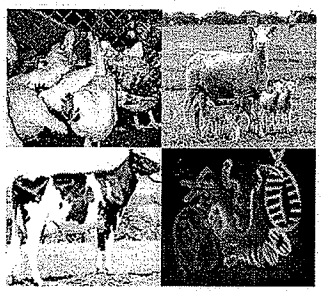
Gene discovery



Karen Marshall

UNE

IAEA, Korea, April, 2006

Gene Mapping

Ultimate aim is to identify gene & functional mutation

- Structural / functional studies on gene product → knowledge of biochemical pathway controlling trait of interest
- Functional mutation is a 'perfect marker' → GAS
- Significant resources and time are required, particularly for continuous traits

Gene Mapping

An alternative end point for MAS

- Haplotype spanning 1-5 cM
- Information content can be similar to that of a direct marker, depending on extent of linkage disequilibrium

The steps

1cM → ~1 million bp containing ~10 genes

Unknown location to ~20cM region
- achievable via 'broad-scale' linkage mapping

~20cM region to <2cM region
- various approaches, including LD mapping
- usually require significant animal resources
- possibly the most difficult step

<2cM region to gene and functional mutation
- positional candidate and other approaches
- may need to sequence through large regions for a number of animals

Strategies for refining region from 20cM → 2cM

Fine-scale linkage mapping

Linkage disequilbrium mapping (also linkage disequilibrium – linkage mapping)

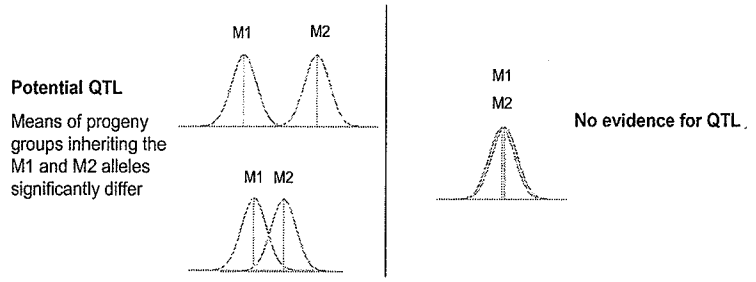
Multi-generational QTL mapping e.g. targeted recombinant progeny

Gene discovery

Fine-scale linkage mapping

Basis of linkage mapping

- As an example consider a half-sib mapping design and single marker analysis



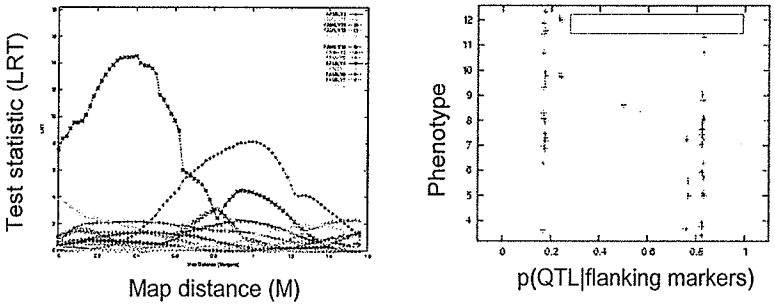
Trait distribution of progeny inheriting either a M1 or M2 allele from sire

Gene discovery

Fine-scale linkage mapping

Basis of linkage mapping

- More usual implementation is interval mapping via maximum likelihood or regression



Gene discovery

Fine-scale linkage mapping

Basis of 'fine-scale' linkage mapping

- as for linkage mapping but with additional markers within the region of interest

- aim is to reduce the confidence interval (CI) of the position estimate
- CI calculated from LRT: decrease in one LOD score either side of best position → 96.8% CI

Gene discovery

Fine-scale linkage mapping

Reality

- At some point increasing marker density will not refine the QTL position
- This is because very large half-sib families are required to generate recombinants between closely spaced markers
  - e.g. markers 2cM apart → only 20 out of 1000 progeny are recombinant within this region

Gene discovery

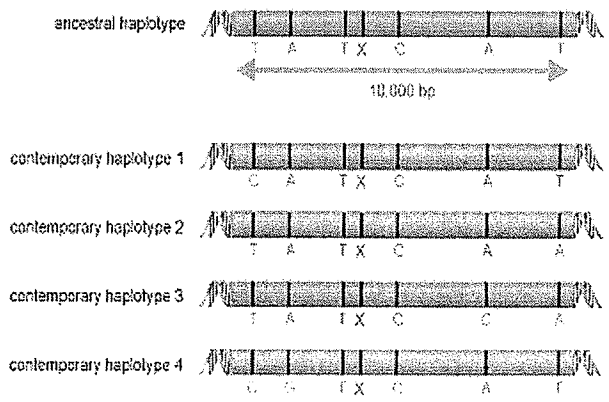
Linkage disequilibrium (LD) mapping

Basis

- Linkage mapping considers the linkage disequilbrium that exists *within families*
- LD mapping considers the linkage disequilbrium that occurs across the *entire population.* Pure LD mapping disregards pedigree structure
- *LD and linkage mapping can be combined → LDLA mapping*

Gene discovery

LD mapping



Gene discovery

Linkage disequilibrium (LD) mapping

Basis

- For LD to occur across the entire population, and not be broken down over generations, the QTL and marker must be closely linked
- LD mapping is applicable to
  - region of ~20cM or less (i.e. in LD) for sheep / cattle
  - historical data, where analysis is performed over generations
  - industry data, where analysis is performed over families
  - half-sib data, if QTL is assumed to be segregating in dams

Gene discovery

Linkage disequilibrium (LD) mapping

Reality

- Powerful method, although merit of *linkage* vs *LD* vs *LDLA* depends on underlying extent of LD / mutation age and data structure, and continues to be evaluated by simulation
- Only recent move to storage of DNA from breeding animals / experimental flocks, thus historical pedigree and phenotypes may be available but DNA is often not
- Successfully used to refine QTL positions e.g.
  - QTL for milk traits refined to 3cM by LD
  - QTL for twinning rate refined to <1cM by LDLA
  - numerous QTL in human literature

Gene discovery

Targeted Recombinant Progeny

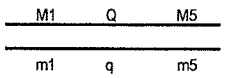
Basis

- Essentially 'multi-generational QTL mapping' but optimised to reduce genotypes / phenotypes
- Steps
  - Produce many progeny from a heterozygous sire
  - Identify those individuals that are recombinant within the region of interest
  - Progeny test these individuals to determine if segregating for the QTL
  - Determine QTL location via 'breakpoint analysis'

Heifetz, Fernando and Soller 7th WCGALP

Gene discovery

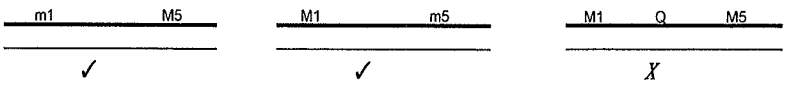
G0H



Sire known to be heterozygous for the qTL

G1 (G0H x dams)

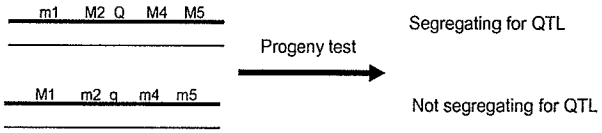
Identify sons recombinant in regions of interest (G1R)



G2 (G1R x dams)

Progeny test recombinant individuals to determine if segregating for QTL

Also type these individuals with additional markers within the region of interest



Gene discovery

Breakpoint analysis

Sire QTL phase (G0H)	M1	M2	M3	M4	M5	M6	M7	M8	M9	
q	A	A	A	A	A	A	A	A	A	
Q	B	B	B	B	B	B	B	B	B	
										Segregating
Sons (G1R)	A	-	B	B	-	B	B	B	B	Y
	A	B	B	B	-	B	B	B	B	Y
	A	-	A	B	-	B	B	B	B	Y
	A	-	A	A	A	A	A	B	B	N
	A	-	A	A	A	A	A	A	B	N
	B	B	A	A	-	A	A	A	A	N
	B	-	B	A	-	A	A	A	A	N
	B	B	B	B	B	A	A	A	A	N
	B	B	B	B	B	B	B	A	A	Y
	B	B	B	B	-	B	B	A	A	Y
	B	B	B	B	B	B	B	B	A	Y
										Region in common

Gene discovery

Targeted recombinant progeny

Reality

- Powerful method, but need sufficient G1R
  - 200 G1 progeny gives ~20 recombinants in 10cM region, half of which are male
- No need to progeny test if G1R individuals can be classified as segregating or not on their own phenotype
- Long time-line (several years)
- Success
  - Carwell narrowed to <1cM region using 8 G1R sires (AgResearch)
  - Callipyge (used own phenotype on G1R)

Gene discovery

From 20cM → 2cM Approach taken depends on

Animal resources

- for fine-scale mapping, the number of required animals is large
- industry / historical data can be used if DNA is available

Map and markers

- new markers within the region of interest will likely be required

Time-scale

- different approaches have different time-scales

Gene discovery

From <2cM to gene and functional mutation

General gene identification strategies

- Positional cloning
  - Uses knowledge of the mapped location of the gene
- Functional cloning
  - Uses knowledge of the protein encoded by the gene
- Candidate gene
  - Gene identified as good candidate
- Approaches can be taken in combination
  - "Positional candidate" approach

Gene discovery

Positional candidate approach

Basis: Sequence comparisons

- Search public databases for genes within region of interest via comparisons to human and other sequences
- Expect to find a number of genes, some of which may be candidates
  - ~10 genes / cM (human)
  - candidates identified after literature search for evidence of gene involvement in trait biology

Gene discovery

Positional candidate approach

Reality: Sequence comparisons

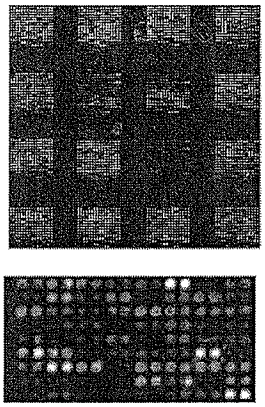
- Better success with smaller confidence interval about QTL
- Better success if homologous region in human (or other) sequence is well defined
- Success:
  - Inverdale QTL spanning ~10cM region to gene (BMP15) on positional candidate basis
  - Milk composition QTL spanning ~3cM to gene (DGAT-1) on positional candidate basis

Gene discovery

Positional candidate approach

Basis: Expression data

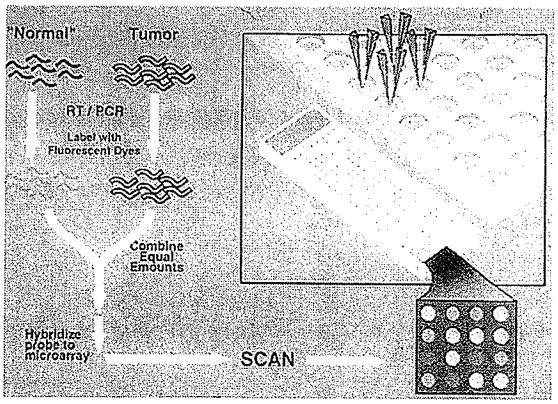
- Detection of differential expression of genes in 'Q' versus 'q' animals
- If differentially expressed gene is then mapped to region of interest → positional candidate
- Popular current technology is microarrays



Microarray showing differential display

Gene discovery

Microarrays



From, Bioinformatics course notes, J. McEwan, AABSS 2004

Positional candidate approach

Reality: expression data

- Microarrays generally produces numerous candidates, a number of which may also be positional, but
  - Arrays are expensive, therefore mRNA often represents tissue from single space / time
  - Not all genes in region may be represented in array
  - Genes identified in microarray may be 'downstream'
    e.g. mutation in gene A (the QTL) causes change in expression level of gene B (detected in the microarray analysis), can only trace back to gene A if pathway knowledge exists

Positional approach

If genes within region of interest are unknown (and thus no candidates)

- Obtain clones of genomic DNA within region of interest
- Identify which of these contain protein coding (gene) sequences by various molecular techniques
  - zoo blots
  - exon trapping
  - cDNA selection (also achieves point below)
- Attempt to determine which of these represents the gene of interest
  - for example, test clones for hybridisation to cDNA derived from cells known to express the mRNA of interest

Gene discovery

Identifying functional mutation

Identifying the functional mutation is usually required as 'proof' that the candidate gene is actually the gene of interest

Achieved by

- sequence 'Q' and 'q' individuals and look for mutations
- predict whether mutation will make a functional difference
- confirm by e.g. sequencing different populations, transgenic studies

Gene discovery

Functional mutation

Note:

Not all mutations result in an altered phenotype

- Mutations may be silent, cause a conserved amino acid substitution, or alter a non-critical part of the protein

Not all functional mutations are in protein coding sequences

- Mutation may be in regulatory region

Gene discovery: is it successful?

Numerous examples in literature for discrete traits

- Identification of the Inverdale gene was described as

"the culmination of many years of research involving breeding and segregation studies, genetic linkage mapping, physiology, molecular biology and comparative links to studies in humans and mice" (Galloway et al., 2001)

Limited examples for quantitative traits

- QTL for milk yield (DGAT1 gene) identified by positional cloning approach (Grisart et al., 2002)

Timeframe and resources is substantial

- e.g. DGAT1 seven + years