Gene Expression & Annotation

Clare Foyle, Nick Fieller & AstraZeneca R & D

University of Sheffield & Alderley Park

October 2005

Outline

- Basic Biology:–
  - DNA, Genes, Proteins, ….
- Measuring Gene Expression
- Gene Expression Data
- Annotation
- Mesh Terms
- MeSH DAG
- Quantifying Annotation
- ……

Basic Biology

- Genes ⇒ Proteins ⇒ Bioactivity
  - bioactivity includes disease (e.g. cancer)

  Aim is to design drugs to interfere with bioactivity by interrupting pathways of protein activity
- DNA: [linear] sequence of ~3 billion base pairs of nucleotides (C G A T)
  - Genes: subsequence of few hundred base pairs
    - recognised by key start & end sequences
    - genes don't vary [much] between people
    - most base pairs are identical across the population
  - ~30,000 genes in total
- ~ 10% of DNA consists of genes ('apparently')

Genes ⇒ Proteins

- Mechanism involves genes '*expressing*' RNA which in turn 'constructs' proteins
- Genes do not 'express' continuously but only in relation to biological events
  - i.e. if 'switched on'

⇒ **key interest: –** determine which genes are (or are not) expressed in various circumstances

- e.g. cancer tissue may have different gene expression profile from other types of tissue
- differences may reveal processes involved in that type of cancer

Measuring Gene Expression

- 3 Basic technologies:
  - Affymetrix® GeneChips® oligonucleotide assays
    - expression levels of each gene in sample
  - Red/green spot assays
    - differential expression levels in 2 samples
  - Radionucleotide assays
    - differential expression levels in 2 samples

  –All involve 'labelling' RNA from tissue
    - e.g. fluorescent markers

Affymetrix GeneChips

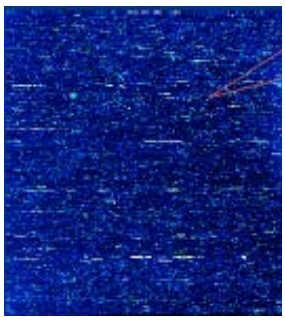
- Sequences of [all] genes is known
  - Can be manufactured to order
    - Just need to stick individual molecules of C G T & A together in right order
    - possible to do with very great precision
  - Subsequence of ~25 bases unique within the genome to that gene is a ***Probe Set*** for that gene
- Affymetrix GeneChips have many thousands of probe sets placed at known coordinates on a chip

7

- Chip has single strands of DNA probe sets
- Tissue containing mRNA processed by reverse transcription to cDNA incorporating fluorescent marker
  - (only contains DNA from 'expressing genes')
- Denatured to produce a single strand
  - (from the double strand helix)
- Solution flushed over chip and strands in sample bind to single strands of probe set
  - fluorescence measured

⇒ measures of expression of genes in sample

8



Affymetrix chip

intensity of fluorescence at each location indicates level of gene expression

9

Points to Consider:–

- Each chip has several copies of each probe set
  - result is mean and variance of measures
    - Affymetrix software provides mean level & a p-value
    - (Better ways of initial processing widely investigated)
- May be several probe sets for single gene
  - e.g. subsequences from beginning/middle/end
  - may happen that only some probe sets from a gene will indicate expression (& not others)
- genes ⇒ proteins
  - but single gene may produce several different proteins under different circumstances
  - different genes may produce same protein

10

Summary so far

- Gene expression data:–
  - typically 'gene expression data sets' have relatively few subjects, little or no replication and quantitative measures of many thousands of gene [or probe sets]
  - data are typically of 'poor quality'
    - many outliers, high variability, 'contamination'
  - data are 'expensive' to obtain per person
  - each chip ~$100
  - one chip can be used for only one sample

⇒ data sets typically are "n < < p" with no replication but there may be covariates

11

Need some dimensionality reduction

- Most common is to proceed in 2 stages:
  - initial selection of 'interesting genes'
  - standard dimensionality reduction on these
  - Stage 1:
    - alternative 1: select the N genes with highest expression levels
    - alternative 2: select the N genes with greatest inter-group variability
      - groups defined by covariates (e.g. cancel 'level')
        - » *c.f. "differential expression"*
      - N=500 and N=100 were tried
  - Stage 2: PCA, LDA, PLS,……

12

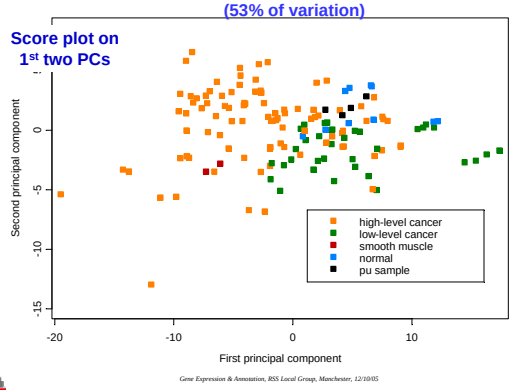
- **example:–**
  - ~60 cancer subjects (+ some controls),
    - ~13,000 probe sets (coded)
    - one sample from each, measured twice,
    - (note these are not full replicates)
  - + background information sex/severity/age/…
- Aside: 3rd party confidentiality meant we had only code numbers for probe sets
  - no information on which probe sets were from same gene
  - inhibits some avenues of analysis

13

PCA on 100 top expressing genes

(53% of variation)

Score plot on 1st two PCs

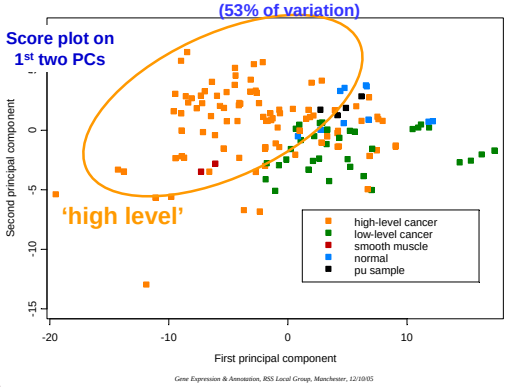


14

PCA on 100 top expressing genes

(53% of variation)

Score plot on 1st two PCs

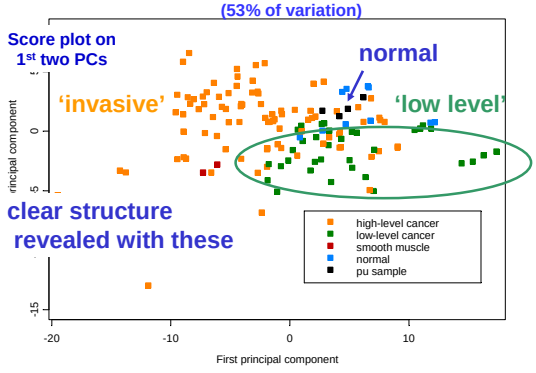


15

PCA on 100 top expressing genes

(53% of variation)

Score plot on 1st two PCs



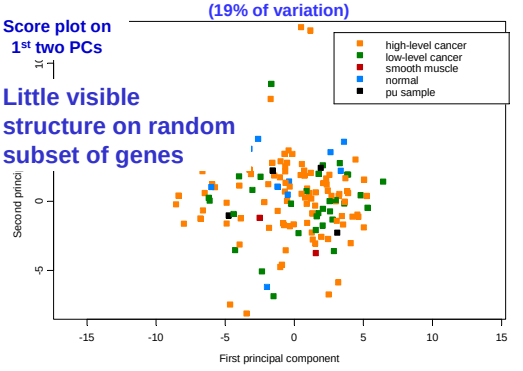
16

PCA with 100 randomly selected genes

(19% of variation)

Score plot on 1st two PCs

Little visible structure on random subset of genes

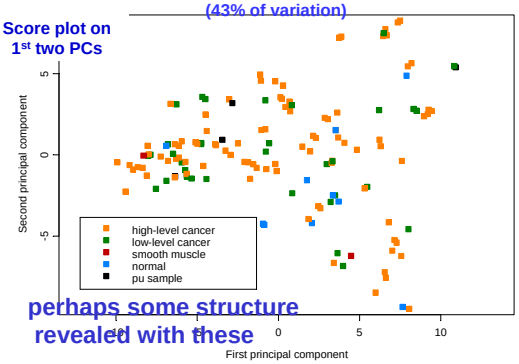


17

PCA on 100 'most variable' genes

(43% of variation)

Score plot on 1st two PCs



perhaps some structure revealed with these

18

- Analysis with top 100 'most variable' gives some structure on 1st two PCs but not as clear as top expressing
  - 3rd & 4th PCs needed (+16% variation)
    - (not shown)
- Analysis with 500 genes gives less clear results on high order PCs than with only first 100
  - (not shown)
- Similar pictures and conclusions from other categorizations of samples
  - e.g. demographic features
    - (not shown)

19

- Results suggest that initial selection methods capture at least some of available information in genes
  - *(actually probe sets in this case)*
  - suggests may be able to reduce variability by 'averaging' over subjects with common covariates — perhaps literally or by modelling
- What else is known?

ANNOTATION

20

ANNOTATION

- 'Annotation' — information on genes
  - typically text, possibly published, free form
  - position on chromosome
  - role in intracellular pathway
  - citation in published paper or tech report
  - ……
  - ……
- may be different types & quantities of information on different genes
  - very extensive

21

- Annotation clearly important in **follow-up** to analysis when 'statistically important' genes have been identified
  - in which pathway does it have a role?
  - which proteins does it 'produce'?
  - ……
- Can annotation be used in the analysis?
  - requires some form of ***quantification*** of the annotation information
    - *(I presume…)*

22

- **Initial ideas on annotation**

structural information on genes (annotation)

structural information on samples (stage, demographics etc)

Sample × Gene data matrix of expression levels

23

- **Initial ideas on annotation**

inherent symmetry: Gene × Sample or Sample × Gene matrix

structural information on genes (annotation)

structural information on samples (stage, demographics etc)

Gene × Sample data matrix of expression levels

24

- Initial ideas on annotation

invites iterative analysis of genes × samples & samples × genes to update selection of genes related to structure (work in progress)

can also compare structure with annotation on randomly selected genes

Gene × Sample data matrix of expression levels

25

- A simple approach
  - » *(simplistic?)*
  - underlying idea is the genes associated with same [or related] 'biofunctions' are themselves 'related'
  - start by looking for co-citations of genes with specific biofunctions
    - i.e. how many papers refer to both gene X and biofunction A
    - need extensive text mining literature search
    - unrealistic to try for all 13,000 genes so restrict [initially] to 'preselected interesting' ones

26

- Next step:
  - determine which groups of genes are linked to same biofunction term
    - regard these as 'related' genes
- Problem:
  - very sparse information
    - SO, link together biofunctions
    - genes linked to 'related biofunctions' are [distantly] related
- Finally
  - adjust gene × sample analysis by 'averaging' over related genes
    - c.f. 'averaging' over related samples

27

- biofunctions???
  - need definitive list
  - need declared hierarchy to provide relationships
- Two (+) possibilities
  - GO
    - **G**ene **O**ntology Consortium
      - » http://www.geneontology.org/
  - MeSH
    - **Me**dical **S**ubject **H**eadings
      - US National Library of Medicine
        - » http://www.nlm.nih.gov/mesh/meshhome.html

28

- biofunctions???
  - need definitive list
  - need declared hierarchy to provide relationships
- Two (+) possibilities
  - GO
    - **G**ene **O**ntology Consortium
      - » http://www.geneontology.org/
  - → MeSH ←
    - **Me**dical **S**ubject **H**eadings
      - US National Library of Medicine
        - » http://www.nlm.nih.gov/mesh/meshhome.html

29

- MeSH
  - **Me**dical **S**ubject **H**eadings
    - US National Library of Medicine
      - » http://www.nlm.nih.gov/mesh/meshhome.html
  - Terms under 15 broad headings organised in hierarchical tree structure (DAG)
  - Selected 2 headings
    - Diseases
    - Biological Sciences
  - scanned literature for co-citation of each of our 'interesting' genes with each term under these headings
  - i.e. two [overlapping] sets of 500 of probe sets
    - » (< 500 distinct genes)
  - ~6000 terms

30

MeSH Tree Structures - 2005

Return to Entry Page

1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊞ Diseases [C]
4. ⊞ Chemicals and Drugs [D]
5. ⊞ Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. ⊞ Psychiatry and Psychology [F]
7. ⊞ Biological Sciences [G]
8. ⊞ Physical Sciences [H]
9. ⊞ Anthropology, Education, Sociology and Social Phenomena [I]
10. ⊞ Technology and Food and Beverages [J]
11. ⊞ Humanities [K]
12. ⊞ Information Science [L]
13. ⊞ Persons [M]
14. ⊞ Health Care [N]
15. ⊞ Geographic Locations [Z]

Return to Entry Page

MeSH Tree Structures - 2005

Return to Entry Page

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
   - Bacterial Infections and Mycoses [C01] +
   - Virus Diseases [C02] +
   - Parasitic Diseases [C03] +
   - Neoplasms [C04] +
   - Musculoskeletal Diseases [C05] +
   - Digestive System Diseases [C06] +
   - Stomatognathic Diseases [C07] +
   - Respiratory Tract Diseases [C08] +
   - Otorhinolaryngologic Diseases [C09] +
   - Nervous System Diseases [C10] +
   - Eye Diseases [C11] +
   - Urologic and Male Genital Diseases [C12] +
   - Female Genital Diseases and Pregnancy Complications [C13] +
   - Cardiovascular Disea
   - Hemic and Lymphatic

&c., &c., &c…… until

2005 MeSH

MeSH Descriptor Data

Return to Entry Page

MeSH Heading	Gift Giving
Tree Number	F01.145.813.208
Scope Note	The bestowing of tangible or intangible benefits, voluntarily and usually without expectation of anything in return. However, gift giving may be motivated by feelings of ALTRUISM or gratitude, by a sense of obligation, or by the hope of receiving something in return.
Entry Term	Financial Contributions
Entry Term	Gifts, Financial
Entry Term	Giftgiving
Entry Term	Gifts
See Also	Altruism
See Also	Charities
See Also	Financial Support
See Also	Fund Raising
Allowable Qualifiers	ES
History Note	2003; for GIFTS, FINANCIAL use FUND RAISING 1978-2002
Unique ID	D037921

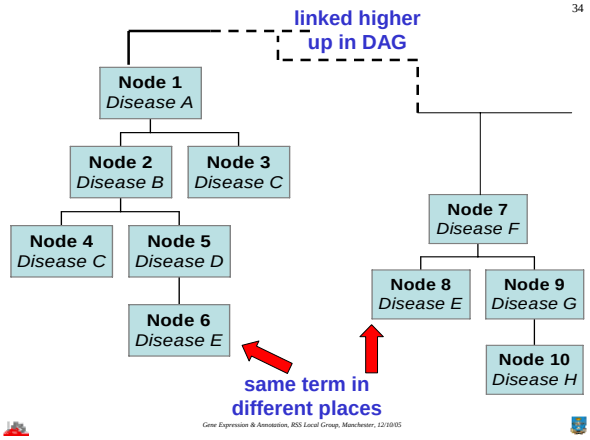
- Some MeSH terms can appear in several places in the DAG

Parasitic Diseases [C03]
Helminthiasis [C03.335]
Nematode Infections [C03.335.508]
Adenophorea Infections [C03.335.508.100] +
Larva Migrans [C03.335.508.523]
Larva Migrans, Visceral [C03.335.508.523.780]
Secernentea Infections [C03.335.508.700] +

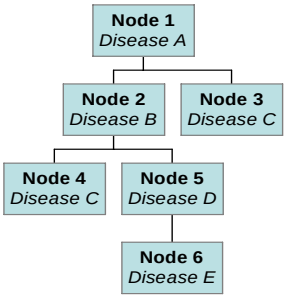
Parasitic Diseases [C03]
Skin Diseases, Parasitic [C03.858]
Larva Migrans [C03.858.424]
Leishmaniasis [C03.858.560] +
Lice Infestations [C03.858.580]
Scabies [C03.858.800]

Skin and Connective Tissue Diseases [C17]
Skin Diseases [C17.800]
Skin Diseases, Infectious [C17.800.838]
Skin Diseases, Parasitic [C17.800.838.775]
Larva Migrans [C17.800.838.775.424]
Leishmaniasis [C17.800.838.775.560] +
Lice Infestations [C17.800.838.775.580]

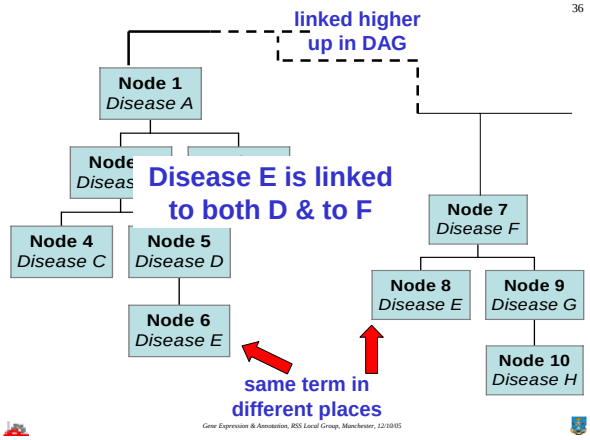
numbers give sequence of nodes in DAG



- Adjacent terms in DAG are 'linked'



- Disease D is linked to both B and E
- 'Up-propagation' involves combining nodes with categories one level higher in DAG
- complication with multi-occurrence of terms



37

Algebraic Manipulations

- Annotation information, ***A***, *n x p* matrix
  - with *n* probesets and *p* annotation terms.
  - gives number of associations found linking an annotation term to a probe set
  - **convert to binary form:**

1 link between annotation & probe set

0 is if no link has been found

  - ***(as yet!)***

38

- $\boldsymbol{AA^T}$ gives the co-occurrence (or 'similarity') matrix of the probe sets
  - Typically large & sparse
    - provides few links between genes of interest
    - to provide further (but weaker) links up-propagation of the MeSH DAG is used to establish further relationships
  - Initially, genes linked via mother-daughter pairs of descriptors can be determined
    - Repetition of process allows successively weaker links to be determined via successively less closely related descriptors

39

- **Up-propagation:**
  - Up-propagation matrix, ***P***
    - *p x p* matrix of the DAG
      - indicates which nodes are directly linked
  - $\boldsymbol{AA^T}$ gives probe sets directly linked via co-occurrence with a single MeSH term
  - $\boldsymbol{APA^T}$ gives the probe sets related by links to one-generation related MeSH descriptors
  - $\boldsymbol{AP^rA^T}$ gives those linked via r-generation related descriptor

40

- **Issues:–**
  - Probe Sets / Genes
    - data in probe sets, annotation re genes
      - (+ have to go through 3rd party for annotation on probe sets)
  - multiple occurrence of MeSH terms
  - both require 'fiddling' with rows/columns in matrix manipulation
  - large sparse matrices (10,000 × 500)
    - computational care required
  - co-occurrence may be a false measure
    - "This gene has nothing whatsoever to do with that disease"
  - only have annotation for genes which are 'interesting' & preselected – not on all genes
  - ……

41

- http://www.shef.ac.uk/paspgr/foyle
- http://www.shef.ac.uk/nickfieller