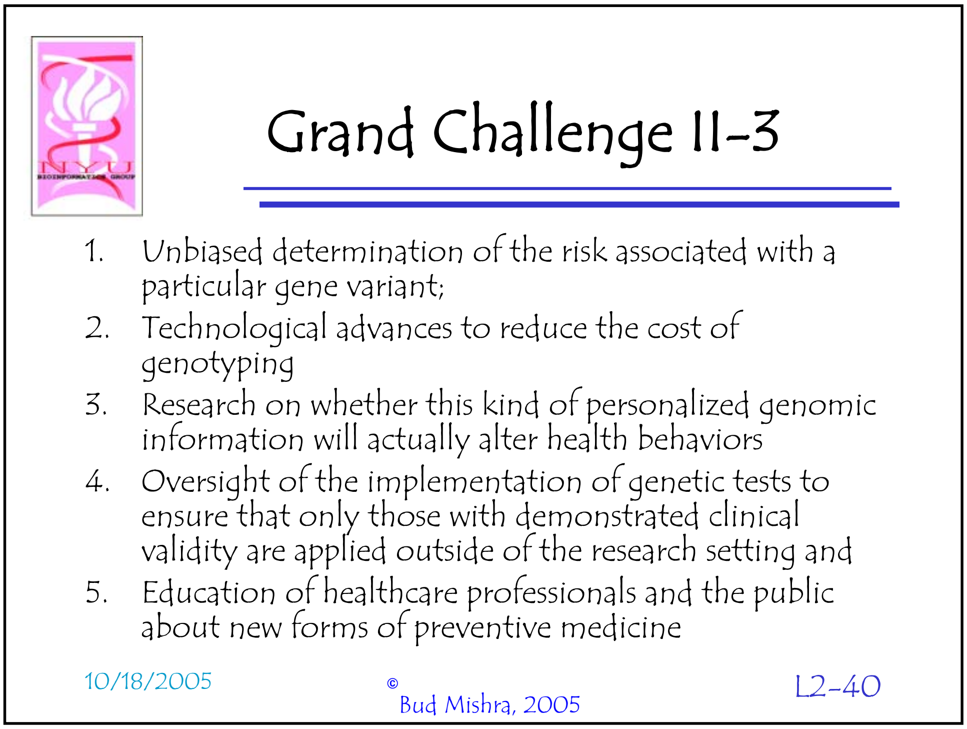
10/18/2005 © Bud Mishra, 2005 L2-38



Grand Challenge II-3

- Develop genome-based approaches to prediction of disease susceptibility and drug response, early detection of illness, and molecular taxonomy of disease states

10/18/2005 © Bud Mishra, 2005 L2-39



Grand Challenge II-3

1. Unbiased determination of the risk associated with a particular gene variant;
2. Technological advances to reduce the cost of genotyping
3. Research on whether this kind of personalized genomic information will actually alter health behaviors
4. Oversight of the implementation of genetic tests to ensure that only those with demonstrated clinical validity are applied outside of the research setting and
5. Education of healthcare professionals and the public about new forms of preventive medicine

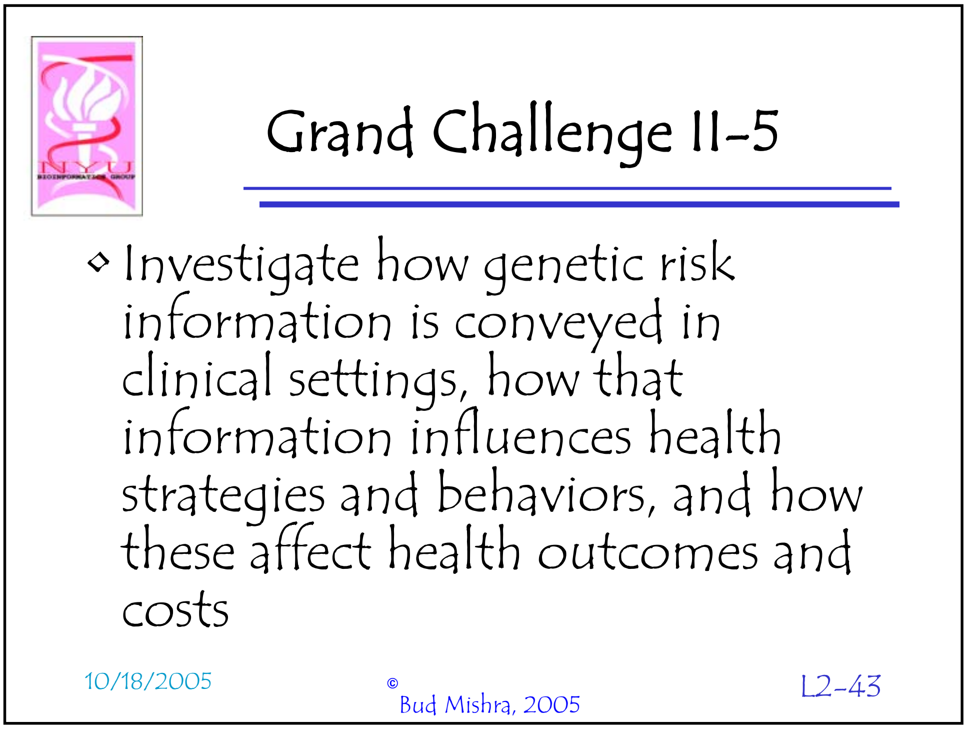
10/18/2005 © Bud Mishra, 2005 L2-40

Grand Challenge II-4

- Use new understanding of genes and pathways to develop powerful new therapeutic approaches to disease

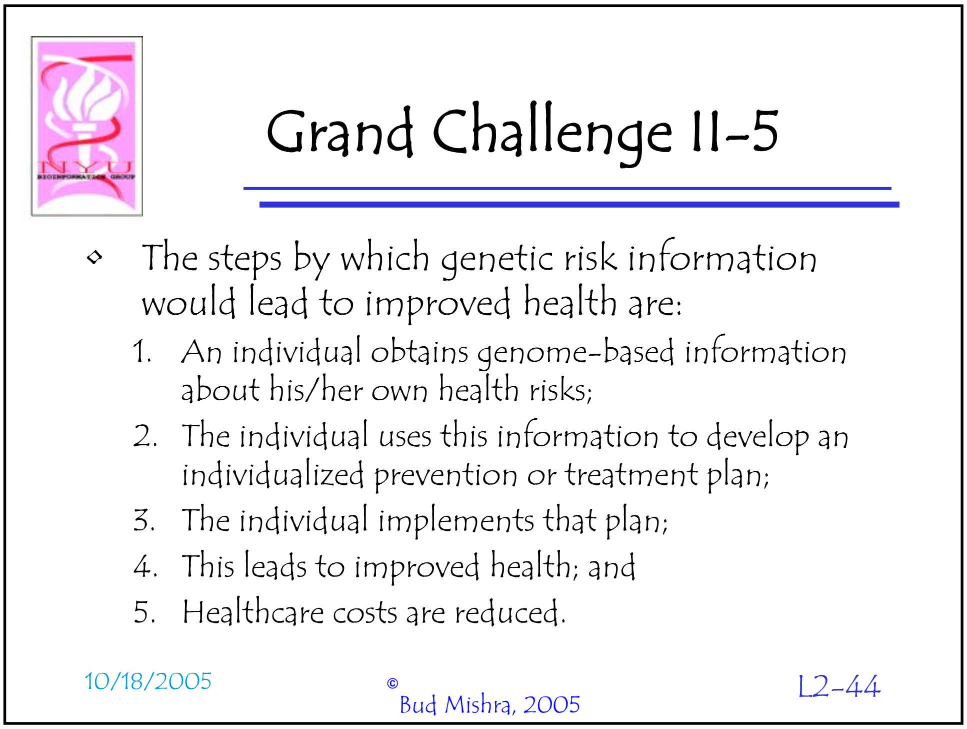
Grand Challenge II-4

- Pharmaceuticals on the market target fewer than 500 human gene products— out of the 30,000 or so human protein-coding genes
- The new understanding of biological pathways should open itself to better therapeutic design.
  - Imatinib mesylate (Gleevec), an inhibitor of the BCRABL tyrosine kinase, in treating chronic myelogenous leukaemia
- Application of 'chemical genomics'.
  - Using libraries of small molecules (natural compounds, aptamers or the products of combinatorial chemistry) and high-throughput screening to advance understanding of biological pathways and to identify compounds that act as positive or negative regulators of individual gene products, pathways or cellular phenotypes.



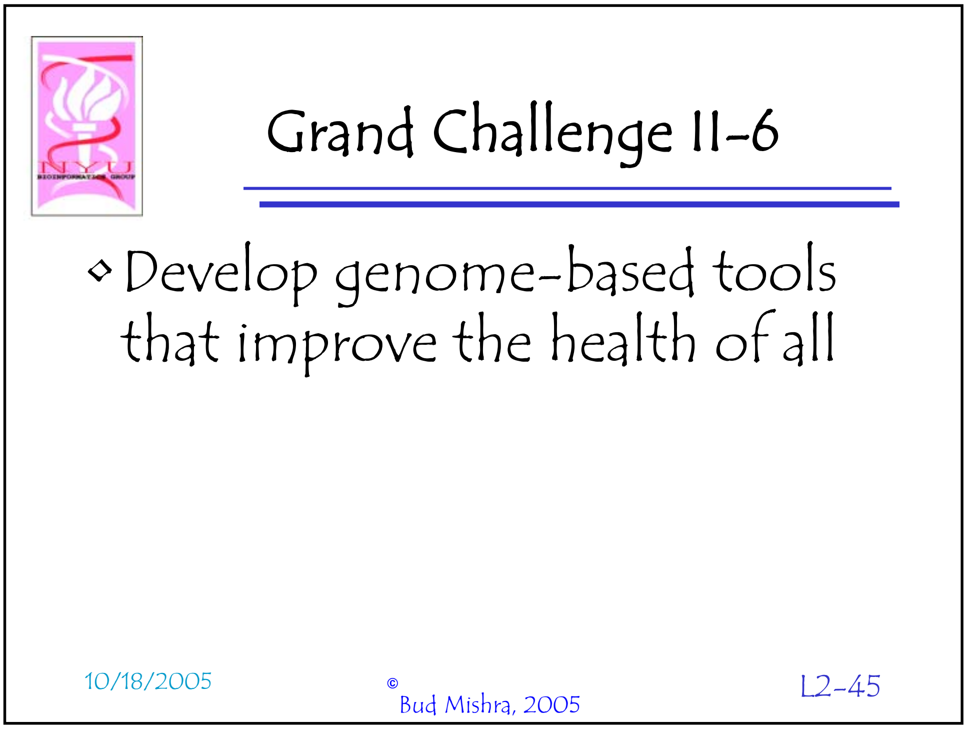
Grand Challenge II-5

- Investigate how genetic risk information is conveyed in clinical settings, how that information influences health strategies and behaviors, and how these affect health outcomes and costs



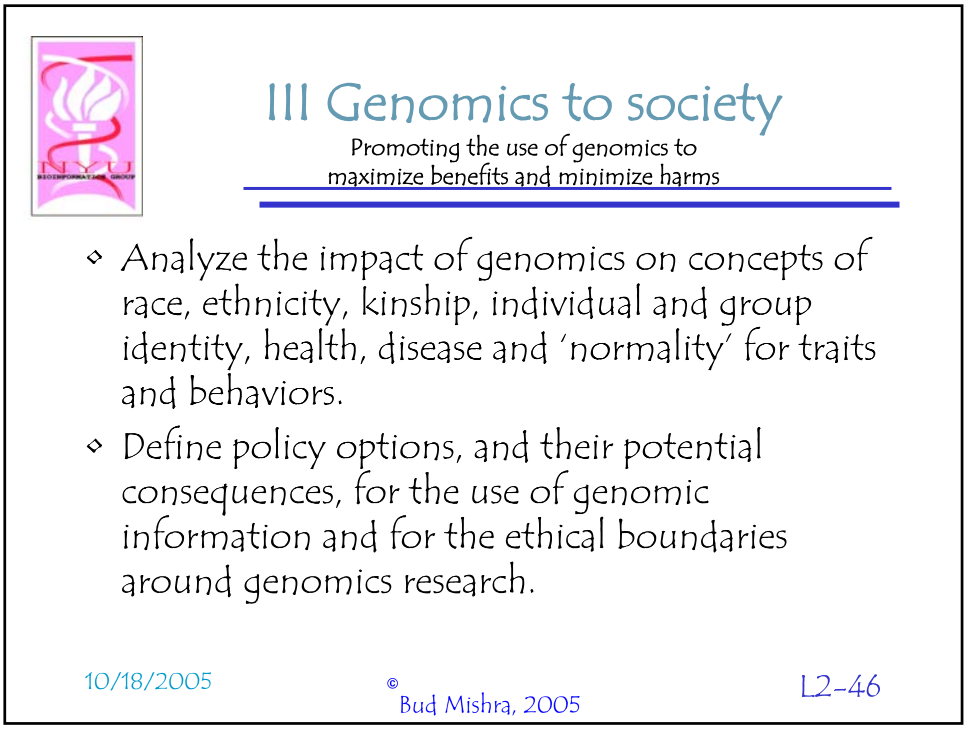
Grand Challenge II-5

- The steps by which genetic risk information would lead to improved health are:
  1. An individual obtains genome-based information about his/her own health risks;
  2. The individual uses this information to develop an individualized prevention or treatment plan;
  3. The individual implements that plan;
  4. This leads to improved health; and
  5. Healthcare costs are reduced.



Grand Challenge II-6

- Develop genome-based tools that improve the health of all



III Genomics to society

Promoting the use of genomics to
maximize benefits and minimize harms

- Analyze the impact of genomics on concepts of race, ethnicity, kinship, individual and group identity, health, disease and 'normality' for traits and behaviors.
- Define policy options, and their potential consequences, for the use of genomic information and for the ethical boundaries around genomics research.

Grand Challenge III-1

- Develop policy options for the uses of genomics in medical and non-medical settings

Grand Challenge III-2

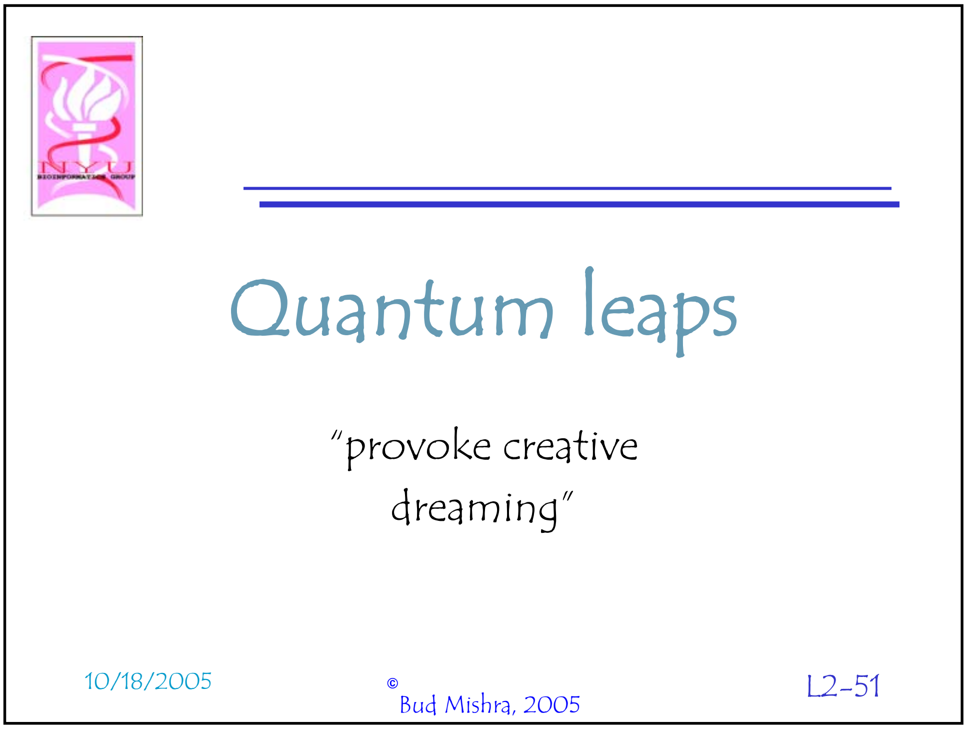
- Understand the relationships between genomics, race and ethnicity, and the consequences of uncovering these relationships

Grand Challenge III-3

- Understand the consequences of uncovering the genomic contributions to human traits and behaviors

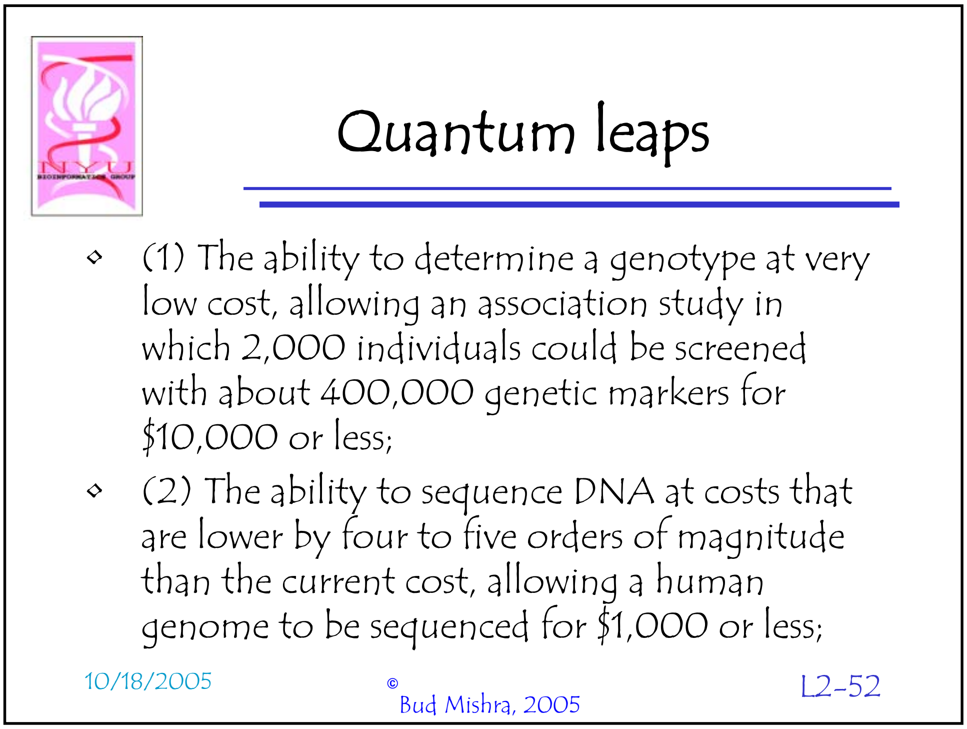
Grand Challenge III-4

- Assess how to define the ethical boundaries for uses of genomics



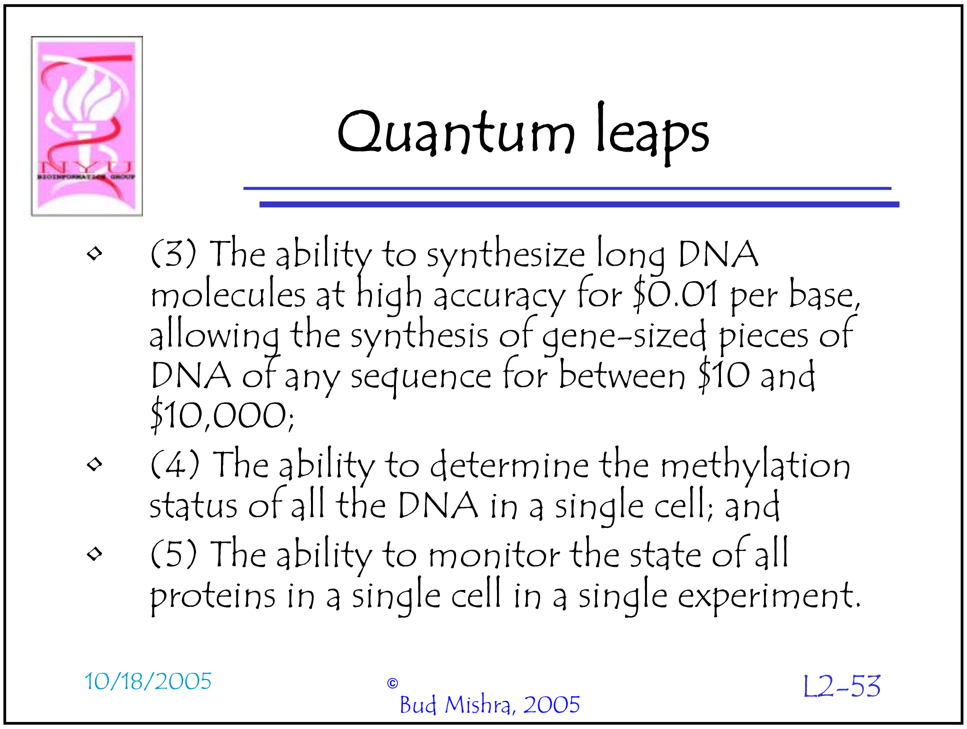
Quantum leaps

"provoke creative dreaming"



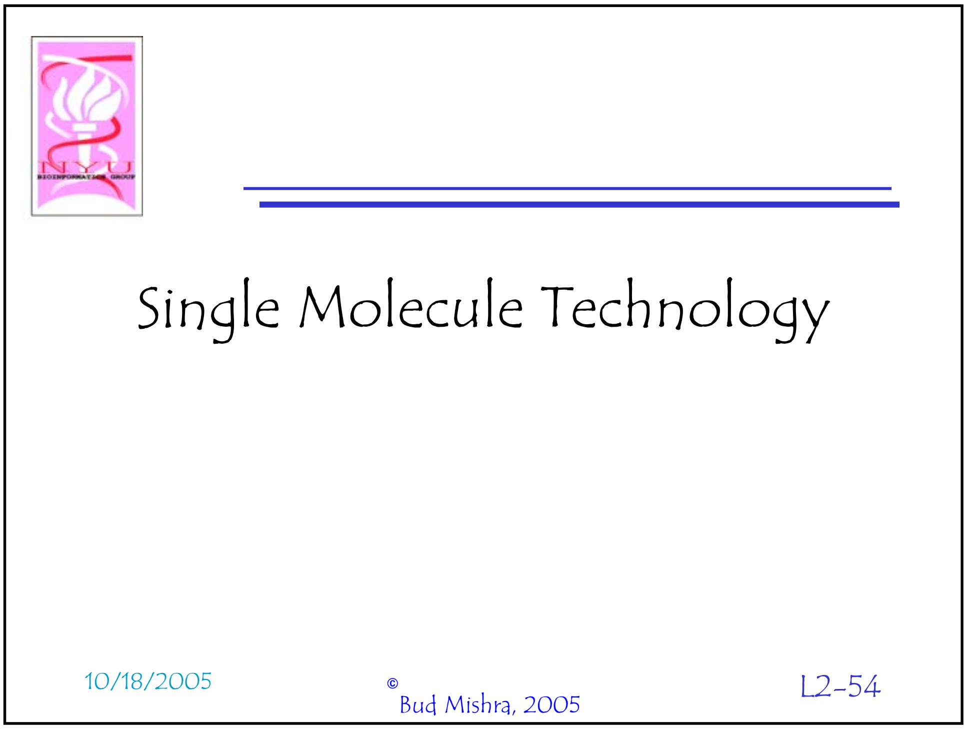
Quantum leaps

- (1) The ability to determine a genotype at very low cost, allowing an association study in which 2,000 individuals could be screened with about 400,000 genetic markers for $10,000 or less;
- (2) The ability to sequence DNA at costs that are lower by four to five orders of magnitude than the current cost, allowing a human genome to be sequenced for $1,000 or less;



Quantum leaps

- (3) The ability to synthesize long DNA molecules at high accuracy for $0.01 per base, allowing the synthesis of gene-sized pieces of DNA of any sequence for between $10 and $10,000;
- (4) The ability to determine the methylation status of all the DNA in a single cell; and
- (5) The ability to monitor the state of all proteins in a single cell in a single experiment.



Single Molecule Technology

Error Sources

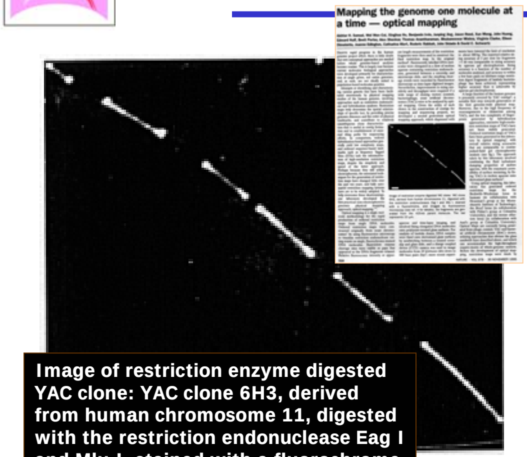
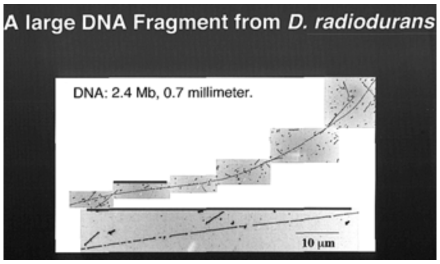


Image of restriction enzyme digested YAC clone: YAC clone 6H3, derived from human chromosome 11, digested with the restriction endonuclease Eag I and Mlu I, stained with a fluorochrome and imaged by fluorescence microscopy.

- Sizing Error
  - (Bernoulli labeling, absorption cross-section, PSF)
- Partial Digestion
- False Optical Sites
- Orientation
- Spurious molecules, Optical chimerism, Calibration

Shotgun Optical Mapping


A large DNA Fragment from D. radiodurans
DNA: 2.4 Mb, 0.7 millimeter.
10 µm

- Large fragments of genomic DNA of length from 2Mb to 12Mb are optically mapped
- The resulting ordered restriction maps are automatically contiged by "Gentig"
- The consensus map computed by Gentig is free of errors due to partial digestion, sizing error and false cuts

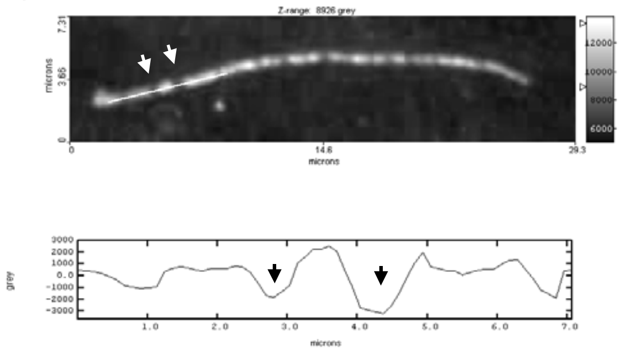
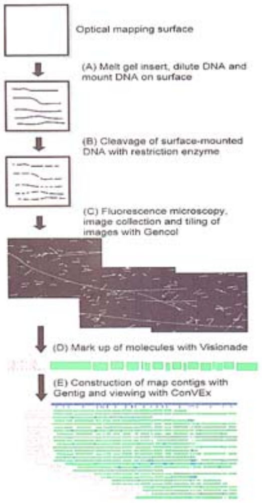


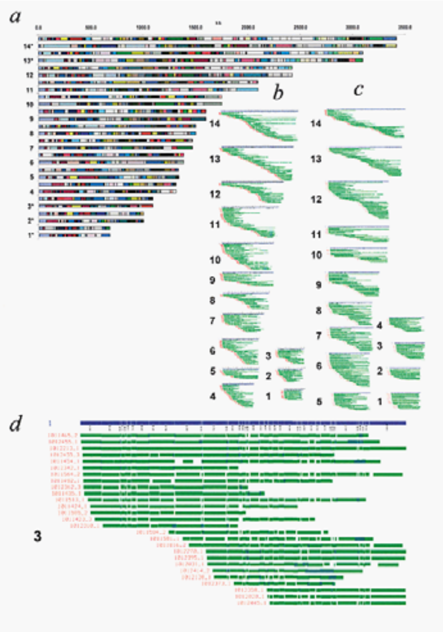
Figure 2

Shotgun Mapping


Optical mapping surface
(A) Melt gel insert, dilute DNA and mount DNA on surface
(B) Cleavage of surface-mounted DNA with restriction enzyme
(C) Fluorescence microscopy, image collection and tiling of images with Gencol
(D) Mark up of molecules with Visionade
(E) Construction of map contigs with Gentig and viewing with ConVEx

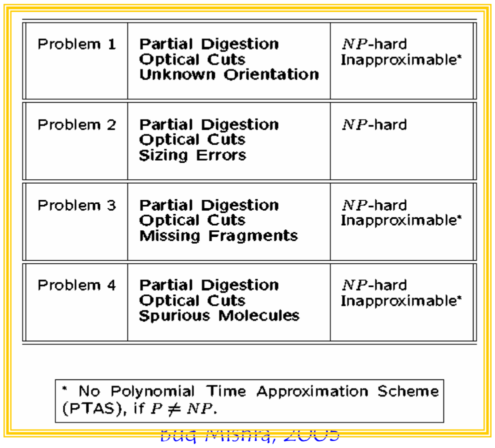
- Schematics
  - Experiment Design
  - Robotics
  - BioChemistry
  - Imaging
  - Image Analysis
  - Statistical Algorithms
  - Visualization

Malaria Parasite: P. falciparum



Complexity Issues

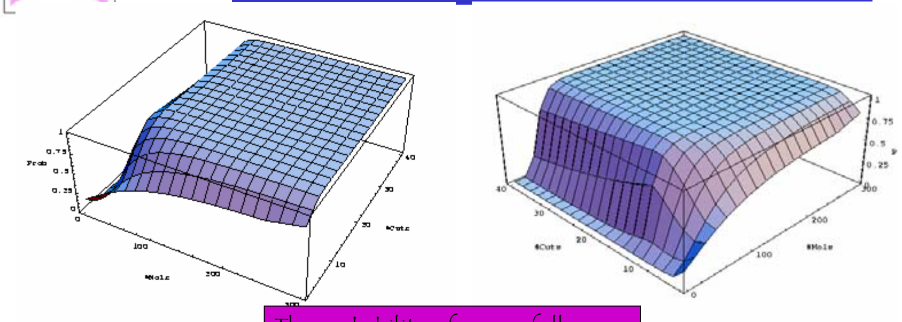
Various combinations of error sources lead to NP-hard Problems



Problem 1	**Partial Digestion Optical Cuts Unknown Orientation**	NP-hard Inapproximable*
Problem 2	**Partial Digestion Optical Cuts Sizing Errors**	NP-hard
Problem 3	**Partial Digestion Optical Cuts Missing Fragments**	NP-hard Inapproximable*
Problem 4	**Partial Digestion Optical Cuts Spurious Molecules**	NP-hard Inapproximable*

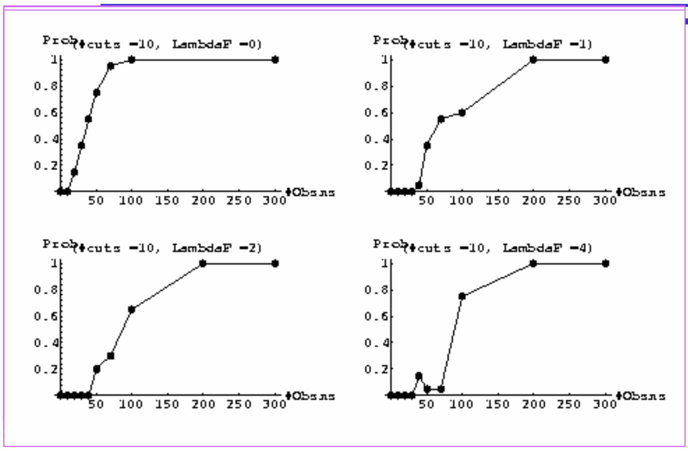
* No Polynomial Time Approximation Scheme (PTAS), if $P \neq NP$.

Prediction

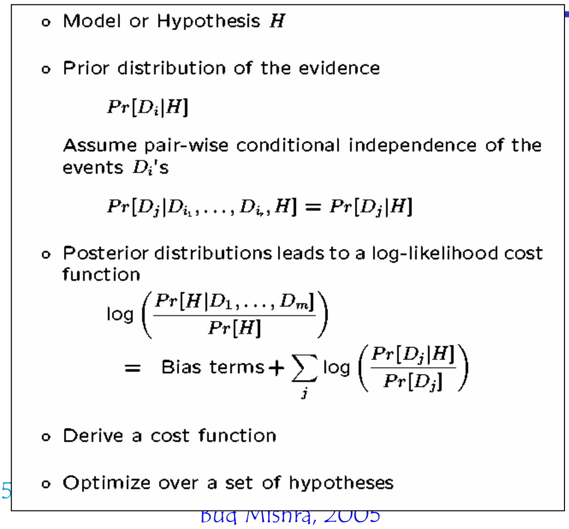


The probability of successfully computing the correct restriction map as a function of the number of cuts in the map and number of molecules used in creating the map…

Experimental Results



Bayesian Approach



- Model or Hypothesis H
- Prior distribution of the evidence

  $$Pr[D_i|H]$$

  Assume pair-wise conditional independence of the events D_i's

  $$Pr[D_j|D_{i_1},\ldots,D_{i_k},H] = Pr[D_j|H]$$

- Posterior distributions leads to a log-likelihood cost function

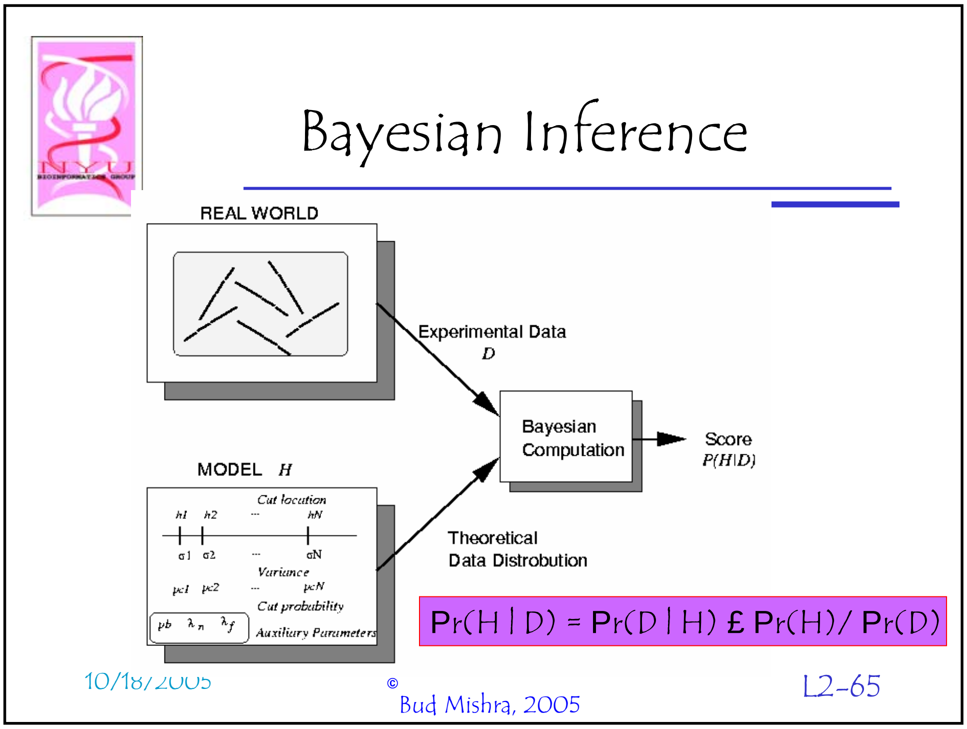
  $$\log\left(\frac{Pr[H|D_1,\ldots,D_m]}{Pr[H]}\right) = \text{Bias terms} + \sum_j \log\left(\frac{Pr[D_j|H]}{Pr[D_j]}\right)$$

- Derive a cost function
- Optimize over a set of hypotheses

Robustness of Optical Mapping Algorithm

- Parameters:
  - Digestion rate can be as low as 10%
  - Orientation of DNA need not be known.
  - 40% foreign DNA
  - 85% DNA partially broken
  - Relative sizing error up to 30%
  - 30% spurious randomly located cuts…

Algorithm Design and
Analysis jointly with
T.S.Anantharaman



Bayesian Inference

REAL WORLD

Experimental Data
D

Bayesian Computation

Score
$P(H|D)$

MODEL H

Cut location

$h1$ $h2$ ... hN

$\sigma 1$ $\sigma 2$ ... σN

Variance

$pc1$ $pc2$ ... pcN

Cut probability

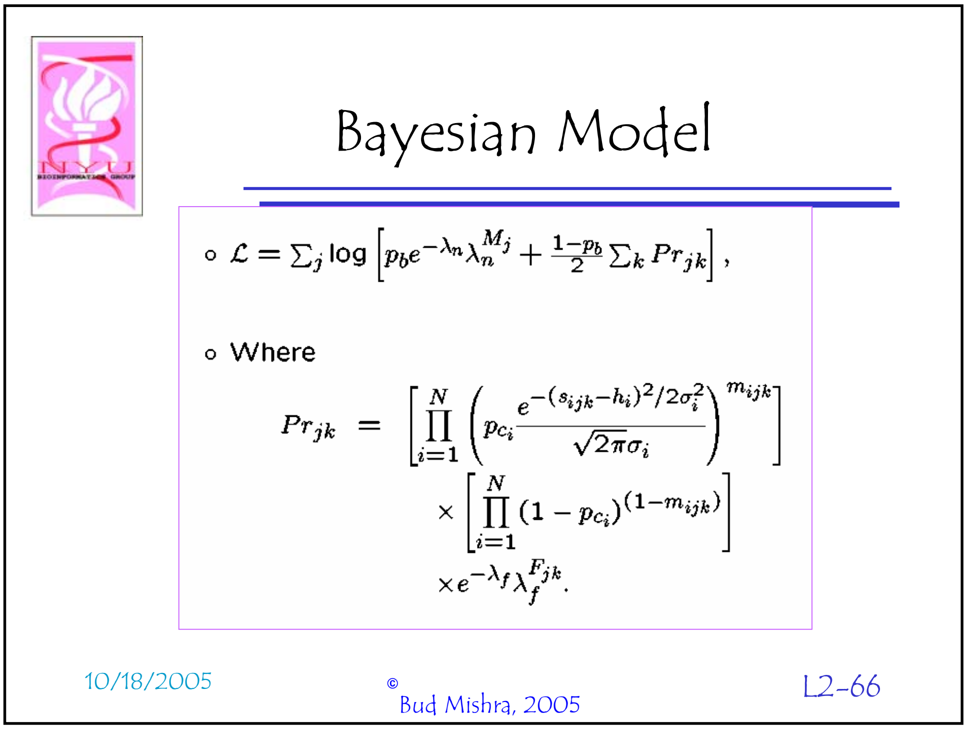
pb λ_n λ_f

Auxiliary Parameters

Theoretical
Data Distrobution

Pr(H | D) = Pr(D | H) £ Pr(H)/ Pr(D)

10/18/2005 © Bud Mishra, 2005 L2-65



Bayesian Model

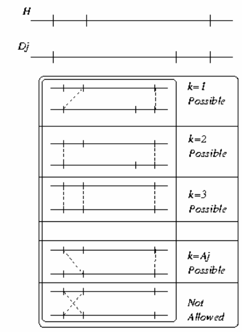
- $\mathcal{L} = \sum_j \log \left[p_b e^{-\lambda_n} \lambda_n^{M_j} + \frac{1-p_b}{2} \sum_k Pr_{jk} \right],$

- Where

$$Pr_{jk} = \left[\prod_{i=1}^{N} \left(p_{c_i} \frac{e^{-(s_{ijk}-h_i)^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i} \right)^{m_{ijk}} \right] \times \left[\prod_{i=1}^{N} (1-p_{c_i})^{(1-m_{ijk})} \right] \times e^{-\lambda_f} \lambda_f^{F_{jk}}.$$

10/18/2005 © Bud Mishra, 2005 L2-66

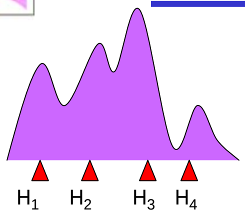
Multiple Alignement



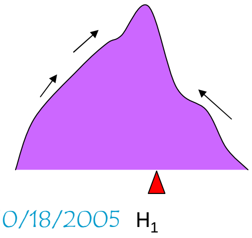
- Various alignments of cuts have to be considered.
- Fast computation is possible... via *Dynamic Programming* and additional heuristics

—Key to our fast implementation.

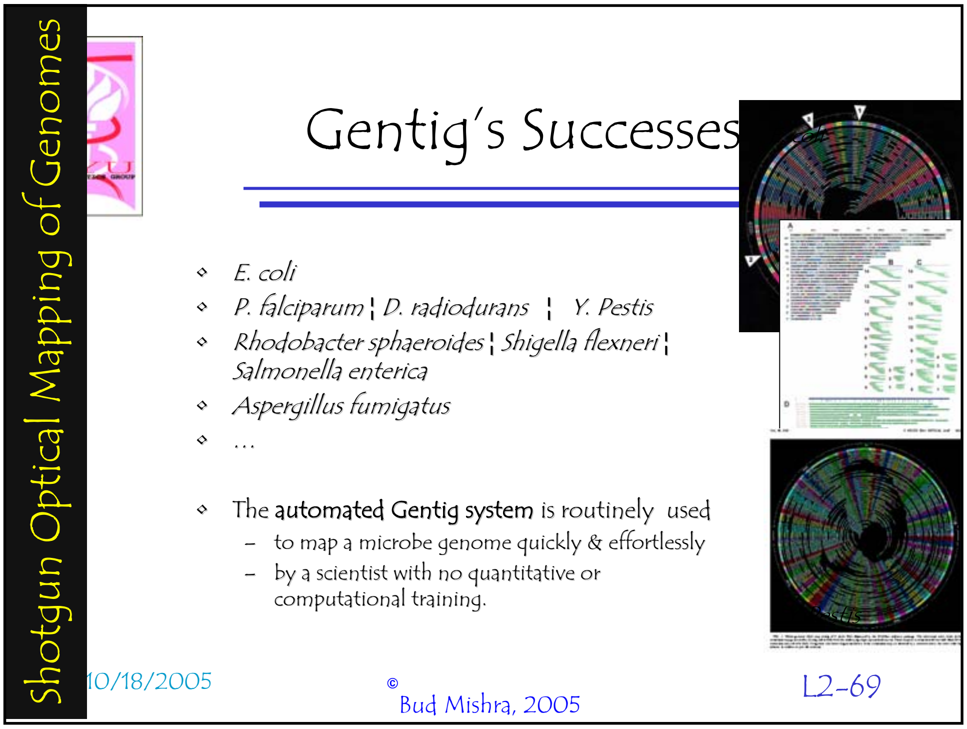
Bayesian Optimization...



Gradient search for good parameters



Local gradient optimization



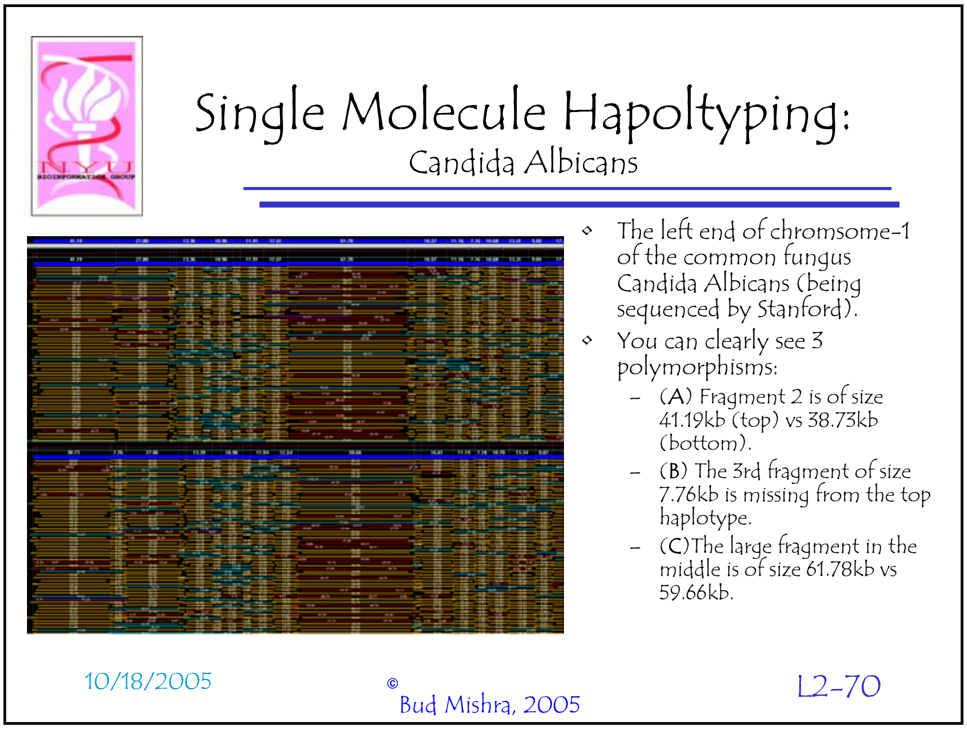
Shotgun Optical Mapping of Genomes

Gentig's Successes

- *E. coli*
- *P. falciparum* ¦ *D. radiodurans* ¦ *Y. Pestis*
- *Rhodobacter sphaeroides* ¦ *Shigella flexneri* ¦ *Salmonella enterica*
- *Aspergillus fumigatus*
- …

- The **automated Gentig system** is routinely used
  - to map a microbe genome quickly & effortlessly
  - by a scientist with no quantitative or computational training.

10/18/2005 © Bud Mishra, 2005 L2-69



Single Molecule Hapoltyping:

Candida Albicans

- The left end of chromsome-1 of the common fungus Candida Albicans (being sequenced by Stanford).
- You can clearly see 3 polymorphisms:
  - (A) Fragment 2 is of size 41.19kb (top) vs 38.73kb (bottom).
  - (B) The 3rd fragment of size 7.76kb is missing from the top haplotype.
  - (C)The large fragment in the middle is of size 61.78kb vs 59.66kb.

10/18/2005 © Bud Mishra, 2005 L2-70

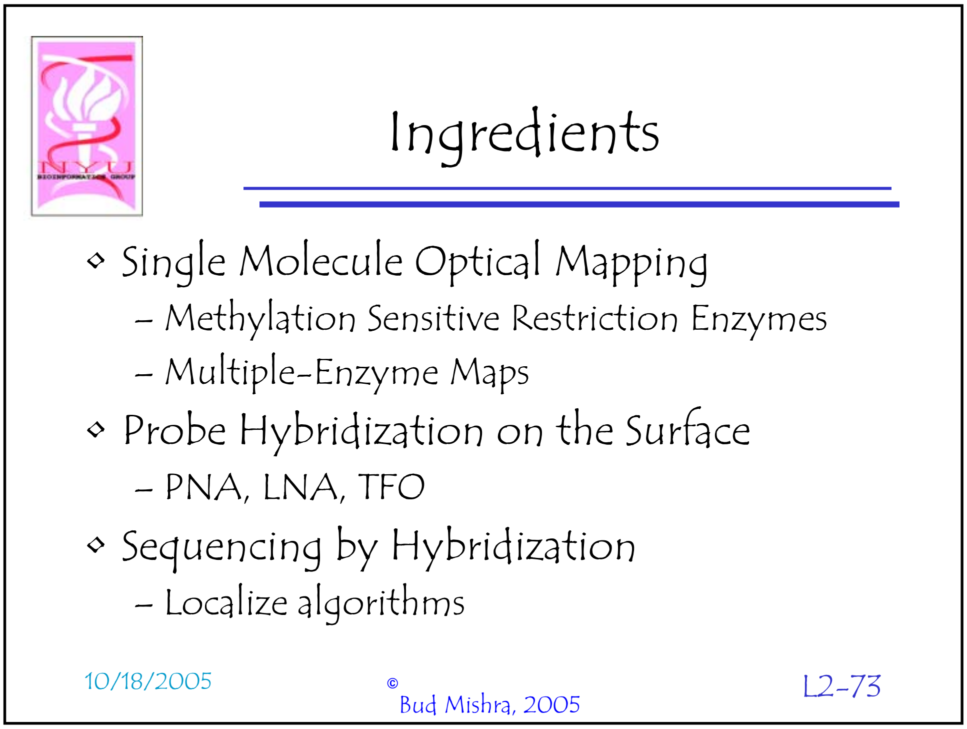
Some Interesting Applications

- Sequence Validation
- Haplotyping
- Sequencing
- Comparative Genomics
- Rearrangement events
- Hemizygous Deletions
- Epigenomics
- Characterizing cDNAs
  - Expression Profiling
  - Alternate Splicing

Sequencing in Post-Genomic Era

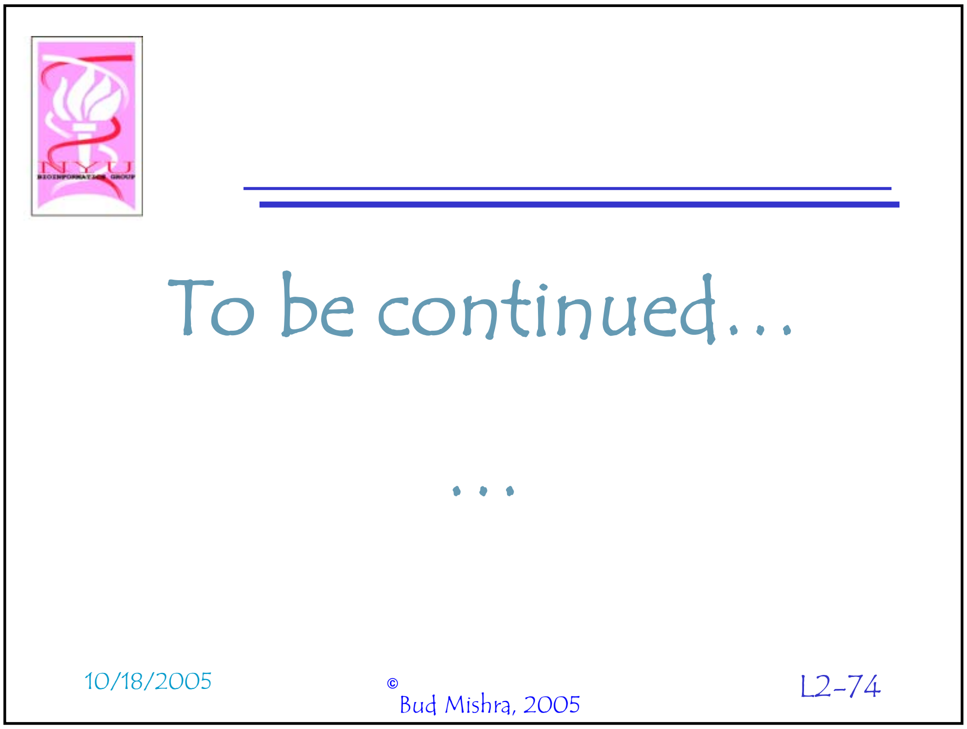
- Haplotypic Sequencing of 6.6 Billion Base Pairs in a Diploid Human genome
- Less than $700
- Less than 24 Hours
- Draft Quality (Not Resequencing)

  – with Anantharaman, Cantor, Demidov, Gimzewski, Reed, Teitell



Ingredients

- Single Molecule Optical Mapping
  - Methylation Sensitive Restriction Enzymes
  - Multiple-Enzyme Maps
- Probe Hybridization on the Surface
  - PNA, LNA, TFO
- Sequencing by Hybridization
  - Localize algorithms



To be continued…

…