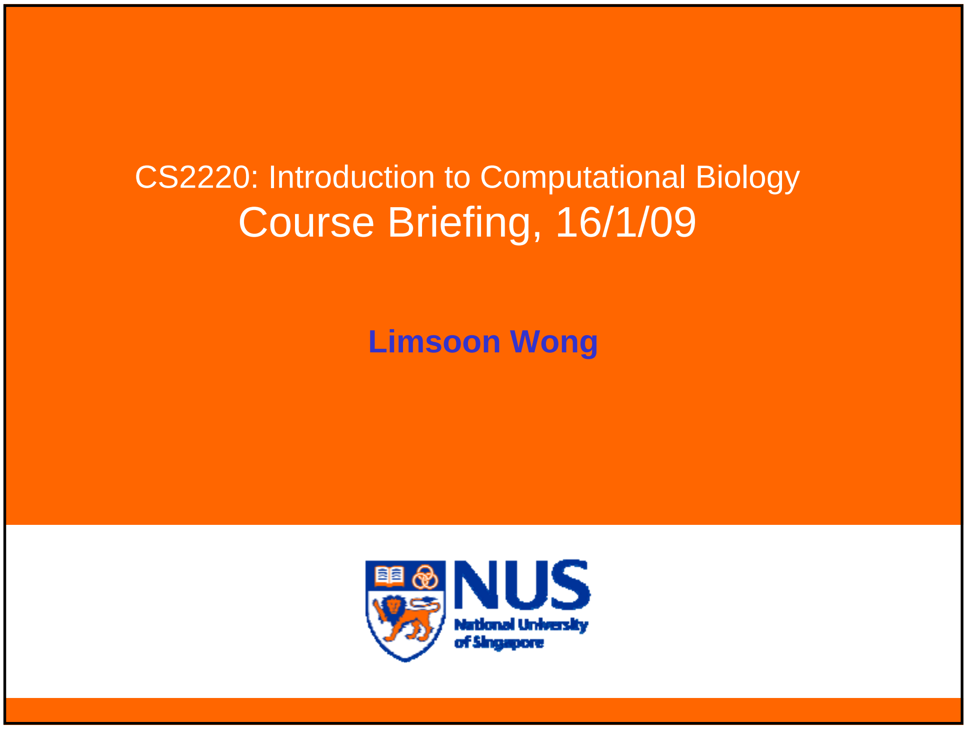
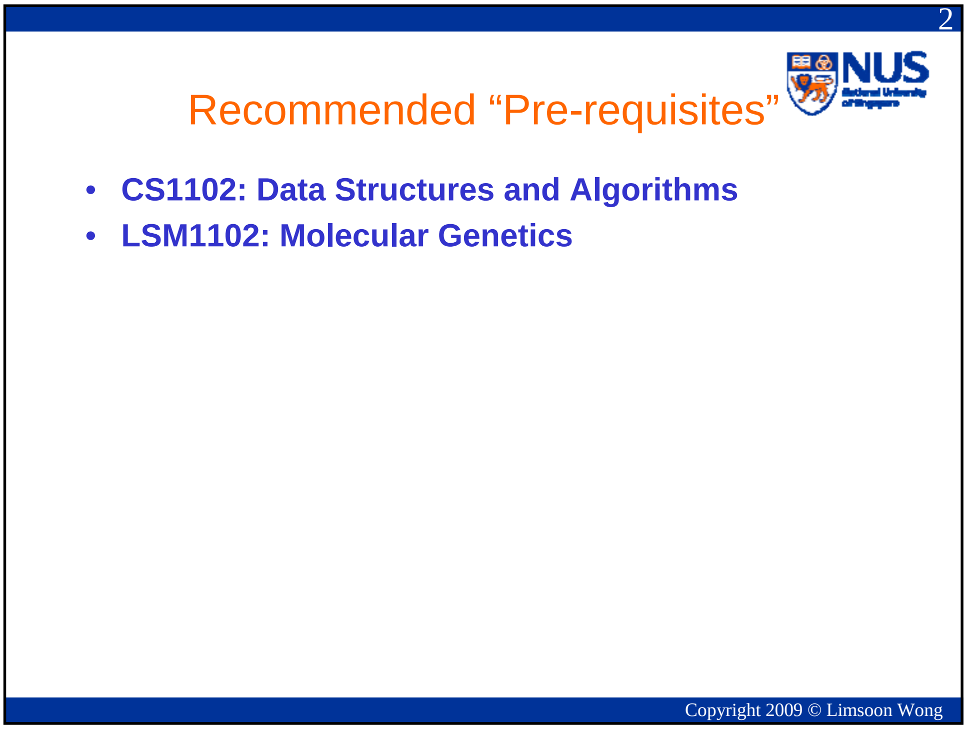


CS2220: Introduction to Computational Biology

Course Briefing, 16/1/09

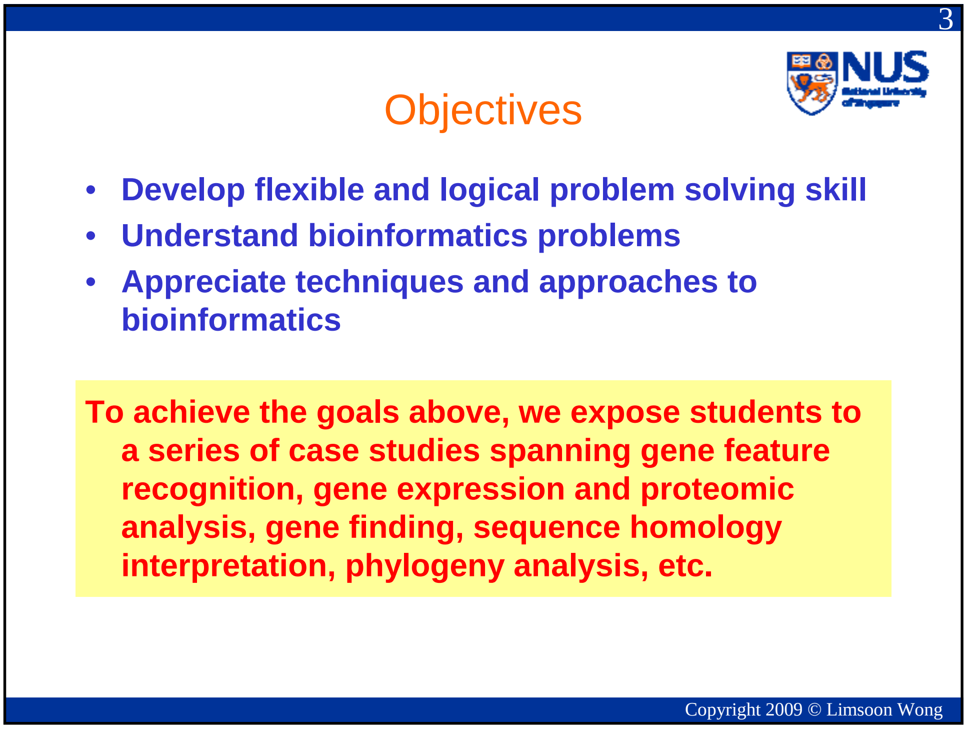
Limsoon Wong



2

Recommended “Pre-requisites”

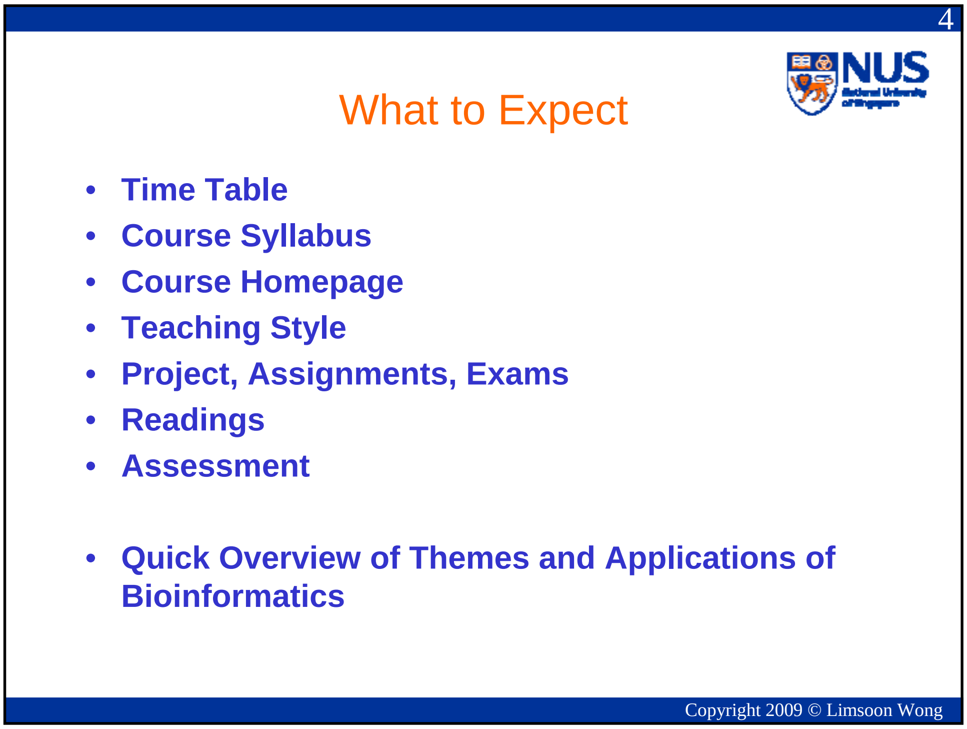
- **CS1102: Data Structures and Algorithms**
- **LSM1102: Molecular Genetics**



Objectives

- Develop flexible and logical problem solving skill
- Understand bioinformatics problems
- Appreciate techniques and approaches to bioinformatics

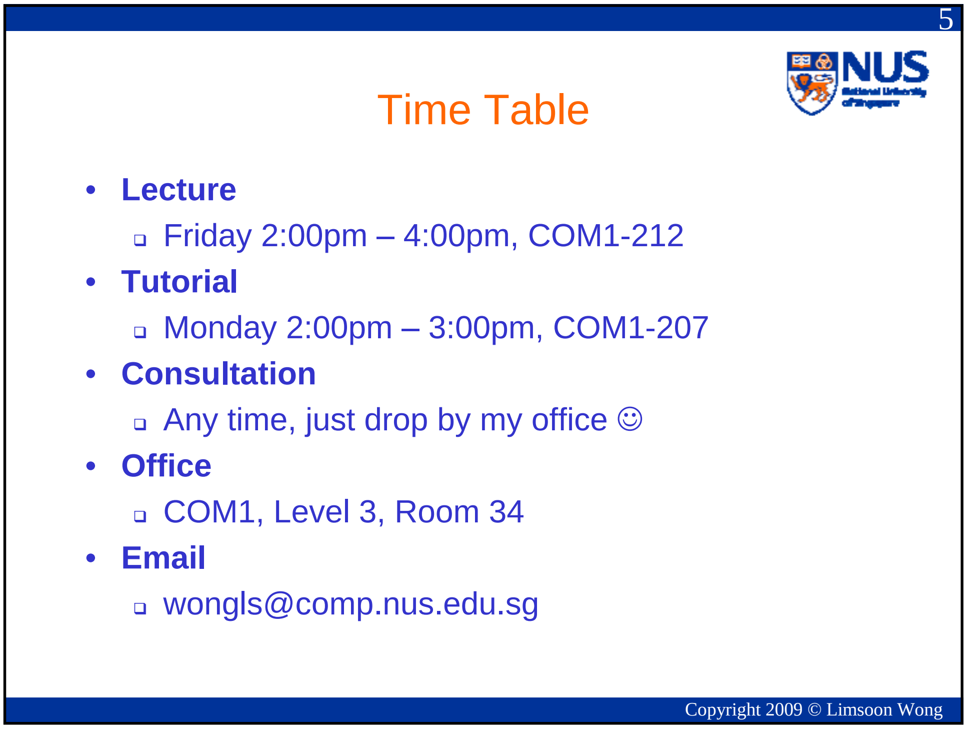
To achieve the goals above, we expose students to a series of case studies spanning gene feature recognition, gene expression and proteomic analysis, gene finding, sequence homology interpretation, phylogeny analysis, etc.



What to Expect

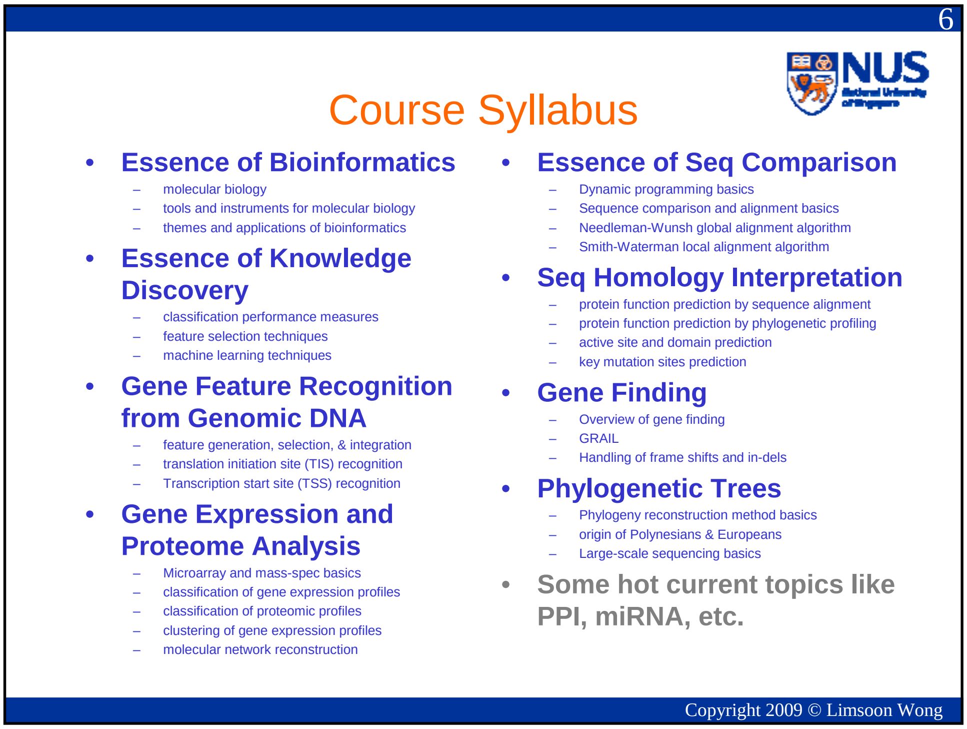
- Time Table
- Course Syllabus
- Course Homepage
- Teaching Style
- Project, Assignments, Exams
- Readings
- Assessment

- Quick Overview of Themes and Applications of Bioinformatics



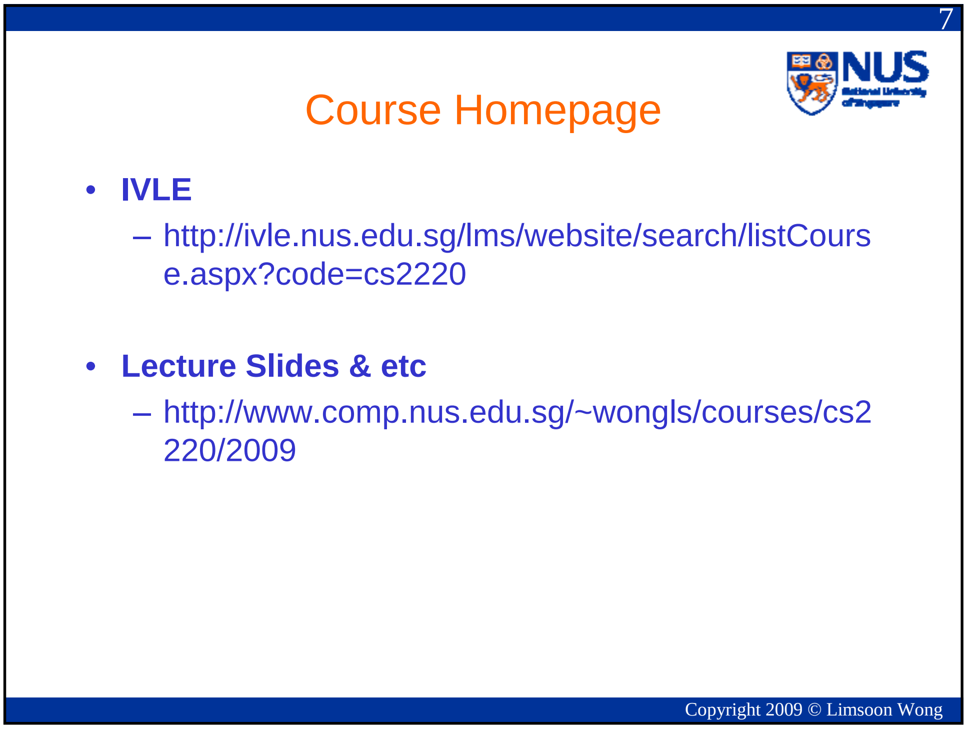
Time Table

- **Lecture**
  - Friday 2:00pm – 4:00pm, COM1-212
- **Tutorial**
  - Monday 2:00pm – 3:00pm, COM1-207
- **Consultation**
  - Any time, just drop by my office ☺
- **Office**
  - COM1, Level 3, Room 34
- **Email**
  - wongls@comp.nus.edu.sg



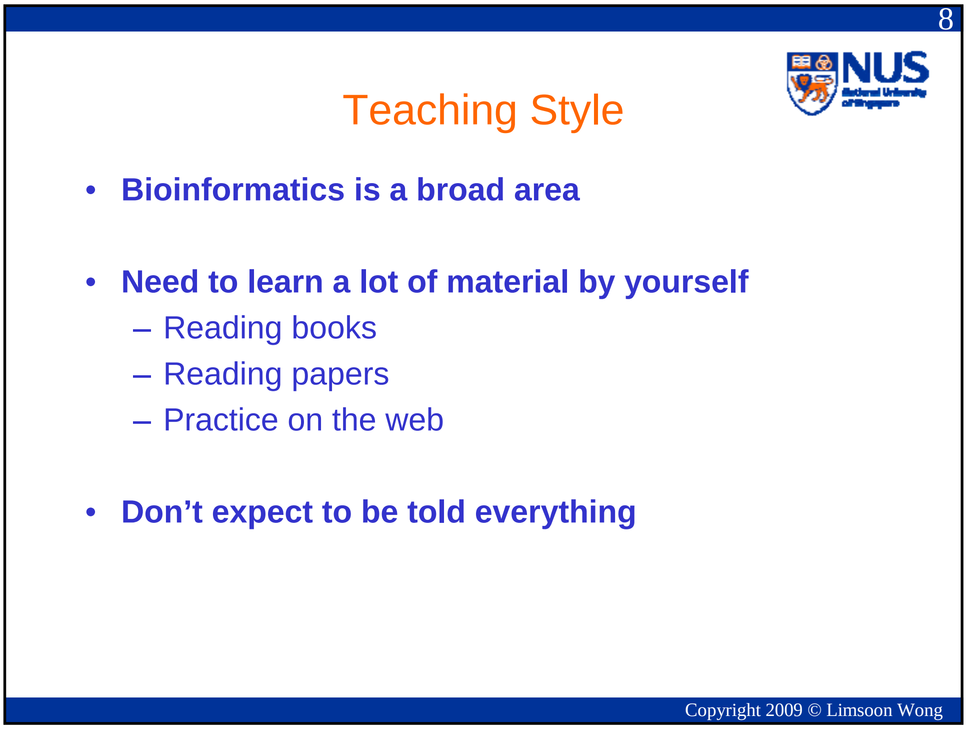
Course Syllabus

- **Essence of Bioinformatics**
  - molecular biology
  - tools and instruments for molecular biology
  - themes and applications of bioinformatics
- **Essence of Knowledge Discovery**
  - classification performance measures
  - feature selection techniques
  - machine learning techniques
- **Gene Feature Recognition from Genomic DNA**
  - feature generation, selection, & integration
  - translation initiation site (TIS) recognition
  - Transcription start site (TSS) recognition
- **Gene Expression and Proteome Analysis**
  - Microarray and mass-spec basics
  - classification of gene expression profiles
  - classification of proteomic profiles
  - clustering of gene expression profiles
  - molecular network reconstruction
- **Essence of Seq Comparison**
  - Dynamic programming basics
  - Sequence comparison and alignment basics
  - Needleman-Wunsh global alignment algorithm
  - Smith-Waterman local alignment algorithm
- **Seq Homology Interpretation**
  - protein function prediction by sequence alignment
  - protein function prediction by phylogenetic profiling
  - active site and domain prediction
  - key mutation sites prediction
- **Gene Finding**
  - Overview of gene finding
  - GRAIL
  - Handling of frame shifts and in-dels
- **Phylogenetic Trees**
  - Phylogeny reconstruction method basics
  - origin of Polynesians & Europeans
  - Large-scale sequencing basics
- **Some hot current topics like PPI, miRNA, etc.**



Course Homepage

- **IVLE**
  - http://ivle.nus.edu.sg/lms/website/search/listCourse.aspx?code=cs2220
- **Lecture Slides & etc**
  - http://www.comp.nus.edu.sg/~wongls/courses/cs2220/2009



Teaching Style

- **Bioinformatics is a broad area**
- **Need to learn a lot of material by yourself**
  - Reading books
  - Reading papers
  - Practice on the web
- **Don’t expect to be told everything**

9

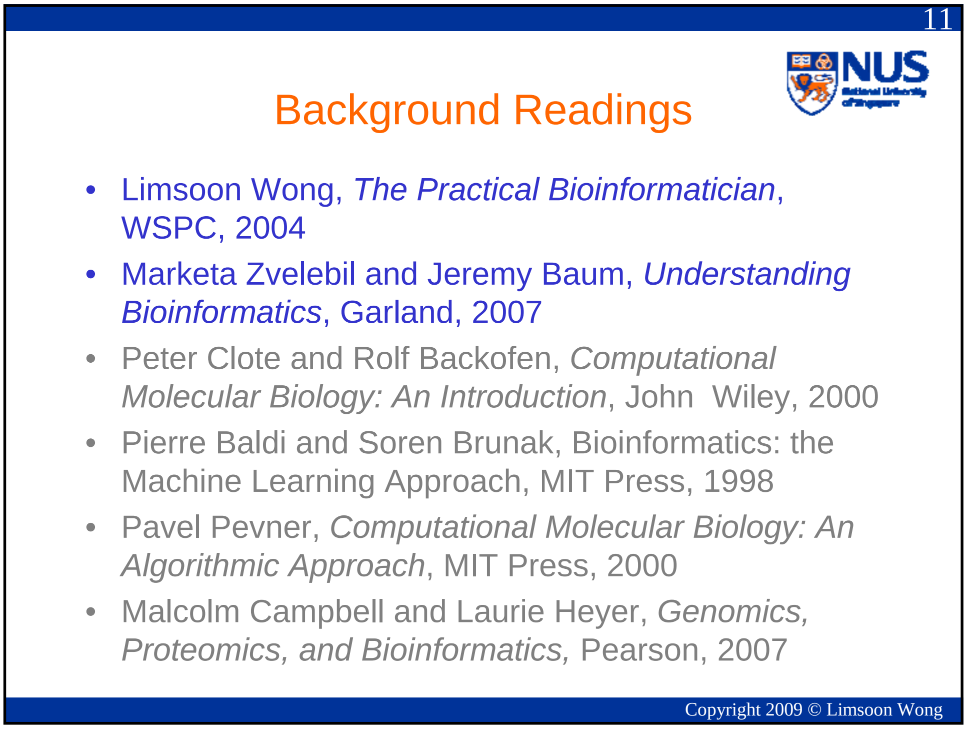
Assignments, Project, & Exam

- **Assignments**
  - Probably 3-4 assignments
  - Some are simple programming assignments
- **Project**
  - Based on a case study in the class
  - 8-10 pages of report expected
- **Exam**
  - 1 final open-book exam

10

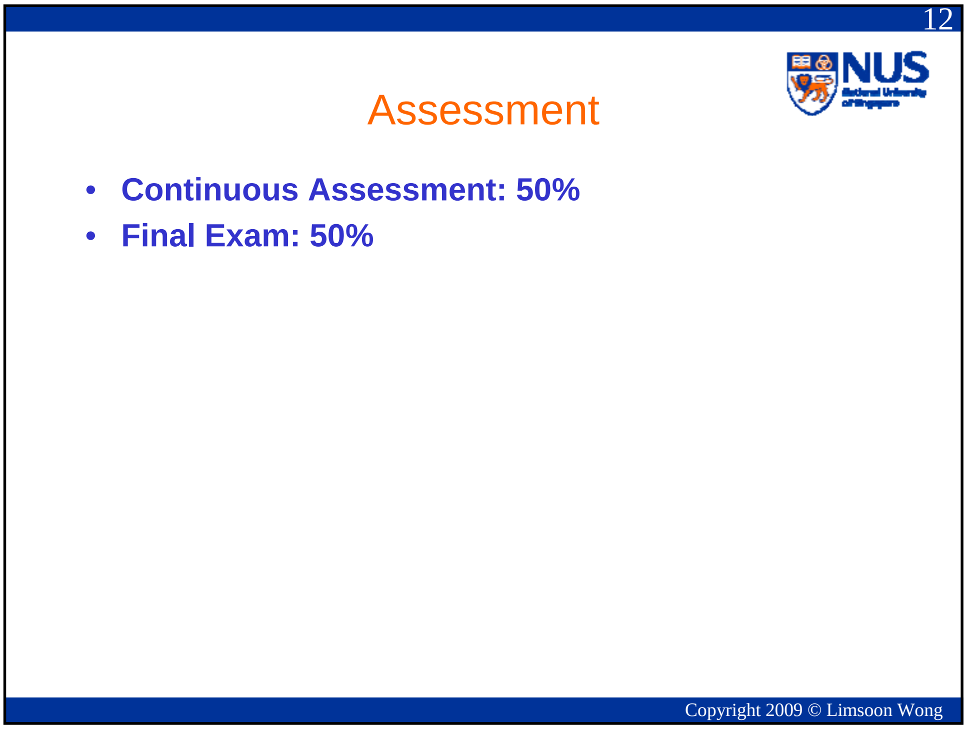
Be Honest

- **Exam**
  - Absence w/o good cause results in ZERO mark
  - Cheating results in ZERO mark
- **Discussion on assignments is allowed**
- **Blatant plagiarism is not allowed**
  - Offender gets ZERO mark for assignment or exam
  - Penalty applies to those who copied AND those who allowed their assignments to be copied



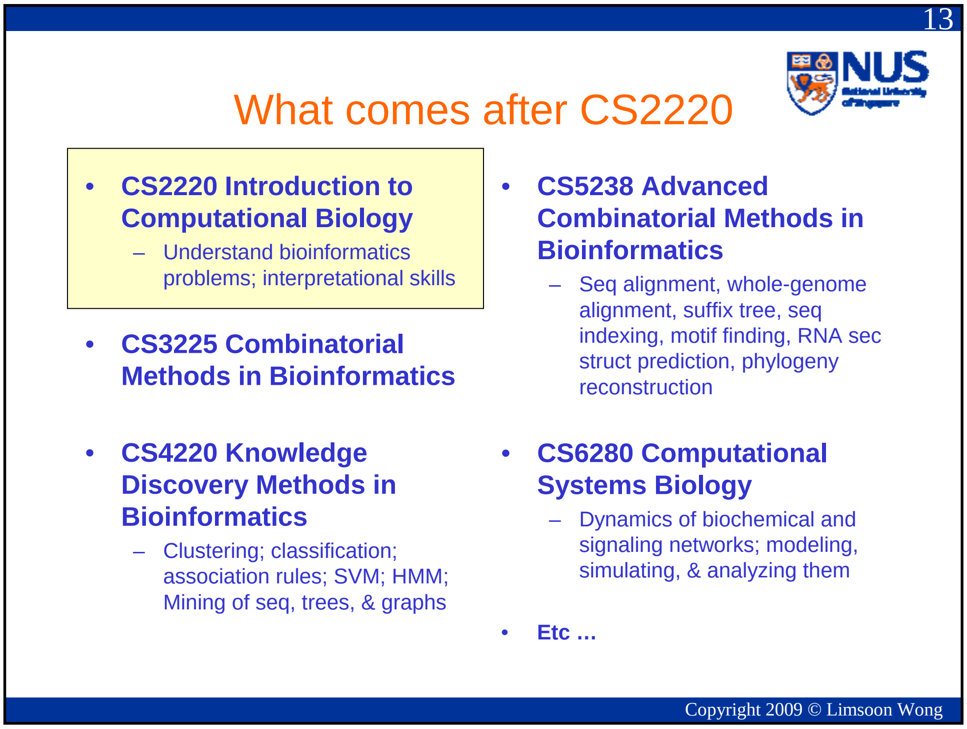
Background Readings

- Limsoon Wong, *The Practical Bioinformatician*, WSPC, 2004
- Marketa Zvelebil and Jeremy Baum, *Understanding Bioinformatics*, Garland, 2007
- Peter Clote and Rolf Backofen, *Computational Molecular Biology: An Introduction*, John Wiley, 2000
- Pierre Baldi and Soren Brunak, Bioinformatics: the Machine Learning Approach, MIT Press, 1998
- Pavel Pevner, *Computational Molecular Biology: An Algorithmic Approach*, MIT Press, 2000
- Malcolm Campbell and Laurie Heyer, *Genomics, Proteomics, and Bioinformatics,* Pearson, 2007



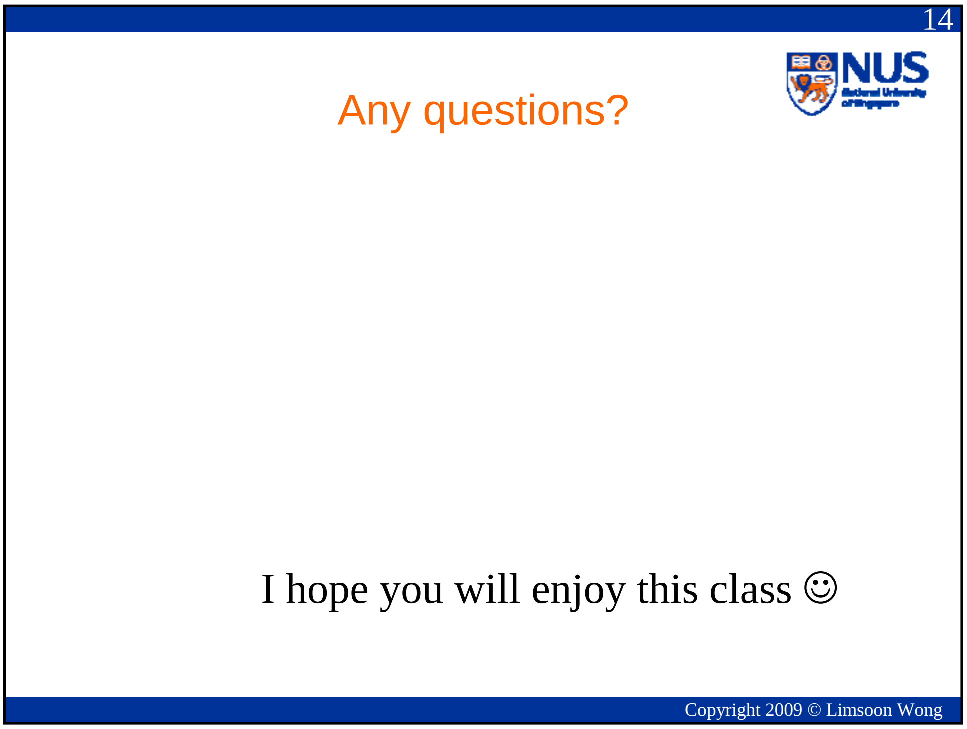
Assessment

- **Continuous Assessment: 50%**
- **Final Exam: 50%**



What comes after CS2220

- **CS2220 Introduction to Computational Biology**
  - Understand bioinformatics problems; interpretational skills
- **CS3225 Combinatorial Methods in Bioinformatics**
- **CS4220 Knowledge Discovery Methods in Bioinformatics**
  - Clustering; classification; association rules; SVM; HMM; Mining of seq, trees, & graphs
- **CS5238 Advanced Combinatorial Methods in Bioinformatics**
  - Seq alignment, whole-genome alignment, suffix tree, seq indexing, motif finding, RNA sec struct prediction, phylogeny reconstruction
- **CS6280 Computational Systems Biology**
  - Dynamics of biochemical and signaling networks; modeling, simulating, & analyzing them
- **Etc …**

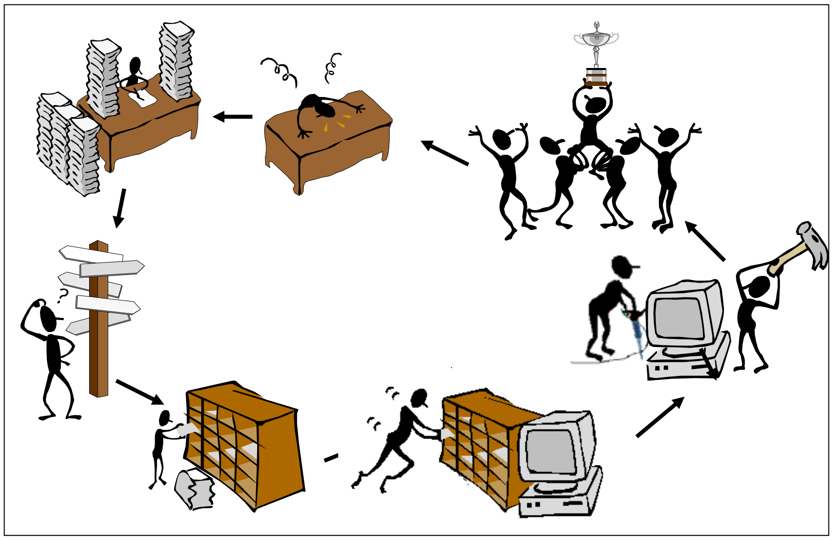


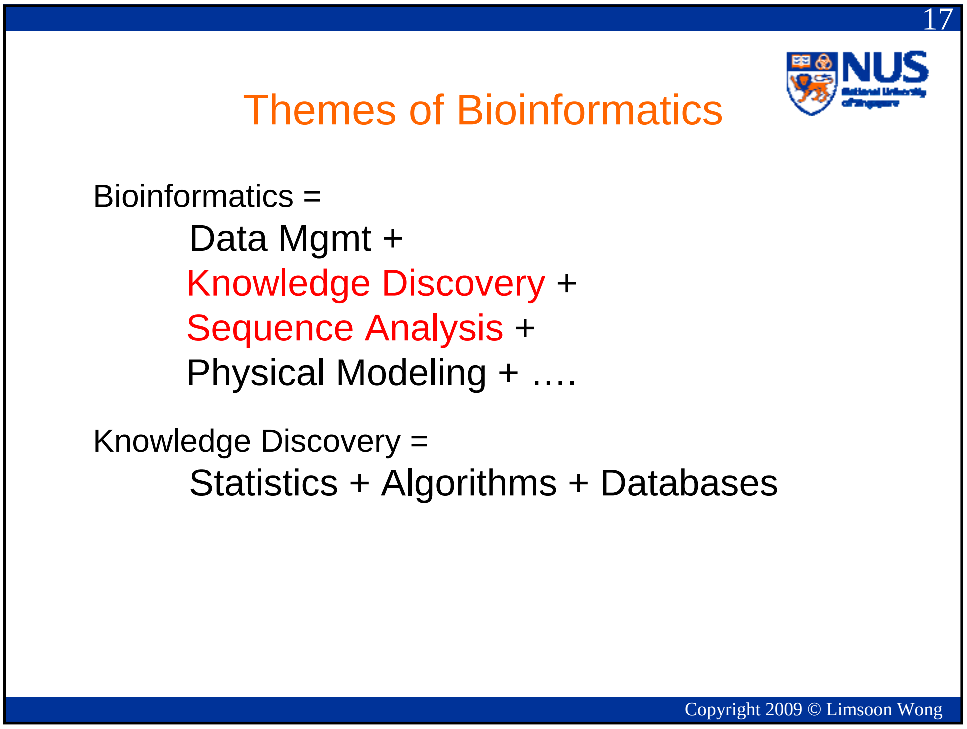
Any questions?

I hope you will enjoy this class ☺

Themes and Applications of Bioinformatics

What is Bioinformatics?





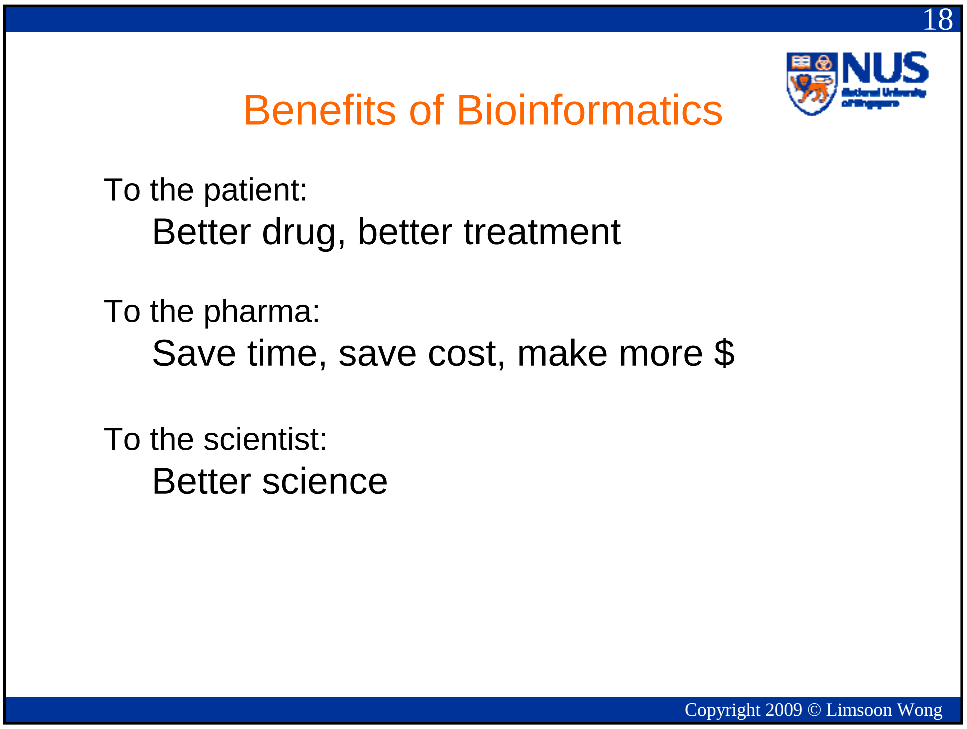
Themes of Bioinformatics

Bioinformatics =

Data Mgmt +
Knowledge Discovery +
Sequence Analysis +
Physical Modeling + ….

Knowledge Discovery =

Statistics + Algorithms + Databases



Benefits of Bioinformatics

To the patient:

Better drug, better treatment

To the pharma:

Save time, save cost, make more $

To the scientist:

Better science

Some Bioinformatics Problems

- Biological Data Searching
- Gene/Promoter finding
- Cis-regulatory DNA
- Gene/Protein Network
- Protein/RNA Structure Prediction
- Evolutionary Tree reconstruction
- Infer Protein Function
- Disease Diagnosis
- Disease Prognosis
- Disease Treatment Optimization, ...

Biological Data Searching

- Biological Data is increasing rapidly
- Biologists need to locate required info
- Difficulties:
  - Too much
  - Too heterogeneous
  - Too distributed
  - Too many errors
  - Due to mutation, need approximate search

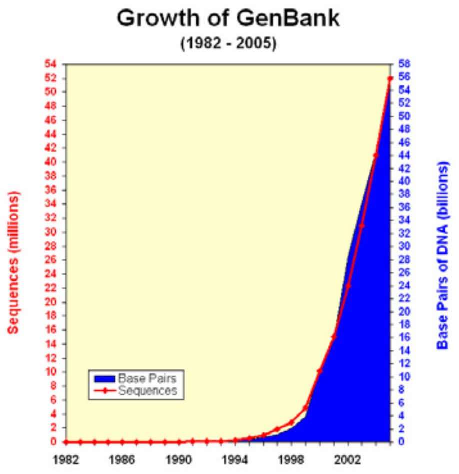


Image credit: NCBI

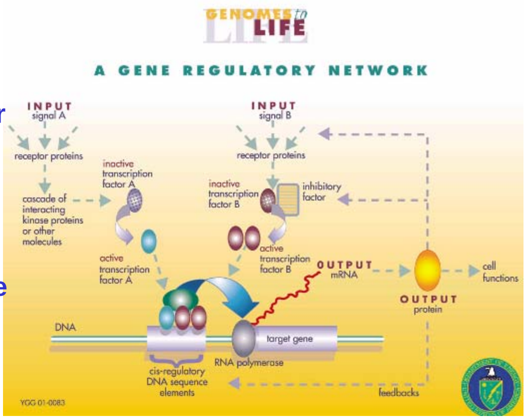
Cis-Regulatory DNAs

- Cis-regulatory DNAs control whether genes should express or not

- Cis-regulatory DNAs may locate in promoter region, intron, or exon

- Finding and understanding cis-regulatory DNAs is one of the key problem in coming years

Image credit: US DOE


GENOMES to LIFE
A GENE REGULATORY NETWORK
INPUT signal A
INPUT signal B
receptor proteins
receptor proteins
inactive transcription factor A
cascade of interacting kinase proteins or other molecules
inactive transcription factor B
inhibitory factor
active transcription factor A
active transcription factor B
OUTPUT mRNA
OUTPUT protein
cell functions
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks
YGG 01-0083

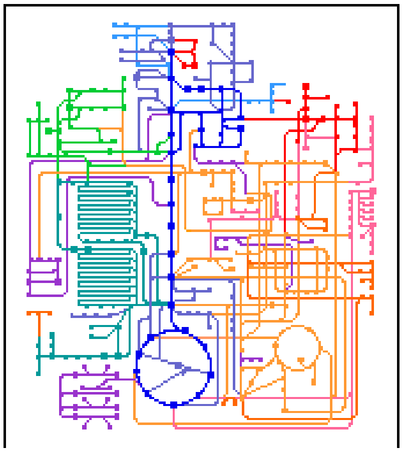
Gene Networks

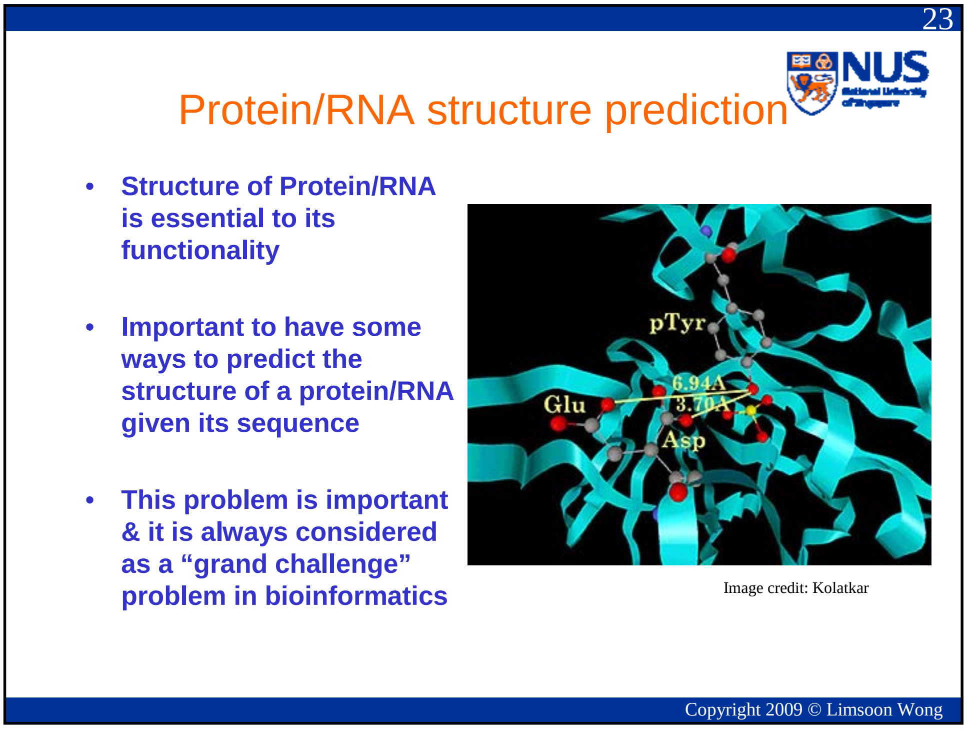
- Inside a cell is a complex system

- Expression of one gene depends on expression of another gene

- Such interactions can be represented using gene network

- Understanding such networks helps identify association betw genes & diseases

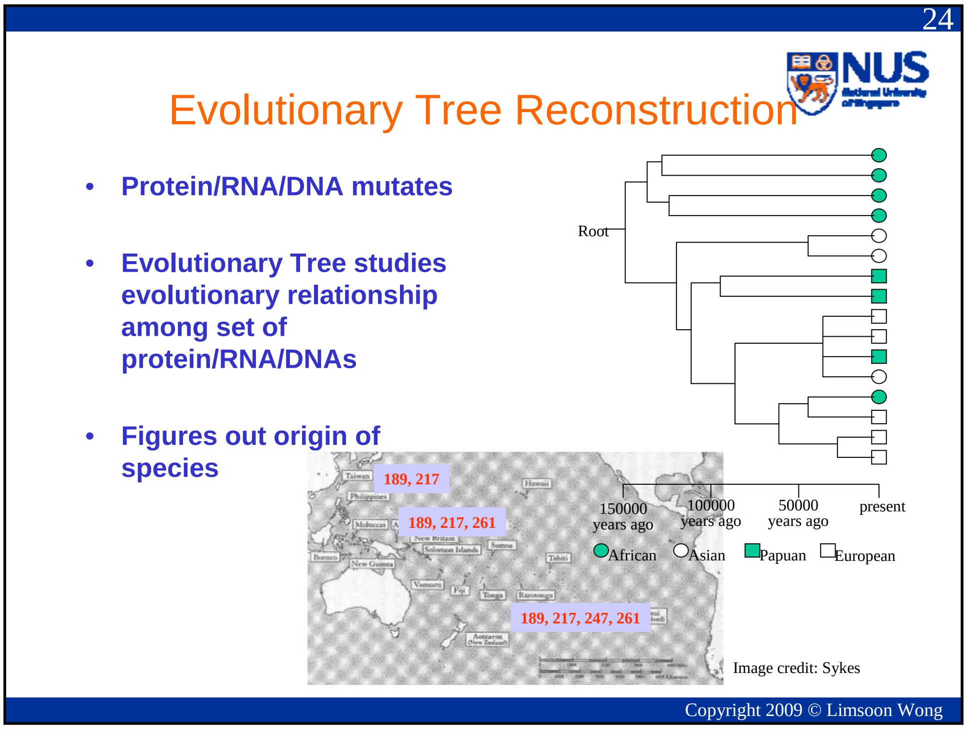




Protein/RNA structure prediction

- Structure of Protein/RNA is essential to its functionality
- Important to have some ways to predict the structure of a protein/RNA given its sequence
- This problem is important & it is always considered as a “grand challenge” problem in bioinformatics

Image credit: Kolatkar



Evolutionary Tree Reconstruction

- Protein/RNA/DNA mutates
- Evolutionary Tree studies evolutionary relationship among set of protein/RNA/DNAs
- Figures out origin of species

Image credit: Sykes

Breast Cancer Outcome Prediction

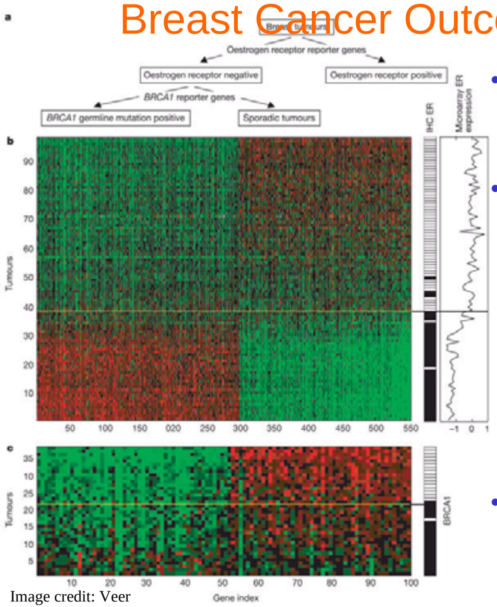
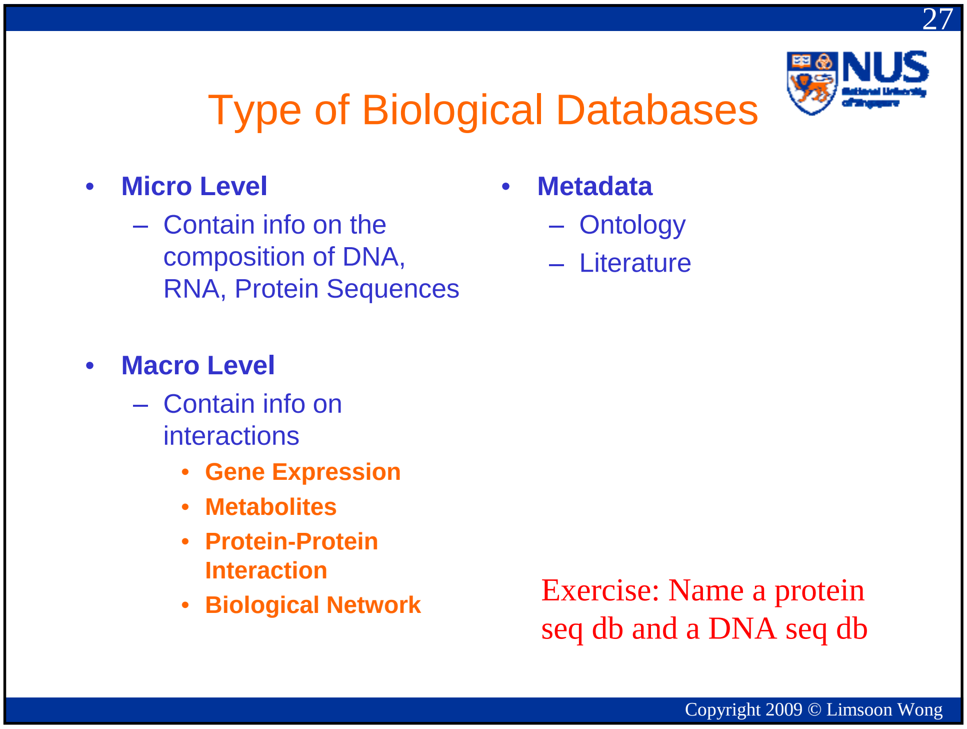


Image credit: Veer

- **Van't Veer et al., *Nature* 415:530-536, 2002**
- **Training set contains 78 patient samples**
  - 34 patients develop distance metastases in 5 yrs
  - 44 patients remain healthy from the disease after initial diagnosis for >5 yrs
- **Testing set contains 12 relapse & 7 non-relapse samples**

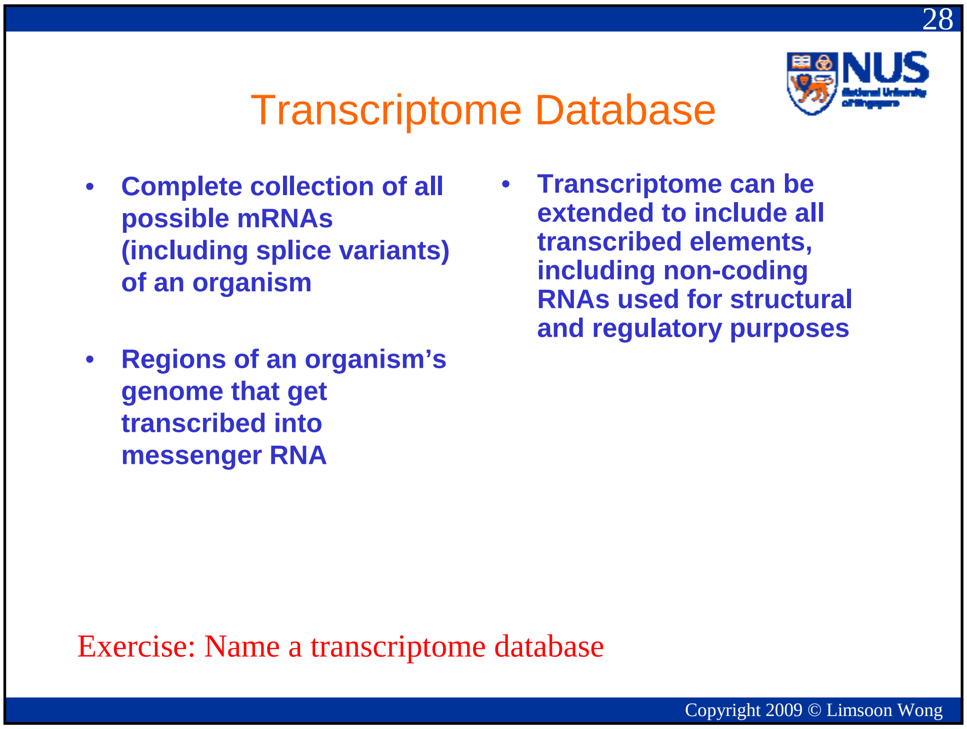
Commonly Used Data Sources



Type of Biological Databases

- **Micro Level**
  - Contain info on the composition of DNA, RNA, Protein Sequences
- **Macro Level**
  - Contain info on interactions
    - **Gene Expression**
    - **Metabolites**
    - **Protein-Protein Interaction**
    - **Biological Network**
- **Metadata**
  - Ontology
  - Literature

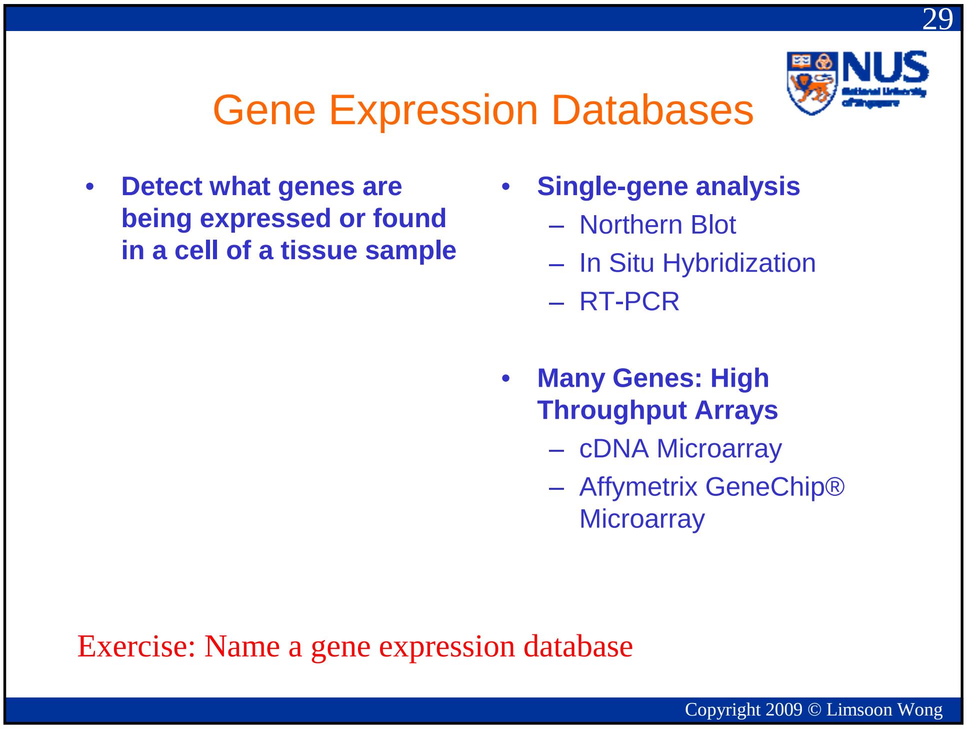
Exercise: Name a protein seq db and a DNA seq db



Transcriptome Database

- **Complete collection of all possible mRNAs (including splice variants) of an organism**
- **Regions of an organism's genome that get transcribed into messenger RNA**
- **Transcriptome can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes**

Exercise: Name a transcriptome database

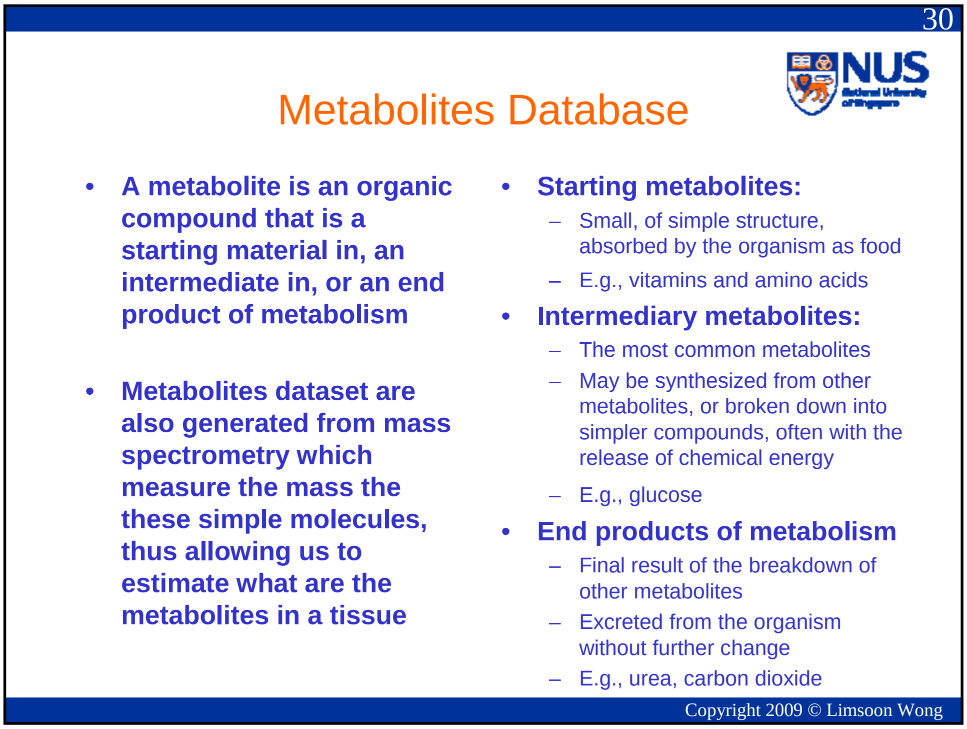


29

Gene Expression Databases

- **Detect what genes are being expressed or found in a cell of a tissue sample**
- **Single-gene analysis**
  - Northern Blot
  - In Situ Hybridization
  - RT-PCR
- **Many Genes: High Throughput Arrays**
  - cDNA Microarray
  - Affymetrix GeneChip® Microarray

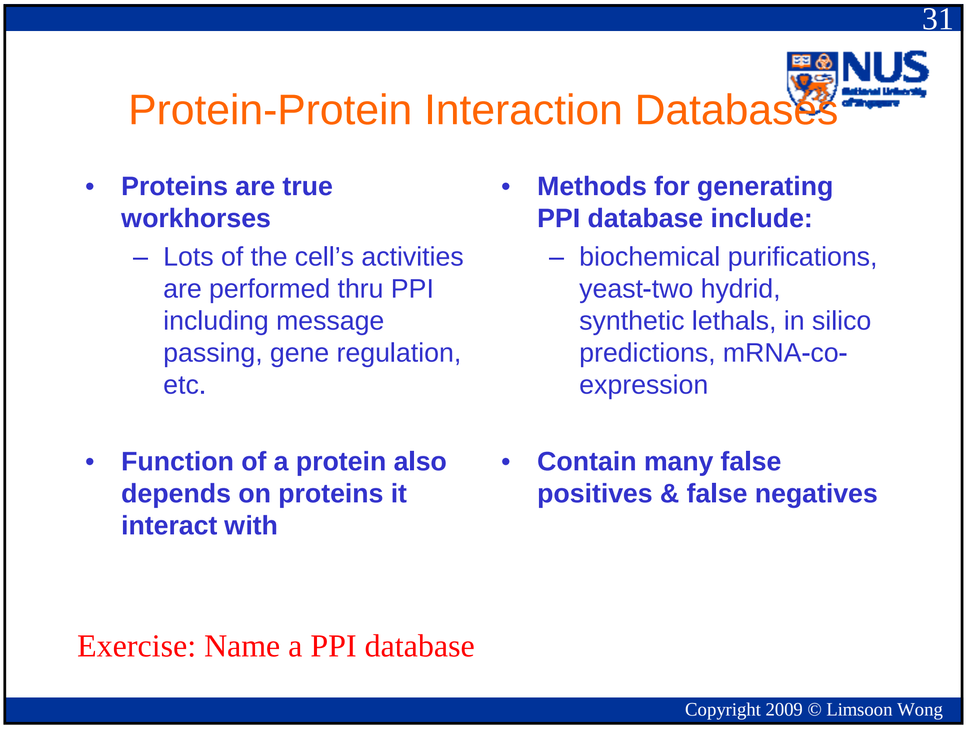
Exercise: Name a gene expression database



30

Metabolites Database

- **A metabolite is an organic compound that is a starting material in, an intermediate in, or an end product of metabolism**
- **Metabolites dataset are also generated from mass spectrometry which measure the mass the these simple molecules, thus allowing us to estimate what are the metabolites in a tissue**
- **Starting metabolites:**
  - Small, of simple structure, absorbed by the organism as food
  - E.g., vitamins and amino acids
- **Intermediary metabolites:**
  - The most common metabolites
  - May be synthesized from other metabolites, or broken down into simpler compounds, often with the release of chemical energy
  - E.g., glucose
- **End products of metabolism**
  - Final result of the breakdown of other metabolites
  - Excreted from the organism without further change
  - E.g., urea, carbon dioxide

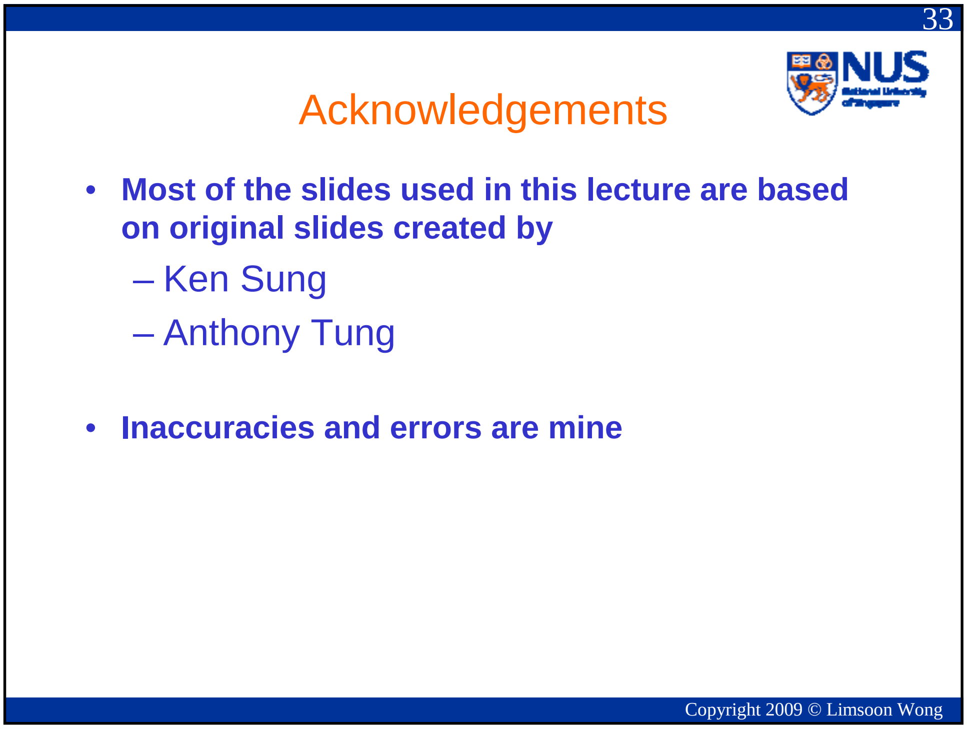


Protein-Protein Interaction Databases

- **Proteins are true workhorses**
  - Lots of the cell’s activities are performed thru PPI including message passing, gene regulation, etc.
- **Function of a protein also depends on proteins it interact with**
- **Methods for generating PPI database include:**
  - biochemical purifications, yeast-two hydrid, synthetic lethals, in silico predictions, mRNA-co-expression
- **Contain many false positives & false negatives**

Exercise: Name a PPI database

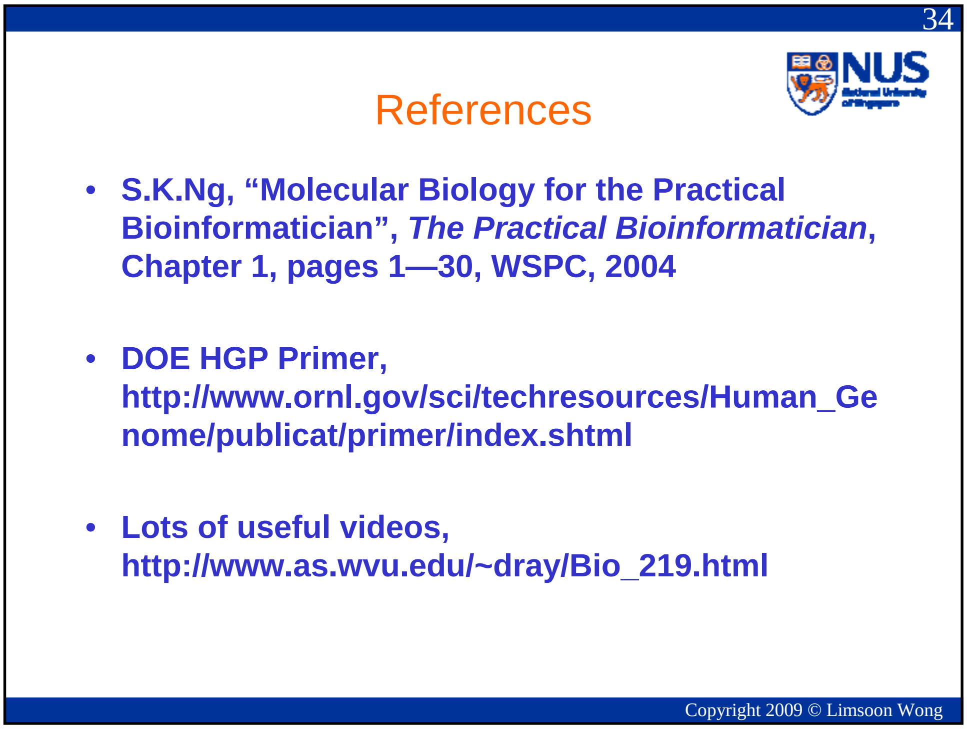
Any Question?



33

Acknowledgements

- **Most of the slides used in this lecture are based on original slides created by**
  - Ken Sung
  - Anthony Tung
- **Inaccuracies and errors are mine**



34

References

- **S.K.Ng, “Molecular Biology for the Practical Bioinformatician”, *The Practical Bioinformatician*, Chapter 1, pages 1—30, WSPC, 2004**
- **DOE HGP Primer, http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/index.shtml**
- **Lots of useful videos, http://www.as.wvu.edu/~dray/Bio_219.html**